REDESIGNING DRUG DESIGN

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

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

DISCLOSURES:
- Scientific Advisory Board, Schrödinger

15 Dec 2016 - GSK - Collegeville, PA

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


Dana Pe'er


John Chodera


Christina Leslie

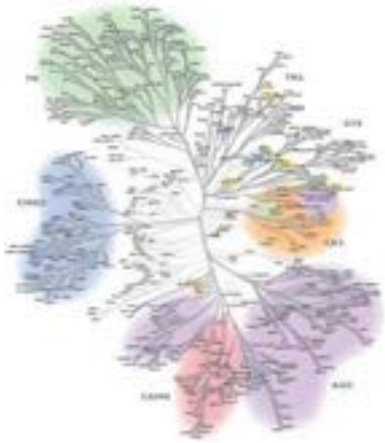

Joao Xavier

WE'RE HIRING!

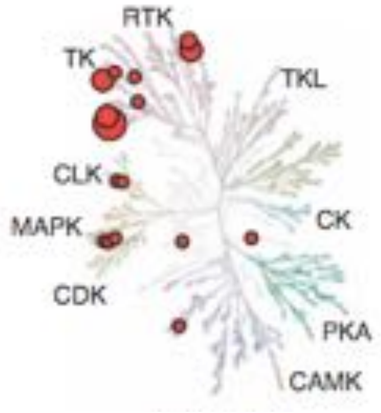
SOMETIMES, DRUG DISCOVERY WORKS WELL

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

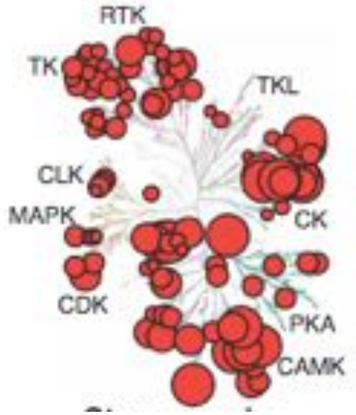
imatinib bound to c-Abl [PDB:1IEP]



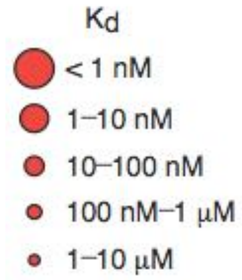
human kinome
[518 kinases]



imatinib
[blockbuster drug]



staurosporine
[toxic natural product]



Nature Biotech 23:329, 2005

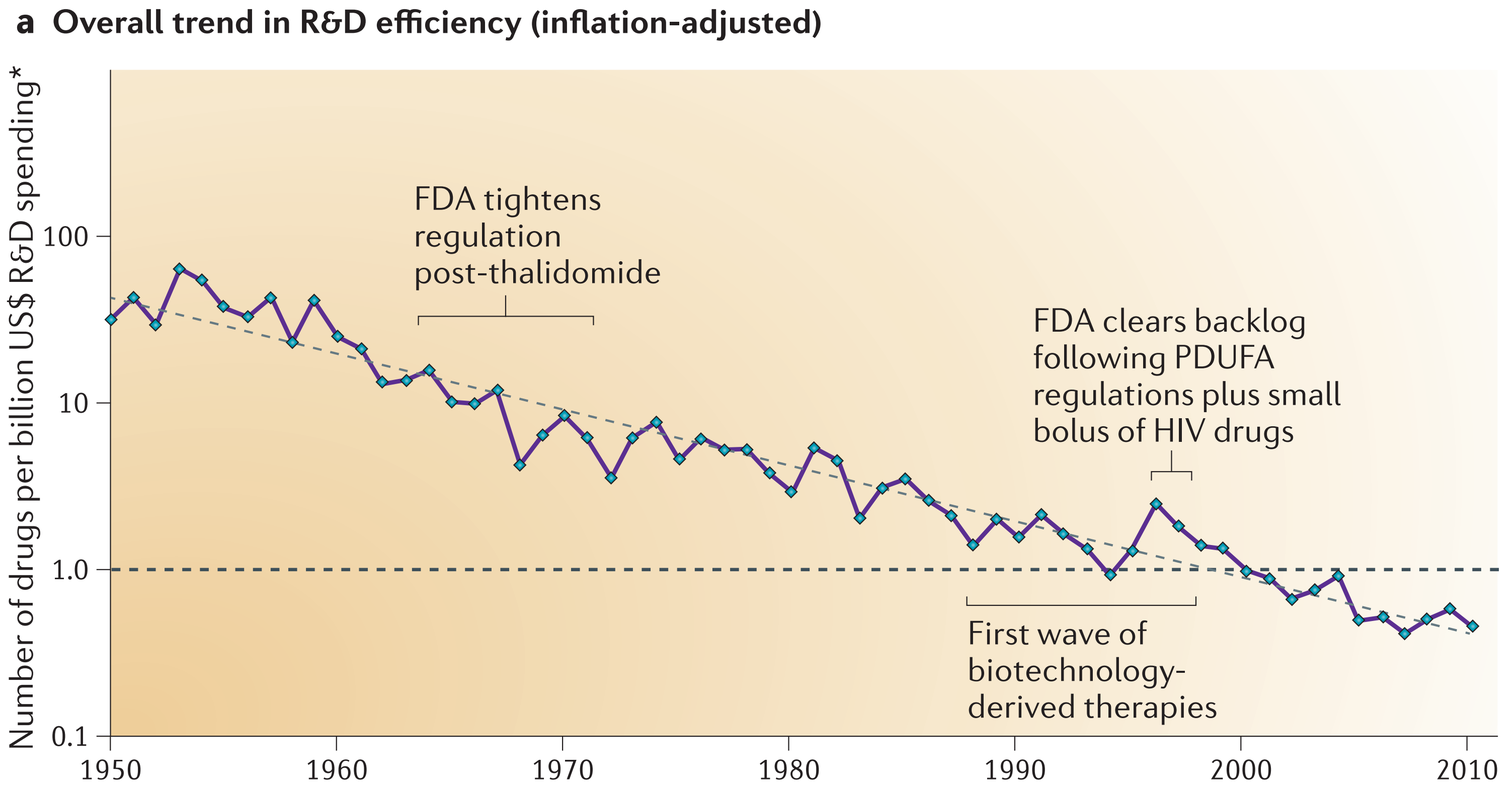
DRUG DISCOVERY USUALLY ENDS IN FAILURE

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

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

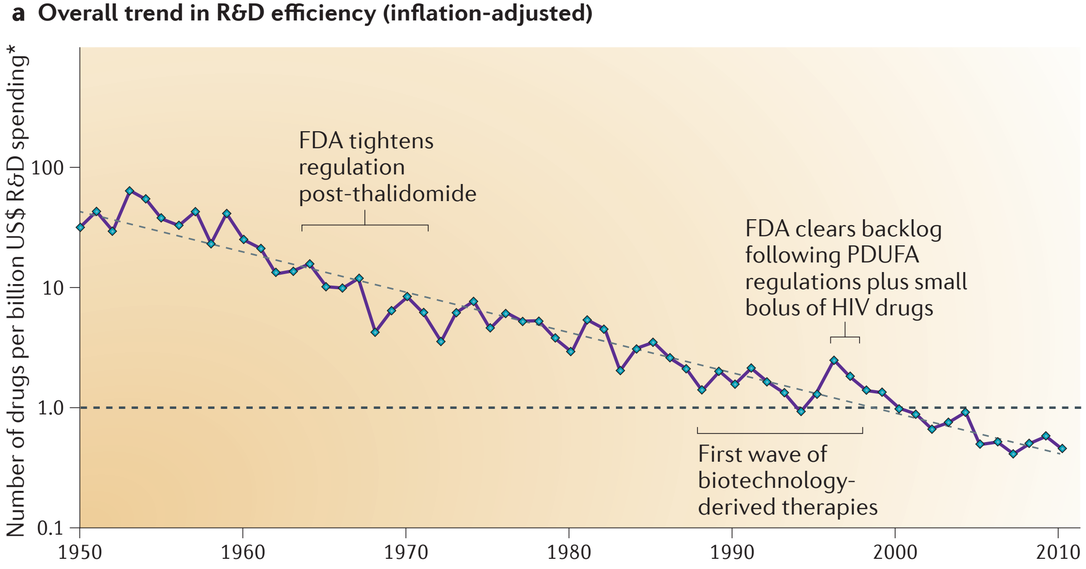
Number of truly innovative new molecules **remained constant at 5-6/year**

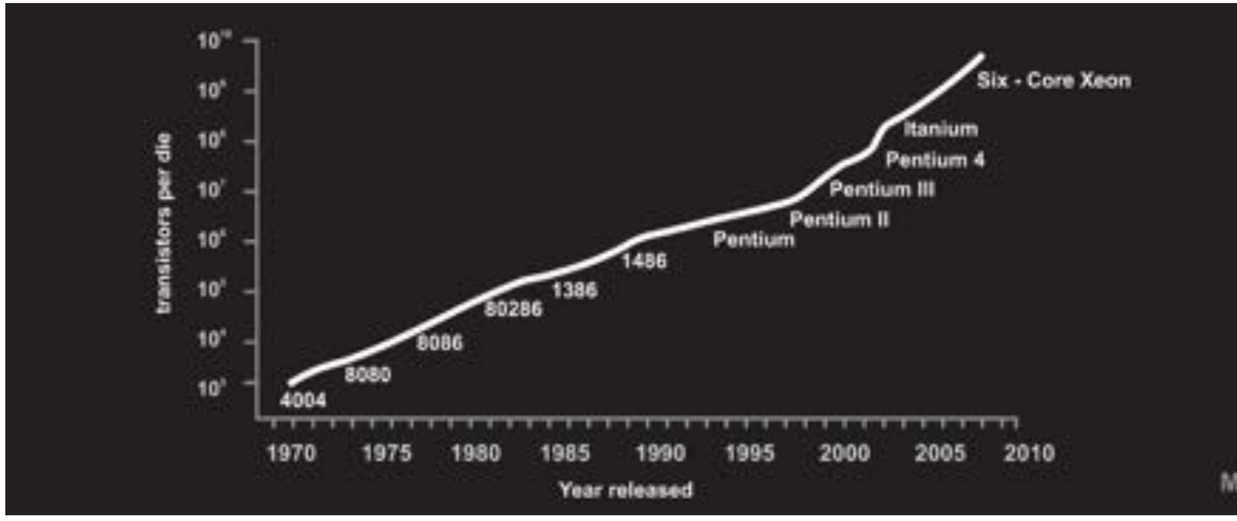
2010-2015 has seen large reductions in pharma R&D in the US



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

DRUG DISCOVERY USUALLY ENDS IN FAILURE

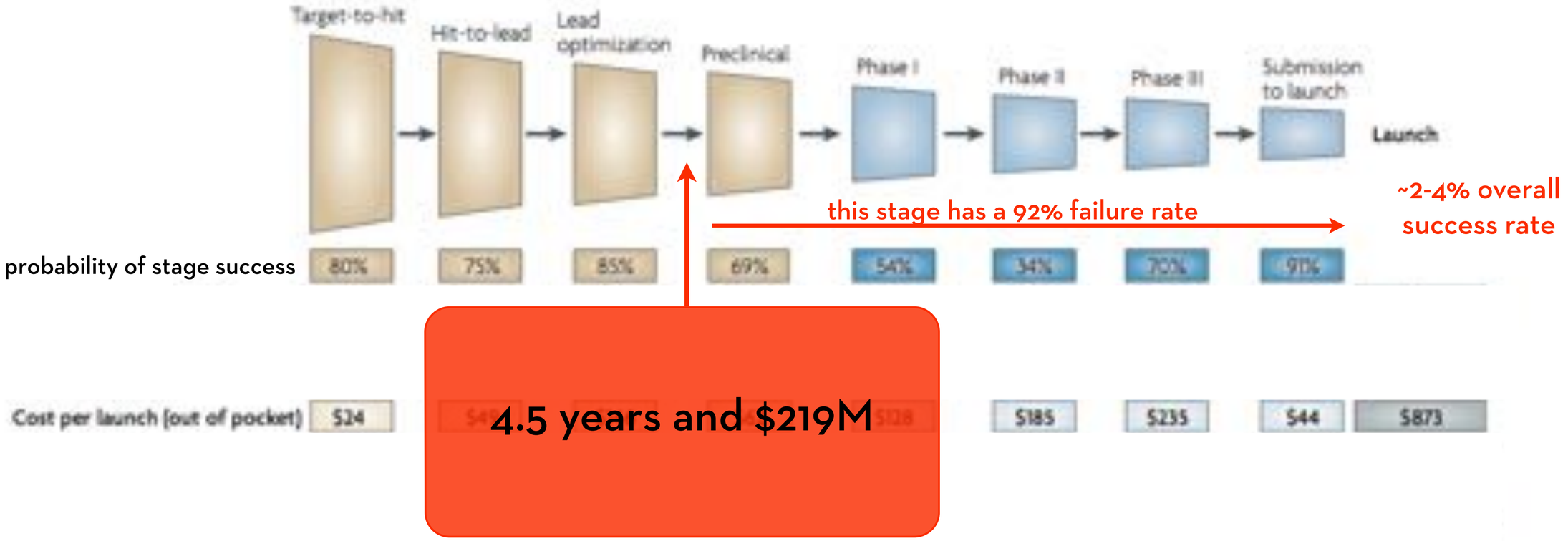




EROOM'S LAW

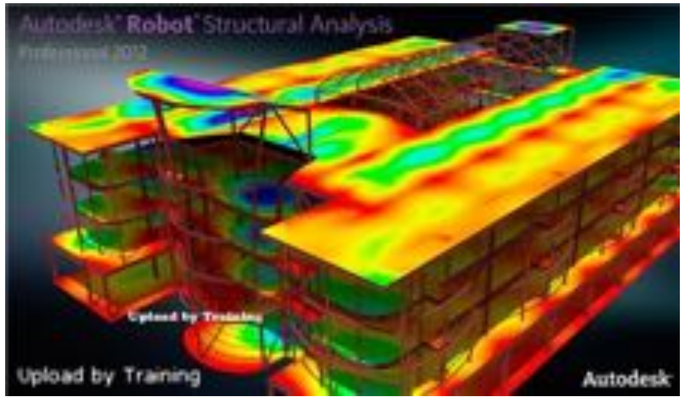
MOORE'S LAW

DRUG DISCOVERY USUALLY ENDS IN FAILURE



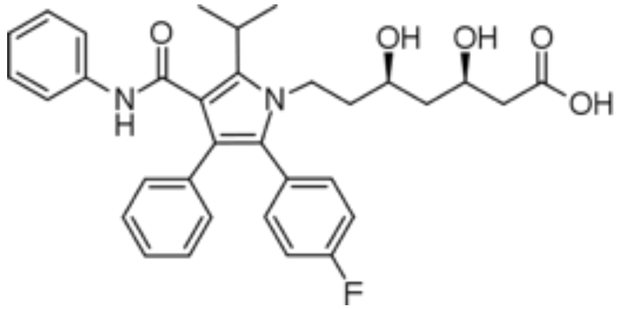
Paul et al. Nat. Rev. Drug Discover. 9:203, 2010.
Chodera et al. Curr. Opin. Struct. Biol., 21:150, 2011.

WE REGULARLY DESIGN PLANES, BRIDGES, AND BUILDINGS ON COMPUTERS



10^3 - 10^6 parts

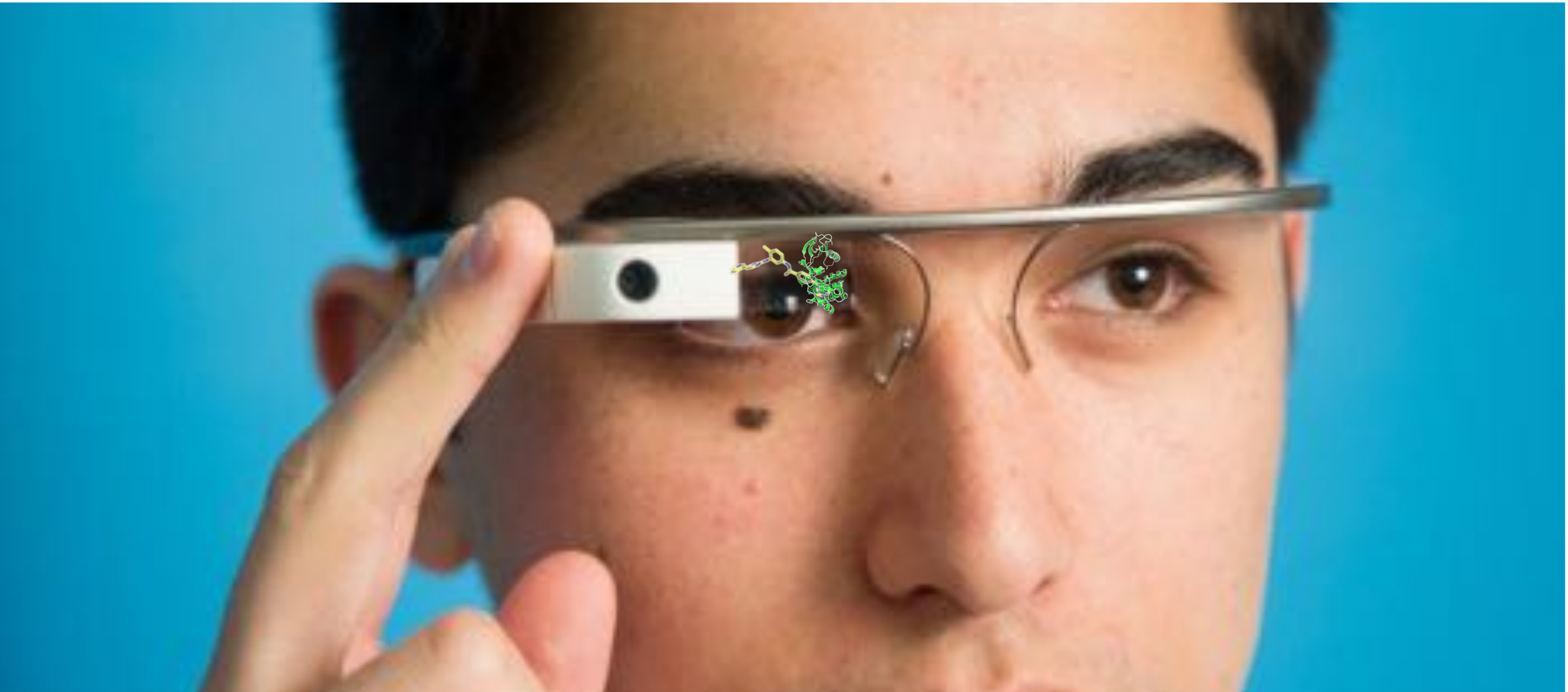
WHY NOT SMALL MOLECULE DRUGS?



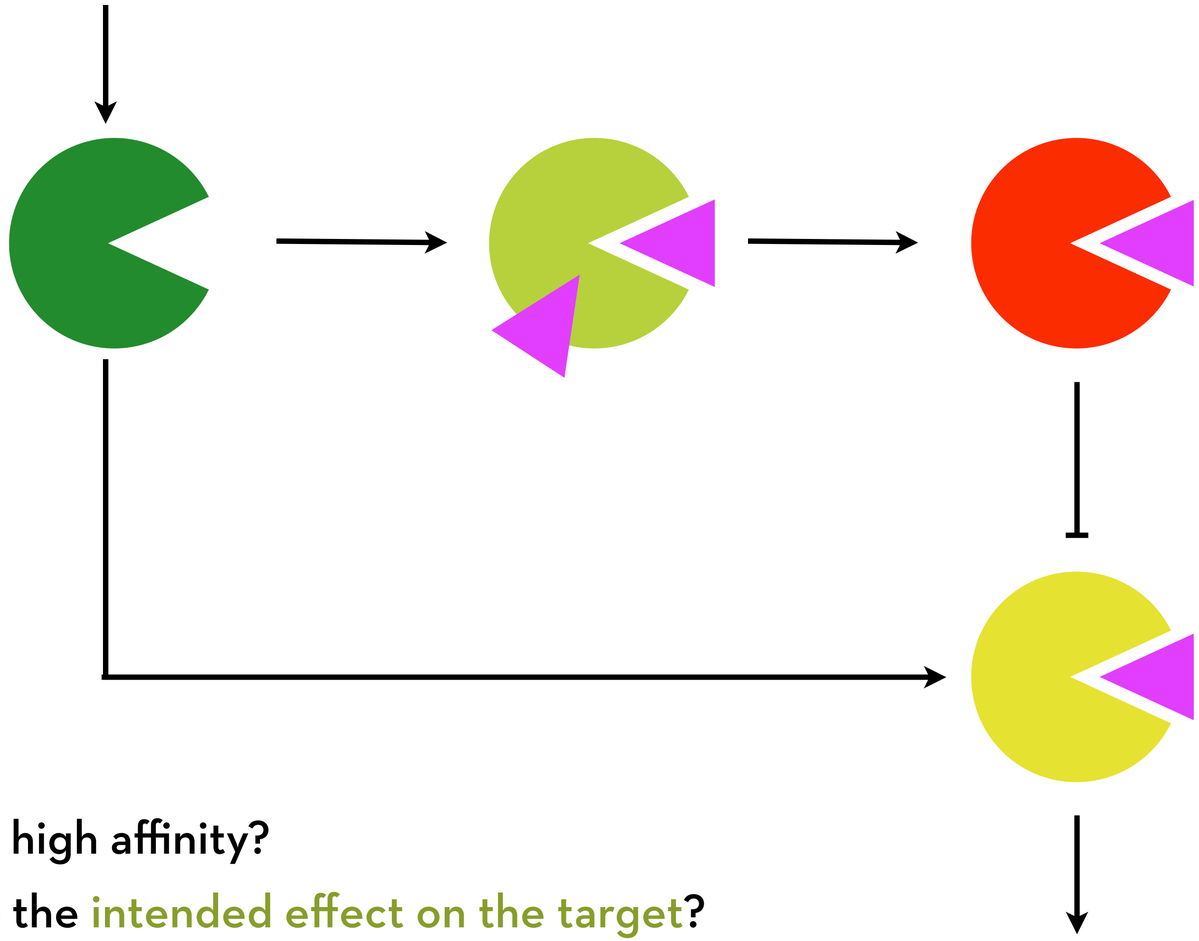
< 10^2 atoms

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

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



HOW CAN WE DESIGN SMALL MOLECULES TO HAVE INTENDED BIOLOGICAL EFFECTS?



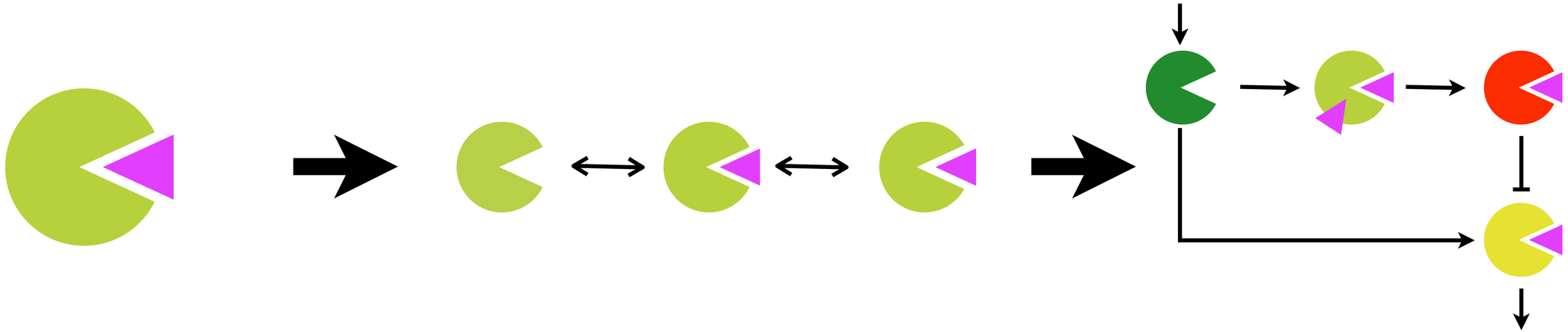
Will it bind the target with high affinity?

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

Does it produce the desired effect on cellular

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

MULTISCALE PHYSICAL MODELS CAN DRIVE SMALL MOLECULE DESIGN



physical binding constant

K_d

catalytic life cycle

$K_{i,app}$

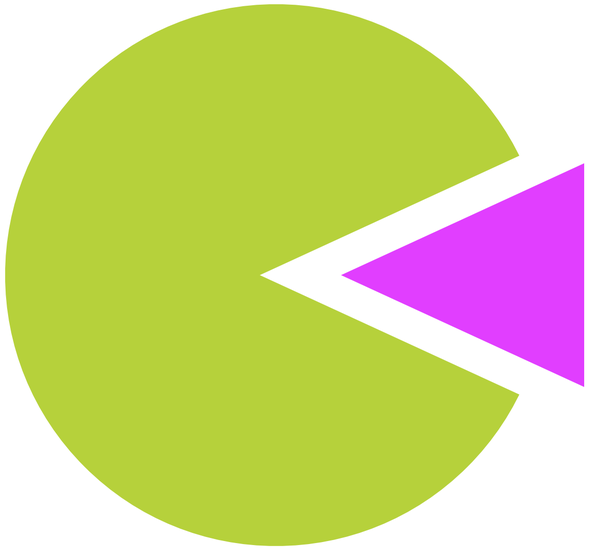
cellular pathways

EC_{50}

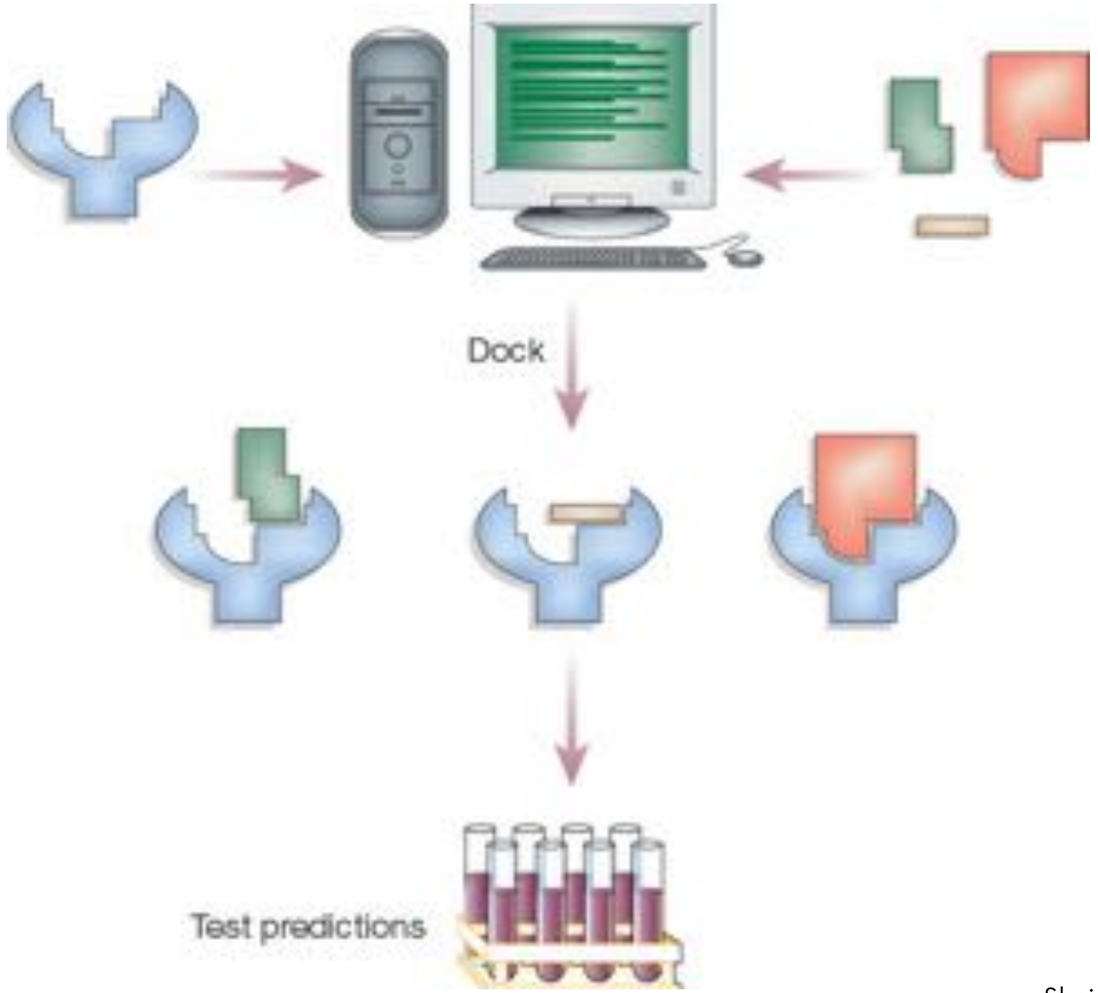
We use physical modeling and statistical mechanics to build predictive models

HOW CAN WE COMPUTE BINDING AFFINITIES FOR MOLECULES THAT HAVE YET TO BE SYNTHESIZED OR TESTED?

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



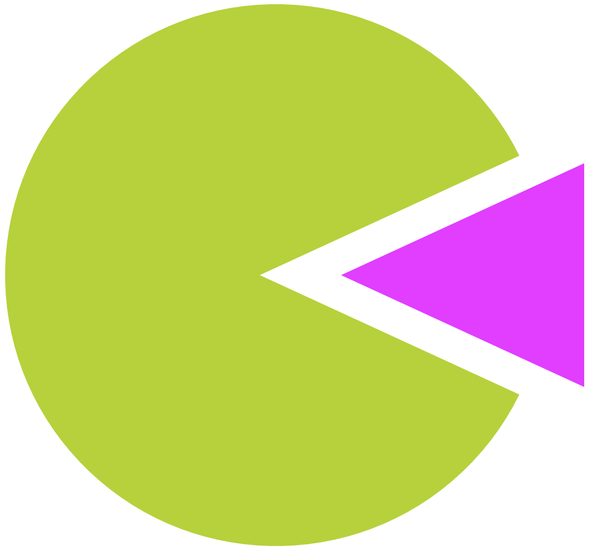
Virtual screening methods are in widespread use in drug discovery efforts today. They must work well, right?



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?

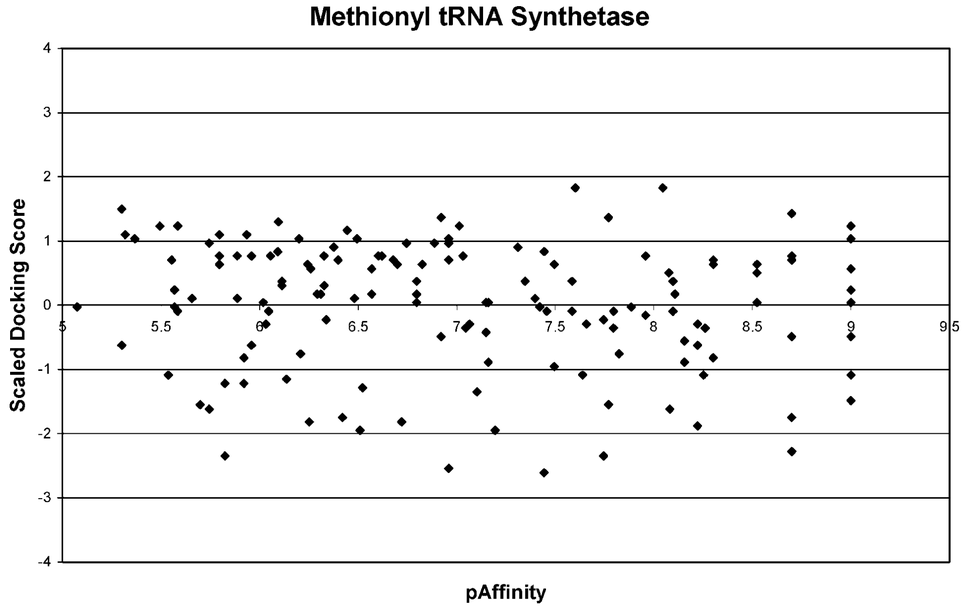
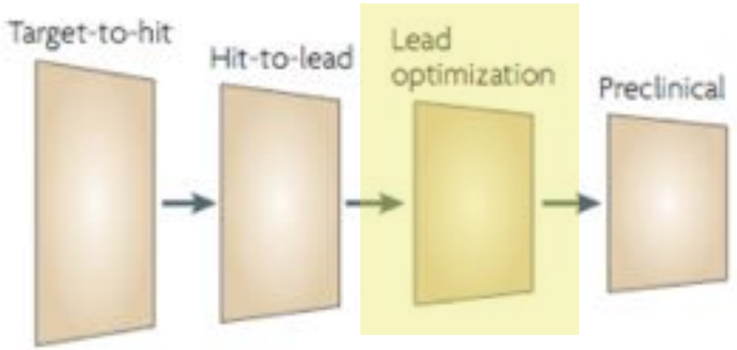


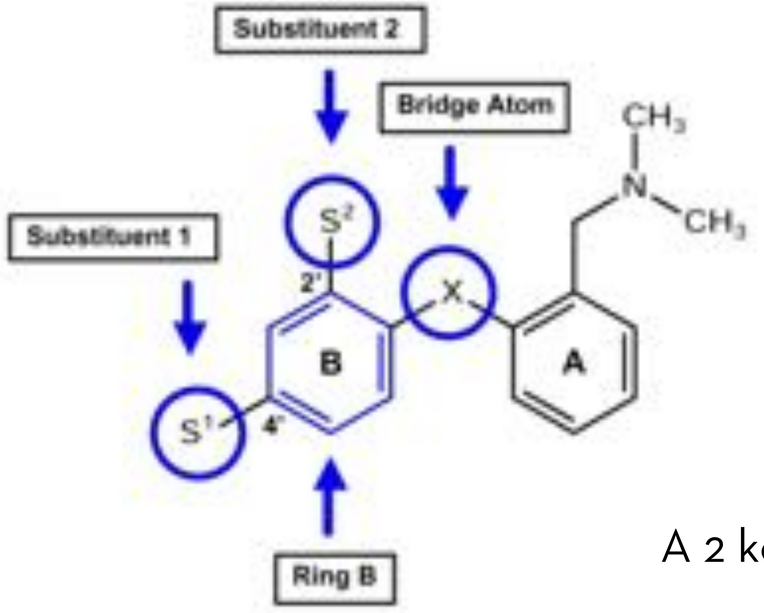
Figure 11. Plot of scaled score vs pAffinity for MRS and PPARδ. While the calculated correlation coefficient for the data shown for MRS is $r = -0.28$, this plot clearly demonstrates that these values are meaningless. No useful correlation exists between the docking score and compound affinity.

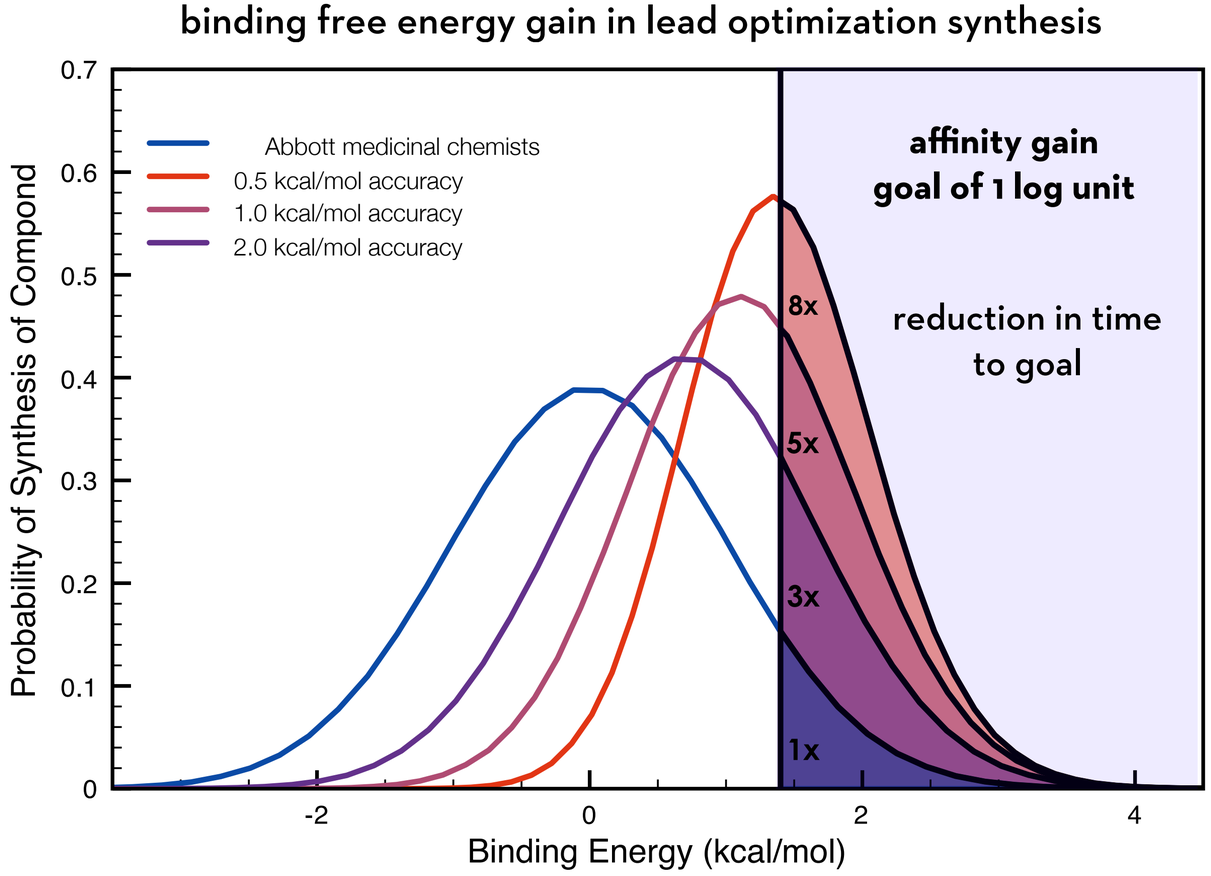
"For prediction of compound affinity, none of the docking programs or scoring functions made a useful prediction of ligand binding affinity."

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

HOW ACCURATE DOES ONE NEED TO BE TO HAVE AN IMPACT ON DRUG DISCOVERY?



