

REDESIGNING DRUG DESIGN

John D. Chodera

MSKCC Computational Biology Program

<http://www.choderalab.org>



Memorial Sloan-Kettering
Cancer Center



Memorial Sloan-Kettering Cancer Center

Sloan-Kettering Institute

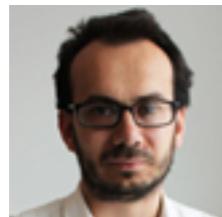
In more than 100 laboratories, our scientists are conducting innovative research to advance understanding in the biological sciences and improve human health.



cBio@MSKCC



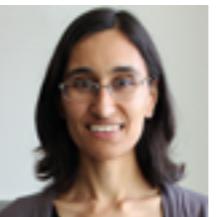
Chris
Sander



Gregoire
Altan-Bonnet



John
Chodera



Christina
Leslie



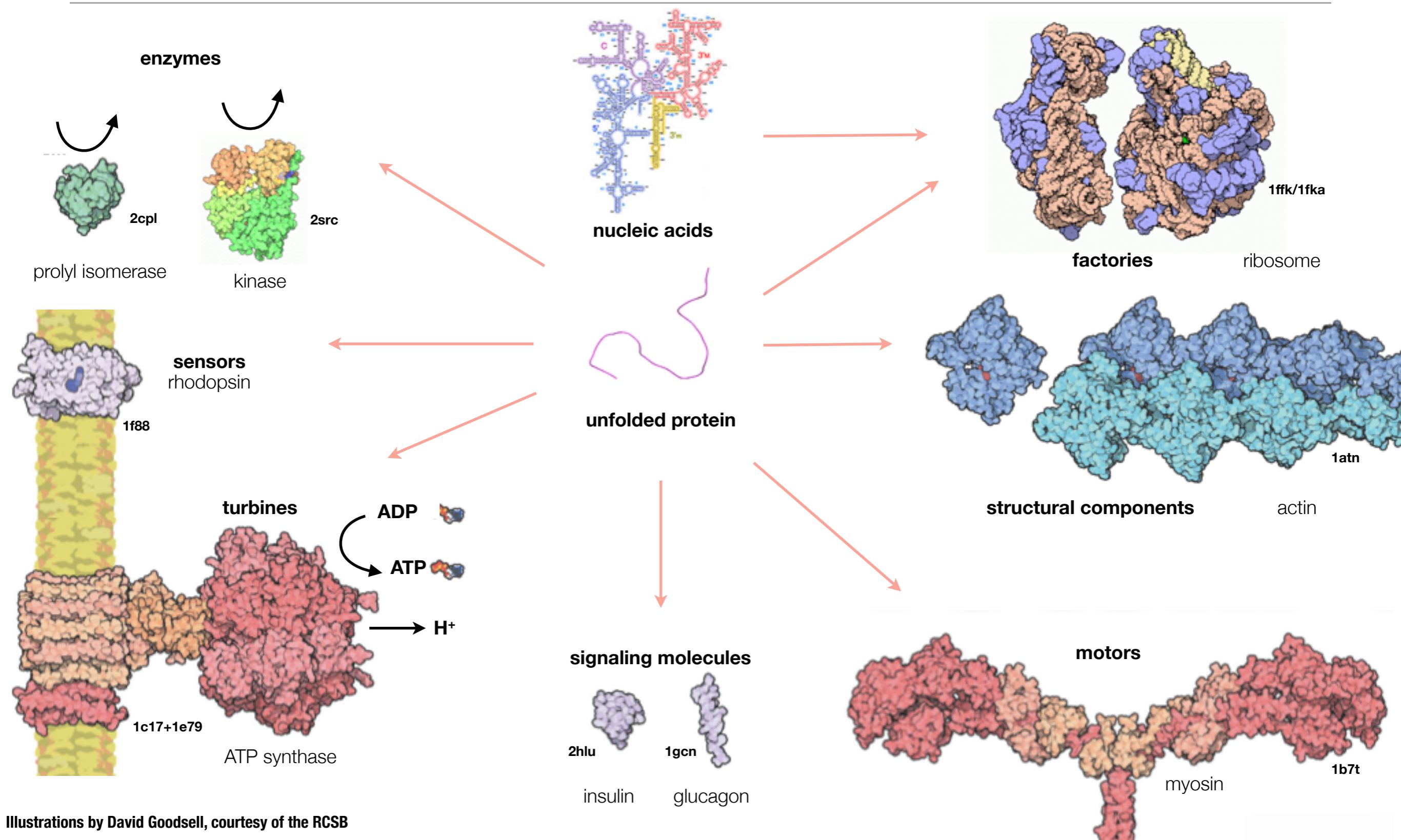
Gunnar
Rätsch



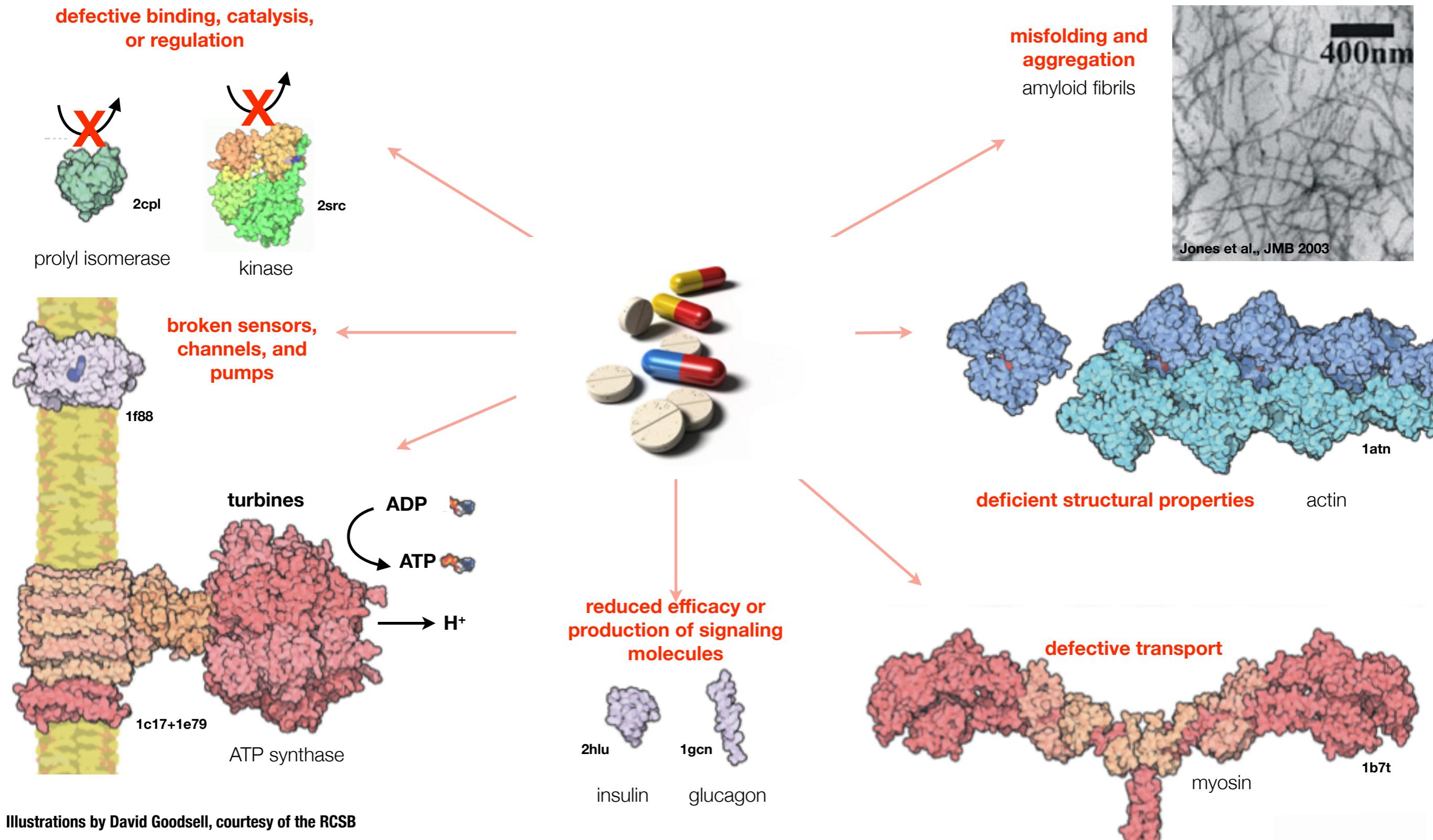
Joao
Xavier

We're hiring!

Biological macromolecules are the molecular machines of life



Perturbations can cause disease by disrupting this machinery



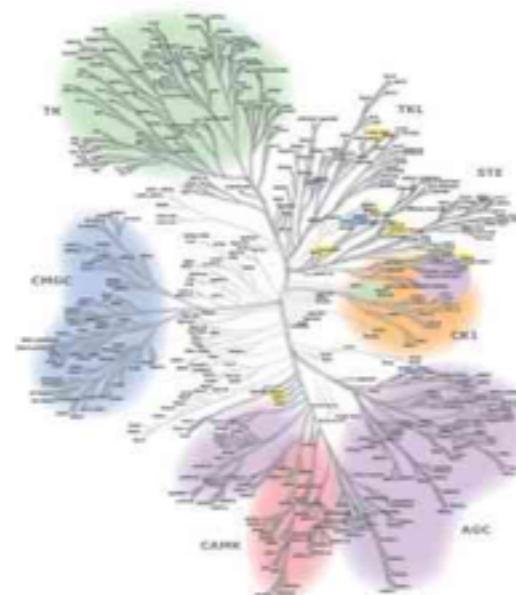
Sometimes, the drug discovery process works well: Imatinib as a success story

Bcr-Abl fusion event observed in majority of CML patients causes **constitutive activation** of ABL kinase, resulting in unchecked white blood cell proliferation.

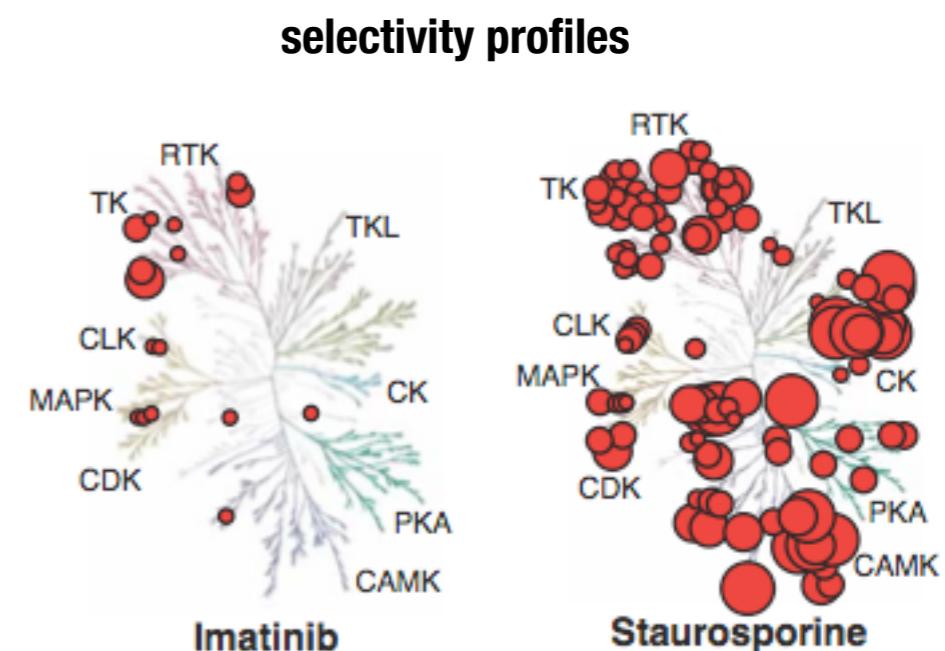
Selective inhibition by imatinib (Gleevec) has been very successful in treating CML.



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



human kinome
(518 kinases)



- K_d
- < 1 nM
- 1–10 nM
- 10–100 nM
- 100 nM–1 μM
- 1–10 μM

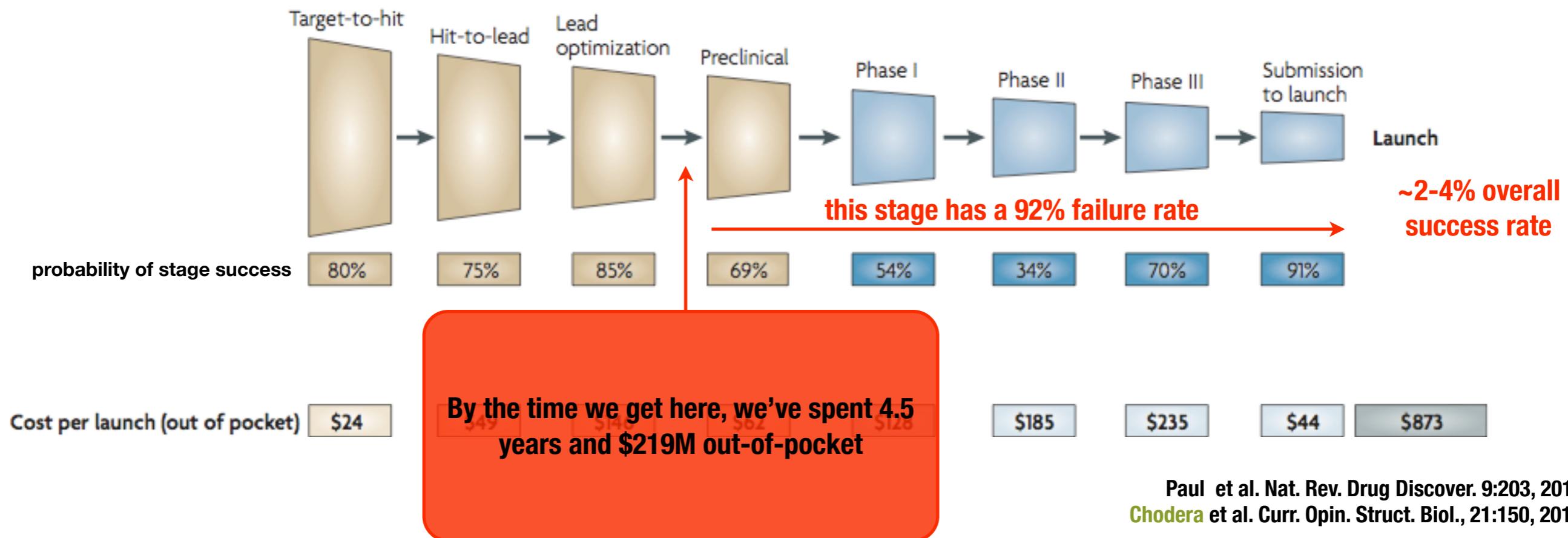
There was great hope this success could be repeated for other cancers by selective targeting of other kinases for cancer.

Usually, drug discovery projects don't go so well

Total pharma research spending has **doubled** to \$65.3B over 2000-2010

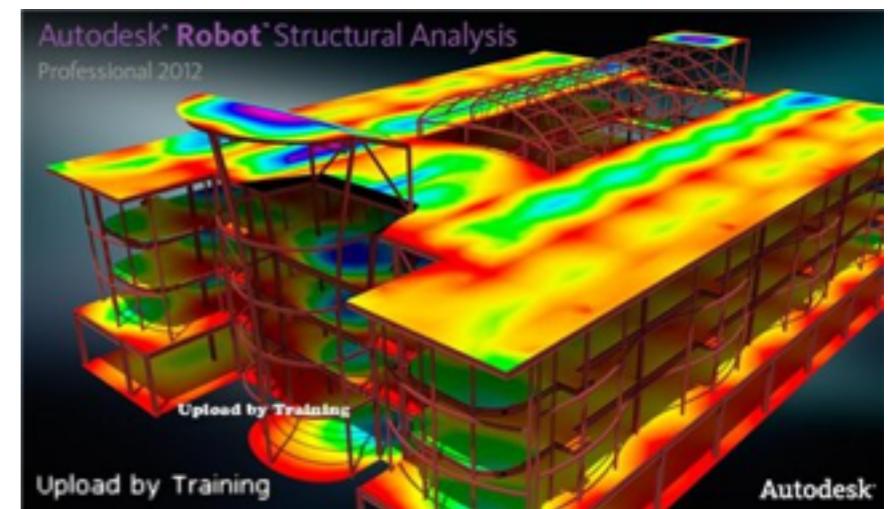
Number of new molecular entities approved by FDA 2005-2009 is **half** that from previous five years

Number of truly innovative new molecular entities has **remained constant** at 5-6/year



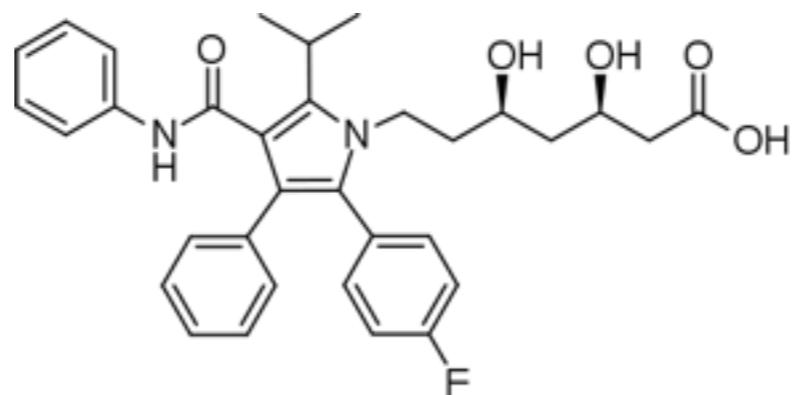
Complex design objectives (efficacy, selectivity, ADME-Tox) make design problem especially complex.

We regularly **design** planes, bridges, and buildings on computers
to satisfy complex design objectives



$10^3 - 10^6$ parts

Why not small molecule drugs?

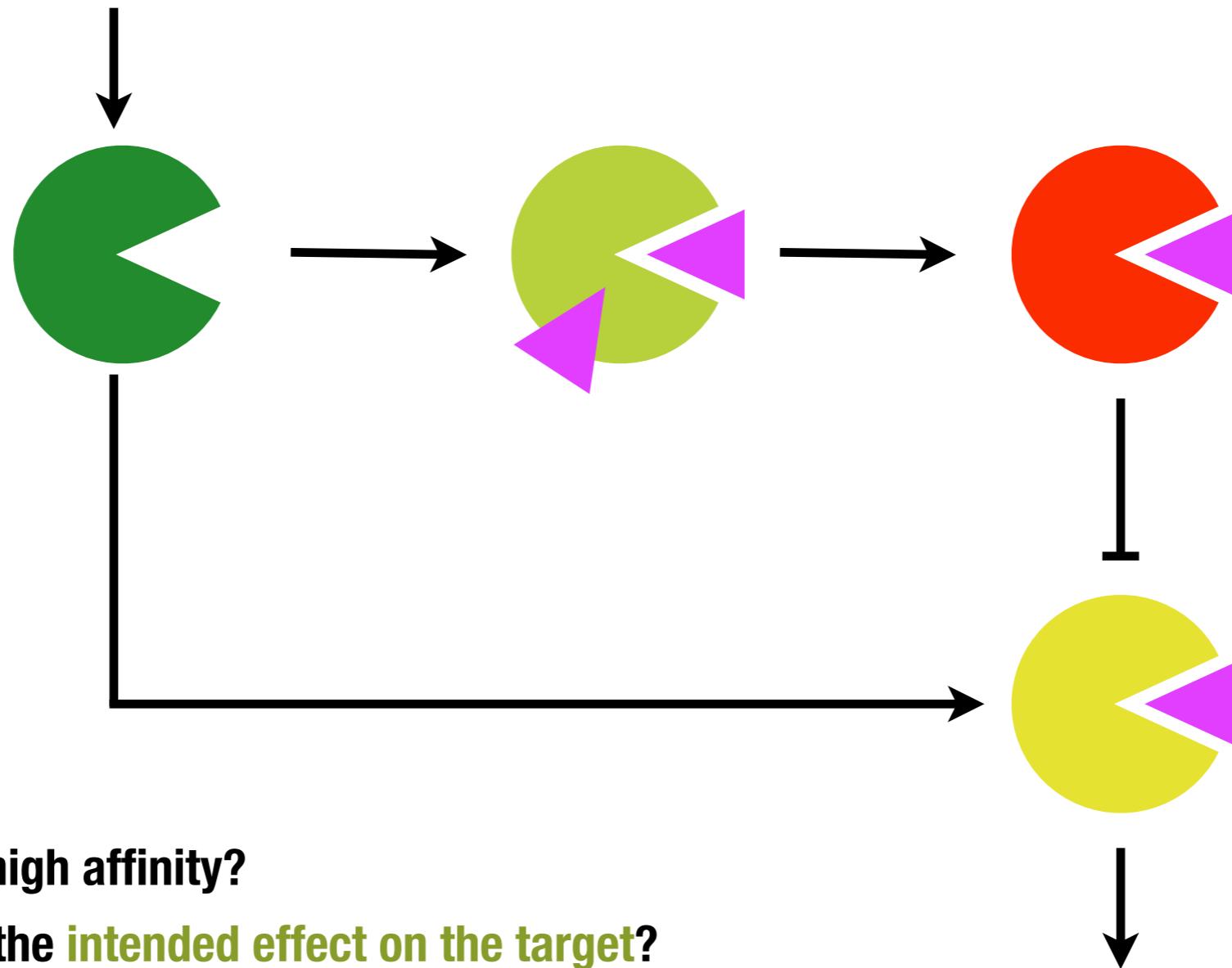


< 10^2 atoms

How can we bring drug design into the 21st century?



To design a small molecule with intended effects, we must **predict** how it will modulate cellular pathways



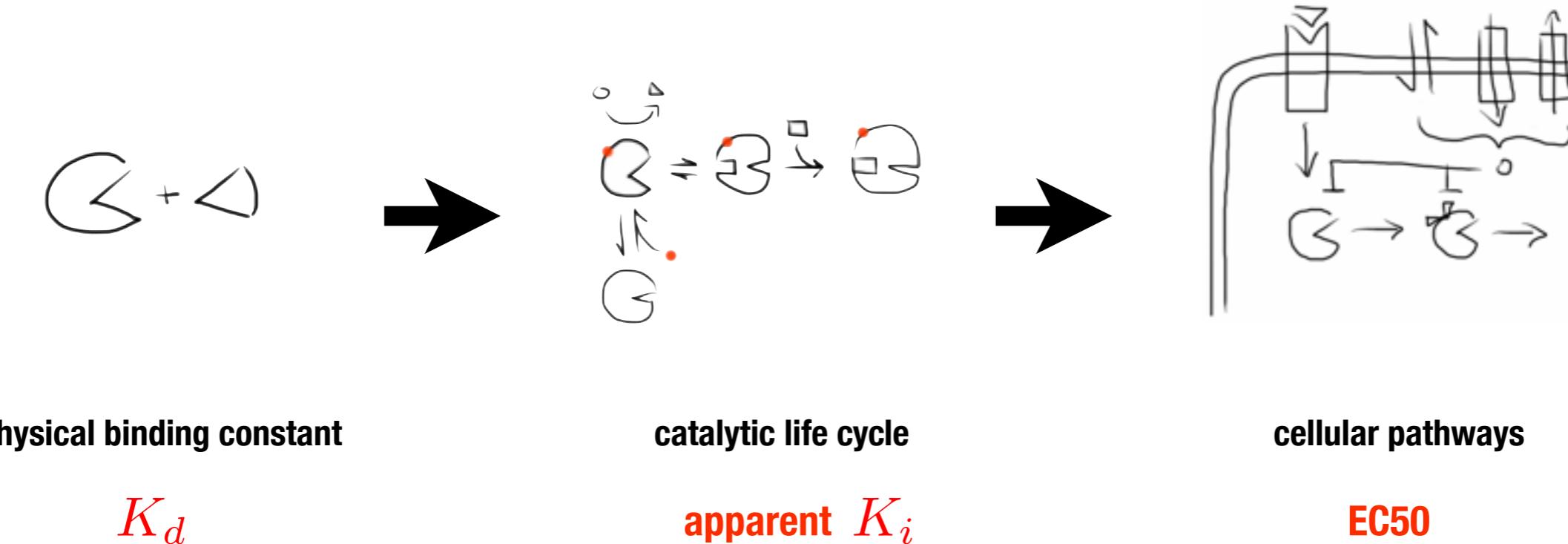
Will it bind the target with high affinity?

Will its binding mode have the intended effect on the target?

Does it produce the desired effect on cellular pathways?

Will it bind unintended targets? Are the resulting effects unacceptably toxic?

Multiscale physical models based on statistical mechanics can potentially aid small-molecule design efforts

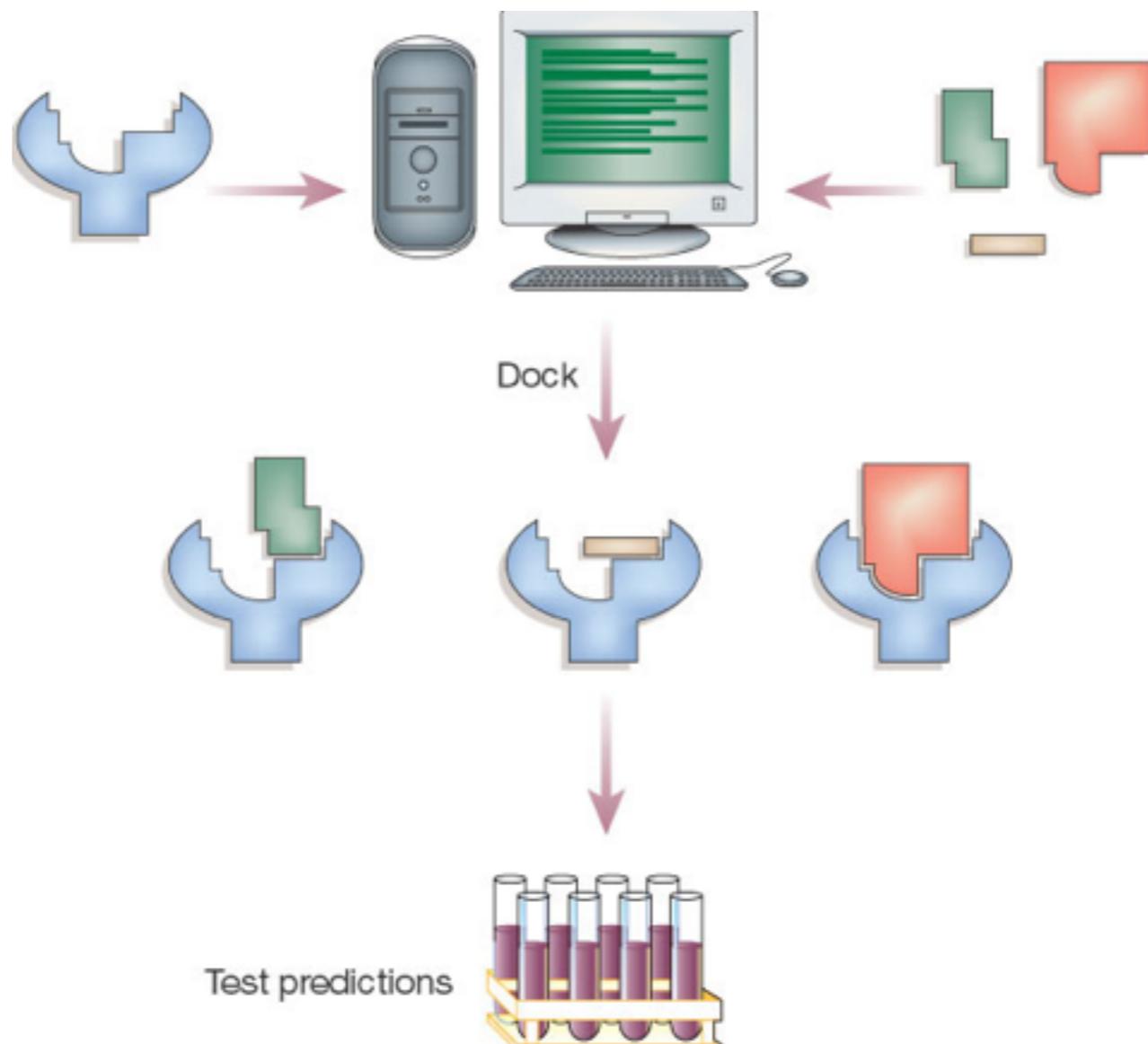


We use **physical modeling** and **statistical mechanics** to build predictive models at each of these scales.

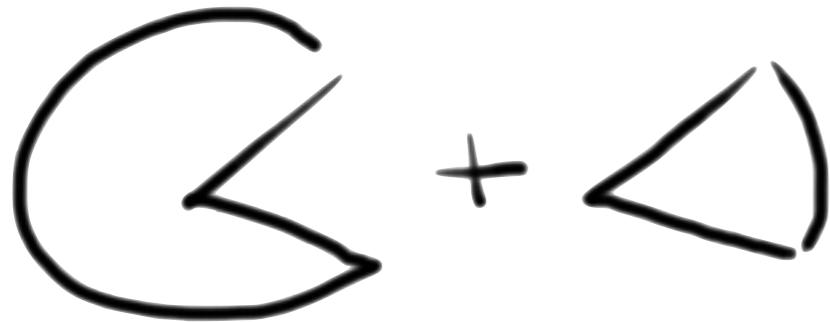
How can we compute binding affinities for molecules that have yet to be synthesized or tested?



Virtual screening methods are in widespread use in drug discovery efforts today. They must work well, right?



How can we compute binding affinities for molecules that have yet to be synthesized or tested?



Virtual screening methods are in widespread use in drug discovery efforts today. They must work well, right?

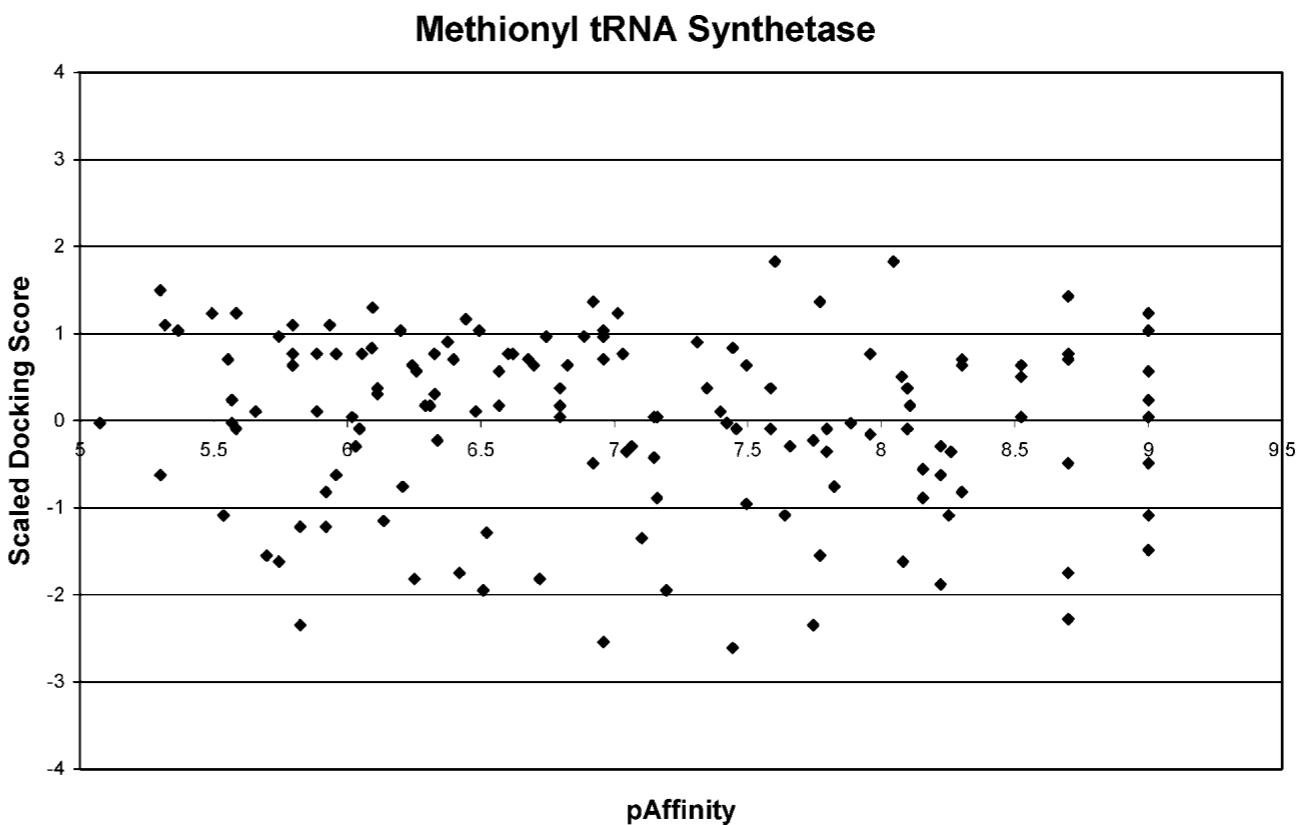
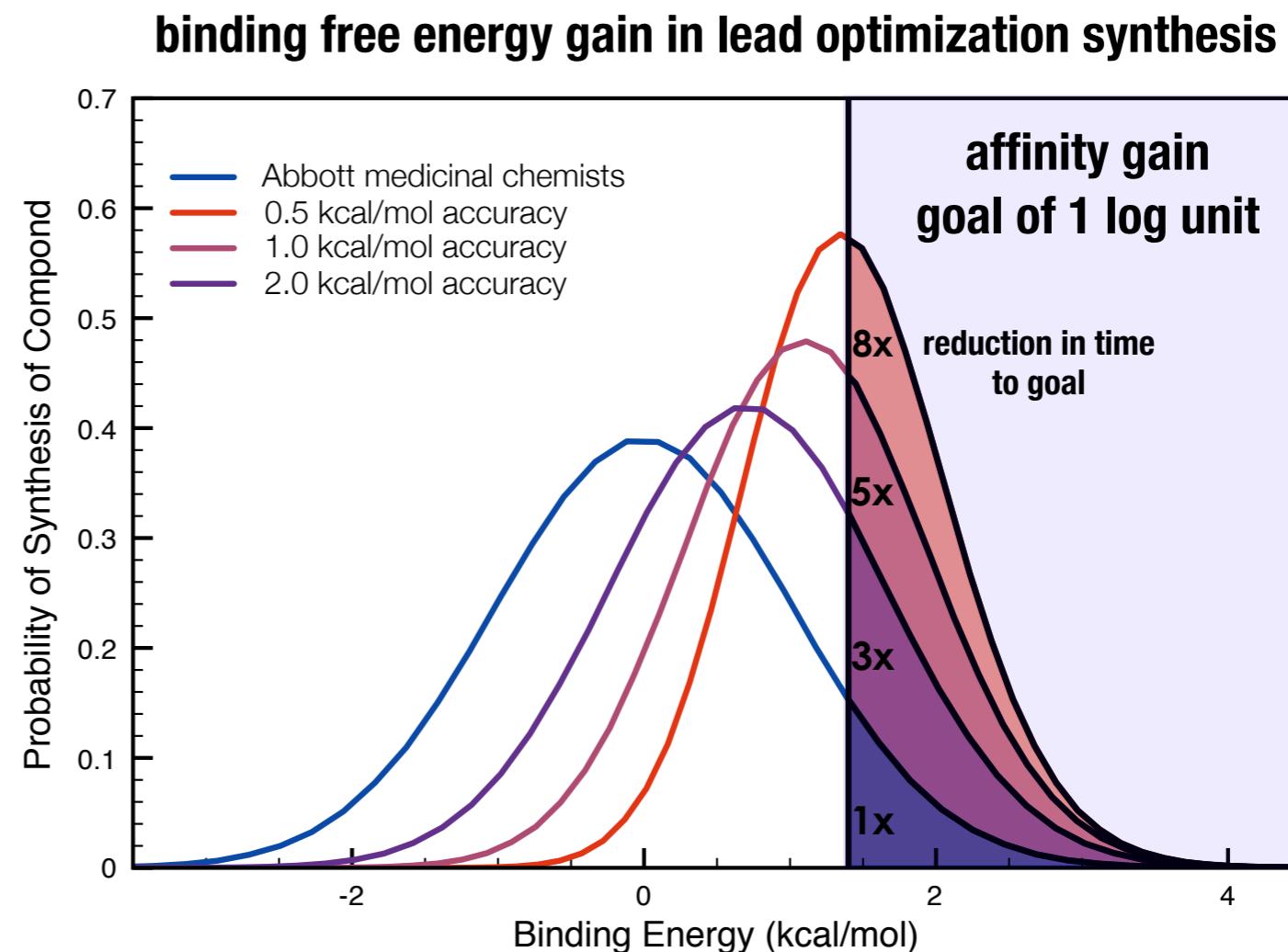
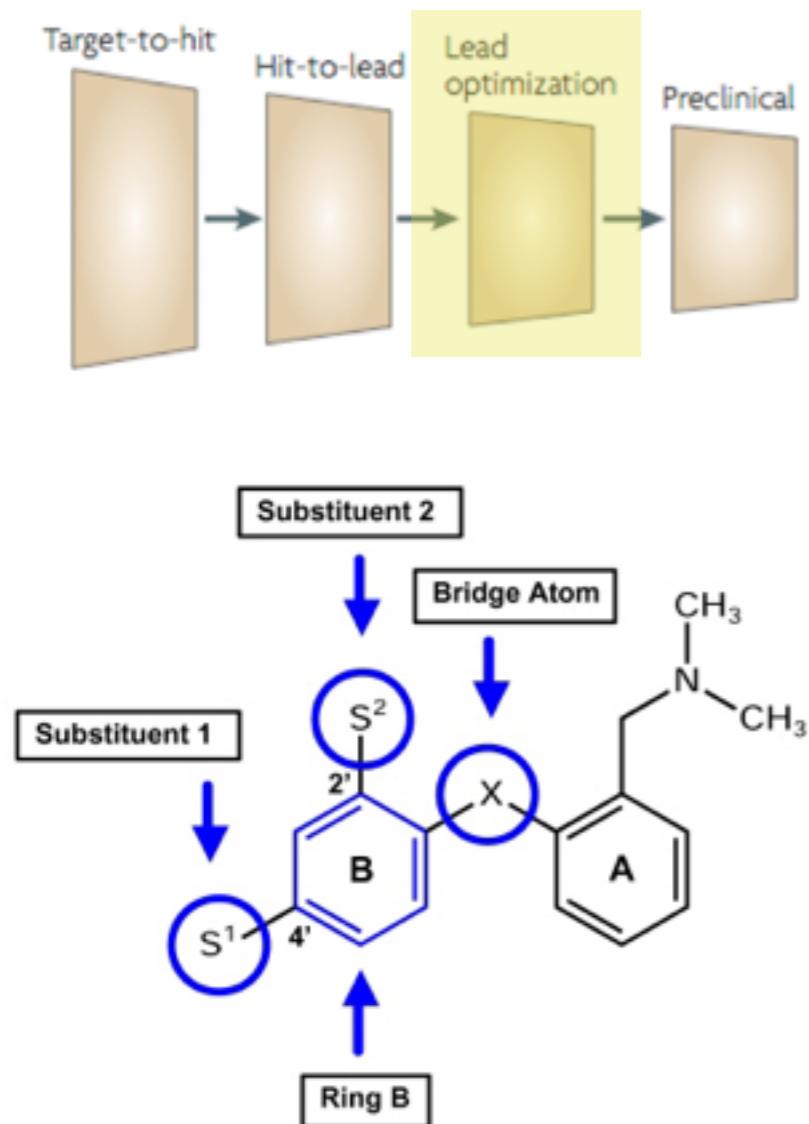


Figure 11. Plot of scaled score vs pAffinity for MRS and PPAR δ . While the calculated correlation coefficient for the data shown for MRS is $r = -0.28$, this plot clearly demonstrates that these values are meaningless. No useful correlation exists between the docking score and compound affinity.

