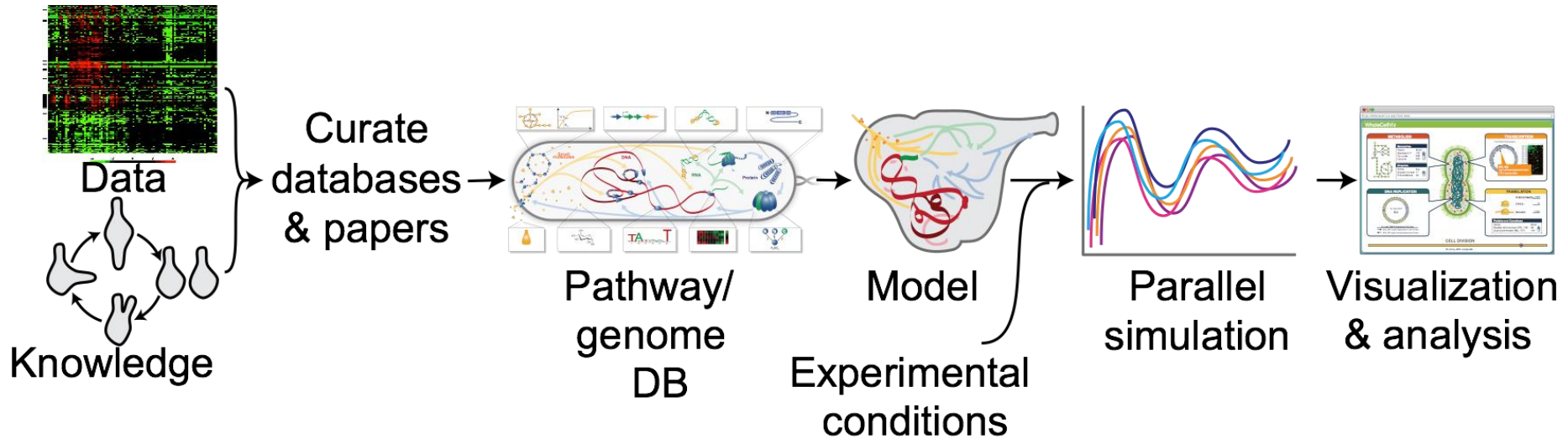


Systemizing and accelerating whole-cell modeling



WC modeling principles



Single-cell



Temporally
complete

GATCCA

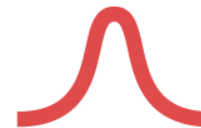
Species-specific



Mechanistic/dynamic



Genetically
complete



Stochastic

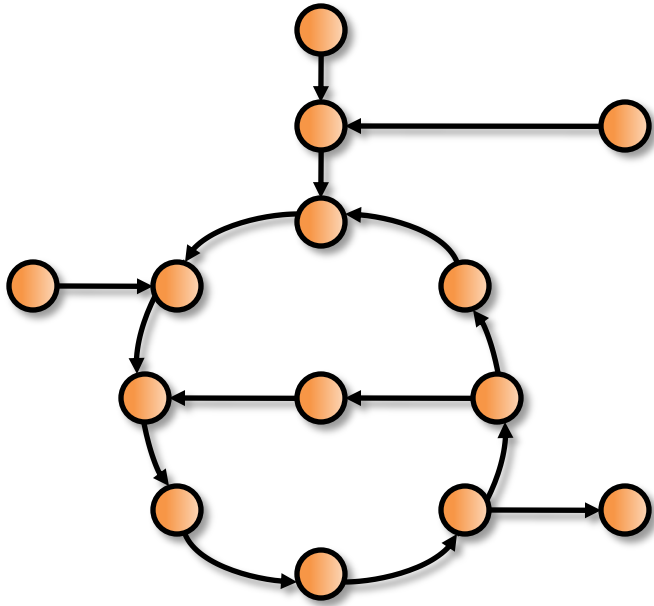


Molecularly
precise

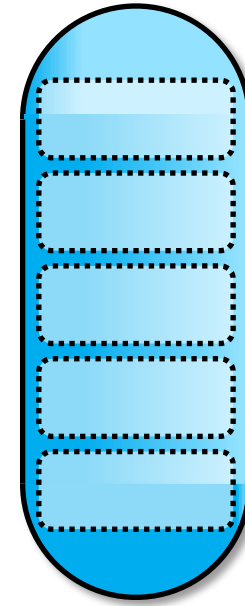
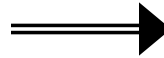


Accurate

WC modeling = genomics + integrative modeling

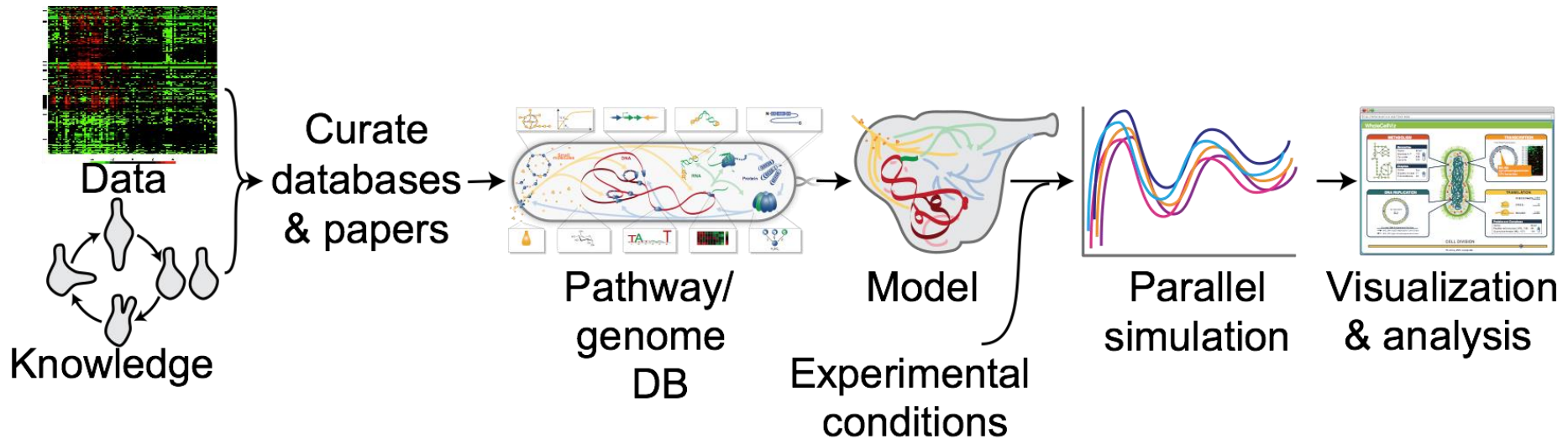


Molecular data



Integrative model

WC modeling process



1. Characterize organism
2. Aggregate data
3. Organize data
4. Design submodels
5. Merge submodels
6. Simulate model

7. Estimate parameters
8. Verify model
9. Validate model
10. Visualize/analyze predictions
11. Applications: Engineering, medicine

Limitations of existing methods

Limited scope and accuracy

- Don't represent several cell functions
- Don't predict several phenotypes

Methods are not rigorous

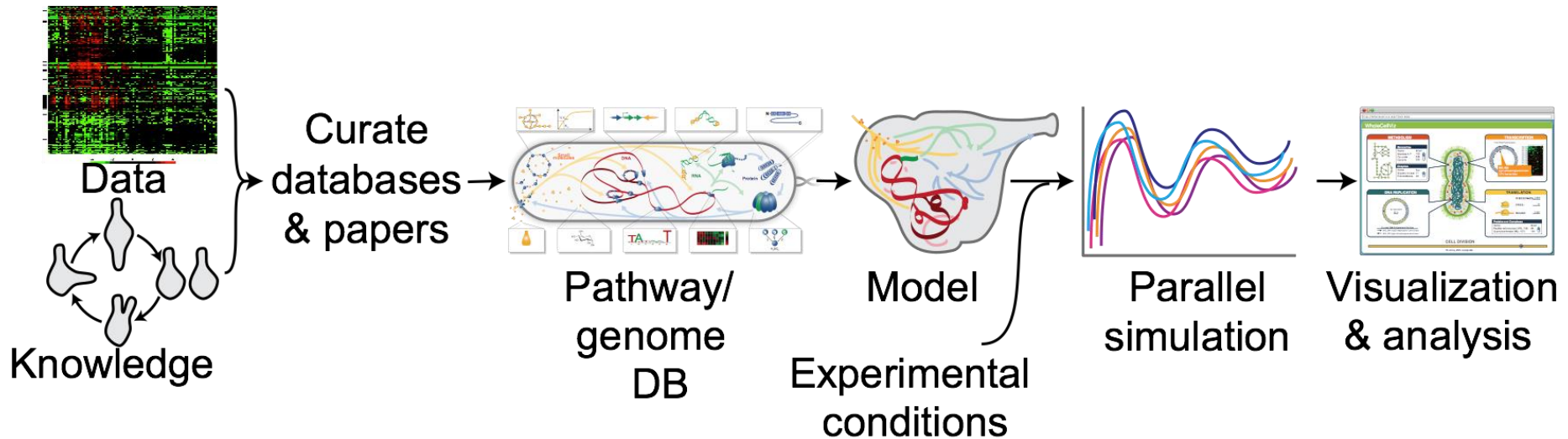
- Multi-algorithm simulation
- Parameter estimation
- Verification
- Model reduction

Time-consuming to construct

- Curate data
- Design model
- Verify model

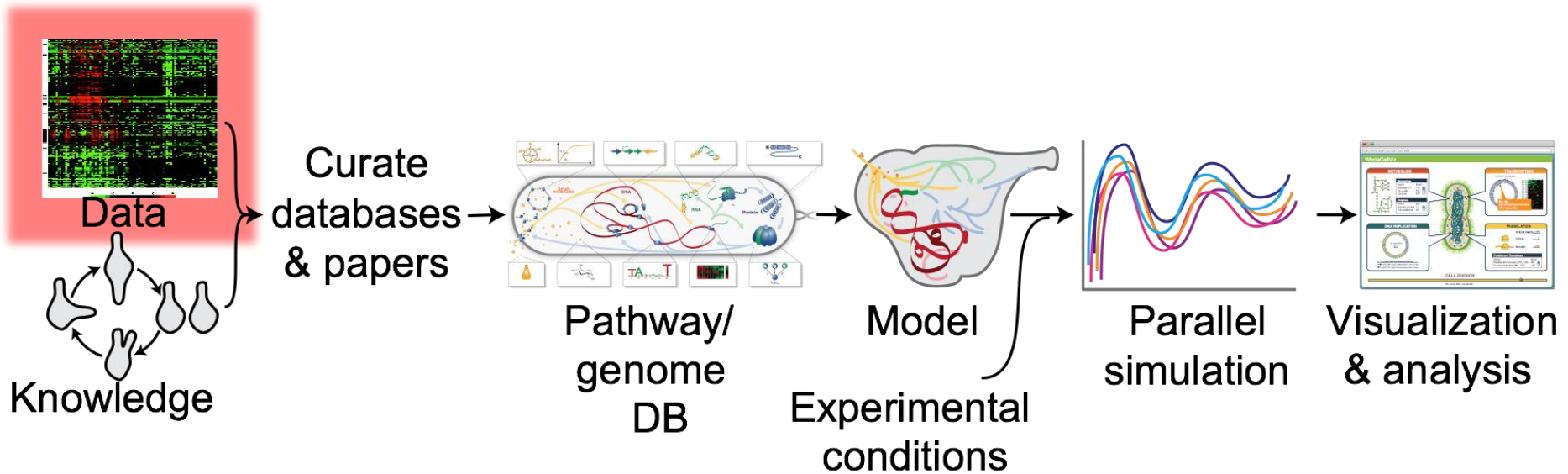
Hard to understand, reuse, reproduce

Systemizing and accelerating WC modeling

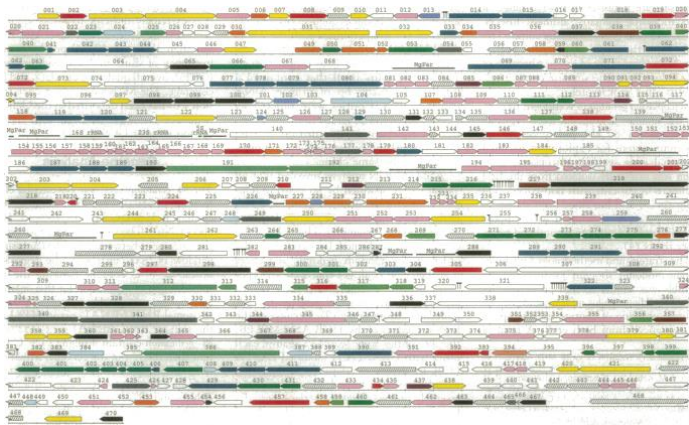


1. Characterize organism
2. Aggregate data
3. Organize data
4. Design submodels
5. Merge submodels
6. Simulate model
7. Estimate parameters
8. Verify model
9. Validate model
10. Visualize/analyze predictions
11. Applications: Engineering, medicine

