Simulating Biomolecules with Variable Protonation State: Constant-pH Molecular Dynamics Simulations with NAMD

Brian Radak

Leadership Computing Facility, Argonne National Laboratory

Computational Biophysics Workshop –
Enhanced Sampling and Free Energy Calculations
University of Illinois, Urbana-Champaign
September 26, 2017

Acknowledgements (other people to blame)

Univ of Chicago

- Benoît Roux
- Donghyuk Suh
- Huan Rui
- Yunjie Chen

ALCF

- Sunhwan Jo (SilcsBio)
- Wei Jiang

UIUC

- Jim Phillips
- Chris Chipot
- Abhi Singharoy

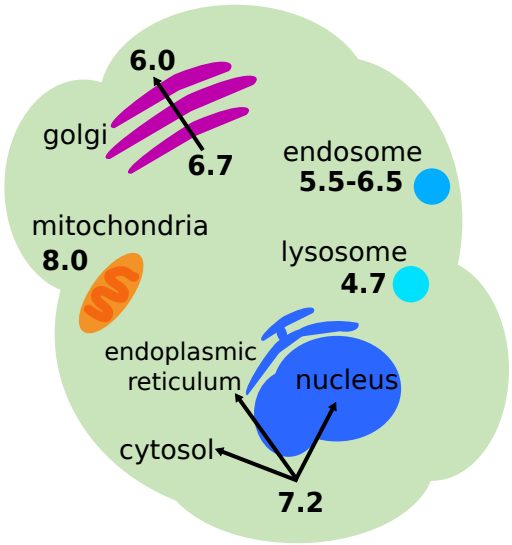
Argonne NATIONAL LABORATORY

Theta Early Science Program

npsdc MEMBRANE PROTEIN STRUCTURAL DYNAMICS GATEWAY

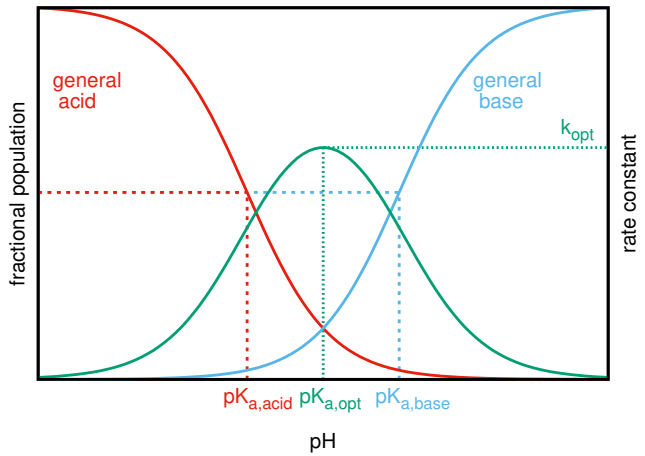
pH Effects in Biochemistry

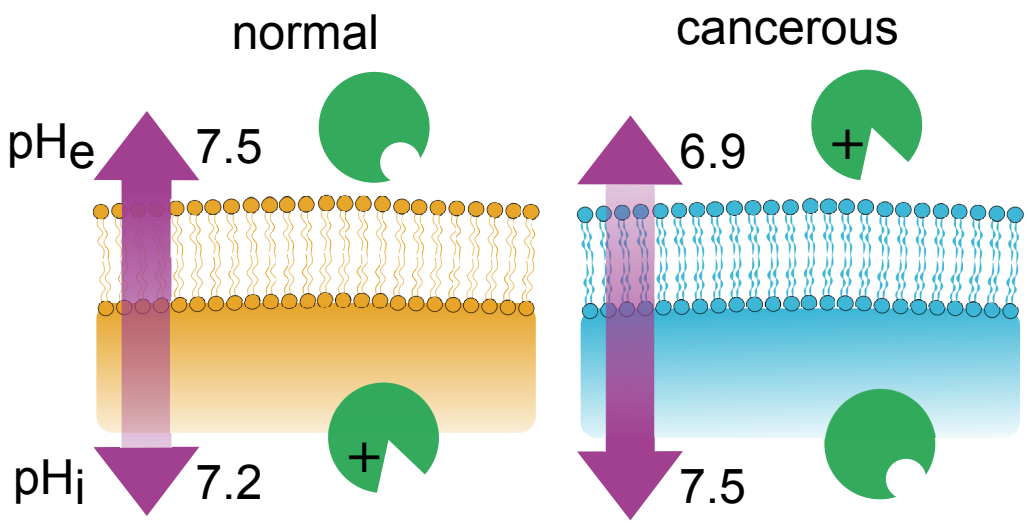
Casey, *et al Nat Rev Mol Cell Biol*, **2010**



variability of pH by region

enzyme rate vs. pH





Webb, *et al Nat Rev Cancer*, **2011**

pH gradients at cell surfaces

Constant pH and the semi-grand canonical ensemble

- Conventional MD samples a canonical ensemble:

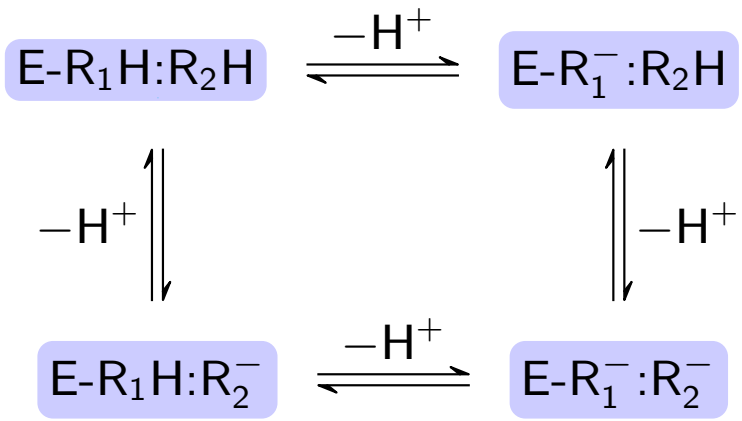
$$Q = \int d\boldsymbol{x}\, e^{-\beta U(\boldsymbol{x})}$$

- Constant-pH MD samples a semi-grand canonical ensemble:

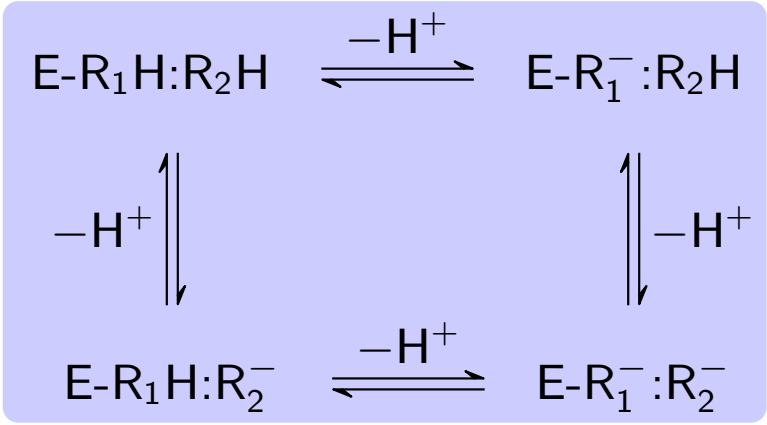
$$\Xi(\mathrm{pH}) = \sum_{\boldsymbol{\lambda} \in \mathcal{S}} Q_{\boldsymbol{\lambda}} 10^{-n_{\boldsymbol{\lambda}} \mathrm{pH}}$$

The added interaction is between the number of protons, $n_{\boldsymbol{\lambda}}$, and a pH bath. $\boldsymbol{\lambda}$ is a new variable designating the protonation state.

