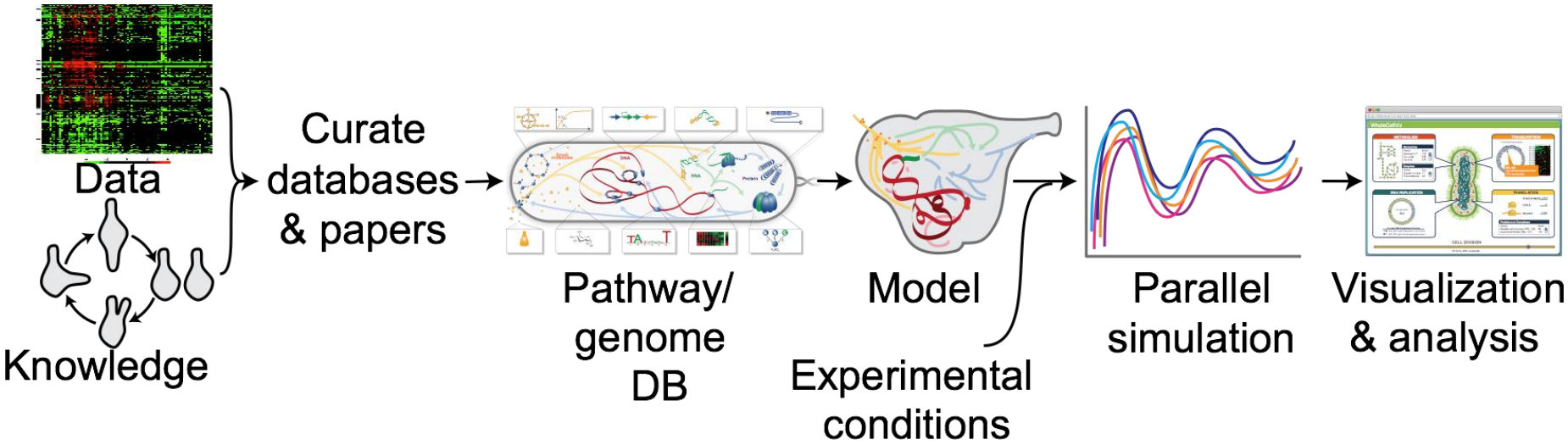
Systemizing and accelerating whole-cell modeling



Jonathan Karr
April 8, 2016

Mount Sinai

KarrLab.org
karr@mssm.edu

WC modeling principles

Single-cell

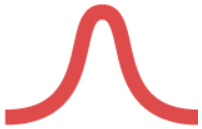
Temporally complete

GATCCA

Species-specific

Mechanistic/dynamic

Genetically complete



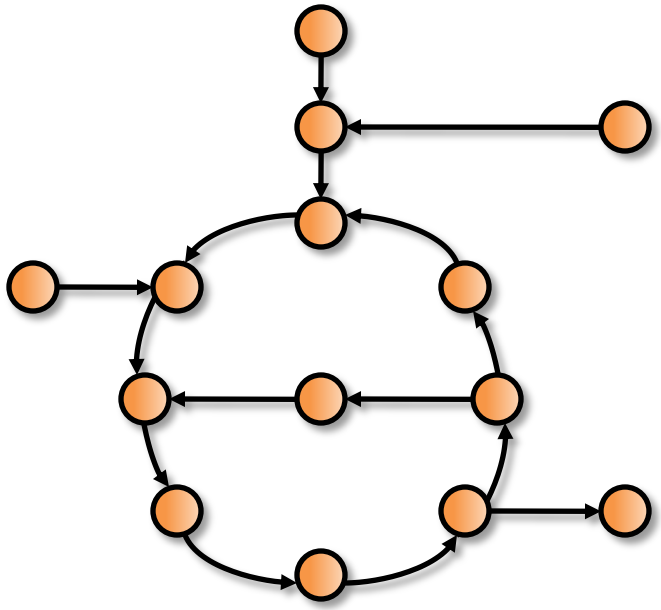
Stochastic

Molecularly precise

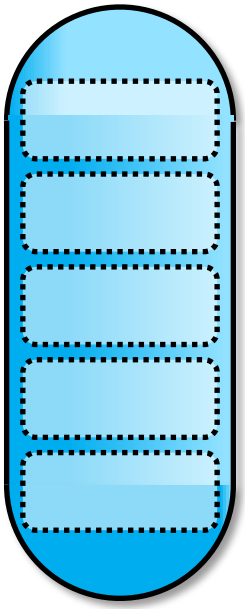
Accurate

Karr et al., 2015

WC modeling = genomics + integrative modeling

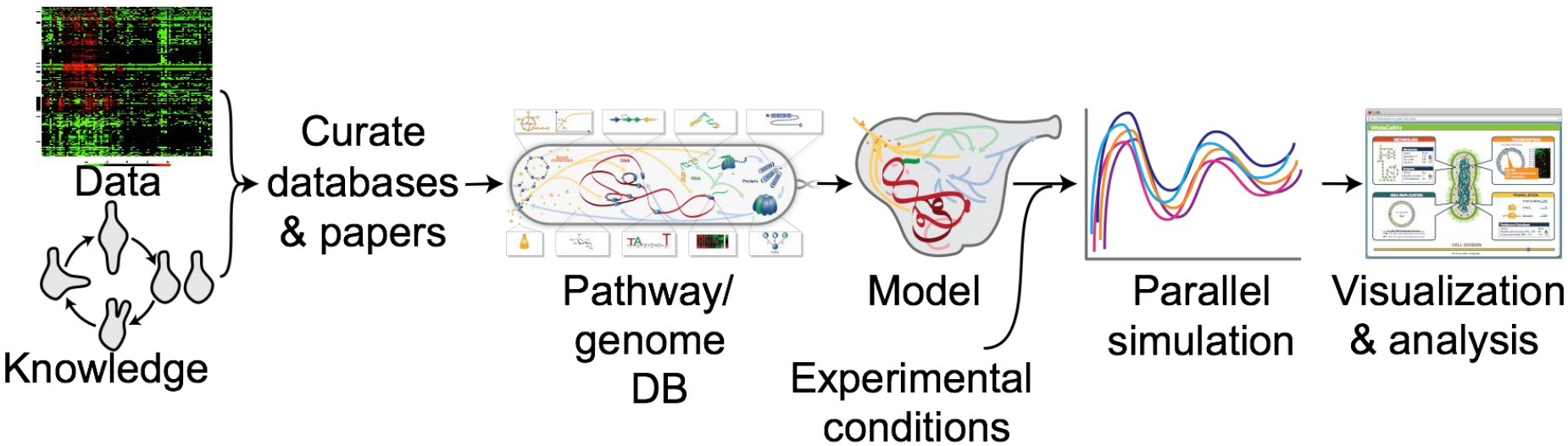


Molecular data



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

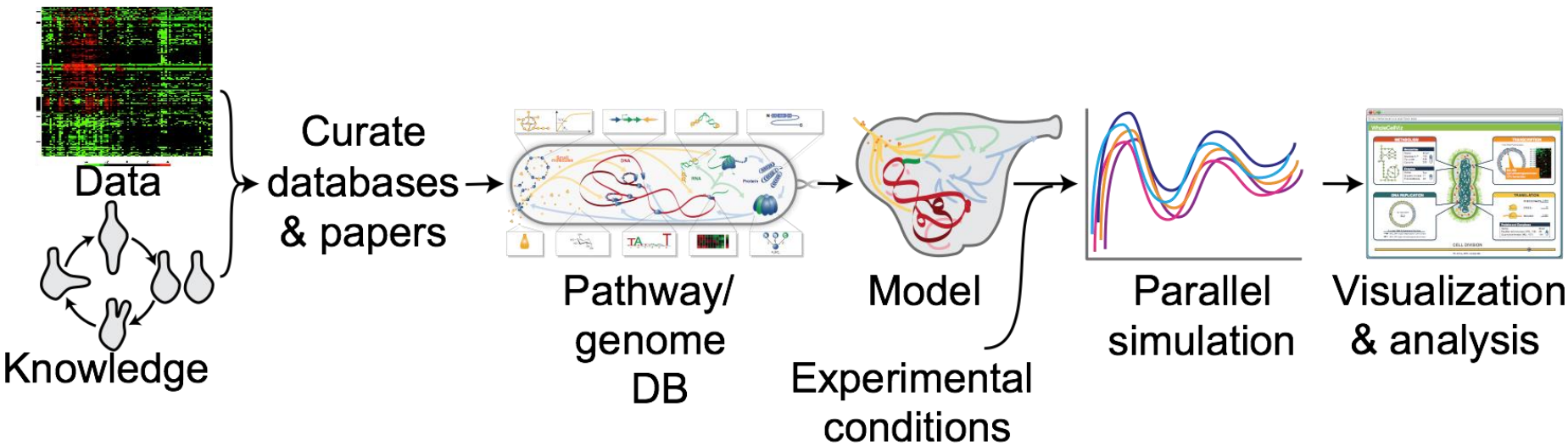
- Multi-algorithm simulation
- Parameter estimation
- Verification
- Model reduction

Time-consuming to construct

- Curate data
- Design model
- Verify model

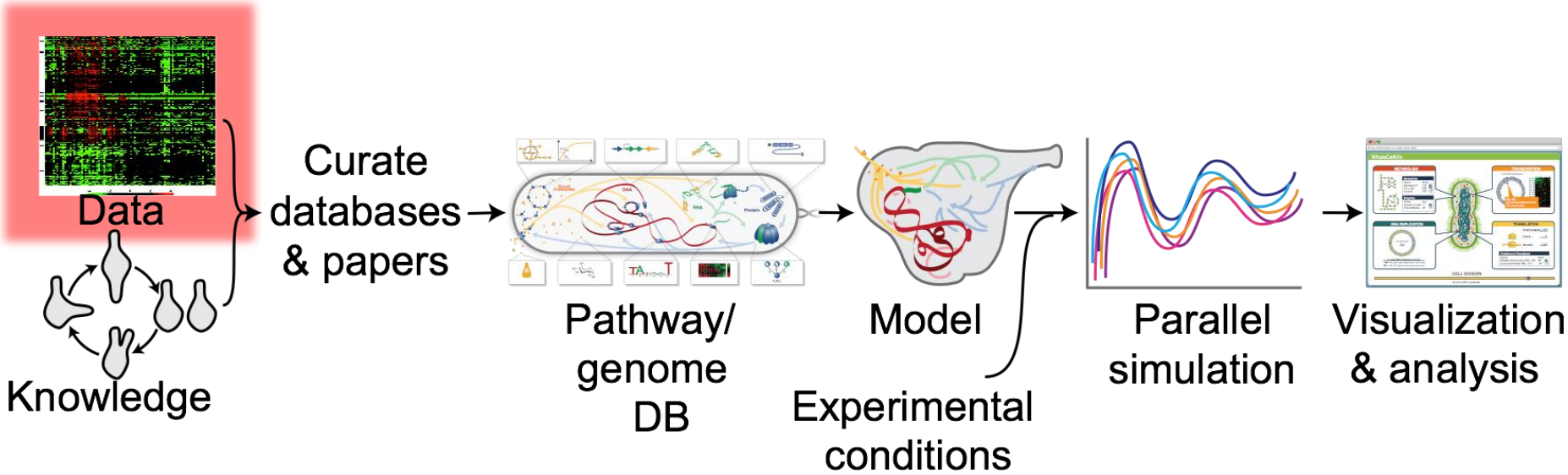
Hard to understand, reuse, reproduce

Systemizing and accelerating WC modeling



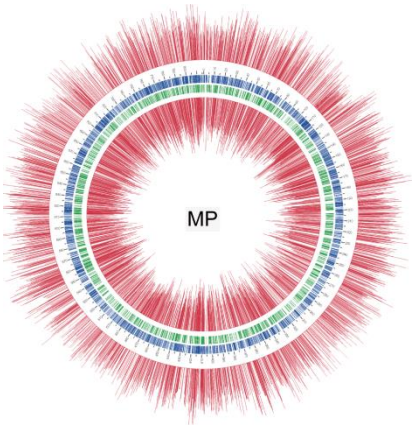
1. Characterize organism
2. Aggregate data
3. Organize data
4. Design submodels
5. Merge submodels
6. Simulate model
7. Estimate parameters
8. Verify model
9. Validate model
10. Visualize/analyze predictions
11. Applications: Engineering, medicine

