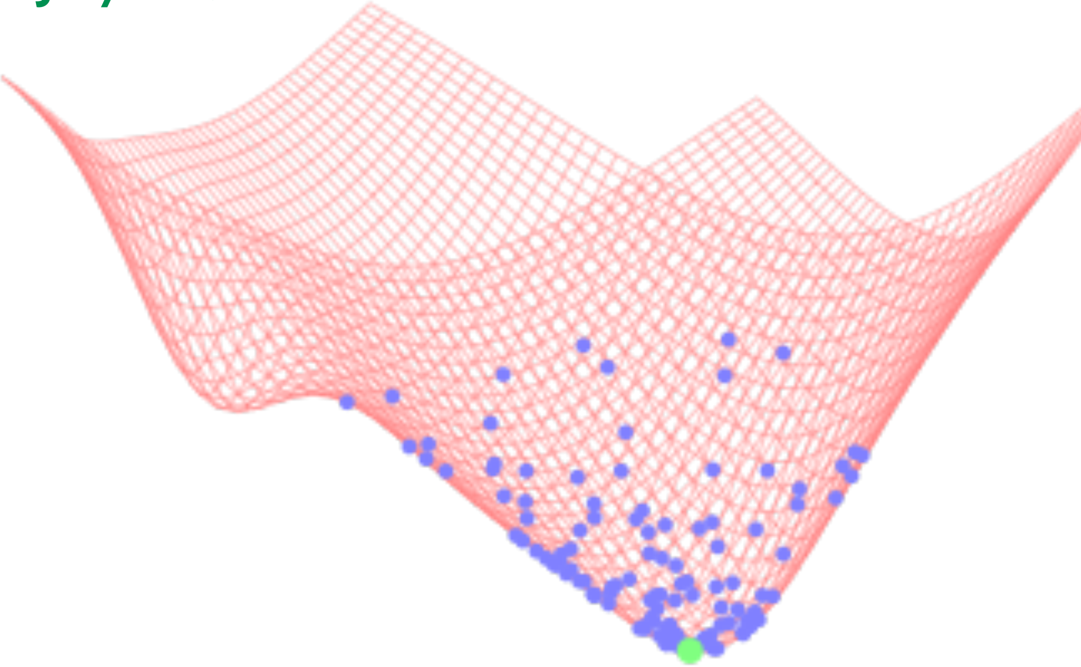
Systems biology's dirty secret:

Parameter estimation, sensitivity analysis, and sloppiness

Ryan Gutenkunst

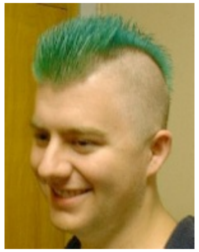
Molecular and Cellular Biology
University of Arizona

q-bio school - July 27, 2015

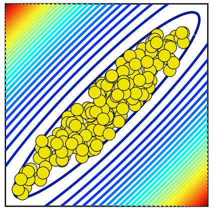


My story

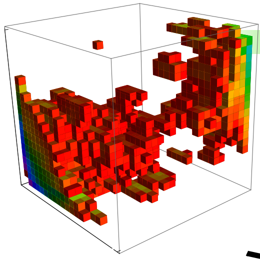
B.S. in physics

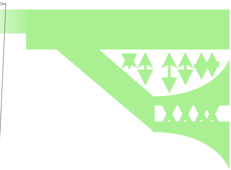


Ph.D. in physics, minor in biophysics with Jim Sethna

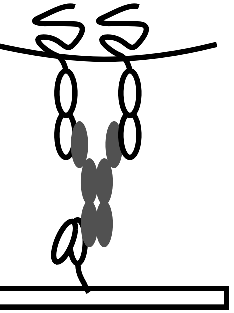


Postdoc in population genetics with Carlos Bustamante





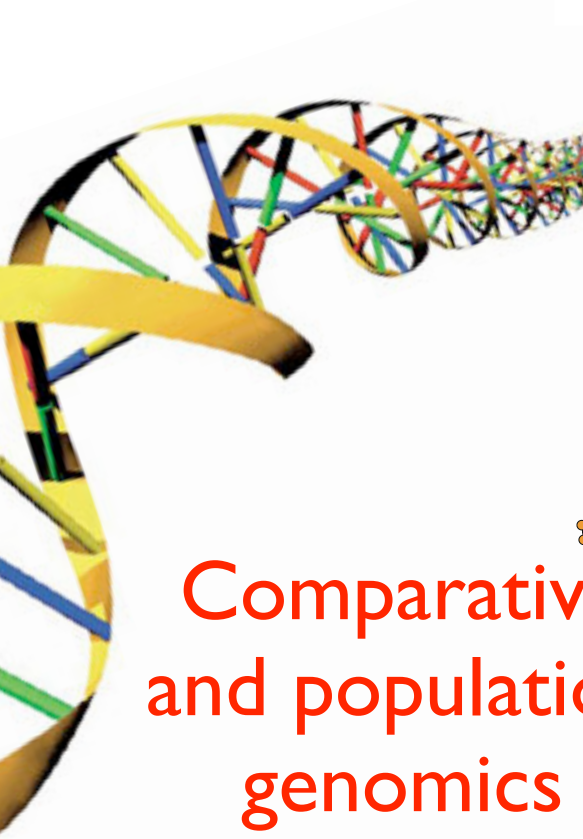
Postdoc in immune signal transduction with Byron Goldstein

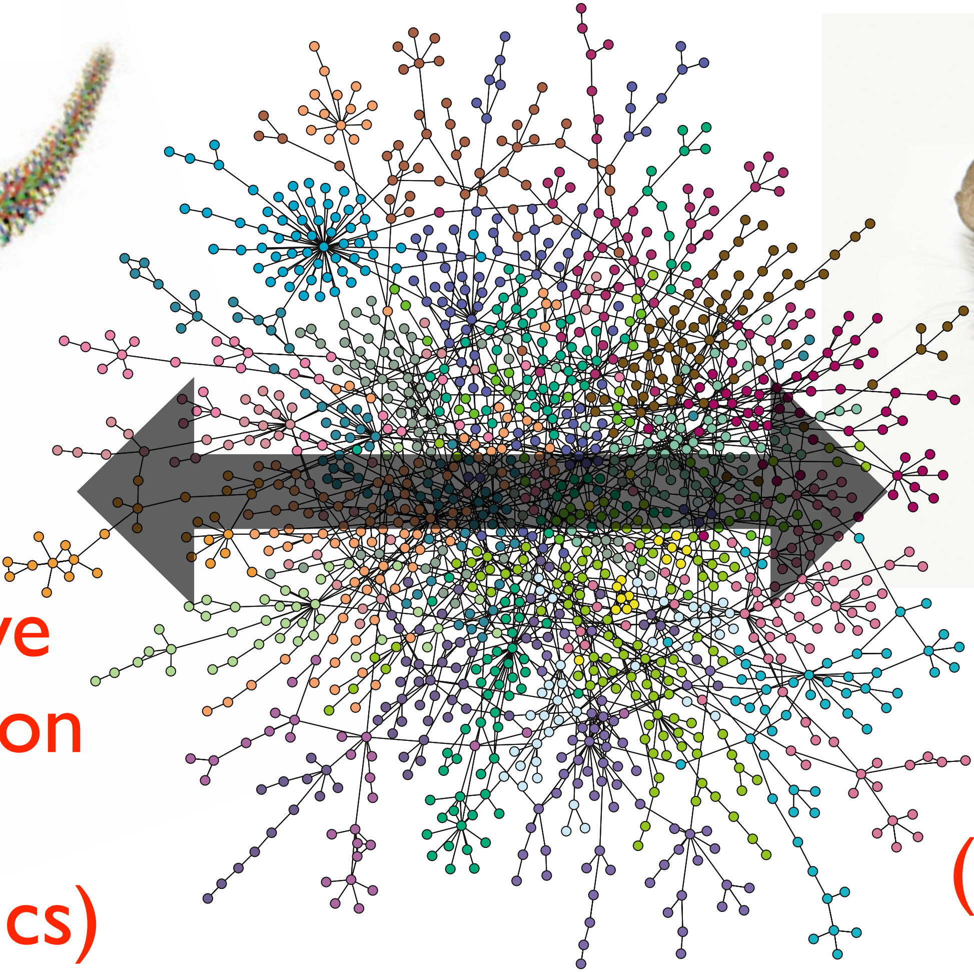


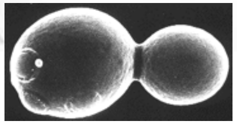
Faculty in Molecular and Cellular Biology

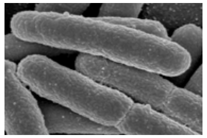
Affiliations: Applied Mathematics, Statistics, Ecology & Evolutionary Biology, Genetics, BIO5 Institute

Gutengroup









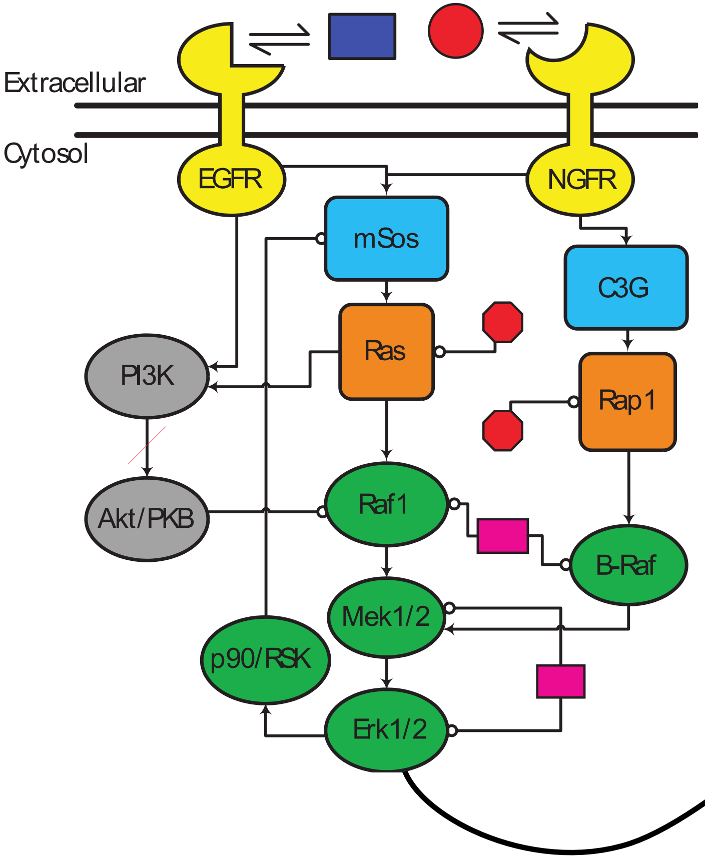
Comparative and population genomics (bioinformatics)

Systems biology (modeling)

http://gutengroup.mcb.arizona.edu

Networks, models, and parameters

Growth factor signaling in PC12 cells



Brown et al.
Phys Biol (2004)

Biochemically detailed models

Often very complex, but....