Networks of protonation states

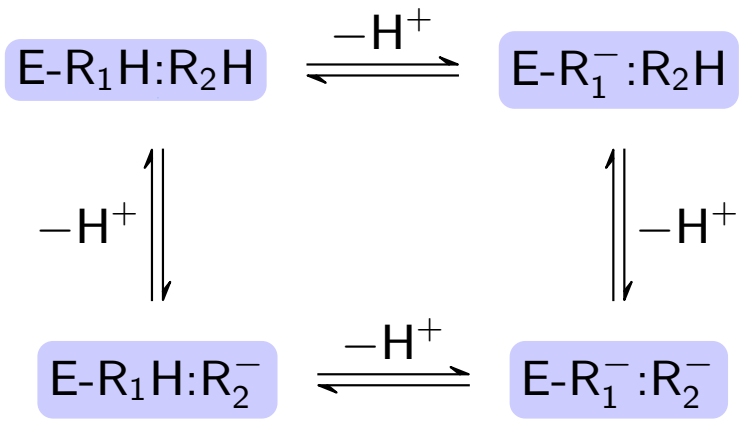


conventional MD

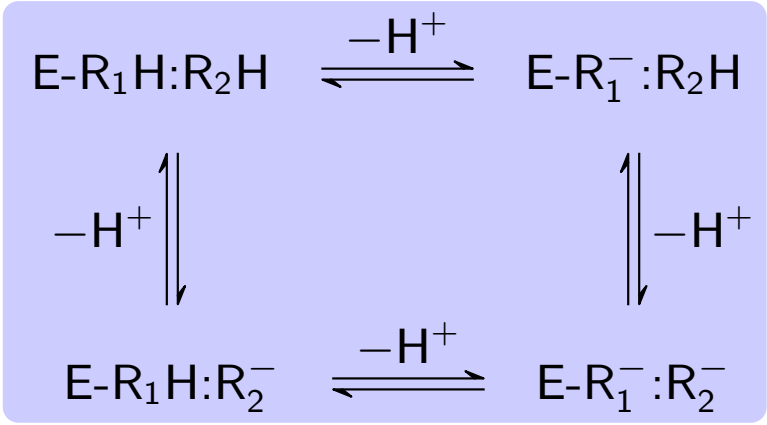


constant pH MD

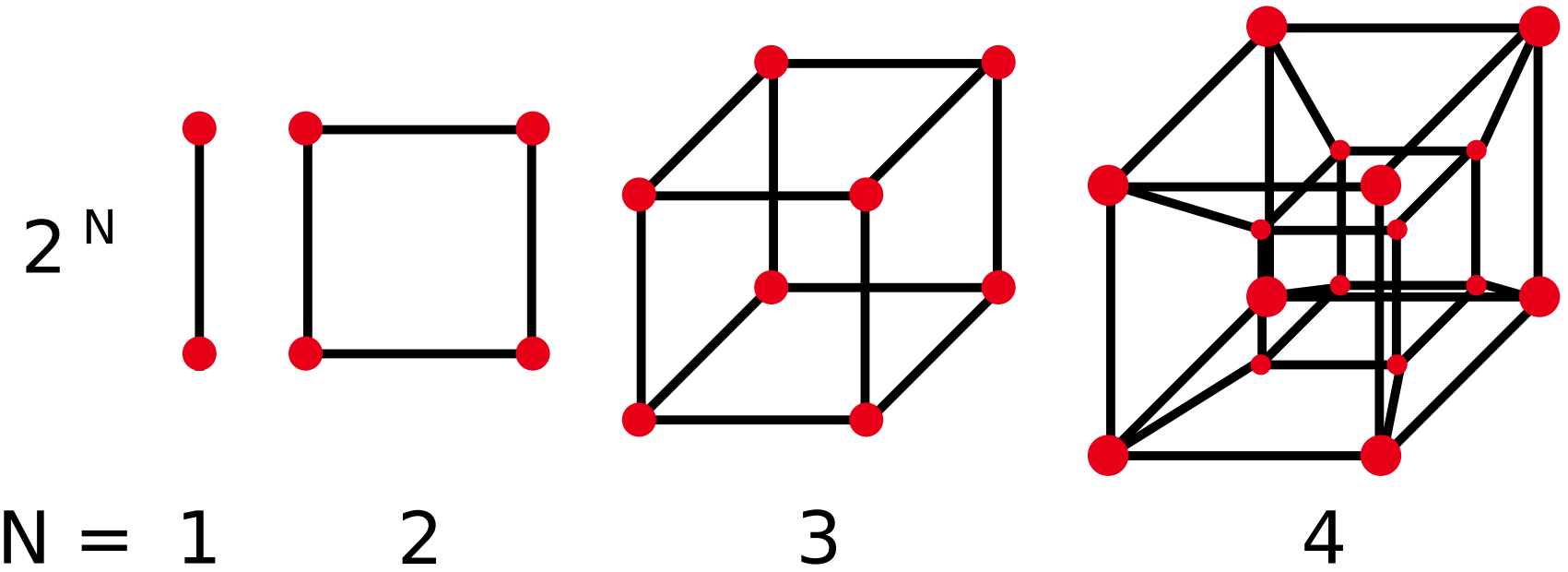
Networks of protonation states



conventional MD



constant pH MD



pH as a *thermodynamic* force

- Classical MD utilizes *mechanical* forces

$$\boldsymbol{F} = -\nabla U[\boldsymbol{x}(t)] = m\frac{\partial \boldsymbol{v}}{\partial t}; \quad \boldsymbol{v} = \frac{\partial \boldsymbol{x}}{\partial t}$$

- pH may be regarded as a *thermodynamic* force

$$\ln 10 \mathrm{pH} = -\frac{\partial \ln \Xi}{\partial n_{\boldsymbol{\lambda}}}$$

Mechanical forces – deterministic/stochastic dynamics
Thermodynamic forces – probabilistic "dynamics"

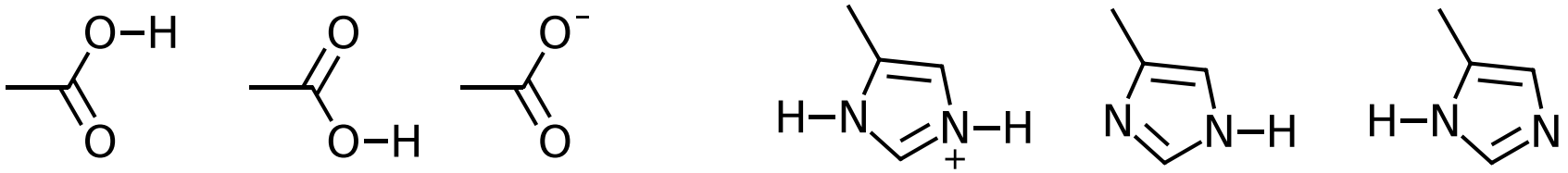
$$P_{\boldsymbol{\lambda}}(\mathrm{pH}) \propto Q_{\boldsymbol{\lambda}} 10^{-n_{\boldsymbol{\lambda}} \mathrm{pH}}$$

How do we define nodes in the network?

Consider a system with m sites:

$$\boldsymbol{\lambda} = \{\underbrace{\lambda_1, \lambda_2;}\ldots\underbrace{\lambda_s, \lambda_{s+1};}\ldots\lambda_m\}$$

$(1,0)$ $(0,1)$ $(0,0)$ $(1,1)$ $(1,0)$ $(0,1)$



Protonation state probabilities/populations

$$\langle A(\boldsymbol{x},\boldsymbol{\lambda})\rangle_{\mathrm{pH}} = \frac{\sum_{\boldsymbol{\lambda}\in\mathcal{S}} \int d\boldsymbol{x}\, A(\boldsymbol{x},\boldsymbol{\lambda}) e^{-\beta U(\boldsymbol{x};\boldsymbol{\lambda})} 10^{-n_{\boldsymbol{\lambda}}\mathrm{pH}}}{\Xi(\mathrm{pH})}$$

$P_{\lambda_s} = \langle \lambda_s \rangle_{\mathrm{pH}}$ – the probability that site s is occupied

There are two kinds of terms in the summation, $\lambda_s = 0/1$

$$\Xi(\mathrm{pH}) = \Xi_0(\mathrm{pH}) + \Xi_1(\mathrm{pH}) 10^{-\mathrm{pH}}$$

thus,

$$\langle \lambda_s \rangle_{\mathrm{pH}} = \frac{\Xi_1(\mathrm{pH}) 10^{-\mathrm{pH}}}{\Xi_0(\mathrm{pH}) + \Xi_1(\mathrm{pH}) 10^{-\mathrm{pH}}} = \frac{1}{1 + \frac{\Xi_0(\mathrm{pH})}{\Xi_1(pH)} 10^{\mathrm{pH}}}$$

Connection to thermodynamics

$$\langle \lambda_s \rangle_{\mathrm{pH}} = \frac{1}{1 + \frac{\Xi_0(\mathrm{pH})}{\Xi_1(pH)} 10^{\mathrm{pH}}}$$

compares to the Henderson-Hasselbalch equation such that

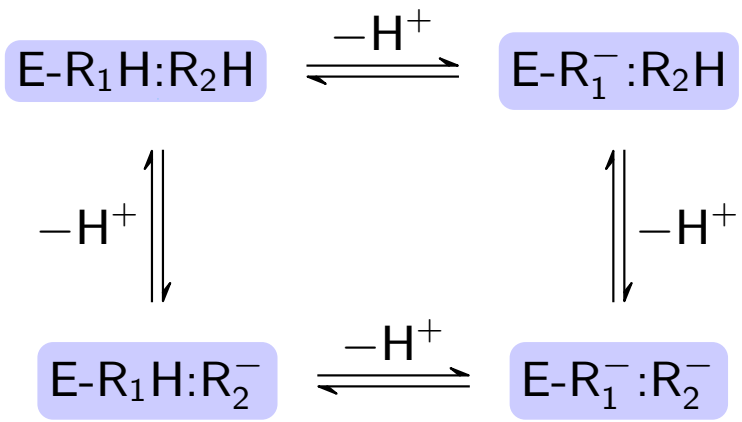
$$\mathrm{p}K_{\mathrm{a}}(\mathrm{pH}) = -\log \frac{\Xi_0(\mathrm{pH})}{\Xi_1(\mathrm{pH})},$$

except that now $\mathrm{p}K_{\mathrm{a}}(\mathrm{pH})$ is pH *dependent*. One often uses the approximation:

