REDESIGNING DRUG DESIGN

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

Memorial Sloan-Kettering
Cancer Center

28th Molecular Modelling Workshop 2014 - Erlangen - 17 Mar 2014

Memorial Sloan-Kettering
Cancer Center

Memorial Sloan-Kettering
Cancer Center

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

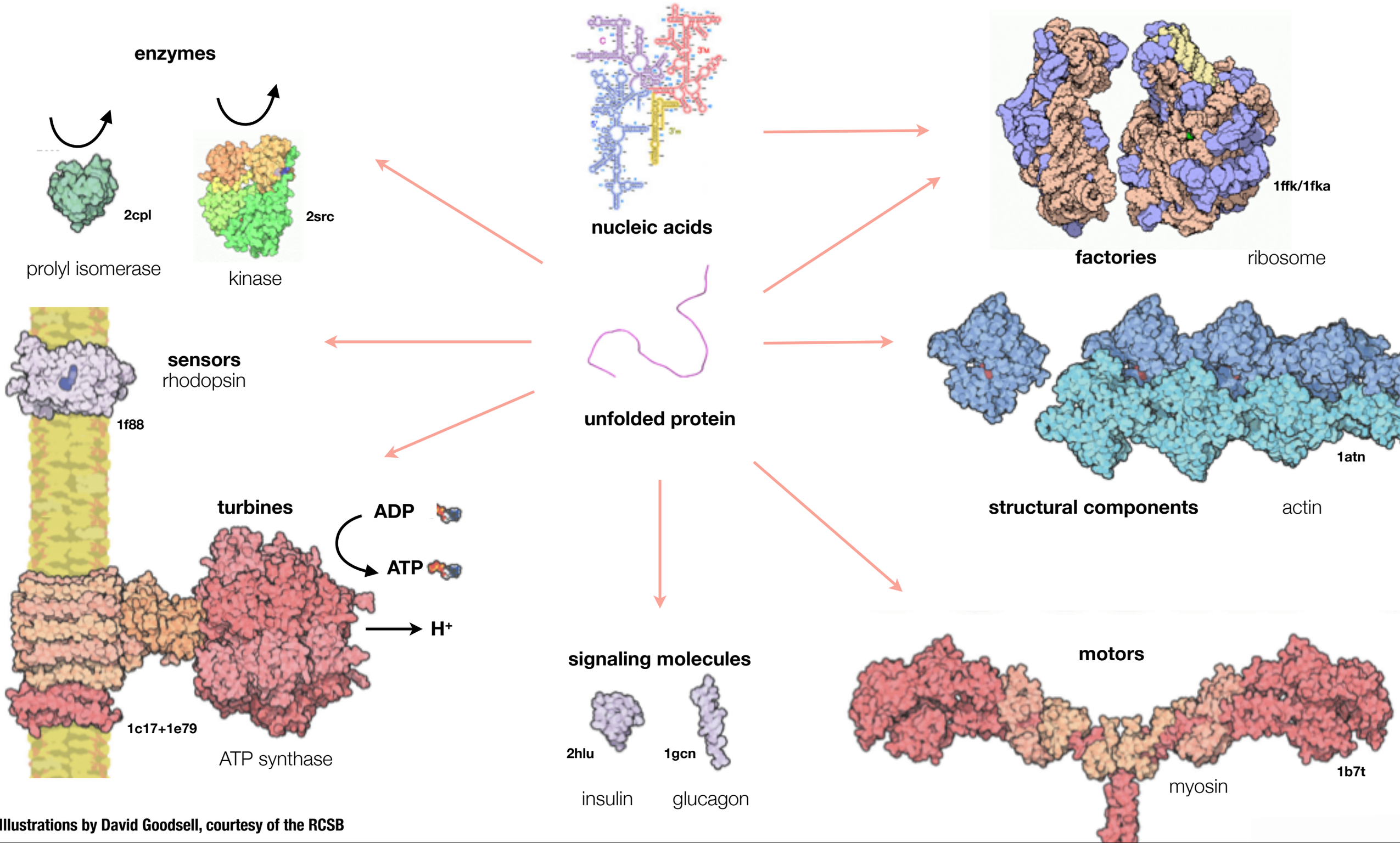
Christina Leslie

Gunnar Rätsch

Joao Xavier

We're hiring!

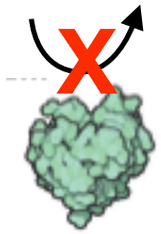
Biological macromolecules are the molecular machines of life



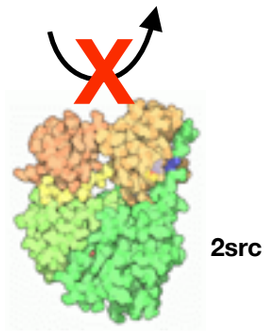
Illustrations by David Goodsell, courtesy of the RCSB

Perturbations can cause disease by disrupting this machinery

defective binding, catalysis, or regulation



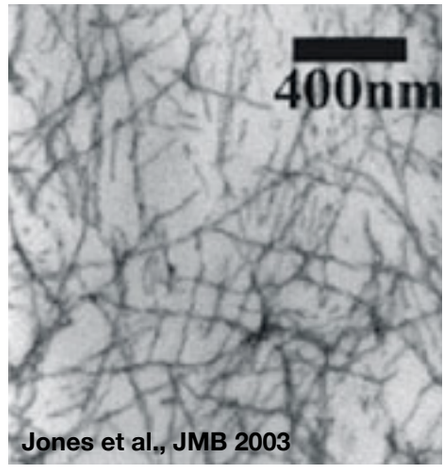
prolyl isomerase



kinase

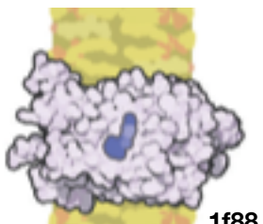
misfolding and aggregation

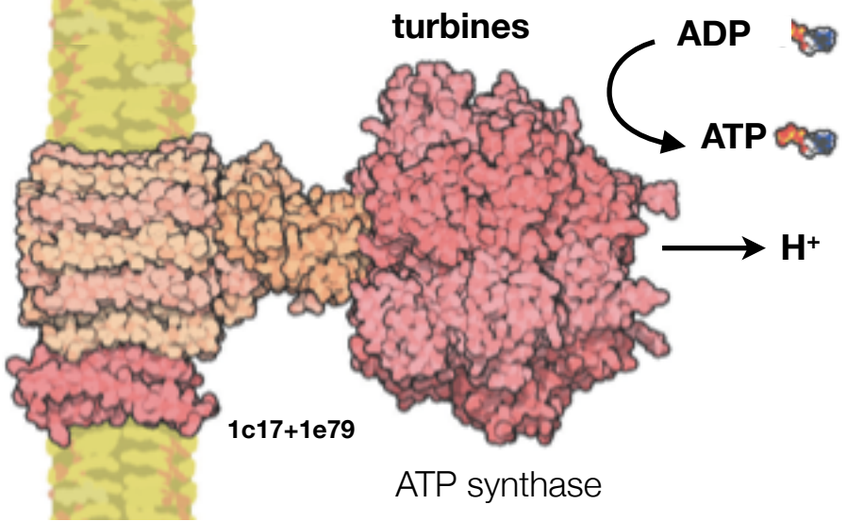
amyloid fibrils



unstable protein degraded by cell

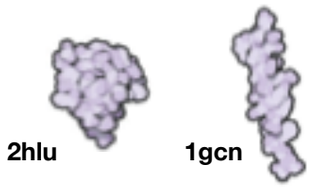
broken sensors, channels, and pumps



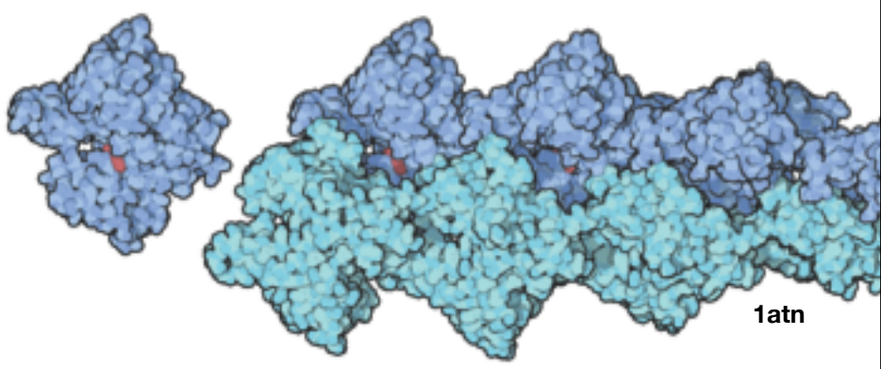


ATP synthase

reduced efficacy or production of signaling molecules

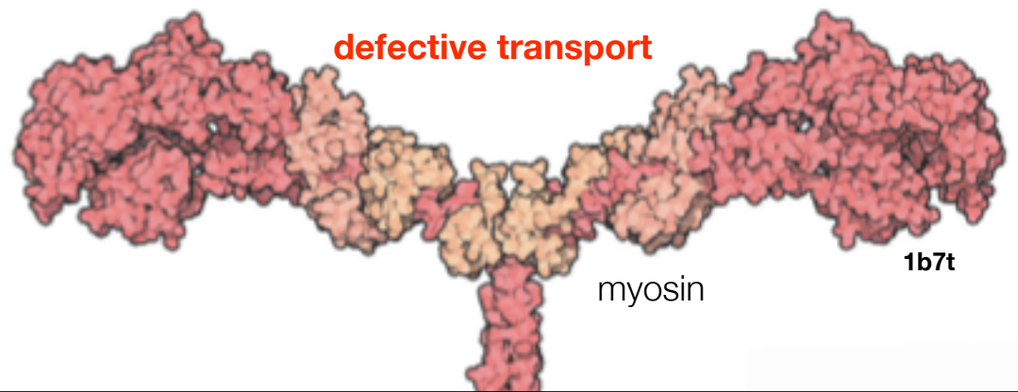


insulin glucagon



deficient structural properties actin

defective transport

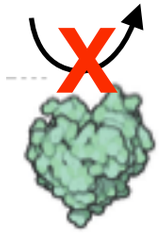


myosin

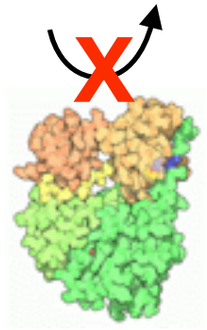
Illustrations by David Goodsell, courtesy of the RCSB

Perturbations can cause disease by disrupting this machinery

defective binding, catalysis, or regulation



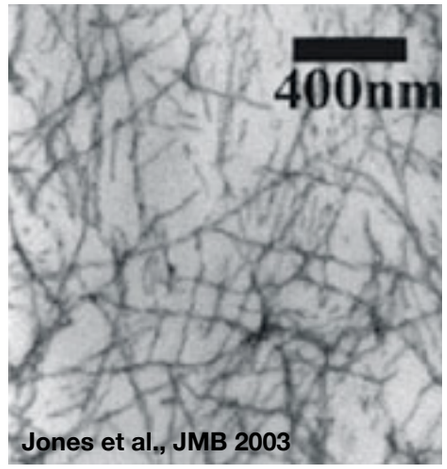
prolyl isomerase



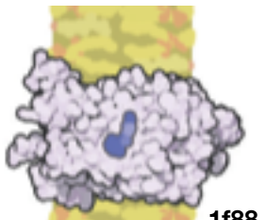
kinase

misfolding and aggregation

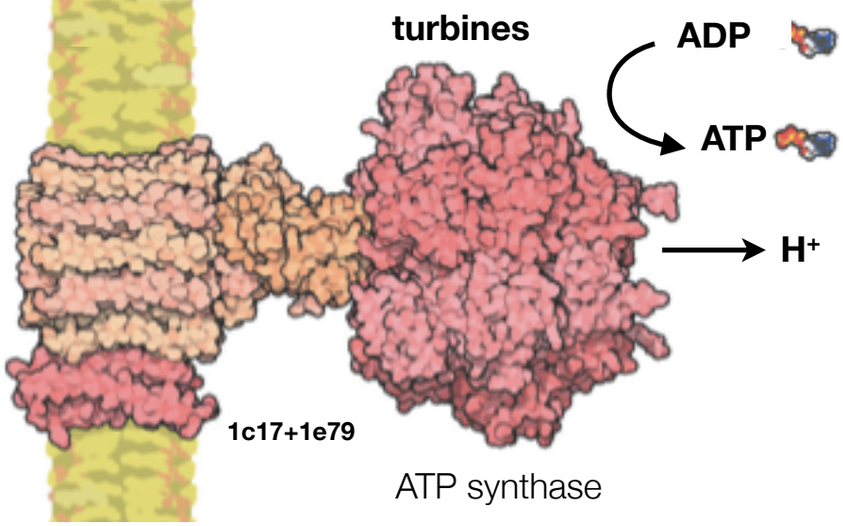
amyloid fibrils



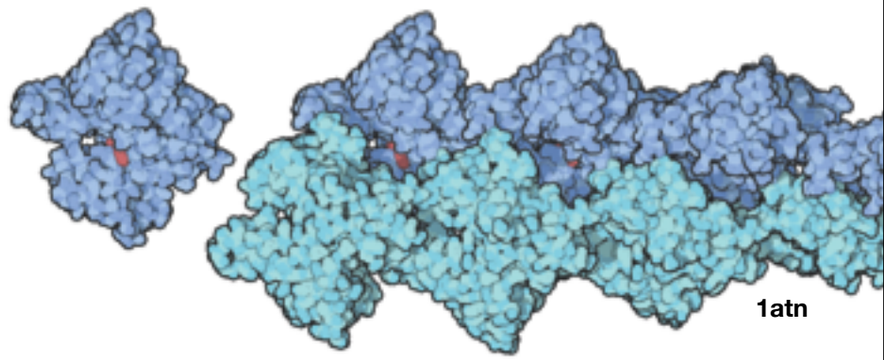
broken sensors, channels, and pumps



turbines

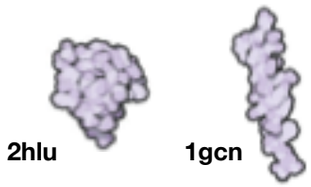


ATP synthase



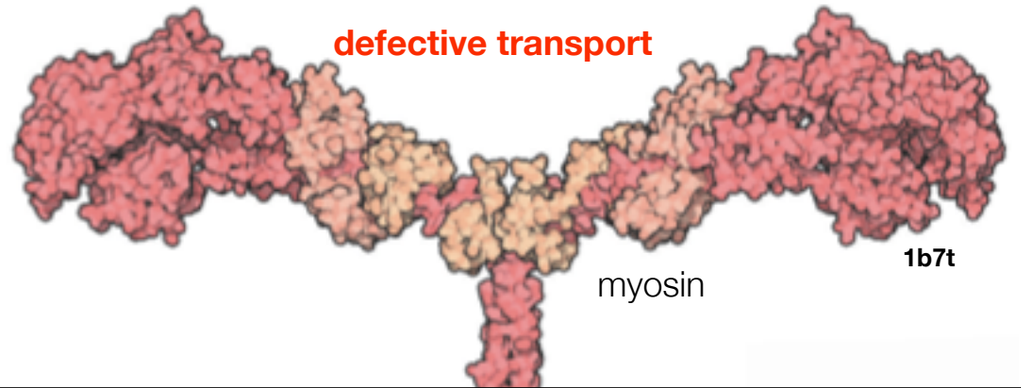
deficient structural properties actin

reduced efficacy or production of signaling molecules



insulin glucagon

defective transport



myosin

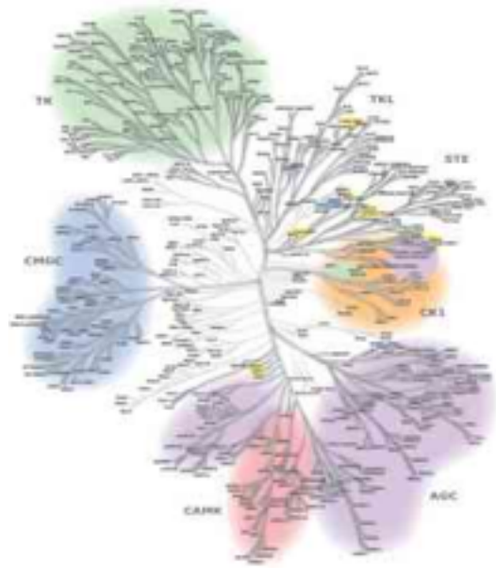
Illustrations by David Goodsell, courtesy of the RCSB

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.

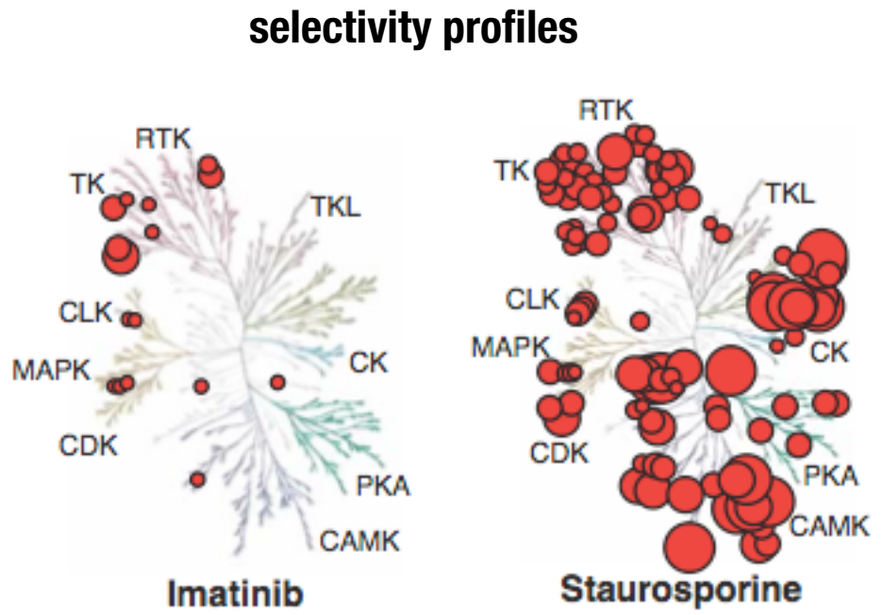
Selective inhibition by imatinib (Gleevec) has been very successful in treating CML.

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



human kinome
(518 kinases)

selectivity profiles



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

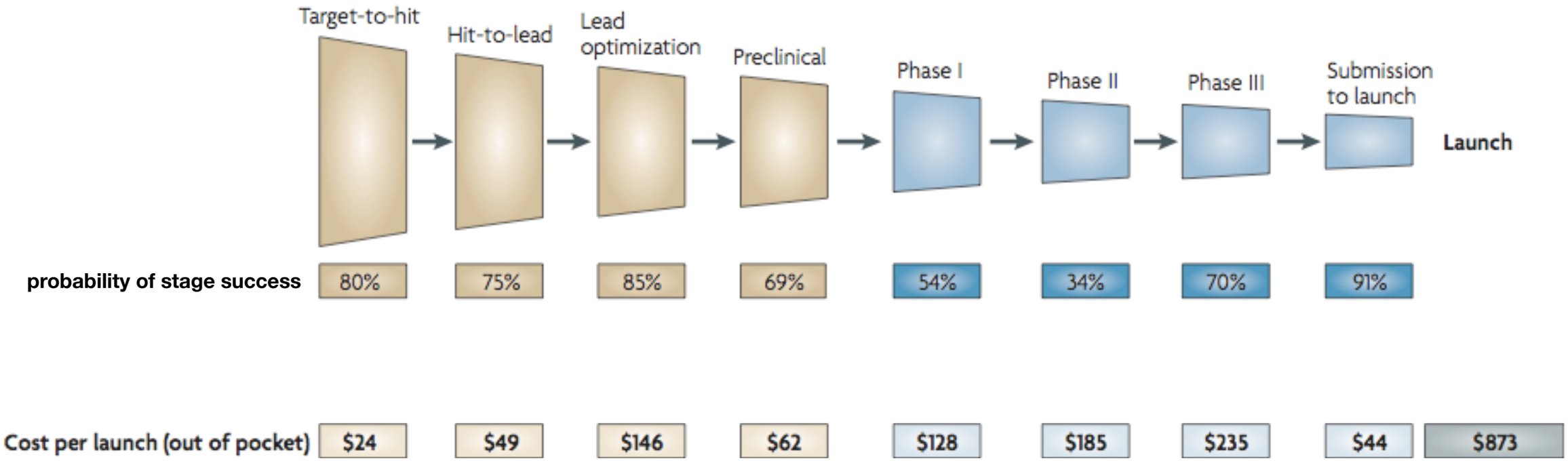
Nature Biotech 23:329, 2005

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



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

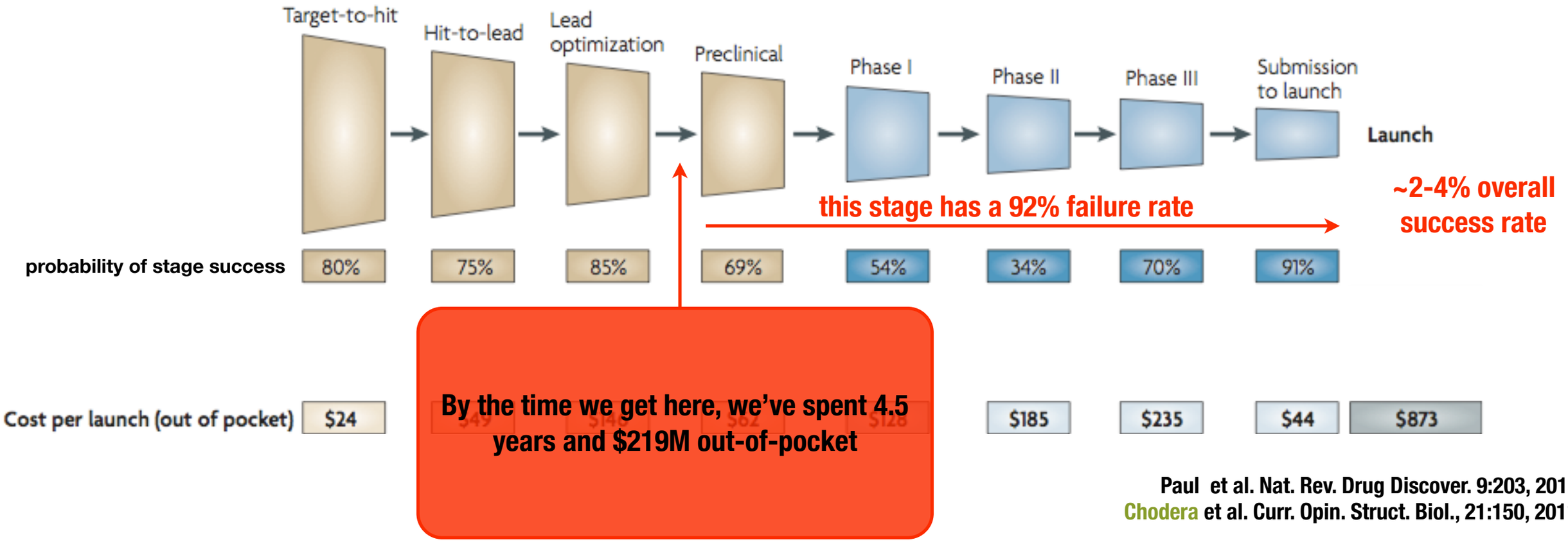
Complex design objectives (efficacy, selectivity, ADME-Tox) make design problem especially complex.

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

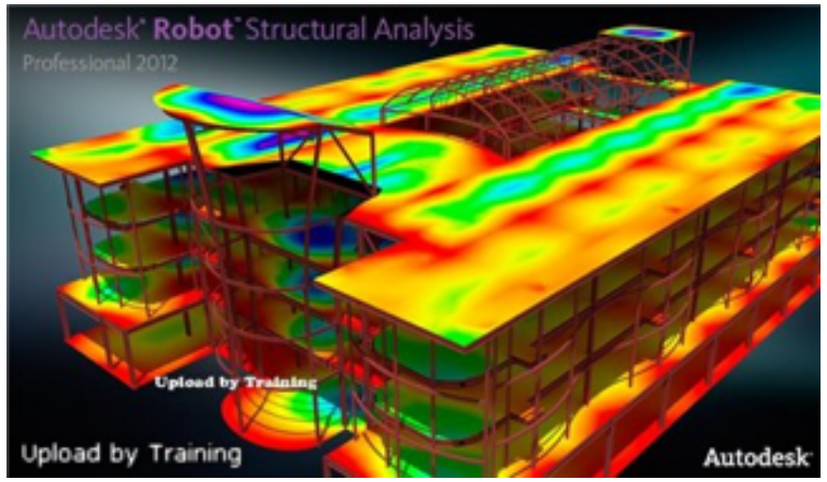
Number of truly innovative new molecular entities has remained constant at 5-6/year



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

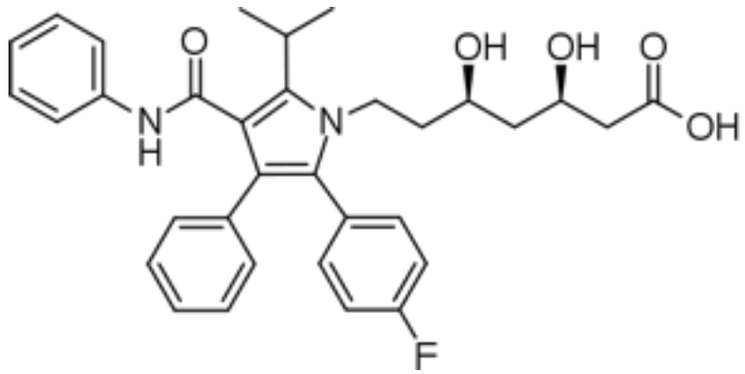
Complex design objectives (efficacy, selectivity, ADME-Tox) make design problem especially complex.

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



10^3 - 10^6 parts

Why not small molecule drugs?

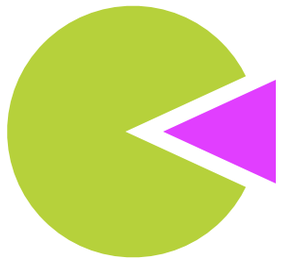


< 10^2 atoms

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

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



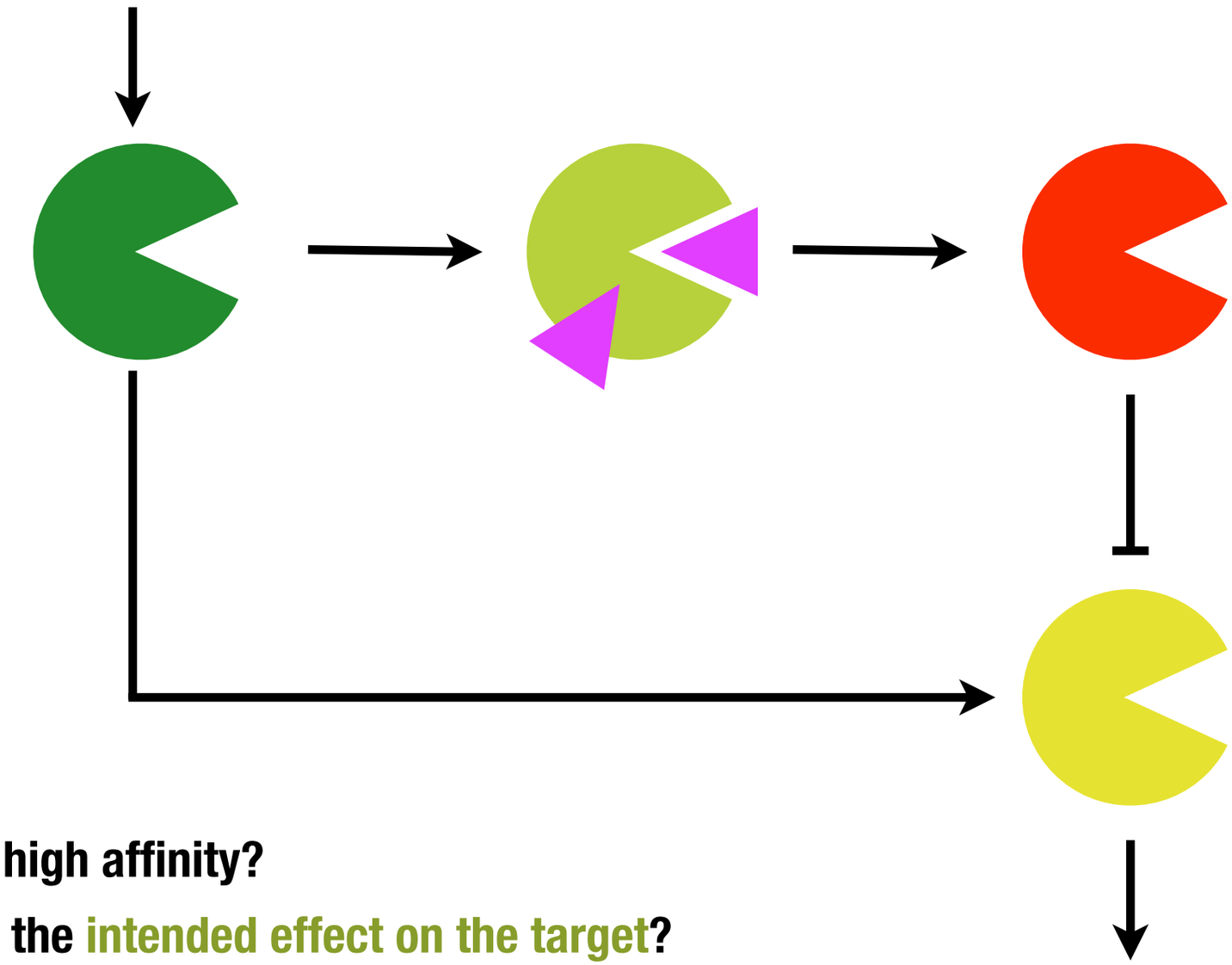
Will it bind the target with high affinity?

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?

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

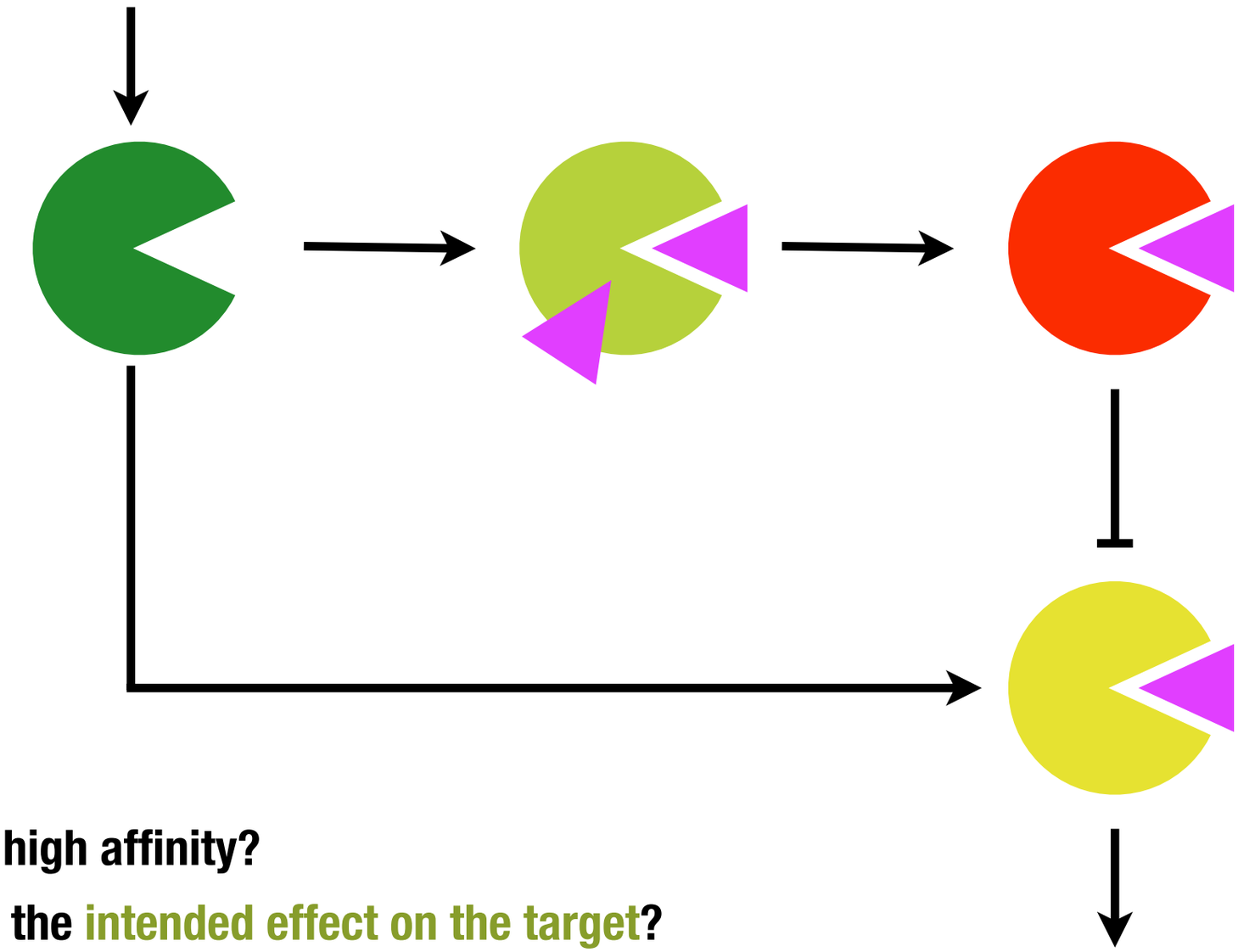


Will it bind the target with high affinity?

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

Does it produce the desired effect on cellular pathways?

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



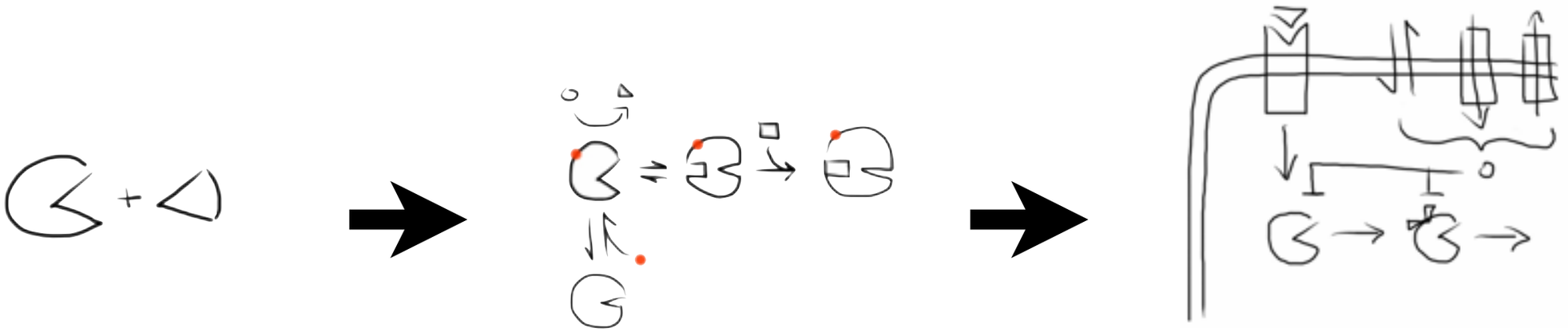
Will it bind the target with high affinity?

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



physical binding constant

K_d

catalytic life cycle

apparent K_i

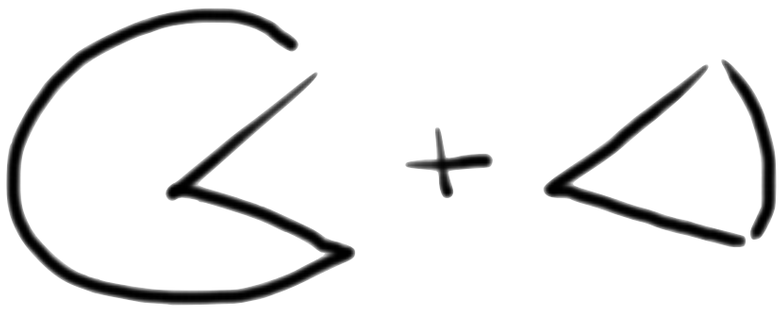
cellular pathways

EC50

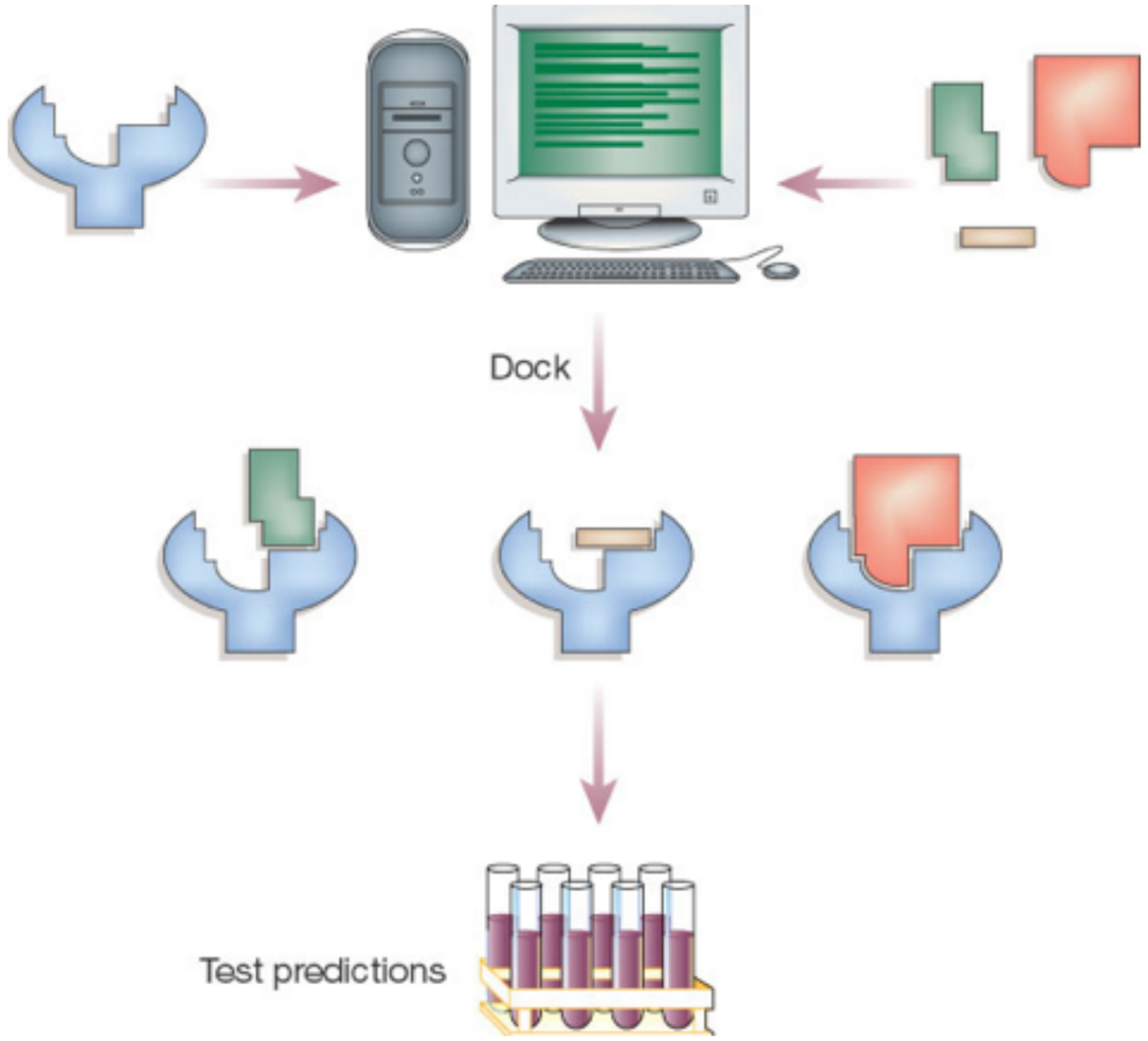
We use physical modeling and statistical mechanics to build predictive models at each of these scales.

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

$$P + L \overset{K_d}{\rightleftharpoons} PL$$



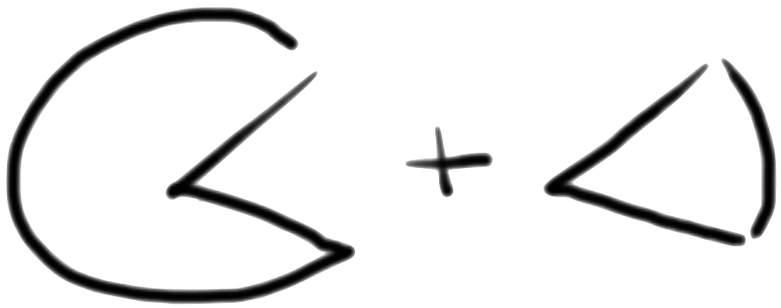
Virtual screening methods are in widespread use in drug discovery efforts today. They must work well, right?



Shoichet BK. Nature 432:862, 2004.

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

$$P + L \overset{K_d}{\rightleftharpoons} PL$$



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

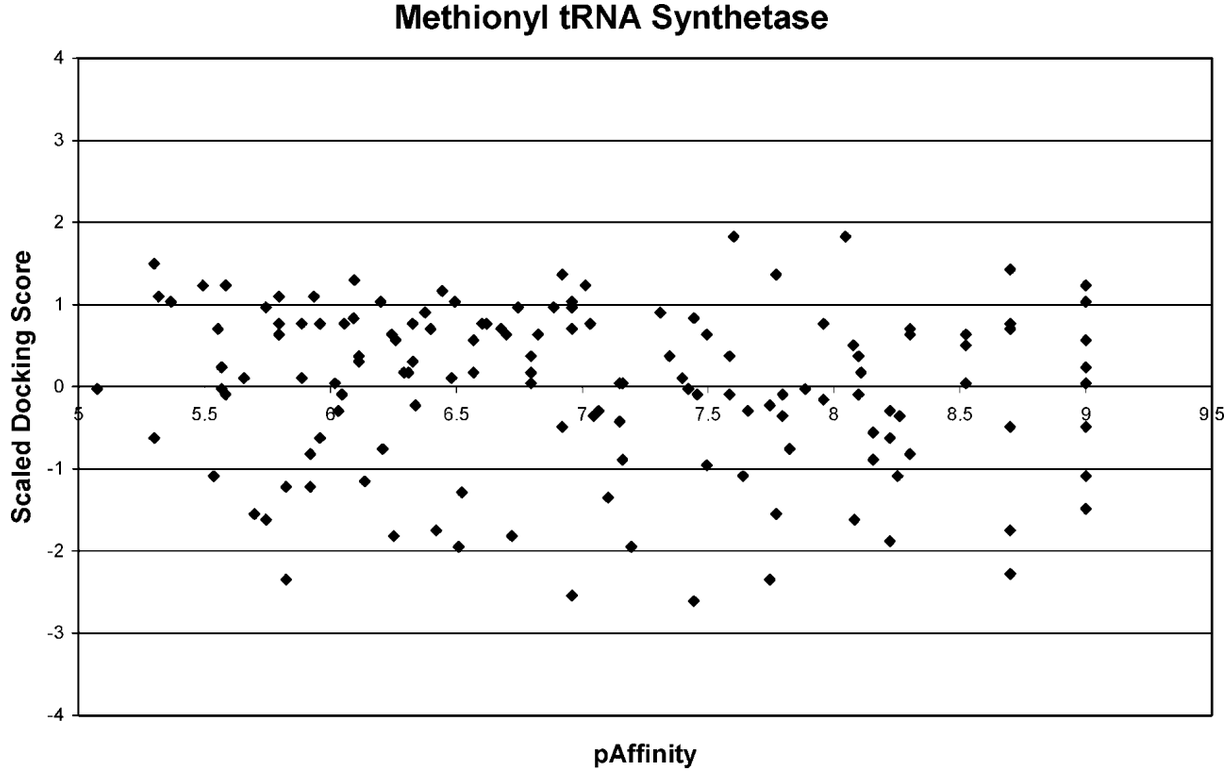
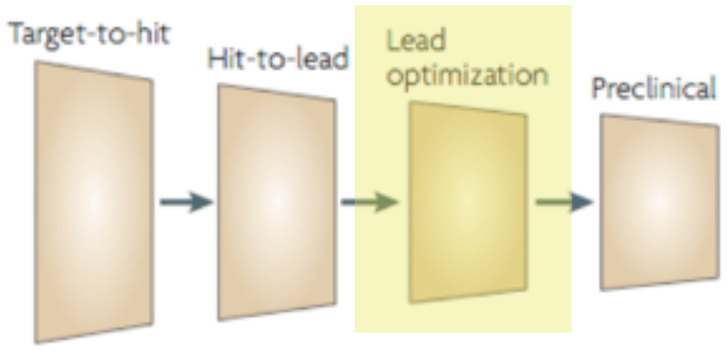


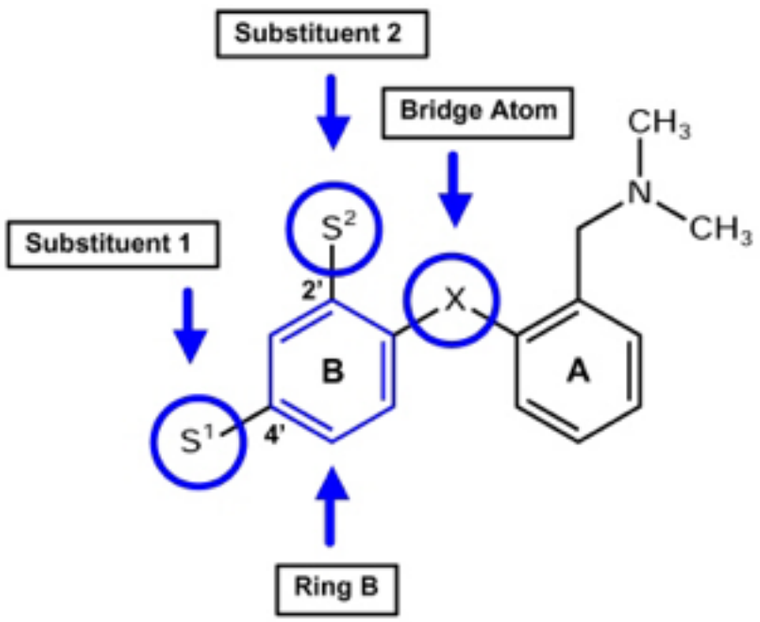
Figure 11. Plot of scaled score vs pAffinity for MRS and PPARδ. While the calculated correlation coefficient for the data shown for MRS is $r = -0.28$, this plot clearly demonstrates that these values are meaningless. No useful correlation exists between the docking score and compound affinity.

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

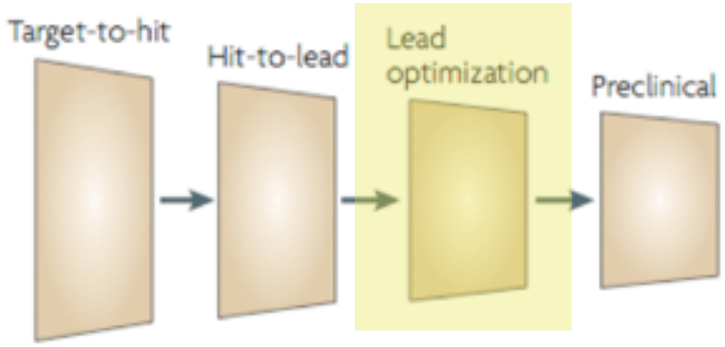
From Warren et al. [GSK], "A Critical Assessment of Docking Programs and Scoring Functions", J. Med. Chem. 49:5912, 2006.

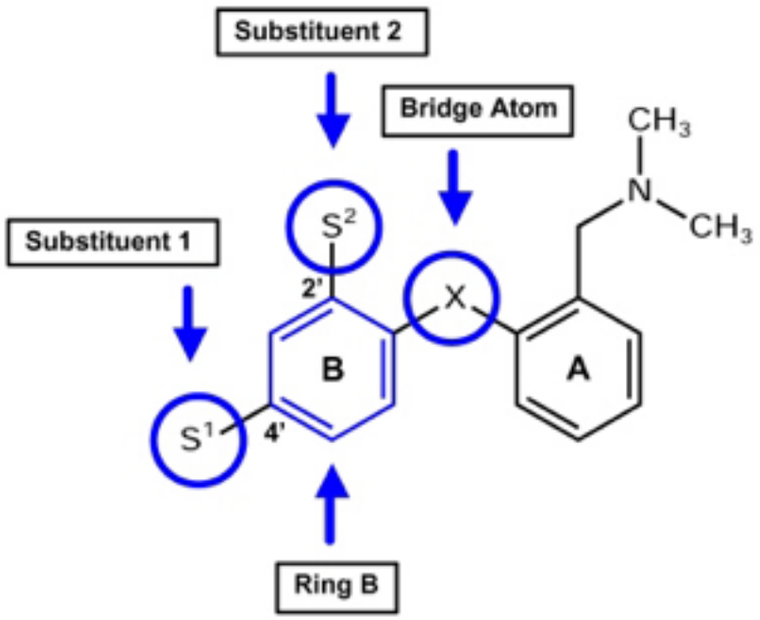
How accurate does a model need to be to aid rational drug design?

