Metainference Metadynamics with PLUMED

Max Bonomi
University of Cambridge
mb2006@cam.ac.uk

Outline

- Molecular Dynamics as a computational microscope
  - sampling problems
  - accuracy of force fields

- Tackling sampling issues by biasing MD simulations
  - umbrella sampling
  - metadynamics
  - recent developments

- The open source library PLUMED

- Combining simulations with experiments: Metainference Metadynamics

A computational microscope

Molecular Dynamics evolves a system in time under the effect of a force field

How? By integration of the Newton's equations of motion

$$M_i \frac{d^2 \mathbf{X}_i}{dt^2} = -\Delta_i V(\mathbf{X})$$

The potential is derived from

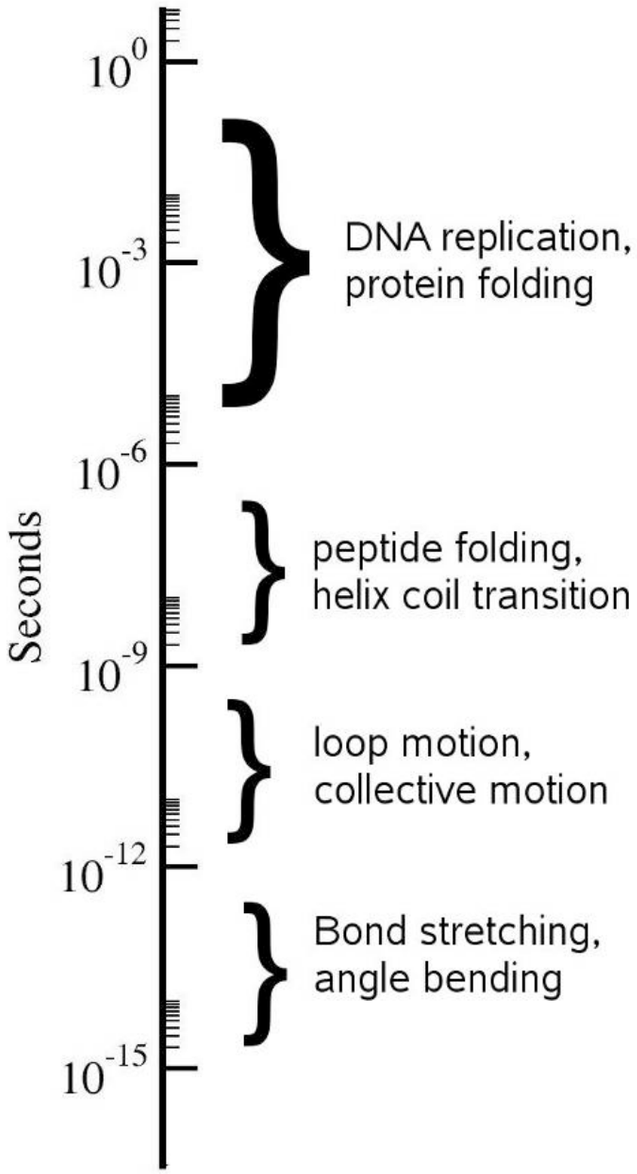
- fitting experimental observables
- first principles calculations

Limitations:

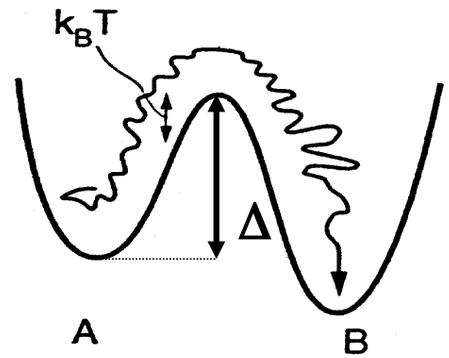
- time scale accessible
- accuracy of the force field

The time scale problem

MD/MC sampling efficiency is limited by the time scale accessible in typical simulations:



★ Activated events



★ Slow diffusion

Dimensional reduction

It is often possible to describe a physical/chemical process in terms of a small number of coarse descriptors of the system:

$$\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{R}) = (S_1(\boldsymbol{R}), \ldots, S_d(\boldsymbol{R}))$$

Key quantity of thermodynamics is the free energy as a function of these variables:

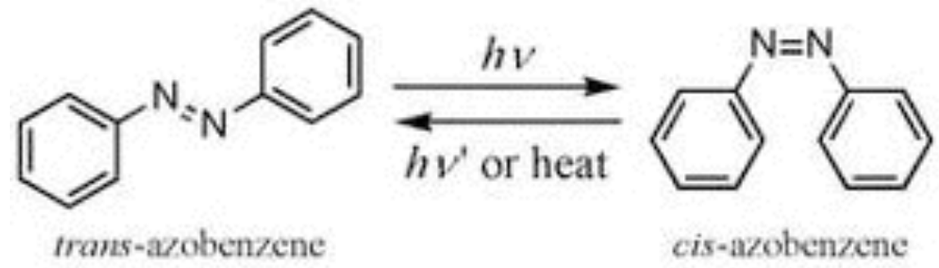
$$F(\boldsymbol{S}) = -\frac{1}{\beta} \ln P(\boldsymbol{S}) \quad \text{where} \quad \beta = \frac{1}{k_B T}$$

canonical ensemble

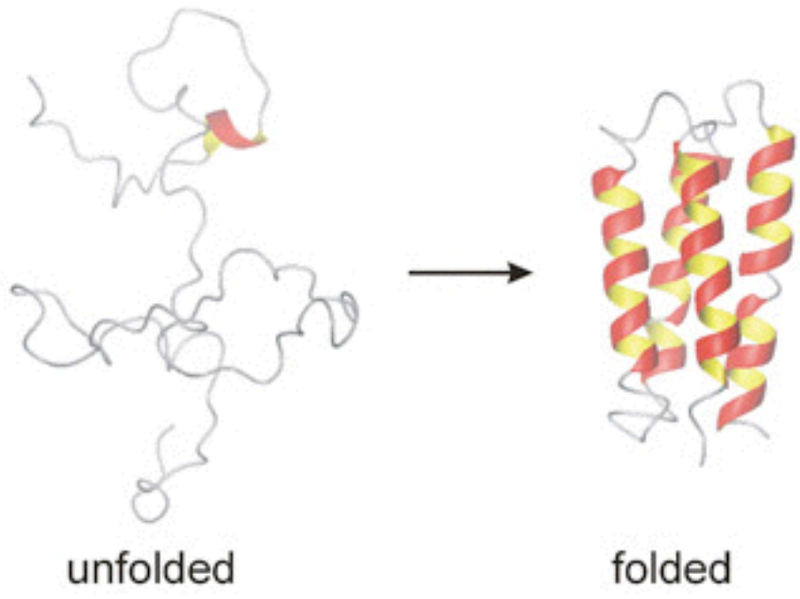
$$P(\boldsymbol{S}) = \frac{\int d\boldsymbol{R}\, \delta(\boldsymbol{S} - \boldsymbol{S}(\boldsymbol{R}))\, e^{-\beta U(\boldsymbol{R})}}{\int d\boldsymbol{R}\, e^{-\beta U(\boldsymbol{R})}}$$

Examples

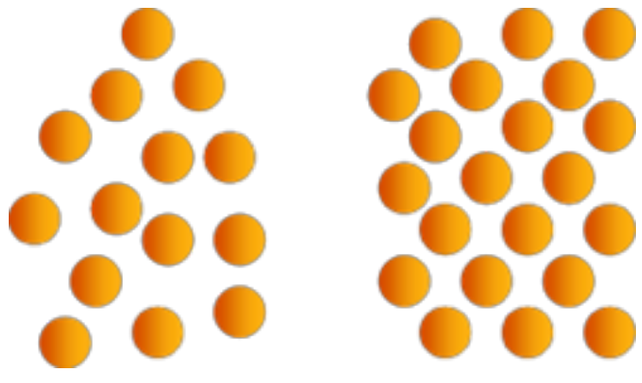
Isomerization:
dihedral angle



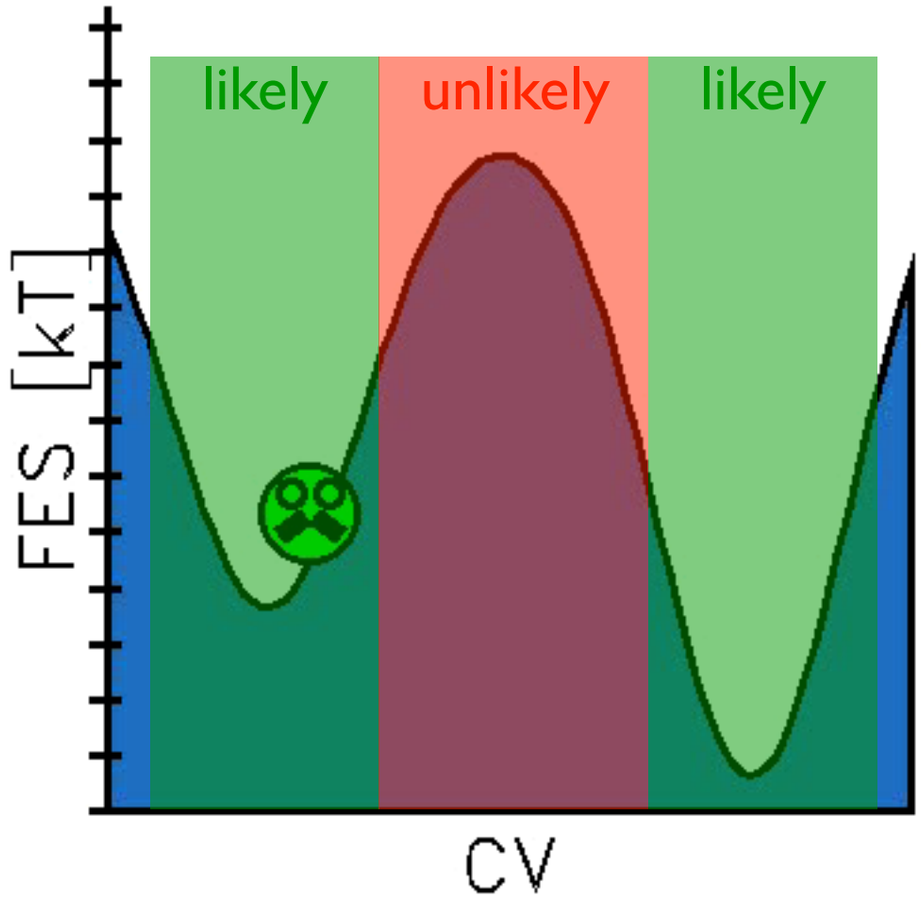
Protein folding:
gyration radius,
number of contacts,
...



