Inferential Structure Determination: Overview and new developments

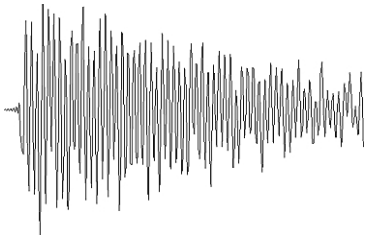
Michael Habeck

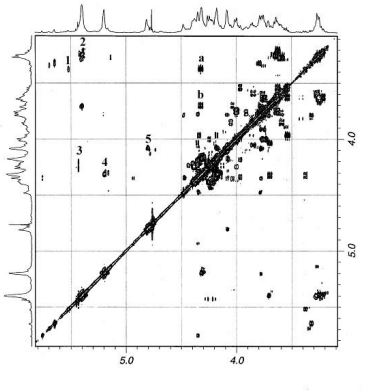
Max Planck Institutes for Developmental Biology and for Biological Cybernetics, Tübingen, Germany

michael.habeck@tuebingen.mpg.de

NMR structure determination flowchart

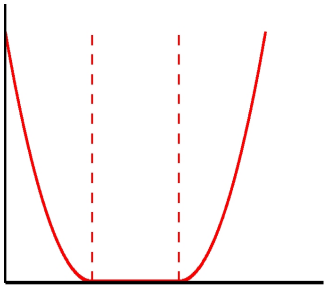
data acquisition and processing

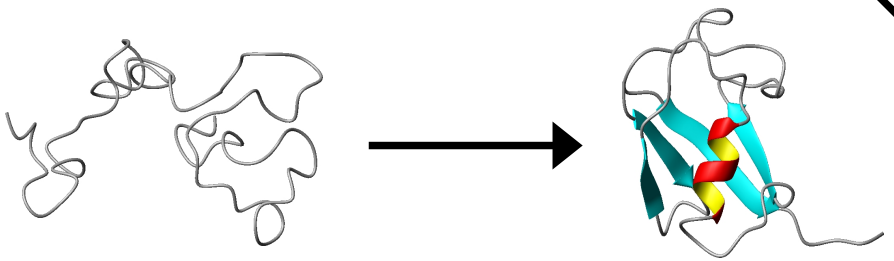




peak picking and assignment

derivation of spatial restraints





structure calculation

Subjective judgment in NMR structure determination

- derivation of conformational restraints
  - NOE classification, calibration, distance bounds
  - parameters of Karplus curve
  - determination of alignment tensor

Subjective judgment in NMR structure determination

- derivation of conformational restraints
- choice of restraint potential
  - harmonic
  - flat-bottom harmonic-wall

Subjective judgment in NMR structure determination

- derivation of conformational restraints
- choice of restraint potential
- choice of weighting factors
  - “force constants” for NOEs and other data
  - empirical defaults
  - crossvalidation

Subjective judgment in NMR structure determination

- derivation of conformational restraints
- choice of restraint potential
- choice of weighting factors
- structure calculation
  - how many?
  - starting structure?

Subjective judgment in NMR structure determination

- derivation of conformational restraints
- choice of restraint potential
- choice of weighting factors
- structure calculation
- selection of representative structures
  - selection criterion: energy based?
  - restraint violation?

Subjective judgment in NMR structure determination

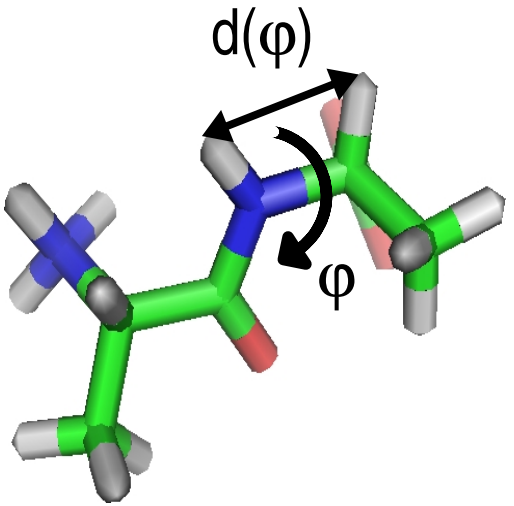
- derivation of conformational restraints
- choice of restraint potential
- choice of weighting factors
- structure calculation
- selection of representative structures

➡ structures depend on subjective decisions rather than just the data

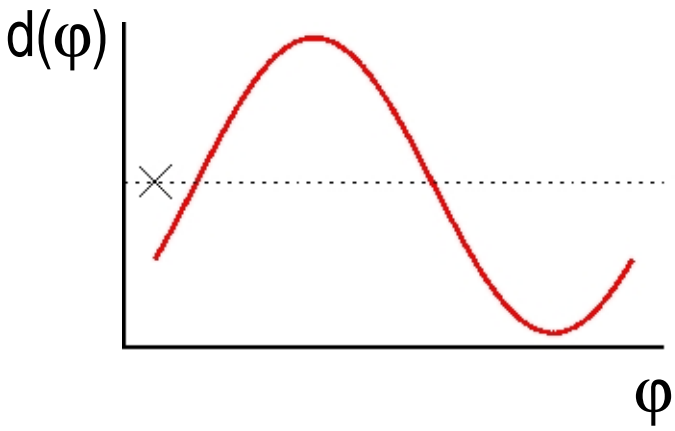
Goals

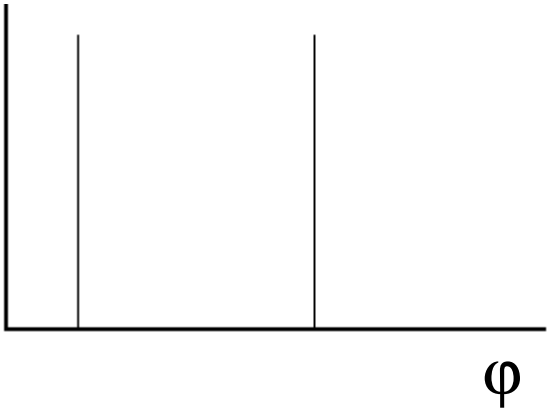
- reduce subjective judgment
- quantify the uncertainty that's inherent in NMR structure determination
- calculate “objective” NMR structures
- make statements about reliability

A simple example

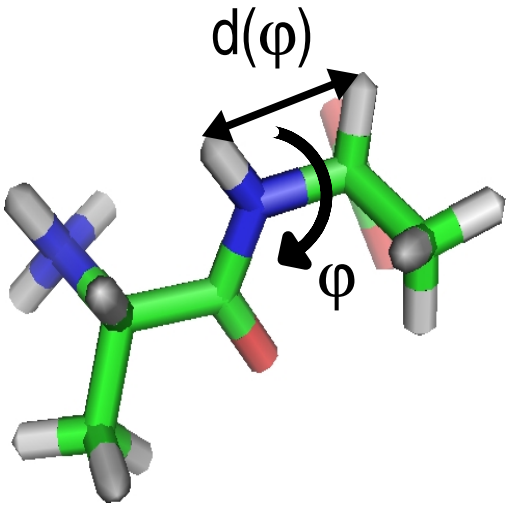


- a single conformational degree of freedom (φ angle)
- data: NOEs for HN-HA distance d(φ)

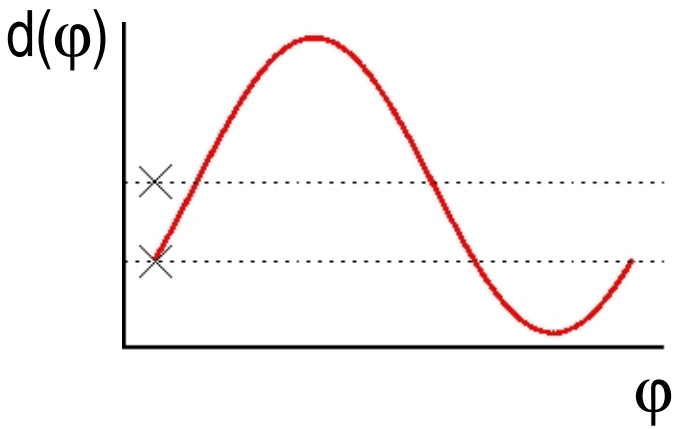


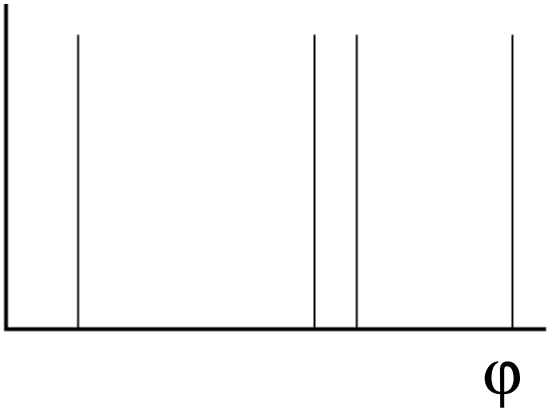


A simple example



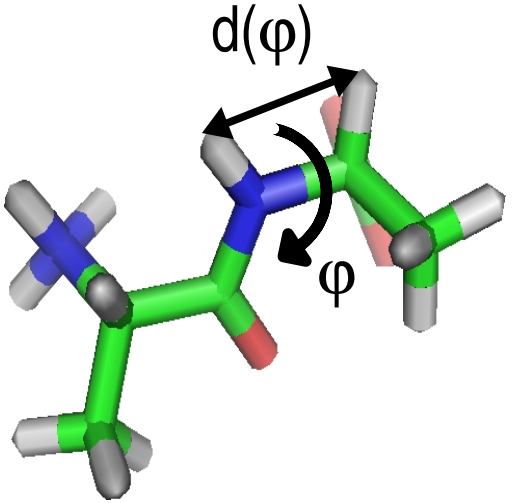
- a single conformational degree of freedom (φ angle)
- data: NOEs for HN-HA distance d(φ)

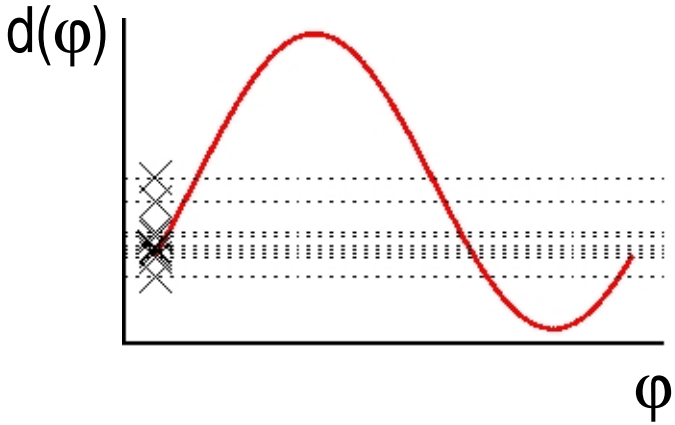


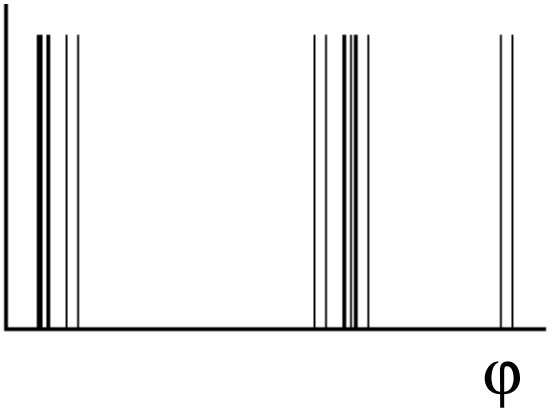


A simple example

- a single conformational degree of freedom (φ angle)
- data: NOEs for HN-HA distance d(φ)

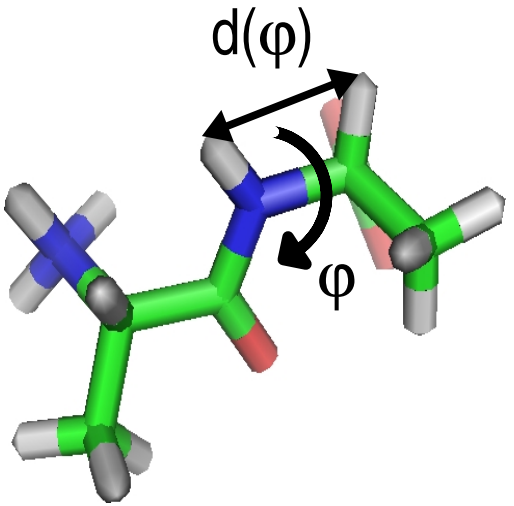


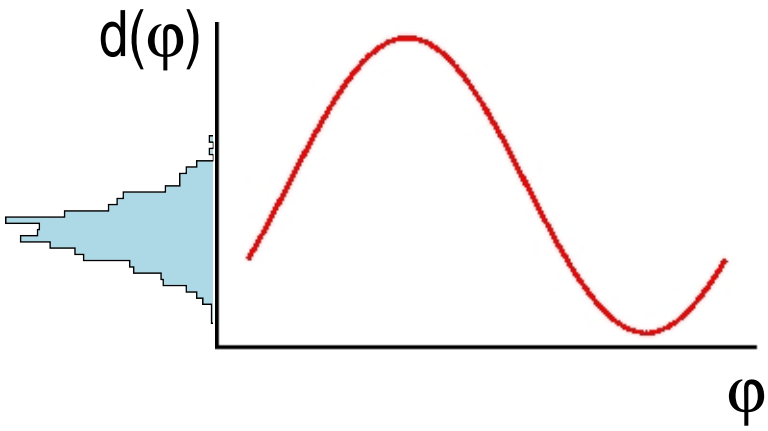


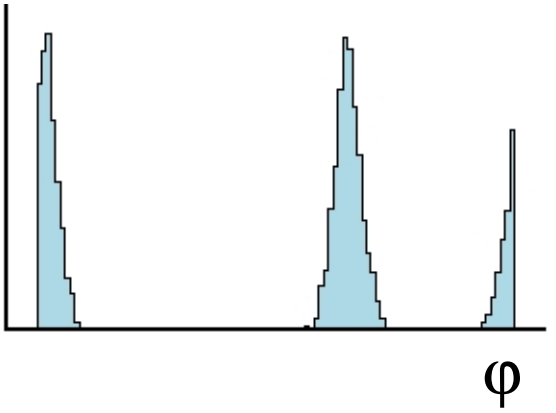


A simple example

- a single conformational degree of freedom (φ angle)
- data: NOEs for HN-HA distance d(φ)