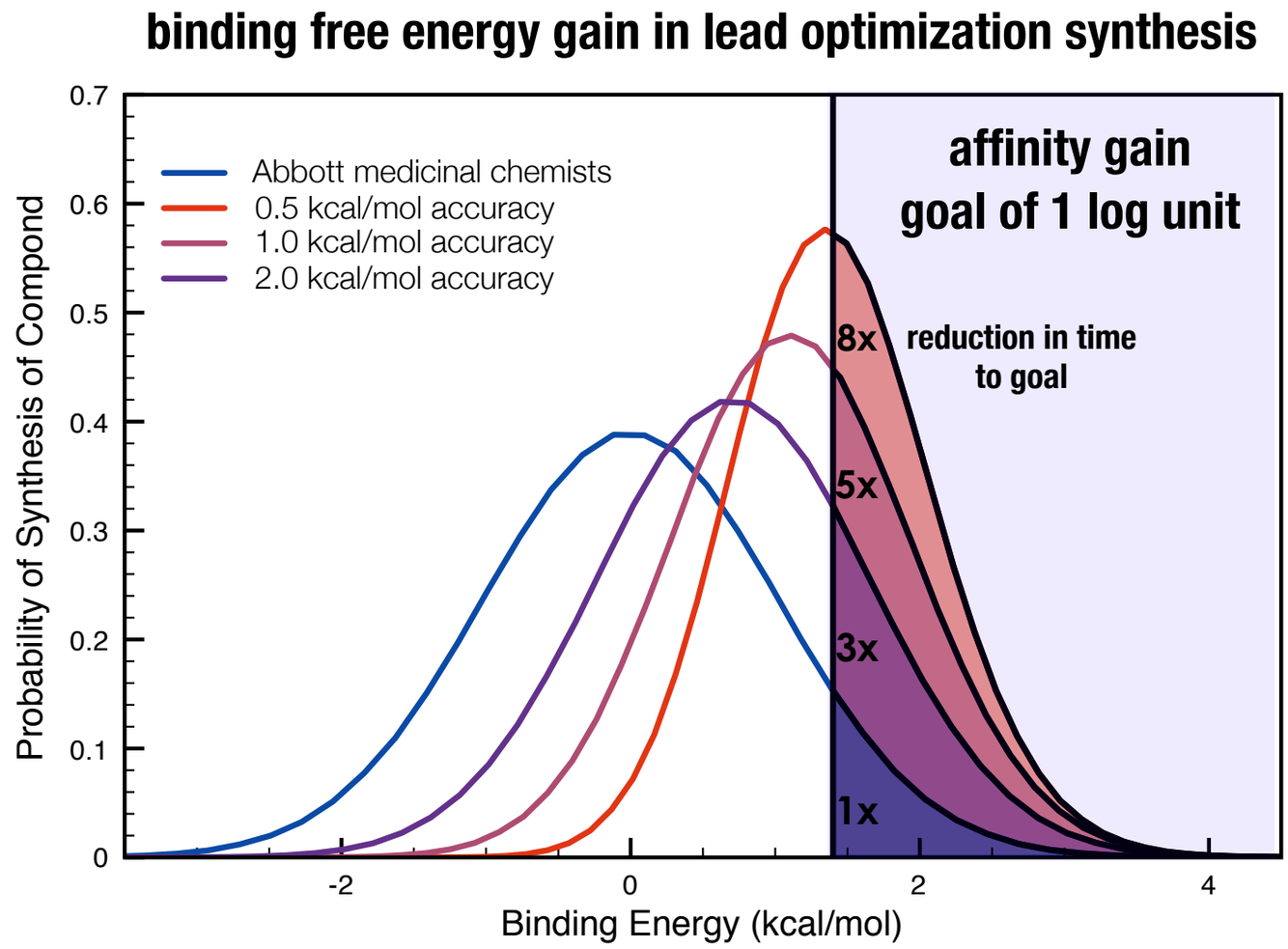




How accurate does a model need to be to aid rational drug design?



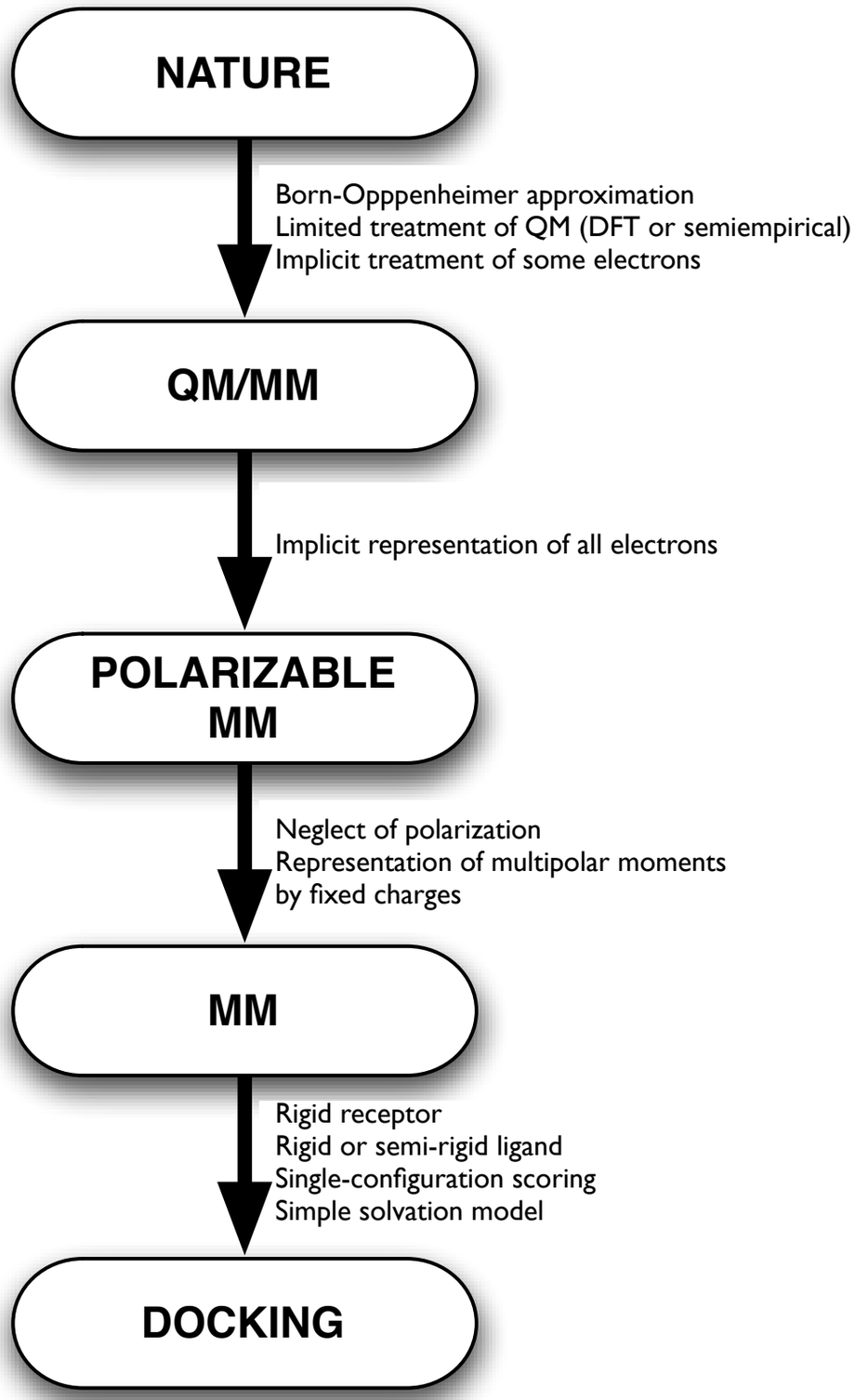




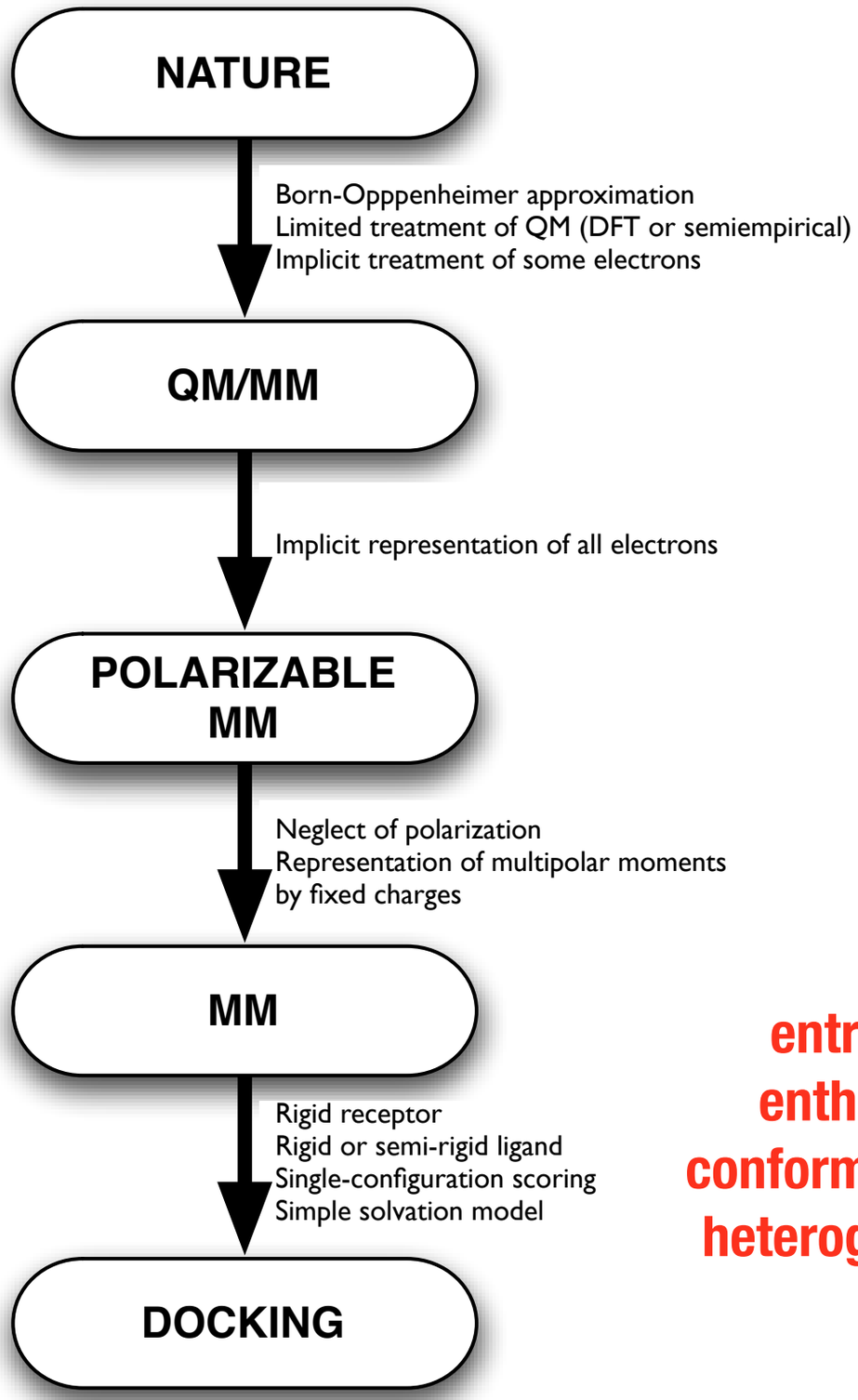
A 2 kcal/mol error in prioritizing lead synthesis would speed lead optimization by 3x but even 10% improvements would be of tremendous benefit

M. R. Shirts, D. L. Mobley and Scott P. Brown. "Free energy calculations in structure-based drug design", in **Drug Design: Structure- and Ligand-Based Approaches**, pgs. 61-86, 2010.

What details are crucial for useful accuracy?
Virtual screening involves a number of approximations

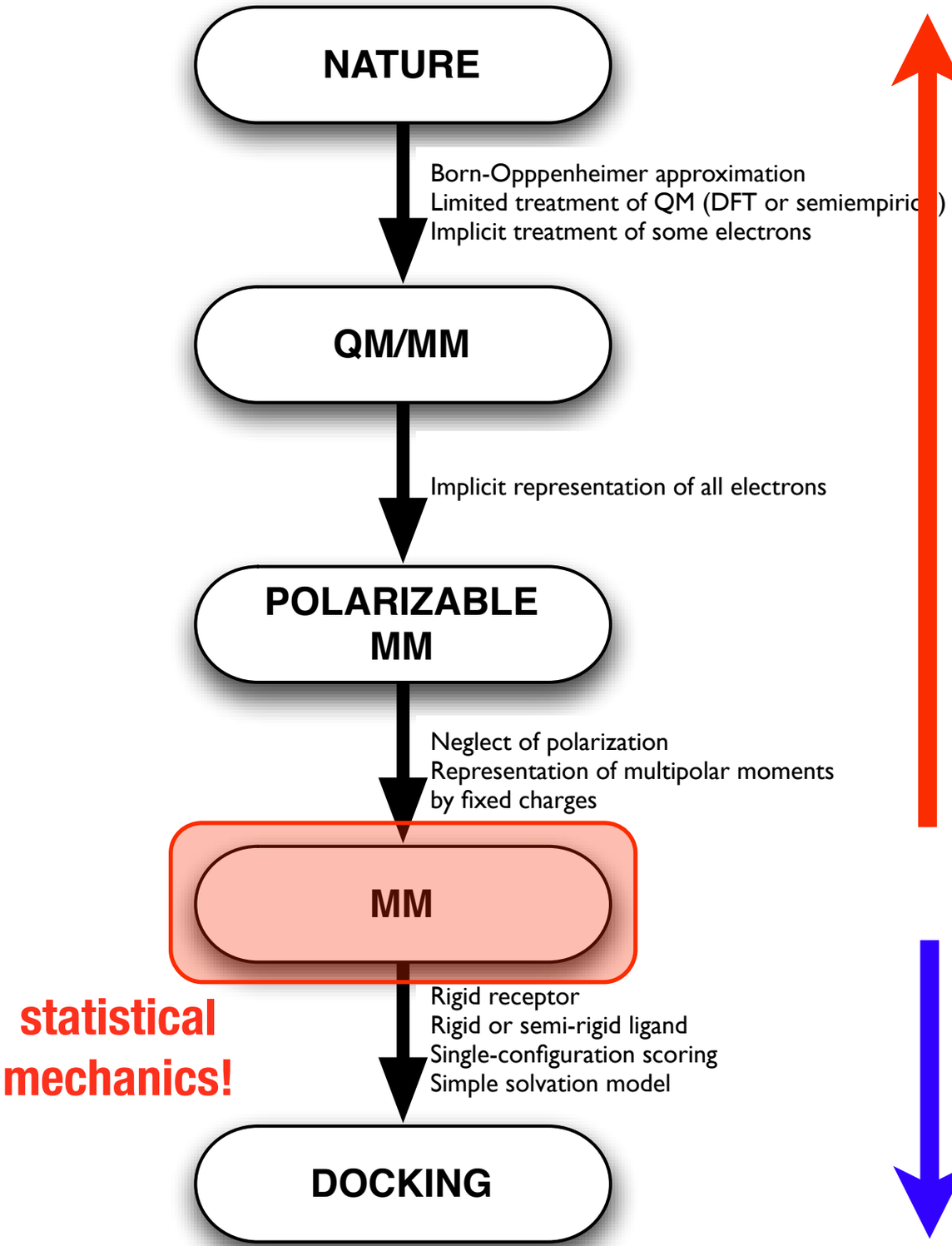


What details are crucial for useful accuracy?
Virtual screening involves a number of approximations



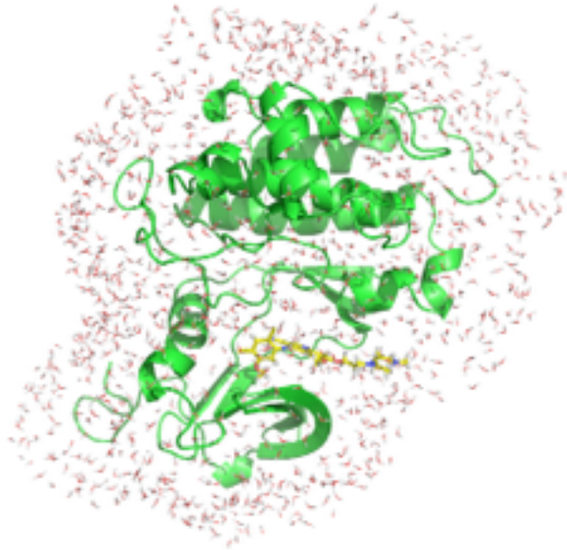
entropy
enthalpy
conformational
heterogeneity

What details are crucial for useful accuracy?
Virtual screening involves a number of approximations



statistical mechanics!

if insufficiently accurate, add detail in perturbative manner



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

$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2$$

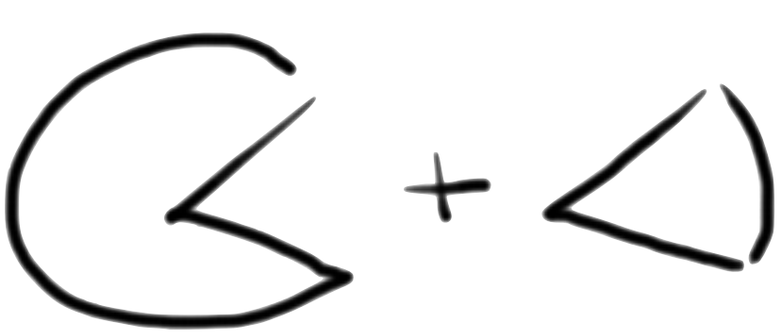
$$+ \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

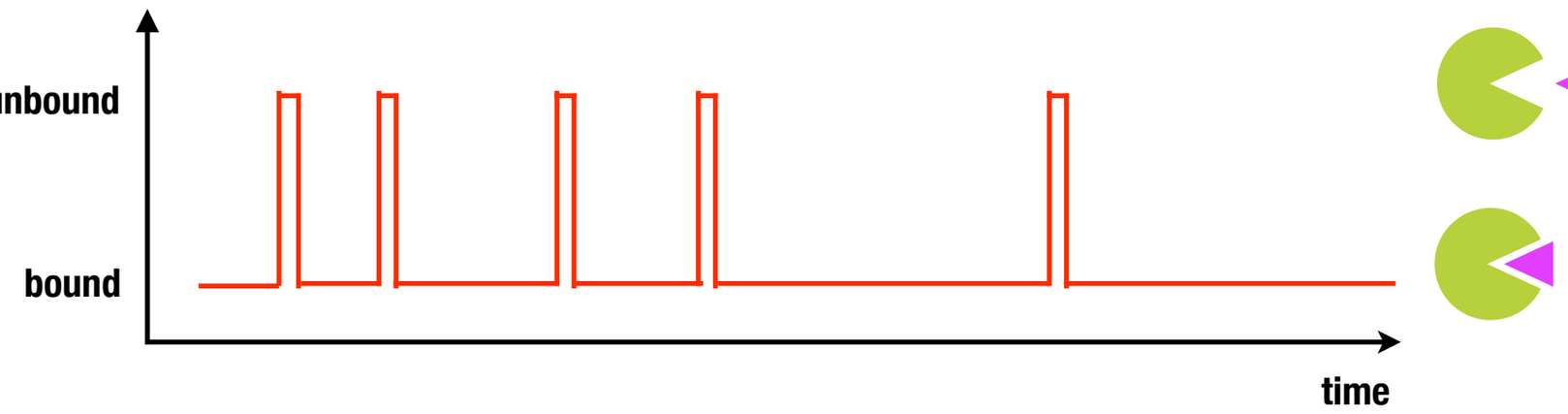
if accurate enough, systematically remove detail

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

$$\mathrm{P} + \mathrm{L} \overset{K_d}{\rightleftharpoons} \mathrm{PL}$$





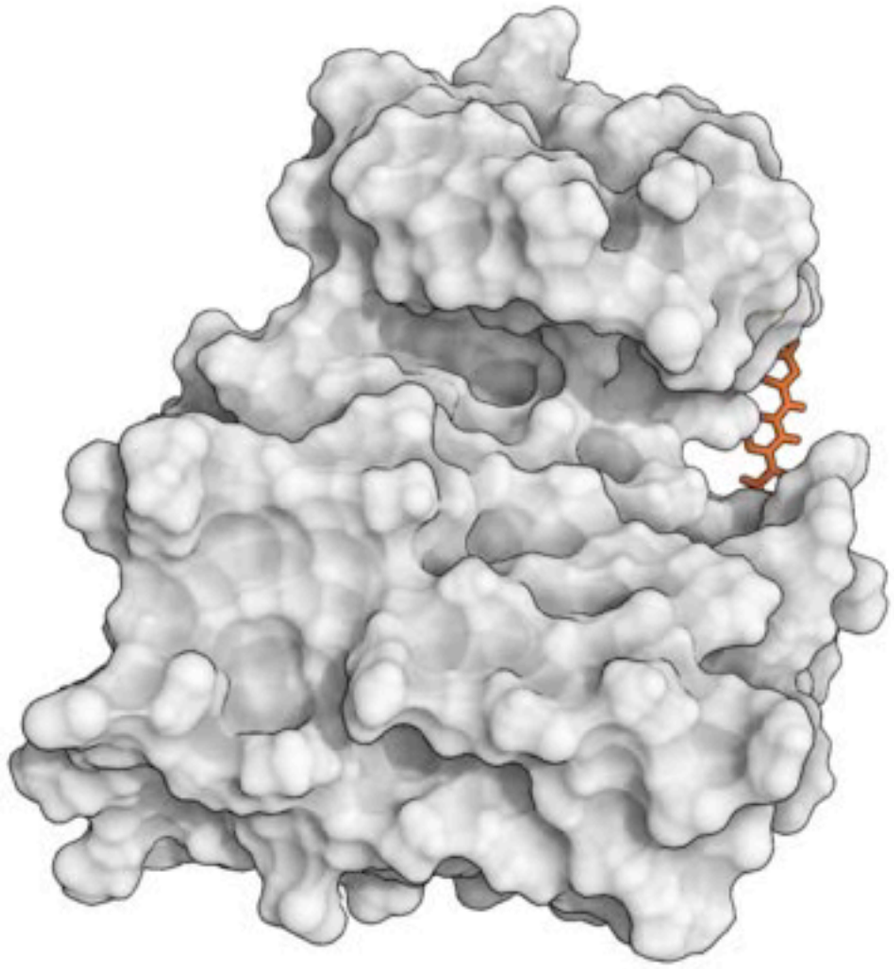
$$K_d \propto \frac{\tau_{\mathrm{unbound}}}{\tau_{\mathrm{bound}}}$$

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

David E. Shaw

ANTON @ DESRES

Src:dasatanib
(4 us simulation)



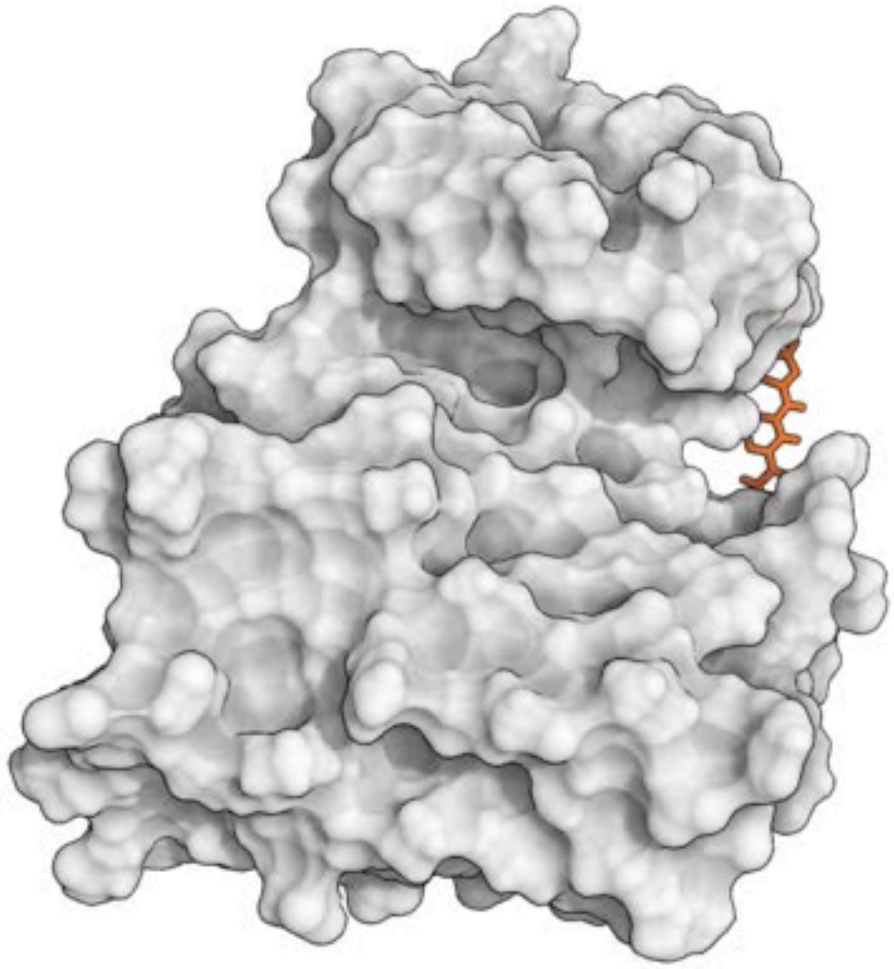
Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. JACS 133:9181, 2011.

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

David E. Shaw

ANTON @ DESRES

Src:dasatanib
(4 us simulation)



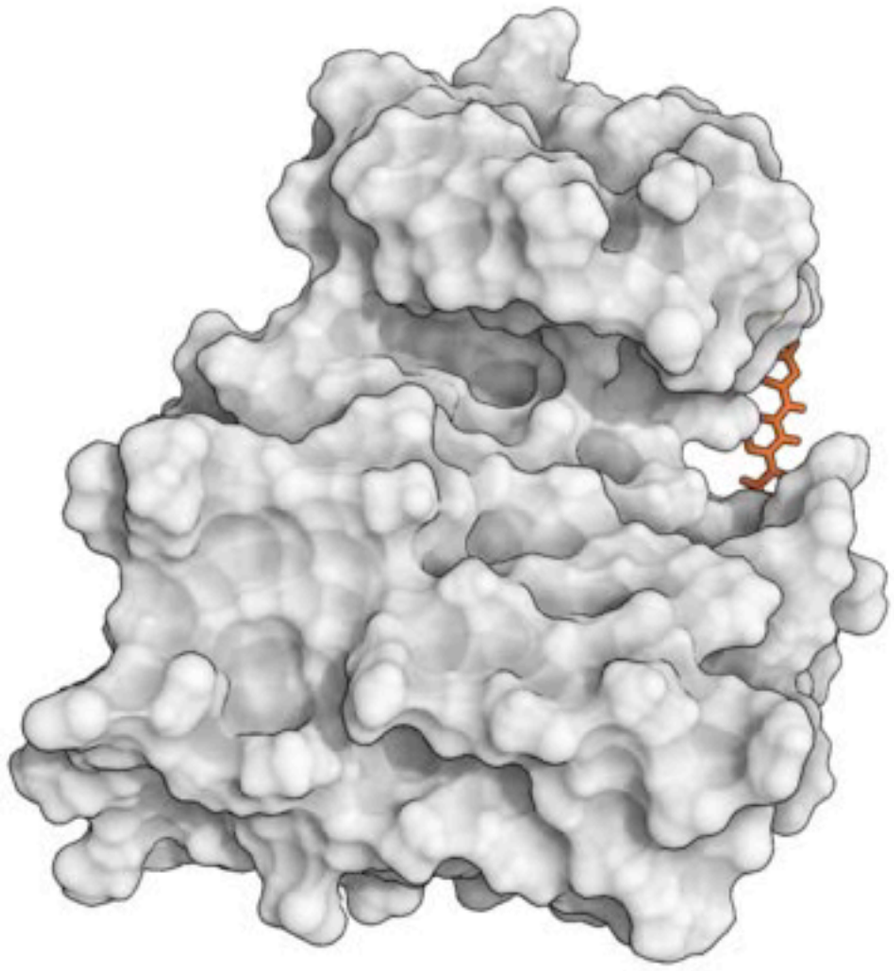
Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. JACS 133:9181, 2011.

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

David E. Shaw

ANTON @ DESRES

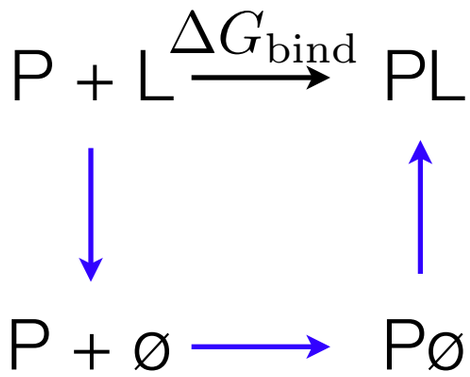
Src:dasatanib
(4 us simulation)



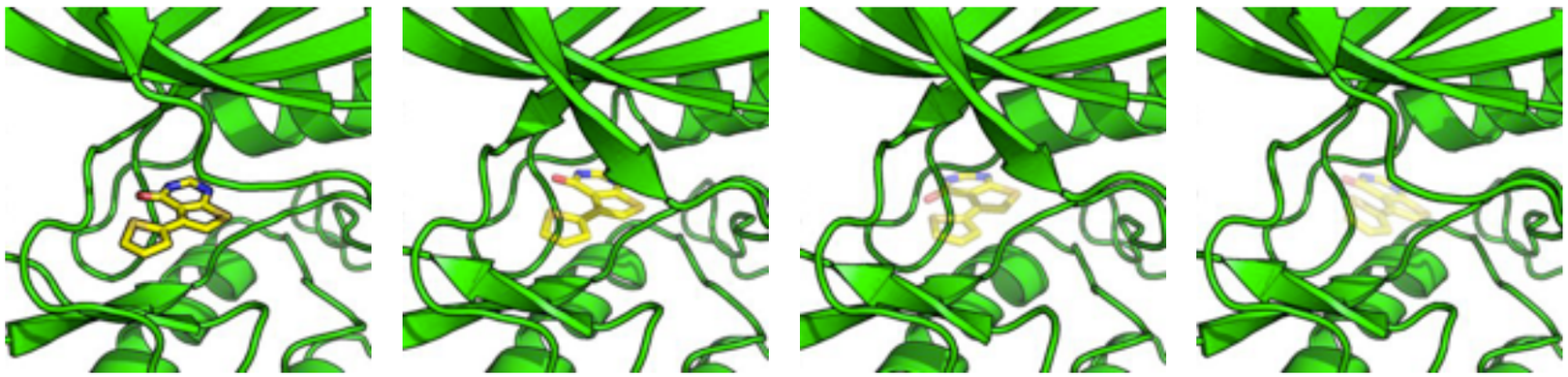
For typical drug off-rates (10^{-4} s^{-1}), trajectories would need to be impractically long (hours) to reliably estimate affinities, requiring ~10^6 years to simulate.

Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. JACS 133:9181, 2011.

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



multiple simulations of "alchemical intermediates" are run



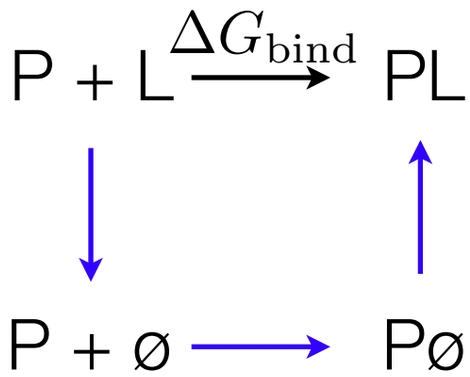
imposition of restraints **discharging** **steric decoupling** **noninteracting**

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

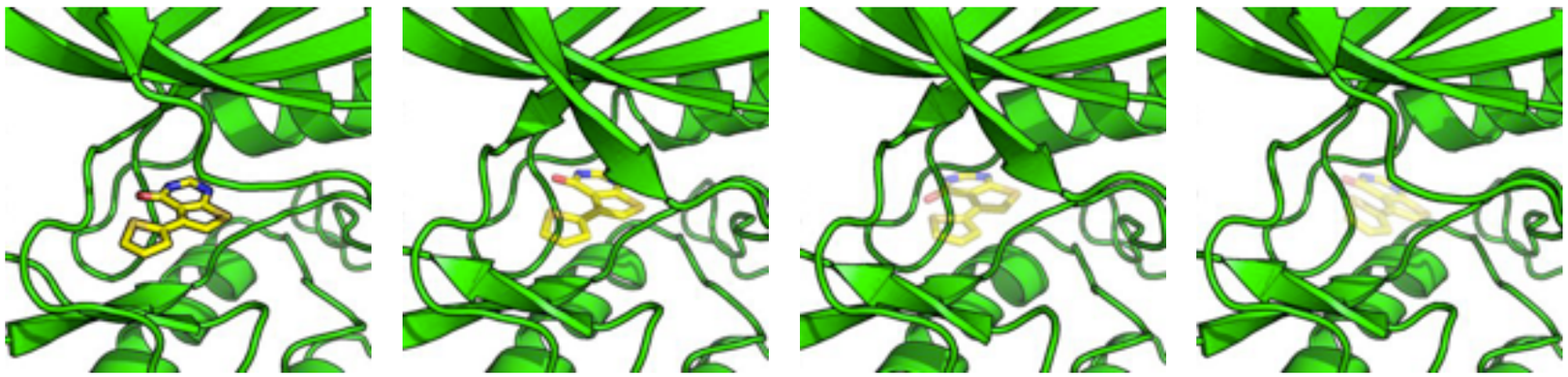
$$\Delta F_{1\to N} = -\beta^{-1} \ln \frac{Z_N}{Z_1} = -\beta^{-1} \ln \frac{Z_2}{Z_1} \cdot \frac{Z_3}{Z_2} \cdots \frac{Z_N}{Z_{N-1}} = \sum_{n=1}^{N-1} \Delta F_{n\to n+1} \qquad Z_n = \int d\mathbf{x}\, e^{-\beta U(\mathbf{x})}$$

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

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



multiple simulations of "alchemical intermediates" are run



imposition of restraints **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.

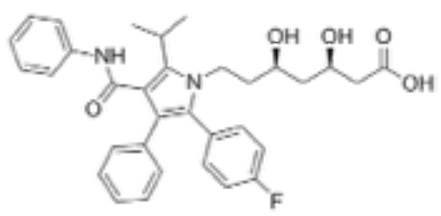
$$\Delta F_{1\to N} = -\beta^{-1} \ln \frac{Z_N}{Z_1} = -\beta^{-1} \ln \frac{Z_2}{Z_1} \cdot \frac{Z_3}{Z_2} \cdots \frac{Z_N}{Z_{N-1}} = \sum_{n=1}^{N-1} \Delta F_{n\to n+1} \qquad Z_n = \int d\mathbf{x}\, e^{-\beta U(\mathbf{x})}$$

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

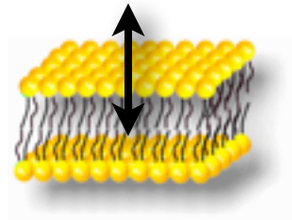
Alchemical free energy calculations can in principle also compute other relevant physical properties

---

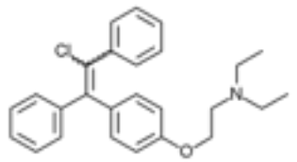
partition coefficients (logP, logD) and permeabilities



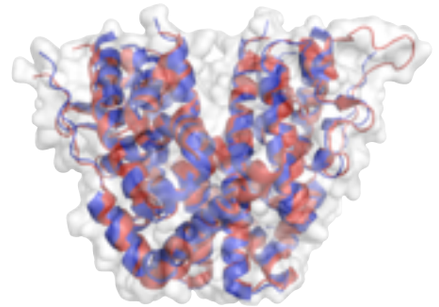
lipitor



selectivity for subtypes or related targets/off-targets



clomifene



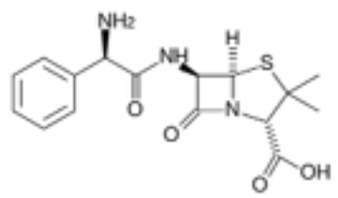
ERα/β

lead optimization of affinity and selectivity

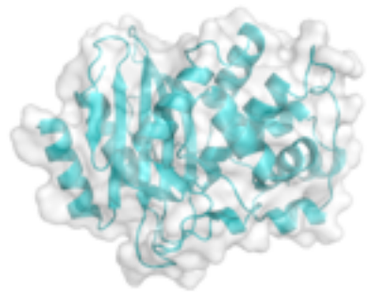
Imatinib

Dasatinib

susceptibility to resistance mutations



ampicillin



β-lactamase

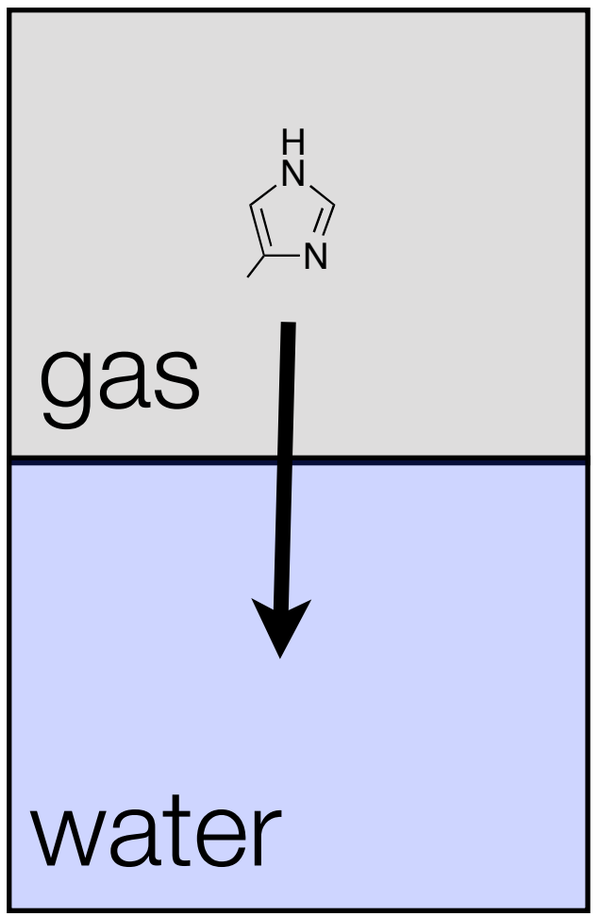
also: solubilities, polymorphs, etc.

with David L. Mobley

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

with David L. Mobley

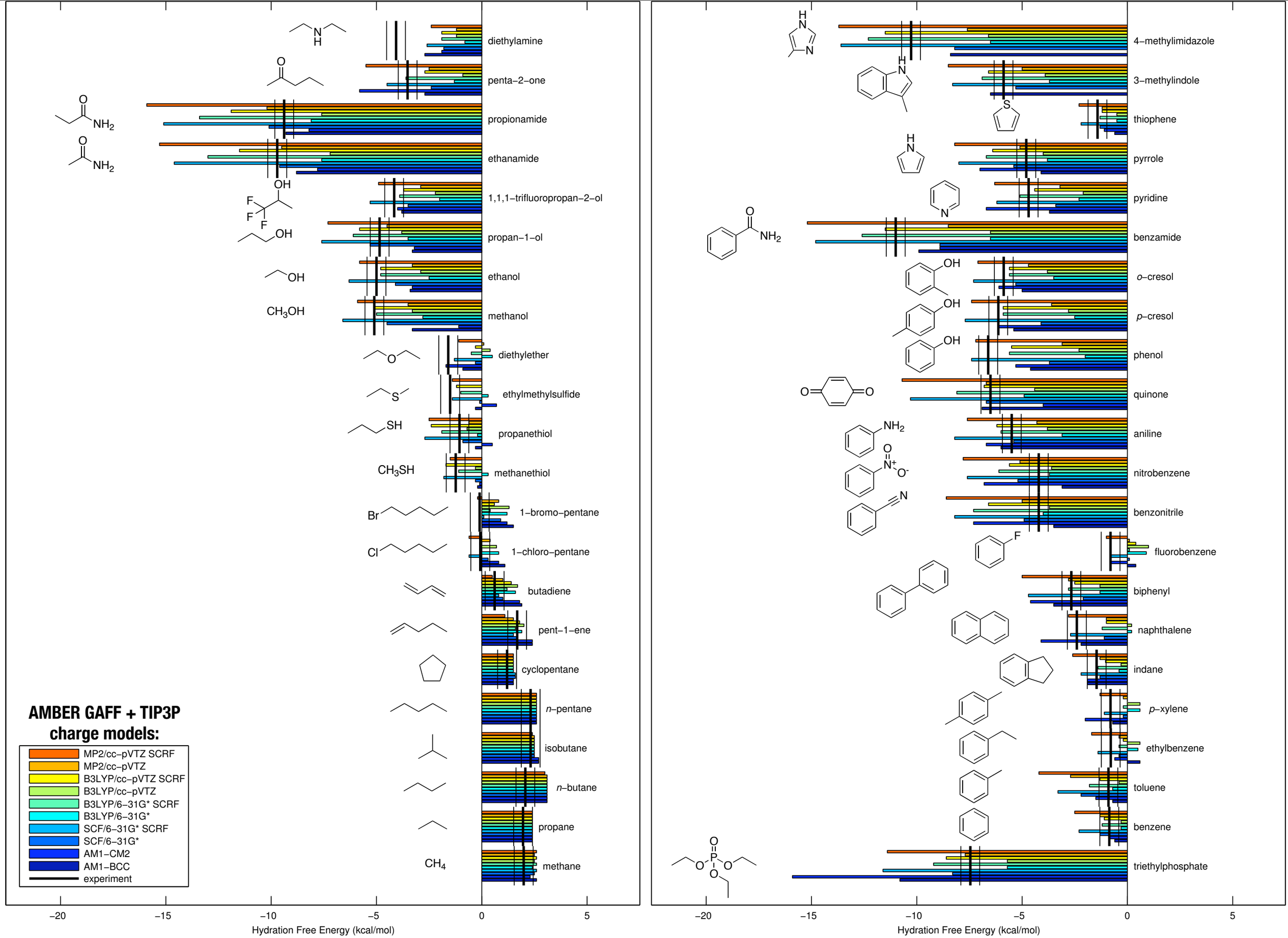
A warmup: hydration free energies



Why hydration free energies?

A test of accuracy of half of binding reaction (withdrawal from water)
A tractable system for studying precision and potential issues

D. L. Mobley, E. Dumont, J. D. Chodera and K. A. Dill. *J. Phys. Chem. B.* **111**:2242-2254 (2007) (DOI)



Thick vertical line shows experimental measurement from Abraham et al., while thin likesshow 95% confidence interval for agrement with eperiment, factoring in both experimental uncertainties (0.2 kcal/mol) and computed uncertaties (less than 0.1 kcal/mol).

D. L. Mobley, E. Dumont, J. D. Chodera and K. A. Dill. *J. Phys. Chem. B.* **111**:2242-2254 (2007)

How accurate are computed hydration free energies?

water model	charge model	all compounds	subset by functionalities		
			alkanes (6)	arenes (6)	alcohols (6)
implicit[25]	AM1BCC	1.35	0.26	0.29	1.41
	AMSOL	2.46	0.31	1.34	2.08
	RESP (SCF/6-31G*)	1.28	0.49	0.23	2.00
explicit TIP3P	AM1BCC	0.92	0.51	0.30	1.43
	AMSOL	1.38	0.41	0.73	1.49
	RESP (SCF/6-31G*)	0.82	0.47	1.05	0.84
	RESP (B3LYP/cc-pVTZ+SCRF)	0.78	0.47	0.60	0.46

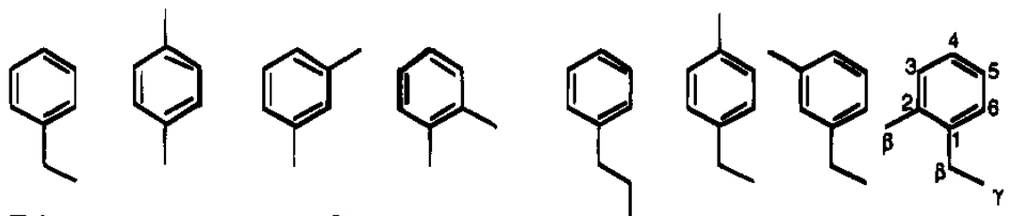
Accuracy of about 1 kcal/mol for best charge models;
High-level QM works well, but only if you also include SCRF for solution-phase dipoles

D. L. Mobley, E. Dumont, J. D. Chodera and K. A. Dill. *J. Phys. Chem. B.* **111**:2242-2254 (2007) (DOI)

with David L. Mobley

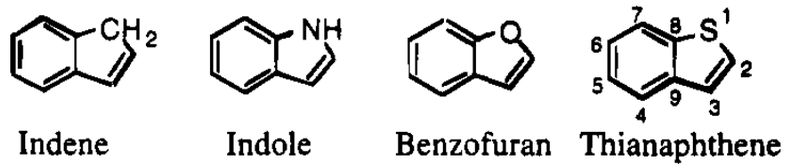
How accurate are binding free energy calculations?
T4 lysozyme L99A as a simple model binding site

Class I - "Isophobic" Ligands



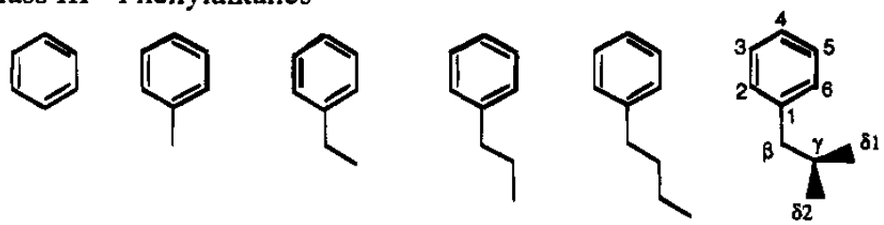
Ethyl benzene | p-, m-, o-Xylene | Propyl benzene | p-, m-, o-Ethyltoluene

Class II - "Isosteric" Ligands

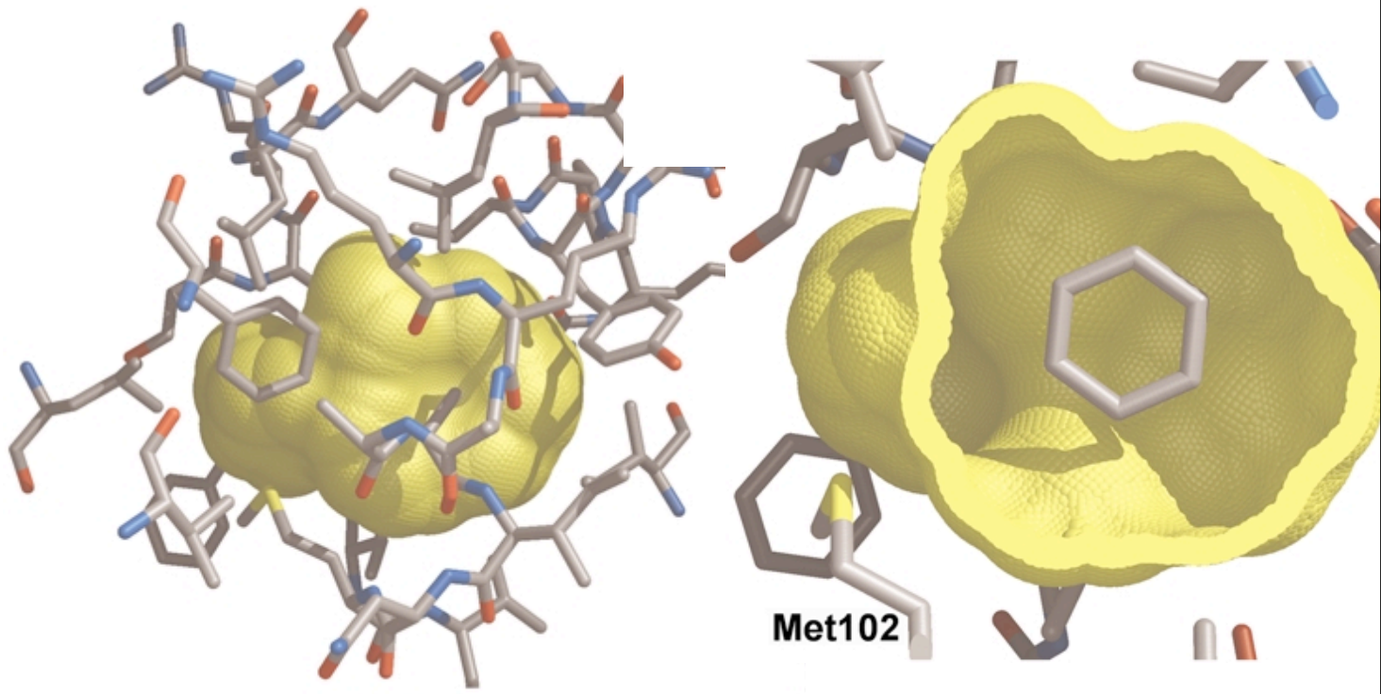


Indene | Indole | Benzofuran | Thianaphthene

Class III - Phenylalkanes



Benzene | Toluene | Ethyl- | Propyl- | Butyl- | iso-Butylbenzene



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

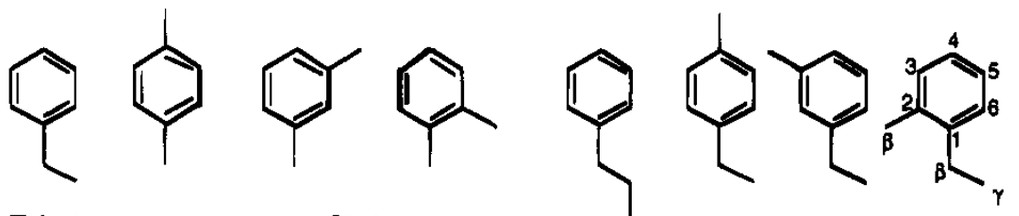
Surprisingly challenging for docking codes to discriminate binders (uM) from non-binders (>>uM)
Polar version of this cavity (L99A/M102Q) even more challenging for docking.

