slides: http://choderalab.org/news

ALCHEMICAL ACADEMY: TEACHING FREE ENERGY CALCULATIONS TO LEARN

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

16 Mar 2023 - OpenEye CUP XXII - Santa Fe

1884

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.

E 70th St
E 71st St
1st Av
York Ave
Helmsley Medical Tower
Weill Cornell Medical College
E 69th St
Le Pain Quotidien
St Catherine of Siena Church
Memorial Sloan-Kettering Cancer Center
MSKCC
New York-Presbyterian University Hospital of Columbia and Cornell
Cornell Weill
NewYork-Presbyterian Hospital
Hospital For Special Surgery
1246
400
1278
E 67th St
St John Nepomucene Church
E 68th St
501
1248
70
FDR Drive
East River Esplanade
E 66th St
Rockefeller
Rockefeller University Hospital
York Ave
The Rockefeller University
1192
Franklin D. Roosevelt East River Drive
East River Esplanade
E 63rd St
12

Dana
Pe’er

Quaid
Morris

Christina
Leslie

Joao
Xavier

Kushal
Dey

John
Chodera

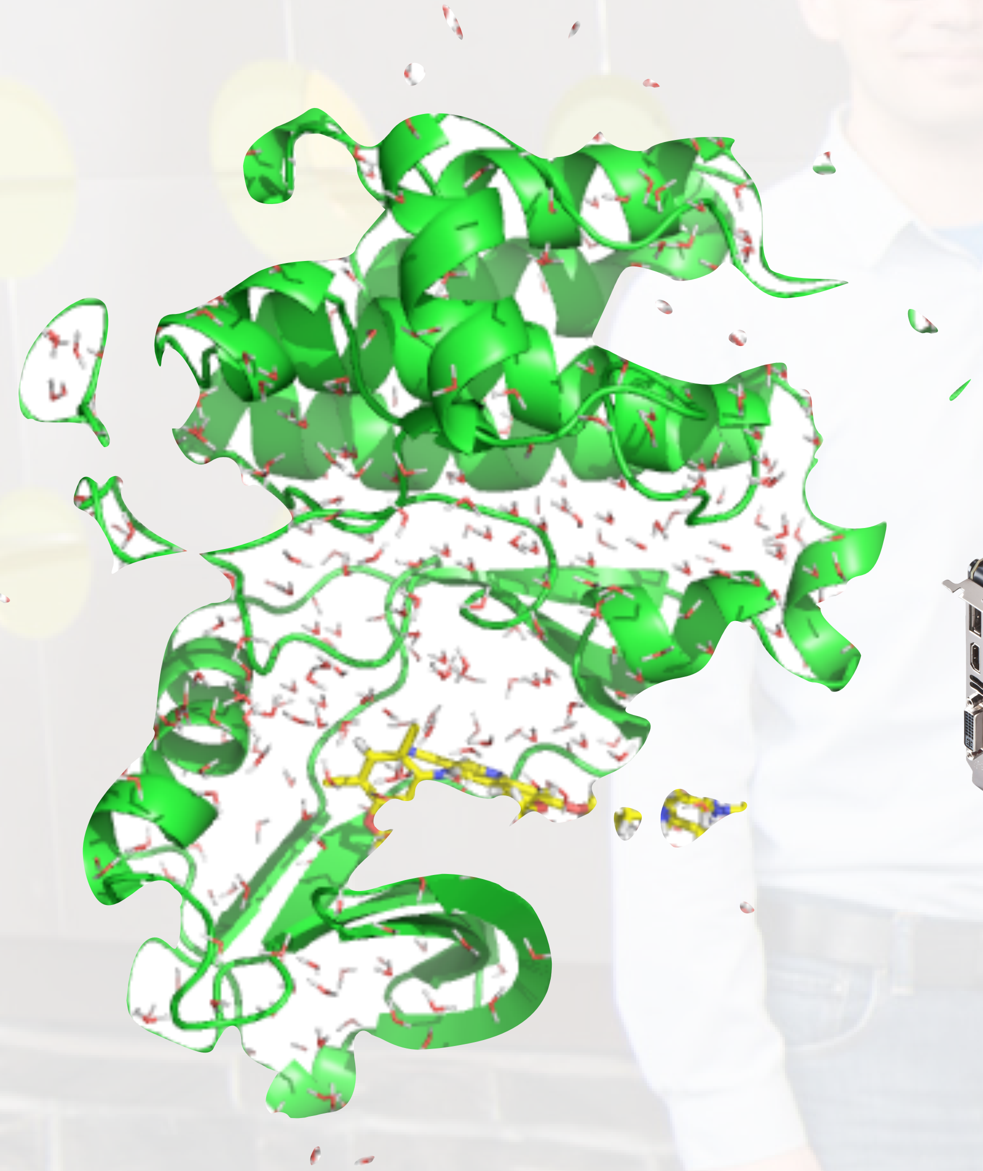
Thomas
Norman

csbio@MSKCC

CHODERA LAB

HOW CAN COMPUTATIONAL BIOPHYSICS AND MACHINE LEARNING ADVANCE DISCOVERY AND TREATMENT IN THE ERA OF CANCER GENOMICS?

MODELING



FOLDING @HOME

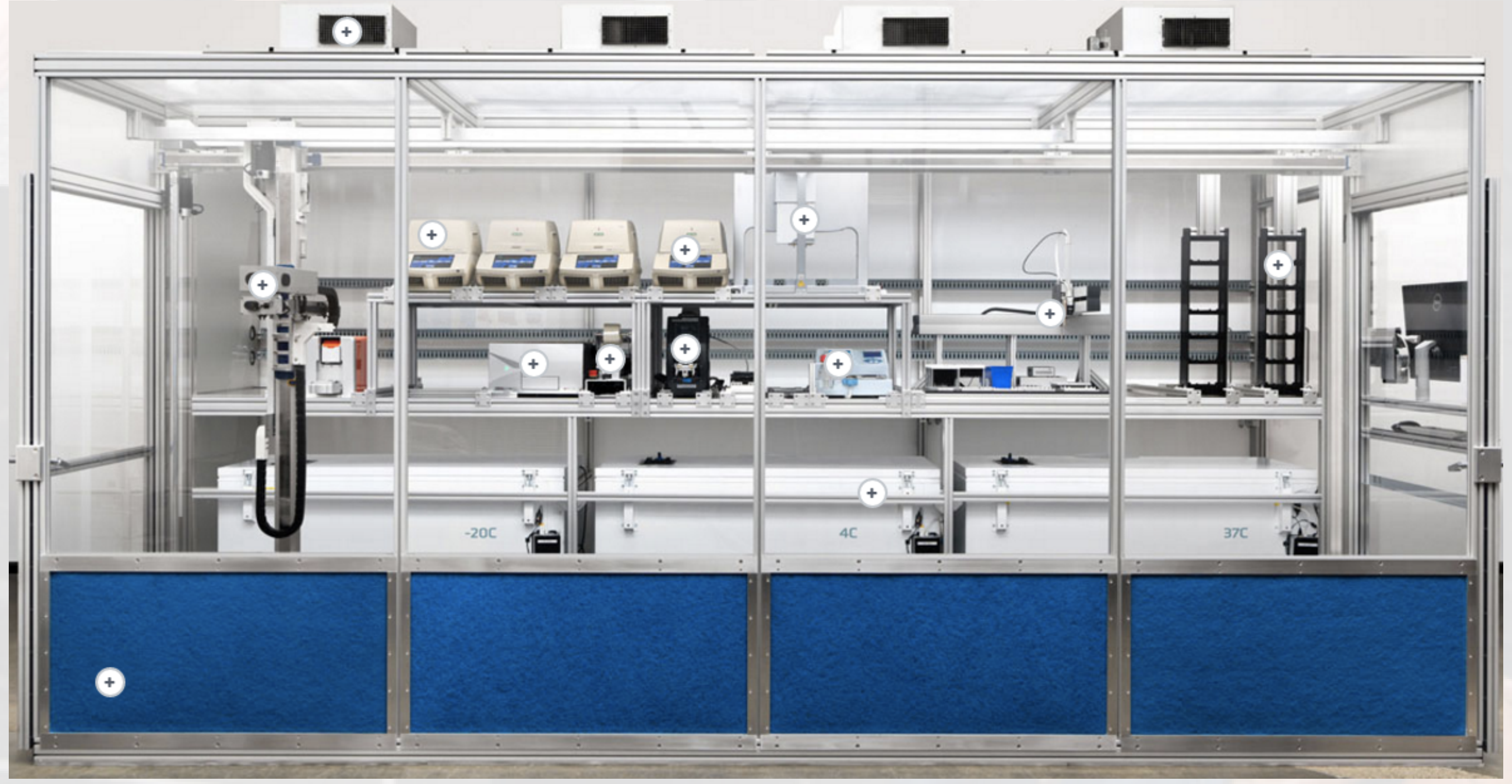
amazon web services EC2

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

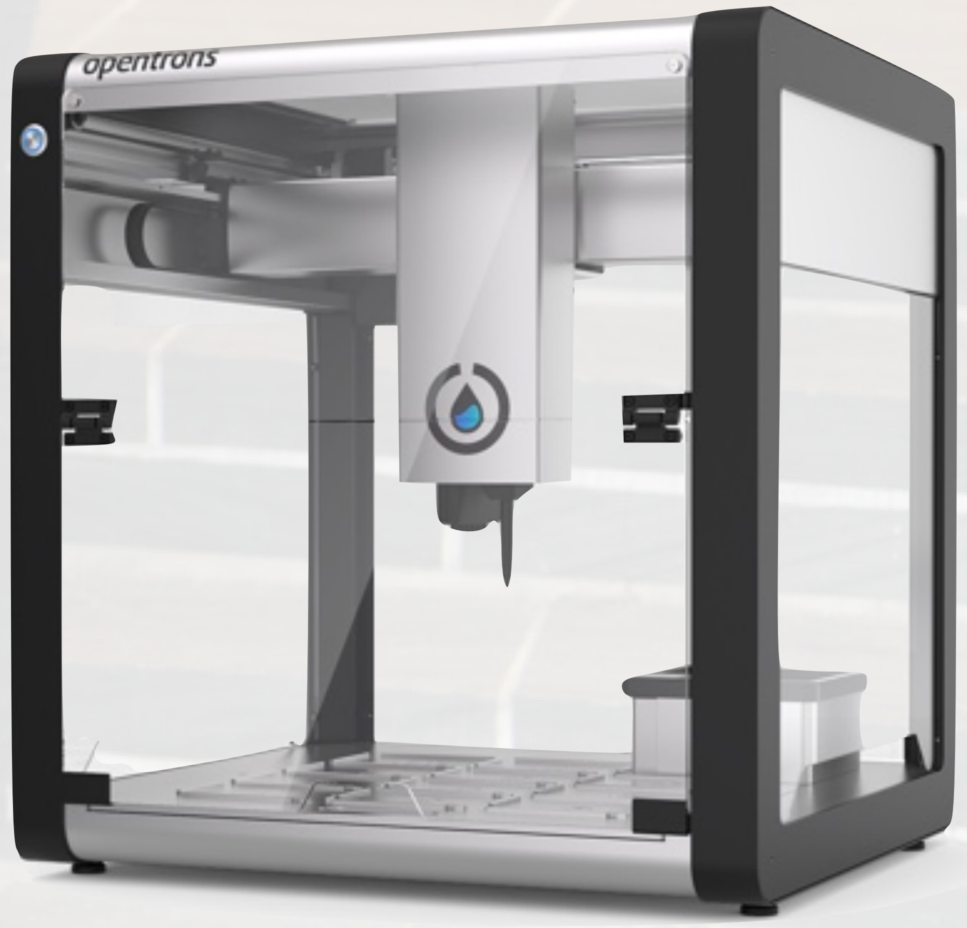
$$+ \sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

AUTOMATION

CHODERA LAB, Z17



STRATEOS CLOUD WETLAB



OPENTRONS

WE COLLABORATE BROADLY TO ADVANCE THE STATE OF DRUG DISCOVERY

BAYER

INTERLINE therapeutics

open forcefield consortium

Open Free Energy

OpenMM

Open Molecular Software Foundation

MolSSI
Molecular Sciences Software Institute

Folding@home

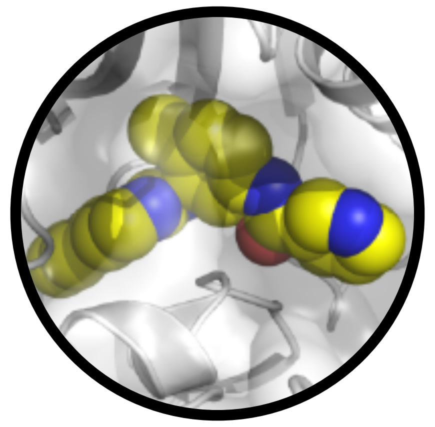
NIH
National Center for Advancing Translational Sciences

open source software development initiatives

Ventus THERAPEUTICS

SILICON Therapeutics

industry collaborations



choderalab
(algorithms and open source software)

data generators, community challenges, and resources

ASAP
Accessible Antivirals to Prevent Pandemics

The SAMPL challenges
Challenge information and announcements

Drug Design Data Resource

COVID Moonshot

OpenEye SCIENTIFIC

academia

CACHE CHALLENGE

XCHEM Diamond Fragment Screening

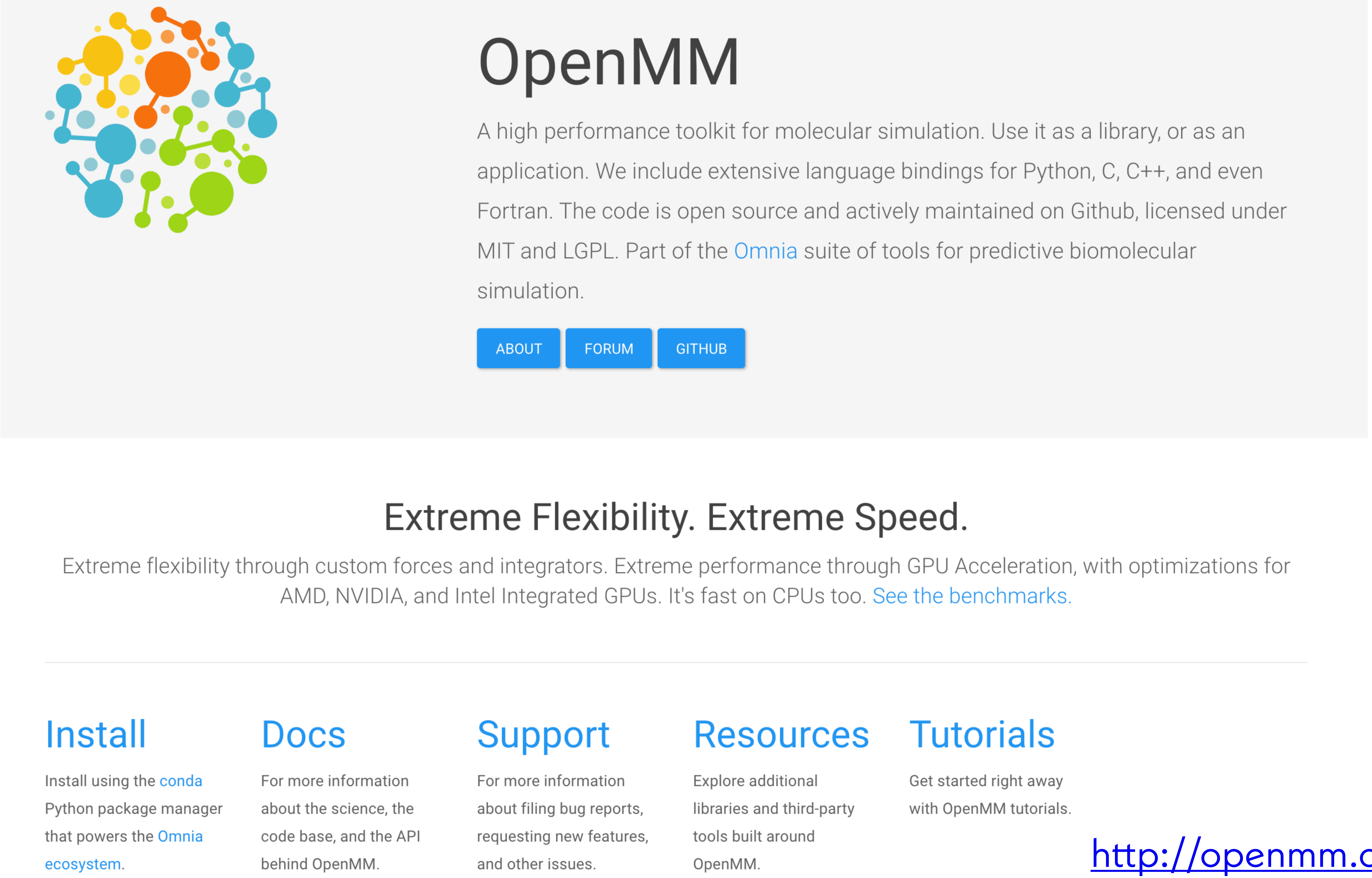
Diamond Light Source / XChem

VIR

SGC

IP-generating collaborations

open science / open source software



OpenMM

A high performance toolkit for molecular simulation. Use it as a library, or as an application. We include extensive language bindings for Python, C, C++, and even Fortran. The code is open source and actively maintained on Github, licensed under MIT and LGPL. Part of the Omnia suite of tools for predictive biomolecular simulation.

Extreme Flexibility. Extreme Speed.

Extreme flexibility through custom forces and integrators. Extreme performance through GPU Acceleration, with optimizations for AMD, NVIDIA, and Intel Integrated GPUs. It's fast on CPUs too. See the benchmarks.

Install

Install using the conda Python package manager that powers the Omnia ecosystem.

Docs

For more information about the science, the code base, and the API behind OpenMM.

Support

For more information about filing bug reports, requesting new features, and other issues.

Resources

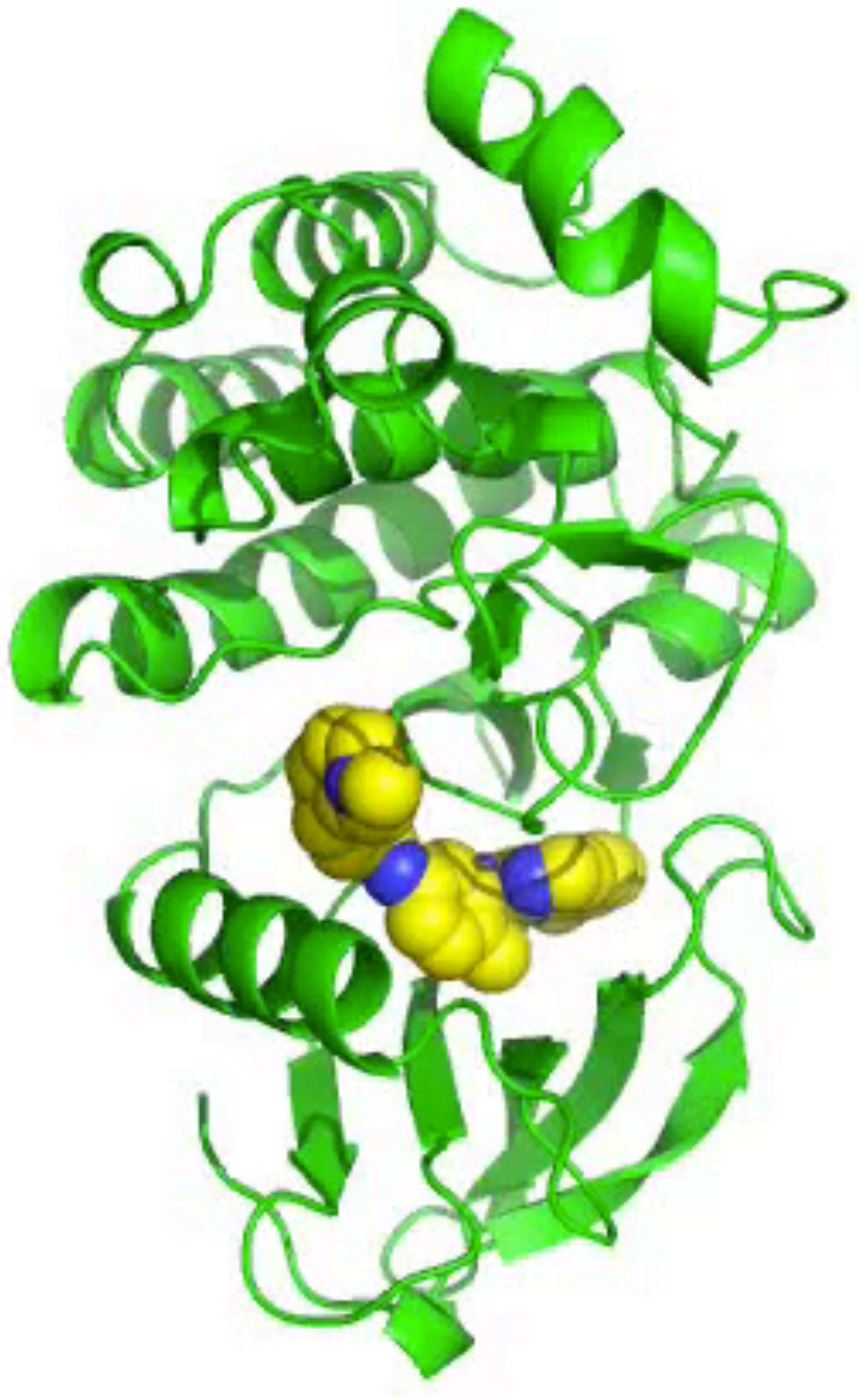
Explore additional libraries and third-party tools built around OpenMM.

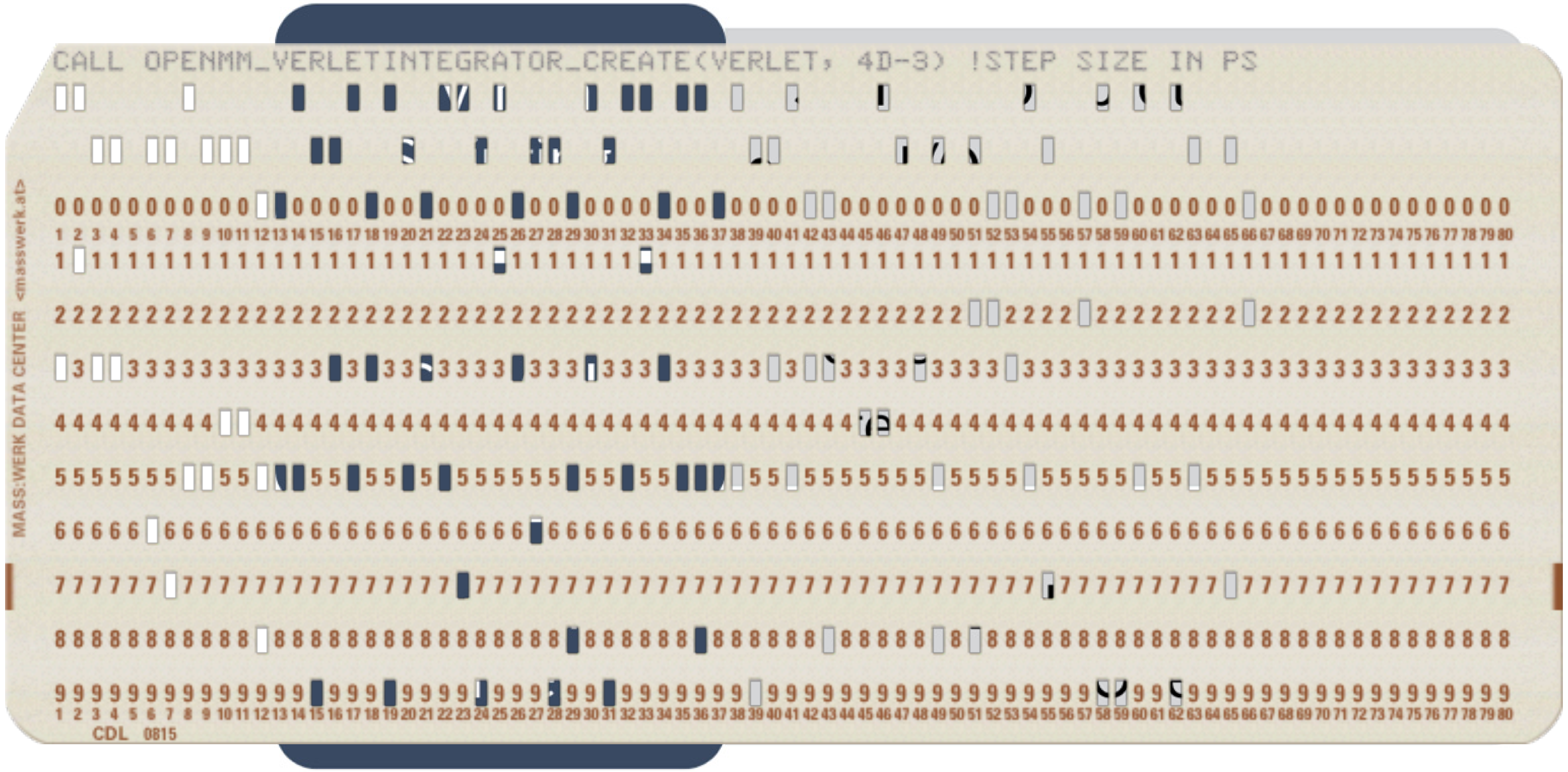
Tutorials

Get started right away with OpenMM tutorials.

http://openmm.org/

OPENMM ARCHITECTURE MAKES DEVELOPMENT SIMPLE



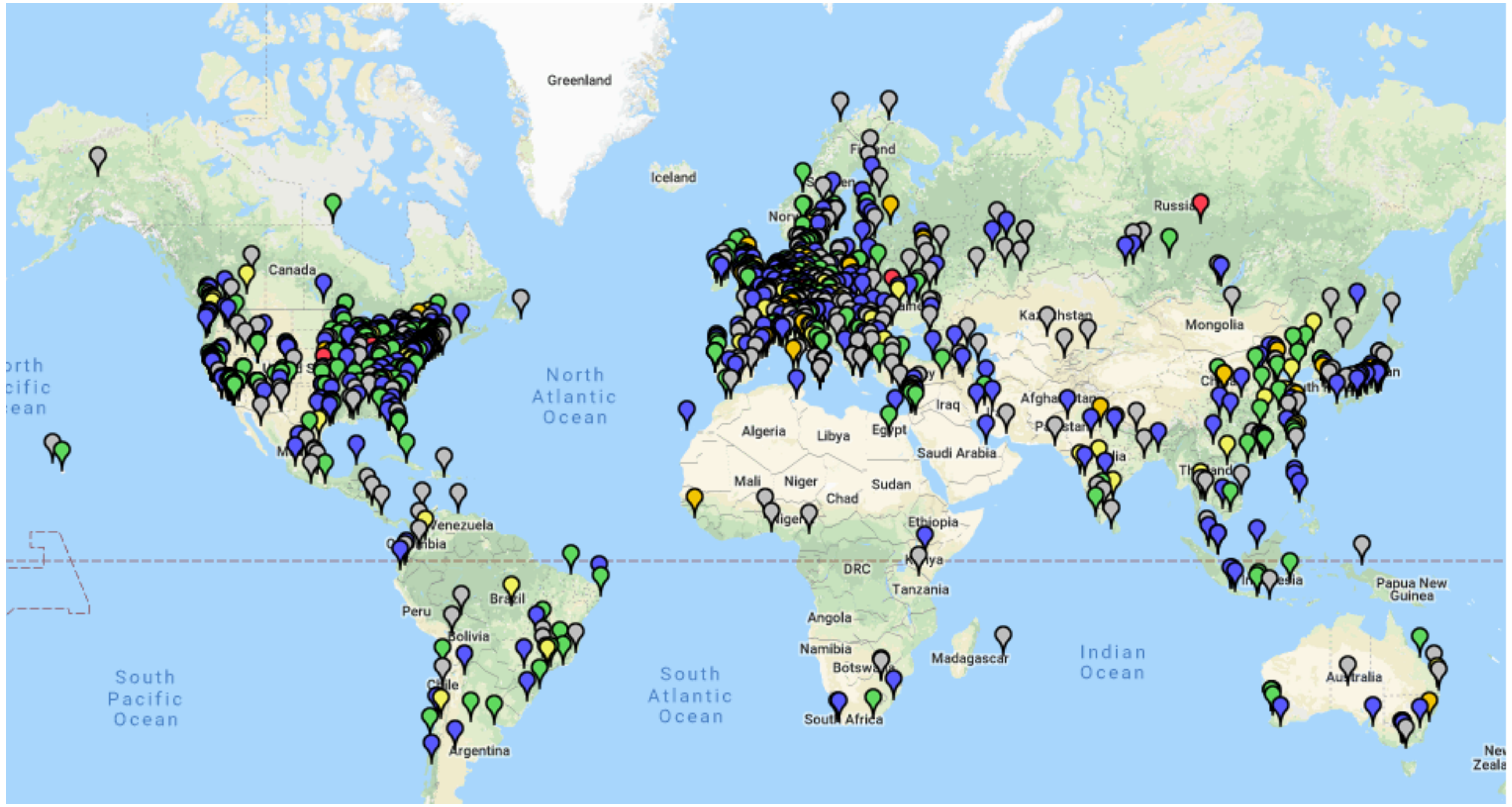


```
CALL OPENMM_VERLETINTEGRATOR_CREATE(VERLET, 4D-3) !STEP SIZE IN PS
```

OpenMM also has bindings for **C++**, **C**, and **FORTRAN**

https://www.masswerk.at/keypunch/

OPENMM IS USED BY RESEARCHERS ALL OVER THE WORLD



Geographic statistics from http://simtk.org

CONDA-FORGE

ANACONDA
CLOUD

OpenMM

http://openmm.org

downloads 1M total

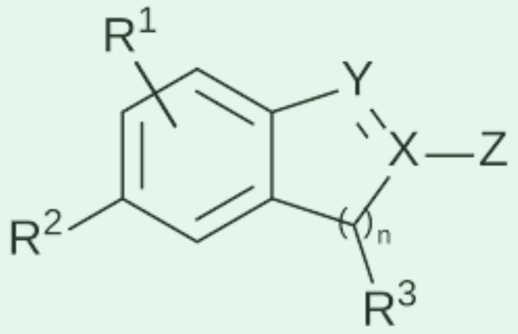
OpenMMTools

http://github.com/choderalab/openmmtools

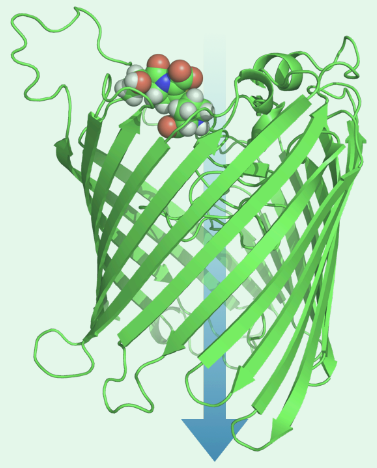
downloads 174k total

OPENMM CAN BE USED AS A LIBRARY TO ENABLE APPLICATIONS TO INTEGRATE PHYSICAL MODELING

isolde



perses



iapetus

targeted domain-specific applications
(Python, C++, C, or Fortran)

APPLICATIONS

openmmtools

high-level simulation algorithms, alchemical tools
(Python to enable rapid development)

ALGORITHMS

OpenMM

general GPU-accelerated MD simulation engine
(C++/CUDA/OpenCL with Python API)

CORE

DESIGNING REAL PRECLINICAL DRUG CANDIDATES IS CHALLENGING

Ed Griffen

Medchemica

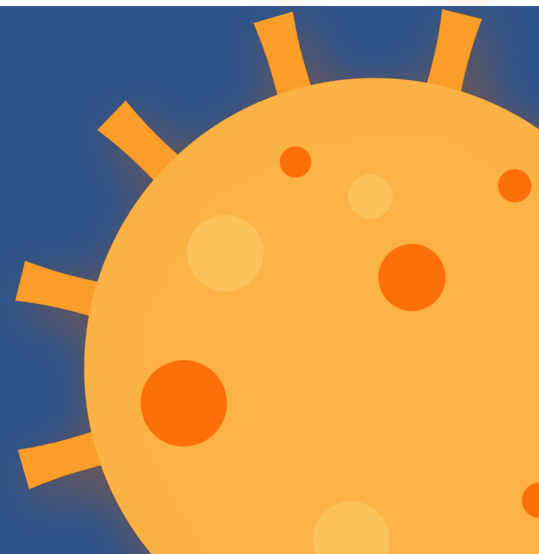
Target Candidate Profile (TCP) for oral SARS-CoV-2 main viral protease (Mpro) inhibitor

Property	Target range	Rationale
protease assay	$IC_{50} < 10$ nM	Extrapolation from other anti-viral programs
viral replication assay	$EC_{50} < 5$ μM	Suppression of virus at achievable blood levels
plaque reduction assay	$EC_{50} < 5$ μM	Suppression of virus at achievable blood levels
route of administration	oral	bid/tid - compromise PK for potency if pharmacodynamic effect achieved
solubility	> 5 mg/mL	Aim for biopharmaceutical class 1 assuming <= 750 mg dose
half-life	> 8 h (human) est from rat and dog	Assume PK/PD requires continuous cover over plaque inhibition for 24 h max bid dosing
safety	Only reversible and monitorable toxicities No significant DDI - clean in 5 CYP450 isoforms hERG and NaV1.5 $IC_{50} > 50$ μM No significant change in QTc Ames negative No mutagenicity or teratogenicity risk	No significant toxicological delays to development DDI aims to deal with co-morbidities / therapies, cardiac safety for COVID-19 risk profile cardiac safety for COVID-19 risk profile Low carcinogenicity risk reduces delays in manufacturing Patient group will include significant proportion of women of childbearing age

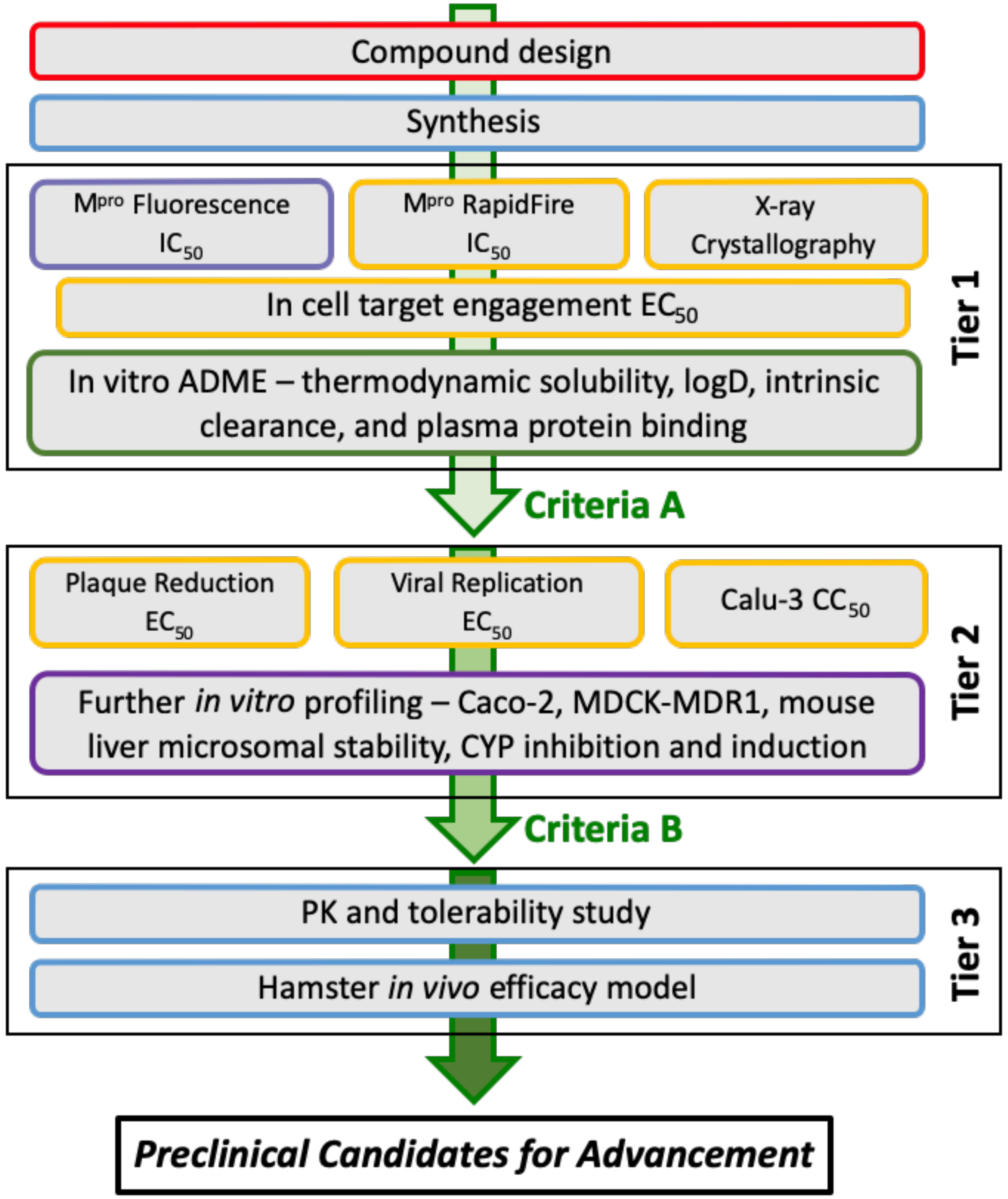
COVID Moonshot

An international effort to
DISCOVER A COVID ANTIVIRAL

https://covid.postera.ai/covid



TO GET THERE, DRUG DESIGN INVOLVES MAKING A LOT OF DECISIONS ABOUT WHICH MOLECULES WILL ACHIEVE CERTAIN OBJECTIVES



Compound design

Synthesis

M^{pro} Fluorescence IC$_{50}$

M^{pro} RapidFire IC$_{50}$

X-ray Crystallography

In cell target engagement EC$_{50}$

In vitro ADME – thermodynamic solubility, logD, intrinsic clearance, and plasma protein binding

Tier 1

Criteria A

Plaque Reduction EC$_{50}$

Viral Replication EC$_{50}$

Calu-3 CC$_{50}$

Further *in vitro* profiling – Caco-2, MDCK-MDR1, mouse liver microsomal stability, CYP inhibition and induction

Tier 2

Criteria B

PK and tolerability study

Hamster *in vivo* efficacy model

Tier 3

Preclinical Candidates for Advancement

assay purpose

Does it inhibit the target? How does it bind?

