The Robot Scientist Genesis: Abduction for Metabolic Modelling

AAAI Spring Symposium
Computational Approaches to Scientific Discovery
San Francisco, 28 March 2023

Alexander H. Gower, Konstantin Korovin, Daniel Brunnsåker, Filip Kronstrom, Gabriel K. Reder, Ievgeniia A. Tiukova, Ronald S. Reiserer, John Wikswo, Ross D. King

- The original discovery problem
- How we formulated the problem in computational terms
- What data and knowledge we provided to our system
- How we represented the system's inputs and outputs
- The space of candidate models that the system searched
- What criteria it used to evaluate candidate models
- How we interpreted results that the system generated

Metabolic modelling *Saccharomyces cerevisiae*

- Yeast is the model eukaryote
- Exist tools to conduct experiments (e.g. CRISPR/Cas9)
- Cell factory

Photo by Jeff Siepman on Unsplash

Photo by Jude Infantini on Unsplash

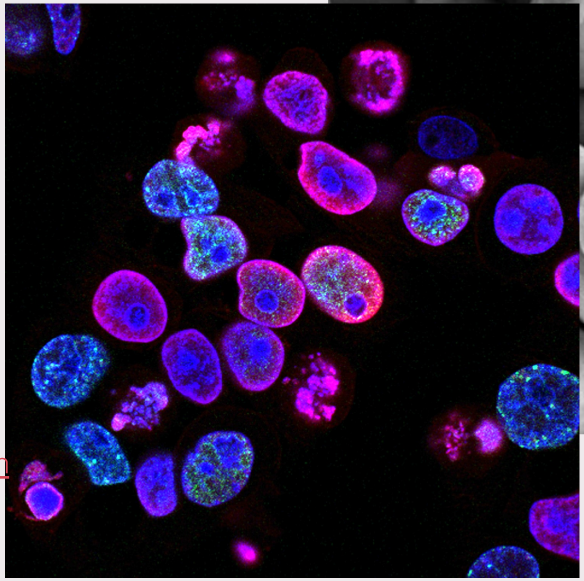
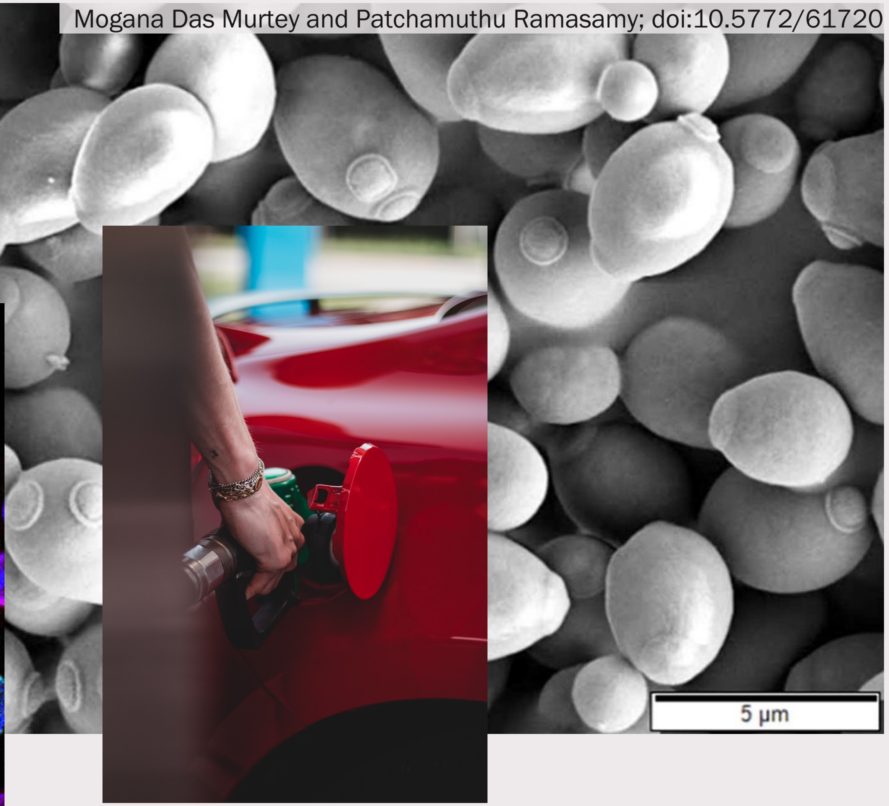


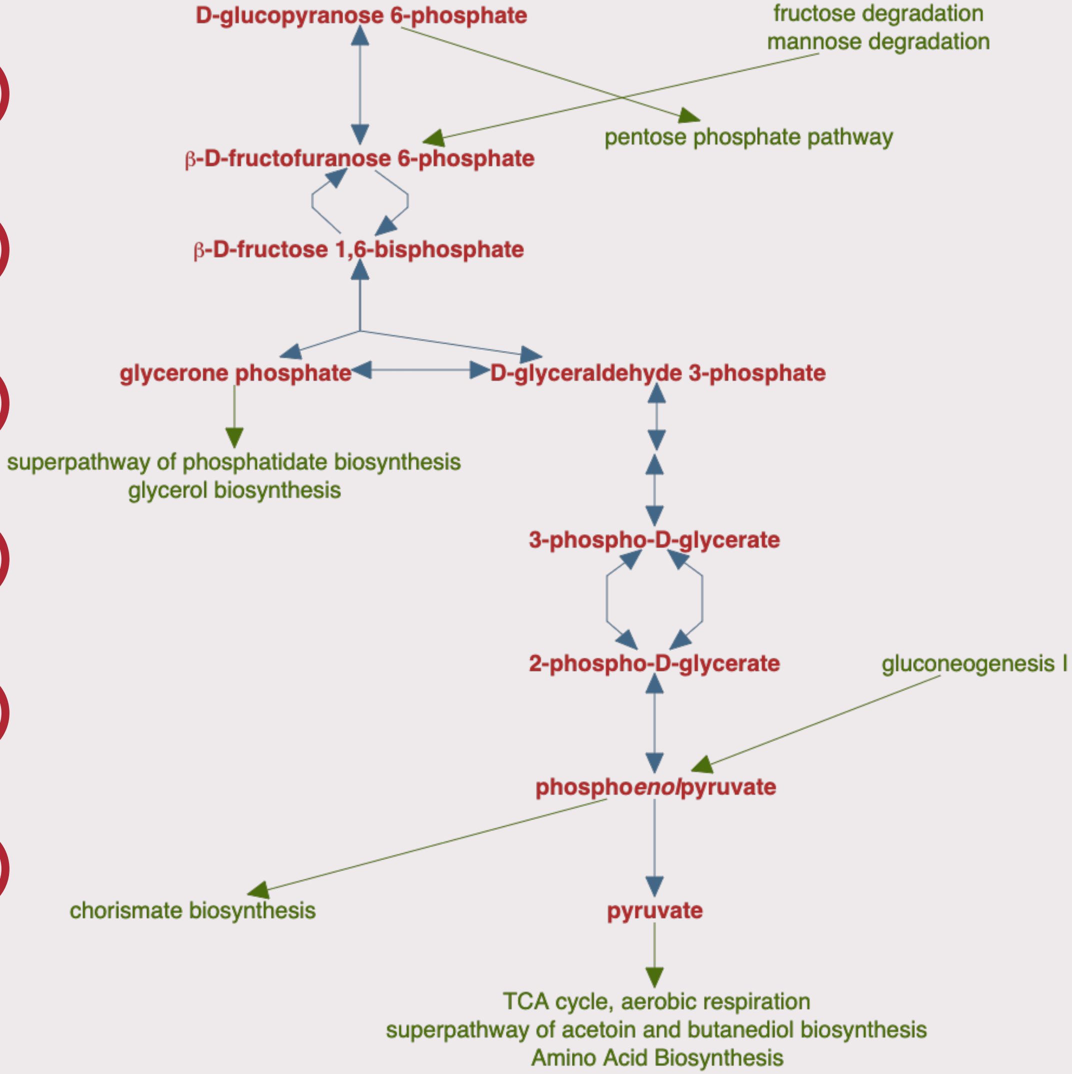
Photo by National Cancer Institute on Unsplash

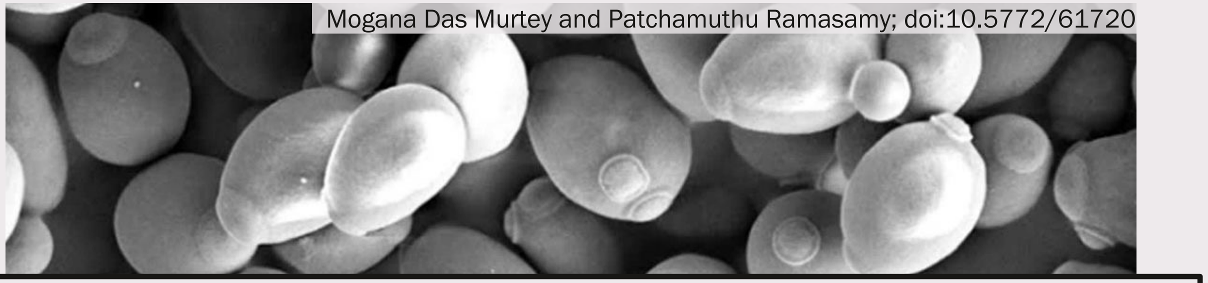


Mogana Das Murtey and Patchamuthu Ramasamy; doi:10.5772/61720

Photo by Wassim Chouak on Unsplash

Metabolic modelling *Saccharomyces cerevisiae*


D-glucopyranose 6-phosphate
fructose degradation
mannose degradation
pentose phosphate pathway
β-D-fructofuranose 6-phosphate
β-D-fructose 1,6-bisphosphate
glycerone phosphate
D-glyceraldehyde 3-phosphate
superpathway of phosphatidate biosynthesis
glycerol biosynthesis
3-phospho-D-glycerate
2-phospho-D-glycerate
gluconeogenesis I
phospho*enol*pyruvate
chorismate biosynthesis
pyruvate
TCA cycle, aerobic respiration
superpathway of acetoin and butanediol biosynthesis
Amino Acid Biosynthesis