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

with David L. Mobley

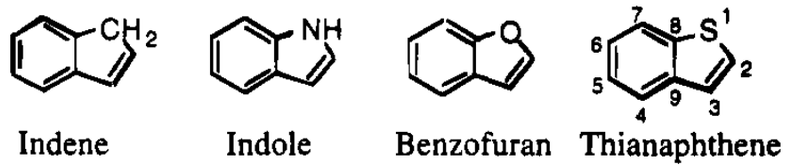
How accurate are binding free energy calculations?
T4 lysozyme L99A as a simple model binding site

Class I - "Isophobic" Ligands



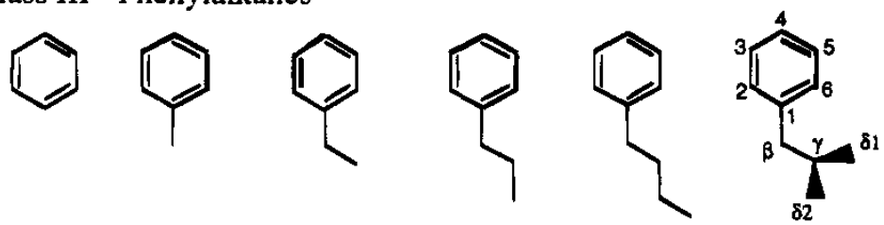
Ethyl benzene | p-, m-, o-Xylene | Propyl benzene | p-, m-, o-Ethyltoluene

Class II - "Isosteric" Ligands

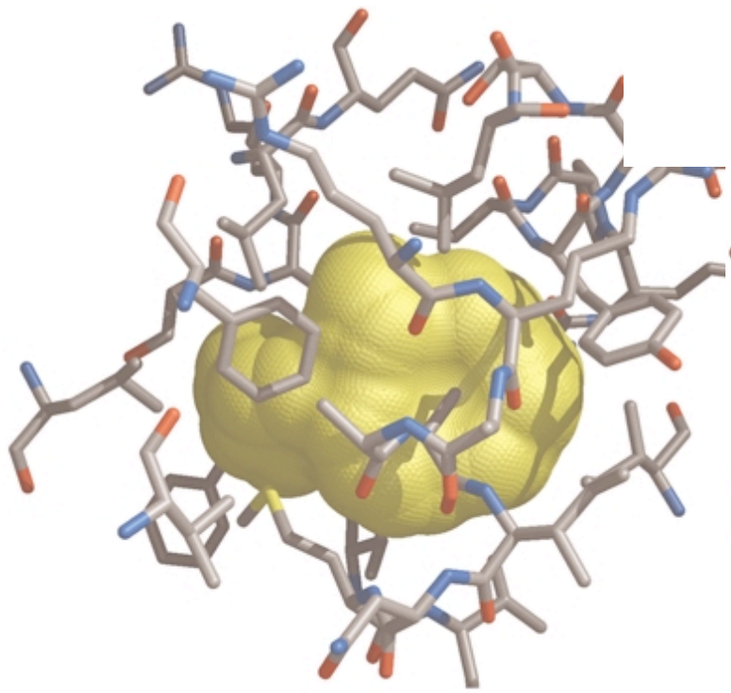


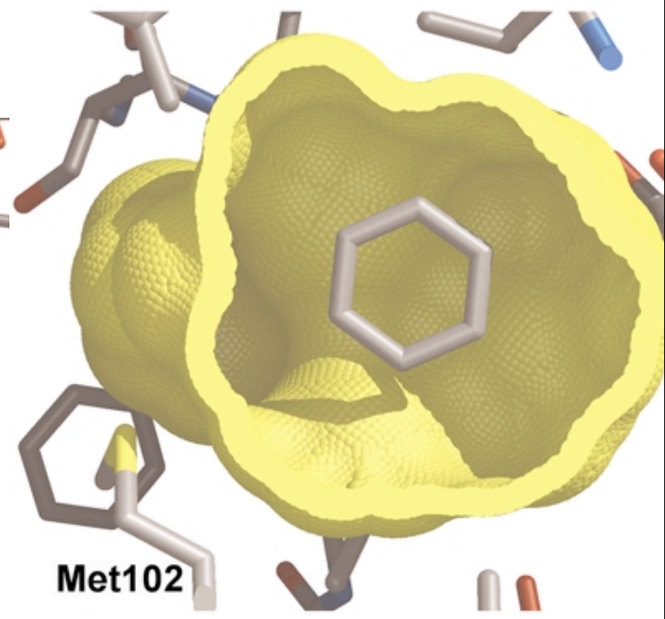
Indene | Indole | Benzofuran | Thianaphthene

Class III - Phenylalkanes



Benzene | Toluene | Ethyl- | Propyl- | Butyl- | iso-Butylbenzene





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

Surprisingly challenging for docking codes to discriminate binders (uM) from non-binders (>>uM)
Polar version of this cavity (L99A/M102Q) even more challenging for docking.

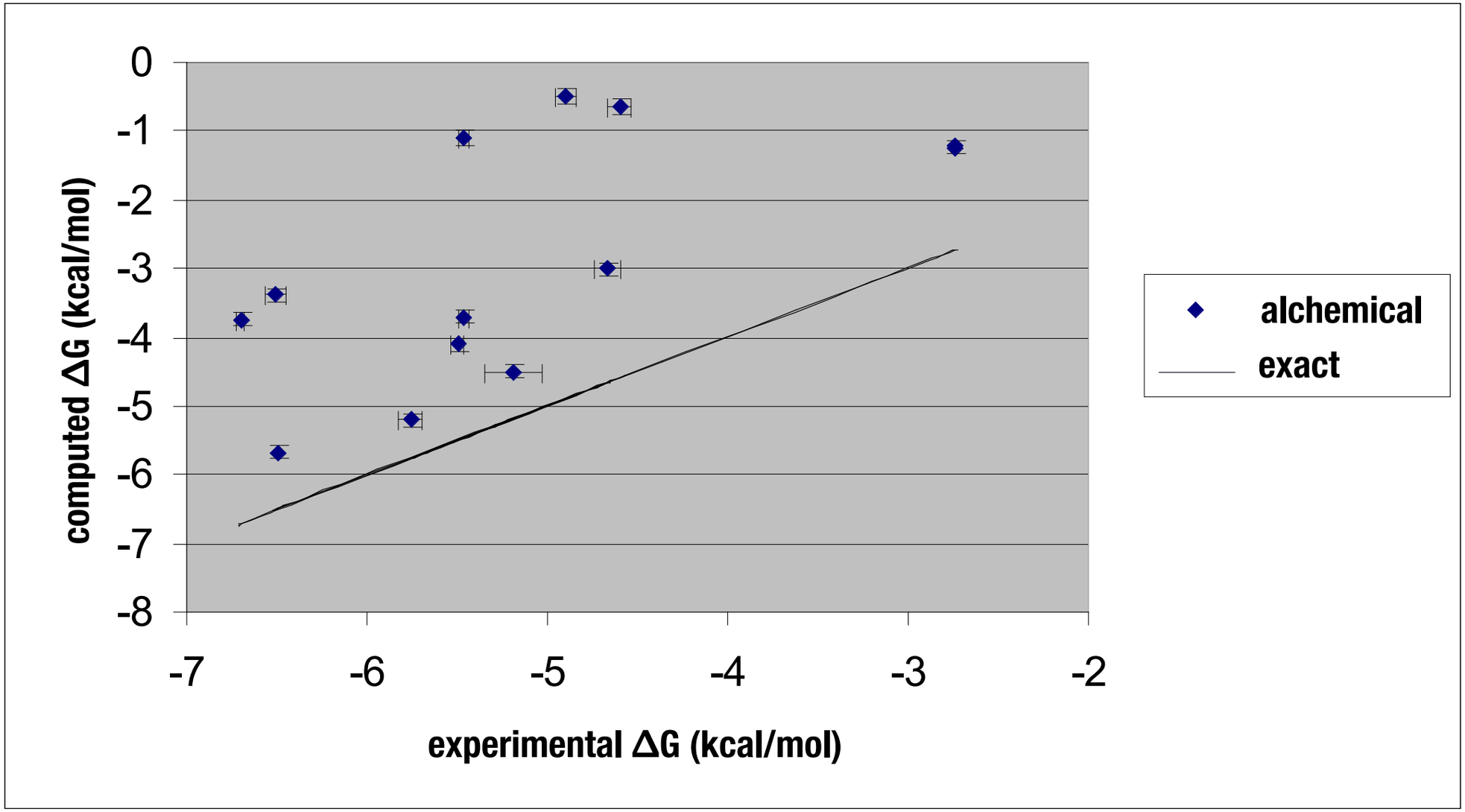
Let's test how well free energy calculations can reproduce measured binding free energies.

"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

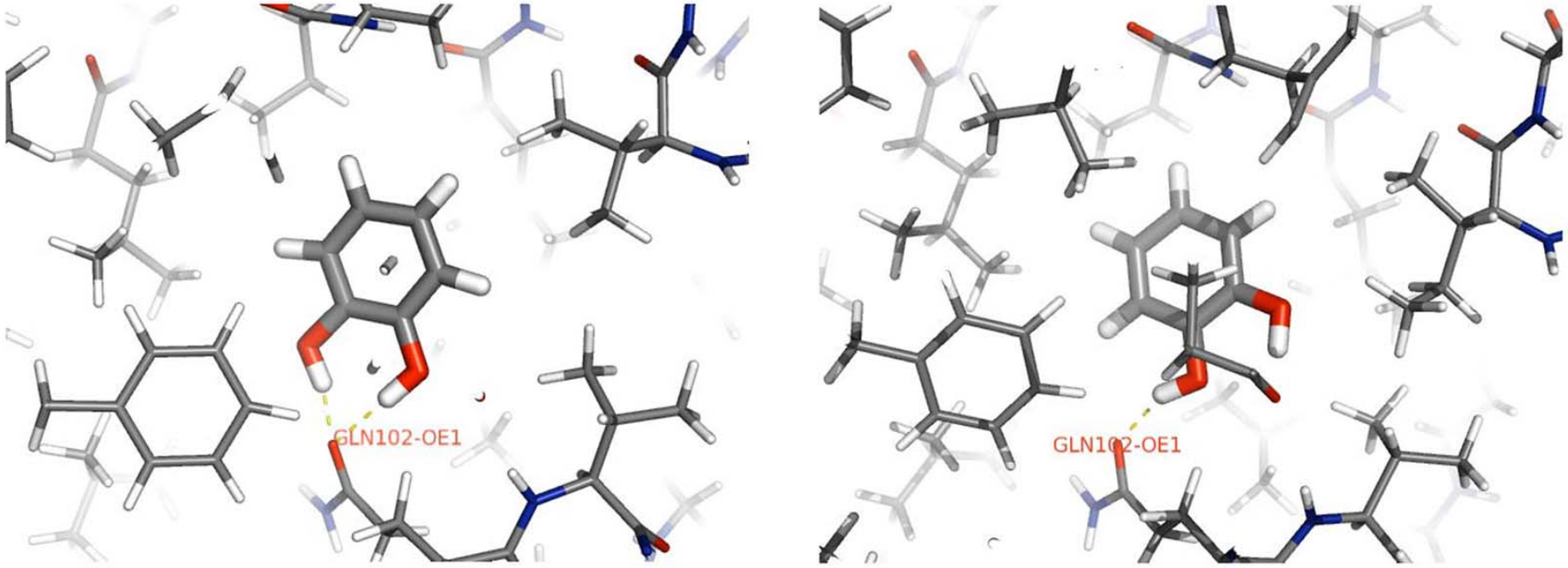


Obviously, we’re missing something.
What’s going on here?

Mobley, Graves, Chodera, McReynolds, Shoichet, Dill. J Mol Biol 371:1118, 2007

Multiple ligand conformations can contribute

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



L99A/M102Q polar version of cavity

Difference in affinity between different bound orientations is only ~1 kT

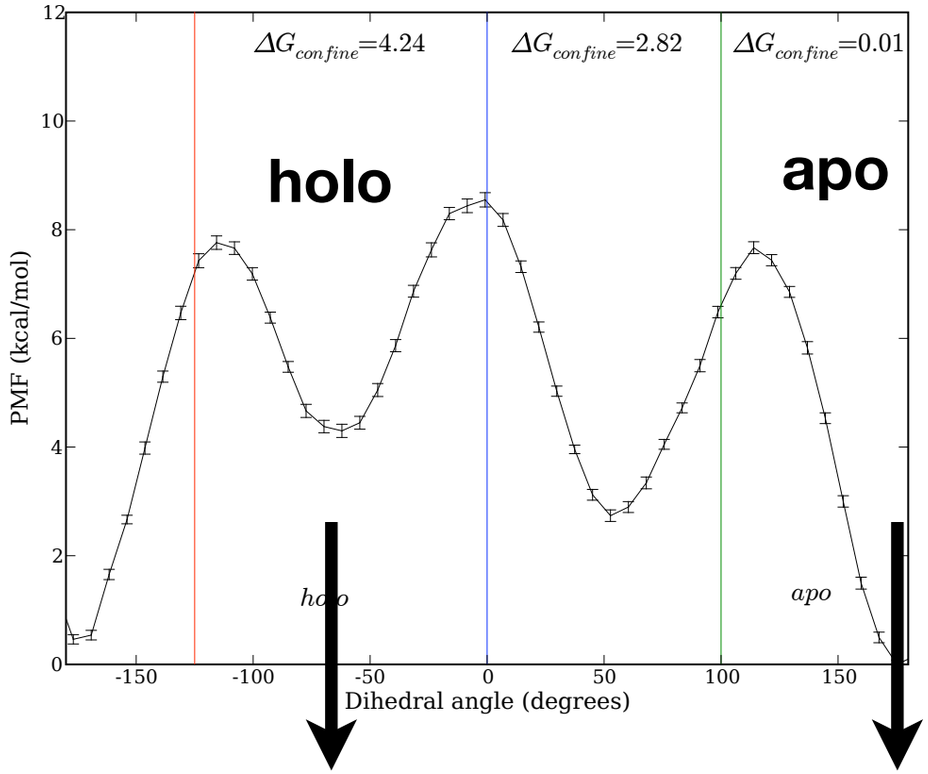
N poses can contribute to the overall binding free energy ~ kT ln N

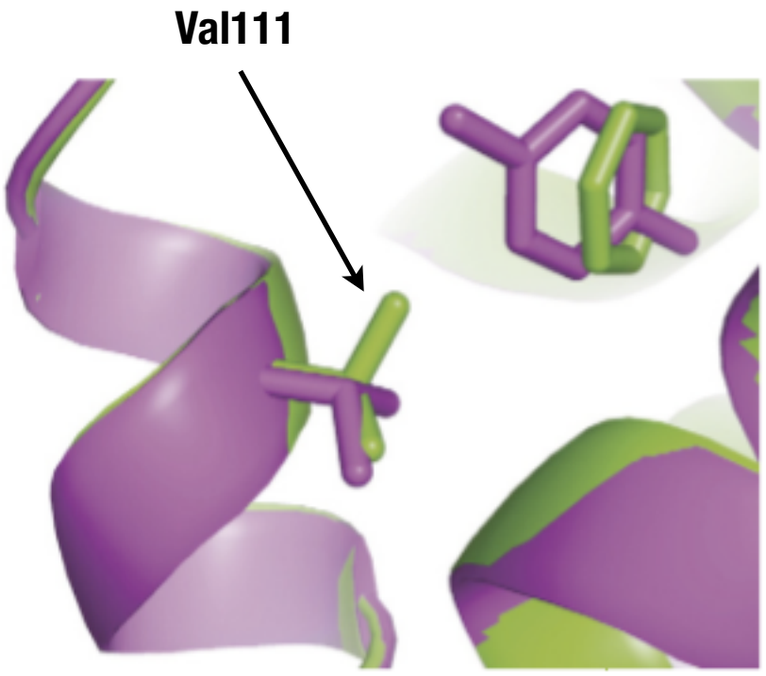
[Also relevant for ligands with pseudosymmetric substituents]

Mobley, Chodera, Dill. JCP 125:084902, 2006.

Multiple protein conformations can contribute

Val111 sidechain χ_1 in apo structure





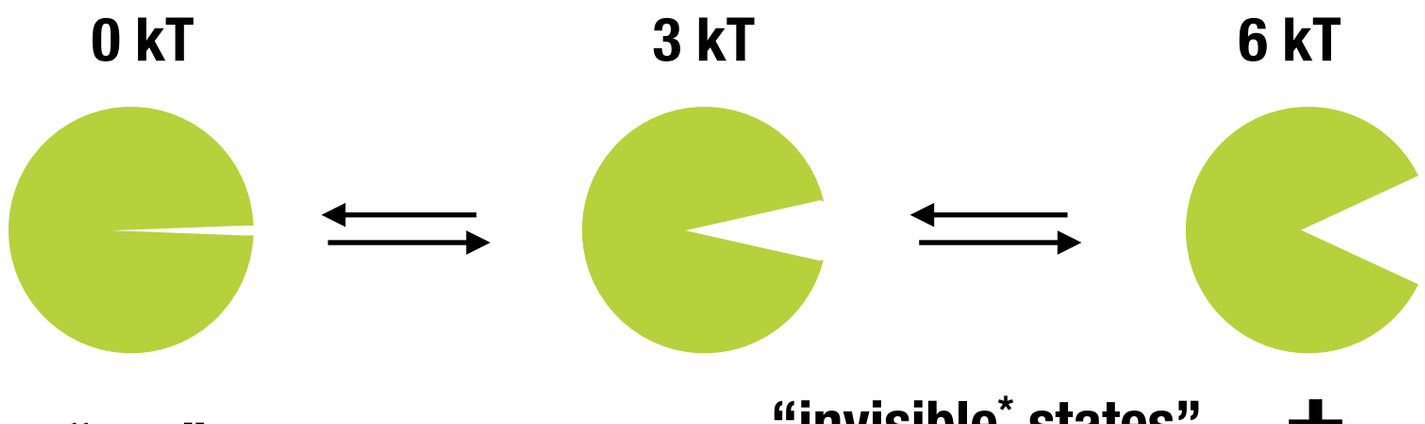
p-xylene vs. benzene bound to T4 lysozyme L99A

binding free energy	-7.3 kcal/mol	-3.0 kcal/mol	**4.3 kcal/mol difference!**

ΔG_{exp} = -4.7 kcal/mol

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

Multiple protein conformations can contribute

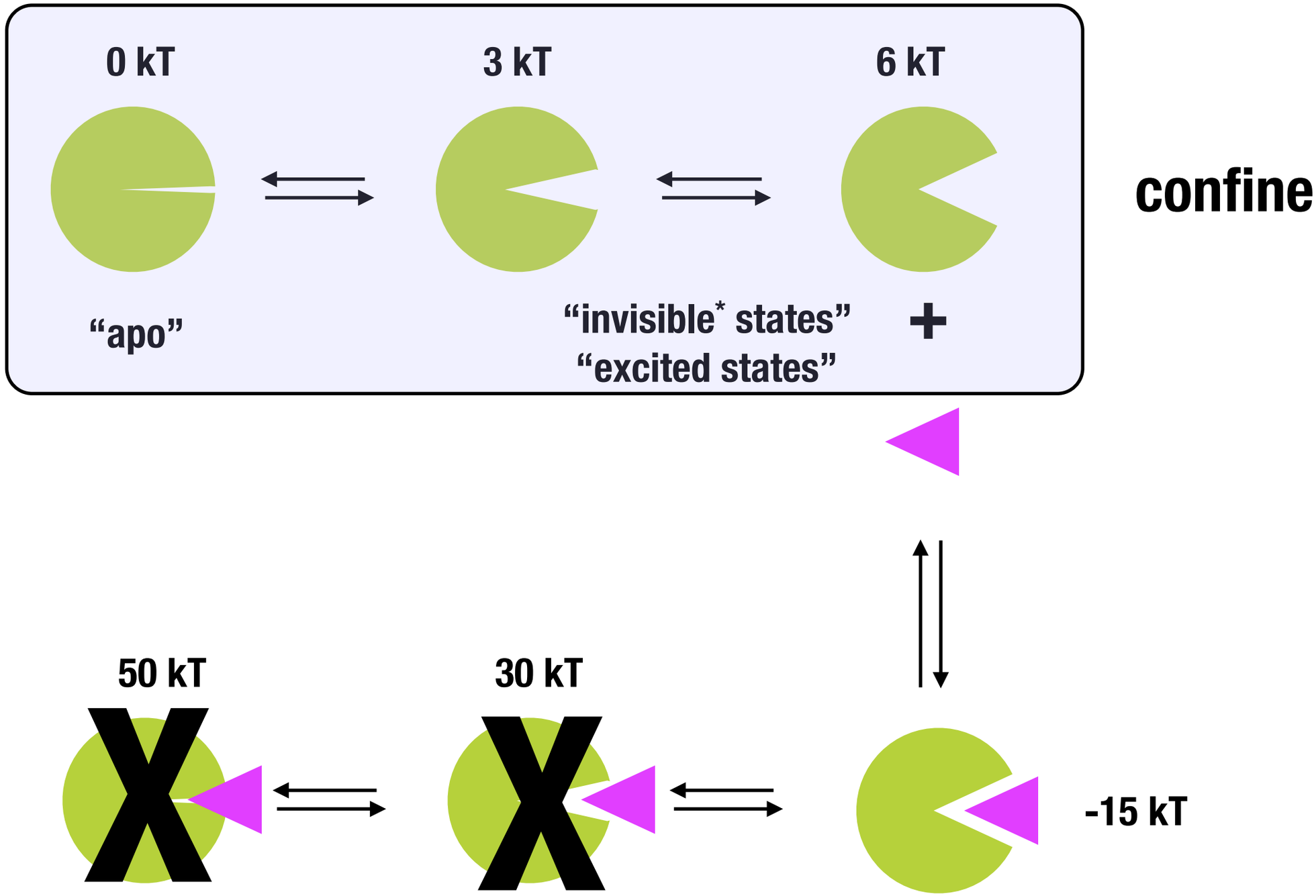


$$K_d \propto \frac{\sum_{i \in \text{unbound}} e^{-\beta F_i}}{\sum_{j \in \text{bound}} e^{-\beta F_j}}$$

*generally invisible to structural biology techniques unless trapped by a ligand

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

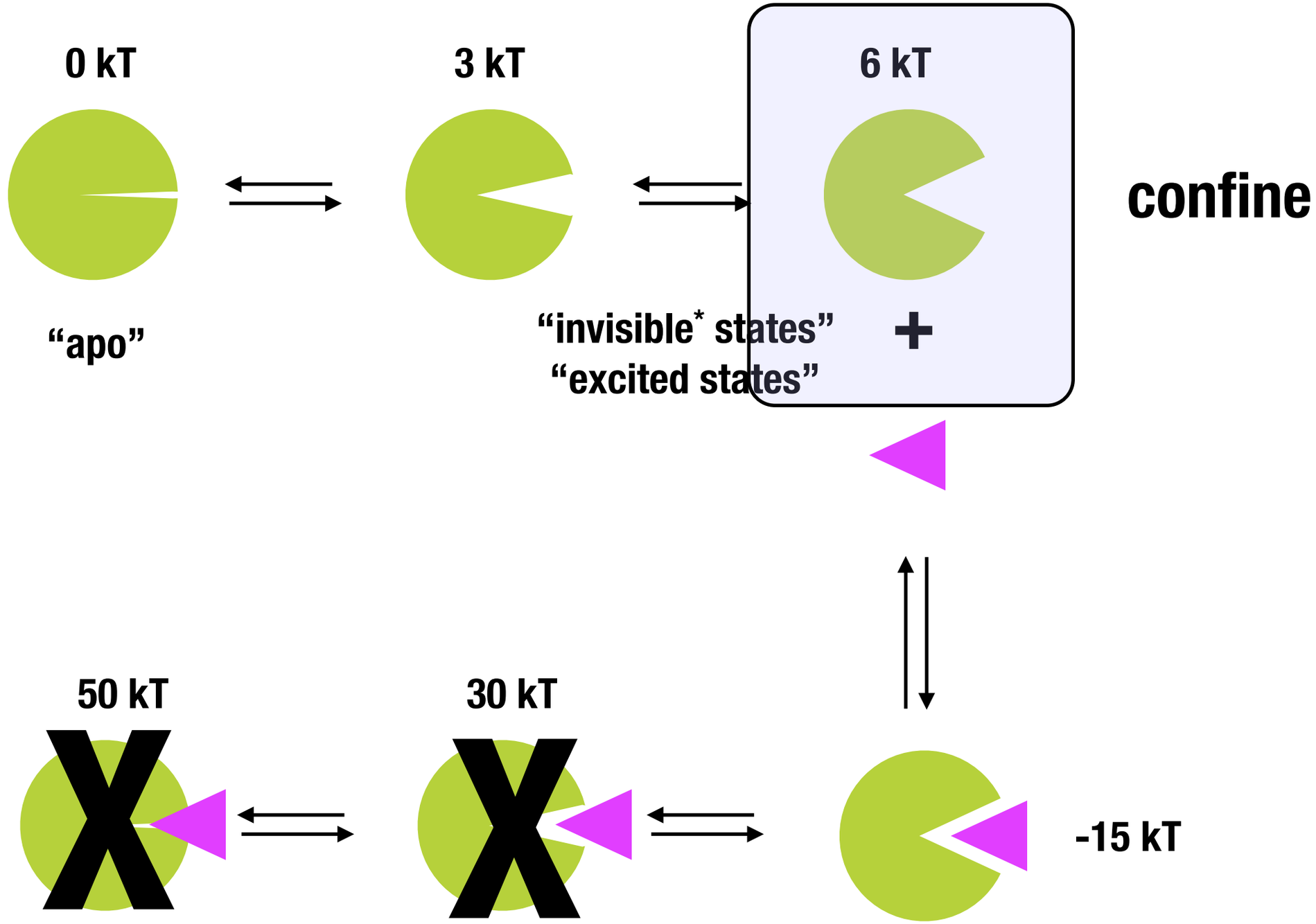
Confine and release: A ‘divide-and-conquer’ strategy



*generally invisible to structural biology techniques unless trapped by a ligand

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

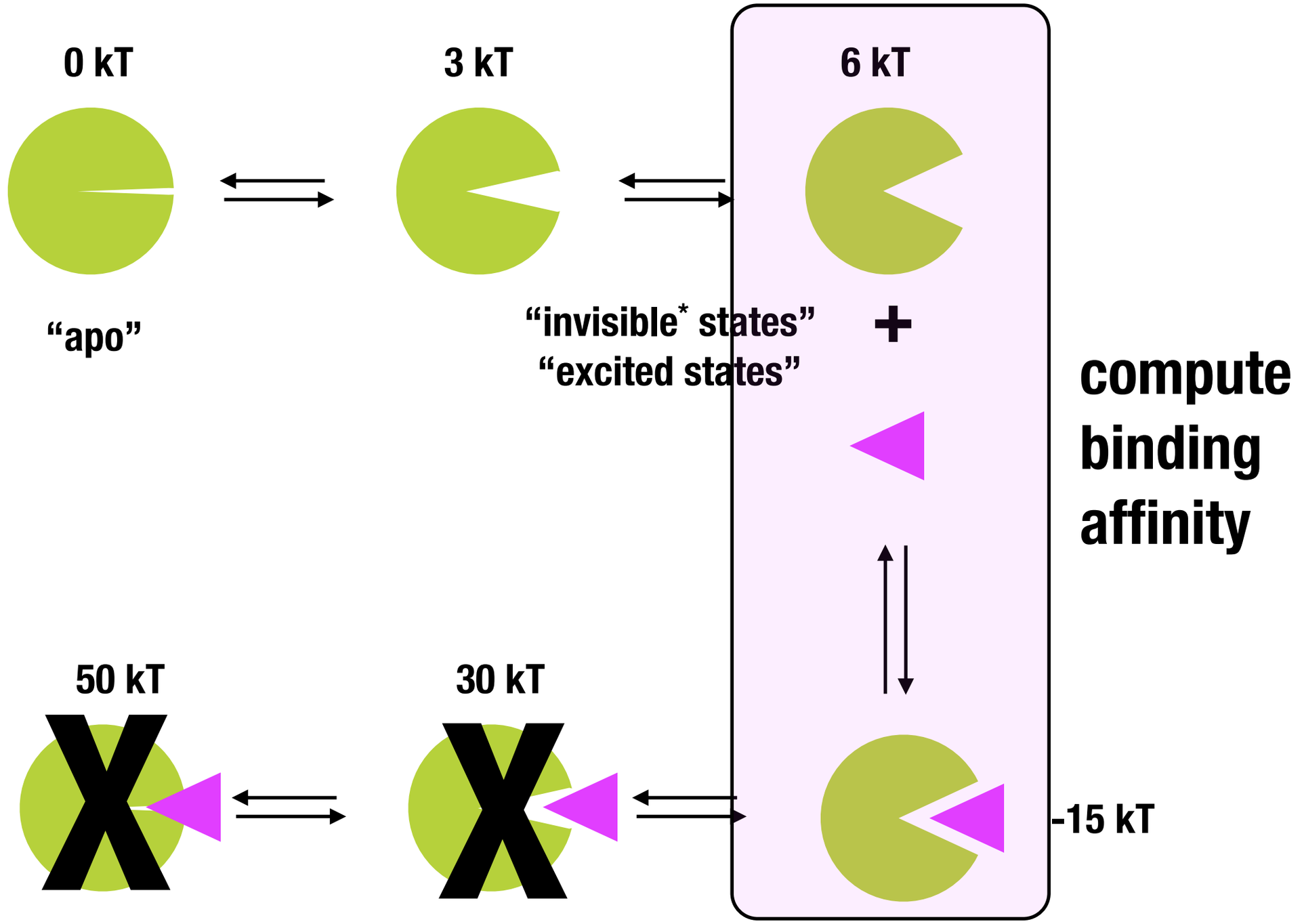
Confine and release: A ‘divide-and-conquer’ strategy



*generally invisible to structural biology techniques unless trapped by a ligand

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

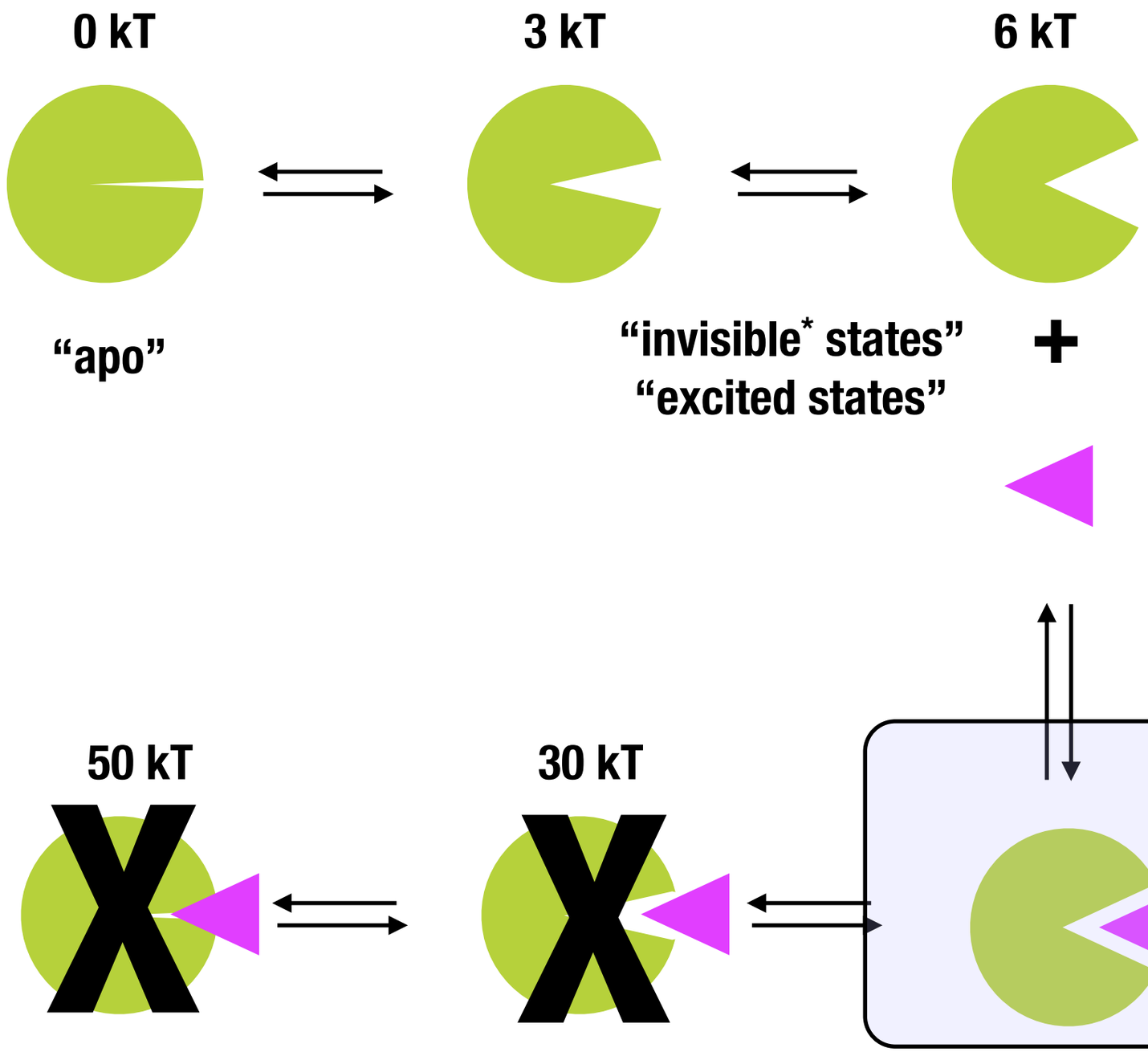
Confine and release: A ‘divide-and-conquer’ strategy



*generally invisible to structural biology techniques unless trapped by a ligand

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

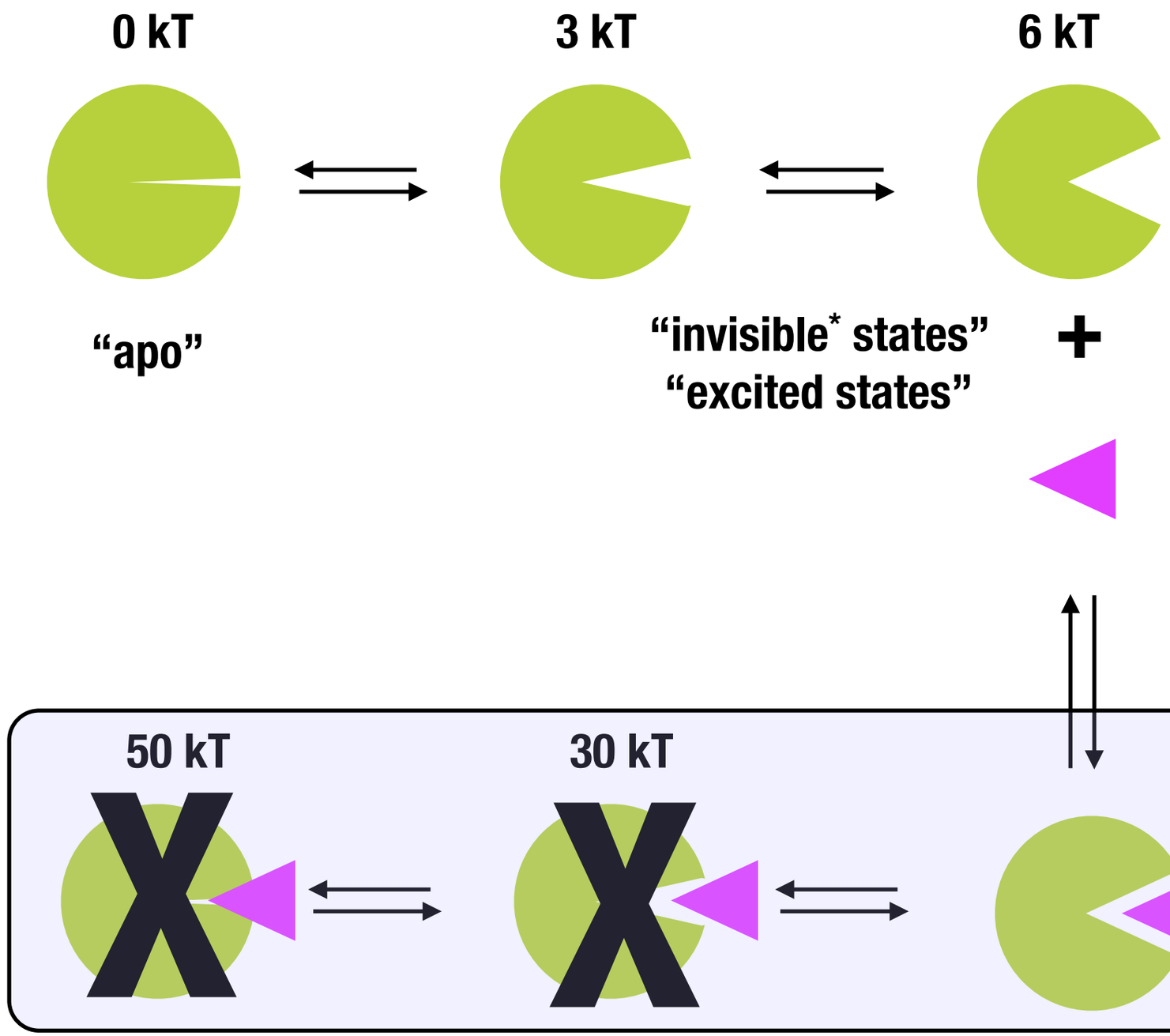
Confine and release: A ‘divide-and-conquer’ strategy



*generally invisible to structural biology techniques unless trapped by a ligand

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

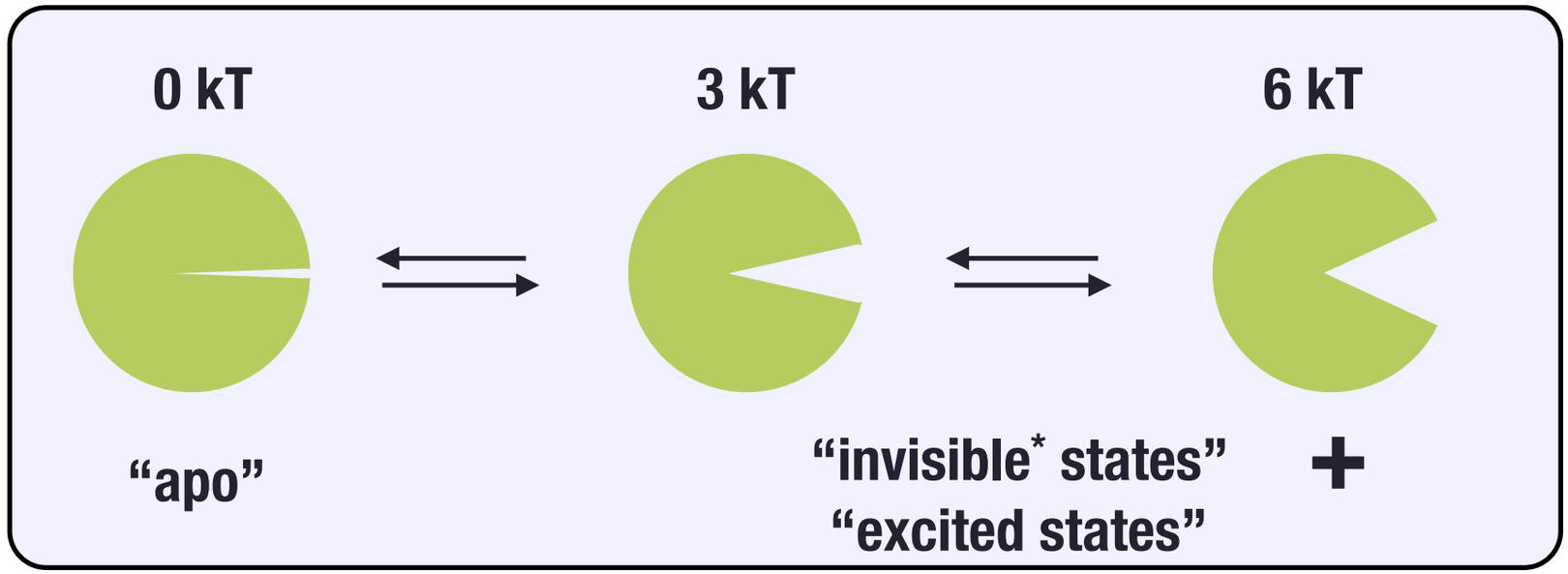
Confine and release: A ‘divide-and-conquer’ strategy



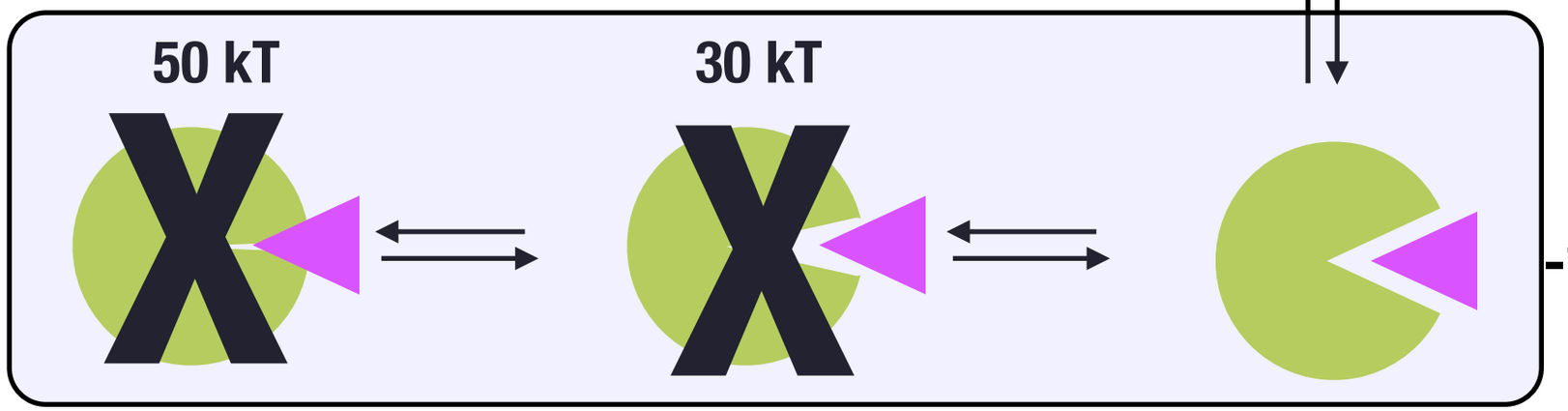
*generally invisible to structural biology techniques unless trapped by a ligand

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

Confine and release: A ‘divide-and-conquer’ strategy



recover overall binding affinity $K_d \propto \dfrac{\sum_{i \in \text{unbound}} e^{-\beta F_i}}{\sum_{j \in \text{bound}} e^{-\beta F_j}}$

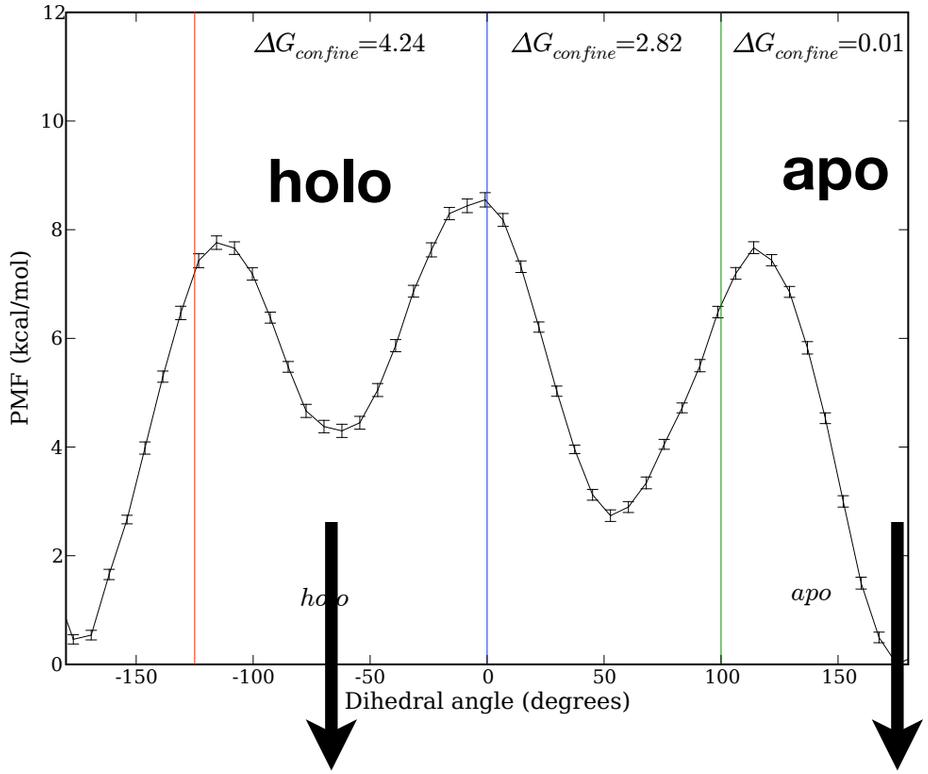


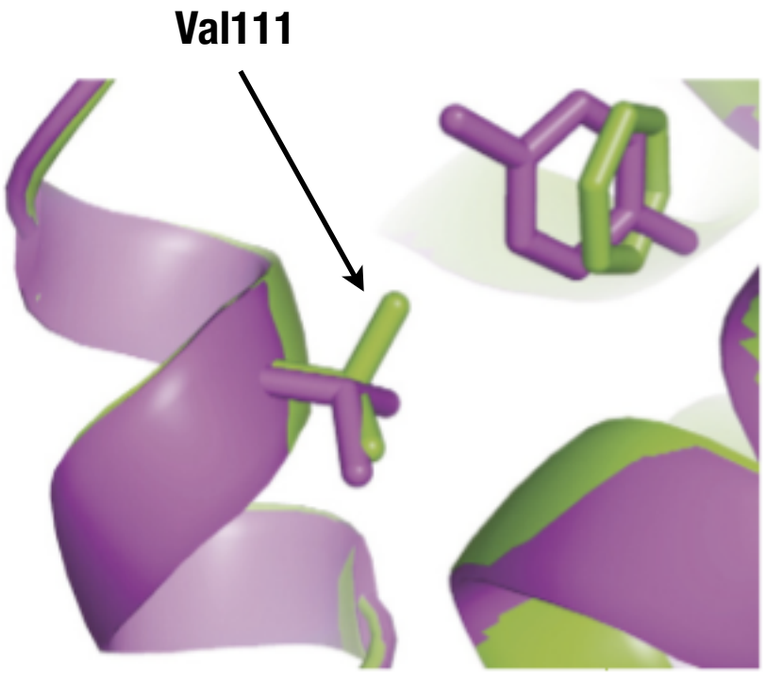
*generally invisible to structural biology techniques unless trapped by a ligand