“For prediction of compound affinity, none of the docking programs or scoring functions made a useful prediction of ligand binding affinity.”

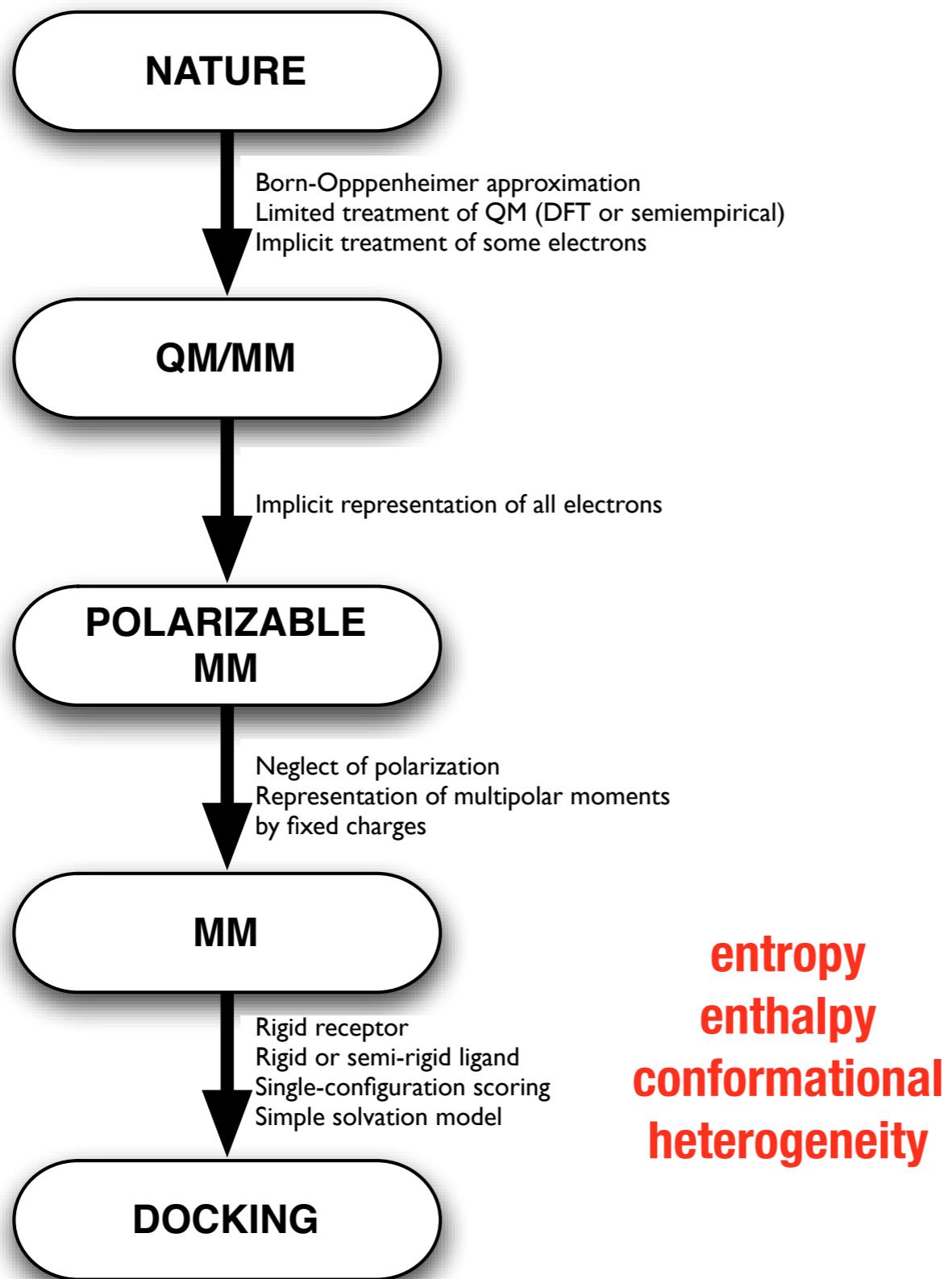
How accurate does a model need to be to aid rational drug design?



A 2 kcal/mol error in prioritizing lead synthesis would speed lead optimization by 3x
but even 10% improvements would be of tremendous benefit

What details are **crucial** for useful accuracy?

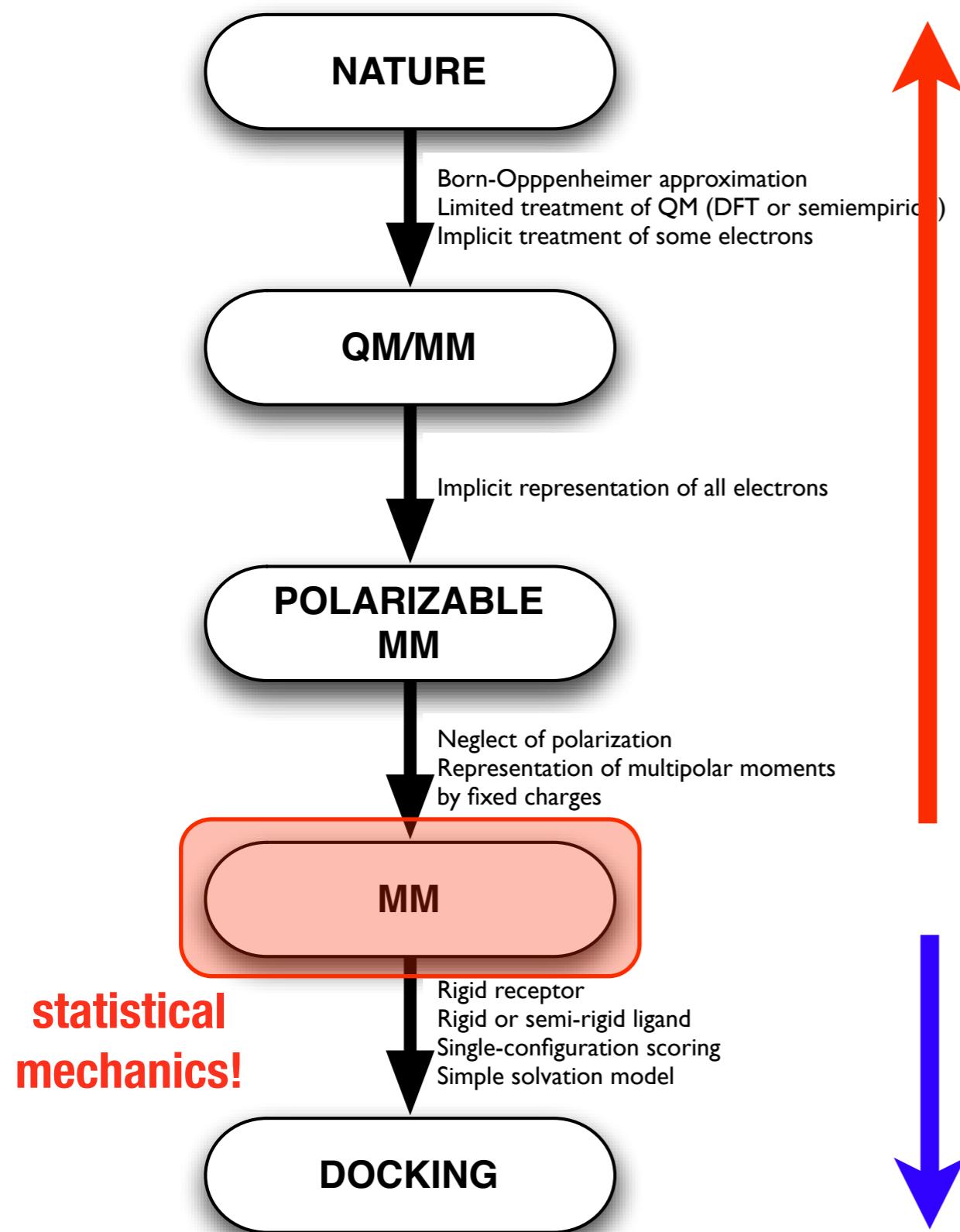
Virtual screening involves a number of approximations



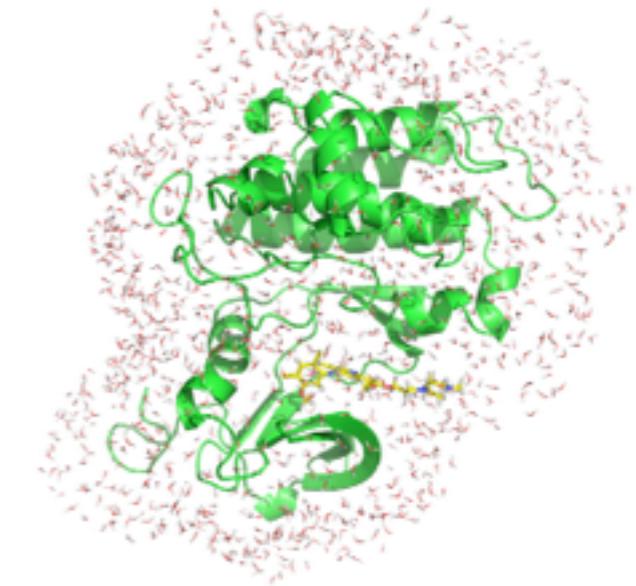
entropy
enthalpy
conformational
heterogeneity

What details are crucial for useful accuracy?

Virtual screening involves a number of approximations



**if insufficiently accurate,
add detail in perturbative
manner**



molecular mechanics potential energy function (e.g. AMBER)

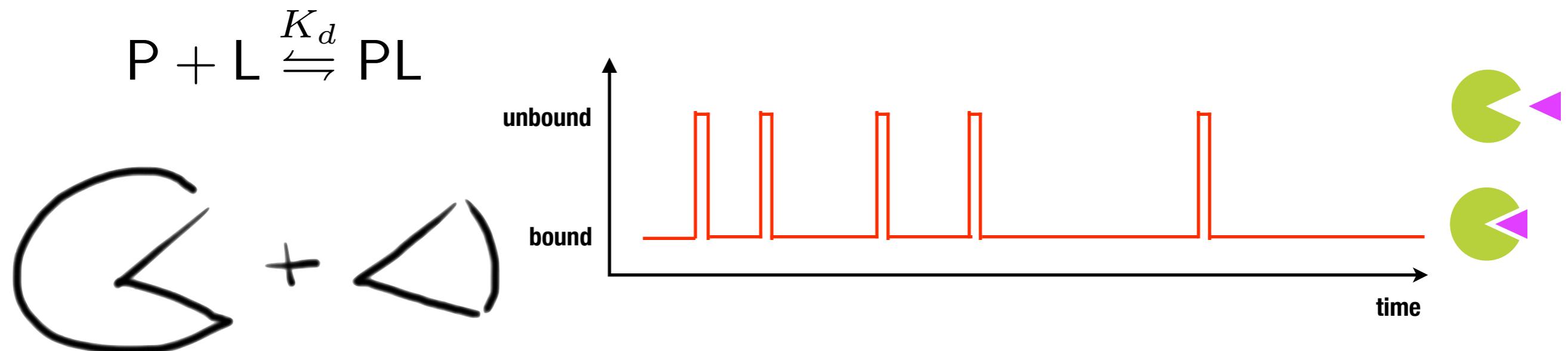
$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

**if accurate enough,
systematically
remove detail**

**statistical
mechanics!**

How can we compute binding affinities that include all relevant **statistical mechanical effects**?

In principle, we could watch many binding/unbinding events to estimate a binding affinity



$$K_d \propto \frac{\tau_{\text{unbound}}}{\tau_{\text{bound}}}$$

ANTON
\$50M special-purpose
supercomputer
(D.E. Shaw Research)

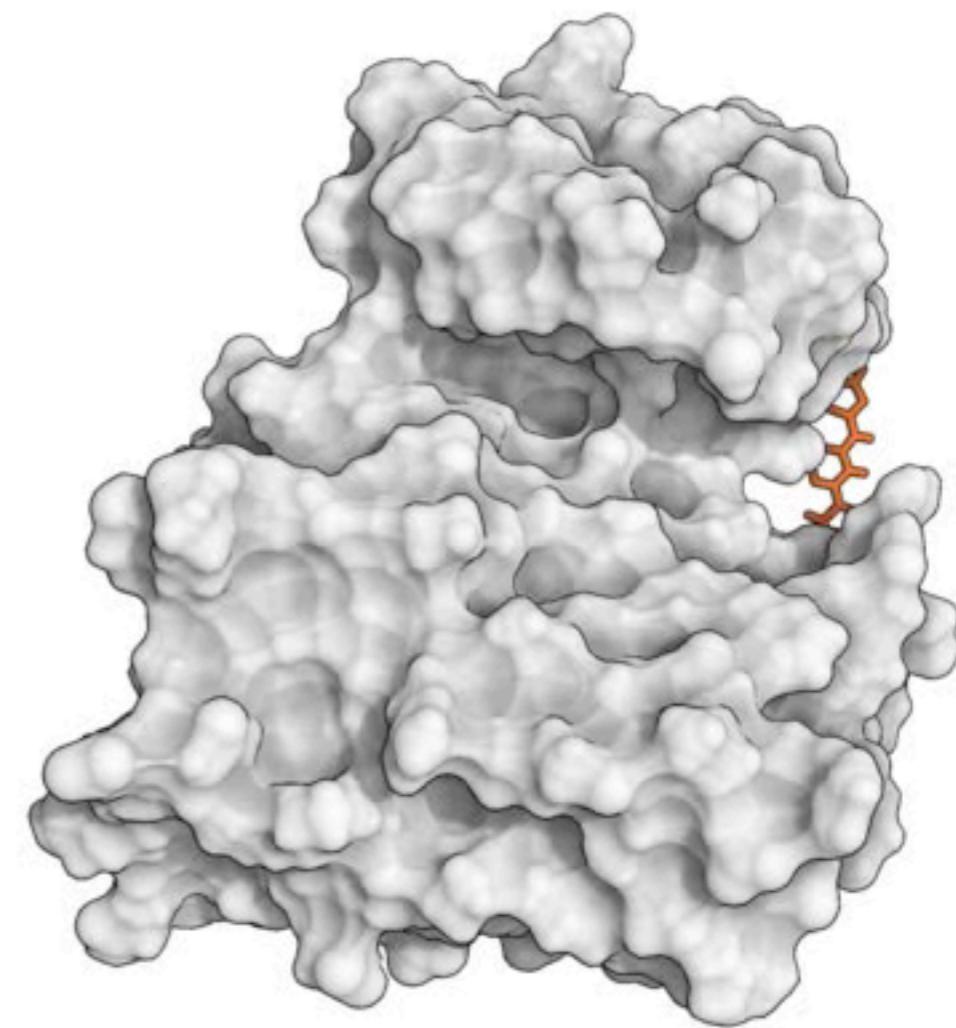


David E. Shaw



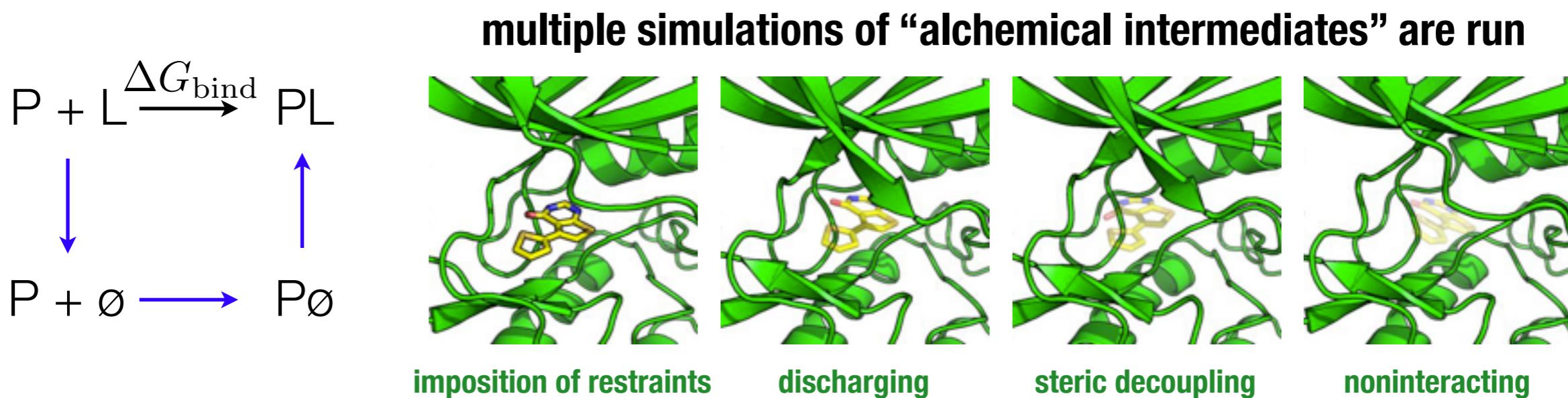
ANTON @ DESRES

Src:dasatanib
(4 ns simulation)



To reliably estimate affinities for molecules with typical off-rates (10^{-4} s^{-1}),
trajectories would need to be impractically long (hours)
requiring $\sim 10^6$ years to simulate.

Alchemical free energy calculations provide a rigorous way to efficiently compute binding affinities for a given forcefield

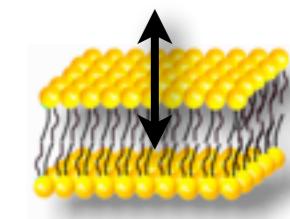
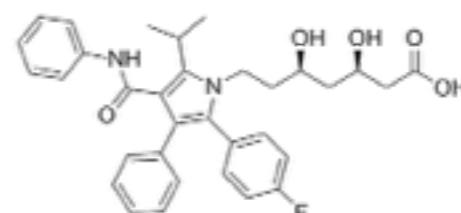


Requires **orders of magnitude** less effort than simulating direct association process, but still includes all enthalpic/entropic contributions to binding free energy.

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

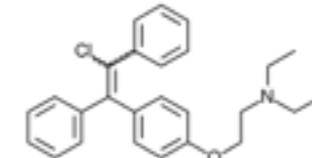
Alchemical free energy calculations can in principle also compute other **relevant physical properties**

partition coefficients (logP, logD) and permeabilities

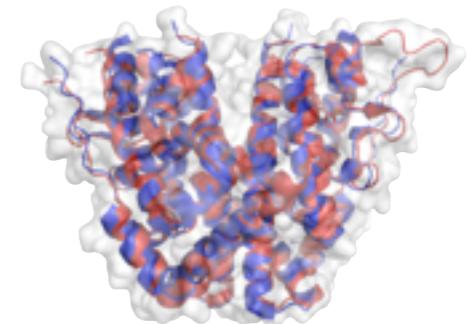


lipitor

selectivity for subtypes or related targets/off-targets



clomifene

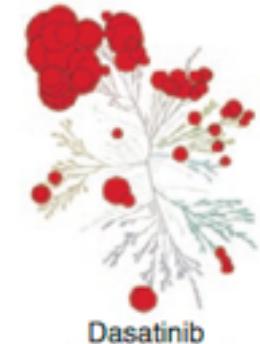


ER α / β

lead optimization of affinity and selectivity

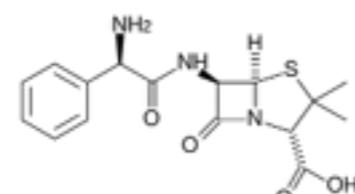


Imatinib

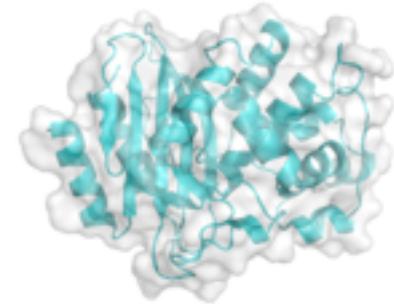


Dasatinib

susceptibility to resistance mutations



ampicillin



β -lactamase

also: solubilities, polymorphs, etc.

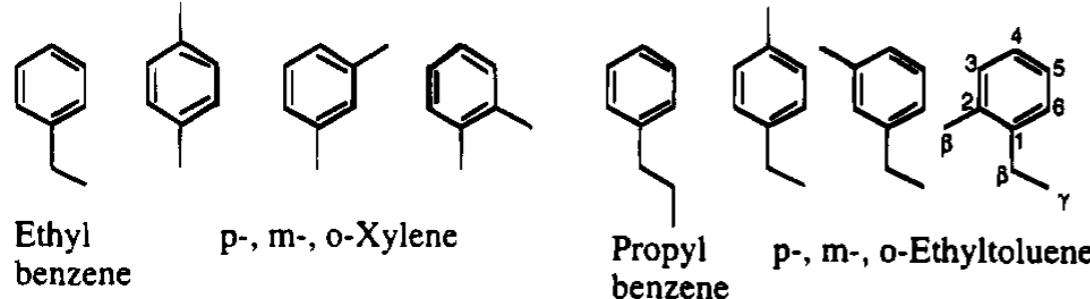


**How accurate are alchemical binding free calculations?
Do they meet this 2 kcal/mol threshold for useful accuracy?**

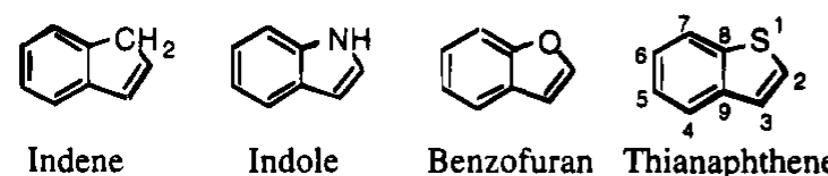


How accurate are binding free energy calculations? T4 lysozyme L99A as a simple model binding site

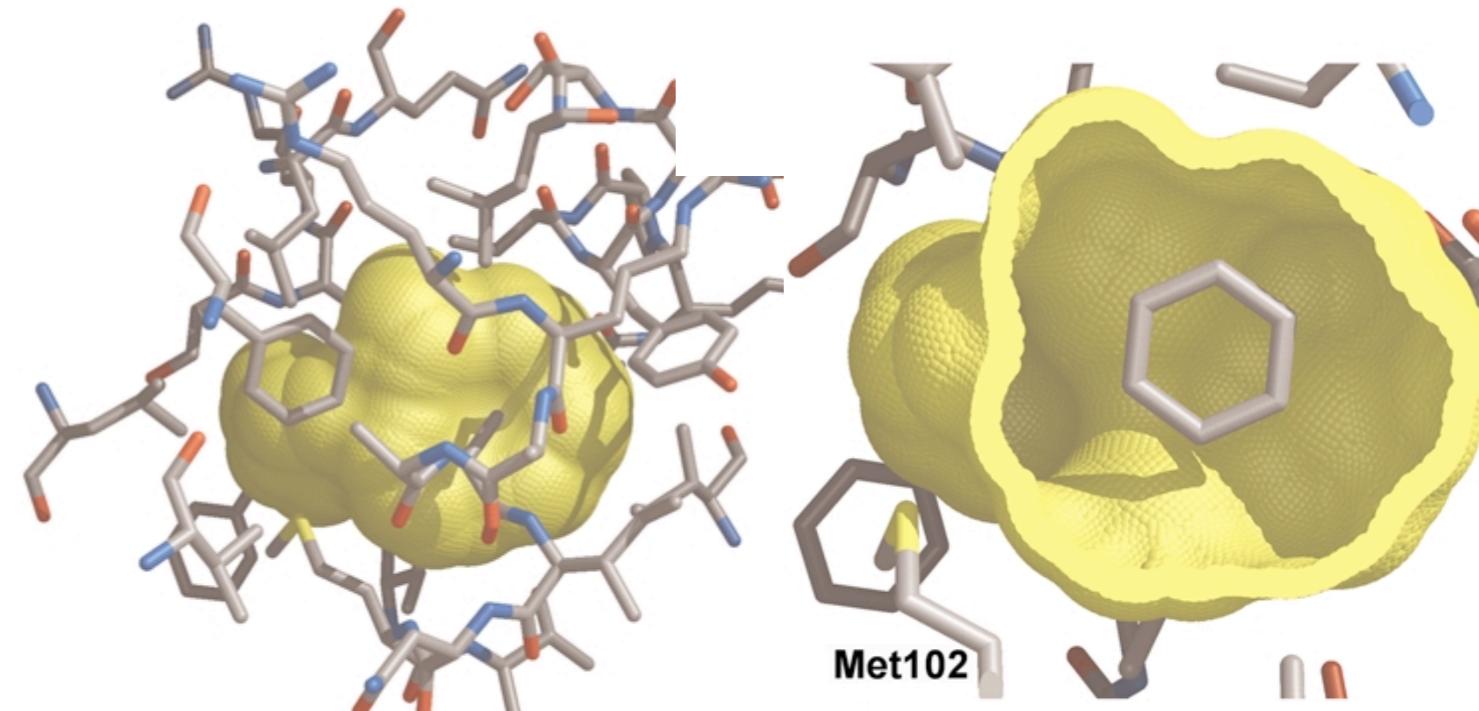
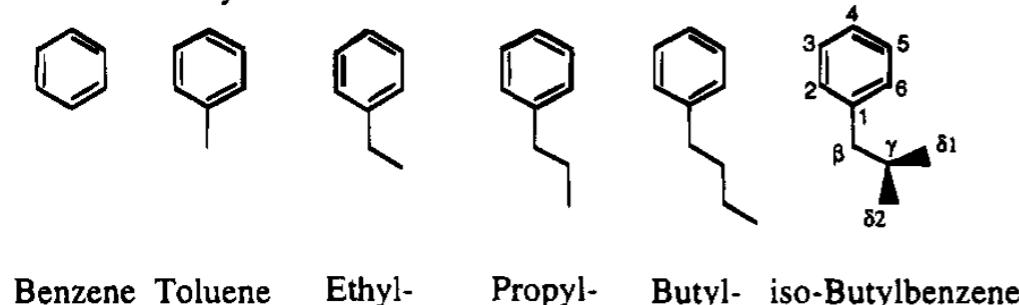
Class I - "Isophobic" Ligands



Class II - "Isosteric" Ligands



Class III - Phenylalkanes



Wei, Baase, Weaver, Matthews, and Shoichet. JMB 322:339, 2002.

**Surprisingly challenging for docking codes to discriminate binders (μM) from non-binders ($>>\mu\text{M}$)
Polar version of this cavity (L99A/M102Q) even more challenging for docking.**

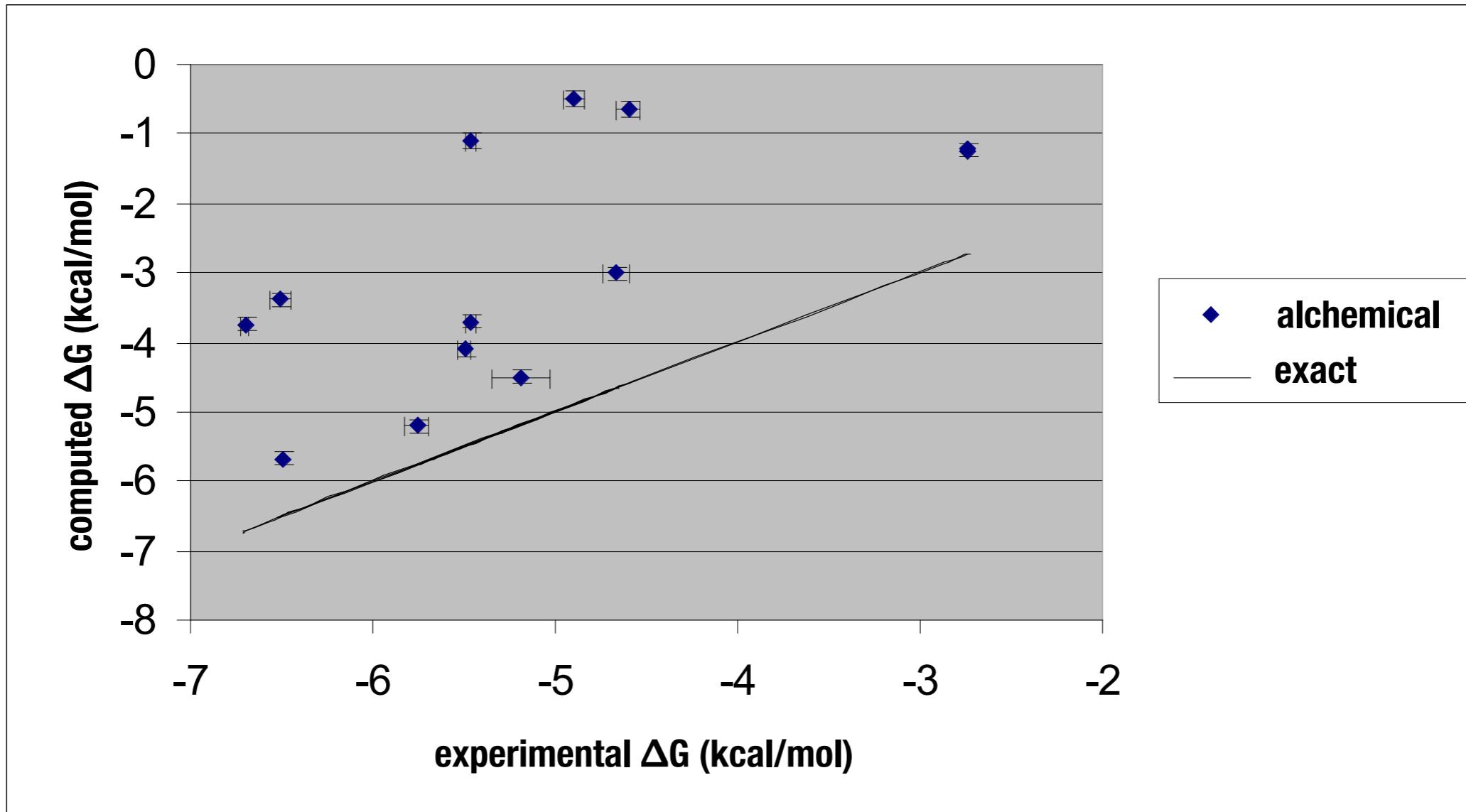
Let's test how well free energy calculations can reproduce measured binding free energies.

"Surely, there's no way we can screw this up."

- Famous last words

“Nothing is foolproof to a sufficiently talented fool”

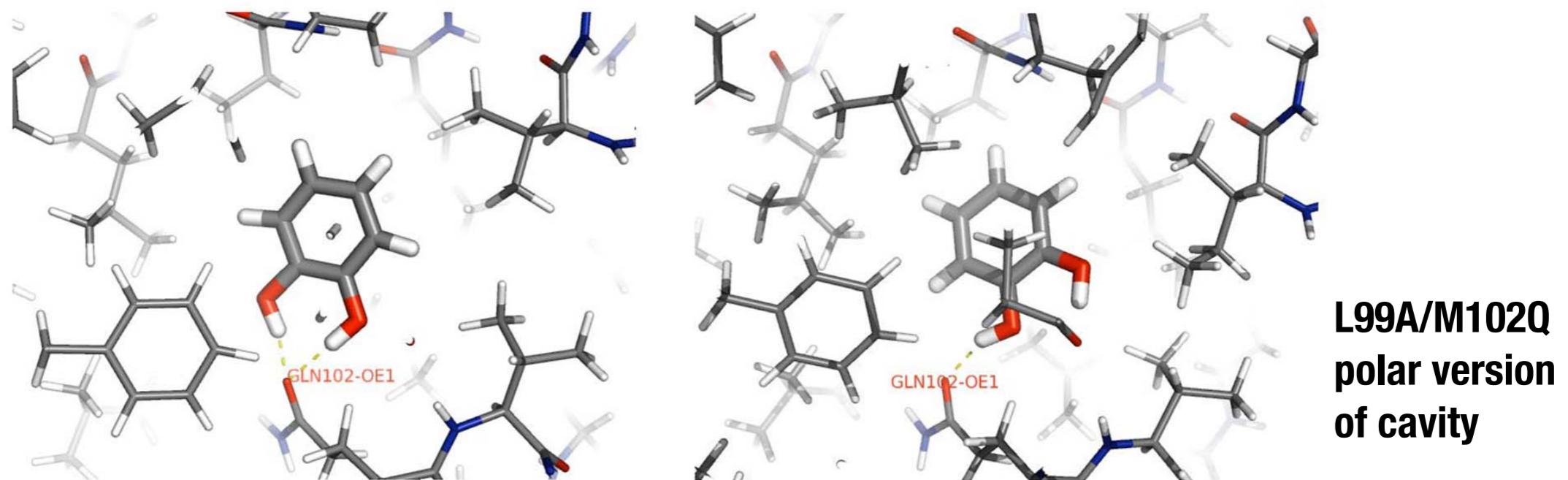
- Old but pertinent philosophy



Obviously, we're missing something.
What's going on here?

Multiple **ligand** conformations can contribute

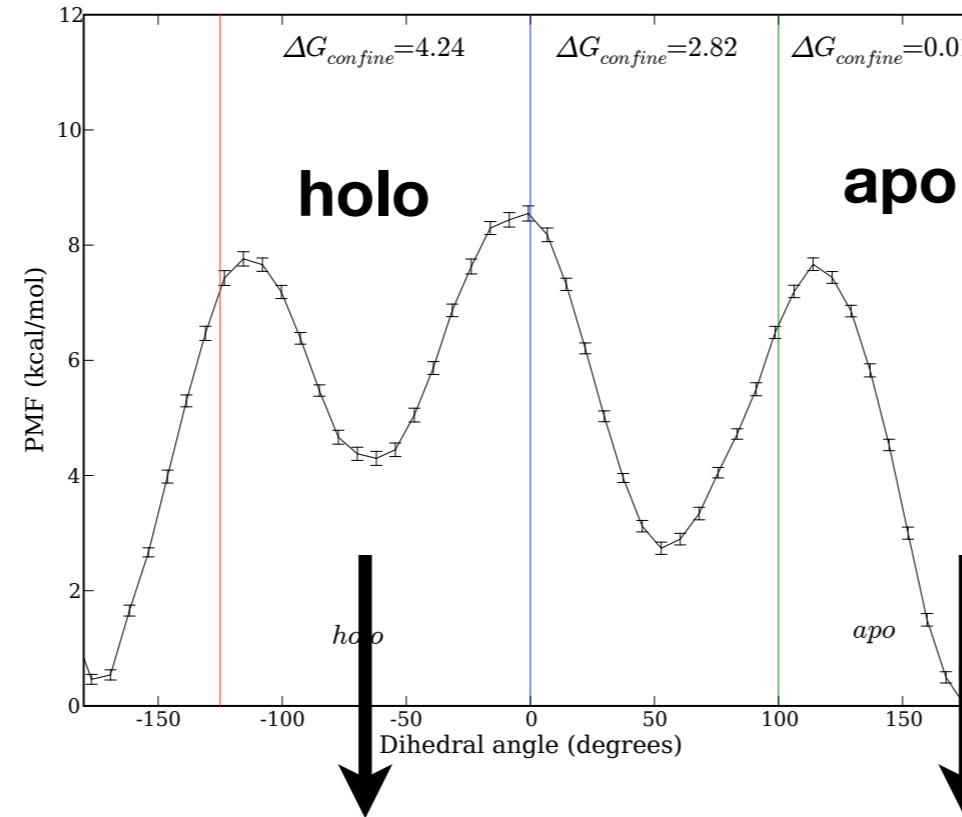
Switching between relevant ligand orientations can occur on a timescale of many nanoseconds!



