Statistical model selection and prediction of systems' responses to exogenous perturbations
- or -
Making predictions with limited data

Bryan Daniels, Ilya Nemenman

The Goal: prediction, control

- By learning from available data, we want to predict the result of exogenous perturbations, with the goal of control
- Do we necessarily want the most detailed model possible?

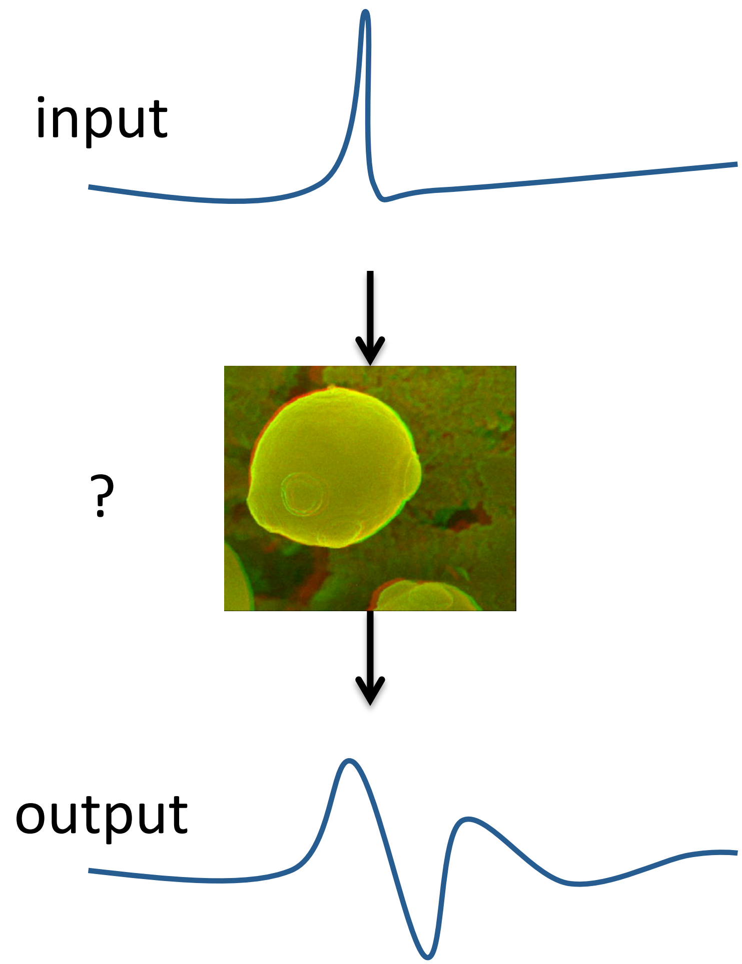
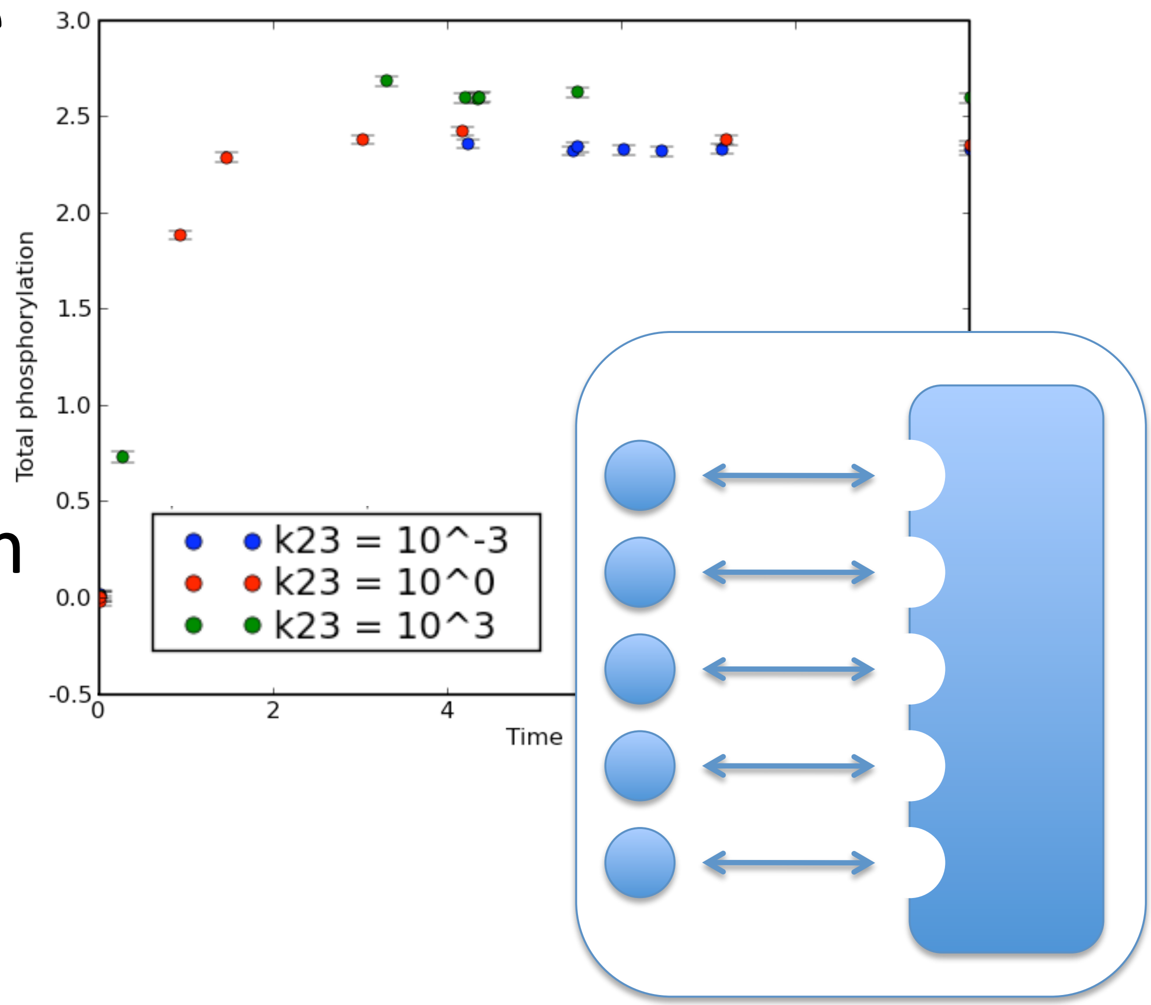