- Close correspondence with expts
- Can integrate with other pathways
- Close to evolutionary mechanism

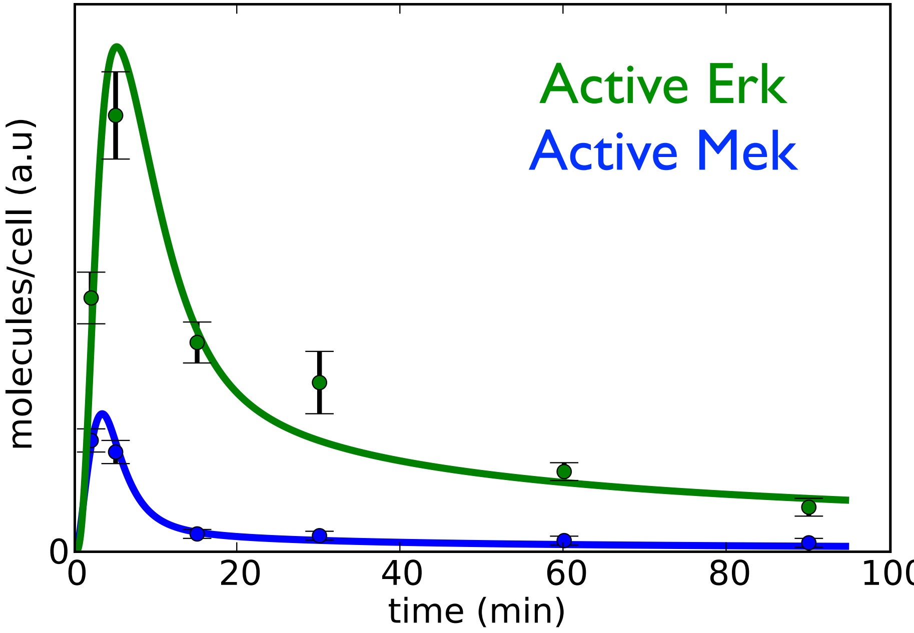
15 nonlinear differential equations

$$\frac{d\,[\mathrm{ErkActive}]}{dt} = \frac{\overset{?}{\mathrm{kpMekCytoplasmic}} \cdot [\mathrm{MekActive}] \cdot [\mathrm{ErkInactive}]}{[\mathrm{ErkInactive}] + \underset{?}{\mathrm{KmpMekCytoplasmic}}} - \frac{\overset{?}{\mathrm{kdErk}} \cdot [\mathrm{PP2AActive}] \cdot [\mathrm{ErkActive}]}{[\mathrm{ErkActive}] + \underset{?}{\mathrm{KmdErk}}}$$

But... 48 biochemical parameters $\vec{k}$, none quantitatively measured

Parameter fitting

- Biochemical parameters are difficult to measure directly
  - Need to express and purify protein
  - Measure *in vitro*, questionable extrapolation to *in vivo*



- Measuring cellular responses often easier (and more interesting)
  - Model parameters need to be fit

What to extremize?

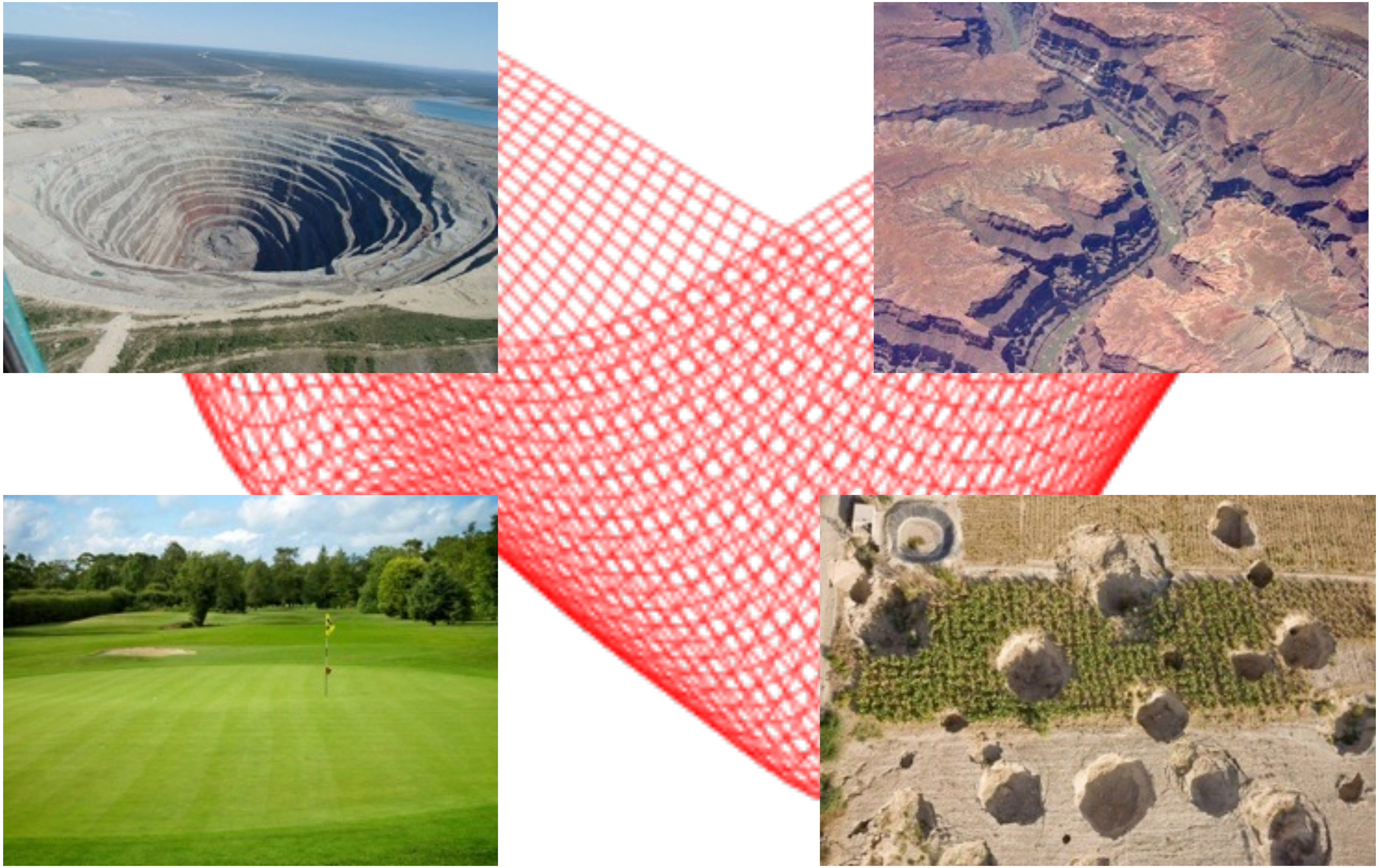
- Maximizing the likelihood of the data given the model extracts maximal information about parameters.
- Likelihood: probability of generating the observed data given your model and parameter values.
- Independent data points with Gaussian noise:

$$\mathcal{L} = \prod_i \exp\left[-\frac{\left(y_i(\vec{\theta}) - d_i\right)^2}{2\sigma_i^2}\right]$$

$$-\log\mathcal{L} = \frac{1}{2}\sum_i \frac{\left(y_i(\vec{\theta}) - d_i\right)^2}{\sigma_i^2} \equiv \sum_i r_i^2 \equiv C(\vec{\theta})$$

Inhomogenous data typically demands a more ad-hoc approach (e.g. fitting Western blots + flow cytometry)

Cost landscape



Optimization methods

- “Local” optimizers
  - Nelder-Mead simplex (“amoeba”)
  - Steepest descent, Conjugate gradient
  - Levenberg-Marquardt

- “Global” optimizers
  - Simulated annealing
  - Genetic algorithms

See Numerical Recipes
or Ashyraliyev et al. *FEBS Lett* (2009)

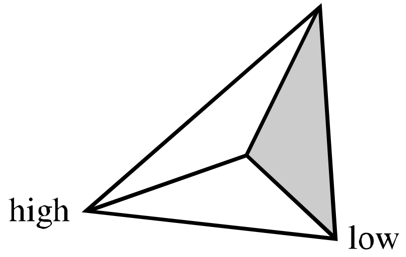
General advice

- An art, rather than a science
- Method comparisons are dubious, since performance can be very problem-specific

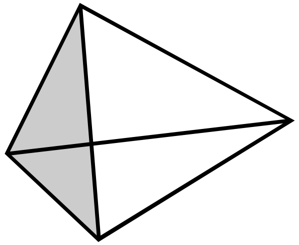
- Hand-fiddling to use your brain is useful, both to develop understanding and to find a starting point
- Most optimizers work best if all parameters have similar scale

Nelder-Mead simplex (“amoeba”)

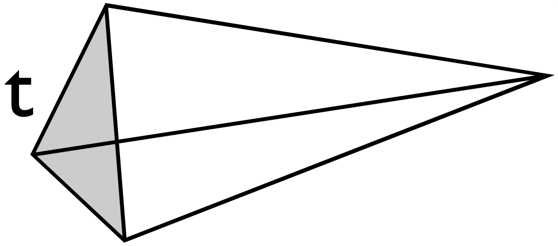
N+1 points define a tetrahedron
in N-dimensional parameter space.



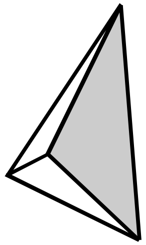
- Reflect worst point across tetrahedron



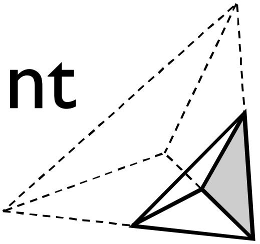
- Reflect and expand worst point



- Contract worst point



- Contract whole tetrahedron toward lowest point



Derivative-free, so very robust,
but slower than gradient-based methods

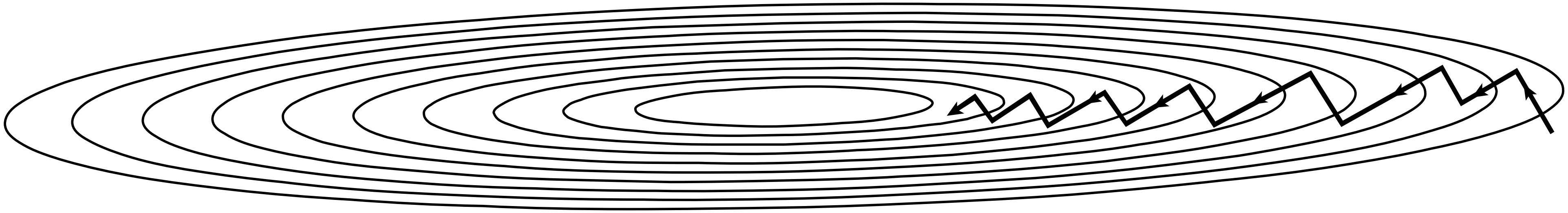
Steepest descent

1. Calculate gradient

2. Minimize along gradient direction

Simple and intuitive

Performs very poorly, because each step must be orthogonal to the previous.



Solution: conjugate gradient, to pick more productive directions.