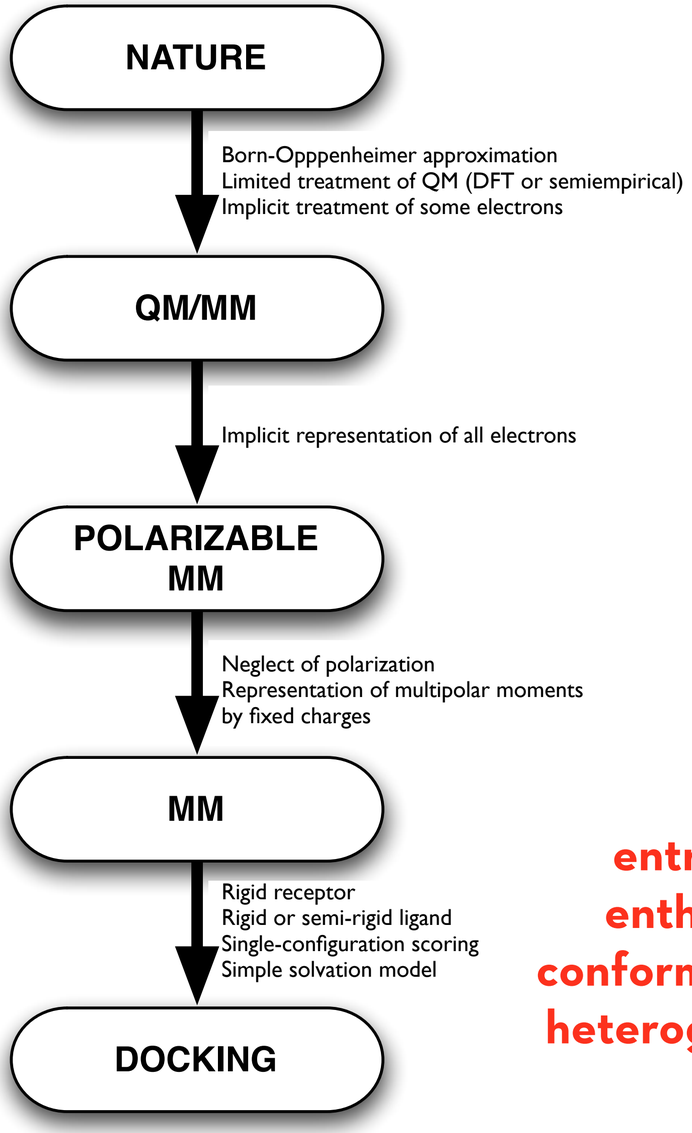




A 2 kcal/mol error in prioritizing lead synthesis would speed lead optimization by 3x but even 10% improvements would be of tremendous benefit

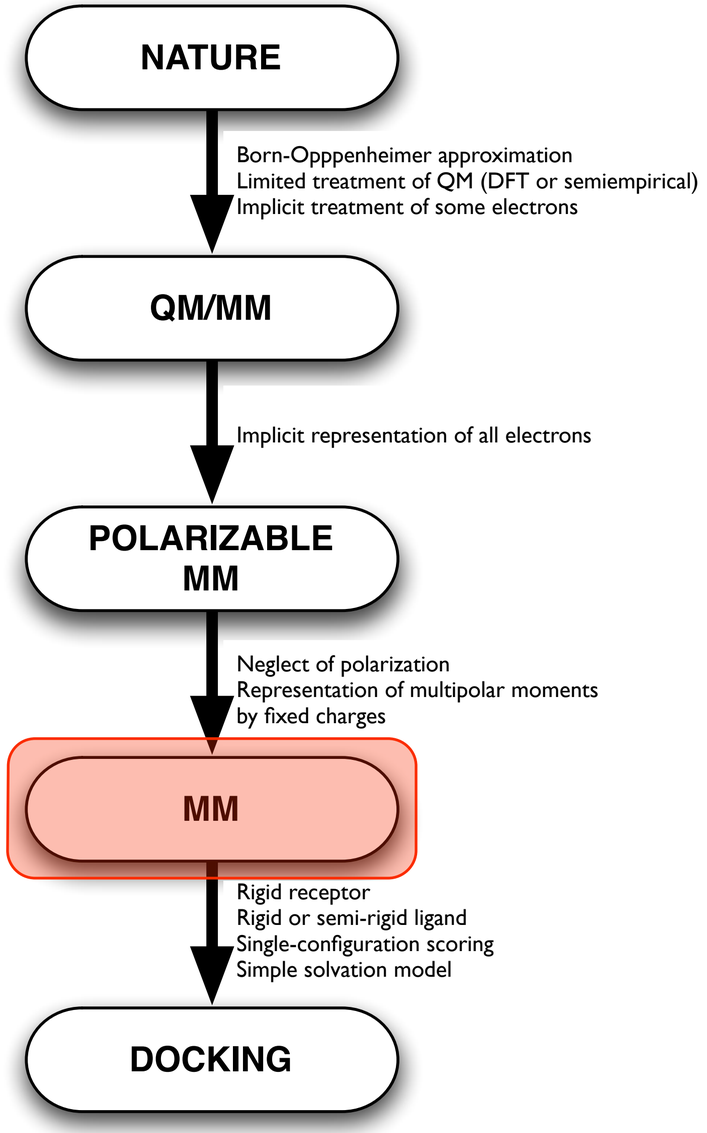
M. R. Shirts, D. L. Mobley and Scott P. Brown. "Free energy calculations in structure-based drug design", in Drug Design: Structure- and Ligand-Based Approaches, pgs. 61-86, 2010.

WHAT DETAILS ARE CRUCIAL FOR ACCURACY?



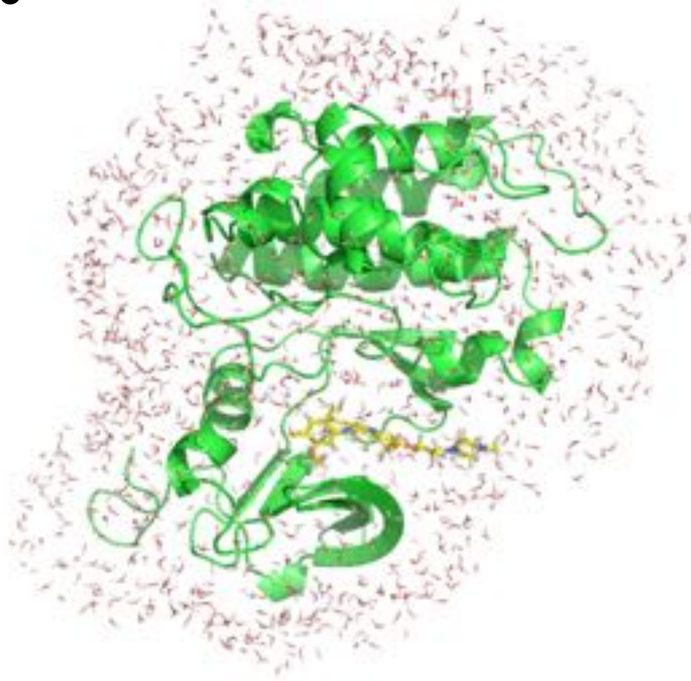
entropy
enthalpy
conformational
heterogeneity

WHAT DETAILS ARE CRUCIAL FOR ACCURACY?



if insufficiently accurate, systematically add detail

if accurate enough, systematically remove detail

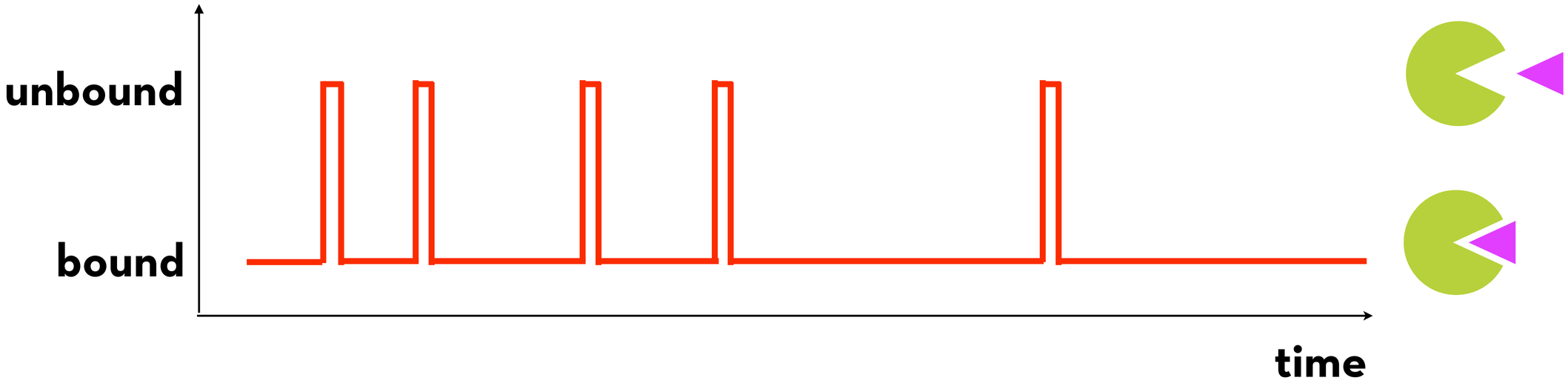


$$V(\mathbf{q}) = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2$$

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

molecular mechanics potential energy

HOW CAN WE COMPUTE A BINDING AFFINITY INCLUDING RELEVANT STATISTICAL MECHANICS?

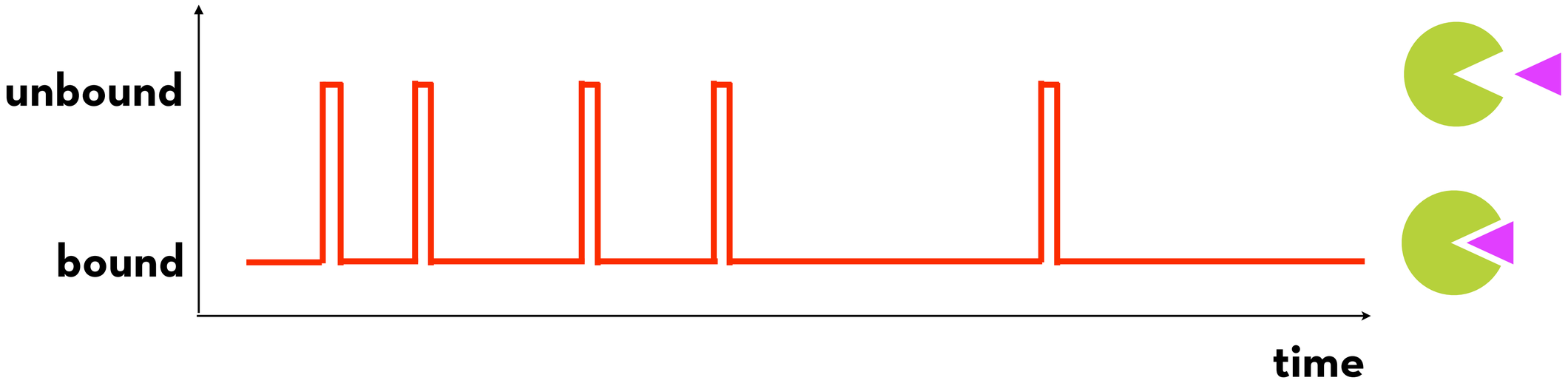


$$\mathrm{P} + \mathrm{L} \overset{K_d}{\rightleftharpoons} \mathrm{PL}$$

dissociation constant

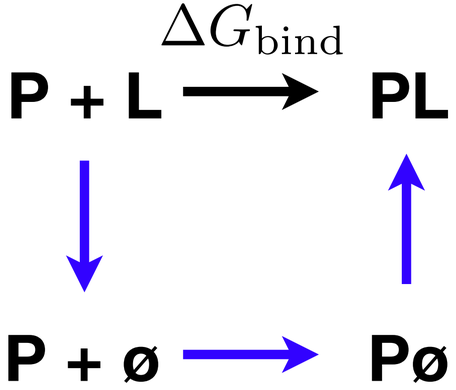
$$K_d \propto \frac{\tau_{\mathrm{unbound}}}{\tau_{\mathrm{bound}}}$$

HOW CAN WE COMPUTE A BINDING AFFINITY INCLUDING RELEVANT STATISTICAL MECHANICS?

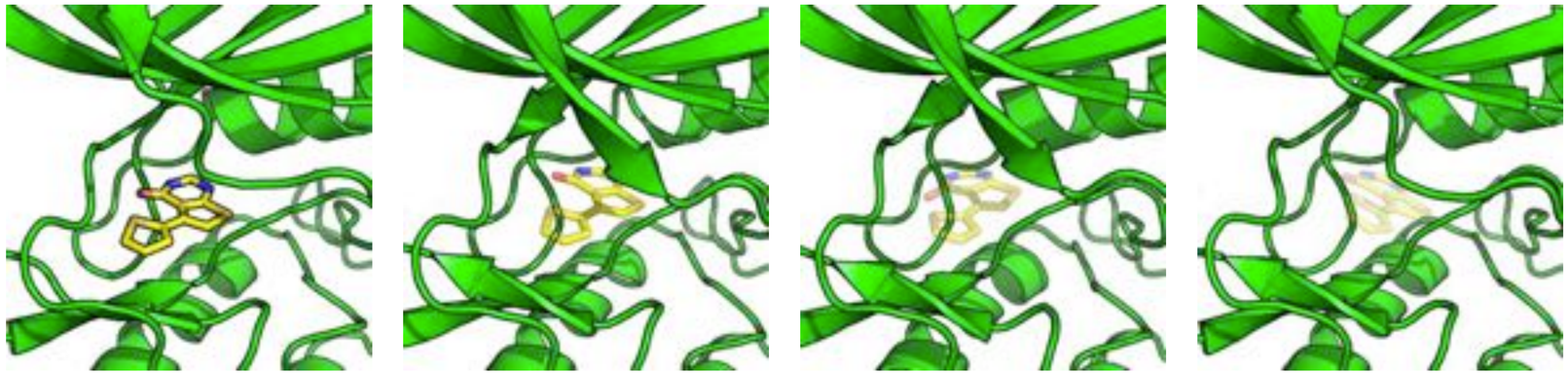


For typical drug off-rates (10^{-4} s^{-1}),
reliable calculation of binding affinities would require hour trajectories,
requiring ~10^{6} years to simulate.

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



multiple simulations of alchemical intermediates



restraint imposition | discharging | steric decoupling | noninteracting

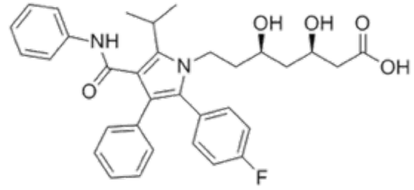
Requires orders of magnitude less effort than simulating direct association process, but still includes all enthalpic/entropic contributions to binding free energy.

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

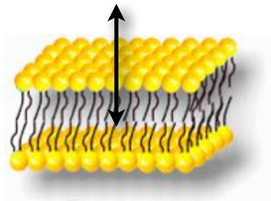
Pioneering work from many: McCammon, van Gunsteren, Kollman, Jorgensen, Chipot, Roux, Boresch, Fujitani, Pande, Shirts, Swope, Christ, Mobley, Schrödinger, and many more
Recent review: Chodera, Mobley, Shirts, Dixon, Branson, Pande. Curr Opin Struct Biol 21:150, 2011.

ALCHEMICAL METHODS CAN ALSO COMPUTE MANY OTHER USEFUL PROPERTIES

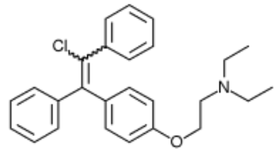
partition coefficients (logP, logD) and permeabilities



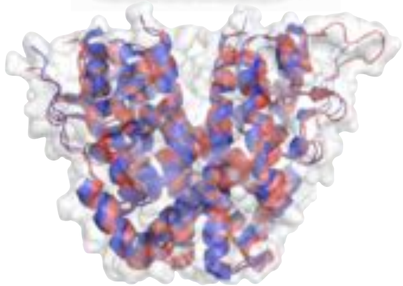
lipitor



selectivity for subtypes or related targets/off-targets



clomifene



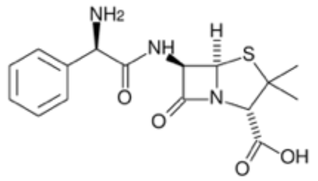
ERα/β

lead optimization of affinity and selectivity

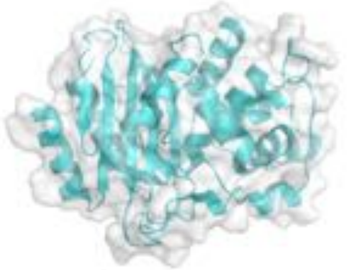
Imatinib

Dasatinib

susceptibility to resistance mutations



ampicillin

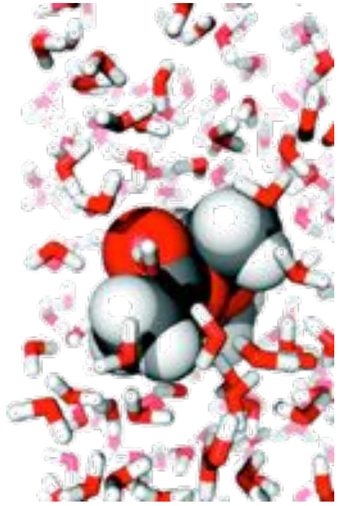


β-lactamase

also solubilities, polymorphs, etc.

ALCHEMICAL FREE ENERGY METHODS CAN WORK RELIABLY IN SIMPLE SYSTEMS, BUT COMPLEX SYSTEMS REMAIN CHALLENGING

model systems ⟷ **pharmaceutically relevant**



hydration free energies

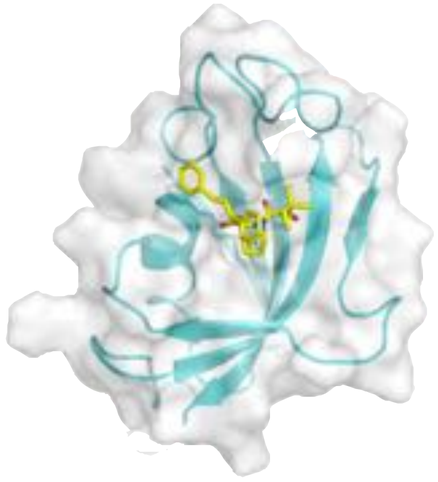
1.04±0.03 kcal/mol (N=44)
Mobley et al. JPC B, 2007
1.23±0.01 kcal/mol (N=502)
Mobley et al. JPC B 2009.

1.33±0.05 kcal/mol (N=17)
Nicholls et al. J Med Chem 2008.

T4 lysozyme L99A

1.89±0.04 kcal/mol (N=13)
Mobley et al. J Mol Biol
371:1118, 2007

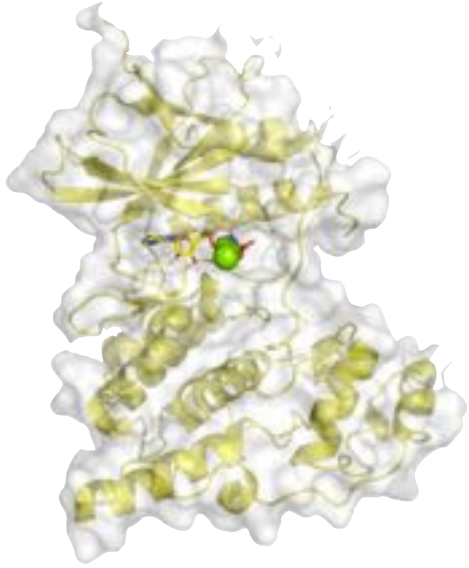
0.6±0.2 kcal/mol (N=3)
Mobley et al. J Mol Biol
371:1118, 2007



FKBP12

0.4 kcal/mol* (N=8)
Fujitani et al.
JCP 123:084108, 2005
*** with 3.2 kcal/mol offset**

. . .



JNK3 kinase

Anecdotal literature reports of success
(publication bias?)

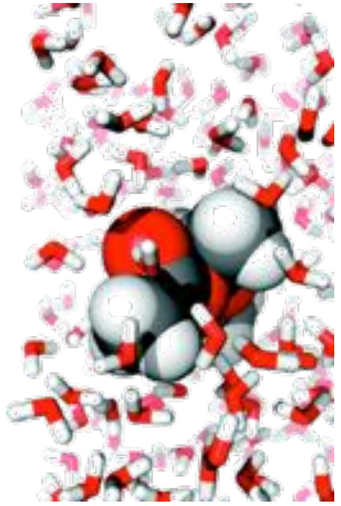
Calculations are notoriously unreliable.
(e.g. SAMPL predictive challenges)

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

(not to scale)

ALCHEMICAL FREE ENERGY METHODS CAN WORK RELIABLY IN SIMPLE SYSTEMS, BUT COMPLEX SYSTEMS REMAIN CHALLENGING

model systems ⟷ pharmaceutically relevant

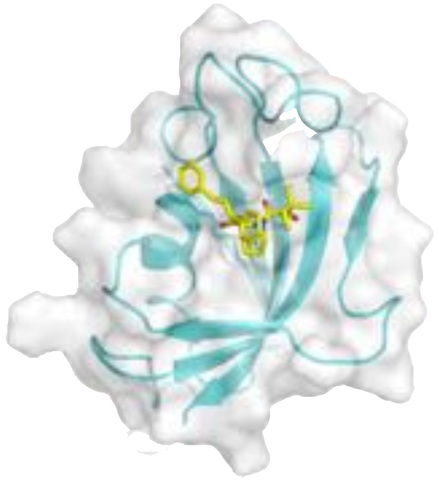


hydration free energies

solvent only
small, neutral molecules
fixed protonation states

T4 lysozyme L99A

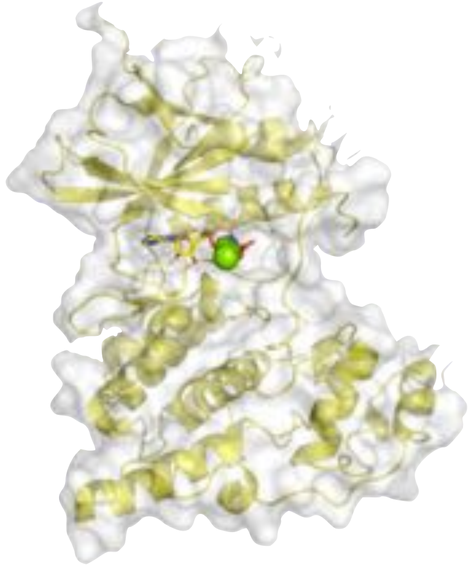
small, rigid protein
small, neutral ligands
fixed protonation states
multiple sidechain orientations
multiple ligand binding modes



FKBP12

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

. . .

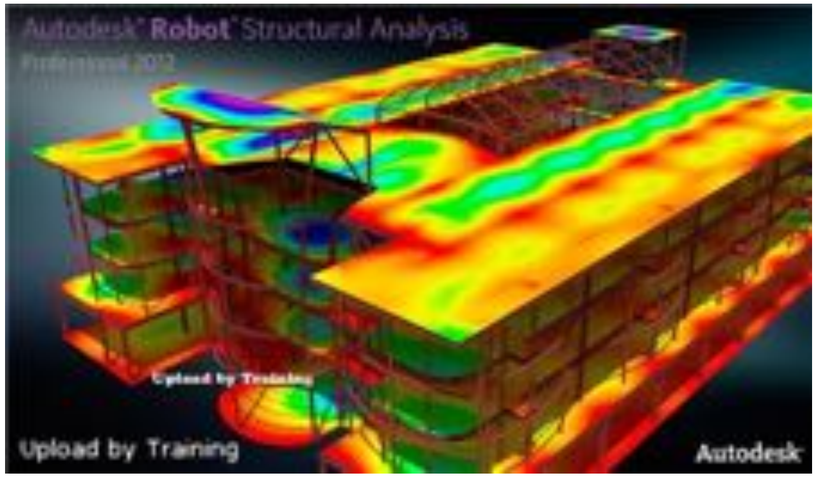


JNK3 kinase

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
Upload by Training
Autodesk

STRUCTURAL ENGINEERING WASN'T ALWAYS SO SUCCESSFUL

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."*

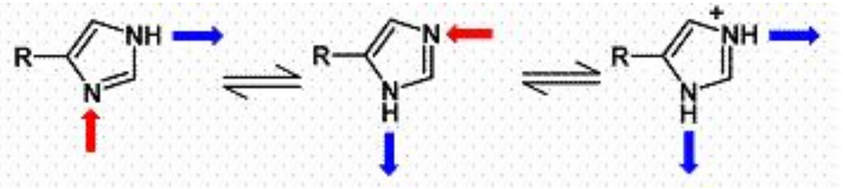
- George Thomson, 1888

PREDICTIONS FAIL FOR THREE REASONS

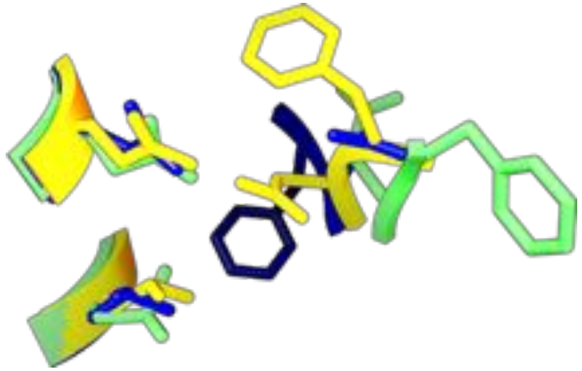
1. The forcefield does a poor job of modeling the physics of our system

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

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



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



WE NEED TO UNDERSTAND WHY FAILURES OCCUR TO IMPROVE THE ROBUSTNESS OF OUR PREDICTIVE MODELS

THE DOMAIN OF APPLICABILITY OF FREE ENERGY CALCULATIONS IS CURRENTLY LIMITED

Multiple high-quality crystal structures of target

Congeneric series of ligands with
all ligands binding in same pose

Only one dominant protonation state
unchanged throughout binding process

No ligand or sidechain tautomerism

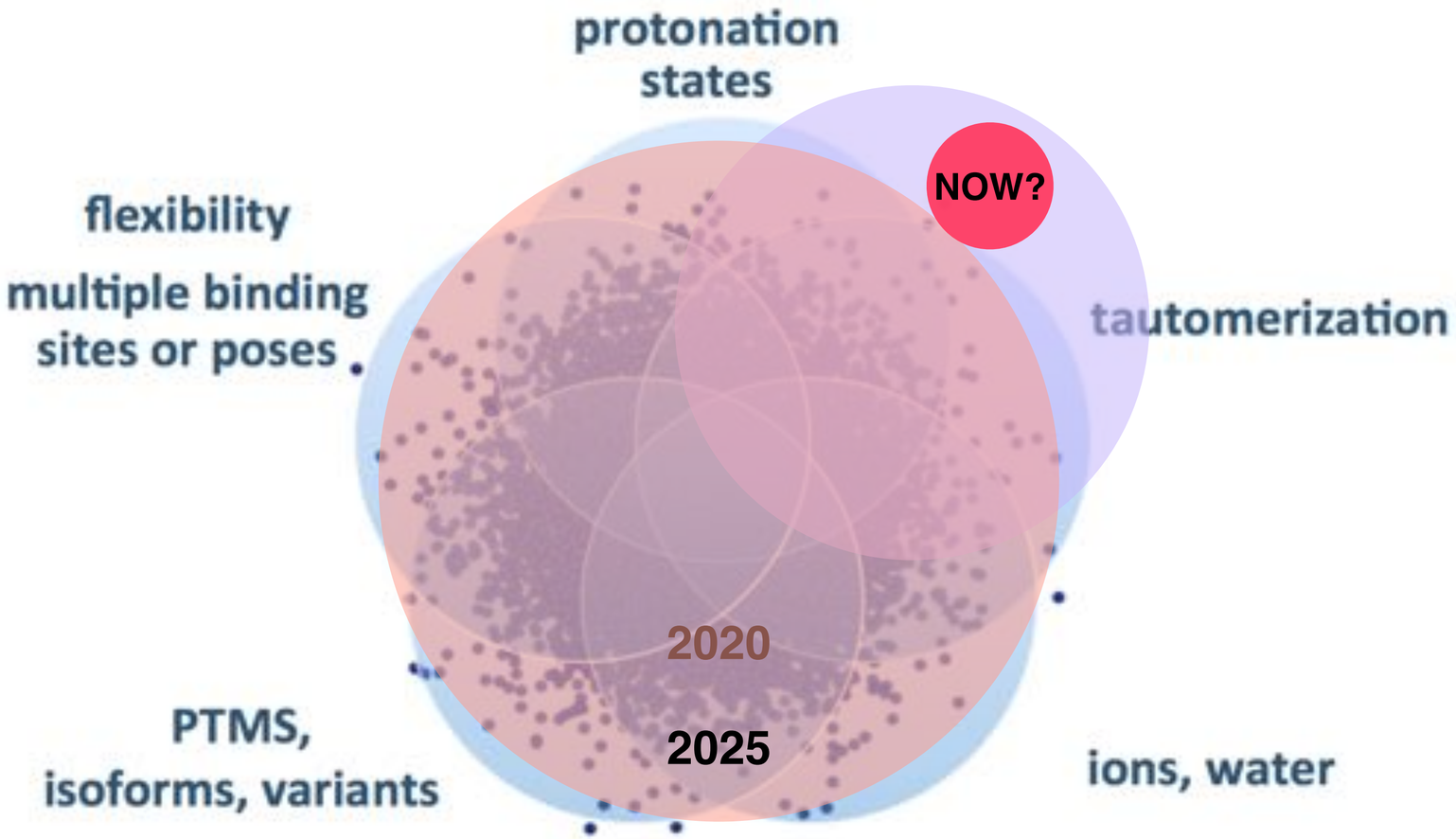
One well-specified, well-resolved isoform/species

No complex cosolvents, binding partners,
slow binding site desolvation events

No exotic chemistries

No metals or prosthetic groups

No membranes?



TARGETS
CHALLENGES

FAIL FAST, FAIL CHEAP

computational predictions

experimental confirmation

HOW CAN WE SPEED UP FREE ENERGY CALCULATIONS?

$65K
9 TFLOP/S

DOUBLING
~18 MONTHS

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 FREE ENERGY CALCULATIONS?

$65K
9 TFLOP/S

DOUBLING EVERY 18 MONTHS

many CPU-weeks/ calculation

$650
9 TFLOP/S

BEATING MOORE'S LAW?

overnight on a workstation?

WE CAN EXPLOIT NEW GPU TECHNOLOGIES TO REACH PRACTICAL COMPUTATION TIMES

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

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

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

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

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

Docs » YANK

YANK

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

Features

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

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

http://www.getyank.org

Omnia
empowering molecular simulation

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

OpenMM

PDB FIXER

MSM BUILDER

PyEMMA

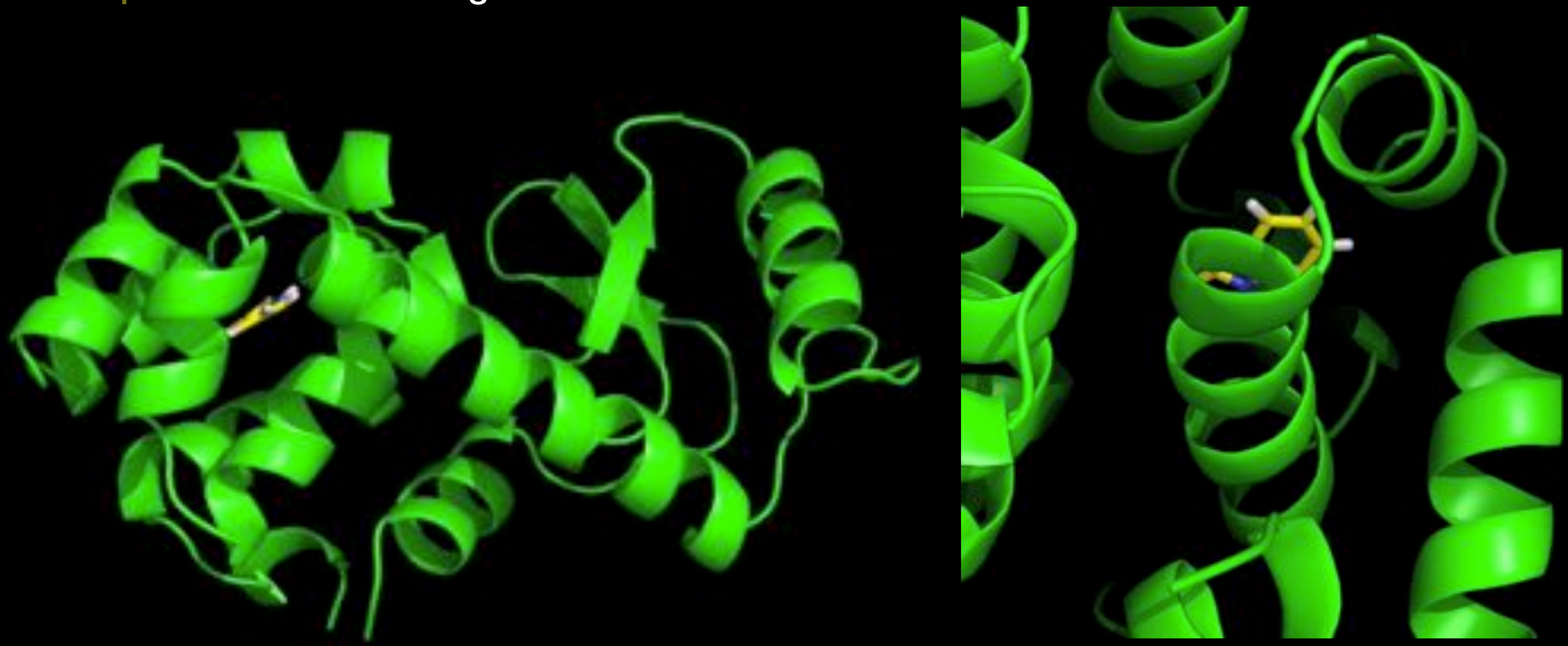
MDTRAJ

ENSEMBLER

http://omnia.md

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

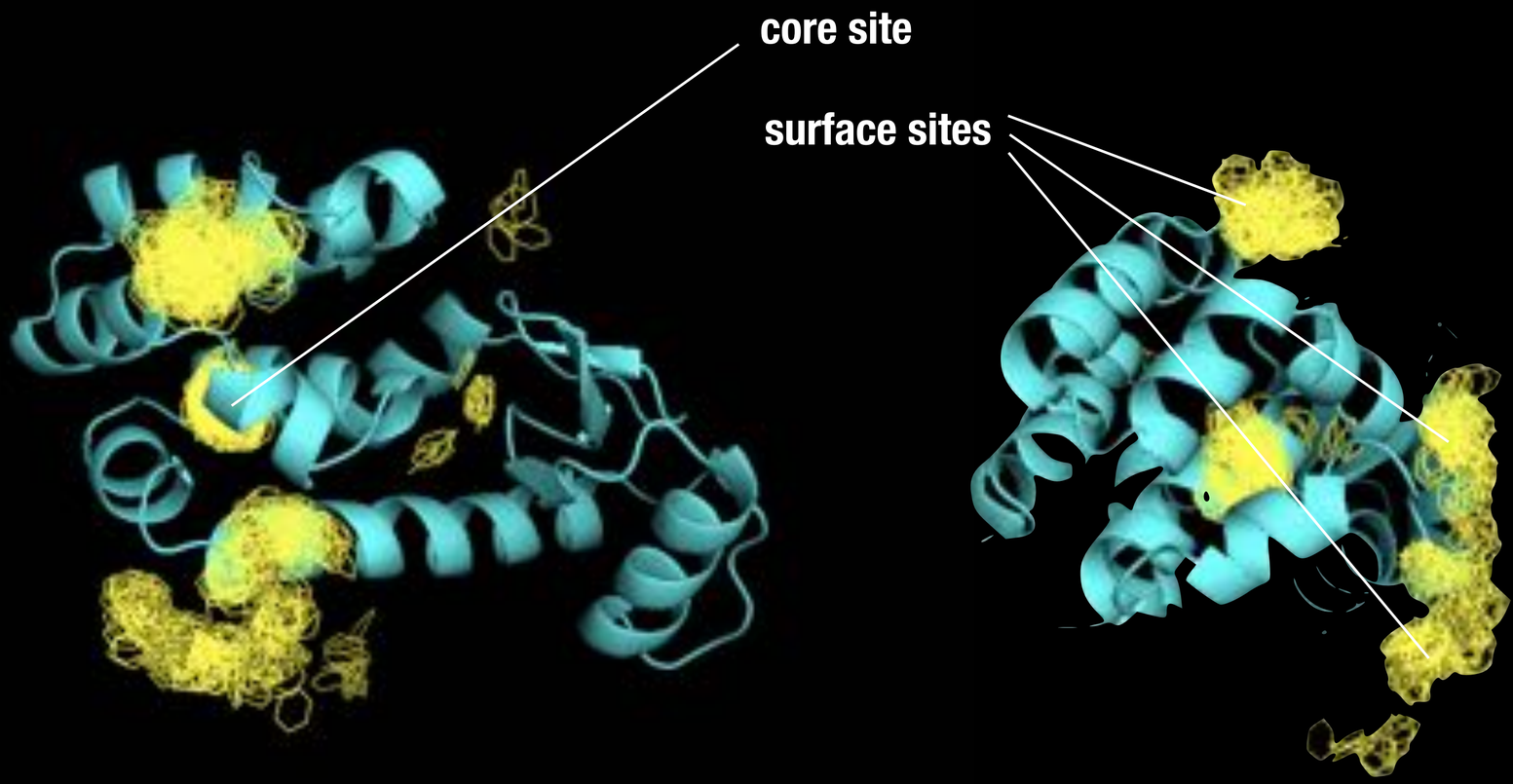
solid fully interacting
transparent noninteracting



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

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

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

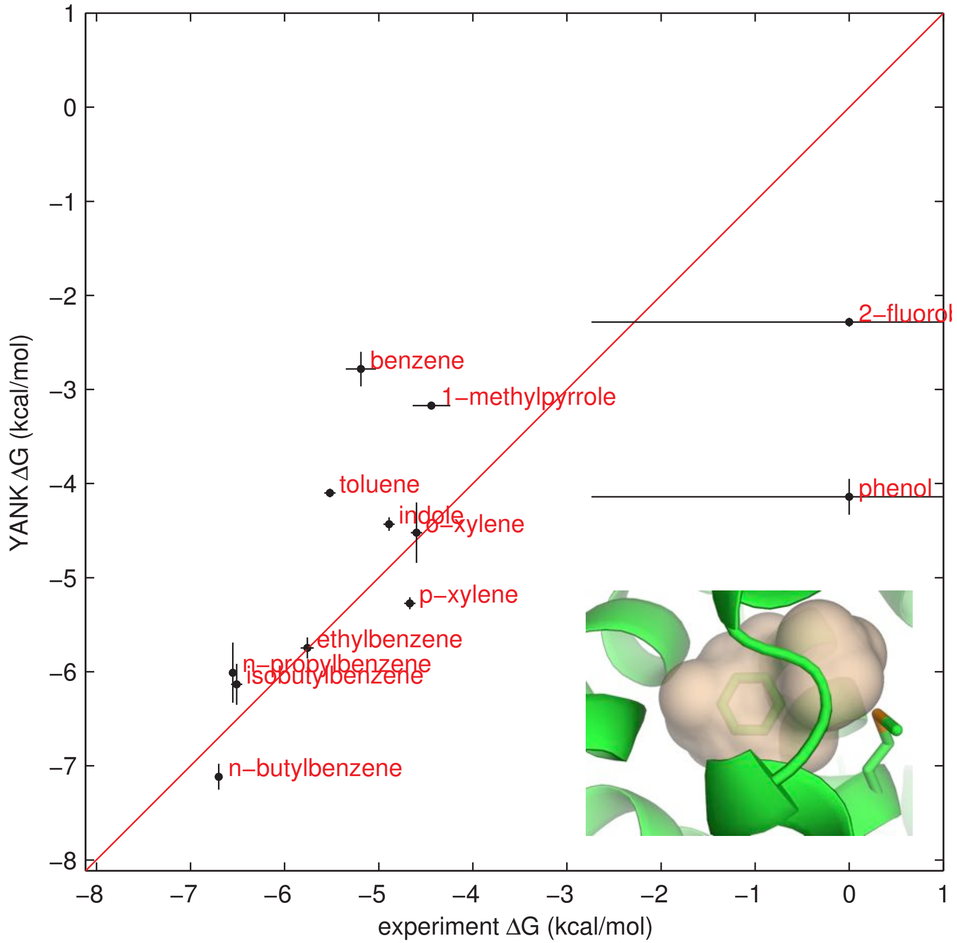


benzene bound to T4 lysozyme L99A
AMBER96 + OBC GBSA

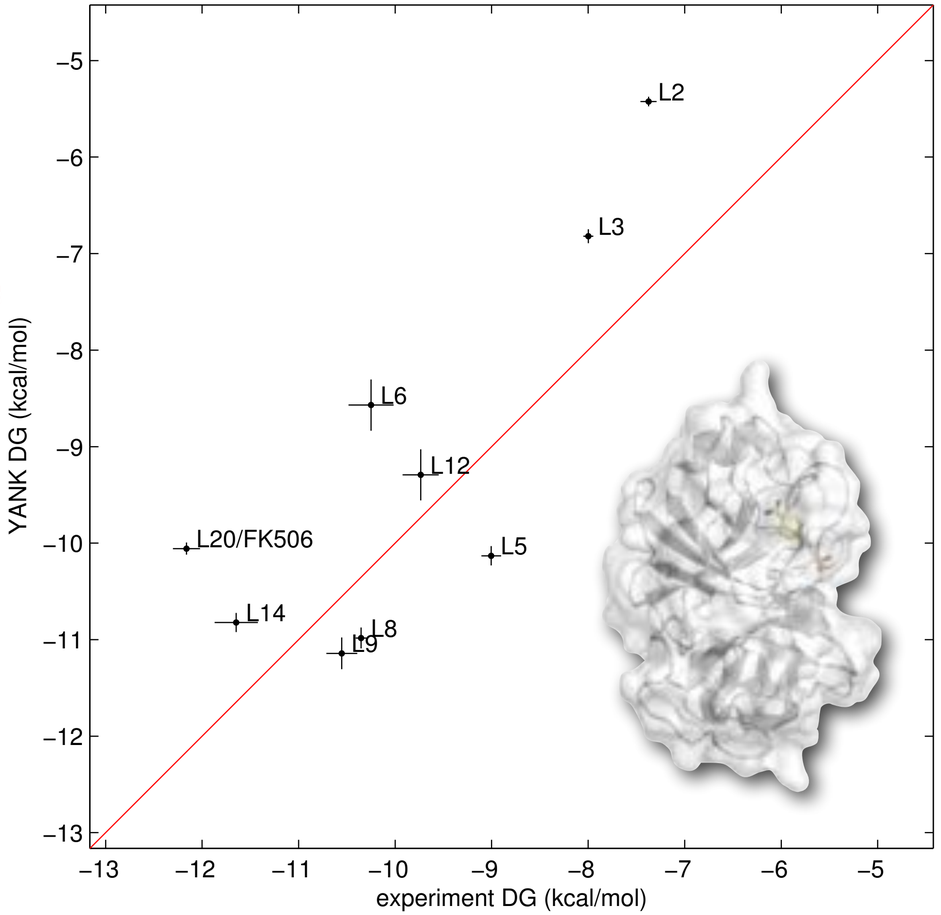
Chodera and Shirts. JCP 135:194110, 2011
Wang, Chodera, Yang, and Shirts. JCAMD 27:989, 2013.
http://github.org/choderalab/yank