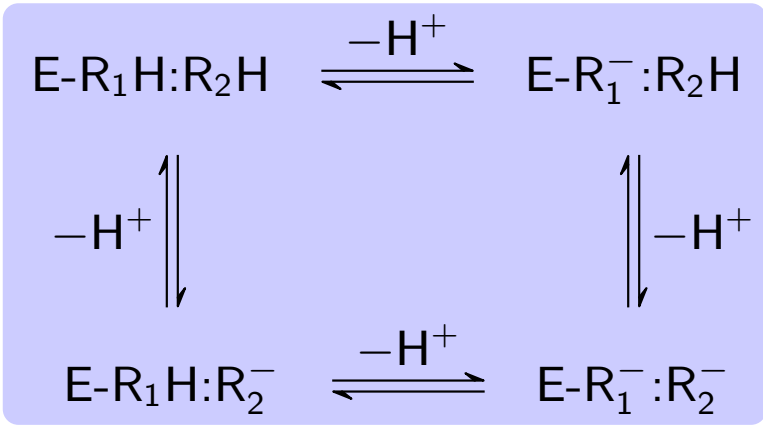
$$\mathrm{p}K_{\mathrm{a}}(pH) \approx \mathrm{p}K_{\mathrm{a}}^{(\mathrm{a})} + (1 - n)\left(\mathrm{pH} - \mathrm{p}K_{\mathrm{a}}^{(\mathrm{a})}\right),$$

where n is the Hill coefficient and $\mathrm{p}K_{\mathrm{a}}^{(\mathrm{a})}$ is the "apparent" $\mathrm{p}K_{\mathrm{a}}$.

Networks of protonation states



conventional MD



constant pH MD

We can now see that the fraction of simulation time spent in a given protonation state is directly impacted by the *difference* of the pK_a of a residue/site and the pH.

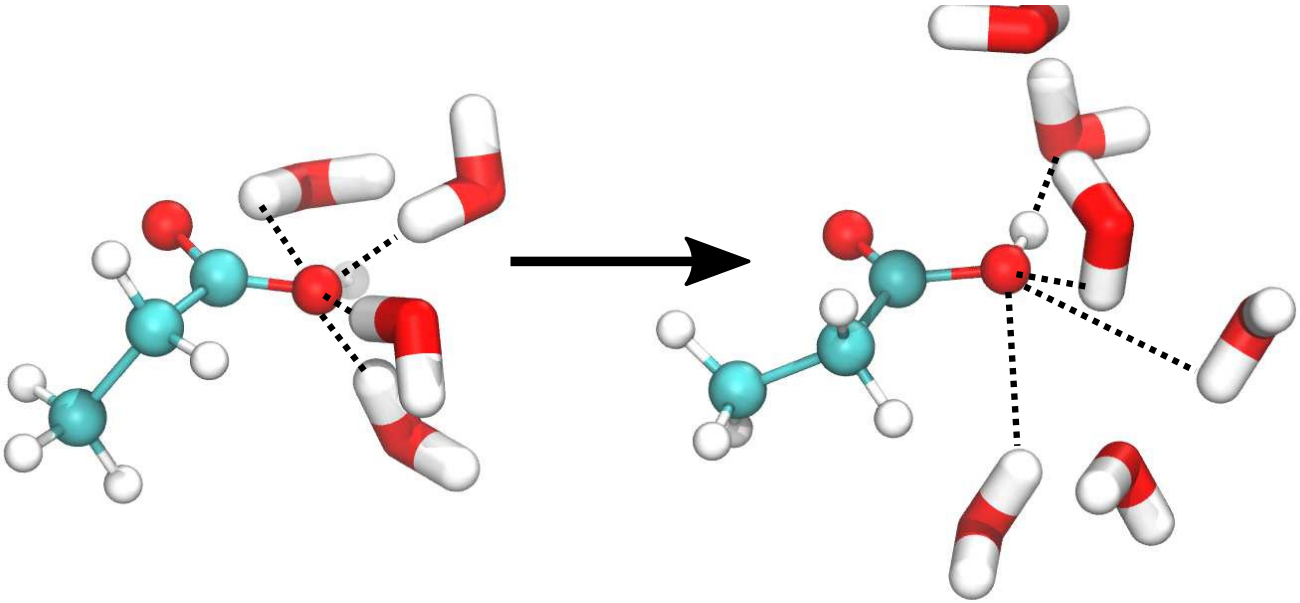
That's great – how do we sample the states?

1. Sample the configuration space of a given state
   (*i.e.*, sample $\boldsymbol{x}$ for a given $Q_{\boldsymbol{\lambda}}$)

2. Change between protonation states according to the number of protons and the given pH
   (*i.e.*, sample $\boldsymbol{\lambda}$ and choose a new $Q_{\boldsymbol{\lambda}}$)

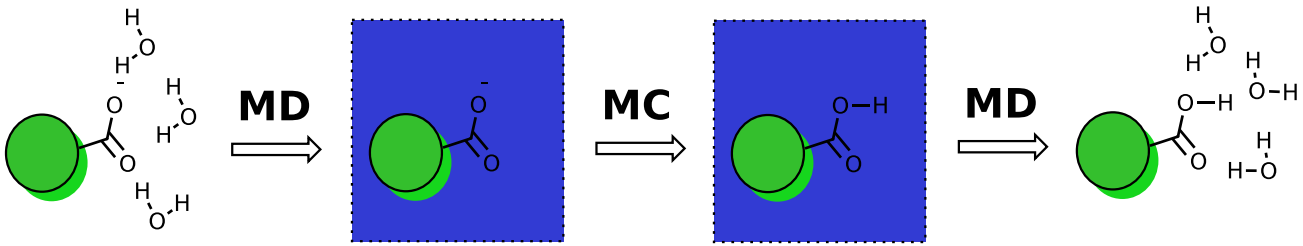
This may be regarded as a **Gibbs sampling**, whereby the configuration and state are sampled in an *alternating* fashion.

A problem! Environmental response



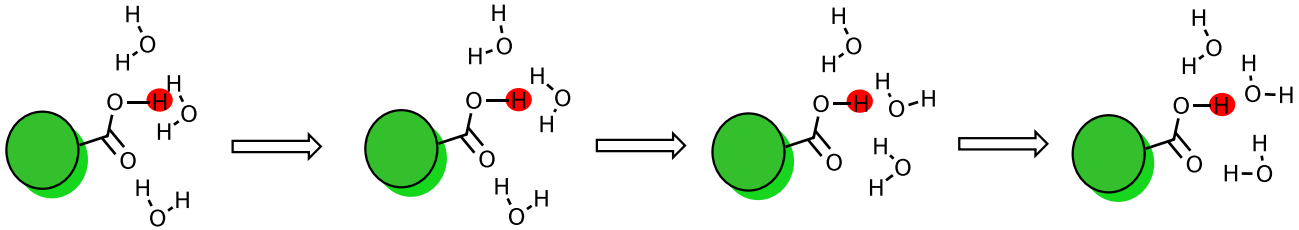
- (De)Protonation is a significant electrostatic event.
- Non-trivial reorganization of solvent, possibly solute.
- Naive sudden changes in protonation are likely to cause high energy configurations and/or steric clashes.

