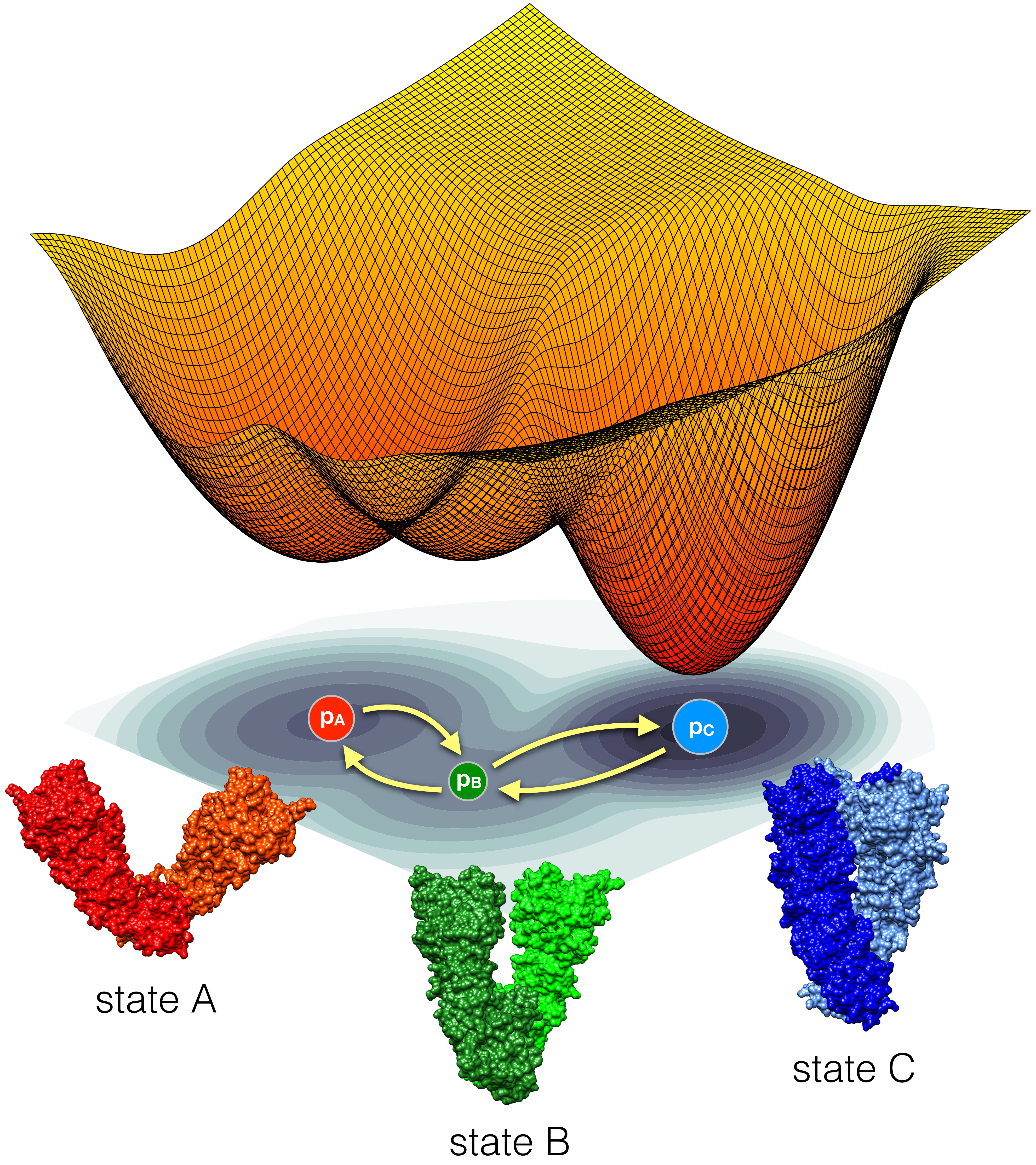
A Bayesian approach to integrate cryo-electron microscopy data into MD simulations with PLUMED

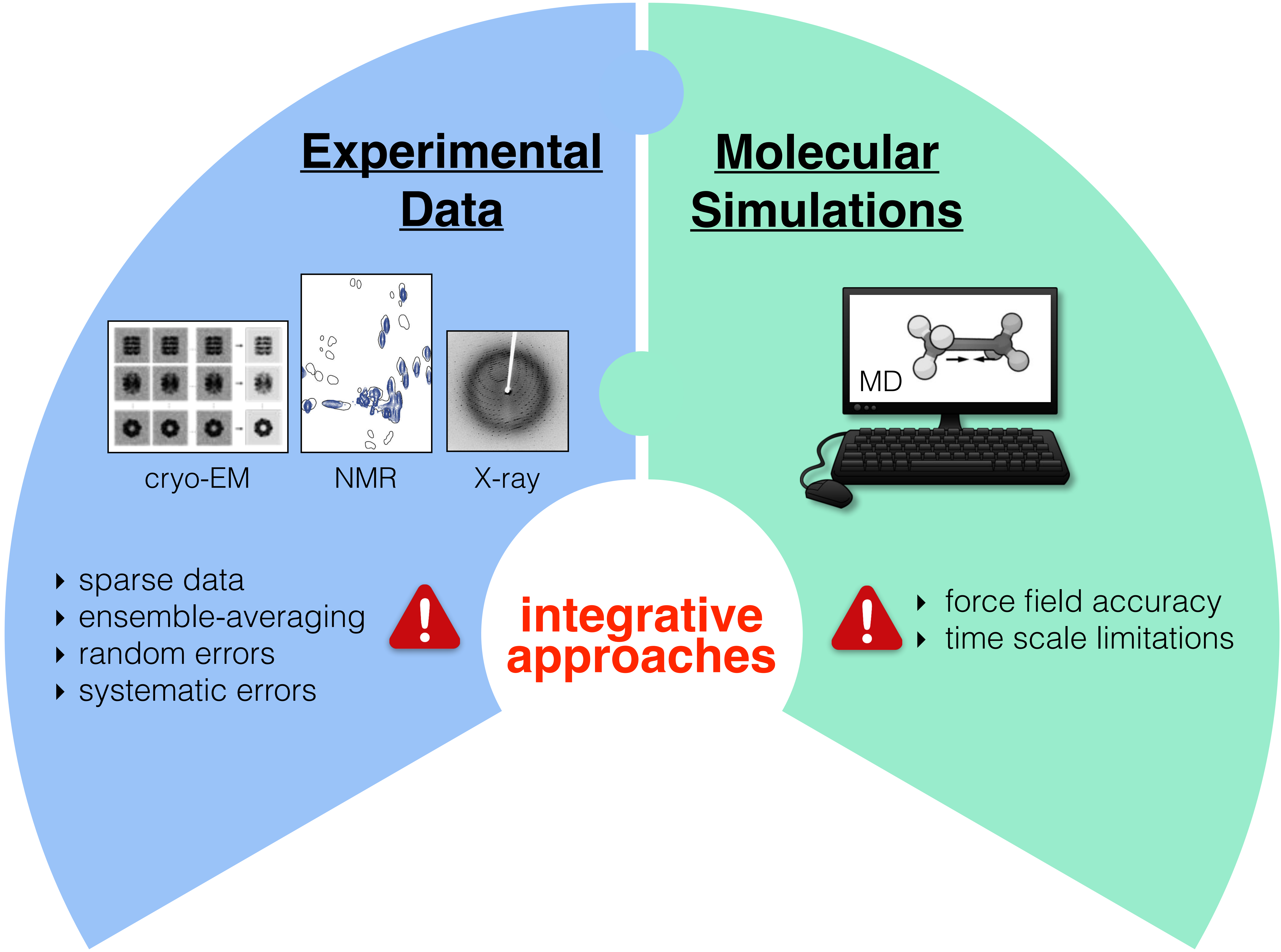
Max Bonomi
Institut Pasteur - CNRS
mbonomi@pasteur.fr

@BonomiMax

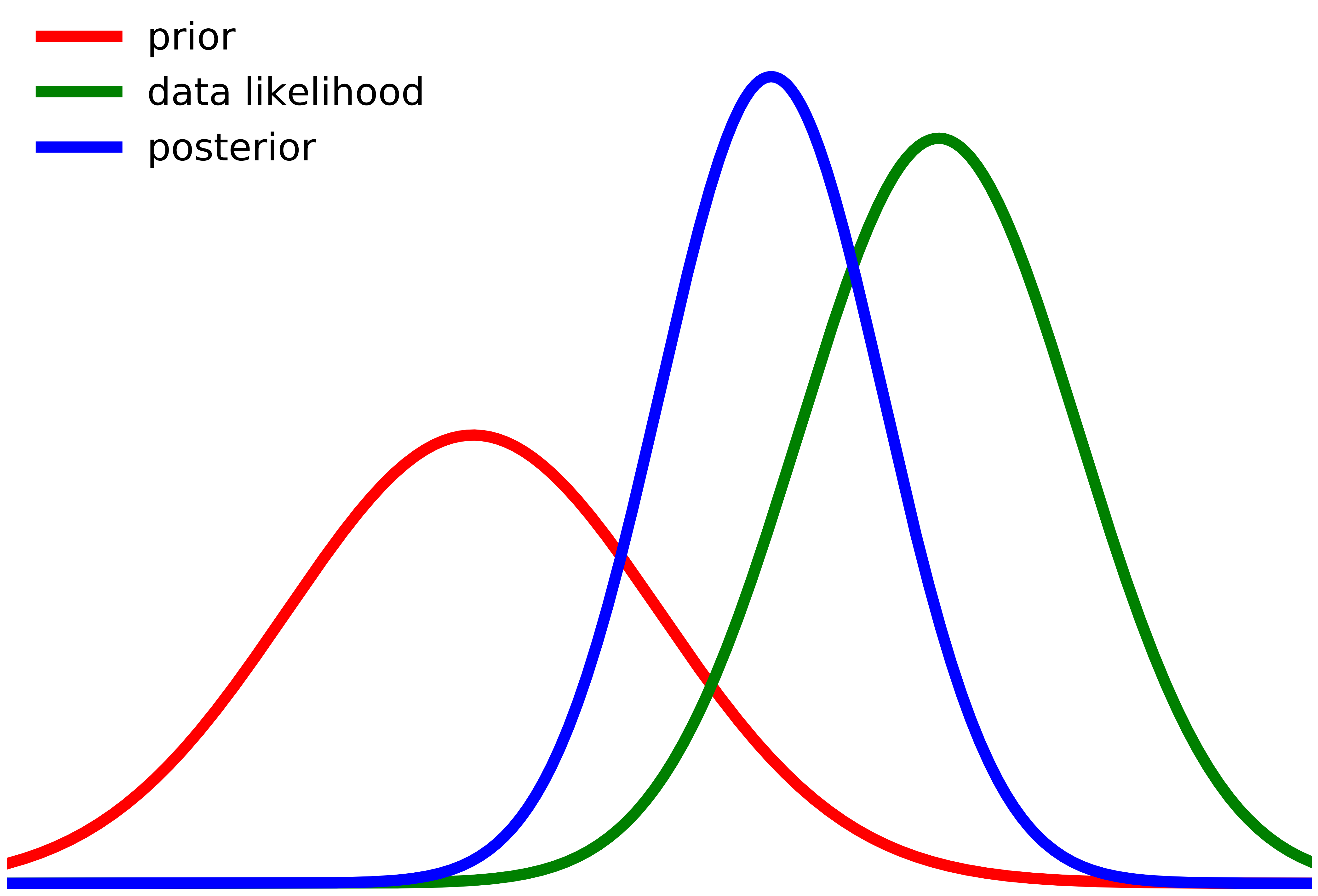
Biological functions from structure and dynamics



