slides: http://choderalab.org/news

THE FUTURE OF FREE ENERGY: CALCULATIONS THAT CAN LEARN FROM EXPERIMENT

John D. Chodera
MSKCC Computational and Systems Biology Program
Sildes will be posted to http://www.choderalab.org/news

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

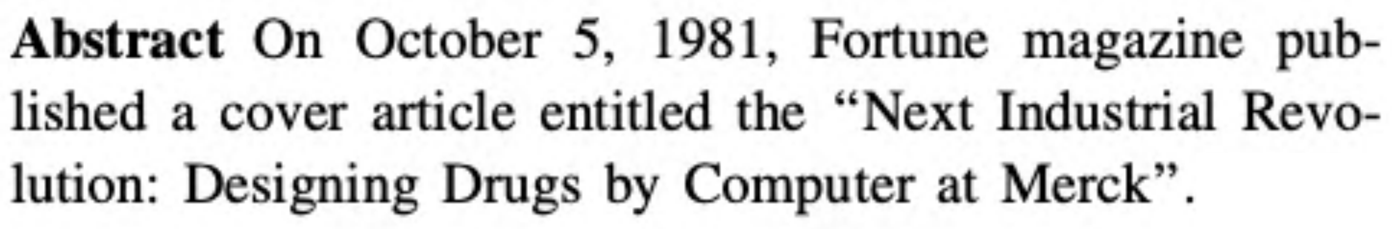
21 Dec 2022 - ACS San Diego - Kate Holloway Award Symposium

CONGRATULATIONS, KATE!

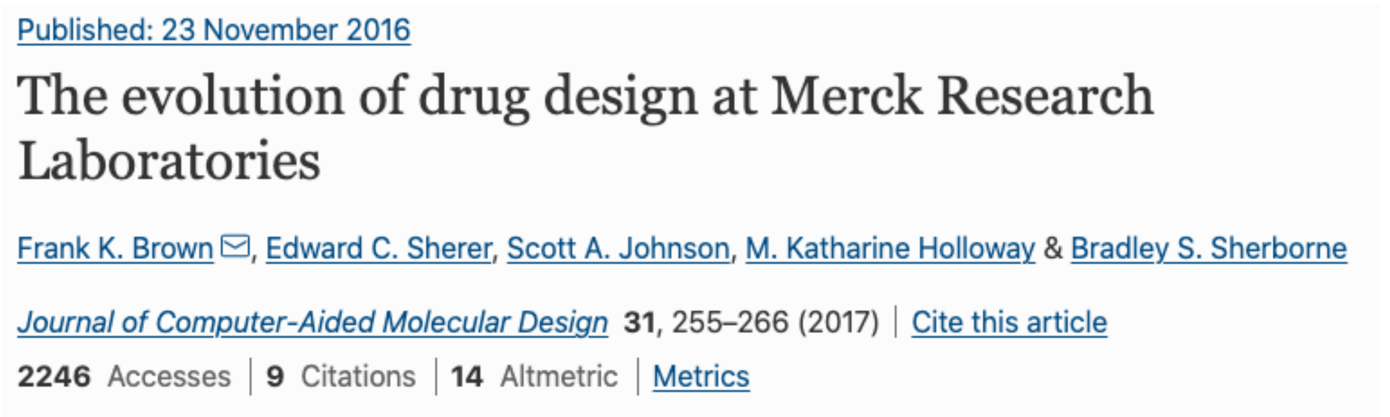
KATHERINE HOLLOWAY

A CAREER OF INSPIRATIONAL AND VISIONARY WORK

WHAT WILL IT TAKE FOR COMPUTATIONAL CHEMISTRY TO DRIVE DISCOVERY PROGRAMS?



Abstract On October 5, 1981, Fortune magazine published a cover article entitled the “Next Industrial Revolution: Designing Drugs by Computer at Merck”.



