Musing about q-bio

Ilya Nemenman

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

nemenmanlab.org

EMORY UNIVERSITY

What to expect

- A few dozen really exciting talks.
- Each one sounds more important, more interesting than the previous one.
- Each one seems like this is the problem you absolutely must switch to and work on.
- So: are they worth switching?

Paraphrasing Hamming…

1. What are the most important problems in your field?
   - (let's write them on the blackboard)
2. What are the problems that you are working on?
   - (let's write them on the blackboard)
3. Why are the answers to (1) and (2) different?

We all want to do something great

- But you cannot do something great when you are working to answer a so-so problem.
  - You won't find gold if you are digging in a place where there isn't one.
- So where do great problems come from?

A word of caution

- **Never take seriously any advice that anyone gives you, including this one.**
- So this whole talk is just an opinion of one man, and probably won't apply to you. But I wish someone discussed some of this with me when I was a student.
- And maybe one should do simply what s/he finds exciting: because how else can you do something for 30+ years and still not get bored?
  - But this is also an advice.

How to maximize chances of finding gold in science?

- Science is collective. There are few “lone wolves”. The needle is moved no by “I” but by “all of us”.
  - 90% (or 99%) of science is stamp collecting, and it’s OK.
  - You wont find gold if you disregard what the others are doing.
  - Reality check: Gold is not common. You likely won’t be the one finding it even if you are doing all the right things.
- But it’s important to be idiosyncratic.
  - You won’t be the one finding gold if you are digging where everyone else is digging.
  - You won’t find gold if you are not the first, and you try to do better what others have already done.
- Dreaming, and dreaming big.
  - “When I examined myself and my methods of thought, I came to the conclusion that the gift of fantasy has meant more to me than my talent for absorbing positive knowledge.” (A. Einstein)

How to maximize chances of finding gold in science?

- Science is not math. It's experimental. It studies the world as it is, but not the world as it could have been.
  - "If simple perfect laws uniquely rule the universe, should not pure thought be capable of uncovering this perfect set of laws without having to lean on the crutches of tenuously assembled observations? True, the laws to be discovered may be perfect, but the human brain is not. Left on it's own, it's prone to stray, as many past examples sadly prove. In fact, we have missed few chances to err until new data freshly gleaned from nature set us right for the next steps. Thus pillars rather than crutches are the observations on which we base our theories…" (K. Schwarzchild)
  - Never solve an approximate problem exactly.
- "Science is only worthy of the name to the extent that mathematics finds a place in it" (I. Kant)
  - Is the "effectiveness of mathematics in natural sciences" so "unreasonable"?
  - Probably not, since the only things we know how to compare are numbers.

So where could gold be?

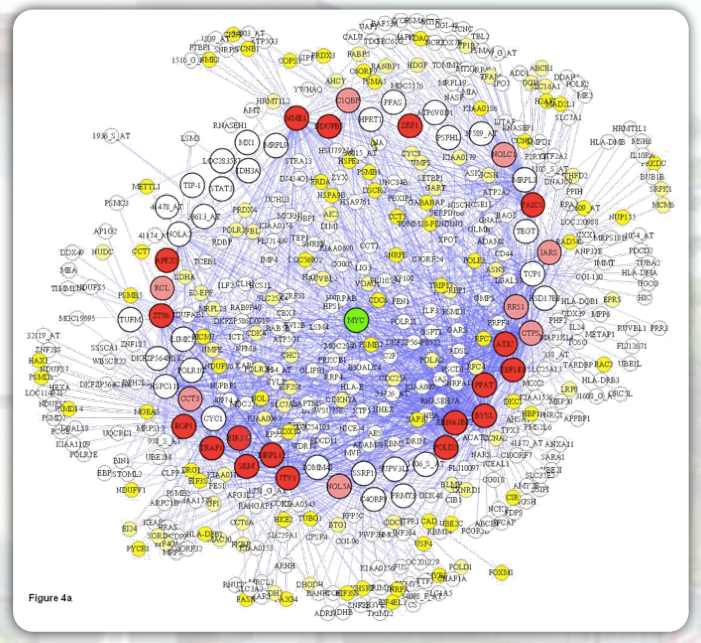
- Kuhnian revolution:
  - Normal science — development by accumulation, or returning to Rutherford and his stamp collection.
  - Episodic revolutions, anomalies — isn't this where the gold is?
- Weisskopf, extensive vs. intensive research:
  - Extensive: goes for explanation of phenomena in terms of known fundamental laws.
  - Intensive: goes for fundamental laws. Presumably, this is where revolutions are.
- Modern biology is ridiculously extensive, connecting everything on organismal and even ecological scales back to molecules.
  - No place for intensive research — there cannot be new fundamental laws away from elementary particles and cosmology.
  - No place for Kuhnian revolution.
  - No place for fantasy and big dreams.
- Modern biology is extensive to the extent that it is becoming not a Western science anymore.

What do I mean by this?

- Western tradition:
  - There are laws (of nature, of god, whatever). A rock is a rock everywhere. It falls the same in Pisa and in Atlanta.
  - There're causes and there are effects.
  - There is "useless information" (Oscar Wilde).
  - But this belief requires closing one's eyes to minor discrepancies
    - Two balls dropped from the Leaning Tower didn't land at the same time.
    - "If we had the STM in the 1920s, there wouldn't be the Debye theory of solids." (H. Levine)
- Non-western tradition, e.g., buddhism
  - Pratityasamutpada: dependent origination: "*Pratitya samutpada* is sometimes called the teaching of cause and effect, but that can be misleading, because we usually think of cause and effect as separate entities, with cause always preceding effect, and one cause leading to one effect. According to the teaching of Interdependent Co-Arising, cause and effect co-arise (*samutpada*) and everything is a result of multiple causes and conditions... " — Thich Nhat Hanh

The end of Western science: The -omics revolution in biology

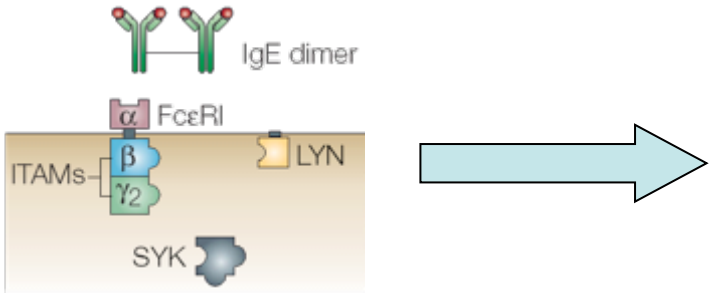
- Breaking life into ever more accurate "fundamental" parts lists
  - Sequences: genomics, metagenomincs, epigenomics,…
  - Activities: gene expression, metabolic profiling, phosphoproteomics, electrophysiology …
  - **From zoology to zoology of molecules.**
- Putting it all back into a network of interactions
  - Metabolic, transcriptional, protein signaling, neural, ecological…
  - Which things go together?
  - Number of possible interactions is astronomically large.
  - **Towards zoology of circuits/networks.**

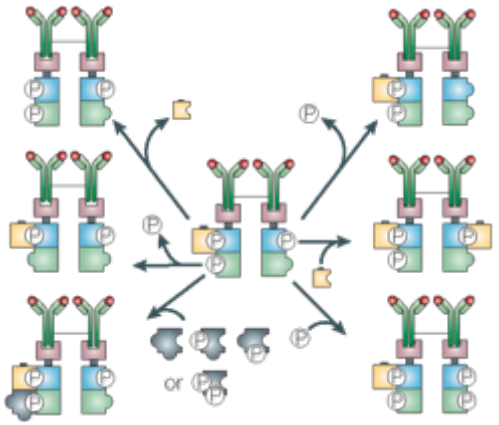


Califano et al., Nat Gen 2005; BMC Bioinf 2006

Is this program feasible?

- To predict **dynamics**, we will need to measure details of many interactions with excruciating details.
- The number of interaction details is combinatorially large!





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

354 species / 3680 reactions
(2954 for trimers)

- Where do we stop? Chemical kinetics? MD simulations? QFT?…
- Can we do better if we only need the **macroscopic dynamics**, but not the microscopic accuracy per se?

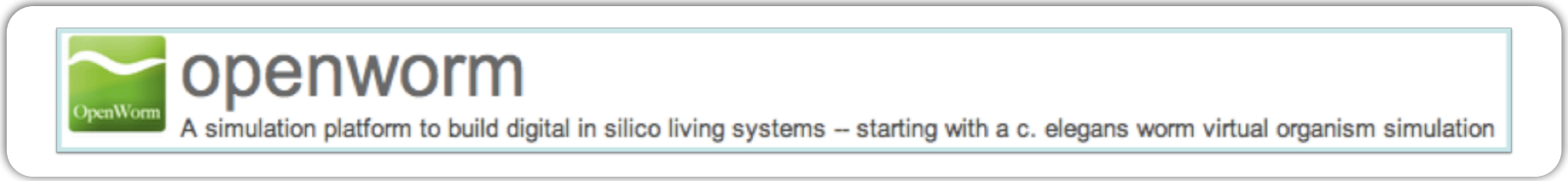
Of exactitude in science

...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.

