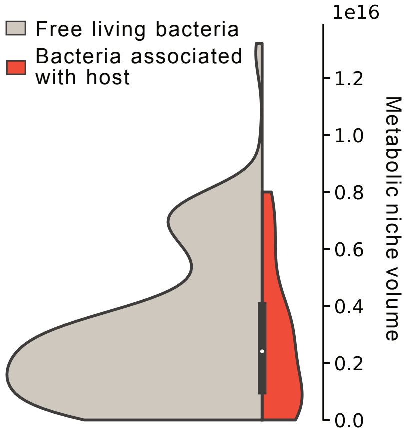
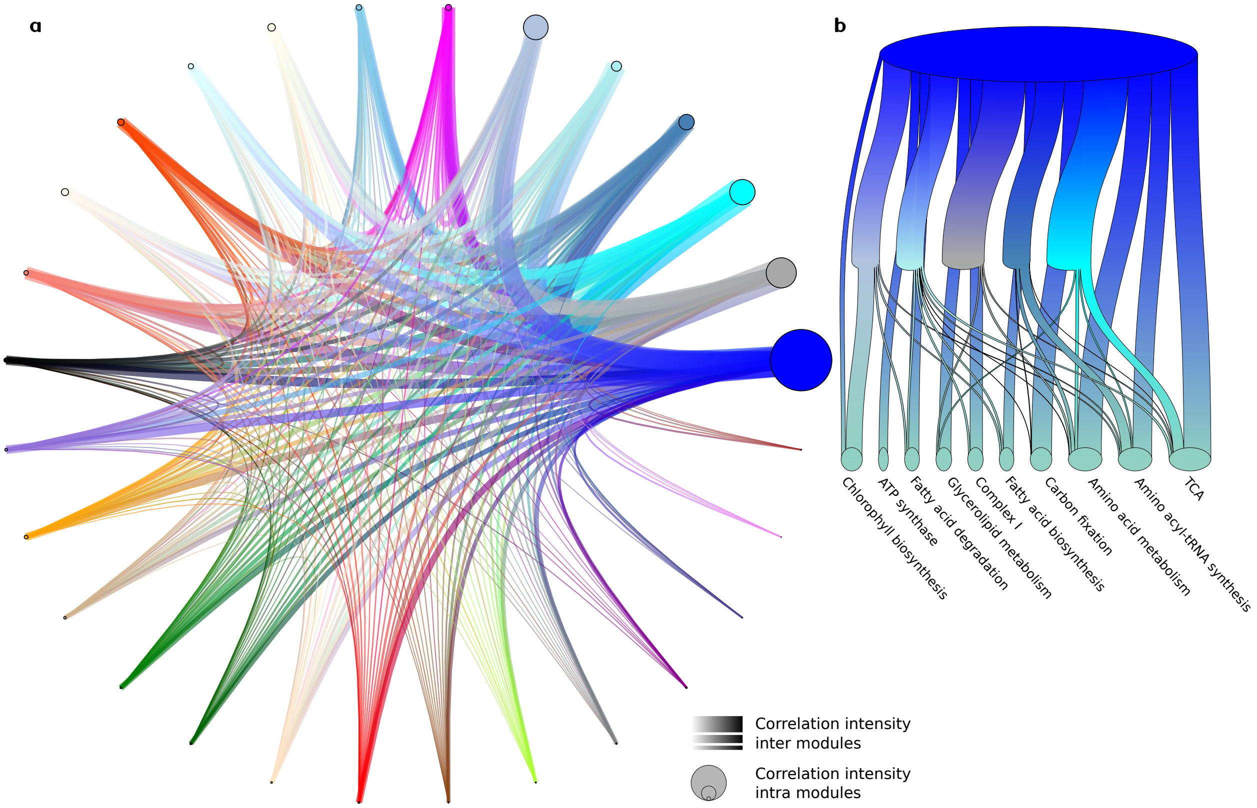