1. Characterize organism

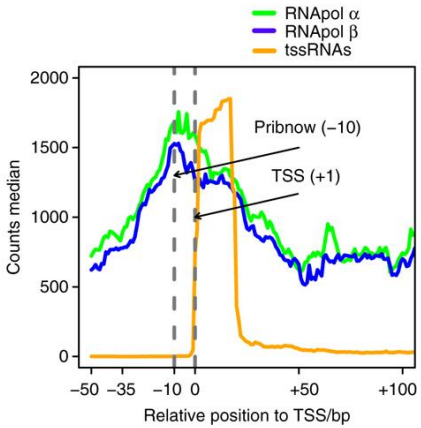


1. Characterize organism

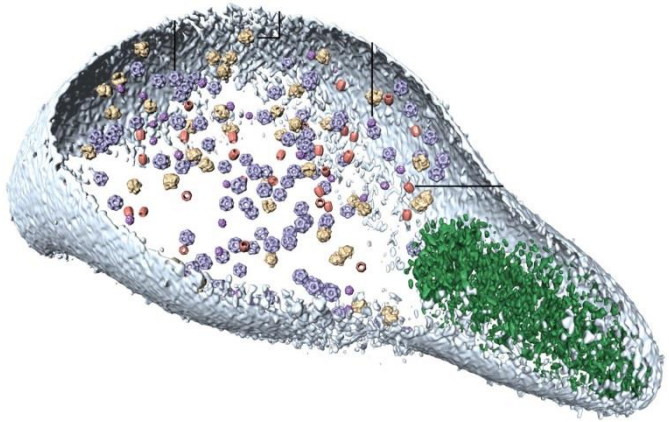
Genome
DNA-seq



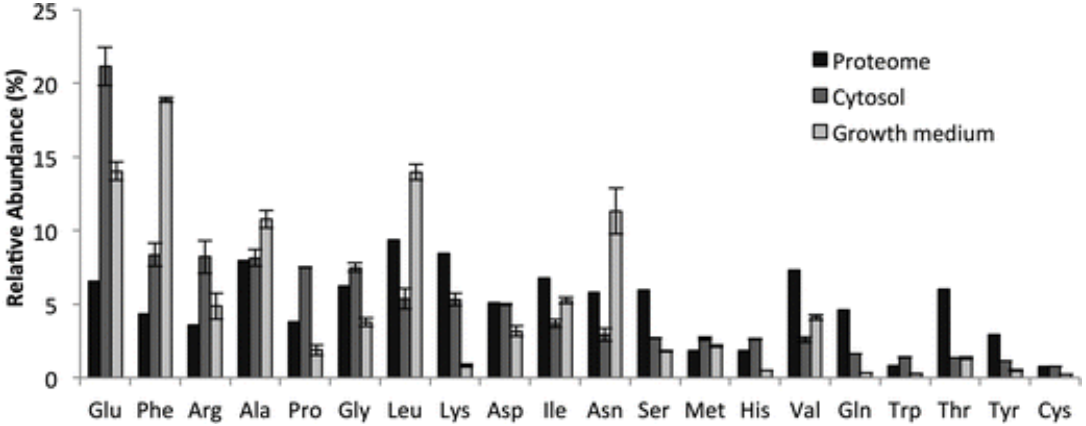
Epigenome
Meth-seq



Transcriptome
RNA-seq



Proteome
Mass-spectrometry



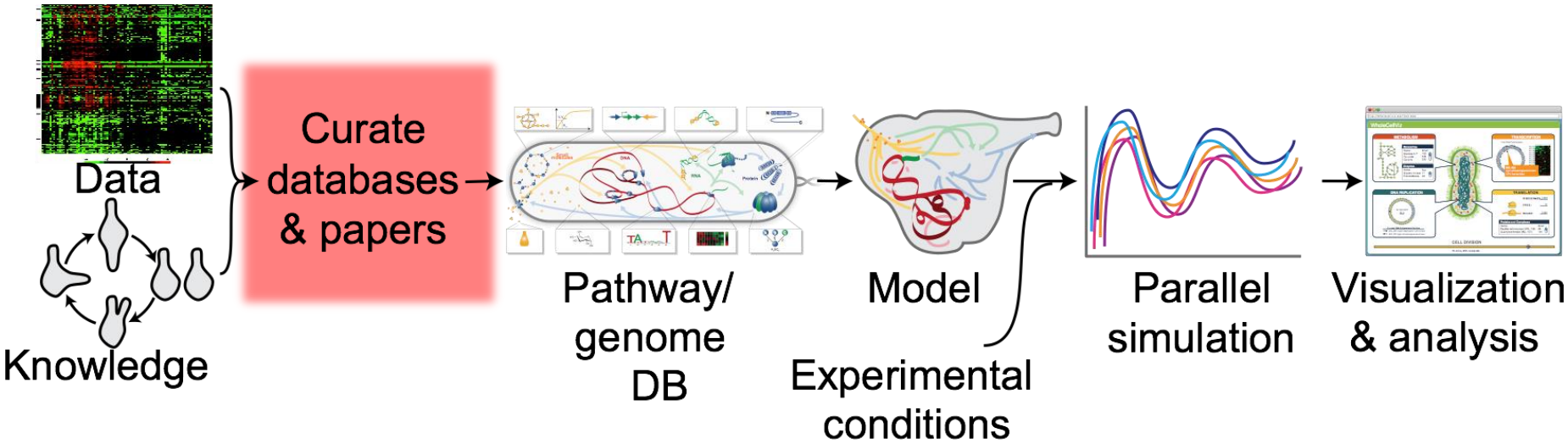
Metabolome
Mass-spectrometry

Fraser et al., 1995; Kühner et al., 2009; Lluch-Senar et al., 2013; Maier et al., 2013; Yus et al. 2012