At a recent meeting...

- Many have expressed an opinion that: "**The final** theory of biological systems will be a large multiscale computational model. We need more and more experimental data to specify details of these models."



- There's something wrong with this statement.
  - The "**final**" theory?
  - Do we need the theory of "**everything**" in any biological (or physical) system?

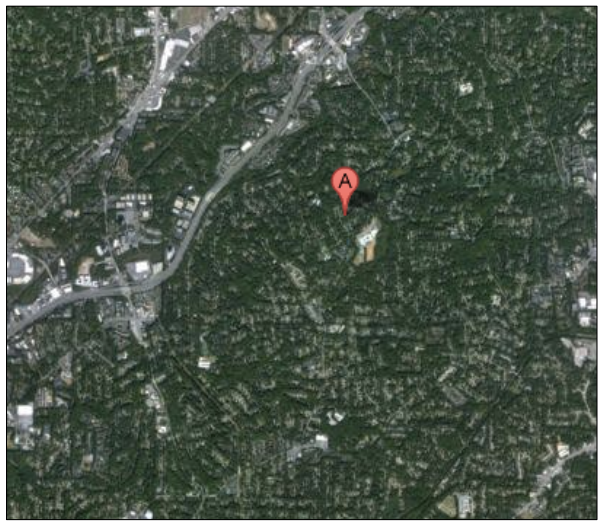
- The best material model of a cat is another, or preferably the same, cat.
  (*Philosophy of Science*, Wiener and Rosenblueth, 1945)

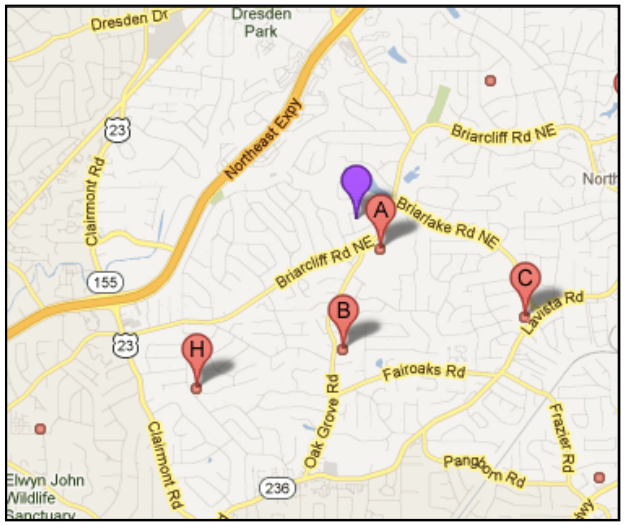
Physics analogy

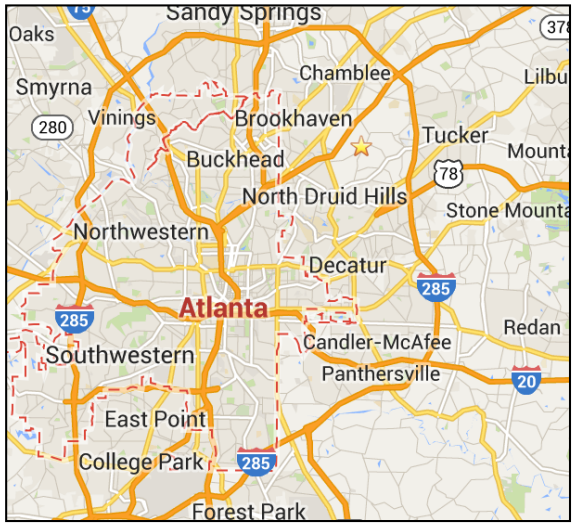
- What is the final, complete theory of the chair you are sitting in?
  - How does it fall from the second floor?
  - How does the cloth seat age and tear?
  - How much weight would the chair hold before it breaks?
  - How does it conduct electricity?
  - How much food can I cook when I burn it?
  - …
- There's no such thing as "the full theory of the chair".
  - We build models tailored to answer **specific questions**.
  - The complete theory that answers **every** question would need to include quarks, superstrings…
- Models must loose details. Otherwise, just use the same cat…

Where would new laws come from?

- **More is different!** (PW Anderson)
  - The **law of large numbers** produces universalities if the right questions are asked (e.g., about *large-scale* quantities).
- Intensive science is in putting things together, not just breaking apart.
  - Coarse-graining: Each modeling level needs its own **effective** degrees of freedom.
  - "Don't model bulldozers with quarks." (Goldenfeld and Kadanoff, 1999)
- This is already common in your **every-day life,** not just physics
  - Which level of description is **better for** driving to a local school?







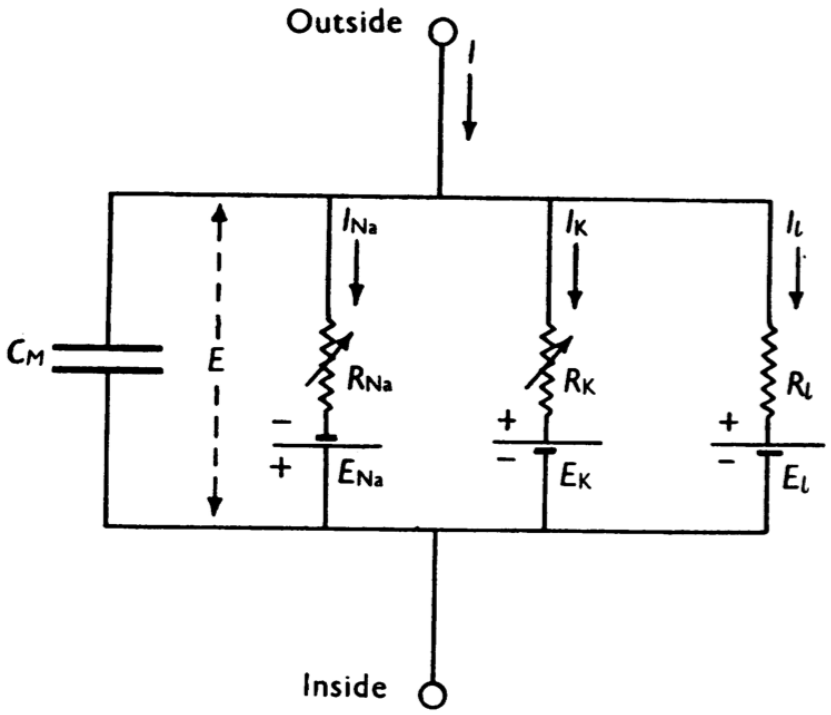
So where is gold in systems biology? (according to IN)

- It's in the **more,** but only if one tries to make it **different**, to compress it
  - Rissanen (paraphrasing): good theories compress data.
  - Good theories (unlike even a good model) compress representation of multiple systems at the same time.
  - The information bottleneck/rate distortion curve (sketch on the board).
- Are there examples of this approach in well-established biology?

Has theory been useful in biology? A poll:

- What are the most important biology (or rather physiology/ medicine) Nobel prizes?
  - One per branch of biology (Mol/cell, Neuro, Evo)
- My prediction is that most of you answered
  - Watson and Crick
  - Luria and Delbruck
  - Hodgkin and Huxley

Hodgkin-Huxley



$$I = C_M \frac{dV}{dt} + I_i, \quad (1)$$

$$I_{Na} = g_{Na}(V - V_{Na}), \quad (3)$$
$$I_K = g_K(V - V_K), \quad (4)$$
$$I_l = \bar{g}_l(V - V_l), \quad (5)$$

$$g_K = \bar{g}_K n^4, \quad (6)$$
$$\frac{dn}{dt} = \alpha_n(1-n) - \beta_n n, \quad (7)$$

$$g_{Na} = m^3 h \bar{g}_{Na}, \quad (14)$$
$$\frac{dm}{dt} = \alpha_m(1-m) - \beta_m m, \quad (15)$$
$$\frac{dh}{dt} = \alpha_h(1-h) - \beta_h h, \quad (16)$$

- A good theory! (and only roughly correct)

Luria and Delbruck

(6a) $$r = (t - t_0)aN_t.$$

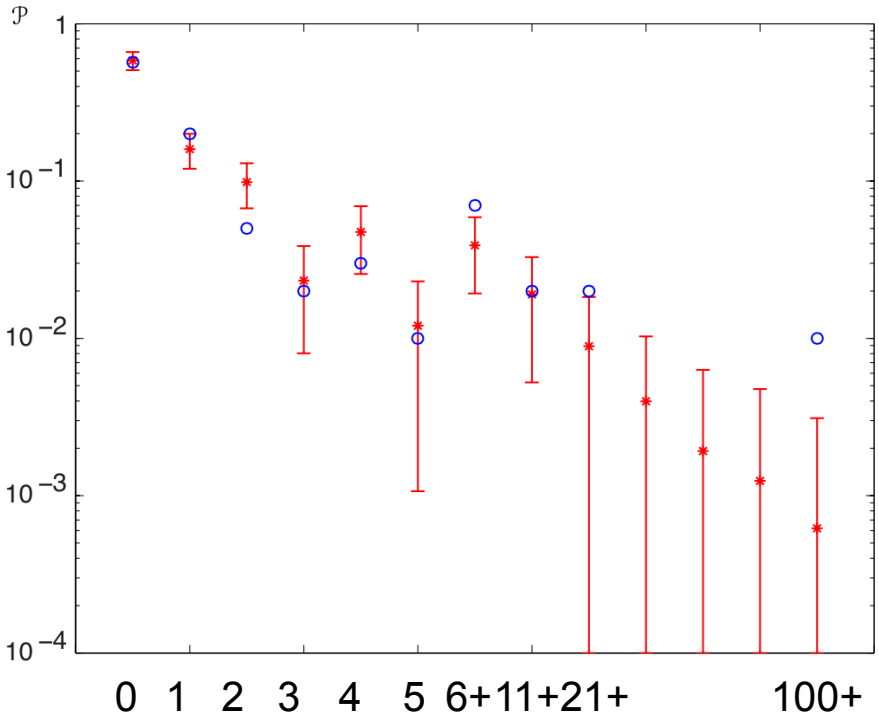
It now remains to choose an appropriate value for the time interval $t - t_0$.