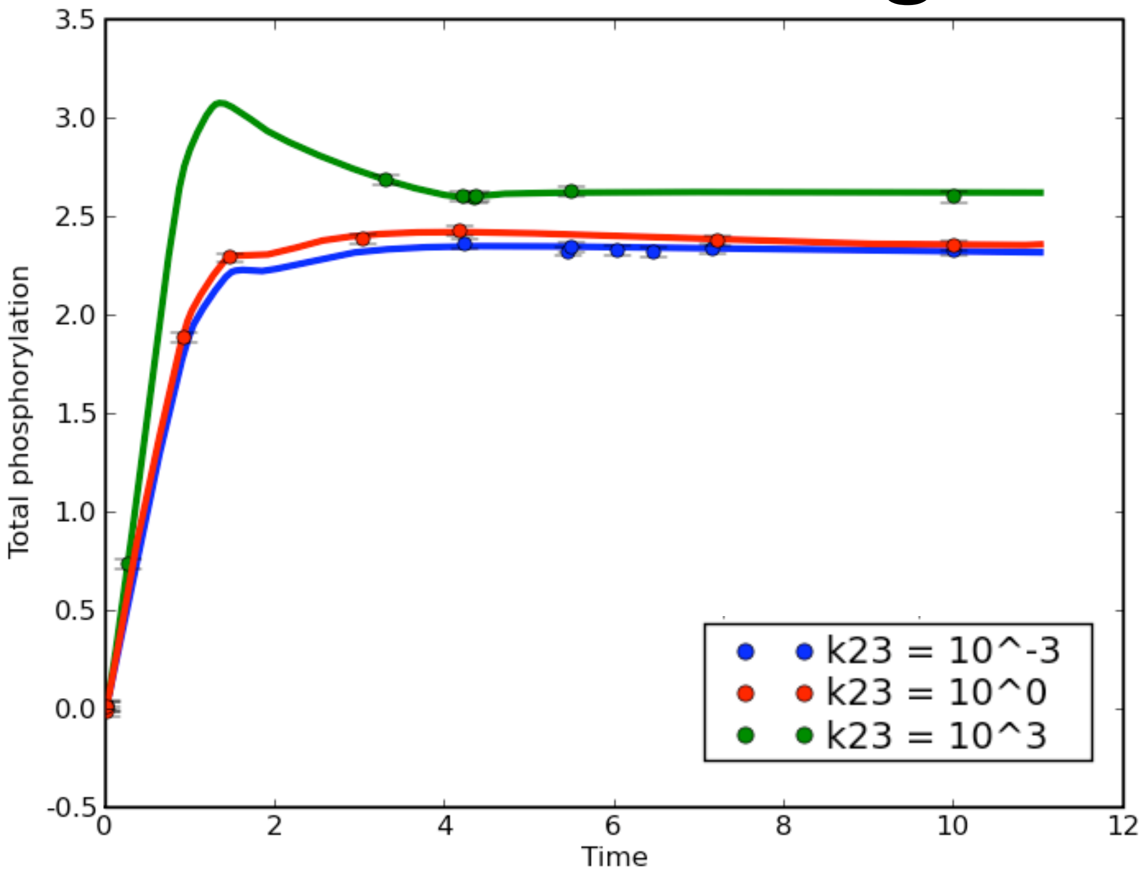
Image: http://www2.biomed.cas.cz/~benada/lem117/eng/stereo.htm

A foreboding example

- Suppose we are trying to fit experimental data with a model…
- Phosphorylation on 5 sites with independent MM rates

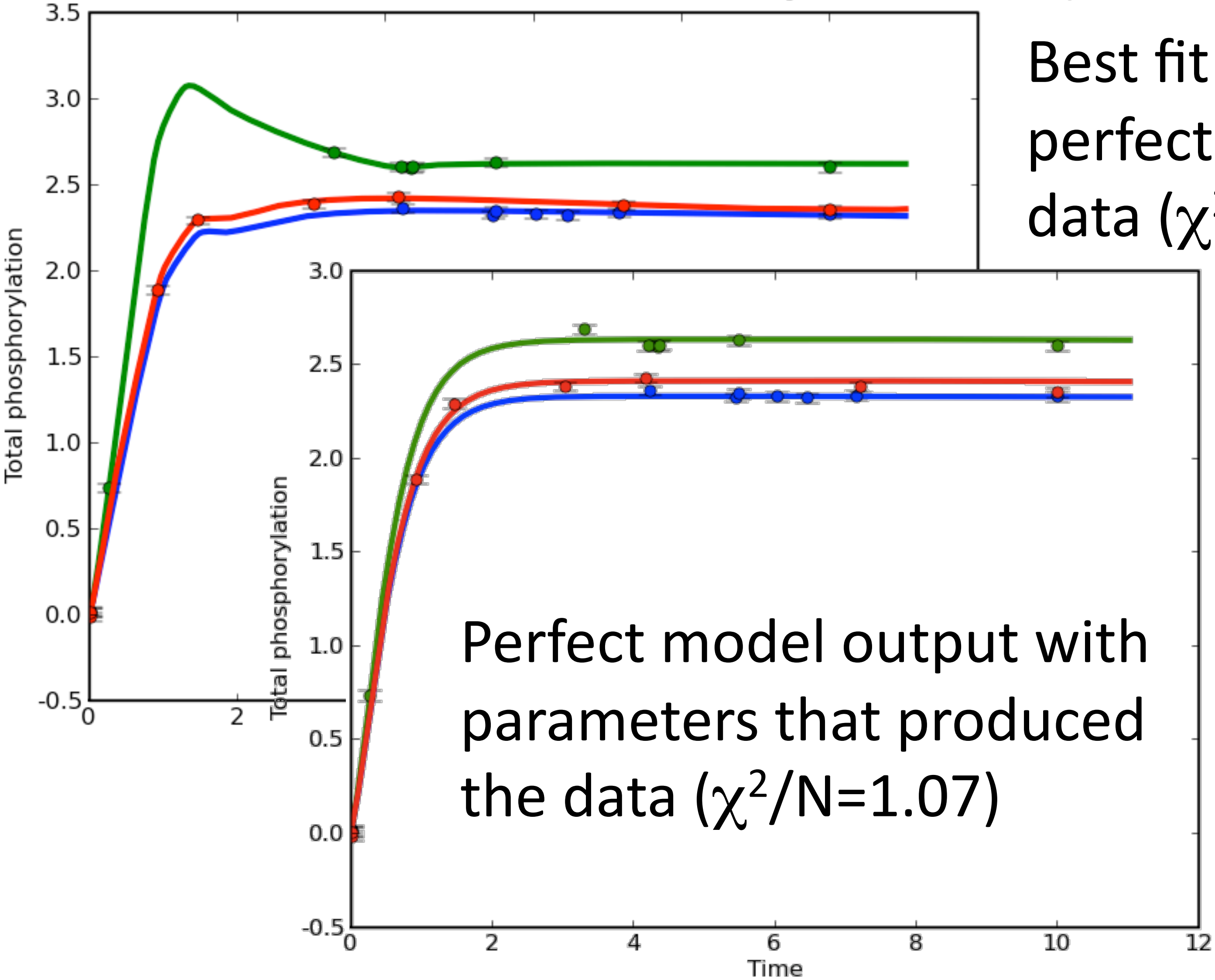


A foreboding example



Best fit of the perfect model to data ($\chi^2/N=0.2$)

A foreboding example



Best fit of the perfect model to data ($\chi^2/N=0.2$)

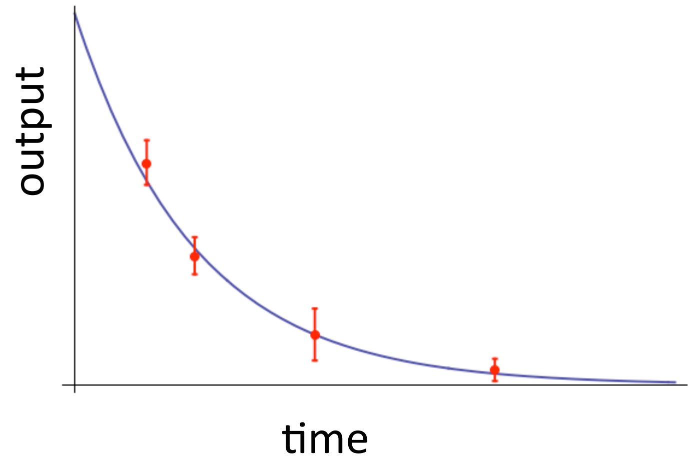
Perfect model output with parameters that produced the data ($\chi^2/N=1.07$)

