cMFA for multi-omics data integration in microbial community models.

Sthyve Tatho, Simon Labarthe, Valentina Baldazzi

PhD Student
3rd June 2025
SMAI 2025

Sthyve Tatho
Univ. Bordeaux, INRAE, BIOGECO

Univ. Bordeaux, Inria-Inrae, Pléiade

Sthyve.tatho@inrae.fr

STHYVE TATHO

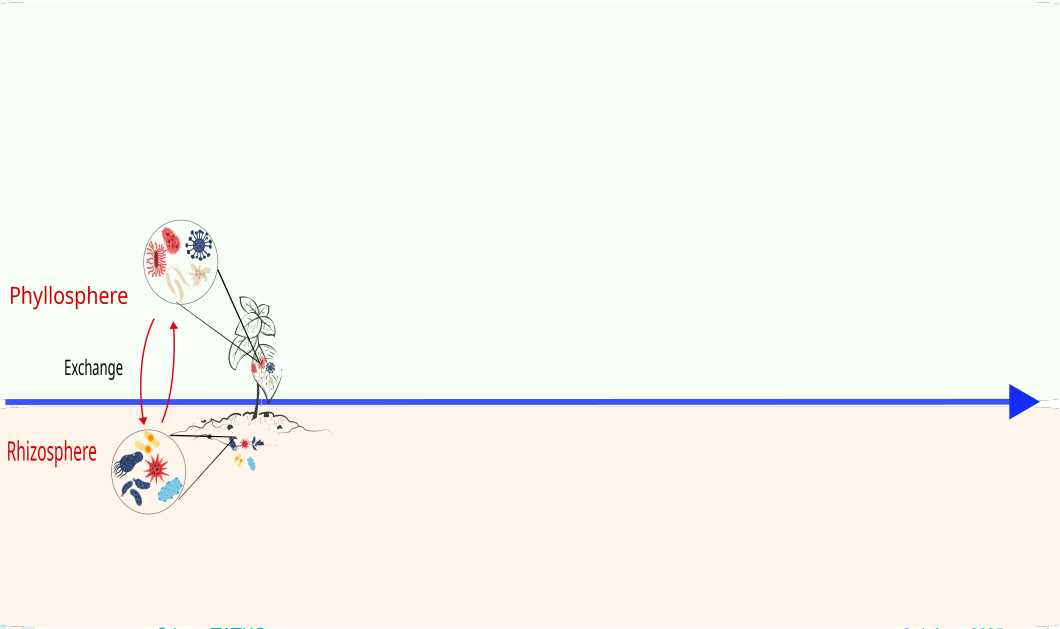
Complexity of natural microbial communities

Growth Stimulation (Phytohormone Production)

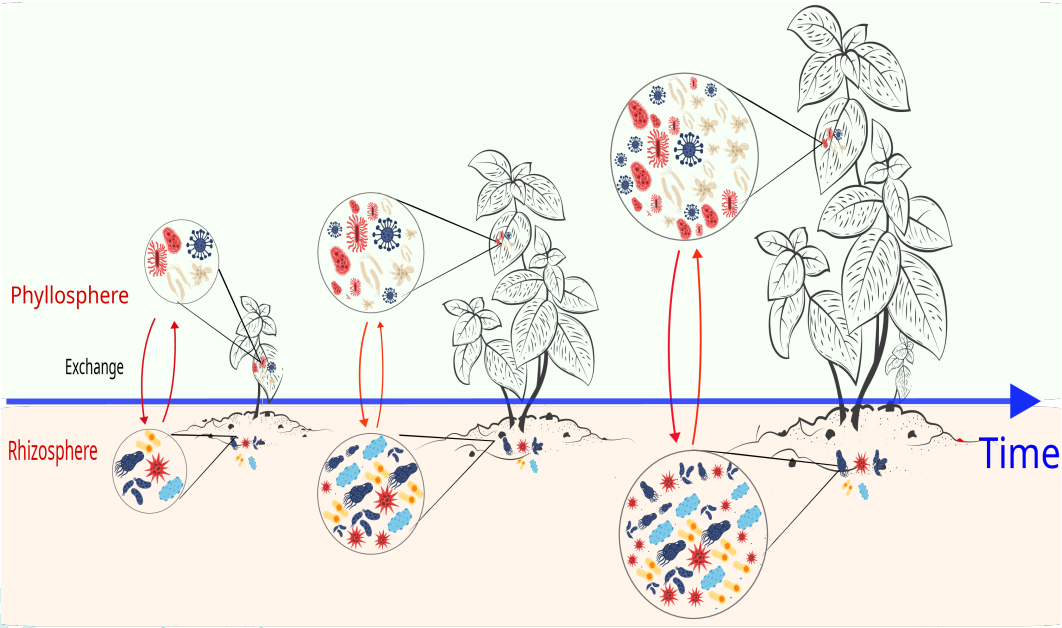
Fertilisation (nitrogene fixation)

Protection (Pathogen resistence)