Possible solutions to the solvent clash problem



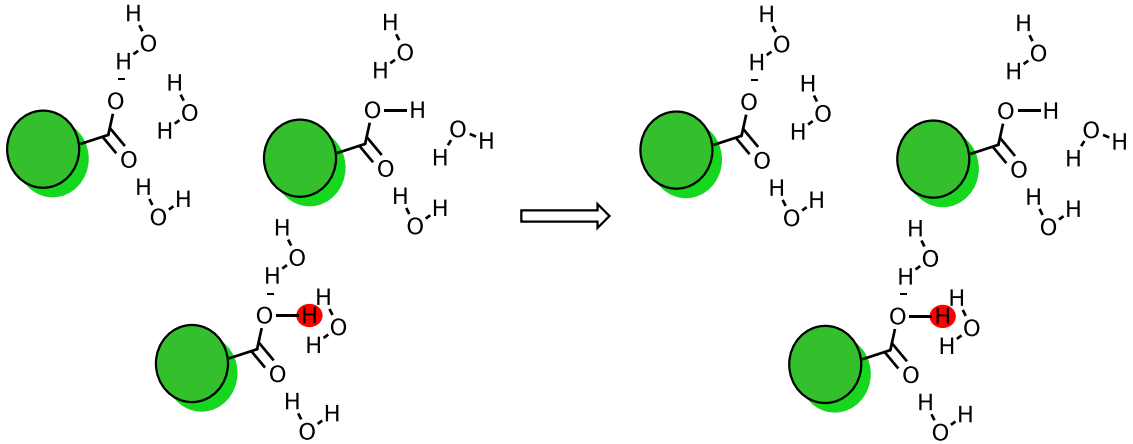
auxillary implicit solvent

Baptista, et al. **2002**.
Swails, et al. **2014**.



continuous fractional proton

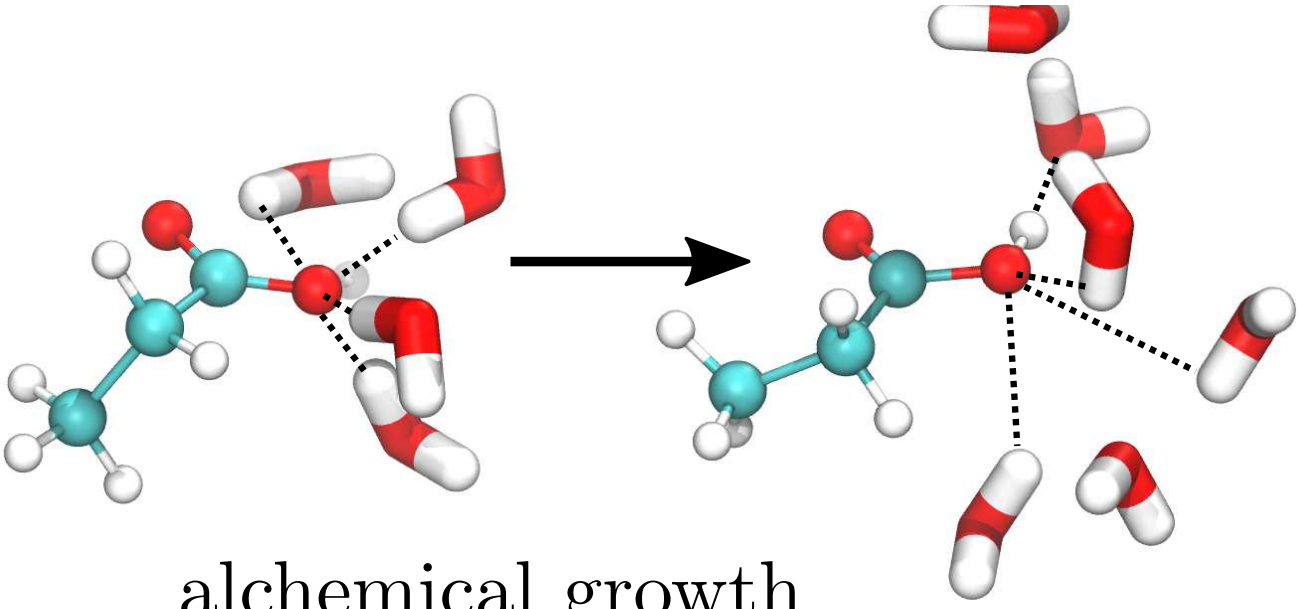
Lee, et al. **2004**,
Donnini, et al. **2011**.



discrete copy fractional proton

Lee, et al. **2014**.

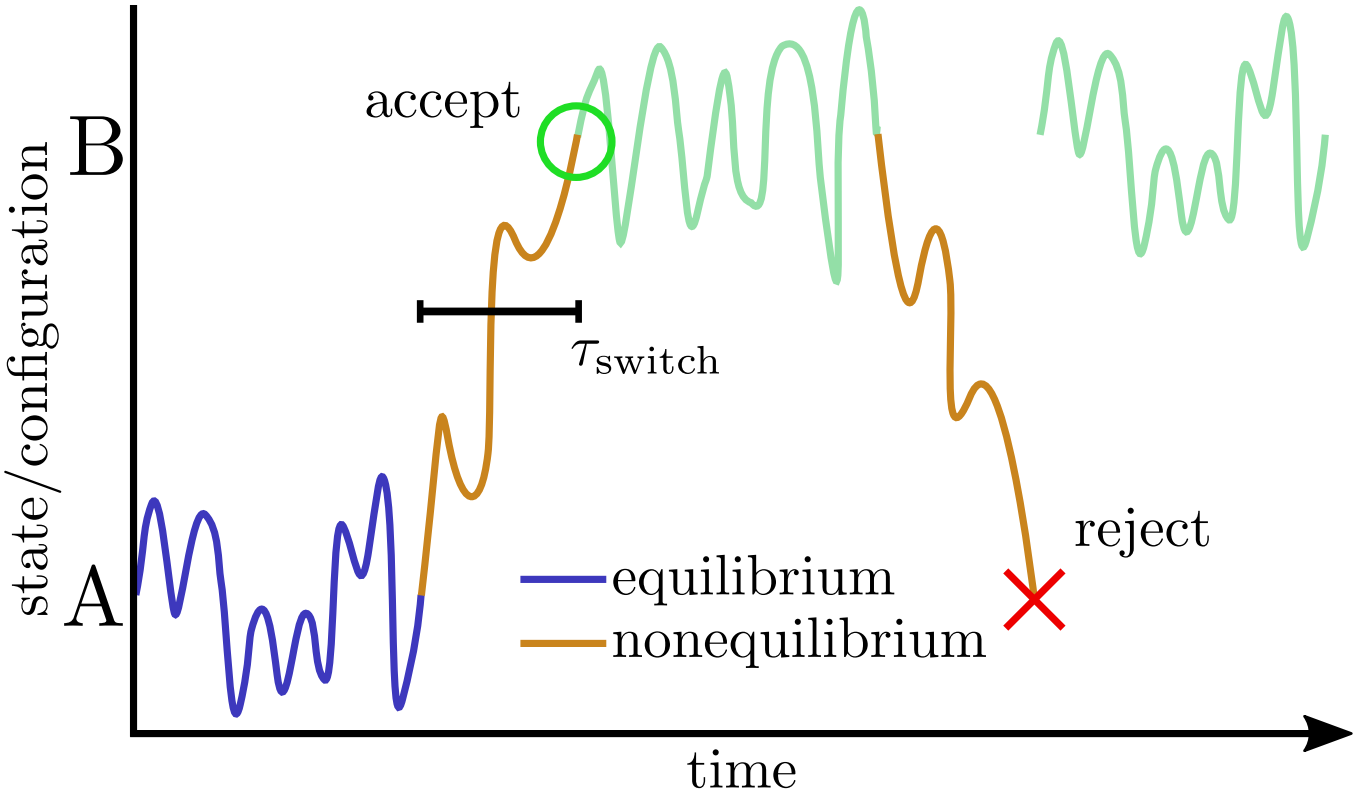
“Fast” alchemical growth



alchemical growth

- Swap the protonation state by using time-dependent interactions.
- Gradually stronger interactions will induce solvent response.
- Clashes are avoided by using the natural dynamics of the model.

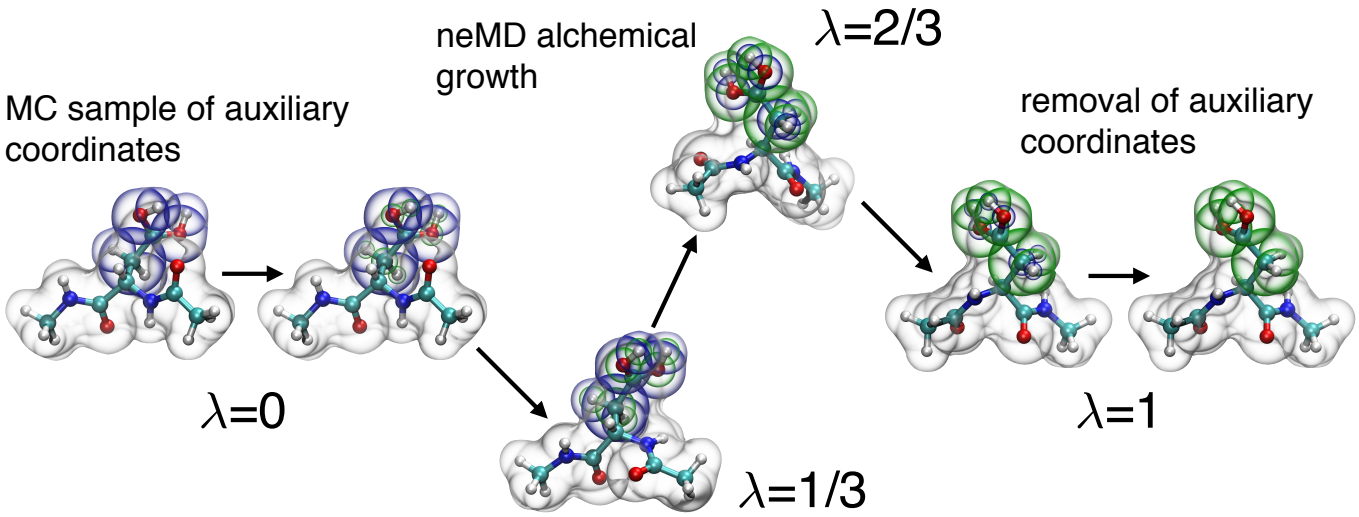
The neMD/MC constant pH paradigm



- Drive alchemical growth with *nonequilibrium* work
- Accept/reject with a generalized Metropolis criterion

