Slides and video licensed under CC-BY 4.0

slides: http://choderalab.org/news

ML/MM REPEX/ATM FEP/MBAR RBFES AND YOU

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

DISCLOSURES:
Scientific Advisory Board, OpenEye Scientific, Redesign Science*, Interline Therapeutics*, Ventus Therapeutics
All funding sources: http://choderalab.org/funding
* Denotes equity interests

7 Mar 2024 - OpenEye CUP XXIII - Santa Fe

YOU MIGHT (JUSTIFIABLY) BE WONDERING...

YOU MIGHT (JUSTIFIABLY) BE WONDERING...

1. WTF?

YOU MIGHT (JUSTIFIABLY) BE WONDERING...

1. WTF?

2. How on earth did we get here?

YOU MIGHT (JUSTIFIABLY) BE WONDERING...

1. WTF?

2. How on earth did we get here?

3. Why is this person keeping me from margaritas?

YOU MIGHT (JUSTIFIABLY) BE WONDERING...

1. WTF?

2. How on earth did we get here?

3. Why is this person keeping me from margaritas?

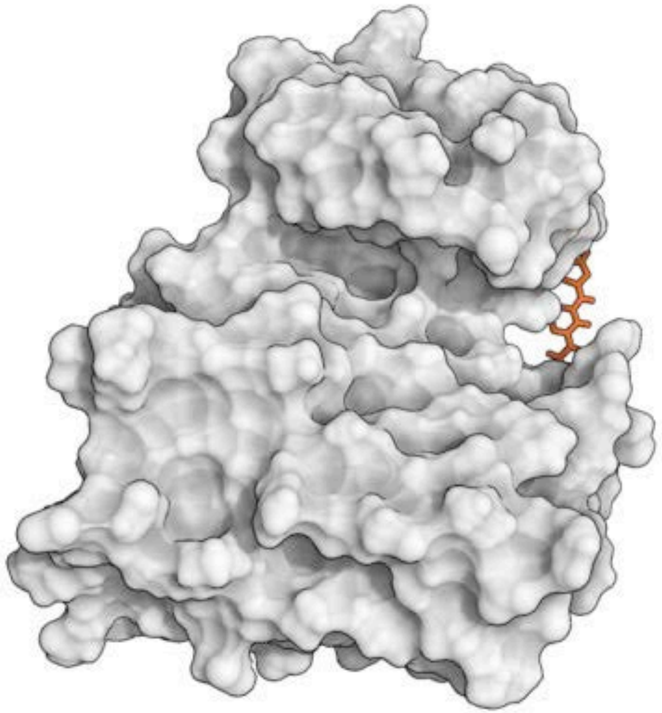
I will answer **at least one** of these questions in this talk.

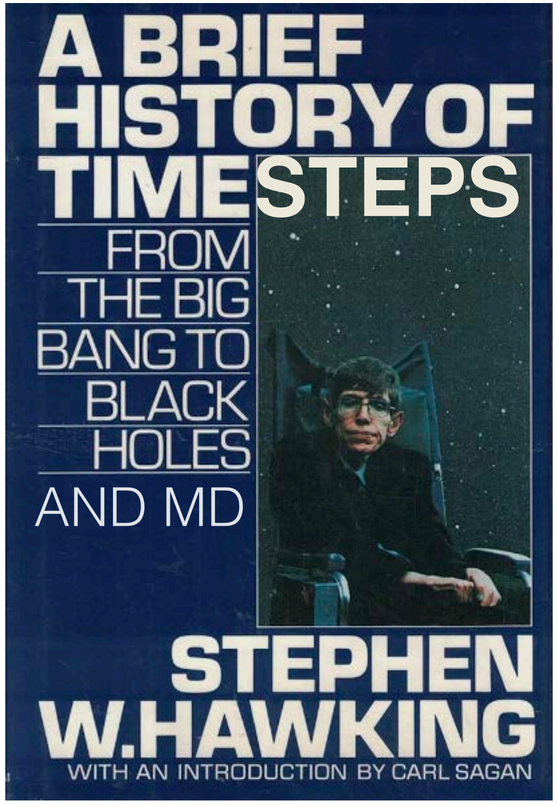
WHAT IS ML/MM REPEX/ATM FEP/MBAR RBFE AND WOULD ANYONE WANT TO USE THEM?

To understand this, we first need to review:

1. **How we got here**
2. **Where we are now**
3. **Where we might be headed**

A BRIEF HISTORY OF TIME(STEPS)





discrete timestep Langevin integrator

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

$$r_t = r_t^* + \Delta t\, v_t'$$

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

Where do we get the **forces**?

Shan, Kim, Eastwood, Dror, Seeliger, Shaw. JACS 133:9181, 2011
Durrant, McCammon. Molecular dynamics simulations and drug discovery. BMC Biology, 2011

MOLECULAR MECHANICS FORCE FIELDS WERE DEVELOPED FOR THINGS CALLED “MINICOMPUTERS”

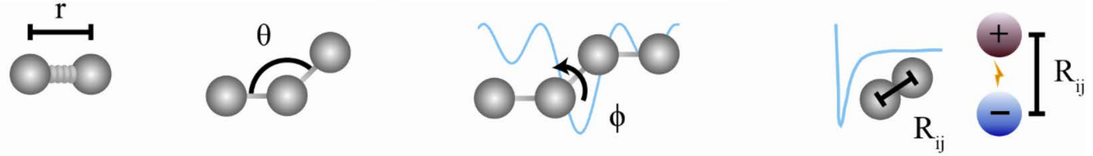
DEC PDP-11

~45 years old

typical class I molecular mechanics force field (ca. 1986 - 2024)

Bonded Non-bonded

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



shitty Taylor series
truncated at lowest order

crappy Fourier series
truncated at n=6

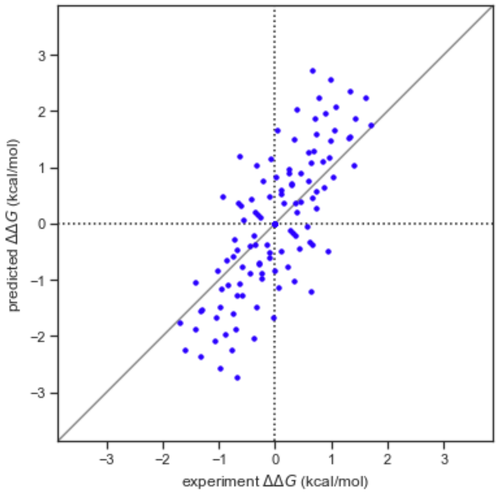
don’t even get me
started on this fucker

Durrant, McCammon. Molecular dynamics simulations and drug discovery. BMC Biology, 2011

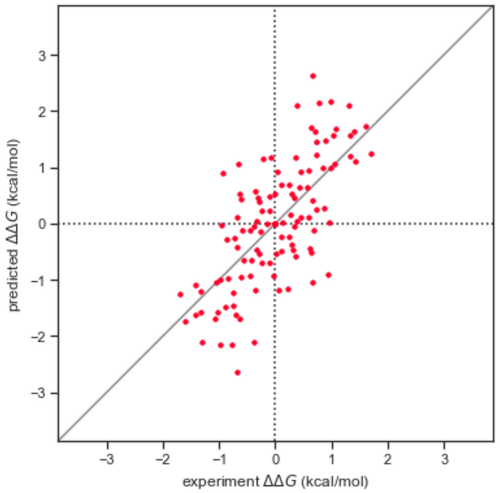
WE'VE MADE SIGNIFICANT PROGRESS IN PARAMETERS SINCE 1986, BUT WE'VE STILL BEEN STUCK WITH THE SAME FUNCTIONAL FORM

Open Force Field Initiative →

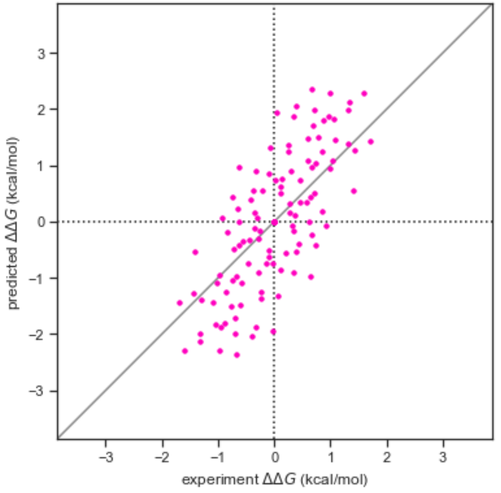
GAFF 1 (1999)



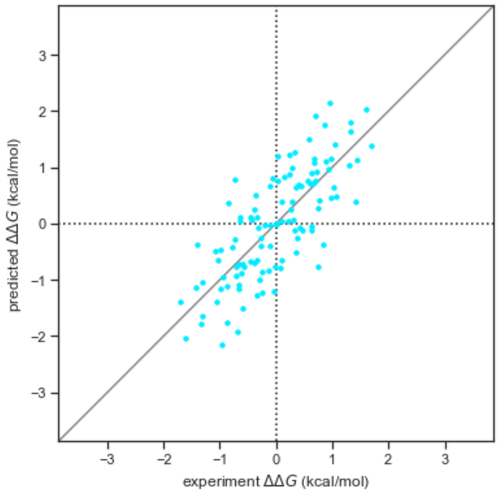
OPLS2.1 (2015)



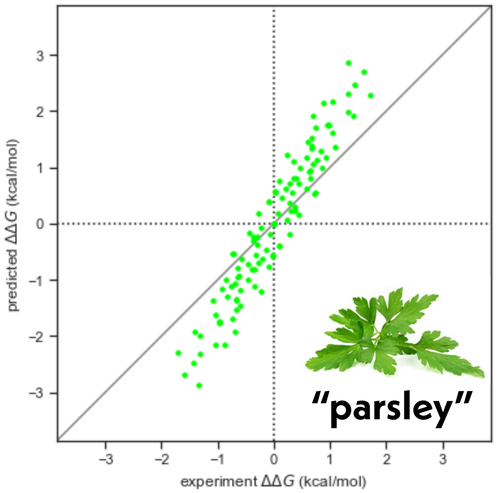
GAFF 2 (2016)



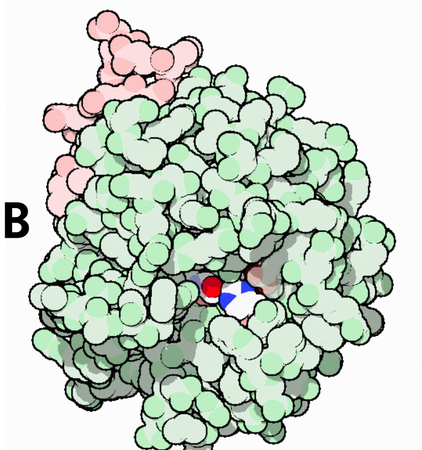
smirnoff99Frosst (2018)



openff 1.0 (2019)



thrombin
PDB101: 1PPB



open forcefield initiative

open forcefield

An open and collaborative approach to better force fields

OPEN SOURCE

Software permissively licensed under the MIT License and developed openly on GitHub.

OPEN SCIENCE

Scientific reports as blog posts, webinars and preprints

OPEN DATA

Curated quantum chemical and experimental datasets used to parameterize and benchmark Open Force Fields.

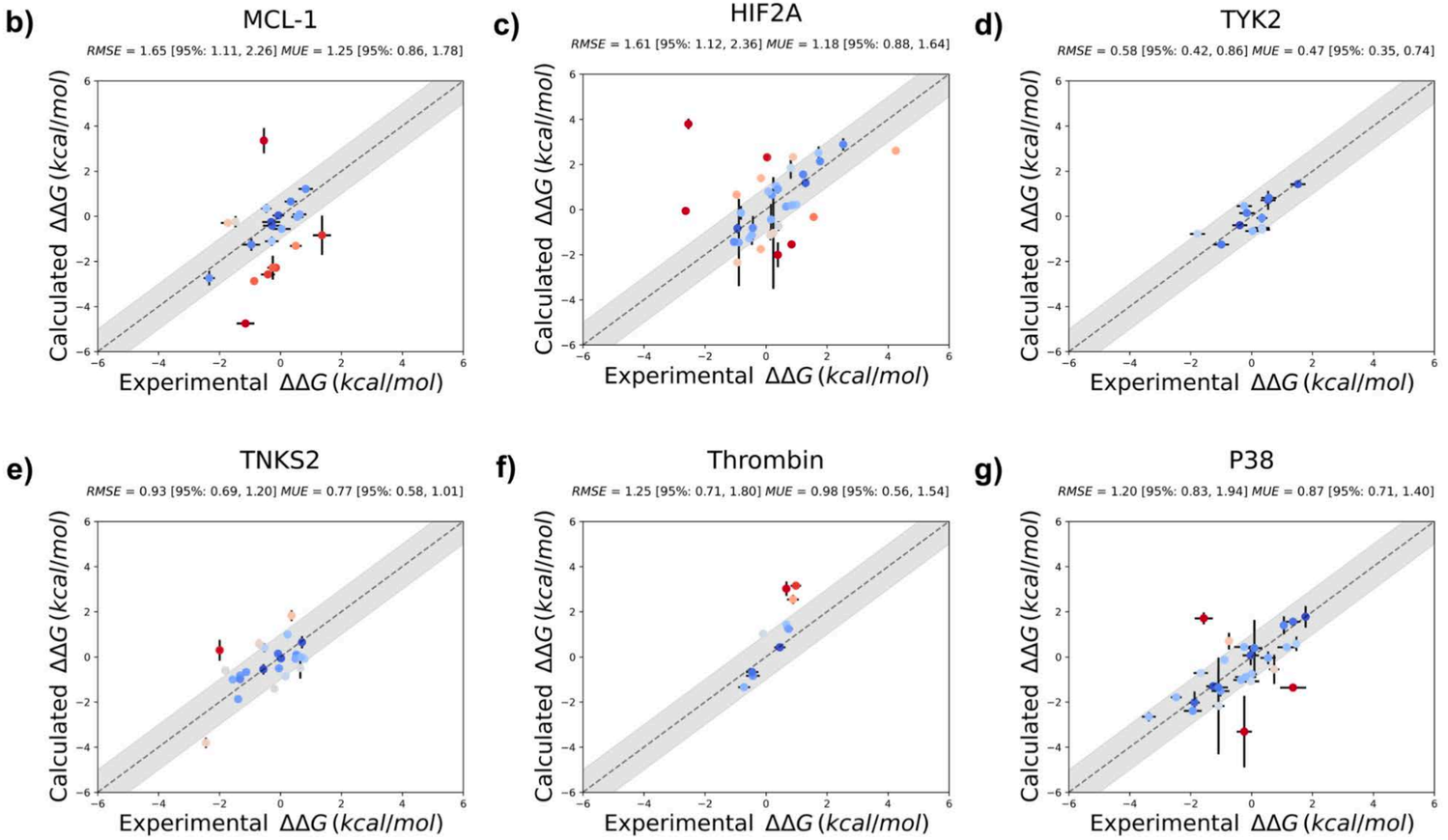
NEWS | TUTORIALS | ROADMAP

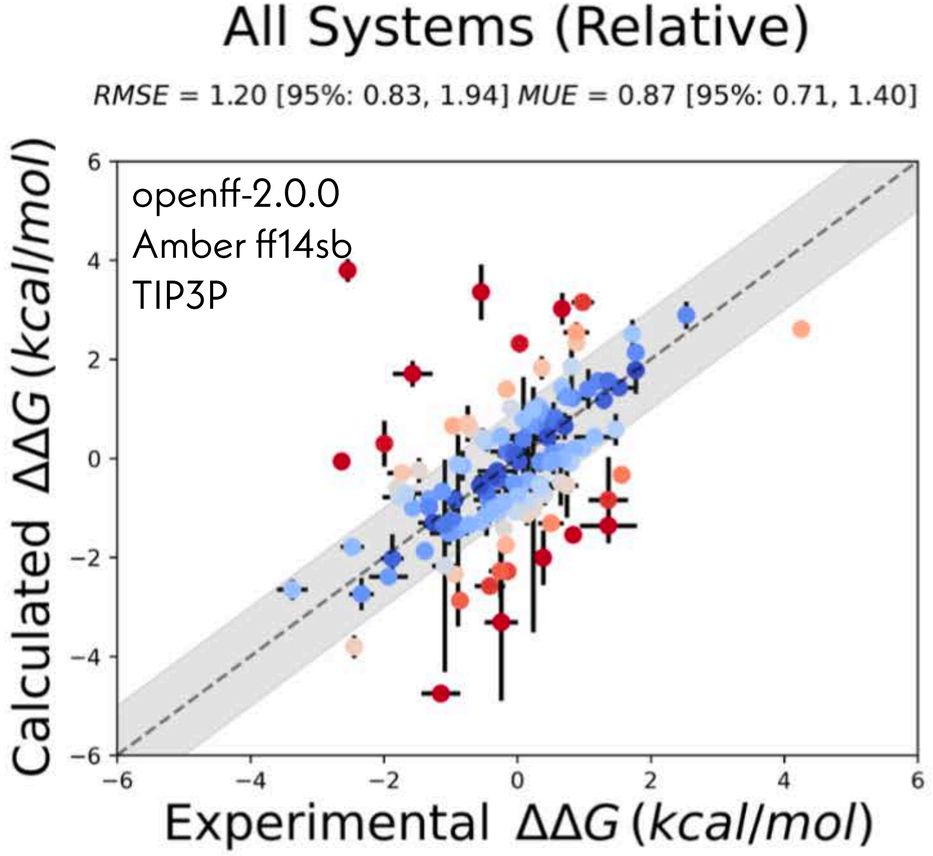
http://openforcefield.org

MM FORCE FIELDS WORK OK.
BUT WE WOULD LOVE TO DO BETTER.