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

Multiple protein conformations can contribute

Val111 sidechain χ_1 in apo structure





p-xylene vs. benzene bound to T4 lysozyme L99A

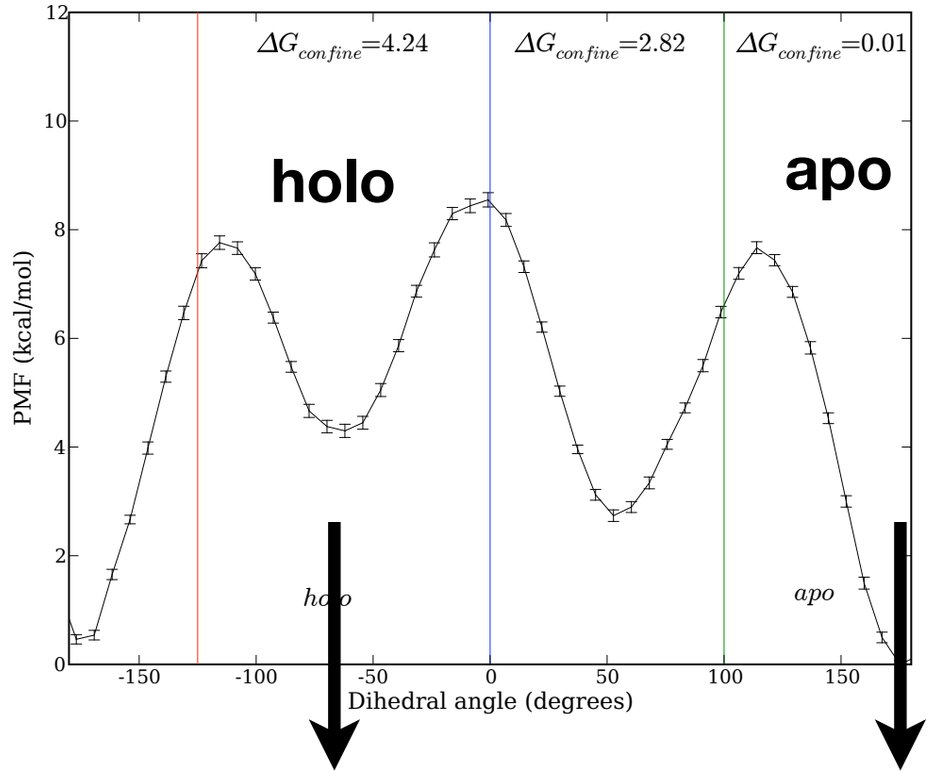
binding free energy	-7.3 kcal/mol	-3.0 kcal/mol	**4.3 kcal/mol difference!**

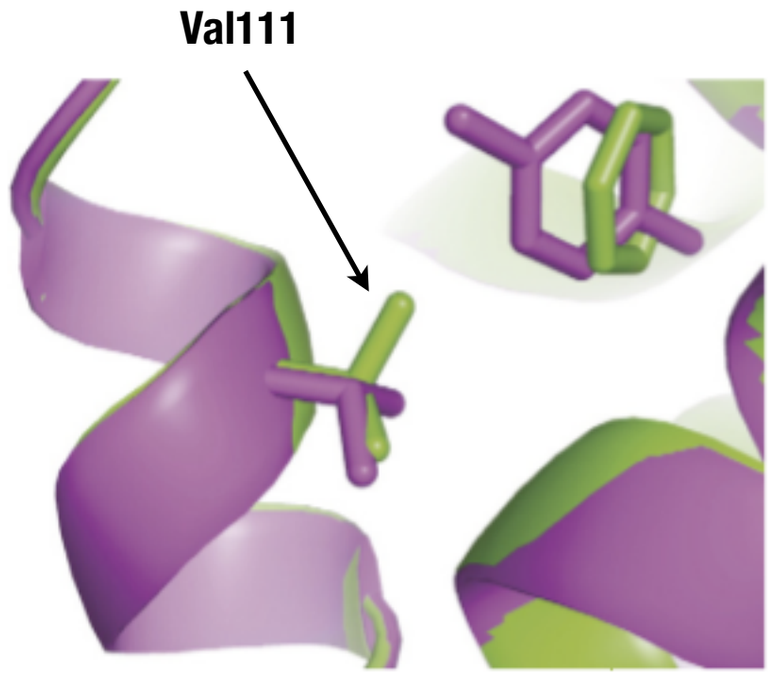
ΔG_{exp} = -4.7 kcal/mol

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

Multiple protein conformations can contribute

Val111 sidechain χ_1 in apo structure





p-xylene vs. benzene bound to T4 lysozyme L99A

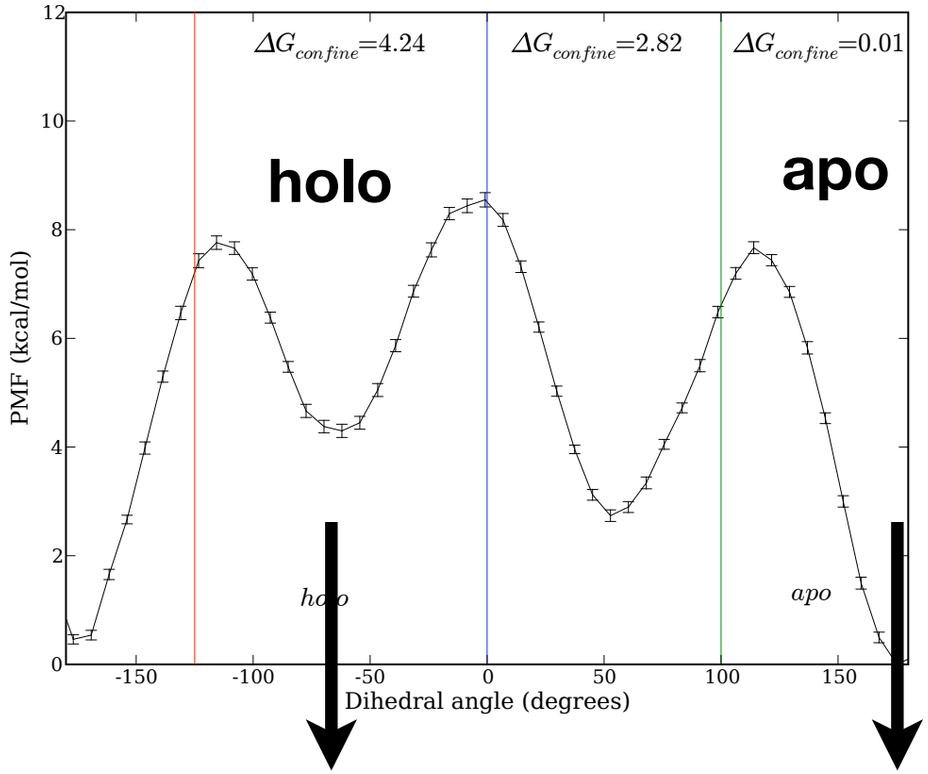
binding free energy	-7.3 kcal/mol	-3.0 kcal/mol	**4.3 kcal/mol difference!**
confinement free energy	4.2 kcal/mol	0.0 kcal/mol	

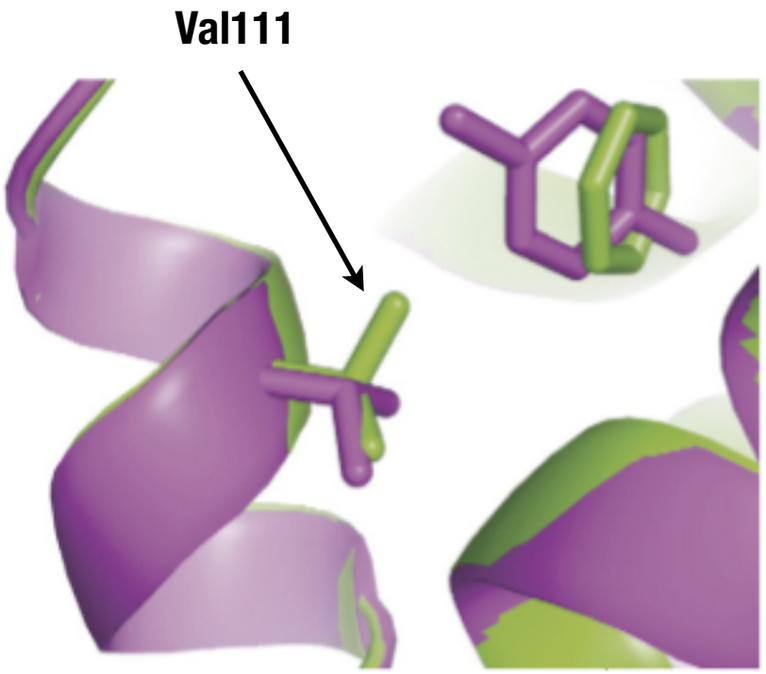
ΔG_{exp} = -4.7 kcal/mol

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

Multiple protein conformations can contribute

Val111 sidechain χ_1 in apo structure





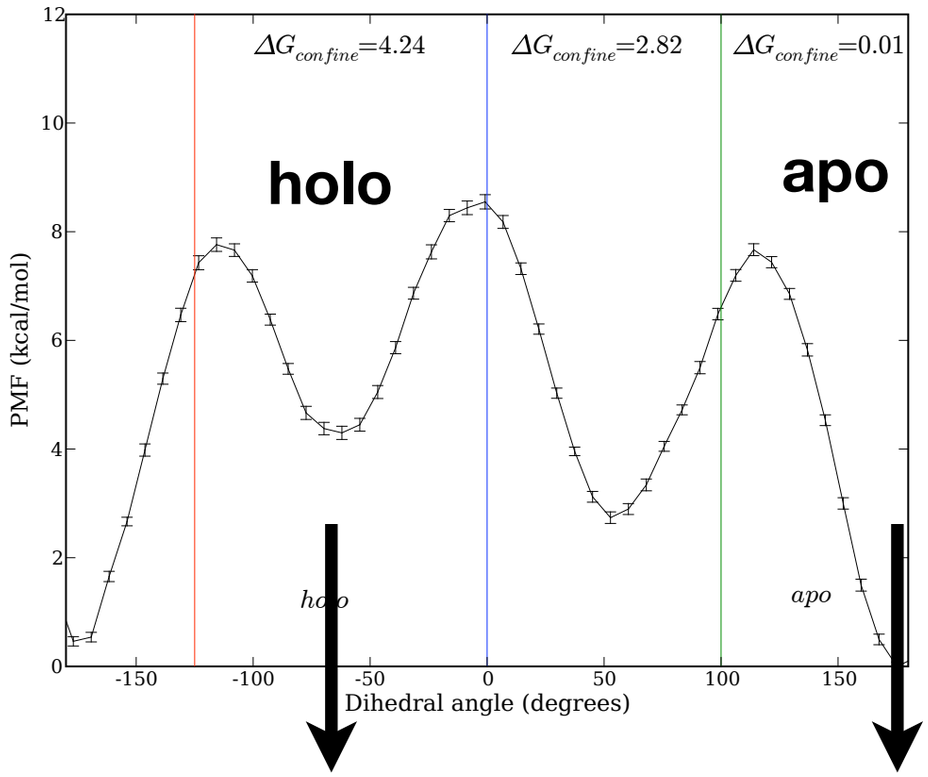
p-xylene vs. benzene bound to T4 lysozyme L99A

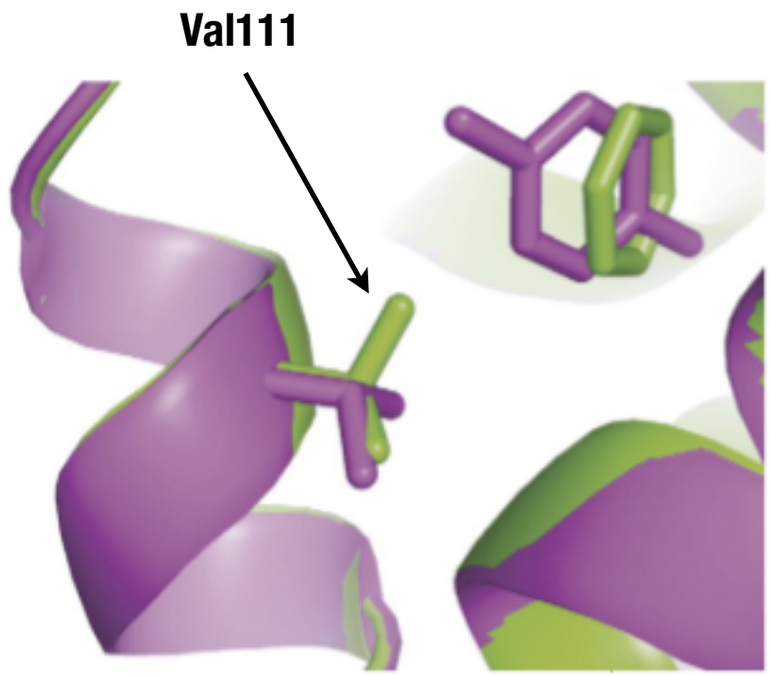
binding free energy	-7.3 kcal/mol	-3.0 kcal/mol	**4.3 kcal/mol difference!**
confinement free energy	4.2 kcal/mol	0.0 kcal/mol	
release free energy following binding	-0.3 kcal/mol	-0.6 kcal/mol	

ΔG_{exp} = -4.7 kcal/mol

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

Multiple protein conformations can contribute





p-xylene vs. benzene bound to T4 lysozyme L99A

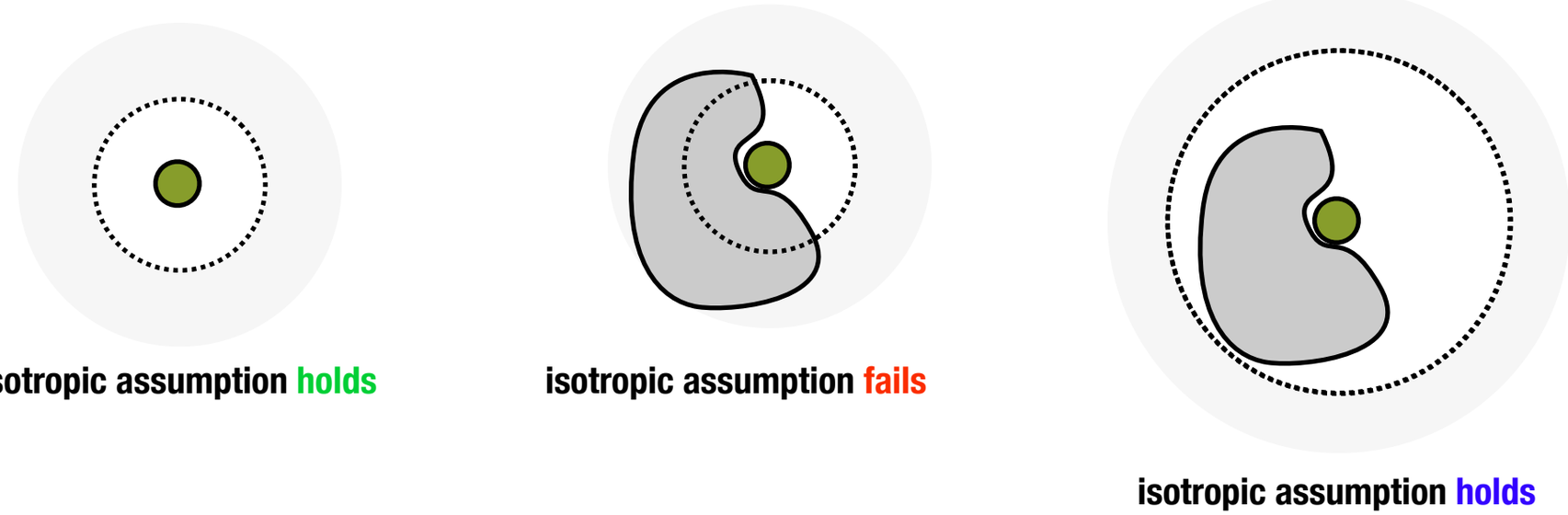
binding free energy	-7.3 kcal/mol	-3.0 kcal/mol	**4.3 kcal/mol difference!**
confinement free energy	4.2 kcal/mol	0.0 kcal/mol	
release free energy following binding	-0.3 kcal/mol	-0.6 kcal/mol	
total binding free energy	-3.4+-0.3	≈ -3.6+-0.3	**ΔG_{exp} = -4.7 kcal/mol**

Mobley, Chodera, and Dill. JCTC 3:1231, 2007.

Anisotropic long-range dispersion correction

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

This correction assumes isotropic distribution of Lennard-Jones sites throughout system.



isotropic assumption holds

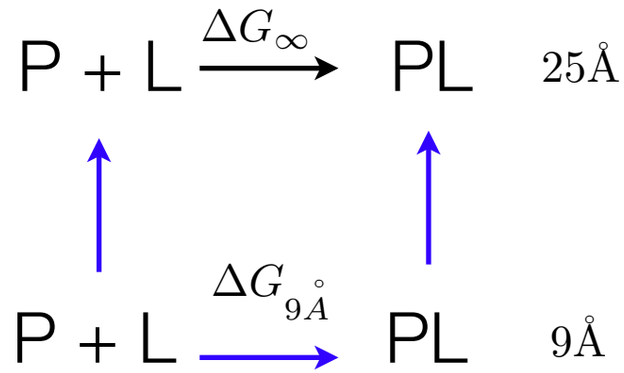
isotropic assumption fails

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



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).

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?

Mobley, Graves, Chodera, McReynolds, Shoichet, Dill. J Mol Biol 371:1118, 2007

How accurately can we predict unknown binding affinities?

Brian Shoichet, UCSF

Blinded test of prediction power with new molecules:

Ligand	Prediction[1]	ΔG^{o}_{calc}[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

Mobley, Graves, Chodera, McReynolds, Shoichet, Dill. J Mol Biol 371:1118, 2007

How accurately can we predict unknown binding affinities?

Brian Shoichet, UCSF

Blinded test of prediction power with new molecules:

Ligand	Prediction[1]	ΔG^o_{calc}[2] (kcal/mol)	Experiment	$\Delta G^o_{expt.}$ (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!

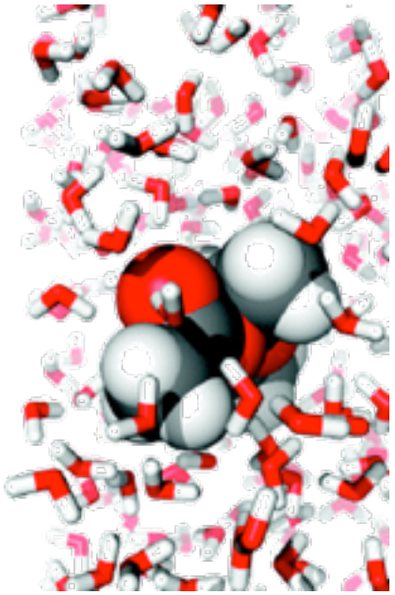
RMS error: 0.6 kcal/mol

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

Mobley, Graves, Chodera, McReynolds, Shoichet, Dill. J Mol Biol 371:1118, 2007

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?)

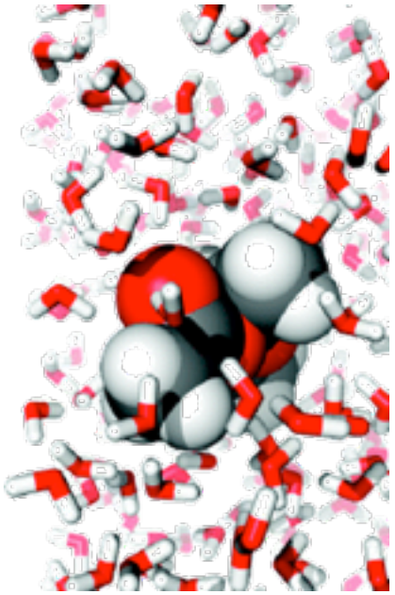
Calculations are notoriously unreliable. (e.g. SAMPL challenges)

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

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

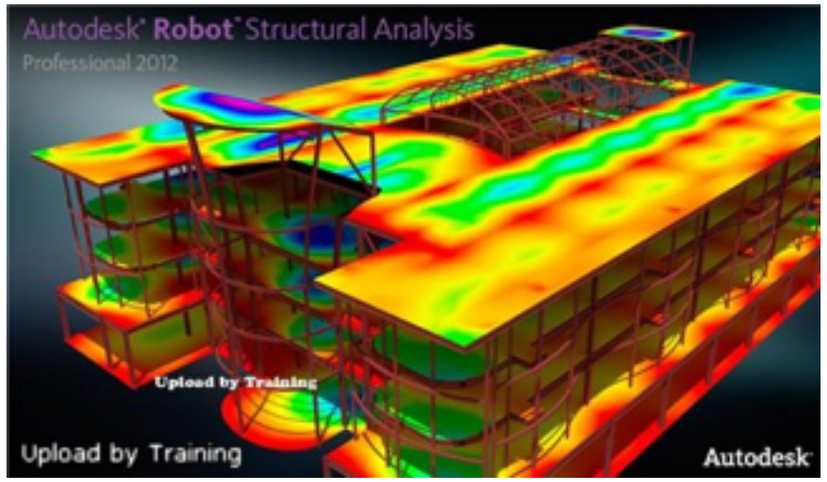
small, rigid protein
fixed protonation states
larger drug-like ligands
with few rotatable bonds

JNK3 kinase

large protein, multiple conformations
large drug-like ligands, rotatable bonds
multiple protonation states? tautomers?
phosphorylation and activation
peptide substrate?
$MgCl_2$ salt effects?

easy
hard

(not to scale)


Autodesk® Robot™ Structural Analysis
Professional 2012
Upload by Training
Upload by Training
Autodesk

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”

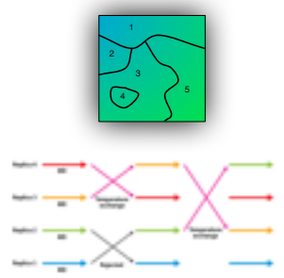
Make rigorous calculations of affinity fast and accurate

GPU acceleration

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

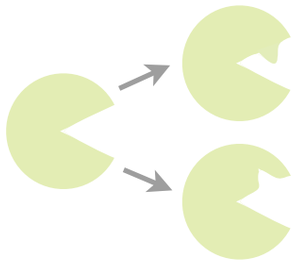
$$AH \leftrightharpoons A^- + H^+$$

modular MCMC for sampling and chemical effects



enhanced sampling algorithms

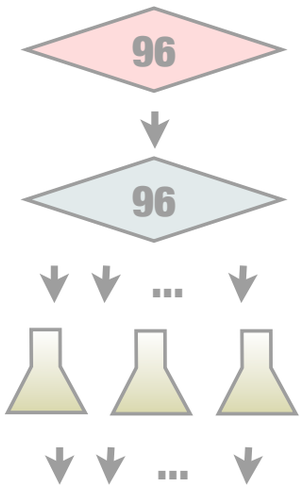
Test and improve models quickly and cheaply by inverting the drug discovery problem



mutate proteins instead of ligands

¢

buy inexpensive ligands



high-throughput experiments

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

How can we speed up the calculations?

$50K
5 TFLOP
doubling every 18 mo

$500
5 TFLOP
doubling every 12 mo

many CPU-weeks/calculation

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

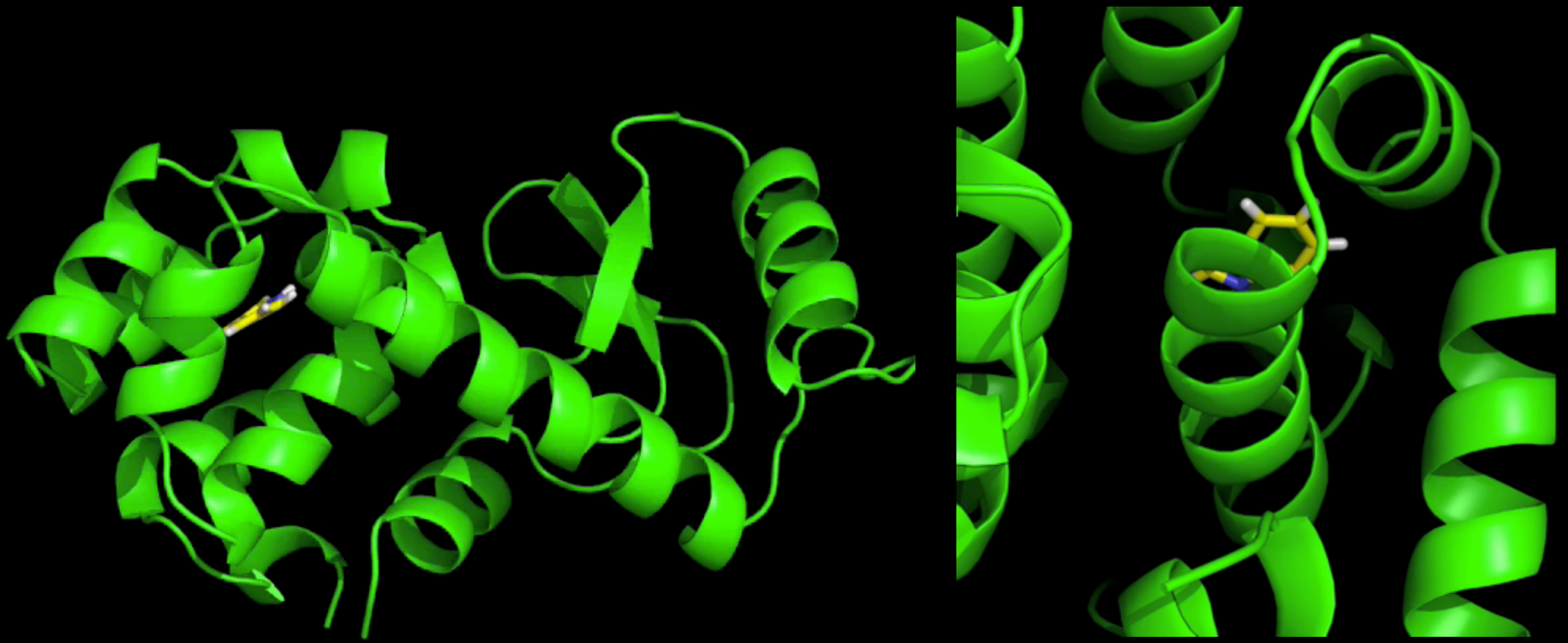
A free, open-source, extensible platform for free energy calculations and ligand design

http://github.com/choderalab/yank

Replica-exchange algorithms facilitate sampling of multiple binding modes

---

solid fully interacting
transparent noninteracting



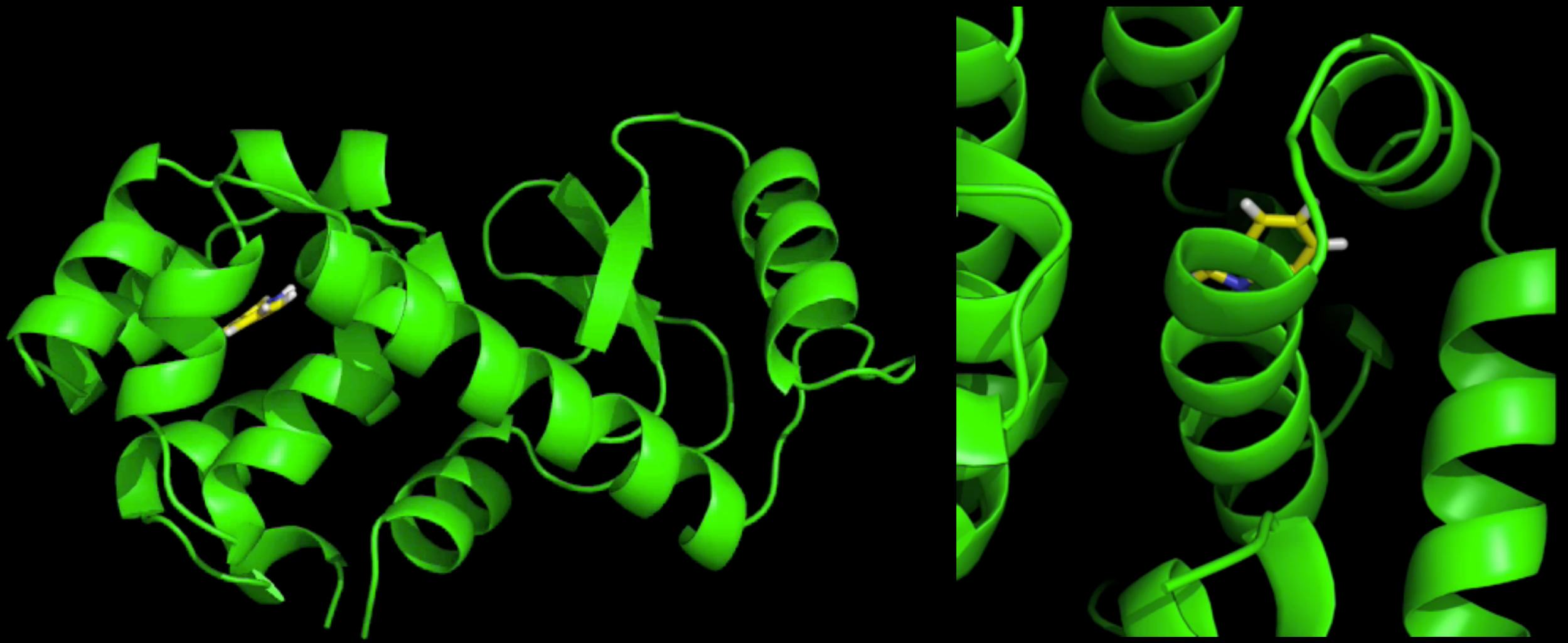
indole binding to T4 lysozyme L99A
12 h on 2 NVIDIA Tesla M2090 GPUs
Hamiltonian exchange with Gibbs sampling

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

Replica-exchange algorithms facilitate sampling of multiple binding modes

---

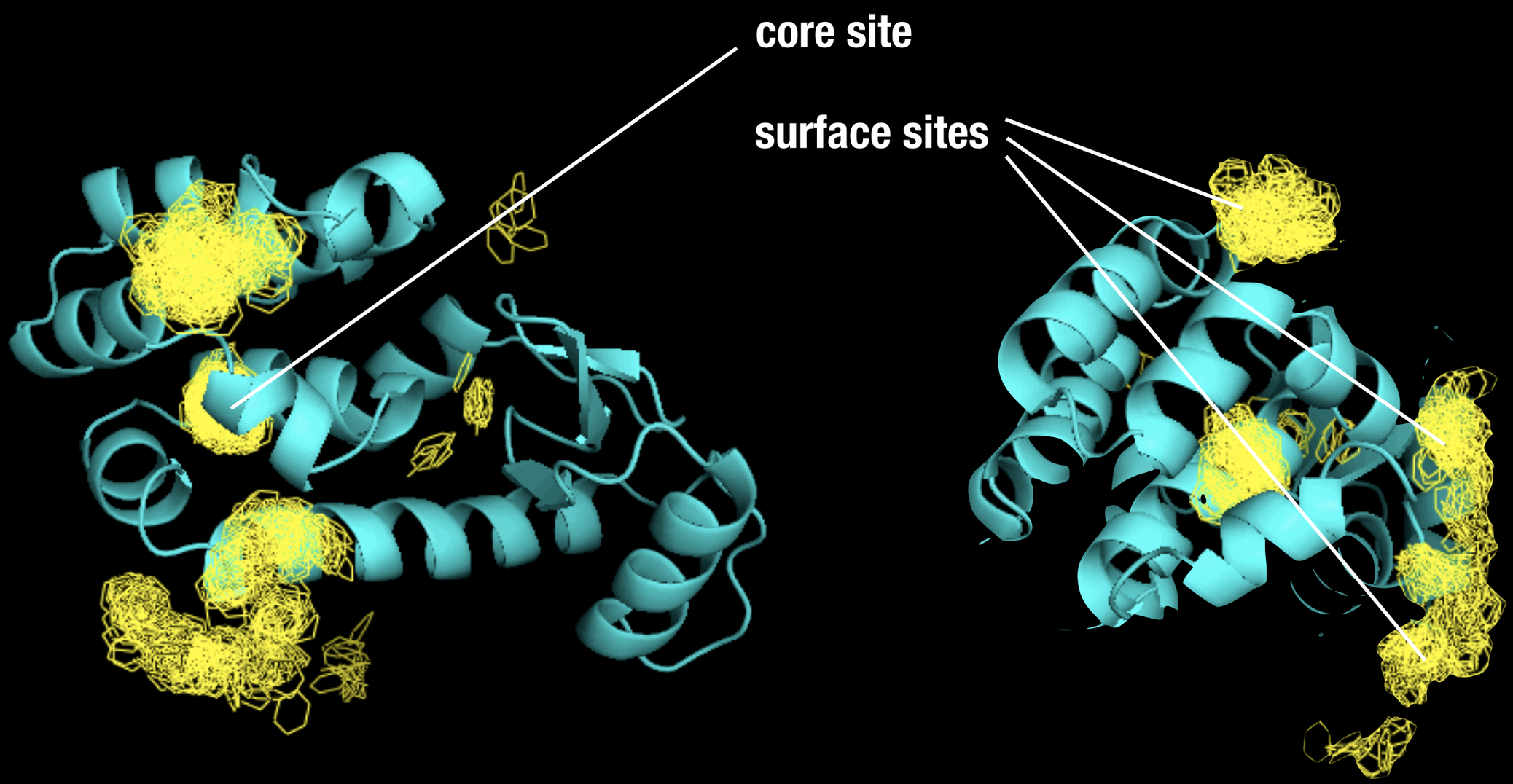
solid fully interacting
transparent noninteracting



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

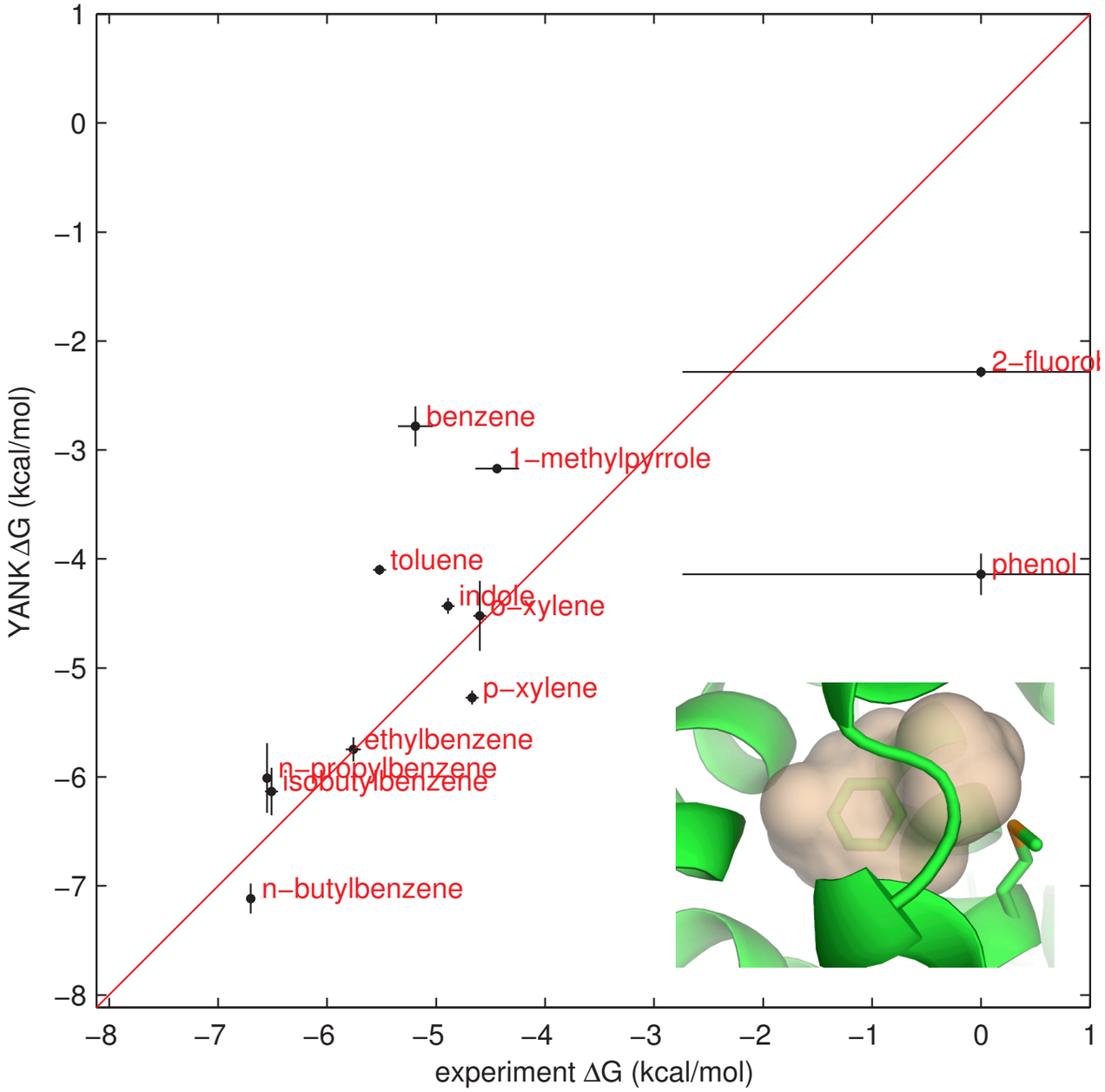


benzene bound to T4 lysozyme L99A
AMBER96 + OBC GBSA

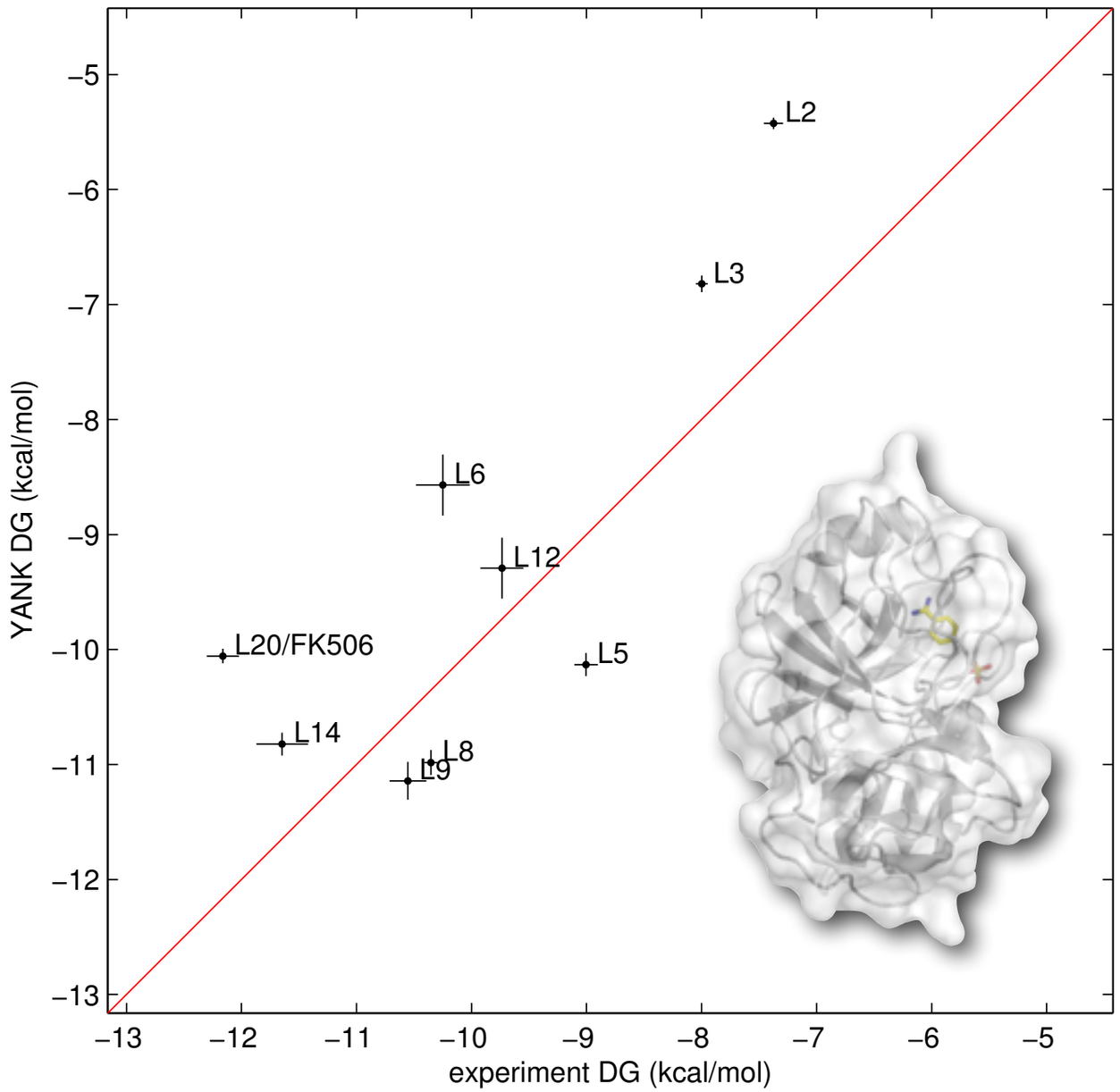
Chodera and Shirts. JCP 135:194110, 2011
Wang, Chodera, Yang, and Shirts. JCAMD 27:989, 2013.
http://github.org/choderalab/yank

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

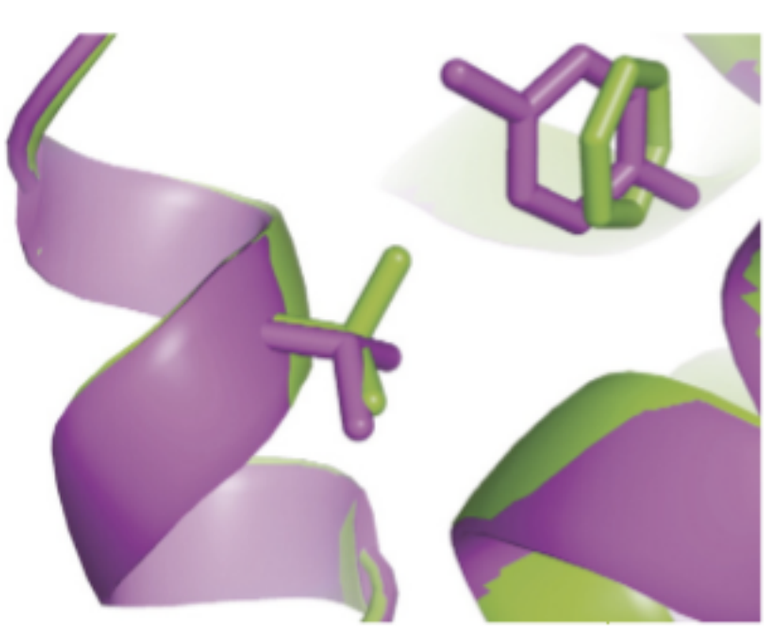
http://github.org/choderalab/yank

[+ Jerome Nilmeier, David D. L. Minh, and Gavin Crooks]

How can we deal with slow sidechains and protonation states? Can Monte Carlo moves help more generally?

Well-designed Monte Carlo moves can be highly efficient.
Good design usually requires knowledge of favorable conformations of system.

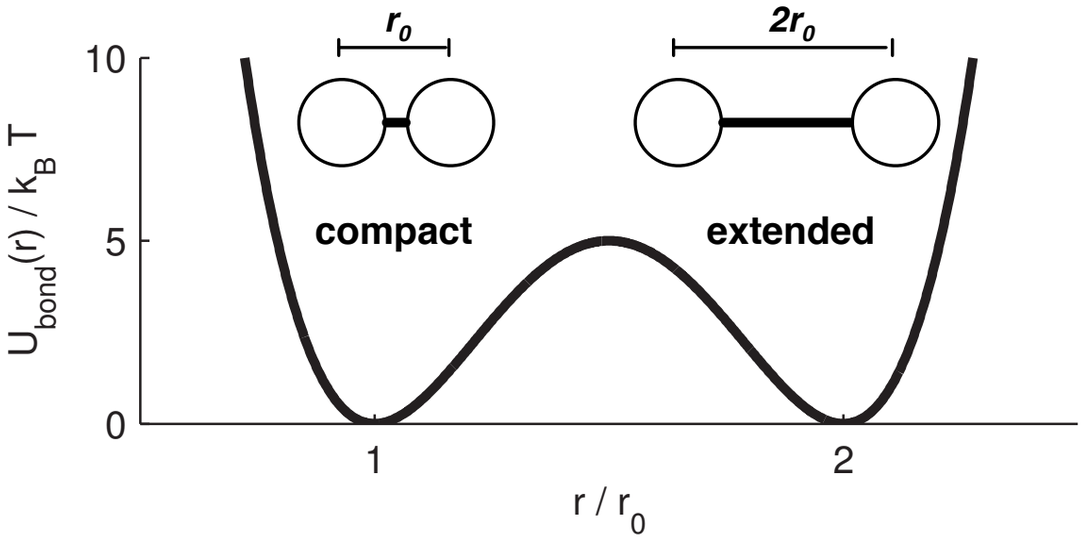
In the implicit solvent, moves like protein sidechain rotamer flips work fantastically well, but are horrendously bad in explicit solvent.



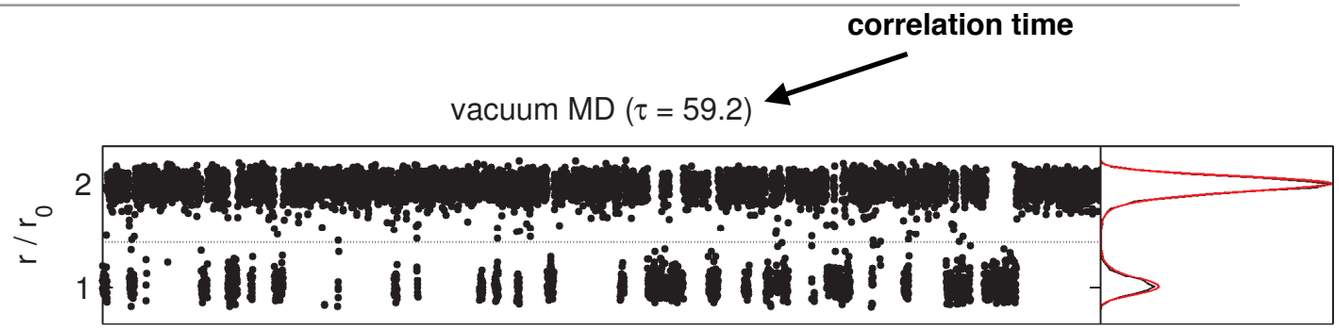
p-xylene vs. benzene bound to T4 lysozyme L99A

How can we deal with this in explicit solvent, or condensed phase systems in general?

Mixing in clever Monte Carlo moves can speed sampling

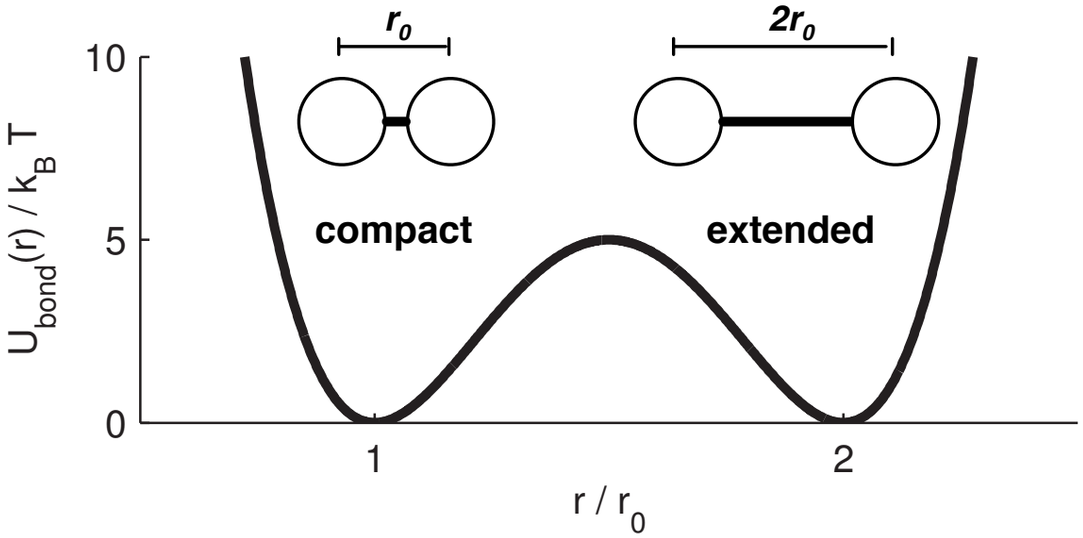


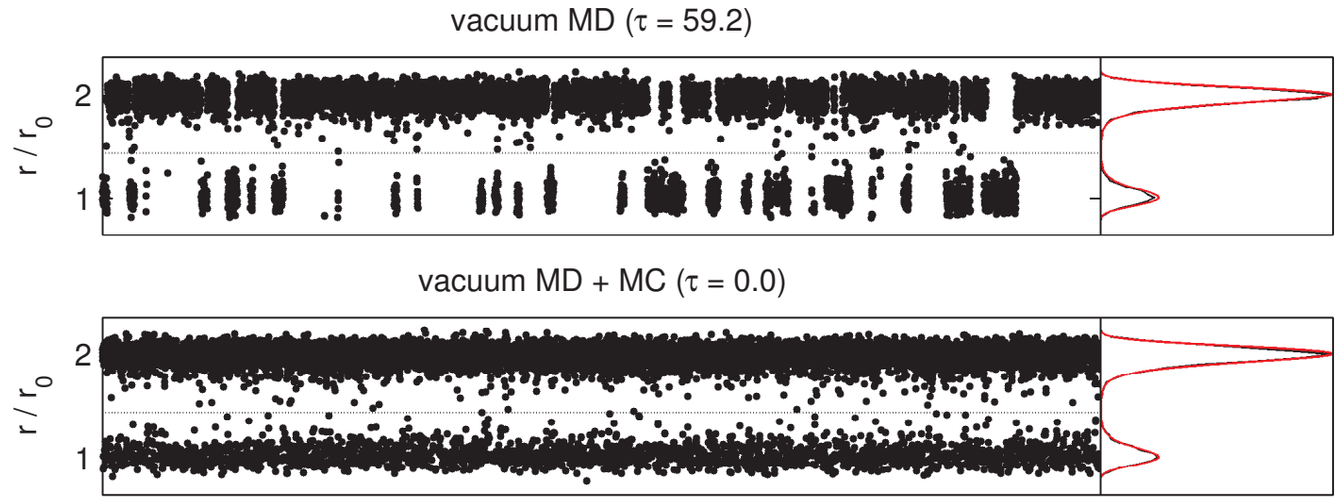
“Geisslerene”

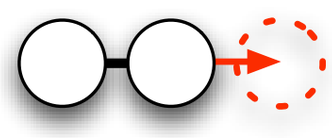


Straightforward molecular dynamics can be slow.