Phase transitions:
lattice vectors,
bond order parameters,
...



Rare events simplified



How can we estimate a free energy difference if we never see a transition?

$$F(A) - F(B) = -k_B T \ln \frac{N_A}{N_B}$$

Biased sampling

The idea is to add a bias potential that acts on the collective variables:

$$U(\mathbf{R}) \rightarrow U(\mathbf{R}) + V(\mathbf{S}(\mathbf{R}))$$

In this biased ensemble the free energy becomes:

$$F'(\mathbf{S}) = -\frac{1}{\beta} \ln P'(\mathbf{S}) + C$$

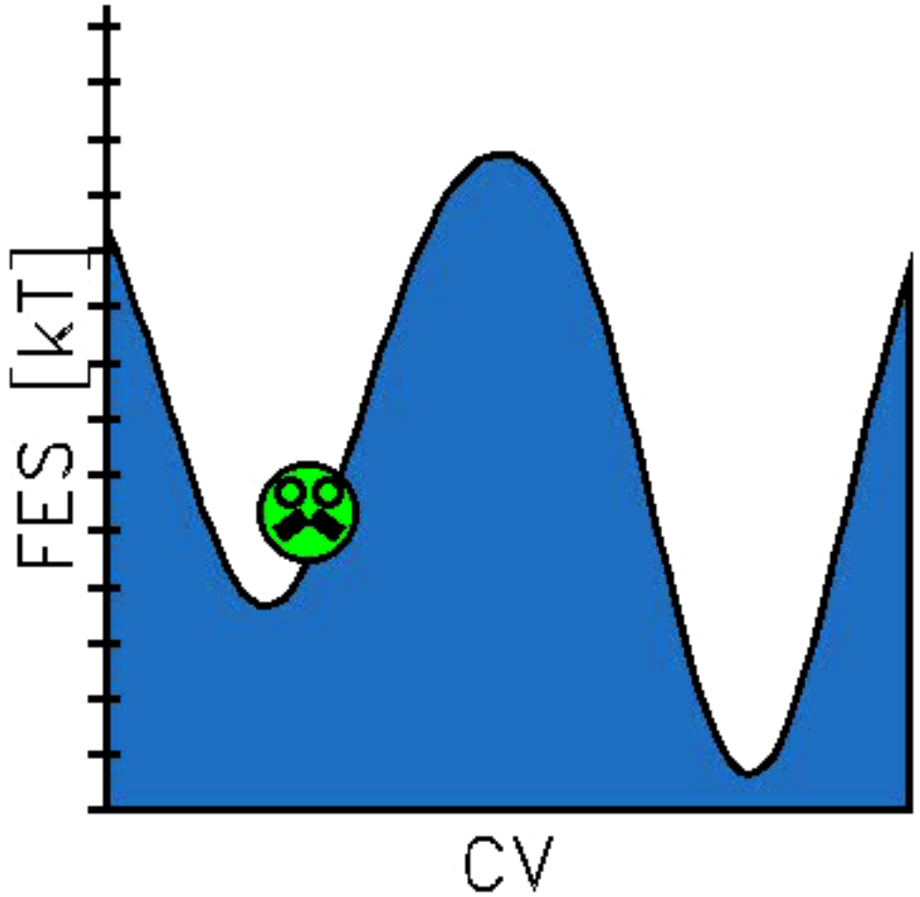
where $$P'(\mathbf{S}) = \frac{\int d\mathbf{R}\, \delta(\mathbf{S} - \mathbf{S}(\mathbf{R}))\, e^{-\beta[U(\mathbf{R}) + V(\mathbf{S}(\mathbf{R}))]}}{\int d\mathbf{R}\, e^{-\beta[U(\mathbf{R}) + V(\mathbf{S}(\mathbf{R}))]}}$$

which leads to:

$$F'(\mathbf{S}) = F(\mathbf{S}) + V(\mathbf{S})$$

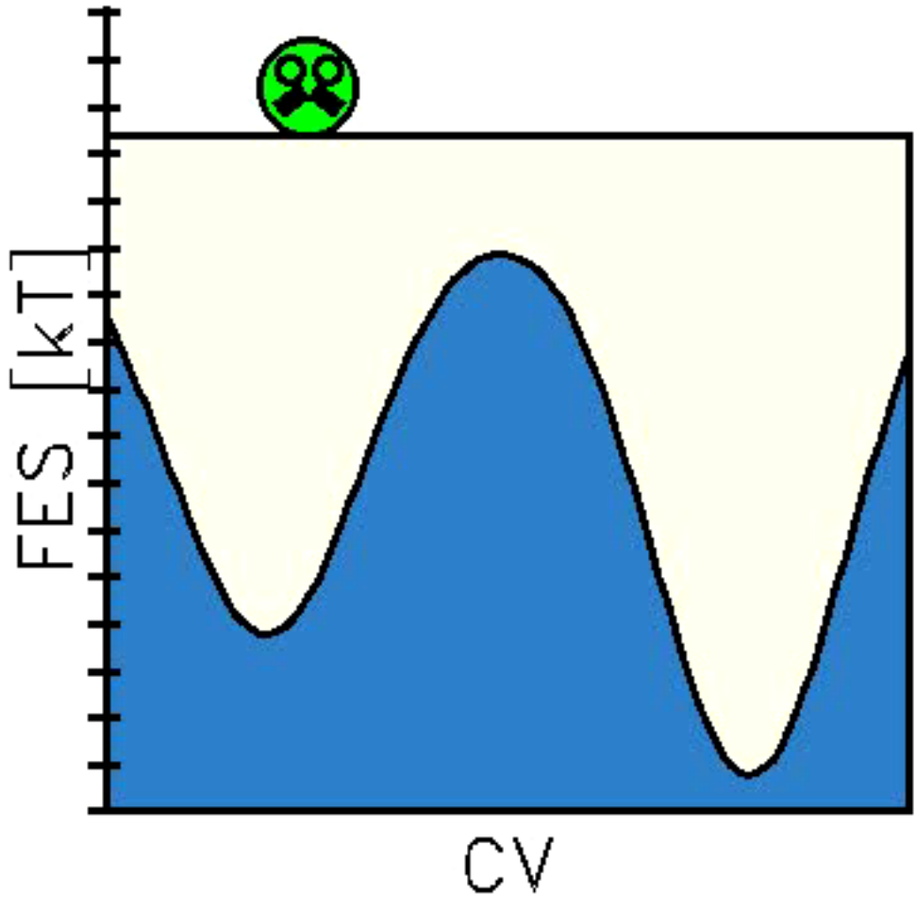
Umbrella sampling

What is a good choice of bias potential?



Umbrella sampling

What is a good choice of bias potential?