For this purpose we return to equation (4), in which it was stated that the average number of mutations which occur in a culture is equal to the mutation rate multiplied by the increase of the number of bacteria. Let us then choose t_0 such that up to that time just one mutation occurred, on the average, in a group of C similar cultures, or

$$1 = aC(N_{t_0} - N_0).$$

In this equation we may neglect N_0, the number of bacteria in each inoculum, in comparison with N_{t_0}, the number of bacteria in each culture at the critical time t_0. We may also express N_{t_0} in terms of N_t, the number of bacteria at the time of observation, applying equation (1):

$$N_{t_0} = N_t e^{-(t-t_0)}.$$



P. Nelson, 2014

- And this theory throws a lot of details out - e.g., it uses discrete generations.

Watson and Crick

The previously published X-ray data[3,4] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

Watson and Crick, 1953

Crick, 1958

V. CONCLUSIONS

I hope I have been able to persuade you that protein synthesis is a central problem for the whole of biology, and that it is in all probability closely related to gene action. What are one's overall impressions of the present state of the subject? Two things strike me particularly. First, the existence of general ideas covering wide aspects of the problem. It is remarkable that one can formulate principles such as the Sequence Hypothesis and the Central Dogma, which explain many striking facts and yet for which proof is completely lacking. This gap between theory and experiment is a great stimulus to the imagination. Second, the extremely active state of the

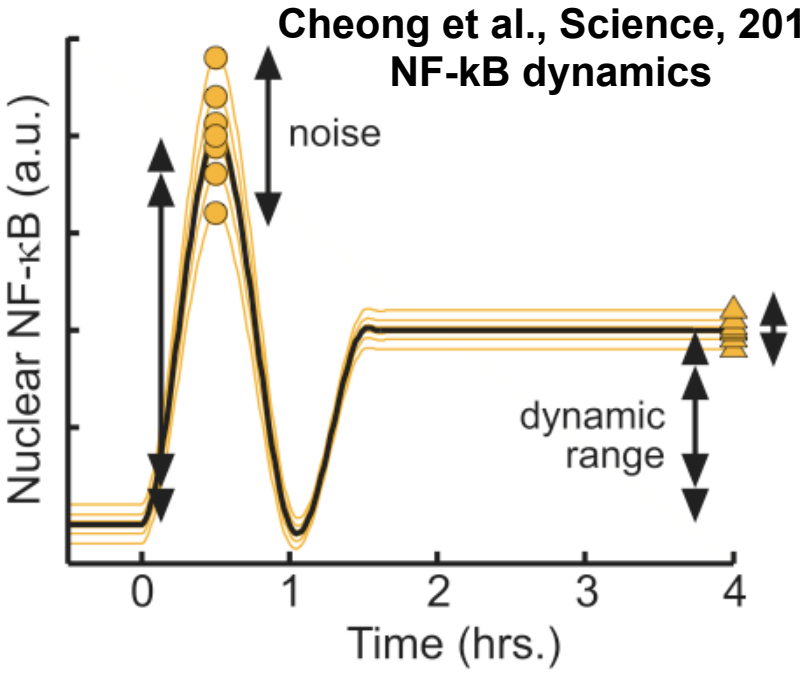
Theory in biology

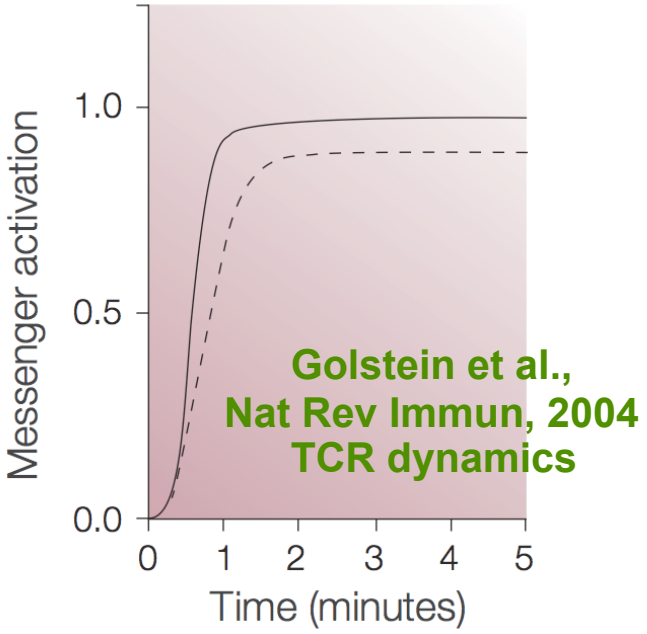
- “He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may cast.”
- And it’s been like this since Leonardo, in physics and in biology.

Can we do more?

- Are there other examples of where functionally accurate representations of biological processes emerge in a large *N* limit, or are we forever doomed to every detail mattering?
  - Can you give me some examples?
  - Can we derive theories top down rather than bottom up? (Refine rather than coarse-grain)

Let's look at dynamics in systems biology





- **Macroscopic dynamics are often simpler than the network structure!**
- Relation of phenomenological to mechanistic parameters often unclear.

First steps…

- 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

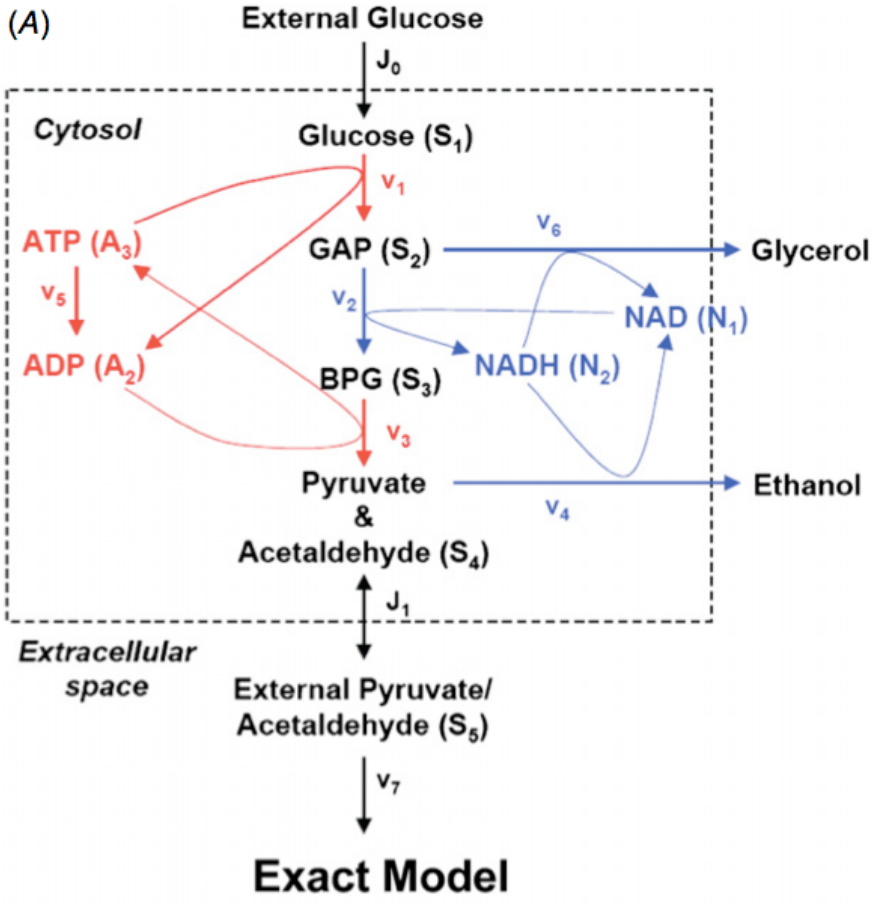
$$\begin{cases} \frac{dx_1}{dt} = f_1(x_1, x_2, \ldots, x_n) \\ \ldots \\ \frac{dx_n}{dt} = f_n(x_1, x_2, \ldots, x_n) \end{cases}$$

- 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

- 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.

