YANK

A free, open source, extensible platform for GPU-accelerated binding free energy calculations

PRIMARY MAINTAINERS

John D. Chodera, Patrick Grinaway, Bas Rustenburg

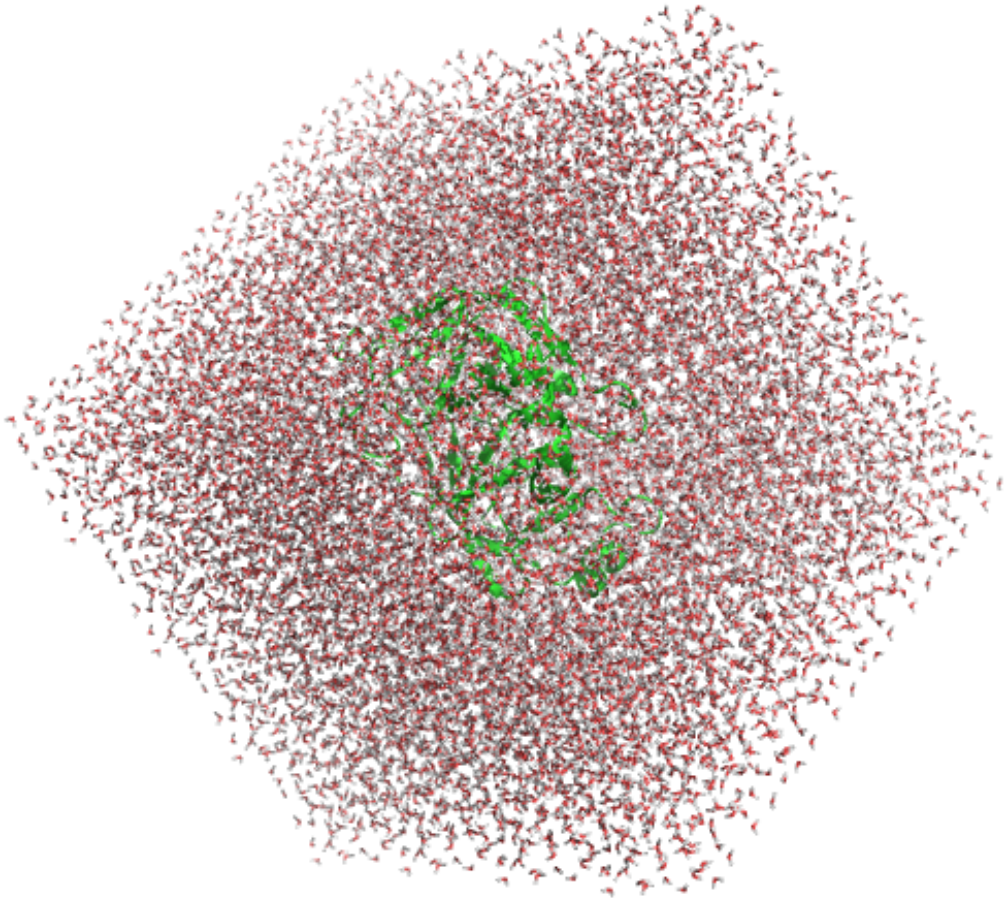
CONTRIBUTORS

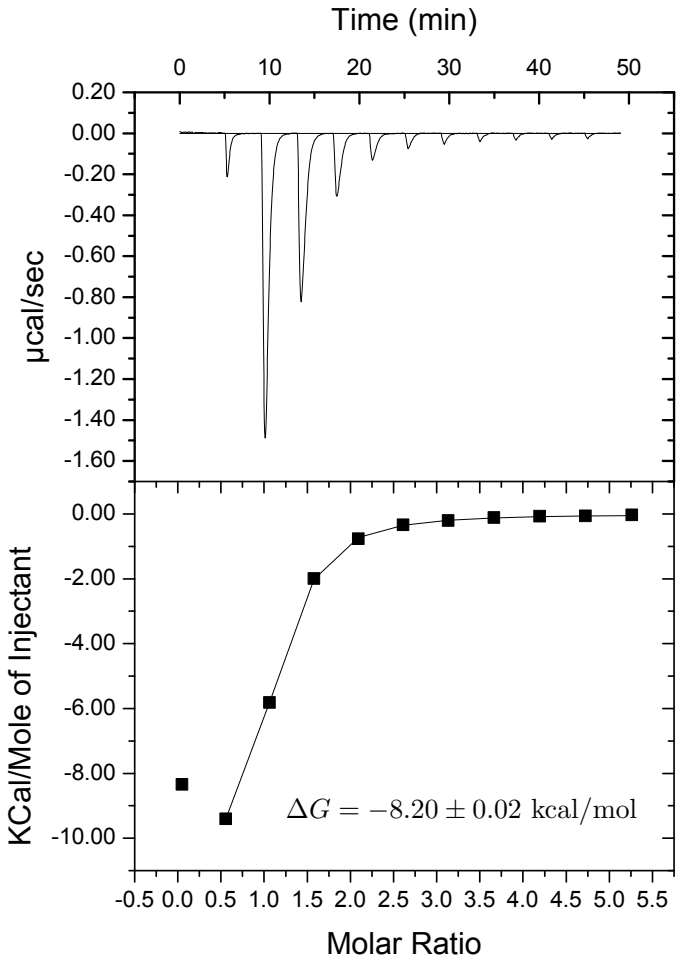
Kim Branson, Kyle A. Beauchamp, Peter M. Eastman, Mark Friedrichs, Imran S. Haque, Christoph Klein, Levi Naden, Daniel Parton, Randy Radmer, Andrea Rizzi, Michael Shirts, Kai Wong

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

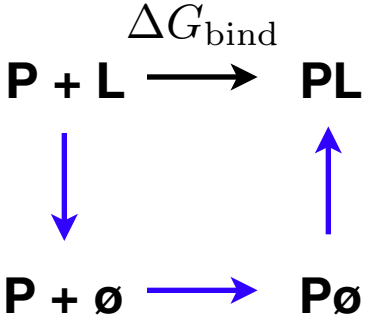
DISCLOSURES:
- Scientific Advisory Board, Schrödinger