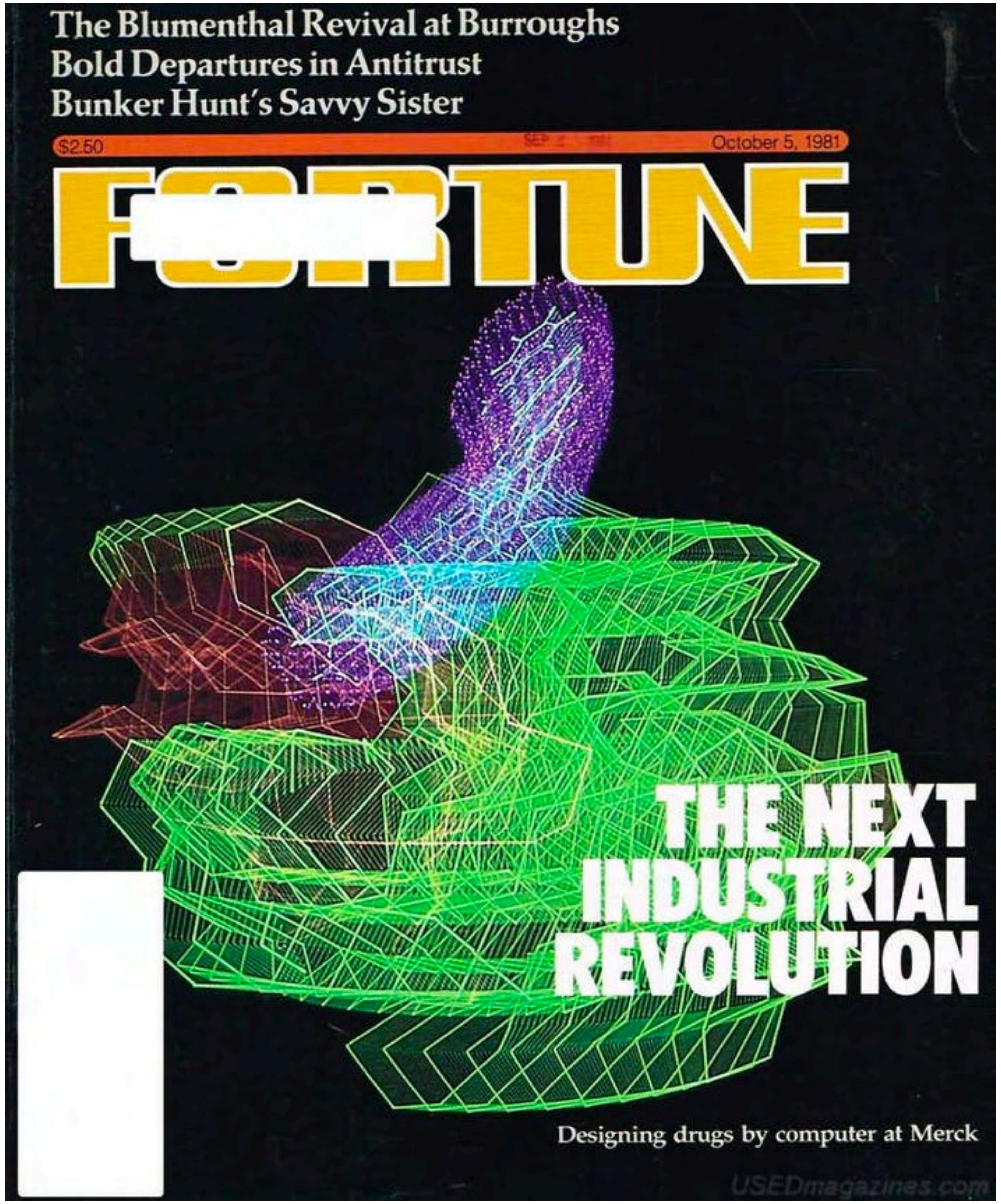
Published: 23 November 2016

The evolution of drug design at Merck Research Laboratories

Frank K. Brown, Edward C. Sherer, Scott A. Johnson, M. Katharine Holloway & Bradley S. Sherborne

Journal of Computer-Aided Molecular Design **31**, 255–266 (2017) | Cite this article

2246 Accesses | **9** Citations | **14** Altmetric | Metrics


The Blumenthal Revival at Burroughs
Bold Departures in Antitrust
Bunker Hunt's Savvy Sister
$2.50
October 5, 1981
FORTUNE
THE NEXT INDUSTRIAL REVOLUTION
Designing drugs by computer at Merck
USEDmagazines.com

5 Oct 1981

WE'RE FACING COMPLEX MULTI-OBJECTIVE DESIGN PROBLEMS

Ed Griffen

Medchemica

Target Product Profile (TPP) for oral SARS-CoV-2 main viral protease (Mpro) inhibitor