Does it work in cells?

Does it have a chance of working in humans?

Does it kill the virus in cells?

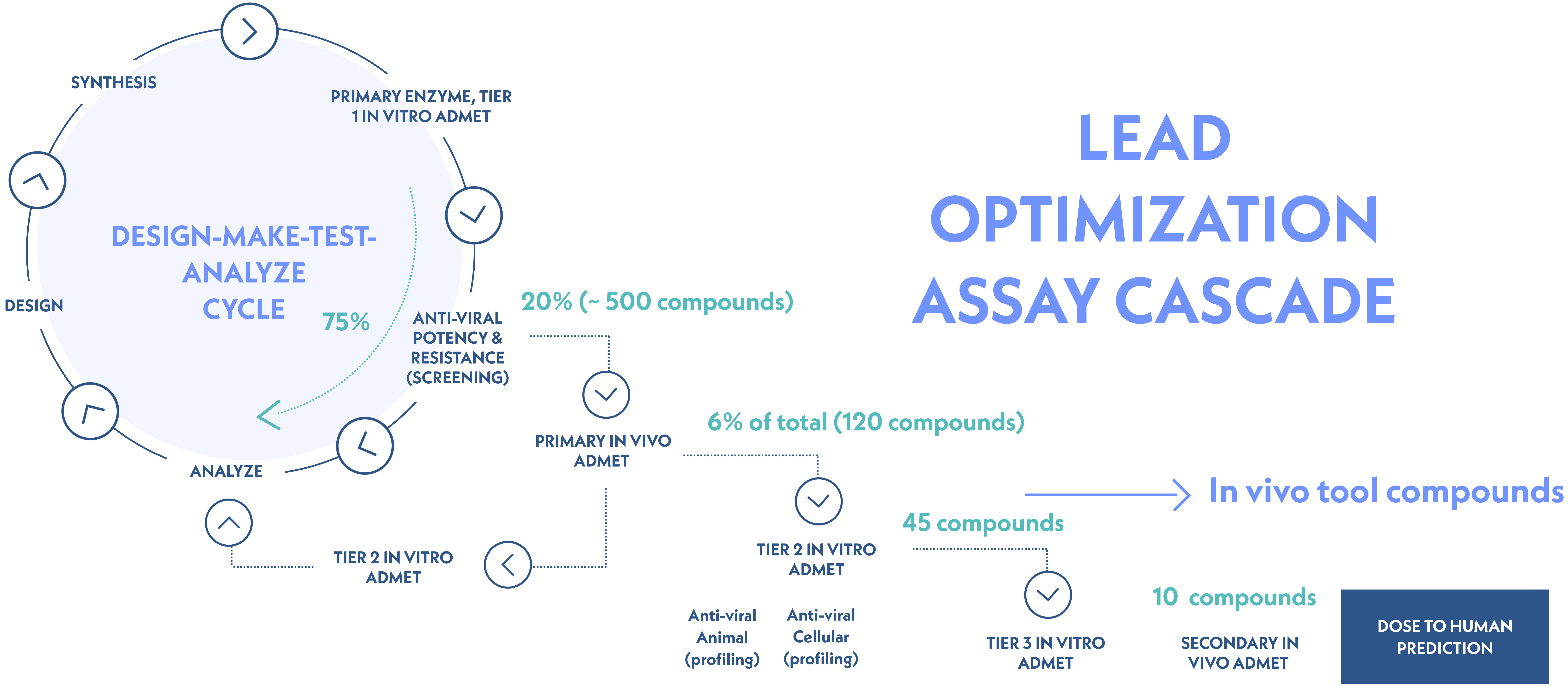
Could it cause bad side effects?

Can oral dosing deliver sufficient drug?

Does it actually work against the disease?

MUCH OF THE TIME IS SPENT IN PREDICTING COMPOUNDS THAT WILL IMPROVE OR MAINTAIN POTENCY

LEAD OPTIMIZATION ASSAY CASCADE



DESIGN-MAKE-TEST-ANALYZE CYCLE

SYNTHESIS

PRIMARY ENZYME, TIER 1 IN VITRO ADMET

ANTI-VIRAL POTENCY & RESISTANCE (SCREENING)

75%

ANALYZE

DESIGN

TIER 2 IN VITRO ADMET

20% (~ 500 compounds)

PRIMARY IN VIVO ADMET

6% of total (120 compounds)

TIER 2 IN VITRO ADMET

Anti-viral Animal (profiling)

Anti-viral Cellular (profiling)

45 compounds

In vivo tool compounds

TIER 3 IN VITRO ADMET

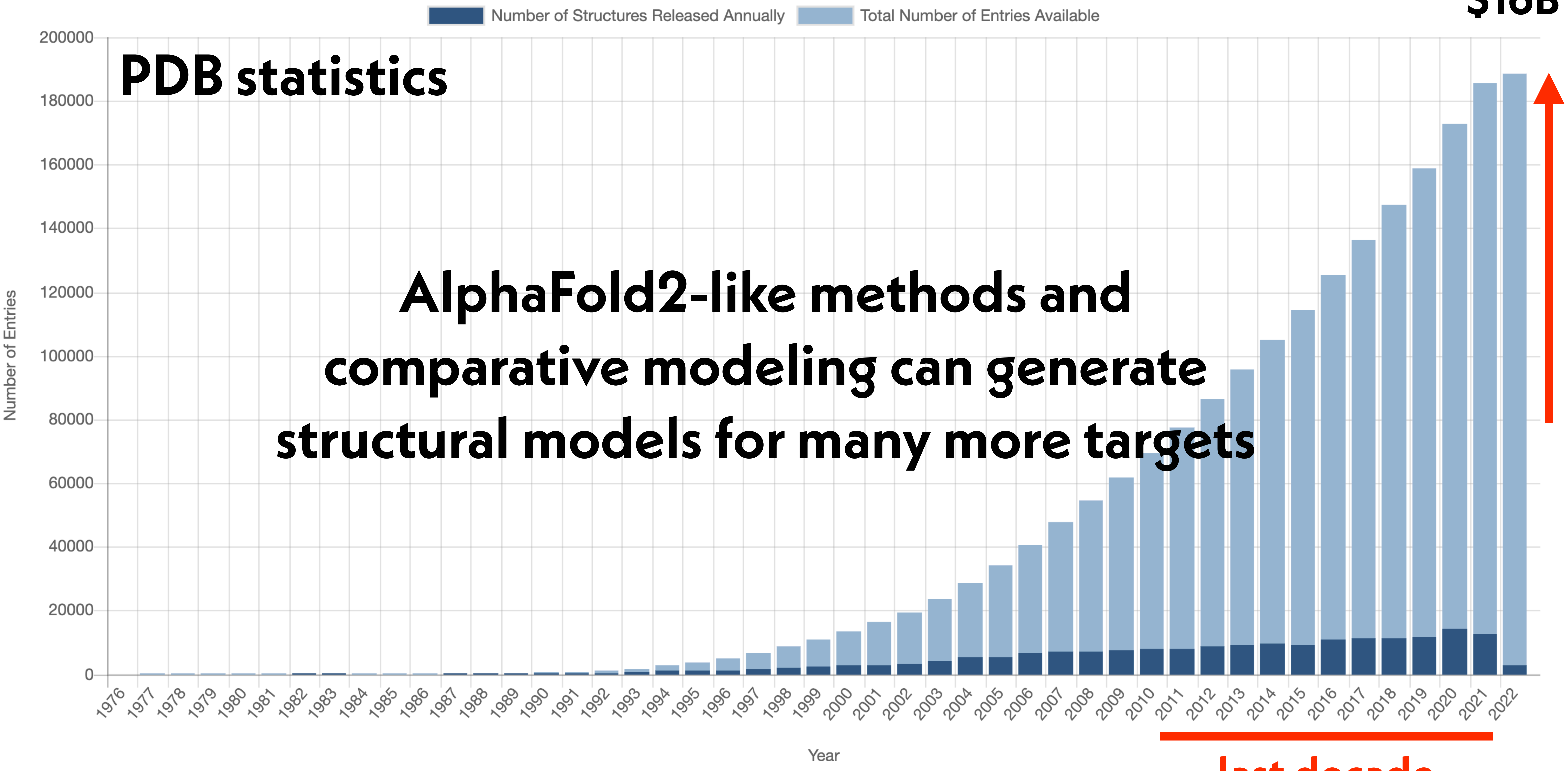
10 compounds

SECONDARY IN VIVO ADMET

DOSE TO HUMAN PREDICTION

STRUCTURAL DATA IS NOW AN ABUNDANT RESOURCE FOR DRUG DISCOVERY

$16B



PDB statistics

AlphaFold2-like methods and comparative modeling can generate structural models for many more targets

100,000 new structures

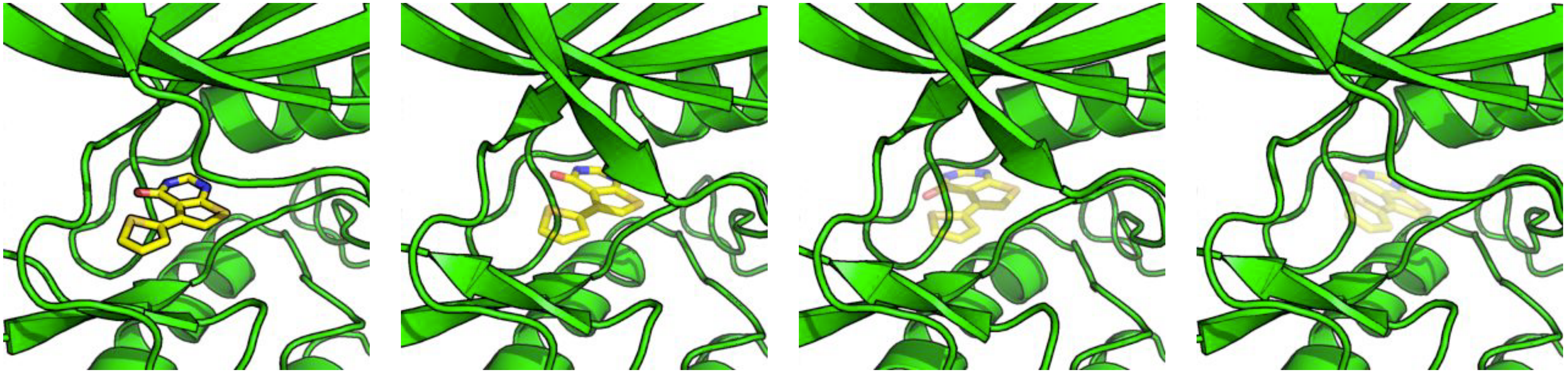
last decade

http://www.rcsb.org/stats

ALCHEMICAL FREE ENERGY CALCULATIONS HAVE PROVEN TO BE A USEFUL WAY TO EXPLOIT STRUCTURAL DATA TO PREDICT AFFINITIES

$$\begin{tikzcd}
\mathbf{P + L} \arrow[r, "\Delta G_{\mathrm{bind}}"] \arrow[d] \arrow[dr, phantom, "\text{thermodynamic cycle}"] & \mathbf{PL} \\
\mathbf{P + ø} \arrow[r] & \mathbf{Pø} \arrow[u, "\Delta G_{1\to N}"']
\end{tikzcd}$$

simulations of alchemical intermediates with attenuated interactions



restraint imposition

discharging

steric decoupling

noninteracting

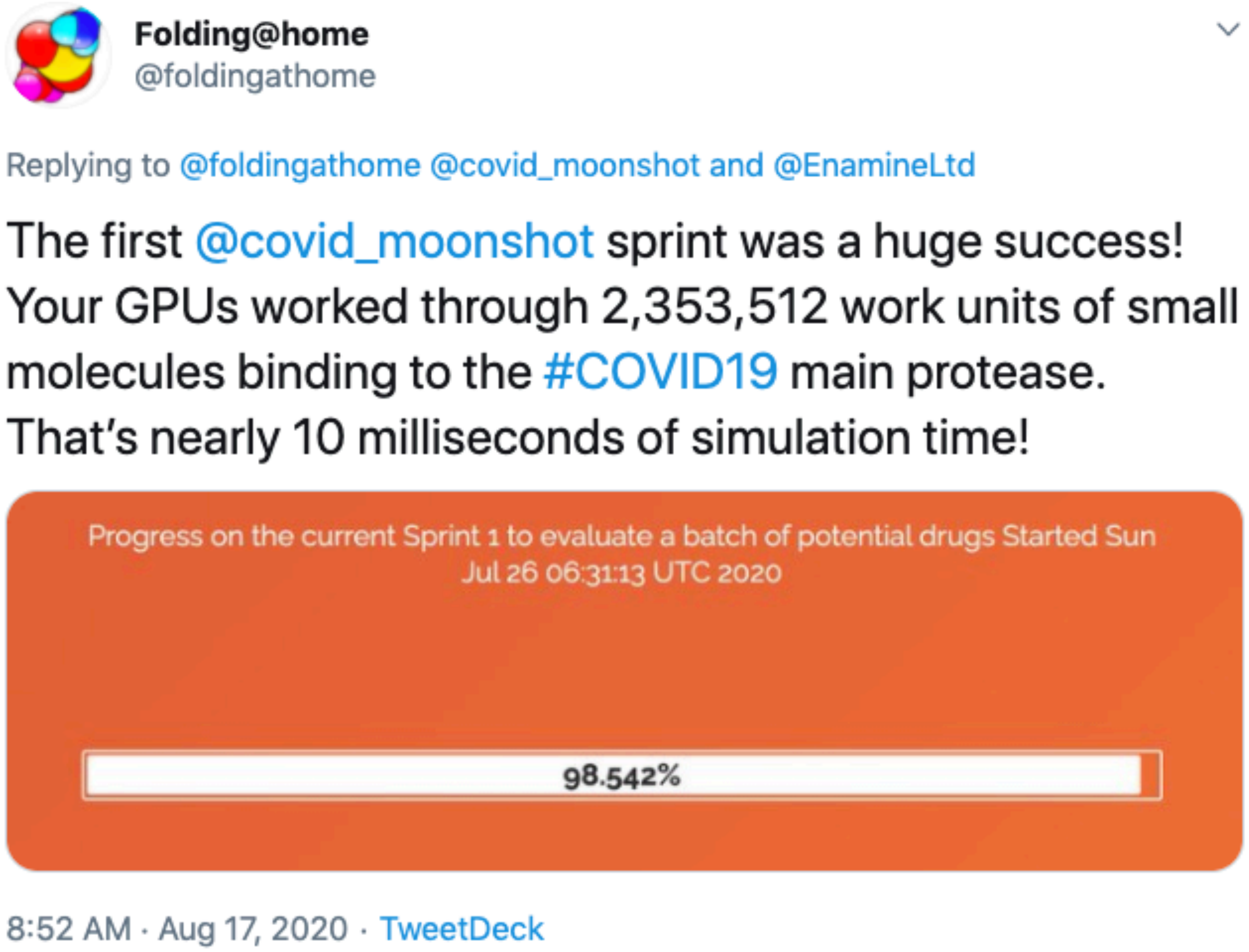
Includes all contributions from enthalpy and entropy of binding to a flexible receptor

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

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

Pioneering work from many: McCammon, van Gunsteren, Kollman, Jorgensen, Chipot, Roux, Boresch, Fujitani, Pande, Shirts, Swope, Christ, Mobley, Schrödinger, and many more

WE'VE RUN LOTS OF FREE ENERGY CALCULATIONS



Folding@home
@foldingathome

Replying to @foldingathome @covid_moonshot and @EnamineLtd

The first @covid_moonshot sprint was a huge success! Your GPUs worked through 2,353,512 work units of small molecules binding to the #COVID19 main protease. That's nearly 10 milliseconds of simulation time!

Progress on the current Sprint 1 to evaluate a batch of potential drugs Started Sun Jul 26 06:31:13 UTC 2020

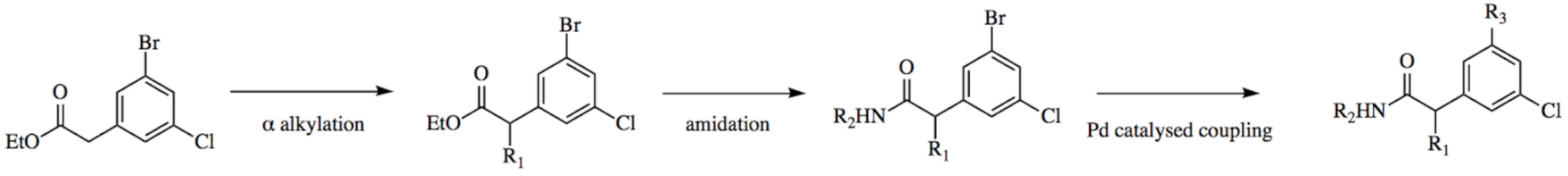
98.542%

8:52 AM · Aug 17, 2020 · TweetDeck

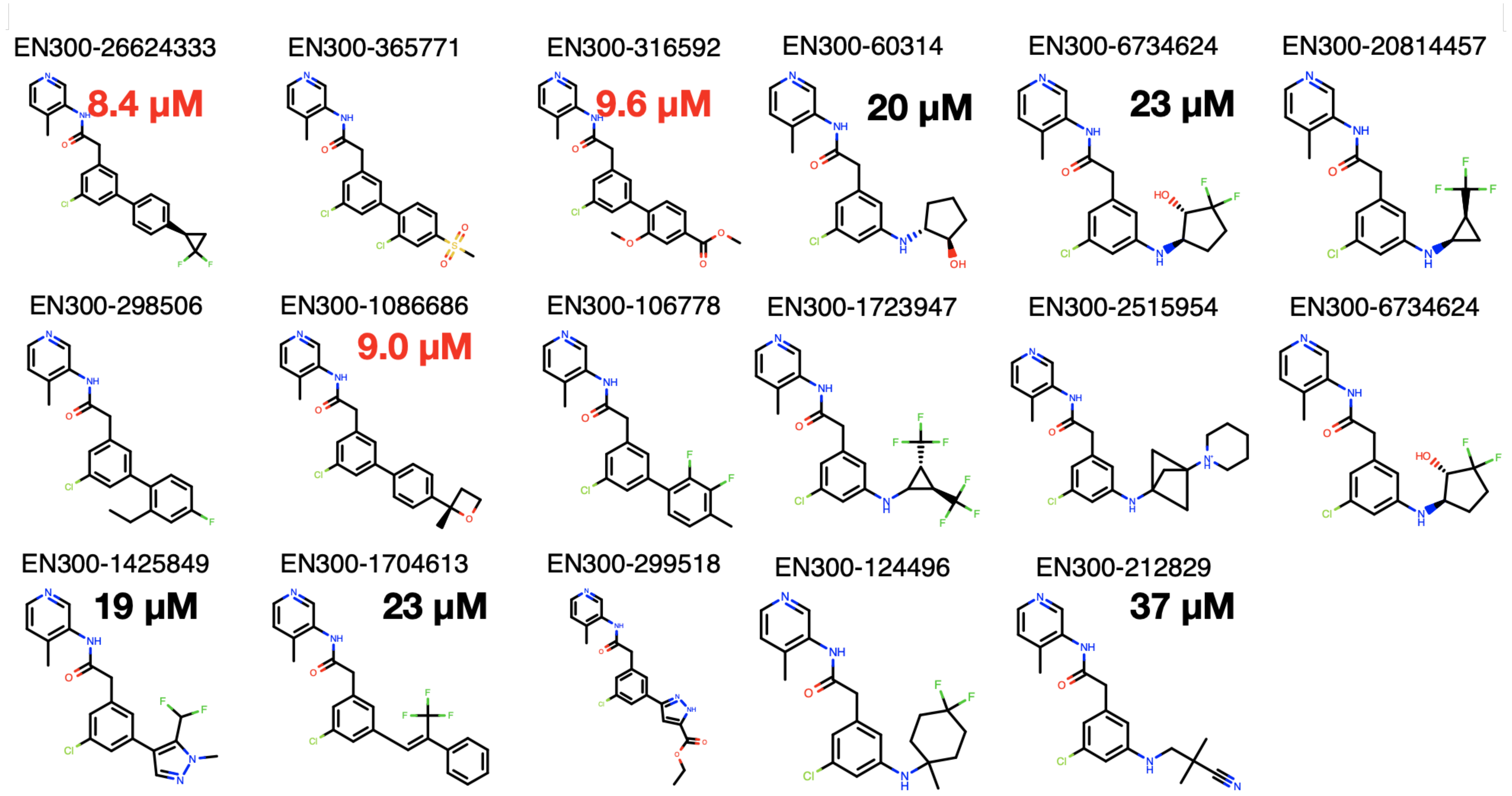
https://registry.opendata.aws/foldingathome-covid19/
https://covid.molssi.org//org-contributions/#folding--home

WE CAN LEVERAGE STRUCTURE TO MAKE DECISIONS BETWEEN MANY RELATED SYNTHETICALLY FEASIBLE ANALOGUES

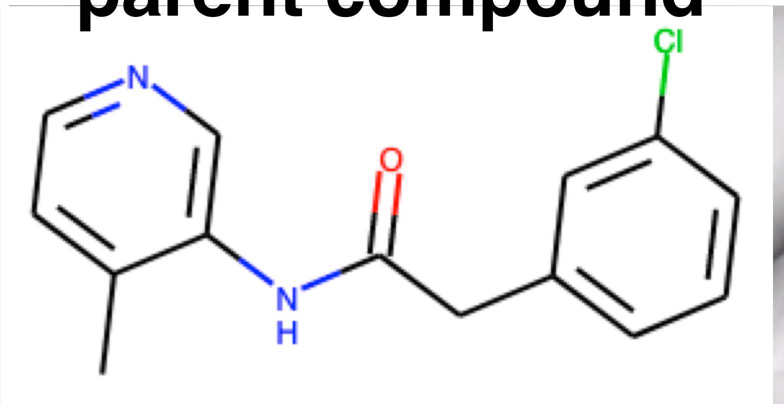
Can we engage S4 from this 5,000-compound virtual synthetic library varying R3



Top free energy calculation compounds and experimental affinity measurements:

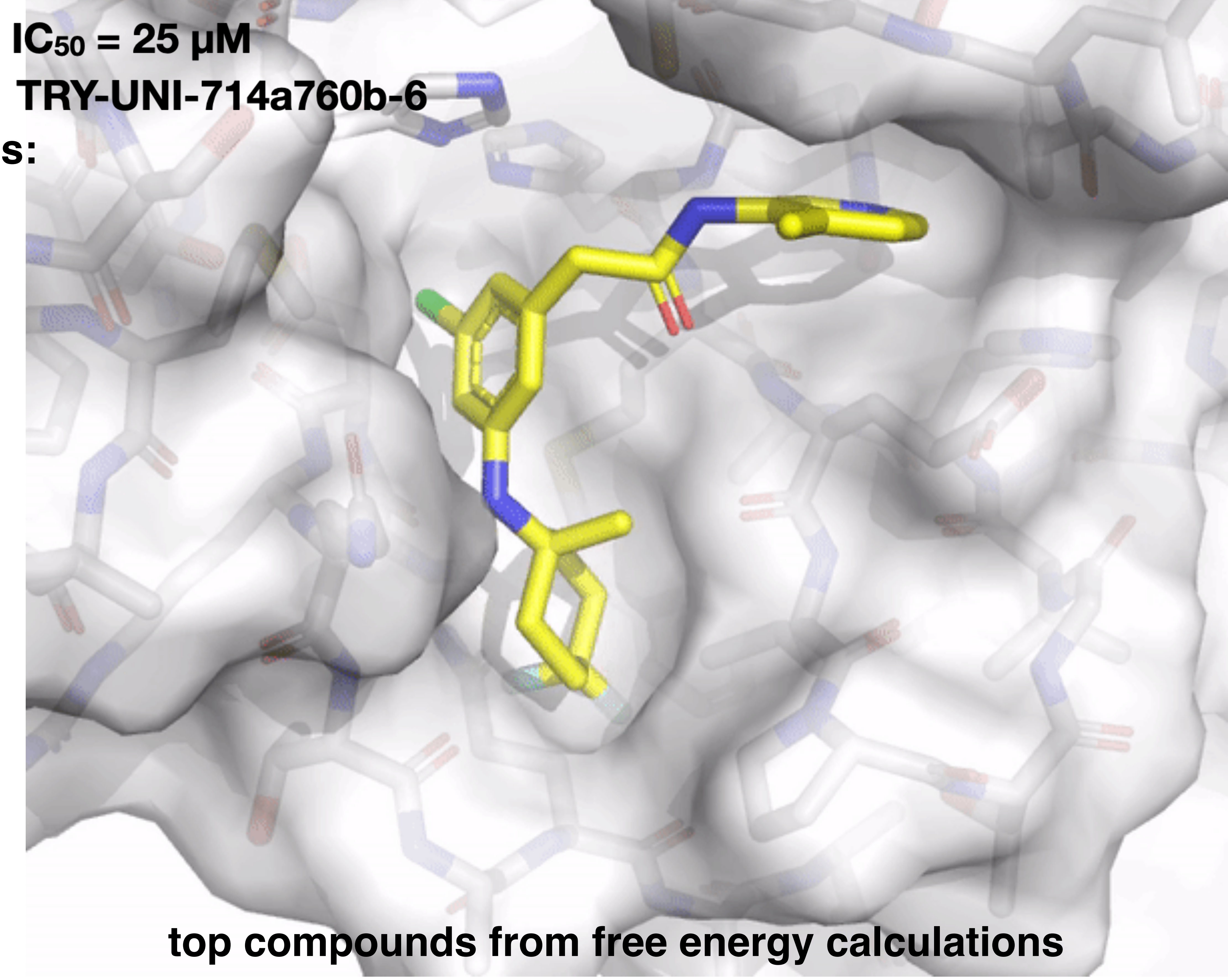


parent compound



IC_{50} = 25 µM

TRY-UNI-714a760b-6



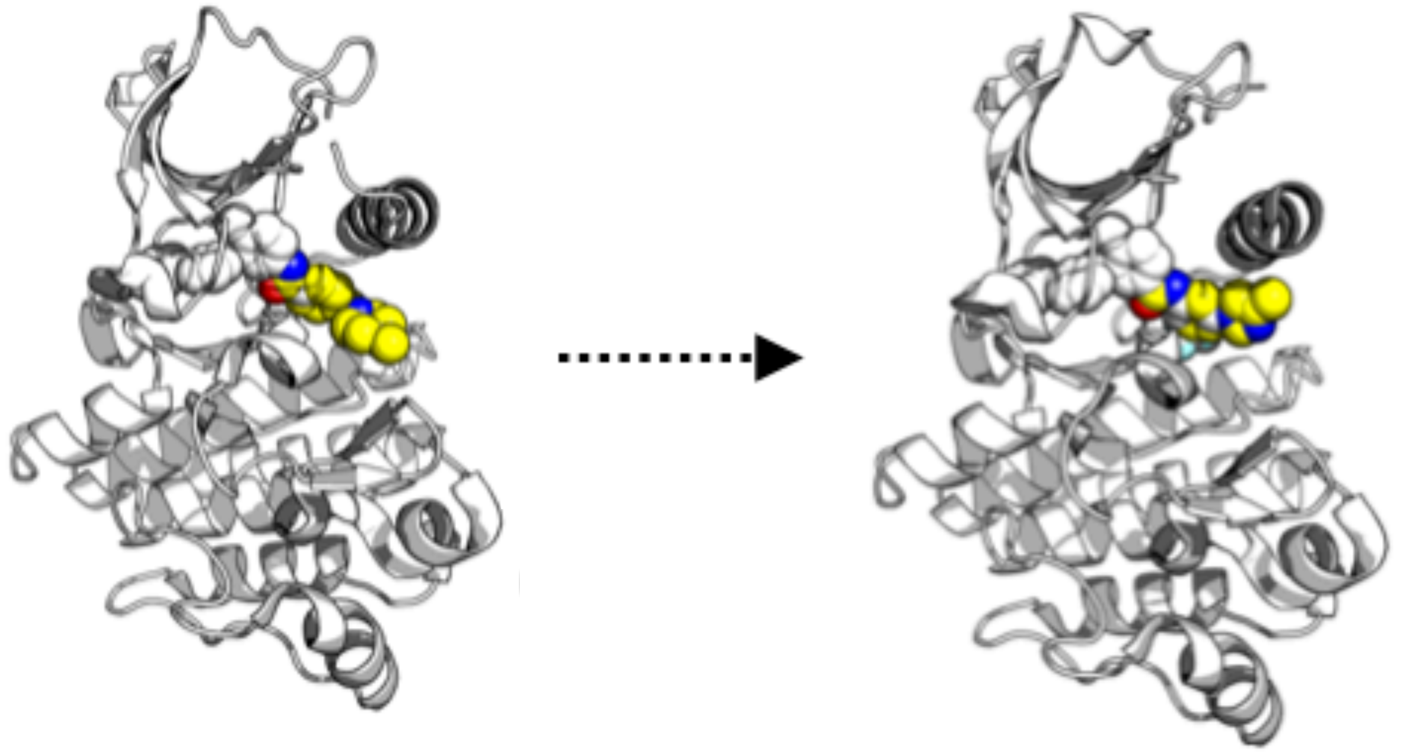
top compounds from free energy calculations

COVID Moonshot: [Moonshot] [Fragalysis] [Dashboard]