YANK COMPUTES
ABSOLUTE BINDING FREE ENERGIES
TO COMPARE DIRECTLY WITH EXPERIMENT

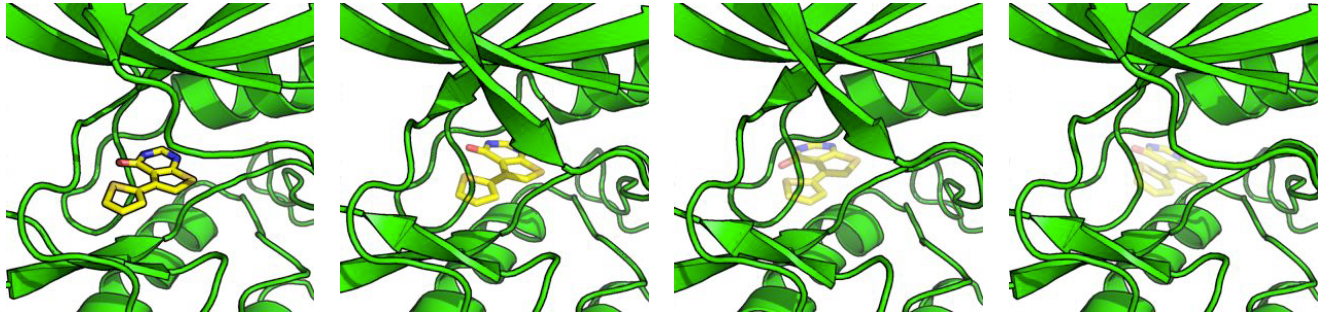




ALCHEMICAL FREE ENERGY CALCULATIONS PROVIDE A RIGOROUS WAY TO EFFICIENTLY COMPUTE BINDING AFFINITIES



multiple simulations of alchemical intermediates



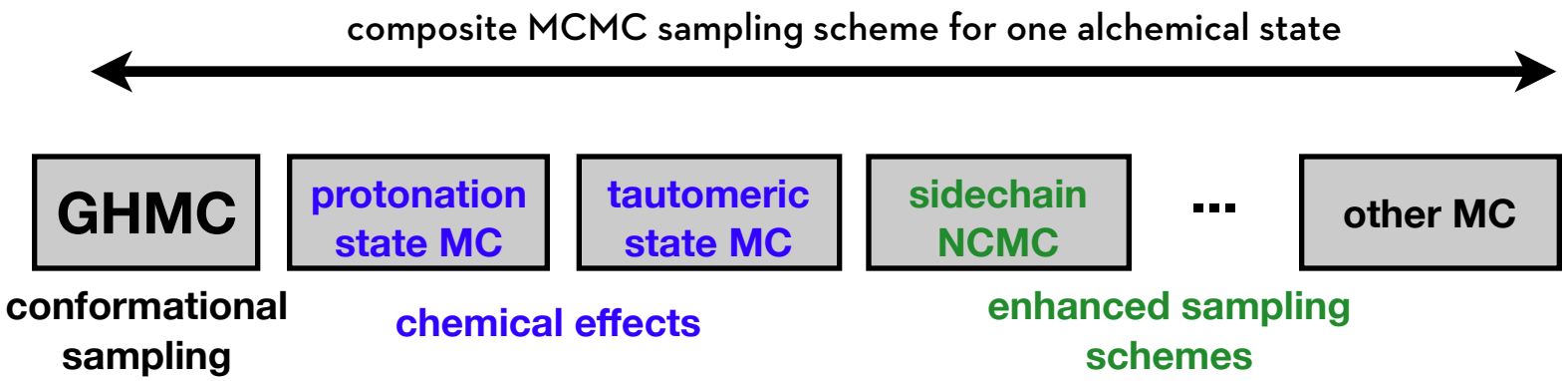
restraint imposition | discharging | steric decoupling | noninteracting

Requires orders of magnitude less effort than simulating direct association process, but still includes all enthalpic/entropic contributions to binding free energy.

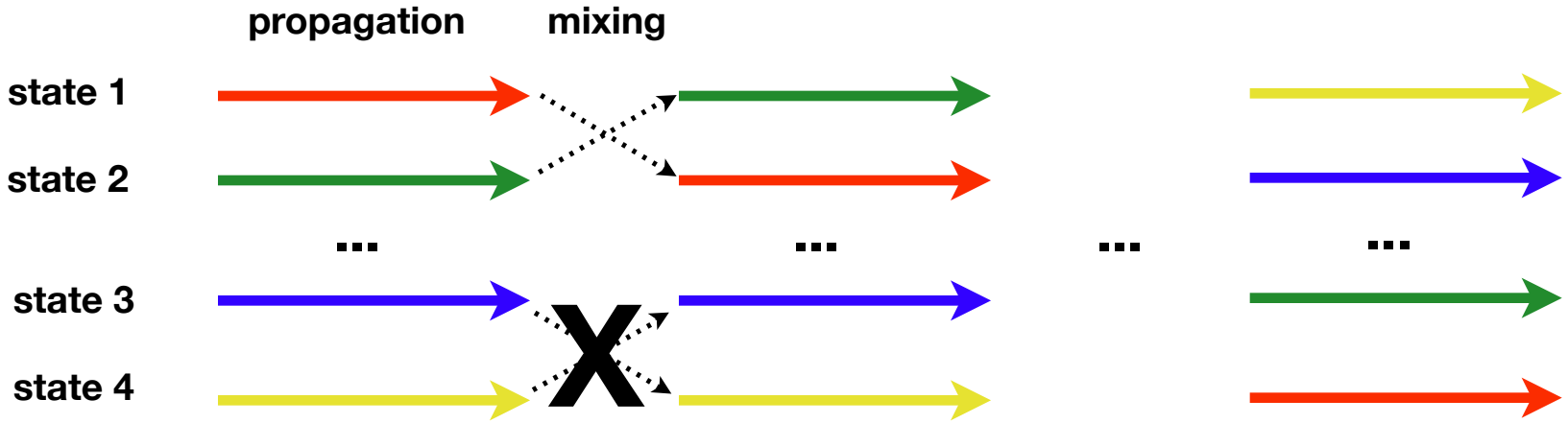
$$\Delta F_{1\to N} = -\beta^{-1} \ln \frac{Z_N}{Z_1} = -\beta^{-1} \ln \frac{Z_2}{Z_1} \cdot \frac{Z_3}{Z_2} \cdots \frac{Z_N}{Z_{N-1}} = \sum_{n=1}^{N-1} \Delta F_{n\to n+1} \qquad Z_n = \int d\mathbf{x}\, e^{-\beta U(\mathbf{x})}$$

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.

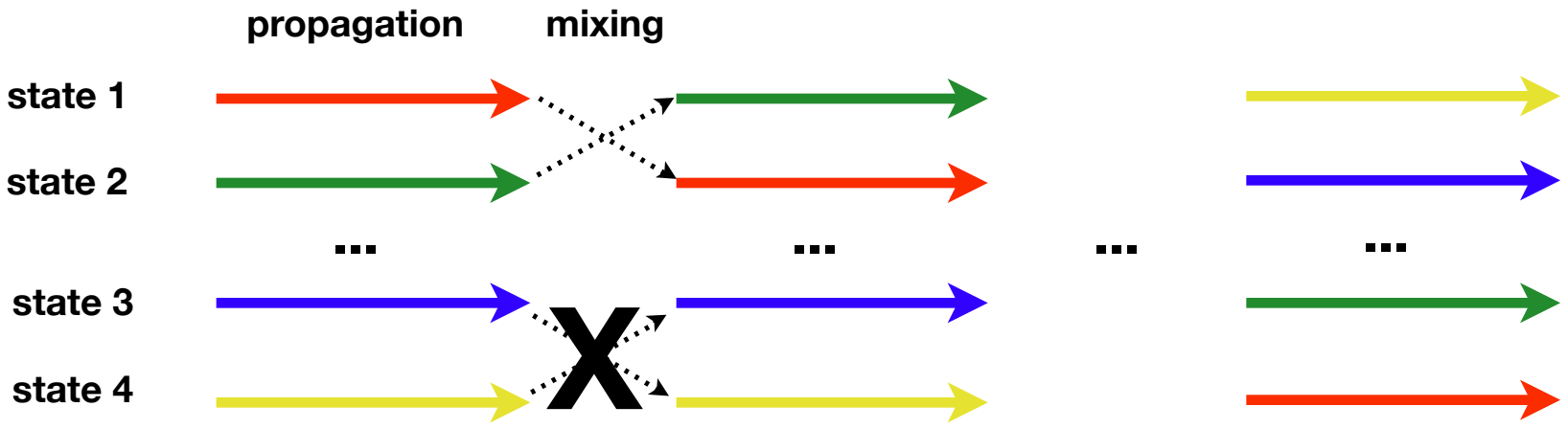
MARKOV CHAIN MONTE CARLO (MCMC) FRAMEWORK ALLOWS FOR FLEXIBLE INCLUSION OF ENHANCED SAMPLING SCHEMES AND CHEMICAL EFFECTS



