How to learn metabolic regulation rules from time series data?

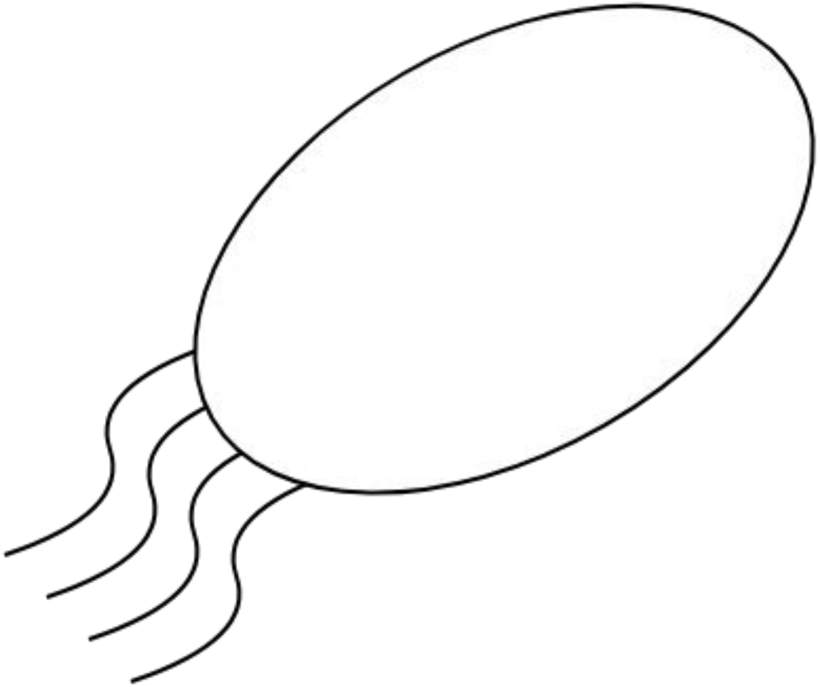
MERRIN: *MEtabolic Regulation Rules INference from time series data*

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

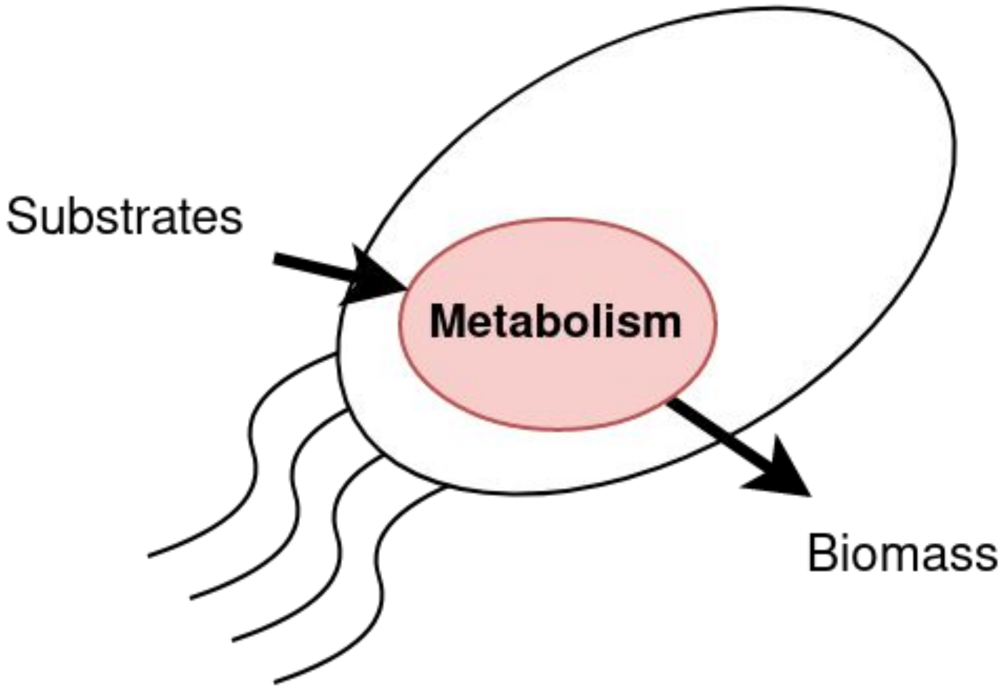
17th November 2022

Cells: hybrid multi-layered structures



Model as two interconnected systems

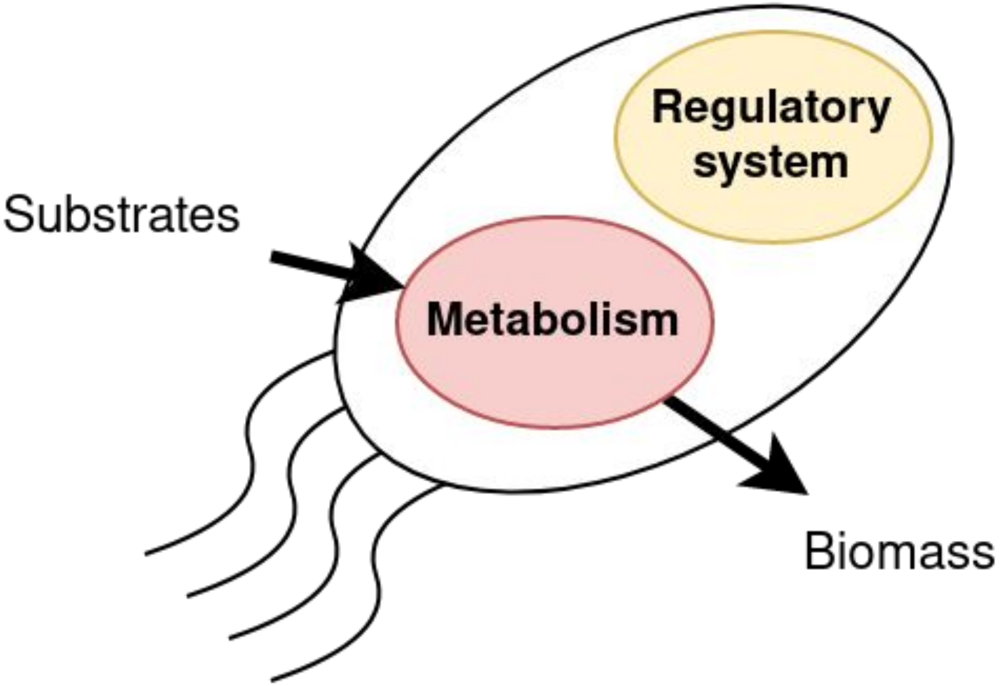
Cells: hybrid multi-layered structures



Model as two interconnected systems

1. **Metabolic system**
   Chemicals reactions converting substrates to energy and biomass

Cells: hybrid multi-layered structures

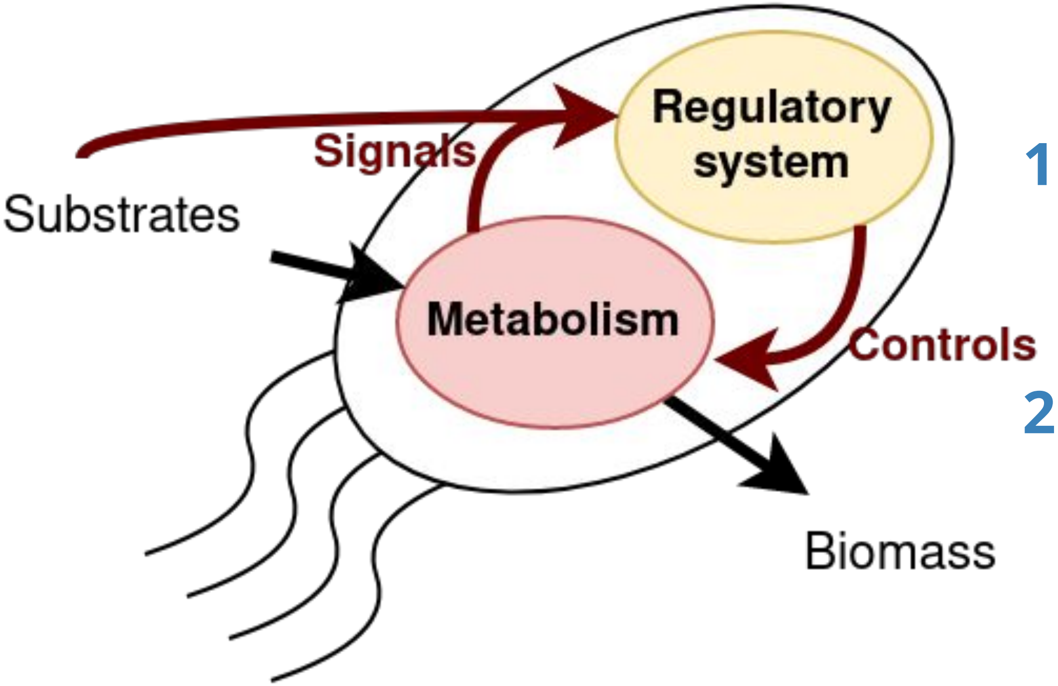


Model as two interconnected systems

1. **Metabolic system**
   Chemicals reactions converting substrates to energy and biomass

2. **Regulatory system**
   Rules constraining the metabolism to adapt itself to its environment

