Noise in bacterial gene expression: Experiments and models

July 19, 2007

David McMillen

Students: Sangram Bagh, Colin Guangqiang Dong, Marco Iafolla, Luke Jakobowski, Mostafizur Mazumder, Ashok Movva, Vandit Sardana, and Tharsan Velauthapillai

Quantitative Biology Laboratory

Dept of Chemical and Physical Sciences

University of Toronto at Mississauga

My history

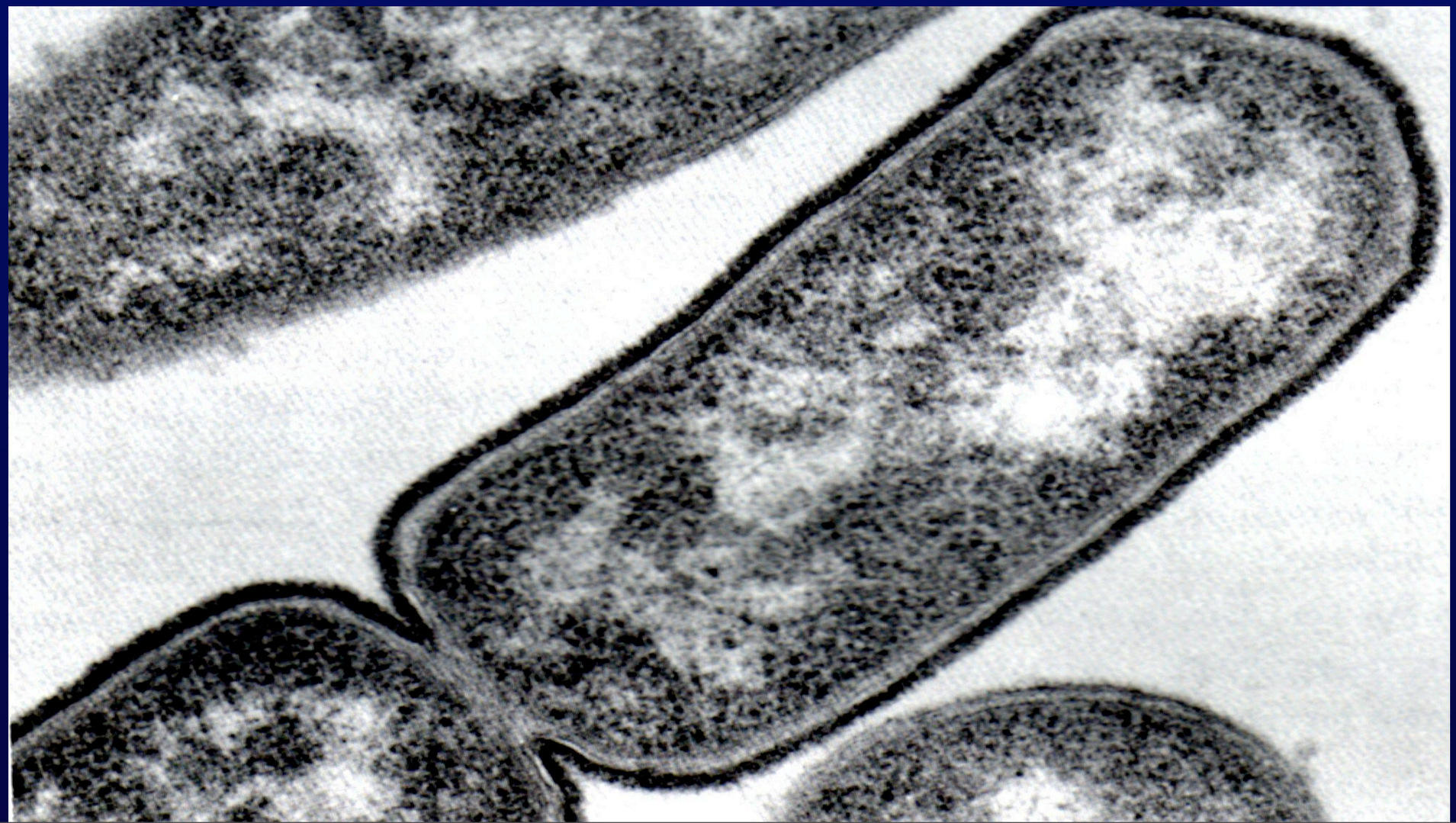
- PhD: University of Toronto engineering
  - Gabe D'Eleuterio, Space Robotics group (aerospace)
  - Models of coupled neural oscillators to control a walking robot
- Postdoc: Center for BioDynamics, Boston University
  - Neural/genetic oscillator synchronization with Nancy Kopell
  - Gene modelling with Jim Collins
- Faculty: Chemistry/Physics,Toronto (since July 2003)

People

People

The cell

- A little package of goo, in a wrapper
- Key molecules:
  - Proteins: do most of the work
  - DNA: codes for proteins



Genome sequencing projects

Genome sequencing projects

Secret of Life Solved!

Cells fully understood!

Molecular biology finished!

- Why is this not true?
- Having the pieces doesn’t mean we know how they function and fit together (dynamics, network behaviour)

Questions

- Big Question #1: How do we predict cellular behaviour?
  - Sequence data gives us the components, now how do we understand the system?
- Big Question #2: How can we *control* cellular behaviour?
  - Diseases, pathogenic invasions: involve alterations of natural dynamics
  - Can we reestablish normal function?

Dynamics of the cell

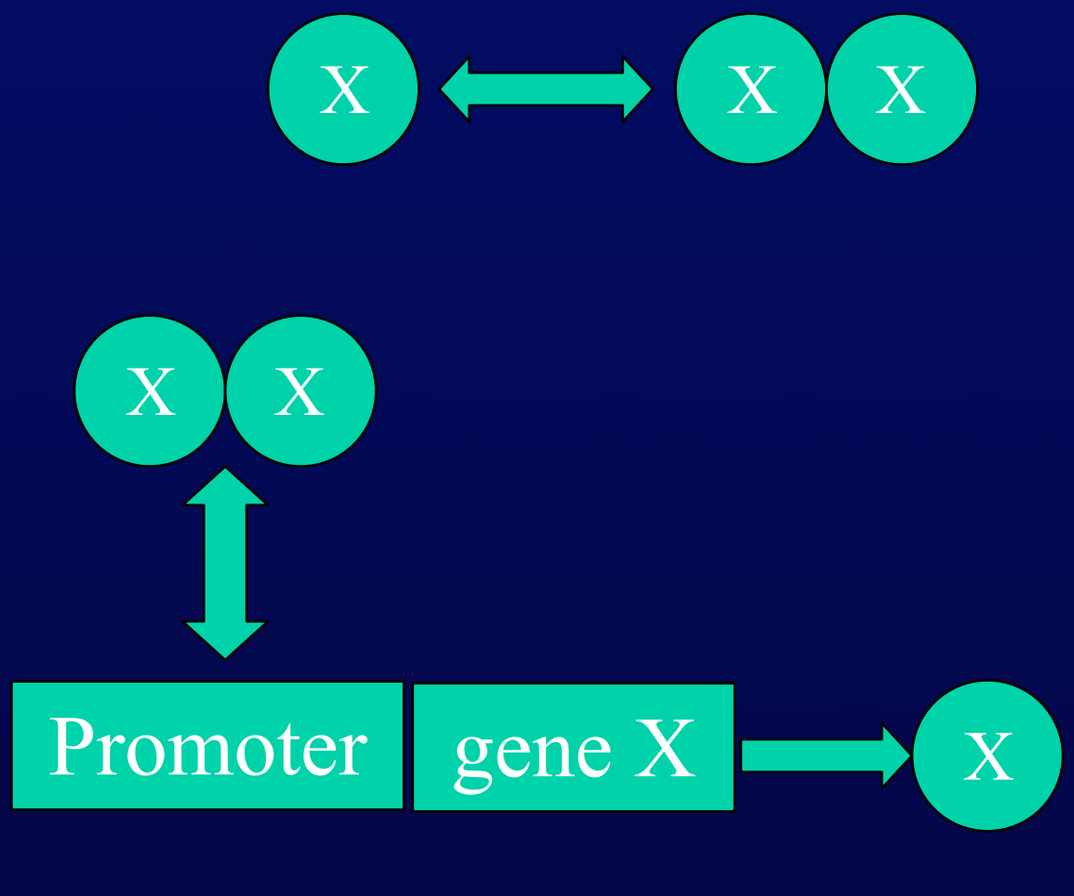
- Dynamics: How a system's state evolves over time
- State of a cell:
  - Numbers of biomolecules/complexes
  - Proteins, mRNA, DNA-protein complexes ...
- Time evolution:
  - Driven by biochemical reactions
  - Transcription, translation, binding ...

If cells were beakers ...

1. Use standard kinetics to form models
2. Identify the set of relevant biochemical reactions
   - List all species of interest
   - Include all reactions that affect those species
3. Determine rate constants
   - Production, degradation, binding
4. Derive dynamics from the chemical kinetics
5. Main problem: sheer scale

Formulating models

- Represent production, binding, degradation as a set of reactions
- Use chemical kinetics to turn reactions into a mathematical model



$$D + X_2 \xleftrightarrow{k_1 / k_{-1}} D_X$$

$$X + X \xleftrightarrow{k_2 / k_{-2}} X_2$$

$$D \xrightarrow{k_T} D + X$$

$$D_X \xrightarrow{\alpha\, k_T} D_X + X$$

$$X \xrightarrow{k_x}$$

Formulating models

- Translate reactions into rate equations for each species

$$D + X_2 \xleftrightarrow{k_1 / k_{-1}} D_X$$

$$X + X \xleftrightarrow{k_2 / k_{-2}} X_2$$

$$D \xrightarrow{k_T} D + X$$

$$D_X \xrightarrow{\alpha\, k_T} D_X + X$$

$$X \xrightarrow{k_x}$$

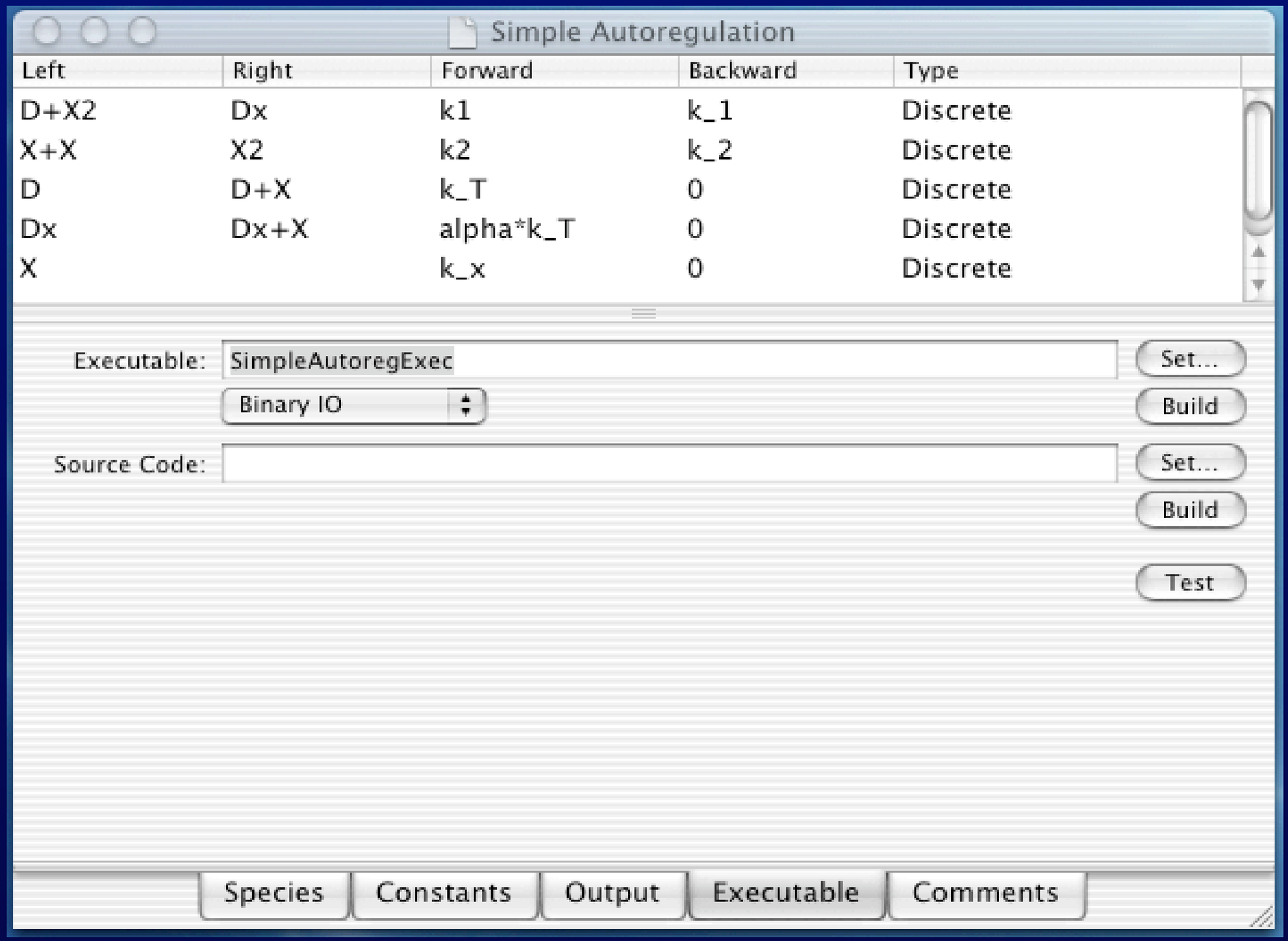
$$\dot{x} = -2k_2x^2 + 2k_{-2}x_2$$

$$+ k_T d$$

$$+ \alpha\, k_T d_x$$

$$- k_x x$$

Autoregulation in BioNetS



Adalsteinsson, McMillen, and Elston. BMC Bioinformatics 5:24 (2004).

But cells differ from beakers ...

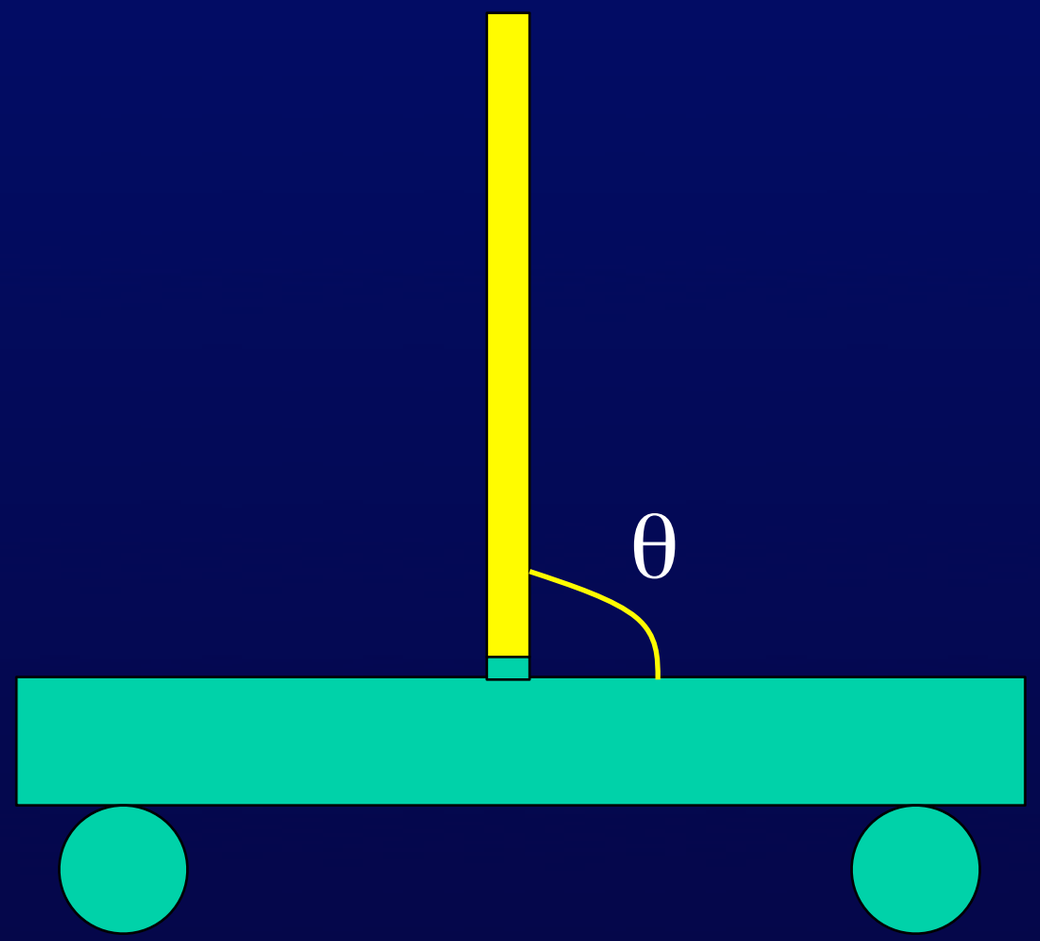
1. Hard to reach into and control
   - Design of synthetic network controllers
2. Small size / small numbers
3. Growth and division
   - Cells double, cut in half
4. Individual histories and identities
   - Cell-to-cell variation
5. High complexity
   - Model reduction methods are valuable
6. Crowded molecular environment
   - How do crowded kinetics change?

