Whole-Cell Modeling

PREDICTING PHENOTYPE FROM GENOTYPE
FOR SCIENCE, MEDICINE & ENGINEERING

Toward WC models for predicting cellular phenotypes

Jonathan Karr

karr@mssm.edu

July 17, 2019

Mount Sinai

Join us at Mount Sinai!

karr@mssm.edu

KarrLab.org

Acknowledgements

Yassmine Chebaro

Yin Hoon Chew

Arthur Goldberg

Zhouyang Lian

Yosef Roth

John Sekar

Bilal Shaikh

Balazs Szigeti

Whole-Cell Modeling

Mount Sinai

ERA SynBio

NIH

NSF

Outline

Genotype to cellular phenotype

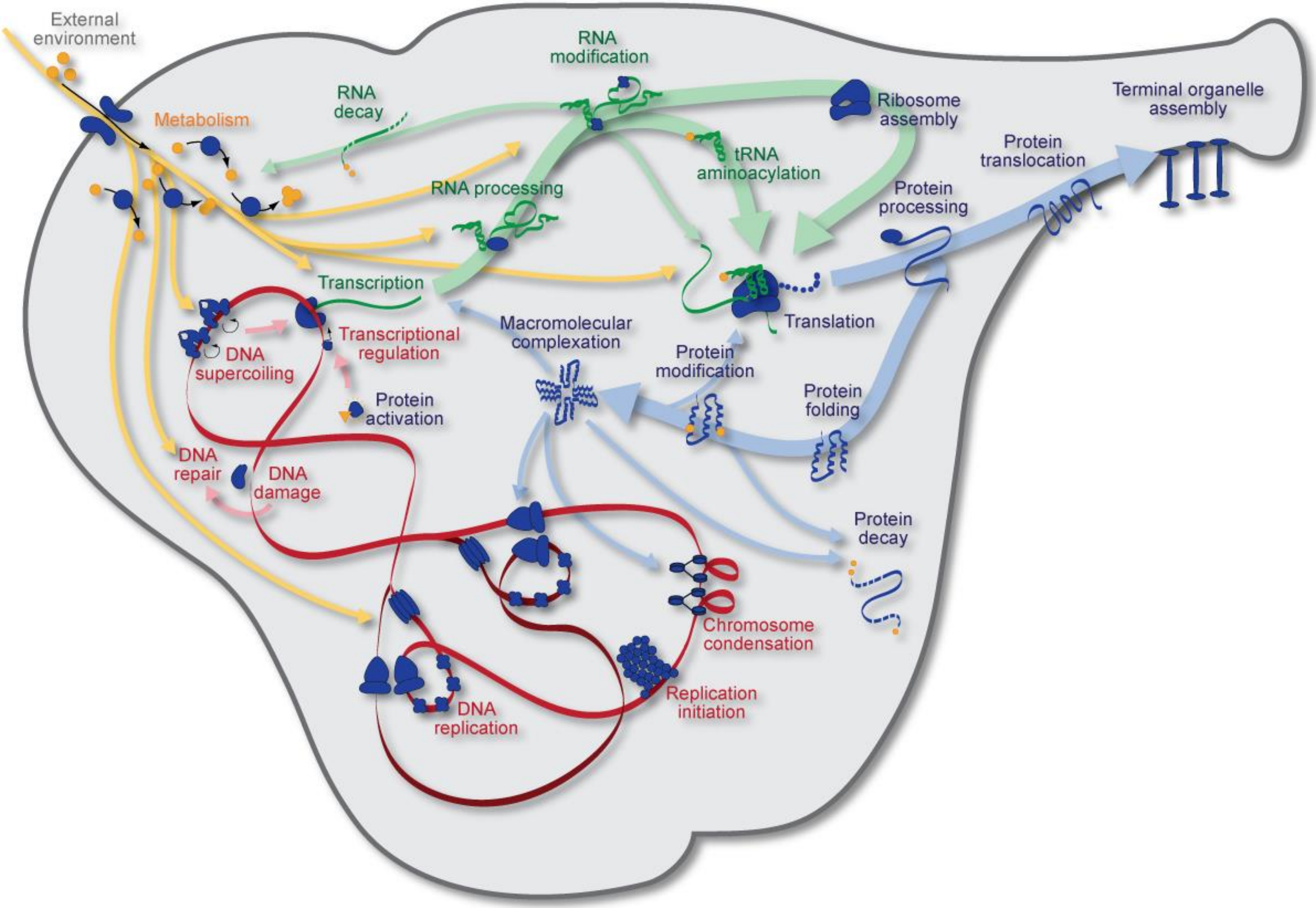
- What is a WC model?
- Why do we need WC models?
- Challenges & feasibility
- Foundational principles and state of the art
- Progress toward comprehensive models

Tips for modeling complex systems

What is a WC model?

Whole-Cell Modeling

Mount Sinai


External environment
Metabolism
RNA decay
RNA modification
Ribosome assembly
Terminal organelle assembly
tRNA aminoacylation
Protein translocation
RNA processing
Protein processing
Transcription
Translation
Macromolecular complexation
Transcriptional regulation
DNA supercoiling
Protein modification
Protein folding
Protein activation
DNA repair
DNA damage
Protein decay
Chromosome condensation
Replication initiation
DNA replication

Goals of WC modeling

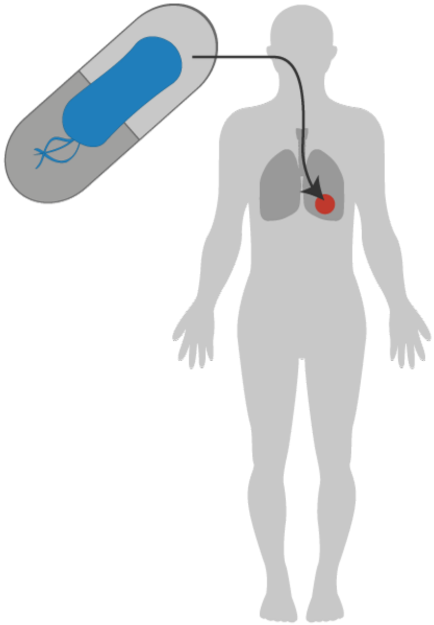
- AGTC Species-specific
- Whole cell
- Whole genome
- Whole cell cycle
- Mechanistic
- Dynamic
- Stochastic