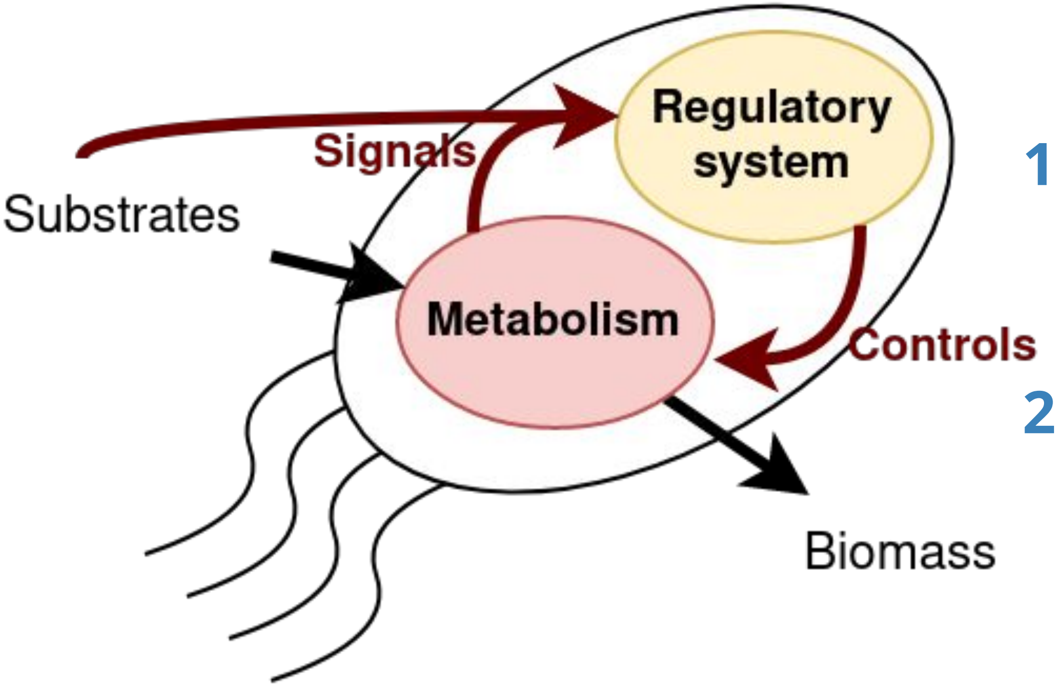
Cells: hybrid multi-layered structures



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

Cells: hybrid multi-layered structures



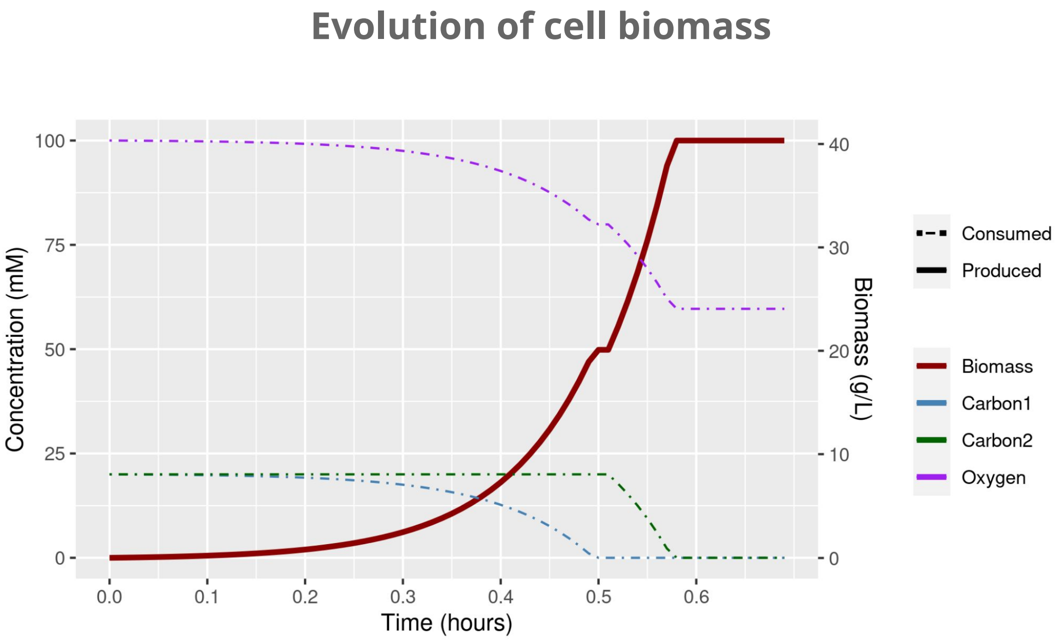
Model as two interconnected systems

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 system** from time series observations of the cells *(metabolism and regulation)*

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

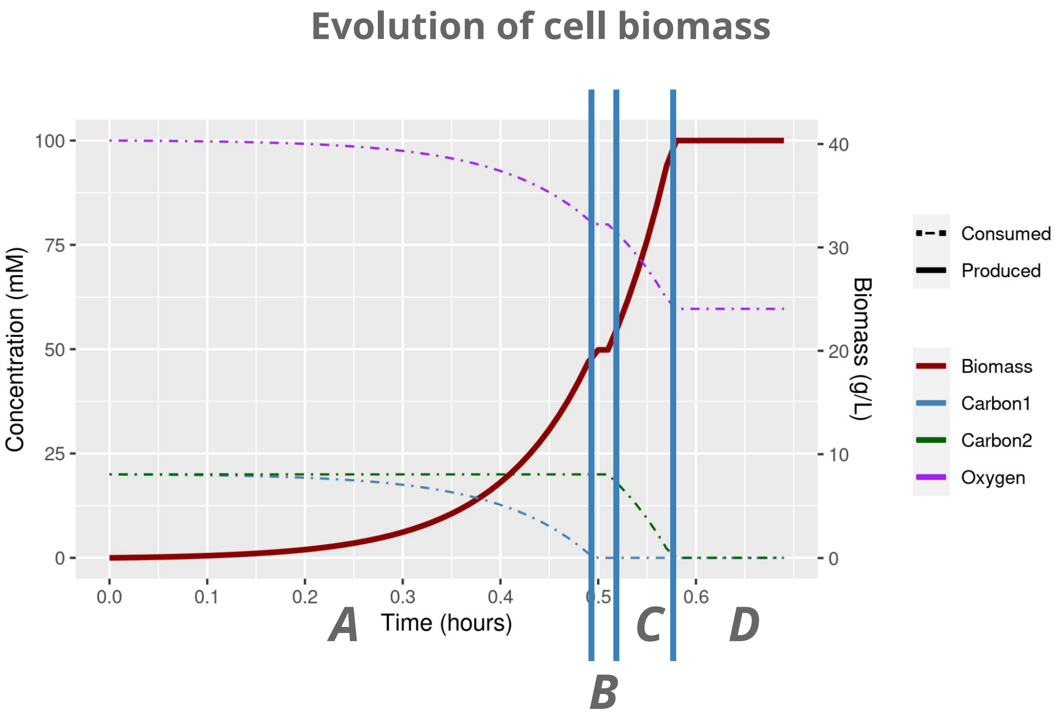


Diauxic shift

⟶ *Successive growth phases on different mediums*

⟶ *Control by regulations*

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



Evolution of cell biomass

Diauxic shift

⟶ Successive growth phases on different mediums

⟶ Control by regulations

Divided in 4 phases

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

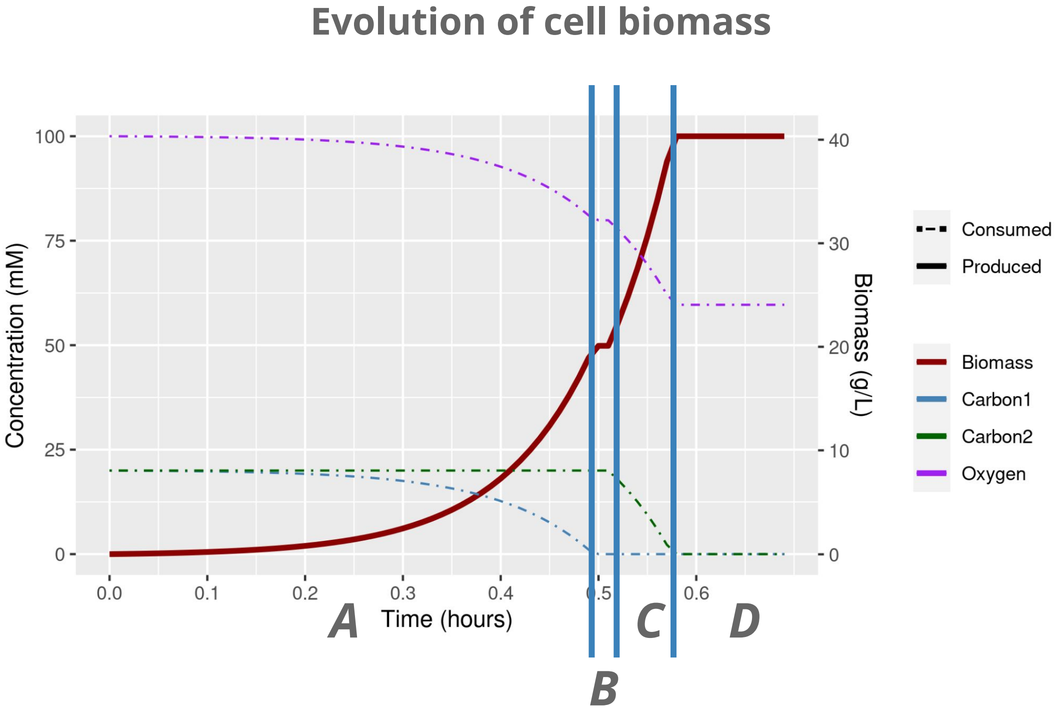
A *⟶ Growth on* ***Carbon1 only***

B *⟶* ***No growth*** *due to regulations*

C *⟶ Growth on* ***Carbon2 only***

D *⟶* ***No growth****, no growth medium*

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



Diauxic shift

→ Successive growth phases on different mediums

→ Control by regulations

Divided in 4 phases

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

A *→ Growth on* ***Carbon1 only***

B *→* ***No growth*** *due to regulations*

C *→ Growth on* ***Carbon2 only***

D *→* ***No growth****, no growth medium*

Objective:

Learning all the regulatory rules explaining the cell behaviour

Taking knowledge into account

Only specific interactions (*activations*, *inhibitions*) between proteins and enzymes can be used in Boolean regulatory rules

Taking knowledge into account

Only specific interactions (*activations*, *inhibitions*) between proteins and enzymes can be used in Boolean regulatory rules

Prior Knowledge Network:

Set of authorised interactions: activation and inhibition effects

Example:

Regulatory rule of RPh can only depend on:

- ***activation of Hext***
- ***inhibition of R8a***

Taking knowledge into account

Only specific interactions (*activations*, *inhibitions*) between proteins and enzymes can be used in Boolean regulatory rules

Prior Knowledge Network:

Set of authorised interactions: activation and inhibition effects

Example:

$$\underbrace{\begin{array}{l}\mathrm{RPh} = 0\\ \mathrm{RPh} = 1\\ \mathrm{RPh} = \mathrm{Hext}\\ \mathrm{RPh} = \neg\mathrm{R8a}\\ \mathrm{RPh} = \mathrm{Hext} \wedge \neg\mathrm{R8a}\\ \mathrm{RPh} = \mathrm{Hext} \vee \neg\mathrm{R8a}\end{array}}_{\textit{6 rules}}$$

Regulatory rule of RPh can only depend on:
- ***activation of Hext***
- ***inhibition of R8a***

***6 potential Boolean regulatory rules** for these 2 interactions*

Taking knowledge into account

Only specific interactions (*activations*, *inhibitions*) between proteins and enzymes can be used in Boolean regulatory rules

Prior Knowledge Network:

Set of authorised interactions: activation and inhibition effects

Example:

$$\underbrace{\begin{array}{l}\mathrm{Hext}=0\\ \mathrm{Hext}=1\end{array}}_{\textit{2 rules}}
\qquad
\underbrace{\begin{array}{l}\mathrm{RPh}=0\\ \mathrm{RPh}=1\\ \mathrm{RPh}=\mathrm{Hext}\\ \mathrm{RPh}=\neg\mathrm{R8a}\\ \mathrm{RPh}=\mathrm{Hext}\wedge\neg\mathrm{R8a}\\ \mathrm{RPh}=\mathrm{Hext}\vee\neg\mathrm{R8a}\end{array}}_{\textit{6 rules}}
\qquad
\underbrace{\begin{array}{l}\mathrm{R8a}=0\\ \mathrm{R8a}=1\\ \mathrm{R8a}=\mathrm{RPh}\end{array}}_{\textit{3 rules}}$$

Regulatory rule of RPh can only depend on:
- activation of Hext
- inhibition of R8a

***6 potential Boolean regulatory rules** for these 2 interactions*

$2 \times 6 \times 3 = 36$ ***compatible regulatory networks***

Taking knowledge into account

Only specific interactions (*activations*, *inhibitions*) between proteins and enzymes can be used in Boolean regulatory rules

Prior Knowledge Network:

Set of authorised interactions: activation and inhibition effects

Example:

$$\underbrace{\begin{array}{l}\mathrm{Hext} = 0 \\ \mathrm{Hext} = 1\end{array}}_{\text{2 rules}}$$

$$\underbrace{\begin{array}{l}\mathrm{RPh} = 0 \\ \mathrm{RPh} = 1 \\ \mathrm{RPh} = \mathrm{Hext} \\ \mathrm{RPh} = \neg\mathrm{R8a} \\ \mathrm{RPh} = \mathrm{Hext} \wedge \neg\mathrm{R8a} \\ \mathrm{RPh} = \mathrm{Hext} \vee \neg\mathrm{R8a}\end{array}}_{\text{6 rules}}$$

$$\underbrace{\begin{array}{l}\mathrm{R8a} = 0 \\ \mathrm{R8a} = 1 \\ \mathrm{R8a} = \mathrm{RPh}\end{array}}_{\text{3 rules}}$$

Regulatory rule of RPh can only depend on:

- activation of Hext

- inhibition of R8a

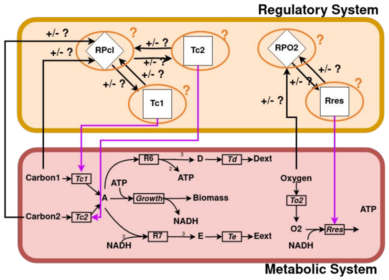
***6 potential Boolean regulatory rules** for these 2 interactions*

Size of the search space

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

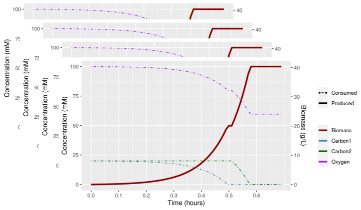
Problems tackled by MERRIN

Inferring regulatory rules from time series observations



Prior Knowledge Network

+



Time series
(kinetics, fluxomics, transcriptomics)

=?