1. Cells are stubborn

- By “stubborn,” we mean: hard to control (fiendishly uncooperative!)
- Want to be able to exert control over cells from the inside
  - “Killer app”: the in-cell cancer detector that kicks cell into apoptosis
- Hope is to learn to design synthetic regulatory networks capable of this
  - We’re working on a system that cures disease - but only in bacteria

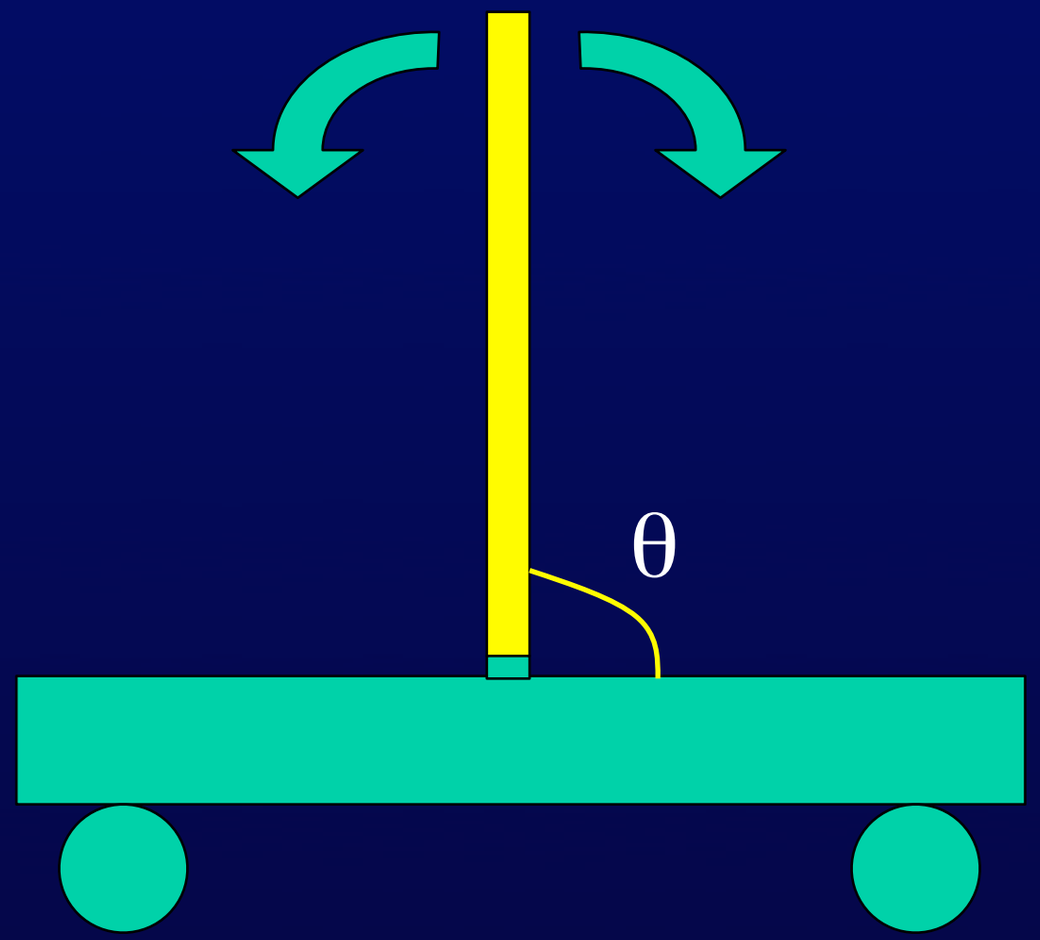
Control systems

- The "game" in control systems is to alter a system's dynamics so that it has desired properties



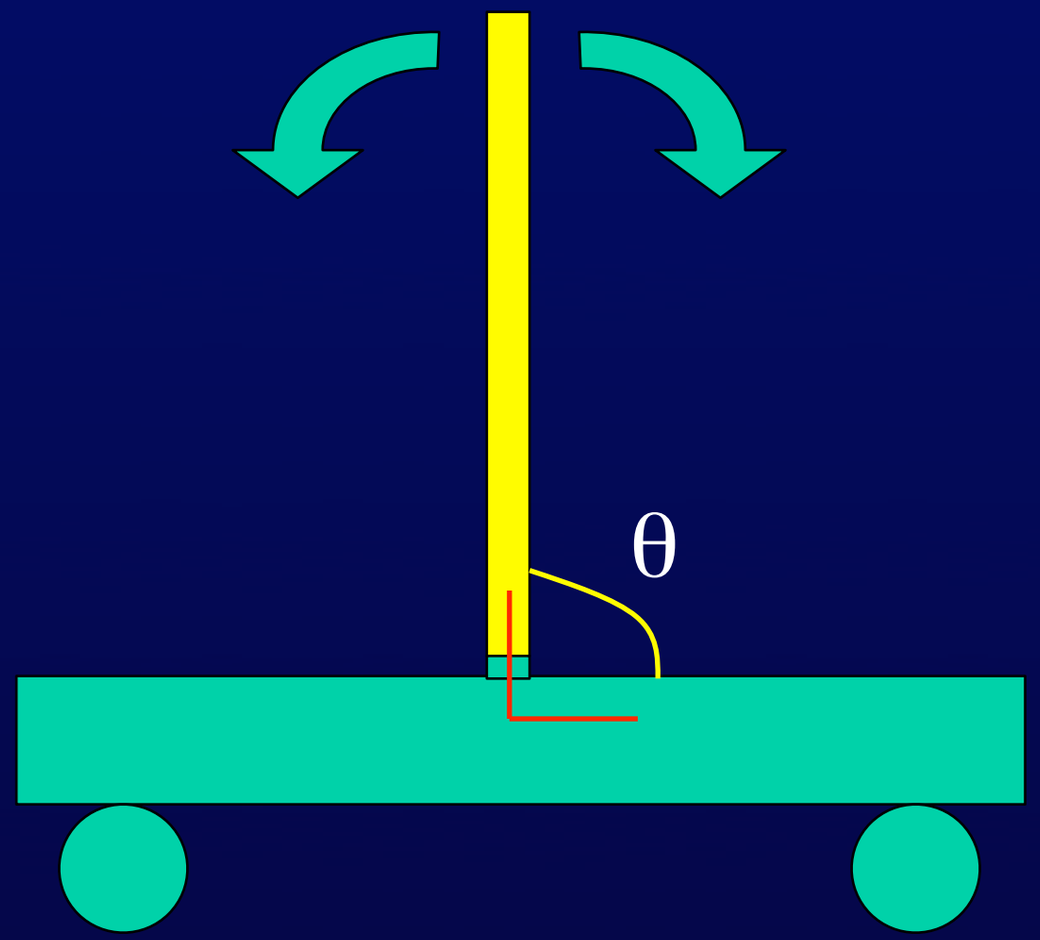
Control systems

- The "game" in control systems is to alter a system's dynamics so that it has desired properties



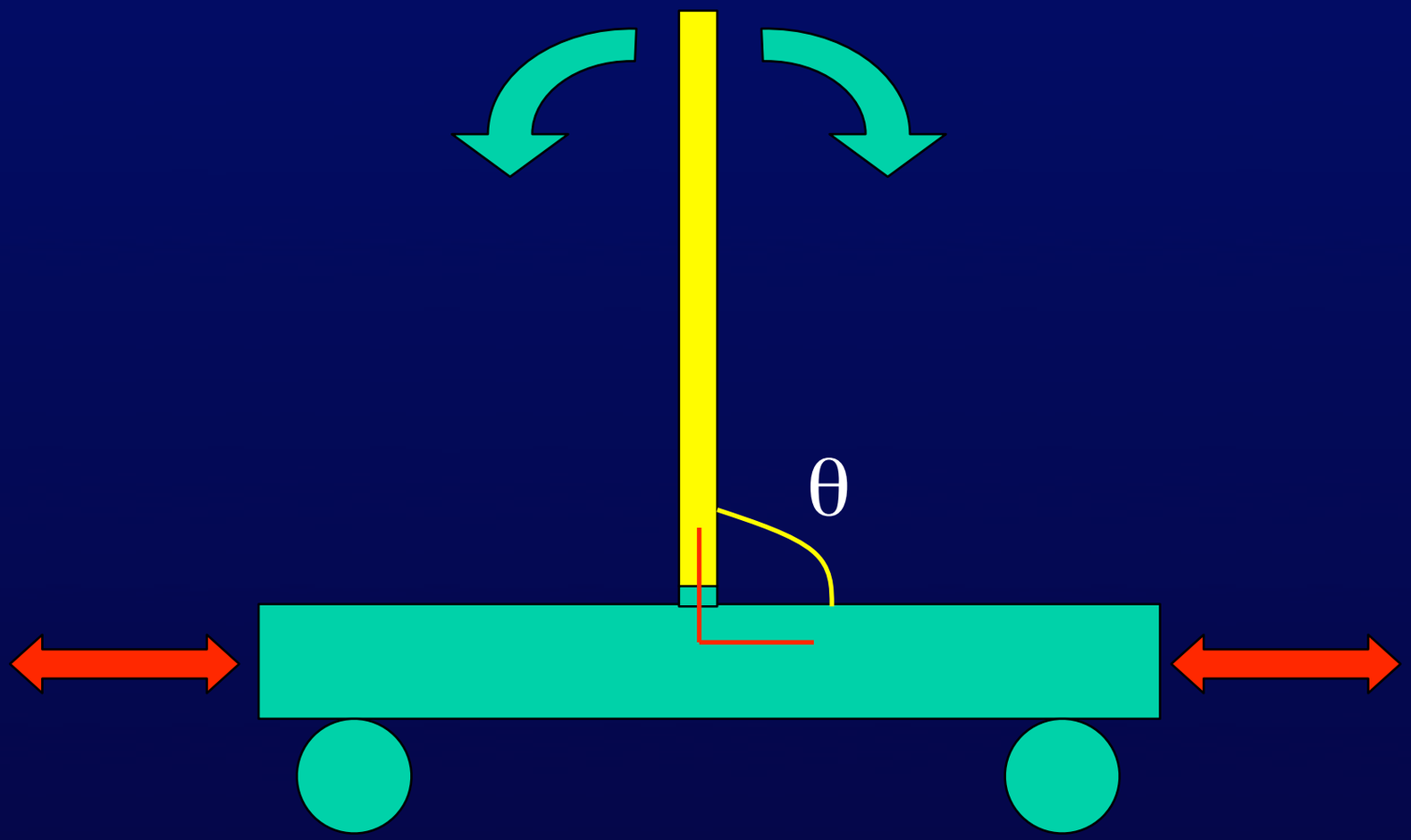
Control systems

- The "game" in control systems is to alter a system's dynamics so that it has desired properties

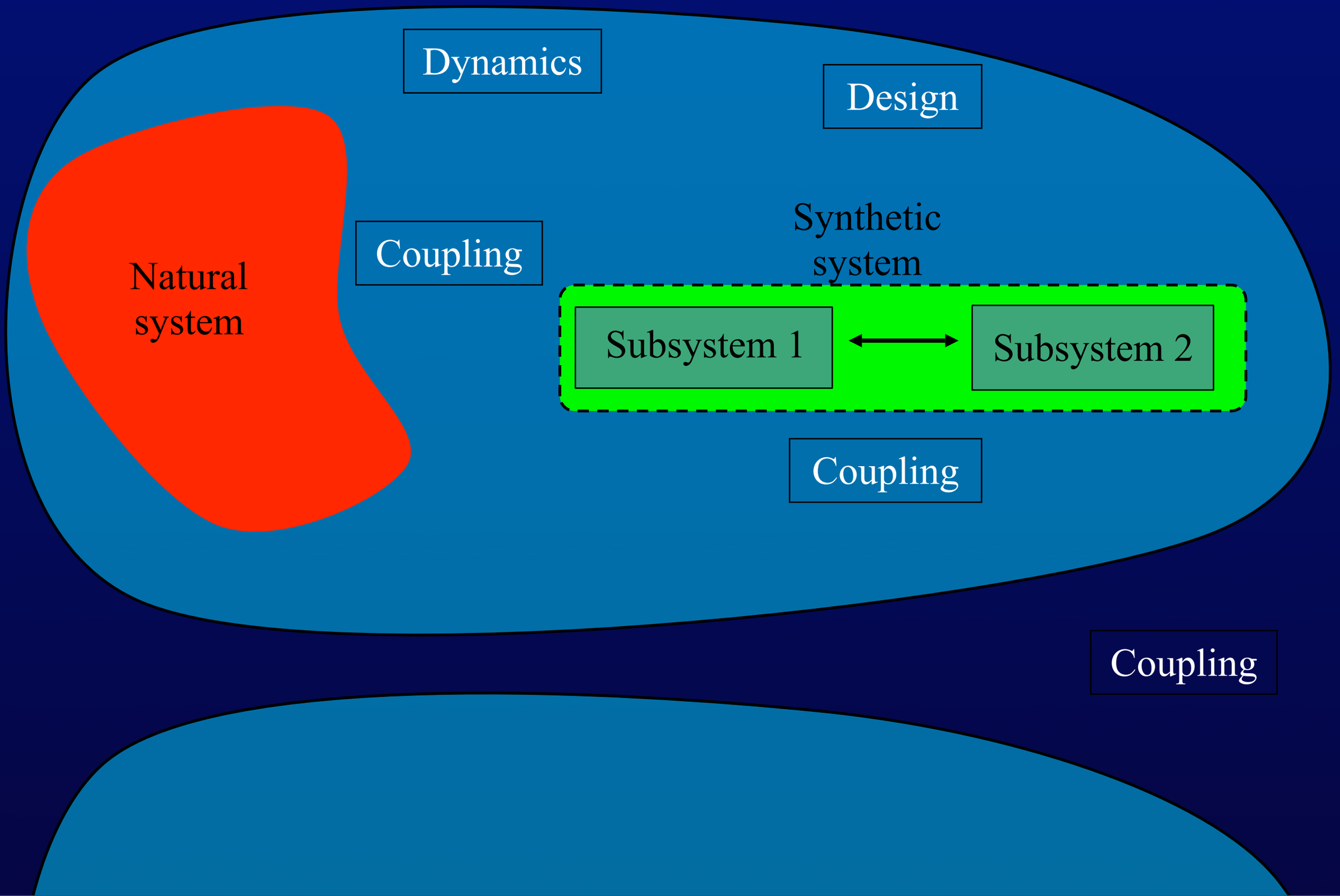


Control systems

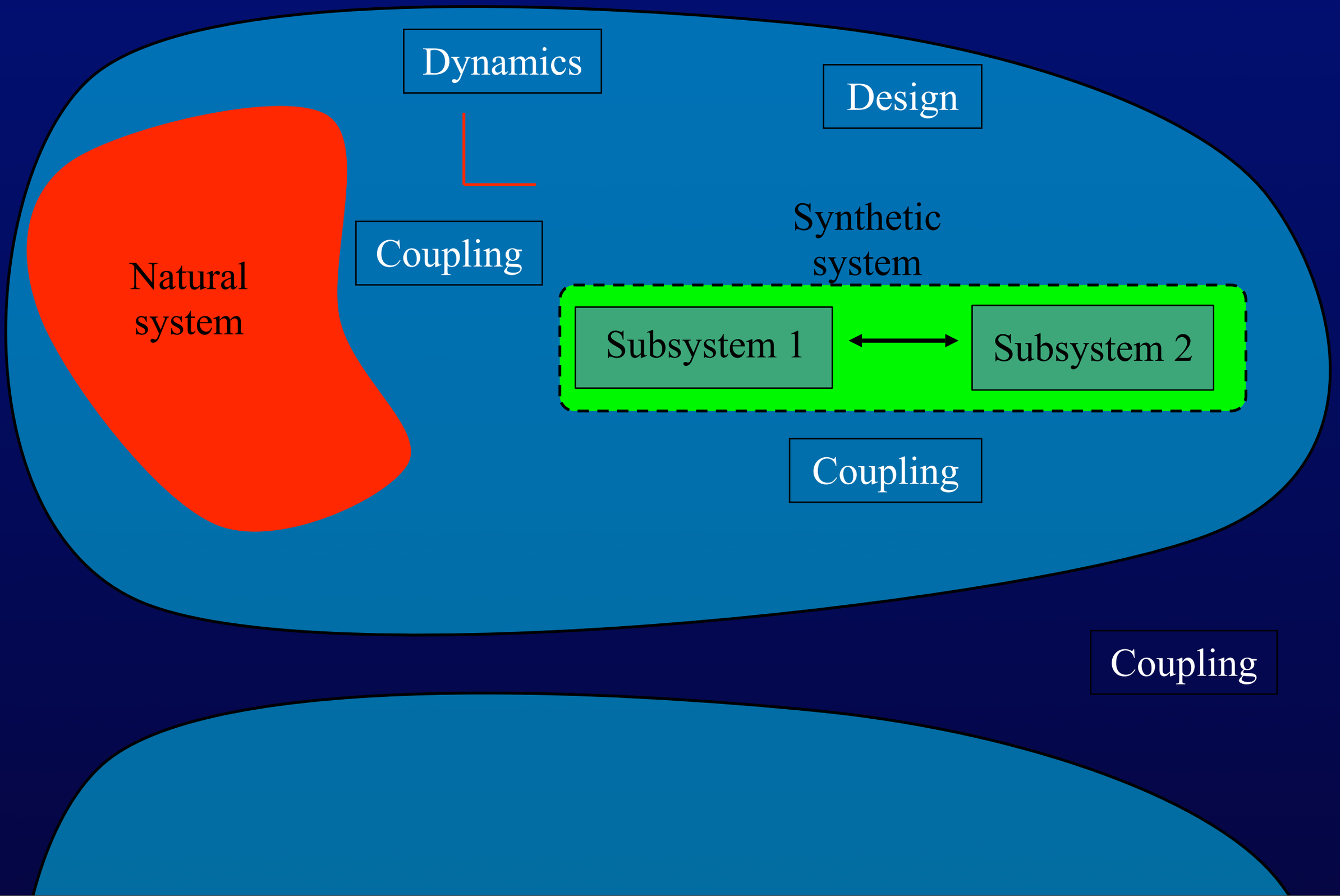
- The "game" in control systems is to alter a system's dynamics so that it has desired properties



