Analyzing Simulation Data Using Bayes' Theorem

$$P(H|DI) = \frac{P(D|HI)\,P(H|I)}{P(D|I)}$$

An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.Phil. Trans. January 1, 1763 53:370-418; doi:10.1098/rstl.1763.0053

David M. Rogers
Dr. Thomas Beck Lab
University of Cincinnati

Questions

- How do we use inference on our simulation data?
- Can we use it to get solvation free energies from atomic simulations?

$$\beta \mu_{\alpha}^{\text{ex}} = \ln \frac{[\alpha]_{\text{aq.}}}{[\alpha]_{\text{gas}}}$$

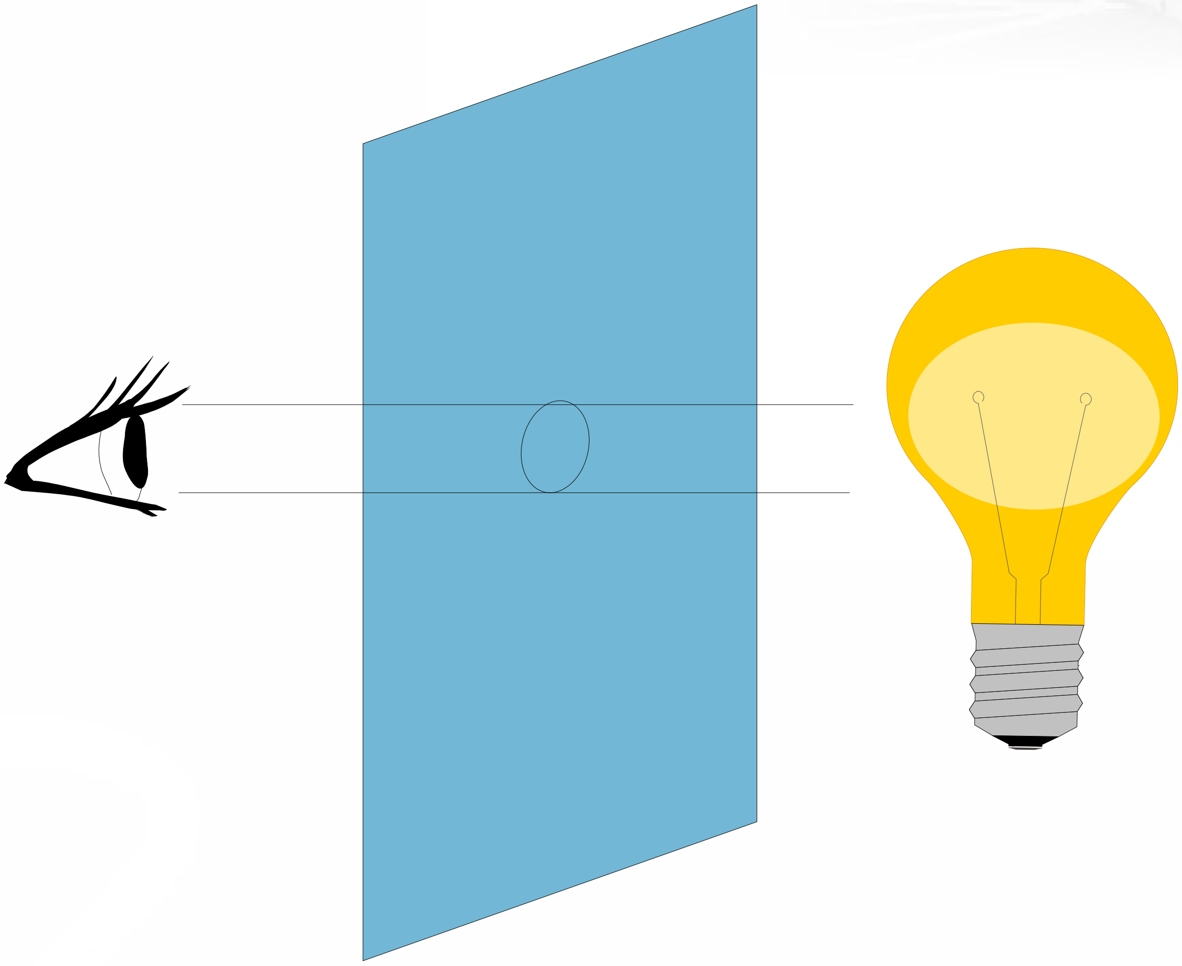
  - solute partitioning and PMFs
- Mesoscale simulation parameters from atomic simulations?
  - more efficient dynamics (multiscale)

Talk Outline

- Inference for real answers,

  experimental outcomes vs. probability
- Re-consider the original problem of Bayes, 245 years later
  - to calculate the likelihood of a molecule dissolving in water using stat. mech.
- Use inference to build a mesoscale simulation model
  - by calculating the probability of a dynamical equation.

Laboratory Inference: World's smallest spectrophotometer



Laboratory Inference: World's smallest spectrophotometer

Problem definition: find the concentration of solute α in the nanoscale volume.

1. Beer's law $A = \epsilon b N$
2. We are lead immediately to the “frequentist” picture, divide the concentration by the absorbtivity constant.
3. Remember that for small volumes, N can fluctuate and re-do the experiment.
4. If we only have samples from a finite amount of time, the concentration must be inferred from the measurements!
5. This is exactly the situation we face in making conclusions from a (necessarily) limited amount of simulation data.

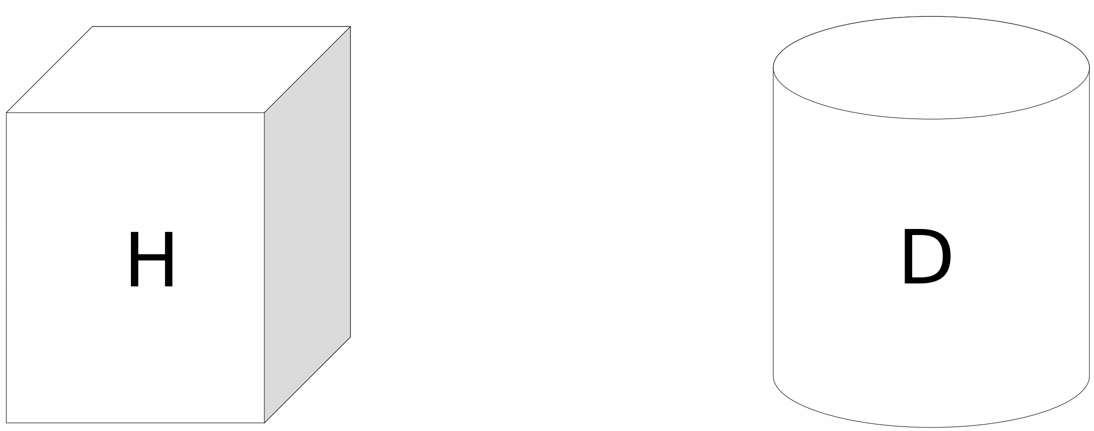
How does Bayes' Theorem Help?