Stern *J Chem Phys*, **2007**; Chen & Roux *J Chem Theory Comput*, **2015**

The neMD/MC constant pH paradigm



- Drive alchemical growth with *nonequilibrium* work
- Accept/reject with a generalized Metropolis criterion

Stern *J Chem Phys*, **2007**; Chen & Roux *J Chem Theory Comput*, **2015**

Beyond Gibbs sampling: Hybrid MD and neMD/MC

We now alternate conventional sampling with MD ($\boldsymbol{x}$) and Metropolis Monte Carlo sampling ($\boldsymbol{x}$ *and* $\boldsymbol{\lambda}$):

$$\rho(\boldsymbol{x}, \boldsymbol{\lambda}) T(\boldsymbol{x}, \boldsymbol{\lambda} \to \boldsymbol{x}', \boldsymbol{\lambda}') = \rho(\boldsymbol{x}', \boldsymbol{\lambda}') T(\boldsymbol{x}', \boldsymbol{\lambda}' \to \boldsymbol{x}, \boldsymbol{\lambda})$$

such that the neMD/MC transition probability is:

$$T(\boldsymbol{x}, \boldsymbol{\lambda} \to \boldsymbol{x}', \boldsymbol{\lambda}') = \min \left[1, \frac{\rho(\boldsymbol{x}', \boldsymbol{\lambda}')}{\rho(\boldsymbol{x}, \boldsymbol{\lambda})} \right] = \min \left[1, e^{-\beta W} 10^{-\Delta n \mathrm{pH}} \right]$$

(If you'd like, MD uses the probability $T(\boldsymbol{x} \to \boldsymbol{x}') = 1$.)

Important considerations

- How long should I sample the equilibrium stage?
- How long should I sample the nonequilibrium stage (the "switch time," τ_{switch})
- Rejecting a nonequilibrum trajectory is expensive, how can we avoid doing that so much?

The two-step "inherent" pK_a algorithm

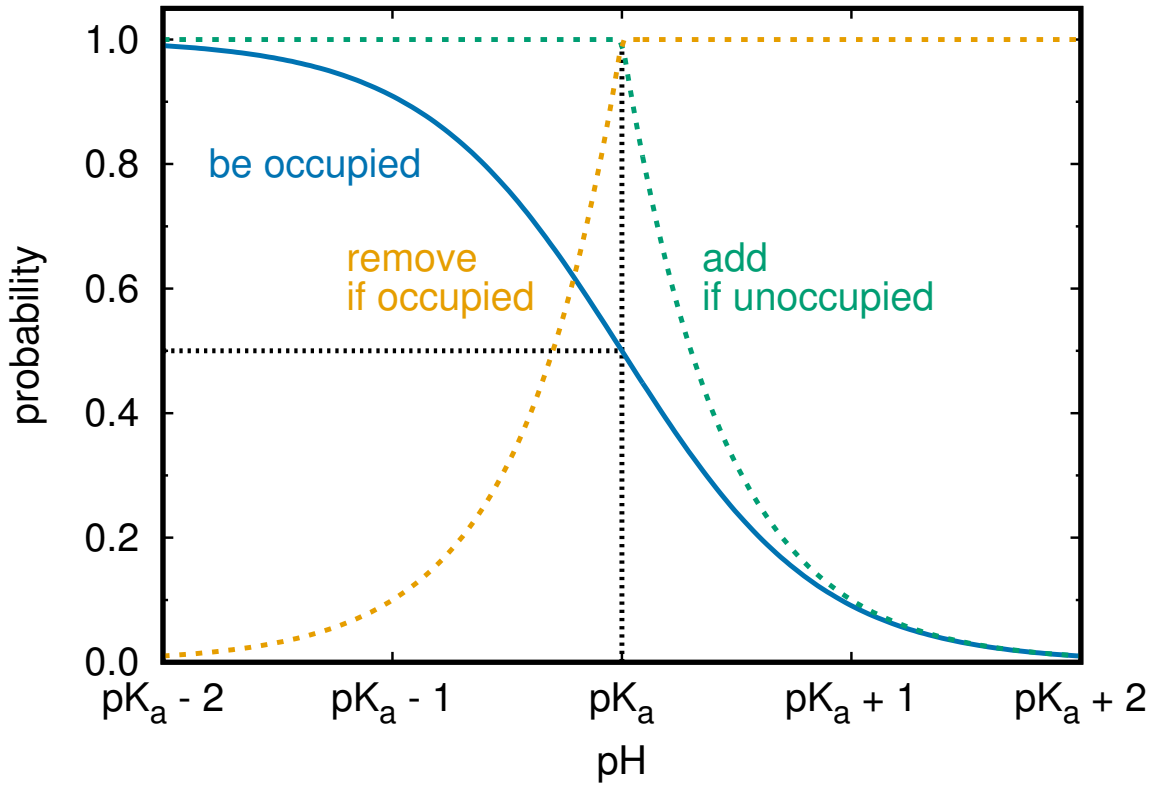
$$T(\boldsymbol{x}, \boldsymbol{\lambda} \to \boldsymbol{x}', \boldsymbol{\lambda}') = T^{(i)}(\boldsymbol{\lambda} \to \boldsymbol{\lambda}') T^{(s)}(\boldsymbol{x} \to \boldsymbol{x}' | \boldsymbol{\lambda} \to \boldsymbol{\lambda}')$$

$$T^{(i)}(\boldsymbol{\lambda} \to \boldsymbol{\lambda}') = \min \left[1, 10^{\mathrm{p}K_a^{(i)}(\boldsymbol{\lambda}, \boldsymbol{\lambda}') - \Delta n \mathrm{pH}} \right]$$

- neMD/MC can be split into *two* parts
  1. $T^{(i)}$ – only depends on $\boldsymbol{\lambda}$ and the pH – CHEAP
  2. $T^{(s)}$ – depends on the switch (W) – COSTLY
- Effort is shifted by estimating a parameter, p$K_a^{(i)}$
- Optimal efficiency achieved for exact pK_a
- Dramatically improved performance on wide pH ranges!
- Can do even better for systems with more than two states.