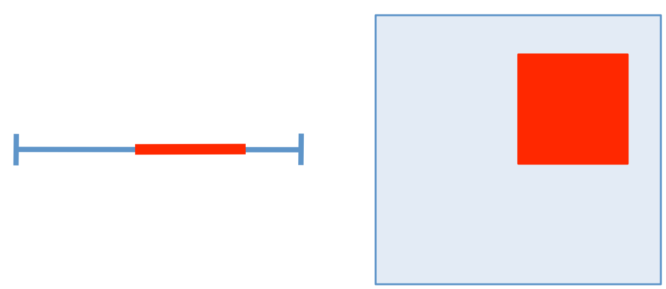
How to proceed

1) How are we supposed to know how good our predictions should be?
   – One approach: find *all* the parameter sets consistent with the data
   – Or use a simpler approximation: the Bayesian Information Criterion (BIC)
2) How can we make better predictions?
   – It is likely that a phenomenological model of lower complexity will produce better predictions

BIC : The idea

- The sum of two terms:
  - Maximum likelihood error
    - How well does the model fit the data?
  - Penalty for complexity
    - How much of parameter space adequately fits the data?





1D: $\propto 1/N^{1/2}$ 2D: $\propto 1/N$ ...

BIC : The derivation

Probability of a model M given the data

Integrate over unknown parameters α

$$P(M \mid \text{data}) = \int d^K\alpha \; P(M \mid \text{data}; \alpha) \; P(\alpha)$$

Gaussian errors $\Rightarrow$ χ^2 is sum of squared residuals

$$\begin{aligned} P\left(M \mid \text{data}; \alpha\right) &= \frac{P\left(M\right)}{P\left(\text{data}\right)} P\left(\text{data} \mid M(\alpha)\right) \\ &= \text{consts} \;\; P\left(\text{data} \mid M(\alpha)\right) \\ &= \text{consts} \;\; \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - M(t_i, \alpha)}{\sigma_i}\right)^2\right] \\ &= \text{consts} \;\; \exp\left[-\frac{1}{2}\chi^2(\alpha)\right] \end{aligned}$$

BIC : The derivation

$$\mathcal{H}_{ij} = \left.\frac{d\chi^2(\alpha)}{d\alpha_i d\alpha_j}\right|_{\alpha_{\text{best}}}$$

$$\begin{aligned} P(M \mid \text{data}) &= \text{consts} \int d^K\alpha \; P(\alpha) \; \exp\left[-\frac{1}{2}\chi^2(\alpha)\right] \\ &\approx \text{consts} \; \exp\left[-\frac{1}{2}\chi^2(\alpha_{\text{best}})\right] \sqrt{\frac{(2\pi)^K}{\det \mathcal{H}}} \end{aligned}$$

$$\boxed{\mathcal{L} \equiv -\log P(M \mid \text{data}) \approx \text{consts} + \frac{1}{2}\chi^2(\alpha_{\text{best}}) + \frac{1}{2}\sum_{\mu=1}^{K} \log \frac{\lambda_\mu}{2\pi} \qquad \text{Usual BIC}}$$

Log posterior probability

Least-squares “cost”

Higher penalty for:

1) More parameters
2) Larger eigenvalues of $\mathcal{H}$

BIC : The derivation

$$P(M \mid \text{data}) = \ \text{consts} \int d^K\alpha \ P(\alpha) \ \exp\left[-\frac{1}{2}\chi^2(\alpha)\right]$$

$$\approx \ \text{consts} \ \exp\left[-\frac{1}{2}\chi^2(\alpha_{\text{best}})\right] \sqrt{\frac{(2\pi)^K}{\det \mathcal{H}}}$$

$$\mathcal{H}_{ij} = \left.\frac{d\chi^2(\alpha)}{d\alpha_i d\alpha_j}\right|_{\alpha_{\text{best}}}$$

$$\mathcal{L} \equiv -\log P(M \mid \text{data}) \approx \ \text{consts} + \frac{1}{2}\chi^2(\alpha_{\text{best}}) + \frac{1}{2}\sum_{\mu=1}^{K} \log \frac{\lambda_\mu}{2\pi}$$

Usual BIC

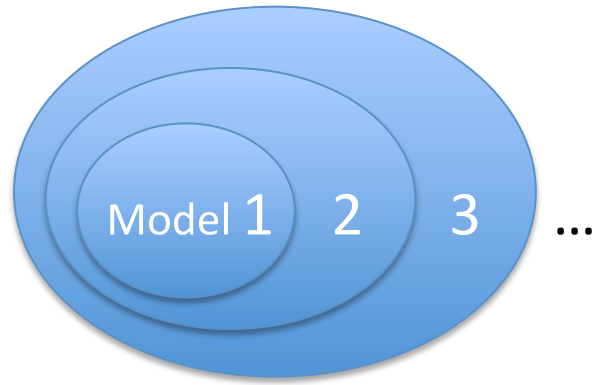
$$\boxed{\mathcal{L} \equiv -\log P(M \mid \text{data}) \approx \ \text{consts} + \frac{1}{2}\chi^2(\alpha_{\text{best}}) + \frac{1}{2}\sum_{\lambda_\mu > \lambda_c} \log \frac{\lambda_\mu}{2\pi}}$$

Modified BIC

Don’t include tiny ‘sloppy’ eigenvalues that are cut off by priors

Model hierarchy

- Next: systematically build up the complexity of a phenomenological model
- We need a hierarchy that is:
  1. Nested
  2. One-dimensional
  3. Guaranteed to eventually fit any data arbitrarily well



We know a single model will win, and we won't have to backtrack. [1]

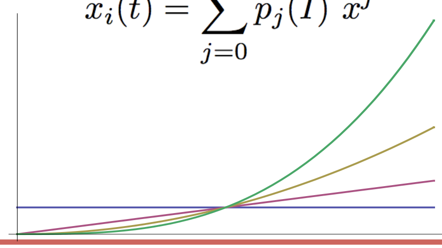
[1] Nemenman I. Fluctuation-Dissipation Theorem and Models of Learning. *Neural Comp* **17**, 2006 (2005)

Model hierarchy

- If we have little knowledge of the microscopic kinetics, what type of model should we use?

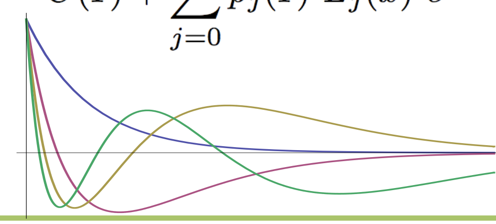
Polynomials

$$x_i(t) = \sum_{j=0}^{J} p_j(I)\ x^j$$



Laguerre polynomials

$$x_i(t) = C(I) + \sum_{j=0}^{J} p_j(I)\ L_j(x)\ e^{-t/\alpha(I)}$$

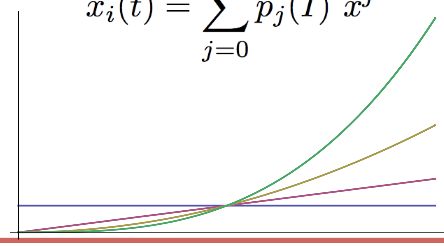


Model hierarchy

- If we have little knowledge of the microscopic kinetics, what type of model should we use?