Can be combined with replica exchange schemes to decrease correlation times



ALCHEMICAL REPLICA-EXCHANGE WITH GIBBS SAMPLING REDUCES CORRELATION TIMES



samples from joint equilibrium distribution of all *K* replicas: $\pi(X,S) \propto \prod_{i=1}^{K} \pi_{s_i}(x_i)$

Replica exchange can be considered a form of Gibbs sampling:

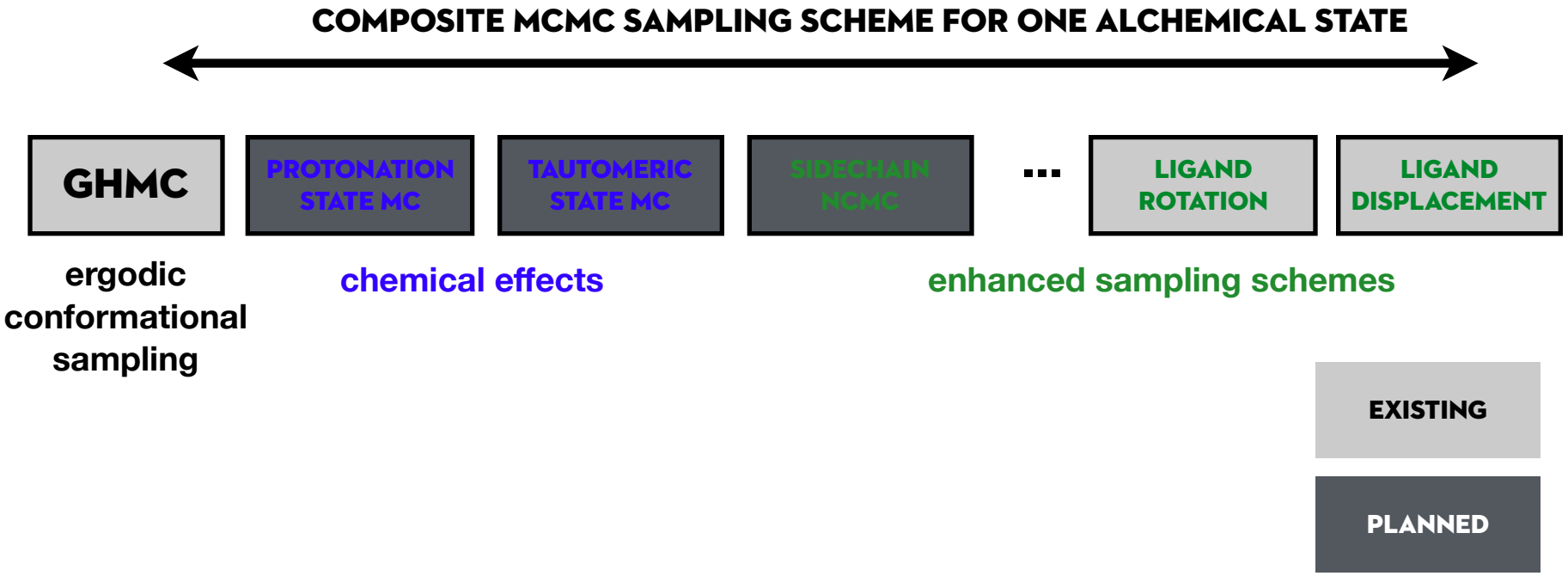
1. update configurations $X_n \sim p(X|S_n)$ [expensive molecular dynamics]
2. update permutation of state labels $S_{n+1} \sim p(S|X_{n+1})$ [inexpensive Monte Carlo swaps]

	Mixing time (1-λ	Autocorrelation of state index variable	End-to-end time
Metropolis	95.1 ± 0.2 ps	211 ± 60 ps	508 ± 20 ps
Gibbs	25.8 ± 0.1 ps	67 ± 4 ps	196 ± 6 ps

2.5x speedup!

Chodera JD and Shirts MR. JCP 135:194110, 2011

MARKOV CHAIN MONTE CARLO (MCMC) FRAMEWORK FOR ENHANCED SAMPLING SCHEMES AND CHEMICAL EFFECTS



THE MULTISTATE BENNETT ACCEPTANCE RATIO (MBAR) ESTIMATOR EXTRACTS ALL INFORMATION FROM THE DATA

Statistically optimal analysis of samples from multiple equilibrium states

Michael R. Shirts[1,a)] and John D. Chodera[2,b)]

– HIDE AFFILIATIONS

[1] Department of Chemical Engineering, University of Virginia, Charlottesville, Virginia 22904, USA

[2] Department of Chemistry, Stanford University, Stanford, California 94305, USA

a) Electronic mail: michael.shirts@virginia.edu.

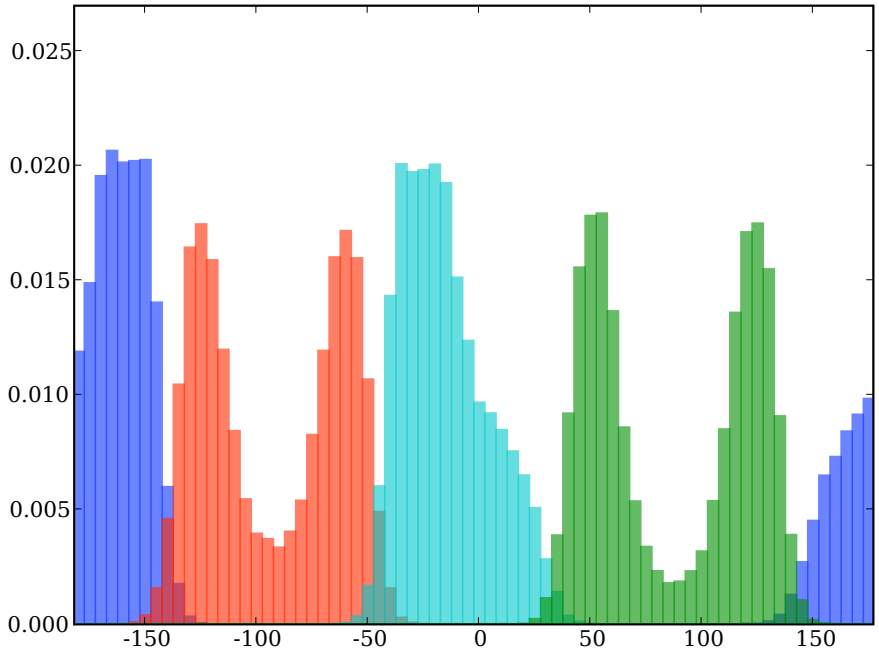
b) Electronic mail: jchodera@gmail.com

J. Chem. Phys. **129**, 124105 (2008); http://dx.doi.org/10.1063/1.2978177

$$\hat{f}_i = -\ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{\exp[-u_i(\boldsymbol{x}_{jn})]}{\sum_{k=1}^{K} N_k \exp[\hat{f}_k - u_k(\boldsymbol{x}_{jn})]}$$

$$\delta^2 \Delta \hat{f}_{ij} = \hat{\Theta}_{ii} - 2\hat{\Theta}_{ij} + \hat{\Theta}_{jj}$$

$$\hat{\boldsymbol{\Theta}} = \mathbf{W}^{\mathrm{T}}(\mathbf{I}_N - \mathbf{W}\mathbf{N}\mathbf{W}^{\mathrm{T}})^{+}\mathbf{W}$$