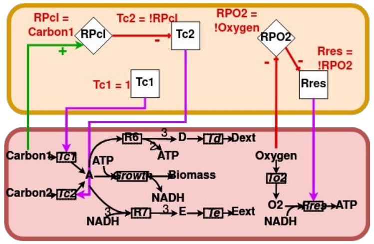
Input:

➔ **Prior Knowledge Network (PKN)**

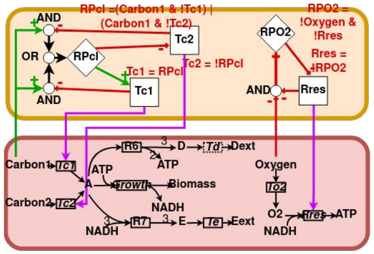
Set of admissible interactions between components of the regulatory network

➔ **Time series data**

Kinetics, fluxomics and/or transcriptomics



Solution n°1



Solution n°2

Output:

➔ **Set of consistent regulatory networks**

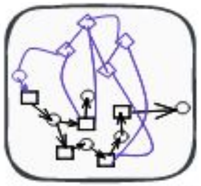
Respecting the admissible interactions
Allowing to reproduce the input time series

State of the art: learning regulatory rules

Several inferring approaches[1,2]

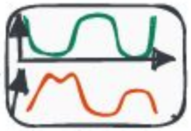
→ Based on **constraint programming**
→ **Discrete modelling** of the regulatory system dynamics
→ **Observations of the regulatory system + PKN**

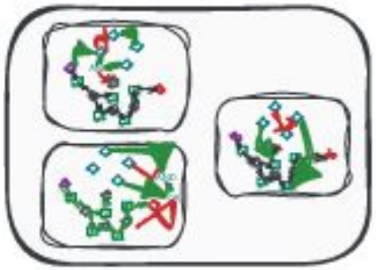
PKN



Observations

Regulatory





Set of regulatory networks that verify time series data

Infer the regulatory systems

---

[1] M. Ostrowski et al., **BioSystems**, 2016 – *CASPOTS*
[2] L. Paulevé et al., **Nature Communication**, 2020

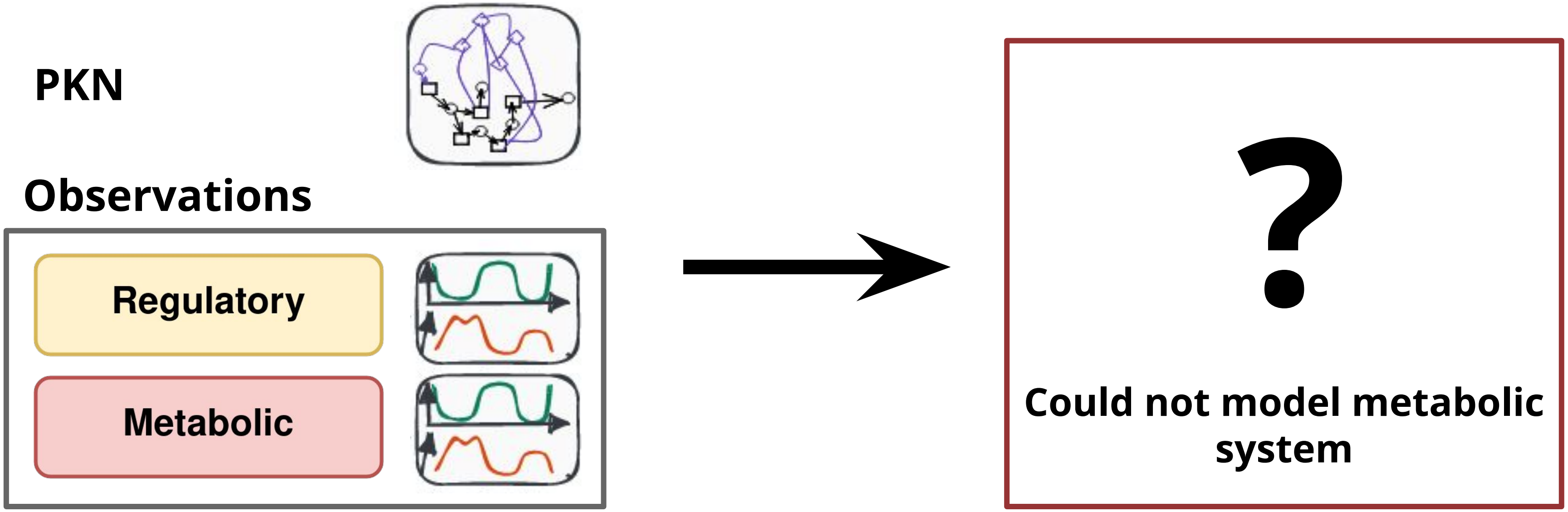
State of the art: learning regulatory rules

Several inferring approaches[1,2]

→ Based on **constraint programming**

→ **Discrete modelling** of the regulatory system dynamics

→ **Observations of the regulatory system + PKN**



PKN

Observations

Regulatory

Metabolic

?

Could not model metabolic system

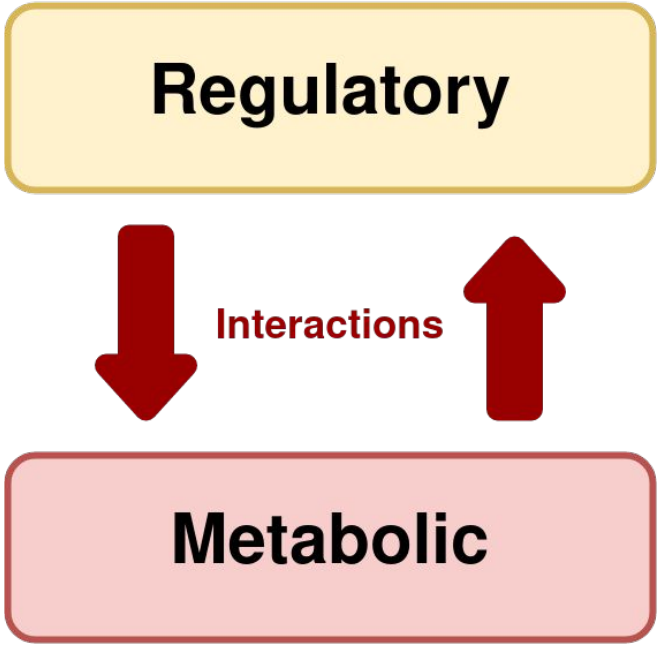
---

[1] M. Ostrowski et al., **BioSystems**, 2016 – *CASPOTS*

[2] L. Paulevé et al., **Nature Communication**, 2020

Multiplicity of modelling formalisms

Two models with different dynamics

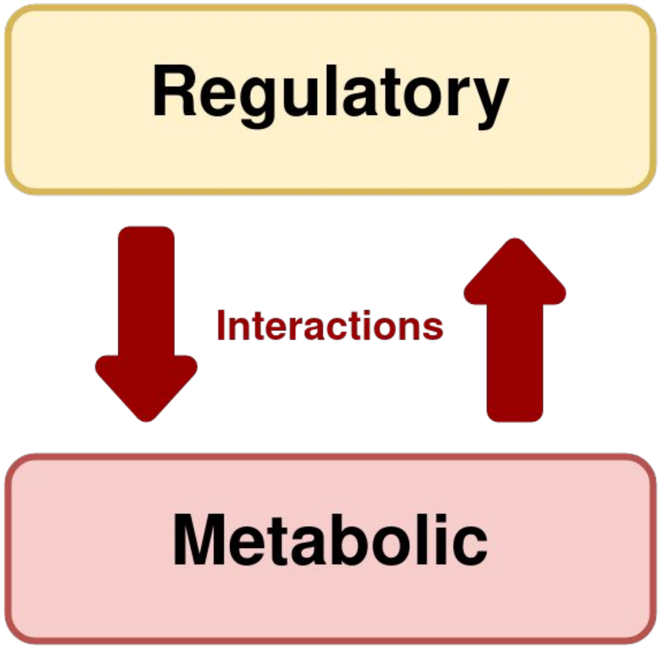

Regulatory

Interactions

Metabolic

Multiplicity of modelling formalisms

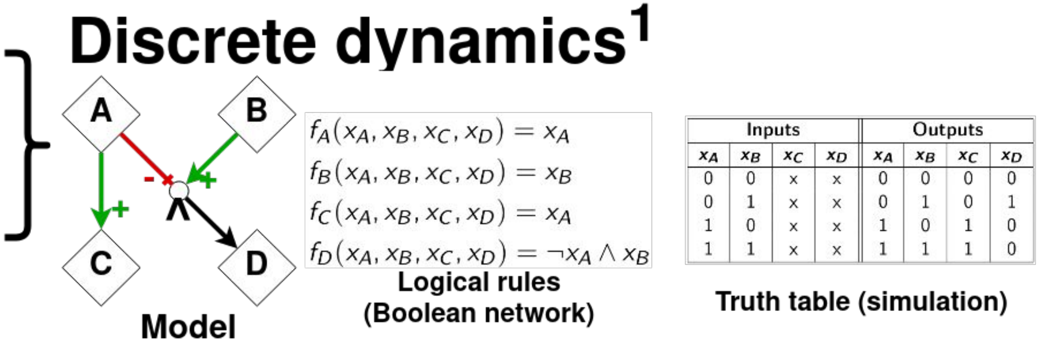
Two models with different dynamics



Regulatory

Interactions

Metabolic



Discrete dynamics[1]

$f_A(x_A, x_B, x_C, x_D) = x_A$

$f_B(x_A, x_B, x_C, x_D) = x_B$

$f_C(x_A, x_B, x_C, x_D) = x_A$

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

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

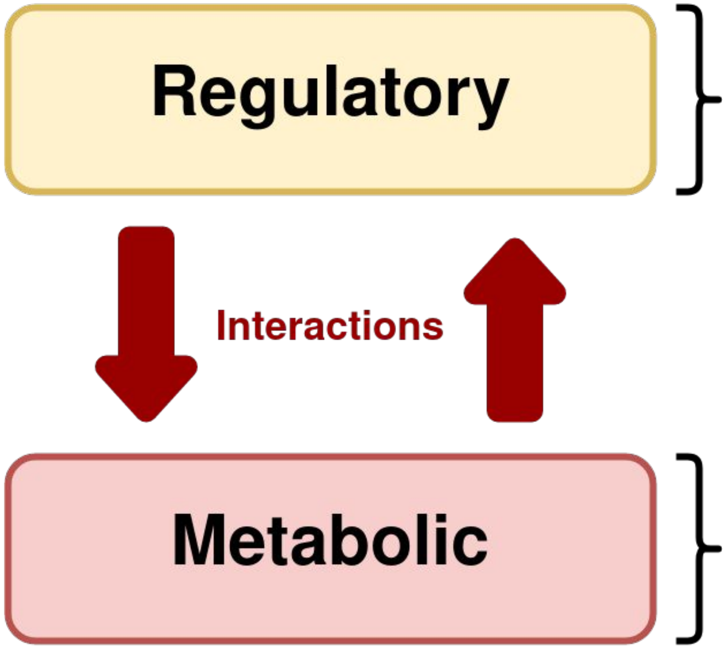
Truth table (simulation)

---

[1] S. Videla et al., **Bioinformatics**, 2016

Multiplicity of modelling formalisms

Two models with different dynamics

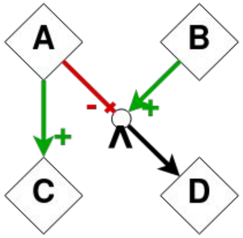


Regulatory

Interactions

Metabolic

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)

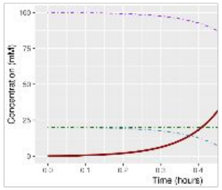
Inputs				Outputs			
x_A	x_B	x_C	x_D	x_A	x_B	x_C	x_D
0	0	x	x	0	0	0	0
0	1	x	x	0	1	0	1
1	0	x	x	1	0	1	0
1	1	x	x	1	1	1	0

