TOWARD REALIZING THE DREAM OF FREE-ENERGY BASED SMALL MOLECULE DESIGN

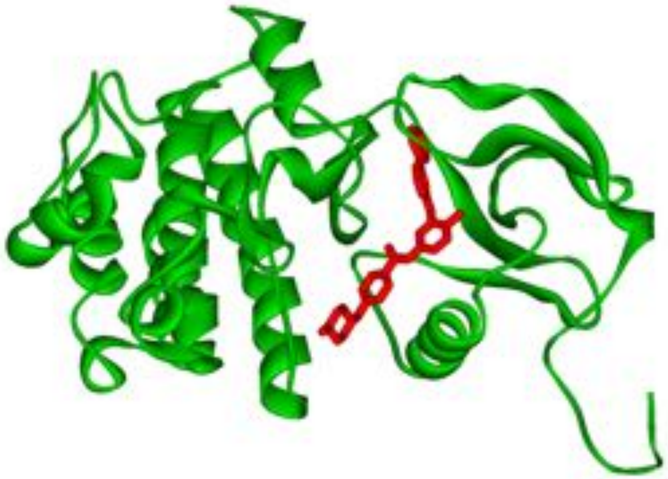
John D. Chodera
MSKCC Computational Biology Program
http://www.choderalab.org

Slides available at http://www.choderalab.org

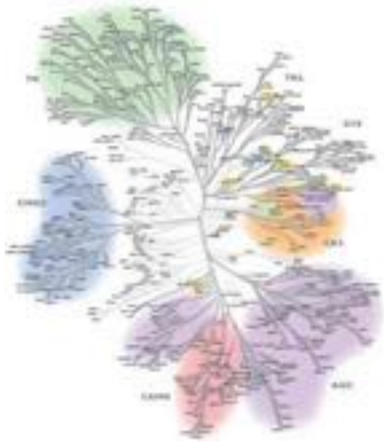
DISCLOSURES:
- Scientific Advisory Board, Schrödinger

SOMETIMES, DRUG DISCOVERY WORKS WELL

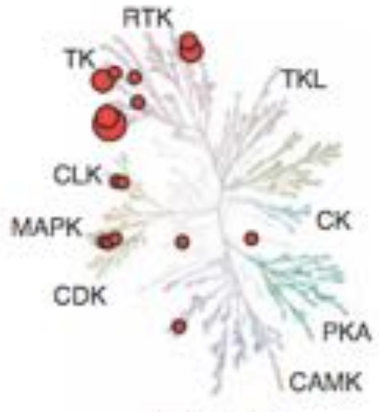
Bcr-Abl fusion constitutively activates ABL in CML patients,
resulting in unchecked white blood cell proliferaton



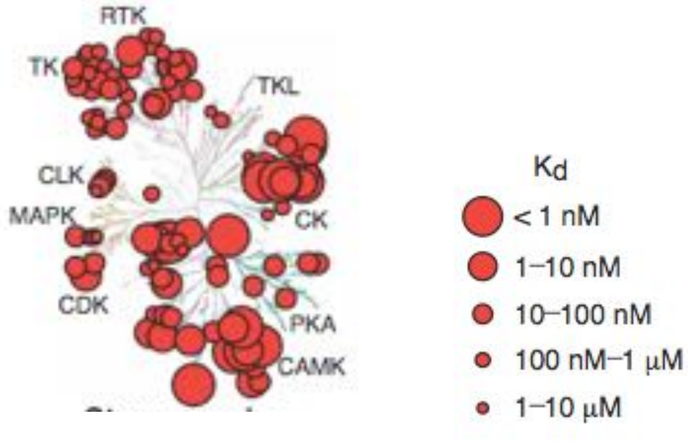
imatinib bound to c-Abl [PDB:1IEP]



human kinome
[518 kinases]



imatinib
[blockbuster drug]



staurosporine
[toxic natural product]

Nature Biotech 23:329, 2005

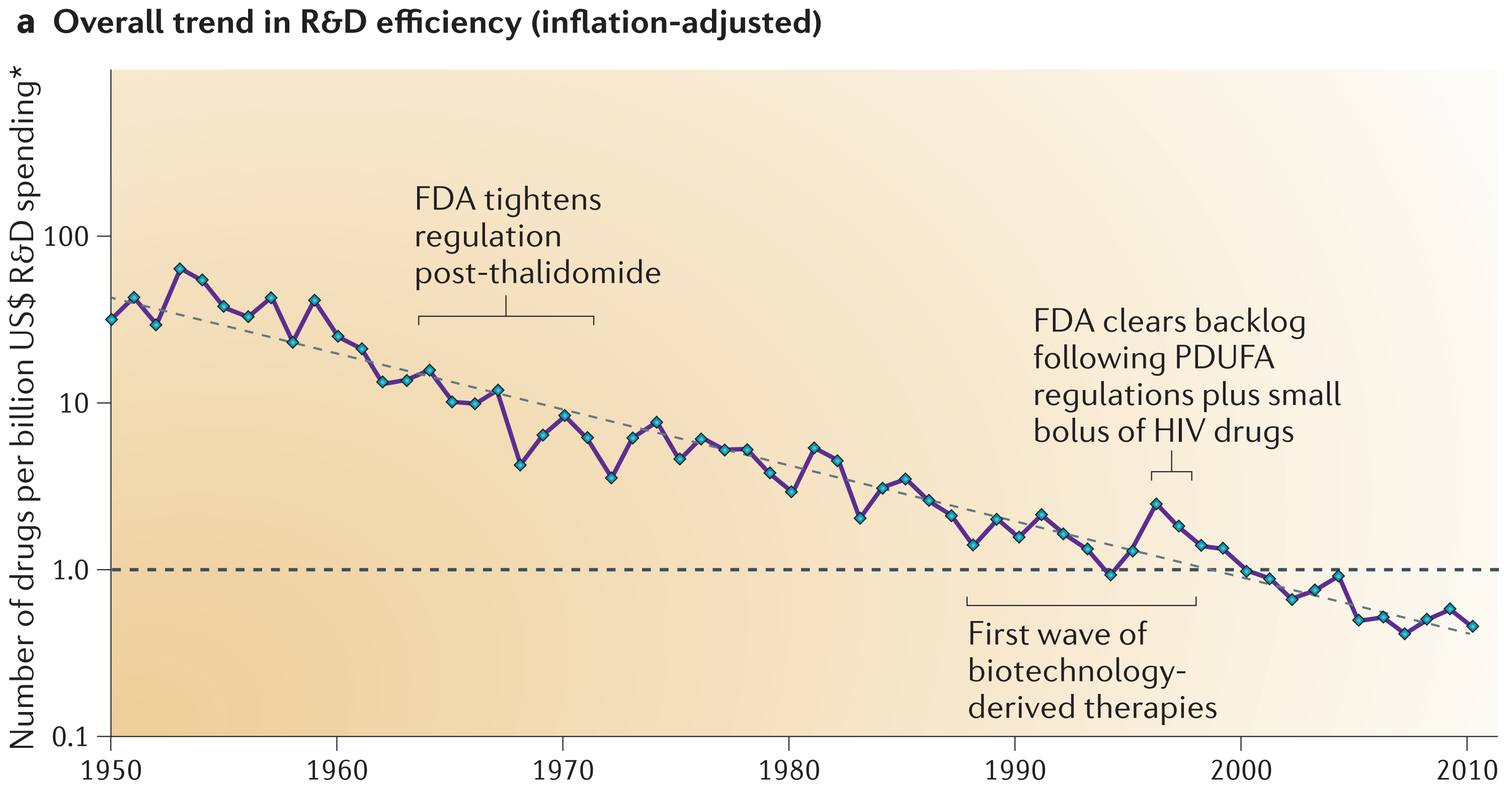
DRUG DISCOVERY USUALLY ENDS IN FAILURE

Total pharma R&D spending **doubled** to $65B over 2000-2010

FDA approvals of new molecular entities **went down by half**

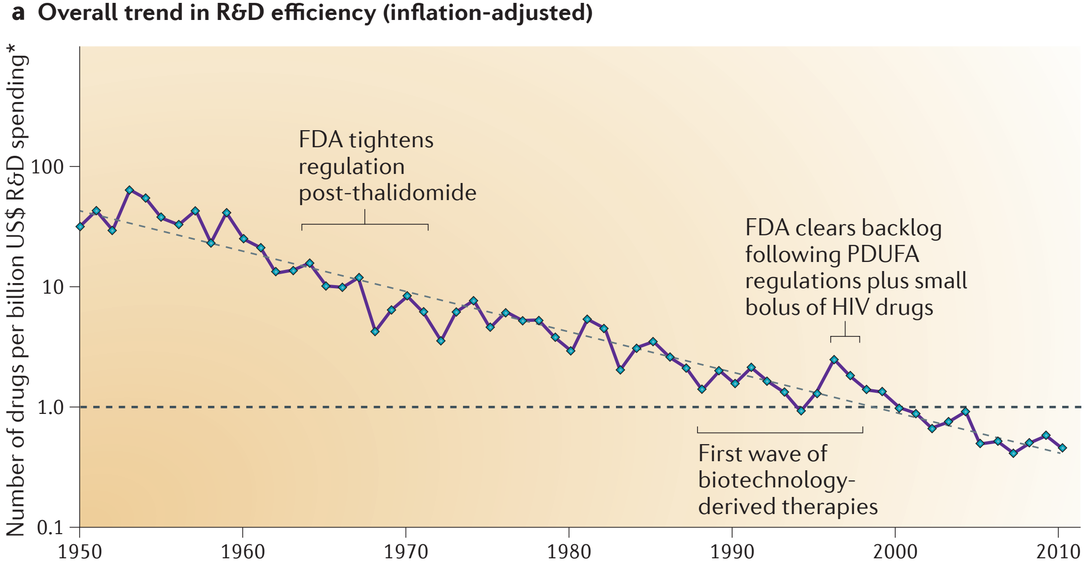
Number of truly innovative new molecules **remained constant at 5-6/year**

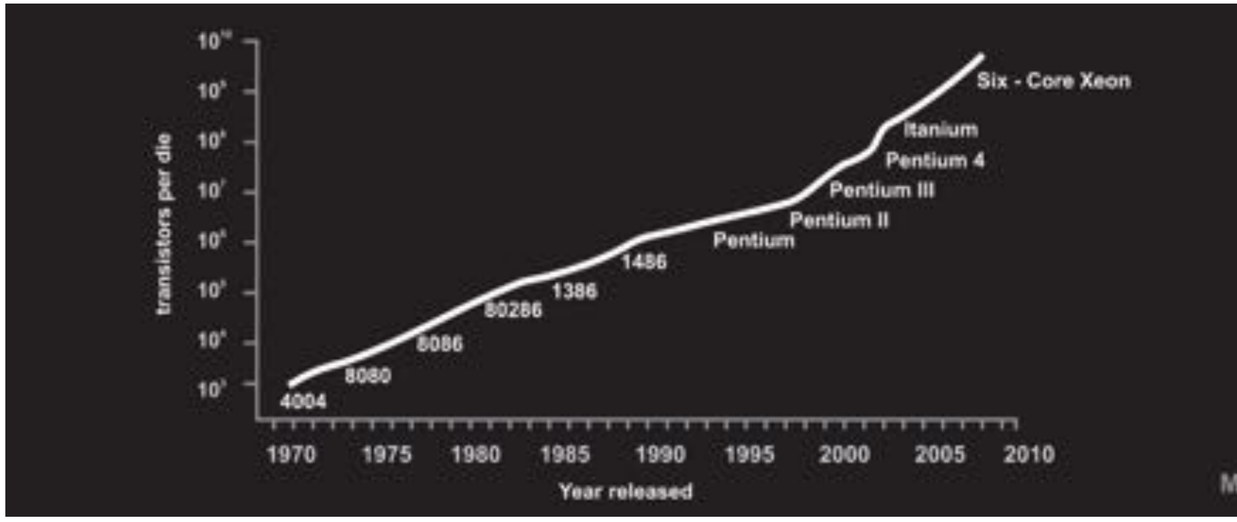
2010-2015 has seen large reductions in pharma R&D in the US



Scannell et al. Nature Rev Drug Disc 11:191, 2012

DRUG DISCOVERY USUALLY ENDS IN FAILURE

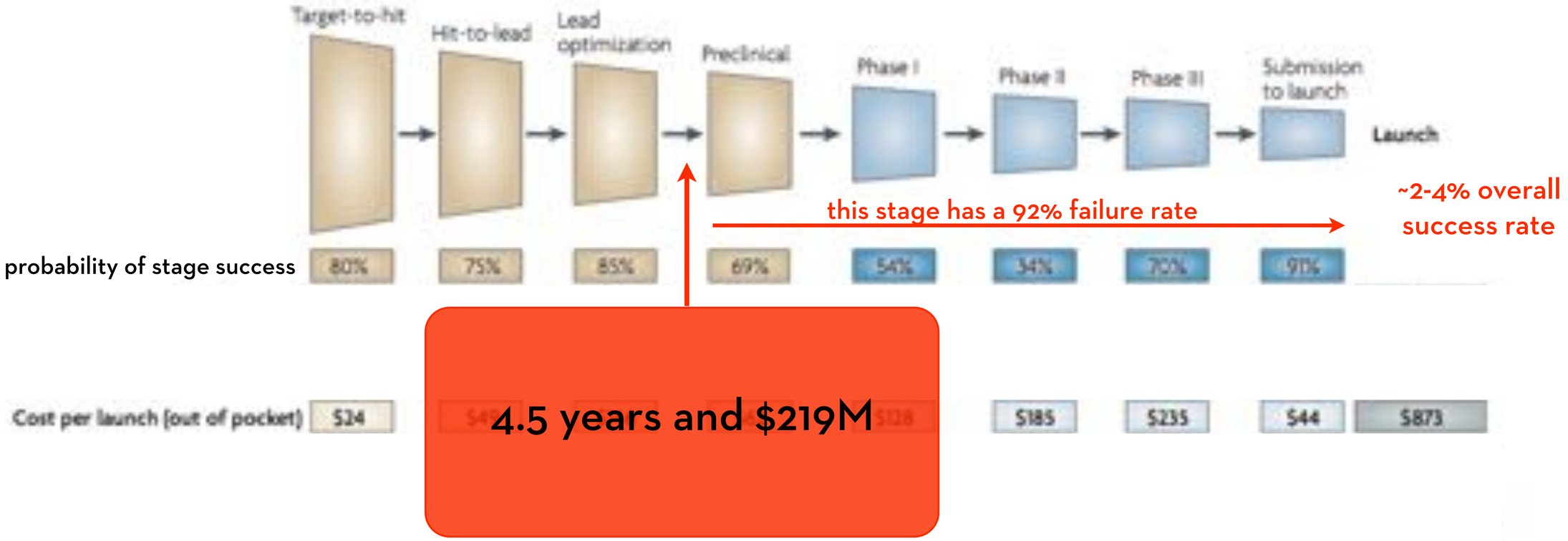




EROOM'S LAW

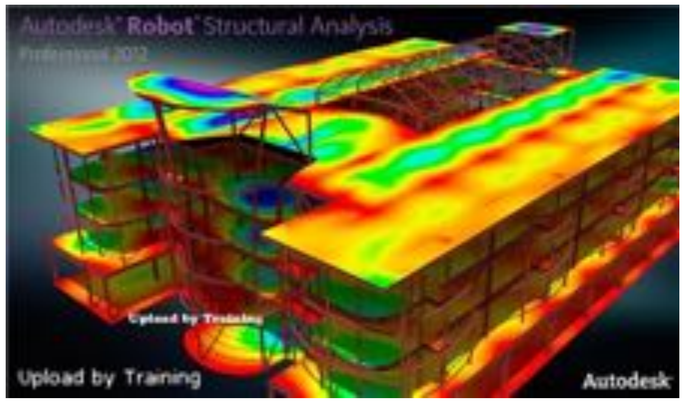
MOORE'S LAW

DRUG DISCOVERY USUALLY ENDS IN FAILURE



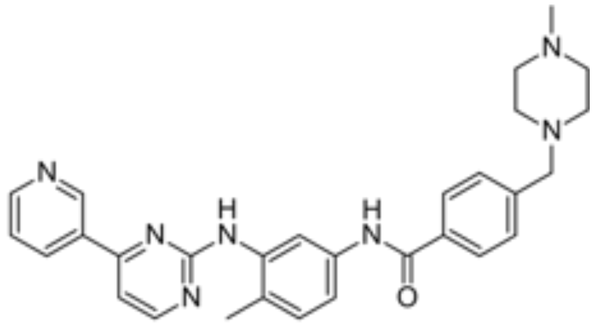
Paul et al. Nat. Rev. Drug Discover. 9:203, 2010.
Chodera et al. Curr. Opin. Struct. Biol., 21:150, 2011.

WE REGULARLY DESIGN PLANES, BRIDGES, AND BUILDINGS ON COMPUTERS



10^3 - 10^6 parts

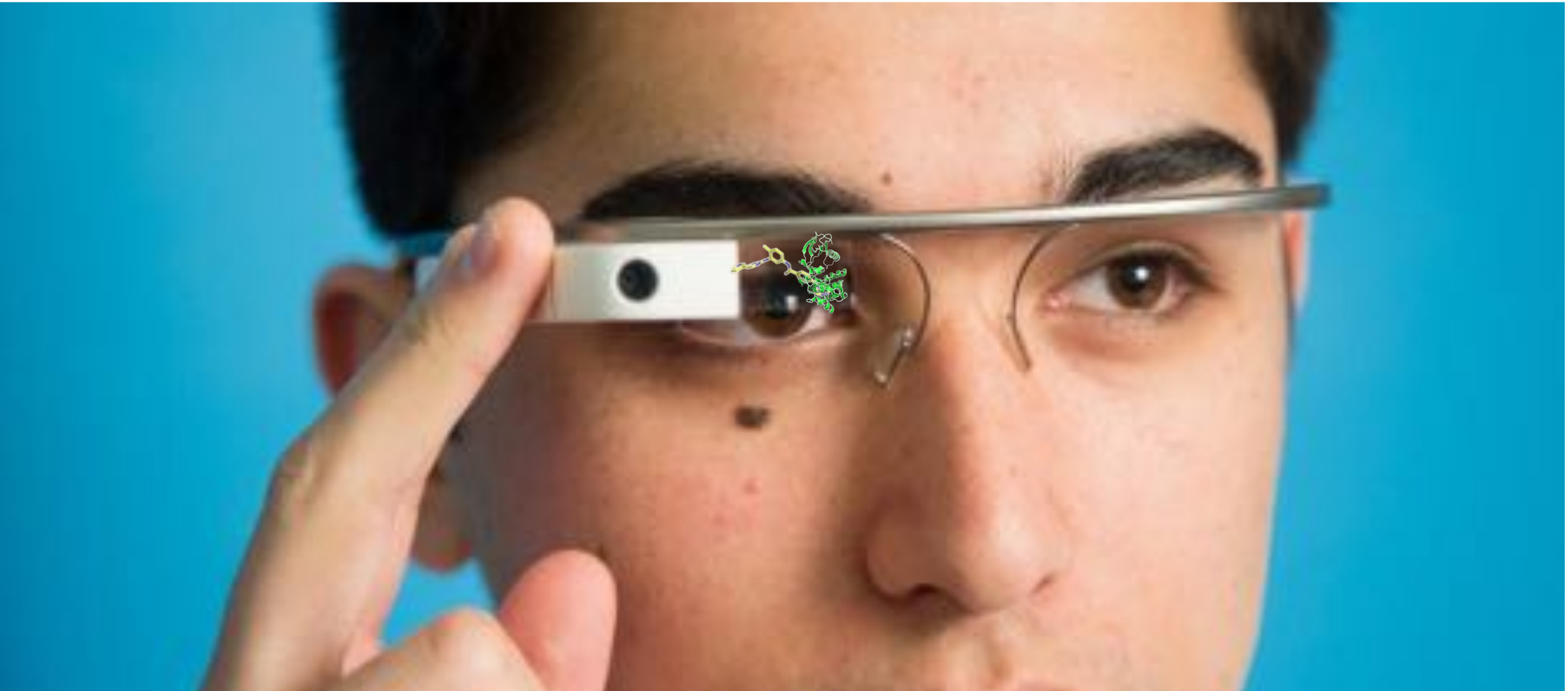
WHY NOT SMALL MOLECULE DRUGS?



< 10^2 atoms

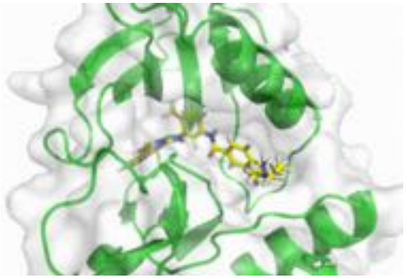
HOW CAN WE BRING DRUG DESIGN INTO THE 21ST CENTURY?

HOW CAN WE BRING DRUG DESIGN INTO THE 21ST CENTURY?

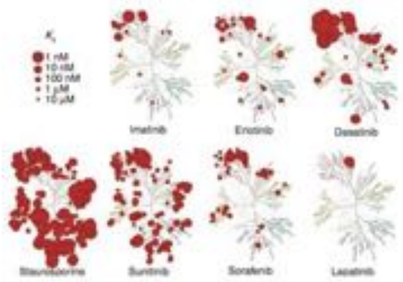


CHODERA LAB

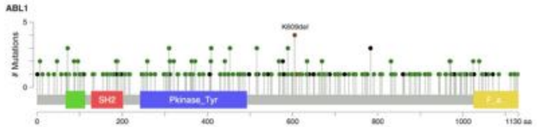
HOW CAN STATISTICAL MECHANICS PLAY A ROLE IN THE ERA OF GENOMICS AND BIOMEDICAL BIG DATA?



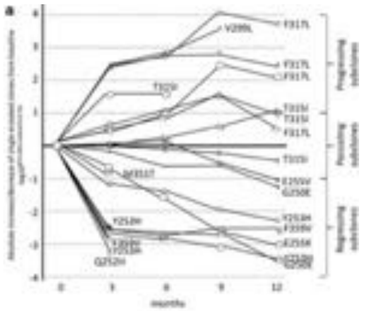
SELECTIVE INHIBITOR DESIGN: TARGETS/ANTITARGETS



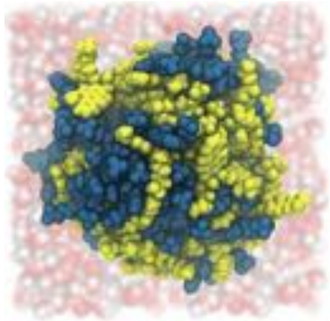
KINASE INHIBITOR SELECTIVITY



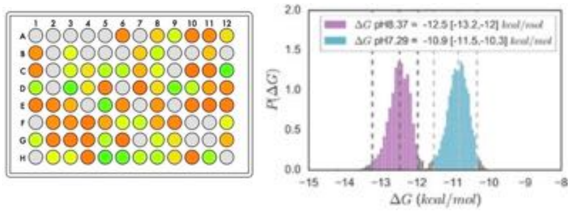
PREDICTING DRUG SENSITIVITY/RESISTANCE



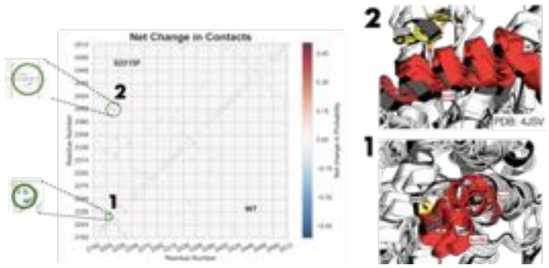
ANTICIPATING DRUG RESISTANCE



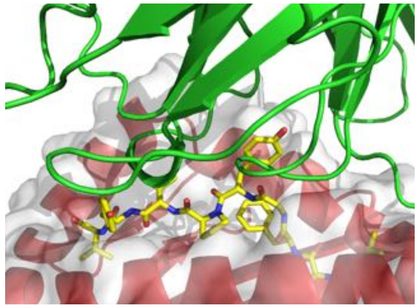
NOVEL DRUG DELIVERY MODALITIES



AUTOMATED BIOPHYSICAL ASSAYS AND INFERENCE



MECHANISMS OF ONCOGENIC ACTIVATION

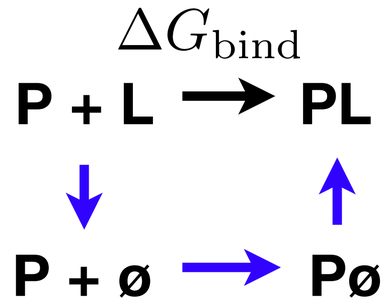


CANCER IMMUNOTHERAPY

COMPUTATION OF BINDING AFFINITIES TRANSFORMED INTO COMPUTATION OF RATIOS OF PARTITION FUNCTIONS

$$P + L \leftrightharpoons PL$$

$$K_d = \frac{[P][L]}{[PL]} \; ; \; K_a = \frac{[PL]}{[P][L]}$$



$$\Delta G^{o}_{PL} = -k_B T \ln \left[\frac{C^o}{8\pi^2} \frac{Z_{PL}}{Z_{P\varnothing}} \frac{Z_{\varnothing}}{Z_L} \right]$$

(strong binding limit)

standard state

$C_0 = 1\ M$

ligand in solvent

$$\frac{Z_L}{Z_{\varnothing}} = \frac{\int_\Gamma dx\, e^{-\beta H_L(x;\lambda=1)}}{\int_\Gamma dx\, e^{-\beta H_L(x;\lambda=0)}}$$

ligand in complex

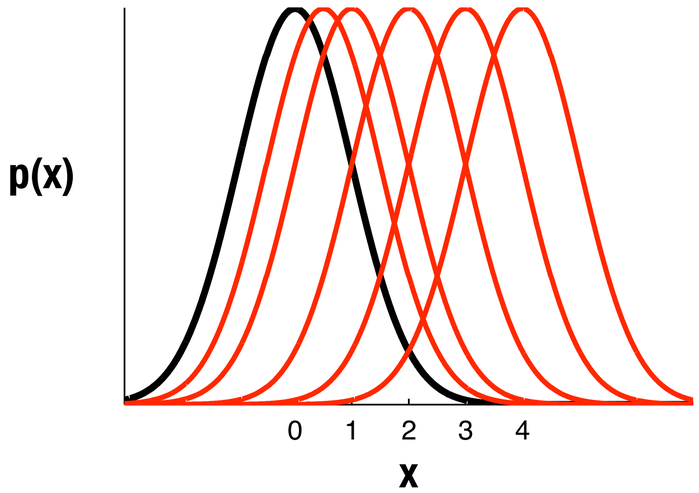
$$\frac{Z_{PL}}{Z_{P\varnothing}} = \frac{\int_\Gamma dx\, e^{-\beta H_{PL}(x;\lambda=1)}}{\int_\Gamma dx\, e^{-\beta H_{PL}(x;\lambda=0)}}$$

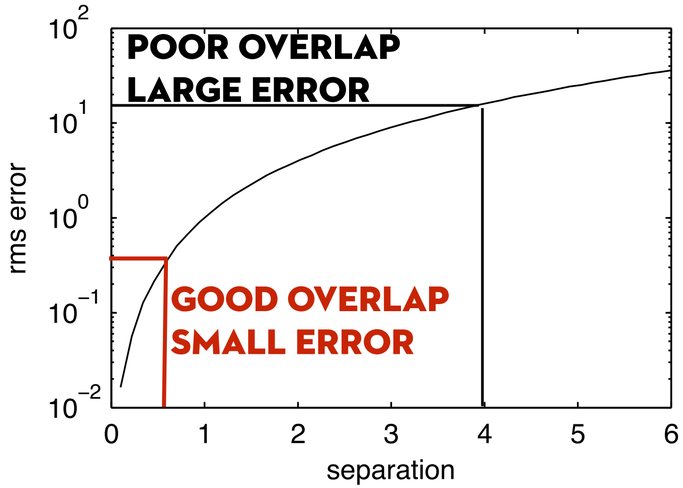
ISOMORPHIC TO COMPUTING RATIOS OF NORMALIZING CONSTANTS IN STATISTICAL INFERENCE OR RATIOS OF MODEL EVIDENCES IN INFERENTIAL MACHINE LEARNING

(A FIELD THAT IS BASICALLY PRINTING MONEY RIGHT NOW)

THE ADDITION OF ALCHEMICAL INTERMEDIATES GREATLY IMPROVES ESTIMATION EFFICIENCY

COMPUTING FREE ENERGY DIFFERENCES BECOMES HARDER AS PHASE SPACE OVERLAP DECREASES; ERROR INCREASES RAPIDLY WITH DIMINISHING PHASE SPACE OVERLAP



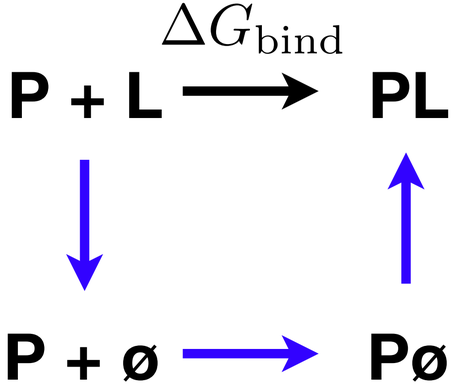


INTRODUCING ALCHEMICAL INTERMEDIATES TO ENSURE RATIOS CAN BE EFFICIENTLY ESTIMATED

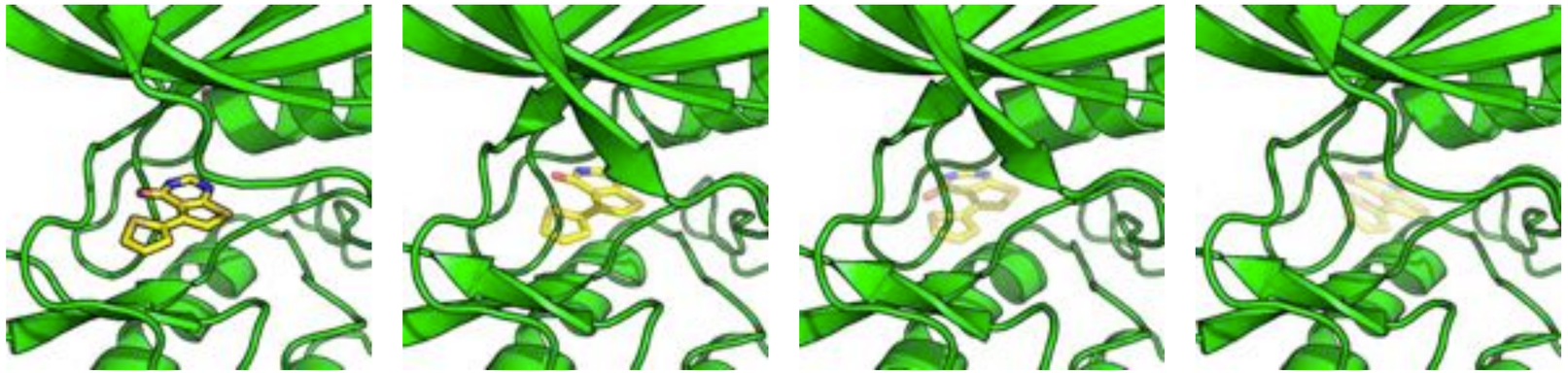
$$\Delta F_{1\to N} = -\beta^{-1}\ln\frac{Z_N}{Z_1} = -\beta^{-1}\ln\frac{Z_2}{Z_1}\cdot\frac{Z_3}{Z_2}\cdots\frac{Z_N}{Z_{N-1}} = \sum_{n=1}^{N-1}\Delta F_{n\to n+1} \qquad Z_n = \int d\mathbf{x}\, e^{-\beta U(\mathbf{x})}$$

D. Wu and D. A. Kofke. "Phase-space overlap measures. I. Fail-safe bias detection...", *J. Chem. Phys.* **123**: 054103 (2005).