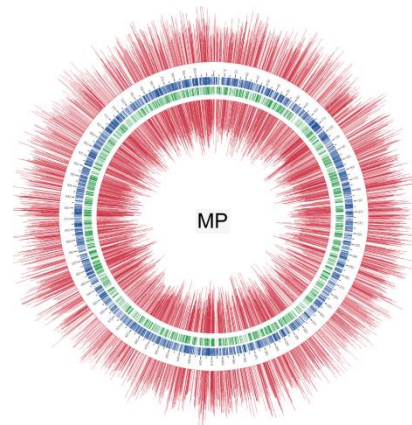
1. Characterize organism



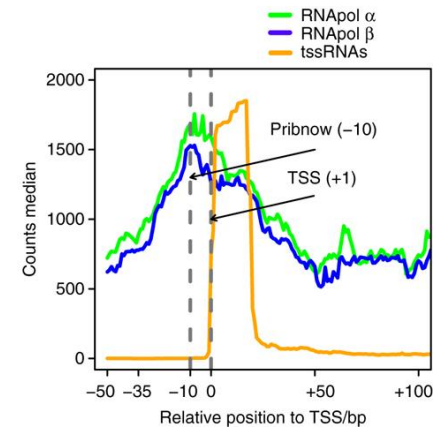
1. Characterize organism



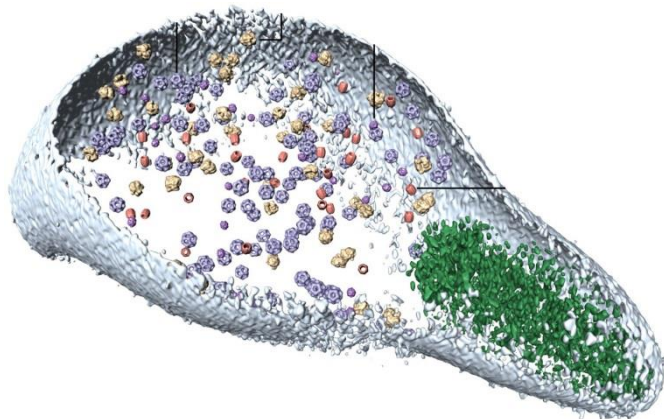
Genome
DNA-seq



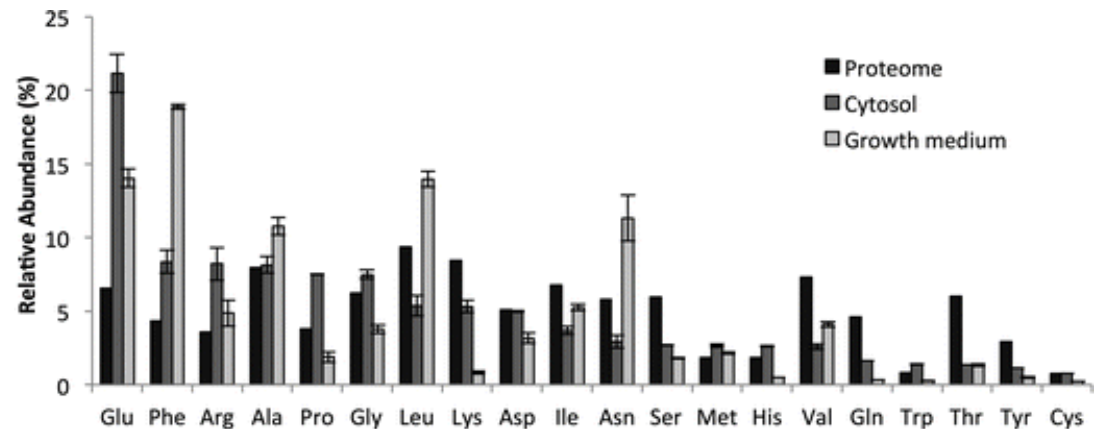
Epigenome
Meth-seq



Transcriptome
RNA-seq



Proteome
Mass-spectrometry

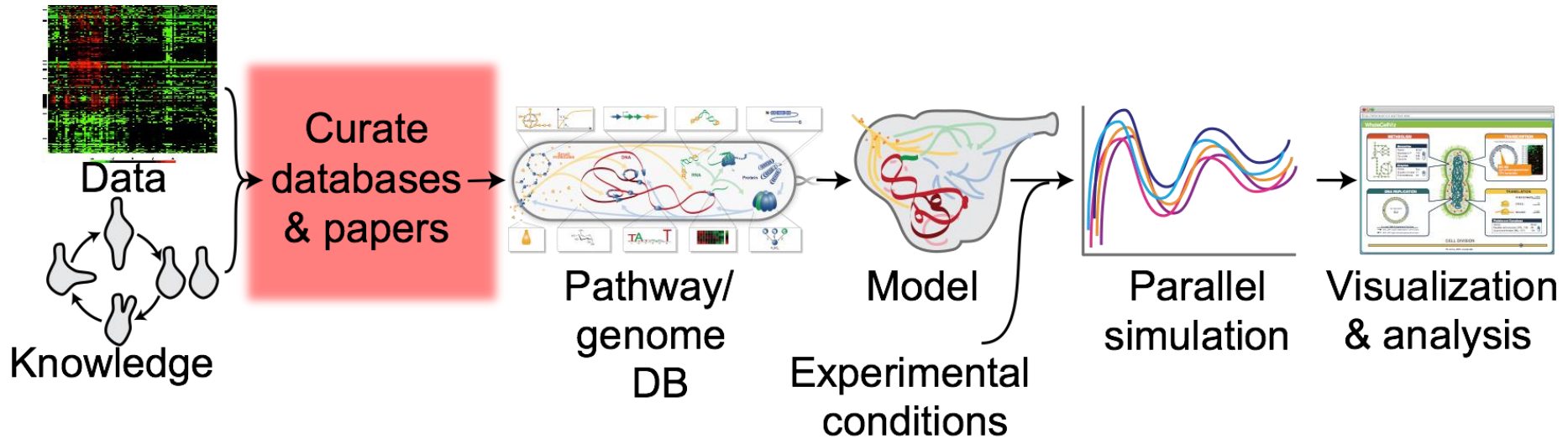


Metabolome
Mass-spectrometry

1. Characterize organism

- Chaperones
- Complex composition
- DNA binding sites
- DNA footprints
- DNA methylation
- DNA sequence
- Gene-drug interactions
- Genome annotation
- Growth rates
- Metabolite concentrations
- Protein cofactors
- Protein expression
- Protein half-lives
- Protein localization
- Protein modification
- RNA editing
- RNA expression
- RNA half-lives
- RNA modification
- RNA maturation
- Reaction fluxes
- Reaction kinetics
- Reaction stoichiometries
- Signaling pathways
- DNA mutations

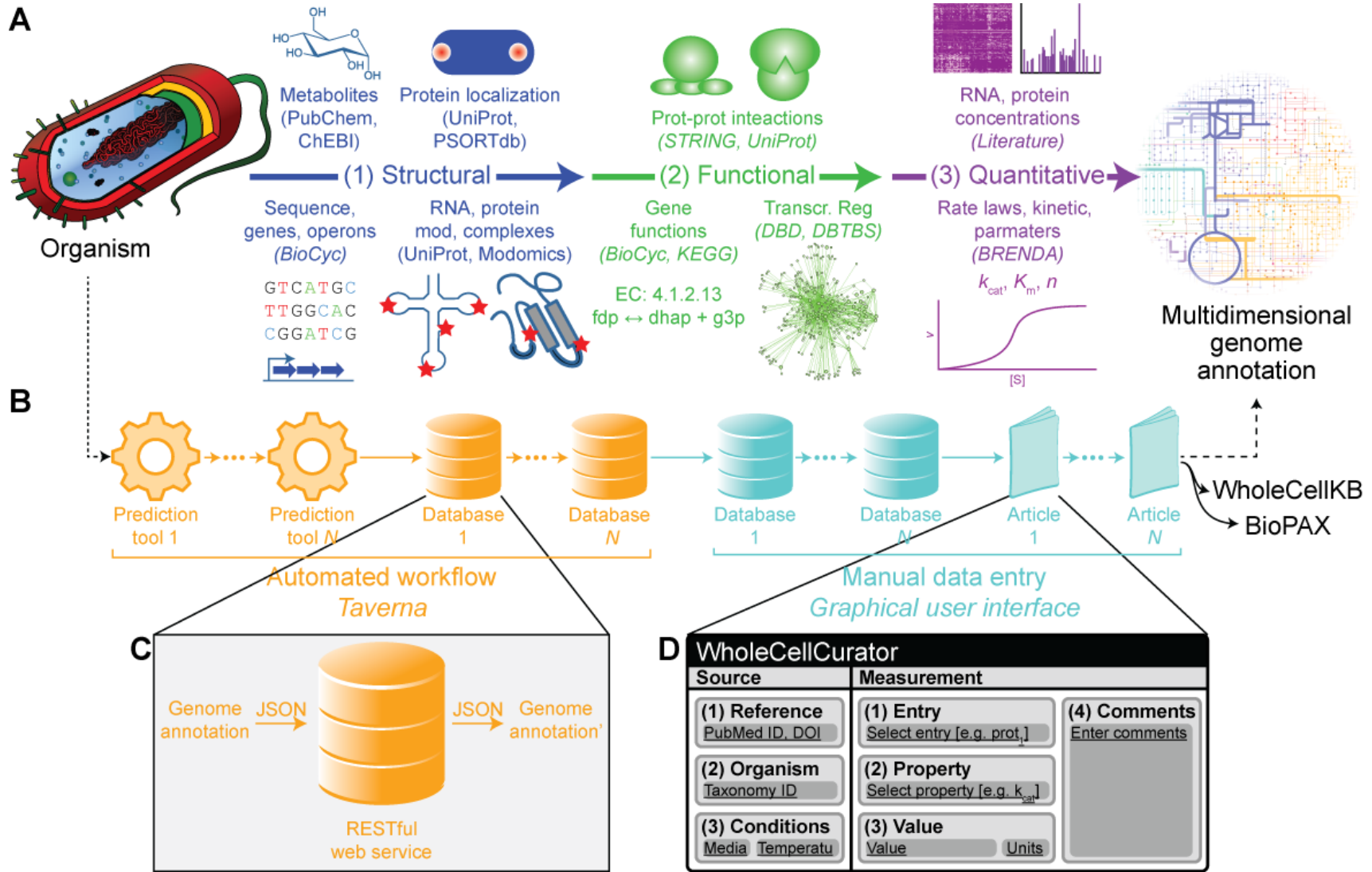
2. Aggregate data



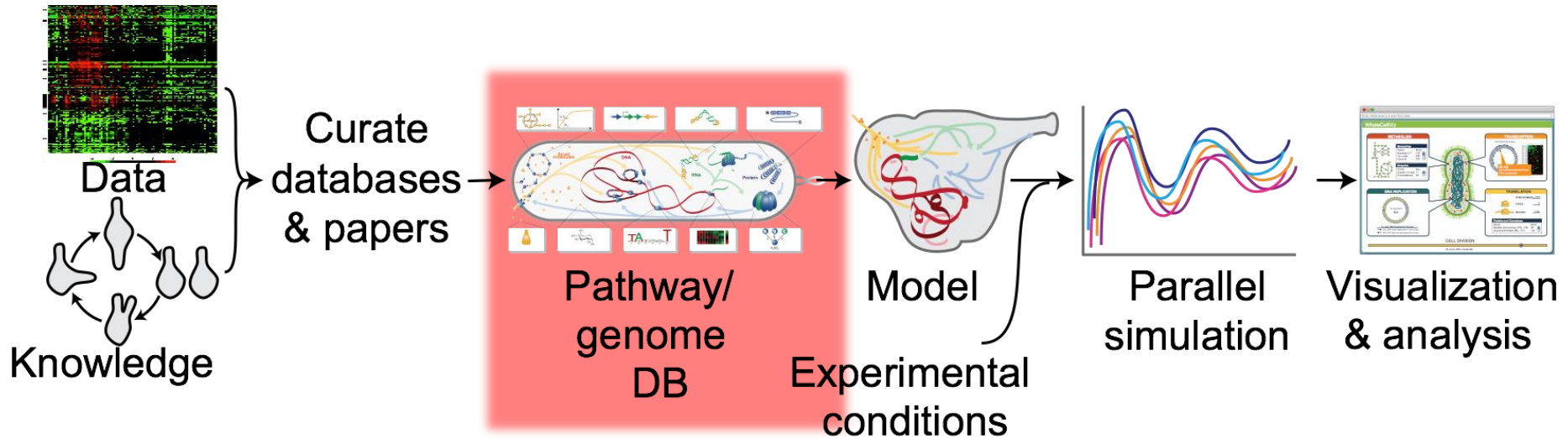
2. Aggregate data

	Data type	Source
Baseline	Chaperones	Literature
	Complex composition	Human Soluble Protein Complexes DB
	DNA binding sites	TRANSFAC, AnimalTFDB
	DNA footprints	Literature
	DNA methylation	MethBase
	DNA sequence	Genome Reference Consortium
	Gene-drug interactions	DrugBank, PharmaGKB
	Genome annotation	Ensembl
	Growth rates	Hapmap, NCI-60
	Metabolite concentrations	Human Metabolome Database
	Protein cofactors	UniProt
	Protein expression	Human Protein Atlas
	Protein half-lives	Literature
	Protein localization	Human Protein Atlas
	Protein modification	Human Protein Reference DB
	RNA editing	RADAR, DARNED
	RNA expression	GEO, Human Protein Atlas
	RNA half-lives	Literature
	RNA modification	RNA Modification DB, MODOMICS
	RNA maturation	RNApathwaysDB
	Reaction fluxes	Literature
	Reaction kinetics	SABIO-RK, BRENDA
	Reaction stoichiometries	Recon X, UniProt, HumanCyc
	Signaling pathways	Literature
Disease	DNA mutations	CCLE, COSMIC
	DNA methylation	TCGA
	Gene-drug interactions	CCLE
	Growth rates	NCI-60
	Metabolite concentrations	Literature
	Protein expression	TCGA
	RNA expression	CCLE