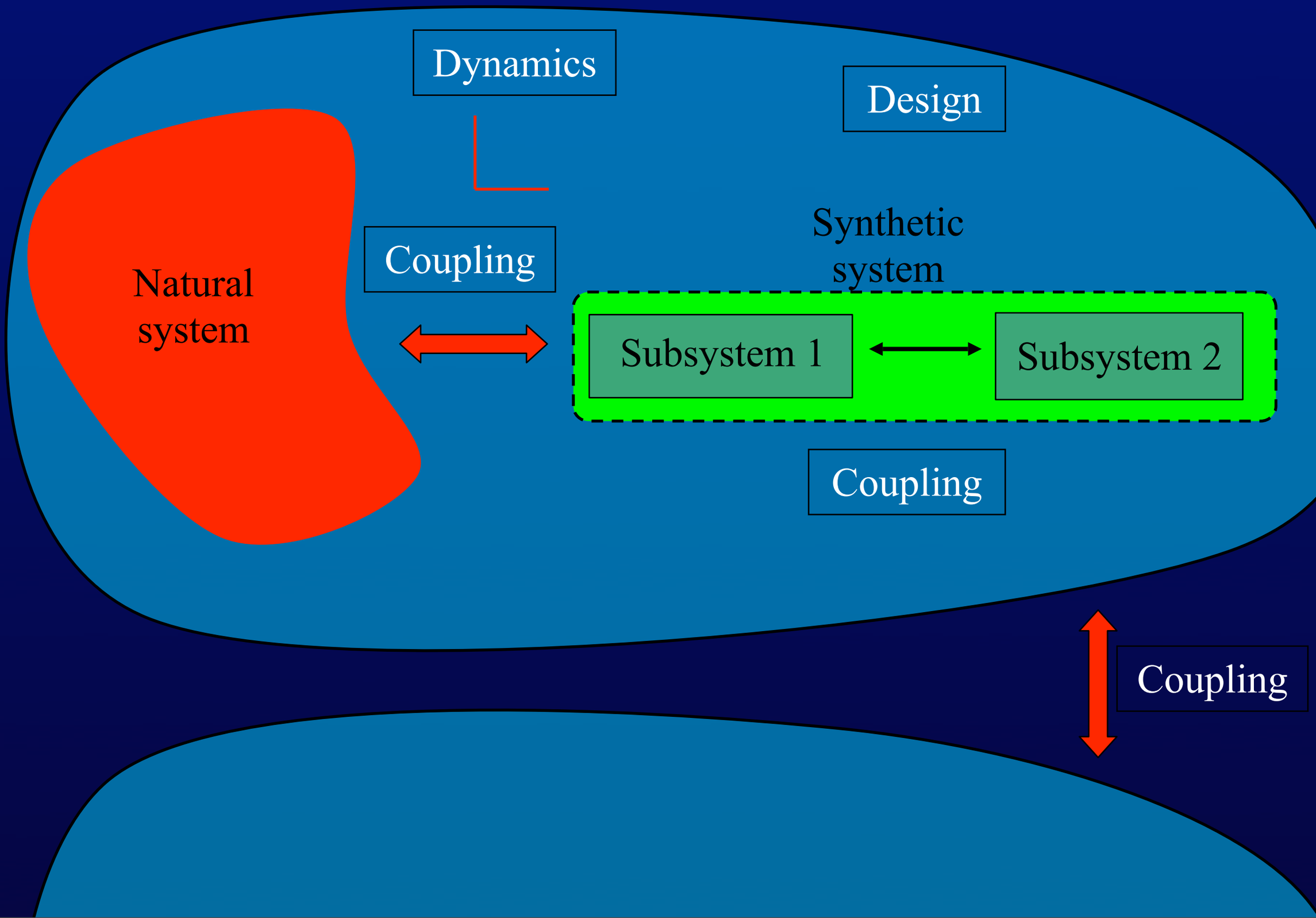
Cellular control - requirements



Cellular control - requirements

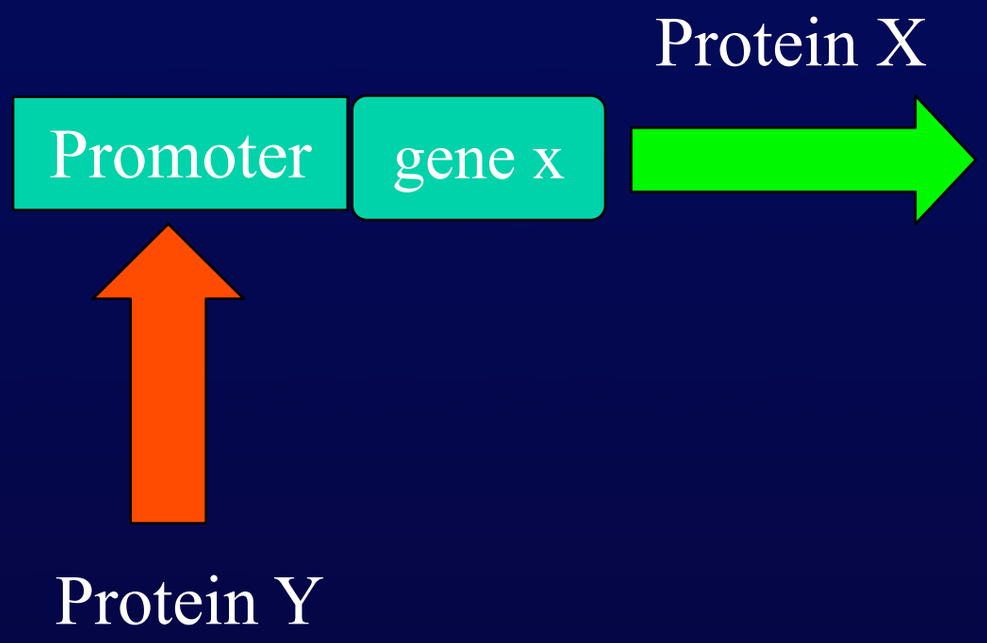


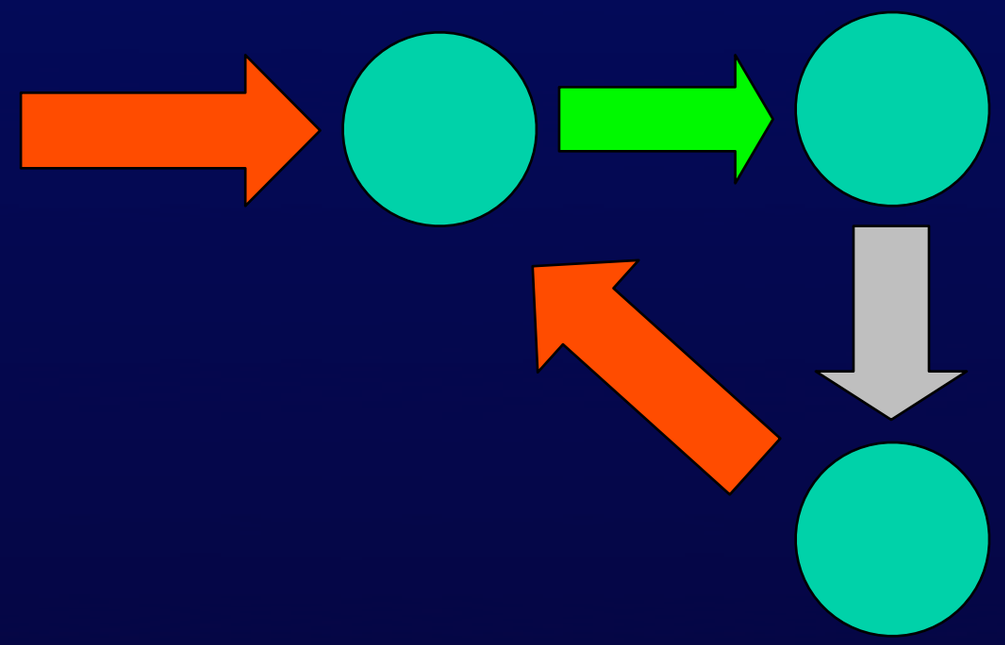
Cellular control - requirements



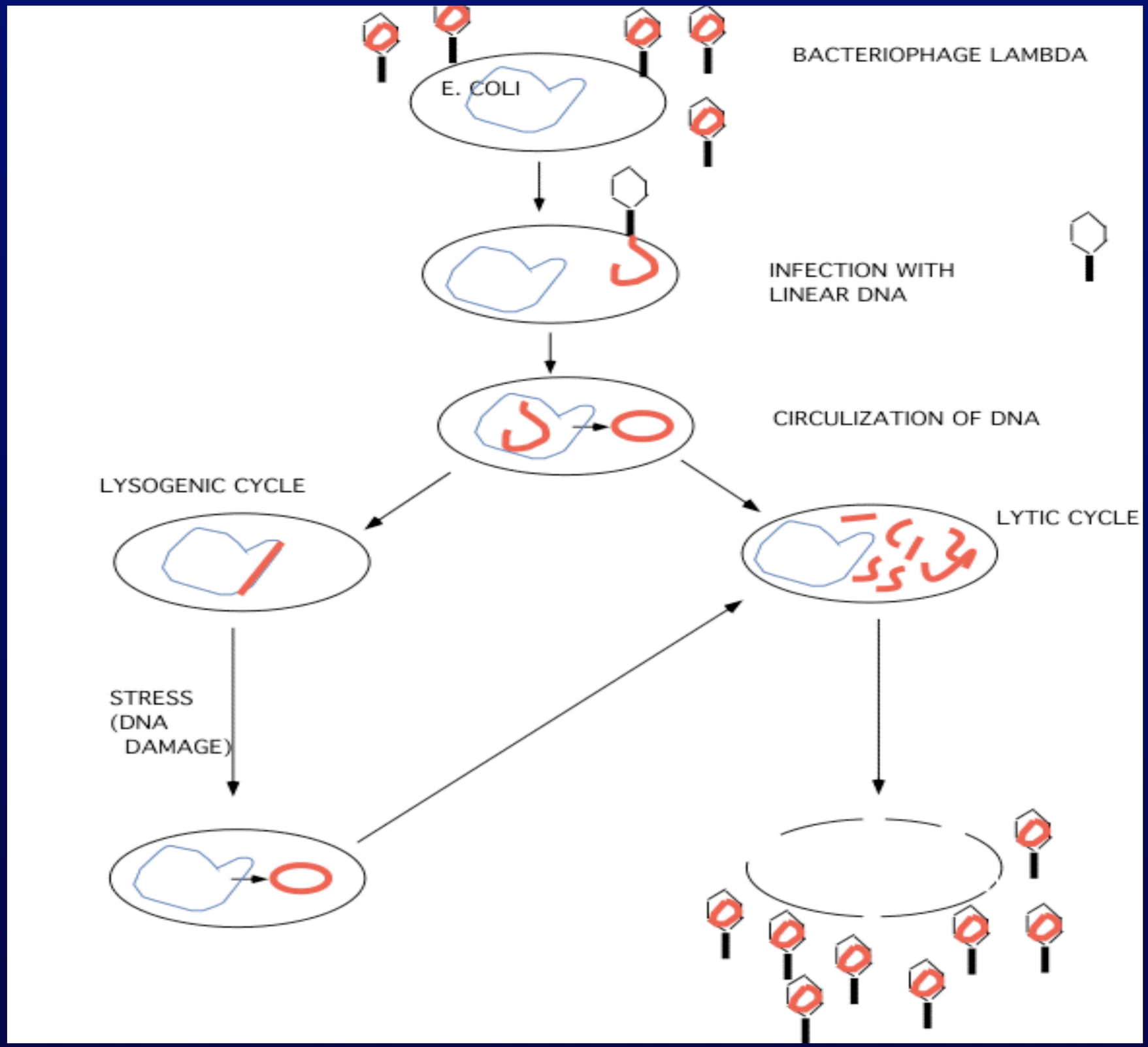
Designing synthetic networks

- Node in the network = a promoter/gene pair
- Choice of promoter = choice of input(s)
  - Sets which proteins affect the node
- Choice of gene = choice of output(s)
  - Sets which genes are affected by the node





Bacteriophage lambda

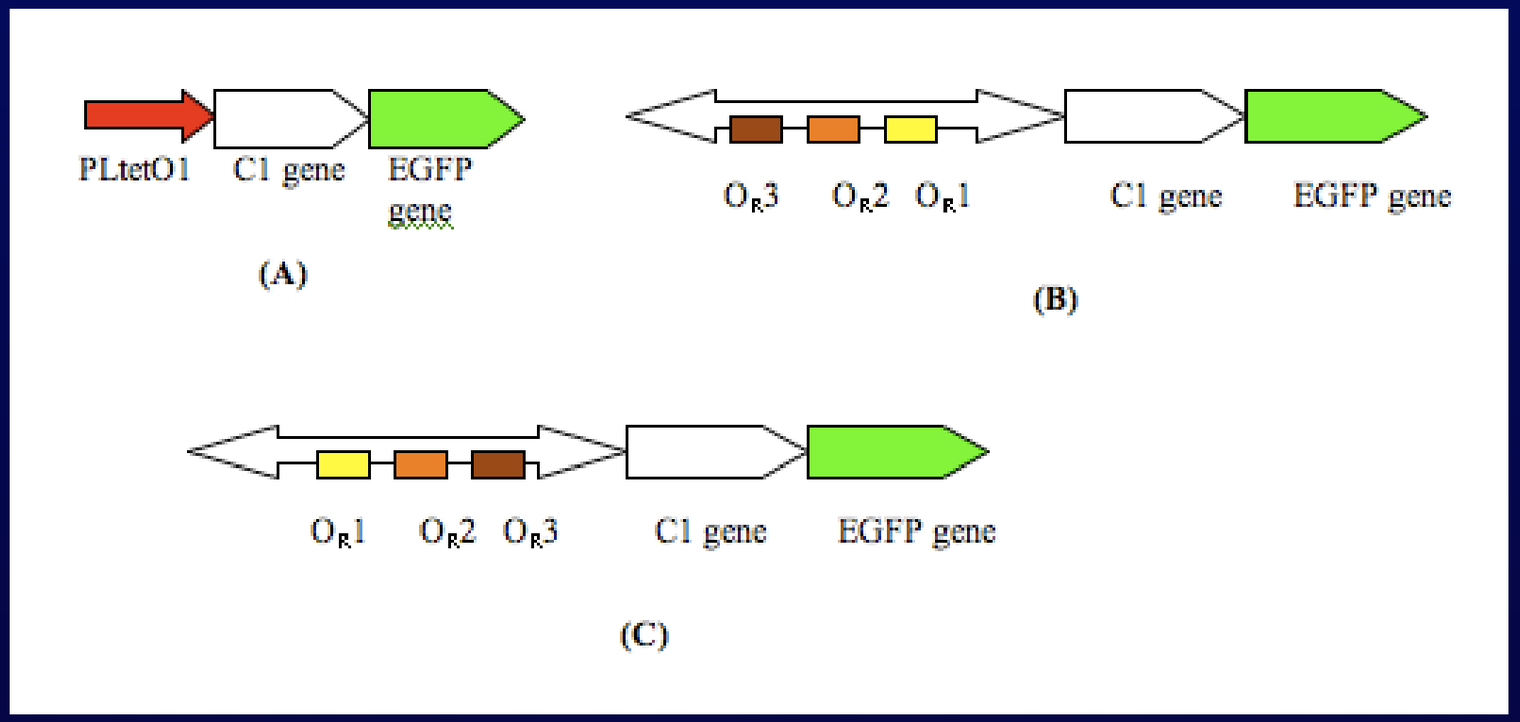


Lysis prevention

- Use a cell-resident network to prevent a fatal disease
  - Currently in bacteria, but looking to the future, and human medical applications
- Prevent lysis in E. coli infected by λ
  - Protein CI: maintains lysogeny, prevents lysis; SOS response causes RecA to cleave CI monomers, CI drops
  - Protein Cro: expression leads to lysogeny

Lysis prevention

- Sense onset of lytic pathway using a CI-repressed promoter
- When CI level drops, (lysis coming!), produce extra CI to maintain lysogeny
- Status: Working towards it



Sangram Bagh

2. Cells are small

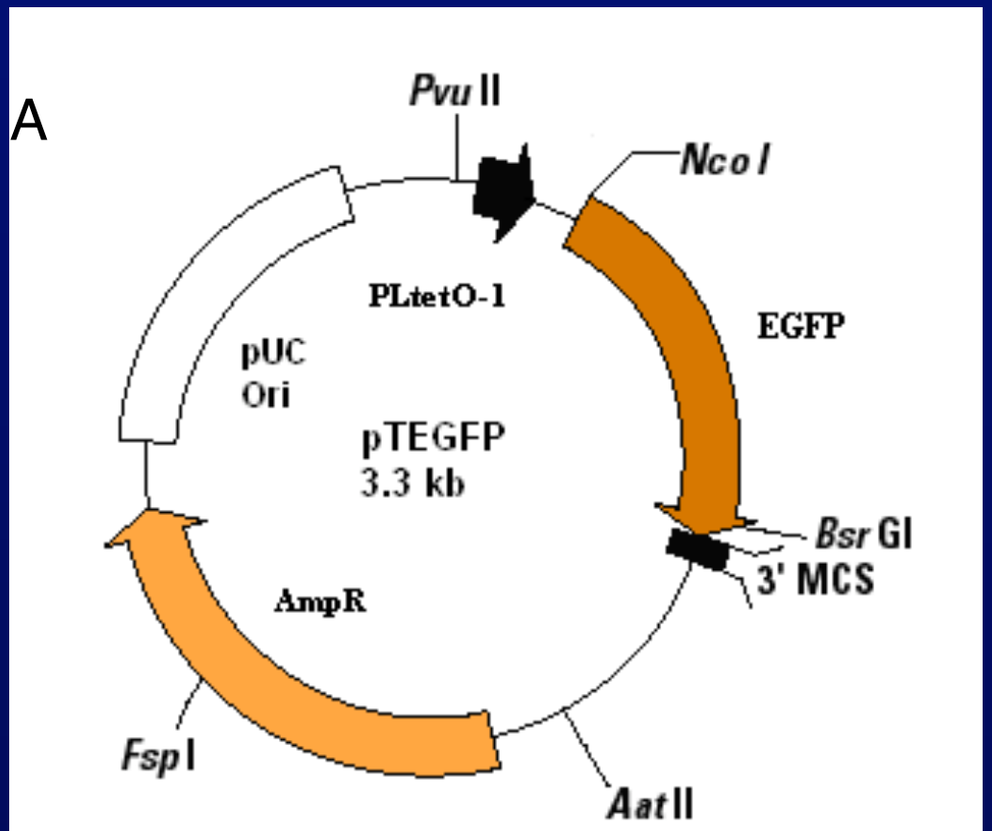
- Chemical kinetics:
  - Random interactions of molecules
  - For moles of particles, rates effectively deterministic
- But for small numbers of molecules, fluctuations become significant
- Inside a cell:
  - Small numbers of molecules, not moles
- Question: What are the actual numbers?
  - Recent work: Swain and Elowitz (binomial division); Ghaemmaghami (blotting); Barkai (fluorescence)

Unregulated system

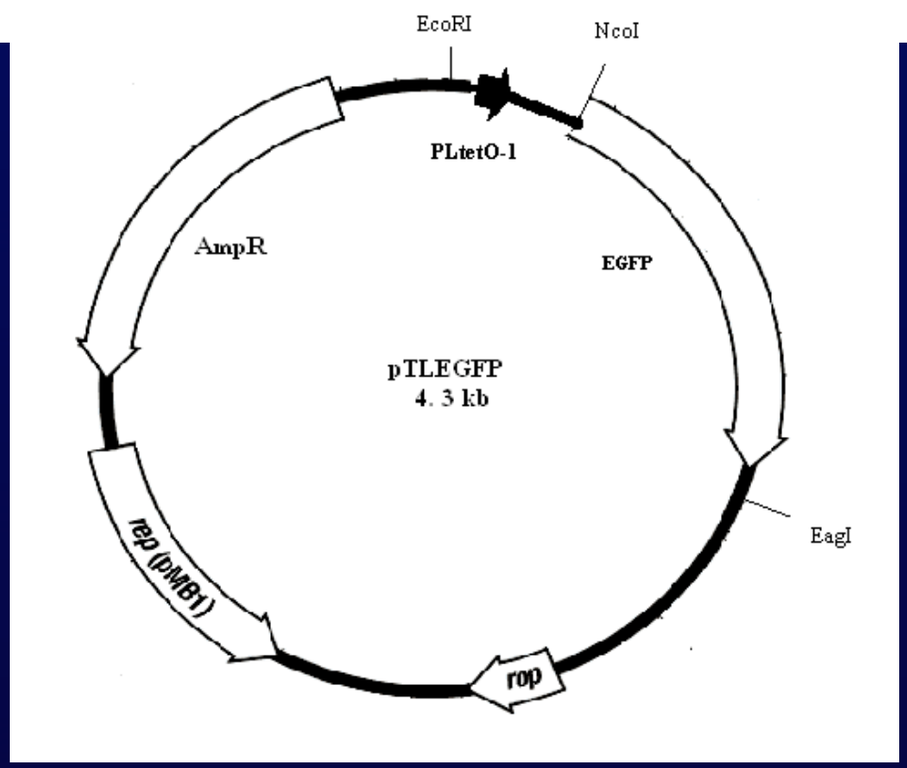
- Constructed a simplified system:
  - Promoter with no (known) regulatory feedback present, $P_L$tetO1
  - Expressing EGFP directly
- Inserted into *E. coli* cells on plasmids with two different replication patterns:
  - High copy number plasmid: replicates as rapidly as it can, circa 400 copies/cell
  - Medium copy number plasmid: feedback reduces replication, circa 40 copies/cell
- Examined four strains of bacteria