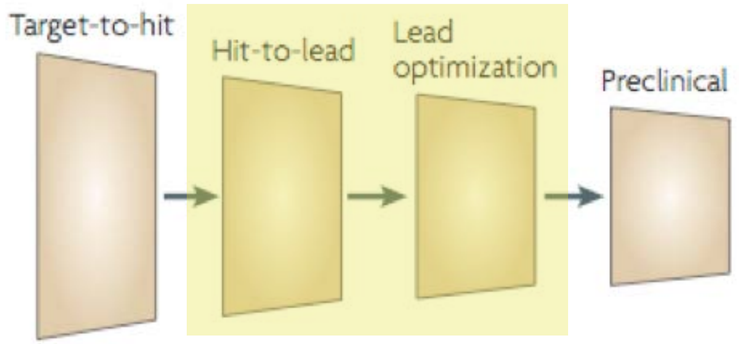
Open Free Energy Consortium Annual Report 2022

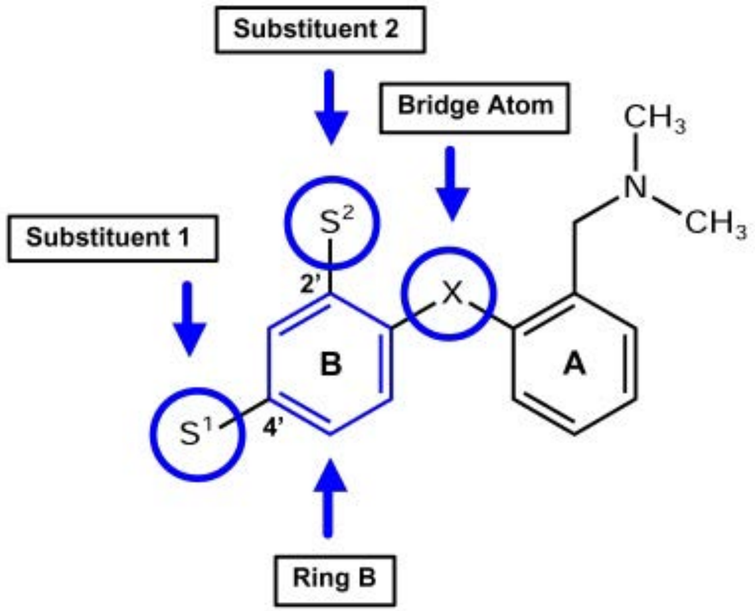
http://openfree.energy

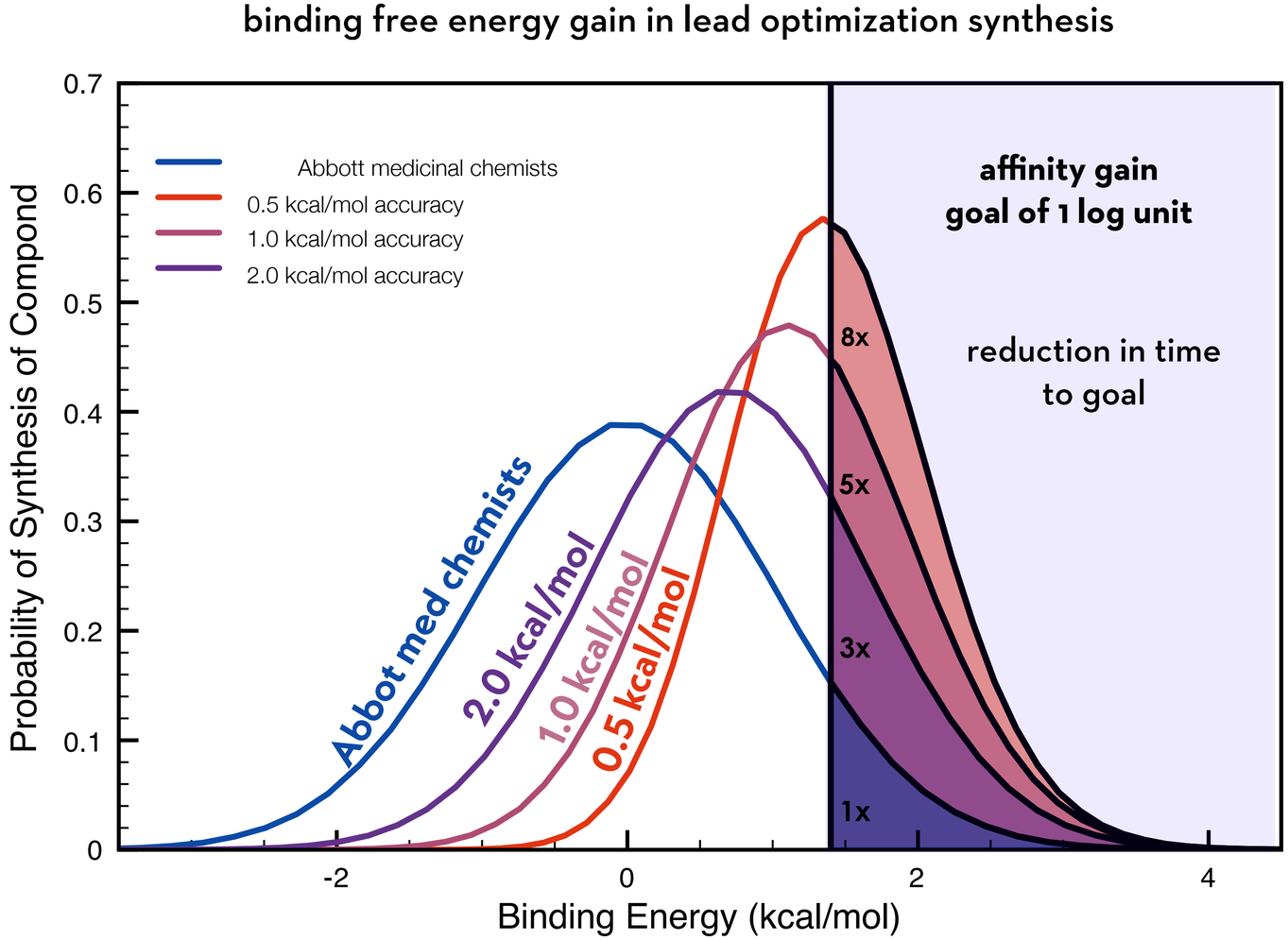




MUCH GREATER IMPACT IS POSSIBLE IF WE COULD REDUCE OUR PREDICTIVE MODEL ERRORS







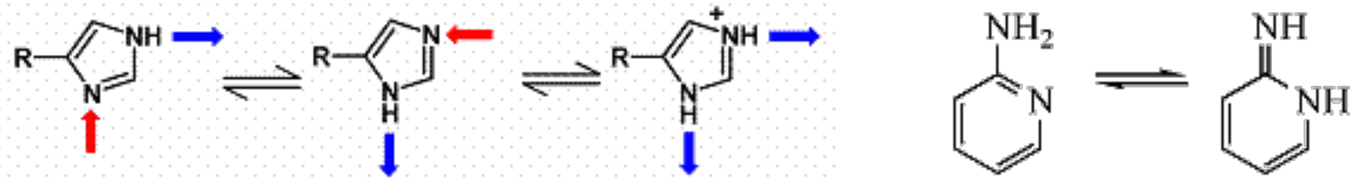
M. R. Shirts, D. L. Mobley and Scott P. Brown. "Free energy calculations in structure-based drug design", in Drug Design: Structure- and Ligand-Based Approaches, pgs. 61-86, 2010.

WHAT IS HOLDING FREE ENERGY CALCULATIONS BACK?

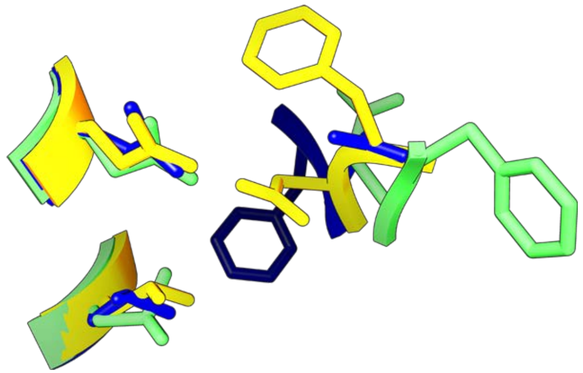
1. The forcefield may do a poor job of modeling the physics of our system (because it is constrained by choices made 40 years ago)

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

2. We're missing some essential chemical in our simulations because we don't bother to model them (e.g. protonation states, tautomers, redox chemistry, PTMs, etc.))



3. We haven't sampled all of the relevant conformations because we can't simulate for long enough

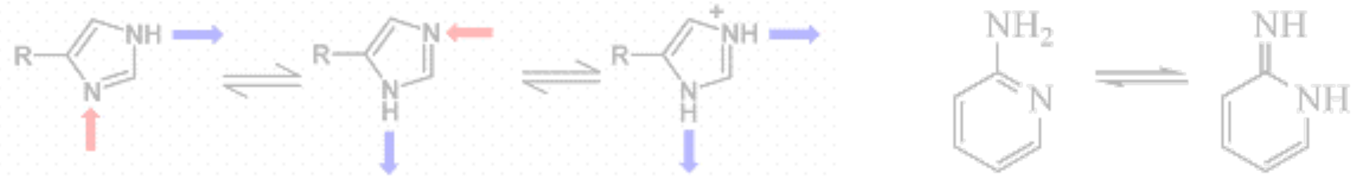


WHAT IS HOLDING BACK FREE ENERGY CALCULATIONS?

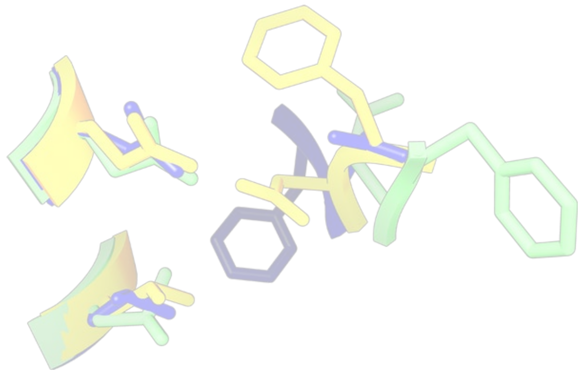
1. **The forcefield may do a poor job of modeling the physics of our system (because it is constrained by choices made 40 years ago)**

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

2. We're missing some essential chemical in our simulations because we don't bother to model them (e.g. protonation states, tautomers, redox chemistry, PTMs, etc.))



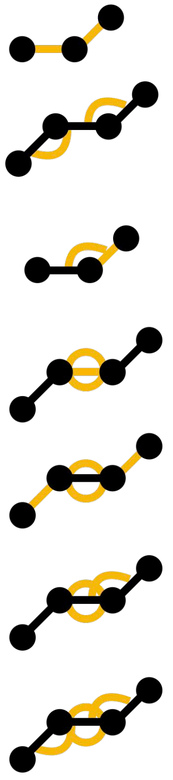
3. We haven't sampled all of the relevant conformations because we can't simulate for long enough



WE COULD GO TO CLASS II FORCE FIELDS…
BUT THE NUMBER OF TERMS EXPLODES COMBINATORIALLY

$$
\begin{aligned}
E = & \sum_b [{}^2K_b(b-b_0)^2 + {}^3K_b(b-b_0)^3 + {}^4K_b(b-b_0)^4] \\
& + \sum_\theta [{}^2K_\theta(\theta-\theta_0)^2 + {}^3K_\theta(\theta-\theta_0)^3 + {}^4K_\theta(\theta-\theta_0)^4] \\
& + \sum_\phi [{}^1K_\phi(1-\cos\phi) + {}^2K_\phi(1-\cos 2\phi) + {}^3K_\phi(1-\cos 3\phi)] \\
& + \sum_\chi K_\chi \chi^2 + \sum_{i>j} \frac{q_i q_j}{r_{ij}} + \sum_{i>j} \epsilon \left[2\left(\frac{r^*}{r_{ij}}\right)^9 - 3\left(\frac{r^*}{r_{ij}}\right)^6 \right] \\
& + \sum_b \sum_{b'} K_{bb'}(b-b_0)(b'-b'_0) + \sum_\theta \sum_{\theta'} K_{\theta\theta'}(\theta-\theta_0) \times \\
& (\theta'-\theta'_0) \\
& + \sum_b \sum_\theta K_{b\theta}(b-b_0)(\theta-\theta_0) \\
& + \sum_\phi \sum_b (b-b_0)[{}^1K_{\phi b}\cos\phi + {}^2K_{\phi b}\cos 2\phi + {}^3K_{\phi b}\cos 3\phi] \\
& + \sum_\phi \sum_{b'} (b'-b'_0)[{}^1K_{\phi b'}\cos\phi + {}^2K_{\phi b'}\cos 2\phi + \\
& {}^3K_{\phi b'}\cos 3\phi] \\
& + \sum_\phi \sum_\theta (\theta-\theta_0)[{}^1K_{\phi\theta}\cos\phi + {}^2K_{\phi\theta}\cos 2\phi + {}^3K_{\phi\theta}\cos 3\phi] \\
& + \sum_\phi \sum_\theta \sum_{\theta'} K_{\phi\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0)\cos\phi \qquad (1)
\end{aligned}
$$

- bond-bond: angle node
- angle-angle: torsion node
- bond-angle: angle node
- torsion-(center) bond: torsion
- torsion-(side) bond: torsion
- torsion-angle: torsion
- torsion-angle-angle: torsion



Can we do a better job of modeling true many-body local valence terms, and set ourselves up to solve the other challenges too?

Hwang et al. (1994) http://doi.org/10.1021/ja00085a036

WE HAVE REAL COMPUTERS NOW

$1599 MSRP*

Why not put them to work?

* A new PDP-11 in ~1975 would cost $160,000 in today's dollars

A NEW GENERATION OF MACHINE LEARNING POTENTIALS PROVIDE MUCH MORE FLEXIBILITY IN FUNCTIONAL FORM (AT HIGHER COST)

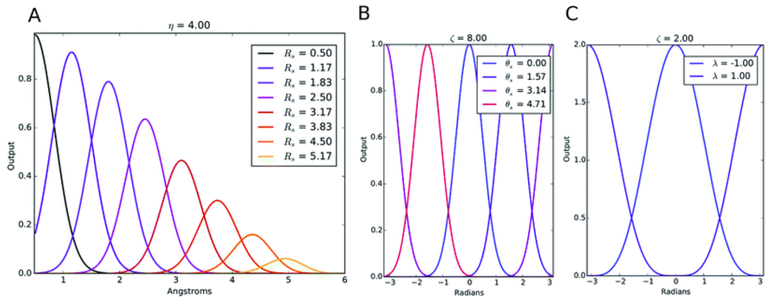
ANI family of quantum machine learning potentials

radial and **angular** features

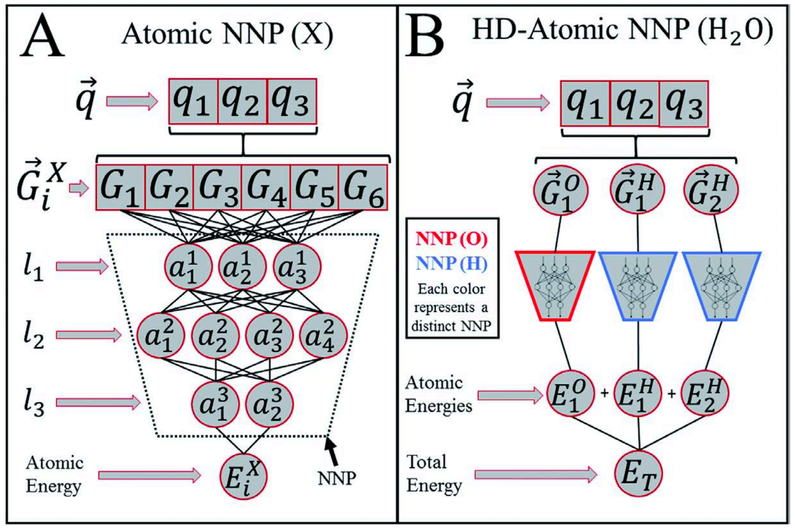
$$f_C(R_{ij}) = \begin{cases} 0.5\times\cos\left(\frac{\pi R_{ij}}{R_C}\right)+0.5 & \text{for } R_{ij} \le R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases}$$

$$G_m^R = \sum^{\text{all atoms}} e^{-\eta\left(R_{ij}-R_s\right)^2} f_C\left(R_{ij}\right)$$