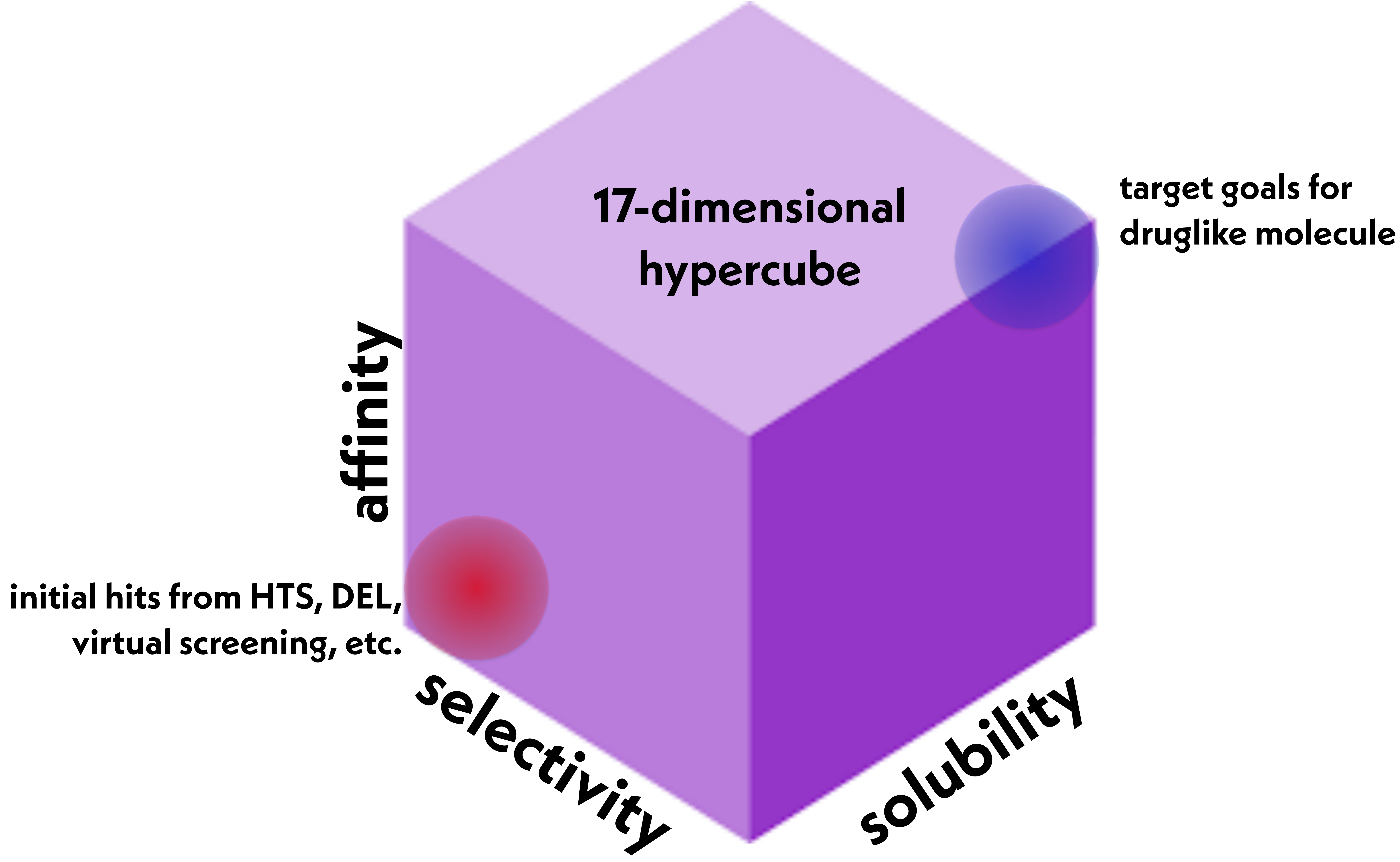
Property	Target range	Rationale
protease assay	$IC_{50} < 10$ nM	Extrapolation from other anti-viral programs
viral replication assay	$EC_{50} < 5$ µM	Suppression of virus at achievable blood levels
plaque reduction assay	$EC_{50} < 5$ µM	Suppression of virus at achievable blood levels
route of administration	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

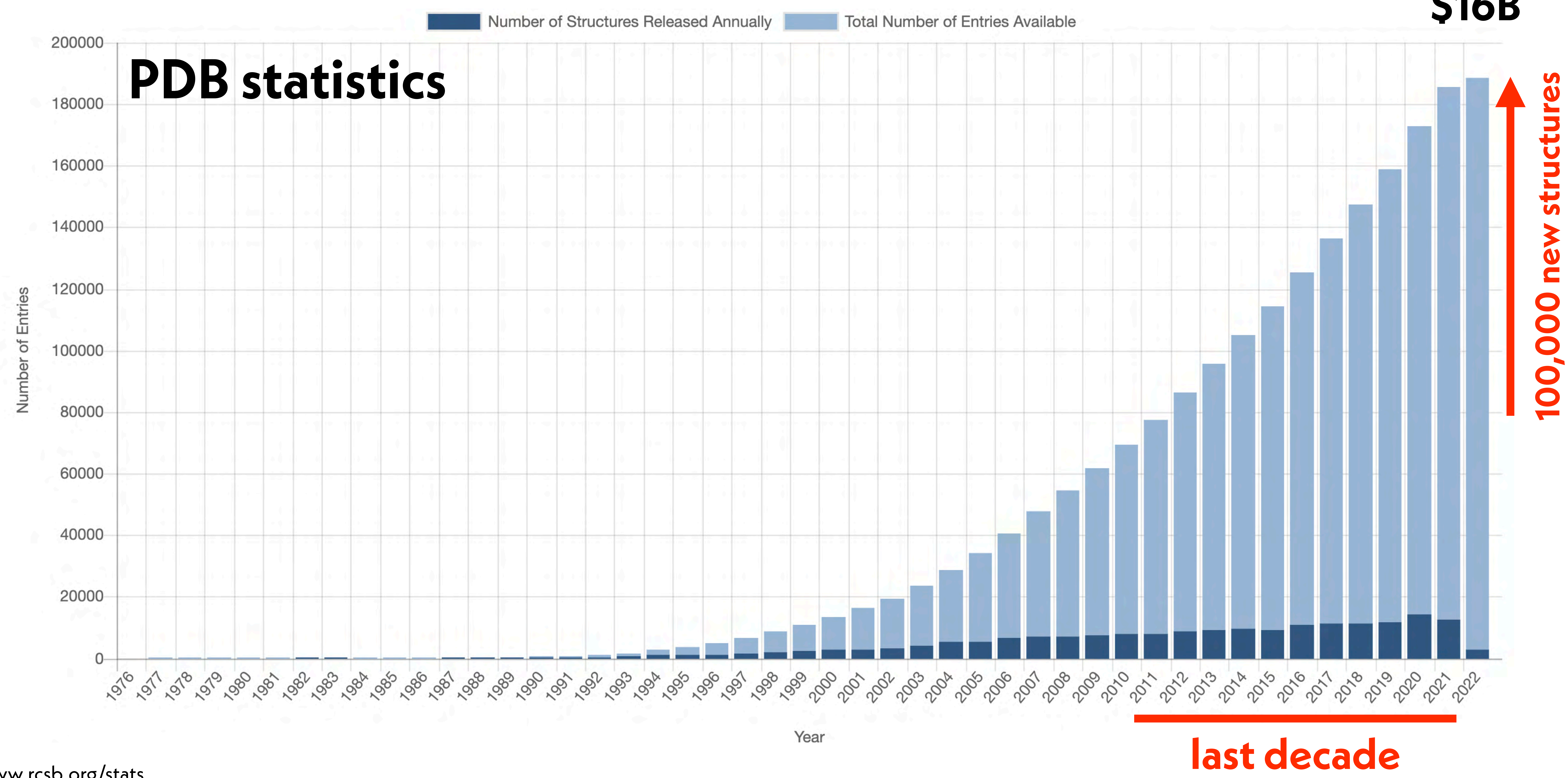
An international effort to
DISCOVER A COVID ANTIVIRAL