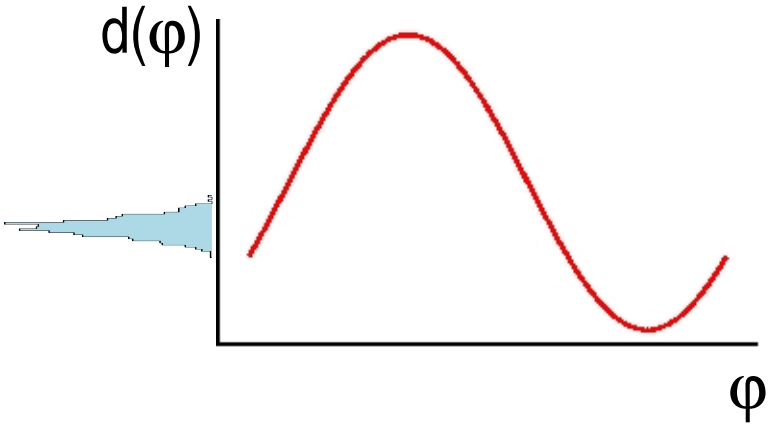


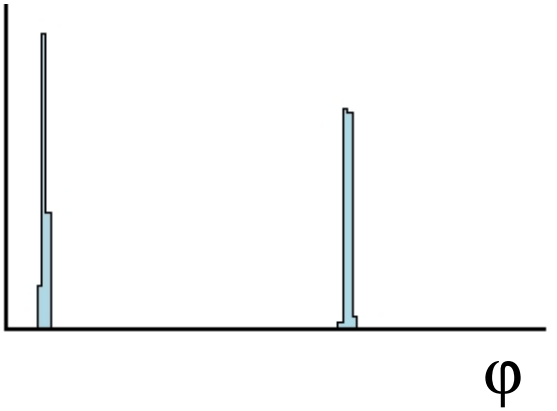




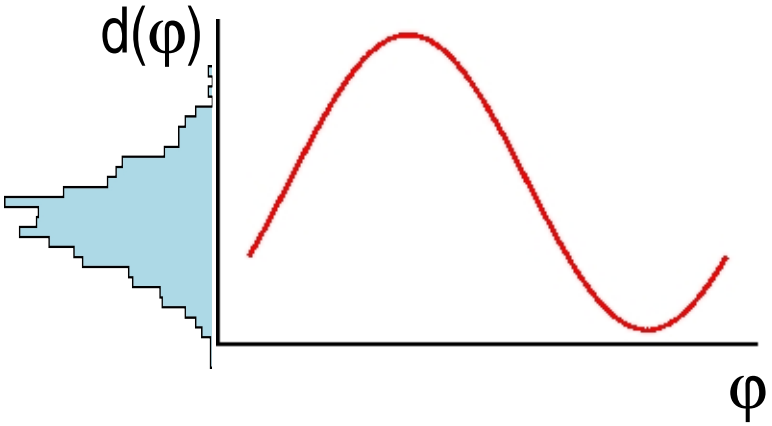
Probability as a measure of uncertainty

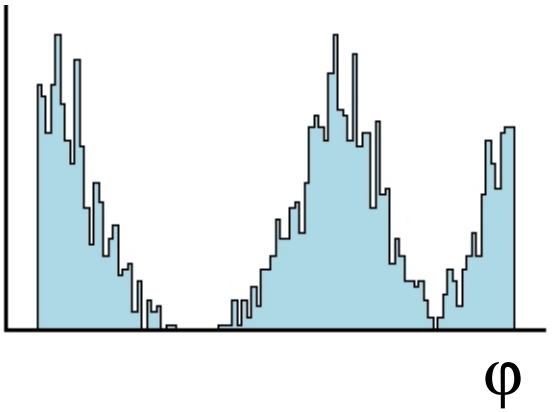
- certain φ



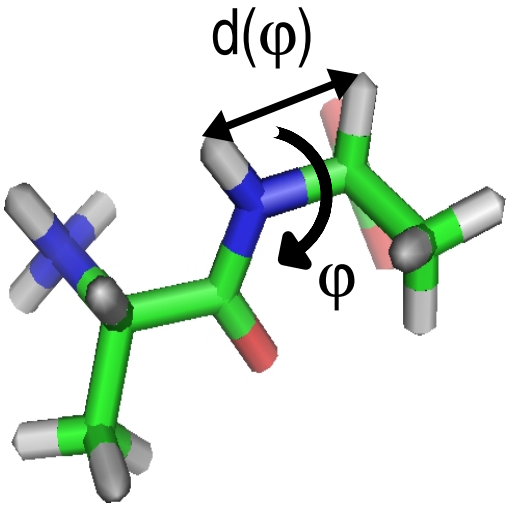


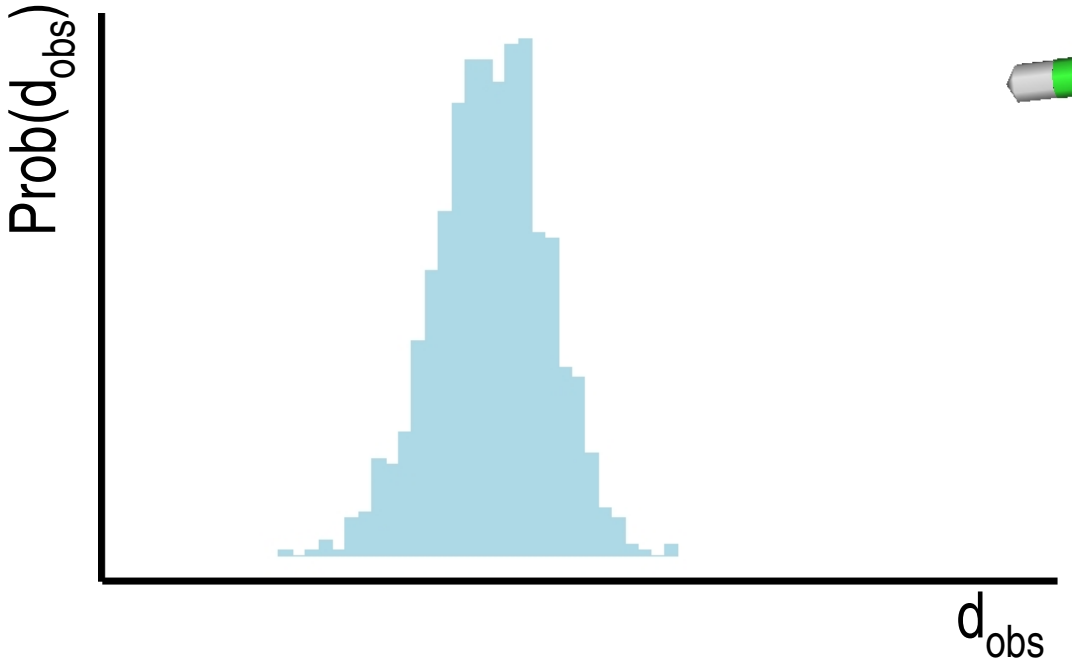
- uncertain φ





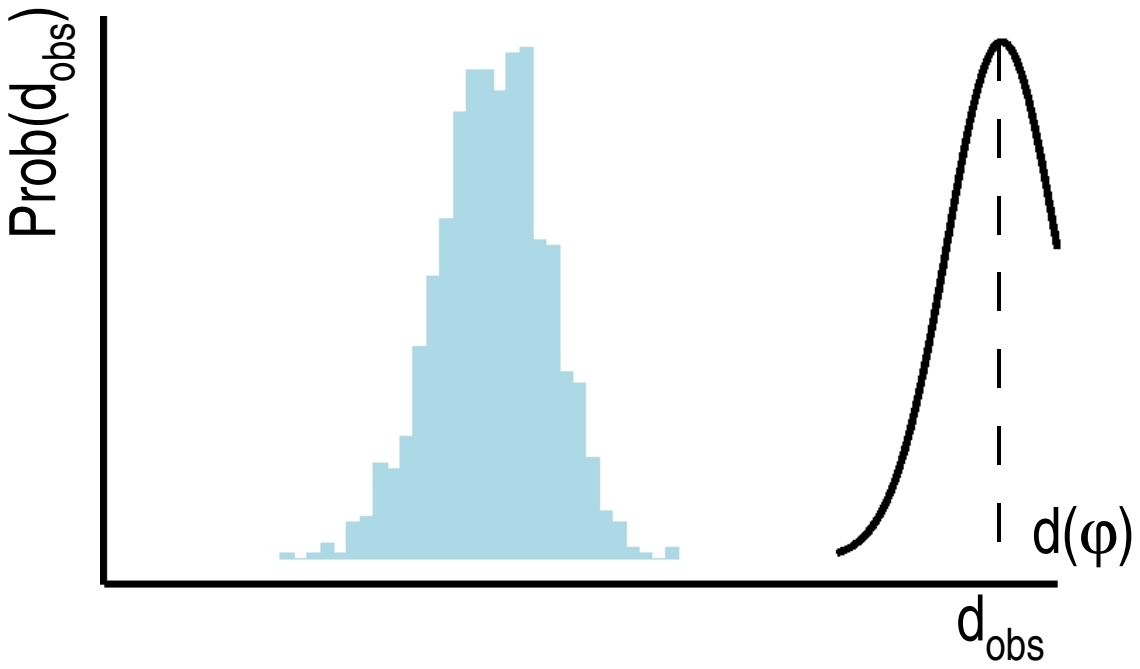
Describing data with probabilities





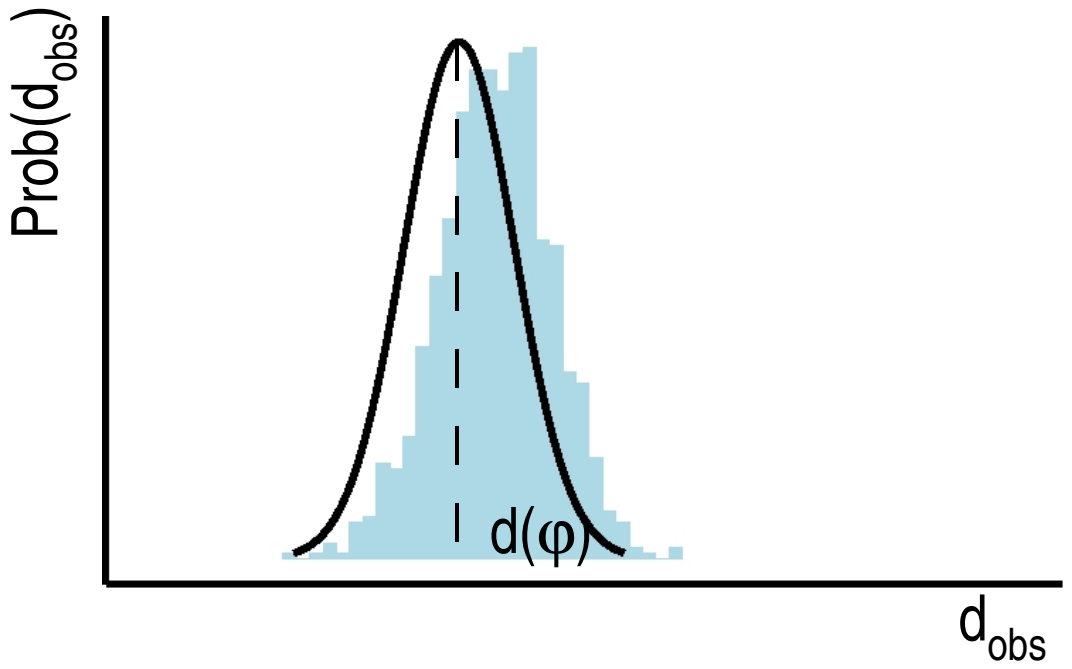
$$\mathrm{Prob}(d_{obs}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_{obs}-d(\varphi))^2}$$

Matching φ



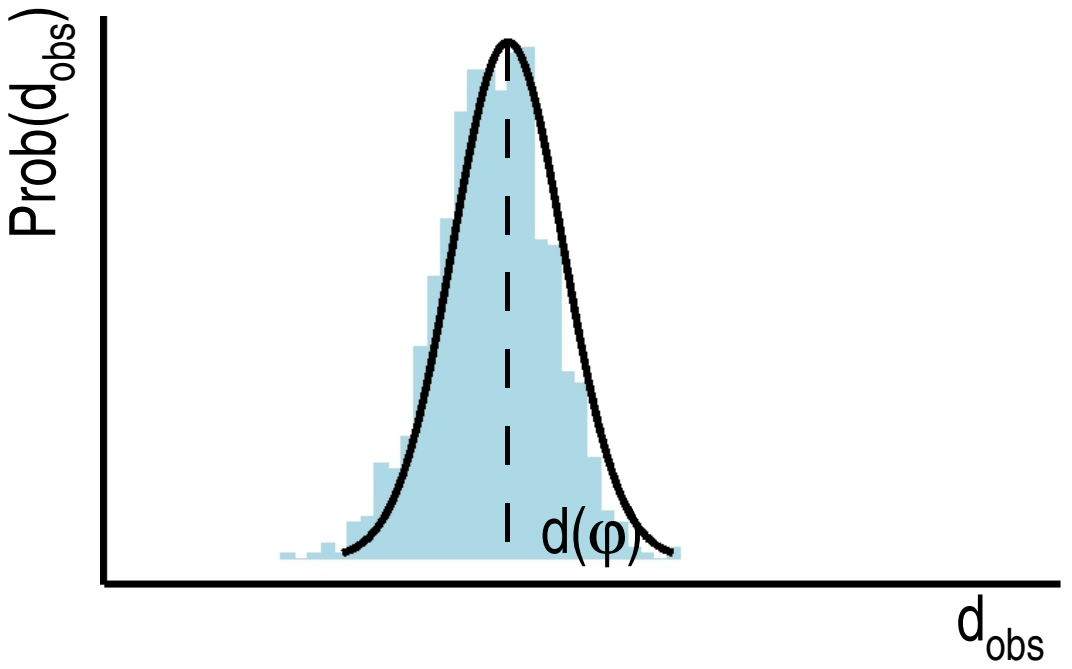
$$\text{Prob}(d_{obs}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_{obs}-d(\varphi))^2}$$

Matching φ



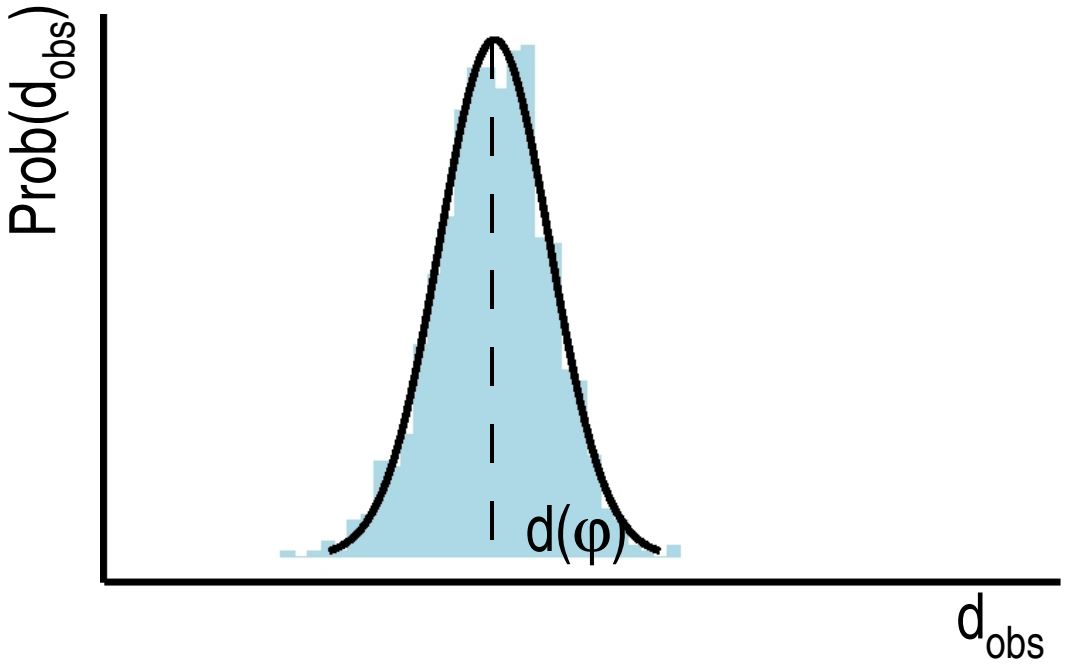
$$\mathrm{Prob}(d_{obs}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_{obs}-d(\varphi))^2}$$

Matching φ



$$\mathrm{Prob}(d_{obs}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_{obs}-d(\varphi))^2}$$

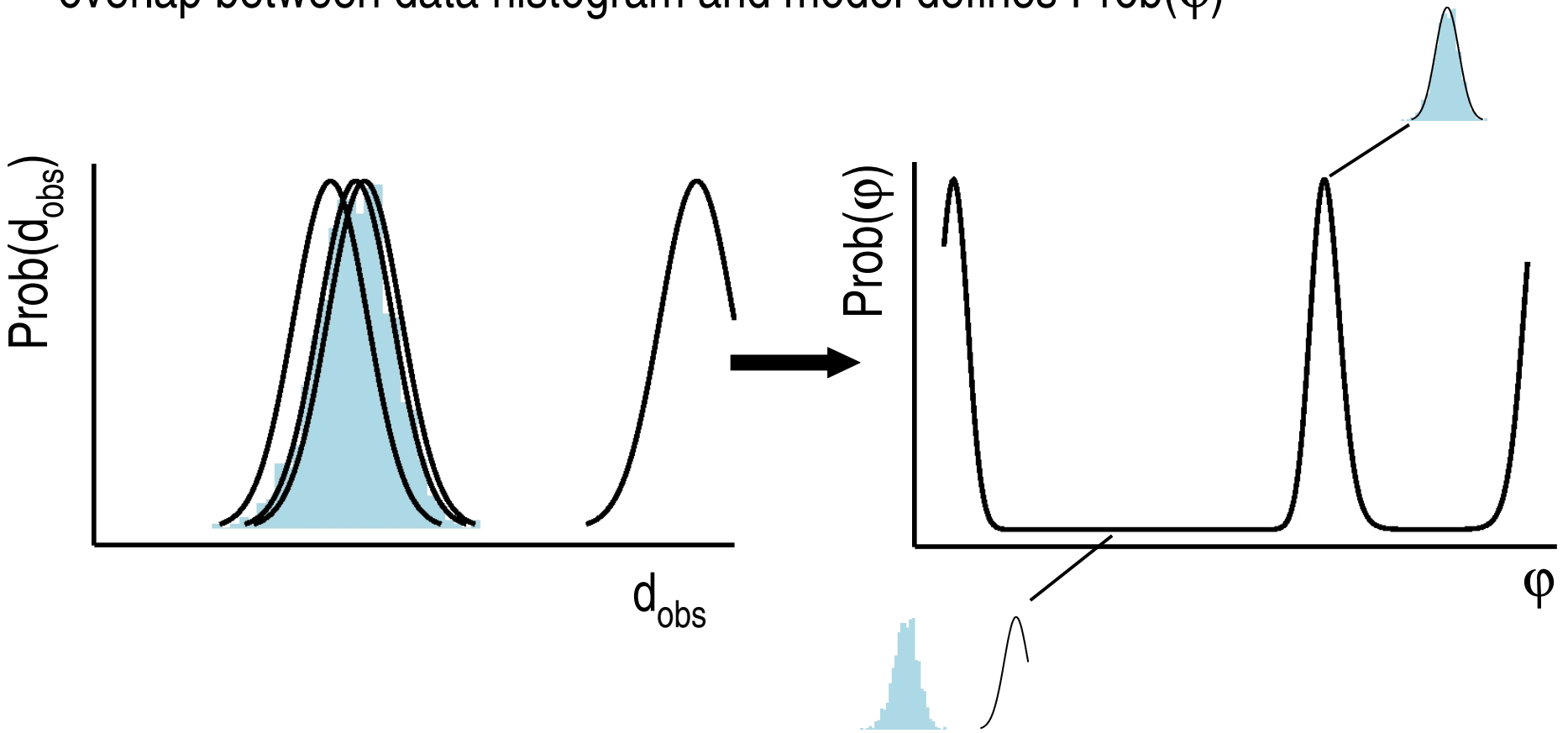
Matching φ



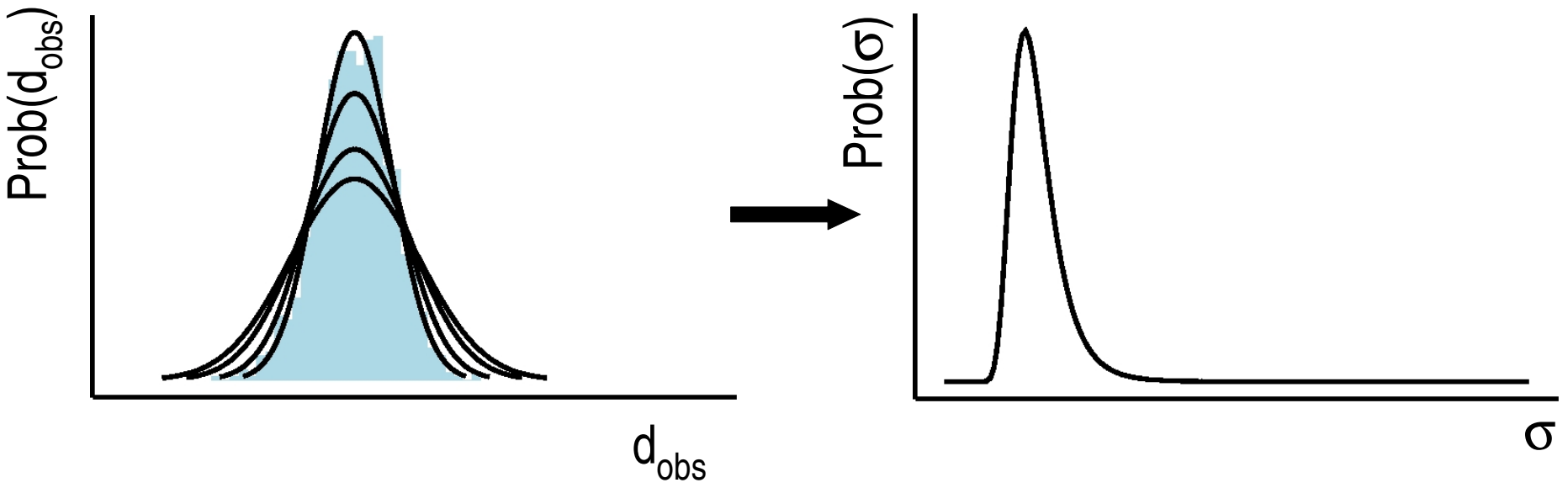
$$\mathrm{Prob}(d_{obs}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_{obs}-d(\varphi))^2}$$

Matching φ

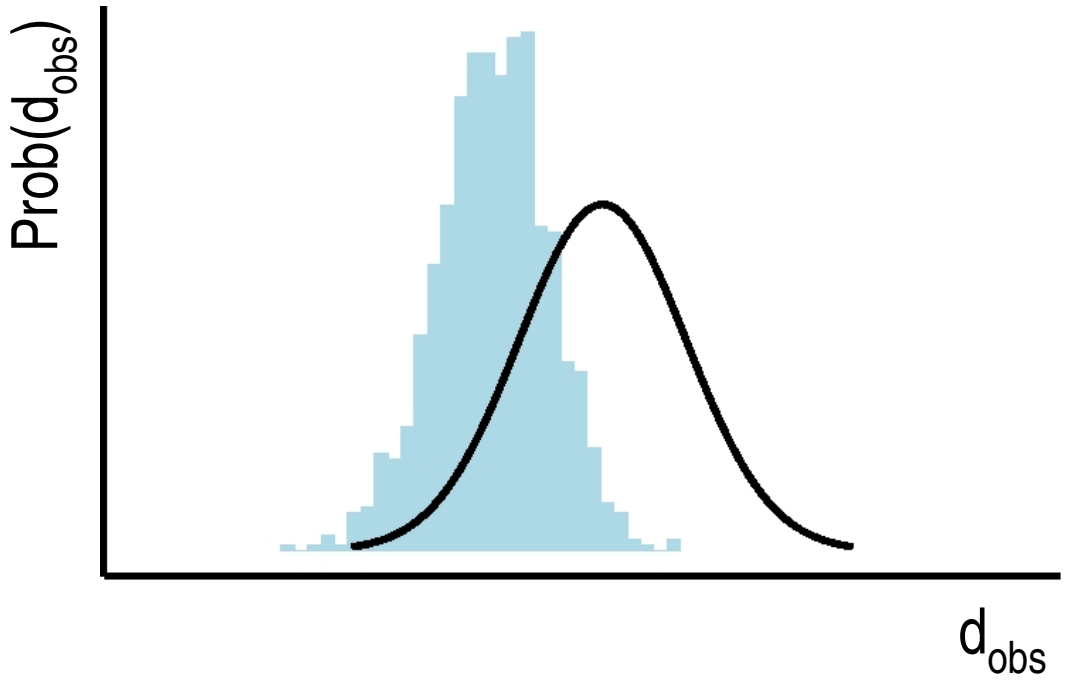
overlap between data histogram and model defines Prob(φ)