Complexity of natural microbial communities



Complexity of natural microbial communities

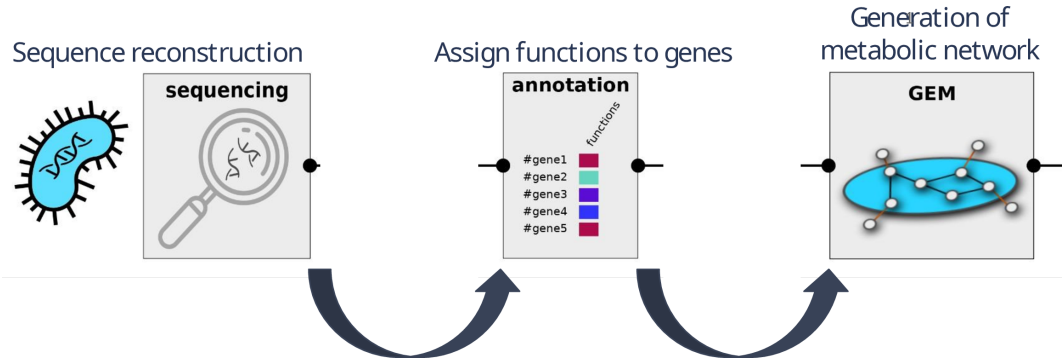


Biology-informed inference from multi-omics data

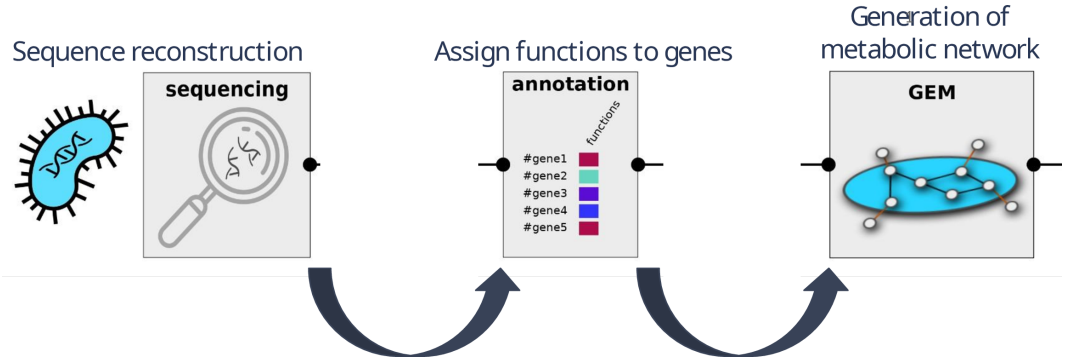
Context :

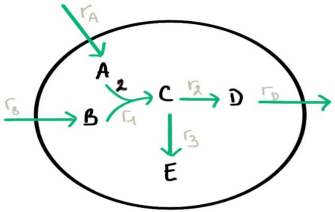
- Developing a mathematical inference method based on metabolic networks.
- Inferring metabolic fluxes by integrating multi-omics data.

Workflow : metabolic network creation



Workflow : metabolic network creation





$$S = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccccc} r_A & r_B & r_1 & r_2 & r_3 & r_D \end{array} \\ \begin{pmatrix} 1 & 0 & -2 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

$$S\nu = 0$$

Multi-omics data

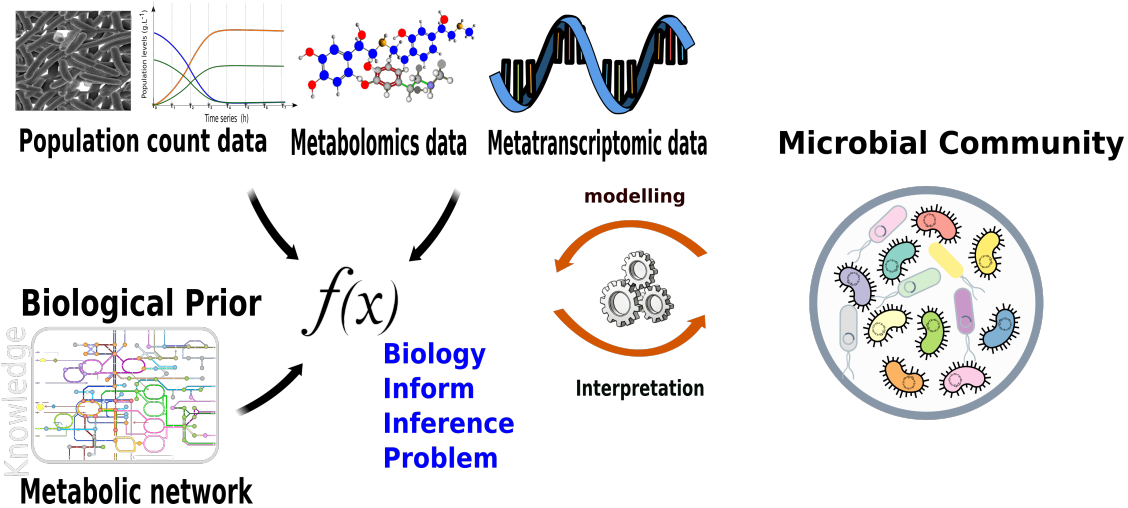


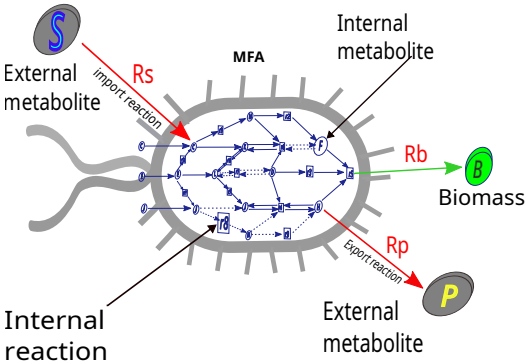
Figure – Differents types of multi-omics data

Biology-informed inference from multi-omics data

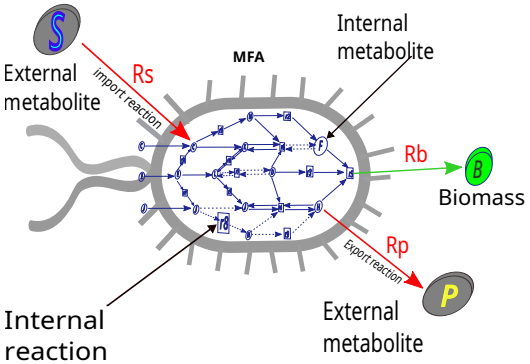
Objectives :

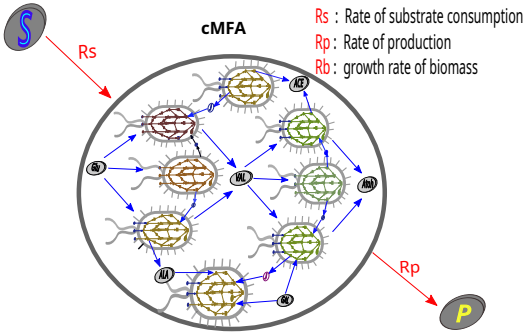
- Estimating the contribution of each microorganism to the community.
- Reconstructing internal fluxes of individual micro-organisms.