The best of both worlds: integrative or *hybrid* approaches



How? A little help from Bayes



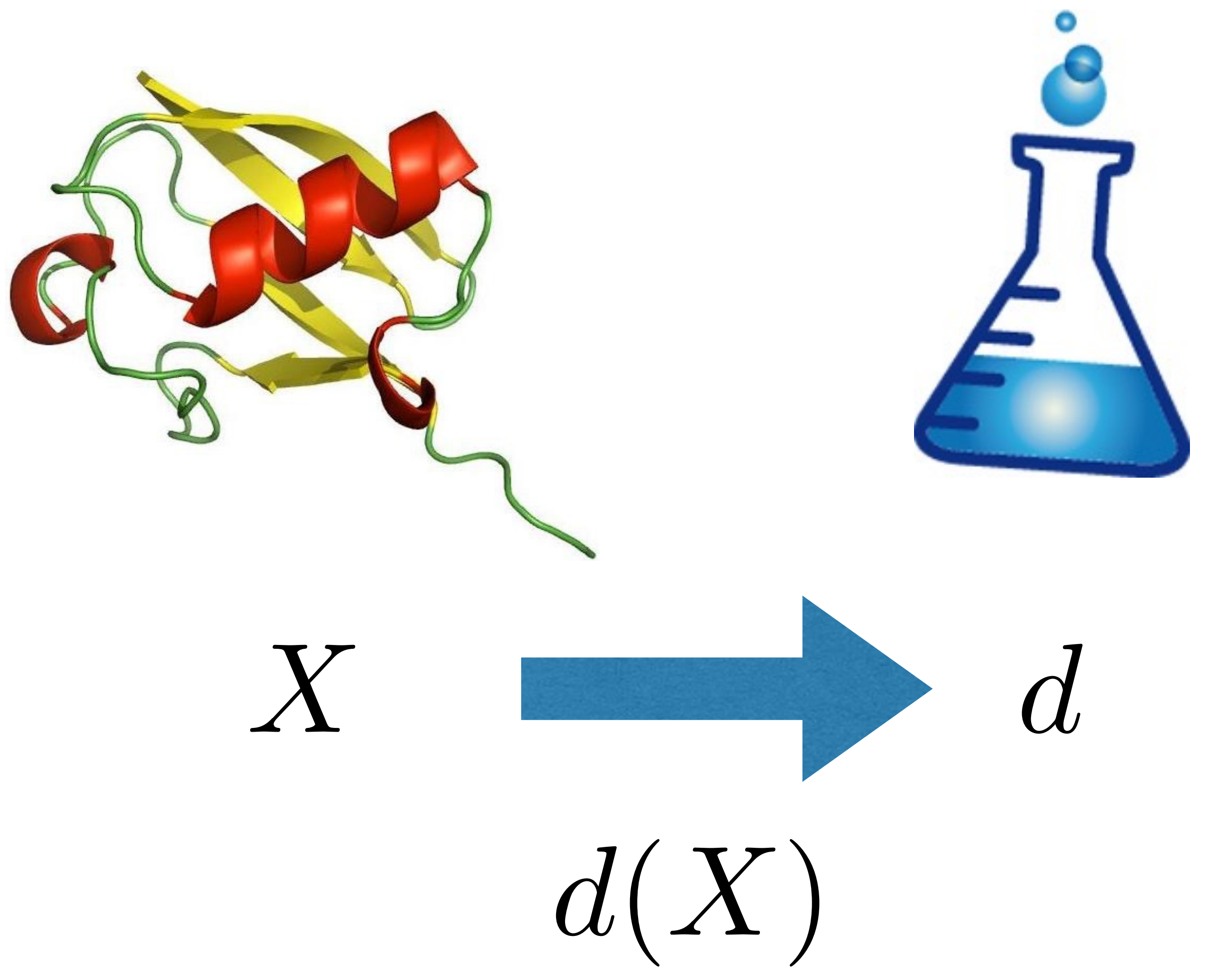
Thomas Bayes (maybe)
1702-1761

$$p(X, \sigma | d) \propto p(d | X, \sigma) \cdot p(X) \cdot p(\sigma)$$

posterior likelihood priors

The ingredients I

1) Define a *forward model*, i.e. a predictor of the experimental data from a structural model:



<u>Examples</u>

NOE: 1/distance^6
Chemical Shifts: SPARTA, SHIFTX, …
SAXS: CRYSOL, FoXS, …
Cross-links: distances, …
Cryo-EM: Gaussians centered on atoms positions

2) Define a *noise model*, i.e. a model of the deviation between predicted and measured data. Typically Gaussian or LogNormal distribution:

$$p(d|X,\sigma,I) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{[d-d(X)]^2}{2\sigma^2}}$$

large error = high tolerance for deviations
small error = small tolerance for deviations

errors in experimental data **and** predictors

The ingredients II

3) Choose the *priors*, i.e. your favorite MD force field

$$p(X|I) = e^{-\frac{E_{FF}(X)}{k_B T}}$$

and a prior for the uncertainty parameter(s):

$$p(\sigma|I) = \frac{1}{\sigma}$$

Jeffrey's prior

$$p(\sigma_i|\sigma_0, I) = \frac{2\sigma_0}{\sqrt{\pi\sigma_i^2}} exp\left(-\frac{\sigma_0^2}{\sigma_i^2}\right)$$

Outliers prior

4) Sample (MD, MC, or combined methods) using the following potential hybrid energy function:

large error = weak restraint

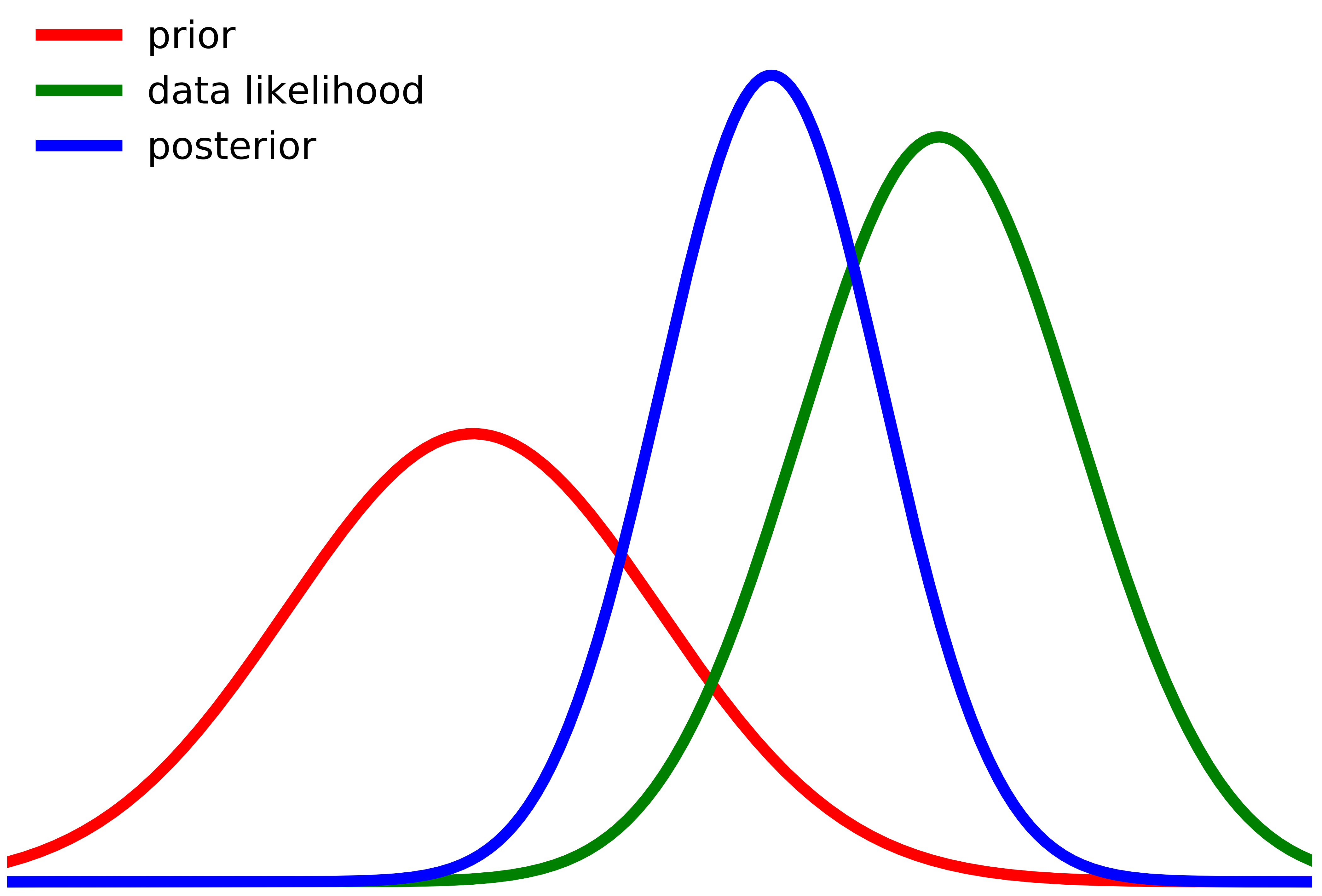
small error = strong restraint

$$E = -k_B T \log p(X, \sigma|d) = E_{FF}(X) + \frac{k_B T}{2\sigma^2}[d - d(X)]^2 + E(\sigma)$$