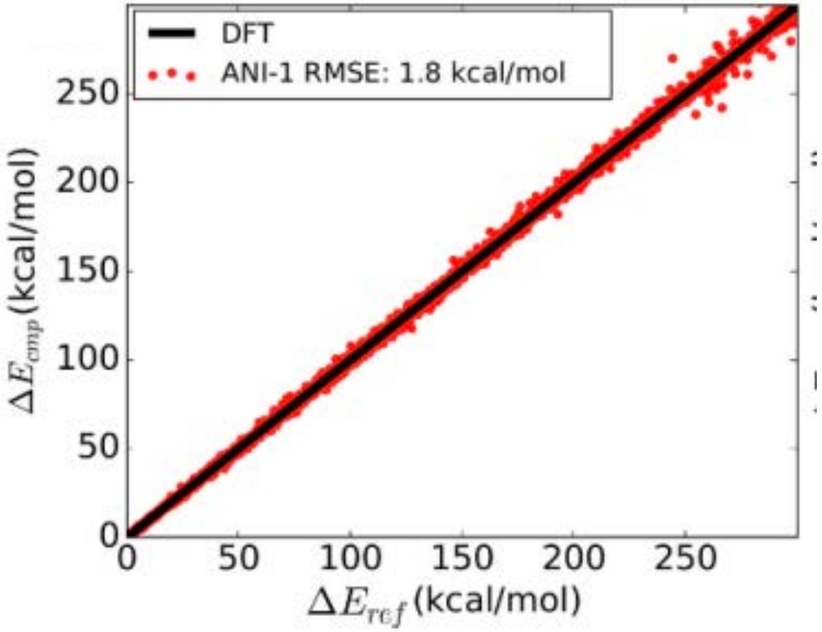
$$G_m^{A_{mod}} = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all atoms}} \left(1+\cos\left(\theta_{ijk}-\theta_s\right)\right)^{\zeta} \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_s\right)^2\right] f_C\left(R_{ij}\right) f_C\left(R_{ik}\right)$$



deep neural network for each atom



excellent agreement with DFT



Can train an ANI model in **~1 day**

OLEXANDR ISAYEV **ADRIAN ROITBERG**

Smith, Isayev, Roitberg. Chemical Science 8:3192, 2017.
http://doi.org/10.1039/c6sc05720a

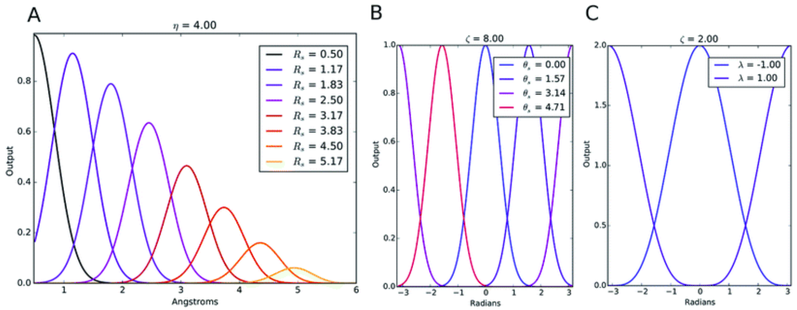
ML POTENTIALS ARE SEEING RAPID EVOLUTION IN ARCHITECTURES THAT ENCODE PHYSICAL INVARIANCES

ANI
distance and angle

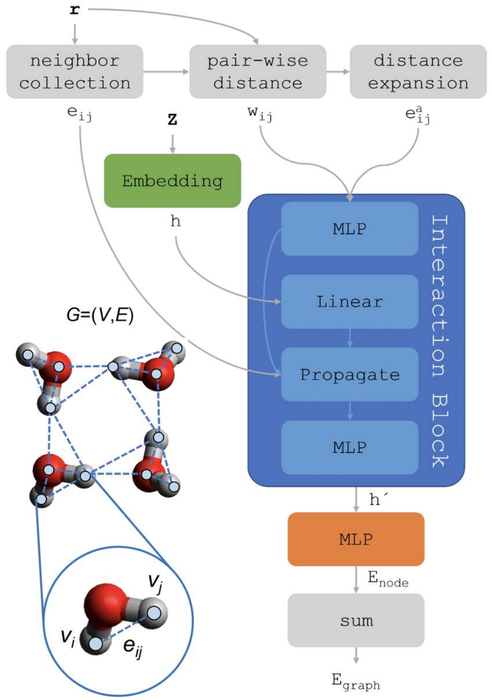
$$f_C(R_{ij}) = \begin{cases} 0.5\times\cos\left(\frac{\pi R_{ij}}{R_C}\right)+0.5 & \text{for } R_{ij} \le R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases}$$

$$G_m^R = \sum^{\text{all atoms}} e^{-\eta\left(R_{ij}-R_s\right)^2} f_C(R_{ij})$$

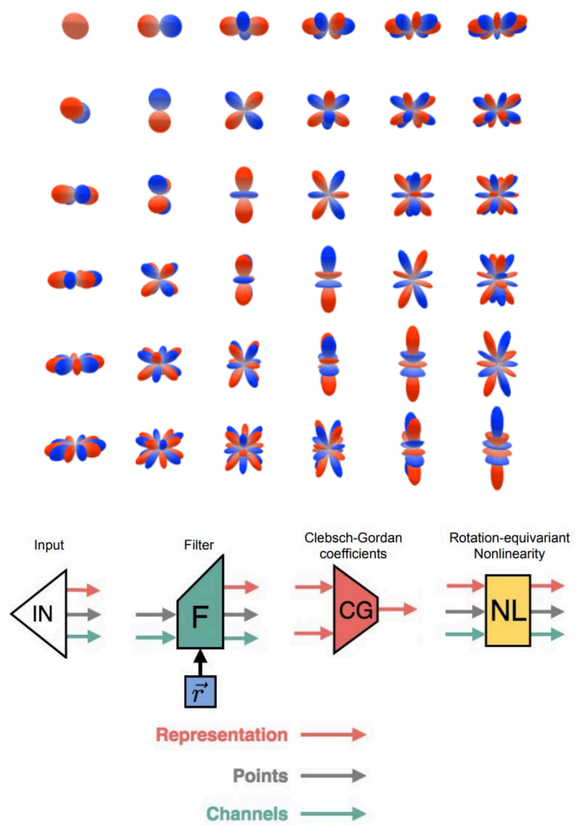
$$G_m^{A_{mod}} = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all atoms}} \left(1+\cos\left(\theta_{ijk}-\theta_s\right)\right)^{\zeta} \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_s\right)^2\right] f_C(R_{ij}) f_C(R_{ik})$$



SchNet
E(3) invariant



Tensor Field Networks
E(3) equivariant



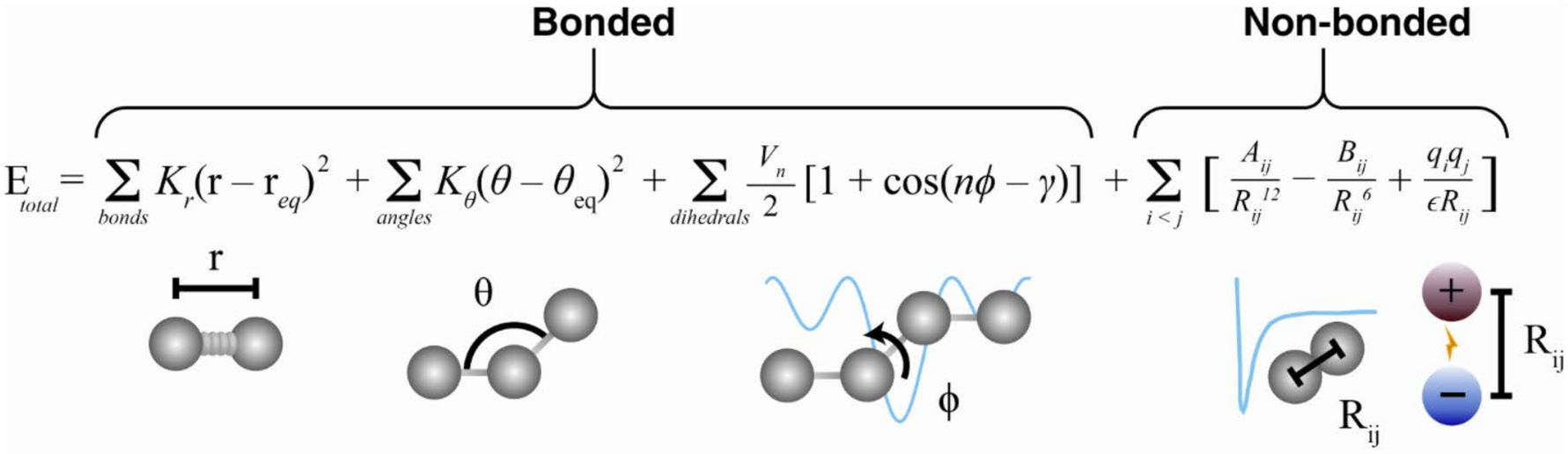
The **ANI** class of models uses distance- and angle-based features [http://doi.org/10.1039/c6sc05720a].
SchNet uses distance-based features for continuous convolutions [https://doi.org/10.1038/ncomms13890].
Tensor Field Networks and Clebsch-Gordon nets use spherical harmonics [https://arxiv.org/abs/1802.08219; https://bit.ly/2SRVS67].

ML/MM REPEX/ATM FEP/MBAR RBFE

OK, so what does this mean?

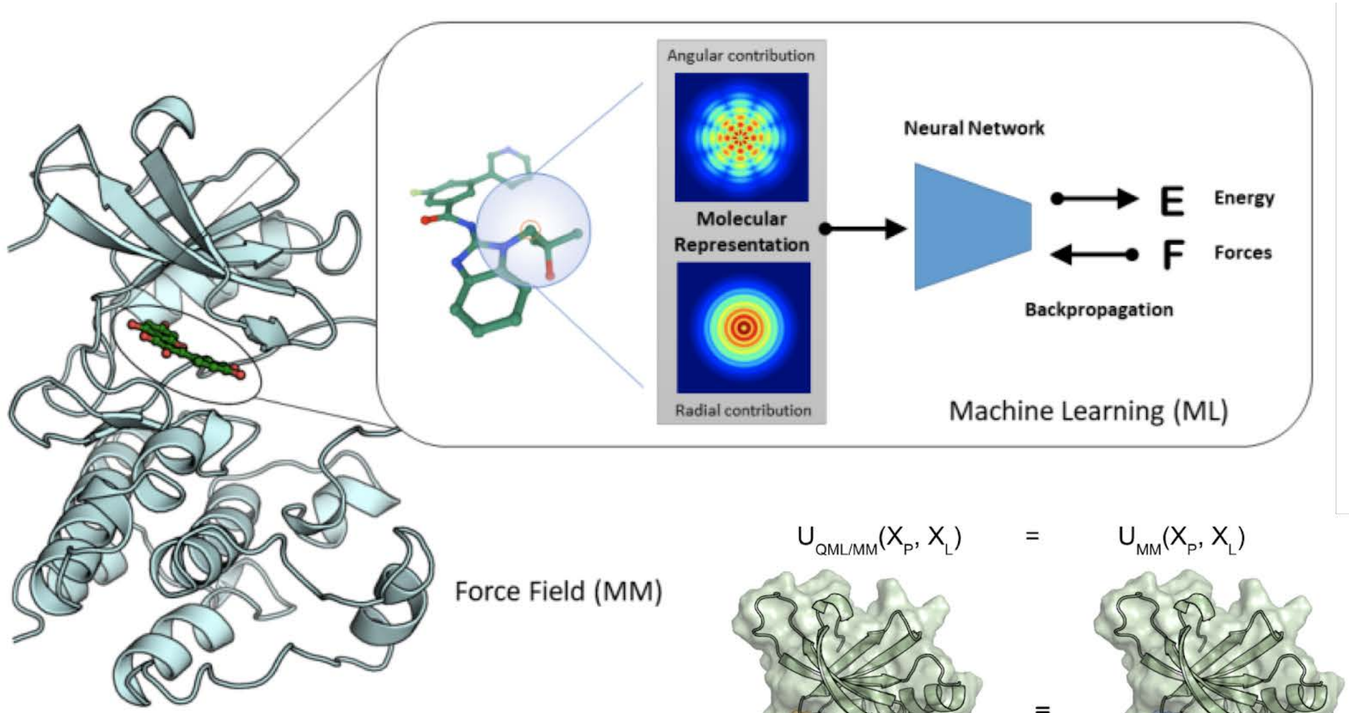
ML/MM REPEX/ATM FEP/MBAR RBFE

molecular mechanics force field

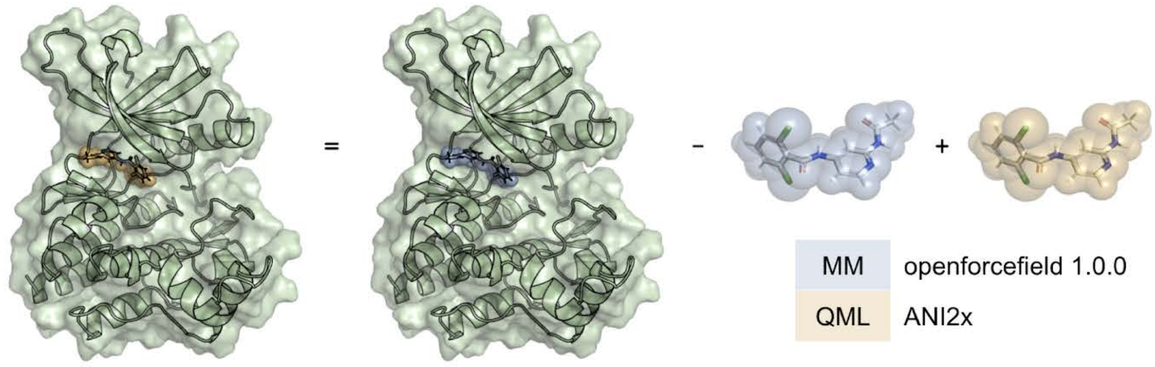


ML/MM REPEX/ATM FEP/MBAR RBFE

hybrid machine learning / molecular mechanics force field



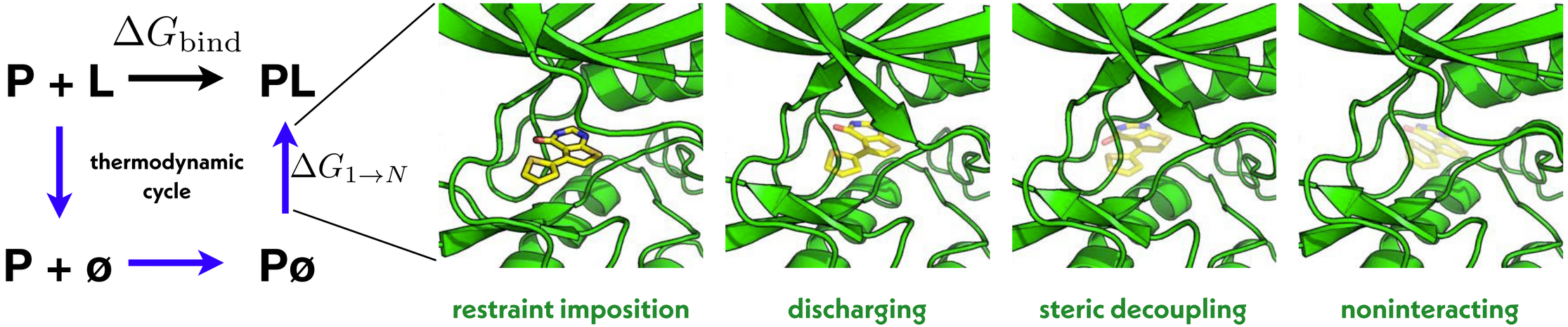
$U_{QML/MM}(X_P, X_L) = U_{MM}(X_P, X_L) - U_{MM}^{vacuum}(X_L) + U_{QML}^{vacuum}(X_L)$



MM	openforcefield 1.0.0
QML	ANI2x

ML/MM REPEX/ATM FEP/MBAR RBFE

Free energy perturbation uses alchemical intermediates to compute binding free energies



Includes all contributions from enthalpy and entropy of binding to a flexible receptor

$$\Delta G_{0\to 1} = -k_B T \ln \frac{Z_1}{Z_0} = -k_B T \ln \frac{Z_{\lambda_2}}{Z_{\lambda_1}} \frac{Z_{\lambda_3}}{Z_{\lambda_2}} \cdots \frac{Z_{\lambda_N}}{Z_{\lambda_{N-1}}}$$

$$Z_n = \int dx\, e^{-\beta U_n(x)}$$ partition function

From the lab of Emilio Gallicchio (Brooklyn College, CUNY)

