(Some) coarse-grained modeling in systems biology

Ilya Nemenman

Departments of Physics and Biology
Computational and Life Sciences Initiative
Emory University

nemenmanlab.org

JAMES S. MCDONNELL FOUNDATION

NATIONAL CANCER INSTITUTE

National Institute of General Medical Sciences

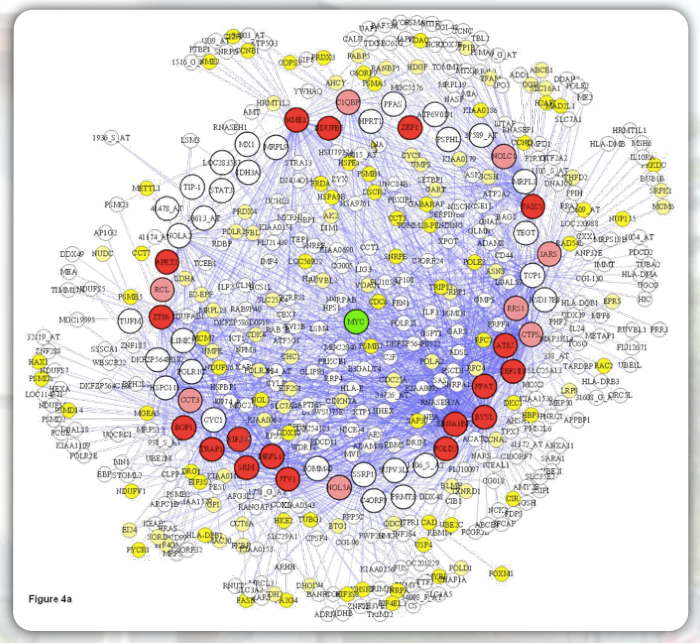
HFSP

NSF

EMORY UNIVERSITY

The -omics revolution in biology

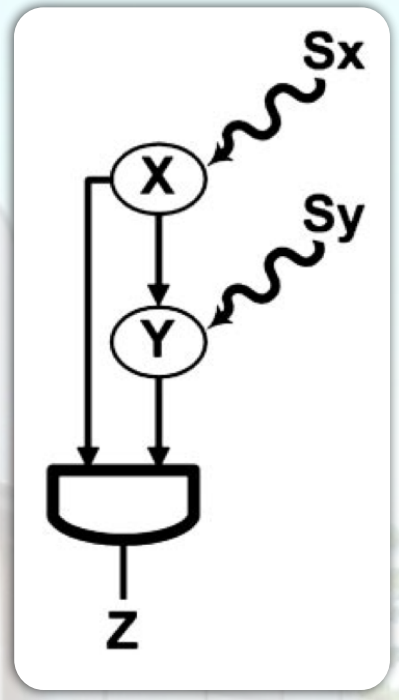
- Breaking life into ever more accurate parts lists
  - Sequences: genomics, metagenomincs, epigenomics,…
  - Activities: gene expression, metabolic profiling, phosphoproteomics, electrophysiology …
  - Zoology of molecules — like cataloging high energy resonances in 1970s.
- Putting it all back into a network of interactions
  - Metabolic, transcriptional, protein signaling, neural, networks…
  - **Which things go together?**
  - **Number of possible interactions is astronomically large.**



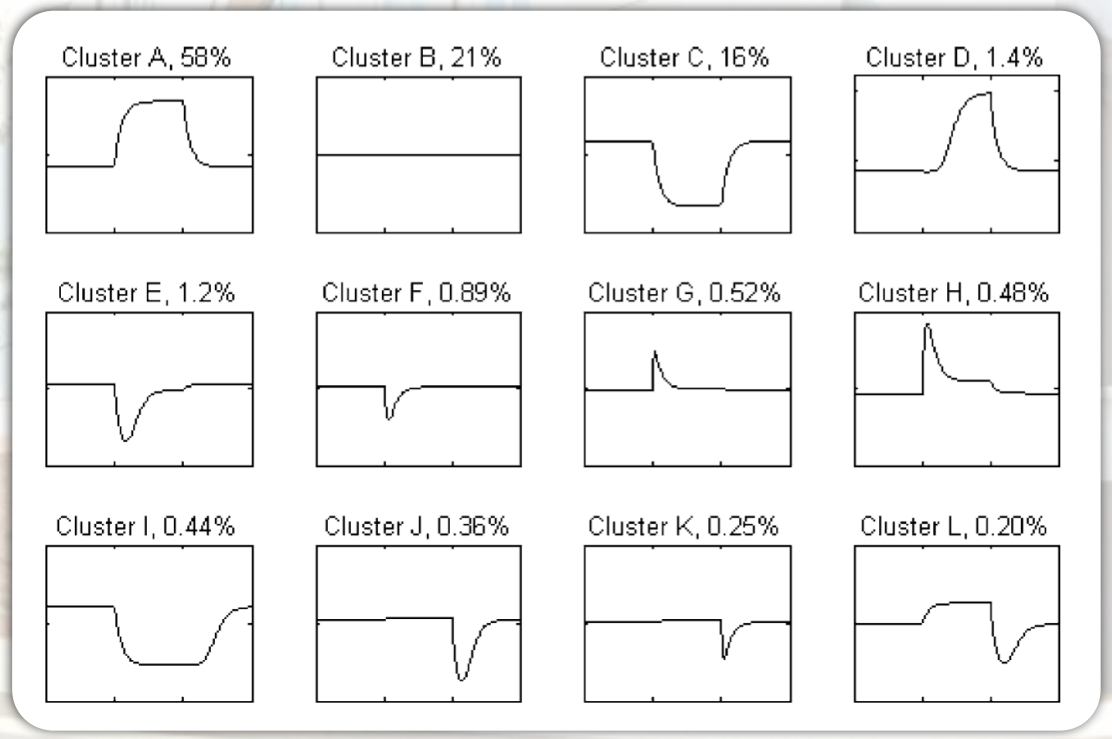
Califano et al., Nat Gen 2005; BMC Bioinf 2006

Attempting to address the complexity: Coarse-graining networks into functional modules

- Groups of interacting molecules are like *modules* in engineering systems.
- But function **(dynamics)** of a module doesn't easily follow from its constitutive parts.



Mangan and Alon, PNAS 2003

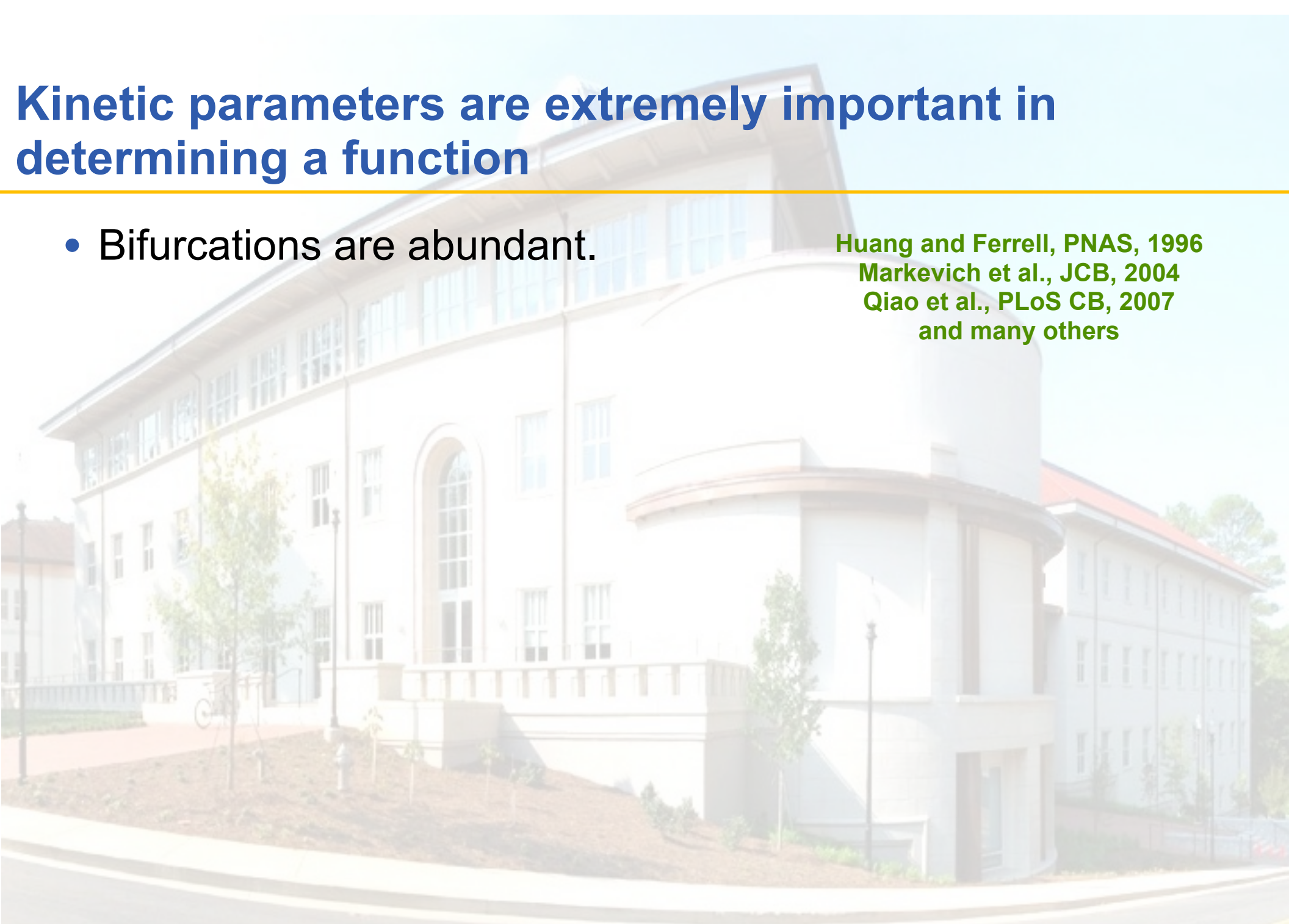


Wall et al., JMB 2005

Kinetic parameters are extremely important in determining a function

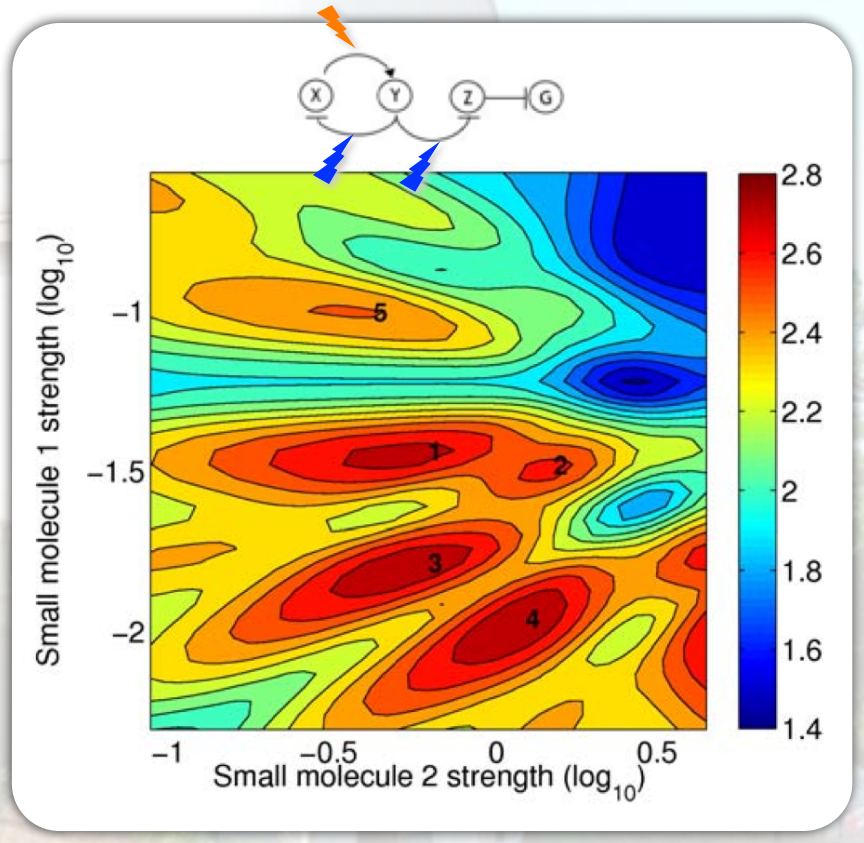
- Bifurcations are abundant.