https://covid.postera.ai/covid

WE'RE FACING COMPLEX MULTI-OBJECTIVE DESIGN PROBLEMS


17-dimensional hypercube
target goals for druglike molecule
affinity
initial hits from HTS, DEL, virtual screening, etc.
selectivity
solubility

WE CAN LEVERAGE AN ENORMOUS AMOUNT OF STRUCTURAL DATA

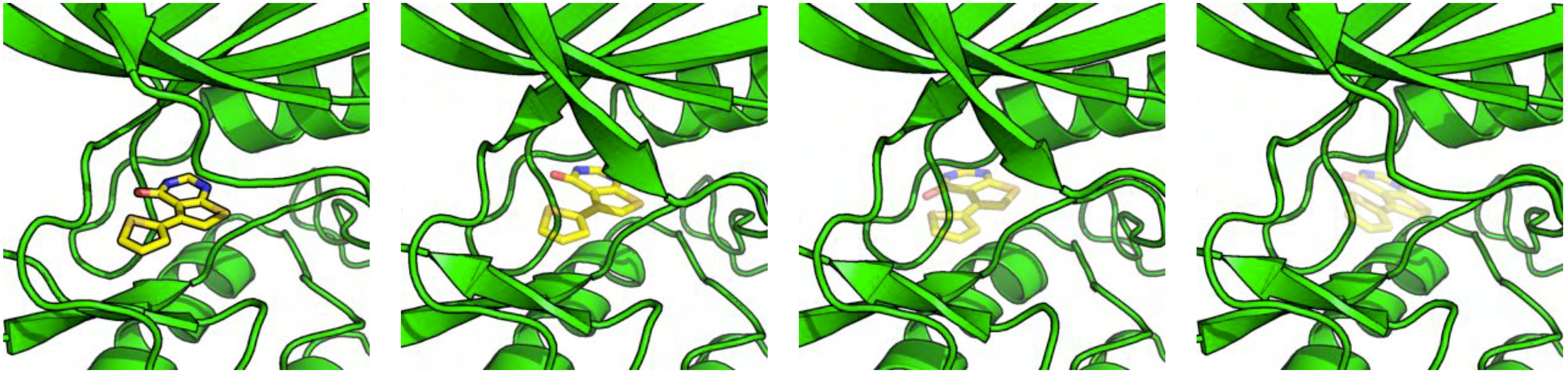


http://www.rcsb.org/stats

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

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

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

$$\Delta G_{1\to N} = -\beta^{-1} \ln \frac{Z_N}{Z_1} = -\beta^{-1} \ln \frac{Z_2}{Z_1} \cdot \frac{Z_3}{Z_2} \cdots \frac{Z_N}{Z_{N-1}}$$

