# Un exemple d'intégration de données omiques hétérogènes par/pour les modèles de processus cellulaires

Anne Goelzer

anne.goelzer@inra.fr

Applied Mathematics and Computer Science,
from Genomes to the Environment
INRA Jouy-en-Josas, FRANCE

INRA SCIENCE & IMPACT

MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES MaIAGE DU GÉNOME À L'ENVIRONNEMENT

17/10/2019

# Equipe BioSys de MaIAGE

## Objectifs de l'équipe BioSys

- Développer **des méthodes et des outils** permettant **de modéliser, d'analyser et de simuler** des systèmes biologiques, de l'échelle des processus subcellulaires à celle de l'individu (ou des communautés) dans leurs environnements
- Développer **des outils** permettant de **représenter, d'exploiter et d'intégrer** les connaissance, les informations et les nombreuses données, allant des nouvelles `omiques' au phénotypage haut-débit et en intégrant l'imagerie spatiale et temporelle de l'échelle subcellulaire à celle d'une population de cellules

⟹ Des modèles prédictifs pour la simulation fine de processus biologiques

### Modèle de la traduction des protéines (déterministe & stochastique)

Données utilisées (déterm.): physiologie, quantification RNAtot, rRNA, transcriptomique, protéomique quantitative

### Modèle d'allocation de ressources entre les processus cellulaires

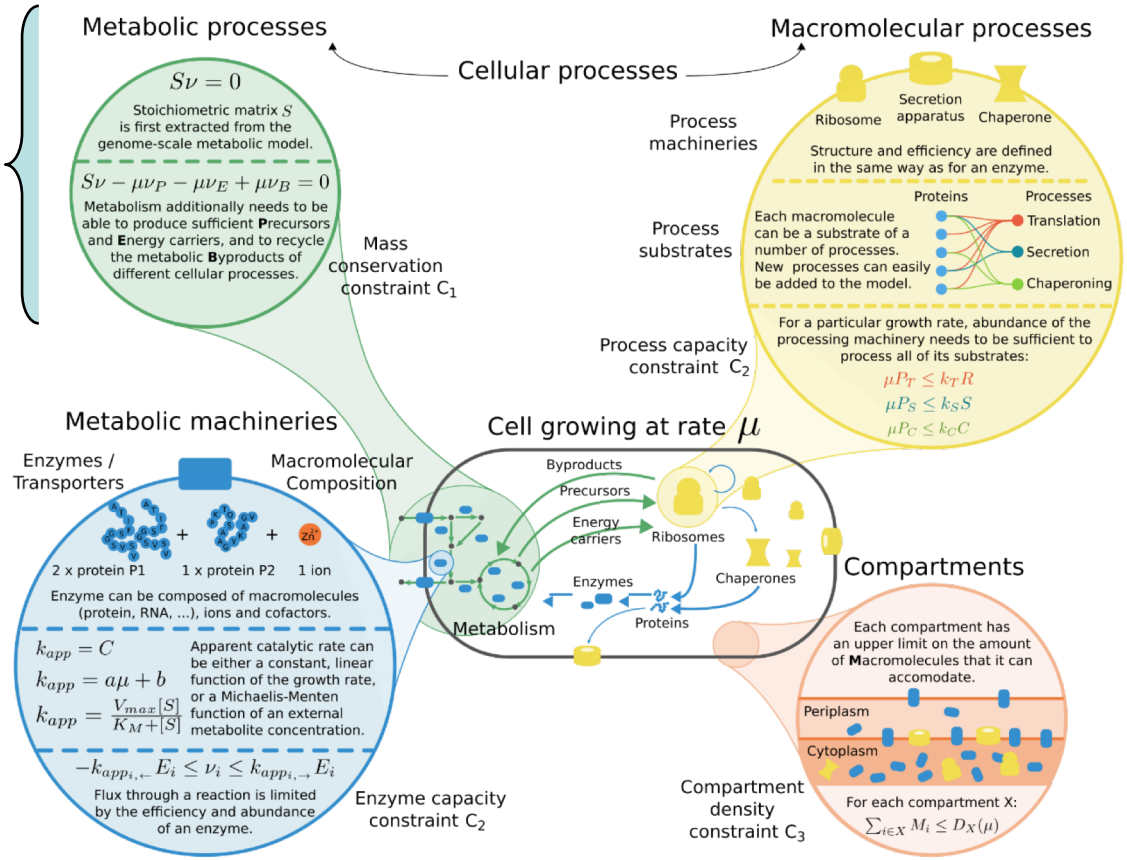
Données utilisées: physiologie, métabolomique, fluxomique, protéomique quantitative

### Modèle de la transcription des ARNm

Données utilisées: DNA, RNAseq, Tiling arrays, métabolomique

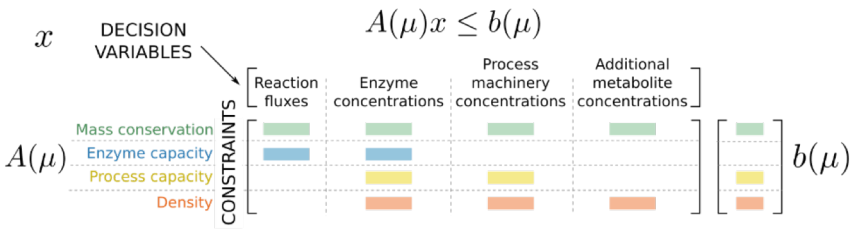
# Resource Balance Analysis in a nutshell

Relate molecular entities




Cellular description

Linear inequalities and equalities to be satisfied




For a particular value of $\mu$, a linear constraint feasibility problem is defined. If the problem is feasible for that particular growth rate, we increase $\mu$ until finding the highest value for which the problem is still feasible.