Plasmids

A


Pvu II
Nco I
PLtetO-1
EGFP
pUC
Ori
pTEGFP
3.3 kb
Bsr GI
3' MCS
AmpR
Fsp I
Aat II

B


EcoRI
NcoI
PLtetO-1
AmpR
EGFP
pTLEGFP
4. 3 kb
EagI
rop

Strains

Strain	Characteristics
DH5a	F^- endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG 80dlacZΔM15 Δ(lacZYA-argF)U169, hsdR17 (r_K^- m_K^+), λ–
Top10	F^- mcrA (mrr-hsdRMS-mcrBC) 80dlacZM15 lacX74 recA1 ara139 (ara-leu)7697 gal/U gal/k rpsL (str^R) endA1 nupG
B/r	F26 his thy
BL21*	(r_k^-, m_k^+) phoA supE44 $^-$ thi-1 gyrA96 relA1

Measurement/calibration

- Measure the "output" (protein expression) by quantifying fluorescence, through flow cytometry and microscopy
- Calculate mean number of EGFP per cell using bulk fluorimetry:
  - Calibrate against known numbers of EGFP in PBS solution (matches cellular pH): yields equivalent # of EGFP in cell culture
  - Use optical absorbance to get cells/ml
  - Divide one by the other to get <EGFP>/cell

Protein numbers

Cell strain	Plasmid	<EGFP>/cell	Div time (min)	Proteins /min
DH5a	High	156,000	36.8	2900
	Medium	23,400	30.5	530
Top10	High	119,000	29.5	2800
	Medium	17,400	29.8	400
B/r	High	144,000	27.9	3600
	Medium	11,000	33.1	230
BL21*	High	46,400	49.9	640
	Medium	5,800	31.6	130

- Order of magnitude: tens to hundreds of thousands of proteins/cell
- Substantial differences across strains

Sangram Bagh, Mostafizur Mazumder

3. Cells grow and divide

- Unlike beakers, our bacteria double in size every 20-40 min, then cut themselves in half
- Implications for kinetics
  - May need to work in number space rather than concentration space
  - Rates are volume-dependent
  - Differential equations --> maps (maybe?)
- Range of cell sizes complicates matters if you can’t reliably scale away cell size (as in flow cytometry)

“Sawtooth” gene expression

- For high rates of expression, the process of gene expression becomes near-deterministic
  - (In simple-ish models)
- Growth and division induces a sawtooth pattern of protein vs time
- If you see only total fluorescence intensity rather than intensity/size, there’s a contribution to “variability” from the size range

Very simple model

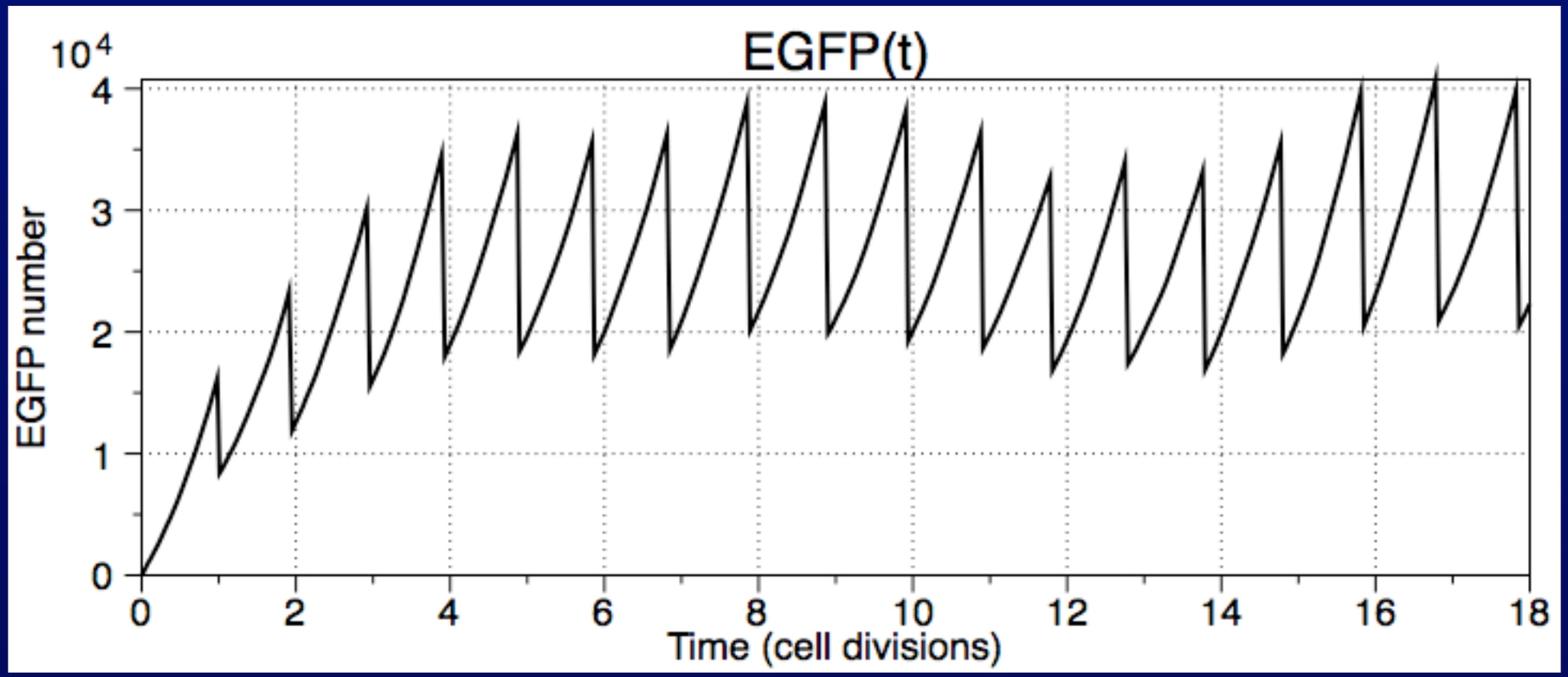
Sawtooth.bnet

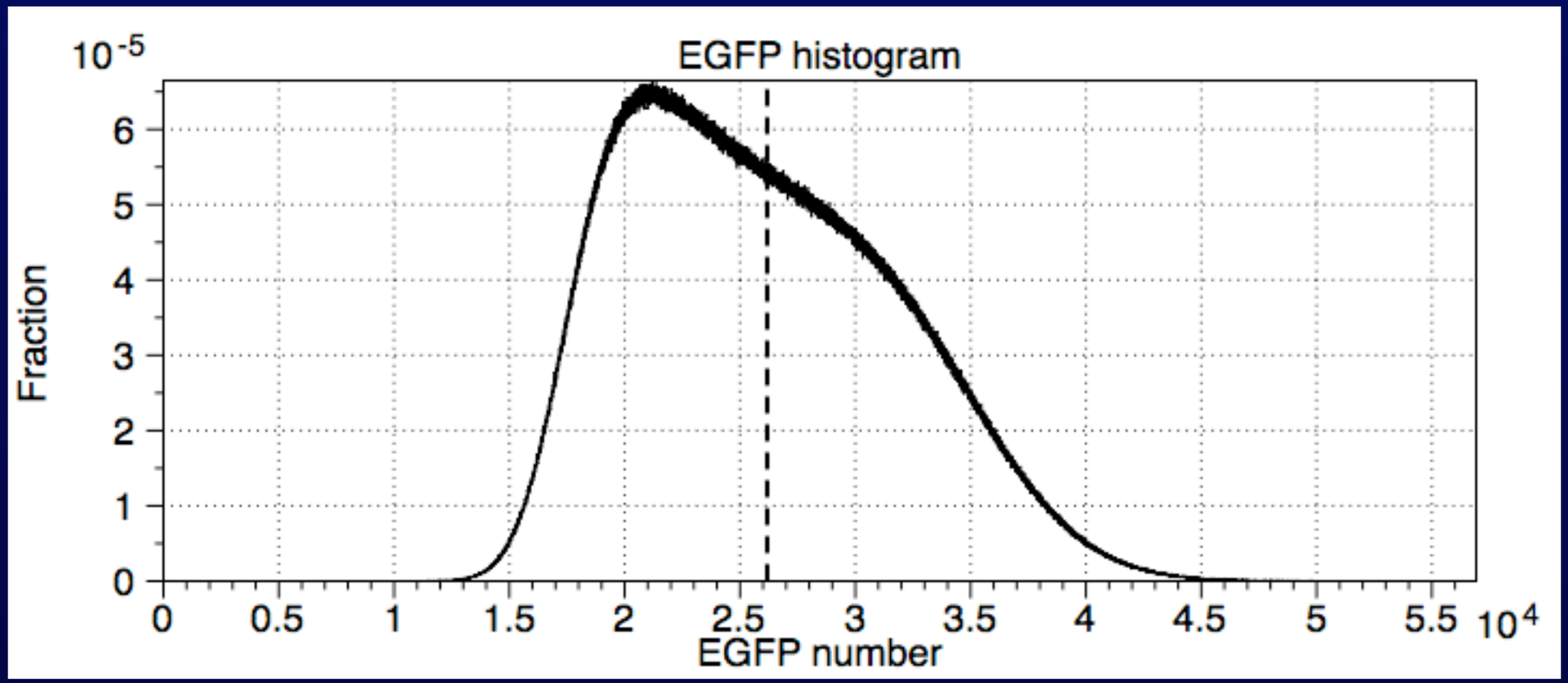
Left	Right	Forward	Backward	Type
	D	k1		Discrete
D	D+G	k2		Discrete
G		k3		Discrete
V	2V	k_v		Discrete

Name	In #	Type	Save	Out #	Hist	Range	Divide
D	1	☑ Discrete	☑ Save	2	☑ Hist.		☑ Divide
G	2	☑ Discrete	☑ Save	3	☑ Hist.		☑ Divide
V	3	☑ Discrete	☑ Save	4	☑ Hist.	200	☑ Divide

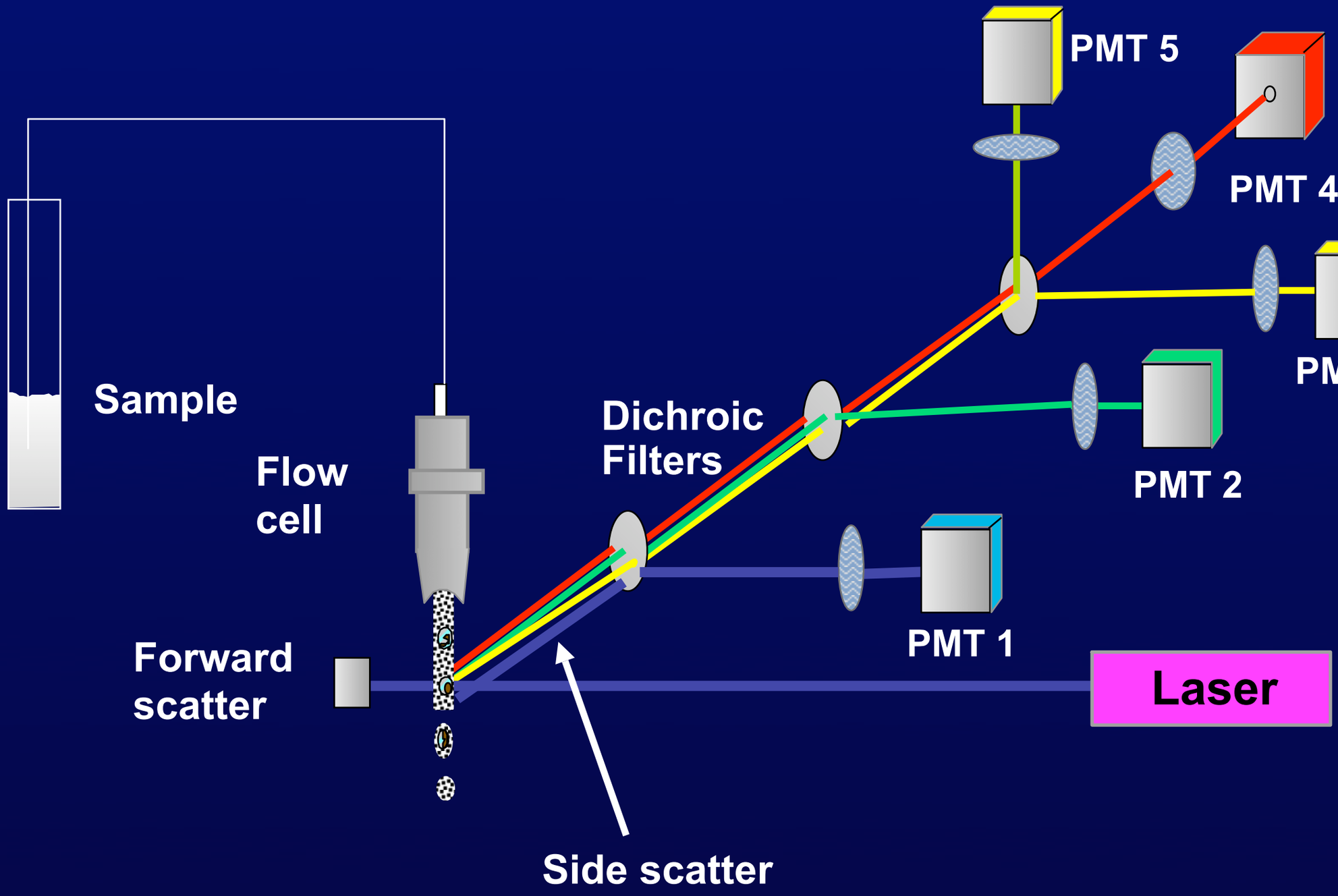
Species | Constants | Output | Expert | Executable | Comments