Huang and Ferrell, PNAS, 1996
Markevich et al., JCB, 2004
Qiao et al., PLoS CB, 2007
and many others



Kinetic parameters are extremely important in determining a function

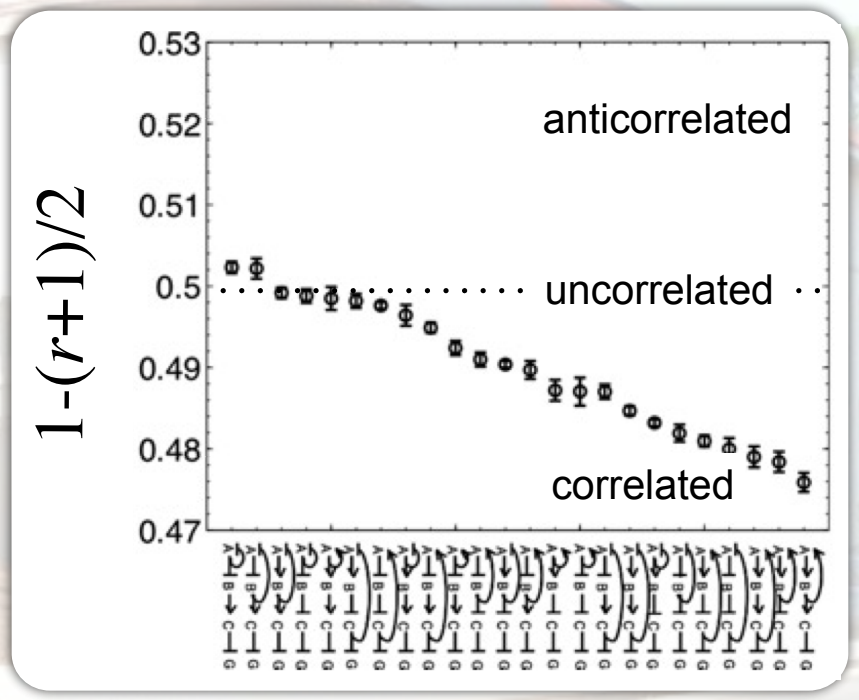
- Bifurcations are abundant.
- The same cellular network can perform multiple accurate (logical) functions.
  - see also **Tikhonov and Bialek, arXiv 2013.**



Ziv, IN, Wiggins, PLoS ONE, 2007

Kinetic parameters are extremely important in determining a function

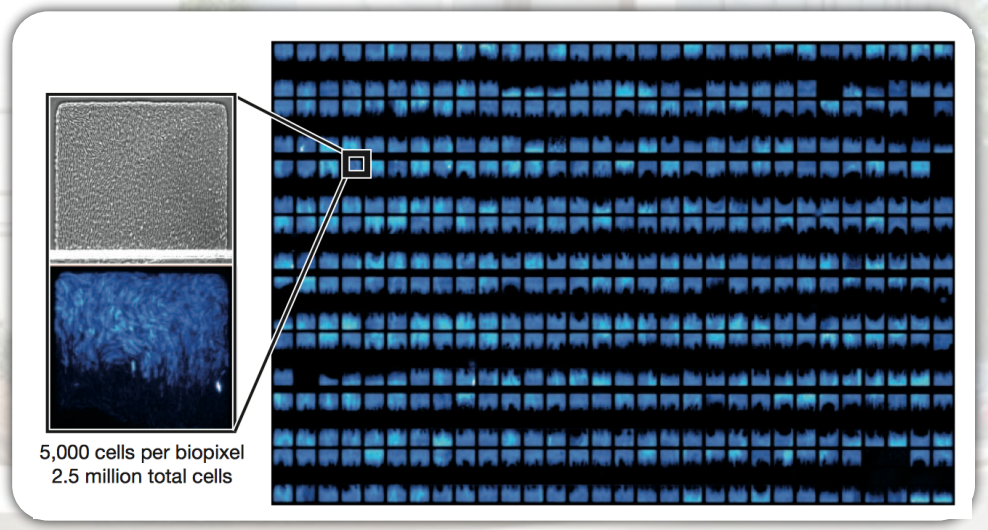
- Bifurcations are abundant.
- The same cellular network can perform multiple accurate (logical) functions.
- Correlations between parameter changes and the resulting function changes are weak.
  - see also **Sethna et al., PLoS CB 2007, …, Science, 2013.**



Mugler, Ziv, IN, Wiggins, IET SB, 2009

Kinetic parameters are extremely important in determining a function

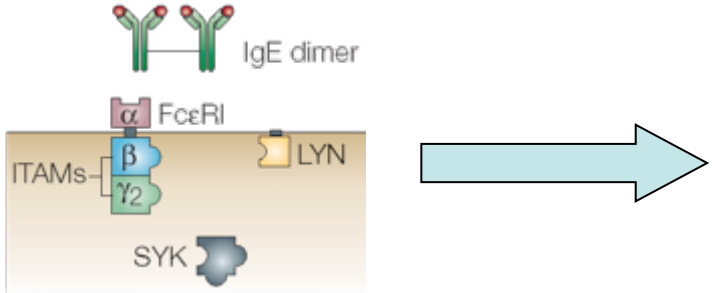
- Bifurcations are abundant.
- The same cellular network can perform multiple accurate (logical) functions.
- Correlations between parameter changes and the resulting function changes are weak.
- Small un-anticipated interactions can have dramatic functional effects.

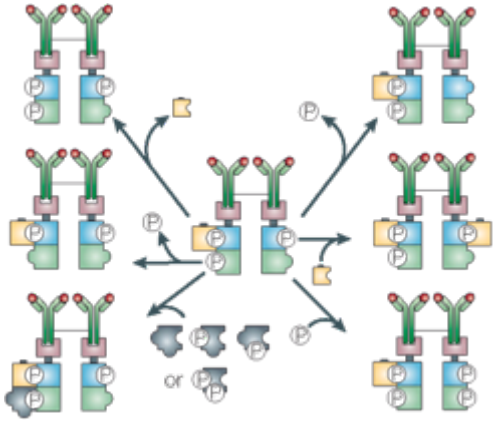


Prindle et al., Nature 2012

Have I made a case for dyna-omics?

- To predict dynamics, we will need to measure details of many interactions with excruciating details.
- The number of interactions is combinatorially large compared to the (large) number of interacting components.





Goldstein, Hlavacek, Faeder, et al., 2000-2009

354 species / 3680 reactions
(2954 for trimers)

- Is this program feasible? Can we do better if we only need the macroscopic **dynamics**, but not the microscopic accuracy per se?

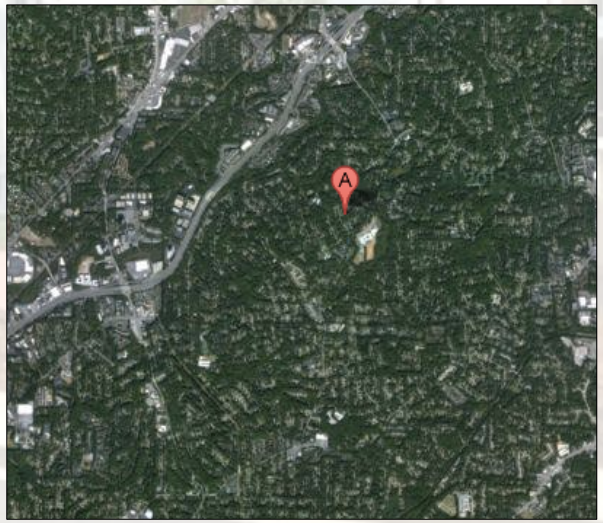
Of exactitude in science

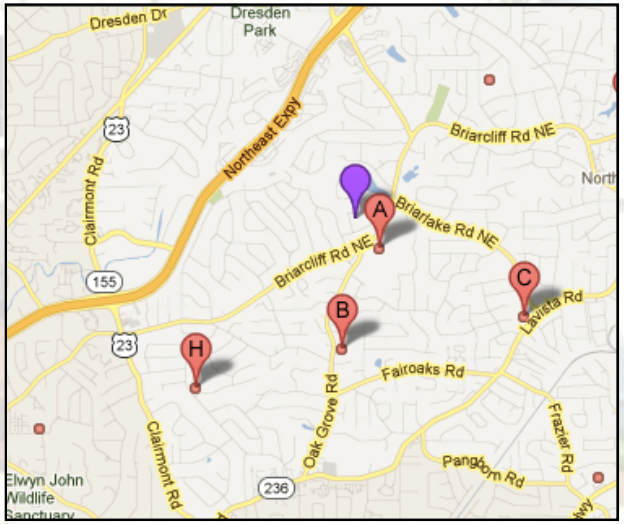
...In that Empire, the craft of Cartography attained such Perfection that the Map of a Single province covered the space of an entire City, and the Map of the Empire itself an entire Province. In the course of Time, these Extensive maps were found somehow wanting, and so the College of Cartographers evolved a Map of the Empire that was of the same Scale as the Empire and that coincided with it point for point. Less attentive to the Study of Cartography, succeeding Generations came to judge a map of such Magnitude cumbersome, and, not without Irreverence, they abandoned it to the Rigours of sun and Rain. In the western Deserts, tattered Fragments of the Map are still to be found, Sheltering an occasional Beast or beggar; in the whole Nation, no other relic is left of the Discipline of Geography.

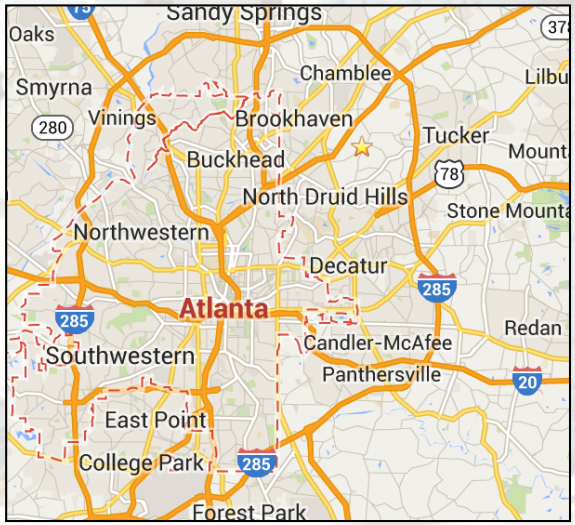
From Travels of Praiseworthy Men (1658) by J. A. Suarez Miranda (a fictional reference).
By Jorge Luis Borges and Adolfo Bioy Casares.
English translation quoted from J. L. Borges, A Universal History of Infamy,
Penguin Books, London, 1975.