Mogana Das Murtey and Patchamuthu Ramasamy; doi:10.5772/61720

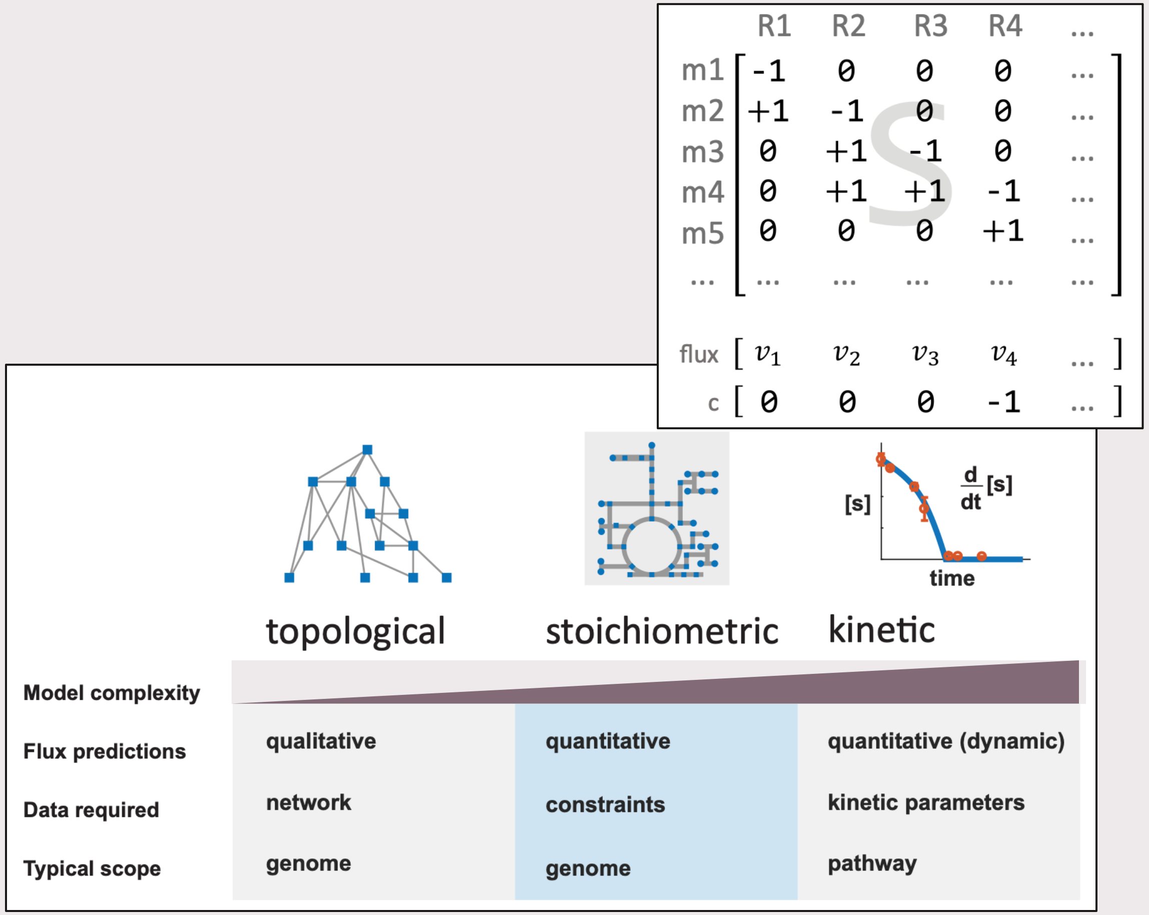
“The ultimate goal of **genome-scale metabolic network** reconstruction in the future is to have a well-annotated network including all parts of the metabolism without any missing reactions or gaps; however it is not yet possible due to incomplete knowledge of the yeast metabolism.”

— *Österlund et. al (2012)*

Metabolic modelling *Saccharomyces cerevisiae*

<u>Genome scale metabolic model</u>

- Some quantities can be measure directly
- Others are abstractions (e.g. metabolic fluxes)
- Common approach is to encode biological knowledge as constraints
    - Either from observed experiments or
    - Biophysical knowledge



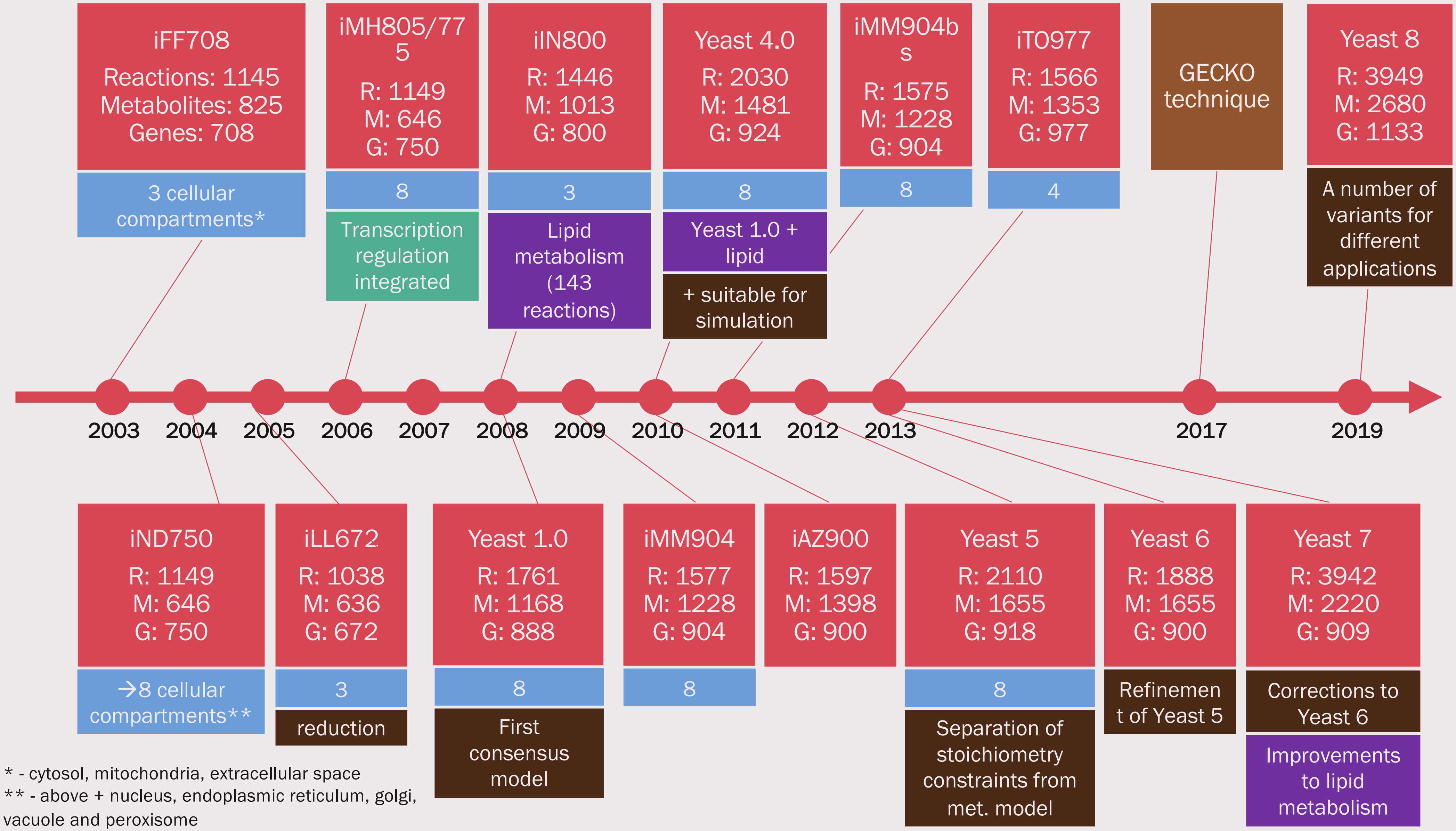
$$
\begin{array}{c|cccccc}
 & R1 & R2 & R3 & R4 & \dots \\
\hline
m1 & -1 & 0 & 0 & 0 & \dots \\
m2 & +1 & -1 & 0 & 0 & \dots \\
m3 & 0 & +1 & -1 & 0 & \dots \\
m4 & 0 & +1 & +1 & -1 & \dots \\
m5 & 0 & 0 & 0 & +1 & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots
\end{array}
$$

$$\text{flux } [\, v_1 \quad v_2 \quad v_3 \quad v_4 \quad \dots \,]$$

$$\text{c } [\, 0 \quad 0 \quad 0 \quad -1 \quad \dots \,]$$

	topological	stoichiometric	kinetic
Model complexity			
Flux predictions	qualitative	quantitative	quantitative (dynamic)
Data required	network	constraints	kinetic parameters
Typical scope	genome	genome	pathway

Figures modified from Avlant Nilsson PhD thesis (2019)



2003
iFF708
Reactions: 1145
Metabolites: 825
Genes: 708
3 cellular compartments*

2004
iND750
R: 1149
M: 646
G: 750
→8 cellular compartments**

2005
iLL672
R: 1038
M: 636
G: 672
3
reduction

2006
iMH805/775
R: 1149
M: 646
G: 750
8
Transcription regulation integrated

2008
iIN800
R: 1446
M: 1013
G: 800
3
Lipid metabolism (143 reactions)

Yeast 1.0
R: 1761
M: 1168
G: 888
8
First consensus model

2009
iMM904
R: 1577
M: 1228
G: 904
8

2010
Yeast 4.0
R: 2030
M: 1481
G: 924
8
Yeast 1.0 + lipid

iAZ900
R: 1597
M: 1398
G: 900

2011
+ suitable for simulation

2012
iMM904bs
R: 1575
M: 1228
G: 904
8

Yeast 5
R: 2110
M: 1655
G: 918
8
Separation of stoichiometry constraints from met. model

2013
iTO977
R: 1566
M: 1353
G: 977
4

Yeast 6
R: 1888
M: 1655
G: 900
Refinement of Yeast 5

Yeast 7
R: 3942
M: 2220
G: 909
Corrections to Yeast 6
Improvements to lipid metabolism