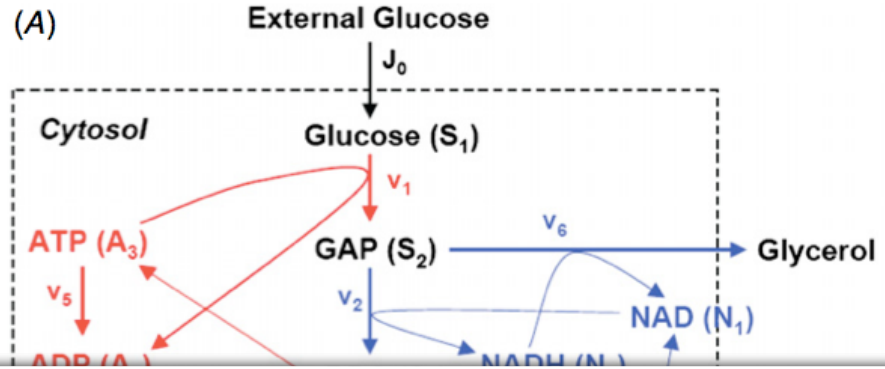
Testing Model: Yeast Glycolytic Oscillator



- 7 species, 28 parameters
- Complex rational dynamical laws

Ruoff et al., 2003

Testing Model: Yeast Glycolytic Oscillator

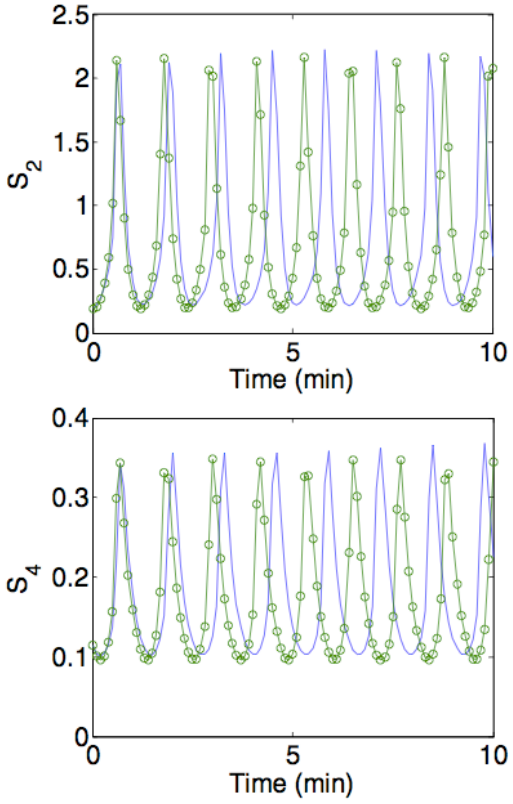


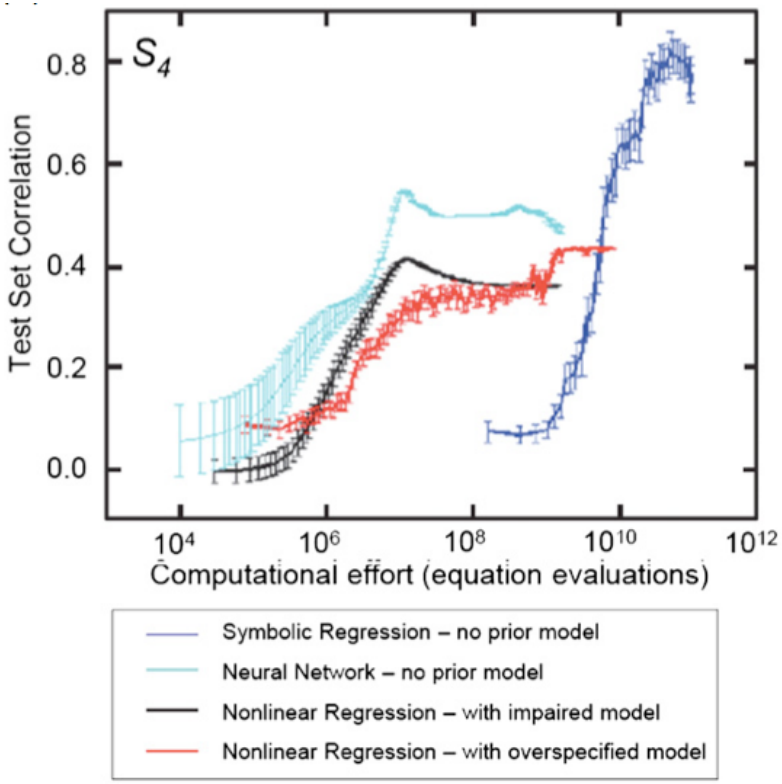
Amazing accuracy!

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…



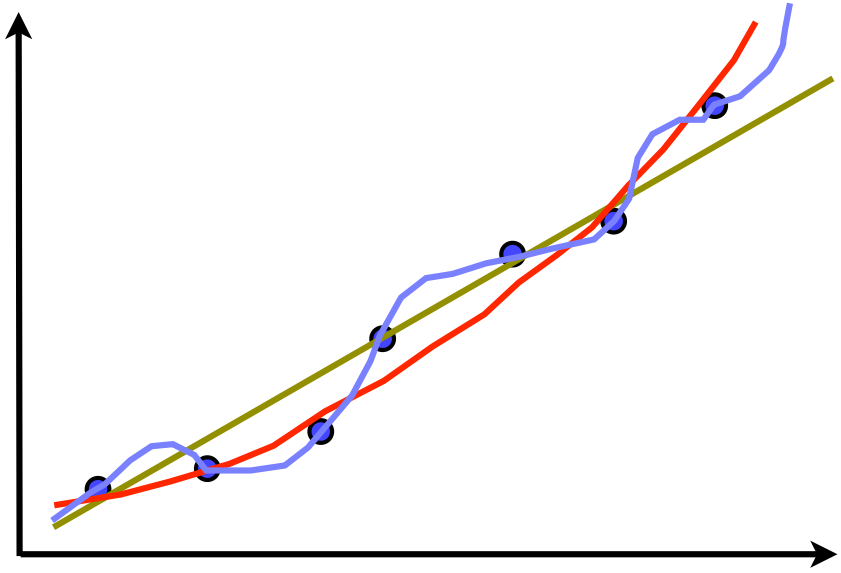


- Astronomical computation times **-- exhaustive search.**
  - **Overfitting** -- need astronomical sample sizes.
- Two exponential costs: **selecting** the best model family, **fitting** the best family with the model.

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



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

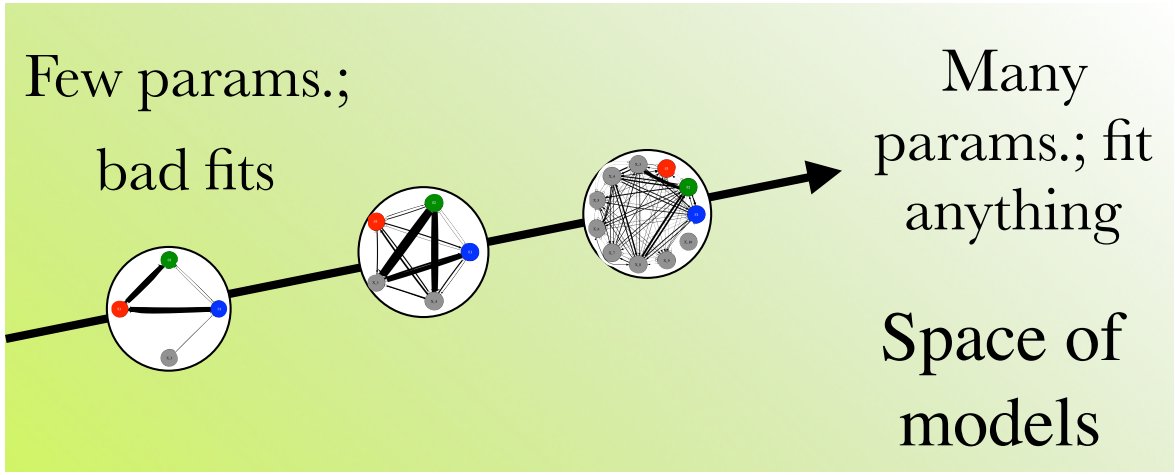
  - Use **nested**, **complete** model families, e.g., Taylor series.
  - Use Bayesian model selection to limit the complexity of the search space (the value of maximum *K*).
  - **Consistency** guaranteed iff nested!

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



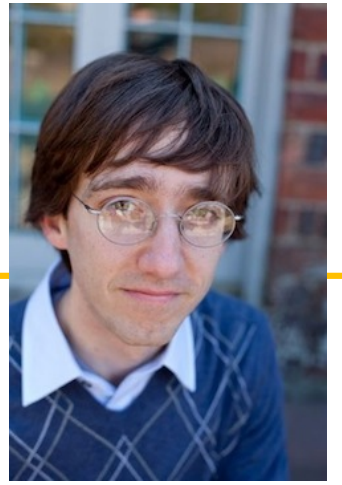
More hidden variables

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

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

- 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



Bryan Daniels

- Both nested and complete.
- Account for nonlinearities **and** hidden variables.
- Biochemically reasonable.