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

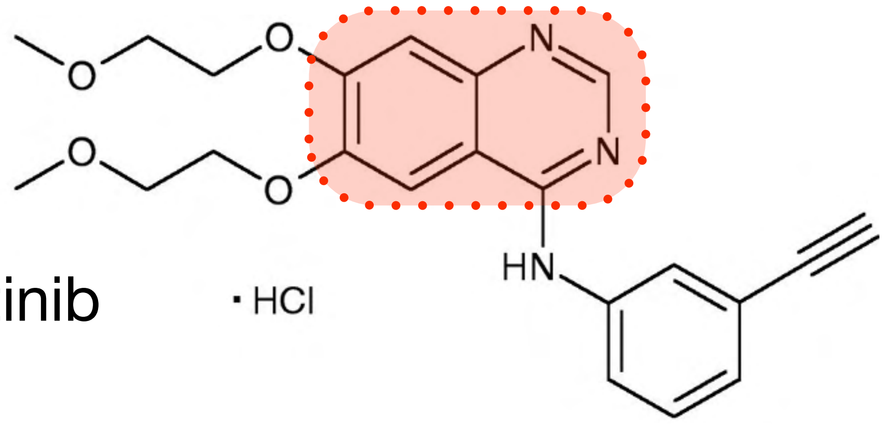
Pioneering work from many: McCammon, van Gunsteren, Kollman, Jorgensen, Chipot, Roux, Boresch, Fujitani, Pande, Shirts, Swope, Christ, Mobley, Schrödinger, and many more

ALCHEMICAL FREE ENERGY CALCULATIONS COME IN TWO FLAVORS: RELATIVE AND ABSOLUTE

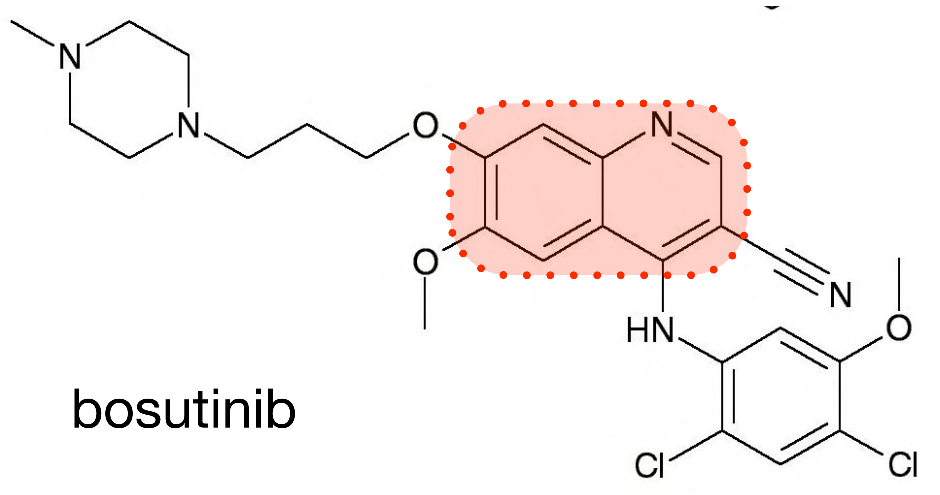
RELATIVE

$$\begin{tikzcd}
P + L_1 \arrow[r] \arrow[d] \arrow[dr, phantom, "\Delta\Delta G"] & PL_1 \\
P + L_2 \arrow[r] & PL_2 \arrow[u]
\end{tikzcd}$$

capable of **transforming a few atoms**
good for comparing **similar ligands**
requires same or **similar scaffolds**
requires **common scaffold to anchor series**



erlotinib



bosutinib

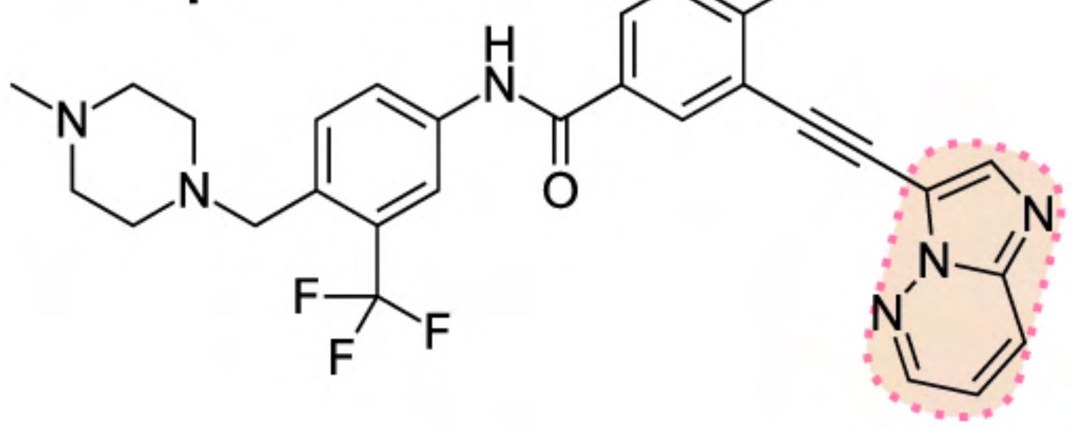
Cournia, Allen, Sherman 2017: http://dx.doi.org/10.1021/acs.jcim.7b00564