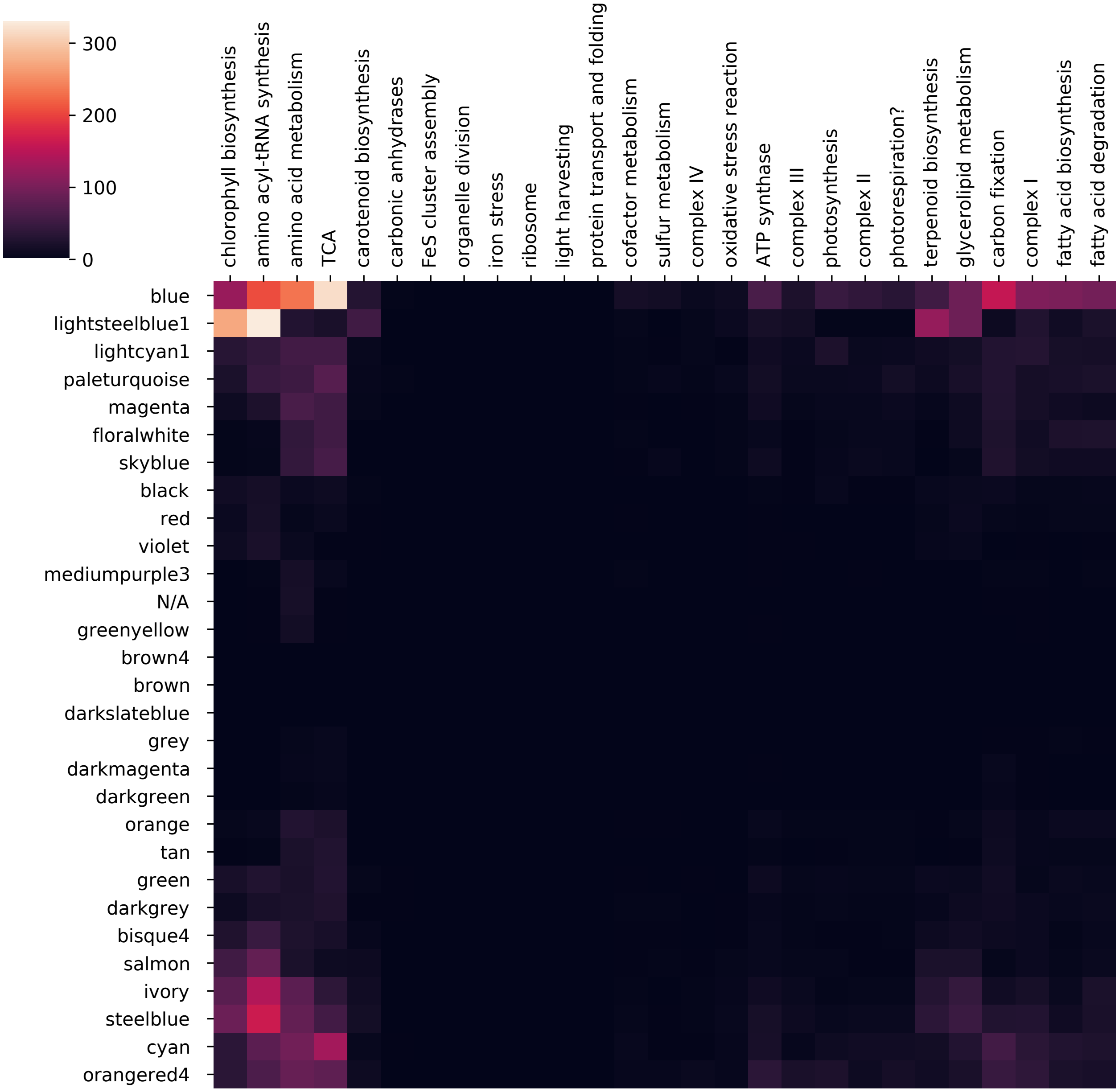
Supporting Information