Difference in affinity between different bound orientations is only $\sim 1 \text{ kT}$
N poses can contribute to the overall binding free energy $\sim \text{kT} \ln N$
[Also relevant for ligands with pseudosymmetric substituents]

Multiple protein conformations can contribute

Val111 sidechain χ_1 in apo structure



binding free energy

-7.3 kcal/mol

-3.0 kcal/mol

4.3 kcal/mol difference!

confinement free energy

4.2 kcal/mol

0.0 kcal/mol

**release free energy
following binding**

-0.3 kcal/mol

-0.6 kcal/mol

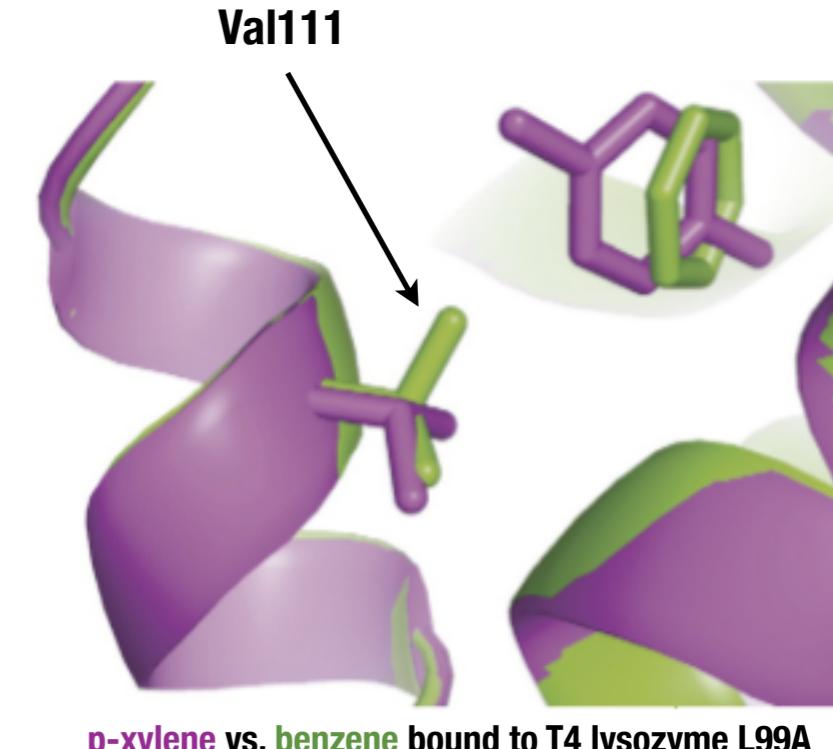
total binding free energy

-3.4+-0.3

\approx

-3.6+-0.3

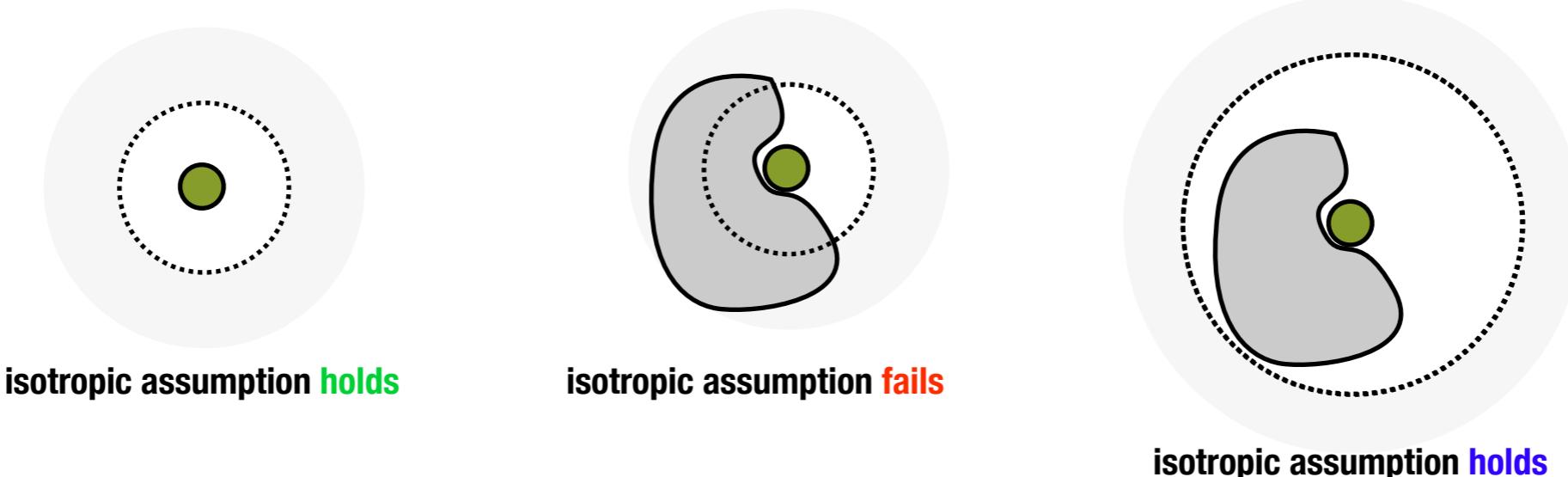
$\Delta G_{exp} = -4.7 \text{ kcal/mol}$



Anisotropic long-range dispersion correction

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

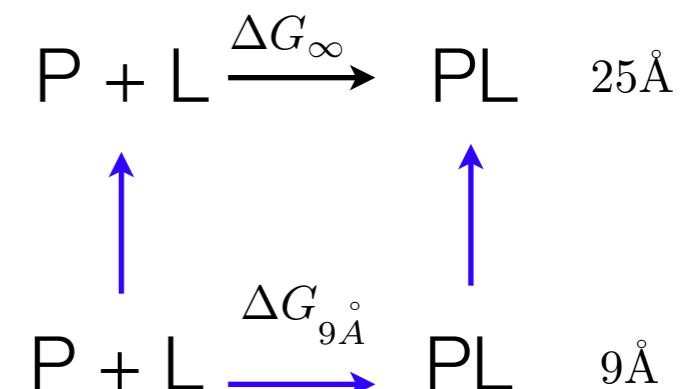
This correction assumes **isotropic** distribution of Lennard-Jones sites throughout system, but protein/water mixtures are not homogeneous and isotropic!



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



Numerous improvements were required for T4 lysozyme L99A

- ✓ multiple long-lived ligand orientations Mobley, Chodera, Dill. JCP 125:084902, 2006.
- ✓ multiple long-lived protein conformations Mobley, Chodera, and Dill. JCTC 3:1231, 2007.
- ✓ anisotropic dispersion correction Shirts, Mobley, Chodera, Pande. JPC B 111:13052, 2007.
- ✓ optimal use of all data in analysis Shirts and Chodera. JCP 129:124105, 2008.
- ✓ binding site restraints to reduce simulation times Mobley, Chodera, Dill. JCP 125:084902, 2006.
- ✓ improved ligand charge models Mobley, Dumont, Chodera, Dill. 111:2242, 2007.

These issues are very general, and algorithmic improvements that address them should universally improve accuracy for protein-ligand systems.

Resulting RMS error: 1.89 ± 0.04 kcal/mol [originally 3.51 ± 0.04 kcal/mol]

But it's easy to fool ourselves when working with a known dataset.
How well do we do on data we've never seen?

How accurately can we predict unknown binding affinities?

Brian Shoichet, UCSF



Blinded test of prediction power with new molecules:

Ligand	Prediction ¹	ΔG_{calc}^o ² (kcal/mol)
1,2-dichlorobenzene	Binder	-5.66 ± 0.15
n-methylaniline	Binder	-5.37 ± 0.11
1-methylpyrrole	Binder	-4.32 ± 0.08
1,2-benzenedithiol	Binder	-2.79 ± 0.13
thieno-[2,3-c]pyridine	Nonbinder	-2.56 ± 0.07

How accurately can we predict unknown binding affinities?

Brian Shoichet, UCSF



Blinded test of prediction power with new molecules:

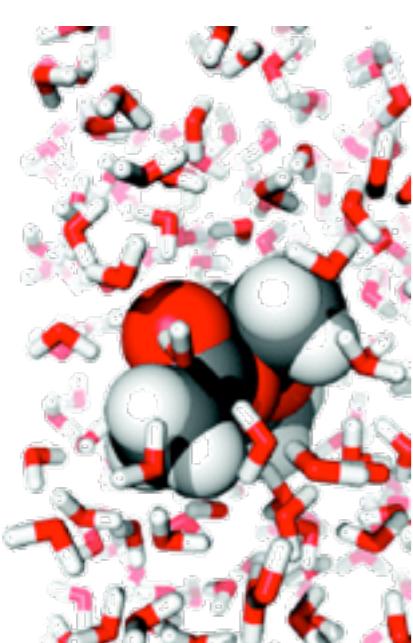
Ligand	Prediction ¹	ΔG_{calc}^o ² (kcal/mol)	Experiment	$\Delta G_{expt.}^o$ (kcal/mol)
1,2-dichlorobenzene	Binder	-5.66 ± 0.15	Binder	-6.37
n-methylaniline	Binder	-5.37 ± 0.11	Binder	-4.70
1-methylpyrrole	Binder	-4.32 ± 0.08	Binder	-4.44
1,2-benzenedithiol	Binder	-2.79 ± 0.13	Binder	N.D.
thieno-[2,3-c]pyridine	Nonbinder	-2.56 ± 0.07	Nonbinder	N.D.

All binding predictions confirmed!
RMS error: 0.6 kcal/mol

(but we could only convince them to do N=3 ITC measurements)

Alchemical free energy methods can work reliably in simple systems but complex systems remain challenging

model systems ← → pharmaceutically relevant



hydration free energies
of small neutral molecules

1.04 ± 0.03 kcal/mol (N=44)

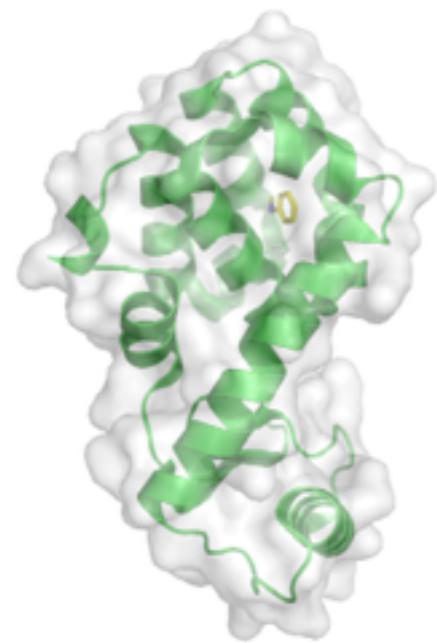
Mobley, Dumont, Chodera, Dill. JPC B, 2007

1.23 ± 0.01 kcal/mol (N=502)

Mobley, Bayly, Cooper, Dill. JPC B 2009.

1.33 ± 0.05 kcal/mol (N=17)

Nicholls, Mobley, Guthrie, Chodera, Pande. J Med Chem 2008.



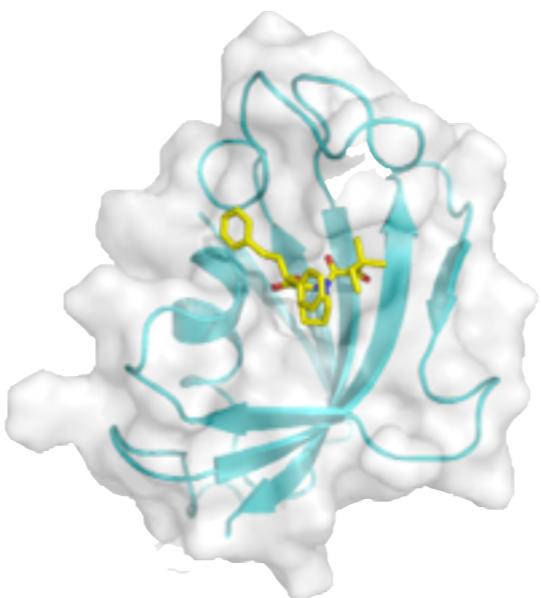
small apolar ligands
T4 lysozyme L99A

1.89 ± 0.04 kcal/mol (N=13)

Mobley, Graves, Chodera, McReynolds, Shoichet Dill. JMB 2007

0.6 ± 0.2 kcal/mol (N=3)

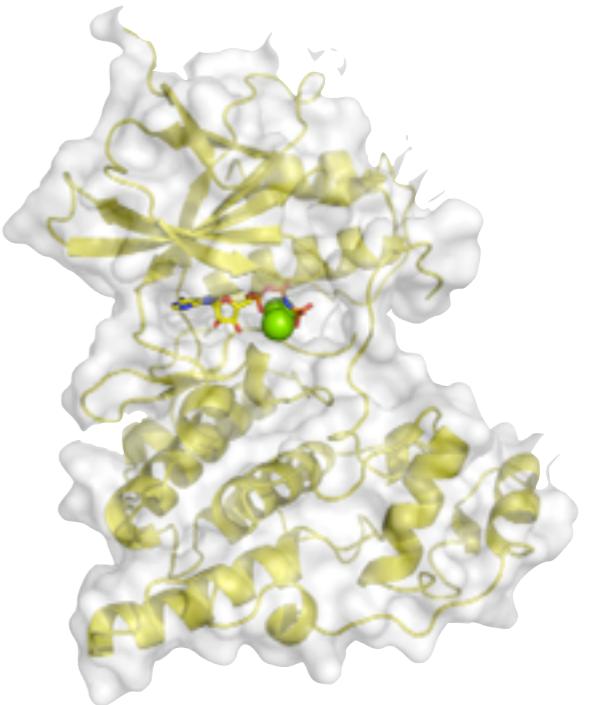
Mobley, Graves, Chodera, McReynolds, Shoichet Dill. JMB 2007



polar ligands
FKBP12

1.42 kcal/mol (N=9)

0.94 kcal/mol (N=7)
(Shirts et al., in preparation)



JNK3 kinase

Anecdotal literature reports of success
(publication bias?)

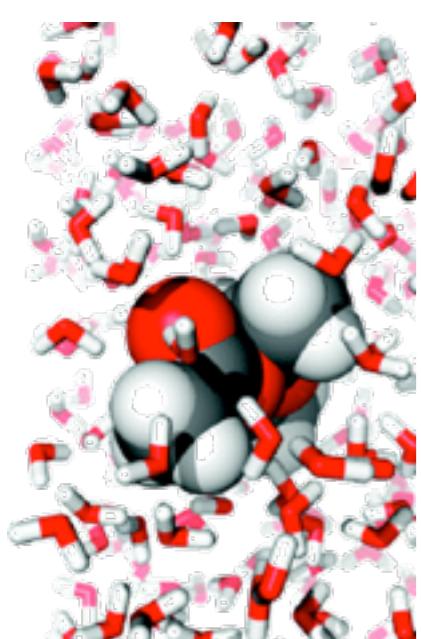
Calculations are notoriously
unreliable.
(e.g. SAMPL challenges)

retrospective RMS error [sample size]
prospective RMS error [sample size]

(not to scale)

Alchemical free energy methods work reliably in simple systems but complex systems have remained challenging

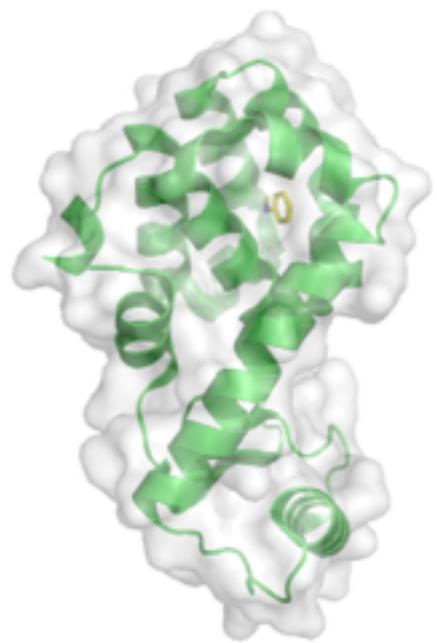
model systems ← → pharmaceutically relevant



hydration free energies
of small neutral molecules

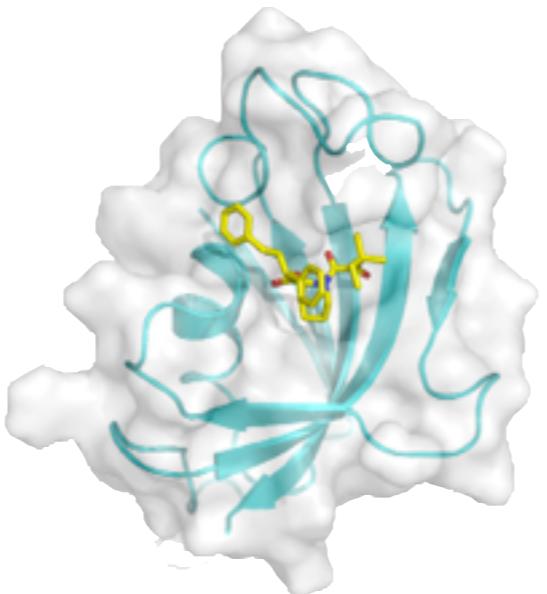
solvent only
small, neutral molecules
fixed protonation states

easy
hard



small apolar ligands
T4 lysozyme L99A

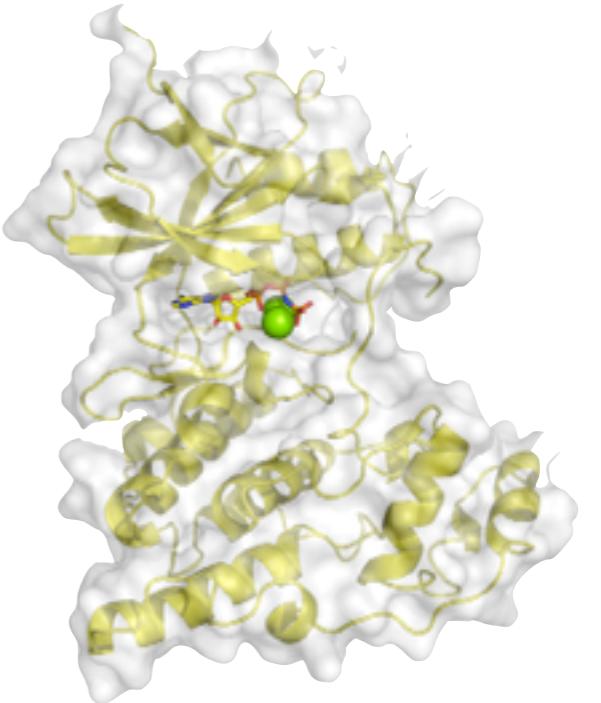
small, rigid protein
small, neutral ligands
fixed protonation states
multiple sidechain orientations
multiple ligand binding modes



polar ligands
FKBP12

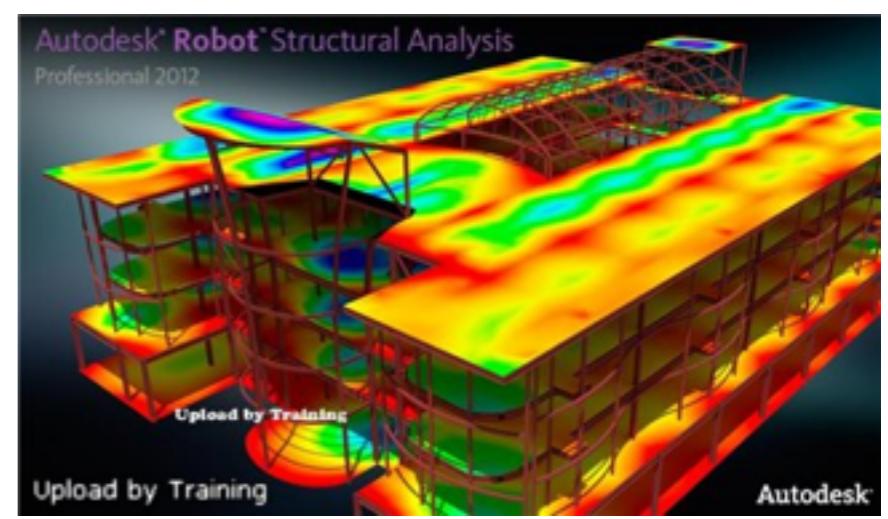
small, rigid protein
fixed protonation states
larger drug-like ligands
with few rotatable bonds

(not to scale)



JNK3 kinase

large protein, multiple conformations
large drug-like ligands, rotatable bonds
multiple protonation states? tautomers?
phosphorylation and activation
peptide substrate?
MgCl₂ salt effects?



Structural engineering wasn't always so successful, either



There were 250 bridge failures in the US and Canada between 1878-1888.

“The subject of mechanical pathology is relatively as legitimate and important a study to the engineer as medical pathology is to the physician. While we expect the physician to be familiar with physiology, without pathology he would be of little use to his fellow-men, and it [is] as much within the province of the engineer to investigate causes, study symptoms, and find remedies for mechanical failures as it is to direct the sources of power in nature for the use and convenience of man.”

- George Thomson, 1888

Computational predictions fail for one of three reasons

1. The **forcefield** does a poor job of modeling the physical system
2. We're missing some **essential chemical effects** in our simulations
(e.g. protonation states, tautomers, covalent association)
3. We haven't **sampled** all of the relevant conformations

We need to figure out **why failures occur** and how we can improve our algorithms to be more robust for prediction.

Achieving reliable accuracy on small systems required cycles of prediction and experiment

computational
predictions



experimental
confirmation

A successful collaboration between experimentalists and theorists:

- * Mobley, Graves, Chodera, McReynolds, Shoichet, Dill. J Mol Biol 371:1118, 2007
- * Nicholls, Mobley, Guthrie, Chodera, Bayly, Cooper, Pande. J Med Chem 51:769, 2008.
- * Boyce, Mobley, Rocklin, Graves , Dill, and Shoichet. JMB 394:747, 2009.

Learning from failures led directly to improvements in:

- * algorithms to speed conformational sampling
- * treatment of physical effects
- * forcefields

But cycle time is ~ 1 year!

- * Calculations take cluster-weeks.
- * Synthesis of new compounds time-consuming and costly.

How can we speed this up?

Speeding up the cycle of learning from failure can accelerate progress toward rational ligand design

computational predictions



experimental confirmation

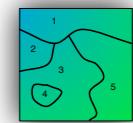
“Fail fast, fail cheap”

Make rigorous calculations of affinity fast and accurate



GPU acceleration

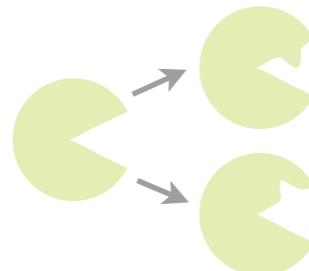
$$\pi(x)K(x,y) = \pi(y)K(y,x)$$



enhanced sampling algorithms

modular MCMC for sampling and chemical effects

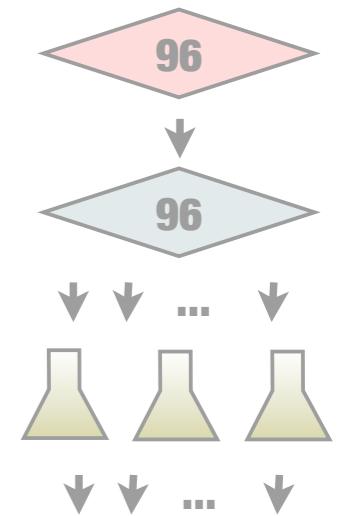
Test and improve models quickly and cheaply by inverting the drug discovery problem



mutate proteins instead of ligands



buy inexpensive ligands



high-throughput experiments

How can we speed up the calculations?



**\$50K
3 TFLOP
doubling every 18 mo**

many CPU-weeks/calculation

Doesn't fit neatly in a synthetic chemist's timeframe to wait weeks for an answer.



How can we speed up the calculations?



\$50K
5 TFLOP
doubling every 18 mo

many CPU-weeks/calculation



\$500
5 TFLOP
doubling every 12 mo

overnight on a workstation?

Can we exploit new GPU technologies to reach practical computation times?

YANK: An open-source, community-oriented platform for GPU-accelerated free energy calculations



NVIDIA GTX-Titan (\$1000)

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://simtk.org/home/openmm>

gromacs benchmarks from <http://biowulf.nih.gov/apps/gromacs-gpu.html>

The screenshot shows the SimTK.org website with the YANK project page. The header includes the SimTK logo, navigation links for Home, About Simtk.org, How to Contribute, Advanced Search, News, Create Project, and Log In. The main content features a large image of a soldier from a World War II era comic book, with the text "YANK THE ARMY MAGAZINE" and the tagline "WE WILL ACCEPT NOTHING LESS THAN FULL VICTORY". The page title is "YANK: GPU-accelerated calculation of ligand binding affinities" and the subtitle is "Project Overview". On the left, there's a sidebar with links for Overview, Team, Downloads, Documents, Publications, Advanced, and Downloads & Source Code. The "Downloads & Source Code" section notes that the project has no public downloads but provides source code. The main text describes YANK as a code for estimating free energies of ligand binding using free energy perturbation methods, utilizing the OpenMM library for GPU-accelerated molecular dynamics. It aims to accelerate calculations and make them simple enough to run while maintaining rigor. On the right, there are profiles for the Project Lead (John Chodera), Kim Branson, and Michael Shirts, each with a photo, contact link, and a "Contact" button.

Initially, YANK will focus on free energy calculations in implicit solvent, being extended to explicit solvent as OpenMM adds support for long-range electrostatics treatments such as PME. While not as accurate as explicit solvent simulations, implicit solvent offers a number of advantages (including straightforward constant pH treatments, enhanced conformational sampling, and simpler incorporation of large-scale Monte Carlo moves), and the speed advantages offered by GPU acceleration are expected to provide significant utility for medicinal chemists, chemical biologists, and biochemists.

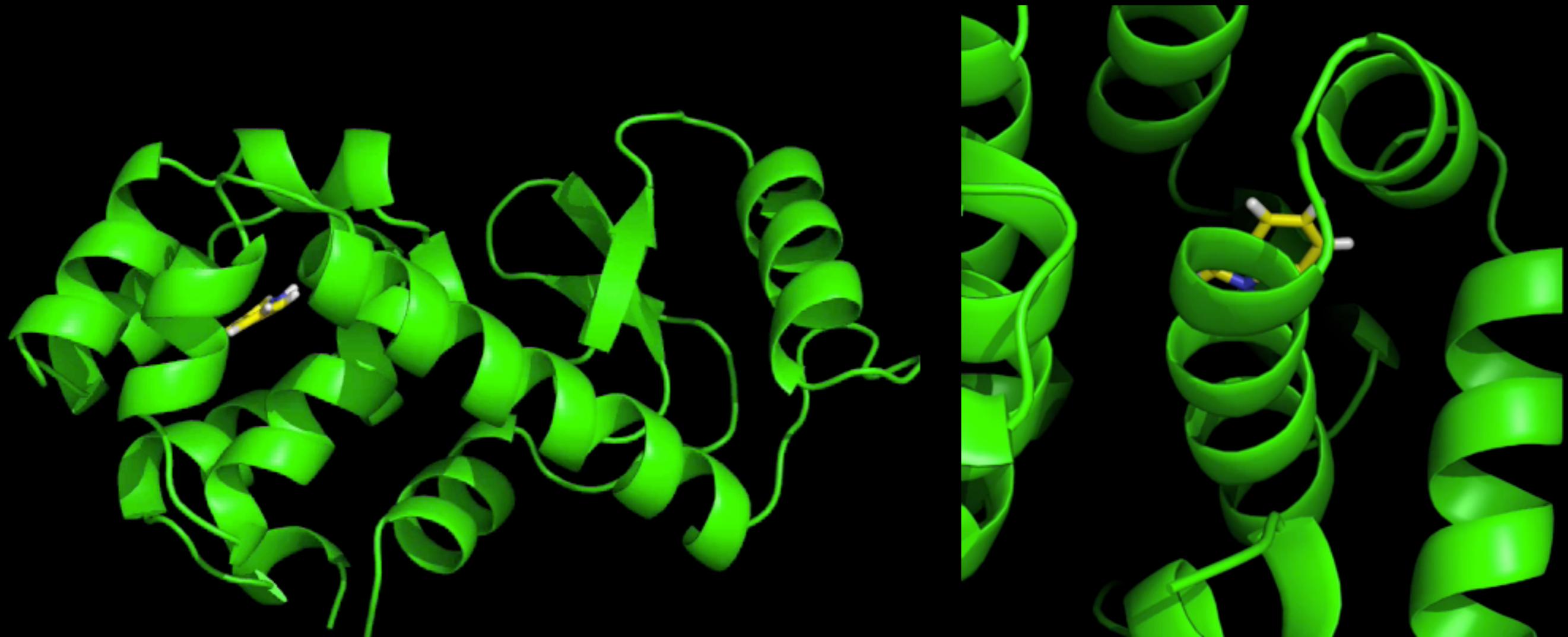
We are working toward a first public release of YANK. We are currently still testing the code and completing detailed documentation on its use. Later revisions will also feature conversion of non-time-critical portions into Python, to facilitate expansion and incorporation into other applications.

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

<http://github.com/choderalab/yank>

Replica-exchange algorithms facilitate sampling of multiple binding modes

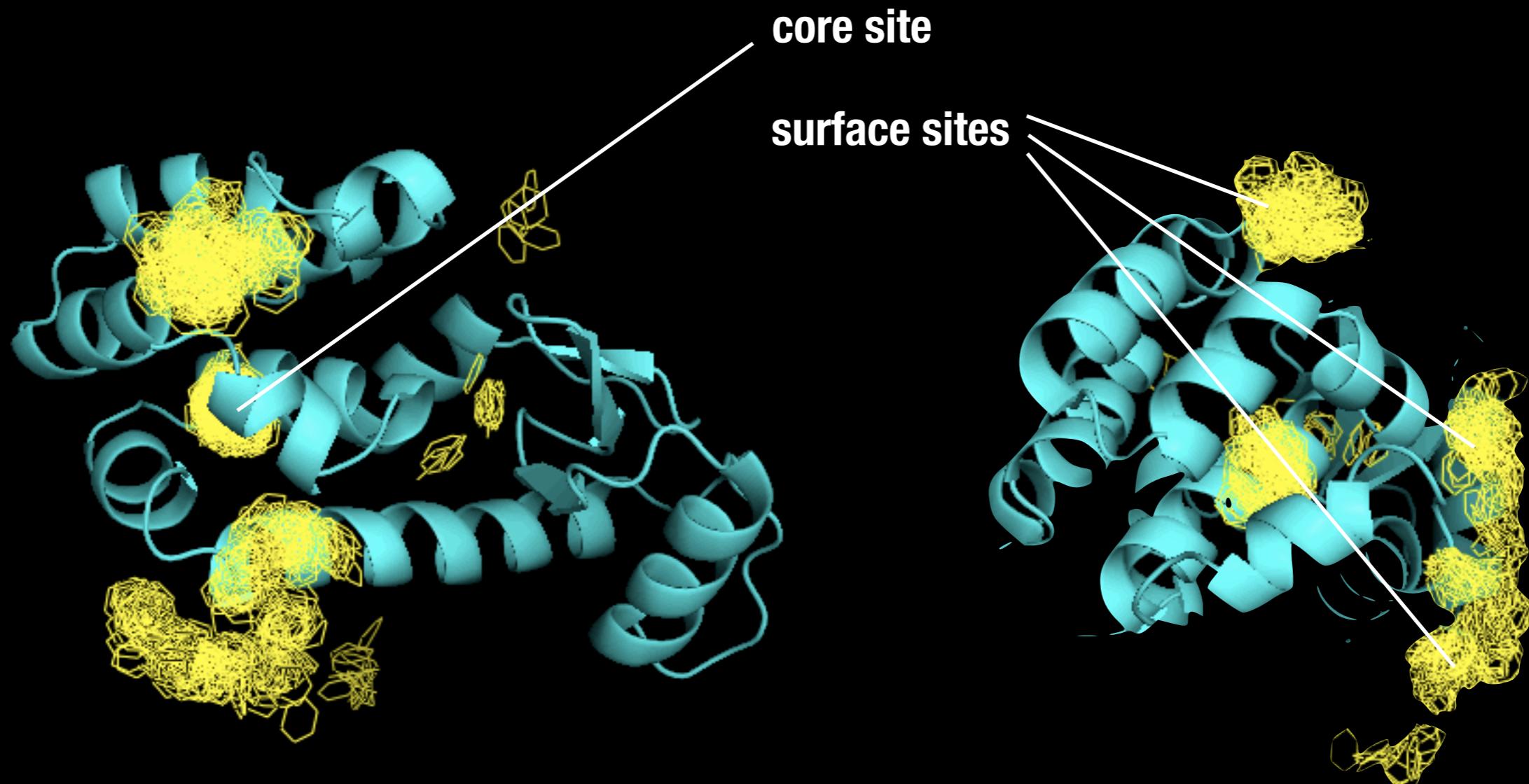
solid fully interacting
transparent noninteracting



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

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

Additional and unknown binding sites can be identified, and their individual affinities estimated by the addition of MC moves

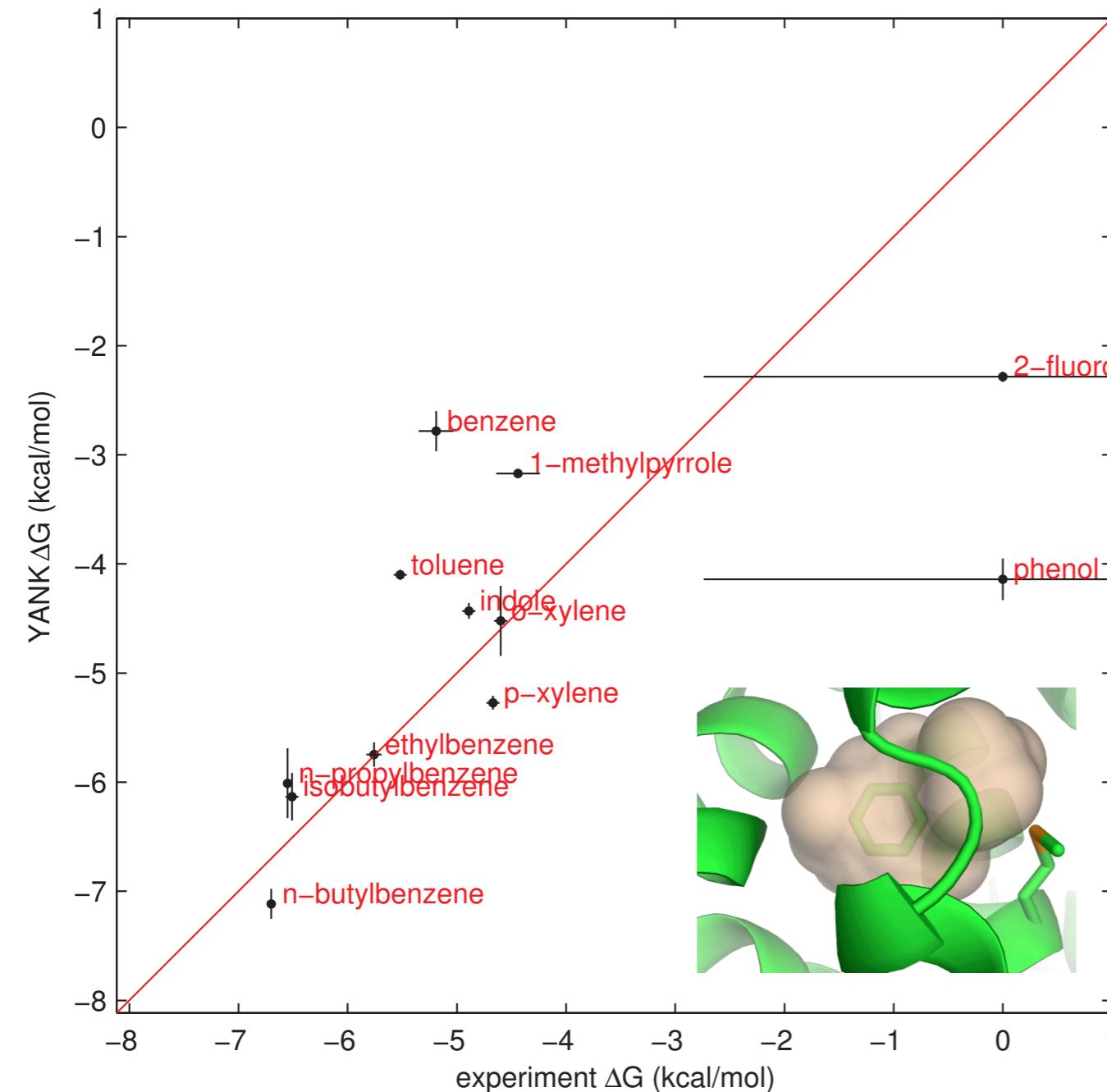


**benzene bound to T4 lysozyme L99A
AMBER96 + OBC GBSA**

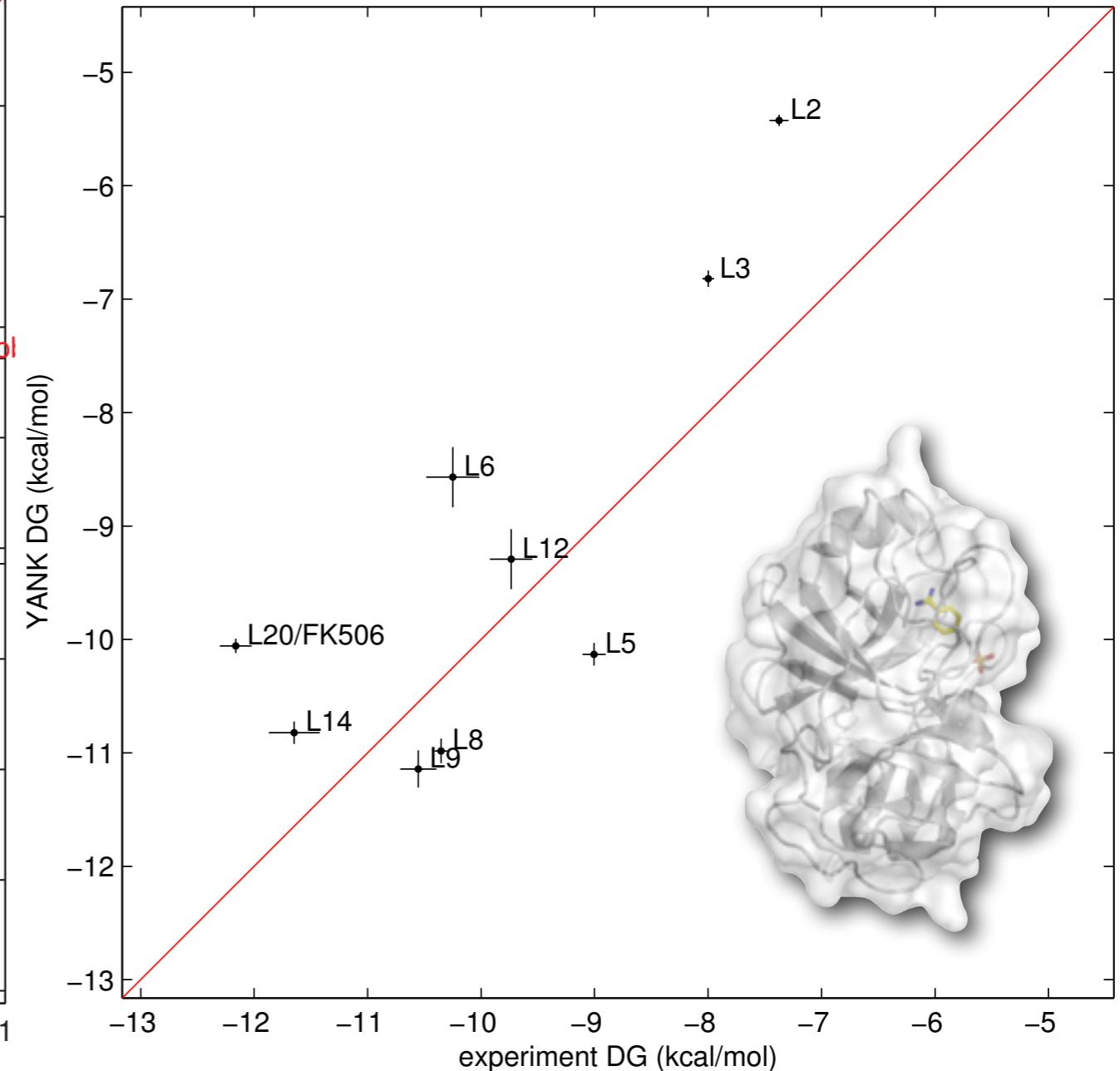
Chodera and Shirts. JCP 135:194110, 2011
Wang, Chodera, Yang, and Shirts. JCAMD 27:989, 2013.
<http://github.org/choderalab/yank>