ALCHEMICAL FREE ENERGY CALCULATIONS PROVIDE A RIGOROUS WAY TO EFFICIENTLY COMPUTE BINDING AFFINITIES



RUN 10-40 EQUILIBRIUM SIMULATIONS OF ALCHEMICAL INTERMEDIATES



RESTRAINT IMPOSITION | DISCHARGING | STERIC DECOUPLING | NONINTERACTING

SAMPLE $x_{kn} \sim \pi(x; \lambda_k)$ **FROM** $\pi(x; \lambda) = Z_k^{-1} e^{-\beta H(x;\lambda)}$

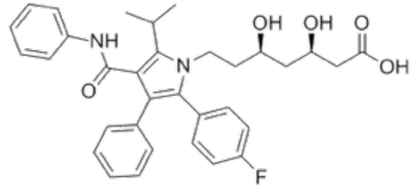
REDUCES EFFORT BY ORDERS OF MAGNITUDE OVER SIMULATING DIRECT ASSOCIATION PROCESS

$$\Delta F_{1\to N} = -\beta^{-1} \ln \frac{Z_N}{Z_1} = -\beta^{-1} \ln \frac{Z_2}{Z_1} \cdot \frac{Z_3}{Z_2} \cdots \frac{Z_N}{Z_{N-1}} = \sum_{n=1}^{N-1} \Delta F_{n\to n+1} \qquad Z_n = \int d\mathbf{x}\, e^{-\beta U(\mathbf{x})}$$

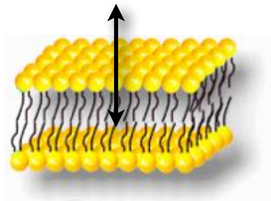
Pioneering work from many: McCammon, van Gunsteren, Kollman, Jorgensen, Chipot, Roux, Boresch, Fujitani, Pande, Shirts, Swope, Christ, Mobley, and many more
Recent review: Chodera, Mobley, Shirts, Dixon, Branson, Pande. Curr Opin Struct Biol 21:150, 2011.

THE SAME ALCHEMICAL METHODS CAN ALSO COMPUTE MANY OTHER USEFUL PROPERTIES

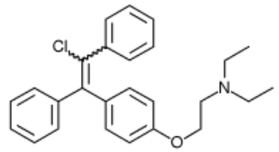
PARTITION COEFFICIENTS (LOGP, LOGD) AND PERMEABILITIES



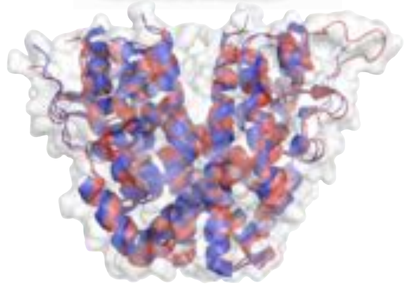
lipitor



SELECTIVITY FOR SUBTYPES OR RELATED TARGETS/OFF-TARGETS



clomifene



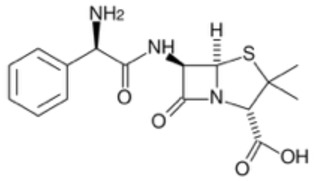
ERα/β

LEAD OPTIMIZATION OF AFFINITY AND SELECTIVITY

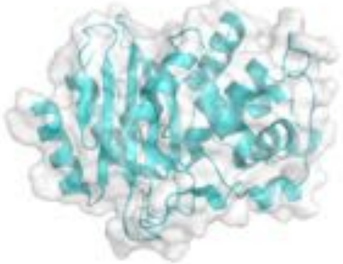
Imatinib

Dasatinib

SUSCEPTIBILITY TO RESISTANCE MUTATIONS



ampicillin



β-lactamase

ALSO SOLUBILITIES, POLYMORPHS, ETC.

WHAT DO WE NEED TO DO TO ENABLE DESIGN BY FREE ENERGY METHODS?

1. Speed up calculations without sacrificing fidelity

2. Include relevant chemical effects

3. Use a sufficiently accurate forcefield

WHAT TIMESTEP* SHOULD I USE?

- WHICH INTEGRATOR?
- HYDROGEN MASS REPARTITIONING?
- MULTIPLE TIMESTEP INTEGRATION?
- WHICH FORCE SPLITTING?
- SOLUTE/SOLVENT SPLITTING?

TO INTEGRATE LANGEVIN DYNAMICS ON A COMPUTER, WE HAVE TO DISCRETIZE EQUATIONS OF MOTION

continuous Langevin dynamics

$$dr = v\ dt$$

$$dv = \frac{f(t)}{m}\,dt - \gamma v\ dt + \sqrt{\frac{2\gamma}{\beta m}}\,d\mathsf{W}(t)$$

discrete timestep Langevin integrator

$$v'_t = v^*_t + \frac{\Delta t}{2m}\left(F_t(r^*_t) - \gamma m v^*_t + \sqrt{\frac{2\gamma m}{\Delta t}}\xi_t\right)$$

$$r_t = r^*_t + \Delta t\, v'_t$$

$$v_t = \frac{1}{1+\frac{\gamma\Delta t}{2}}\left[v'_t + \frac{\Delta t}{2m}\left(F_t(r_t) + \sqrt{\frac{2\gamma m}{\Delta t}}\xi'_t\right)\right]$$

NOT ALL INTEGRATORS ARE EQUAL

Continuous Langevin dynamics equations of motion

$$d\begin{bmatrix}\mathbf{x}\\ \mathbf{v}\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{v}\\ 0\end{bmatrix}}_{\mathsf{R}}\mathrm{d}t + \underbrace{\begin{bmatrix}0\\ -M^{-1}\nabla U(\mathbf{x})\end{bmatrix}}_{\mathsf{V}}\mathrm{d}t + \underbrace{\begin{bmatrix}0\\ -\gamma\mathbf{v}\mathrm{d}t + \sigma M^{1/2}dW\end{bmatrix}}_{\mathsf{O}}$$

Some of these parts can be integrated exactly:

$$(e^{t\mathcal{L}_{\mathsf{R}}}\phi)(\mathbf{q},\mathbf{p}) = \phi(\mathbf{q}+t\mathbf{M}^{-1}\mathbf{p},\mathbf{p})$$
$$(e^{t\mathcal{L}_{\mathsf{V}}}\phi)(\mathbf{q},\mathbf{p}) = \phi(\mathbf{q},\mathbf{p}-t\nabla\mathrm{U}(\mathbf{q}))$$
$$(e^{t\mathcal{L}_{\mathsf{O}}}\phi)(\mathbf{q},\mathbf{p}) = \int_{\mathcal{P}}\phi(\mathbf{q},e^{-\gamma t}\mathbf{p}+\eta\mathbf{M}^{1/2}\mathbf{x})\frac{e^{-|\mathbf{x}|^2/2}}{(2\pi)^{N/2}}d\mathbf{x}$$

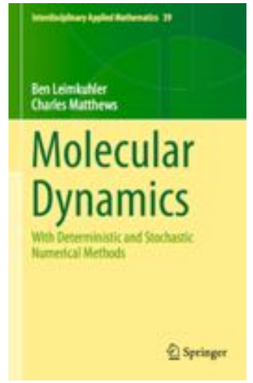
$$\mathsf{R}:(\mathbf{x},\mathbf{v};h) \mapsto (\mathbf{x}+\mathbf{v}h,\mathbf{v})$$
$$\mathsf{V}:(\mathbf{x},\mathbf{v};h) \mapsto (\mathbf{x},\mathbf{v}+(f(\mathbf{x})/m)h)$$
$$\mathsf{O}:(\mathbf{x},\mathbf{v};h) \mapsto (\mathbf{x},a_h\mathbf{v}+b_h\sqrt{k_BT/m}\xi)$$

We can approximate the Langevin propagator by various splittings:

$$e^{\Delta t[\mathcal{L}_o+\mathcal{L}_v+\mathcal{L}_r+\mathcal{L}_h]} \simeq$$
$$e^{\frac{\Delta t}{2}\mathcal{L}_r}e^{\frac{\Delta t}{2}\mathcal{L}_h}e^{\frac{\Delta t}{2}\mathcal{L}_v}e^{\Delta t\,\mathcal{L}_o}e^{\frac{\Delta t}{2}\mathcal{L}_v}e^{\frac{\Delta t}{2}\mathcal{L}_h}e^{\frac{\Delta t}{2}\mathcal{L}_r}$$ **RVOVR**

$$e^{\Delta t[\mathcal{L}_o+\mathcal{L}_v+\mathcal{L}_r+\mathcal{L}_h]} \simeq$$
$$e^{\frac{\Delta t}{2}\mathcal{L}_r}e^{\frac{\Delta t}{2}\mathcal{L}_o}e^{\frac{\Delta t}{2}\mathcal{L}_h}e^{\Delta t\,\mathcal{L}_v}e^{\frac{\Delta t}{2}\mathcal{L}_h}e^{\frac{\Delta t}{2}\mathcal{L}_o}e^{\frac{\Delta t}{2}\mathcal{L}_r}$$ **ROVOR**

...



NOT ALL INTEGRATORS ARE EQUAL

Some integrators are superior for purely kinetic properties:

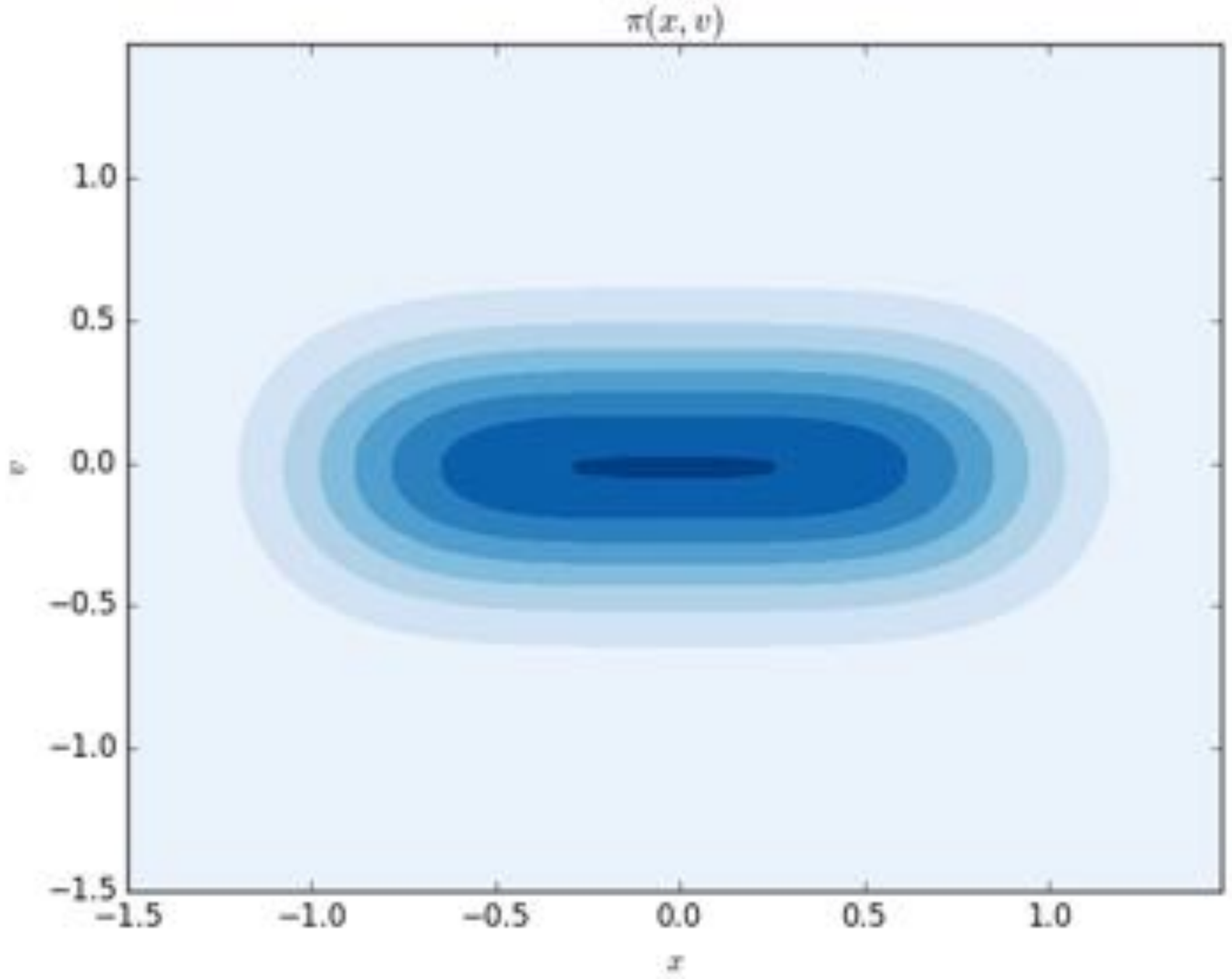
Desideratum	OVRVO	ORVRO	RVOVR	VRORV	VOROV	ROVOR
zero-force MSD	exact	exact	exact	exact	exact	exact
zero-force MSV	exact	exact	exact	exact	exact	exact
zero-force VAC	exact	exact	exact	exact	exact	exact
uniform-force terminal drift	exact	exact	exact	exact	$O(\Delta t^2)$ error	$O(\Delta t^2)$ error
linear-force MSD	exact at $n+\frac{1}{2}$	exact at $n+\frac{1}{2}$	exact at n	exact at n	$O(\Delta t^2)$ error	$O(\Delta t^2)$ error
linear-force MSV	exact at n	exact at n	exact at $n+\frac{1}{2}$	exact at $n+\frac{1}{2}$	$O(\Delta t^4)$ error at n	$O(\Delta t^4)$ error at $n+\frac{1}{2}$
linear-force virial	$O(\Delta t^2)$ error	$O(\Delta t^2)$ error	$O(\Delta t^2)$ error	$O(\Delta t^2)$ error	$O(\Delta t^2)$ error	$O(\Delta t^2)$ error
irreducible	yes	yes	no	no	yes	no
path action	simple	requires values at $n+\frac{1}{2}$	may be infinite	may be infinite	simple	may be infinite
maintains Hamiltonian dependence for large $\gamma\Delta t$	yes	yes	yes	yes	no	no
for reducible dynamics, needs half as many random variables	yes	yes	yes	yes	no	no
in various limits, reduces to other popular integrators	yes	no	no	no	no	no

TABLE II. Comparison of properties for different splittings.

But what if we're only interested in configurational properties (like free energies)?

Sivak, Chodera, Crooks. PRX 3:011007, 2013.

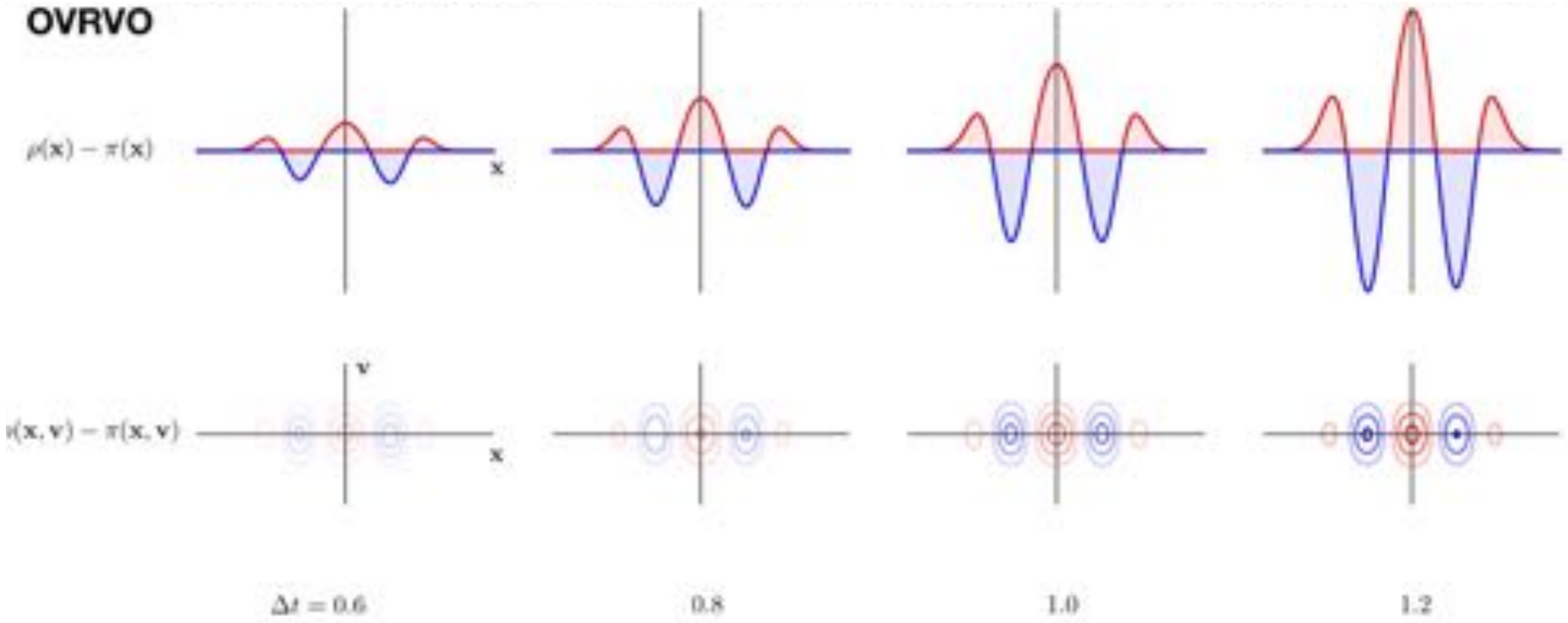
ALL INTEGRATORS INTRODUCE TIMESTEP-DEPENDENT ERROR



SIMPLE QUARTIC POTENTIAL

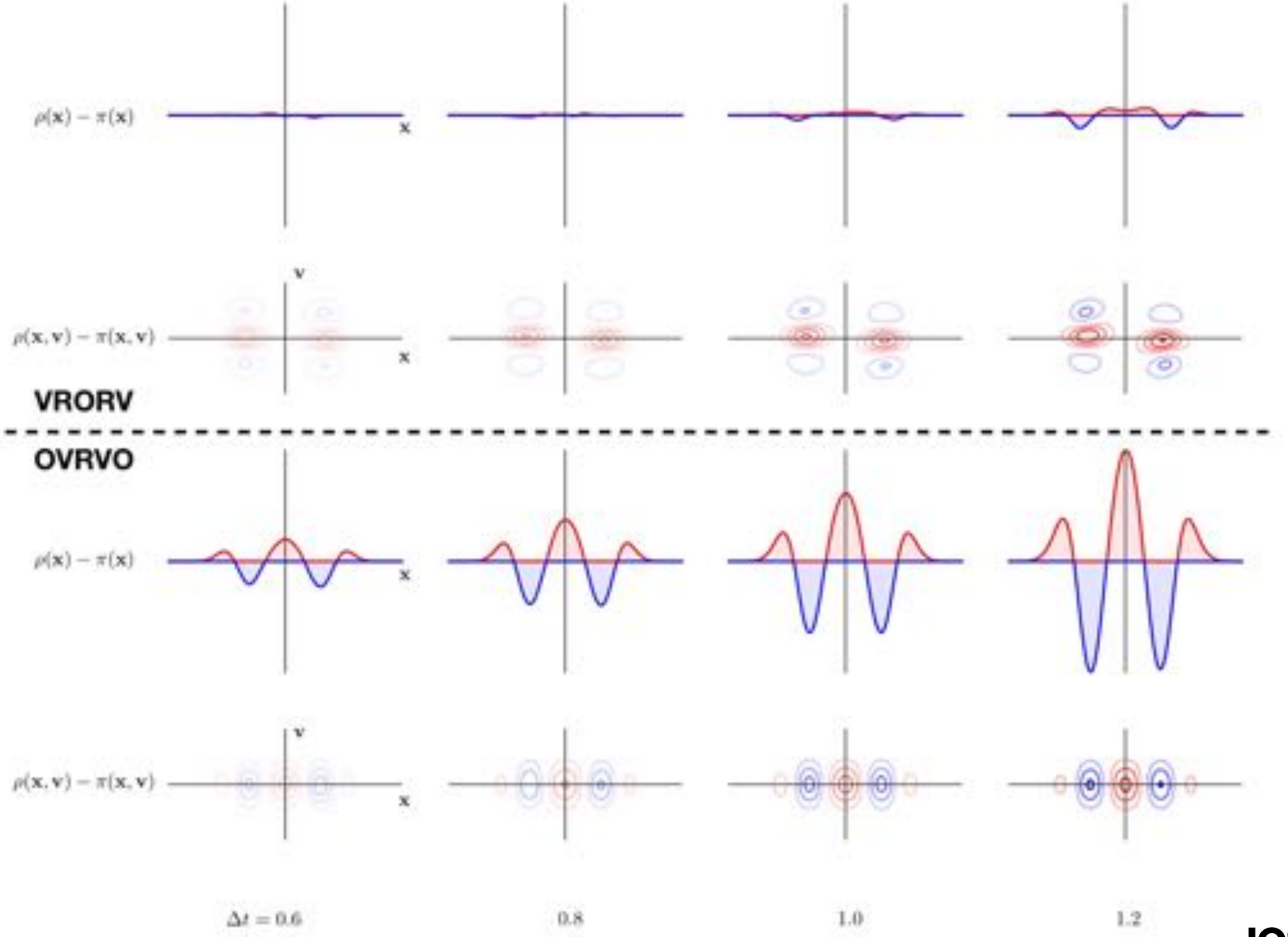
JOSH FASS

ALL INTEGRATORS INTRODUCE TIMESTEP-DEPENDENT ERROR



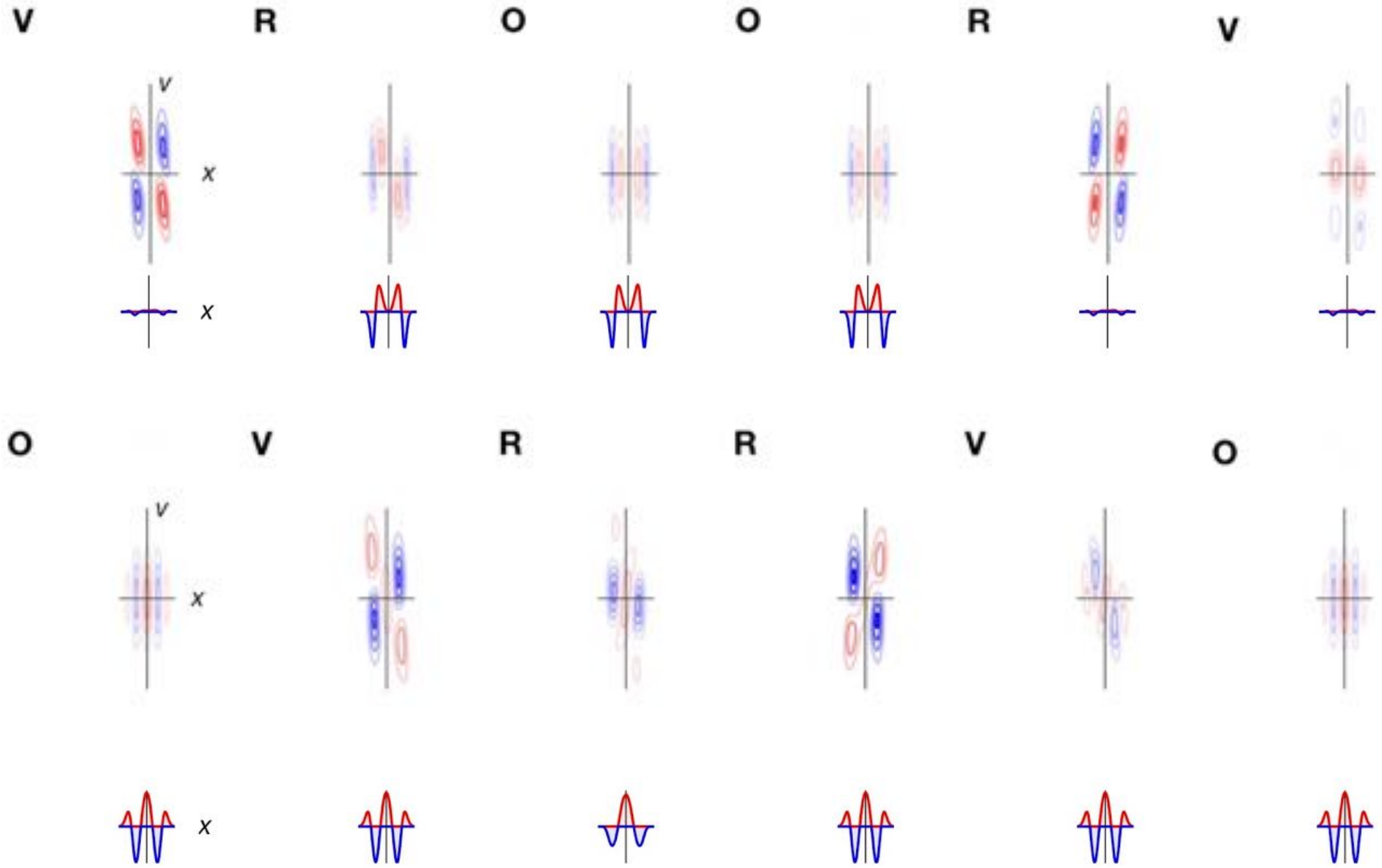
JOSH FASS

NOT ALL INTEGRATORS ARE EQUAL



JOSH FASS

ERRORS MOVE AROUND AT DIFFERENT STAGES OF THE INTEGRATOR CYCLE



JOSH FASS

WHAT IF WE COULD MEASURE INTEGRATOR ERROR VIA A COMMON YARDSTICK?

Here's an example from the literature where the authors pick their favorite order parameters:

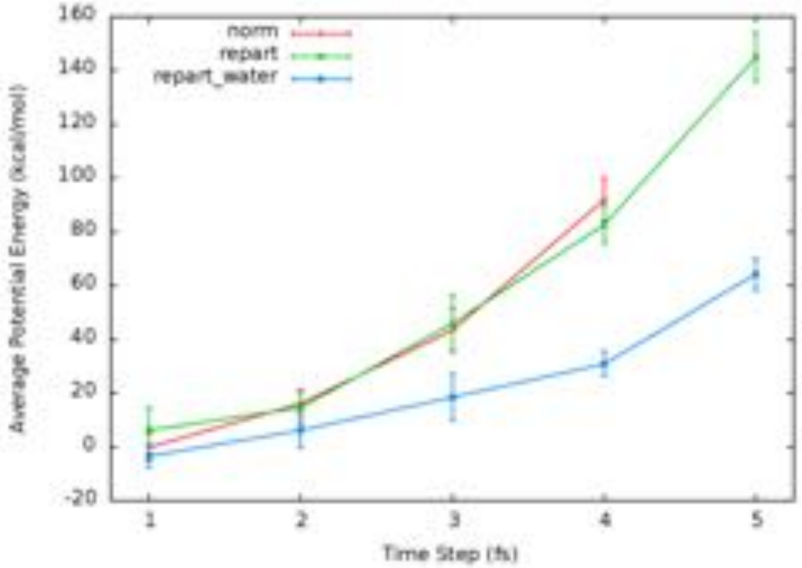


Figure 7. Average potential energies with different mass topology/time step combinations for the $(Ala)_3$ peptide, relative to the norm/1 fs average, averaged over 10 trajectories of each trajectory type. The error bars show the standard deviation in the mean over the 10 trajectories.