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

T4 LYSOZYME L99A



FKBP12



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

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

YANK ROADMAP

Q4 2016
- YANK 1.0 PREVIEW RELEASE OUT
- TUTORIALS / BEST PRACTICES / DOCS / TESTS

Q1 2017
- STABLE PYTHON API
- ROBUST WORKFLOW PIPELINE
- SINGLE REPLICA CALCULATIONS

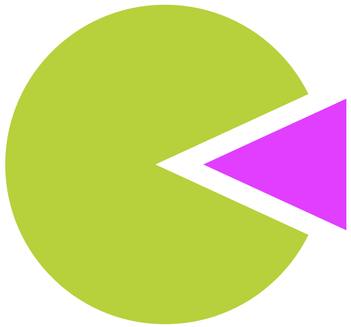
Q2 2017
- RELATIVE FREE ENERGY CALCULATIONS
- MULTIPLE FORCEFIELD SUPPORT
- MULTIPLE ALCHEMICAL REGIONS
- DYNAMIC PROTOMERS, TAUTOMERS, COUNTERIONS

Q3 2017
- PERSES AUTOMATED DESIGN
- SIMULTANEOUS KINETICS AND FREE ENERGIES

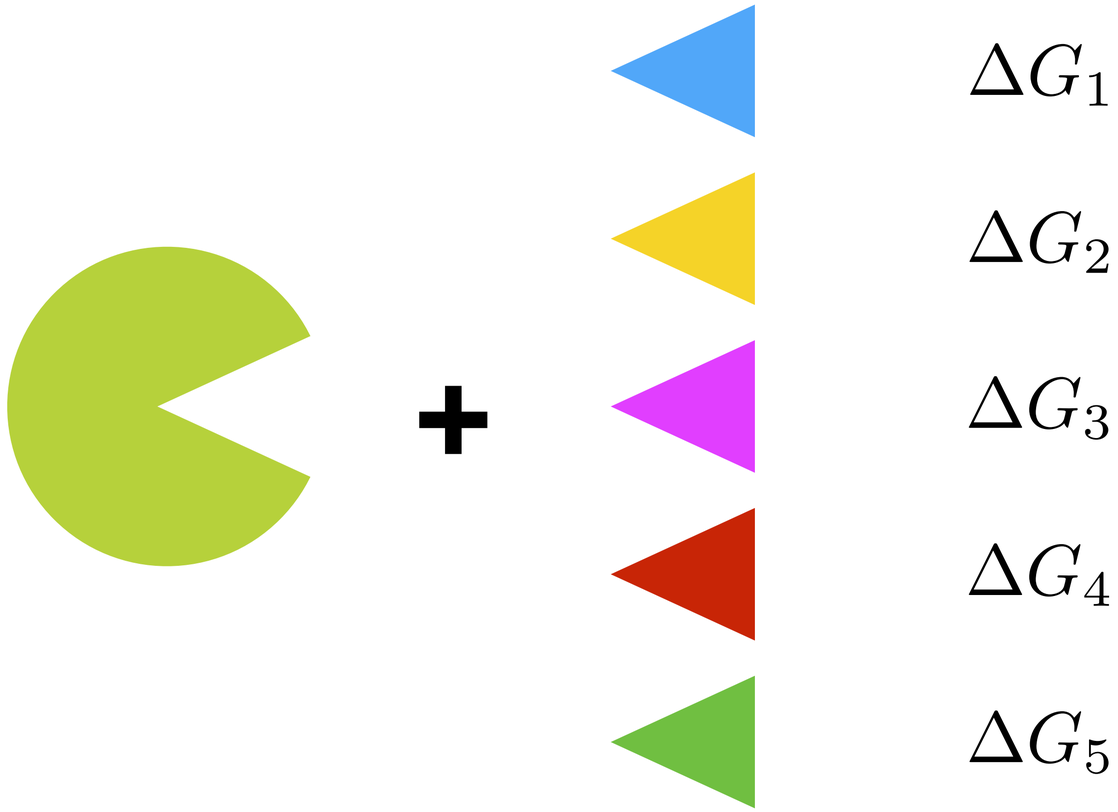
FAIL FAST, FAIL CHEAP

computational predictions

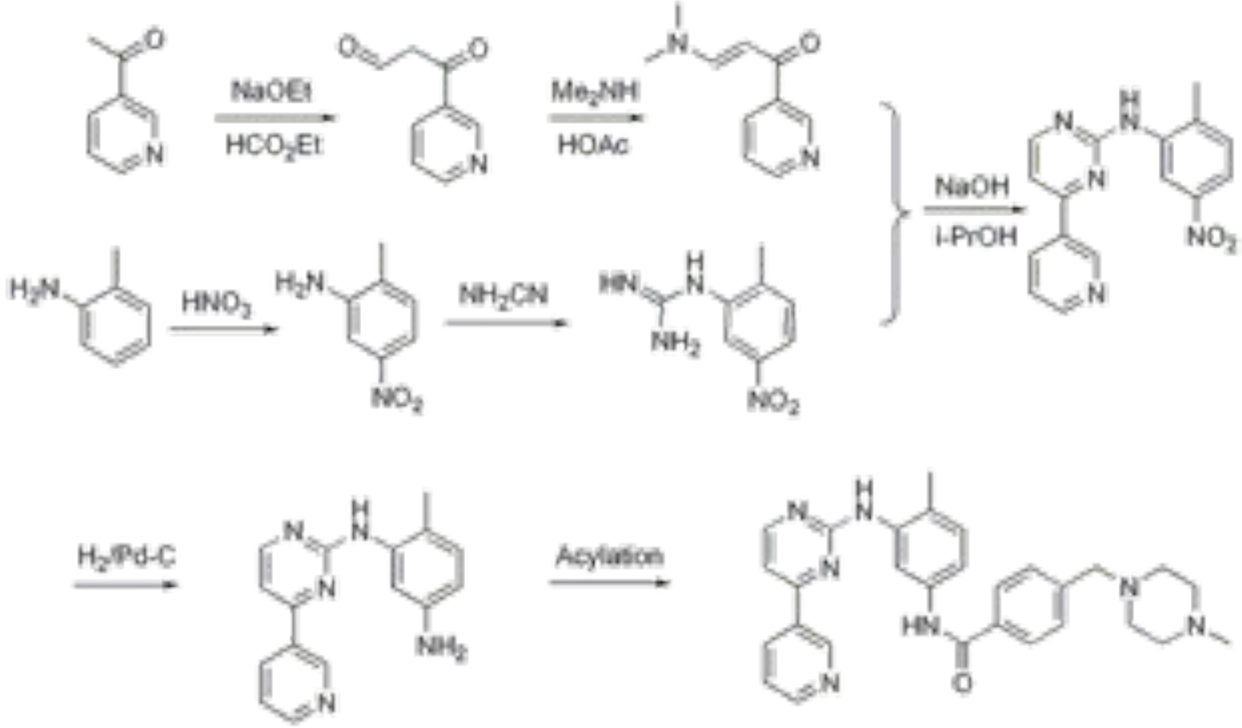
experimental confirmation



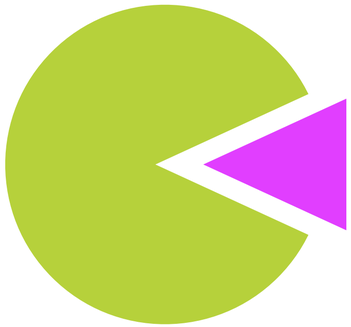
1. PREDICT HOW MODIFICATIONS TO LIGAND WILL CHANGE AFFINITY
2. TEST EXPERIMENTALLY



SYNTHESIS OF NEW COMPOUNDS TO TEST HYPOTHESES IS EXPENSIVE AND TIME-CONSUMING



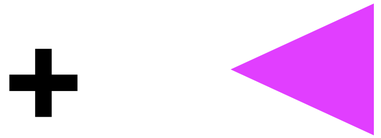
SYNTHESIS OF IMATINIB



1. PREDICT HOW MODIFICATIONS TO PROTEIN WILL CHANGE AFFINITY
2. TEST EXPERIMENTALLY

ΔG_1

ΔG_2



ΔG_3

INVERTING THE DRUG DISCOVERY PROBLEM ALLOWS US TO FAIL QUICKLY AND CHEAPLY

quick-change mutagenesis

expression assessment in 1 ml culture

expression and purification in ~2L culture

\+ purchase known ligands

ISOTHERMAL TITRATION CALORIMETRY

SURFACE PLASMON RESONANCE

FLUORESCENCE BINDING ASSAYS

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

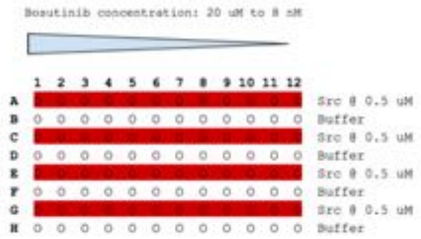
precise
structured
consistent
reproducible
round-the-clock

AUTOMATE. EVERYTHING.

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

ASSAY AUTOMATION CAN CONTROL ERROR

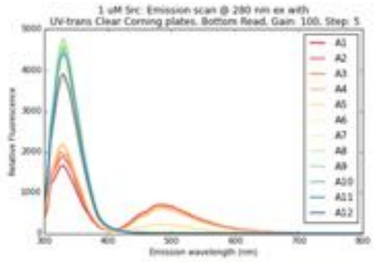
Dispense



Shake/Mix

Spin

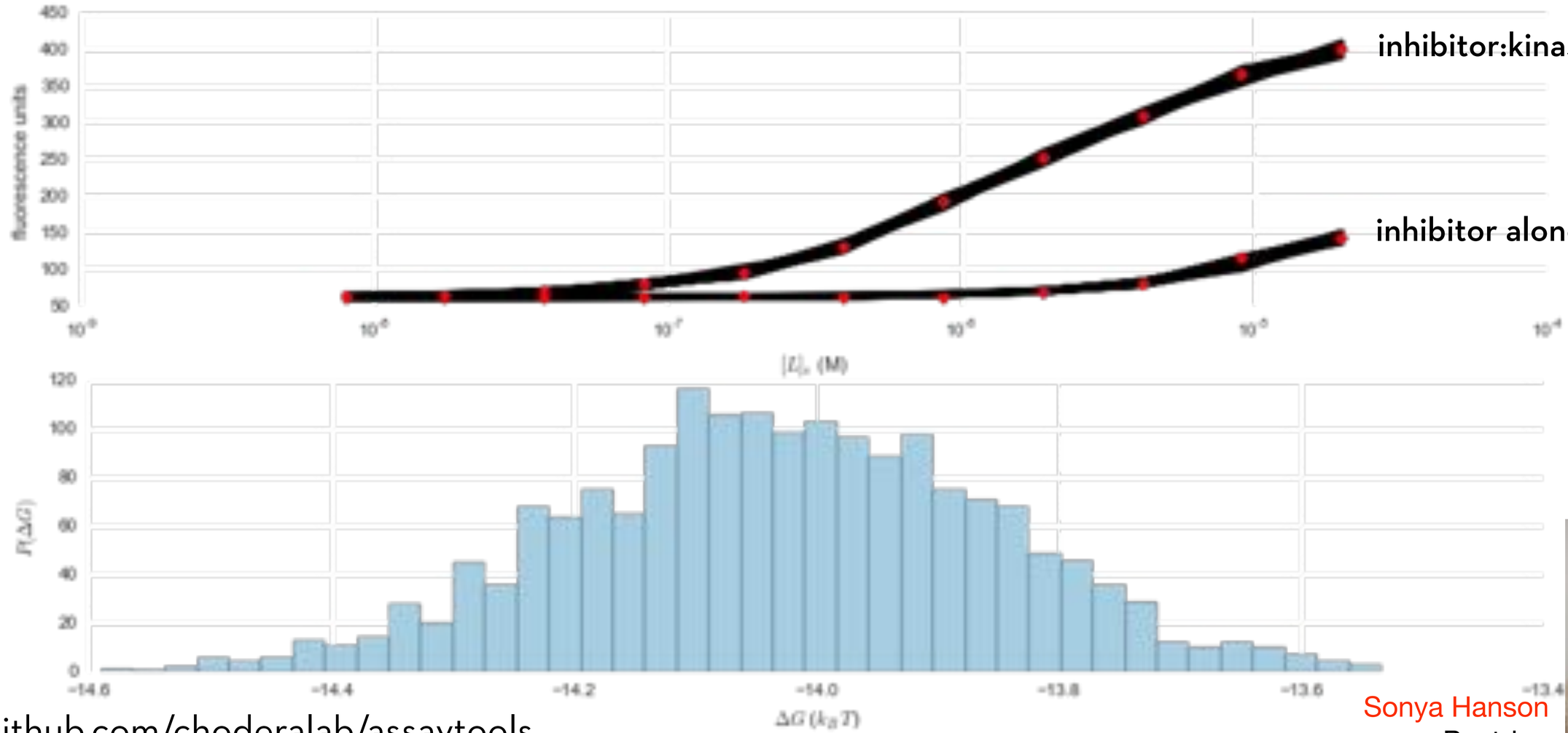
Read Fluorescence



TECAN.

BAYESIAN INFERENCE ALLOWS US TO QUANTIFY EXPERIMENTAL UNCERTAINTY

```
# Sample with MCMC
mcmc = pymc.MCMC(pymc_model, db='ram', name='Sampler', verbose=True)
mcmc.sample(iter=100000, burn=10000, thin=50, progress_bar=False)
```



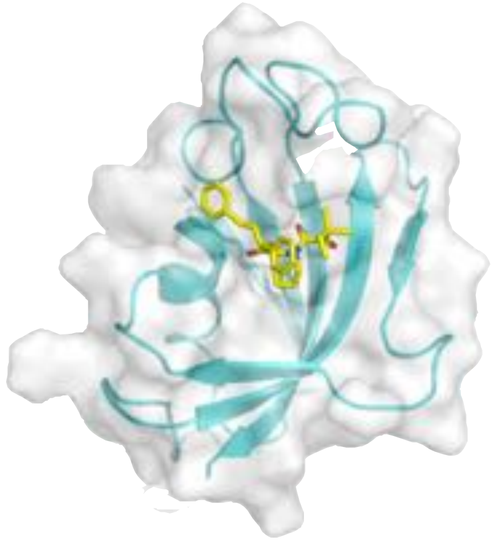
https://github.com/choderalab/assaytools

Sonya Hanson
Postdoc

MODEL SYSTEMS CAN TEACH US VALUABLE LESSONS

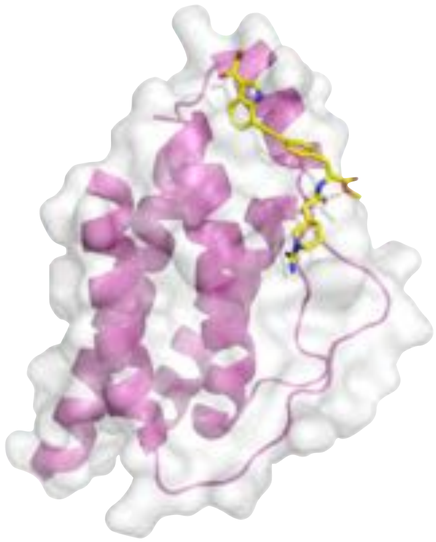
T4 LYSOZYME L99A

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



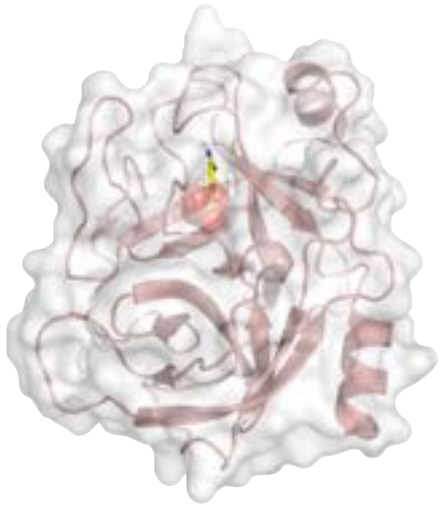
FKBP-12

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



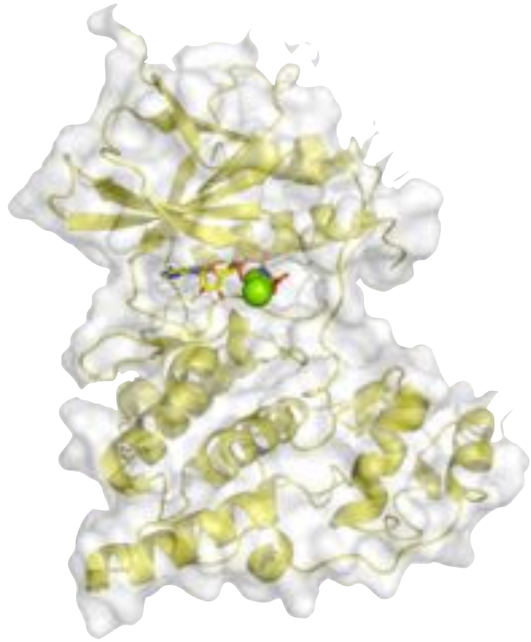
IL-2

small protein
fixed protonation states
some allostery and
binding site plasticity



TRYPSIN

small, rigid protein
small ligands
charged ligands
protonation state changes



KINASES

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

easy
hard

WHERE DO MODEL SYSTEMS COME FROM?

- Word of mouth (“Hey, you should really look at aspartyl proteases…”)
- My old advisor worked on this (T4 lysozyme mutants)
- I got the plasmid from the lab down the hall (chicken Src)
- Everybody else is working on it! (Abl)

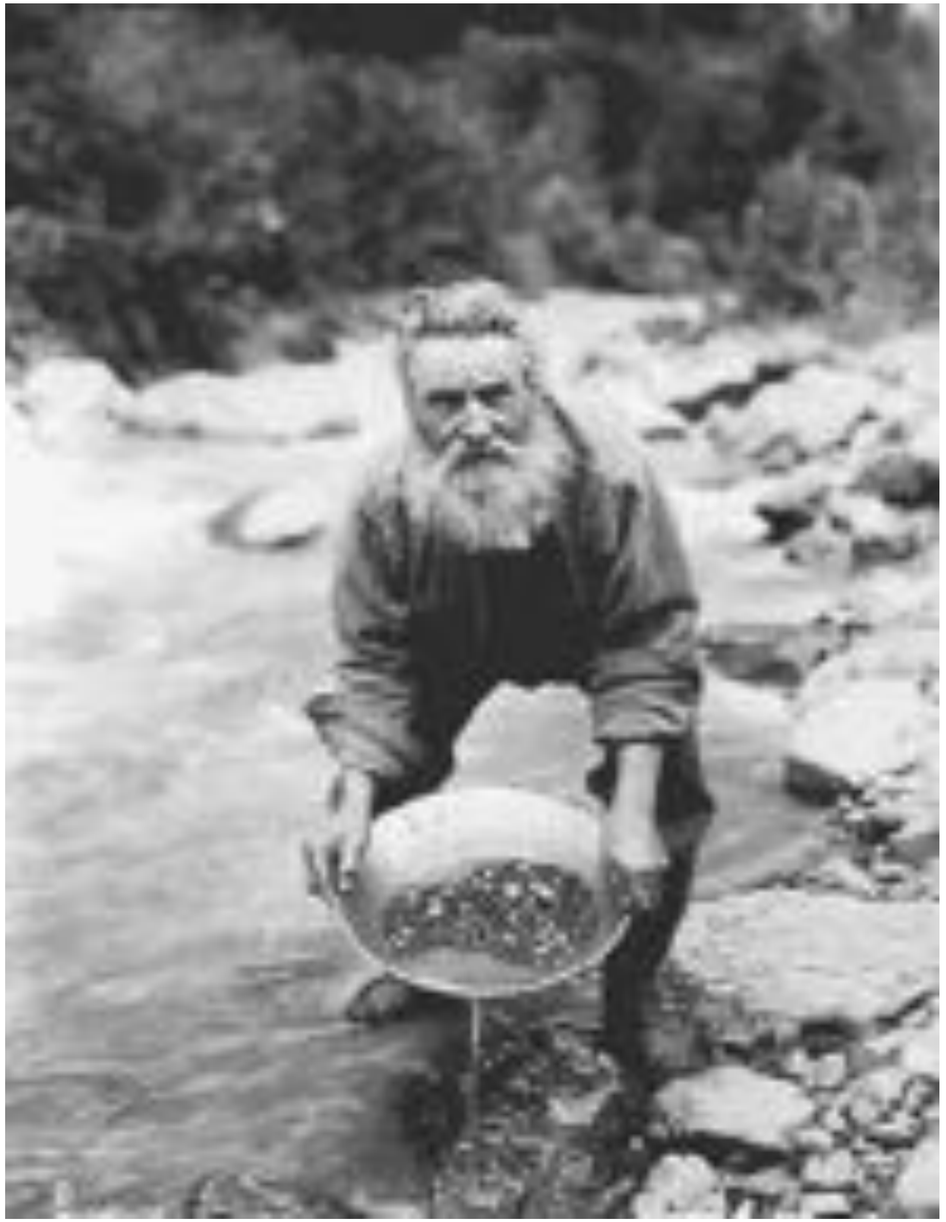
SURELY THERE MUST BE A BETTER WAY!

CAN WE MINE PUBLIC DATASETS FOR GOOD MODEL SYSTEMS?

Desiderata:

- good bacterial expression (for cheap protein production)
- multiple structures available in PDB
- a variety of known ligands available for purchase
- large dynamic range of binding affinities (>3 kcal/mol)
- accessibility to biophysical assays (fluorescence, SPR, ITC)
- known point mutants (e.g. UniProt)
- disease relevance (for funding!)
- properties characteristic of real challenging targets

PANNING FOR MODEL SYSTEMS



initial set of UniProt IDs

retrieve all UniProt metadata

retrieve all known structures and ligands

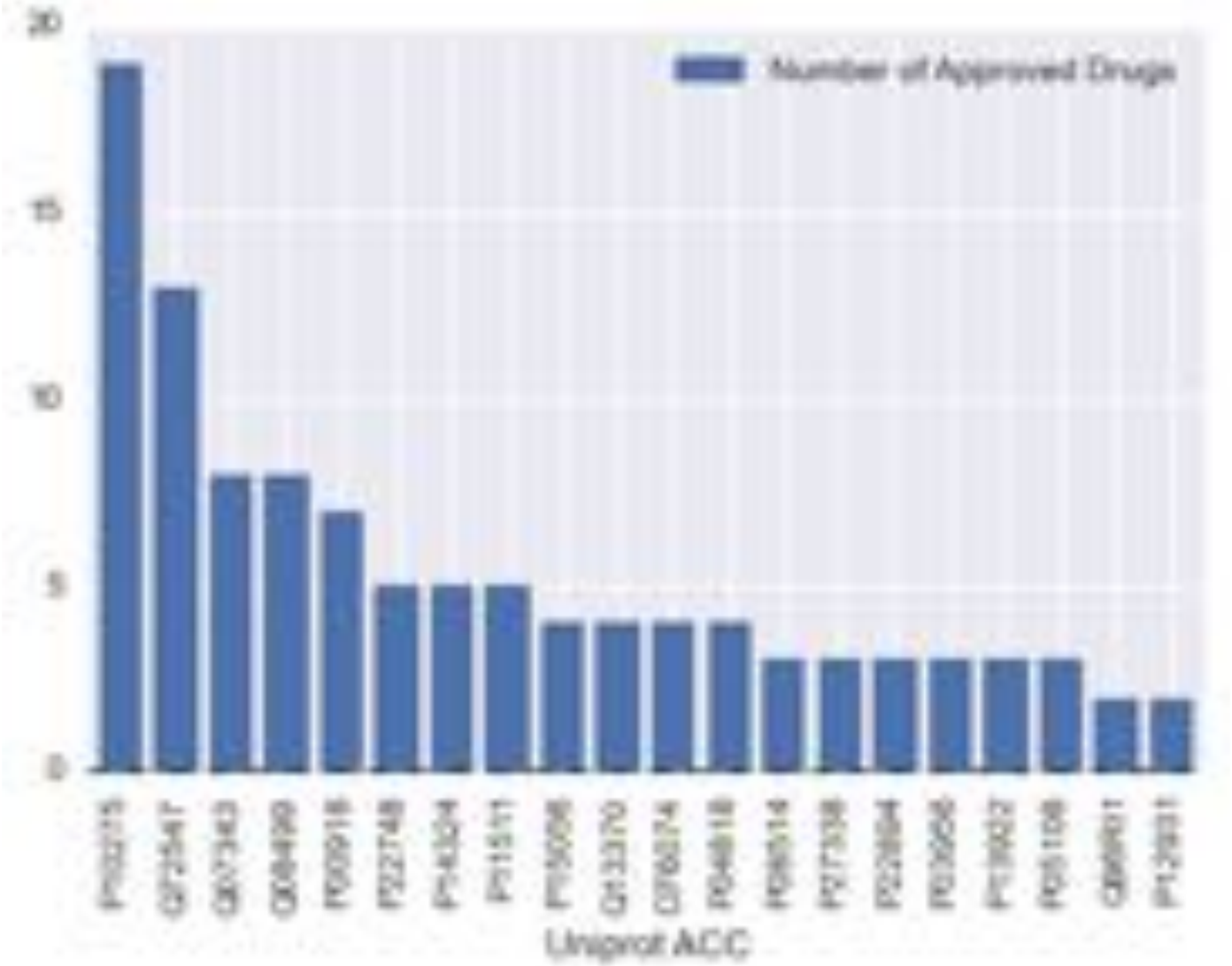
retrieve additional data

filter by criteria of interest

Mehtap Isik

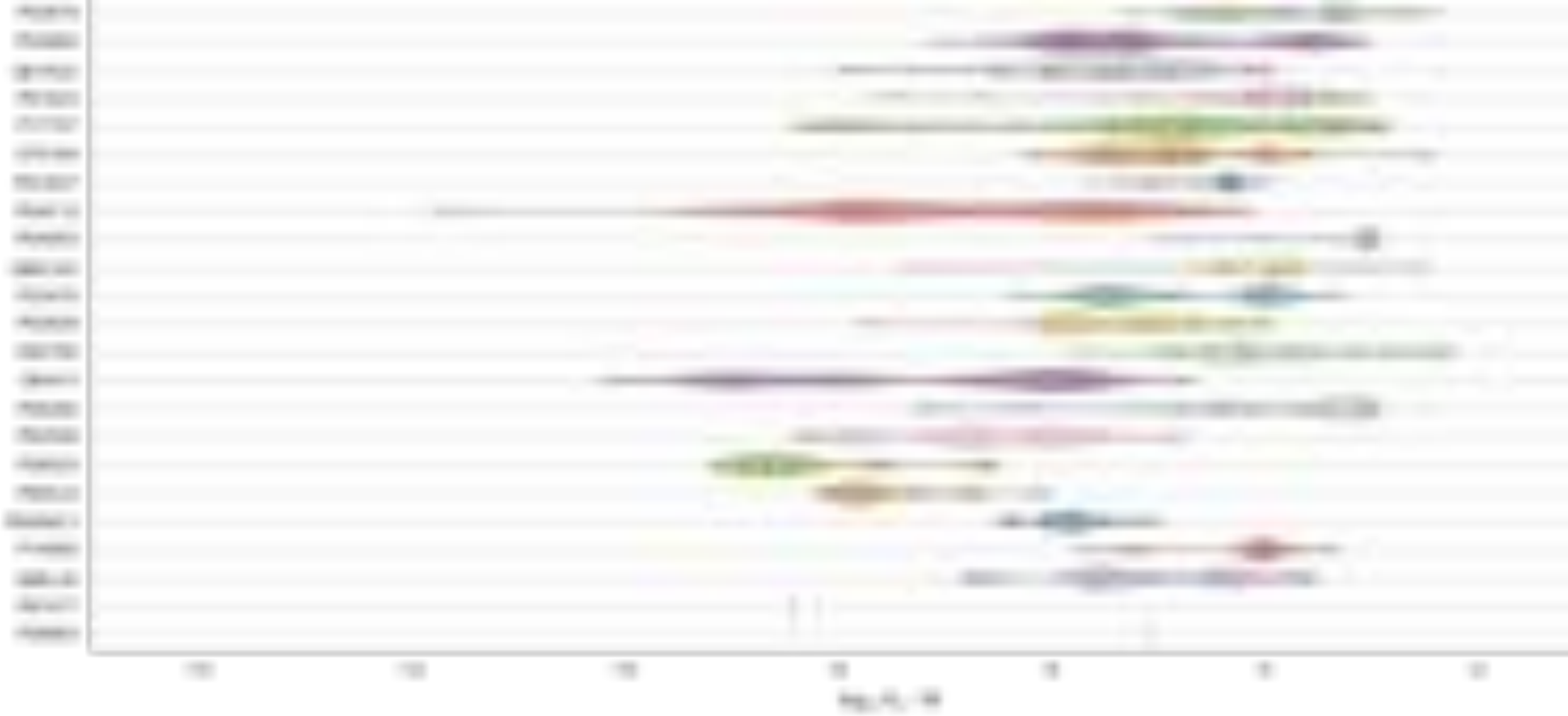
Sonya Hanson

SOME TARGETS HAVE BIOASSAY DATA FOR MULTIPLE FDA-APPROVED DRUGS



Mehtap Isik

Sonya Hanson



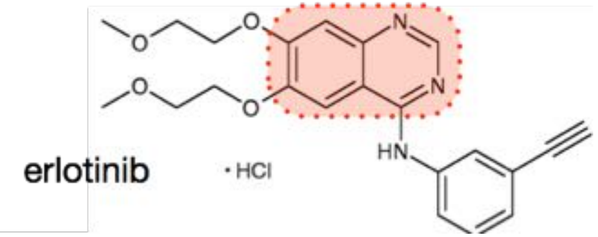
Mehtap Isik

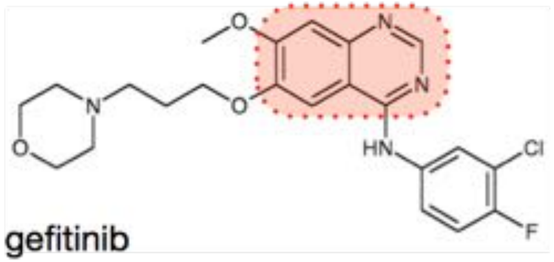
Sonya Hanson

Many targets have usefully large dynamic ranges of known affinities

CAN WE SEARCH FOR POTENTIAL FLUORESCENT PROBE COMPOUNDS?

Quinazoline scaffolds are often fluorescent



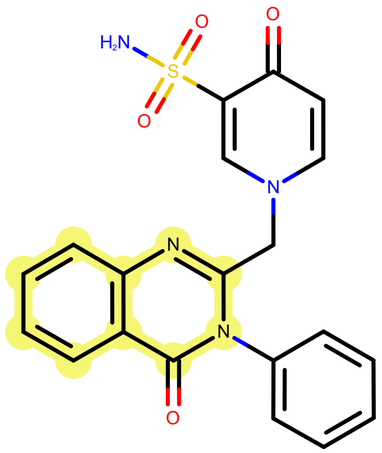


...which can be expressed as a SMARTS query

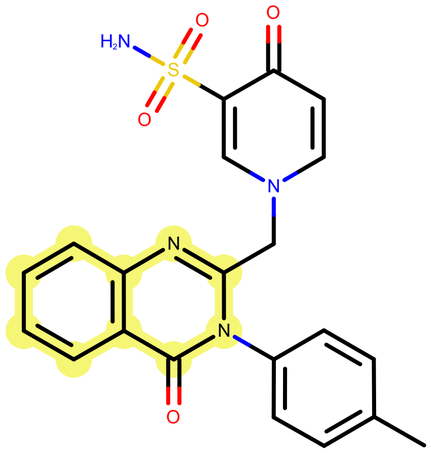
```
c1cccc2c1cncn2
```

Thanks OpenEye!

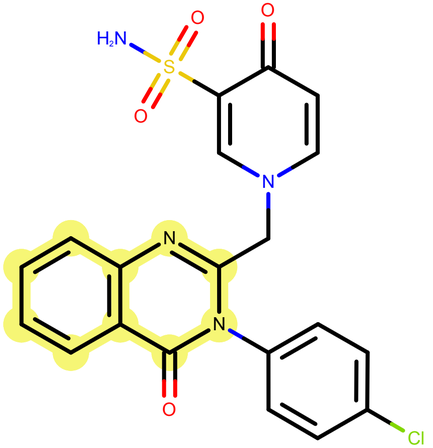
...and used to find some quinazoline scaffold inhibitors of Uniprot P00918 (carbonic anhydrase II) to serve as probes:



567 nM



490 nM



700 nM

CAN WE EXPAND THIS SEARCH TO ALL KNOWN FLUORESCENT SCAFFOLDS?

MANY OF SYSTEMS CAN BE EXPRESSED BY ROBOTS USING A STANDARD PROTOCOL

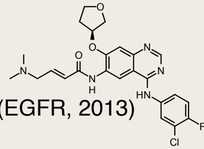
ID	Gene	Protein	PCR Result	Expression test	ug/mL culture	Tag (Best)
1	Abl1	Abl kinase	yes	NO		N/A
2	AK1	adenylate kinase 1	yes	High	121	SUMO
3	AK2	adenylate kinase 2	yes	High	121	H6
4	AMPC_WT	AmpC beta-lactamase	NO	---		---
5	CAH2	carbonic anhydrase II	yes	High	198	H6
6	CALM1	calmodulin	yes	High	107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

Sarah Boyce

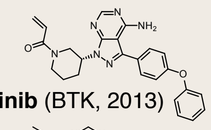
BioMek FXP
(QB3 MacroLab)

FDA APPROVED

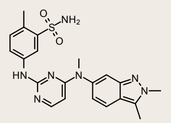
* * * KINASE INHIBITORS * * *



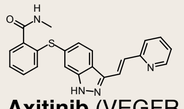
Afatinib (EGFR, 2013)



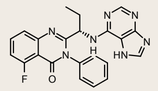
Ibrutinib (BTK, 2013)



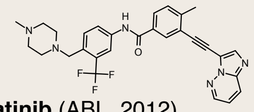
Pazopanib (VEGFR, 2009)



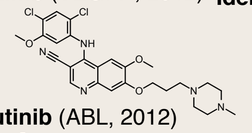
Axitinib (VEGFR, 2012)



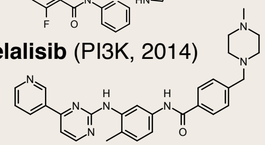
Idelalisib (PI3K, 2014)



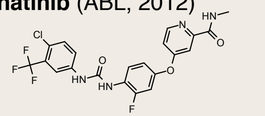
Ponatinib (ABL, 2012)



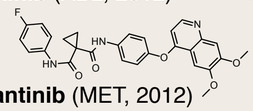
Bosutinib (ABL, 2012)



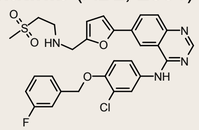
Imatinib (ABL, 2001)



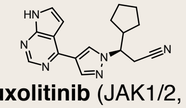
Regorafenib (VEGFR, 2012)



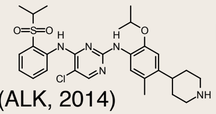
Cabozantinib (MET, 2012)



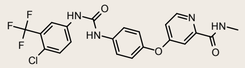
Lapatinib (HER2, 2007)



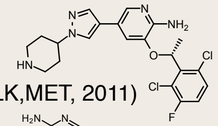
Ruxolitinib (JAK1/2, 2011)



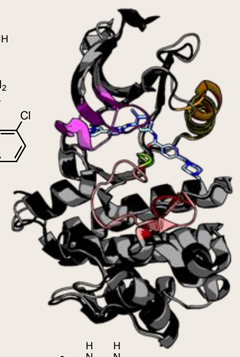
Ceritinib (ALK, 2014)

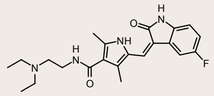


Sorafenib (VEGFR, 2005)

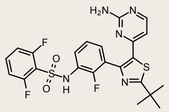


Crizotinib (ALK,MET, 2011)

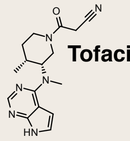




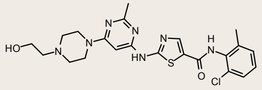
Sunitinib (VEGFR, 2006)



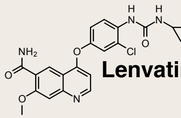
Dabrafenib (BRAF, 2013)



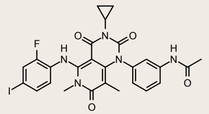
Tofacitinib (JAK3, 2012)



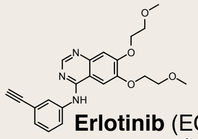
Dasatinib (ABL, 2006)



Lenvatinib (VEGFR, 2015)

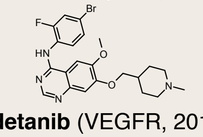


Trametinib (MEK, 2013)

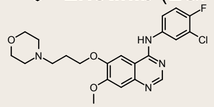


Erlotinib (EGFR, 2005)

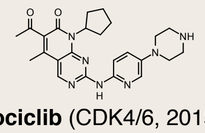
Nilotinib (ABL, 2007)



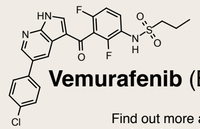
Vandetanib (VEGFR, 2011)



Gefitinib (EGFR, 2003)



Palbociclib (CDK4/6, 2015)



Vemurafenib (BRAF, 2011)

Find out more at: accessdata.fda.gov
and dx.doi.org/10.1021/cb500129t

Last updated: June 1, 2015

CHODERA LAB // MSKCC

WHAT MATTERS?