Simplifying complexity?

- Models must loose details. Otherwise…
  - The best material model of a cat is another, or preferably the same, cat. (***Philosophy of Science*, Wiener and Rosenblueth, 1945**)
- Each modeling level needs its own **effective** degrees of freedom
  - "Don't model bulldozers with quarks." **(Goldenfeld and Kadanoff, Science, 1999)**
- Adaptive coarsening is common in **physics** and **every-day life**
  - Which level of description is better for driving to a local school?







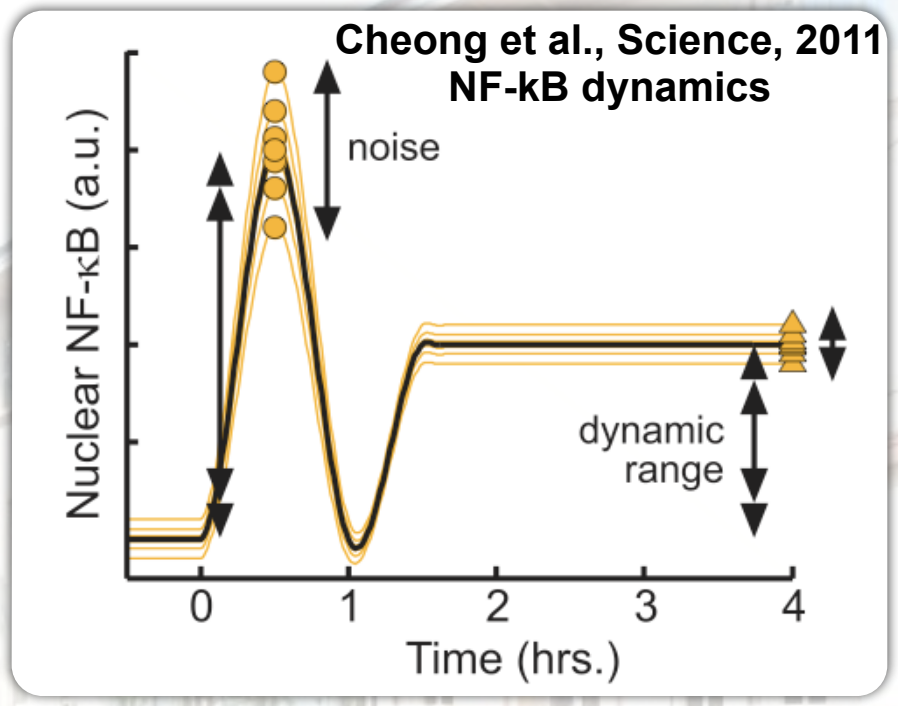
So...

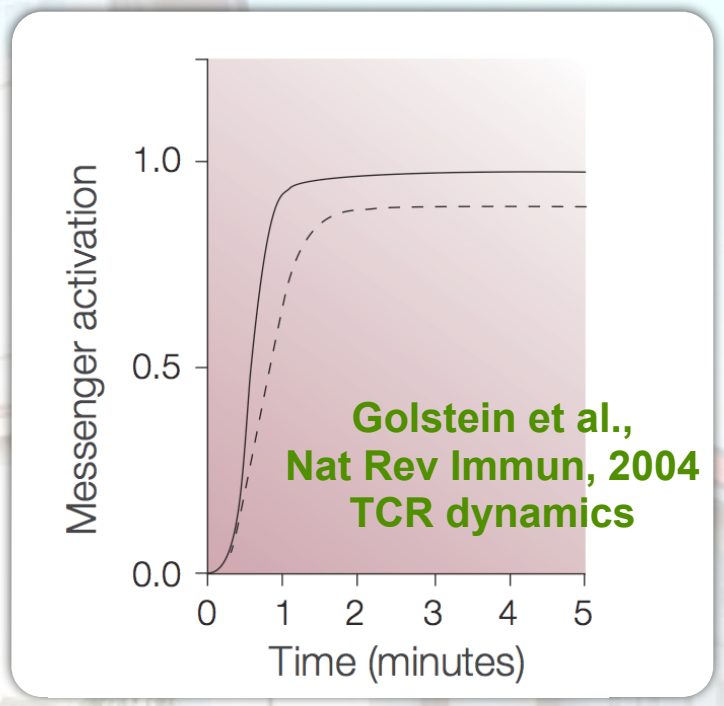
- Can we build **adaptive, phenomenological, coarse-grained,** and yet functionally accurate representations of (some) biological dynamics, or are we forever doomed to every detail mattering?
  - Examples of effective phenomenological models in physics that do not obviously follow from the microscopic description:
    - Ohm's law, Hooke's law;
    - Ideal gas law;
    - Second law of thermodynamics;
    - Newton's law of universal gravitation.

Overview

- Introduction
- Hope: Why do we expect coarse-grained models to work?
- Approach: inference of coarse-grained (deterministic) dynamics

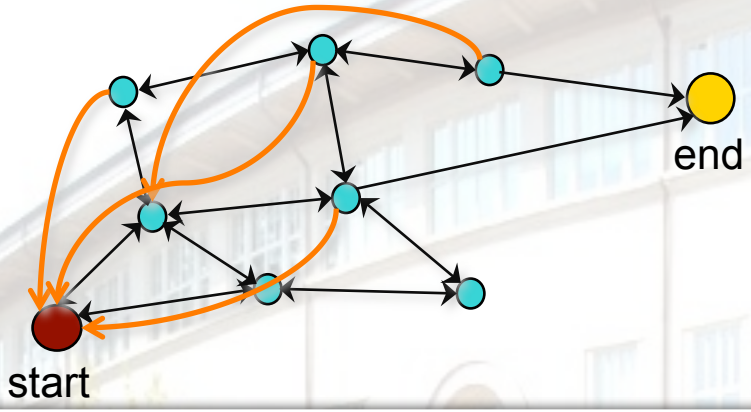
Part 1:
Why would this be possible?



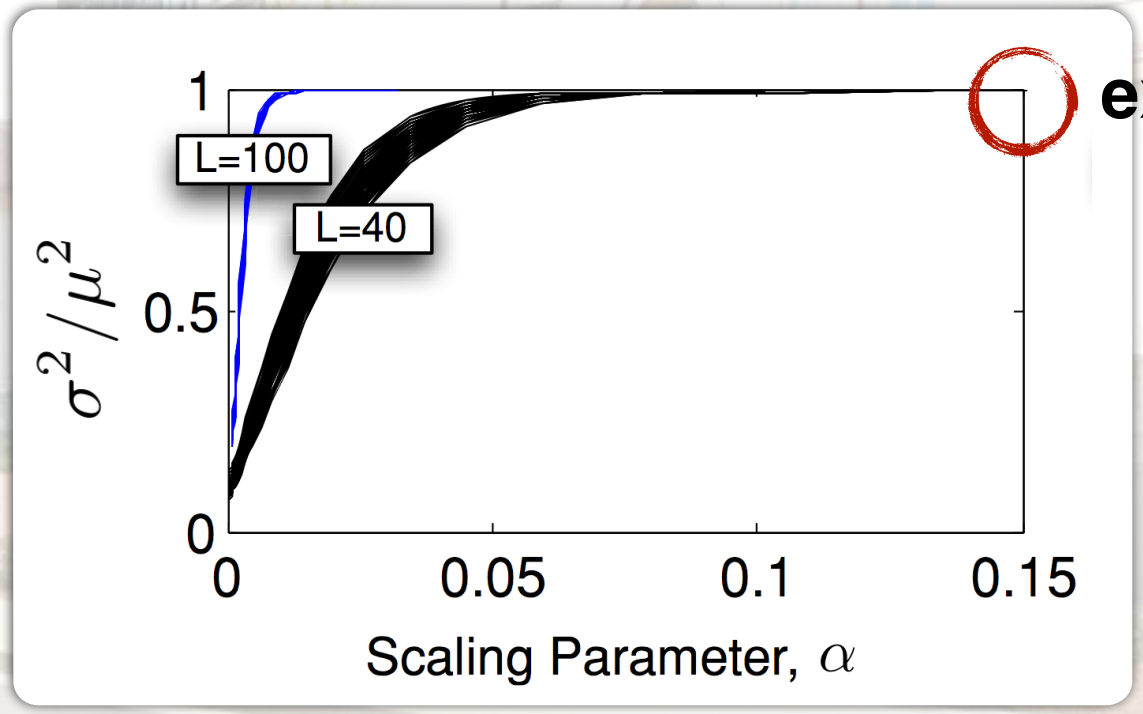


- **Macroscopic dynamics are often simpler than the network structure**: a handful of phenomenological parameters describe responses to most experimentally accessible perturbations.
- Relation of phenomenological to mechanistic parameters often unclear.

One explicitly calculated example: Macromolecules assembly through kinetic proofreading



α - relative strength of proofreading shortcuts



exponential pdf

- exponential completion time PDF for KPR
- very narrow completion time PDF for competing KPR systems

Bel, Munsky, Cheng, IN, *PhysBiol, JCP,* 2009-13

Another example (with more details): Emergence of apparent criticality for free

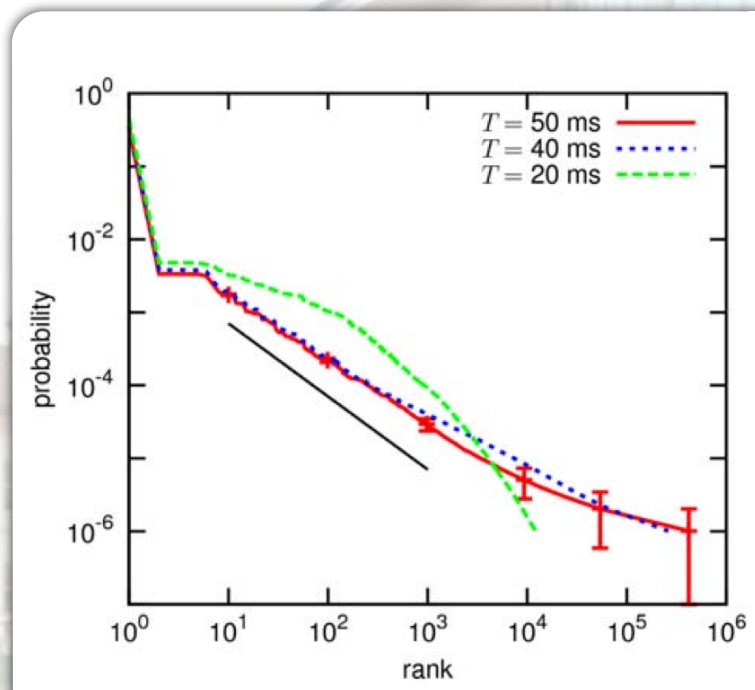
- Modern biology experiments measure multidimensional vectors of "states" of biological systems
  - Neurons firing/not firing at an instance t_i in a time window $i = 1 \ldots N$. Activity: $\sigma_i = \pm 1$. Firing at different times correlated.
  - Neuron i in a set of N neurons firing. Activity: $\sigma_i = \pm 1$. Firing of pre/post synaptic neurons correlated.
  - Genetic sequences of length N, with $\sigma_i = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ the letter at position i. Different nearby letters are correlated.
- Estimate the distribution of the activities $P(\vec{\sigma}) = P(\{\sigma_i\})$ from data and study its properties — **often see Zipf**.