2017
GECKO technique

2019
Yeast 8
R: 3949
M: 2680
G: 1133
A number of variants for different applications

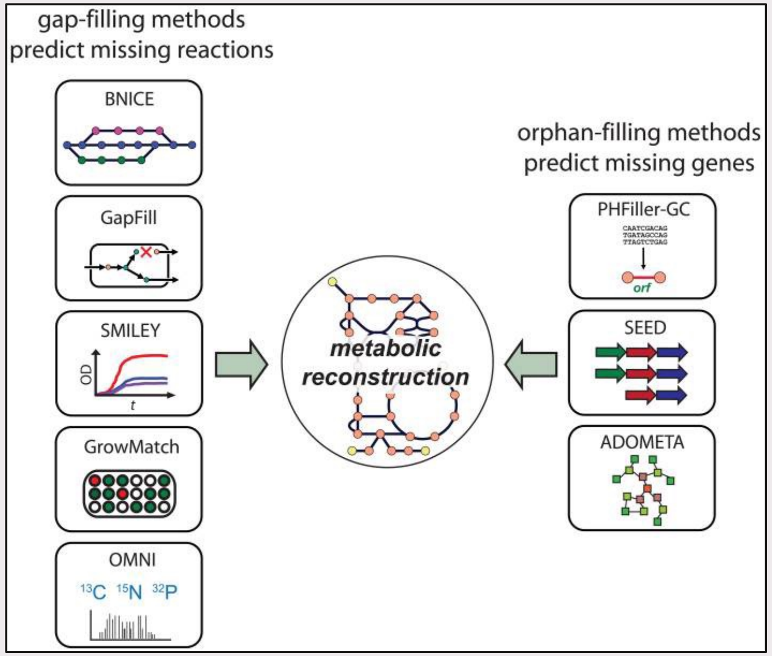
* - cytosol, mitochondria, extracellular space
** - above + nucleus, endoplasmic reticulum, golgi, vacuole and peroxisome

Model improvement

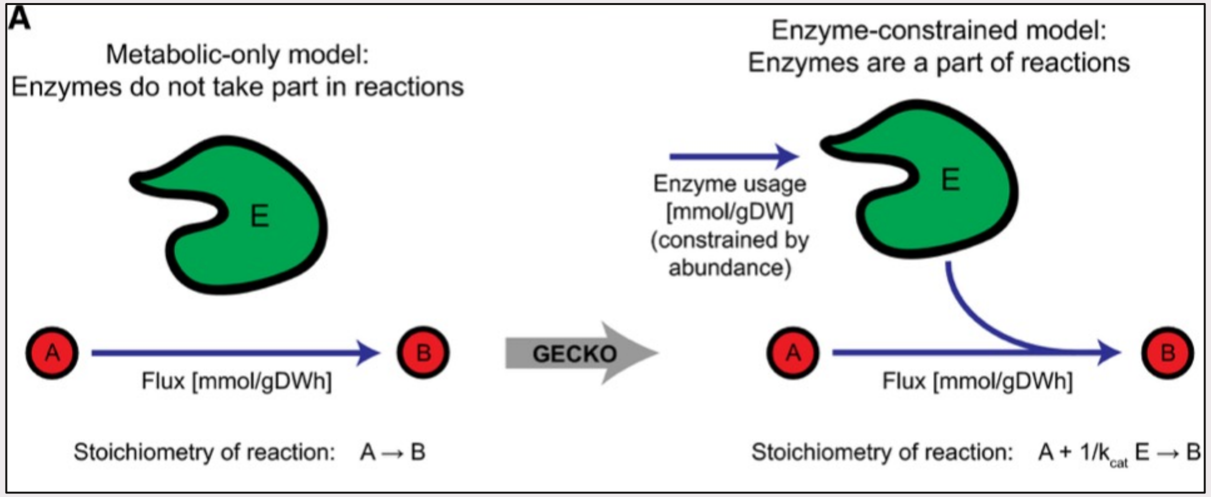
- Model reduction
- Model expansion (new annotation)
  - Regulatory interaction
  - Assign gene to known reaction
  - Predict missing reaction
- Condition-specific effects
- Compartmentalisation (split model over parts of cell)
- New mathematical rules (enzymatic rate equation

$$v_i = \overline{k_{cat,i}} \cdot \sum E_i \cdot \rho_i$$

- New constraint mechanism (e.g. introducing enzymes explicitly into equations, GECKO)



Orth & Palsson (2010)


A
Metabolic-only model:
Enzymes do not take part in reactions
E
A
B
Flux [mmol/gDWh]
Stoichiometry of reaction: A → B
GECKO
Enzyme-constrained model:
Enzymes are a part of reactions
Enzyme usage [mmol/gDW] (constrained by abundance)
E
A
B
Flux [mmol/gDWh]
Stoichiometry of reaction: $A + 1/k_{cat}\ E \rightarrow B$

Sanchez, Zhang et. al. (2017)

Humans and machines working together

- Model reduction
- Model expansion (new annotation)
  - Regulatory interaction
  - Assign gene to known reaction
  - Predict missing reaction
- Condition-specific effects

Algorithms exist or are being developed for these methods

- Compartmentalisation (split model over parts of cell)
- New mathematical rules (enzymatic rate equation

$$v_i = \overline{k_{cat,i}} \cdot \sum E_i \cdot \rho_i$$

- New constraint mechanism (e.g. introducing enzymes explicitly into equations, GECKO)

Currently proposed by human scientists – require a level of abstraction

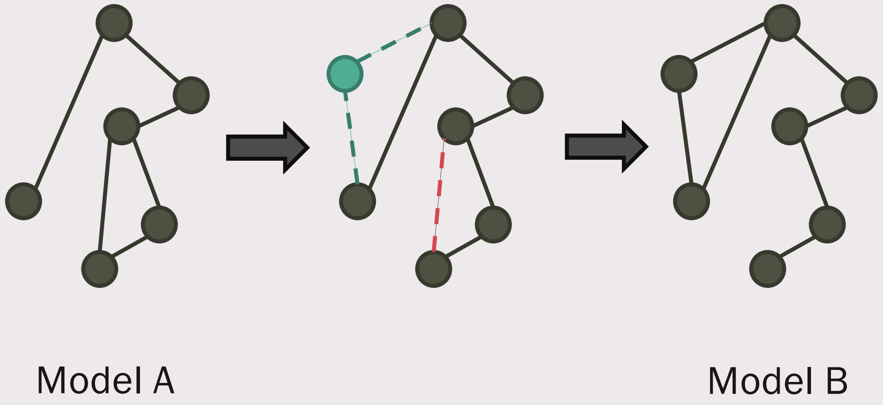
How to compare models

Good models have:

- explanatory power
- predictive power
- consistency across contexts
- consistency with other scientific models

Many possible metrics one could use:

- genomic coverage;
- overlap of annotated metabolites;
- predictive ability for single gene essentiality;
- biomass production prediction;
- …



- It is difficult to find a single metric that can summarise a model’s quality
- Among *S. cerevisiae* models there is evidence of tradeoffs between predictive accuracy and gene network coverage (Heavner and Price, 2015)
- Models are often developed for specific applications

- ● The original discovery problem
- ● How we formulated the problem in computational terms
- ○ What data and knowledge we provided to our system
- ○ How we represented the system's inputs and outputs
- ○ The space of candidate models that the system searched
- ○ What criteria it used to evaluate candidate models
- ○ How we interpreted results that the system generated

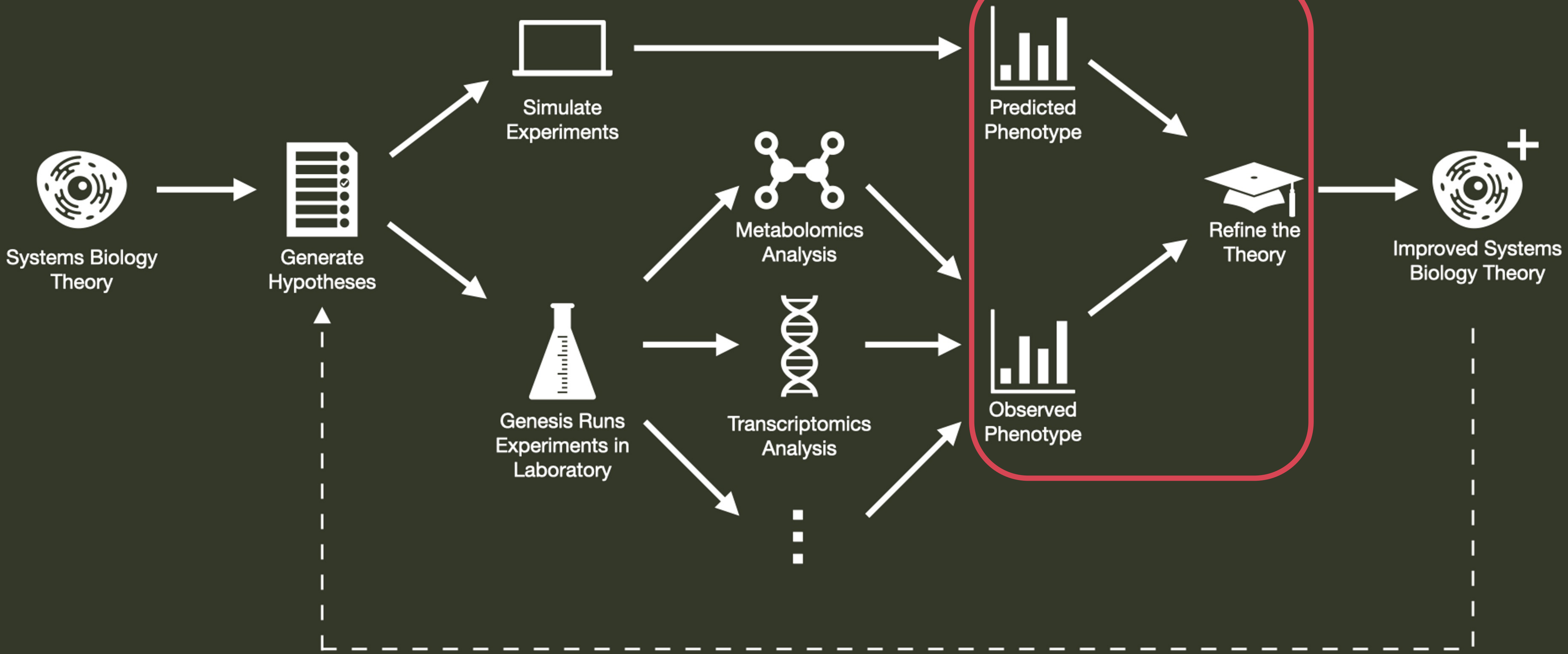
Logical inference

induction	allows us to generalise models from data
deduction	given a theory what conclusions can we draw
abduction	how can we “fix” the theory to be consistent with empirical data?

