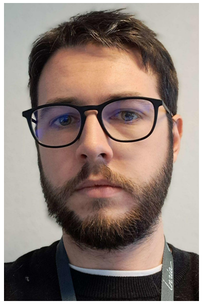
Learning metabolic regulatory rules from time series data

MERRIN: *MEtabolic Regulation Rules INference from time series data*[*]



Kerian Thuillier[1], *Caroline Baroukh*[2], *Alexander Bockmayr*[3], *Ludovic Cottret*[2], *Loïc Paulevé*[4], *Anne Siegel*[1]

[1] Univ Rennes, Inria, CNRS, IRISA, Rennes, France

[2] LIPME, INRAe, CNRS, Université de Toulouse, Castanet-Tolosan, France

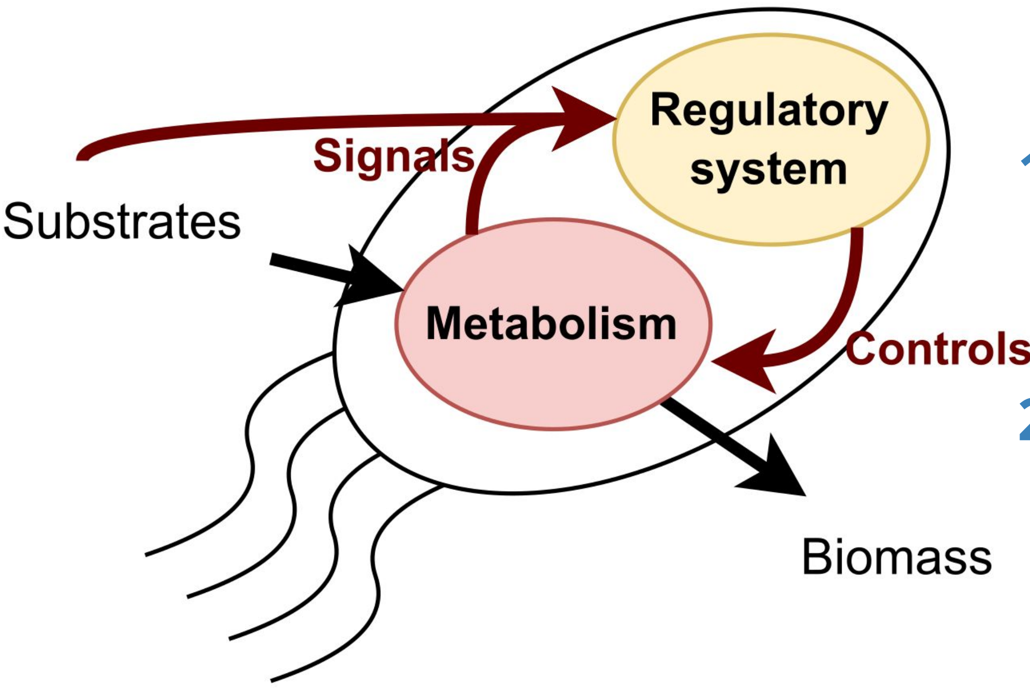
[3] Freie Universität Berlin, Institute of Mathematics, D-14195 Berlin, Germany

[4] Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR5800, F-33400 Talence, France

26th March 2024

[*] K. Thuillier et al., **Oxford Bioinformatics**, 2022

Cells: hybrid multi-layered structures



Model as two interconnected systems

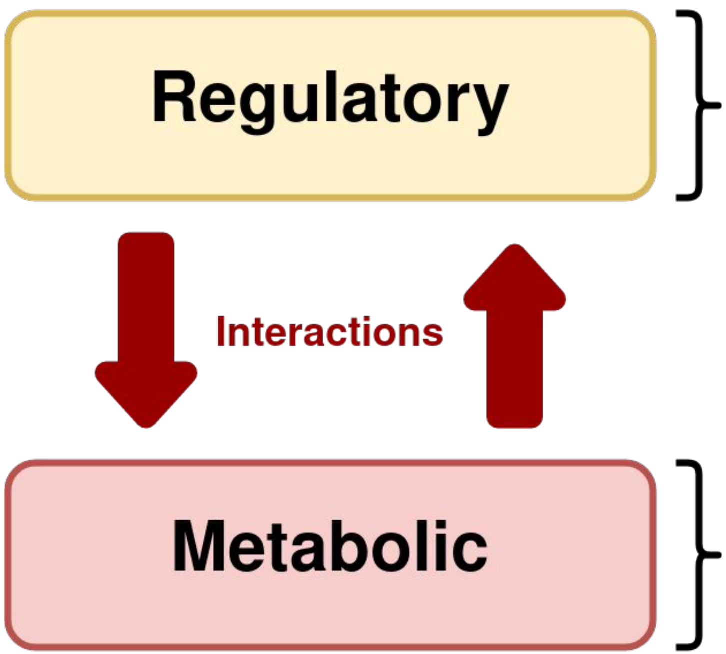
1. **Metabolic system**
   Chemicals reactions converting substrates to energy and biomass
2. **Regulatory system**
   Rules constraining the metabolism to adapt itself to its environment

Objective:

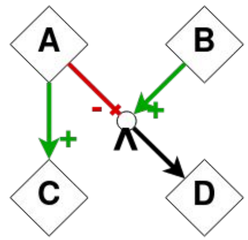
Inferring the **regulatory systems** from time series observations of the cells *(metabolism and regulation)*

Multiplicity of modelling formalisms

Two systems models with different dynamics



Discrete dynamics[1]



Model

$f_A(x_A, x_B, x_C, x_D) = x_A$
$f_B(x_A, x_B, x_C, x_D) = x_B$
$f_C(x_A, x_B, x_C, x_D) = x_A$
$f_D(x_A, x_B, x_C, x_D) = \neg x_A \wedge x_B$

Logical rules
(Boolean network)