Motivation

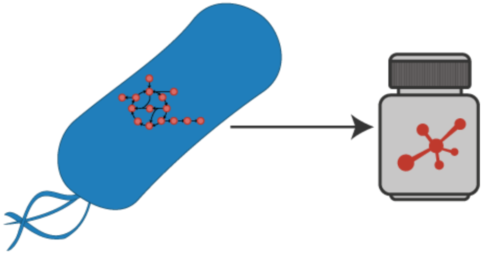
Whole-Cell Modeling

Mount Sinai

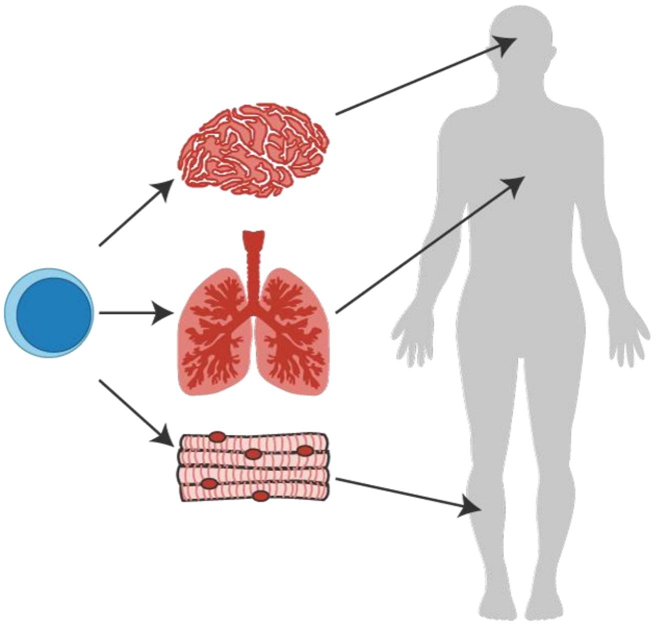
Synthetic biology requires WC models



Biosensors

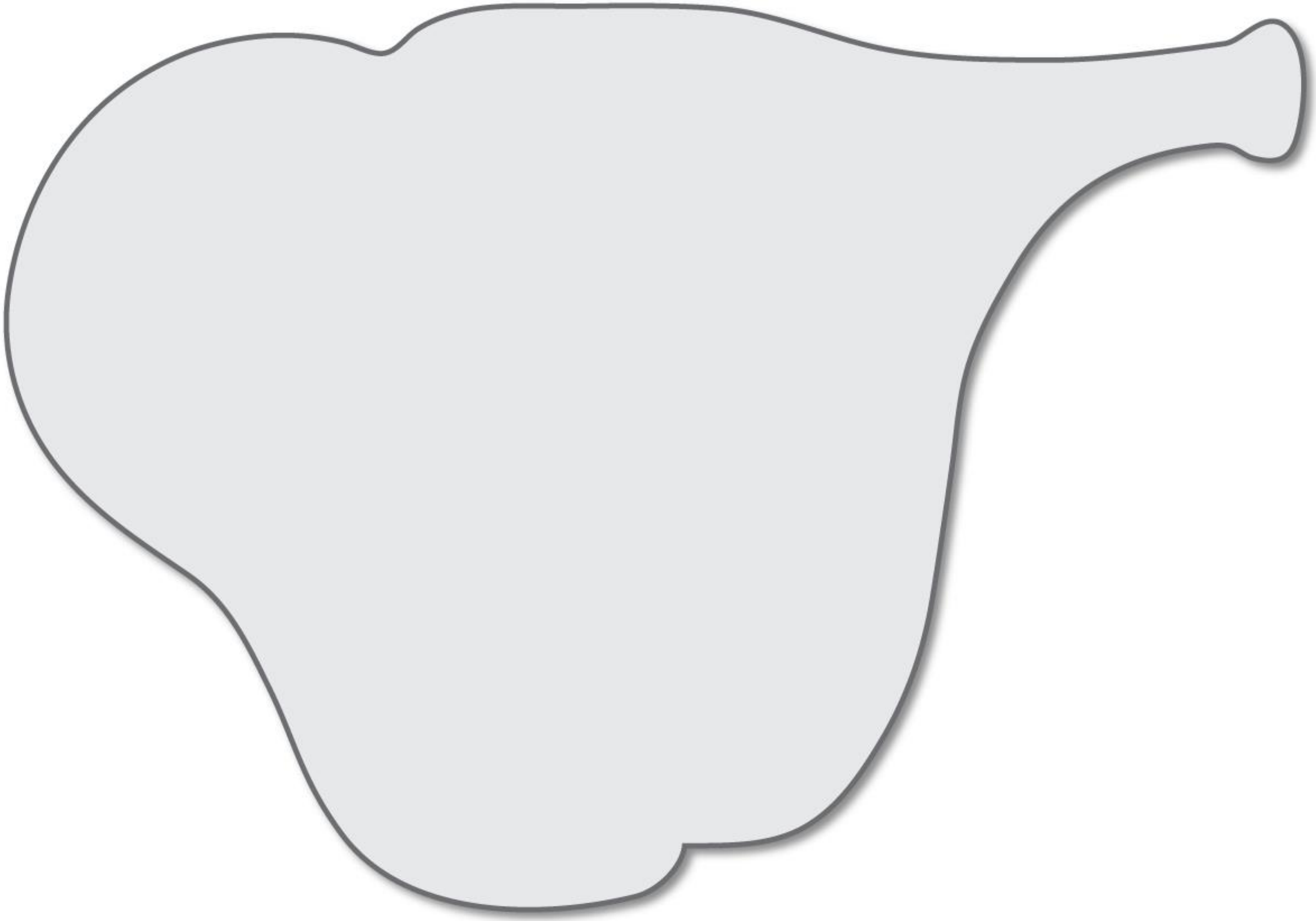


Biofactories

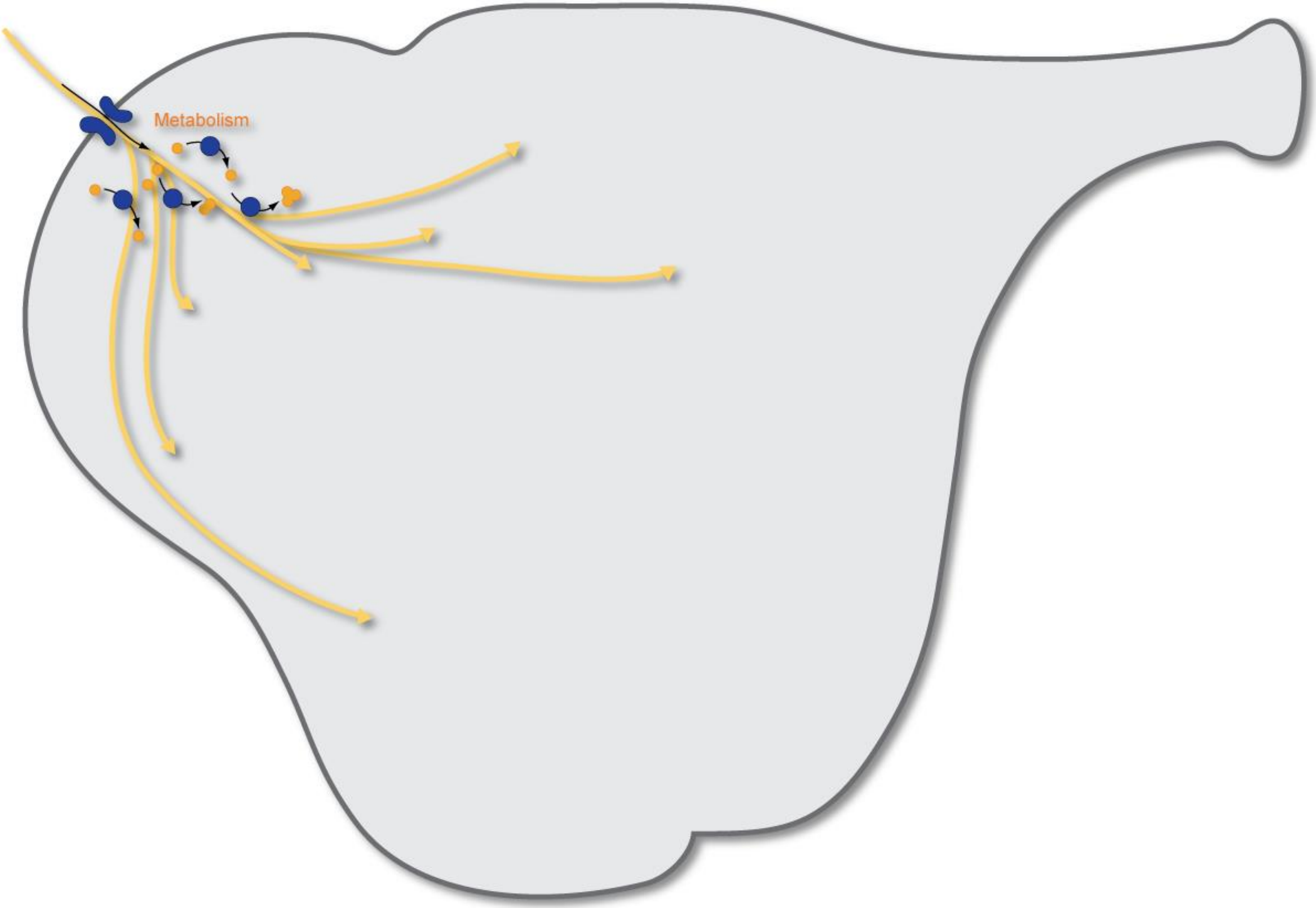


Tissue engineering

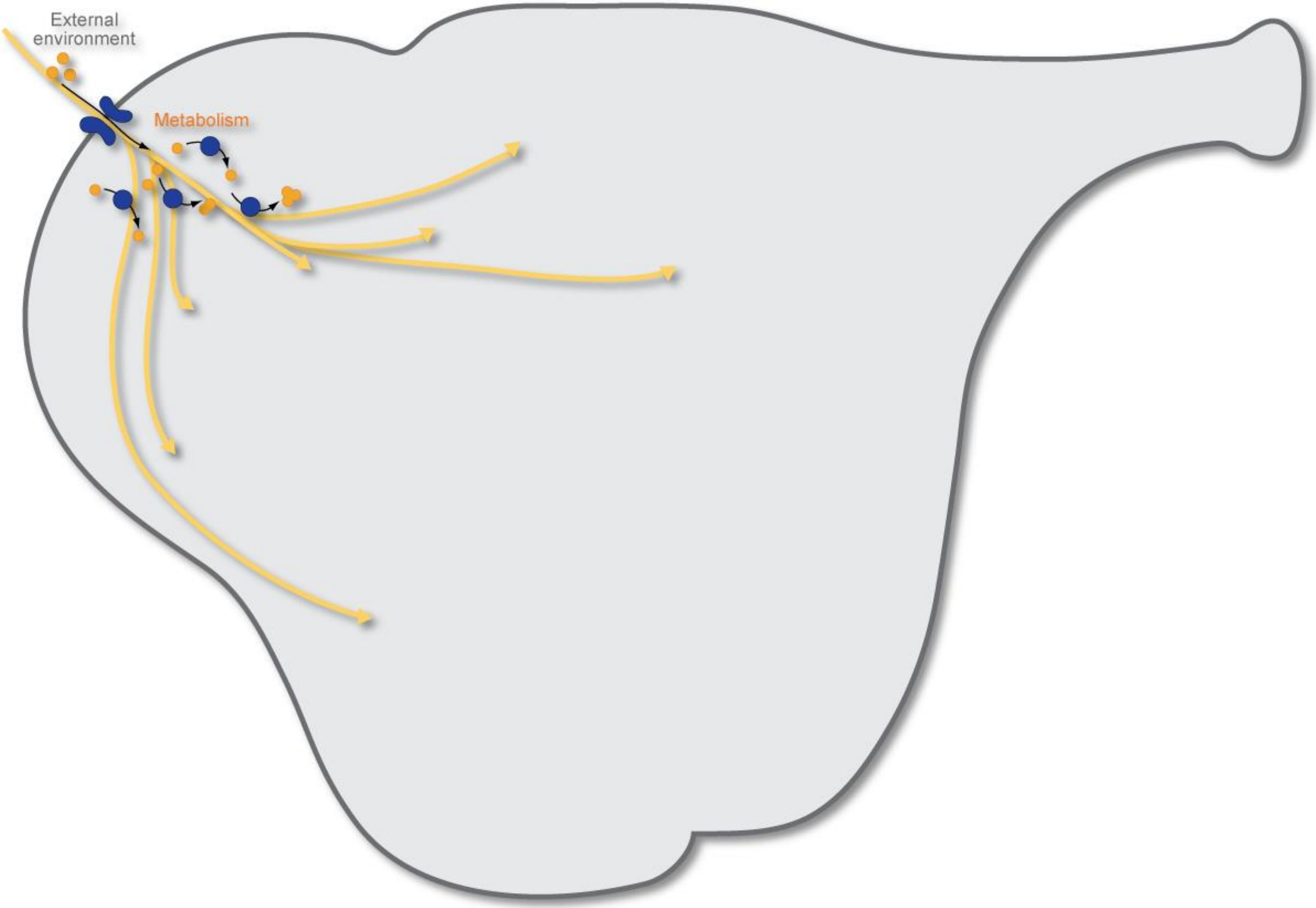
Example: drug biosynthesis



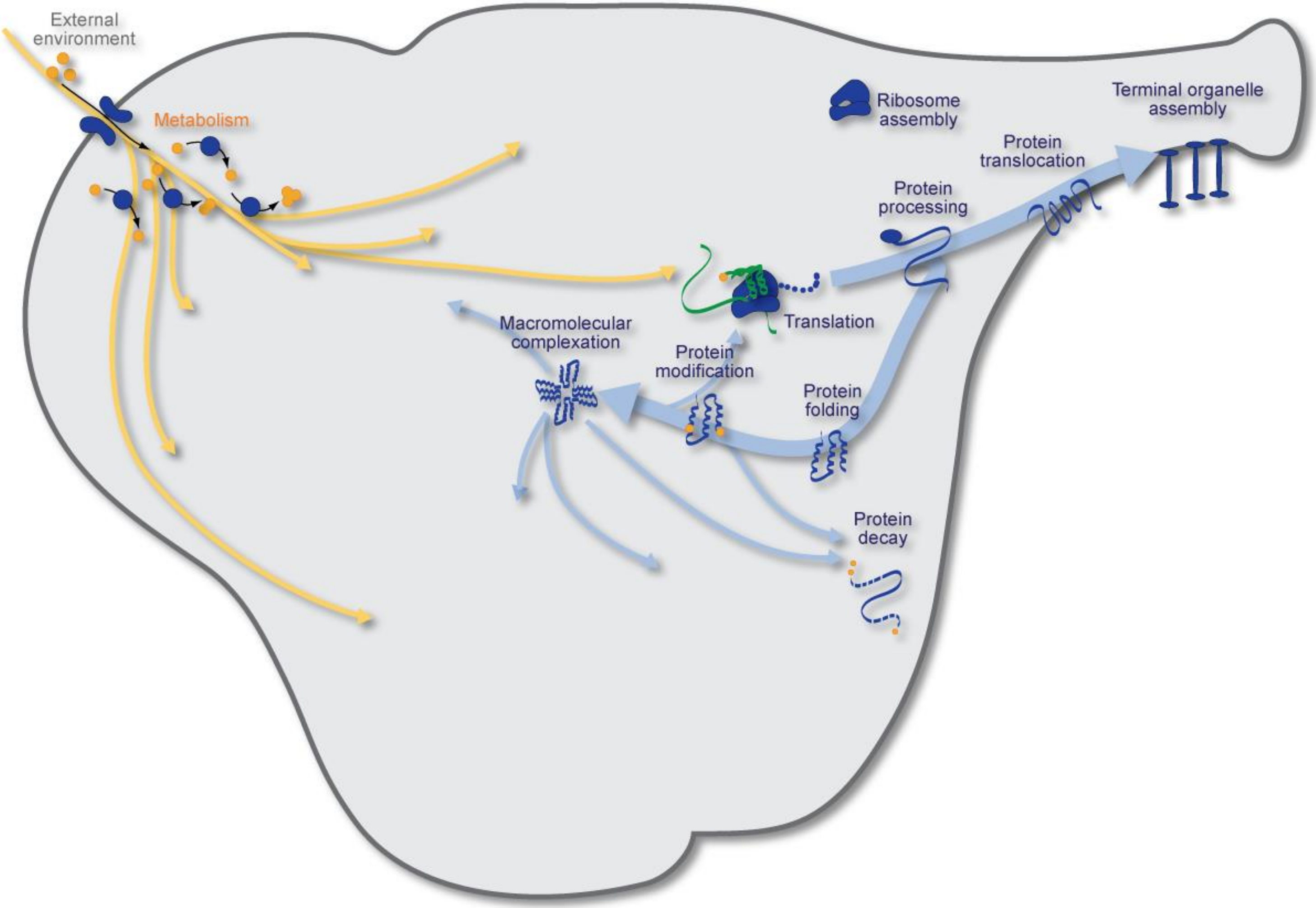
Example: drug biosynthesis



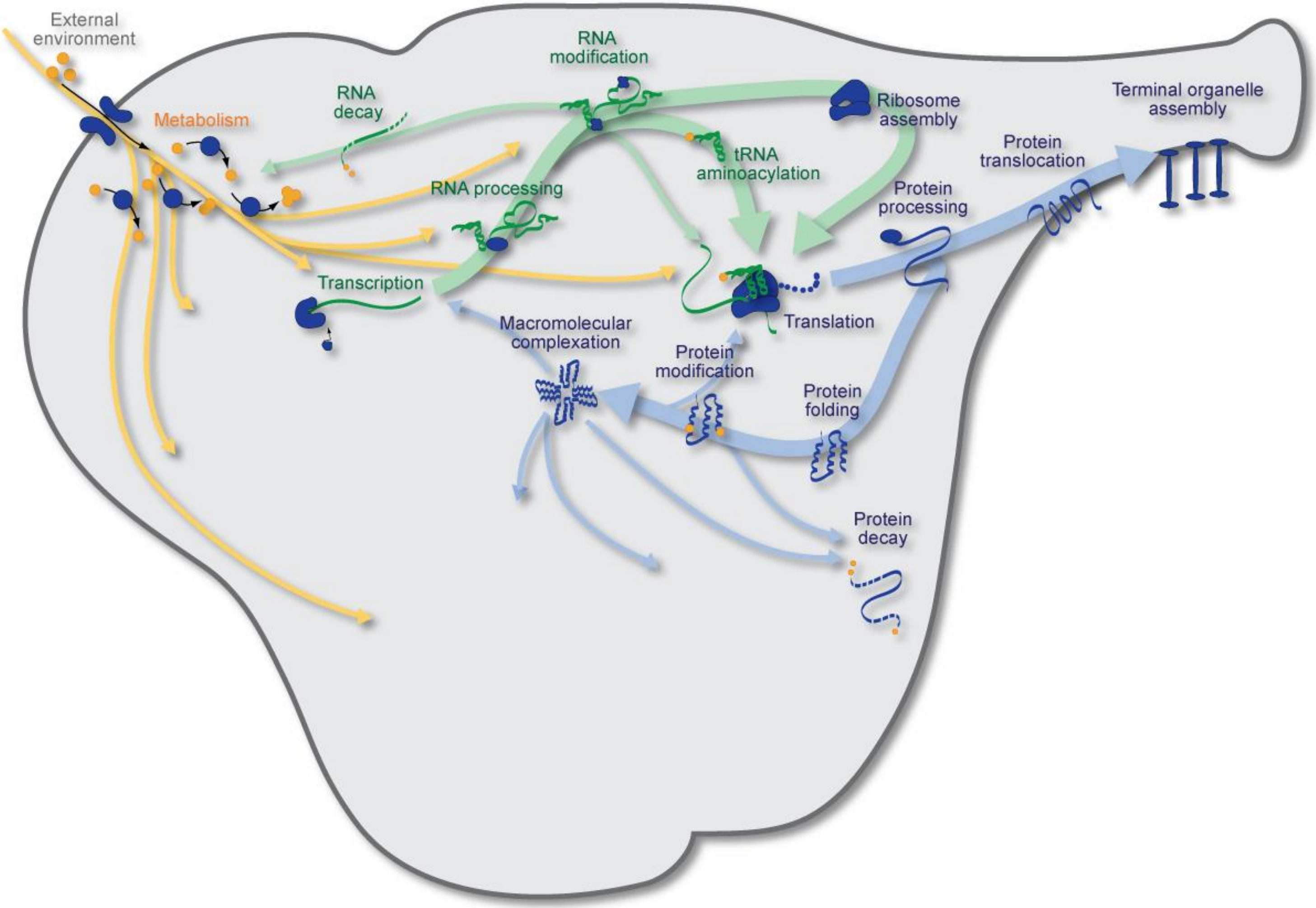
Example: drug biosynthesis



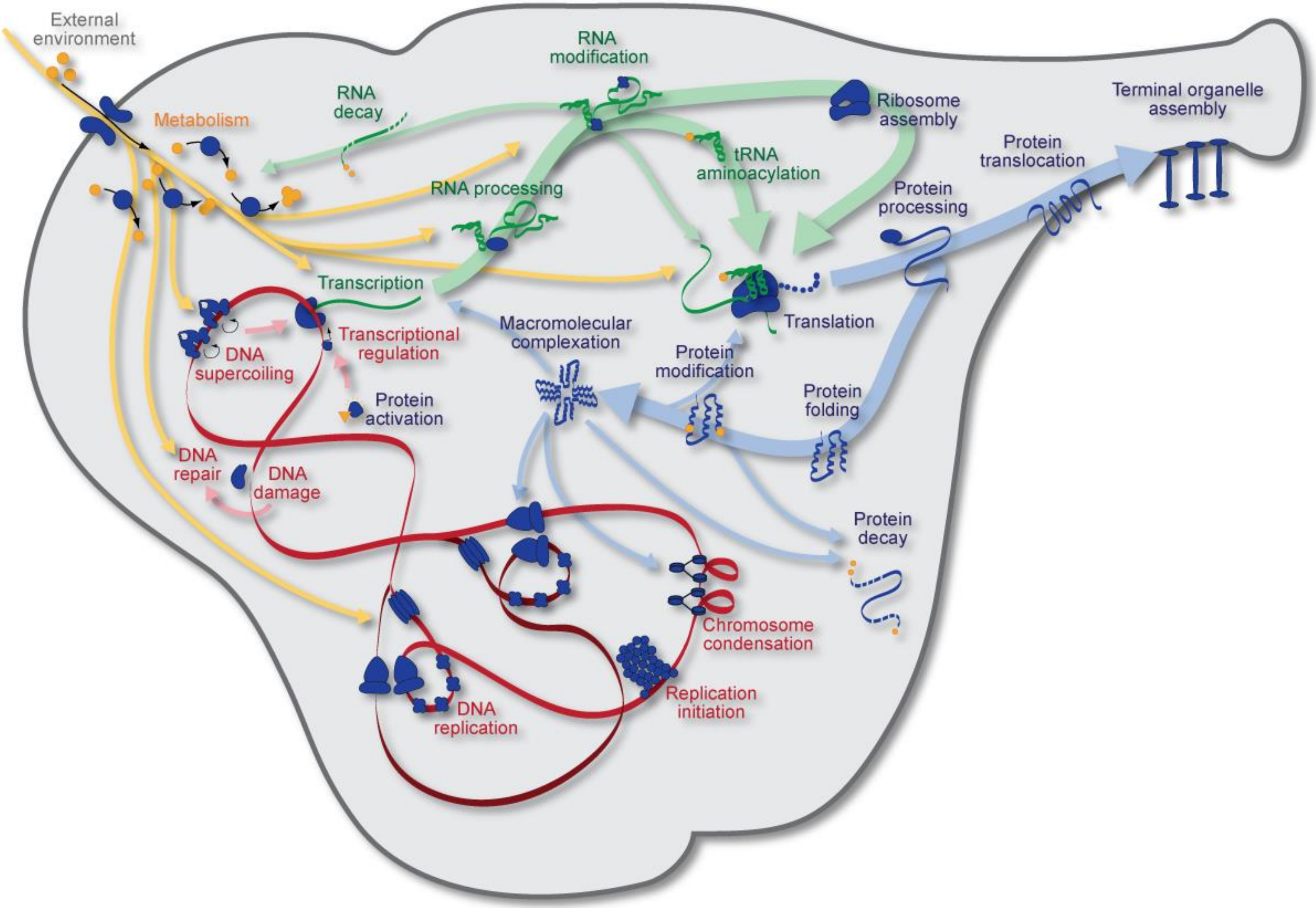
Example: drug biosynthesis



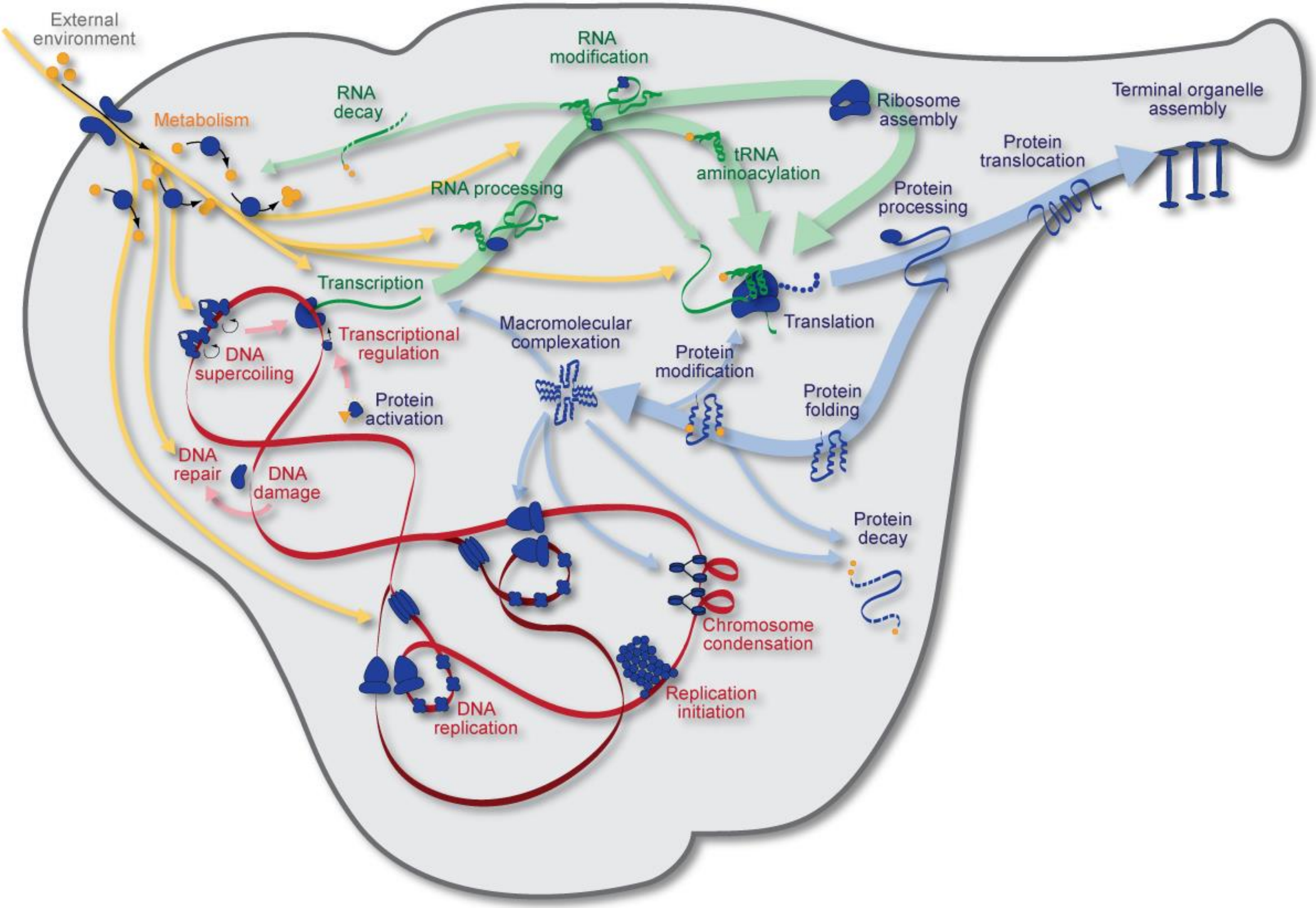
Example: drug biosynthesis



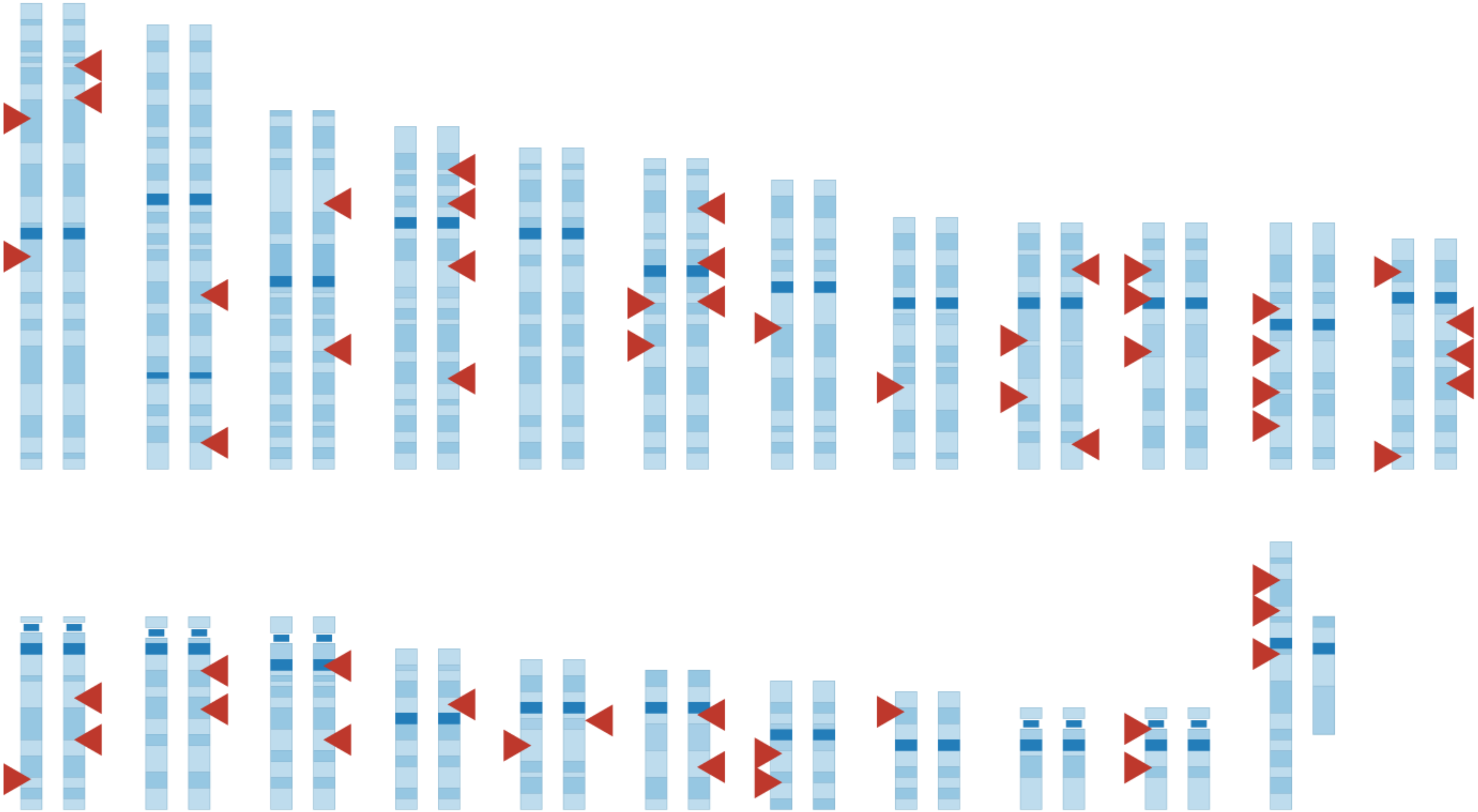
Example: drug biosynthesis



Example: drug biosynthesis



Precision medicine requires WC models

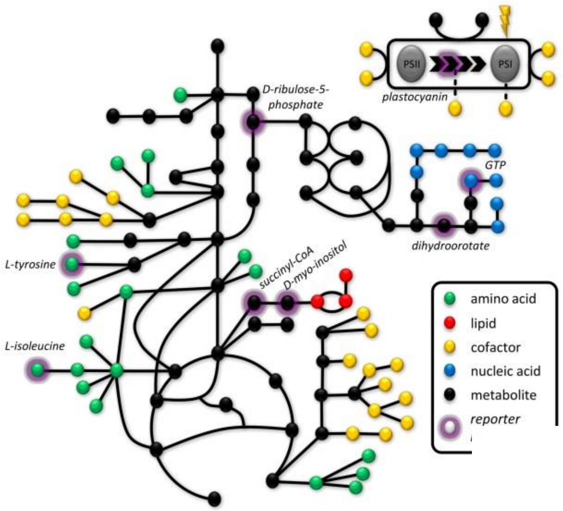


Challenges

Whole-Cell Modeling

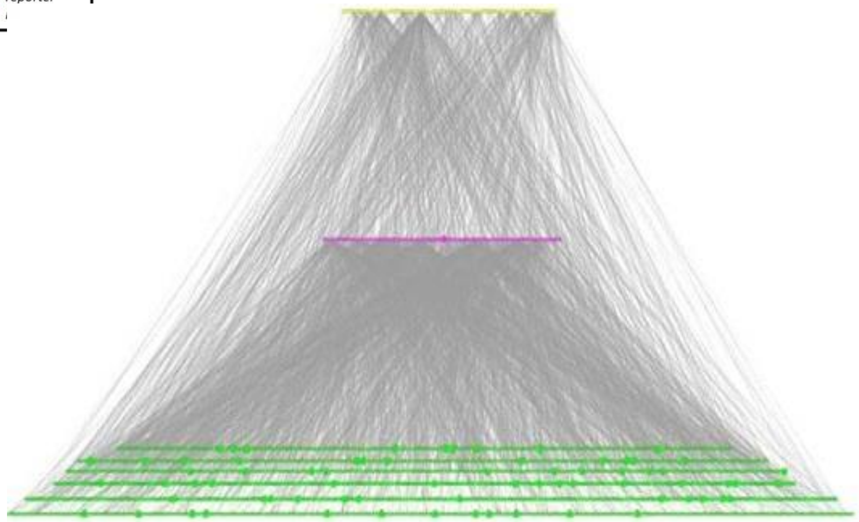
Mount Sinai

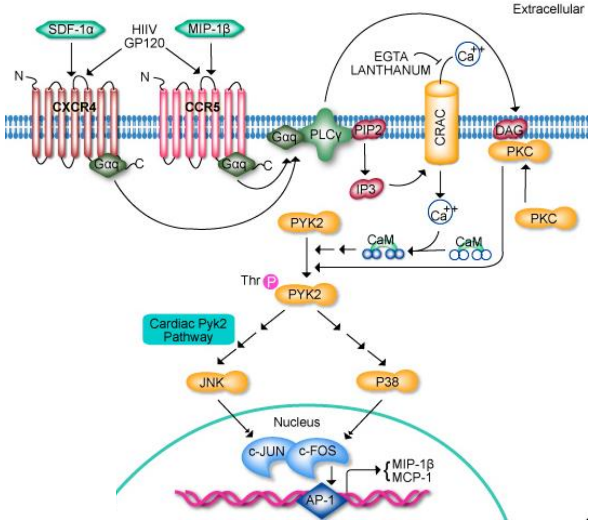
Challenge: explain diverse chemistry



Metabolism
FBA