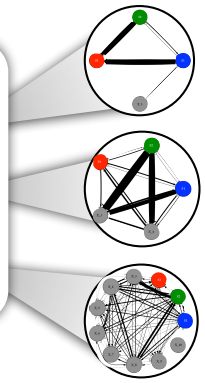
Sigmoidal recurrent networks

Daniels and Beer, arXiv 2010

Degradation Interactions Input

$$\frac{dx_i}{dt} = -x_i/\tau_i + \sum_{j=1}^{J} W_{ij}\ \xi(x_j + \theta_j) + \sum_{k=1}^{K} V_{ik} I_k$$

$$\text{with } \xi(y) = 1/(1+e^{-y})$$

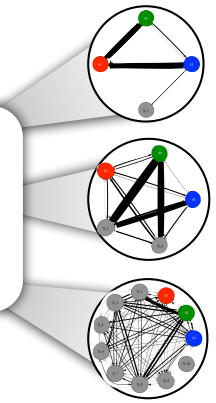


S-systems

Savageau et al., 1976-…

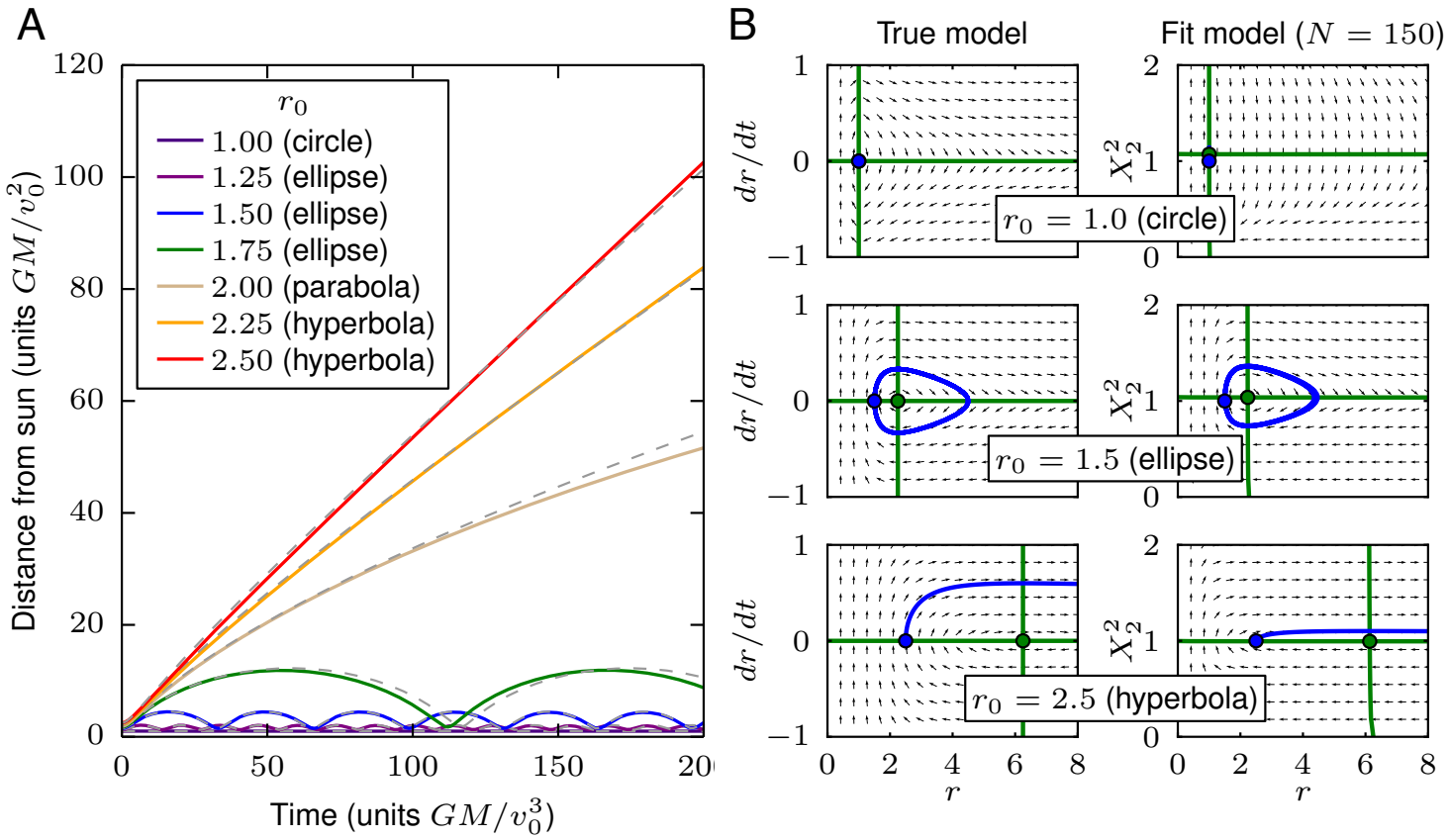
Interactions and input dependence

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



Daniels and Nemenman, Nature Comm 2015; PLoS ONE 2015

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

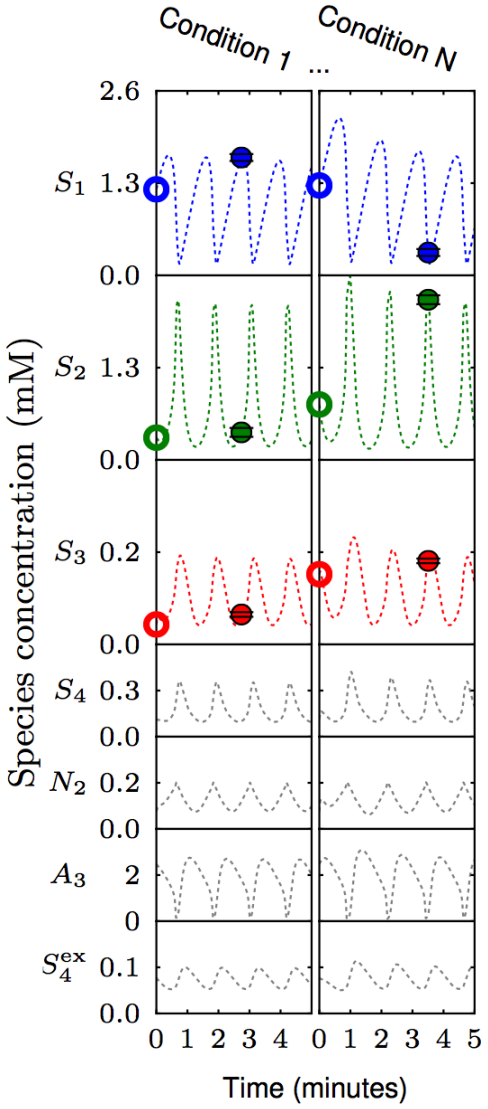


- Finds the hidden variable needed to account for the Newton's laws.
- Accounts for different classes of trajectories.