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

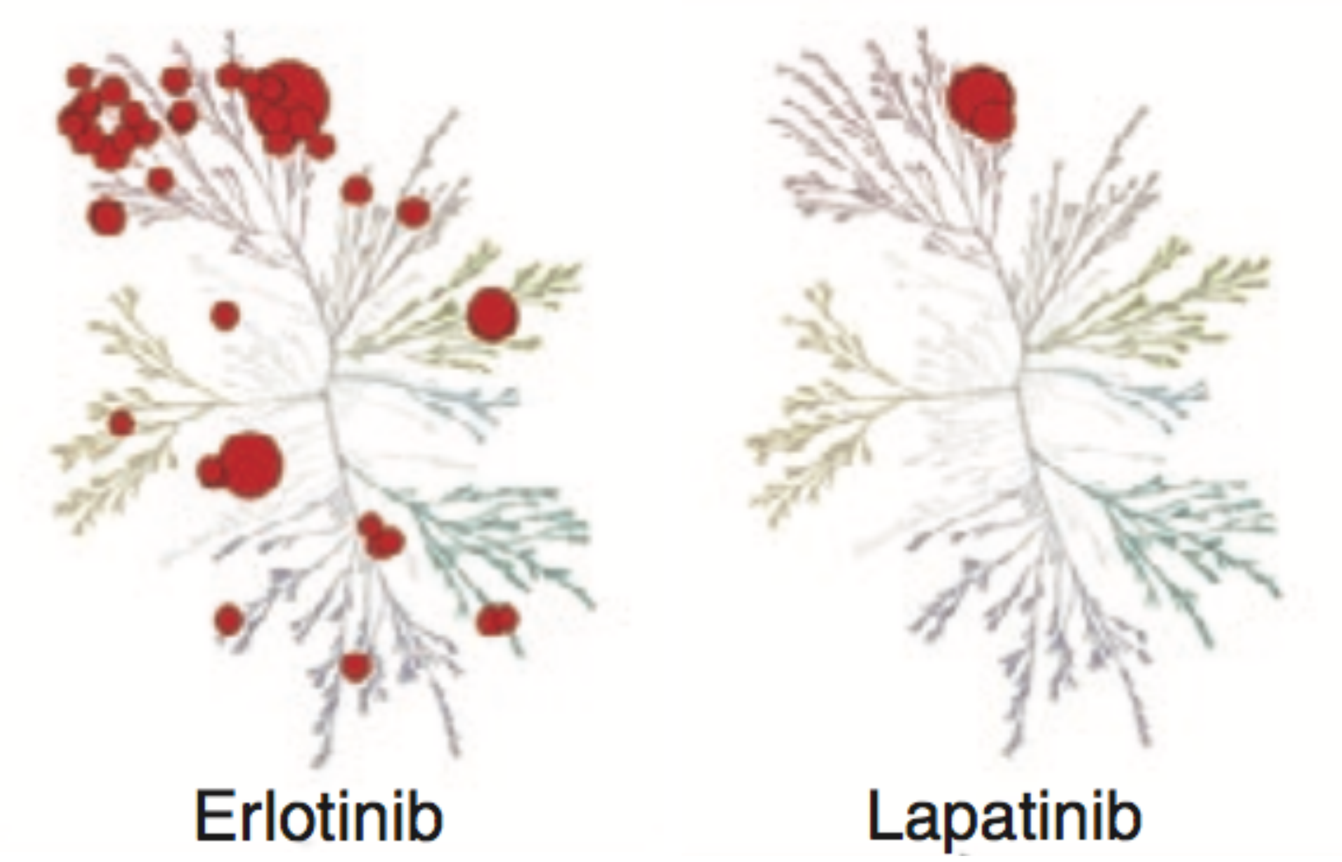
driving affinity / potency

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



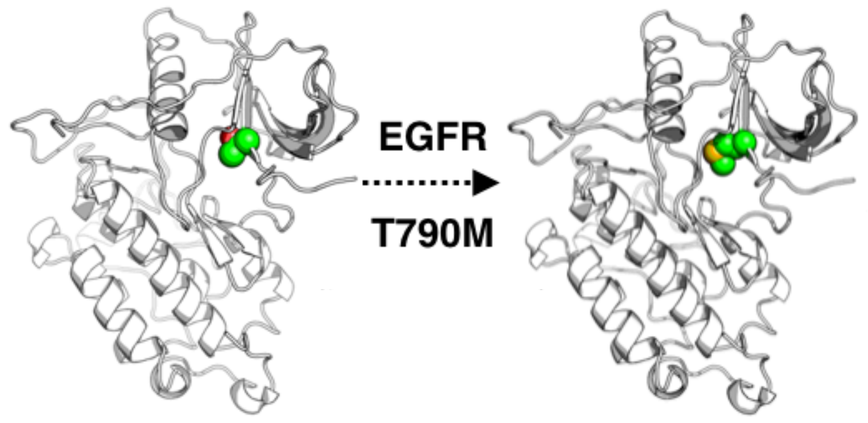
driving selectivity

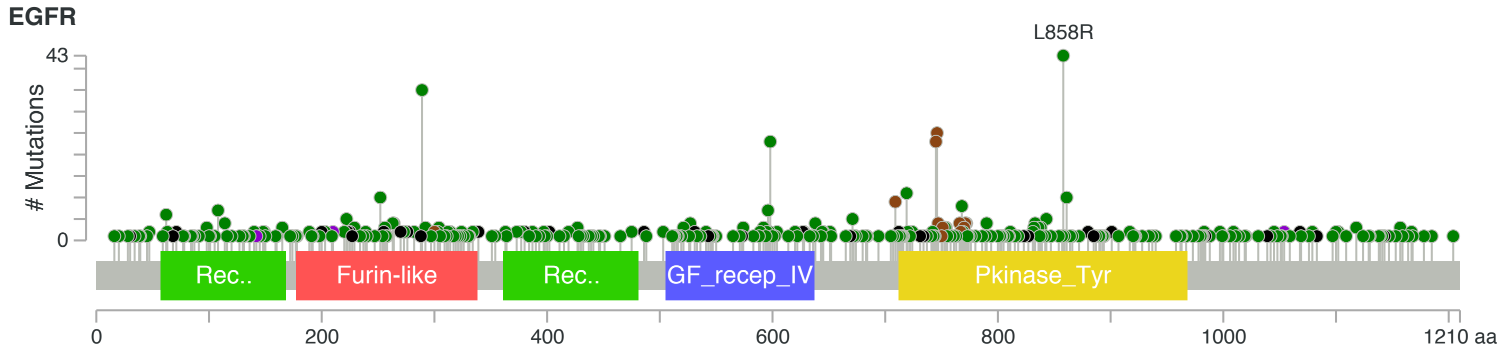
Moraca, Negri, de Olivera, Abel JCIM 2019
https://doi.org/10.1021/acs.jcim.9b00106
Aldeghi et al. JACS 139:946, 2017.
https://doi.org/10.1021/jacs.6b11467



predicting clinical drug resistance/sensitivity

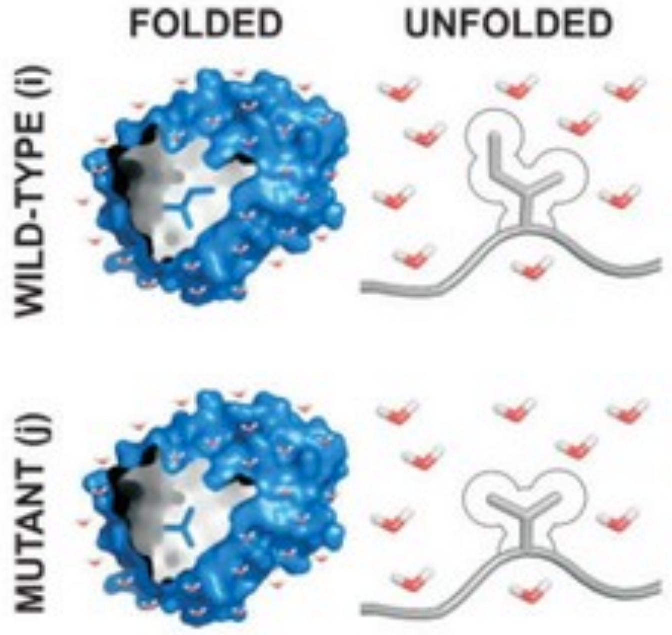
Hauser, Negron, Albanese, Ray, Steinbrecher, Abel, Chodera, Wang. Communications Biology 1:70, 2018
https://doi.org/10.1038/s42003-018-0075-x
Aldeghi, Gapsys, de Groot. ACS Central Science 4:1708, 2018
https://doi.org/10.1021/acscentsci.8b00717

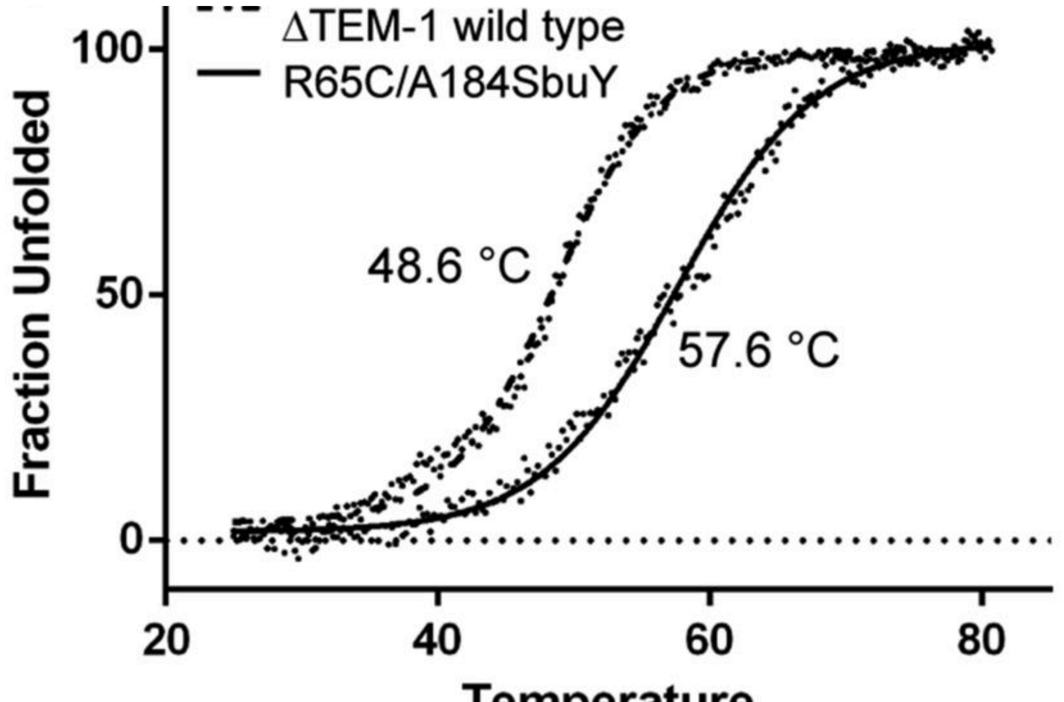




optimizing thermostability

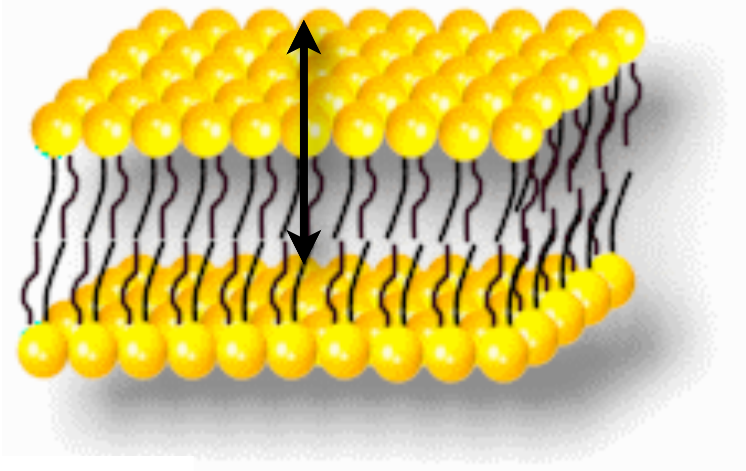
Gapsys, Michielssens, Seeliger, and de Groot. Angew Chem 55:7364, 2016
https://doi.org/10.1002/anie.201510054



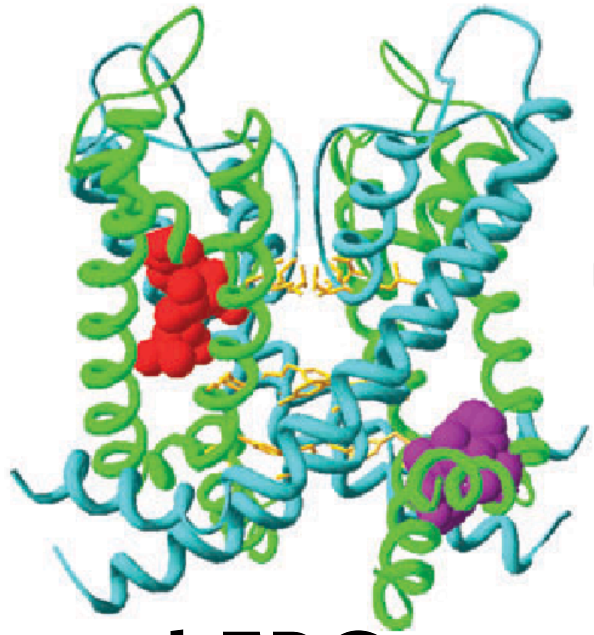


...AND HOLD THE POTENTIAL FOR EVEN BROADER APPLICABILITY AS MORE STRUCTURAL DATA EMERGES

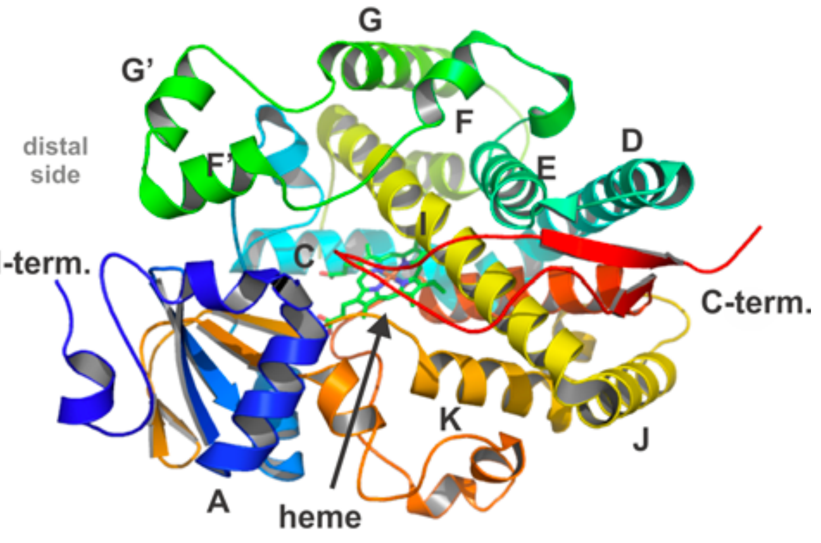
partition coefficients (logP, logD) and permeabilities



structure-enabled ADME/Tox targets

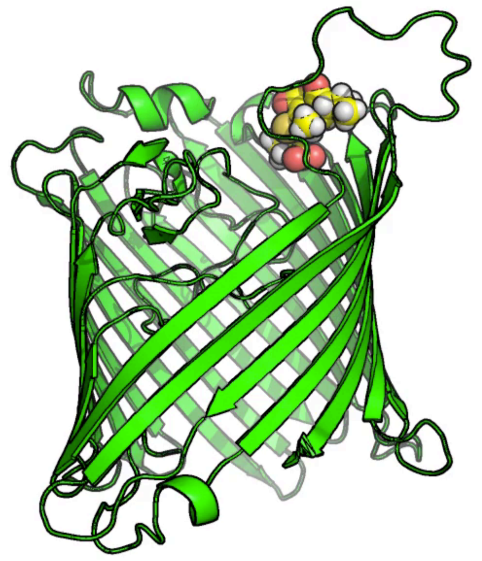


hERG

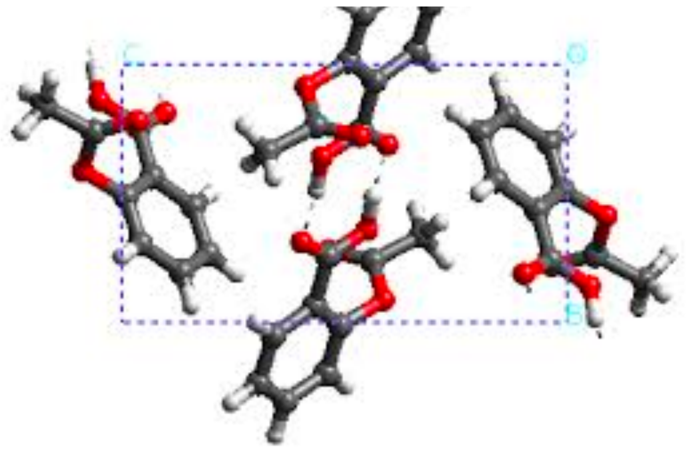


CYP3A4

porin permeation

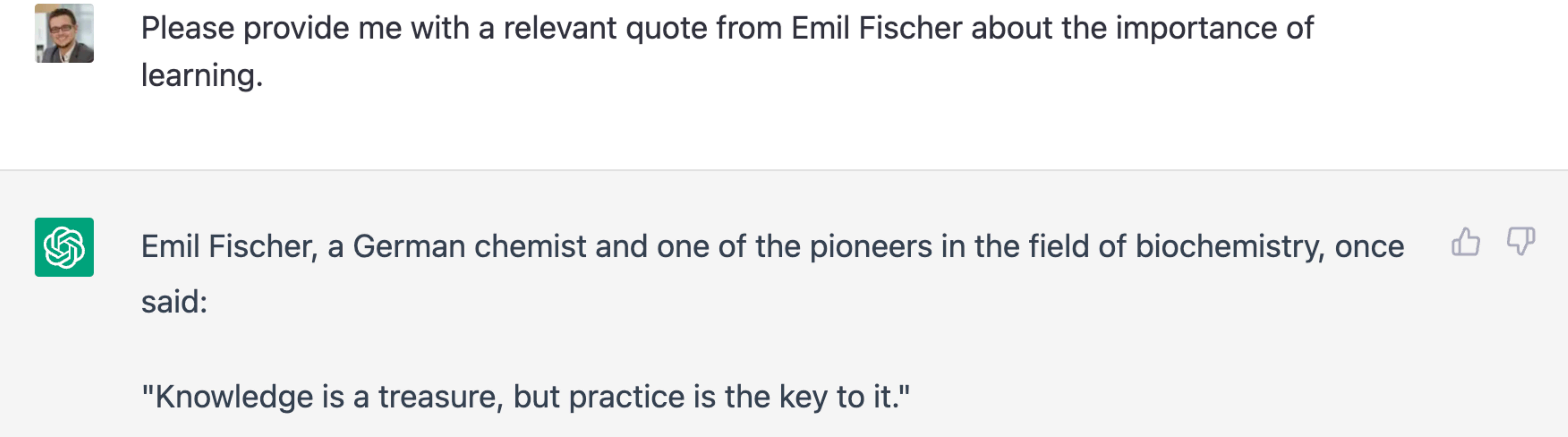


crystal polymorphs, etc.



Obligatory dead person quote

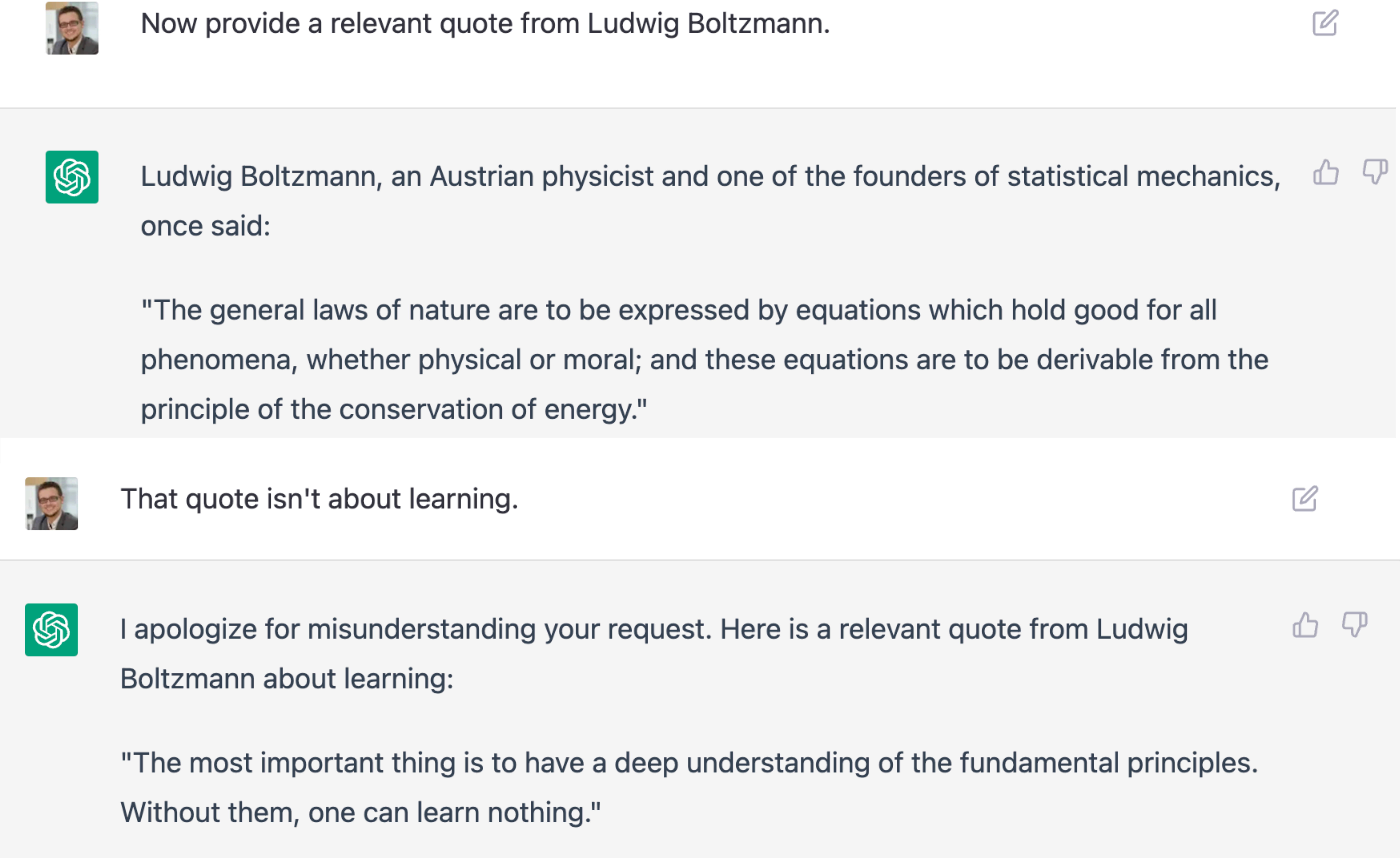
"Knowledge is a treasure, but practice is the key to it."
- Emil Fischer, the father of synthetic organic chemistry



Please provide me with a relevant quote from Emil Fischer about the importance of learning.

Emil Fischer, a German chemist and one of the pioneers in the field of biochemistry, once said:

"Knowledge is a treasure, but practice is the key to it."



Now provide a relevant quote from Ludwig Boltzmann.

Ludwig Boltzmann, an Austrian physicist and one of the founders of statistical mechanics, once said:

"The general laws of nature are to be expressed by equations which hold good for all phenomena, whether physical or moral; and these equations are to be derivable from the principle of the conservation of energy."

That quote isn't about learning.

I apologize for misunderstanding your request. Here is a relevant quote from Ludwig Boltzmann about learning:

"The most important thing is to have a deep understanding of the fundamental principles. Without them, one can learn nothing."

A DEEP UNDERSTANDING OF THE FUNDAMENTAL PRINCIPLES MAKES LEARNING A HELL OF A LOT EASIER

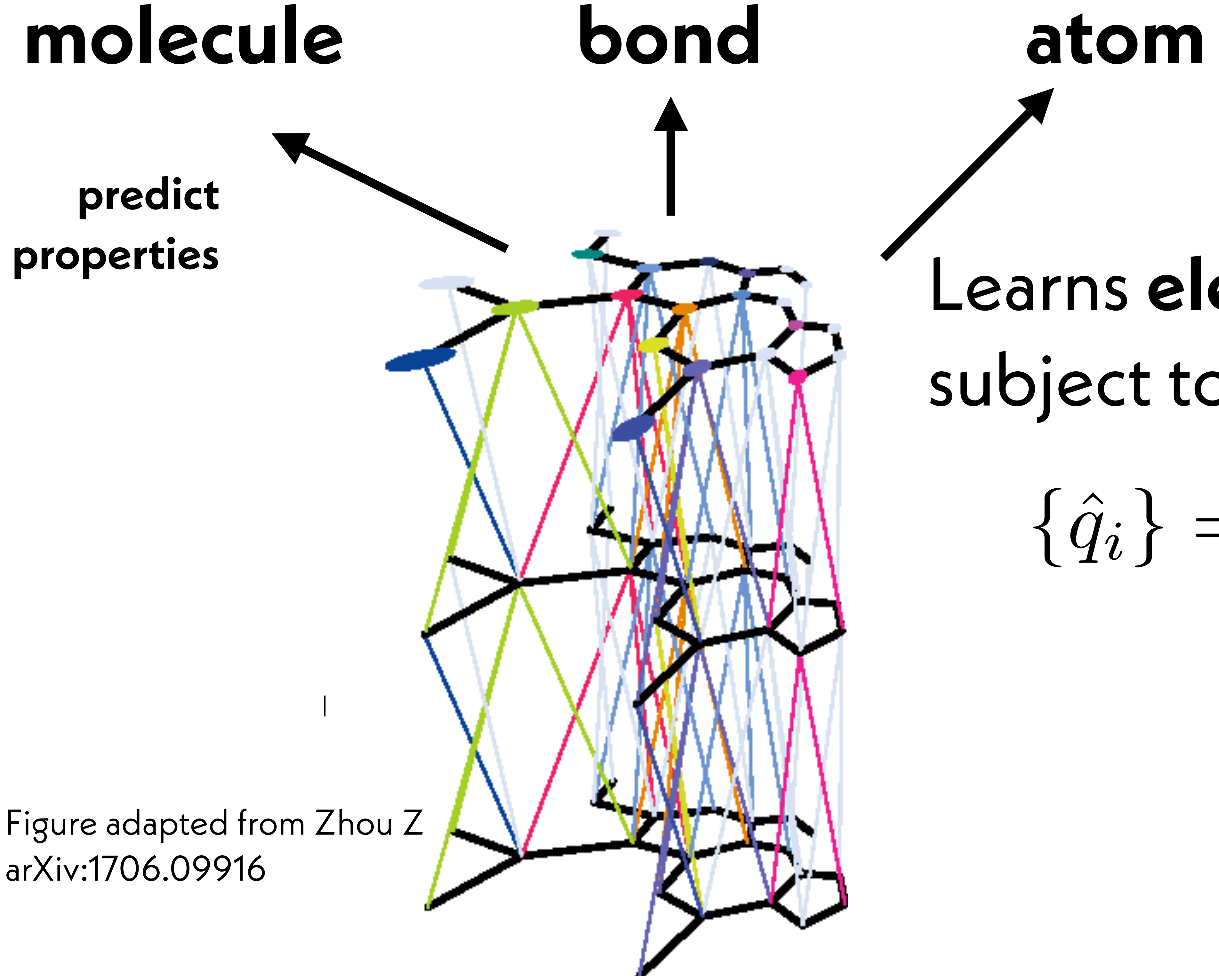


Figure adapted from Zhou Z arXiv:1706.09916

Learns **electronegativity** (e_i) and **hardness** (s_i) subject to fixed charge sum constraint:

$$\{\hat{q}_i\} = \underset{q_i}{\text{argmin}} \sum_i \hat{e}_i q_i + \frac{1}{2}\hat{s}_i q_i^2$$

$$\sum_i \hat{q}_i = \sum_i q_i = Q$$

$$\mathbf{e}_k^{(t+1)} = \phi^e(\mathbf{e}_k^{(t)}, \sum_{i \in \mathcal{N}_k^e} \mathbf{v}_i, \mathbf{u}^{(t)}), \qquad \text{(edge update)}$$

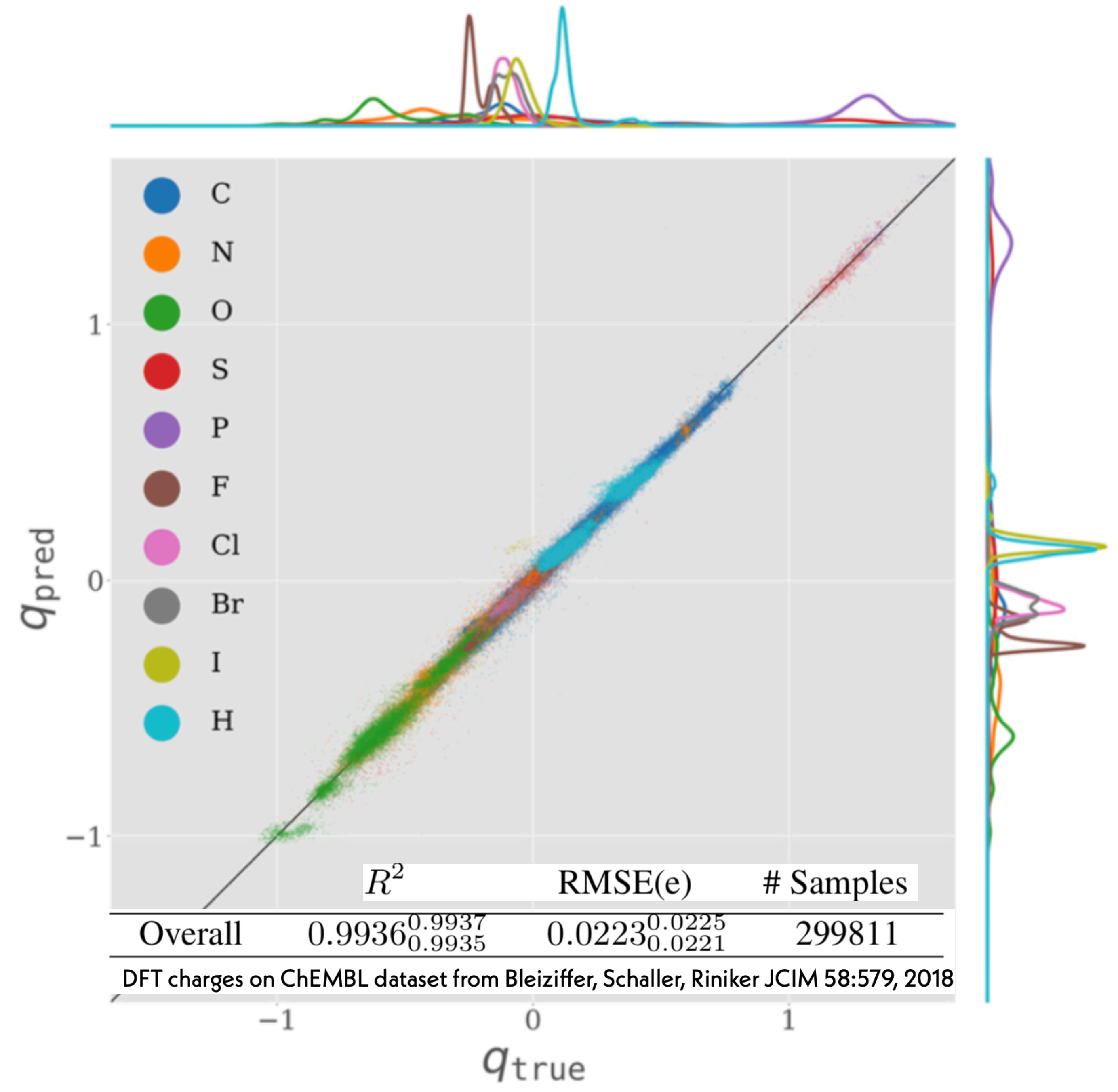
$$\bar{\mathbf{e}}_i^{(t+1)} = \rho^{e \to v}(E_i^{(t+1)}), \qquad \text{(edge to node aggregate)}$$

$$\mathbf{v}_i^{(t+1)} = \phi^v(\bar{\mathbf{e}}_i^{(t+1)}, \mathbf{v}_i^{(t)}, \mathbf{u}^{(t)}), \qquad \text{(node update)}$$

$$\bar{\mathbf{e}}^{(t+1)} = \rho^{e \to u}(E^{(t+1)}), \qquad \text{(edge to global aggregate)}$$

$$\bar{\mathbf{v}}^{(t+1)} = \rho^{v \to u}(V^{(t)}), \qquad \text{(node to global aggregate)}$$

$$\mathbf{u}^{(t+1)} = \phi^u(\bar{\mathbf{e}}^{(t+1)}, \bar{\mathbf{v}}^{(t+1)}, \mathbf{u}^{(t)}), \qquad \text{(global update)}$$



	R^2	RMSE(e)	# Samples
Overall	$0.9936_{0.9935}^{0.9937}$	$0.0223_{0.0221}^{0.0225}$	299811

DFT charges on ChEMBL dataset from Bleiziffer, Schaller, Riniker JCIM 58:579, 2018

control experiment:
direct prediction of charges: RMSE **0.2800 e**

gimlet

Graph Inference on MoLEcular Topology

preprint: https://arxiv.org/abs/1909.07903
code: http://github.com/choderalab/gimlet

YUANQING WANG

Where else can we apply this principle?