Truth table (simulation)

Steady-states approximation[2]

$$\boxed{\begin{array}{ll} \text{maximise} & x_{\text{Growth}} \\ \text{such that:} & S \cdot x = 0 \\ & l_r \cdot f_r \leq x_r \leq u_r \cdot f_r \quad \forall r \in \mathcal{R} \end{array}}$$

Dynamic flux balance analysis (dFBA)



Metabolic traces

---

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

Contribution: MERRIN[1]

Input:

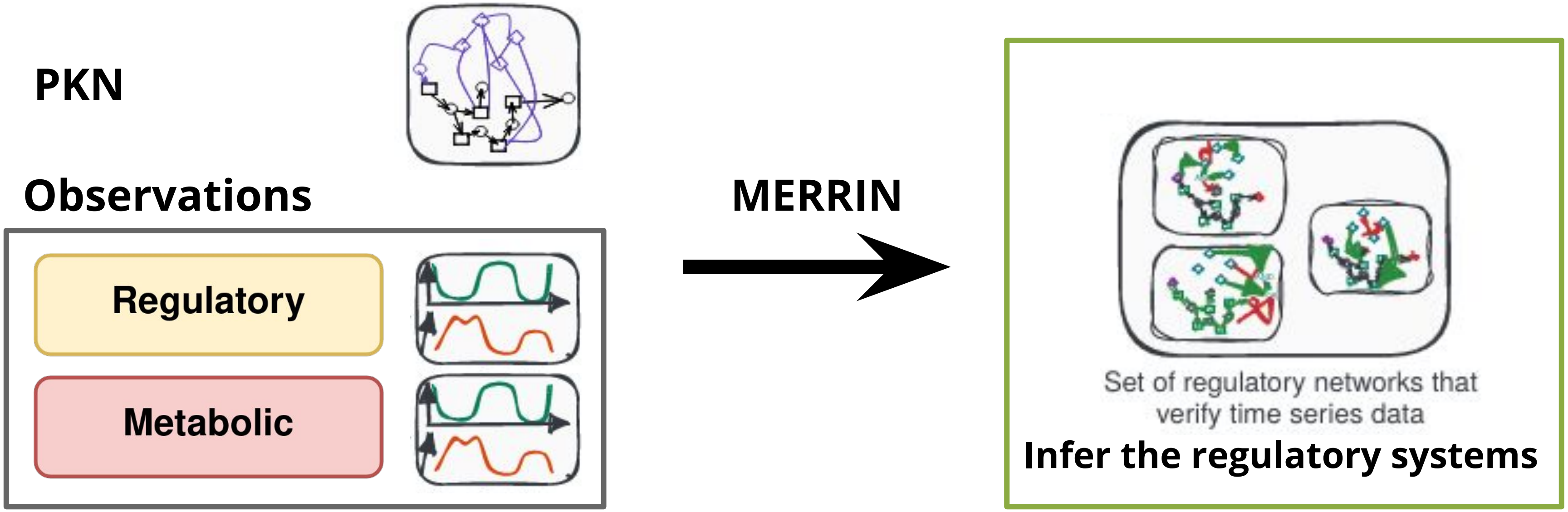
⟶ *Metabolic network,*

⟶ *Prior Knowledge Network (PKN),*

⟶ *Time series data*

Output:

All regulatory networks s.t:
⟶ ***consistent with the PKN***
⟶ *and **time series data***



PKN

Observations

Regulatory

Metabolic

MERRIN

Set of regulatory networks that verify time series data

Infer the regulatory systems

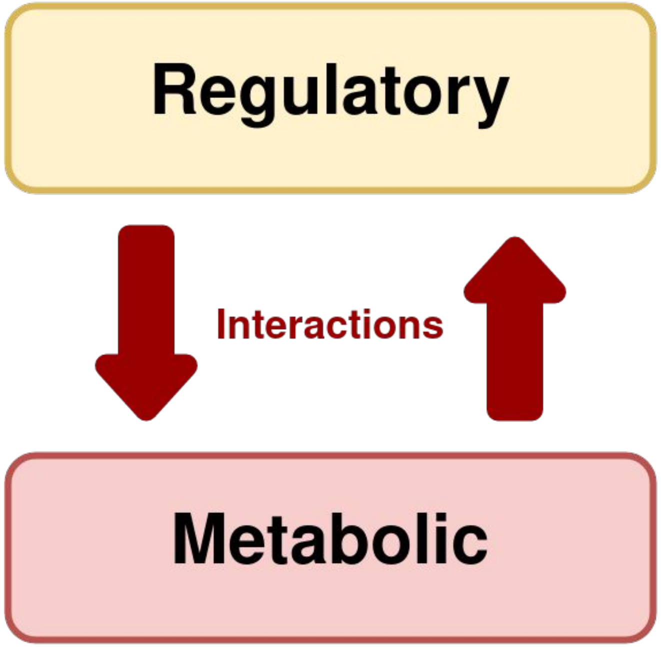
---

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

Underlying simulation formalism

Several simulations approaches

*Based on **regulatory Flux Balance Analysis**[1] (dynamic + regulations)*



rFBA timestep:

1. Update the **regulatory system**

   1 synchronous update of the Boolean network

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

2. Update the **metabolic system**

   Solve FBA — LP problem

   $$\boxed{\begin{array}{lll} \text{maximise} & x_{\text{Growth}} & \\ \text{such that:} & S \cdot x = 0 & \\ & l_r \cdot f_r \leq x_r \leq u_r \cdot f_r & \forall r \in \mathcal{R} \end{array}}$$

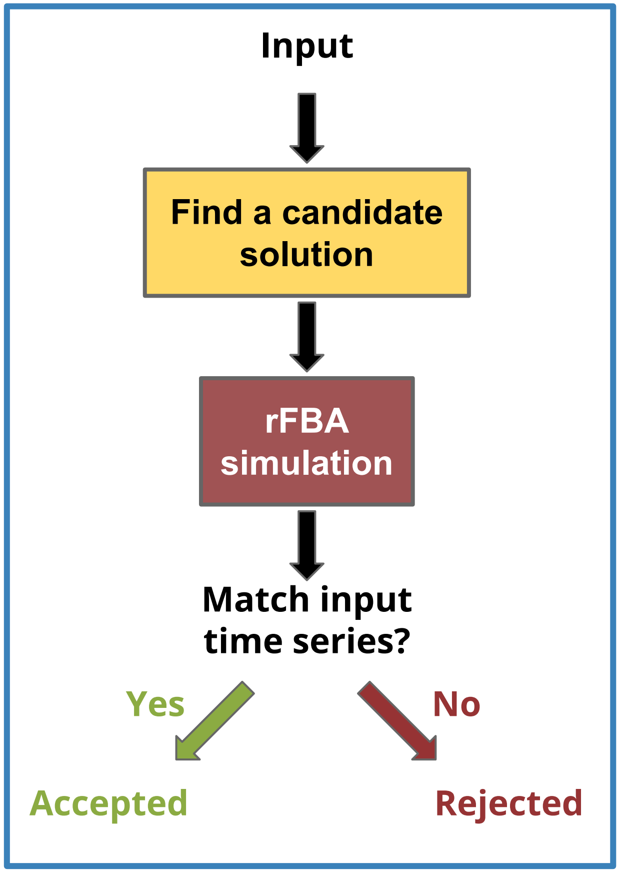
3. Update the cell environment

---

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

Inferring workflow

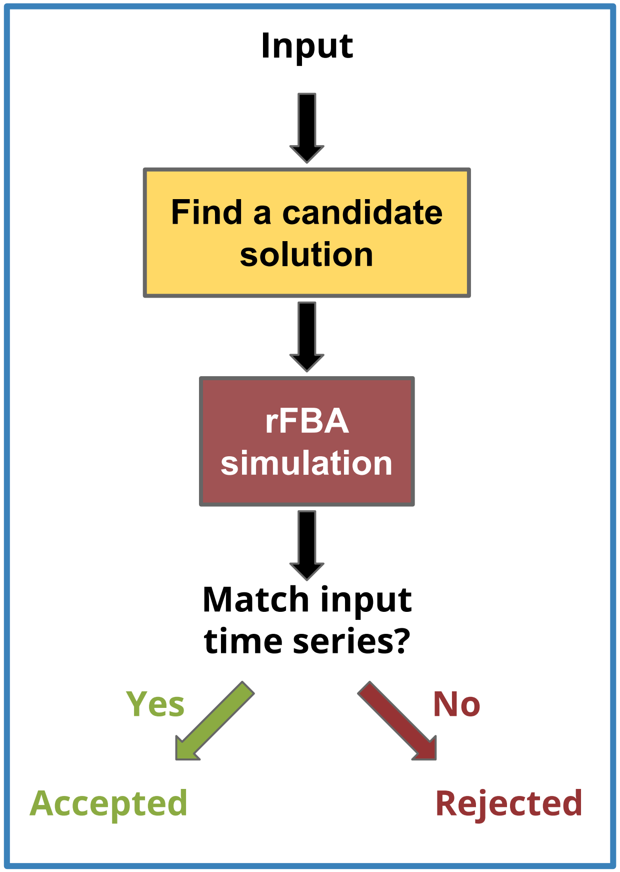
rFBA is directly used in the inferring process



Inferring workflow in a nutshell

Inferring workflow

rFBA is directly used in the inferring process



Input

Find a candidate solution

rFBA simulation

Match input time series?