Levenberg-Marquardt

$$C = \frac{1}{2}\sum_i r_i^2$$

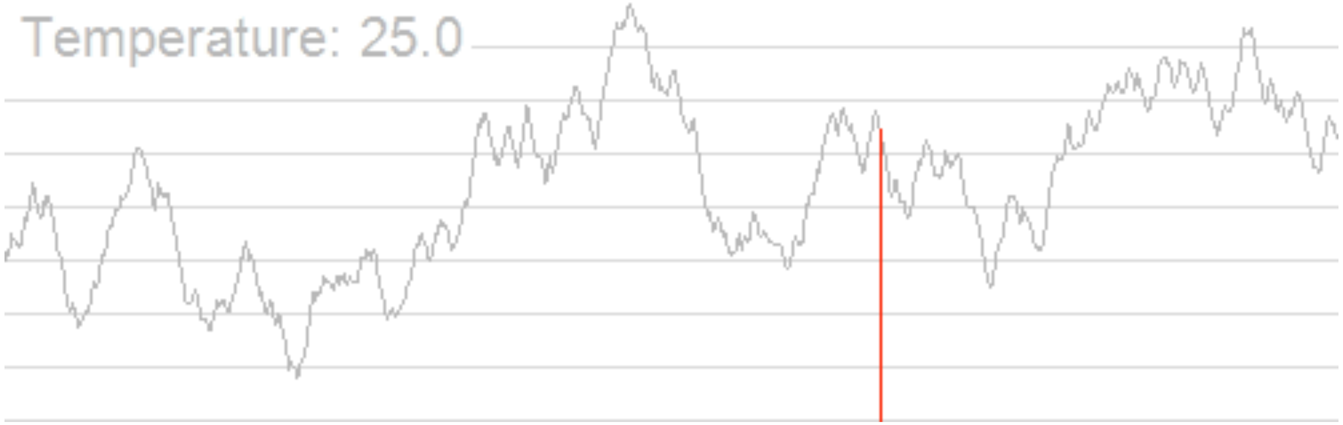
$$\frac{\partial^2 C}{\partial\theta_j \partial\theta_k} = \sum_i \frac{\partial r_i}{\partial\theta_j}\frac{\partial r_i}{\partial\theta_k} + \cancelto{\sim 0}{\sum_i r_i \frac{\partial^2 r_i}{\partial\theta_j \partial\theta_k}}$$

- Direct estimate of quadratic form, using only single derivatives

- Very efficient when started “close to” local optimum

Simulated annealing

- Each step test a new set of parameters sampled from a proposal density.
- If C' < C accept move with probability 1, otherwise accept with probability exp[(C - C')/T].
- Slowly reduce T to zero, via cooling schedule.
- Guaranteed convergence if cooling is "slow enough"
- Robust, applicable to discrete optimization, but slow



Evolutionary optimization

- Population of "individuals", each a set of parameters
- Apply mutations (changes in single parameter values) and recombinations (swaps of multiple values between individuals)
- Fitness of each individual is inversely proportional to cost
- Next generation reproduce according to fitness
- Robust, very easy to parallelize.

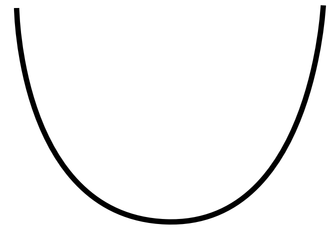
Sensitivity analysis

- How sensitive is your model to parameter changes?
- Conversely, how reliable are your parameter estimates?

- 1-D
- Multi-dimensional

1-dimensional sensitivities

- Transects of the cost function
  - Width is proportional to uncertainty



- First derivatives of interesting quantities are “easy” with ODEs

$$\frac{d\vec{y}}{dt} = f(\vec{y}, t, \vec{p})$$

$$\frac{d}{dt}\frac{dy_i}{dp_i} = \frac{\partial f}{\partial p_i} + \sum_j \frac{\partial f}{\partial y_j}\frac{dy_j}{dp_i}$$

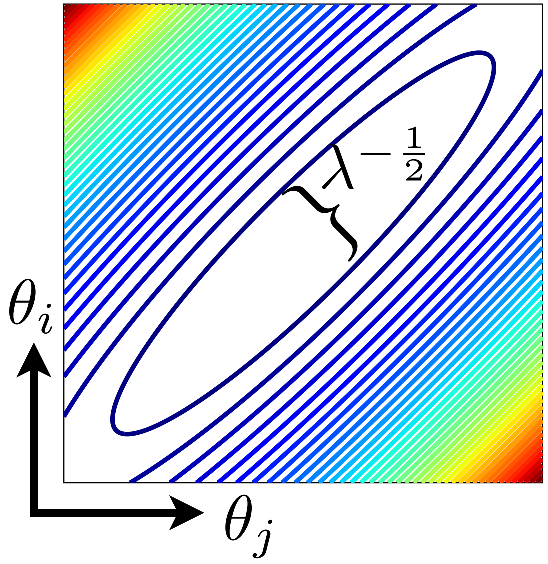
Multidimensional sensitivities

- Quadratic form

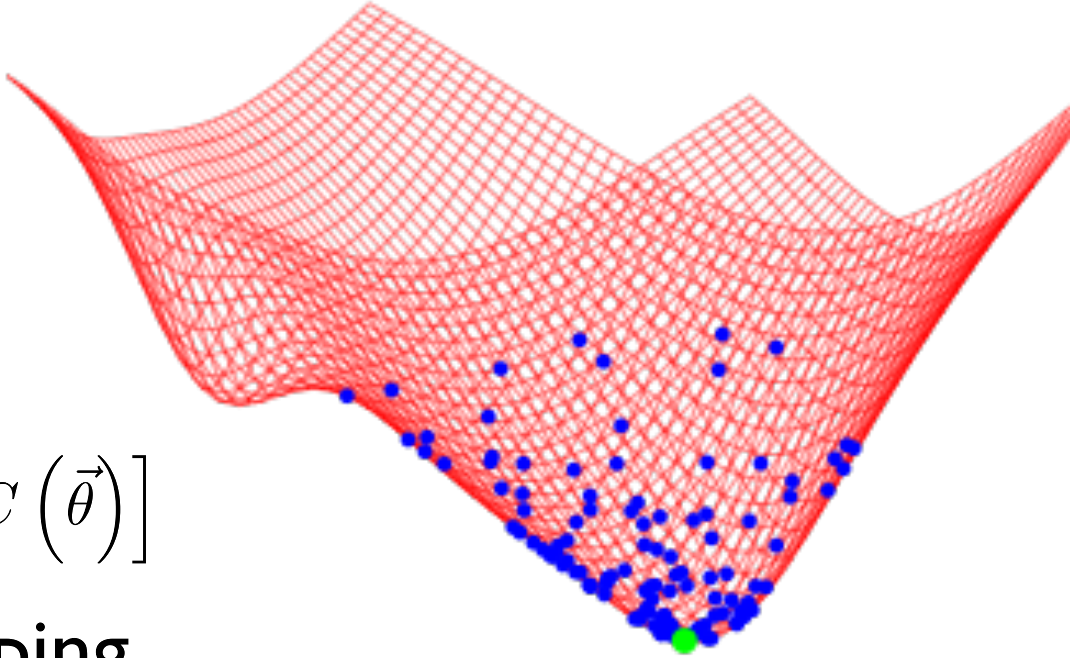
$$C(\theta) = C(\theta^*) + (\theta - \theta^*) \cdot H \cdot (\theta - \theta^*) + \cdots$$

$$H_{ij} = \frac{d^2 C}{d\theta_i d\theta_j}$$

- Approximating probability distributions as multidimensional normal or log-normal



Multidimensional sensitivities



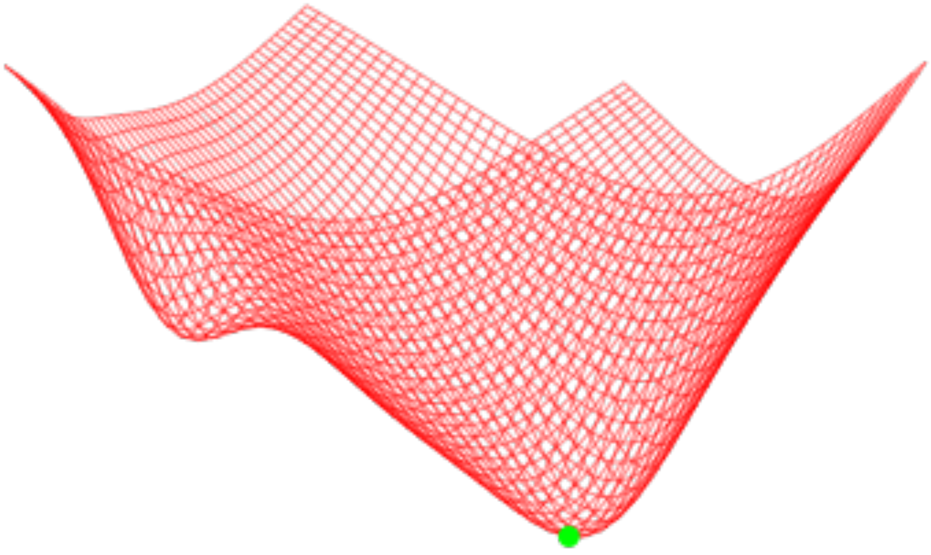
- Parameter ensembles
  - Bayesian MCMC

$$P\left(\vec{\theta}|D\right) = \frac{P\left(D|\vec{\theta}\right)}{P\left(D\right)} P\left(\vec{\theta}\right) \propto \exp\left[-C\left(\vec{\theta}\right)\right]$$

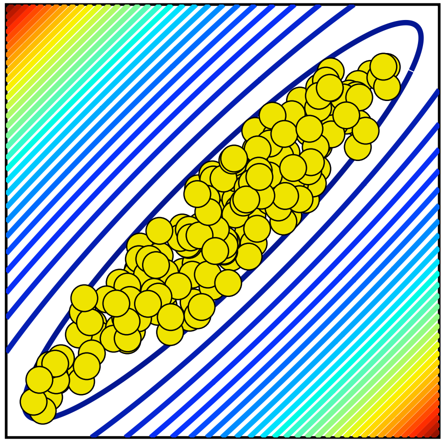
  - Frequentist bootstrapping (resampling of data)
  - Approximate Bayesian Computation (when can’t compute the likelihood, use summary statistics)

Summary

- Parameter optimization is hard
- Your toolbox should contain a variety of algorithms, both local and global
- Algorithms are no substitute for understanding your model and your data
- Even trickier for stochastic systems