# Formalization into an optimization problem

## Resource Balance Analysis (RBA)

For fixed $P_G \geq 0$, $\mu \geq 0$,

Find $R \geq 0, C \geq 0, \nu^x \in \mathcal{R}^m$,

subject to

$|\nu_i| = k_{E_i} E_i$

Energy & precursors production

$(C_{1a})$ For all $i \in I_p$,
$-\sum_{j=1}^{m} S_{p_{ij}} \nu_j^x + \mu \left( \sum_{j=1}^{m} C_{M_{ij}}^{M_p} |\nu_j^x| + C_{R_i}^{M_p} R + C_{C_i}^{M_p} C + C_{G_i}^{M_p} P_G^{x,T} \right) - \nu_Y = 0$

$(C_{1b})$ For all $i \in I_c$,
$-\sum_{j=1}^{m} S_{c_{ij}} \nu_j^x + \mu \bar{X}_{c_i} = 0$

$(C_{1c})$ For all $i \in I_r$,
$\sum_{j=1}^{m} S_{r_{ij}} \nu_j^x + \mu \left( \sum_{j=1}^{m} C_{M_{ij}}^{M_r} |\nu_j^x| + C_{R_i}^{M_r} R + C_{C_i}^{M_r} C + C_{G_i}^{M_r} P_G^{x,T} \right) = 0$

$(C_{1d})$ For all $i \in I_i$,
$\sum_{j=1}^{m} S_{I_{ij}} \nu_j^x = 0$

$(C_{2a})$ $\mu(\sum_{j=1}^{m} C_{M_j}^{R} |\nu_j^x| + C_R^R R + C_C^R C + C_G^R P_G^{x,T}) - k_T R = 0$

Protein production

$(C_{2b})$ $\alpha_c \mu(\sum_{j=1}^{m} C_{M_j}^{R} |\nu_j^x| + C_R^R R + C_C^R C + C_G^R P_G^{x,T}) - k_C C = 0$

Protein folding

$(C_{3a})$ $\sum_{j=1}^{m} C_{M_j}^{D} |\nu_j^c| + C_R^D R + C_C^D C + C_G^D P_G^{c,T} - \bar{D}_c \leq 0$

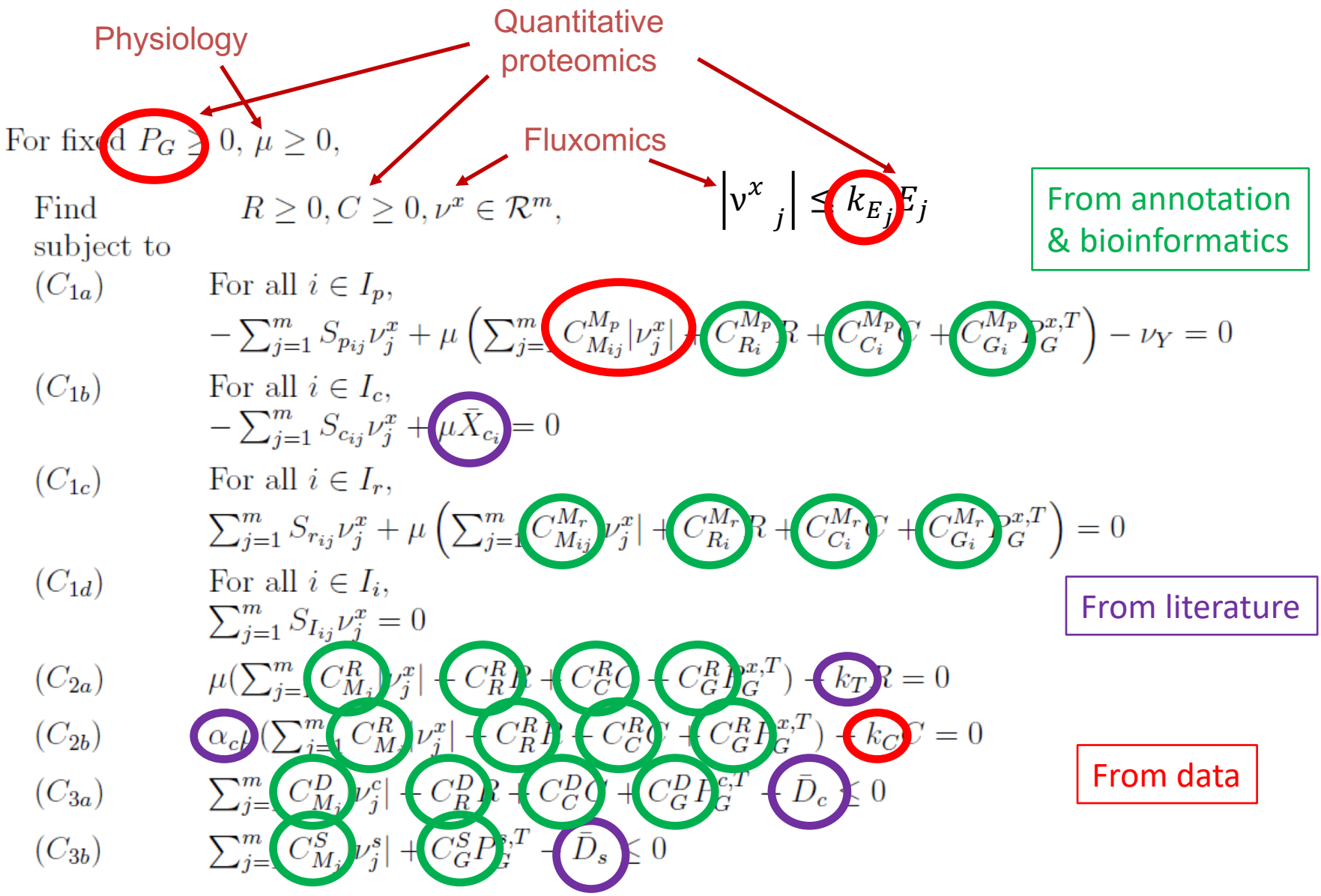
Cytosol occupancy

$(C_{3b})$ $\sum_{j=1}^{m} C_{M_j}^{S} |\nu_j^s| + C_G^S P_G^{s,T} - \bar{D}_s \leq 0$

Membrane occupancy


A. Goelzer , V. Fromion and G. Scorletti *Cell design in bacteria as a convex optimization problem*. 48th IEEE Conference on Decision and Control, China, 4517 -22. **2009**.
A. Goelzer , V. Fromion and G. Scorletti *Cell design in bacteria as a convex optimization problem*. Automatica,47(6):1210-1218. **2011**.