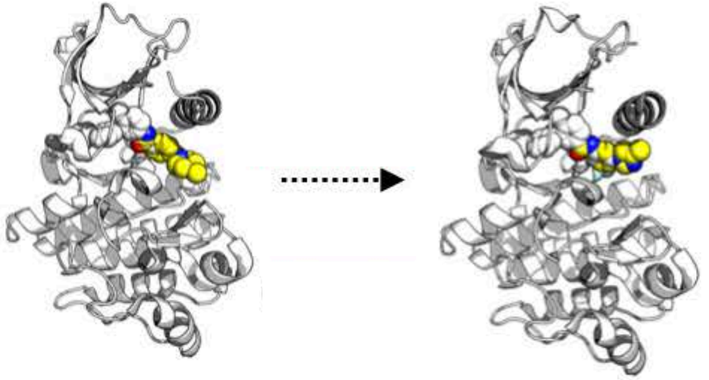
JCIM 17:3309, 2021, JCIM 62:309, 2022, JCTC 18:6438, 2022, JCTC 19:4780, 2023, JCIM 64:8501, 2024

Roitberg, Fujitani, Pande, Shirts, Swope, Christ, Mobley, Schrödinger, and many more

ALCHEMICAL FREE ENERGY CALCULATIONS HAVE A BROAD DOMAIN OF APPLICABILITY IN DRUG DISCOVERY

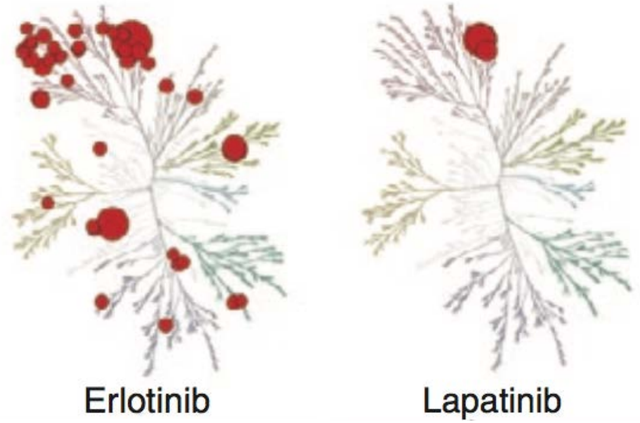
driving affinity / potency

Schindler, Baumann, Blum et al. JCIM 11:5457, 2020
https://doi.org/10.1021/acs.jcim.0c00900



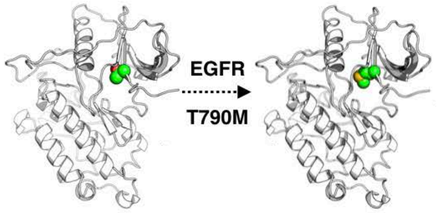
driving selectivity

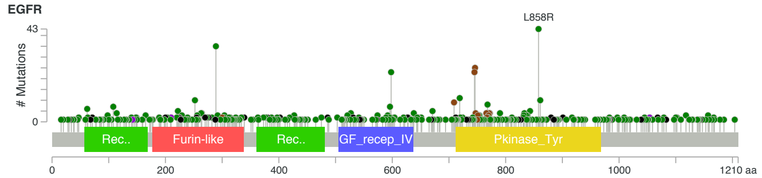
Moraca, Negri, de Olivera, Abel JCIM 2019
https://doi.org/10.1021/acs.jcim.9b00106
Aldeghi et al. JACS 139:946, 2017.
https://doi.org/10.1021/jacs.6b11467



predicting clinical drug resistance/sensitivity

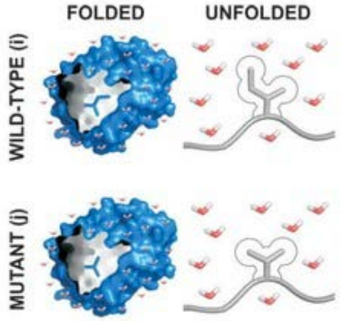
Hauser, Negron, Albanese, Ray, Steinbrecher, Abel, Chodera, Wang.
Communications Biology 1:70, 2018
https://doi.org/10.1038/s42003-018-0075-x
Aldeghi, Gapsys, de Groot. ACS Central Science 4:1708, 2018
https://doi.org/10.1021/acscentsci.8b00717

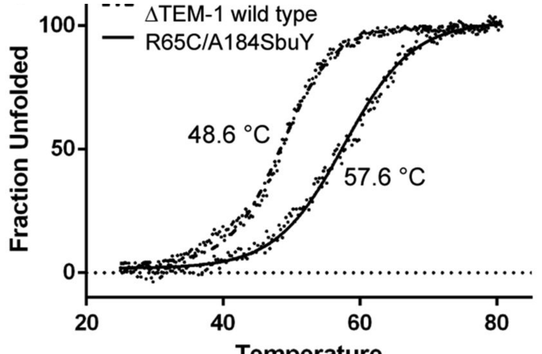




optimizing thermostability

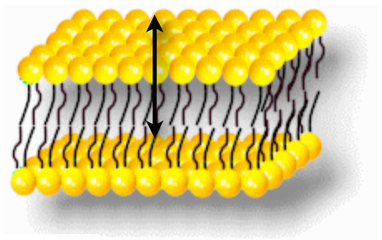
Gapsys, Michielssens, Seeliger, and de Groot. Angew Chem 55:7364, 2016
https://doi.org/10.1002/anie.201510054



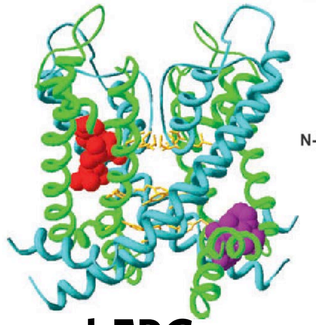


...AND HOLD THE POTENTIAL FOR EVEN BROADER APPLICABILITY AS MORE STRUCTURAL DATA EMERGES

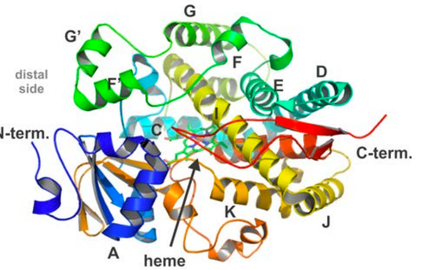
partition coefficients (logP, logD) and permeabilities



structure-enabled ADME/Tox targets

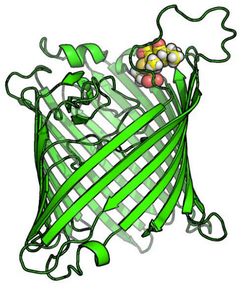


hERG

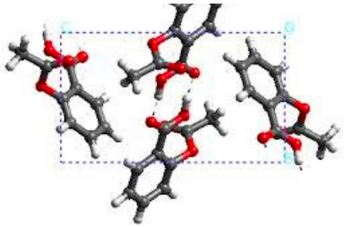


CYP3A4

porin permeation



crystal polymorphs, etc.



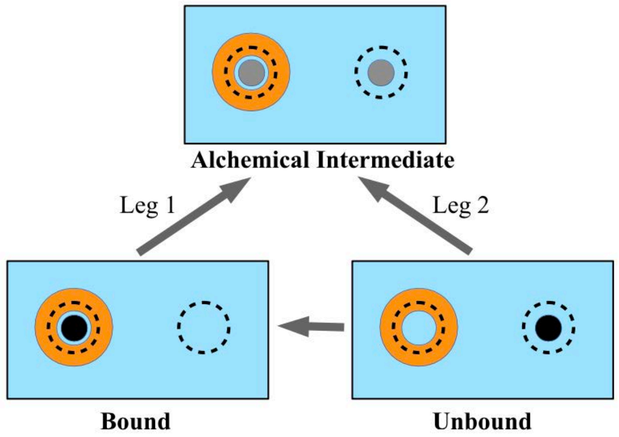
ML/MM REPEX/ATM FEP/MBAR RBFE

Alchemical Transfer Method (ATM) defines alchemical intermediates in a surprisingly simple way:

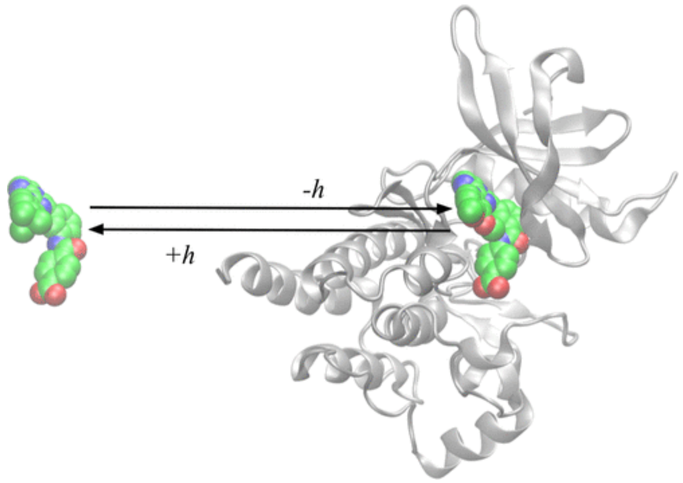
$$U(x;\lambda) = (1-\lambda)\,U_0(x) + \lambda\,U_0(x+\Delta x)$$

unmodified potential

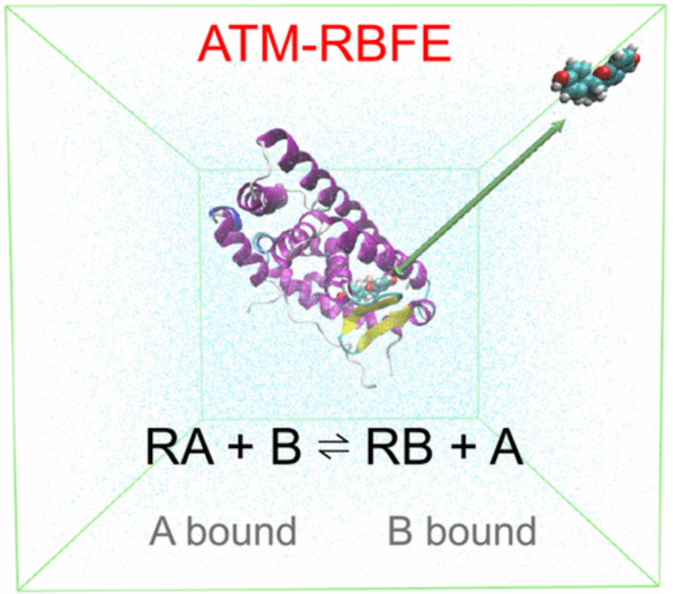
displaced ligand potential



absolute binding free energies



relative binding free energies



ATM works with molecular mechanics and machine learning force fields without any special changes!

From the lab of Emilio Gallicchio (Brooklyn College, CUNY)

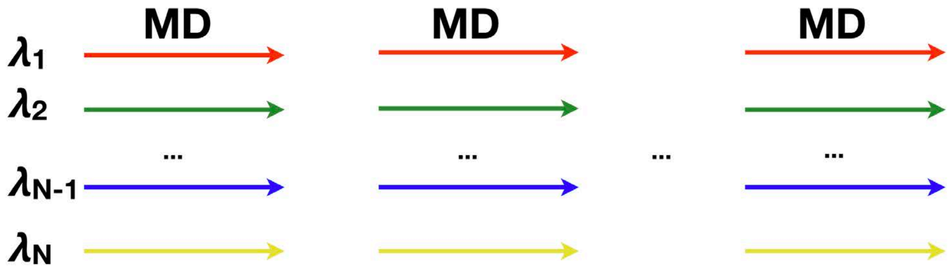
JCIM 17:3309, 2021 ; JCIM 62:309, 2022 ; JCIM 63:2438, 2023 ; JCIM 64:250, 2024

ML/MM REPEX/ATM FEP/MBAR RBFE

replica exchange sampling of multiple alchemical states

Independent simulations
Easy to parallelize, but sampling problems at any λ can make calculations unreliable

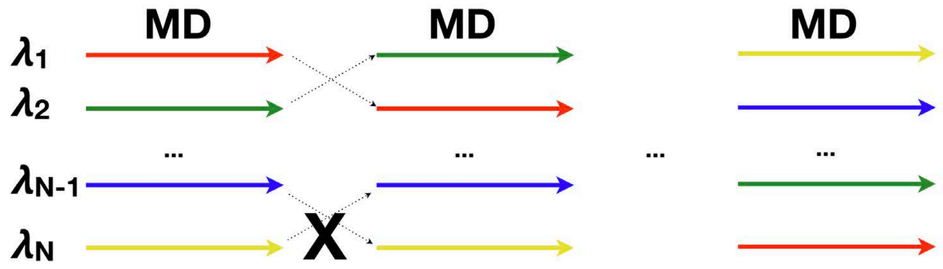
simple but dangerous due to poor sampling of conformational changes coupled to λ



AMBER18 TI
Song, Lee, Zhu, York, Merz 2019
https://doi.org/10.1021/acs.jcim.9b00105

Replica exchange (REPEX)
Good sampling at any λ can rescue problems at other λ if good λ overlap

reliable but communication heavy



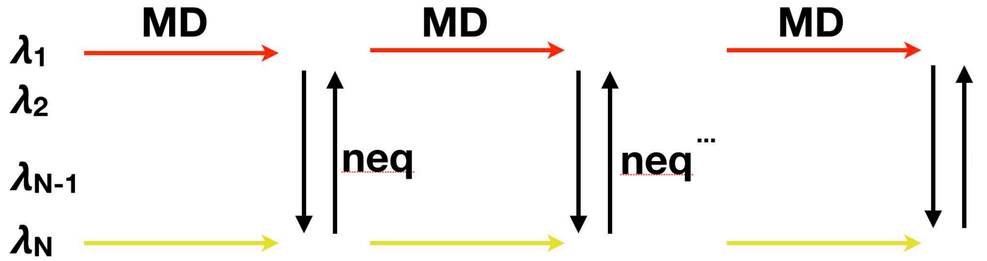
Schrödinger FEP+
Wang, Wu, Deng, Kim, ... Abel 2015
https://doi.org/10.1021/ja512751q

NAMD
Jiang, Thirman, Jo, Roux 2018
http://doi.org/10.1021/acs.jpcb.8b03277

OpenMM
Chodera, Shirts
https://doi.org/10.1063/1.3660669

Nonequilibrium methods
Less efficient than equilibrium calculations, but can work robustly and scalably if properly tuned

cloud- and wall clock friendly



pmx / gromacs
Aldeghi, Gapsys, de Groot 2018
https://doi.org/10.1021/acscentsci.8b00717

Orion NES!