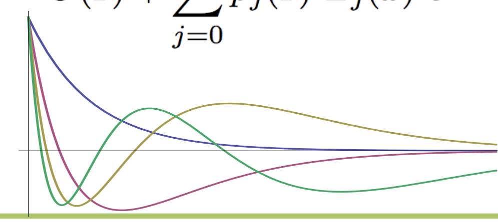
Polynomials

$$x_i(t) = \sum_{j=0}^{J} p_j(I)\ x^j$$



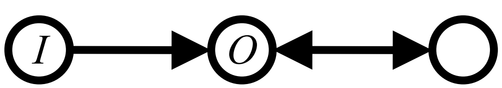
Laguerre polynomials

$$x_i(t) = C(I) + \sum_{j=0}^{J} p_j(I)\ L_j(x)\ e^{-t/\alpha(I)}$$



S-system power-law networks

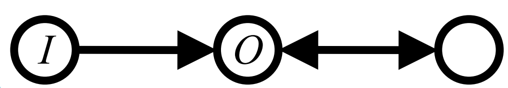
$$\frac{dx_i}{dt} = \delta_i \left(\prod_{j=1}^{J+K} x_j^{g_{ij}} - \gamma_i \prod_{j=1}^{J+K} x_j^{h_{ij}} \right)$$

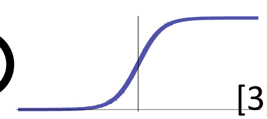


[2]

Sigmoidal networks

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





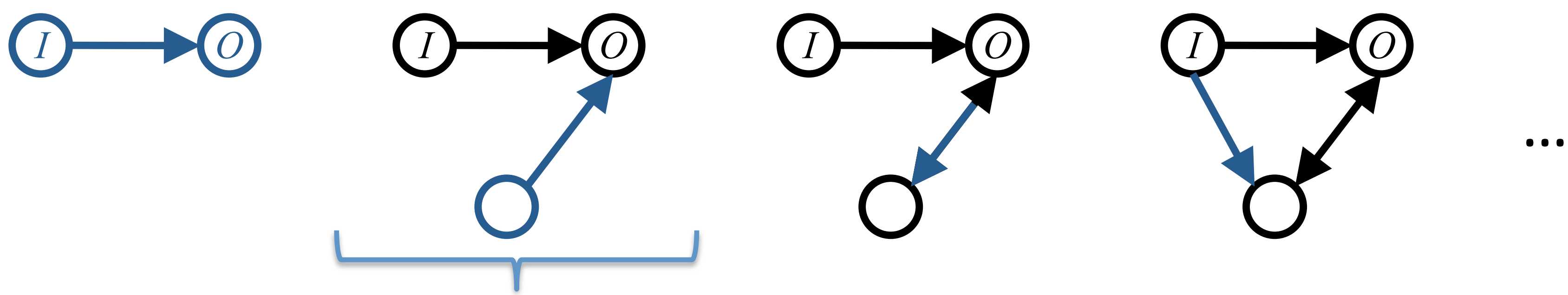
[3]

[2] Savageau, MA, Voit EO. *Math Biosci* **87**, 83 (1987). [3] Beer, RD. *Neural Comp* **18**, 3009 (2006).

Model hierarchy

- For network models, we need a way of "turning on" both parameters and topology

7 14 16 18



...

10

$$\frac{dx_1}{dt} = \delta_1 \left(x_I^{g_{10}} x_1^{g_{11}} x_2^{g_{12}} - \gamma_1 x_I^{h_{10}} x_1^{h_{11}} \right)$$

$$\frac{dx_2}{dt} = x_2^{g_{22}} - 1$$

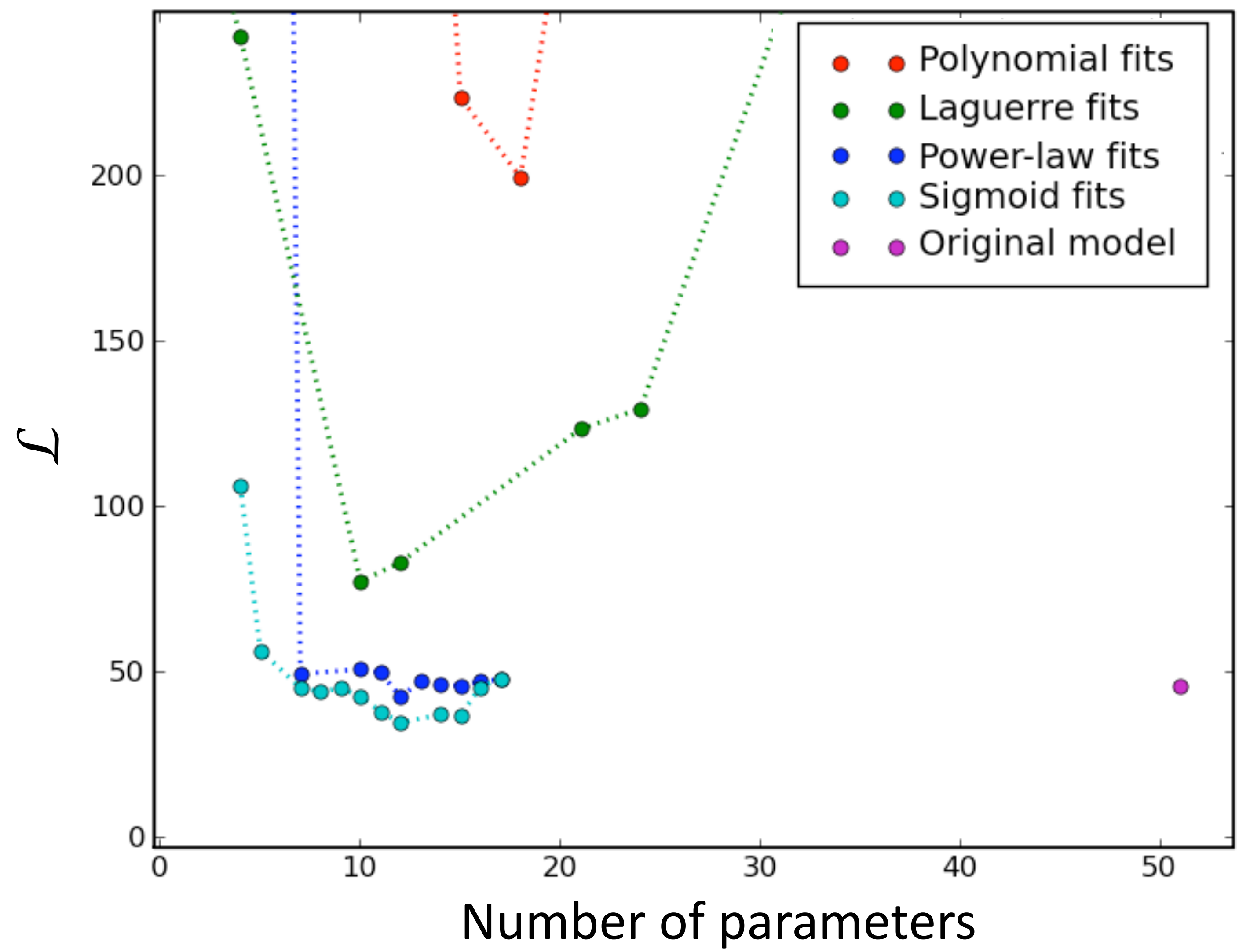
11

$$\frac{dx_1}{dt} = \delta_1 \left(x_I^{g_{10}} x_1^{g_{11}} x_2^{g_{12}} - \gamma_1 x_I^{h_{10}} x_1^{h_{11}} x_2^{h_{12}} \right)$$