2. Aggregate data



3. Organize data

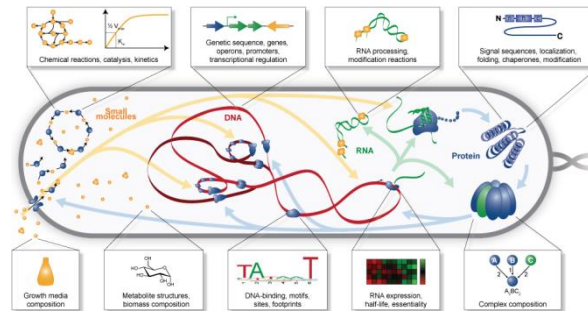


3. Organize data

Aggregate data: Scripts, Excel



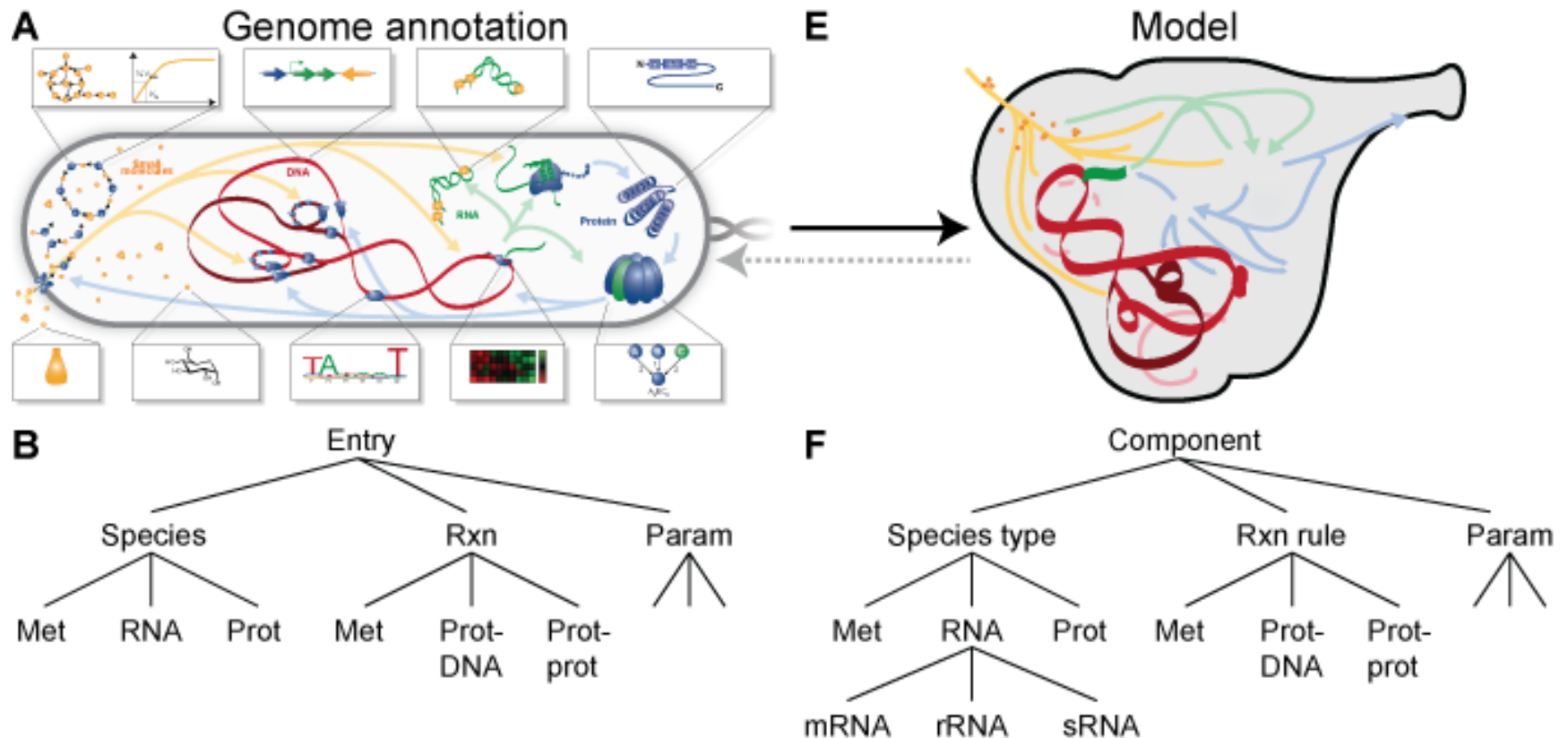
PGDB: Pathway Tools, WC-KB



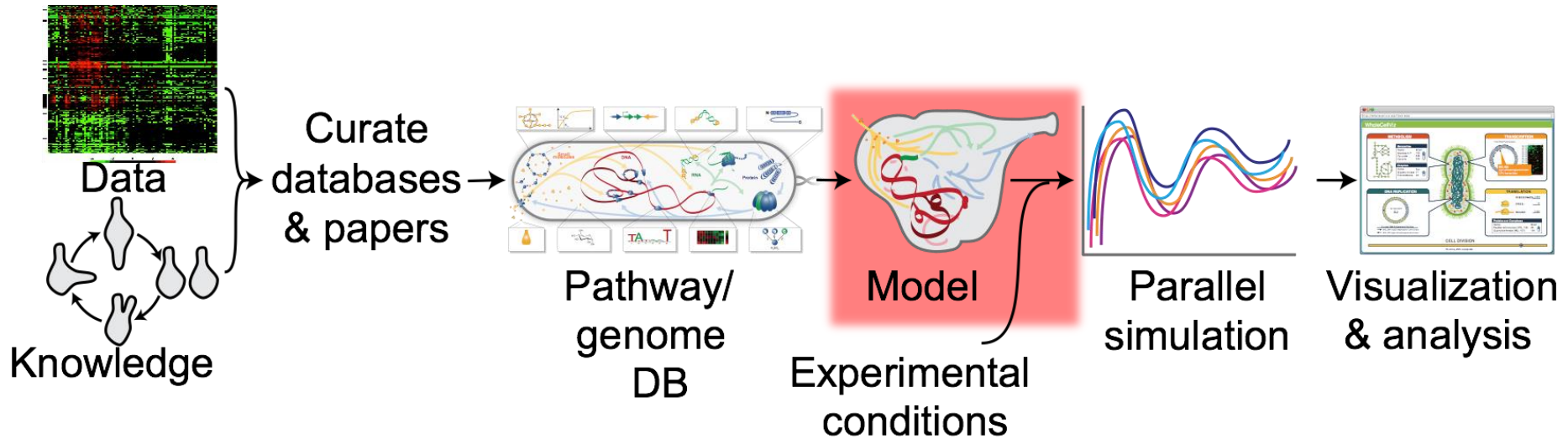
PGDB Browser

A screenshot of the EcoCyc database interface. The top navigation bar includes 'Search', 'Genome', 'Metabolism', 'Analysis', 'SmartTables', and 'Help'. The main content area shows details for the gene 'dnaC' (chromosome replication; initiation and chain elongation) in *Escherichia coli*. It lists accession IDs (EG10217, EC64351, PQ4E70), length (738 bp / 245 aa), map position, and location (cytosol). A 'Regulation Summary Diagram' shows a regulatory network involving RNAP, dnaX, and dnaC. The 'Summary' section states that DnaC is an accessory protein that loads the DnaB replicative helicase onto duplex DNA to initiate replication.