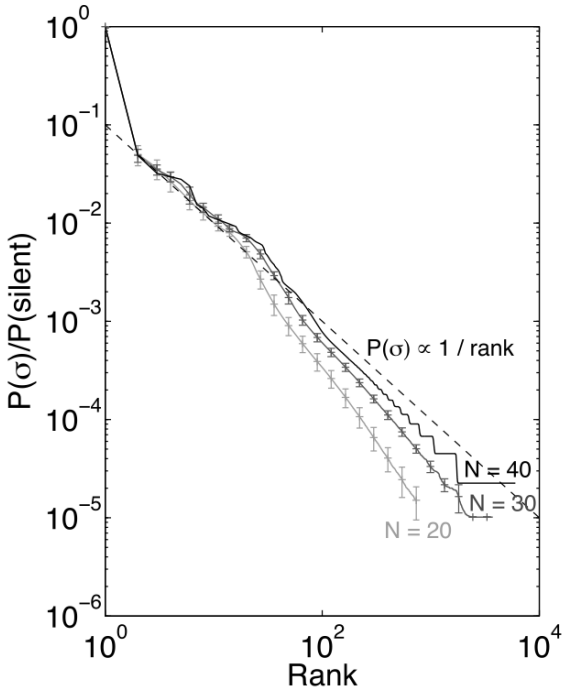
Zipf law (frequency~1/rank) is observed! Why such universality?

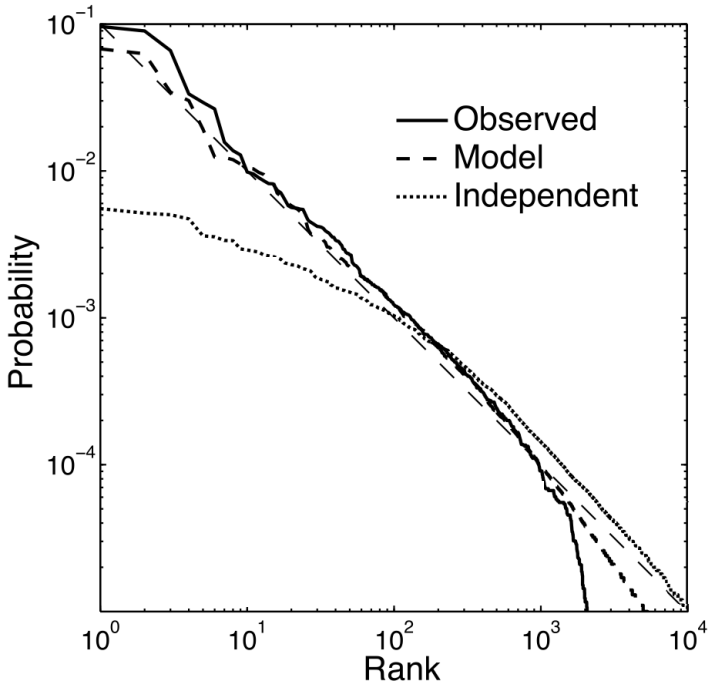
Activity of the fly H1 neuron in a time window

Activity of *N* neurons in salamander retina

Zebrafish antibody sequences







Nemenman et al., 2008

Tkacik et al., 2007
Mora and Bialek, 2011

Mora et al., 2010
Mora and Bialek, 2011

First, why is Zipf significant?
A signature of very special criticality!

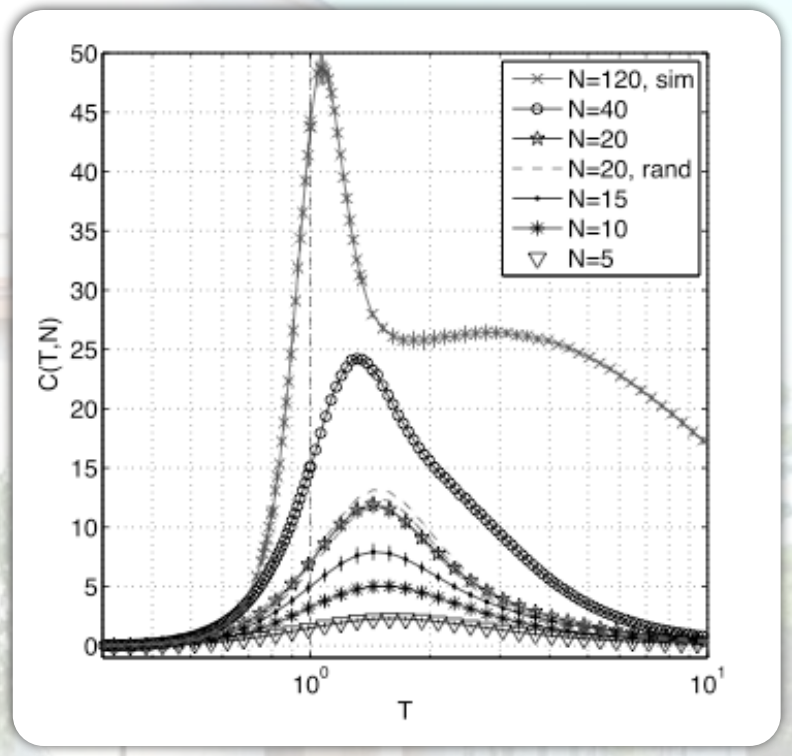
- Define energy, temperature, density of states, and micro canonical entropy.

$$P(\vec{\sigma}) = e^{\log P} = e^{-E}$$

$$P(\vec{\sigma}|T) = \frac{1}{Z} e^{-E/T}$$

$$P(\vec{\sigma}) \propto \frac{1}{r(\vec{\sigma})^{\alpha}} \Rightarrow S(E) = \frac{E}{\alpha} + o(N)$$

$$C(T) = \frac{N}{T^2}\left[-\frac{d^2 S}{dE^2}\right]^{-1}$$



Tkacik et al., 2007
Mora and Bialek, 2010

Why should these diverse biological systems be Zipf-critical?

- There are many arguments for why brain should be critical. But why should the brain be Zipfian specifically?
- Such arguments are harder to come by for cellular or genetic data.
- So could there be another explanation?
  - But: **we never record *everything* about the system…**

Coupling to unobserved variables

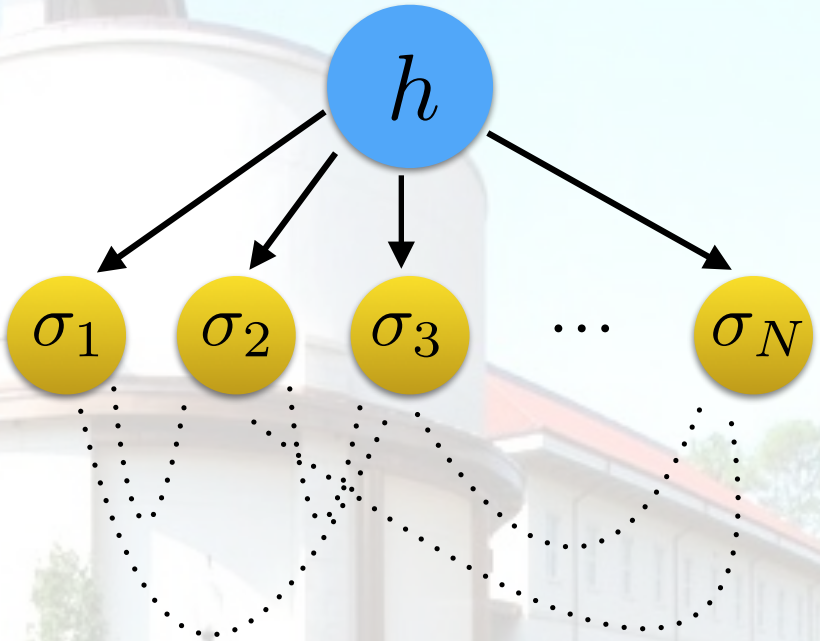
Schwab, IN, Mehta, 2014

$$P(\vec{\sigma}|h) = \prod_{i=1}^{N} P(\sigma_i|h) = \prod_{i=1}^{N} \frac{e^{h\sigma_i}}{2\cosh h};$$

$$P(\vec{\sigma}) = \frac{1}{2^N} \int dh\, p(h|h_0) e^{N(hm - \log\cosh h)}$$

$$\equiv e^{E(m,h_0)};$$

$$m = \sum \sigma_i/N, \quad \bar{h} = h_0$$



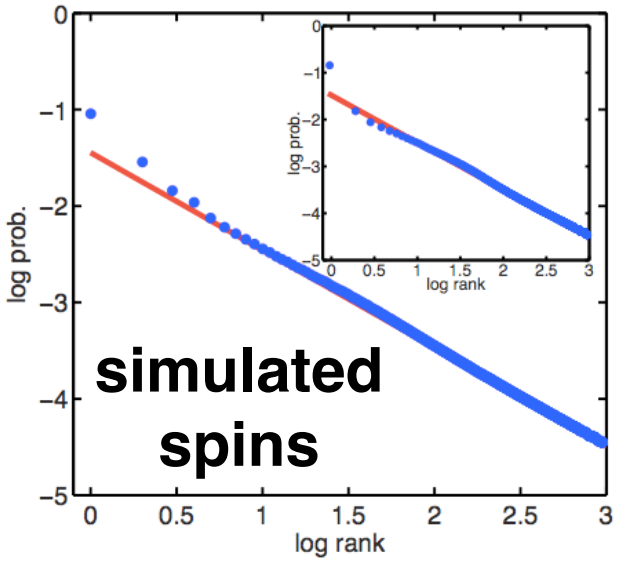
- Saddle point for large N, to the leading order.
- There's always $N_0(h_0, \mathrm{var}\, h)$, such that for $N > N_0$

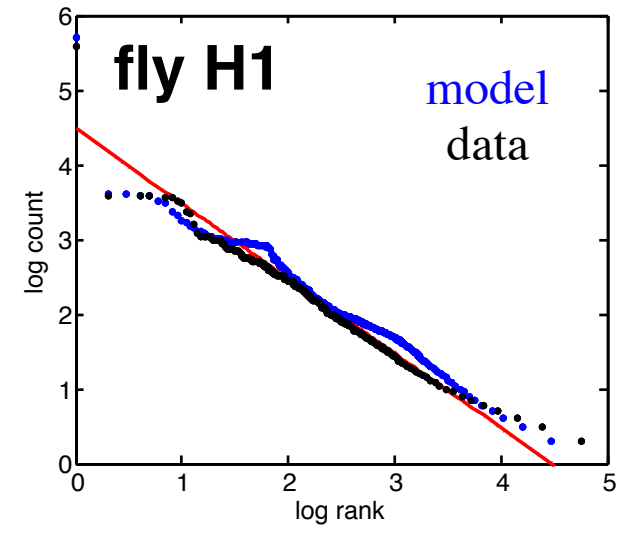
$$E(m) = S(m) + o(N)$$

Coupling to a hidden variable produces Zipf's scaling

This is general, but non-generic

- Large deviations theory generalizes the result (with caveats) to
  - Multiple external variables
    - infections, complex neural stimuli.
  - Nonuniform coupling.
  - Field-independent terms in energy
    - e.g., response to stimulus by a spike train with refractoriness.
- Only works if N is large enough to infer the field h from the spins
  - for moderate N requires **adaptation**, so that the spins are affected by the field.