cMFA for community models

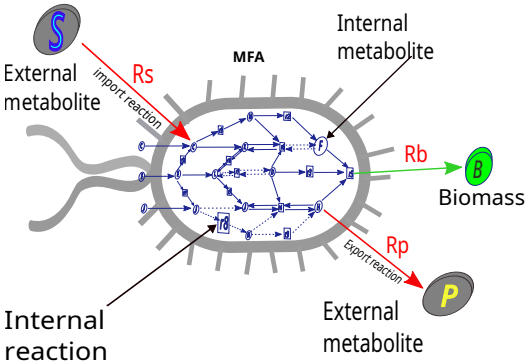


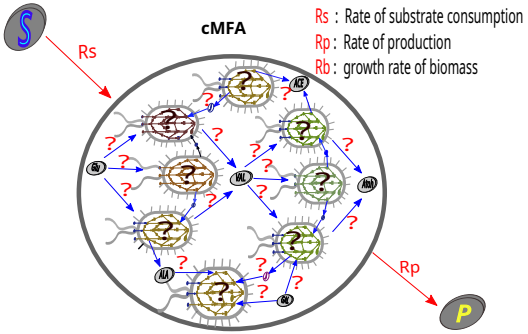
cMFA for community models



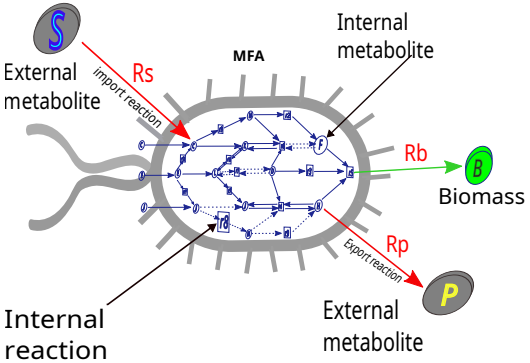


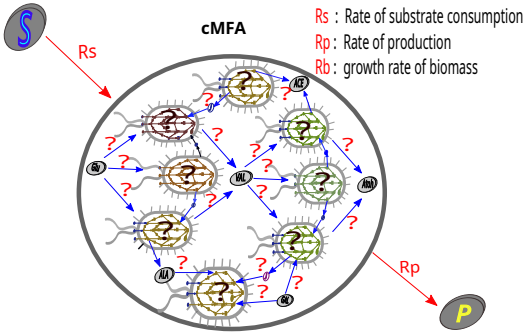
cMFA for community models





cMFA for community models

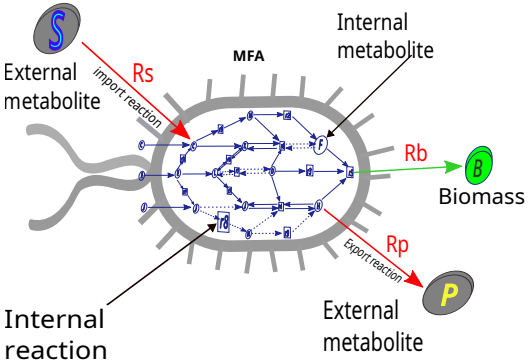


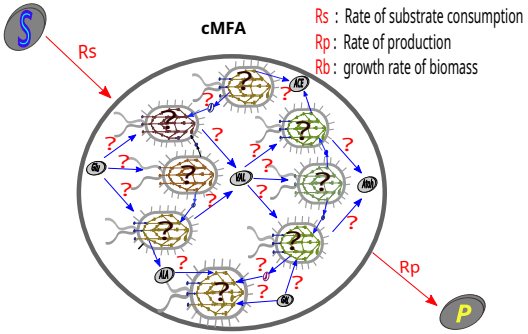


$$\hat{\nu}(t) := \arg\min \sum_{m \in \mathcal{M}} \left\| \overbrace{R_{m,t}}^{\text{Data rate}} - \overbrace{\sum_{b \in \mathcal{B}_m} \overset{\text{Biomass}}{\mathcal{X}_t^{(b)}} \nu_{m,t}^{(b)}}^{\text{Model}} \right\|_2^2 + \underbrace{\lambda \mathcal{R}(\nu)}_{\text{Regularization term}}$$

Set of (Metabolites & Micro organisms)

cMFA for community models





$$\hat{\nu}(t) := \arg\min \sum_{m \in \mathcal{M}} \left\| \overbrace{R_{m,t}}^{\text{Data rate}} - \overbrace{\sum_{b \in \mathcal{B}_m} \overset{\text{Biomass}}{\mathcal{X}_t^{(b)}} \nu_{m,t}^{(b)}}^{\text{Model}} \right\|_2^2 + \underbrace{\lambda \mathcal{R}(\nu)}_{\text{Regularization term}}$$

Set of (Metabolites & Micro organisms)

biological Information

$$\begin{cases} \overbrace{S^{(b)}}^{\text{stochiometry matrix}} \cdot \nu^{(b)} = 0, \\ C_{min}^{(b)} \leq \underbrace{\nu^{(b)}}_{\text{Fluxes}} \leq C_{max}^{(b)}. \end{cases}$$

Meta_transcriptomic data inclusion

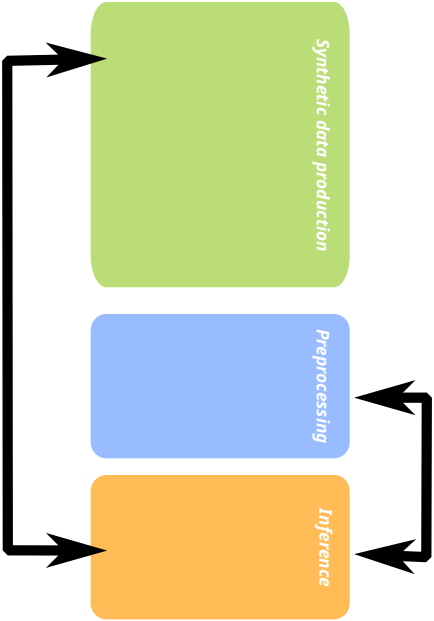
$$\hat{\nu}(t) := \arg\min_{\nu_t} \sum_{m \in \mathcal{M}} \| R_{m,t} - \sum_{b \in \mathcal{B}_m} \mathcal{X}_t^{(b)} \nu_{m,t}^{(b)} \|_2^2 + \lambda \mathcal{R}(\nu)$$
$$\text{subject to} : \begin{cases} S^{(b)} \cdot \nu^{(b)} & = 0 \\ C_{min}^{(b)} \leq \nu^{(b)} & \leq C_{max}^{(b)}, \end{cases} \quad \text{for} \quad b = 1, \cdots, N_b \tag{1}$$