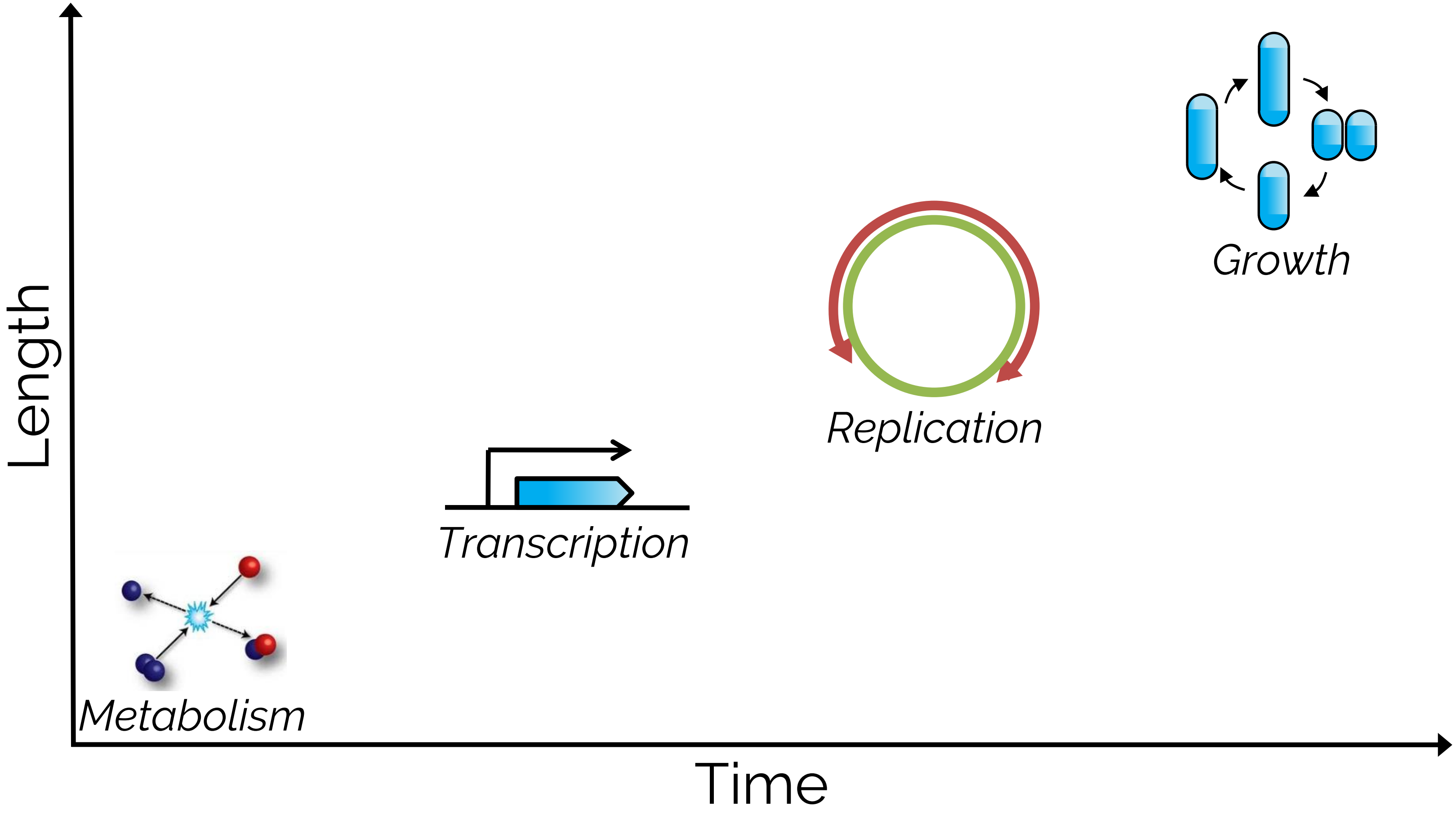
Transcriptional
regulation
Logical



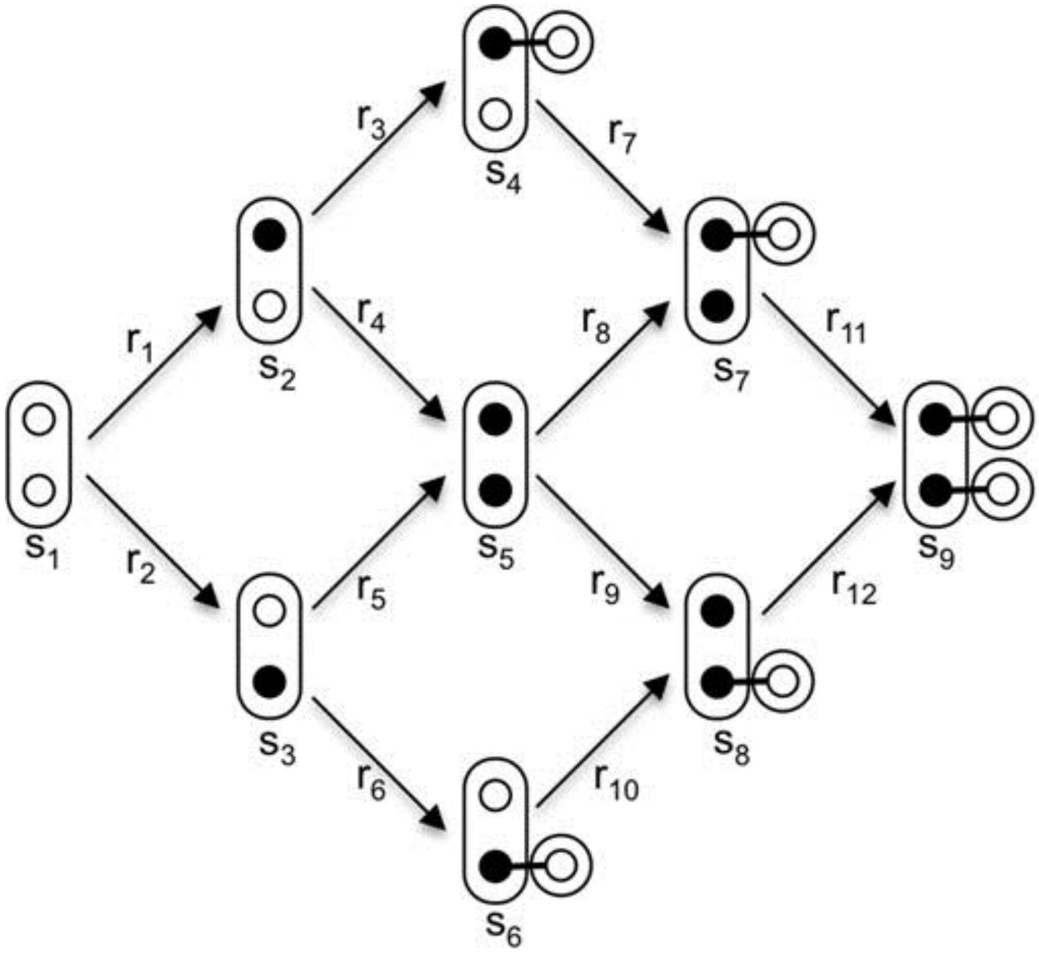


Signaling
ODE, SSA

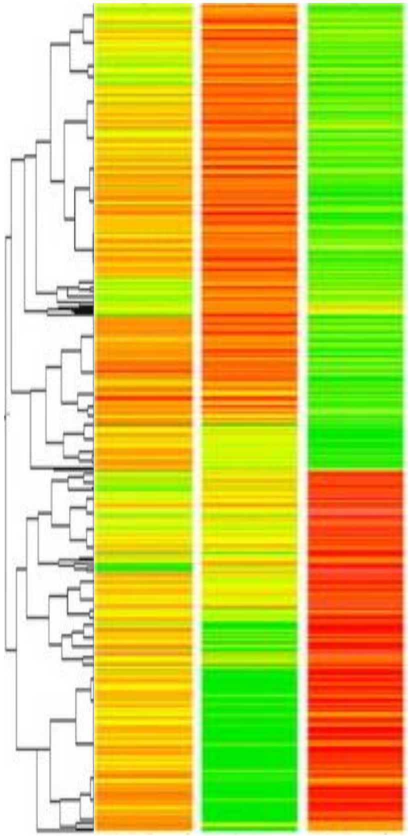
Challenge: explain multiple scales



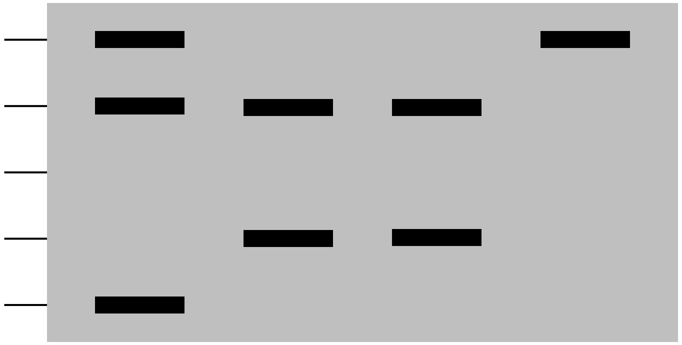
Challenge: capture chemical complexity



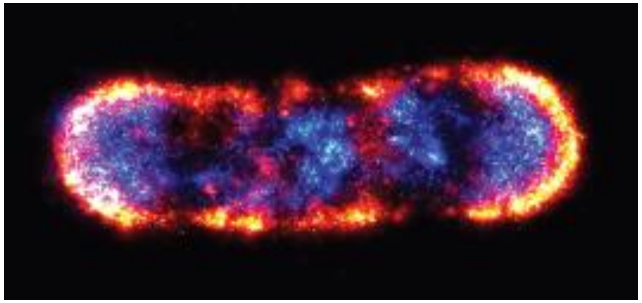
Challenge: heterogeneous data



Transcription
RNA-seq

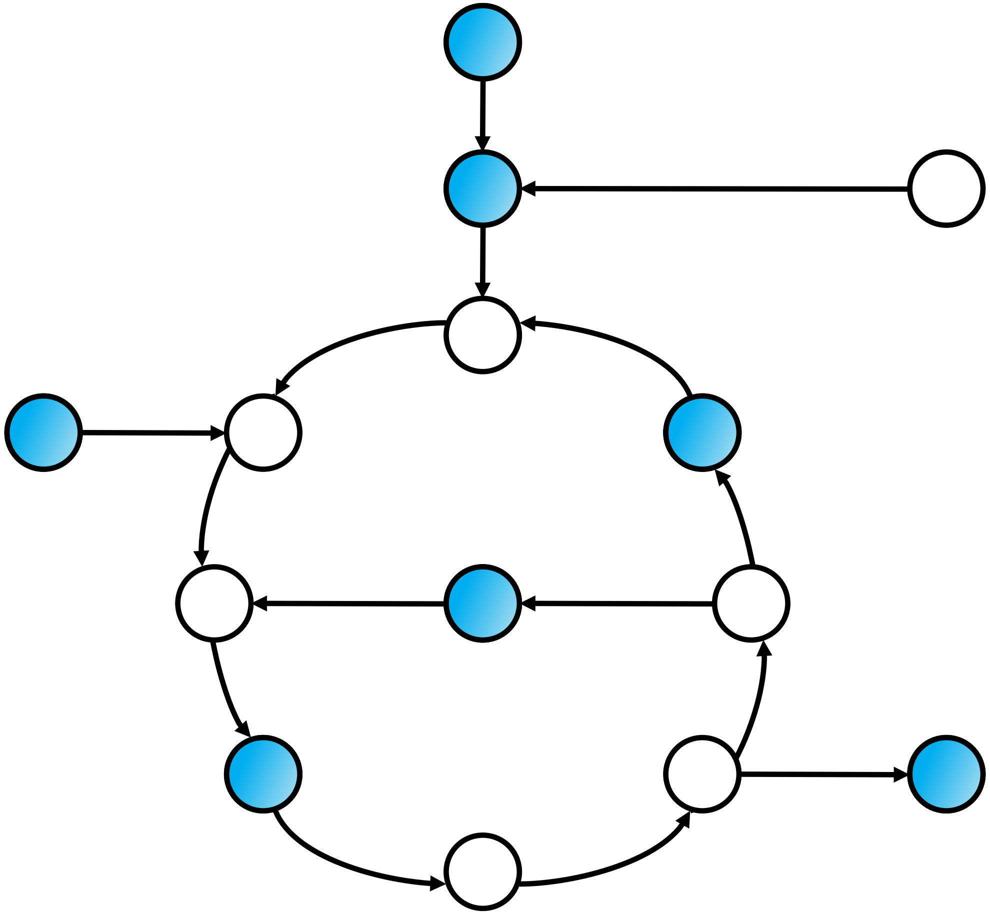


Protein expression
Mass-spec, Western blot



Single-cell variation
Microscopy

Challenge: incomplete data

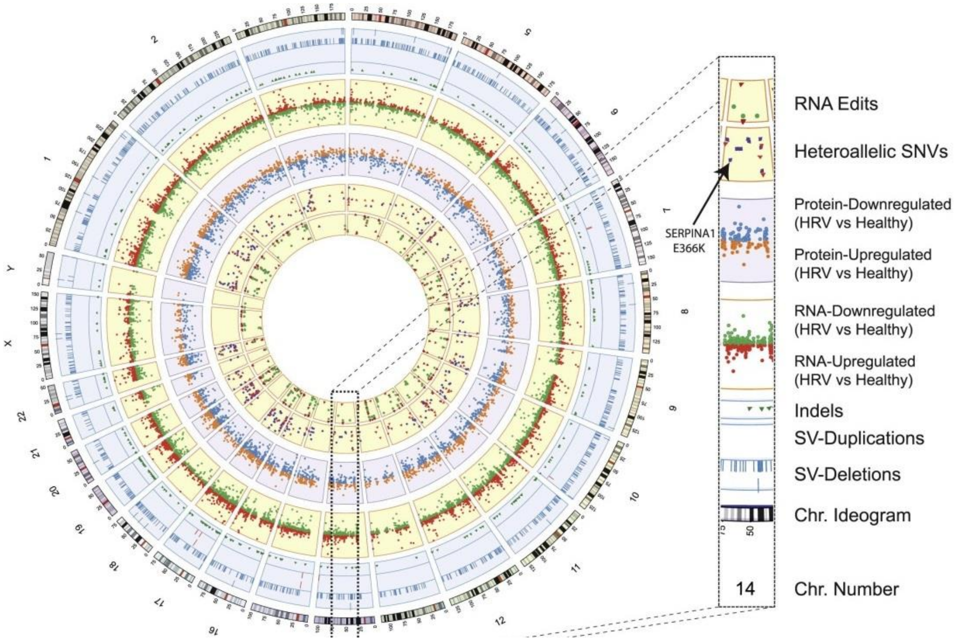


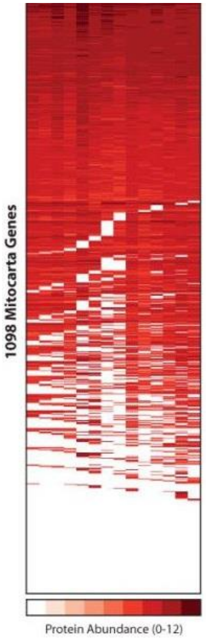
Feasibility

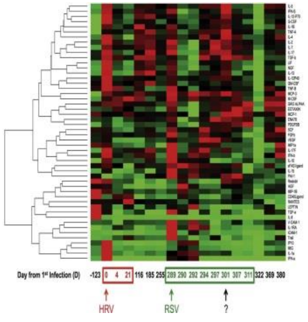
Whole-Cell Modeling

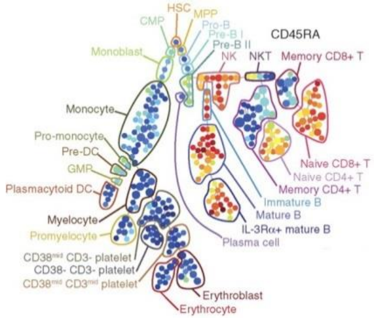
Mount Sinai

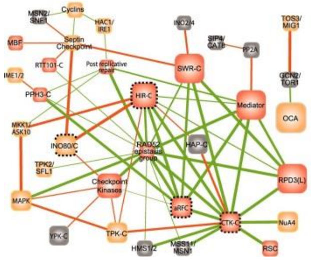
Feasibility: Extensive data

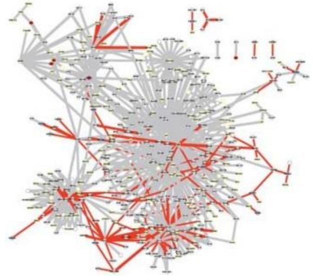




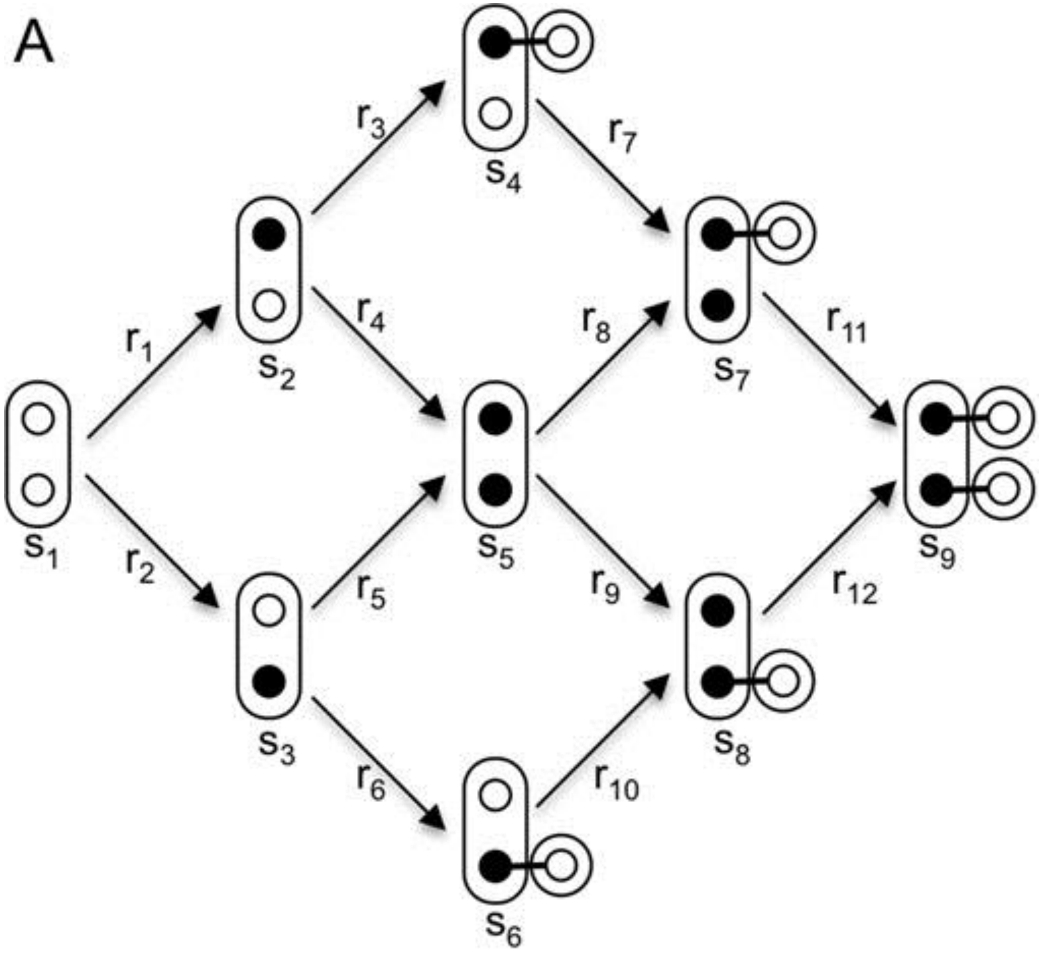




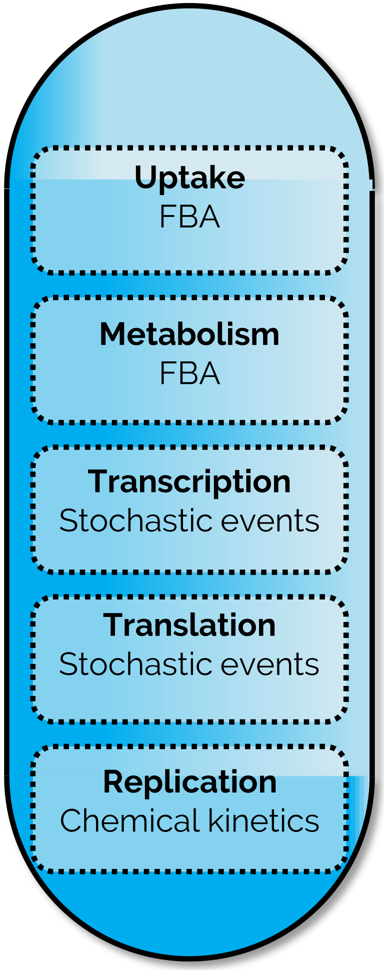




Feasibility: Rule-based modeling



Feasibility: Multi-algorithm simulation