H|I Prior Probability

D|HI Likelihood

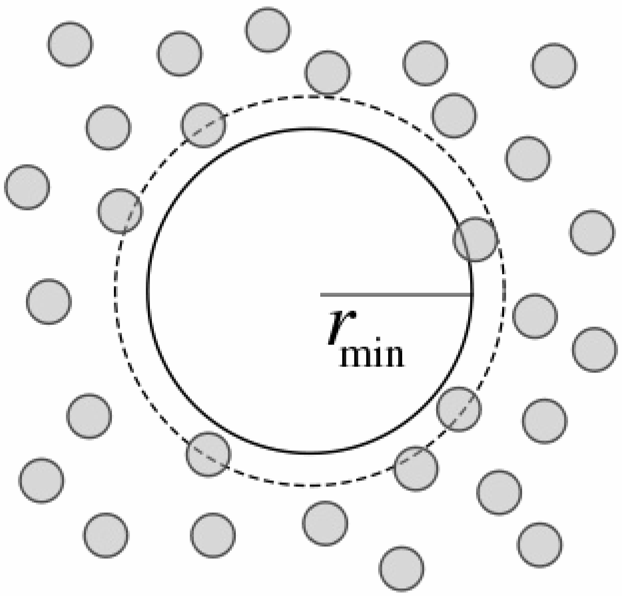
H|DI Posterior Probability

$$P(H|DI) = \frac{P(D|HI)\,P(H|I)}{P(D|I)}$$



Binomial Distribution

N_{tot}	$r_{min} > \lambda$	$r_{min} < \lambda$
50	x=17	33

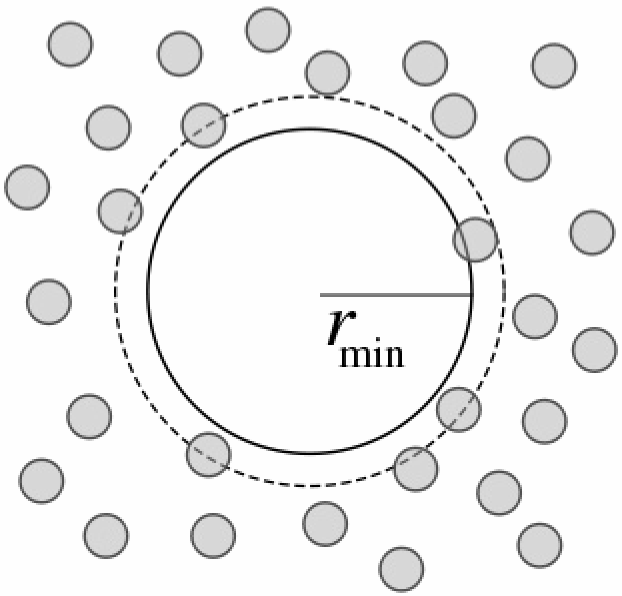


$$\beta \mu_{HS}^{ex}(\lambda) = -\ln p$$

Rogers and Beck, *J. Chem. Phys.* **129**:134505, 2008.

Binomial Distribution

N_{tot}	$r_{min} > \lambda$	$r_{min} < \lambda$
50	x=17	33



$$\beta \mu_{HS}^{ex}(\lambda) = -\ln p$$

Prior Probability

$$P(p|I) \propto p^{-1}(1-p)^{-1}$$

Likelihood

$$P(x|p\,I) = \binom{N}{x} p^{x}(1-p)^{N-x}$$

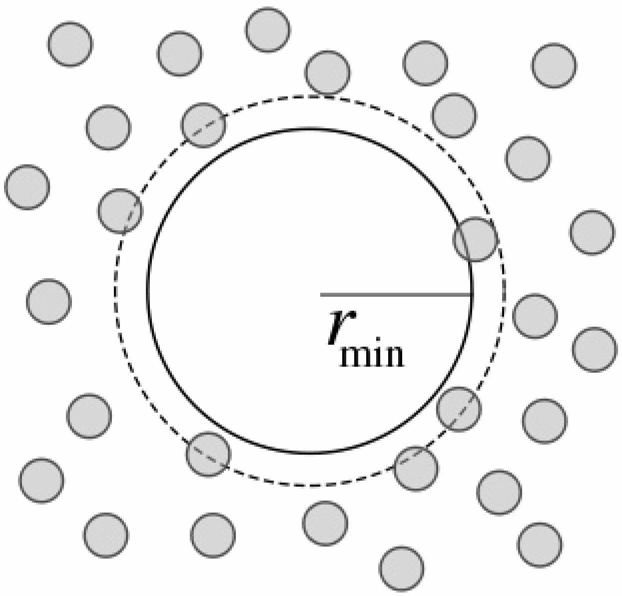
Posterior Probability

Beta(x,N-x)