ABSOLUTE

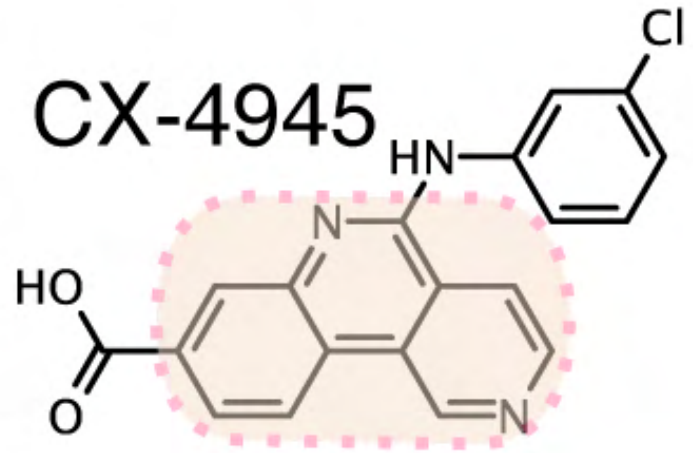
$$\begin{tikzcd}
P + L \arrow[r, "\Delta G_{\text{bind}}"] \arrow[d] & PL \\
P + ø \arrow[r] & Pø \arrow[u]
\end{tikzcd}$$

capable of **disappearing a few atoms**
good for comparing **dissimilar ligands**
can use entirely **disparate scaffolds**
requires use of **restraints to anchor ligand**