Sloppiness in biochemical networks



Ryan Gutenkunst

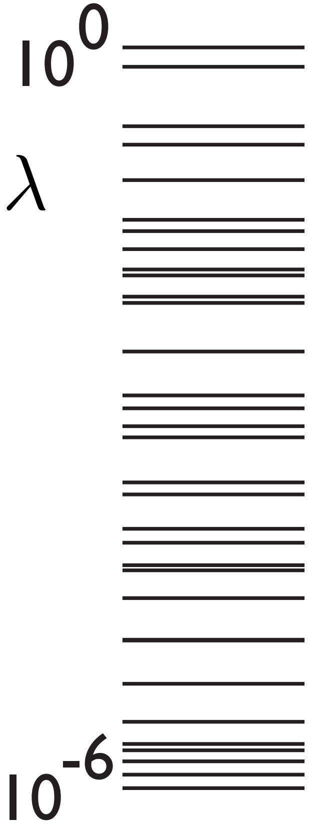
Molecular and Cellular Biology

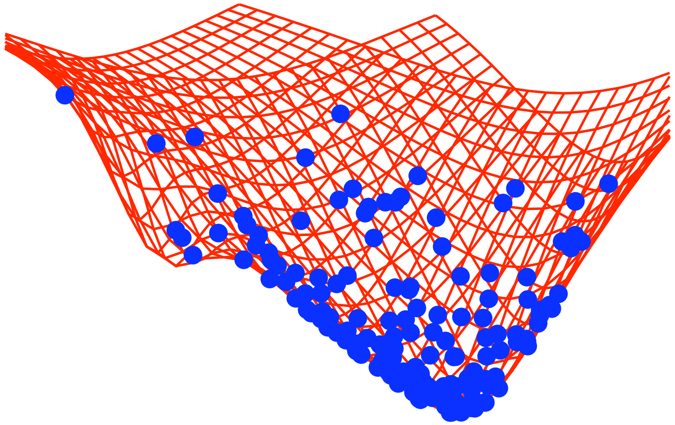
University of Arizona

July 27, 2015

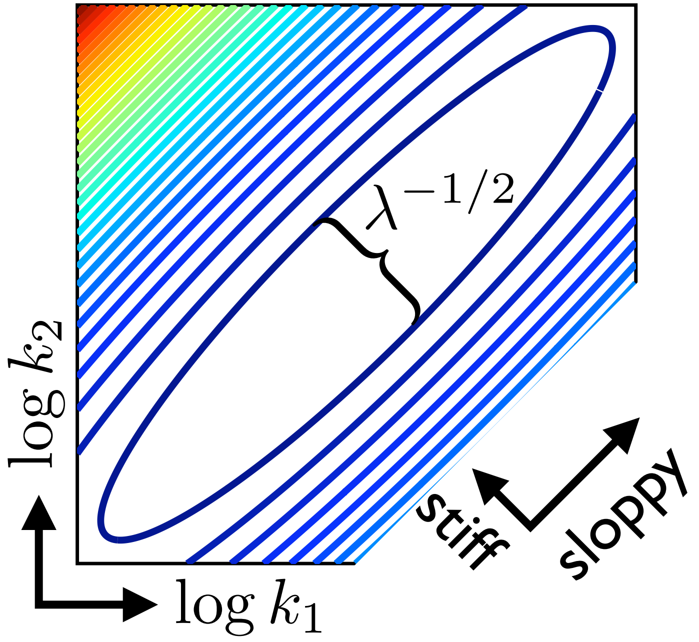
with Jim Sethna, Chris Myers

Kevin Brown, Josh Waterfall, Fergal Casey





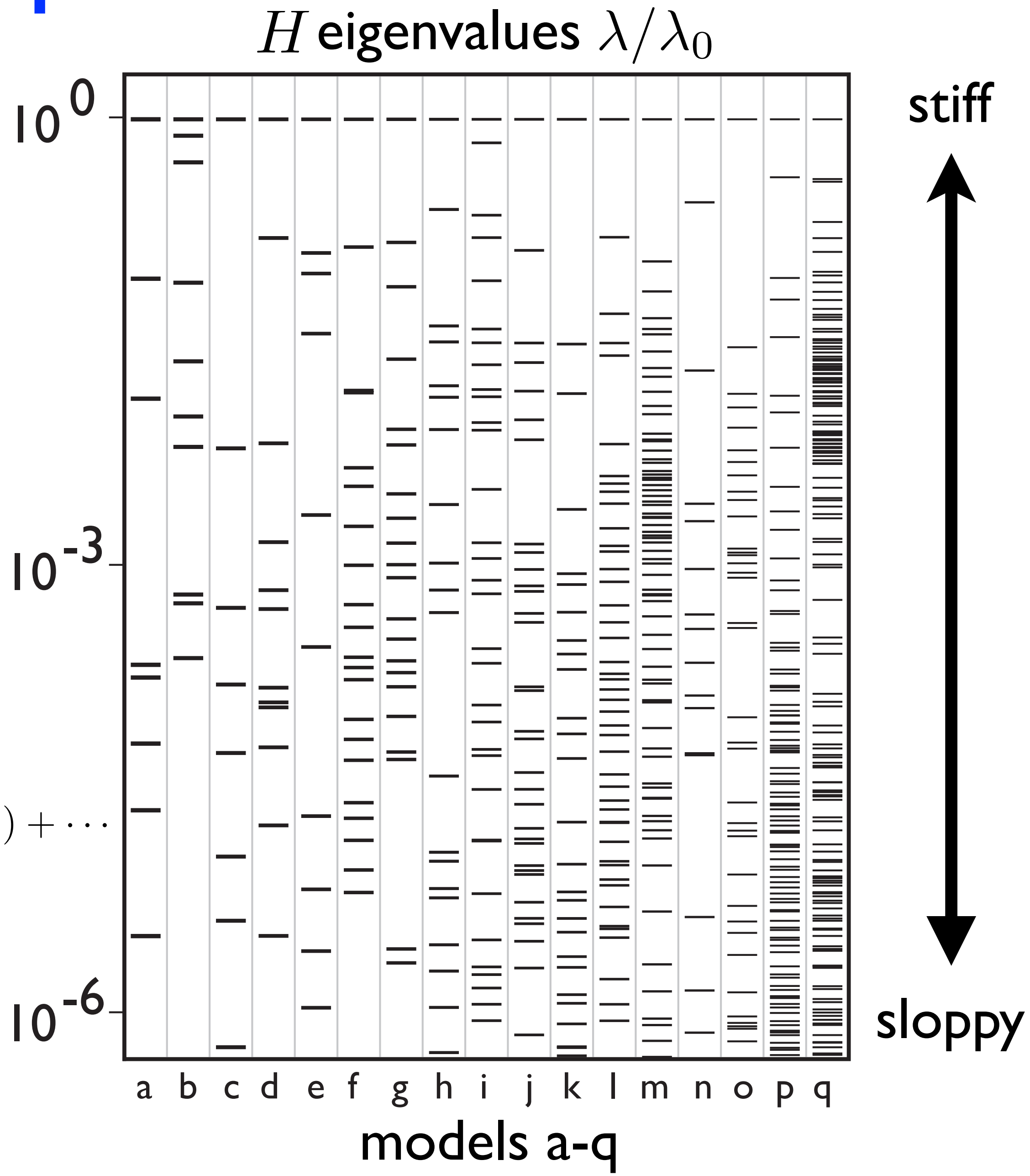
Sloppiness



$$\chi^2(\vec{k}) \propto \sum_y \int \left(\frac{y(t,\vec{k}) - y(t,\vec{k}_0)}{\sigma_y} \right)^2 \mathrm{d}t$$

$$\chi^2(\vec{k}) = \chi^2(\vec{k}_0) + (\log \vec{k} - \log \vec{k}_0) \cdot H \cdot (\log \vec{k} - \log \vec{k}_0) + \cdots$$

$$H_{ij} = \frac{d^2\chi^2}{d \log k_i \, d \log k_j}$$

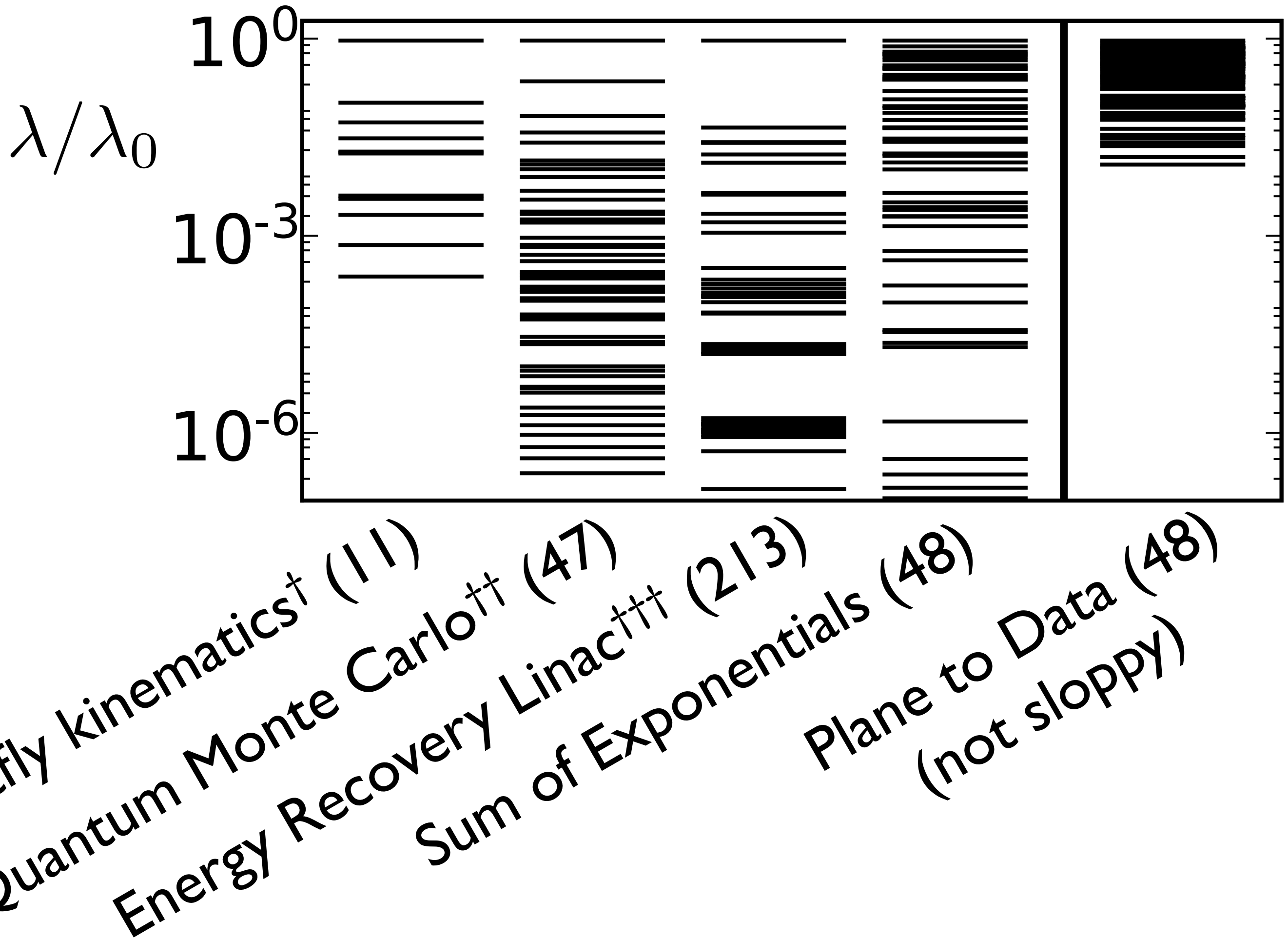


Sloppiness is universal in biochem. network models.