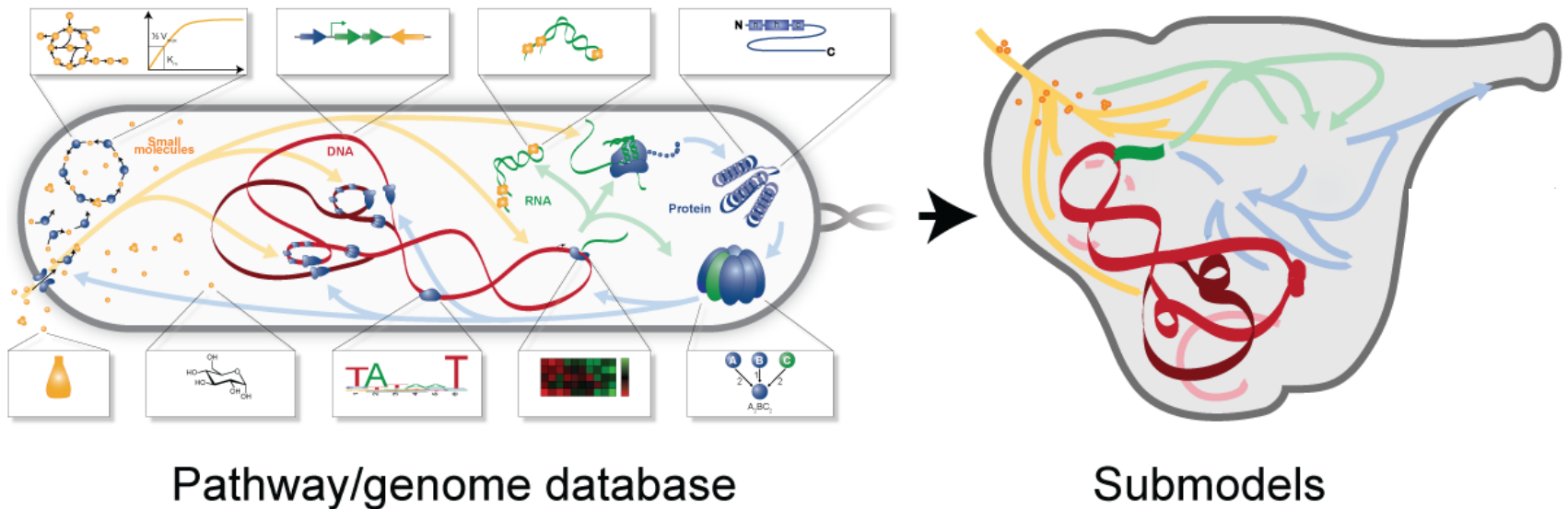
3. Organize data



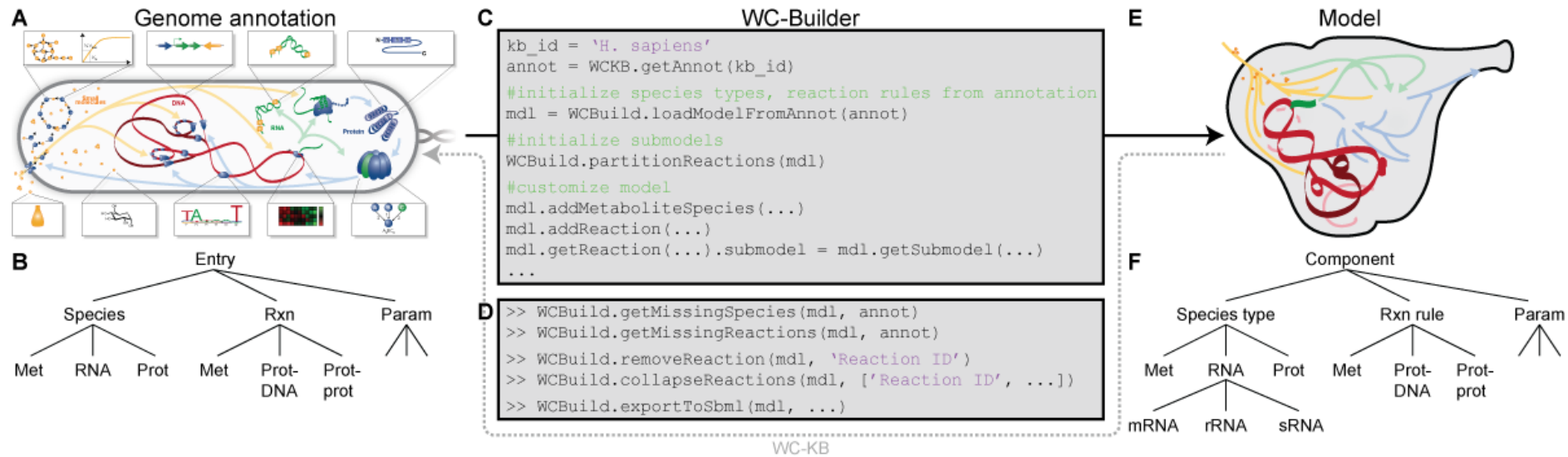
4. Design submodels



4. Design submodels



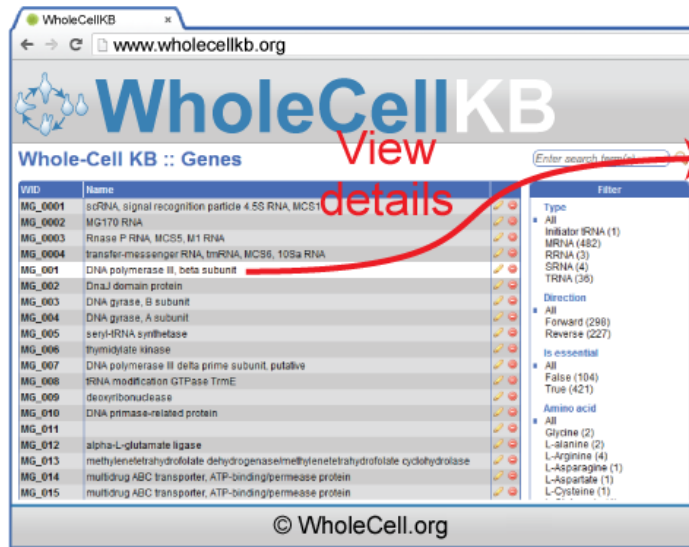
4. Design submodels



WC-ML, SBML, CellML

4. Design submodels

A



WholeCellKB

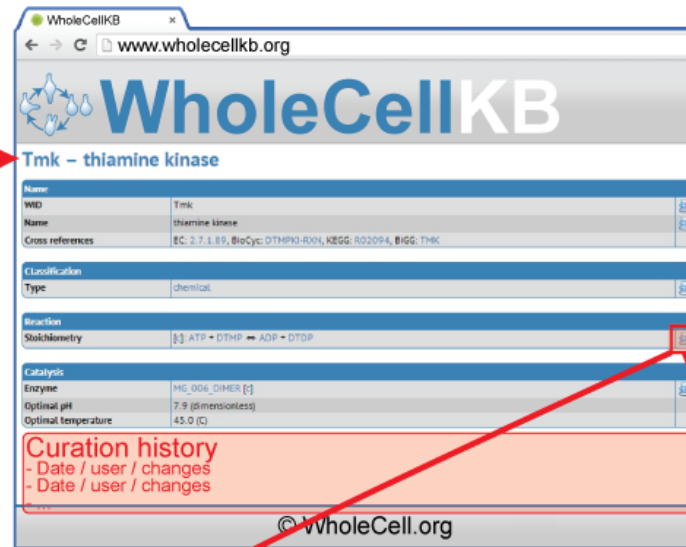
Whole-Cell KB :: Genes

View details

WID	Name	Type
MG_0001	sRNA, signal recognition particle 4.5S RNA, MCS1	All
MG_0002	MG170 RNA	Initiator (RNA) (1)
MG_0003	Rnase P RNA, MCS5, M1 RNA	MRNA (482)
MG_0004	transfer-messenger RNA, tmRNA, MCS6, 10Sa RNA	SRNA (3)
MG_001	DNA polymerase III, beta subunit	TRNA (36)
MG_002	DNA domain protein	
MG_003	DNA gyrase, B subunit	
MG_004	DNA gyrase, A subunit	
MG_005	seri-RNA synthetase	
MG_006	thymidylate kinase	
MG_007	DNA polymerase III delta prime subunit, putative	
MG_008	tRNA modification GTPase TrmE	
MG_009	deoxynucleosidase	
MG_010	DNA primase-related protein	
MG_011		
MG_012	alpha-L-glutamate ligase	
MG_013	methyltetrahydrofolate dehydrogenase/methyltetrahydrofolate cyclohydrolase	
MG_014	multidrug ABC transporter, ATP-binding/permease protein	
MG_015	multidrug ABC transporter, ATP-binding/permease protein	

© WholeCell.org

B



WholeCellKB

Tmk - thiamine kinase

Name: Tmk

Type: thiamine kinase

Cross references: EC: 2.7.1.85, BioCyc: DTHMP3-RXN, KEGG: R02094, BiGG: TMK

Classification: Type: chemical

Reaction: Stoichiometry: $ATP + DTHP \rightleftharpoons ADP + DTP$

Catalysis: Enzyme: MG_006_DIMER [-]

Optimal pH: 7.9 (dimensionless)

Optimal temperature: 45.0 (C)

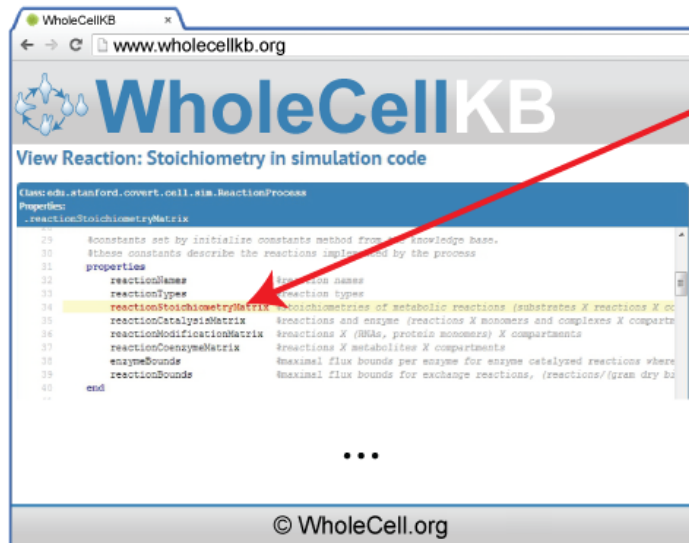
View in model

Curation history

- Date / user / changes
- Date / user / changes

© WholeCell.org

C



WholeCellKB

View Reaction: Stoichiometry in simulation code

```

@name edo.stanford.covart.cell.eia.ReactionProcess
Properties:
  .reactionStoichiometryMatrix
29
30 $constants set by initialize constants method from knowledge base.
31 $these constants describe the reactions implemented by the process
32
33 properties
34   reactionNames      reaction names
35   reactionTypes      reaction types
36   reactionStoichiometryMatrix  stoichiometries of metabolic reactions (substrates X reactions X R)
37   reactionCatalysisMatrix  reactions and enzyme (reactions X monomers and complexes X compartments)
38   reactionModificationMatrix  reactions X (RNAs, protein monomers) X compartments
39   reactionCofactorMatrix  reactions X metabolites X compartments
40   enzymeBounds      maximal flux bounds per enzyme for enzyme catalyzed reactions where
41   reactionBounds      maximal flux bounds for exchange reactions, (reactions/gram dry bi
42
43 end
  
```

© WholeCell.org

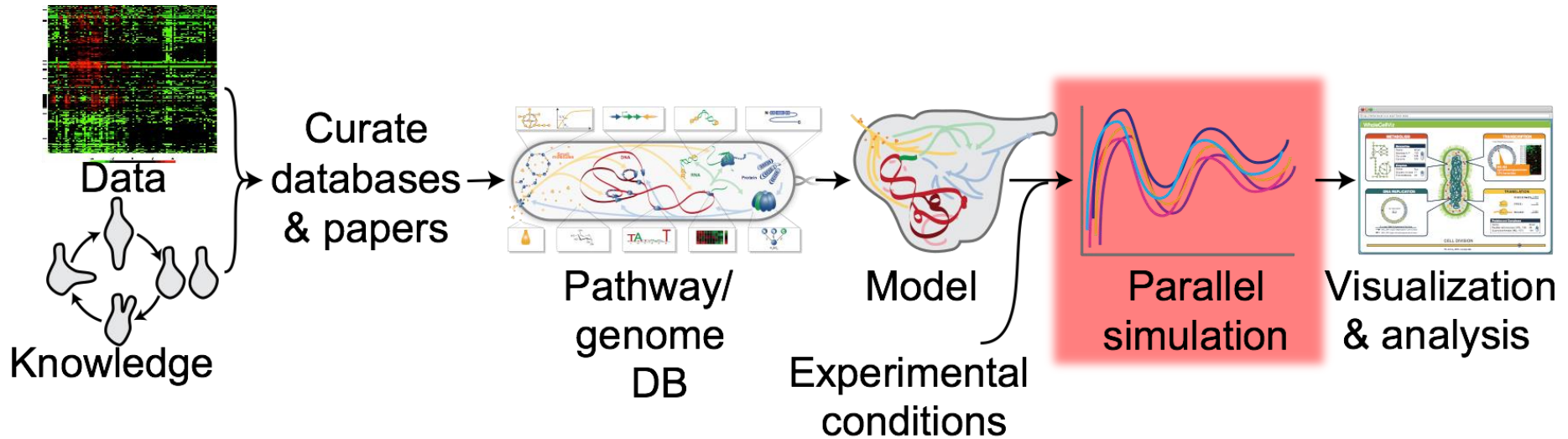
D

Entry	ModelComp.	User
organism: <i>Organism</i> id: string name: string synonyms: [string] comments: string evidence: [Evidence] modelComp: [ModelComp] modifiedUser: User validFrom: date validTo: date	modelId: string submodelId: string compld: string linesInSbml: [int]	firstName: string lastName: string affiliation: string email: string creation: date