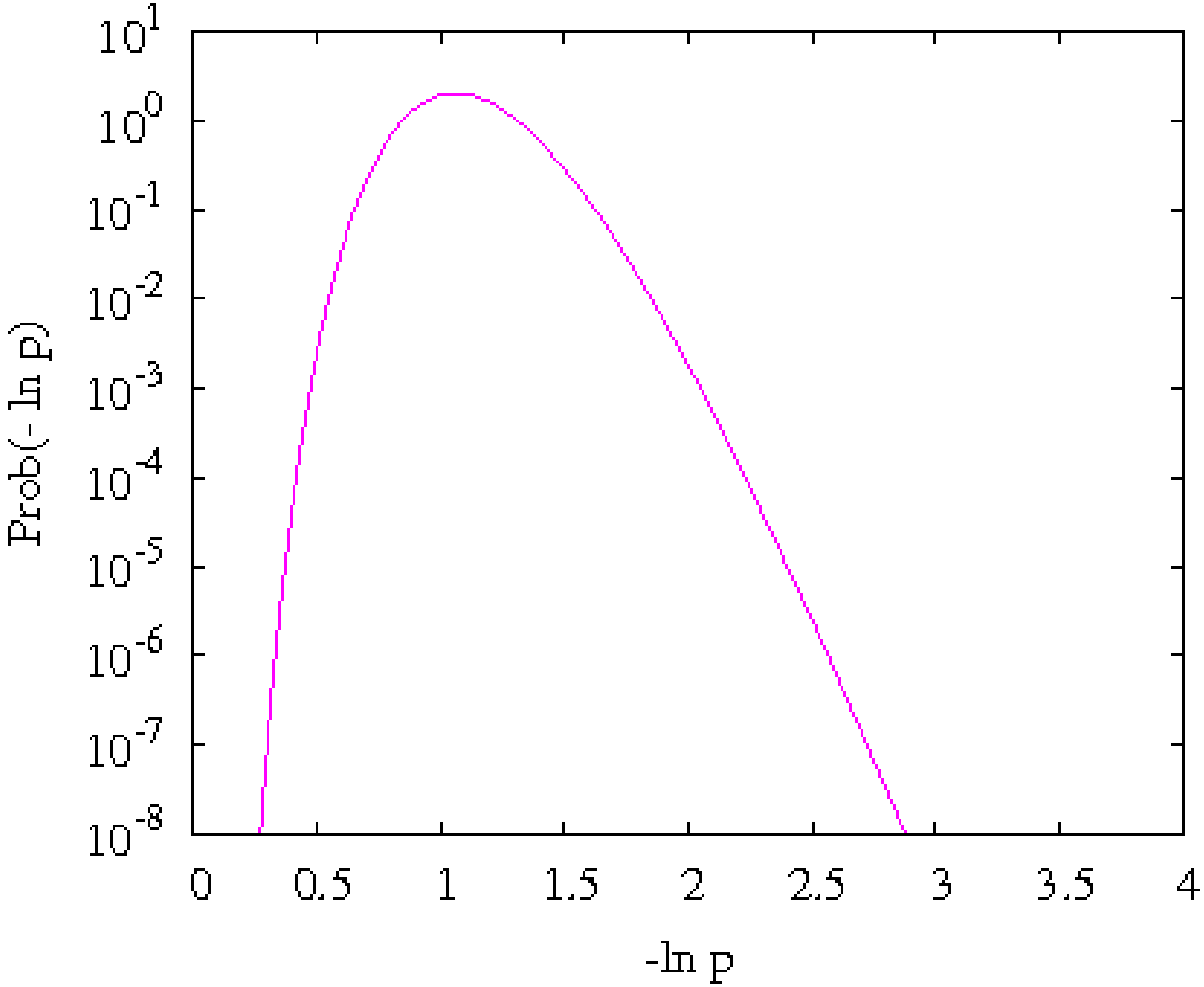
Probability for a Free Energy

$r_{min} > \lambda$	$r_{min} < \lambda$
17	33

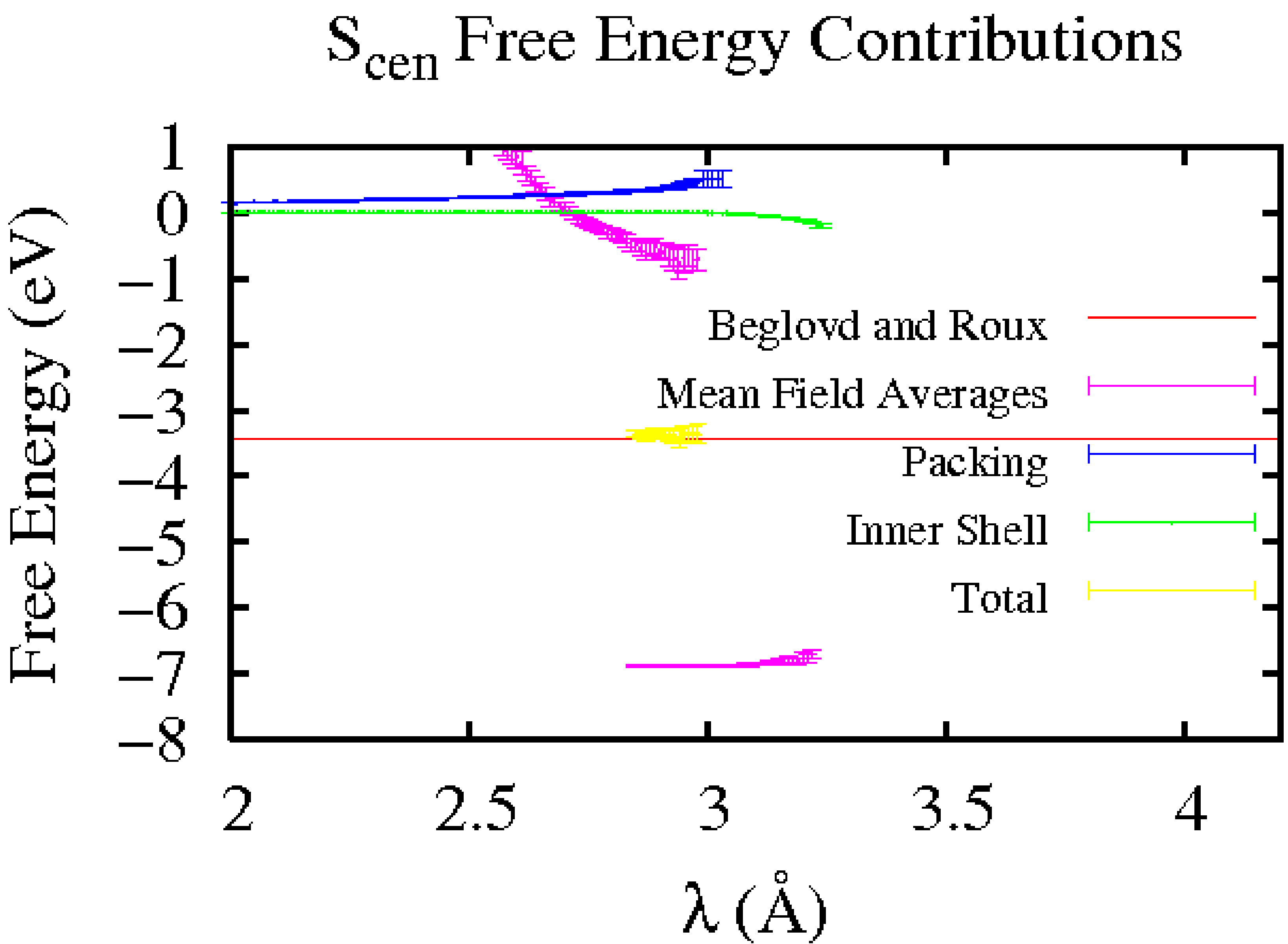
$\beta \mu^{ex}_{HS} = 1.1$ +/- 0.2



$$\beta \mu^{ex}_{HS}(\lambda) = -\ln p$$



Chloride Transporter Site S_{cen}



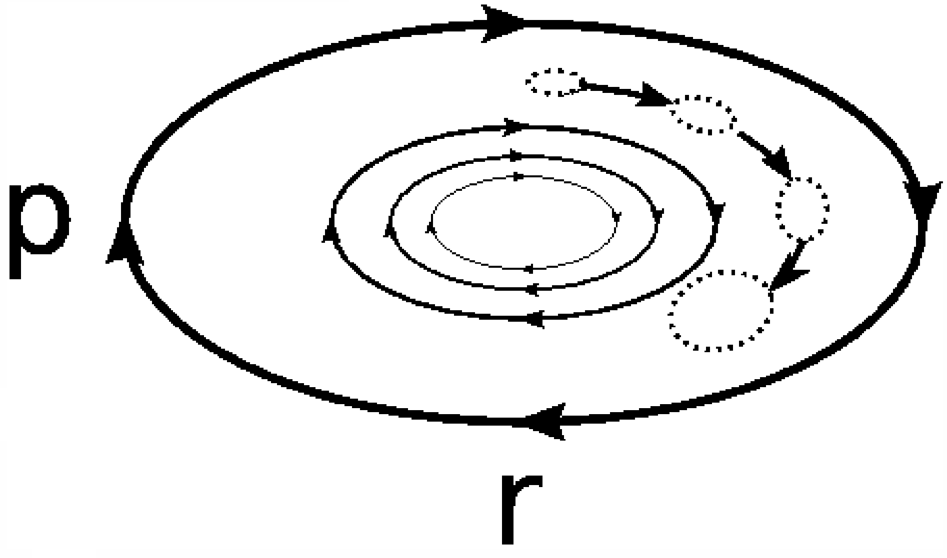