hybrid energy | **MD force field** | **data restraint** | **error restraint**

Inferential structure determination

Rieping *et al.* Science 2005

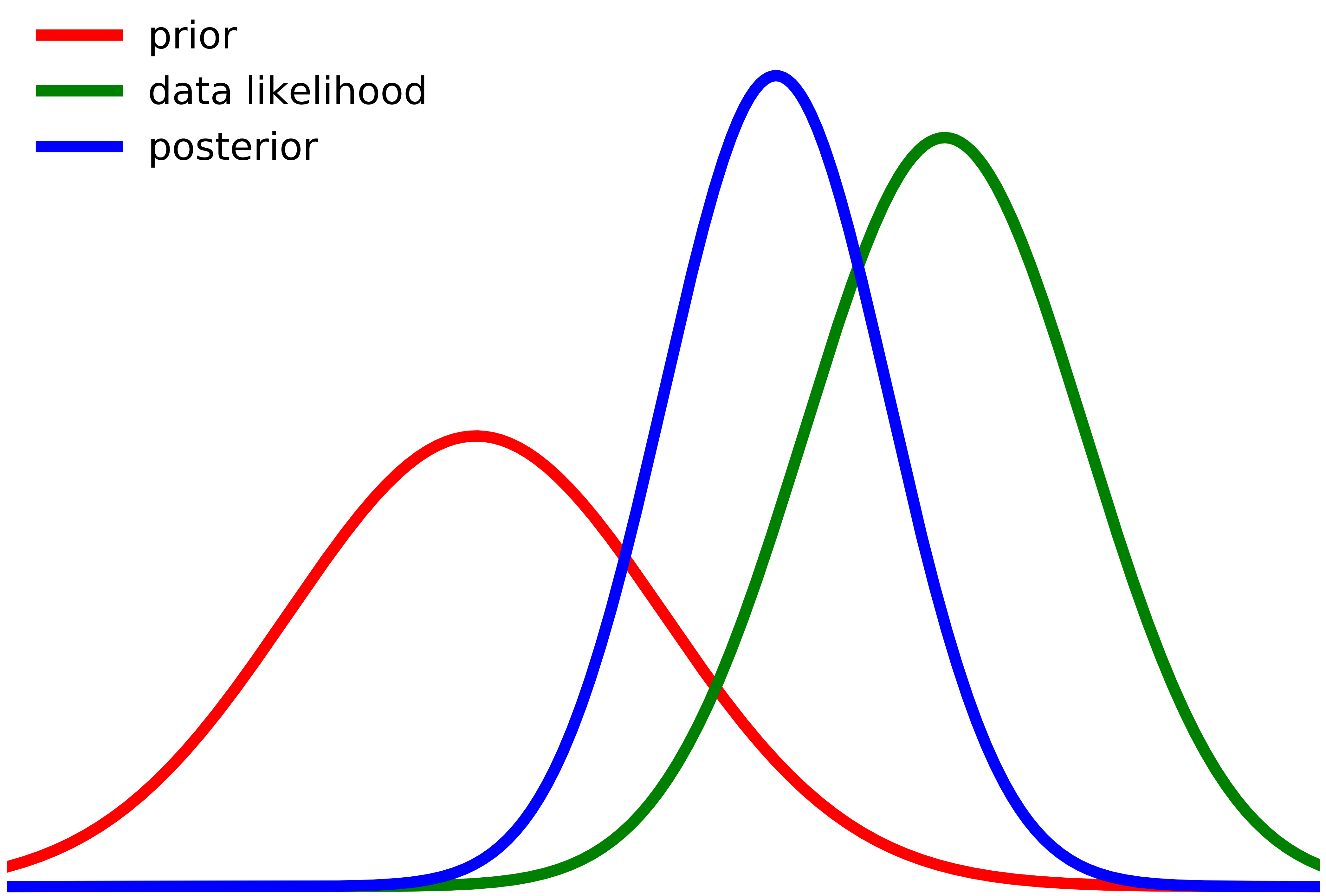


$$p(X,\sigma|d) \propto \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[d-d(X)]^2}{2\sigma^2}} \cdot e^{-\frac{E_{FF}(X)}{k_B T}} \cdot p(\sigma)$$

posterior **likelihood** **priors**

Inferential structure determination

Rieping *et al.* Science 2005



$$E = -k_B T \log p(X, \sigma | d) = E_{FF}(X) + \frac{k_B T}{2\sigma^2}[d - d(X)]^2 + E(\sigma)$$

hybrid energy

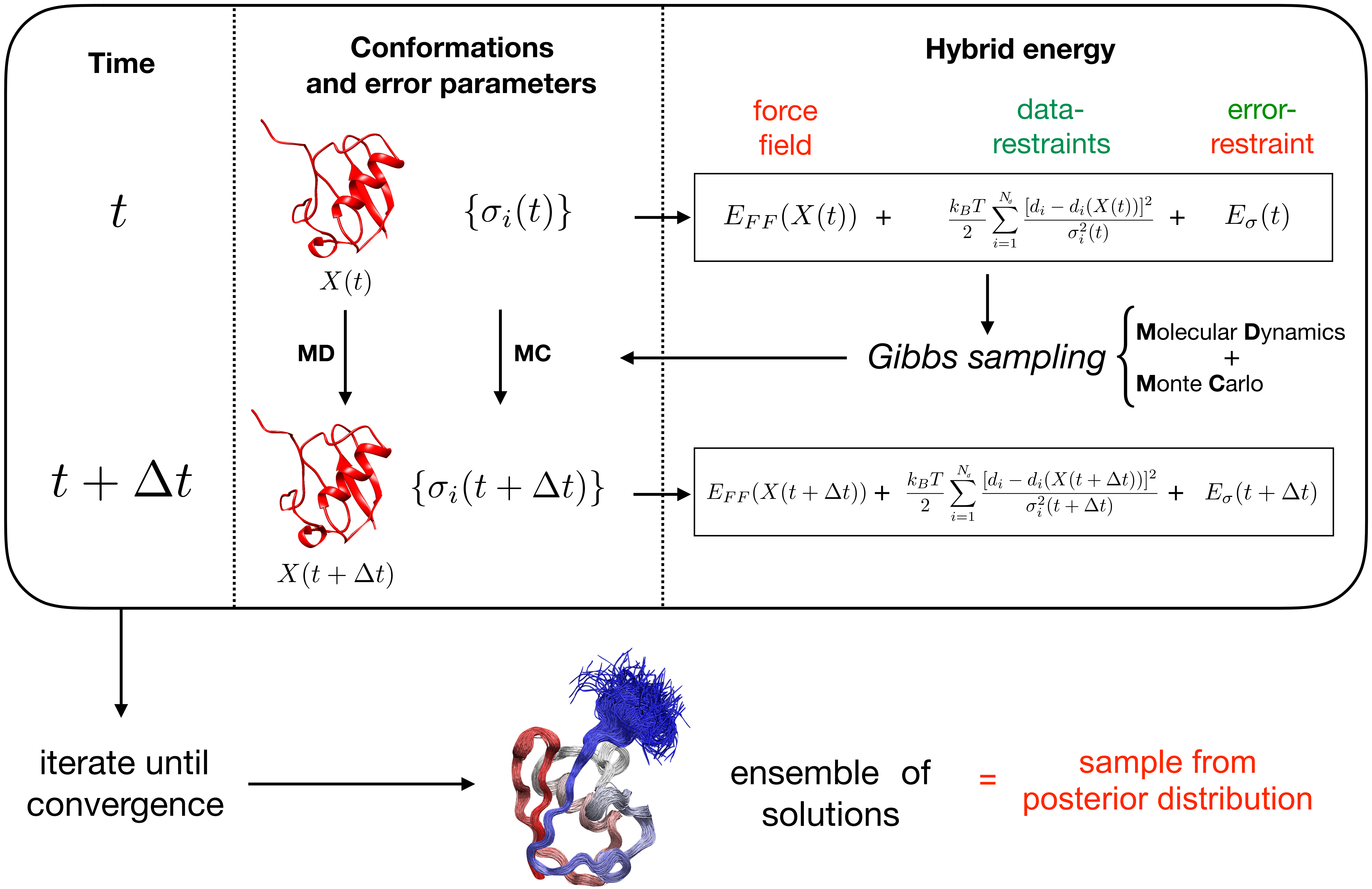
MD force field

data restraint

error restraint

Inferential structure determination

Rieping *et al.* Science 2005



The advantages of the Bayesian approach

- Information other than experimental data can be rigorously incorporated into the modelling

- <u>Errors in the data</u> are accounted for

An *uncertainty parameter* σ is associated to the data and inferred along with the structure X

$$p(X, \sigma | D, I) \propto p(D | X, \sigma, I) \cdot p(\sigma | I) \cdot p(X | I)$$

- Data from different experiments is automatically <u>balanced</u>

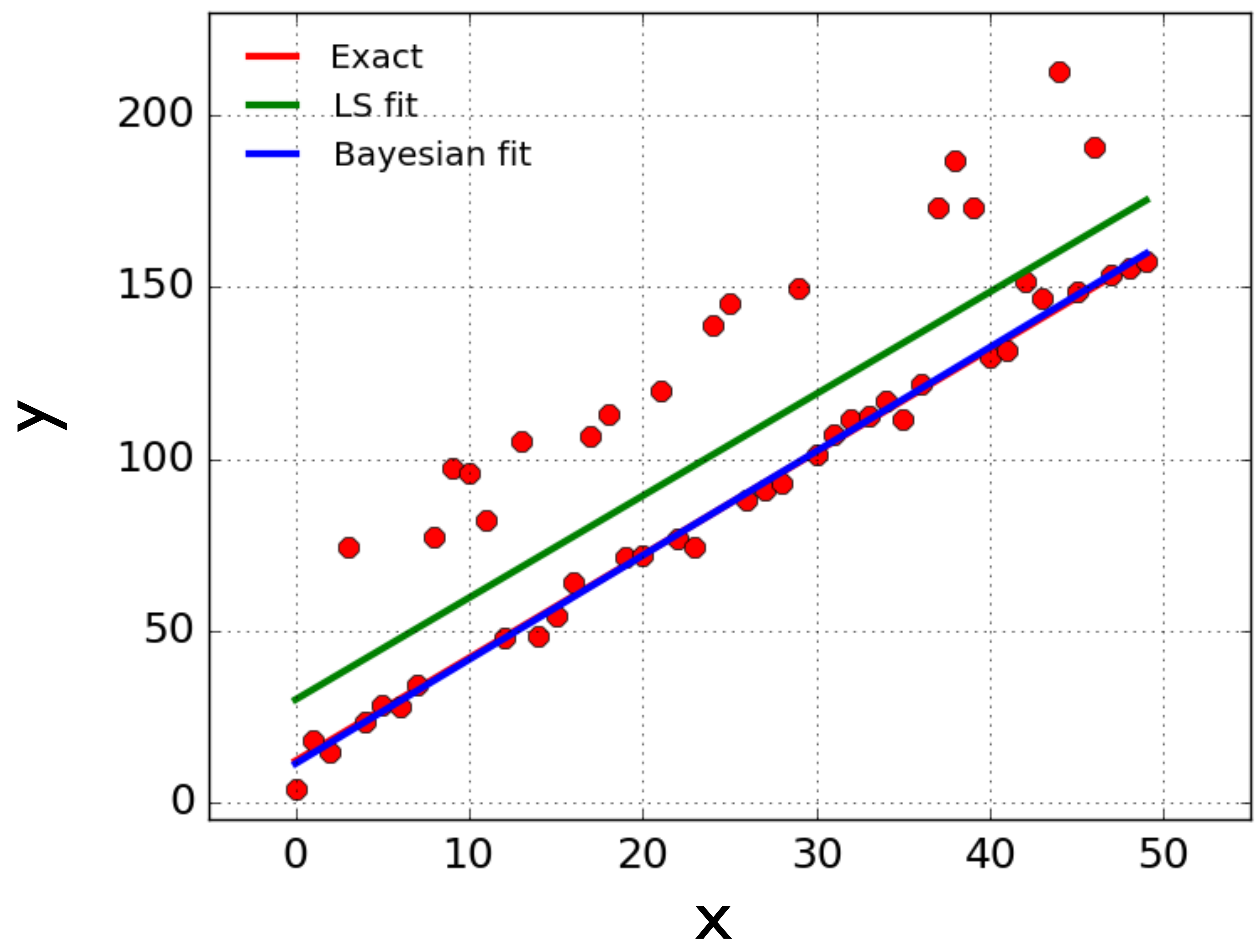
$$p(X, \{\sigma_i\} | \{d_i\}, I) \propto \left[\prod_i p(d_i | X, \sigma_i, I) \cdot p(\sigma_i | I) \right] \cdot p(X | I)$$

<u>The *weight* of each experimental data point in the construction of the structural model(s) is automatically determined based on its accuracy</u>