Gutenkunst et al. (2007) *PLoS Comp Biol*

Erguler et al. *Mol Biosyst* (2011) - 160 more sloppy models

Sloppiness elsewhere

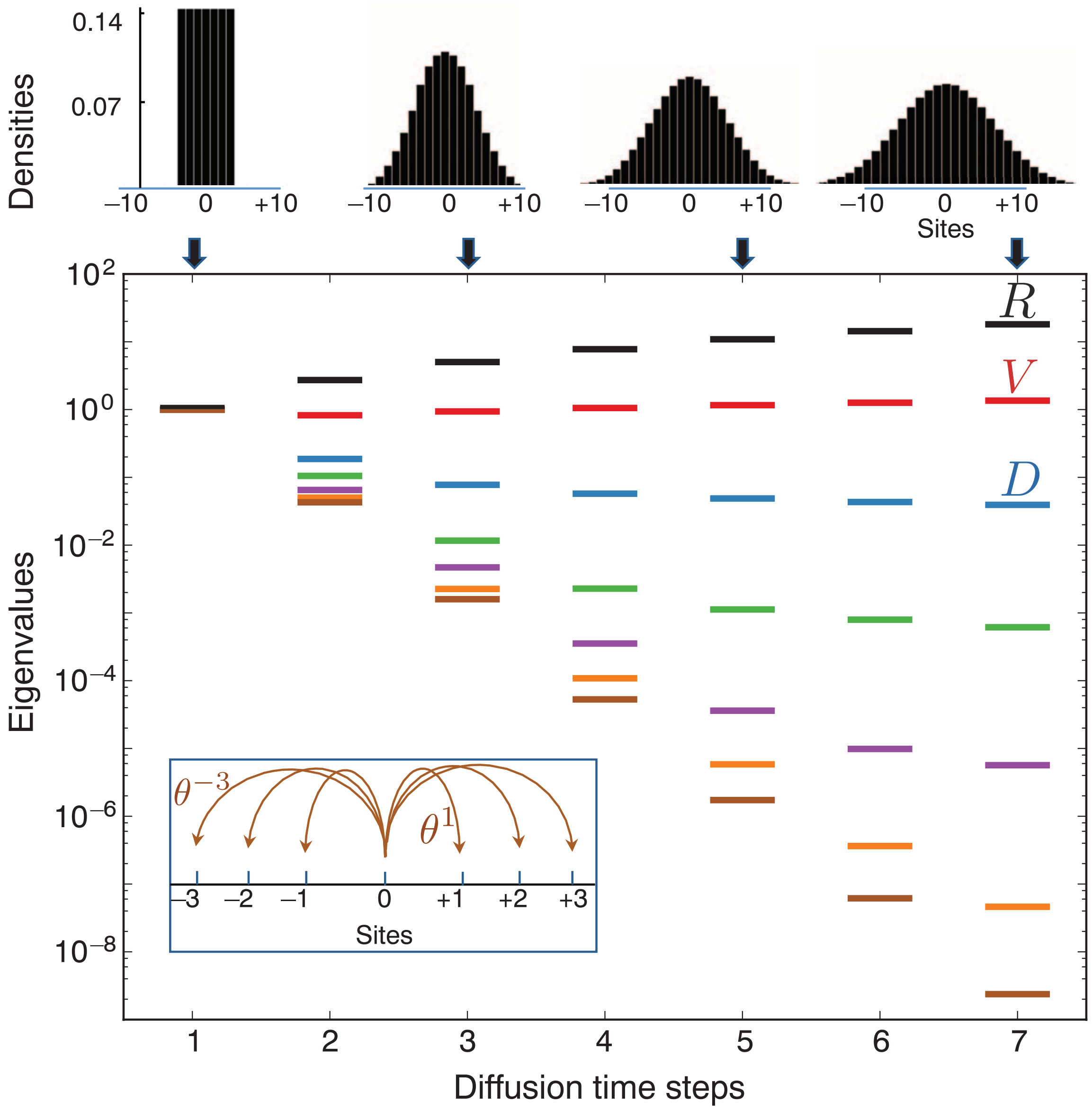


Sloppiness is a general feature of nonlinear least-squares fits.

† Gordan Berman, Jane Wang
†† Cyrus Umrigar
††† Chris Mayes, Georg Hoffstaetter

Origins of sloppiness

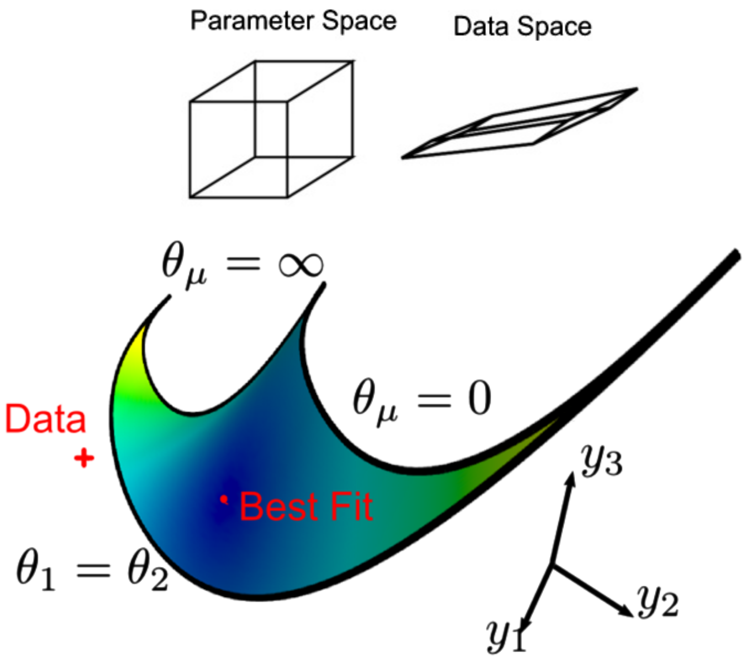
In some simple models, sloppiness can be shown to arise from macroscopic observations that obscure microscopic parameter effects.

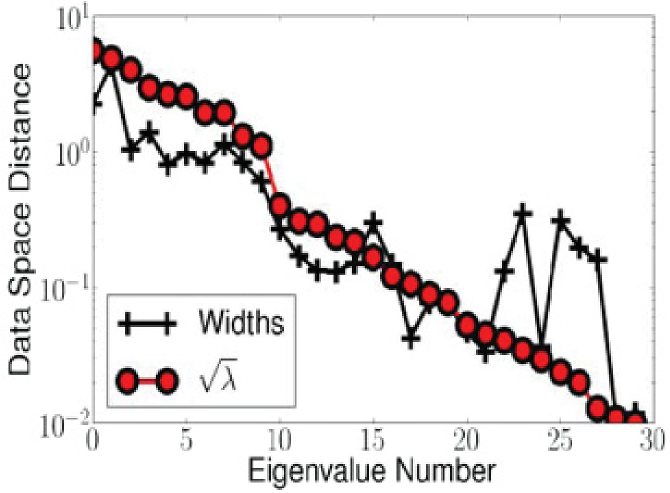


Machta et al. (2013) *Science*

Information geometry

- A model is a mapping from M-dimensional parameter space to a manifold within N-dimensional data space (N > M)
- For non-linear models, these manifolds are often bounded and contain singular points.
- Local sloppy analysis predicts the global shape of this manifold.
- These torture optimizers, but clever algorithms can work around them.



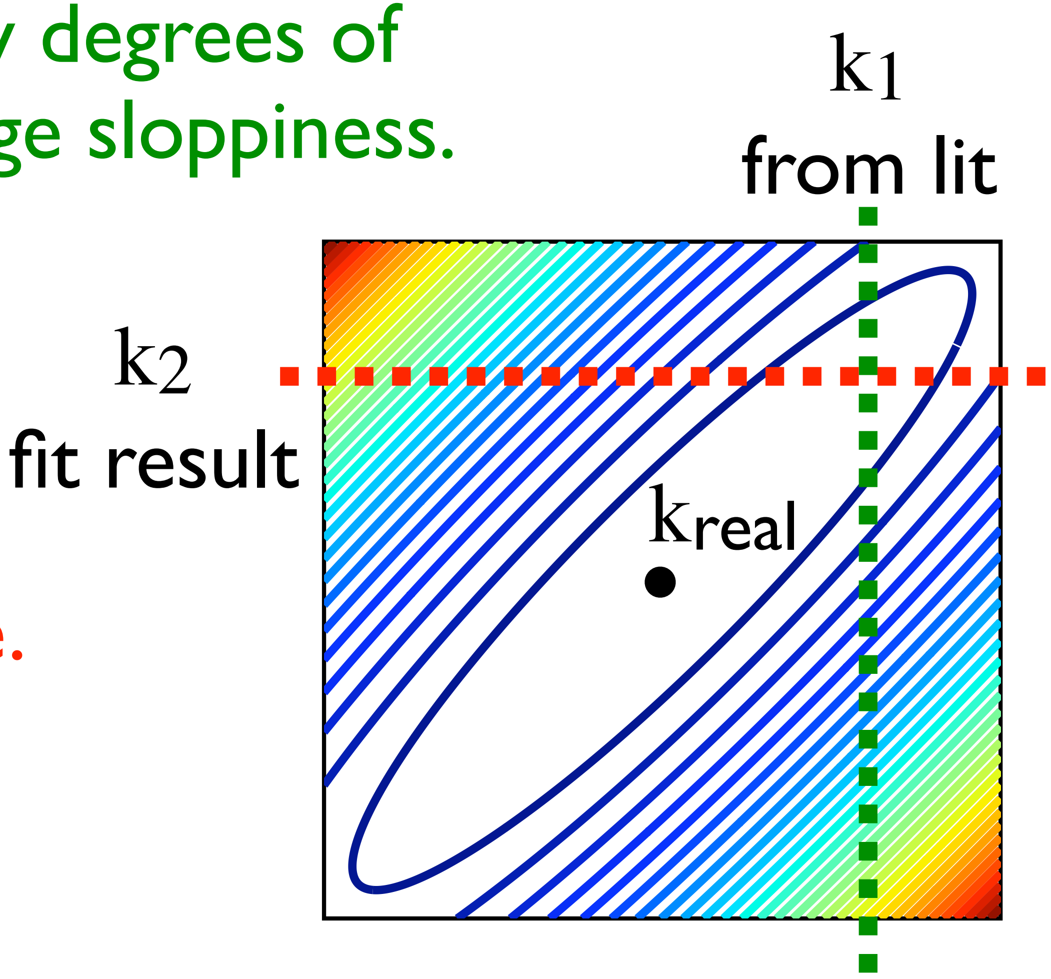


Transtrum, Machta, Sethna
(2010) *Phys Rev Lett*
(2011) *Phys Rev E*

Why do literature params work?

Often, previous experiments were done in a different cell type or *in vitro*. Why do those parameter values work in other model contexts?

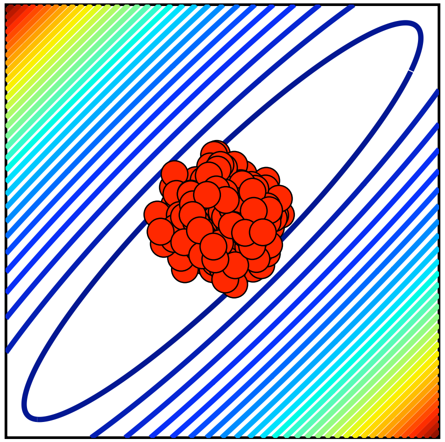
Usually, at least a few degrees of freedom left to leverage sloppiness.



In sloppy basin, so fit is still reasonable.