# 752 parameters to be estimated



# Data dedicated to RBA validation (5 conditions)

| Data type | Group |
|---|---|
| **Physiological data: length, width, volume Growth rate, Cell dry weight** | **Inra - Jouy/Grignon Greifswald** |
| **DNA concentration (5 conditions)** | **Inra – Grignon** |
| **Transcriptomic data** | **Inra – Grignon** |
| **Protein quantification (absolute) [quantification some key membrane proteins]** | **Greifswald** |
| **Concentration of ribosomes** | **Greifswald** |
| **Amount of total mRNAs** | **Greifswald/Inra - Grignon** |
| **mRNA half life** | **Greifswald** |
| **Polymerase activity (chip on chip)** | **Inra – Jouy** |
| **Metabolic data - external/internal** | **ETH – Zurich** |

J. Muntel , et al. *Comprehensive absolute quantification of the cytosolic proteome of Bacillus subtilis by multiplexed LC/MS (LC/MSE).* Molecular and Cellular Proteomics, 13(4):1008-1019. **2014** .

# La difficulté de croiser des données quantitatives absolues et hétérogènes

## Trois groupes impliqués dans la génération des données expérimentales

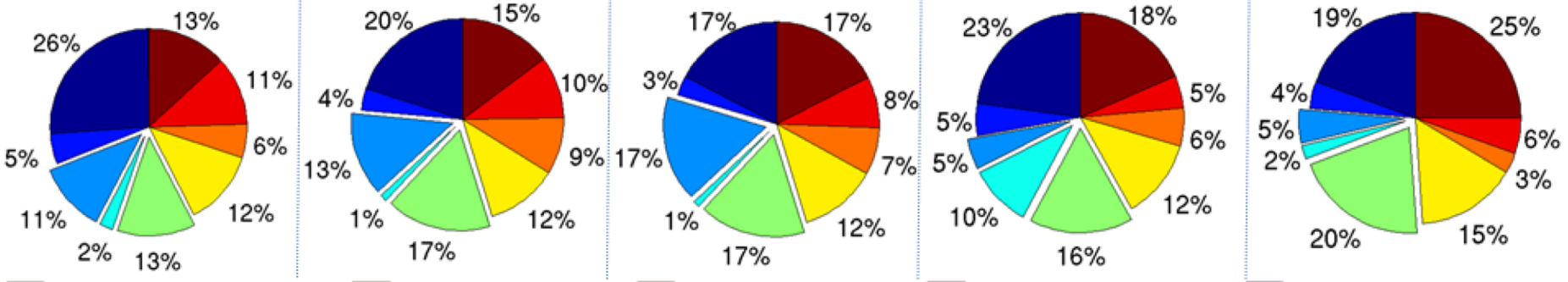
- Les conditions de culture doivent être « strictement » identiques
  - → Par ex. même taux de croissance pour les cultures dans les différents labos
- Les informations communes doivent être identiques
  - → Par ex. une DO600 de 0,4 doit être comparable entre les différents labos

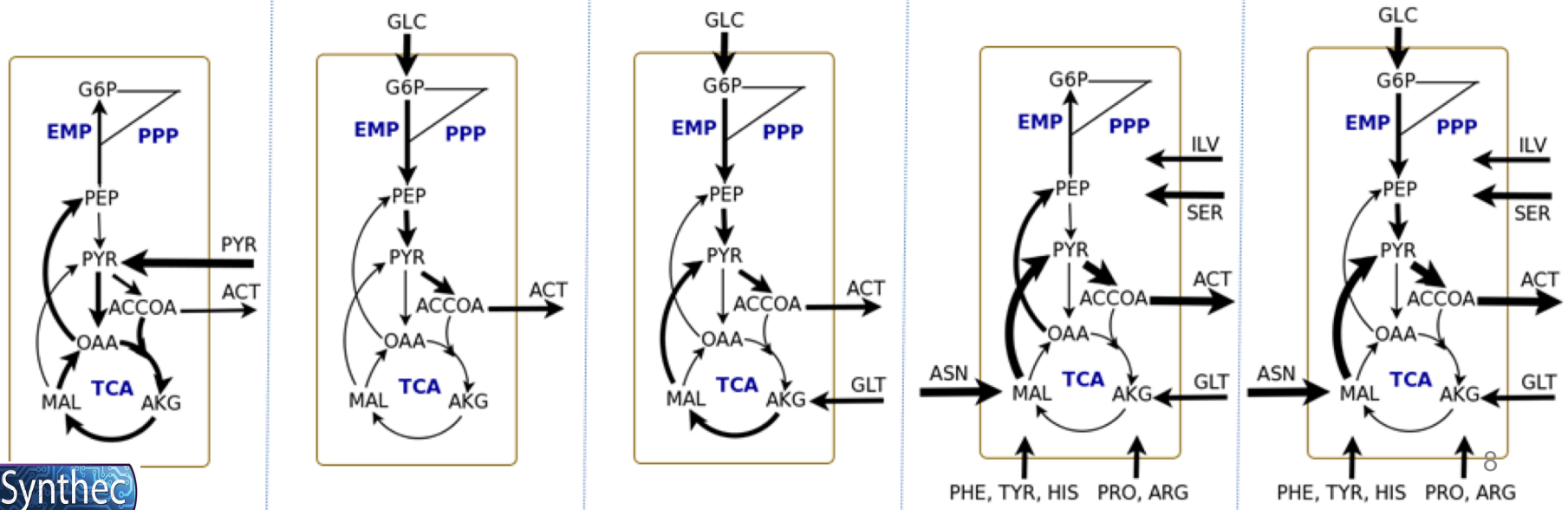
⇒ Nécessité d'établir des SOP précises, et des stratégies pour contrôler que les SOP sont bien respectées

## Du point de vue de la modélisation:

- Le cauchemar des conversions d'unités entre données expérimentales et entités modélisées
- Des informations essentielles pour les modèles quantitatifs mais souvent absentes
  - → Souvent des données « macro » à l'échelle de la cellule:
    Par ex: conversion DO/poids sec, DO/nombre de cellules, volume, quantité totale de protéines, etc.
- Evaluer la cohérence du jeu de données par rapport à l'existant

| Minimal medium Pyruvate (PYR) | Minimal medium Glucose, Citrate (S) | Medium S + Glutamate (TS) | Complex medium 18 amino acids (CH) | Medium CH + Glucose (CHG) |
|---|---|---|---|---|
| μ: 0,3 $h^{-1}$ | μ: 0,6 $h^{-1}$ | μ:0,9 $h^{-1}$ | μ: 1,1 $h^{-1}$ | μ:1,5 $h^{-1}$ |
| 797 proteins | 1127 proteins | 1158 proteins | 984 proteins | 1103 proteins |