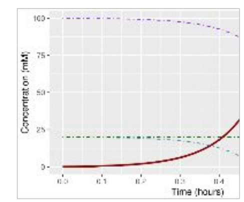
Inputs				Outputs			
x_A	x_B	x_C	x_D	x_A	x_B	x_C	x_D
0	0	x	x	0	0	0	0
0	1	x	x	0	1	0	1
1	0	x	x	1	0	1	0
1	1	x	x	1	1	1	0

Truth table (simulation)

Steady-states approximation[2]

$$\begin{aligned} \text{maximise} \quad & v_{\text{Growth}} \\ \text{such that:} \quad & S \cdot v = 0 \\ & l_r \cdot x_r \leq v_r \leq u_r \cdot x_r \quad \forall r \in \text{reactions} \end{aligned}$$

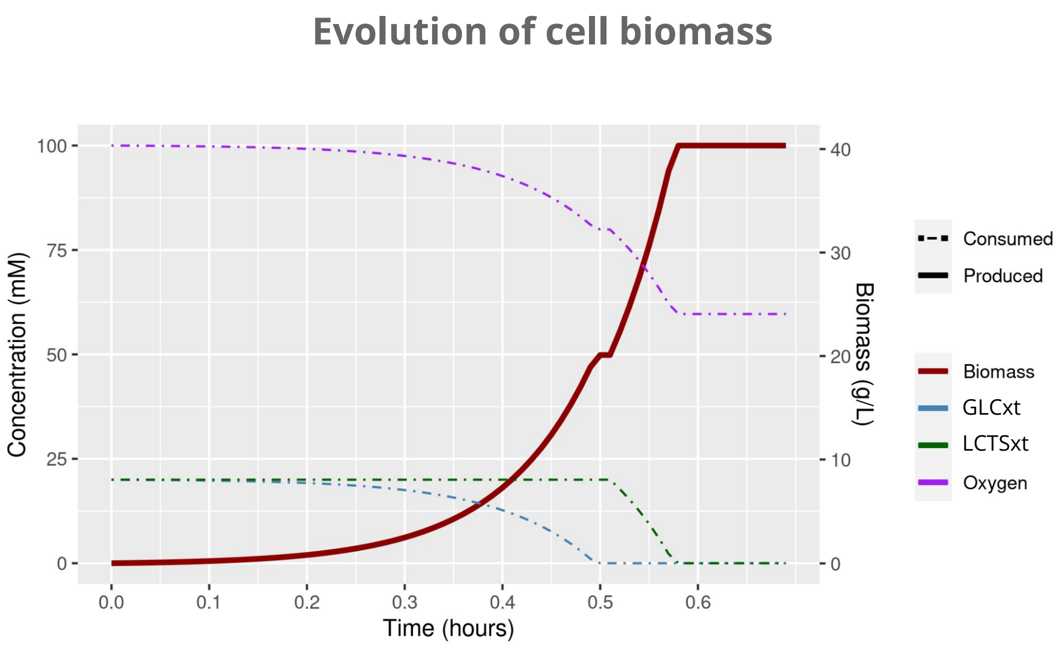
Regulatory flux balance analysis (rFBA)



Metabolic traces

[1] S. Videla et al., **Bioinformatics**, 2016
[2] M. W. Covert et al., **Journal of theoretical biology**, 2001