Schwab, IN, Mehta, 2014

Summary #1:

- For many biological systems, deterministic or stochastic, dynamics is simpler than the network structure.
  - And one such simplification could be Zipfianity.
- Hope: it should be possible to infer low-dimensional dynamics directly from data, rather than building a detailed model first, and then coarse-graining it.

Part 2: Can we fit simple, phenomenological models to biological data?

- We will assume that dynamics of cellular networks is given by local **ordinary differential equations**.
  - Do not fit curves; **fit dynamics**.
- We will neglect stochasticity, and spatial structure for now

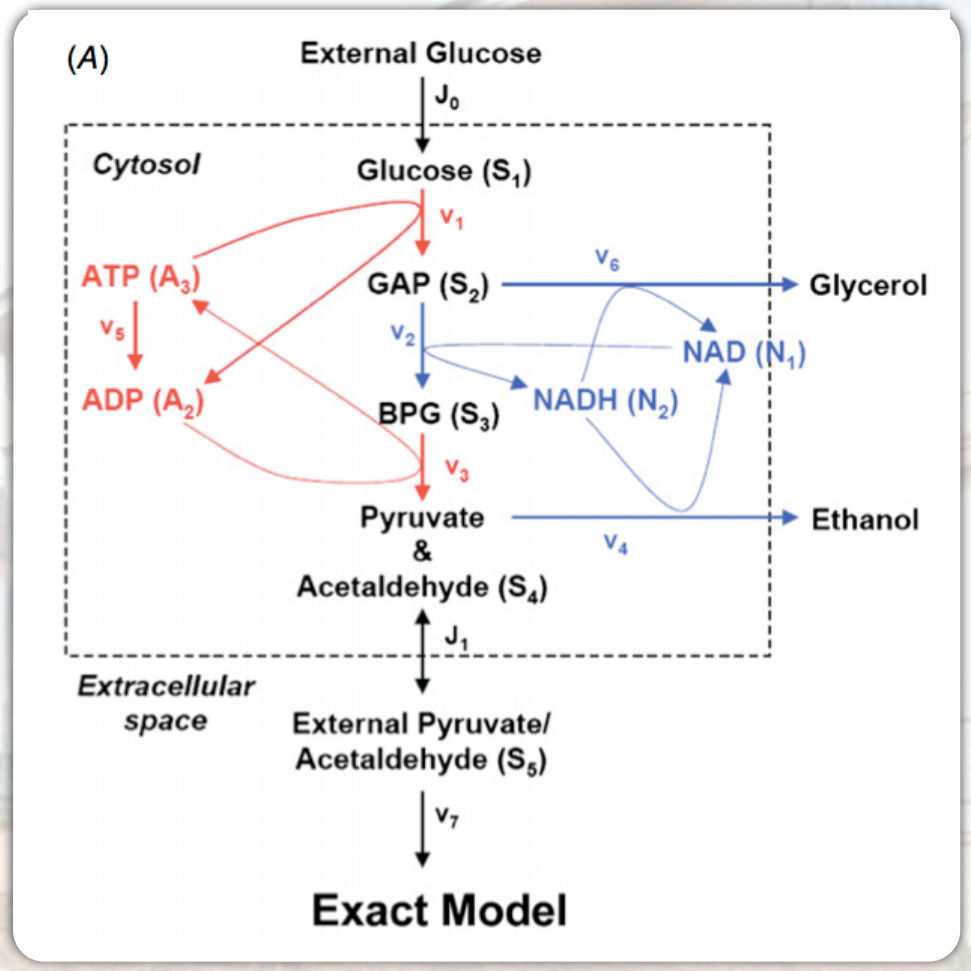
$$\begin{cases} \frac{dx_1}{dt} = f_1(x_1, x_2, \ldots, x_n) \\ \ldots \\ \frac{dx_n}{dt} = f_n(x_1, x_2, \ldots, x_n) \end{cases}$$

- Data: a few points per trajectory; *not derivatives*.
- Can we automatically fit these functions f_i from data?
  - How do we enumerate the set of all possible multivariate functions?
  - How do we search through this list? How do we not overfit?

Prior art in systems biology

- The full search approach for an exact model
  - Small systems dynamics — search for all possible models using S-systems formalism **(Voit et al, Theor Biol Med Model 2006).**
  - Searching for a control model from a (small) set of *a priori* allowed models **(Lillacci and Khammash, PLoS CB 2010).**
  - Searching for a stochastic model from a (small) set of *a priori* allowed models **(Munsky, et al., MSB 2009, Science 2013).**
  - **Eureqa:** exhaustive genetic algorithm search through all possible elementary function combinations, with selection of new experiments to optimize discriminability among models **(Lipson et al., Science 2009, Phys Biol 2011).**
- Phenomenological search **(Crutchfield and McNamara, Compl Syst 1987).**
- Problems (limiting the analysis to only a few variables)
  - **data/computing demands** explode with the number of variables;
  - cannot handle **unobserved** variables.

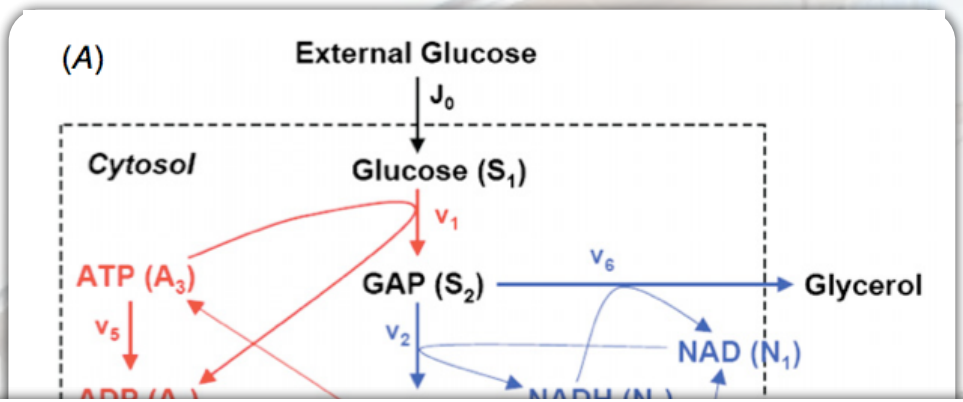
Testing Model: Yeast Glycolytic Oscillator



- 7 species, 28 parameters
- Complex rational dynamical laws

Ruoff et al., 2003

Testing Model: Yeast Glycolytic Oscillator

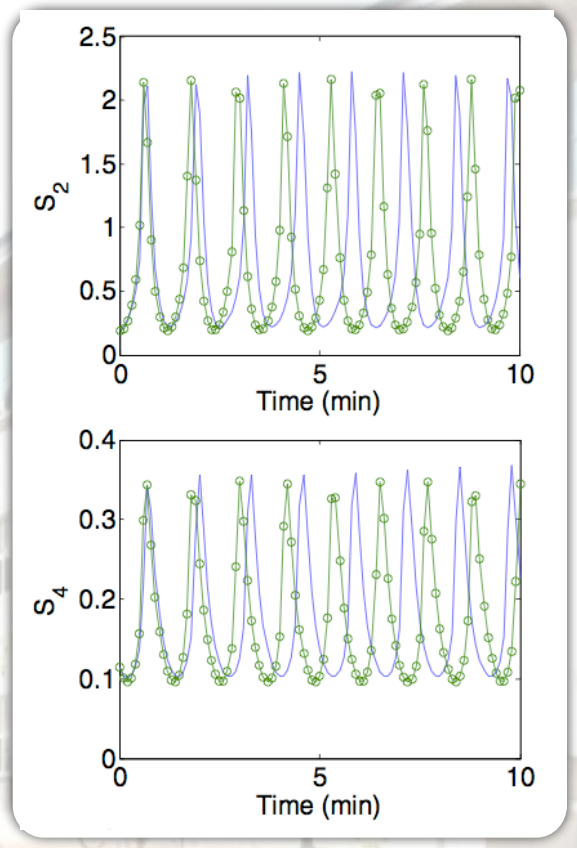


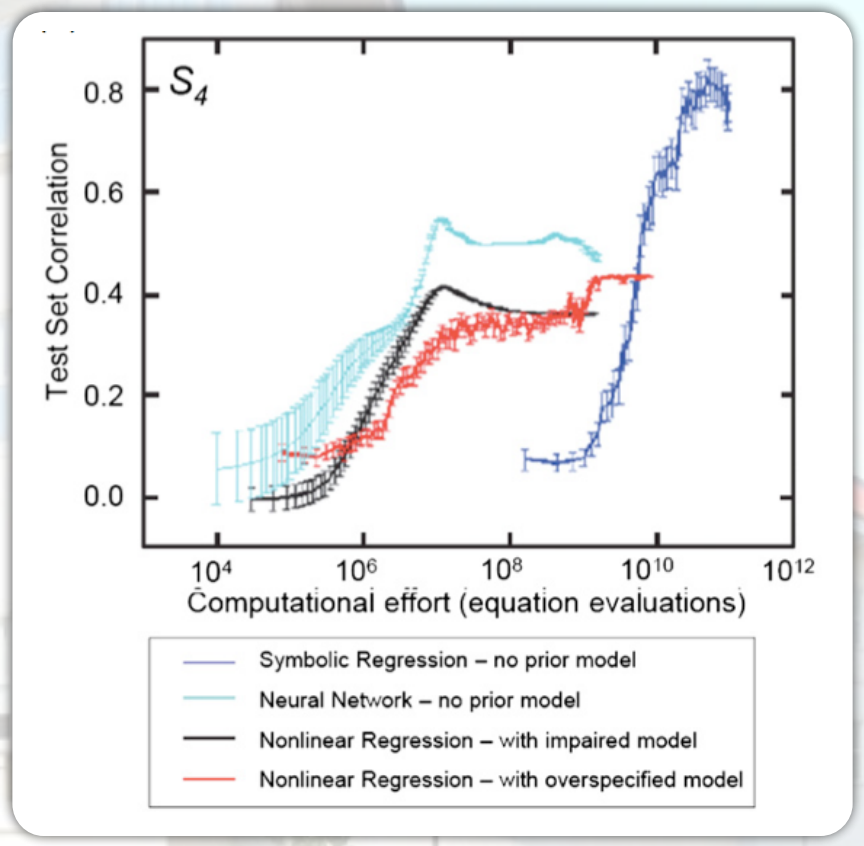
Amazing accuracy!