NEED SENSITIVITY ANALYSIS

- KINASE CONFORMATION
- PROTONATION STATES
  - KINASE
  - INHIBITOR
- SALT ENVIRONMENT
- SOLVENT MODEL
- ELECTROSTATIC TREATMENT
- FORCEFIELD
- KINASE PHOSPHORYLATION STATE

SONYA HANSON

JULIE BEHR

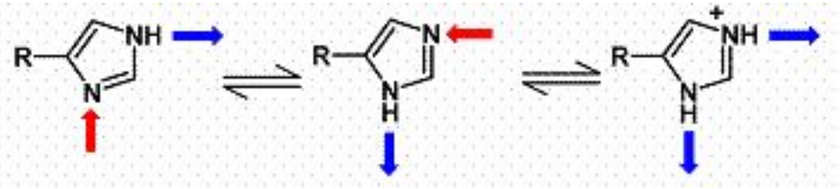
ANDREA RIZZI

PREDICTIONS FAIL FOR THREE REASONS

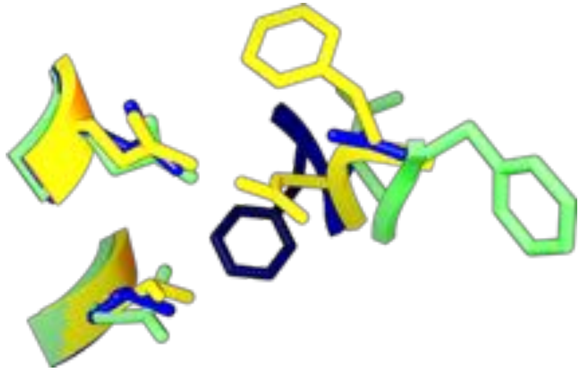
1. The forcefield does a poor job of modeling the physics of our system

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

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



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



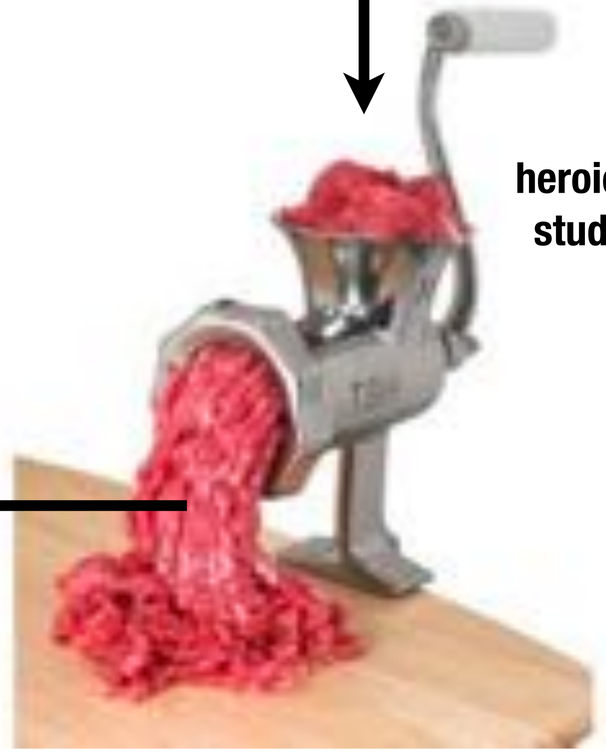
PREDICTIONS FAIL FOR THREE REASONS

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

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

HOW ARE FORCEFIELDS MADE?

experimental data
quantum chemistry
keen chemical intuition

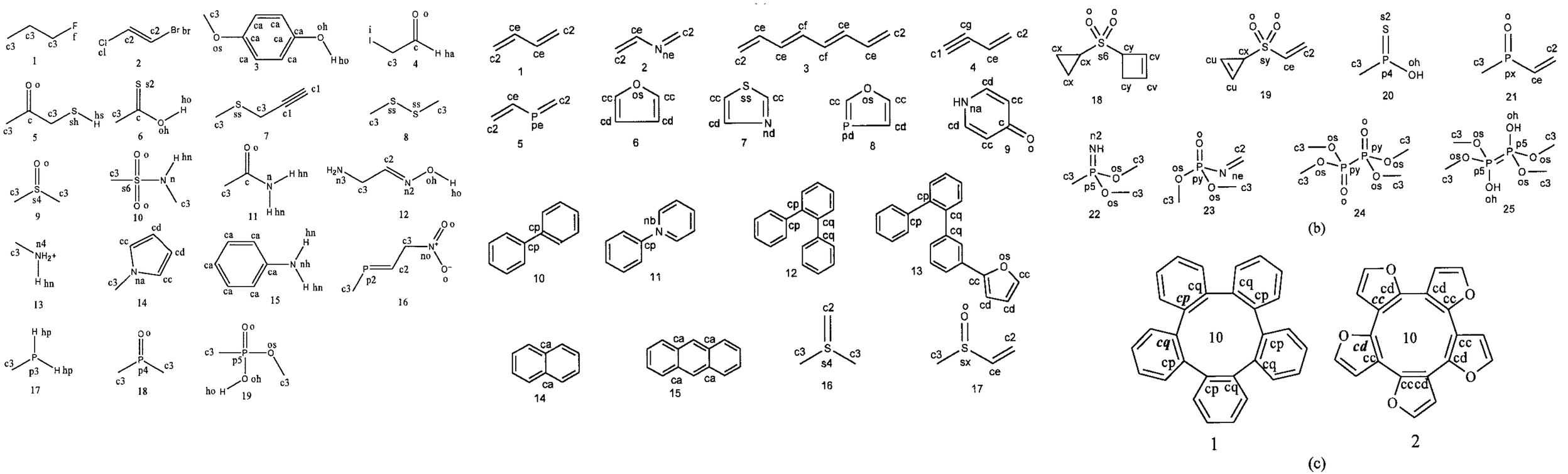


heroic effort by graduate students and postdocs

a parameter set we desperately hope someone actually uses

AS DRUG DISCOVERY EXPLORES NEW PARTS OF CHEMICAL SPACE, HOW CAN FORCEFIELDS KEEP UP?

The Generalized Amber Forcefield (GAFF) was parameterized with this chemical universe:



Extension of this universe is nontrivial because parameter fitting code never released!

Wang J, Wolf RM, Caldwell JW, Kollman PA, and Case DA. J Comput Chem 25:1157, 2004.

THE APPROACH TO PARAMETERIZATION HAS EVOLVED OVER TIME, BUT IT'S STILL NOT COMPLETELY AUTOMATED BY ANY MEASURE

year	forcefield	parameter fitting	atom types
1990s	AMBER parm96	lots of "hand tweaking"	hand-picked
early 2000s	GAFF	genetic algorithm	hand-picked
mid 2000s	TIP4P-Ew	least-squares optimization	hand-picked

Torsion barrier for peptide bond from parm96.dat

```
X -C -N -X    4    10.00        180.0              2.          AA|check Wendy?&NMA
```

How can we move to automated schemes that are easy to grow and refine?

WHAT DO WE WANT OUT OF A FORCEFIELD PARAMETERIZATION SCHEME?

Everything is automatic; don't need to tweak things by hand.

Stupendous feats of chemical insight are not required.

Automatically chooses optimal functional forms.

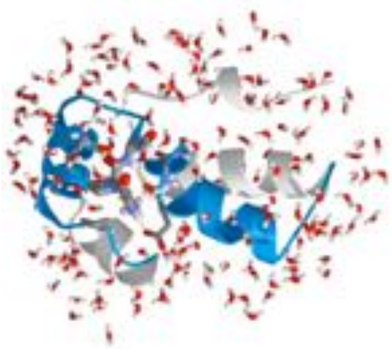
We can can add more data when we reach uncharted parts of chemical space.

Would give us an idea of how reliable it new predictions are expected to be.

We can build a map of what data we should try to collect to improve accuracy.

Is there a procedure that could fit these criteria?

THE OLD WAY

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

One set of parameters in, one computed result out

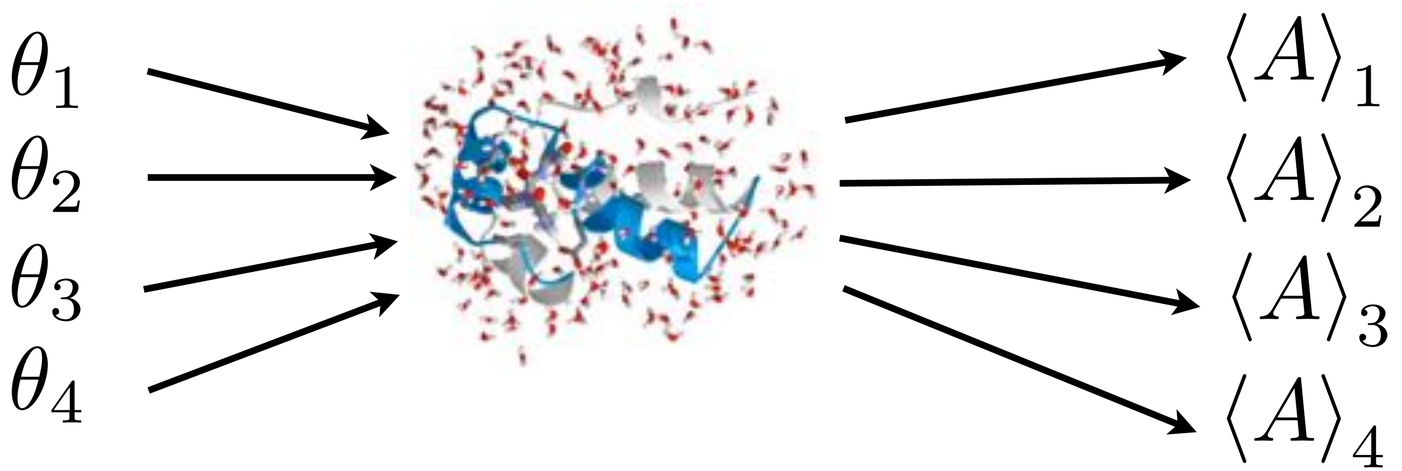
THE BAYESIAN WAY

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

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

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

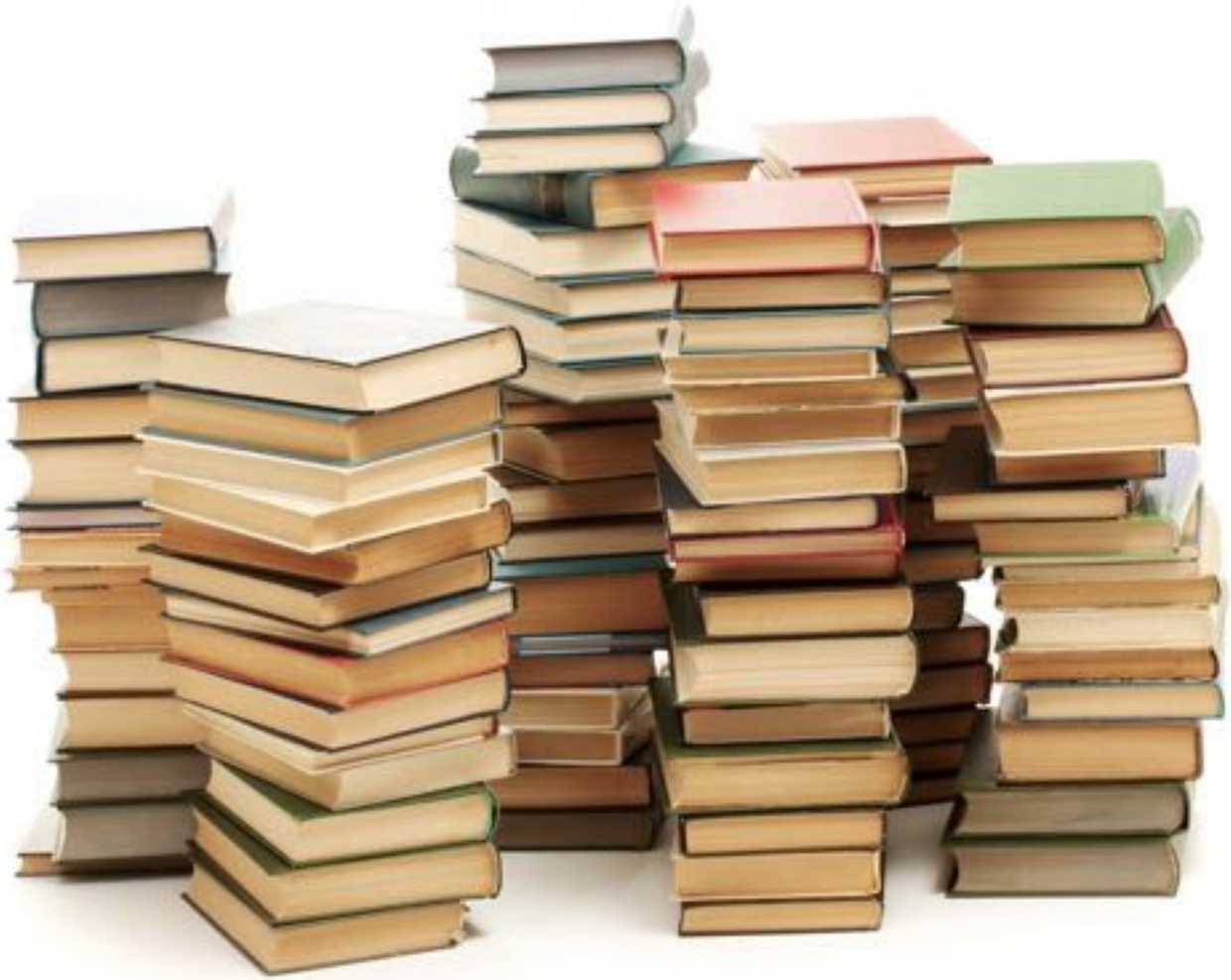
THE BAYESIAN WAY



Multiple parameter sets in, multiple estimates out

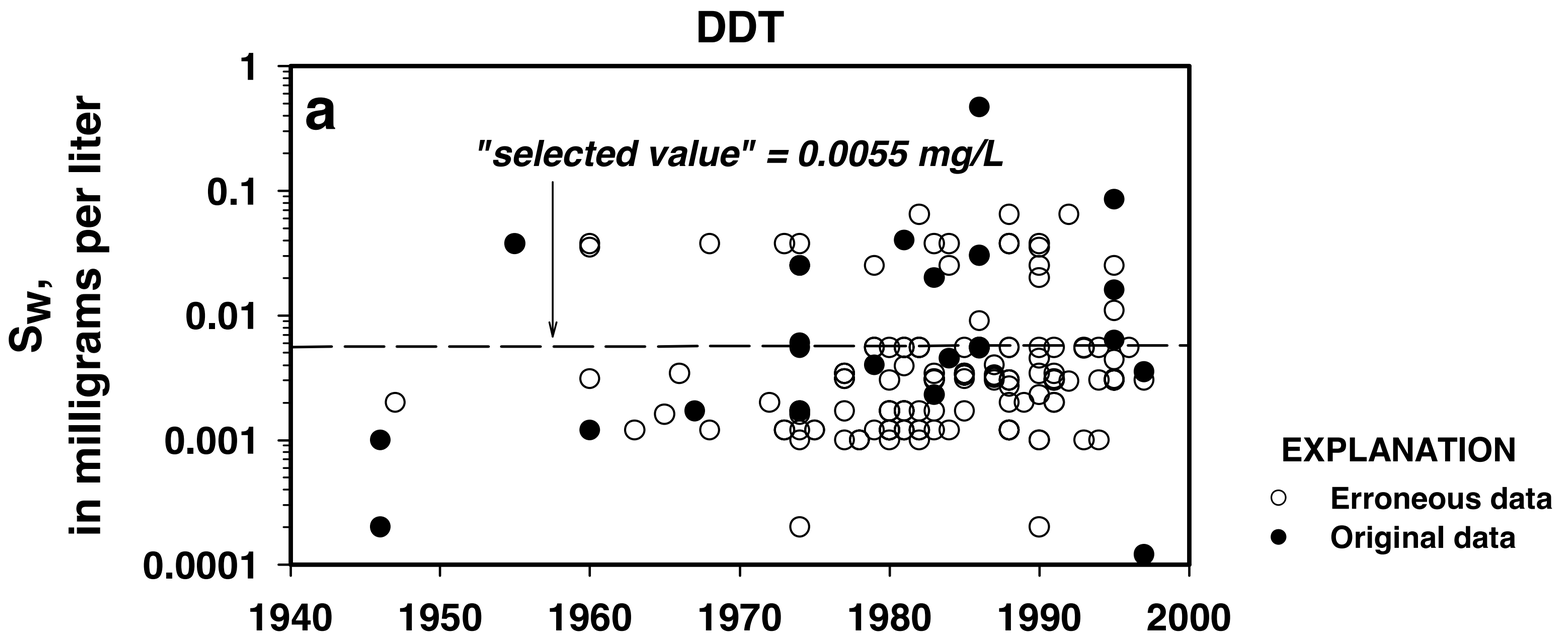
W can estimate both statistical and systematic components of computed results

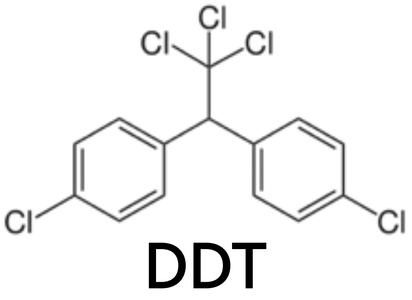
WHERE DO WE GET THE DATA?

WHER  DATA?

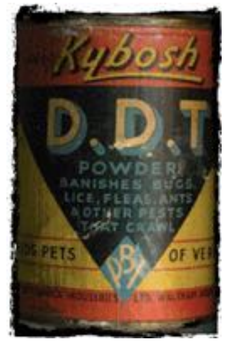
“ANALOGUE DATABASES”

THE LITERATURE IS FILLED WITH ERRONEOUS DATA



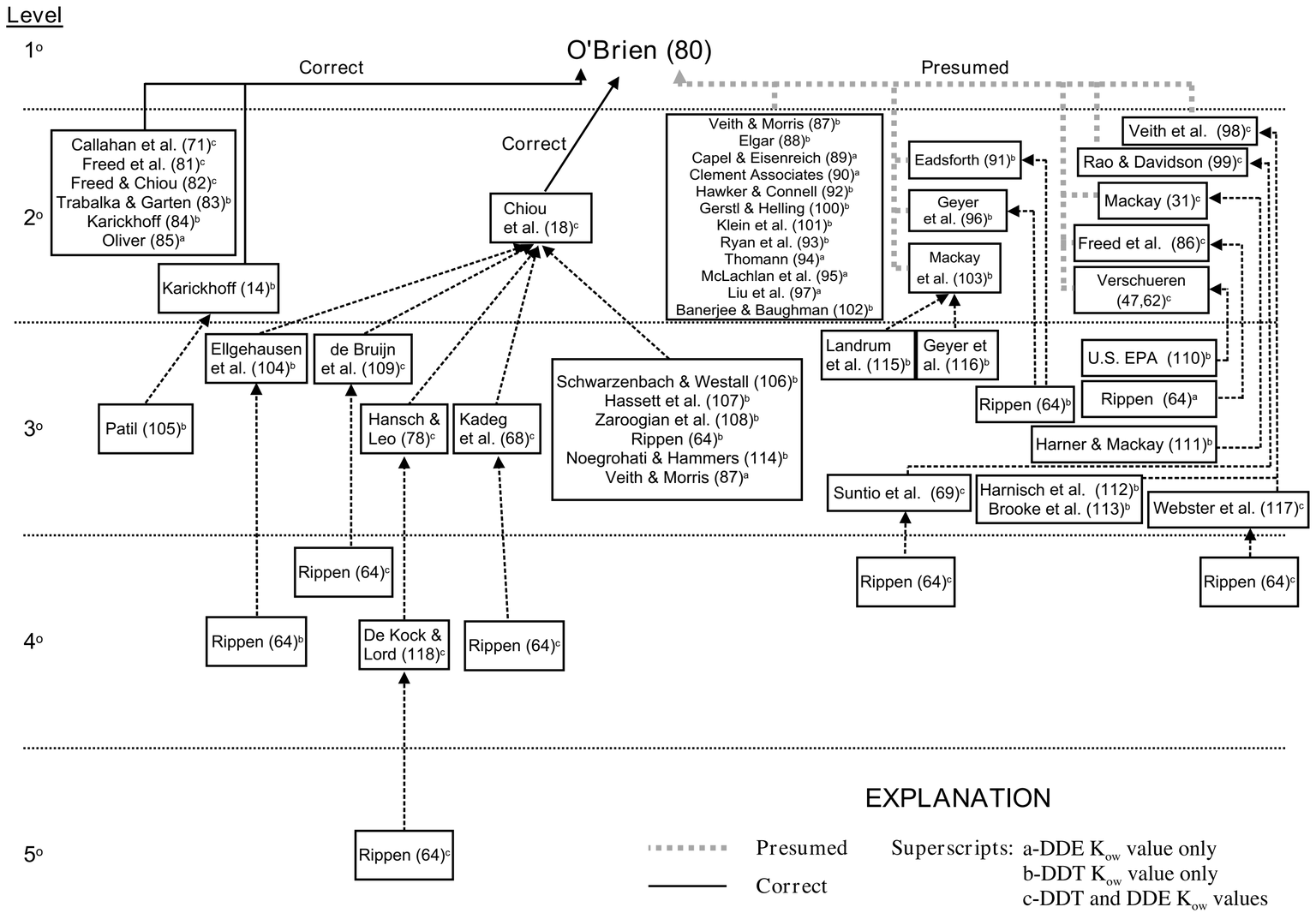


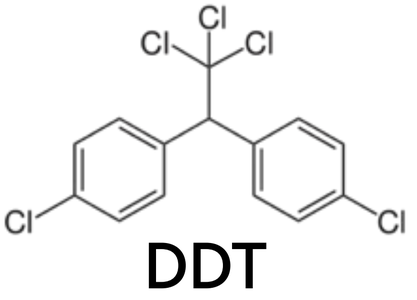
DDT



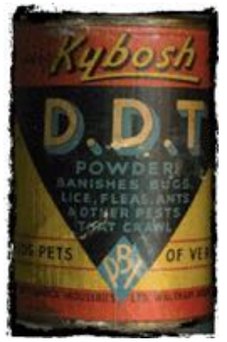
USGS Water-Resources Investigations Report 01-4201, 2001.

DATA HAS A HABIT OF BEING RE/MISREPORTED: THE GENEALOGY OF A SINGLE MEASUREMENT





DDT



USGS Water-Resources Investigations Report 01-4201, 2001.

NIST HAS A SOLUTION

Kyle Beauchamp

Applied Chemicals and Materials Division
TRC Group

NIST National Institute of Standards and Technology

ThermoML

AN XML-BASED IUPAC STANDARD FOR STORAGE AND EXCHANGE OF EXPERIMENTAL THERMOPHYSICAL AND THERMOCHEMICAL PROPERTY DATA

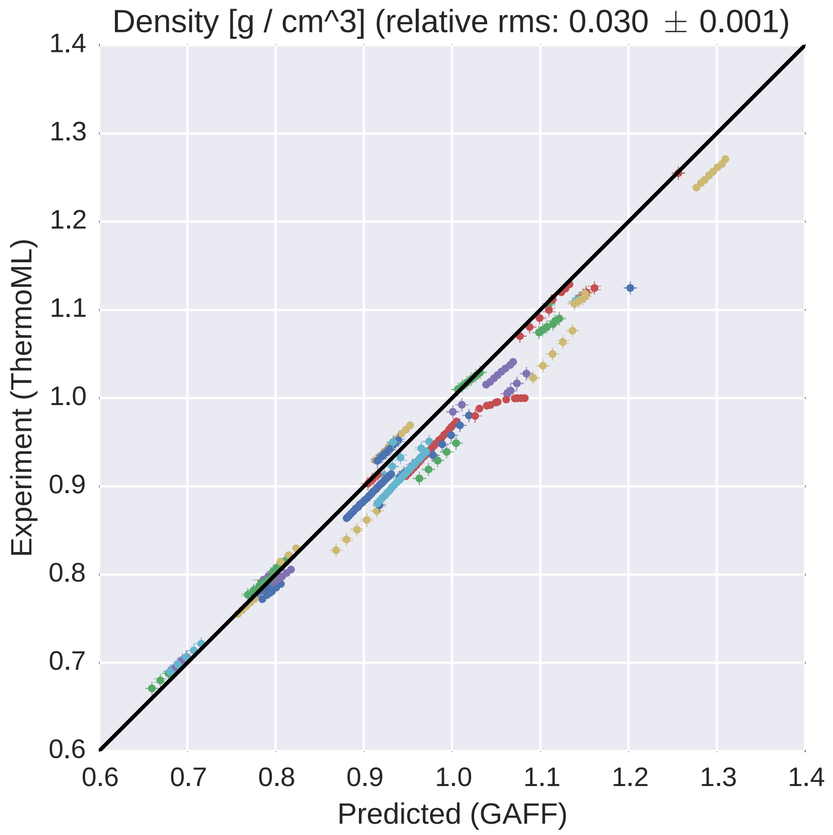
with Kenneth Kroenlein, NIST TRC

	Number of measurements remaining	
Filter step	**Mass density**	**Static dielectric**
1. Single Component	136212	1651
2. Druglike Elements	125953	1651
3. Heavy Atoms	71595	1569
4. Temperature	38821	964
5. Pressure	14103	461
6. Liquid state	14033	461
7. Aggregate T, P	3592	432
8. Density+Dielectric	246	246

JPC B ASAP. DOI:10.1021/acs.jpcb.5b06703

DENSITIES OF MOLECULAR LIQUIDS ARE REASONABLY WELL MODELED

Kyle Beauchamp



Towards Automated Benchmarking of Atomistic Forcefields:
Neat Liquid Densities and Static Dielectric Constants from the ThermoML Data Archive

Kyle A. Beauchamp, Julie M. Behr, Ariën S. Rustenburg, Christopher I. Bayly, Kenneth Kroenlein, and John D. Chodera

Abstract

Atomistic molecular simulations are a powerful way to make quantitative predictions, but the accuracy of these predictions depends entirely on the quality of the forcefield employed. While experimental measurements of condensed phase physical properties offer a straightforward approach for evaluating forcefield quality, the bulk of this information has been tied up in formats that are not machine-readable. Compiling benchmark datasets of physical properties from non-machine-readable sources requires substantial human effort and is prone to accumulation of human errors, hindering the development of reproducible benchmarks of forcefield accuracy. Here, we examine the feasibility of benchmarking atomistic forcefields against the NIST ThermoML data archive of physicochemical measurements, which aggregates thousands of experimental measurements in a portable, machine-readable, self-annotating format. As a proof of concept, we present a detailed benchmark of the generalized Amber small-molecule forcefield (GAFF) using the AM1-BCC charge model against experimentally reported liquid density and static dielectric constant at ambient pressure, automatically extracted from the archive, and discuss the extent of available data. The results of this benchmark highlight a general problem with fixed-charge forcefields in the representation of low-dielectric environments such as those seen in binding pockets or biological membranes.

Keywords: molecular mechanics forcefields, broadband parameterization, forcefield accuracy, forcefield validation, mass density, static dielectric constant, thermodynamic simulation

I. INTRODUCTION

Recent advances in hardware and software for molecular dynamics simulation now permit routine access to atomistic simulations at the ms to timescale and beyond [1]. Leveraging these advances in combination with consumer GPU clusters, distributed computing, or custom hardware has brought nanosecond and millisecond simulation timescales within reach of many laboratories. These dramatic advances in sampling, however, have revealed deficiencies in forcefields as a critical barrier to enabling truly predictive simulations of physical properties of biomolecular systems.

Protein and water forcefields have been the subject of numerous benchmarks [2–4] and enhancements [5–7], with key outcomes including the ability to fold fast-folding proteins [8–10], improved fidelity of water thermodynamic properties [11], and improved prediction of NMR observables. Although small molecule forcefields have also been the subject of benchmarks [12] and improvements [13], such work has typically focused on small perturbations to specific functional groups. For example, a recent study found that modified hydroxyl monovalent parameters led to improved prediction of static dielectric constants and hydration free energies [13]. There are also outstanding questions of generalizability of these targeted perturbations: it is uncertain whether changes to the parameters for a specific chemical moiety will be compatible with seemingly unrelated rearrangements to other groups. Addressing these questions requires establishing community agreement upon shared benchmarks that can be easily replicated among laboratories to test proposed forcefield enhancements and expanded as the body of experimental data grows.

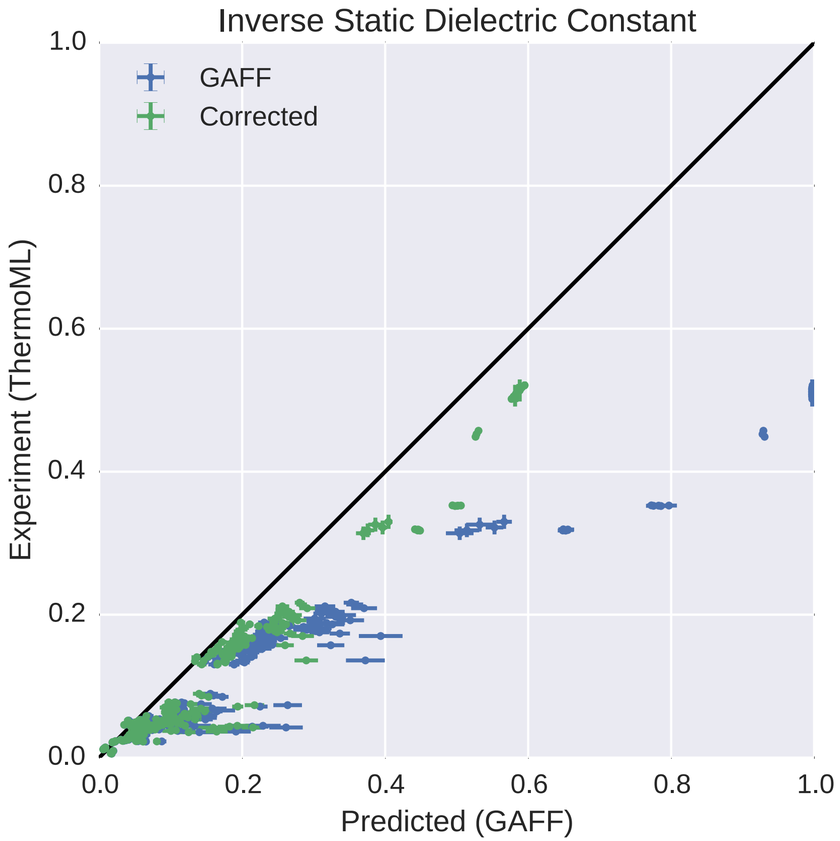
A key barrier to establishing reproducible and extensible forcefield accuracy benchmarks is that many experimental datasets are heterogeneous, paywalled, and unavailable in machine-readable formats (although notable counterexamples exist, e.g. the NCSB [14], ChemSpider [15], and the BMRB [16]). While this inconvenience is relatively minor for benchmarking forcefield accuracy for a single target system (e.g. water), it becomes prohibitive for studies spanning the large relevant chemical spaces, such as forcefields intended to describe a large variety of druglike small organic molecules.

In addition to inconvenience, the number and kind of human-induced errors that can corrupt hand-compiled benchmarks are legion. A United States Geological Survey (USGS) case study examining the reporting and use of literature values of the aqueous solubility [17,] and octanol-water partition coefficients (K_{ow}) for DDT and its analo-

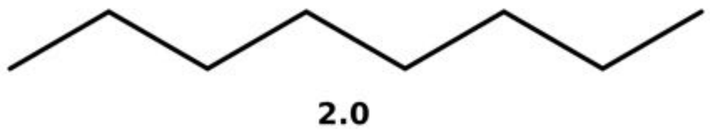
JPC B ASAP. DOI:10.1021/acs.jpcb.5b06703

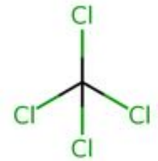
LOW-DIELECTRIC MOLECULES ARE POORLY MODELED

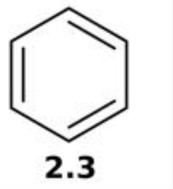
Kyle Beauchamp



$$U(r) = \frac{q_1 q_2}{\epsilon\, r} \propto \frac{1}{\epsilon}$$

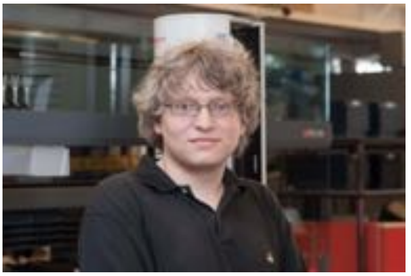




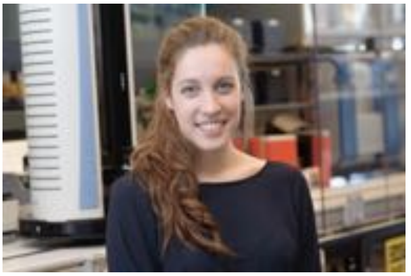


JPC B ASAP. DOI:10.1021/acs.jpcb.5b06703

NEW DATA WILL GREATLY IMPROVE FORCEFIELD QUALITY

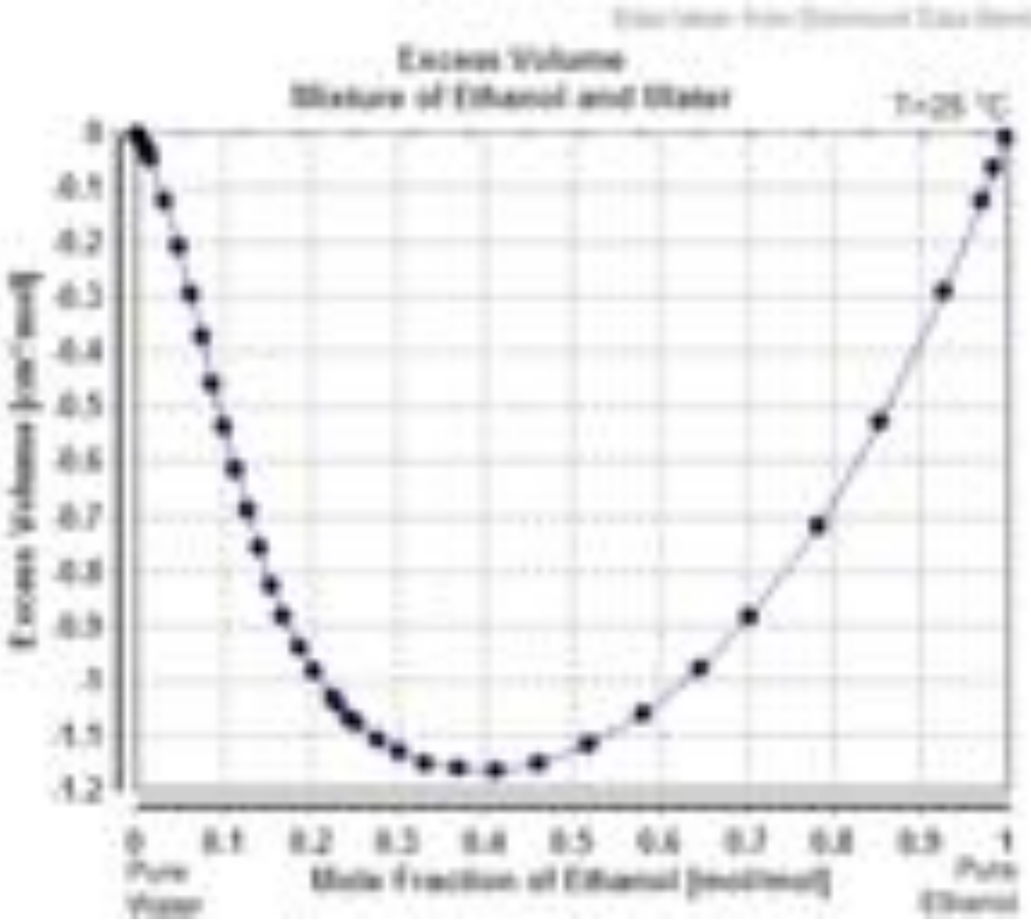


Patrick Grinaway



Julie Behr

Temperature-dependent densities of binary mixtures provides valuable information about atomic interactions



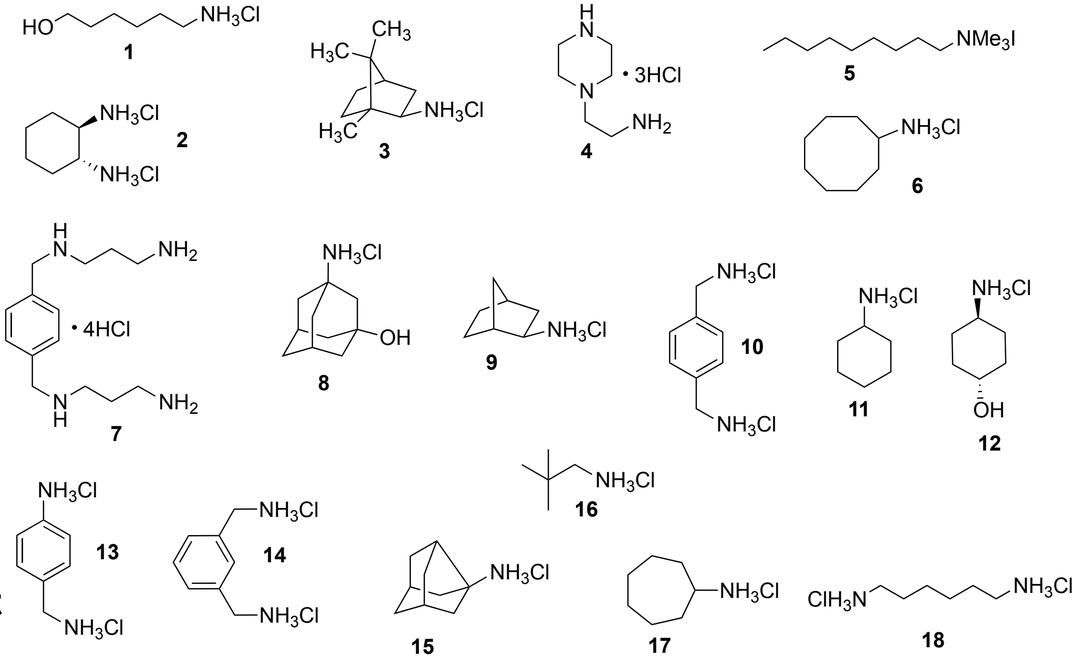
Mettler-Toeldo DM40 density meter
accuracy: 0.0001 g/cm3
range: 0-3 g/cm3
temperature: 0-91 C
sample volume: 1 mL

Mettler-Toledo SC30
automated 30-sample changer

HOST-GUEST SYSTEMS ARE AN EXCELLENT SYSTEM FOR BENCHMARKING SMALL-MOLECULE AFFINITIES

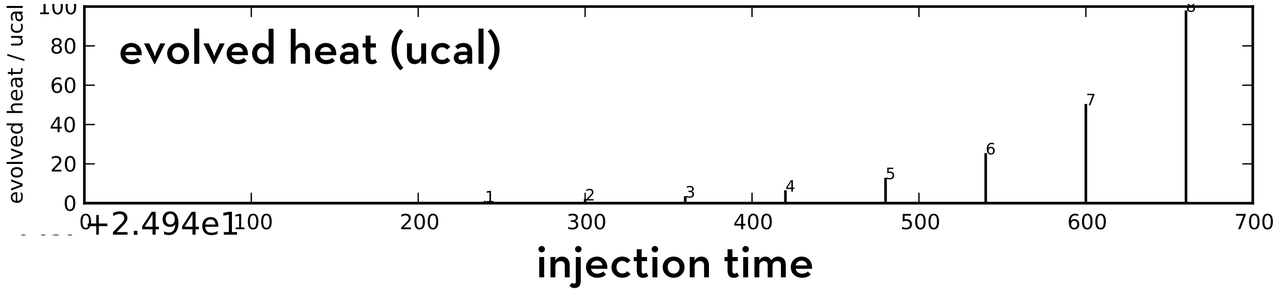
Bas Rustenburg

cucurbit[7]uril (CB7) with small molecule guest
(Lyle Isaacs and Michael Gilson)



small-molecule guests

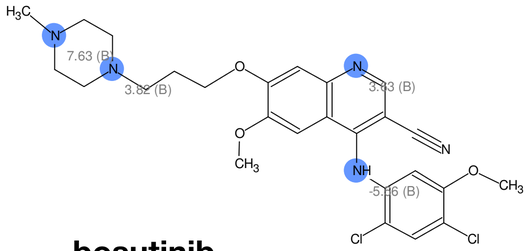
GE/MicroCal Auto iTC-200
isothermal titration calorimeter



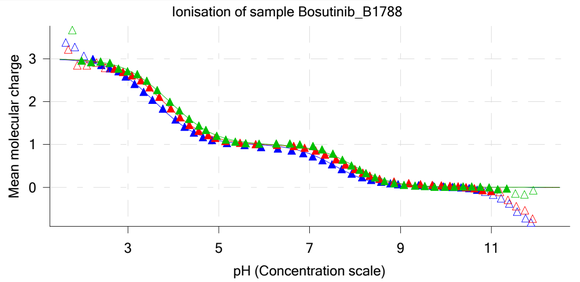
AQUEOUS SOLUBILITIES, PKAS, AND PARTITION COEFFICIENTS CAN PROVIDE HIGHLY USEFUL INFORMATION

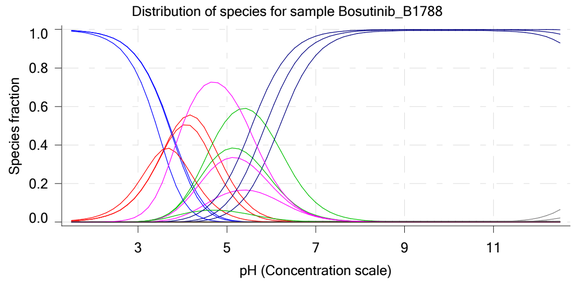
Sirius Analytical T3

Up to 192 measurements per
fully automated walk-away session
~1 ug dry compound
or ~7 uL of 10 mM DMSO stock

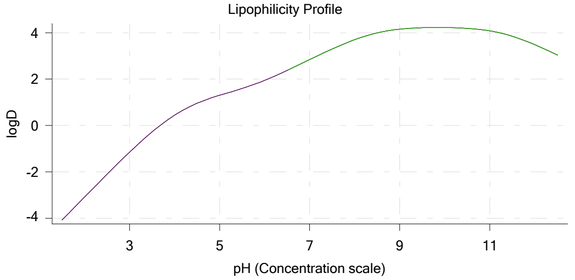


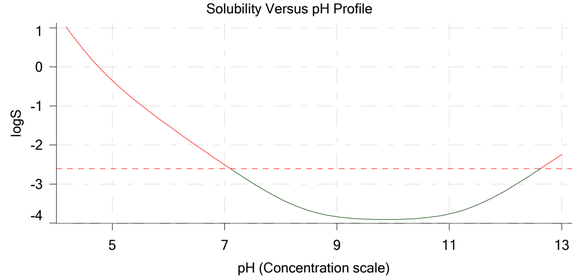
bosutinib





* aqueous **solubilities**
* UV- and electrochemical **pKa**
* **partition coefficients** between water and any immiscible phase (octanol, cyclohexane, hexane, etc.)