Initial results using **implicit** models of solvent are promising: Could have a role in rapid affinity prediction

T4 lysozyme L99A



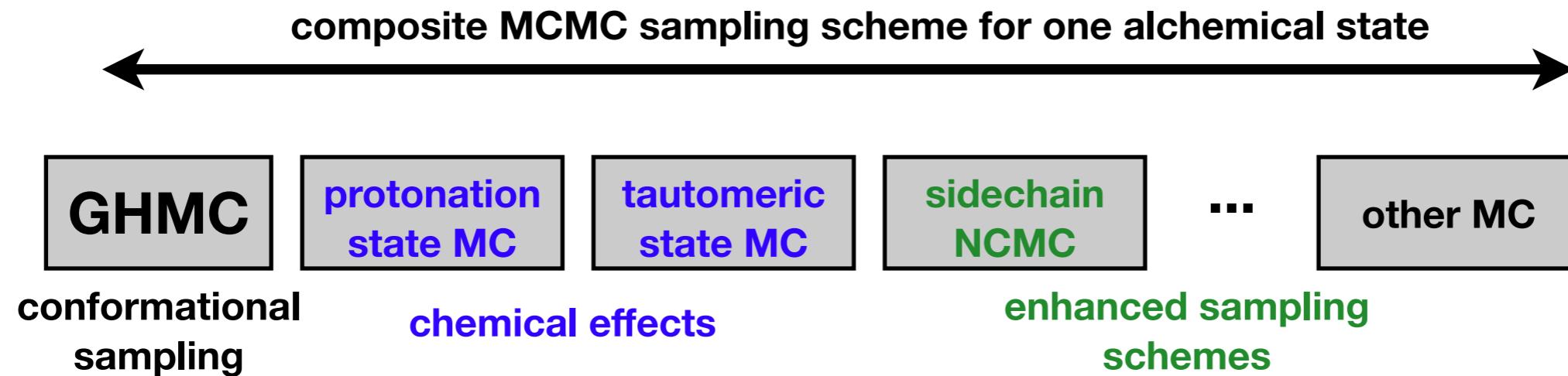
FKBP12



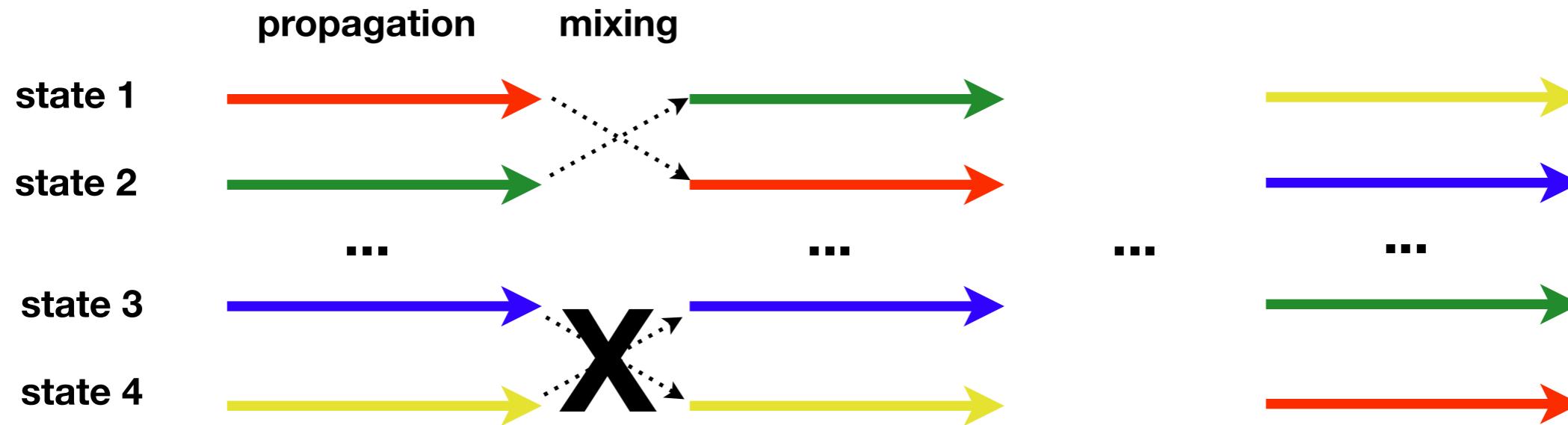
AMBER ff96 + OBC GBSA (no cutoff) + GAFF/AM1-BCC
12 h on 2 GPUs

<http://github.org/choderalab/yank>

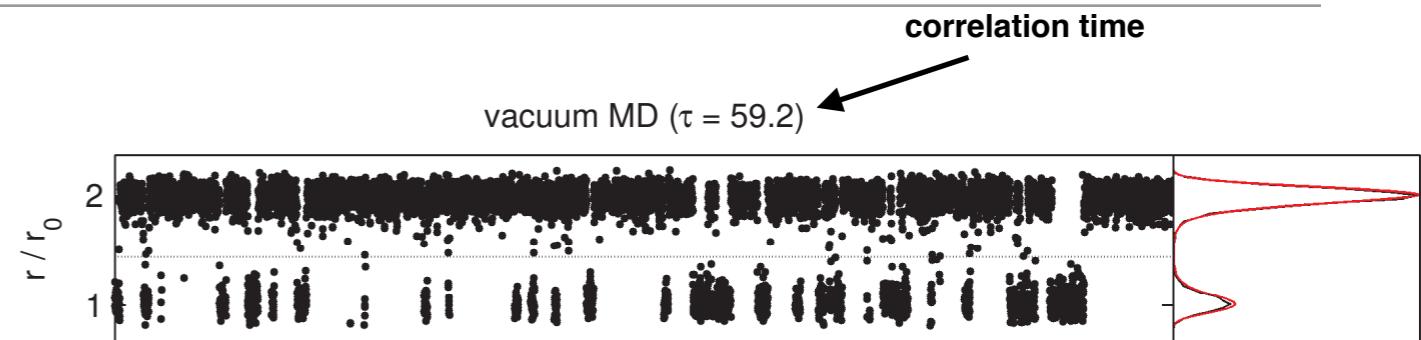
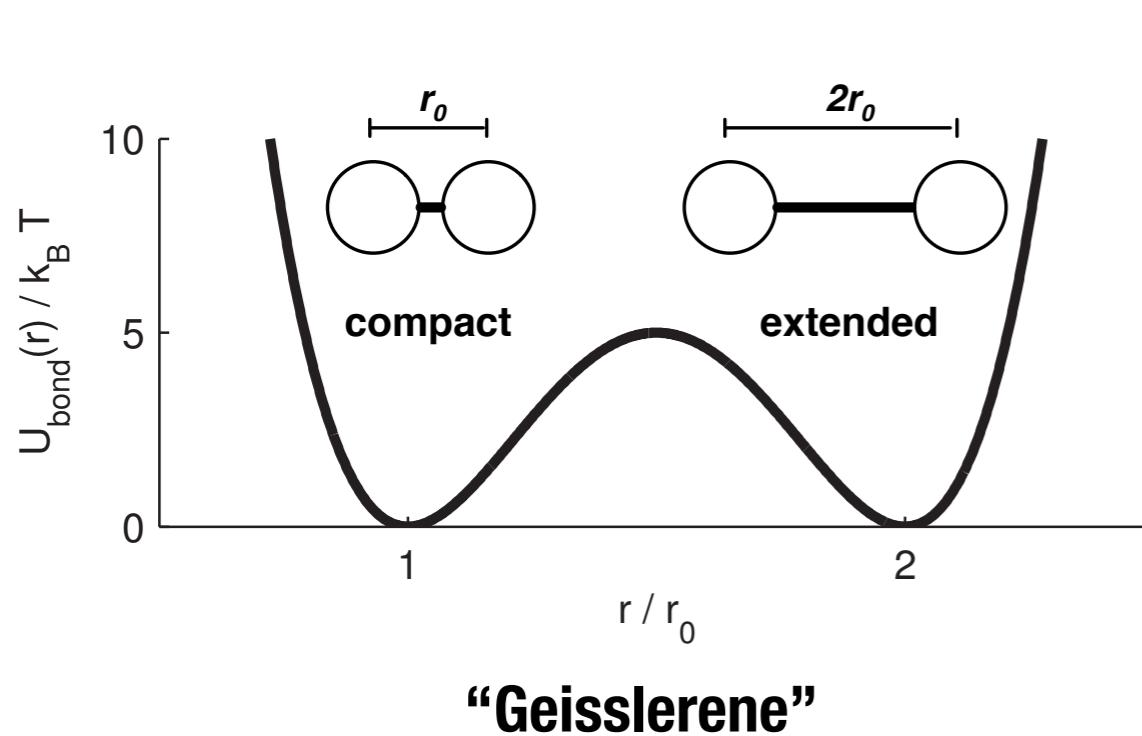
Markov chain Monte Carlo framework can allow flexible inclusion of enhanced sampling schemes and chemical effects



Can be combined with replica exchange schemes to decrease correlation times

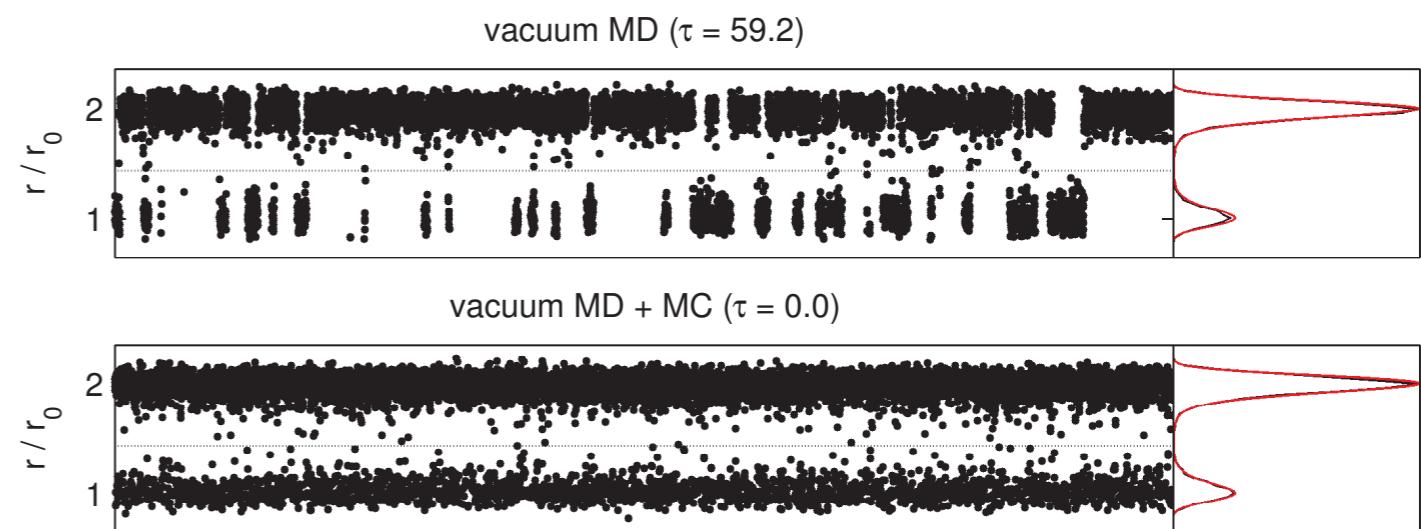
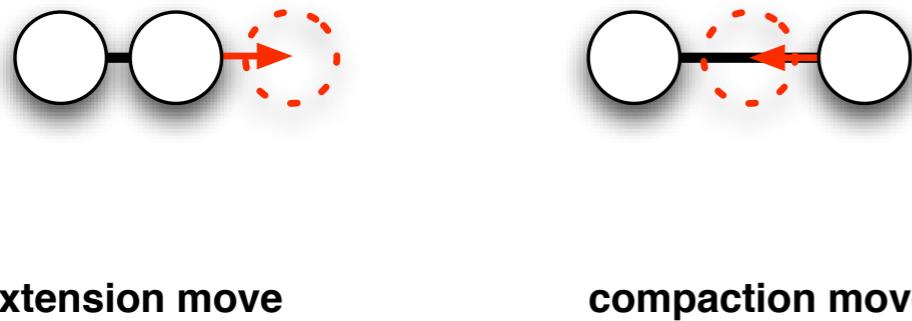
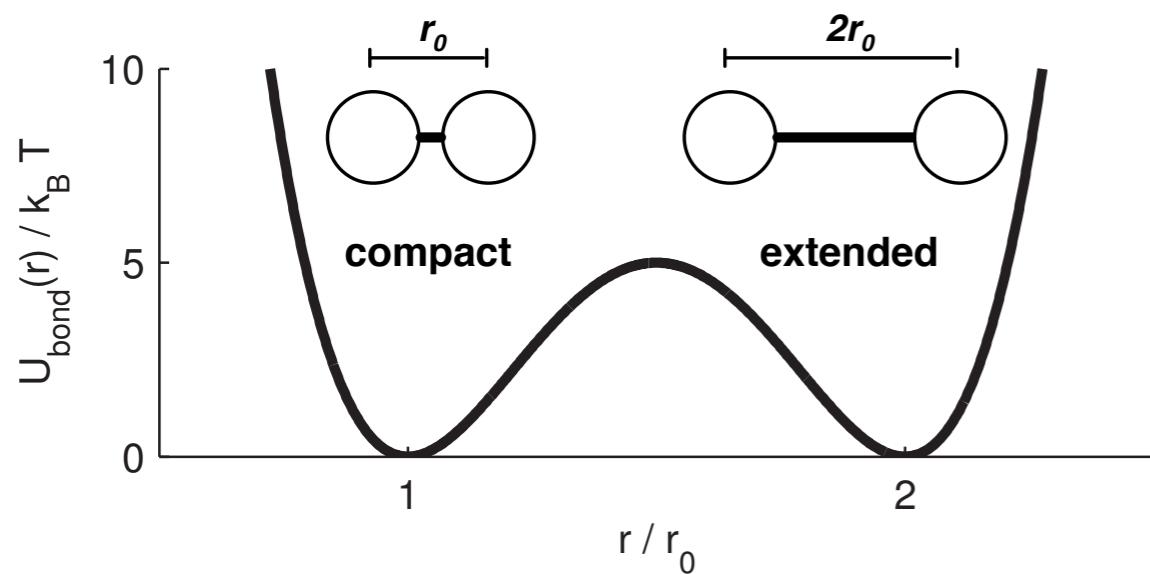


Mixing in clever Monte Carlo moves can speed sampling

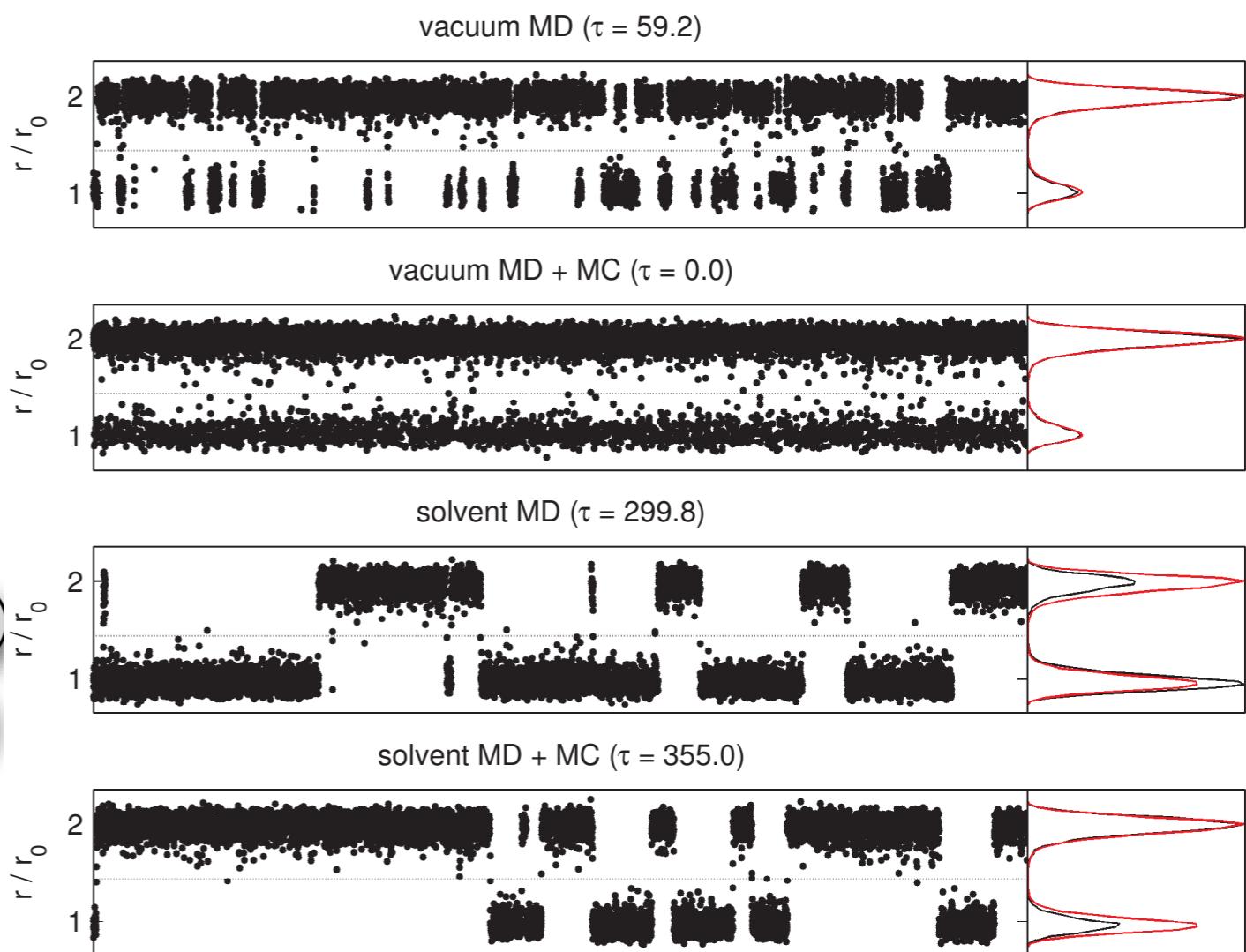
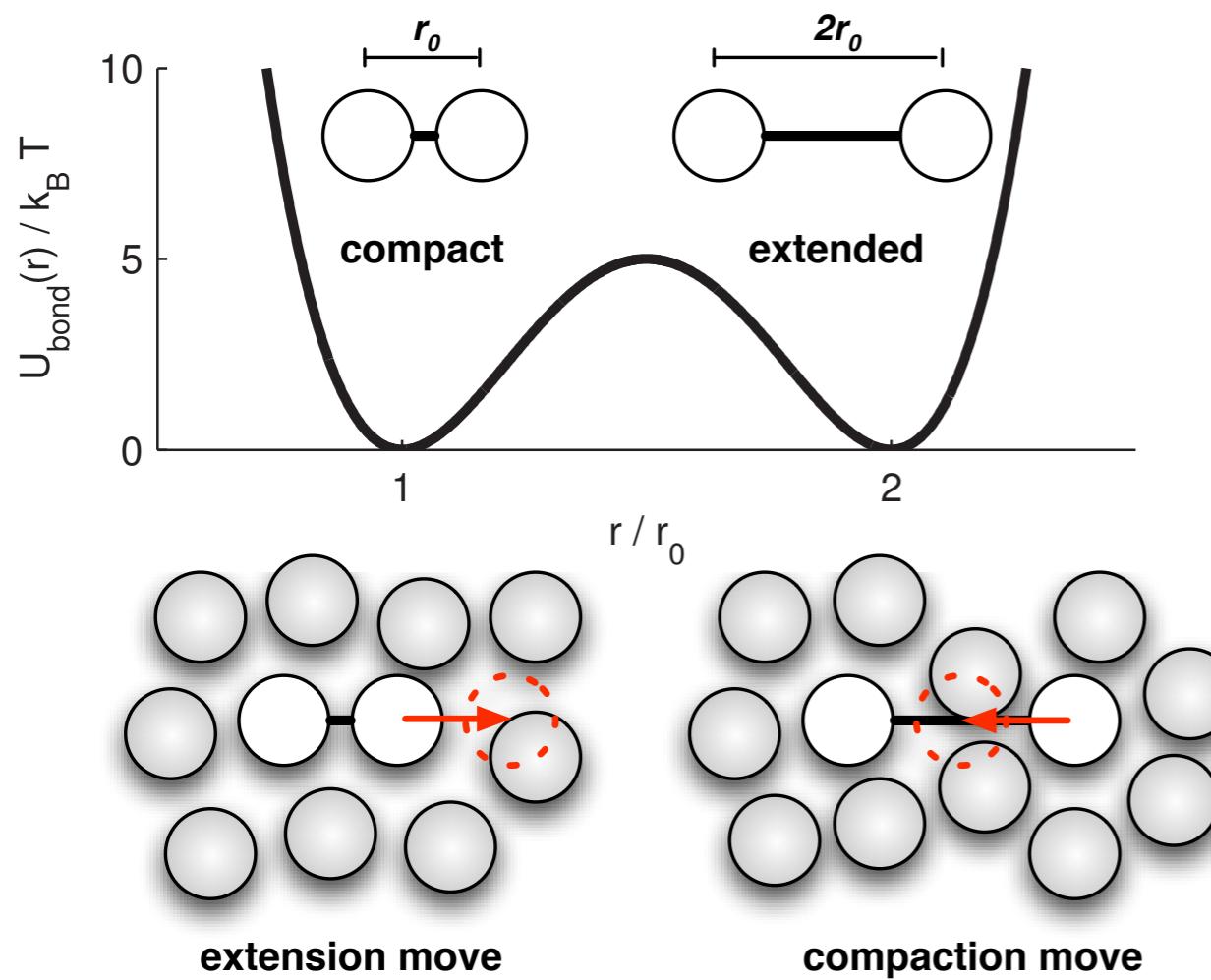


Straightforward molecular dynamics can be slow.

Monte Carlo moves can speed barrier crossing in vacuum...



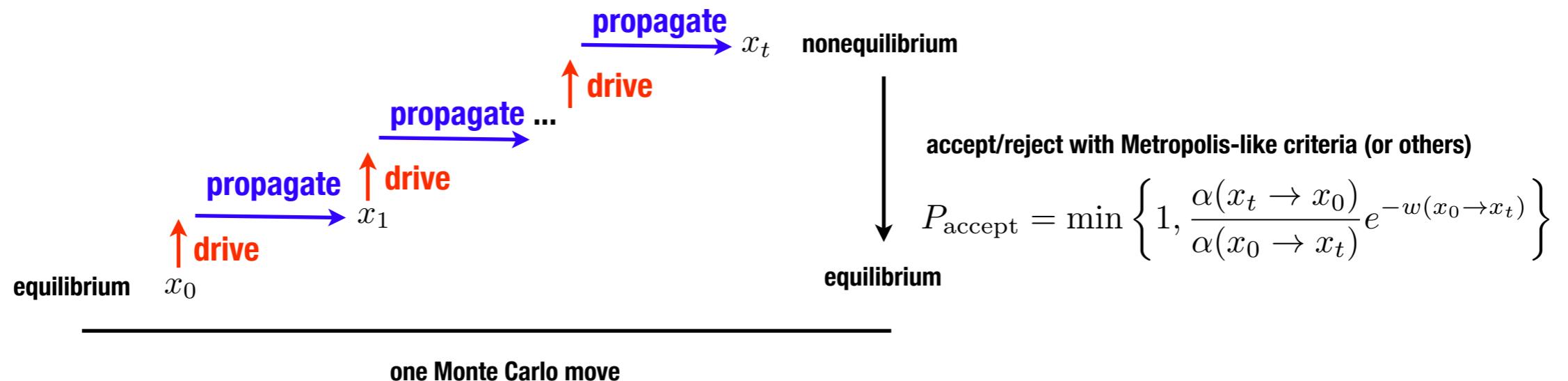
...but Monte Carlo moves are useless in dense solvent



What if we **drive** some degrees of freedom, but **propagate** the rest? Can we correct for this in a Markov chain Monte Carlo framework?

Algorithm:

Drive some degrees of freedom or thermodynamic parameters in small steps, accumulating work
In between driven steps, **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).

Equilibrium Monte Carlo with nonequilibrium moves

$x_t \in \mathcal{X}$

configurations in some state space

$\alpha_t(x, y)$

perturbation kernel

$$\sum_y \alpha_t(x, y) = 1$$

$K_t(x, y)$

propagation kernel

$$\sum_x \pi_t(x) K_t(x, y) = \pi_t(y)$$

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

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

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

$$\alpha_t(x, y) > 0 \Leftrightarrow \alpha_t(y, \tilde{x}) > 0$$

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

$\tilde{\Lambda}$ **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):

$A(X|\Lambda)$
acceptance probability

$\Pi(X|x_0, \Lambda)$
path generation probability

$P(\Lambda|x_0, \lambda_0)$
protocol selection

$\pi(x_0, \lambda_0)$
equilibrium

$A(\tilde{X}|\tilde{\Lambda})$
acceptance probability

$\Pi(\tilde{X}|\tilde{x}_T, \tilde{\Lambda})$
path generation probability

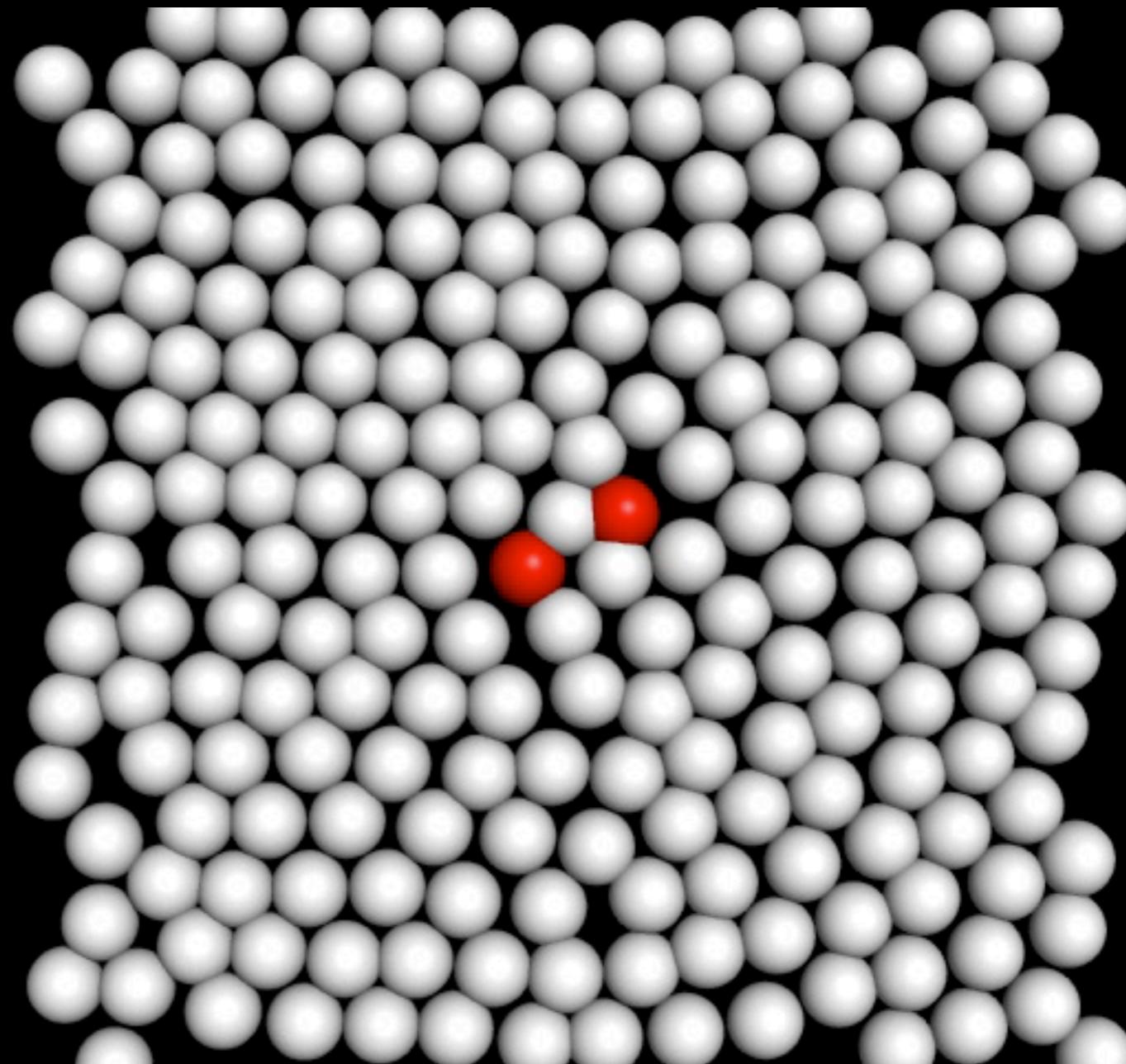
$P(\tilde{\Lambda}|\tilde{x}_T, \lambda_T)$
protocol selection

$\pi(\tilde{x}_T, \lambda_T)$
equilibrium

Result is a general acceptance criteria for any nonequilibrium perturbation:

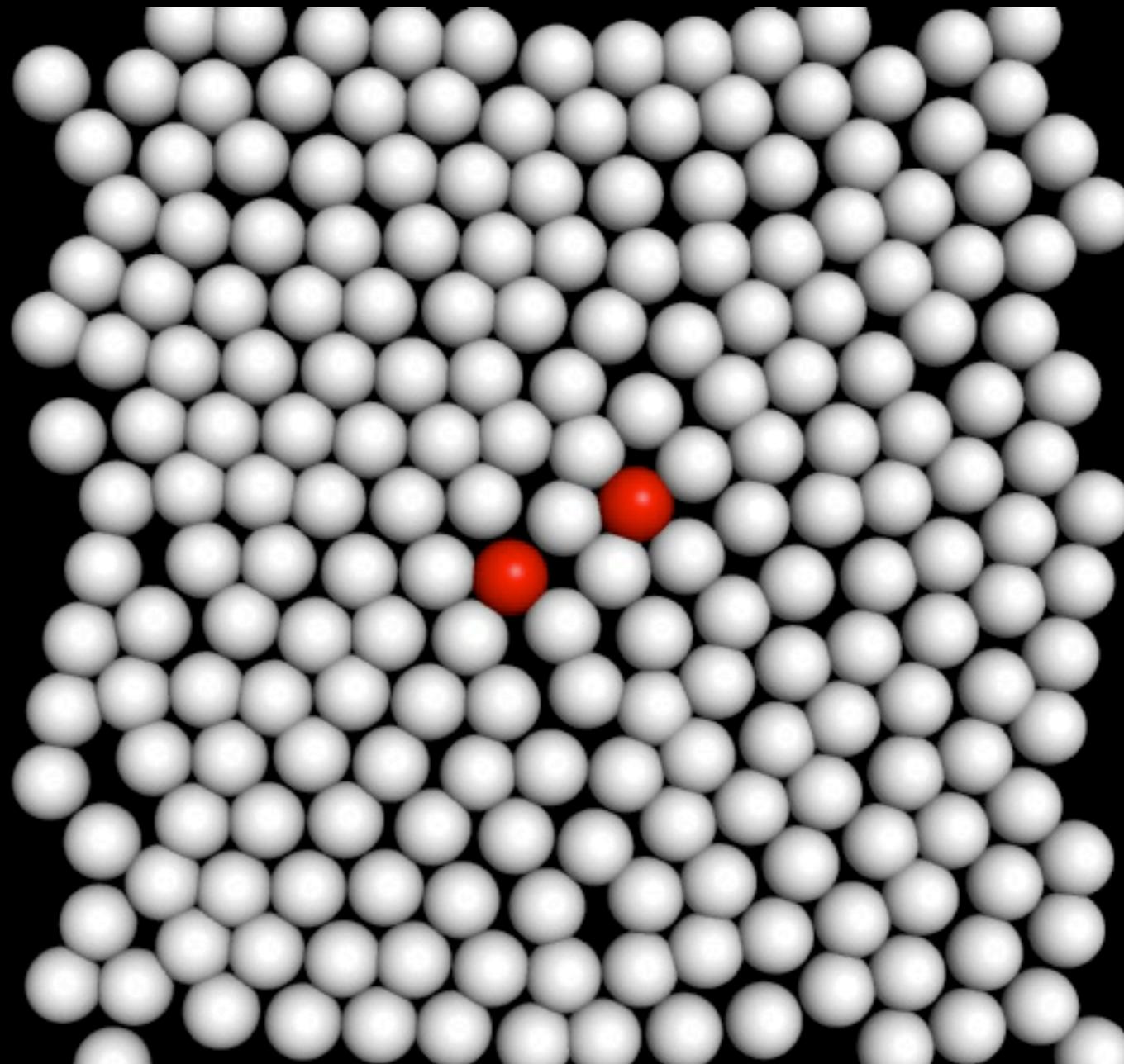
$$A(X|\Lambda) = \min \left\{ 1, \frac{\pi(x_T, \lambda_T)}{\pi(x_0, \lambda_0)} \frac{P(\tilde{\Lambda}|\tilde{x}_T, \lambda_T)}{P(\Lambda|x_0, \lambda_0)} \frac{\tilde{\alpha}(\tilde{X})}{\alpha(X)} e^{-\Delta S(X)} \right\}$$