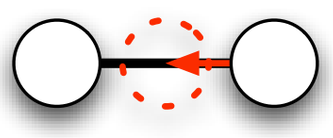
Monte Carlo moves can speed barrier crossing in vacuum...





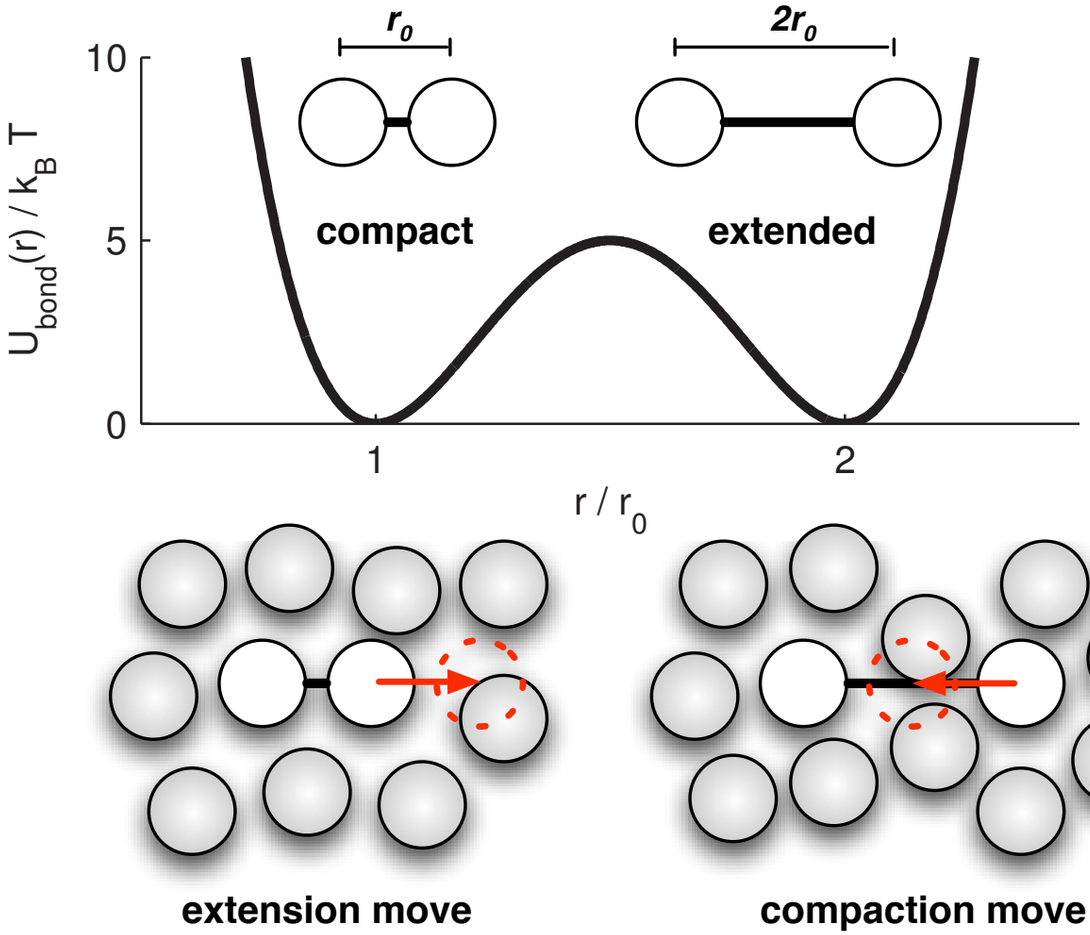


extension move



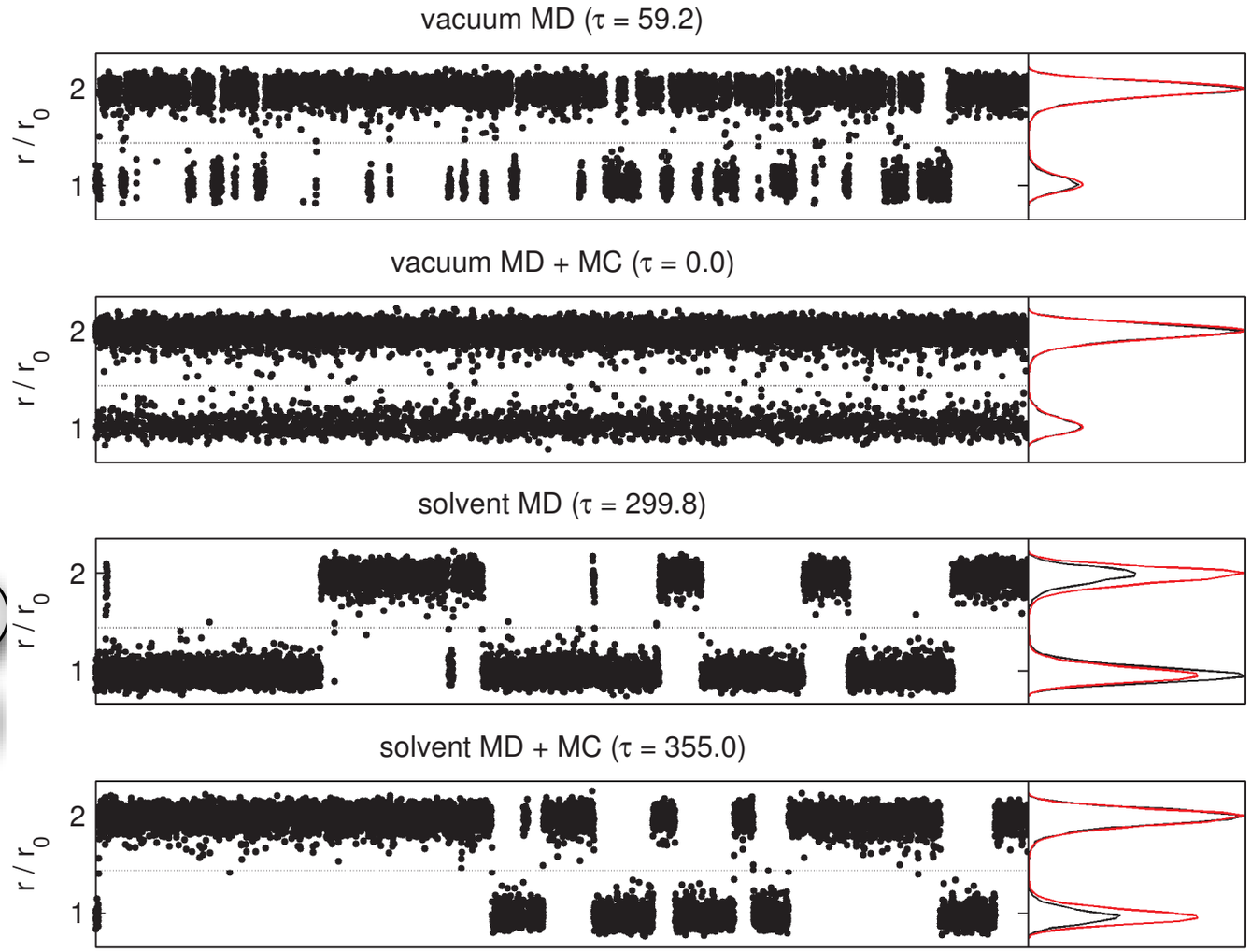
compaction move

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



extension move

compaction move



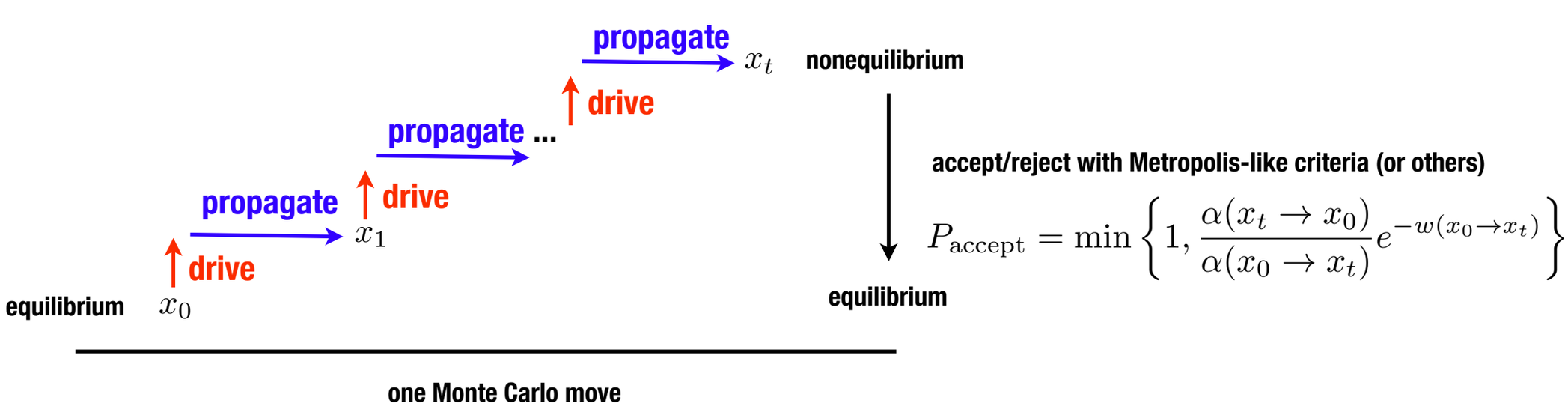
What if we drive some degrees of freedom, but propagate the rest? Can we correct for this in a Markov chain Monte Carlo framework?

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

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

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

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

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

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

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

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

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

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

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

both must be selected with nonzero probability!

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

$$A(X|\Lambda)\,\Pi(X|x_0,\Lambda)\,P(\Lambda|x_0,\lambda_0)\,\pi(x_0,\lambda_0) = A(\tilde{X}|\tilde{\Lambda})\,\Pi(\tilde{X}|\tilde{x}_T,\tilde{\Lambda})\,P(\tilde{\Lambda}|\tilde{x}_T,\lambda_T)\,\pi(\tilde{x}_T,\lambda_T)$$

Result is a general acceptance criteria for any nonequilibrium perturbation:

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

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

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

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

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

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

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

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

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

forward process

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

reverse process

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

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

both must be selected with nonzero probability!

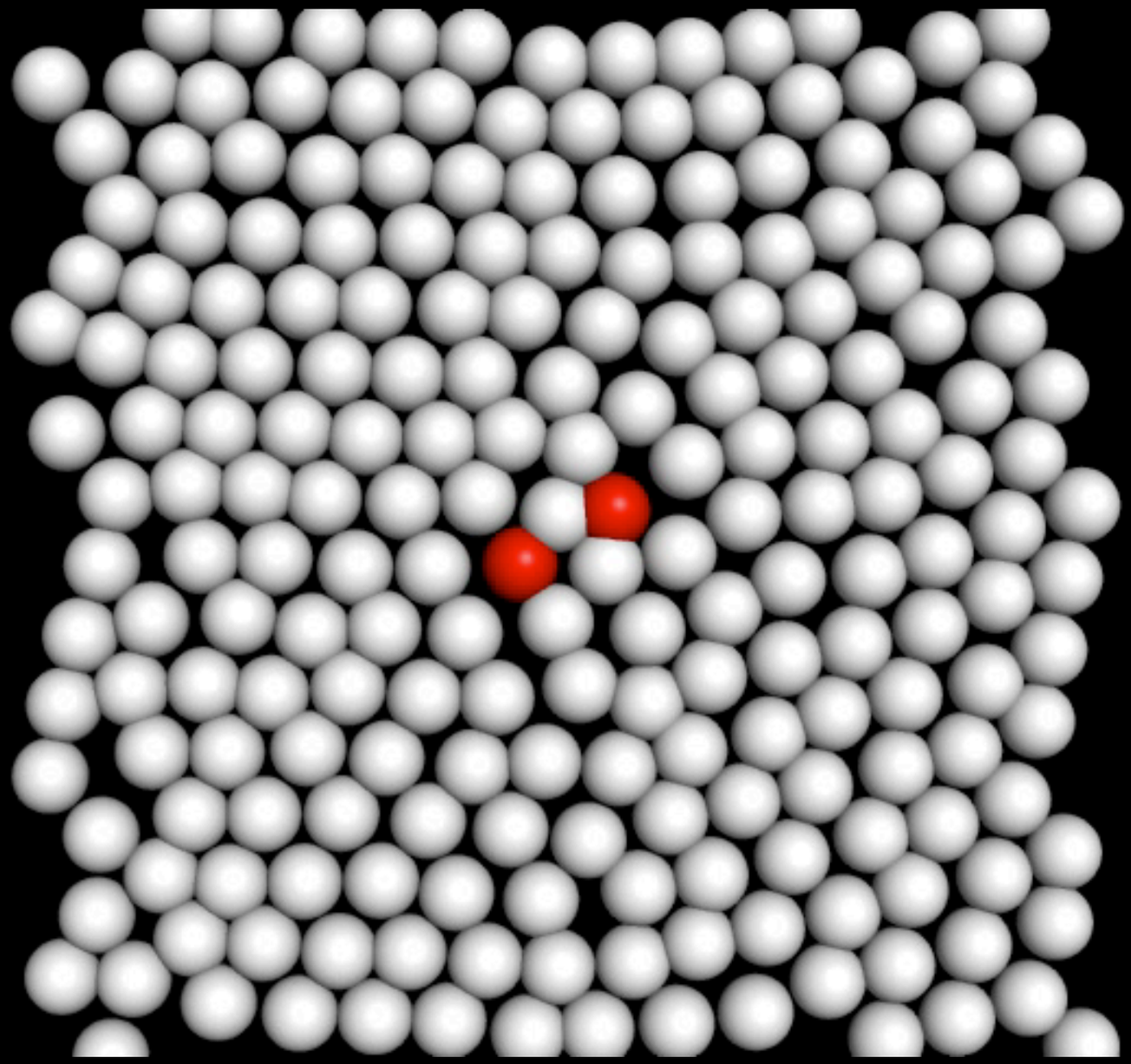
Enforce strict "pathwise" form of detailed balance (which also ensures detailed balance is satisfied):

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

Result is a general acceptance criteria for any nonequilibrium perturbation:

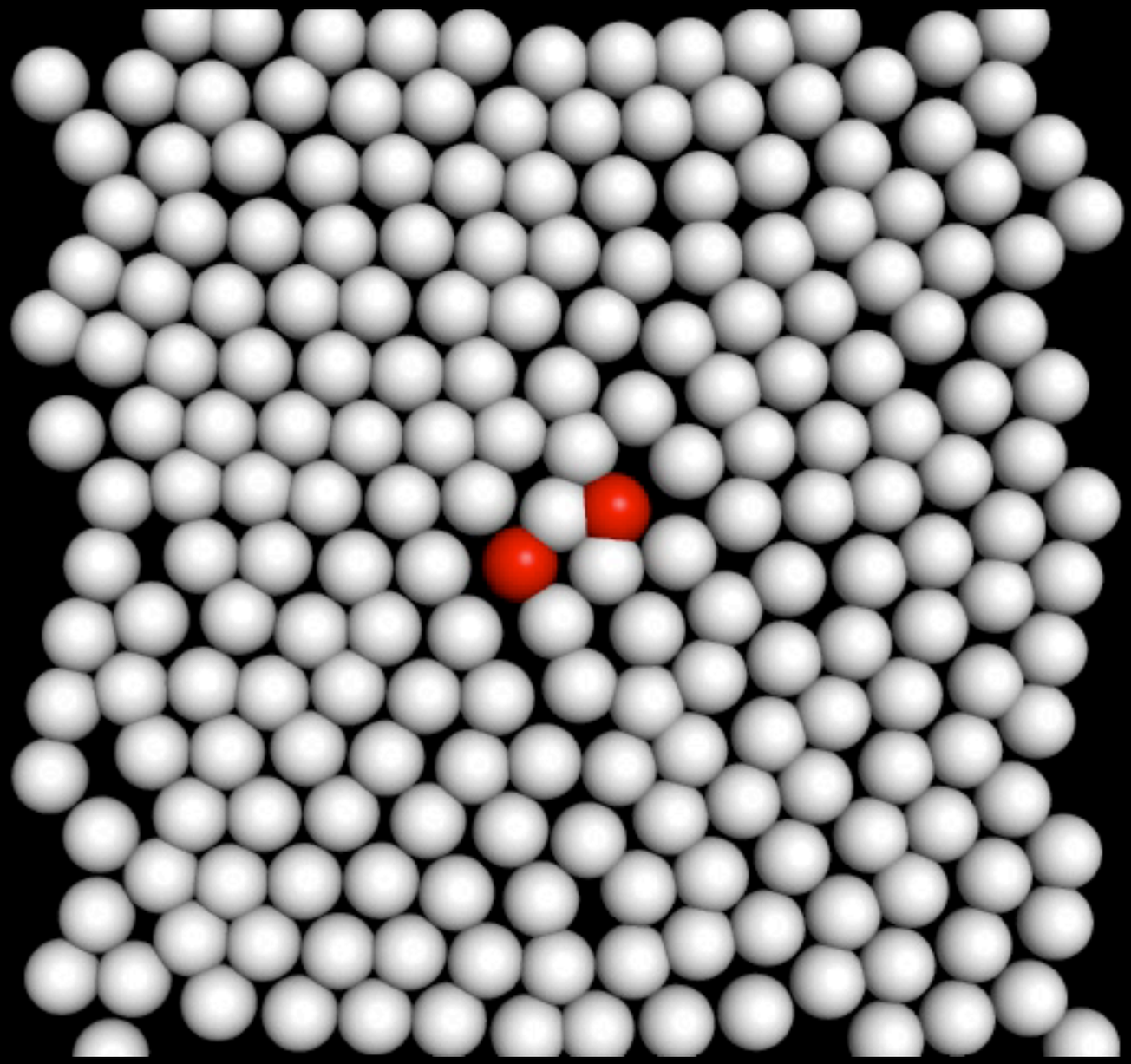
$$A(X|\Lambda) = \min\left\{1, \underbrace{\frac{\pi(x_T, \lambda_T)}{\pi(x_0, \lambda_0)}}_{\text{equilibrium}} \underbrace{\frac{P(\tilde{\Lambda}|\tilde{x}_T, \lambda_T)}{P(\Lambda|x_0, \lambda_0)}}_{\text{selection}} \underbrace{\frac{\tilde{\alpha}(\tilde{X})}{\alpha(X)}}_{\text{perturbation}} \underbrace{e^{-\Delta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

Nonequilibrium MC dimer moves in a 2D WCA fluid



Instantaneous Metropolis Monte Carlo

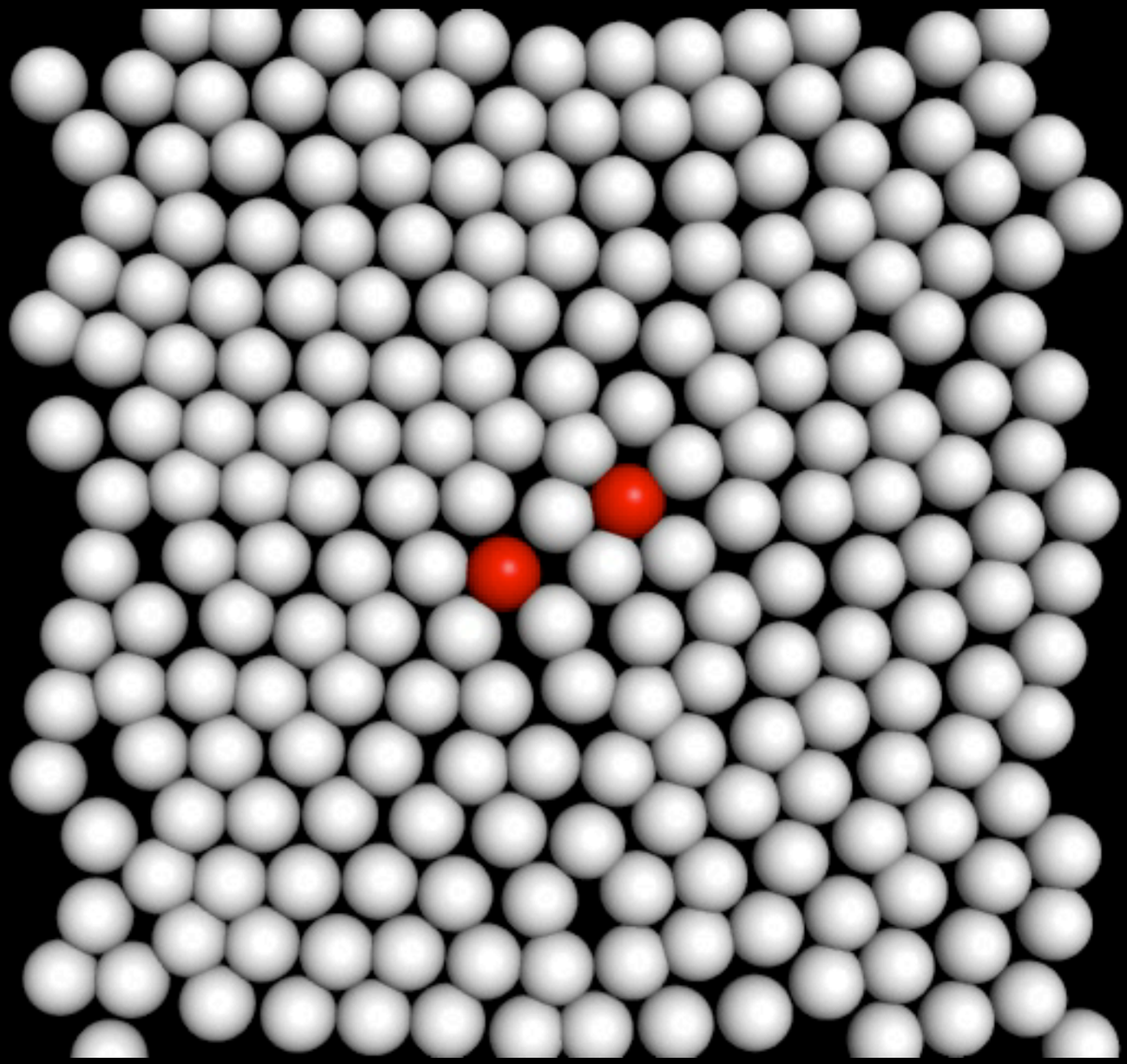
Nonequilibrium MC dimer moves in a 2D WCA fluid



Bad solvent overlap:
Always rejected :(

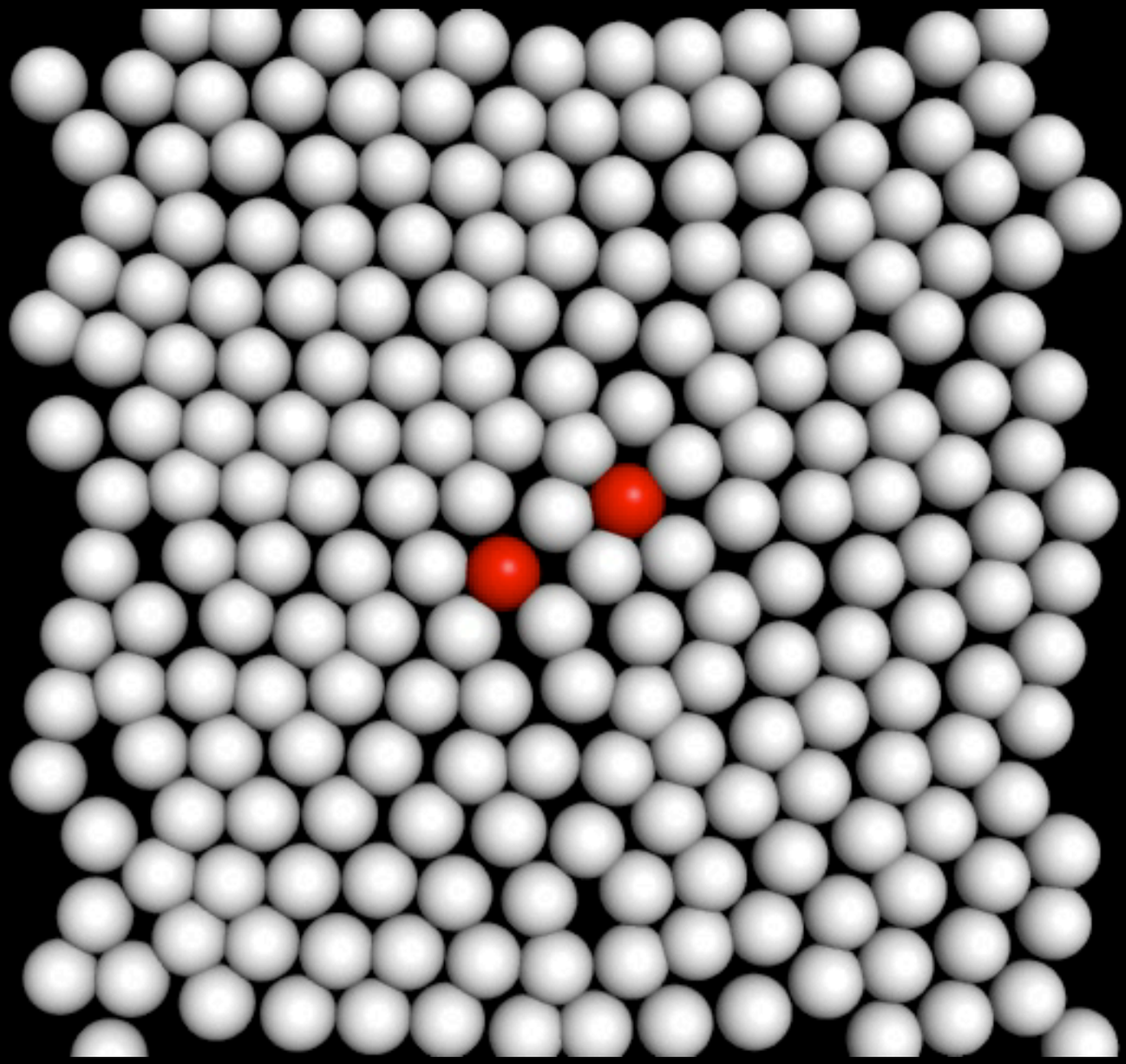
Instantaneous Metropolis Monte Carlo

Nonequilibrium MC dimer moves in a 2D WCA fluid



32 switching steps with velocity Verlet dynamics as “propagation kernel”

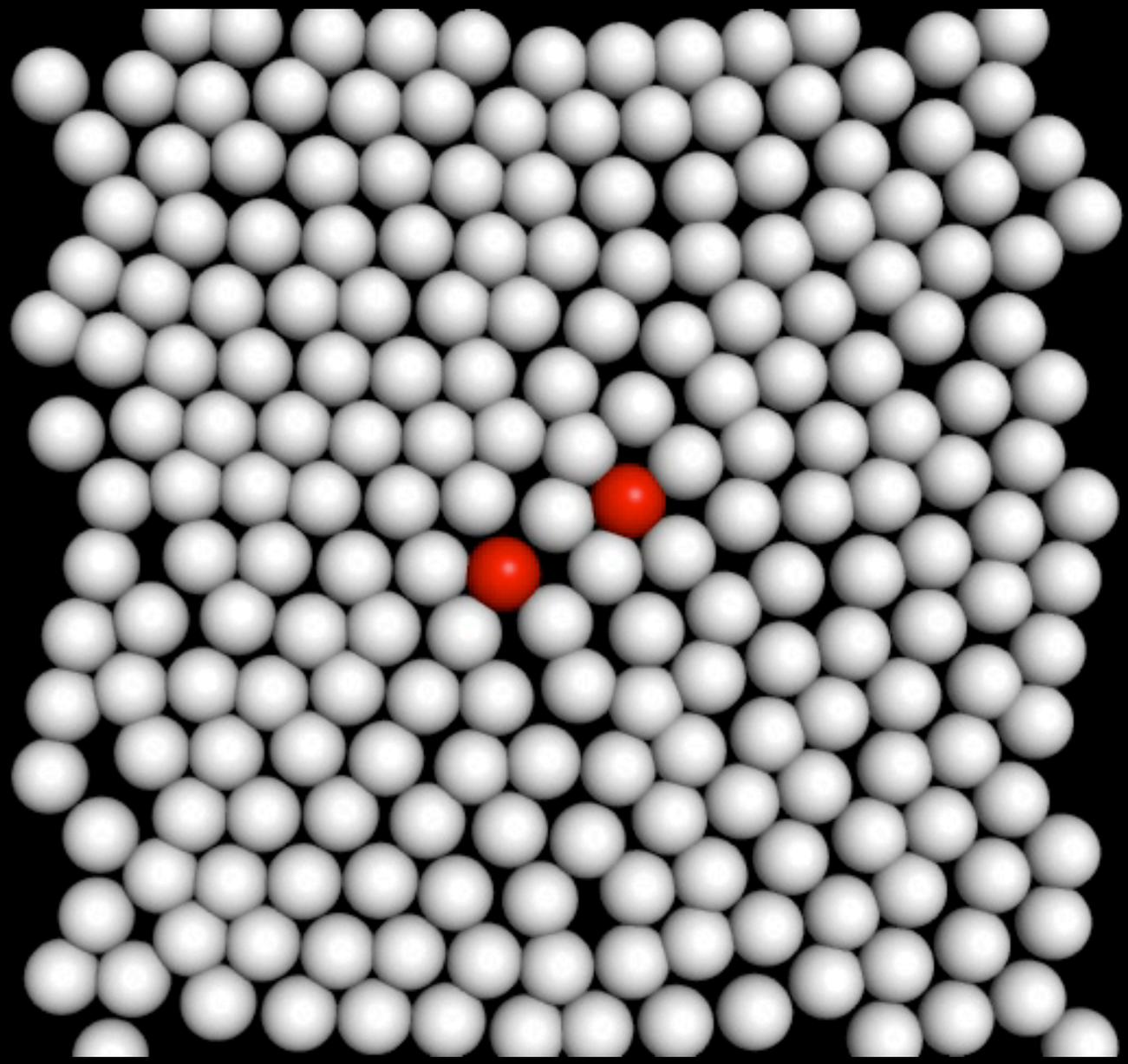
Nonequilibrium MC dimer moves in a 2D WCA fluid



Solvent is “squeezed” out
but significantly perturbed :/

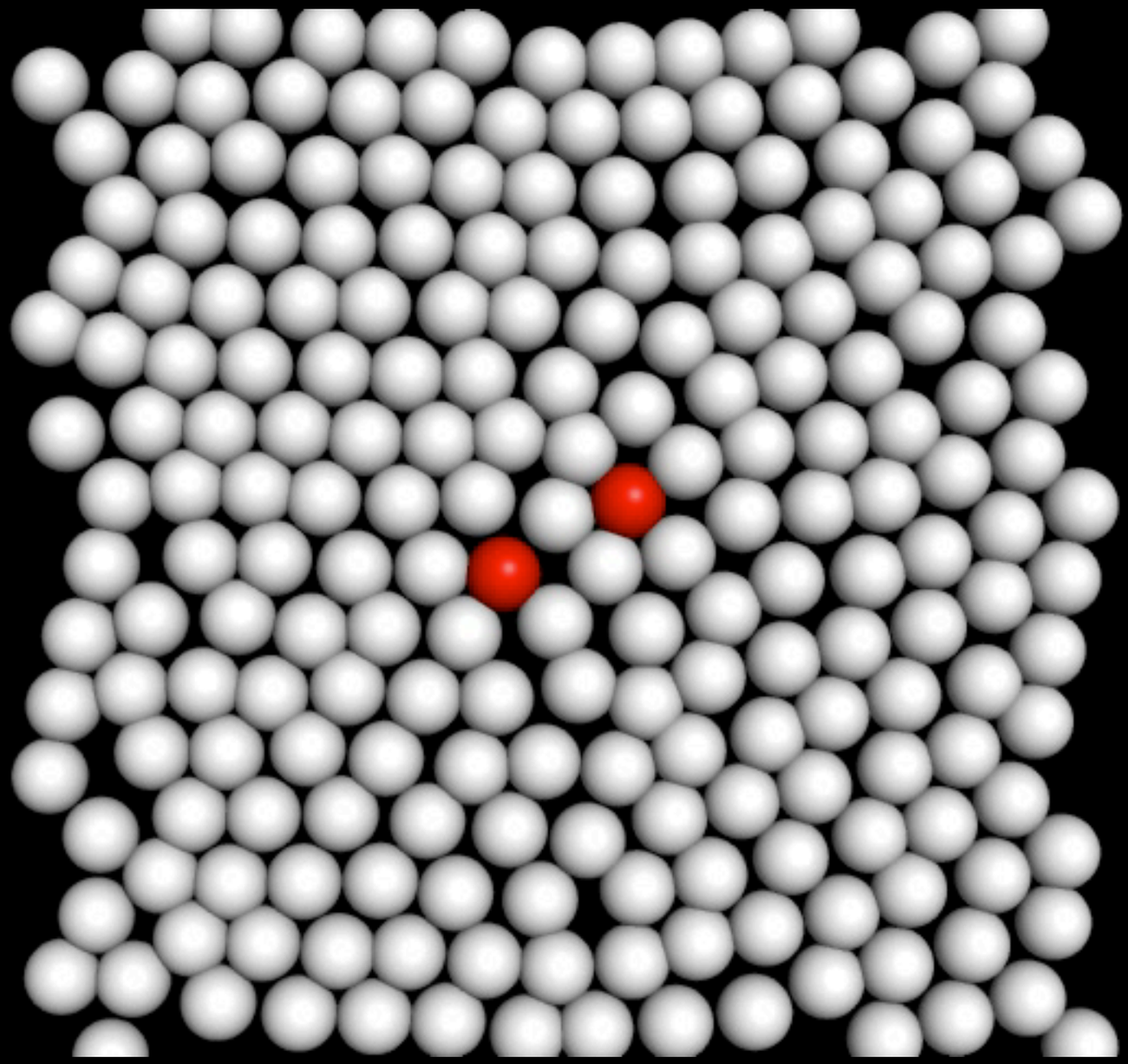
32 switching steps with velocity Verlet dynamics as “propagation kernel”

Nonequilibrium MC dimer moves in a 2D WCA fluid



64 switching steps with velocity Verlet dynamics as “propagation kernel”

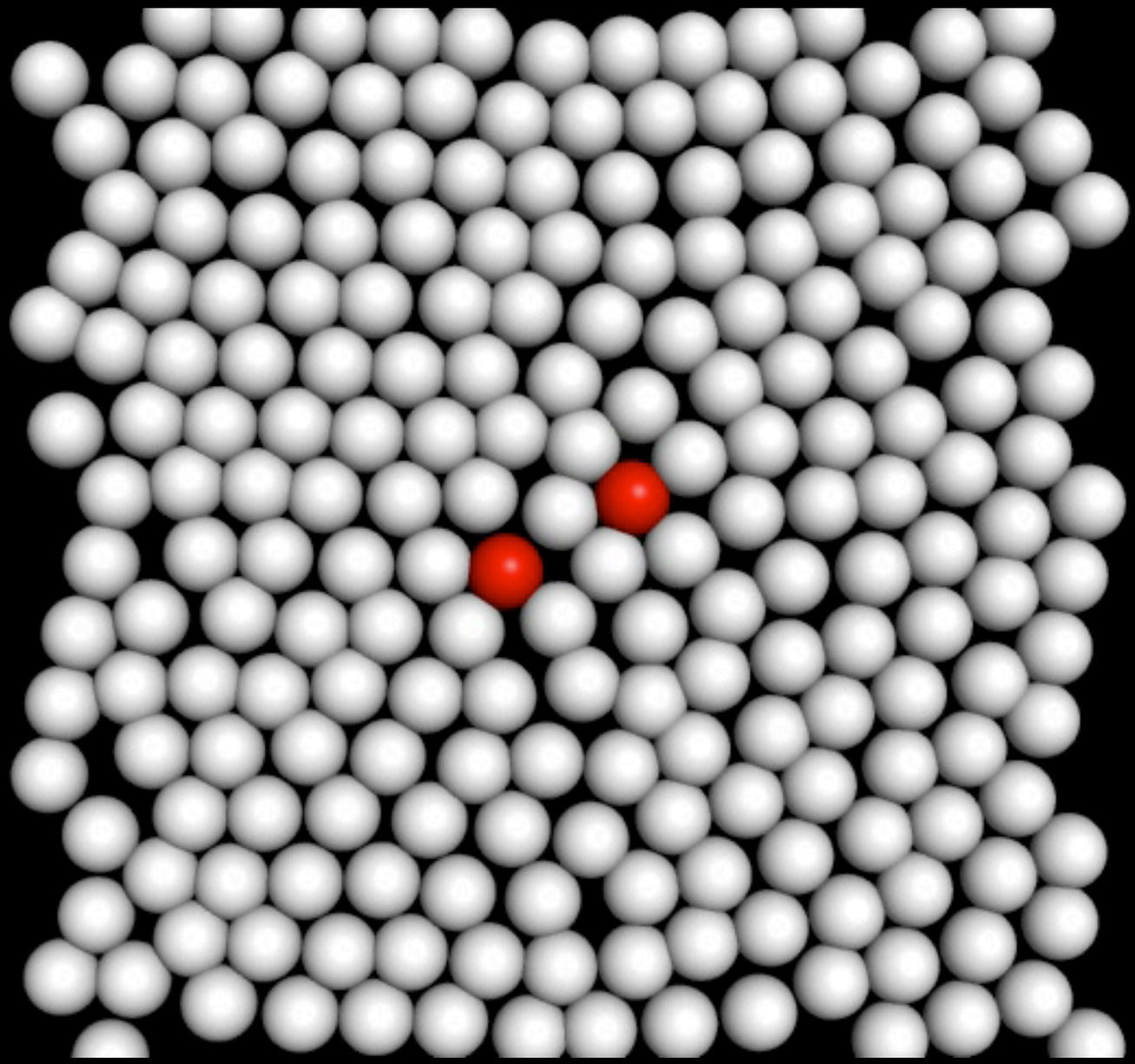
Nonequilibrium MC dimer moves in a 2D WCA fluid



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

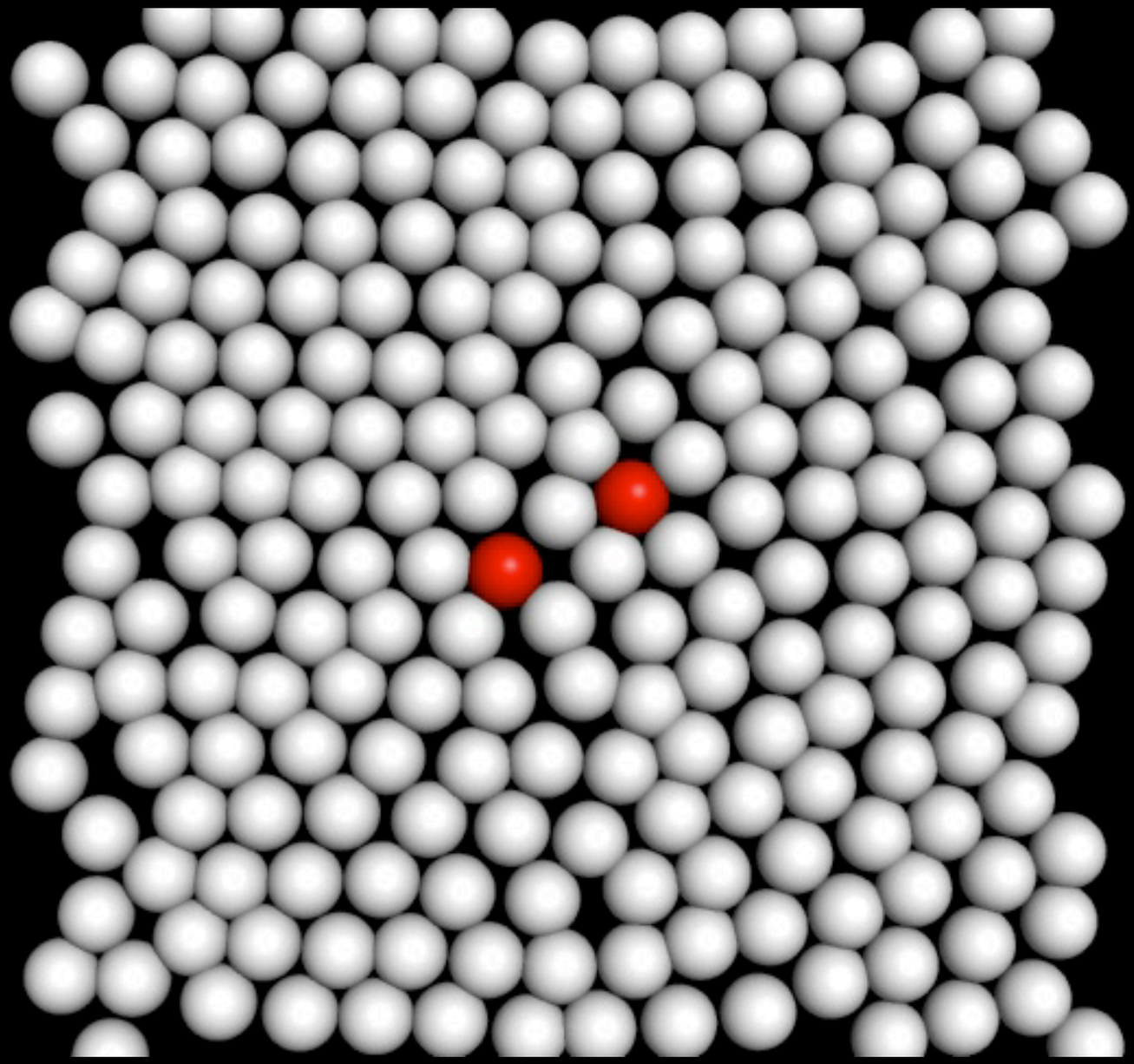
64 switching steps with velocity Verlet dynamics as “propagation kernel”

Nonequilibrium MC dimer moves in a 2D WCA fluid



256 switching steps with velocity Verlet dynamics as “propagation kernel”

Nonequilibrium MC dimer moves in a 2D WCA fluid

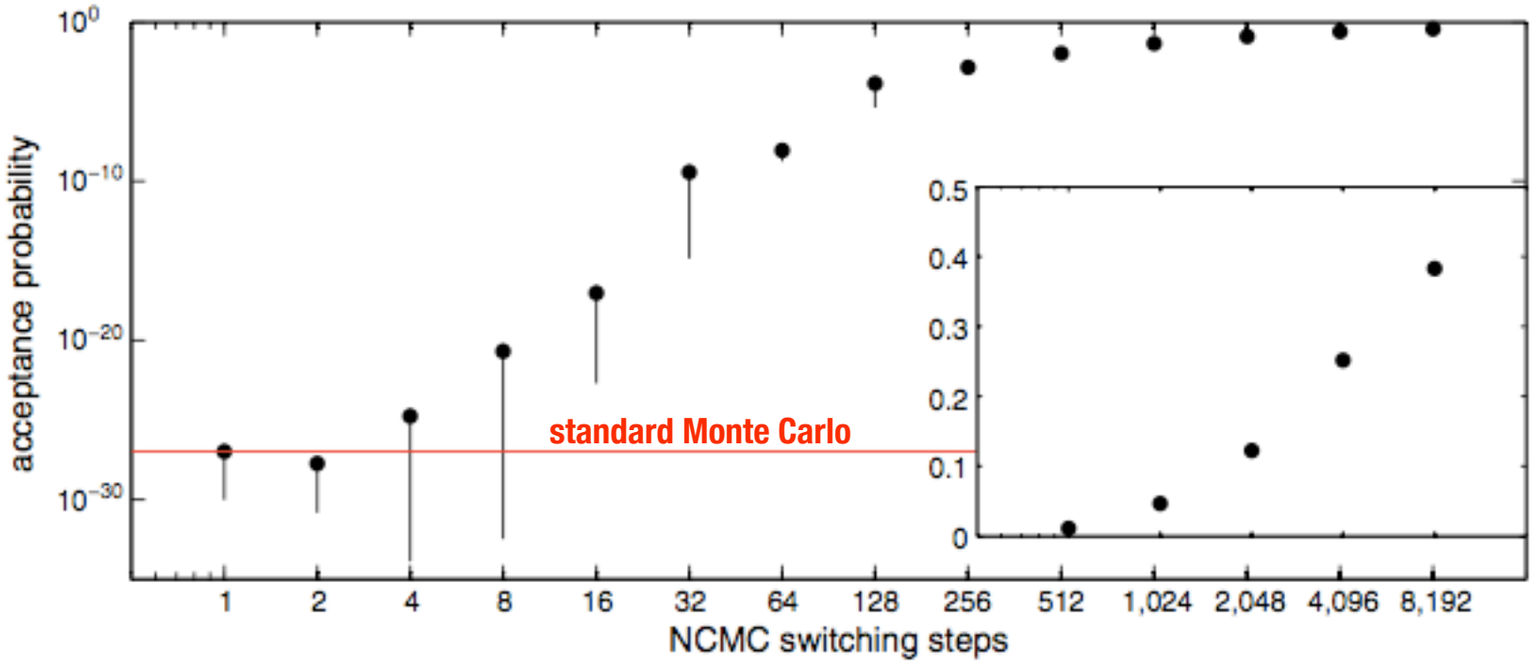


Solvent is “squeezed” out close to reversible limit; high acceptance rates! :D

256 switching steps with velocity Verlet dynamics as “propagation kernel”

Acceptance probability can be astronomically boosted

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



PNAS PLUS

Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation

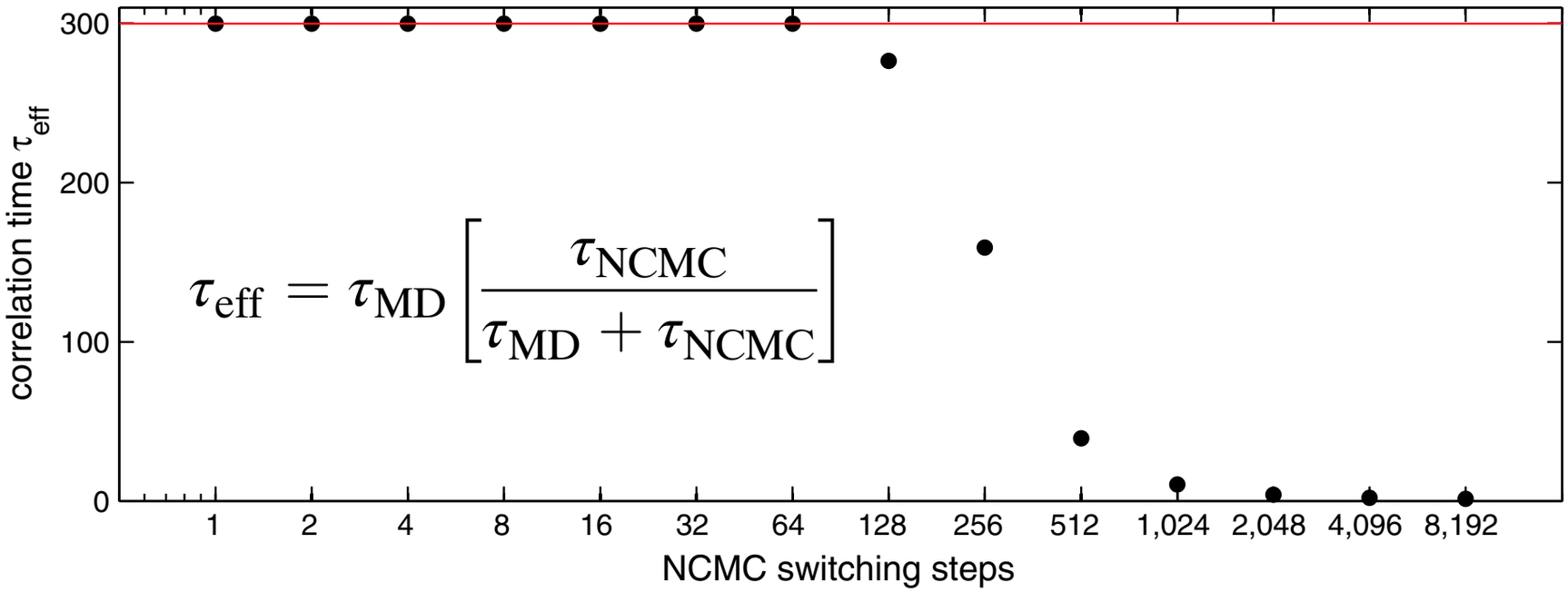
Jerome P. Nilmeier[a], Gavin E. Crooks[b], David D. L. Minh[c], and John D. Chodera[d,1]