Stochastic Dynamics

$$\frac{\Delta v}{\Delta t} = \frac{F(x)}{m} + R - \gamma v$$

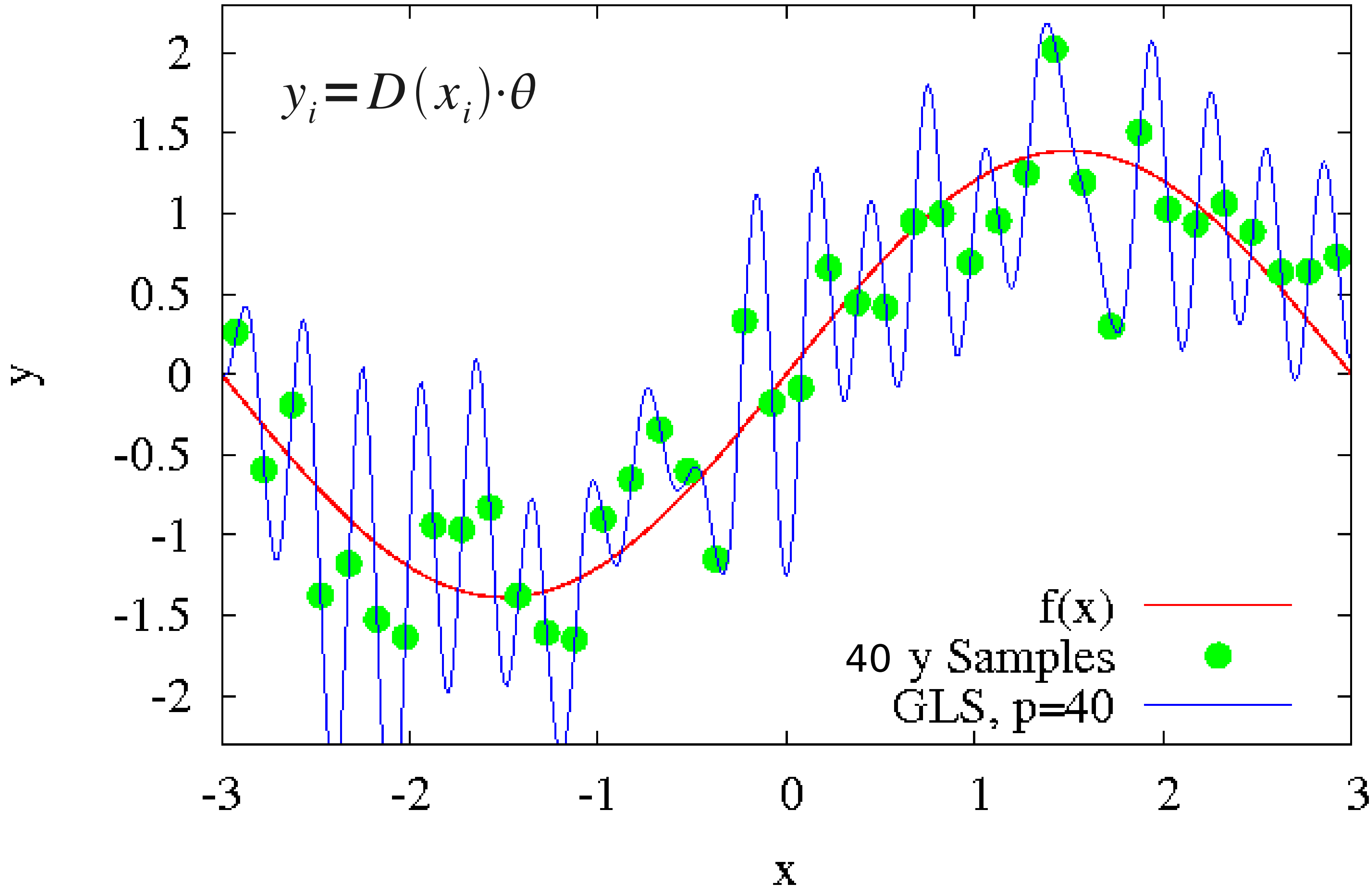
Mori, Prog. Theor. Phys. 1965.

$$\theta = \text{argmin} \| F_{\text{obs.}} - F(x;\theta) \|^2$$

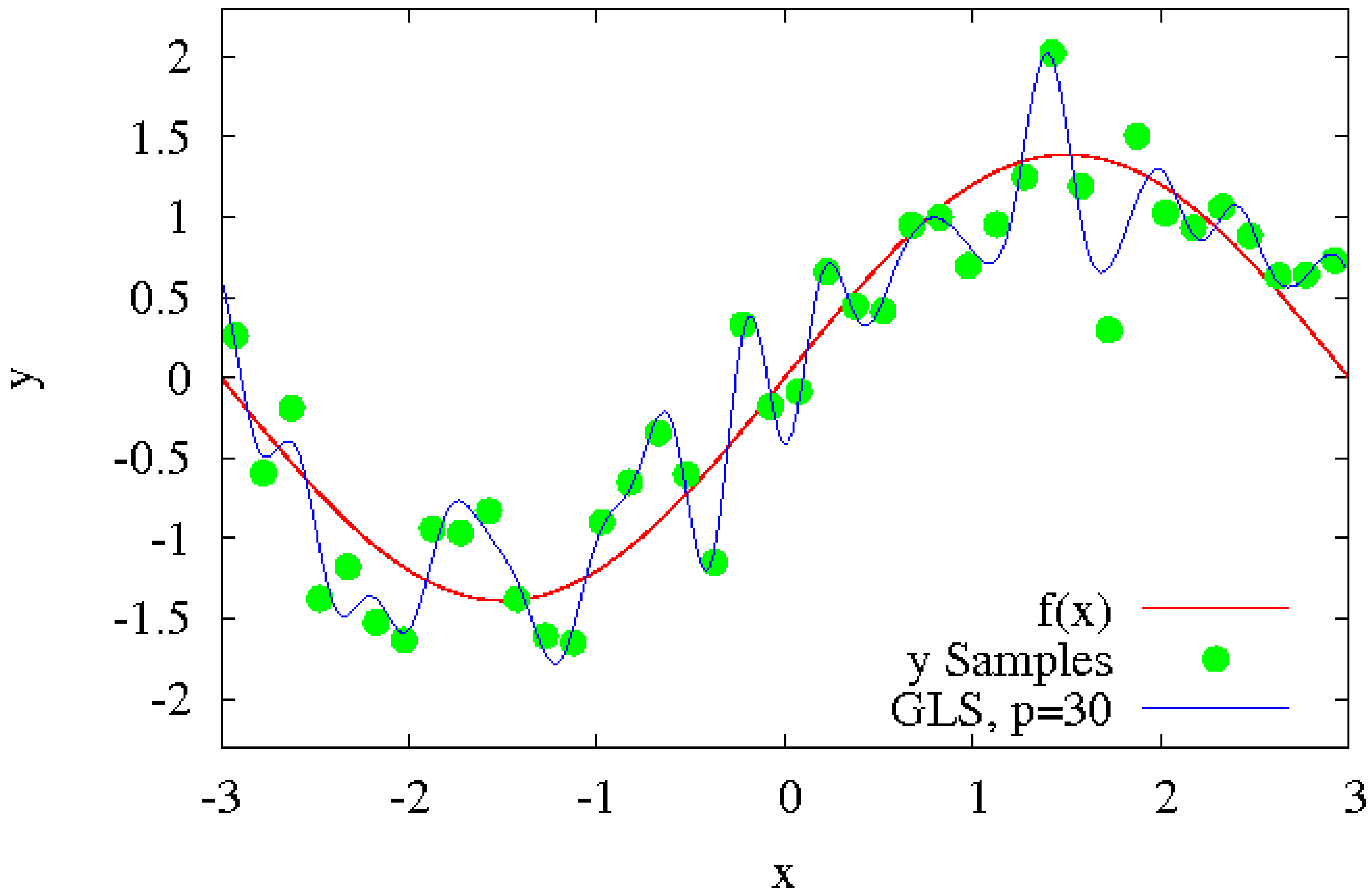
Ercolessi and Adams, Europhys. Lett. 1994.