# Before calibration ⟶ Analysis of the datasets

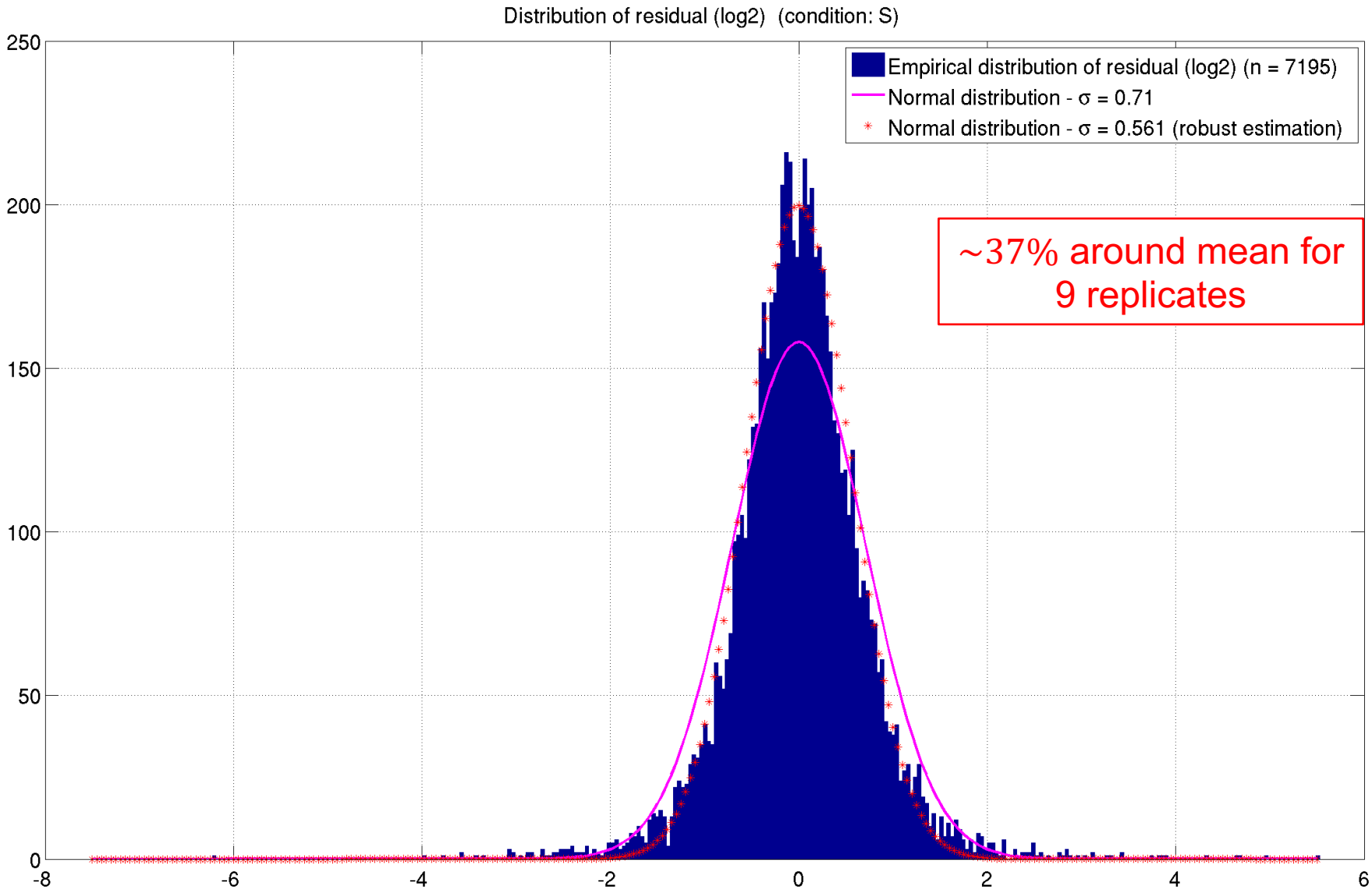
## Metabolomics/Fluxomics/Physiology

- HPLC, LC/MS, GC/MS, 2 replicates per condition
- Computation of the uptake rates & excretion rates of nutrients + confidence intervals
- Comparison with known fluxes in similar medium and solving discrepancies if any… (pb of conversion OD/cell dry weight)

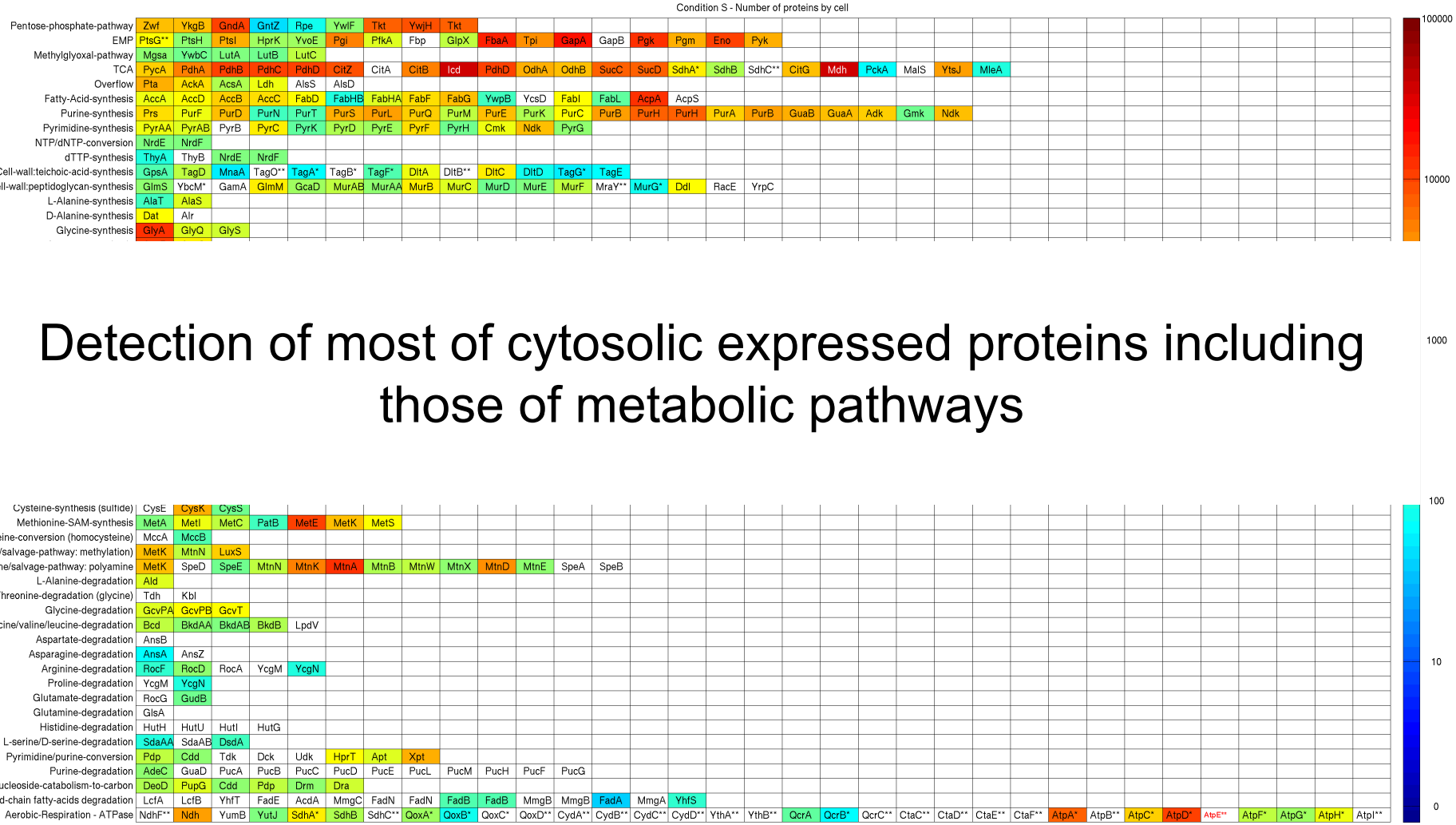
## Quantitative proteomics (1st dataset on *B. subtilis*)