Nonequilibrium MC dimer moves in a 2D WCA fluid



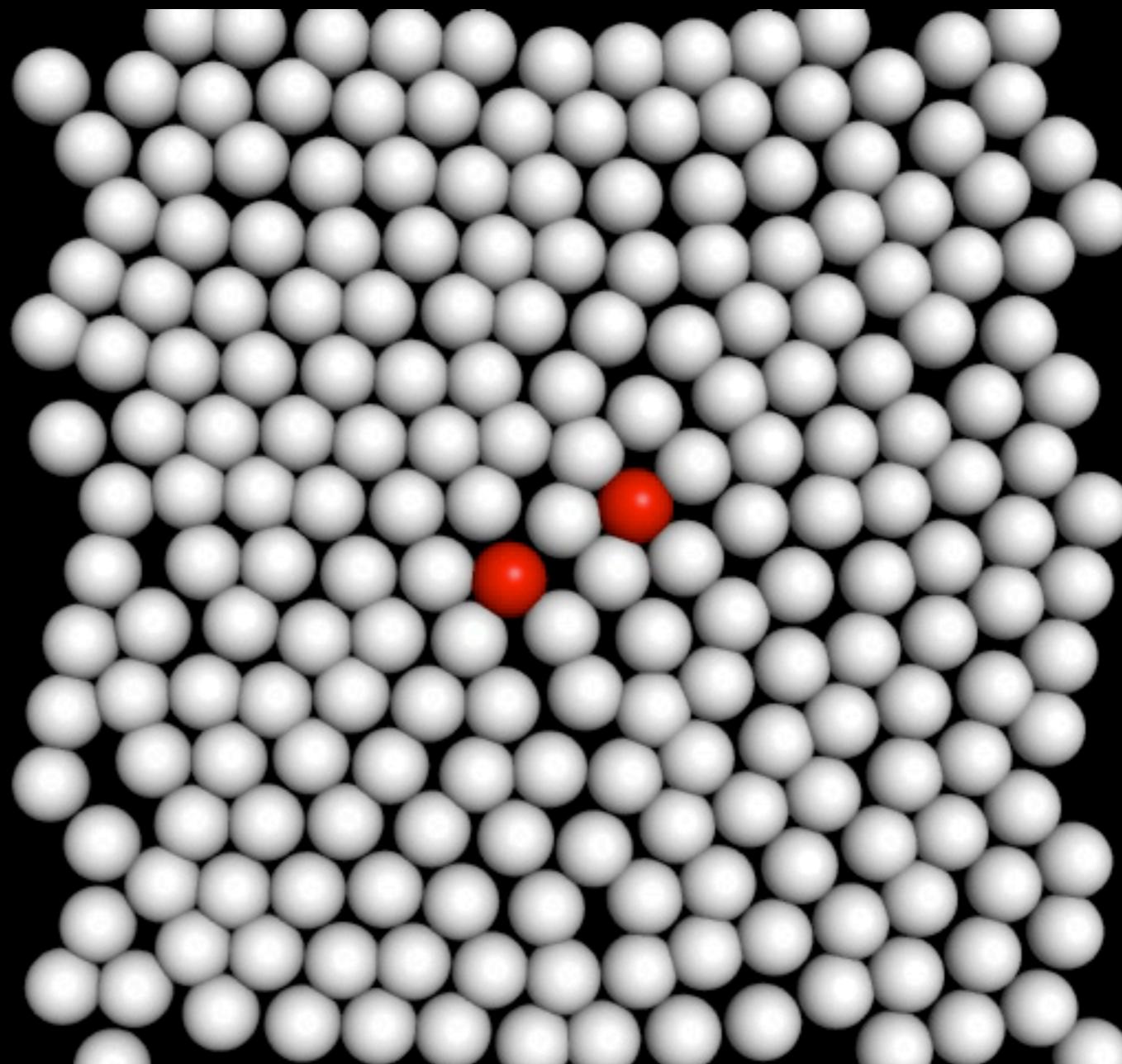
**Bad solvent overlap:
Always rejected :(**

Nonequilibrium MC dimer moves in a 2D WCA fluid



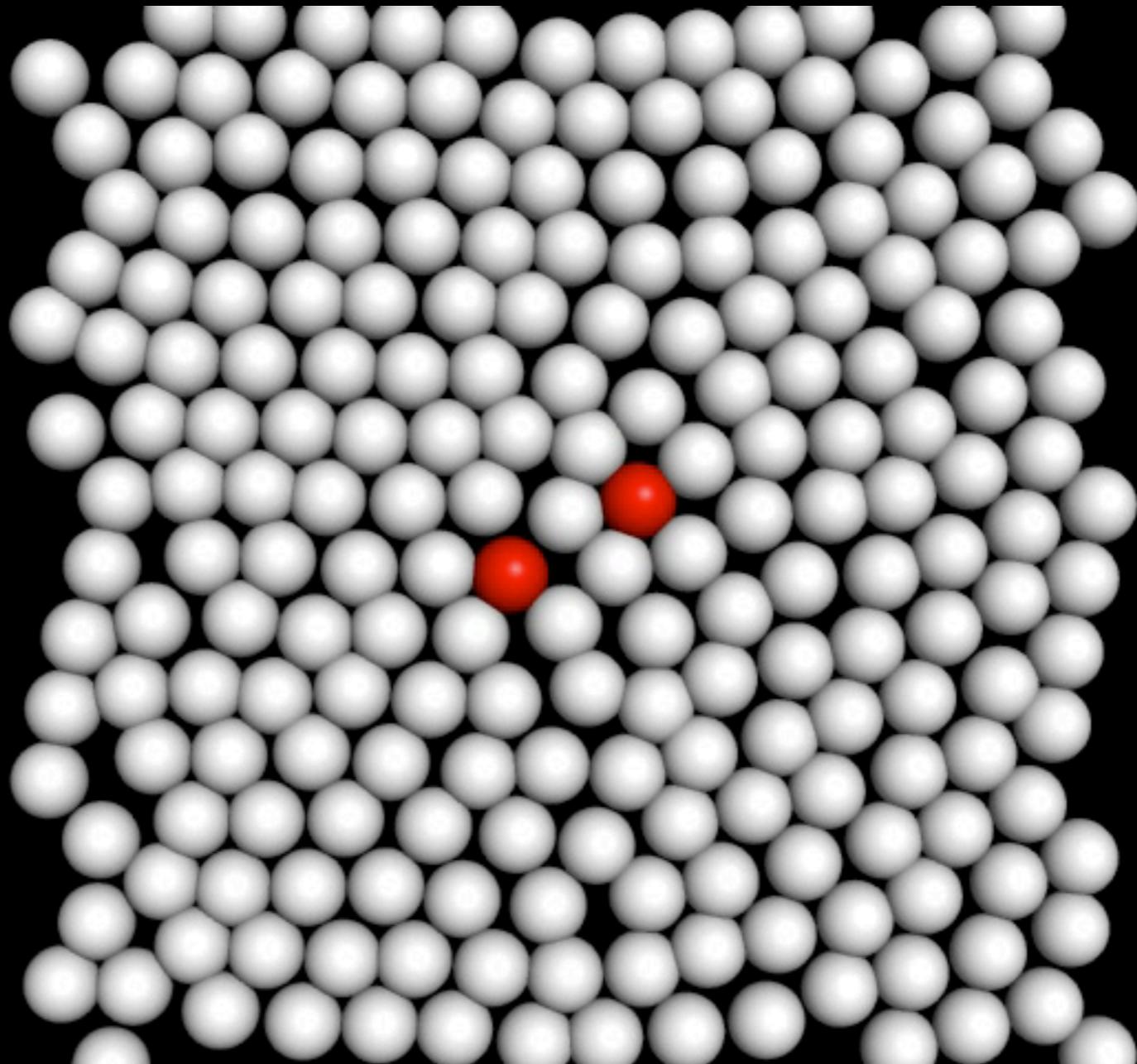
Solvent is “squeezed” out
but significantly perturbed :/

Nonequilibrium MC dimer moves in a 2D WCA fluid



Solvent is “squeezed” out
with time for reorganization :)

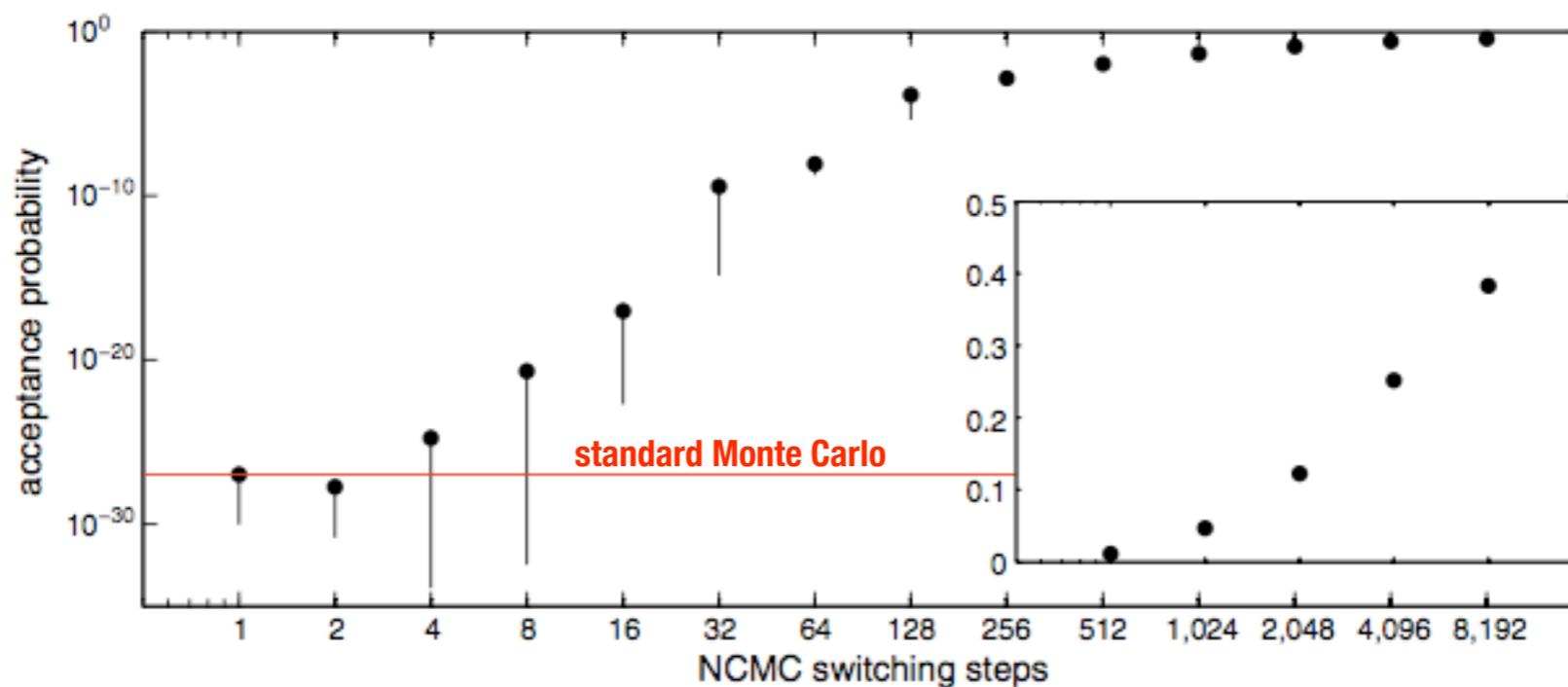
Nonequilibrium MC dimer moves in a 2D WCA fluid



Solvent is “squeezed” out
close to reversible limit;
high acceptance rates! :D

Acceptance probability can be **astronomically** boosted

Acceptance probability can be increased from 10^{-27} to 10^0 (38%)!



Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation

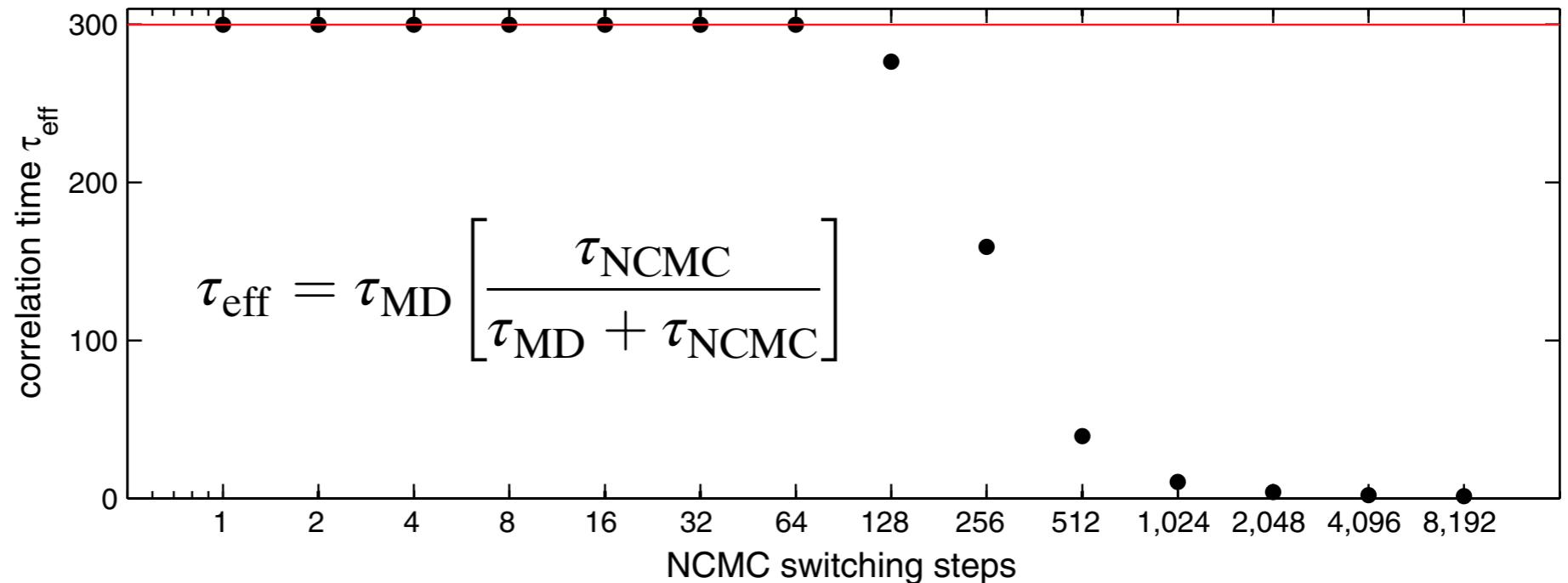
Jerome P. Nilmeier^a, Gavin E. Crooks^b, David D. L. Minh^c, and John D. Chodera^{d,e}

Full 3D system
216 WCA particles
Reduced density $\rho\sigma^3 = 0.96$
Reduced temperature $k_B T/\epsilon = 0.824$
5 kT barrier

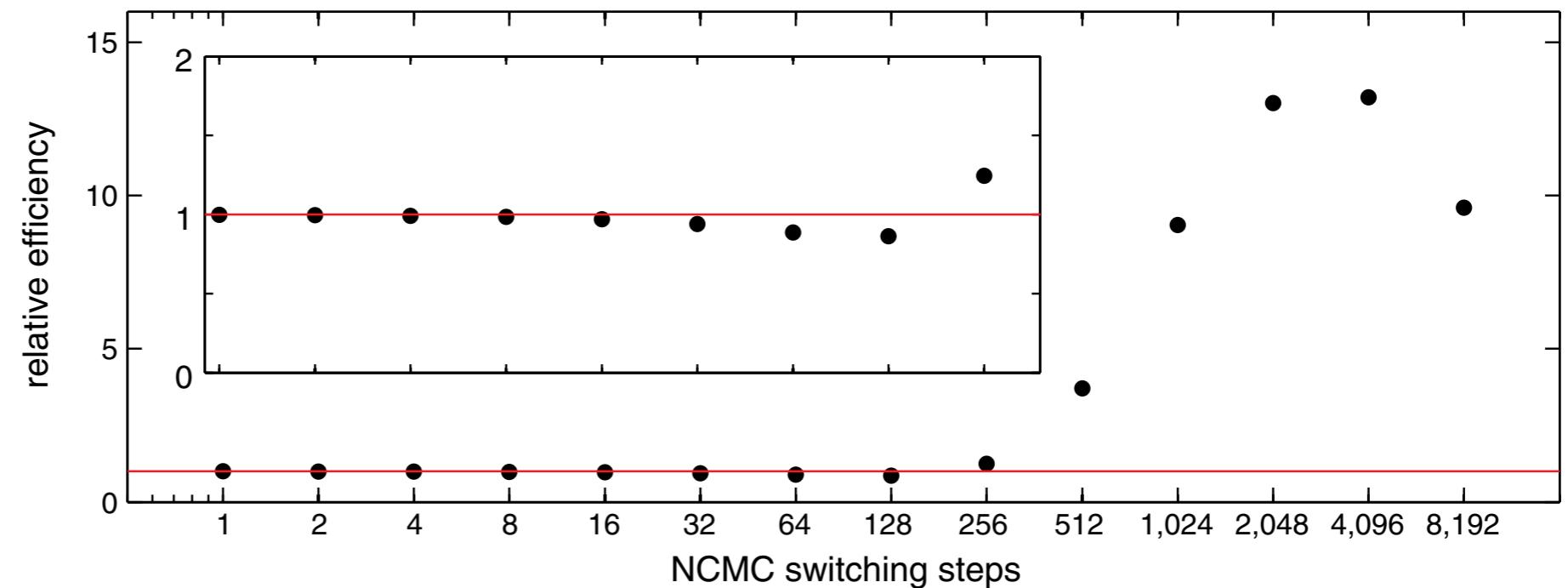
^aBiosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550; ^bPhysical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; ^cBiosciences Division, Argonne National Laboratory, Argonne, IL 60439; and ^dCalifornia Institute for Quantitative Biosciences (QB3), University of California, Berkeley, CA 94720

Overall efficiency gain can still be large, even when the extra work required for NCMC switching is included

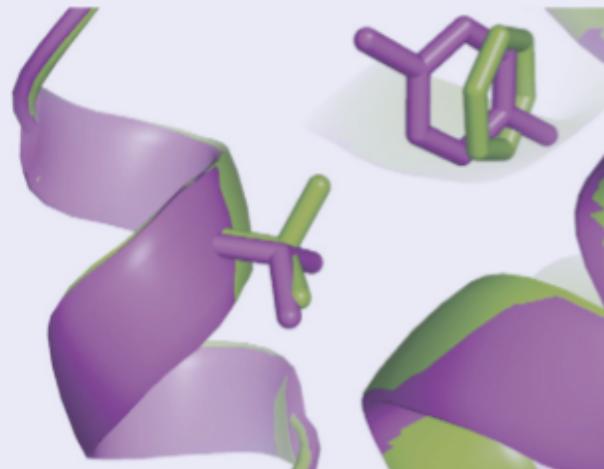
measure of number of iterations to reduce correlation to $1/e$



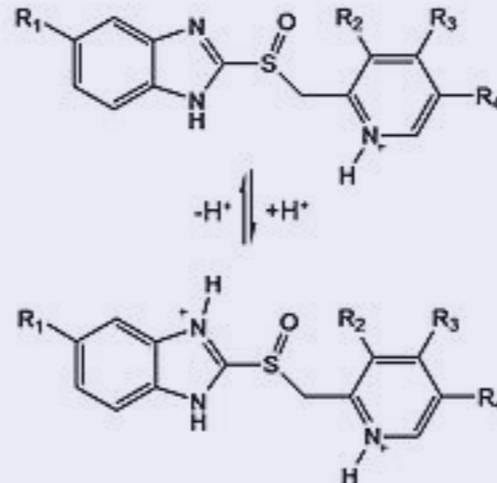
ratio of computer time required to generate one uncorrelated sample without and with NCMC



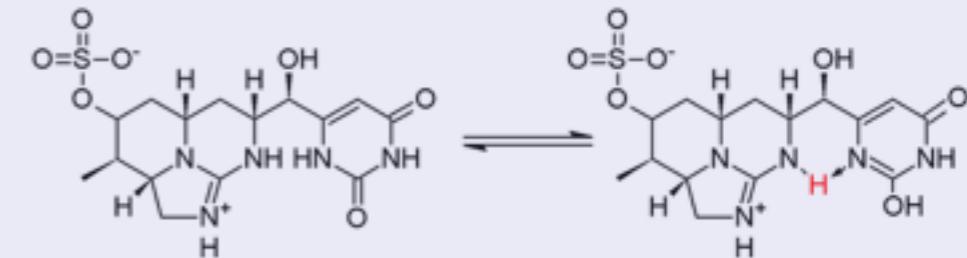
New Monte Carlo techniques open up new possibilities for treating physical effects and facilitating **design**



sidechain rotamer sampling

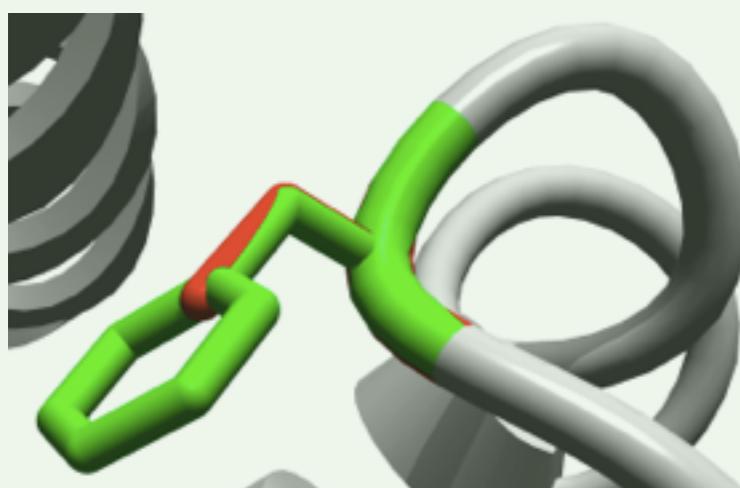


dynamic protonation states
(protein and ligand)

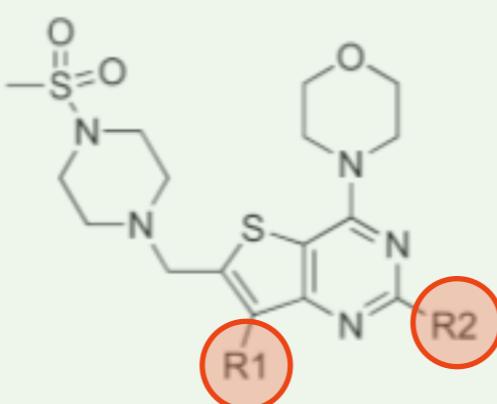


dynamic tautomerization
(ligand)

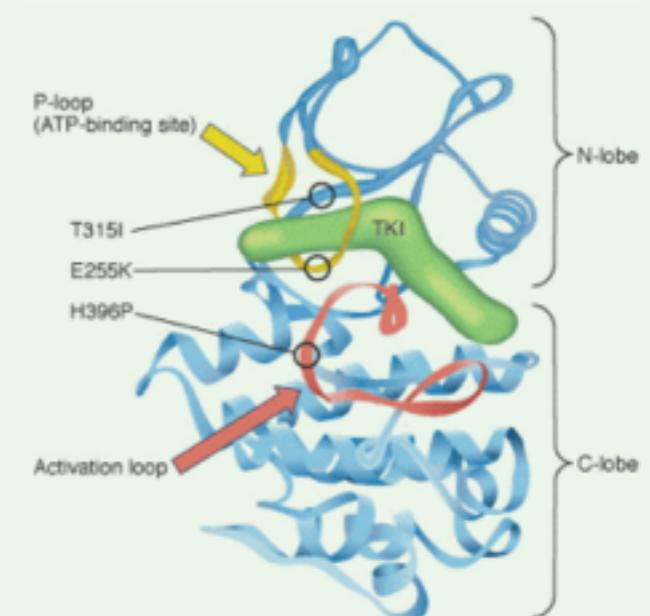
ENHANCING CHEMICAL ACCURACY
FACILITATING DESIGN



sampling over
target mutations

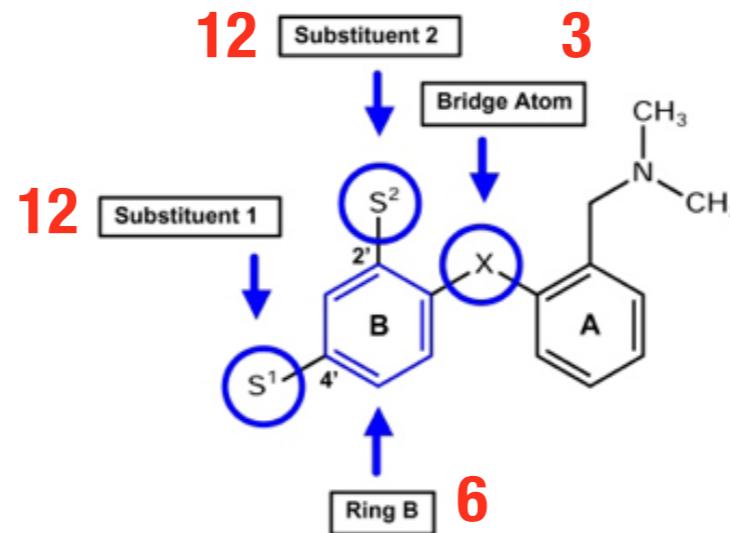


sampling over substituents
for ligand design



ligand design considering
potential resistance mutations

Expanded-ensemble methods will allow **combinatorially large** chemical spaces to be sampled to solve complex design problems



$12 \cdot 12 \cdot 6 \cdot 3 = 2592$ possibilities!

Expanded ensemble:

$$p(\mathbf{x}, k) = Z^{-1} \exp[-u_k(\mathbf{x}) + g_k]$$

\mathbf{x} configuration
 k chemical species

Monte Carlo proposals between chemical species
with clever choices of biasing potential
could “hunt” for good binders

Speeding up the cycle of learning from failure can accelerate progress toward rational ligand design

computational predictions



experimental confirmation

“Fail fast, fail cheap”

Make rigorous calculations of affinity fast and accurate



GPU acceleration

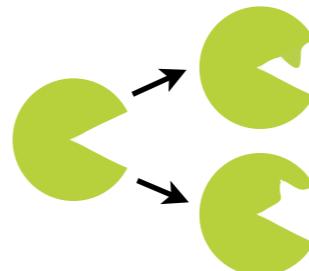
$$\pi(x)K(x,y) = \pi(y)K(y,x)$$



modular MCMC for sampling and chemical effects

enhanced sampling algorithms

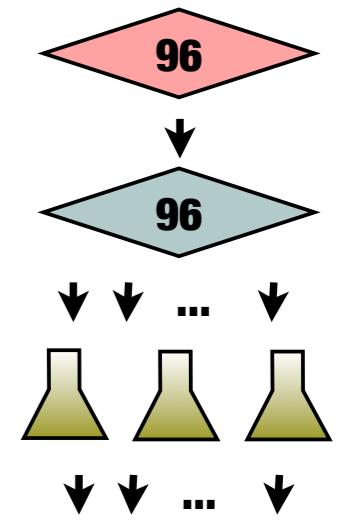
Test and improve models quickly and cheaply by inverting the drug discovery problem



mutate proteins instead of ligands



buy inexpensive ligands



high-throughput experiments

Inverting the drug discovery problem: A faster, cheaper way to build (and break) computational models



quick change mutagenesis

~ 1-2 weeks, >70% efficiency



expression assessment in 1 ml culture

~ 2-3 weeks, select mutants that express >25 ug/ml



expression and purification in ~2L culture

~ 4/week, >25 mg



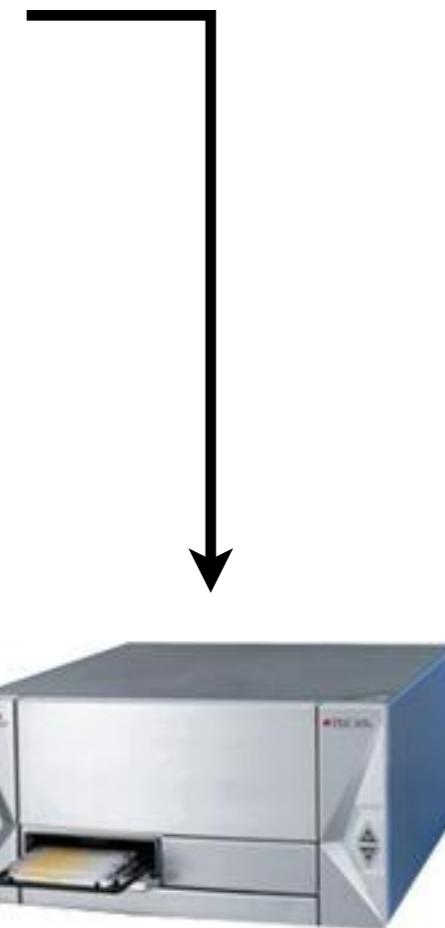
purchase known ligands several ligands/receptor



isothermal titration calorimetry
~40/day (automated); <0.5 mg/experiment



surface plasmon resonance
6 x 6 parallel experiments; ~10 ug protein

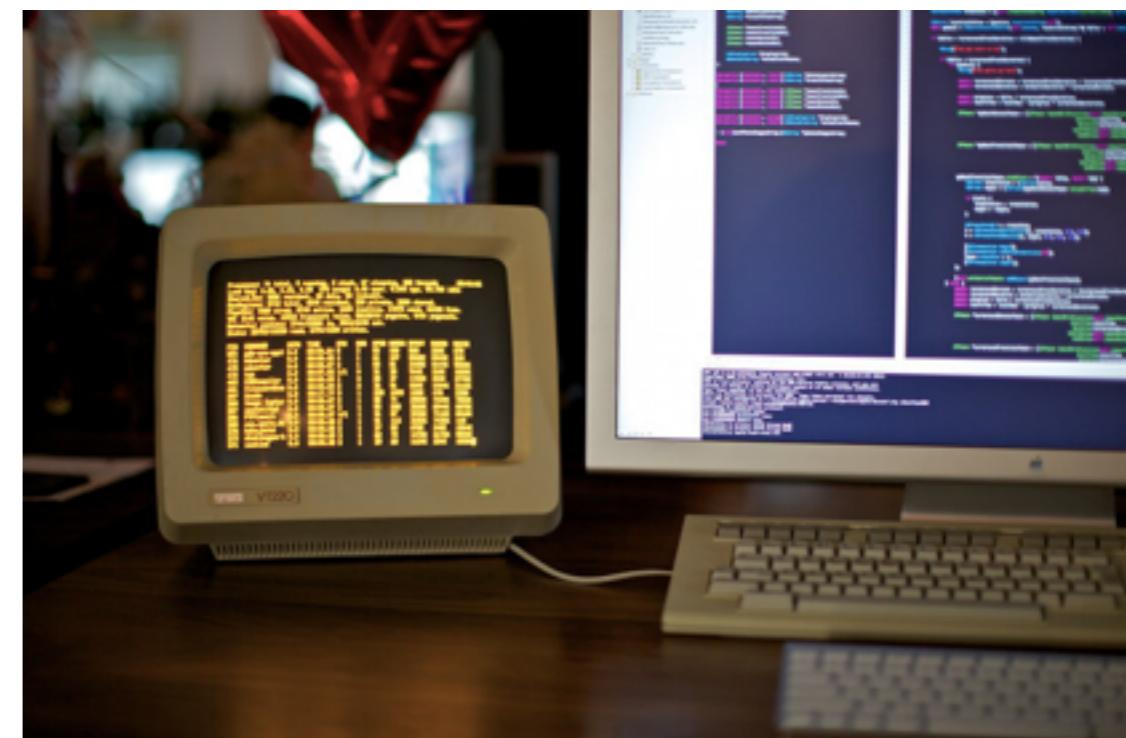
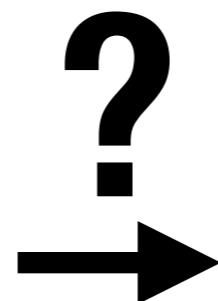


fluorescence binding assays
96 assays/plate; ~1 ug protein

How can we make wetlab experiments look more like problems we know how to solve efficiently?



messy
laborious
inconsistent
skill-dependent
9 am - 5 pm



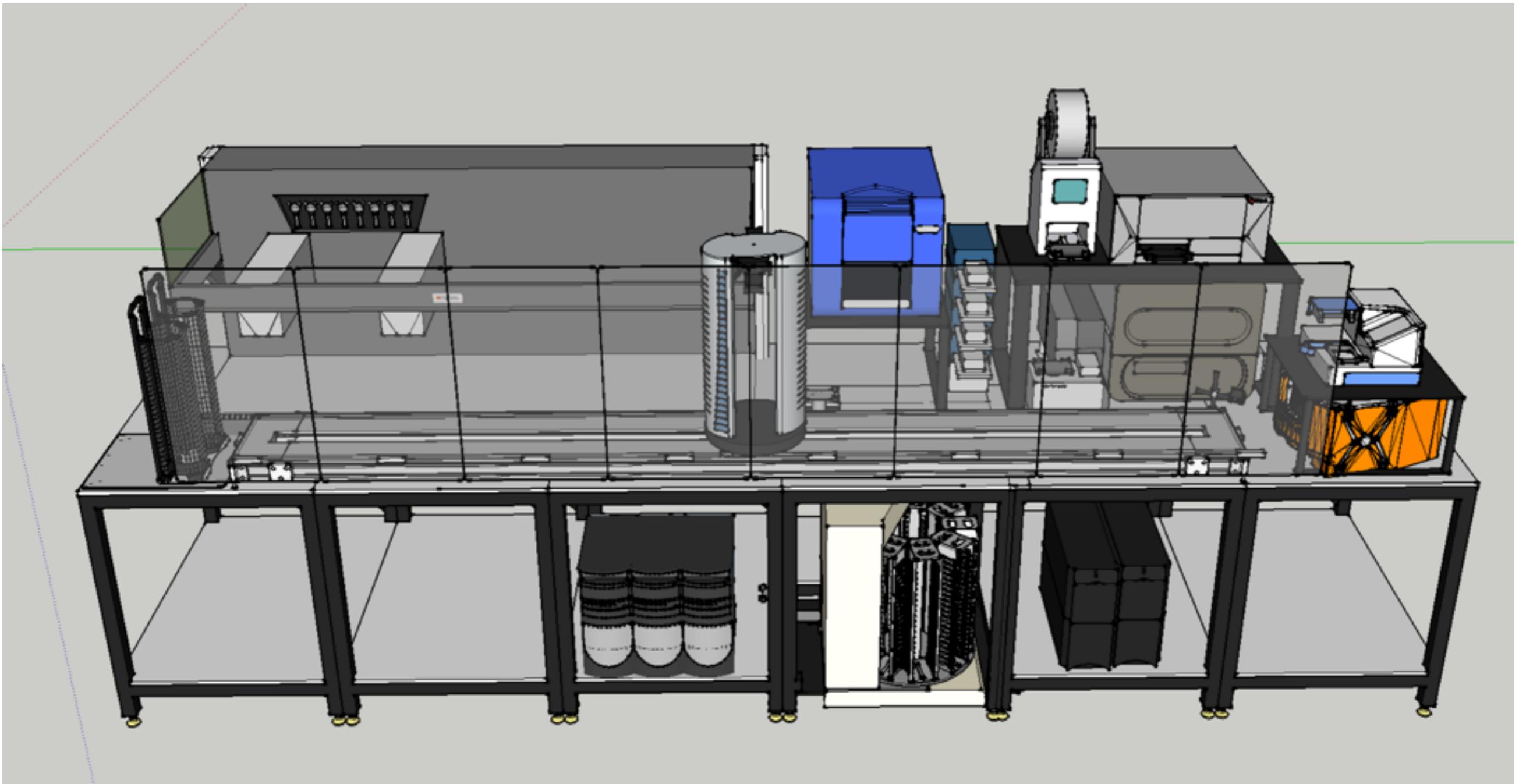
precise
structured
consistent
reproducible
round-the-clock

AUTOMATE. EVERYTHING.