Yes

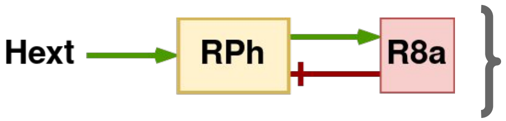
No

Accepted

Rejected

Inferring workflow in a nutshell

No issue for small instance

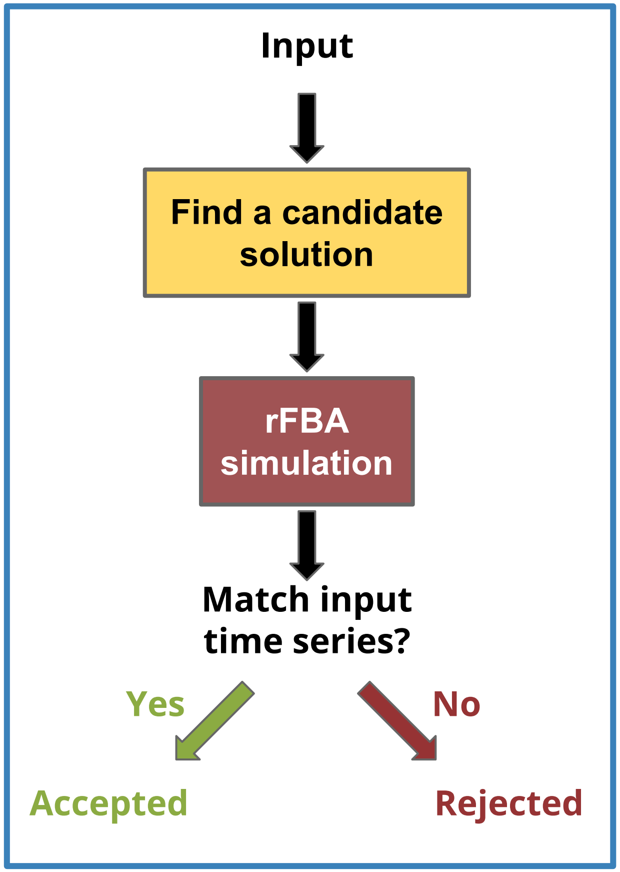


36 compatible regulatory networks

At most 36 rFBA simulations

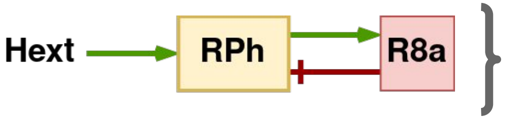
Inferring workflow

rFBA is directly used in the inferring process



Inferring workflow in a nutshell

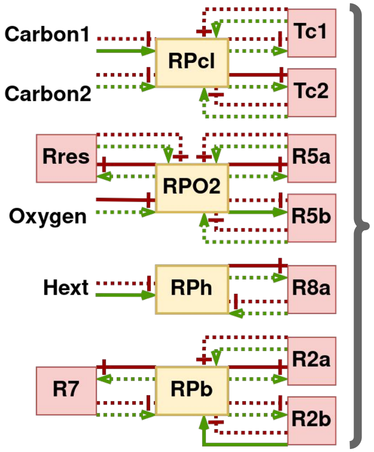
No issue for small instance



36 compatible regulatory networks

At most 36 rFBA simulations

Not tractable with bigger ones

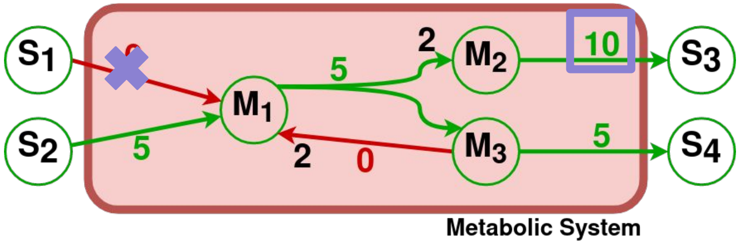


~2.9x10^{12} compatible regulatory networks

Solving ~2.9x10^{12} rFBA simulations is not tractable

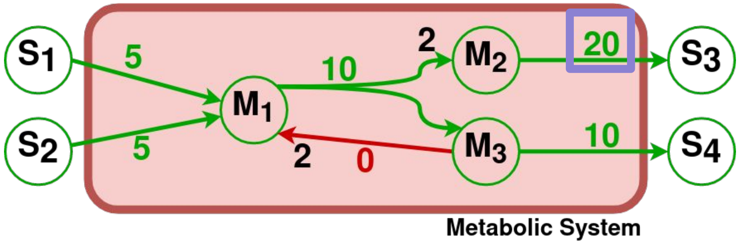
Counter-examples generalisation

Monotone property over network structure



With *an inhibited reaction*
Less biomass produced

≤

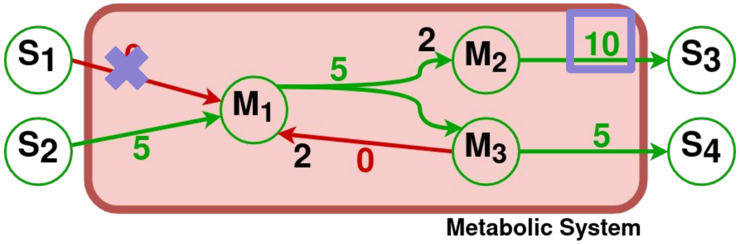


Without *an inhibited reaction*
More biomass produced

Inhibiting a new reaction can only decrease the biomass production

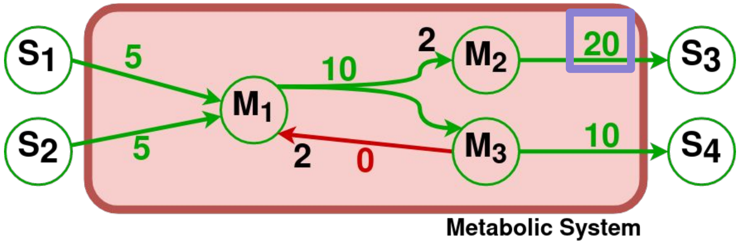
Counter-examples generalisation

Monotone property over network structure



With an inhibited reaction
Less biomass produced

$\leq$



Without *an inhibited reaction*
More biomass produced

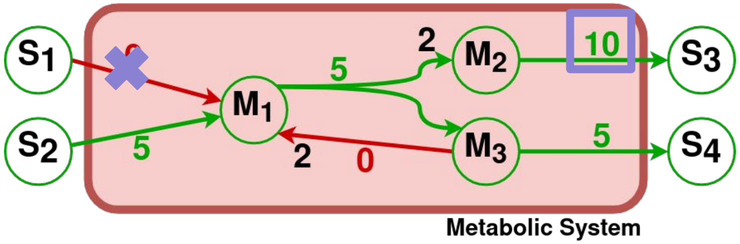
New constraints for candidate solution can be added

- **Too much biomass produced → not enough inhibited reactions**
  *All **subsets of inhibited reactions** will produce too much biomass*

Inhibiting a new reaction can only decrease the biomass production

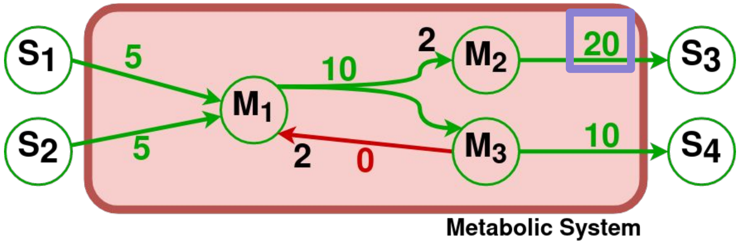
Counter-examples generalisation

Monotone property over network structure



With *an inhibited reaction*
Less biomass produced

$\leq$



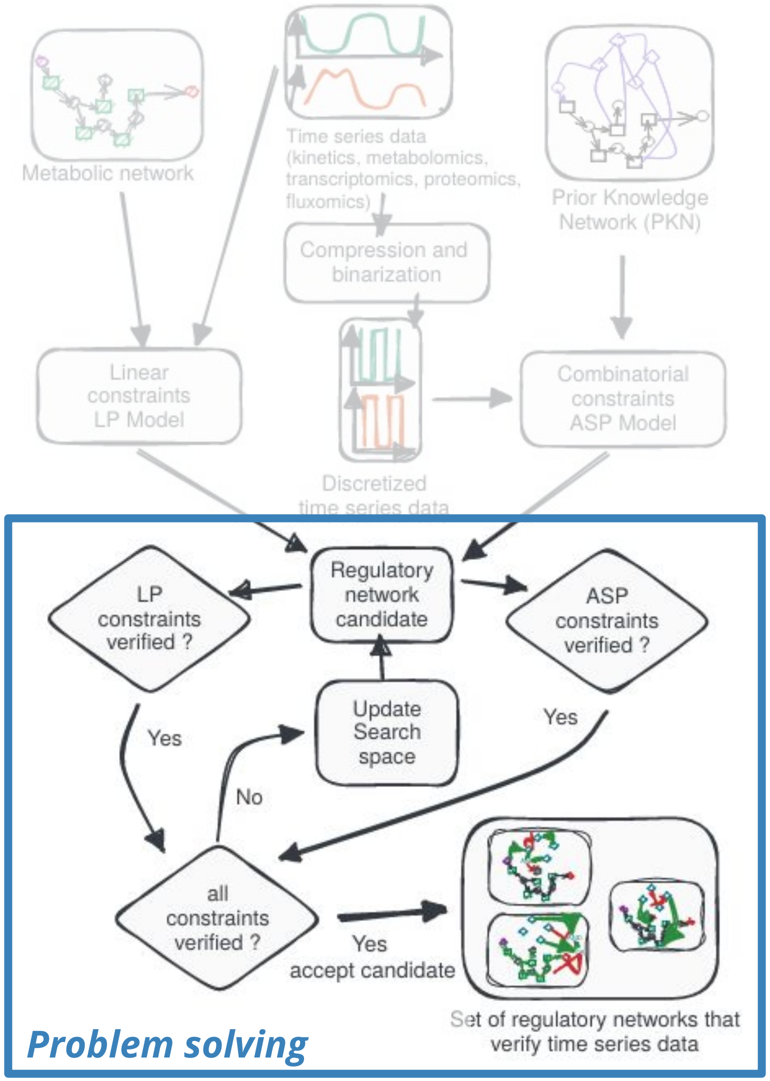
Without *an inhibited reaction*
More biomass produced

New constraints for candidate solution can be added

- **Too much biomass produced → not enough inhibited reactions**
  All ***subsets of inhibited reactions*** *will produce too much biomass*
- **Not enough biomass produced → too much inhibited reactions**
  All ***supersets of inhibited reactions*** *will produce not enough biomass*

Inhibiting a new reaction can only decrease the biomass production

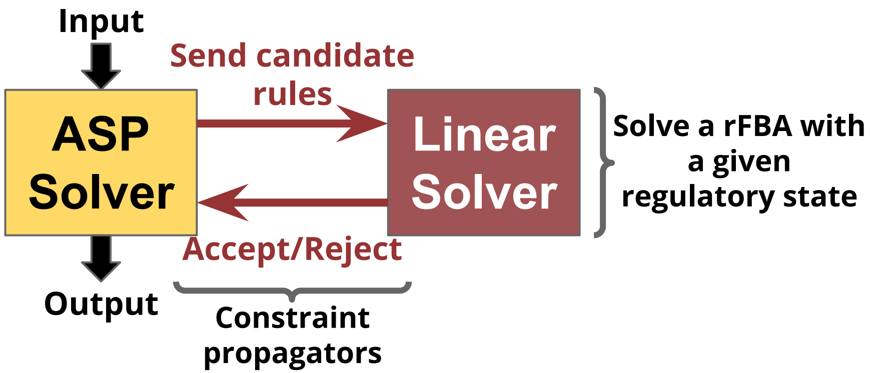
MERRIN's workflow



Problem solving

Rely on hybrid solving approach

(Answer Set Programming + Linear Programming)



Input

Send candidate rules

ASP Solver

Linear Solver

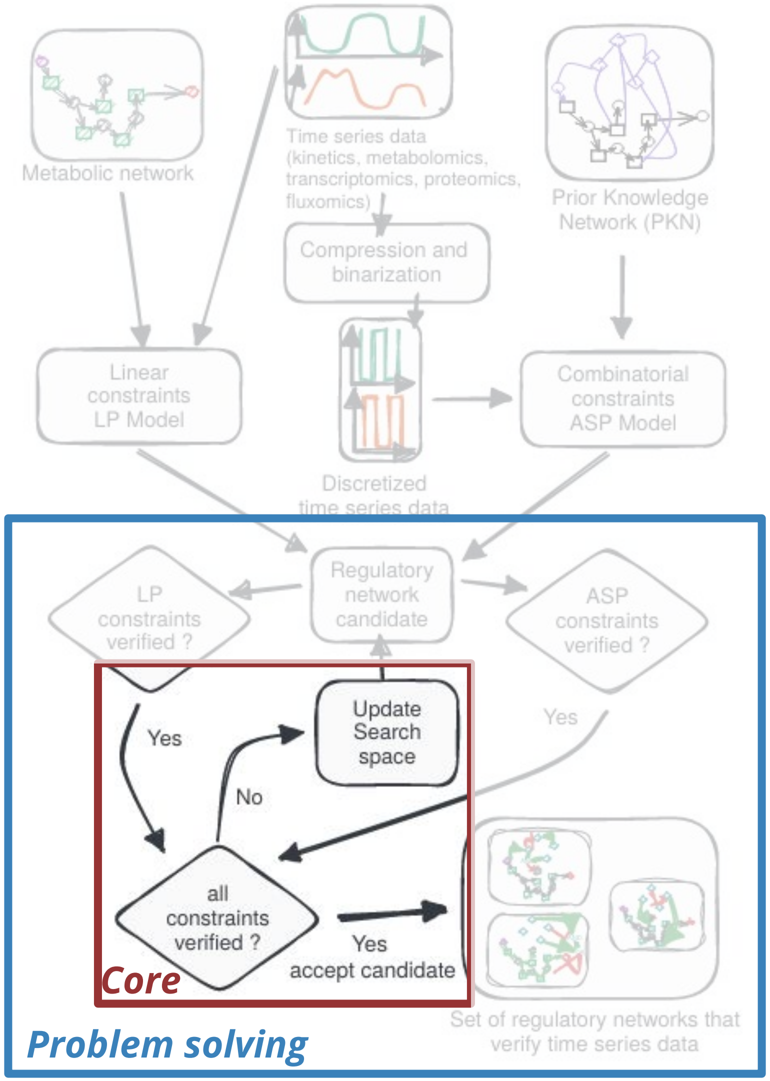
Solve a rFBA with a given regulatory state