ponatinib



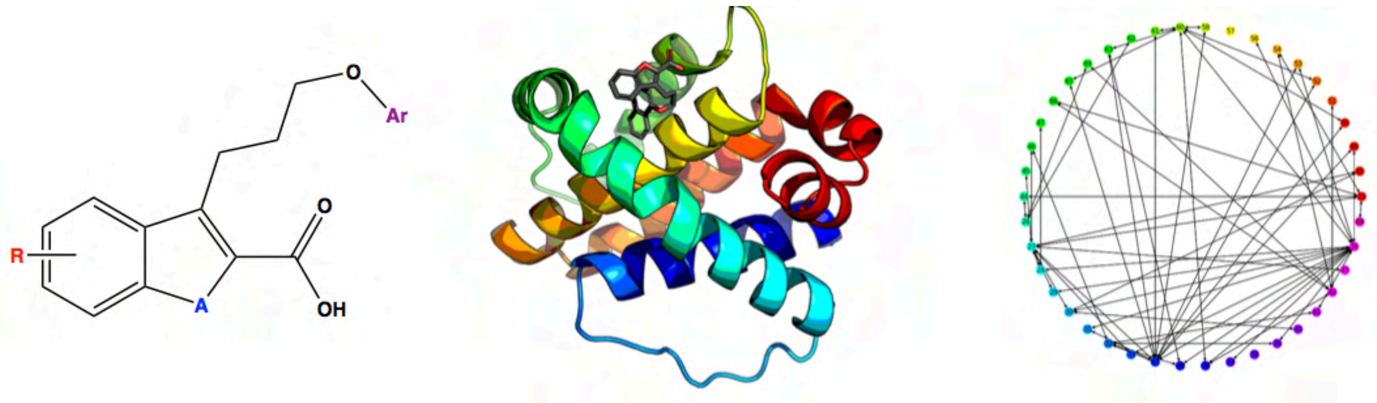
CX-4945



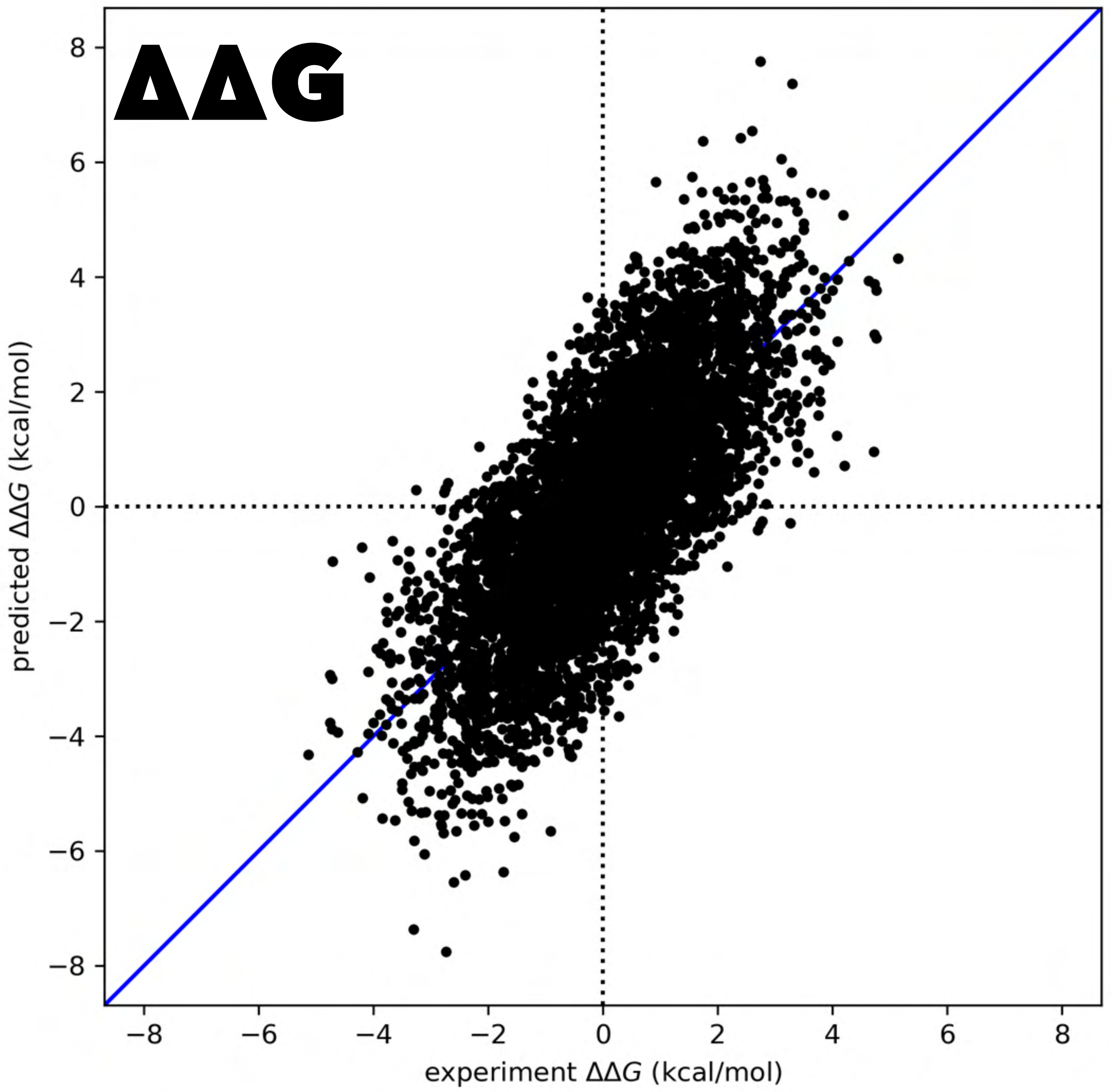
Aldeghi, Bluck, Biggin 2018: https://doi.org/10.1007/978-1-4939-7756-7_11

USEFUL ACCURACY IS SOMETIMES ACHIEVABLE

RELATIVE



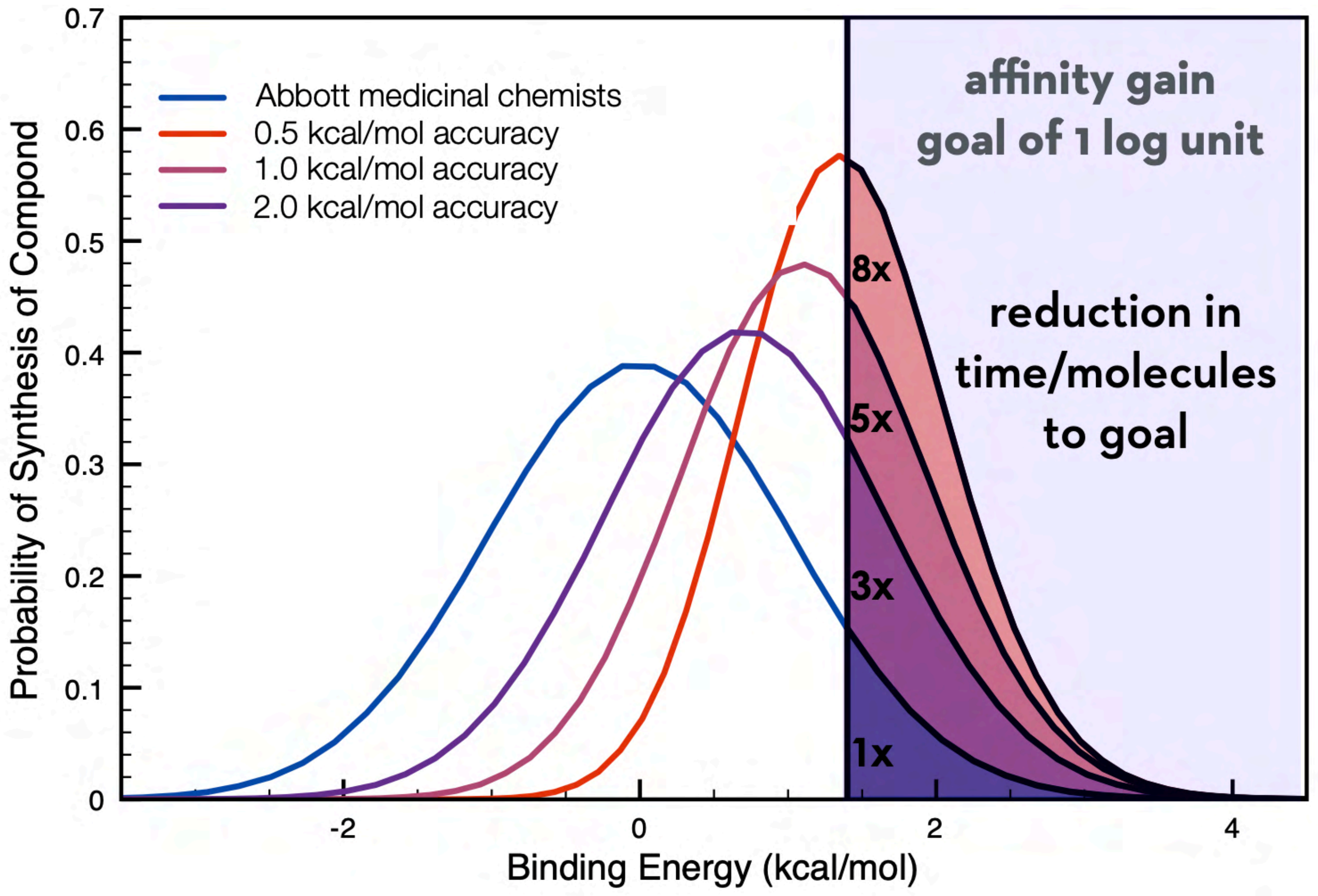
```
                 all within-target pairs ΔΔG (N = 5620)
RMSE: OPLS       1.37 [95%:   1.34,   1.39] kcal/mol
MUE : OPLS       1.09 [95%:   1.07,   1.11] kcal/mol
R2  : OPLS       0.10 [95%:   0.06,   0.15] kcal/mol
rho : OPLS       0.73 [95%:   0.72,   0.74] kcal/mol
```