Example: Bayesian linear fit

Let's suppose our data can be described by a linear relation, with some error

Gaussian noise + 30% outliers



$$y=Ax+B$$

$A_0=3.0$ $B_0=12.0$

$A_{LS}=2.9$ $B_{LS}=29.8$

$A_B=3.0$ $B_B=11.3$

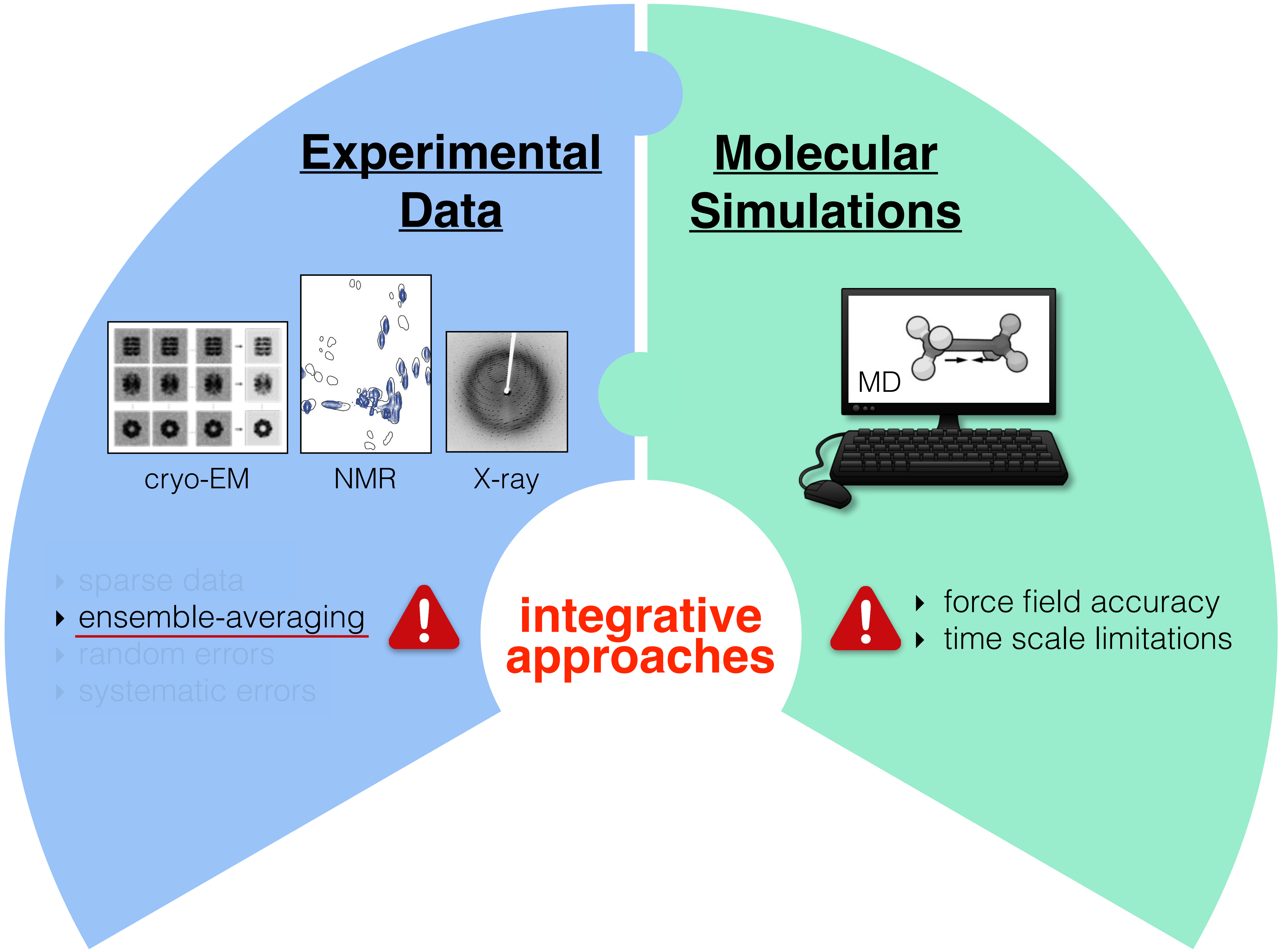
$$p(A, B, \sigma_0 | \{Y_k\}, I) \propto \prod_k \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{[f(x_k)-Y_k]^2}{2\sigma_k^2}} \cdot \frac{2\sigma_0}{\sqrt{\pi \sigma_k^2}} e^{-\frac{\sigma_0^2}{\sigma_k^2}}$$

Posterior probability

$$f(x_k) = Ax_k + B$$

Linear forward model

The best of both worlds: integrative or *hybrid* approaches



Inferential determination of structural ensembles with Metainference

Bonomi *et al.* Sci Adv 2016

META INFERENCE

$$E_{MI}(\mathbf{X},\sigma) = k_B T \cdot \sum_{r=1}^{N} \left\{ -\log p(X_r) + \sum_{i=1}^{N_d} (d_i - f_i(\mathbf{X}))^2 \frac{1}{2\sigma_{r,i}^2} + \log \sigma_{r,i} - \log p(\sigma_{r,i}) \right\}$$

hybrid energy

*multiple-replica sampling

MD force field

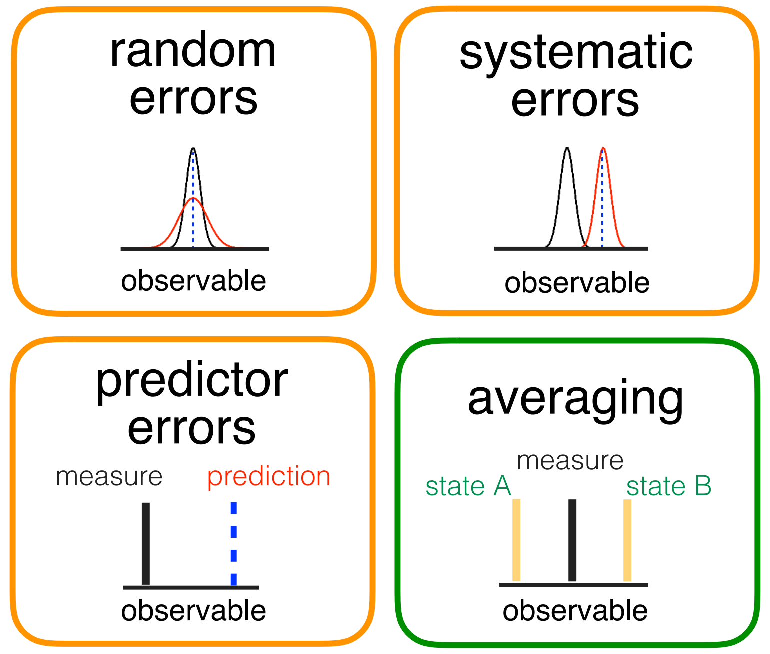
prior knowledge

data restraints

agreement with <u>experimental data</u>

error restraints

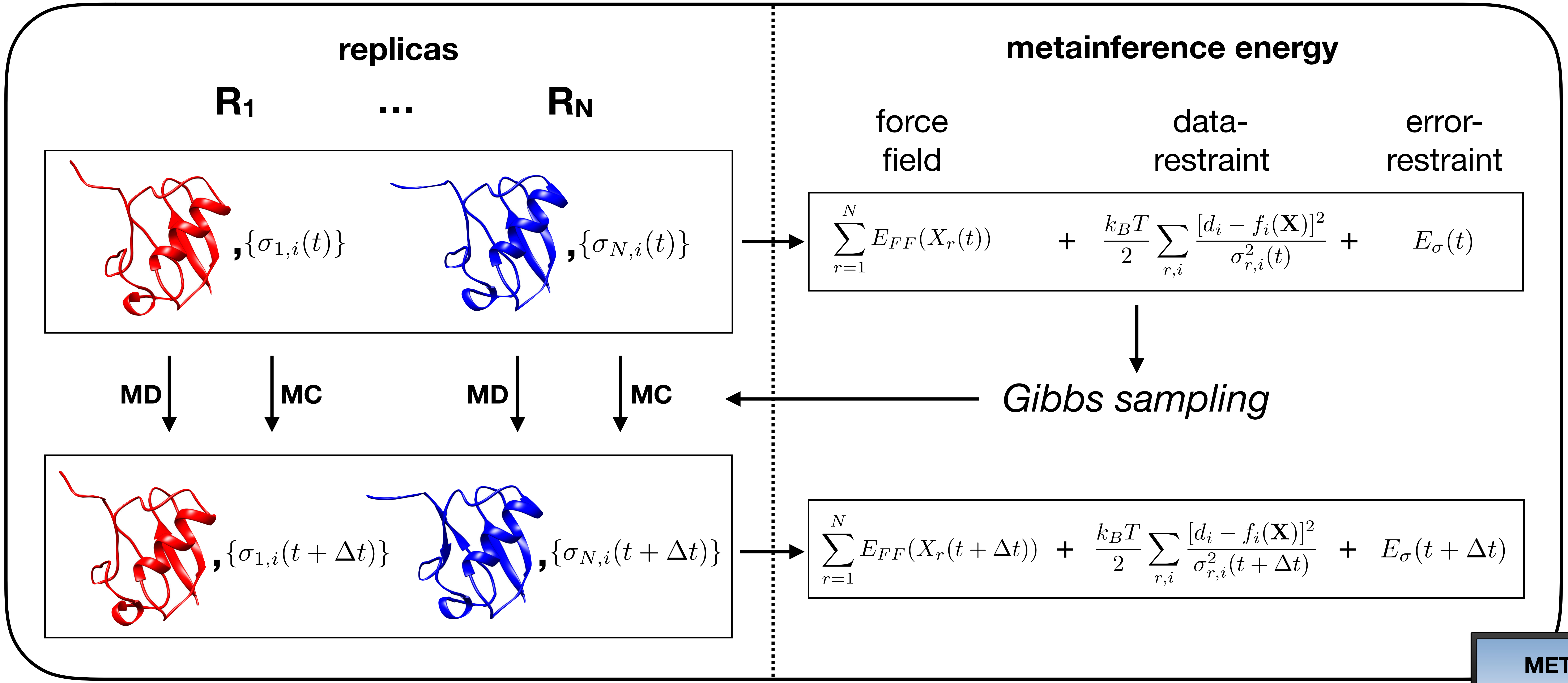
information about all the errors at play

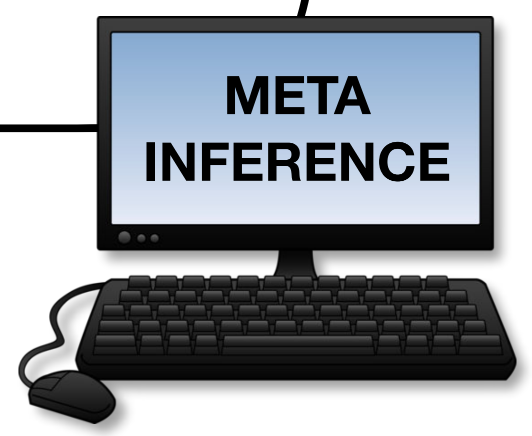


*Lindorff-Larsen *et al.* Nature 2005
Cavalli, Camilloni, Vendruscolo. J Chem Phys 2013
Hummer, Köfinger. J Chem Phys 2015