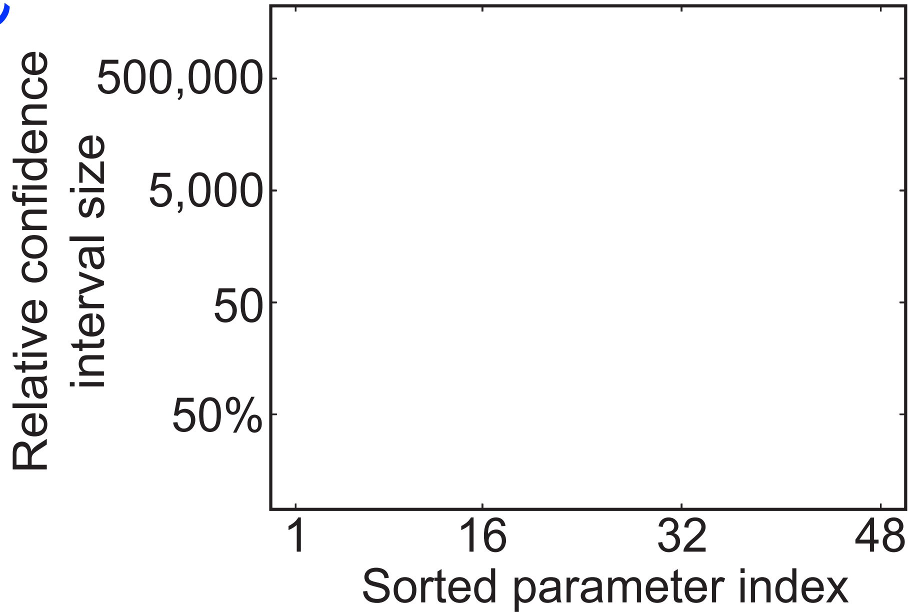
Sloppiness & uncertainties

(All uncertainties by MCMC)

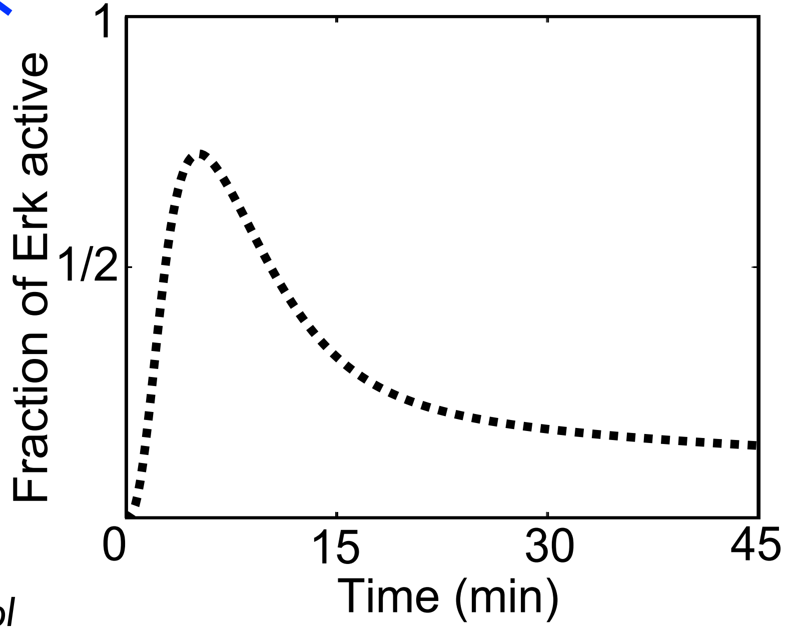


All measured

Parameters



Prediction

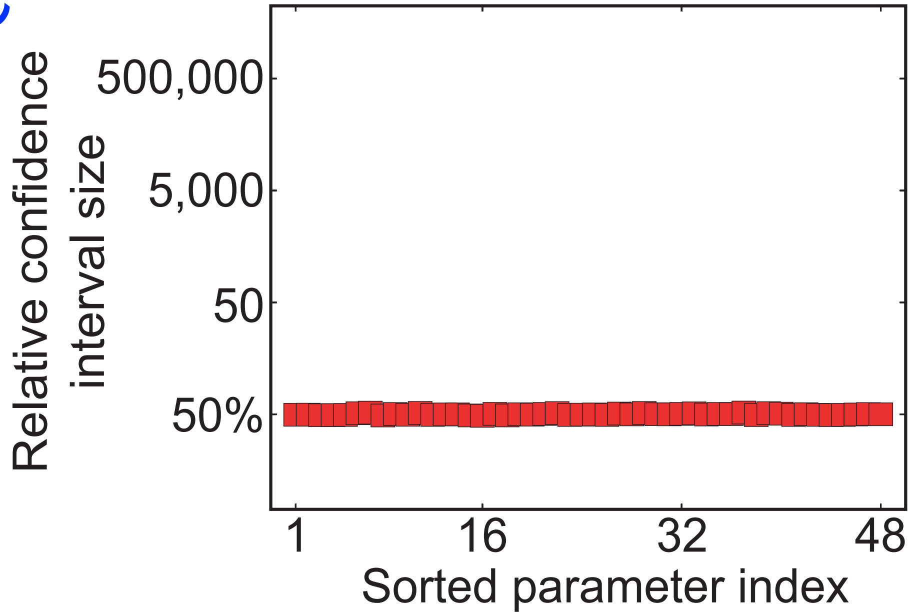


Gutenkunst et al.
(2007) *PLoS Comput Biol*

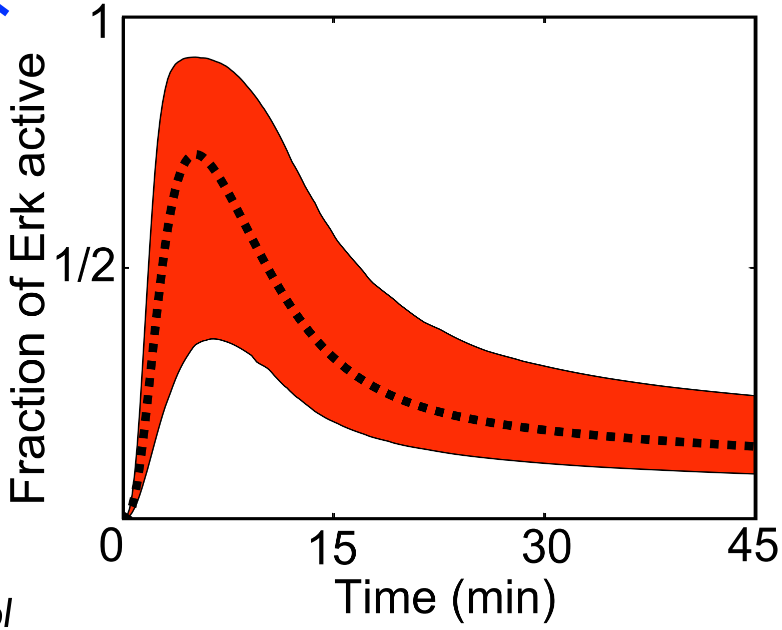
Sloppiness & uncertainties

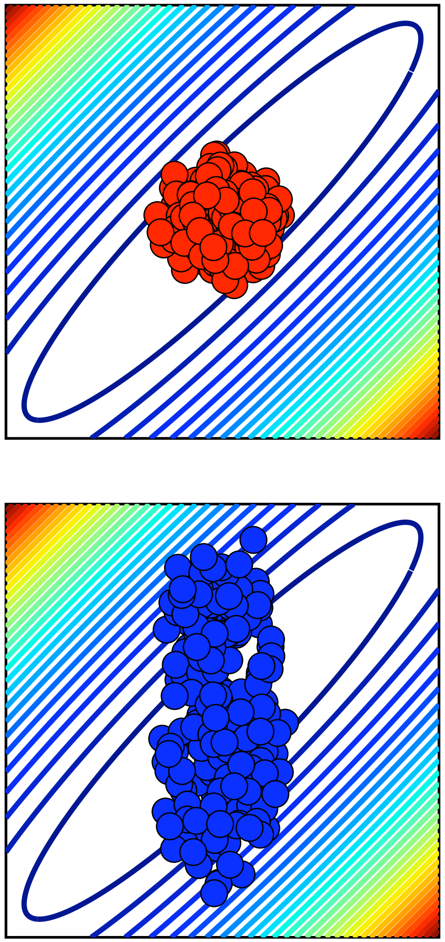
(All uncertainties by MCMC)