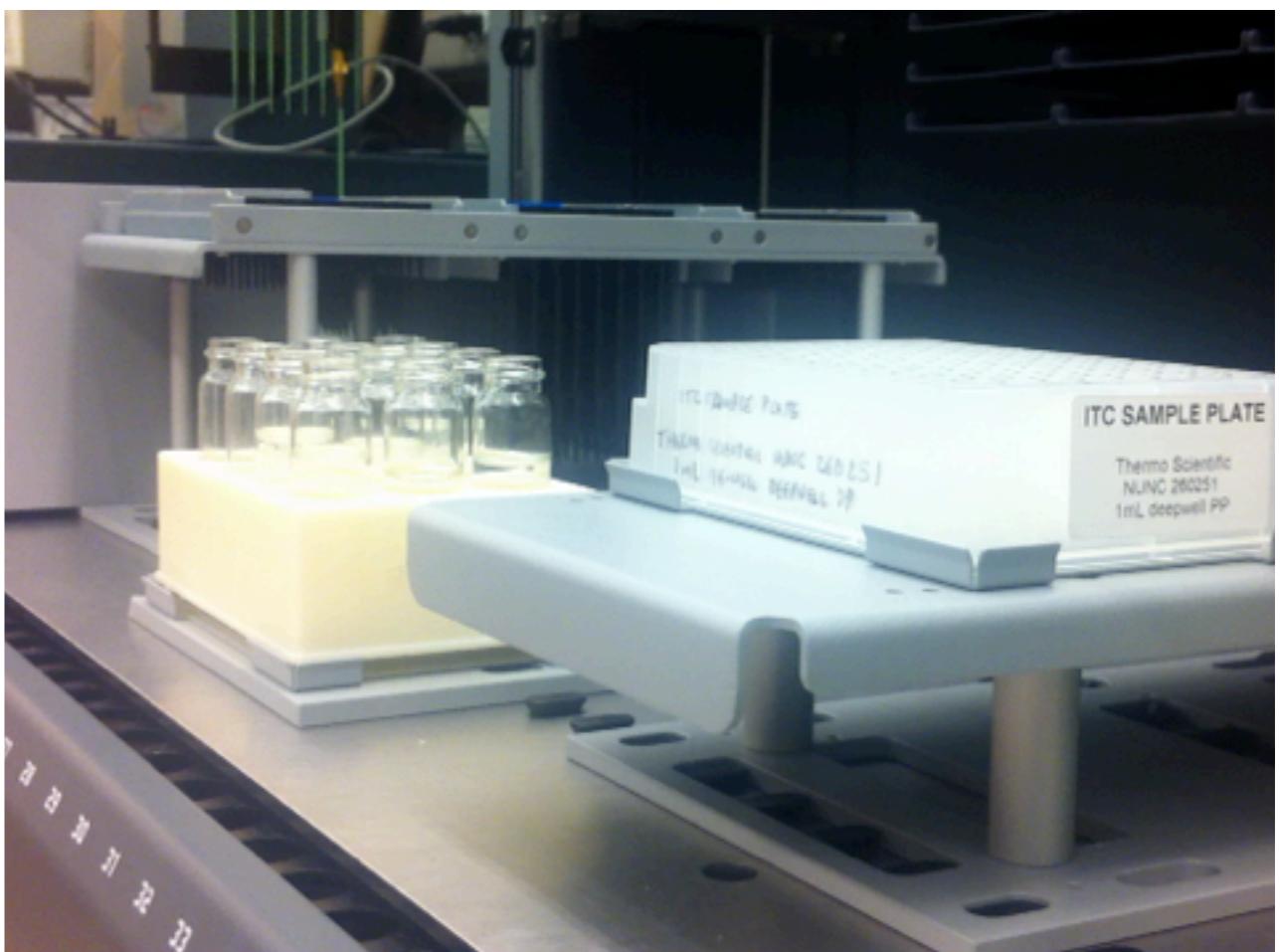
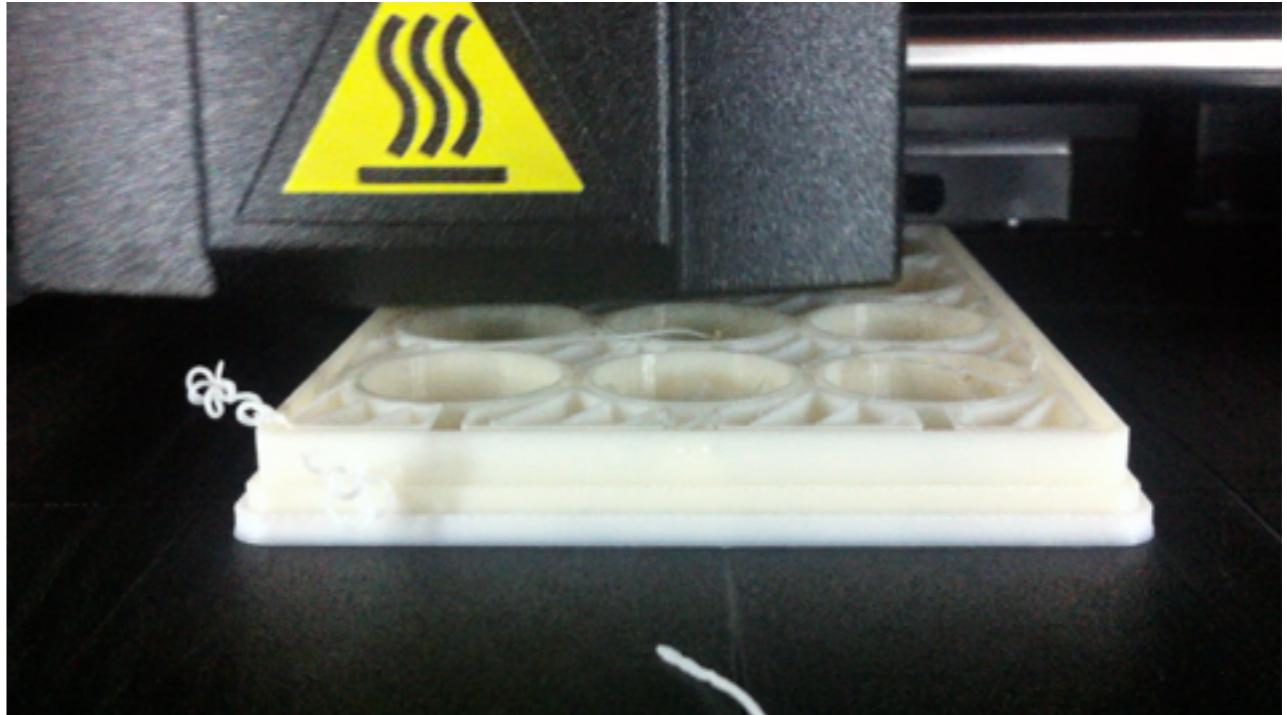
Automated platform for bacterial cloning, mutagenesis, expression, purification, and binding affinity measurement with 24/7 operational capacity

AUTOMATE. EVERYTHING.

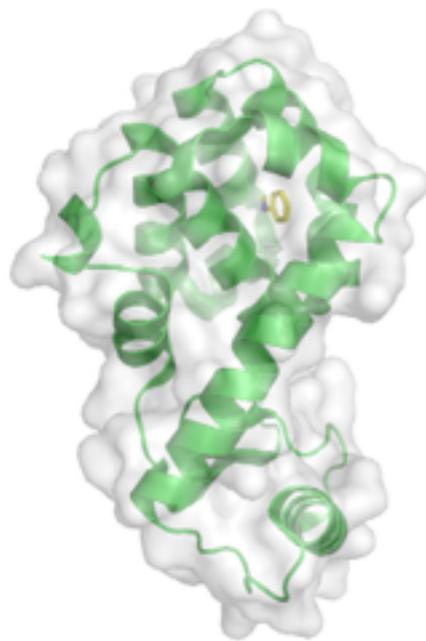


Automated platform for bacterial cloning, mutagenesis, expression, purification, and binding affinity measurement with 24/7 operational capacity

New technologies facilitate rapid progress

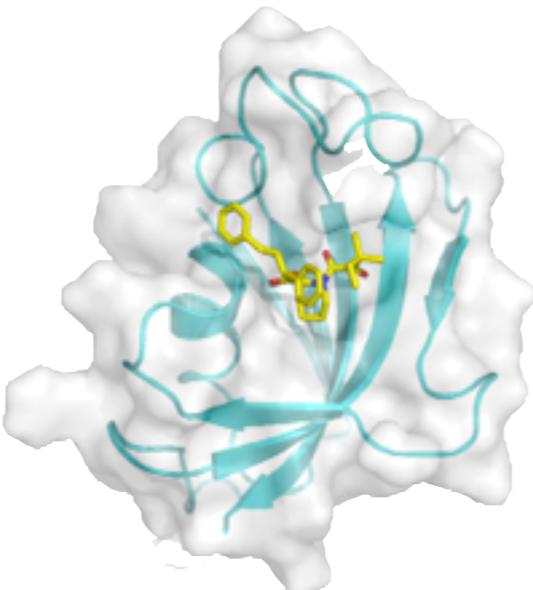


We're building a benchmark set covering a range of complexity, isolating individual challenges to modeling



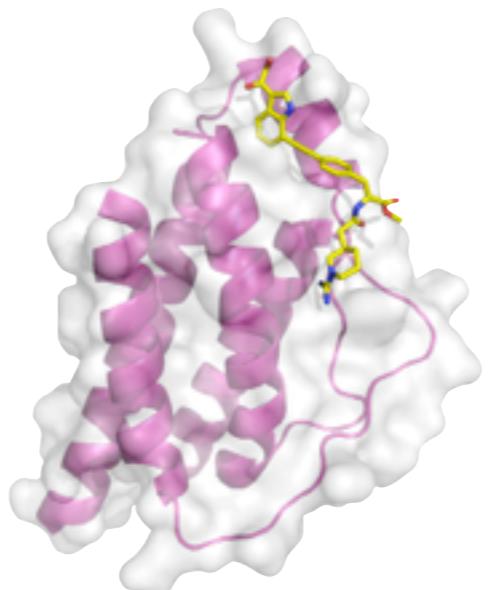
small apolar ligands
T4 lysozyme L99A

small, rigid protein
small, neutral ligands
fixed protonation states
multiple sidechain orientations
multiple ligand binding modes



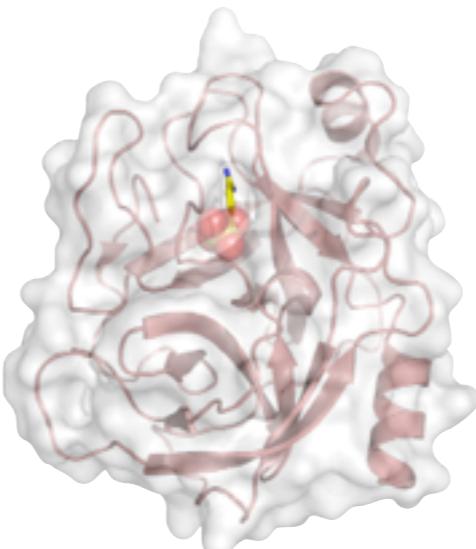
polar ligands
FKBP12

small, rigid protein
fixed protonation states
larger natural product-like
ligands with rotatable bonds



IL-2

small protein
fixed protonation states
some allostery and
binding site plasticity



trypsin proteases

small, rigid protein
small ligands
charged ligands
potential for protonation state changes



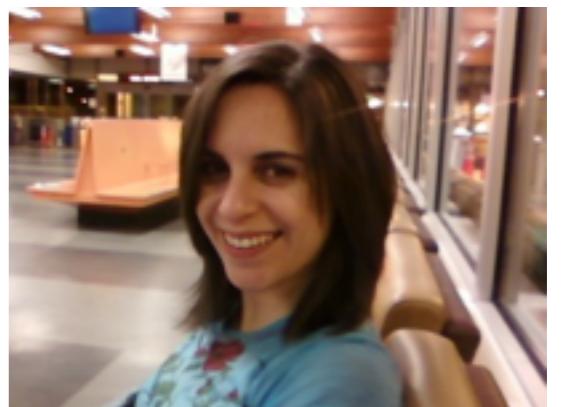
JNK3 kinase

large protein, multiple conformations
large drug-like ligands, rotatable bonds
multiple protonation states? tautomers?
phosphorylation and activation
peptide substrate?
MgCl₂ salt effects?

easy
hard

many other good model systems to choose from: DHFR, cytochrome C peroxidase, AmpC, adenylate kinase, etc.

We can take a “shotgun approach” keeping only those systems and mutants that express well



Sarah Boyce

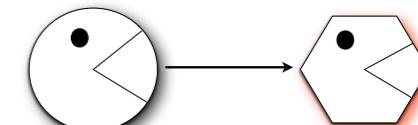
ID	Gene	Protein	PCR Result	Expression test	ug/mL culture	Tag (Best)
1	Abl1	Abl kinase	yes	NO		N/A
2	AK1	adenylate kinase 1	yes	High	121	SUMO
3	AK2	adenylate kinase 2	yes	High	121	H6
4	AMPC_WT	AmpC beta-lactamase	NO	---		---
5	CAH2	carbonic anhydrase II	yes	High	198	H6
6	CALM1	calmodulin	yes	High	107	SUMO
7	CCP_WT	cytochrome c peroxidase	yes	High	224	H6
8	CCP_GA	cytochrome c peroxidase	yes	High	198	H6
9	CDK2	cyclin-dependent kinase 2	yes	Low	19	SUMO
10	DHFR	dihydrofolate reductase	yes	Low	5	MOCR
11	ESR1	estrogen receptor α	yes	NO		N/A
12	ESR2	estrogen receptor β	yes	NO		N/A
13	FKBP12	FK506 binding protein	WRONG ORF	WRONG ORF		
14	GYRB	E. coli DNA gyrase B	yes	High	152	MOCR
15	HSV1_TK	thymidine kinase	yes	High	161	MOCR
16	IL2	interleukin 2	yes	Low-Moderate	29	MOCR
17	JNK3	JNK3 kinase	yes	Low	9	MOCR
18	p38MAPK14	p38 kinase	yes	High	170	MOCR
19	PIM1	PIM1 kinase	yes	High	102	MOCR
20	RXRa	retinoic acid receptor	yes	High	88	TRX
21	SRC	Src kinase	NO	---		---
22	T4_L99A	T4 lysozyme L99A	yes	Low-Moderate	42	H6
23	T4_L99A/M102Q	T4 lysozyme L99A/M102Q	yes	Low-Moderate	18	H6
24	TRYP3 (PRSS3)	trypsin	yes	NO- Very Low		TRX

How **sensitive** are binding free energies to various physical effects?

experimental details

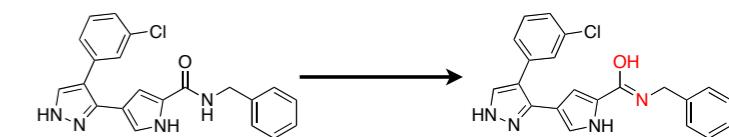
Protein conformation

- Which conformation is most likely?
- Conformational change upon binding
- Multiple conformations contributing to binding



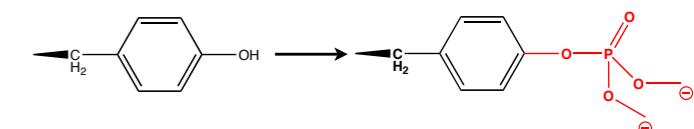
Ligand protonation/tautomeric state

- Appropriate choice of protonation/tautomeric state
- Change in protonation/tautomeric state upon binding
- Mixture of protonation/tautomeric states relevant to binding



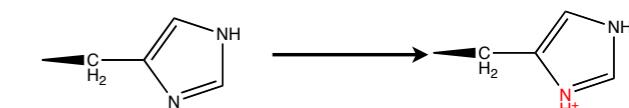
Phosphorylation state

- Conformational change upon phosphorylation
- Change in electrostatic environment



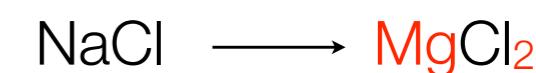
Protein protonation state

- Appropriate choice of protonation state
- Change in protonation state upon binding
- Mixture of protonation states relevant to binding



Salt environment

- Salt required for function
- Appropriate salt parameters
- Other cosalts, cosolvents, and chelators



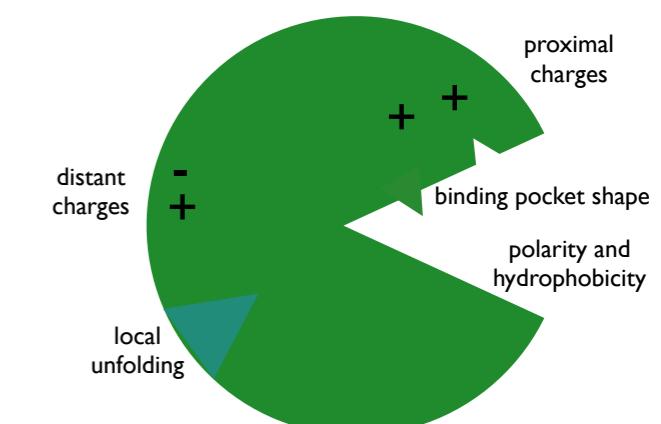
Proximal/distal charged residues

Binding pocket shape

Polarity and hydrophobicity

Local unfolding

can probe with point mutations



Could have immediate impact on current design efforts.

Iterative rounds of prediction and assessment will drive improvement of algorithms and forcefields

computational
predictions

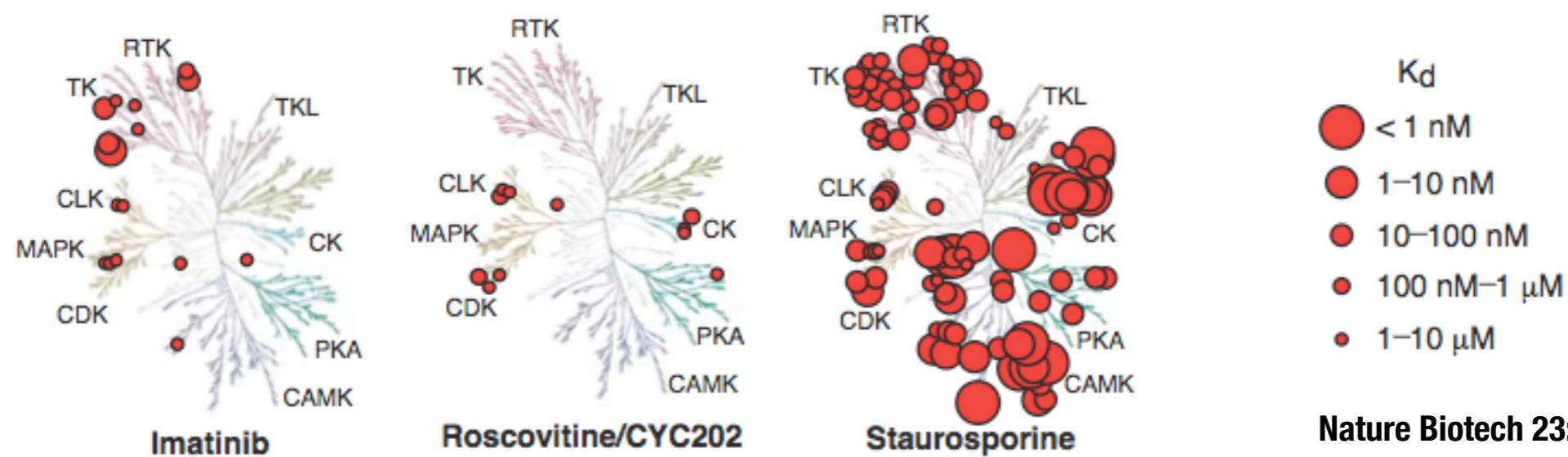


experimental
confirmation

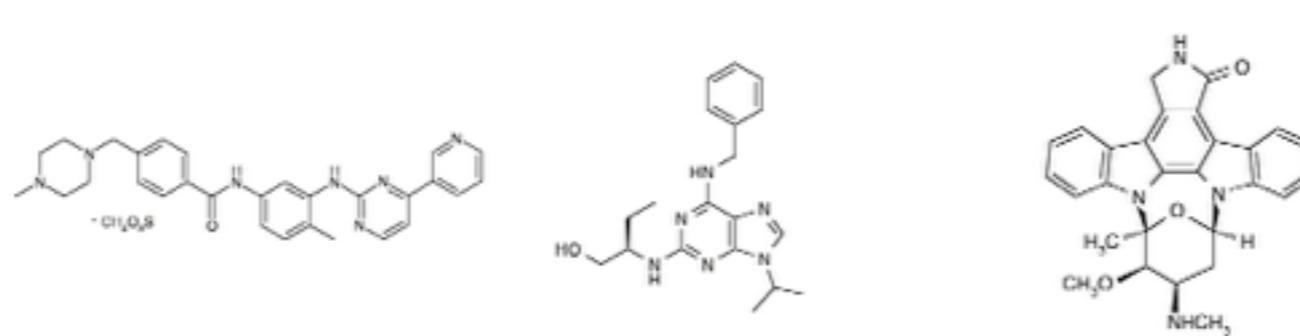
* Use combination of mutations selected from literature, by hand, and through expanded ensemble calculations, we can test mutations that will **push the limits** of current models and identify critical physical effects.

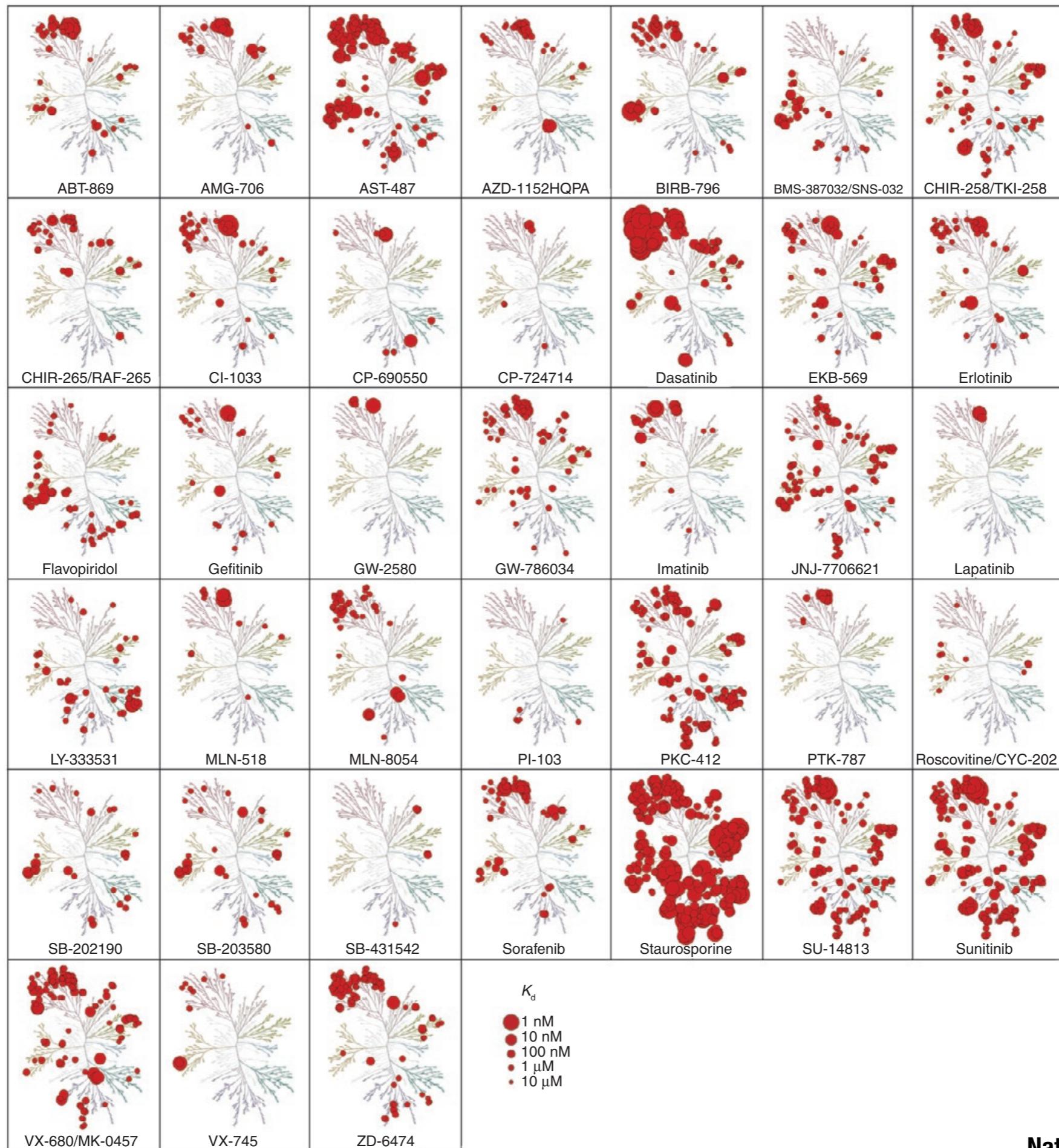
- * Improvement of algorithms to **sample conformational changes**.
- * Inclusion of **chemical effects** using nonequilibrium Monte Carlo methods.
- * **Automated Bayesian methods** to improve forcefields using data.

How can we quantitatively understand (and later design) the selectivity of inhibitors for kinases?

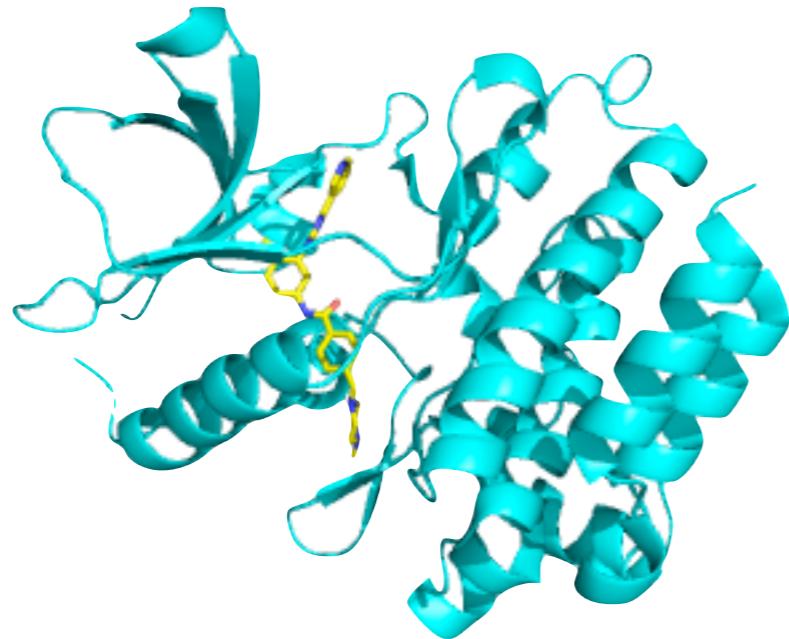


Nature Biotech 23:329, 2005





Differences in stabilities of inactive states may be responsible for origin of some kinase inhibitor selectivity



imatinib bound to Abl kinase [2HYY]

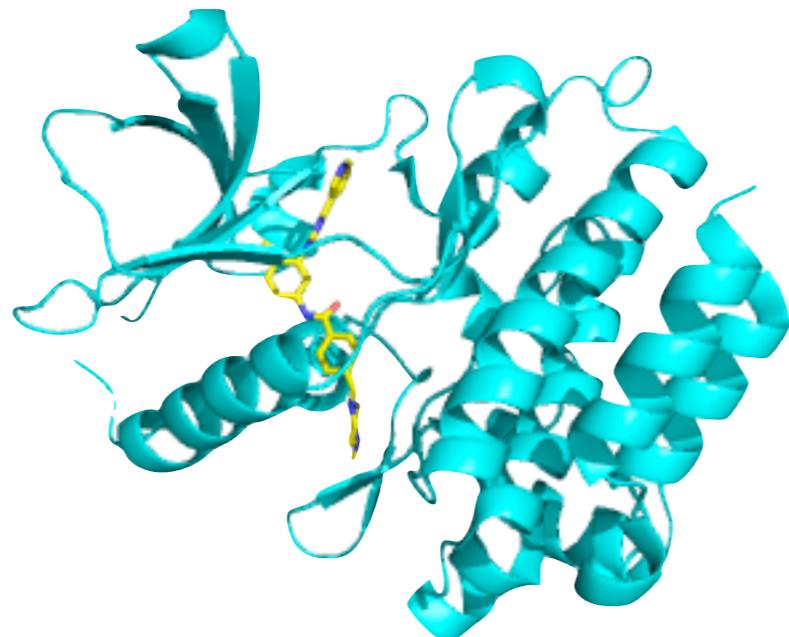


imatinib bound to Src kinase [2OIQ]

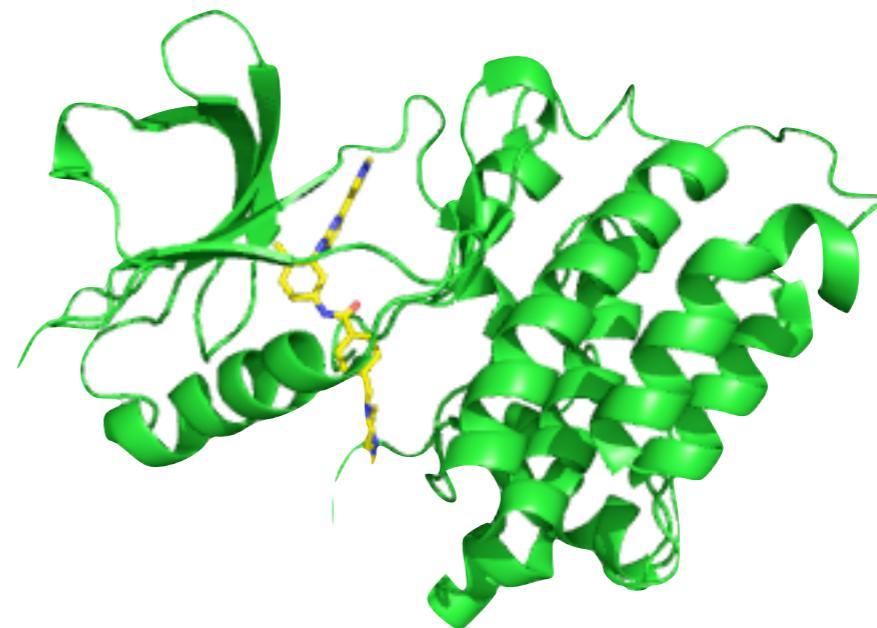
$$\Delta\Delta G = 4.6 \text{ kcal/mol} \text{ (favoring Abl binding)}$$

- * essentially same binding mode in X-ray structure!
- * essentially same interactions
- * calculations suggest no difference in binding free energy for this conformation
[Aleksandrov and Simonson , J Biol Chem 285:13807, 2010]

Differences in stabilities of inactive states may be responsible for origin of some kinase inhibitor selectivity

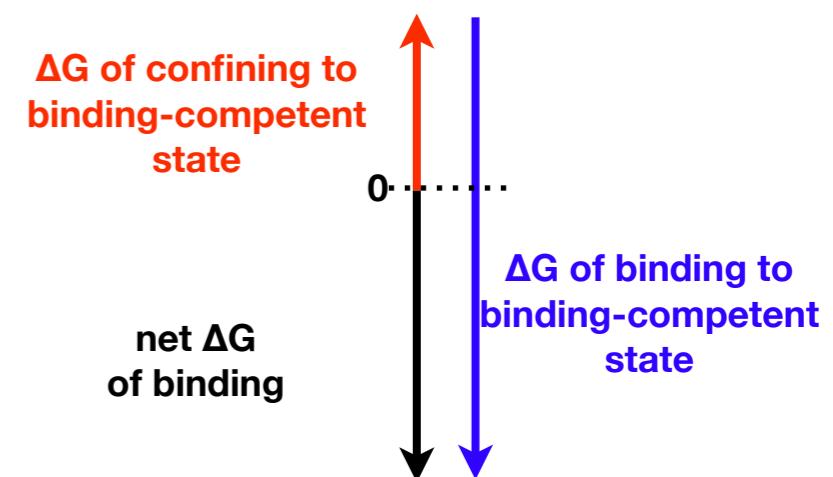
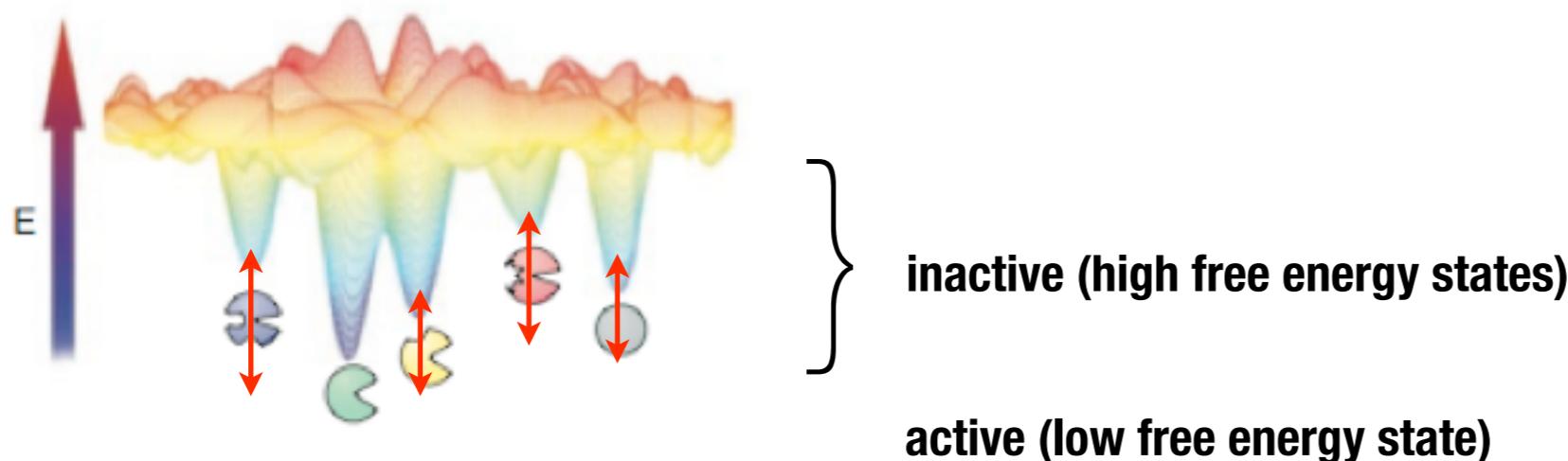


imatinib bound to Abl kinase [2HYY]



imatinib bound to Src kinase [2OIQ]

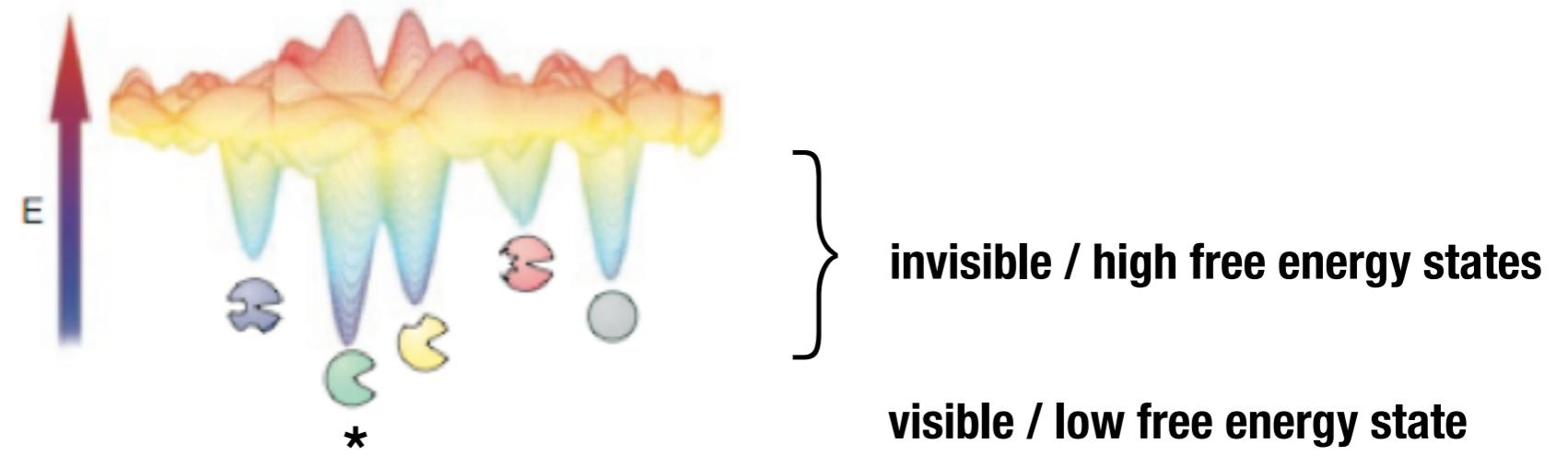
$$\Delta\Delta G = 4.6 \text{ kcal/mol} \text{ (favoring Abl binding)}$$



Enumeration of metastable states will be crucial to successful design of selective kinase inhibitors.