- 9 replicates (3 biological x 3 technical replicates), LC-MS/MS
- Per condition:
  - ✓ Analysis of the variance with respect to the mean (variance stabilization by log transformation), to the number of replicates
  - ✓ Computation of the 95% confidence interval
  - ✓ Coverage of the (active) metabolic pathways, and other macro-molecular processes
  - ✓ Correlation analysis between known operons and regulons
- Between conditions:
  - ✓ Differential analysis between two conditions including a volume correction by a multi-test Benjamin-Hochberg procedure with a threshold of 1%

# Analysis of protein abundance (9 replicates)

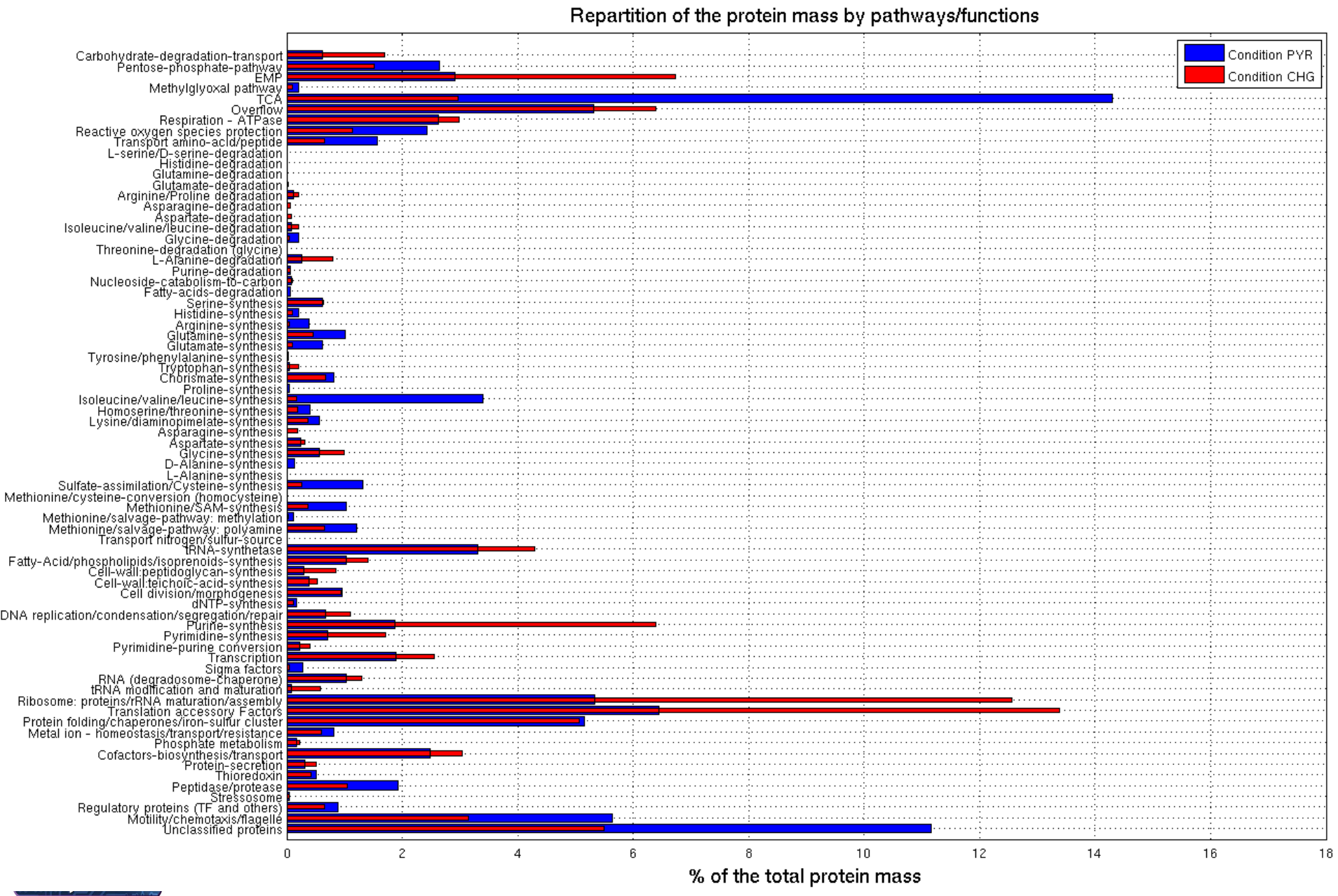


# Number of protein by cell among the core metabolism (S)

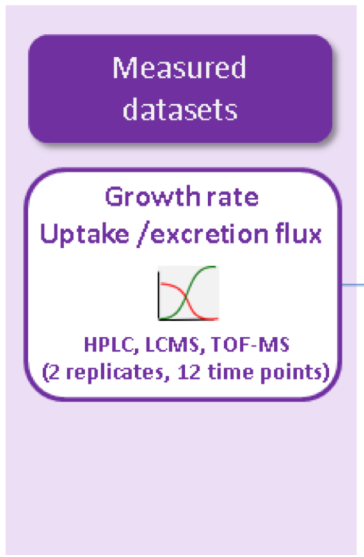


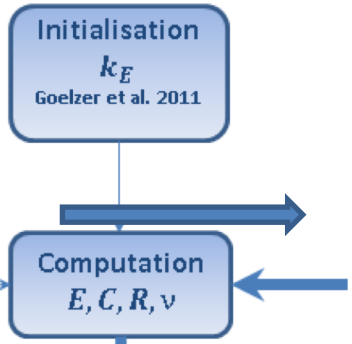

# Detection of most of cytosolic expressed proteins including those of metabolic pathways

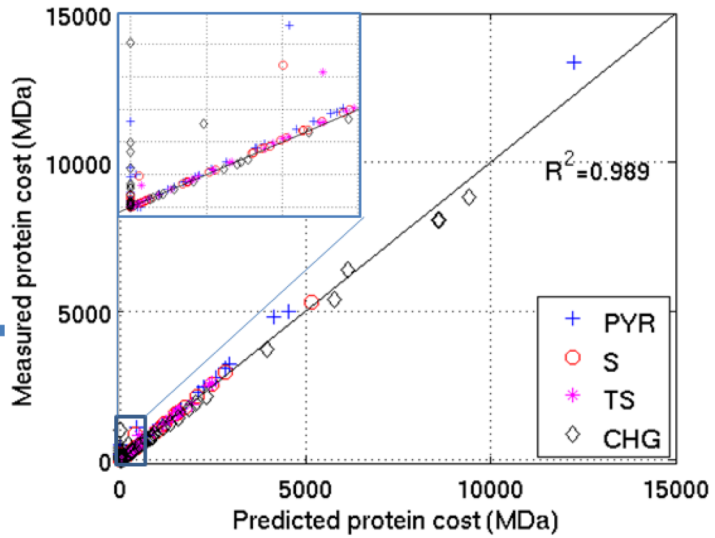
# Protein allocation among all the cell processes (PYR vs. CHG)