Extended Data Fig. 1. Distribution of metabolic niche volumes for bacteria known for being associated with a host (in red), and bacteria not defined as being associated with host (in grey), as stored in PATRIC database.



Extended Data Fig. 2. Graph of transcriptomic clusters correlation On (a) a general view of the clusters correlations is represented. Each cluster is a vertex, which size depends on the intra-correlation of its genes. The width of the edges represents the value of the correlation sum that is computed between clusters. In (b) more than 60% of the blue module correlation sum is detailed. A ribbon from the blue ellipse is a correlation that leads to another module and then split into different pathways or directly to a pathway if not related to one of the five represented modules. The intensity of the correlation is proportional to the size of the ribbon.



Extended Data Fig. 3. Correlation heatmap of pathways versus modules

Supplementary Text 1: Computational implementation of the metabolic niche

Genome-scale metabolic networks We applied the metabolic niche formalism on different metabolic models available in the BiGG database (King et al., 2016). *Escherichia coli str. K-12 substr. MG1655* core is a heterotrophic microbial model organism suitable for bioinformatics benchmarking, comprises 95 reactions of which 20 are exchanges reactions and 72 metabolites. *Phaeodactylum tricornutum CCAP 1055/1* is an ubiquitous eukaryotic organism for which a genome-scale metabolic model is already well described (Broddrick et al., 2019). Finally, the metabolic niche was computed on several prokaryotic metabolic models reconstructed with Carveme (Machado et al., 2018) available at the following repository github.com/cdanielmachado/embl_-gems.

Pipeline From the metabolic network of the considered organism, we identify the exchange reactions (**1**) we want to see as parameters of the niche (whose flux will correspond to the axis of the metabolic niche). With that information, we formulate the problem as a Vector Linear Program (VLP) (**2**), that once solved, results in a list of vertices. The vertices fully characterize the niche as a volume (**3**) in the defined space.

(1) Identification of Significant exchange reactions Exchange reactions are explicit from the metabolic network description. However, for the sake of simplicity, the metabolic niche definition requires minimizing the number of exchange reactions. For this purpose, we run a Flux Variability Analysis (FVA) (Heirendt et al., 2019) with *COBRApy* that computes both lower and upper bounds of each flux while respecting our constraints. Our constraints are the same as a classical constraint-based model, plus the survival condition, which imposes a minimum flux through the biomass reaction. In our formalism, exchange reactions are considered to have only a reactant and no product. A negative flux describes a consumption by the organism (the metabolite "appears" in the organism), and a positive flux is a production (the metabolite "disappears" from the organism). Thus every reaction having a negative FVA minimal bound is a reaction responsible for consuming a nutrient. This preliminary check allows us to narrow the number of exchange reactions to consider. For instance, blocked or fixed exchange reactions can bring numerical error in the next step of the niche computation and should be avoided.

(2) VLP formulation Once identified, the reactions define a space on which we want to project the niche. From the stoichiometric matrix, the projection matrix, and the reaction bounds, the solver *Bensolve* (Löhne & Weißing, 2017) allows us to solve the problem formulated as follow:

$$\begin{cases} min \quad \mathbf{Px} \\ \text{subject to} \quad \mathbf{a} \leq \mathbf{Qx} \leq \mathbf{b} \quad | \quad \mathbf{l} \leq \mathbf{x} \leq \mathbf{s} \end{cases}$$

Where $\mathbf{Q} = \mathbf{S}$, $\mathbf{a} = \mathbf{b} = \mathbf{0}$ represent the quasi steady states approximation and $\mathbf{l}$ and $\mathbf{s}$ are respectively lower and upper bound defined in (5). The matrix $\mathbf{P}$ is the one defined in (7), with adjustment because the $\mathbf{x}$ vector considered in *Bensolve* is our $\mathbf{v}$. The formulation is done through *Benpy* a python wrapping of *Bensolve* that can be found at gitlab.univ-nantes.fr/mbudinich/benpy. This resolution gives us the upper image of the solution space with the last component that need to be removed. Once removed, only vertices are relevant as directions come from the last component.

(3) Volume computation The *Bensolve* solver allows us to get the vertex of the polytope. We then have the V-representation of the polytope. Depending on the network complexity and the size of the projected space, the number of vertices might be vast and difficult to manipulate. To avoid that, one can apply agglomerate clustering to reduce the number of points that simplify the polytope. Once simplified, we can use *lrs* (Avis, 2000) that gives us the volume of the defined polygon.