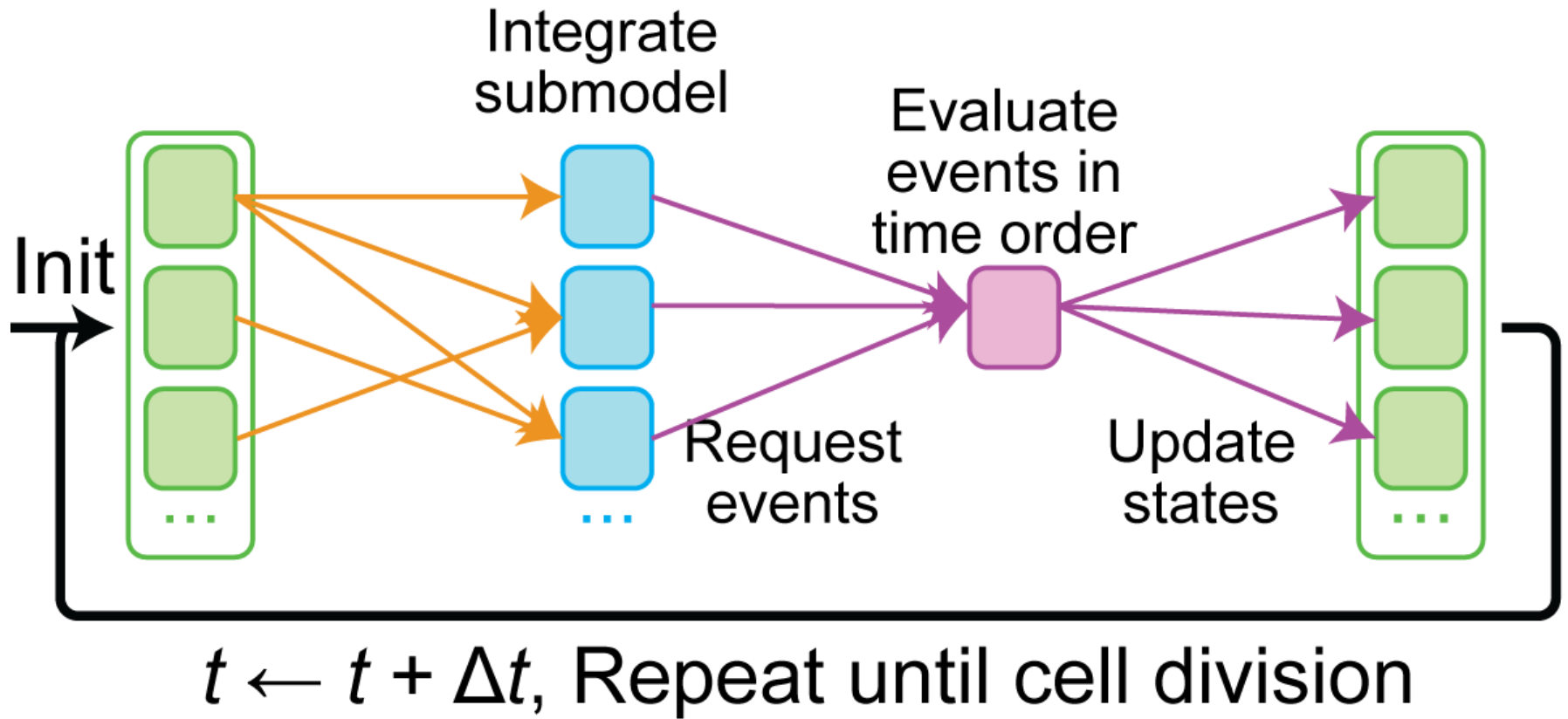
5. Map submodels onto common state

*Automatically handled by designing
submodels from common PGDB*

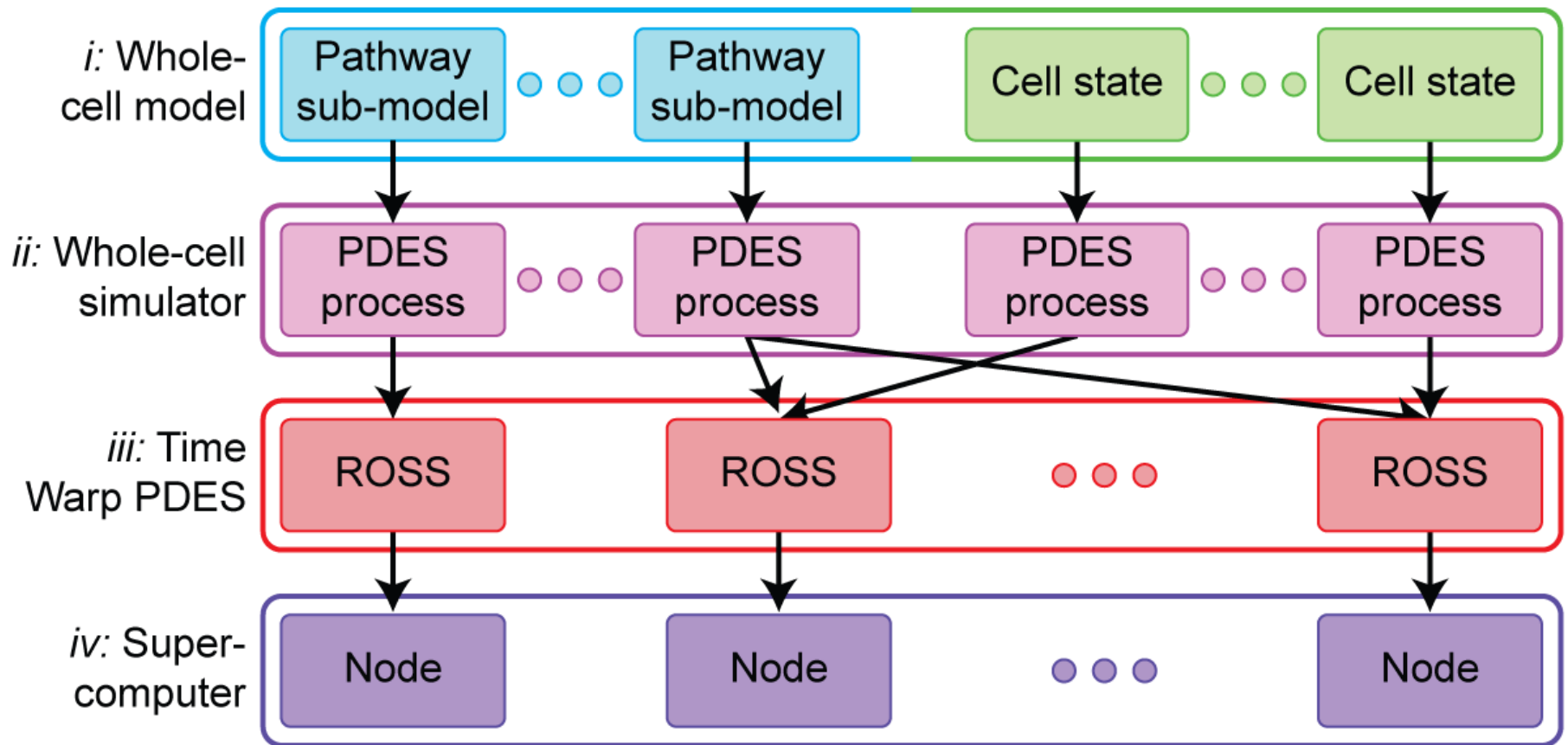
6. Simulate



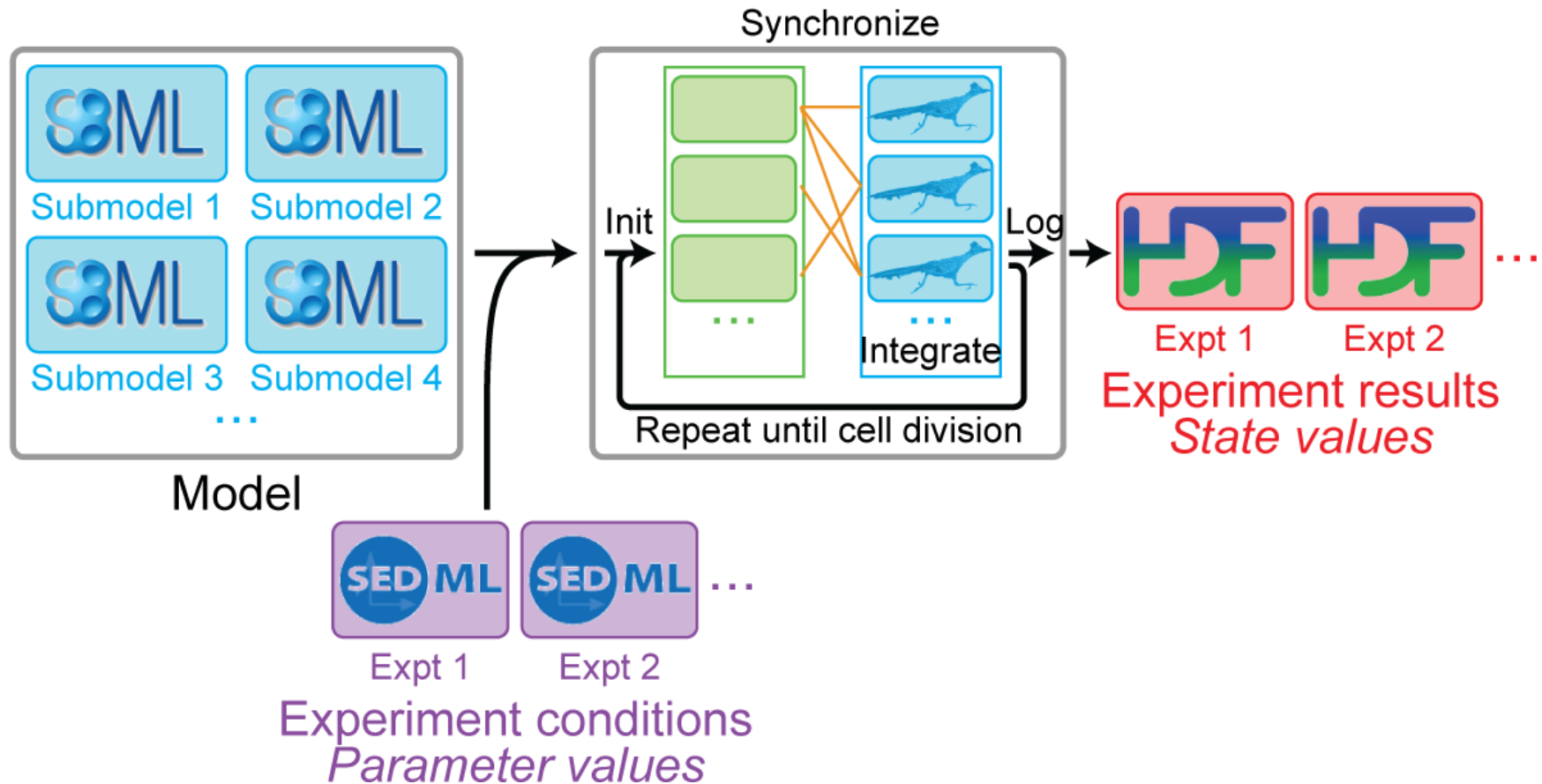
6. Simulate: Concurrently integrate submodels



6. Simulate: Parallel discrete event simulation



6. Simulate: High-performance simulator



6. Simulate: High-performance simulator

Whole-cell simulator

Config simulations

(1) Select model
Select Mge.sbml

(2) Parameter vals
 $[ATP]_i$ 1,000
 k_{cat} 10
 τ 10


(3) Stat sampling
Simulations 1,000
Length (h) 10

(4) Cluster config
Nodes 100

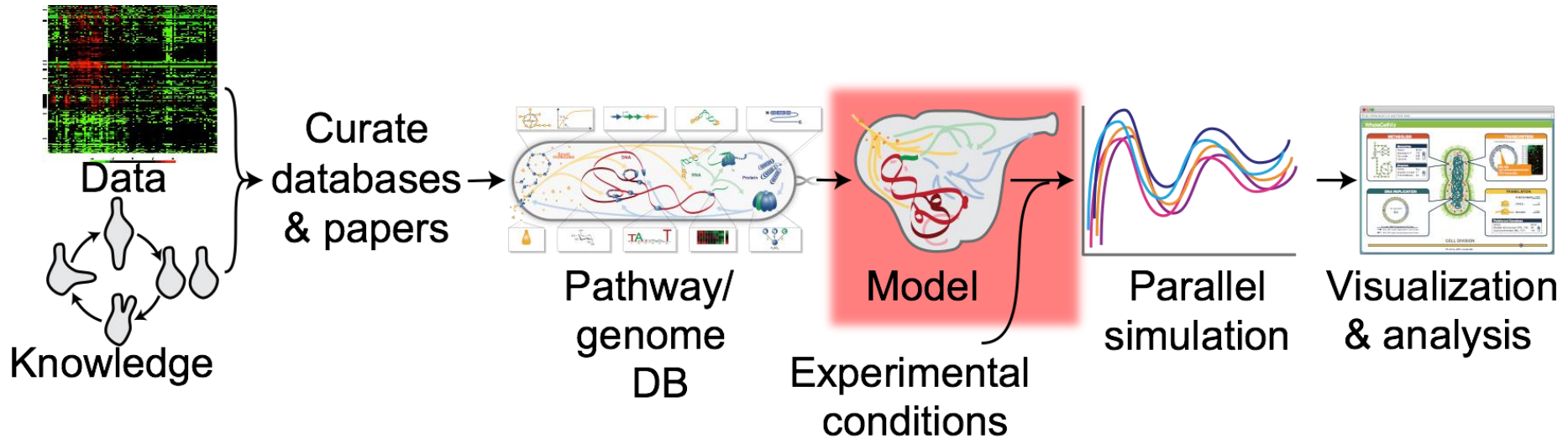
Results

(1) Select simulations
☒1 ☐2
☒3 ☐4

(2) Select states
☒RNA-1 ☐RNA-3
☐RNA-2 ☐Protein-1

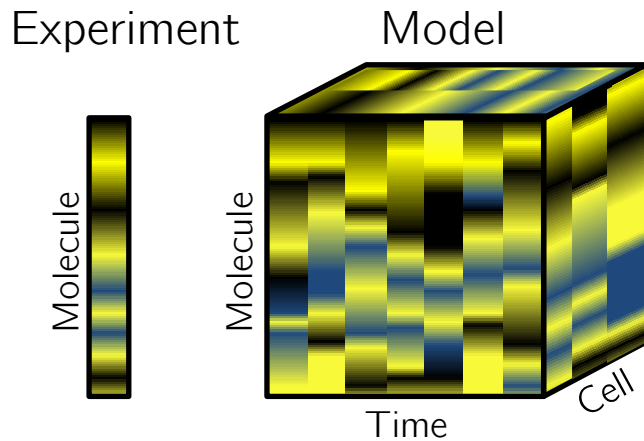
(3) View results


7. Estimate parameters



7. Estimate parameters

1. Reduce model



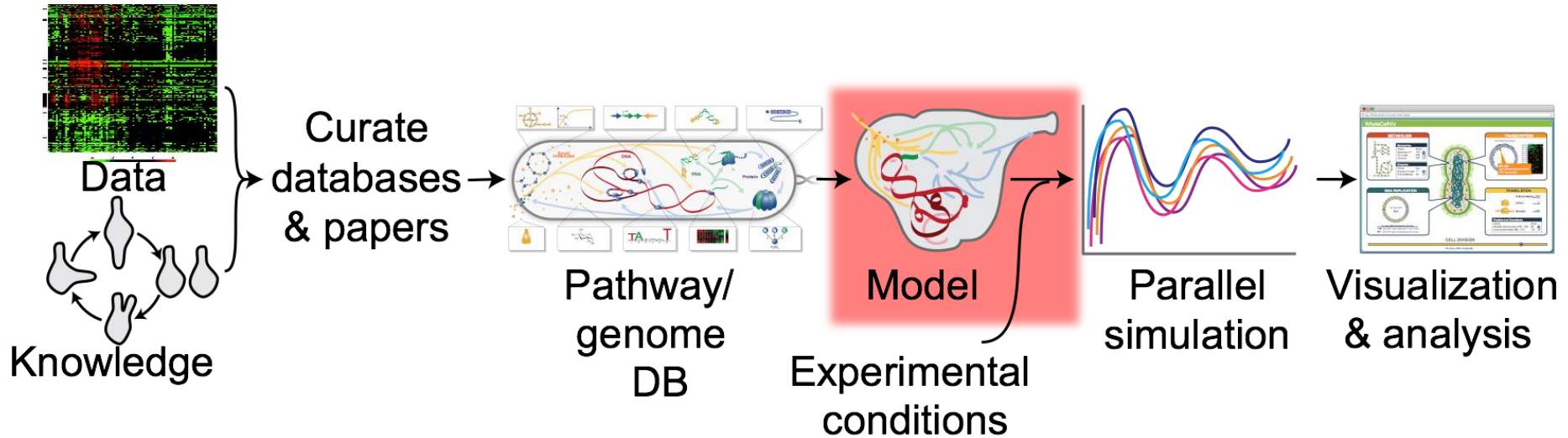
2. Identify reduced model parameters using traditional methods