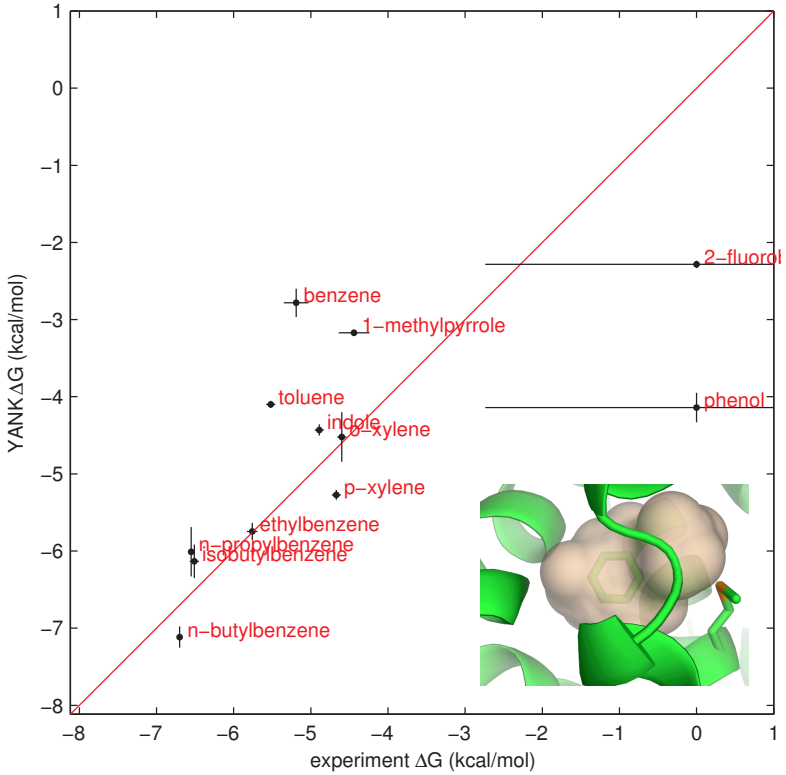


- asymptotically optimal estimator for free energy differences from equilibrium data
- robust estimates of uncertainties
- combine data from multiple temperatures, pressures, bias potentials
- freely-available Python implementation installable via **conda**
- batteries included: comes with tools to subsample correlated data to extract independent data

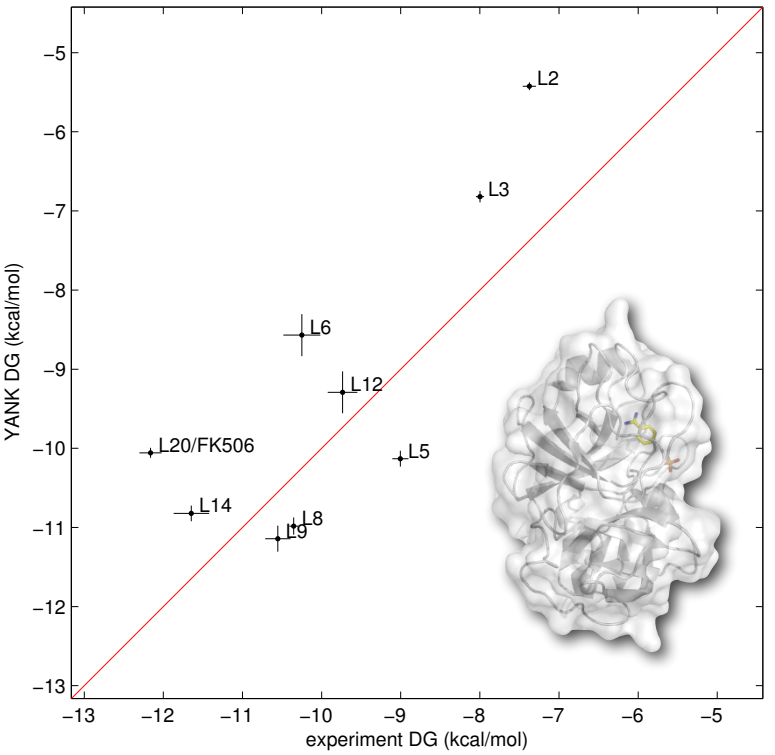
http://github.org/choderalab/pymbar

FREE ENERGIES WITH IMPLICIT MODELS OF SOLVENT ARE PROMISING: COULD PLAY A ROLE IN RAPID AFFINITY PREDICTION

T4 LYSOZYME L99A



FKBP12



AMBER ff96 + OBC GBSA (no cutoff) + GAFF/AM1-BCC
12 h on 2 GPUs

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

INSTALLING YANK

```
conda install -c http://conda.binstar.org/omnia yank
```

(Is this simple enough now, Paul?)