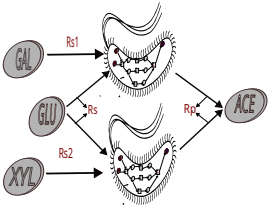
Inclusion by Penalization

$$\mathcal{T}_t^{(b)}(r) = f(T_t^{(b)}(r))$$
$$\mathcal{R} := \mathcal{R}(\mathcal{T}\nu)$$

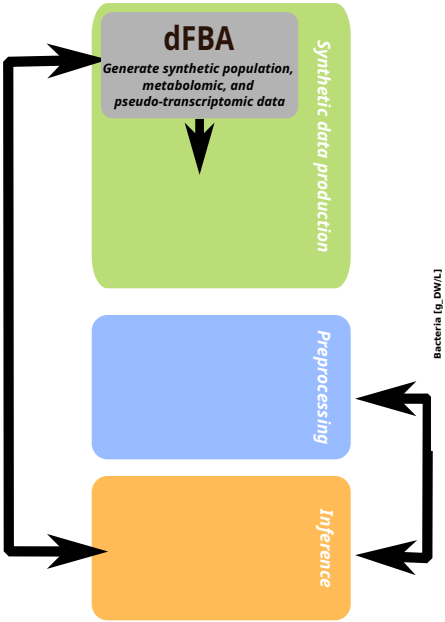
$\mathcal{T}_t^{(b)}(r)$: the linear coefficient of ponderation associated with reaction r , which is small if the reaction is activated and high if it is deactivated.
f : is the mapping function : transcript of reaction r to linear coefficient
T : trancript level