Accept/Reject

Output

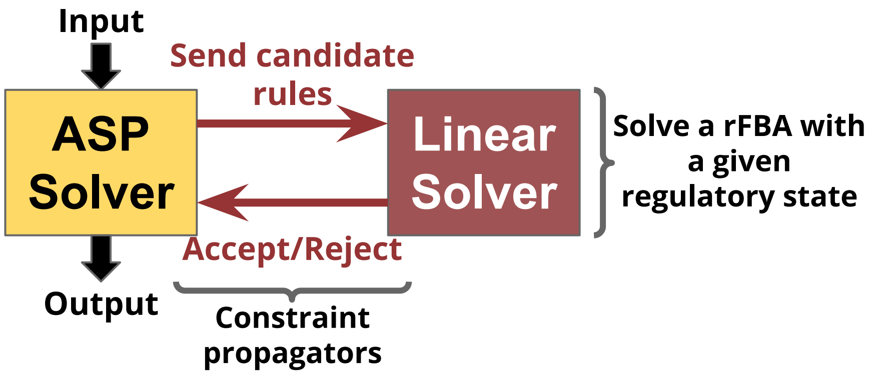
Constraint propagators

MERRIN’s workflow



Rely on hybrid solving approach

(Answer Set Programming + Linear Programming)

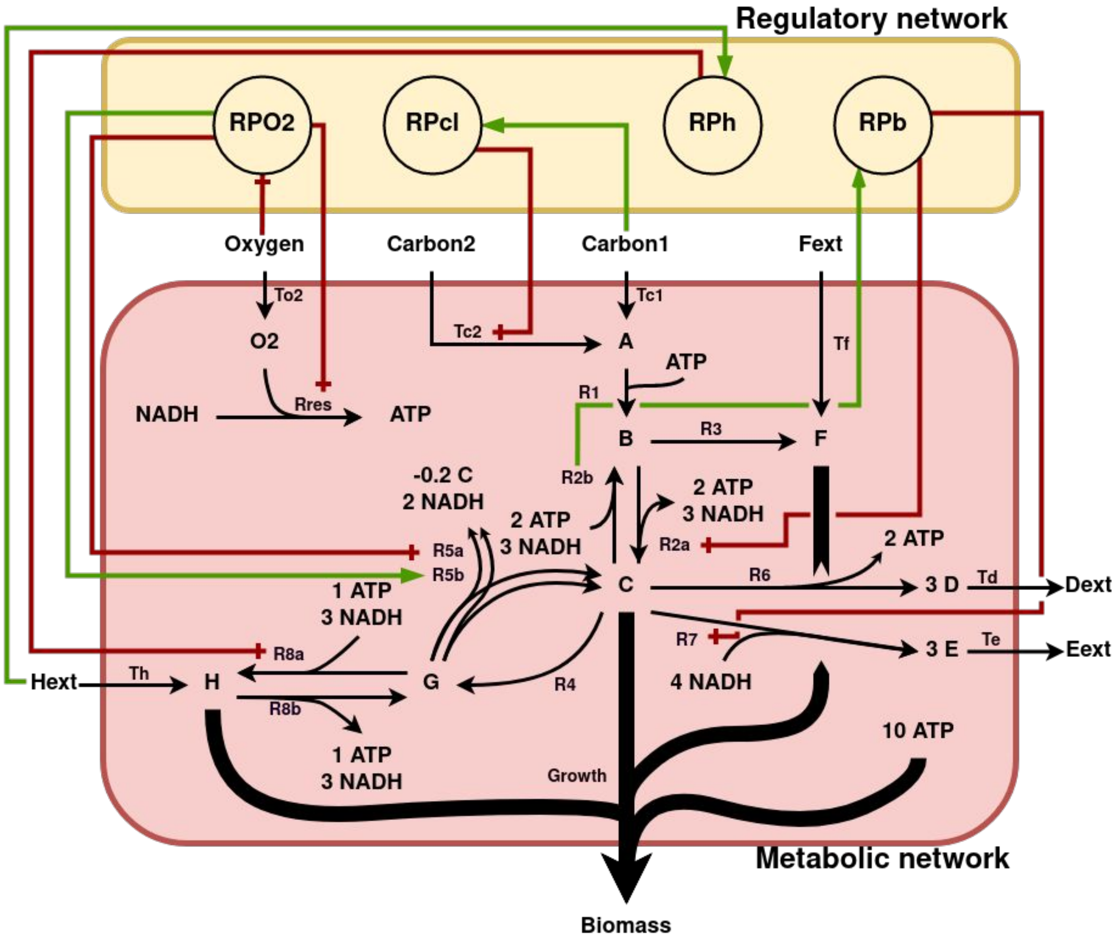


Accelerates the solving process by pre-filtering wrong solutions

→ ***Monotone properties on regulatory network structures***

→ ***Over-generalisation of counter-examples***

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

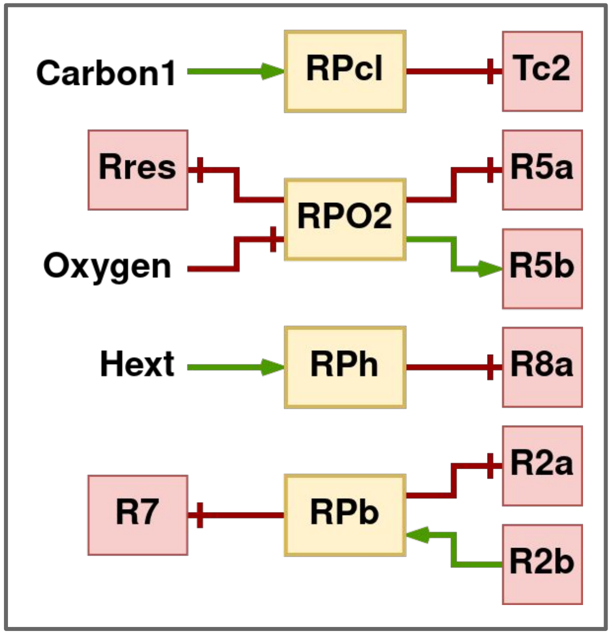


Toy model based on *E.coli*

20 reactions, 4 regulatory proteins,
11 regulations

Model complex behaviours

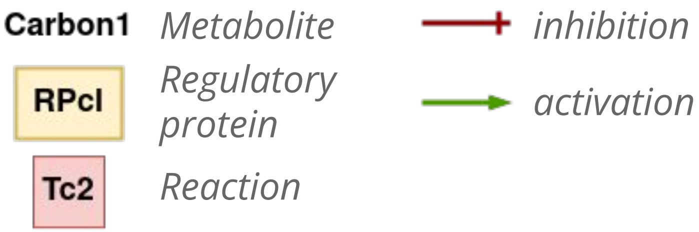
Diauxic shift, aerobic/anaerobic growth, etc.

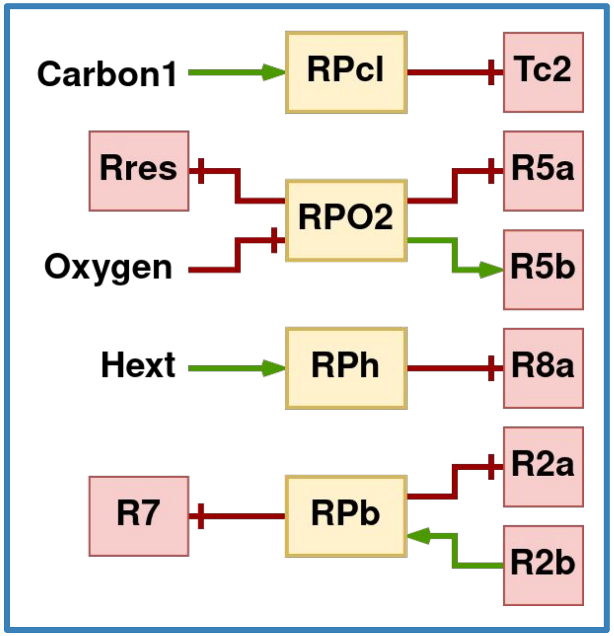


Influence graph

MERRIN inputs

Prior Knowledge Network





Gold standard influence graph

---

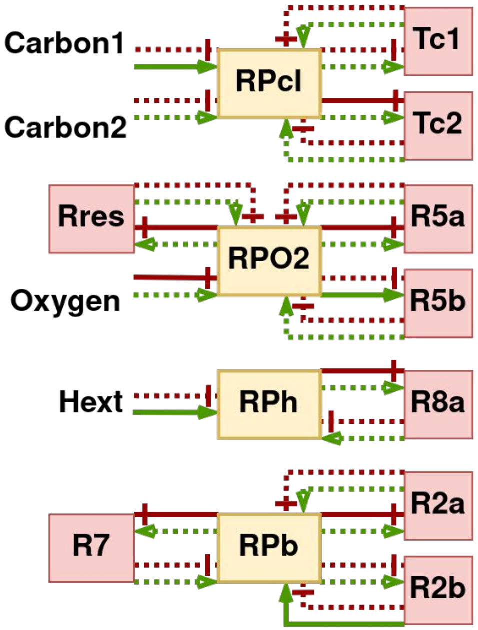
[1] M. W. Covert et al., **Journal of theoretical biology**, 2001

MERRIN inputs

Prior Knowledge Network

Prior Knowledge Network

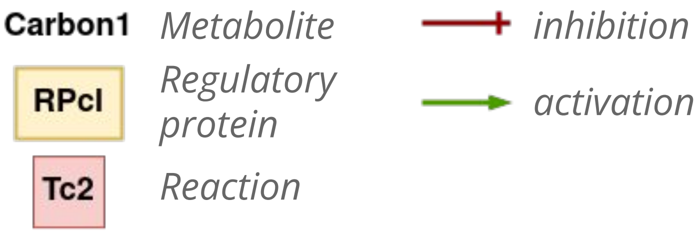
Set of admissible interactions
Superset of Gold Standard influence graph

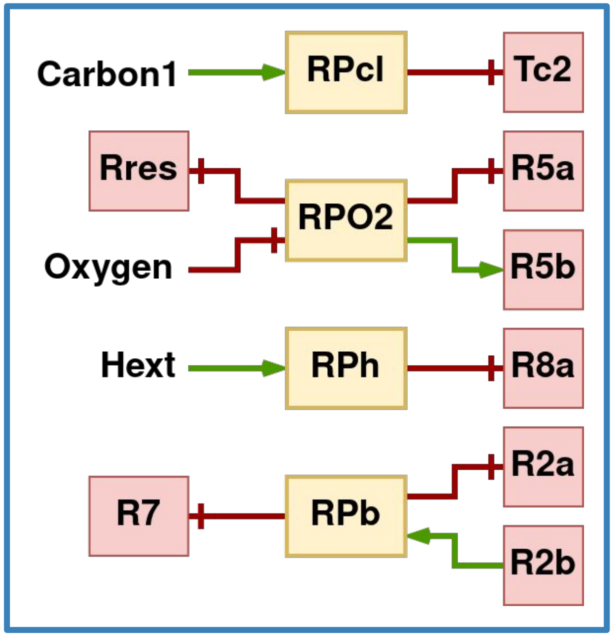


Remove interaction ***signs***
Remove interaction ***directions***

Add hypothetical regulations
(e.g. RPcl and Tc1)