THE OPEN FORCEFIELD GROUP

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

THE CAST OF CHARACTERS (SO FAR):

DAVID MOBLEY
UCI

MICHAEL GILSON
UCSD

MICHAEL SHIRTS
UNIVERSITY OF COLORADO, BOULDER

CHRISTOPHER BAYLY
OPENEYE SCIENTIFIC

JOHN CHODERA
SKI/MSKCC

SMARTS-BASED FORCEFIELD ASSIGNMENT

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

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

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


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

https://github.com/open-forcefield-group/smarty/blob/master/The-SMIRFF-force-field-format.md

SAMPLING OVER TYPES

PARENT TYPES

```
% atom types
[#1]    hydrogen
[#6]    carbon
[#7]    nitrogen
[#8]    oxygen
[#9]    fluorine
[#15]   phosphorous
[#16]   sulfur
[#17]   chlorine
[#35]   bromine
[#53]   iodine
```

X

DECORATORS

```
% total connectivity
X1              connections-1
X2              connections-2
X3              connections-3
X4              connections-4
% total-h-count
H0              total-h-count-0
H1              total-h-count-1
H2              total-h-count-2
H3              total-h-count-3
% formal charge
+0              neutral
+1              cationic+1
-1              anionic-1
% aromatic/aliphatic
a               aromatic
A               aliphatic
```

=

PROPOSED CHILD TYPES

```
[#6X4:1]        tetrahedral carbon
[#6:1]~[#7]     carbon nitrogen-adjacent
```

INDEX	ATOMS	MOLECULES	TYPE NAME	SMARTS	REF TYPE	FRACTION OF REF TYPED MOLECULES MATCHED
1 :	464	42	c_hydrogen	[#1]	HC	244 / 244 (100.000%)
2 :	0	0	c_carbon	[#6]		
3 :	232	42	c_carbon neutral	[#6&+0]	CT	232 / 232 (100.000%)
4 :	107	42	c_oxygen	[#8]	OH	68 / 68 (100.000%)
TOTAL :	803	42				

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

THE FUTURE OF FORCEFIELD PARAMETERIZATION?

EXPERIMENTAL DATA
QUANTUM CHEMISTRY
UNCERTAINTIES

BAYESIAN INFERENCE;
CONTINUAL AUTOMATIC UPDATING

ENSEMBLE OF
PARAMETER SETS

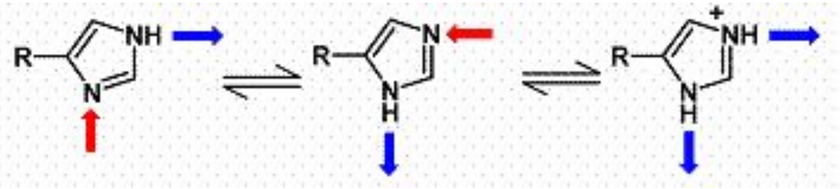
PULL THE TRIGGER AND GO!

PREDICTIONS FAIL FOR THREE REASONS

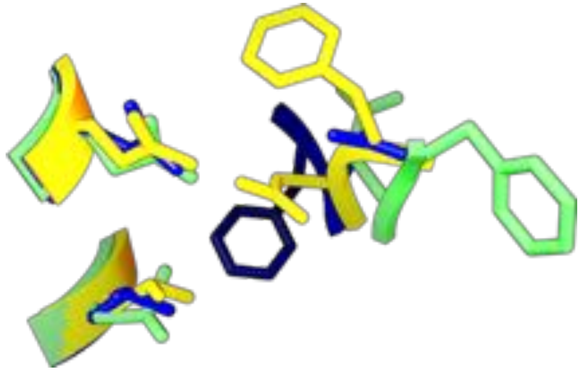
1. The forcefield does a poor job of modeling the physics of our system

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

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

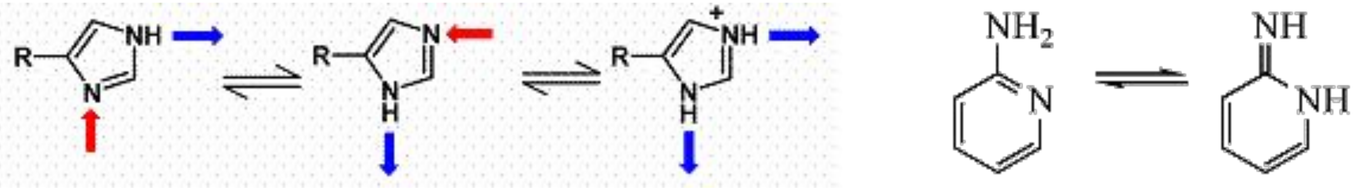


3. We haven't sampled all of the relevant conformations

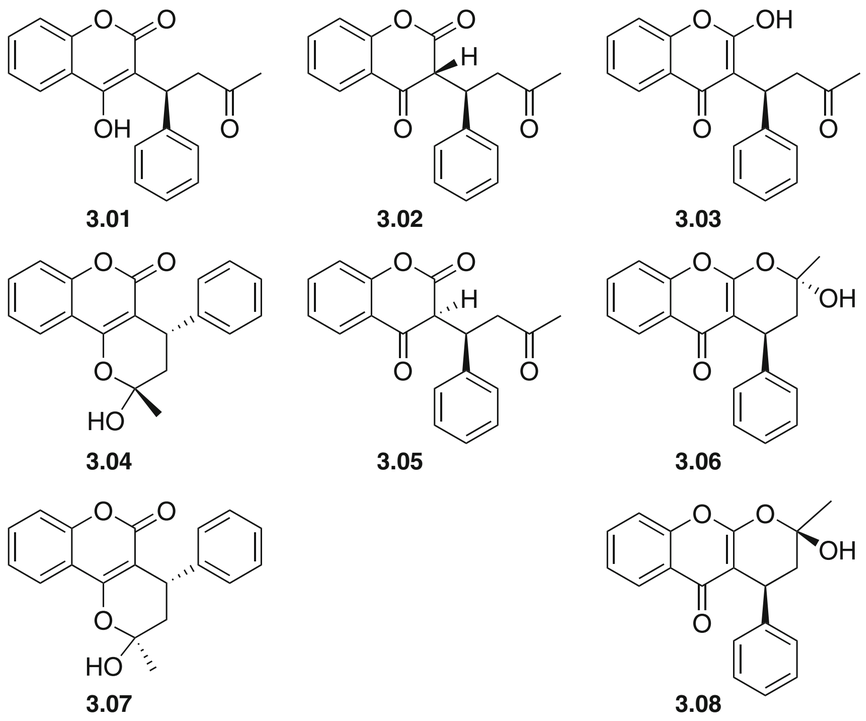


PREDICTIONS FAIL FOR THREE REASONS

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



LET'S NOT FORGET TAUTOMERS



tautomers of **warfarin**

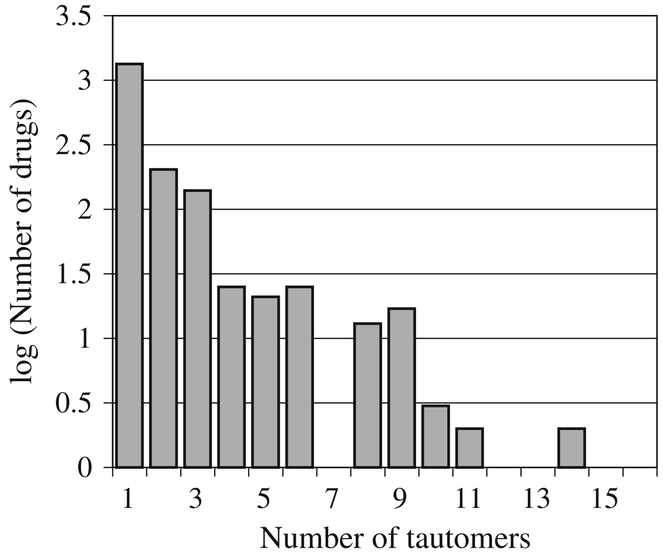
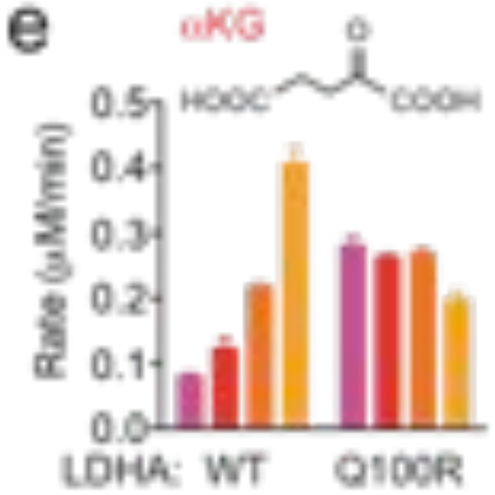


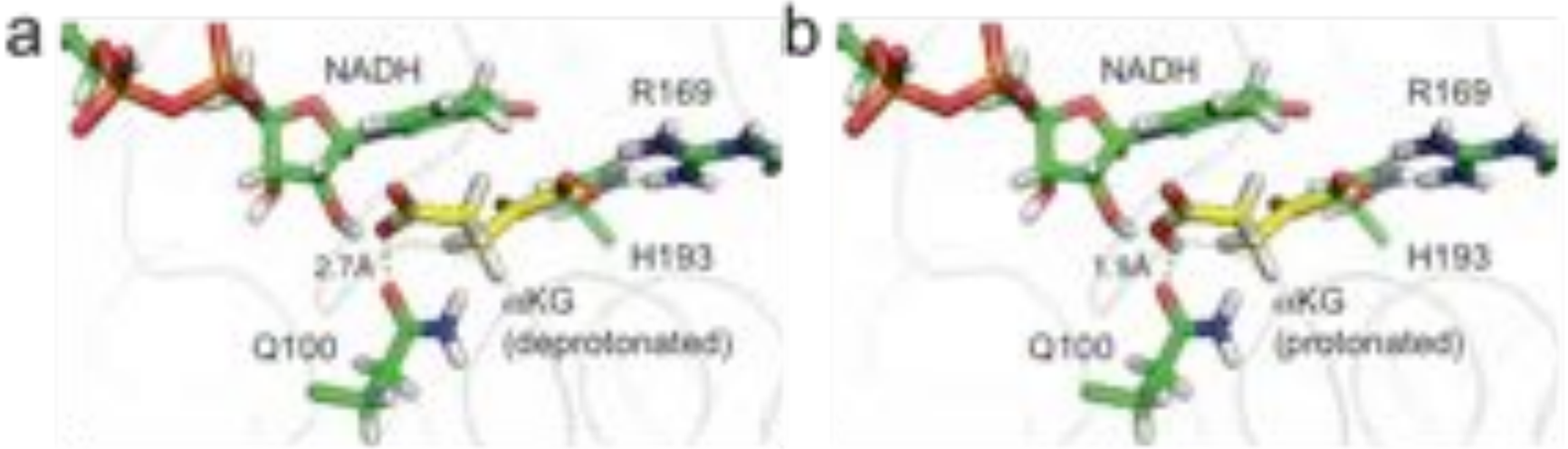
Fig. 13 The frequency distribution of tautomers of marketed drugs

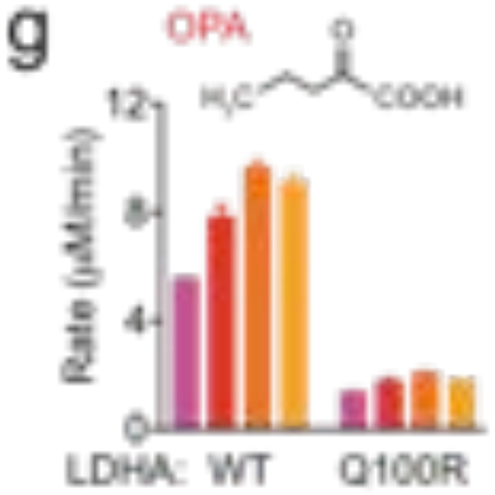
MORE THAN HALF OF ALL DRUGS HAVE 2 OR MORE TAUTOMERS

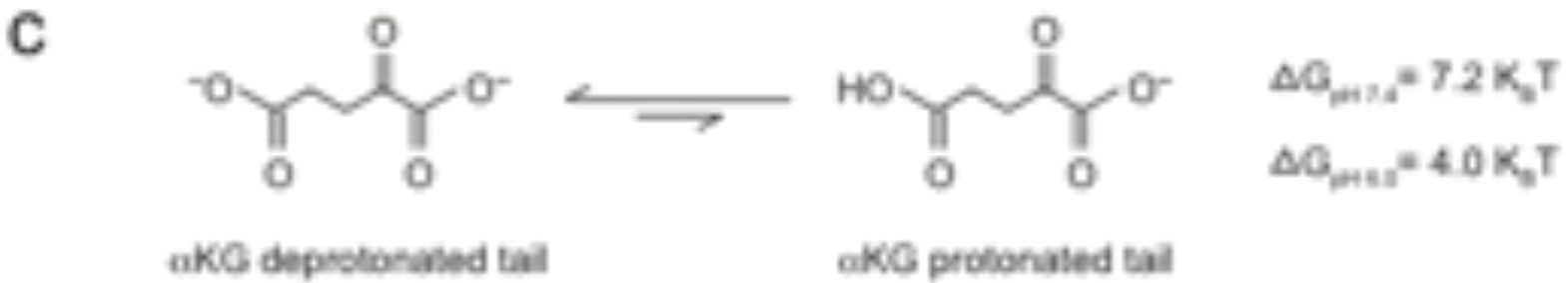
Martin Y. JCAMD 23:693, 2009. Let's not forget tautomers.

PROTONATION STATE EFFECTS CAN BE IMPORTANT FOR LIGAND BINDING



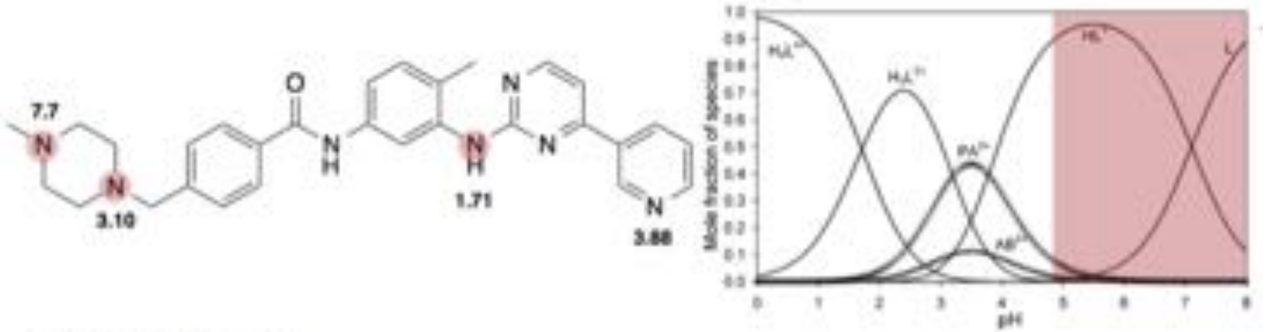






Intlekofer AM, Wang B, Liu H, Carmona-Fontaine C, Rustenburg AS, Salah S, Gunner MR, **Chodera JD,** Cross JR, and Thompson CB. *Nature Chemical Biology,* in press.

KINASES AND KINASE INHIBITORS ARE FULL OF TITRATABLE GROUPS



(A) imatinib

inhibitor

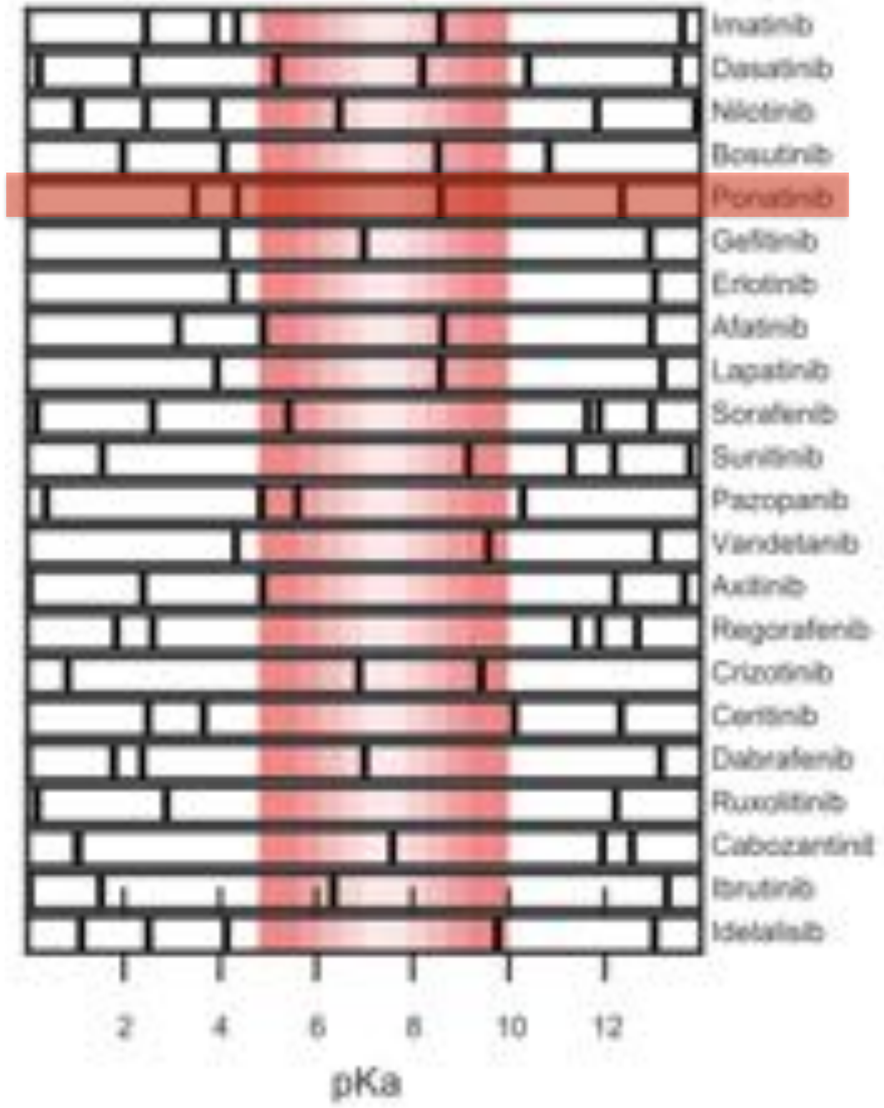
$$\overset{0\ kT}{L} \rightleftharpoons \overset{0.6\ kT}{HL^+} \rightleftharpoons \cdots$$

kinase

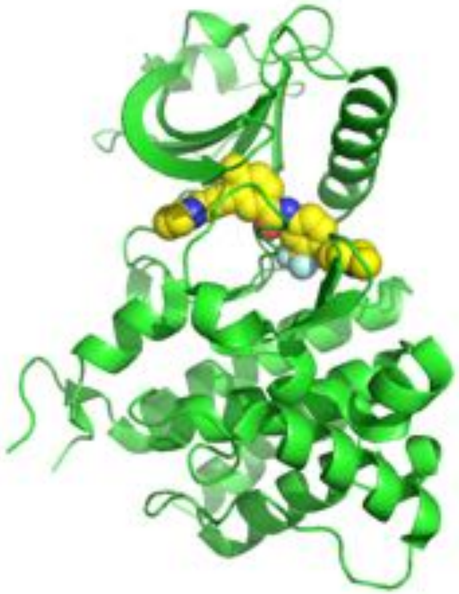
$$\begin{array}{ccc} \overset{1.8\ kT}{Abl_{(DFG\text{-}in;AspH)}} & \rightleftharpoons & \overset{0\ kT}{Abl_{(DFG\text{-}out;AspH)}} \\ \overset{0\ kT}{\uparrow\downarrow} & & \overset{2\ kT}{\uparrow\downarrow} \\ Abl^-_{(DFG\text{-}in;Asp\text{-})} & \rightleftharpoons & Abl^-_{(DFG\text{-}out;Asp\text{-})} \end{array}$$

complex

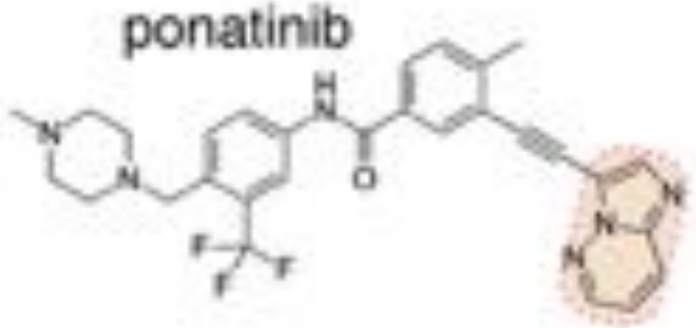
$$\overset{2\ kT}{HL^+:Abl_{(DFG\text{-}out;AspH)}} \rightleftharpoons \overset{0\ kT}{HL^+:Abl^-_{(DFG\text{-}out;Asp\text{-})}} \rightleftharpoons \cdots$$

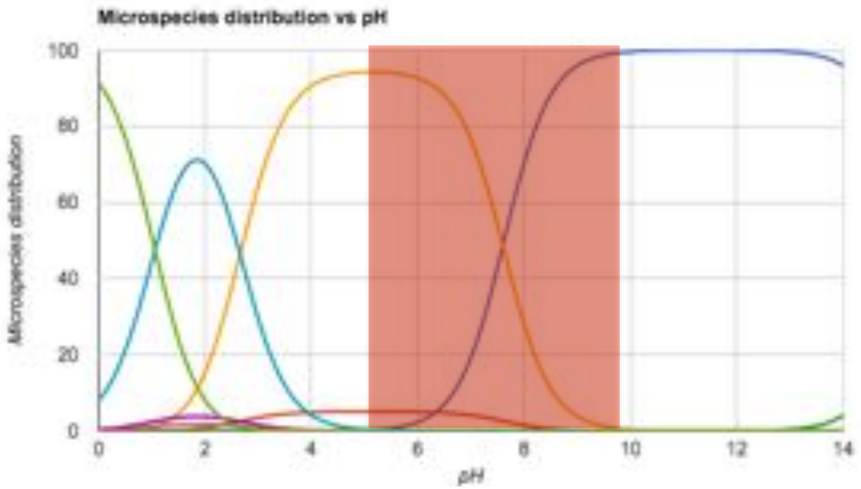


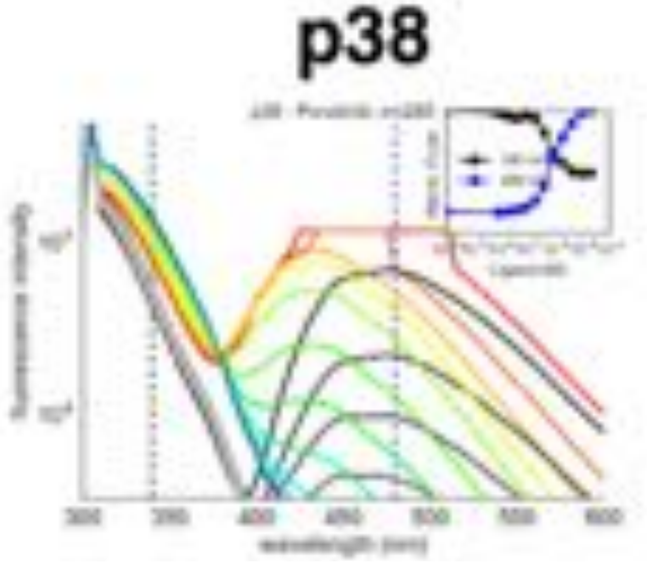
KINASE:INHIBITOR BINDING CAN BE PH-SENSITIVE

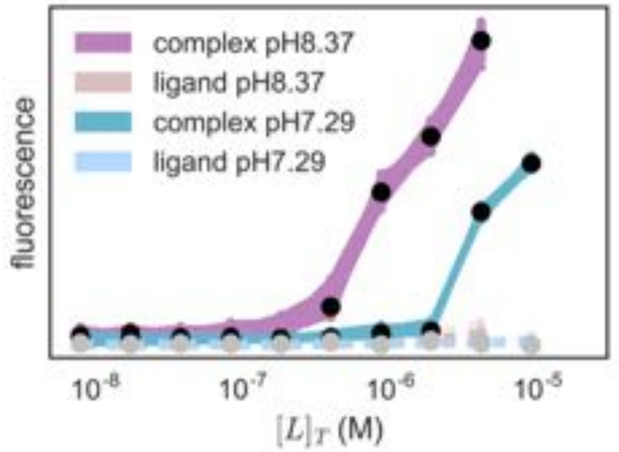


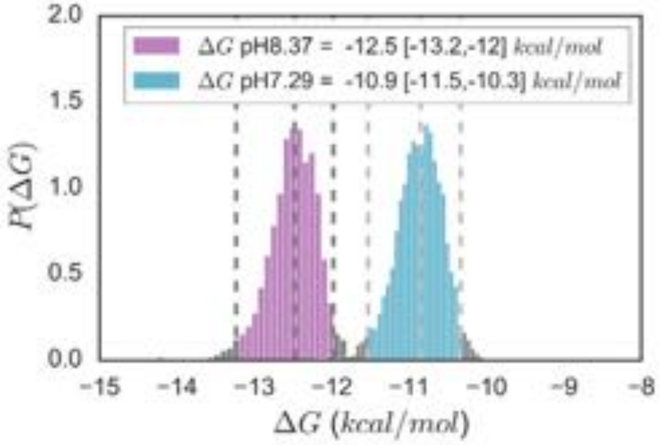
ponatinib:DDR1
(pdbid:3ZOS)











Sonya Hanson
Postdoc

PROTONATION STATE EFFECTS CAN BE IMPORTANT FOR LIGAND BINDING

Marilyn Gunner and Salah Salah (CCNY) in collaboration with Markus Seeliger (Stony Brook) and Paul Czodrowski (Merck Serono)

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

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

proton gain

proton loss

tautomer shift

PROTON-DRIVE

Protons

Protonation states and tautomers for OpenMM

Watch

Navigation

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

Protons: Protonation states and tautomers for OpenMM

Note:

This is module is undergoing heavy development. None of the API calls are final.

Introduction

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

Installation

Use the command

```
python setup.py install
```

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

To test the installation, run

```
nosetests protons
```

Requirements

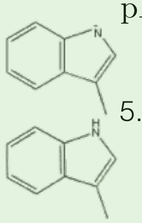
BAS RUSTENBURG

GREGORY ROSS

https://github.com/choderalab/protons

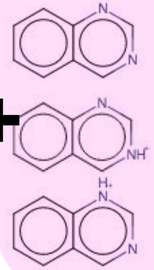
CONSTANT-PH FOR BOTH LIGAND AND PROTEIN IN EXPLICIT SOLVENT

amino acids



pK_a	$\Delta G_{aq}^{\mathrm{pH\ 7.4}}$
	0.0 kT
5.88	
	3.5 kT

small molecules



$\Delta G_{aq}^{\mathrm{pH\ 7.4}}$
0.0 kT
2.3 kT
4.7 kT

SAMS calibration of protomer log weights

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

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

pH correction of protomer log weights

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

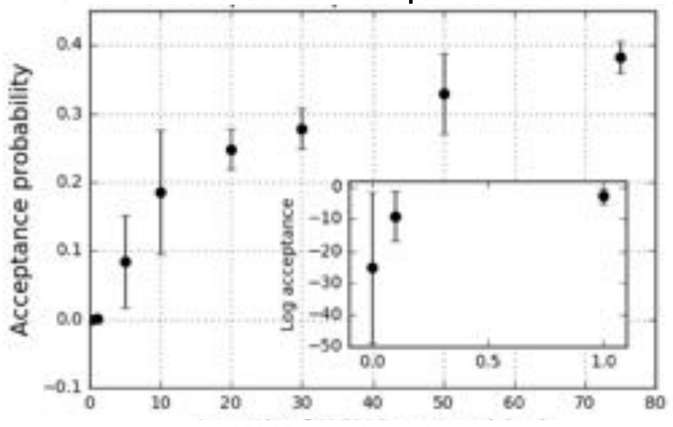
$\mathbf{g}(\mathrm{pH})$ pH-corrected log weight
$\pi_s(\mathrm{pH}) \propto e^{-\beta \Delta G_{aq}^{\mathrm{pH}}}$
π_s^o protomer population in calibration simulation

MCMC protonation state sampling

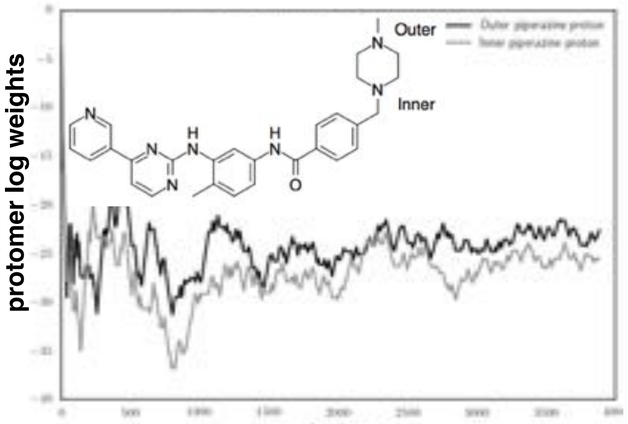
$$\mathbf{x}_{t+1} \sim \pi(\mathbf{x}|\mathbf{s}_t, \mathbf{g}(\mathrm{pH}), \beta, p)$$ **MD / GHMC**
$$\mathbf{s}_{t+1} \sim \pi(\mathbf{s}|\mathbf{x}_{t+1}, \mathbf{g}(\mathrm{pH}), \beta, p)$$ **NCMC**

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

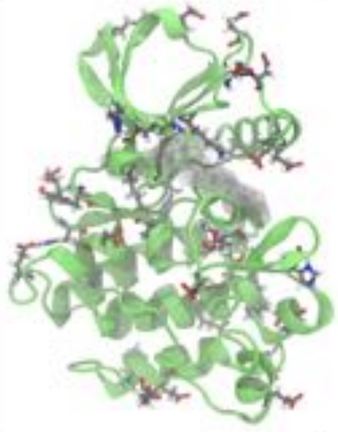
A NCMC acceptance probability for histidine in explicit solvent



B SAMS calibration of imatinib log weights

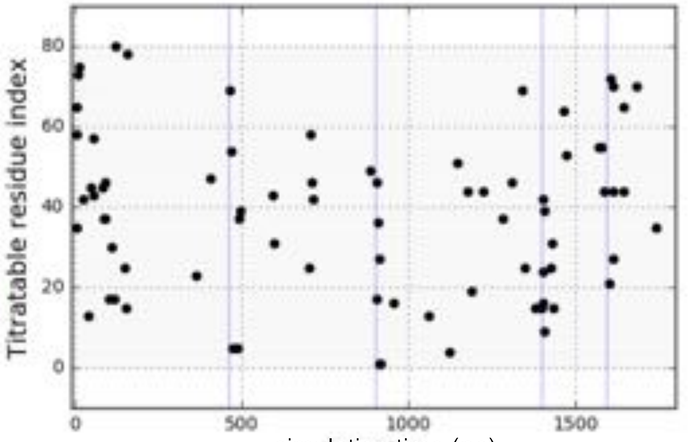


C

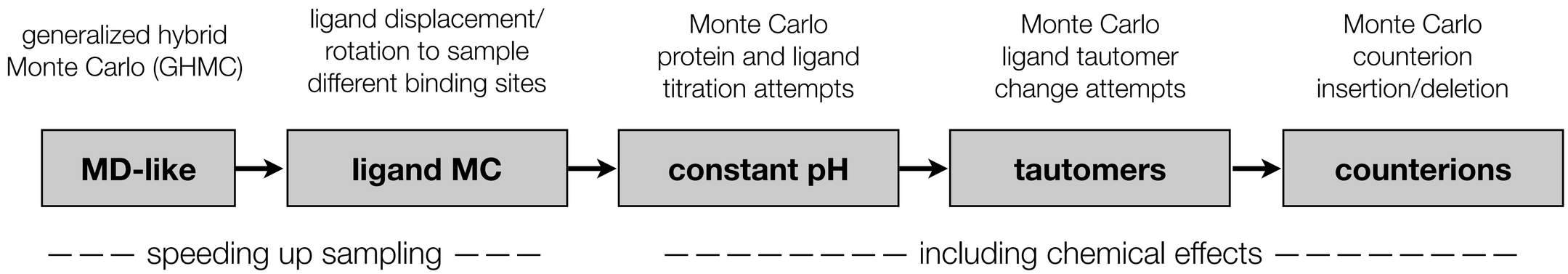


Abl:imatinib

D Abl constant-pH simulation protomer change events



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



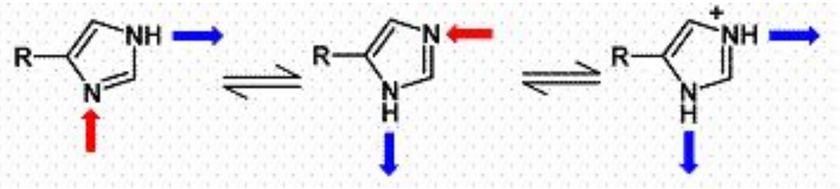
WE CAN USE FREE ENERGY CALCULATIONS AND EXPERIMENTS TO QUANTIFY THE ERROR IN NEGLECTING OF PROTOMERS AND TAUTOMERS

PREDICTIONS FAIL FOR THREE REASONS

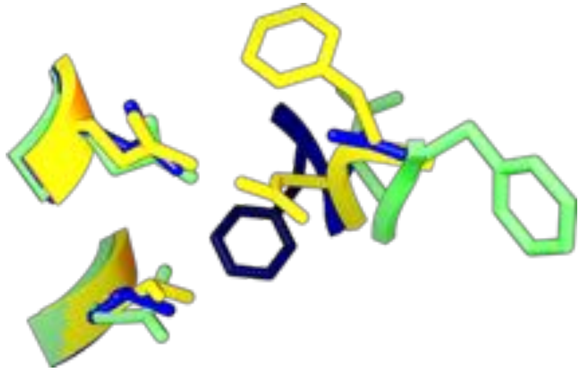
1. The forcefield does a poor job of modeling the physics of our system

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

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

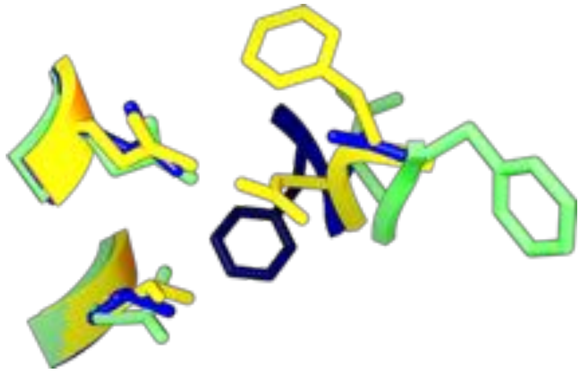


3. We haven't sampled all of the relevant conformations

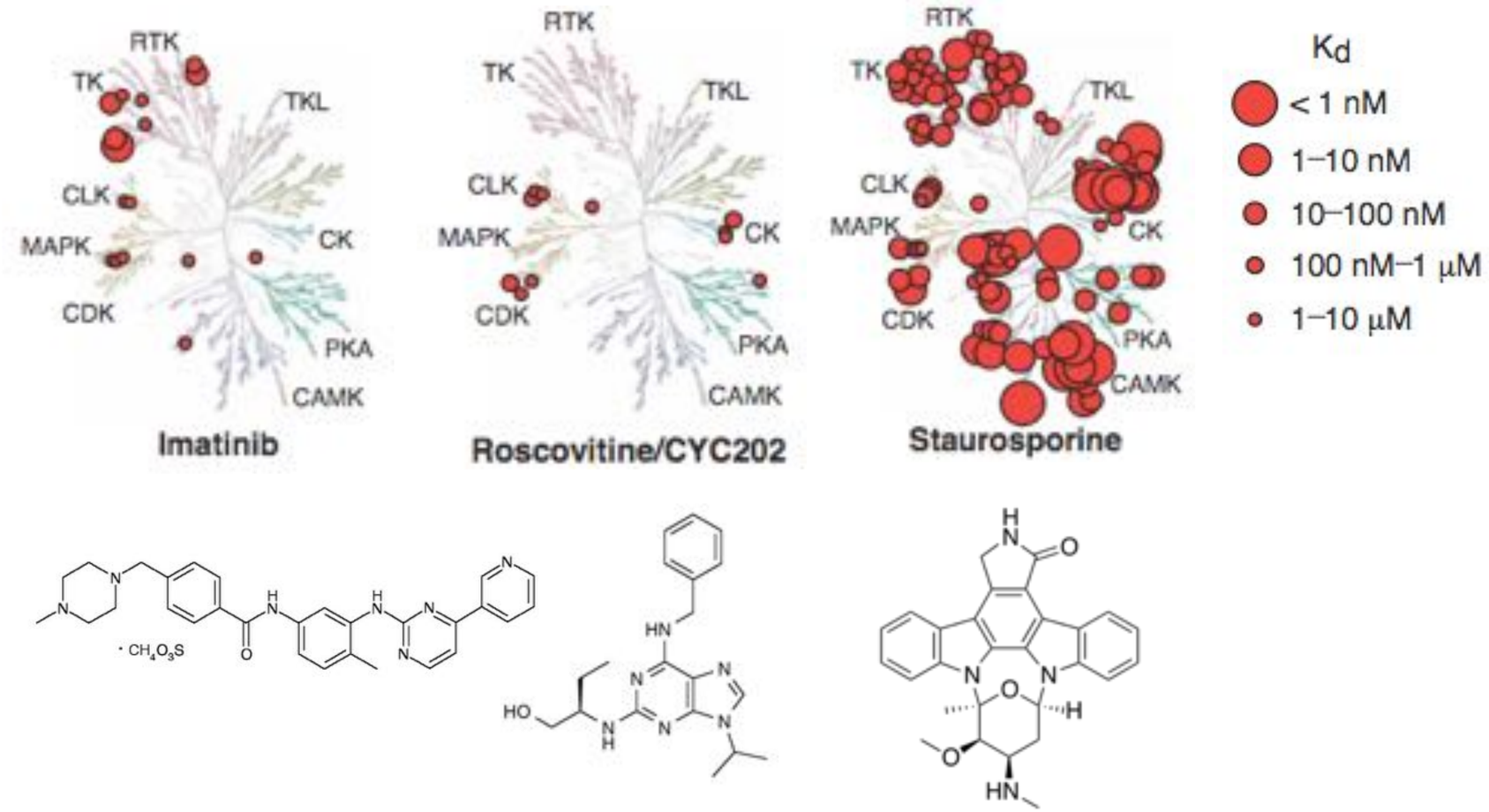


PREDICTIONS FAIL FOR THREE REASONS

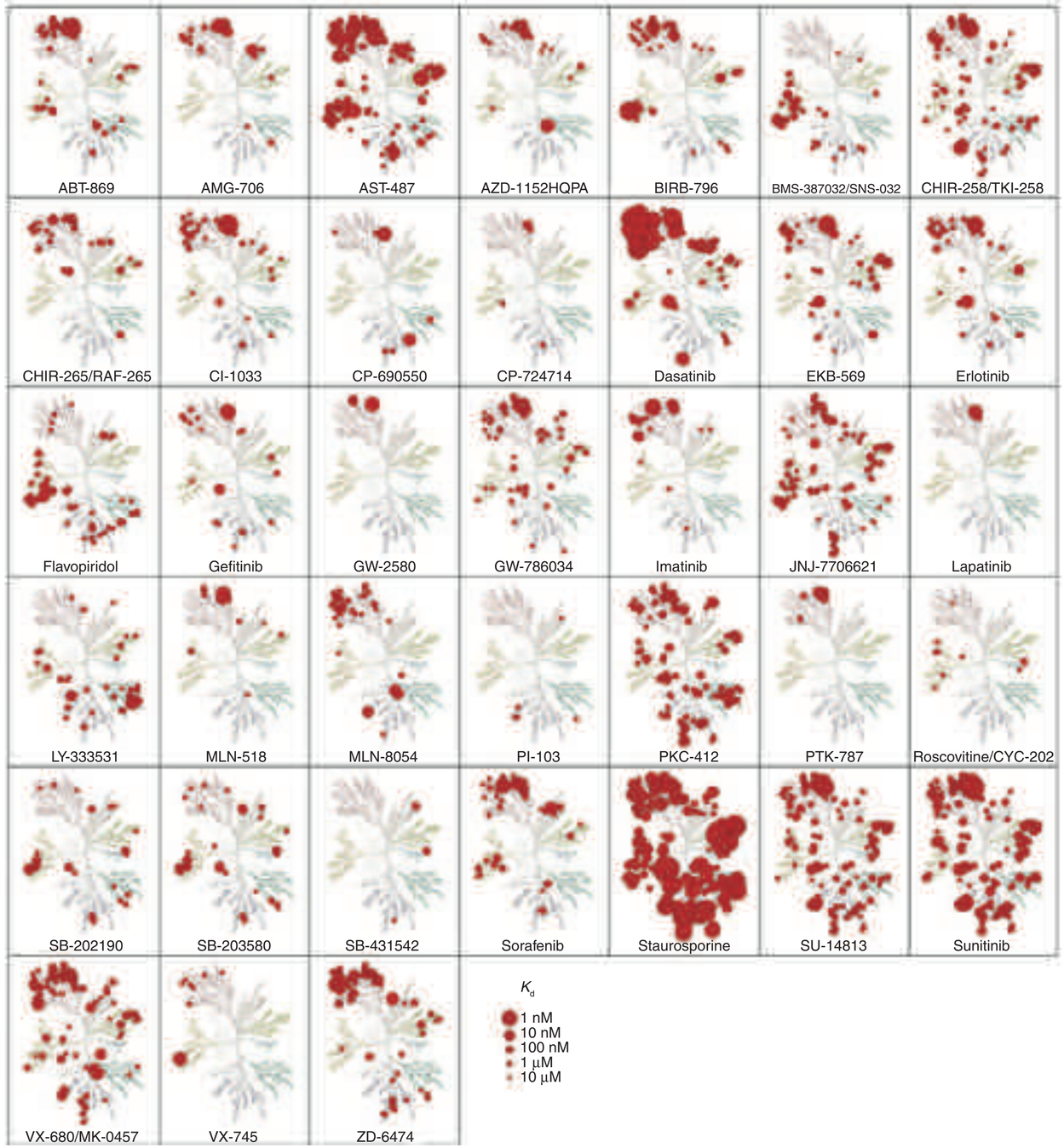
3. We haven't sampled all of the relevant conformations



HOW CAN WE QUANTITATIVELY UNDERSTAND (AND DESIGN) THE SELECTIVITY OF KINASE INHIBITORS?