Daniels and Nemenman, Nature Comm 2015, PLoS ONE 2015

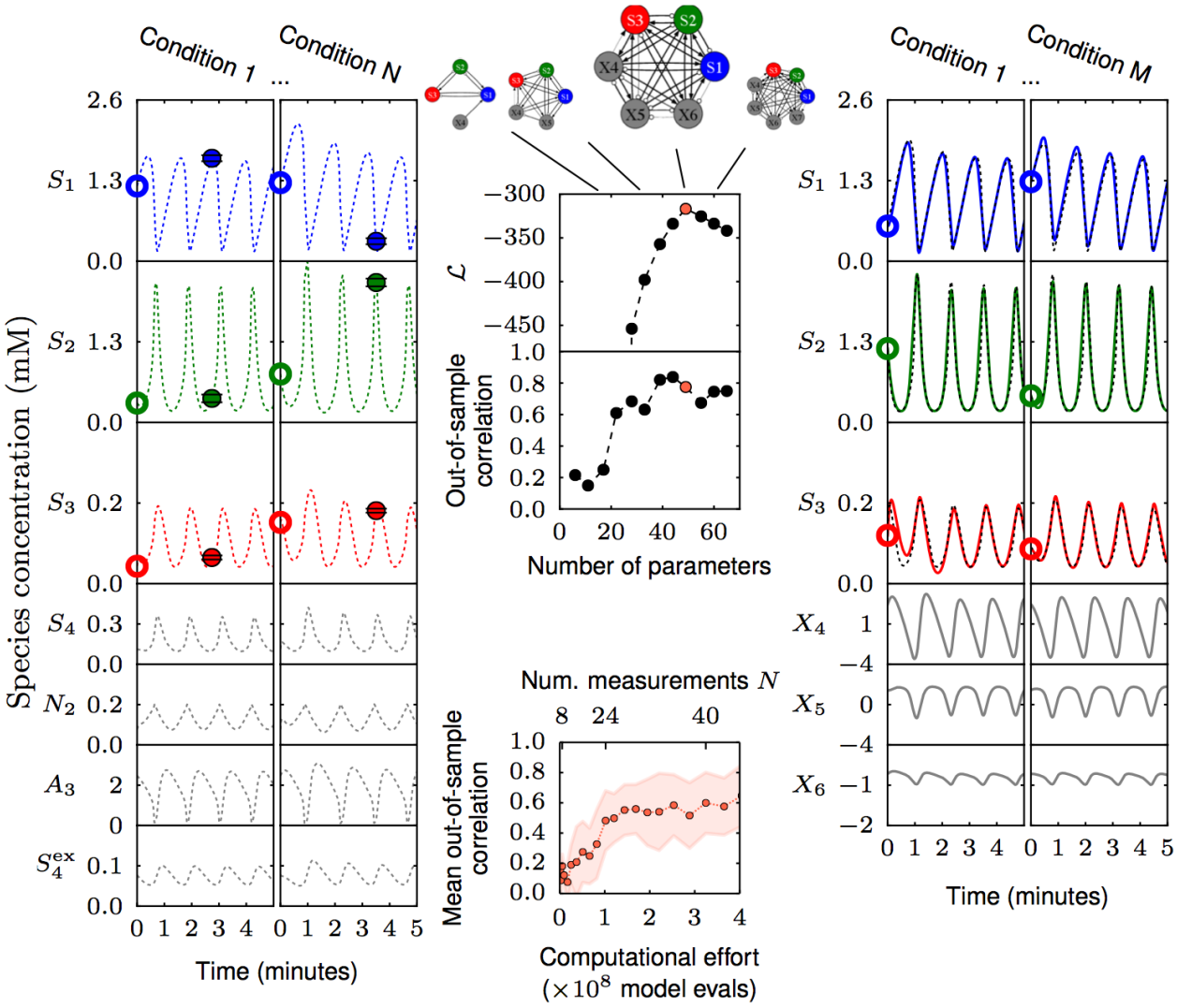
The yeast glycolytic oscillations: Complex dynamics needing complex structure

- 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; PLoS ONE 2015

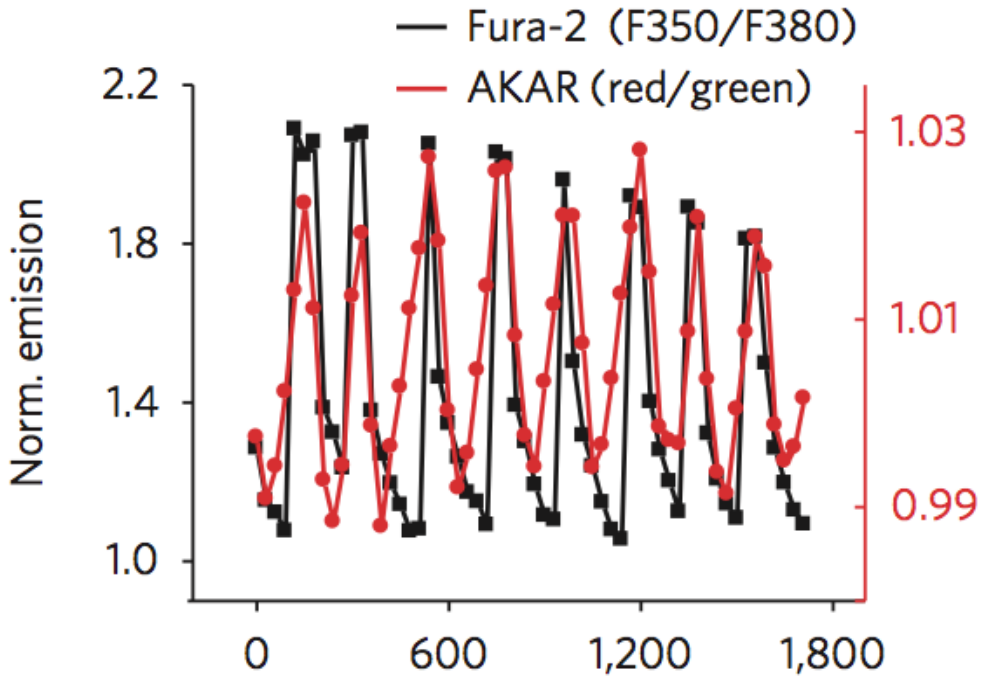


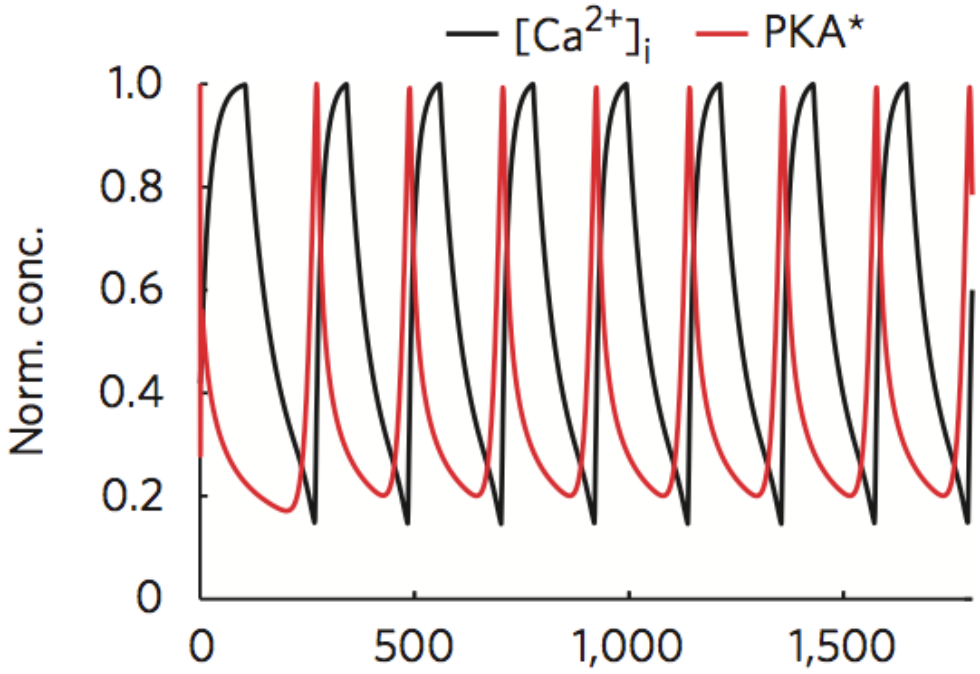
- ~100x fewer evaluations for the same accuracy compared to full search.
- ~1000x fewer data points than full search.
- Better accuracy than curve fitting.
- Linear scaling with the amount of data and with the number of variables.

Calcium-PKA oscillatory dynamics in beta cells

Experiment

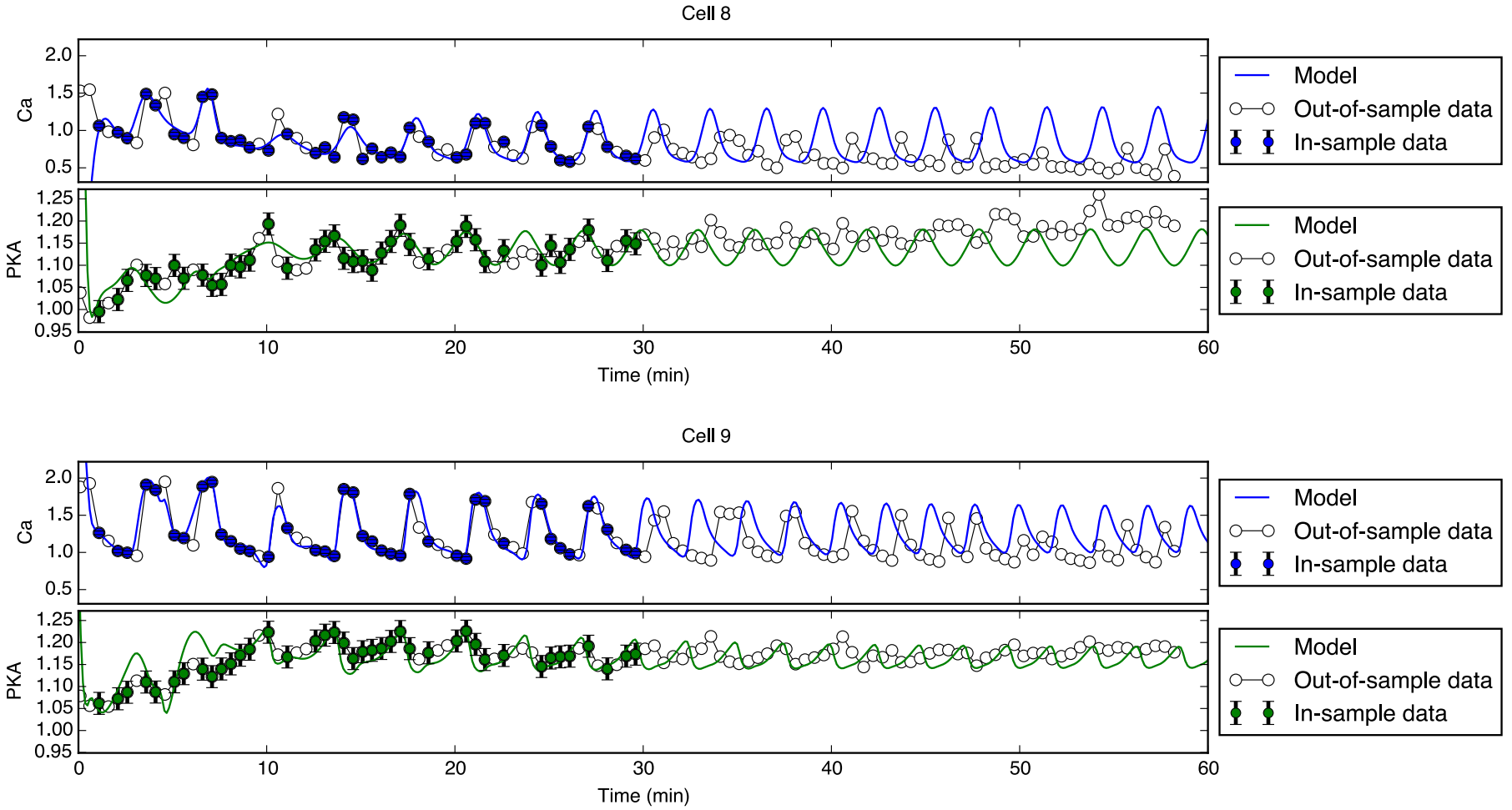
Mechanistic model:
11 ODEs, 78+ parameters,
fits all cells with parameter adjustments