Inferential determination of structural ensembles with Metainference

Bonomi *et al.* Sci Adv 2016

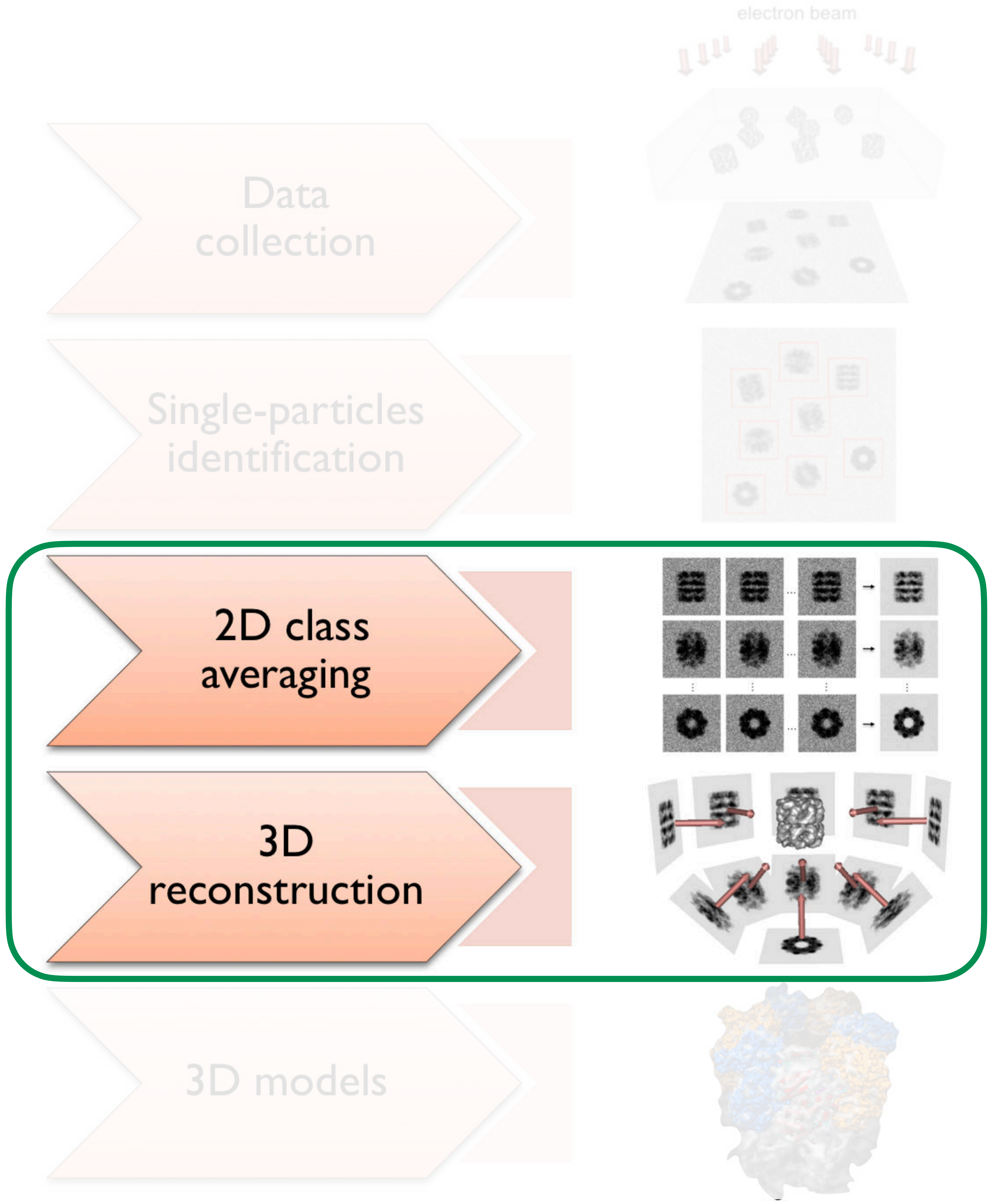


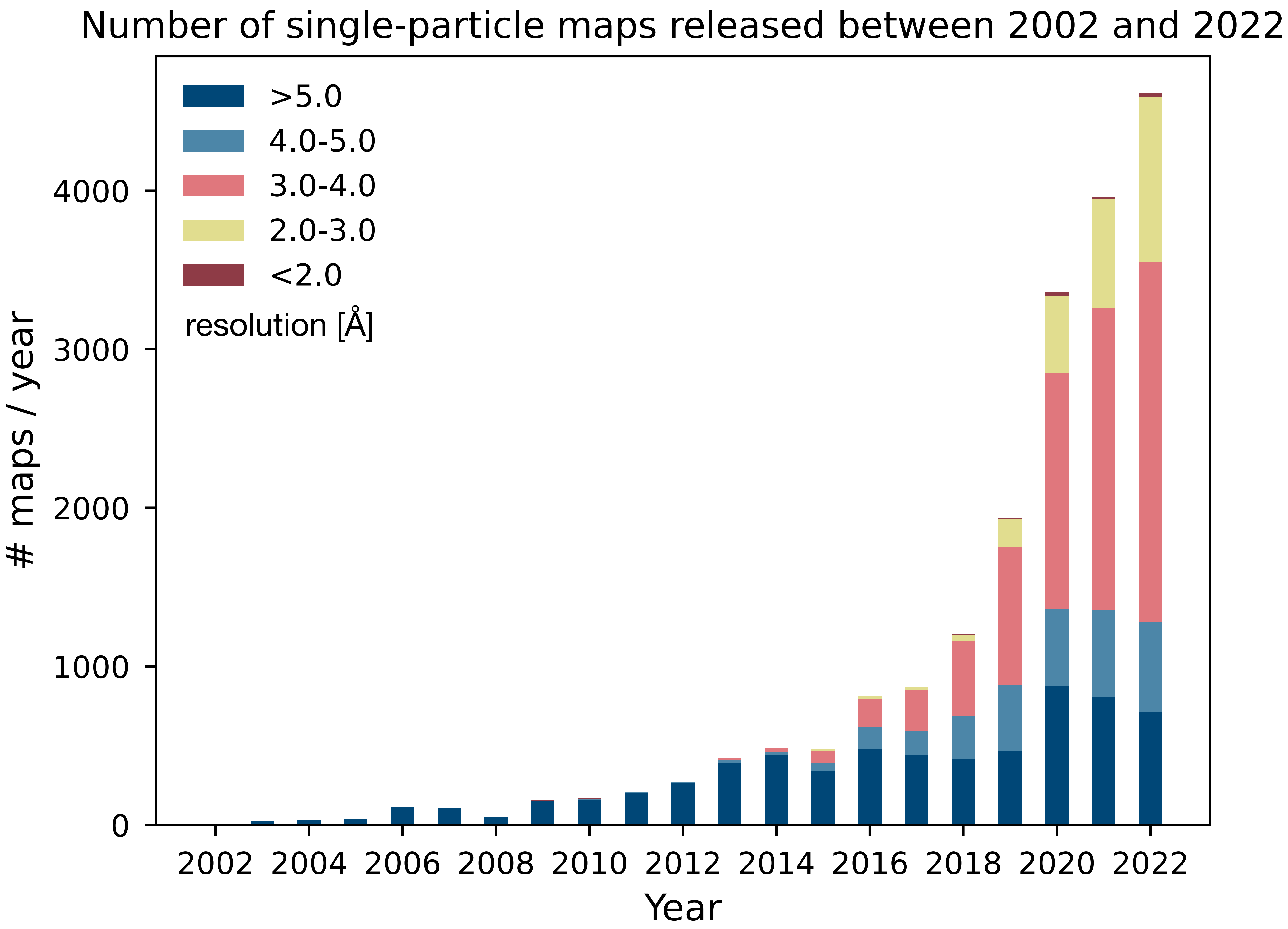


Minimal perturbation of the prior ensemble
Maximum Entropy principle*

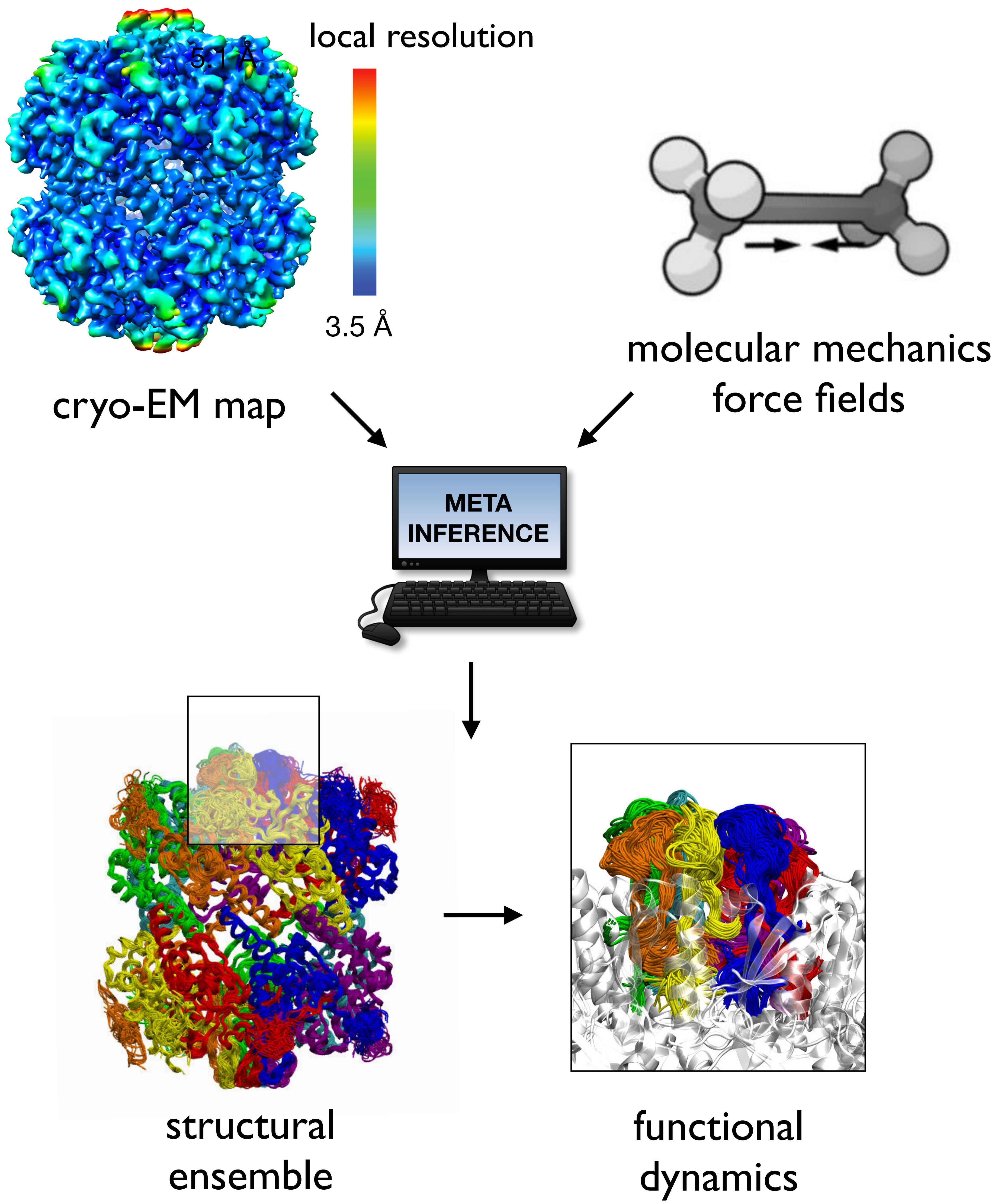
*see Hummer & Köfinger. J Chem Phys 2015