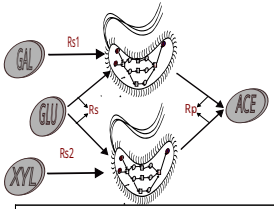
Benchmark method

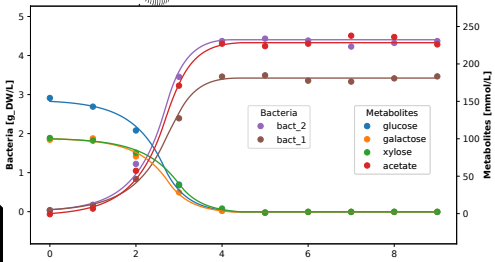




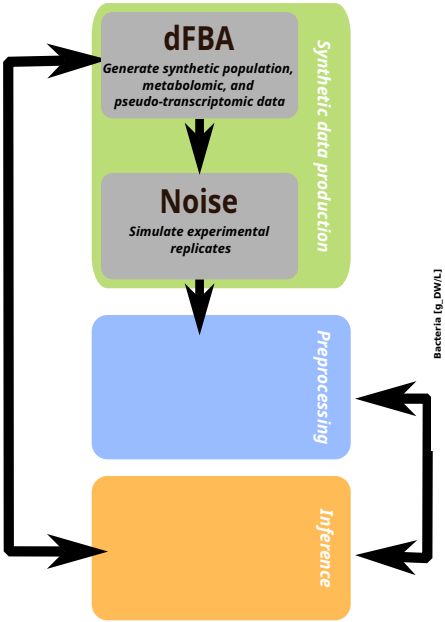
Benchmark method

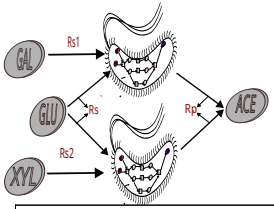


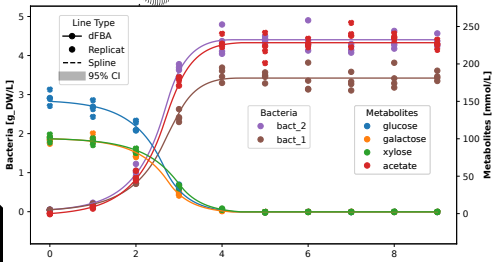




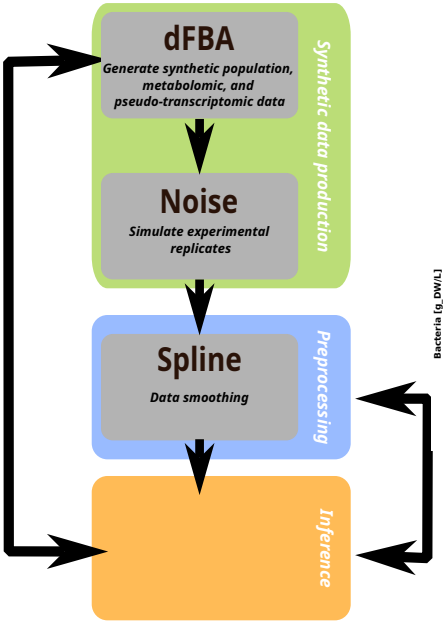
Benchmark method

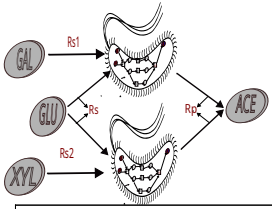


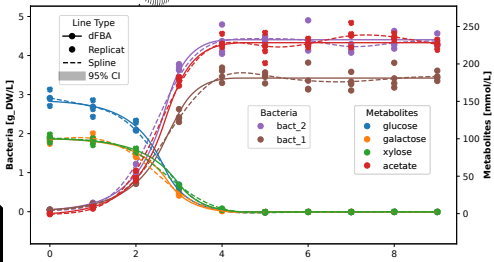




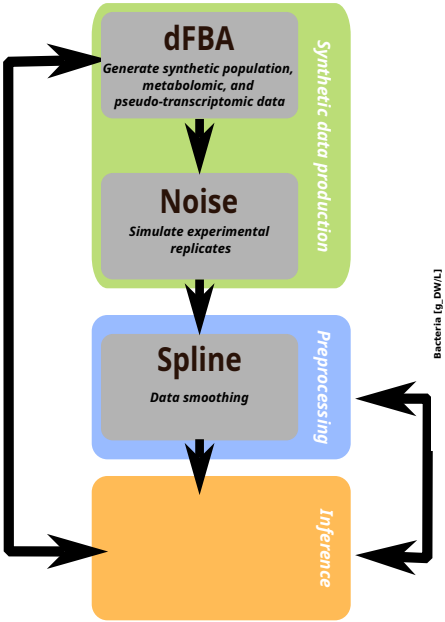
Benchmark method

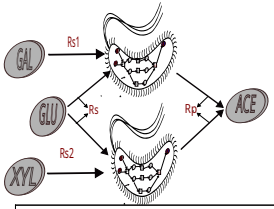


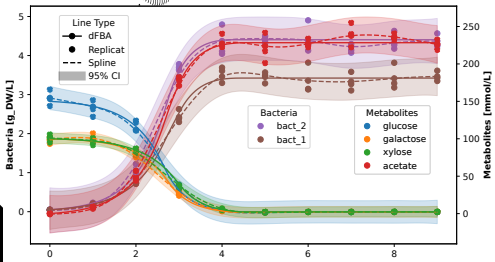




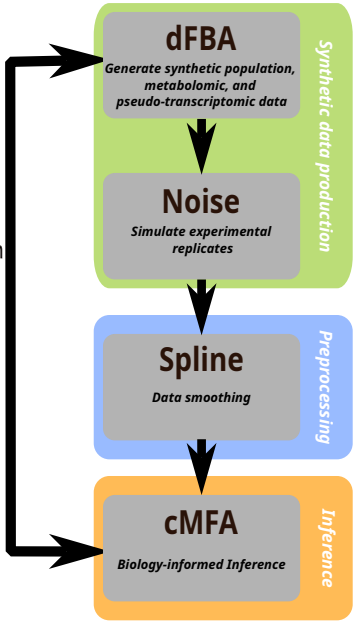
Benchmark method



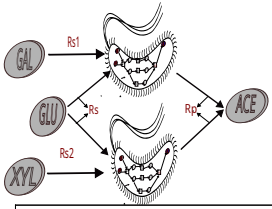


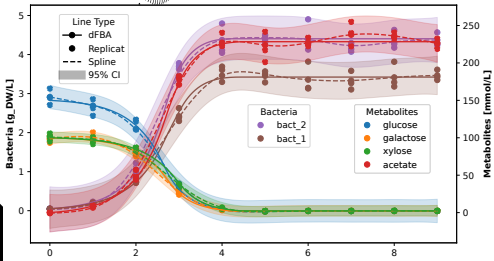


Benchmark method



Comparison between intracellular rates in **dFBA** and **cMFA**





Comparison between extracellular rates in **spline** and **cMFA**

Synthetic Metatranscriptomic Data Generation

Linear coefficient

$$\mathcal{T}_t^{(b)}(r) = f(r) \quad \forall r \in \mathcal{A}_t^{(b)} \text{ where } \mathcal{A}_t^{(b)} = \{r \in \mathcal{R}_b \mid \left|\mathcal{F}_r^b(t)\right| > 10^{-15}\}$$

$$\mathcal{T}_t^{(b)} \mapsto f\left(\mathcal{F}_r^b(t)\right) := \left(\max(C_r^b) - \min(C_r^b) - C_r^b\right) \tag{2}$$

with $C_r^b = \frac{\mathcal{F}_r^b}{\max \mathcal{F}_r^b} \quad r \in \mathcal{R}_b$

$\boldsymbol{\mathcal{T}^{(b)}}_t$: linear coefficient
$\mathcal{F}_r^b(t)$:= value of bacterial reaction flux r in response to gene expression levels.
$\mathcal{R}_b$:= set of reactions of the microorganism b

Preprocessing : Rescaling

Penalization + Transcriptomic data

$$\hat{\nu}(t) := \arg\min_{\nu_t} \sum_{m \in \mathcal{M}} \|R_{m,t} - \sum_{b \in \mathcal{B}_m} \mathcal{X}_t^{(b)} \nu_{m,t}^{(b)}\|_2^2 + \lambda \mathcal{R}(\mathcal{T}\nu)$$
$$\text{subject to : } \begin{cases} S^{(b)} \cdot \nu^{(b)} & = 0 \\ \nu_{min}^{(b)} \leq \nu^{(b)} & \leq \nu_{max}^{(b)}, \end{cases} \quad \text{for} \quad b = 1, \cdots, N_b \tag{3}$$

Preprocessing : Rescaling

Penalization + Transcriptomic data

$$\hat{\nu}(t) := \arg\min_{\nu_t} \sum_{m \in \mathcal{M}} \|R_{m,t} - \sum_{b \in \mathcal{B}_m} \mathcal{X}_t^{(b)} \nu_{m,t}^{(b)}\|_2^2 + \lambda \mathcal{R}(\mathcal{T}\nu)$$

$$\text{subject to : } \begin{cases} S^{(b)} \cdot \nu^{(b)} & = 0 \\ \nu_{min}^{(b)} \leq \nu^{(b)} & \leq \nu_{max}^{(b)}, \end{cases} \quad \text{for} \quad b = 1, \cdots, N_b \tag{3}$$

Vectorization + promotion of sparsity

$$\forall t \geq 0, \quad \hat{\nu}(t) := \arg\min_{\nu_t} \|R_t - B_t \nu_t\|_2^2 + \lambda \|\mathcal{T}\nu_t\|_1$$

$$\text{subject to : } \begin{cases} S \cdot \nu_t & = 0 \\ \nu_{\min} \leq \nu_t & \leq \nu_{\max} \end{cases} \tag{4}$$

$$\text{with} \quad B_t \nu_t = \sum_{b \in \mathcal{B}_m} \mathcal{X}_t \nu_t$$

Preprocessing : improving solver convergence

Renormalization and Dimensionless

$$\sigma_{1_t}^{-1} = \frac{1}{R_t}, \quad \sigma_{2_t} = \frac{R_t}{B_t}$$

so that after renormalization, the dimensionless problem becomes

$$\hat{\nu}_t = \arg\min_{\tilde{\nu}_t} \|\tilde{R}_t - \tilde{B}_t \tilde{\nu}_t\|_2^2 + \lambda \|\tilde{\nu}_t\|_1$$
$$\text{subject to : } \begin{cases} \tilde{S}\tilde{\nu}_t = 0 \\ \tilde{\nu}_{\min} \leq \tilde{\nu}_t \leq \tilde{\nu}_{\max} \end{cases} \tag{5}$$

$$\tilde{R}_t = \sigma_{1_t}^{-1} R_t, \quad \tilde{B}_t = \sigma_{1_t}^{-1} B_t \sigma_{2_t}, \quad \tilde{\nu}_t = \sigma_{2_t}^{-1} \nu_t$$
$$\tilde{S} = S\sigma_{2_t}, \quad \tilde{\nu}_{\min} = \sigma_{2_t}^{-1} \nu_{\min}, \quad \tilde{\nu}_{\max} = \sigma_{2_t}^{-1} \nu_{\max}$$