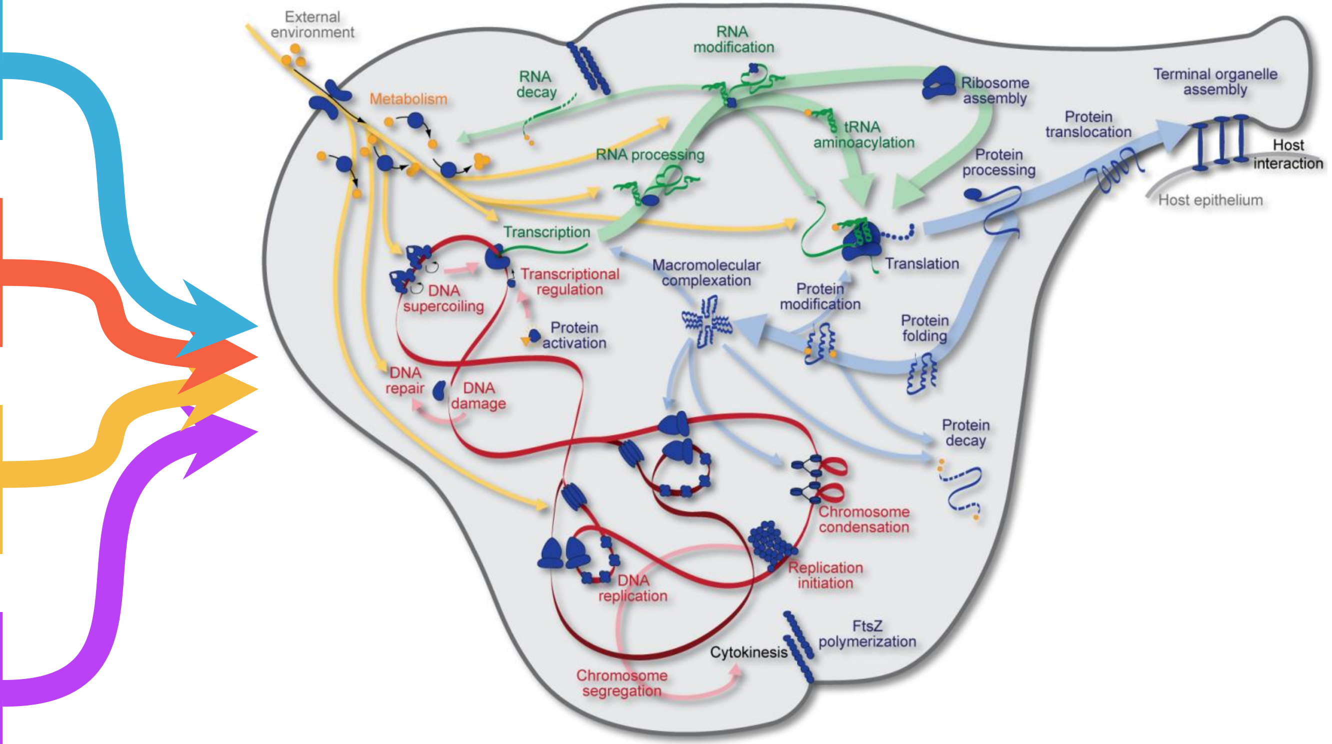
WC modeling is becoming feasible

Genomic and biochemical data

Pathway submodels

Rule-based modeling

Multi-algorithmic simulation

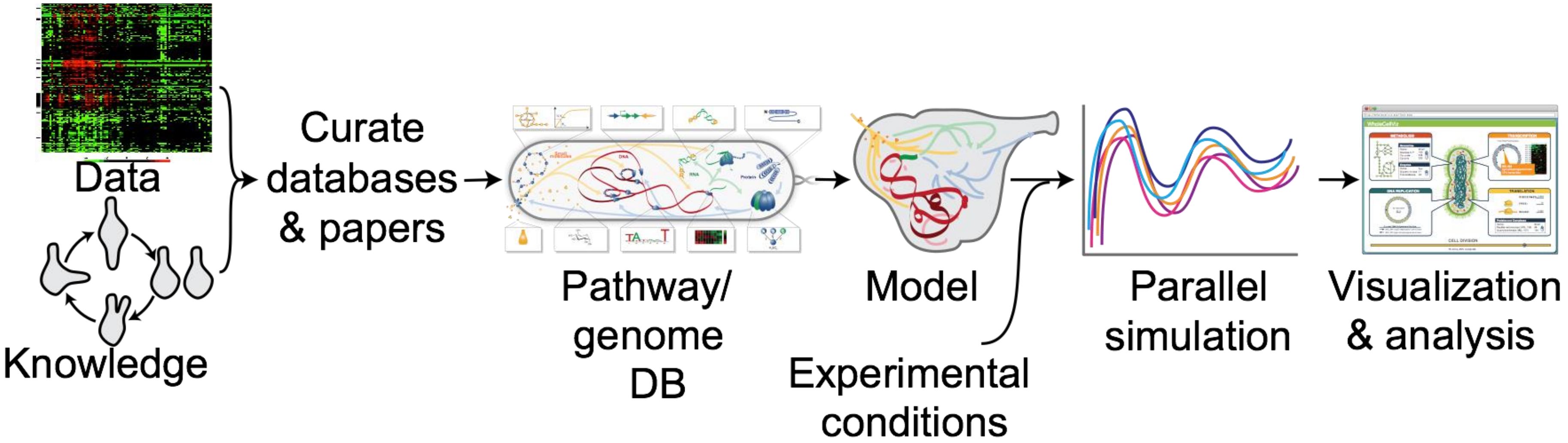


Workflow

Whole-Cell Modeling

Mount Sinai

Workflow



Data

Knowledge

Curate databases & papers

Pathway/ genome DB

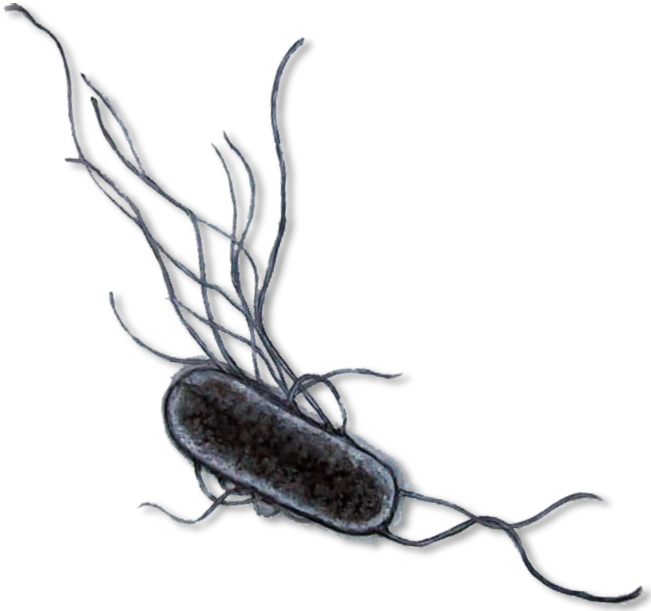
Model

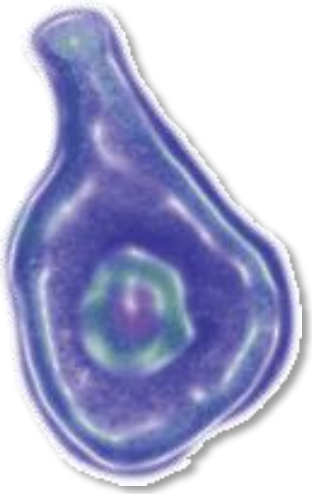
Experimental conditions

Parallel simulation

Visualization & analysis

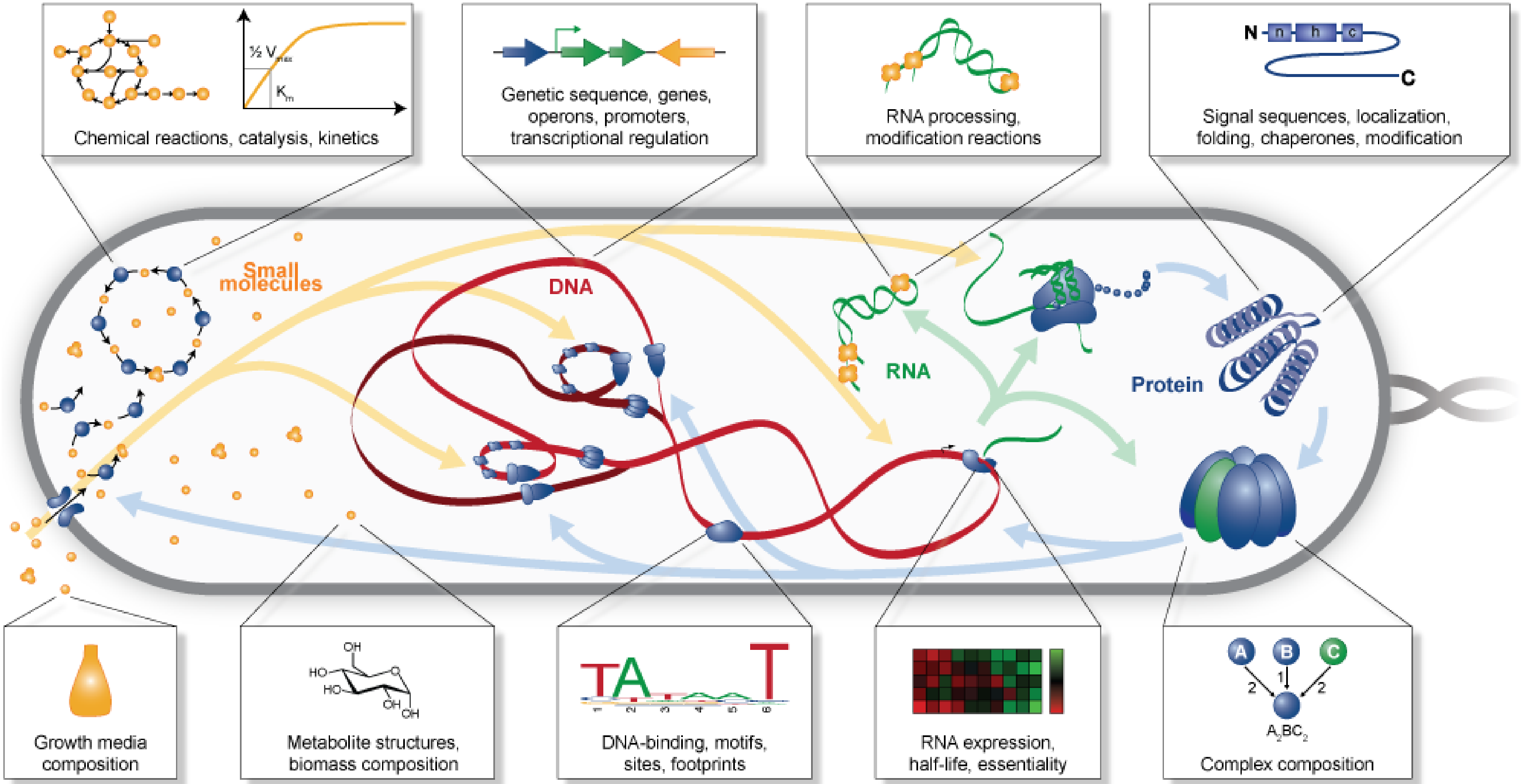
1. Focus on simple cells





	E. coli	*M. genitalium*
Genome	4700 kb	580 kb
Genes	4461	525
Size	2 µm × 0.5 µm	0.2-0.3 µm

2. Aggregate and integrate data



Chemical reactions, catalysis, kinetics

Genetic sequence, genes, operons, promoters, transcriptional regulation

RNA processing, modification reactions

Signal sequences, localization, folding, chaperones, modification

Growth media composition

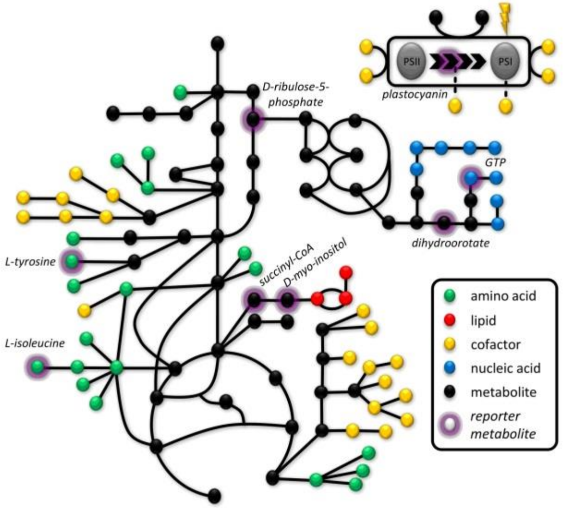
Metabolite structures, biomass composition