Matching σ

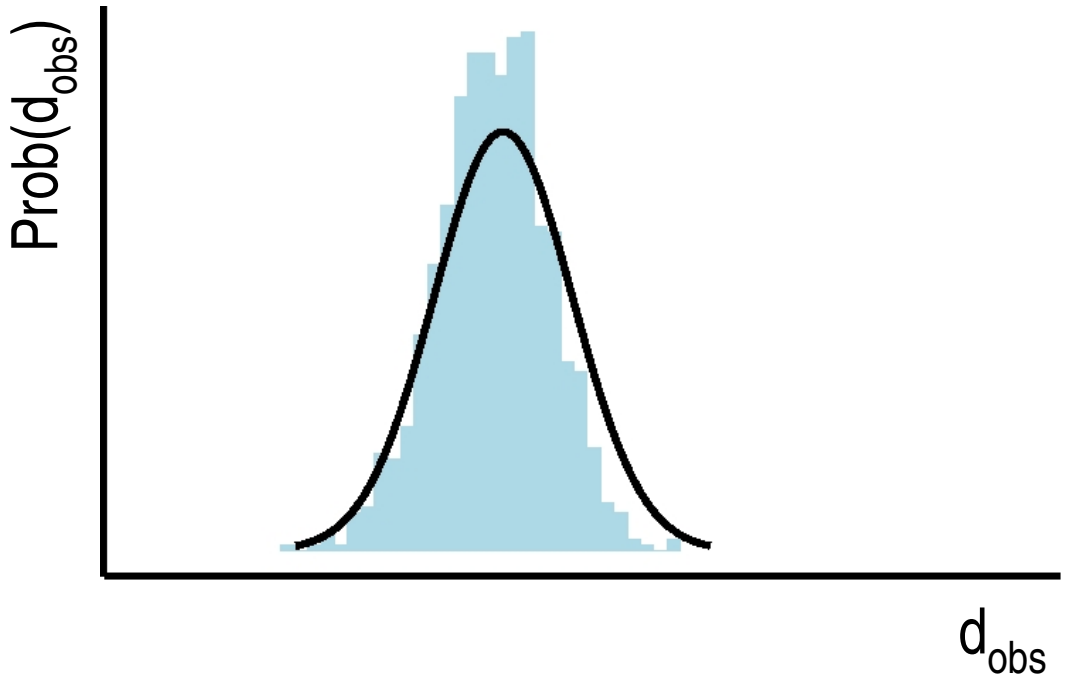


Matching φ and σ simultaneously



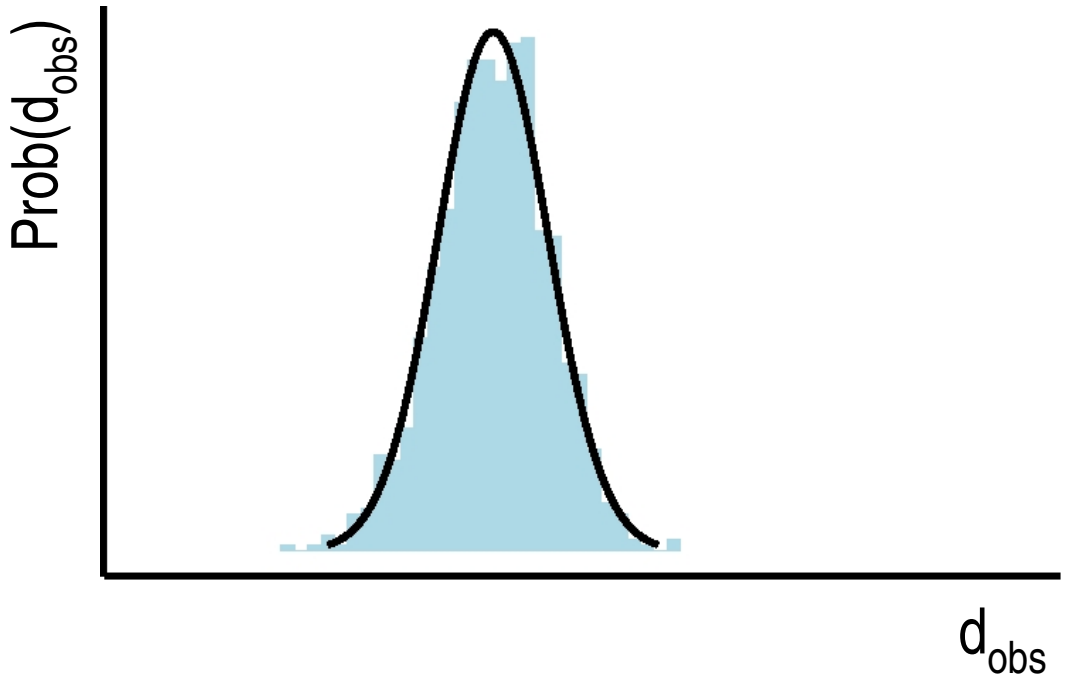
$$\mathrm{Prob}(d_{obs}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_{obs}-d(\varphi))^2}$$

Matching φ and σ simultaneously



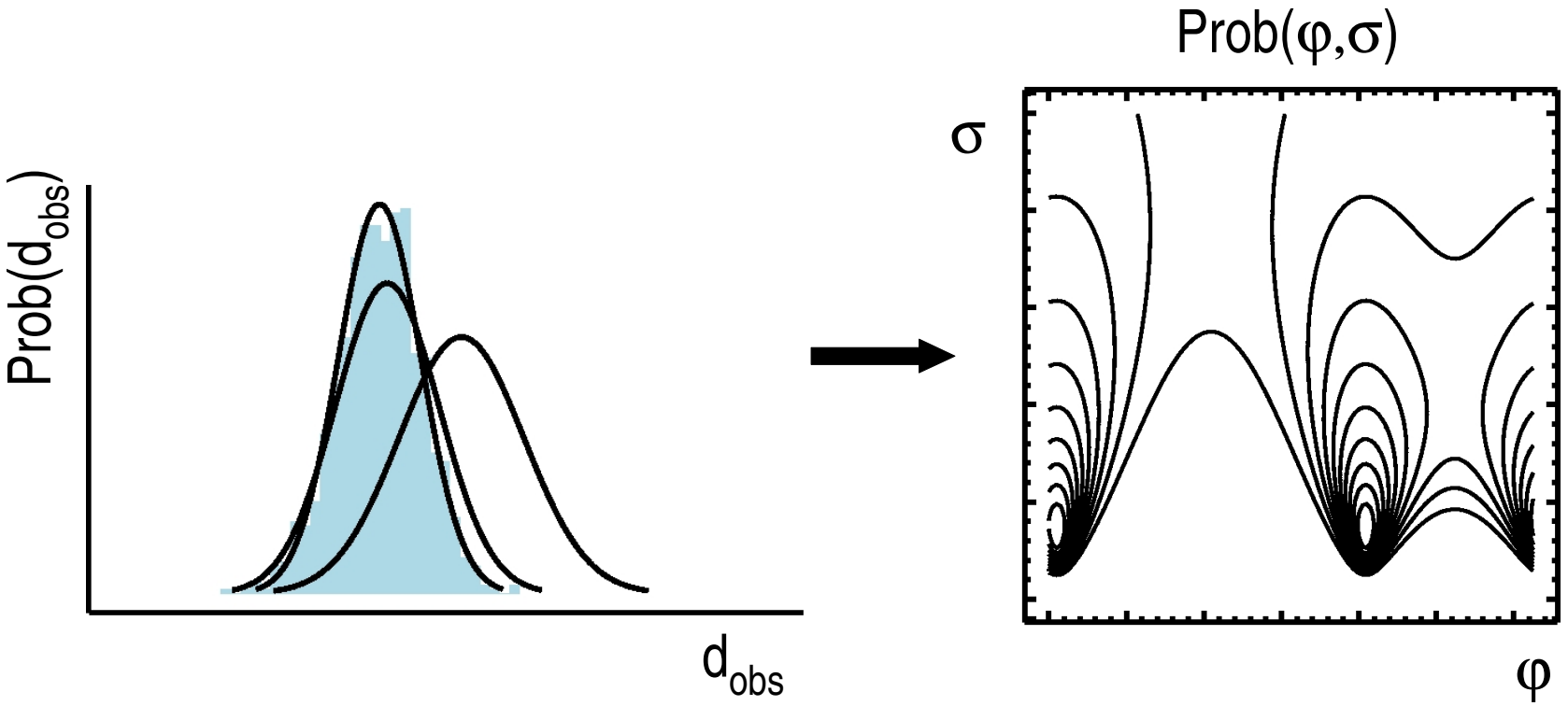
$$\mathrm{Prob}(d_{obs}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_{obs}-d(\varphi))^2}$$

Matching φ and σ simultaneously

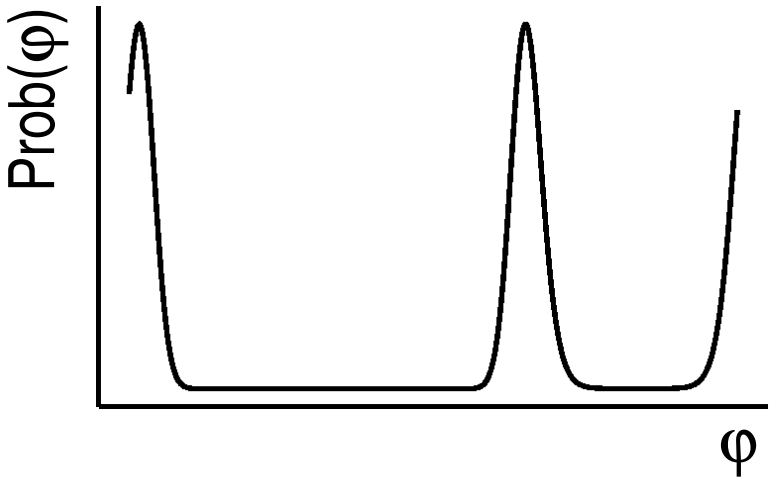


$$\mathrm{Prob}(d_{obs}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_{obs}-d(\varphi))^2}$$

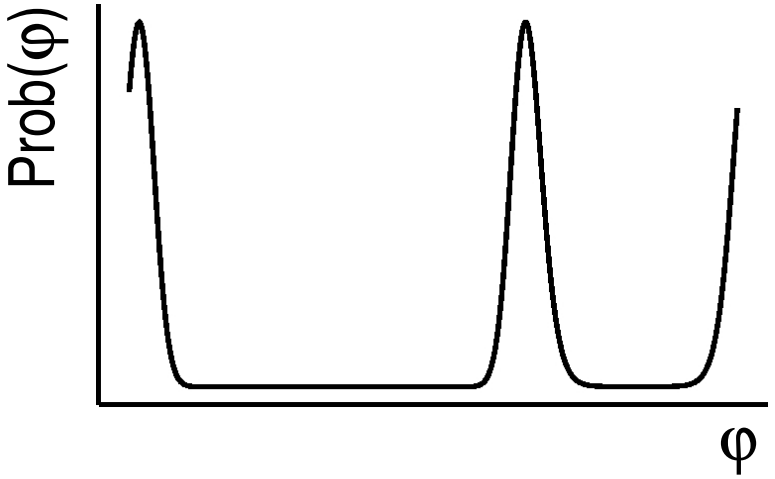
Matching φ and σ simultaneously

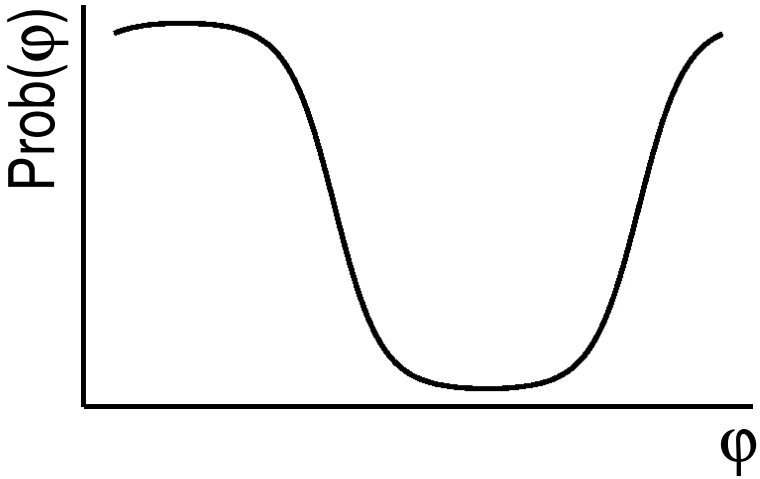


Prior information

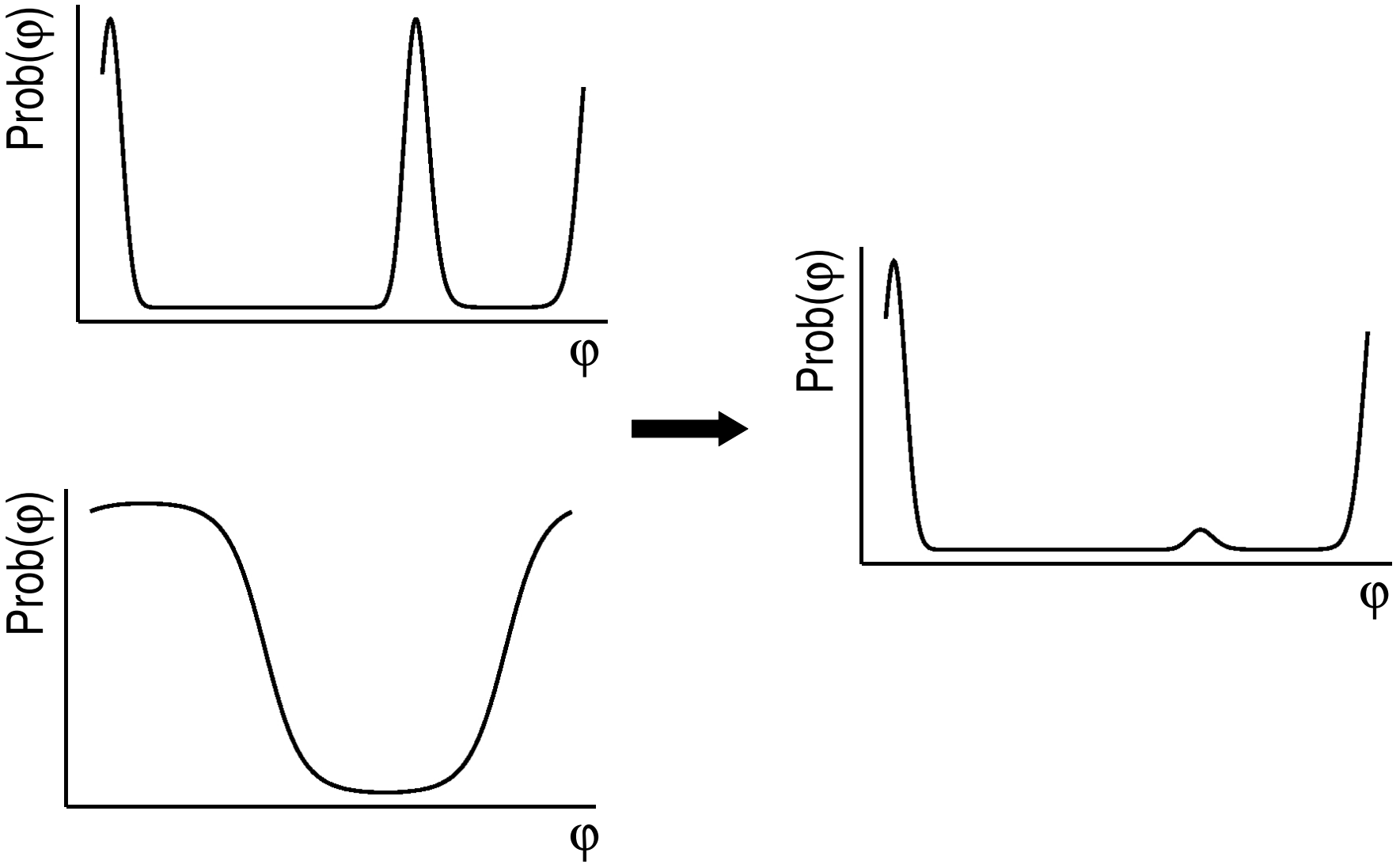


Prior information

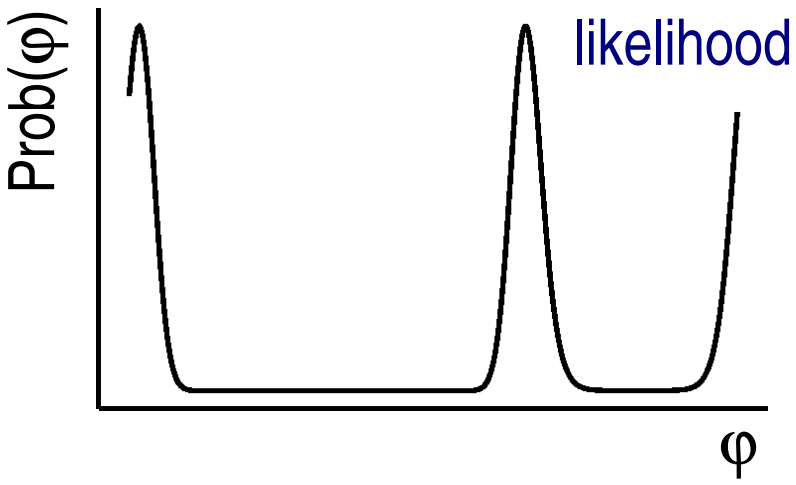


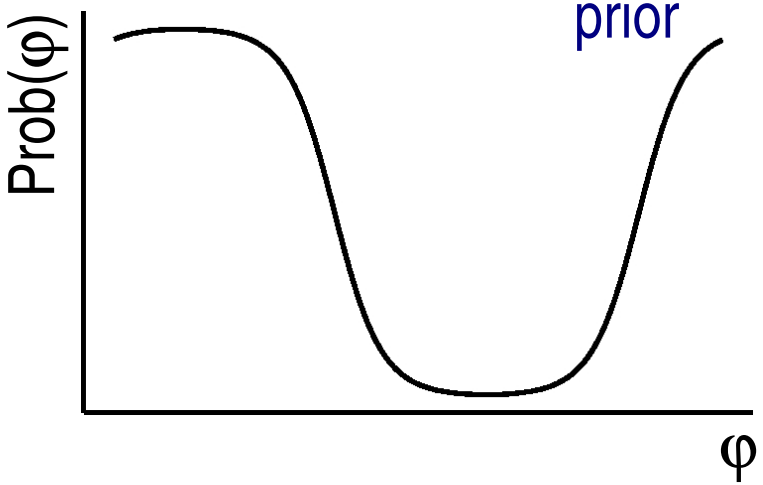


Prior information



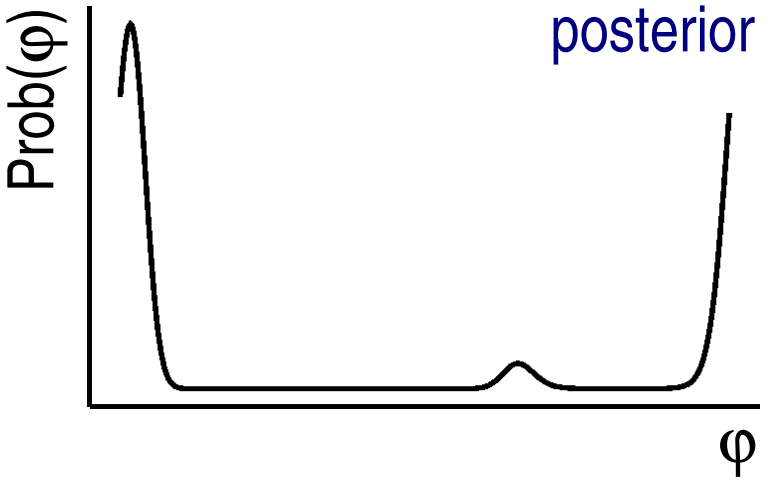
Prior information





posterior ~ likelihood x prior

(Bayes' theorem)