Protein–Structure. Figure 8 shows a histogram of the total backbone root-mean-square deviation (RMSD) to the 4LYT

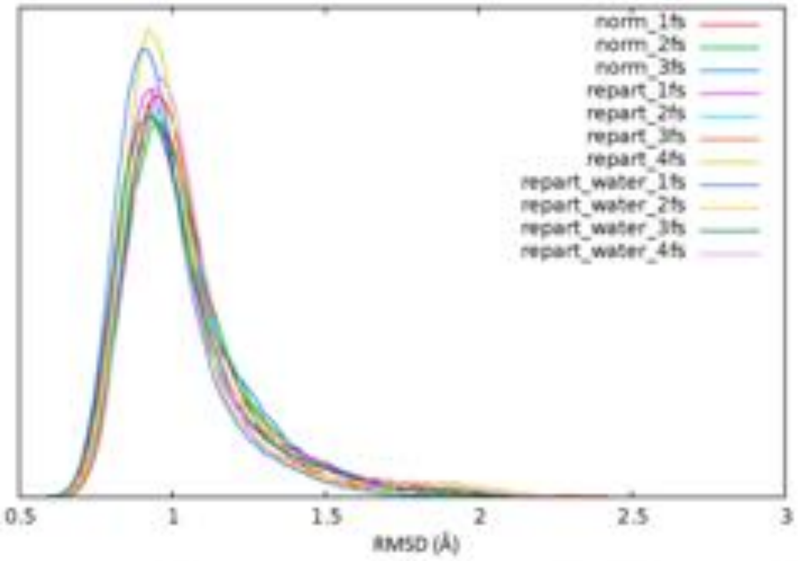


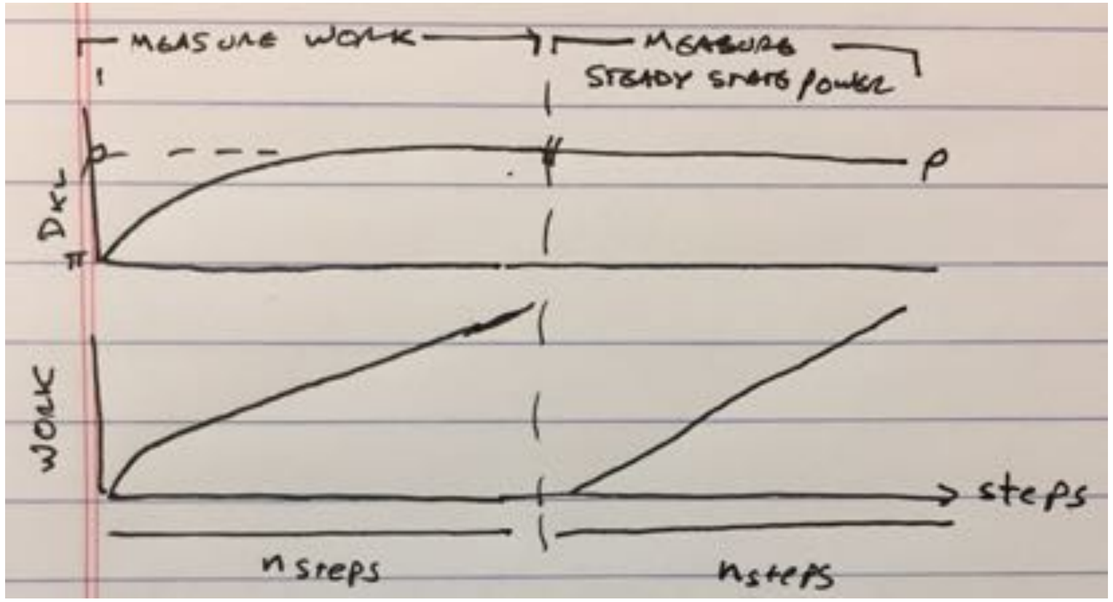
Figure 8. Backbone root-mean-square deviation (RMSD) to a crystal structure in HEWL for each trajectory type.

[Reference not given to protect the guilty]

WHAT IF WE COULD MEASURE INTEGRATOR ERROR VIA A COMMON YARDSTICK?

There is: the **KL-divergence**, which is isomorphic with a **nonequilibrium free energy difference from equilibrium**

$$\Delta F_{neq}(\Delta t) = (\langle E \rangle_{\Delta t} - TS_{\Delta t}) - (\langle E \rangle - TS) = \mathcal{D}_{\mathrm{KL}}(\rho_{\Delta t} \| \pi) \quad \equiv \quad \int d\mathbf{x} \int d\mathbf{v}\, \rho_{\Delta t}(\mathbf{x}, \mathbf{v}) \ln \frac{\rho_{\Delta t}(\mathbf{x}, \mathbf{v})}{\pi(\mathbf{x}, \mathbf{v})}$$



Sivak, Chodera, Crooks. PRX 3:011007, 2013.

We can estimate the nonequilibrium free energy deviation:

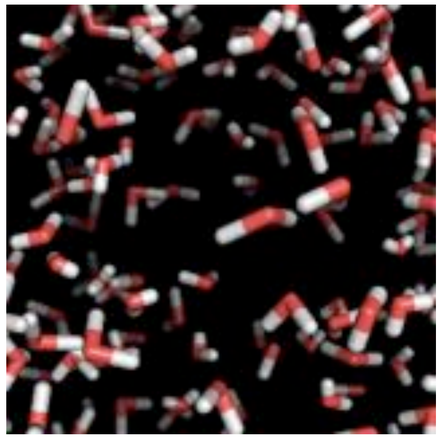
$$\Delta F_{\mathrm{neq}} \equiv F_{\mathrm{neq}} - F_{\mathrm{eq}} = \tfrac{1}{2}\left[\langle W_{\mathrm{shad}} \rangle - (t_{\mathrm{f}} - t_{\mathrm{i}})\mathcal{P}_{\mathrm{ss}}\right]$$

David A. Sivak
Simon Fraser

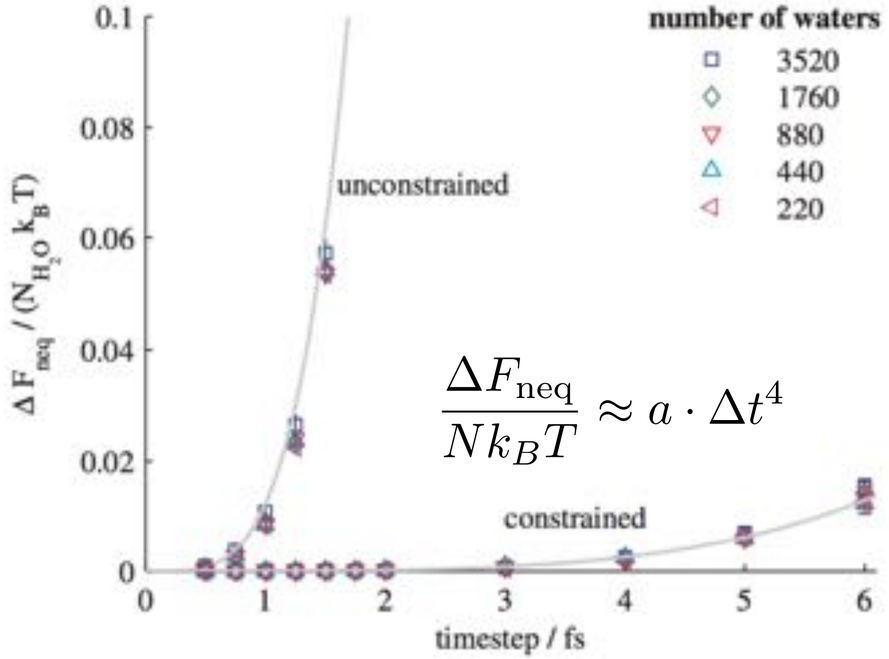
Gavin E. Crooks
Rigetti Computing

WHAT IF WE COULD MEASURE INTEGRATOR ERROR VIA A COMMON YARDSTICK?

$$\Delta F_{neq}(\Delta t) = (\langle E \rangle_{\Delta t} - TS_{\Delta t}) - (\langle E \rangle - TS) = \mathcal{D}_{\mathrm{KL}}(\rho_{\Delta t} \| \pi) \equiv \int d\mathbf{x} \int d\mathbf{v}\, \rho_{\Delta t}(\mathbf{x},\mathbf{v}) \ln \frac{\rho_{\Delta t}(\mathbf{x},\mathbf{v})}{\pi(\mathbf{x},\mathbf{v})}$$



220 TIP3P waters at 298K and 1 atm



$$\frac{\Delta F_{\mathrm{neq}}}{N k_B T} \approx a \cdot \Delta t^4$$

David A. Sivak
Simon Fraser

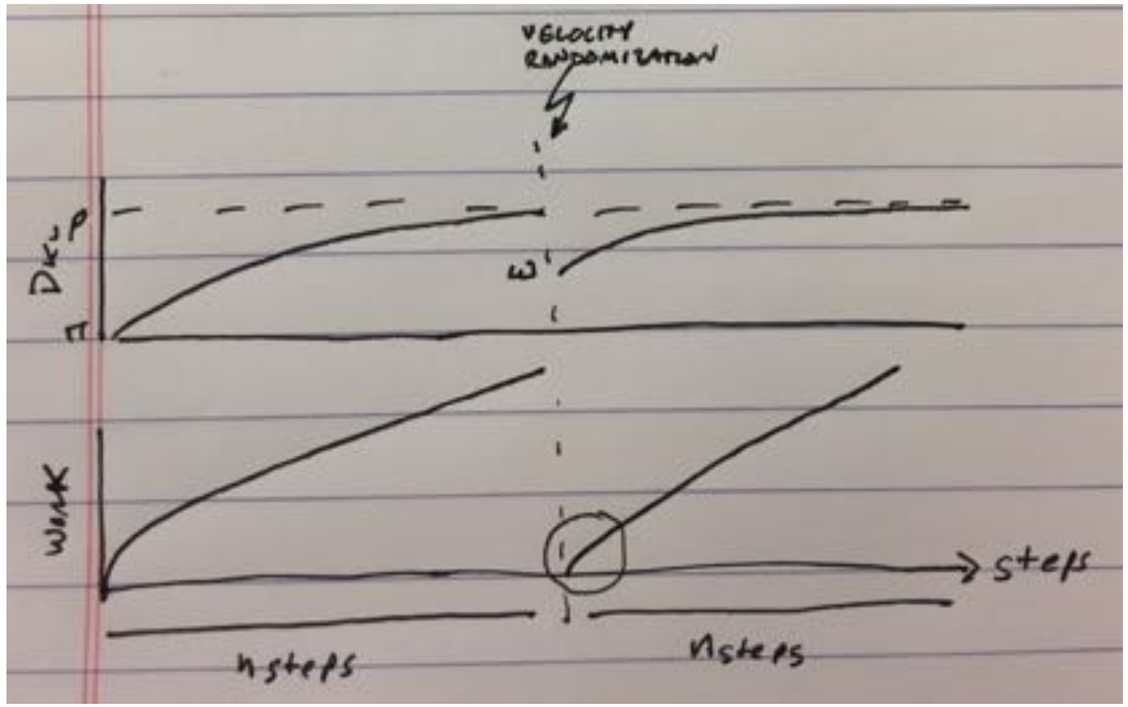
Phase space error grows as $O(\Delta t^4)$!

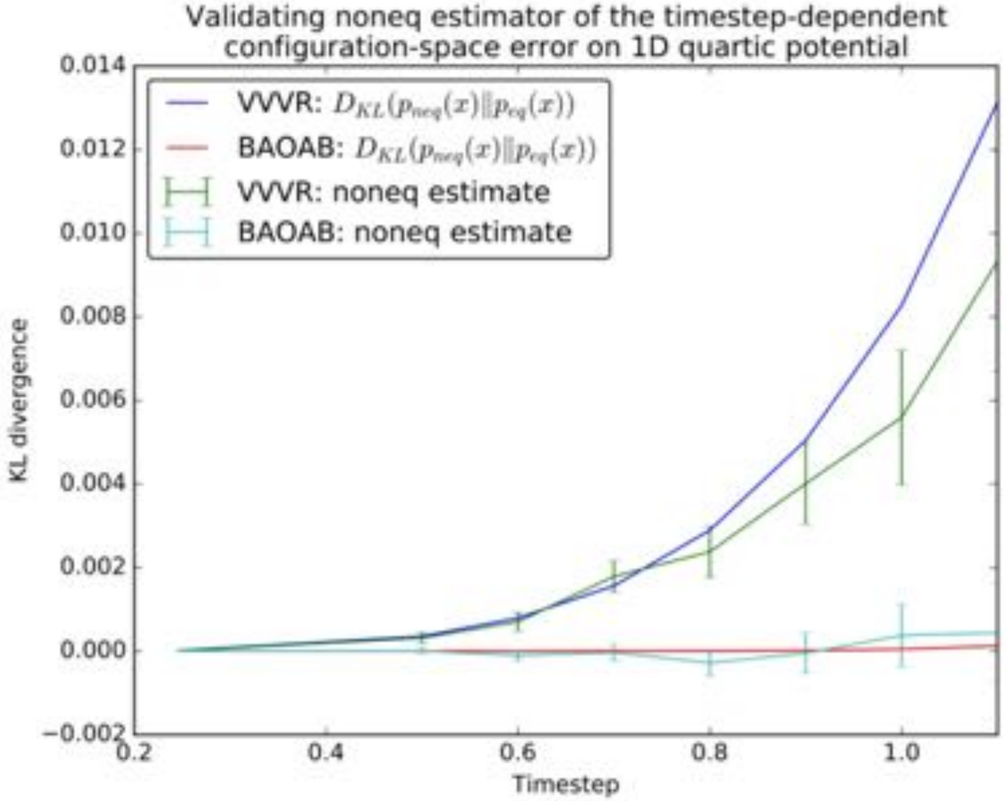
But what about configuration space error?

Sivak, Chodera, Crooks. PRX 3:011007, 2013.

Gavin E. Crooks
Rigetti Computing

WE CAN MEASURE ERROR IN JUST CONFIGURATION SPACE TOO





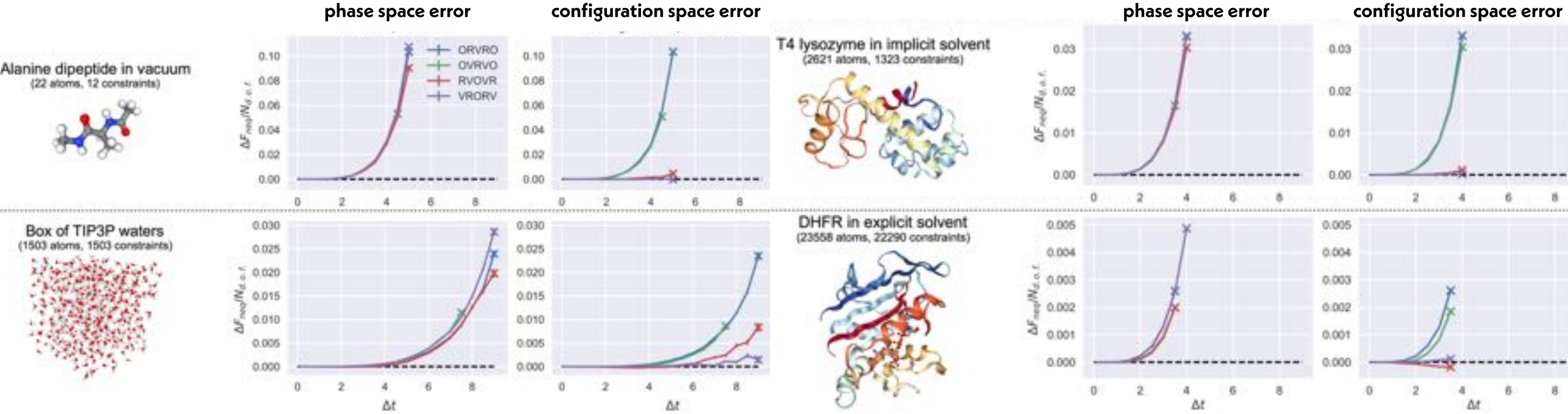
Gavin E. Crooks
Rigetti Computing

1. We can actually measure configuration space error!
2. The error of BAOAB (VRORV) seems really, really small...

JOSH FASS

CERTAIN INTEGRATOR SPLITTINGS LEAD TO EXCEPTIONALLY SMALL CONFIGURATION-SPACE ERROR

$$\Delta F_{neq}(\Delta t) = (\langle E \rangle_{\Delta t} - TS_{\Delta t}) - (\langle E \rangle - TS) = \mathcal{D}_{\mathrm{KL}}(\rho_{\Delta t} \| \pi) \equiv \int d\mathbf{x} \int d\mathbf{v}\, \rho_{\Delta t}(\mathbf{x}, \mathbf{v}) \ln \frac{\rho_{\Delta t}(\mathbf{x}, \mathbf{v})}{\pi(\mathbf{x}, \mathbf{v})}$$



Take-home messages:

1. Error in VRORV almost too small to measure; MUCH better than other integrators
2. We can probably take larger timesteps than we're used to, but it really depends on the system!
   2 fs is not "good enough for everything": sometimes it's too big, often too small

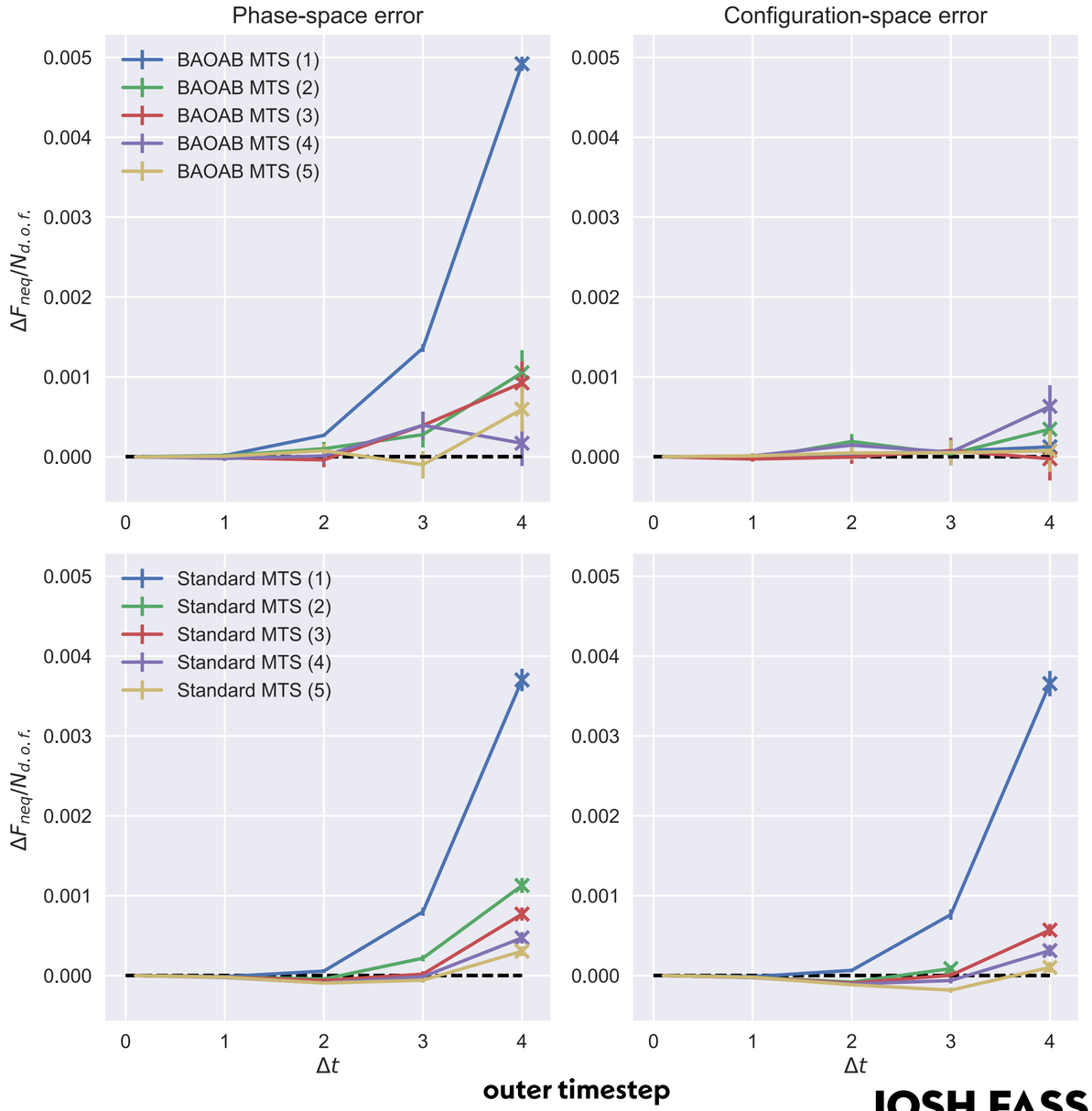
JOSH FASS

MULTIPLE TIMESTEP METHODS CAN HELP!

Multiple timestep scheme:
- inner timestep: valence forces
- outer timestep: nonbonded forces

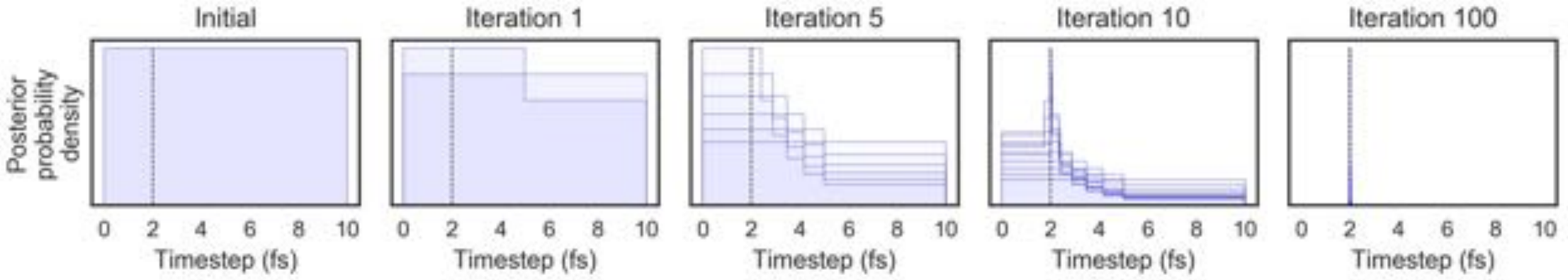
GAFF / AMBER99SB-ILDN / TIP3P
PME elecrostatics

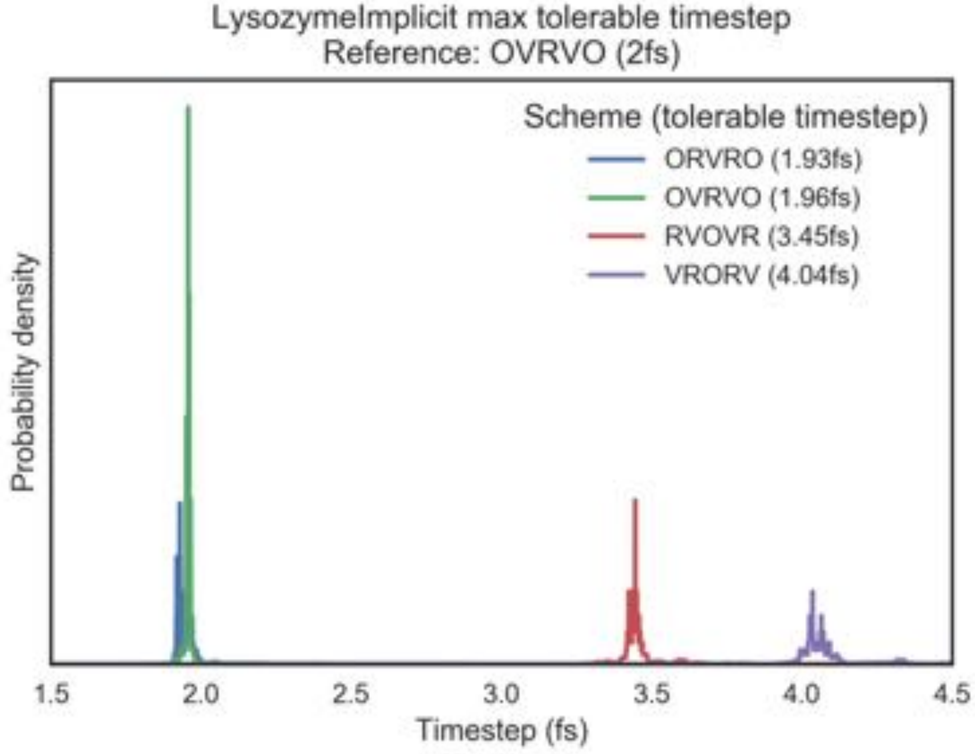
[WARNING: PRELIMINARY DATA; NOT FULLY CONVERGED]



JOSH FASS

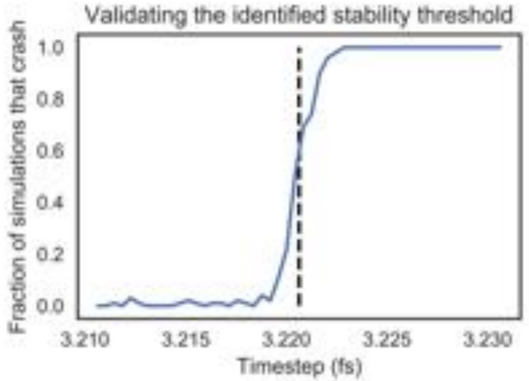
PROBABILISTIC BISECTION ALLOWS FOR AUTOMATED SYSTEM-SPECIFIC DETERMINATION OF MAXIMUM TOLERATED TIMESTEP

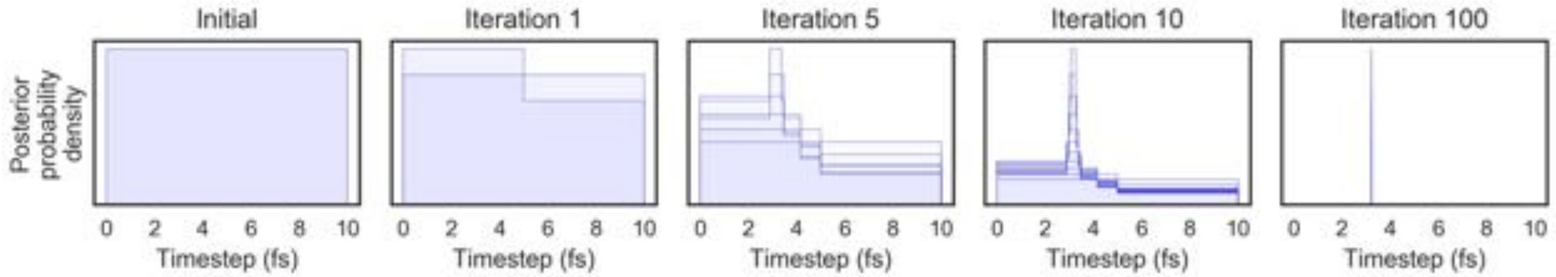


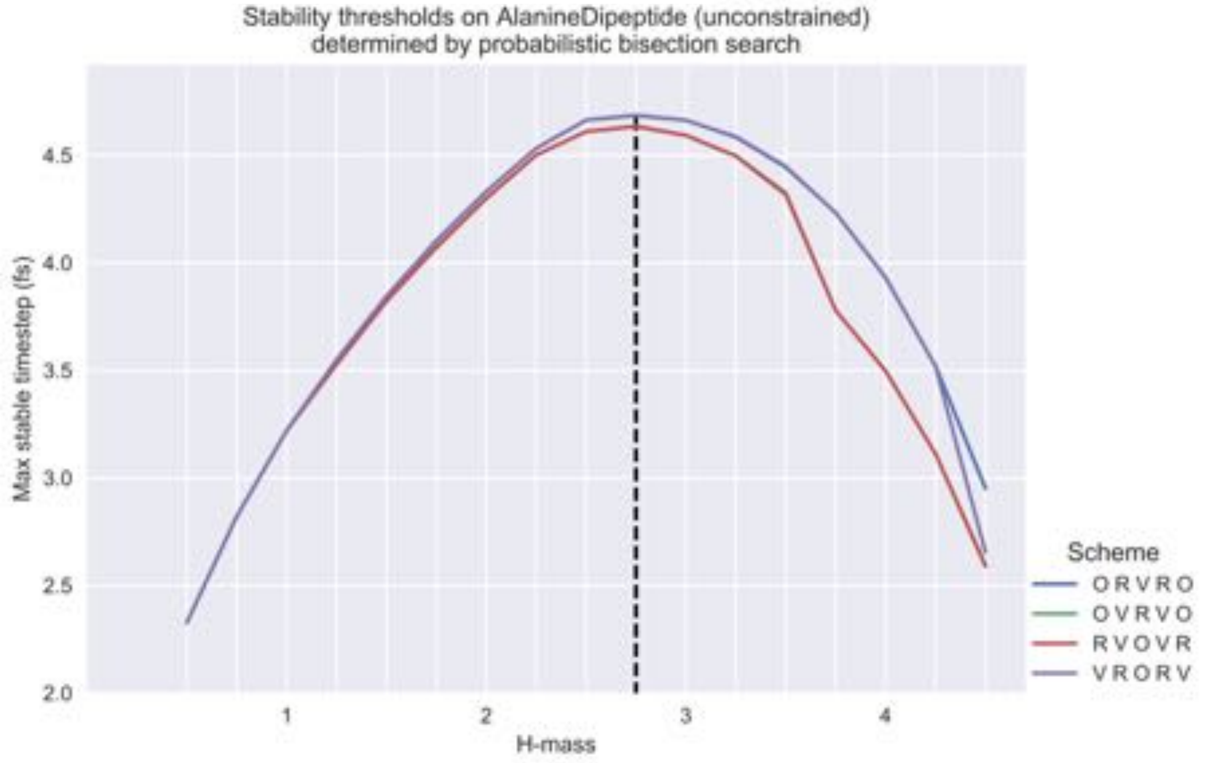


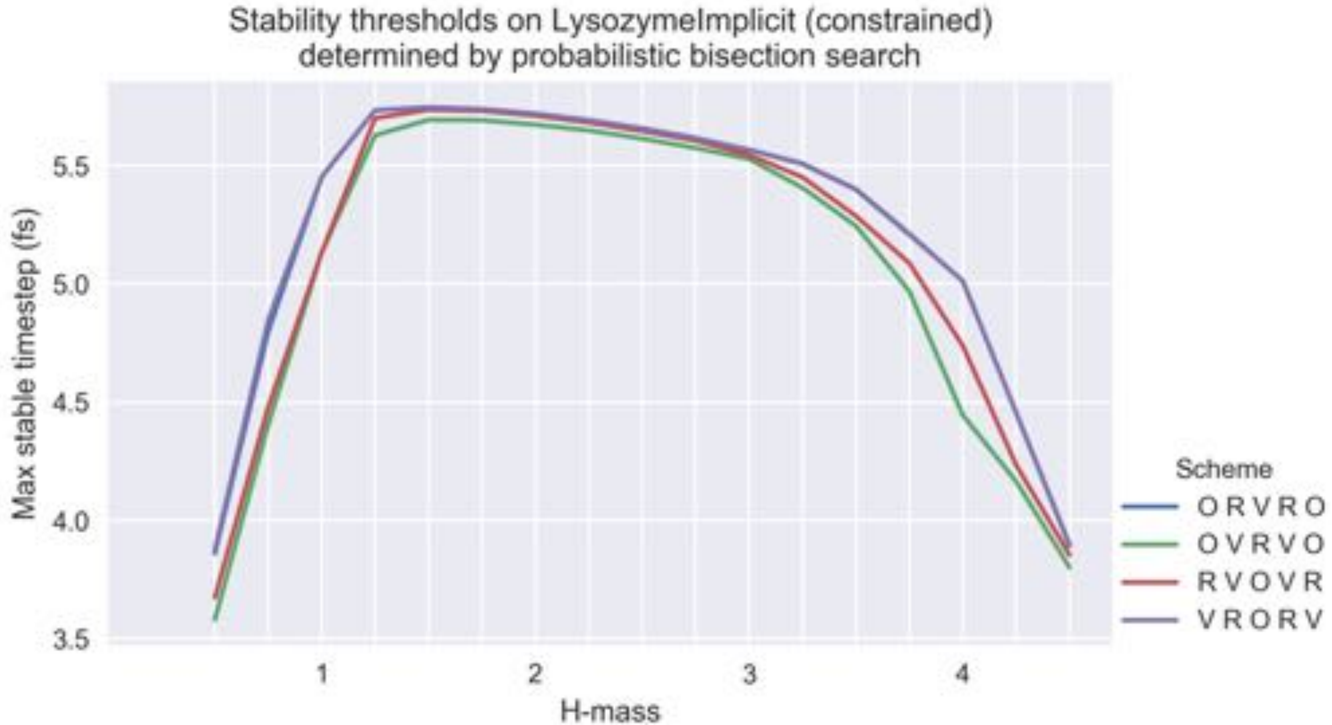
JOSH FASS

PROBABILISTIC BISECTION CAN ALSO IDENTIFY MAXIMUM STABLE TIMESTEP









HOW CAN WE IDENTIFY OPTIMAL THERMODYNAMIC PROTOCOLS?