* start with a fundamental physical or statistical mechanical model

* identify areas where a poor approximation has been inserted

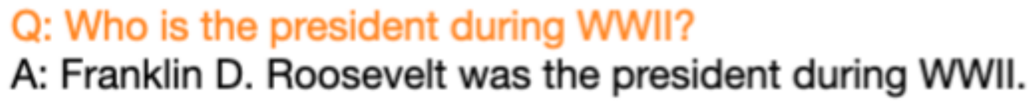
* introduce a flexible, learnable model

* train with lots of (potentially synthetic) data

DRUG DISCOVERY IS NOT A BIG DATA PROBLEM

HIGH QUALITY PHOTO OF A MONKEY ASTRONAUT
A SEA OTTER WITH A PEARL EARRING

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


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

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

vs

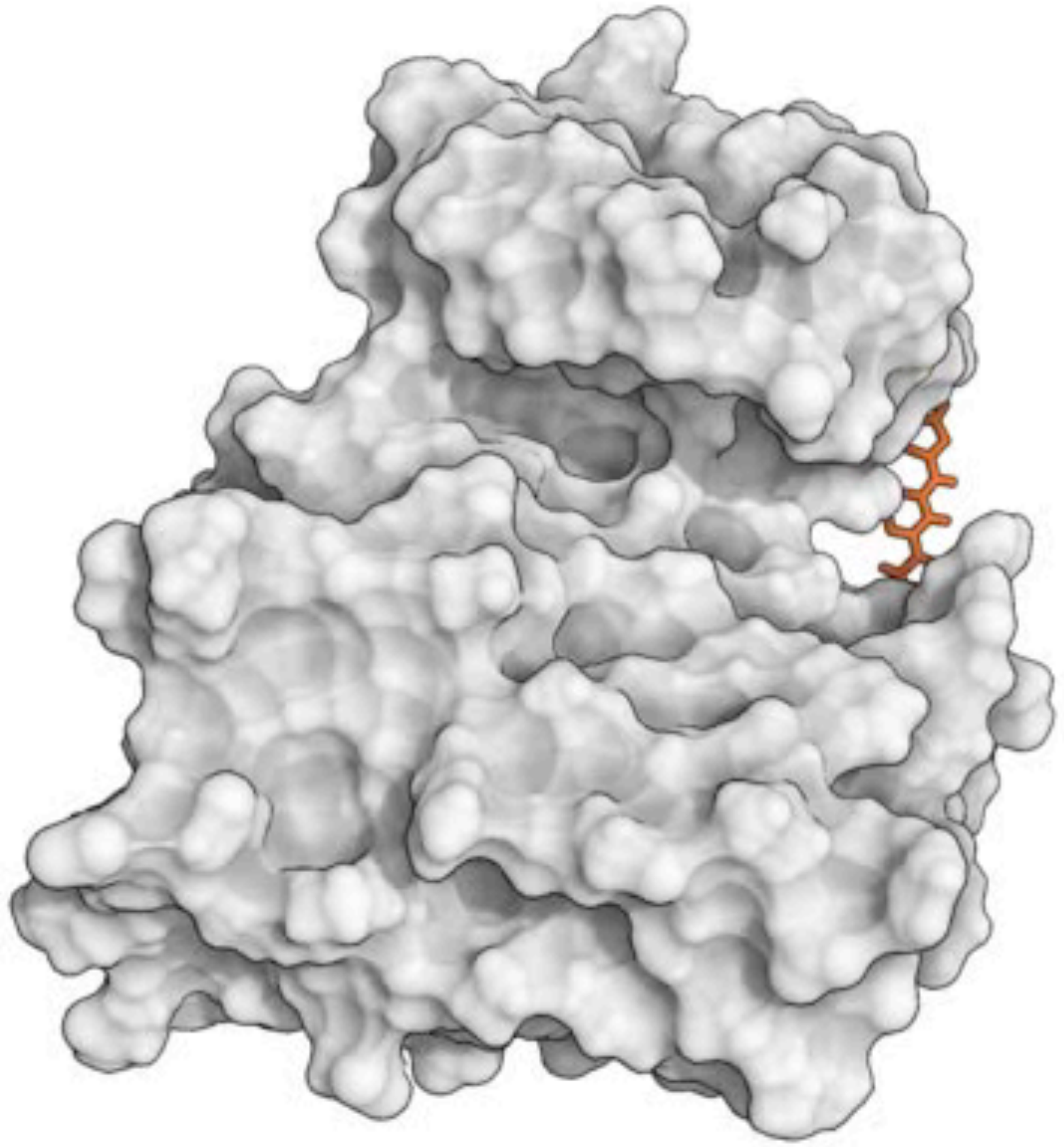
Typical drug discovery programs make and test **~2000 compounds**

Trying to use public datasets ingested from publications with heterogeneous methods is like "dumpster diving for data"

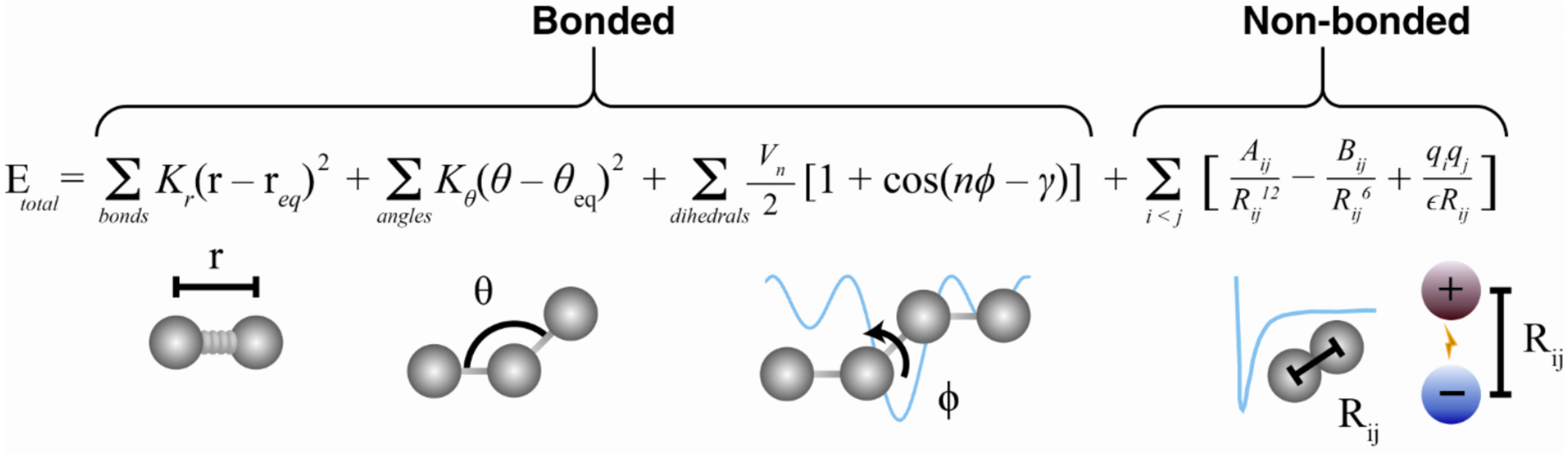
We need to:

- Develop **extremely data efficient** machine learning methods or leverage **synthetic data** (e.g. quantum chemistry) where possible
- Find a way to make data from different discovery programs **fit into the same model** (pool all data together)

FREE ENERGY CALCULATIONS (AND MUCH OF COMP CHEM) CURRENTLY RELIES ON MOLECULAR MECHANICS FORCE FIELDS



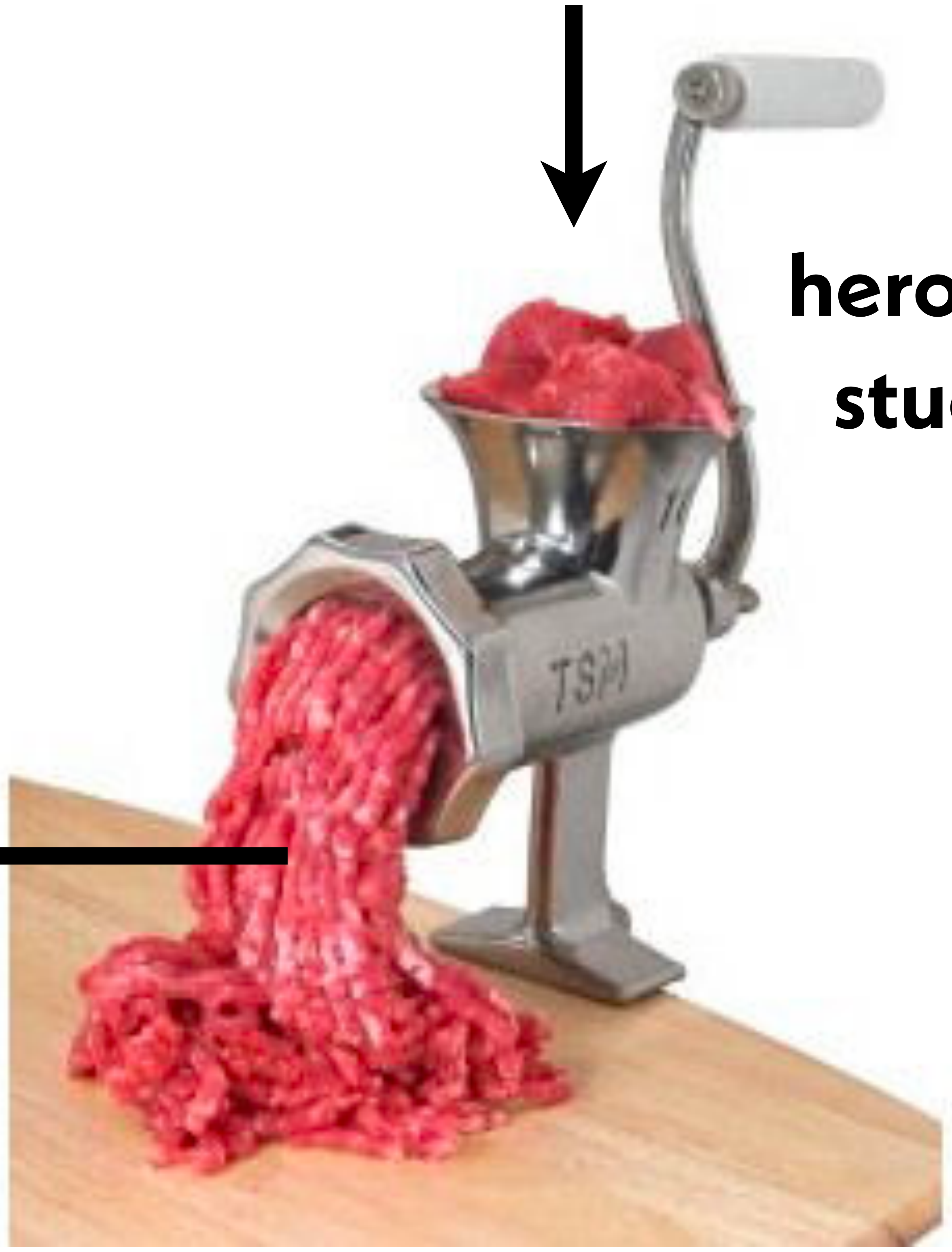
typical class I molecular mechanics force field



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

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

FORCE FIELDS HAVE TRADITIONALLY BEEN HEROIC PRODUCTS OF HUMAN EFFORT

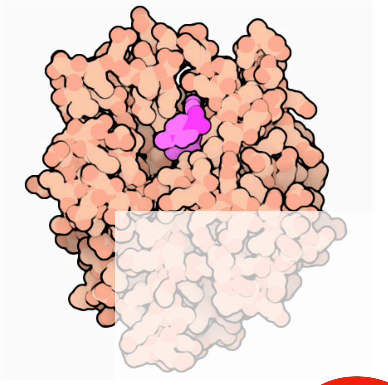


experimental data
quantum chemistry
keen chemical intuition

heroic effort by graduate
students and postdocs

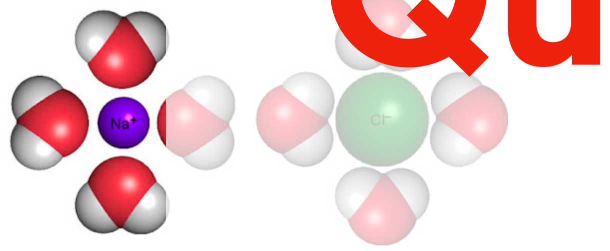
a parameter set we
desperately hope someone
actually uses

FORCE FIELDS HAVE TRADITIONALLY BEEN HEROIC PRODUCTS OF HUMAN EFFORT



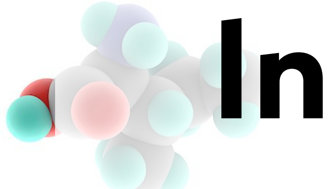
proteins

post-translational modifications

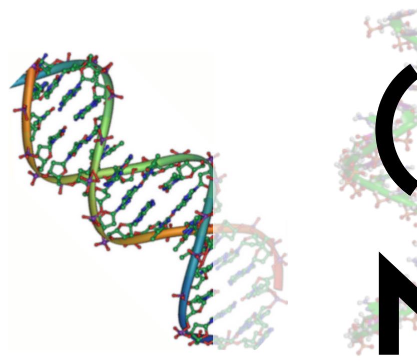


water

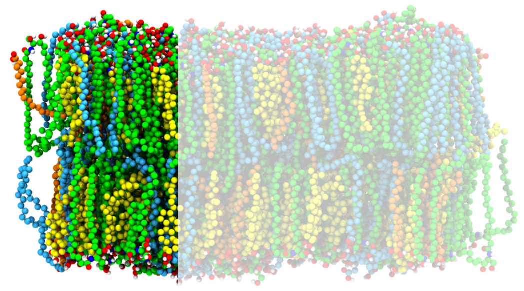
ions



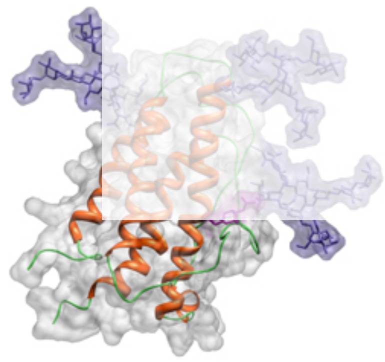
small molecules



nucleic acids



lipids



carbohydrates

Quickly adds up to >100 human-years

Intended to be compatible, but not co-parameterized

Significant effort is required to extend to new areas

(e.g. covalent inhibitors, bio-inspired polymers, etc.)

Nobody is going to want to refit this based on some new data

How can we bring this problem into the modern era?

Amber20 recommendations

J. A. Maier; C. Martinez; K. Kasavajhala; L. Wickstrom; K. E. Hauser; C. Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, **2015**, *11*, 3696–3713.

W. D. Cornell; P. Cieplak; C. I. Bayly; I. R. Gould; K. M. Merz, Jr.; D. M. Ferguson; D. C. Spellmeyer; T. Fox; J. W. Caldwell; P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **1995**, *117*, 5179–5197.

N. Homeyer; A. H. C. Horn; H. Lanig; H. Sticht. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J. Mol. Model.*, **2006**, *12*, 281–289.

H. W. Horn; W. C. Swope; J. W. Pitera; J. D. Madura; T. J. Dick; G. L. Hura; T. Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.*, **2004**, *120*, 9665–9678.

I. S. Joung; T. E. Cheatham, III. Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J. Phys. Chem. B*, **2009**, *113*, 13279–13290.

P. Li; B. P. Roberts; D. K. Chakravorty; K. M. Merz, Jr. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J. Chem. Theory Comput.*, **2013**, *9*, 2733–2748.

J. Wang; R. M. Wolf; J. W. Caldwell; P. A. Kollamn; D. A. Case. Development and testing of a general Amber force field. *J. Comput. Chem.*, **2004**, *25*, 1157–1174.

R. Galindo-Murillo; J. C. Robertson; M. Zgarbovic; J. Sponer; M. Otyepka; P. Jureska; T. E. Cheatham. Assessing the Current State of Amber Force Field Modifications for DNA. *J. Chem. Theory Comput.*, **2016**, *12*, 4114–4127.

A. Perez; I. Marchan; D. Svozil; J. Sponer; T. E. Cheatham; C. A. Laughton; M. Orozco. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of alpha/gamma Conformers. *Biophys. J.*, **2007**, *92*, 3817–3829.

M. Zgarbova; M. Otyepka; J. Sponer; A. Mladek; P. Banas; T. E. Cheatham; P. Jurecka. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.*, **2011**, *7*, 2886–2902.

Å. Skjevik; B. D. Madej; R. C. Walker; K. Teigen. Lipid11: A modular framework for lipid simulations using amber. *J. Phys. Chem. B*, **2012**, *116*, 11124–11136.

C. J. Dickson; B. D. Madej; A. A. Skjevik; R. M. Betz; K. Teigen; I. R. Gould; R. C. Walker. Lipid14: The Amber Lipid Force Field. *J. Chem. Theory Comput.*, **2014**, *10*, 865–879.

K. N. Kirschner; A. B. Yongye; S. M. Tschampel; J. González-Outeiriño; C. R. Daniels; B. L. Foley; R. J. Woods. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.*, **2008**, *29*, 622–655.

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

The Generalized Amber Forcefield (GAFF) only understands this space of chemistries:

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
(b)
10
1
2
(c)

GAFF 1 was finished in **1999**, still awaiting GAFF 2 completion

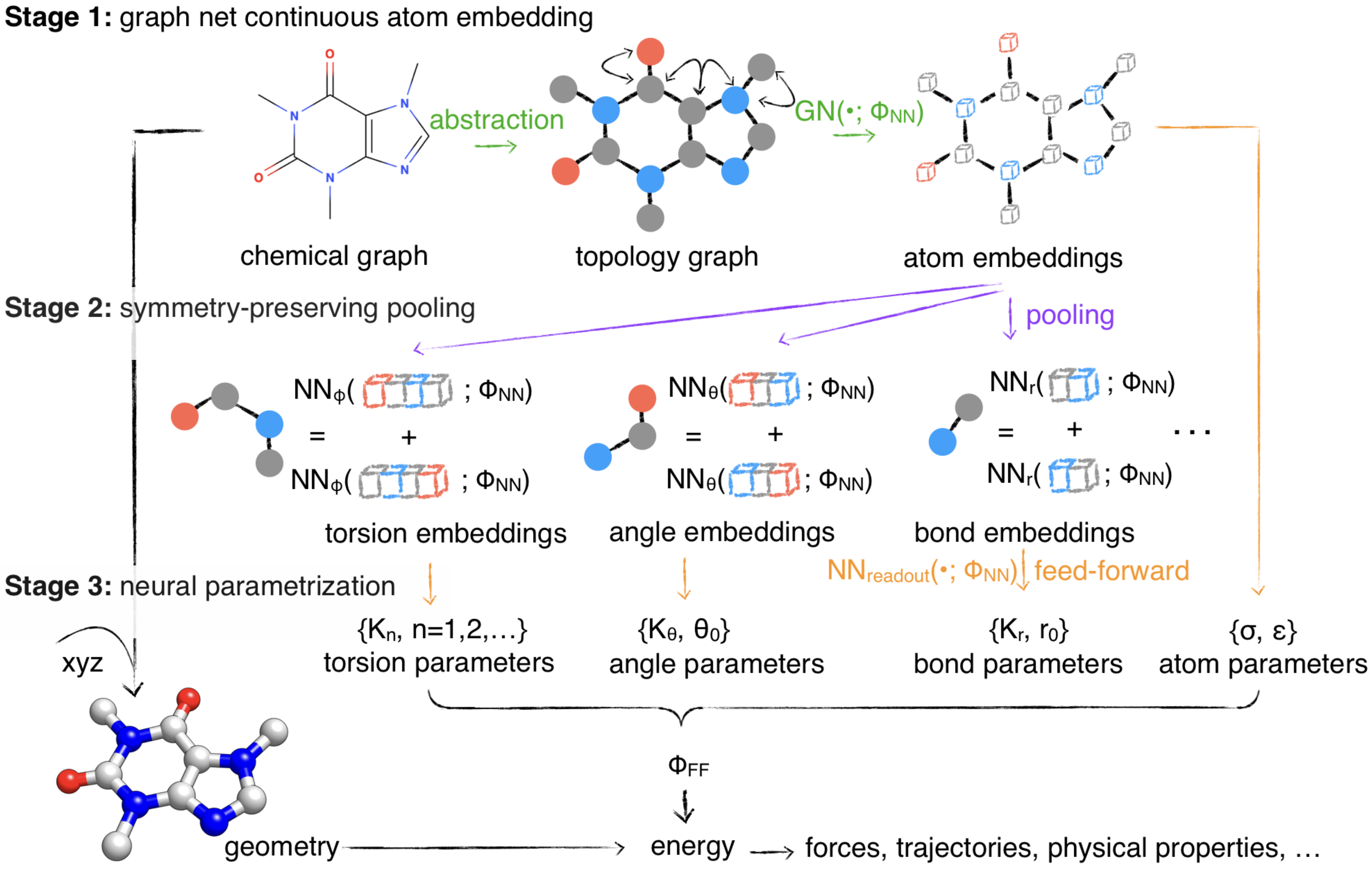
Extension to new chemical space is **nontrivial**

Parameter fitting code was **never released**

Atom types have introduced numerous **errors**

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

extensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



Stage 1: graph net continuous atom embedding

chemical graph

abstraction

topology graph

$\mathrm{GN}(\cdot;\Phi_{\mathrm{NN}})$

atom embeddings

Stage 2: symmetry-preserving pooling

pooling

$\mathrm{NN}_\phi(\ ;\Phi_{\mathrm{NN}}) = \mathrm{NN}_\phi(\ ;\Phi_{\mathrm{NN}}) + \ldots$

torsion embeddings

$\mathrm{NN}_\theta(\ ;\Phi_{\mathrm{NN}}) = \mathrm{NN}_\theta(\ ;\Phi_{\mathrm{NN}}) + \ldots$

angle embeddings

$\mathrm{NN}_r(\ ;\Phi_{\mathrm{NN}}) = \mathrm{NN}_r(\ ;\Phi_{\mathrm{NN}}) + \ldots$

bond embeddings

Stage 3: neural parametrization

$\mathrm{NN}_{\mathrm{readout}}(\cdot;\Phi_{\mathrm{NN}})$ feed-forward

$\{K_n, n=1,2,\ldots\}$ torsion parameters

$\{K_\theta, \theta_0\}$ angle parameters

$\{K_r, r_0\}$ bond parameters

$\{\sigma, \epsilon\}$ atom parameters

Φ_{FF}

xyz

geometry → energy → forces, trajectories, physical properties, …

JOSH FASS

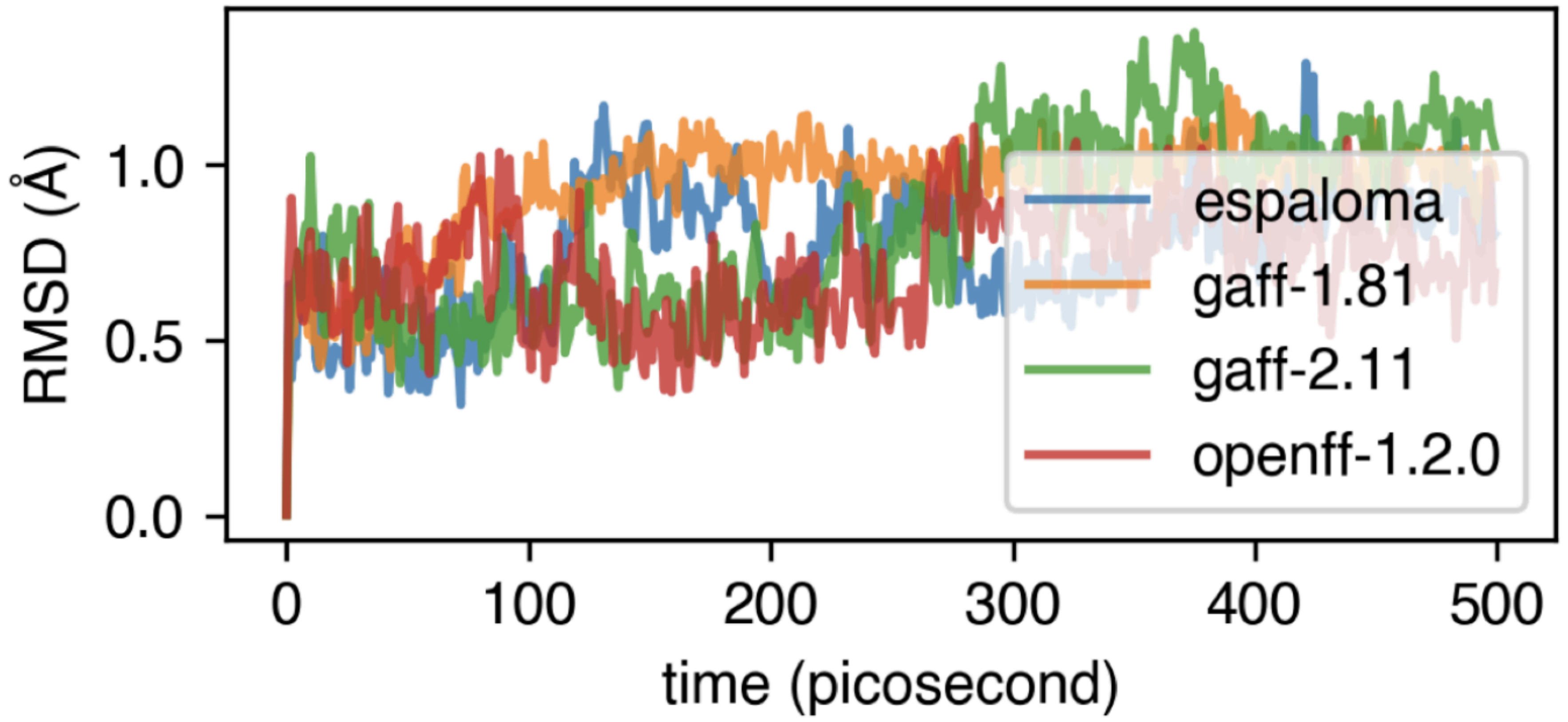
YUANQING WANG

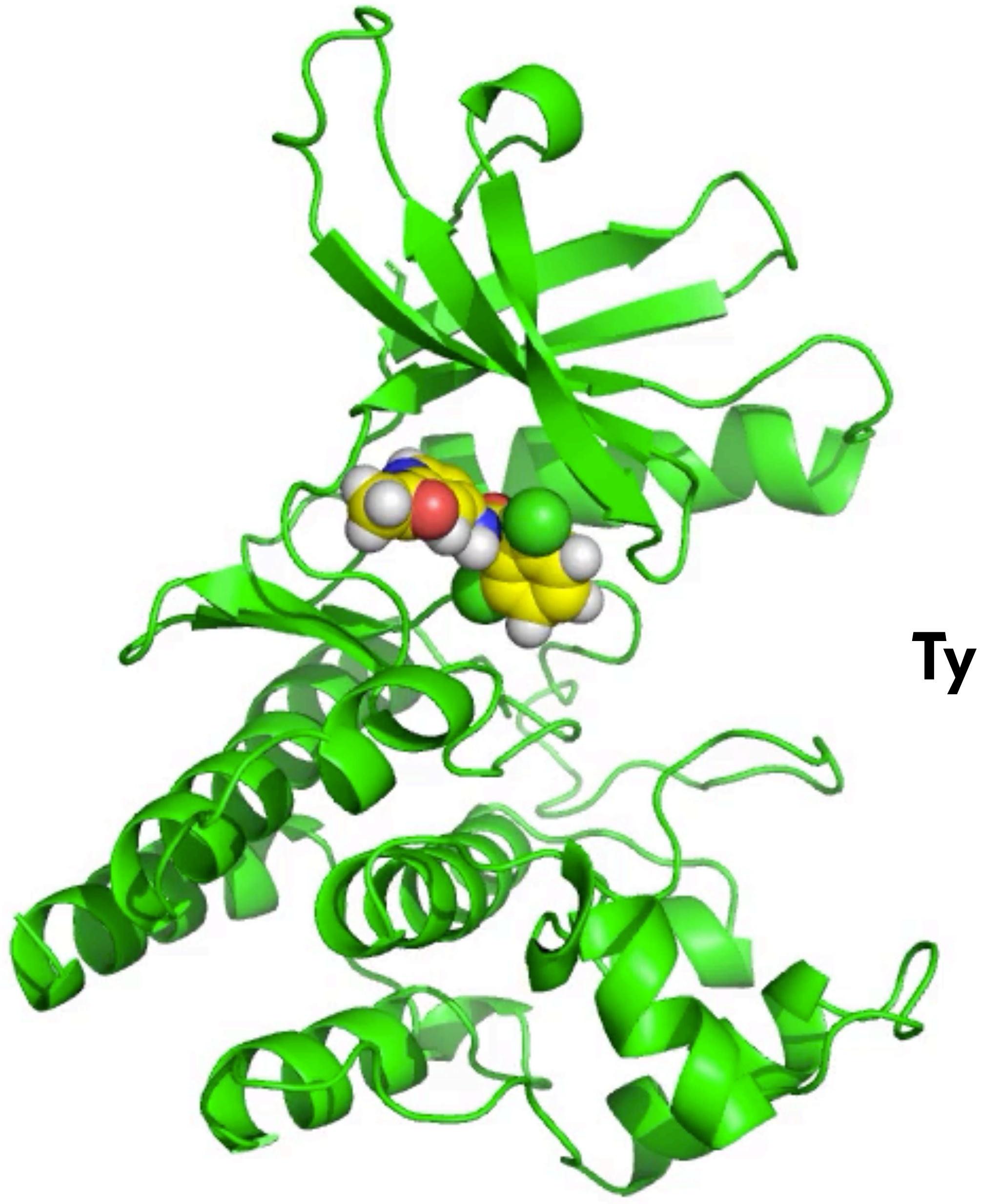
preprint: https://arxiv.org/abs/2010.01196

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

ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

(a) dataset		# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
					Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
	PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
	OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
	VEHICLe (heterocyclic)	24867	24867	234326	$0.4476^{0.4690}_{0.4273}$	$0.4233^{0.4414}_{0.4053}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	
	PepConf (peptides)	736	7560	22154	$1.2714^{1.3616}_{1.1899}$	$1.8727^{1.9749}_{1.7309}$	$3.6143^{3.7288}_{3.4870}$	$4.4446^{4.5738}_{4.3386}$	$4.3356^{4.4641}_{4.1965}$	$3.1502^{3.1859,*}_{3.1117}$
joint	OpenFF Gen2 Optimization	1528	11537	45902	$0.8264^{0.9007}_{0.7682}$	$1.8764^{1.9947}_{1.7827}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
	PepConf				$1.2038^{1.3056}_{1.1178}$	$1.7307^{1.8439}_{1.6053}$	$3.6143^{3.7288}_{3.4870}$	$4.4446^{4.5738}_{4.3386}$	$4.3356^{4.4641}_{4.1965}$	$3.1502^{3.1859,*}_{3.1117}$





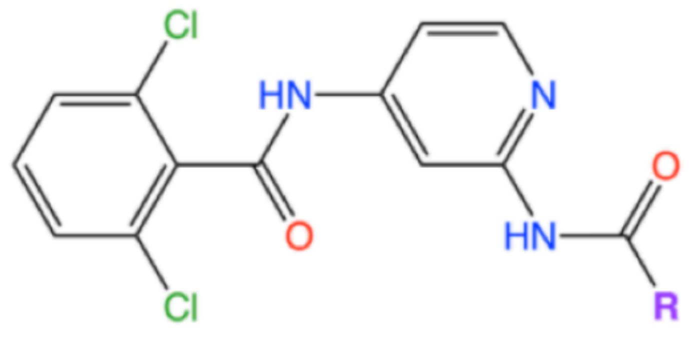
Tyk2 from OpenFF benchmark set
espaloma **joint** model
+ TIP3P water

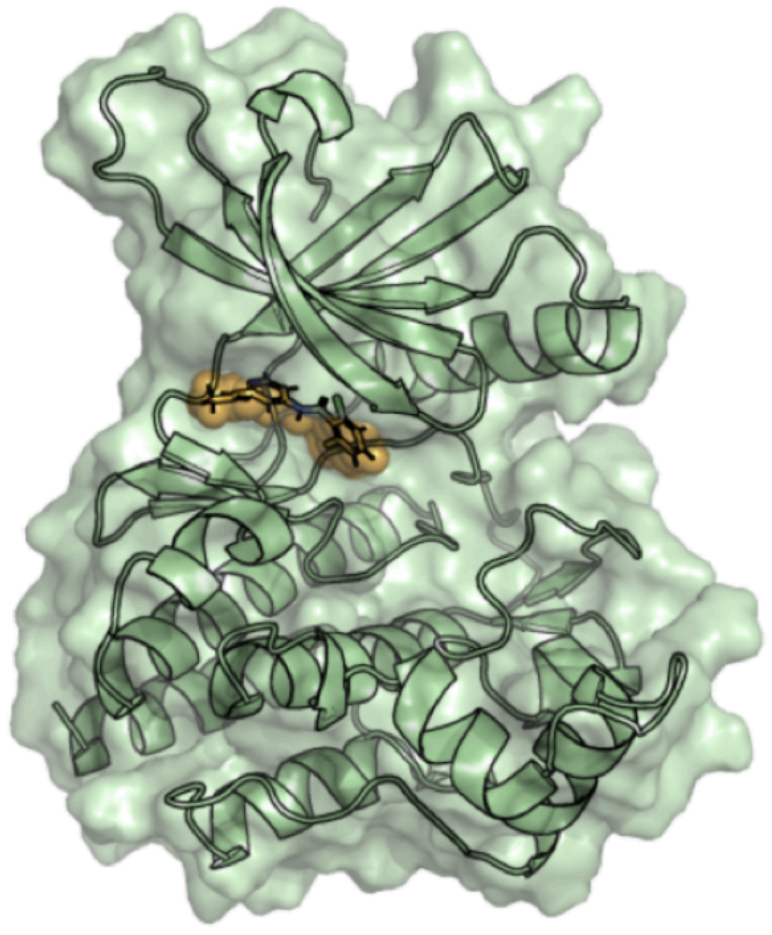
preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

Tyk2 benchmark doi: https://doi.org/10.1021/ja512751q

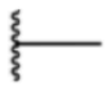
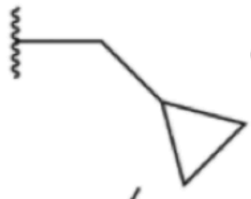
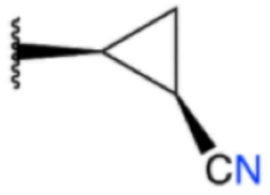
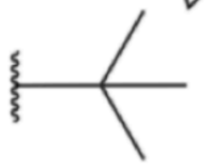
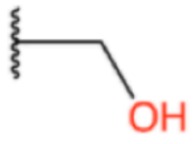
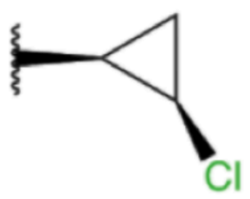
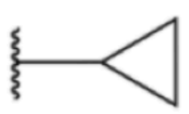
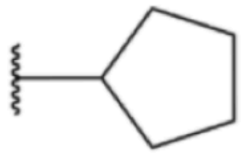
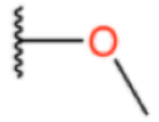
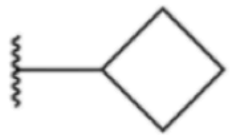
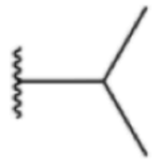
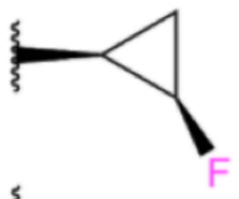
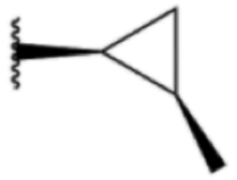
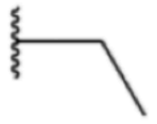
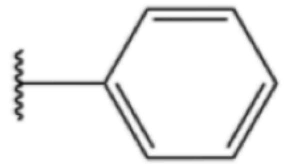
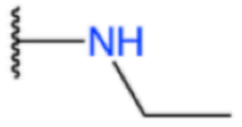
YUANQING WANG