Simple model outputs



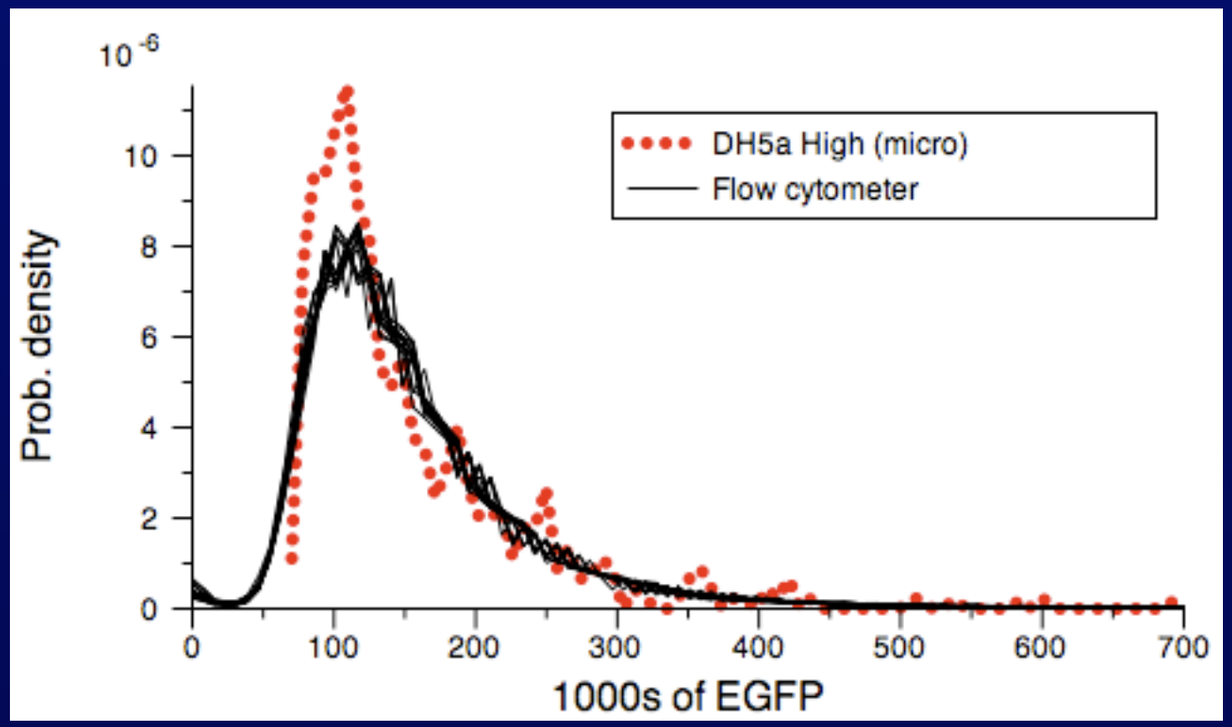


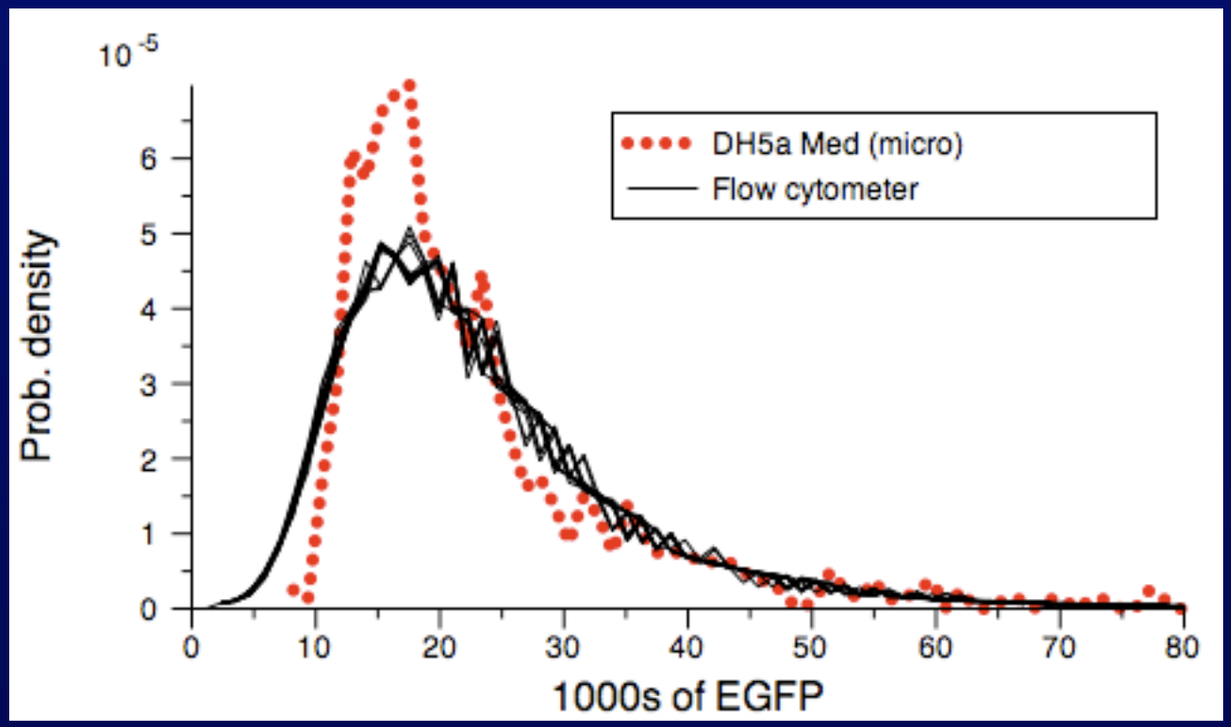
Experiments: Flow cytometry



Protein distributions

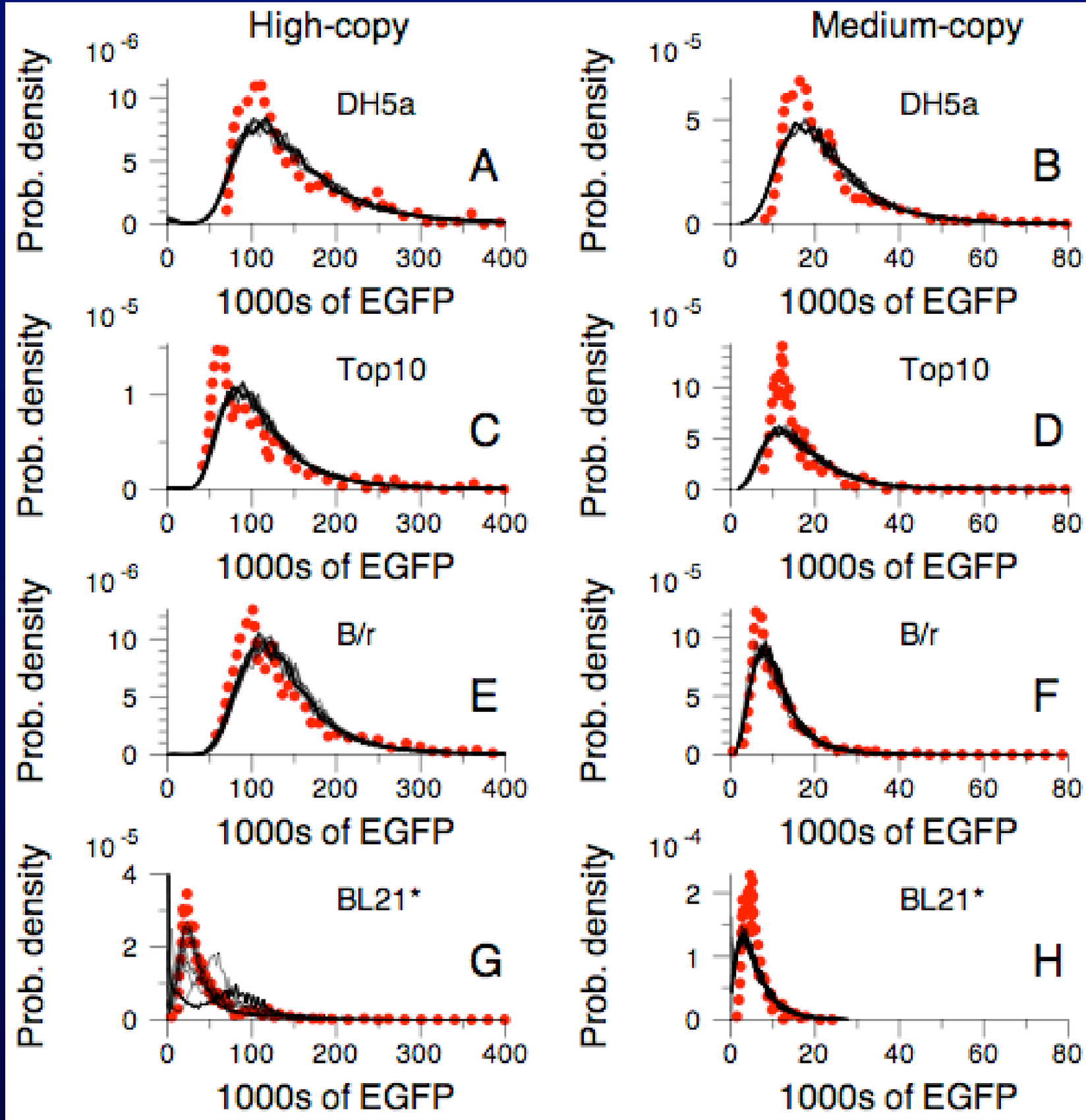
- By flow cytometry and microscopy:



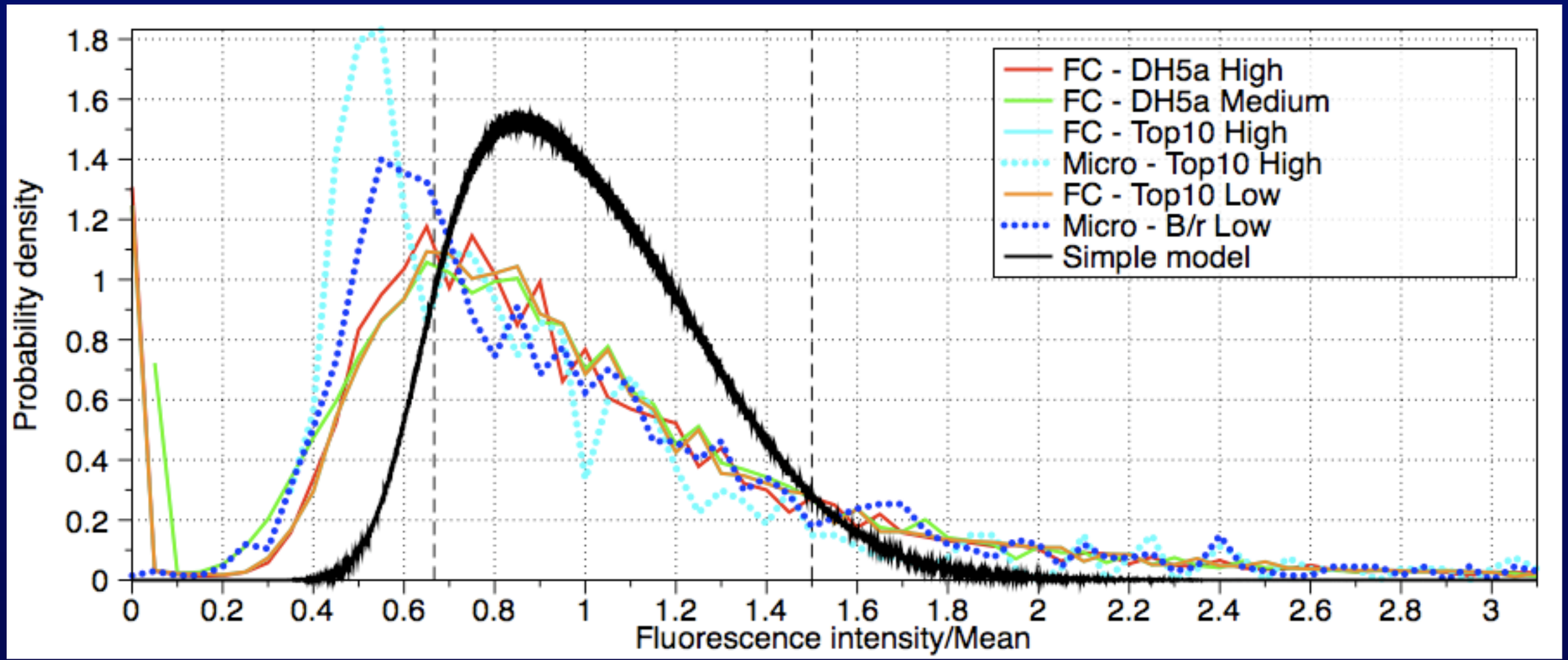


Sangram Bagh, Mostafizur Mazumder

Distributions: all strains

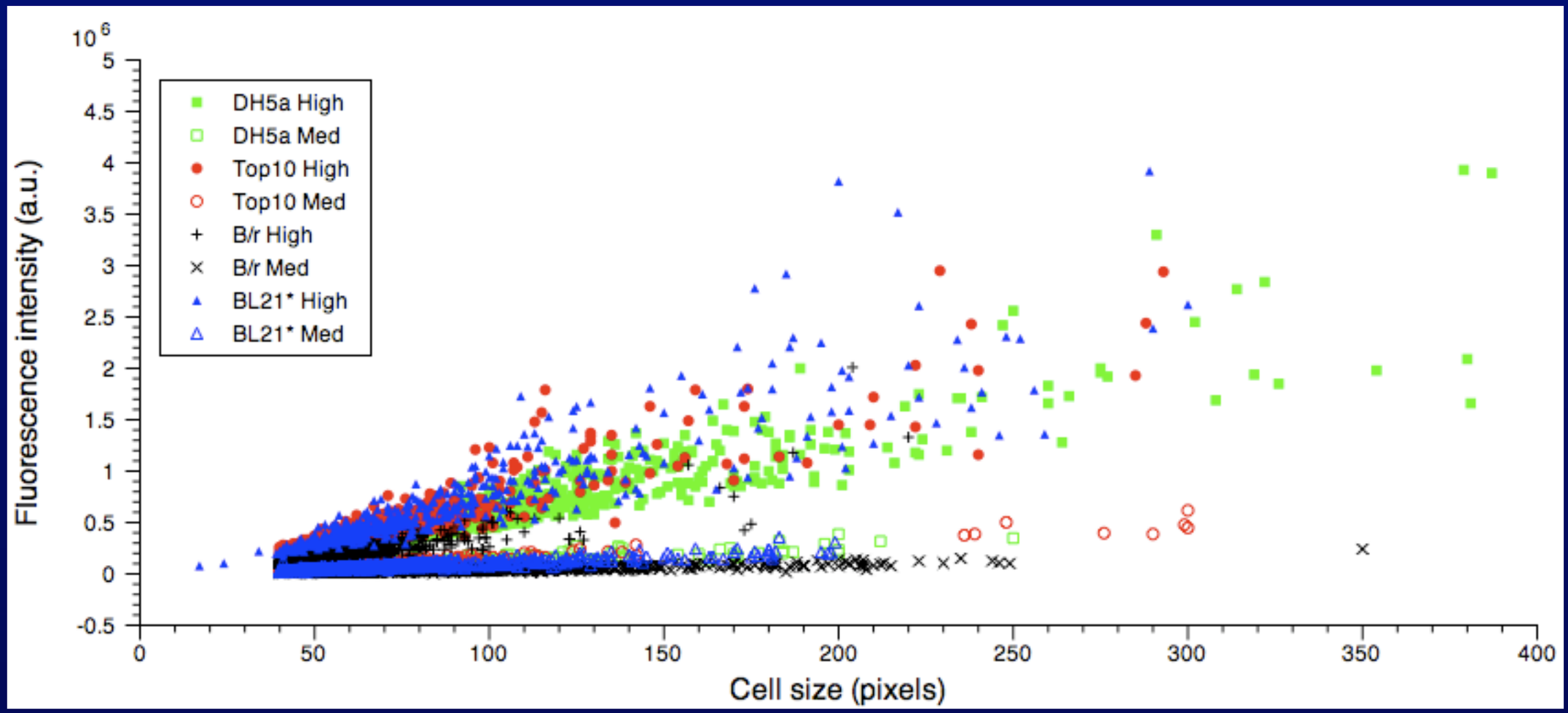


Comparing to model



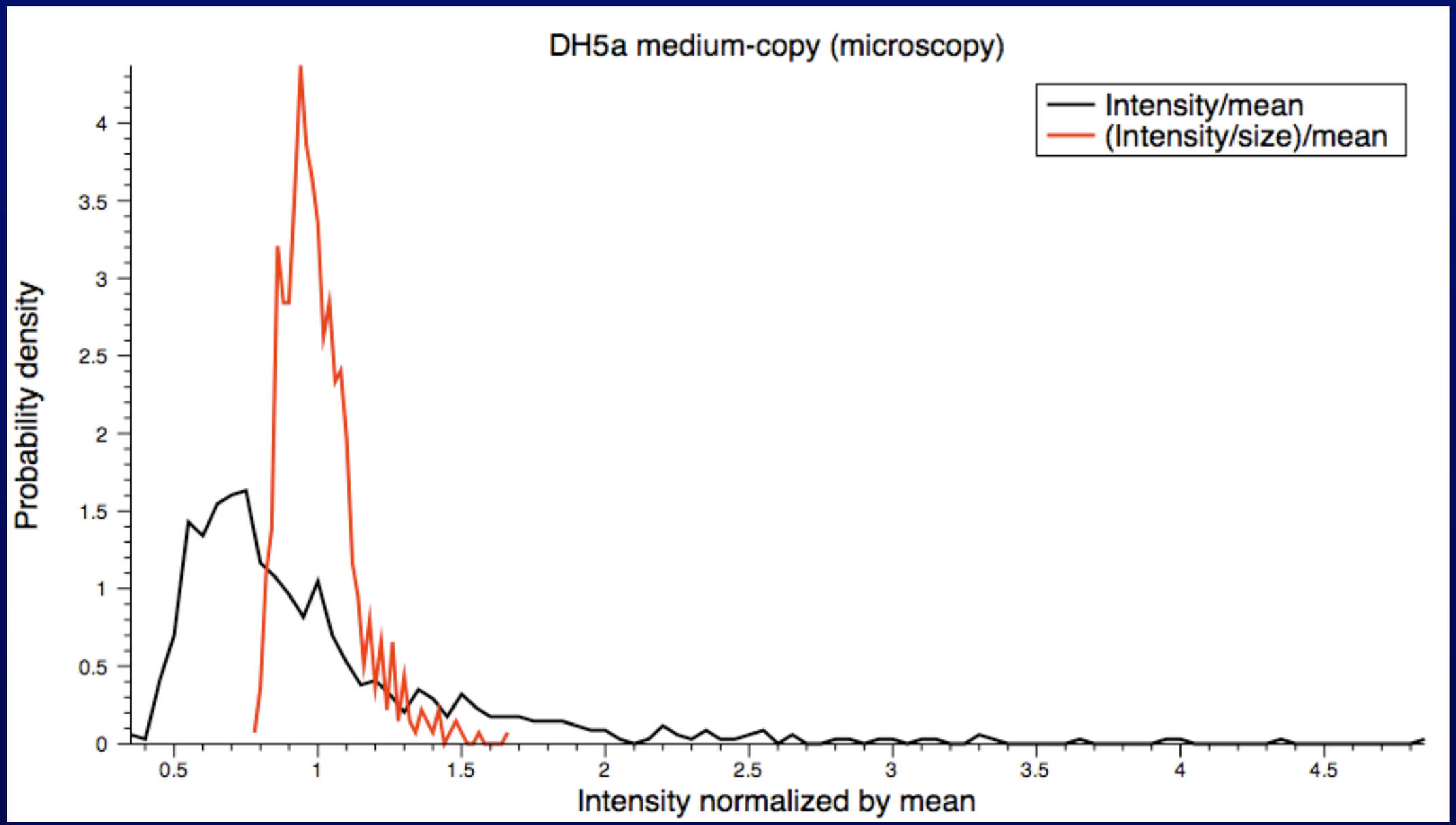
- Simple cell growth picture (V->2V) doesn’t suffice to reproduce actual cell size distributions

Fluorescence vs size



- Strong correlation
- Stronger for medium-copy than high-
- Effect of cell division

Size-scaled histogram



Variability of gene expression

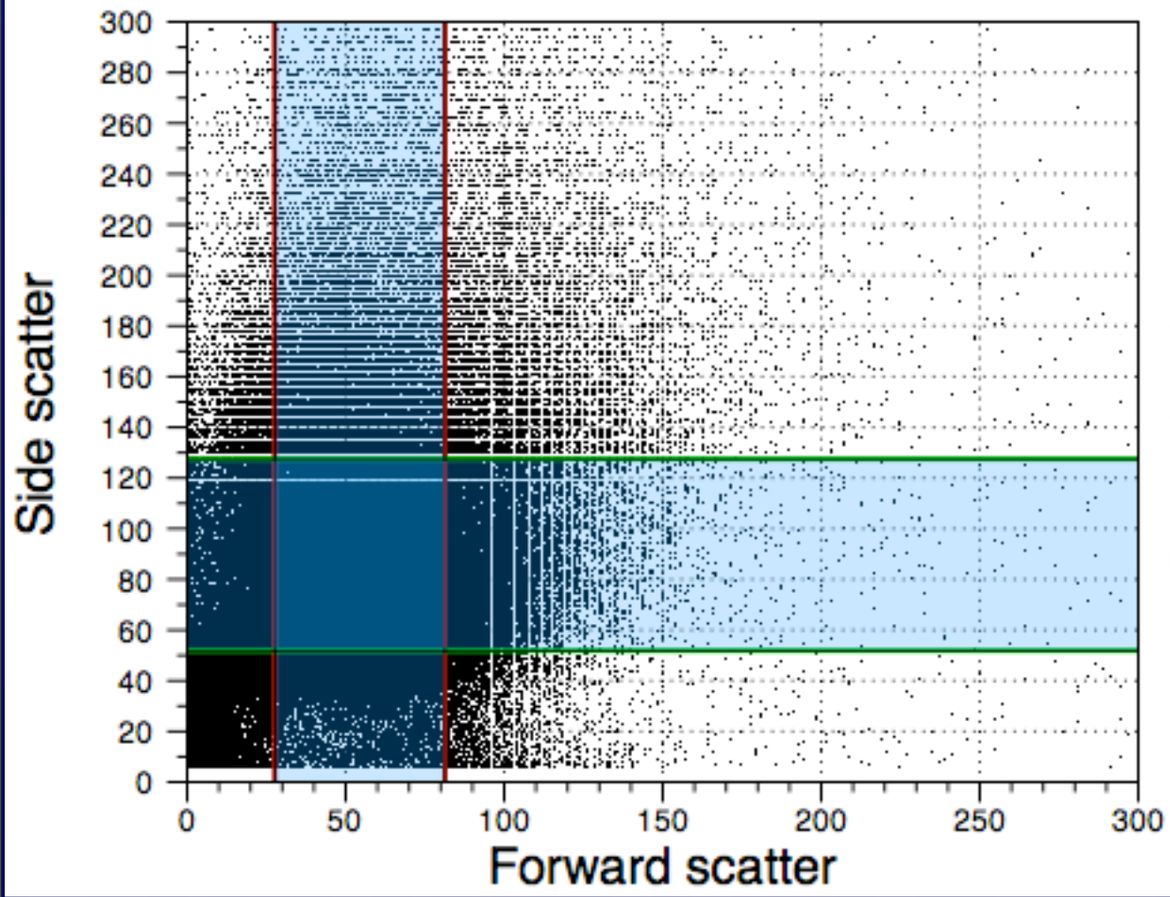
Cell strain	Plasmid	%CV (cytom.)	%CV (micro.)	%CV (micro.) (size scaled)
DH5a	High	55.0	58.4	20.2
	Medium	52.9	56.0	12.6
Top10	High	55.7	76.4	25.0
	Medium	57.7	74.1	12.8
B/r	High	51.5	66.8	24.2
	Medium	58.5	61.0	12.4
BL21*	High	86.0	78.4	29.0
	Medium	75.2	51.7	12.1