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

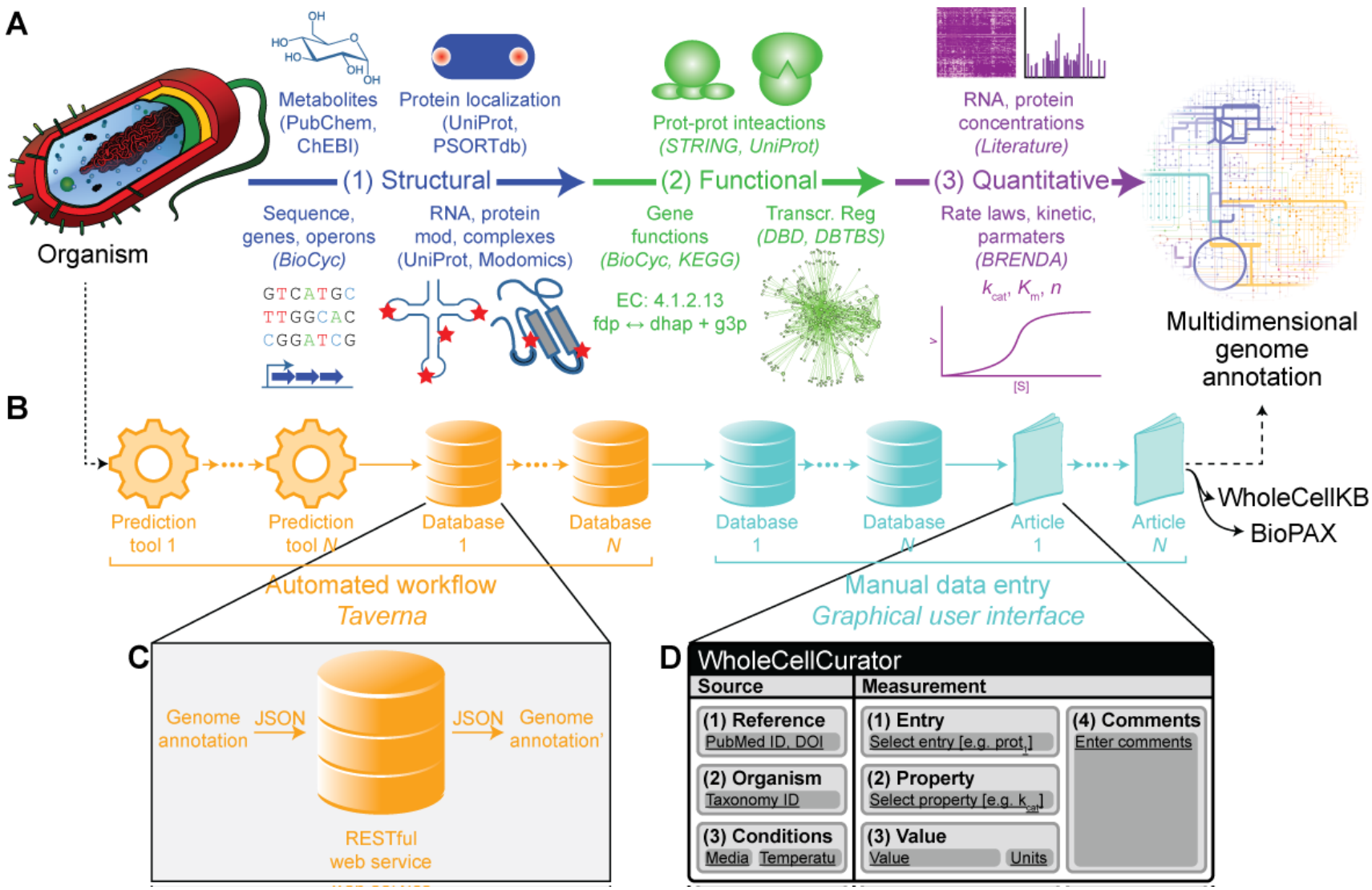
2. Aggregate data



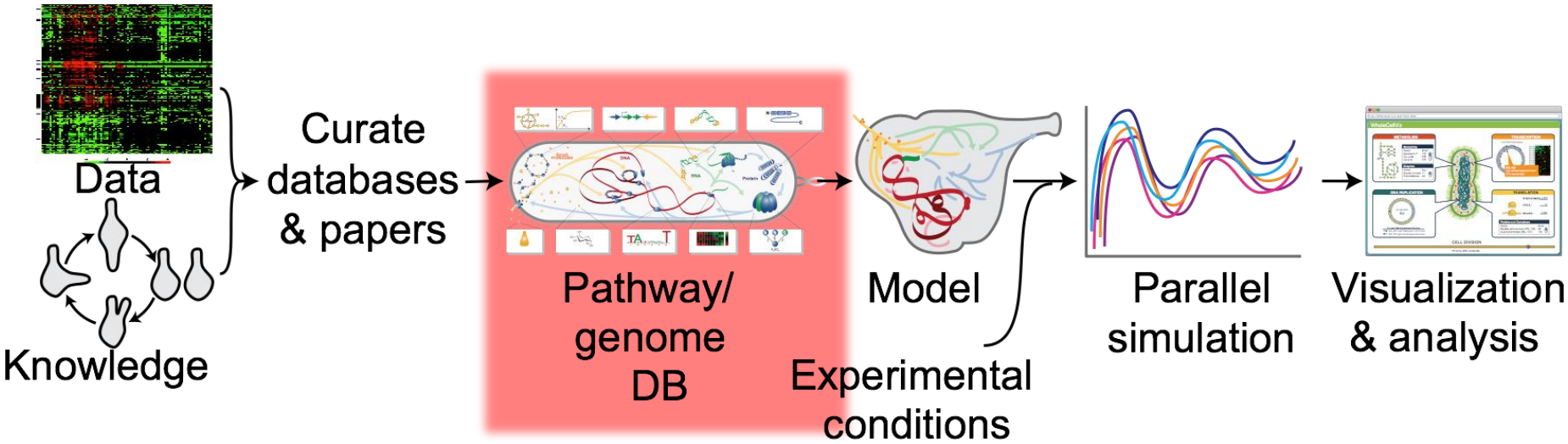
2. Aggregate data

	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

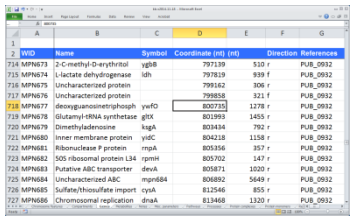


3. Organize data

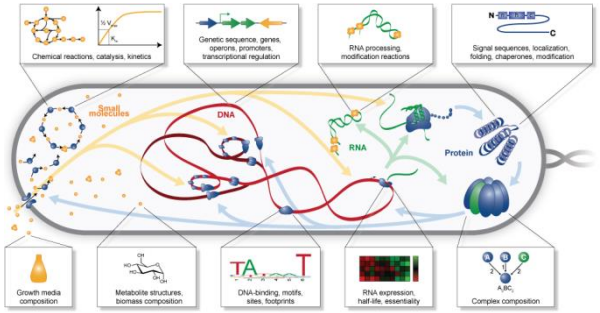


3. Organize data

Aggregate data: Scripts, Excel



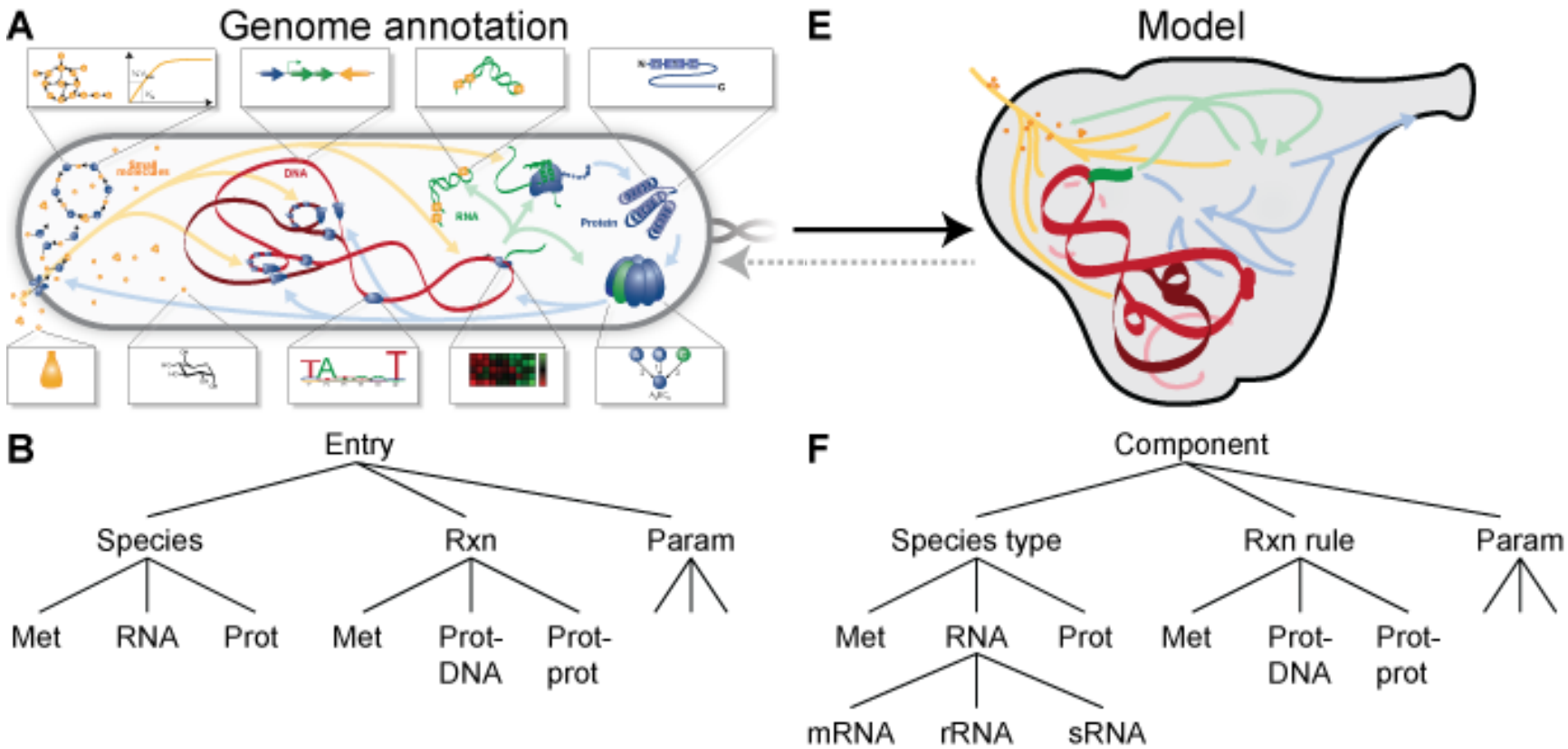
PGDB: Pathway Tools, WC-KB



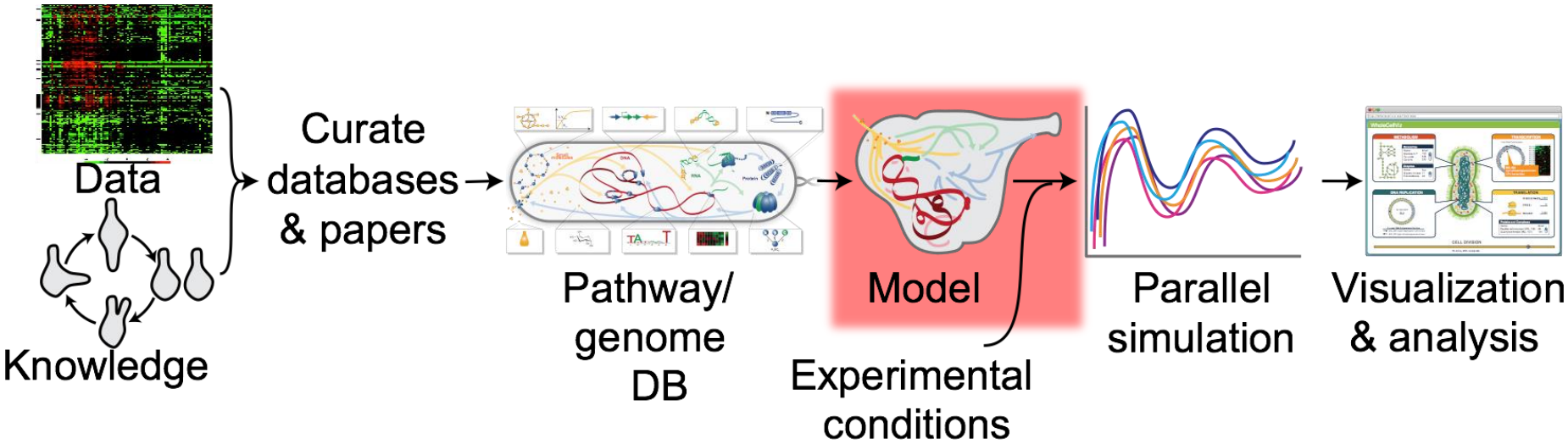
PGDB Browser

Caspi et al., 2016

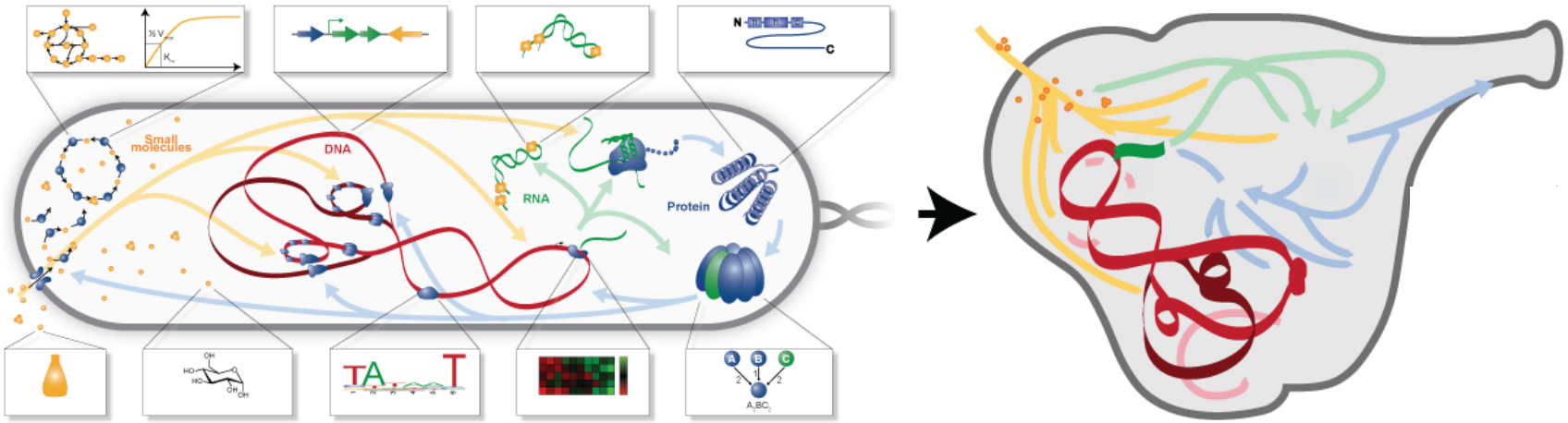