Supplementary Text 2: Formalization of the metabolic niche pairwise comparison

Comparing niches As p-dimensional volumes, niches can be compared and characterized with different measures. To formally compare such volumes, one can use a pseudo distance based on the Jaccard index (Conci & Kubrusly, 2017). The Jaccard index is a similarity measure applied on ensembles, looking at the intersection ratio over the union of the two compared ensembles. The distance is computed as follows:

$$d(V_a, V_b) = 1 - J(V_a, V_b) = 1 - \frac{|V_a \cap V_b|}{|V_a \cup V_b|}$$

, where $|.|$ is an operator measuring the size of the ensemble. For the niche, it is the volume. Biologically speaking, the intersection of two metabolic niches represents all the conditions (fluxes distribution through exchange reactions) that allow both species to survive.

The intersection of multidimensional polytope can be computationally intensive, so we developed a method based on metabolic networks to allow a more simple estimation of the intersection.

Let us consider two different species. We then consider the metabolic networks and the associated stoichiometric matrices $\mathbf{T} \in \mathbb{R}^{m_T, n_T}$ and $\mathbf{B} \in \mathbb{R}^{m_B, n_B}$. To compute the intersection of the two niches, one need to make the assumption that they have exchange reactions in common. Let us note $R_1..R_p$ the p exchange reactions we want to consider for the niche computation, and $M_1..M_p$ the p corresponding metabolites (for clarity we are omitting here the subscripts T and B that tell from which organism we are talking about).

We are going to order the matrix $\mathbf{T}$ and $\mathbf{B}$, so that the p reactions and metabolites are placed at the top left of the matrix:

$$\mathbf{T} = \begin{pmatrix} \begin{matrix} -\mathbf{I}_p \\ \mathbf{O}_{m_\mathbf{T}-p,p} \end{matrix} & \widetilde{\mathbf{T}} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \begin{matrix} -\mathbf{I}_p \\ \mathbf{O}_{m_\mathbf{B}-p,p} \end{matrix} & \widetilde{\mathbf{B}} \end{pmatrix}$$

Here the exchange reactions of the two species are distinct axes in the niche space. We need to modify the model so that there is only one set of p exchange reactions responsible for the intake of the p metabolites of both species. Exchange reaction i should look like : $M_{Bi} + M_{Ti} \longleftrightarrow M_i ex$. In term of matrix this gives us:

$$\mathbf{S} = \begin{pmatrix} \begin{matrix} -\mathbf{I}_p \\ \mathbf{O}_{m_\mathbf{T}-p,p} \end{matrix} & \widetilde{\mathbf{T}} & \mathbf{O}_{m_\mathbf{T}, n_\mathbf{B}-p} \\ \begin{matrix} -\mathbf{I}_p \\ \mathbf{O}_{m_\mathbf{B}-p,p} \end{matrix} & \mathbf{O}_{m_\mathbf{B}, n_\mathbf{T}-p} & \widetilde{\mathbf{B}} \end{pmatrix}$$

The model has then $m_T + m_B$ metabolites and $n_T + n_B - p$ reactions. The first m_T line correspond to the network T, and the last m_B lines to the network B. The corresponding flux vector $\mathbf{x}$ will have its first p component responsible for the intake of the p metabolites in T and B, the following $m_t - p$ components will be the inner mechanism of T that we want to abstract, and the last $m_B - p$ lines the ones of B.

Let us see the implication of such a formalism in term of flux bounds. If we rearranged the bounds order to correspond to the matrix $\mathbf{T}$ and $\mathbf{B}$ defined earlier we have:

$$\mathbf{ub_T} = \begin{pmatrix} \mathbf{ub_T}_p \\ \mathbf{ub_{T(n_T-p)}} \end{pmatrix} \quad \mathbf{ub_B} = \begin{pmatrix} \mathbf{ub_B}_p \\ \mathbf{ub}_{\mathbf{B}(n_\mathbf{B}-p)} \end{pmatrix} \quad \mathbf{lb_T} = \begin{pmatrix} \mathbf{lb_T}_p \\ \mathbf{lb}_{\mathbf{T}(n_\mathbf{T}-p)} \end{pmatrix} \quad \mathbf{lb_B} = \begin{pmatrix} \mathbf{lb_B}_p \\ \mathbf{lb}_{\mathbf{B}(n_\mathbf{B}-p)} \end{pmatrix}$$