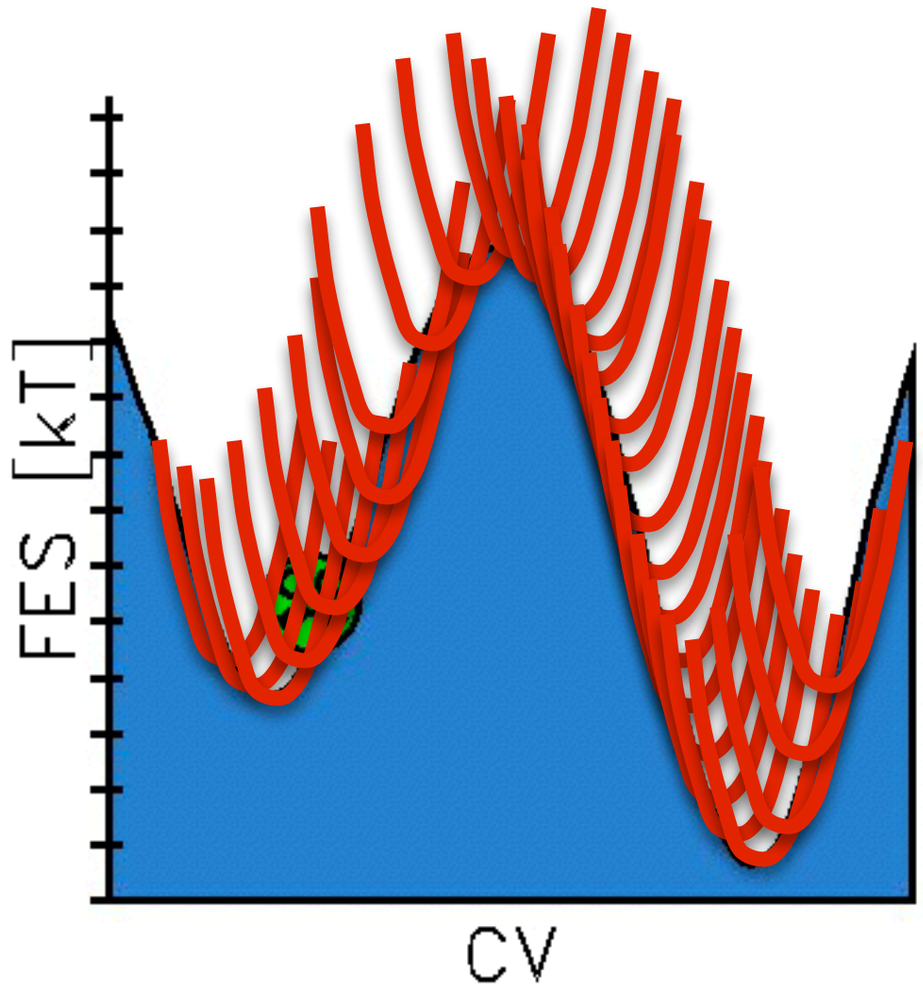
The one that leads to $F'(\mathbf{S}) = 0 \longrightarrow V(\mathbf{S}) = -F(\mathbf{S})$

Let's use an approximation of the free energy as bias potential

Torrie & Valleau JCP 1977

Multiple restraints + WHAM

The idea is to do multiple umbrella sampling calculations using harmonic restraints as bias potentials

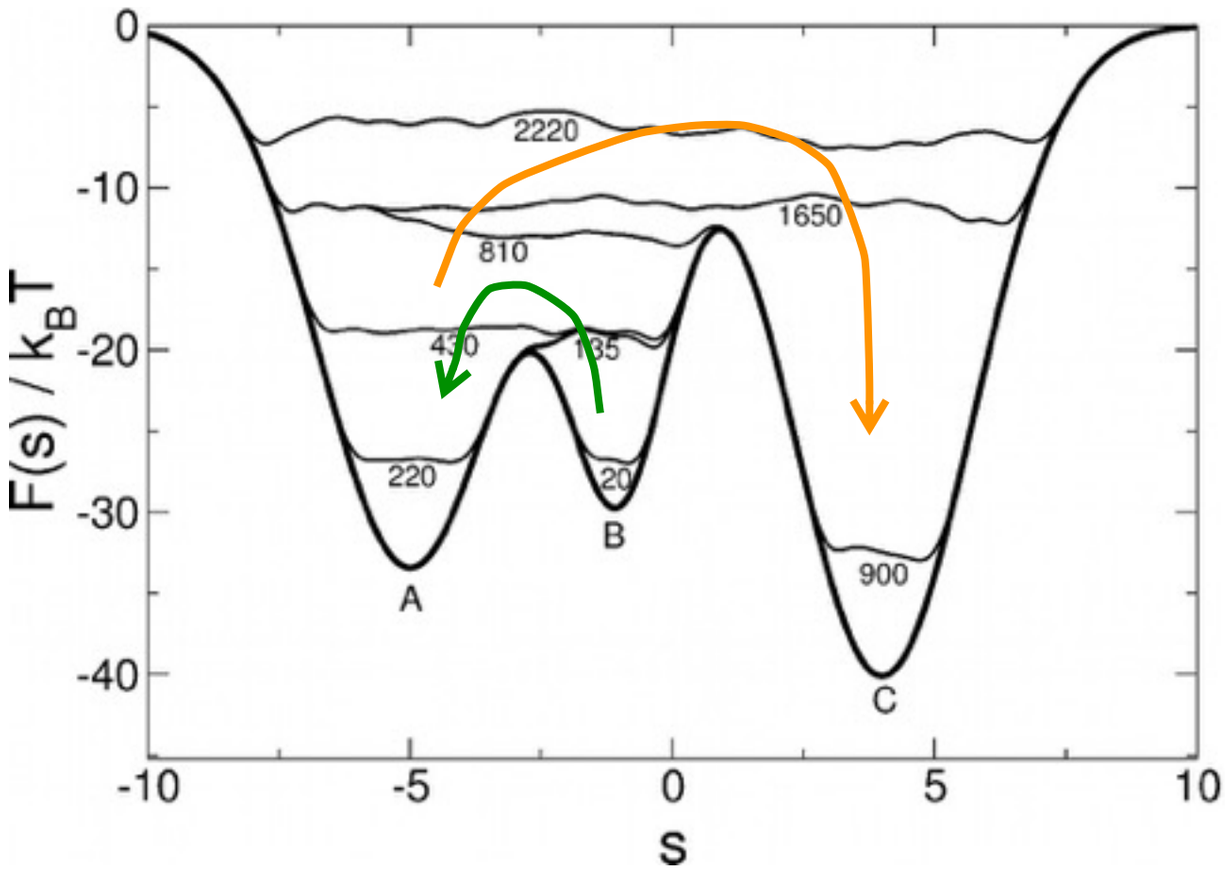


And use WHAM* to merge the biased simulations

*Ferrenberg & Swendsen PRL 1989

Metadynamics

History-dependent bias potential acting on selected degrees of freedom or Collective Variables (CVs)



$$\boldsymbol{S} = (S_1(\boldsymbol{R}), ..., S_d(\boldsymbol{R}))$$

$$V_G(\mathbf{S}, t) = W \sum_{t'=\tau_G, 2\tau_G, \ldots}^{t'<t} exp\left(-\sum_{i=1}^{d} \frac{(S_i - S_i(\mathbf{R}(t')))^2}{2\sigma_i^2}\right)$$

$$V_G(\boldsymbol{S}, t \to \infty) = -F(\boldsymbol{S}) + C$$

Laio & Parrinello PNAS 2002

REVIEW: Barducci, Bonomi, Parrinello WIREs Comput Mol Sci 2011

Pros and Cons

Advantages

- Enhanced sampling along the CVs
- Reconstruction of the FES:

$$V_G(\boldsymbol{S}, t \to \infty) = -F(\boldsymbol{S}) + C$$ Bussi, Laio, Parrinello PRL 2006

- *A priori* knowledge of the landscape not required

Disadvantages

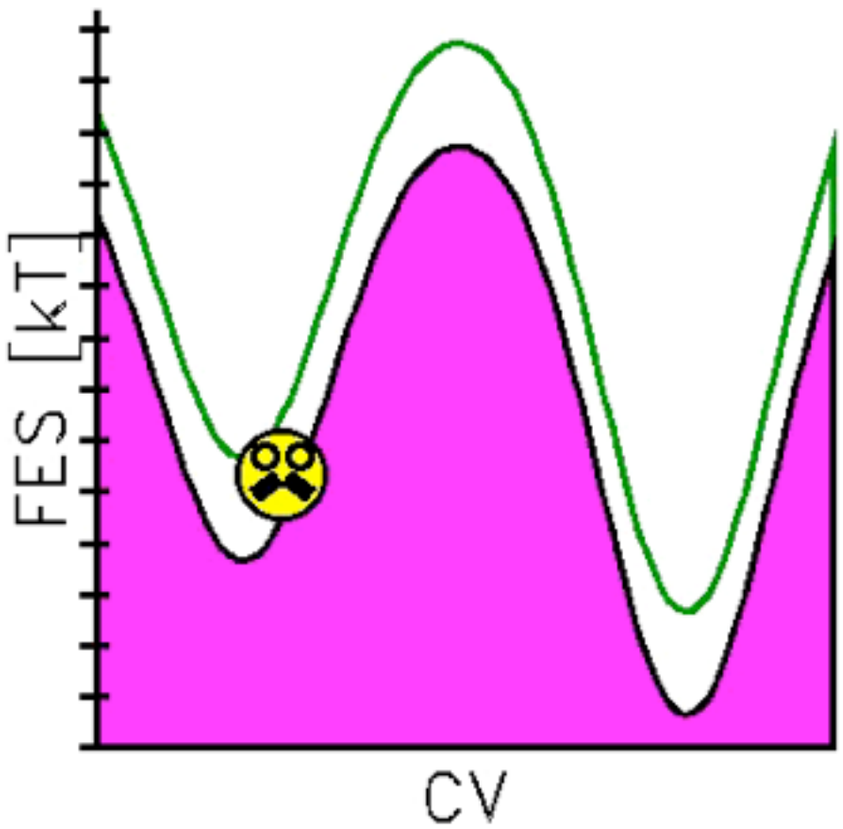
- Lack of convergence in a single run
- Overfilling
- The choice of the CVs is not trivial

Well-Tempered Metadynamics

The initial Gaussian height w_0 is rescaled during the simulation:

$$w = w_0 \, e^{-\frac{V(\boldsymbol{s},t)}{k_B \Delta T}}$$

where $T + \Delta T$ is a fictitious CV temperature.