This renormalization harmonizes the scales of fluxes, particularly for external exchange reactions

Application : Benchmark Results for Two-Species Communities in Three Biological Cases

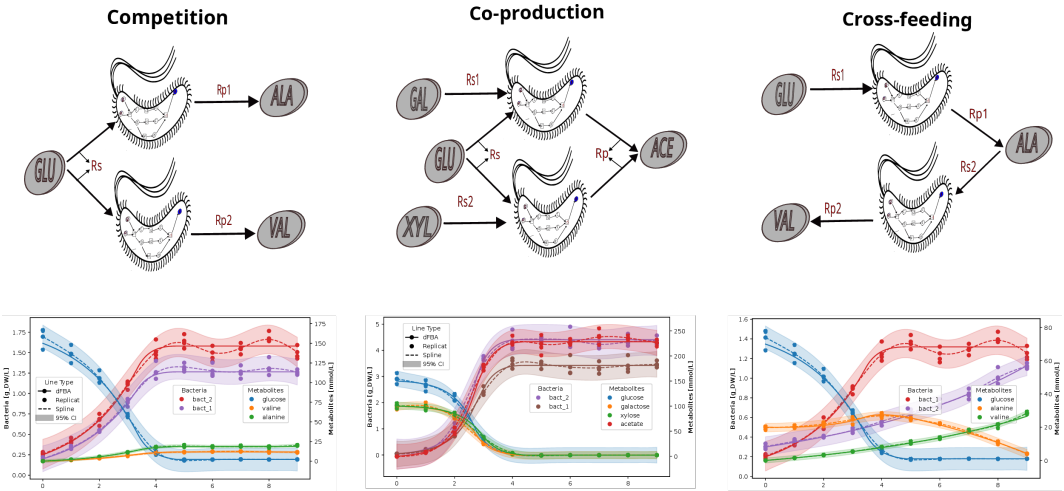
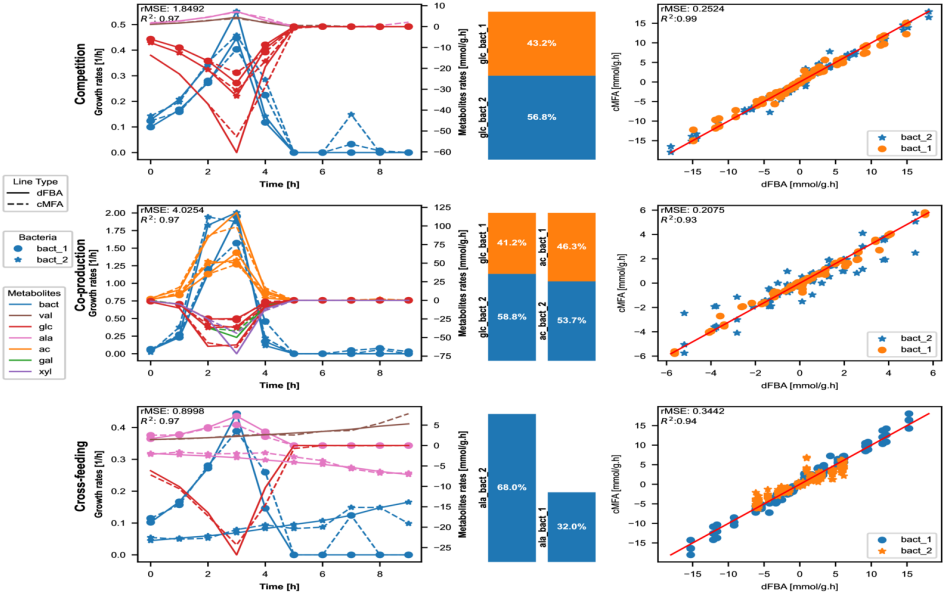
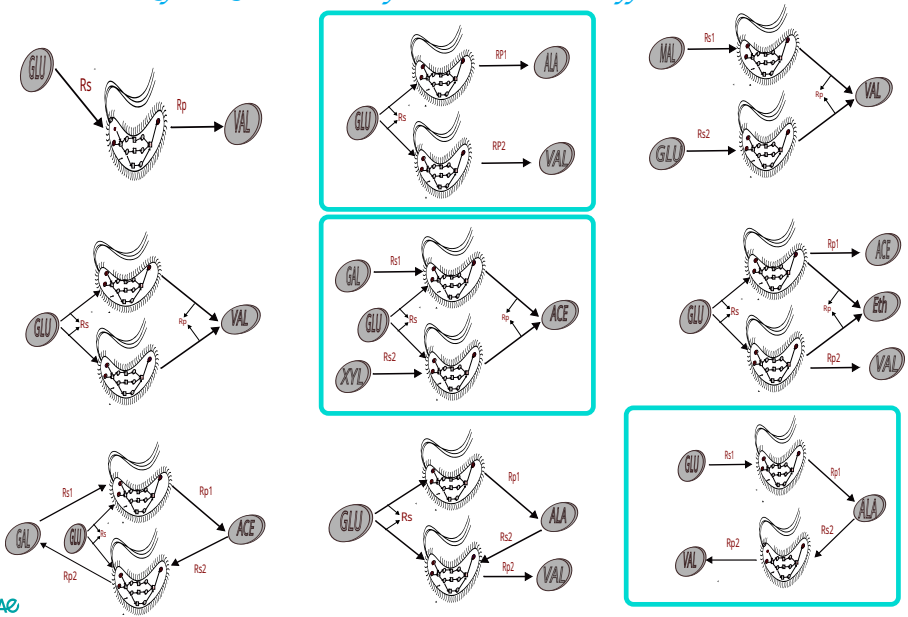