Original system	Automatically inferred system
$\frac{dS_1}{dt} = 2.5 - \frac{100*A_3S_1}{1+13.68*A_3^4}$	$\frac{dS_1}{dt} = 2.53 - \frac{98.79\cdot A_3S_1}{1+12.66\cdot A_3^4}$
$\frac{dS_2}{dt} = \frac{200*A_3S_1}{1+13.68*A_3^4} - 6*S_2 - 6*S_2N_2$	$\frac{dS_2}{dt} = \frac{200.23\cdot A_3S_1}{1+13.80\cdot A_3^4} - 6.87\cdot S_2 - 6.87\cdot N_2 + 0.95$
$\frac{dS_3}{dt} = 6*S_2 - 6*N_2S_2 - 64*S_3 + 16*A_3S_3$	$\frac{dS_3}{dt} = 6.00\cdot S_2 - 6.00\cdot N_2S_2 - 64.16\cdot S_3 + 16.08\cdot A_3S_3$
$\frac{dS_4}{dt} = 64*S_3 - 16*A_3S_3 - 13*S_4 - 100*N_2S_4 + 13*S_5$	$\frac{dS_4}{dt} = 64.04\cdot S_3 - 16.03\cdot A_3S_3 - 13.03\cdot S_4 - 100.11\cdot N_2S_4 + 13.21\cdot S_5$
$\frac{dN_2}{dt} = 6*S_2 - 18*N_2S_2 - 100*N_2S_4$	$\frac{dN_2}{dt} = -0.05 + 5.99\cdot S_2 - 17.94\cdot N_2S_2 - 98.82\cdot N_2S_4$
$\frac{dA_3}{dt} = -1.28*A_3 - \frac{200*A_3S_1}{1+13.68*A_3^4} + 128*S_3 + 32*A_3S_3$	$\frac{dA_3}{dt} = -1.12\cdot A_3 - \frac{192.24\cdot A_3S_1}{1+12.50\cdot A_3^4} + 124.92\cdot S_3 + 31.69\cdot A_3S_3$
$\frac{dS_5}{dt} = 1.3*S_4 - 3.1*S_5$	$\frac{dS_5}{dt} = 1.23\cdot S_4 - 2.91\cdot S_5$

Schmidt et al., Phys Biol 2011

But at the same time…



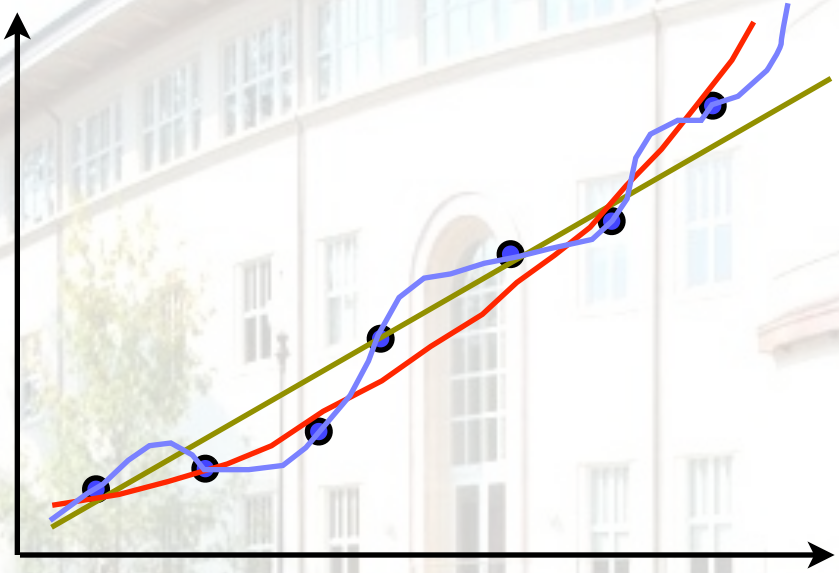


- Astronomical computation times **-- exhaustive search.**
  - **Overfitting** -- need astronomical sample sizes.
- Two exponential costs: **selecting** the best model family, **fitting** the best family with the model.

Schmidt et al., Phys Biol 2011

Can we avoid the exhaustive search?

- We don't need to do an exhaustive search when fitting 1-dimensional curves



$$y_K(x) = \sum_{k=1}^{K} A_k x^k + \text{noise}$$

  - Use Bayesian model selection to limit the complexity of the search space (the value of maximum *K*).

Schwartz, Ann Stat 1978; MacKay, Neural Comp, 1992
Balasubramanian, Neural Comp1996; Nemenman, Neural Comp, 2005

Bayesian Model selection

$$P(K|\{x_i\}) = \int d^K\vec{\alpha} P(\vec{\alpha}|\{x_i\}) = \int d^K\vec{\alpha} \frac{P(\{x_i\}|\vec{\alpha})\mathcal{P}(\alpha)}{P(\{x_i\})}$$
$$= \int d^k\vec{\alpha} \exp(-N\mathcal{L})$$

$$\log P(K|\{x_i\}) = \log P(\{x_i\}|\vec{\alpha}_{\mathrm{ML}}) - \tfrac{1}{2}\log\det N\mathcal{F} + O(N^0)$$

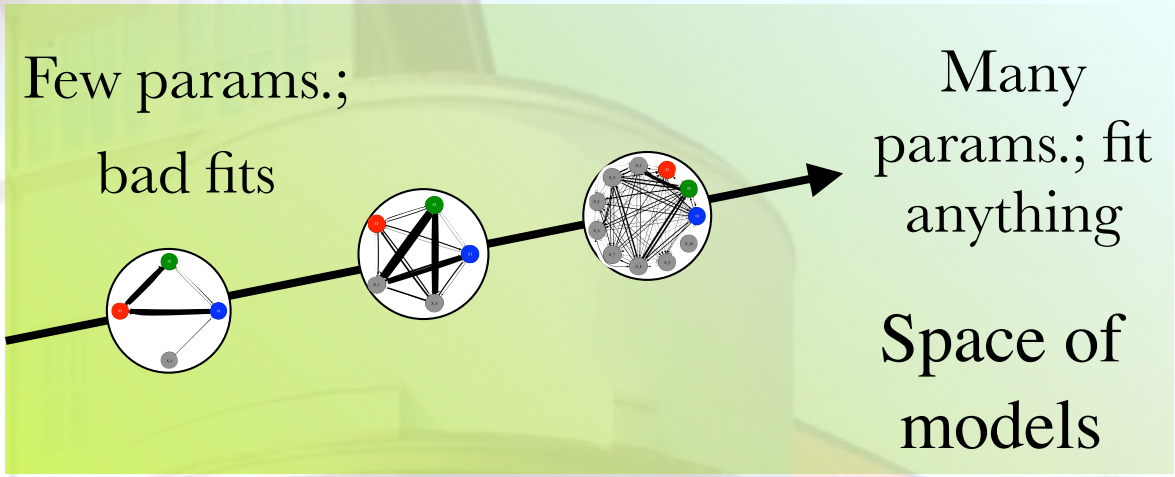
- For large sample size *N*, averages done in the Laplace (saddle point) limit.
- Penalty for model complexity (the log term) "selects" the best model family.
- Not that simple in detail, but this description is roughly accurate.
- Consistency properties for **nested, complete (infinite)** model families.

Schwartz, Ann Stat 1978; MacKay, Neural Comp, 1992
Balasubramanian, Neural Comp1996; Nemenman, Neural Comp, 2005

Why is fitting dynamics so hard?

$$\frac{d\vec{x}}{dt} = A_{\{xx\}}\vec{x} + A^{(2)}_{\{xx\}}\vec{x}\odot\vec{x} + \dots + A^{(K)}_{\{xx\}}\vec{x}\odot\cdots\odot\vec{x}$$

More nonlinearities



More hidden variables

$$\frac{d\vec{x}}{dt} = A_{\{xx\}}\vec{x}$$

$$\begin{cases} \frac{d\vec{x}}{dt} = A_{\{xx\}}\vec{x} + B_{\{x\}1}\xi_1 + \cdots + B_{\{x\}K}\xi_K \\ \frac{d\xi_1}{dt} = A_{1\{x\}}\vec{x} + B_{11}\xi_1 + \cdots + B_{1K}\xi_K \\ \cdots \\ \frac{d\xi_K}{dt} = A_{K\{x\}}\vec{x} + B_{1K}\xi_1 + \cdots + B_{KK}\xi_K \end{cases}$$

- Hidden degrees of freedom and nonlinearities breaks nestedness -- no consistency.
- Choose any (reasonable) **complete** path through the model space
  - Good choice — good fits with few data; Bad choice — not worse than exhaustive search.

Two types of model families

- Both nested and complete.
- Account for nonlinearities **and** hidden variables as more variables are added.
- Biochemically reasonable.

Why S-systems? Recasting!