- Convergence and overfilling issues solved:

$$V(\mathbf{s},t) \rightarrow -\frac{\Delta T}{T + \Delta T} F(\mathbf{s})$$

- ΔT used to tune the extent of exploration

Barducci, Bussi, Parrinello PRL 2008

Parallel Bias Metadynamics

Biasing a large number of CVs with WTMetaD is inefficient

In PBMetaD we apply multiple low-dimensional bias potentials:

$$V(S_1, t), ..., V(S_N, t)$$

one at a time:

$$P_t(\mathbf{R}, \eta) \propto \exp\left[-\beta\left(U(\mathbf{R}) + \sum_i \eta_i V(S_i, t)\right)\right]$$

where $\eta = (\eta_1, ..., \eta_N)$ switches on and off (and allows updating) one bias potential at a time

Each bias potential converges to the corresponding free energy:

$$V(S_i, t) \to -\frac{\Delta T}{T + \Delta T} F(S_i)$$

Pfaendtner & Bonomi JCTC 2015

Parallel Bias Metadynamics

Since we are not interested in the η-distribution, we can marginalize this variable:

$$P_t(\mathbf{R}) = \int d\eta P_t(\mathbf{R}, \eta) \propto \exp\left[-\beta \left(U(\mathbf{R}) + V_{PB}(\mathbf{S}, t)\right)\right]$$

where:

$$V_{PB}(\mathbf{S}, t) = -\frac{1}{\beta} \log \sum_{i=1}^{N} \exp - \left[\beta V(S_i, t)\right]$$

In order for each bias potential to converge to the corresponding free energy, we need a new rescaling rule:

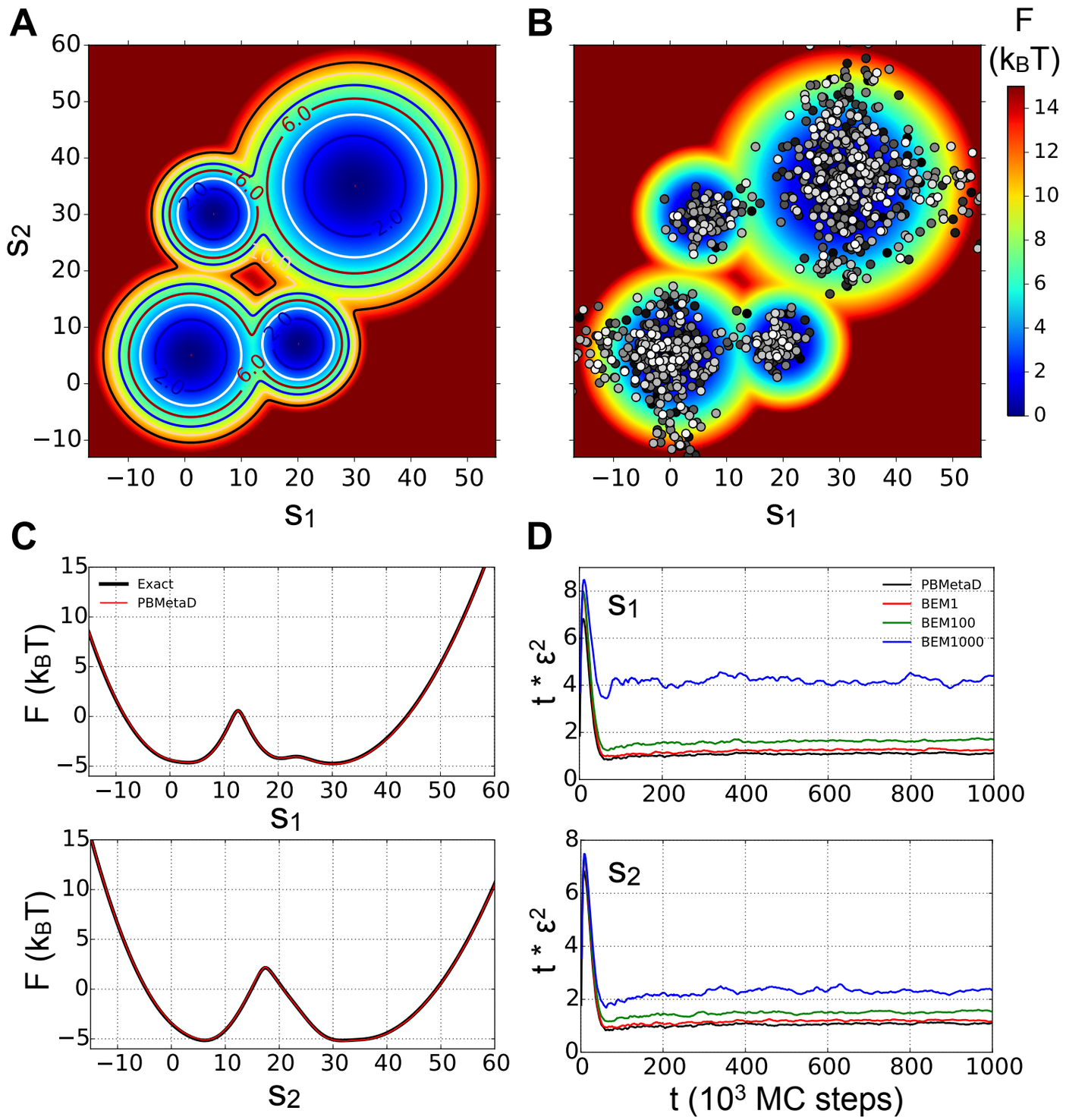
$$\omega_i = \omega_{0,i}\, e^{-\frac{V(S_i,t)}{k_B \Delta T_i}}\, P(\eta_i = 1|\mathbf{R})$$

where:

$$P(\eta_i = 1|\mathbf{R}) = \frac{\exp\left[-\beta V(S_i, t)\right]}{\sum_{j=1}^{N} \exp\left[-\beta V(S_j, t)\right]}$$

Pfaendtner & Bonomi JCTC 2015

Benchmark on a model system

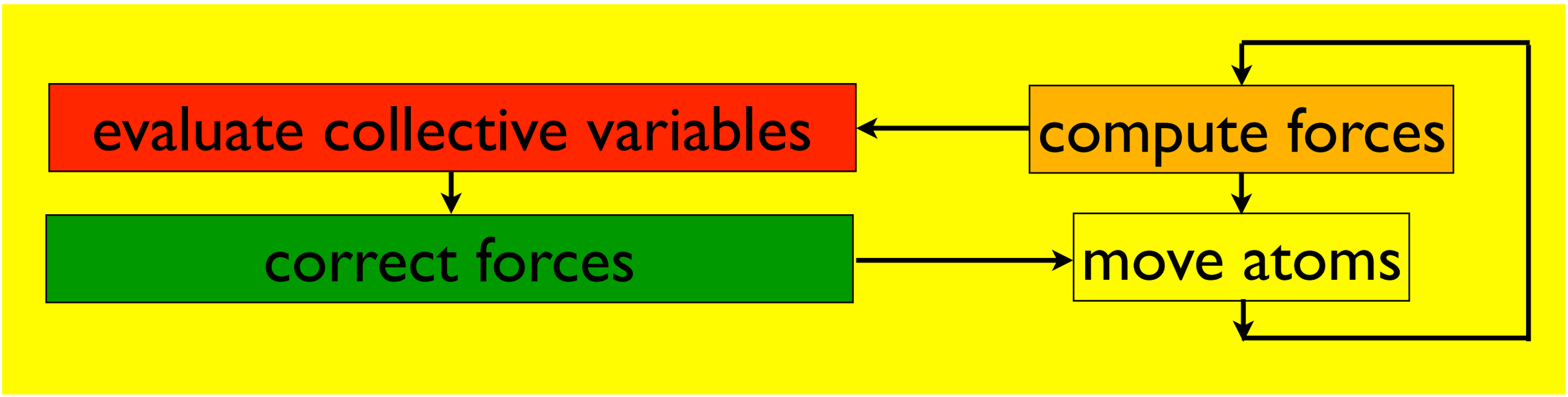


Pfaendtner & Bonomi JCTC 2015

The implementation

depending on the physical problem:
distances, angles, ...

depending on physical
problem/type of machine/...



several possible algorithms
e.g. umbrella sampling, metadynamics, ...

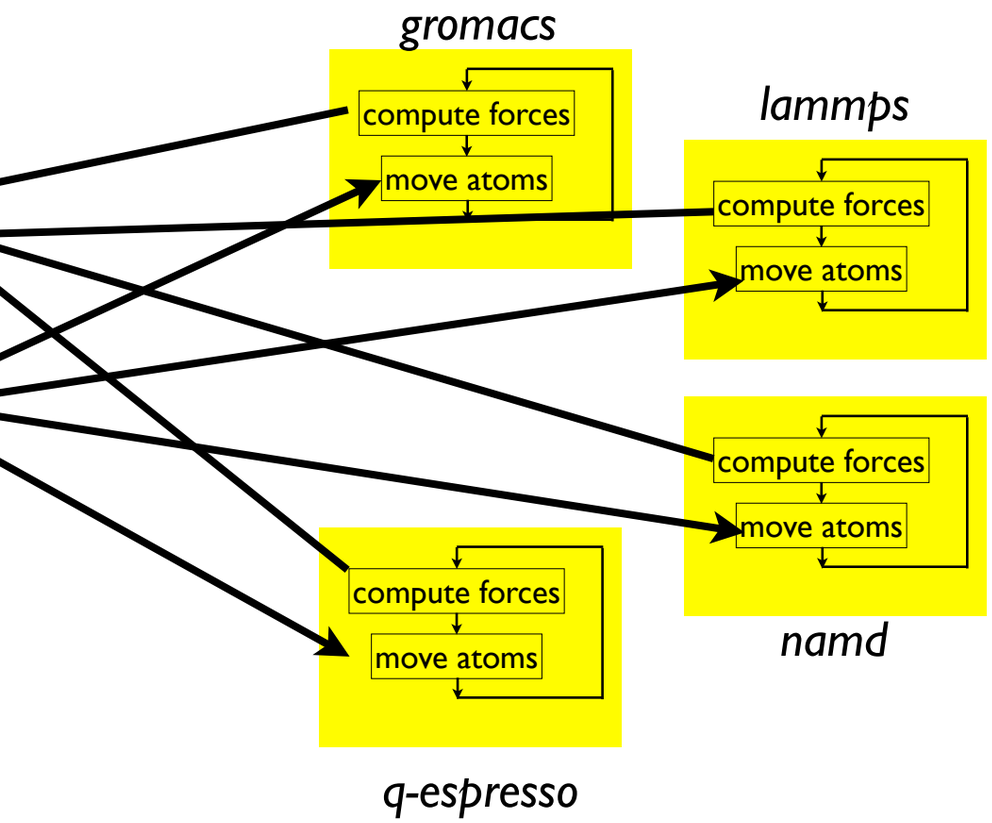
PLUMED

PLUGIN

evaluate collective variables

correct forces

MD codes



One open source pluging
for several MD codes!