The cryo-EM resolution revolution





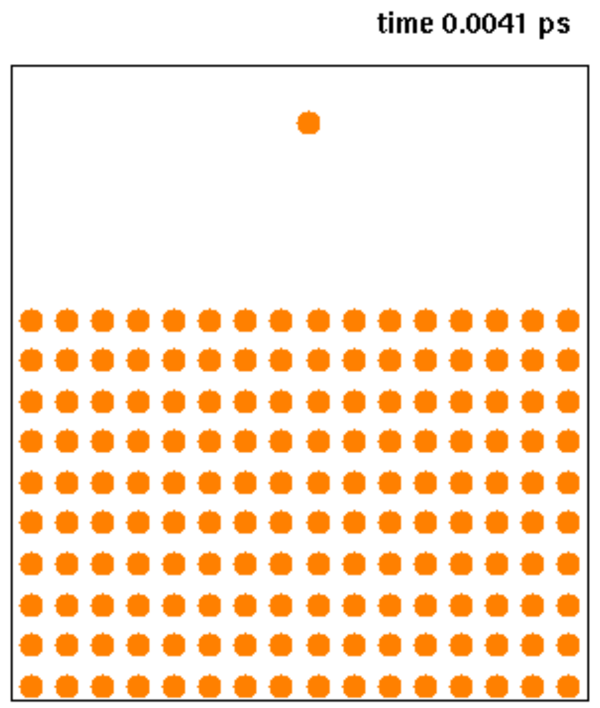
The cryo-EM resolution revolution



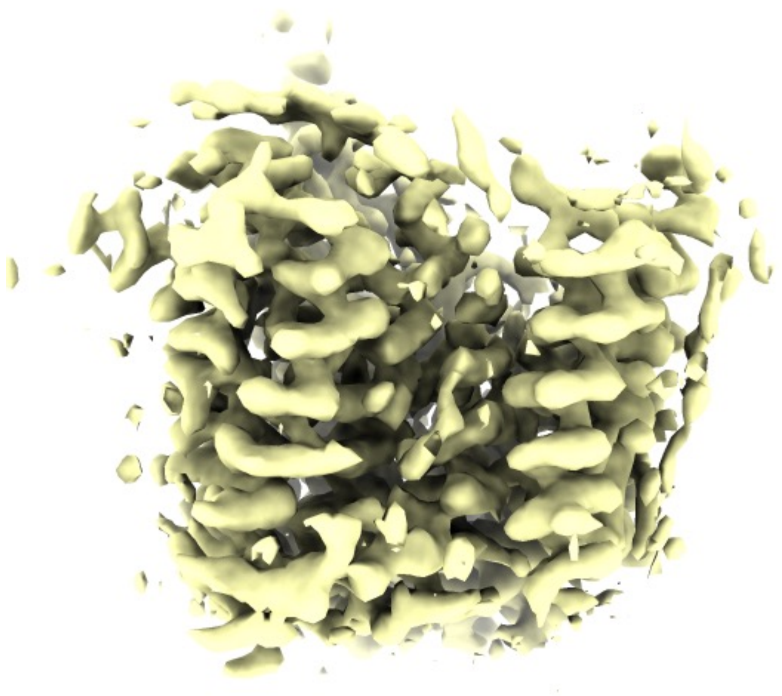
Integrate cryo-EM into MD simulations with EMMIVox

Hoff & Bonomi. In preparation; Bonomi, …,Pellarin. Structure 2019

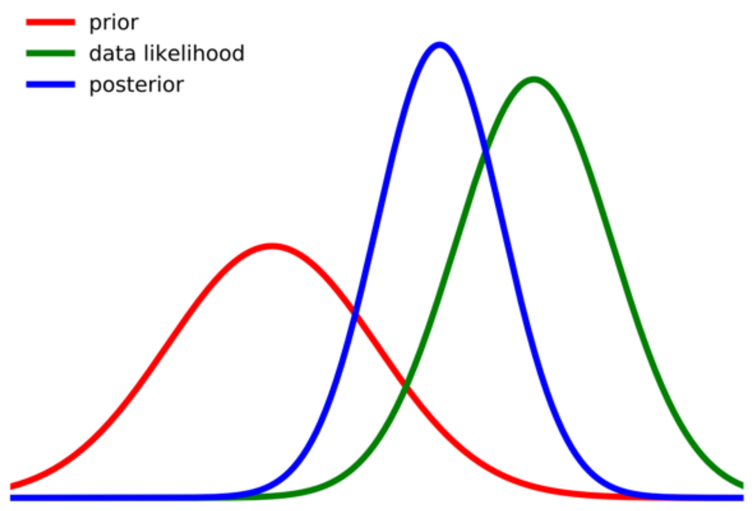
Accuracy of MD is limited!



Experimental data

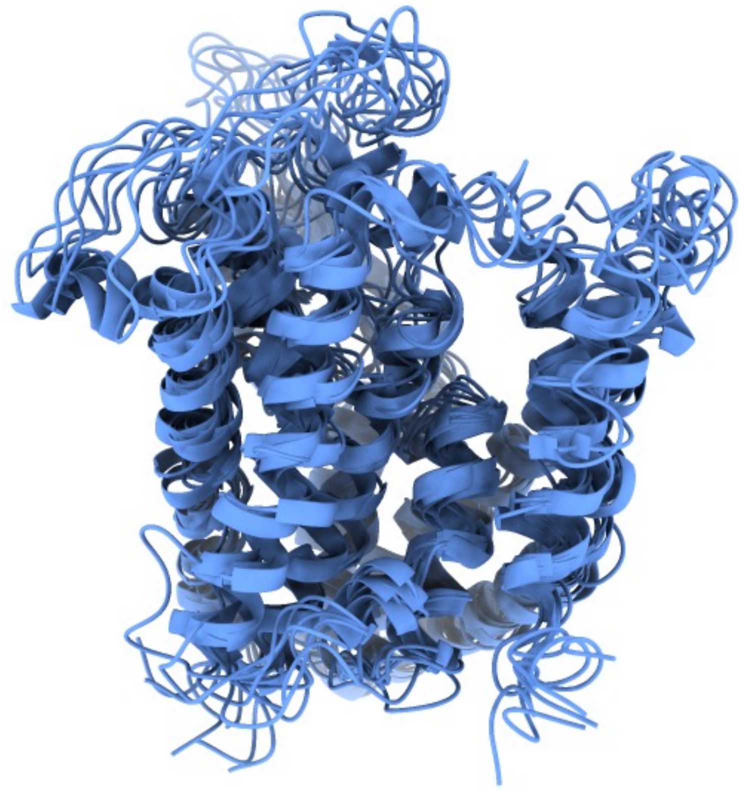


Bayesian inference



$$p(X,\sigma|d) \propto p(d|X,\sigma) \cdot p(X) \cdot p(\sigma)$$

Structural models



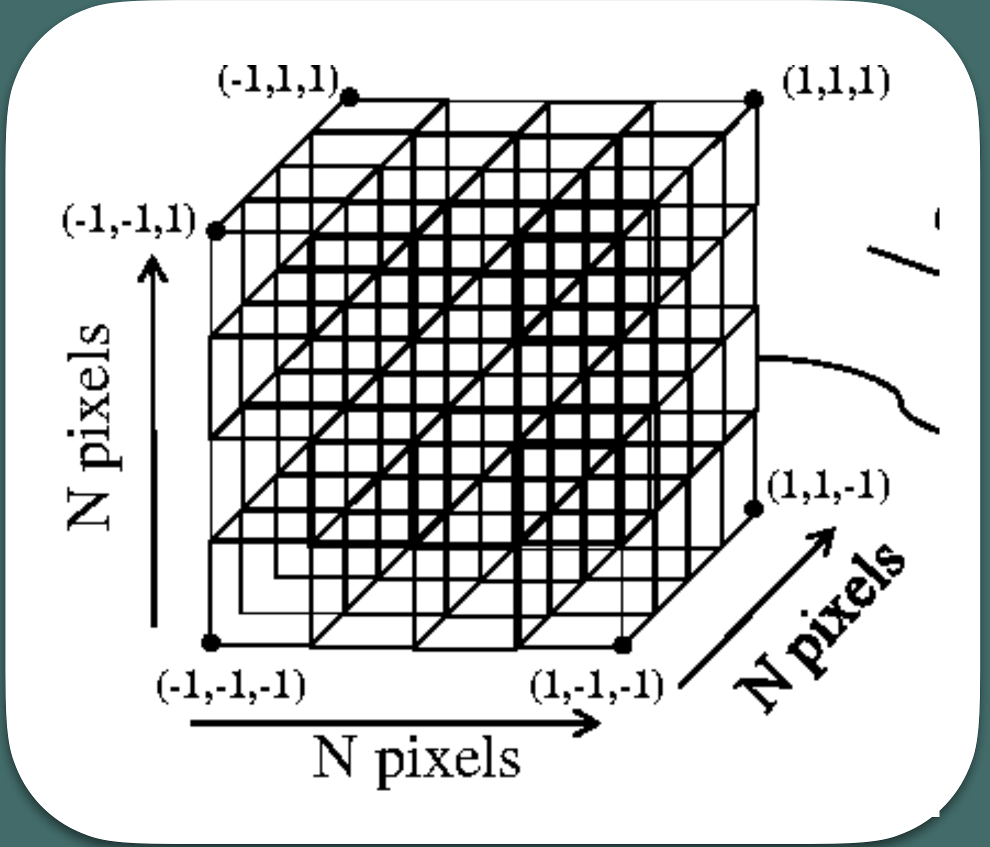
Objectives:

- Improving stereochemical quality of deposited PDBs (fewer clashes, …)
- Modelling structural ensembles with Metainference
- Integrating other experimental data (NMR, SAXS, EPR, …)

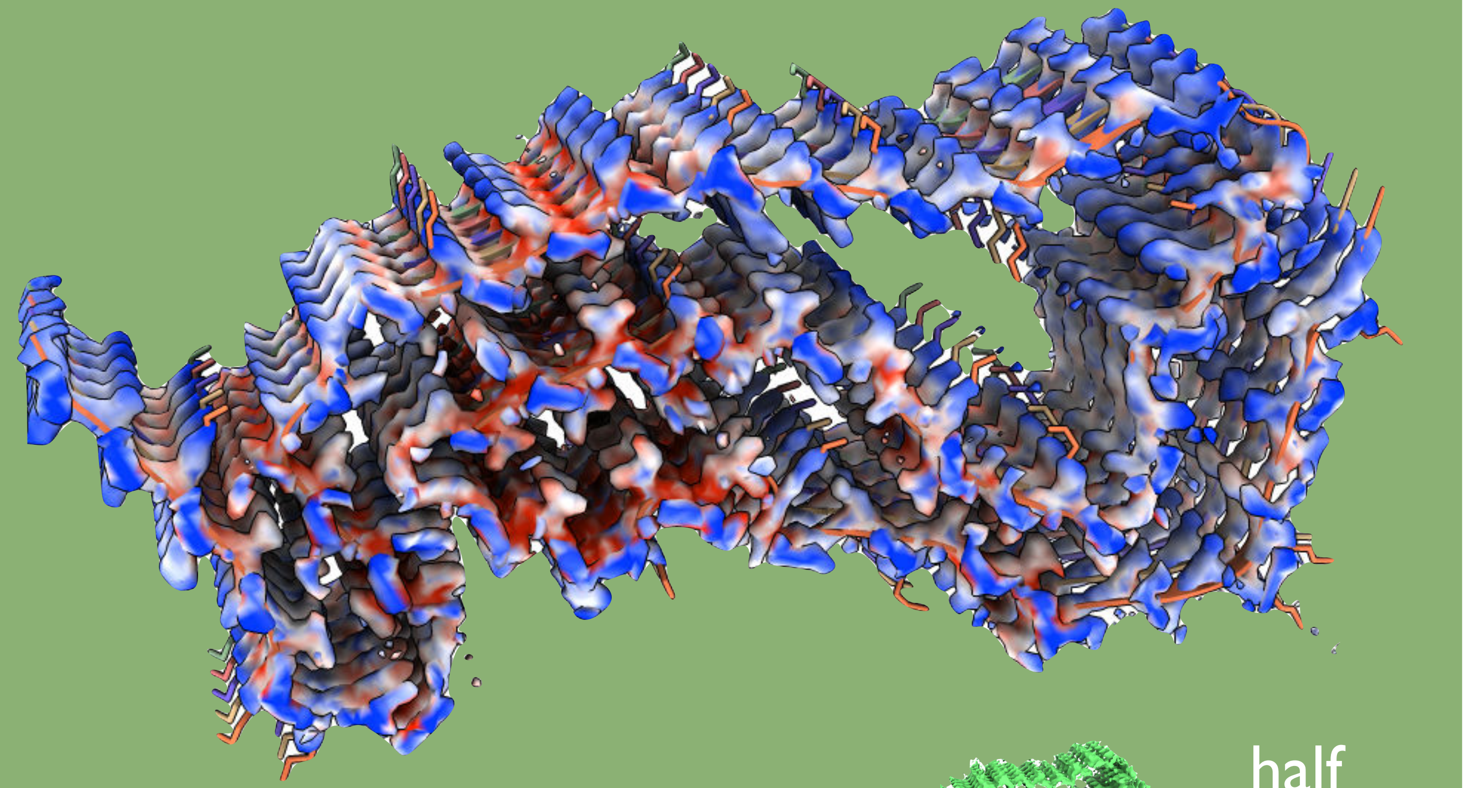
Integrate cryo-EM into MD simulations with EMMIVox