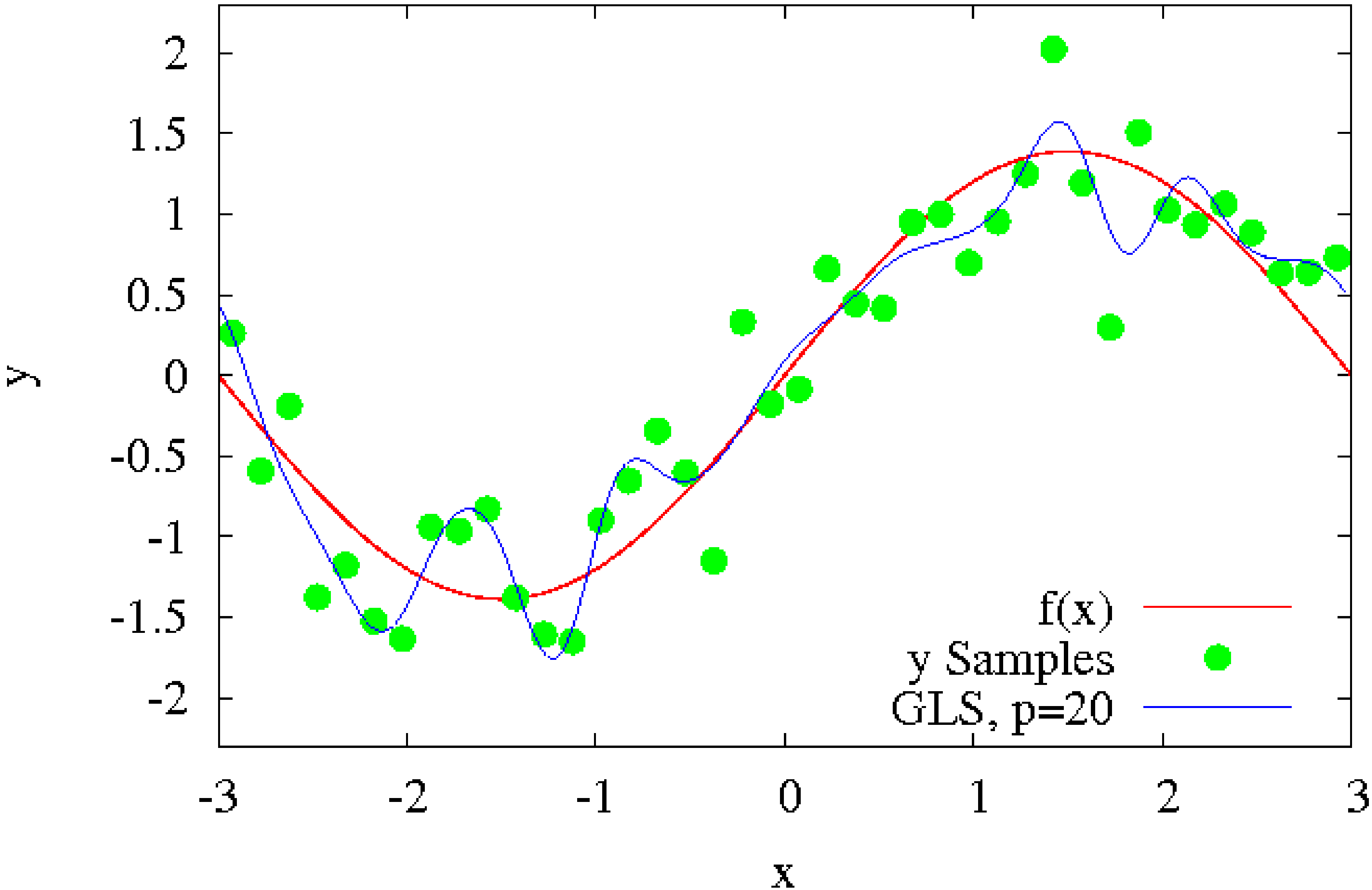
Modeling Functions with Splines: sine test function



Modeling Functions with Splines: sine test function



Modeling Functions with Splines: sine test function



Apply Bayes' Theorem

Likelihood

$$\ln \mathrm{P}(y_i | x_i \theta \sigma) = \text{const.} - \frac{\sigma^{-2}}{2} \| D(x_i) \cdot \theta - y_i \|^2$$

Prior Probability

$$\ln \mathrm{P}(\theta | \lambda \sigma I) = \text{const.} - \frac{\lambda}{2} \int f'(x)^2 \, dx, \quad \mathrm{P}(\lambda \sigma | I) \propto \sigma^{-1} \lambda^{-1} + \ldots$$