Why **PLUMED**?

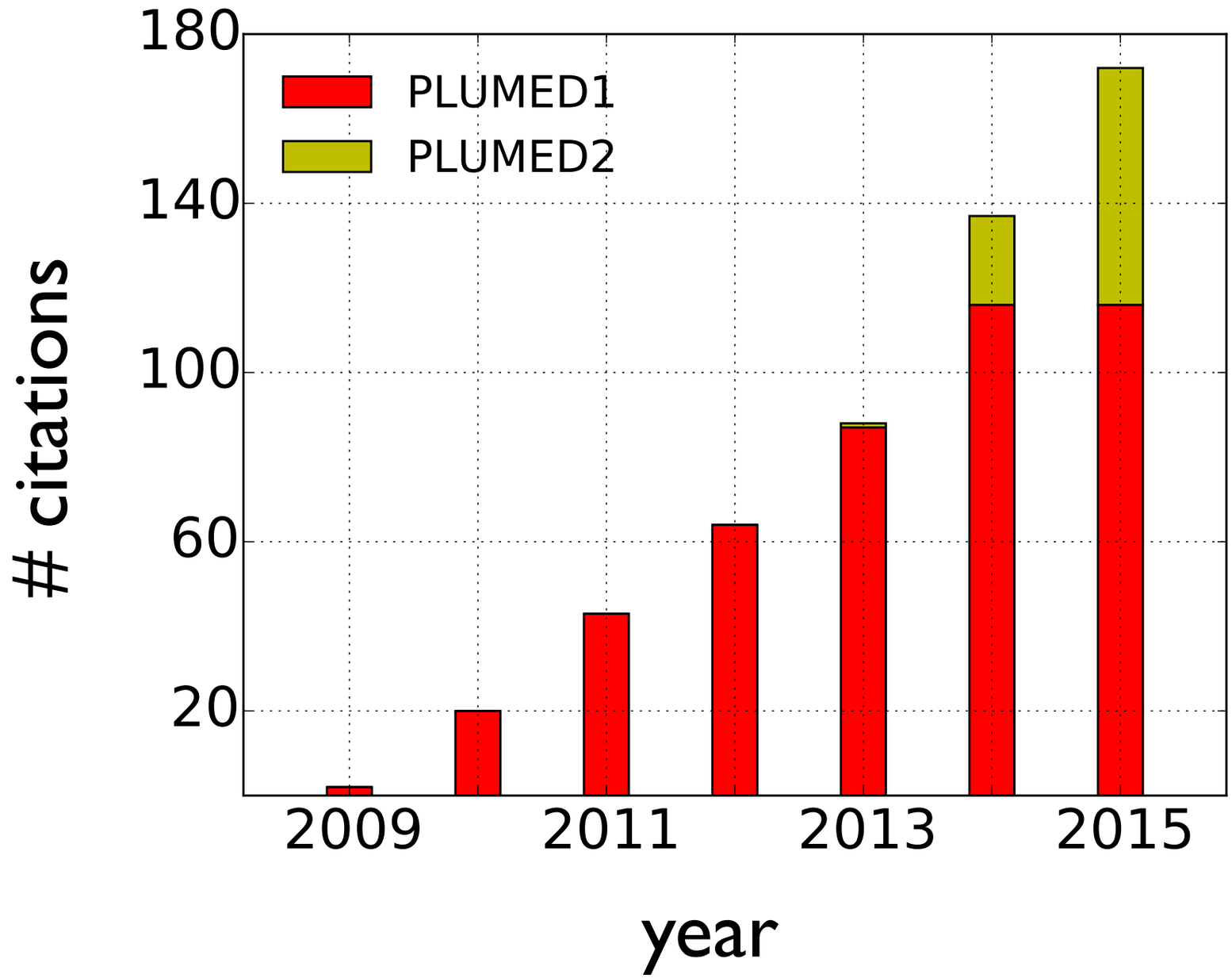
PLUgin for **MEta**Dynamics

PLUgin for free-energy **METho**Ds

PLUgin for **Mol**Ecular Dynamics

Bonomi *et al.* CPC 2008
Tribello *et al.* CPC 2014

A quickly growing community



PLUMED1 = Bonomi *et al.* CPC 2008
PLUMED2 = Tribello *et al.* CPC 2014

Source: Google Scholar (Nov 2015)

What can you do with PLUMED?

Analyze trajectories[$]

```
                    # using plumed as a standalone tool
plumed driver --igro traj.gro --plumed plumed.dat
```

Analyze simulations on the fly*

```
# e.g. using gromacs:
mdrun -plumed plumed.dat
```

Bias simulations on the fly*

```
# e.g. using gromacs:
mdrun -plumed plumed.dat
```

[$]from command line or from VMD - Giorgino, CPC (2014), http://github.com/tonigi/vmd_plumed

*used in combination with a supported MD engine, e.g. GROMACS, NAMD, LAMMPS, Q-ESPRESSO, AMBER + others

PLUMED+MD

PLUMED

MD code

read from a separate file

initialization

evaluate collective variables

correct forces

initialization

compute forces

move atoms

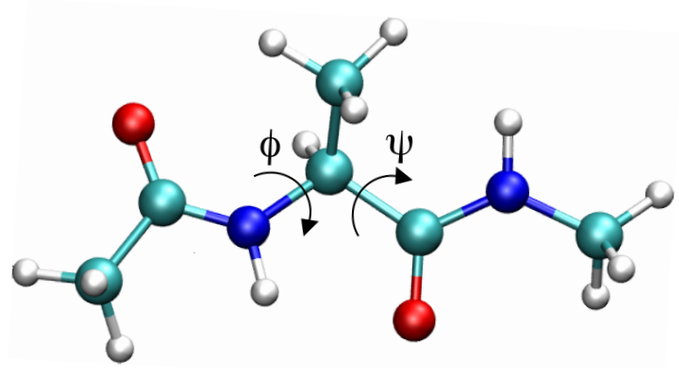
also derivatives w.r.t. atom positions

sometime using history-dependent schemes

Example of PLUMED input file

CV

```
# collective variables definition
phi:    TORSION ATOMS=5,7,9,15
psi:    TORSION ATOMS=7,9,15,17
```



BIAS

```
# activate Parallel Bias Metadynamics
PBMETAD ...
 ARG=phi,psi
 PACE=500 HEIGHT=1.2
 SIGMA=0.35,0.35
 FILE=HILLS_PHI,HILLS_PSI
 BIASFACTOR=8.0
... PBMETAD
```