Active learning

Machine learning paradigm where the learning agent has agency over the selection of the next data to learn from — analagous to the scientific method

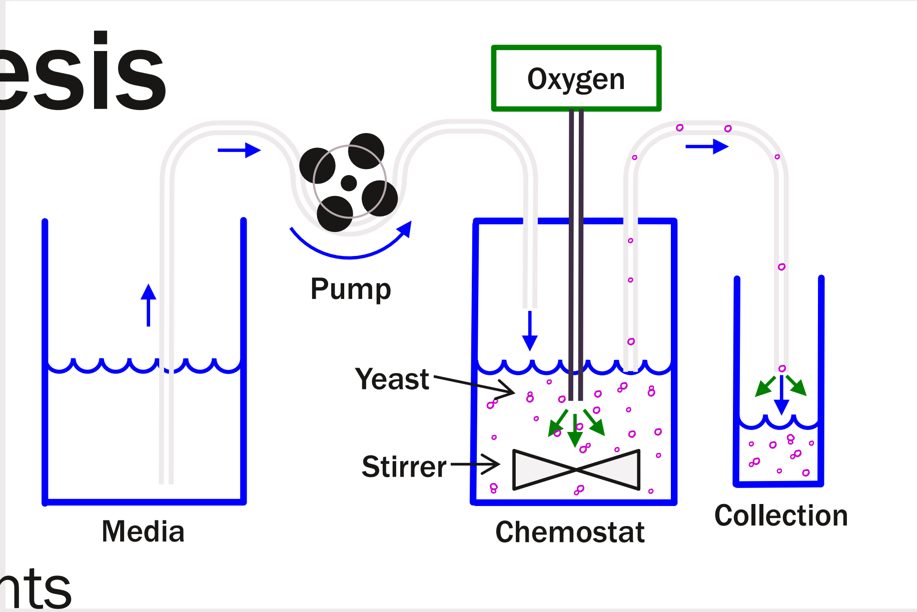
Genesis flow

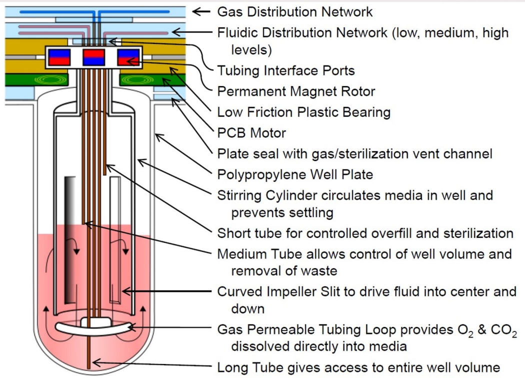


Background: Genesis

- Scalable automated biological experimentation
- Small volume chemostat cultivations – vision is for thousands of parallel experiments
- AI-driven laboratory machine
- Measurements via high-throughput metabolomics and transcriptomics

John Wikswo group VIIBRE




Gas Distribution Network

Fluidic Distribution Network (low, medium, high levels)

Tubing Interface Ports

Permanent Magnet Rotor

Low Friction Plastic Bearing

PCB Motor

Plate seal with gas/sterilization vent channel

Polypropylene Well Plate

Stirring Cylinder circulates media in well and prevents settling

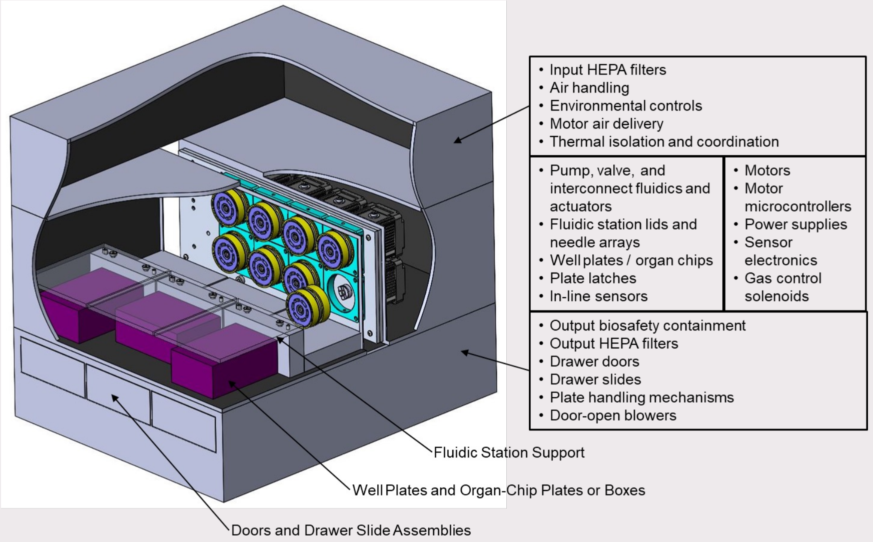
Short tube for controlled overfill and sterilization

Medium Tube allows control of well volume and removal of waste

Curved Impeller Slit to drive fluid into center and down

Gas Permeable Tubing Loop provides O_2 & CO_2 dissolved directly into media

Long Tube gives access to entire well volume


- Input HEPA filters
- Air handling
- Environmental controls
- Motor air delivery
- Thermal isolation and coordination

- Pump, valve, and interconnect fluidics and actuators
- Fluidic station lids and needle arrays
- Well plates / organ chips
- Plate latches
- In-line sensors

- Motors
- Motor microcontrollers
- Power supplies
- Sensor electronics
- Gas control solenoids

- Output biosafety containment
- Output HEPA filters
- Drawer doors
- Drawer slides
- Plate handling mechanisms
- Door-open blowers

Fluidic Station Support

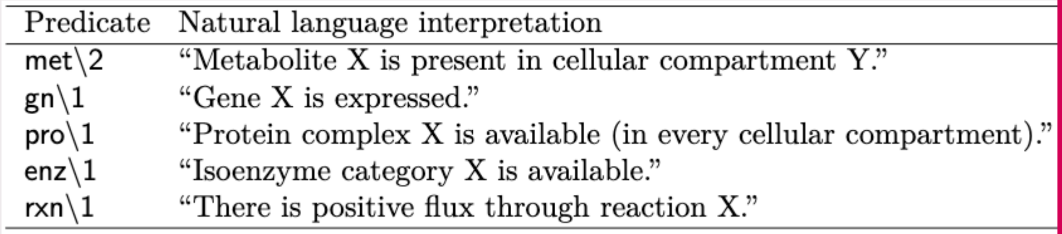
Well Plates and Organ-Chip Plates or Boxes

Doors and Drawer Slide Assemblies

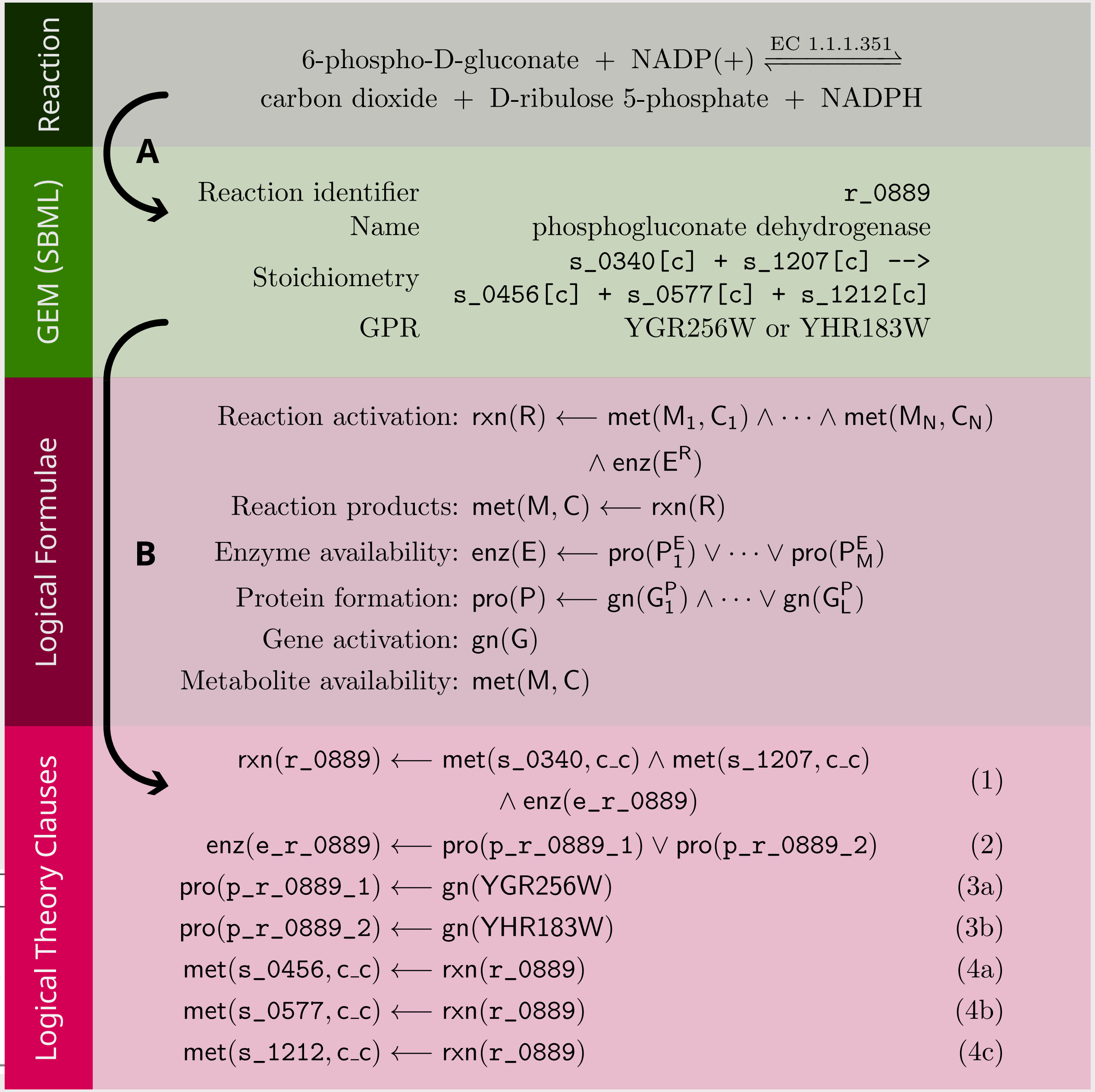
- ● The original discovery problem
- ● How we formulated the problem in computational terms
- ● What data and knowledge we provided to our system
- ○ How we represented the system's inputs and outputs
- ○ The space of candidate models that the system searched
- ○ What criteria it used to evaluate candidate models
- ○ How we interpreted results that the system generated