3. Organize data



4. Design submodels



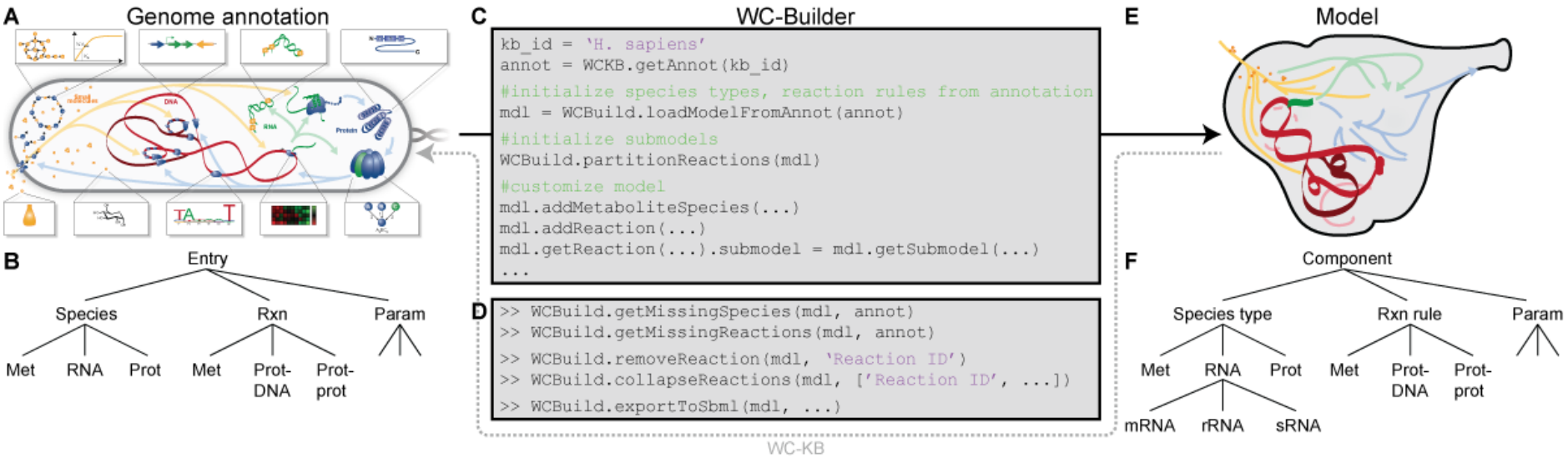
4. Design submodels



Pathway/genome database

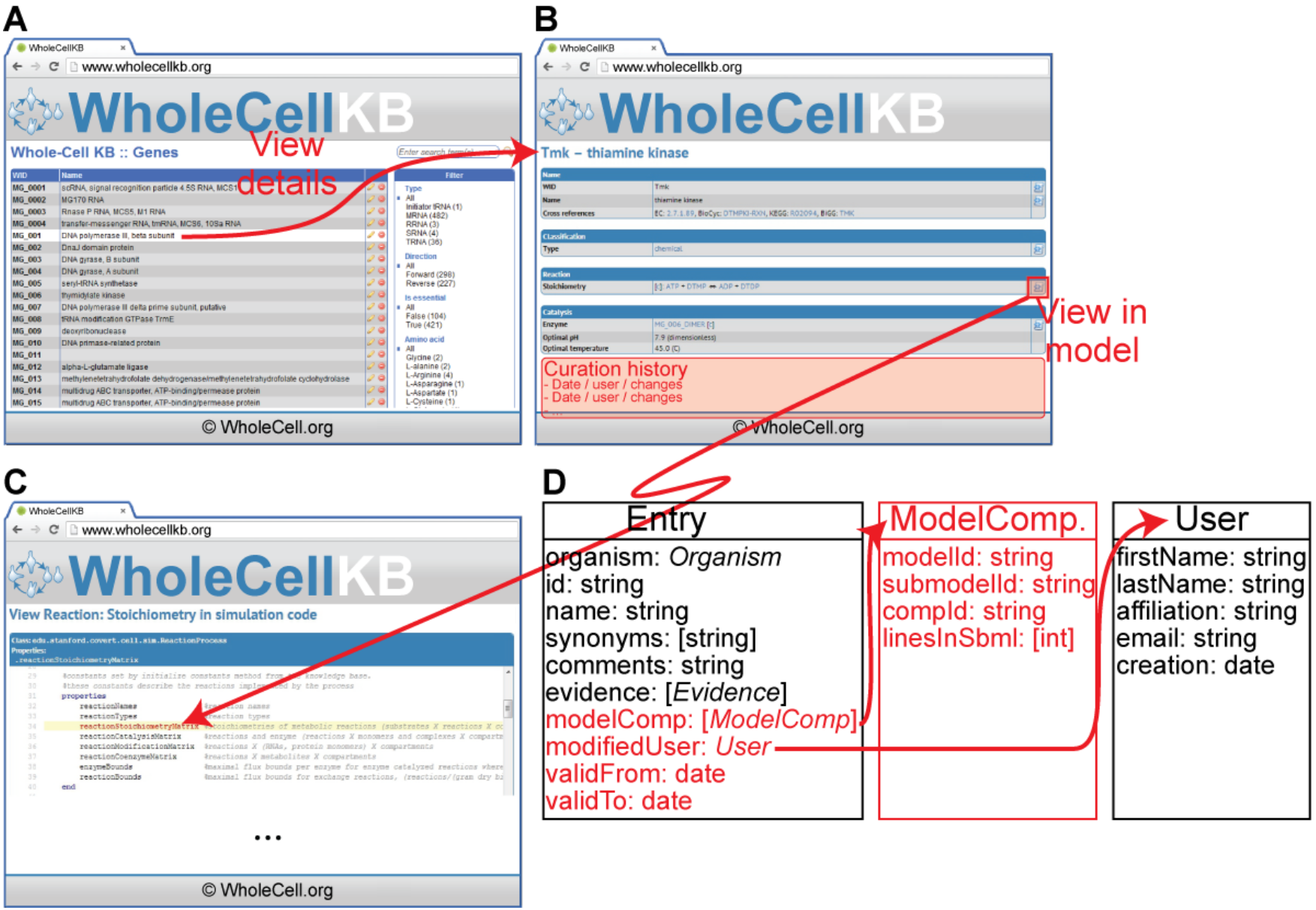
Submodels

4. Design submodels



WC-ML, SBML, CellML

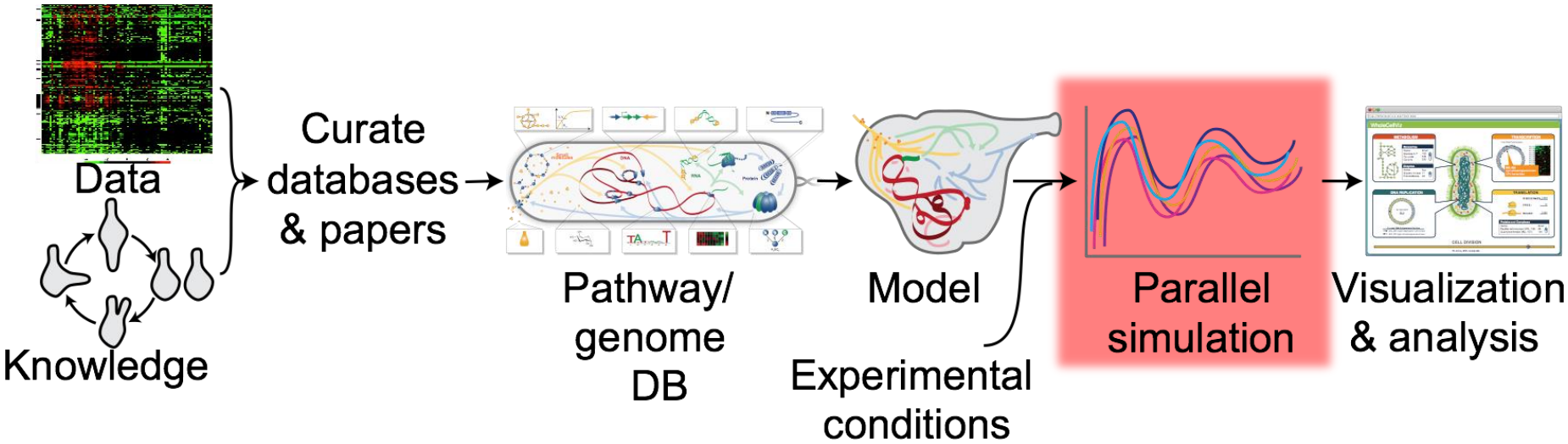
4. Design submodels



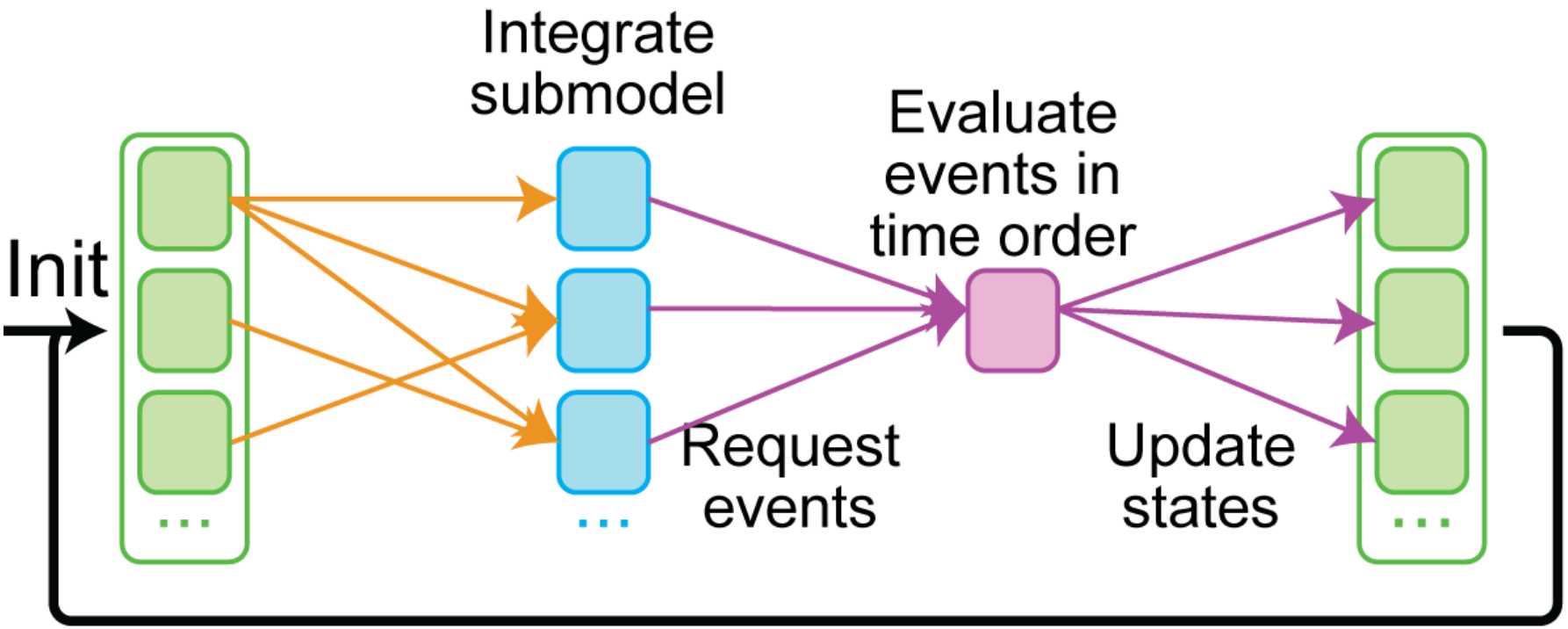
5. Map submodels onto common state

Automatically handled by designing submodels from common PGDB

6. Simulate


Data
Knowledge
Curate databases & papers
Pathway/ genome DB
Model
Experimental conditions
Parallel simulation
Visualization & analysis