OUTPUT

```
# printout
PRINT ARG=phi,psi,pbmetad.bias STRIDE=500 FILE=COLVAR
```

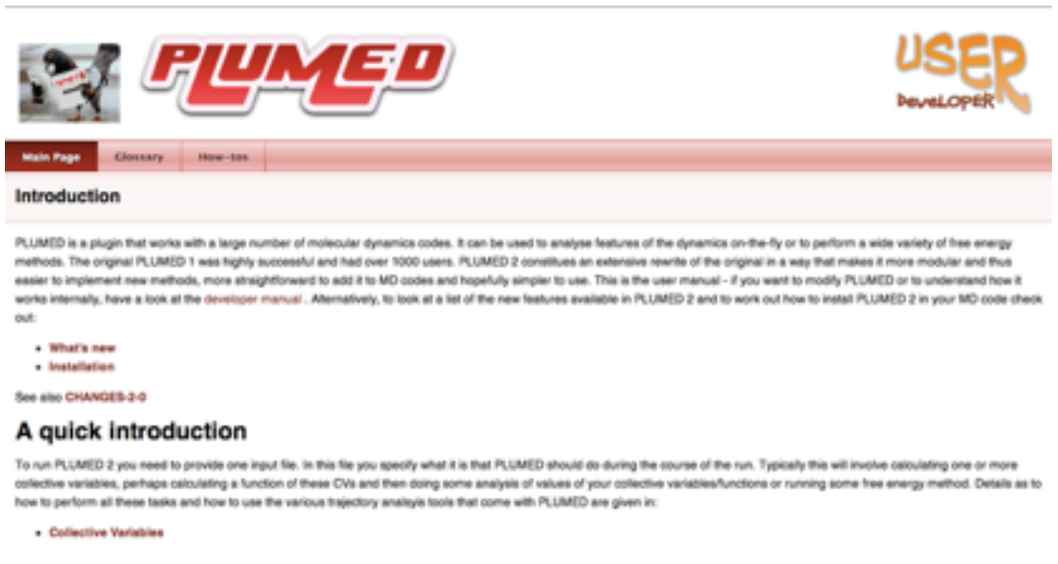
On the WEB

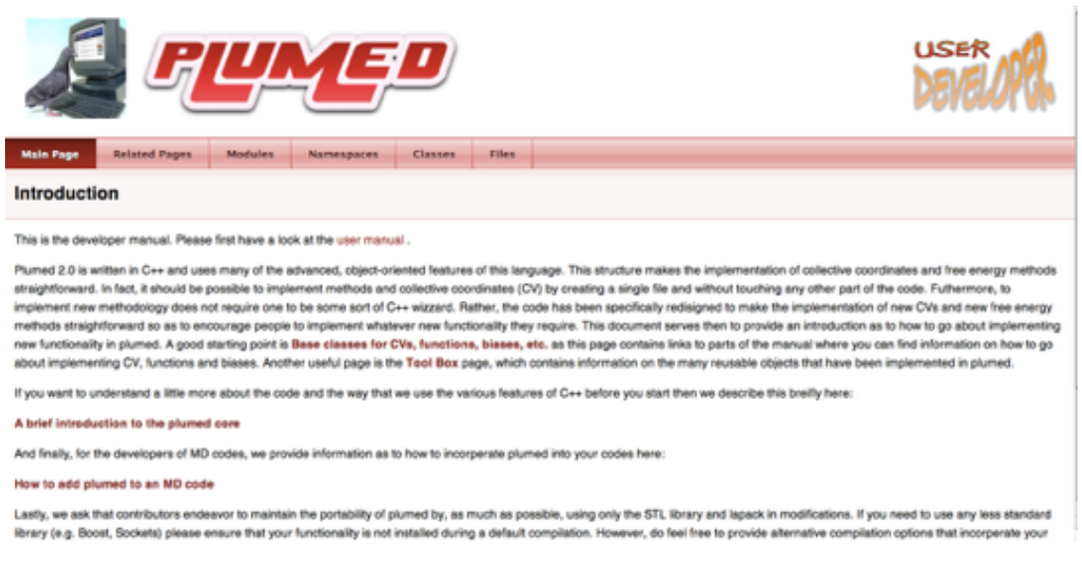
Website: http://www.plumed.org/

Github: http://github.com/plumed/plumed2

User & developer mailing lists

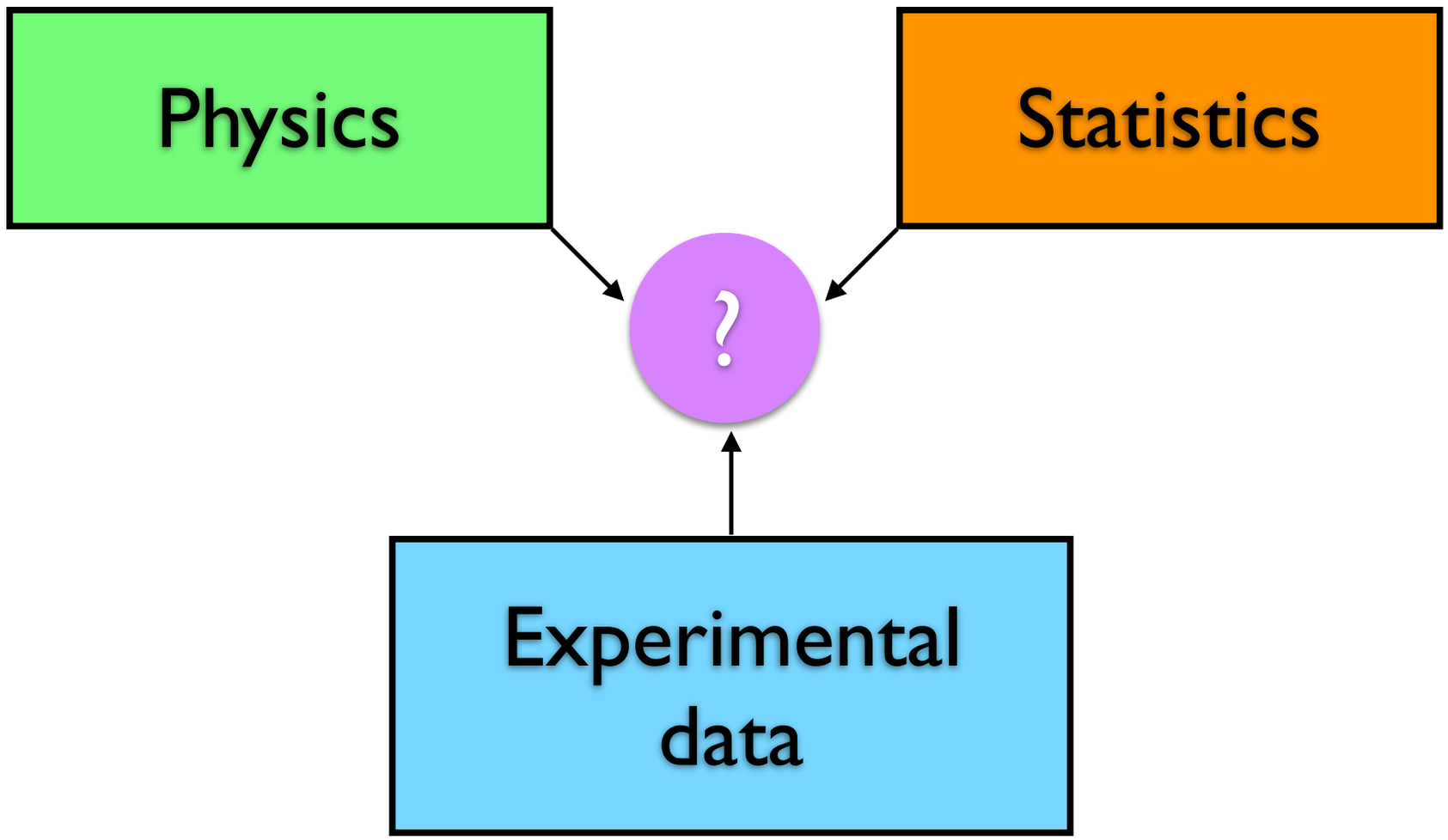
User & developer manuals + tutorials





A sneak peak at the future of PLUMED

A more accurate description of a system can be achieved if we combine all the sources of information available



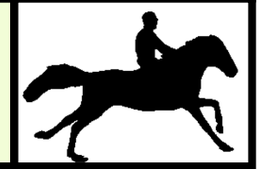
How can we properly combine them?

The challenges of data modelling

a) Input

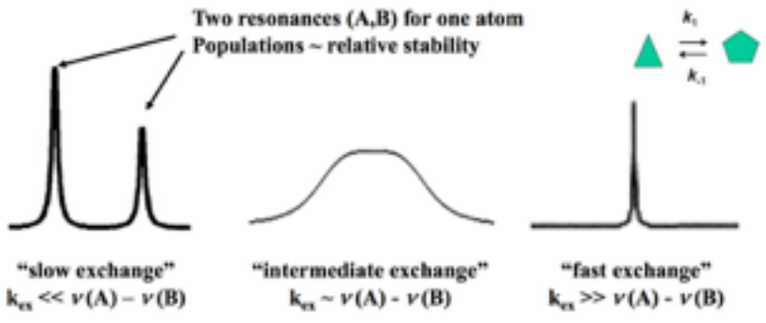
Measurement+Prior

b) Errors

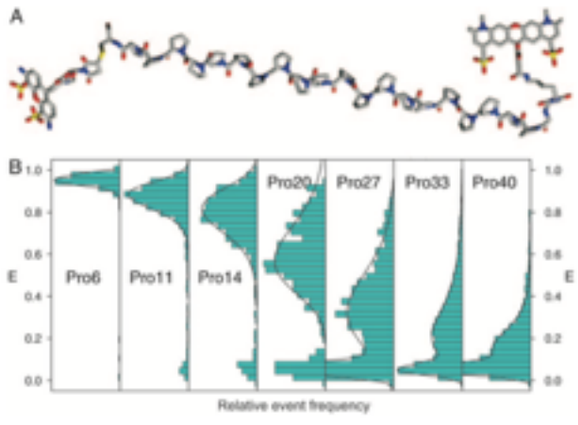
Mixing	
Random	
Systematic	
Forward model	

EXAMPLES

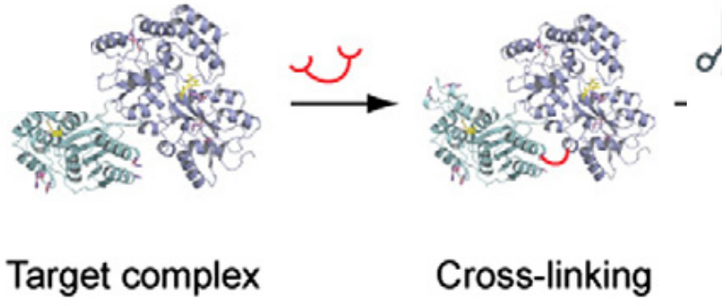
NMR



FRET distributions



XL/MS



Addressing these challenges

Mixing		Replica-Averaged Modelling+ +Camilloni *et al.* JACS 2012
Random		Bayesian Modelling* *Rieping *et al.* Science 2005
Systematic		
Forward model		

Replica-Averaged Modelling

Find the minimal perturbation of the prior information that reproduces exactly the experimental data, assuming that data are averaged over multiple conformations

The maximum entropy principle (MEP) recipe is to add an harmonic restraint between experimental and predicted data:

$$E_{RAM}(\mathbf{X}) = E_{MD}(\mathbf{X}) + \frac{1}{2}k(d - f(\mathbf{X}))^2$$

where predicted data is averaged over multiple replicas of the system:

$$f(\mathbf{X}) = \frac{1}{N}\sum_{r=1}^{N} f(X_r)$$

The intensity of the restraint should be as strong as possible and should scale more than linearly with the number of replicas *N*.