Example: diauxic shift *(Monod et al., 1953)*

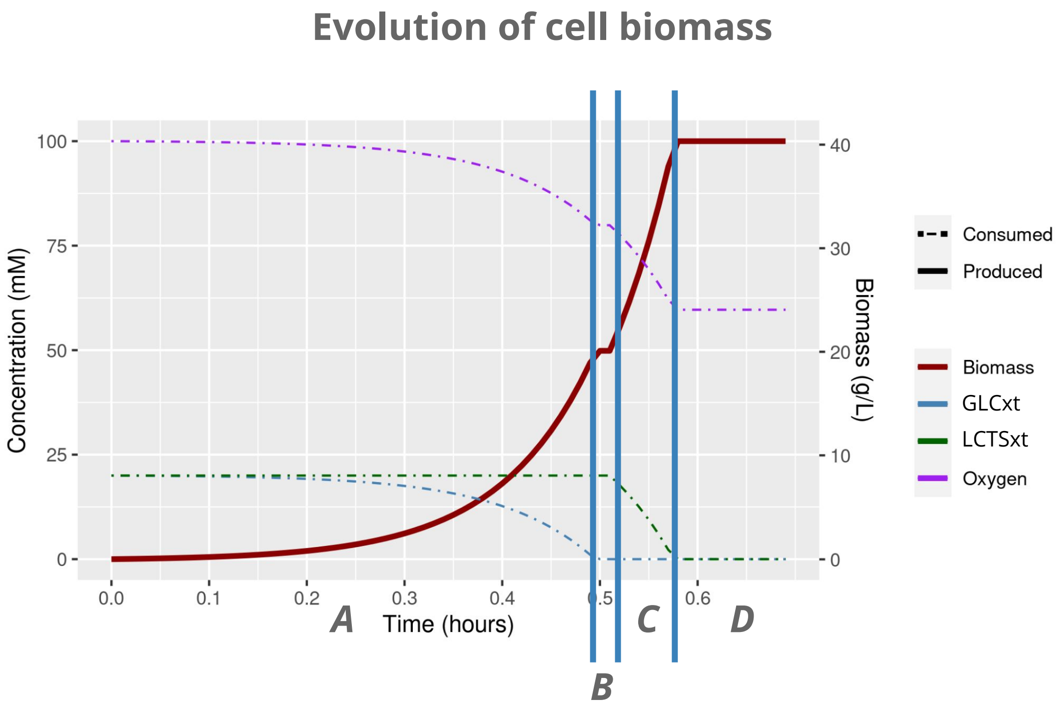


Diauxic shift

→ *Successive growth phases on different mediums*

→ *Control by regulations*

Example: diauxic shift *(Monod et al., 1953)*



Diauxic shift

→ *Successive growth phases on different mediums*

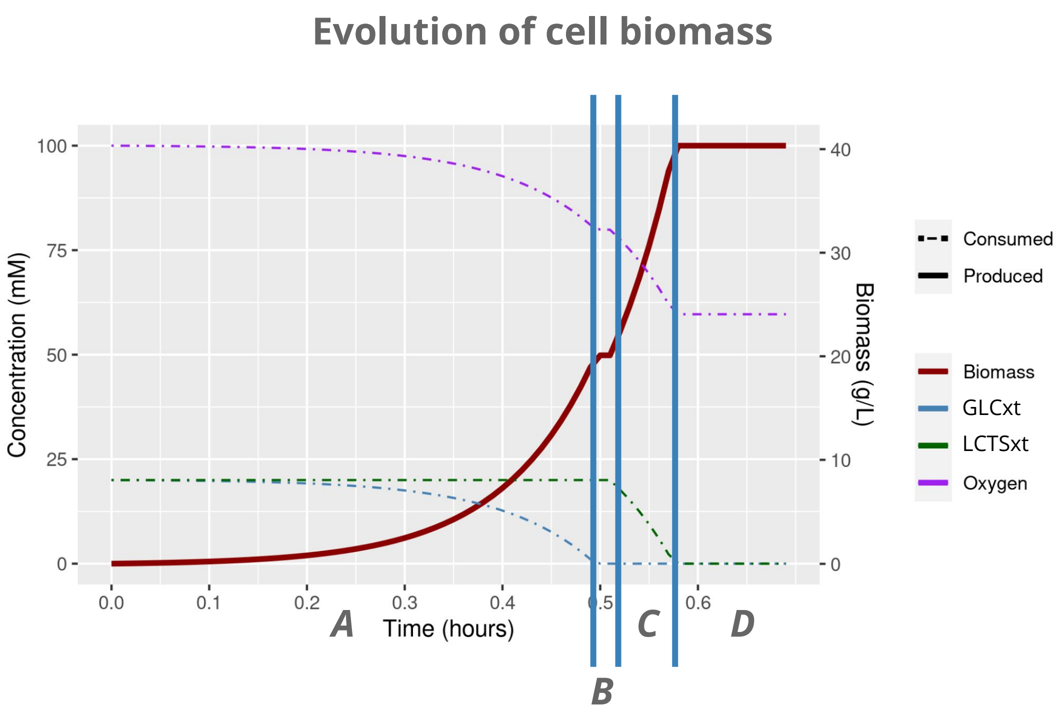
→ *Control by regulations*

Divide in 4 phases

Characterize by different qualitative behaviours (e.g. growth medium)

A → *Growth*	***B*** → ***No growth***
C → *Growth*	***D*** → ***No growth***

Example: diauxic shift *(Monod et al., 1953)*



Diauxic shift

→ *Successive growth phases on different mediums*

→ *Control by regulations*

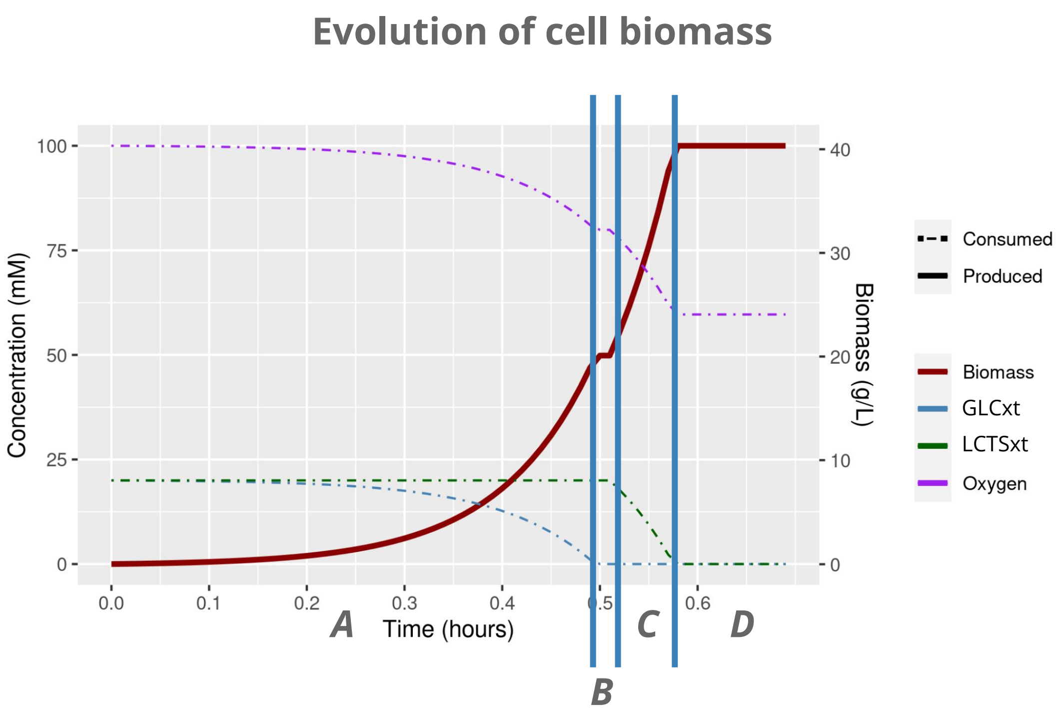
Divide in 4 phases

Characterize by different qualitative behaviours (e.g. growth medium)

A → *Growth*	***B* → *No growth***
C → *Growth*	***D* → *No growth***

Both regulatory system and metabolic system dynamics must be considered to reproduce experimental observations