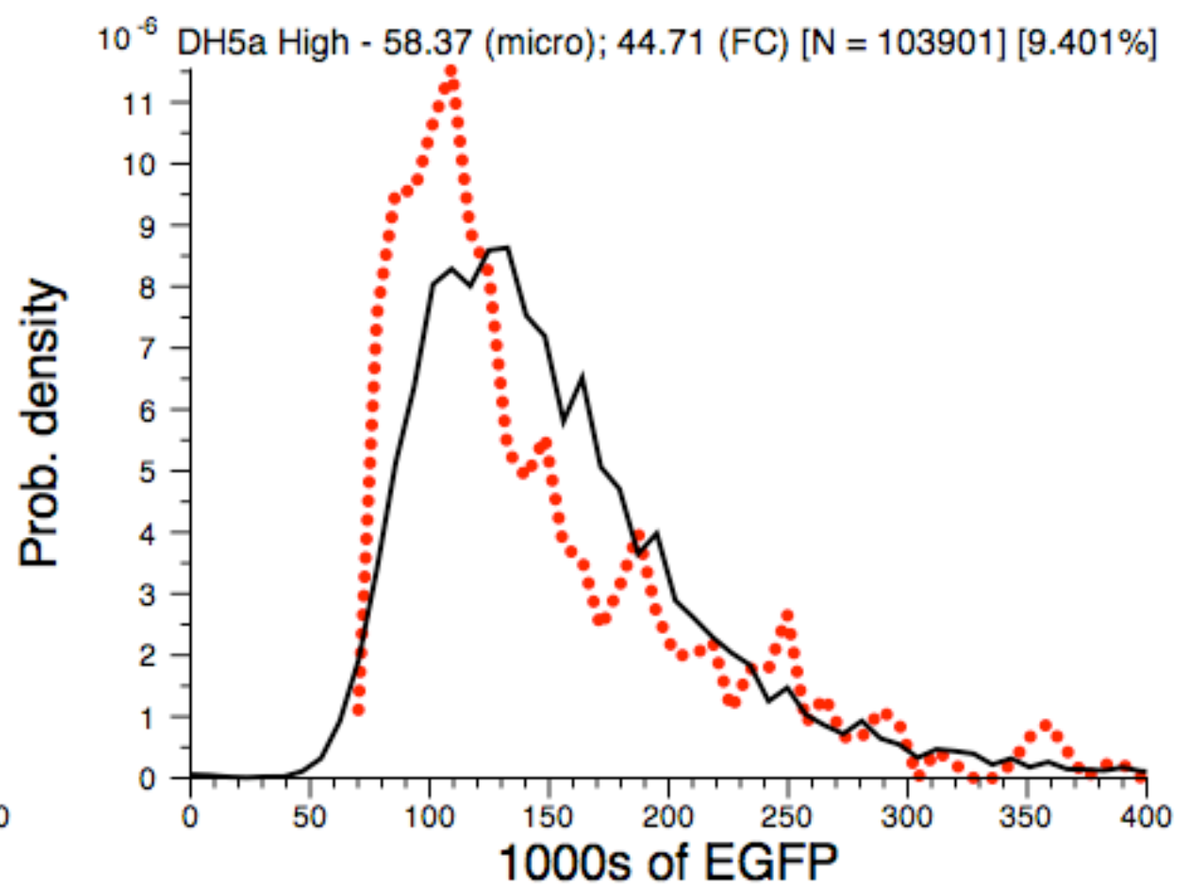
- Substantial drop in CV (std dev/mean) when size is scaled away; note also the consistently lower CVs for the medium plasmid (fewer proteins, but less variable!)

Negative feedback in plasmid copy number

- High-copy plasmid (pUC ORI) replicates as fast as it can, constrained only by resources
- Medium-copy (ColE1 ORI) plasmid incorporates negative feedback: interferes with its own replication
- Lower size-scaled CVs for the medium-copy plasmid suggest that the negative feedback keeps copy number less variable, and that shows up in protein CV

Size from cytometry?





- Restricting forward/side scatter region does not have same size-scaling effect as in microscopy

Movie done in DataTank (Visual Data Tools)

“Dark proteins” I

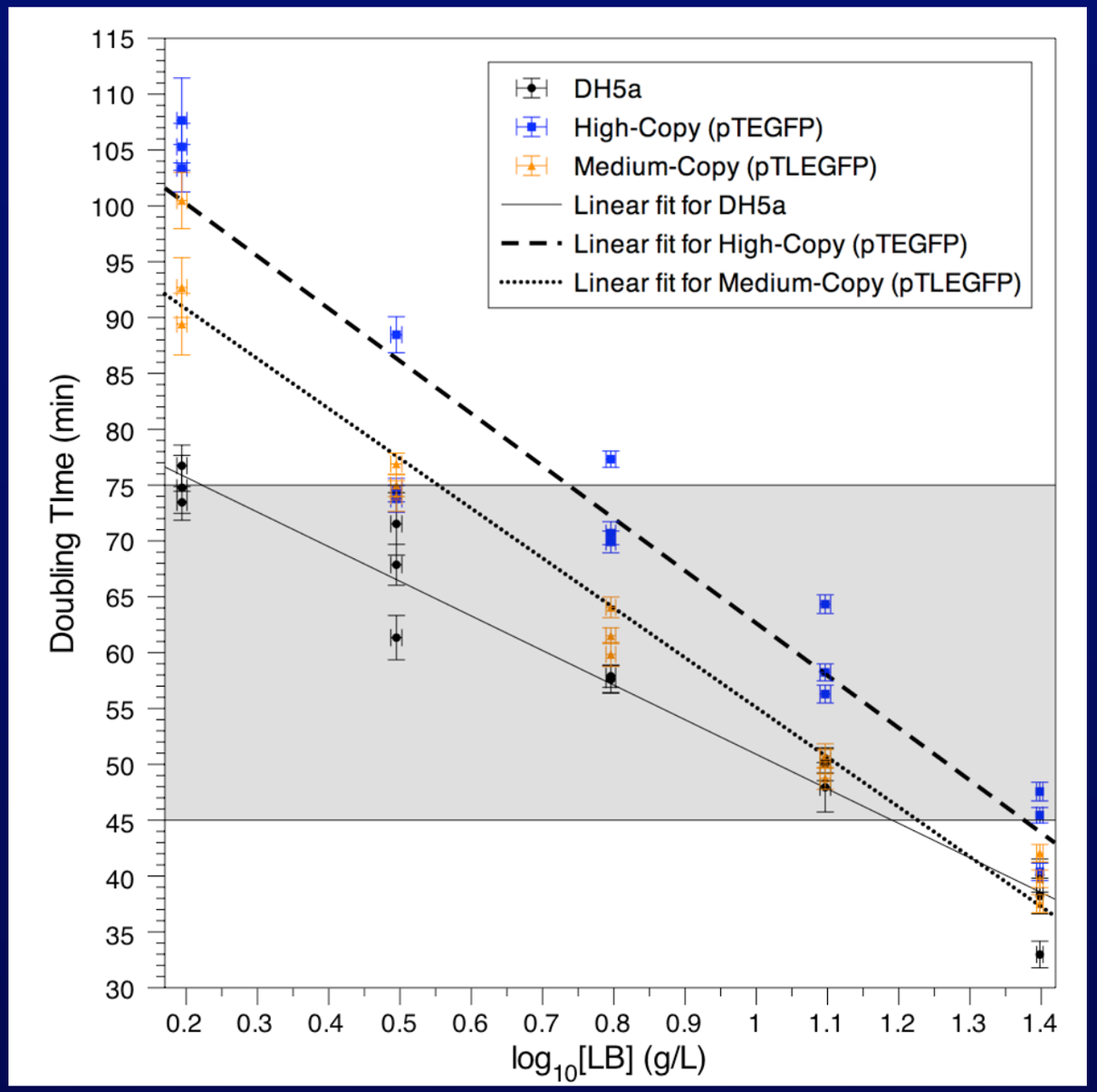
- A problem affecting all studies using fluorescence as a measure of gene expression:
  - You can (of course) only see the proteins that have become fluorescent
- Proteins may be invisible (“dark”):
  - If they are misfolded
  - If they have yet to mature

Inclusion body formation

- Many studies (including ours) use plasmid-based expression
  - Useful as a means of inserting tailored gene networks, to study gene dynamics or as control mechanisms
- High expression levels from plasmids can lead to formation of inclusion bodies
  - Insoluble aggregates of misfolded, non-functional proteins
  - Fluorescent proteins caught in these bodies will not fluoresce

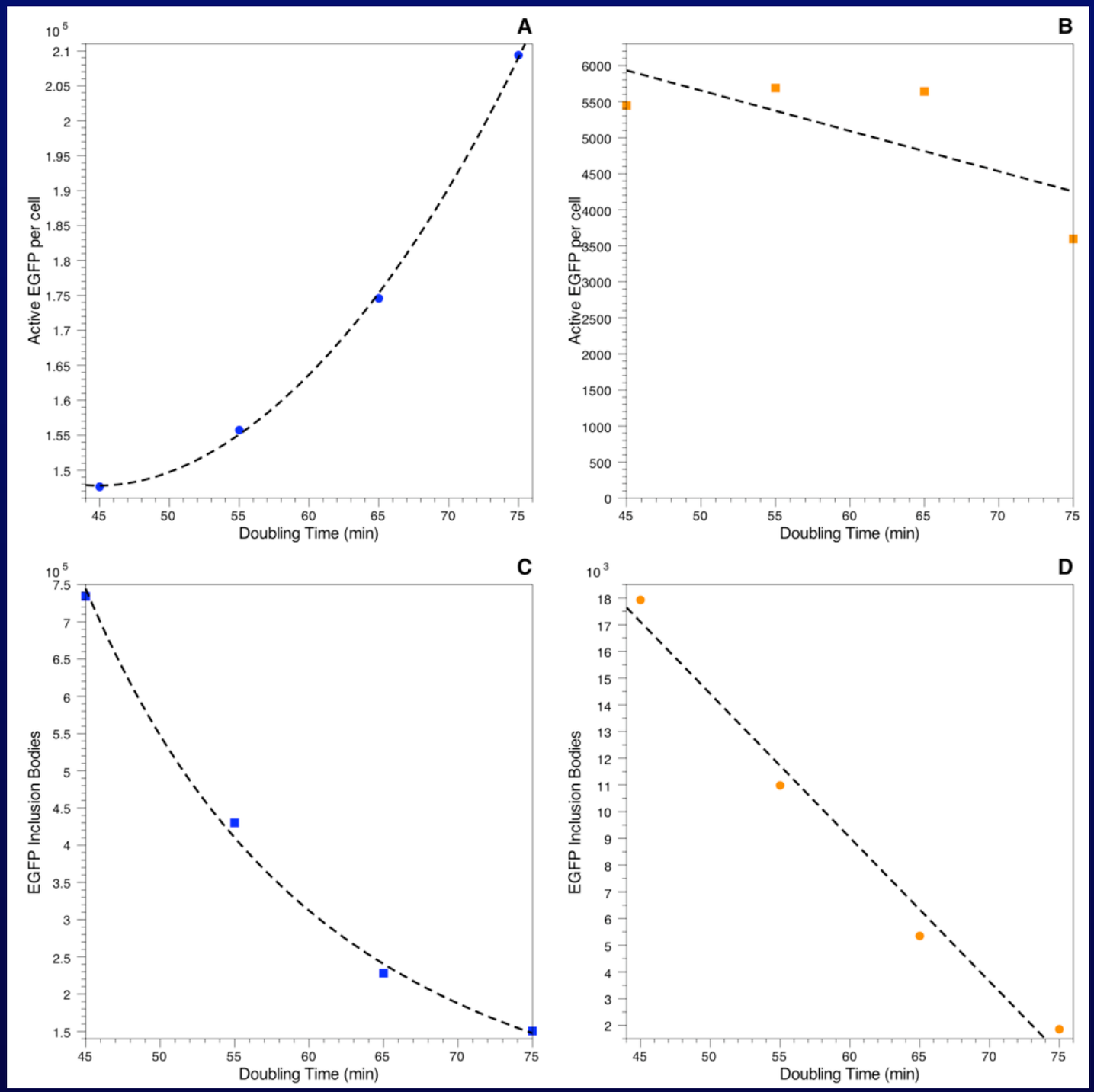
Marco Iafolla

Varying growth rates



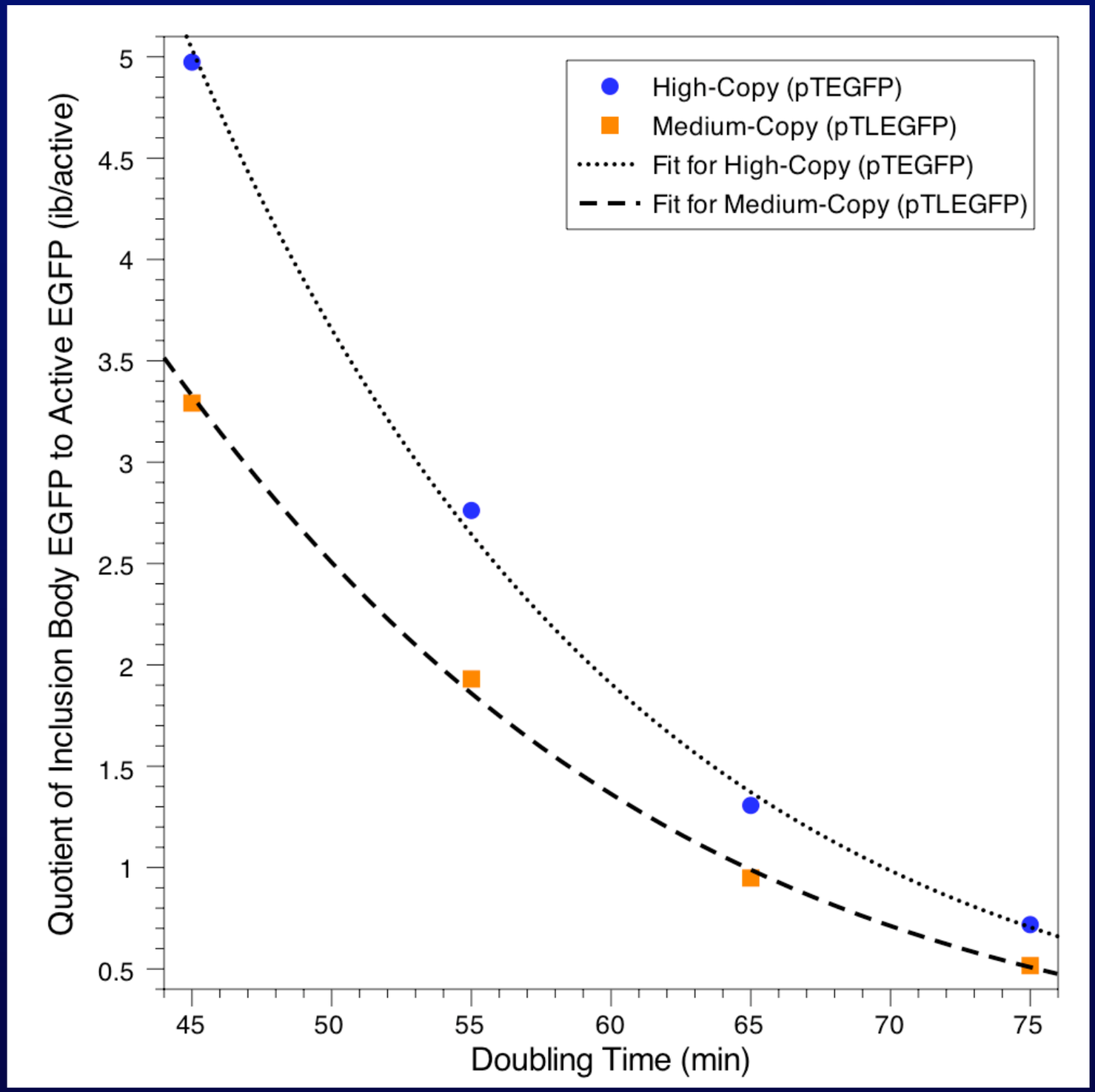
- Bremer and Dennis (1996): a very handy paper with many cellular parameters - all vary with growth rate