Camilloni *et al.* JACS 2012

Bayesian Modelling

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

posterior **likelihood** **priors**

The **model** comprises the structure coordinates and additional parameters (noise level, calibration...).

The **likelihood** function encodes the agreement with the data *d* (through a **forward model**) and provides a model for the noise.

The **priors** define the probability of model, given any knowledge other than the data.

The **Bayesian score** is: $E_{Bayes}(X, \sigma) = -k_B T \cdot \log p(X, \sigma | d)$

Rieping *et al.* Science 2005

Addressing these challenges

Mixing	
Random	
Systematic	
Forward model	

Metainference*

*Bonomi *et al.* Arxiv 2015

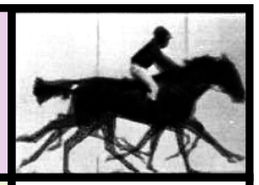
To produce ensemble of models

and determine their populations

a) Input

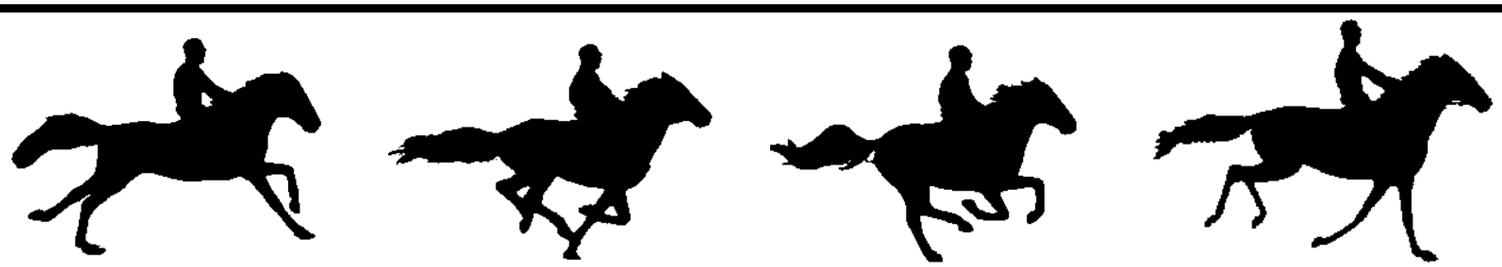
Measurement+Prior

b) Errors

Mixing	
Random	
Systematic	
Forward model	

c) Output

Metainference: Ensemble of models



30% 20% 45% 5%

Metainference

Inspired by replica-averaged modelling, we consider a finite sample of the distribution of models (N replicas):

$$f(\mathbf{X}) = \frac{1}{N}\sum_{r=1}^{N} f(X_r)$$

The Metainference energy function (or score) is:

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

where $\sigma_{r,i}$ includes all sources of errors:

$$\sigma_{r,i} = \sqrt{(\sigma_{r,i}^{SEM})^2 + (\sigma_{r,i}^{B})^2}$$

and $\sigma_{r,i}^{SEM} \propto 1/\sqrt{N}$

errors are negligible

Replica-Averaged Modelling

data is not generated by an ensemble

Bayesian Modelling

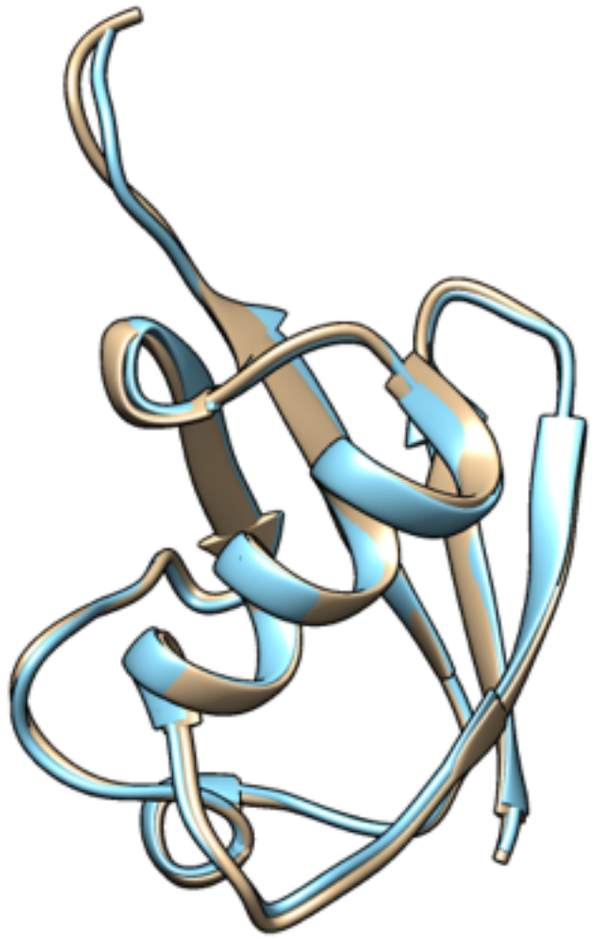
Integrative Dynamical Biology

We compare Metainference and replica-averaged modeling with real experimental data collected on ubiquitin:

- Chemical Shifts + RDCs

We also compare the Metainference ensemble with single structures:

- X-ray (1UBQ)
- NMR (1D3Z)



Cα-RMSD = 0.52 Å

and with the ensemble generated by standard MD

Models are evaluated by fit with other exp data (RDCs, J3)

Technical details

- CHARMM22* with TIP3P explicit solvent, ~25000 atoms

  Piana *et al.* Biophys. J. 2011 Jorgensen *et al.* JCP 1983

- GROMACS 4.6.7 + PLUMED 2 (development branch)

  Hess *et al.* JCTC 2008 Tribello *et al.* CPC 2014

- Double parallelization: 8 replicas (ensemble modelling) x 8 cores per replica

- Non-bonded interactions cutoff at 0.9 nm + PME

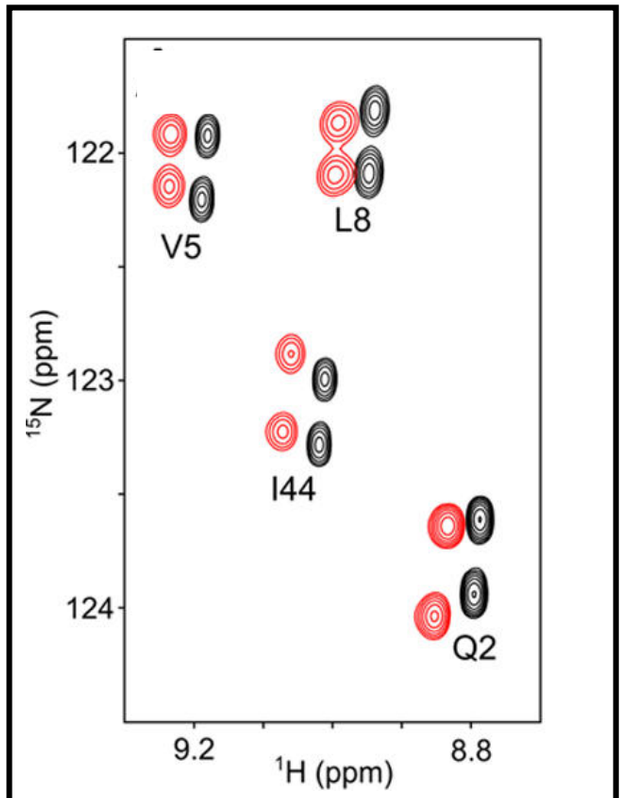
- NVT with Bussi-Donadio-Parrinello thermostat Bussi *et al.* JCP 2007