[a]Biosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550; [b]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [c]Biosciences Division, Argonne National Laboratory, Argonne, IL 60439; and [d]California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, CA 94720

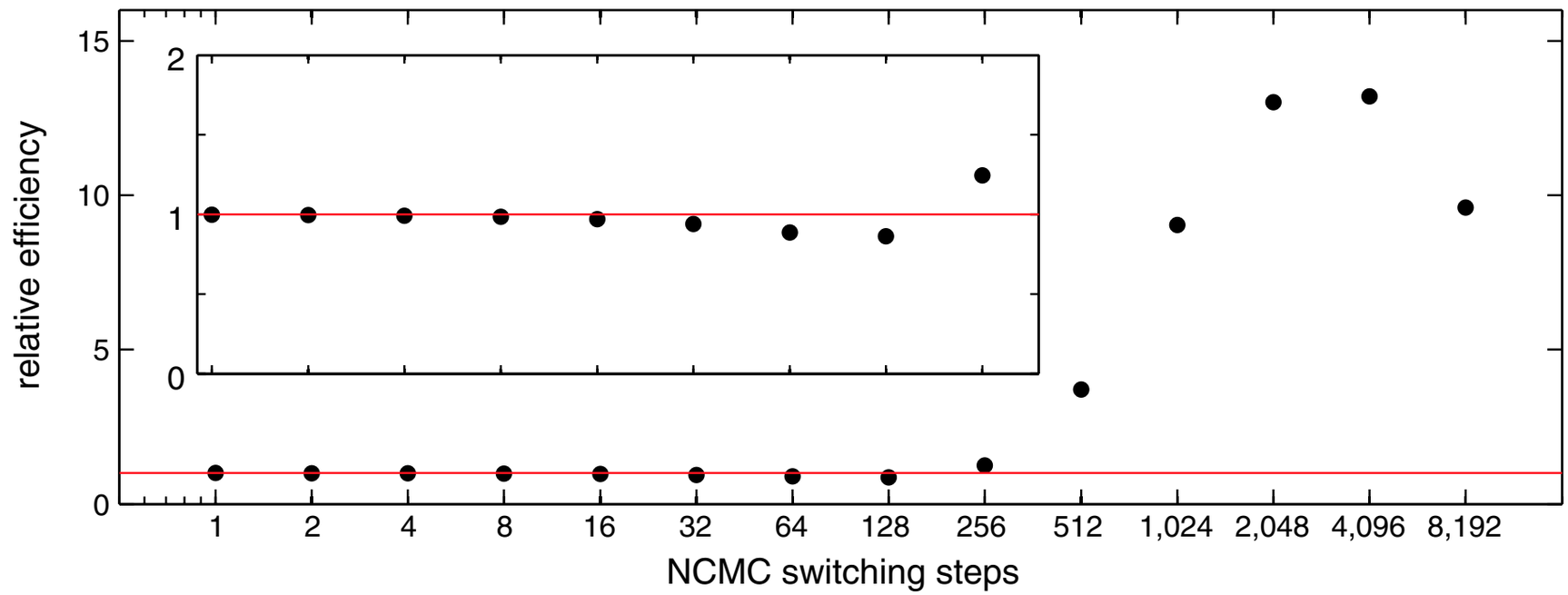
Full 3D system
216 WCA particles
Reduced density $\rho\sigma^3 = 0.96$
Reduced temperature $k_B T/\epsilon = 0.824$
5 kT barrier

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

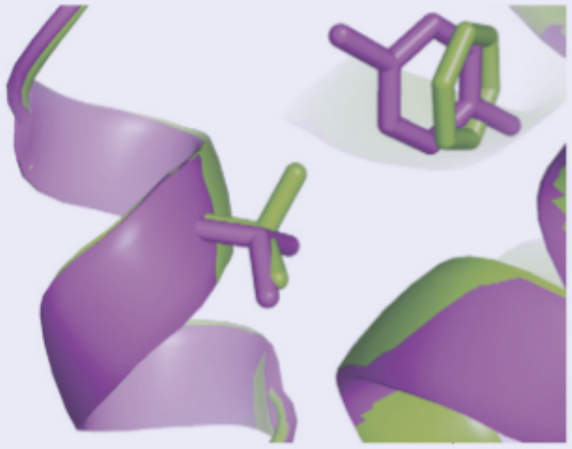
measure of number of iterations to reduce correlation to 1/e



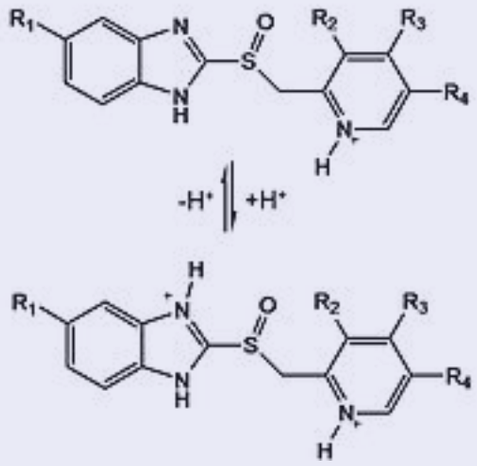
ratio of computer time required to generate one uncorrelated sample without and with NCMC



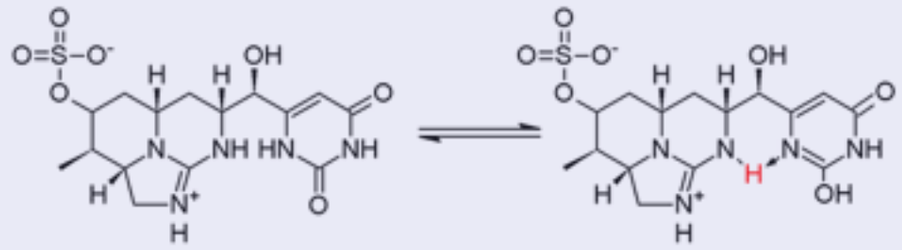
New Monte Carlo techniques open up new possibilities for treating physical effects and facilitating design



sidechain rotamer sampling



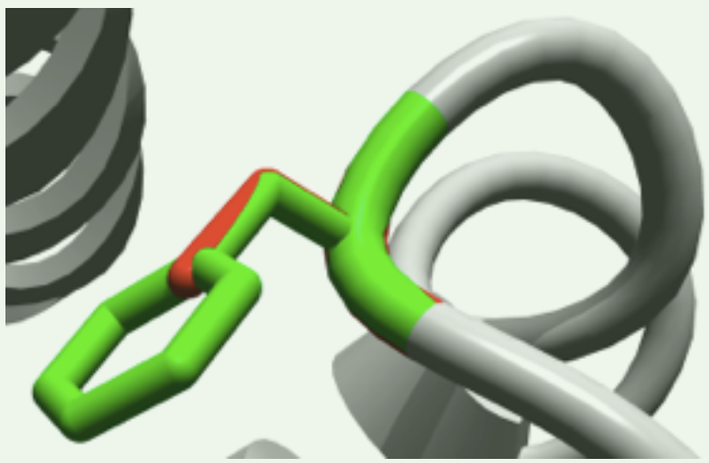
dynamic protonation states (protein and ligand)



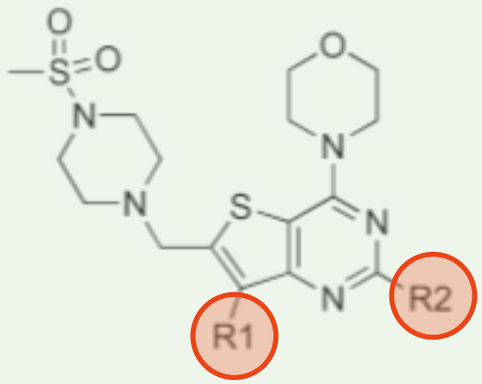
dynamic tautomerization (ligand)

ENHANCING CHEMICAL ACCURACY

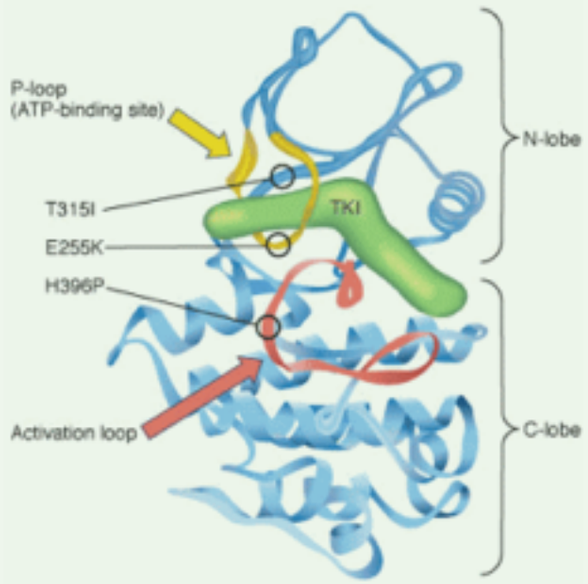
FACILITATING DESIGN



sampling over target mutations

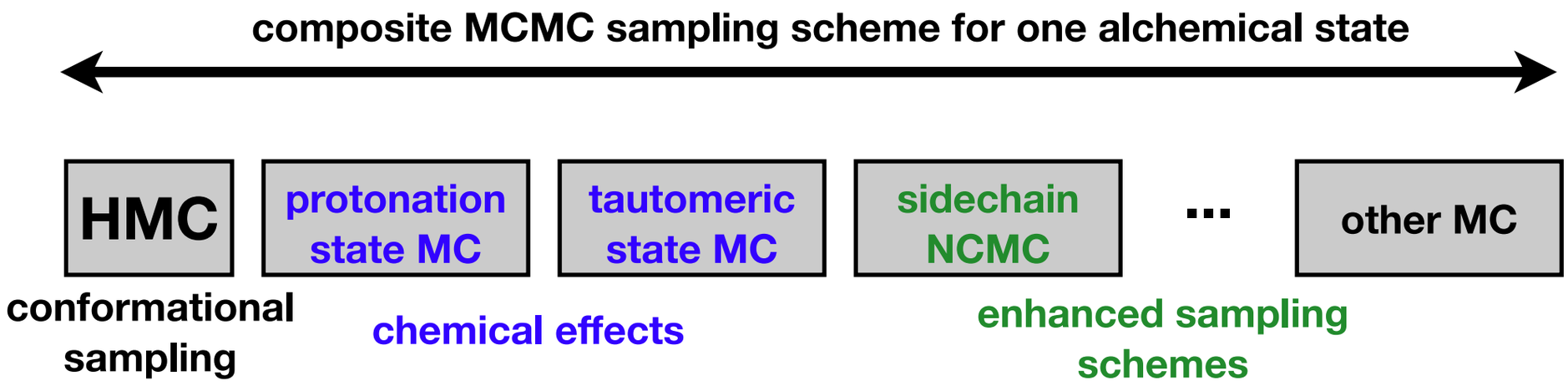


sampling over substituents for ligand design

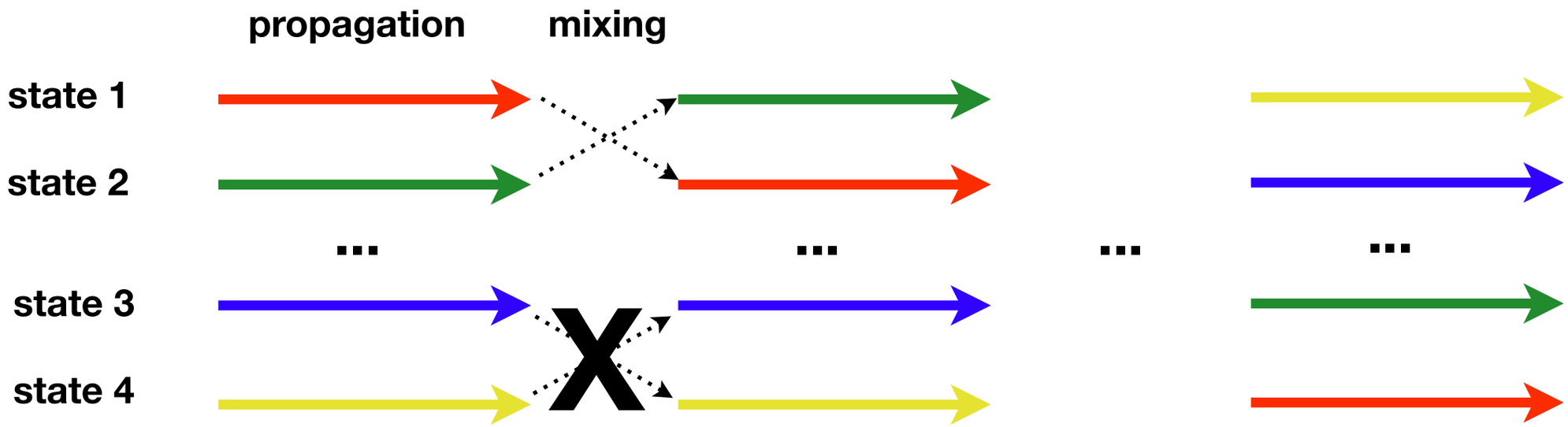


ligand design considering potential resistance mutations

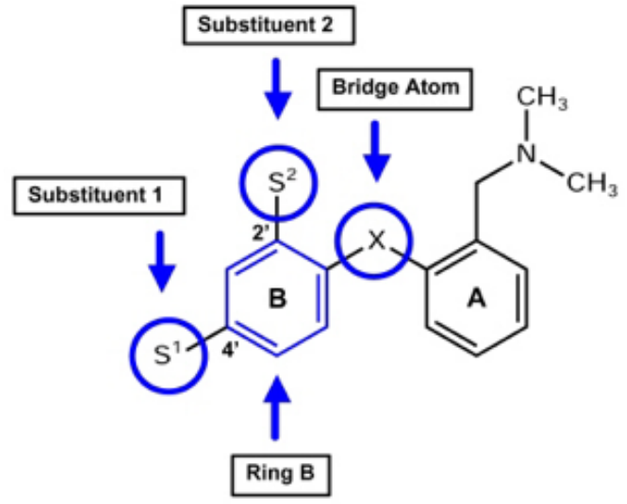
Markov chain Monte Carlo framework allows flexible inclusion of enhanced sampling schemes and chemical effects



Can be combined with replica exchange schemes to decrease correlation times



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



Expanded ensemble:

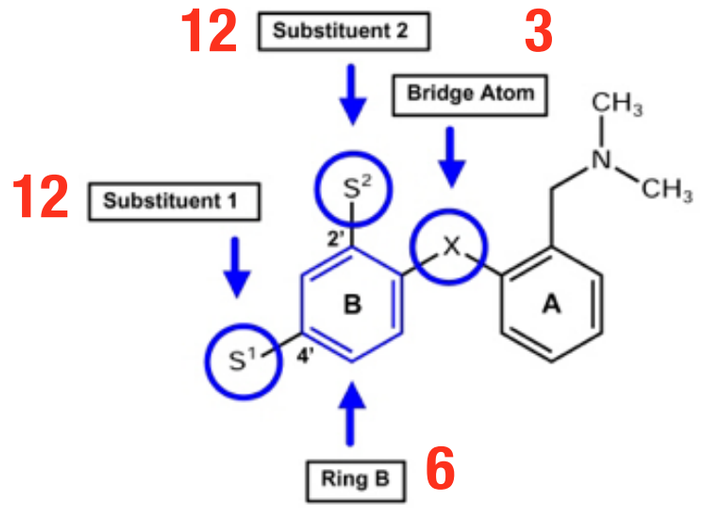
$$p(\mathbf{x}, k) = Z^{-1} \exp[-u_k(\mathbf{x}) + g_k]$$

$\mathbf{x}$ configuration

k chemical species

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

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



12 · 12 · 6 · 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

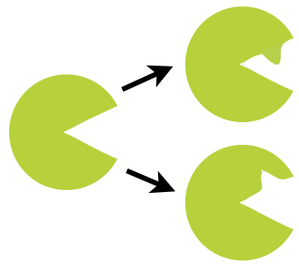
$$\pi(x)K(x,y) = \pi(y)K(y,x)$$

$$AH \leftrightharpoons A^- + H^+$$

modular MCMC for sampling and chemical effects

enhanced sampling algorithms

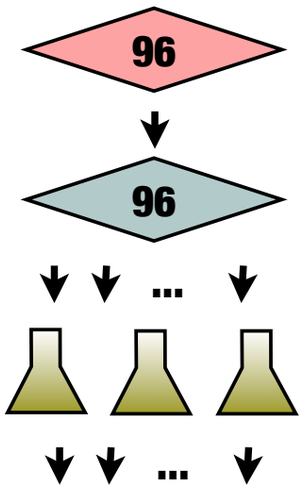
Test and improve models quickly and cheaply by inverting the drug discovery problem



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

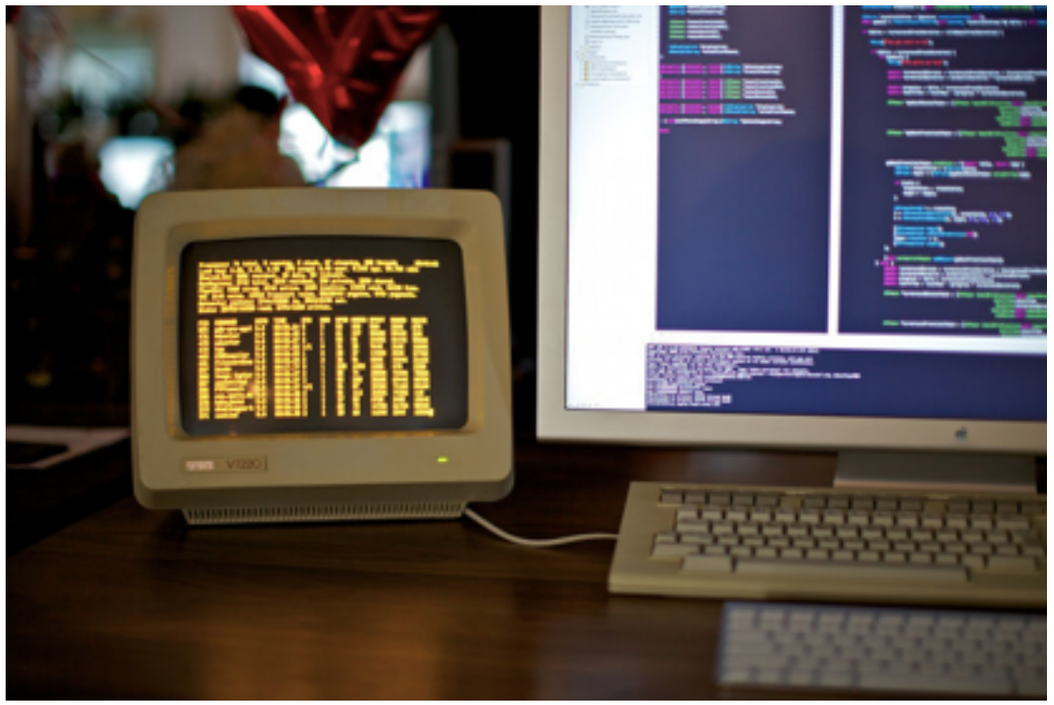
How can we make wetlab experiments look more like problems we know how to solve efficiently?

messy
laborious
inconsistent
skill-dependent
9 am - 5 pm

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

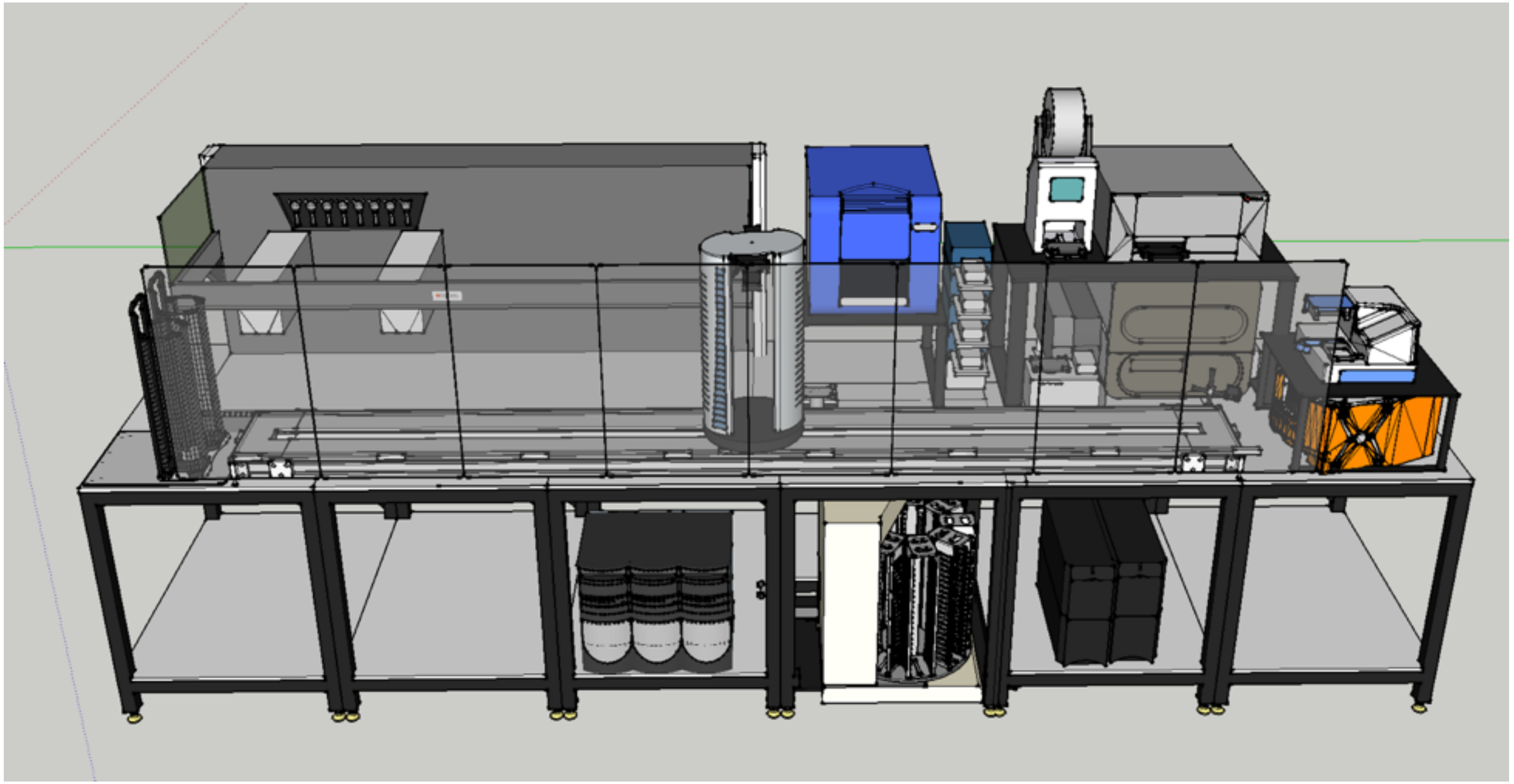
messy
laborious
inconsistent
skill-dependent
9 am - 5 pm

?



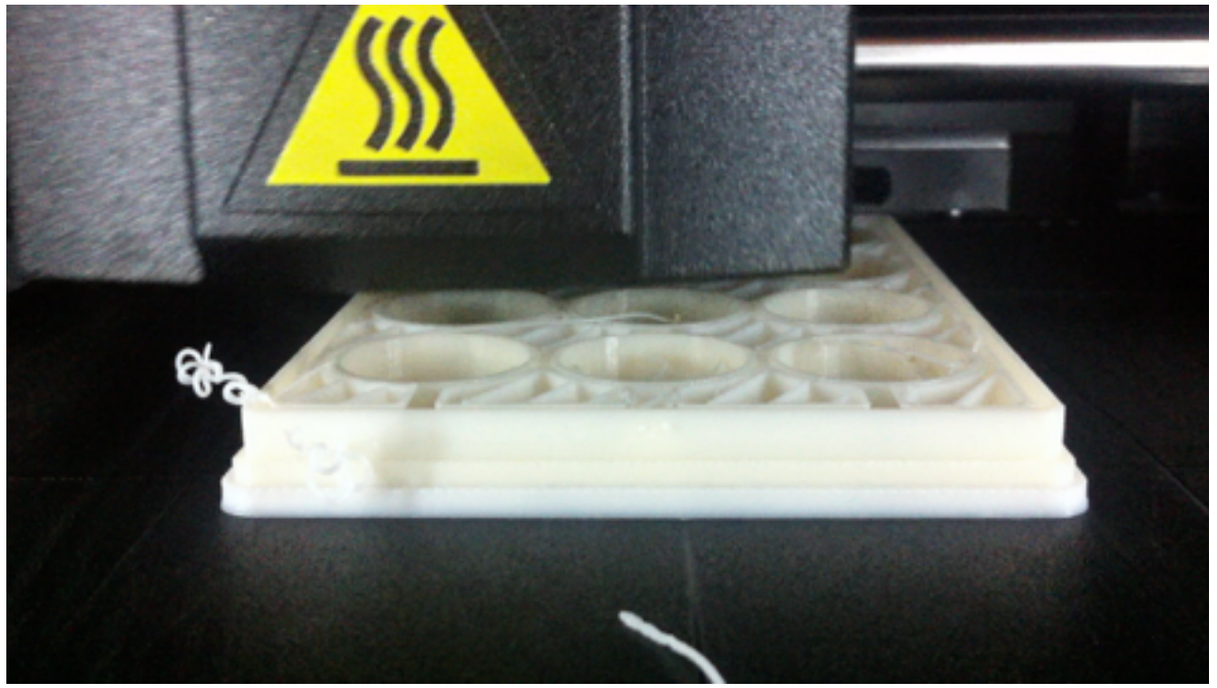
precise
structured
consistent
reproducible
round-the-clock

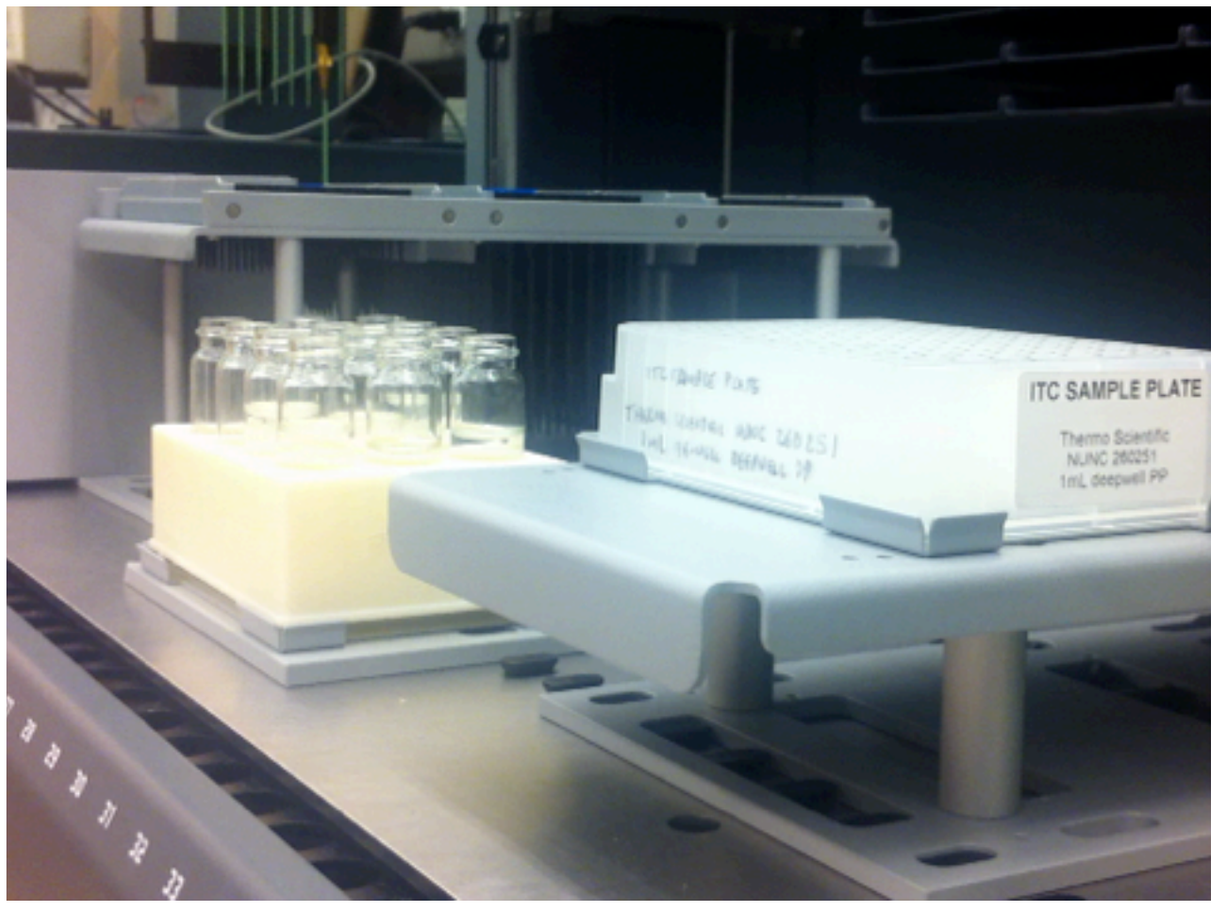
AUTOMATE. EVERYTHING.



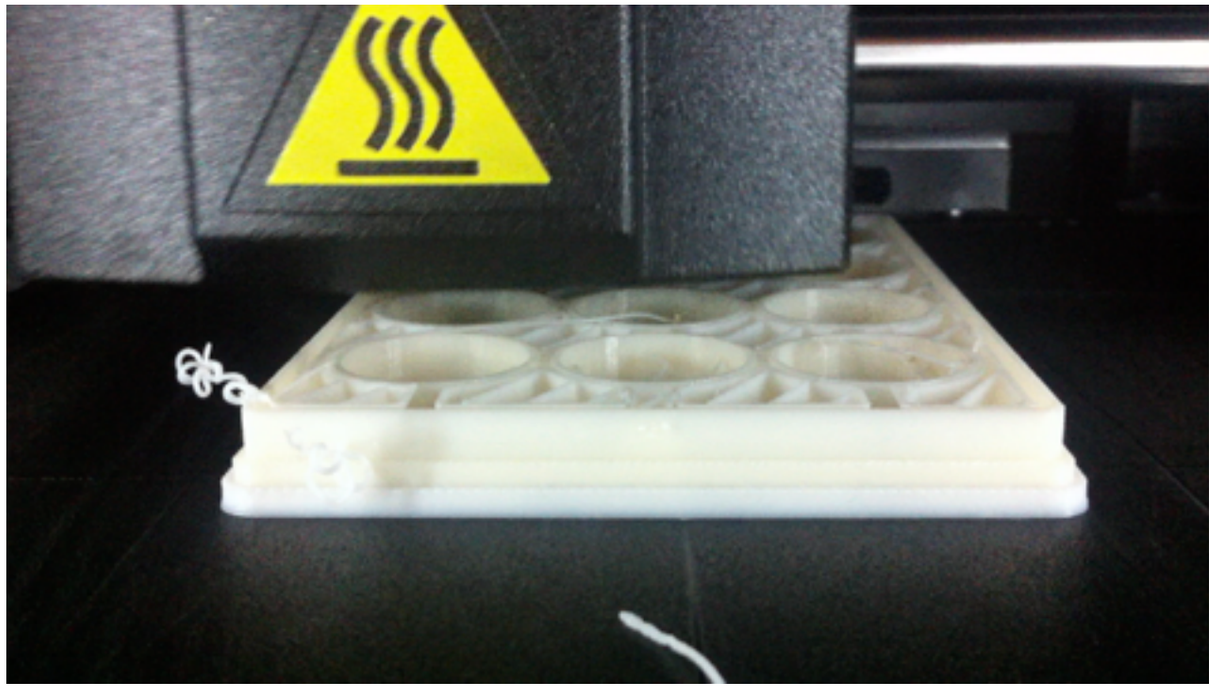
Automated platform for bacterial cloning, mutagenesis, expression, purification, and binding affinity measurement with 24/7 operational capacity

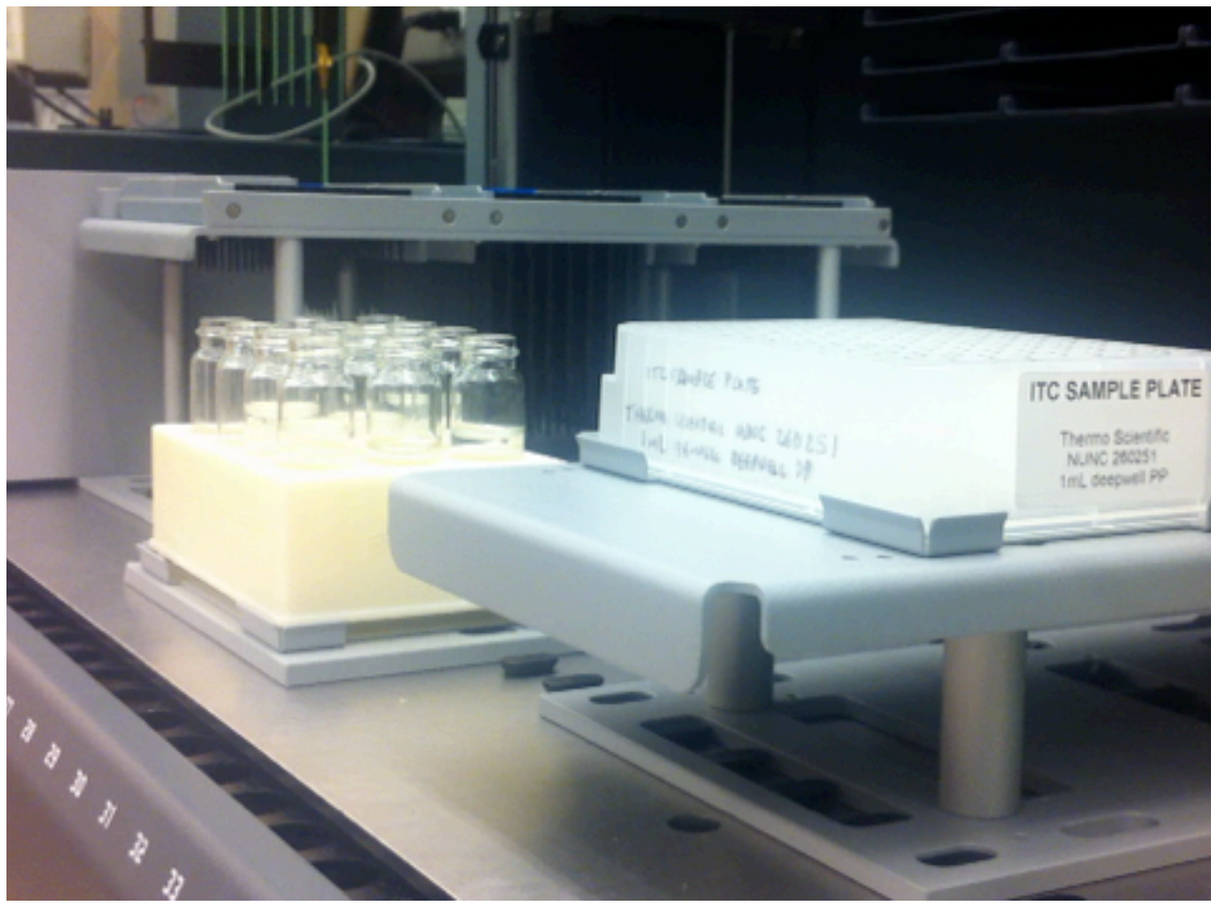
New technologies facilitate rapid progress





New technologies facilitate rapid progress





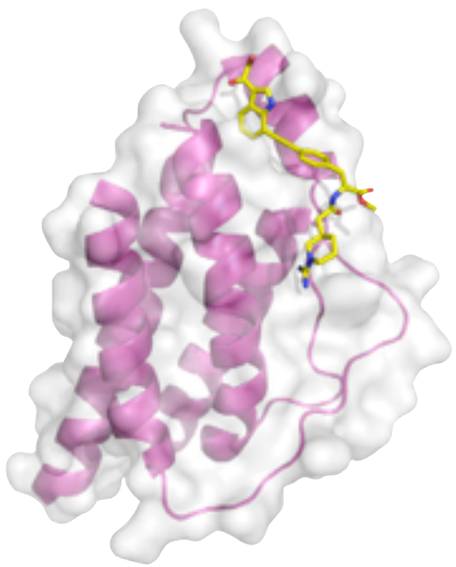
We're building a benchmark set covering a range of complexity, isolating individual challenges to modeling

small apolar ligands
T4 lysozyme L99A

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

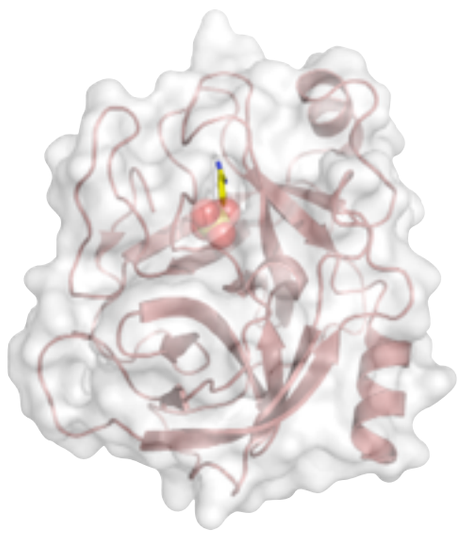
polar ligands
FKBP12

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?
$MgCl_2$ salt effects?

easy
hard

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

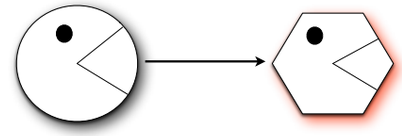
How sensitive are binding free energies to various physical effects?

simulation details

experimental details

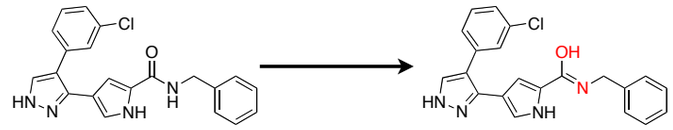
Protein conformation

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



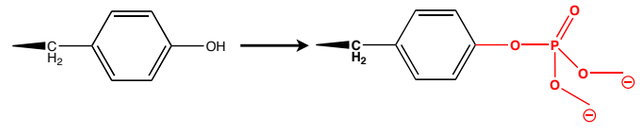
Ligand protonation/tautomeric state

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



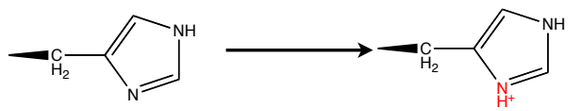
Phosphorylation state

- Conformational change upon phosphorylation
- Change in electrostatic environment



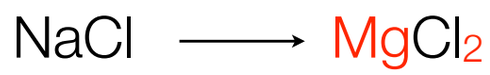
Protein protonation state

- 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



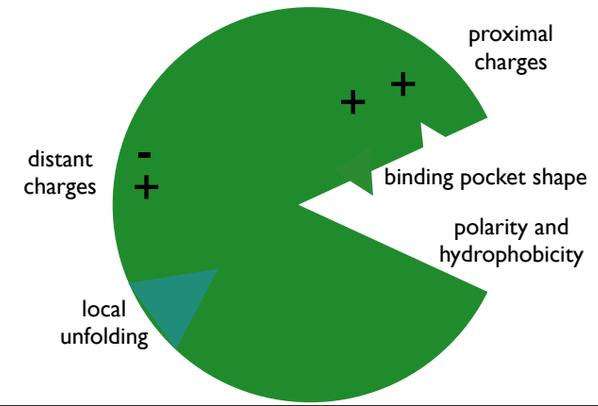
Proximal/distal charged residues

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.

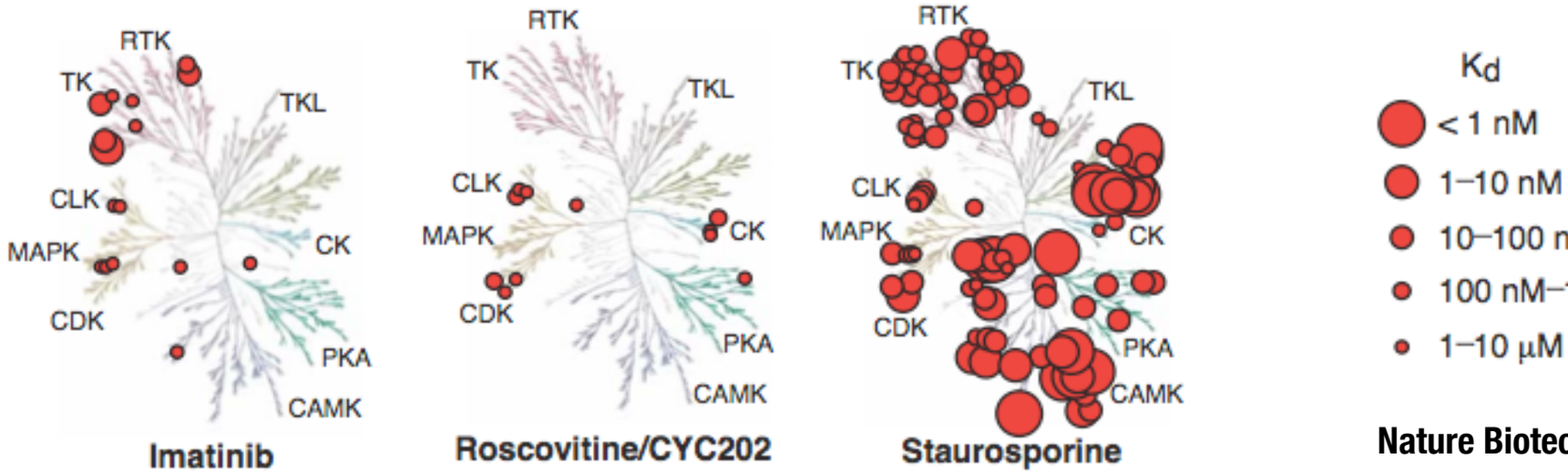
* Improvement of algorithms to sample conformational changes.

* Inclusion of chemical effects using nonequilibrium Monte Carlo methods.

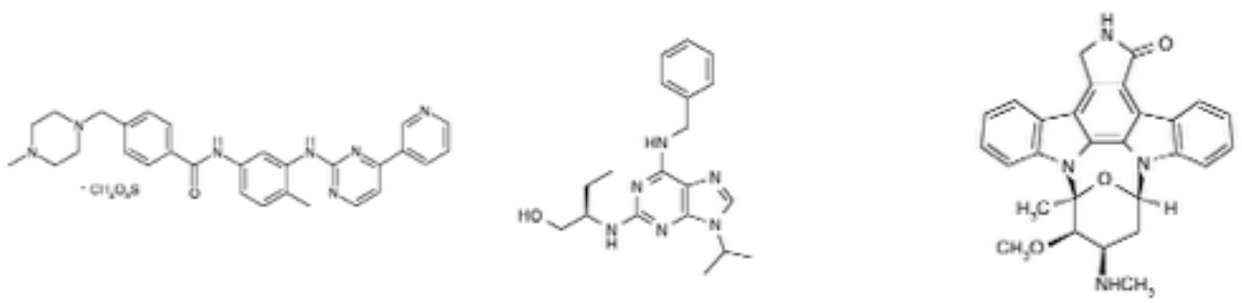
* Automated Bayesian methods to improve forcefields using data.