Constructing a logical theory of metabolism

- Background knowledge is encoded in the curated genome-scale metabolic models (GEMs)
- Each reaction in the GEM is translated to a set first-order logic clauses
- Clauses are written in conjugate normal form to produce theory
- Inference is performed using iProver: a theorem prover for first-order logic with support for arithmetical reasoning (Korovin, 2008)
- iProver has the efficiency required for a logical theory on this scale



Predicate	Natural language interpretation
met\2	"Metabolite X is present in cellular compartment Y."
gn\1	"Gene X is expressed."
pro\1	"Protein complex X is available (in every cellular compartment)."
enz\1	"Isoenzyme category X is available."
rxn\1	"There is positive flux through reaction X."



$$\text{6-phospho-D-gluconate} + \text{NADP(+)} \xrightleftharpoons{\text{EC 1.1.1.351}} \text{carbon dioxide} + \text{D-ribulose 5-phosphate} + \text{NADPH}$$

Reaction identifier	r_0889
Name	phosphogluconate dehydrogenase
Stoichiometry	s_0340[c] + s_1207[c] --> s_0456[c] + s_0577[c] + s_1212[c]
GPR	YGR256W or YHR183W

Reaction activation: $\mathsf{rxn}(R) \longleftarrow \mathsf{met}(M_1, C_1) \wedge \cdots \wedge \mathsf{met}(M_N, C_N) \wedge \mathsf{enz}(E^R)$

Reaction products: $\mathsf{met}(M, C) \longleftarrow \mathsf{rxn}(R)$

Enzyme availability: $\mathsf{enz}(E) \longleftarrow \mathsf{pro}(P_1^E) \vee \cdots \vee \mathsf{pro}(P_M^E)$

Protein formation: $\mathsf{pro}(P) \longleftarrow \mathsf{gn}(G_1^P) \wedge \cdots \vee \mathsf{gn}(G_L^P)$

Gene activation: $\mathsf{gn}(G)$

Metabolite availability: $\mathsf{met}(M, C)$

$$\mathsf{rxn}(\mathtt{r\_0889}) \longleftarrow \mathsf{met}(\mathtt{s\_0340}, \mathsf{c\_c}) \wedge \mathsf{met}(\mathtt{s\_1207}, \mathsf{c\_c}) \wedge \mathsf{enz}(\mathtt{e\_r\_0889}) \tag{1}$$

$$\mathsf{enz}(\mathtt{e\_r\_0889}) \longleftarrow \mathsf{pro}(\mathtt{p\_r\_0889\_1}) \vee \mathsf{pro}(\mathtt{p\_r\_0889\_2}) \tag{2}$$

$$\mathsf{pro}(\mathtt{p\_r\_0889\_1}) \longleftarrow \mathsf{gn}(\mathsf{YGR256W}) \tag{3a}$$

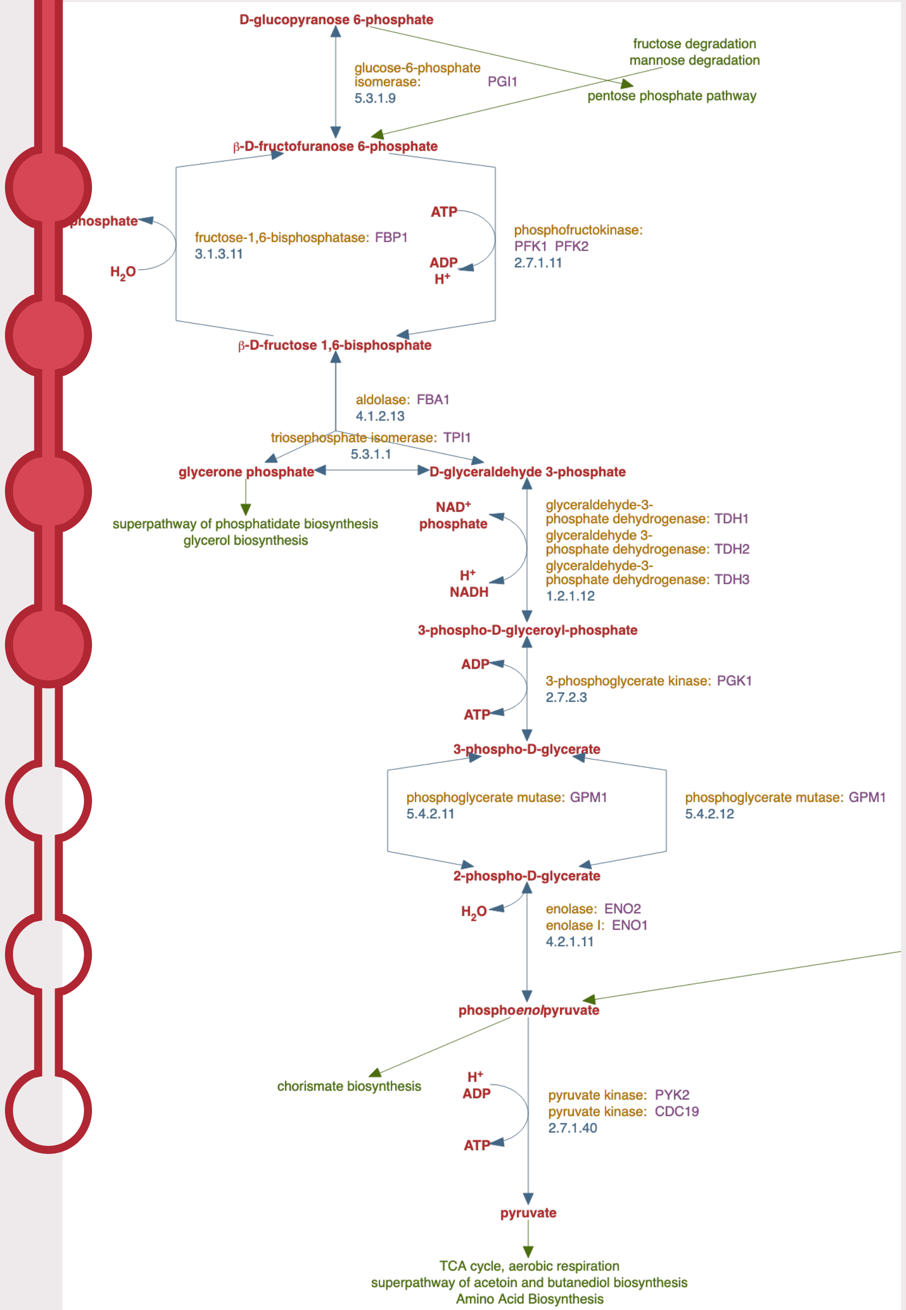
$$\mathsf{pro}(\mathtt{p\_r\_0889\_2}) \longleftarrow \mathsf{gn}(\mathsf{YHR183W}) \tag{3b}$$

$$\mathsf{met}(\mathtt{s\_0456}, \mathsf{c\_c}) \longleftarrow \mathsf{rxn}(\mathtt{r\_0889}) \tag{4a}$$