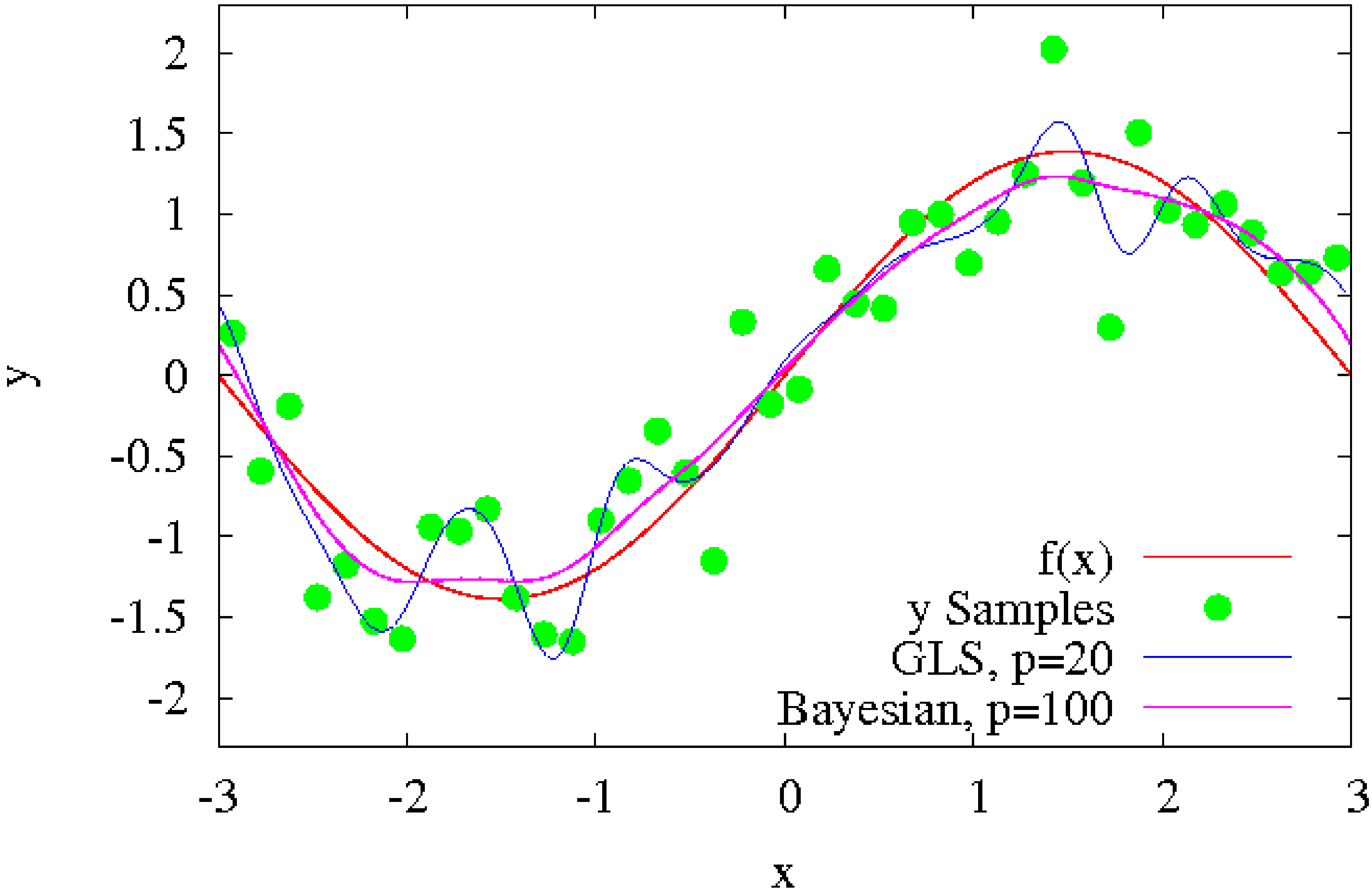
Posterior Distribution

$$\mathrm{P}(\theta \sigma \lambda | \{y, x\} I) \propto \mathrm{P}(\{y, x\} | \theta \sigma) \mathrm{P}(\theta \lambda \sigma | I)$$

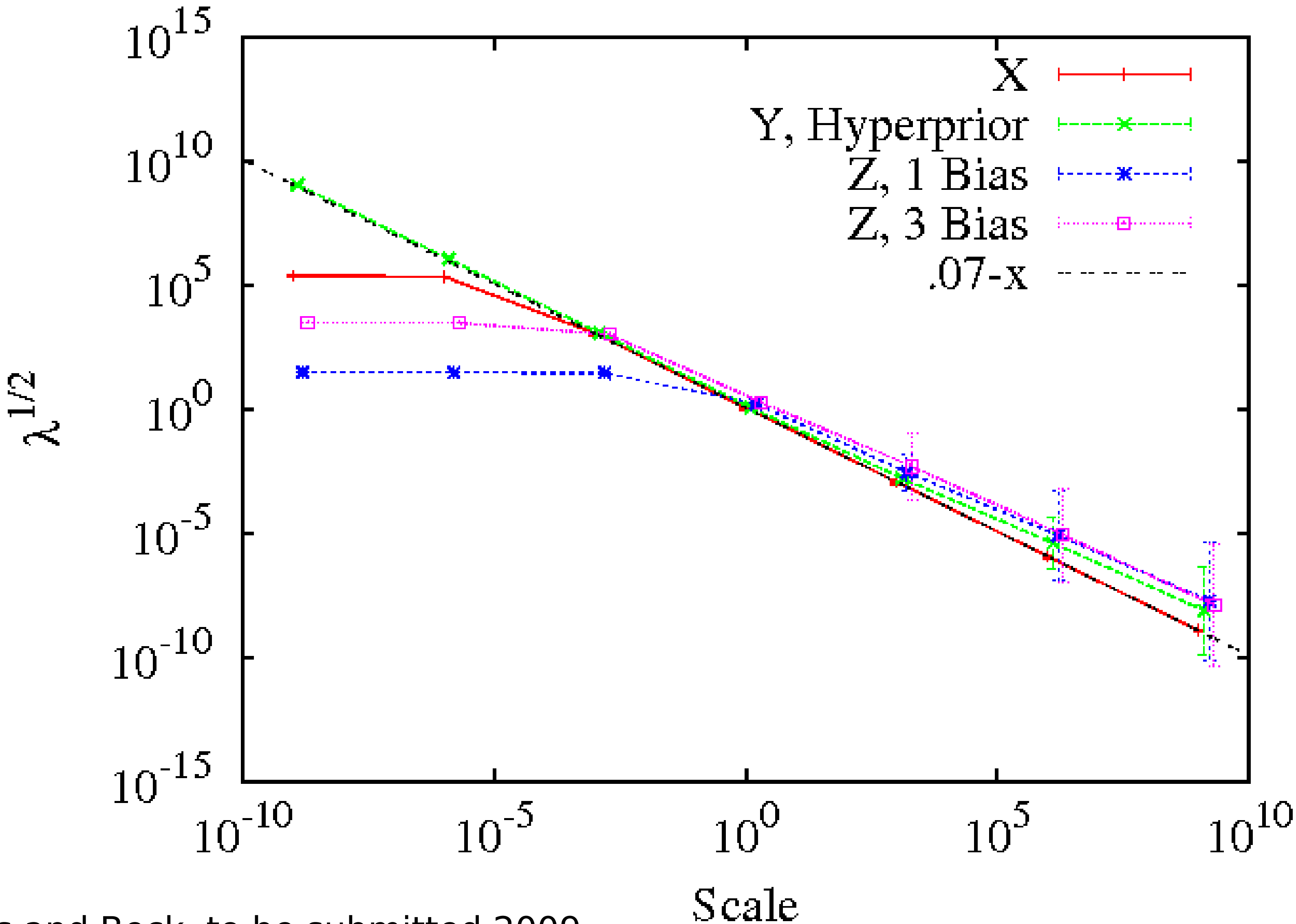
Probability for a force equation

= inference on coarse dynamics!

Modeling Functions with Splines: sine test function



Scale Independence?