Ni et al, NCB, 2010

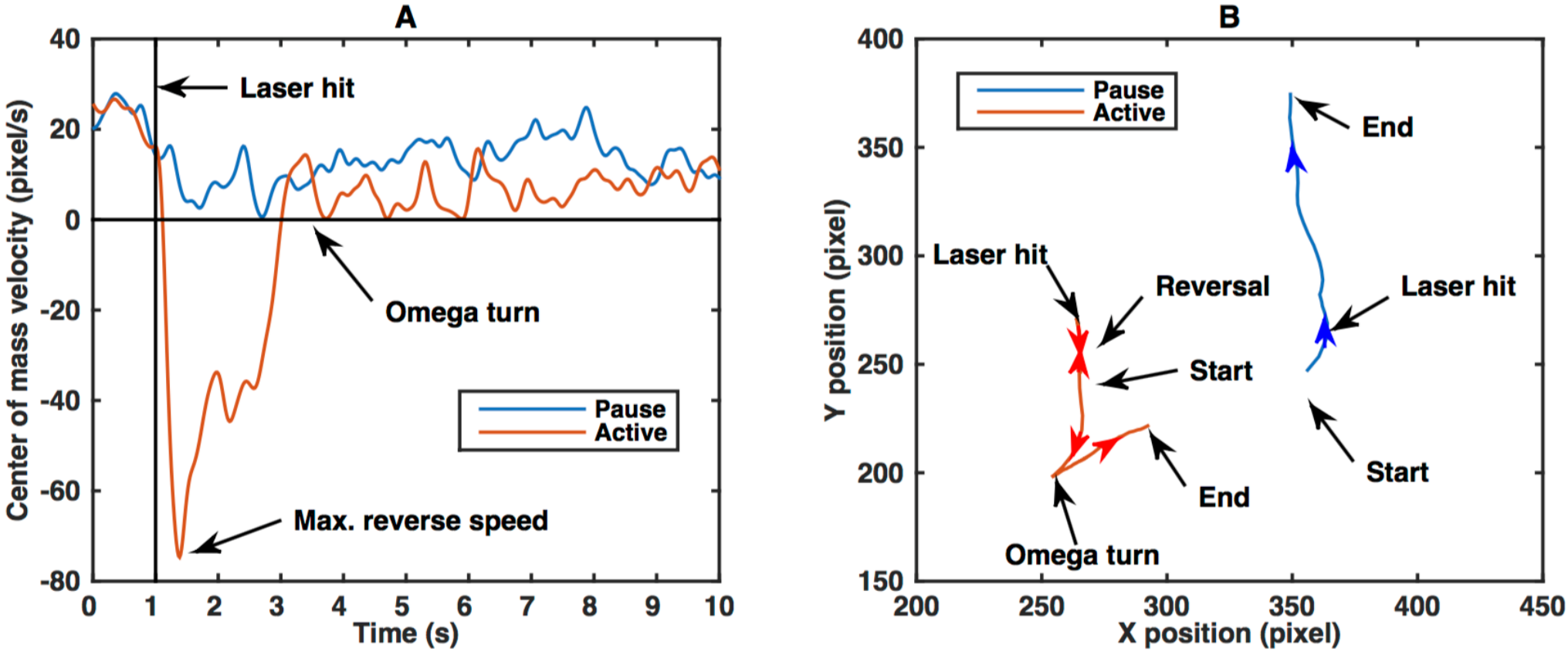
Calcium-PKA oscillatory dynamics in beta cells: automatic inference



6 ODEs
<40 parameters

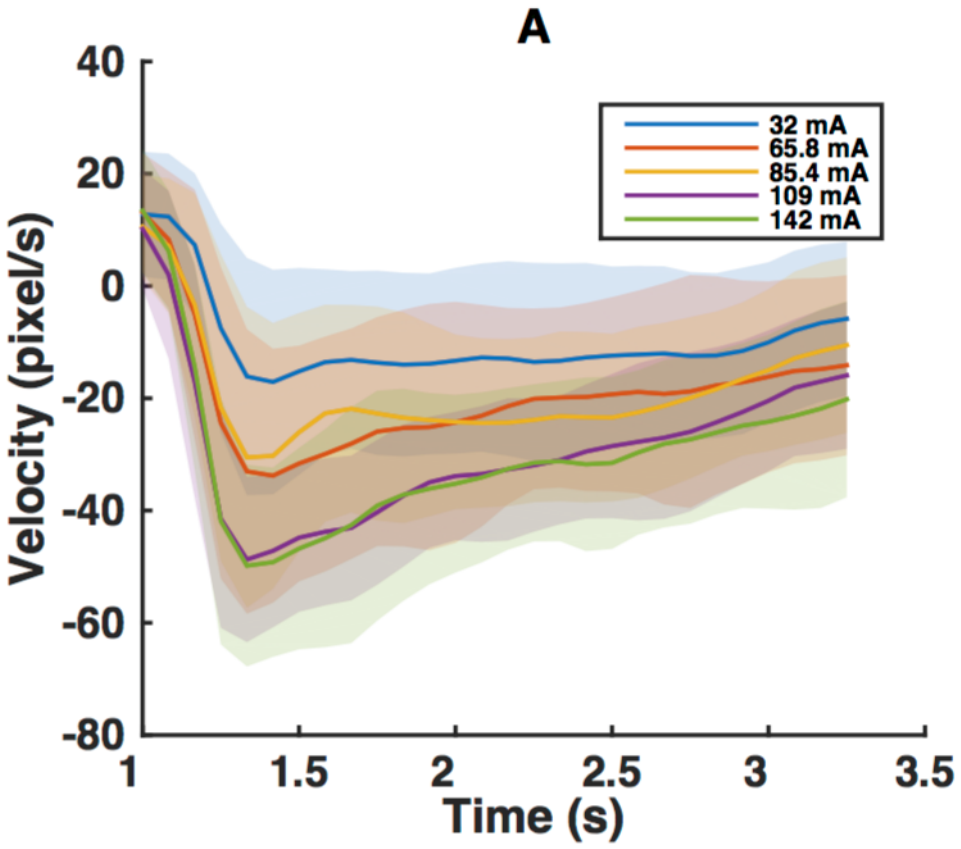
Daniels, IN, Levchenko, in prep.