ML/MM REPEX/ATM FEP/MBAR RBFE

Multistate Bennett Acceptance Ratio (MBAR) provides an optimal way to analyze data to estimate free energy differences

$$q_k(x) \equiv e^{-\beta[U_0(x)+U_k(x)]}$$

$U_0(x)$ Unperturbed potential

$U_k(x)$ perturbed potential

unnormalized probability distribution

We can use optimal bridge sampling estimator machinery (Z. Tan, Meng, Wong, others) to produce the multistate generalization of Bennett acceptance ratio (BAR) that provides efficient estimators for

free energy differences

$$\Delta f_{ij} \equiv f_j - f_i = -\ln \frac{\int d\boldsymbol{x}\, q_j(\boldsymbol{x})}{\int d\boldsymbol{x}\, q_i(\boldsymbol{x})}$$

exact

equilibrium expectations

$$\langle A \rangle \equiv \frac{\int d\boldsymbol{x}\, A(\boldsymbol{x})\, q(\boldsymbol{x})}{\int d\boldsymbol{x}\, q(\boldsymbol{x})}$$

estimators from data

$$\hat{f}_i = -\ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{q_i(\boldsymbol{x}_{jn})}{\sum_{k=1}^{K} N_k\, e^{\hat{f}_k}\, q_k(\boldsymbol{x}_{jn})}$$

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

$$\hat{A} = \sum_{n=1}^{N} W_{na}\, A(\boldsymbol{x}_n)$$

where n now runs from 1 to $N = \sum_{k=1}^{K} N_k$

$$W_{na} \propto \frac{q(\boldsymbol{x}_n)}{\sum_{k=1}^{K} N_k\, e^{\hat{f}_k} q_k(\boldsymbol{x}_n)}$$

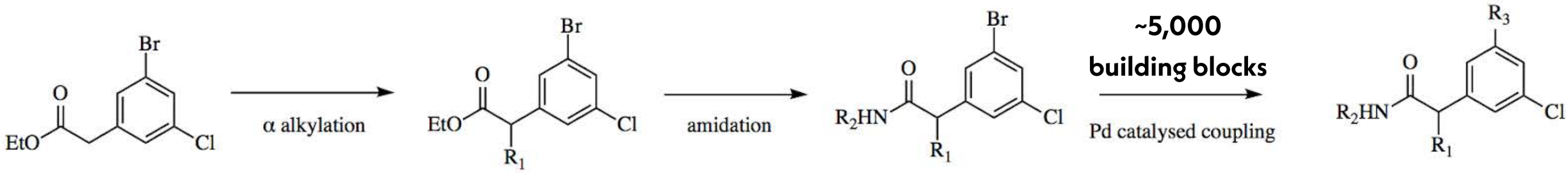
$$\delta^2 \hat{A} = \hat{A}^2 (\hat{\Theta}_{AA} + \hat{\Theta}_{aa} - 2\,\hat{\Theta}_{Aa})$$

MATH

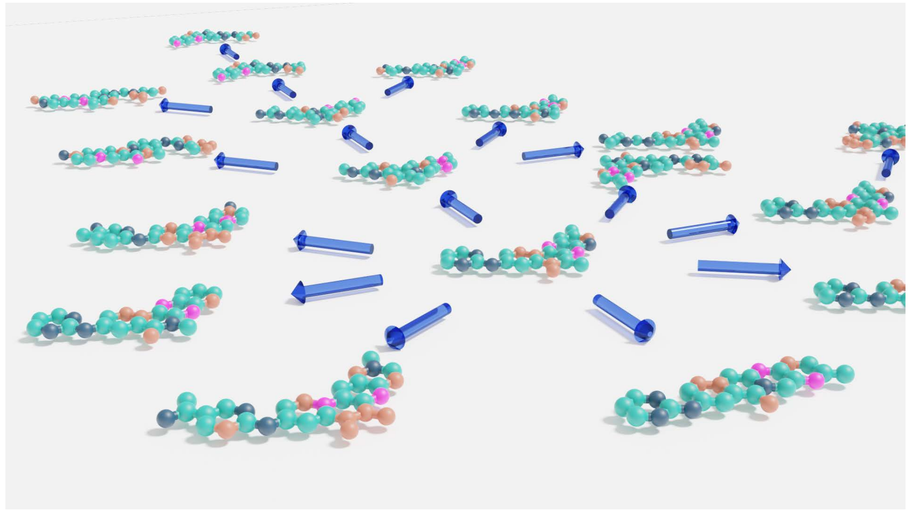
Shirts and Chodera. JCP 129:124105, 2008

ML/MM REPEX/ATM FEP/MBAR RBFE

Relative Binding Free Energy (RBFE) calculations are a useful way to make decisions about which synthetically tractable molecules to make

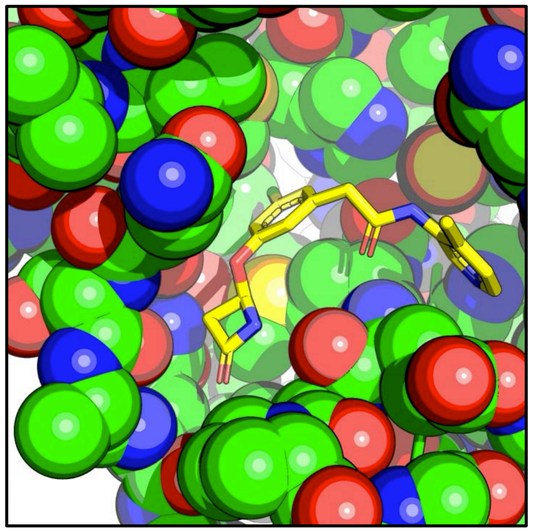


relative alchemical transformation network



Open Free Energy Consortium: https://openfree.energy/

docked poses



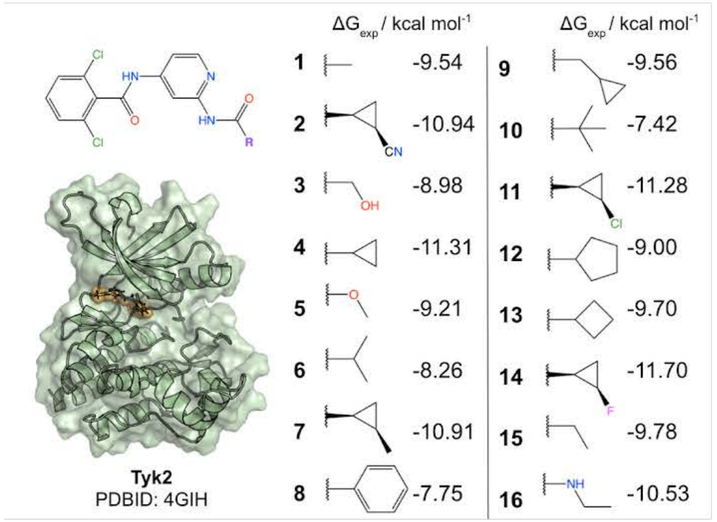
Best Practices for Alchemical Free Energy Calculations
https://doi.org/10.33011%2Flivecoms.2.1.18378

WE HAD PREVIOUSLY SEEN MM TO ML/MM CORRECTIONS HAD SHOWN SIGNIFICANT PROMISE…

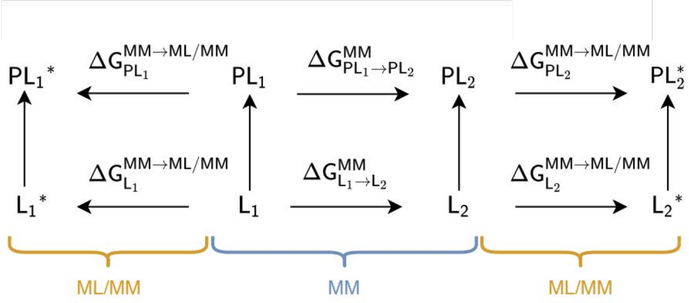
MM (OPLS2.1 + CM1A-BCC charges)
Missing torsions from LMP2/cc-pVTZ(-f) QM calculations
SPC water

MM (OpenFF 1.0.0 "Parsley")
AMBER14SB protein force field
TIP3P; Joung and Cheatham ions

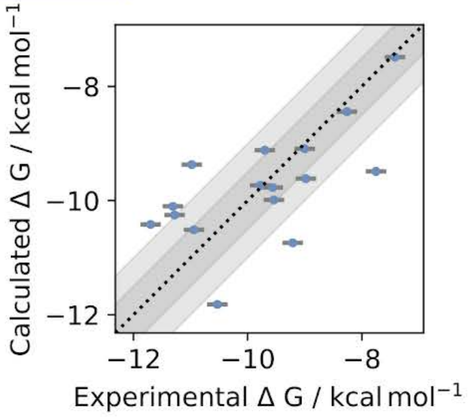
QML/MM (OpenFF 1.0.0 + ANI2x)
AMBER14SB protein force field
TIP3P; Joung and Cheatham ions



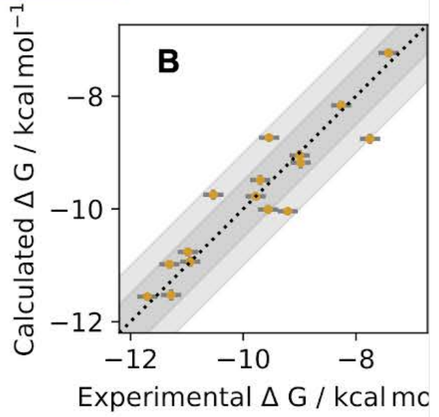
	Tyk2
no. of compds	16
binding affinity range (kcal/mol)	4.3
crystal structure	4GIH
series ref	52,53
no. of perturbations	24
MUE FEP	0.75 ± 0.11
RMSE FEP	0.93 ± 0.12




```
                    MM: openff-1.0.0 (N = 16)
RMSE:   0.97 [95%: 0.68, 1.22]
MUE:    0.77 [95%: 0.51, 1.08]
R2:     0.42 [95%: 0.08, 0.75]
rho:    0.65 [95%: 0.25, 0.88]
```




```
                    ML/MM: openff-1.0.0 with ANI2x (N = 16)
RMSE:   0.47 [95%: 0.32, 0.68]
MUE:    0.35 [95%: 0.24, 0.56]
R2:     0.86 [95%: 0.66, 0.95]
rho:    0.93 [95%: 0.79, 0.97]
```

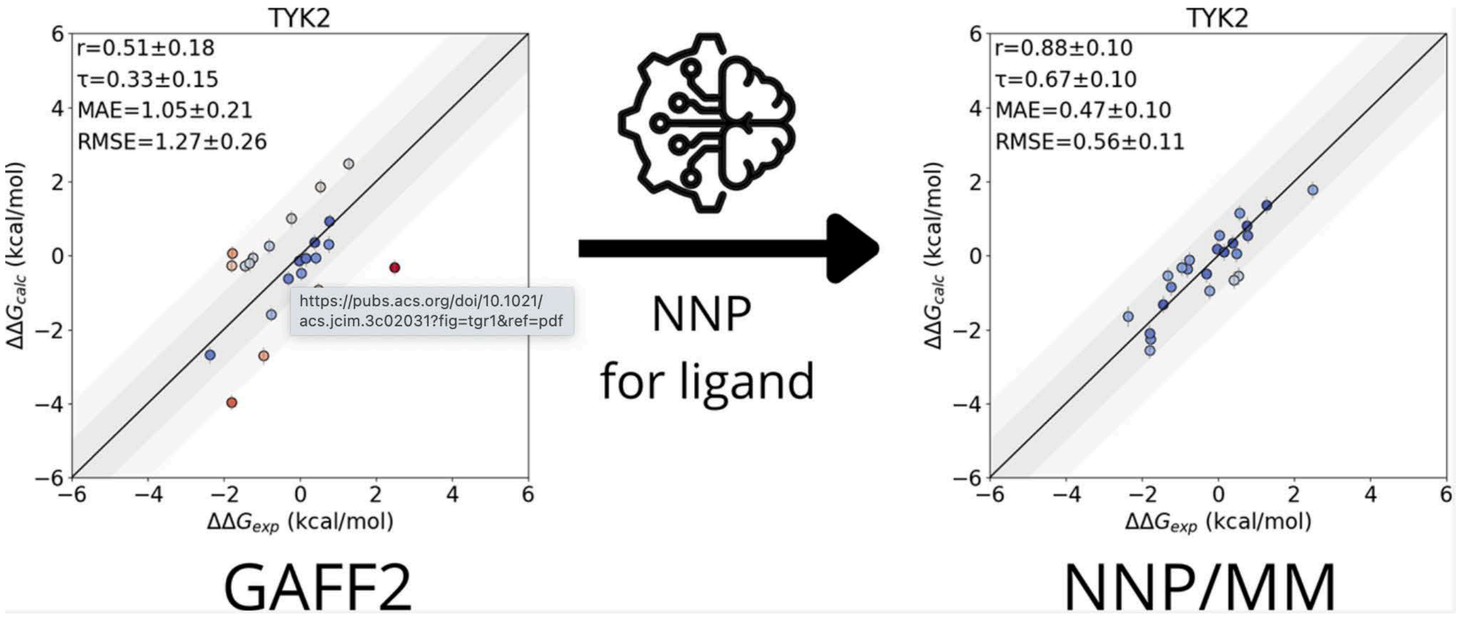


Tyk2 benchmark system from Wang et al. JACS 137:2695, 2015
replica-exchange free energy calculations with solute tempering (FEP/REST)

replica-exchange free energy calculations with perses

Rufa, Bruce Macdonald, Fass, Wieder, Grinaway, Roitberg, Isayev, and **Chodera**.
preprint: https://doi.org/10.1101/2020.07.29.227959
code: https://github.com/choderalab/qmlify

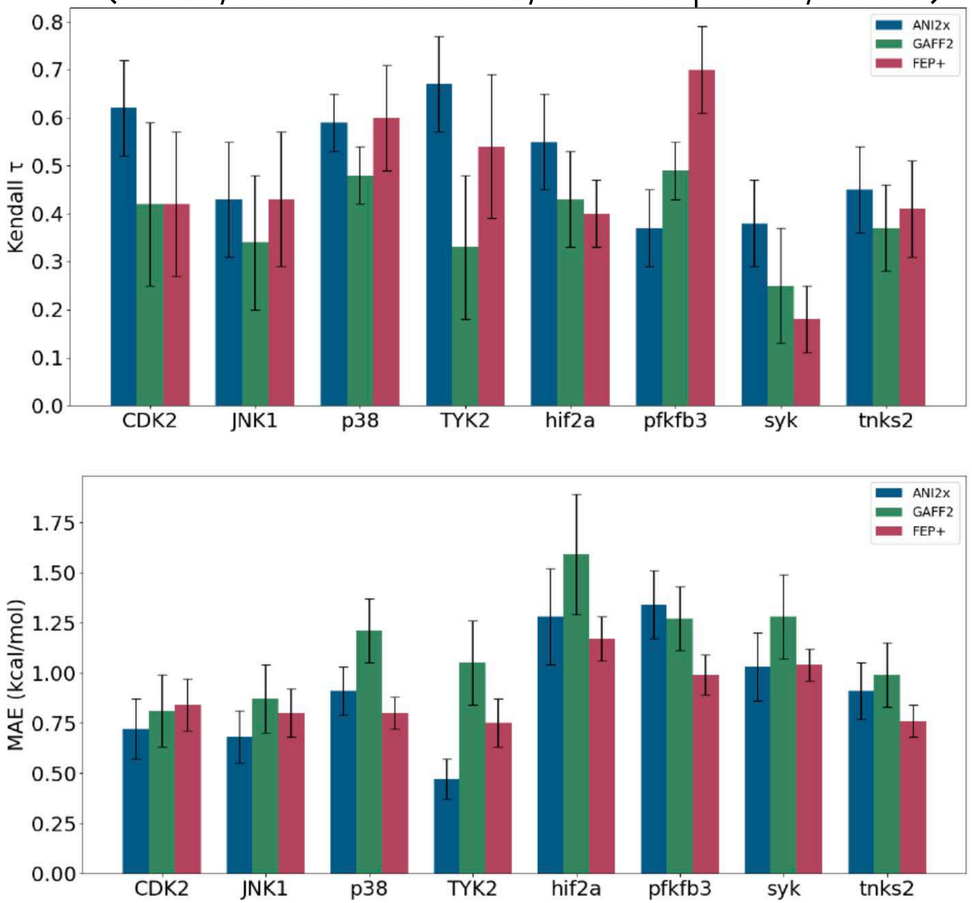
ML/MM REPEX/ATM FEP/MBAR RBFE
APPEARS TO WORK SURPRISINGLY WELL



significantly increased utility compared to GAFF2.11

ANI2x vs **GAFF2.11** vs **OPLS3e** (FEP+)

(ANI2x/GAFF used FF14SB/TIP3P for protein/solvent)



Zariquiey, Galvelis, Gallicchio, Chodera, Markland, and **De Fabritiis**
Enhancing protein-ligand binding affinity predictions using neural network potentials
JCIM 2024 (in press)

ML/MM REPEX/ATM FEP/MBAR RBFE
CAN BE SURPRISINGLY FAST

RTX 4090 benchmarks

PDB ID	# res	# heavy atoms	OpenMM ns/day (4 fs timestep)	TorchANI QML/MM ns/day (2 fs timestep)	OpenMM QML/MM* ns/day (2 fs timestep)
3BE9	328	48	995	14.0	151 / 74.2
2P95	286	50	1006	12.2	147 / 73.5
1HPO	198	64	1227	13.4	152 / 65.9
1AJV	198	75	1382	12.6	155 / 60.1

* ANI ensemble size: 1 / 8

NNPOps library
https://github.com/openmm/nnpops

* CUDA/CPU accelerated kernels
* API for inclusion in MD engines
* Ops wrappers for ML frameworks (PyTorch so far)
* Community-driven, package agnostic

~3x slower than GPU MD right now, but need 2x smaller timestep

Notably, MD will not get much faster for small systems as hardware improves.

ML *will* continue to get much faster.

paper: https://arxiv.org/abs/2201.08110
code: https://github.com/openmm/nnpops

OPENMM 8 MAKES ML/MM SIMULATIONS INCREDIBLY EASY

```
conda install -c conda-forge openmm-ml
```

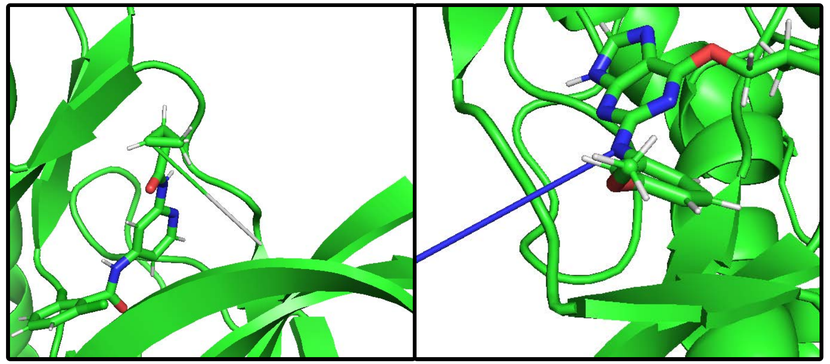
```python
# Use Amber 14SB and TIP3P-FB for the protein and solvent
forcefield = ForceField('amber14-all.xml', 'amber14/tip3pfb.xml')
# Use OpenFF for the ligand
from openmmforcefields.generators import SMIRNOFFTemplateGenerator
smirnoff = SMIRNOFFTemplateGenerator(molecules=molecules)
# Create an OpenMM MM system
mm_system = forcefield.createSystem(topology)
# Replace ligand intramolecular energetics with ANI-2x
potential = MLPotential('ani2x')
ml_system = potential.createMixedSystem(topology, mm_system, ligand_atoms)
```

https://github.com/openmm/openmm-ml

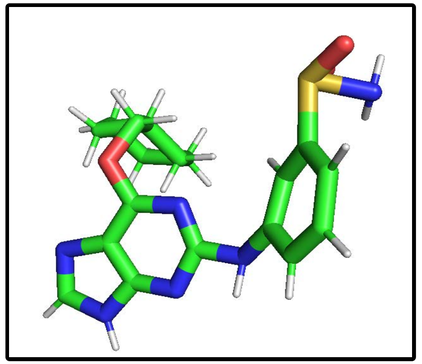
ML POTENTIALS ARE NOT WITHOUT CHALLENGES. IT'S STILL EARLY DAYS.