The bounds for the p exchange reactions will be the $\mathbf{lb}_i = max(\mathbf{lb}_{\mathbf{T}i}, \mathbf{lb}_{\mathbf{B}i})$ and $\mathbf{ub}_i = min(\mathbf{ub}_{\mathbf{T}i}, \mathbf{ub}_{\mathbf{B}i})$. The rest of the bounds are defined by the network that possesses the reaction. Thus we have:

$$\mathbf{ub} = \begin{pmatrix} min(\mathbf{ub}_{\mathbf{T}p}, \mathbf{ub}_{\mathbf{B}p}) \\ \mathbf{ub}_{\mathbf{T}(n_{\mathbf{T}}-p)} \\ \mathbf{ub}_{\mathbf{B}(n_{\mathbf{B}}-p)} \end{pmatrix} \quad \mathbf{lb} = \begin{pmatrix} max(\mathbf{lb}_{\mathbf{T}p}, \mathbf{lb}_{\mathbf{B}p}) \\ \mathbf{lb}_{\mathbf{T}(n_{\mathbf{T}}-p)} \\ \mathbf{lb}_{\mathbf{B}(n_{\mathbf{B}}-p)} \end{pmatrix}$$

The newly created system can then be formulated as a VLP and the previously described pipeline allows computation of the solution which is the intersection of the two metabolic niches of T and B computed on the p exchange reactions.

Pairwise comparison of marine prokaryotes Due to numerical imprecisions or error we made some approximations to make our results more robust. Inclusion where considered when the intersection was covering at least 999‰ of the volume of one of the two considered niches. That means, if we consider niche i and j, with a volume vol_i and vol_j, with an intersection $inter$ the inclusion was determined if:

$$|1 - \frac{inter}{vol_i}| < 10^{-3} \quad \text{or} \quad |1 - \frac{inter}{vol_j}| < 10^{-3} \quad \text{or} \quad d_{Jacquard}(i, j) < 10^{-3}$$

We consider the computation as an error if:

$$\frac{inter}{\min(vol_i, vol_j)} > 1.001$$

We had 502 species. That means 125751 comparisons. Among those comparisons, 111775 (89%) where computed correctly, 4286 (less than 4% of computed comparisons) results in error because of an intersection that were too big. The rest were either not computed because of an error of the solver, or was taking too long (over than 2 days) during computation (9% of all the comparisons). When computing the inclusion graph we have 47287 edges, an edge is an inclusion. The inclusion relation is transitive, that means that if A include B, and B include C, then A include C. We can applied a transitive closure on the graph. When we do so we have a graph of 56322 edges. This means that around 10000 comparisons (computed or not) should be inclusion, whereas we got an other results. Half of the newly found edges belongs to not computed comparisons. 15% of them where found in error, which results in 35% of them that are not inclusion in our computations (less than 3% of the computed comparisons). Those errors come from the *Bensolve* solver during computation of the vertices coordinates or from the *Lrs library* during computation of the volume.

Supplementary Text 3: Sampling of *Phaeodactylum tricornutum* niche space

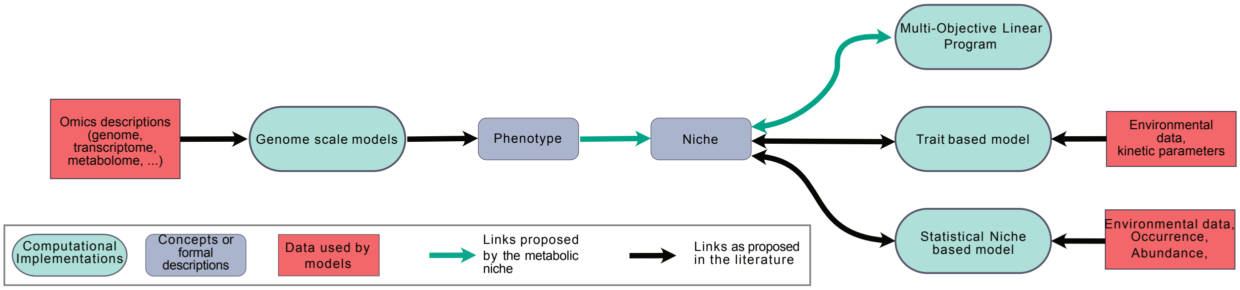
Exploring niches The niche flux space investigation emphasizes how the organism allocates its resources and its energy for the sake of its survival. However, the formal investigation of this space is a challenging task that we propose to overcome via the use of OptGP sampler (Megchelenbrink et al., 2014) from *COBRApy* (Heirendt et al., 2019). This technique computes different points that belong to the niche flux space. Each point is a distribution of flux values over all the reactions. Considering fluxes as random variables, we computed pairwise correlations between reactions over the extracted samples to create a weighted correlation graph. It summarizes the metabolic niche's organization and highlights dependencies between metabolic reactions motivated by its survival. As a final metric, we sum all the correlations associated with one reaction. A high value for a given reaction indicates that this reaction plays a pivotal role as a flux variation of this reaction will imply large changes in several other reactions.