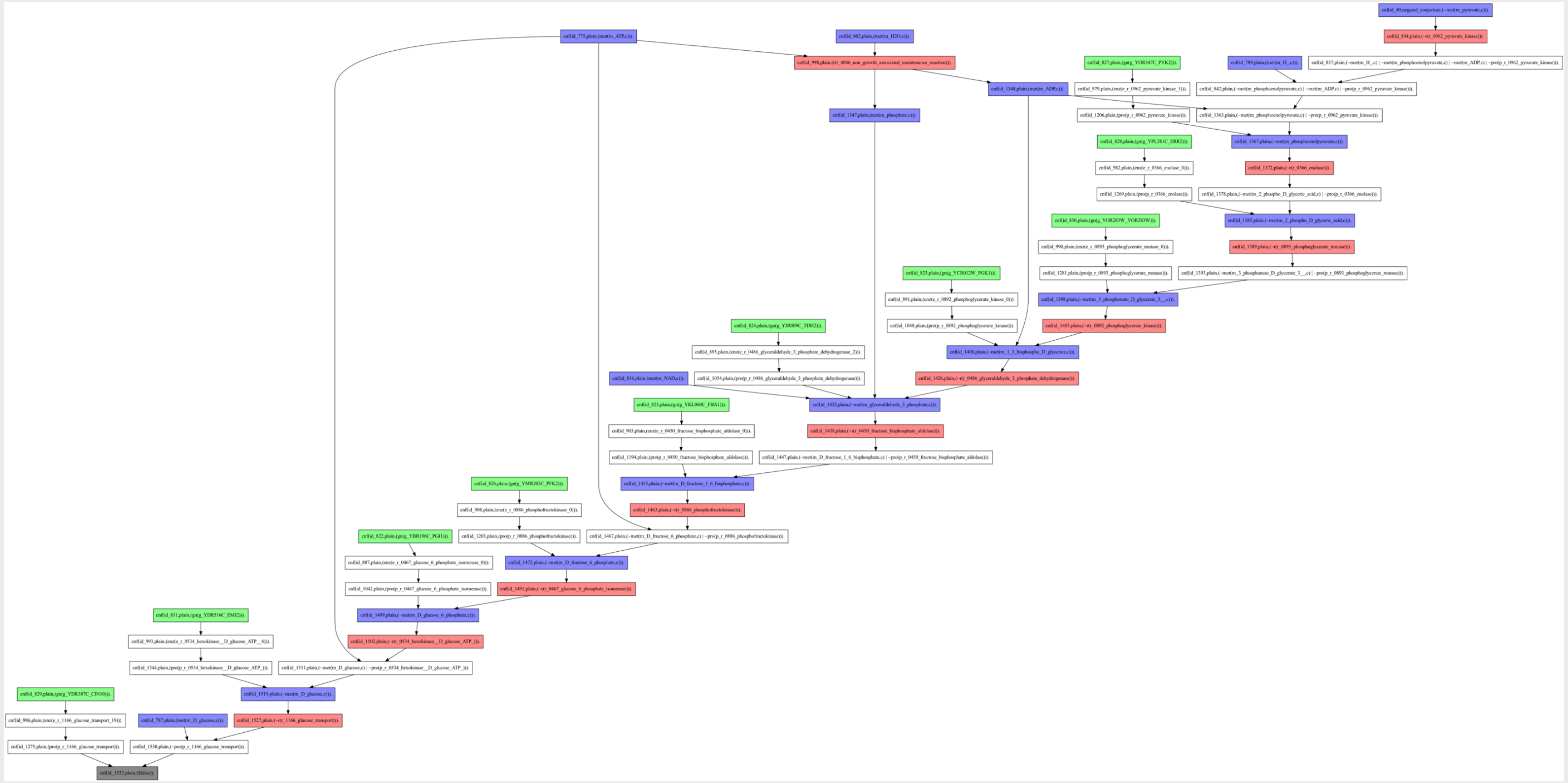
$$\mathsf{met}(\mathtt{s\_0577}, \mathsf{c\_c}) \longleftarrow \mathsf{rxn}(\mathtt{r\_0889}) \tag{4b}$$

$$\mathsf{met}(\mathtt{s\_1212}, \mathsf{c\_c}) \longleftarrow \mathsf{rxn}(\mathtt{r\_0889}) \tag{4c}$$

- The original discovery problem
- How we formulated the problem in computational terms
- What data and knowledge we provided to our system
- How we represented the system's inputs and outputs
- The space of candidate models that the system searched
- What criteria it used to evaluate candidate models
- How we interpreted results that the system generated


D-glucopyranose 6-phosphate
glucose-6-phosphate isomerase: PGI1
5.3.1.9
fructose degradation
mannose degradation
pentose phosphate pathway
β-D-fructofuranose 6-phosphate
phosphate
H_2O
fructose-1,6-bisphosphatase: FBP1
3.1.3.11
ATP
ADP
H^+
phosphofructokinase:
PFK1 PFK2
2.7.1.11
β-D-fructose 1,6-bisphosphate
aldolase: FBA1
4.1.2.13
triosephosphate isomerase: TPI1
5.3.1.1
glycerone phosphate
D-glyceraldehyde 3-phosphate
superpathway of phosphatidate biosynthesis
glycerol biosynthesis
NAD^+
phosphate
H^+
NADH
glyceraldehyde-3-phosphate dehydrogenase: TDH1
glyceraldehyde 3-phosphate dehydrogenase: TDH2
glyceraldehyde-3-phosphate dehydrogenase: TDH3
1.2.1.12
3-phospho-D-glyceroyl-phosphate
ADP
ATP
3-phosphoglycerate kinase: PGK1
2.7.2.3
3-phospho-D-glycerate
phosphoglycerate mutase: GPM1
5.4.2.11
phosphoglycerate mutase: GPM1
5.4.2.12
2-phospho-D-glycerate
H_2O
enolase: ENO2
enolase I: ENO1
4.2.1.11
phosphoenolpyruvate
chorismate biosynthesis
H^+
ADP
ATP
pyruvate kinase: PYK2
pyruvate kinase: CDC19
2.7.1.40
pyruvate
TCA cycle, aerobic respiration
superpathway of acetoin and butanediol biosynthesis
Amino Acid Biosynthesis

Glycolysis pathway (from https://pathway.yeastgenome.org/)



Proof graph from iProver (simplified)

● The original discovery problem

● How we formulated the problem in computational terms

● What data and knowledge we provided to our system

● How we represented the system's inputs and outputs

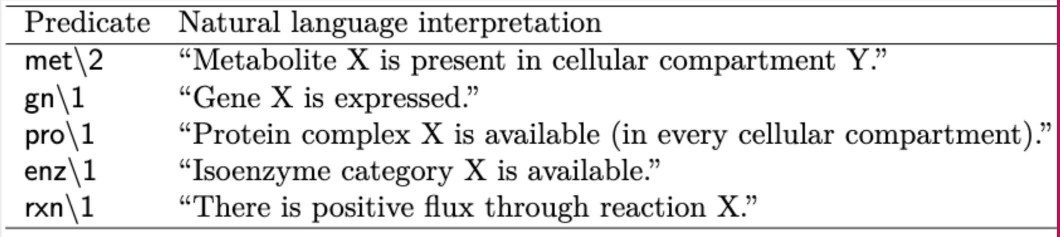
● The space of candidate models that the system searched

○ What criteria it used to evaluate candidate models

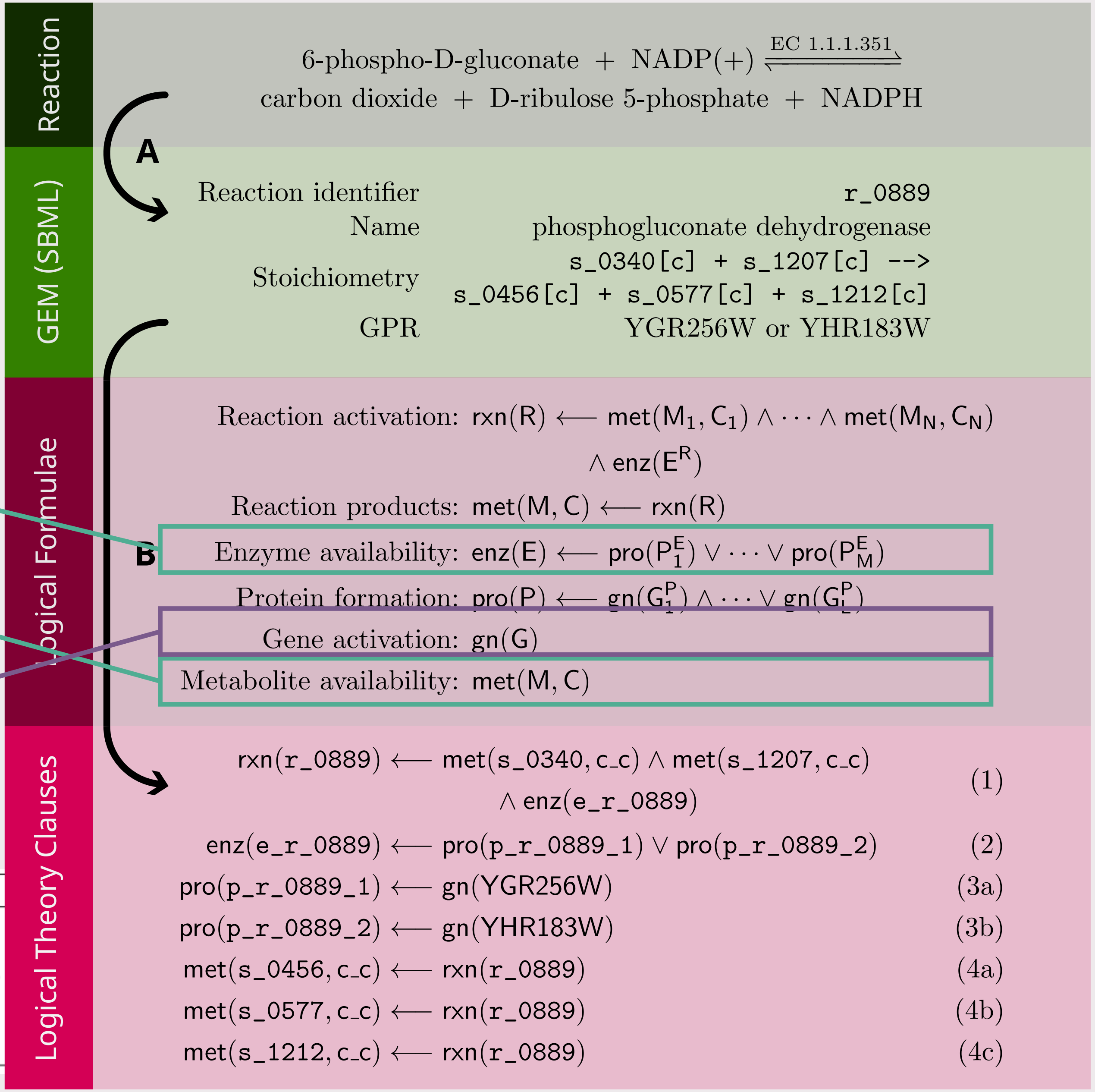
○ How we interpreted results that the system generated

Approach: abduction opportunities

- Some parts of the model are well-known due to the chemistry (reaction stoichiometry, protein formation)
- We seek to:
  a) learn rules about which enzymes catalyse which reactions
  b) identify possible missing reactions by finding compound presence that will repair a broken pathway
- Future work will be to learn rules for gene expression and activation



Predicate	Natural language interpretation
met\2	"Metabolite X is present in cellular compartment Y."
gn\1	"Gene X is expressed."
pro\1	"Protein complex X is available (in every cellular compartment)."
enz\1	"Isoenzyme category X is available."
rxn\1	"There is positive flux through reaction X."



$$\text{6-phospho-D-gluconate} + \text{NADP(+)} \xrightleftharpoons{\text{EC 1.1.1.351}} \text{carbon dioxide} + \text{D-ribulose 5-phosphate} + \text{NADPH}$$