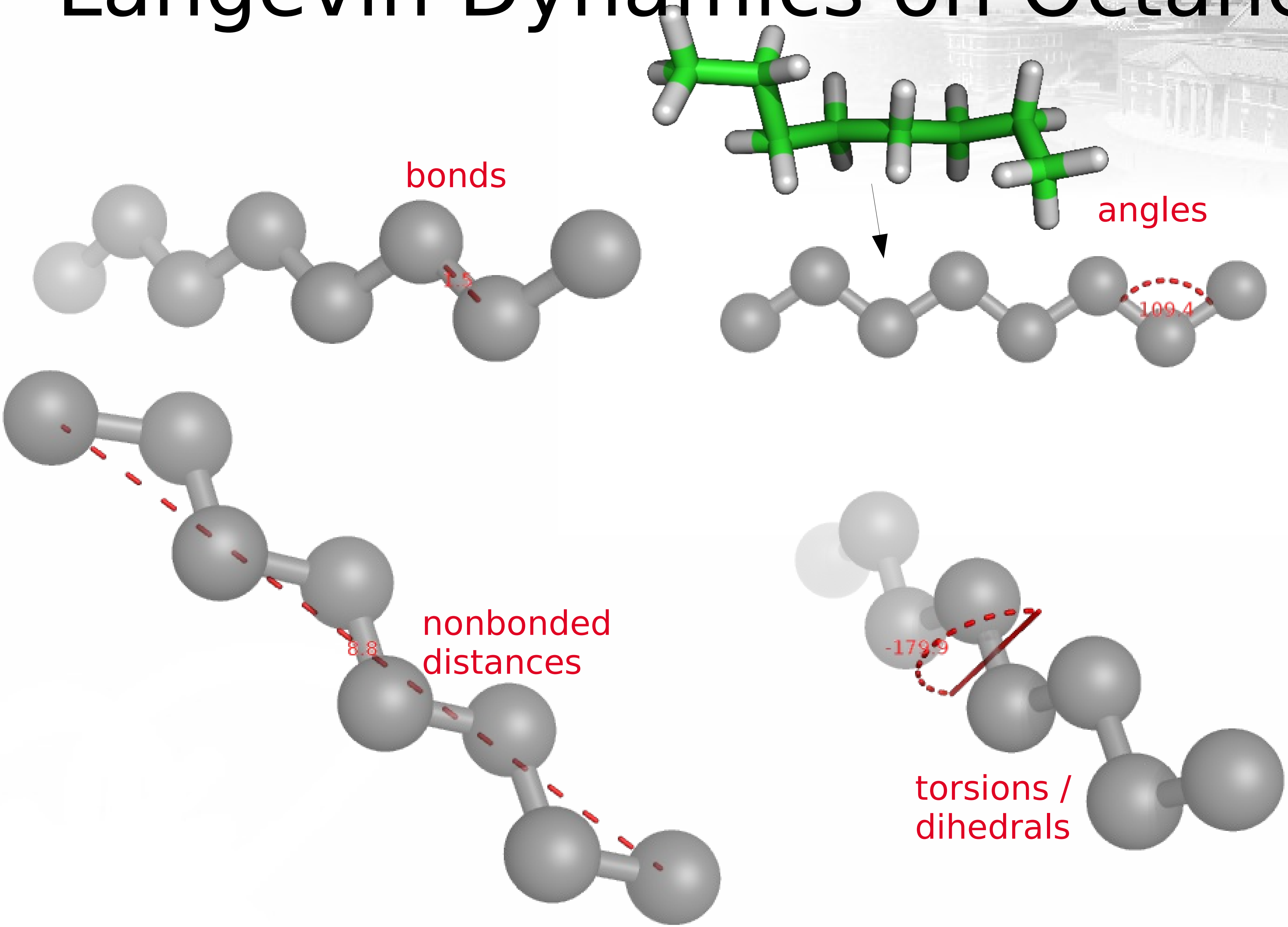


X: Rogers and Beck, to be submitted 2009.

Y: Jullion and Lambert, Comp. Stat. Anal. 2007.

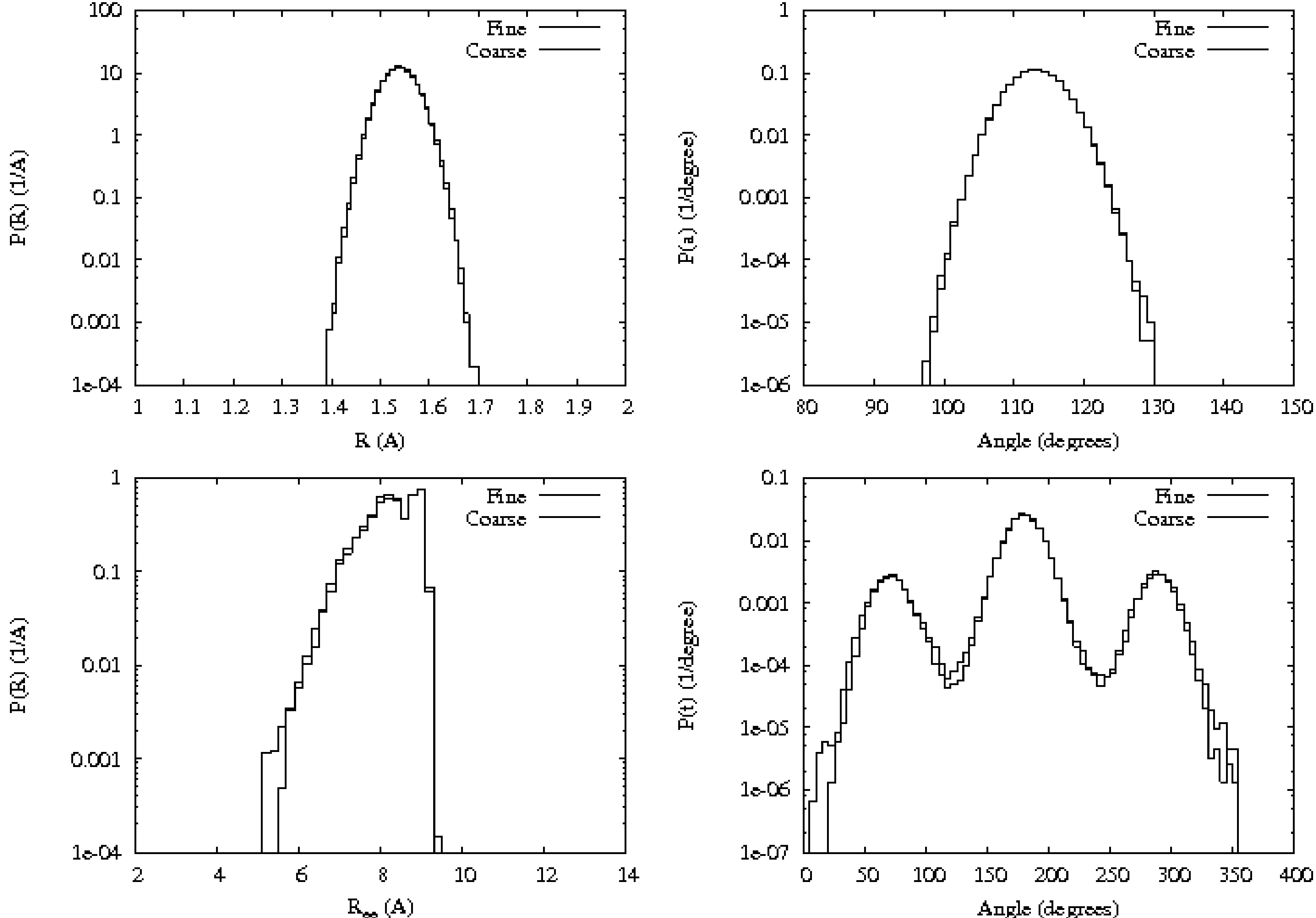
Z: Lang and Brezger, J. Comp. Graph. Stat. 2004.