Parameters



Prediction





All measured

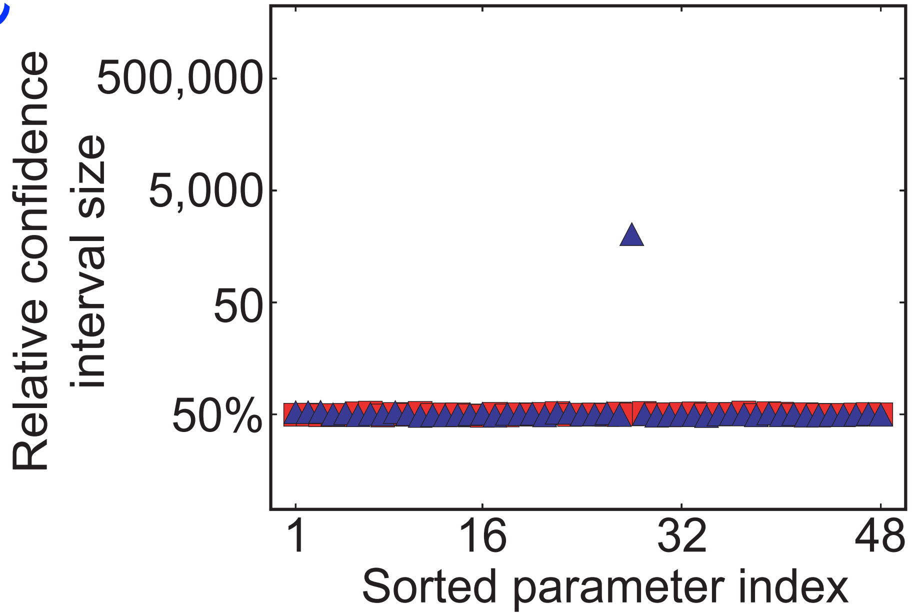
One unmeasured

Gutenkunst et al.
(2007) *PLoS Comput Biol*

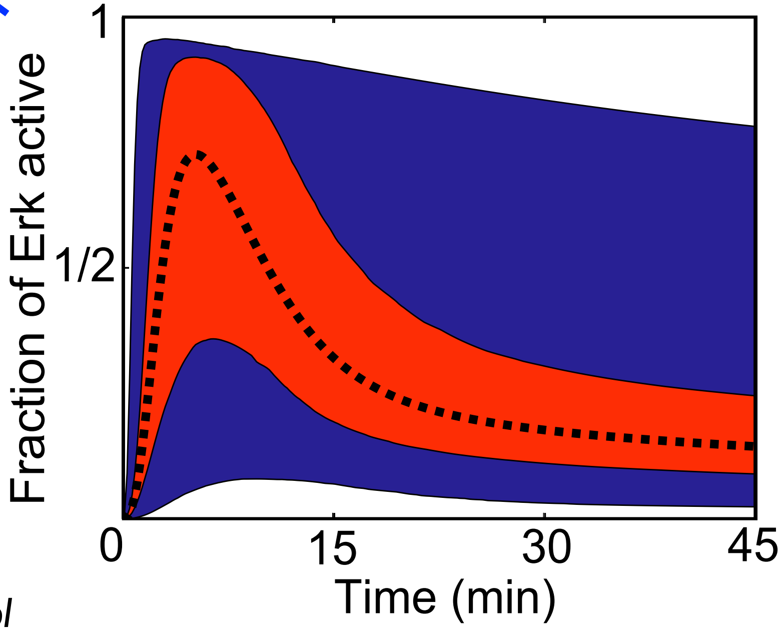
Sloppiness & uncertainties

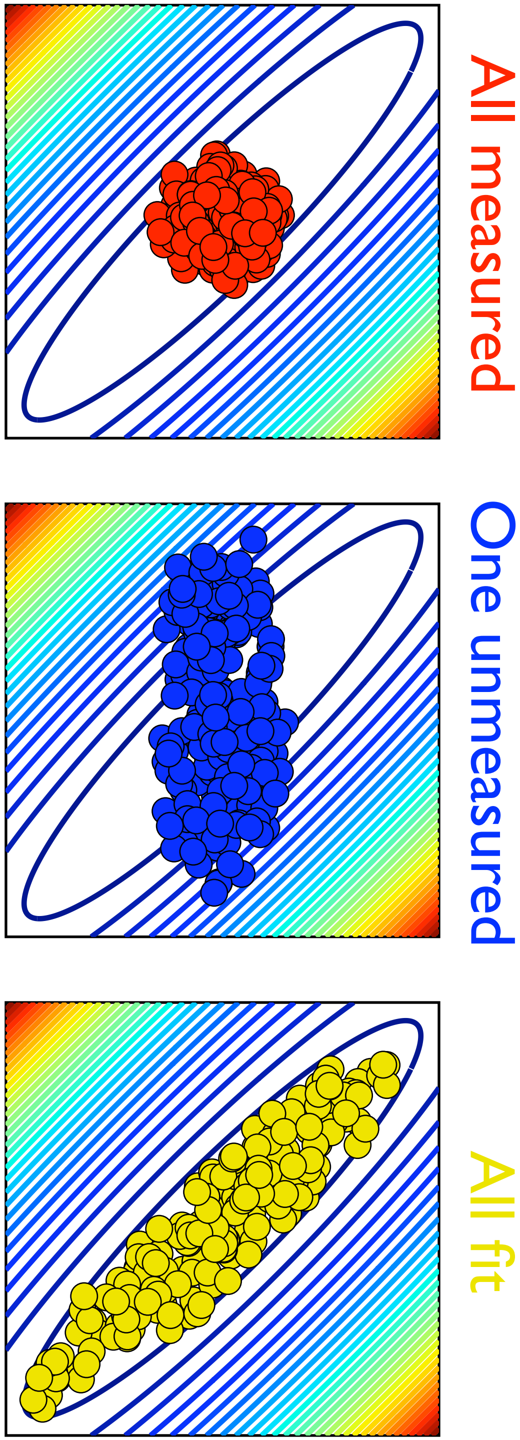
(All uncertainties by MCMC)

Parameters



Prediction



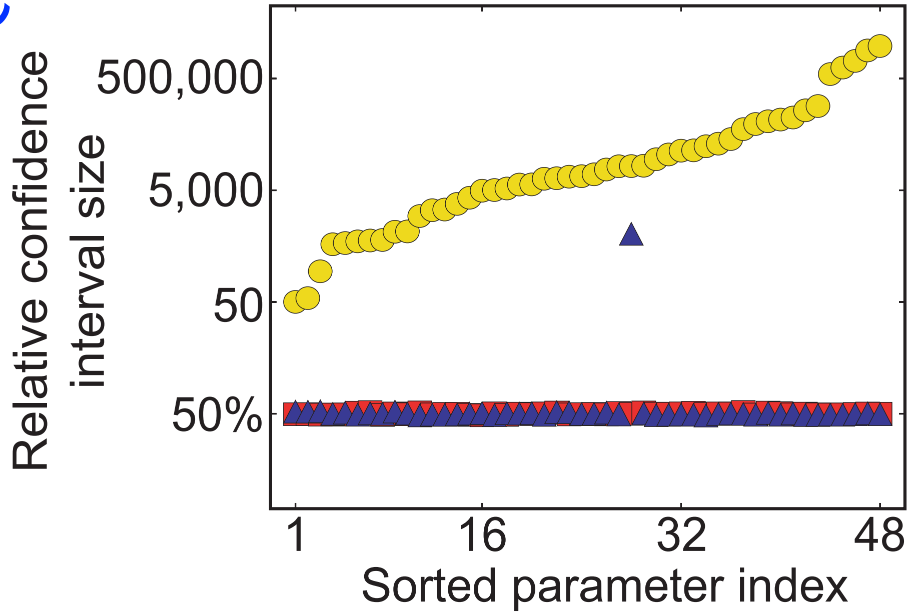


Gutenkunst et al.
(2007) *PLoS Comput Biol*

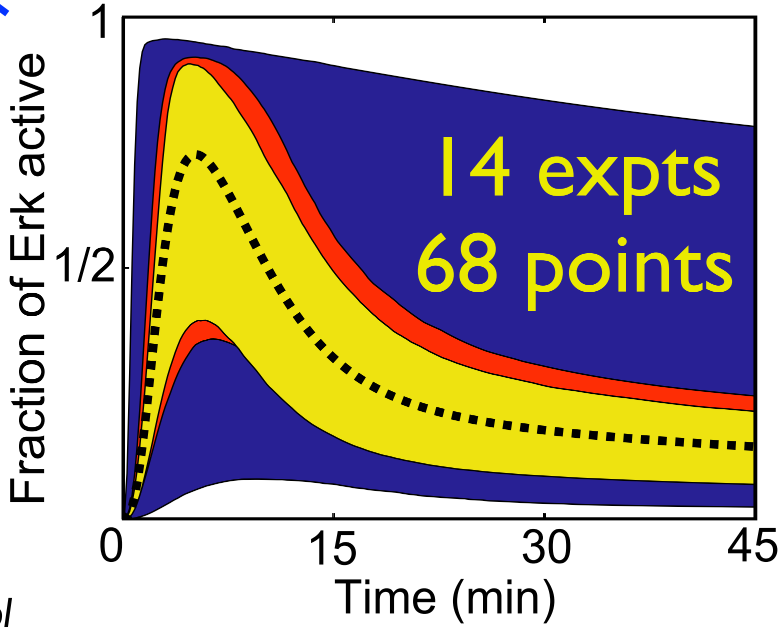
Sloppiness & uncertainties

(All uncertainties by MCMC)