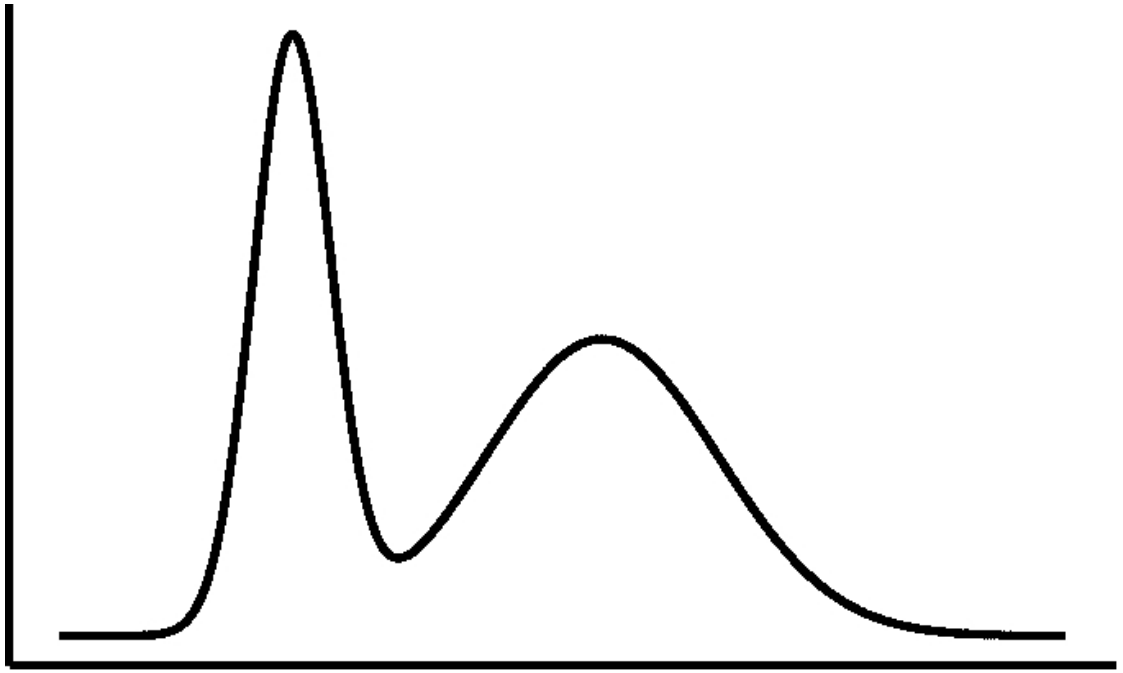
Inferential Structure Determination (ISD)

- quantify uncertainties by probabilities
- describe NMR data probabilistically
- express background knowledge by prior probabilities
- consistently combine the probabilities using probability calculus (Bayes' theorem)
- analyse the joint posterior probability of **all** unknowns (coordinates + model parameters + errors)

Science, 2005; Phys Rev E, 2005

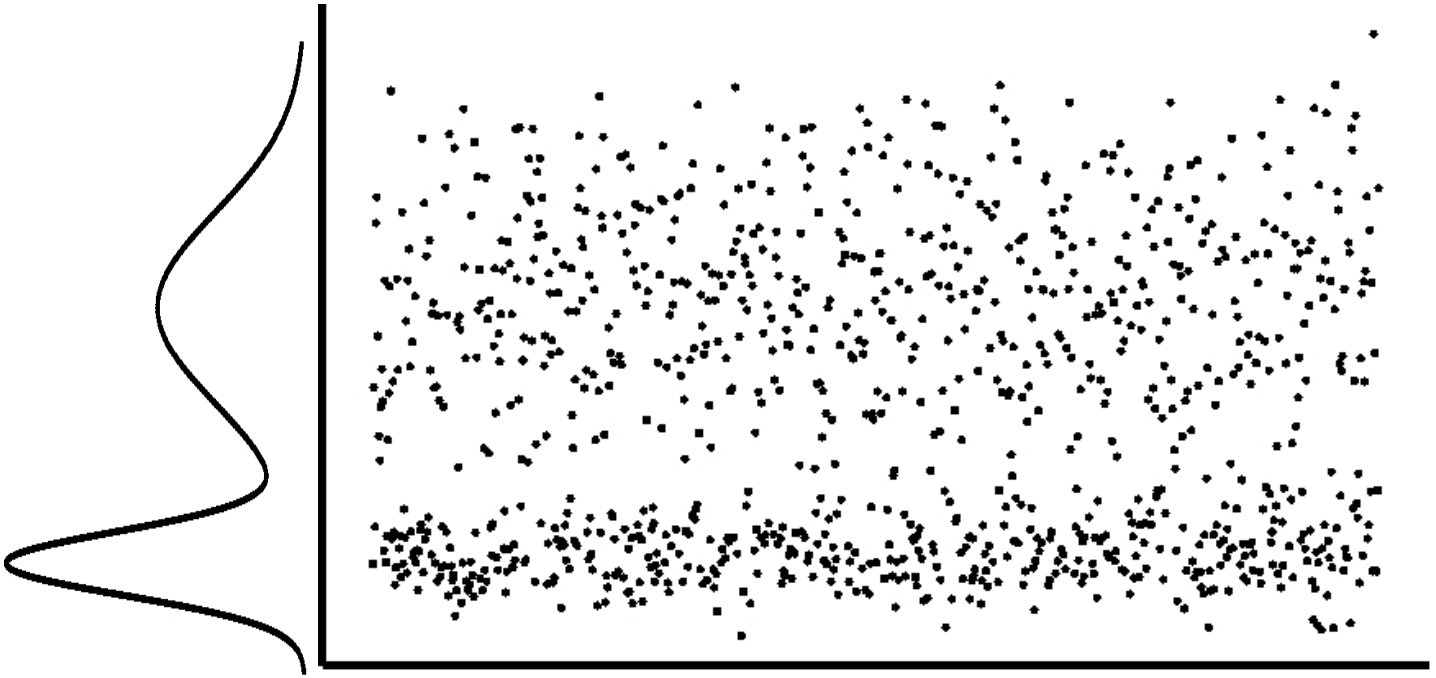
Drawing random samples from probabilities

- in real-world problems probabilities are too complex for visual or analytical analysis
- idea of sampling: pick a set of representatives



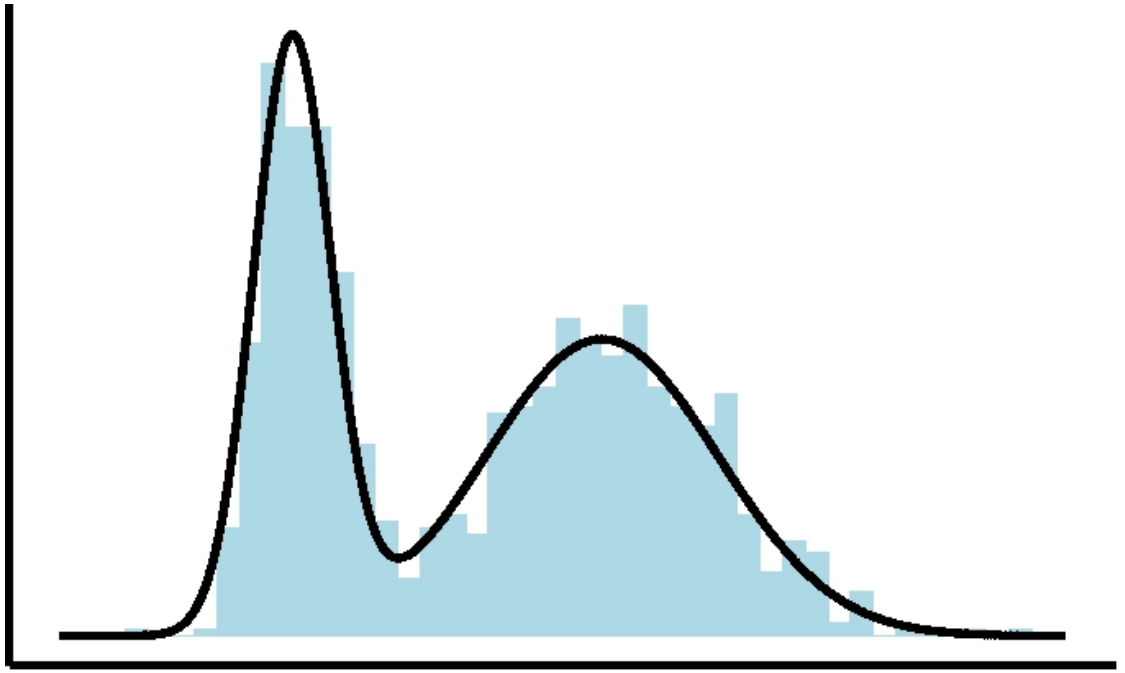
Drawing random samples from probabilities

- in real-world problems probabilities are too complex for visual or analytical analysis
- idea of sampling: pick a set of representatives

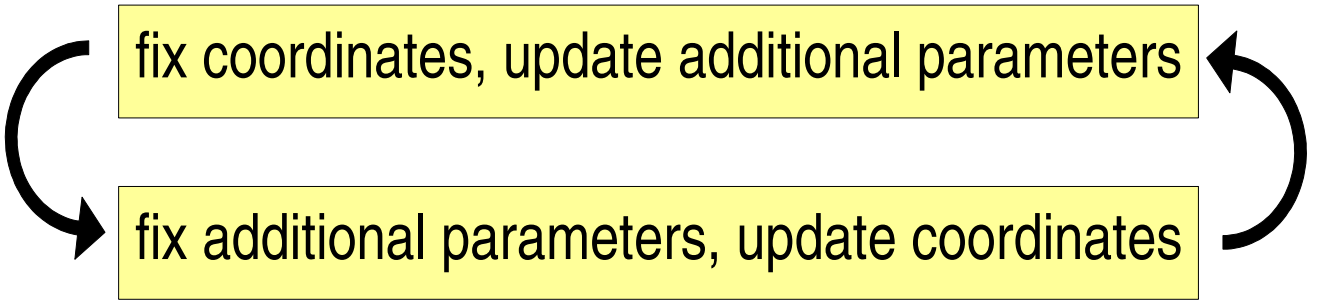


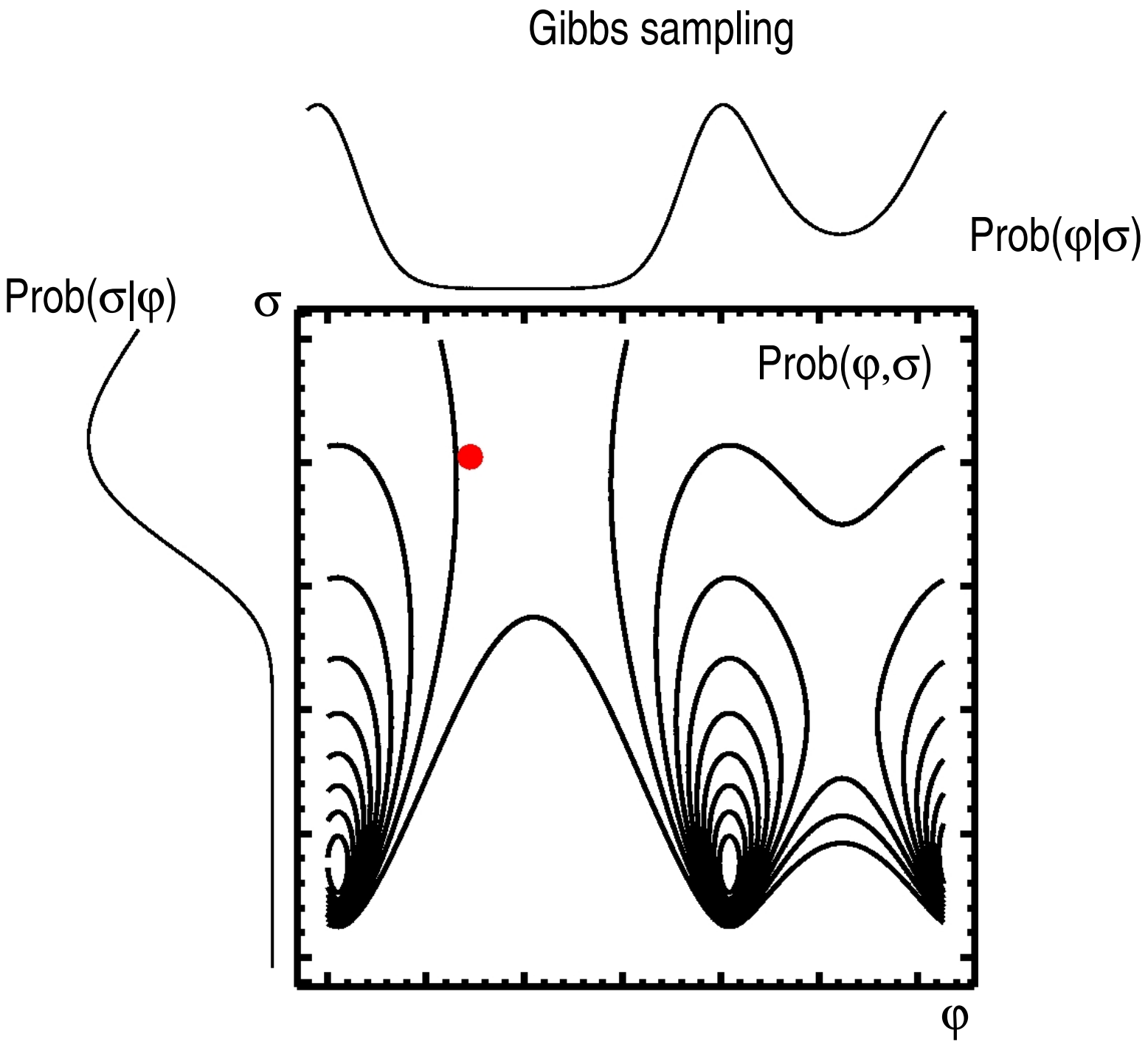
Drawing random samples from probabilities

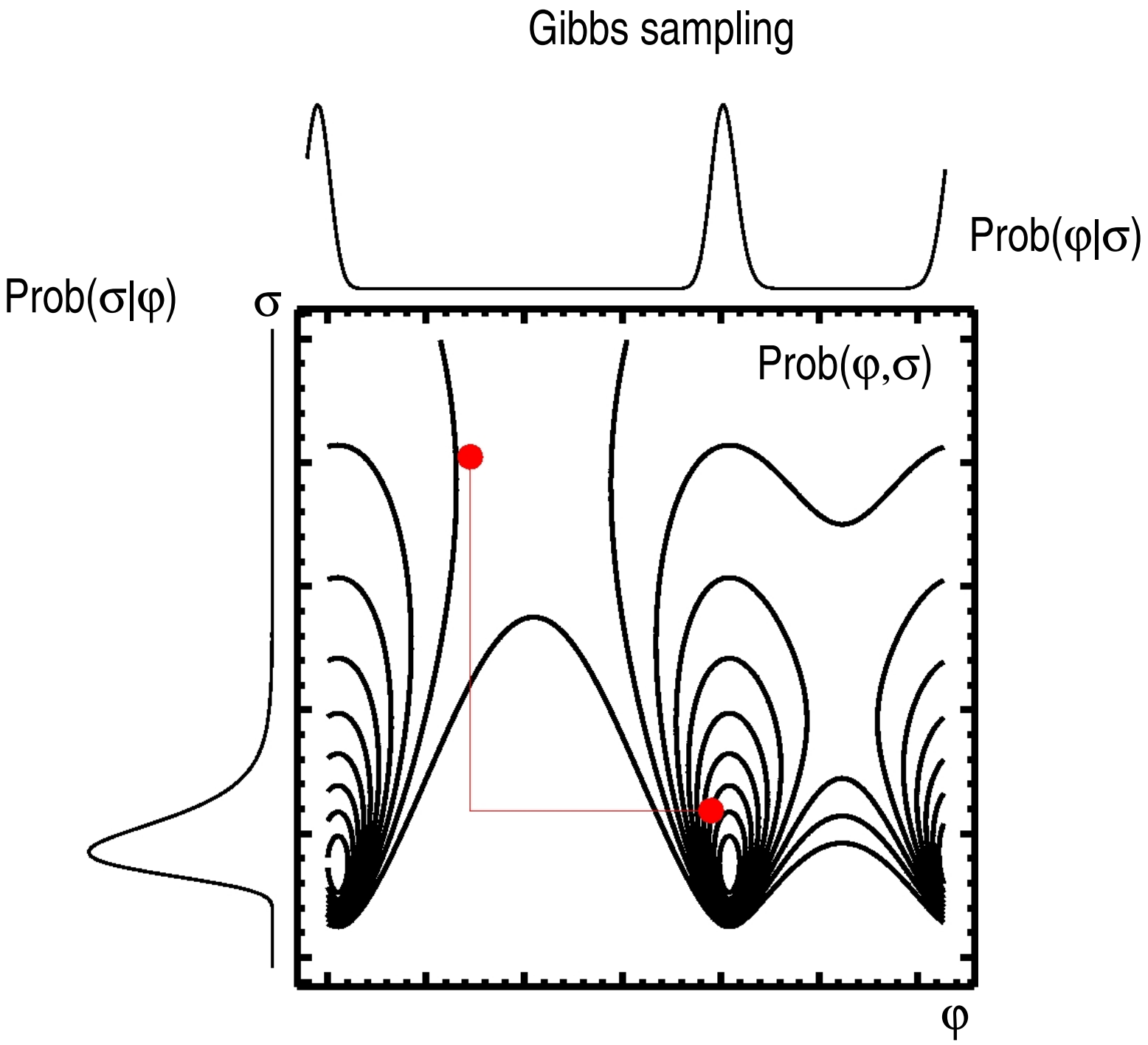
- in real-world problems probabilities are too complex for visual or analytical analysis
- idea of sampling: pick a set of representatives