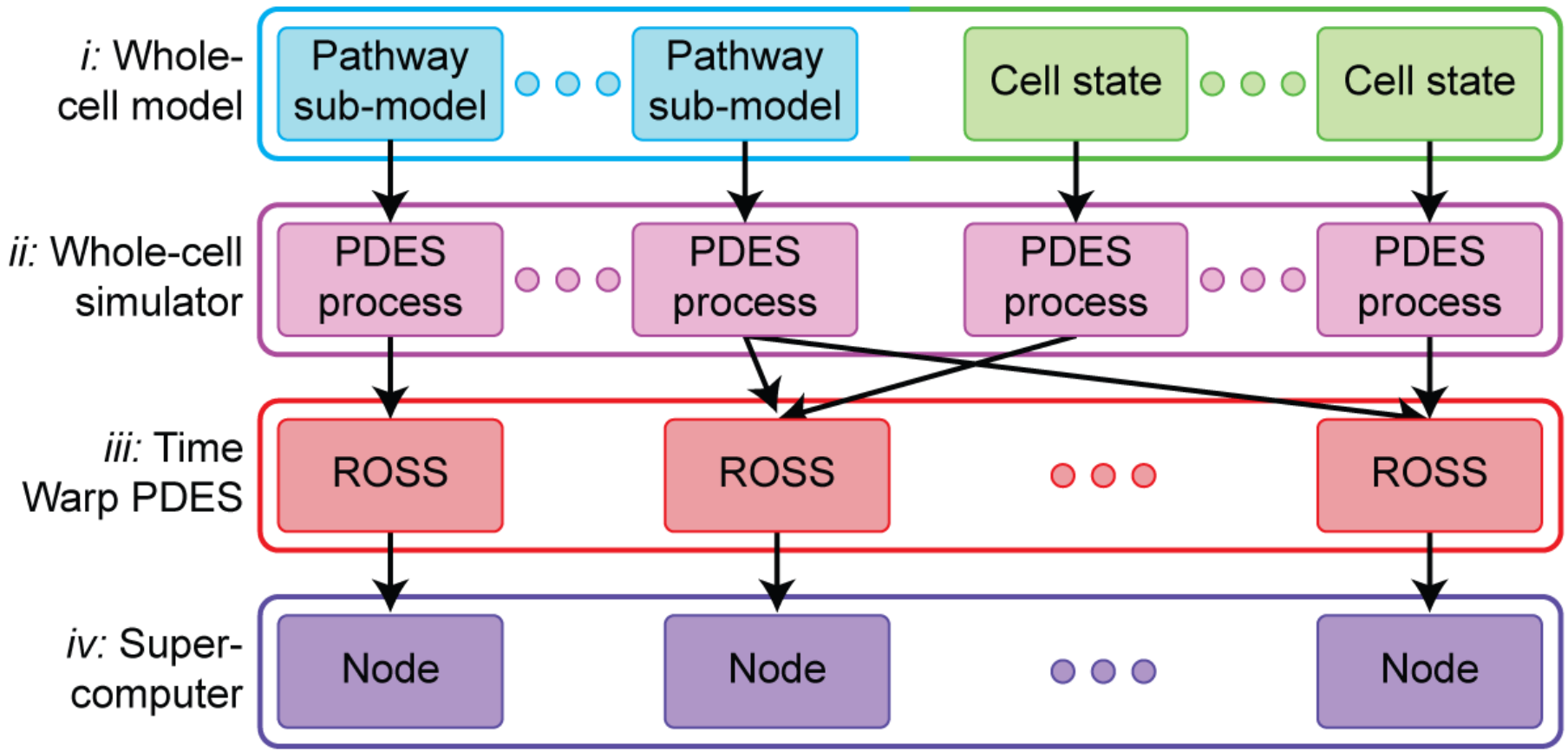
6. Simulate: Concurrently integrate submodels