Example: diauxic shift *(Monod et al., 1953)*

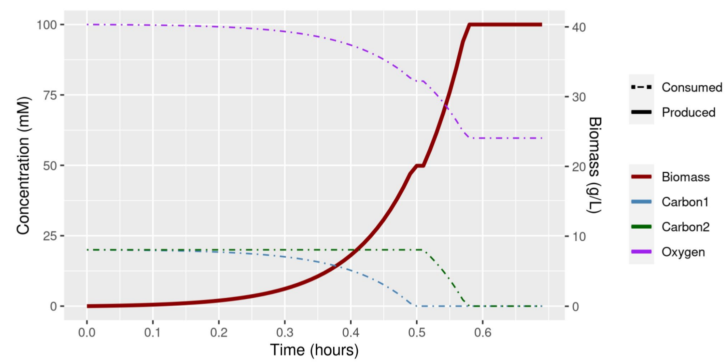


How can we learn the regulatory rules from such observations?

$$f_{\mathrm{LacI}}(x) = \neg x_{\mathrm{LCTSxt}}$$
$$f_{\mathrm{GalR}}(x) = \neg x_{\mathrm{LCTSxt}}$$
$$f_{\mathrm{lacZ}}(x) = \neg x_{\mathrm{GLCxt}} \wedge \neg x_{\mathrm{LacI}}$$
$$f_{\mathrm{galKTEU}}(x) = \neg x_{\mathrm{GLCxt}} \wedge \neg x_{\mathrm{GalR}}$$

Both regulatory system and metabolic system dynamics must be considered to reproduce experimental observations

Time series data

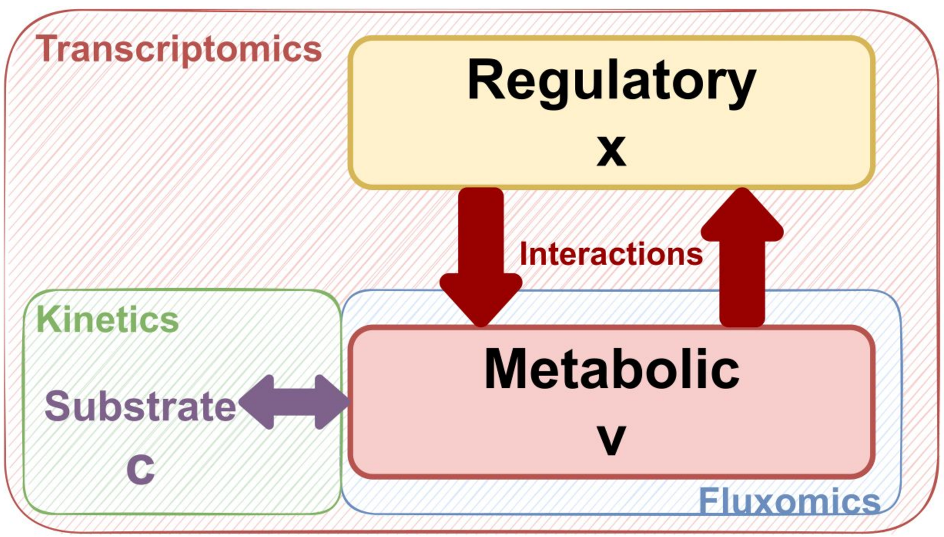


GLCxt	LCTSxt	lacZ	galP	galKTEU	LacI	GalR
0	0	0	0	0	1	1

Observations of the *regulatory* *and* *metabolic* system activities

⟶ ***Quantitative*** *and* ***qualitative*** *measurements*

⟶ ***Compatible*** *with observation from different* ***mutant strains***



3 data types:

➔ **Transcriptomics** (qualitative)
Analysis of the RNA transcripts

➔ **Fluxomics** (quantitative)
Rates of metabolic reactions

➔ **Kinetics** (quantitative)
Substrate concentrations

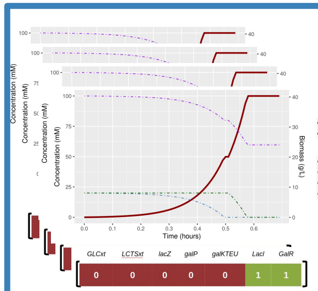
Compatible with any combination of those datatypes

Problems tackled by MERRIN

Inputs:

Metabolic

Metabolic network $\mathcal{N}$



Pair of successive observations of:

1. *Metabolic fluxes* v
2. *Substrate* w
3. *Regulatory state* x

Set of time series $\{T_i\}_i$

(kinetics, fluxomics, transcriptomics)

Set of authorised interactions: activation and inhibition effects

$\text{Hext} = 0$
$\text{Hext} = 1$

$\text{RPh} = 0$
$\text{RPh} = 1$
$\text{RPh} = \text{Hext}$
$\text{RPh} = \neg\text{R8a}$
$\text{RPh} = \text{Hext} \wedge \neg\text{R8a}$
$\text{RPh} = \text{Hext} \vee \neg\text{R8a}$

$\text{R8a} = 0$
$\text{R8a} = 1$
$\text{R8a} = \text{RPh}$

36 compatible regulatory networks

$O(2^{2^n})$ ***in the number*** n ***of interactions***

Prior Knowledge Network (PKN)

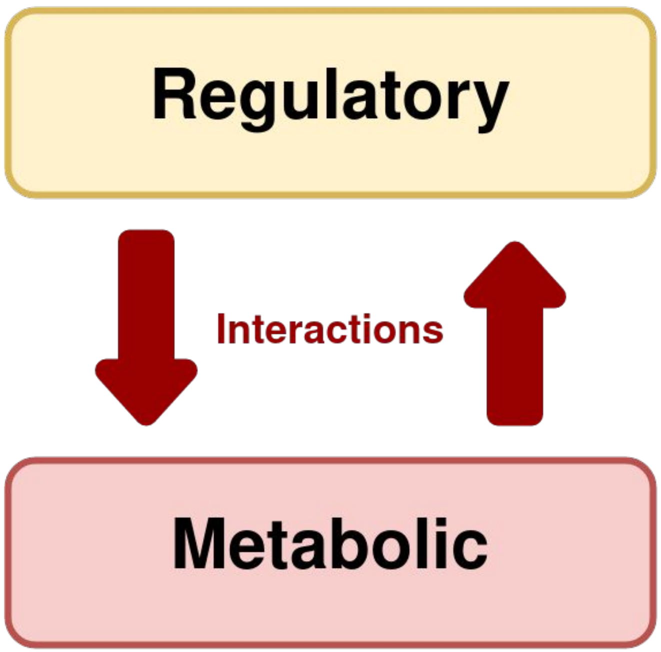
Define a search space $\mathcal{F}$

Outputs:

All the regulatory networks $f \in \mathcal{F}$ compatible with the PKN and matching with the time series $\{T_i\}_i$

Underlying formalism

Regulatory Flux Balance Analysis[1] – rFBA



rFBA timestep:

1. Update the **regulatory system**

   1 synchronous update of the Boolean network

   $$\begin{aligned} f_A(x_A, x_B, x_C, x_D) &= x_A \\ f_B(x_A, x_B, x_C, x_D) &= x_B \\ f_C(x_A, x_B, x_C, x_D) &= x_A \\ f_D(x_A, x_B, x_C, x_D) &= \neg x_A \wedge x_B \end{aligned}$$

2. Update the **metabolic system**

   Solve a FBA — LP problem

   $$\begin{aligned} \text{maximise} \quad & v_{\text{Growth}} \\ \text{such that:} \quad & S \cdot v = 0 \\ & l_r \cdot x_r \leq v_r \leq u_r \cdot x_r \quad \forall r \in \text{reactions} \end{aligned}$$

3. Update the cell environment

[1] M. W. Covert et al., **Journal of theoretical biology**, 2001

Inferring problem — *formal definition*

Input: *metabolic network* $\mathcal{N}$*, PKN* $\mathcal{F}$*, set of time series* $\{T_i\}_i$

Output: *all regulatory networks* $f \in \mathcal{F}$ *such that:*

$$\bigwedge_{T_i} \bigwedge_{(s,s')\in T_i} \Bigg(f(x) = x'$$

$$\wedge \exists \hat{v} \in \mathbb{R}^{\text{Reactions}}, \left(S \cdot \hat{v} = 0 \wedge \bigwedge_{r\in\text{Reactions}} l_r \cdot x'_r \leq \hat{v}_r \leq u_r \cdot x'_r \wedge \hat{v}_{\text{growth}} \geq v'_{\text{growth}} - \epsilon \right)$$

$$\wedge \forall \hat{v} \in \mathbb{R}^{\text{Reactions}}, \left(S \cdot \hat{v} = 0 \wedge \bigwedge_{r\in\text{Reactions}} l_r \cdot x'_r \leq \hat{v}_r \leq u_r \cdot x'_r \right) \implies \hat{v}_{\text{growth}} \leq v'_{\text{growth}} + \epsilon \Bigg)$$

Hybrid problem: combinatorial + quantified linear constraints

Inferring problem — *formal definition*

Input: *metabolic network* $\mathcal{N}$*, PKN* $\mathcal{F}$*, set of time series* $\{T_i\}_i$

Output: *all regulatory networks* $f \in \mathcal{F}$ *such that:*

Boolean constraints

$$\bigwedge_{T_i} \bigwedge_{(s,s')\in T_i} \Bigg(f(x) = x'$$

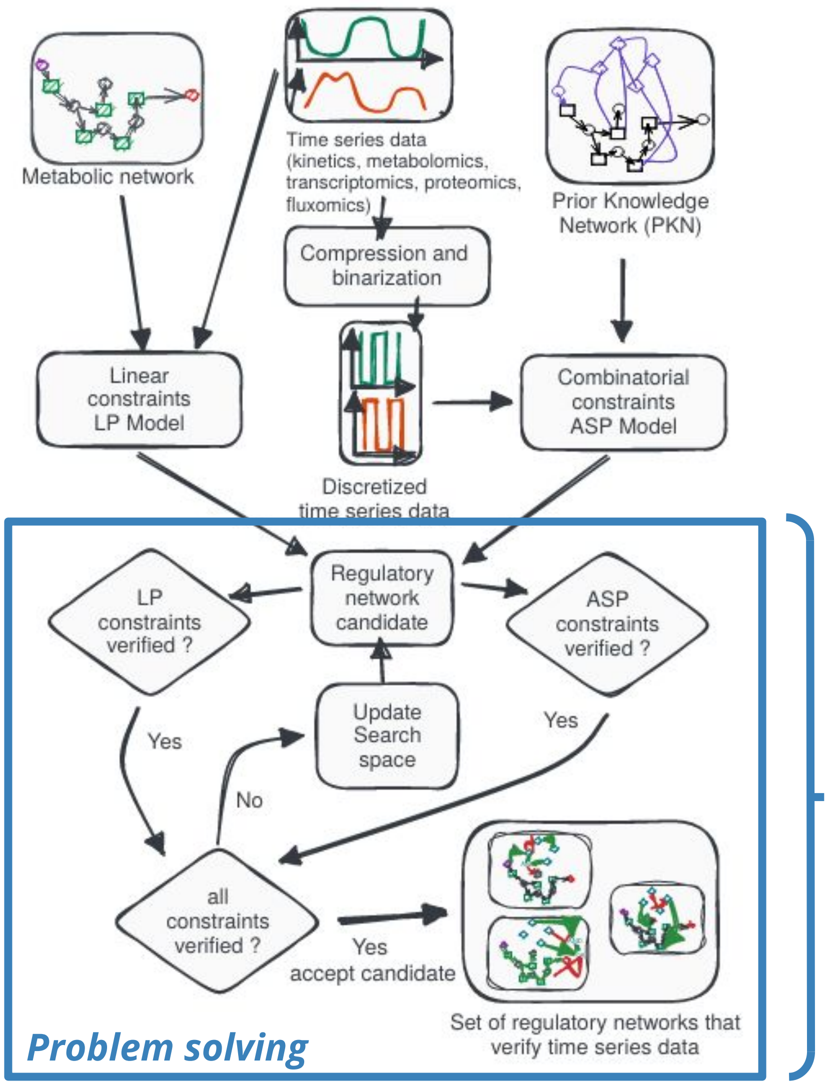
$$\wedge\ \exists \hat{v} \in \mathbb{R}^{\text{Reactions}}, \left(S \cdot \hat{v} = 0 \wedge \bigwedge_{r \in \text{Reactions}} l_r \cdot x'_r \le \hat{v}_r \le u_r \cdot x'_r \wedge \hat{v}_{\text{growth}} \ge v'_{\text{growth}} - \epsilon \right)$$