$$\frac{dx_2}{dt} = x_2^{g_{22}} - 1$$

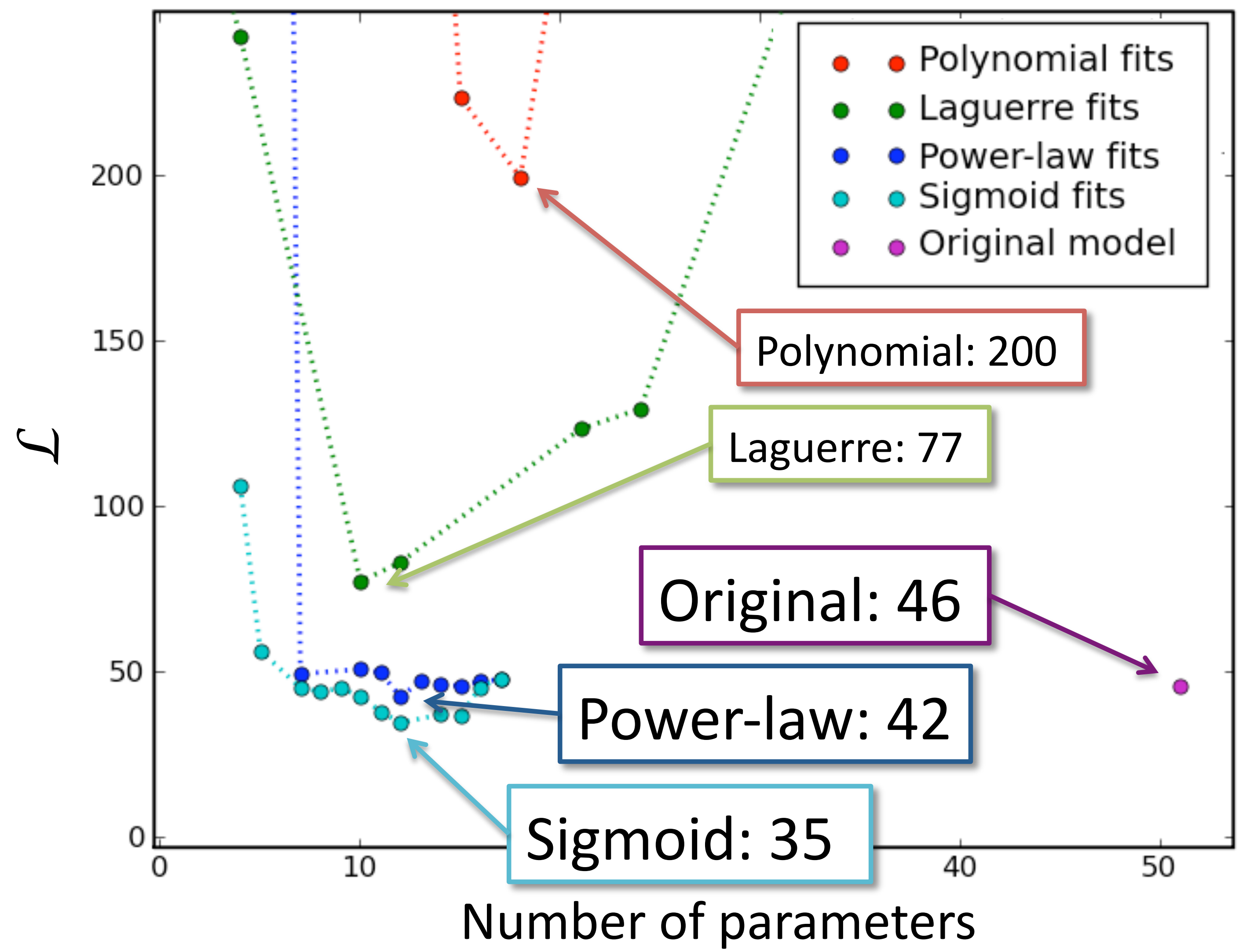
...

Results



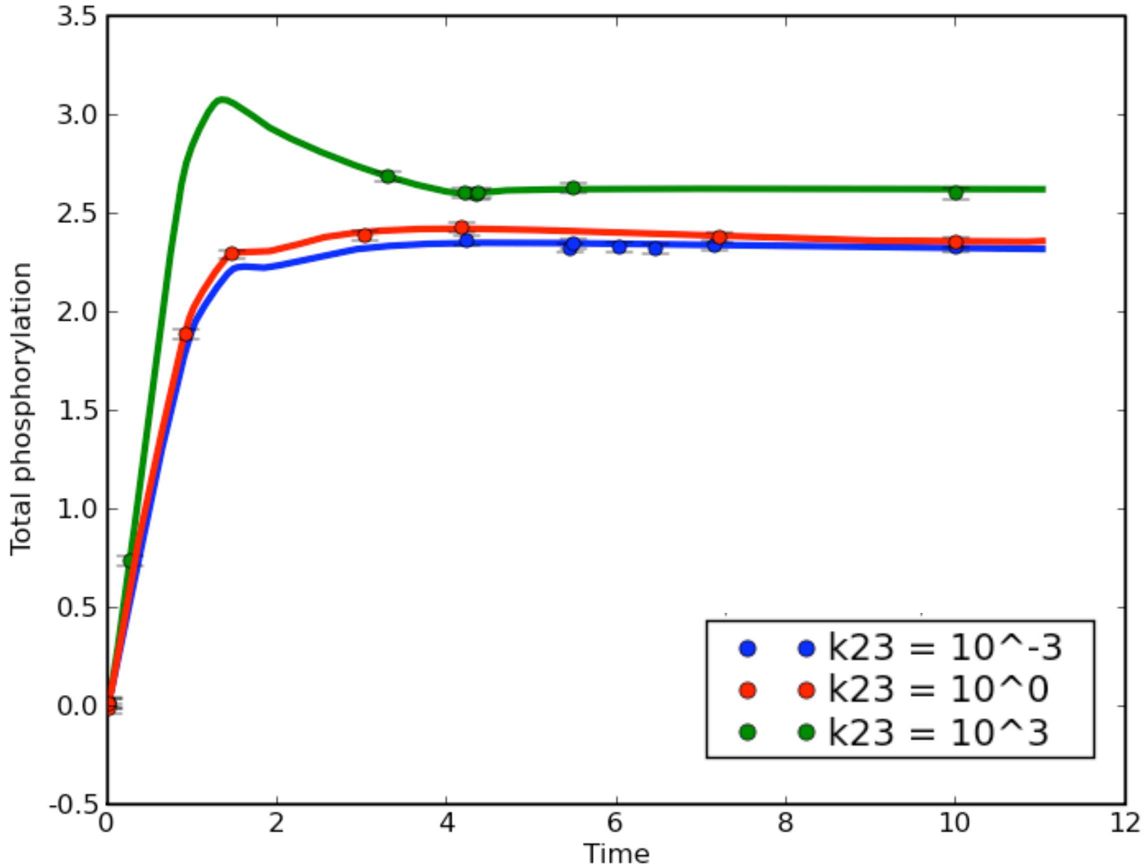
Seung HS, Sompolinsky H, Tishby N. Statistical mechanics of learning from examples. *Phys Rev A* **45**, 6056 (1992).