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

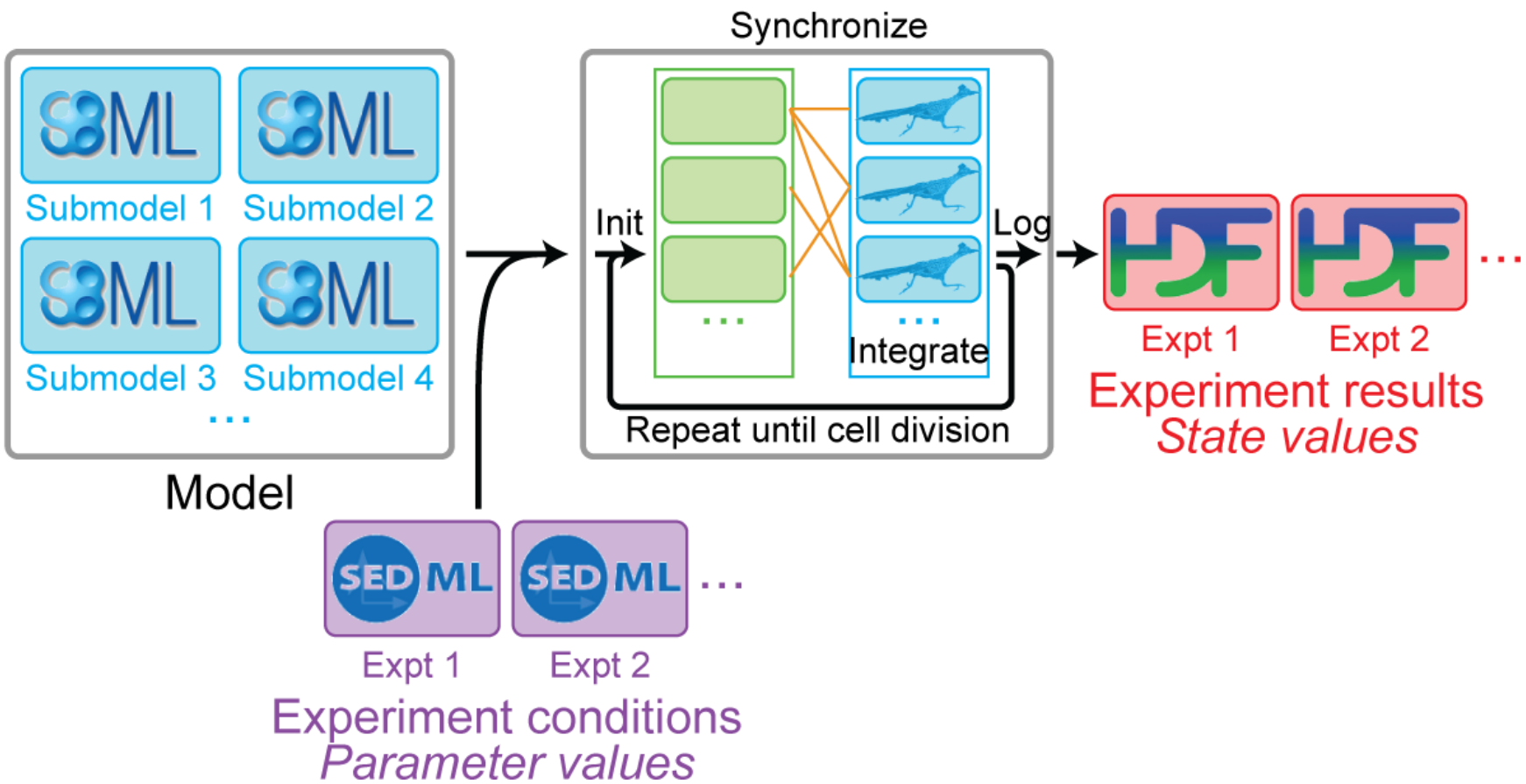
Goldberg et al., 2016

6. Simulate: Parallel discrete event simulation

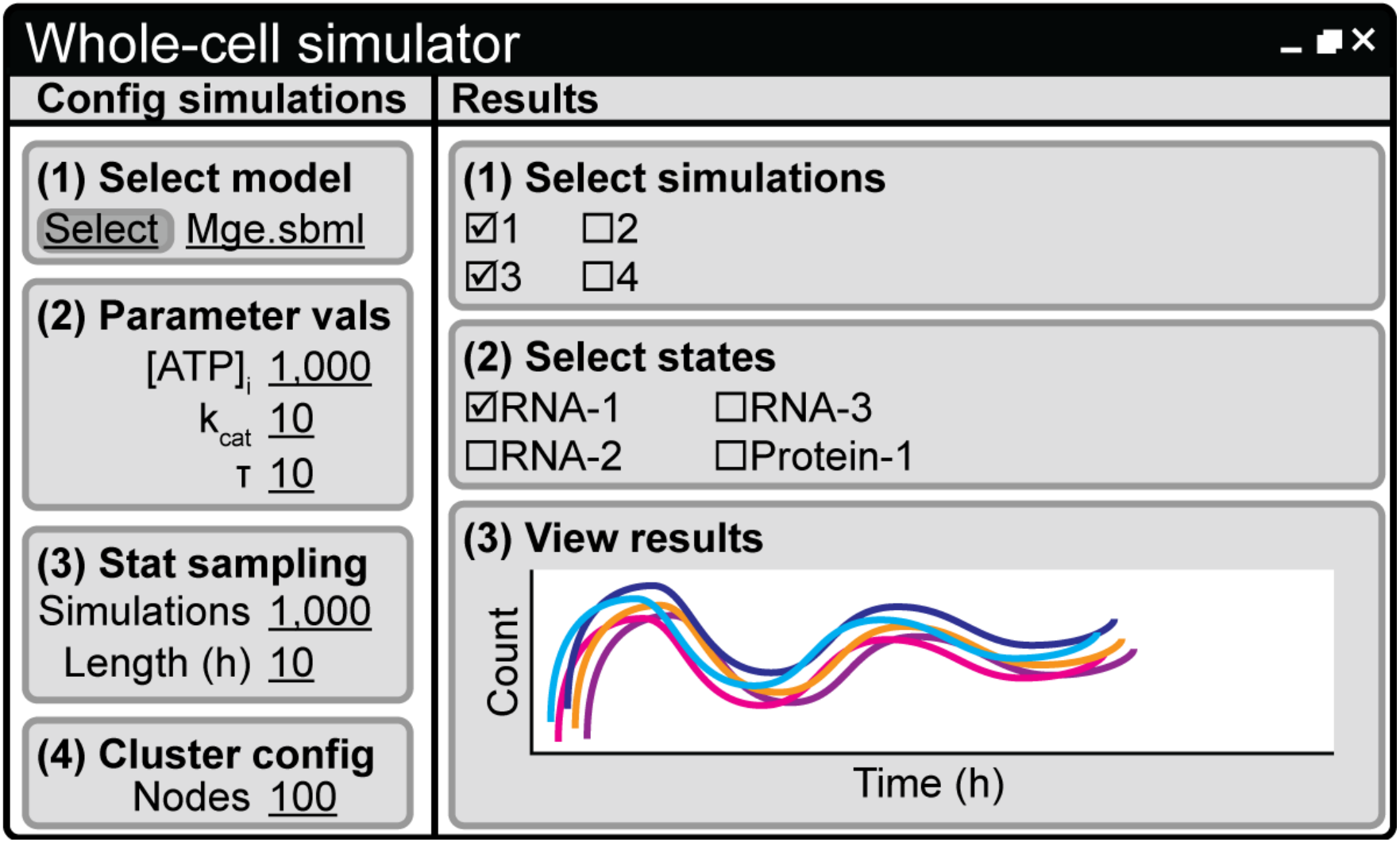


Goldberg et al., 2016

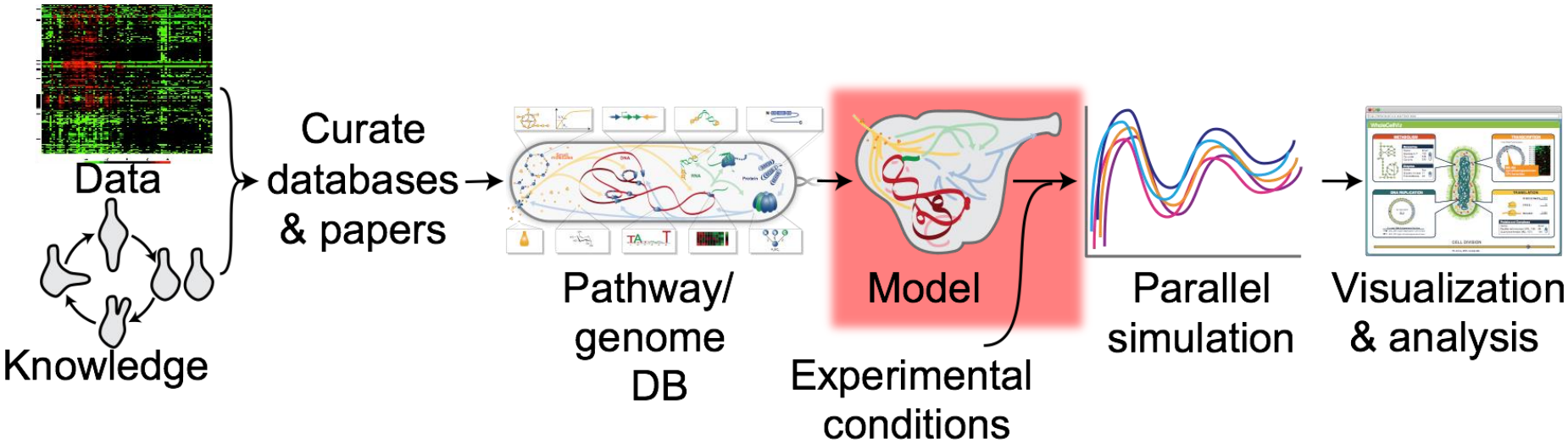
6. Simulate: High-performance simulator



6. Simulate: High-performance simulator

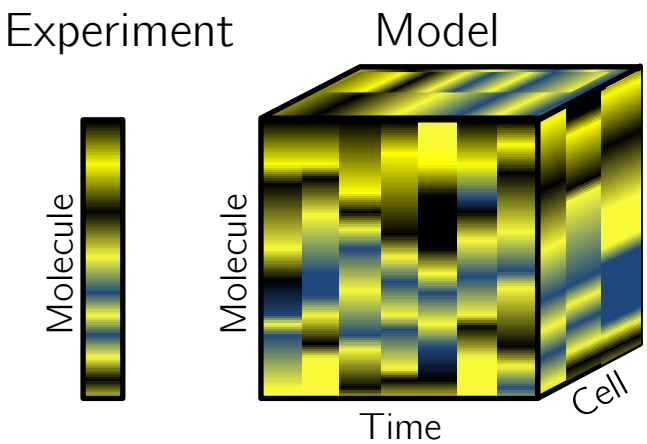


7. Estimate parameters



7. Estimate parameters

1. **Reduce model**



2. **Identify reduced model parameters using traditional methods**

3. **Manually tune parameters using full model**

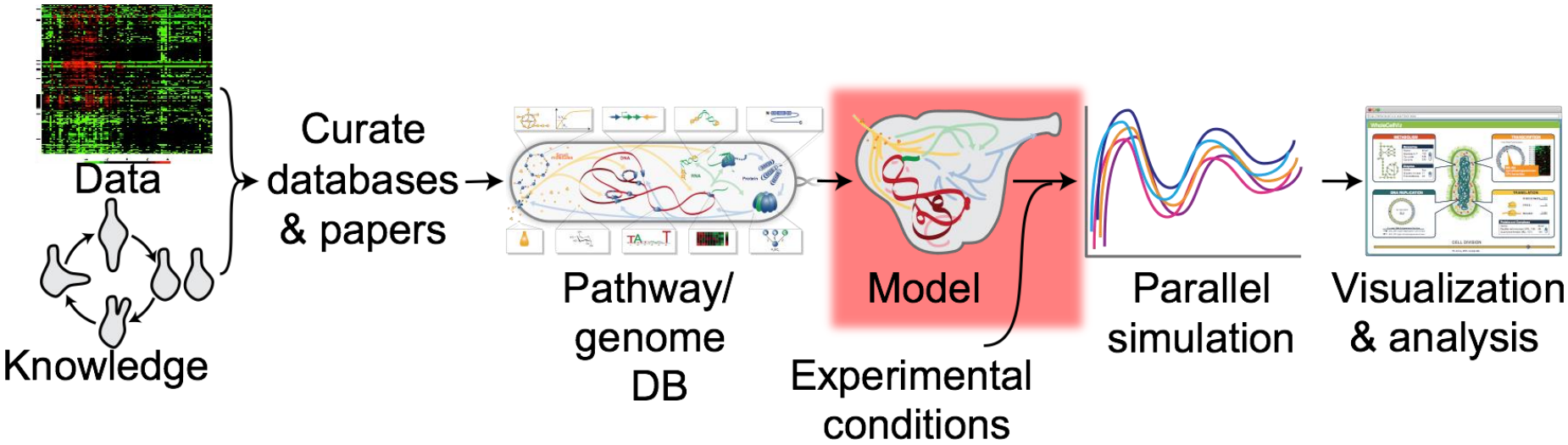
Karr et al., 2015

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

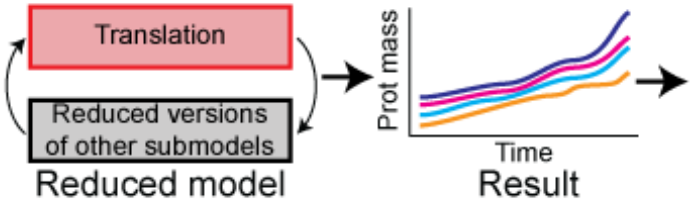
- ☑ Plot model predictions
- ☑ Manually inspect data
- ☑ Compare to known biology

8. Verify model reproduces known biology

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

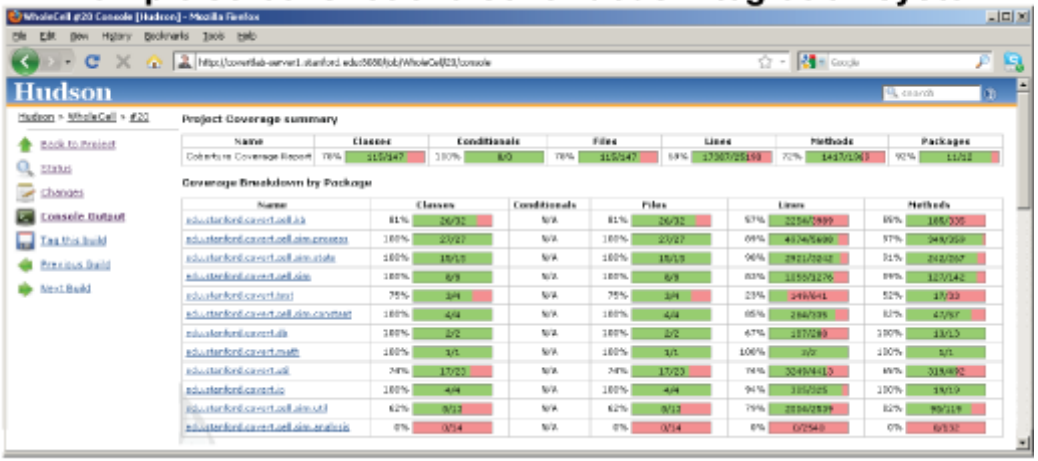
- Mass balance
- Charge balance
- Consistent localization
  - Small molecules
  - Enzymes
- Reactants and products can be produced/recycled by metabolism submodel
- Sufficient small molecule and enzymatic resources to support growth

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

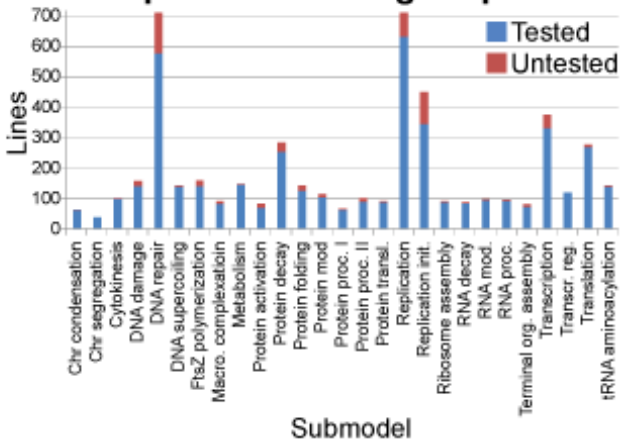


- Protein mass duplicates over cell cycle
- Protein complex duplicates over cell cycle
- No excess accumulation AA
- All gene products that are essential to the modeled pathways are translated

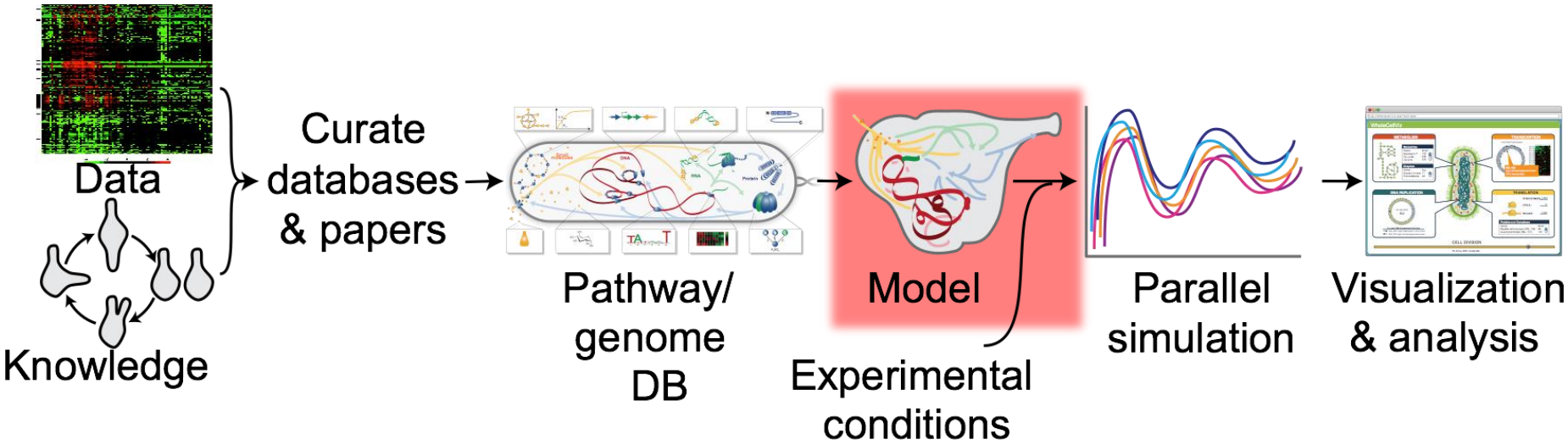
D. Example screenshot of a continuous integration system