MANY CHOICES EXIST FOR HOW TO DECOUPLE LIGAND FROM ENVIRONMENT

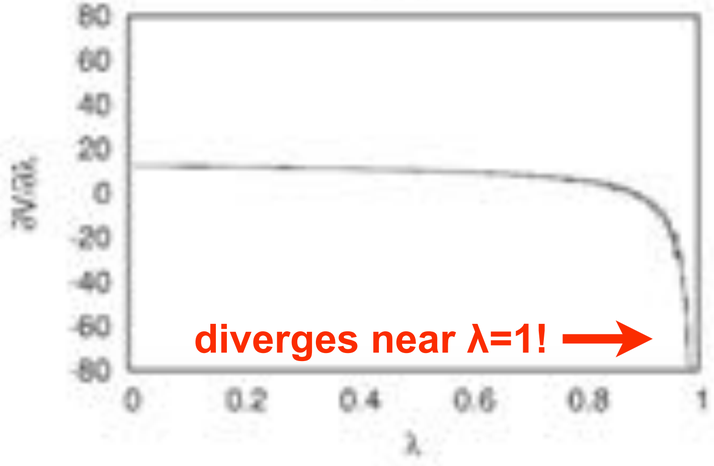
λ-DEPENDENT POTENTIAL MUST BE WELL-BEHAVED THROUGHOUT

$$\Delta F = \int_0^1 d\lambda' \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_{\lambda'}$$

THERMODYNAMIC INTEGRATION

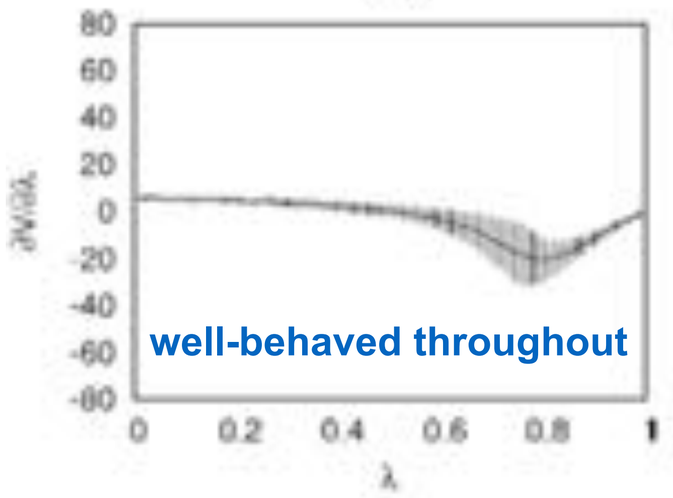
LINEAR ALCHEMICAL SCALING

$$U(r;\lambda) = 4\epsilon(1-\lambda)\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$



"SOFT-CORE" FORM

$$U(r;\lambda) = 4\epsilon(1-\lambda)\left[\frac{1}{[\alpha\lambda + (r/\sigma)^6]^2} - \frac{1}{[\alpha\lambda + (r/\sigma)^6]}\right]$$



T. A. Steinbrecher, D. L. Mobley, and D. A. Case. JCP 127:214108 (2007)

IS THERE AN OPTIMAL PROTOCOL?

THE THERMODYNAMIC METRIC TENSOR MEASURES HOW RAPIDLY THE EQUILIBRIUM DISTRIBUTION CHANGES AS CONTROL PARAMETERS ARE TWIDDLED.

$$g_{ij}(\boldsymbol{\lambda}) \equiv \left\langle \frac{\partial \ln \pi(\boldsymbol{x};\boldsymbol{\lambda})}{\partial \lambda_i} \frac{\partial \ln \pi(\boldsymbol{x};\boldsymbol{\lambda})}{\partial \lambda_j} \right\rangle_{\boldsymbol{\lambda}}$$

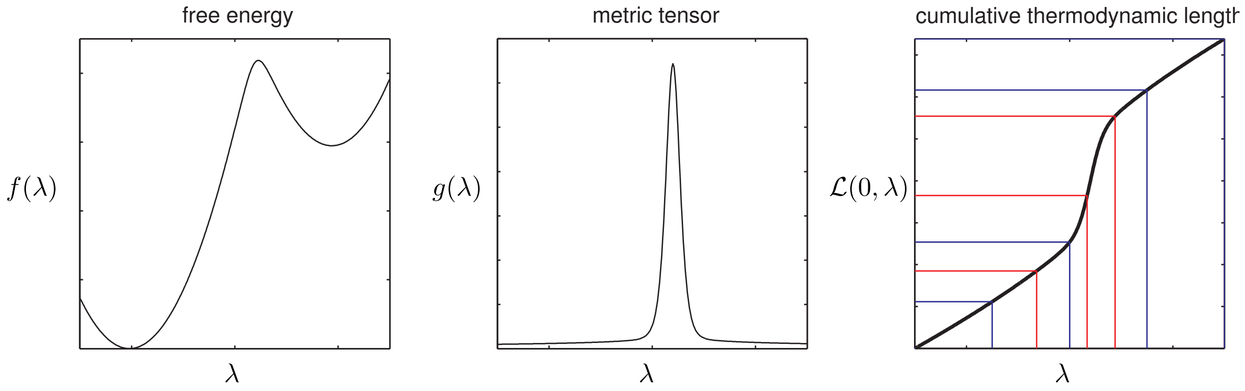
THE THERMODYNAMIC LENGTH MEASURES HOW MUCH THE DISTRIBUTION HAS CHANGES FROM ONE VALUE OF CONTROL PARAMETERS TO ANOTHER.

$$\mathcal{L} \equiv \int_0^\tau dt \left(\dot{\boldsymbol{\lambda}}^{\mathrm{T}} \boldsymbol{g} \, \dot{\boldsymbol{\lambda}} \right)^{1/2}$$

.

OPTIMAL PROTOCOLS ARE GEODESICS IN THERMODYNAMIC METRIC SPACE; THEY EQUALIZE THERMODYNAMIC LENGTH BETWEEN MEASUREMENTS.

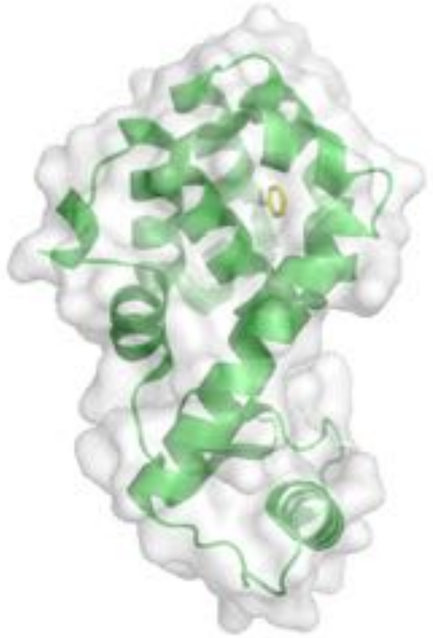
$$\mathrm{var}(\Delta \hat{f}) \geq N^{-1} \mathcal{L}(\lambda_a, \lambda_b)^2$$



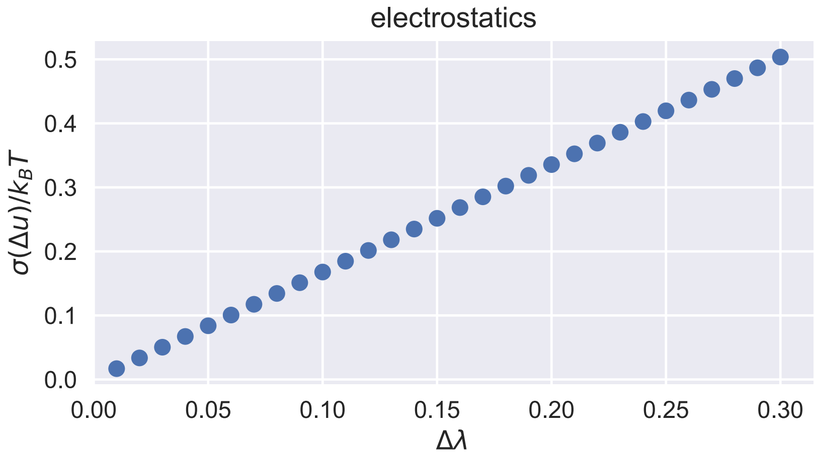
OPTIMIZED PATHS CAN YIELD ORDERS OF MAGNITUDE REDUCTION IN VARIANCE!

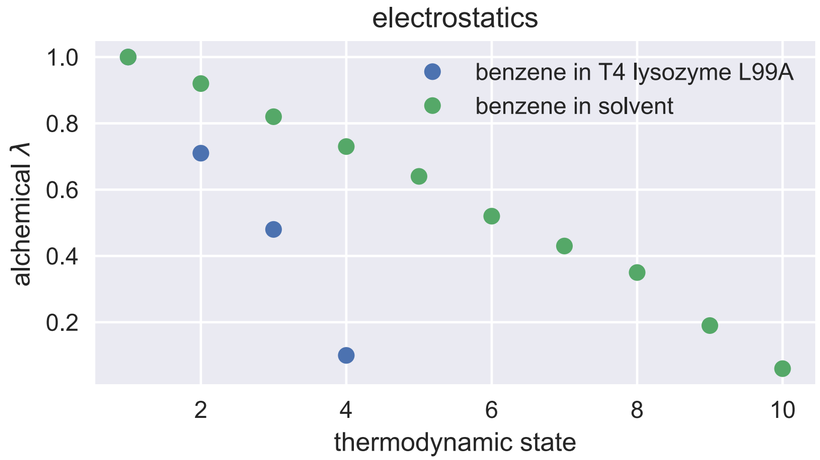
Crooks, PRL 99:100602, 2007; Shenfeld et al. PRE 80:046705, 2009; Sivak and Crooks PRL 108:190602, 2012.

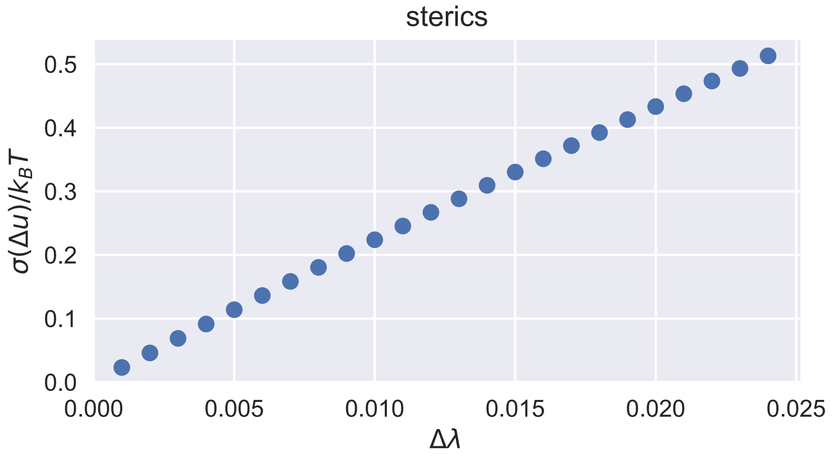
A SIMPLE SCHEME FOR OPTIMIZING PROTOCOL GIVES A GOOD INITIAL GUESS

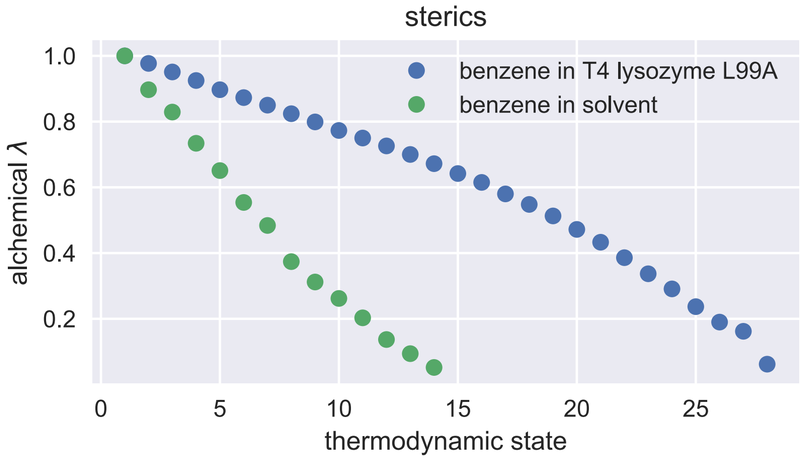


benzene : T4 lysozyme L99A
GAFF / AMBER99SB-ILDN / TIP3P
PME





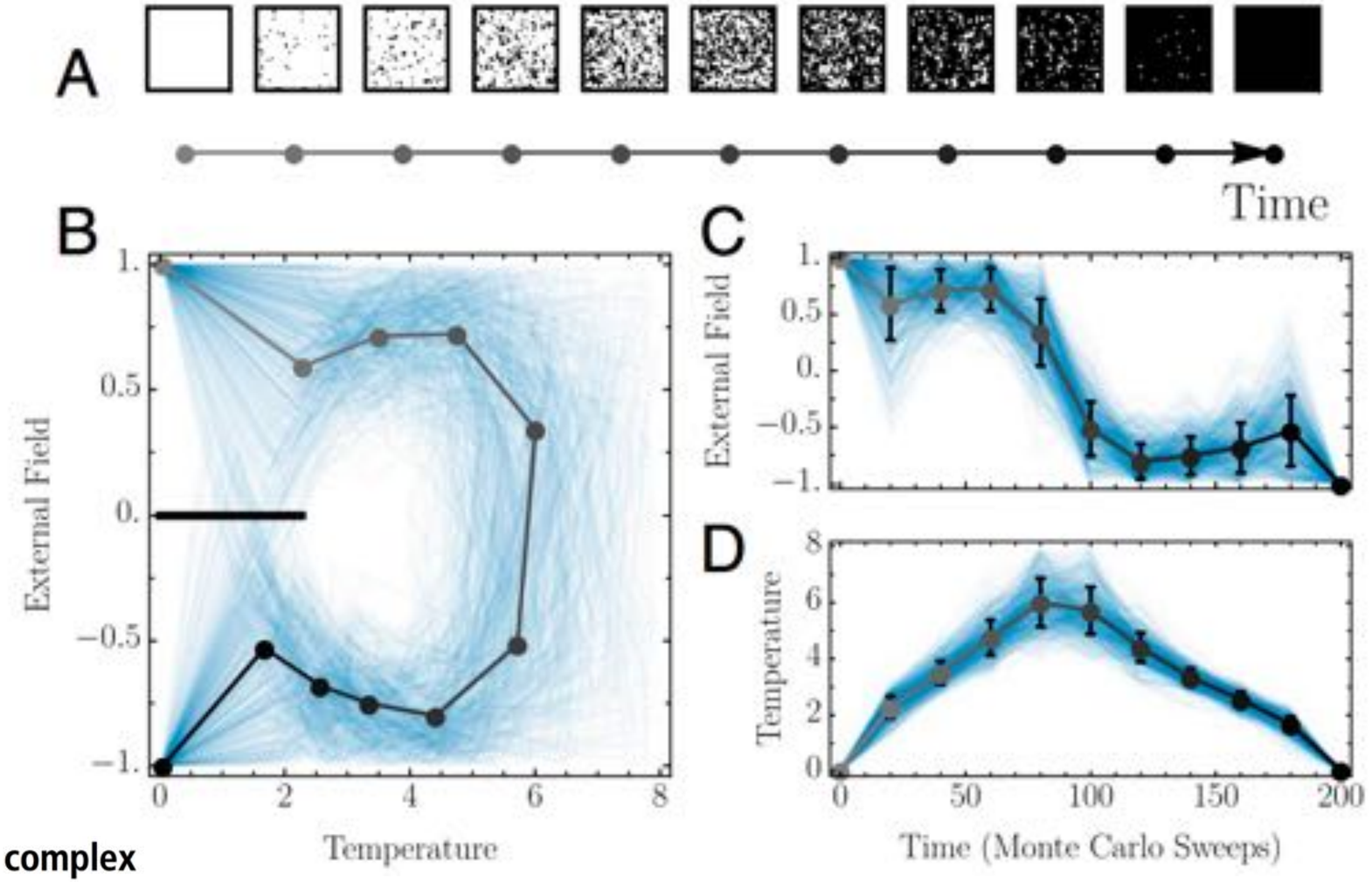




ANDREA RIZZI
CBM graduate student

(Suggested by David Minh and Huafeng Xu)

ADAPTIVE PROTOCOL REFINEMENT WILL LIKELY BE FRUITFUL AREA OF RESEARCH



Near-optimal protocols in complex nonequilibrium transformations

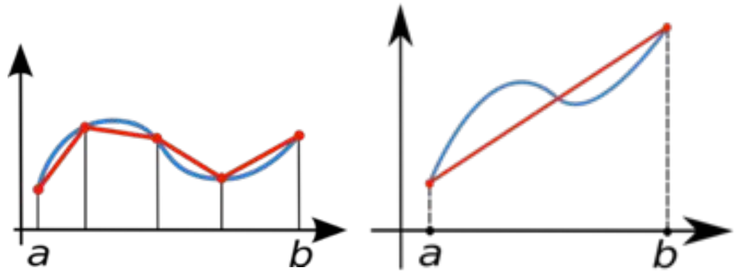
Todd R. Gingrich[a,b,1], Grant M. Rotskoff[c], Gavin E. Crooks[d,e], and Phillip L. Geissler[b,c,d,f]

PNAS 113:10263, 2016.

HOW CAN WE EXTRACT ALL INFORMATION FROM THE DATA?
NOT ALL ESTIMATORS ARE EQUAL
(IN BIAS AND STATISTICAL EFFICIENCY)

THERMODYNAMIC INTEGRATION (TI)

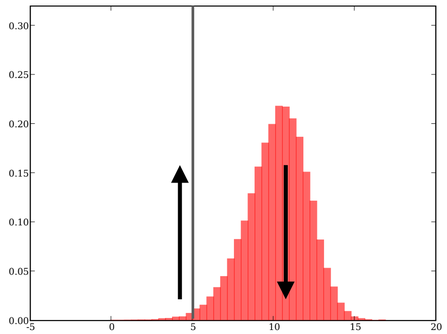
$$\Delta F = \int_{\lambda_1}^{\lambda_2} d\lambda' \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda'} \approx \frac{\Delta\lambda}{2} \left[\left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda_1} + \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_{\lambda_2} \right]$$



QUADRATURE ERROR (BIAS) DIFFICULT TO QUANTIFY;
DERIVATIVES OFTEN NOT AVAILABLE IN SIMULATION CODES

EXPONENTIAL REWEIGHTING (EXP)

$$\Delta F = -\beta^{-1} \ln \left\langle e^{-\beta(U_2 - U_1)} \right\rangle_{\lambda_1} = +\beta^{-1} \ln \left\langle e^{-\beta(U_1 - U_2)} \right\rangle_{\lambda_2}$$

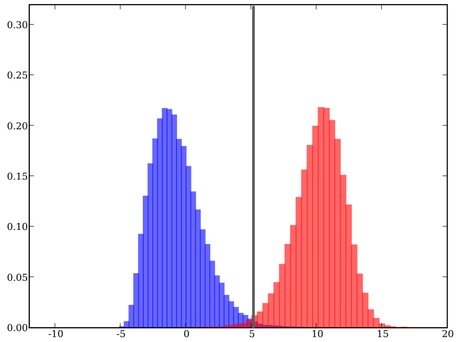


Zwanzig RW. *JCP* **22:**1420, 1954.
Shirts MR and Pande VS. *JCP* **122:**144107, 2005.

SUFFERS FROM LARGE BIAS AND VARIANCE

BENNETT ACCEPTANCE RATIO (BAR)

$$\Delta F = -\beta^{-1} \ln \frac{\langle f(U_2 - U_1) \rangle_{\lambda_1}}{\langle f(U_1 - U_2) \, \exp[-\beta(U_2 - U_1)] \rangle_{\lambda_2}}$$



Bennett CH. *J Comput Phys* **22:**245, 1976.
Shirts MR, Bair E, Hooker G, and Pande VS. *PRL* **91:**140601, 2003.

SUPERIOR, BUT ONLY APPLICABLE TO TWO STATES—WHY CAN'T WE USE ALL THE DATA?

THE REDUCED POTENTIAL GENERALIZES ALL THERMODYNAMIC STATES OF INTEREST

Define the reduced potential for a state k as a combination of terms

$$u_k(\mathbf{x}) = \beta_k[U_k(\mathbf{x}) + p_k V(\mathbf{x}) + \mu_k^{\mathrm{T}} \mathbf{N}(\mathbf{x})]$$

with thermodynamic parameters for each state

β_k	**inverse temperature**
U_k	**potential energy function**
p_k	**external pressure**
μ_k	**chemical potential of exchangeable species**

where

$\mathbf{x}$	**microstate or configuration**
$V(\mathbf{x})$	**volume of simulation box**
$\mathbf{N}(\mathbf{x})$	**number of each chemical species in system**

protonation states
number of counter ions

The target configuration space density is given by

$$\pi_k(\mathbf{x}) = Z_k^{-1} \exp[-u_k(\mathbf{x})] \qquad Z_k = \int d\mathbf{x} \exp[-u_k(\mathbf{x})]$$

Covers many common thermodynamic ensembles: NVT, NPT, μVT, μPT

Shirts MR and Chodera JD. “Statistical optimal analysis of samples from multiple equilibrium states.”, *JCP* **129**:124105 (2008).

EXTENDED BRIDGE SAMPLING ESTIMATORS

DEFINE A FAMILY OF USEFUL ESTIMATORS

Suppose we have K states defined by unnormalized distribution functions q(x):

$$\pi_k(\boldsymbol{x}) = c_k^{-1}\, q_k(\boldsymbol{x}) \qquad \mathrm{q}_k(\boldsymbol{x}) > 0$$

$$\mathrm{c}_k = \int d\boldsymbol{x}\, q_k(\boldsymbol{x})$$

Start with the identity

$$c_i \left\langle \alpha_{ij} q_j \right\rangle_i = \int_{\mathbf{\Gamma}} d\boldsymbol{x}\, q_i(\boldsymbol{x})\, \alpha_{ij}(\boldsymbol{x})\, q_j(\boldsymbol{x}) = c_j \left\langle \alpha_{ij} q_i \right\rangle_j \qquad i,j = 1,\ldots,K$$

$$\alpha_{ij}(\boldsymbol{x}) \ \text{arbitrary}$$

Sum over all K states

$$\sum_{j=1}^{K} c_i \left\langle \alpha_{ij} q_j \right\rangle_i = \sum_{j=1}^{K} c_j \left\langle \alpha_{ij} q_i \right\rangle_j \qquad i = 1,\ldots,K$$

Substitute the empirical estimator for expectations

$$\sum_{j=1}^{K} \frac{\hat{c}_i}{N_i} \sum_{n=1}^{N_i} \alpha_{ij} q_j(\boldsymbol{x}_{in}) \quad = \quad \sum_{j=1}^{K} \frac{\hat{c}_j}{N_j} \sum_{n=1}^{N_j} \alpha_{ij} q_i(\boldsymbol{x}_{jn})$$

Defines a family of estimators parameterized by choice of $\alpha_{ij}(x)$

Ratios of normalization constants c_i/c_j determined by solving a set of K nonlinear equations.

Kong A, McCullagh P, Meng X-L, Nicolae D, and Tan Z. J. Royal. Stat. Soc. B **65:**585 (2003).
Tan Z. "On a likelihood approach for Monte Carlo integration.", *J. Am. Stat. Assoc.* **99**: 468 (2004).

THE ASYMPTOTICALLY LOWEST-VARIANCE ESTIMATOR IS KNOWN!

Optimal choice of $\alpha_{ij}(\boldsymbol{x})$ among large class of functions is:

$$\alpha_{ij}(\boldsymbol{x}) = \frac{N_j \, \hat{c}_j^{-1}}{\sum_{k=1}^{K} N_k \, \hat{c}_k^{-1} \, q_k(\boldsymbol{x})}$$

Covariance of any functions ϕ, ψ of ratios $\hat{c}_i / \hat{c}_j$ can be estimated from asymptotic covariance

$$\mathrm{cov}\left(\hat{\phi}, \hat{\psi}\right) = \sum_{i,j=1}^{K} \frac{\partial \phi}{\partial \theta_i} \hat{\Theta}_{ij} \frac{\partial \psi}{\partial \theta_j} .$$

where the asymptotic covariance estimate is computed in terms of $\hat{\theta}_i \equiv -\ln \hat{c}_i$

$$\hat{\mathbf{\Theta}} = [(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{+} - \mathbf{N}]^{+}$$

$$\hat{\Theta}_{ij} \equiv \mathrm{cov}\left(\hat{\theta}_i, \hat{\theta}_j\right)$$

$$W_{ni} = \frac{\hat{c}_i^{-1} \, q_i(x_n)}{\sum_{k=1}^{K} N_k \, \hat{c}_k^{-1} \, q_k(x_n)}$$

$$\mathbf{N} = \mathrm{diag}\{N_1, N_2, \ldots, N_K\}$$

code available at:
http://pymbar.org

MULTISTATE GENERALIZATION OF BENNETT ACCEPTANCE RATIO (MBAR)

Kong A, McCullagh P, Meng X-L, Nicolae D, and Tan Z. J. Royal. Stat. Soc. B **65:**585 (2003).
Tan Z. "On a likelihood approach for Monte Carlo integration.", *J. Am. Stat. Assoc.* **99**: 468 (2004).
Shirts MR and Chodera JD. “Statistical optimal analysis of samples from multiple equilibrium states.”, *JCP* **129**:124105 (2008).