Identifying these conformational substates structurally and energetically is often difficult

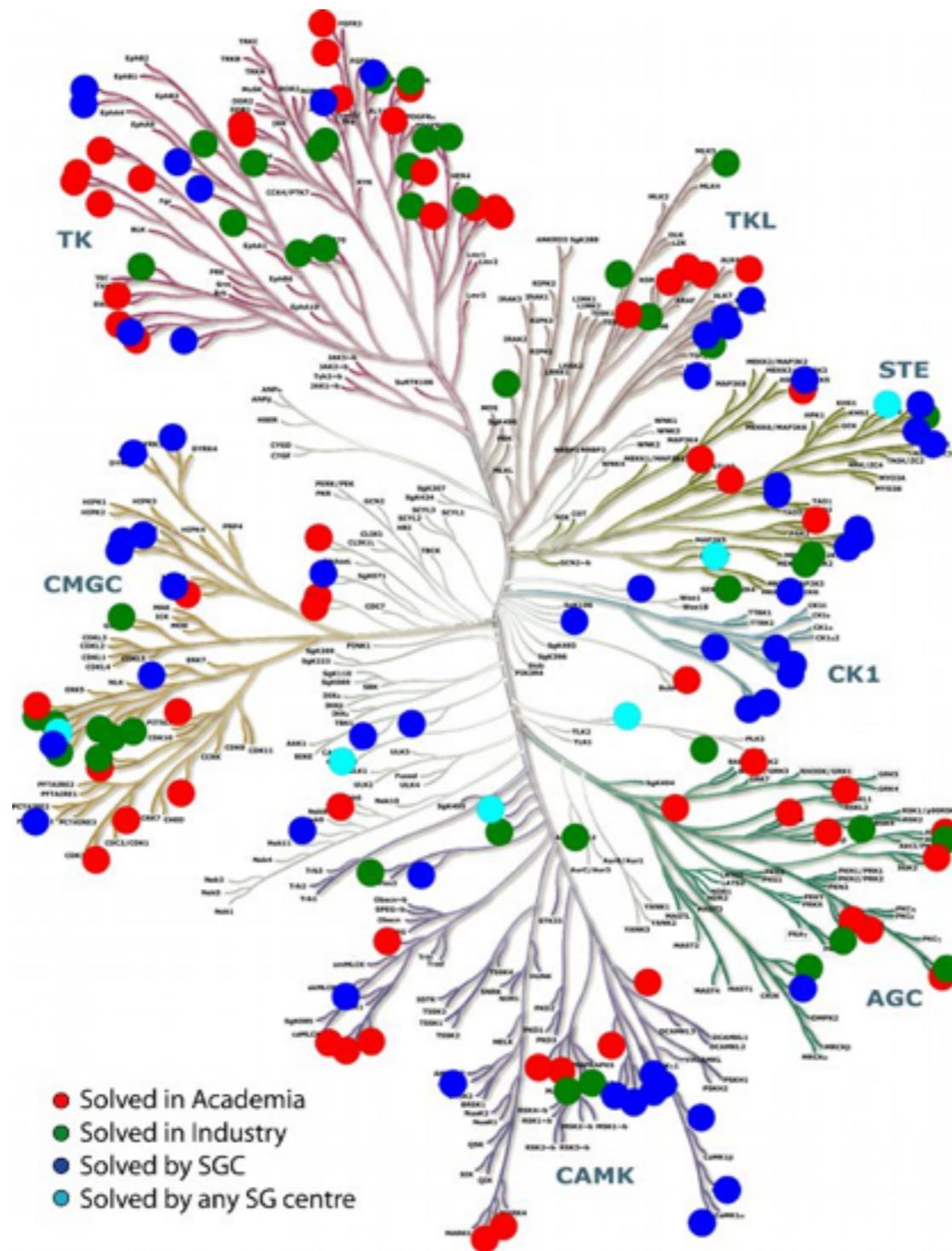
- * allow substrate access
- * expose binding surface
- * expose allosteric regulatory site
- * release product
- * alter catalytic activity
- * transduce signals



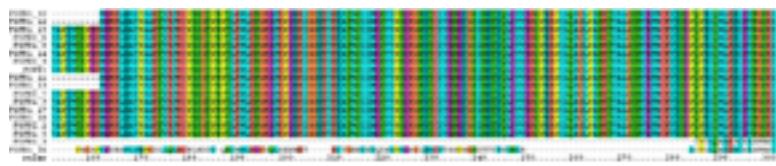
Structural biology approaches can only “see” these states if fortuitously trapped by ligands
Simulations get “stuck” in these conformations for long times

Structural data on human kinases exists, but is incomplete

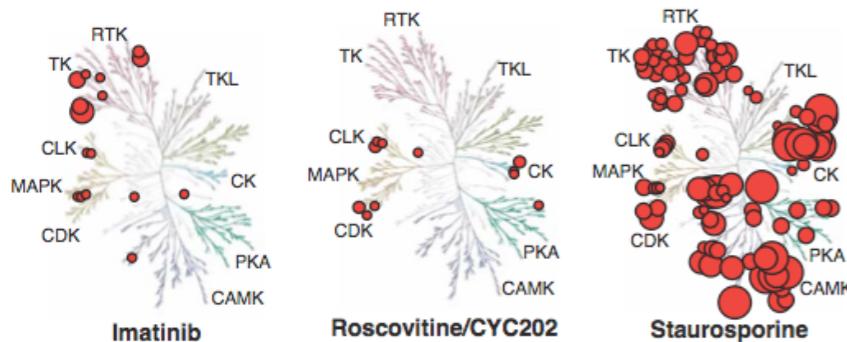
Human kinases with available structural data



sequences of targets/off-targets



template structures from Uniprot



**multiple binding free energy calculations
to compute affinities and selectivities**

Daniel Parton
Postdoc



structural models of many conformations



Patrick Grinaway
PBSB student

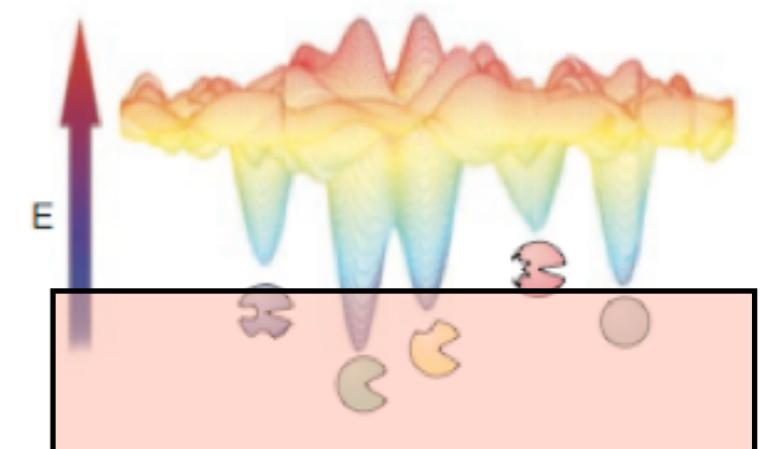
Kyle Beauchamp
Postdoc



Jan-Hendrik Prinz
Postdoc



**conformational ensembles
with relative energetics**



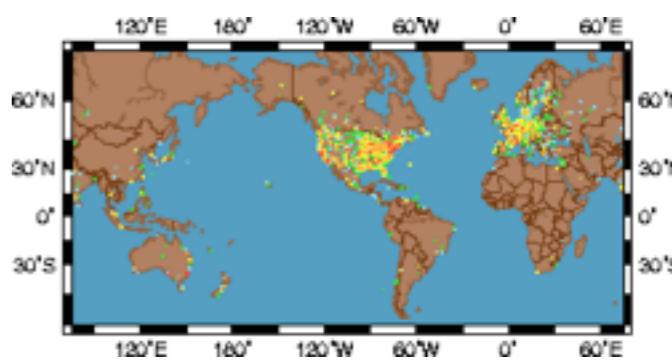
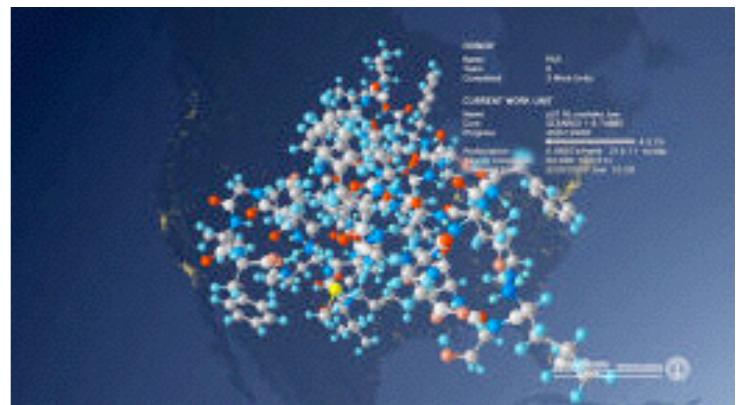
Folding@Home gives us access to enormous computational resources for probing biomolecular dynamics



Client statistics by OS

OS Type	Native TFLOPS*	x86 TFLOPS*	Active CPUs	Active Cores	Total CPUs
Windows	962	962	221141	322418	4605439
Mac OS X	11	11	3976	27804	24237
Linux	14	14	4367	24431	38685
ATI GPU	809	1707	5694	5694	358742
NVIDIA GPU	897	1893	4744	4744	289083
NVIDIA Fermi GPU	4768	10060	14065	14065	298666
Total	7461	14647	253987	399156	5614852

Table last updated at Mon, 20 May 2013 06:14:30



Over 14 PFLOP/s of aggregate computational power!

<http://folding.stanford.edu>

Folding@Home enables whole-kinome simulation

518 human protein kinases
excluding splice and disease variants

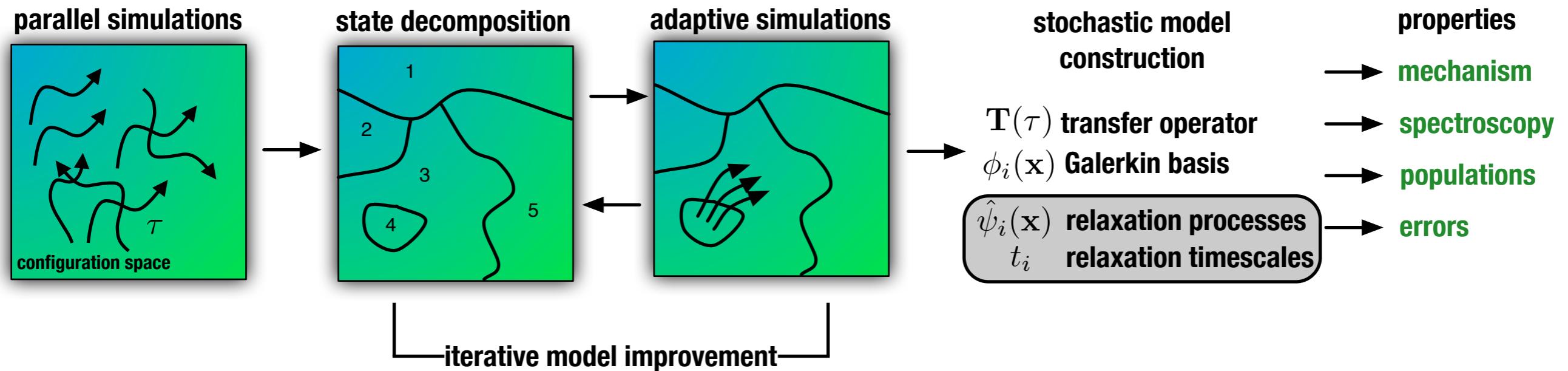
X **3,507** kinase catalytic domain structures
in UniProt

= **1,816,626** kinase models will be built and refined
on new MSKCC compute resources housed at SDSC

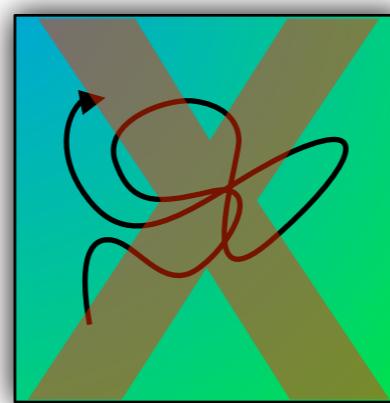
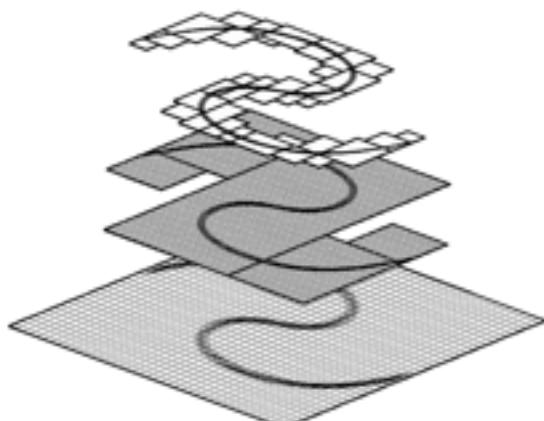
~ 18,166,260 kinase simulations on Folding@Home
over one year



Adaptive algorithms can efficiently build stochastic dynamical models for biomolecules from short trajectories



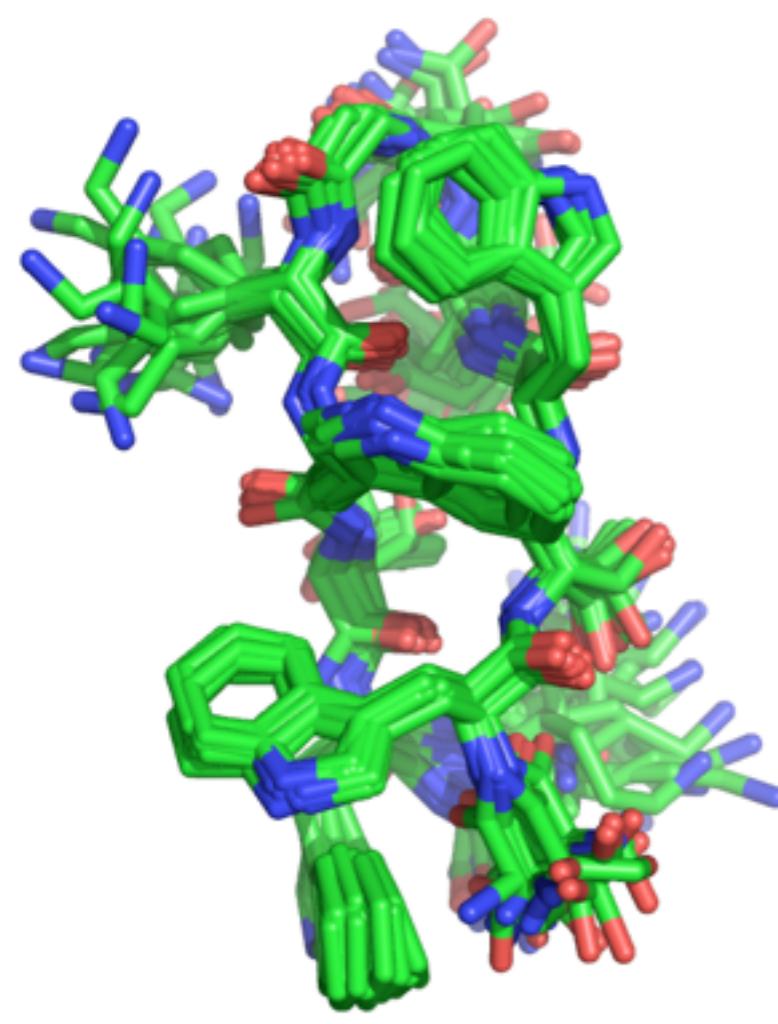
Similar in spirit to **adaptive mesh refinement** algorithms in engineering,
very different from traditional approach of running a single long simulation.



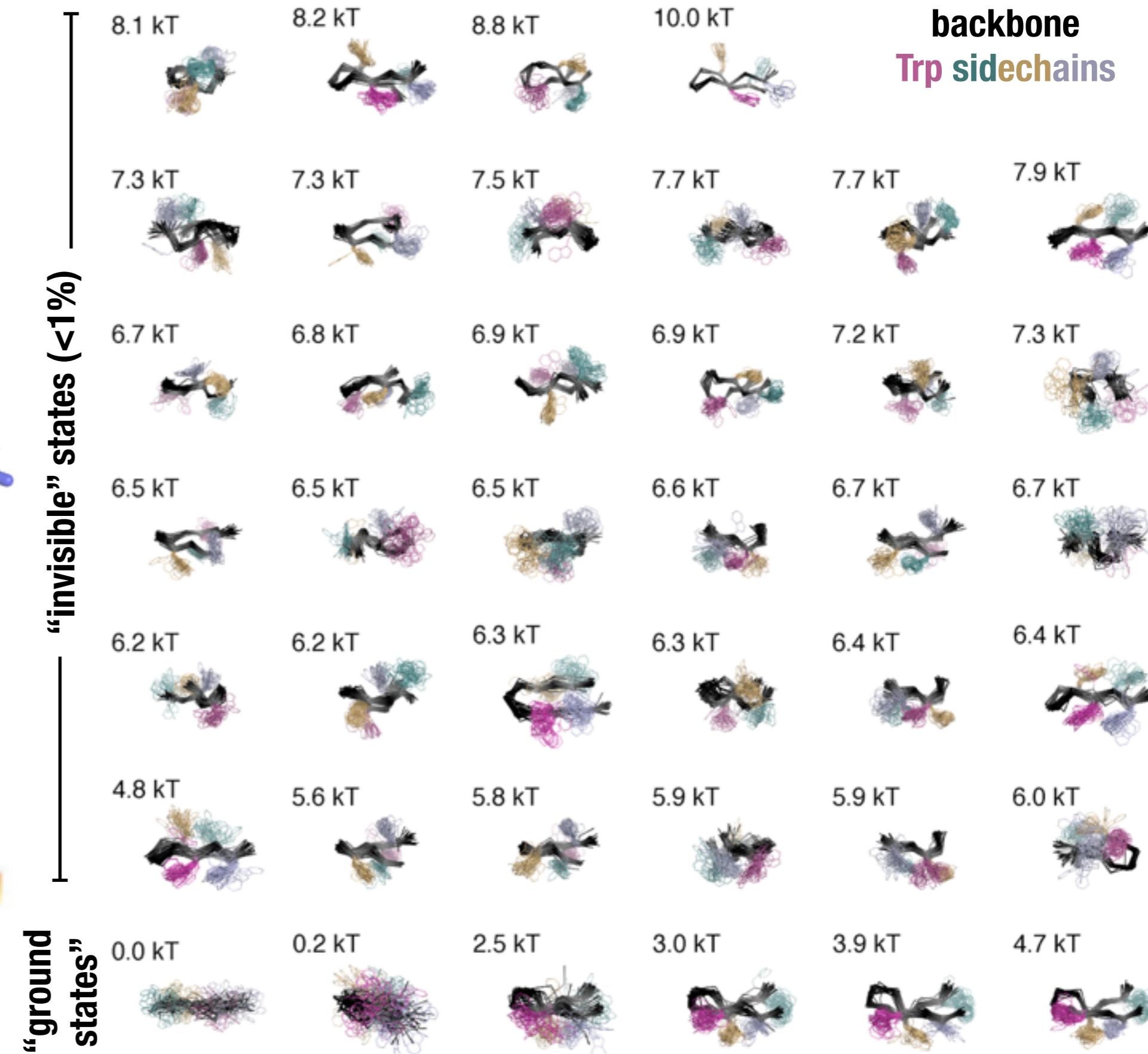
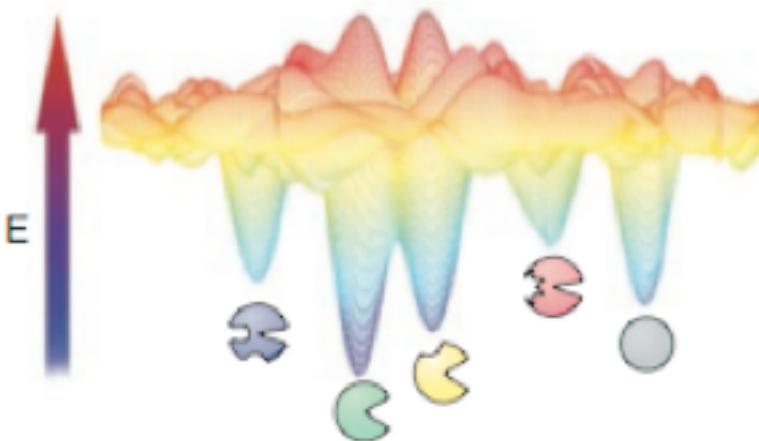
Chodera, Singhal, Swope, Pande, Dill. JCP 126:155101, 2007.
Hinrichs and Pande. JCP 126:244101, 2007.
Bacallado, Chodera, Pande. JCP 131:045106, 2009.
Noé. JCP 128:244103, 2008.
Chodera and Noé. JCP 133:105102, 2010.

NMR model of trpzip2 at 288 K

Many distinct metastable states can be identified at $T \sim T_m$

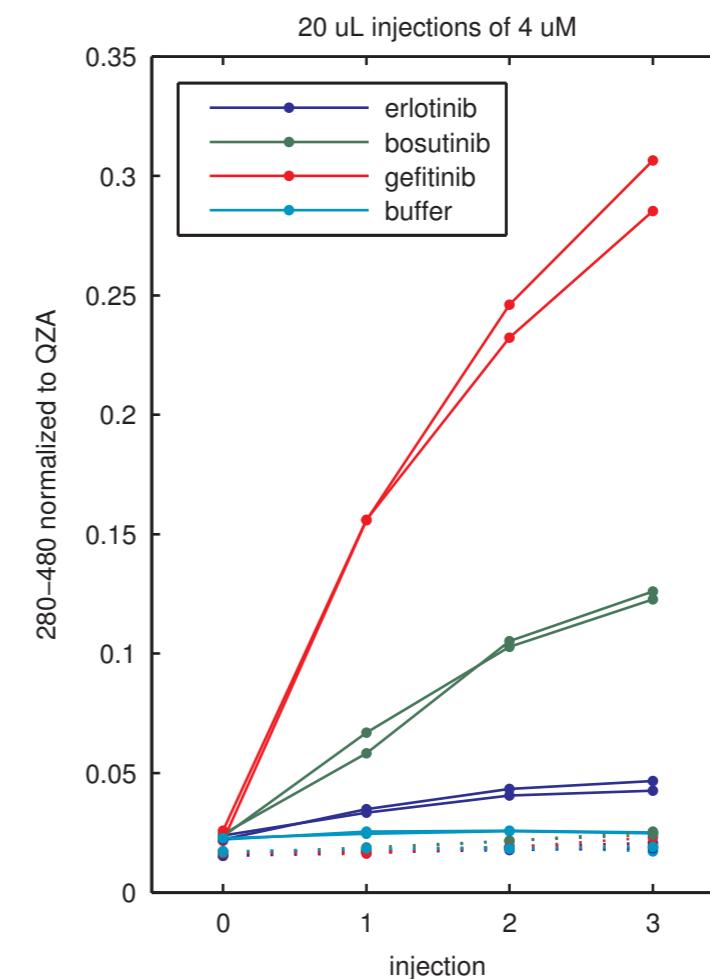
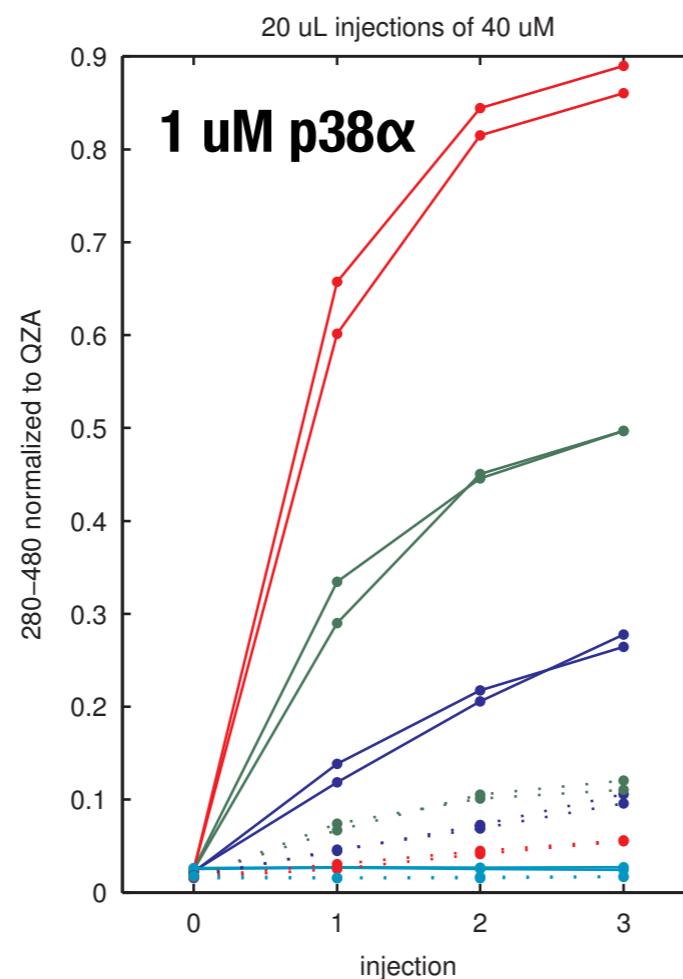
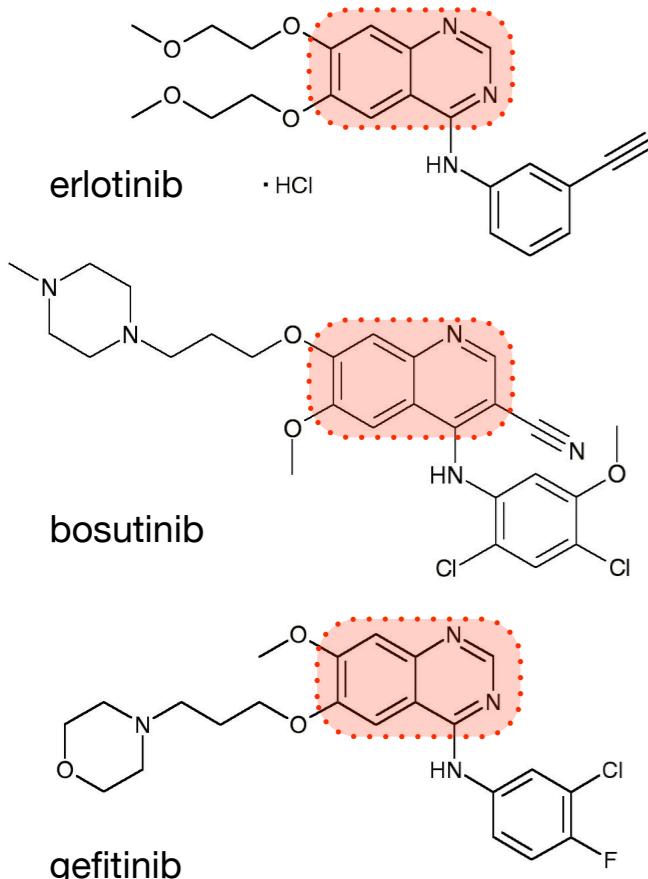


trpzip2
[PDB:1LE1]



High-throughput fluorescence assays can measure binding affinities of a panel of ligands to kinase mutants

fluorescent inhibitors

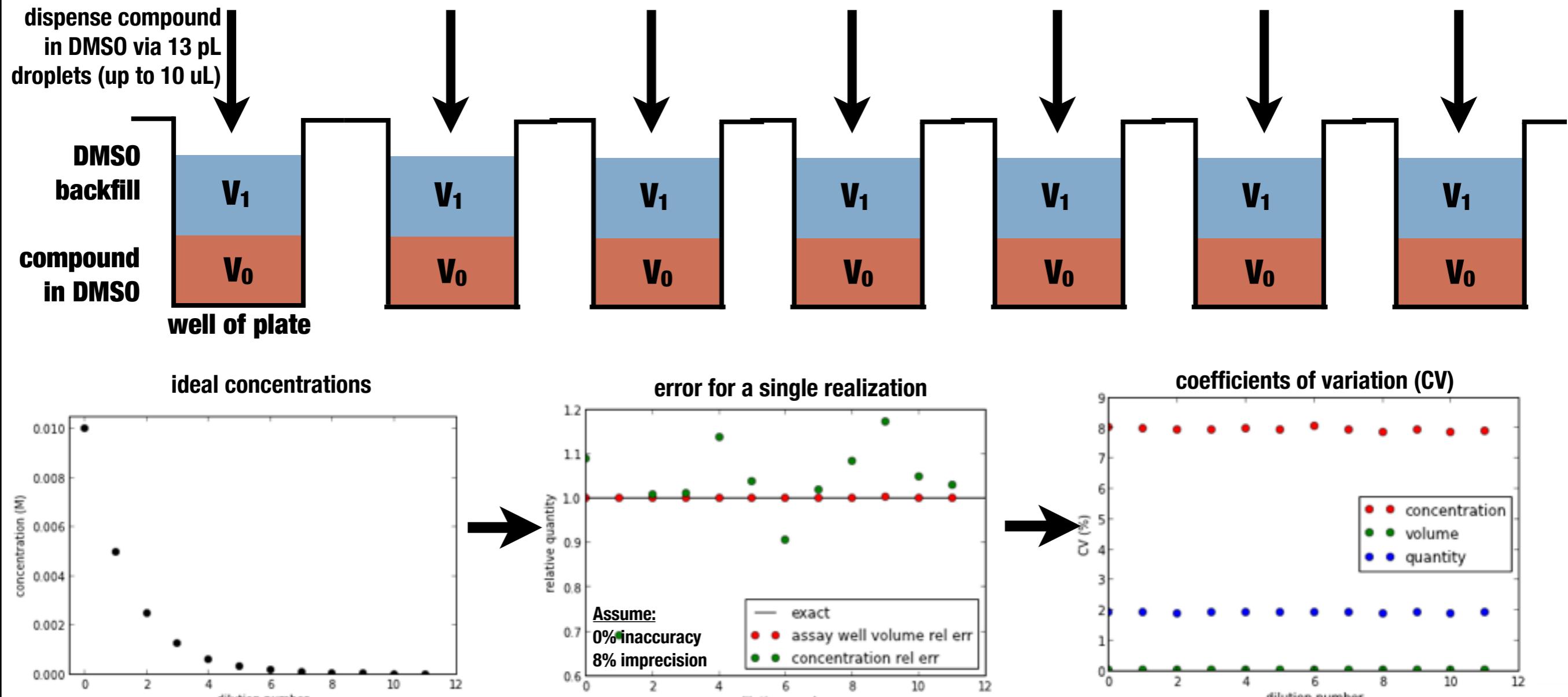


**excite 280 nm
[Trp FRET ~350 nm]
measure 480 nm**

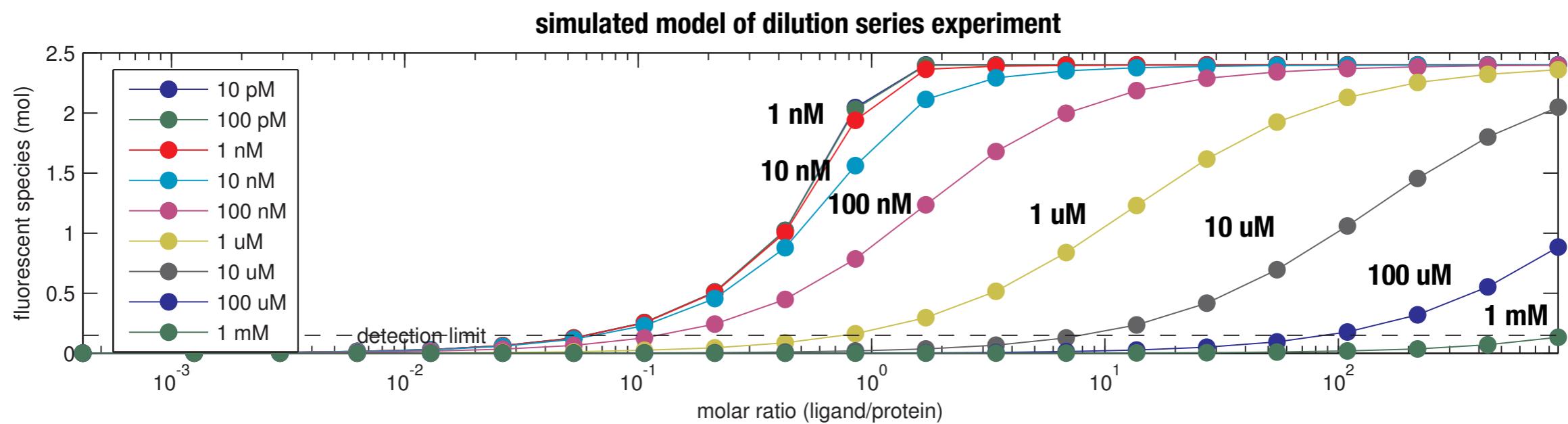
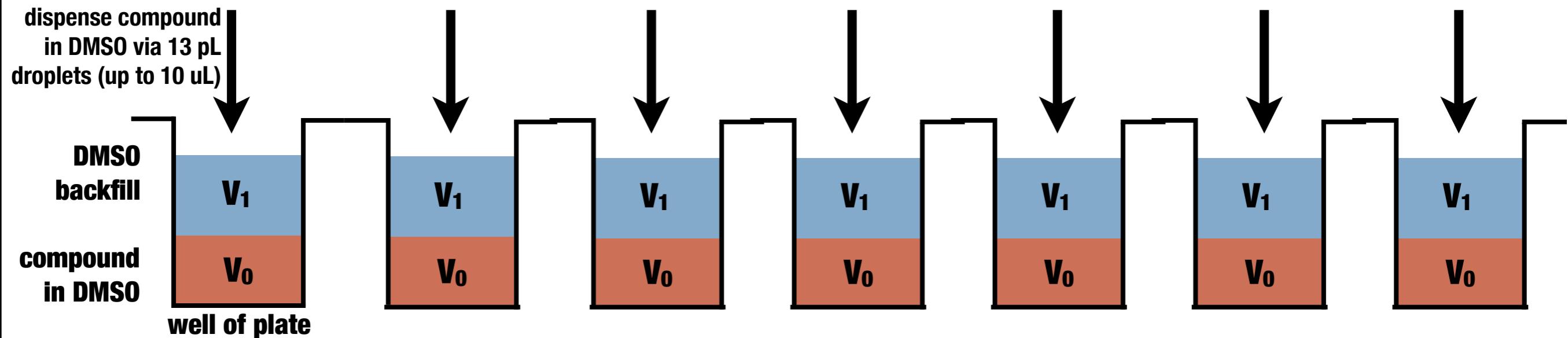


with Nick Levinson, Boxer lab, Stanford

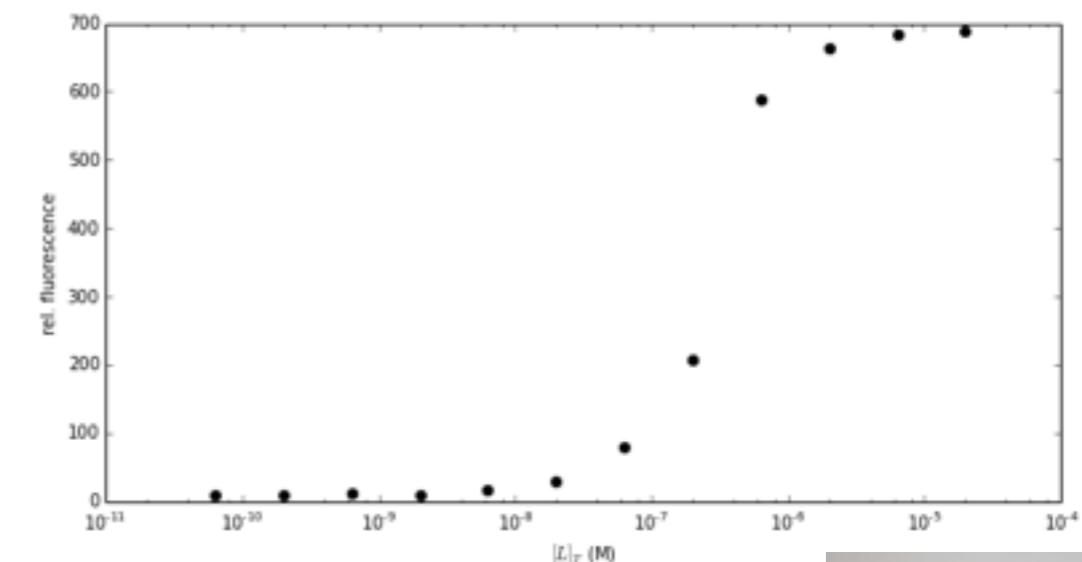
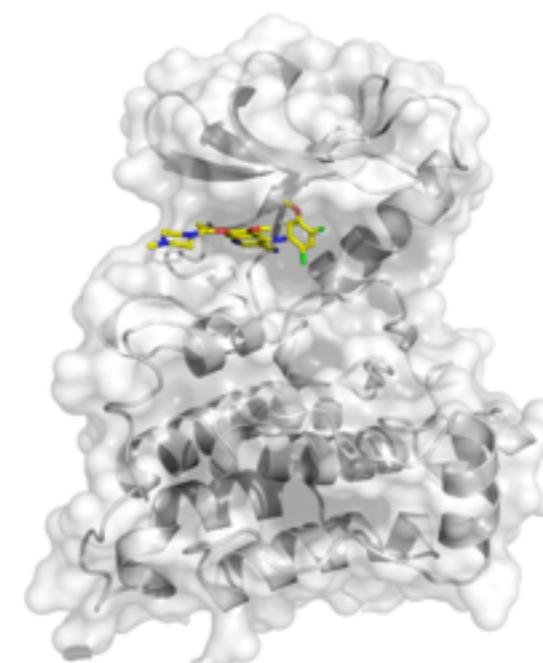
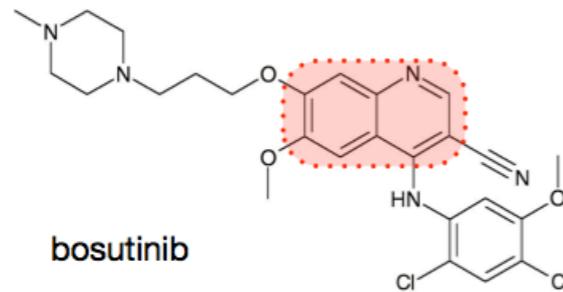
Modeling of biophysical experiments can be a helpful tool for designing protocols for optimal information content



Modeling of biophysical experiments can be a helpful tool for designing protocols for optimal information content



Bayesian inference allows us to use tools we know (equilibrium sampling) to quantify experimental uncertainty