Problem description When formalizing the niche we have a well defined space. The characterization of this space can be done through the interdependences of each pair of reactions. This can be directly measured with the correlation. Method relying on kernel analysis has been proposed to compute correlation between each reaction (Schwartz et al., 2019). But this method did not take into account the boundaries of the system, which is for us one key component of our formalism. A way to circumvent this issue is to compute the correlation between reactions through a sampling method. But sampling such a high dimensional volume is not an easy thing to do.

Tools and approximation We used the OptGP sampler (Megchelenbrink et al., 2014) embedded in *COBRApy* to obtain enough points to compute the correlation. We sampled 10^5 points with a fitting of 10^5. This allow a convergence of the sampled distribution. We verified the convergence by making 10 batches with the same parameters and look at the variance of each correlation. The sampling has been done with the integration of the survival condition in the model, meaning that the lower bound of the flux through the biomass reaction was set to 0.01. When we sampled the niche space we did not get rid of biomass reactions, whereas a strict sampling of the niche space should be a sampling without the biomass reactions that introduce a bias on the distribution. Indeed the biomass reaction is a constraint, it models the survival of the organism, but the niche definition does not need to have its value, as long as it is above the threshold. Unfortunately the OptGP sampler would requires some heavy modifications to allow this sampling.

Visualization Once we have the correlation graph, we need a proper algorithms to visualize it. We used the Graph-Tool library (Peixoto, 2014a). The library implements a hierarchical block structures algorithm (Peixoto, 2014b) which is of great help for module detection in huge graph.

Supplementary Text 4: Metabolic niche versus other niche modelings


Multi-Objective Linear Program
Omics descriptions (genome, transcriptome, metabolome, ...)
Genome scale models
Phenotype
Niche
Trait based model
Environmental data, kinetic parameters
Computational Implementations
Concepts or formal descriptions
Data used by models
Links proposed by the metabolic niche
Links as proposed in the literature
Statistical Niche based model
Environmental data, Occurrence, Abundance,

Extended Data Fig. 4. Illustration of the genome-scale modeling contribution to the concept of niche.

In addition to existing implementations of niche models, the metabolic niche relies on Genome-Scale Models (Price et al., 2004) that are available thanks to recent high-throughput data and systems biology theories. These models aim at abstracting metabolic phenotypes (i.e., Extended Data Fig. 4., left panels). The *phenotype* is the set of observable characteristics or traits of an organism[1]. The term covers the organism's morphology or physical form and structure, its developmental processes, biochemical and physiological properties, behavior, and the organism's effect on the environment. By focusing only on biochemical and physiological properties, metabolic modeling aims to investigate properties that can be linked to metabolic processes called the *metabolic phenotype* (i.e., no consideration for morphology).

Metabolic engineering targets understanding of internal machinery of organisms described originally via their gene content. This knowledge is then synthesized into the organism's metabolic network. It regroups all the metabolic reactions encoded in the genotype and the modeling reactions, such as biomass and exchange reactions, where the biomass reaction models the growth rate of the organism and exchange reactions model the interaction of the organism with its environment. Since the metabolic network models the organism's metabolic phenotype, we used this modeling and focused on the sole exchange flux values satisfying a minimal flux through the biomass reaction. This reduction allows to reach a niche formulation, called the metabolic niche. Indeed, this formulation is inspired by the central niche concept in ecology and proposed by the seminal work of Hutchinson, where the *fundamental niche* defines environmental conditions that allow an organism to survive. This concept inherently removes the mechanistic understanding of the organism's physiology for characterizing a relationship between the growth (or at least the survival) and its abiotic environment.