SETTING UP A YANK CALCULATION

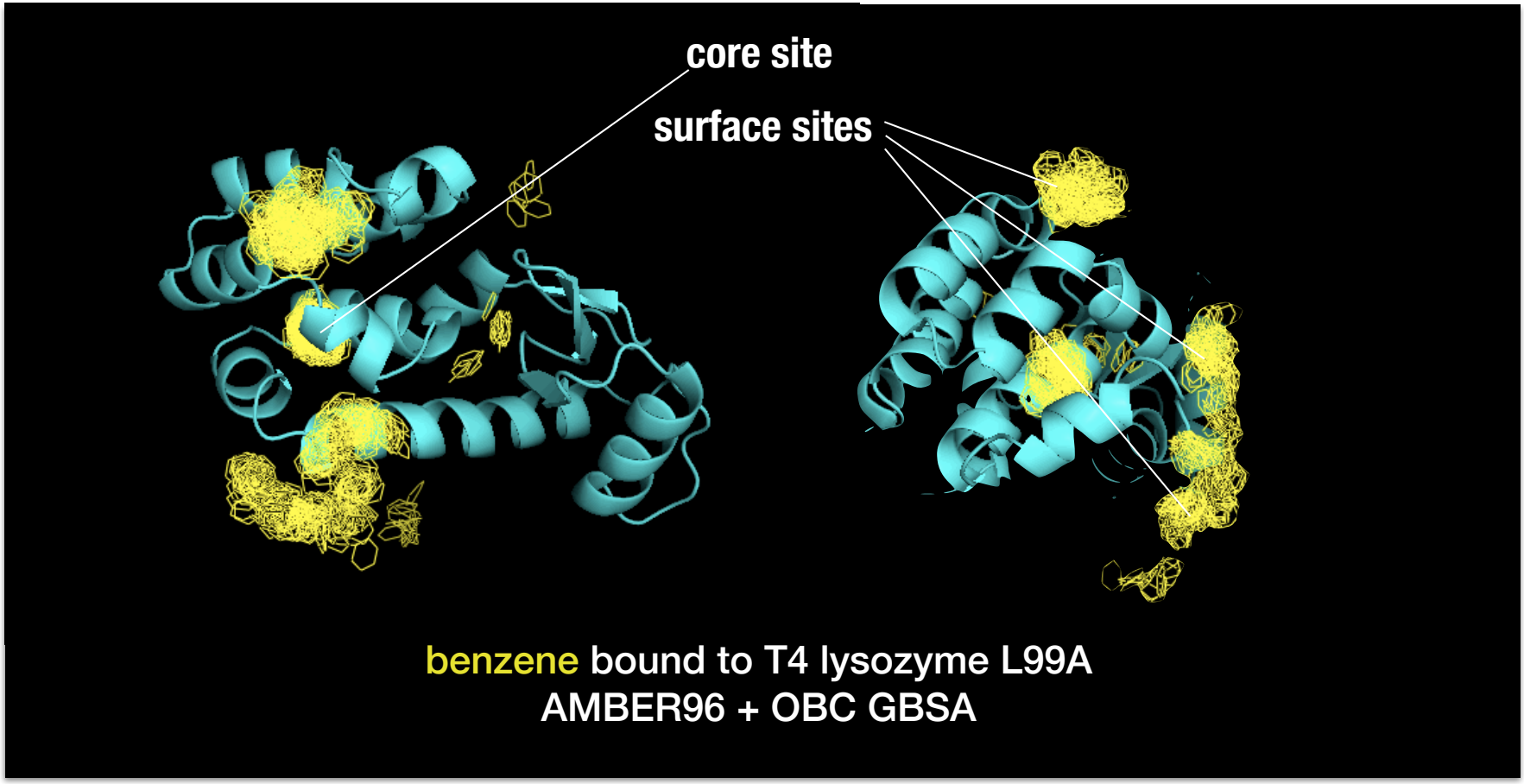
Using the command-line:

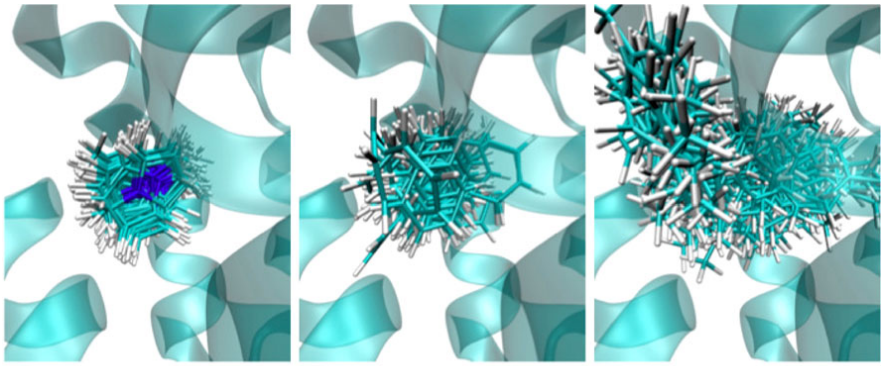
```
#!/bin/bash

# Set up calculation.
echo "Setting up binding free energy calculation..."
yank prepare binding amber --setupdir=setup --ligand="resname MOL" --store=output --iterations=1000 \
    --nbmethod=CutoffPeriodic --temperature="300*kelvin" --pressure="1*atmosphere" --minimize --verbose

# Run the simulation with verbose output:
echo "Running simulation..."
yank run --store=output --verbose

# Analyze the data (can be done asynchronously)
echo "Analyzing data..."
yank analyze --store=output
```

Using the Python API:

```
from yank.yank import Yank

# Initialize YANK object.
yank = Yank(store_dir)

# Set some options.
options = dict()
options['number_of_iterations'] = 1000

# Create reference thermodynamic state.
from yank.repex import ThermodynamicState
thermodynamic_state = ThermodynamicState(temperature=temperature, pressure=pressure)

# Create new simulation.
yank.create(phases, systems, positions, atom_indices, thermodynamic_state, options=options)

# Run the simulation
yank.create(phases, systems, positions, atom_indices, thermodynamic_state, options=options)

# Analyze the simulation
results = yank.analyze()
```

ADDITIONAL BINDING SITES CAN BE IDENTIFIED AND INDIVIDUAL AFFINITIES ESTIMATED BY MIXING IN MONTE CARLO MOVES





Chodera and Shirts. JCP 135:194110, 2011
Wang, Chodera, Yang, and Shirts. JCAMD 27:989, 2013.

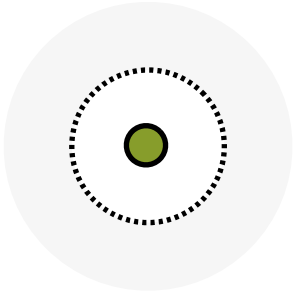
CUSTOM OPENMM FORCES ALLOW EXPERIMENTATION WITH ALCHEMICAL DEFINITIONS

```
if isinstance(reference_force, openmm.NonbondedForce):
    # CustomNonbondedForce will handle softcore interactions with ligand.
    energy_expression = “4*epsilon*lambda*x*(x-1.0);" # softcore potential
    energy_expression += "x = 1.0/(alpha*(1.0-lambda) + (r/sigma)^6);"
    energy_expression += "epsilon = sqrt(epsilon1*epsilon2);” # Lorentz-Berthelot combining rules
    energy_expression += "sigma = 0.5*(sigma1 + sigma2);” # Lorentz-Berthelot combining rules
    energy_expression += "lambda = lambda1*lambda2;" # alchemical combining rule

    force = openmm.CustomNonbondedForce(energy_expression)
    alpha = 0.5 # softcore parameter
    force.addGlobalParameter("alpha", alpha);
    force.addPerParticleParameter("sigma")
    force.addPerParticleParameter("epsilon")
    force.addPerParticleParameter("lambda");
    for particle_index in range(reference_force.getNumParticles()):
        # Retrieve parameters.
        [charge, sigma, epsilon] = reference_force.getParticleParameters(particle_index)
        # Alchemically modify parameters.
        if particle_index in ligand_atoms:
            force.addParticle([sigma, epsilon, vdw_lambda])
        else:
            force.addParticle([sigma, epsilon, 1.0])
    for exception_index in range(reference_force.getNumExceptions()):
        # Retrieve parameters.
        [iatom, jatom, chargeprod, sigma, epsilon] = reference_force.getExceptionParameters(exception_index)
        # All exceptions are handled by NonbondedForce, so we exclude all these here.
        force.addExclusion(iatom, jatom)
    force.setNonbondedMethod( reference_force.getNonbondedMethod() )
    force.setCutoffDistance( reference_force.getCutoffDistance() )
    system.addForce(force)
```

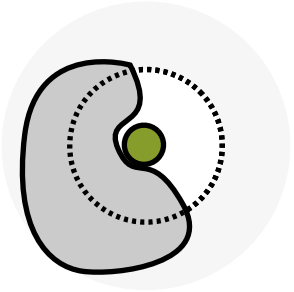
ANISOTROPIC LONG-RANGE DISPERSION CORRECTION IS REQUIRED TO ELIMINATE SYSTEMATIC ERROR IN BINDING AFFINITIES IN EXPLICIT SOLVENT

Simulations in explicit solvent must be run with long-range dispersion correction to ensure results are not sensitive to choice of Lennard-Jones cutoff.