- Chemical Shifts predicted by Camshifts Kohlhoff *et al.* JACS 2009

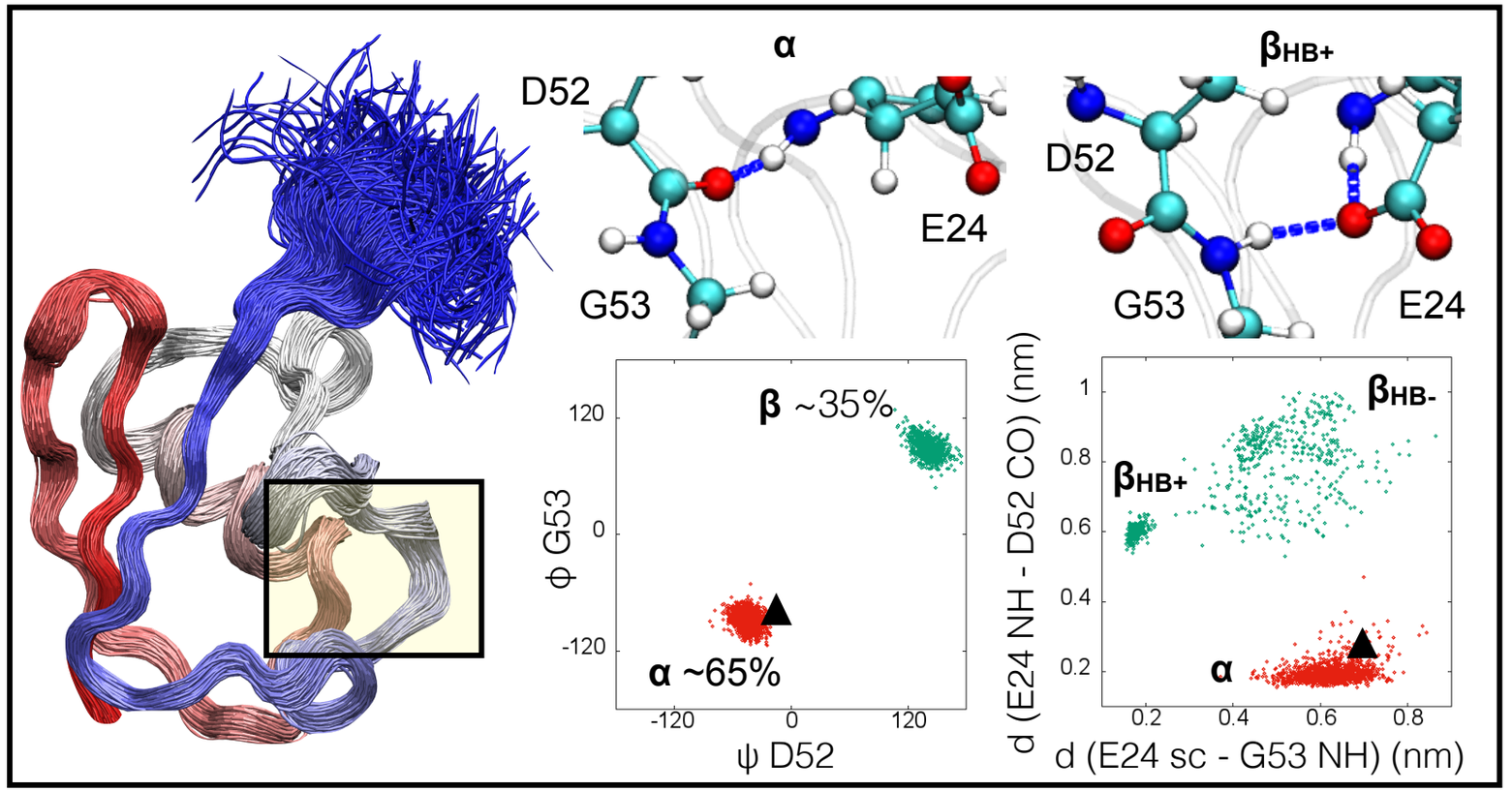
- RDC predicted by θ-method Camilloni & Vendruscolo JPCB 2015

Ubiquitin ensembles

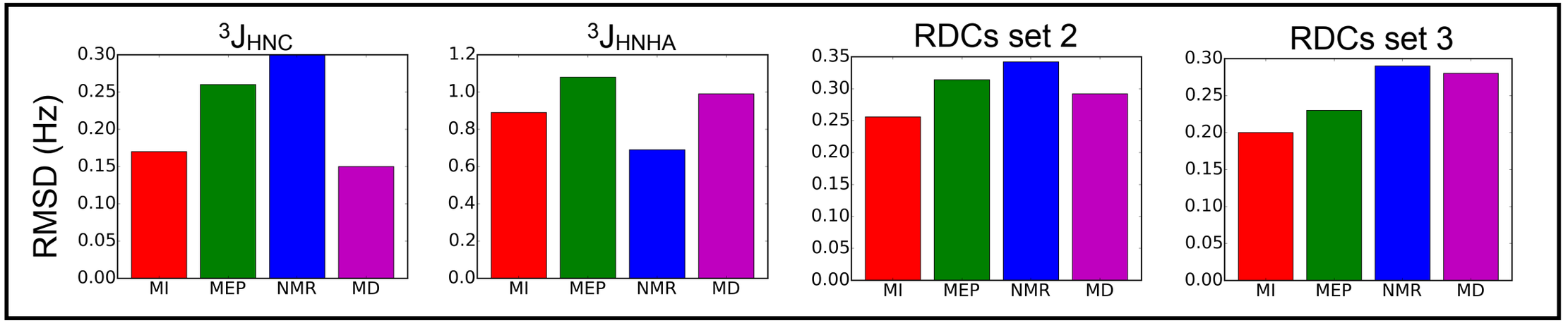
a) Input



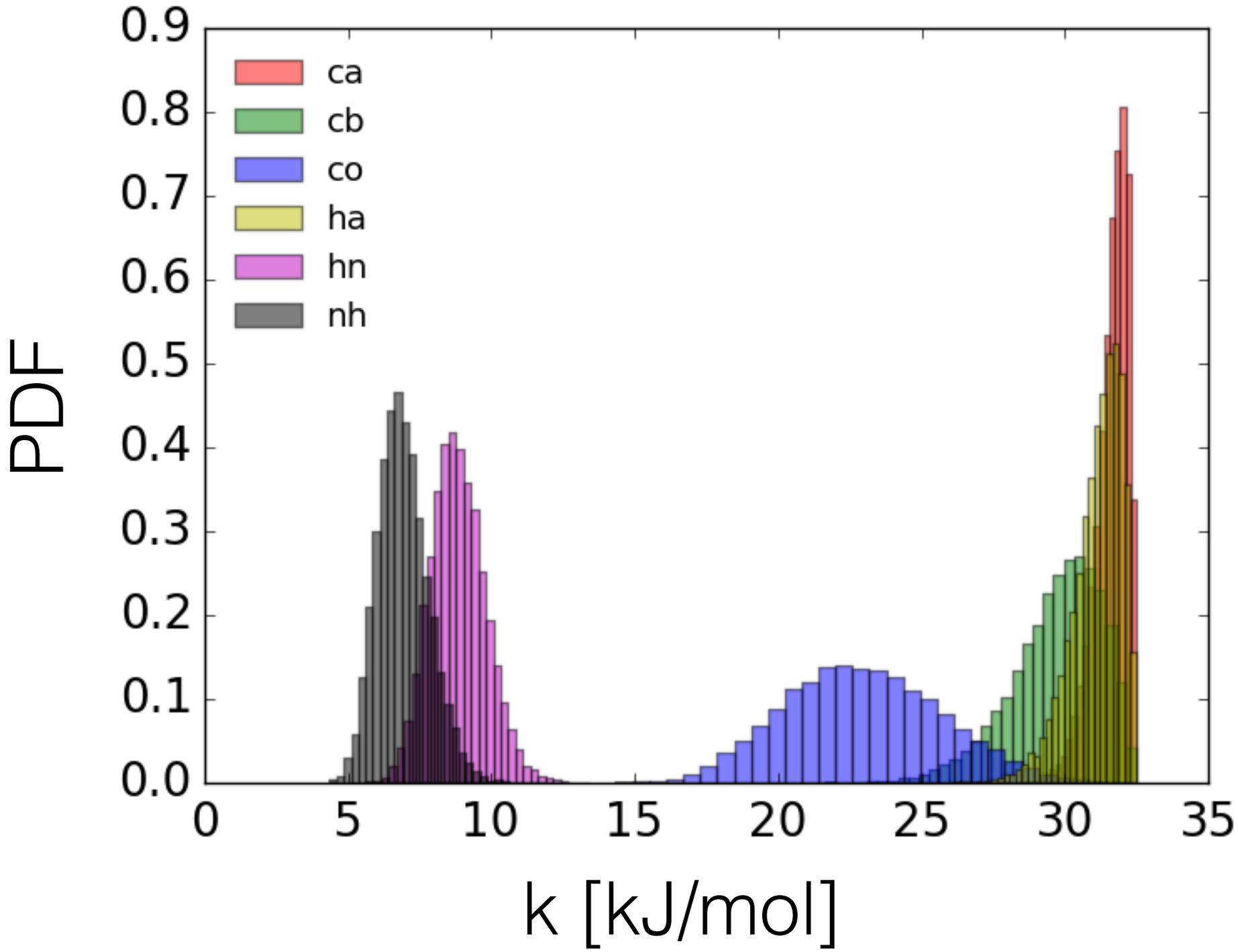
b) Ensemble



c) Validation



Chemical Shifts weights



Conclusions

PLUMED is a open source library:

- to analyze MD simulations, on-the-fly and a posteriori
- to bias MD simulations and accelerate sampling
- compatible with many popular MD codes

Contributions are welcome!

www.plumed.org

Metainference is a method to optimally combine multiple sources of information on a system, including prior information (from physics or statistics) and experimental data collected in equilibrium conditions

To study heterogeneous systems in which relevant states are separated by high free-energy barriers, Metainference can be coupled with Parallel Bias Metadynamics

Acknowledgments

Carlo Camilloni
Andrea Cavalli

Paolo Arosio
Thomas Mueller

Jim Pfaendtner

Michele Vendruscolo

Tuomas Knowles

Davide Branduardi
Giovanni Bussi

Carlo Camilloni
Gareth Tribello

All of you for your attention