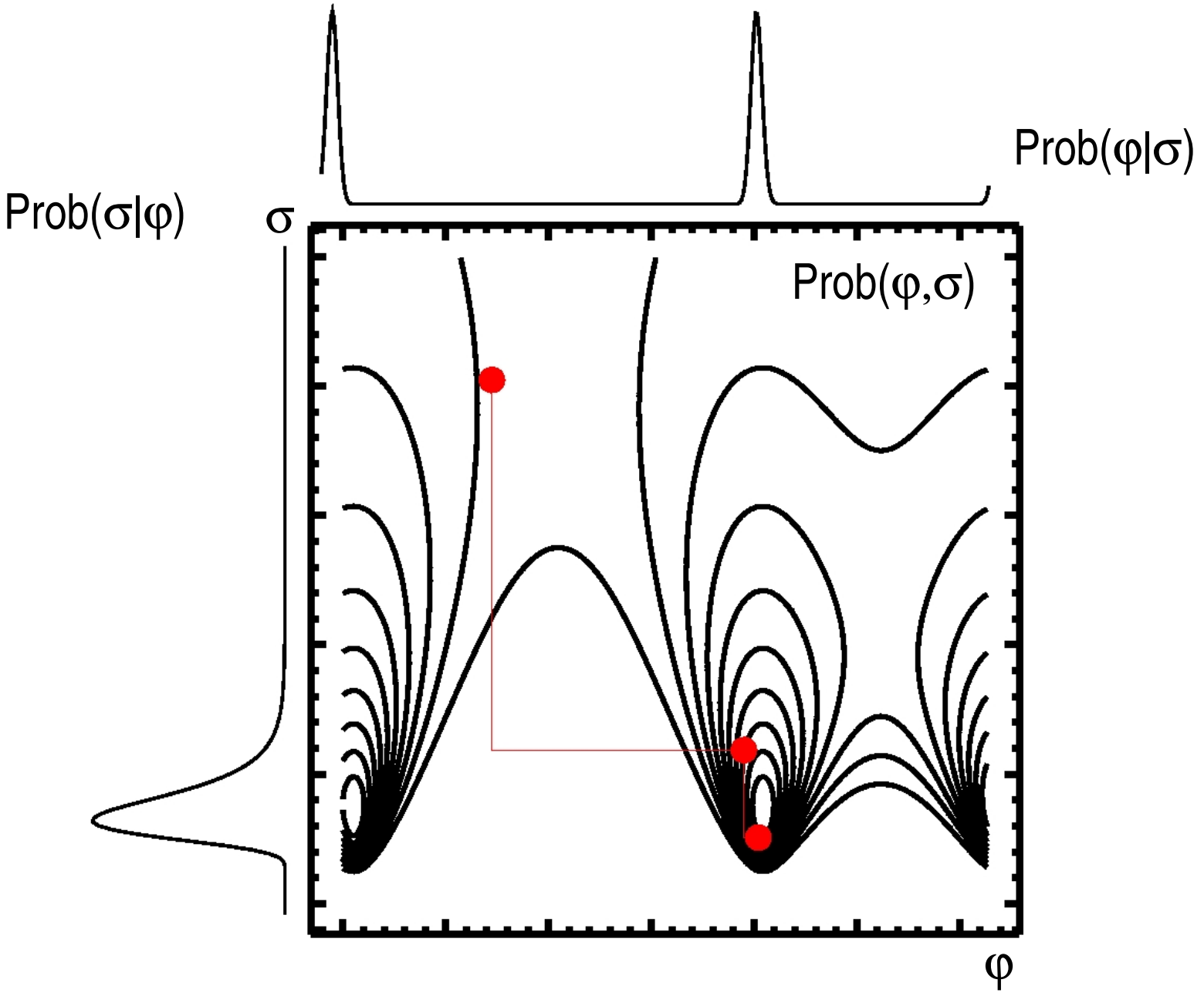
Gibbs sampling



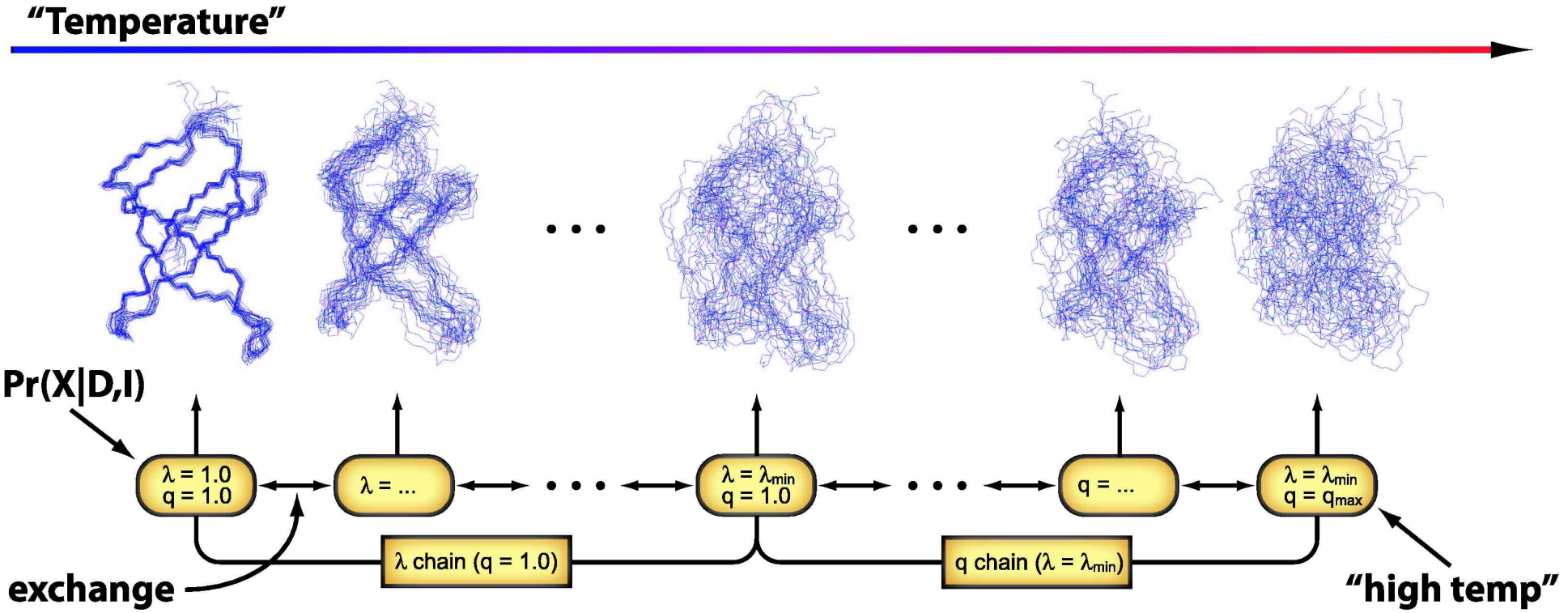

Gibbs sampling
Prob(φ|σ)
Prob(σ|φ)
σ
Prob(φ,σ)
φ


Gibbs sampling
Prob(φ|σ)
Prob(σ|φ)
σ
Prob(φ,σ)
φ

Gibbs sampling

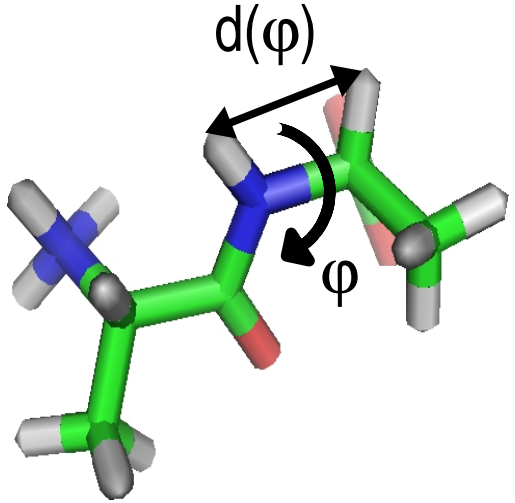


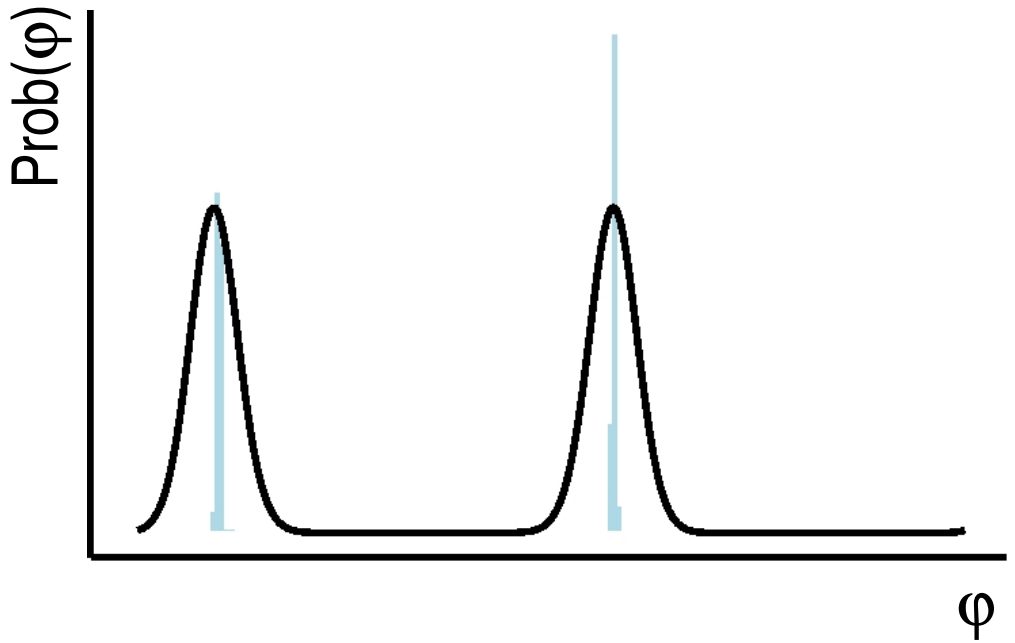
Generation of structure ensembles



Optimization algorithms are not for sampling

- optimization algorithms are often used to “sample” the protein conformation space
- but they are designed to locate (global) optima
- multi-start simulated annealing already fails for example



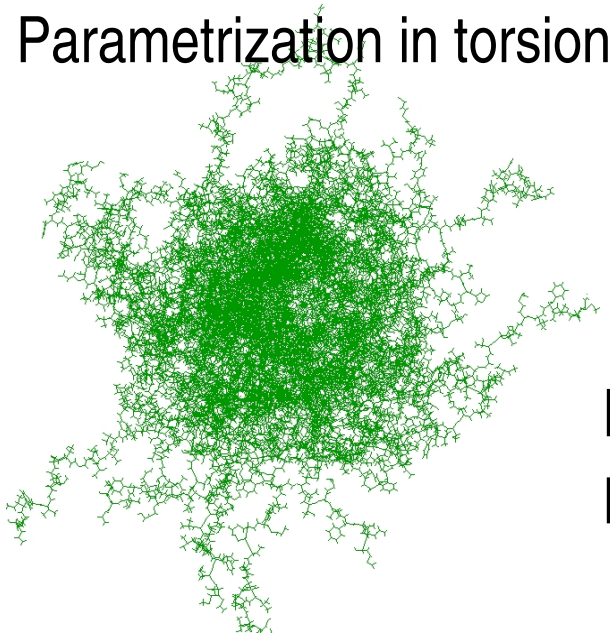


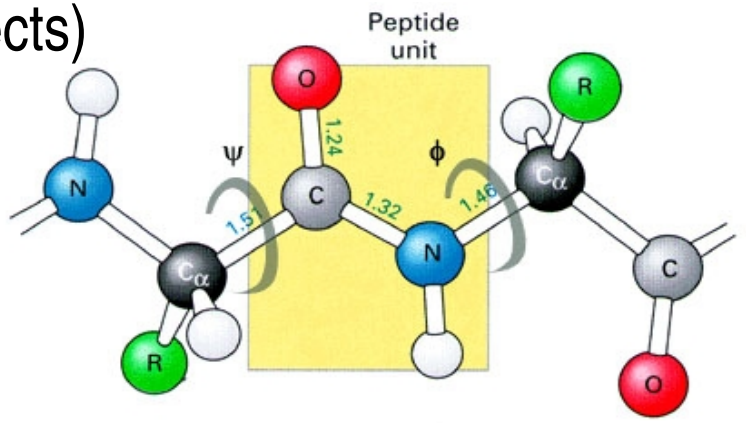
Prior probability of protein conformations

Physical knowledge:

- covalent forces (bond lengths, angles)
- noncovalent forces (van der Waals, electrostatic)
- solvent (hydrophobic forces, entropic effects)

Parametrization in torsion angles





$$\mathsf{Prob}(\mathsf{structure}) \propto \exp\{-\beta \mathsf{E}_{\mathsf{phys}}(\mathsf{structure})\}$$

Prior = Boltzmann ensemble

Probabilistic modelling of NMR data

Principle: imagine a process that could have generated your data

This typically comprises

- a forward model (eg. ISPA, Karplus curve)
- an error model (eg. Gaussian distribution)

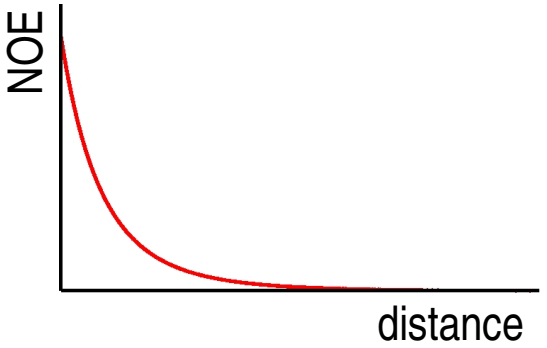
“nuisance” parameters:

- model parameters (eg. A, B, C in Karplus curve)
- error parameters (eg. width σ of the Gaussian)

Modelling NOEs

forward model: ISPA

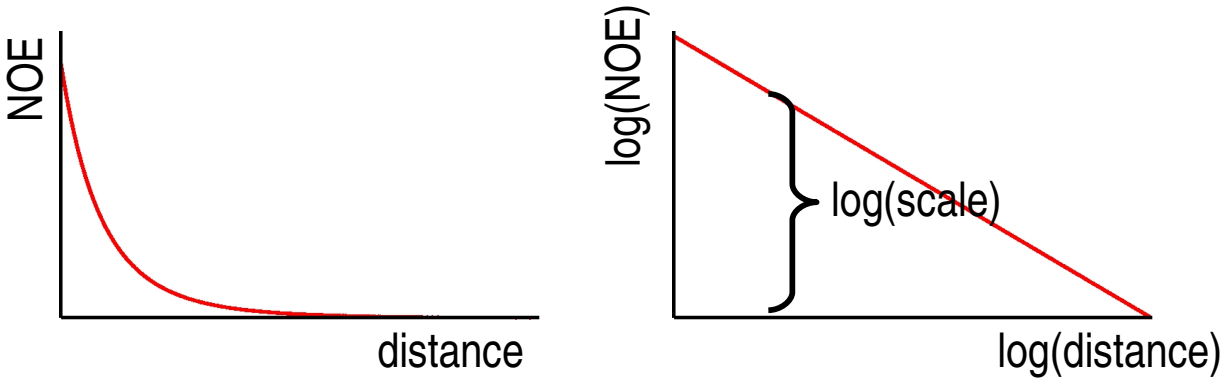
- NOE = scale / distance^6



Modelling NOEs

forward model: ISPA

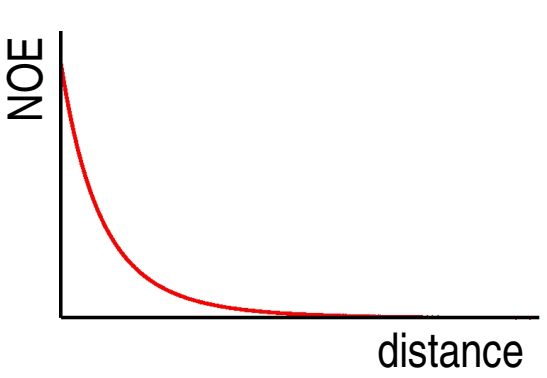
- NOE = scale / distance^6
- log(NOE) = log(scale) – 6 log(distance)

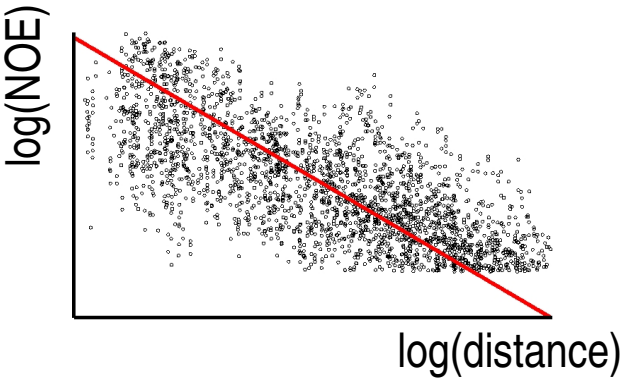


Modelling NOEs

forward model: ISPA

- NOE = scale / distance^6
- log(NOE) = log(scale) – 6 log(distance)

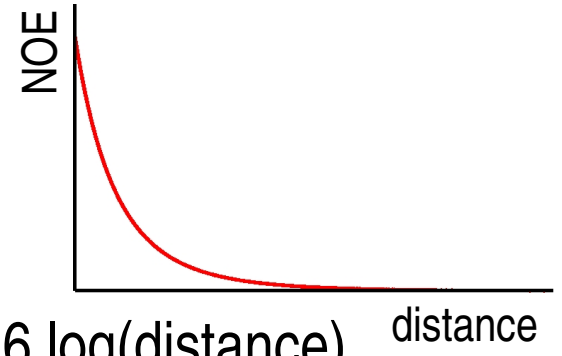


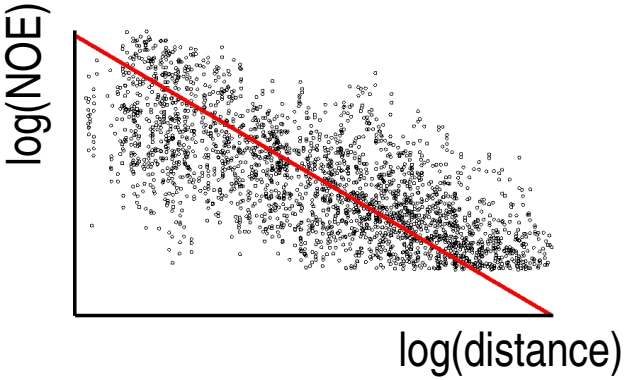


Modelling NOEs

forward model: ISPA

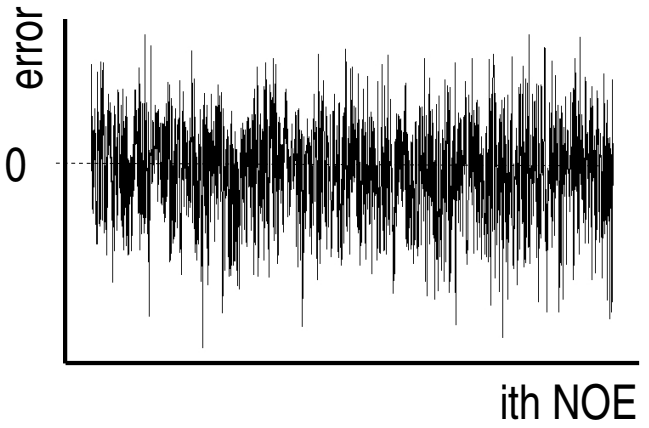
- NOE = scale / distance^6
- log(NOE) = log(scale) – 6 log(distance)





error model: log-normal

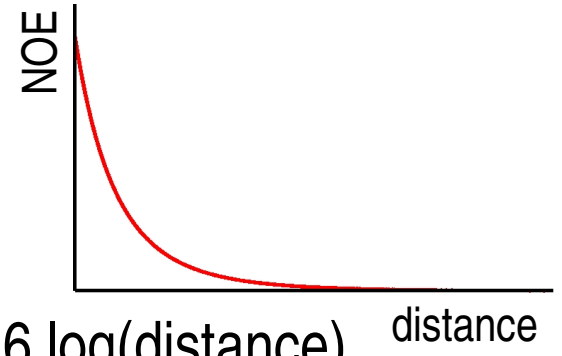
- error = log(NOE) - log(scale) + 6 log(distance)

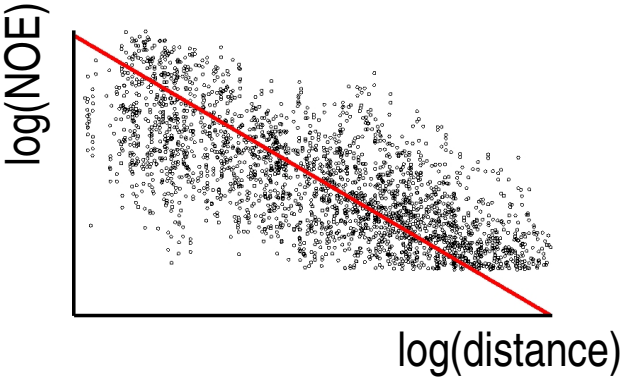


Modelling NOEs

forward model: ISPA

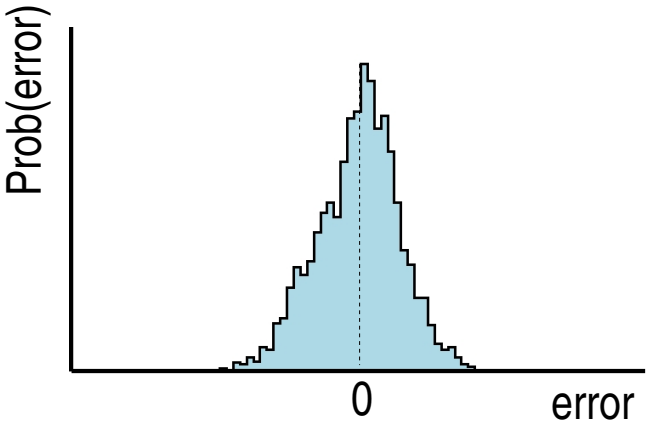
- NOE = scale / distance^6
- log(NOE) = log(scale) – 6 log(distance)





error model: log-normal

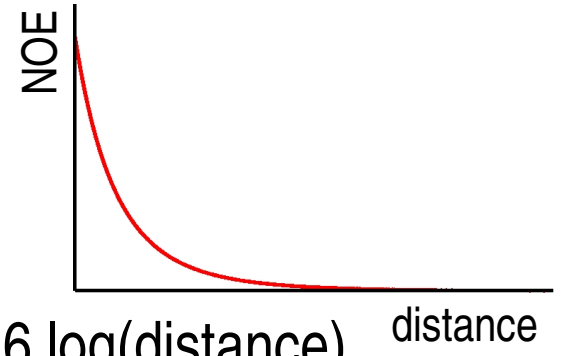
- error = log(NOE) - log(scale) + 6 log(distance)