DNA-binding, motifs, sites, footprints

RNA expression, half-life, essentiality

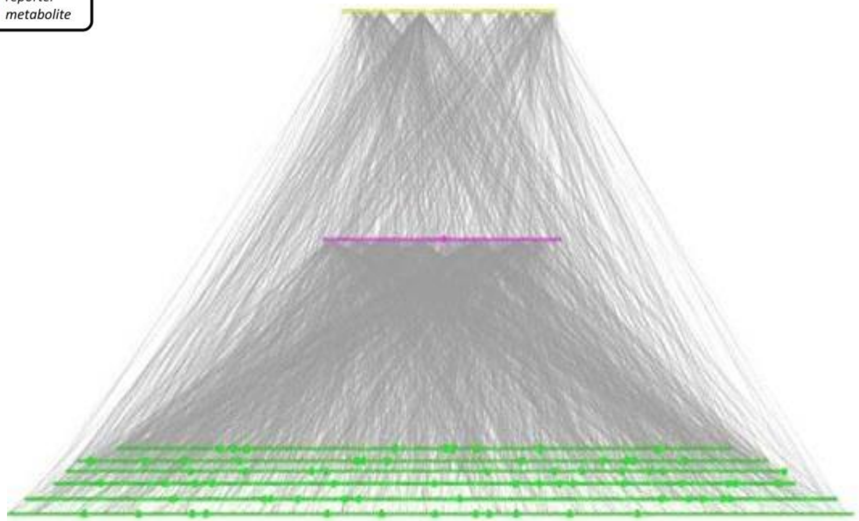
Complex composition

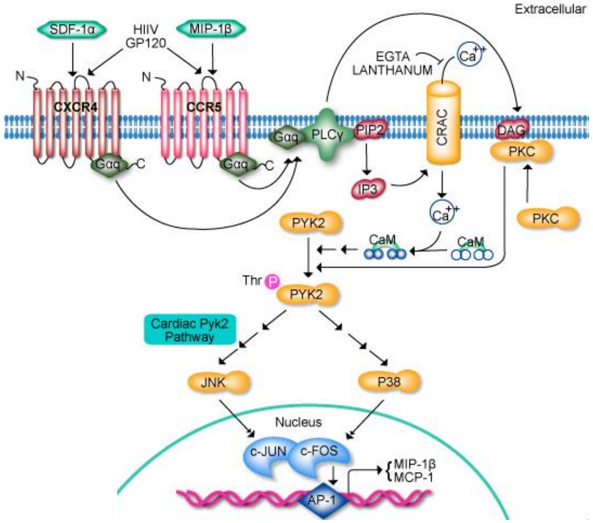
3. Model each process



Metabolism

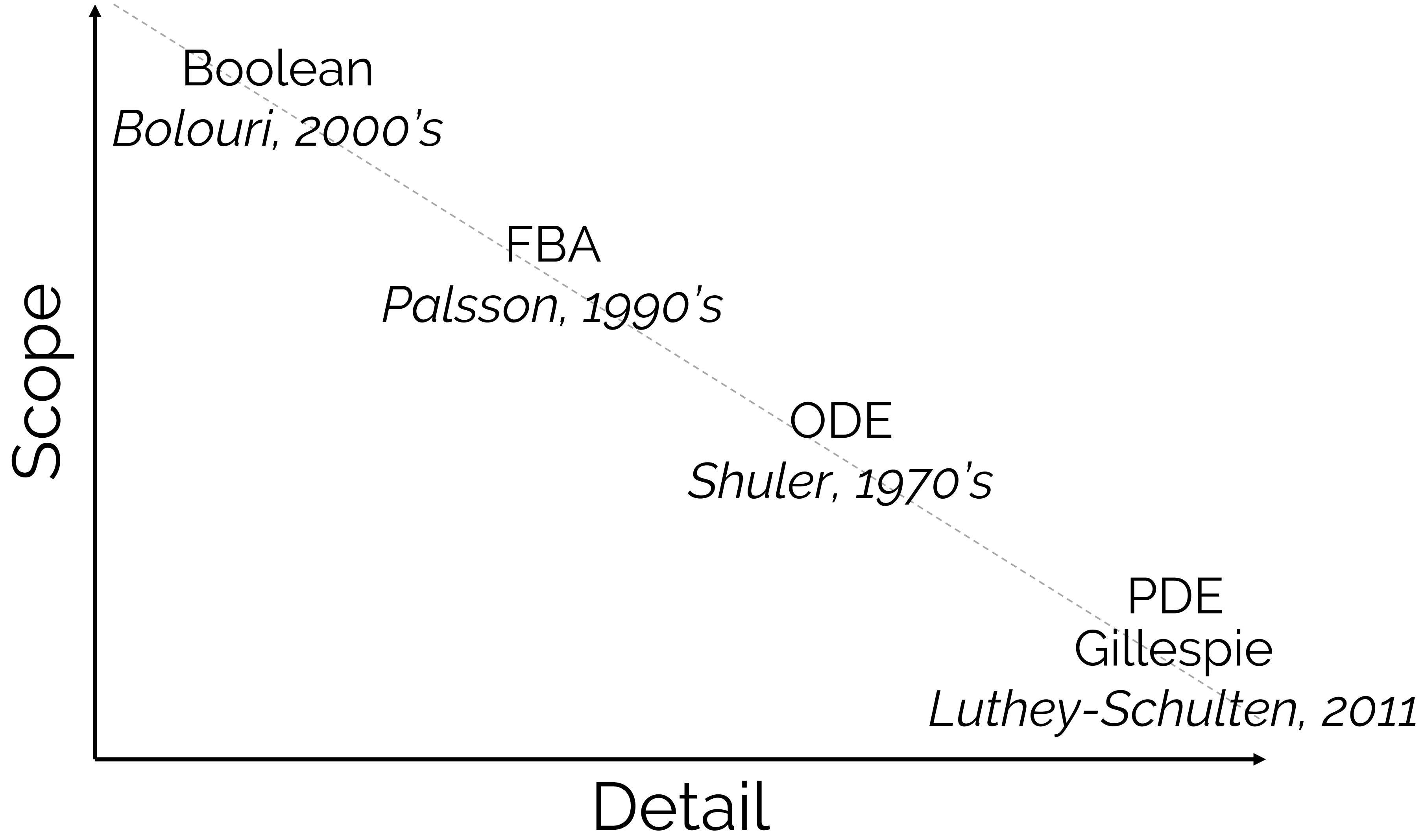
Transcriptional regulation





Signaling

3. Model each process



Boolean

Bolouri, 2000's

FBA

Palsson, 1990's

ODE

Shuler, 1970's

PDE

Gillespie

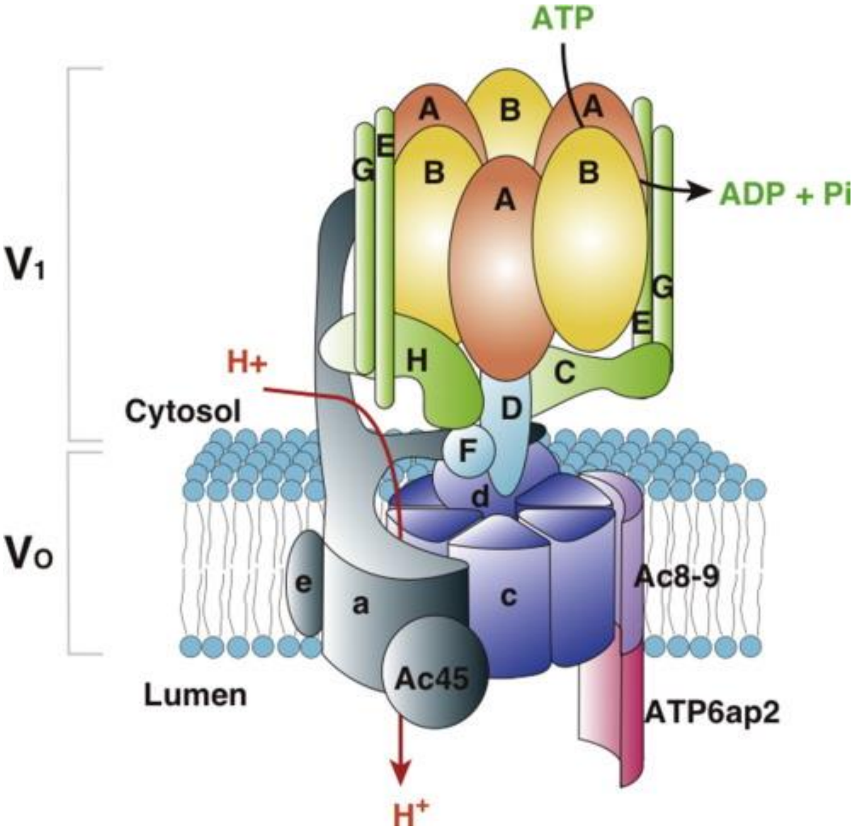
Luthey-Schulten, 2011

Scope

Detail

3. Model each process

Metabolism



Species and reactions

ADP + Pi + 4 H+[p]

↓ ATPase

ATP + H2O + 3 H+[c]

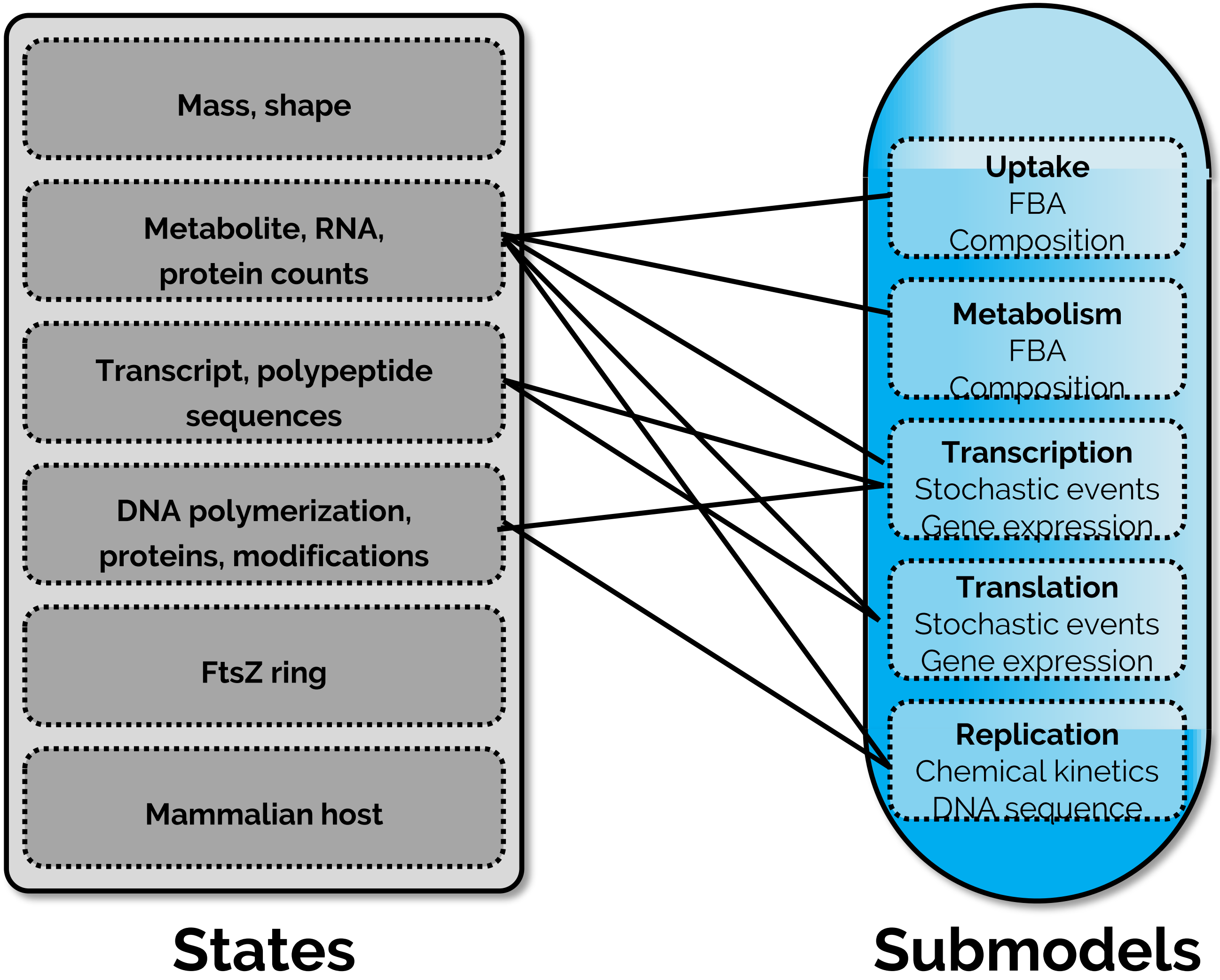
Catalysis

ATPase =	3 *	AtpA
	1 *	AtpB
	1 *	AtpC
	3 *	AtpD
	10 *	AtpE
	2 *	AtpF
	1 *	AtpG
	1 *	AtpH

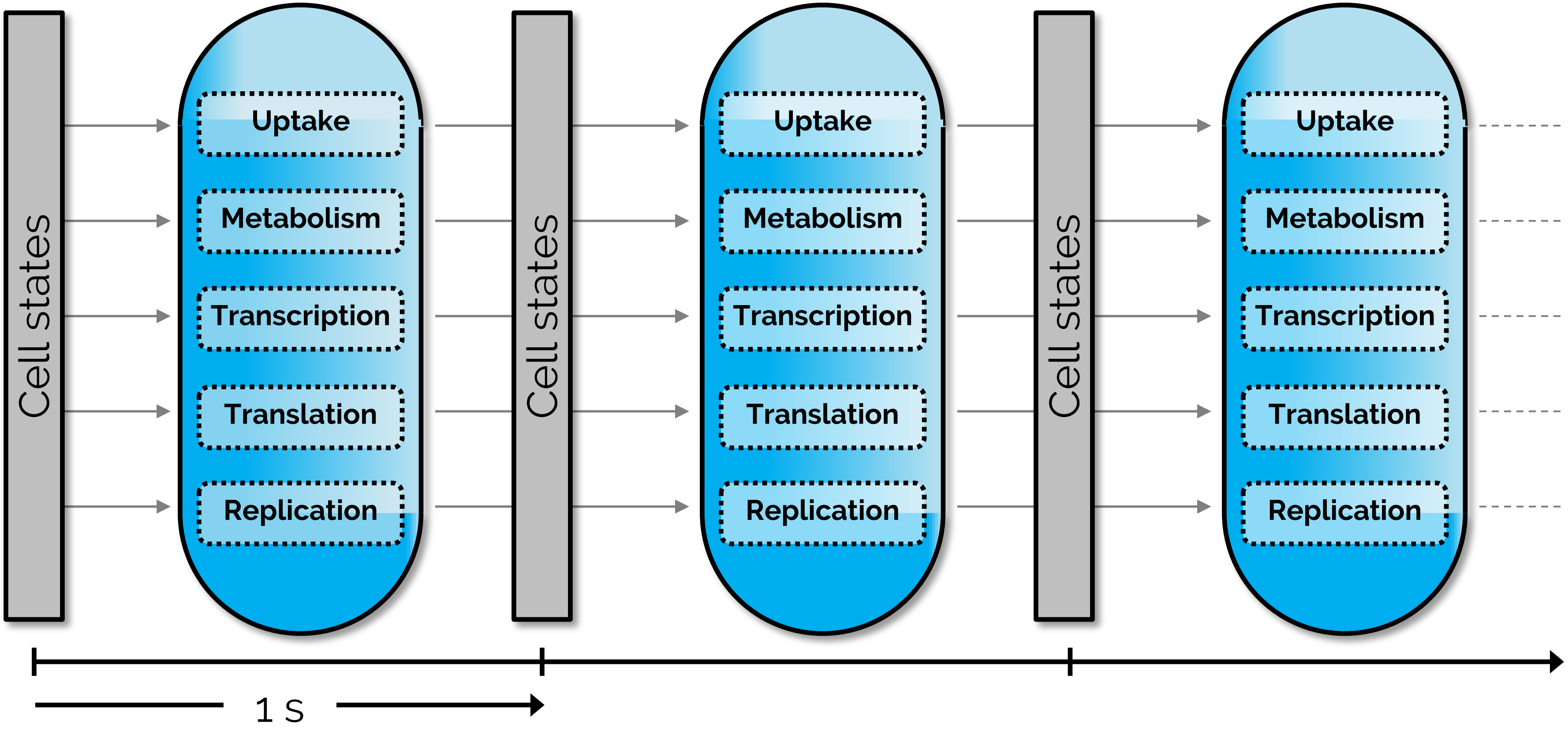
Kinetics

$$v = \mathrm{k_{cat}}[\mathrm{ATPase}]\frac{[\mathrm{ADP}]}{K_m + [\mathrm{ADP}]}$$

4. Merge models into a single model



5. Co-simulate models



6. Verify model

☑Matches training data

- ☑Cell mass, volume
- ☑Biomass composition
- ☑RNA, protein expression, half-lives
- ☑Superhelicity