Langevin Dynamics on Octane



Coordinate Distributions

Rogers and Beck 2008. http://forcesolve.sourceforge.net

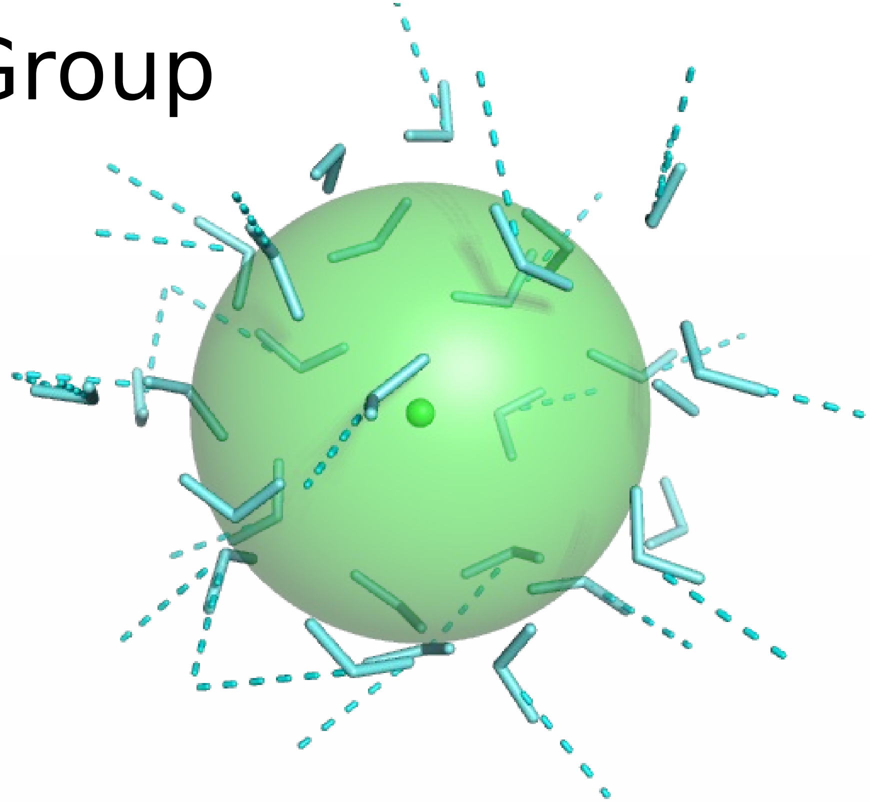


Conclusions

- Bayesian estimation
  - clarifies data analysis process.
  - gives estimates with known precision.
- Thermodynamics of solvation
  - New problem separation, new possibilities.
  - Complicated environments? Larger solutes?
- CG inference can generate parameters
  - QM → forcefield simulations
  - FF → mesoscale
  - and beyond???

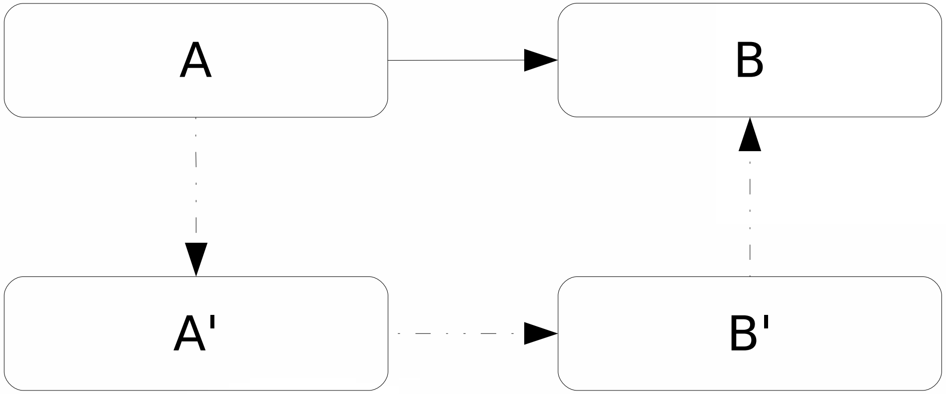
Acknowledgments

- Funding Agencies
  - DOE CSGF
  - NSF and DoD
- Beck Lab Research Group
- Audience at Large



Solvation Thermodynamic Cycle

$$\Delta F_{A->B} = \Sigma_i \, \Delta F_i$$



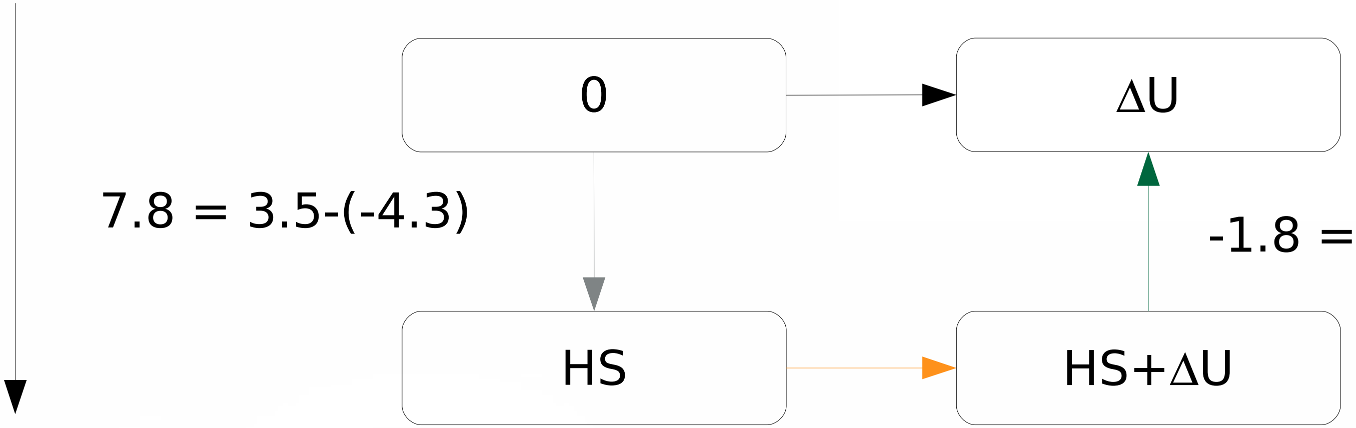
- Use any intermediate steps to break up process
  - We restrict solvent shell structures.

Methane Solvation

- QCT decomposition:

$d\mu = \quad dh - (T\,ds)$
$2.2 = -2.4 - (-4.6)$

CRC Handbook:
$2.0 = -2.9 - (-4.9)$



$7.8 = 3.5-(-4.3)$

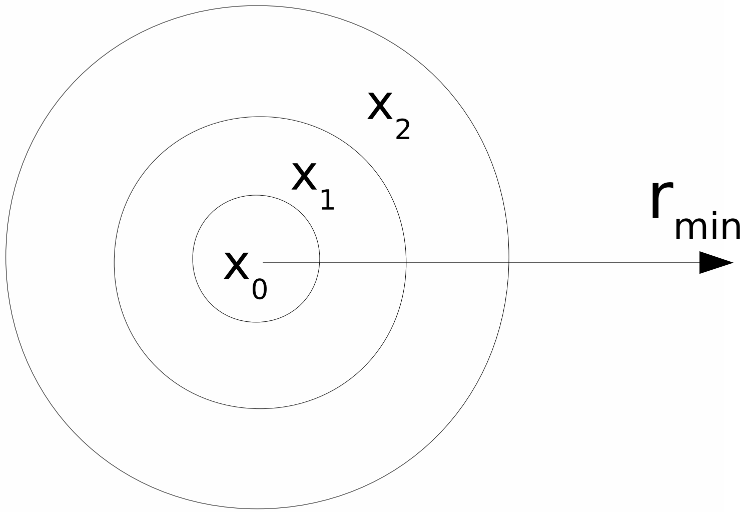
$-1.8 = -0.3 - (1.5)$

$\lambda = R_{HS} = 3.6$ Å

$-3.8 = -5.6 - (-1.8)$

$$\mu^{ex} = \mu^{ex}_{HS} + \mu^{ex}_{LR} + \mu^{ex}_{IS}$$

Multinomial Counting



- Estimates error
- Robust at small sample sizes
- instantaneous growth