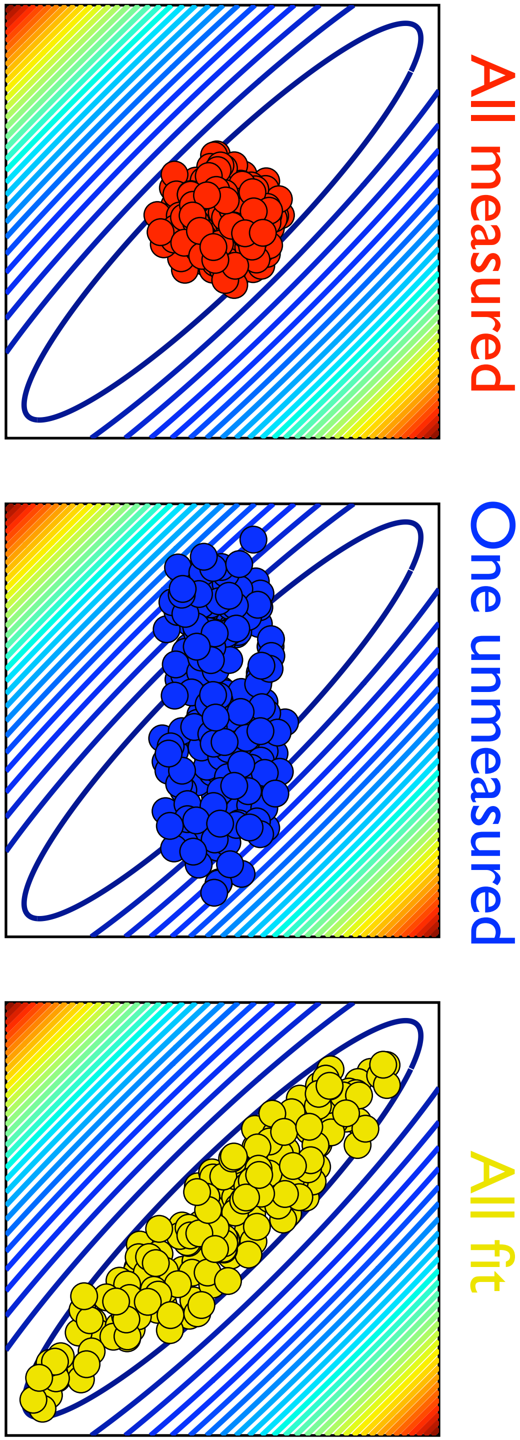
Parameters



Prediction



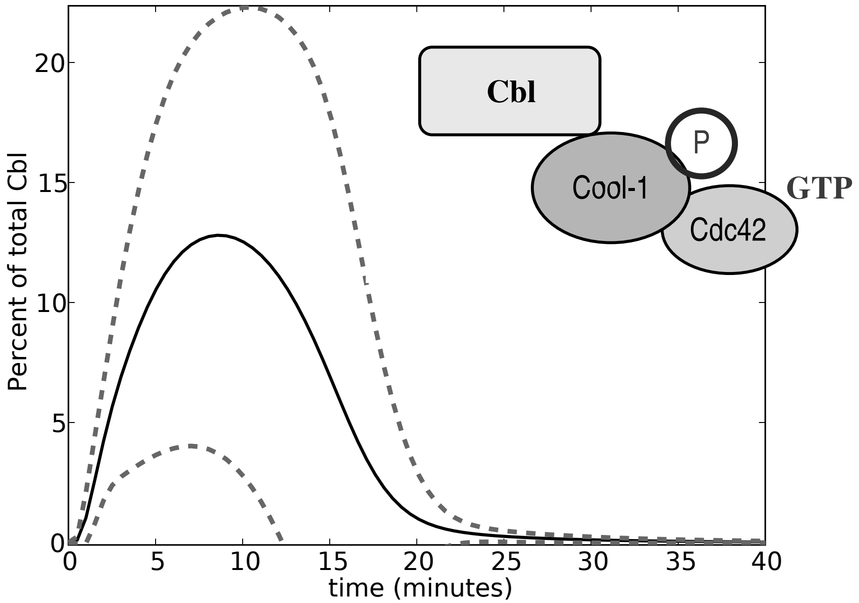
14 expts
68 points



Gutenkunst et al.
(2007) *PLoS Comput Biol*

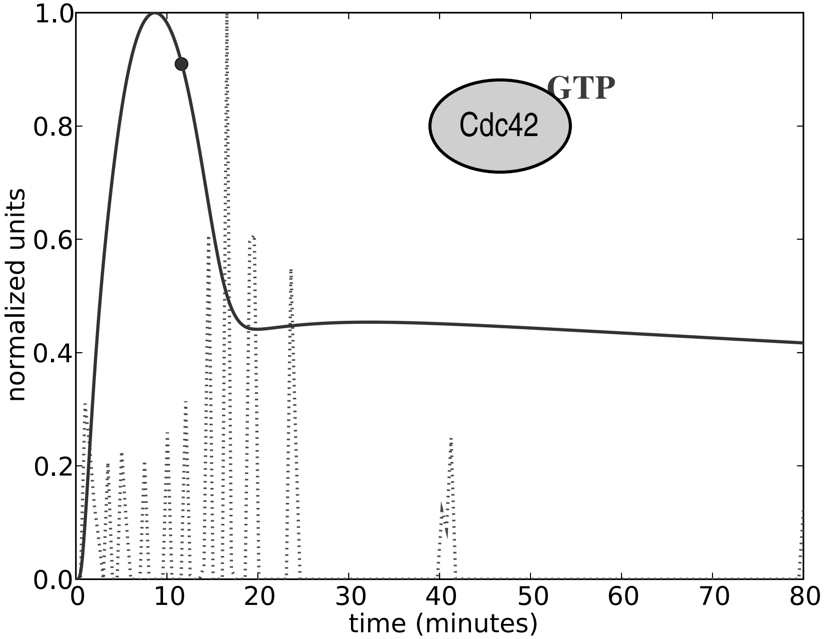
Refining predictions

Loose prediction



Optimized design for variance

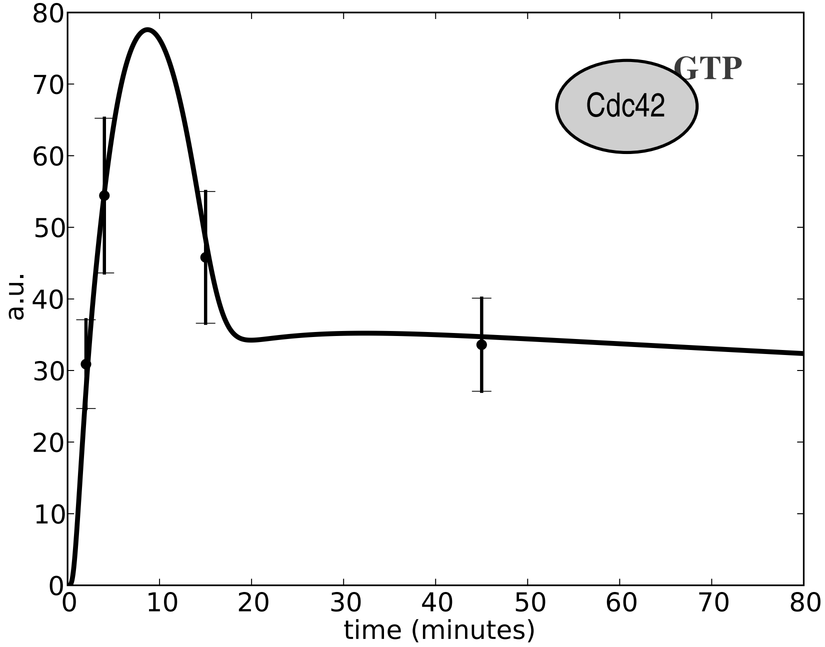
$$Var(y(t)) \sim \left.\frac{\partial\, y(t,\theta)}{\partial\theta}\right|_{\hat{\theta}} H^{-1} \left.\frac{\partial\, y(t,\theta)}{\partial\theta}\right|_{\hat{\theta}}$$



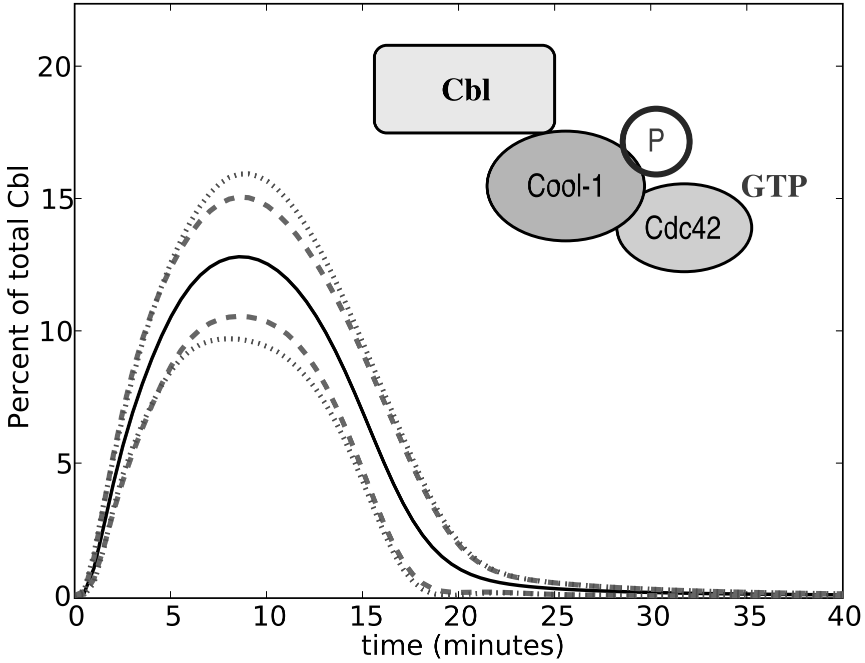
Casey et al.
(2007) *IET Sys Biol*

Design results

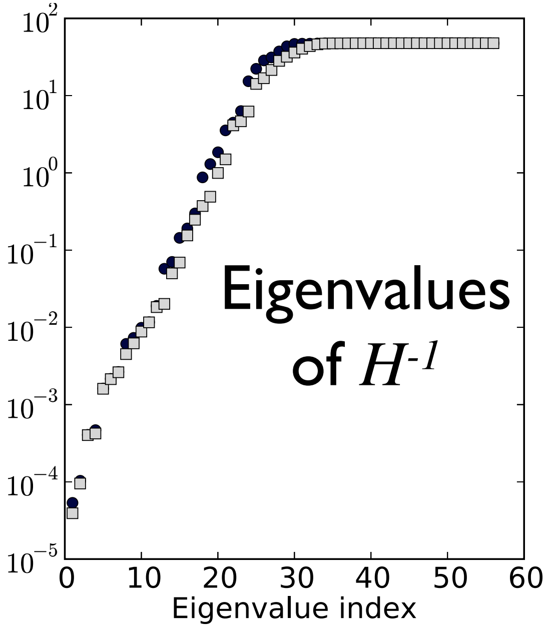
Experiment



Resulting tight prediction



No change in parameter uncertainties

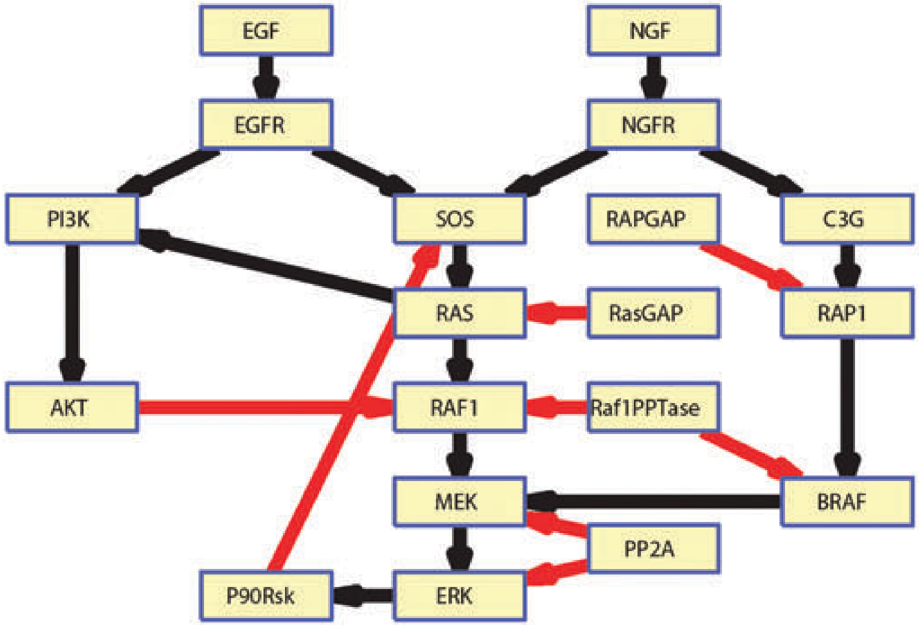


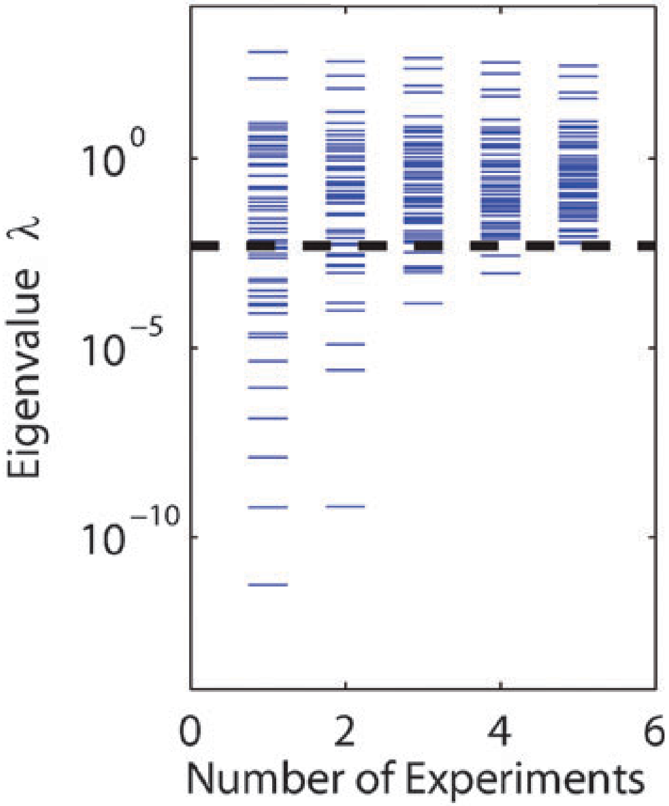
Eigenvalues of H^{-1}

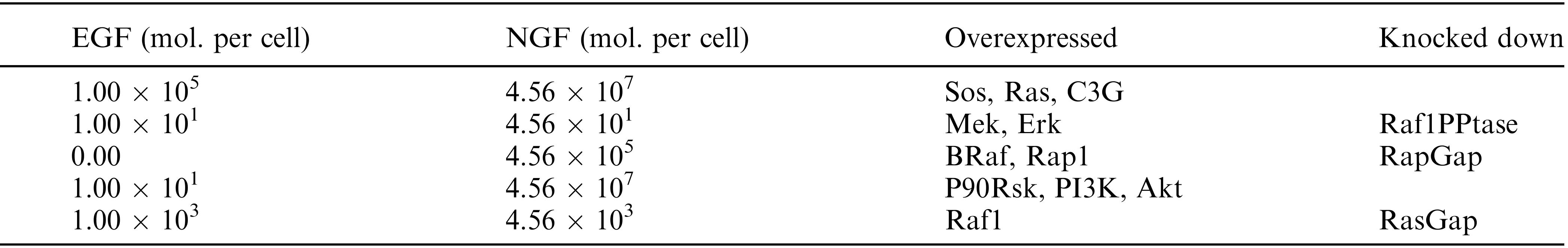
Casey et al. (2007) *IET Sys Biol*

More sophisticated expt design

Apgar et al. (2010) *Mol Biosyst*







EGF (mol. per cell)	NGF (mol. per cell)	Overexpressed	Knocked down
1.00×10^5	4.56×10^7	Sos, Ras, C3G	
1.00×10^1	4.56×10^1	Mek, Erk	Raf1PPtase
0.00	4.56×10^5	BRaf, Rap1	RapGap
1.00×10^1	4.56×10^7	P90Rsk, PI3K, Akt	
1.00×10^3	4.56×10^3	Raf1	RasGap

Conclusions

Parameter estimation ain't easy.

Toolbox should include a variety of optimization algorithms.

Sloppy parameter sensitivities appear to be universal.

Sloppiness implies focusing on predictions not parameters.

Experimental design is key to optimizing experiments

http://gutengroup.mcb.arizona.edu/publications/Mannakee2015.pdf

http://arxiv.org/abs/1501.07668

Join us!

Seeking a computationally skilled postdoc interested in evolutionary systems biology or population genomics

http://gutengroup.mcb.arizona.edu