$$\wedge\ \forall \hat{v} \in \mathbb{R}^{\text{Reactions}}, \left(S \cdot \hat{v} = 0 \wedge \bigwedge_{r \in \text{Reactions}} l_r \cdot x'_r \le \hat{v}_r \le u_r \cdot x'_r \right) \implies \hat{v}_{\text{growth}} \le v'_{\text{growth}} + \epsilon \Bigg)$$

Quantified linear constraints

Hybrid problem: combinatorial + quantified linear constraints

Contribution: MERRIN's workflow

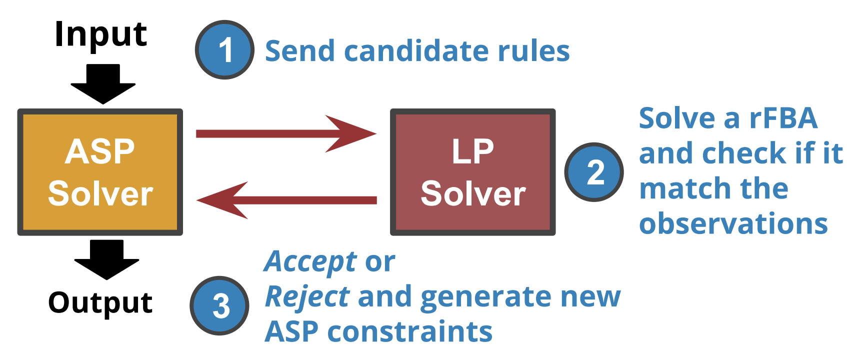


Problem

Input: ***Metabolic network***, ***Prior Knowledge Network (PKN)***, ***Time series data***
+ several solving parameters

Output: *All the subset minimal regulatory metabolic network* ***satisfying the PKN*** *and* ***matching time series data***

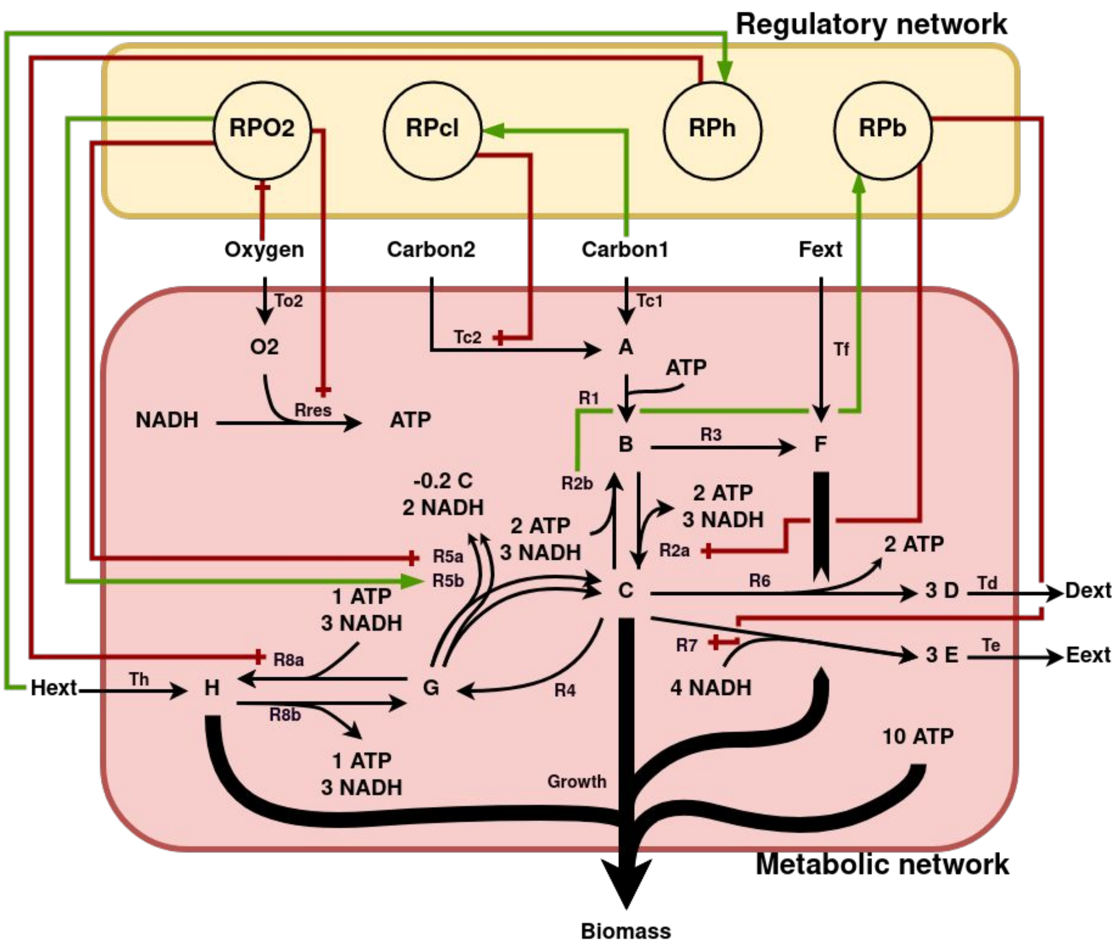
Rely on a hybrid resolution framework[1]