Extraction/quantification of inclusion bodies



Marco Iafolla

Ratios of IB to active EGFP

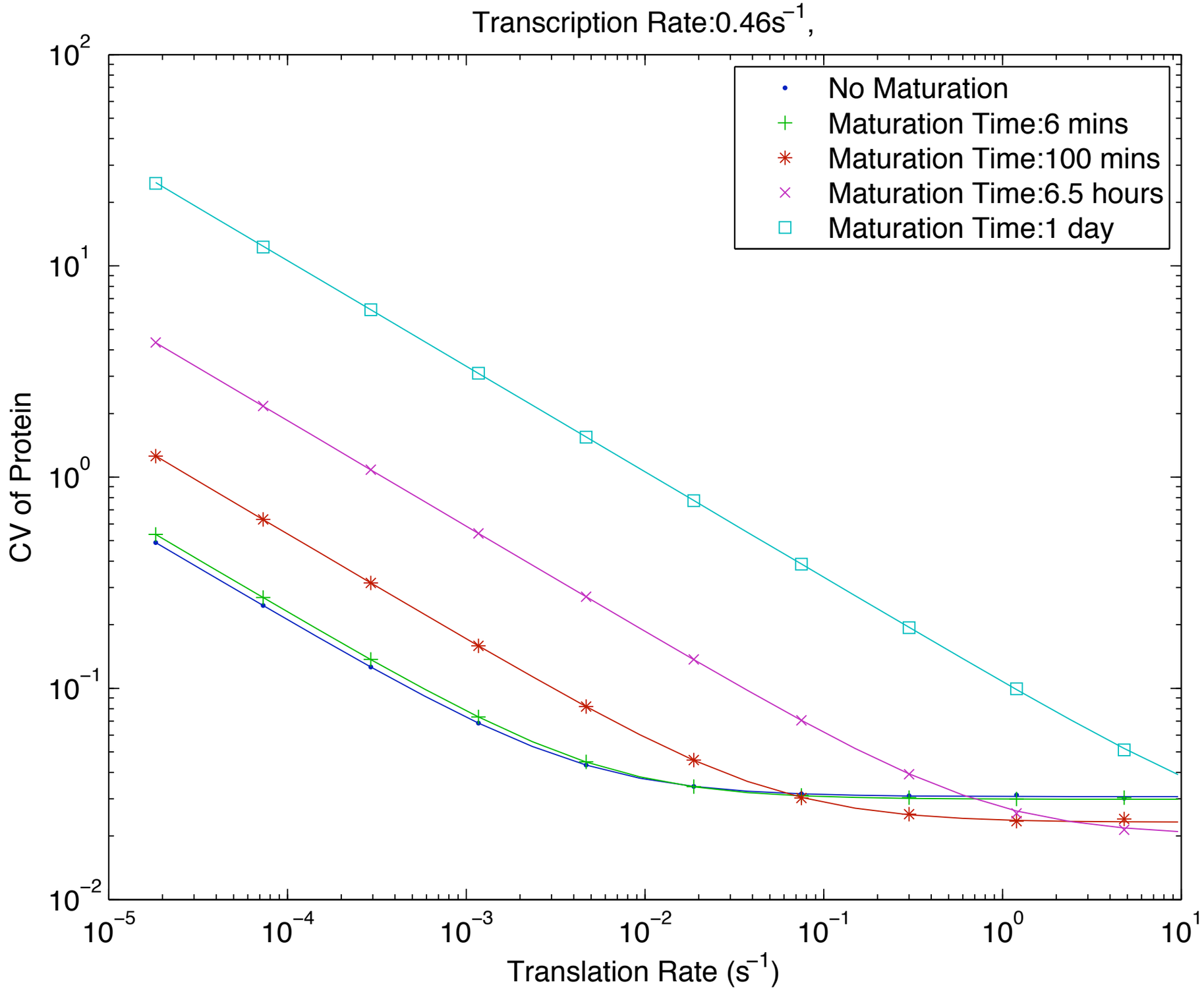


“Dark proteins” II: The Return

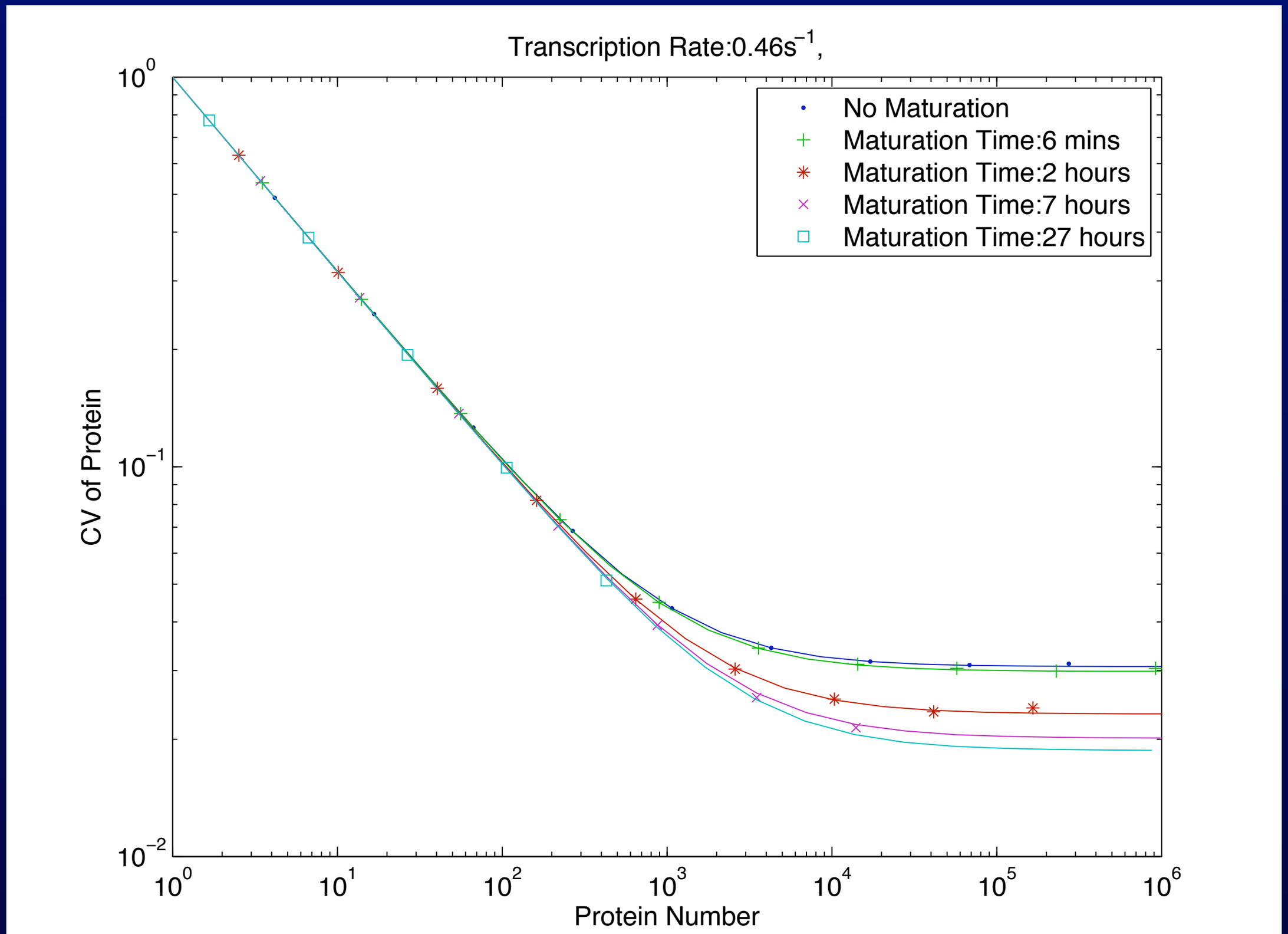
- Proteins are also “dark” in the transient period before they mature and start fluorescing
  - Folding, cyclization, oxidation
  - Maturation rates vary from minutes to days
- Modelling work:
  - At small numbers of proteins expressed, maturation effect increases observed variability (just by reducing numbers)
  - At large numbers, maturation can actually decrease observed variability

C. Guangqiang Dong

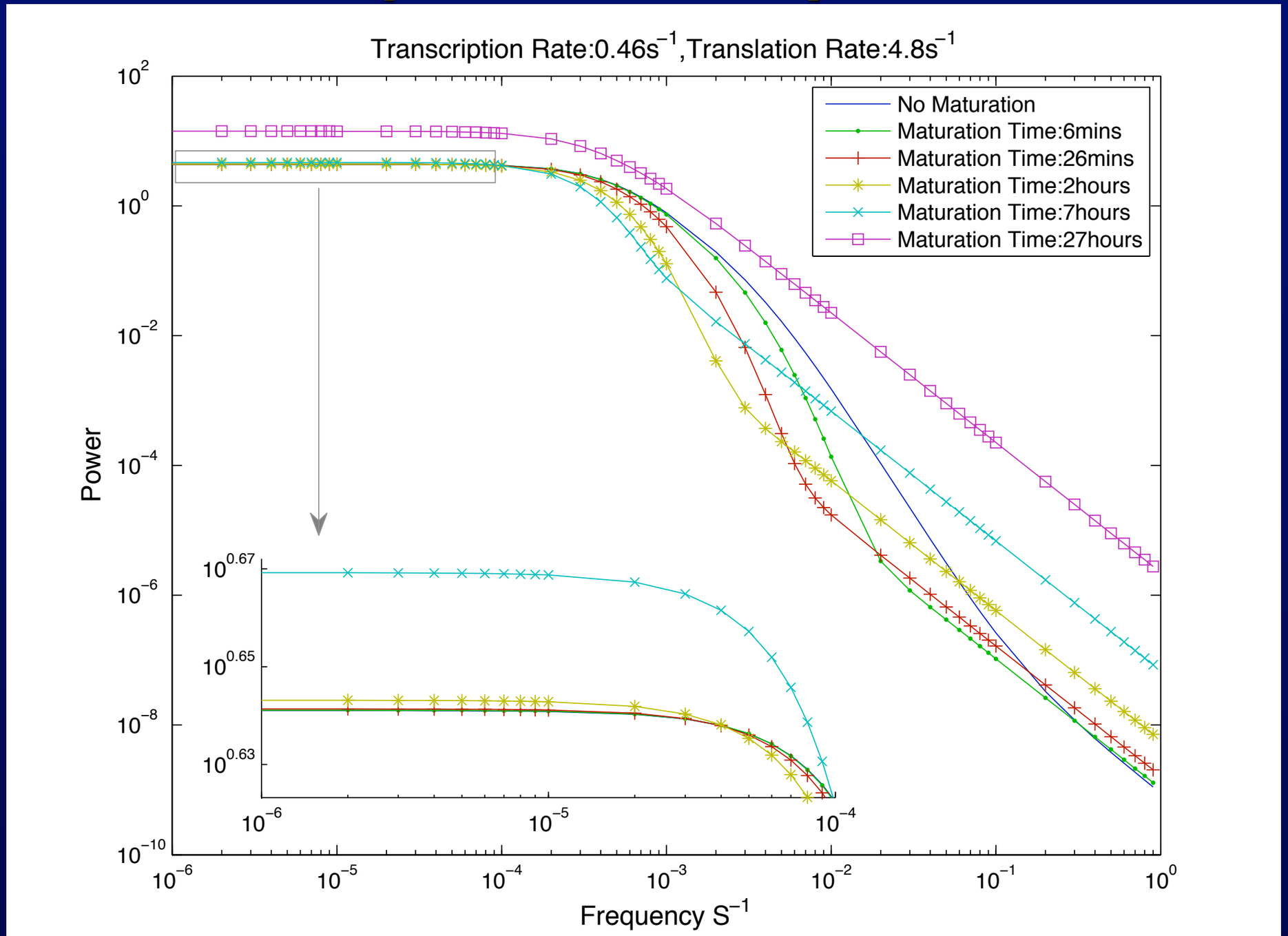
Protein maturation



CV vs <number>


Transcription Rate:0.46s^{-1},
No Maturation
Maturation Time:6 mins
Maturation Time:2 hours
Maturation Time:7 hours
Maturation Time:27 hours
CV of Protein
Protein Number
10^{0}
10^{-1}
10^{-2}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
10^{6}

Noise power spectrum



- Coming soon: experiments! Vary: maturation rate (proteins), transcription rate (promoters), translation rate (RBS) (Sangram Bagh, C. Guangqiang Dong)

4. Cells are “individuals”

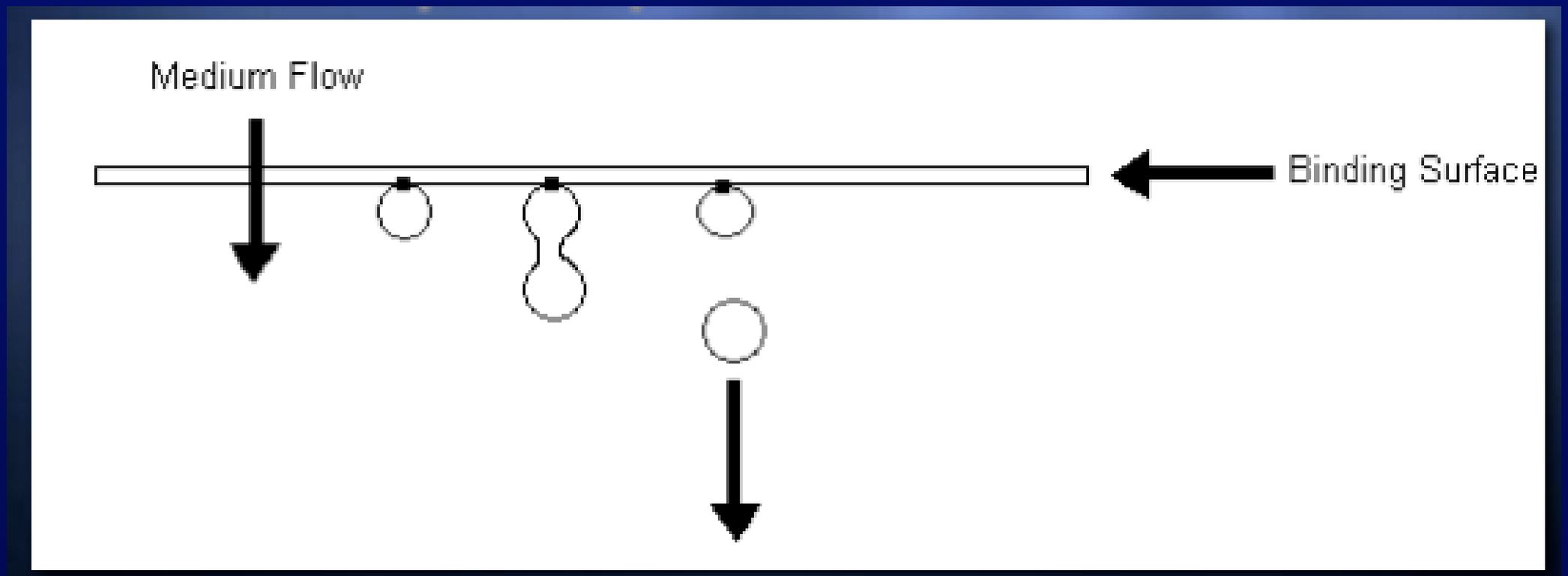
- Cells even in an apparently homogeneous population can have varying individual histories (growth/ division events, nutrient exposure, cumulative effects of fluctuations)
  - Different histories lead to different states
  - Plasmid copy number and cell size: can be viewed as extrinsic or intrinsic noise sources
- We’re interested in the effects of various perturbations on cells’ behaviour

Changing growth conditions

- Using a "baby machine" to examine synchronized cells
  - Early results: synchronized cells (all in nearly same phase of "cell cycle") are less variable than asynchronous cells
  - More than just a size effect: reduction is in the size-scaled variability levels
- Investigating effect of varying nutrient levels by growing cells in a chemostat
  - Early results: chemostat ("continuous culture") cells are more variable than exponentially growing cultures

Luke Jakobowski

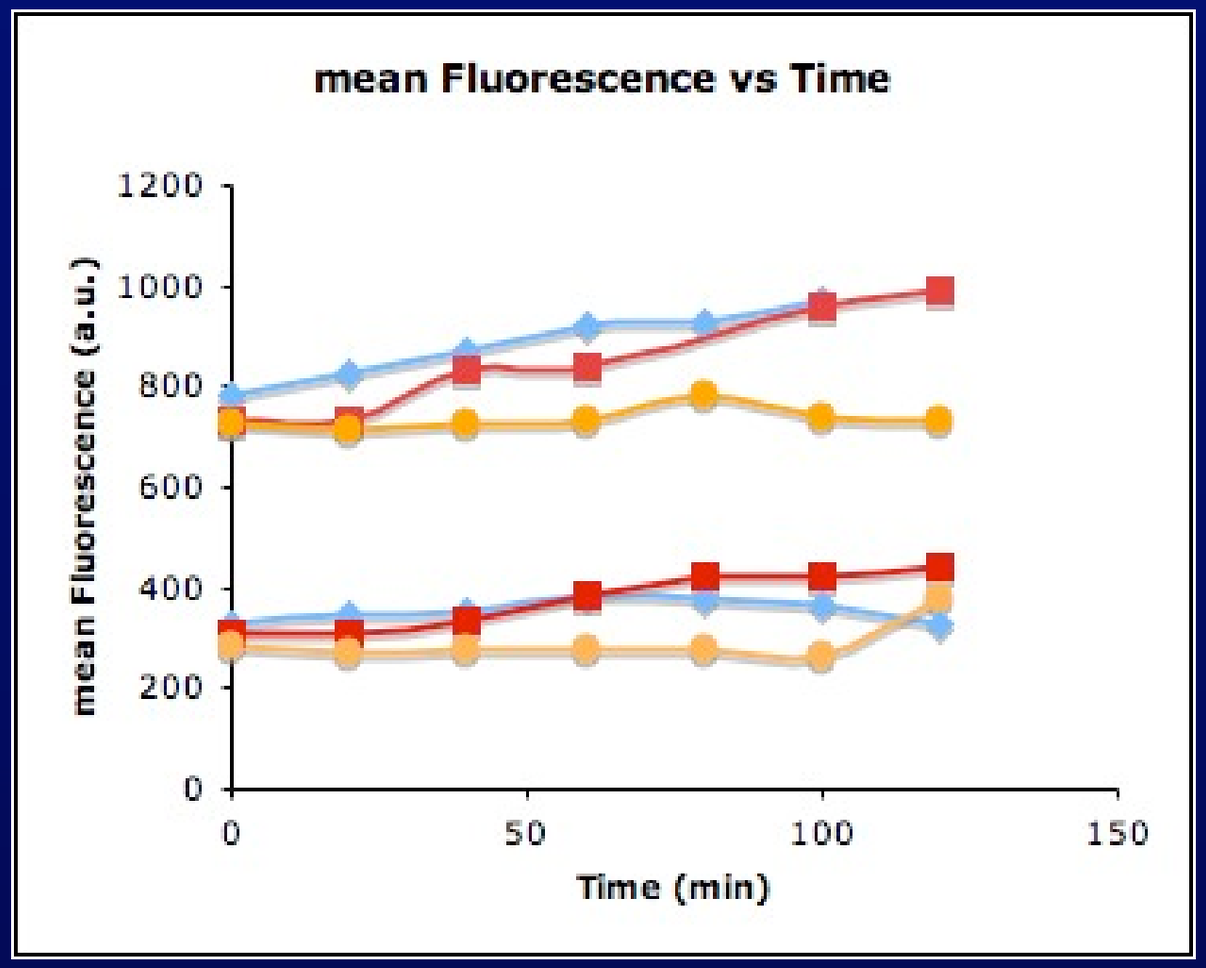
The baby machine



Perturbation by cell sorting

- Cell sorting: can use the flow cytometer to select cells based on any optical output, divert desired cells into a tube
- We’re using this to sort out subpopulations based on brightness (related to EGFP expression level)
  - Observe the distribution of the perturbed population over time, see how (or if) it relaxes back to original distribution

Early cell-sorting results



- Sorted top 10% brightest cells
- Population stays bright over 3 hours
- Grow overnight: back to original dist.(?)