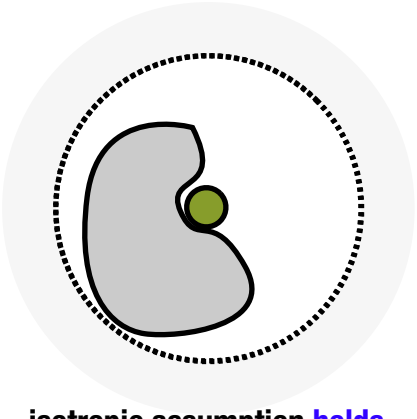
This correction assumes isotropic distribution of Lennard-Jones sites throughout system, but protein/water mixtures are not homogeneous and isotropic!



isotropic assumption holds



isotropic assumption fails

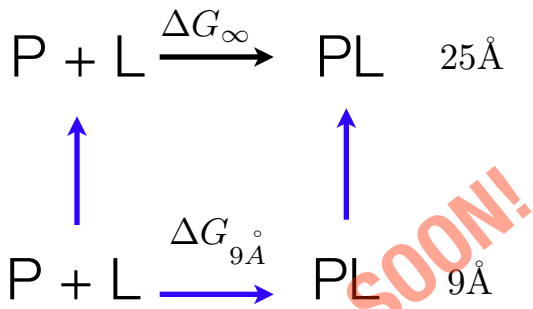


isotropic assumption holds

Instead, we have to enlarge cutoff so that isotropic assumption holds

An explicit postprocessing step recomputes energies with large cutoff and estimates perturbation free energies using exponential reweighting.

Error can be as large as 3 kcal/mol, depending on number of ligand atoms



COMING SOON!

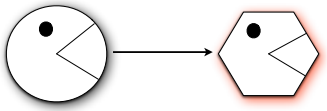
M. R. Shirts*, **D. L. Mobley***, J. D. Chodera, and V. S. Pande. "Accurate and efficient corrections for missing dispersion interactions in molecular simulations", *J. Phys. Chem. B* **111**:13052-13063 (2007).

A MAJOR GOAL OF YANK: QUANTIFY HOW SENSITIVE BINDING AFFINITIES TO VARIOUS PHYSICAL EFFECTS

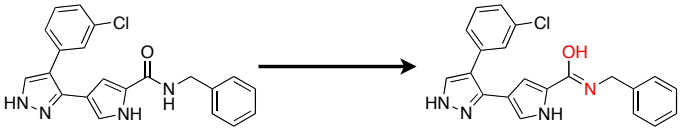
simulation details

experiment

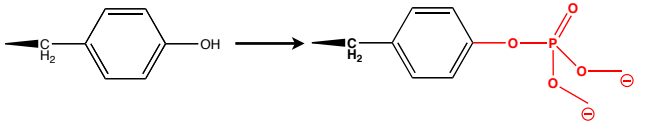
Protein conformation



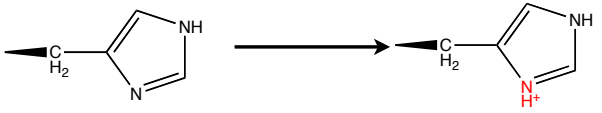
Ligand protonation/tautomeric state



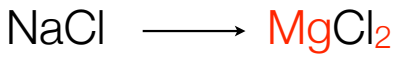
Phosphorylation state



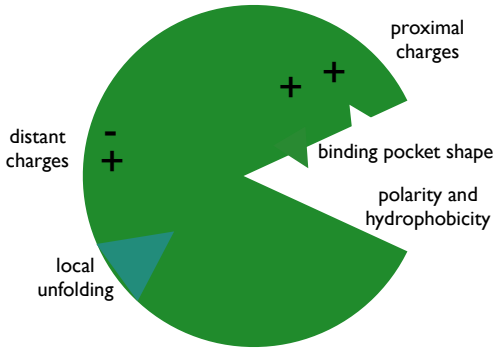
Protein protonation state



Salt environment



Proximal/distal charged residues
Binding pocket shape
Polarity and hydrophobicity
Local unfolding



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)

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

method	natoms	gromacs CPU	OpenMM GPU	speedup
GB/SA	2,489	2.54 ns/day	287 ns/day	**113 x**
RF	23,558	18.8 ns/day	163 ns/day	**8.7 x**
PME	23,558	6.96 ns/day	104 ns/day	**15 x**

http://openmm.org
gromacs benchmarks from http://biowulf.nih.gov/apps/gromacs-gpu.html

NVIDIA GTX-TITAN ($1000)