$$\dot{x} = \sin x$$

equivalent to

$$\dot{x}_1 = x_2,\ \dot{x}_2 = x_3 x_2,\ \dot{x}_3 = -x_2^2$$

Sigmoidal recurrent networks

Daniels and Beer, arXiv 2010

Degradation

$$\frac{dx_i}{dt} = -x_i/\tau_i$$

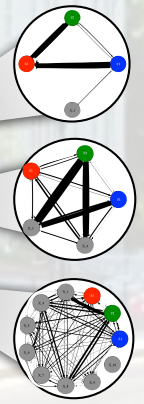
with $\xi(y) =$

S-systems

Savageau et al., 1976-…

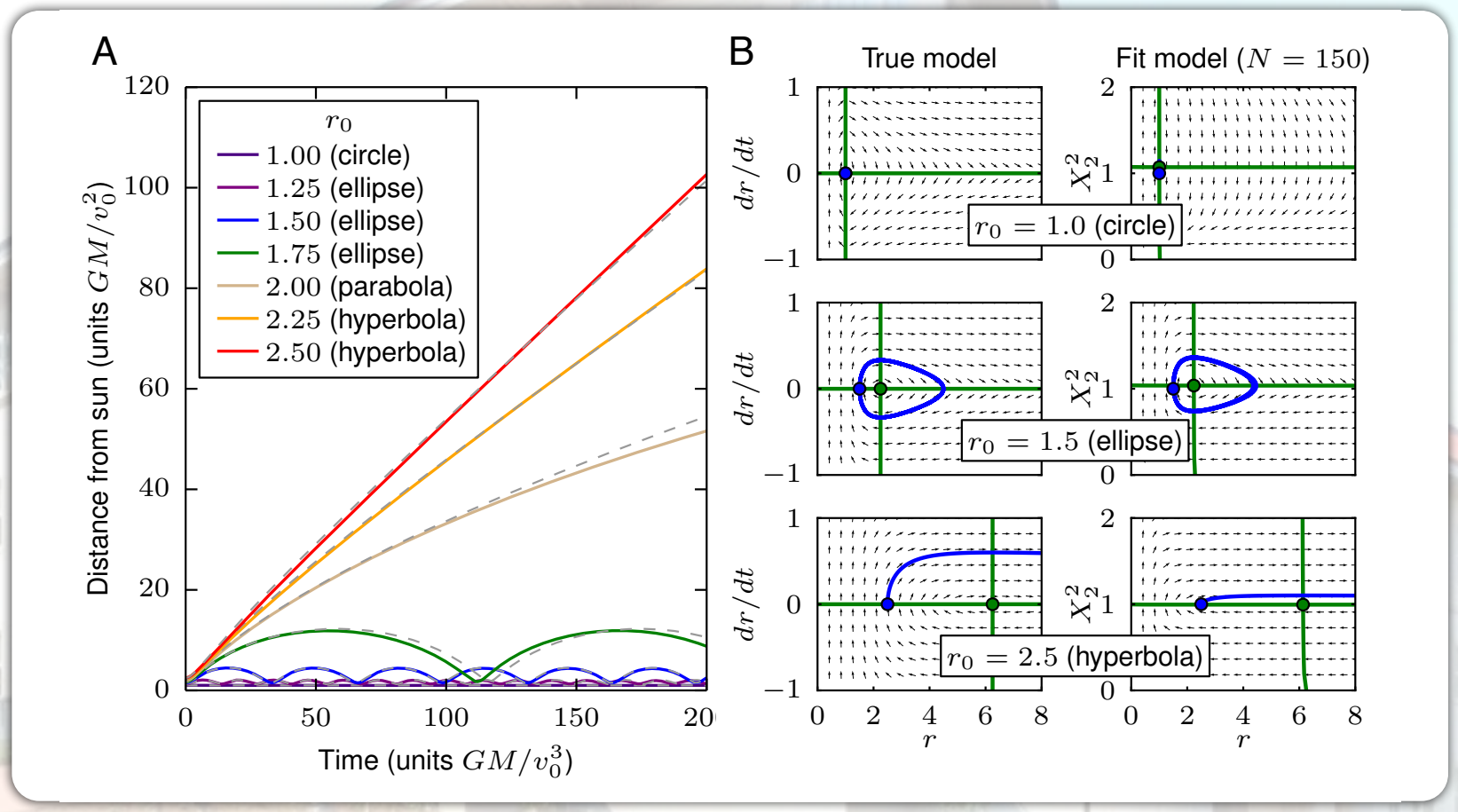
Interactions and input dependence

$$\frac{dx_i}{dt} = A_i \prod x_j^{\alpha_{ij}} \prod_k I_k^{a_{ik}} - B_i \prod x_j^{\beta_{ij}} \prod_k I_k^{b_{ik}}$$



Daniels and Nemenman, arXiv, 2014

Finding laws that we already know: An automated Sir Isaac *(SirIsaac on GitHub)*

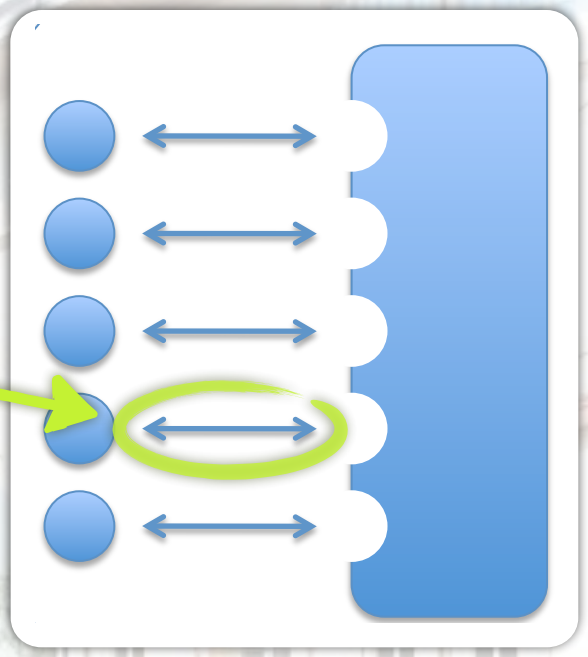


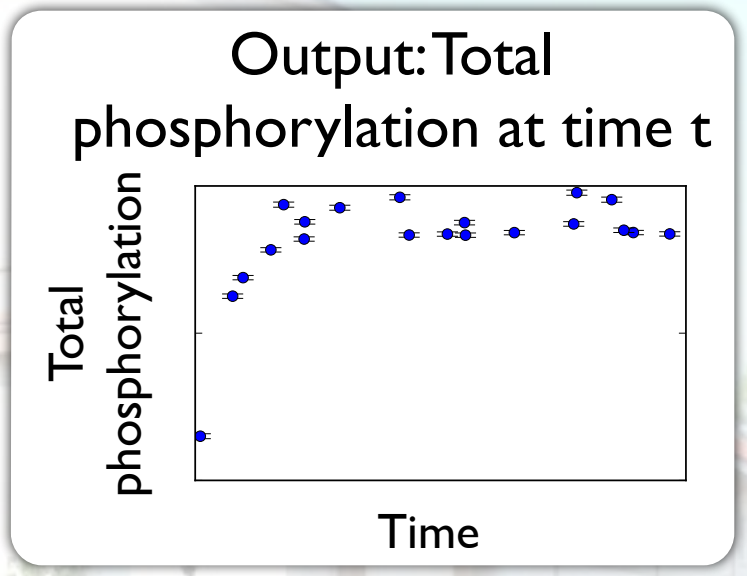
- Finds the hidden variable needed to account for the Newton's laws.
- Accounts for different classes of trajectories.

Daniels and Nemenman, arXiv, 2014

Simple dynamics from a complex network: Combinatorial multisite phosphorylation

Input:
Single on-rate

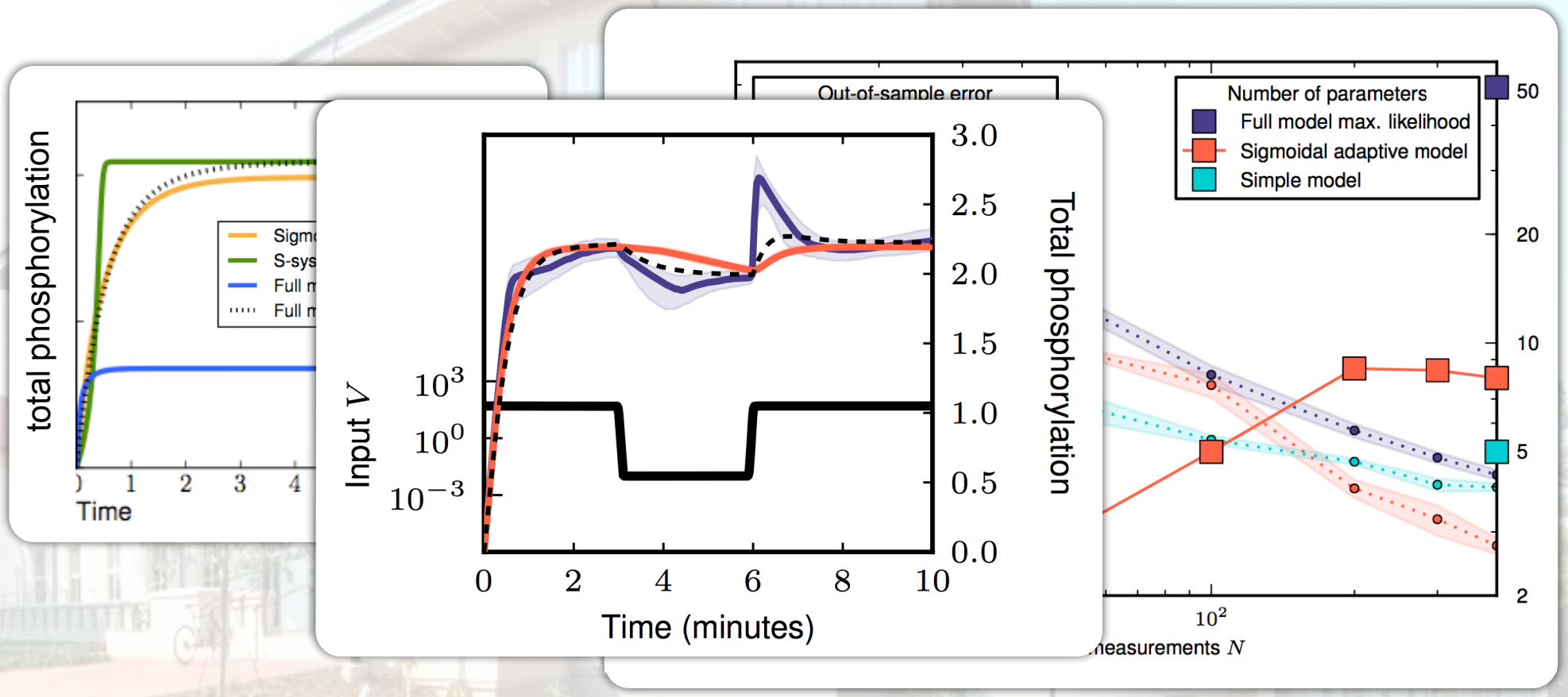




- Rates depend on occupancy of the nearby sites, 32 species, about 50 parameters total.
- Caricature of some of the most combinatorially complex signaling models.
- Typically more parameters than data.

Effective, reduced model of multi-site phosphorylation

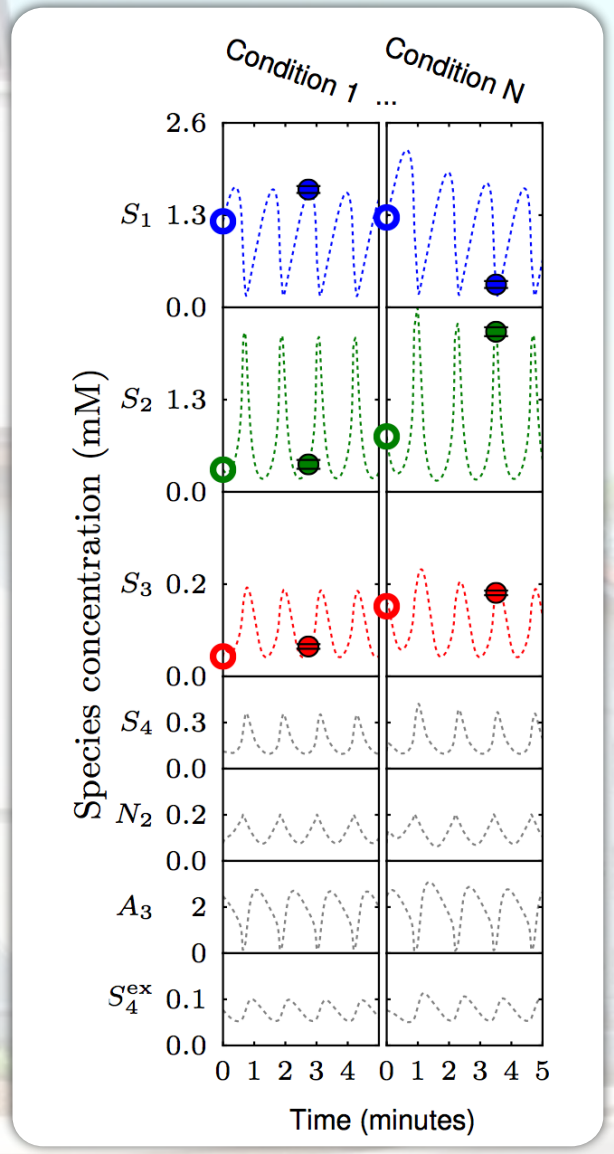
Daniels and Nemenman, arXiv and in review, 2014



- Effective models fit better than the true, full model for small data sets!
- Can *extra*polate to new signal classes, and not just *inter*polate.
- (Of course eventually the full, true model will win).