Reaction identifier: r_0889

Name: phosphogluconate dehydrogenase

Stoichiometry: s_0340[c] + s_1207[c] --> s_0456[c] + s_0577[c] + s_1212[c]

GPR: YGR256W or YHR183W

Reaction activation: $\mathsf{rxn}(\mathsf{R}) \longleftarrow \mathsf{met}(\mathsf{M}_1, \mathsf{C}_1) \wedge \cdots \wedge \mathsf{met}(\mathsf{M}_\mathsf{N}, \mathsf{C}_\mathsf{N}) \wedge \mathsf{enz}(\mathsf{E}^\mathsf{R})$

Reaction products: $\mathsf{met}(\mathsf{M}, \mathsf{C}) \longleftarrow \mathsf{rxn}(\mathsf{R})$

Enzyme availability: $\mathsf{enz}(\mathsf{E}) \longleftarrow \mathsf{pro}(\mathsf{P}_1^\mathsf{E}) \vee \cdots \vee \mathsf{pro}(\mathsf{P}_\mathsf{M}^\mathsf{E})$

Protein formation: $\mathsf{pro}(\mathsf{P}) \longleftarrow \mathsf{gn}(\mathsf{G}_1^\mathsf{P}) \wedge \cdots \vee \mathsf{gn}(\mathsf{G}_\mathsf{L}^\mathsf{P})$

Gene activation: $\mathsf{gn}(\mathsf{G})$

Metabolite availability: $\mathsf{met}(\mathsf{M}, \mathsf{C})$

$$\mathsf{rxn}(\mathtt{r\_0889}) \longleftarrow \mathsf{met}(\mathtt{s\_0340}, \mathtt{c\_c}) \wedge \mathsf{met}(\mathtt{s\_1207}, \mathtt{c\_c}) \wedge \mathsf{enz}(\mathtt{e\_r\_0889}) \quad (1)$$

$$\mathsf{enz}(\mathtt{e\_r\_0889}) \longleftarrow \mathsf{pro}(\mathtt{p\_r\_0889\_1}) \vee \mathsf{pro}(\mathtt{p\_r\_0889\_2}) \quad (2)$$

$$\mathsf{pro}(\mathtt{p\_r\_0889\_1}) \longleftarrow \mathsf{gn}(\mathsf{YGR256W}) \quad (3a)$$

$$\mathsf{pro}(\mathtt{p\_r\_0889\_2}) \longleftarrow \mathsf{gn}(\mathsf{YHR183W}) \quad (3b)$$

$$\mathsf{met}(\mathtt{s\_0456}, \mathtt{c\_c}) \longleftarrow \mathsf{rxn}(\mathtt{r\_0889}) \quad (4a)$$

$$\mathsf{met}(\mathtt{s\_0577}, \mathtt{c\_c}) \longleftarrow \mathsf{rxn}(\mathtt{r\_0889}) \quad (4b)$$

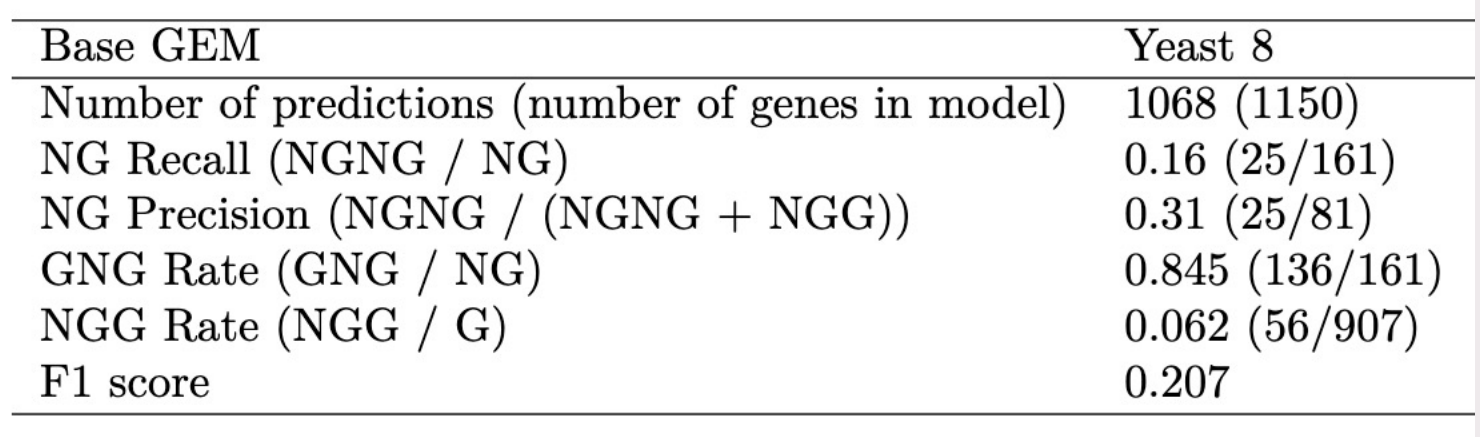
$$\mathsf{met}(\mathtt{s\_1212}, \mathtt{c\_c}) \longleftarrow \mathsf{rxn}(\mathtt{r\_0889}) \quad (4c)$$

- The original discovery problem
- How we formulated the problem in computational terms
- What data and knowledge we provided to our system
- How we represented the system's inputs and outputs
- The space of candidate models that the system searched
- What criteria it used to evaluate candidate models
- How we interpreted results that the system generated

Evaluating models: predicting single-gene essentiality

- Which of the ~6000 genes are necessary for healthy growth?
- We assume the presence of extracellular metabolites that correspond to YNB (yeast nitrogen base) plus glucose
- Systematically remove a gene from the model by negating the relevant clause

  g(g_to_knock_out) becomes ~g(g_to_knock_out)
- Then query for the essential metabolites



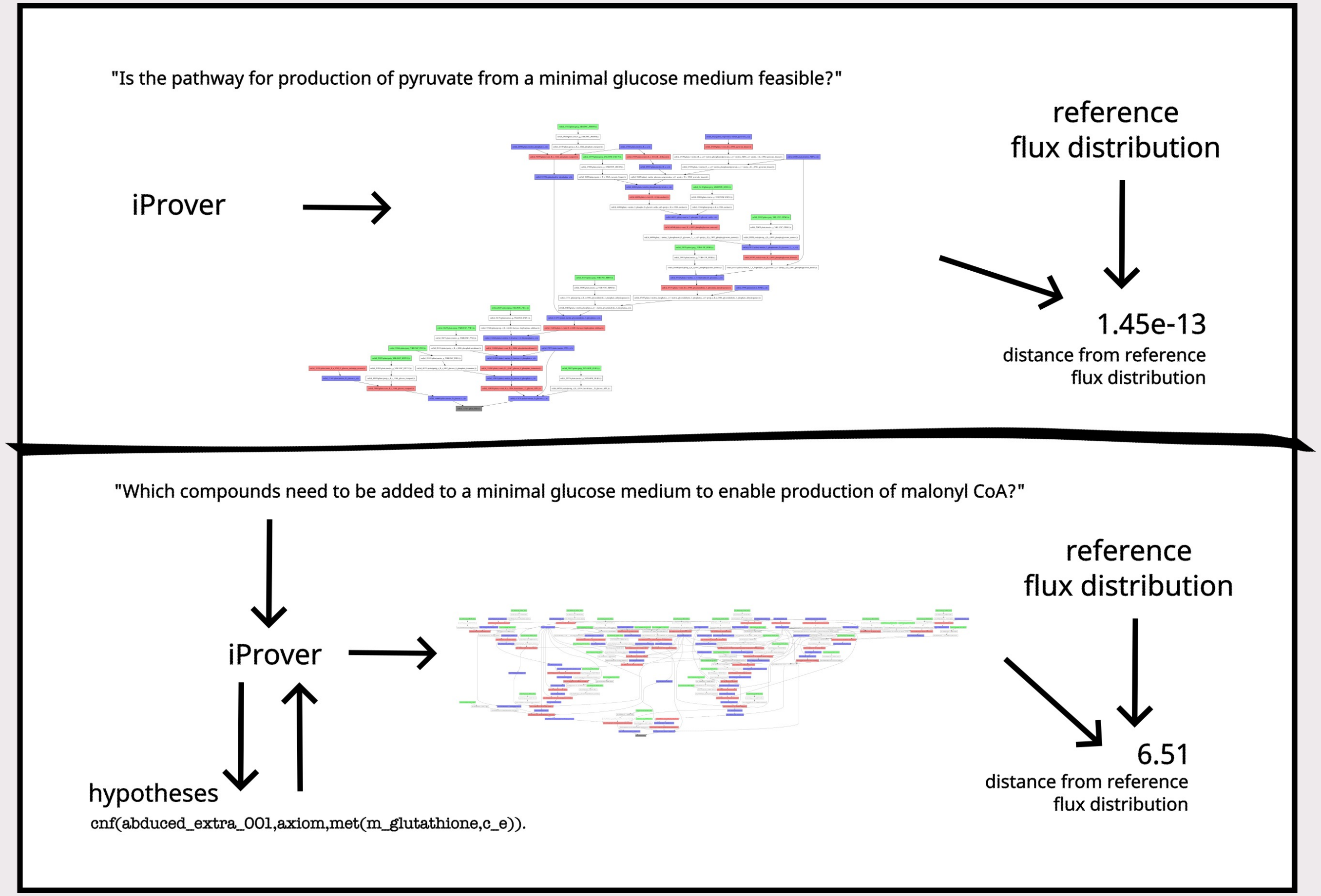
Base GEM	Yeast 8
Number of predictions (number of genes in model)	1068 (1150)
NG Recall (NGNG / NG)	0.16 (25/161)
NG Precision (NGNG / (NGNG + NGG))	0.31 (25/81)
GNG Rate (GNG / NG)	0.845 (136/161)
NGG Rate (NGG / G)	0.062 (56/907)
F1 score	0.207

Evaluating models: constraining flux balance simulations

Stoichiometric matrix – S

Fluxes – $\boldsymbol{\nu}$

$$\begin{aligned} \underset{\boldsymbol{\nu} \in \mathbb{R}^n}{\text{maximize}} \quad & f(\nu_1, \ldots, \nu_n) \\ \text{subject to} \quad & S\boldsymbol{\nu} = \mathbf{0}, \\ & \nu_i^{\mathrm{LB}} \leq \nu_i \leq \nu_i^{\mathrm{UB}}, \quad i = 1, \ldots, n. \end{aligned}$$