3. Manually tune parameters using full model

7. Estimate parameters

- Automatic model reduction
- Distributed numerical optimization
- Enabled by
 - Declarative model description
 - High-performance simulation

8. Verify model reproduces known biology



8. Verify model reproduces known biology

☒ 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

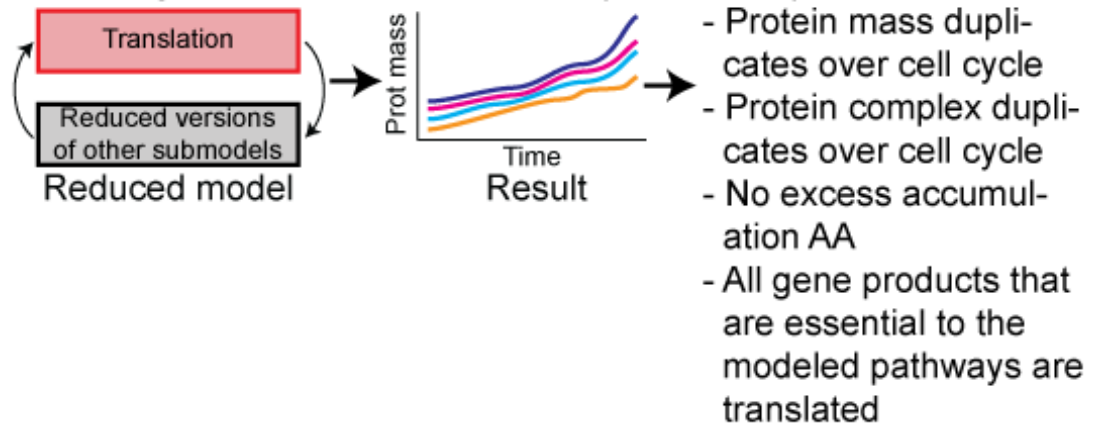
- ☒ Plot model predictions
- ☒ Manually inspect data
- ☒ Compare to known biology

8. Verify model reproduces known biology

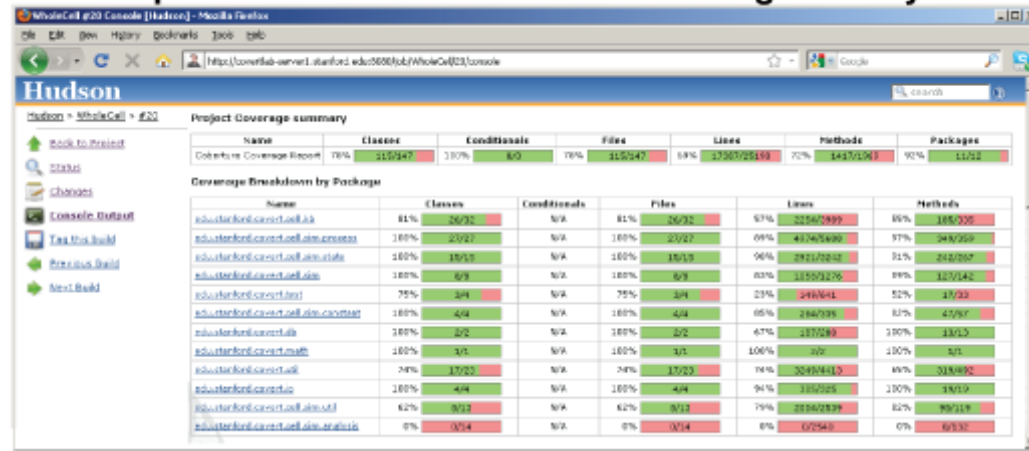
B. Example static test (Subaim 4a-c)

- Mass balance
- Charge balance
- Consistent localization
 - Small molecules
 - Enzymes
- Reactants and products can be produced/recycled by metabolism sub-model
- Sufficient small molecule and enzymatic resources to support growth

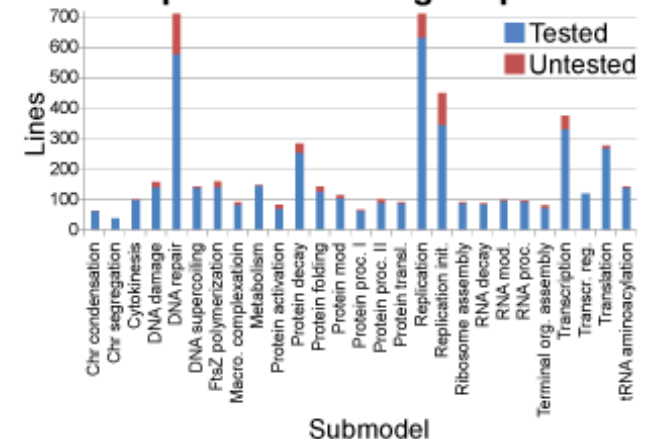
C. Example simulation-based test (Subaim 4d)



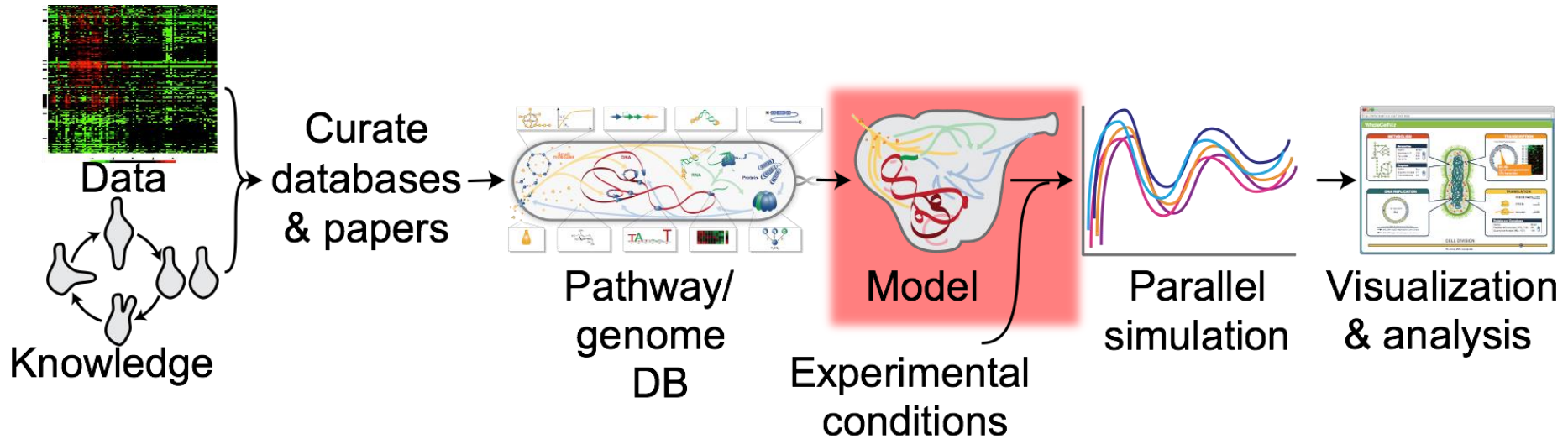
D. Example screenshot of a continuous integration system



E. Example test coverage report



9. Validate model reproduces *true* biology



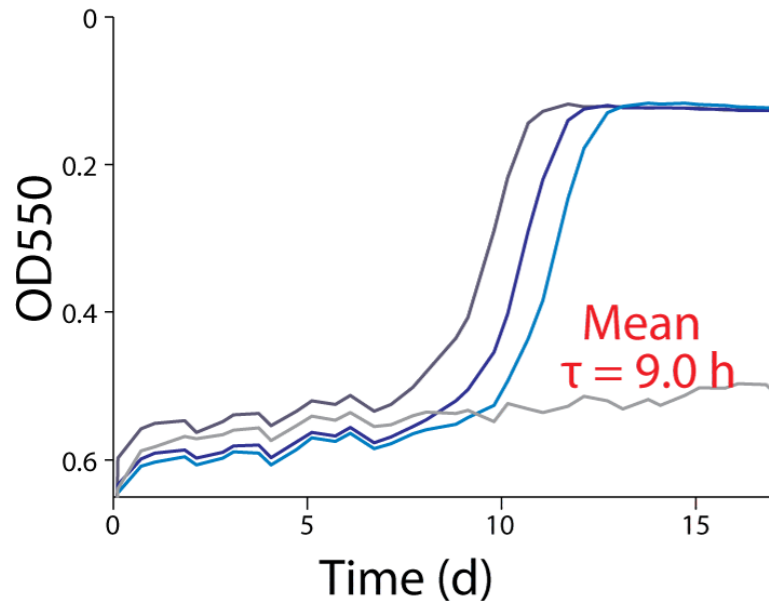
9. Validate model reproduces *true* biology

☑ Matches new data

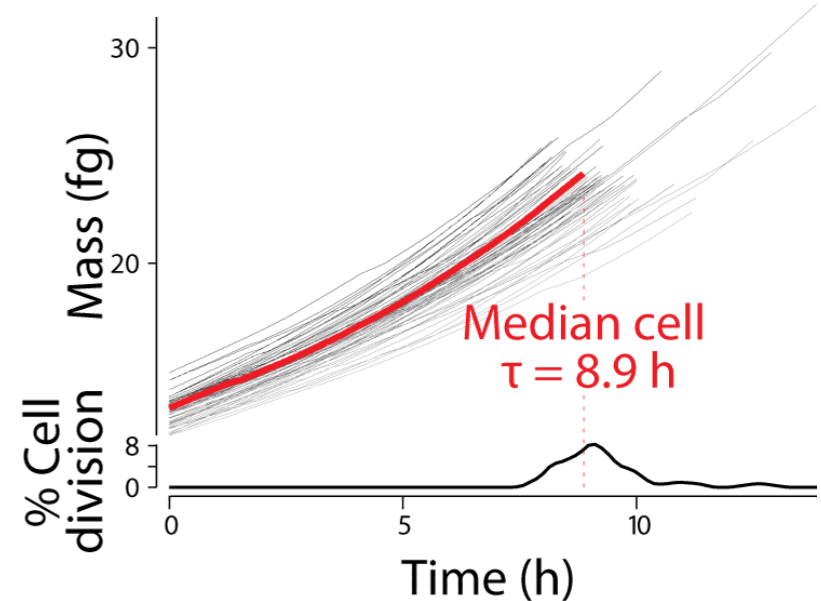
- ☑ Wild-type growth rate
- ☑ Disruption strain growth rates
- ☑ Single-cell division times
- ☑ Single-cell cell cycle phase lengths
- ☑ Single-cell sizes