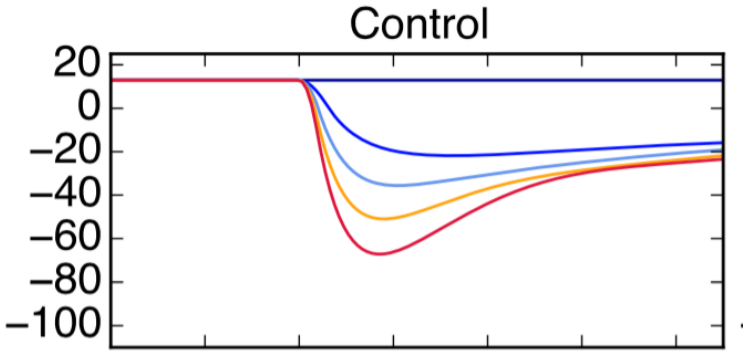
Modeling *C. elegans* escape response

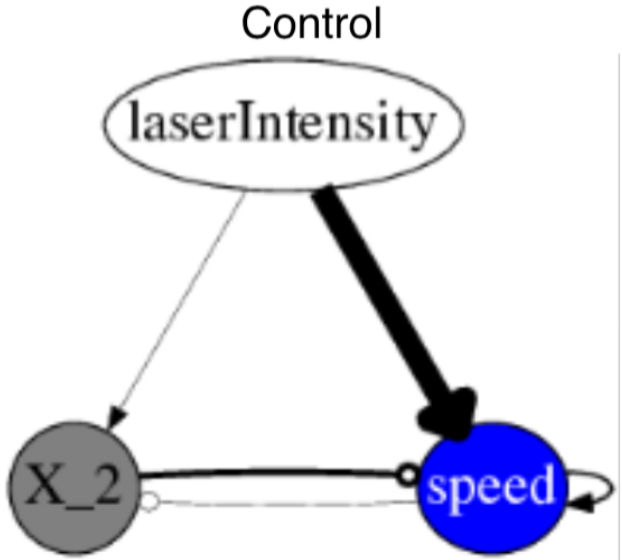


http://www.physics.utoronto.ca/~wryu/ryulab/movies/BJ02_044_run2_LS_NS_v1.mpg

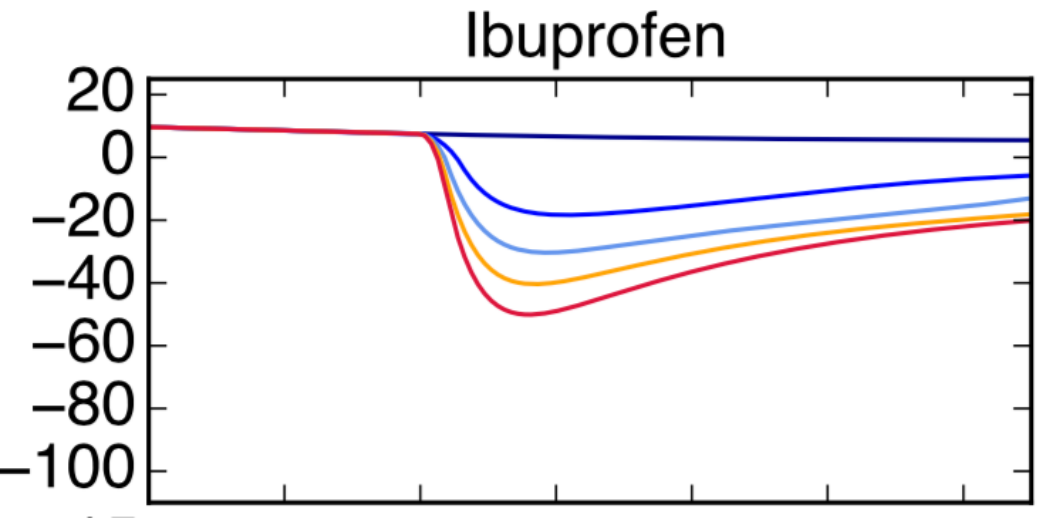
Data and fits

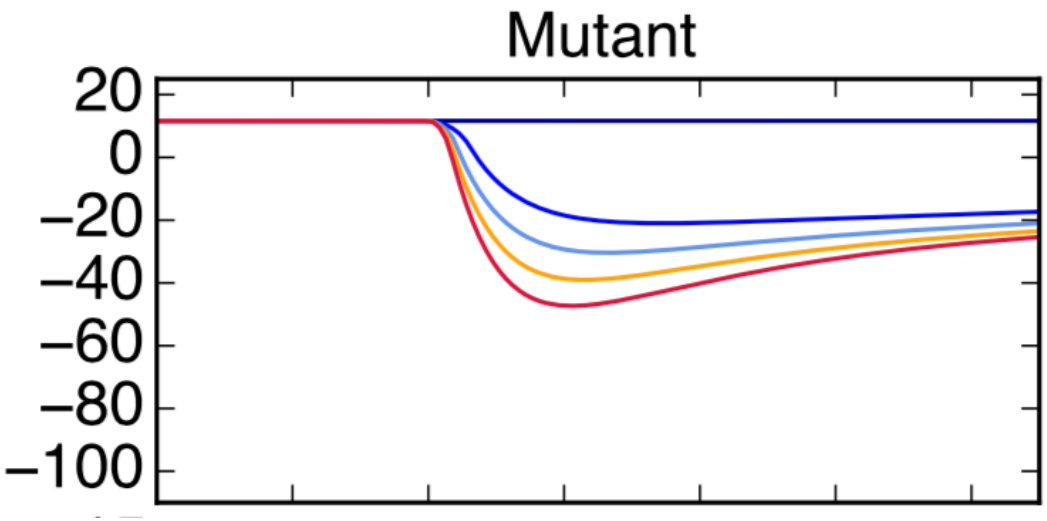






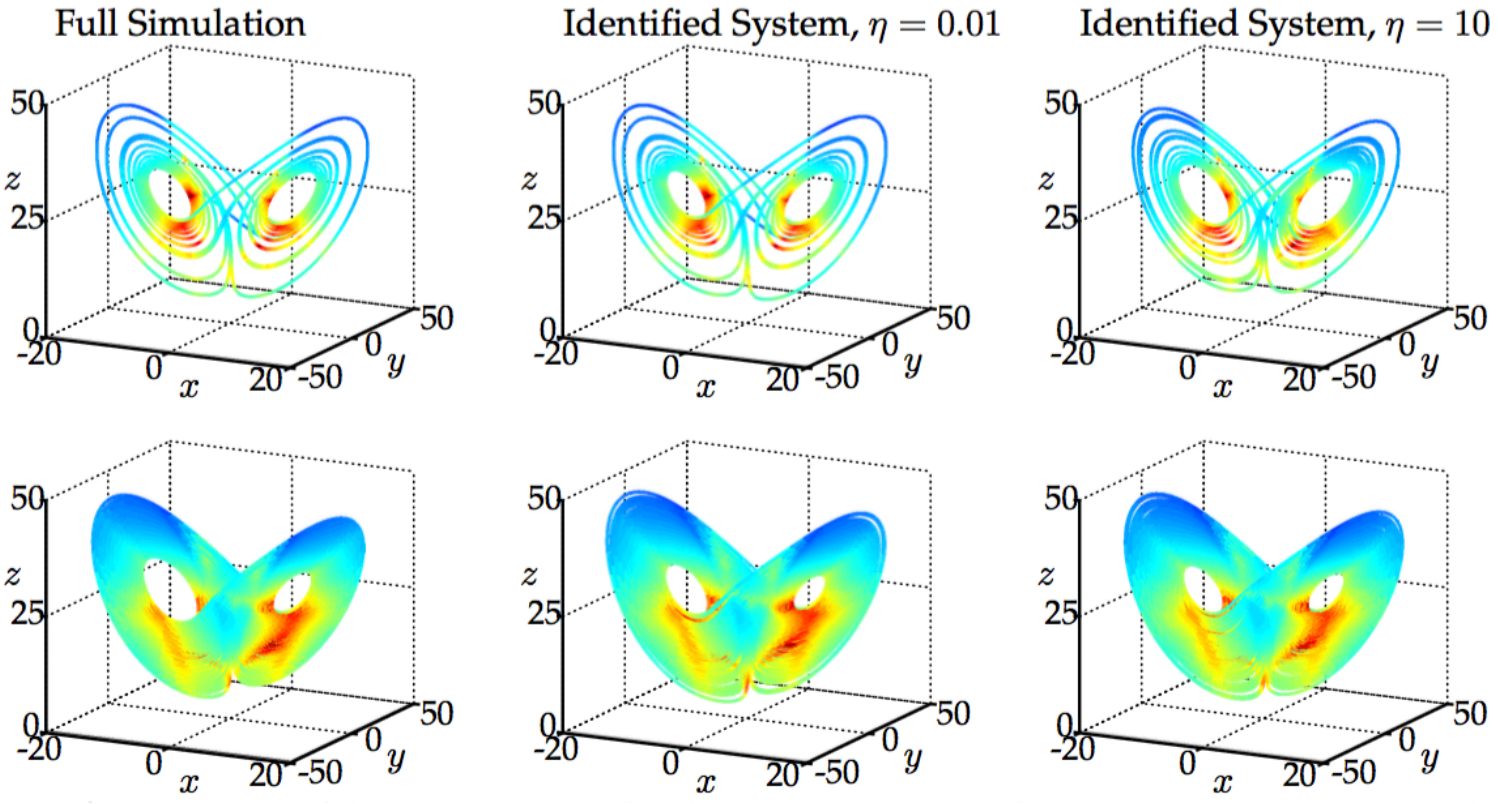
And the same model also explains treads worms





We are not alone

- Sparse regression for automated model inference (Brunton et al., *PNAS* 2016)



We are not alone

- Sparse regression for automated model inference (Brunton et al., *PNAS* 2016)
- Evolutionary search for network models (Francois et al., *PNAS* 2014)
- Dynamical systems approaches for characterization of attractors (Sugihara et al., *Science* 2012)

- Come to APS in New Orleans for more of this.

Conclusions

- Maybe gold in biology is in theories?
  - coarse-grain, refine — but loose details! Don't model a cat by a cat.
- Search for **phenomenological** dynamics instead of exact.
- Why do this?
  - **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*).