Figure – **Synthetic data for cMFA benchmark.** the different cases of increasingly complex ecological interactions are sketched

Application : Benchmark Results for Two-Species Communities in Three Biological Cases



Benchmark for Robustness of the method : Different Interactions



Benchmark for Robustness of the method : Data noise for replication

Noise percent in data

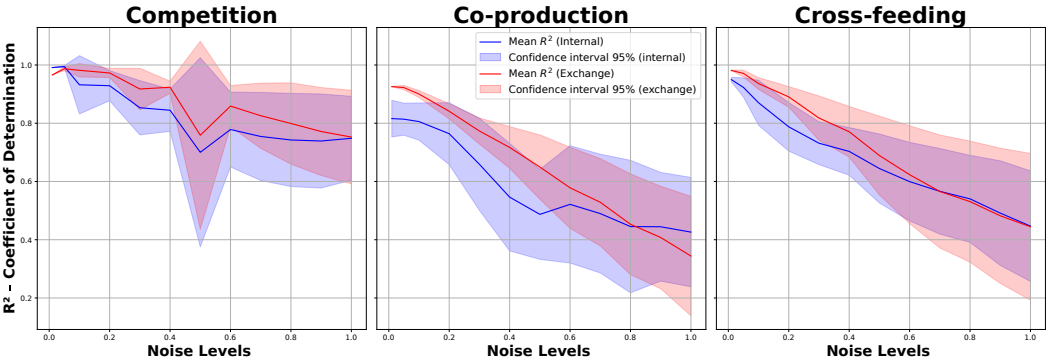
$$D_{\text{final}}(t) = \max\left(0, D_{\text{dFBA}}(t) \times (1 + \epsilon(t))\right), \text{ with }, \quad \epsilon(t) \sim \mathcal{N}(0, \sigma^2)$$

$\sigma = q\sigma_b$, where $\sigma_b = 1$

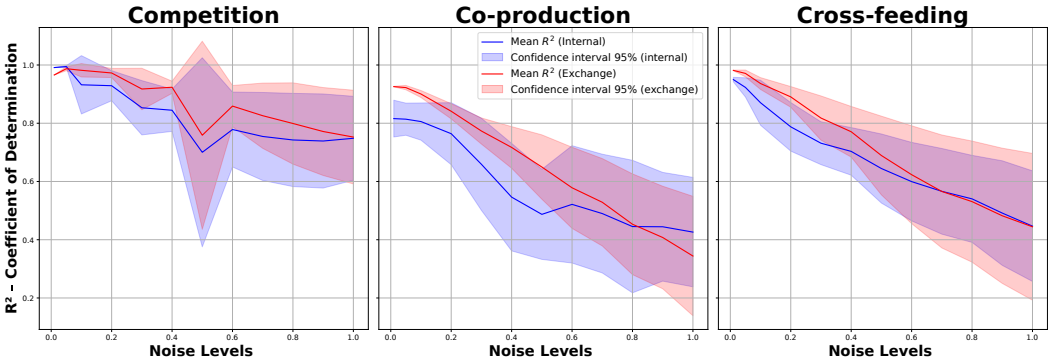
$$q \in \{[1\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%]\}$$

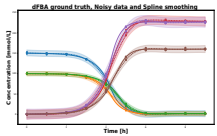
are used for the data noise benchmark
For each noise level, we replicated the whole benchmark pipeline (i.e. data noising, pre-processing with spline-smoothing and cMFA inference) $n = 7$ times,

Benchmark for Robustness of the method : Data noise for replication

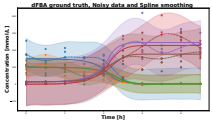


Benchmark for Robustness of the method : Data noise for replication

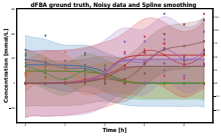




1%



50%



100%

Performance

1- high-performance modelf

2- R^2>0.7 for noise level <0.5

Benchmark for Robustness of the method : Incomplete meta-transcriptomic data

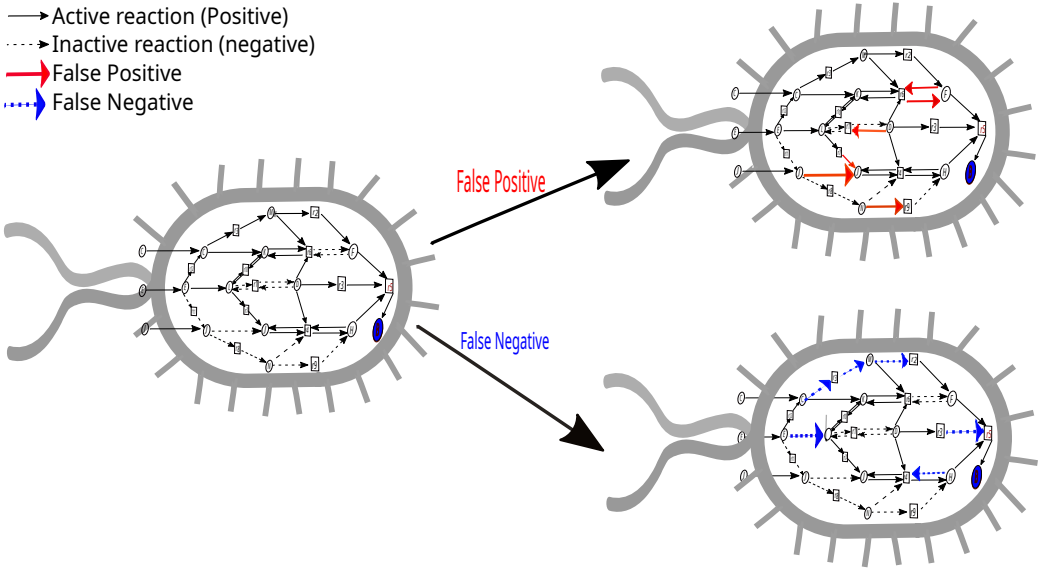


Figure – Incomplete meta-transcriptomic information

Synthetic Metatranscriptomic Data Generation

False Negative and False Positive

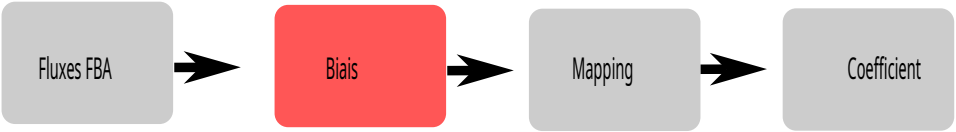
$\mathcal{B}_t^{(b)} \subseteq \mathcal{A}_t^{(b)}$

$$B\left(\mathcal{F}_r^b(t)\right) := 0 \qquad \text{if } r \in \mathcal{B}_t^{(b)}$$