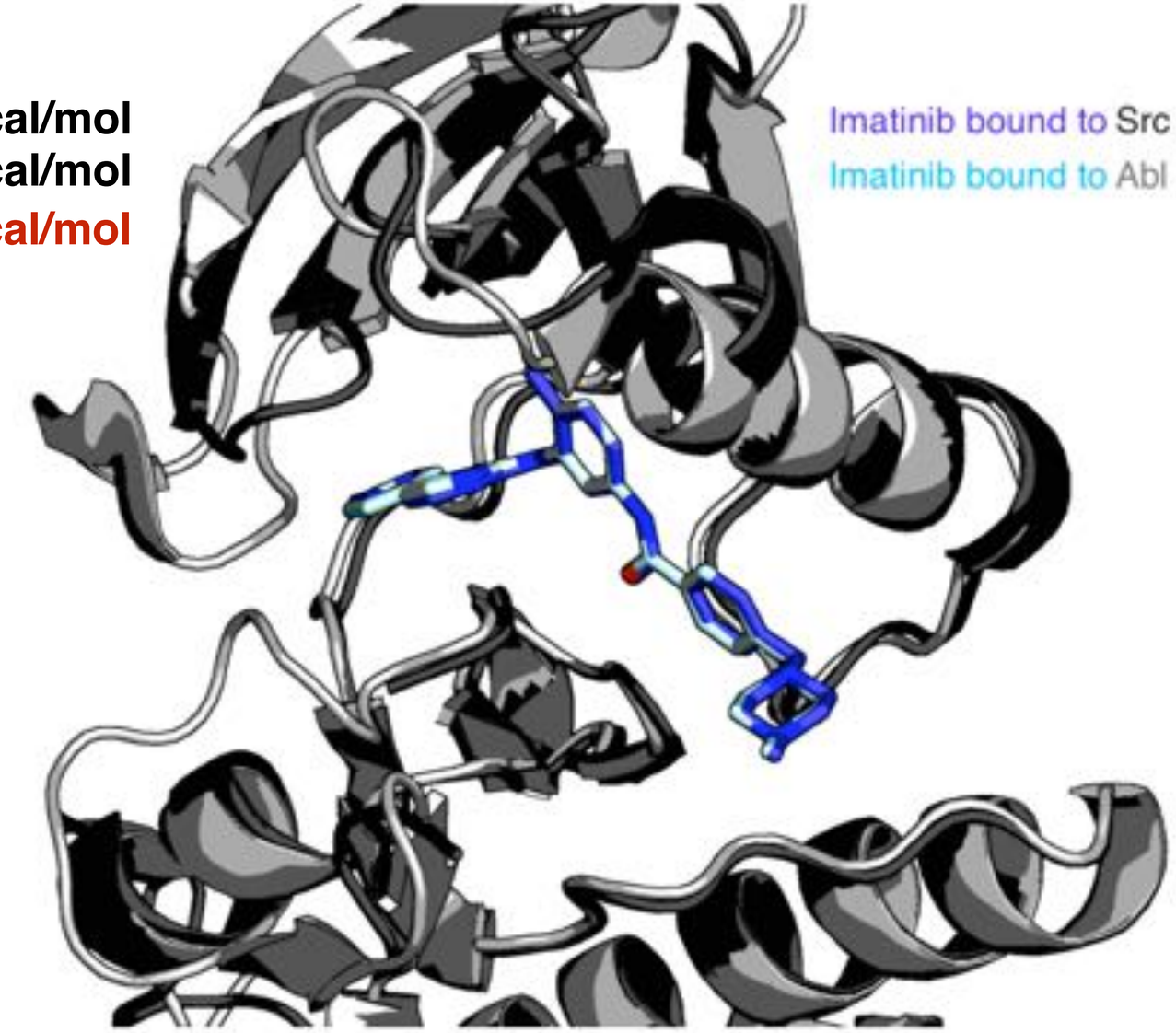
Nature Biotech 26:127, 2008



Nature Biotech 26:127, 2008

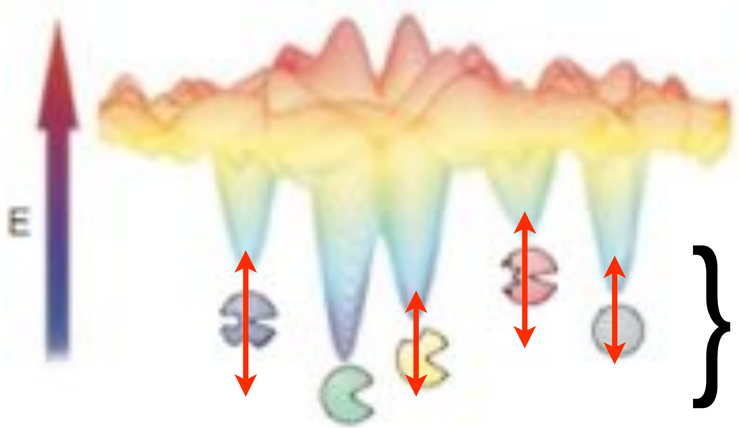
DIFFERENCES IN STABILITIES OF INACTIVE STATES MAY BE RESPONSIBLE FOR ORIGIN OF SOME KINASE INHIBITOR SELECTIVITY

Abl:imatinib ΔG = -10.9 kcal/mol
Src:imatinib ΔG = -6.2 kcal/mol
ΔΔG = 4.7 kcal/mol



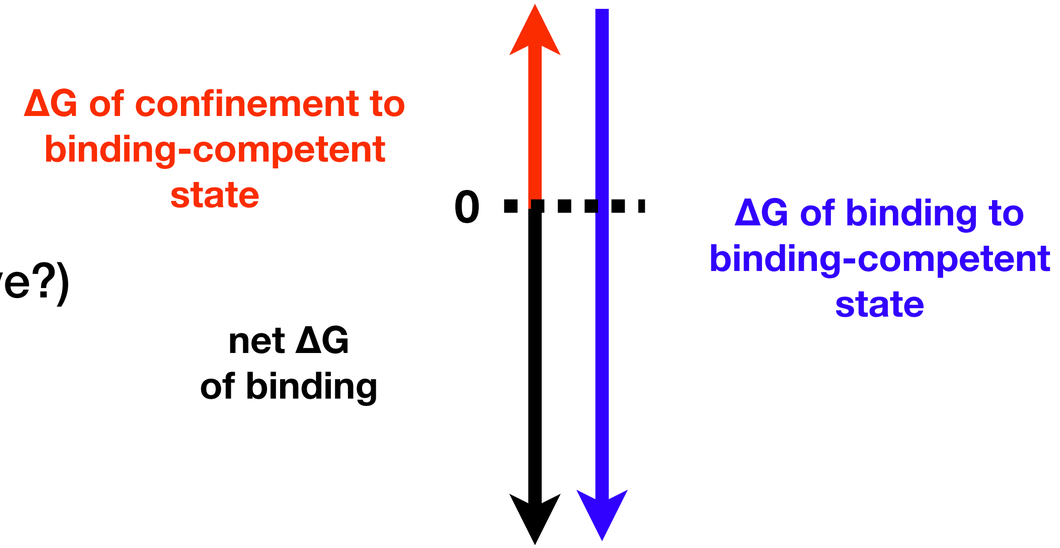
* essentially same binding mode in X-ray structure, same interactions
* calculations suggest no difference in binding free energy [J Biol Chem 285:13807, 2010; PNAS 110:1664, 2013]

DIFFERENCES IN STABILITIES OF INACTIVE STATES MAY BE RESPONSIBLE FOR ORIGIN OF SOME KINASE INHIBITOR SELECTIVITY



high free energy states (inactive?)

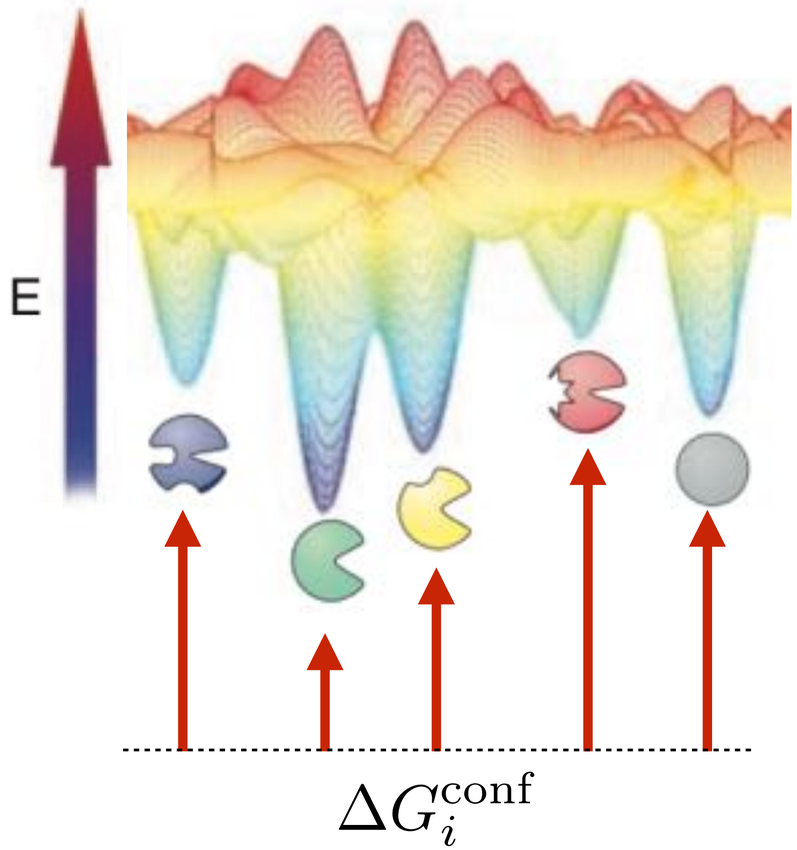
low free energy state (active?)



QUANTIFYING CONFORMATION ENERGETICS MAY BE CRUCIAL TO SUCCESSFUL DESIGN OF SELECTIVE KINASE INHIBITORS

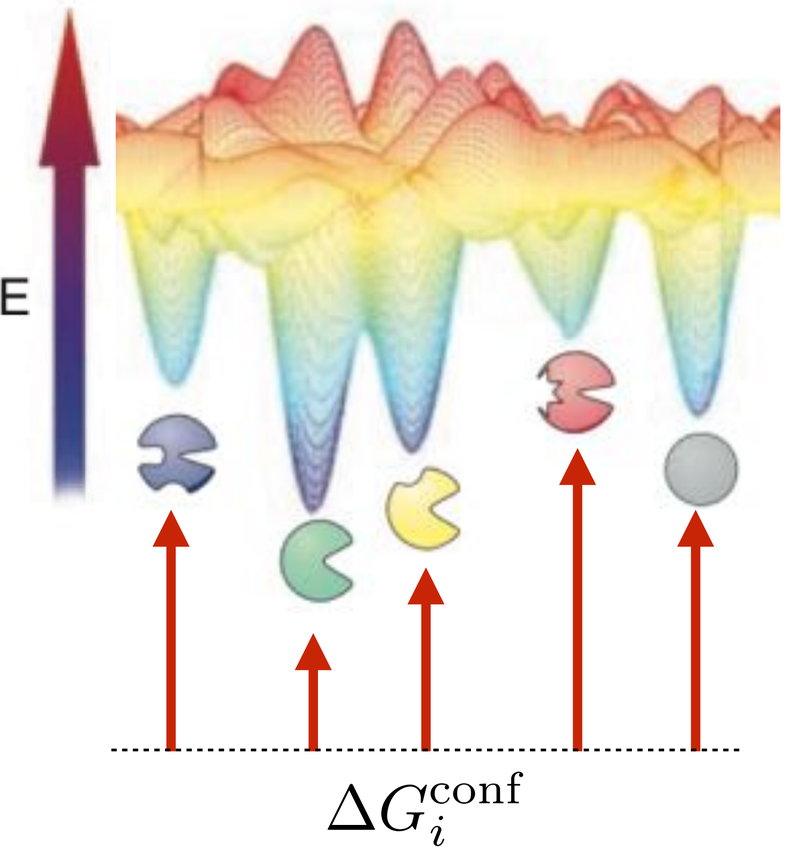
Lee and Craik. Science 324:213, 2009

CAN WE BUILD AN ATLAS OF KINASE STRUCTURES AND ENERGETICS?



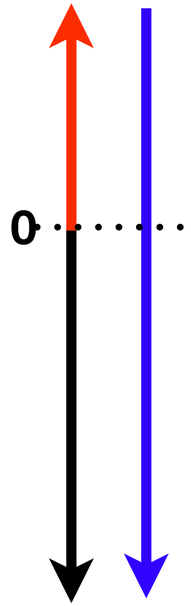
Lee and Craik. Science 324:213, 2009

DIVIDE-AND-CONQUER TO COMPUTE AFFINITIES AND SELECTIVITIES

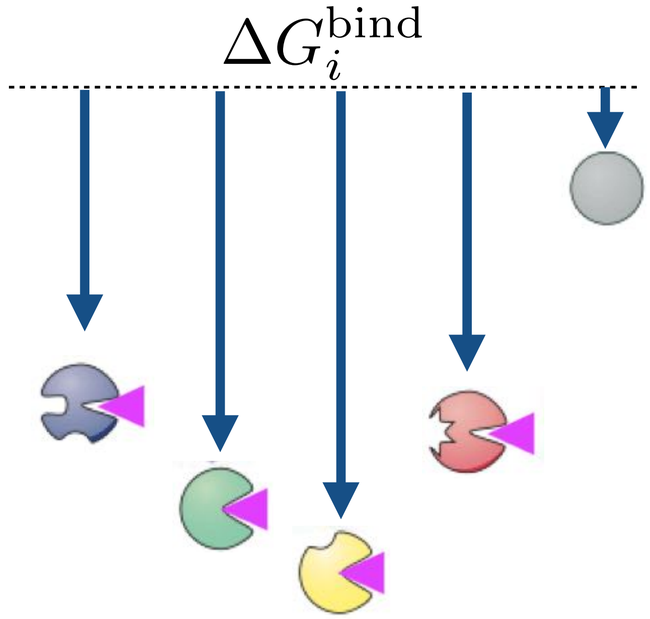


ΔG of trapping conformation

net ΔG of binding



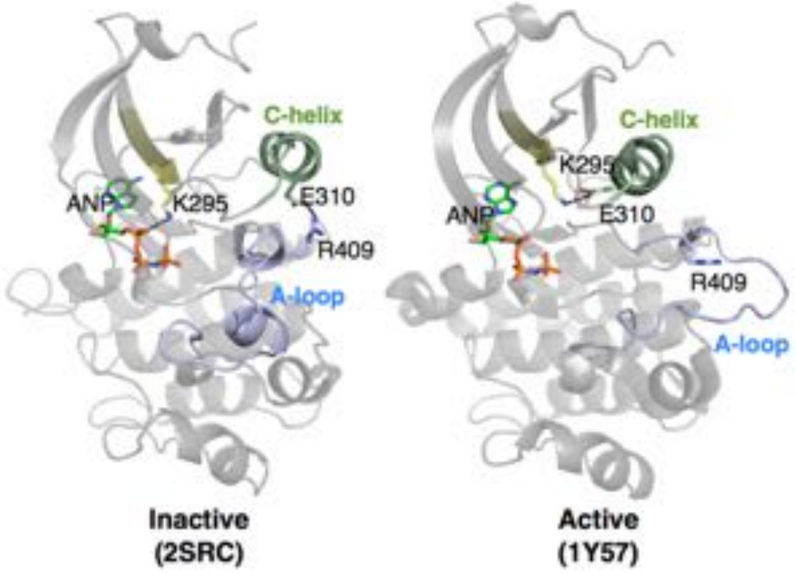
ΔG of binding to trapped conformation



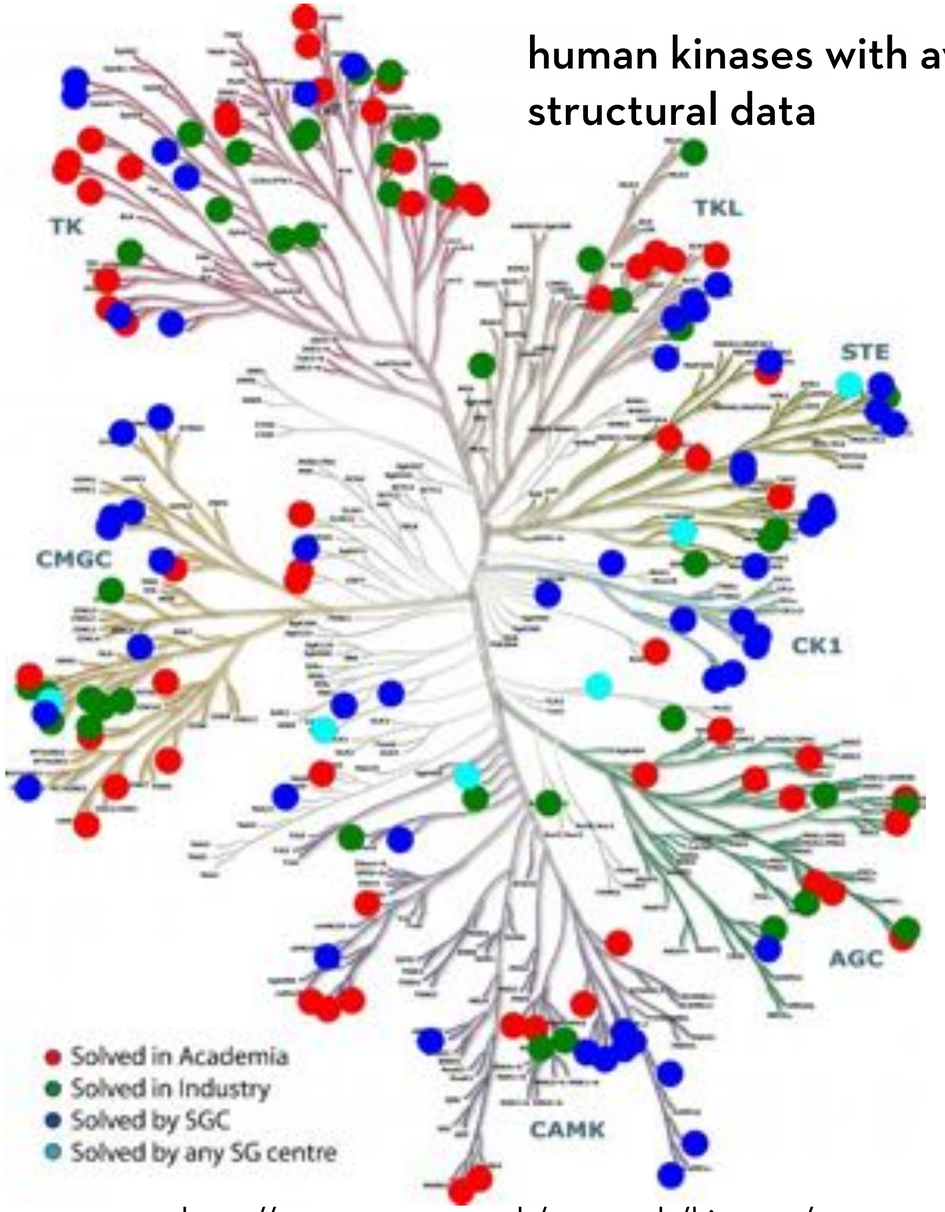
$$\Delta G = -k_B T \ln \sum_i e^{-\beta(\Delta G_i^{\mathrm{conf}} + \Delta G_i^{\mathrm{bind}})}$$

Lee and Craik. Science 324:213, 2009

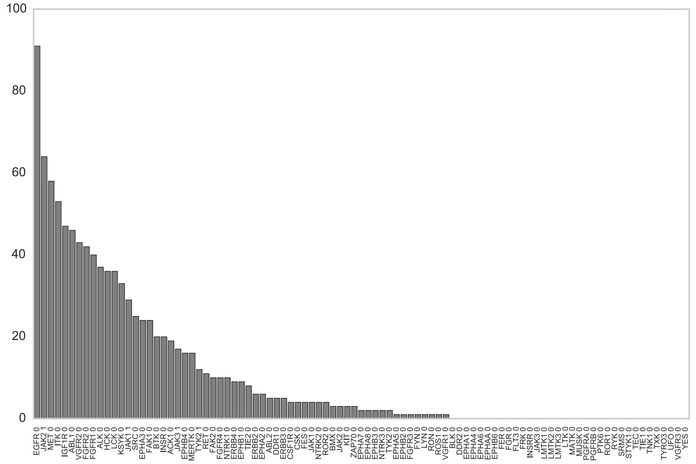
STRUCTURAL DATA ON HUMAN KINASES IS INCOMPLETE



human kinases with available structural data



number of structures/kinase



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

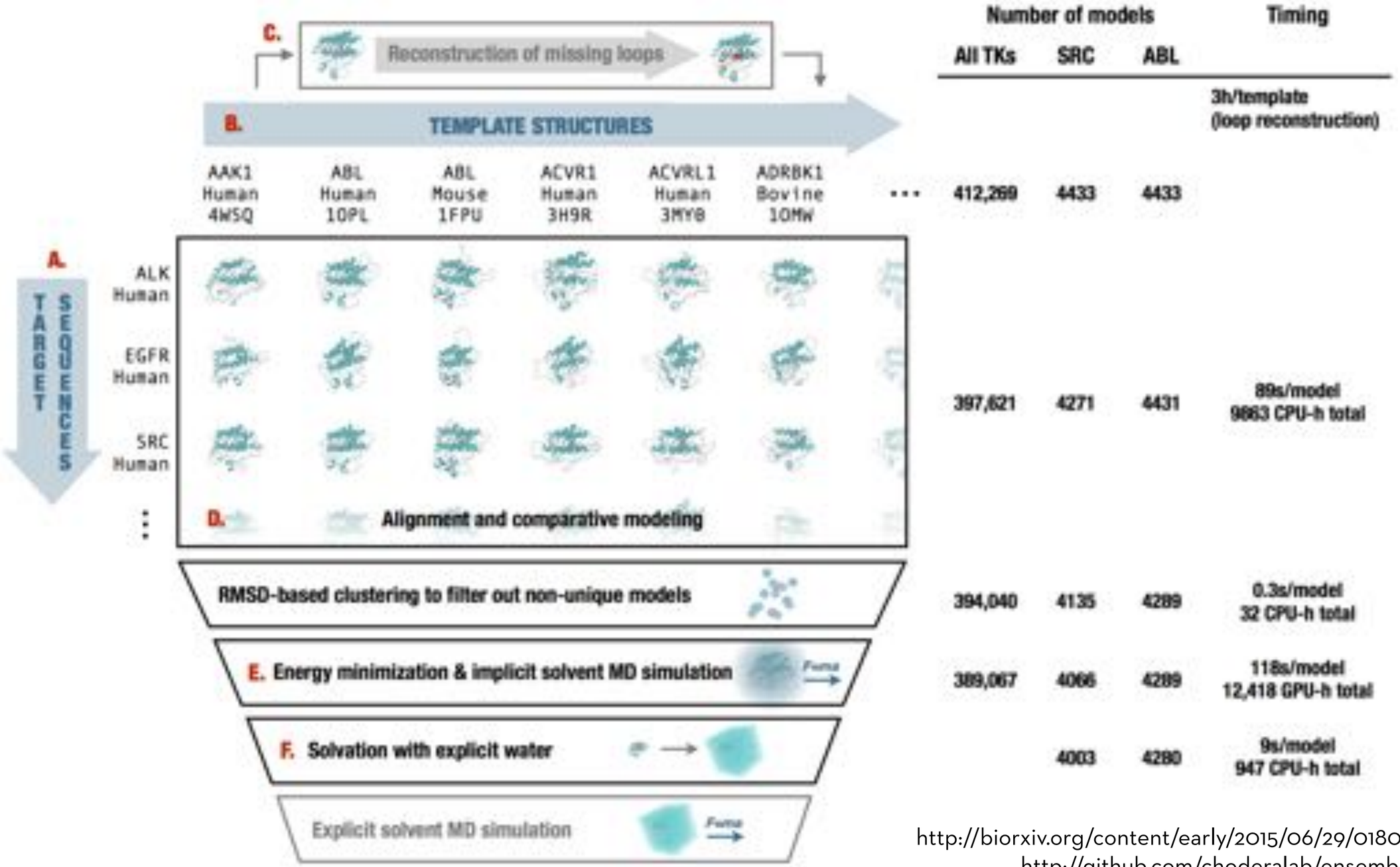
ENSEMBLER: AUTOMATING SIMULATIONS AT THE SUPERFAMILY SCALE

Daniel Parton
Postdoc

Patrick Grinaway
PBSB student

Kyle Beauchamp
Postdoc

Sonya Hanson
Postdoc



http://biorxiv.org/content/early/2015/06/29/018036
http://github.com/choderalab/ensembler

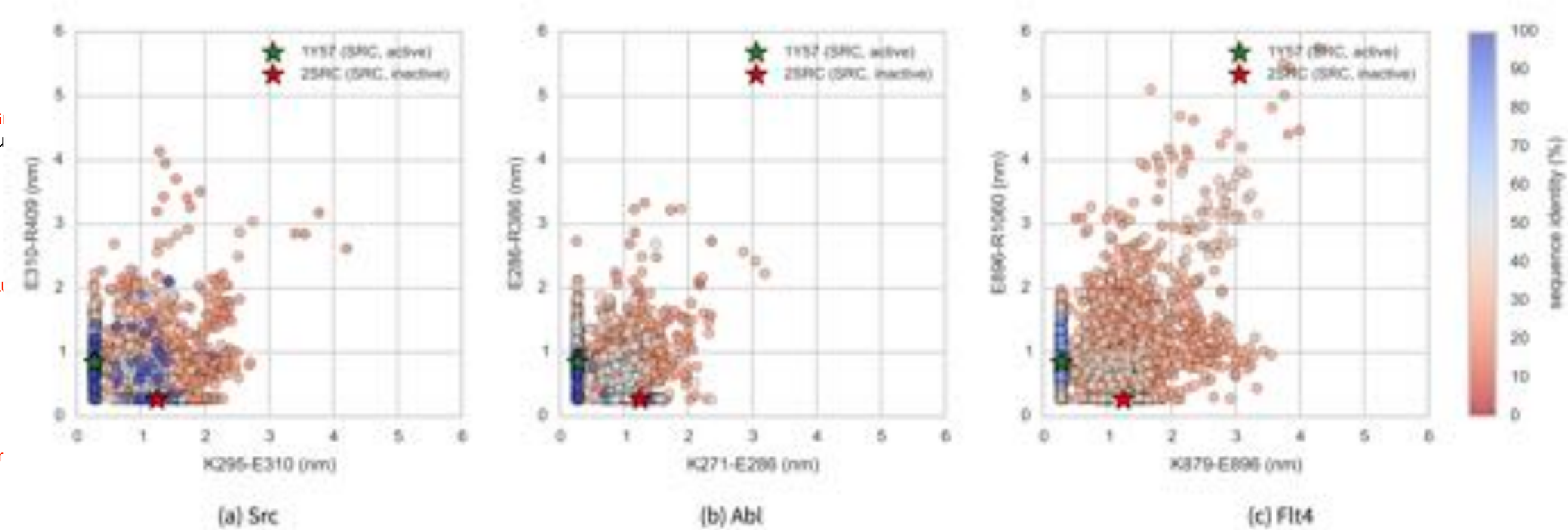
ENSEMBLER: AUTOMATING SIMULATIONS AT THE SUPERFAMILY SCALE

Daniel Parton
Postdoc

Patrick Gi
PBSB stu

Kyle Beau
Postdoc

Sonya Har
Postdoc



http://biorxiv.org/content/early/2015/06/29/018036
http://github.com/choderalab/ensembler

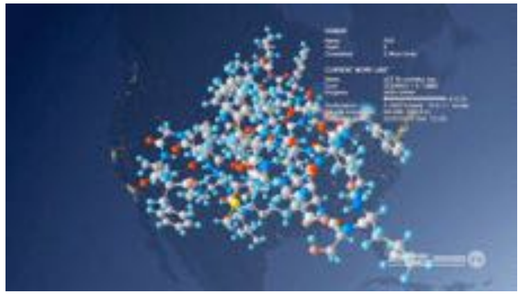
FOLDING@HOME GIVES US ACCESS TO ENORMOUS COMPUTATIONAL RESOURCES FOR PROBING BIOMOLECULAR DYNAMICS

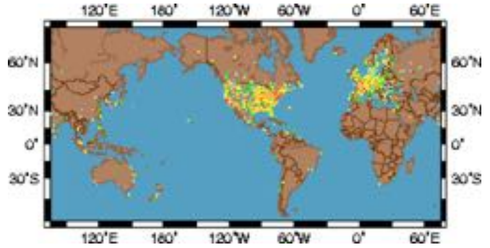
Folding@home DISTRIBUTED COMPUTING

VIJAY S. PANDE
STANFORD UNIVERSITY

OS Type	Native TFLOPS*	x86 TFLOPS*	Active CPUs	Active Cores	Total CPUs
Windows	347	347	79790	199354	5634616
Mac OS X	21	21	7625	55470	184189
Linux	22	22	7011	32192	795767
ATI GPU	1019	2150	7174	7174	399830
NVIDIA GPU	1275	2690	6745	6745	342790
NVIDIA Fermi GPU	12575	26533	37094	135487	535673
Total	15259	31763	145439	436422	7892865

Table last updated at Mon, 01 Jun 2015 23:02:21





OVER 31 PFLOP/S OF AGGREGATE COMPUTATIONAL POWER!

http://folding.stanford.edu

FOLDING@HOME ENABLES WHOLE-KINOME SIMULATION

518 human protein kinases
excluding splice and disease variants

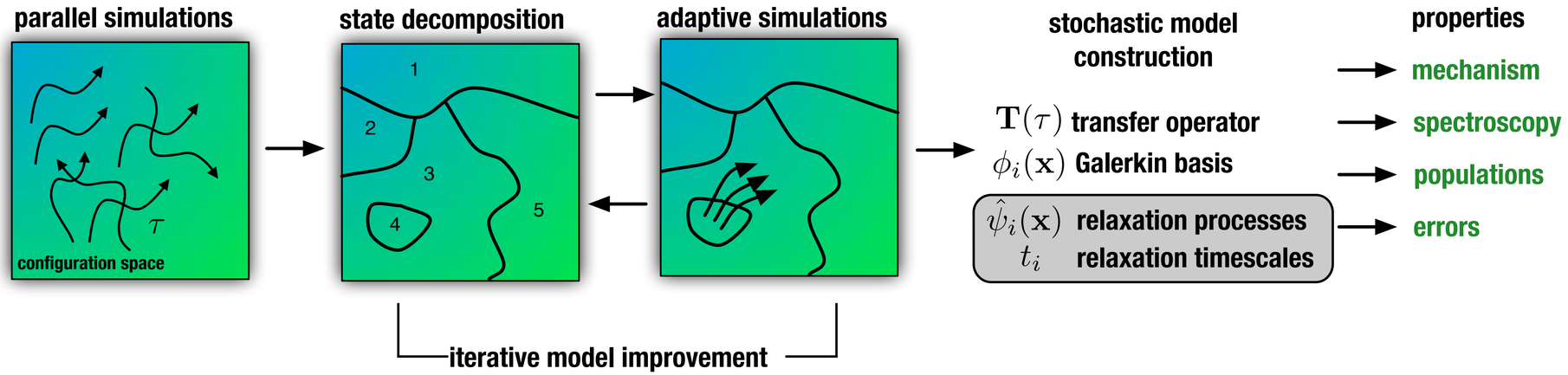
x 3,507 kinase catalytic domain structures
in UniProt

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

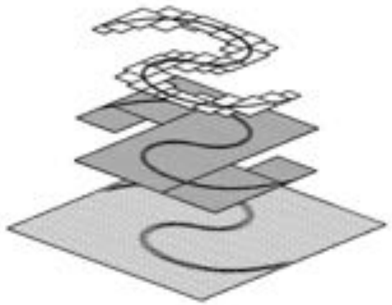
~ 18,166,260 kinase simulations on Folding@Home
over one year

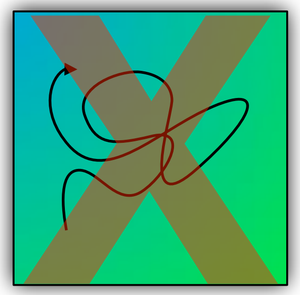
Exploiting scalable, fault-tolerant frameworks (e.g. hadoop, spark, redis, celery, cassandra) is essential to enable scalability to the family scale

AN ADAPTIVE APPROACH CAN BE USED TO REDUCE STATISTICAL DYNAMICS TO A DISCRETE-STATE STOCHASTIC (MARKOV) MODEL



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



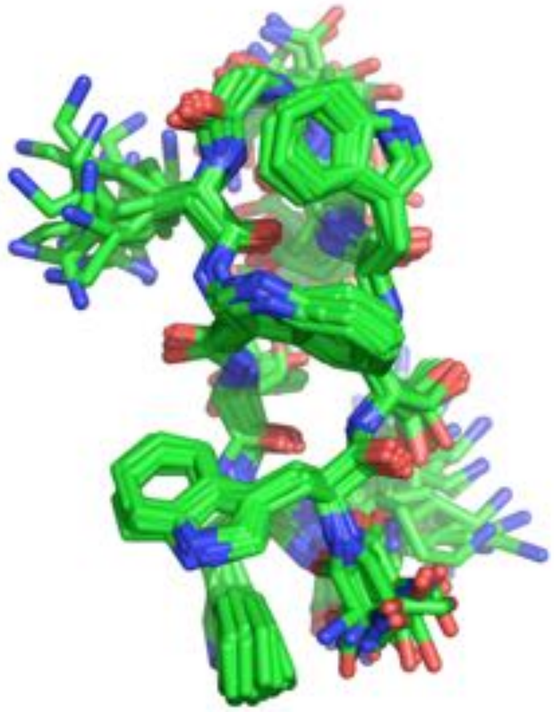


Chodera, Singhal, Swope, Pande, Dill. JCP 126:155101, 2007.
Hinrichs and Pande. JCP 126:244101, 2007.
Bacallado, Chodera, Pande. JCP 131:045106, 2009.
Noé. JCP 128:244103, 2008.
Chodera and Noé. JCP 133:105102, 2010.

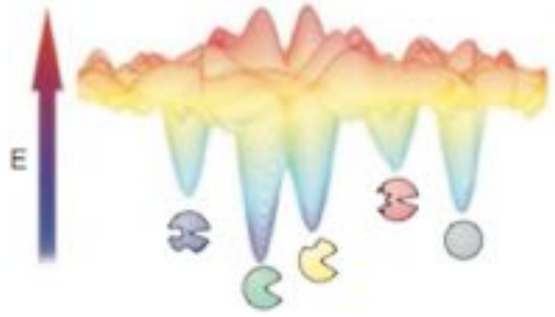
with important contributions from Schütte, Noé, Weber, Hummer, Roux, Vanden-Eijnden

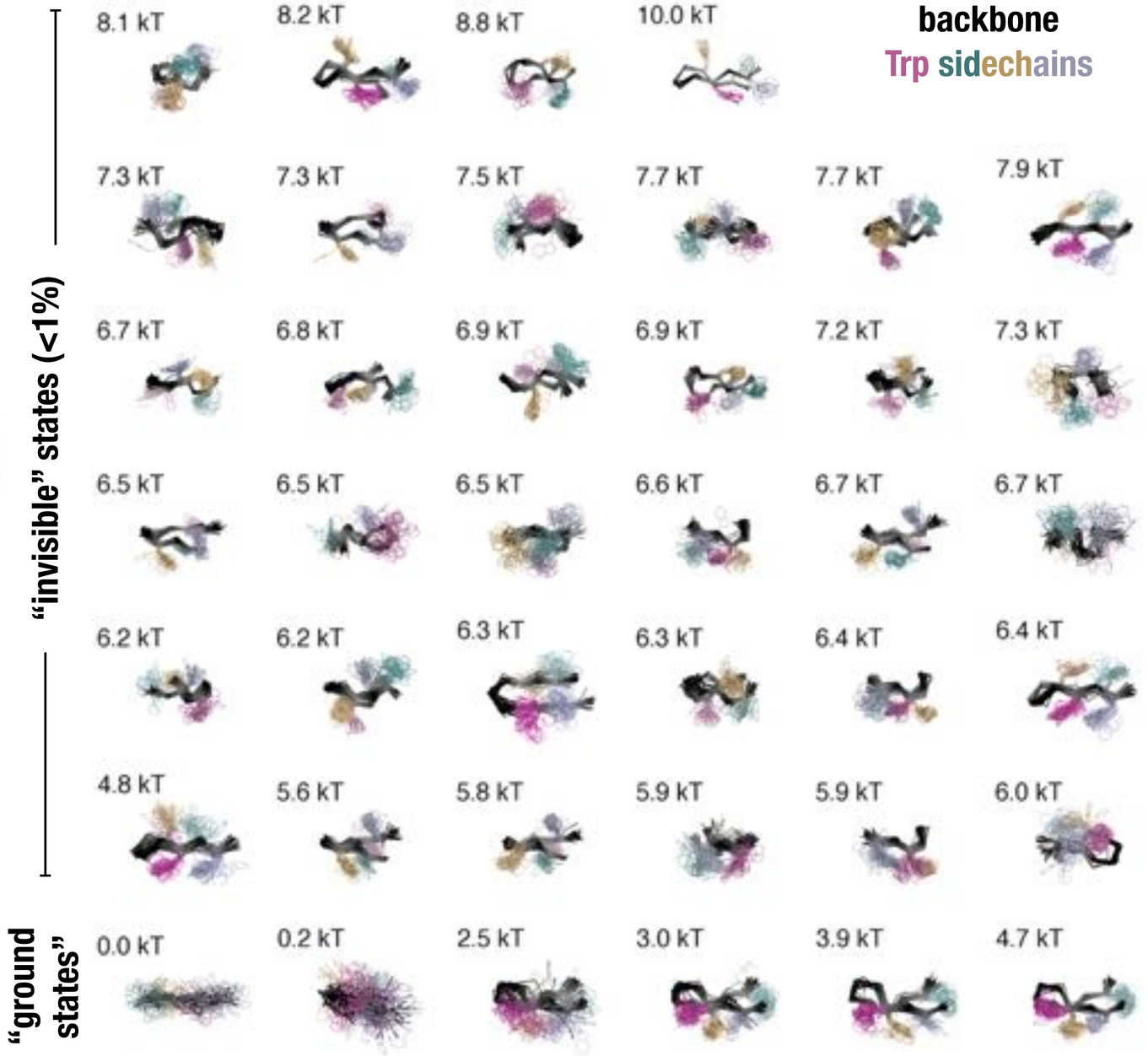
NMR model of trpzip2 at 288 K

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



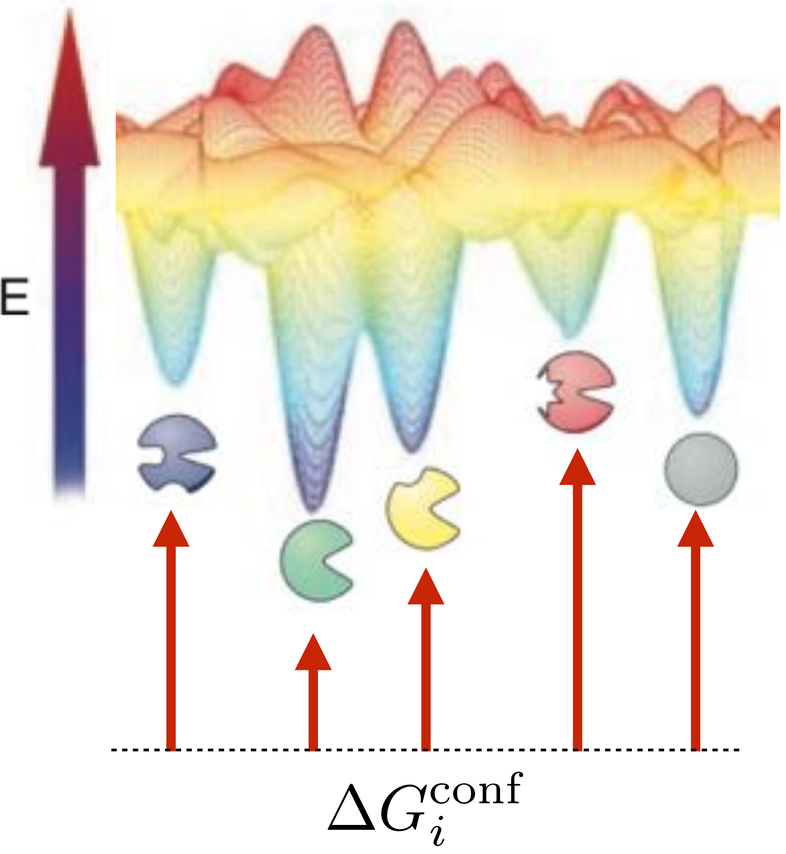
trpzip2
[PDB:1LE1]





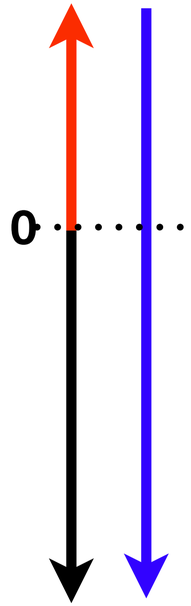
Chodera, Singhal, Swope, Pande, Dill. JCP 126:155101, 2007.