☑Matches published data

- ☑Metabolite concentrations
- ☑DNA-bound protein density
- ☑Gene essentiality

☑Matches theory

- ☑Mass conservation
- ☑Central dogma
- ☑Cell theory
- ☑Evolution

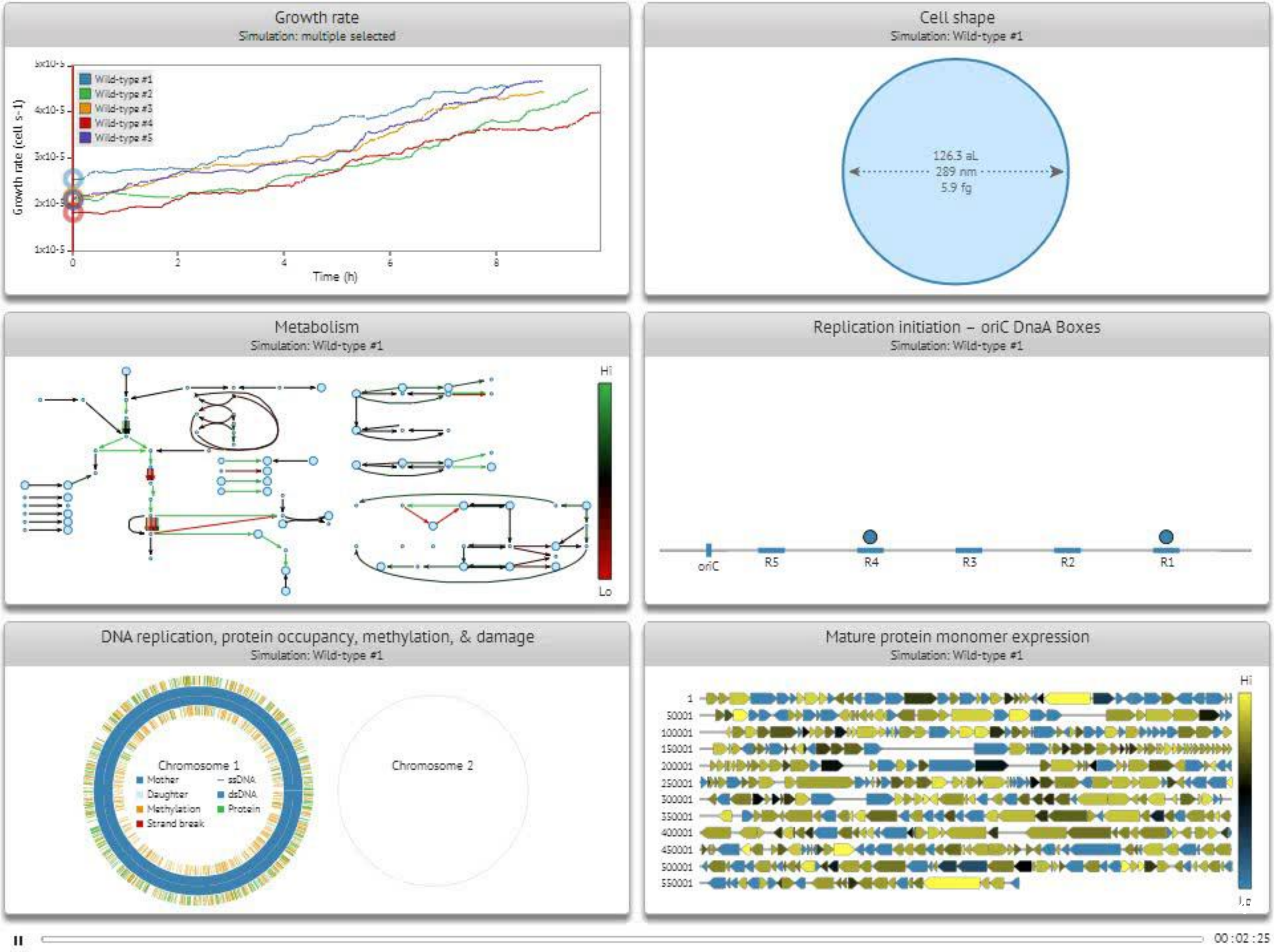
☑No obvious errors

- ☑Plot model predictions
- ☑Manually inspect data
- ☑Compare to known biology

State of the art

Whole-Cell Modeling

Mount Sinai



Growth rate

Simulation: multiple selected

Cell shape

Simulation: Wild-type #1

Metabolism

Simulation: Wild-type #1

Replication initiation – oriC DnaA Boxes

Simulation: Wild-type #1

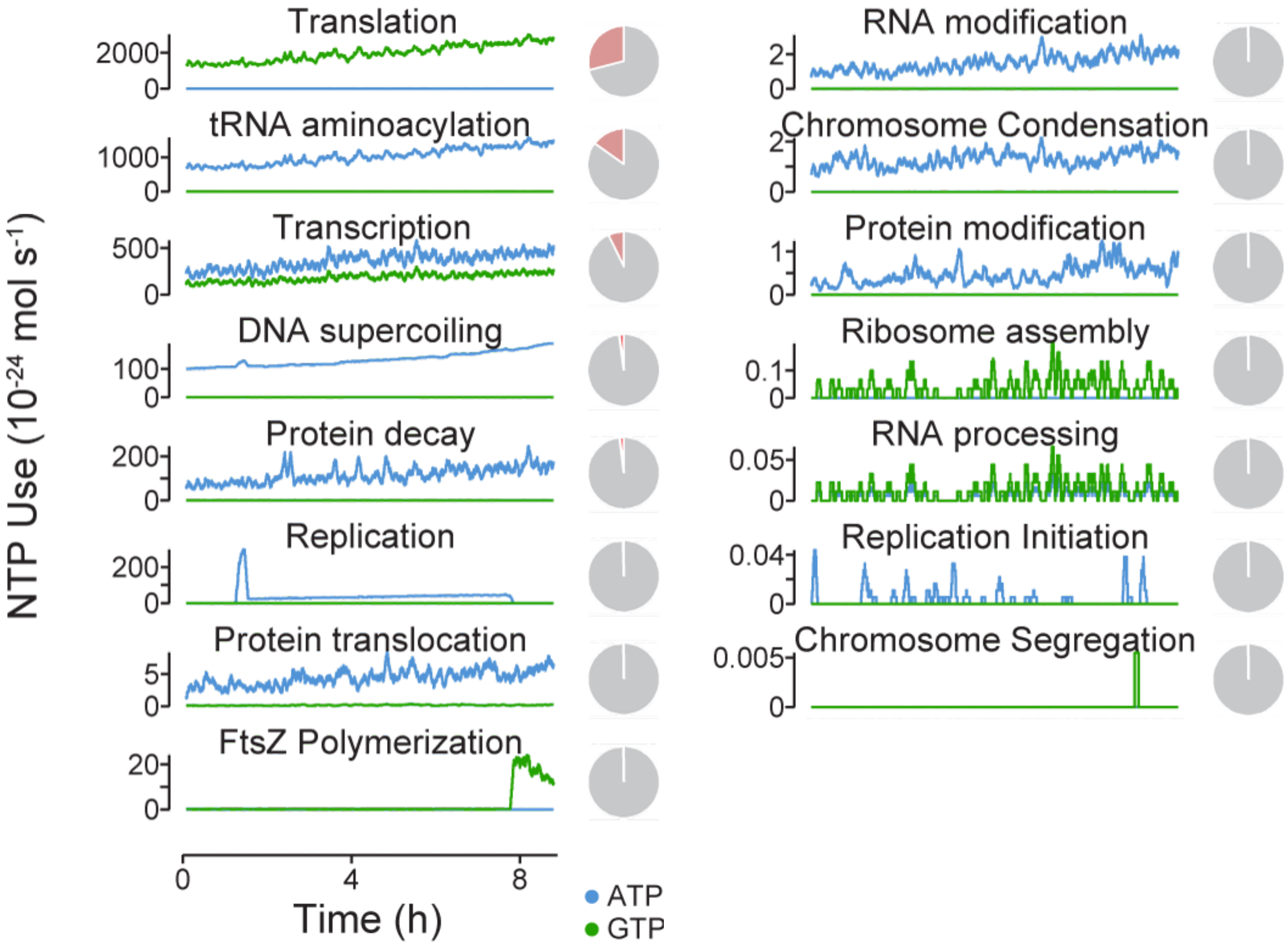
DNA replication, protein occupancy, methylation, & damage

Simulation: Wild-type #1

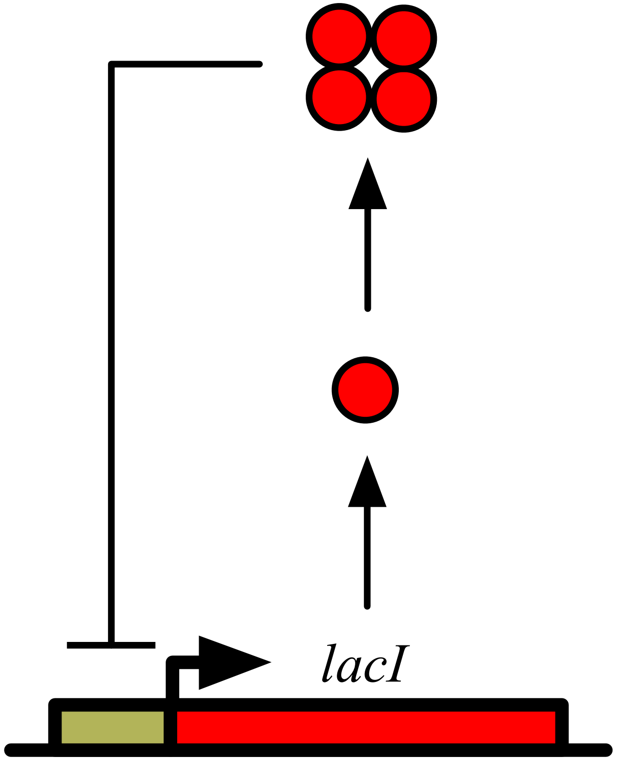
Mature protein monomer expression

Simulation: Wild-type #1

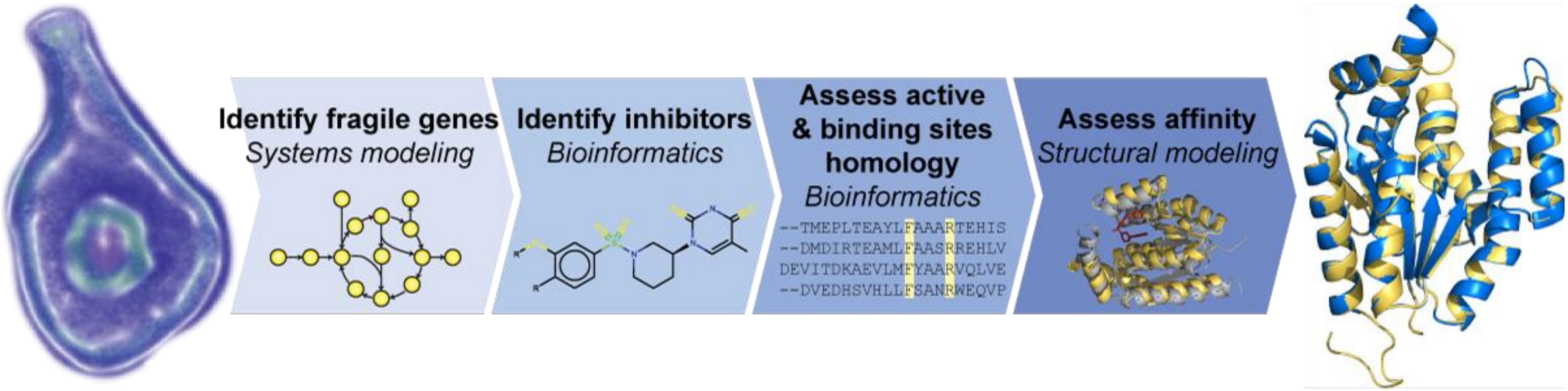
WC models provide novel insights



WC models help design cells



WC models help purpose drugs



M. genitalium

525 genes

Identify fragile genes

Systems modeling

Identify inhibitors

Bioinformatics

Assess active & binding sites homology

Bioinformatics

```
--TMEPLTEAYLFAAARTEHIS
--DMDIRTEAMLFAASRREHLV
DEVITDKAEVLMFYAARVQLVE
--DVEDHSVHLLFSANRWEQVP
```

Assess affinity

Structural modeling

Tmk

1 gene

Limitations of the *Mycoplasma* model

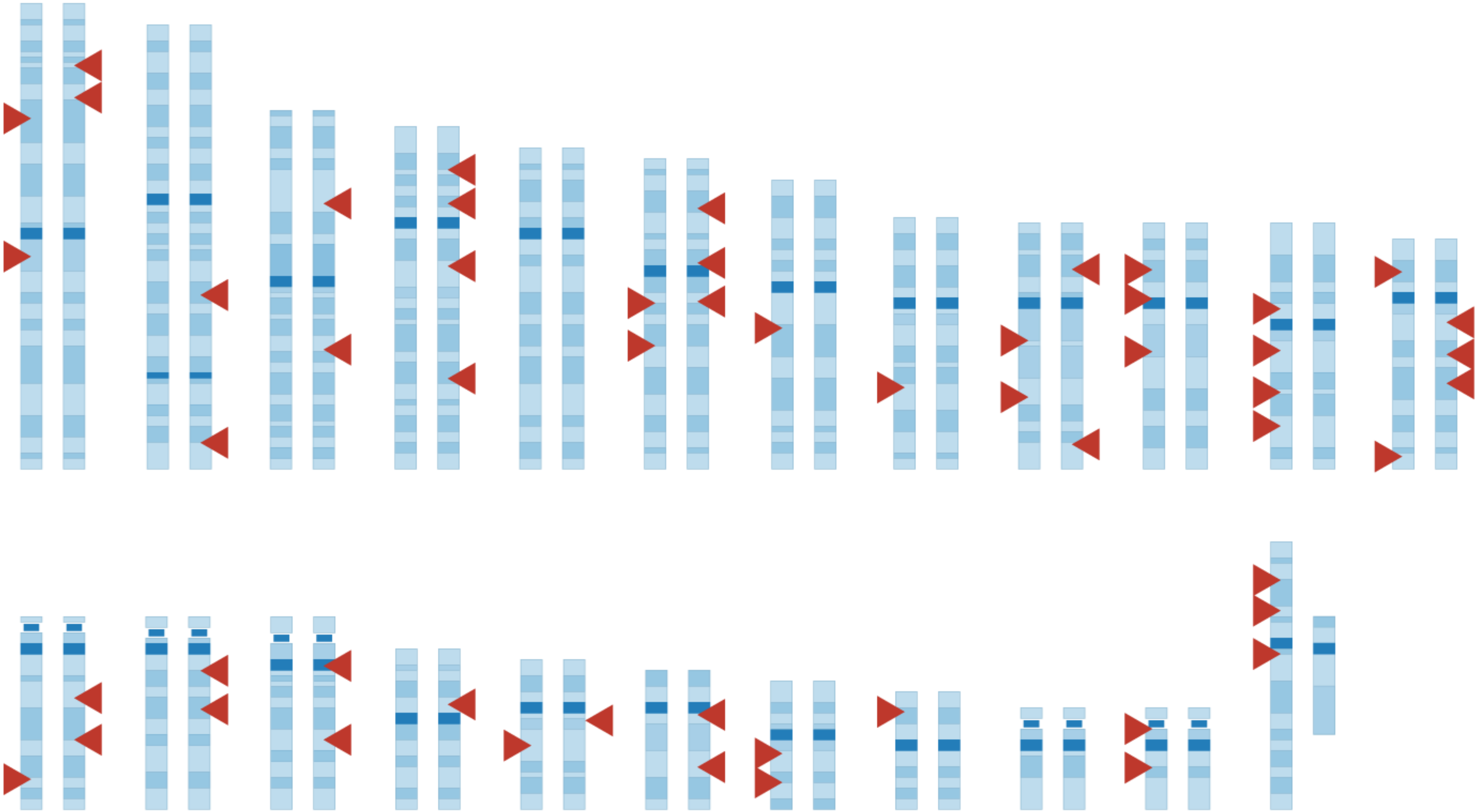
- Represents one of the smallest bacteria
- Ignores several processes
- Mispredicts several phenotypes
- Methods were ad hoc
- Hard to understand, reuse, and expand
- Time-consuming to build

Toward more comprehensive and more accurate models

Whole-Cell Modeling

Mount Sinai

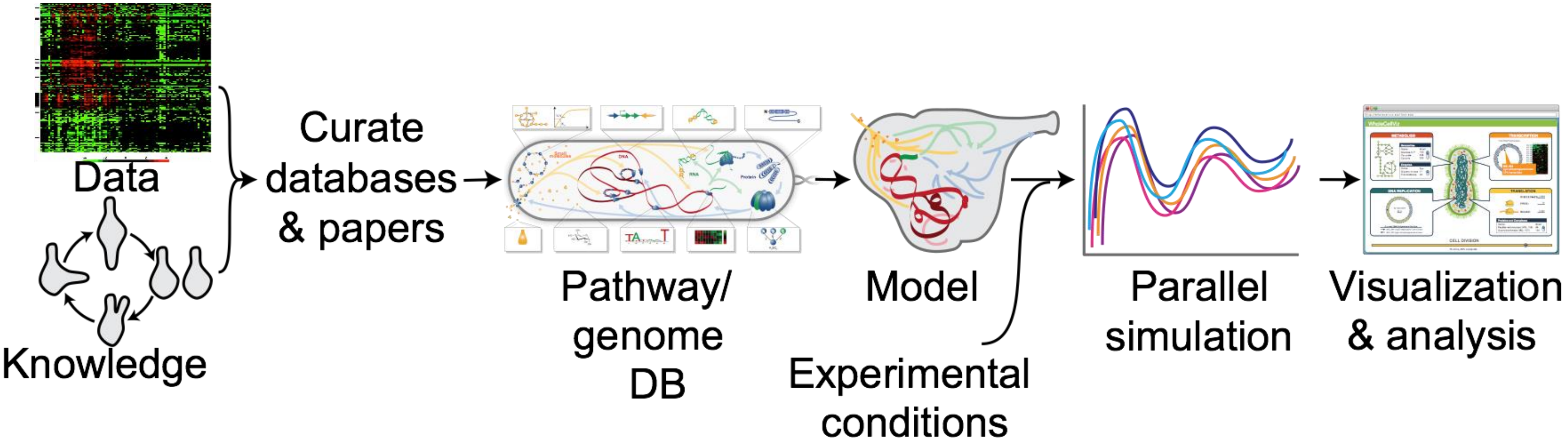
Goal: design precise therapy



Challenge: H1 hESCs

- Karyotypically normal
- Autonomous
- Well-characterized

Bottlenecks



Data

Knowledge

Curate databases & papers

Pathway/ genome DB

Model

Experimental conditions

Parallel simulation

Visualization & analysis

Bottlenecks

- **Data aggregation:** Hard to find relevant data
  - Data is incomplete, scattered, and insufficient annotated
- **Model design:** Hard to capture multiple scales and describe models modularly
  - Insufficient abstraction and metadata
- **Simulation**: Hard to simulate multiple scales
  - Simulators are only support individual formalisms and are slow
- **Verification:** Little formalism or standardization
- **Collaboration:** Difficult to describe the data, assumptions, and decisions that underlie modeling

Data needed for WC modeling

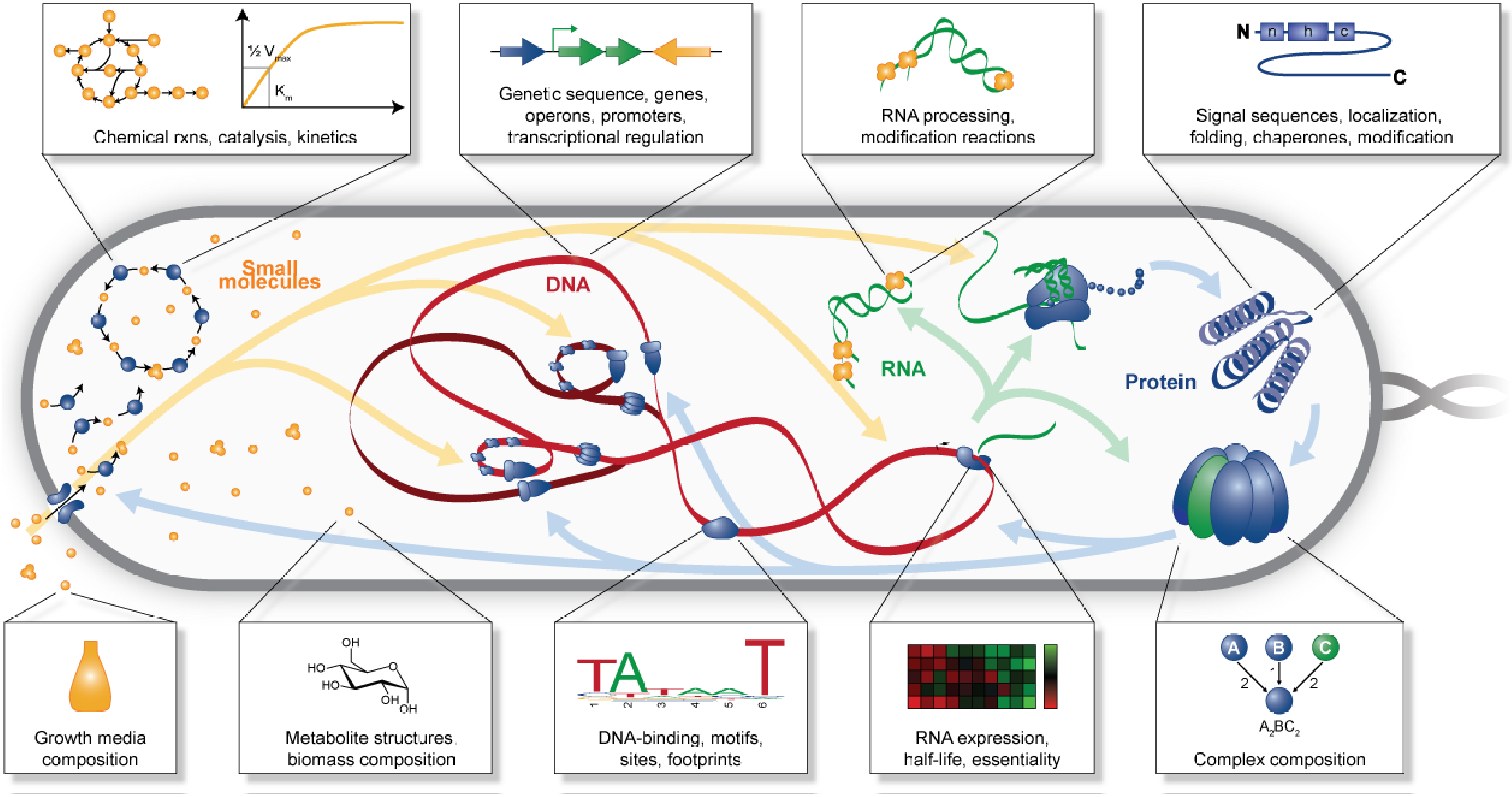
$$v = k_{\mathrm{cat}}\,[\mathrm{enzyme}]\,\frac{[\mathrm{substrate}]}{[\mathrm{substrate}] + K_{\mathrm{m}}}$$