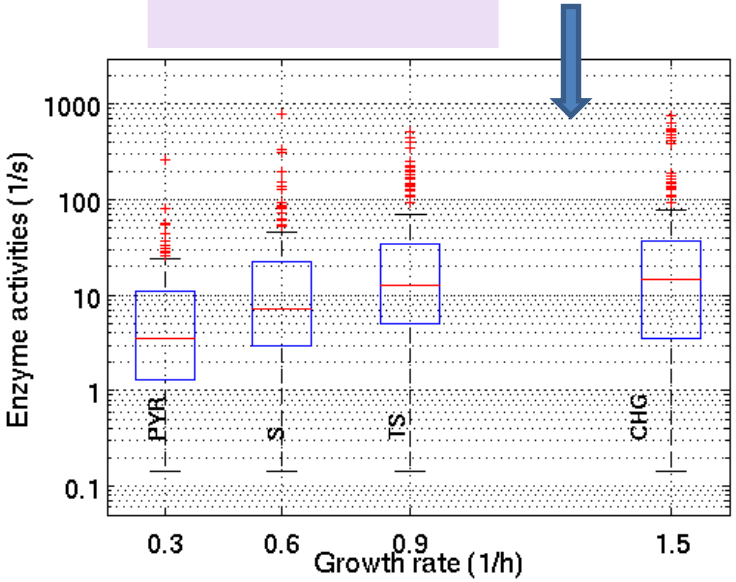


# A dedicated procedure for parameter estimation from fluxome and absolute protein quantification

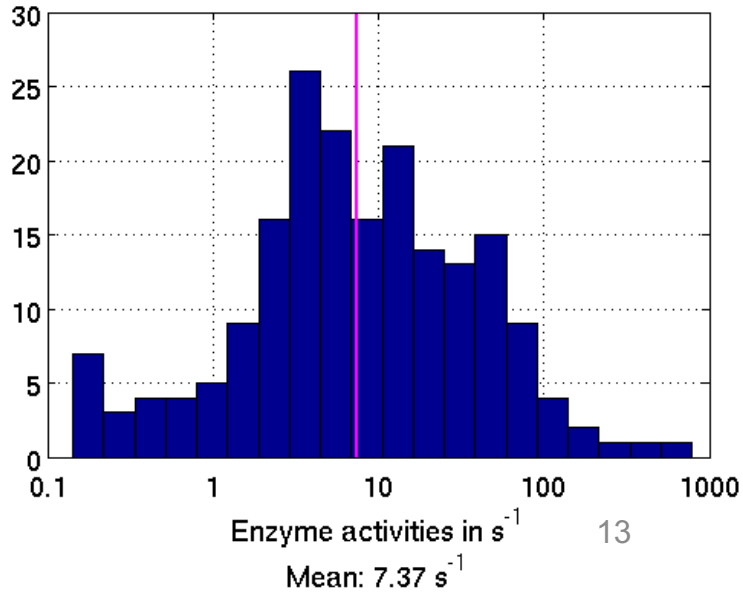


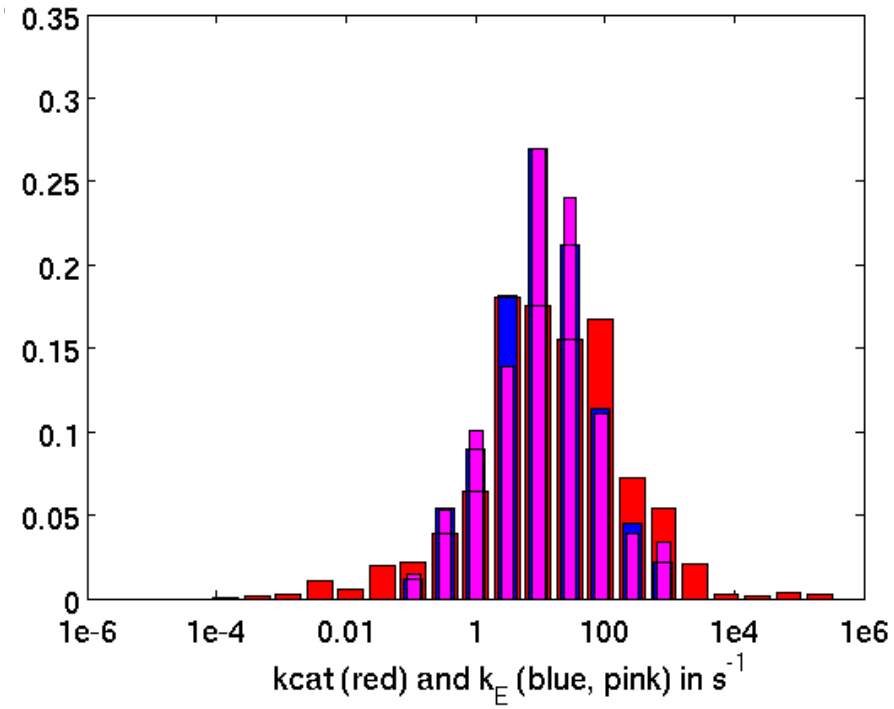



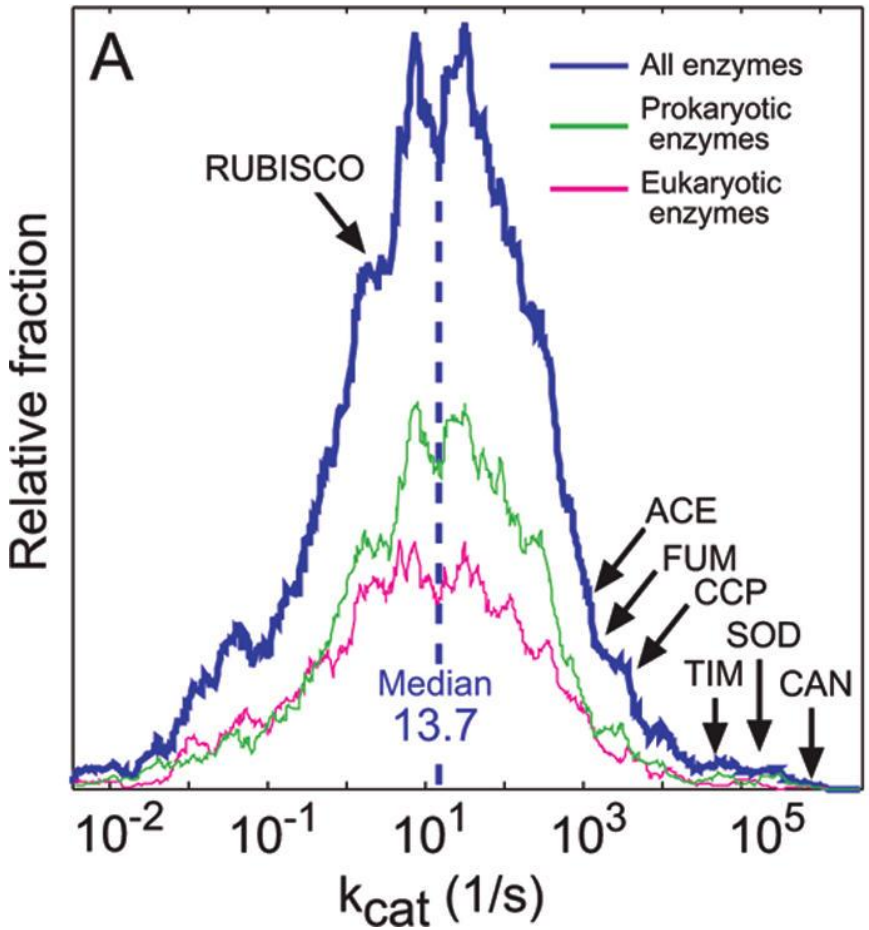




wo step iterative procedu



# Identification of apparent catalytic rate of ≈600 enzymes for each growth condition (Consistency with the expected distribution)