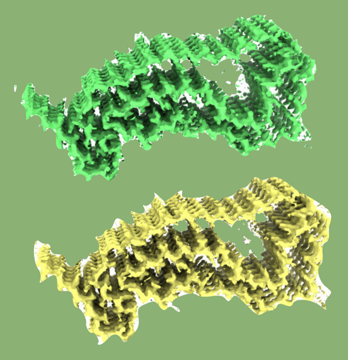
Hoff & Bonomi. In preparation; Bonomi, …,Pellarin. Structure 2019

Estimate correlation between voxels



Estimate "experimental noise" for each voxel





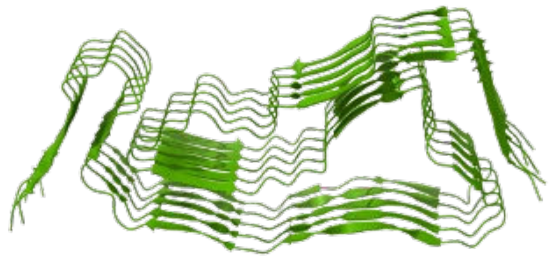
Gaussian noise model for each voxel

$$p(\phi_{D,i}|M) = \frac{1}{\sqrt{2\pi}\sigma_i} \cdot exp\left[-\frac{(\phi_{D,i} - \phi_{M,i})^2}{2\sigma_i^2}\right]$$

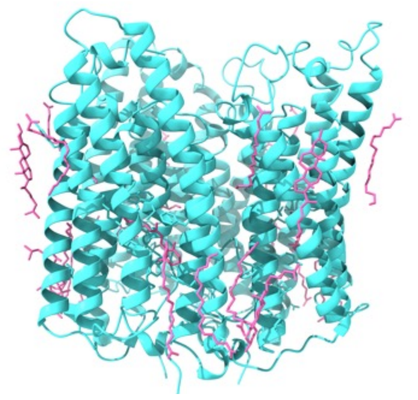
- remove correlated voxels
- one error parameter per voxel
- minimum value from error map
- error marginalization

Better balance between stereochemical quality and fit to the map

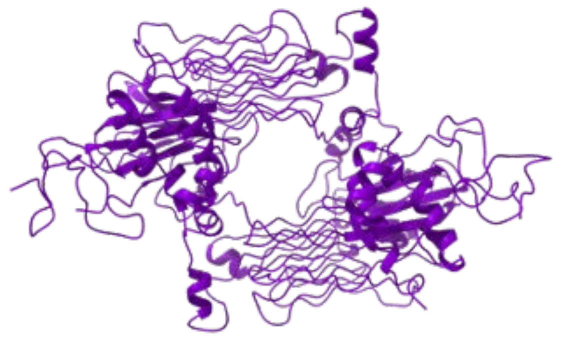
7p6a



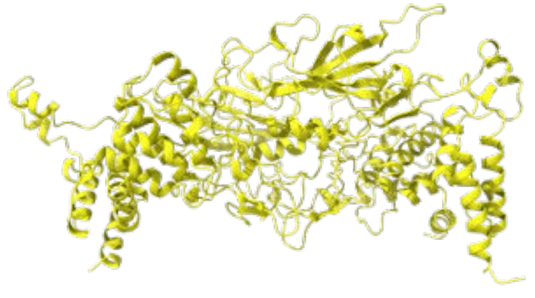
7w9w



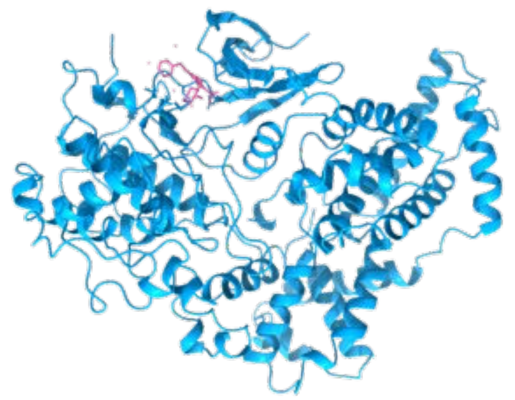
7n00



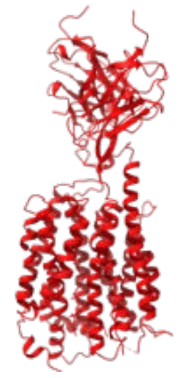
7t4n



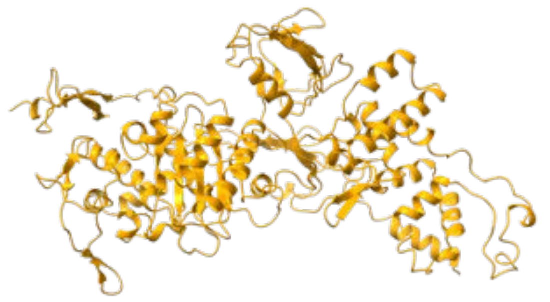
7b5o



7mjs



7lq6



7nqk

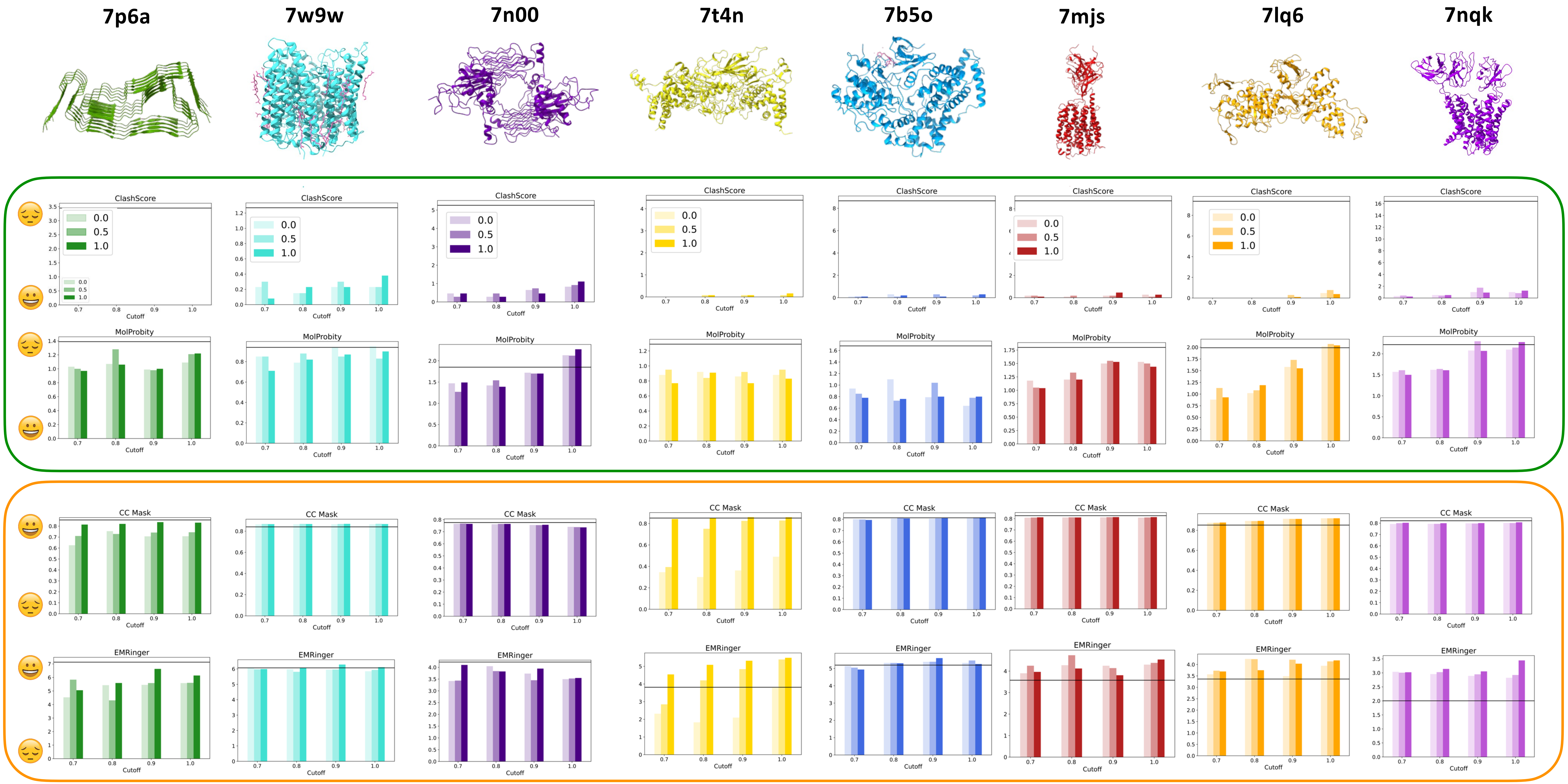
Clash between atoms: Clashscore

Overall sterochemical quality: Molprobity

Fit model/map: CC_mask

Correct positioning of sidechains: EMRinger

Better balance between stereochemical quality and fit to the map



EMMIVox is implemented in PLUMED

Bonomi *et al.* CPC 2009; Tribello *et al.* CPC 2014

PLUMED

www.plumed.org

- open-source, freely-available
- community-developed
- enhanced sampling methods for MD
- analysis tools
- algorithms for integrative biology

NAMD Scalable Molecular Dynamics

DL_POLY

OpenMM

QUANTUMESPRESSO

GROMACS FAST. FLEXIBLE. FREE.