ELIMINATING DISCRETE TYPES APPEARS TO SIGNIFICANTLY IMPROVE ACCURACY IN FREE ENERGY CALCULATIONS



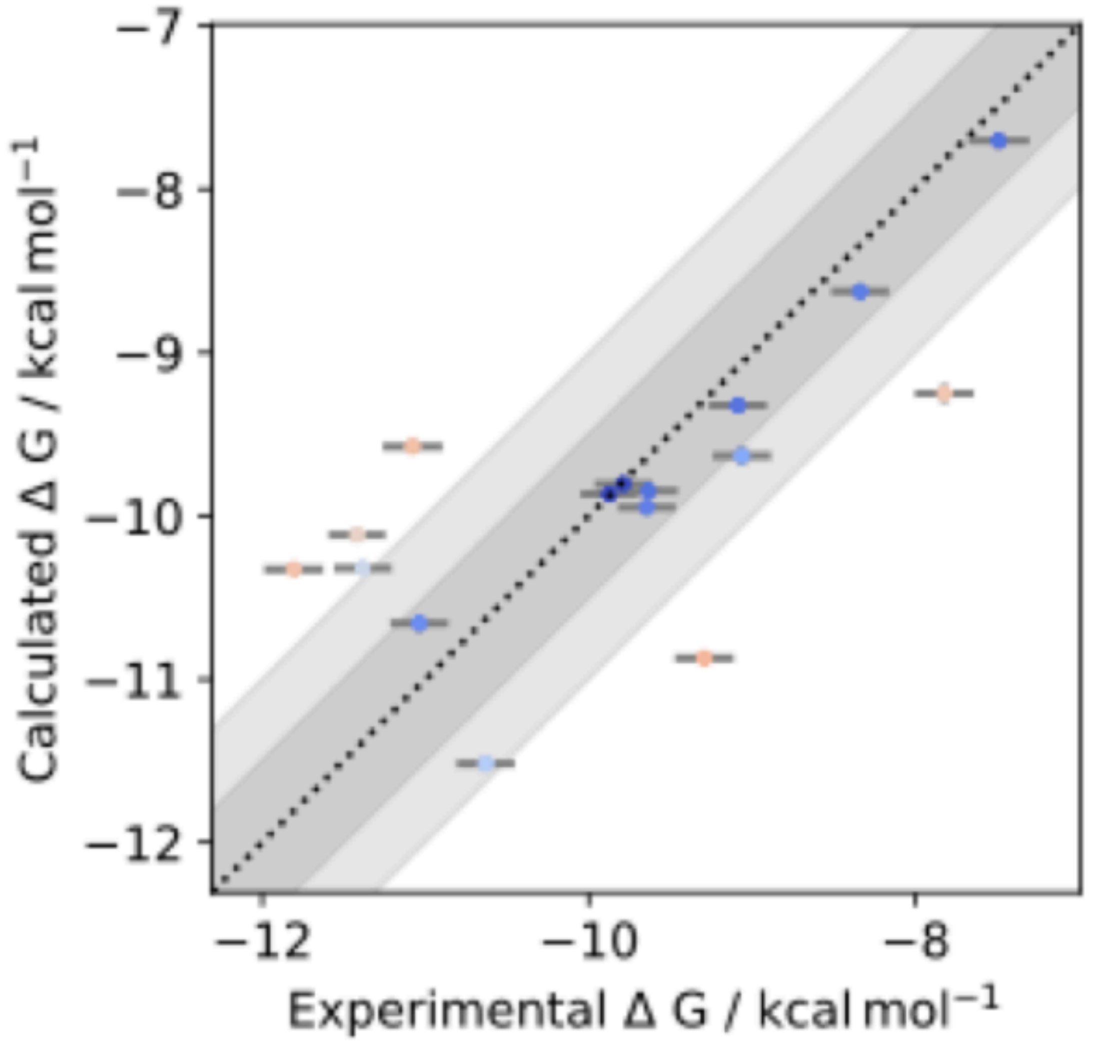


Tyk2
PDBID: 4GIH

		ΔG_{exp} / kcal mol^{-1}			ΔG_{exp} / kcal mol^{-1}
1		-9.54	9		-9.56
2		-10.94	10		-7.42
3		-8.98	11		-11.28
4		-11.31	12		-9.00
5		-9.21	13		-9.70
6		-8.26	14		-11.70
7		-10.91	15		-9.78
8		-7.75	16		-10.53

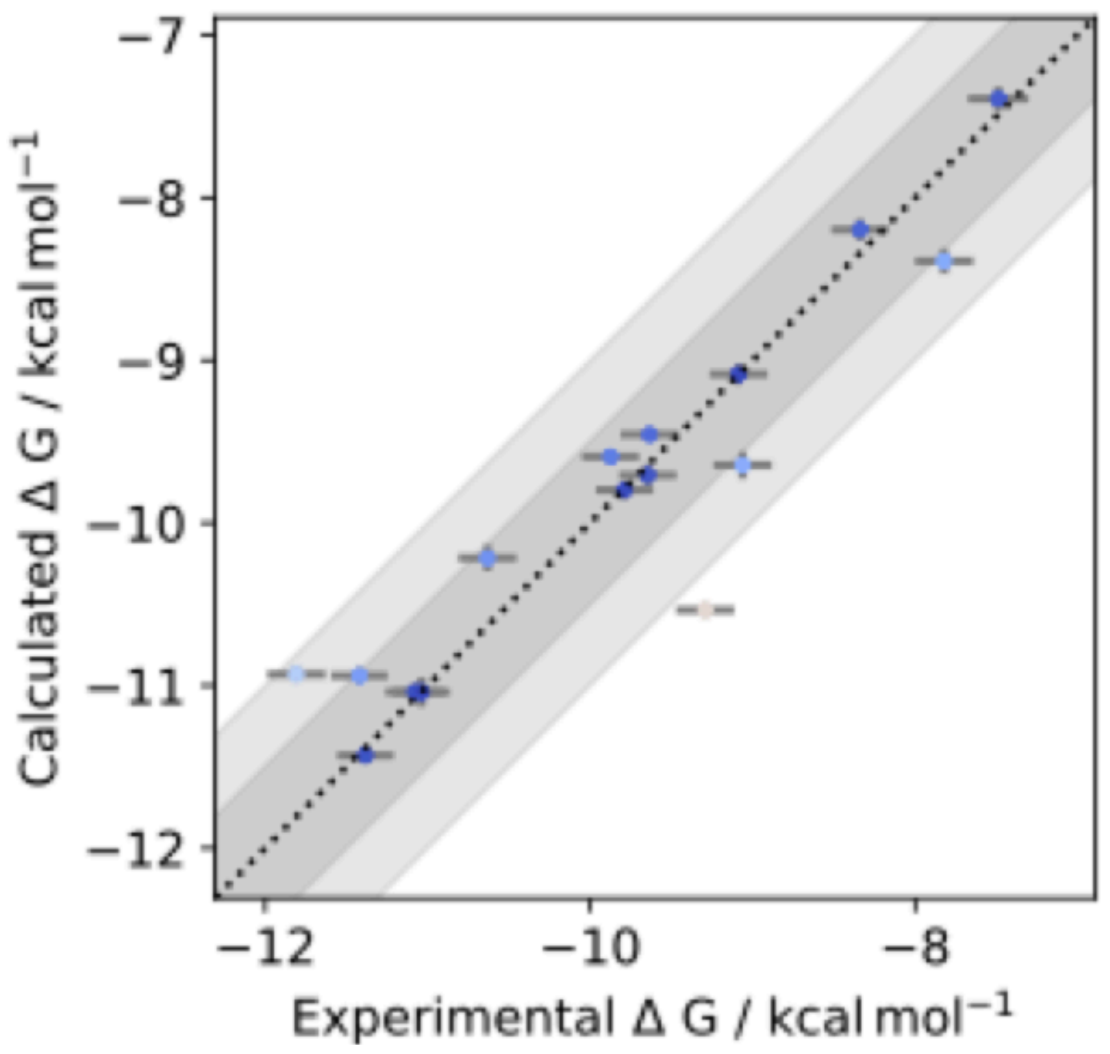
OpenFF 1.2.0 small molecule
Amber ff14SB protein
TIP3P water

```
Absolute binding energies - tyk2
                     tyk2 (N = 16)
 RMSE:   0.91 [95%: 0.66, 1.17]
  MUE:   0.72 [95%: 0.47, 1.03]
   R2:   0.48 [95%: 0.09, 0.78]
  rho:   0.69 [95%: 0.28, 0.89]
```



espaloma "joint" 0.2.2 small molecule
Amber ff14SB protein
TIP3P water

```
Absolute binding energies - tyk2
                     tyk2 (N = 16)
 RMSE:   0.47 [95%: 0.30, 0.70]
  MUE:   0.31 [95%: 0.22, 0.56]
   R2:   0.87 [95%: 0.62, 0.96]
  rho:   0.93 [95%: 0.80, 0.98]
```



MIKE HENRY

IVÁN PULIDO

IVY ZHANG

DOMINIC RUFA

HANNAH BRUCE CDONALD

YUANQING WANG

preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma
free energy calculations with http://github.com/choderalab/perses

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

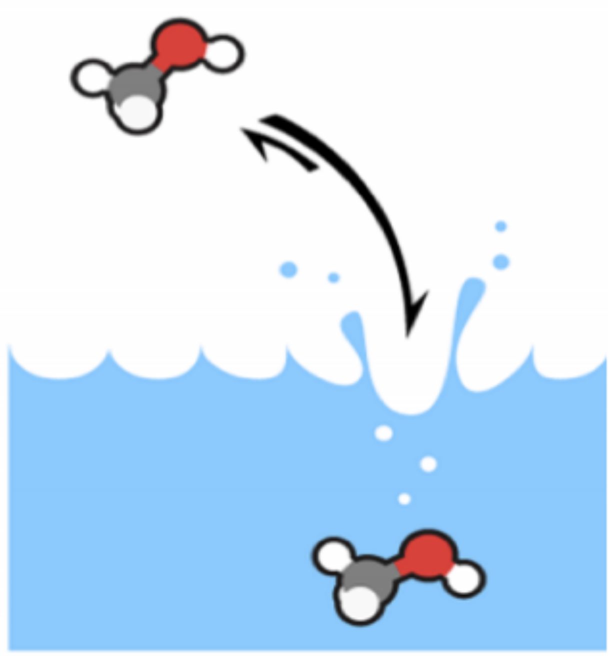
week 2

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

using new model tuned to target

from first week’s data

CAN WE LEARN TO FIT EXPERIMENTAL DATA AS WELL?



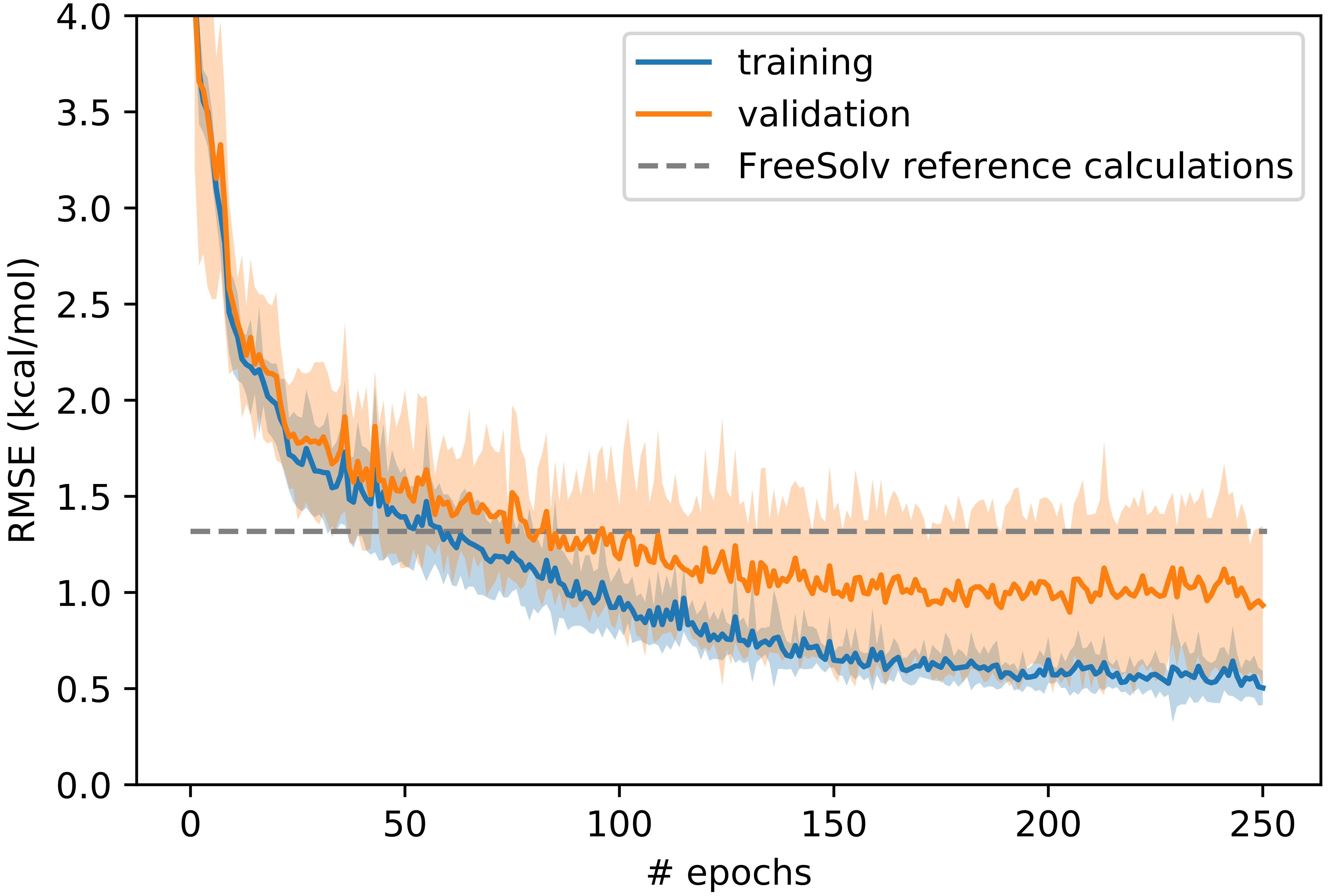
experimental hydration
free energies from **FreeSolv**
https://github.com/MobleyLab/FreeSolv

loss function:

$$L(\Phi_{NN}) = \sum_{n=1}^{N} \frac{[\Delta G_n(\Phi_{NN}) - \Delta G_n^{\text{exp}}]^2}{\sigma_n^2}$$

Here, ΔG estimated via one-step free energy perturbation, but can easily differentiate properties through MBAR

OBC2 GBSA FreeSolv RMSE



JOSH FASS

YUANQING WANG

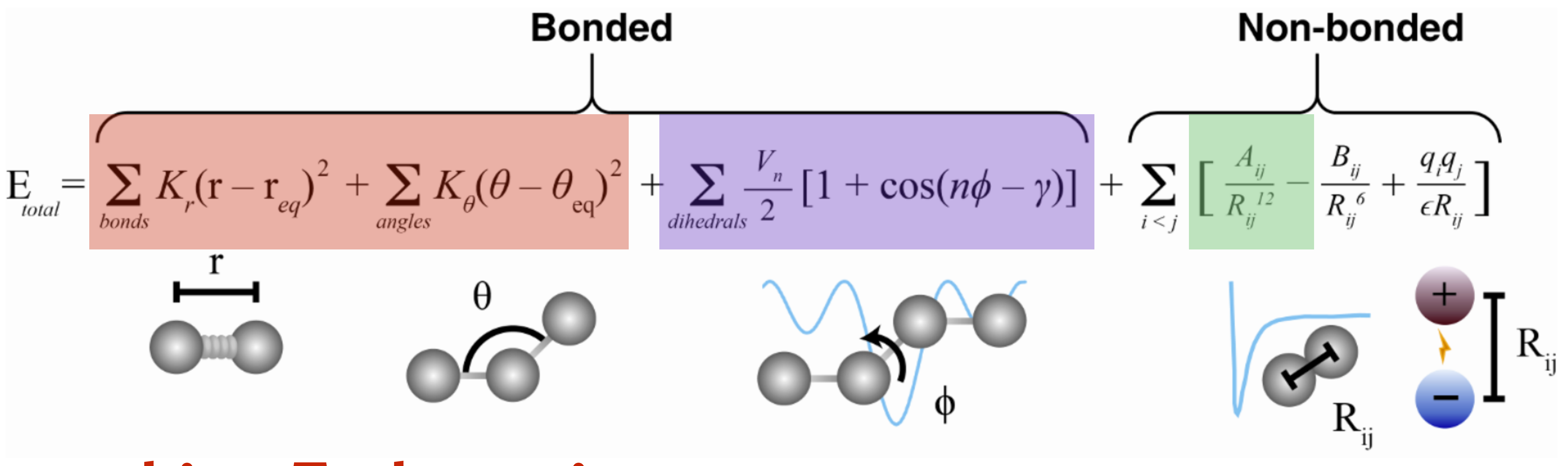
preprint: https://arxiv.org/abs/2010.01196
code: https://github.com/choderalab/espaloma

WHY SHOULD WE BE STUCK WITH A PHYSICAL MODEL THAT CATERED TO THE CAPABILITIES OF A PDP-11?

DEC PDP-11

~45 years old

typical class I molecular mechanics force field



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

shitty Taylor series truncated at lowest order

crappy Fourier series truncated at n=6

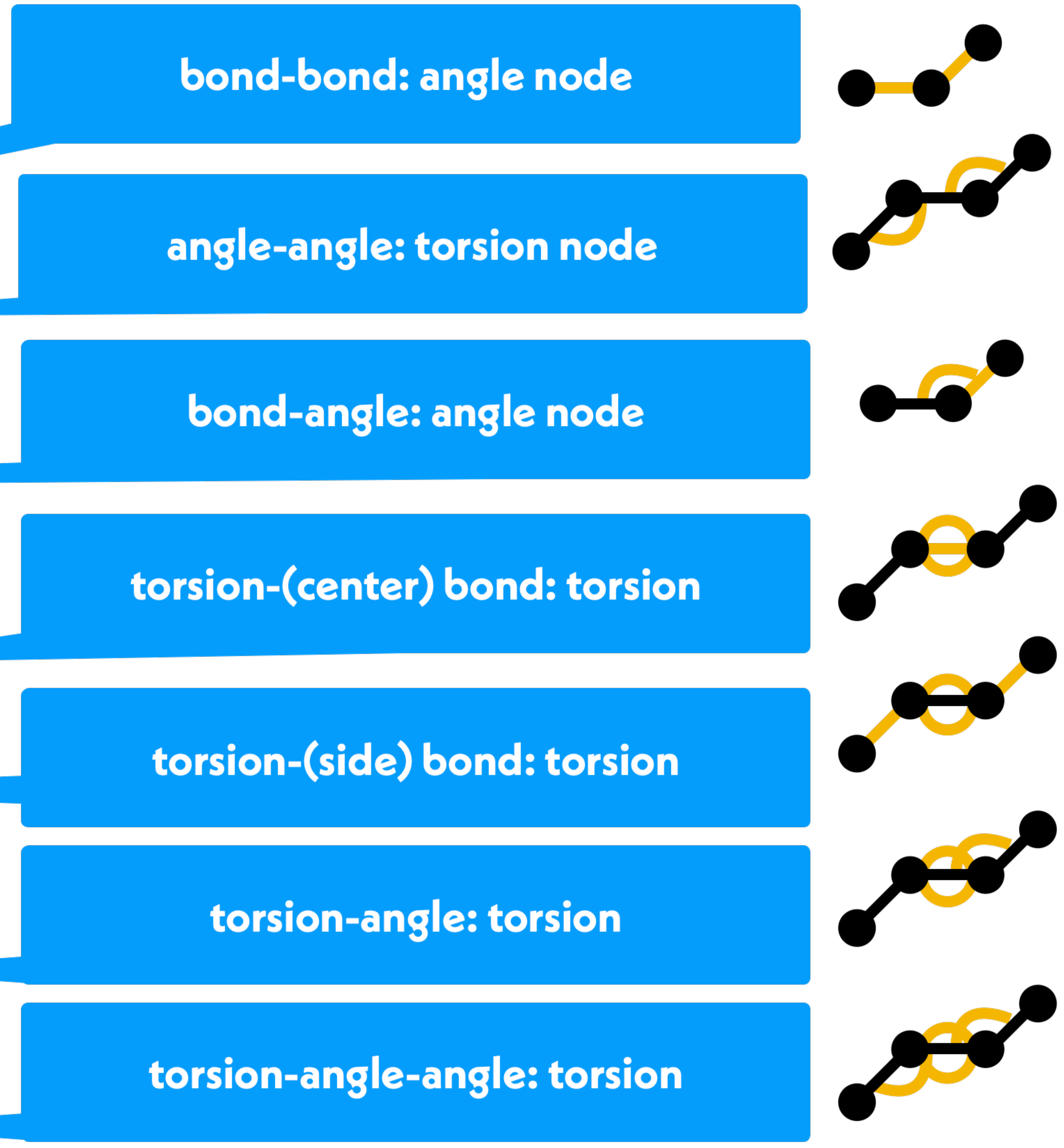
don't even get me started on this fucker

Shan, Kim, Eastwood, Dror, Seeliger, Shaw. JACS 133:9181, 2011

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

WE COULD GO TO CLASS II FORCE FIELDS… IT'S CERTAINLY EASY TO DO NOW

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



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

But can we do a better job of modeling true many-body local valence terms?

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

A NEW GENERATION OF QUANTUM MACHINE LEARNING (QML) POTENTIALS PROVIDE SIGNIFICANTLY MORE FLEXIBILITY IN FUNCTIONAL FORM, THOUGH AT MUCH GREATER COST

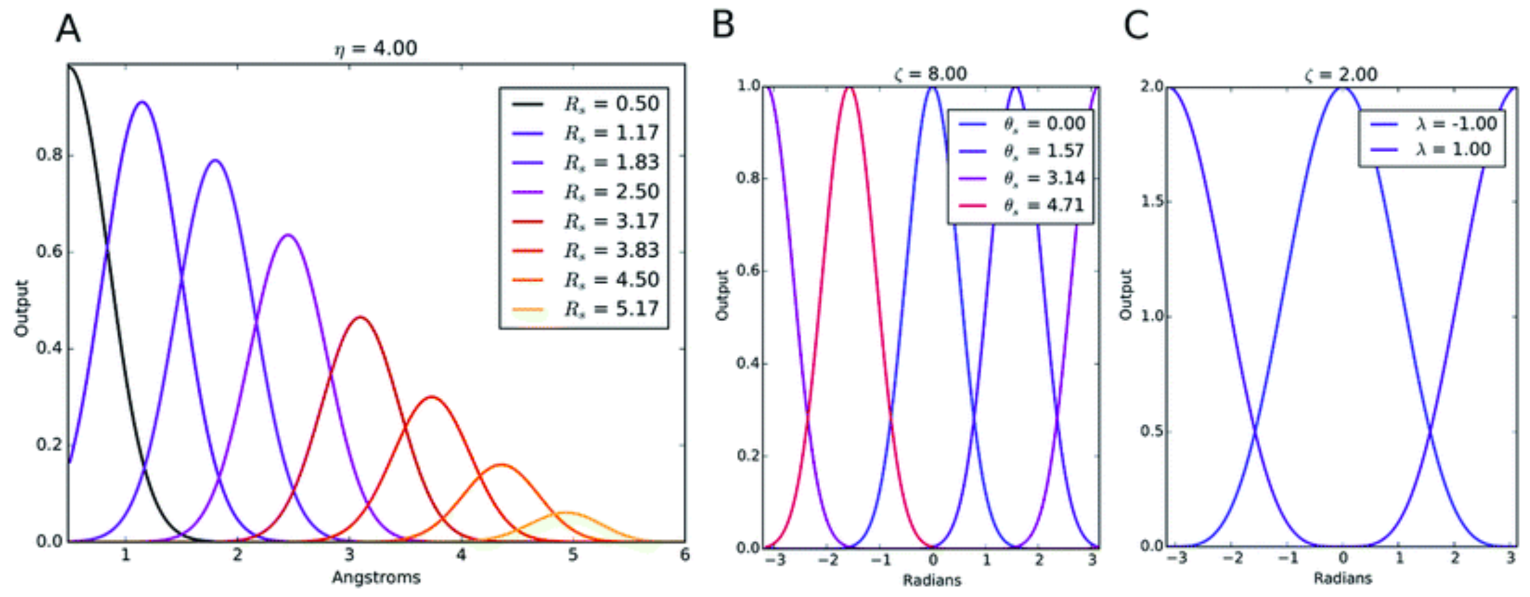
ANI family of quantum machine learning (QML) potentials

radial and **angular** features

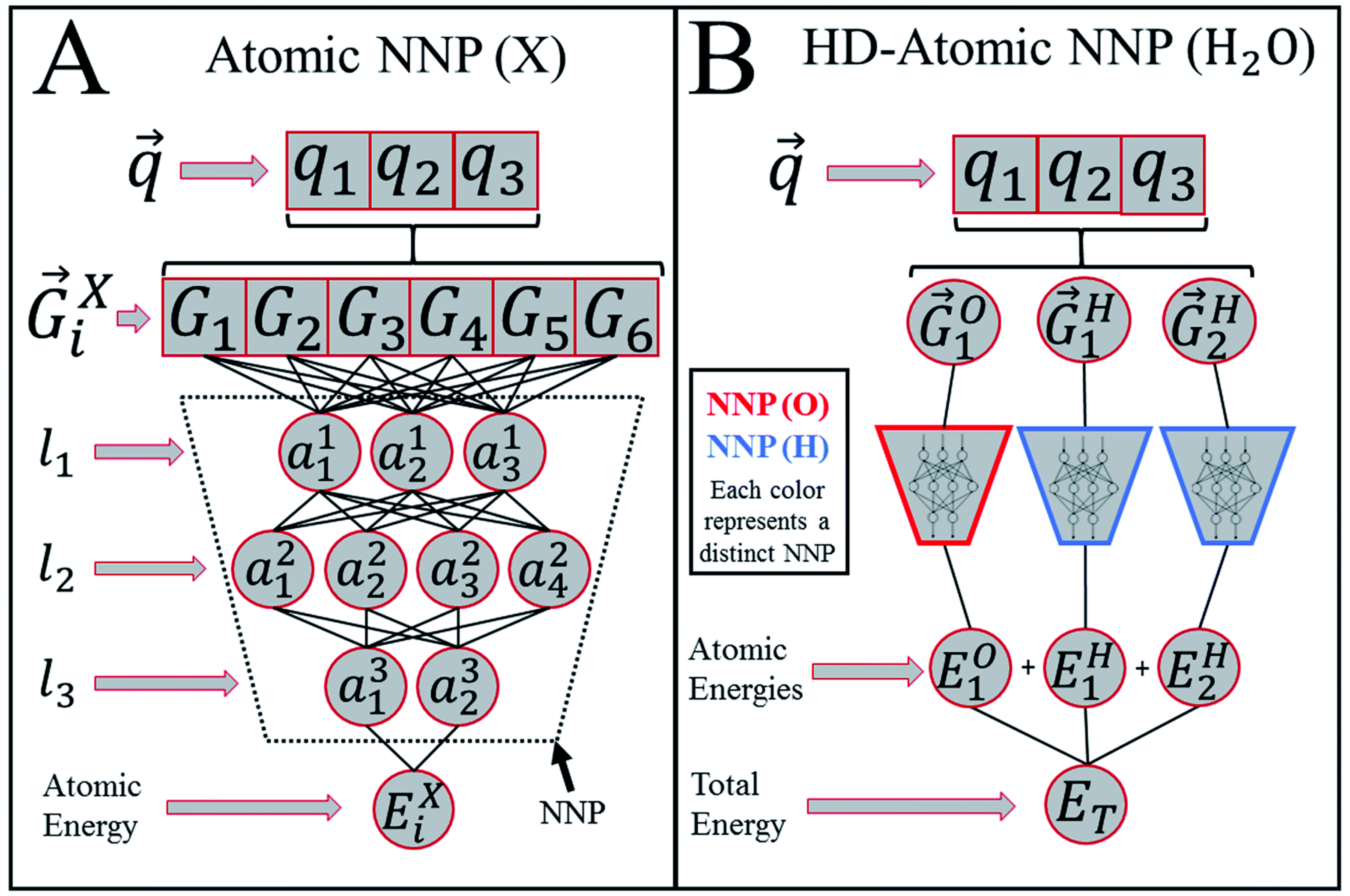
$$f_C(R_{ij}) = \begin{cases} 0.5\times\cos\left(\frac{\pi R_{ij}}{R_C}\right)+0.5 & \text{for } R_{ij} \le R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases}$$