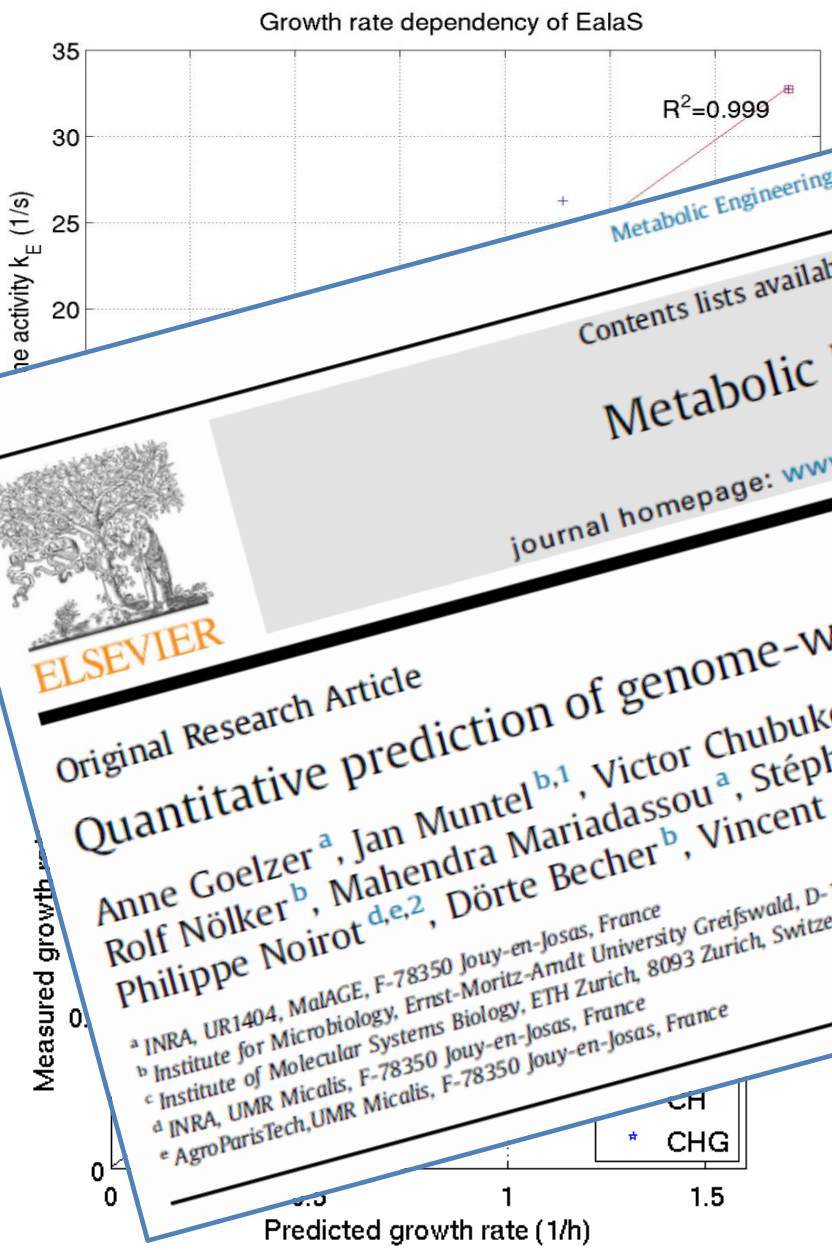




A. Bar-Even, et al. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters, Biochemistry, 2011, 50 (21), pp. 4402–4410



BaSynthec

# Evolution of the apparent catalytic rates



# Also possible for another organism

Use of RBApy workflow to build a RBA model for *Escherichia coli* [1]

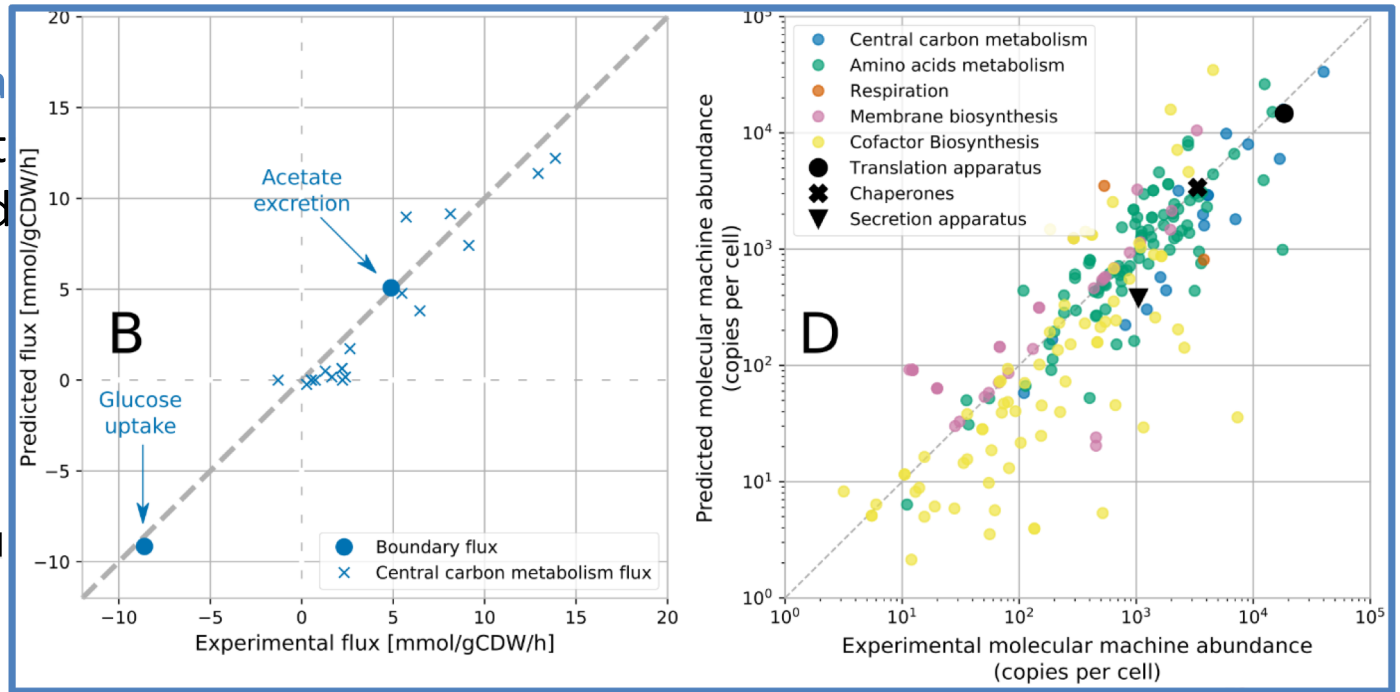
⟶ By applying the same rationale for data analysis/ parameter estimation

Source of information:

- ❑ The iJO1366 metabolic model [2]
- ❑ Quantitative proteomics [3] and fluxomics [4]

Estimation of pa

- ❑ Apparent cat
  minimal med
- ❑ Total protein
  [3]




Institut de Modélisation des Systèmes Vivants

# What we need & what is missing

## In data acquisition/treatments

For some data (from our experience in microbiology), lack of a clear consensus on

- Data units especially between different scientific communities
- Identifiers (e.g. metabolomics, fluxomics) ?

And also

- Missing information on the mathematical treatment between the raw and the treated data
- Missing physiological data
- Alternative splicing: multi-mRNAs, multi-proteins?

# What we need & what is missing

## In the cell description

### The active molecular machine at the basis of the cell description

- Badly and rarely described by itself: ex. Uniprot, TAIR, etc.