E. Example test coverage report



9. Validate model reproduces *true* biology

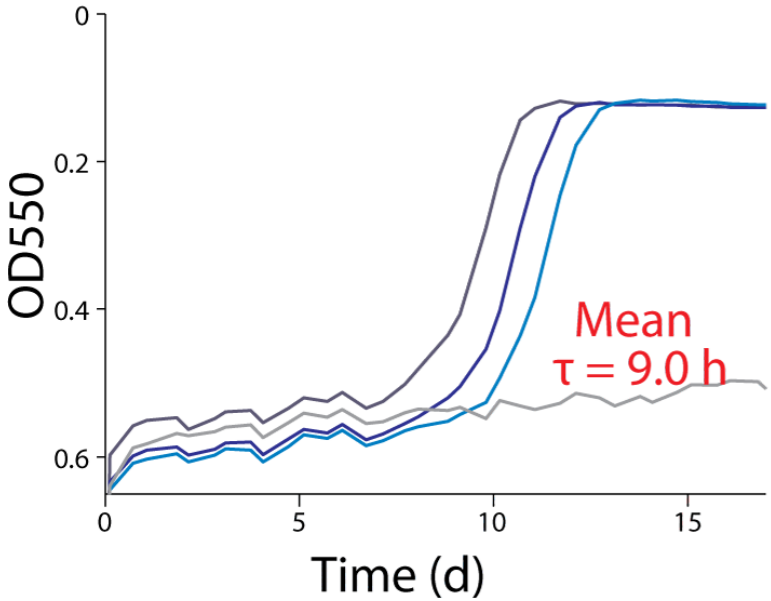


9. Validate model reproduces *true* biology

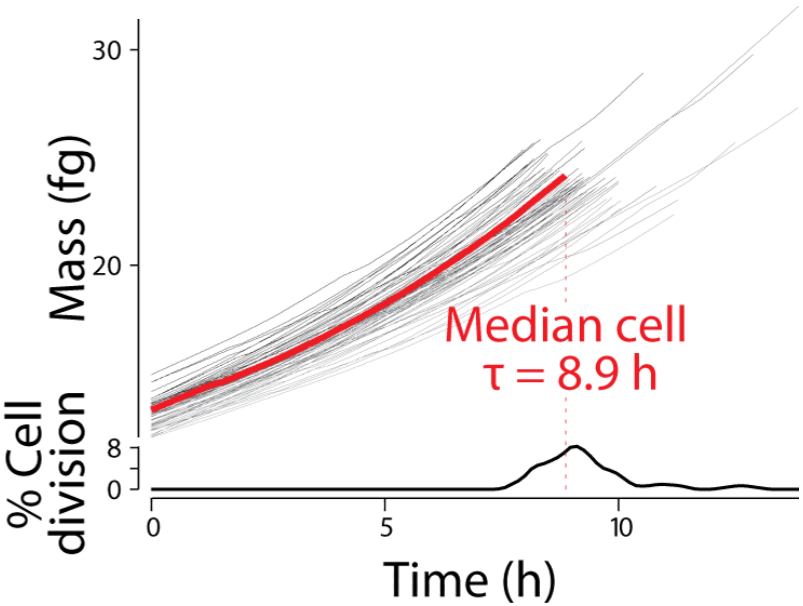
- ☑ **Matches new data**
  - ☑ Wild-type growth rate
  - ☑ Disruption strain growth rates
  - ☑ Single-cell division times
  - ☑ Single-cell cell cycle phase lengths
  - ☑ Single-cell sizes

9. Validate model reproduces *true* biology

Colorimetric growth assay

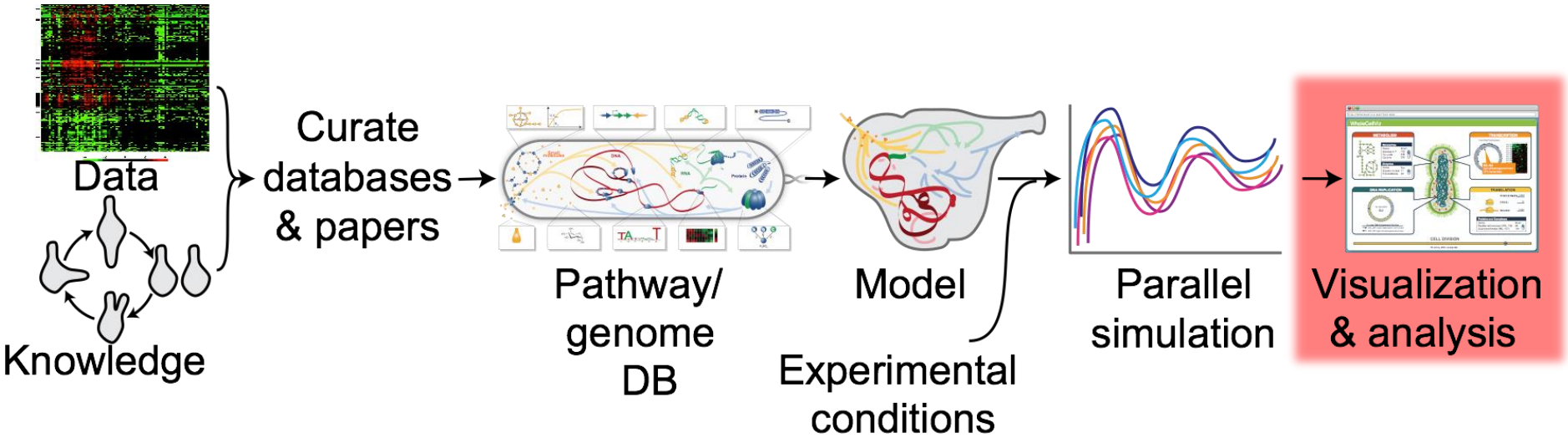


Model predictions

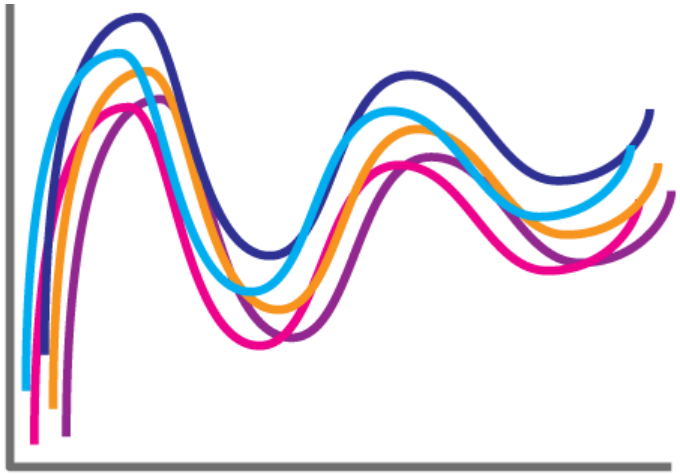


Karr et al., 2012

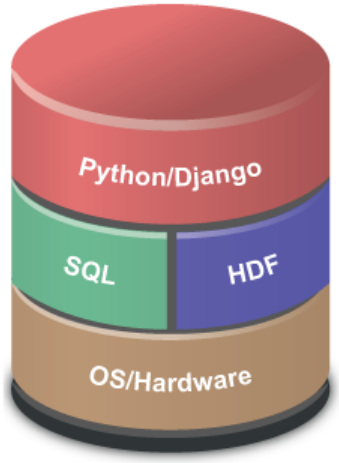
10. Visualization & analysis



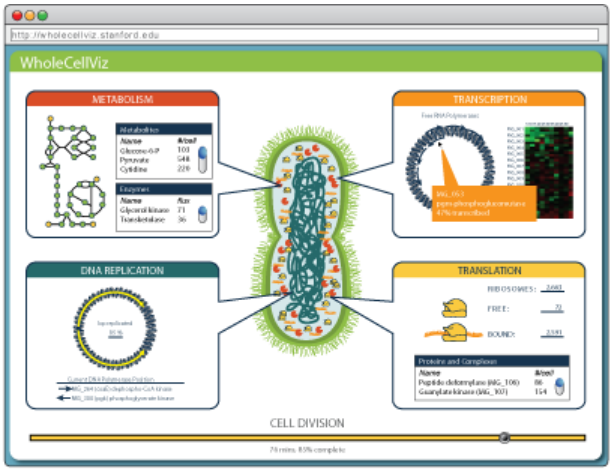
10. Visualization & analysis



Simulation



Simulation database



Visualization & analysis

10. Visualization & analysis

WholeCellSimDB

www.wholecellsimdb.org/list/simulation

WholeCellSimDB

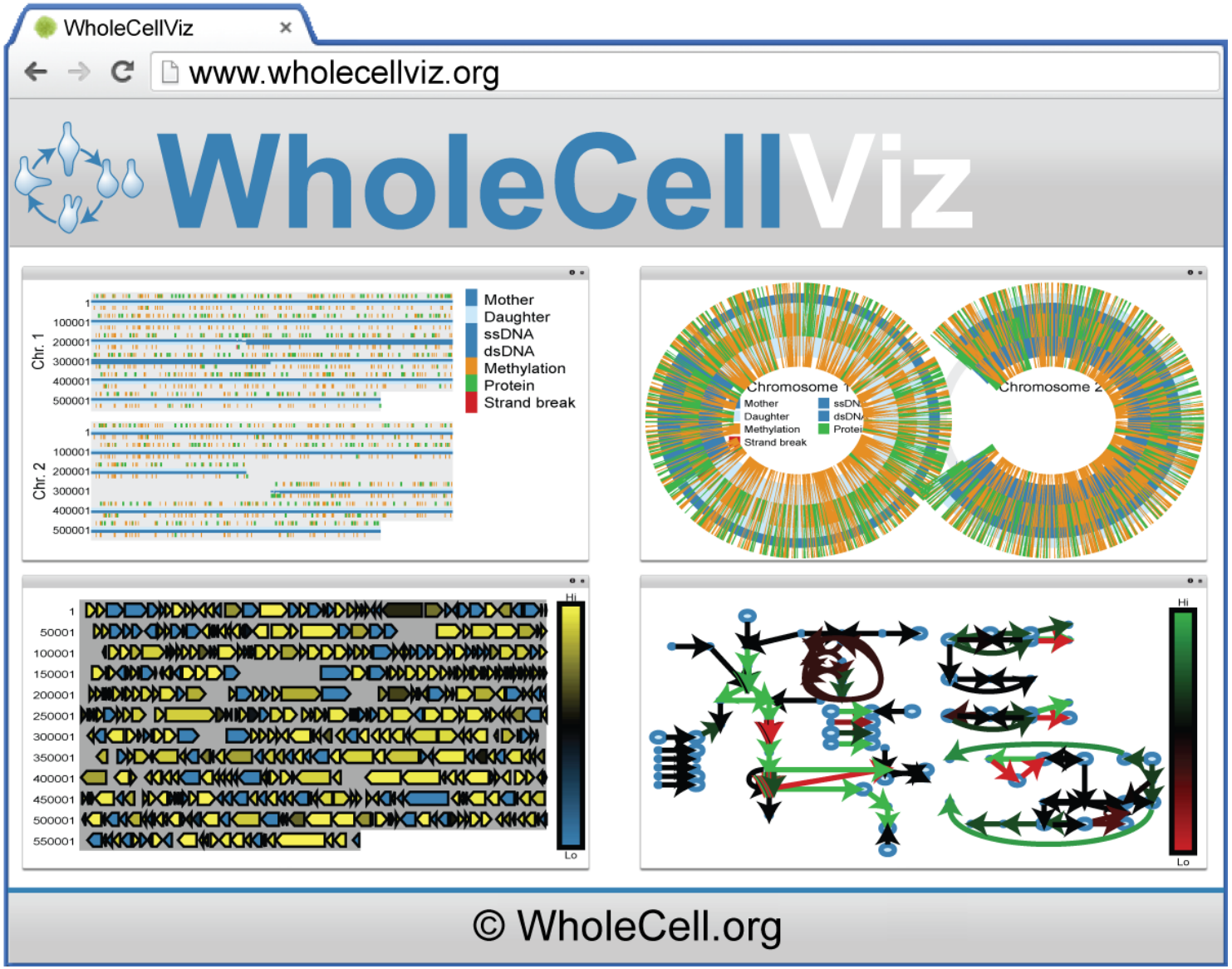
Simulations

Organism Name	Ver	Simulation Name	No.	User
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