Chen & Roux *J Chem Theory Comput*, **2015**; Radak, *et al submitted*

Let's look at this graphically

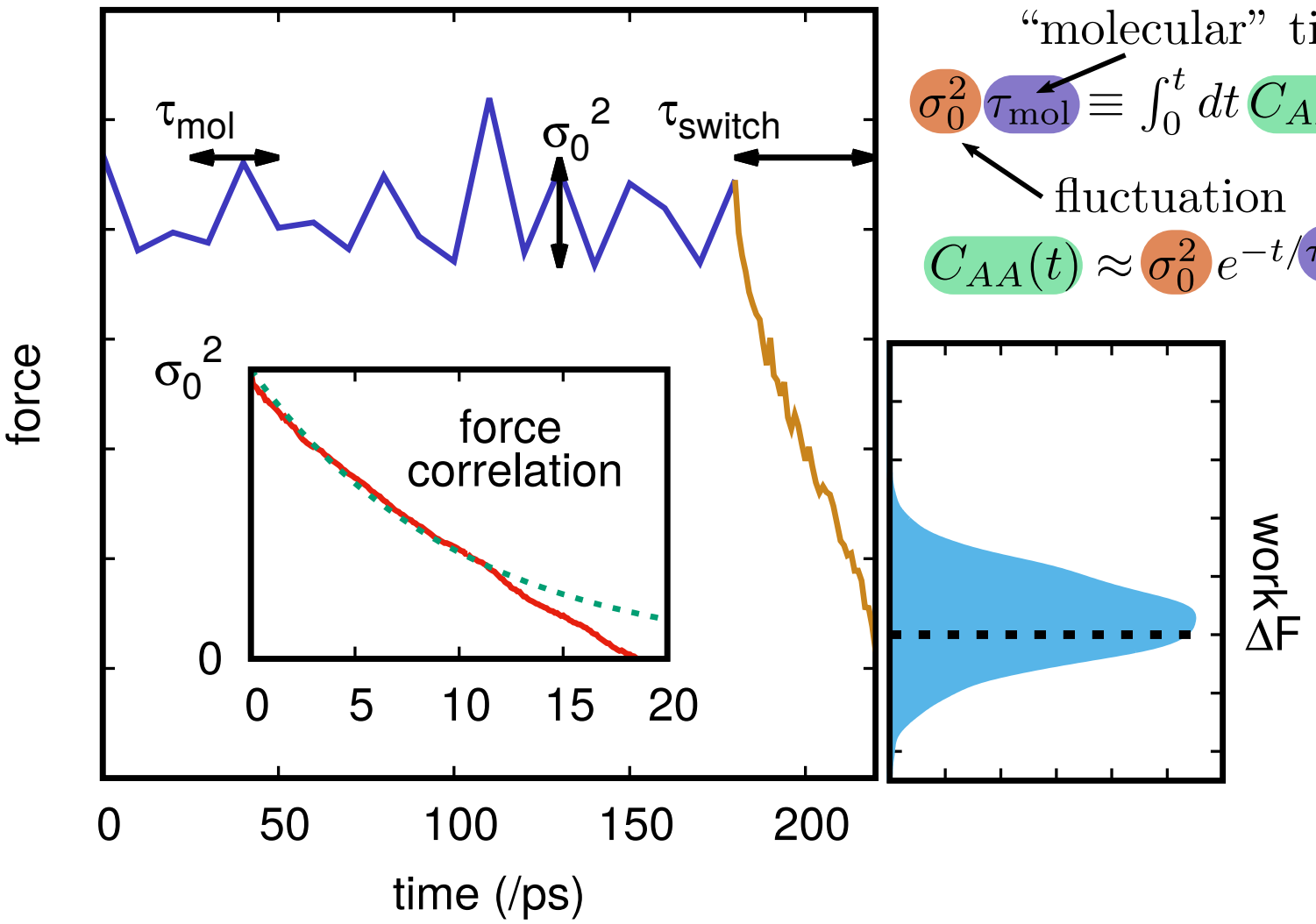


- It's silly to try to add/remove protons to/from acidic/basic residues at high/low pH
- Transitions are proposed in proportion to the estimated population.

What about after we've proposed a switch?

- A short switch will not change much and likely be rejected.
- A long switch is expensive (limit of a single switch – BAD).
- Since the switch success depends on the work, let's analyze that.

Work and force fluctuations – a typical neMD/MC cycle



"molecular" timescale

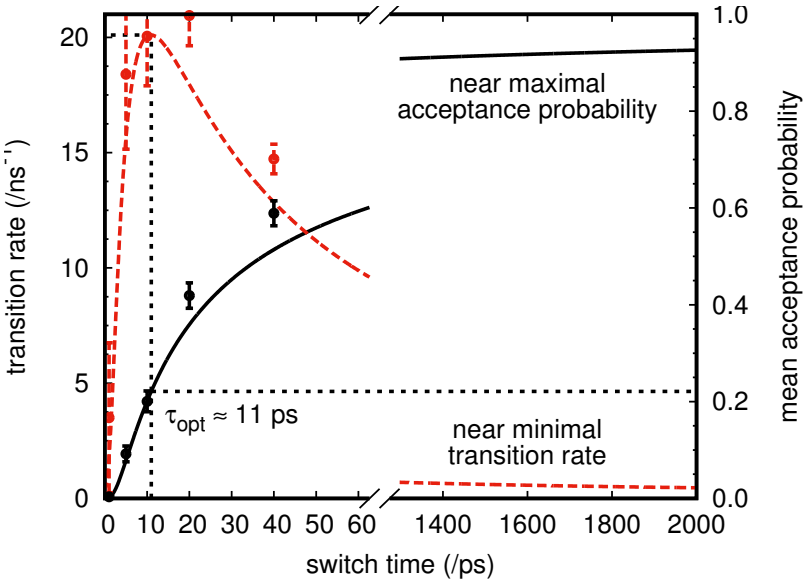
$$\sigma_0^2 \tau_{\mathrm{mol}} \equiv \int_0^t dt\, C_{AA}(t)$$

fluctuation

$$C_{AA}(t) \approx \sigma_0^2 e^{-t/\tau_{\mathrm{mol}}}$$

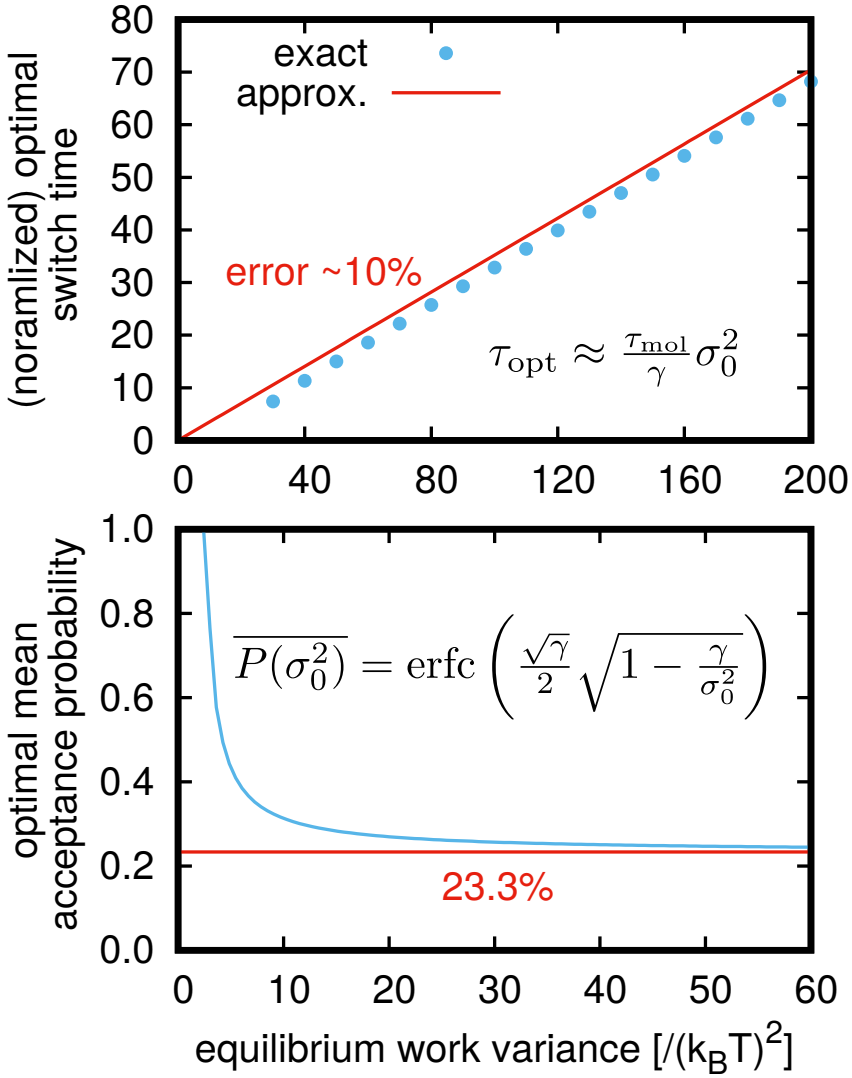
Radak & Roux *J Chem Phys*, **2016**

Theoretical and Empirical Performance Analysis



- Well-defined criteria for optimization.
- Cost is quite tractable.