~2.9x10^{12} potential
Boolean networks





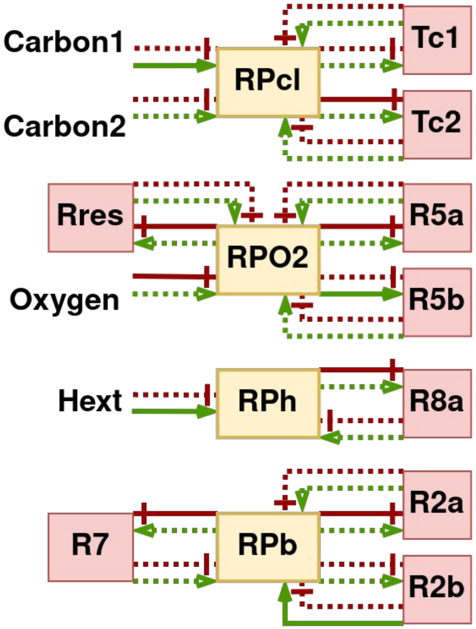
Gold standard influence graph

Instance generation

MERRIN inputs

Prior Knowledge Network

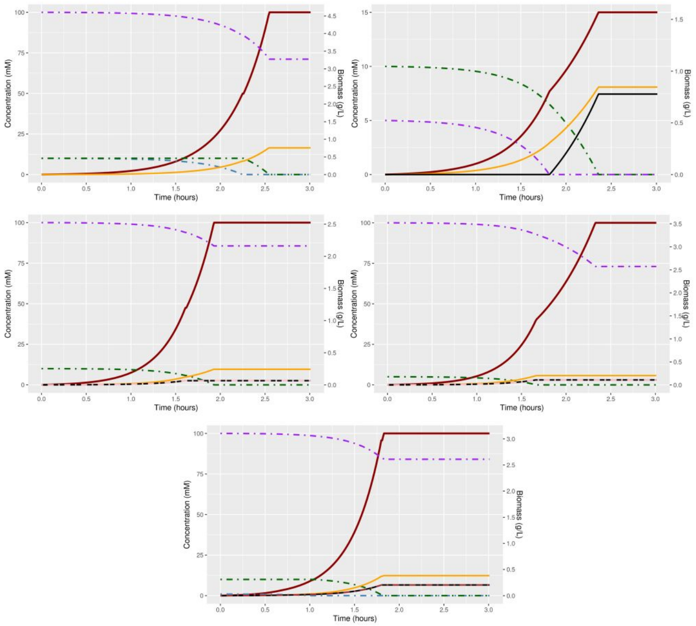
Add hypothetical regulations
Remove sign + direction of interactions



~2.9x10^{12} potential
Boolean networks

5 simulated experiments

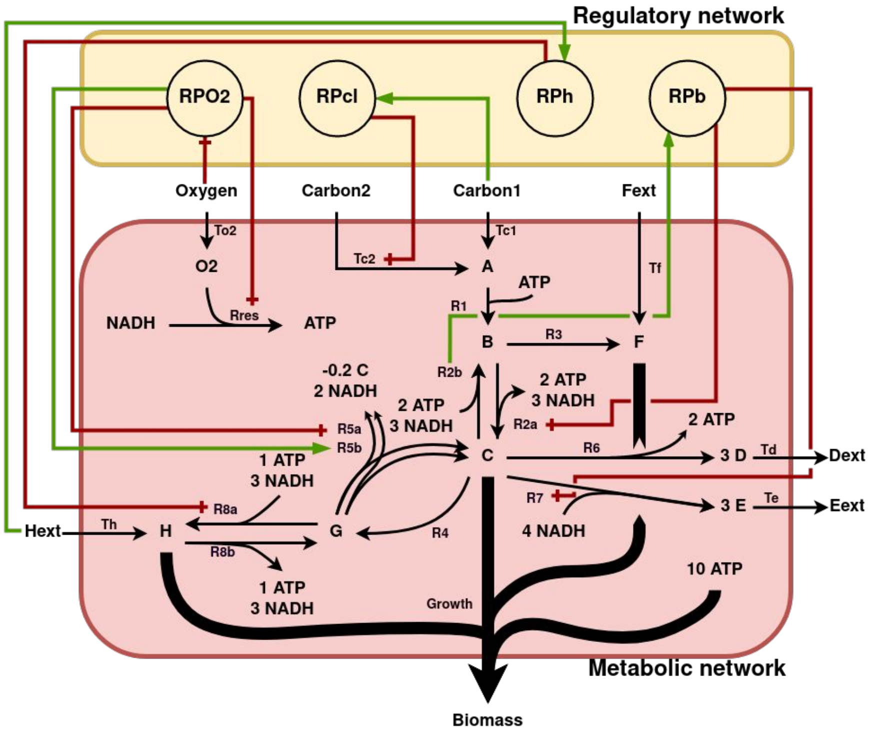
Provided by literature[1]
d-rFBA - made with FlexFlux



---

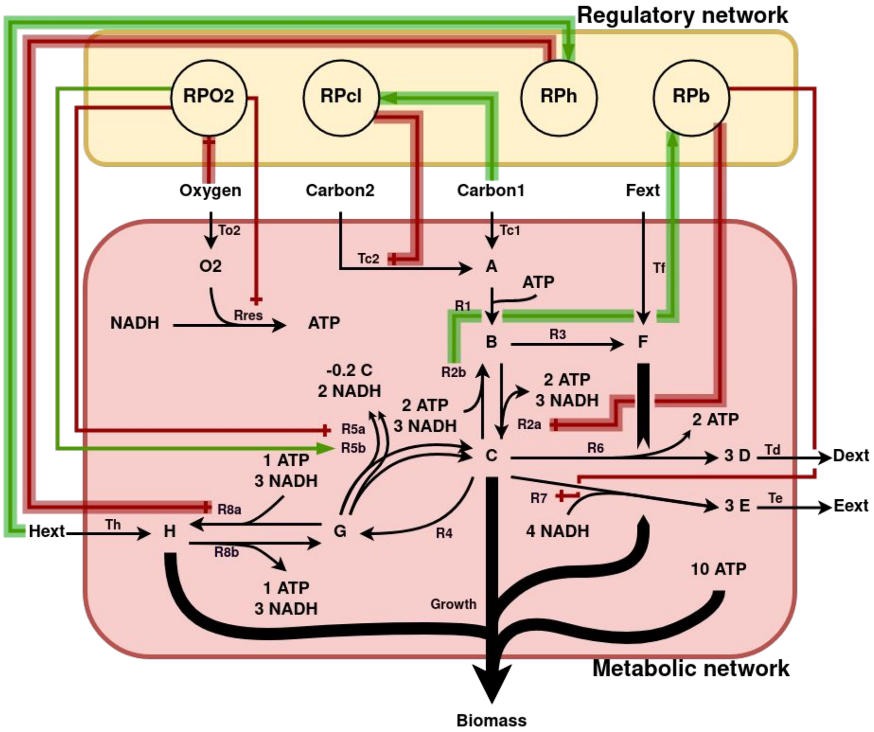
[1] M. W. Covert et al., **Journal of theoretical biology**, 2001

MERRIN validation



Gold standard is in the set of inferred models

MERRIN validation



Focus on smallest BNs

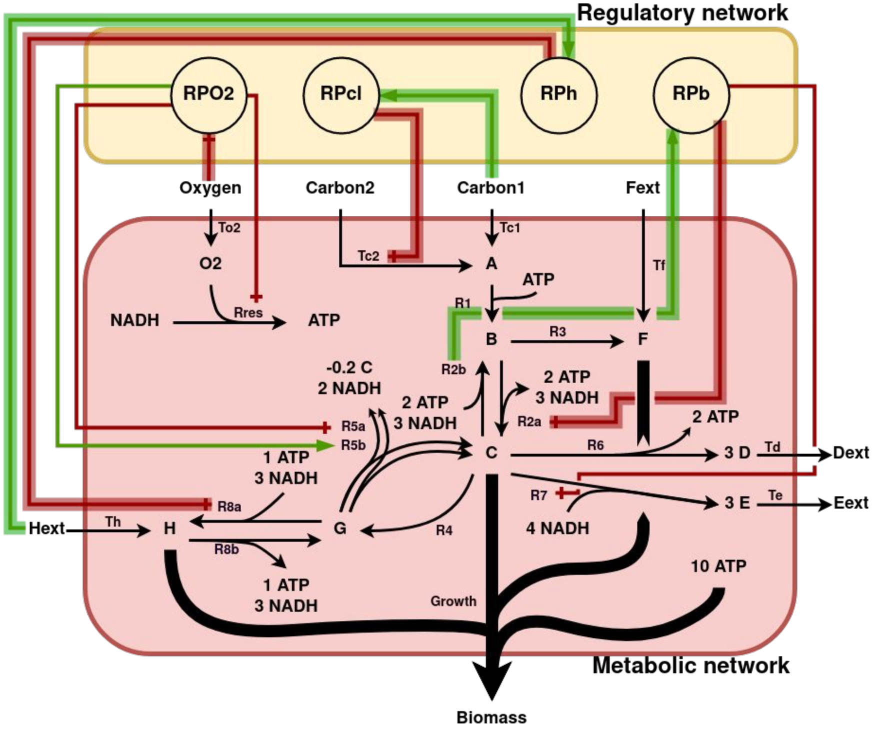
1 smallest BN

Computation time: ~7s

Subset of gold standard BN

Precision: 1.0 / Recall: 0.64

MERRIN validation



Focus on smallest BNs

1 smallest BN

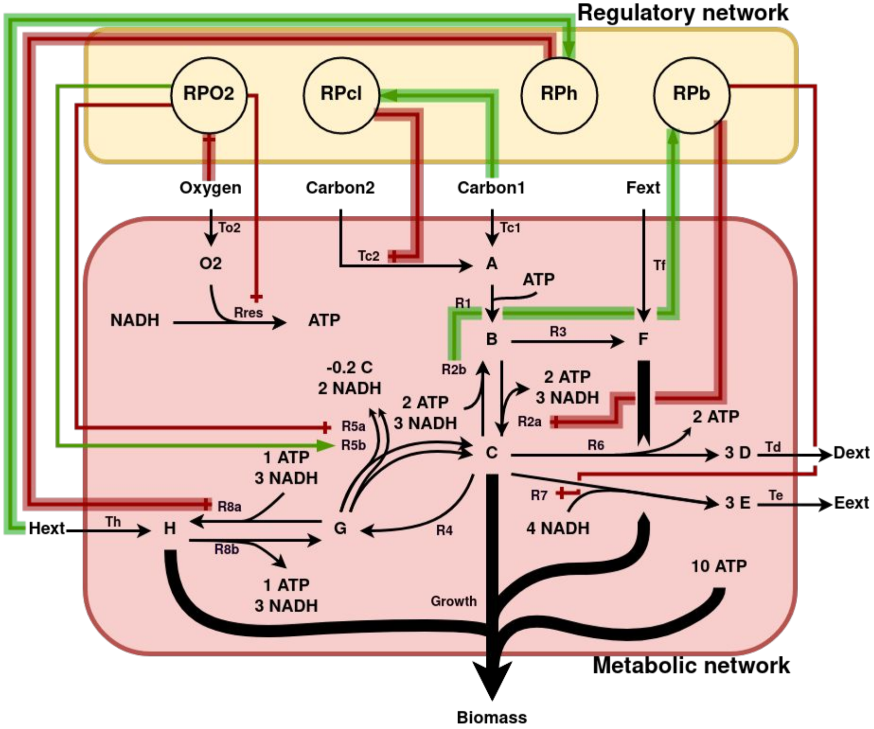
Computation time: ~7s

Subset of gold standard BN

Precision: 1.0 / Recall: 0.64

Did we fail inferring missing regulations ?