HOW CAN WE RUN CALCULATIONS QUICKLY AND RAPIDLY TEST NEW ALGORITHMS?

YANK: AN OPEN-SOURCE, COMMUNITY-ORIENTED PLATFORM FOR GPU-ACCELERATED FREE ENERGY CALCULATIONS

NVIDIA GTX-1080 ($650)
9 TFLOP/S SINGLE PRECISION

OpenMM speedup (GTX-1080) over 12-core Xeon X5650 CPU for DHFR

method	natoms	gromacs CPU	OpenMM GPU	speedup
GB/SA	2,489	2.54 ns/day	789 ns/day	**311 x**
RF	23,558	18.8 ns/day	572 ns/day	**30.4 x**
PME	23,558	6.96 ns/day	337 ns/day	**48.4 x**

http://openmm.org OpenMM 7.1.0 development snapshot benchmark
gromacs benchmarks from http://biowulf.nih.gov/apps/gromacs-gpu.html

YANK 1.0 coming Aug 2017!

Docs » YANK

YANK

A GPU-accelerated Python framework for exploring algorithms for alchemical free energy calculations

Features

- Modular Python framework for easily exploring new algorithms
- GPU-accelerated via the OpenMM toolkit
- Alchemical free energy calculations in both **explicit** and **implicit** solvent
- Hamiltonian exchange among alchemical intermediates with Gibbs sampling framework
- General Markov chain Monte Carlo framework for exploring enhanced sampling methods
- Built-in equilibration detection and convergence diagnostics
- Support for AMBER prmtop/inpcrd files
- Support for absolute binding free energy calculations
- Support for transfer free energies (such as hydration free energies)

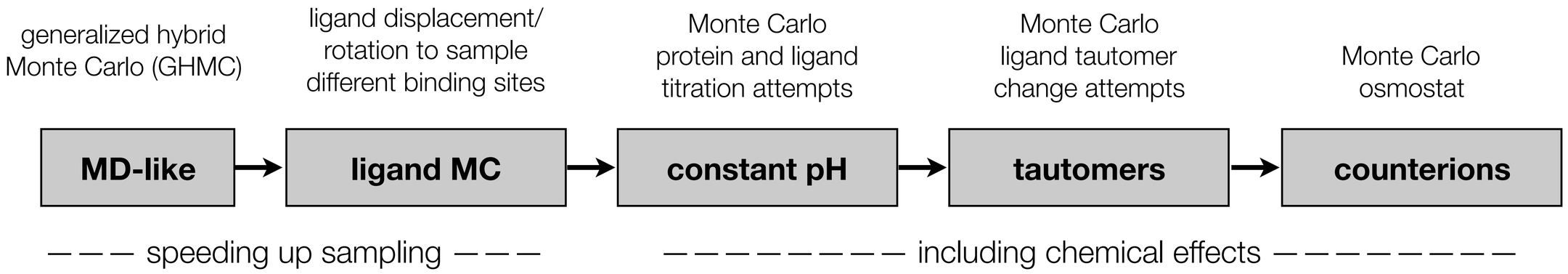
http://www.getyank.org

A free, open-source, extensible platform for free energy calculations and ligand design

ANDREA RIZZI
CBM graduate student

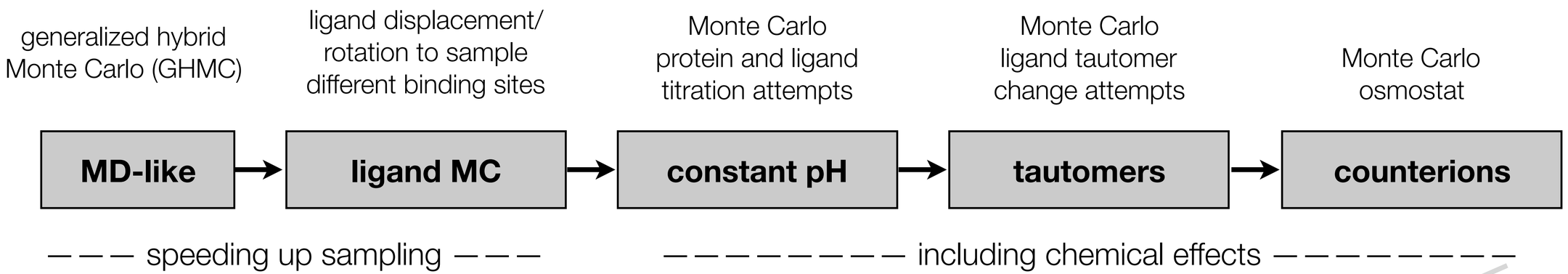
LEVI NADEN
Merck KGaA-sponsored postdoc

MARKOV CHAIN MONTE CARLO (MCMC) FRAMEWORK PERMITS EFFICIENT MIXING* OF MD AND MC

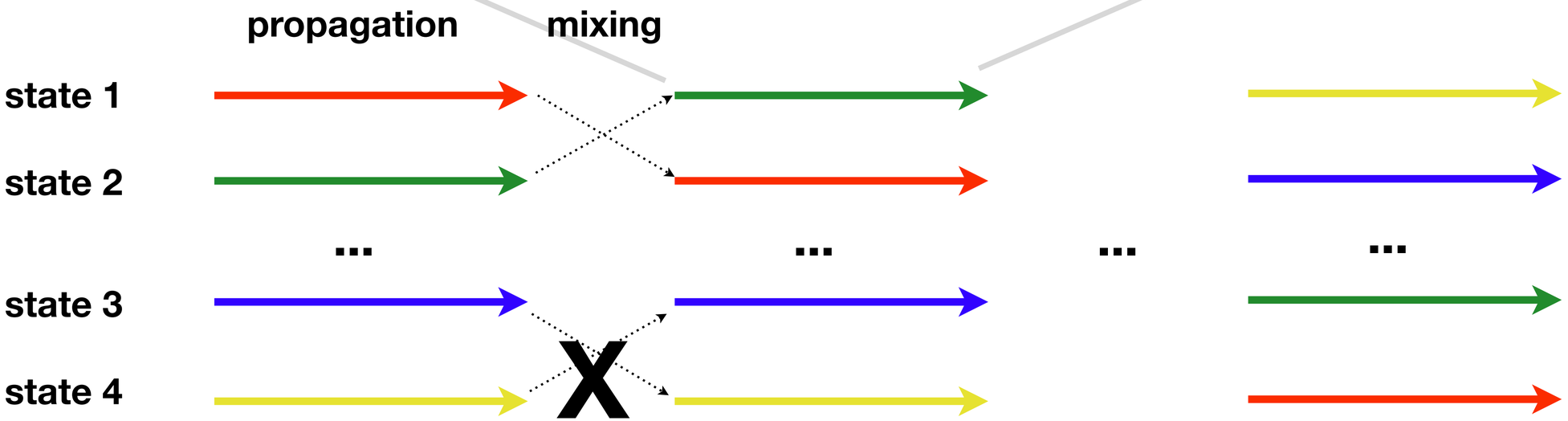


* Must be careful to avoid mismatch between MD/MC densities.

MCMC IS COMBINED WITH REPLICA EXCHANGE TO DECREASE CORRELATION TIMES

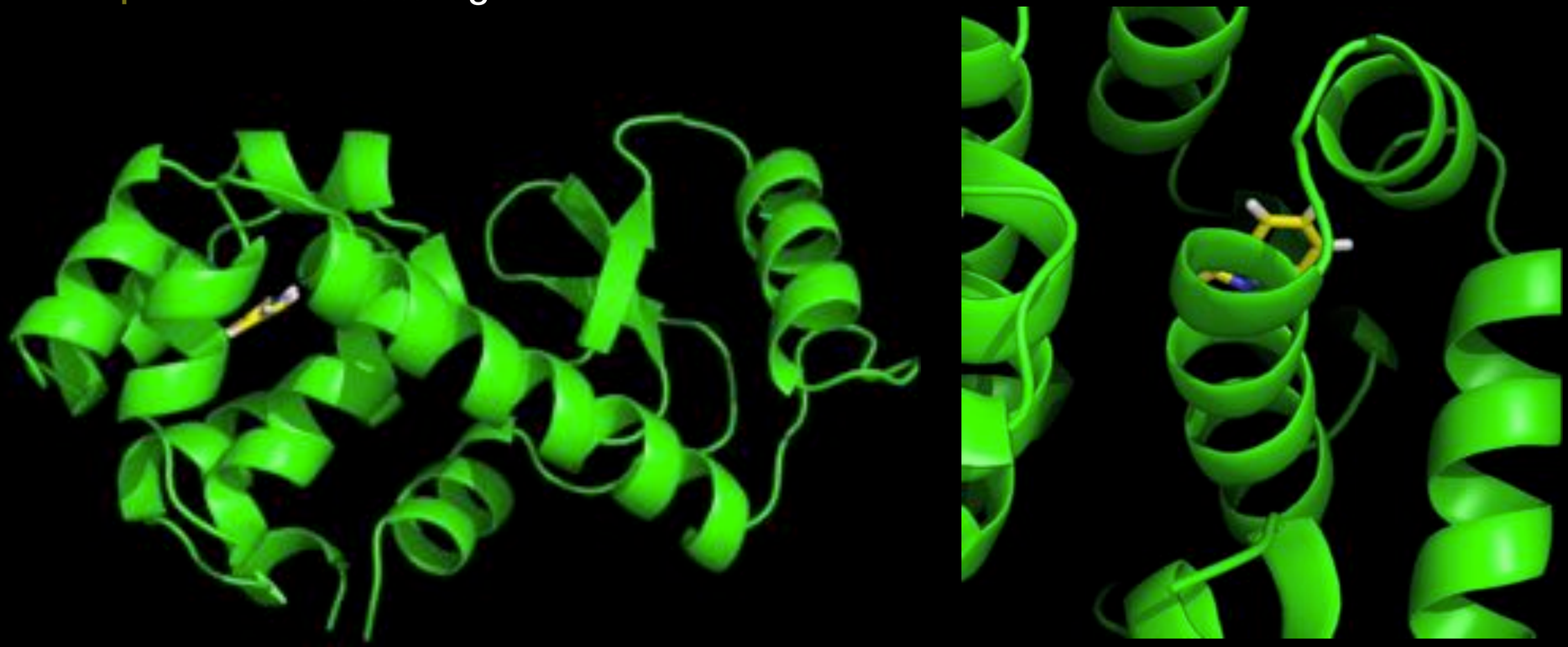


* Must be careful to avoid mismatch between MD/MC densities.



HAMILTONIAN EXCHANGE PROTOCOL ALLOWS FOR REPEATED BINDING/UNBINDING EVENTS AND REORIENTATION IN SITE

solid fully interacting
transparent noninteracting



indole binding to T4 lysozyme L99A
12 h on 2 NVIDIA Tesla M2090 GPUs
Hamiltonian exchange with Gibbs sampling

getyank.org

Chodera and Shirts. JCP 135:194110, 2011
Wang, Chodera, Yang, and Shirts. JCAMD 27:989, 2013.
http://github.org/choderalab/yank

YAML SYNTAX ALLOWS FOR EASY AUTOMATION OF FREE ENERGY CALCULATIONS

```yaml
---
options:
  verbose: true
  output_dir: test_kinase
  temperature: 310*kelvin
  pressure: 1*atmosphere
  constraints: HBonds
  number_of_iterations: 200
  minimize: yes

molecules:
  abl:
    filepath: !Combinatorial [2HYY.pdb, 3C59.pdb]
    strip_protons: yes
  src:
    filepath: !Combinatorial [2OIQ_A.pdb, 3EL7_A.pdb]
    strip_protons: yes

  imatinib:
    filepath: clinical-kinase-inhibitors.csv
    select: 0
    antechamber:
      charge_method: bcc
    epik:
      select: 0
      tautomerize: no
      ph: 7.4
      ph_tolerance: 5
```

```
# Run Schrodinger's tool Epik with default parameters and
# select the most likely protonation state for the molecule
# (in solution). More control over epik parameters is
# possible (see YANK documentation).
```

YAML SYNTAX ALLOWS FOR EASY AUTOMATION OF FREE ENERGY CALCULATIONS

```yaml
solvents:
  PME:
    nonbonded_method: PME
    nonbonded_cutoff: 1*nanometer
    switch_distance: 0.9*nanometer
    ewald_error_tolerance: 0.0003
    clearance: 10*angstroms
    positive_ion: Na+
    negative_ion: Cl-
  GBSA:
    nonbonded_method: NoCutoff                    # Implicit solvent using GBSA/OBC2 model.
    implicit_solvent: OBC2
    implicit_solvent_salt_concentration: 1.0*mole/liter
    solute_dielectric: 1.5
    solvent_dielectric: 80.0
  vacuum:
    nonbonded_method: NoCutoff
```

YAML SYNTAX ALLOWS FOR EASY AUTOMATION OF FREE ENERGY CALCULATIONS

```yaml
systems:
  kinase-inhibitor:
    receptor: !Combinatorial [abl, src]
    ligand: !Combinatorial [imatinib, bosutinib]
    solvent: !Combinatorial [PME, GBSA]
    pack: yes
    leap:
      parameters: [oldff/leaprc.ff99SBildn, leaprc.gaff]

  imatinib-hydration:
    solute: imatinib
    solvent1: PME
    solvent2: vacuum
    leap:
      parameters: [leaprc.ff14SB, leaprc.gaff]

  amber-system:
    phase1_path: [complex.prmtop, complex.inpcrd]        # System files for the complex phase.
    phase2_path: [solvent.prmtop, solvent.inpcrd]        # System files for the solvent phase.
    ligand_dsl: resname MOL                              # MDTraj DSL string to select the ligand.
    solvent: RF                                          # Specify how to model the solvent.

  gromacs-system:
    phase1_path: [complex.top, complex.gro]
    phase2_path: [solvent.top, solvent.gro]
    gromacs_include_dir: include/                        # Optional path to the directory containing the files
                                                         # included in .top files.
    ligand_dsl: resname MOL
    solvent: RF
```

YAML SYNTAX ALLOWS FOR EASY AUTOMATION OF FREE ENERGY CALCULATIONS

```yaml
protocols:
  absolute-binding:
    complex:
      alchemical_path:
        lambda_electrostatics: [1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0, 0.0, 0,0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
        lambda_sterics: [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0]
    solvent:
      alchemical_path:
        lambda_electrostatics: [1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0, 0.0, 0,0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
        lambda_sterics: [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0]

experiments:
  system: kinase-inhibitor
  protocol: absolute-binding
  restraint:
    type: !Combinatorial [Harmonic, FlatBottom]

  options:
    temperature: !Combinatorial [298.0*kelvin, 307.0*kelvin]
```

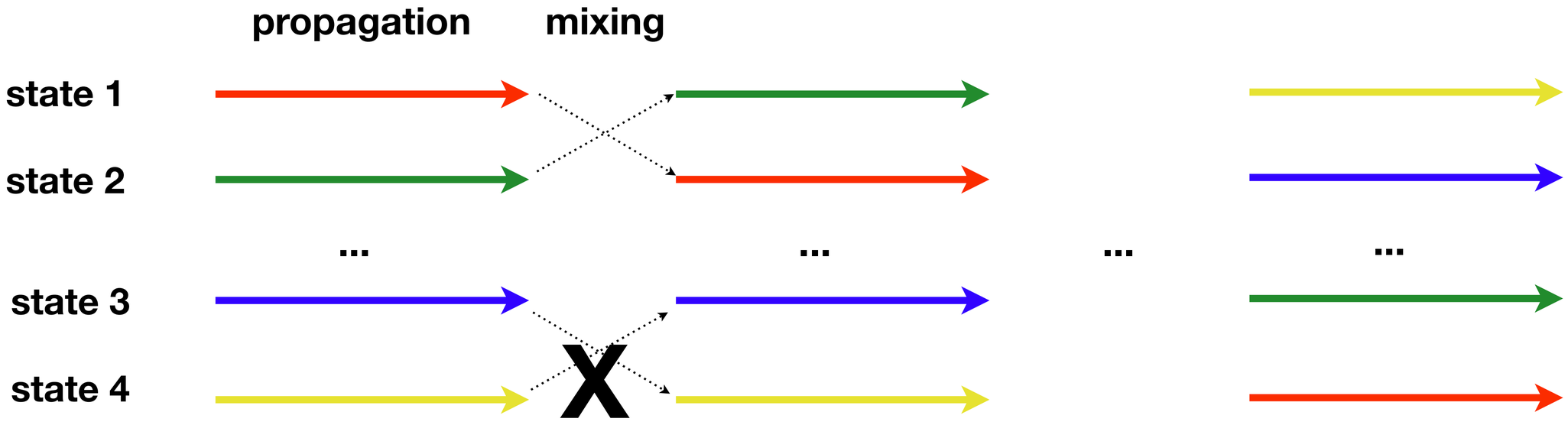
```yaml
# Optionally, apply a restrain to the ligand to keep
# it close to the receptor. Possible types are null,
# Harmonic, FlatBottom, and Boresch.
# All options that can be specified in the "options"
# section can be included here.
```

INSTALLING YANK IS EASY

```
MINICONDA=“Miniconda3-latest-Linux-x86_64.sh”
wget https://repo.continuum.io/miniconda/$MINICONDA
bash $MINICONDA -b -p $HOME/miniconda
export PATH=“$HOME/miniconda/bin:$PATH”
```

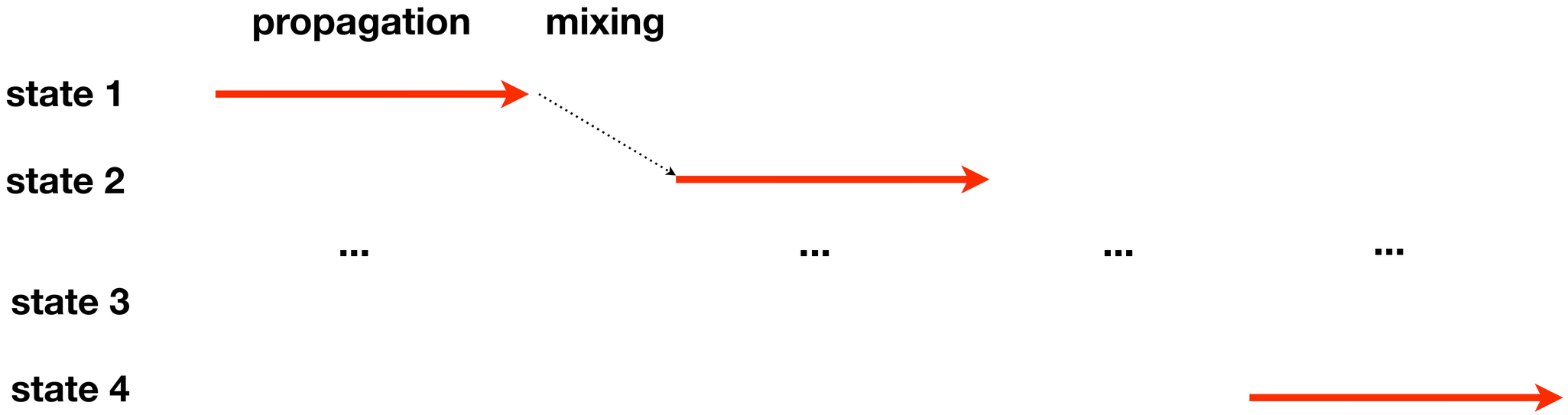
```
conda install -c omnia yank
```

DO WE NEED ALL THOSE REPLICAS?



HAMILTONIAN EXCHANGE

DO WE NEED ALL THOSE REPLICAS?



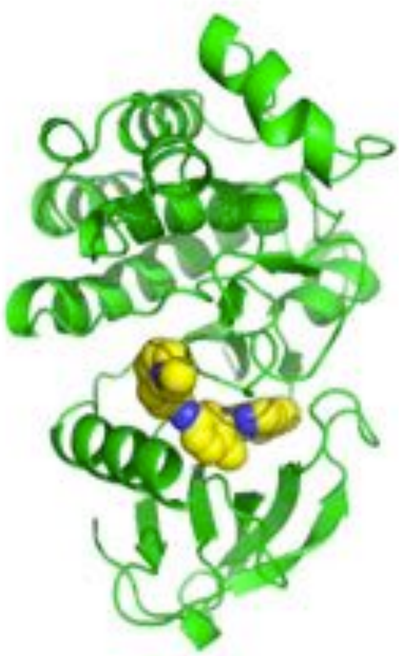
EXPANDED ENSEMBLE

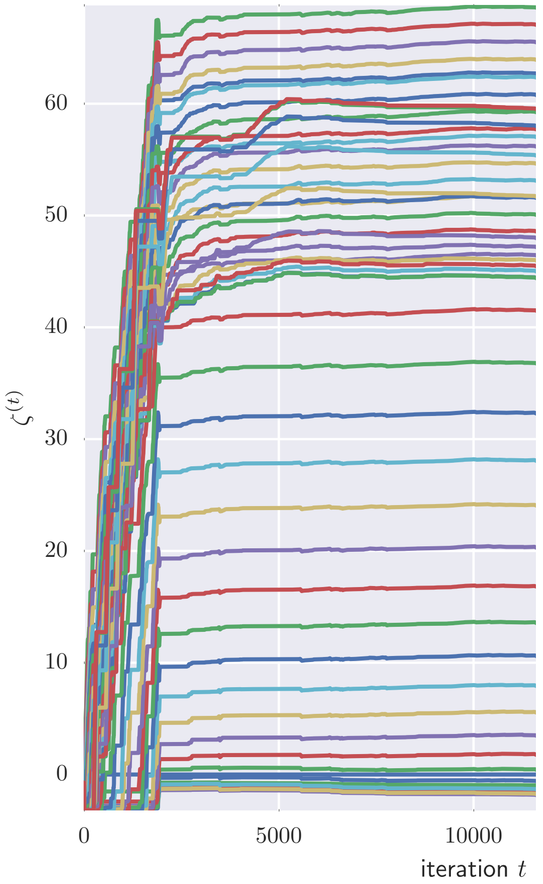
$$\pi(x,k) \propto \exp[-u_k(x) + g_k]$$

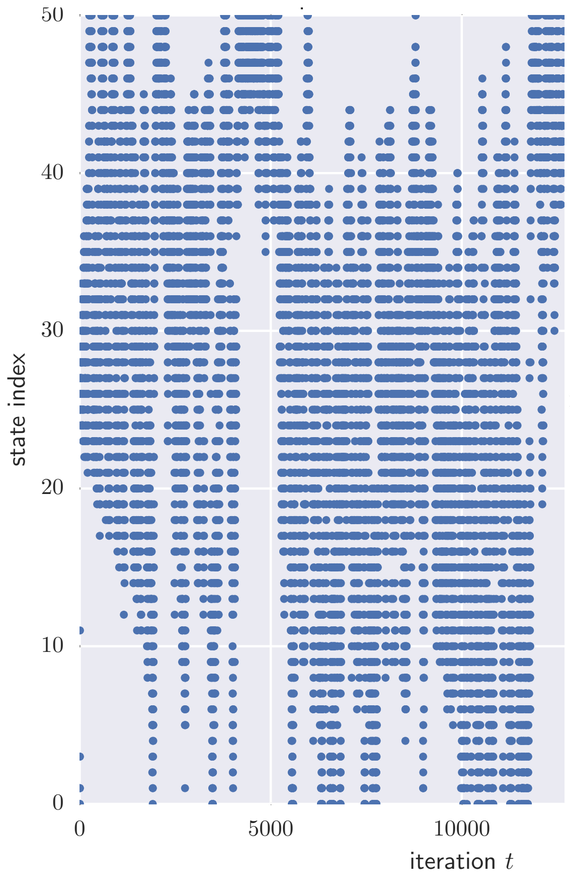
One caveat: We need to guess the weights g
(which are unfortunately the free energies we are trying to compute!)

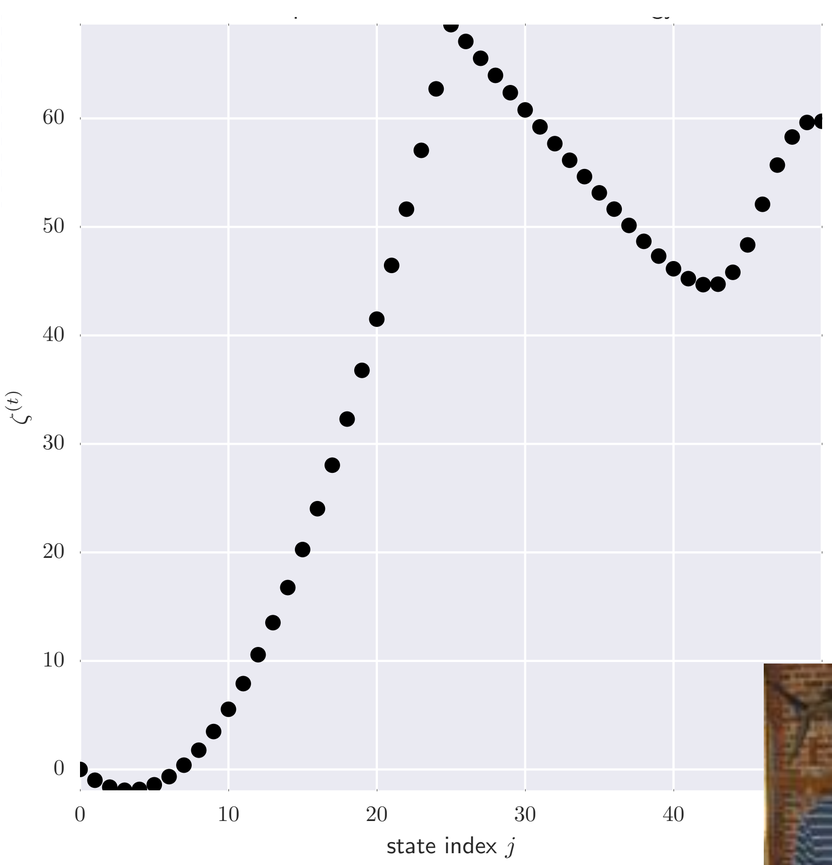
SELF-ADJUSTED MIXTURE SAMPLING (SAMS)

Provably asymptotically optimal algorithm for estimating free energies from a single simulation!
Tan Z. J. Comp. Graph. Stat. http://dx.doi.org/10.1080/10618600.2015.1113975









ZHIQIANG TAN
Rutgers

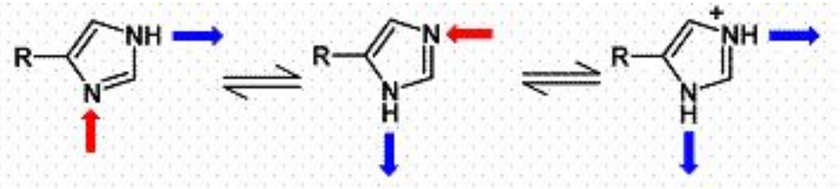
Similar to simulated scaling (Wei Yang), but provably optimal.