Radak & Roux *J Chem Phys*, **2017**

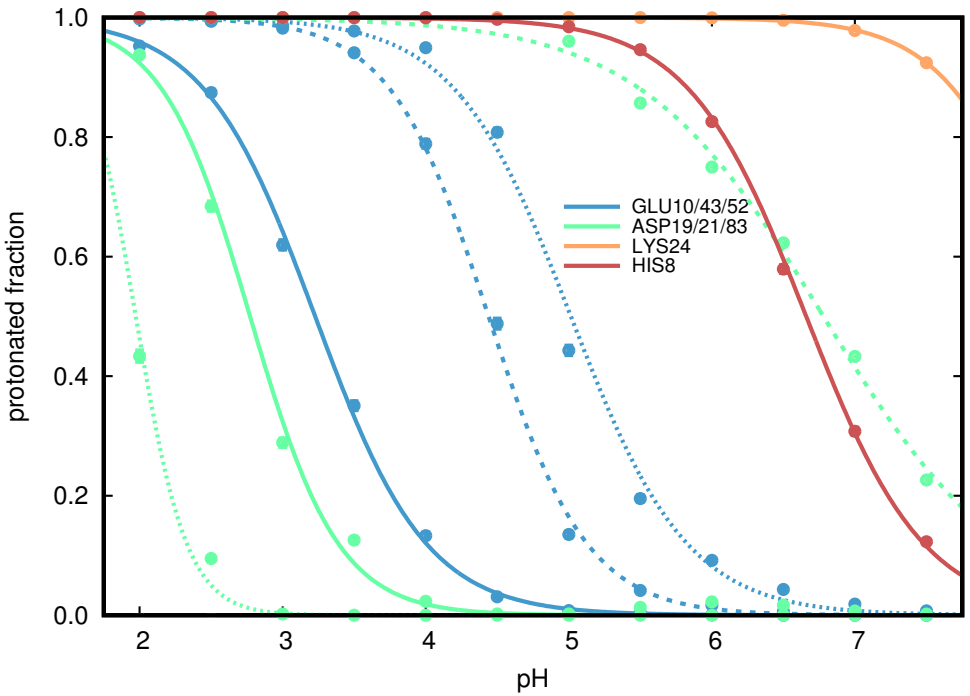


NAMD Constant pH features

- Flexible Tcl interface `source ...lib/namdcph/namdcph.tcl`
- PSF build procedure is unchanged (automated `psfgen`)
- Implemented with PME and full electrostatics
- Normal CPU scaling (no GPU yet) - depends on alchemy
- Companion analysis script `cphanalyze`

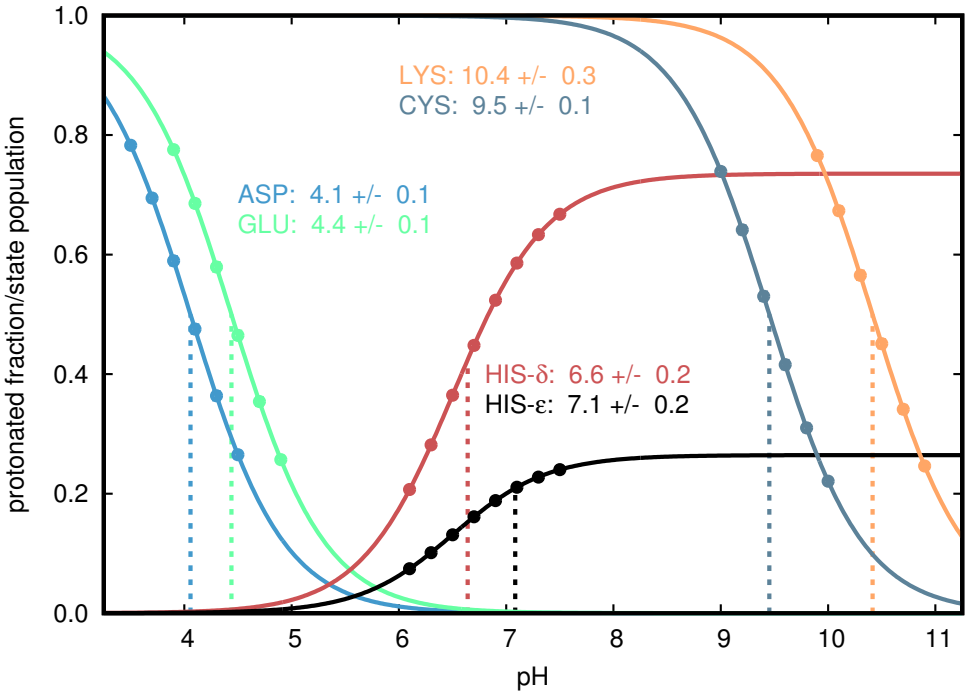
```
pH 7.0
cphNumstepsPerSwitch 7500 ;# run 7500 steps per switch
cphRun 5000 10 ;# run 10 cycles of 5000 MD steps
```

Output and Analysis



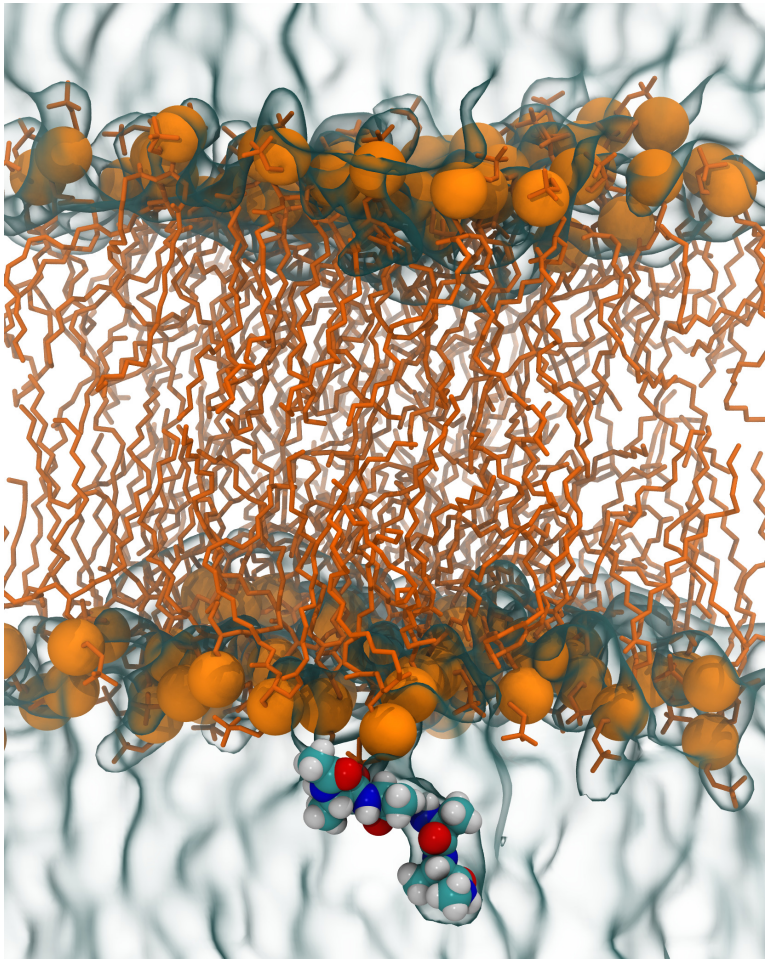
- Normal usage requires multiple pH values ("titration curves")
- New output `cphlog` and `cphrst`
- New checkpoint files `psf/pdb`
- Can boost performance with WHAM (`cphanalyze`)
- Can also analyze residue correlations