9. Validate model reproduces *true* biology

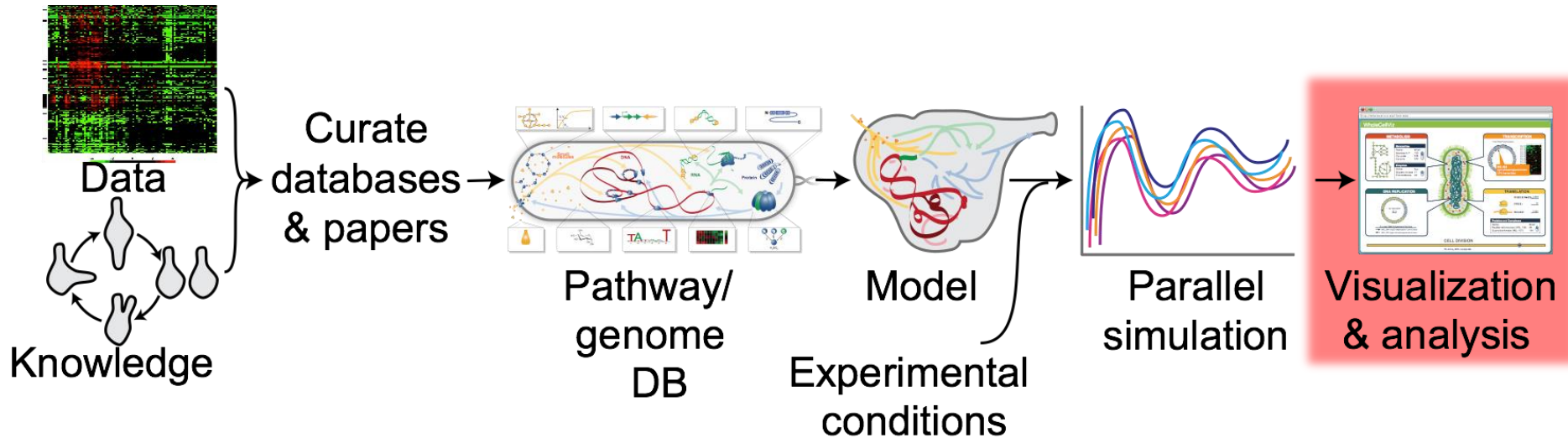
Colorimetric growth assay



Model predictions



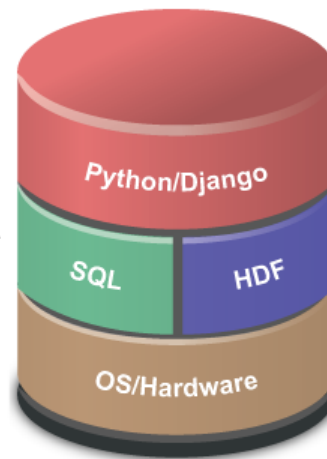
10. Visualization & analysis



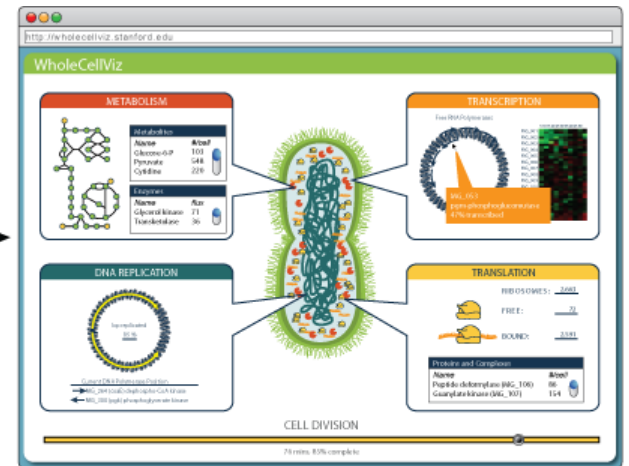
10. Visualization & analysis



Simulation

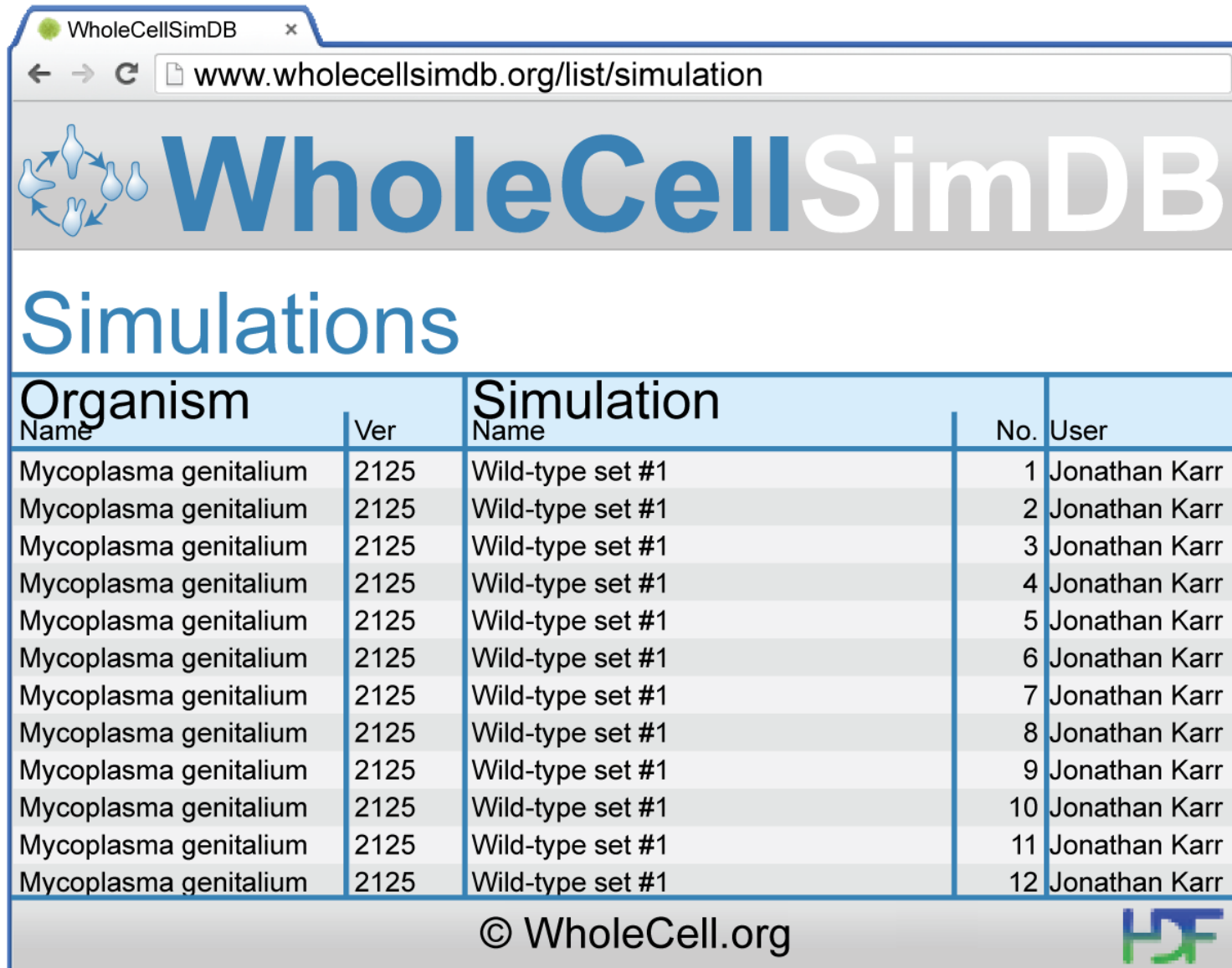


Simulation
database



Visualization & analysis

10. Visualization & analysis

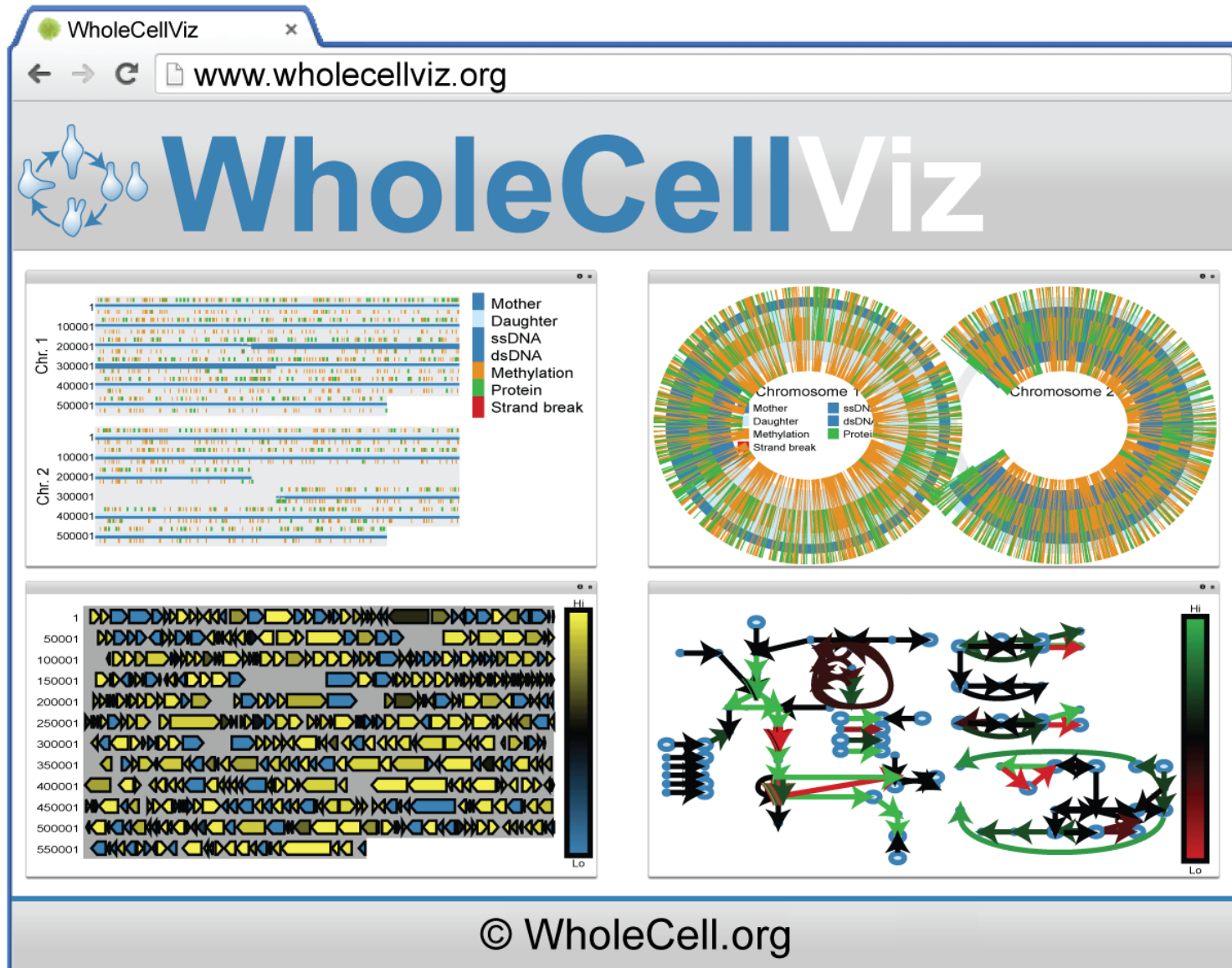


The screenshot shows a web browser window with the address bar displaying www.wholecellsimdb.org/list/simulation. The page features the WholeCellSimDB logo, which includes a diagram of a cell with internal components and arrows indicating interactions. Below the logo, the word "Simulations" is prominently displayed. A table lists various simulations, with columns for Organism Name, Ver, Simulation Name, No., and User. The table contains 12 rows of data, all for Mycoplasma genitalium simulations. At the bottom of the page, there is a copyright notice for WholeCell.org and a logo for the Human Frontier Science Program (HFSP).