[1] K. Thuillier et al., **Proceedings of the AAAI Conference**, 2024

Gold standard instance *(Covert et al, 2001)*

activation effect *inhibition effect*

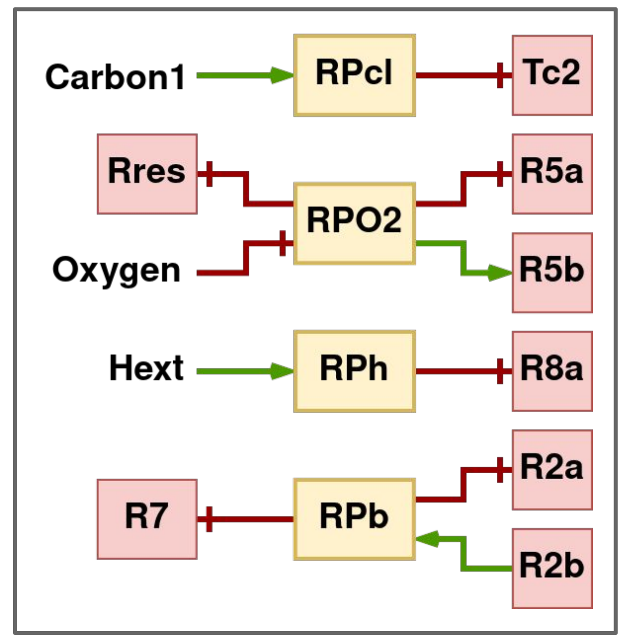


Toy model based on *E.coli*

20 reactions, 4 regulatory protein,
11 regulations

Model complex behaviours

Diauxic shift, aerobic/anaerobic growth, etc.



Influence graph

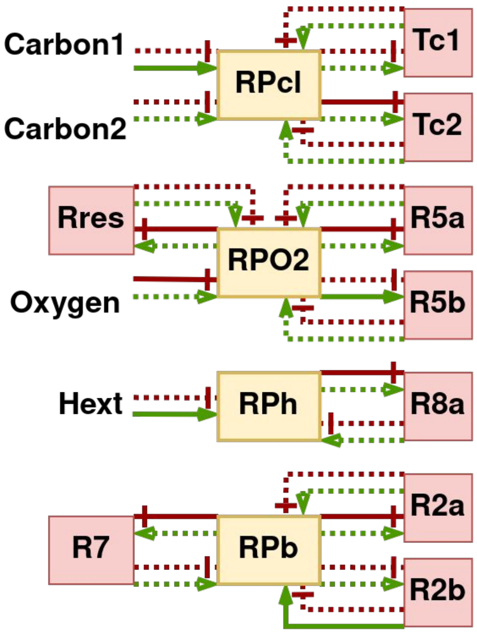
[1] M. W. Covert et al., **Journal of theoretical biology**, 2001

Instance generation

MERRIN inputs

Prior Knowledge Network

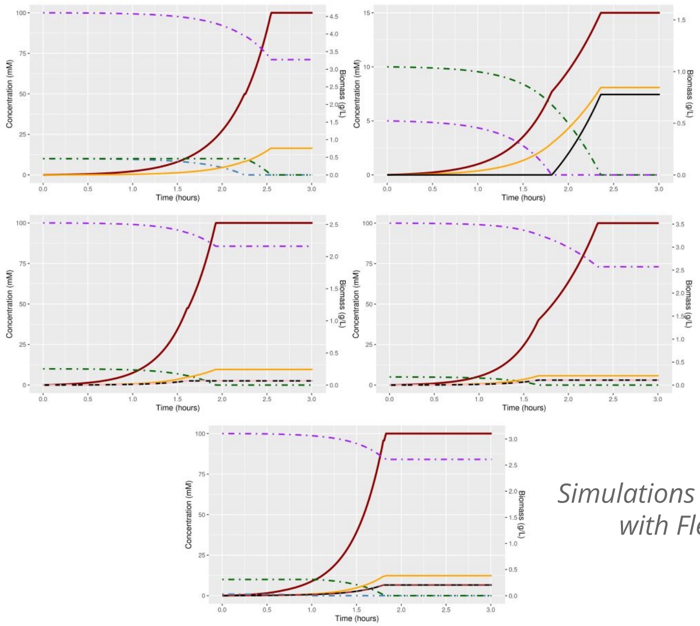
Add hypothetical regulations (e.g. RPcl and Tc1)
Remove sign + direction of interactions



~2.9x10^{12} BNs compatibles

5 simulations[1]

Kinetics, fluxomics and transcriptomics
Perfect observations (no noise)