MERRIN validation



Focus on smallest BNs

1 smallest BN

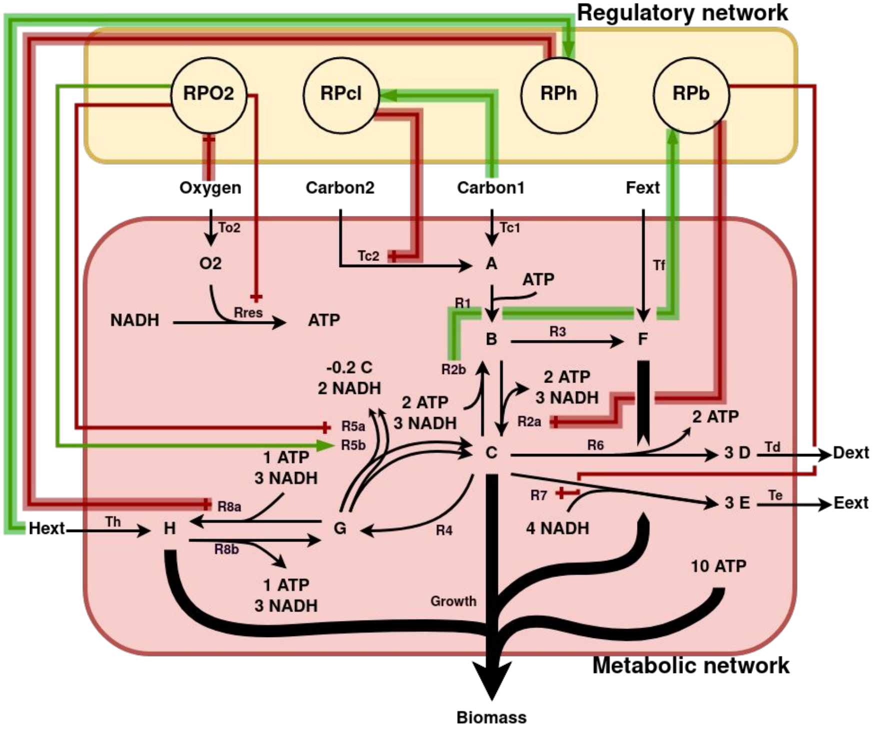
Computation time: ~7s

Subset of gold standard BN

Precision: 1.0 / Recall: 0.64

Reproduce exactly the input time series

MERRIN validation



Focus on smallest BNs

1 smallest BN

Computation time: ~7s

Subset of gold standard BN

Precision: 1.0 / Recall: 0.64

Reproduce exactly the input time series

Unrecovered regulations can be explained

Example Rres:

→ Enzyme decay is needed to capture it

→ Does not impact rFBA optimum

More parsimonious model than gold standard

Application to 3 classes of measures

Observations of the *regulatory and metabolic* system activities

→ ***Quantitative*** *and* ***qualitative*** *measurements*

→ ***Simulations*** *mimicking experimental data*

Application to 3 classes of measures

Observations of the *regulatory and metabolic* system activities

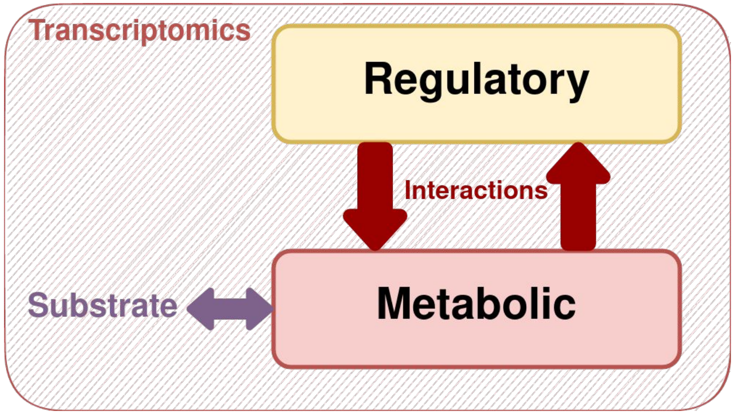
→ ***Quantitative*** *and* ***qualitative*** *measurements*

→ ***Simulations*** *mimicking experimental data*

3 data types supported by MERRIN:

➔ **Transcriptomics** (qualitative)

Analysis of the RNA transcripts



Application to 3 classes of measures

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

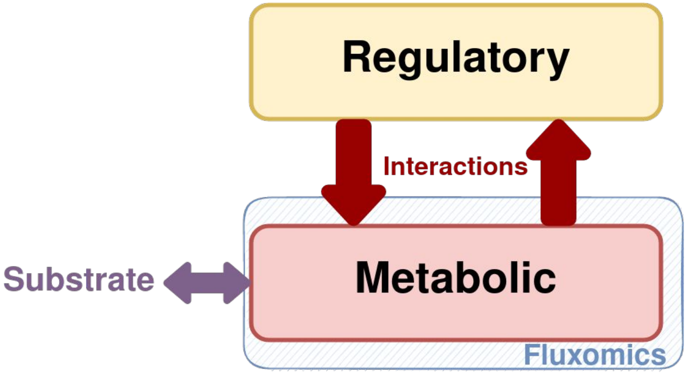
⟶ ***Quantitative*** *and* ***qualitative*** *measurements*

⟶ ***Simulations*** *mimicking experimental data*

3 data types supported by MERRIN:

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

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



Application to 3 classes of measures

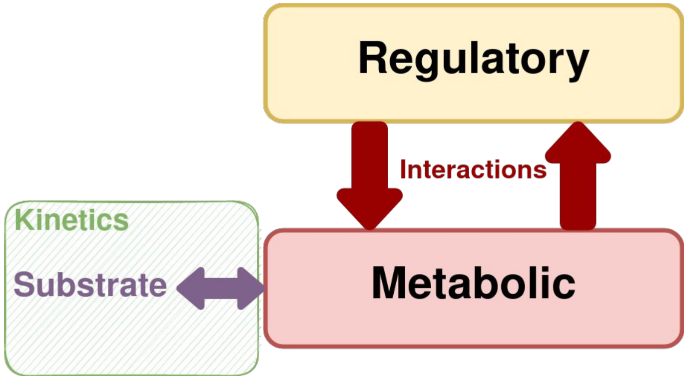
Observations of the *regulatory and metabolic* system activities

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

⟶ ***Simulations*** *mimicking experimental data*

3 data types supported by MERRIN:

- ➔ **Transcriptomics** (qualitative)
  Analysis of the RNA transcripts
- ➔ **Fluxomics** (quantitative)
  Rates of metabolic reactions
- ➔ **Kinetics** (quantitative)
  Substrate concentrations



Application to 3 classes of measures

Observations of the *regulatory and metabolic* system activities

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

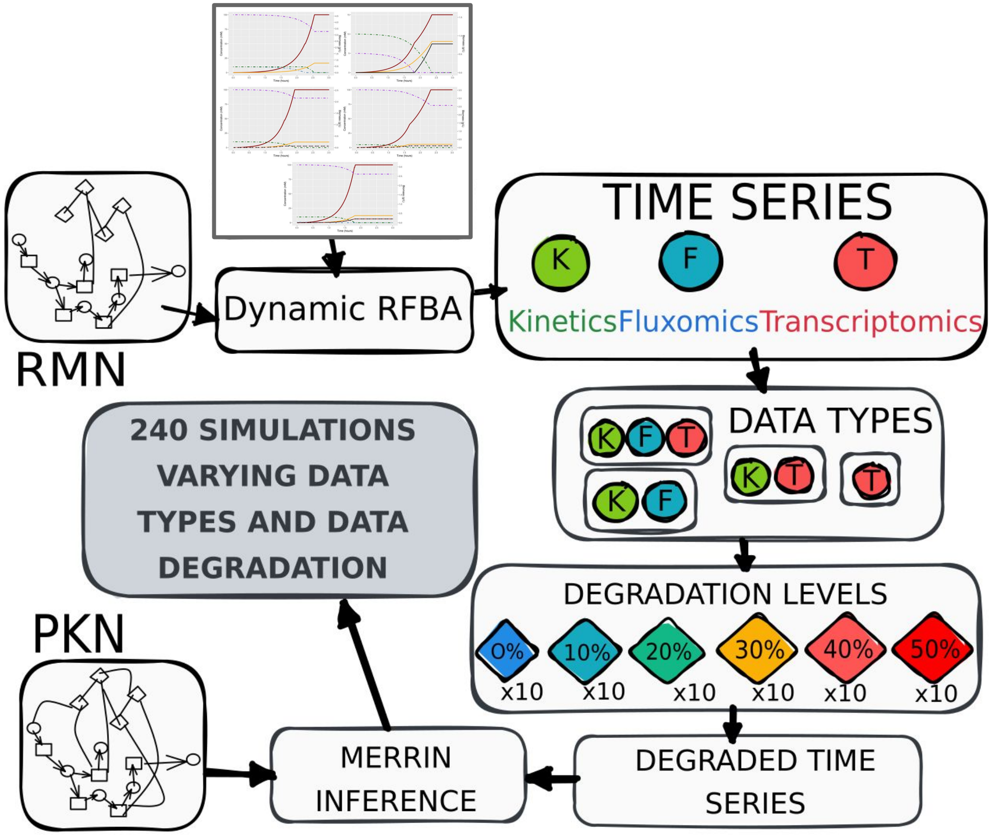
⟶ ***Simulations** mimicking experimental data*

3 data types supported by MERRIN:

- ➔ **Transcriptomics** (qualitative)
  Analysis of the RNA transcripts
- ➔ **Fluxomics** (quantitative)
  Rates of metabolic reactions
- ➔ **Kinetics** (quantitative)
  Substrate concentrations

MERRIN supports any combination of these 3 data types

Benchmark generation



5 simulations[1]:

➔ *Kinetics, fluxomics and transcriptomics*

4 data type combinations:

➔ *Kinetics, fluxomics, transcriptomics*
➔ *Kinetics, fluxomics*
➔ *Kinetics, transcriptomics*
➔ *Transcriptomics*

6 noise rates:

➔ *From **0% to 50%***
➔ *Missing observations and measures*

Benchmark of 240 instances

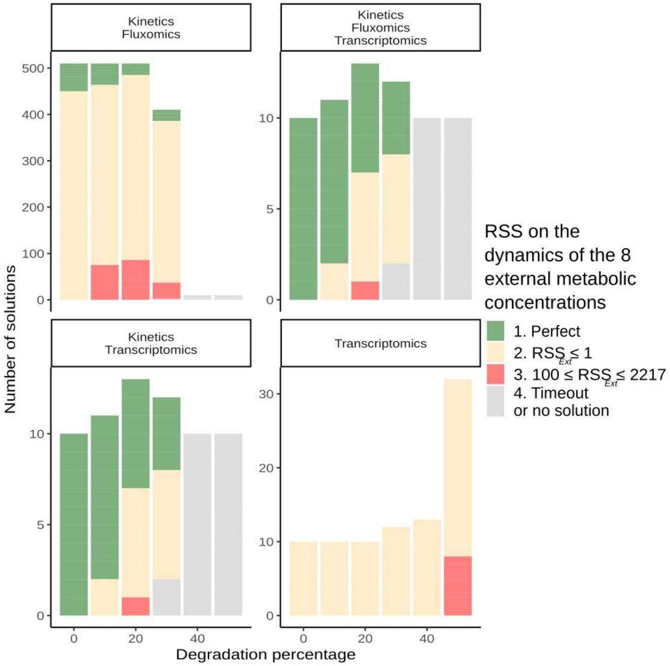
---

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

MERRIN robustness testing

Simulation reproducibility

Residual Sum of Squares – RSS



Average computation time ~25s

- **Robustness testing on**

  ⟶ ***noises:*** *0% – 50%*

  ⟶ ***data types:*** *KFT, KF, KT, T*

- **2 evaluation metrics**
  - ***Simulation reproducibility***

    Residual Sum of Square — RSS

    ➔ ***RSS < 1***
  - ***Regulatory network structure***

    Precision and Recall

    ➔ ***Precision = 1 / Recall = 0.64***

MERRIN optimal inputs

At least **transcriptomic** and **kinetic** data
With a **noise ≤ 20%**

Conclusion

Contribution

- **MERRIN[1]: inferring regulatory rules from time series**
  → ***Hybrid (ASP + LP) resolution***
- **Validation and robustness testing**
  → *Finding a **more parsimonious model** than the gold standard*
  → *Impact of **noise** and **data type** on inferred regulations*

Perspectives

- **Scalability to genome-scale metabolic networks**
  → ***rFBA scales well**, the **PKN size should be the issue***
  → *Next step: validation on **E.coli core medium-scale** model*
- **Study unrecovered regulations**
  → *Impact of experiments and data types on the inferred regulations*
  → *Extend simulation formalisms with **enzyme amounts** and **synthesis** — r-deFBA[2]*

---

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

[2] Z. Liu et al., **Journal of Theoretical Biology**, 2020