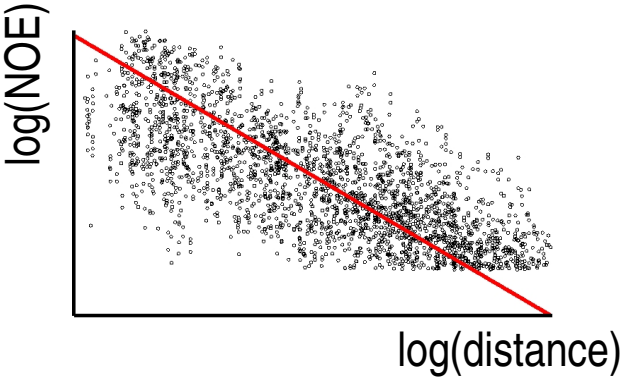


Modelling NOEs

forward model: ISPA

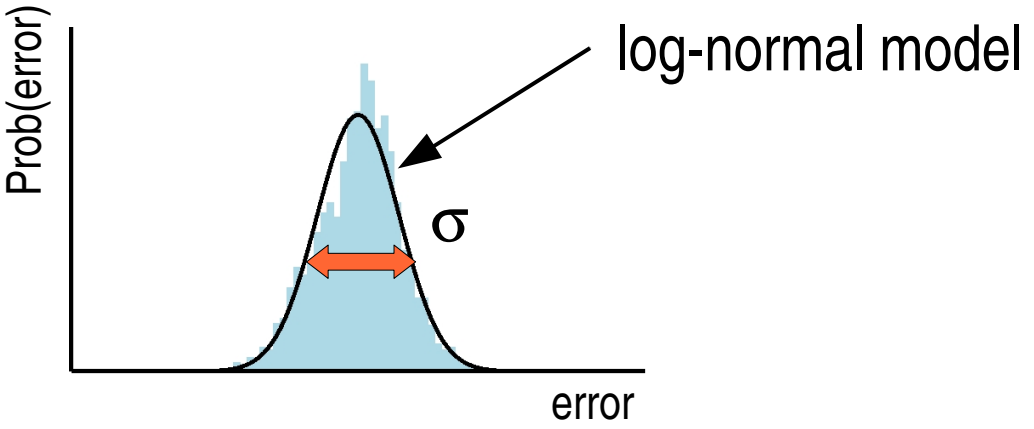
- NOE = scale / distance^6
- log(NOE) = log(scale) – 6 log(distance)





error model: log-normal

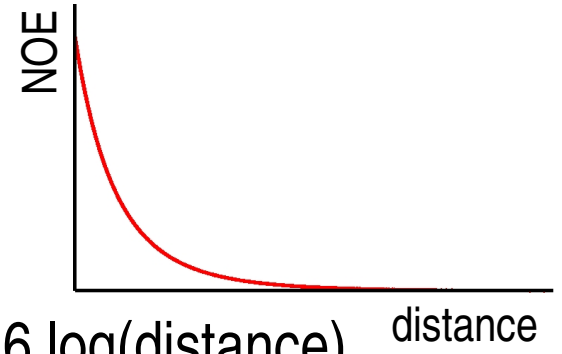
- error = log(NOE) - log(scale) + 6 log(distance)

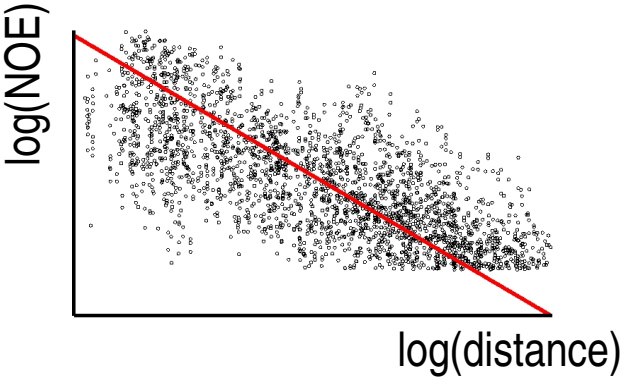


Modelling NOEs

forward model: ISPA

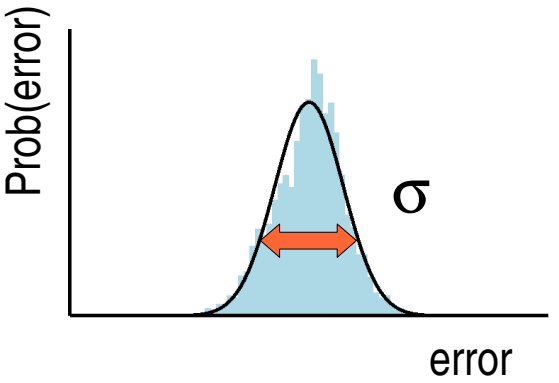
- NOE = scale / distance^6
- log(NOE) = log(scale) – 6 log(distance)



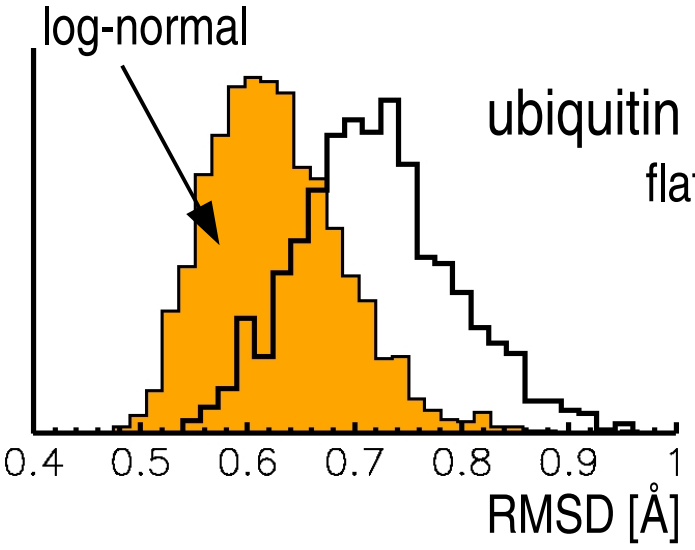


error model: log-normal

- error = log(NOE) - log(scale) + 6 log(distance)



Improvement in accuracy and quality

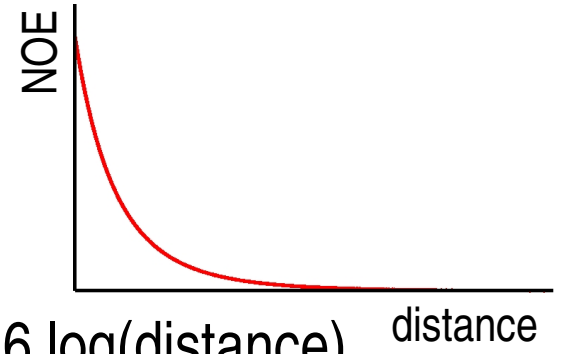


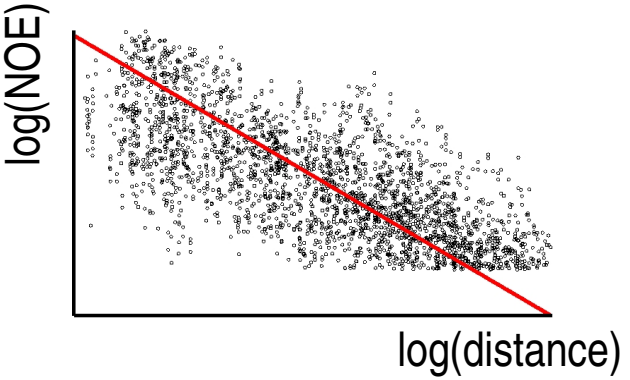
JACS, 2005

Modelling NOEs

forward model: ISPA

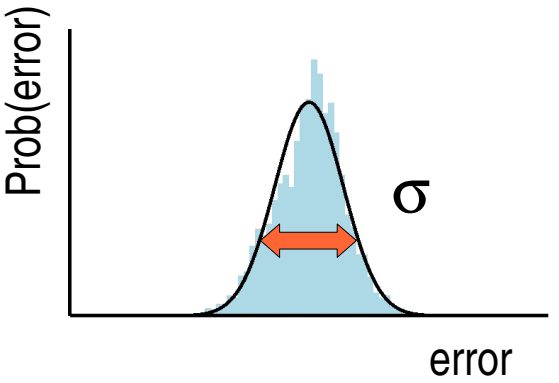
- NOE = scale / distance^6
- log(NOE) = log(scale) – 6 log(distance)



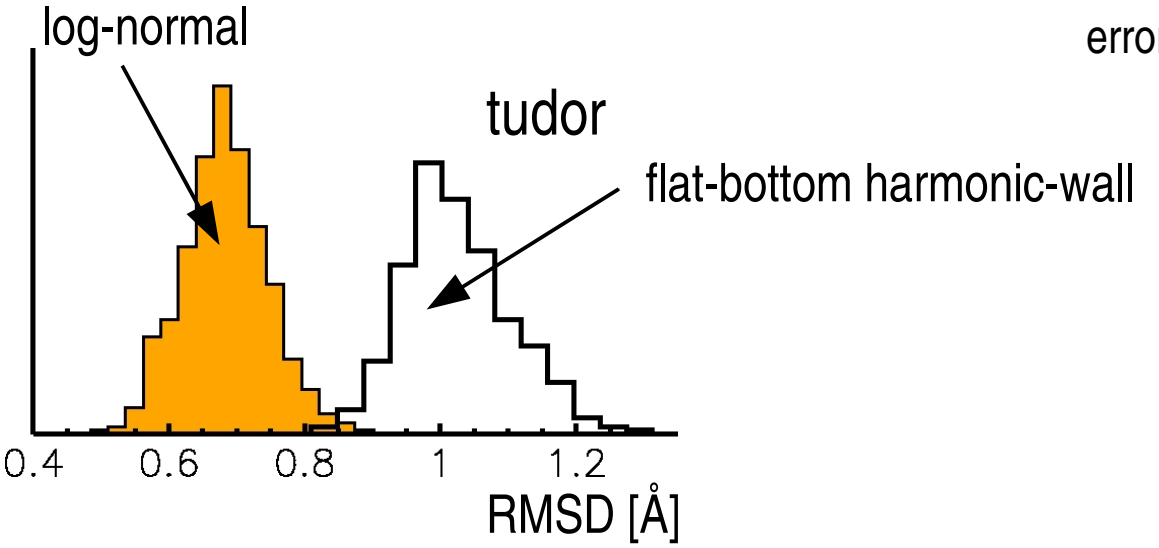


error model: log-normal

- error = log(NOE) - log(scale) + 6 log(distance)