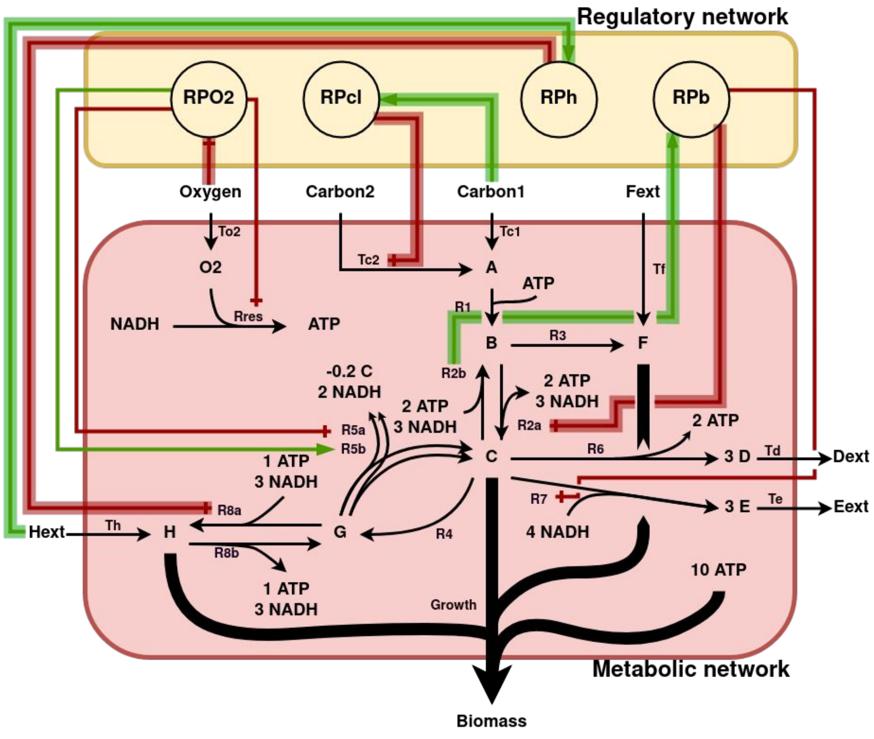


Simulations made with FlexFlux

240 instances with 6 degradation levels

[1] M. W. Covert et al., **Journal of theoretical biology**, 2001

Validation and robustness testing



Learn more parsimonious model than ground truth

- Reproduce exactly the input time series
- Unrecovered regulations can be explained

Validation on a benchmark of 240 instances (*in silico*)

- 4 data types
- 6 level of degradations (0% to 50%)

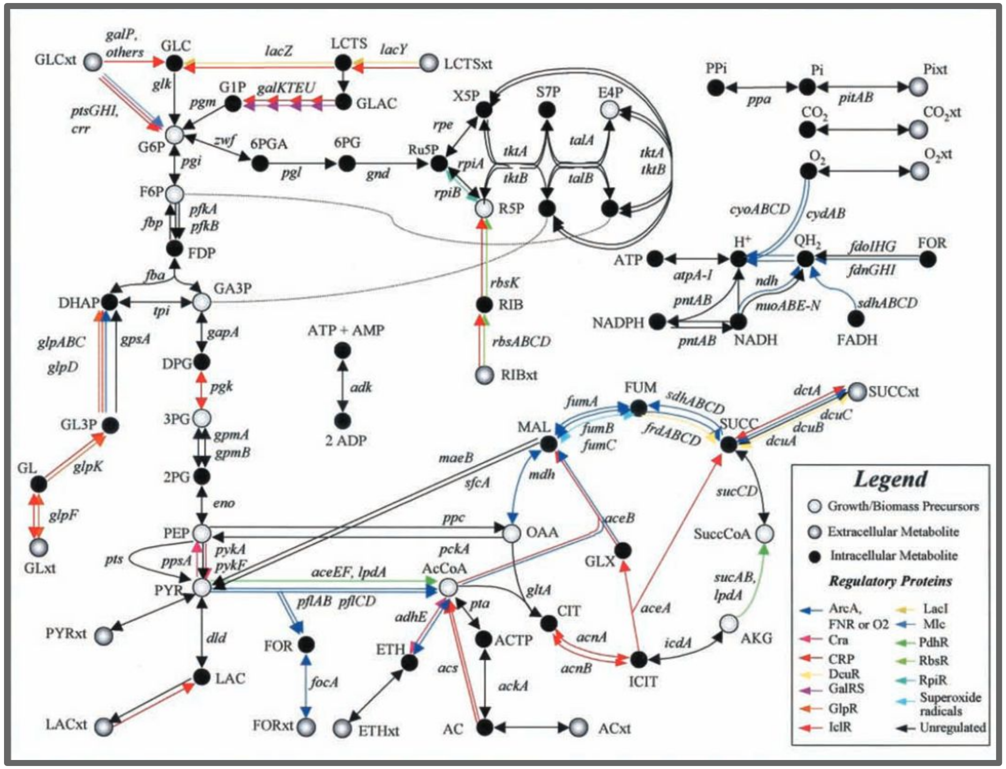
Perfectly reproduce the time series with:

- *kinetics* and *transcriptomics* data
- < *20%* of degradation

[1] M. W. Covert et al., **Journal of theoretical biology**, 2001

Scalability of MERRIN

E.coli medium scale instance[1]



Metabolic network[1]

Description:

- 60 regulatory rules
  - *19 regulatory proteins*
  - *41 (regulated) genes*
- 113 reactions
  - 66 are reversible

3 experimental conditions[1]

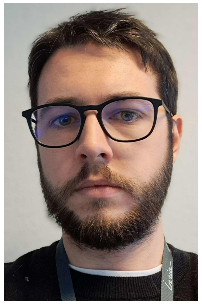
- rFBA time series *in silico*
- ***Mutant strains***

Computation time: *~15 minutes*

MERRIN scales on bigger models

[1] M. W. Covert and B. Ø. Palsson, **Journal of biological chemistry**, 2002

Conclusion



- **MERRIN[1]: inferring regulatory rules from metabolic traces**
  → ***Hybrid (ASP + LP) resolution*** *based on SMT approaches*
  → *Compatible with* ***kinetics, fluxomics and/or transcriptomics data***
  → *Compatible with* ***mutant strains***

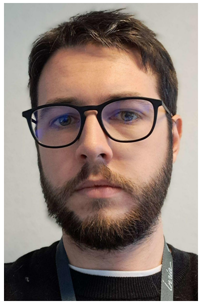
- **Validation on simulated benchmark**
  → *Find smaller RN than gold standard — Consistent with state of the art*
  → *Study the* ***impact of noise and data type*** *on the inferring — 240 instances*

- **Scalability**
  → *E.coli medium scale instance*

[1] Implementation available on *https://github.com/bioasp/merrin/*

Perspective — Work In Progress —

Model correction with MERRIN



Description: | *Escherichia coli str. K-12 substr. MG1655*

- 1.473 regulatory rules
  - *600 genes and regulatory proteins*
  - *873 regulated reactions*
- 1075 reactions

New input datatype: *Biolog data*[1] [2]

- 111 mutant strains
- 124 mediums

| ***Work in progress***

| **13.764 observations**
(*in silico* and *in vivo*)

Existing models can be incompatible with new experimental results

— How to update them? —

[1] M. W. Covert and B. Ø. Palsson, **Nature**, 2004
[2] J. D. Glasner et al. **Nucleic Acids Res.**, 2003