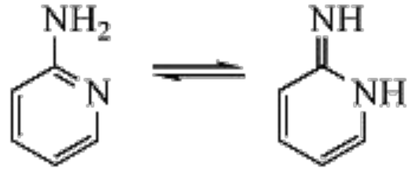
PREDICTIONS FAIL FOR THREE REASONS

1. The forcefield does a poor job of modeling the physics of our system

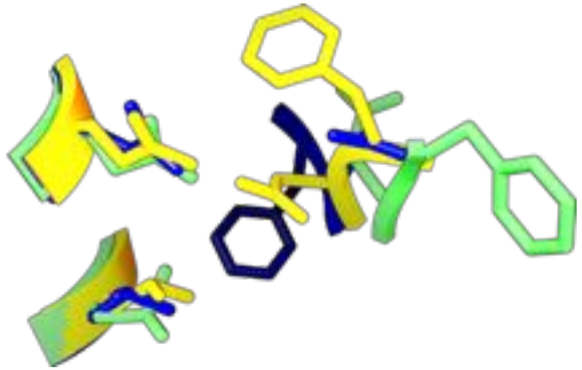
$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

2. We're missing some essential chemical in our simulations
(e.g. protonation states, tautomers, covalent association)





3. We haven't sampled all of the relevant conformations



PREDICTIONS FAIL FOR THREE REASONS

1. The forcefield does a poor job of modeling the physics of our system

$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

THE OPEN FORCEFIELD GROUP

https://github.com/open-forcefield-group

DAVID MOBLEY
UCI

MICHAEL GILSON
UCSD

MICHAEL SHIRTS
UNIVERSITY OF COLORADO, BOULDER

CHRISTOPHER BAYLY
OPENEYE SCIENTIFIC

JOHN CHODERA
SKI/MSKCC

DIRECT CHEMICAL PERCEPTION ESCAPES THE SHACKLES OF ATOM TYPE CENTRIC FORCEFIELDS

```
<NonbondedForce coulomb14scale="0.833333" lj14scale="0.5" sigma_unit="angstroms" epsilon_unit="kilocalories_per_mole">
   <Atom smirks="[#1:1]" rmin_half="1.4870" epsilon="0.0157"/>
   <Atom smirks="[#1:1]-[#6]" rmin_half="1.4870" epsilon="0.0157"/>
   ...
</NonbondedForce>

<HarmonicBondForce length_unit="angstroms" k_unit="kilocalories_per_mole/angstrom**2">
   <Bond smirks="[#6X4:1]-[#6X4:2]" length="1.526" k="620.0"/>
   <Bond smirks="[#6X4:1]-[#1:2]" length="1.090" k="680.0"/>
...
</HarmonicBondForce>

<HarmonicAngleForce angle_unit="degrees" k_unit="kilocalories_per_mole/radian**2">
   <Angle smirks="[a,A:1]-[#6X4:2]-[a,A:3]" angle="109.50" k="100.0"/>
   <Angle smirks="[#1:1]-[#6X4:2]-[#1:3]" angle="109.50" k="70.0"/>
</HarmonicAngleForce>

<BondChargeCorrections method="AM1" increment_unit="elementary_charge">
  <BondChargeCorrection smirks="[#6X4:1]-[#6X3a:2]" increment="+0.0073"/>
  <BondChargeCorrection smirks="[#6X4:1]-[#6X3a:2]-[#7]" increment="-0.0943"/>
  <BondChargeCorrection smirks="[#6X4:1]-[#8:2]" increment="+0.0718"/>
</BondChargeCorrections>
```

SMARTS IS LIKE REGULAR EXPRESSIONS FOR SMILES

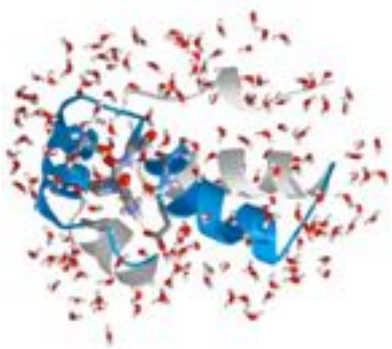
CHECK OUT OUR 332-LINE SMALL MOLECULE FORCEFIELD!

implementation: http://github.com/open-forcefield-group/openforcefield
forcefield: http://github.com/open-forcefield-group/smirnoff99Frosst

EXAMPLE: TIP3P WATER

```
<?xml version='1.0' encoding='ASCII'?>
<SMIRNOFF version="0.1">
  <!-- SMIRks Native Open Force Field (SMIRNOFF) file -->
  <Date>2017-04-29</Date>
  <Author>J. D. Chodera, MSKCC; A. Rizzi, Weill Cornell; C. C. Bannan, UC Irvine</Author>
  <!-- SMIRNOFF file implementing TIP3P water model. -->
  <NonbondedForce coulomb14scale="0.833333" lj14scale="0.5" sigma_unit="nanometers" epsilon_unit=
      "kilojoules_per_mole">
    <!-- TIP3P water oxygen with charge override -->
    <Atom smirks="[#1]-[#8X2H2+0:1]-[#1]" id="n1" sigma="0.31507524065751241" epsilon="0.635968"
        charge="-0.834"/>
    <!-- TIP3P water hydrogen with charge override -->
    <Atom smirks="[#1:1]-[#8X2H2+0]-[#1]" id="n2" sigma="1" epsilon="0" charge="0.417"/>
  </NonbondedForce>
  <Constraints distance_unit="angstroms">
    <!-- constrain water O-H bond to equilibrium bond length (overrides earlier constraint) -->
    <Constraint smirks="[#1:1]-[#8X2H2+0:2]-[#1]" id="c1" distance="0.9572"/>
    <!-- constrain water H...H, calculating equilibrium length from H-O-H equilibrium angle and H-O equilibrium
        bond lengths -->
    <Constraint smirks="[#1:1]-[#8X2H2+0]-[#1:2]" id="c2" distance="1.5139006545247014"/>
  </Constraints>
</SMIRNOFF>
```

THE OLD WAY

θ^* ⟶ **MD** ⟶  ⟶ $\langle A \rangle_*$

One set of parameters in, one computed result out

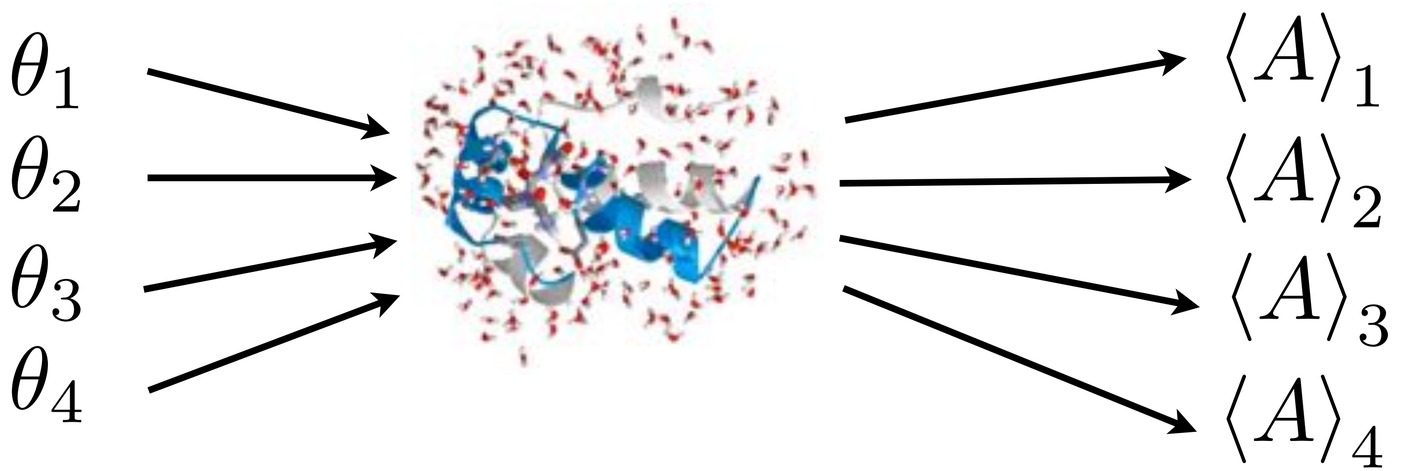
THE BAYESIAN WAY

Bayes rule provides a probability measure over unknown parameters given data and an automated way to update parameters given new experimental data

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

$\mathcal{D}$	data
θ	forcefield
$p(\theta\|\mathcal{D})$	posterior
$p(\mathcal{D}\|\theta)$	data model
$p(\theta)$	prior on forcefield parameters

THE BAYESIAN WAY



Multiple parameter sets in, multiple estimates out

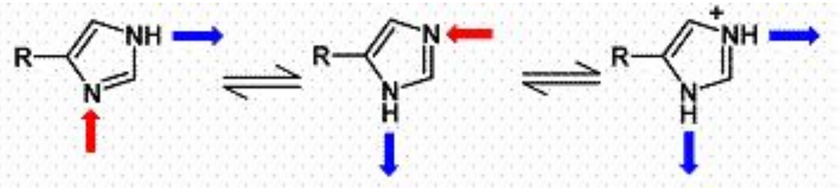
W can estimate both statistical and systematic components of computed results

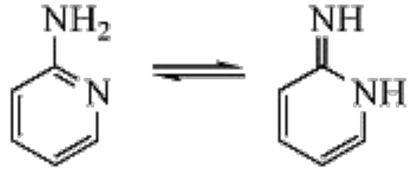
PREDICTIONS FAIL FOR THREE REASONS

1. The forcefield does a poor job of modeling the physics of our system

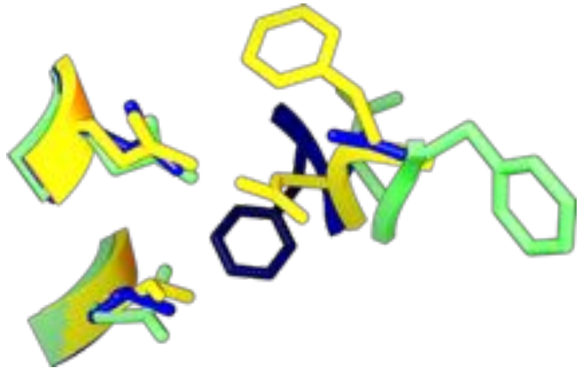
$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

2. We're missing some essential chemical in our simulations
(e.g. protonation states, tautomers, covalent association)



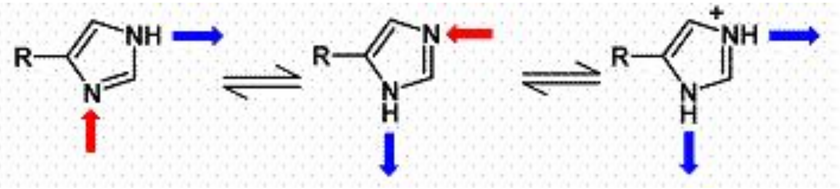


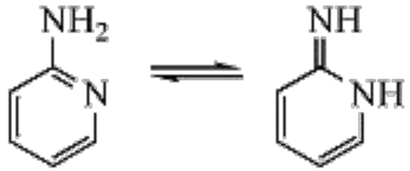
3. We haven't sampled all of the relevant conformations



PREDICTIONS FAIL FOR THREE REASONS

2. We're missing some essential chemical in our simulations
(e.g. protonation states, tautomers, covalent association)

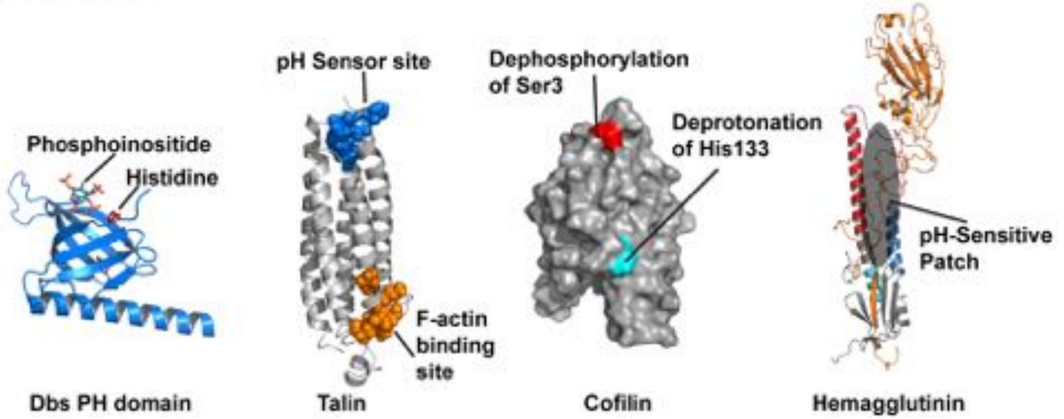




PROTONATION STATE EFFECTS ARE UBIQUITOUS IN BIOLOGY

proteins

ionizable residues constitute 29% of protein residues

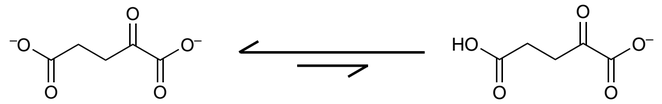


Diane Barber // Considering protonation as a post-translational modification regulating protein structure and function. Annu. Rev. Biophys. 42:289, 2013.

---

metabolites

often possess ionizable groups for solubility



αKG deprotonated tail αKG protonated tail

$\Delta G_{pH\ 7.4} = 7.2\ K_B T$

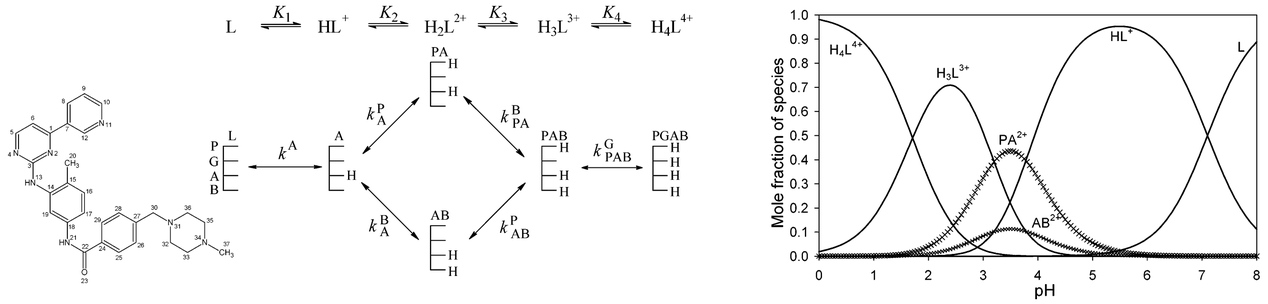
$\Delta G_{pH\ 6.0} = 4.0\ K_B T$

Craig Thompson // L-2-hydroxyglutarate production arises from noncanonical enzyme function at acidic pH. Nat. Chem. Biol. 13:494, 2017.

---

drugs

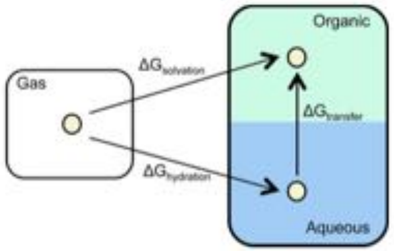
frequently contain many titratable sites
>25% of approved drugs have more than one tautomer

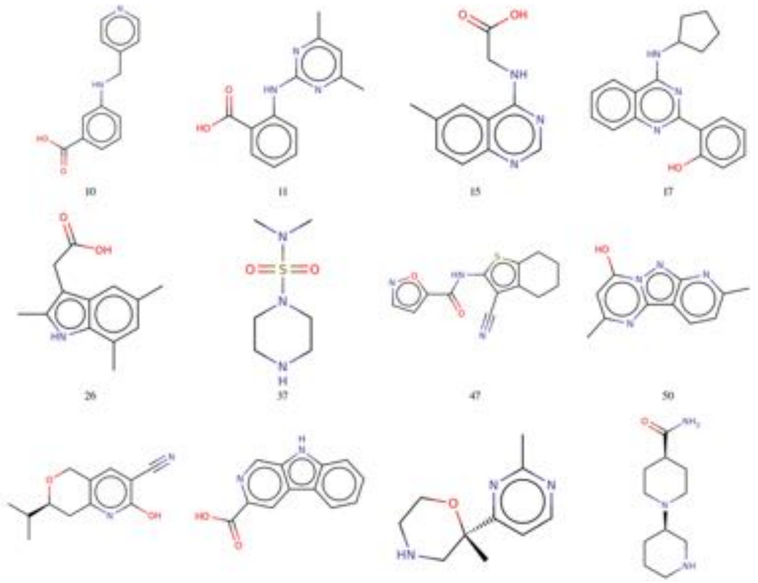


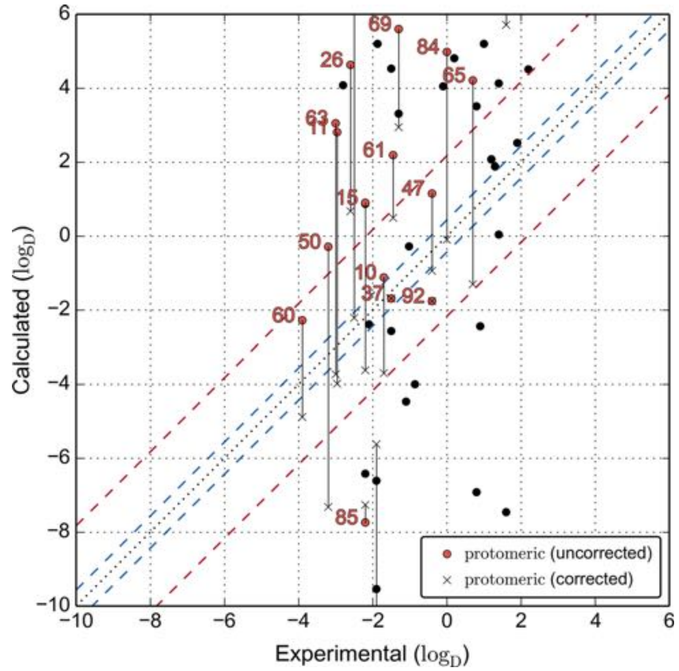
Béla Noszál // Acid-base profiling of imatinib (Gleevec) and its fragments. J. Med. Chem. 48:249. 2005.

NEGLECT OF PROTONATION STATE EFFECTS LEADS TO LARGE ERRORS IN TRANSFER FREE ENERGIES

Blind challenge eliminates slow protein sampling issue to test accuracy of small molecule transfer energetics modeling



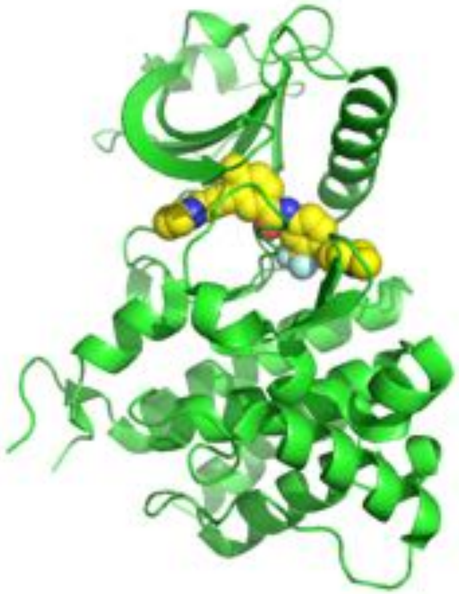




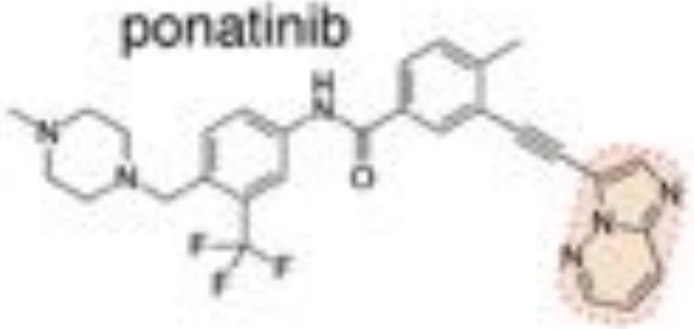
Protonation state effects contribute up to 10 kcal/mol for some compounds
Including protonation state effects reduced RMSE by 3.5 kcal/mol

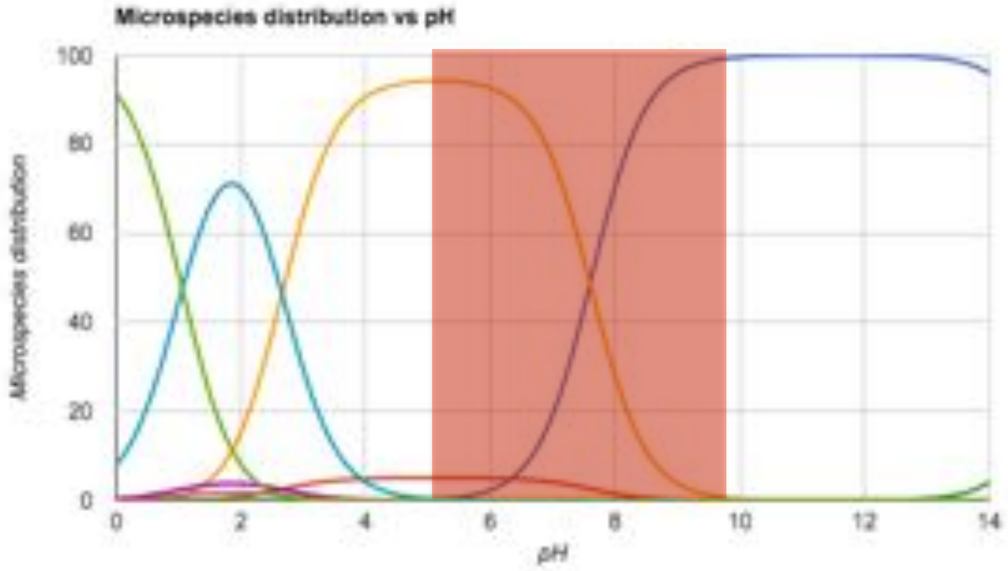
David Mobley // Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. J. Chem. Theor. Comput. 12:4015, 2016.
Bernie Brooks // Blind prediction of distribution in the SAMPL5 challenge with QM based promoter and pKa corrections. J. Comput. Aided Mol. Des. 30:1087, 2016.

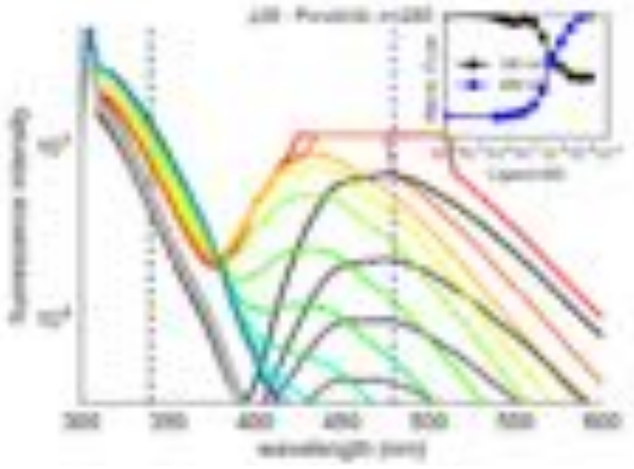
KINASE:INHIBITOR BINDING CAN BE PH-SENSITIVE



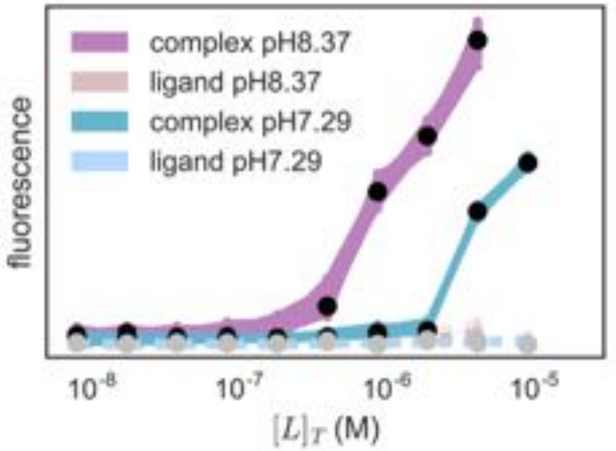
ponatinib:DDR1
(pdbid:3ZOS)

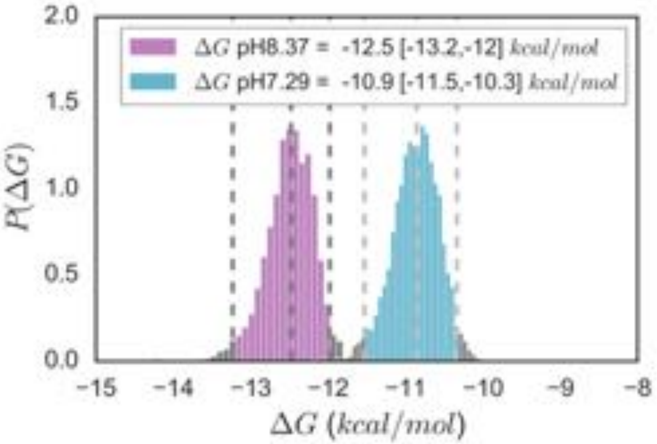






Change pH 7.3 → 8.4





ΔΔG = 1.6 kcal/mol

SONYA HANSON

CONTINUUM ELECTROSTATICS PREDICTIONS SUGGEST PROTONATION STATE EFFECTS ARE PERVASIVE

pdbid	inhibitor	kinase	Δ protein	Δ inhibitor	Δ protomer
3UE4	Bosutinib	ABL	0	0.5	YES
2GQG	Dasatinib	ABL	-0.1	0.6	YES
4XEY	Dasatinib	ABL	0.12	0.82	YES
2HYY	Imatinib	ABL	-0.2	-0.01	NO
3PYY	Imatinib	ABL	-0.28	0.01	NO
3CS9	Nilotinib	ABL	0.1	0.06	NO
3OXZ	Ponatinib	ABL	-0.6	0.02	NO
3IK3	Ponatinib	ABL T315I	-0.63	0.06	NO
3AOX	Alectinib	ALK	0	0.13	NO
4MKC	Ceritinib	ALK	0.7	0	NO
2XP2	Crizotinib	ALK	-0.04	-0.77	YES
4ANQ	Crizotinib	ALK G1269A	-0.1	-0.76	YES
2YFX	Crizotinib	ALK L1196M	-0.1	-0.77	YES
4ANS	Crizotinib	ALK L1196M/G1269A	-0.1	-0.77	YES
4XV2	Dabrafenib	BRAF	0.92	-0.31	NO
5CSW	Dabrafenib	BRAF	0.4	0.65	YES
5HIE	Dabrafenib	BRAF	1	-0.25	NO
2EUF	Palbociclib	CDK6	-0.08	-0.28	NO
3ZOS	Ponatinib	DDR1	-1.5	-0.23	NO
4GSJ	Afatinib	EGFR	-0.18	-0.98	YES
1M17	Erlotinib	EGFR	0.2	0	NO
4WKQ	Gefitinib	EGFR	0	0.65	YES
1XKK	Lapatinib	EGFR	-0.26	-0.54	YES
4ZAU	Osimertinib	EGFR	-0.3	0.02	NO
2ITY	Gefitinib	EGFR	0.08	-0.04	NO
2ITZ	Gefitinib	EGFR	0	0.09	NO

pdbid	inhibitor	kinase	Δ protein	Δ inhibitor	Δ protomer
4HJO	Erlotinib	EGFR (inactive)	0	0	YES
2ITO	Gefitinib	EGFR G719S	0.1	0.48	YES
3UG2	Gefitinib	EGFR G719S/T790M	0.2	0.13	NO
4G5P	Afatinib	EGFR T790M	-0.5	-0.01	NO
4I22	Gefitinib	EGFR T790M/L858R	0.2	-0.18	NO
4V01	Ponatinib	FGFR1	-2.6	0.06	YES
4V04	Ponatinib	FGFR1	-1.2	0.05	YES
4QRC	Ponatinib	FGFR4	-1	0.02	YES
4TYJ	Ponatinib	FGFR4	-0.3	0.03	NO
4UXQ	Ponatinib	FGFR4	-0.36	0.02	NO
3LXK	Tofacitinib	JAK3	-0.02	-0.07	NO
4U0I	Ponatinib	KIT	-0.1	0.03	NO
4AN2	Cobimetinib	MEK1	0	0.01	NO
4LMN	Cobimetinib	MEK1	0	0	NO
2WGJ	Crizotinib	MET	-0.06	-1.05	YES
4AG8	Axitinib	VEGFR2	0.12	0	NO
4AGC	Axitinib	VEGFR2	0	0	NO
3WZD	Lenvatinib	VEGFR2	0.18	0	NO
3CJG	Pazopanib	VEGFR2	0.34	-0.02	NO
2QU5	Regorafenib	VEGFR2	0.5	-0.07	YES
3WZE	Sorafenib	VEGFR2	0.1	-0.01	NO
4ASD	Sorafenib	VEGFR2	0.2	-0.01	NO
4AGD	Sunitinib	VEGFR2	0.32	-0.99	YES

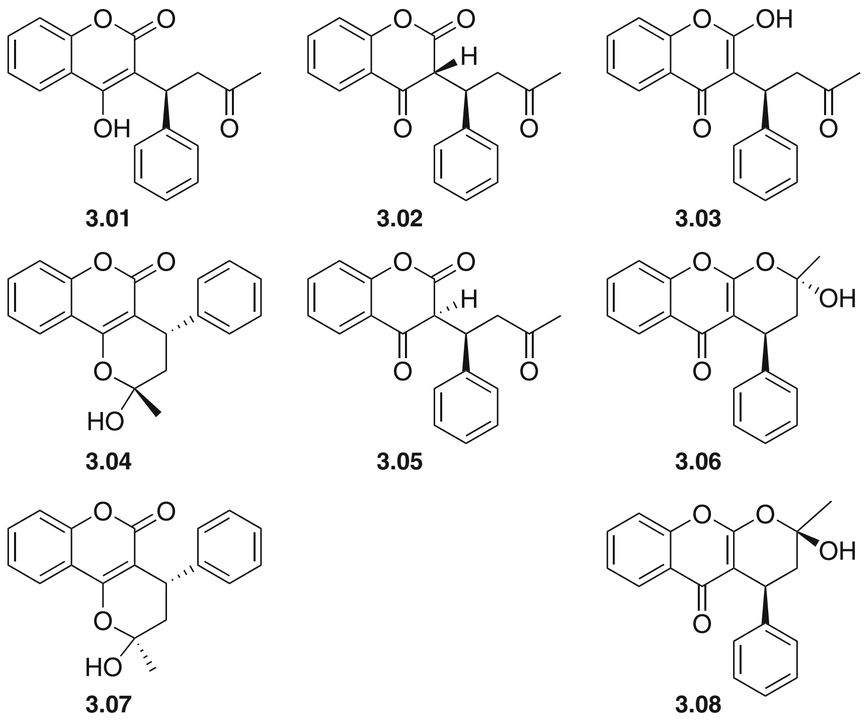
proton gain

proton loss

tautomer shift

MARILYN GUNNER
SALAH SALAH

LET'S NOT FORGET TAUTOMERS



tautomers of **warfarin**

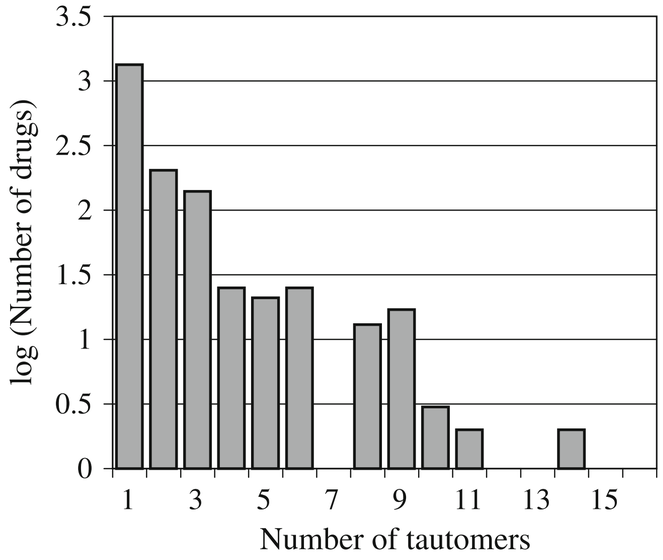


Fig. 13 The frequency distribution of tautomers of marketed drugs

MORE THAN HALF OF ALL DRUGS HAVE 2 OR MORE TAUTOMERS

Martin Y. JCAMD 23:693, 2009. Let's not forget tautomers.