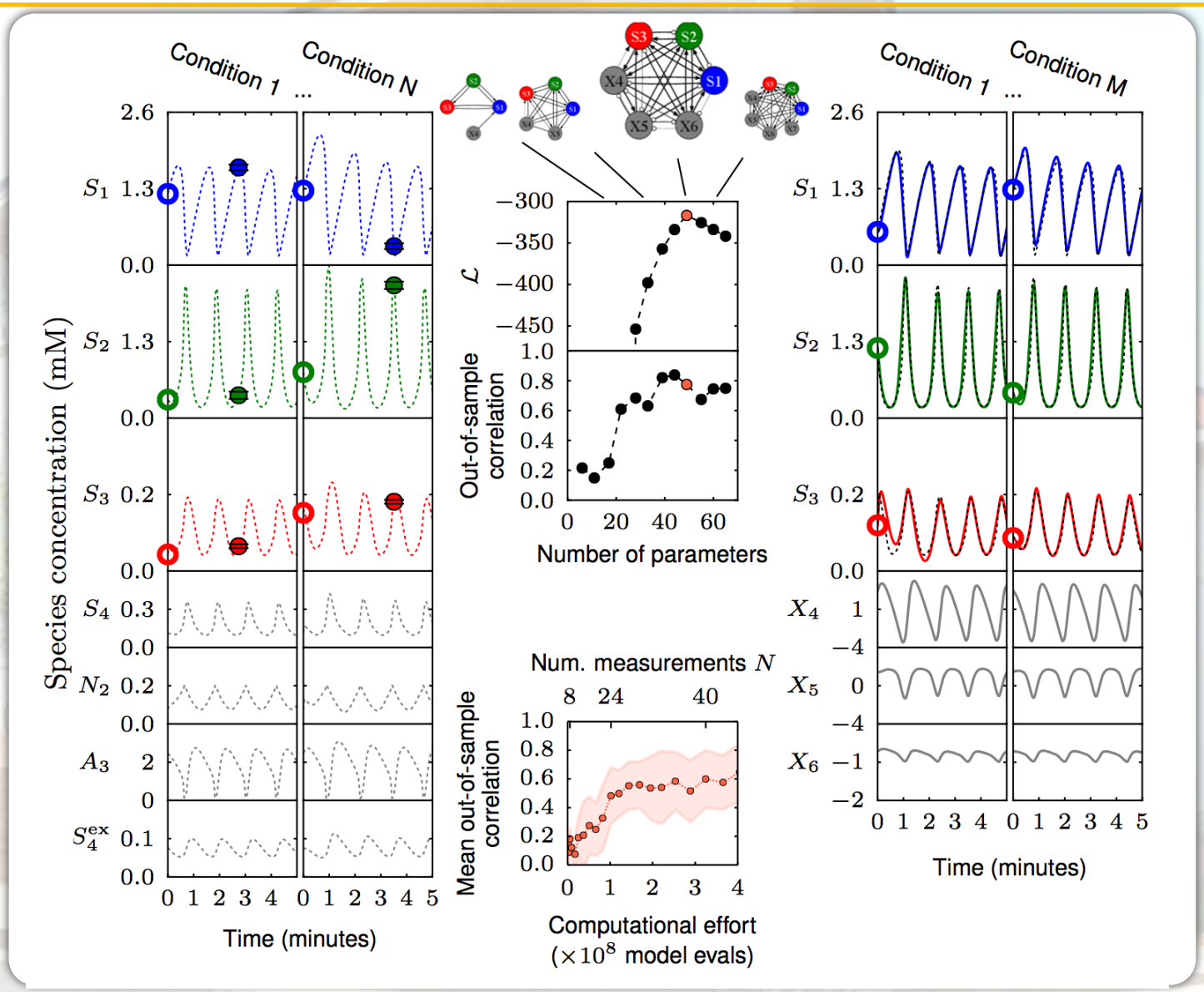
The yeast glycolytic oscillations: Complex dynamics needing complex structure

- Observe only 3/7 of variables; add 10% noise.
- Data: *N* samples of structure
  - Initial condition of the 3 species;
  - Some random time later;
  - The value of these 3 species at that time.

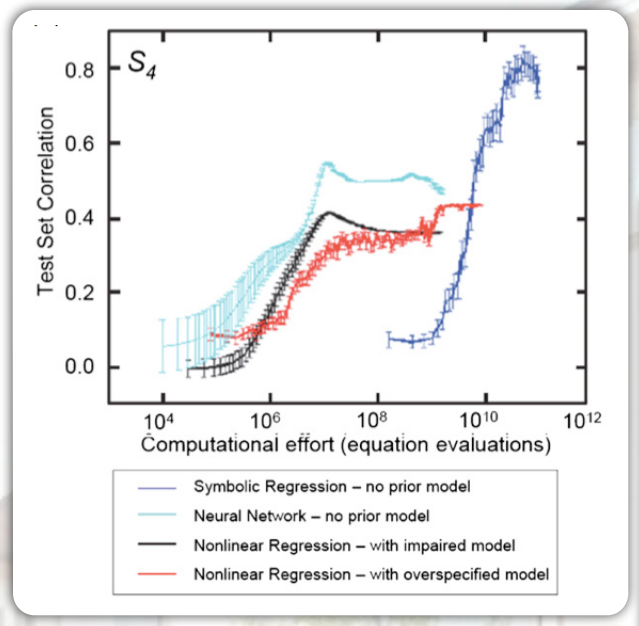


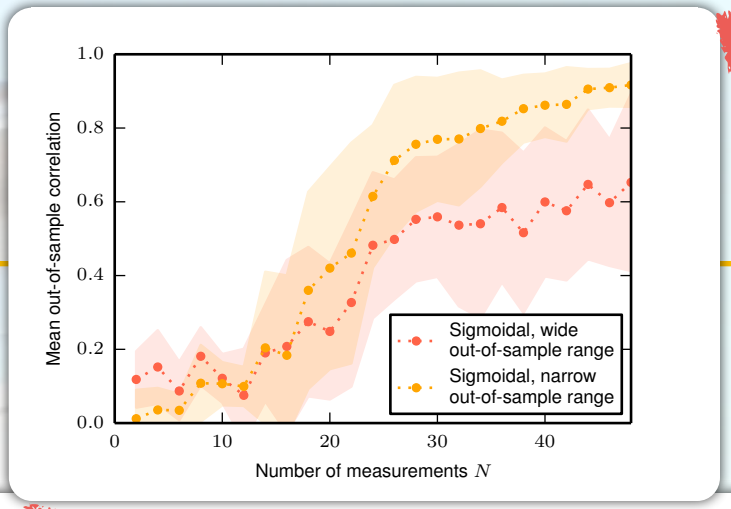
Results

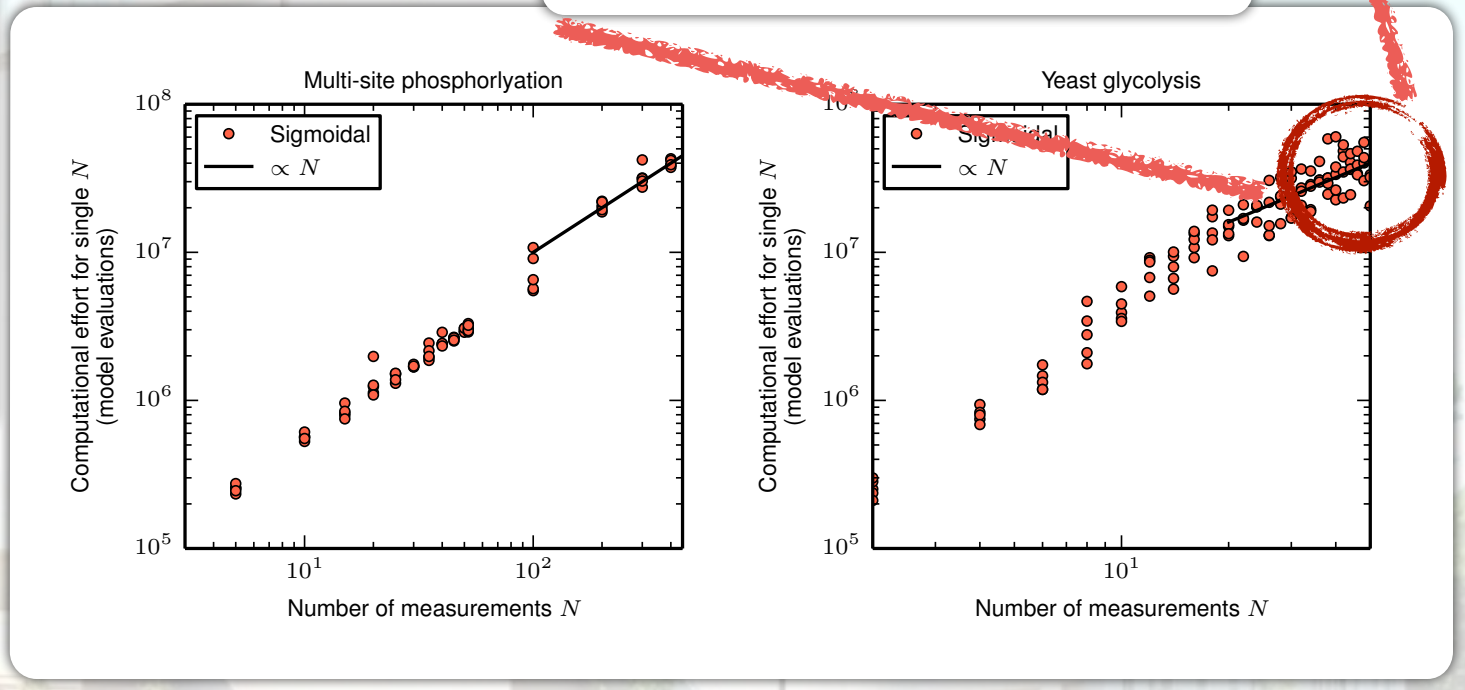
Daniels and Nemenman, arXiv and in review, 2014



Computational effort







- ~100x fewer evaluations for the same accuracy compared to full search.
- Only 50 data points (~1000x fewer than full search).
- Better accuracy than curve fitting.

Conclusions

- Search for **phenomenological** dynamics instead of exact.
- Why do this?
  - Sometimes biological systems do look simpler this way.
  - **The duck test:** If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck.
    - Indeed, **can predict response to yet-unseen perturbations!**
  - **Find new phenomenological laws of nature**
    - Repeat Hookean approach in biology: build effective models of similar systems and look for patterns (e.g., chemotaxis in *C. elegans* and *E. coli*).
- **Complete, nested** model families of dynamics allow to use Bayesian model selection to adapt effective model complexity to the available data.
- Such phenomenological models make accurate predictions in the undersampled regime, where true models overfit.

Announcements

- **The q-bio Conference**
  - Physical modeling in systems biology
  - Aug 10-14, 2014
  - Santa Fe, NM
  - Accepting late-breaking abstracts
  - Registration open
  - 2015 — Blacksburg, VA, 2016 — …
- **2 PD positions immediately available**