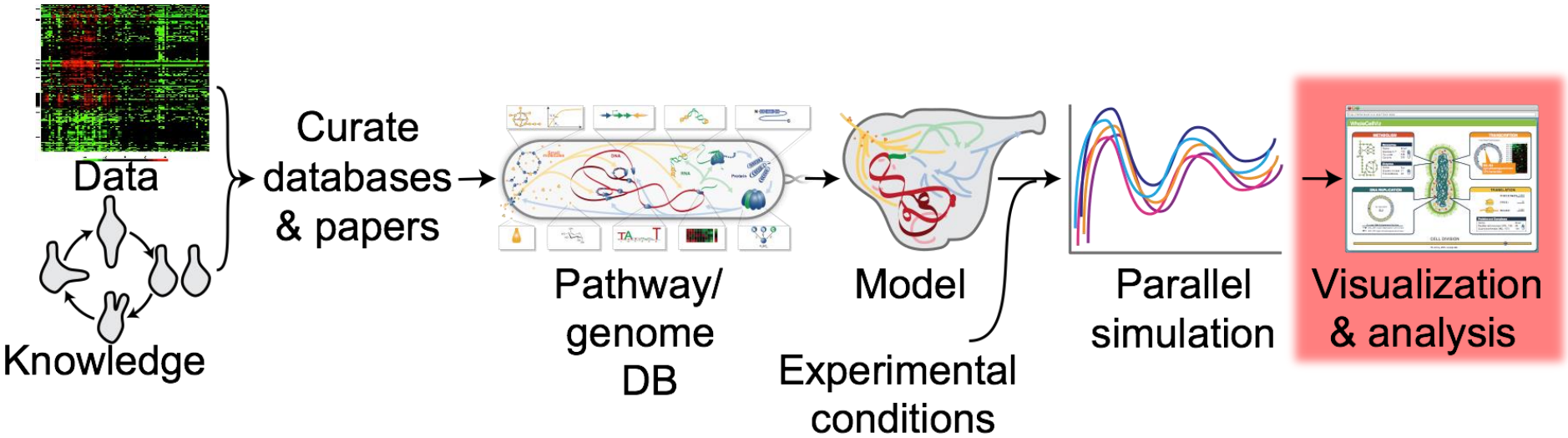
Karr et al., 2014

10. Visualization & analysis

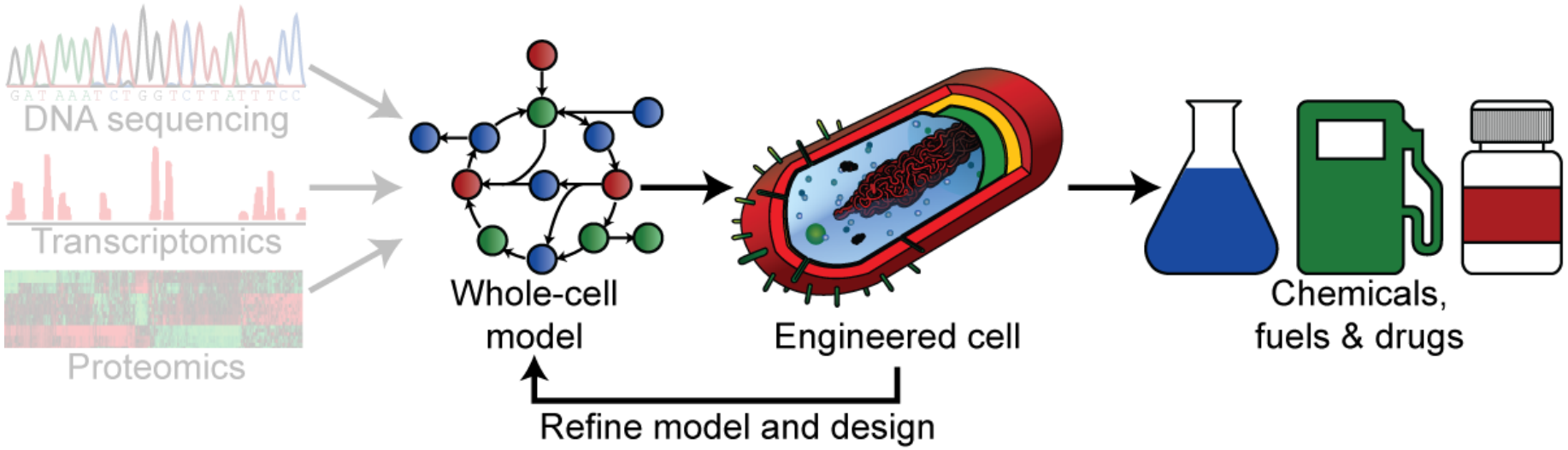


Lee et al., 2013

11. Applications: Engineering



11. Applications: Engineering



11. Applications: Engineering

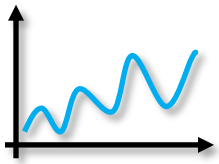
Graphical design tool

Clotho, TinkerCell, GenoCAD

(if (nutrients)
(grow)
(sporulate))

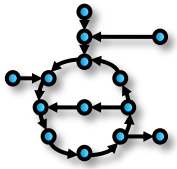
High-level language

BioCompiler



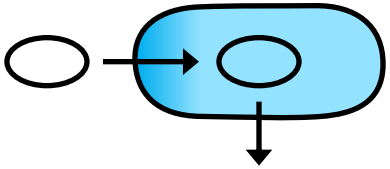
Biophysical model

WC models, SCHEMA, MD



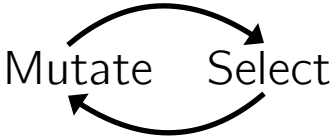
Physical implementation

Gibson assembly, TALENs, ZFNs, CRISPR



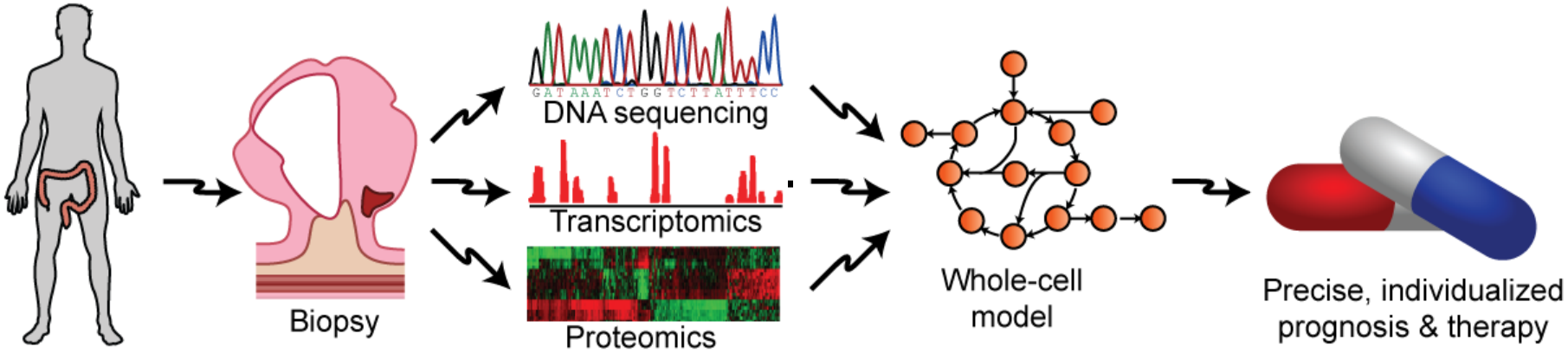
Transplantation

Transplantation



Directed evolution

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

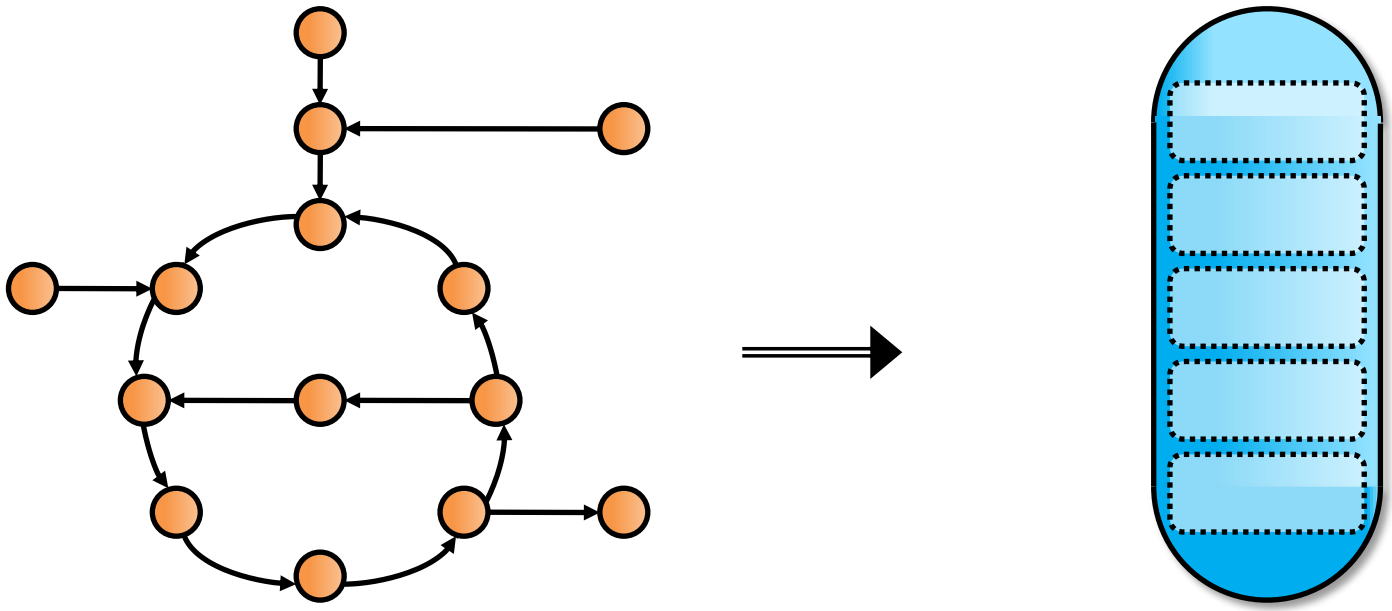
Community benchmark model project

- Contact Jonathan

SysMod ISCB special interest group

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

Summary



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