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

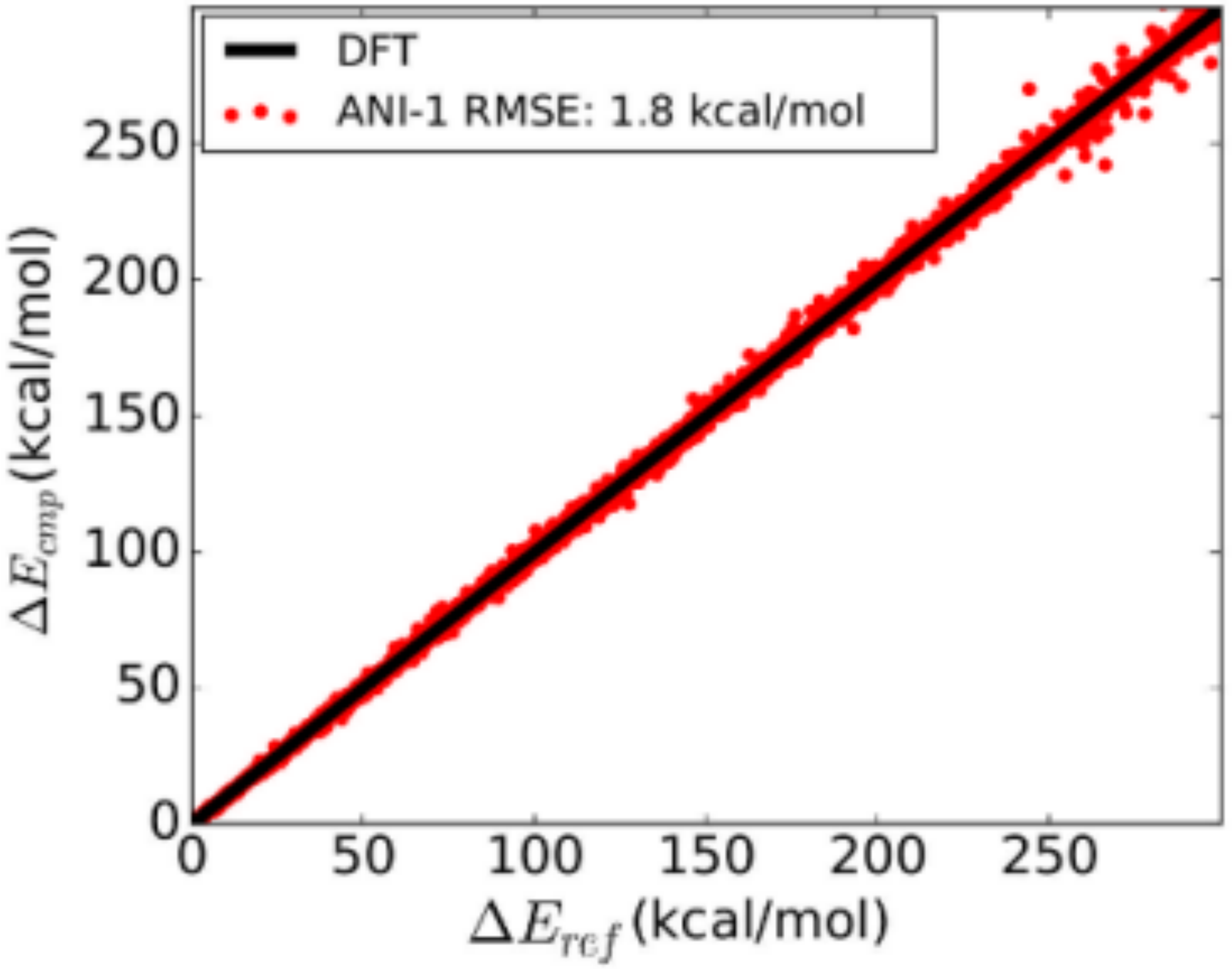
$$G_m^{A_{\text{mod}}} = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all atoms}} \left(1+\cos(\theta_{ijk}-\theta_s)\right)^{\zeta} \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_s\right)^2\right] f_C(R_{ij}) f_C(R_{ik})$$



deep neural network for each atom



excellent agreement with DFT



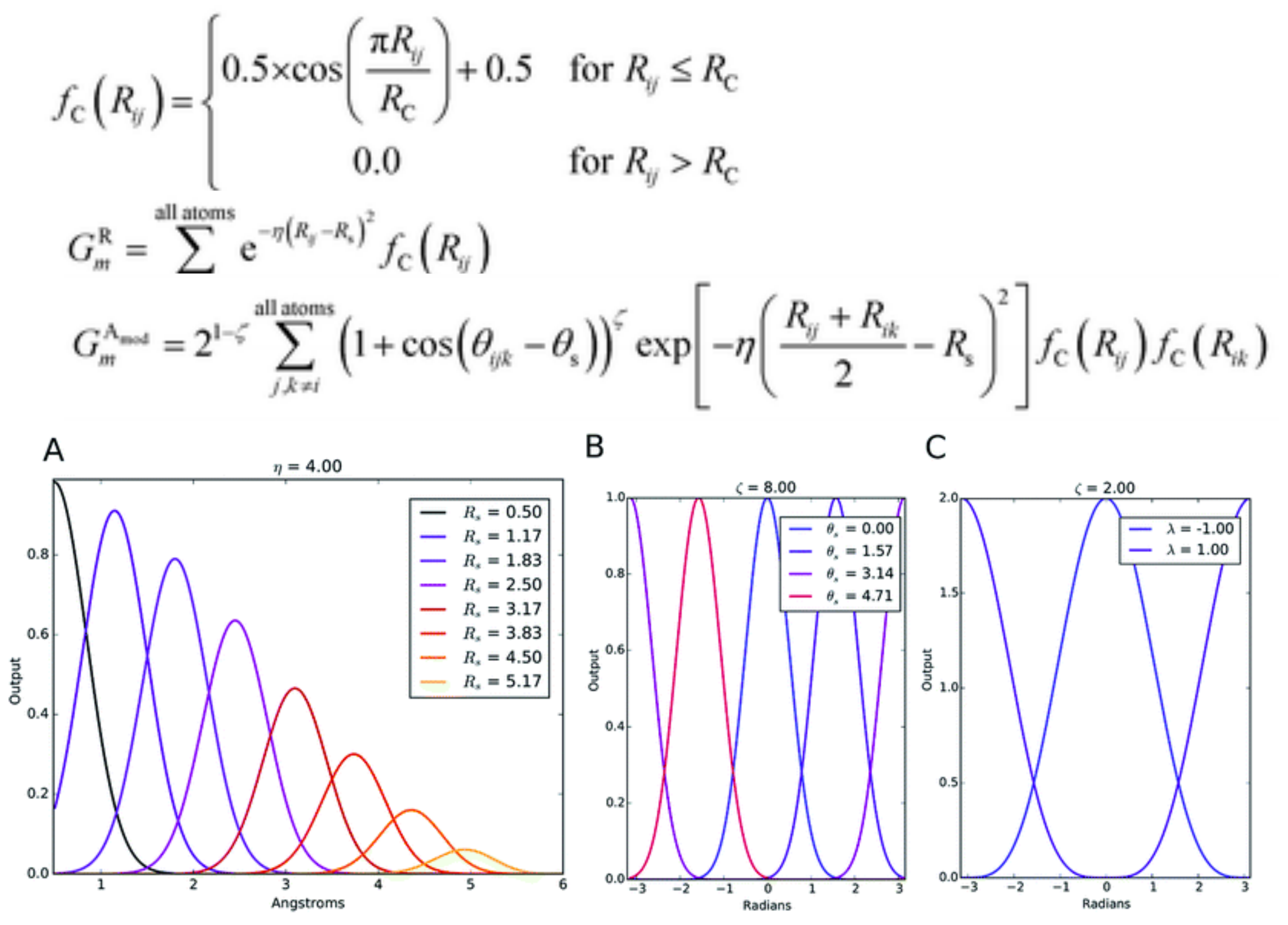
OLEXANDR ISAYEV

ADRIAN ROITBERG

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

QML POTENTIALS ARE SEEING RAPID EVOLUTION IN ARCHITECTURES

ANI

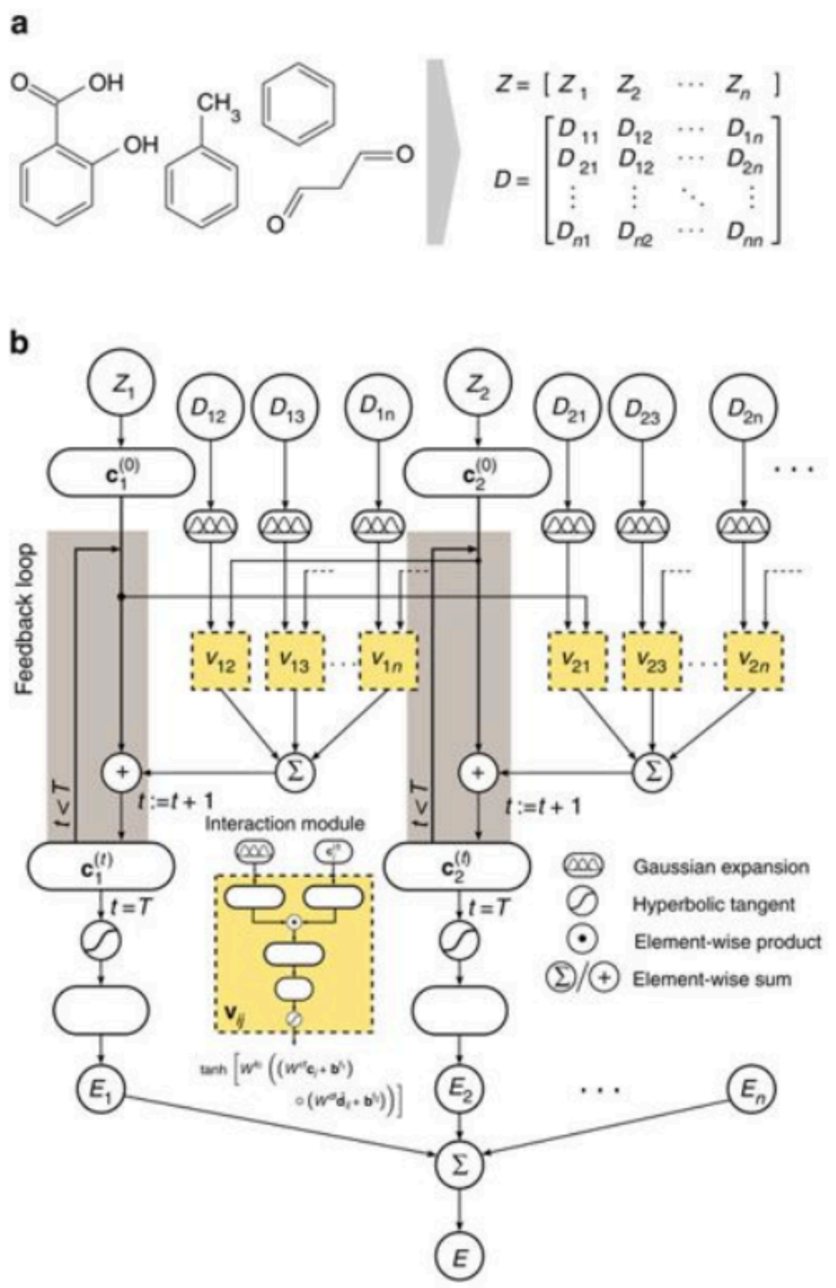


$$f_C(R_{ij}) = \begin{cases} 0.5\times\cos\left(\frac{\pi R_{ij}}{R_C}\right)+0.5 & \text{for } R_{ij}\le R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases}$$

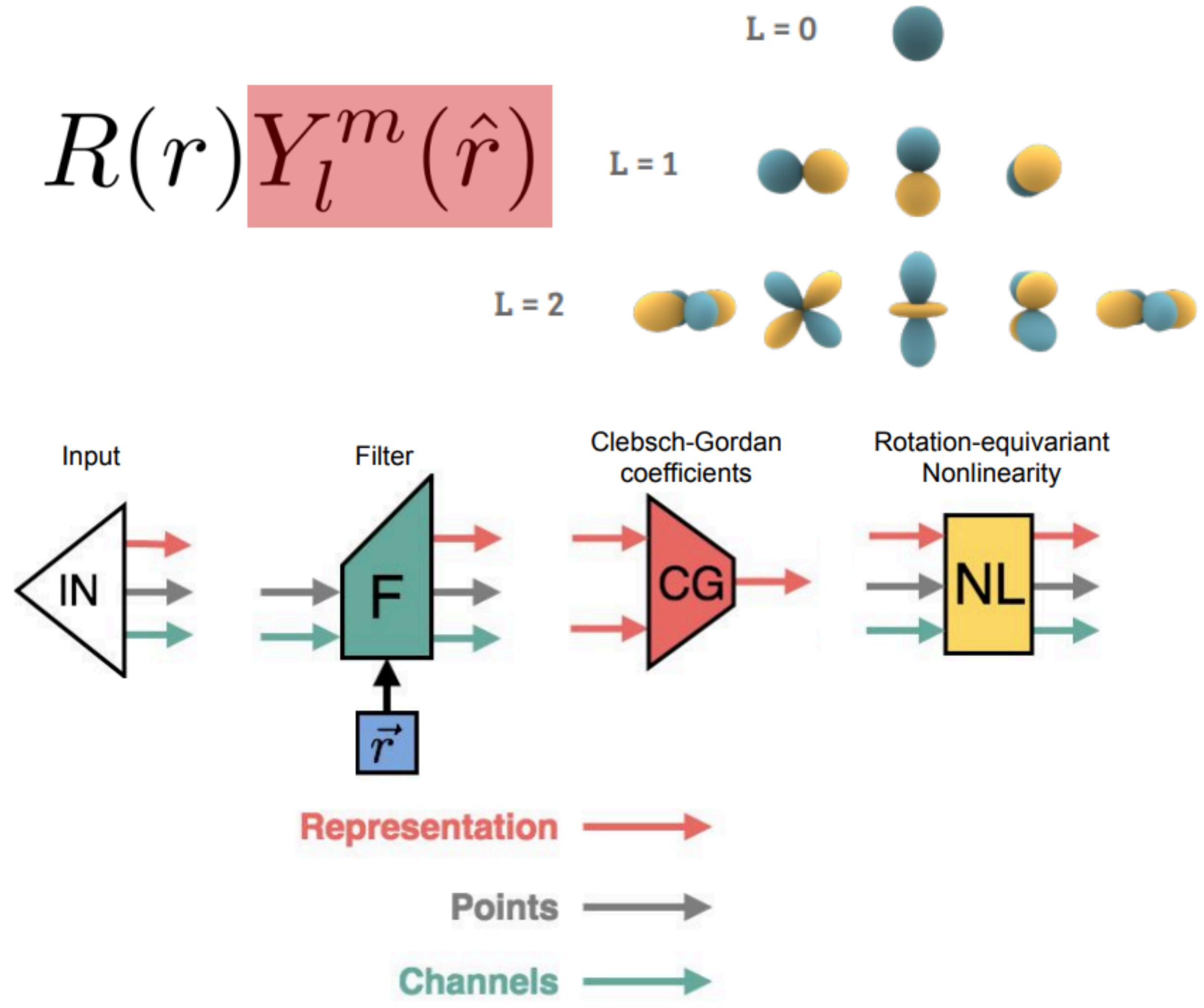
$$G_m^R = \sum_{}^{\text{all atoms}} e^{-\eta(R_{ij}-R_s)^2} f_C(R_{ij})$$

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

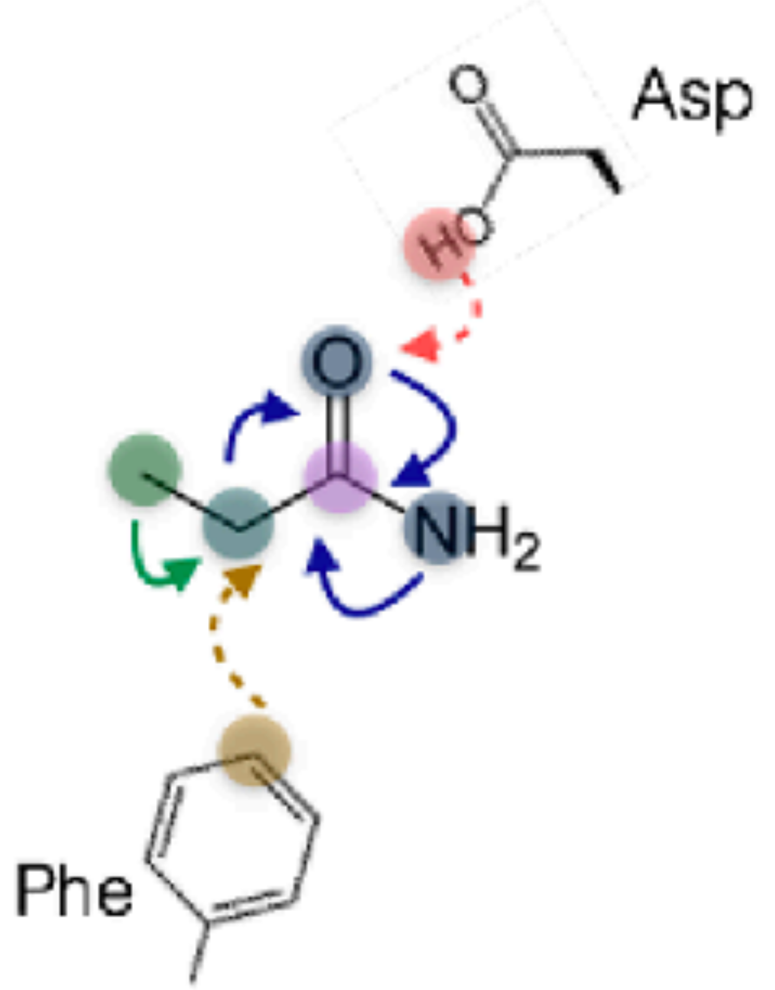
Deep Tensor Networks



Tensor Field Networks

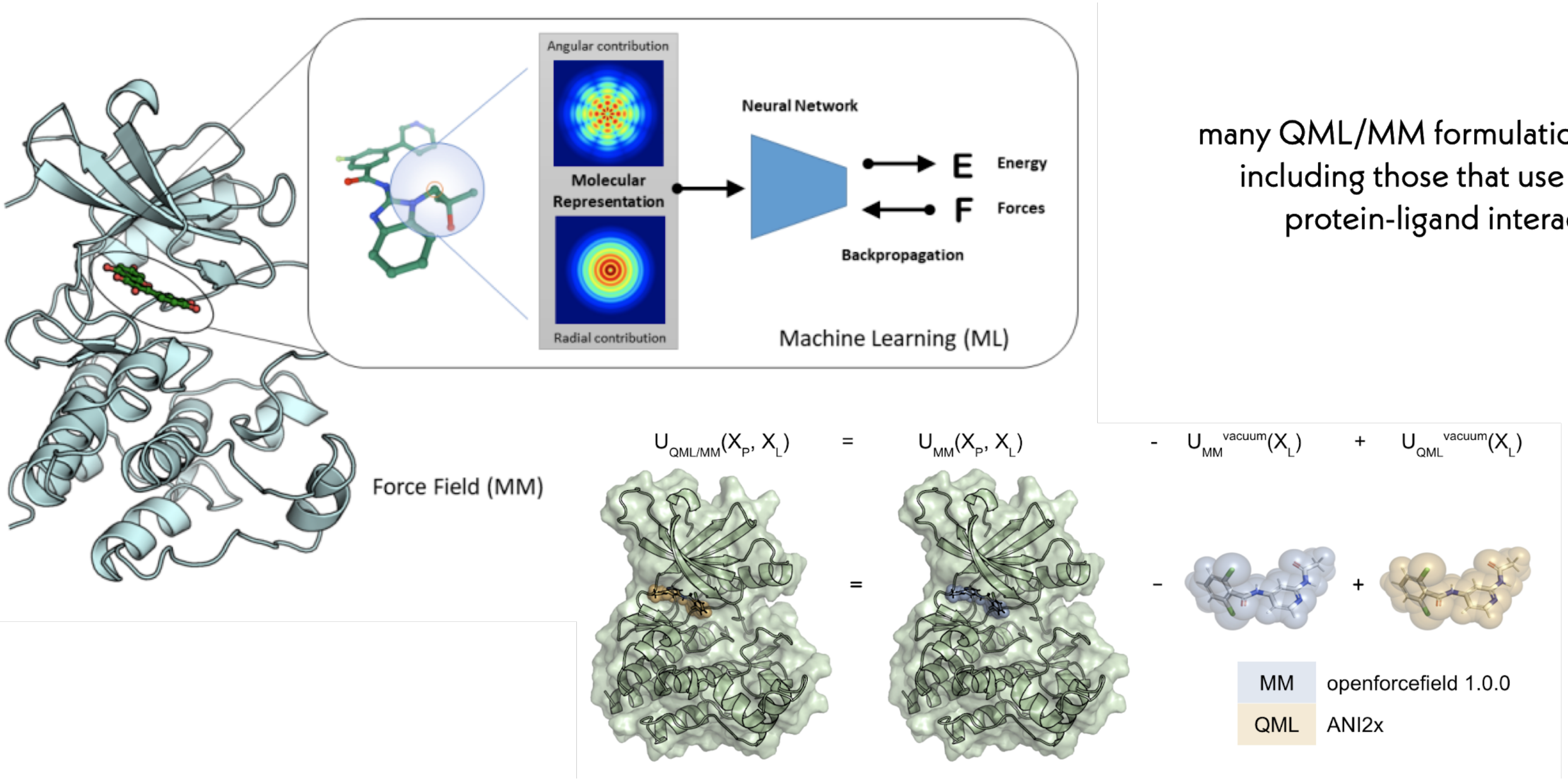


PotentialNet



The **ANI** class of models uses distance- and angle-based features [http://doi.org/10.1039/c6sc05720a].
Deep Tensor Networks and **SchNet** use 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].
PotentialNet uses a graph convolutional network augmented by distance-dependent edges [https://doi.org/10.1021/acscentsci.8b00507].

HYBRID QUANTUM MACHINE LEARNING / MOLECULAR MECHANICS (QML/MM) SIMULATIONS ARE MOST FEASIBLE IN THE NEAR TERM



many QML/MM formulations possible, including those that use QML for protein-ligand interactions

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

MM	openforcefield 1.0.0
QML	ANI2x

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

WE CAN ASSESS HOW WELL QML/MM FREE ENERGY CALCULATIONS MIGHT PERFORM THROUGH A PERTURBATIVE CORRECTION

A ML/MM AUGMENTED THERMODYNAMIC CYCLE

$$\begin{tikzcd}
PL_1{}^* & PL_1 \arrow[l, "\Delta G^{MM\to ML/MM}_{PL_1}"'] \arrow[r, "\Delta G^{MM}_{PL_1\to PL_2}"] & PL_2 \arrow[r, "\Delta G^{MM\to ML/MM}_{PL_2}"] & PL_2^* \\
L_1{}^* \arrow[u] & L_1 \arrow[u] \arrow[l, "\Delta G^{MM\to ML/MM}_{L_1}"'] \arrow[r, "\Delta G^{MM}_{L_1\to L_2}"] & L_2 \arrow[u] \arrow[r, "\Delta G^{MM\to ML/MM}_{L_2}"] & L_2{}^* \arrow[u]
\end{tikzcd}$$

ML/MM MM ML/MM

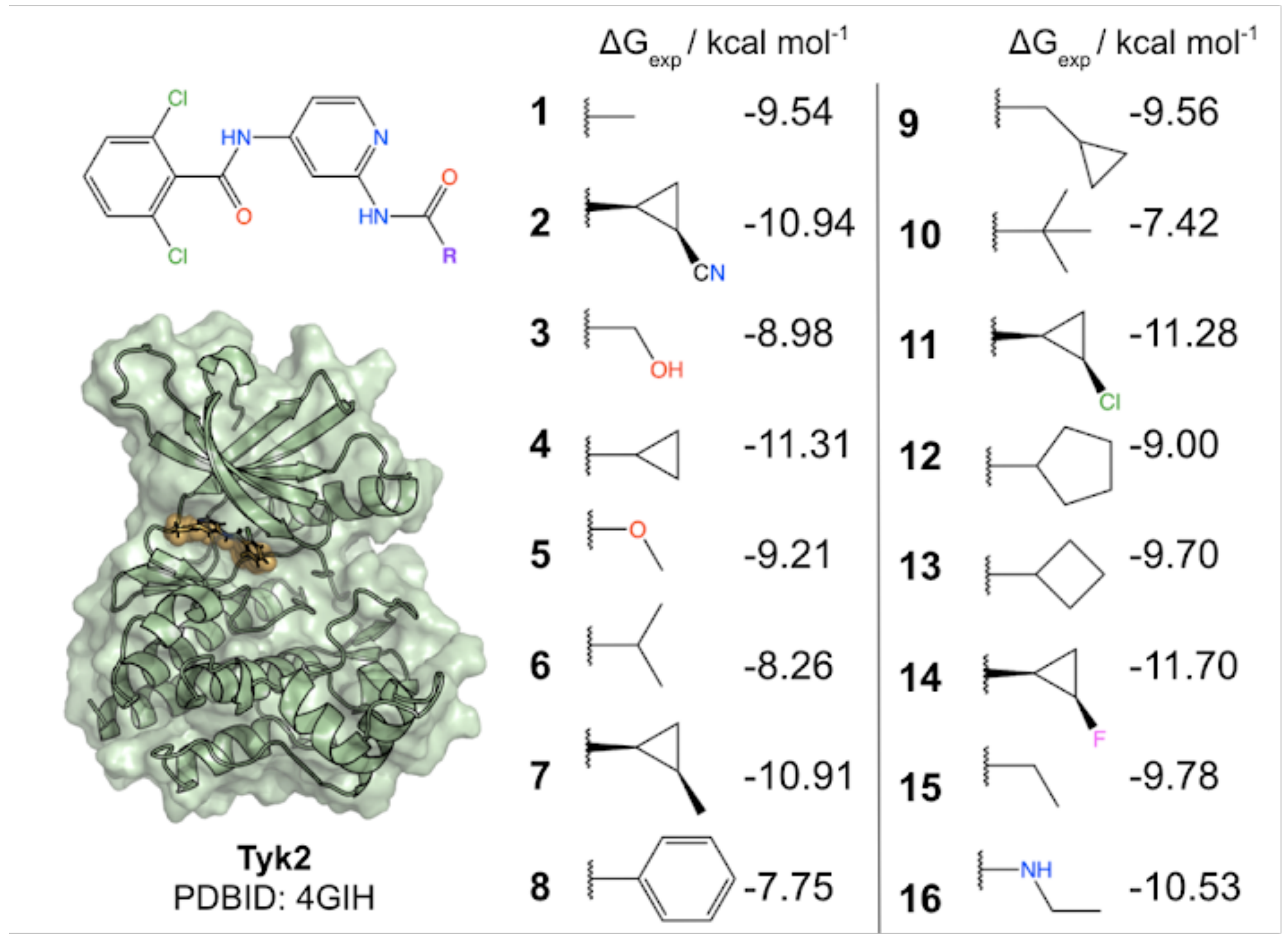
Rufa, Bruce Macdonald, Fass, Wieder, Grinaway, Roitberg, Isayev, and **Chodera**.

preprint: https://doi.org/10.1101/2020.07.29.227959

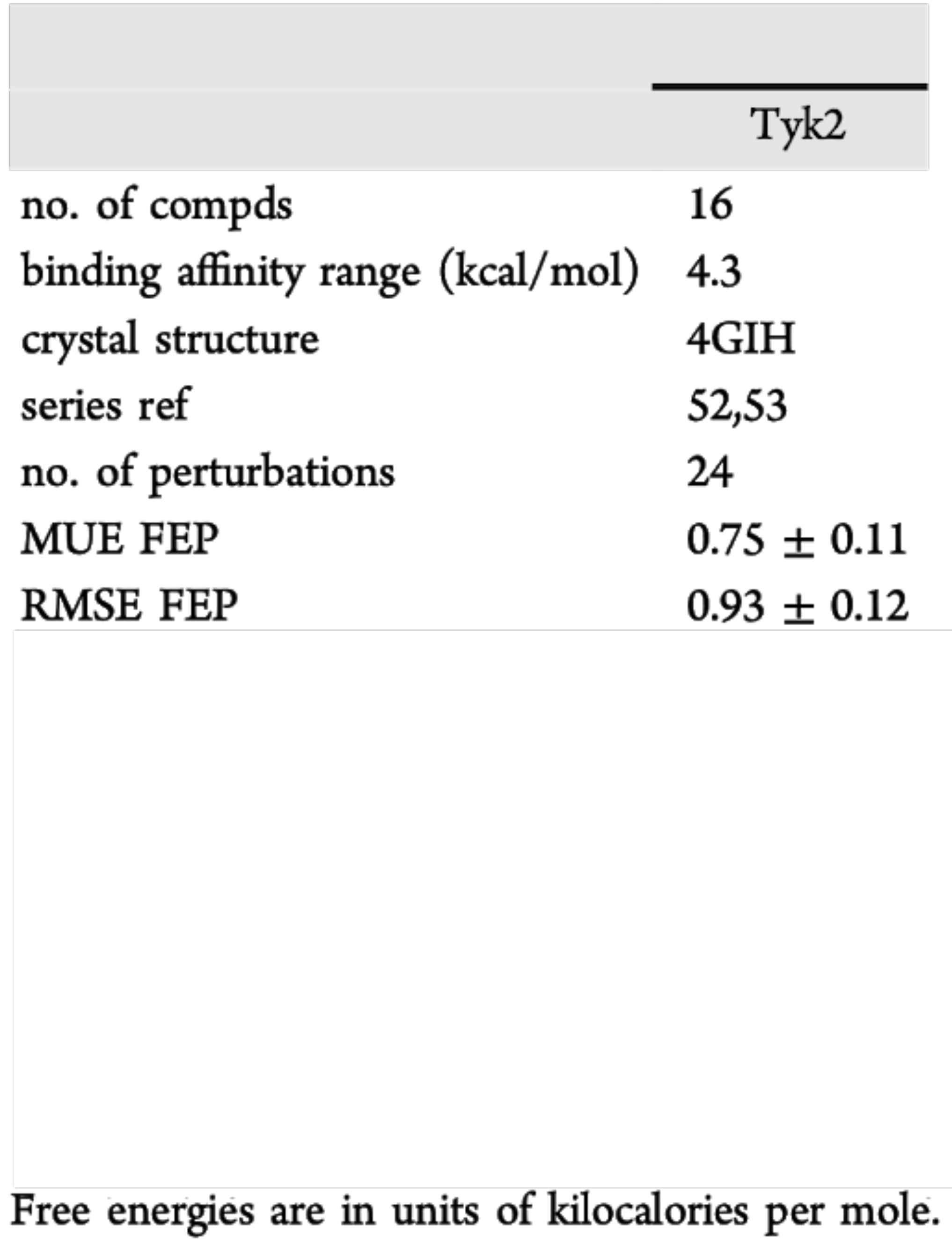
code: https://github.com/choderalab/qmlify

HYBRID QUANTUM MACHINE LEARNING / MOLECULAR MECHANICS (QML/MM) POST-PROCESSING CAN IMPROVE ACCURACY

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



	ΔG_{exp} / kcal mol^{-1}		ΔG_{exp} / kcal mol^{-1}
1	-9.54	9	-9.56
2	-10.94	10	-7.42
3	-8.98	11	-11.28
4	-11.31	12	-9.00
5	-9.21	13	-9.70
6	-8.26	14	-11.70
7	-10.91	15	-9.78
8	-7.75	16	-10.53



	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

Free energies are in units of kilocalories per mole.

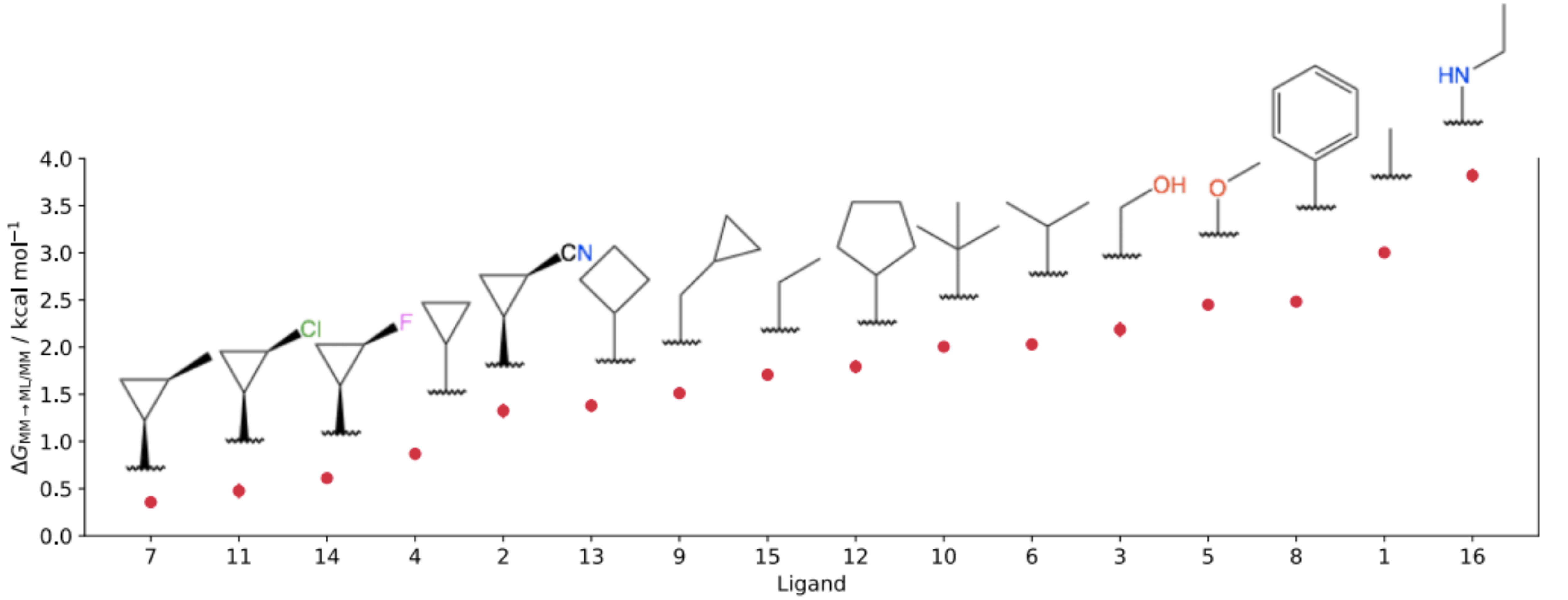


Figure 7. ML/MM corrections to MM binding free energies can be up to 4 kcal mol^{-1} in magnitude. The signed $\Delta G^{MM \to ML/MM}$ corrections for each ligand (with R-group shown) are shown, ordered from least positive (slightly disfavoring binding) to most positive (strongly disfavoring binding).