Sonya Hanson
Postdoc



Bayesian inference allows us to use Monte Carlo sampling to quantify experimental uncertainty

binding model



pymc model

<https://pypi.python.org/pypi/pymc>

pymc sampling

```
import pymc

# Two-component binding.
def two_component_binding(DeltaG, P, L):
    Kd = np.exp(-DeltaG)
    PL = 0.5 * ((P + L + Kd) - np.sqrt((P + L + Kd)**2 - 4*P*L)); # complex concentration (M)
    P = P - PL; # free protein concentration in sample cell after n injections (M)
    L = L - PL; # free ligand concentration in sample cell after n injections (M)
    return [P, L, PL]

# Create a pymc model
def make_model(Pstated, dPstated, Lstated, dLstated, Fobs_i):
    N = len(Lstated)

    # Prior on binding free energies.
    DeltaG = pymc.Uniform('DeltaG', lower=-20, upper=+20, value=0.0) # binding free energy (kT)

    # Priors on true concentrations of protein and ligand.
    Ptrue = pymc.Lognormal('Ptrue', mu=np.log(Pstated**2 / np.sqrt(dPstated**2 + Pstated**2)), tau=np.sqrt(np.log(1.0/dPstated**2 + 1.0/Pstated**2)))
    Ltrue = pymc.Lognormal('Ltrue', mu=np.log(Lstated**2 / np.sqrt(dLstated**2 + Lstated**2)), tau=np.sqrt(np.log(1.0/dLstated**2 + 1.0/Lstated**2)))

    # Priors on fluorescence intensities of complexes (later divided by a factor of Pstated for scale).
    F_background = pymc.Gamma('F_background', alpha=0.1, beta=0.1, value=Fobs_i.min()) # background fluorescence
    F_PL = pymc.Gamma('F_PL', alpha=0.01, beta=0.01, value=Fobs_i.max()) # complex fluorescence
    F_L = pymc.Gamma('F_L', alpha=0.01, beta=0.01, value=Fobs_i.max()) # ligand fluorescence

    # Unknown experimental measurement error.
    log_sigma = pymc.Uniform('log_sigma', lower=-3, upper=+3, value=0.0)
    @pymc.deterministic
    def precision(log_sigma=log_sigma): # measurement precision
        return 1.0 / np.exp(log_sigma)

    # Fluorescence model.
    @pymc.deterministic
    def Fmodel(F_background=F_background, F_PL=F_PL, F_P=F_P, F_L=F_L, Ptrue=Ptrue, Ltrue=Ltrue, DeltaG=DeltaG):
        Fmodel_i = np.zeros([N])
        for i in range(N):
            [P, L, PL] = two_component_binding(DeltaG, Ptrue, Ltrue[i])
            Fmodel_i[i] = (F_PL*PL + F_L*L) / Pstated + F_background
        return Fmodel_i

    # Experimental error on fluorescence observations.
    Fobs_i = pymc.Normal('Fobs_i', mu=Fmodel, tau=precision, size=[N], observed=True, value=Fobs_i) # observed data

    # Construct dictionary of model variables.
    pymc_model = { 'Ptrue' : Ptrue, 'Ltrue' : Ltrue,
                   'log_sigma' : log_sigma, 'precision' : precision,
                   'F_PL' : F_PL, 'F_P' : F_P, 'F_L' : F_L, 'F_background' : F_background,
                   'Fmodel_i' : Fmodel, 'Fobs_i' : Fobs_i, 'DeltaG' : DeltaG }
    return pymc_model

# Uncertainties in protein and ligand concentrations.
dPstated = 0.10 * Pstated # protein concentration uncertainty
dLstated = 0.08 * Lstated # ligand concentration uncertainty (due to gravimetric preparation and HP D300 dispensing)

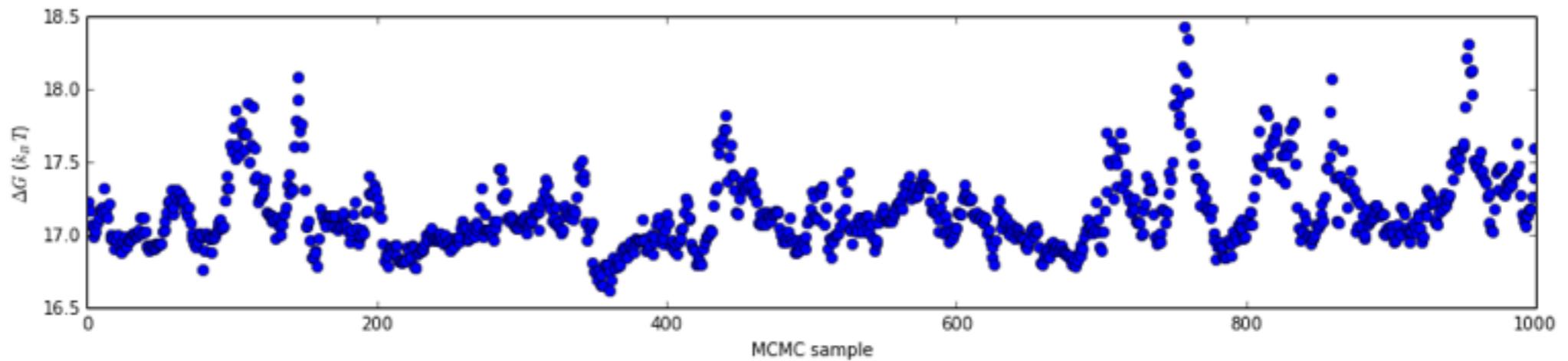
# Build model.
pymc_model = pymc.Model(make_model(Pstated, dPstated, Lstated, dLstated, F_i))

# Sample with MCMC
mcmc = pymc.MCMC(pymc_model, db='ram', name='Sampler', verbose=True)
mcmc.sample(iter=100000, burn=50000, thin=50, progress_bar=False)
```

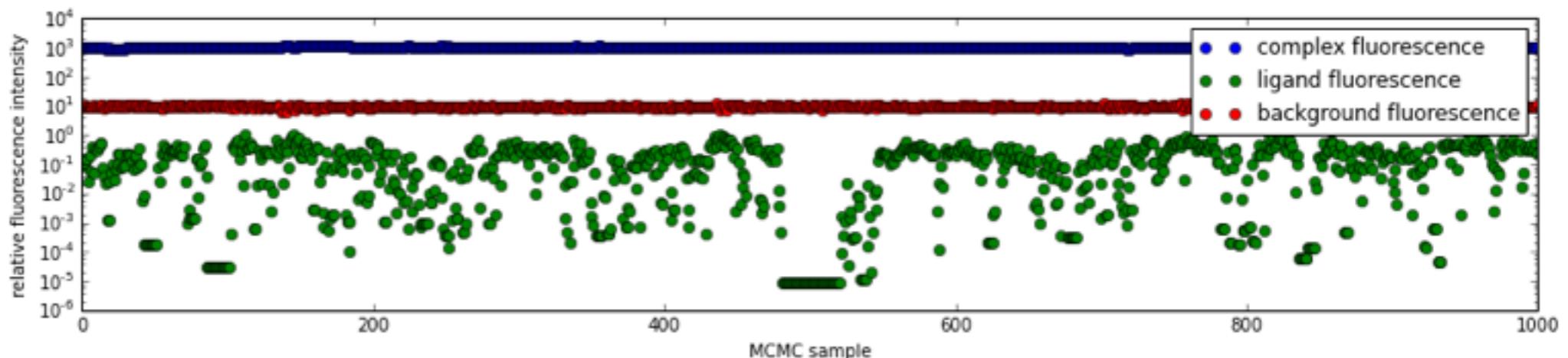
Code at <https://github.com/choderalab/cup-xiv>

Bayesian inference allows us to use tools we know (Monte Carlo sampling) to quantify experimental uncertainty

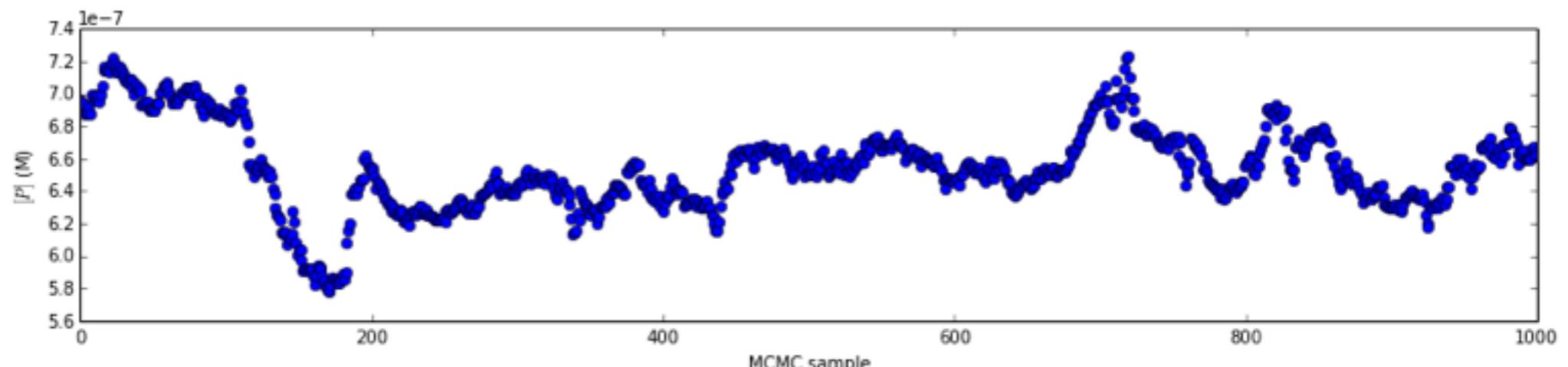
MCMC sampling
of free energies



MCMC sampling
of intrinsic ligand
and complex
fluorescence

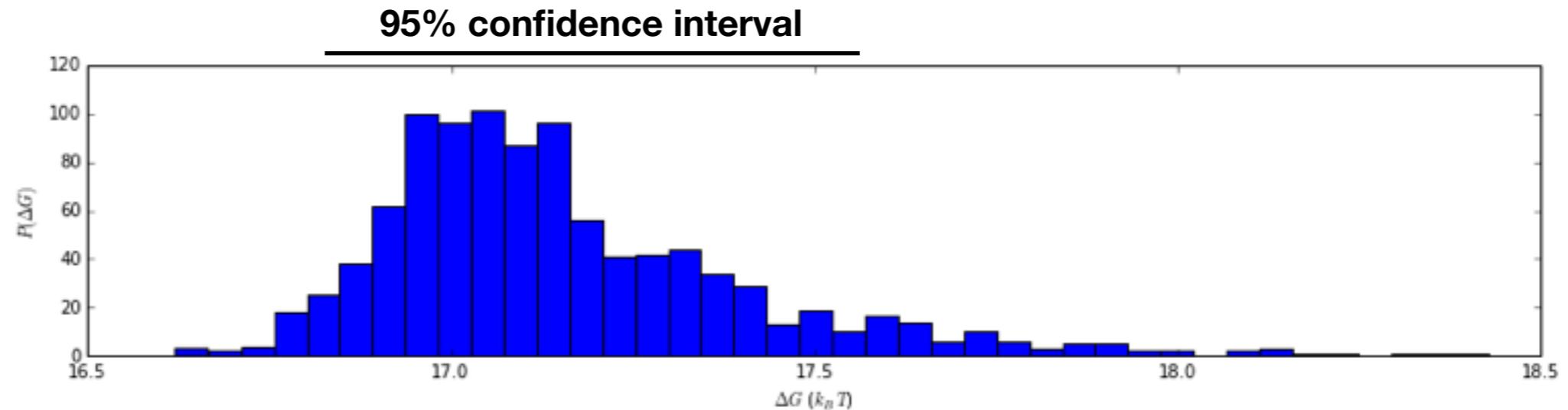


MCMC sampling
of true concentrations

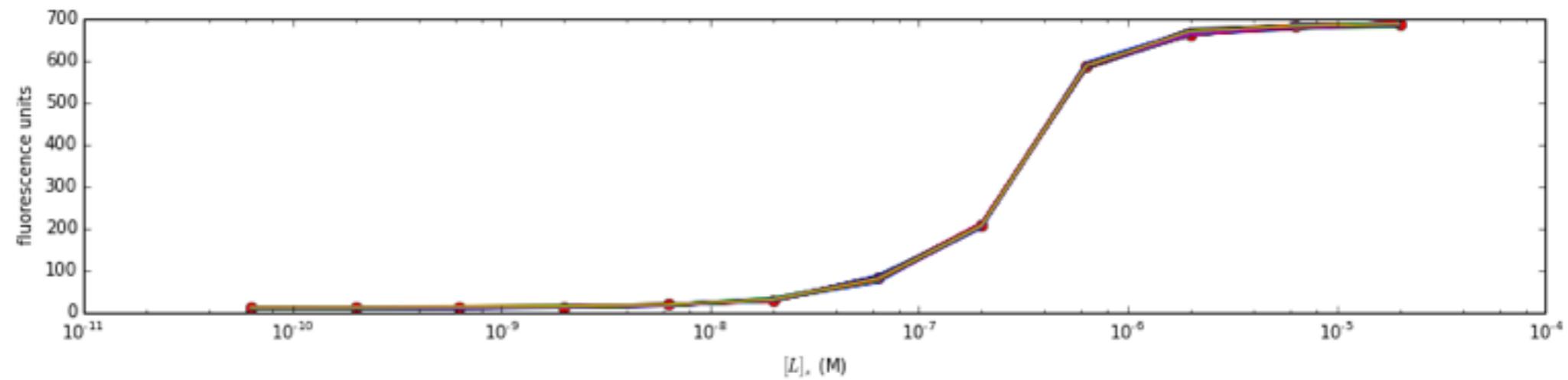


Bayesian inference allows us to use tools we know (Monte Carlo sampling) to quantify experimental uncertainty

posterior distribution
of binding free energies



family of models
fit to experimental
fluorescence data



Assays are expensive: Can we change that?

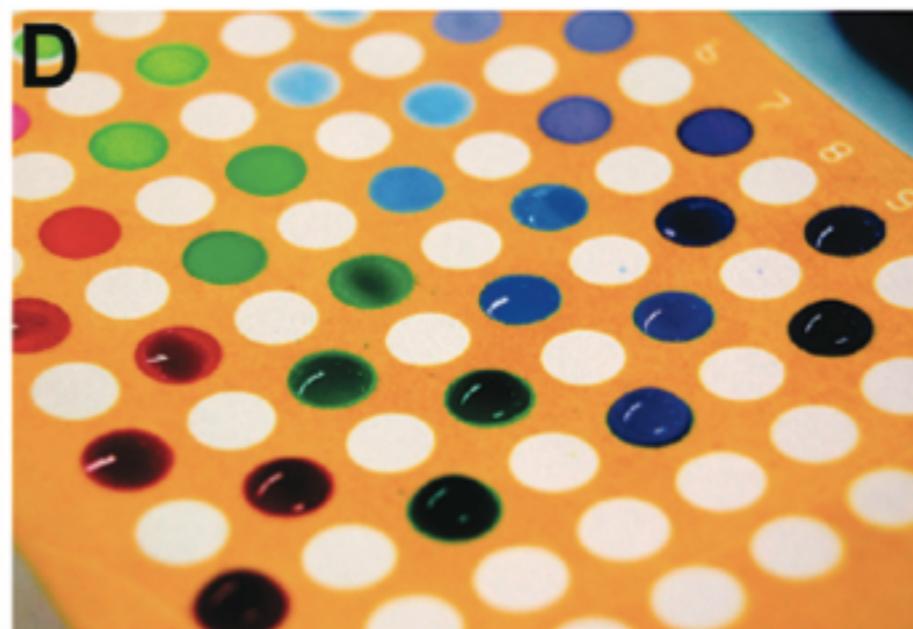
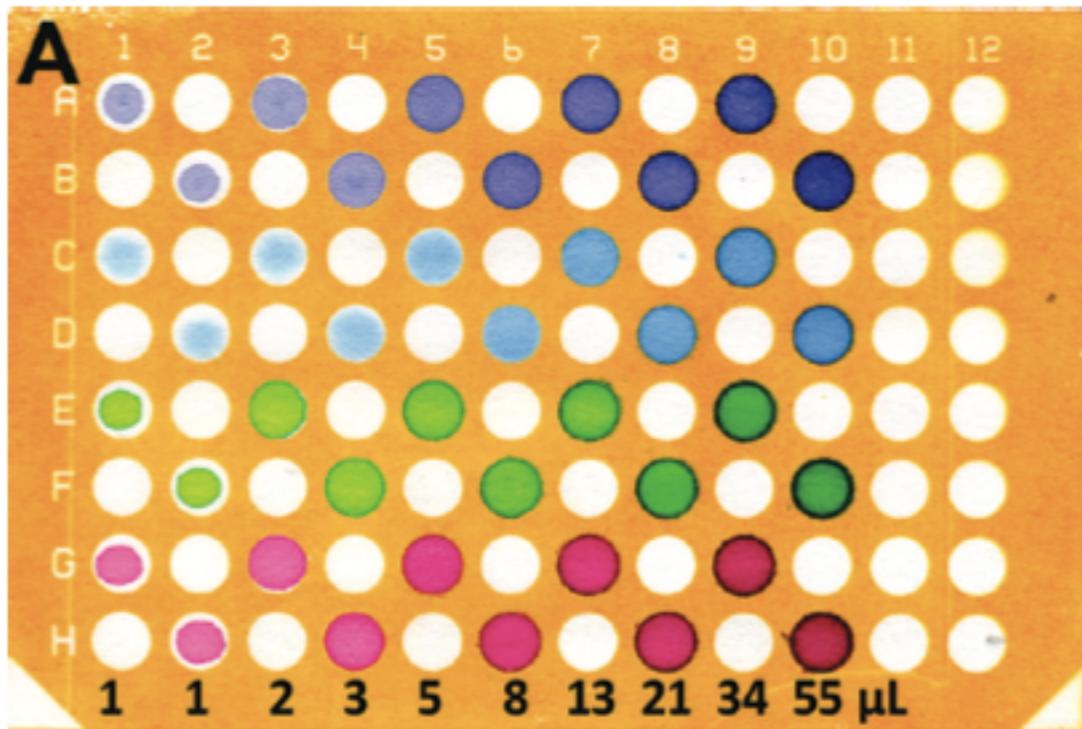


assay plate: \$10/plate
protein cost: \$22/plate

Paper Microzone Plates

Emanuel Carrilho,^{*,†,‡} Scott T. Phillips,^{†,§} Sarah J. Vella,[†] Andres W. Martinez,[†] and George M. Whitesides^{*,†}

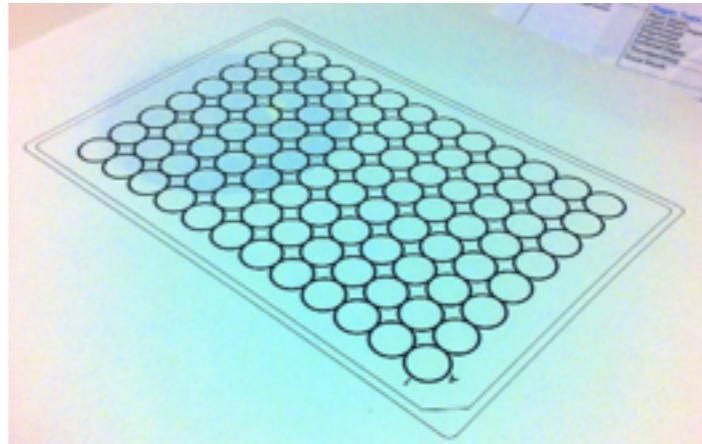
Anal. Chem. 2009, 81, 5990–5998



Assays are expensive: Can we change that?



Xerox ColorQube 8570
wax printer



patterned wax microplate



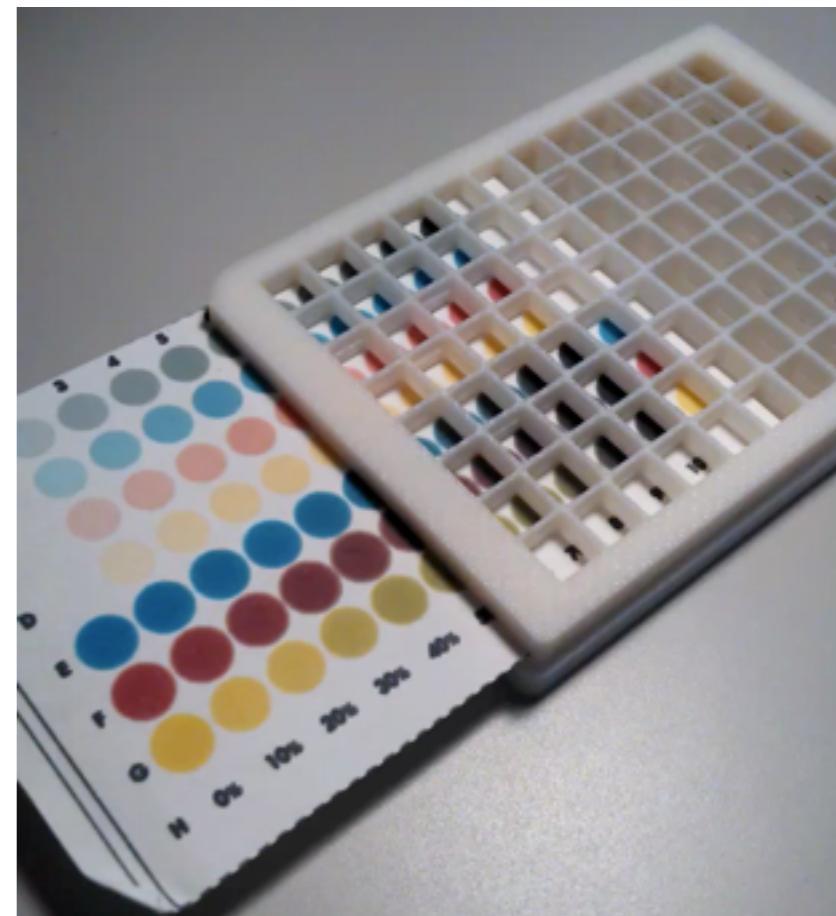
reflow wax through paper
with hot plate



Tecan HP D300
inkjet printer for compounds in DMSO



3D printed plate carrier



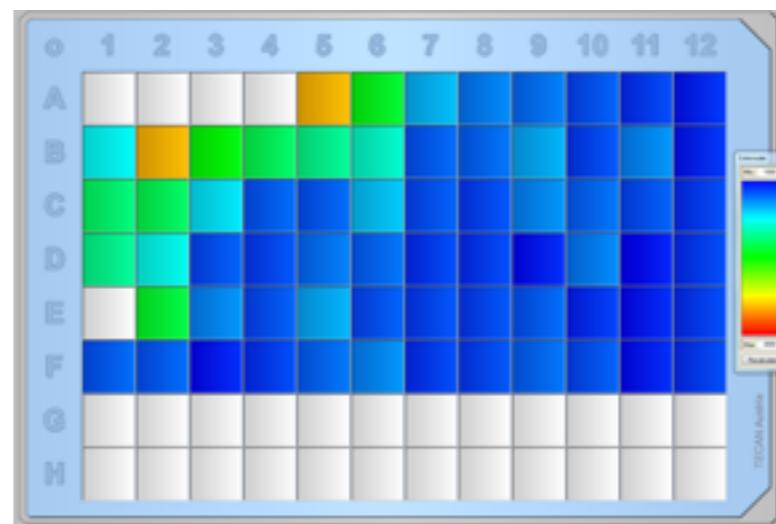
Assays are expensive: Can we change that?



assay plate: \$10/plate
protein cost: \$22/plate

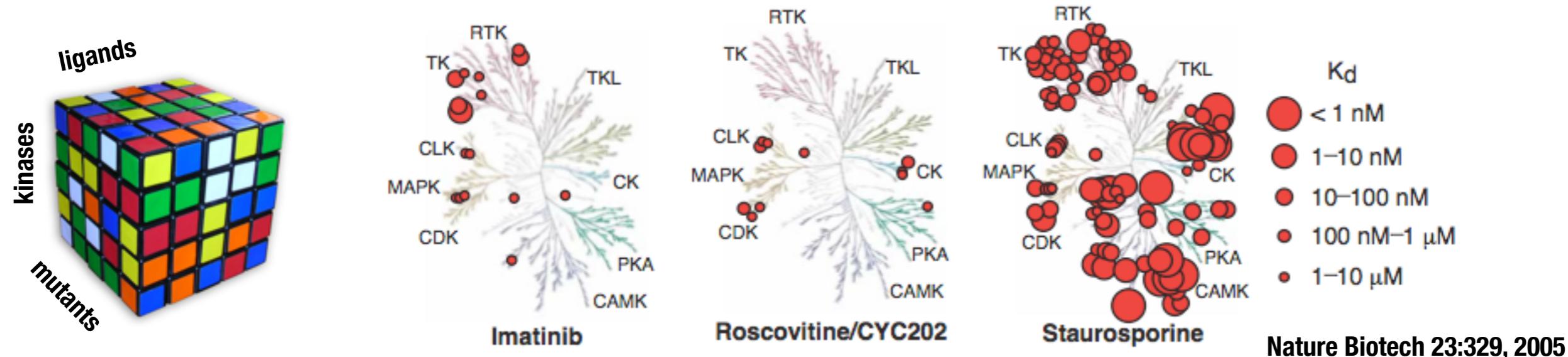


paper microzone plate: \$0.20/plate
protein cost: \$2/plate
10x reduction in costs!

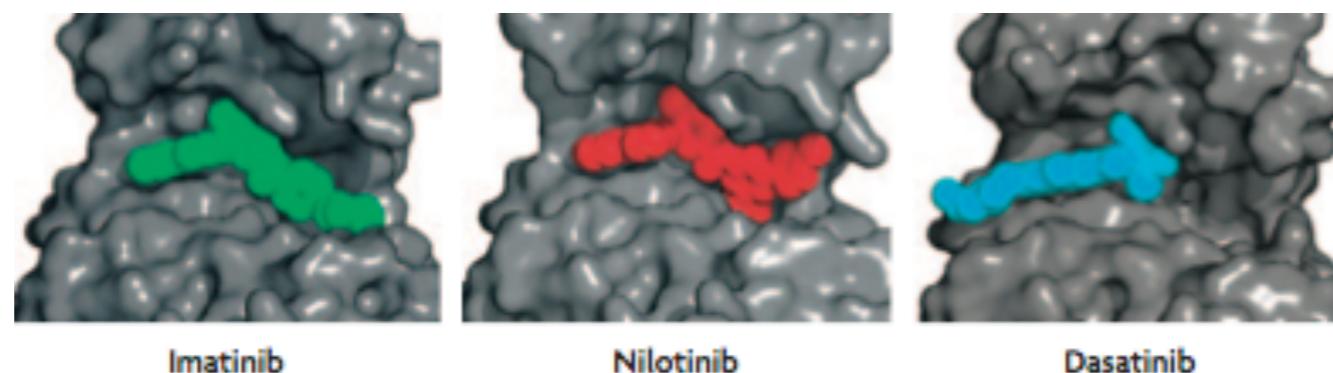


What determines selectivity of inhibitors for kinases?

High-throughput fluorescence measurements and free energy calculations can address physical determinants of kinase inhibitor selectivity:



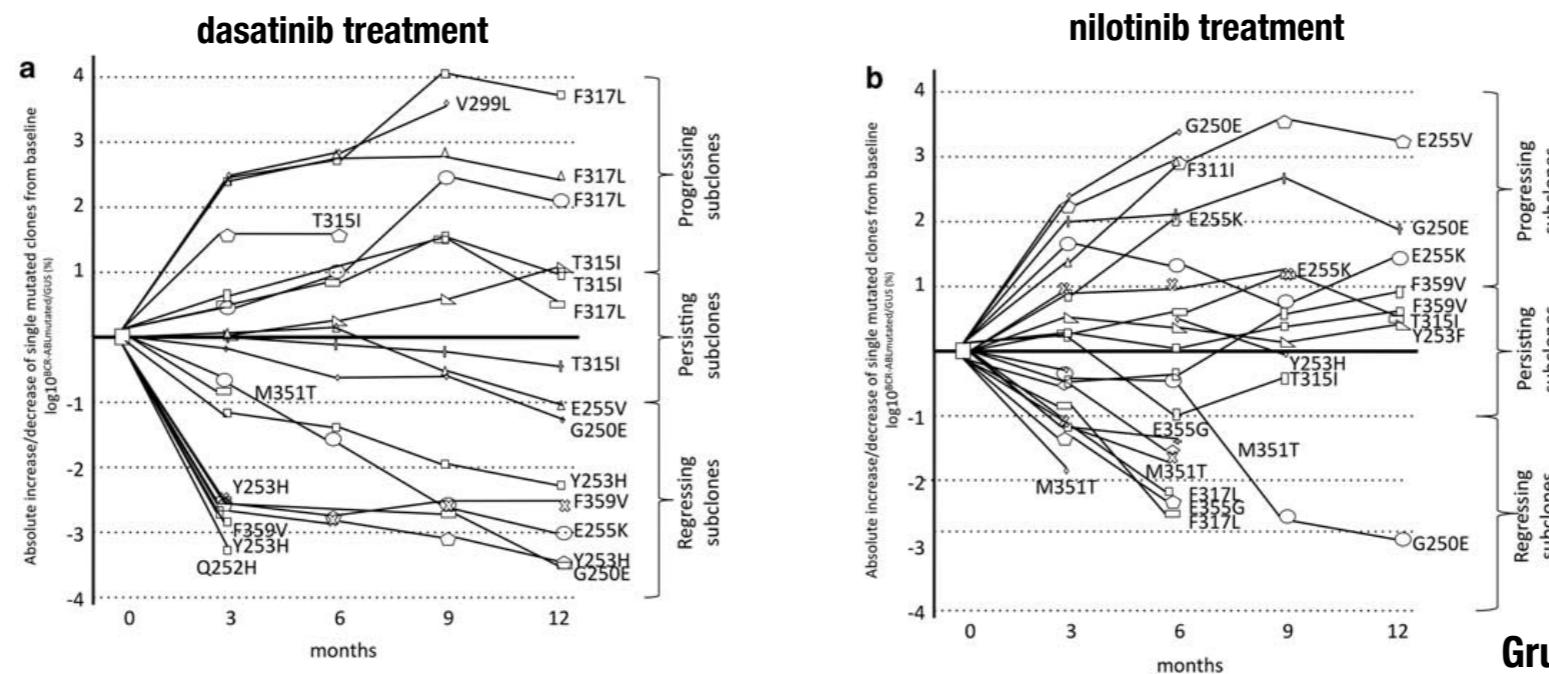
- * Are particular ligand scaffolds privileged with specificity?
- * Are particular binding modes better for specificity?
- * Are certain kinases inherently more promiscuous?



Weisberg et al.
Nature Rev. Cancer 7:345, 2007.

Can we develop a physical model of resistance mutations?

Treatment of CML with imatinib often induces resistance, predominantly E255K, T315I
Second-line drugs elicit further resistance:



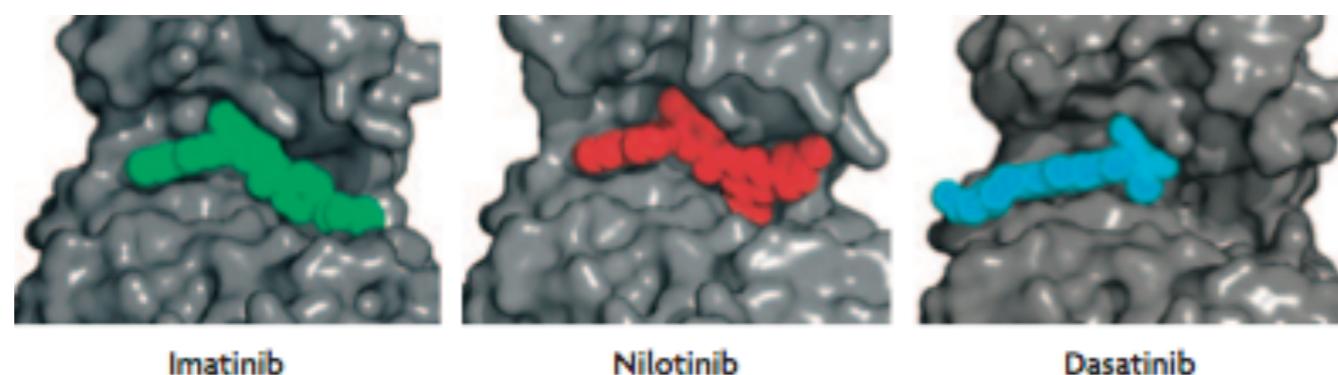
Gruber et al. Leukemia 26:172, 2012.

We can hypothesize and test a simple physical mechanism of resistance:

Resistance mutations reduce inhibitor binding affinity but retain ATP affinity (a surrogate for activity)

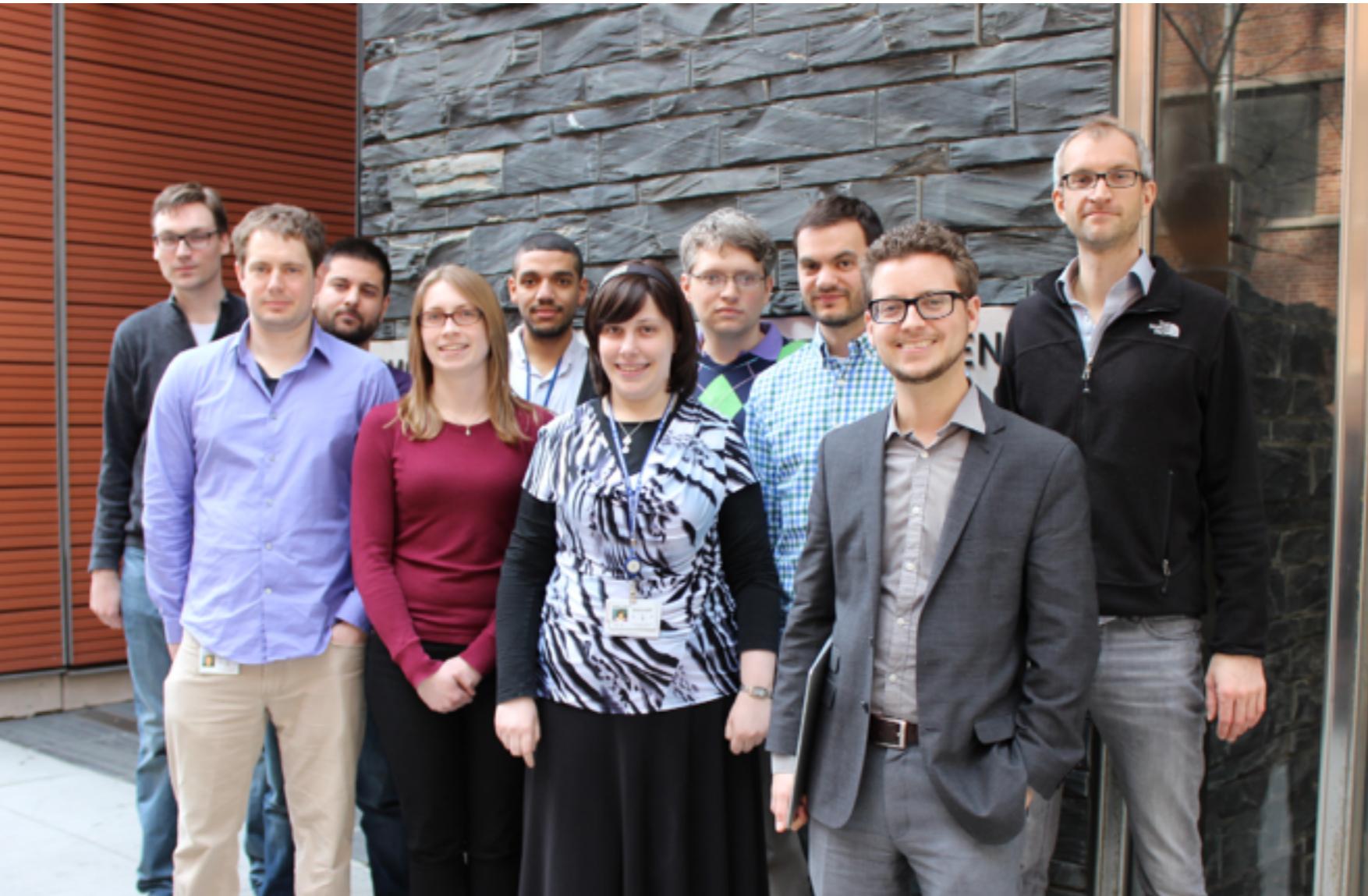
* Are certain inhibitors or binding modes less likely to elicit resistance?

* Can we incorporate likelihood of eliciting resistance mutations into rational ligand design?



Weisberg et al.
Nature Rev. Cancer 7:345, 2007.

The Chodera Lab @ MSKCC



Bas Rustenburg

PBSB

Guillermo Perez-Hernandez

visitor

Kadeem Ho Sang

admin

Danny Parton

postdoc

Jan-Hendrik Prinz

postdoc

Kyle Beauchamp

postdoc

Sonya Hanson

postdoc

Chaya Stern

TPCB

Patrick Grinaway

PBSB

Code and data available at <http://www.choderlab.org>

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

Collaborators

Stanford

Vijay Pande
Sergio Bacallado
NIH SimBios

IBM Almaden

Bill Swope
Jed Pitera

University of Chicago

Nina Singhal Hinrichs

UC Irvine

David Mobley

CCNY

Marilyn Gunner

Stony Brook

Ken Dill
Markus Seeliger

UCSF

Brian Shoichet
David Sivak

University of Virginia

Michael Shirts

Duke

David Minh

Freie Universität Berlin

Frank Noé
Bettina Keller
Jan-Hendrik Prinz
Antonia S. J. S. Mey

Rutgers

Zhiqiang Tan

UC Berkeley

Phillip Elms (BioRad)
Susan Marqusee
Carlos Bustamante
Christian Kaiser
Gheorghe Christol
Suri Vaikuntanathan

LBNL

Gavin Crooks

Vanderbilt

Joel Tellinghuisen

Hessian Informatics

Kim Branson

Vertex Pharmaceuticals

Richard Dixon