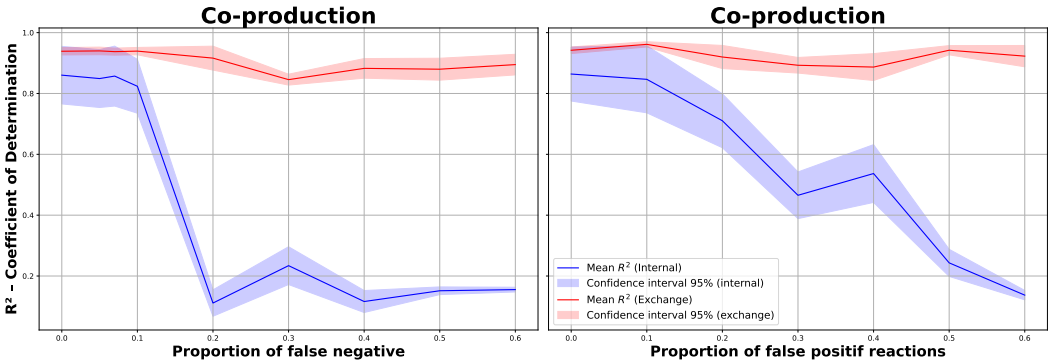
$\mathcal{B}_t^{(b)} \subseteq \mathcal{I}_t^{(b)}$

$$B\left(\mathcal{F}_r^b(t)\right) := \boldsymbol{\mathcal{T}_t^{(b)}(r)} \quad \text{if } r \in \mathcal{A}_t^{(b)}$$

B = : Biais function $\mathcal{I}_b$:= inactive reactions of microorganism b



Benchmark for Robustness of the method : Incomplete meta-transcriptomic data



Performances

- Good estimation of exchange fluxes
- Good estimation of internal fluxes when proportion < 10 %

Difficulty

- Sensitivity of metatranscriptomic data

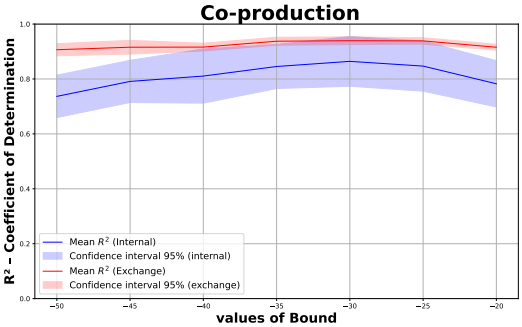
Benchmark for Robustness of the method : metabolic import rates

Intrinsic Flux

$$C_{min}^{(b)} = \max\left(\nu_{intr}, -\frac{m}{\Delta_t \cdot \sum_{b \in \mathcal{B}_m} \mathcal{X}_b}\right)$$

$\nu_{intr} \in [-50, -45, -40, -35, -30, -25, -20]$

reference value $\nu_{intr} = -30$



Performances

- Good estimation of exchange fluxes
- Good estimation of internal fluxes

Benchmark for Robustness of the method : Large Community Benchmark

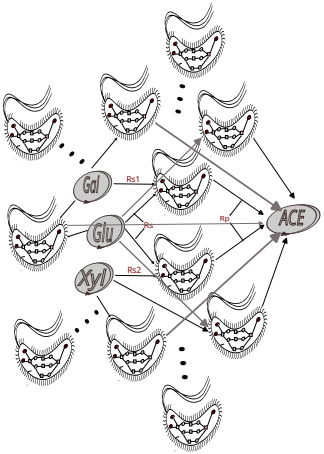


Figure – Same strain

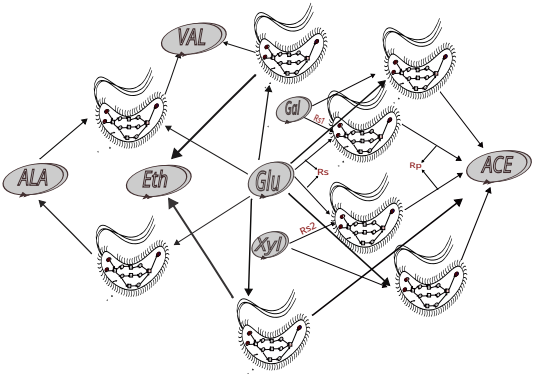


Figure – Different strain

Conclusion

We present a Statistical inference method informed by biology : *cMFA*

Synthetic Data

- Simulate data with dynamic FBA
- Add noise to create replicates
- Smooth the noisy trajectories
- Generate metatranscriptomic data

Work Completed

- Explore different interaction types
- Benchmark noise levels
- Handle incomplete metatranscriptomic data
- Vary metabolic import rates
- Test different community sizes (4–56)

Advantages

- Robust to noise
- Robust to variations in metabolic import rates
- Robust for exchange-rate inference with or without metatranscriptomic data
- Scalable to large communities

Disadvantages

- Internal-flux inference less robust when metatranscriptomic data are absent

References

- Brunner et al. (2023). *PLoS Comput. Biol.*
- Gerken Starepravo, M. et al. (2022). *Digital Chemical Engineering*, **2**.
- Lecomte, A. et al. (2023). *Metabolic Engineering*.
- Mahadevan, R. et al. (2002). *Biophysical Journal*, **83**.
- Pansu, M. (2004). *Courbes B-splines*.
- Pande, A. et al. (2013). *The ISME Journal*, **8**(5).


Thank you!