WE CAN SIMULATE AN EXPANDED ENSEMBLE WHERE PROTONATION STATES ARE DYNAMIC

Define the reduced potential for a state k as a combination of terms

$$u_k(\mathbf{x}) = \beta_k[U_k(\mathbf{x}) + p_k V(\mathbf{x}) + \mu_k^{\mathrm{T}}\mathbf{N}(\mathbf{x})]$$

μ_k **chemical potential of exchangeable species**

$\mathbf{N}(\mathbf{x})$ **number of each chemical species in system** protons

The target configuration space density is given by

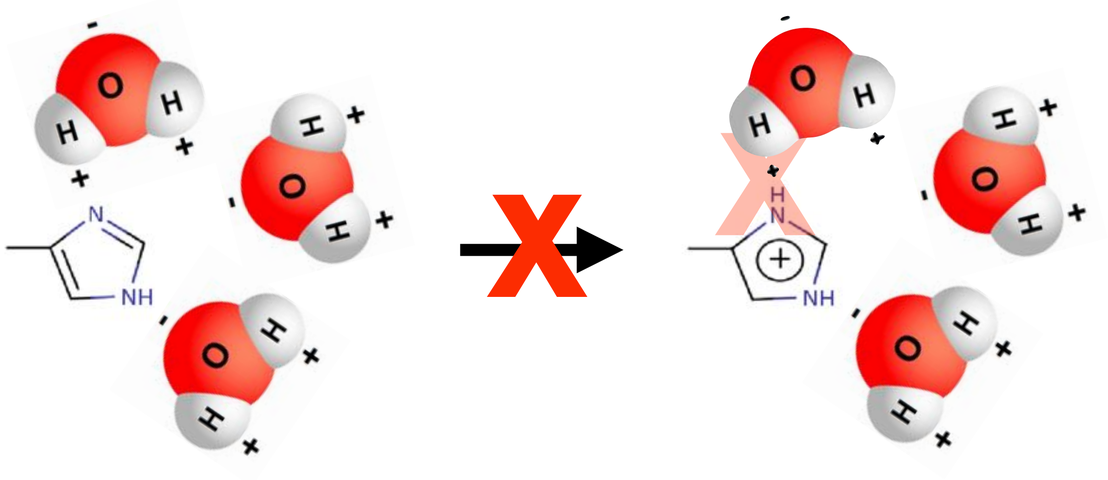
$$\pi_k(\mathbf{x}) = Z_k^{-1}\ \exp[-u_k(\mathbf{x})] \qquad Z_k = \int d\mathbf{x}\ \exp[-u_k(\mathbf{x})]$$

We can sample from this ensemble using Gibbs sampling

$\mathbf{x}_{n+1} \sim \pi(\mathbf{x}|\mathbf{N}_n)$ update positions with MD/MC at fixed protonation state

$N_{n+1} \sim \pi(\mathbf{N}|\mathbf{x}_{n+1})$ update protonation states with MC at fixed position

PROBLEM: ACCEPTANCE RATES FOR PROTONATION STATE CHANGES ARE ESSENTIALLY ZERO IN EXPLICIT SOLVENT



INSTANTANEOUS CHANGES IN PROTONATION STATE ARE HIGHLY UNFAVORABLE DUE TO SOLVATION SHELL

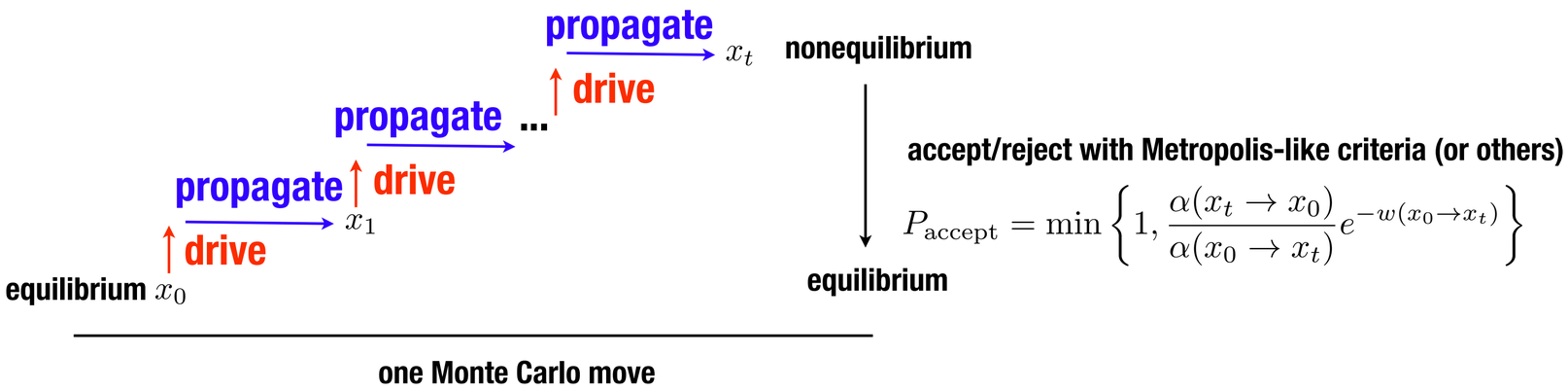
NONEQUILIBRIUM CANDIDATE MONTE CARLO CAN ASTRONOMICALLY BOOST MC ACCEPTANCE RATES

Algorithm:

Drive some degrees of freedom or thermodynamic parameters in small steps, accumulating work

Propagate others using Metropolis MC or molecular dynamics

Accept or reject final configuration with Metropolis-like criterion



Follows earlier ideas by Manuel Athenes (work-bias Monte Carlo), Harry Stern (constant pH simulation), Chris Jarzynski (switching replica temperatures), and Jerome Nilmeier (approximate).

Nilmeier, Crooks, Minh, Chodera. PNAS 108:E1009, 2011.

THE ACCEPTANCE CRITERIA CAN BE GENERALIZED TO ALLOW ARBITRARY PROPAGATION KERNELS

$x_t \in \mathcal{X}$ **configurations in some state space**

$\alpha_t(x, y)$ **perturbation kernel** $\sum_y \alpha_t(x, y) = 1$

$K_t(x, y)$ **propagation kernel** $\sum_x \pi_t(x) K_t(x, y) = \pi_t(y)$

$\Lambda \equiv \{\alpha_0, K_0, \ldots, \alpha_T, K_T\}$ **switching protocol**

$X \equiv \{x_0, x_1^*, x_1, \ldots, x_T\}$ **trajectory**

$\tilde{x}$ **is momentum-reversed** x

$\alpha_t(x, y) > 0 \Leftrightarrow \alpha_t(y, x) > 0$
$K_t(x, y) > 0 \Leftrightarrow K_t(y, x) > 0$

$\tilde{\Lambda}$ **is time-reversed** Λ

$\tilde{X}$ **is time-reversed** X

forward process $x_0 \xrightarrow{\alpha_1} x_1^* \xrightarrow{K_1} x_1 \longrightarrow \cdots \longrightarrow x_{T-1} \xrightarrow{\alpha_T} x_T^* \xrightarrow{K_T} x_T.$

reverse process $\tilde{x}_T \xrightarrow{K_T} \tilde{x}_T^* \xrightarrow{\alpha_T} \tilde{x}_{T-1} \longrightarrow \cdots \longrightarrow \tilde{x}_1 \xrightarrow{K_1} \tilde{x}_1^* \xrightarrow{\alpha_1} \tilde{x}_0.$

$P(\tilde{\Lambda}|\tilde{x}_T) > 0 \Leftrightarrow P(\Lambda|x_0 > 0)$

both must be selected with nonzero probability!

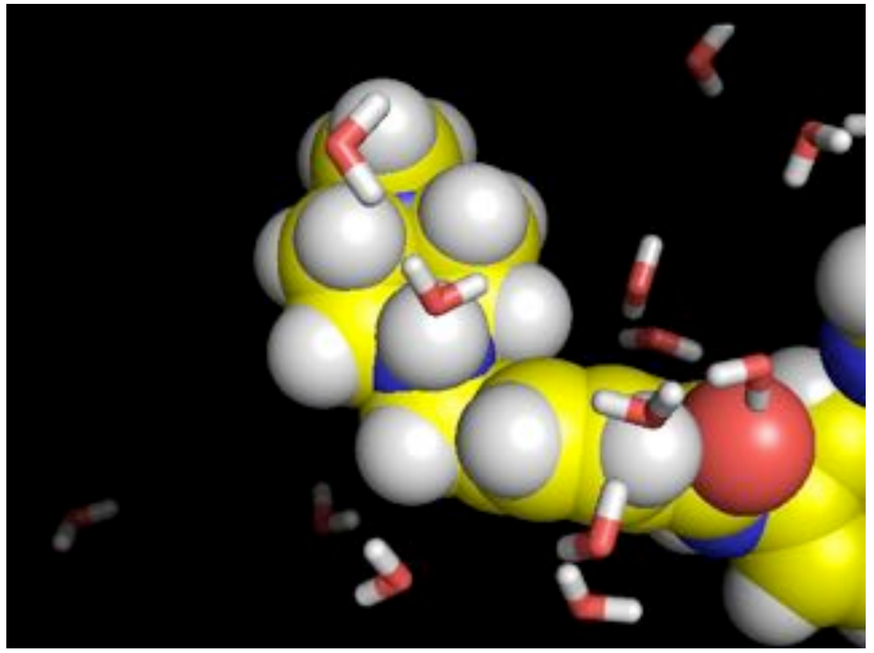
Enforce strict "pathwise" form of detailed balance (which also ensures detailed balance is satisfied):

$$\underbrace{A(X|\Lambda)}_{\text{acceptance probability}} \underbrace{\Pi(X|x_0, \Lambda)}_{\text{path generation probability}} \underbrace{P(\Lambda|x_0, \lambda_0)}_{\text{protocol selection}} \underbrace{\pi(x_0, \lambda_0)}_{\text{equilibrium}} = \underbrace{A(\tilde{X}|\tilde{\Lambda})}_{\text{acceptance probability}} \underbrace{\Pi(\tilde{X}|\tilde{x}_T, \tilde{\Lambda})}_{\text{path generation probability}} \underbrace{P(\tilde{\Lambda}|\tilde{x}_T, \lambda_T)}_{\text{protocol selection}} \underbrace{\pi(\tilde{x}_T, \lambda_T)}_{\text{equilibrium}}$$

Result is a general acceptance criteria for any nonequilibrium perturbation:

$$A(X|\Lambda) = \min\left\{1, \underbrace{\frac{\pi(x_T, \lambda_T)}{\pi(x_0, \lambda_0)}}_{\text{equilibrium}} \underbrace{\frac{P(\tilde{\Lambda}|\tilde{x}_T, \lambda_T)}{P(\Lambda|x_0, \lambda_0)}}_{\text{selection}} \underbrace{\frac{\tilde{\alpha}(\tilde{X})}{\alpha(X)}}_{\text{perturbation}} \underbrace{e^{-\Delta S(X)}}_{\text{action}}\right\}$$

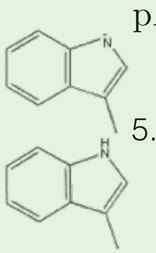
ANNIHILATING A PROTON IN EXPLICIT SOLVENT OVER 20 PS NCMC SWITCHING TRAJECTORY



BAS RUSTENBURG

DYNAMIC LIGAND PROTONATION STATES REQUIRE AN EXPLICIT CALIBRATION STEP

amino acids

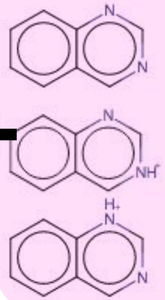


pK_a | $\Delta G_{aq}^{\mathrm{pH\ 7.4}}$

5.88 | 0.0 kT

3.5 kT

small molecules



$\Delta G_{aq}^{\mathrm{pH\ 7.4}}$

0.0 kT

2.3 kT

4.7 kT

SAMS calibration of protomer log weights

$$\mathbf{x}_{t+1} \sim \pi(\mathbf{x}|\mathbf{s}_t, \mathbf{g}_t^o, \beta, p)$$
$$\mathbf{s}_{t+1} \sim \pi(\mathbf{s}|\mathbf{x}_{t+1}, \mathbf{g}_t^o, \beta, p)$$
$$\mathbf{g}_{t+1}^o = f(\mathbf{s}_{t+1}, \mathbf{g}_t^o)$$

$\mathbf{x}$ positions
$\mathbf{s}$ protomer states
$\mathbf{g}^o$ uncorrected log weights

pH correction of protomer log weights

$$g_s(\mathrm{pH}) = g_s^o + \ln \frac{\pi_s(\mathrm{pH})}{\pi_s^o}$$

$\mathbf{g}(\mathrm{pH})$ pH-corrected log weight
$\pi_s(\mathrm{pH}) \propto e^{-\beta \Delta G_{aq}^{\mathrm{pH}}}$
π_s^o protomer population in calibration simulation

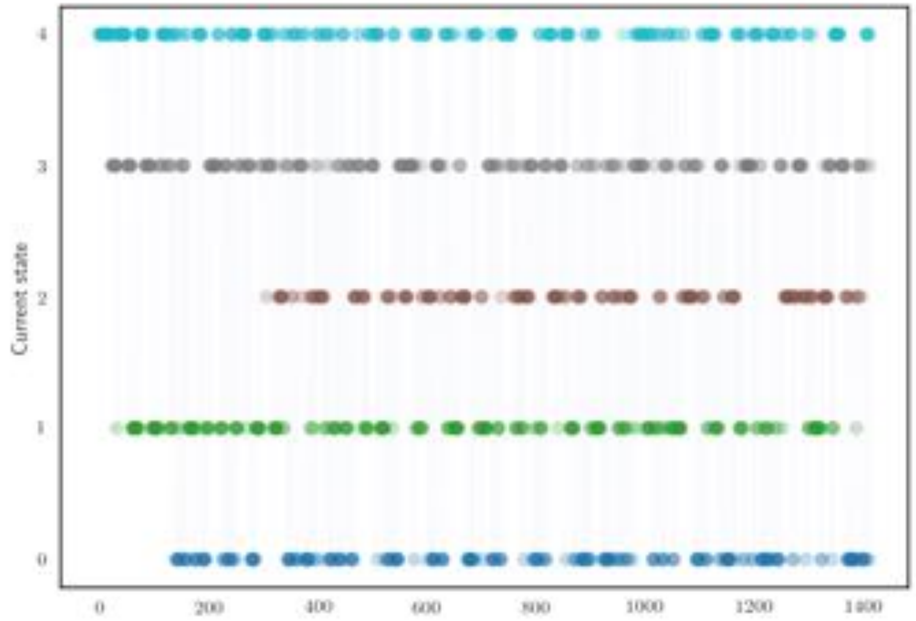
MCMC protonation state sampling

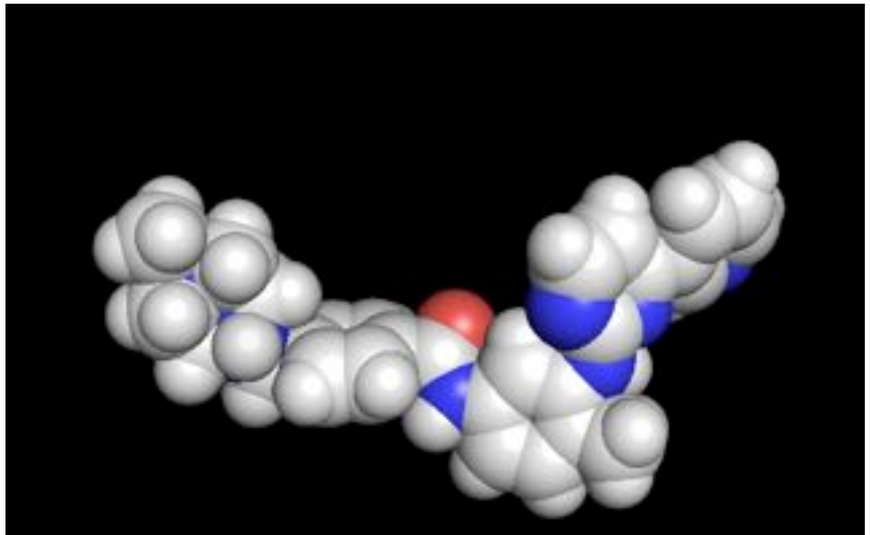
$$\mathbf{x}_{t+1} \sim \pi(\mathbf{x}|\mathbf{s}_t, \mathbf{g}(\mathrm{pH}), \beta, p)$$ **MD / GHMC**
$$\mathbf{s}_{t+1} \sim \pi(\mathbf{s}|\mathbf{x}_{t+1}, \mathbf{g}(\mathrm{pH}), \beta, p)$$ **NCMC**

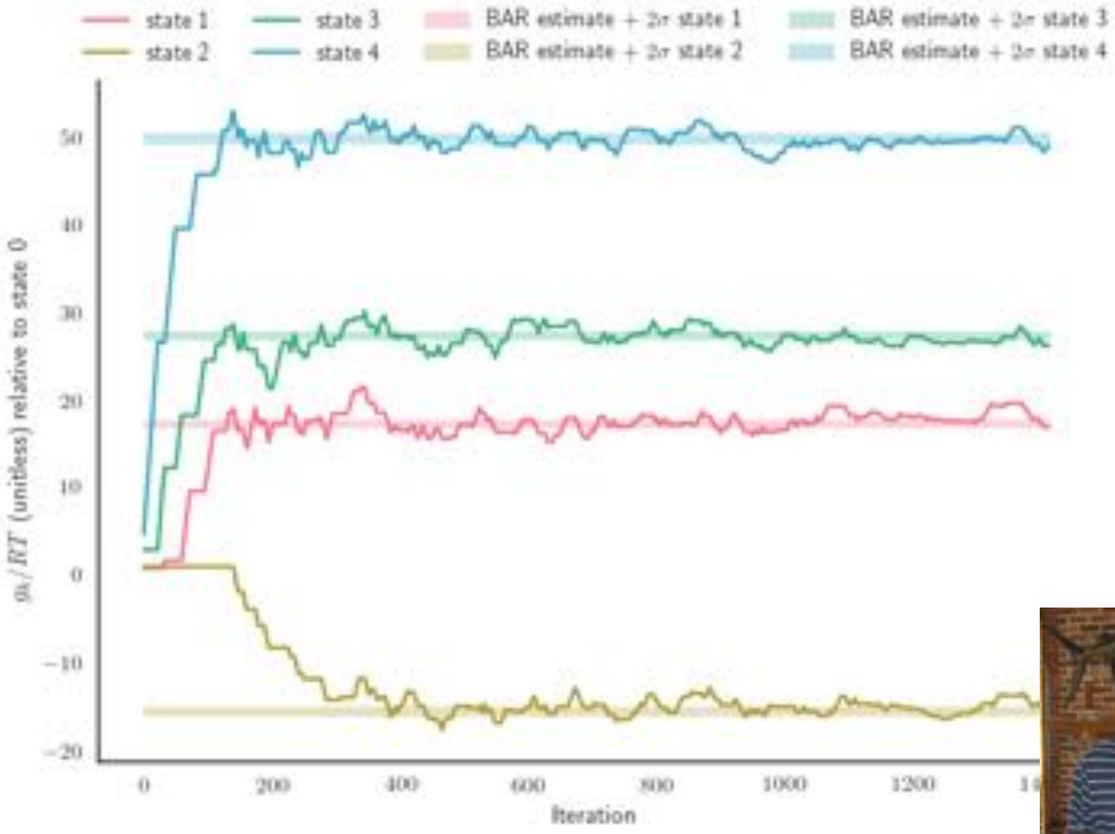
$$\pi(\mathbf{x}, \mathbf{s}) \propto e^{-\beta[U(\mathbf{x};\mathbf{s})+pV(\mathbf{x})]+\mathbf{g_s}(\mathrm{pH})}$$

BAS RUSTENBURG

CALIBRATION USES SAMS + BAR TO ESTIMATE RELATIVE PROTONATION STATE FREE ENERGIES

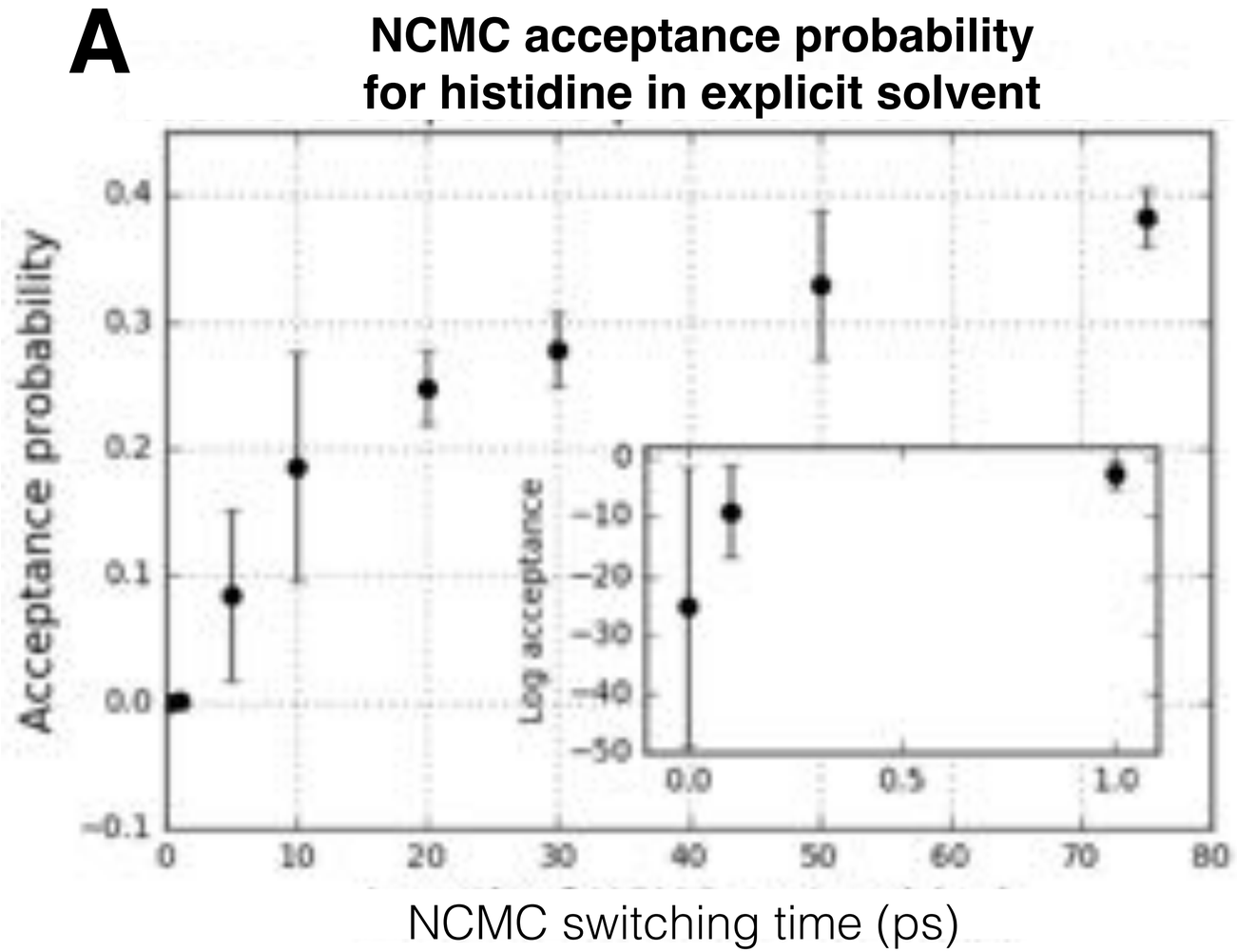






ZHIQIANG TAN
Rutgers

NCMC GIVES USEFUL ACCEPTANCE RATES



BAS RUSTENBURG

PROTON-DRIVE

Protons

Protonation states and tautomers for OpenMM

Watch

Navigation

Setting up a constant-pH MD simulation
Advanced calibration options
The ligutils submodule

Protons: Protonation states and tautomers for OpenMM

> **Note:**
>
> This is module is undergoing heavy development. None of the API calls are final.

Introduction

This python module implements a constant-pH MD scheme for sampling protonation states and tautomers of amino acids and small molecules in Op

Installation

Use the command

```
python setup.py install
```

to install the package. The installation does not automatically check for requirements.

To test the installation, run

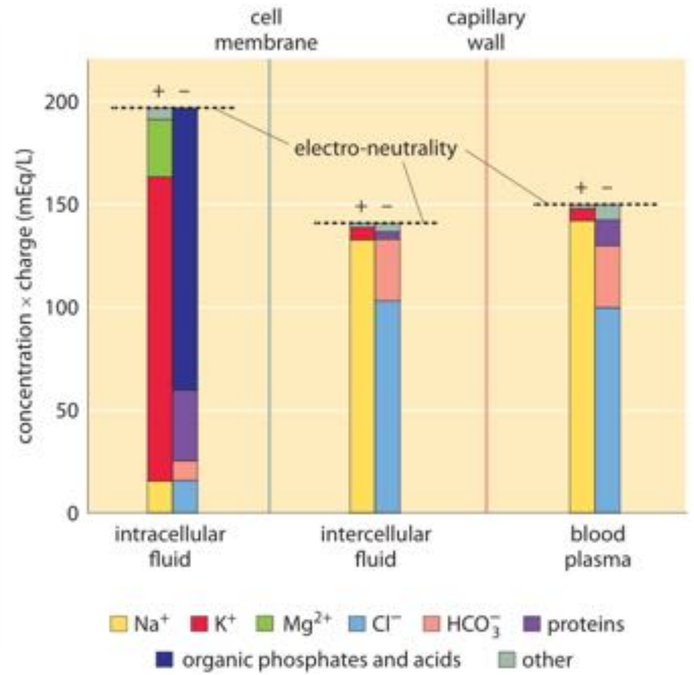
```
nosetests protons
```

Requirements

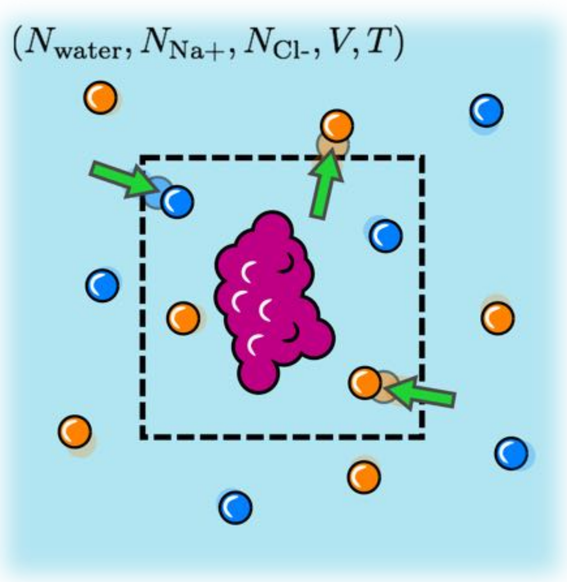
BAS RUSTENBURG

https://github.com/choderalab/protons

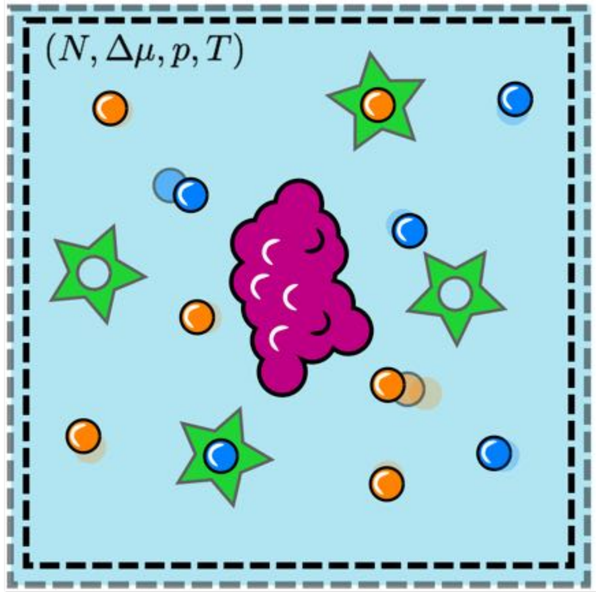
AN NCMC OSMOSTAT CAN MODEL REALISTIC SALT CONCENTRATION FLUCTUATIONS



real system



semigrand ensemble



semigrand ensemble

$$\pi(x, \theta; \Delta\mu, N, p, T) = \frac{1}{\Xi(\Delta\mu, N, p, T)} e^{-\beta[U(x,\theta)+pV(x)+n(\theta)\Delta\mu]}$$