For decades, in practice, the niche was approximated or computationally implemented through two different types of modeling, relying on different types of data (i.e., Extended Data Fig. 4. right panels). Mainly from the estimation of kinetic parameters, *trait-based modeling* proposes a quantitative estimation of the organism abundance. Through various measures and observations on an organism, traits are inferred and modeled to assess the phenotypic plasticity of organisms. Thus, as discussed in *Rebuilding community ecology from functional traits*(McGill et al., 2006), McGill et al. define a trait as a measurable property of an organism that can be linked to its performance. A trait is thus to be linked with optimization features, sometimes called fitness, for seeking parameters values that approximate accurate growth rates or other observations from given environmental conditions. On the other hand, from occurrence data, *statistical niche-based modelings* describe the niche through a statistical inference between environmental conditions and organism occurrences. Worth noticing, the link between traits and statistical niche-based models has been investigated in several studies(Thuiller et al., 2010; Nagaraju et al., 2013).

Contrary to those standard niche modelings, the metabolic niche does not follow an optimal assumption (i.e., an organismal objective to maximize) but explores all exchanges fluxes values that allow an organism to survive. In this context, the metabolic niche is an abstraction of the phenotype's fundamental niche.

[1] https://www.nature.com/scitable/definition/phenotype-phenotypes-35/

However, solutions associated with trait-based models are included in the solution space. Trait-based models often rely on one or more criteria to maximize. Those criteria can be translated into additional constraints in our approach, leading to sub-spaces of the niche situated on the external niche envelope (i.e., corresponding to the maximal biomass fluxes). So there is an explicit dependency between the trait-based models and the metabolic niche formulation (i.e., the inclusion of metabolic trait-based model solutions in the metabolic niche space). However, we can still implement the metabolic niche without maximizing the biomass flux.

Formally, the metabolic niche reformulates semi-quantitative knowledge (i.e., presence/absence of genes or relative abundance of gene transcripts) into a quantitative framework to fit the fundamental niche expectation, which is quantitative by nature. This change of abstraction from semi-quantitative to quantitative is a theoretical and computational challenge necessary and recurrent in omics data. The metabolic niche contributes to this general effort by resolving a complicated mathematical problem and assuming the biological system in quasi-steady-state conditions. It is a strong assumption for modeling a biological system in its environment, but it remains coherent with preliminary metabolic engineering studies. Indeed, previous experimental results showed that microbial metabolisms adapt themselves within an hour. Complementary, other constraint-based modeling techniques take benefit from this assumption for simulating an organisms' adaptation by computing different metabolic fluxes following the evolution of substrates at the minute time-scale (i.e., the systems being at quasi-steady states every minute) (Bulović et al., 2019). These points advocate for the quasi-steady-state assumption and the accuracy of the metabolic niche to investigate the adaptation or acclimation of organisms in environmental conditions.

Limits and approximations Like most models, the metabolic niche has approximations and limits. Major metabolic niche approximations come from the genome-scale metabolic model definition. Among others are the biomass reaction, the temperature, and the metabolic network per se. The metabolic niche is sensitive to the biomass reaction, as its composition would likely change metabolic needs. The temperature is not a parameter of our model as we cannot exhibit differences in metabolic needs depending on the environment's temperature. Indeed, genome-scale metabolic models do not apprehend the temperature as a parameter but use it during the network reconstruction step. The temperature is then used to determine the reversibility of reactions and their bounds that are involved in (2). Hence, our modeling framework heavily relies on GSMs, and their critical reconstruction. To allow (1) we had to assume that the system was at quasi-steady-states which is common in systems biology (Varma & Palsson, 1994). Finally, the metabolic niche needs one parameter value, which is the death rate of the organism. The impact of this value in the metabolic niche computation is straightforward, as a metabolic niche with a lower death rate will increase its metabolic niche space.

The computational limits lie in the dimension of the niche space and the size of the network. As the resolution of (7) is sensible to the number of dimensions of $\mathcal{N}$, and because the solver is computing vertices, computation of the niche on too many dimensions will cause a high computation time and ask for lots of RAM.