"Is the pathway for production of pyruvate from a minimal glucose medium feasible?"

iProver

reference flux distribution

1.45e-13

distance from reference flux distribution

"Which compounds need to be added to a minimal glucose medium to enable production of malonyl CoA?"

iProver

hypotheses

cnf(abduced_extra_001,axiom,met(m_glutathione,c_e)).

reference flux distribution

6.51

distance from reference flux distribution

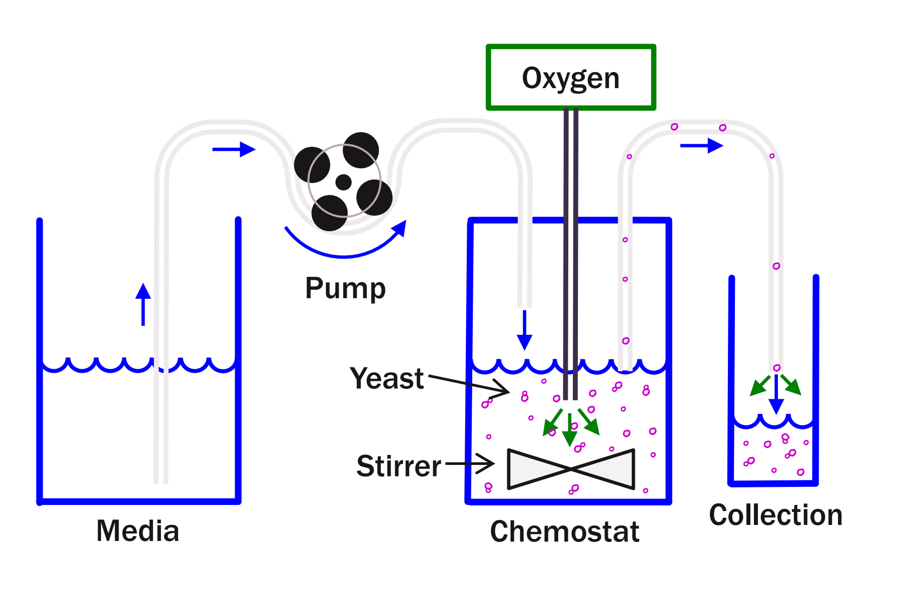
- The original discovery problem
- How we formulated the problem in computational terms
- What data and knowledge we provided to our system
- How we represented the system's inputs and outputs
- The space of candidate models that the system searched
- What criteria it used to evaluate candidate models
- How we interpreted results that the system generated

Interpreting results

- Can use automated theorem prover (iProver) for deduction but also abduction—has the efficiency required for models of this size
- Finding models is easy, finding good models is hard
- Using multiple deduction techniques can help check model consistency—bridge the divide
- Conflicting evidence in literature—e.g. pathway for L-arginine production
- Limits to what one can do with others’ data

Future work

- Hypothesis testing with Genesis platform
- Integration of metabolomics and transciptomics measurements
- Learn rules for gene expression and regulation
- Integrate with signalling pathways



Acknowledgements

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Konstantin Korovin

King Lab

Erik Bjurström
Daniel Brunnsåker
Filip Kronström
Praphapan Lasin
Gabriel Reder
Ievgeniia Tiukova
Ross D. King

CHALMERS
UNIVERSITY OF TECHNOLOGY

Vanderbilt University

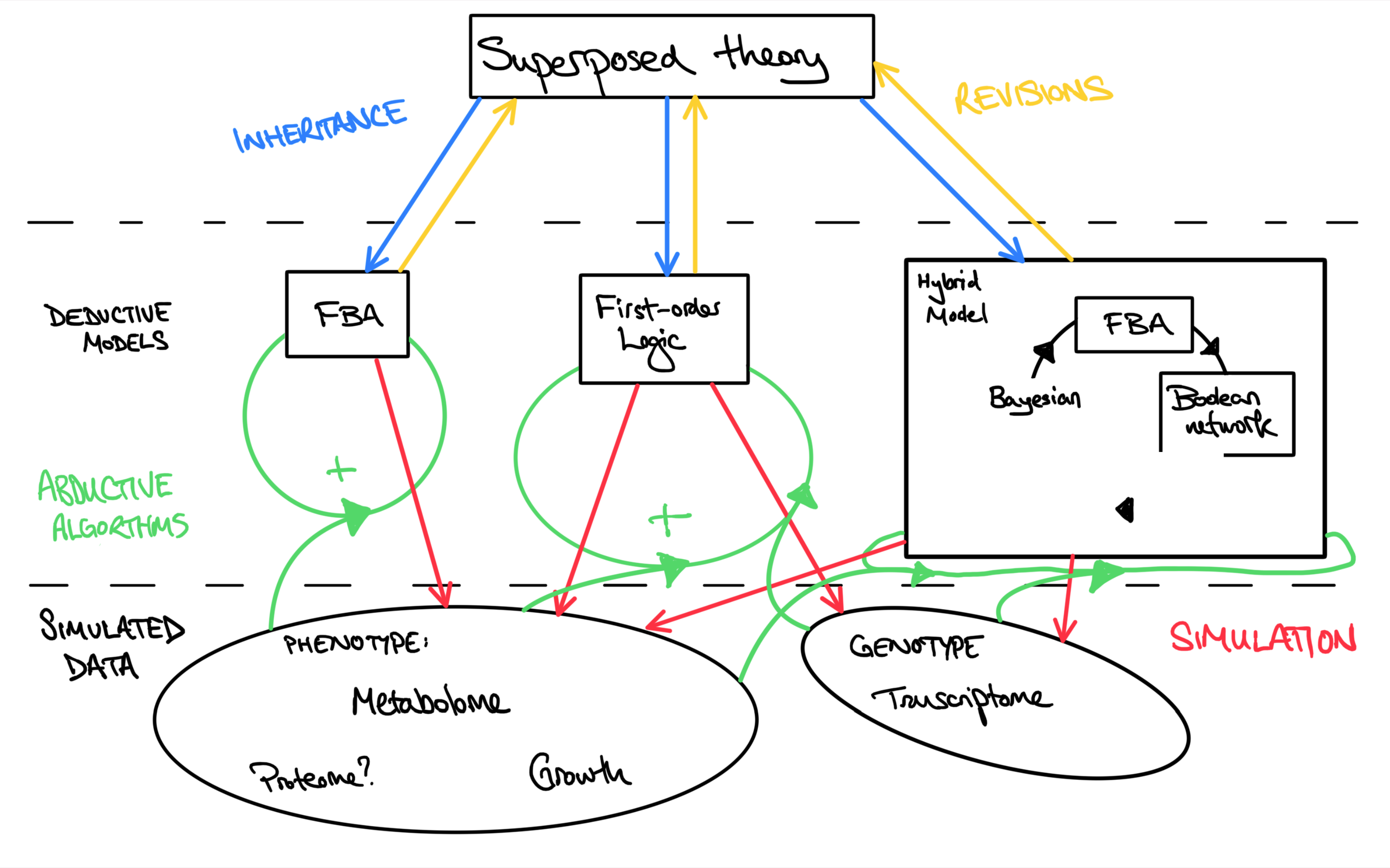
John Wikswo
Ron Reiserer
Clayton Britt
Greg Gerken
Kyle Hawkins
Dmitry Markov
Philip Samson
David Schaffer
Eric Spivey
Erik Werner

Larisa N. Soldatova

Thoughtworks

Rushikesh Halle
Vinay Mahamuni
Amit Patel
Harshal Hayatnagarkar

...and more!


Superposed theory
INHERITANCE
REVISIONS
DEDUCTIVE MODELS
FBA
First-order Logic
Hybrid Model
FBA
Bayesian
Boolean network
ABDUCTIVE ALGORITHMS
+
+
SIMULATED DATA
PHENOTYPE:
Metabolome
Proteome?
Growth
GENOTYPE
Transcriptome
SIMULATION

Areas of improvement

Challenge - _from Chen Y, Li F, Nielsen J (2022)_*	Hypotheses	Techniques to test hypotheses
Annotation of the model	Correct formulae, charge etc. for metabolites; enzyme numbers; reaction directionality	Mass balance analysis; ??
Noise from low-confidence components	Removal of certain elements from model; additional dimension of model “explains” noise	; condition-specific evaluation techniques
“Dead-end” metabolites	Add reactions	Predictive accuracy of metabolic activity with/without reactions; thermodynamic balance analysis
Un- or poorly-annotated reactions (in particular transport reactions)	New or changed gene-protein rule	Comparing transcriptomics or proteomics data with metabolic activity; mutant strain cultivation
Changes to biomass equation itself	Coefficient change; variable addition or removal; condition-specific biomass equations	Comparative prediction analysis
Enzyme turnover rate estimation	Values for enzyme turnover rates	Experimentally measure enzyme levels; comparitive prediction analysis
Integration of **subcellular constraints**	Reaction constraints e.g. within mitochondrion	Quantification of sub-cellular proteomes; ??
Integration of regulation mechanisms	More precise mathematical formulations of regulatory mechanisms; better models of currently understood mechanisms of regulation	Comparative prediction analysis; multi-omics analysis; mutant strain cultivation

* - Y. Chen, F. Li, and J. Nielsen, ‘Genome-scale modeling of yeast metabolism: retrospectives and perspectives’, *FEMS Yeast Research*, vol. 22, no. 1, p. foac003, Jan. 2022, doi: 10.1093/femsyr/foac003.

Role of automation

> “Note that all the computer-predicted additions should be recorded and require expert level verification.” — *Yu et. Al (2022)*

Automated laboratory processes and AI techniques bring:

1. Efficiency
2. Broader reasoning (as opposed to deep reasoning right now)
3. Precision and repeatability

I would summarise by saying the strength of automation (in scientific discovery for *S. cerevisiae* models) will be in embracing complexity by **avoiding excessive simplification** during reasoning and exploiting big datasets to **extract small signals from large noise**, particularly when it comes to testing condition-specific models.