Reference amino acids are well-reproduced

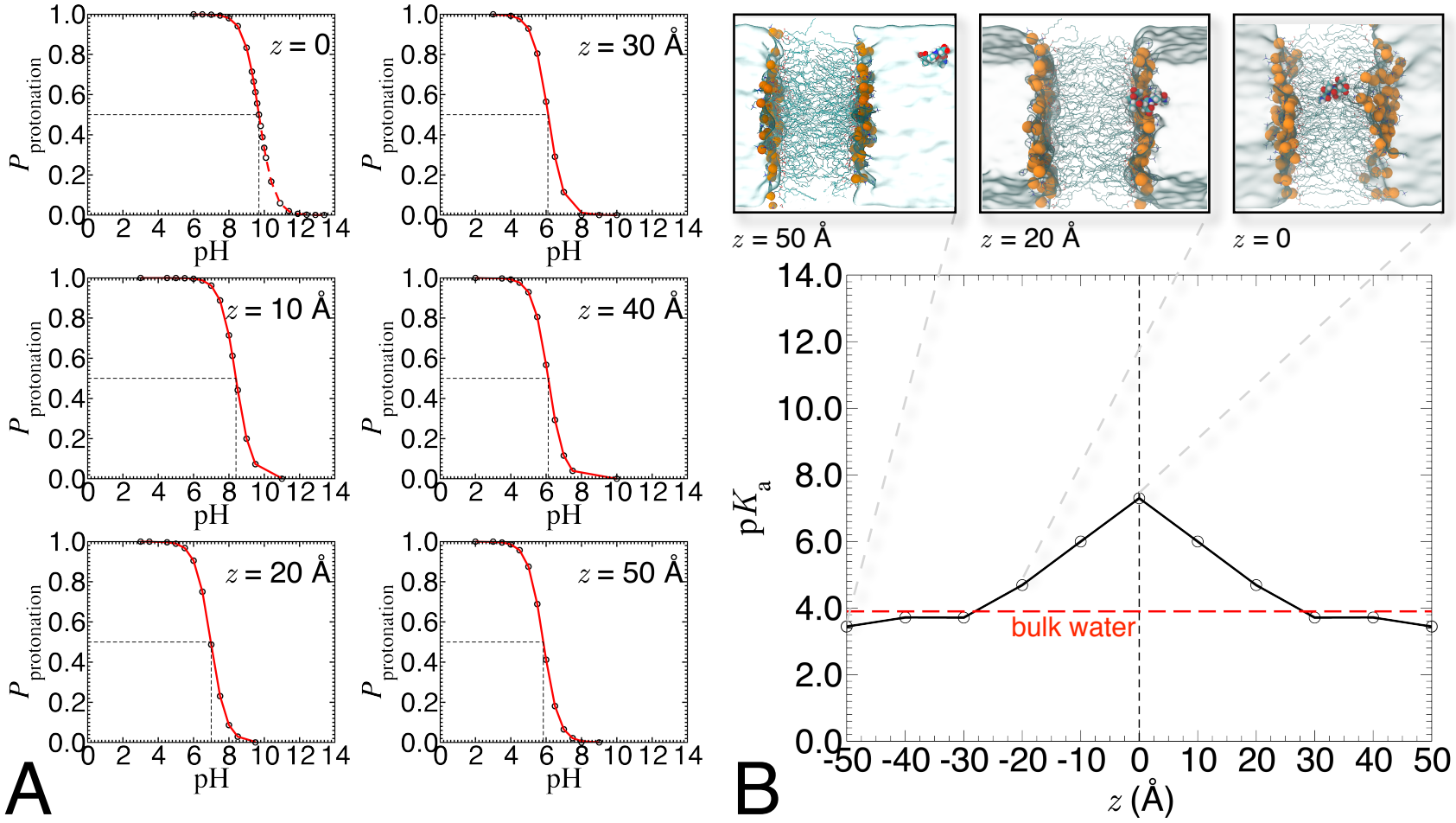


- Adjustments to force field enforce empirical reference values
- Implicitly model solvated proton and bond energy effects
- Bonus - accurate reproduction of tautomeric ratios!

What about challenging environments?



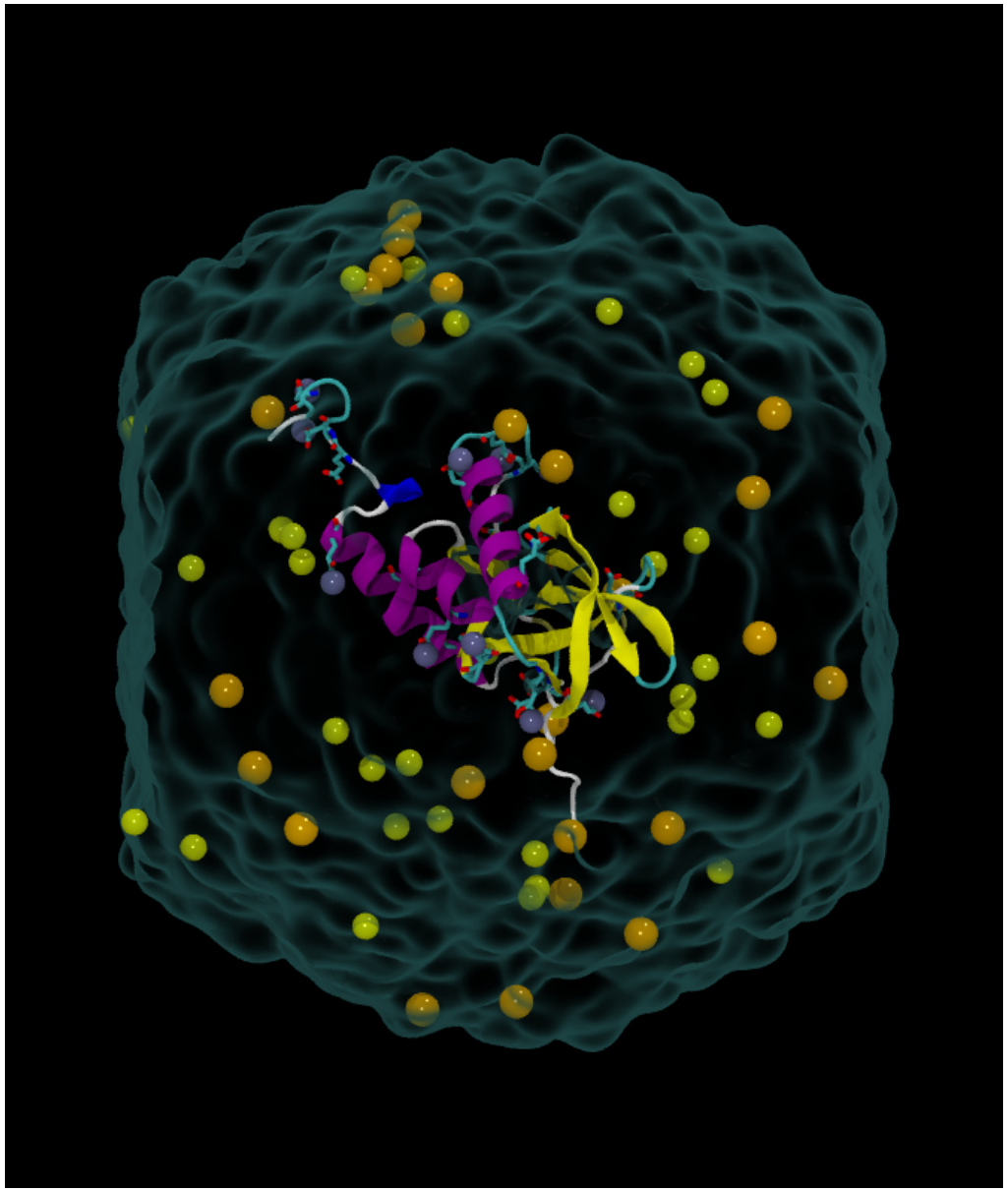
- Single titratable peptide (AADAA)
- Lipid relaxation is *slow* (slower than water)
- Low dielectric region should perturb pKa in obvious way



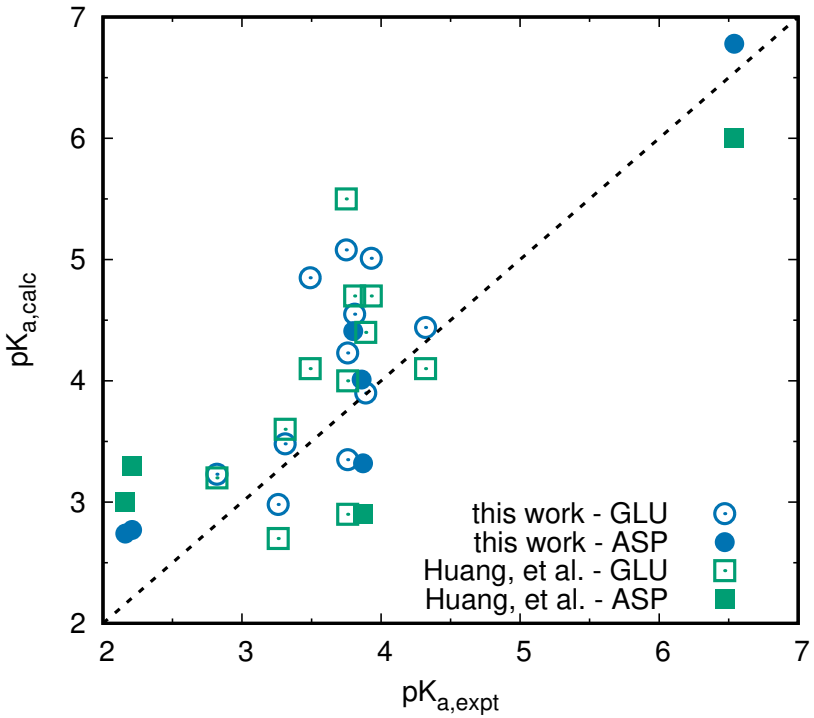
- Significant shifts due to low dielectric region.
- Switch time of 10 ps is sufficient.

Radak, *et al submitted* Teixeira, *et al J Chem Theory Comput*, **2016**

Staph nuclease (SNase) - A constant pH benchmark



Benchmarking of SNase pK_a values



- Good correlation with measured values for carboxylates
- Bonus - estimates for HIS and LYS

residue		this work
LYS	24	8.43 (0.45)
HIS	8	6.66 (0.56)
	121	5.36 (0.50)

Concluding Remarks/Future Directions

- Things we are working on:
  - Performance improvements in alchemy – CUDA support
  - Titratable lipids and phosphates
- Things we would like to work on:
  - `psfgen` improvements – support for Drude
  - Better visualization support in VMD
  - More powerful interface for analysis (PyNAMD)

Acknowledgements (other people to blame)

Univ of Chicago

- Benoît Roux
- Donghyuk Suh
- Huan Rui
- Yunjie Chen

ALCF

- Sunhwan Jo (SilcsBio)
- Wei Jiang

UIUC

- Jim Phillips
- Chris Chipot
- Abhi Singharoy

Argonne NATIONAL LABORATORY

Theta Early Science Program

npsdc MEMBRANE PROTEIN STRUCTURAL DYNAMICS GATEWAY