Metabolite concentrations

Enzyme concentrations

Reaction kinetics

Data needed for WC modeling



Chemical rxns, catalysis, kinetics

Genetic sequence, genes, operons, promoters, transcriptional regulation

RNA processing, modification reactions

Signal sequences, localization, folding, chaperones, modification

Growth media composition

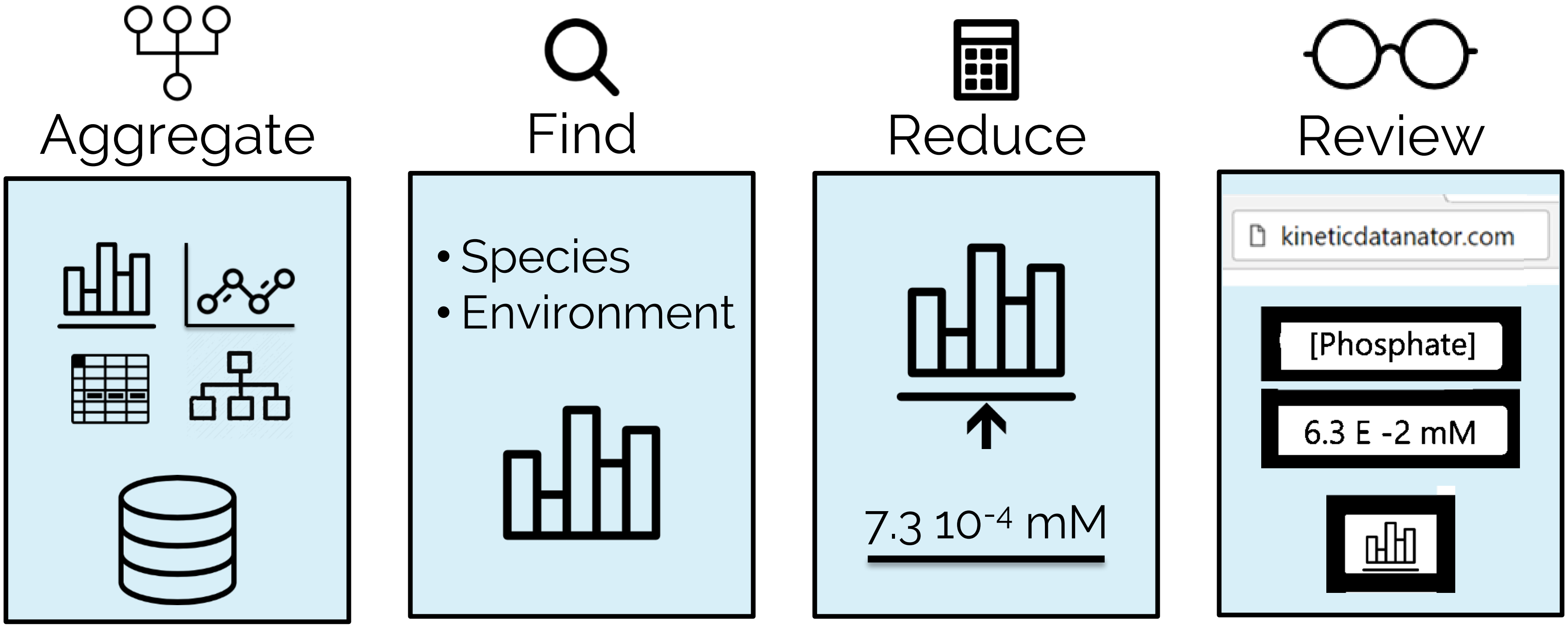
Metabolite structures, biomass composition

DNA-binding, motifs, sites, footprints

RNA expression, half-life, essentiality

Complex composition

Datanator: data integration & discovery



Aggregate

Find

- Species
- Environment

Reduce

$7.3\ 10^{-4}$ mM

Review

kineticdatanator.com

[Phosphate]

6.3 E -2 mM

Datanator: data aggregation

Metabolites

- ChEBI
- ECMDB, YMDB
- PubChem

DNA

- GenBank

RNA

- Array Express
- MODOMICS
- RNALocate
- RNA MOD

Protein

- COMPARTMENTS
- CORUM
- Human Protein Ref. DB
- Pax-DB
- PDB
- PSORTdb
- RESID
- UniProt

Interactions

- BioCyc
- DBTBS
- DrugBank
- JASPAR
- KEGG
- SuperTarget

Taxonomy

- NCBI

Pathways

- KEGG
- Pathway Commons
- Reactome
- WikiPathways

Rates

- BRENDA
- SABIO-RK

Datanator: actionable metadata

Measured entity/property

Measured value, uncertainty, units

Genotype

- Taxon
- Genetic variant
- Cell, tissue type

Environment

- Temperature
- pH
- Growth media

Data generation process

- Experimental design
- Measurement method

Data analysis process

- Software
- Version

Metadata

- Authors
- Curator
- Date
- Citation

Datanator: Finding relevant data

Chemical similarity

- Tanimoto index
- Sequence similarity

Genetic similarity

- Whole-genome similarity
- Taxonomic distance

Environmental similarity

- Temperature
- pH

WC-Lang: scalable model descriptions

- Concretely describe composite multi-algorithmic models
- Concrete descriptions of every model element
- Capture data and assumptions underlying models
- Explicit descriptions of mixed granularity / lumping
- Structured description of initial conditions
- User interfaces suited to large models

WC-Lang: scalable model descriptions

RNA(i, 0) + NTP(i, 1) → RNA(i, 1) + PPi

RNA(i, 1) + NTP(i, 2) → RNA(i, 2) + PPi

RNA(i, 2) + NTP(i, 3) → RNA(i, 3) + PPi

RNA(i, 3) + NTP(i, 4) → RNA(i, 4) + PPi

…

WC-Lang: scalable model descriptions

$$\mathrm{RNA}(i, l) + \mathrm{NTP}(i, l+1) \rightarrow \mathrm{RNA}(i, l+1) + \mathrm{PPi}$$

$$\mathrm{RNA}(I, l) + H_2O \rightarrow \mathrm{RNA}(i, l-1) + \mathrm{NMP}(i, l)$$

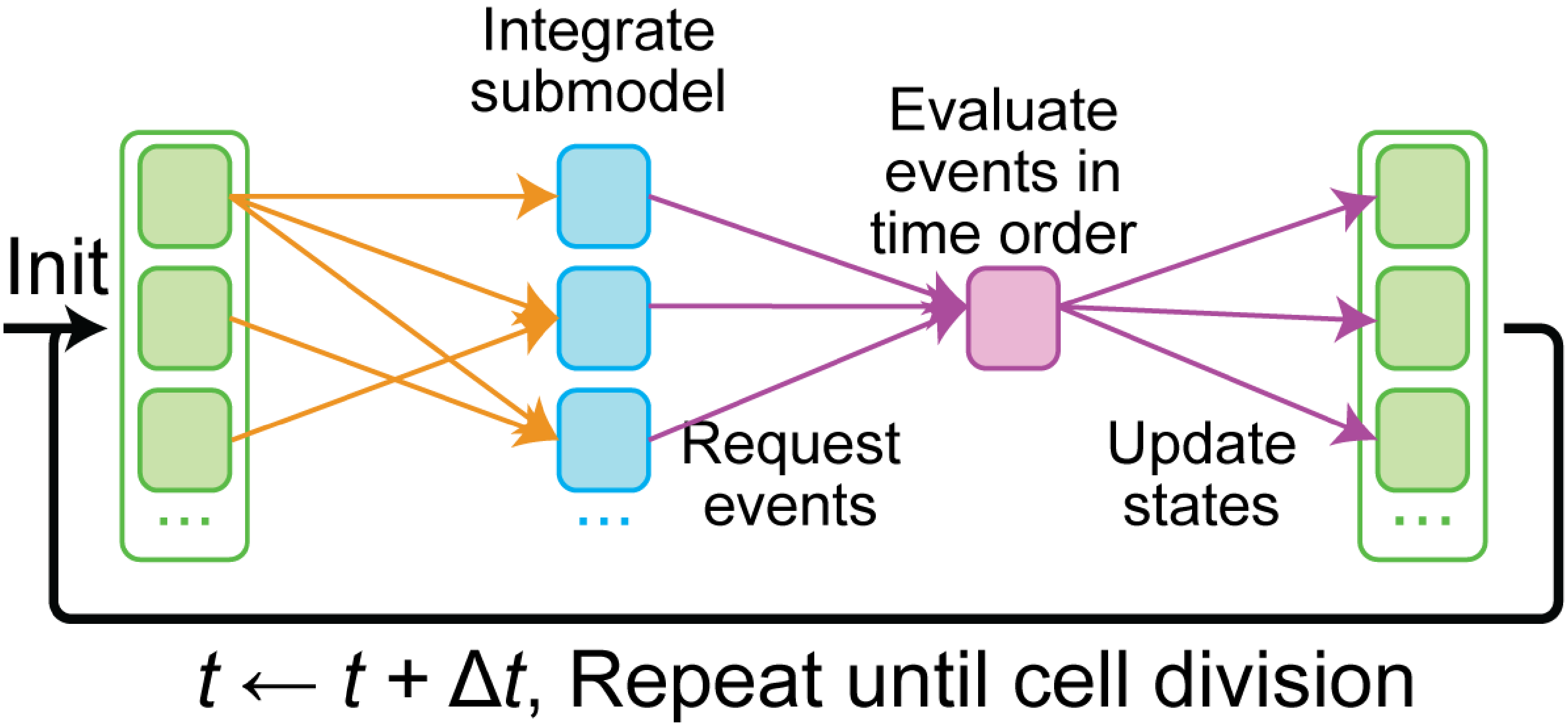
$$\mathrm{Protein}(i, l) + \mathrm{AA}(i, l+1) \rightarrow \mathrm{Protein}(i, l+1) + H_2O$$

$$\mathrm{Protein}(i, l) + H_2O \rightarrow \mathrm{Protein}(i, l-1) + \mathrm{AA}(i, l)$$

WC-Lang: scalable model descriptions

	Initiation		Elongation		Termination	
SBML	1 per RNA	335	1 per base	~500k	1 per RNA	335
Rules	1	1	1	1	1	1

WC-Sim: scalable co-simulation



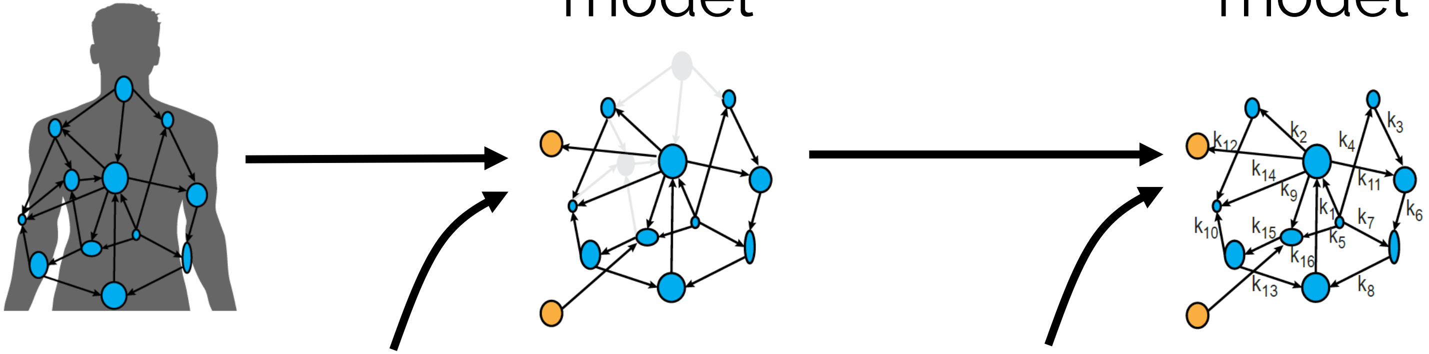
$t \leftarrow t + \Delta t$, Repeat until cell division

H1-hESC model

Recon 2.2

H1 model

Composable model



- H1 transcriptomics data (ENCODE)
- Cell composition
- Media composition

- Kinetic data (SABIO-RK)
- Protein abundance (Phanstiel et al., 2011; PaxDB)

Summary

Whole-Cell Modeling

Mount Sinai

Availability

- **Code:** code.karrlab.org (GitHub, PyPI)
- **Data:** data.karrlab.org (Quilt)
- **Images:** DockerHub
- **Primer and docs:** docs.karrlab.org
- **Tutorials:** sandbox.karrlab.org

Summary

Bioengineering and medicine needs WC models

WC modeling is becoming feasible

New technologies will enable WC modeling

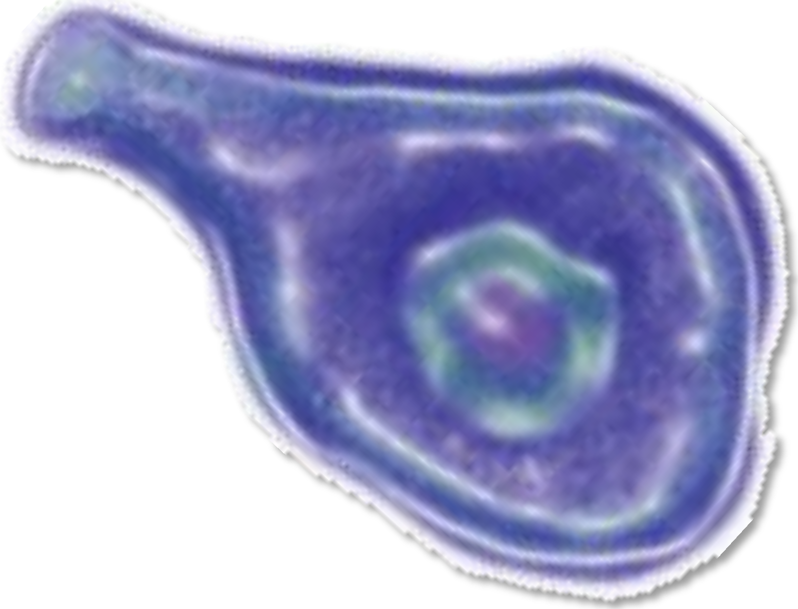
Pilot models will show the feasibility of bacteria and human models

Tips & tricks

Challenges to g2p2pop

- Build models from imperfect data
- Capture complexity within and between scales
- Systematically link scales
- Scalably simulate multiple scales
- Collaborate

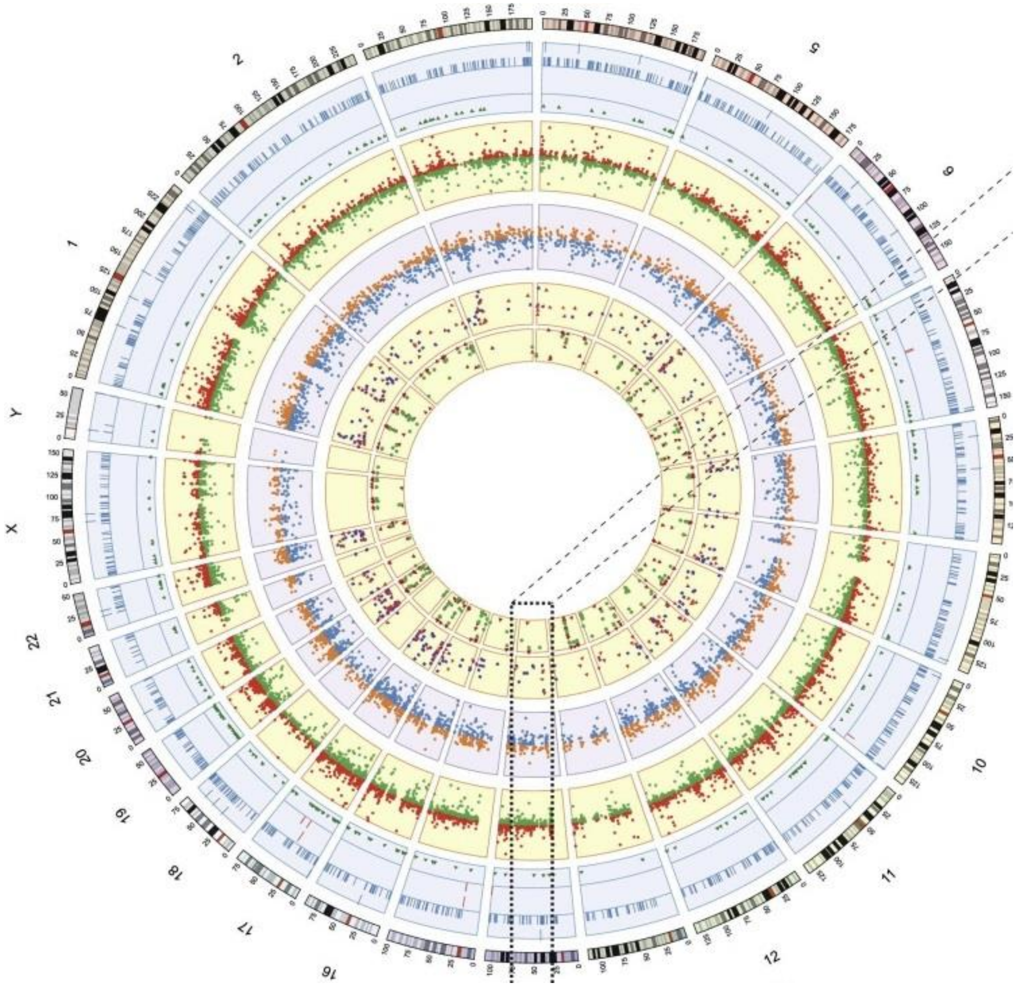
Stretch goals inspire innovation



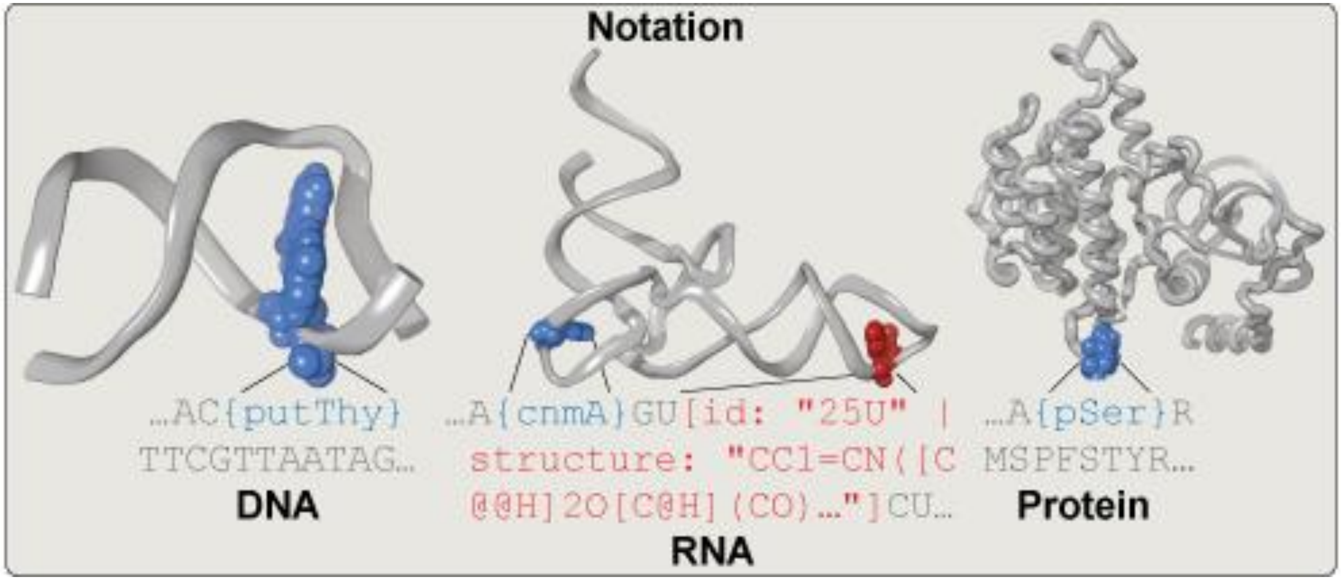
Integration enables great scope and depth