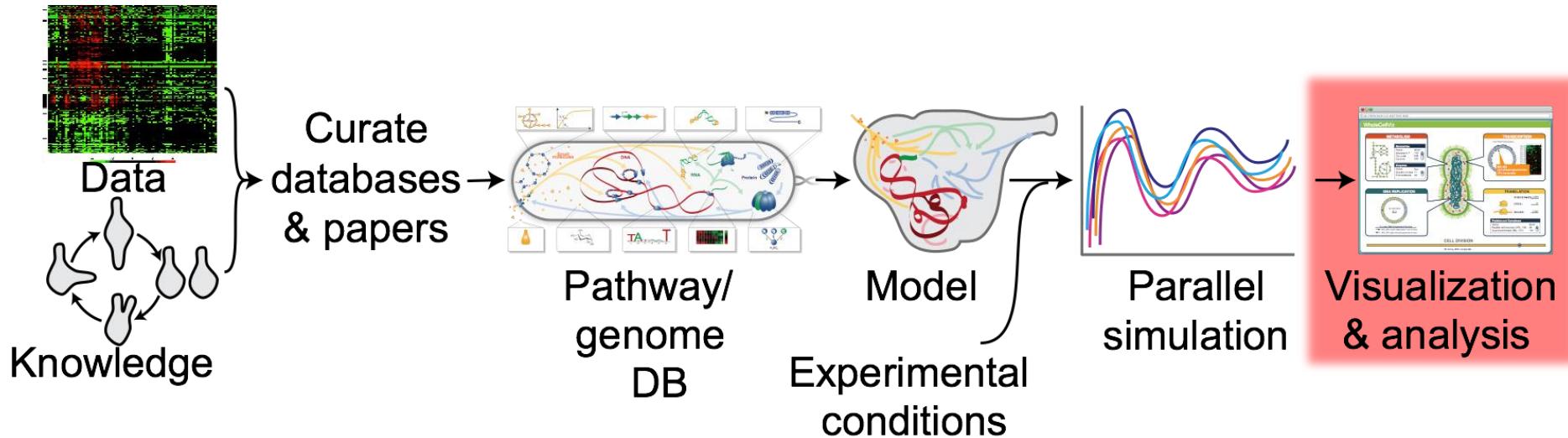
Organism		Simulation		User
Name	Ver	Name	No.	
Mycoplasma genitalium	2125	Wild-type set #1	1	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	2	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	3	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	4	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	5	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	6	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	7	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	8	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	9	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	10	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	11	Jonathan Karr
Mycoplasma genitalium	2125	Wild-type set #1	12	Jonathan Karr

© WholeCell.org

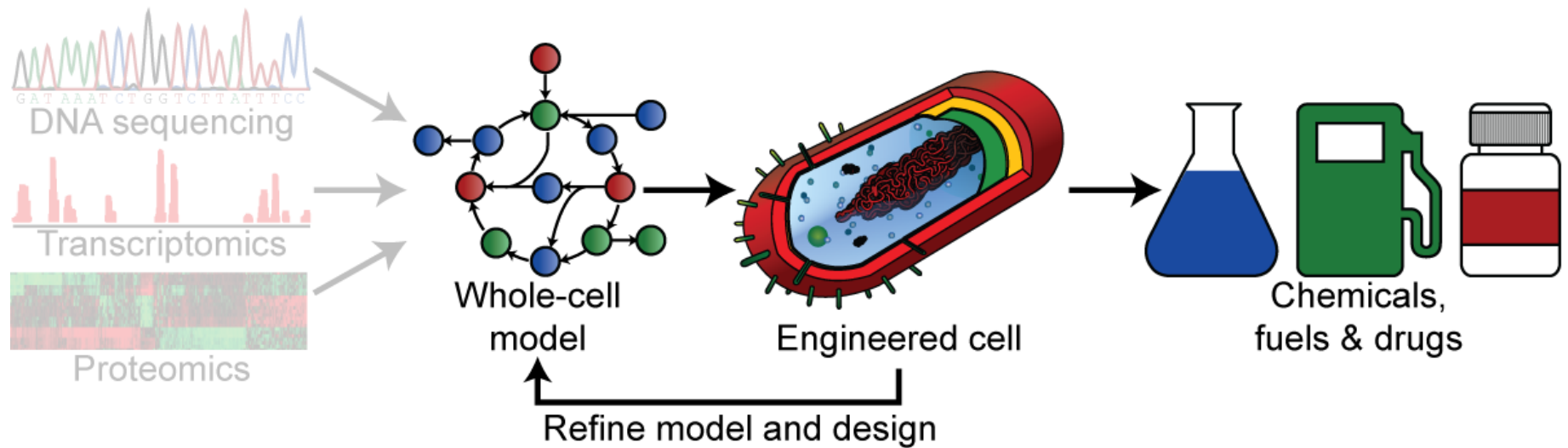
10. Visualization & analysis



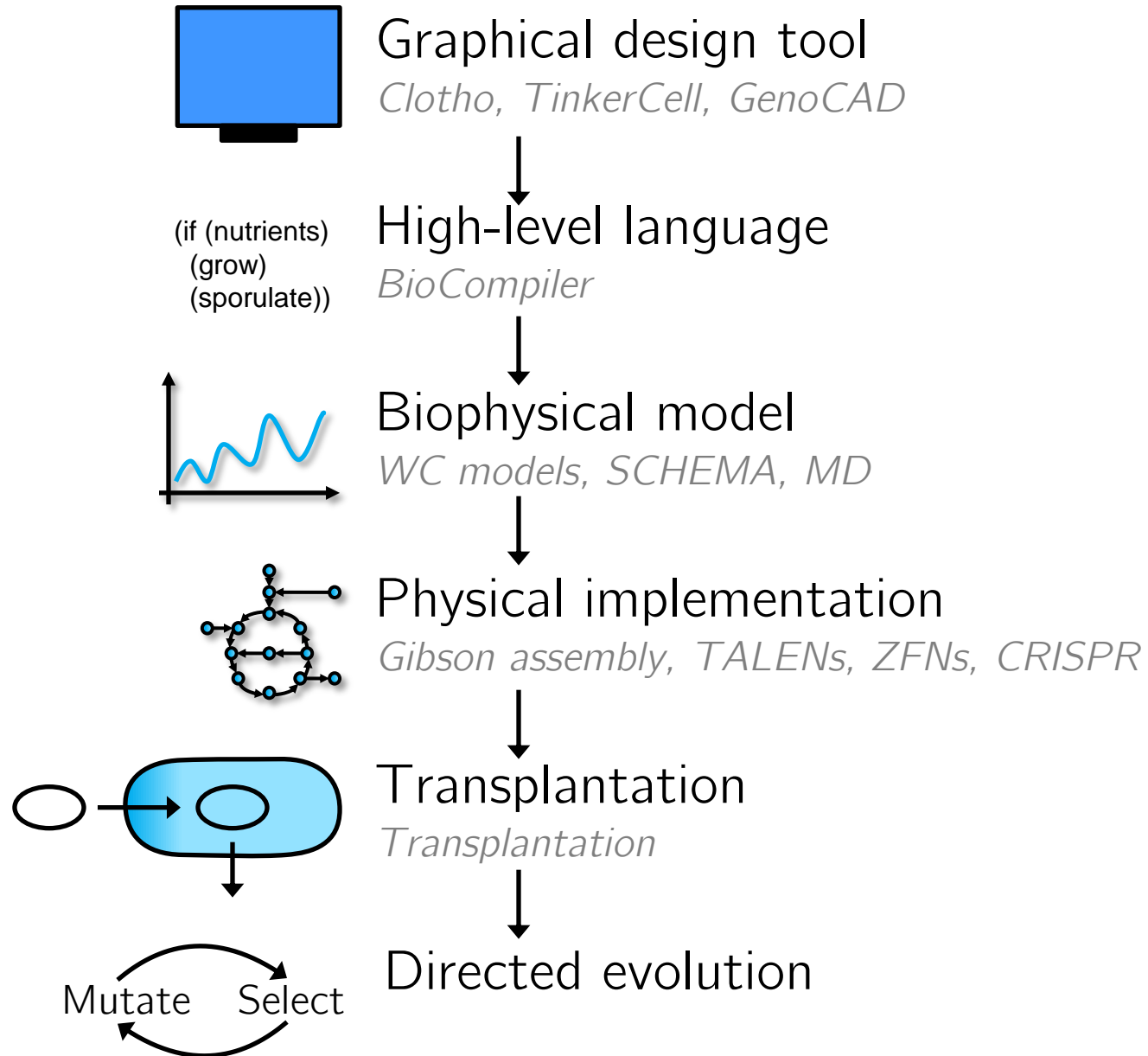
11. Applications: Engineering



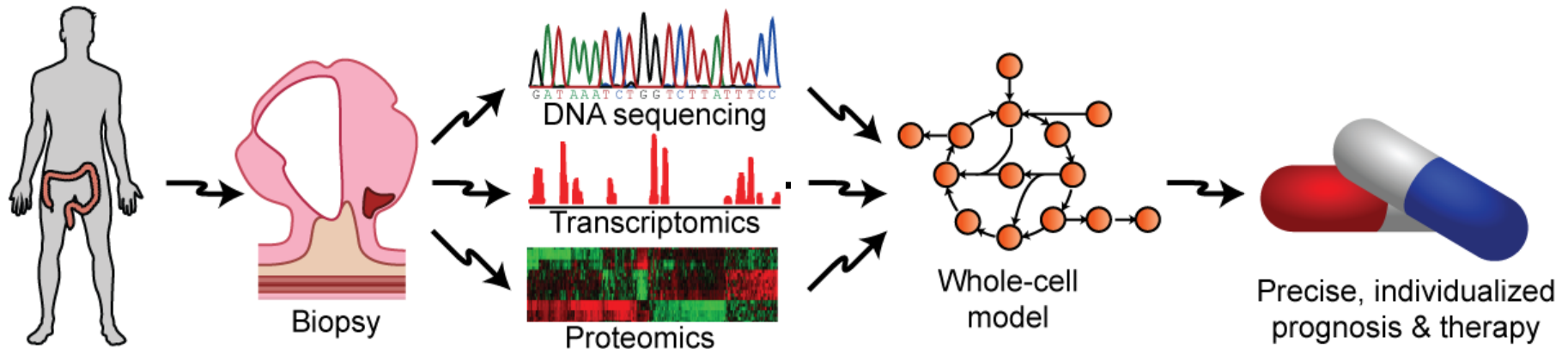
11. Applications: Engineering



11. Applications: Engineering



11. Applications: Medicine



Open challenges

Modeling step	Challenges		
	Computational	Experimental	Community
1. Characterize		• Comprehensive metabolomic, proteomic, kinetic data	• Diversify experimental effort
2. Aggregate data	• Data aggregation software • Natural language processing • Crowdsourced curation		• Annotate data • Deposit raw data
3. Organize data	• Design data model which mirrors models		
4-5. Design models	• Tools to design models from PGDBs		• Standard sequence-based, multi-algorithmic language
6. Simulate model	• Determine how to integrate multi-algorithm models • Develop high-performance simulator		
7. Estimate parameters	• Automate model reduction • Use distributed optimization		
8. Verify model	• Develop test generator • Adopt formal verification • Use continuous integration		
9. Validate model		• Comprehensive single-cell phenomics	• Model validation standard
10. Visualize & analyze	• Improved simulation database • Data exploration tools		
11. Engineer	• Algorithms to optimize predicted phenotypes • Structural integration to design sequences	• Methods for large-scale genome engineering • Design-build-test automation	

Getting involved

Suggested reading

- See school website

Many ongoing projects across the field

- Contact the lecturers

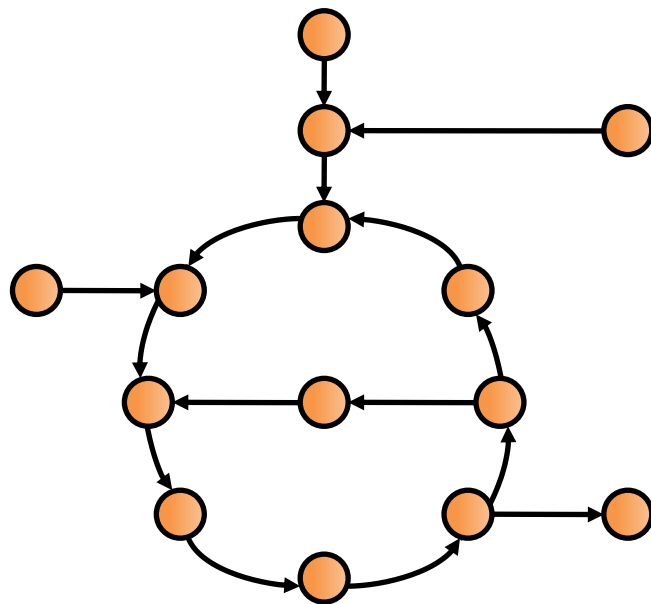
Community benchmark model project

- Contact Jonathan

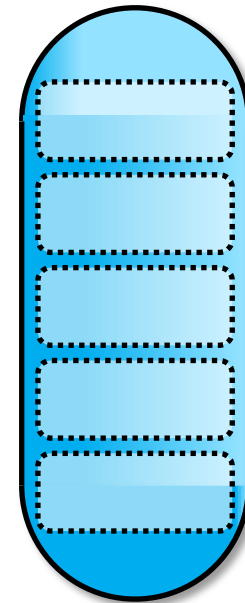
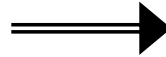
SysMod ISCB special interest group

- SysMod.info
- First meeting @ ISMB, July 2016, Orlando FL

Summary



Data



Integrative model

Broadly **predicts** cell physiology

Integrates heterogeneous data and models

Guides **bioengineering** and **medicine**

Opportunities to **develop improved methods**

Acknowledgements



Anne Marie Barrette



Yin Hoon Chew



Arthur Goldberg



Graeme Gossel



Pablo Meyer
IBM/Sinai



Roger Rodriguez
UNAM

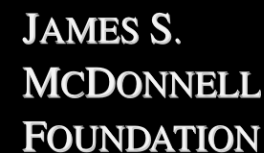


Center for Regulatory Genomics

Luis Serrano

Maria Lluch-Senar

Veronica Llorens



Mount
Sinai