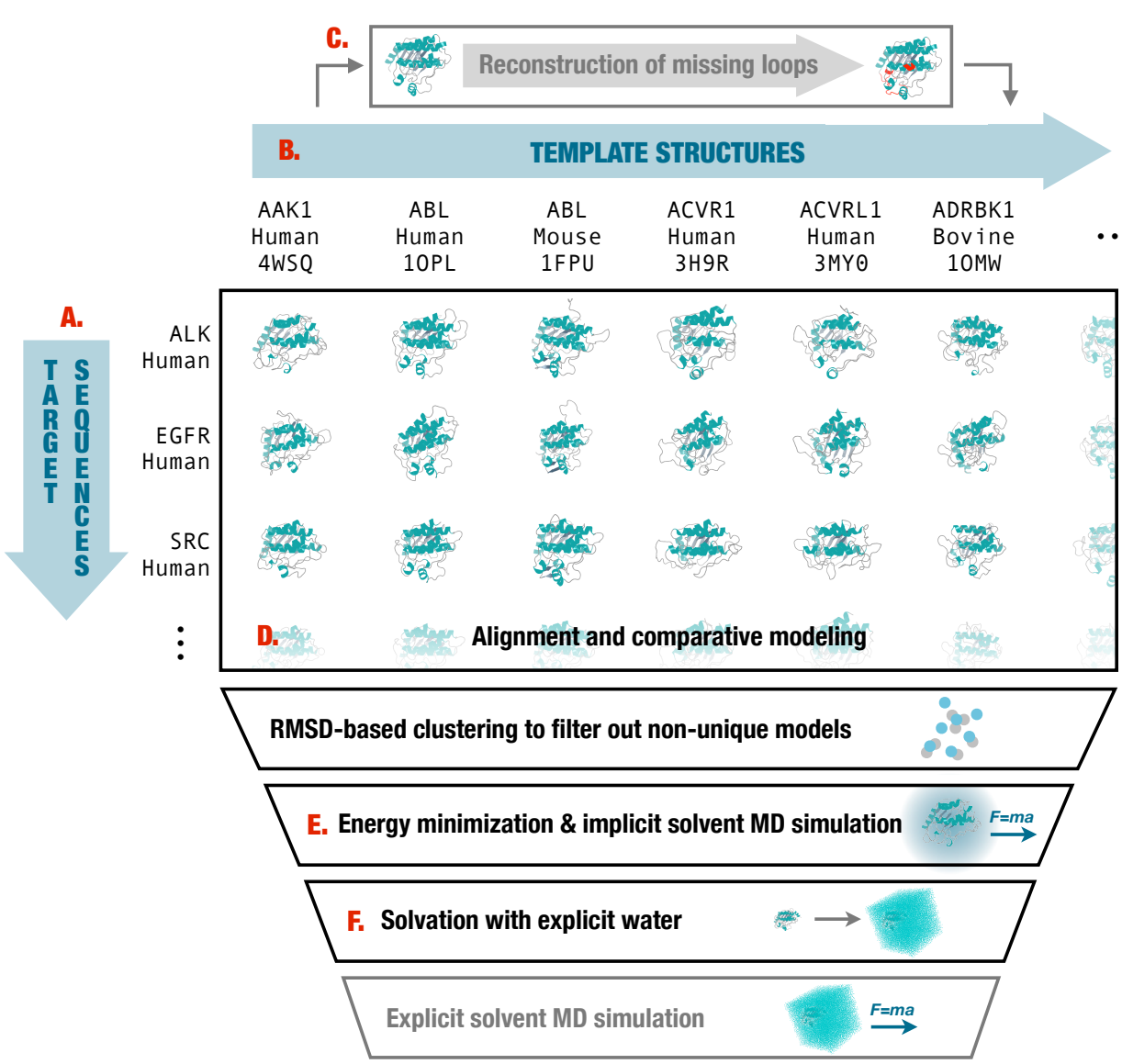
OBLIGATORY HAZARD STATEMENT

YANK 0.7 is research software and still under active development!
There is a suite of unit tests, but we have not yet verified every capability works as expected.
Use at your own risk.

ENSEMBLER

AUTOMATED MODELING AND PREPARATION FOR SUPERFAMILY-SCALE SIMULATIONS

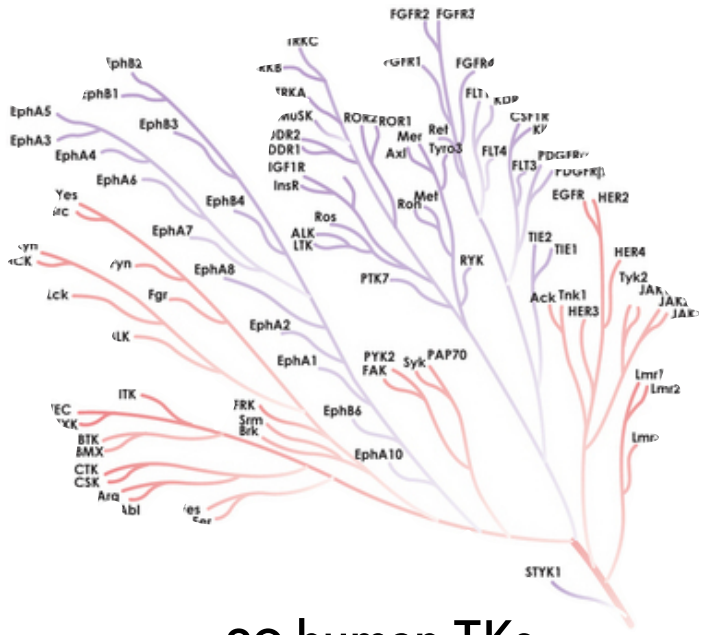
DANIEL PARTON, SONYA HANSON, PATRICK GRINAWAY



	Number of models			Timing
	All TKs	SRC	ABL	
				3h/template (loop reconstruction)
B. TEMPLATE STRUCTURES	412,269	4433	4433	
D. Alignment and comparative modeling	397,621	4271	4431	89s/model 9863 CPU-h total
RMSD-based clustering to filter out non-unique models	394,040	4135	4289	0.3s/model 32 CPU-h total
E. Energy minimization & implicit solvent MD simulation	389,067	4066	4289	118s/model 12,418 GPU-h total
F. Solvation with explicit water		4003	4280	9s/model 947 CPU-h total

bioRxiv preprint:
http://dx.doi.org/10.1101/018036

MODELING ALL 90 HUMAN TYROSINE KINASES ONTO ALL KINASE CATALYTIC DOMAIN PDBS



90 human TKs

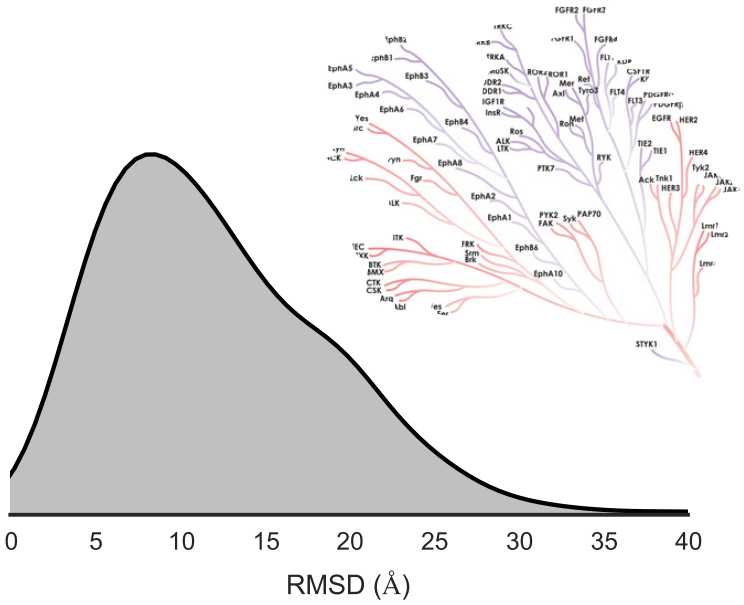
x

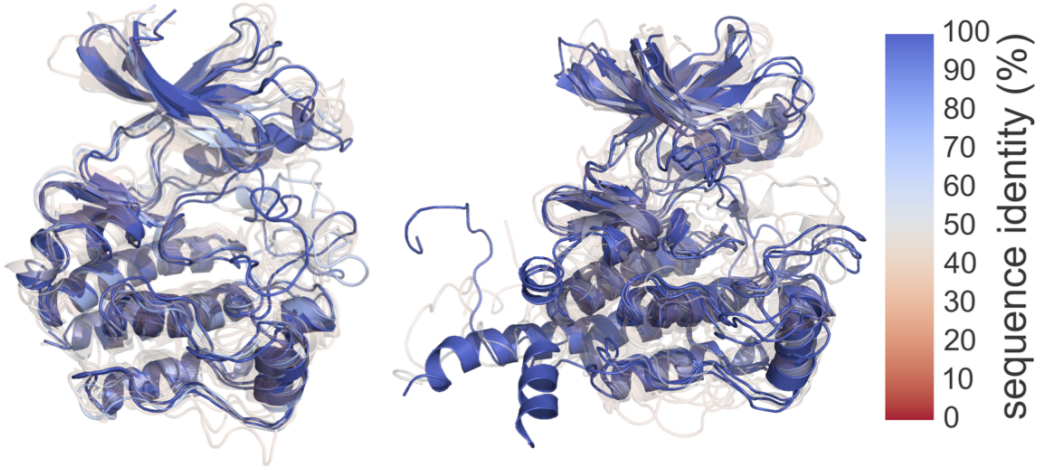
RCSB PDB
PROTEIN DATA BANK

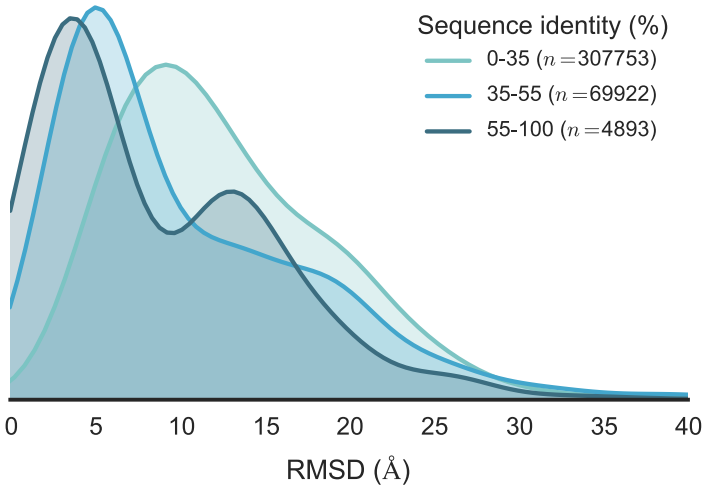
4433 PDB structures of kinase catalytic domains