Results



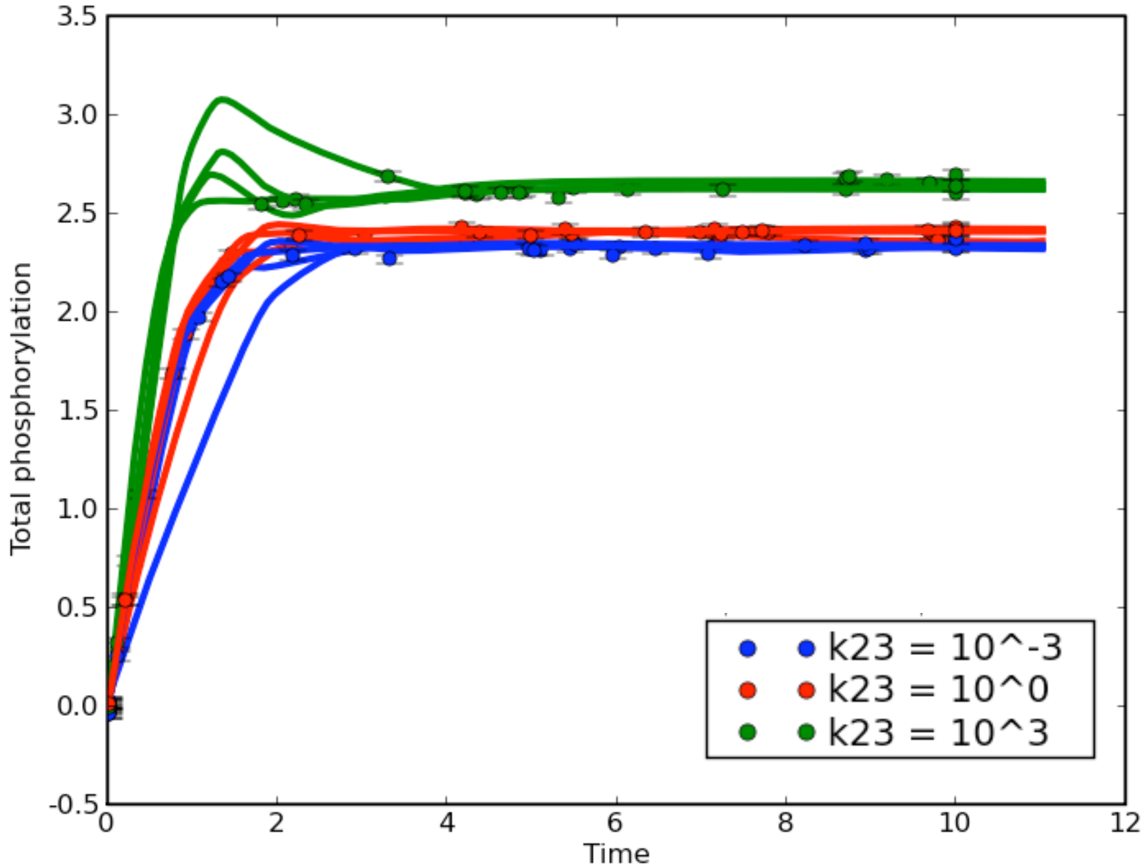
Seung HS, Sompolinsky H, Tishby N. Statistical mechanics of learning from examples. *Phys Rev A* **45**, 6056 (1992).