DIVIDE-AND-CONQUER TO COMPUTE AFFINITIES AND SELECTIVITIES

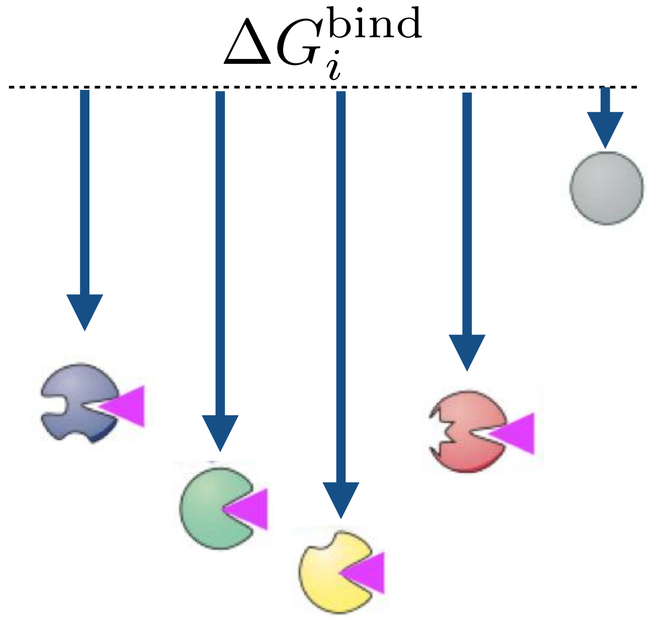


ΔG of trapping conformation

net ΔG of binding



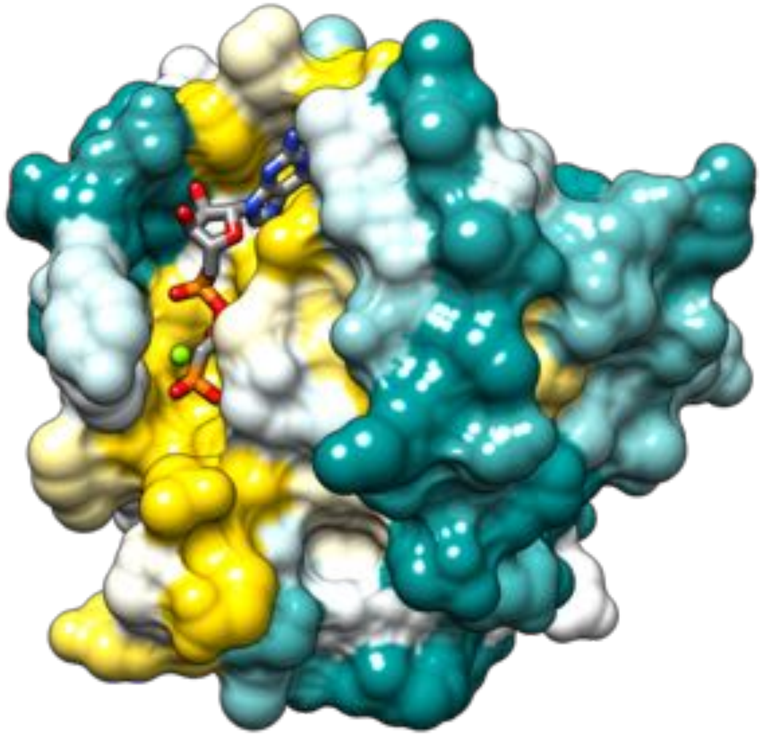
ΔG of binding to trapped conformation



$$\Delta G = -k_B T \ln \sum_i e^{-\beta(\Delta G_i^{\mathrm{conf}} + \Delta G_i^{\mathrm{bind}})}$$

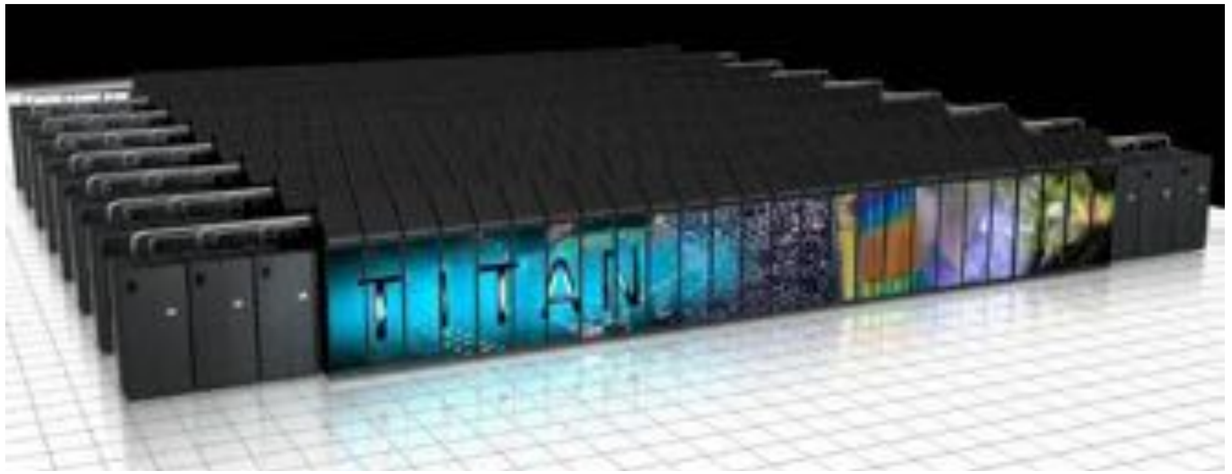
Lee and Craik. Science 324:213, 2009

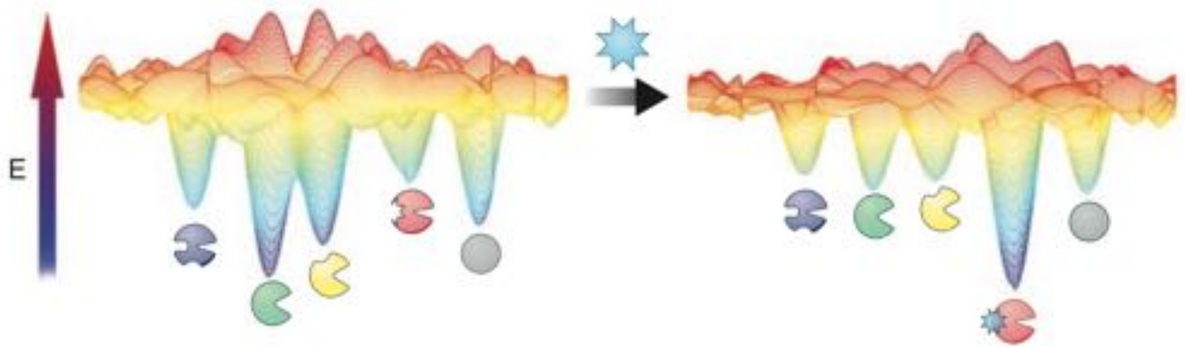
CAN WE DRUG THE UNDRUGGABLE? ALLOSTERIC MODULATORS OF K-RAS MAY OPEN NEW DOORS IN CANCER THERAPY



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)

YANK ROADMAP

Q4 2016

YANK 1.0 RELEASE (LEVI)
TUTORIALS / BEST PRACTICES / DOCS / TESTS

Q1 2017

ROBUST WORKFLOW PIPELINE (LEVI)
SINGLE REPLICA CALCULATIONS (JOHN/ANDREA)
DYNAMIC PROTONATION STATES (BAS)

Q2 2017

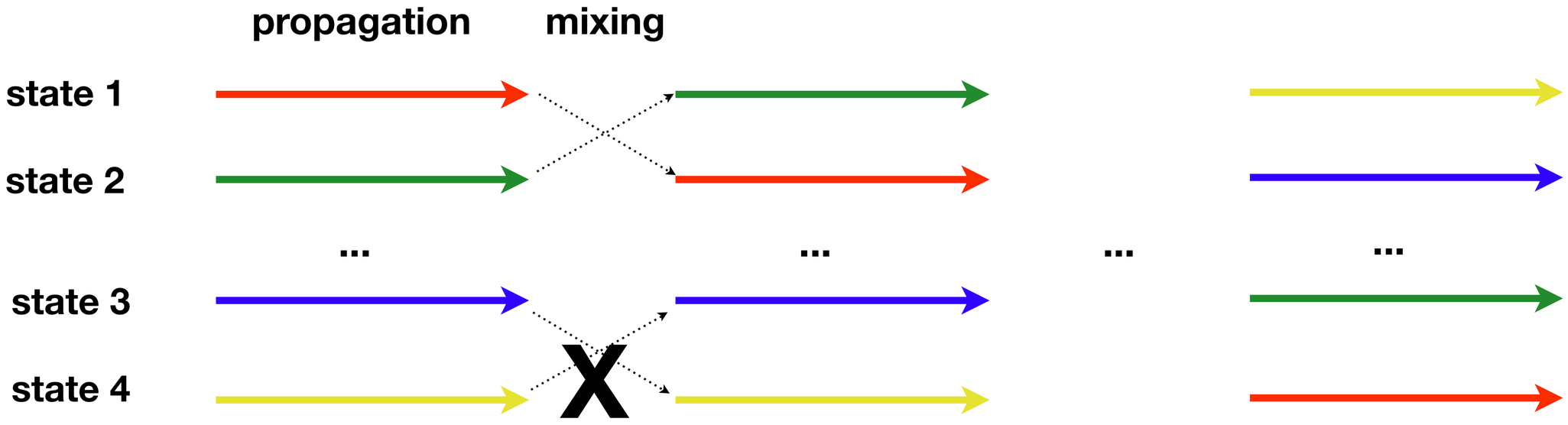
MULTIPLE FORCEFIELD SUPPORT (LEVI)
MULTIPLE ALCHEMICAL REGIONS (JOHN)
DYNAMIC COUNTERIONS (GREG)

Q3 2017

RELATIVE FREE ENERGY CALCULATIONS (JULIE)
PERSES AUTOMATED DESIGN (PATRICK/JULIE/STEVEN)
SIMULTANEOUS KINETICS AND FREE ENERGIES

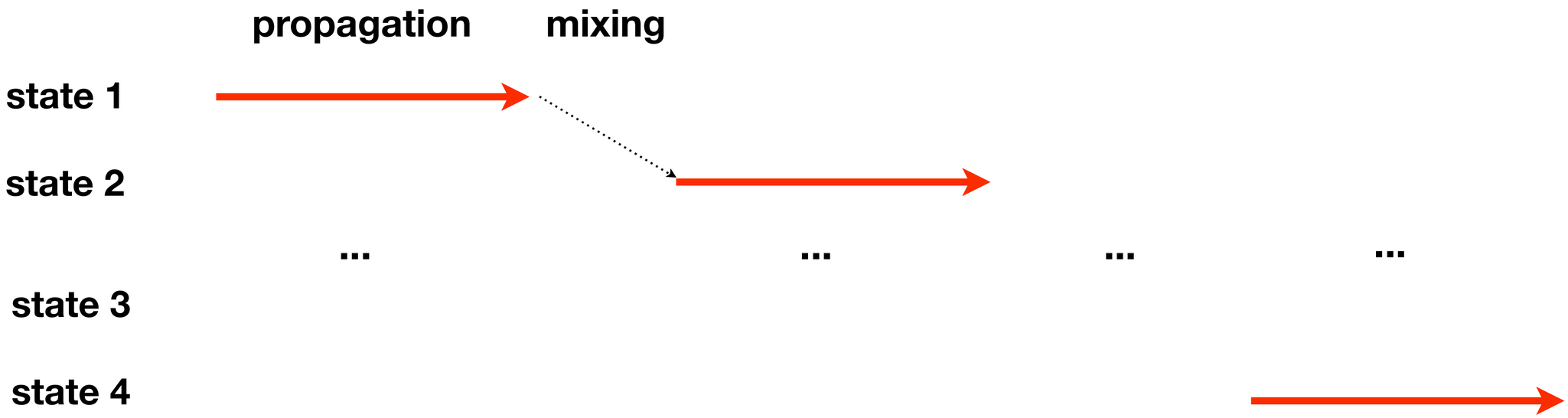
SINGLE-REPLICA METHODS

DO WE NEED ALL THOSE REPLICAS?



HAMILTONIAN EXCHANGE

DO WE NEED ALL THOSE REPLICAS?

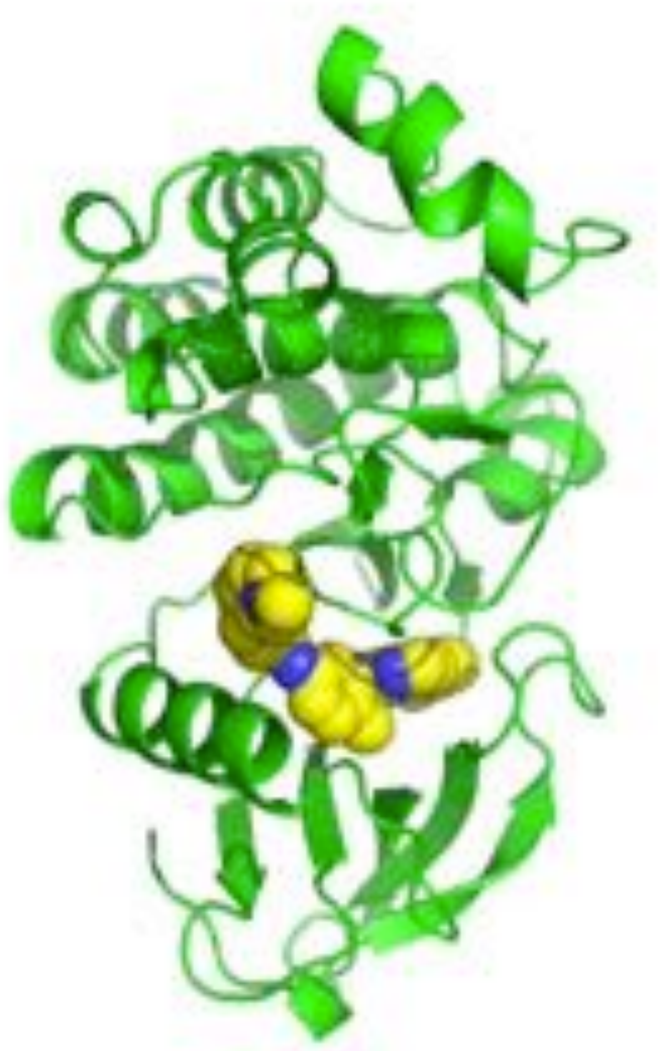


EXPANDED ENSEMBLE

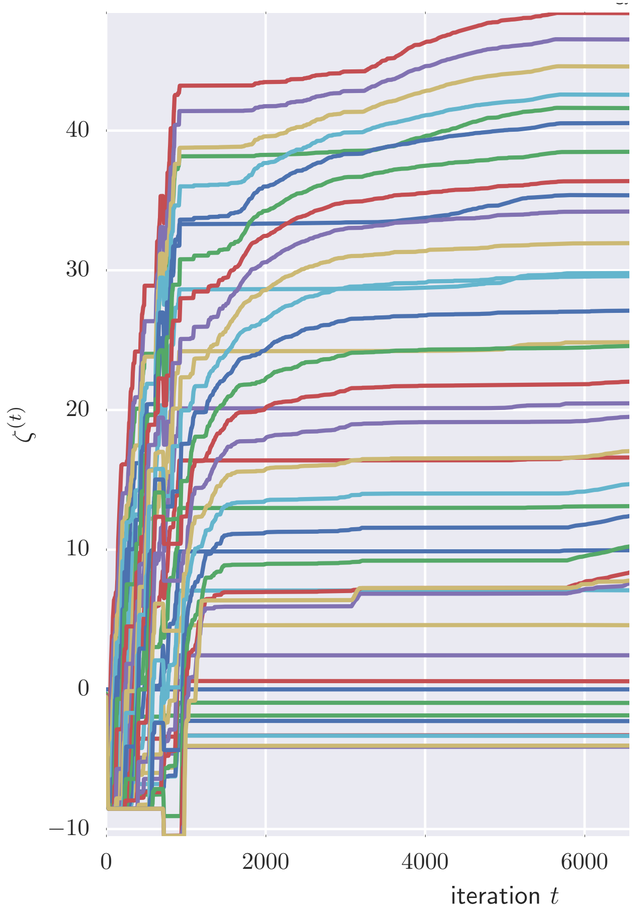
$$\pi(x,k) \propto \exp[-u_k(x) + g_k]$$

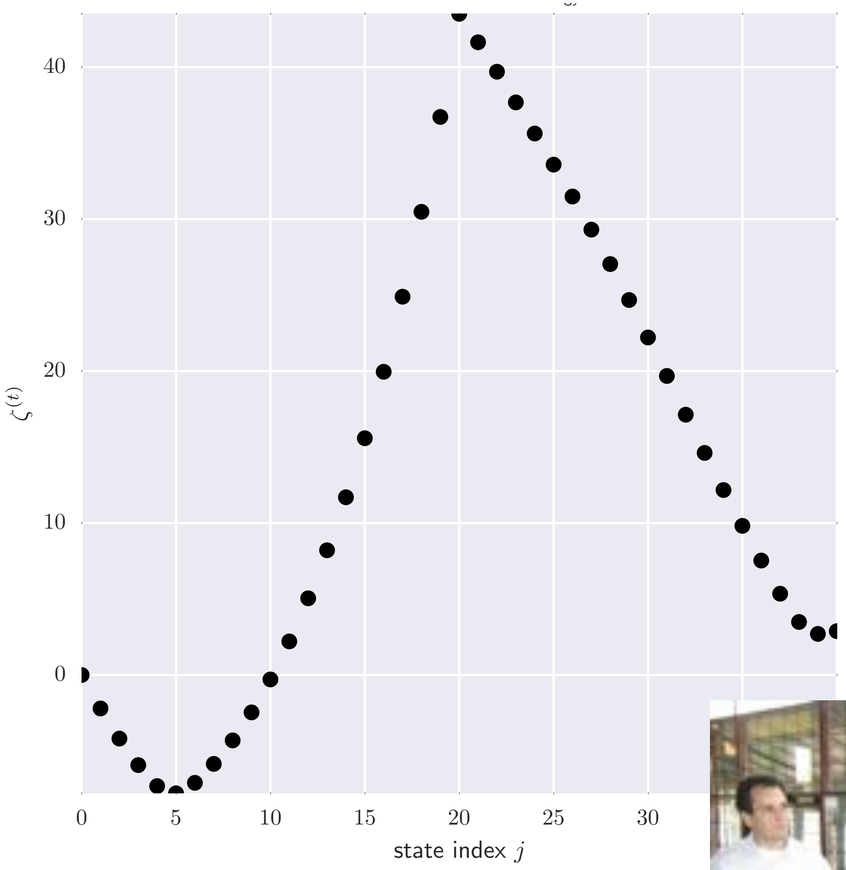
One caveat: We need to guess the weights g
(which are unfortunately the free energies we are trying to compute!)

SELF-ADJUSTED MIXTURE SAMPLING (SAMS)



Provably asymptotically optimal strategy for finding free energy weights!
Tan Z. J. Comp. Graph. Stat. http://dx.doi.org/10.1080/10618600.2015.1113975

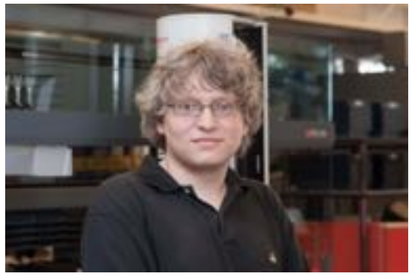




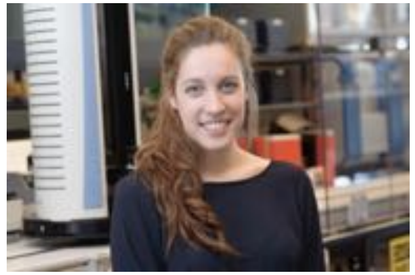
Zhiqiang Tan
Rutgers

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

AUTOMATED DESIGN



PATRICK GRINAWAY
Synthetic route-guided multiobjective ligand optimization
Allosteric inhibition of mutant KRAS

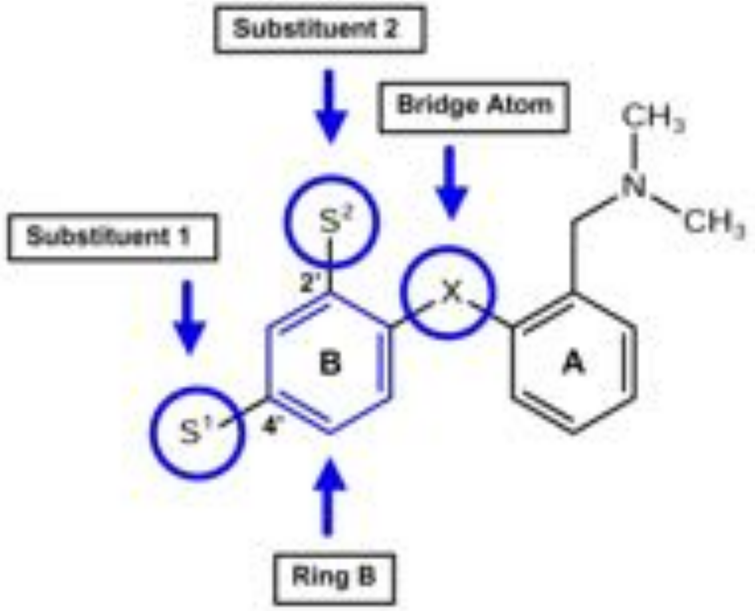


JULIE BEHR
Predicting the evolution of resistance mutations
Protein and peptide design

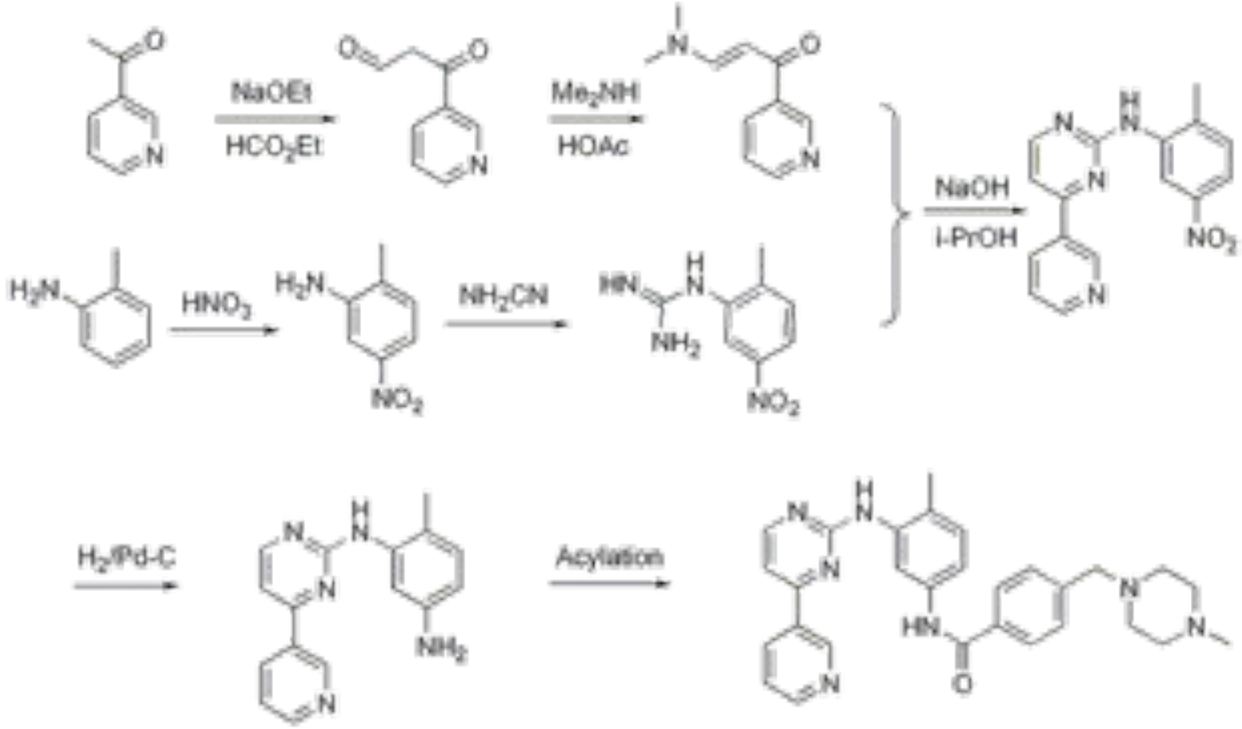
STEVEN ALBANESE
Selective kinase inhibitor design
Designing for polypharmacology

HOW CAN WE DESIGN MOLECULES WITH DESIRED AFFINITIES AND SELECTIVITIES?

alternative substituents



alternative starting materials



How can we search large chemical spaces based on free energy objectives?

HOW CAN WE DESIGN MOLECULES WITH DESIRED AFFINITIES AND SELECTIVITIES?

Can we sample the joint space of configuration x and chemical state i so that the marginal chain maximizes a desired objective function?

$$\pi(i) \propto \int dx\, \pi(x, i)$$

Express objectives in terms of ratios of partition functions:

Maximize target affinity

$$\pi(i) \propto \frac{Z_{PL^{(i)}}}{Z_{L^{(i)}}} \propto K_a^{(i)}$$

Maximize selectivity for target protein (or conformation) P1 over antitarget P2

$$\pi(i) \propto \frac{Z_{P_1 L^{(i)}}}{Z_{P_2 L^{(i)}}} \propto \frac{K_{a,1}^{(i)}}{K_{a,2}^{(i)}}$$

Select resistance mutations that minimize inhibitor affinity while maintaining activity

$$\pi(i) \propto \frac{Z_{P^{(i)} S}}{Z_{P^{(i)} I}} \propto \frac{K_S^{(i)}}{K_i^{(i)}}$$

HOW CAN WE DESIGN MOLECULES WITH DESIRED AFFINITIES AND SELECTIVITIES?

SAMS allows us to construct a NEW recursion scheme to achieve a desired marginal distribution in terms of ratios of partition functions:

sample a new configuration with MCMC (e.g. hybrid Monte Carlo)

$$x_{n+1} \sim p(x|s_n)$$

sample a new chemical state with Monte Carlo

$$s_{n+1} \sim p(s|x_{n+1}, \{Z_s\}, \{\pi_s\}) \propto \frac{\pi_s}{Z_s} e^{-u_s(x)}$$

update free energy estimates using recursion

$$\log Z_{s,n+1} = \log Z_{s,n} - \frac{1}{n} \frac{w_s}{\pi_s}$$

update sampling target probabilities to maximize objective using recursion

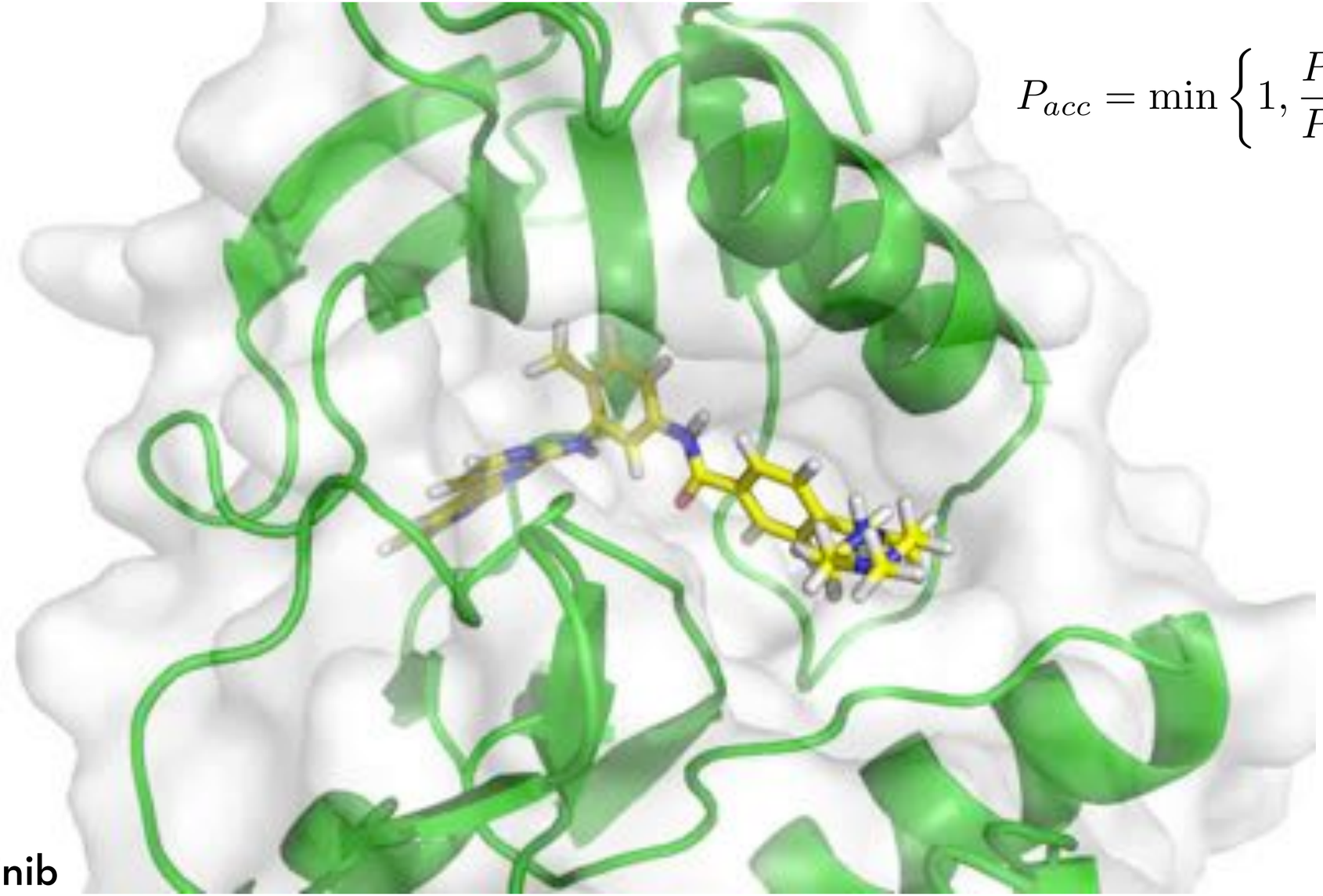
$$\pi_{s,n+1} = \frac{Z_{PL^{(s)}}}{Z_{L^{(s)}}}$$

Stochastic approximation theory: Z. Tan. J. Comp. Graph. Stat. 2015

Inspired by J. W. Pitera Proteins 15:385, 2000.

SAMPLING A NEW CHEMICAL STATE WITH REVERSIBLE JUMP MONTE CARLO (RJMC)

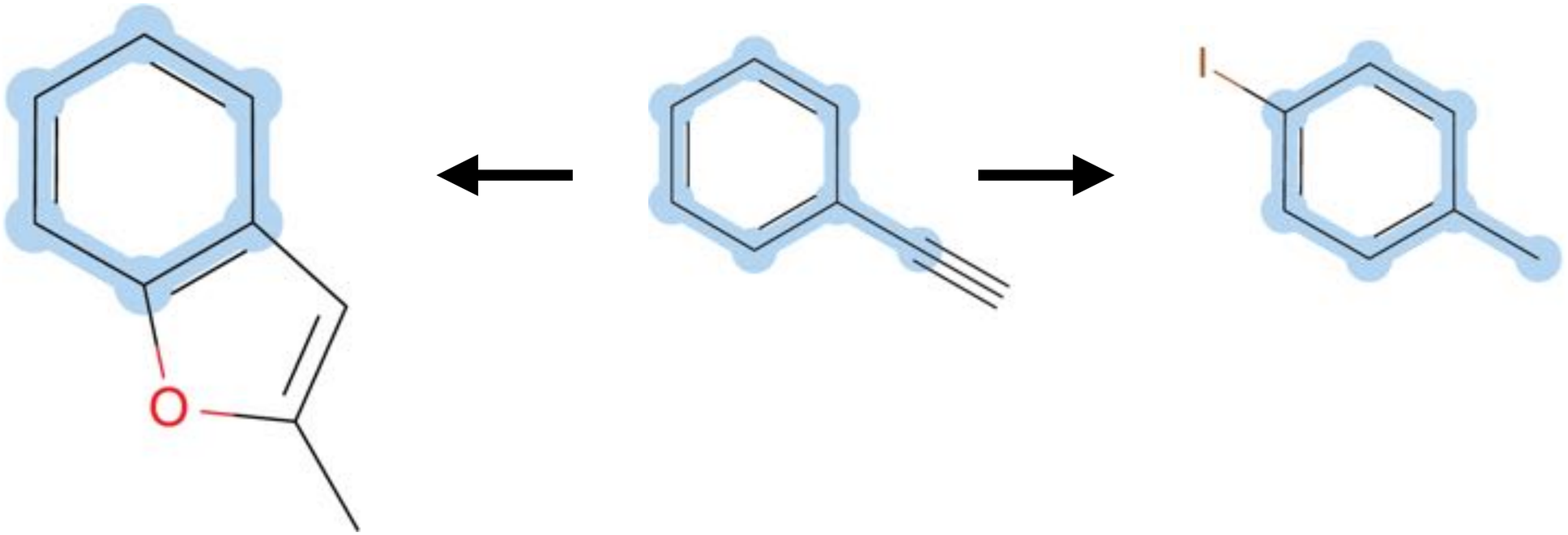
$$P_{acc} = \min\left\{1, \frac{P(\text{old}|\text{new})}{P(\text{new}|\text{old})}\frac{\pi(new)}{\pi(\text{old})}\right\}$$