Tharsan Velauthapillai, Mostafizur Mazumder

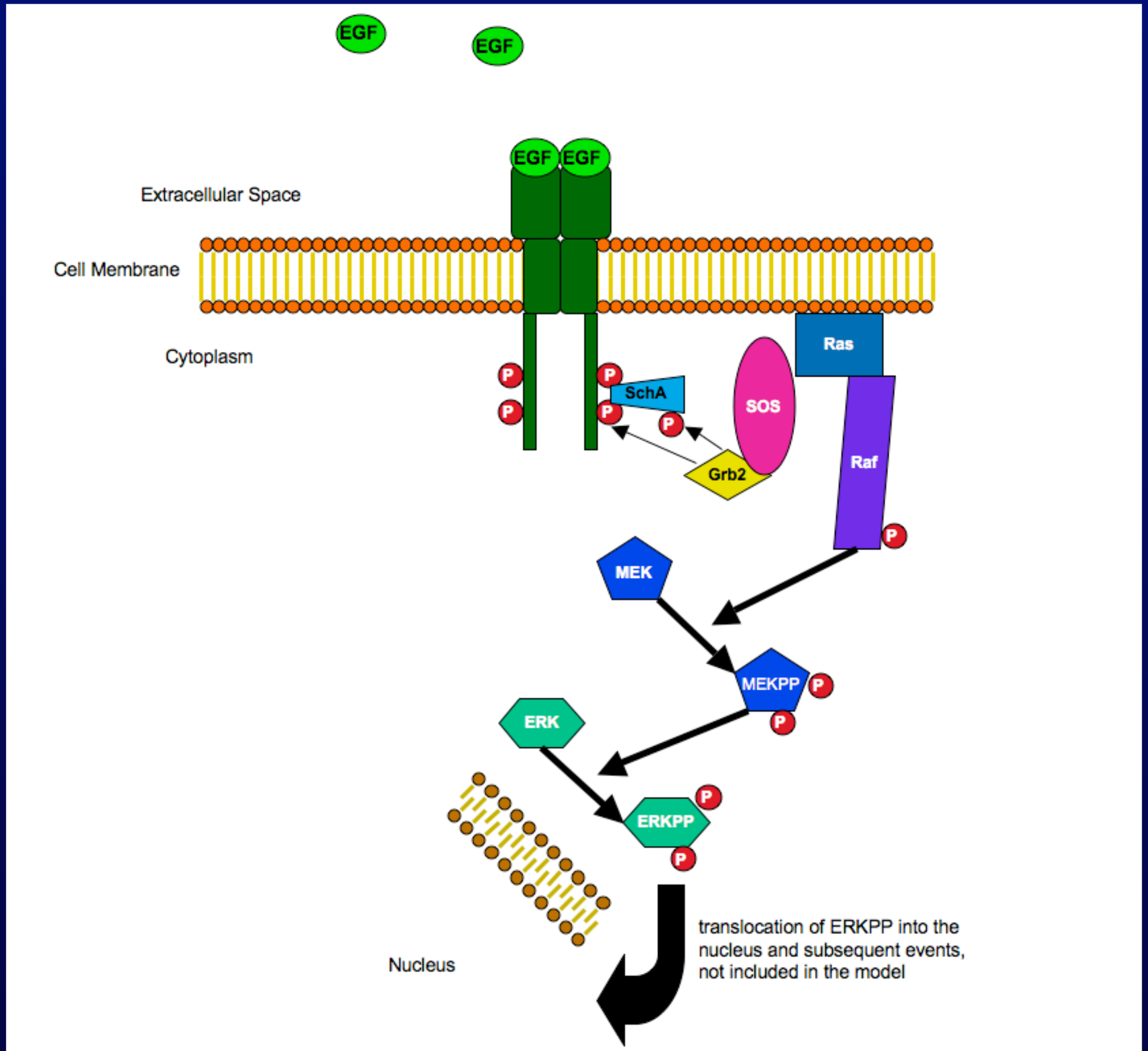
5. Cells are complicated

- Many species and reactions are involved in even a moderately complete description of any biological system
  - Complete model may require dozens (or literally millions!)
- Model reduction methods have been explored for deterministic ODEs
  - Based on time scales, quasi-steady states...
- We're working on methods that can be applied to stochastic systems as well

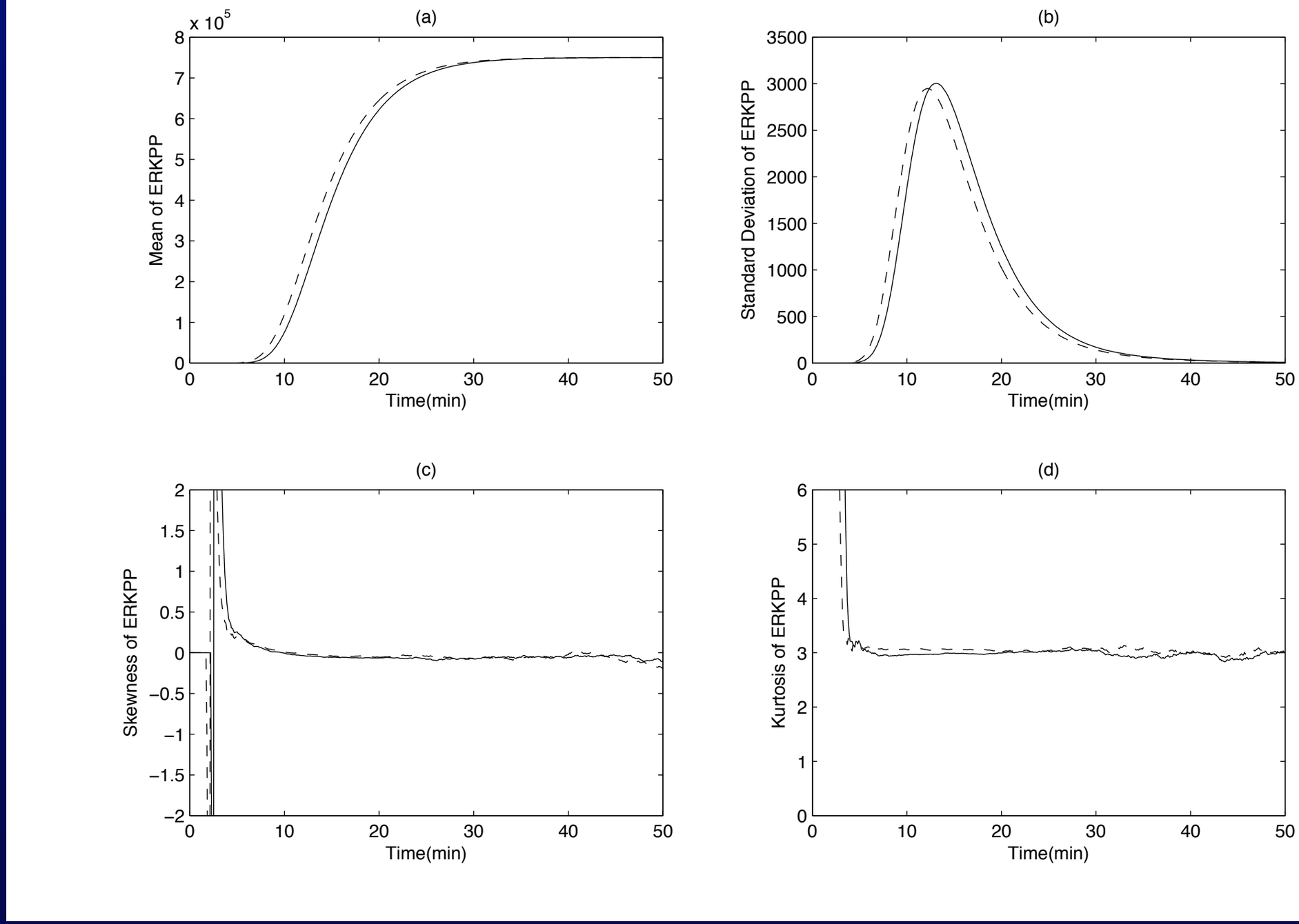
Stochastic model reduction

- Use a known deterministic ODE reduction method:
  - Partition reactions into fast/slow
  - Form reduced system by approximating fast steps as near-instantaneous
- Translate reduced system back into a reaction scheme
- Use stochastic simulator to run it, generate fluctuations
- Result: highly reduced systems still match statistical behaviour of original

Signal cascade model



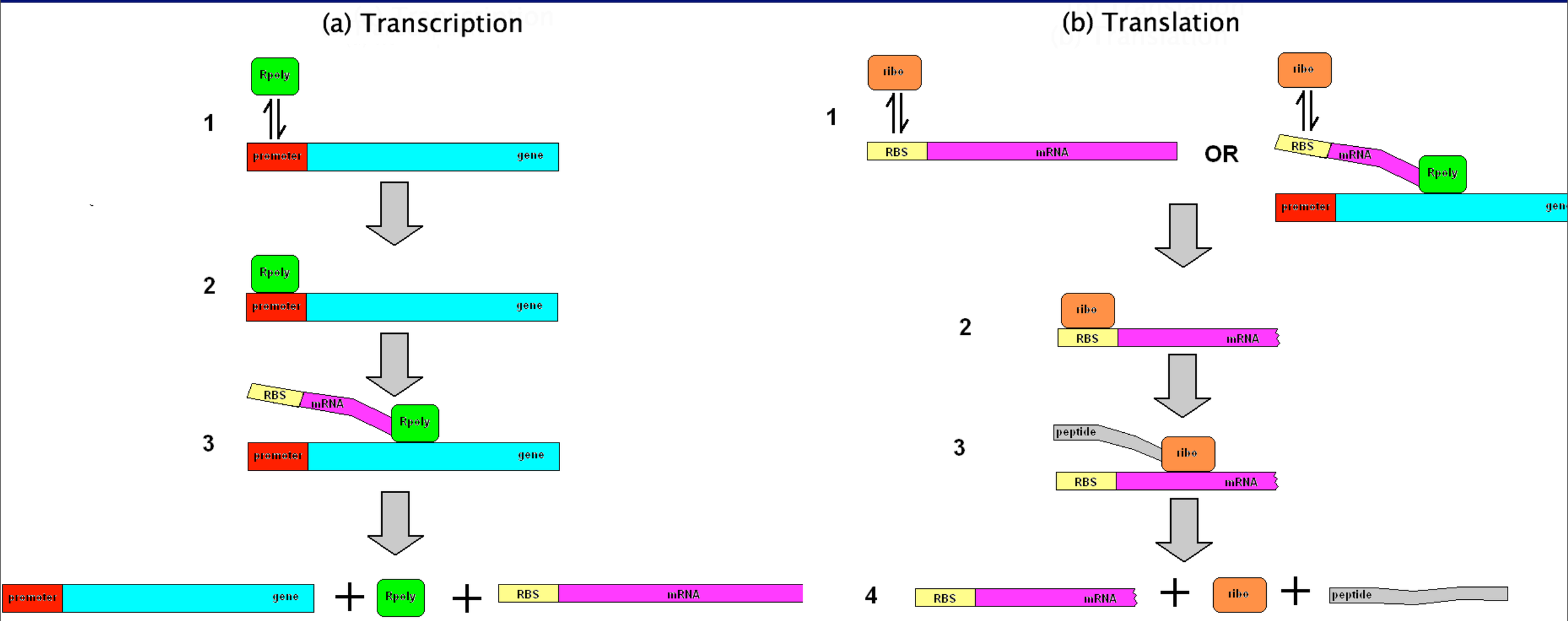
Reduced vs original



- Original: 63 reactions, 41 species
- Reduced: 34 reactions, 27 species
- Dong et al, *J Biol Phys* **32**: 173 (2006)

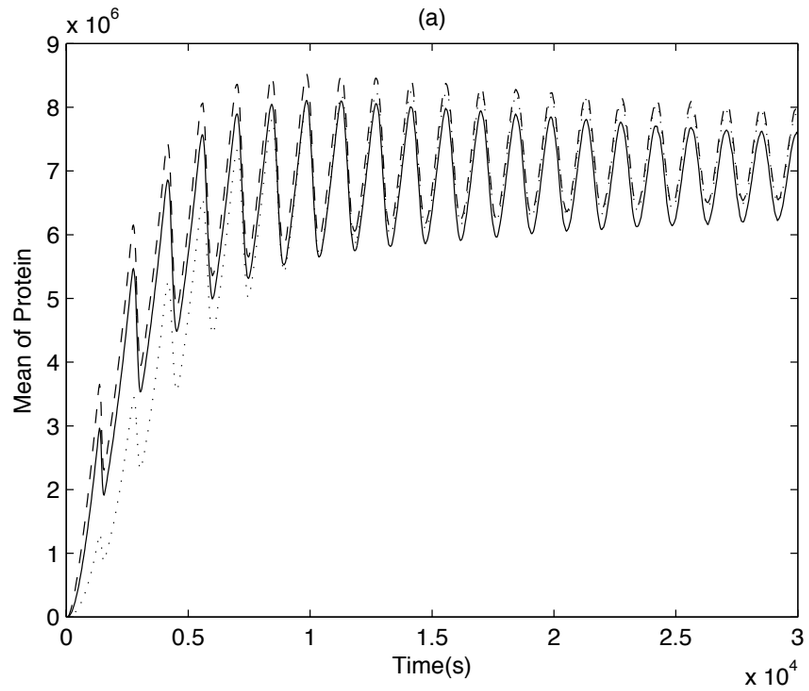
C. Guangqiang Dong

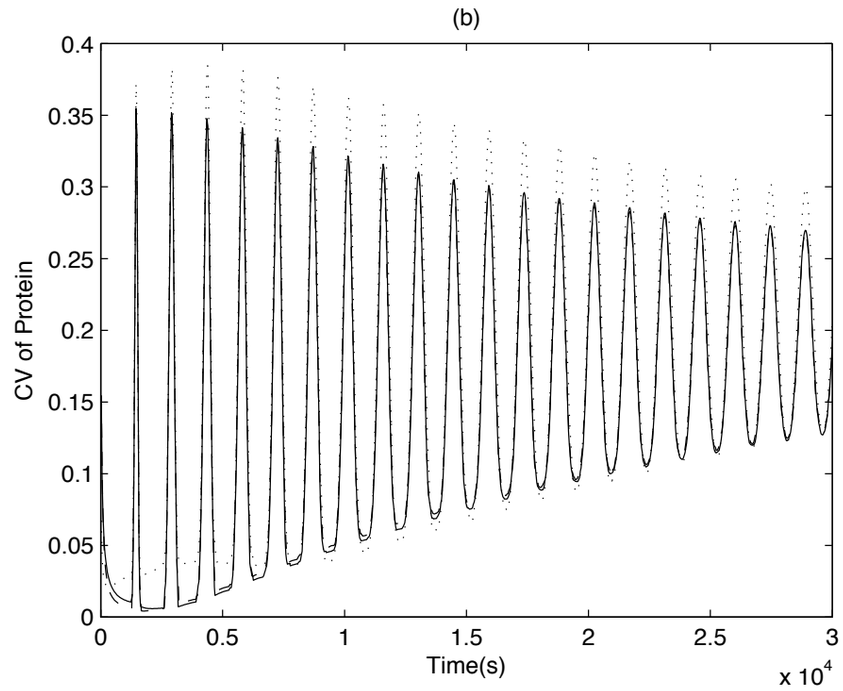
Gene expression model

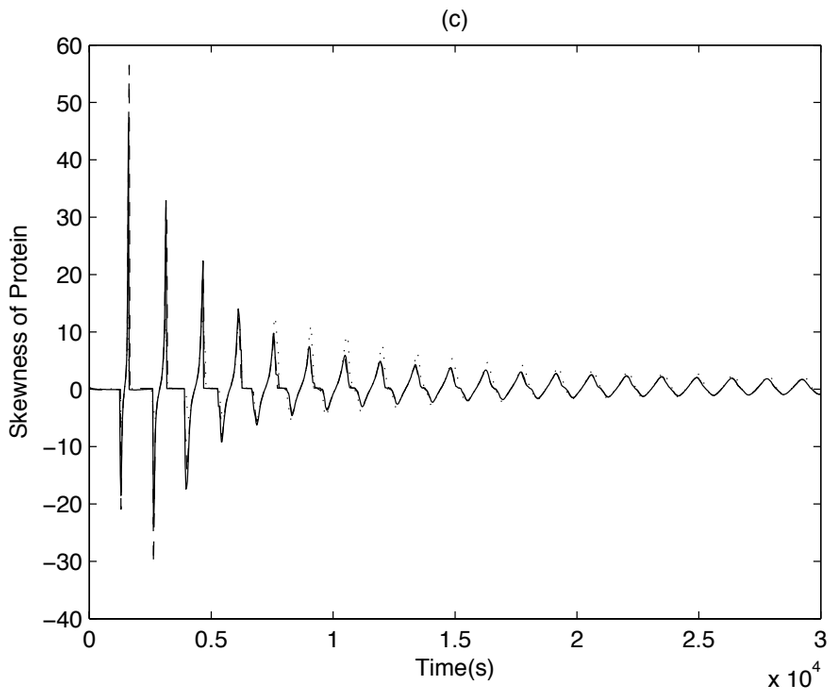


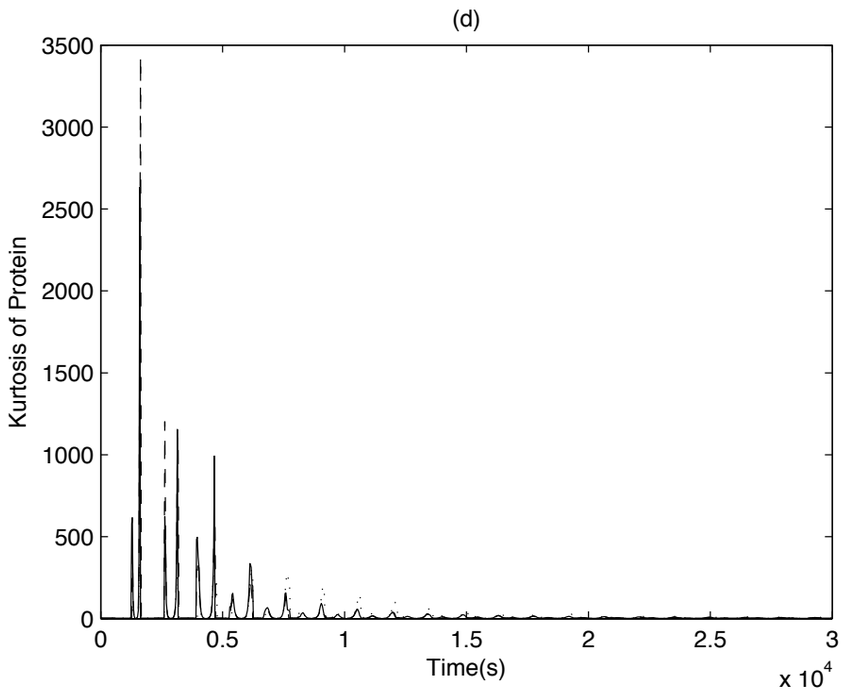
- Original: 71 reactions, 47 species [Iafolla and McMillen, *J Phys Chem-B* **110**: 22019 (2006)]
- Reduced to two different levels:
  - 28 reactions, 29 species
  - 10 reactions, 10 species

Reduced vs original









Caveats

- Works near-perfectly for cascades (feed-forward)
- Still works well for "moderate" feedback
  - But as you crank up feedback strength enough, will eventually break down
- Method can fail with an non-invertible matrix in one of the steps
  - So far, we can't precisely identify the conditions for this to occur
- May need to adjust parameters to achieve an optimal fit

6. Cells are crowded

- Interior of a cell is far from being a dilute solution; more like a paste/gel
- Standard modelling methods tend to assume a well-mixed system
  - Don't keep track of locations of individual molecules, just total concentration or #
  - Crowded environment can have impact on kinetics, and on variability
- We're working on experiments in vitro, to vary crowding level and observe effects
  - Collaboration with Ray Kapral (elegant, fast method of simulating spatial behaviour)

Issues with biochemical kinetic models

- Sometimes I ponder the following two facts:
- Fact #1: Biochemical kinetic models yield useful results
  - Predictive abilities have been demonstrated
- Fact #2: Biochemical kinetics seemingly cannot be *right*, as usually written

Conclusions

- I have no Conclusions
- But some things to note:
  - Biology is interesting, even for physical scientists and mathematicians
  - It's a huge challenge: everything in biology is the least-tractable example of its class
    - Nonlinear, stochastic, far from equilibrium, nonideal, messy, actively uncooperative
  - Fortunately, we tend to love a challenge
  - There's something to be said for working closely with living cells: the daily weirdness tends to promote humility

Acknowledgements

- Funding:
  - NSERC Discovery grant
  - CFI New Opportunities grant (shiny new cell-sorting flow cytometer, baby!)
- People:
  - Christopher McCulloch and Wilson Lee, UofT Dentistry (initial sorter access)
  - The Quantitative Biology Lab students, who do all the real work

Turn back! Extra slides past this point

Expression noise

- Gene expression noise is parcelled into two sources:
  - Intrinsic: from small-number effects on the gene itself
  - Extrinsic: everything else (cell-to-cell variation in components outside gene, e.g. numbers of the enzymes that drive transcription/translation)
- Extrinsic noise: individual cellular environments cause different behaviour
  - Different histories lead to different states
  - Plasmid copy number and cell size: can be viewed as extrinsic or intrinsic noise sources