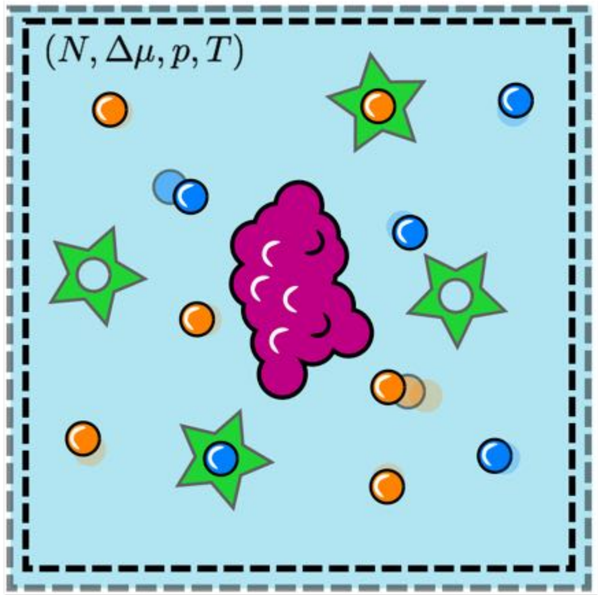
number of ions

difference in chemical potential between salt and water

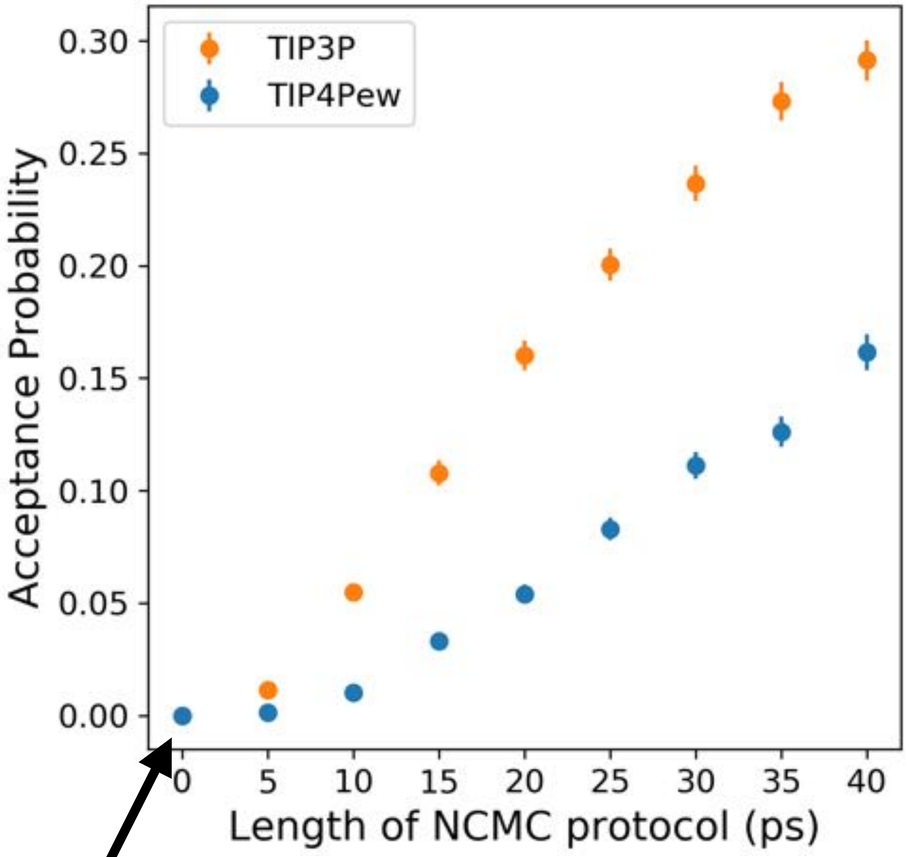
$$\Xi(\Delta\mu, N, p, T) = \sum_{\theta, \text{ subject to } \sum_i^N \theta_i = -z} \int \mathrm{d}x\, \pi(x, \theta, \Delta\mu, N, p, T),$$

AN NCMC OSMOSTAT CAN MODEL REALISTIC SALT CONCENTRATION FLUCTUATIONS

semigrand ensemble



NCMC acceptance probability for water-salt exchange

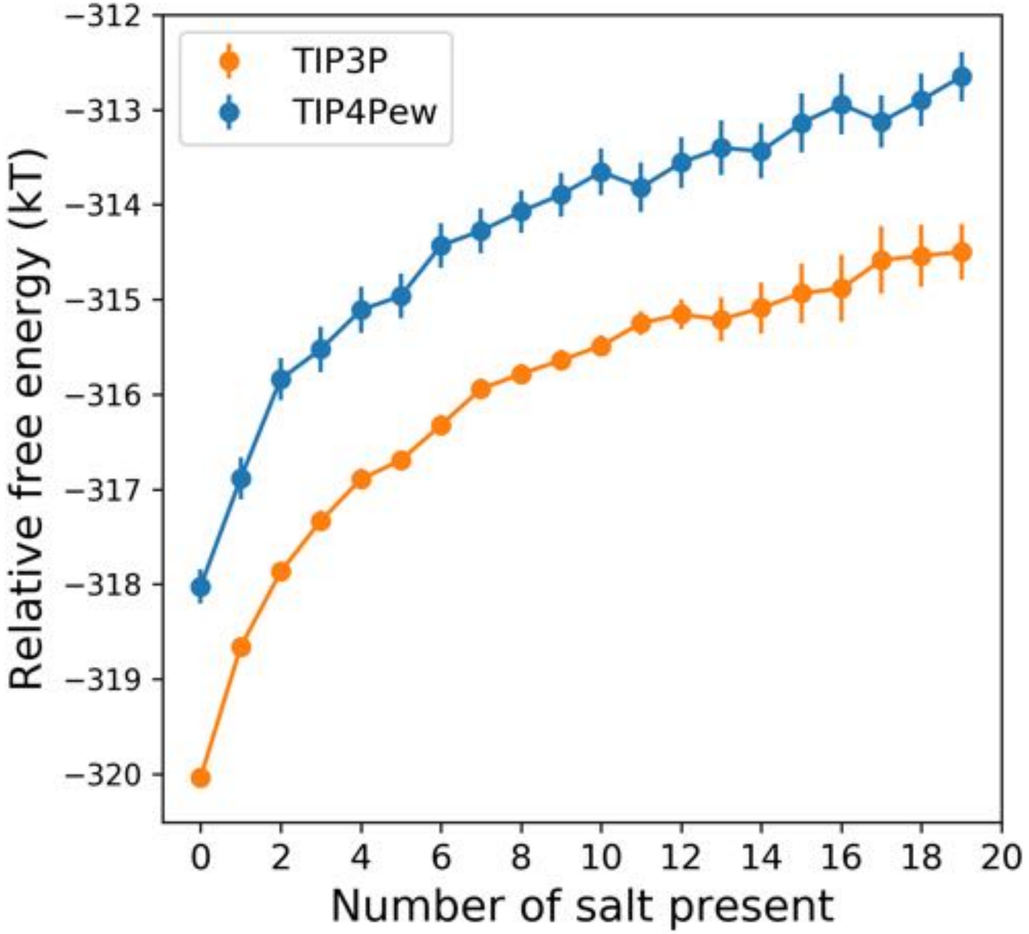


10^{-22} if instantaneous MC used

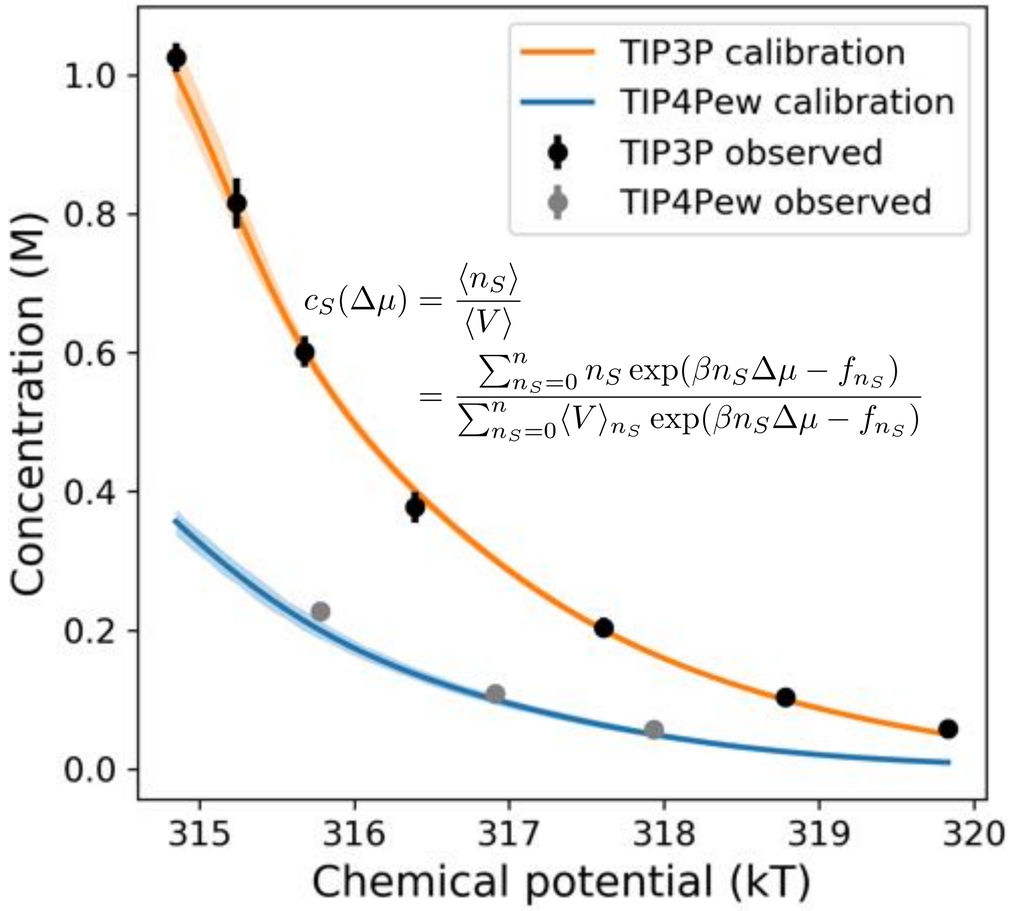
GREGORY ROSS

AN NCMC OSMOSTAT CAN MODEL REALISTIC SALT CONCENTRATION FLUCTUATIONS

SAMS+BAR estimates free energy change for inserting salt pair

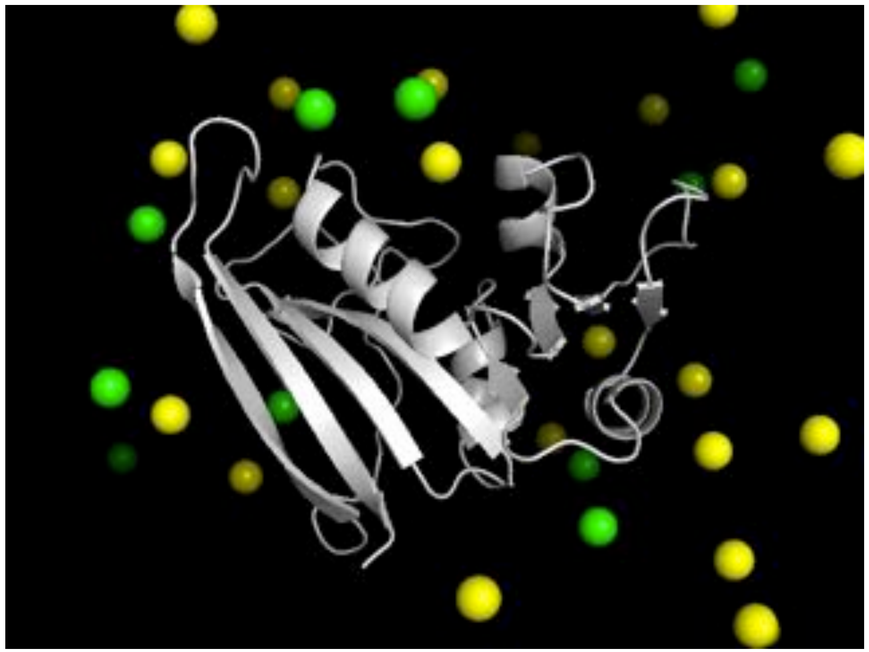


Macroscopic salt concentration vs chemical potential

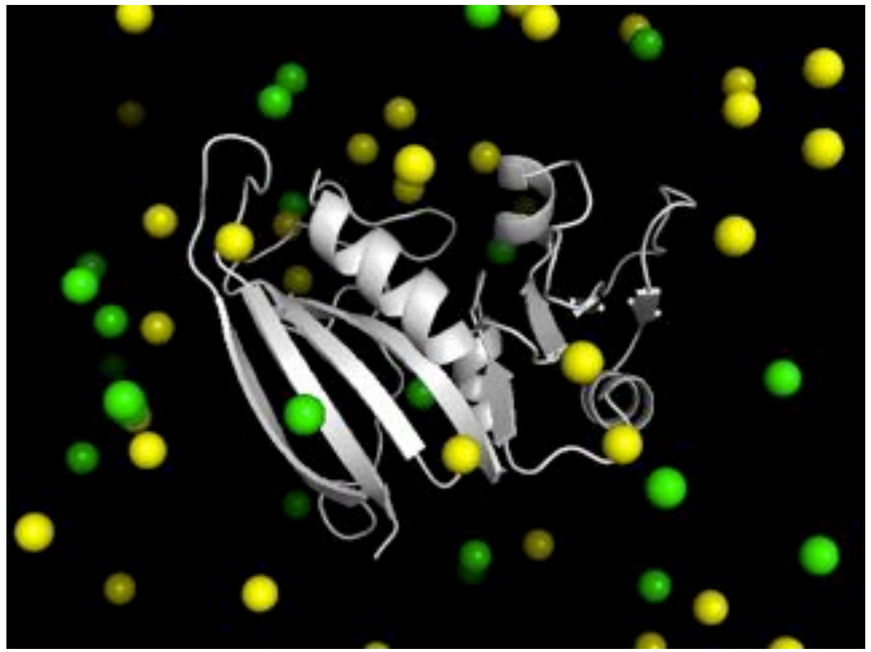


COUNTERION ENVIRONMENTS NEAR BIOMOLECULES ARE HIGHLY DYNAMIC

100 mM NaCl



200 mM NaCl



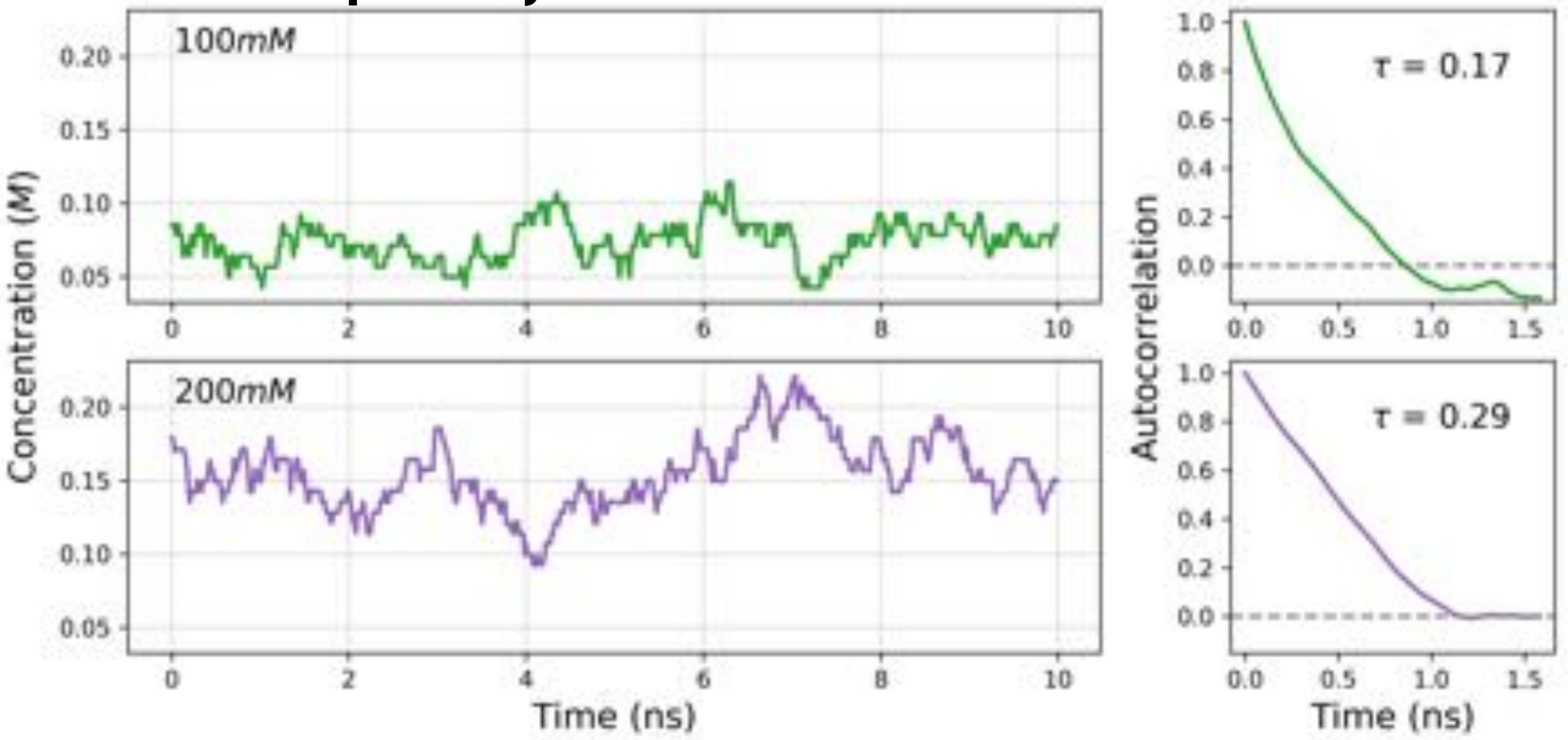
DHFR (from JAC benchmark) in TIP3P with PME

AMBER99SB-ILDN with Cheatham-Joung ion parameters

GREGORY ROSS

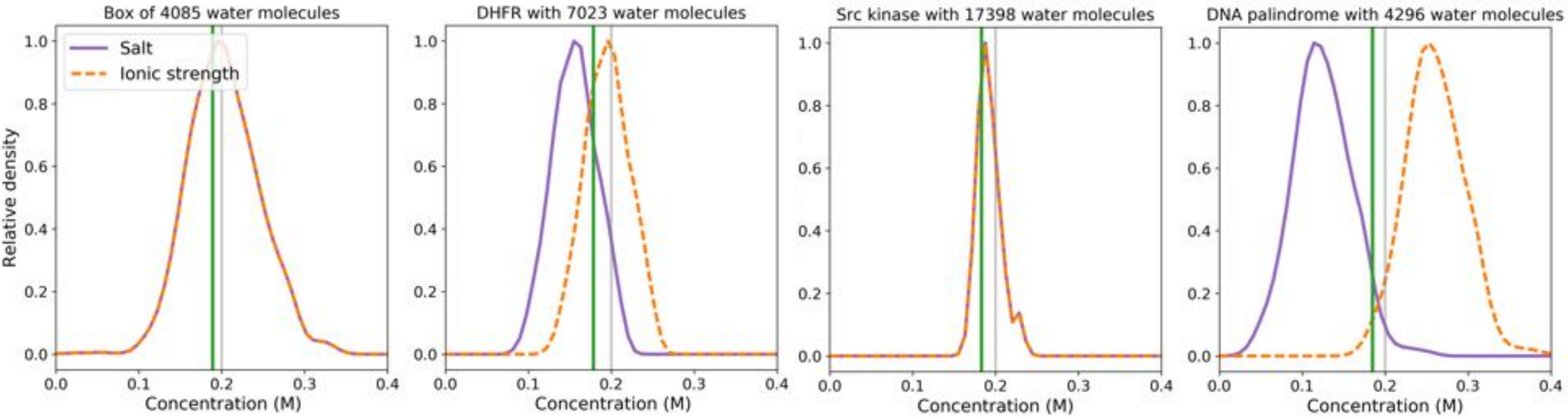
COUNTERION ENVIRONMENTS NEAR BIOMOLECULES ARE HIGHLY DYNAMIC

salt pair trajectories for DHFR



GREGORY ROSS

COUNTERION ENVIRONMENTS NEAR BIOMOLECULES ARE HIGHLY DYNAMIC



Salt concentration

Ionic strength

$$I = \frac{1}{2}\sum_i c_i z_i^2$$

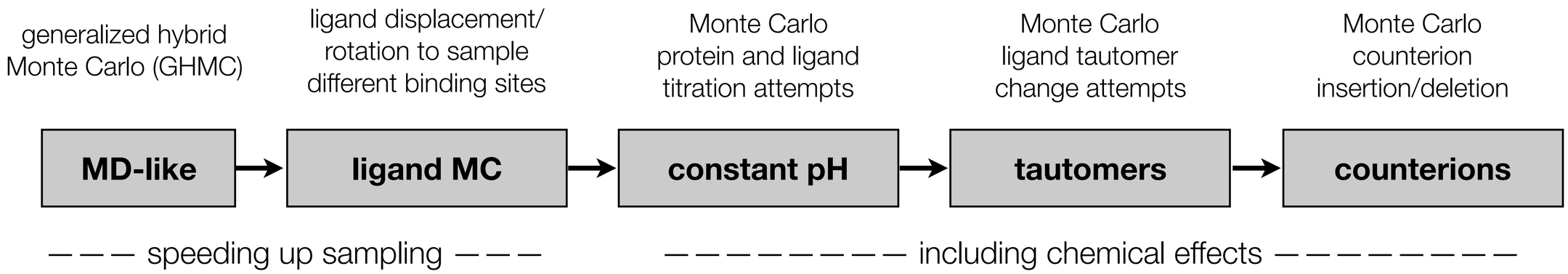
Fixed number guess

Macroscopic concentration

Real salt environments can differ substantially from our typical fixed-number guess

GREGORY ROSS

MARKOV CHAIN MONTE CARLO (MCMC) PROVIDES A FLEXIBLE FRAMEWORK FOR ENHANCEMENTS



WE CAN USE FREE ENERGY CALCULATIONS AND EXPERIMENTS TO QUANTIFY THE ERROR IN NEGLECTING OF PROTOMERS AND TAUTOMERS

SAMPL6 IS COMING SOON

Model systems of **intermediate complexity** to focus community on challenges in blind tests

Model protein-ligand systems

Isolate individual physical challenges (e.g. binding of charged ligands)

Physical properties

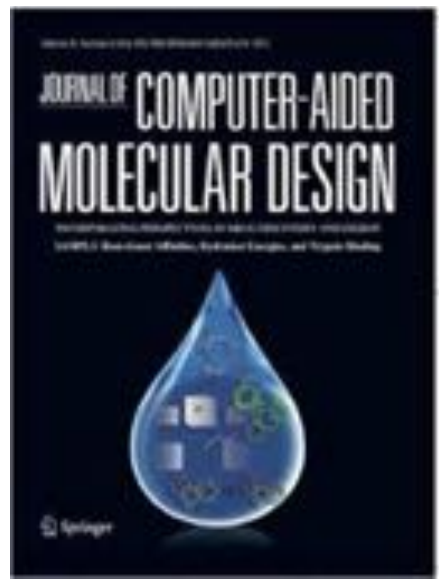
Tests of forcefield accuracy in hydrated or protein-like environments

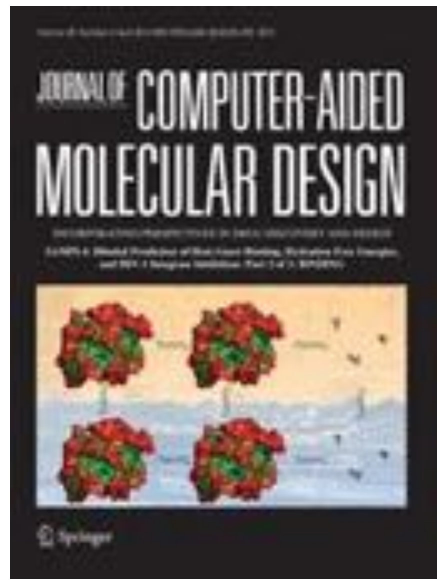
Isolate chemical effects (protonation states, ligand conformations) **without** slow protein timescales

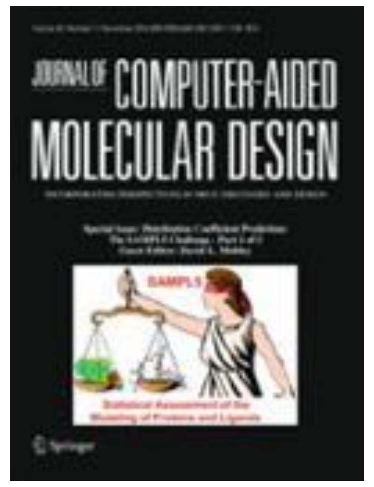
Host-guest systems

Binding of small drug-like molecules with protein-like affinities, **without** slow protein timescales

SAMPL0 2007	SAMPL1 2008	SAMPL2 2009	SAMPL3 2011	SAMPL4 2013	SAMPL5 2016	SAMPL6 2017	SAMPL7 2018
JNK3 kinase inhibitors hydration free energies	CDK2 kinase inhibitors hydration free energies	hydration free energies tautomer ratios	trypsin inhibitors hydration free energies	HIV-1 integrase inhibitors hydration free energies octaacid host-guest CB7 host-guest	distribution coefficients CBClip host-guest CB7 host-guest	distribution coefficients solubilities (CheqSol) octaacid host-guest CB8 host-guest	human serum albumin pKas distribution coefficients solubilities (CheqSol) octaacid host-guest CB8 host-guest







https://drugdesigndata.org/about/sampl

SAMPL7/8: HUMAN SERUM ALBUMIN

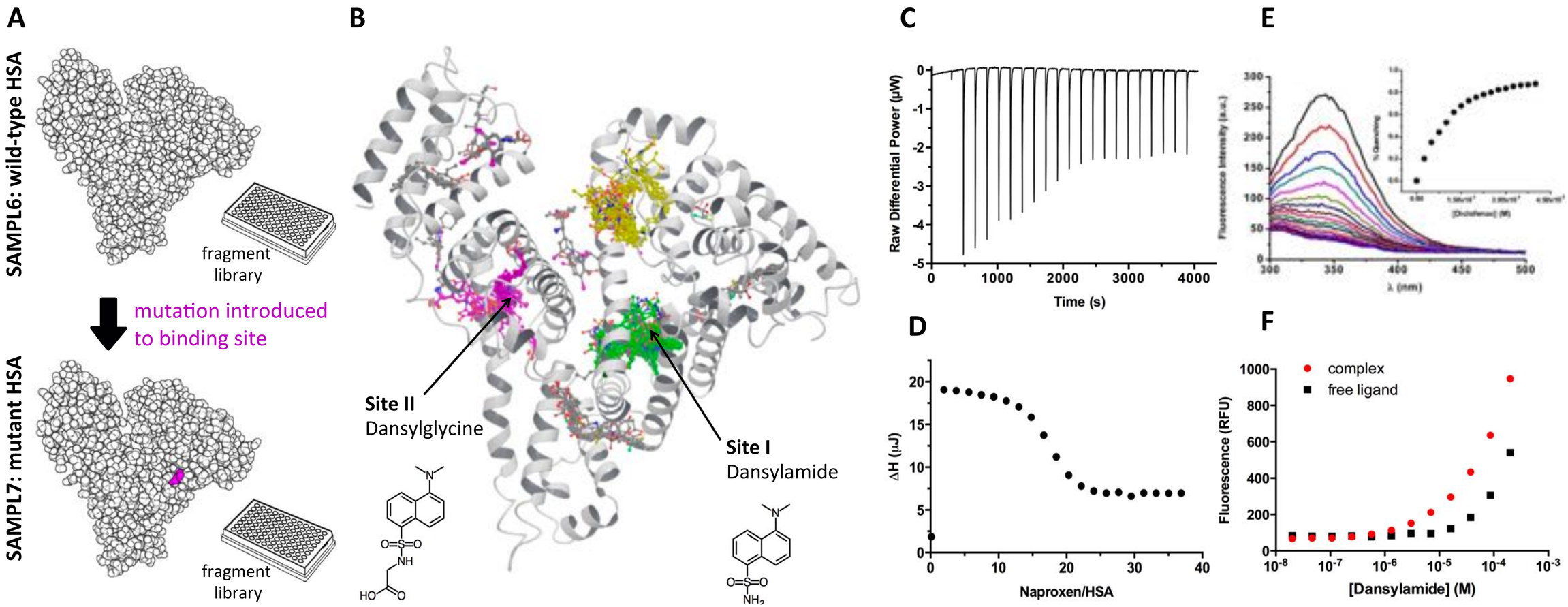


Figure 7. The SAMPL7/8 protein-ligand Challenge focuses on soluble drug fragment binding to human serum albumin (HSA)

https://drugdesigndata.org/about/sampl

THE CHODERA LAB @ MSKCC

Josh Fass
Bas Rustenburg
Andrea Rizzi
Patrick Grinaway
Greg Ross
Levi Naden

Code and data available at http://www.choderalab.org

Start Folding at http://folding.stanford.edu

WE MAKE TOOLS FOR THE COMMUNITY

OpenMM (GPU-accelerated MD with Python API) : https://openmm.org

openmmtools (integrators, alchemy, samplers) : http://openmmtools.readthedocs.io

YANK (absolute alchemical free energy calculations) : http://getyank.org

Software best practices : https://github.com/choderalab/software-development

Omnia consortium : http://omnia.md

Open Forcefield Group : https://github.com/open-forcefield-group