- Data aggregation
- Model composition
- Multi-algorithmic co-simulation
- Modular methods and software
- Interdisciplinary collaboration

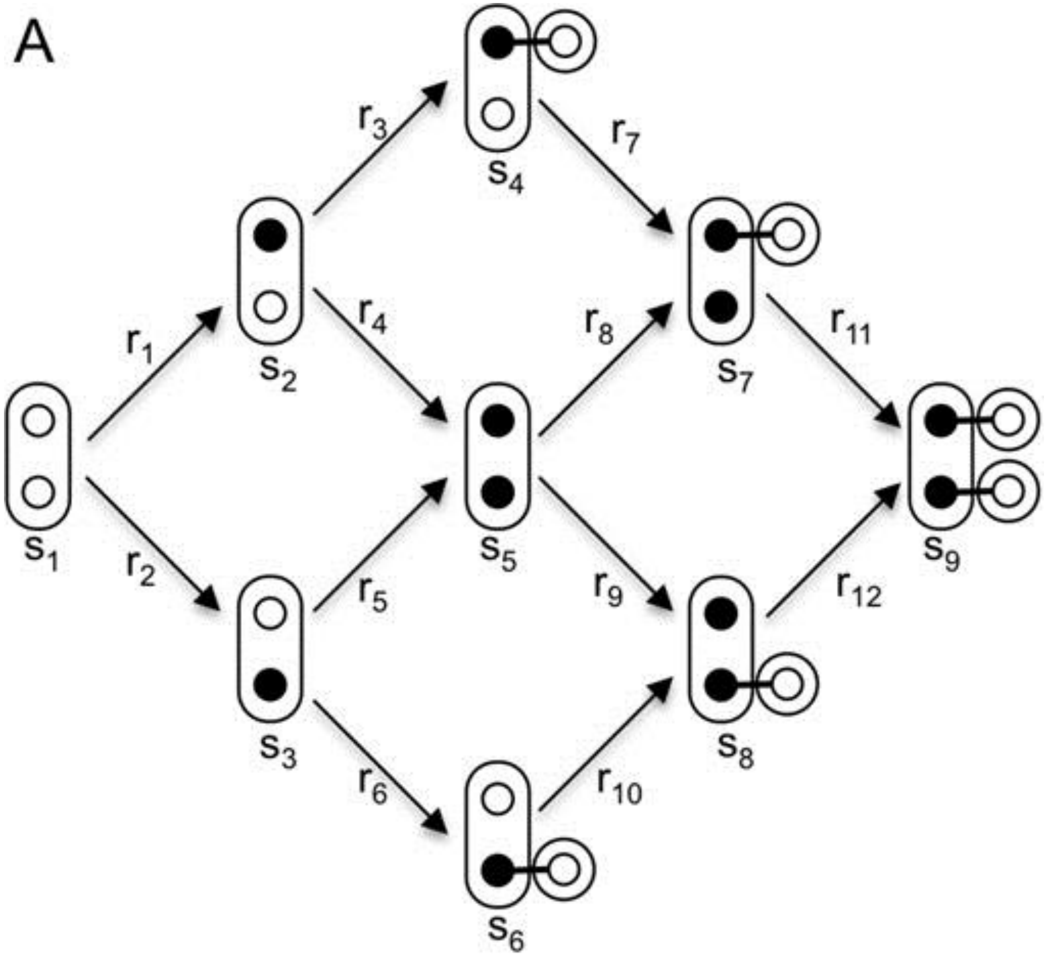
Frameworks enable scalable integration



Common languages enable frameworks

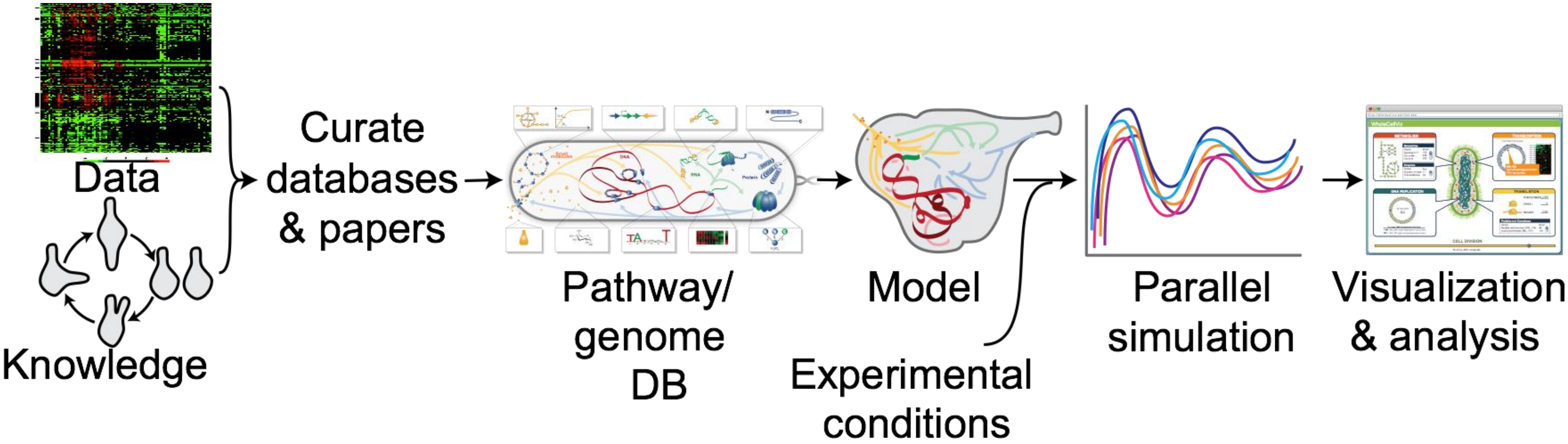


Agent-based modeling can capture complexity



Collaboration enables solutions

Modularity enables collaboration



Data

Knowledge

Curate databases & papers

Pathway/ genome DB

Model

Experimental conditions

Parallel simulation

Visualization & analysis

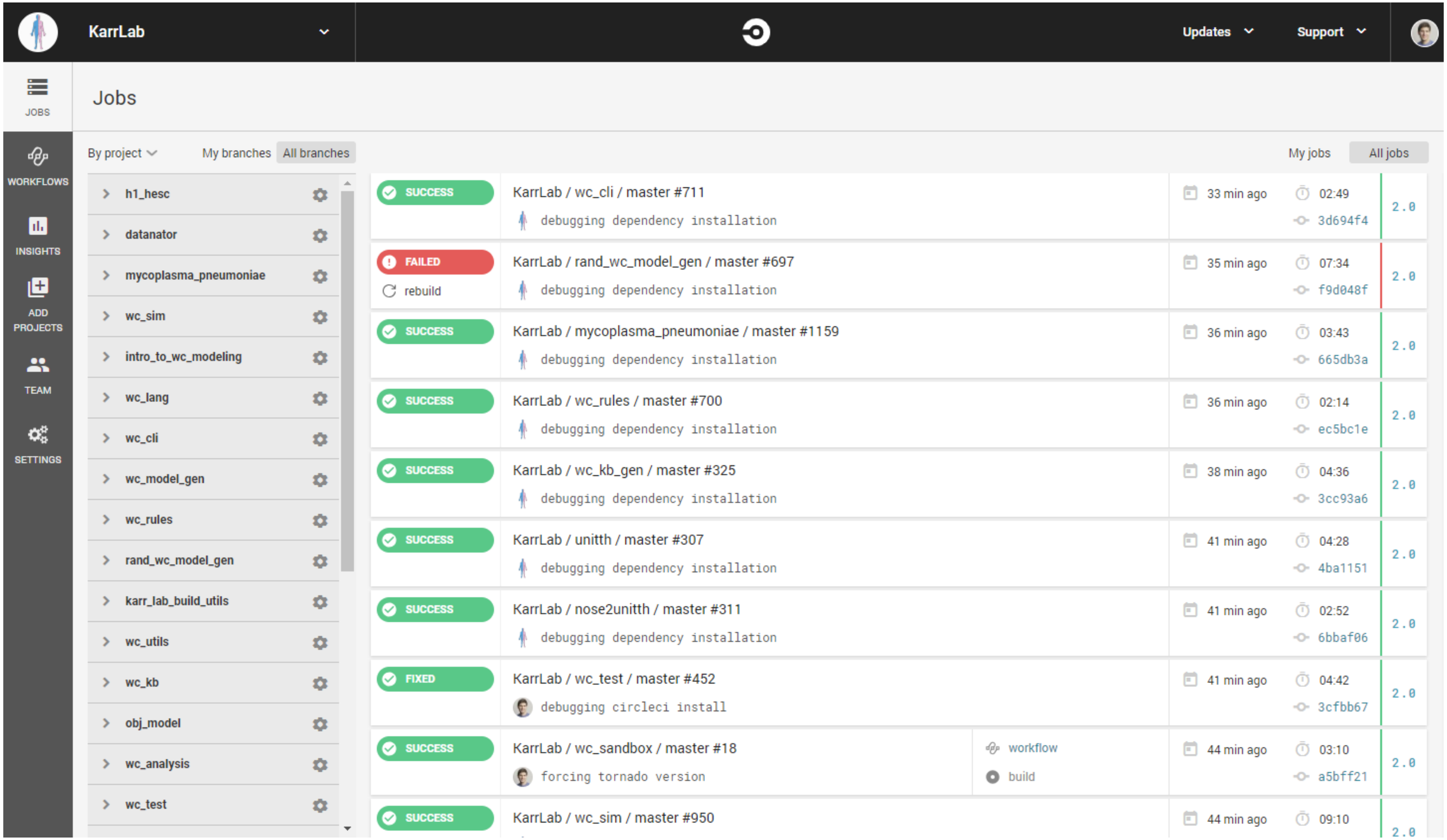
Sharing promotes collaboration

- **Quilt:** data
- **GitHub:** code
- **PyPI:** packaged code
- **Docker:** computing environments
- **Google Docs, Overleaf:** written documents
- **Google Drive:** other files
- **GitHub issues:** tasks

Common practices ease collaboration

- Interfaces between modules
- Coarse-graining
- Package organization
- Coding, documentation styles
- Software libraries

QC inspires trust among collaborators

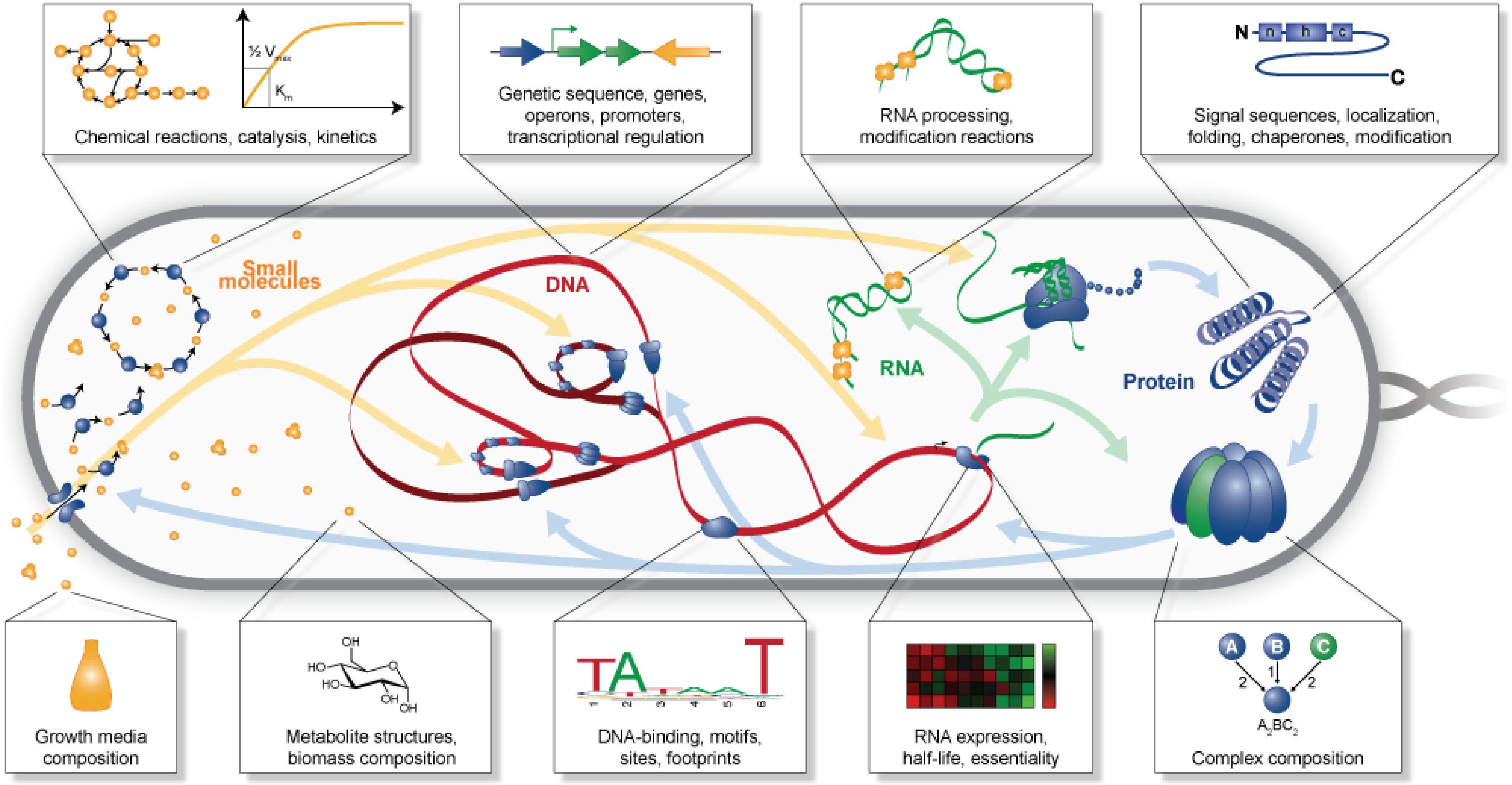


Jobs

- h1_hesc
- datanator
- mycoplasma_pneumoniae
- wc_sim
- intro_to_wc_modeling
- wc_lang
- wc_cli
- wc_model_gen
- wc_rules
- rand_wc_model_gen
- karr_lab_build_utils
- wc_utils
- wc_kb
- obj_model
- wc_analysis
- wc_test

SUCCESS	KarrLab / wc_cli / master #711 debugging dependency installation		33 min ago	02:49 3d694f4	2.0
FAILED rebuild	KarrLab / rand_wc_model_gen / master #697 debugging dependency installation		35 min ago	07:34 f9d048f	2.0
SUCCESS	KarrLab / mycoplasma_pneumoniae / master #1159 debugging dependency installation		36 min ago	03:43 665db3a	2.0
SUCCESS	KarrLab / wc_rules / master #700 debugging dependency installation		36 min ago	02:14 ec5bc1e	2.0
SUCCESS	KarrLab / wc_kb_gen / master #325 debugging dependency installation		38 min ago	04:36 3cc93a6	2.0
SUCCESS	KarrLab / unitth / master #307 debugging dependency installation		41 min ago	04:28 4ba1151	2.0
SUCCESS	KarrLab / nose2unitth / master #311 debugging dependency installation		41 min ago	02:52 6bbaf06	2.0
FIXED	KarrLab / wc_test / master #452 debugging circleci install		41 min ago	04:42 3cfbb67	2.0
SUCCESS	KarrLab / wc_sandbox / master #18 forcing tornado version	workflow build	44 min ago	03:10 a5bff21	2.0
SUCCESS	KarrLab / wc_sim / master #950		44 min ago	09:10	2.0

Data integration enables modelers to drive science



Chemical reactions, catalysis, kinetics

Genetic sequence, genes, operons, promoters, transcriptional regulation

RNA processing, modification reactions

Signal sequences, localization, folding, chaperones, modification

Small molecules

DNA

RNA

Protein

Growth media composition

Metabolite structures, biomass composition

DNA-binding, motifs, sites, footprints

RNA expression, half-life, essentiality

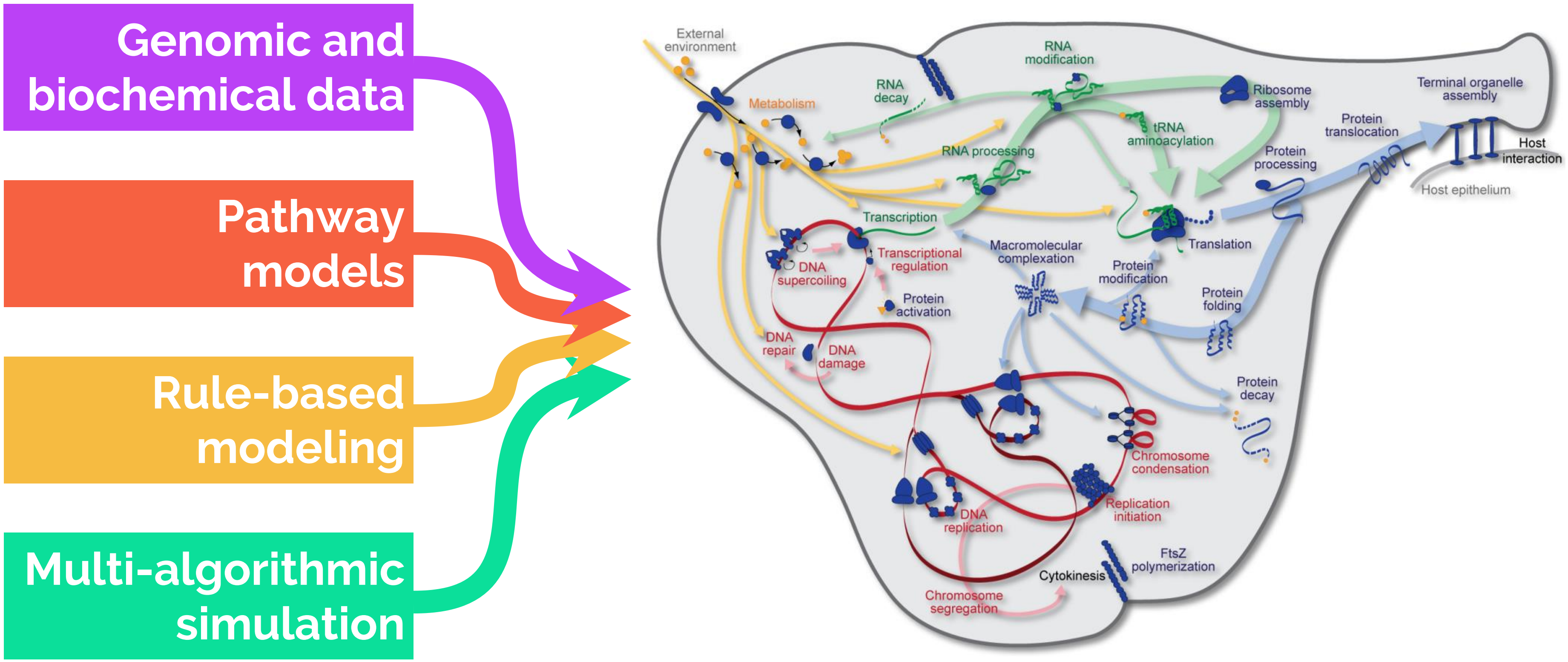
Complex composition

Summary

Whole-Cell
Modeling

Mount
Sinai

Integration is enabling WC modeling



Genomic and biochemical data

Pathway models

Rule-based modeling

Multi-algorithmic simulation

Whole-Cell Modeling

Mount Sinai