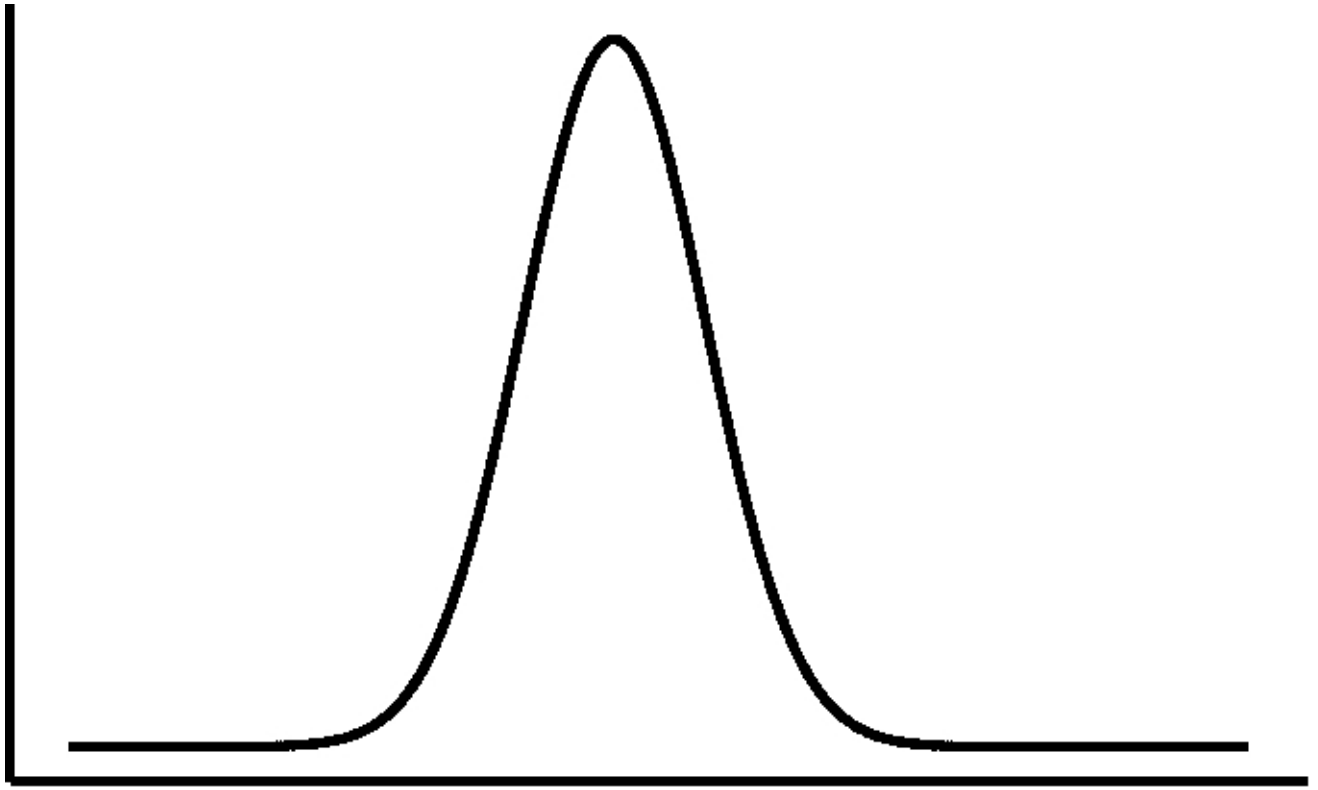
Improvement in accuracy and quality



JACS, 2005

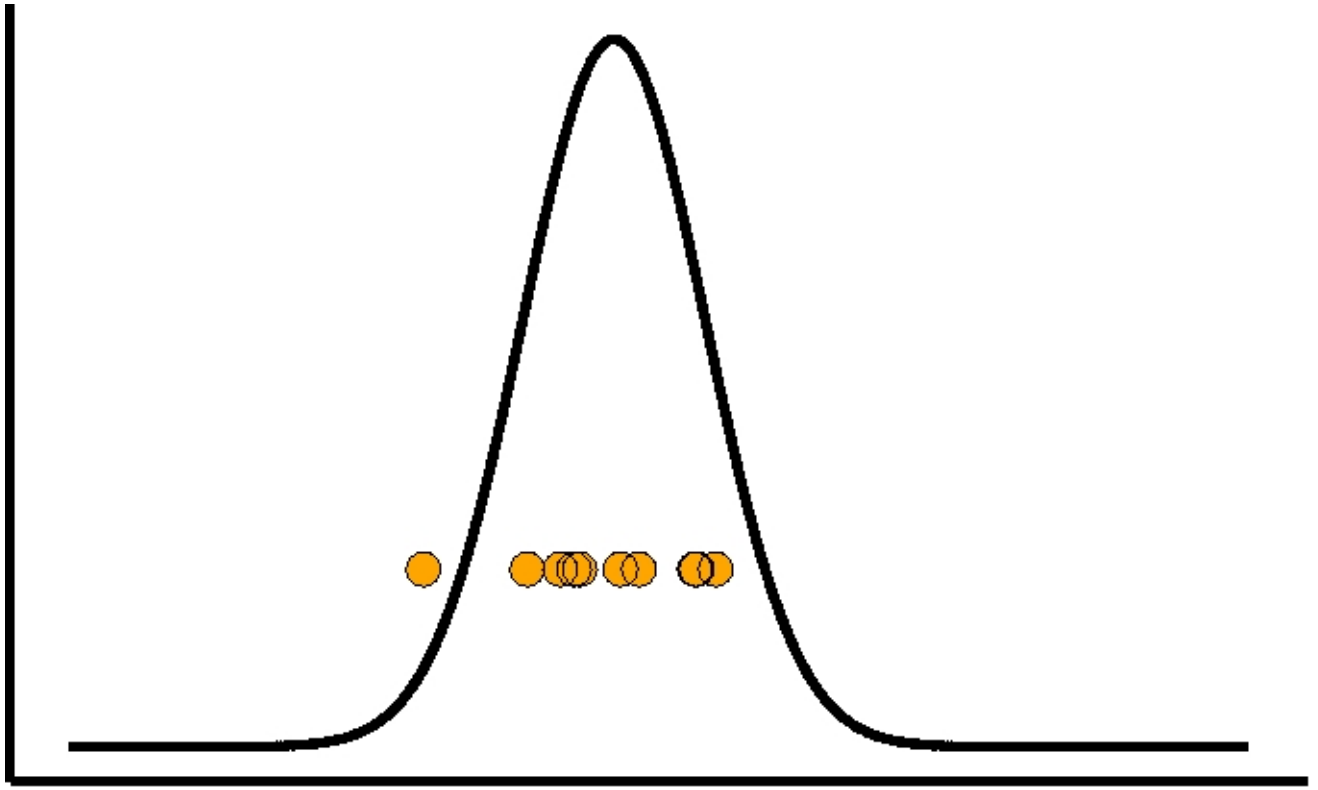
Error bars for NMR structures

Idea: calculate average and variance from random samples



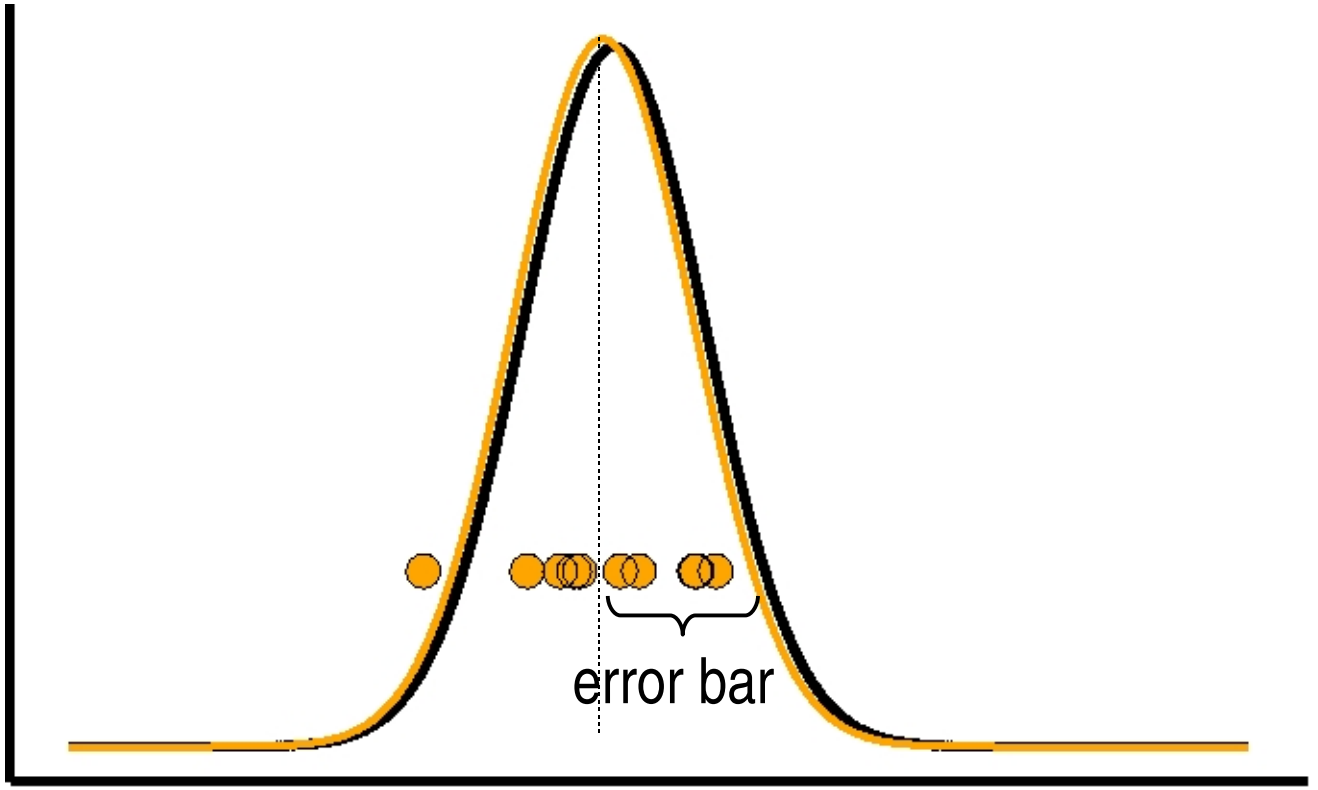
Error bars for NMR structures

Idea: calculate average and variance from random samples

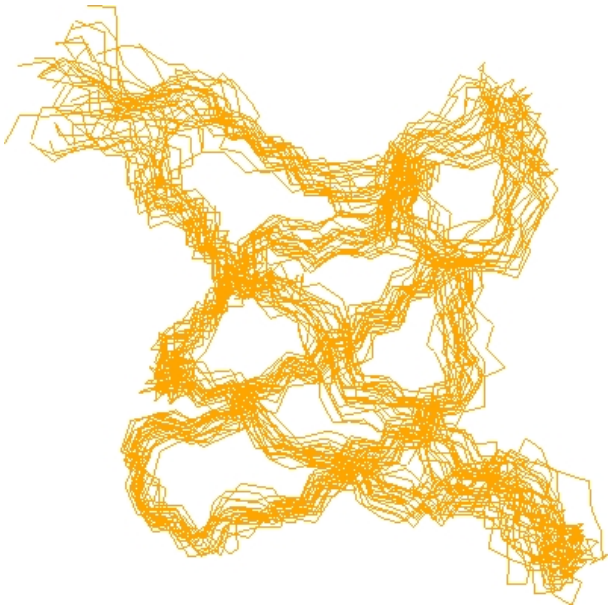


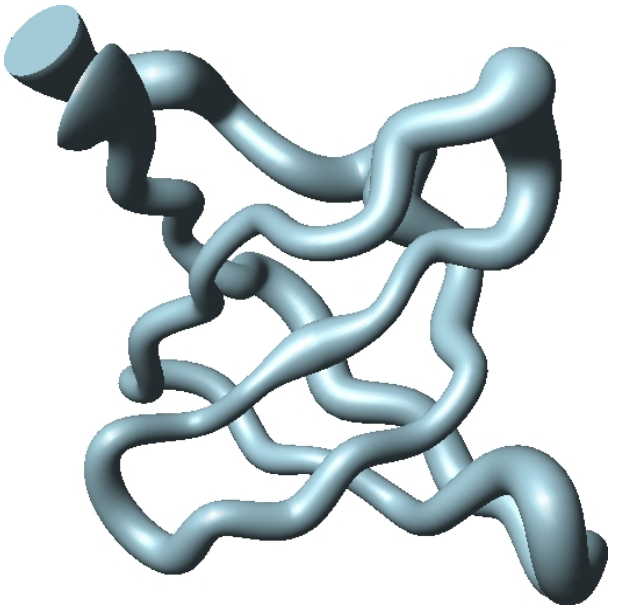
Error bars for NMR structures

Idea: calculate average and variance from random samples



Error bars for NMR structures

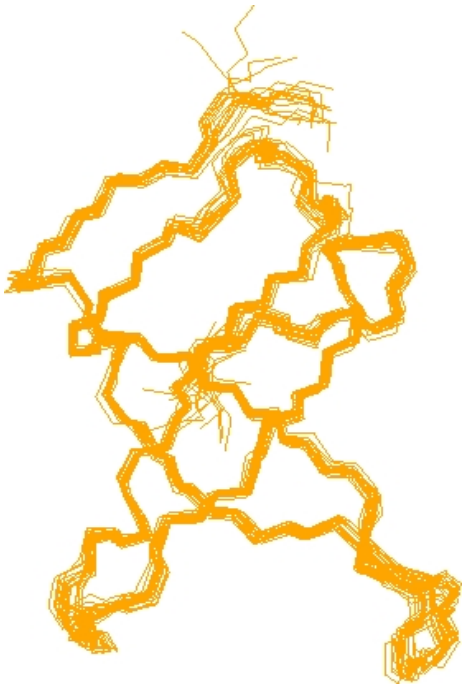


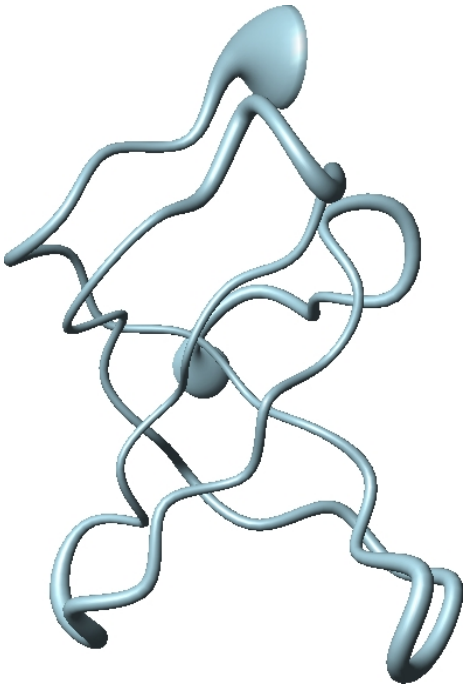


SH3, 154 distances

Science, 2005

Error bars for NMR structures





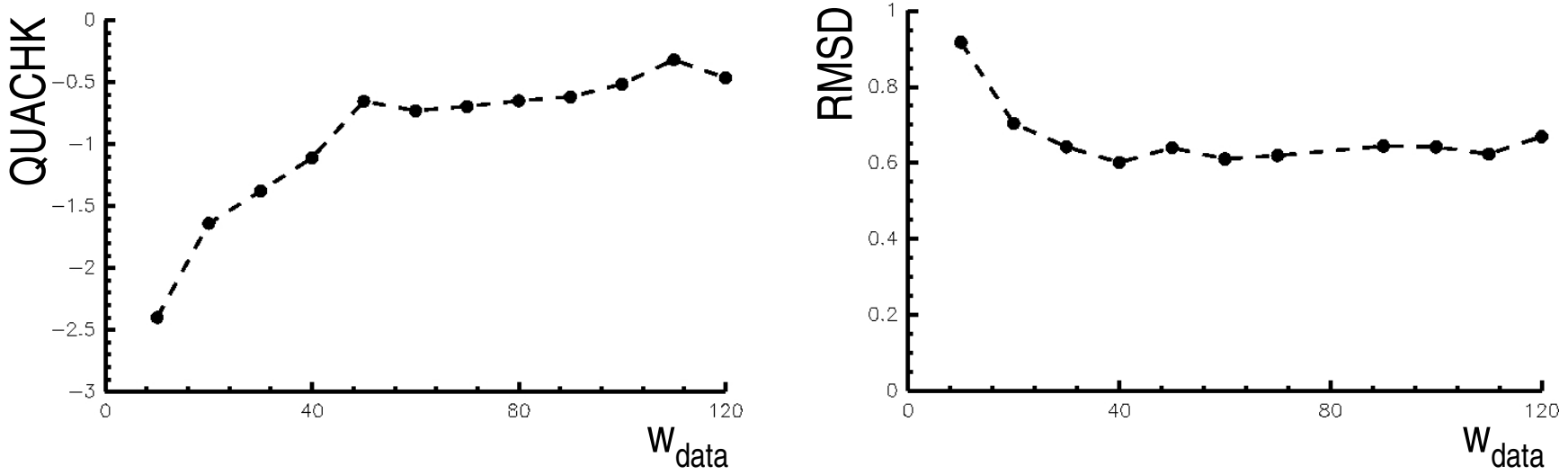
TUDOR, 1875 distances

Phys Rev E, 2005; JACS 2005

Adaptive weighting of data

Standard approach: $E_{hybrid} = w_{data} E_{data} + E_{phys}$

Choice of weight is critical:



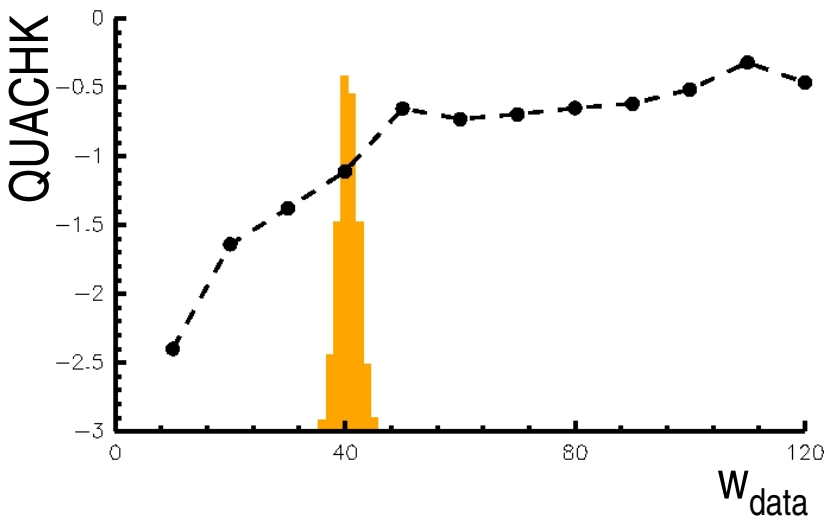
Probabilistic interpretation: $w_{data} = 1/\sigma^2$

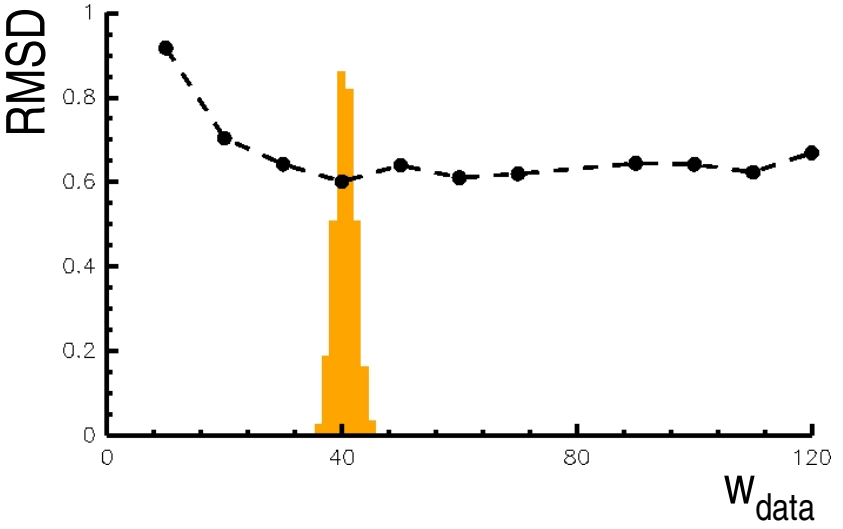
PNAS, 2006

Adaptive weighting of data

Standard approach: $E_{hybrid} = w_{data} E_{data} + E_{phys}$

Choice of weight is critical:

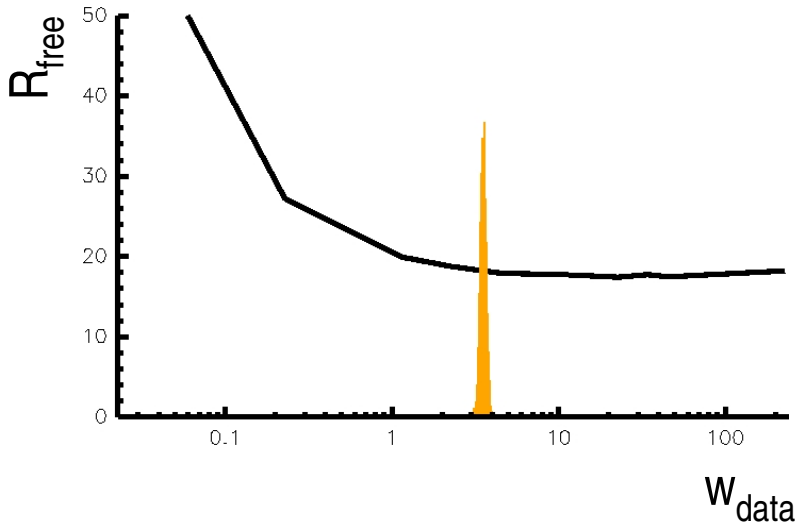


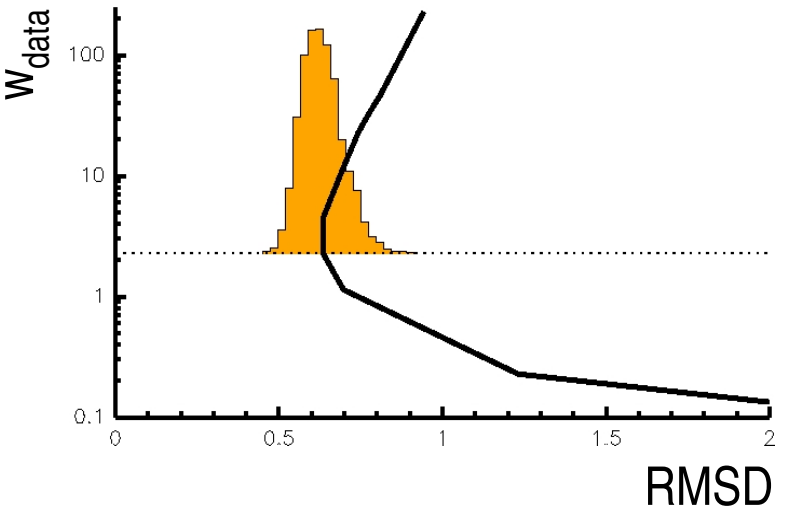


Probabilistic interpretation: $w_{data} = 1/\sigma^2$

PNAS, 2006

Bayes vs. crossvalidation

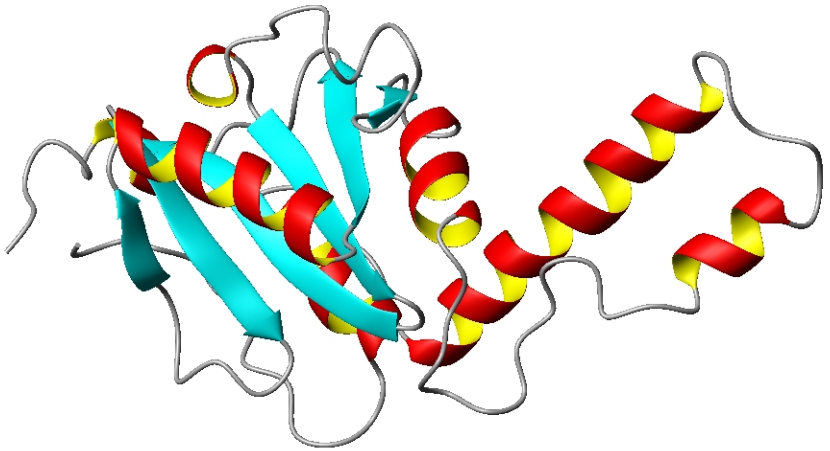




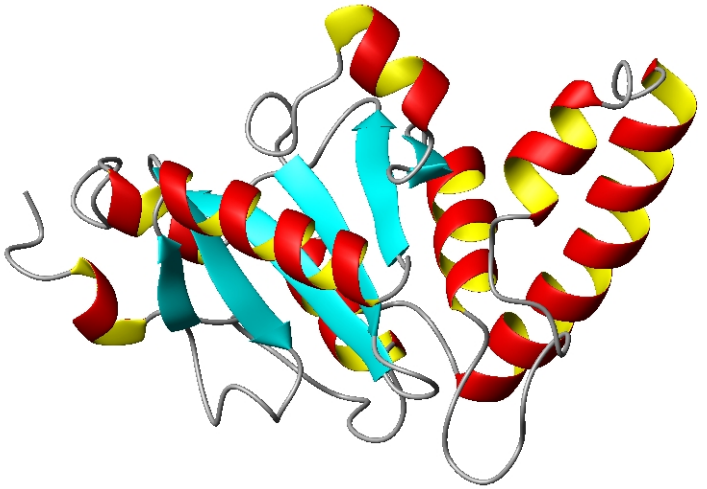
PNAS, 2006

Bayesian error as figure of merit

2 NMR structures of *Josephin*

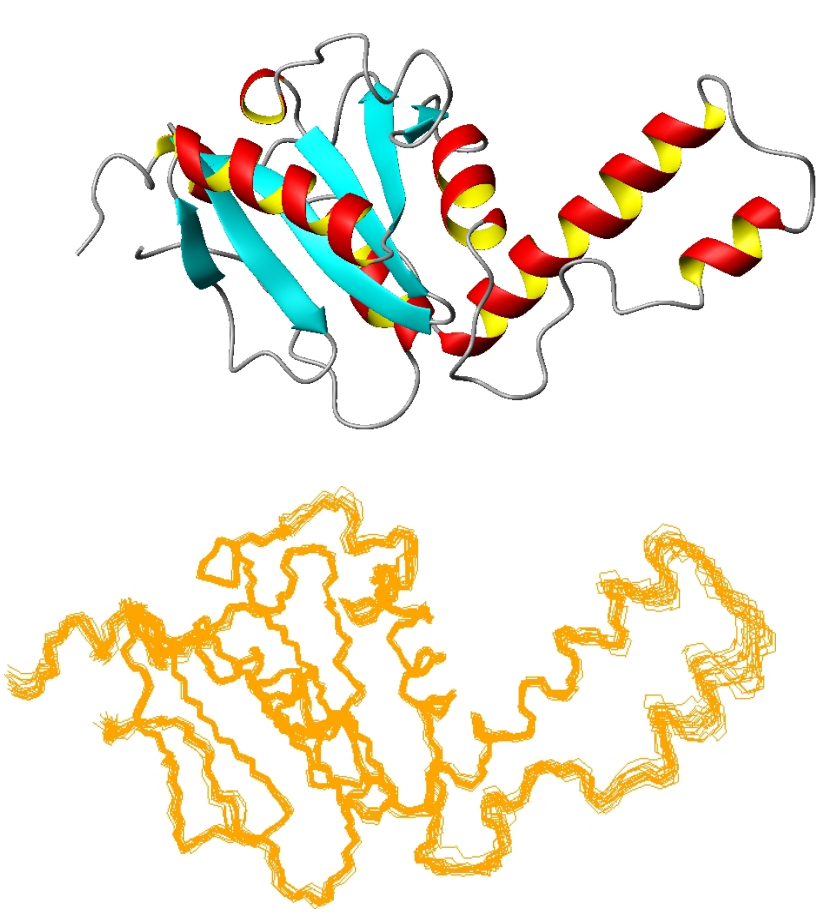


1yzb

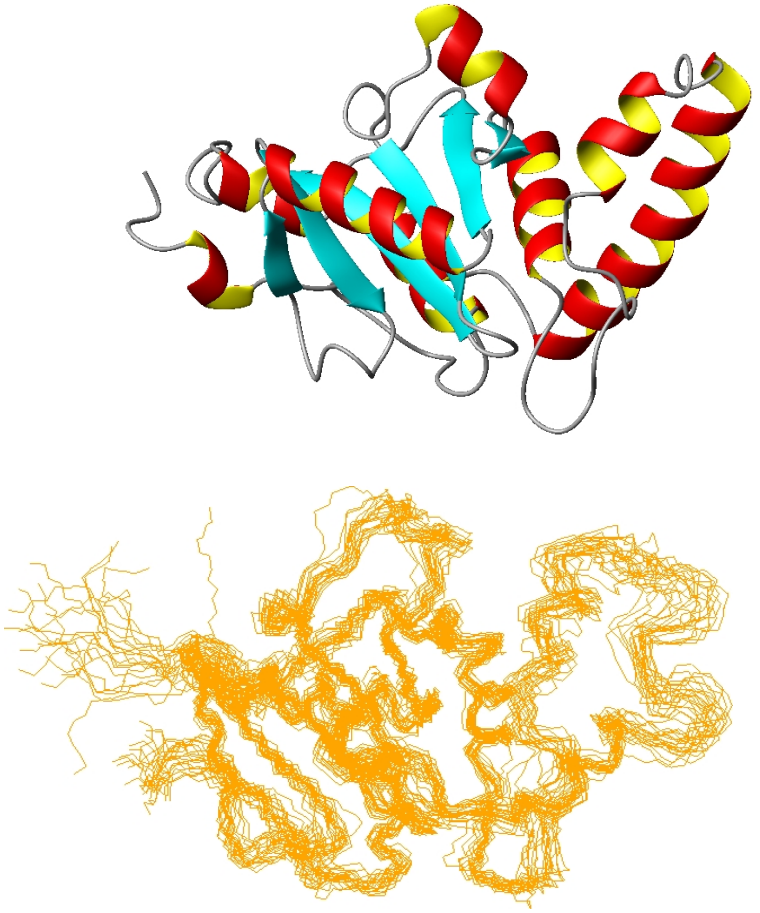


2aga

Bayesian error as figure of merit

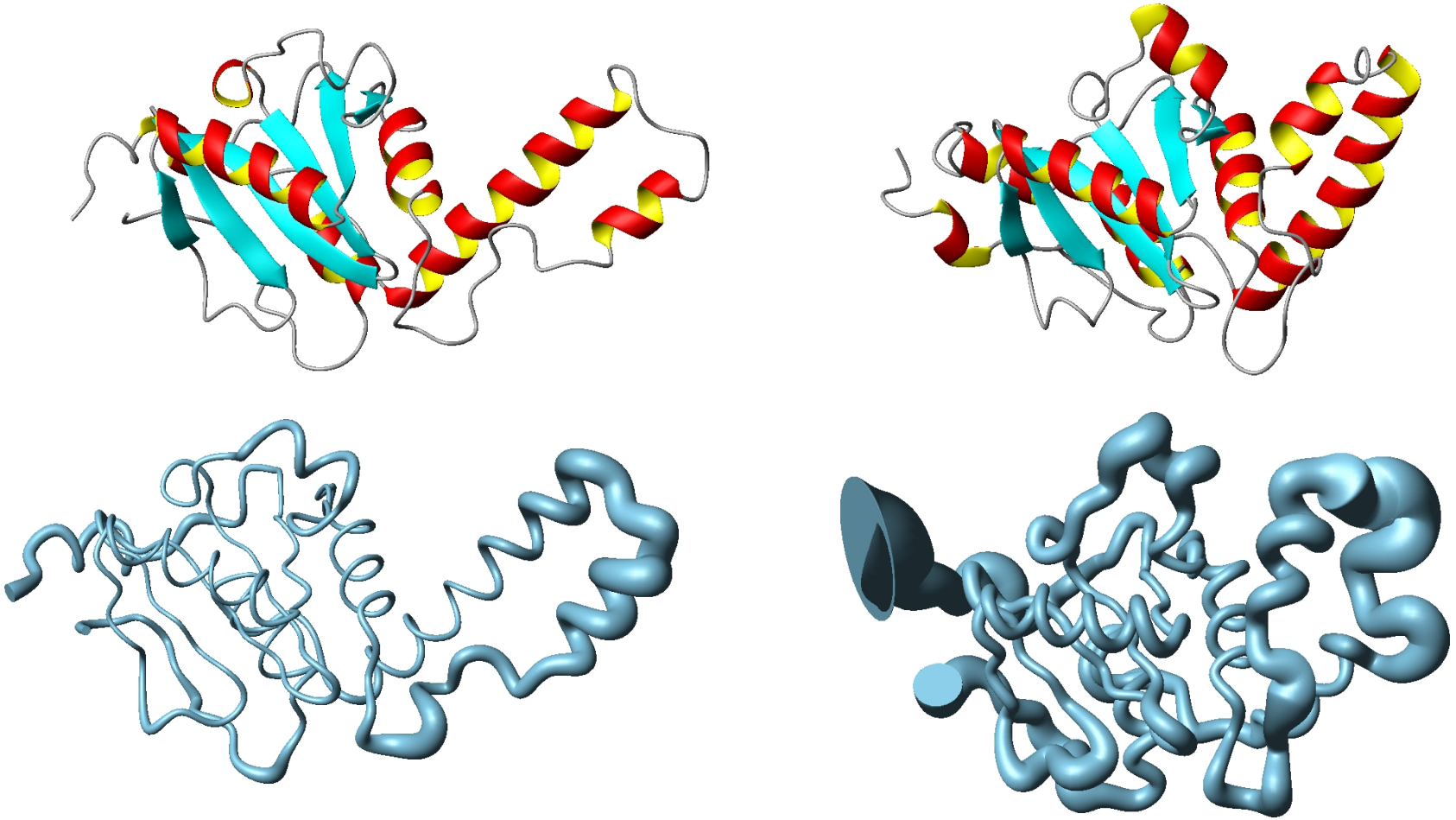


1yzb (5525 + 925)



2aga (2960)

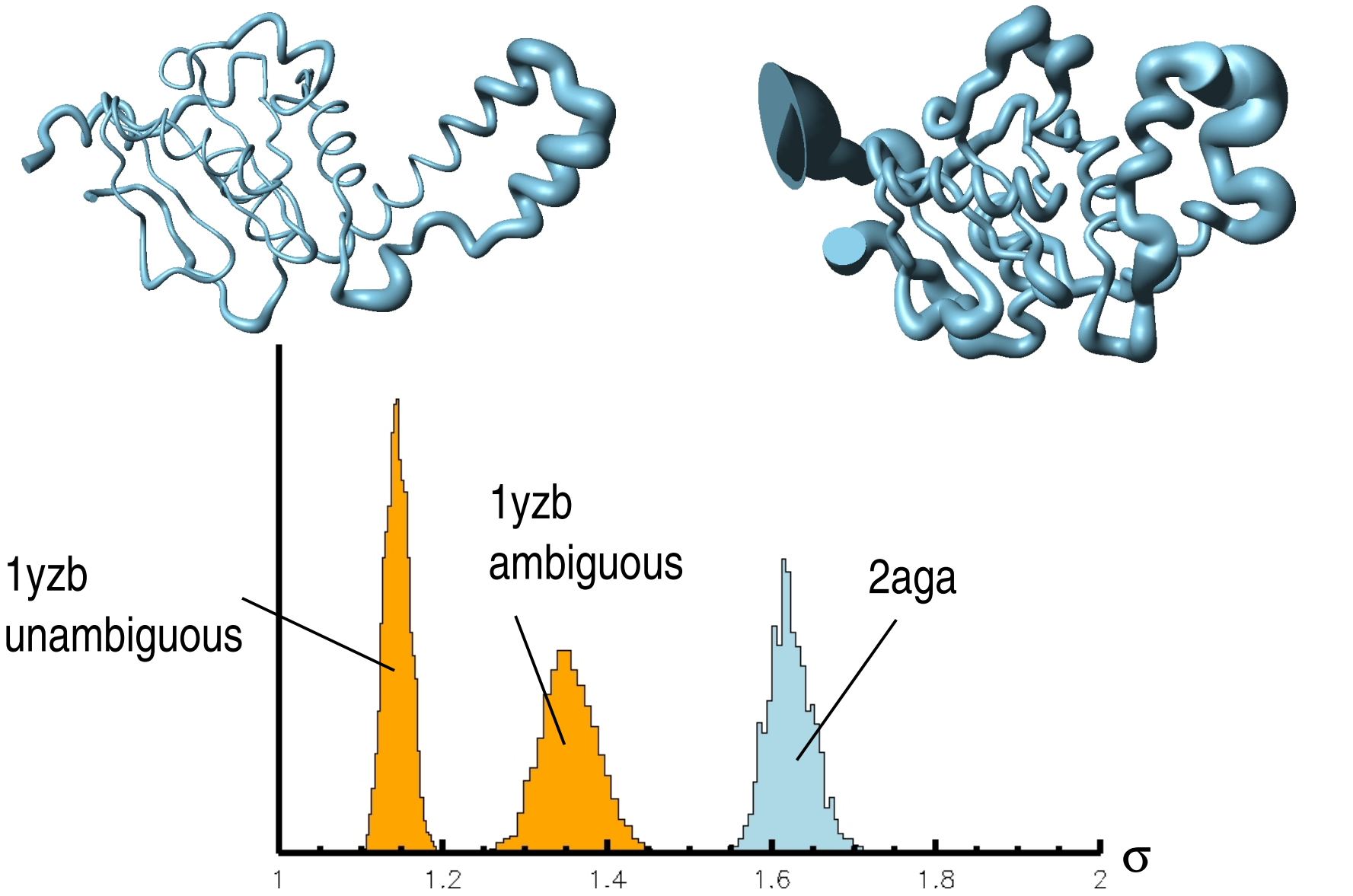
Bayesian error as figure of merit



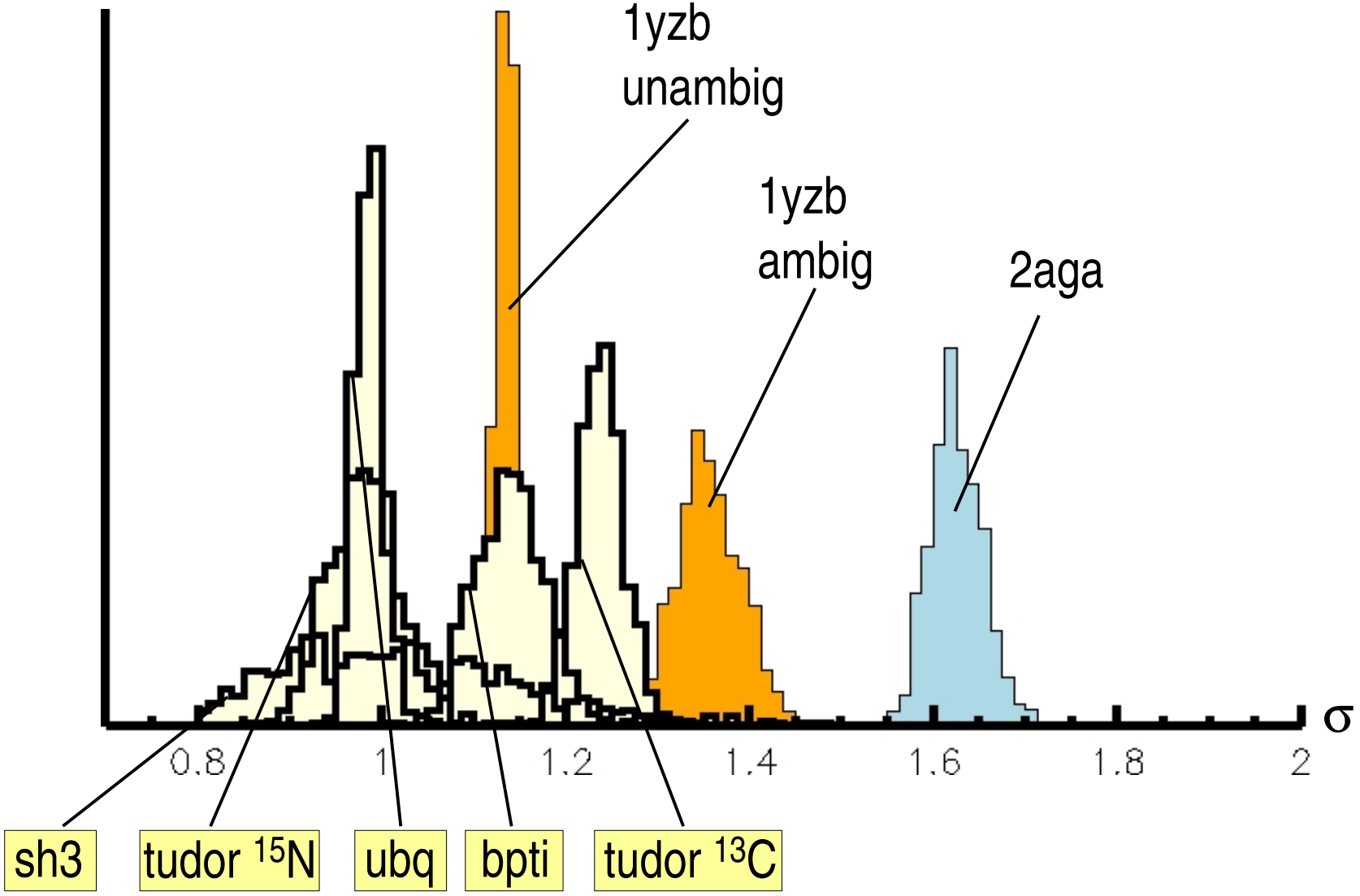
1yzb (5525 + 925)

2aga (2960)

Bayesian error as figure of merit



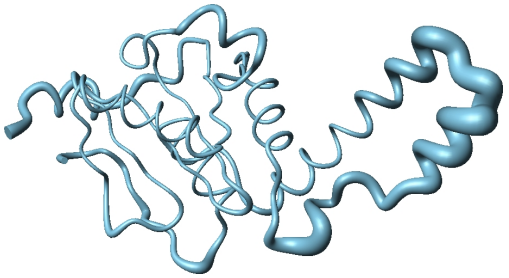
Bayesian error as figure of merit



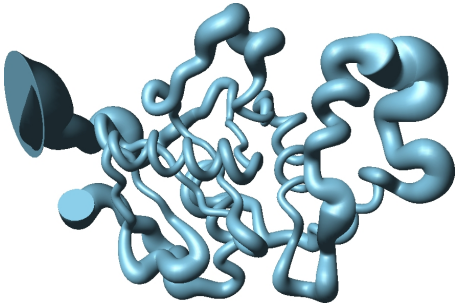
$$\sigma = \langle \text{restraint RMS} \rangle_{\text{unbiased}} \propto \sqrt{R_{\text{free}}}$$

Bayesian error as figure of merit

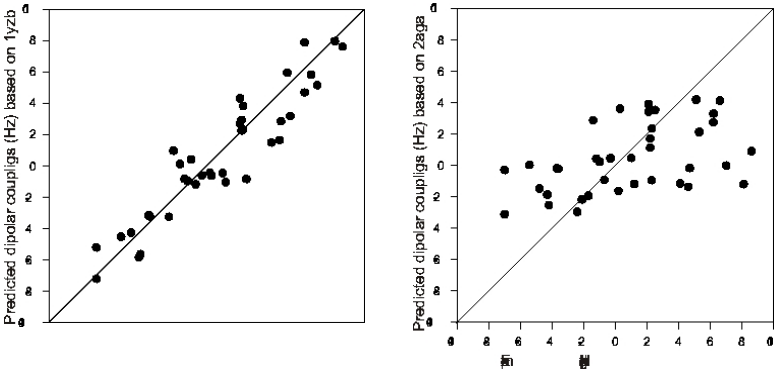
1yzb (5525 + 925)

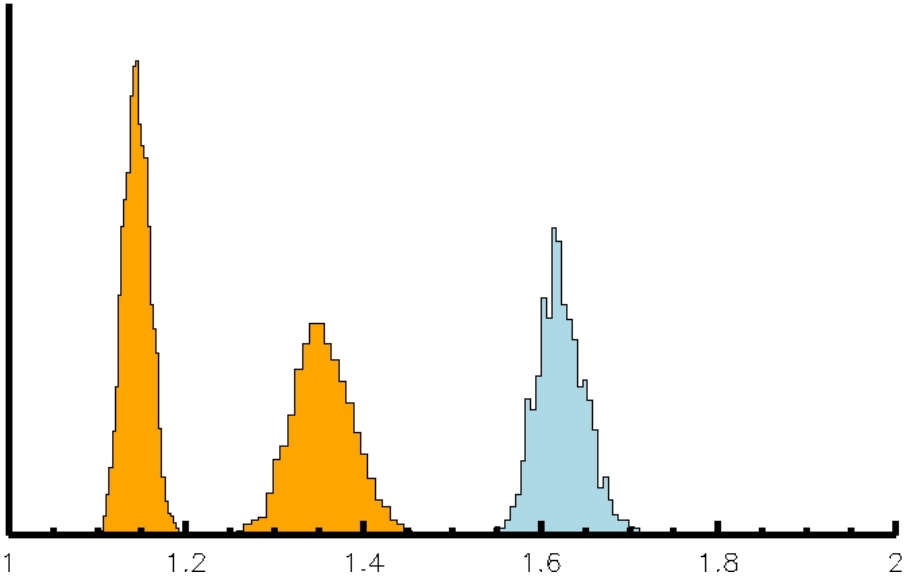


2aga (2960)



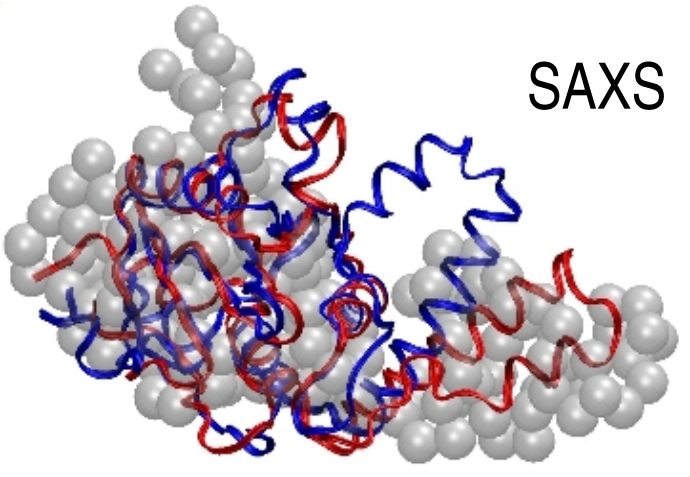
independent exp. evidence





RDCs

SAXS



Summary

- uncertainties in NMR structure determination must be treated probabilistically
- structure calculation by posterior sampling
- ensemble of sampled structures is statistically meaningful
- model parameters can be estimated (eg. NOE scale, Karplus parameters, alignment tensor)
- error parameters can be estimated (effectively: adaptive weighting of the data)
- estimated errors are useful figures of merit and could replace free R values

Acknowledgement

Andrei Lupas (MPI for Developmental Biology, Tübingen)

Bernhard Schölkopf (MPI for Biological Cybernetics, Tübingen)

Annalisa Pastore (NIMR, London)

Michael Nilges (Institut Pasteur, Paris)

Wolfgang Rieping (University of Cambridge)