ESPResSo

CP2K

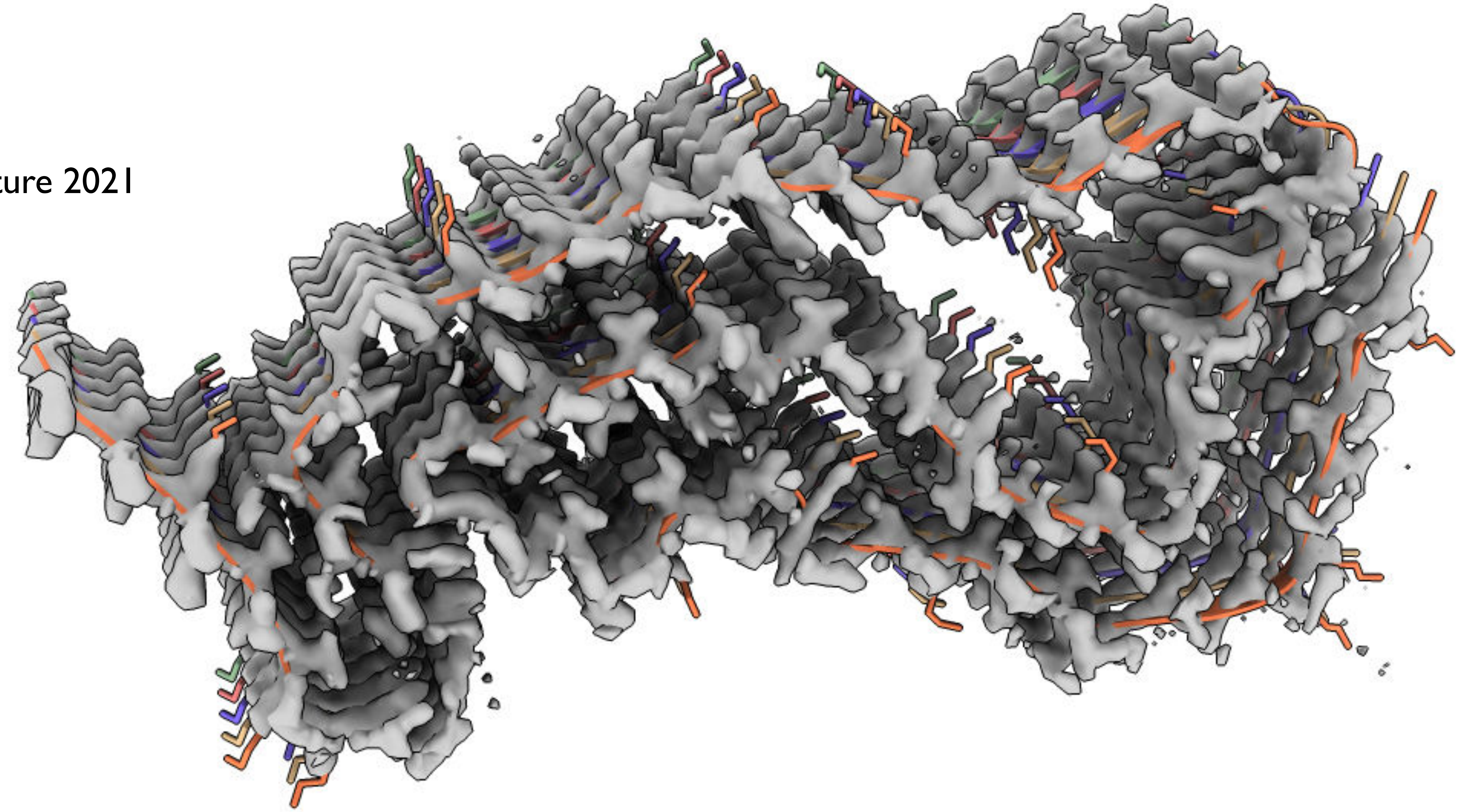
LAMMPS

OpenPathSampling

Today's system
Tau filament from the brains of individuals with PSP

*PDB 7p6a
EMDB 13223
resolution 1.9 Å

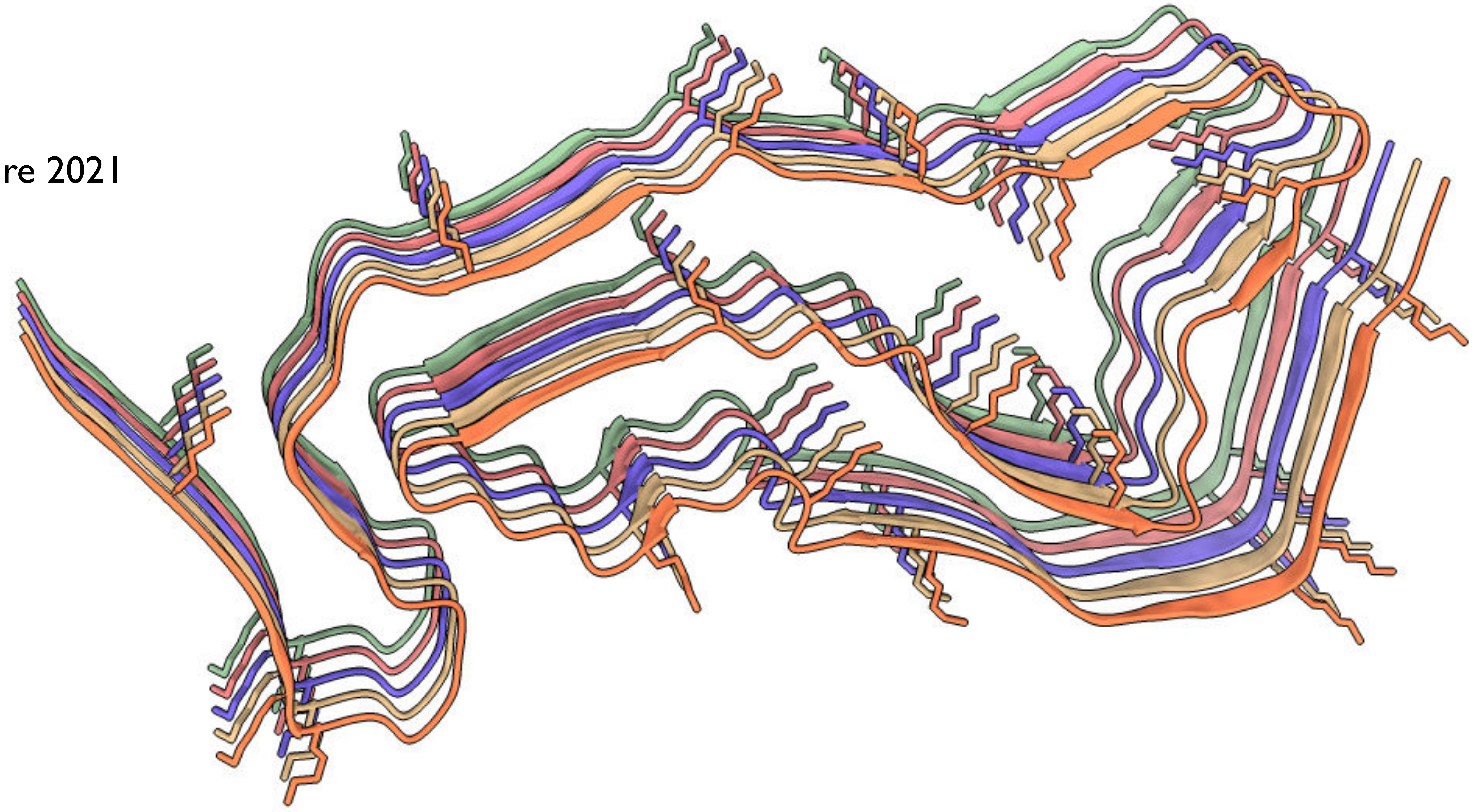
*Shi, …, Scheres. Nature 2021



Today's system
Tau filament from the brains of individuals with PSP

*PDB 7p6a
EMDB 13223
<u>resolution 1.9 Å</u>

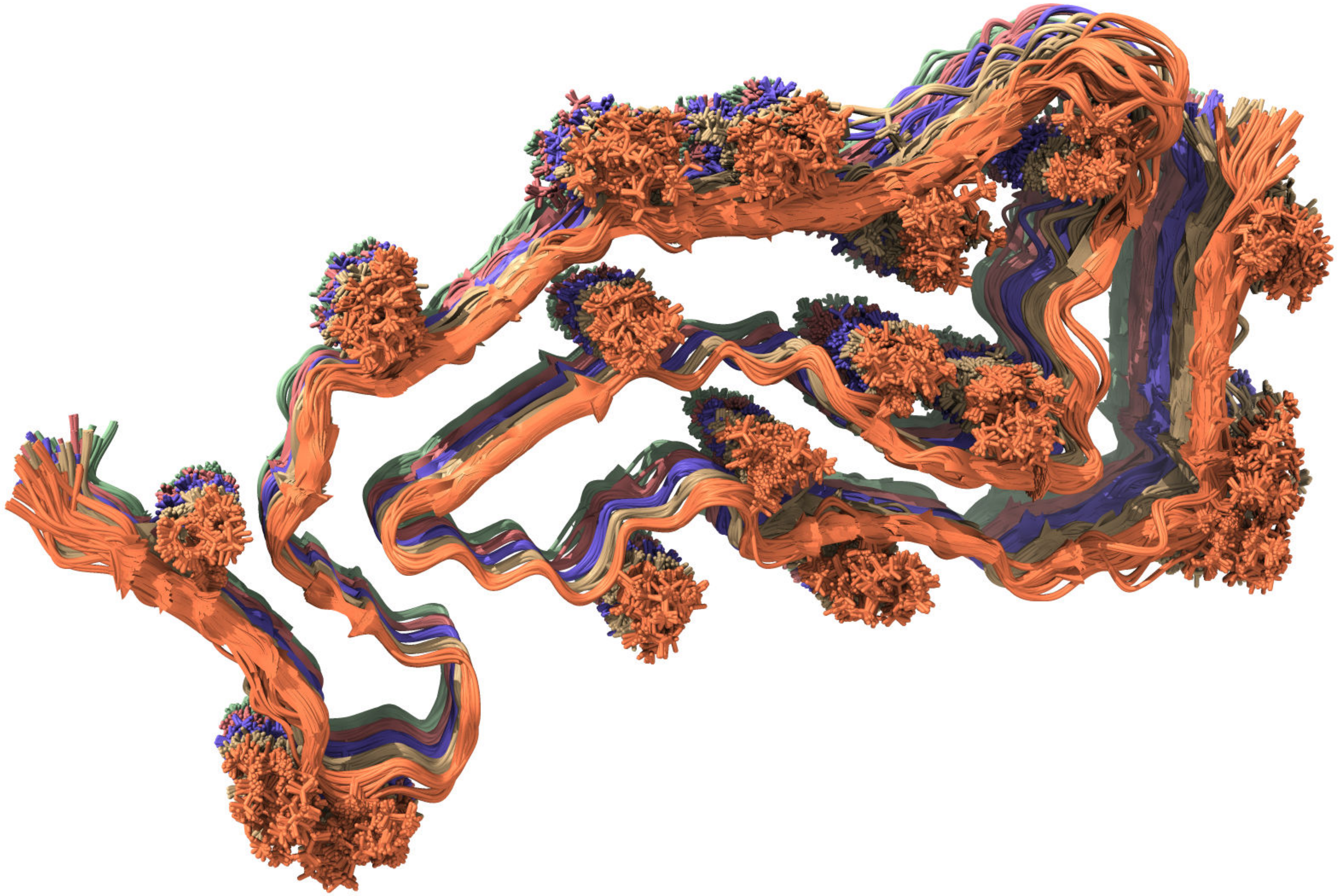
*Shi, …, Scheres. Nature 2021



Exercise 1
Compare EMMIVox refined model and deposited PDB

	Validation metric	deposited PDB	EMMIVox model
Fit model-map	CC_Mask		
	EMRinger		
Stereochemical quality	Clashscore		
	Molprobity		

Exercise 2
Determining structural ensembles with EMMIVox and Metainference



Fit
model-map

Validation metric	deposited PDB	EMMIVox model	EMMIVox ensemble
CC_Mask			

Acknowledgments

INSTITUT PASTEUR

Samuel Hoff
Riccardo Pellarin
Michael Nilges

PLUMED consortium

SISSA
Scuola Internazionale Superiore di Studi Avanzati

Giovanni Bussi

QUEEN'S UNIVERSITY BELFAST

Gareth Tribello

UNIVERSITAS STUDIORUM MEDIOLANENSIS

Carlo Camilloni

All of you for your attention

Funding

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

sanofi

GENCI
Le calcul intensif au service de la connaissance

cnrs

ANR AGENCE NATIONALE DE LA RECHERCHE

PRACE

anrt ASSOCIATION NATIONALE RECHERCHE TECHNOLOGIE