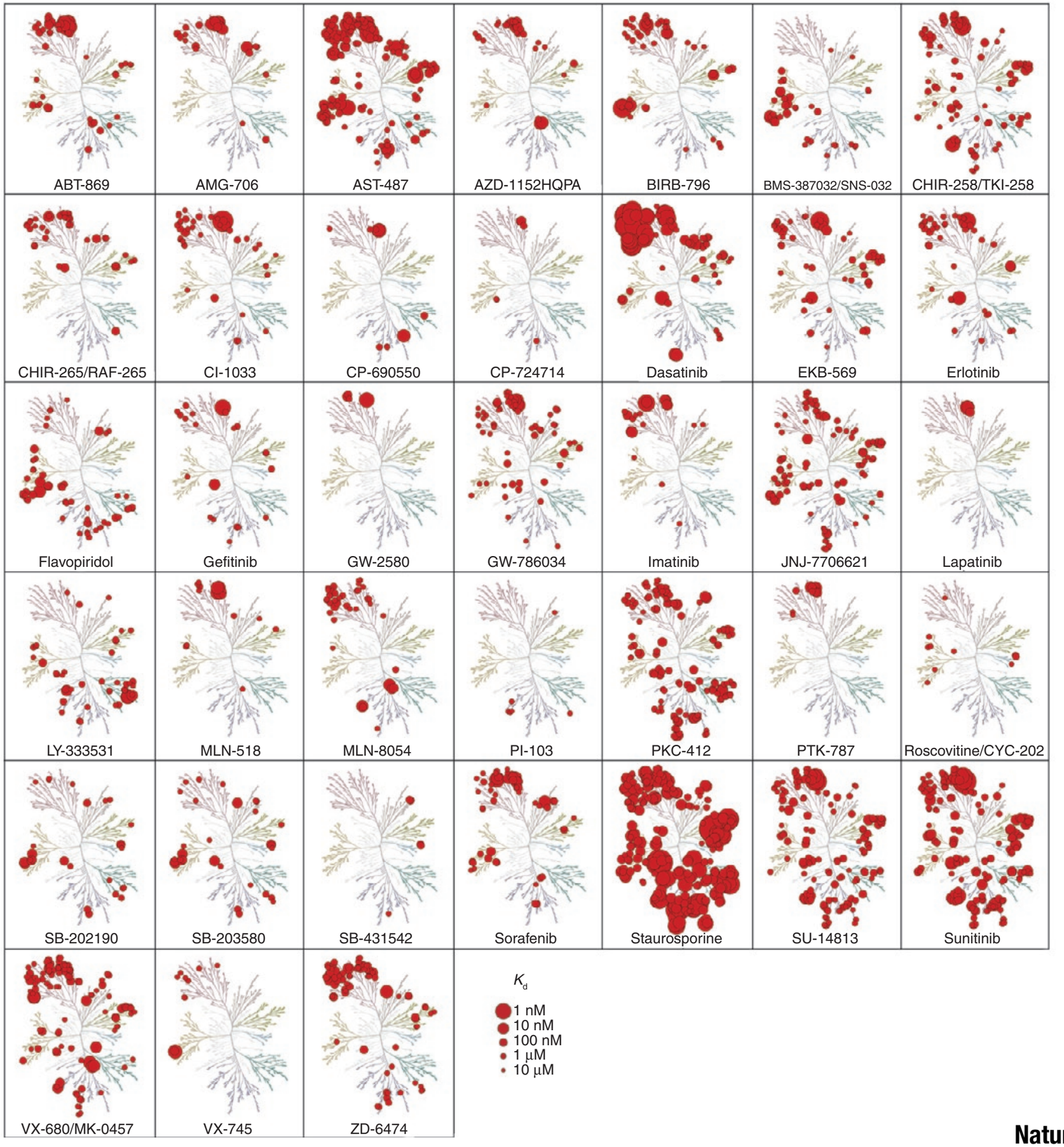
Tackling biological questions (and getting grants funded)

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



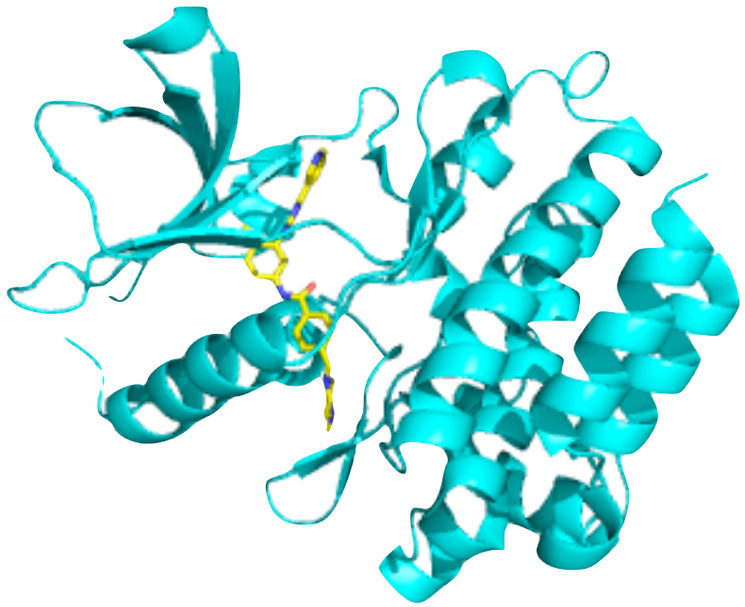
Nature Biotech 23:329, 2005





Nature Biotech 26:127, 2008

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



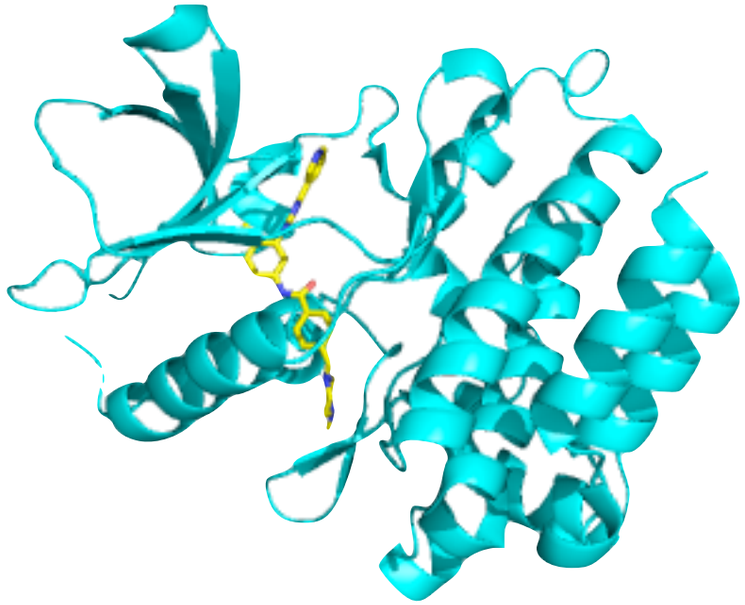
imatinib bound to Abl kinase [2HYY]

imatinib bound to Src kinase [2OIQ]

$$\Delta\Delta G = 4.6 \ \mathrm{kcal/mol}$$ **(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]

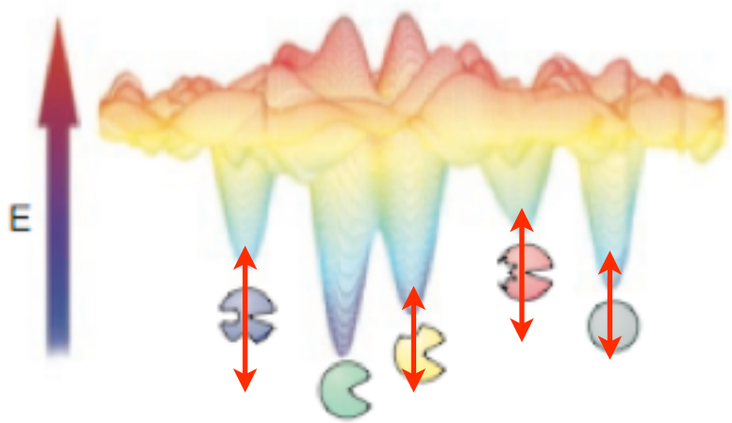
Differences in stabilities of inactive states may be responsible for origin of some kinase inhibitor selectivity



imatinib bound to Abl kinase [2HYY]

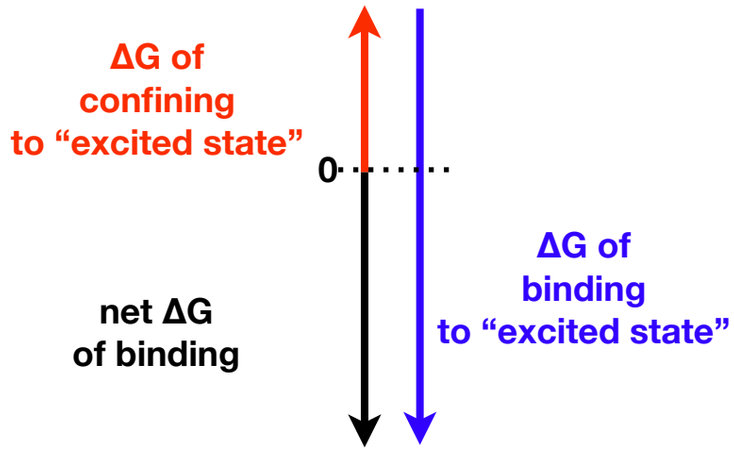
imatinib bound to Src kinase [2OIQ]

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



inactive “excited states”

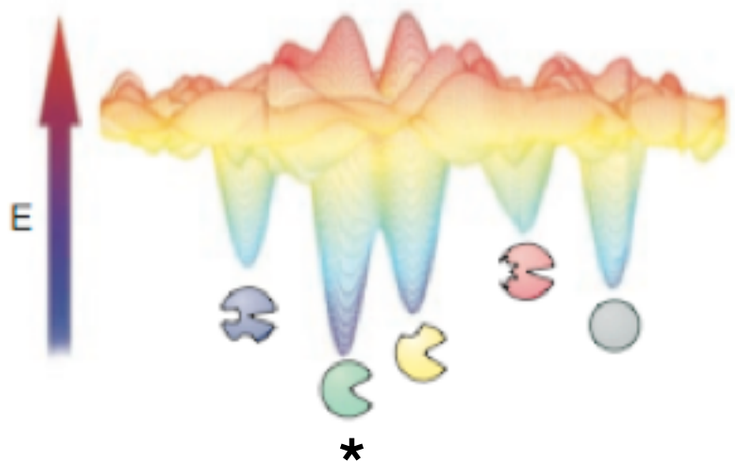
active “ground state”



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



invisible “excited states”

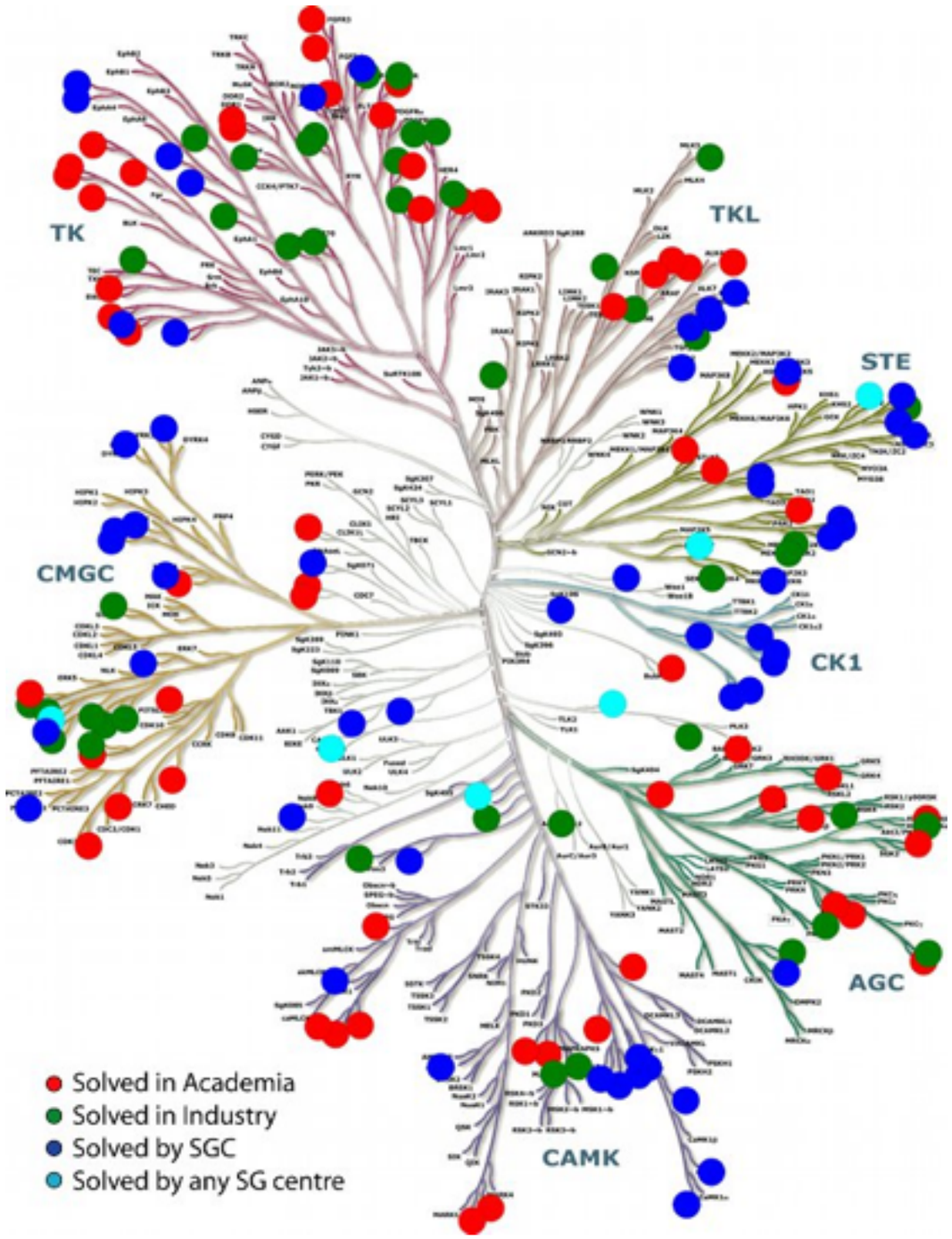
visible “ground state”

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

Lee and Craik. Science 324:213, 2009

Structural data on human kinases exists, but is incomplete

Human kinases with available structural data



http://www.sgc.ox.ac.uk/research/kinases/

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

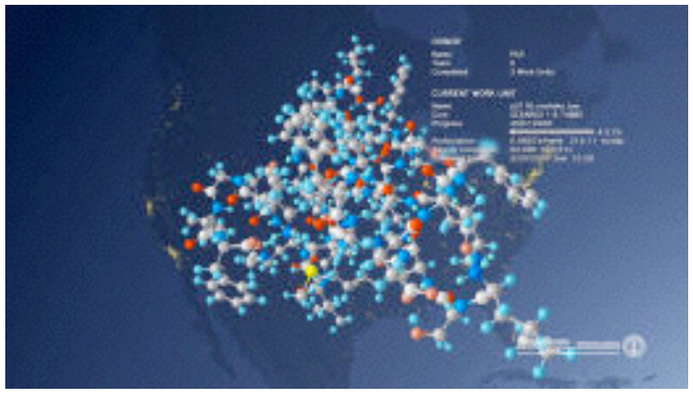
Folding@home DISTRIBUTED COMPUTING

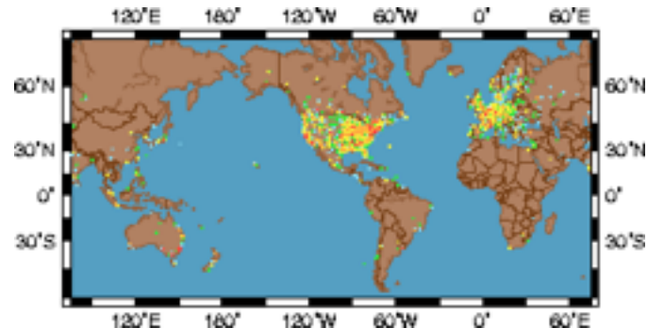
Vijay S. Pande
Stanford University

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

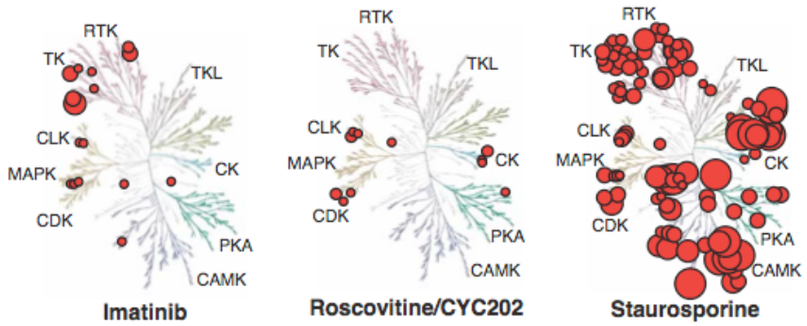
Table last updated at Mon, 20 May 2013 06:14:30



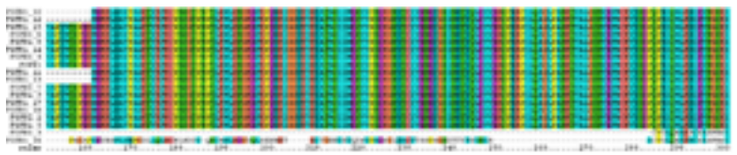


Over 14 PFLOP/s of aggregate computational power!

http://folding.stanford.edu


RTK
TK
TKL
CLK
MAPK
CK
CDK
PKA
CAMK
Imatinib
RTK
TK
TKL
CLK
MAPK
CK
CDK
PKA
CAMK
Roscovitine/CYC202
RTK
TK
TKL
CLK
MAPK
CK
CDK
PKA
CAMK
Staurosporine

sequences of targets/off-targets

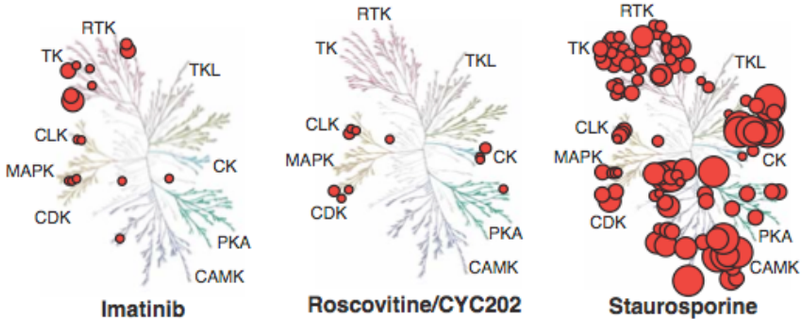


+

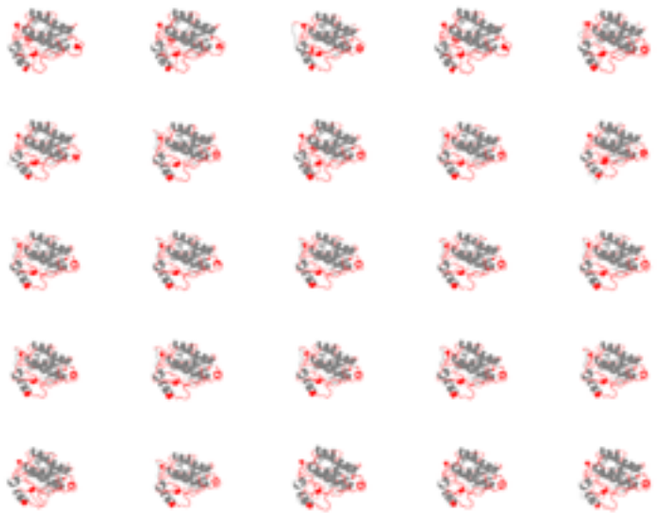
UniProt

RCSB PDB
PROTEIN DATA BANK

template structures from Uniprot



structural models of many conformations



conformational ensembles
with relative energetics

multiple binding free energy calculations
to compute affinities and selectivities

Folding@Home enables whole-kinome simulation

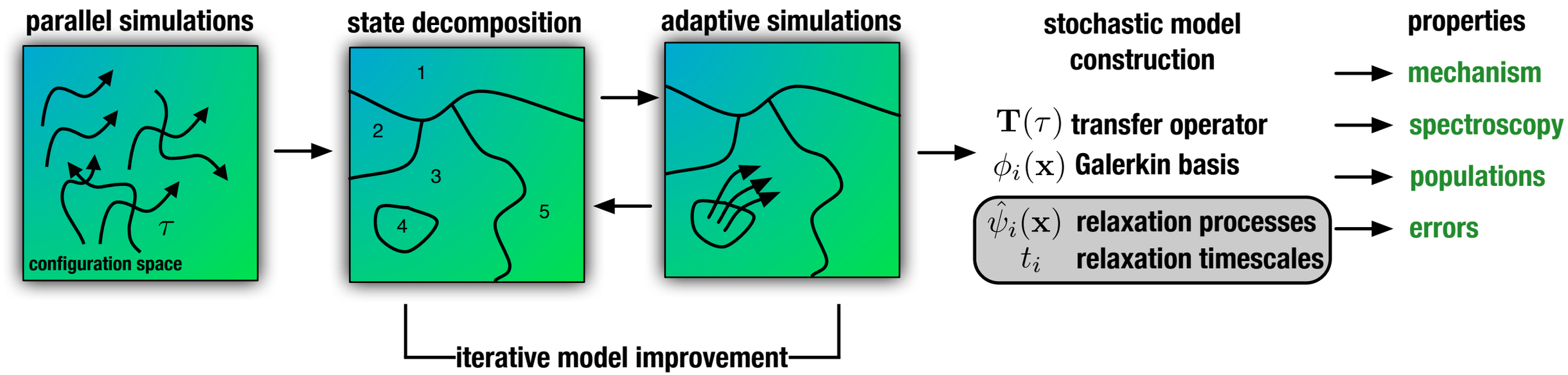
518 **human protein kinases**
excluding splice and disease variants

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

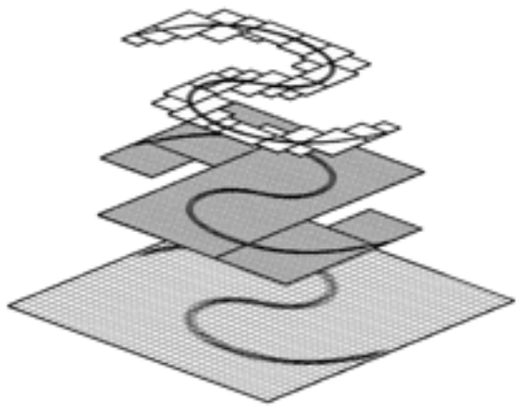
= **1,816,626** **kinase models will be built and refined**
on new MSKCC compute resources housed at SDSC

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

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



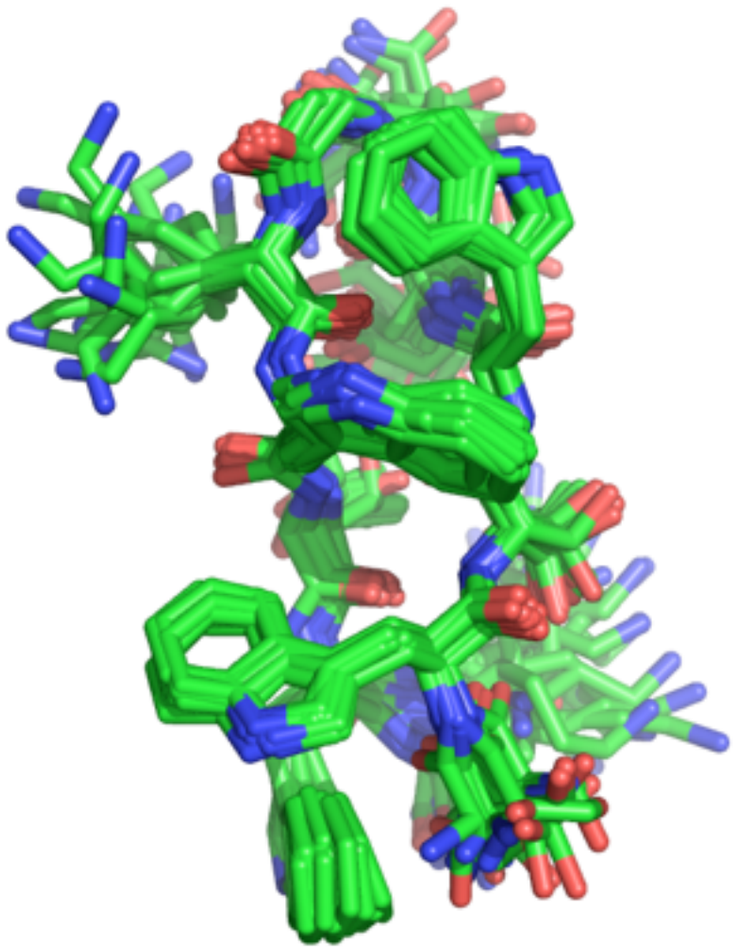
Similar in spirit to adaptive mesh refinement algorithms in engineering, very different from traditional approach of running a single long simulation.



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.

with important contributions from Schütte, Noé, Weber, Hummer, Roux, Vanden-Eijnden

NMR model of trpzip2 at 288 K



trpzip2
[PDB:1LE1]

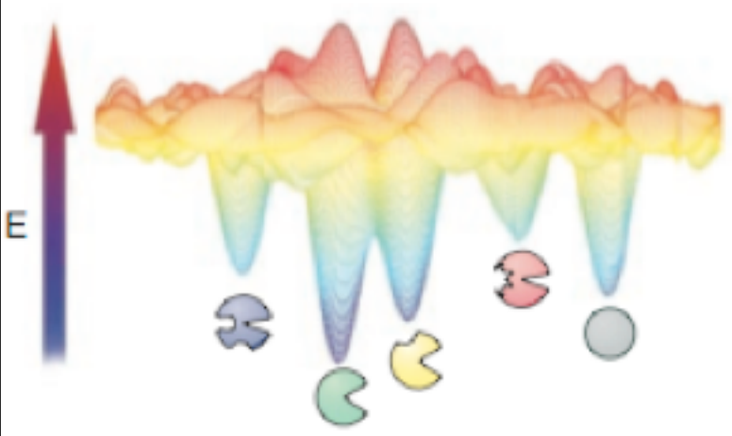
Many distinct metastable states can be identified at T ~ T_m

backbone
Trp sidechains

"invisible" states (<1%)

"ground states"

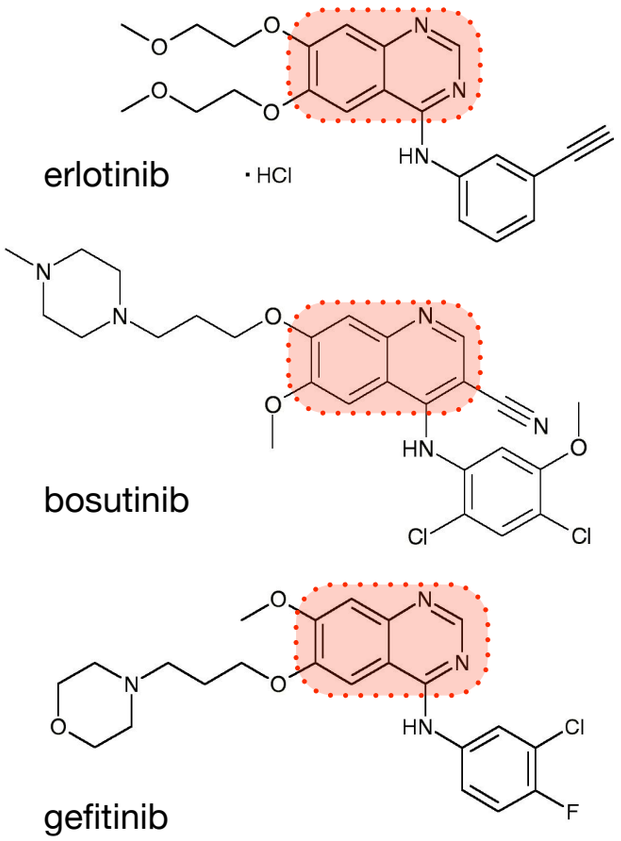
8.1 kT	8.2 kT	8.8 kT	10.0 kT		
7.3 kT	7.3 kT	7.5 kT	7.7 kT	7.7 kT	7.9 kT
6.7 kT	6.8 kT	6.9 kT	6.9 kT	7.2 kT	7.3 kT
6.5 kT	6.5 kT	6.5 kT	6.6 kT	6.7 kT	6.7 kT
6.2 kT	6.2 kT	6.3 kT	6.3 kT	6.4 kT	6.4 kT
4.8 kT	5.6 kT	5.8 kT	5.9 kT	5.9 kT	6.0 kT
0.0 kT	0.2 kT	2.5 kT	3.0 kT	3.9 kT	4.7 kT

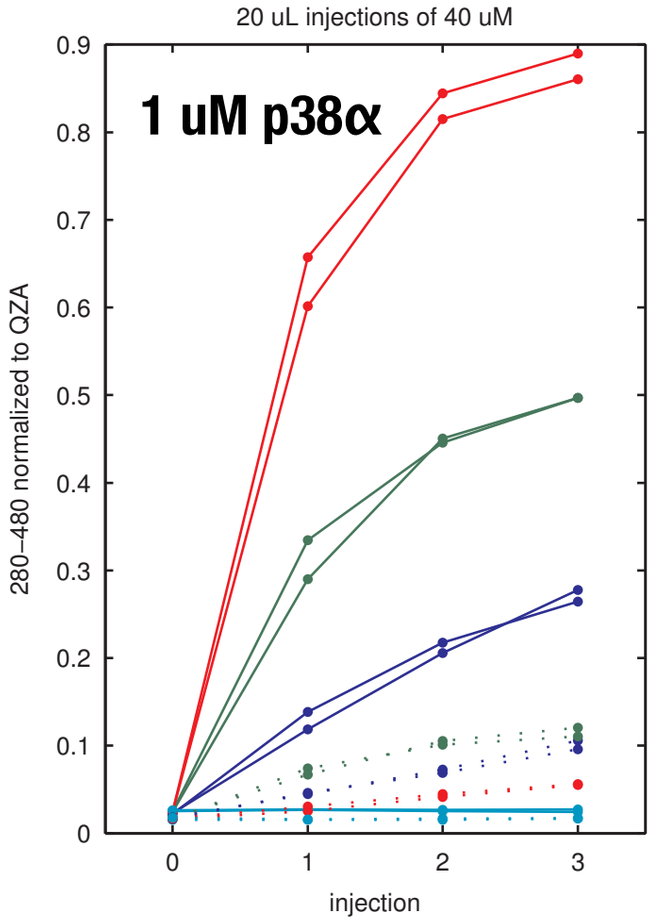


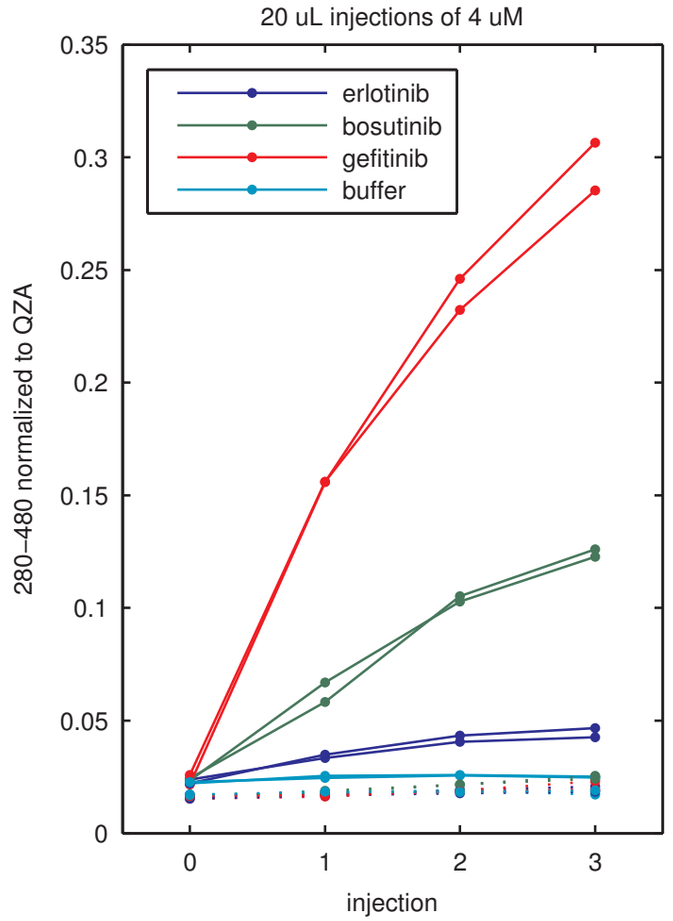
Chodera, Singhal, Swope, Pande, Dill. JCP 126:155101, 2007.

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

fluorescent inhibitors



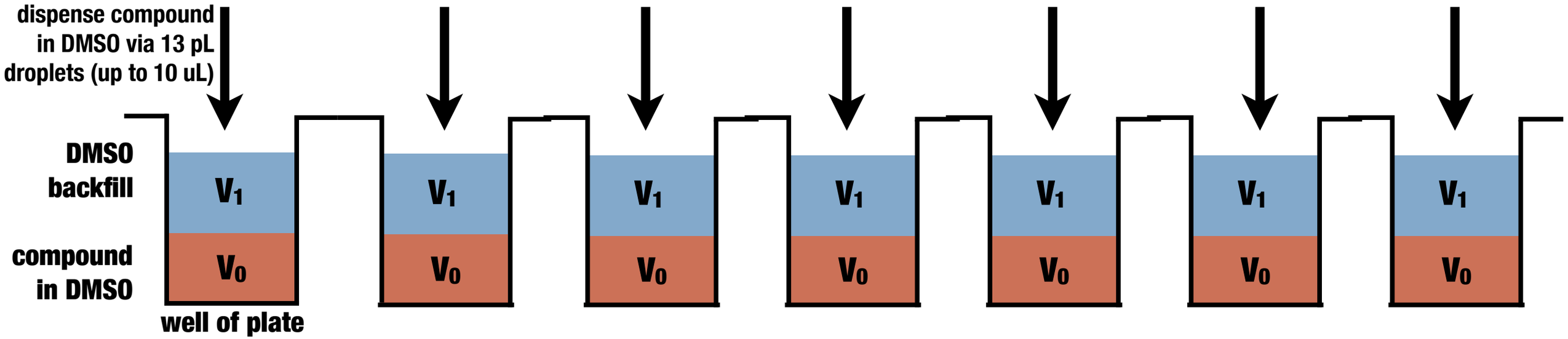




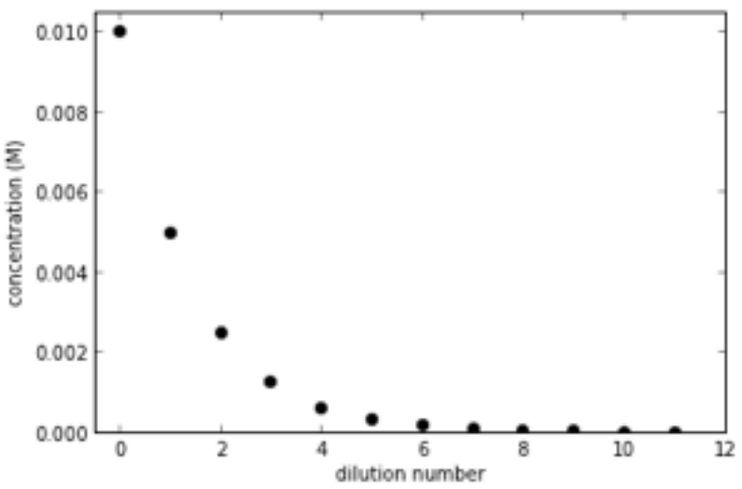
excite 280 nm
[Trp FRET ~350 nm]
measure 480 nm

with Nick Levinson, Boxer lab, Stanford

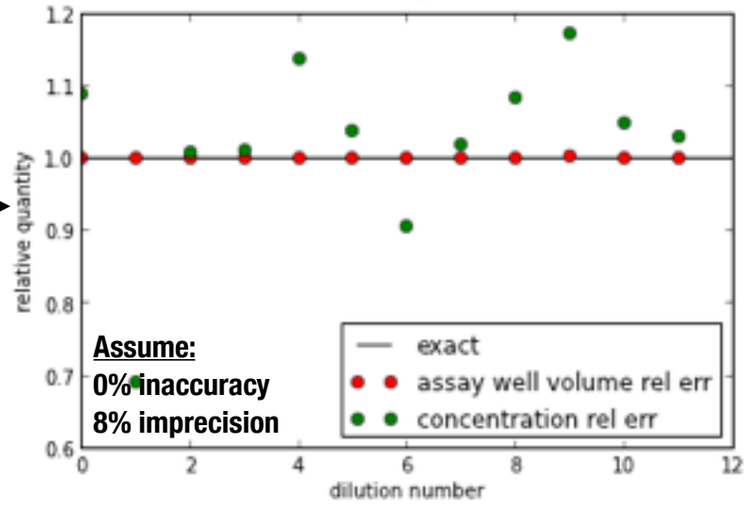
Modeling of biophysical experiments can be a helpful tool for designing protocols for optimal information content



ideal concentrations

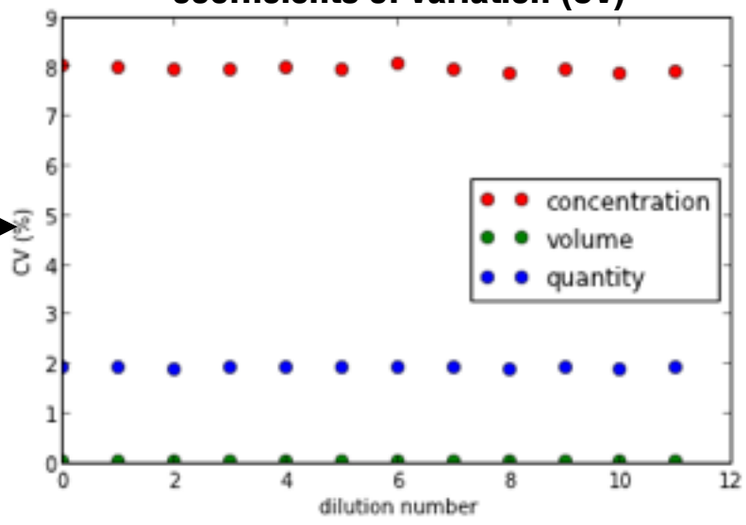


error for a single realization



Assume:
0% inaccuracy
8% imprecision

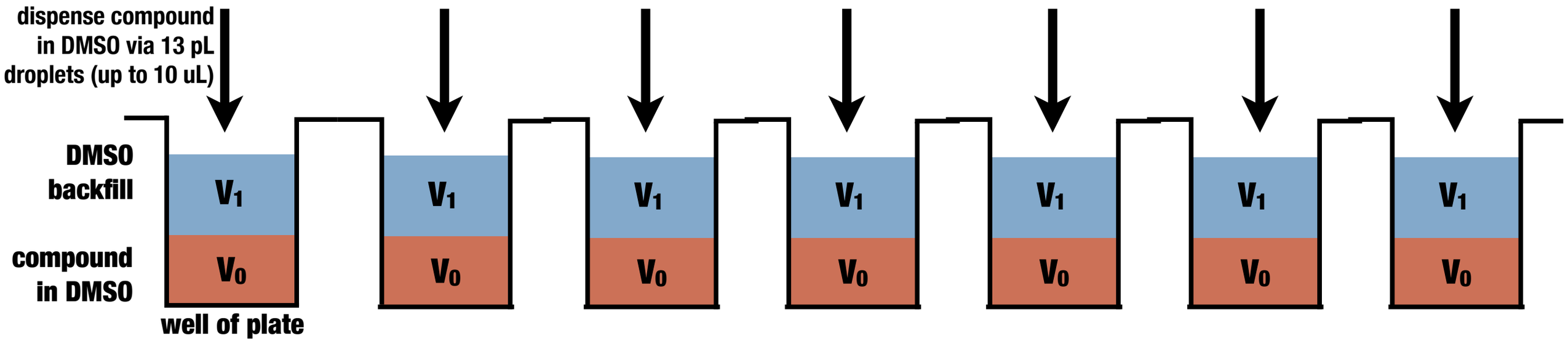
coefficients of variation (CV)

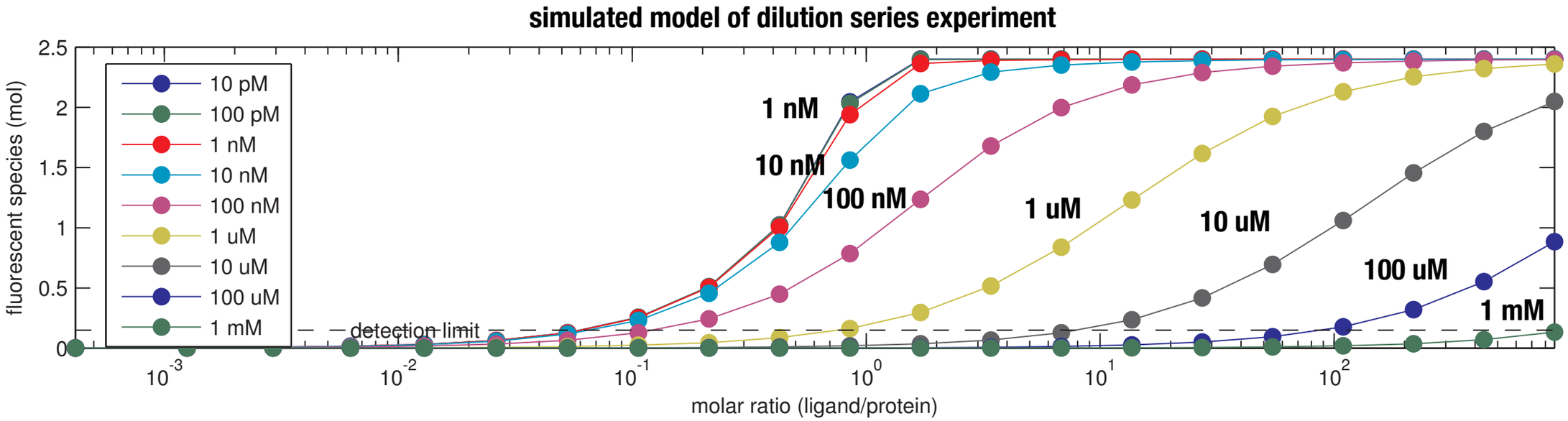


Tecan HP D300

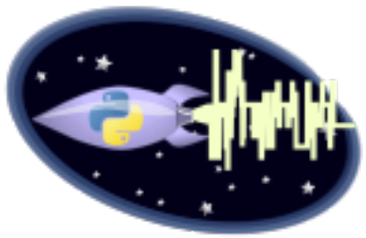
Code at https://github.com/choderalab/cup-xiv

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 (Monte Carlo sampling) to quantify experimental uncertainty



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

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

binding model

```
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]
```

pymc model

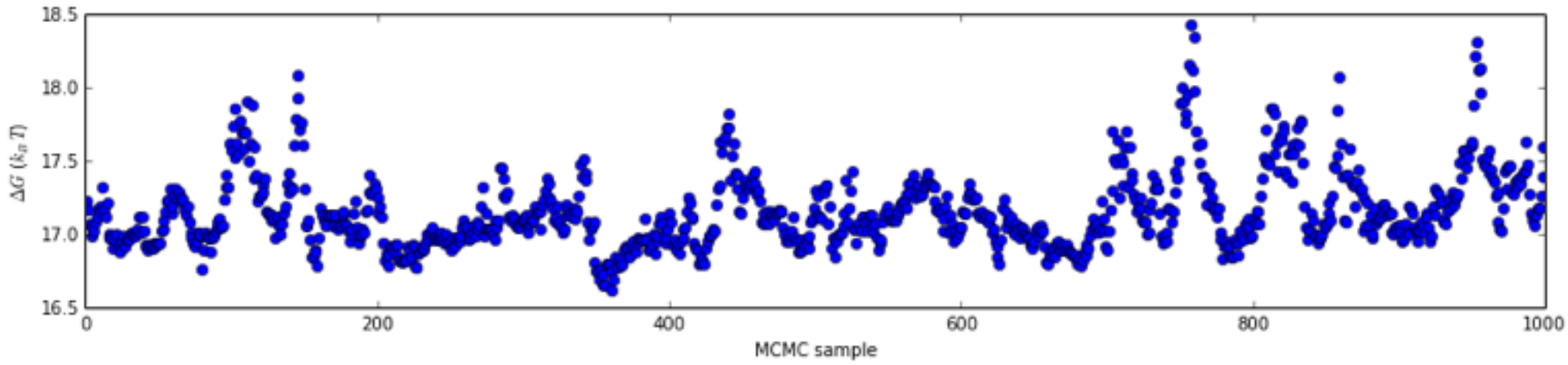
```
# Create a pymc model
def make_model(Pstated, dPstated, Lstated, dLstated, Fobs_i):
    N = len(Lstated)

    # 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.
    Ltrue = pymc.Lognormal('Ltrue', mu=np.log(Lstated**2 / np.sqrt(dLstated**2 + Lstated**2)), tau=np.sqrt(np.log(1.

    # 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 concentraiton uncertainty (due to gravimetric preparation and HP D300 dispensing)
```

pymc sampling

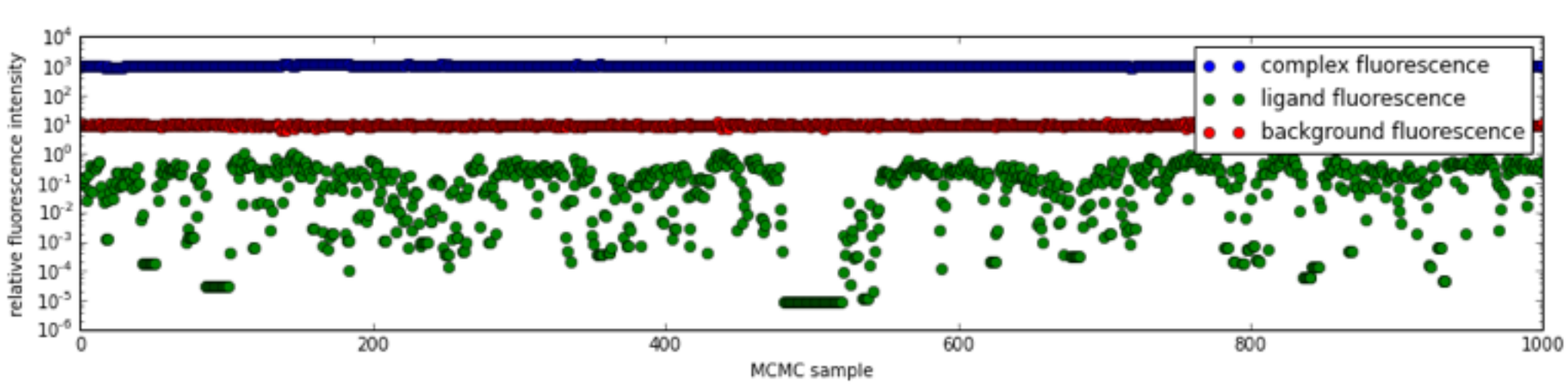
```
# Build model.
pymc_model = pymc.Model(make_model(Pstated, dPstated, Lstated, dLstated, F_i))

# 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)
```

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

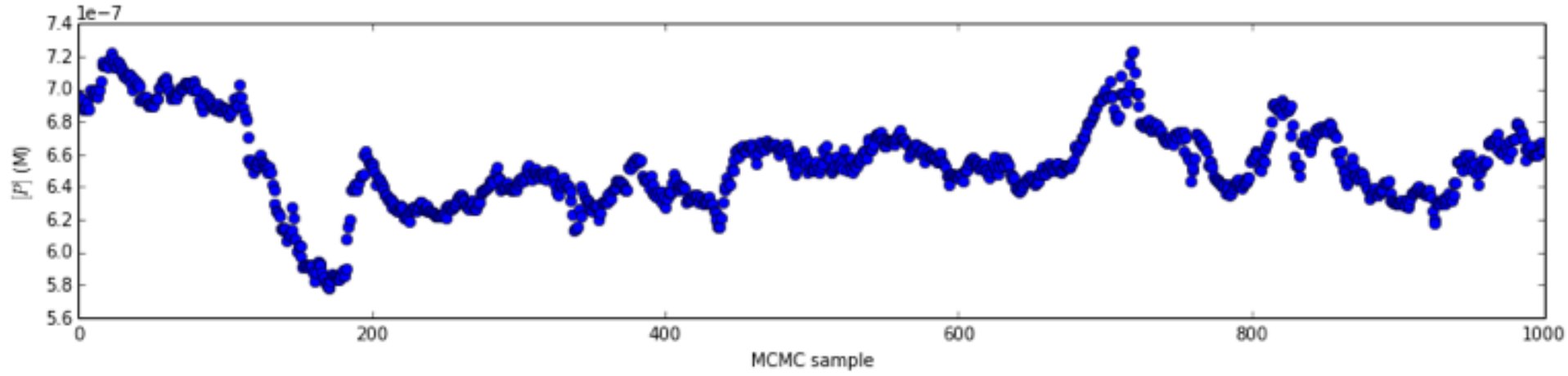
MCMC sampling of free energies



MCMC sampling of intrinsic ligand and complex fluorescence



MCMC sampling of true concentrations

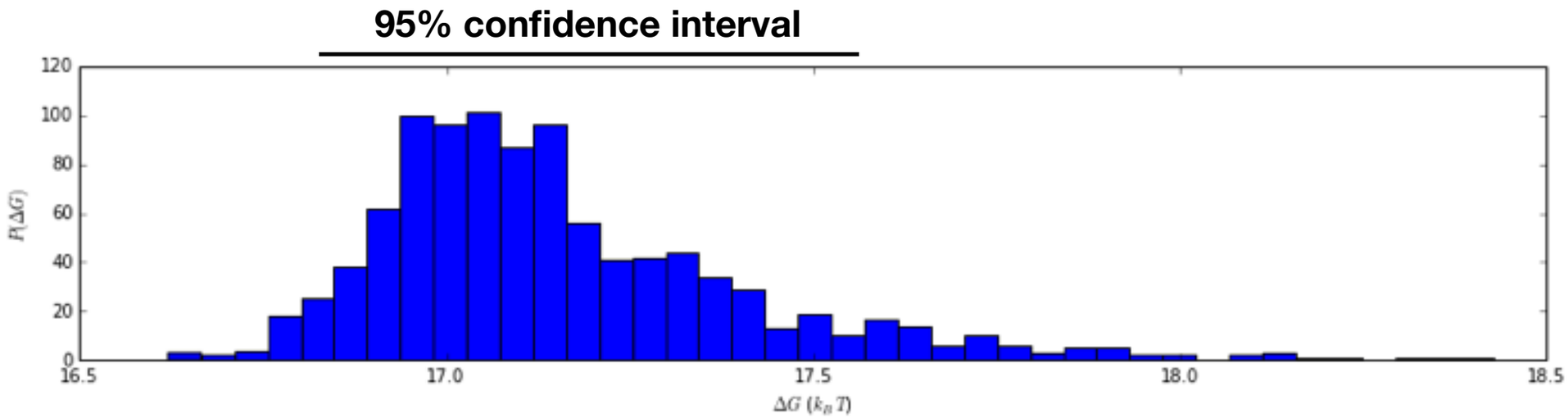


Code at https://github.com/choderala

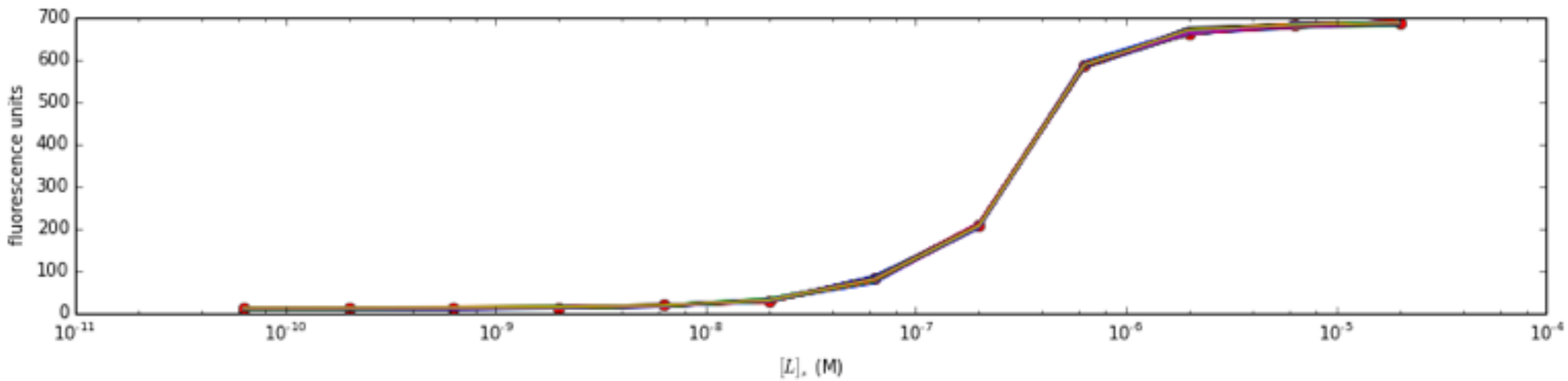
Bayesian inference allows us to use tools we know (Monte Carlo sampling) to quantify experimental uncertainty

---

posterior distribution of binding free energies



family of models fit to experimental fluorescence data



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

Assays are expensive: Can we change that?