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

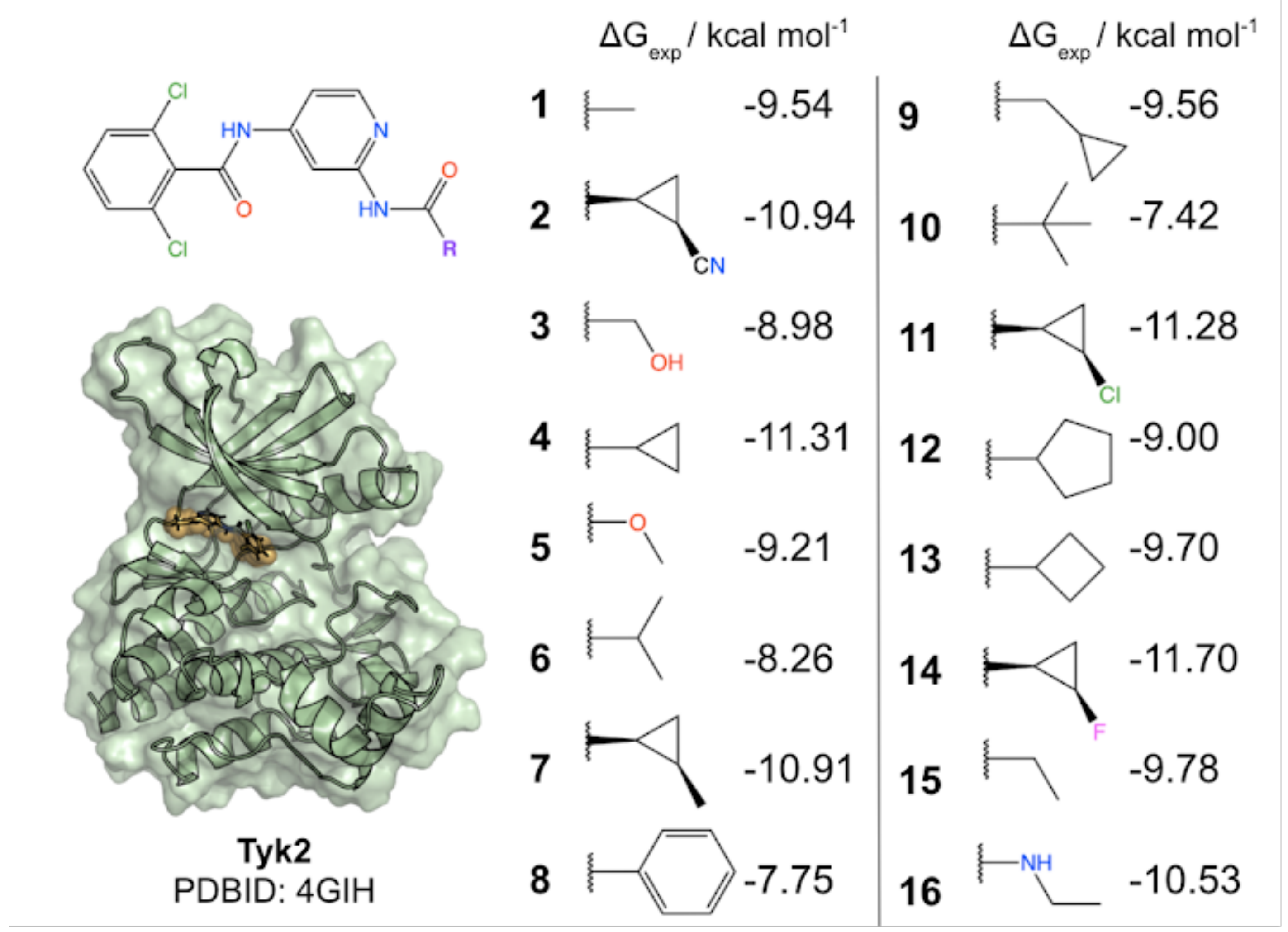
Rufa, Bruce Macdonald, Fass, Wieder, Grinaway, Roitberg, Isayev, and **Chodera**.
preprint: https://doi.org/10.1101/2020.07.29.227959
code: https://github.com/choderalab/qmlify

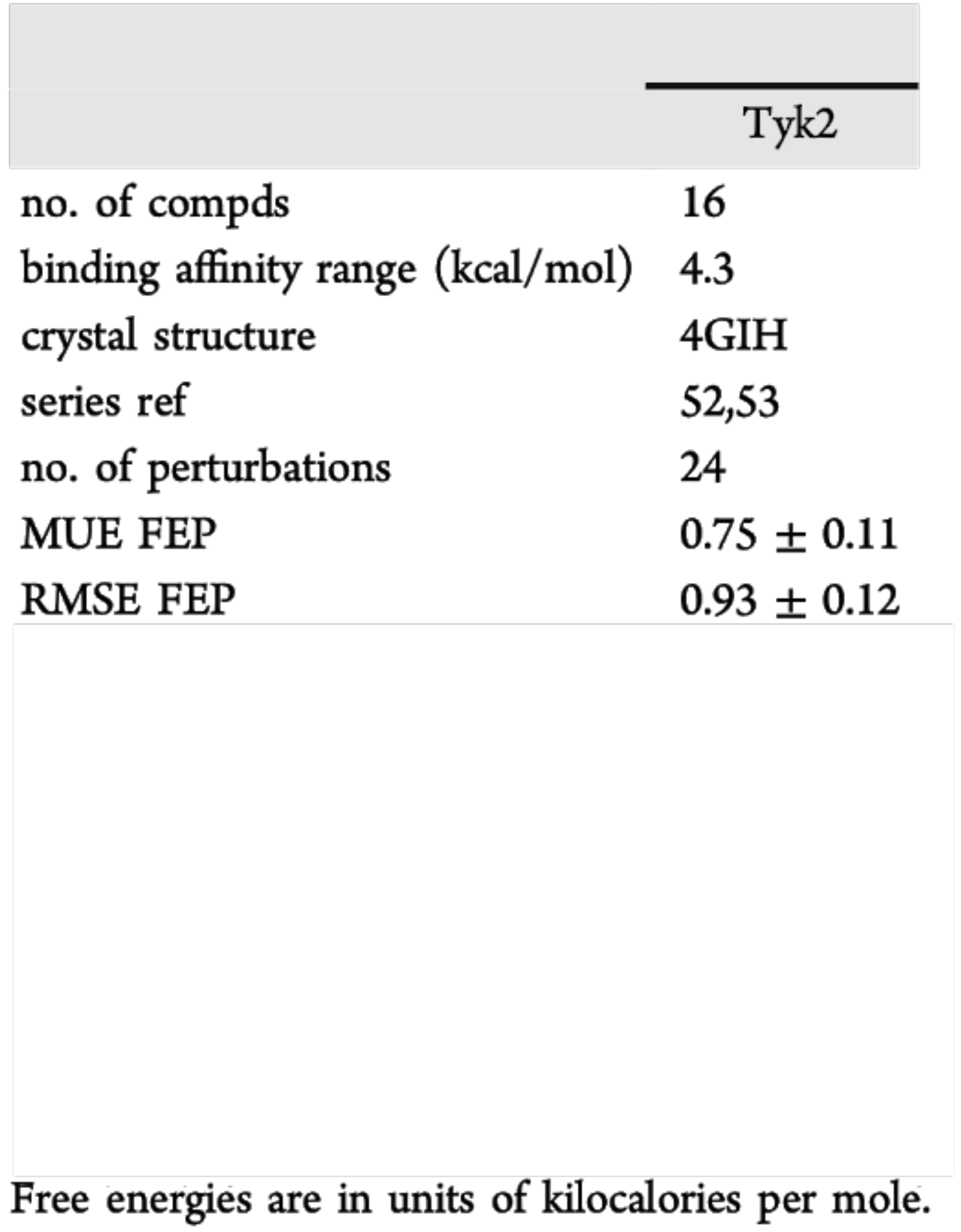
HYBRID QUANTUM MACHINE LEARNING / MOLECULAR MECHANICS (QML/MM) FREE ENERGY CALCULATIONS CUT ERROR IN HALF

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

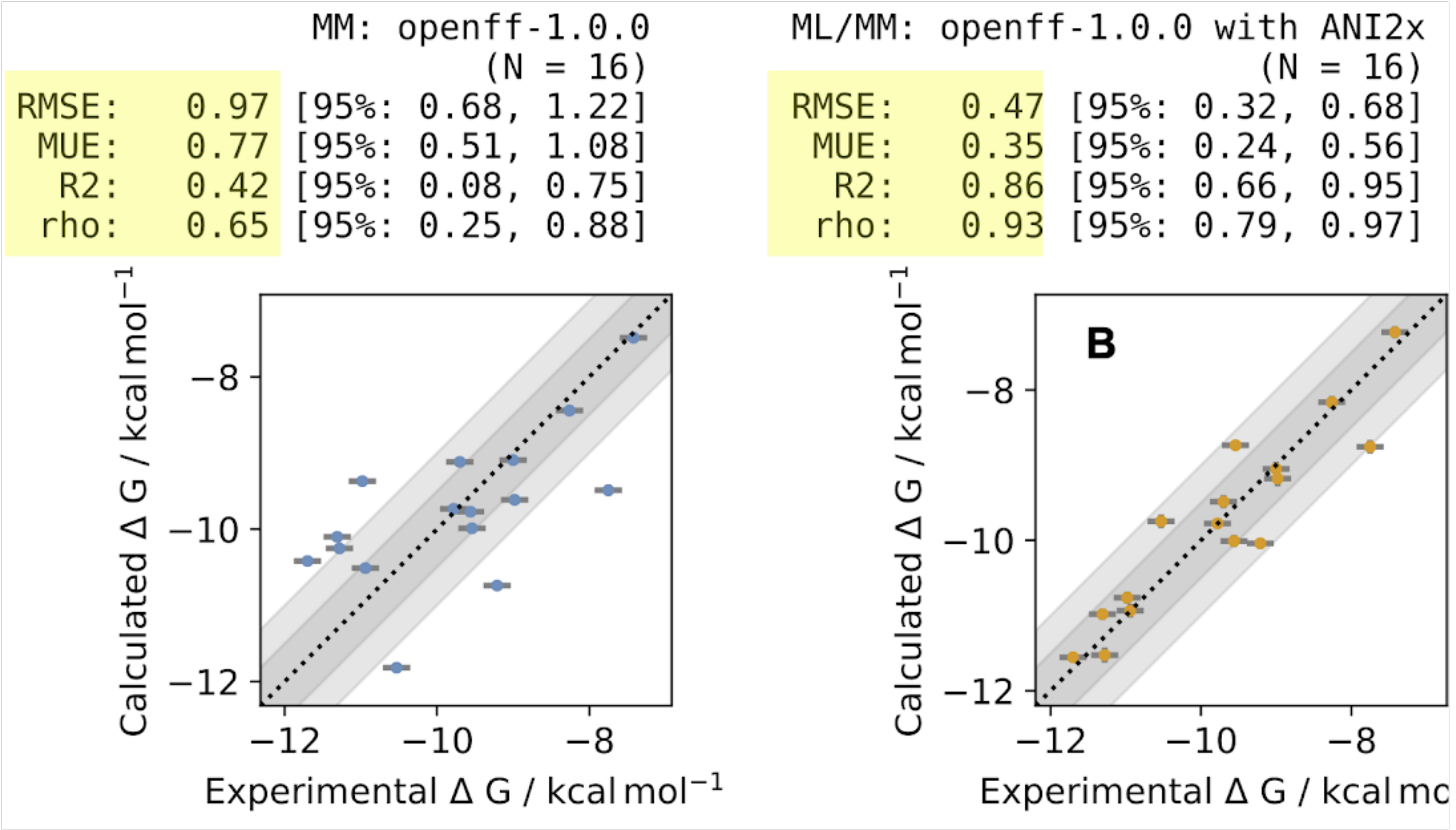




	Tyk2
no. of compds	16
binding affinity range (kcal/mol)	4.3
crystal structure	4GIH
series ref	52,53
no. of perturbations	24
MUE FEP	0.75 ± 0.11
RMSE FEP	0.93 ± 0.12

Free energies are in units of kilocalories per mole.

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



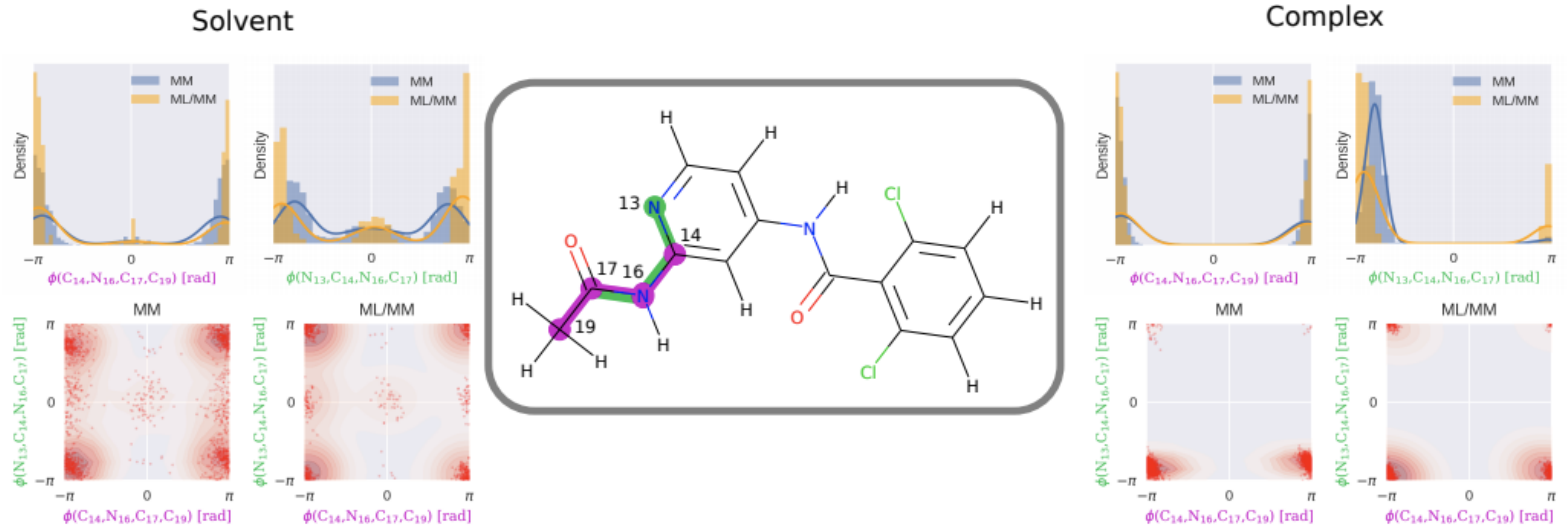
```
MM: openff-1.0.0
(N = 16)
RMSE:   0.97 [95%: 0.68, 1.22]
 MUE:   0.77 [95%: 0.51, 1.08]
  R2:   0.42 [95%: 0.08, 0.75]
 rho:   0.65 [95%: 0.25, 0.88]
```

```
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]
```

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

HYBRID QUANTUM MACHINE LEARNING / MOLECULAR MECHANICS (QML/MM) POST-PROCESSING CAN IMPROVE ACCURACY

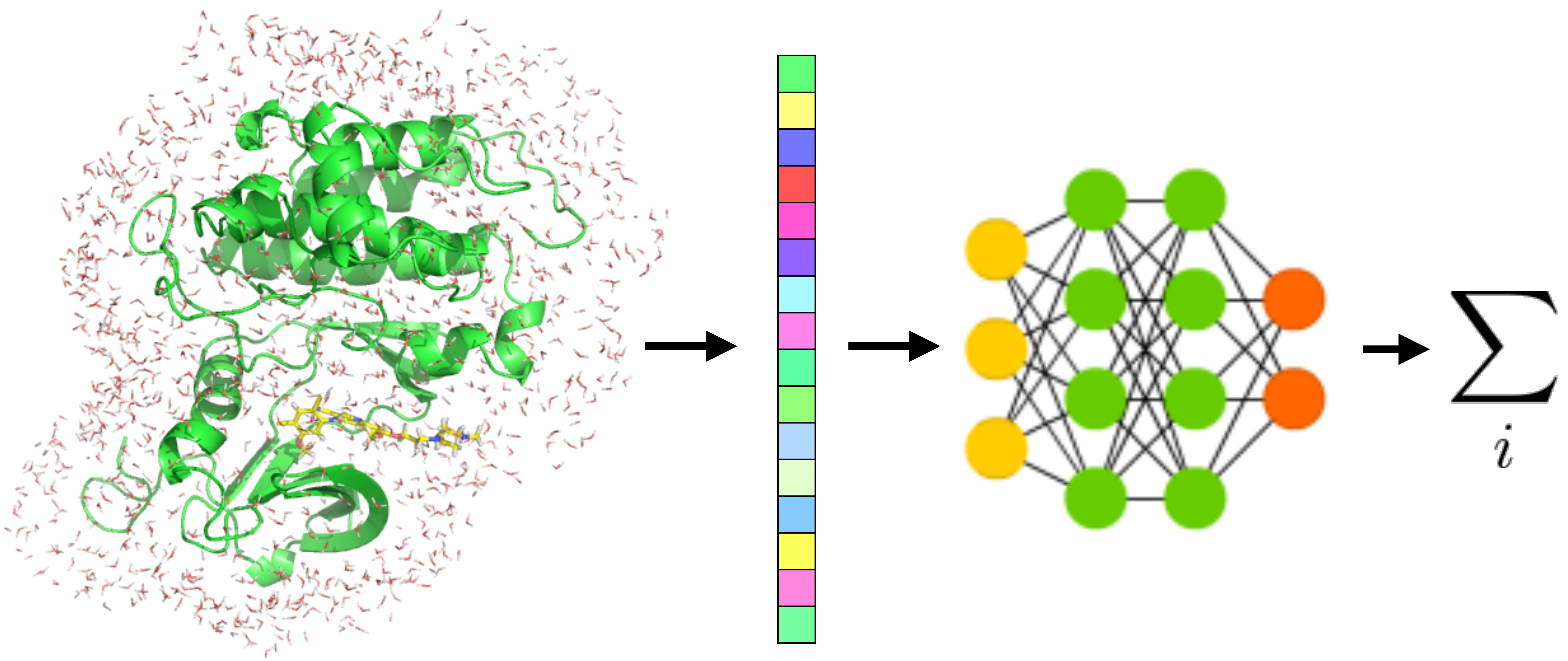


Rufa, Bruce Macdonald, Fass, Wieder, Grinaway, Roitberg, Isayev, and **Chodera**.

preprint: https://doi.org/10.1101/2020.07.29.227959

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

COMPUTATIONAL BOTTLENECKS IN CURRENT QML MODELS CAN BE SPED UP WITH CUSTOM GPU KERNELS



atomic coordinates

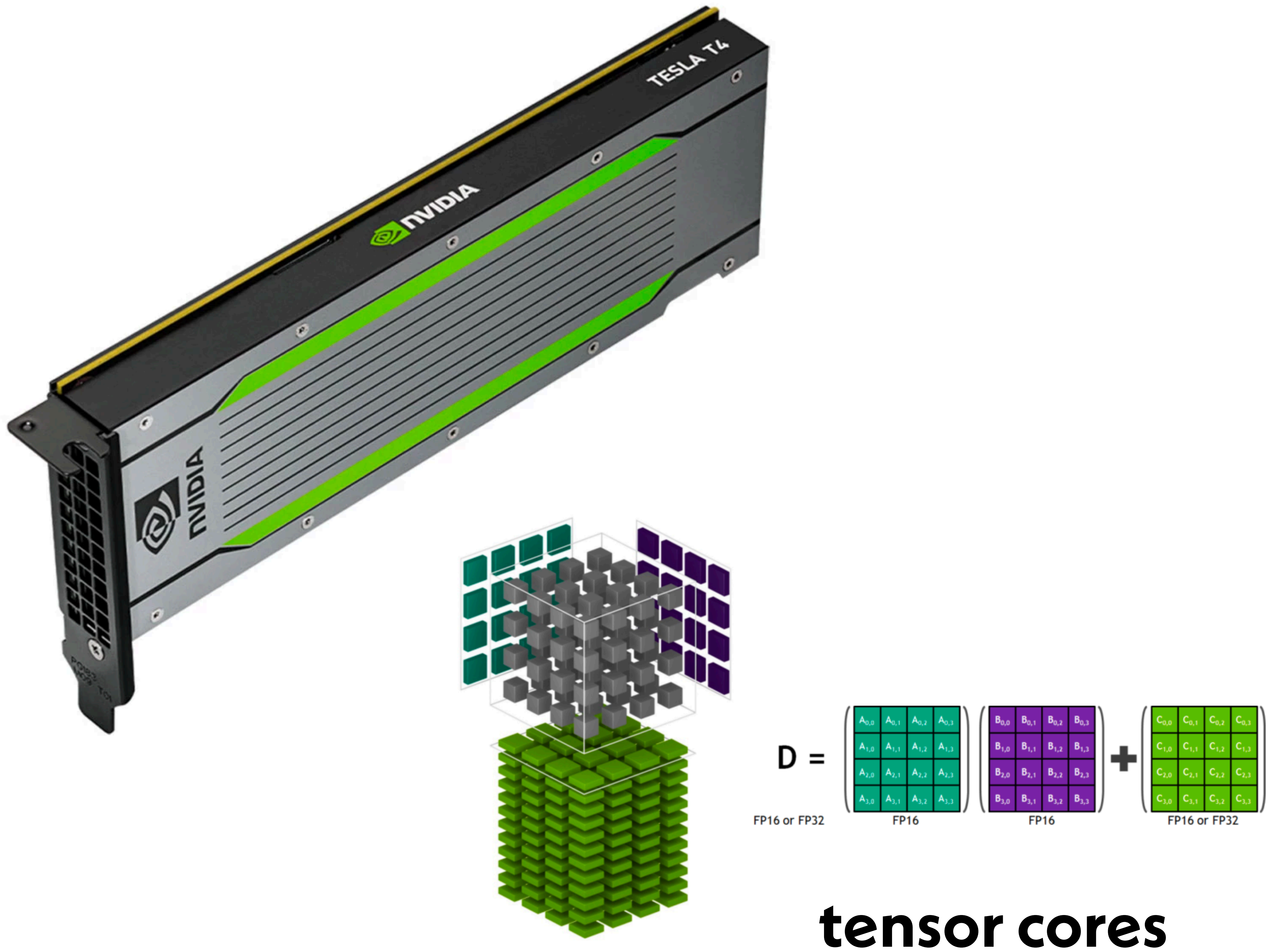
feature computation

NN computation

energy/force accumulation

We can speed this up with OpenMM GPU kernels using common pairlists, etc. (e.g. for ANI models)

TensorFlow/PyTorch do this efficiently, and hardware will keep getting better for this step



tensor cores

COMPUTATIONAL BOTTLENECKS IN CURRENT QML MODELS CAN BE SPED UP WITH CUSTOM GPU KERNELS

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	436	10.4	96.5 / 50.8
2P95	286	50	430	7.93	96.8 / 49.8
1HPO	198	64	547	9.12	101 / 44.6
1AJV	198	75	666	9.19	101 / 40.7

* 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, TensorFlow, JAX)
* Community-driven, package agnostic

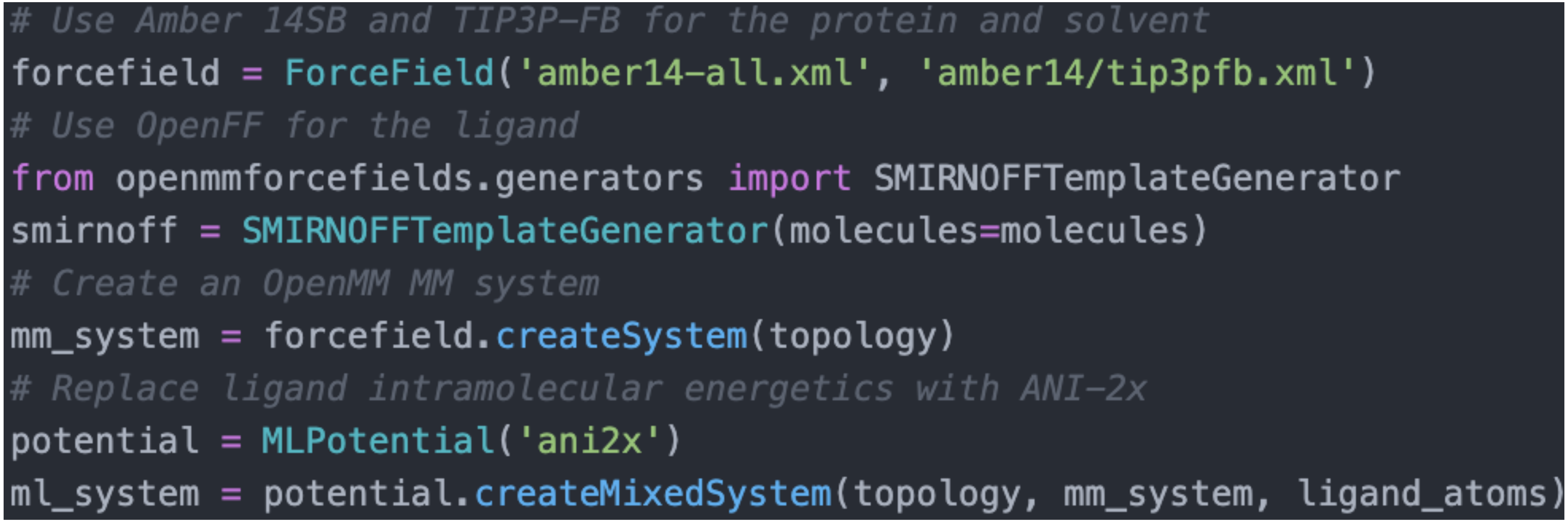
(~2.5x slower than GPU MD right now, but need 2x smaller timestep)

model distillation will become important in building single models that are efficient on hardware

paper: https://arxiv.org/abs/2201.08110

code: https://github.com/openmm/nnpops

OPENMM 8 WILL MAKE QML/MM SIMULATIONS INCREDIBLY EASY

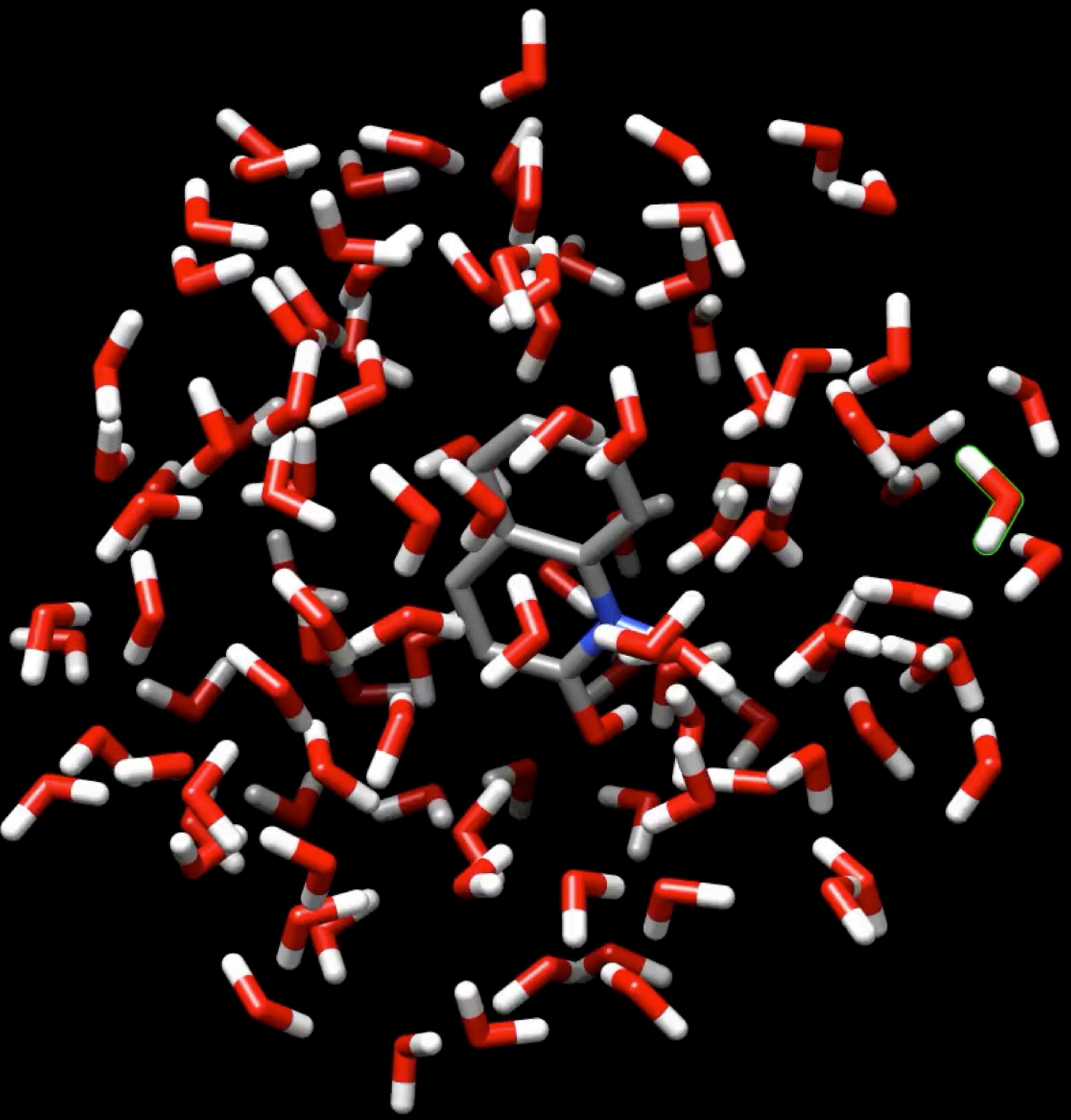


```
# Use Amber 14SB and TIP3P-FB for the protein and solvent
forcefield = ForceField('amber14-all.xml', 'amber14/tip3pfb.xml')
# Use OpenFF for the ligand
from openmmforcefields.generators import SMIRNOFFTemplateGenerator
smirnoff = SMIRNOFFTemplateGenerator(molecules=molecules)
# Create an OpenMM MM 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)
```

OpenMM 8 was just released!

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

WHY DO WE NEED MM AT ALL?



Can we just use ML force fields for everything?

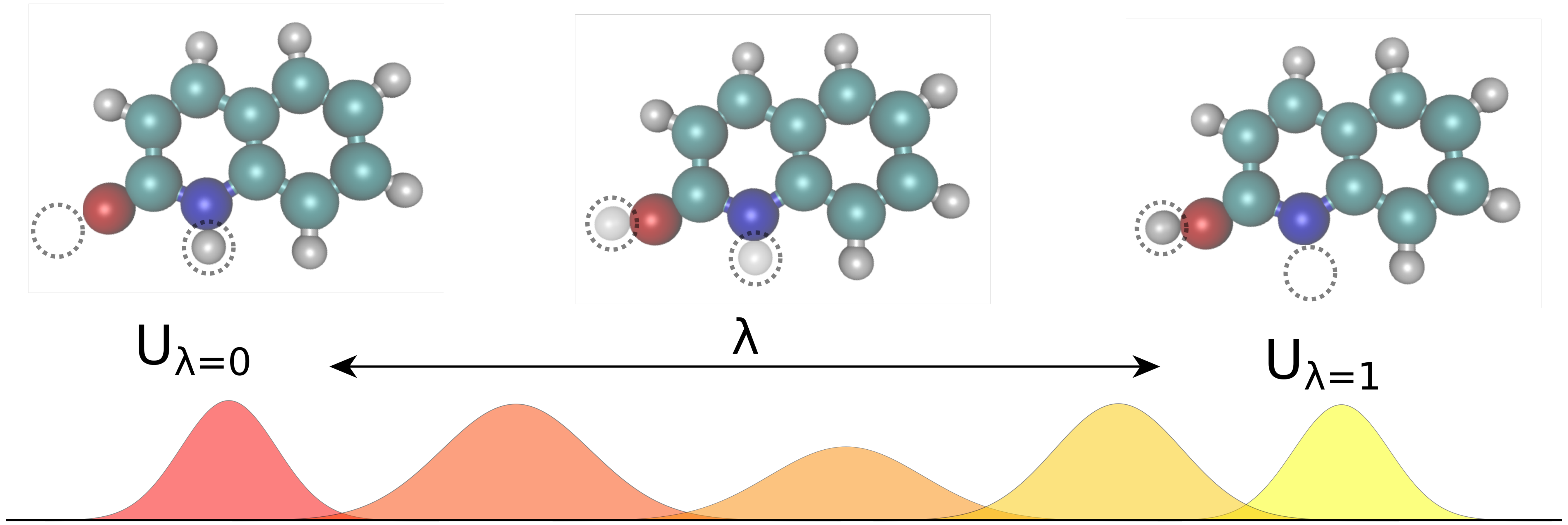
We can finally be free of the hegemony of bonds!

ANI2x

PURE QUANTUM MACHINE LEARNING (QML) POTENTIALS CAN BE USED TO COMPUTE FREE ENERGY DIFFERENCES BETWEEN CHEMICAL SPECIES

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

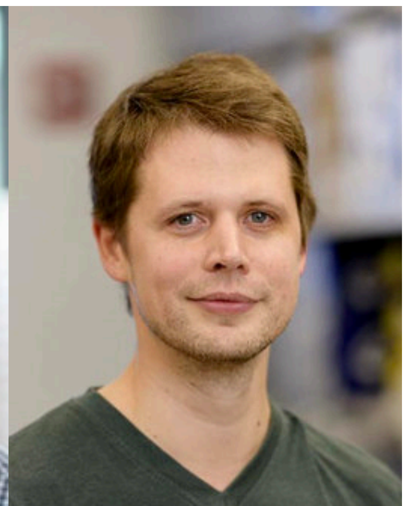
$$U(x;\lambda) = (1-\lambda)U_{\lambda=0}(x) + \lambda U_{\lambda=1}(x)$$



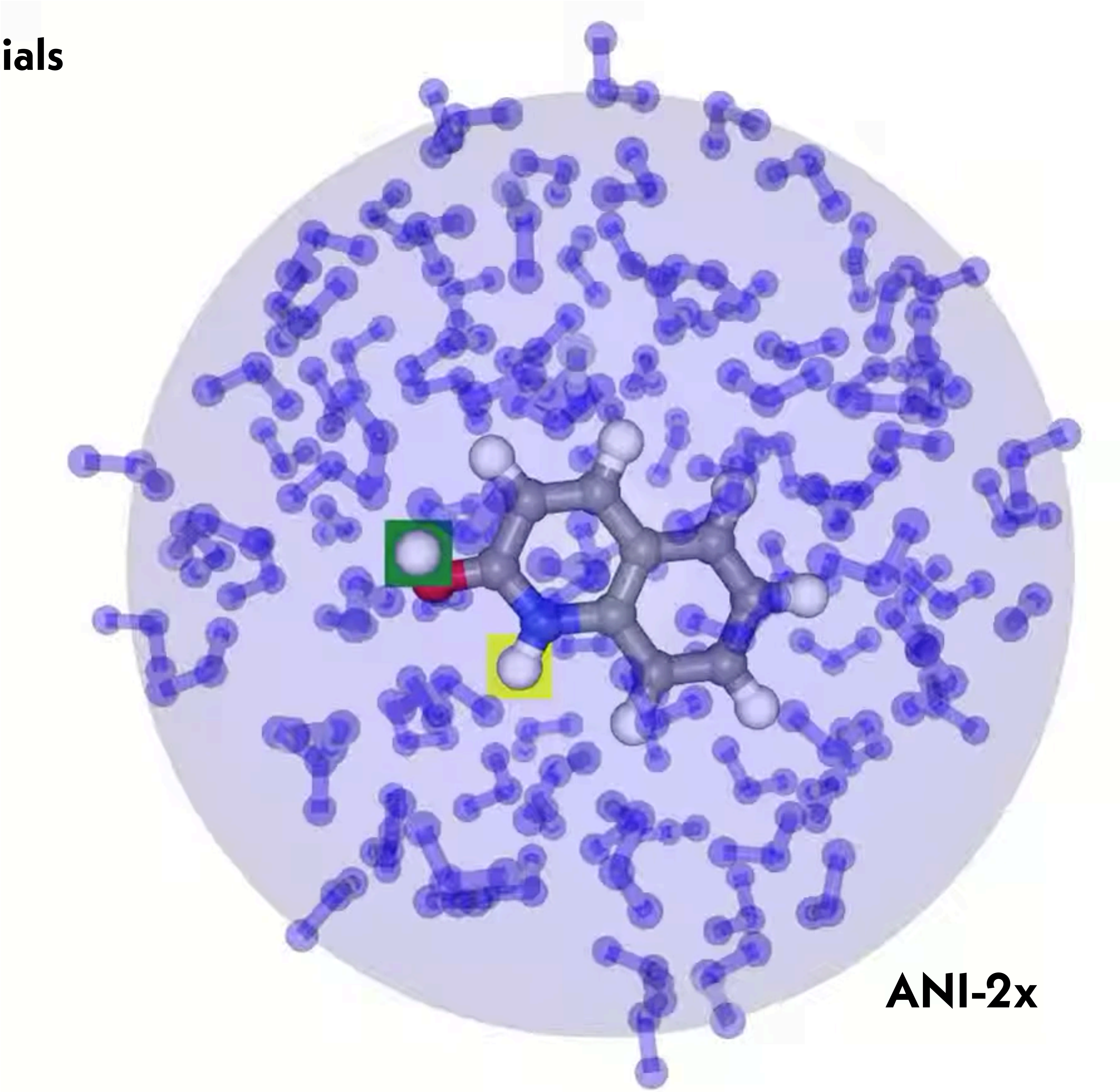
Simple restraints can be used when we need to enforce specific chemical species

JOSH FASS

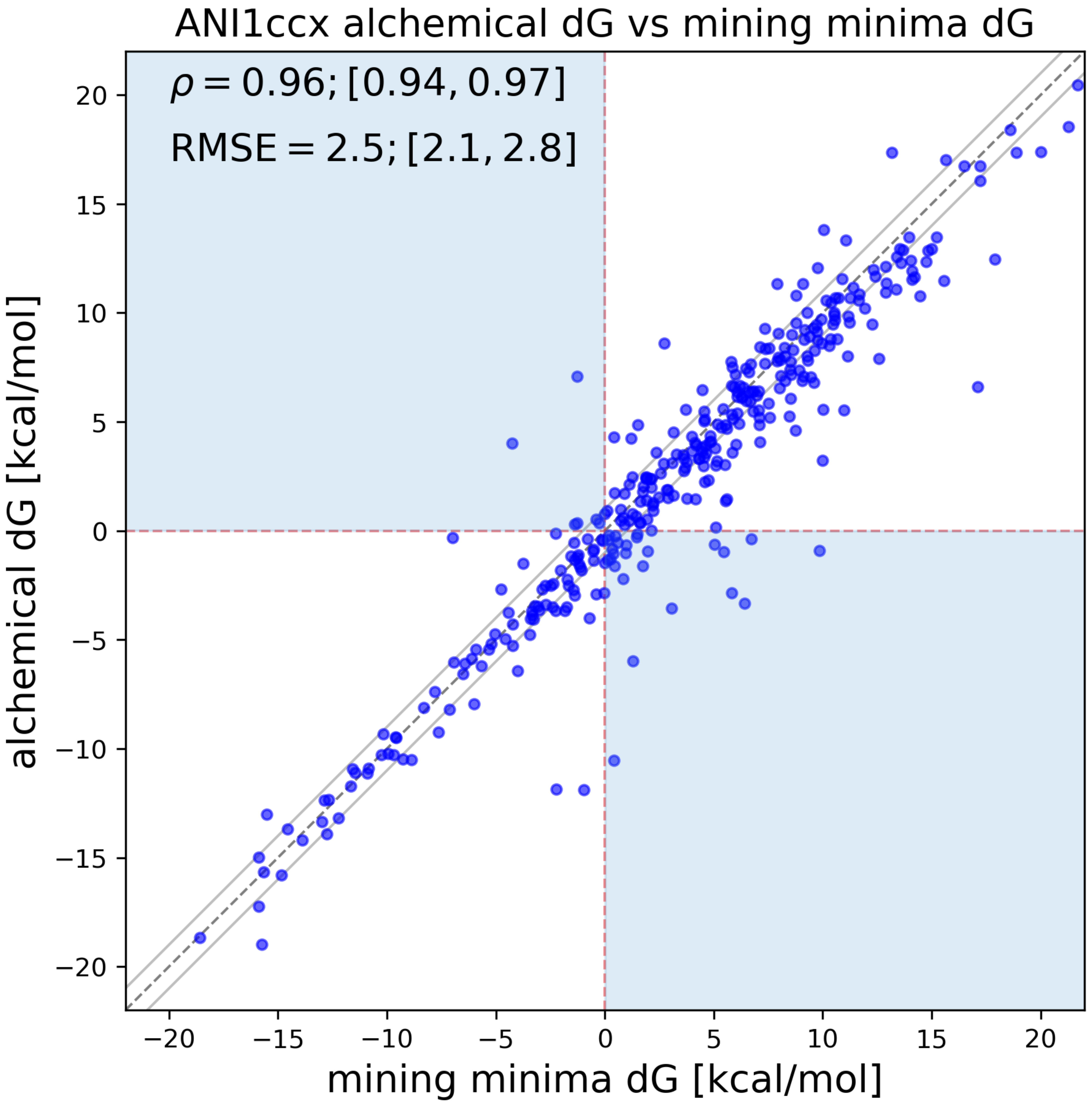
MARCUS WIEDER



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



STATISTICAL MECHANICS IS ESSENTIAL IN TAUTOMER RATIOS. EVEN IN VACUUM, ONLY SUMMING OVER MINIMA INTRODUCES HUGE ERRORS.