~ A gallery of horrors ~

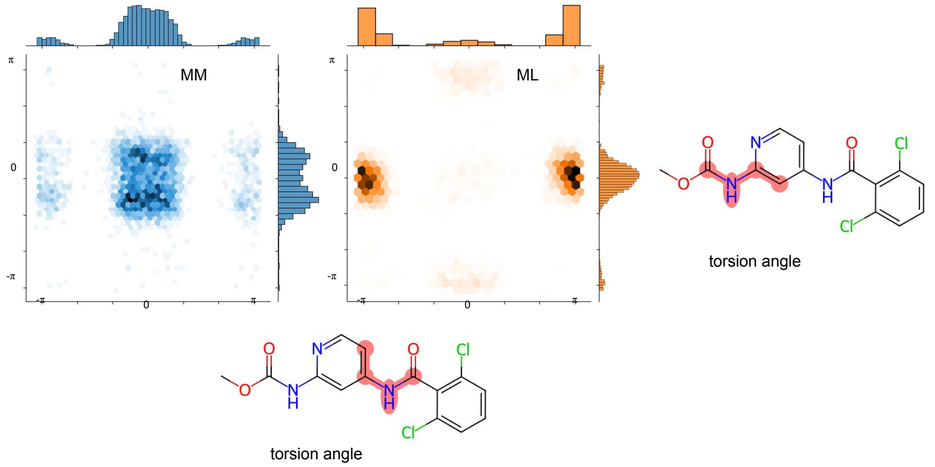
ANI2x proton cannon!



90-degree sulfonamides!

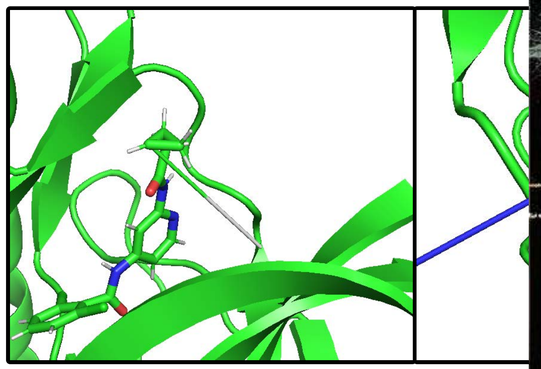


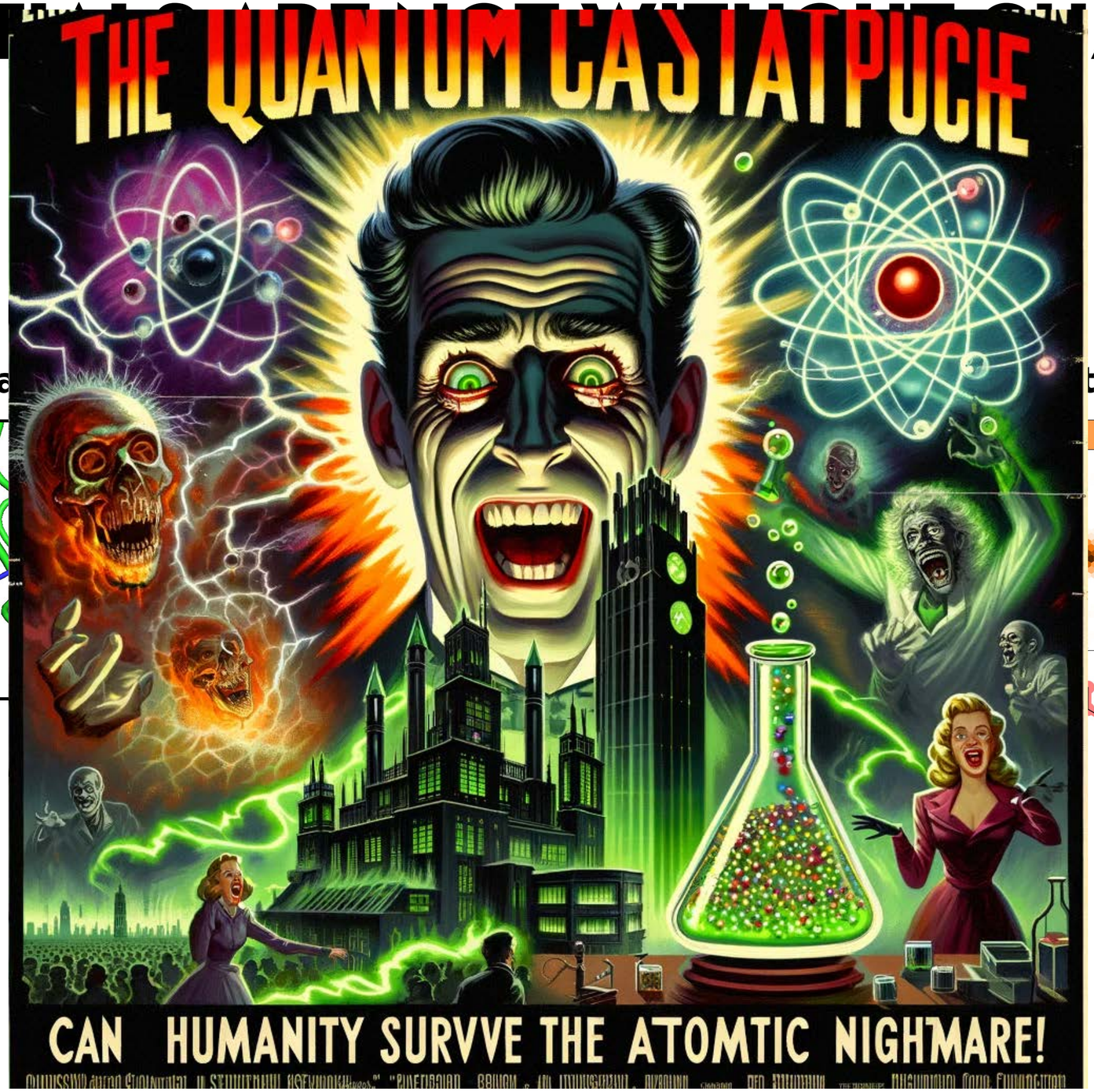
Totally different amide torsions!

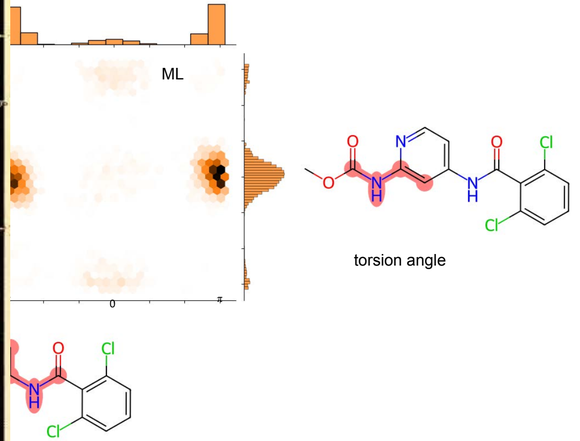


ML POTENT[illegible]ALLENGES.

ANI2x proton ca[illegible]t amide torsions!







CAN WE CHANGE PRACTICE IN STRUCTURE-ENABLED DRUG DISCOVERY BY LEVERAGING DATA WE GENERATE?

2023

week 1

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions	synthesis			new data		

using published force field model

week 2

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions	synthesis			new data		

using the same published force field model!
we haven't learned anything from the data

"Insanity is doing the same thing over and over again and expecting different results"

- Rita Mae Brown (not Albert Einstein)

CAN WE CHANGE PRACTICE IN STRUCTURE-ENABLED DRUG DISCOVERY BY LEVERAGING DATA WE GENERATE?

2023

week 1

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions	synthesis			new data		

using published force field model

week 2

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions	synthesis			new data		

using the same published force field model!
we haven't learned anything from the data

2025

week 1

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions 1.0	synthesis			new data	build model 2.0!	

using force field model
built from public + private data

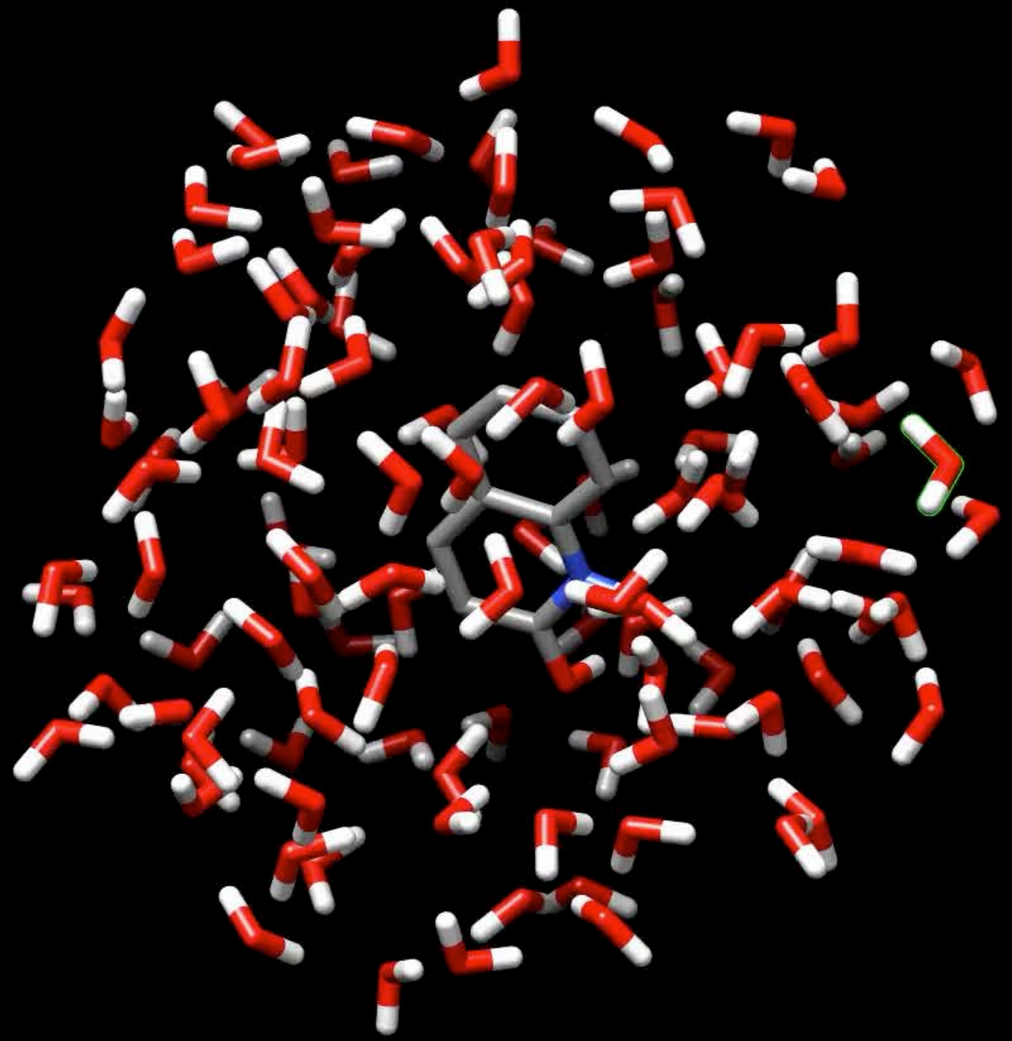
week 2

MON	TUE	WED	THU	FRI	SAT	SUN
designs/ predictions 2.0	synthesis					

using new model tuned to target
from first week's data

We want to introduce more "learnability" into our potentials

WHY DO WE NEED MM AT ALL?



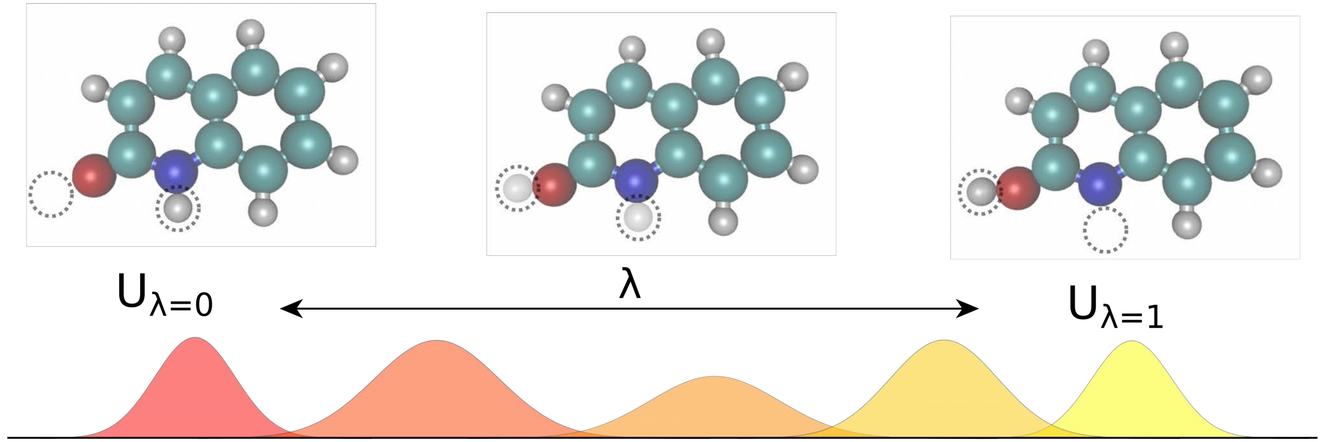
Can we just use ML force fields for everything?
We can finally be free of the hegemony of bonds!

ANI2x

ML POTENTIALS CAN BE USED TO MODEL ENTIRE SYSTEMS IN FREE ENERGY CALCULATIONS

Potentials are free of singularities, so **simple linear alchemical potentials** can robustly compute alchemical free energies

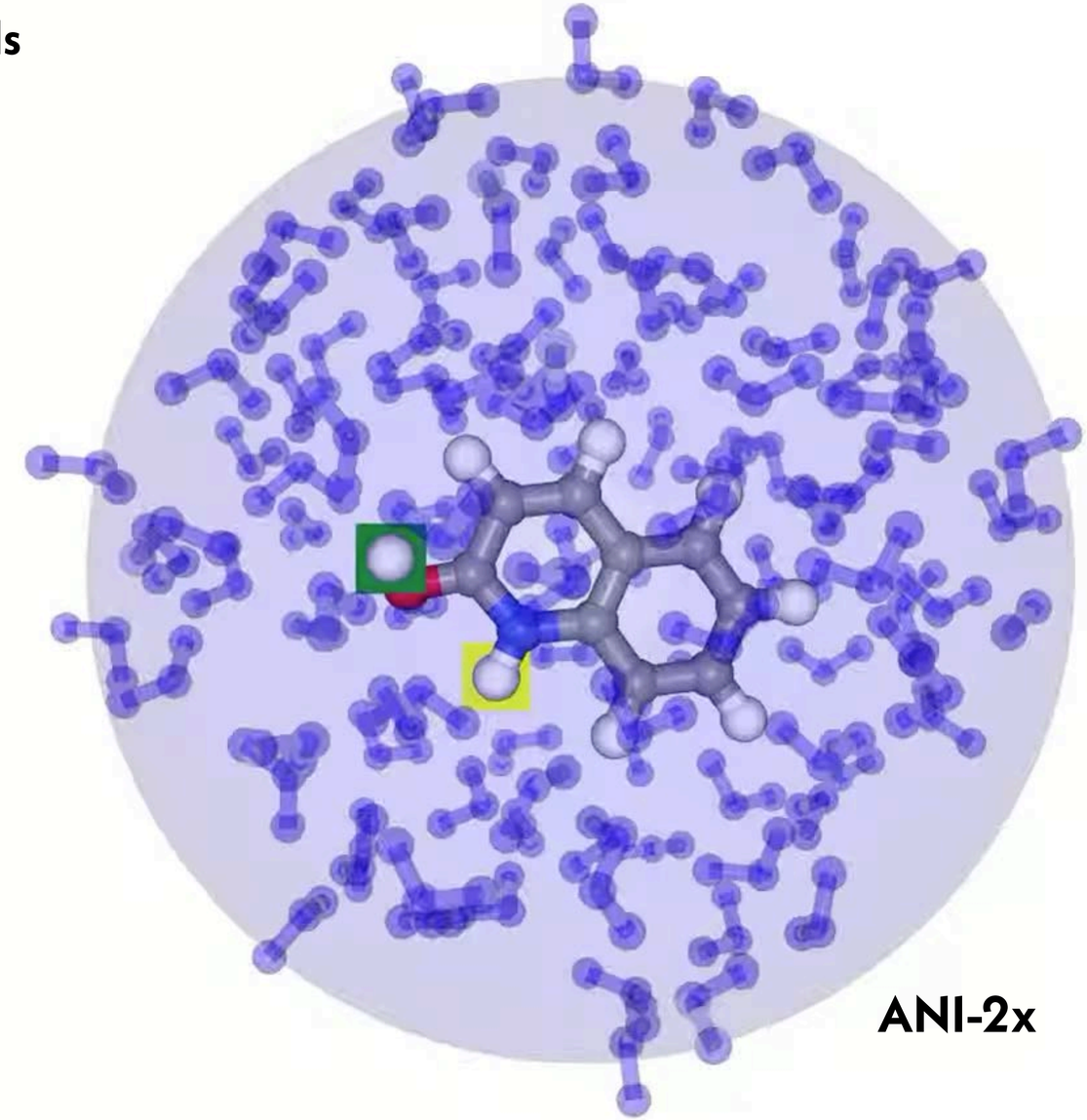
$$U(x;\lambda) = (1-\lambda)U_{\lambda=0}(x) + \lambda U_{\lambda=1}(x)$$



Simple restraints can be used when we need to enforce specific chemical species

JOSH FASS **MARCUS WIEDER**

We can even make and break bonds!