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

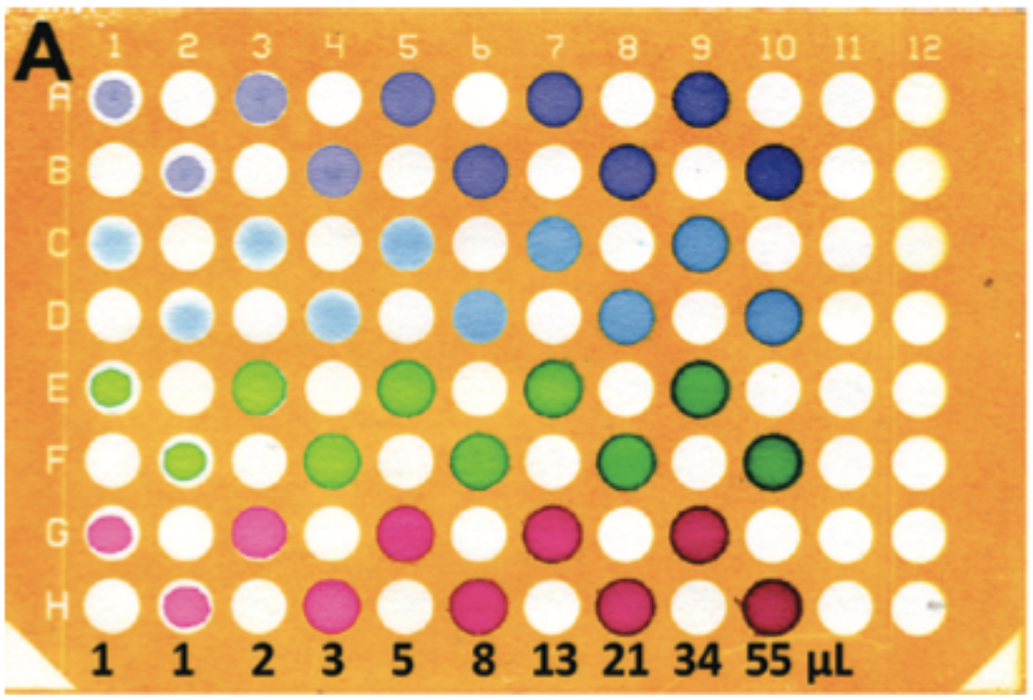
Assays are expensive: Can we change that?

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

Paper Microzone Plates

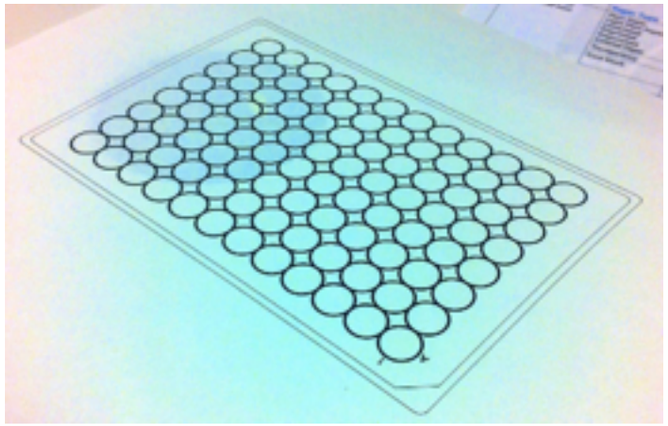
Emanuel Carrilho,*[,†,‡] Scott T. Phillips,[†,§] Sarah J. Vella,[†] Andres W. Martinez,[†] and George M. Whitesides*[,†]

Anal. Chem. **2009,** *81,* 5990–5998



Assays are expensive: Can we change that?

Xerox ColorQube 8570
wax printer



patterned wax microplate

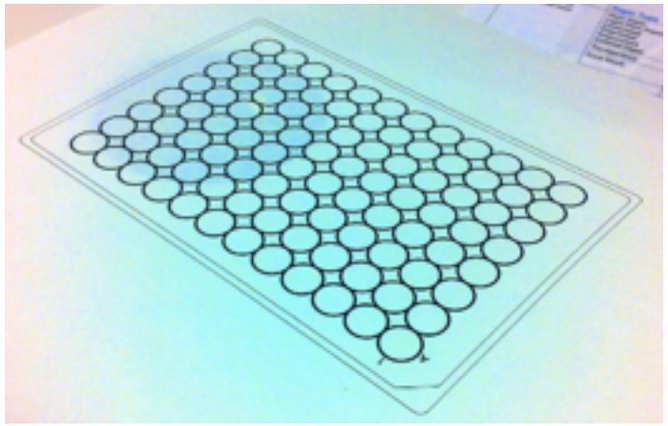
reflow wax through paper
with hot plate

Tecan HP D300
inkjet printer for compounds in DMSO

3D printed plate carrier

Assays are expensive: Can we change that?

Xerox ColorQube 8570
wax printer



patterned wax microplate

reflow wax through paper
with hot plate

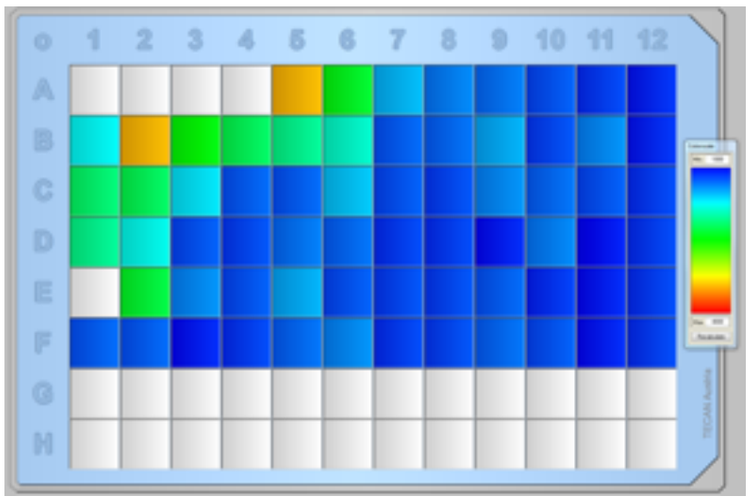
Tecan HP D300
inkjet printer for compounds in DMSO

3D printed plate carrier

Assays are expensive: Can we change that?

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

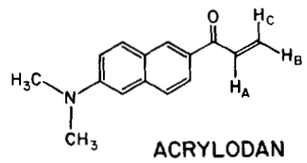
paper microzone plate: $0.20/plate
protein cost: $2/plate
10x reduction in costs!



Fluorescent labeling (FLiK) assays give complementary information about kinase conformational state

acrylodan fluorescent labeling of introduced Cys

[Prendergast et al. J Biol Chem 258:7541, 1983]



type IV: V338C

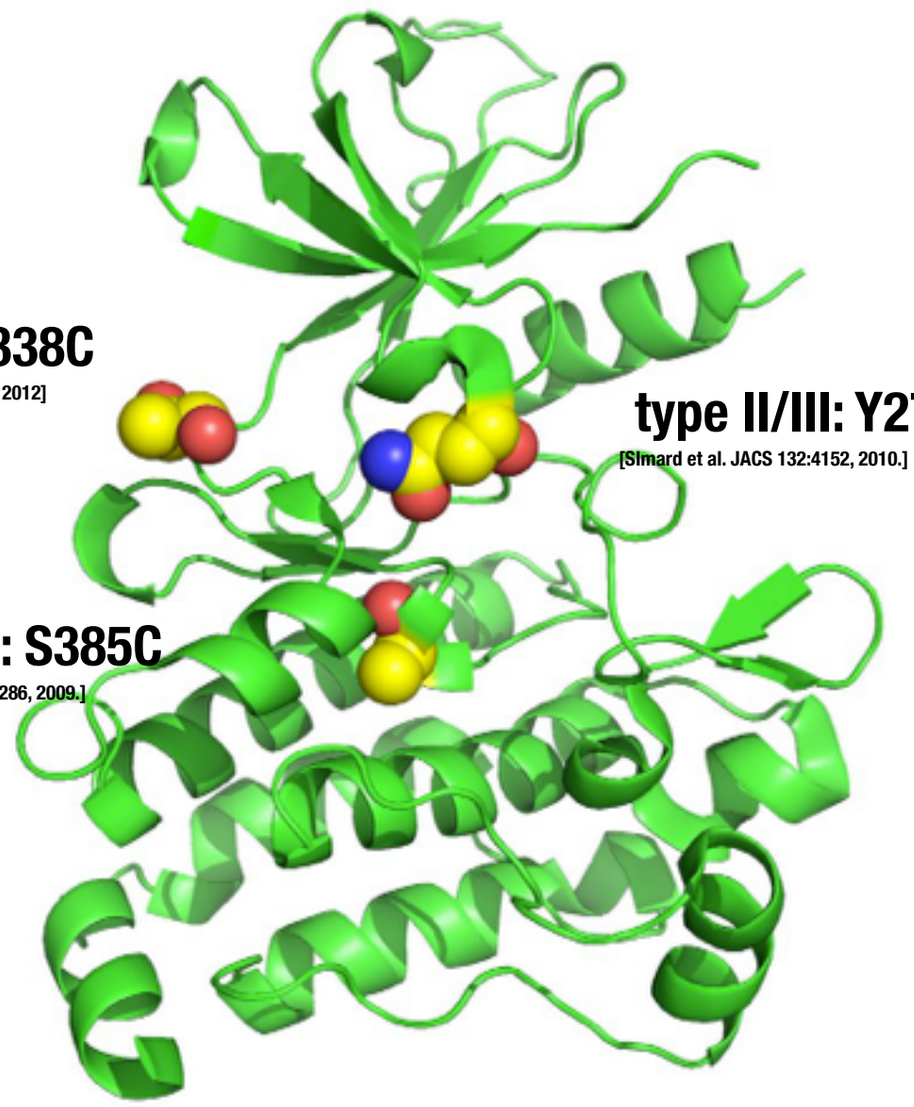
[Schneider et al. JACS 134:9138, 2012]

type II/III: Y271C

[Simard et al. JACS 132:4152, 2010.]

type II/III: S385C

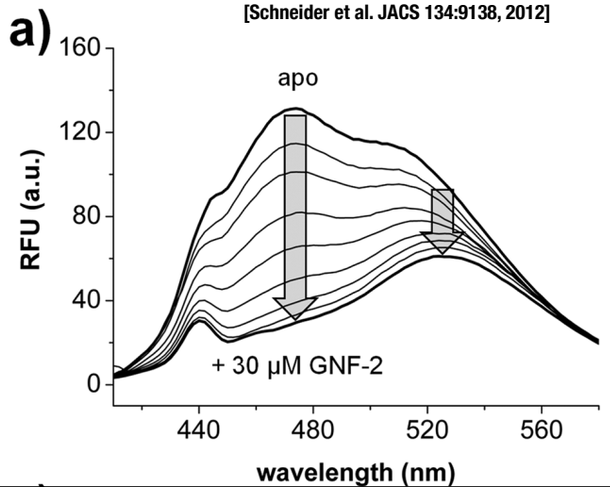
[Simard et al. JACS 131:13286, 2009.]



Abl kinase [3k5v]

type IV: V338C

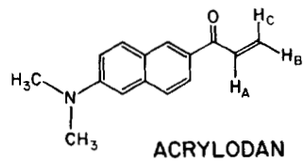
[Schneider et al. JACS 134:9138, 2012]



Fluorescent labeling (FLiK) assays give complementary information about kinase conformational state

acrylodan fluorescent labeling of introduced Cys

[Prendergast et al. J Biol Chem 258:7541, 1983]



type IV: V338C

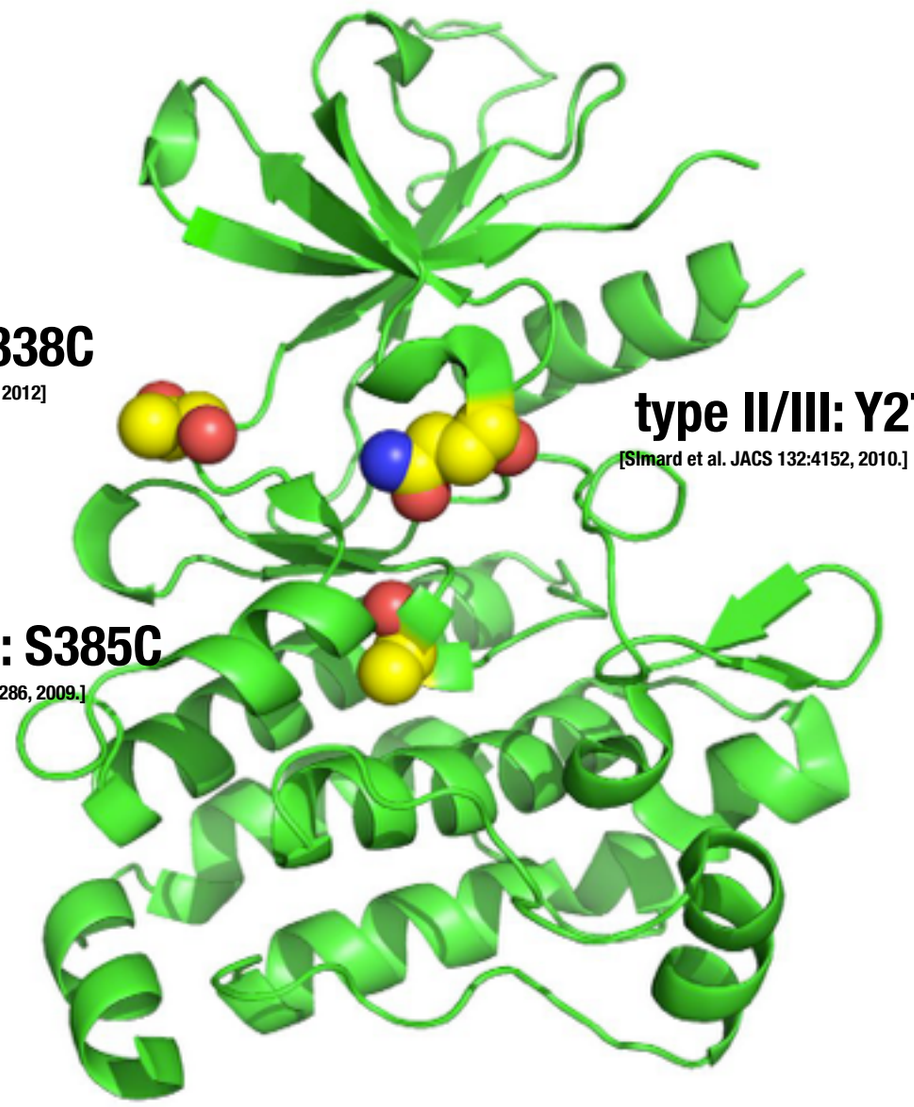
[Schneider et al. JACS 134:9138, 2012]

type II/III: Y271C

[Simard et al. JACS 132:4152, 2010.]

type II/III: S385C

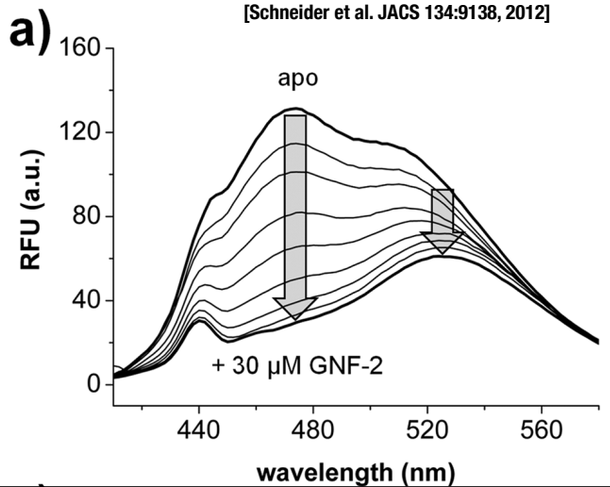
[Simard et al. JACS 131:13286, 2009.]



Abl kinase [3k5v]

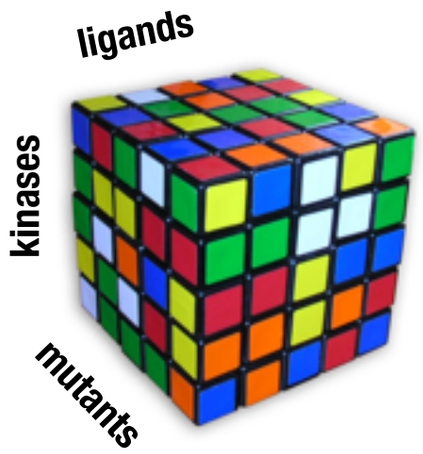
type IV: V338C

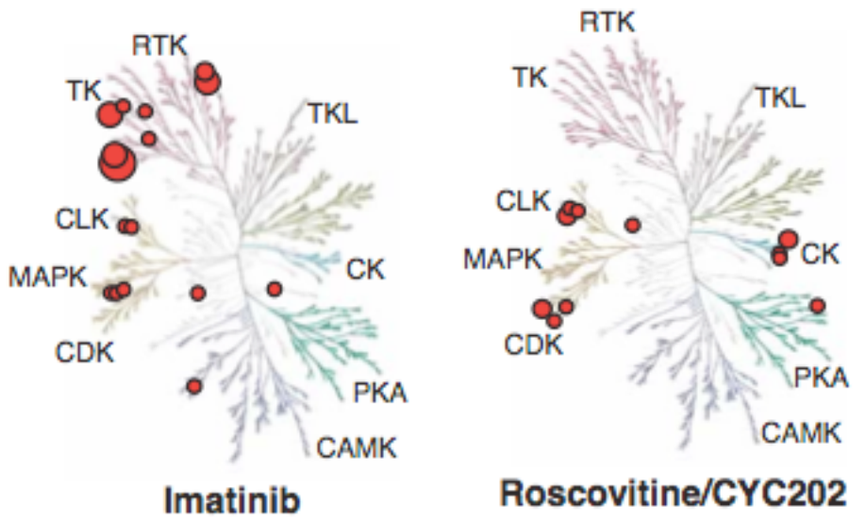
[Schneider et al. JACS 134:9138, 2012]



What determines selectivity of inhibitors for kinases?

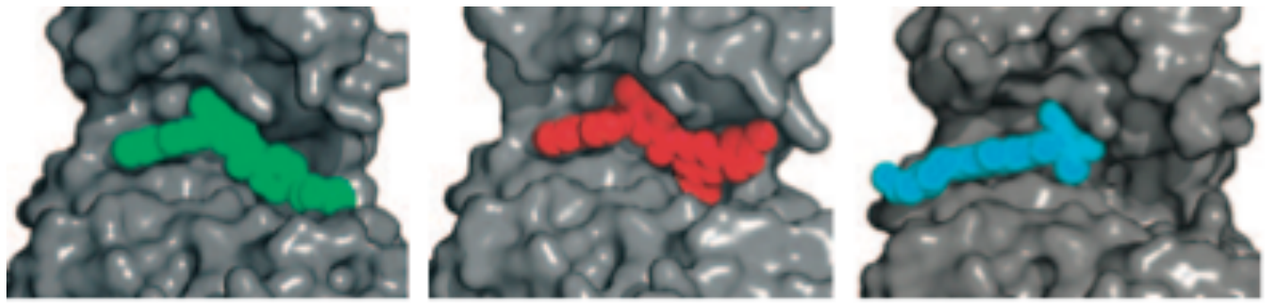
High-throughput fluorescence measurements and free energy calculations can address physical determinants of kinase inhibitor selectivity:





Nature Biotech 23:329, 2005

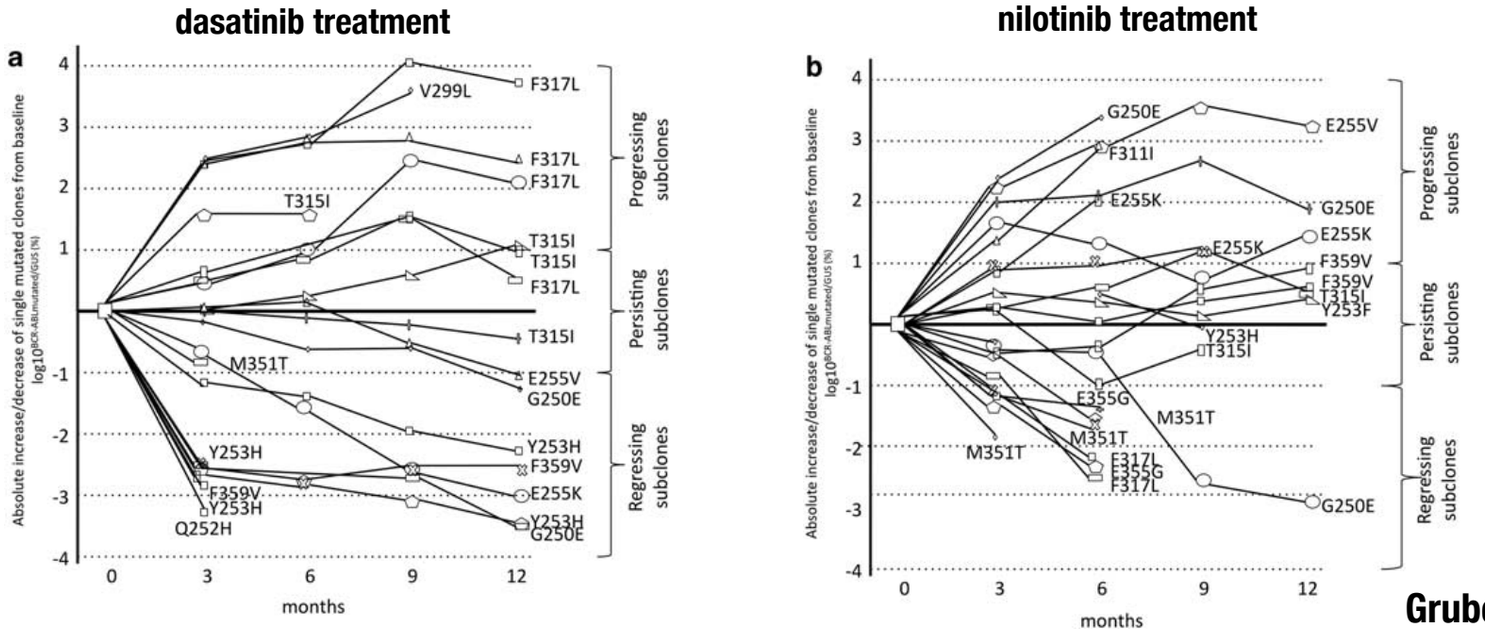
* Are particular ligand scaffolds privileged with specificity?
* Are particular binding modes better for specificity?
* Are certain kinases inherently more promiscuous?



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

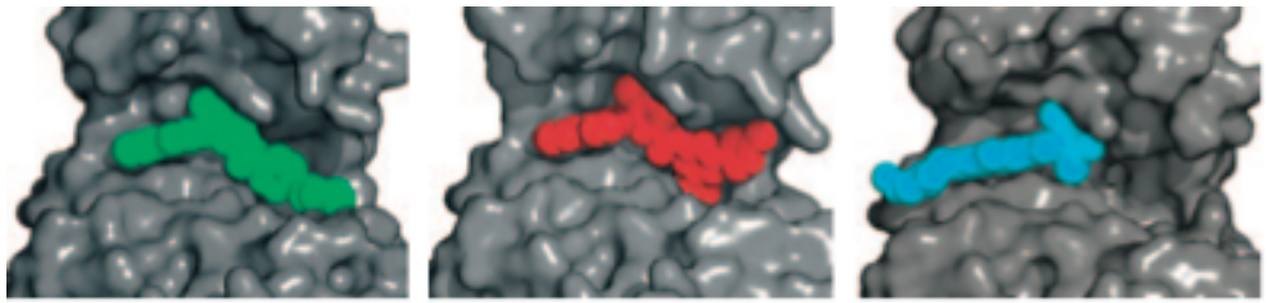
Can we develop a physical model of resistance mutations?

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



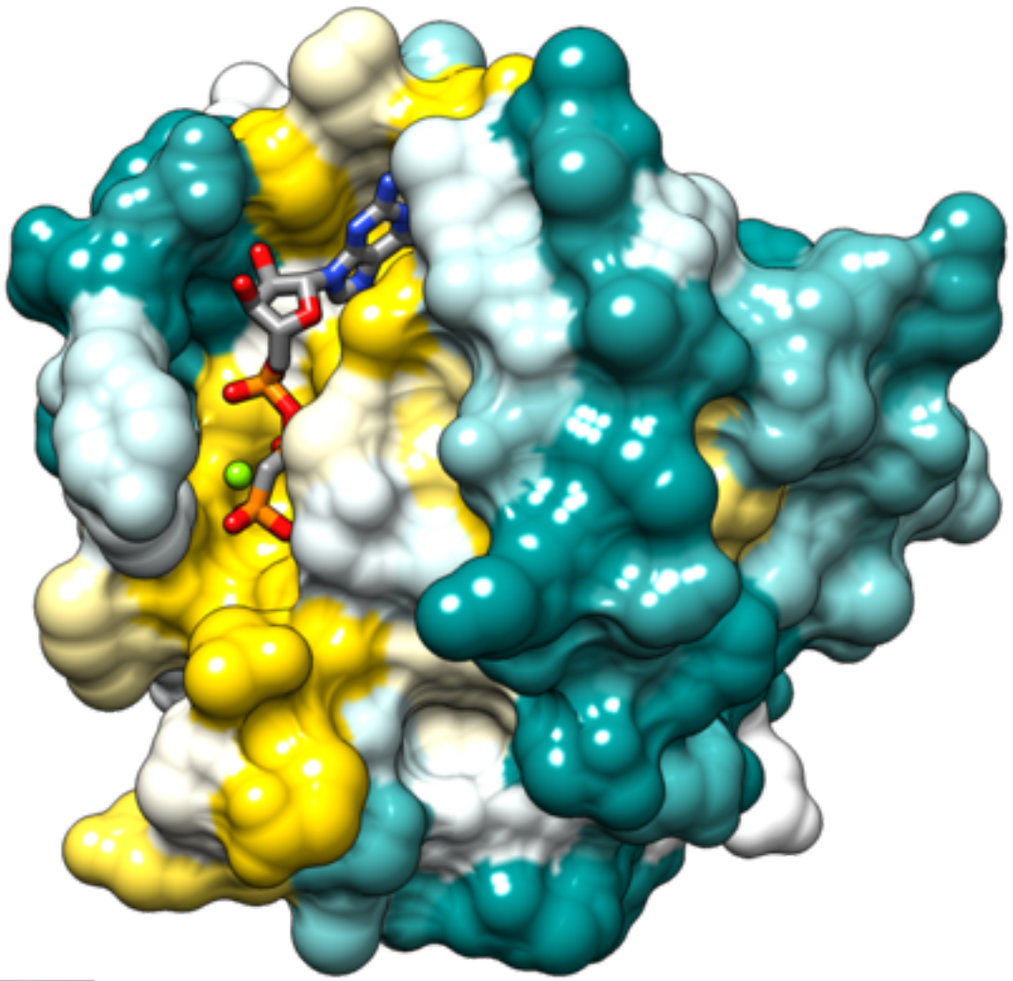
Gruber et al. Leukemia 26:172, 2012.

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.

Can we drug the undruggable? Allosteric modulators of Ras may open new doors in cancer therapy



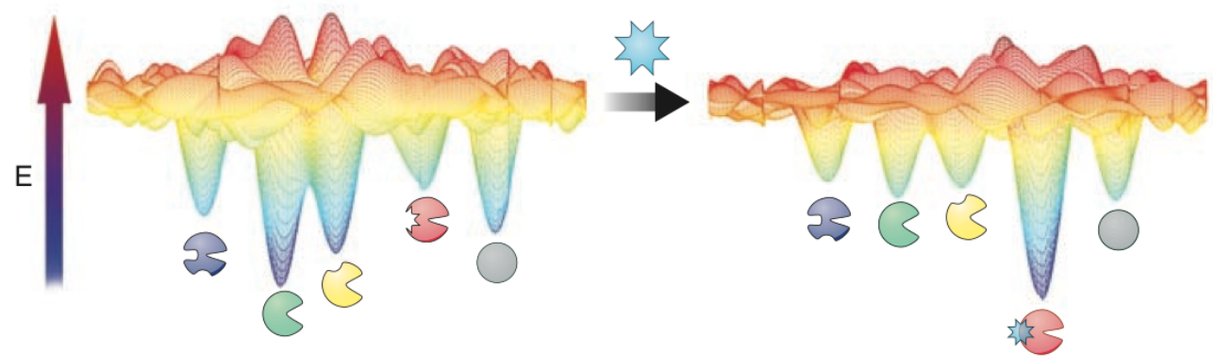
human HRAS with GTP analogue [121p]

Mutant Ras found in 20-30% of all human tumors

Oncogenic mutations constitutively activate Ras
by eliminating catalytic activity (eg Q61K) or rendering Ras insensitive to inactivation by GAP

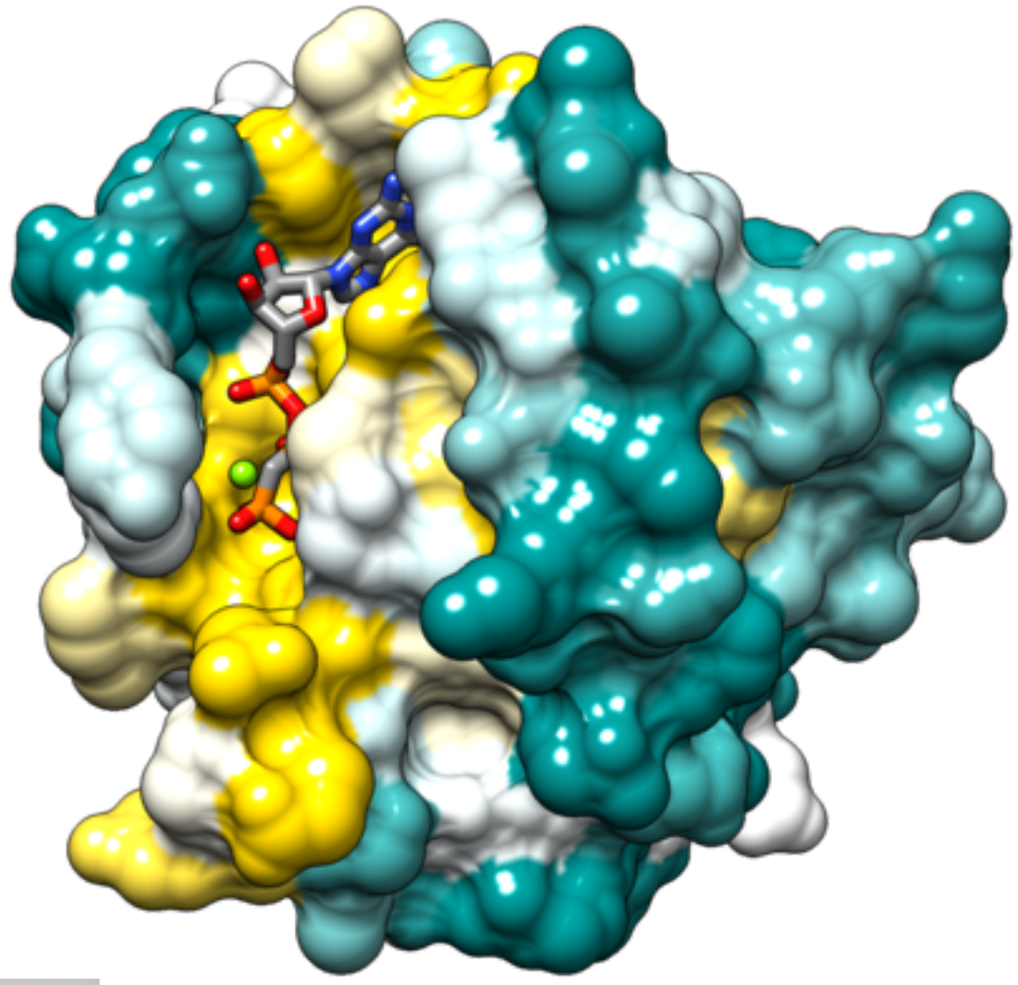
"Undruggable" because these are loss-of-function mutations; hard to conceive of drugs to restore function

Can we inactivate Ras by engineering allosteric modulators that trap Ras in an "off" conformation?



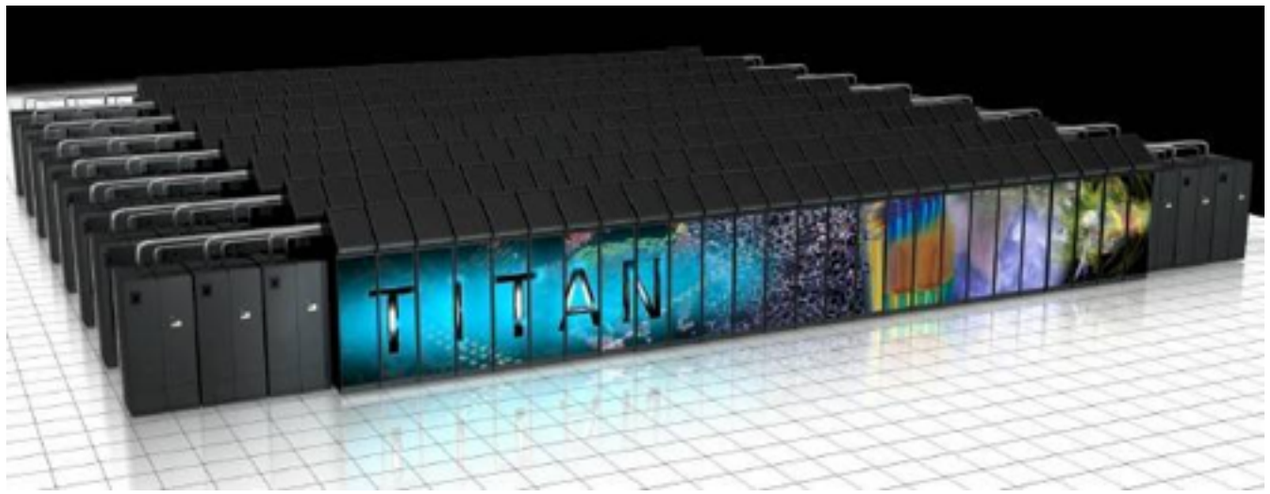
Patrick Grinaway
In collaboration with Jeremy C. Smith (ORNL), Guillermo Perez-Hernandez and Frank Noé (FU Berlin)

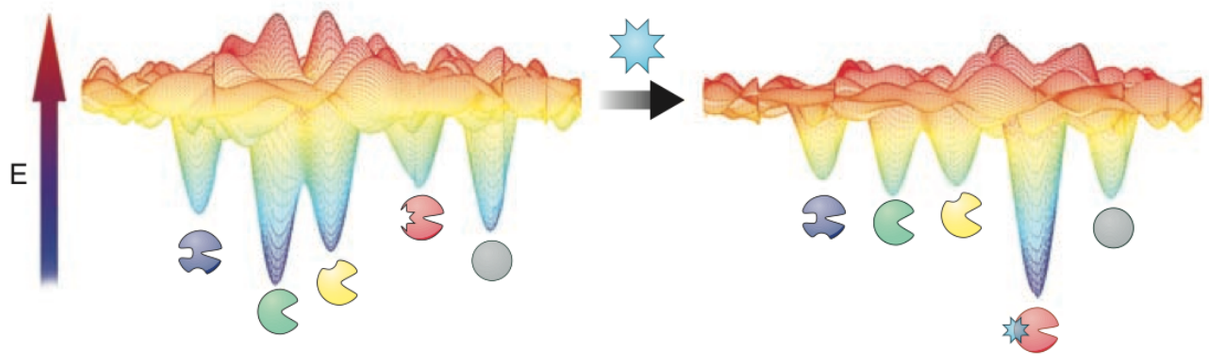
Can we drug the undruggable? Allosteric modulators of Ras may open new doors in cancer therapy



human HRAS with GTP analogue [121p]

ORNL Titan: 18,688 NVIDIA Tesla K20 GPUs





Patrick Grinaway
In collaboration with Jeremy C. Smith (ORNL), Guillermo Perez-Hernandez and Frank Noé (FU Berlin)

What big challenges lie ahead for molecular modeling?

---

* How can we most effectively use cycles of computation and experiment to iteratively improve models?

* How will molecular modeling and physical simulation exploit the era of “cheap genomics”?

* How can we make our tools more accessible and more impactful for experimentalists and synthetic chemists?

Contact us!
choderaj@mskcc.org

The Chodera Lab @ MSKCC

Danny Parton
postdoc

Sonya Hanson
postdoc

Jan-Hendrik Prinz
postdoc

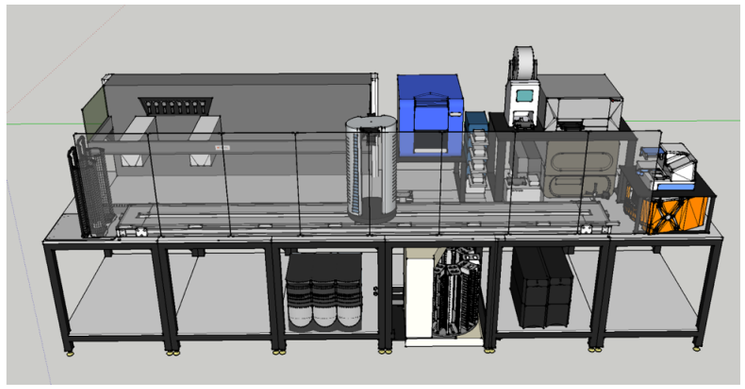
Sarah Boyce
postdoc alumna
(now at Schrödinger)

Patrick Grinaway
PBSB graduate student

Julie Behr
CBM rotation
student

Kyle Beauchamp
postdoc

RUG-1
robot



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

Stony Brook
Ken Dill

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

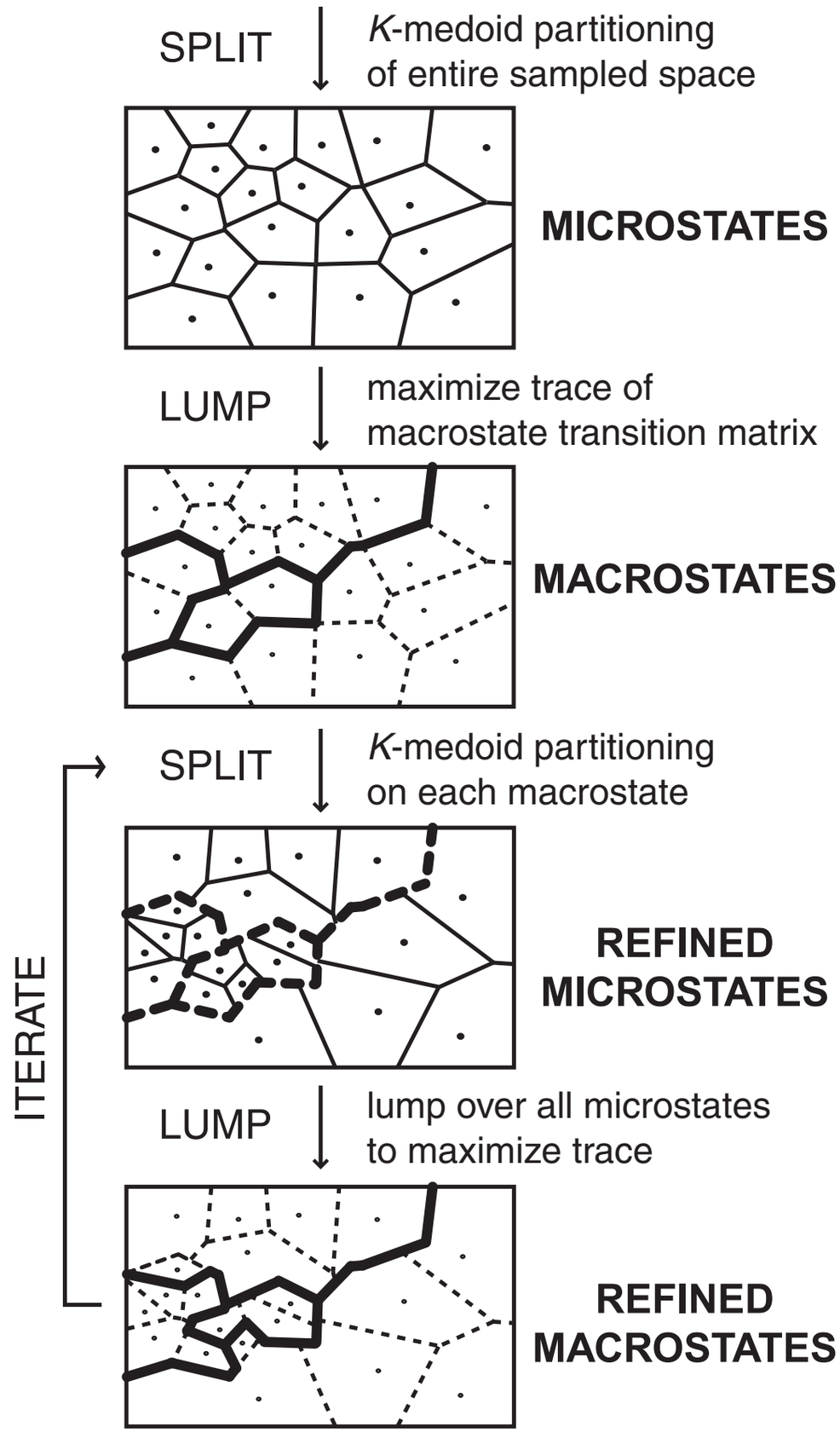
LBNL
Gavin Crooks

Vanderbilt
Joel Tellinghuisen

Hessian Informatics
Kim Branson

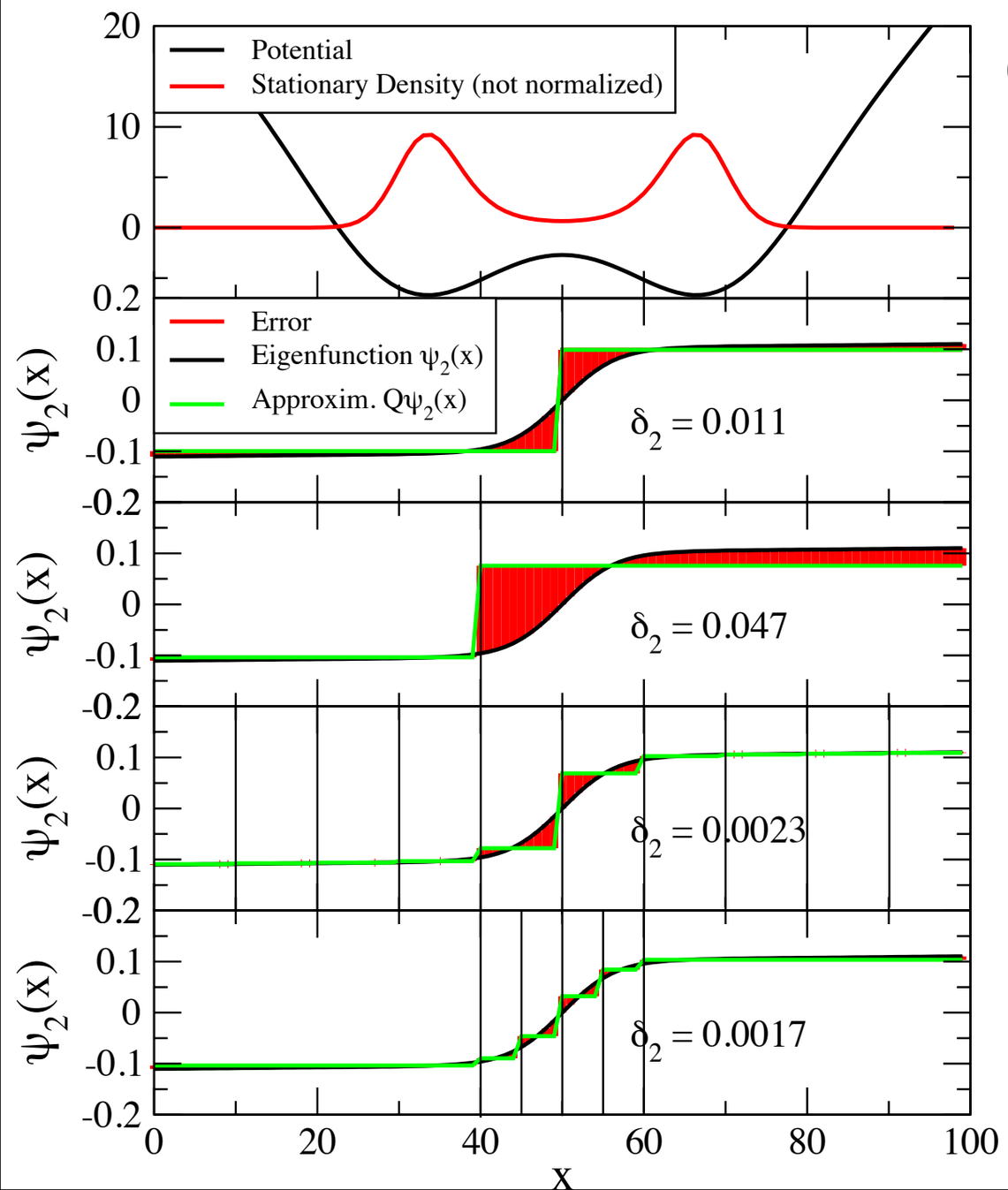
Vertex Pharmaceuticals
Richard Dixon

Code and data available at http://www.choderalab.org
Start folding at http://folding.stanford.edu


SPLIT
K-medoid partitioning
of entire sampled space
MICROSTATES
LUMP
maximize trace of
macrostate transition matrix
MACROSTATES
SPLIT
K-medoid partitioning
on each macrostate
REFINED
MICROSTATES
ITERATE
LUMP
lump over all microstates
to maximize trace
REFINED
MACROSTATES

Approximation quality improves with finer partitioning

$$E(k) := \left\| Q[\mathcal{T}(\tau)]^k Q - Q[\mathcal{T}(\tau)Q]^k \right\|_{\mu,2} \qquad E(k) \leq \min\{2, [m\delta + \eta(\tau)]\,[a(\delta) + b(\tau)]\}\,\lambda_2^k$$

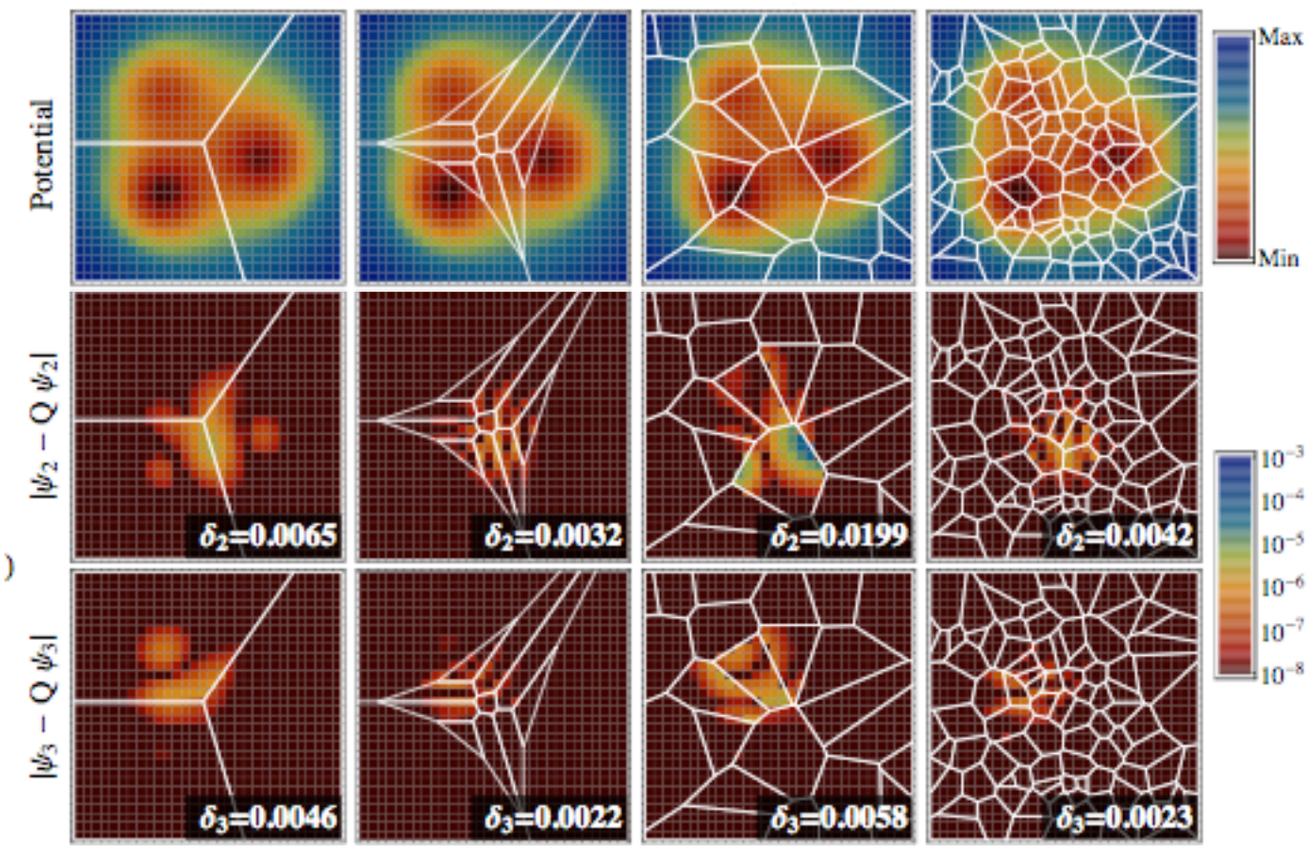


eigenfunction error $a(\delta) = \sqrt{m}(k-1)\delta$

spectral error $b(\tau) = \dfrac{\eta(\tau)}{1-\eta(\tau)}(1-\eta(\tau)^{k-1})$

spectral ratio $\eta(\tau) = \dfrac{\lambda_{m+1}(\tau)}{\lambda_2(\tau)}$

Sarich, Noé, Schütte. Multisc. Model Simul. 2010.



Prinz, Wu, Sarich, Keller, Fischbach, Held, Chodera, Schütte, and Noé. JCP 2011.