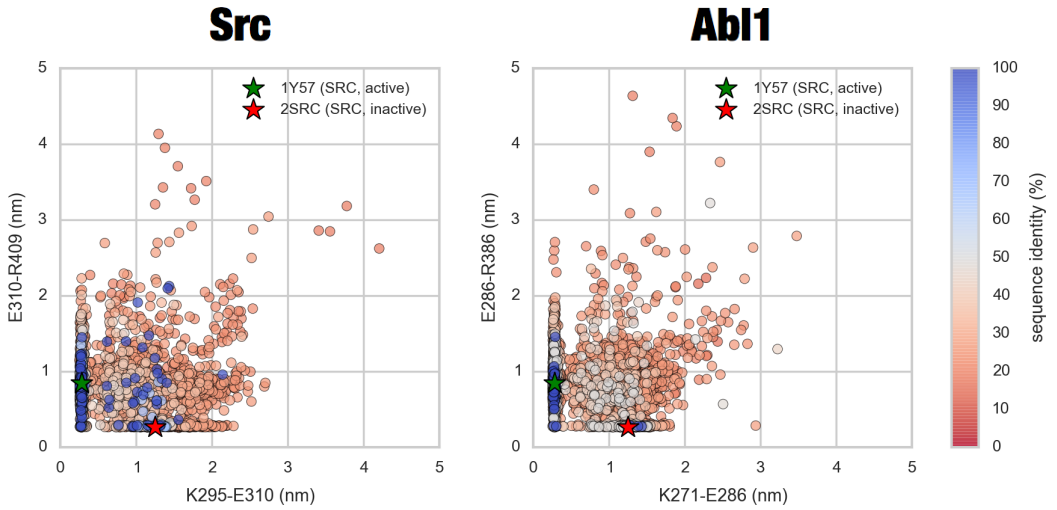
```
#!/bin/bash

conda create -c https://conda.binstar.org/omnia -n ensembler1.0 python=2.7 ensembler=1.0 --yes
source activate ensembler1.0

ensembler init
ensembler gather_targets --query 'family:"tyr protein kinase family" AND organism:"homo sapiens" AND reviewed:yes' \
   --uniprot_domain '^Protein kinase(?!; truncated)(?!; inactive)'
ensembler gather_templates --gather_from uniprot --query 'domain:"Protein kinase" AND reviewed:yes' \
   --uniprot_domain_regex '^Protein kinase(?!; truncated)(?!; inactive)'
ensembler loopmodel
ensembler align
ensembler build_models
ensembler cluster
ensembler refine_implicit
```

MODELING ALL 90 HUMAN TYROSINE KINASES ONTO ALL KINASE CATALYTIC DOMAIN PDBS










```
$ conda config --add channels http://conda.binstar.org/omnia
$ conda install ensembler
```

bioRxiv preprint:
http://dx.doi.org/10.1101/018036

OMNIA:

OPEN SOURCE, HIGH PERFORMANCE, HIGH USABILITY TOOLKITS FOR PREDICTIVE BIOMOLECULAR SIMULATION.

OpenMM

PDBFixer

MSMBuilder

MDTraj

Ensembler

http://omnia.md

INSTALLING OMNIA

```
conda config --add channels http://conda.binstar.org/omnia
conda install omnia
```

That's really it. Seriously.

OMNIA ENABLES REPRODUCIBLE SCIENCE

```
#!/bin/bash

# Create conda environment with exact versions of tools needed to reproduce paper.
if [ ! -d conda-env ]; then
   conda config --add channels http://conda.binstar.org/omnia
   conda create --yes --quiet -p conda-env python=2.7 openmmtools=0.7.0 openmm=6.2 matplotlib=1.4 \
       pymbar=3.0.0.beta2 netCDF4
fi
source activate ./conda-env

# Run simulations.
python simulate.py

# Analyze simulation data to generate figures.
if [ ! -e figures ]; then
    mkdir figures
fi
python analyze-1.py
python analyze-2.py

# Deactivate conda environment.
source deactivate
```

bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

A simple method for automated equilibration detection in molecular simulations

John D Chodera
doi: http://dx.doi.org/10.1101/021659

2016 WORKSHOPS

CAMBRIDGE/BOSTON MA, TENTATIVELY 16-20 MAY 2016

ALCHEMICAL FREE ENERGY METHODS IN DRUG DISCOVERY

Email list signup: https://goo.gl/bLJl1t

orgs: Michael Schnieders, Michael Shirts, David Mobley, John Chodera, Vijay Pande

MARKOV STATE MODELS IN DRUG DISCOVERY

Email list signup: https://goo.gl/bLJl1t

orgs: John Chodera, Rommie Amaro, Benoît Roux, Vijay Pande, Frank Noé

OMNIA COLLABORATORS

Vijay S. Pande, Stanford University

Vijay Pande is professor of Chemistry, Structural Biology, Biophysics, and Computer Science at Stanford University. Vijay is the founder and director of Folding@Home, the world's largest distributed computing project.

Pande lab webpage

Peter Eastman, Stanford University

Peter Eastman is the lead architect and principal developer of the OpenMM molecular dynamics suite, as well as the lead developer of PDBFixer.

John D. Chodera, MSKCC

John Chodera is an assistant professor in the Computational Biology Program at the Memorial Sloan-Kettering Cancer Center, and the lead developer of the YANK package for alchemical binding free energy calculations.

Chodera lab webpage

Justin L. MacCallum, University of Calgary

Justin MacCallum is an assistant professor in the Department of Chemistry at the University of Calgary. He is the lead developer of the MELD package for inferring protein structure from sparse and unreliable data.

MacCallum lab webpage

Kyle Beauchamp, MSKCC

Kyle Beauchamp is a research fellow in the Chodera lab at MSKCC. Kyle is a co-principal developer of MSMBuilder, MDTraj, and other tools for protein simulation and analysis.

Robert T. McGibbon, Stanford University

Robert McGibbon is a graduate student in the Pande lab at Stanford. Robert is the lead developer MDTraj, a co-principle developer of MSMBuilder, and a contributor to OpenMM.

Jason M. Swails, Rutgers University

Jason Swails is a postdoctoral researcher in the Case lab at Rutgers University. He is the principal developer of the ParmEd program to rapidly prototype force field modifications and development. He is also a contributor to the OpenMM and MDTraj projects.

HOW CAN WE MAKE OUR TOOLS MORE USEFUL TO YOU?