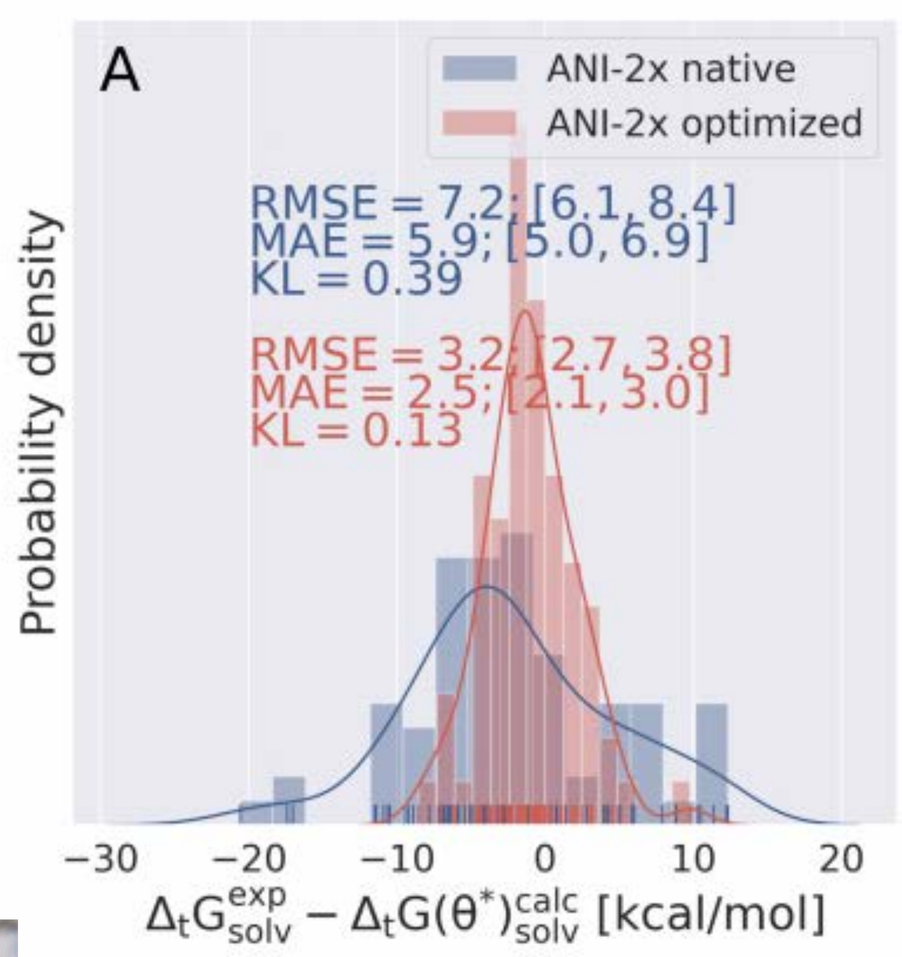
ANI-2x

preprint: https://doi.org/10.1101/2020.10.24.353318

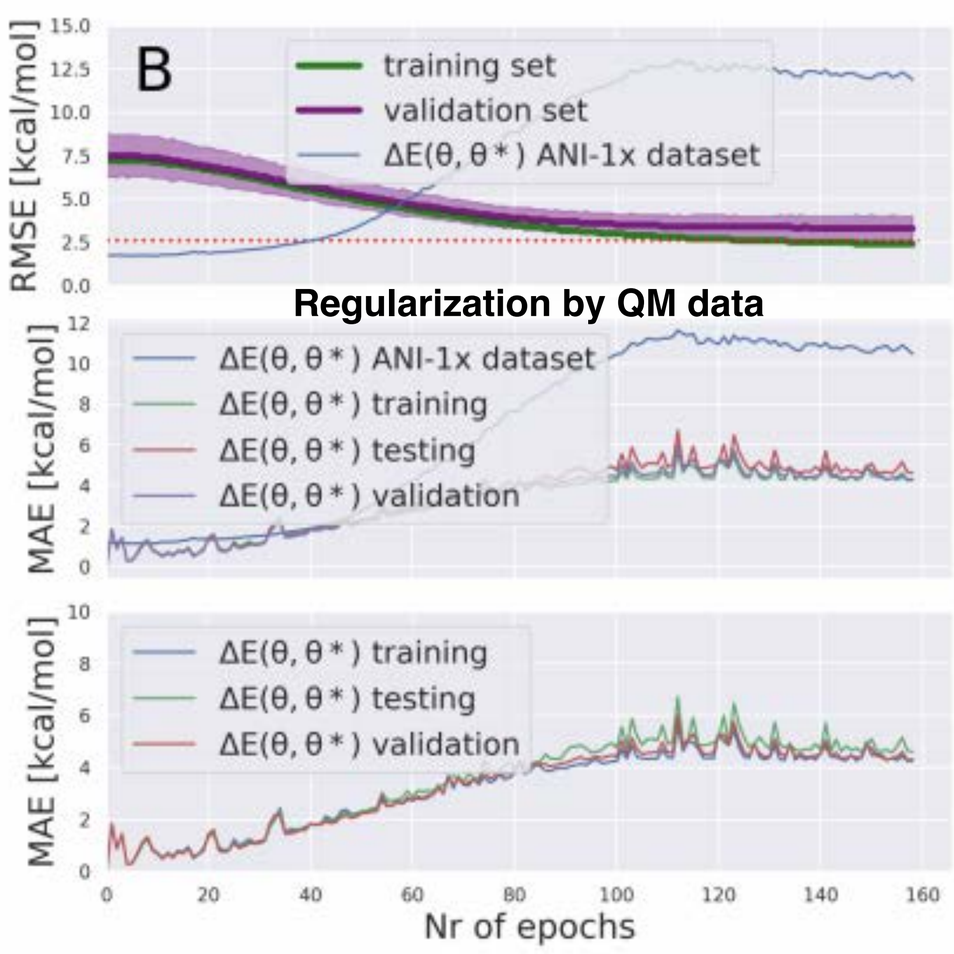
code: https://github.com/choderalab/neutromeratio

PURE ML POTENTIALS ARE NOT HIGHLY ACCURATE FOR CONDENSED PHASE PROPERTIES (YET), BUT CAN LEARN FROM DATA!

test set performance



training / validation optimization

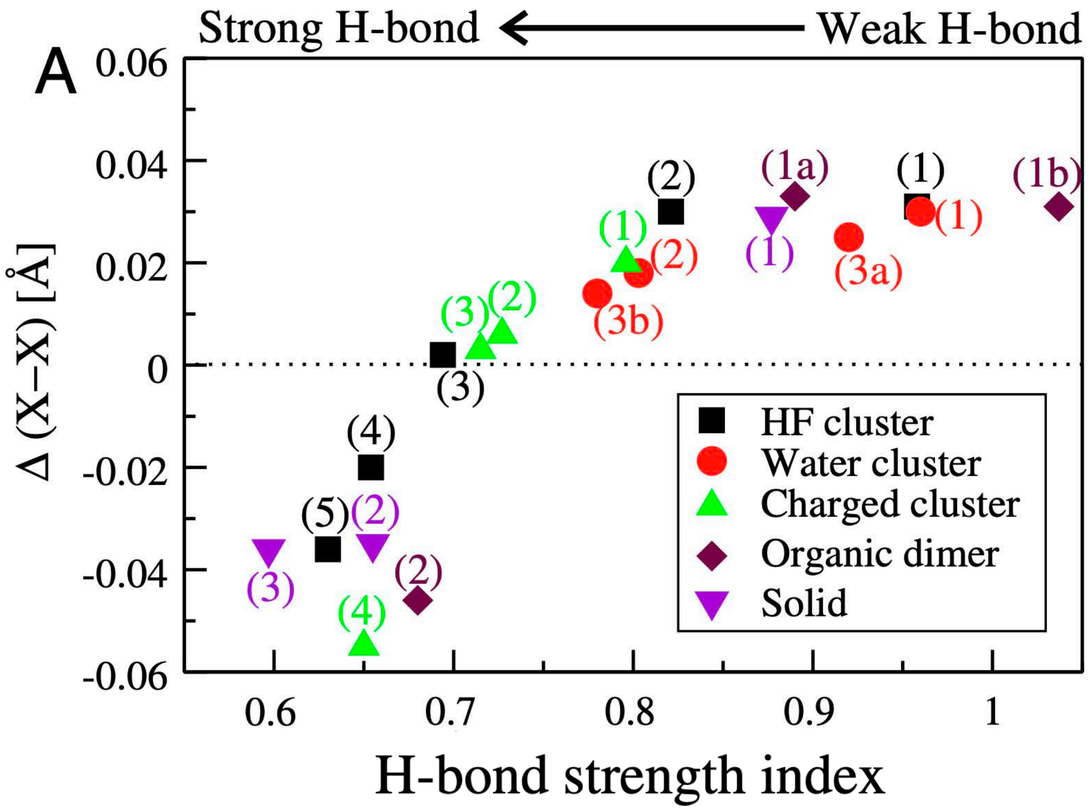


JOSH FASS MARCUS WIEDER

preprint: https://doi.org/10.1101/2020.10.24.353318
code: https://github.com/choderalab/neutromeratio

Fast on-the-fly reweighting enables inexpensive loss/gradient computation without repeating expensive free energy calculation

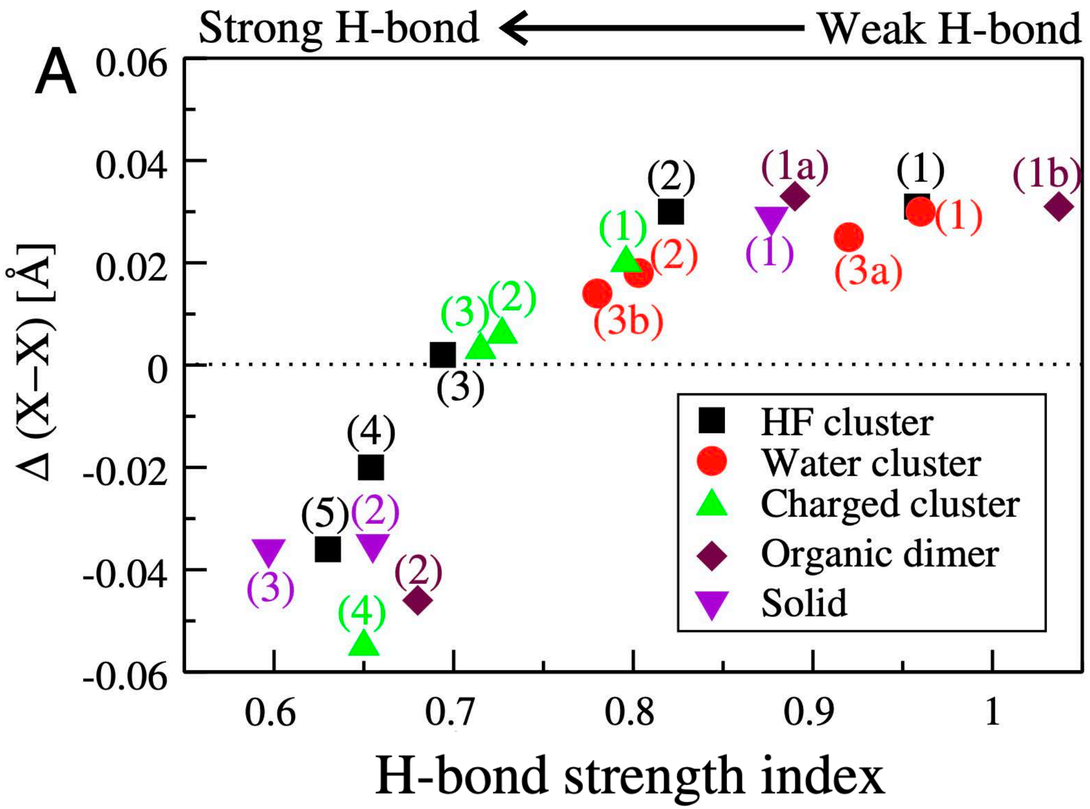
ML POTENTIALS TRAINED ONLY ON QM DATA OMIT QUANTUM NUCLEAR EFFECTS, WHICH ARE IMPORTANT FOR H-BONDS

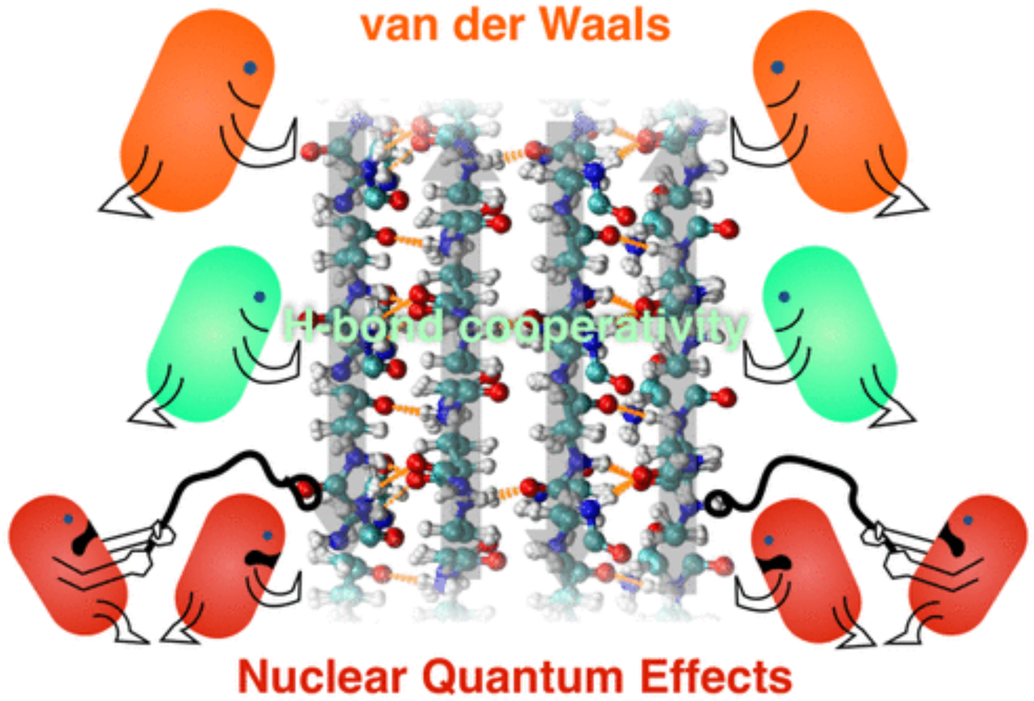


We can fix this by including experimental condensed-phase data in our ML potential training, just like we do with MM force fields

Li, Walker, Michaelides. PNAS 108:6369, 2011.
https://www.pnas.org/doi/abs/10.1073/pnas.1016653108

ML POTENTIALS TRAINED ONLY ON QM DATA OMIT QUANTUM NUCLEAR EFFECTS, WHICH ARE IMPORTANT FOR H-BONDS





We can fix this by including experimental condensed-phase data in our ML potential training, just like we do with MM force fields

Li, Walker, Michaelides. PNAS 108:6369, 2011.
https://www.pnas.org/doi/abs/10.1073/pnas.1016653108