JOSH FASS

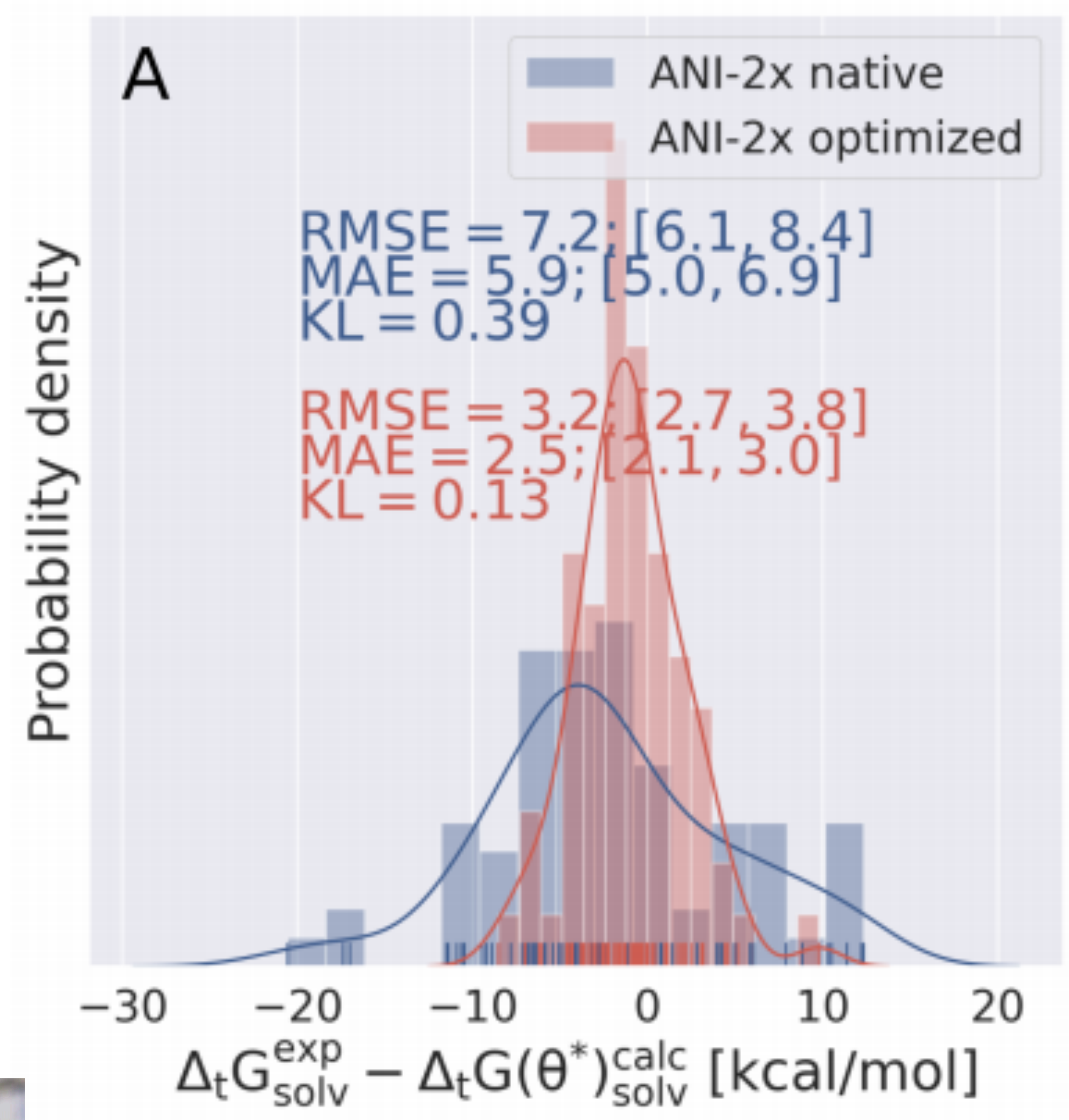
MARCUS WIEDER

OLEXANDR ISAYEV

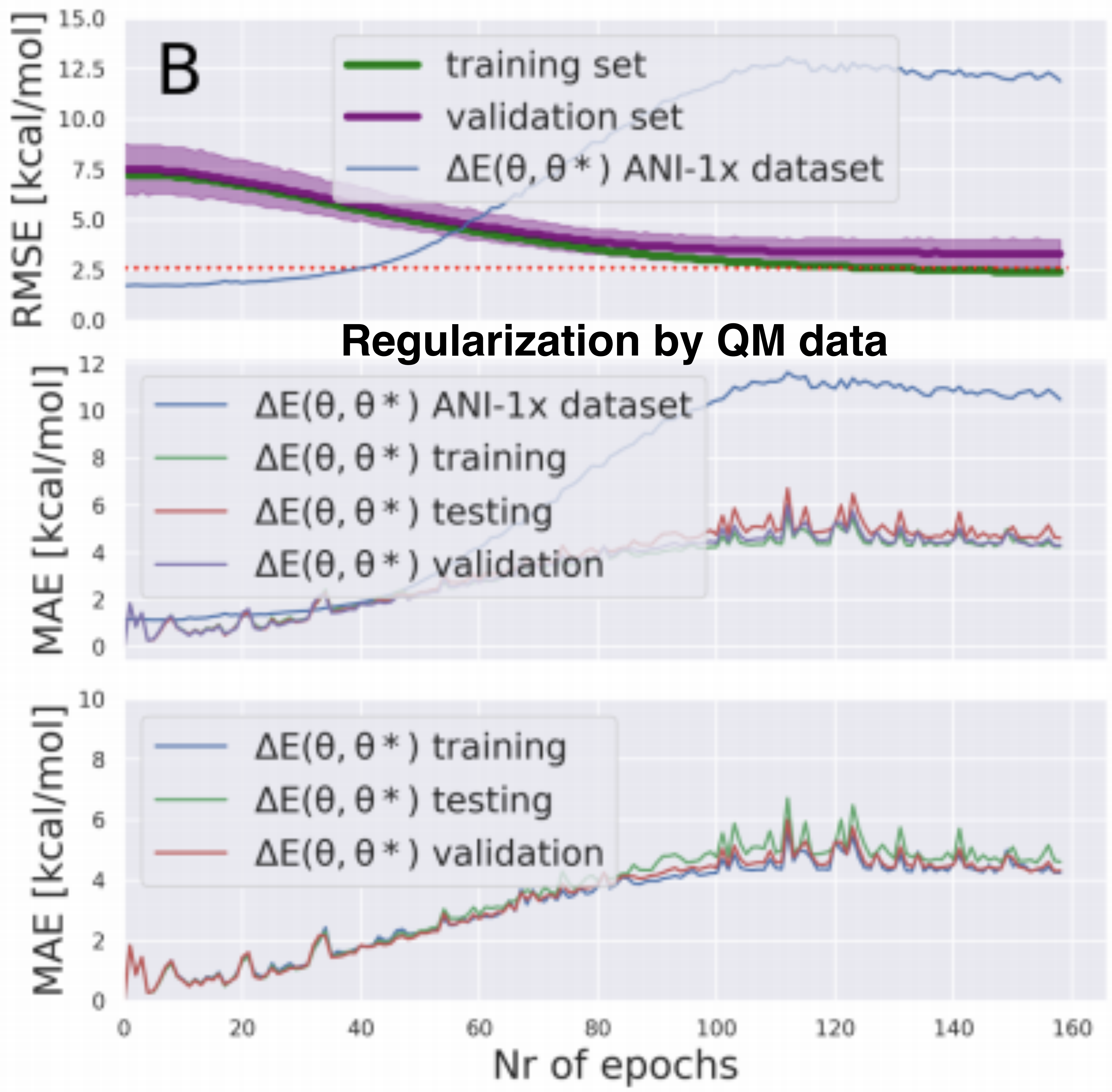
ADRIAN ROITBERG

PURE QUANTUM MACHINE LEARNING (QML) POTENTIALS CAN BE TUNED/RETRAINED BY FREE ENERGIES, REGULARIZED BY QM DATA

test set performance



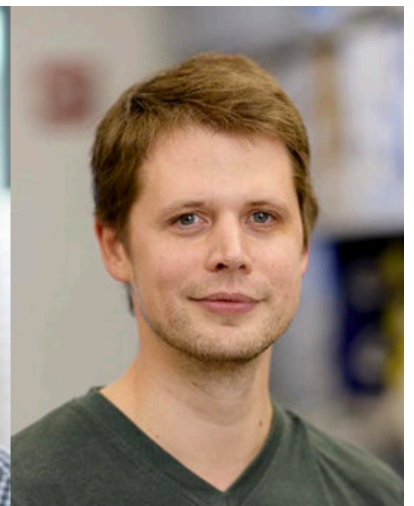
training / validation optimization



Regularization by QM data

JOSH FASS

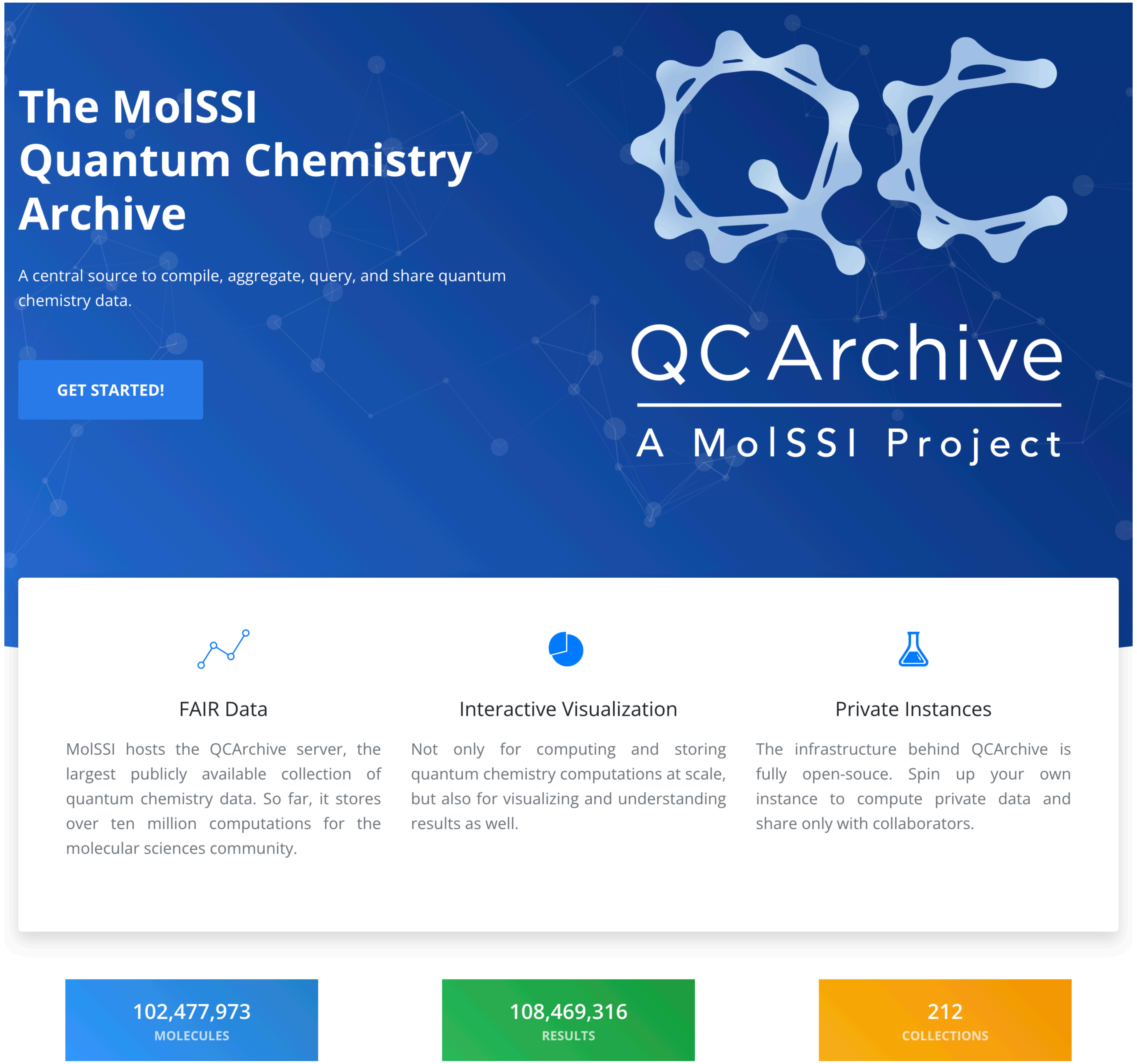
MARCUS WIEDER



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



The MolSSI Quantum Chemistry Archive

A central source to compile, aggregate, query, and share quantum chemistry data.

FAIR Data

MolSSI hosts the QCArchive server, the largest publicly available collection of quantum chemistry data. So far, it stores over ten million computations for the molecular sciences community.

Interactive Visualization

Not only for computing and storing quantum chemistry computations at scale, but also for visualizing and understanding results as well.

Private Instances

The infrastructure behind QCArchive is fully open-souce. Spin up your own instance to compute private data and share only with collaborators.

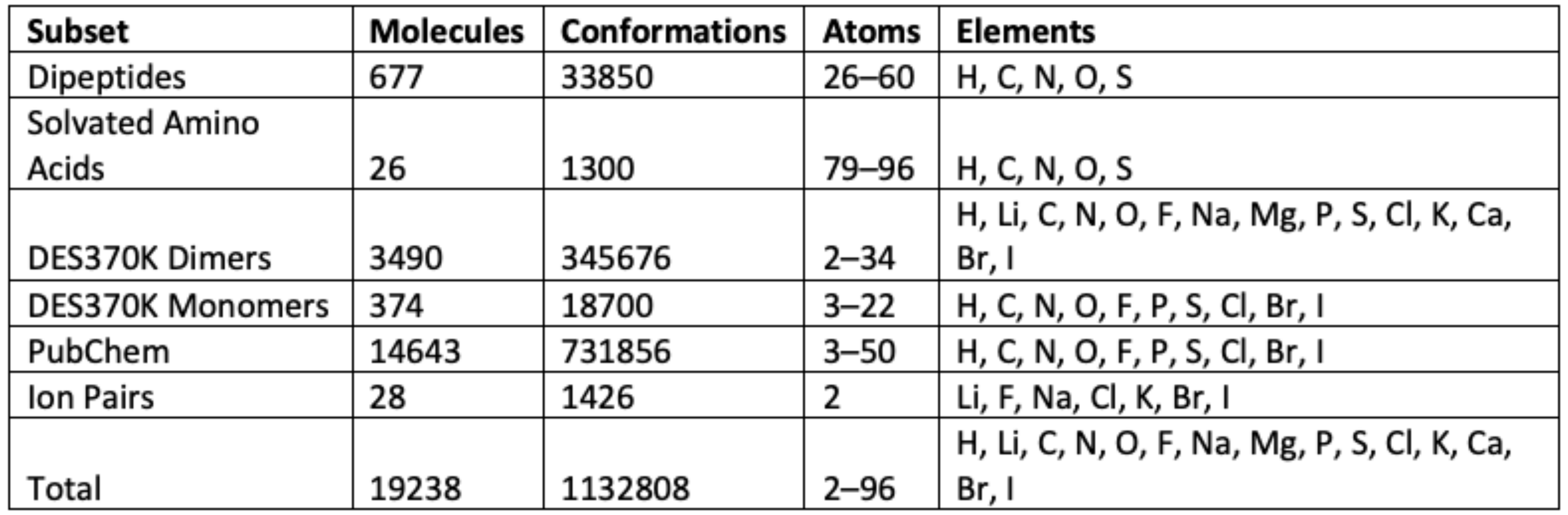
102,477,973 MOLECULES

108,469,316 RESULTS

212 COLLECTIONS

http://qcarchive.molssi.org

OpenMM and the Open Force Field Initiative are working closely with MolSSI to expand the QCArchive to support the construction of next-generation machine learning force fields



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

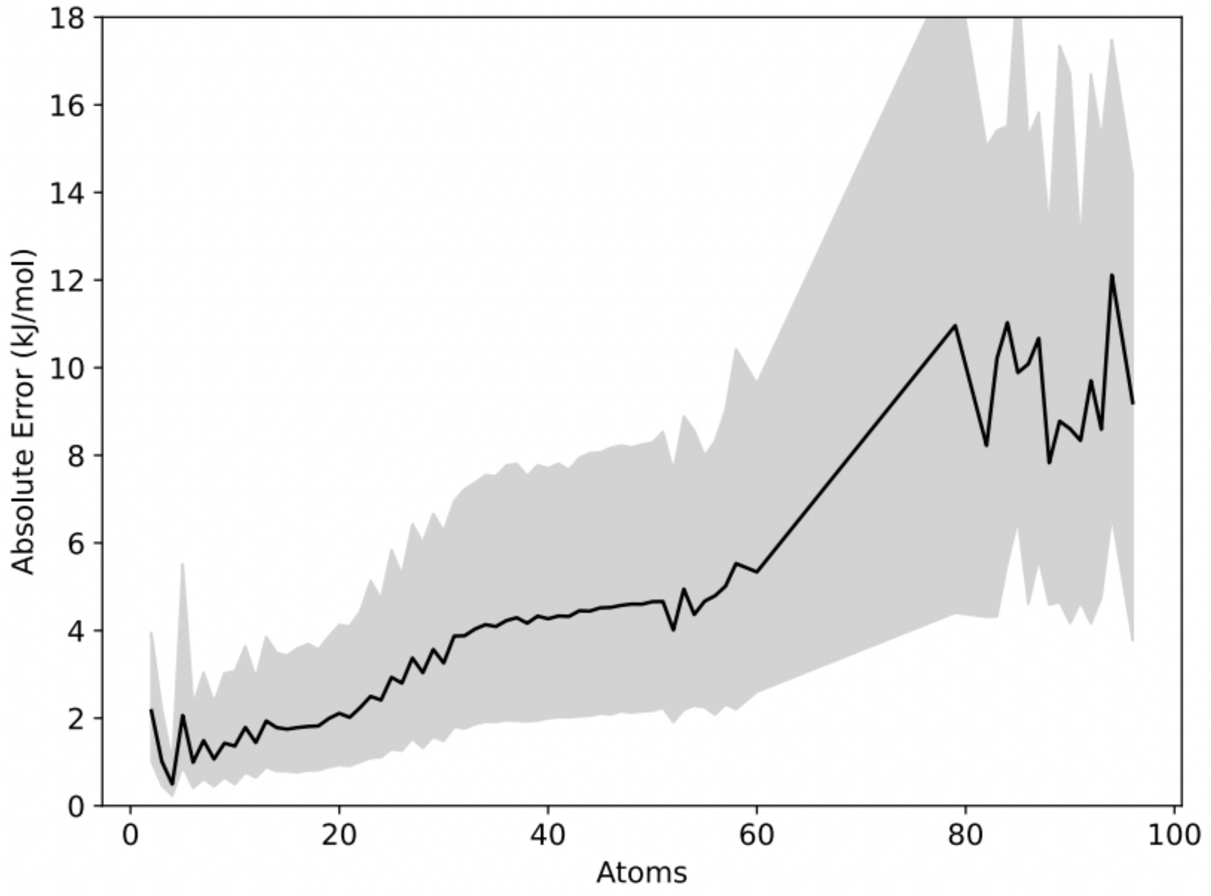


Figure 5. Absolute error as a function of the total number of atoms in a molecule. The line indicates the median for all molecules of a certain size, and the gray region contains the central 50% of samples.

SPICE QML model: 0.7 kcal/mol

median absolute error

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

SPICE IS OUR FIRST STEP TOWARD BUILDING "FOUNDATION MODELS" THAT CAN BE RAPIDLY TAILORED TO DIFFERENT APPLICATIONS

🌶️ + 🍹

Dataset

QC specification	Dataset	Category	# Mols	# Conformations
Openff-default	OpenFF Gen2 Optimization	Small molecules	1022	244944
	OpenFF PepConf Optimization	Di-, Tri-peptides	522	228582
	RNA-BGSU Diverse Dataset	RNA trinucleotides	64	3649
	RNA-BGSU Trinucleotide Dataset	RNA trinucleotides	64	35134
	SPICE Pubchem	Small molecules	14110	601719
	SPICE Dipeptide	Dipeptides	677	25098
	SPICE DES Monomers	Small molecules	369	18450

Name	Dataset	Mols	Conformations	Energy RMSE (kcal/mol) Force RMSE (kcal/mol·Å^{-1})		Baseline FF Energy RMSE (kcal/mol) (Test molecules) Baseline FF Force RMSE (kcal/mol·Å^{-1}) (Test molecules)			
				Train	Test	GAFF-1.81	GAFF-2.11	OpenFF-1.2.0	OpenFF-2.0.0
Joint	Gen2	1022	244944	$1.13_{1.11}^{1.16}$ $90.41_{86.35}^{94.92}$	$\mathbf{1.77}_{\mathbf{1.70}}^{\mathbf{1.85}}$ $35.66_{26.51}^{44.16}$	$2.97_{2.91}^{3.03}$ $10.68_{10.65}^{10.70}$	$2.96_{2.90}^{3.03}$ $10.65_{10.62}^{10.68}$	$2.82_{2.76}^{2.90}$ $11.11_{11.08}^{11.14}$	$2.69_{2.61}^{2.77}$ $\mathbf{10.33}_{\mathbf{10.30}}^{\mathbf{10.36}}$
	PepConf	522	228582	$1.61_{1.58}^{1.63}$ $6.17_{6.17}^{6.17}$	$\mathbf{1.97}_{\mathbf{1.92}}^{\mathbf{2.03}}$ $\mathbf{6.27}_{\mathbf{6.26}}^{\mathbf{6.27}}$	$3.64_{3.60}^{3.68}$ $192.61_{65.86}^{300.66}$	$4.61_{4.55}^{4.66}$ $78.83_{33.44}^{124.76}$	$3.09_{3.04}^{3.14}$ $13.67_{10.49}^{17.91}$	$3.23_{3.18}^{3.28}$ $45.45_{18.05}^{69.14}$
	SPICE-Pubchem	14110	601719	$2.39_{2.38}^{2.40}$ $9.10_{9.06}^{9.16}$	$\mathbf{2.68}_{\mathbf{2.65}}^{\mathbf{2.71}}$ $\mathbf{9.40}_{\mathbf{9.35}}^{\mathbf{9.46}}$	$4.44_{4.40}^{4.48}$ $14.62_{14.59}^{14.65}$	$4.62_{4.58}^{4.66}$ $15.16_{15.11}^{15.21}$	$4.28_{4.24}^{4.32}$ $14.82_{14.80}^{14.85}$	$4.32_{4.28}^{4.36}$ $14.66_{14.64}^{14.69}$
	SPICE-Dipeptide	677	25098	$2.48_{2.45}^{2.51}$ $6.64_{6.63}^{6.65}$	$\mathbf{2.46}_{\mathbf{2.38}}^{\mathbf{2.55}}$ $\mathbf{6.39}_{\mathbf{6.36}}^{\mathbf{6.43}}$	$4.16_{4.03}^{4.29}$ $12.76_{12.70}^{12.81}$	$4.15_{4.02}^{4.27}$ $12.67_{12.62}^{12.73}$	$4.04_{3.91}^{4.17}$ $12.41_{12.36}^{12.46}$	$3.83_{3.70}^{3.96}$ $12.72_{12.67}^{12.77}$
	SPICE-DES-Monomers	369	18450	$1.37_{1.35}^{1.40}$ $7.12_{7.07}^{7.17}$	$\mathbf{1.70}_{\mathbf{1.58}}^{\mathbf{1.84}}$ $\mathbf{8.69}_{\mathbf{8.57}}^{\mathbf{8.82}}$	$2.84_{2.61}^{3.08}$ $13.77_{13.51}^{14.07}$	$2.75_{2.52}^{2.97}$ $13.35_{13.11}^{13.58}$	$2.99_{2.80}^{3.20}$ $14.31_{14.02}^{14.63}$	$3.06_{2.84}^{3.29}$ $14.88_{14.59}^{15.19}$
	RNA-Diverse	64	3649	$3.35_{3.23}^{3.46}$ $17.71_{17.66}^{17.77}$	$\mathbf{3.76}_{\mathbf{3.40}}^{\mathbf{4.09}}$ $17.90_{17.76}^{18.05}$	$6.67_{6.18}^{7.16}$ $\mathbf{16.32}_{\mathbf{16.23}}^{\mathbf{16.42}}$	$6.77_{6.32}^{7.26}$ $17.27_{17.17}^{17.37}$	$6.47_{5.99}^{6.96}$ $18.52_{18.38}^{18.67}$	$6.50_{5.97}^{7.04}$ $18.58_{18.43}^{18.73}$
	RNA-Trinucleotide	64	35134	$3.04_{3.01}^{3.08}$ $17.72_{17.70}^{17.73}$	$\mathbf{3.53}_{\mathbf{3.42}}^{\mathbf{3.64}}$ $17.93_{17.88}^{17.98}$	$6.10_{5.93}^{6.26}$ $\mathbf{16.44}_{\mathbf{16.41}}^{\mathbf{16.47}}$	$6.12_{5.95}^{6.30}$ $17.42_{17.39}^{17.45}$	$6.21_{6.04}^{6.38}$ $18.68_{18.63}^{18.73}$	$6.19_{6.03}^{6.36}$ $18.75_{18.70}^{18.80}$

KEN TAKABA

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

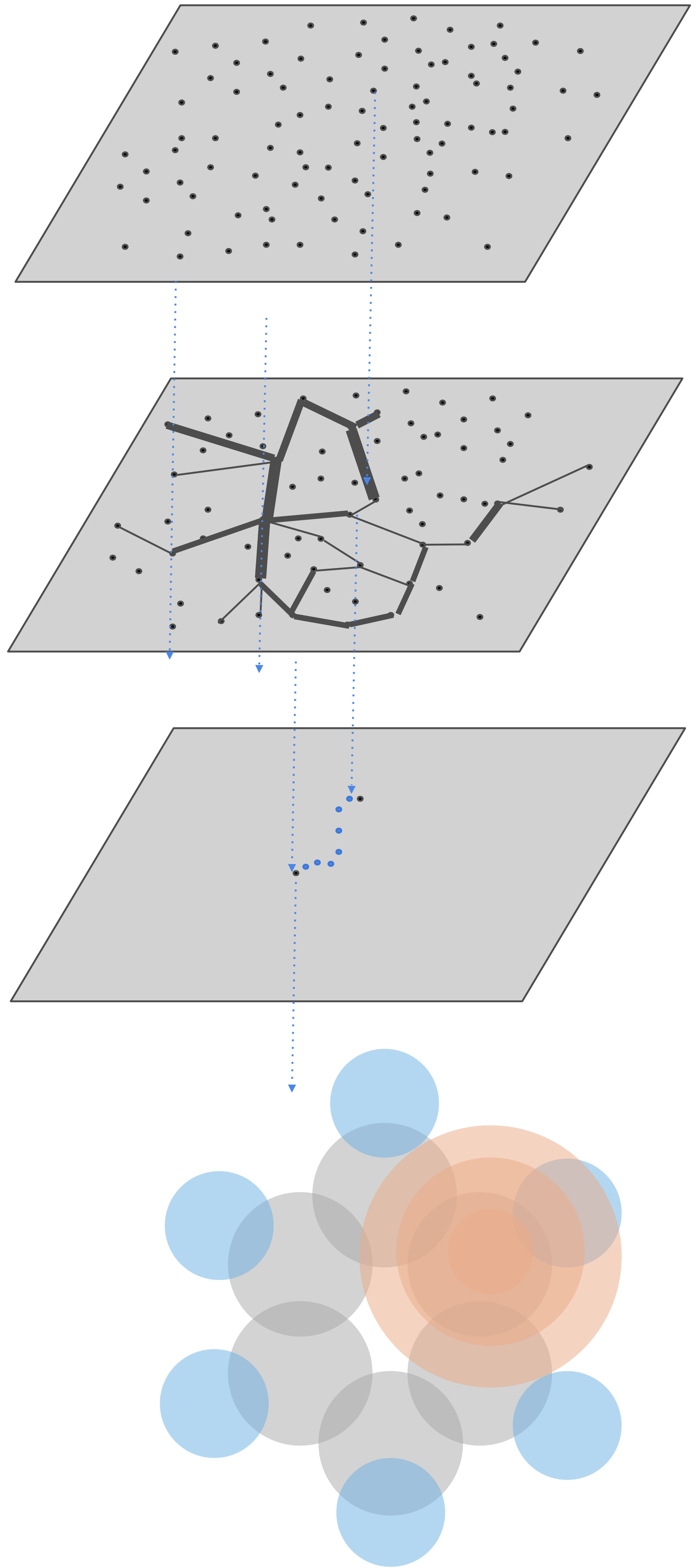
week 2

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

using **new** model tuned to target

from first week's data

HYBRID PHYSICAL / MACHINE LEARNING MODELS COULD DRIVE A NEW ERA OF PRODUCTIVITY IN COMPUTATIONAL CHEMISTRY



- Fast, structure-based **machine learning surrogates** assess designs over vast synthetic chemical spaces prioritize useful calculations
- Adaptive allocation of effort to alchemical free energy calculations guided by **machine learning cost prediction**s
- **Machine learned optimal alchemical transformations** produce faster estimates of free energy differences more cheaply
- **Learnable machine learning potentials** fit to experimental free energy and quantum chemical data produce higher accuracy predictions

Chemical Science 2022 https://doi.org/10.1039/D2SC02739A
bioRxiv https://www.biorxiv.org/content/10.1101/2021.08.24.457513v2

PREPRINTS AND CODE

gimlet: graph convolutional networks for partial charge assignment
preprint: https://arxiv.org/abs/1909.07903
code: http://github.com/choderalab/gimlet

espaloma: end-to-end differentiable assignment of force field parameters
preprint: https://arxiv.org/abs/2010.01196
code: https://github.com/choderalab/espaloma

qmlify: hybrid QML/MM alchemical free energy calculations for protein-ligand binding
preprint: https://doi.org/10.1101/2020.07.29.227959
code: https://github.com/choderalab/qmlify

neutromeratio: alchemical free energy calculations with fully QML potentials for tautomer ratio prediction
preprint: https://doi.org/10.1101/2020.10.24.353318
code: https://github.com/choderalab/neutromeratio

CHODERA LAB

National Institutes of Health

NIH

NSF

OpenEye SCIENTIFIC

SILICON Therapeutics

PARKER INSTITUTE for CANCER IMMUNOTHERAPY

STIFTUNG CHARITÉ

EINSTEIN Foundation.de

Gerstner FAMILY FOUNDATION

STARR CANCER CONSORTIUM

VIR

Open Force Field Consortium

BAYER

XtalPi

CYCLE FOR SURVIVAL

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