Abl kinase
imatinib >> nilotinib

CHEMICAL MONTE CARLO MOVES

PROPOSE NEW MOLECULES WITH COMMON SCAFFOLD



CHEMICAL MONTE CARLO MOVES

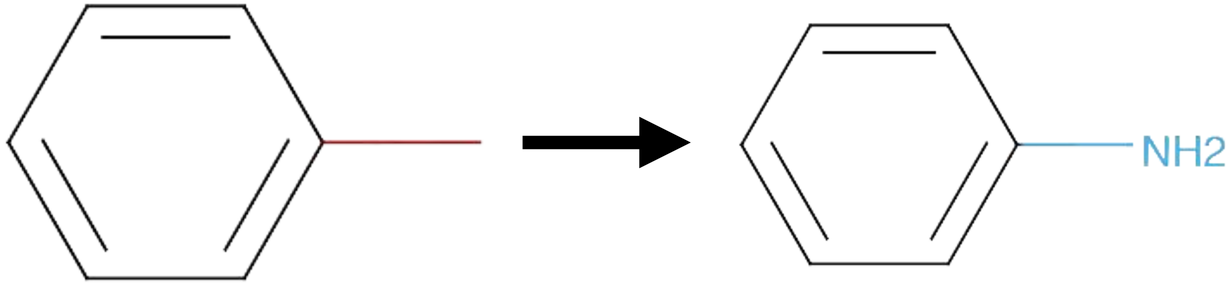
RJMC

DELETE

OLD ATOMS

CREATE

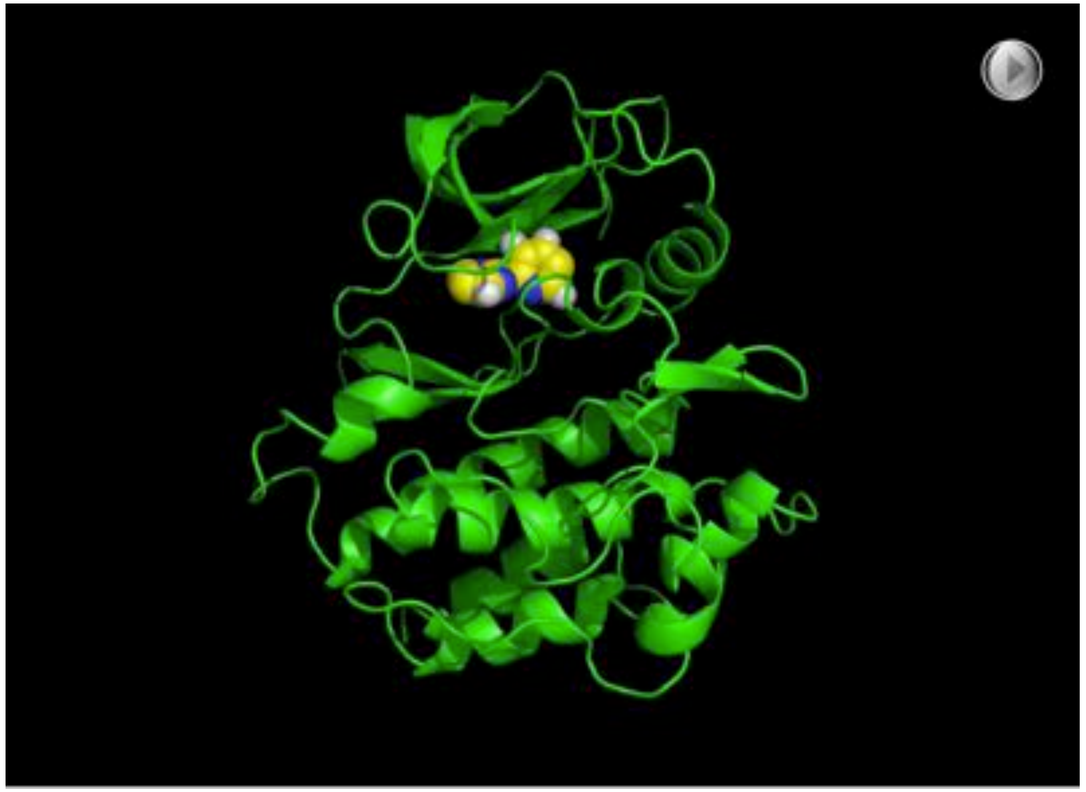
NEW ATOMS



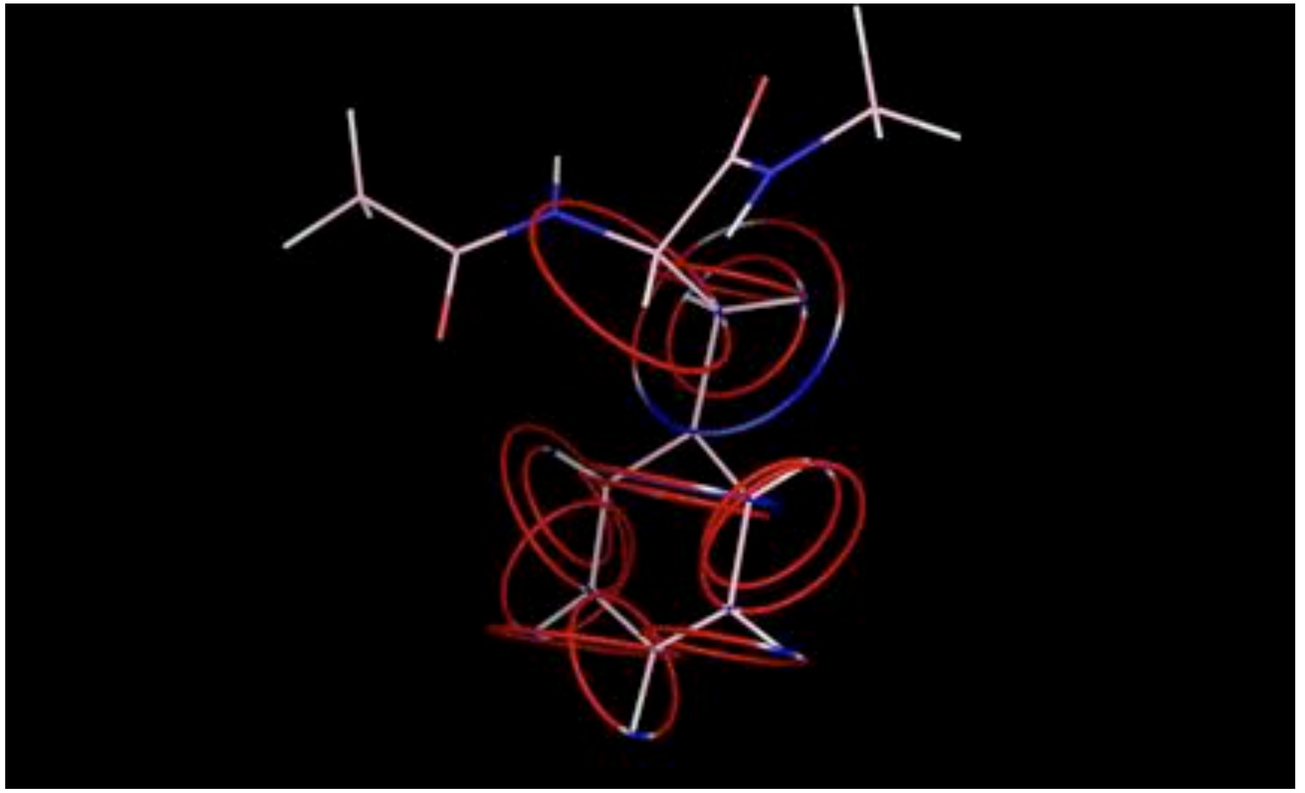
15 atoms

14 atoms

CHEMICAL MONTE CARLO MOVES
RJMC



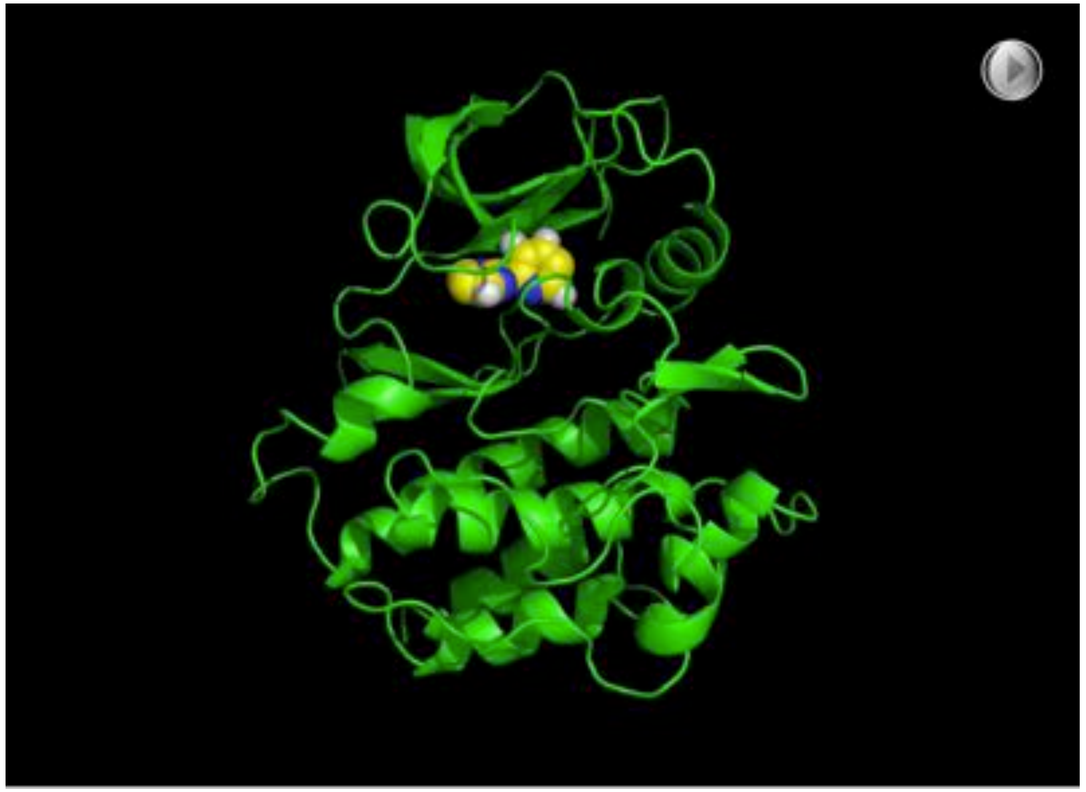
SMALL MOLECULES

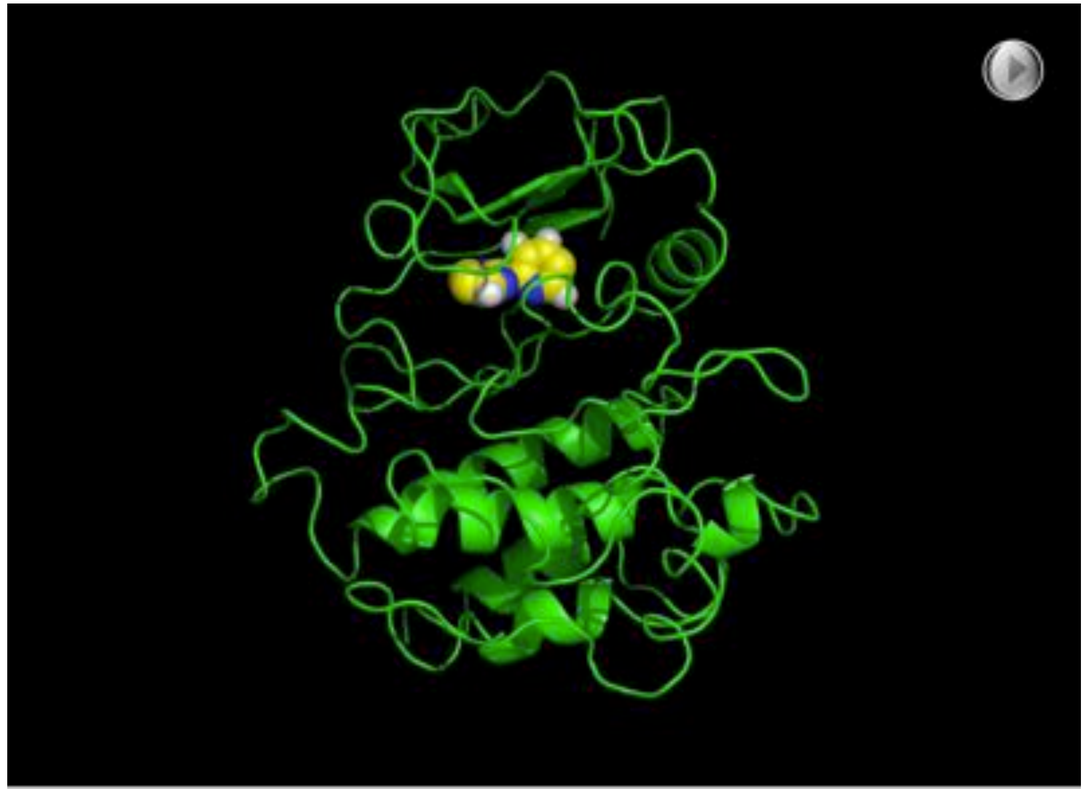


AMINO ACIDS

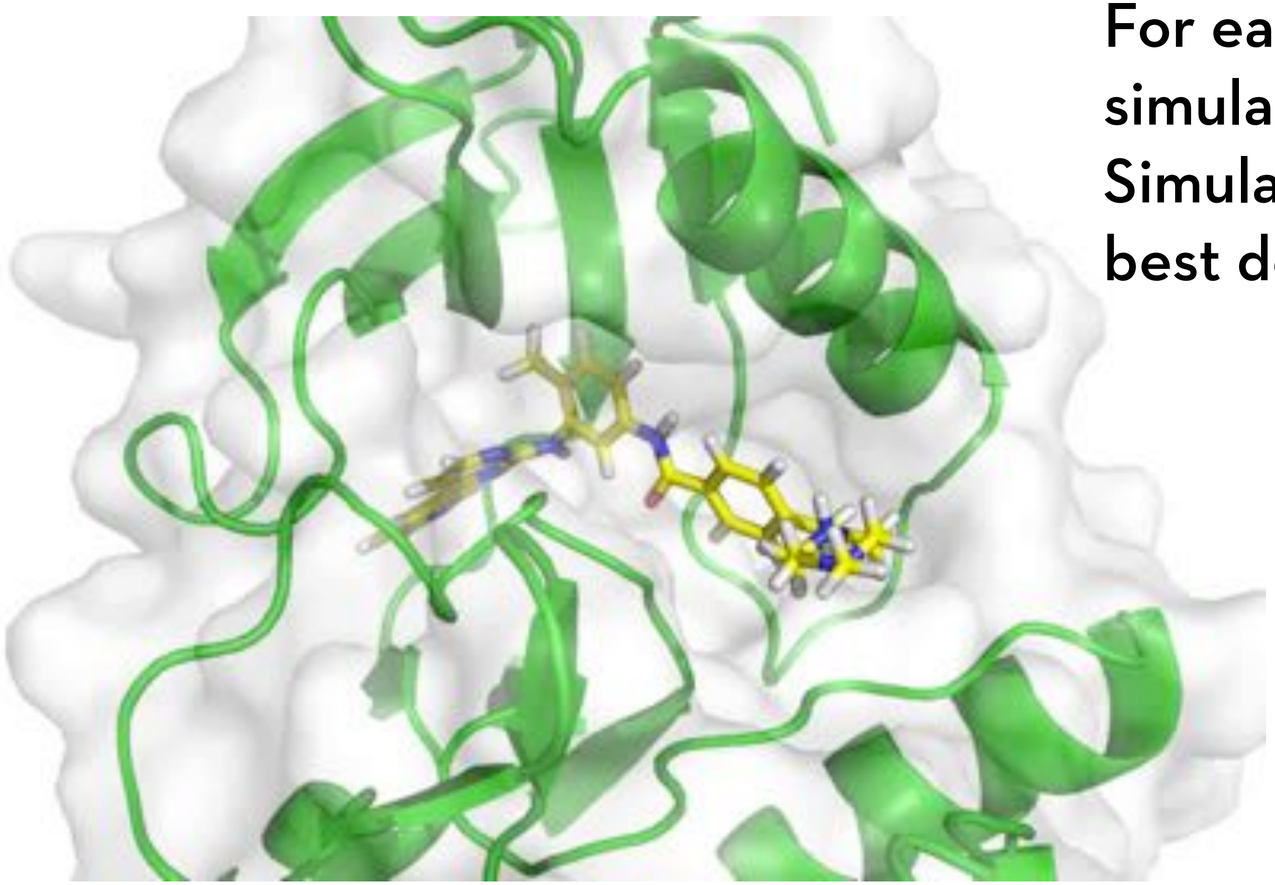
CHEMICAL MONTE CARLO MOVES:

RJMC + NCMC

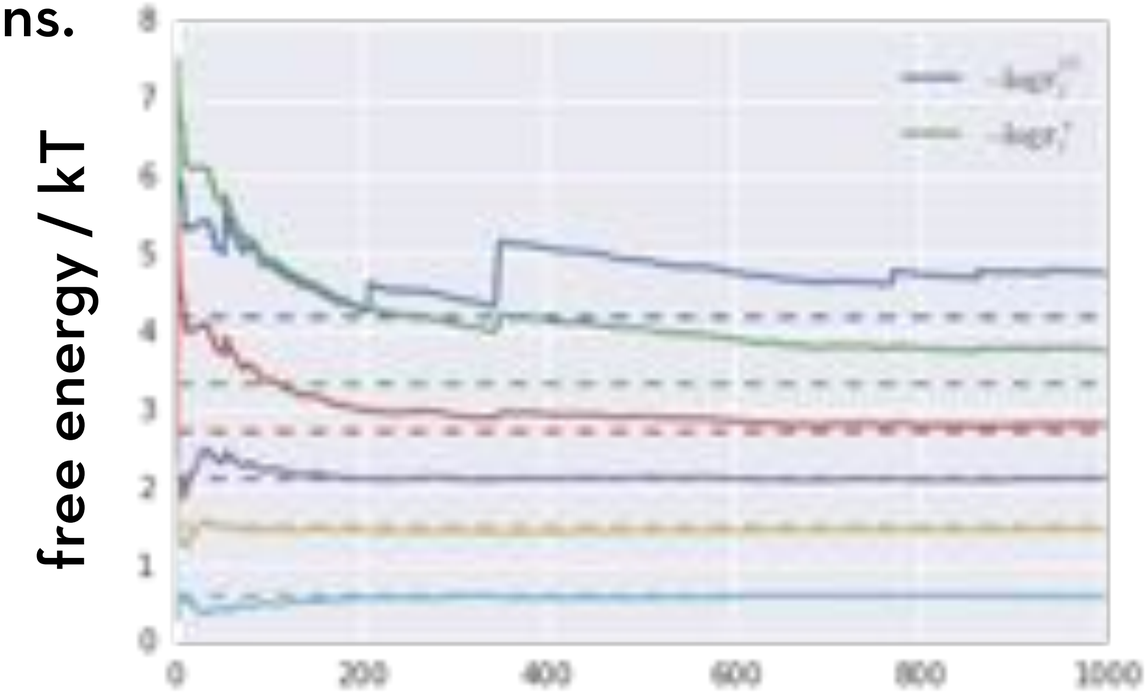




PROJECT PERSES



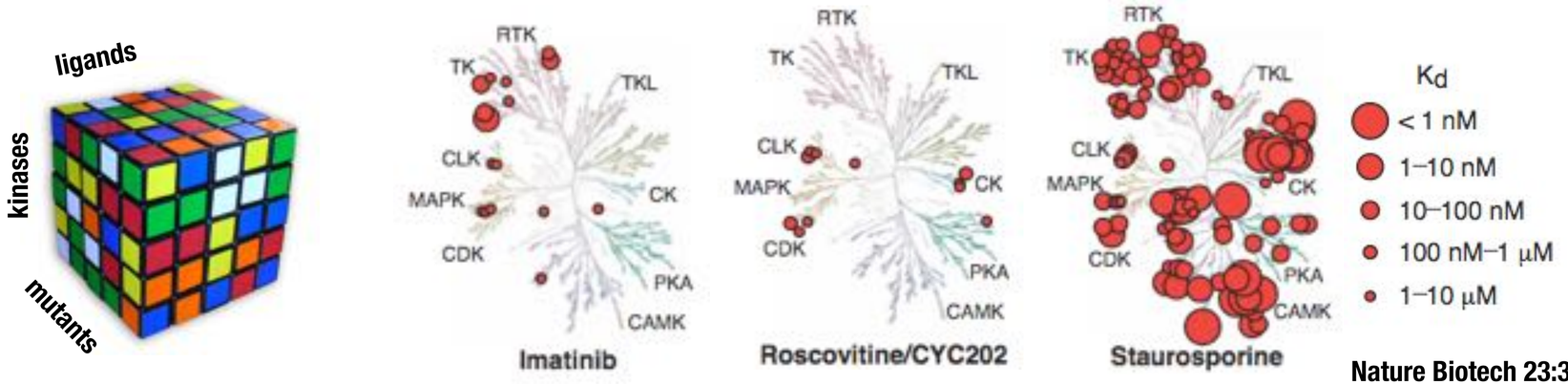
For each target/antitarget, we run a coupled simulation that can visit multiple chemical states. Simulations are biased to spend more time visiting best designs.



https://github.com/choderalab/perses [in development]

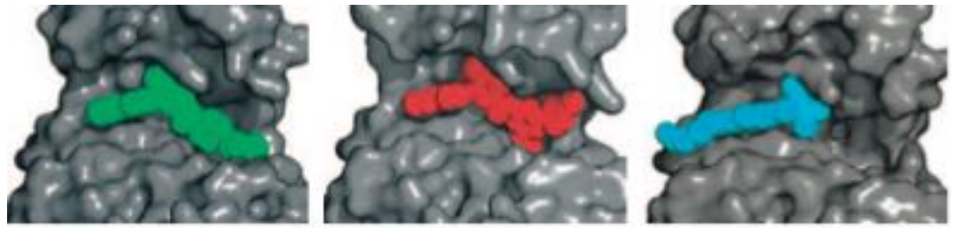
WHAT DETERMINES THE SELECTIVITY OF KINASE INHIBITORS?

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?

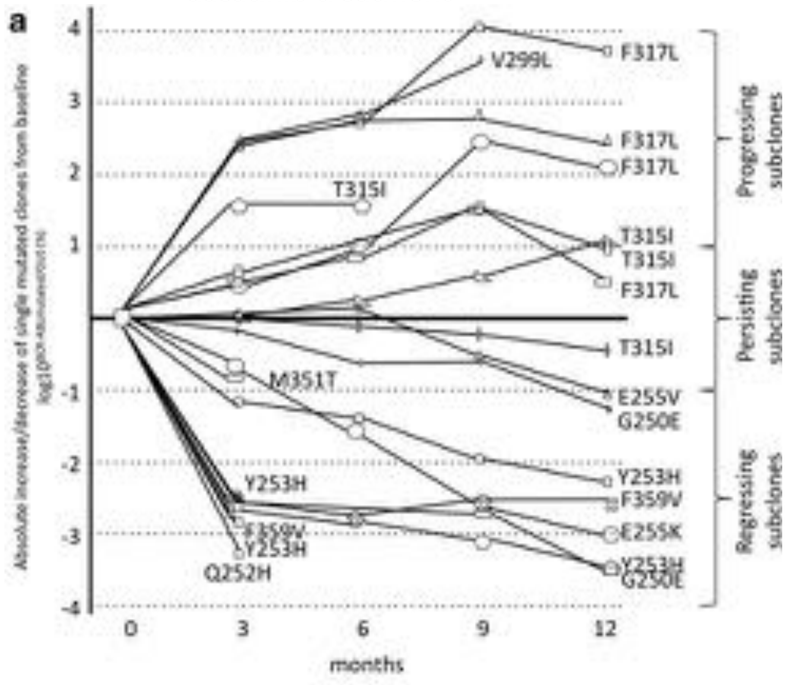


Weisberg et al.
Nature Rev. Cancer 7:345, 2007.

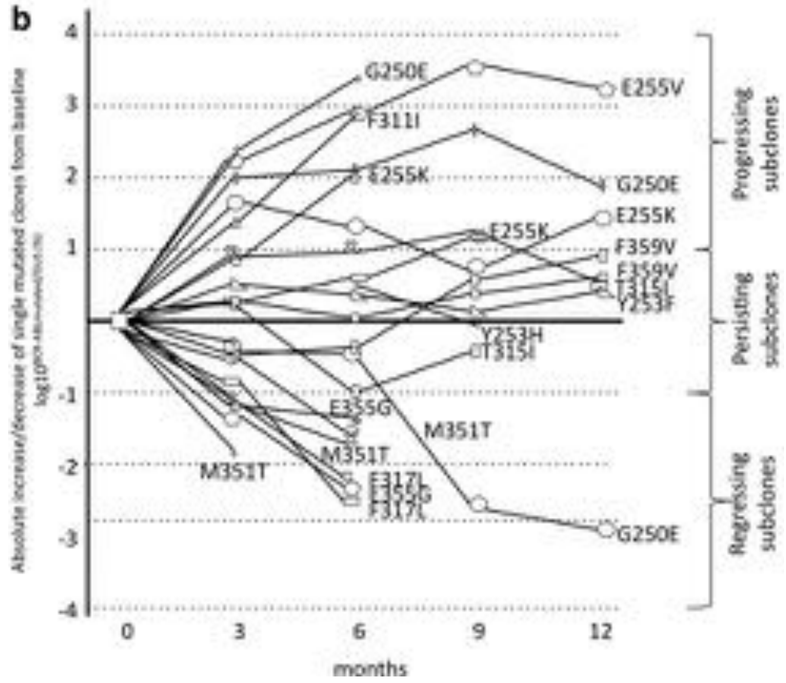
CAN WE DEVELOP A PHYSICAL MODEL OF RESISTANCE?

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

dasatinib treatment



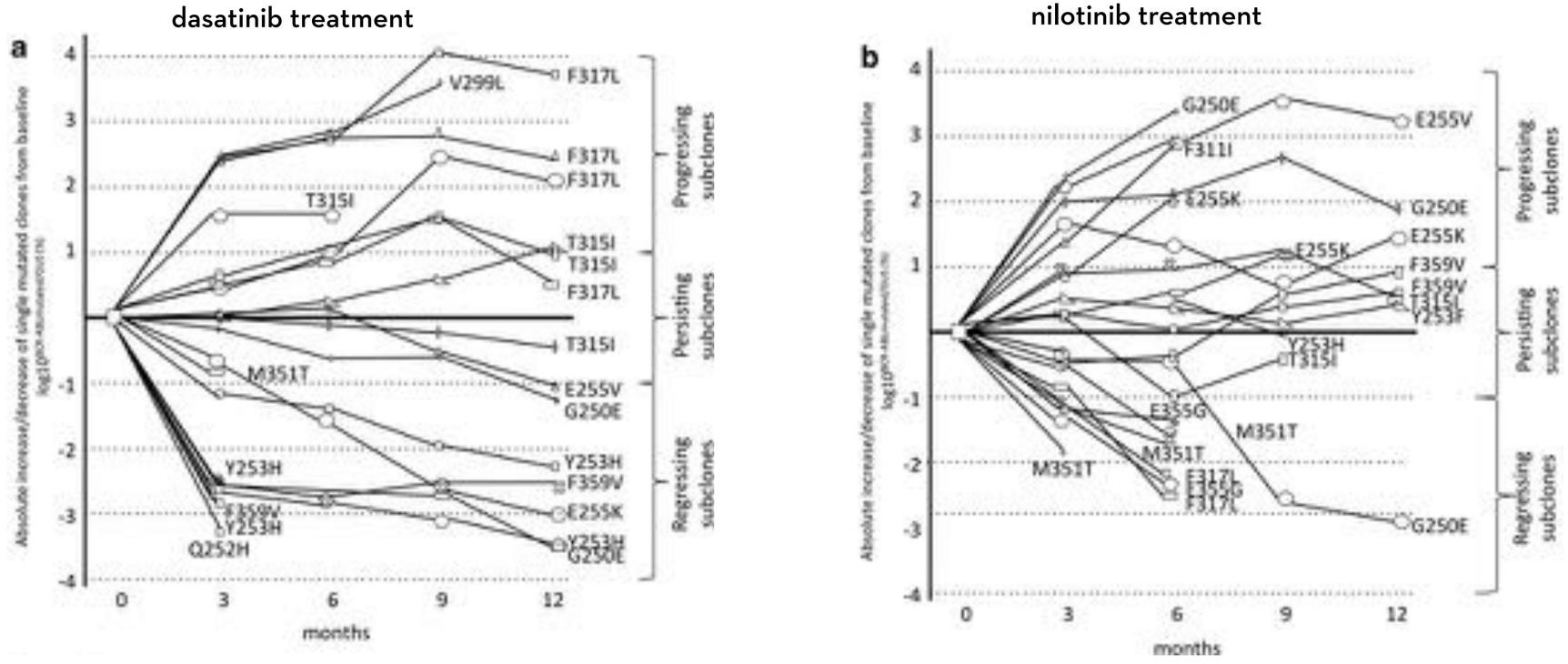
nilotinib treatment



Gruber et al. Leukemia 26:172, 2012.

CAN WE DEVELOP A PHYSICAL MODEL OF RESISTANCE?

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



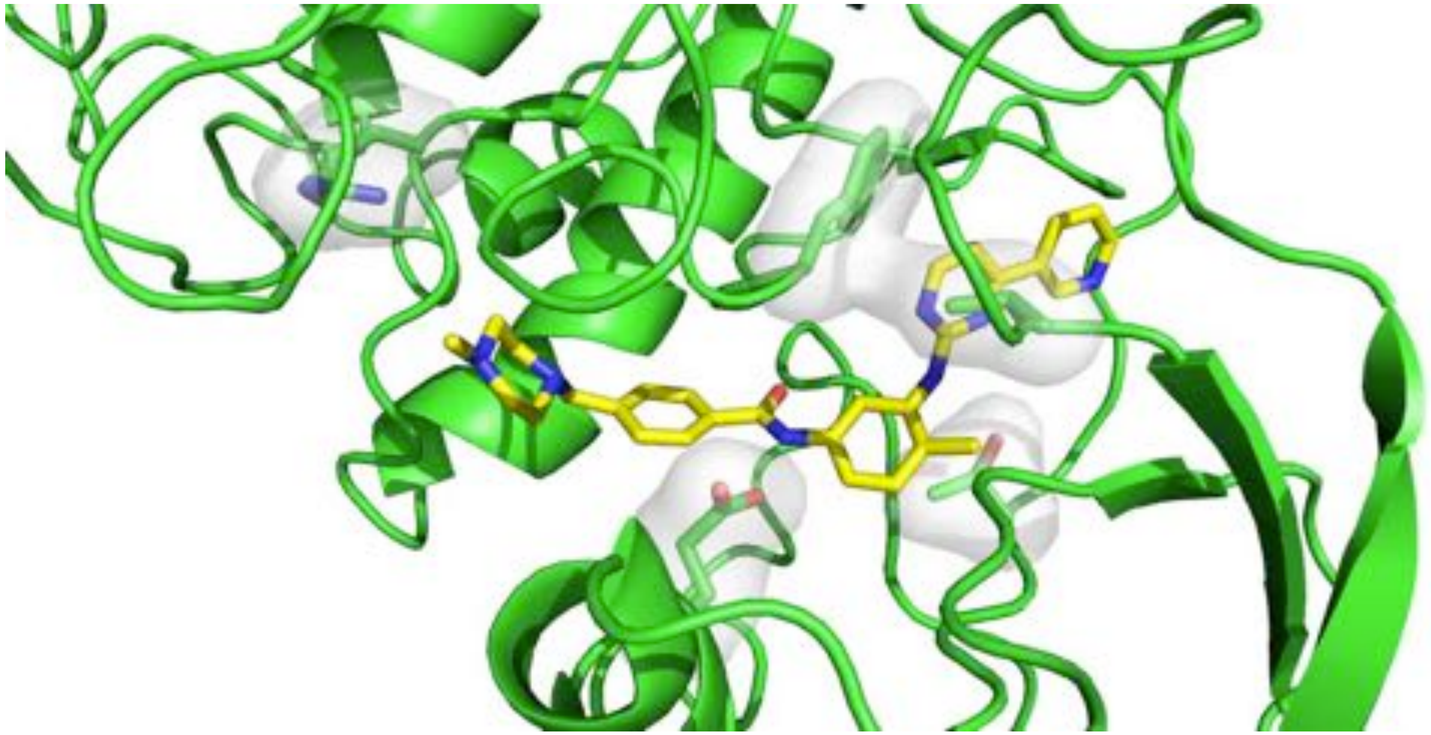
Gruber et al. Leukemia 26:172, 2012.

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?

CAN WE DEVELOP A PHYSICAL MODEL OF RESISTANCE?

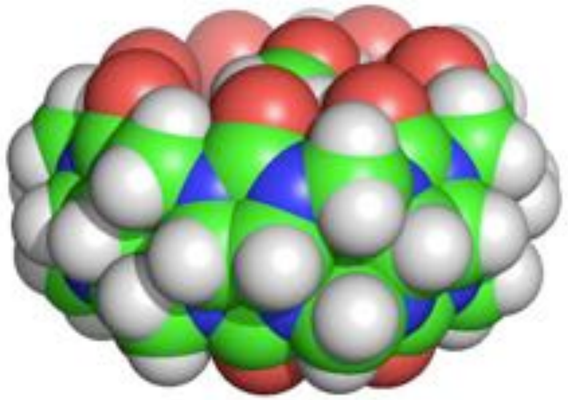
$$P(\text{mutant}) \propto \frac{K_{\text{ATP}}^{\text{mutant}}}{K_{\text{I}}^{\text{mutant}}}$$



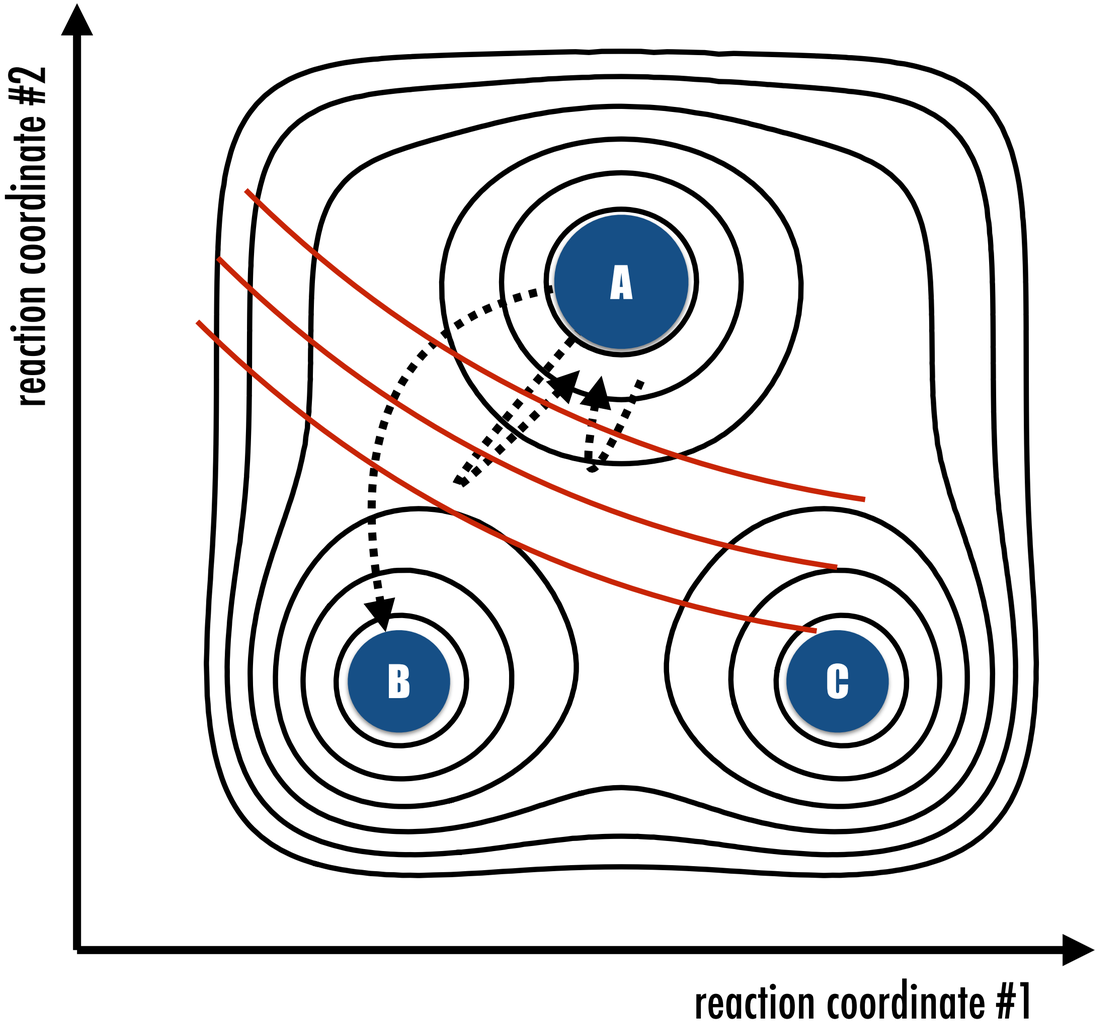
We can hypothesize and test a simple physical mechanism of resistance:
Resistance mutations reduce inhibitor binding affinity but retain ATP affinity (a surrogate for activity)

* Are certain inhibitors or binding modes less likely to elicit resistance?
* Can we incorporate likelihood of eliciting resistance mutations into rational ligand design?

BINDING KINETICS

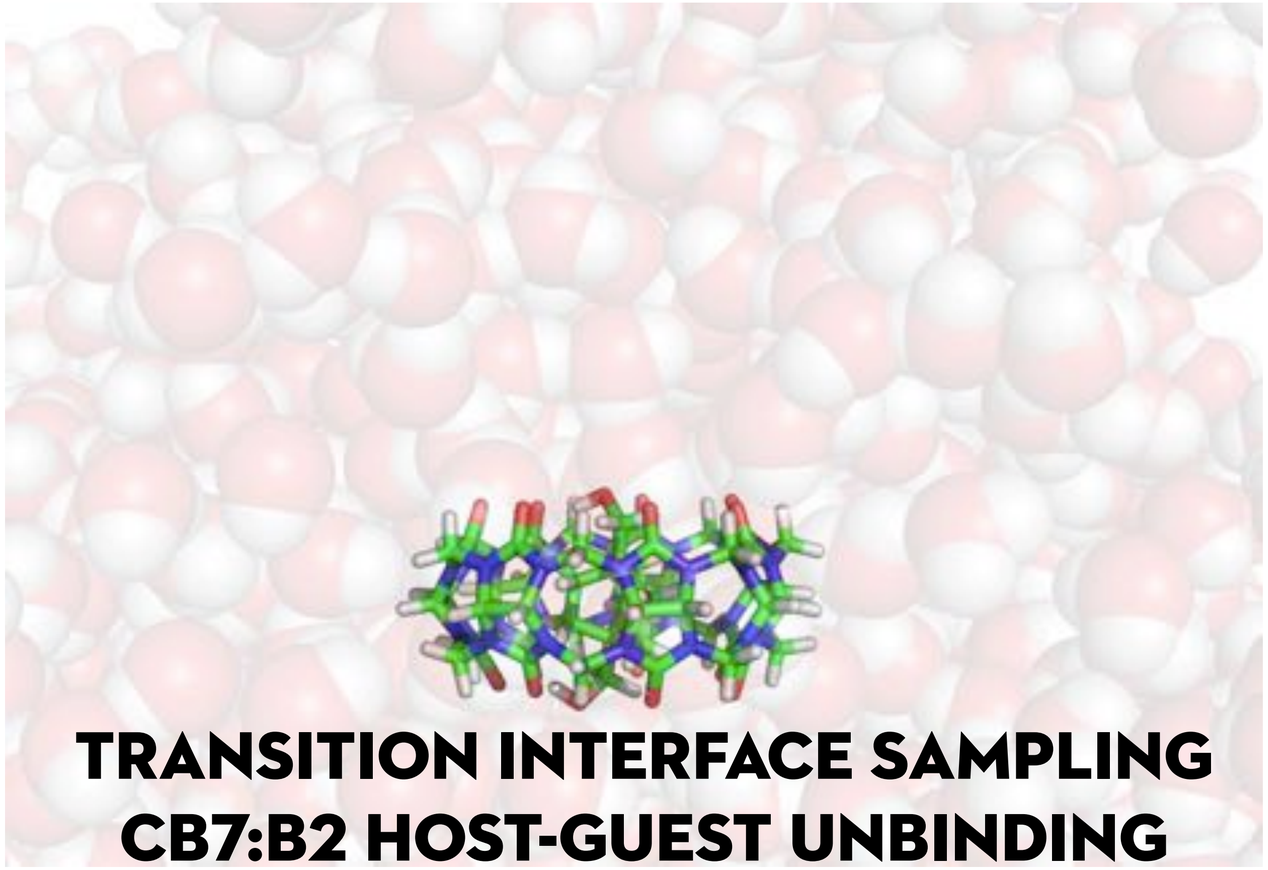


TRANSITION INTERFACE SAMPLING
CB7:B2 HOST-GUEST UNBINDING

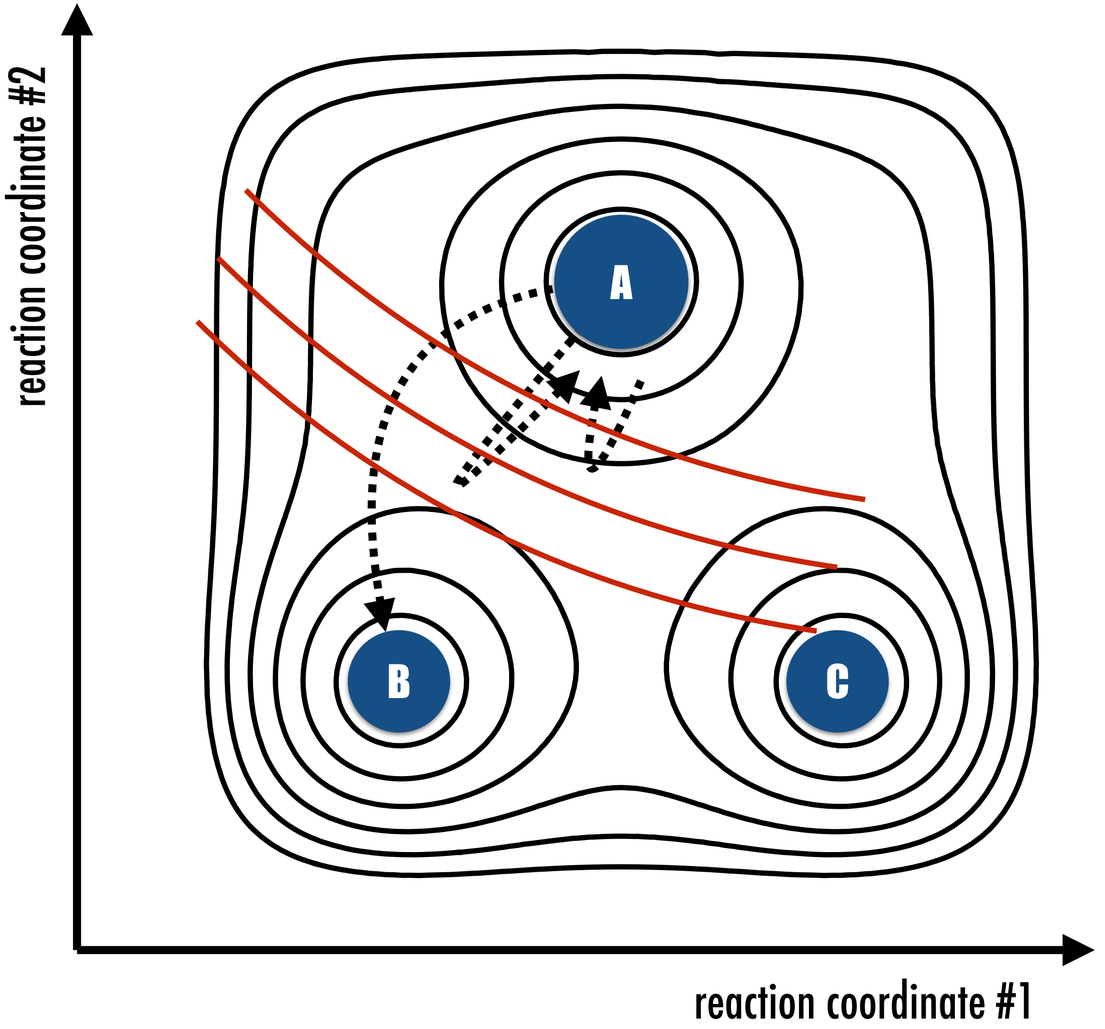


http://openpathsampling.org

BINDING KINETICS



TRANSITION INTERFACE SAMPLING
CB7:B2 HOST-GUEST UNBINDING



http://openpathsampling.org

THE CHODERA LAB @ MSKCC

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

http://folding.stanford.edu

COLLABORATORS

Stanford
Vijay Pande
Sergio Bacallado
NIH SimBios

IBM Almaden
Bill Swope
Jed Pitera
Julia Rice

University of Chicago
Nina Singhal Hinrichs

UC Irvine
David Mobley

CCNY
Marilyn Gunner

OpenEye
Christopher Bayly
Anthony Nicholls

Stony Brook
Ken Dill
Markus Seeliger

UCSF
Brian Shoichet
David Sivak

University of Virginia
Michael Shirts

Duke
David Minh

Freie Universität Berlin
Frank Noé
Bettina Keller
Jan-Hendrik Prinz

Rutgers
Zhiqiang Tan

University of Edinburg
Antonia S. J. S. Mey

UC Berkeley
Phillip Elms (BioRad)
Susan Marqusee
Carlos Bustamante
Christian Kaiser
Gheorghe Christol

University of Chicago
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