Rossi, Fang, Michaeledis. JPC Letters 6:4233, 2015.
https://doi.org/10.1021/acs.jpclett.5b01899

WE NEED FOUNDATION DATASETS

WE ARE BUILDING FOUNDATION QM DATASETS USEFUL FOR BUILDING AND ASSESSING ML AND MM MODELS

OpenMM SPICE v1 (2M QM snapshots)

- fragments of biomolecules (and their dimers)
- dipeptides
- ion pairs
- PubChem (15K molecules)
- Solvated amino acids

Subset	Molecules	Conformations	Atoms	Elements
Dipeptides	677	33850	26–60	H, C, N, O, S
Solvated Amino Acids	26	1300	79–96	H, C, N, O, S
DES370K Dimers	3490	345676	2–34	H, Li, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I
DES370K Monomers	374	18700	3–22	H, C, N, O, F, P, S, Cl, Br, I
PubChem	14643	731856	3–50	H, C, N, O, F, P, S, Cl, Br, I
Ion Pairs	28	1426	2	Li, F, Na, Cl, K, Br, I
Total	19238	1132808	2–96	H, Li, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I

DFT ωB97M-D3(BJ)/def2-TZVPPD level of theory

>4M core-hours computed on QCFractal academic clusters

also DFT B3LYP-D3BJ/DZVP (OpenFF default)

OpenMM SPICE v2 [nearly done]

- water clusters
- PubChem (B, Si)
- amino acid : ligand fragments from the PDB
- solvated PubChem subset

+ huge thanks to Prescient/Genentech (Josh Rackers) and Exscientia

OpenMM SPICE v3 [planning]

- virtual synthetic spaces (Enamine REALSpace, etc.)
- More levels of theory

https://github.com/openmm/spice-dataset

Scientific Data 10:11, 2023
https://doi.org/10.1038/s41597-022-01882-6

THE THERMOML ARCHIVE FROM NIST PROVIDES A WEALTH OF PHYSICAL PROPERTY DATA FOR REFITTING LENNARD-JONES

NIST

Thermodynamics Research Center / ThermoML Archive / Browse | Search

Cooperating Journals

- Journal of Chemical and Engineering Data (JCED)
- The Journal of Chemical Thermodynamics (JCT)
- Fluid Phase Equilibria (FPE)
- Thermochimica Acta (TCA)
- International Journal of Thermophysics (IJT)

General Info | Data Summary | Searching Info

NIST/TRC ThermoML Archive

Summary of ThermoML Archive data points through 2019 for all cooperating journals

The ThermoML Archive includes data through 2019 for all cooperating journals as present in TRC databases on the date 2020-09-30. The /ThermoML/Browse route, linked above in the navigation gray bar, provides browsing by journal issue or by property, collated for all journals, to the source of the data points. A summary of all data point counts collected for each journal, and then overall journals, may be browsed below by property group and name.

Journal	Data Sets	Data Points	Pure Data Sets	Pure Data Points	Binary Data Sets	Binary Data Points	Ternary Data Sets	Ternary Data Points	Reaction Data Sets	Reaction Data Points	Show Details
All Journals	123727	2692934	45855	563741	58302	1385778	18324	740838	1246	2577	Show Details
J. Chem. Eng. Data	57357	1285627	20604	272263	27487	632403	9141	380519	125	442	Show Details
J. Chem. Thermodyn.	36011	857345	14043	176330	16528	464525	4603	214812	837	1678	Show Details
Thermochim. Acta	8284	144269	4109	46674	3088	60471	814	36678	273	446	Show Details
Fluid Phase Equilib.	20531	364651	6241	57466	10635	203716	3647	103461	8	8	Show Details
Int. J. Thermophys.	1544	41042	858	11008	564	24663	119	5368	3	3	Show Details

openff evaluator

https://trc.nist.gov/ThermoML/
https://docs.openforcefield.org/projects/evaluator

WE NEED FOUNDATION MODELS

WE'VE ONLY SEEN THE FIRST STEPS TOWARD FOUNDATION ML POTENTIALS FOR SMALL MOLECULE DRUG DISCOVERY

ANI2x : 9M QM calculations, 7 elements (H, C, N, O, F, Cl, S)
https://pubs.acs.org/doi/10.1021/acs.jctc.0c00121

AIMNet2 : 20M QM calculations, 14 elements, charged and neutral species
https://doi.org/10.26434/chemrxiv-2023-296ch

MACE-OFF23 : 2M QM calculations, trained on SPICE dataset (15 elements)
https://arxiv.org/abs/2312.15211

openmm-ml: https://github.com/openmm/openmm-ml

```
conda install -c conda-forge openmm-ml
```

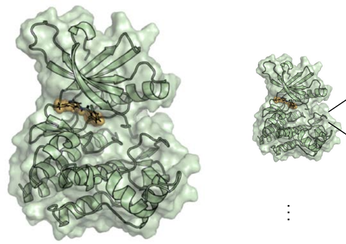
```
from openmmml import MLPotential
potential = MLPotential('ani2x')
system = potential.createSystem(topology)
```

A NEW PARADIGM EMERGES

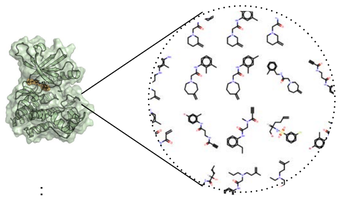
use **foundation model** trained on public data

fine-tune model on internal program data

fine-tune model on internal program data



structure-enabled target



100K REALSpace ideas

initial hits with structural hypotheses
1-10 hits
20 - 50 μM
no need for SAR

ideation of related compounds from Enamine REALSpace
select 100K from 38.5B compounds

100K REALSpace ideas

96 - 192 REALSpace compounds selected

Prioritize compounds for synthesis
[1 week]

Enamine REALSpace synthesis and assay
[4 weeks]

tailoring
[1 day]

100K REALSpace ideas

96 - 192 REALSpace compounds selected

Prioritize compounds for synthesis
[1 week]

Enamine REALSpace synthesis and assay
[4 weeks]

tailoring
[1 day]

100K REALSpace ideas

96 - 192 REALSpace compounds selected

Prioritize compounds for synthesis
[1 week]

Optimized initial leads with progressable Tier 1 ADME and SAR

Target candidate profile (TCP) for acceptance into lead optimization program

Round 1

Round 2

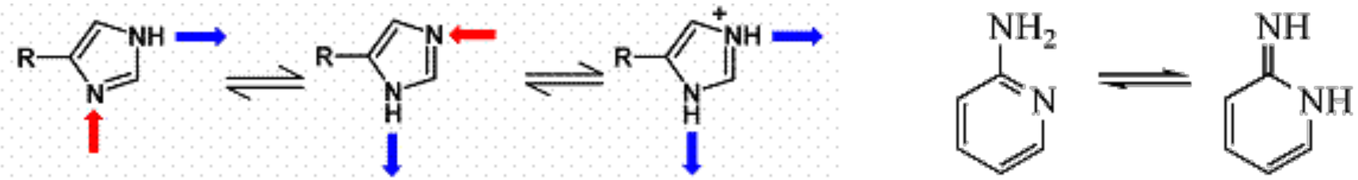
Round 3

WHAT IS HOLDING FREE ENERGY CALCULATIONS BACK?

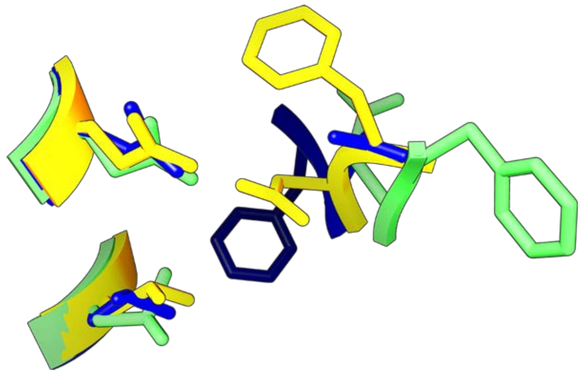
1. The **forcefield** may do a poor job of modeling the physics of our system (because it is constrained by choices made 40 years ago)

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

2. We're missing some **essential chemical** in our simulations because we don't bother to model them (e.g. protonation states, tautomers, redox chemistry, PTMs, etc.))



3. We haven't **sampled** all of the relevant conformations because we can't simulate for long enough



ML POTENTIALS MAKE IT EASIER TO SOLVE THE OTHER CHALLENGES TOO!

1. The forcefield may do a poor job of modeling the physics of our system (because it is constrained by choices made 40 years ago)

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

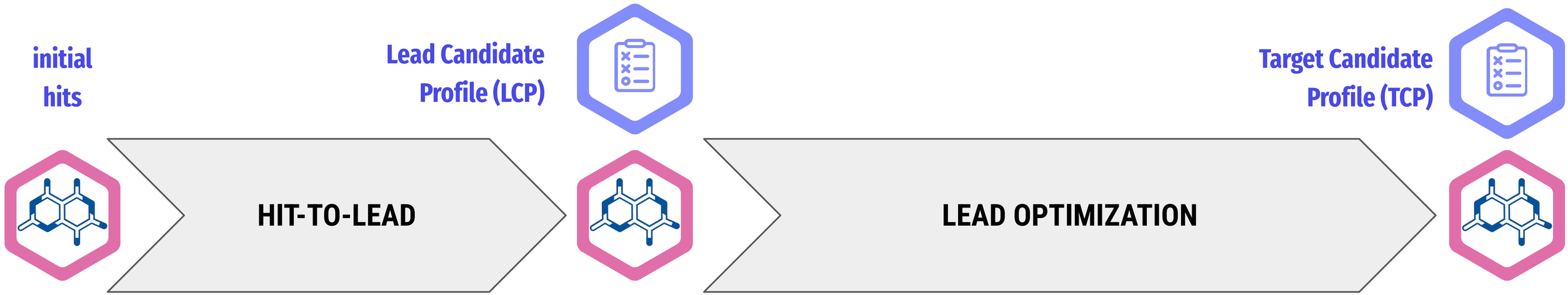
2. We're missing some essential chemical in our simulations because we don't bother to model them (e.g. protonation states, tautomers, redox chemistry, PTMs, etc.))

Incredibly easy to implement constant-pH and related algorithms now that we don't have to worry about bookkeeping MM valence terms!

3. We haven't sampled all of the relevant conformations because we can't simulate for long enough

We can turn generative models of protein conformation into clever Monte Carlo moves!

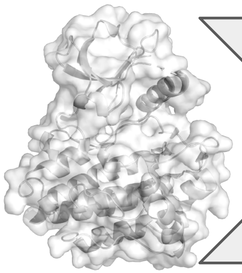
NOBODY LIKES THE CURRENT DRUG DISCOVERY PARADIGM



currently ~3.5 years, ~2000 compounds, $12.5M, ~50% success rate [1]

THERE ARE BETTER STRATEGIES WE COULD EXPLOIT IF OUR SIMULATIONS CAN LEARN FROM DATA

target



INEXPENSIVE CHEMISTRY (e.g. nanoscale chemistry, Enamine REALSpace)

rapidly build accurate model of binding site to inform FTE chemistry

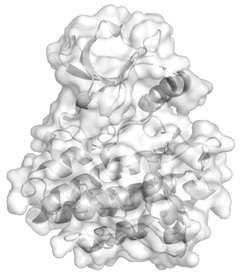
Target Candidate Profile (TCP)

HIT-TO-CANDIDATE

maybe < 1 year, < 500 molecules synthesized by CRO FTE chemists?

THERE'S THE POTENTIAL FOR A COMPLETELY NEW PARADIGM FOR DISCOVERY

target



\+ Target Candidate Profile (TCP)

GENERATIVE DESIGN CONDITIONED ON OBJECTIVES

Multiple candidates for preclinical development

<< 1 year, ~ 10 molecules synthesized by CRO FTE chemists

What does it take to get here?

WE DON'T HAVE THE SCALE OF (EXPERIMENTAL) DATA TO DO THIS

Q: Who is the president during WWII?
A: Franklin D. Roosevelt was the president during WWII.

OpenAI:

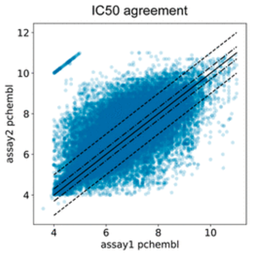
DALL-E 2 was trained on a dataset of **650 million** images

GPT-3 was trained on a corpus of **22.5 billion pages of text** (45 TB)

CADD:

Typical **drug discovery programs** make and test ~2000 compounds

PDBBind contains ~20K protein:ligand complexes



R2 = 0.31
MAE = 0.50
Kendall's tau = 0.51

BigBind contains 538K measurements paired with structures

ChEMBL contains 2.4M compounds, but it's a dumpster fire

Landrum and Riniker, JCIM 2024
https://pubs.acs.org/doi/full/10.1021/acs.jcim.4c00049

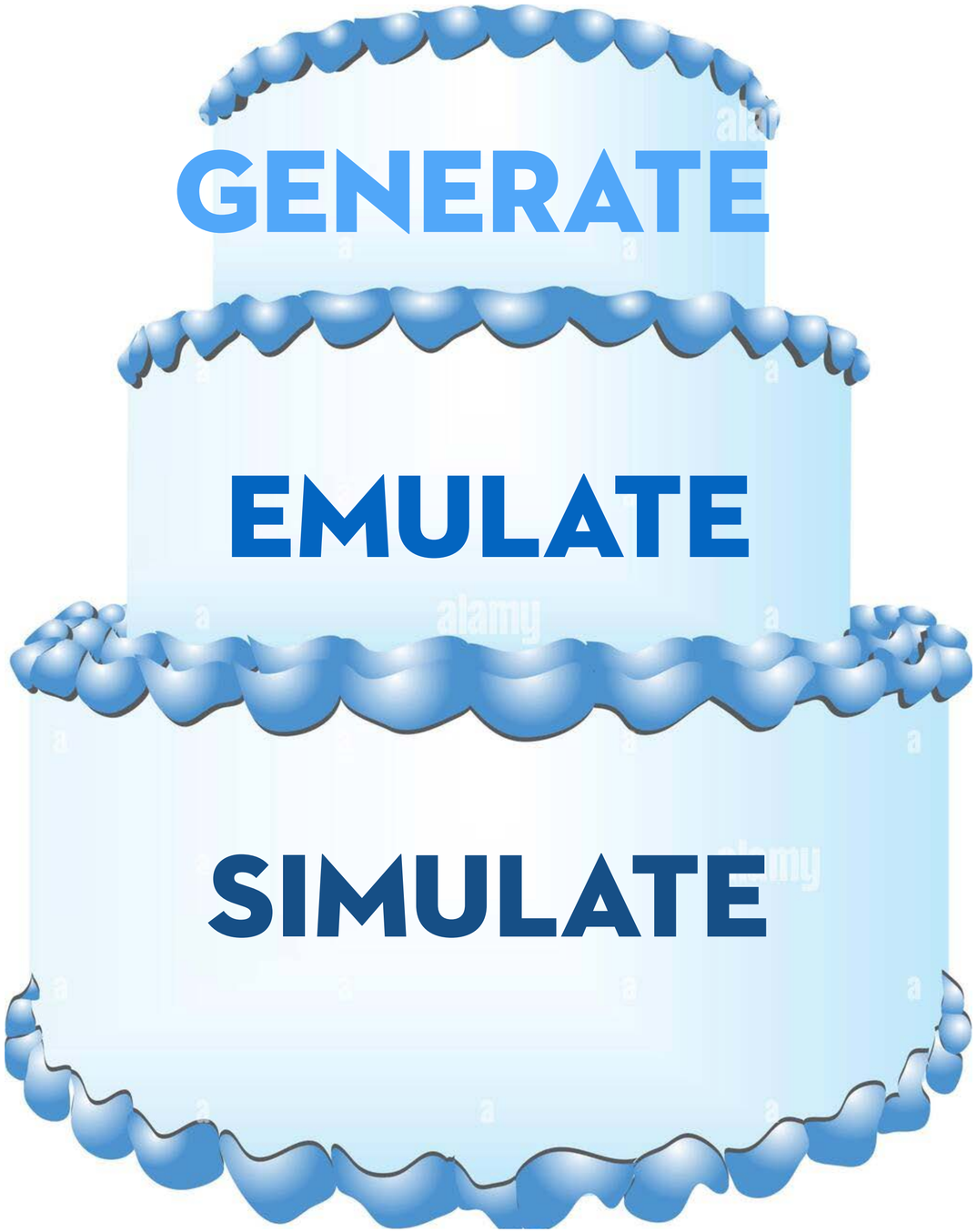
...BUT IF WE HAVE A GOOD ENOUGH SIMULATOR (AND ENOUGH MONEY), WE CAN SIMULATE OUR WAY THERE.

SIMULATE → EMULATE → GENERATE

build accurate phyiscal biomolecular simulation models from limited QM + experimental data

build surrogate models that accurately model biomolecular simulations

build generative ML models that predict molecules conditioned on design goals


GENERATE
EMULATE
SIMULATE

CHODERA LAB

- All funding: http://choderalab.org/funding