- Some promising attempts for pathway-centered repository (Reactome, ChloroKB, etc) … But still ambiguous
  - ✓ e.g. The mitochondrial protein IDH3A of *S. cerevisiae* in Reactome is linked to the protein P28241 (that still contains the signal peptide)

→ A problem of cell description and in fine of cross-references between repositories.

→ Which lead to a problem of data identification during acquisition

→ The problem of cell description is even more difficult for multicellular organism
  - ✓ In genome-scale metabolic models of plants, the gene encodes the reaction

### New omics technologies allow the large-scale identification of new types of « data » (a few examples):

- Molecular machine efficiencies (including apparent catalytic rates )
- Half-lifes of mRNAs, proteins
- Gene-specific translation efficiencies,
- etc.

→ How to store the information ?

# Acknowledgments


INRA-MaIAGE

V. Fromion

M. Mariadassou, P. Nicolas, L. Tournier

S. Fischer, M. Dinh, W. Liebermeister

Ecole Centrale de Lyon

G. Scorletti

INRA-Micalis

M. Jules, S. Aymerich (Grignon)

P. Noirot, E. Prestel, E. Dervyn (Jouy)

A.Bulovic, E. Klipp

**The BaSynthec partners**

J. Muntel R. Nölker, D. Becher M. Hecker and U. Mader (Greifswald),

V. Chubukov and U. Sauer (ETHZ)