Results: fits to data



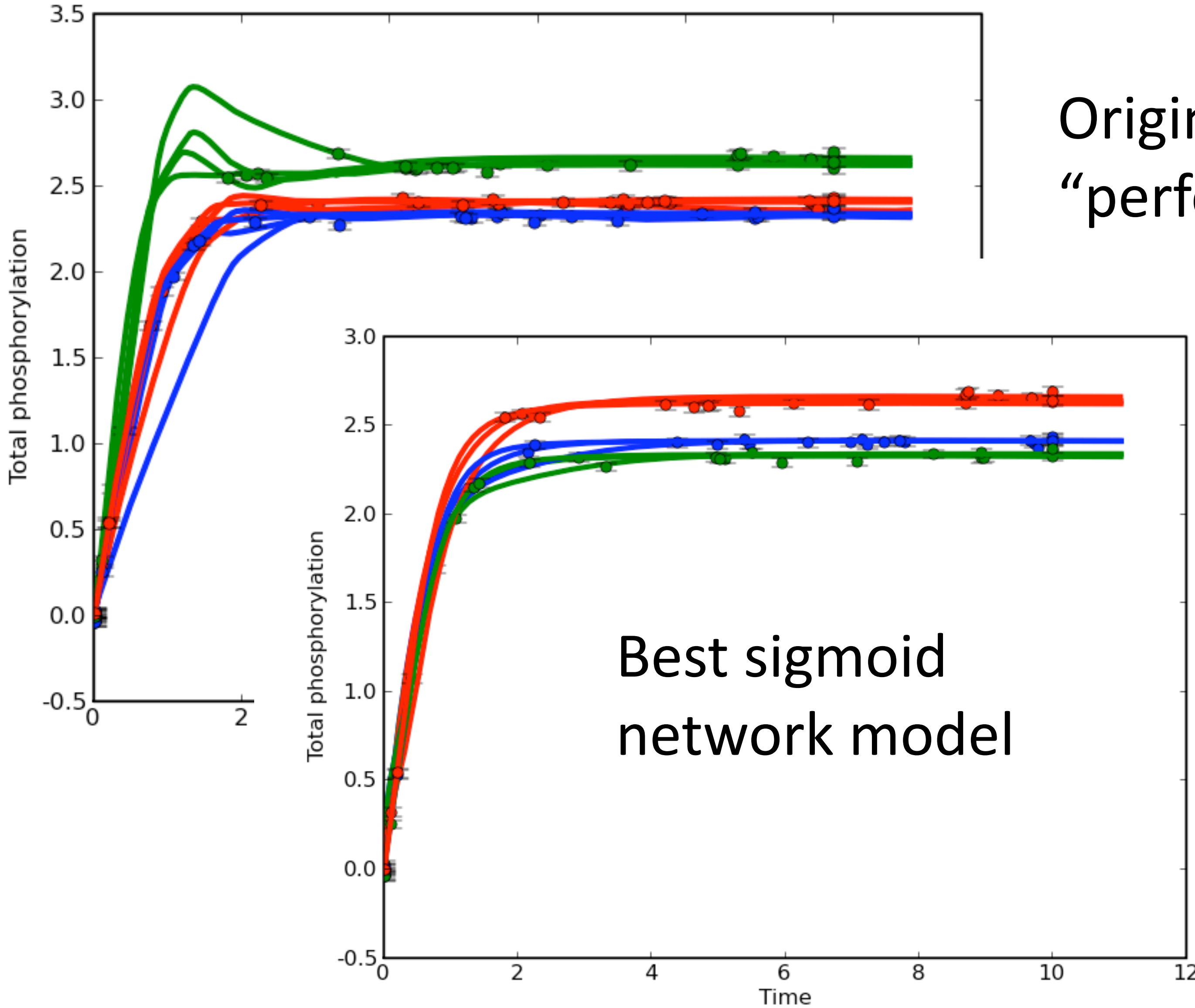
Original, “perfect” model

Results: fits to data



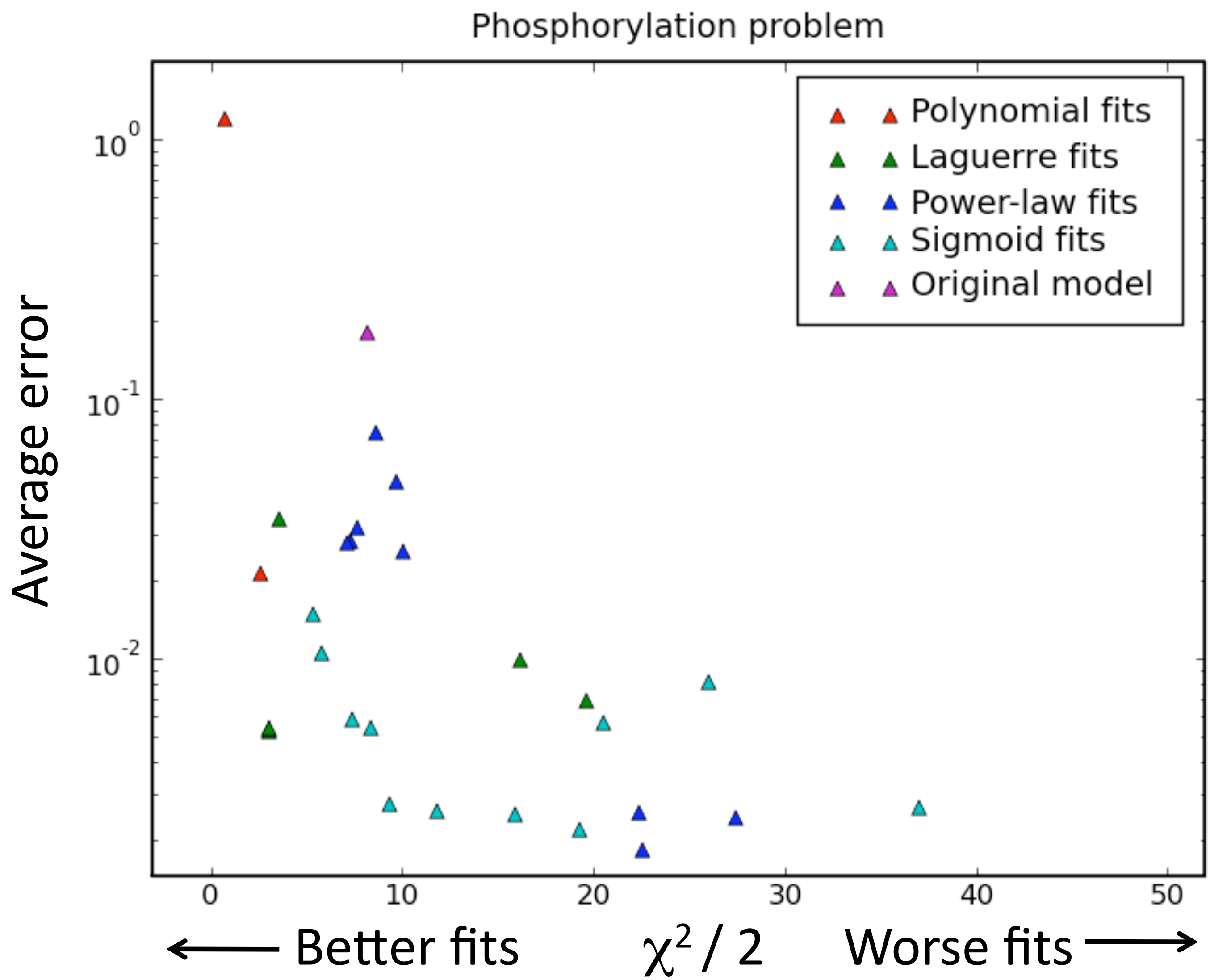
Original,
"perfect" model

Results: fits to data

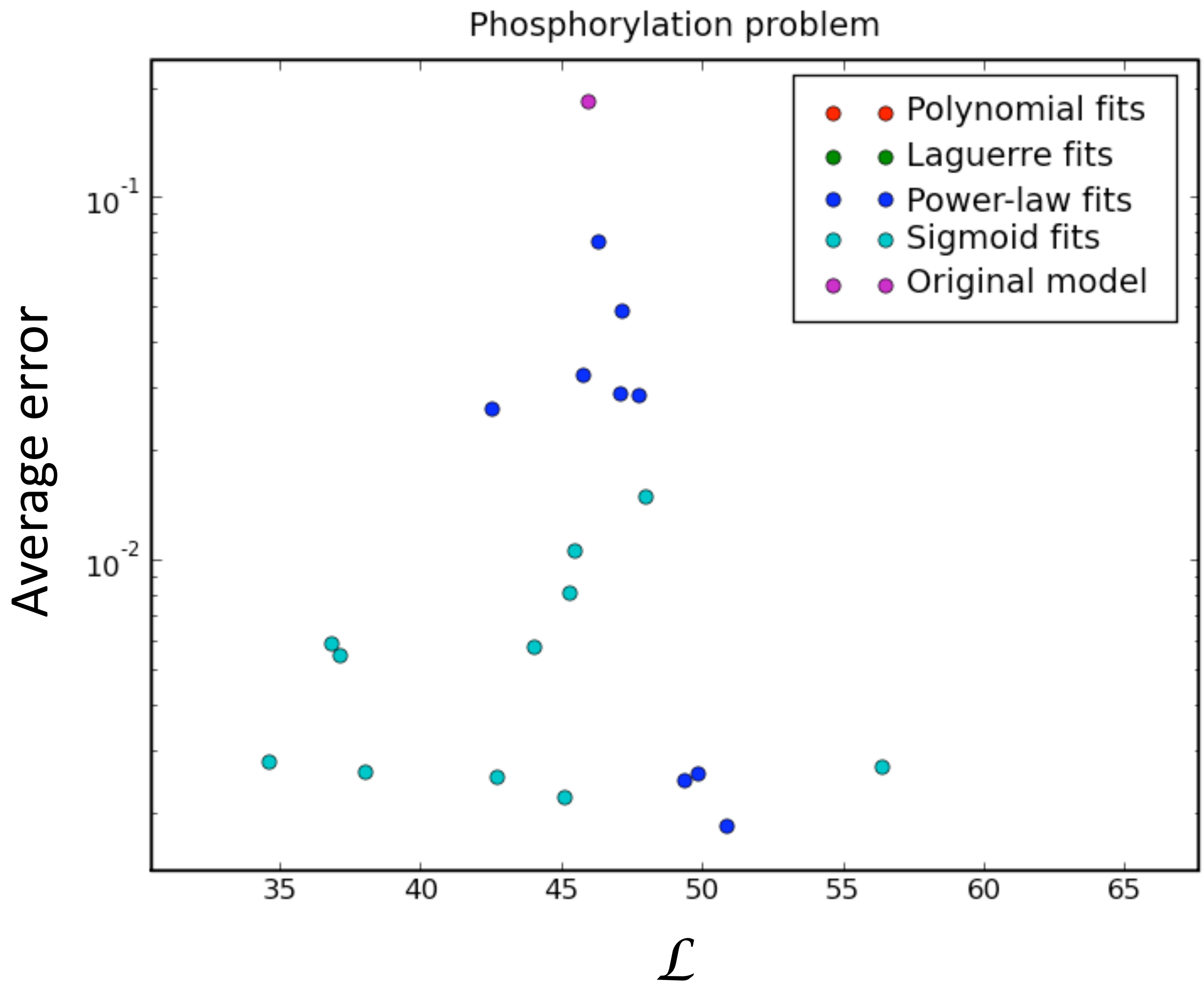


Original, “perfect” model

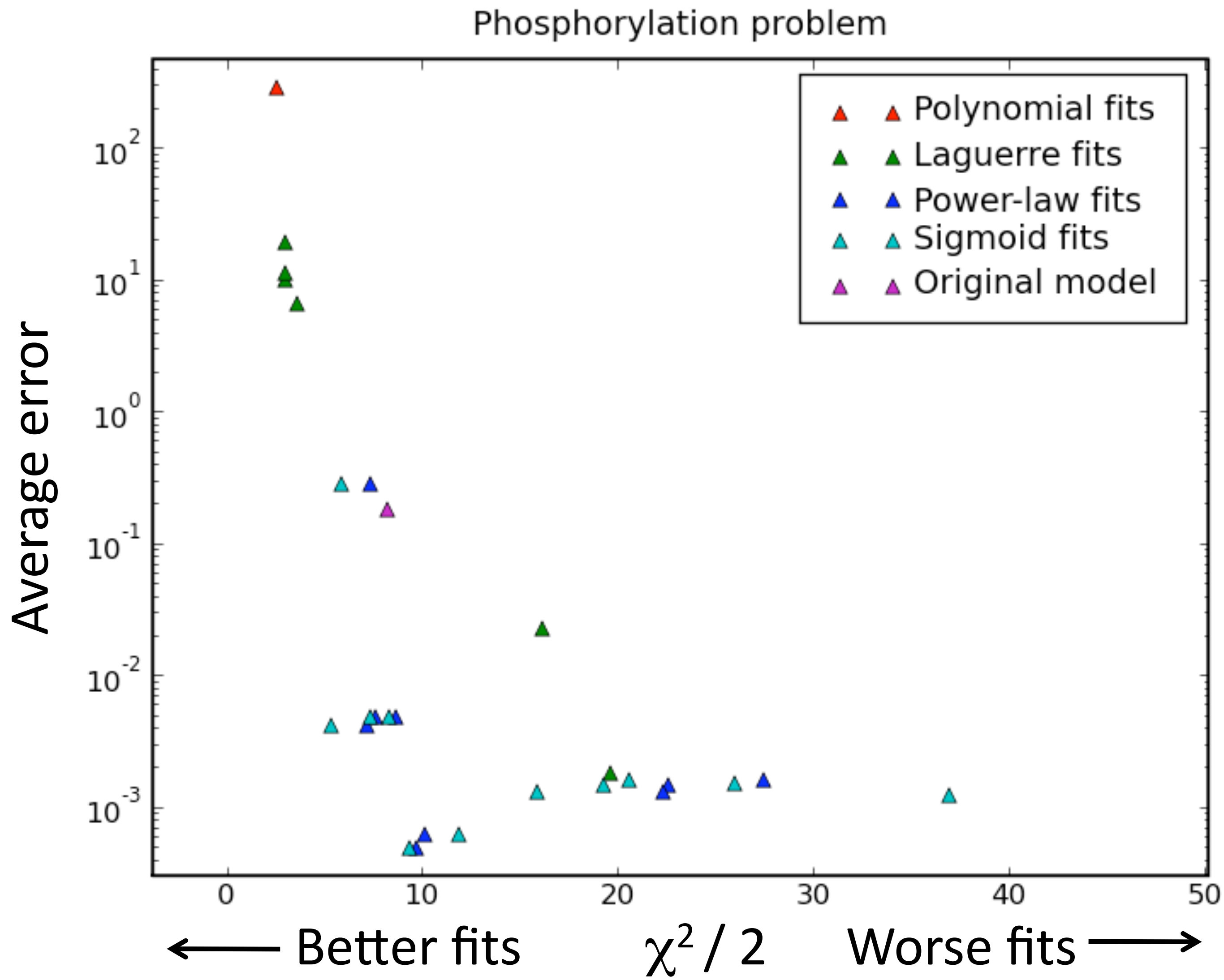
Results, interpolation



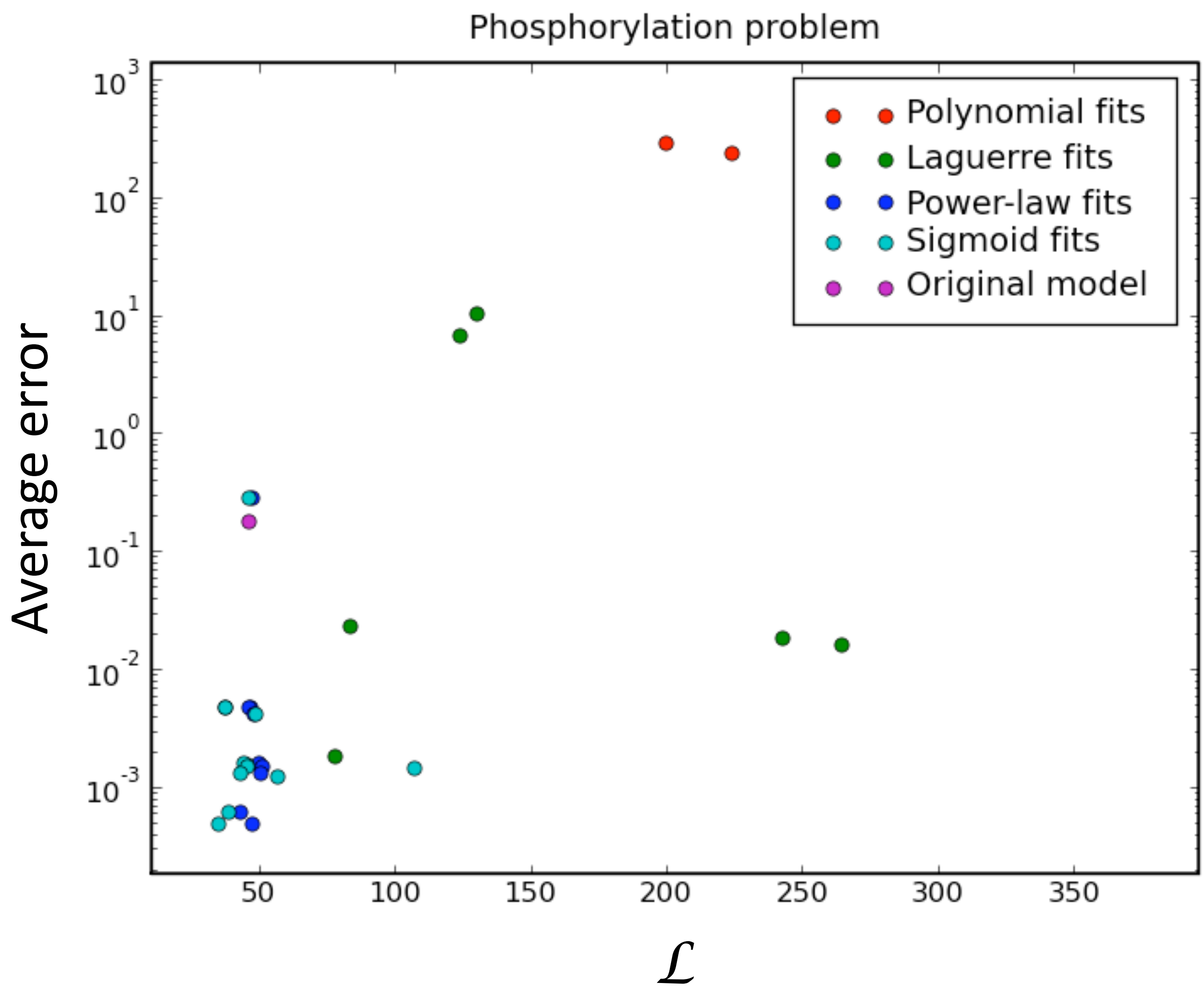
Results, interpolation



Results, extrapolation



Results, extrapolation



Conclusions

- Fitting complex models to limited data is dangerous, and may produce worse predictions
- BIC gives a useful (and theoretically defensible) measure that rewards good fits but penalizes overfitting
- Including more information about the underlying system (while avoiding overfitting) produces better predictions