Wang et al. (Schrödinger) JACS 137:2695, 2015
https://doi.org/10.1021/ja512751q
Reanalysis: http://github.com/jchodera/jacs-dataset-analysis

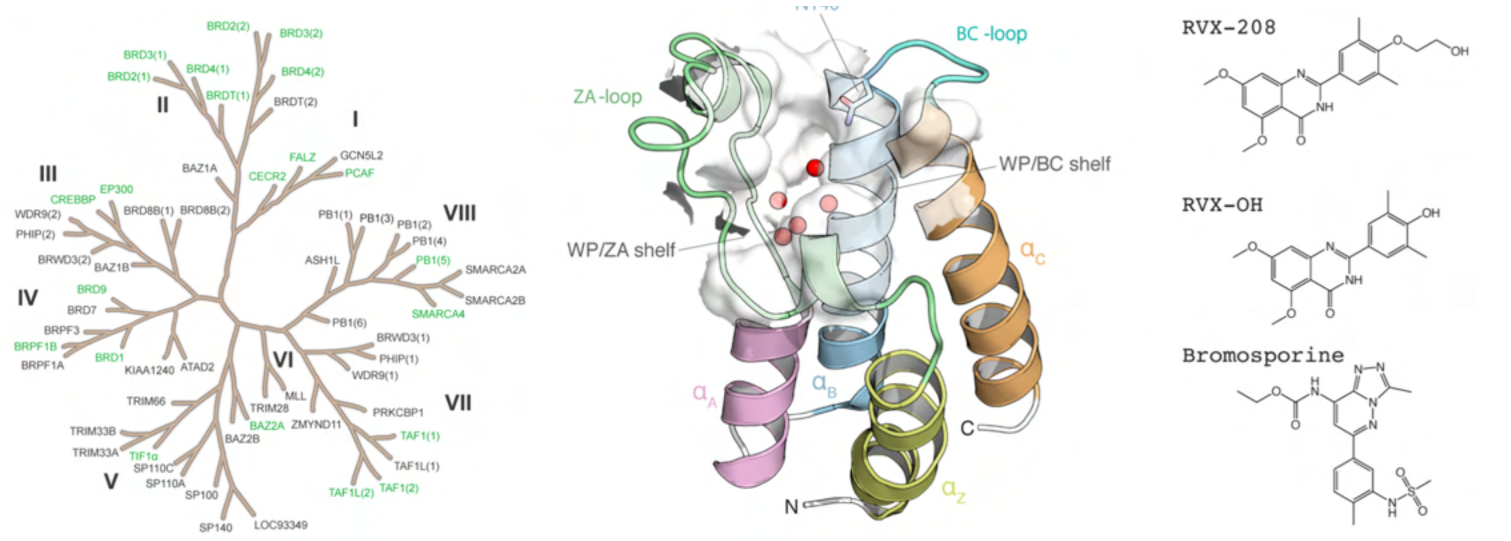
ΔΔG RMSE ~ 1.4 kcal/mol for well-behaved* proteins/chemistries:

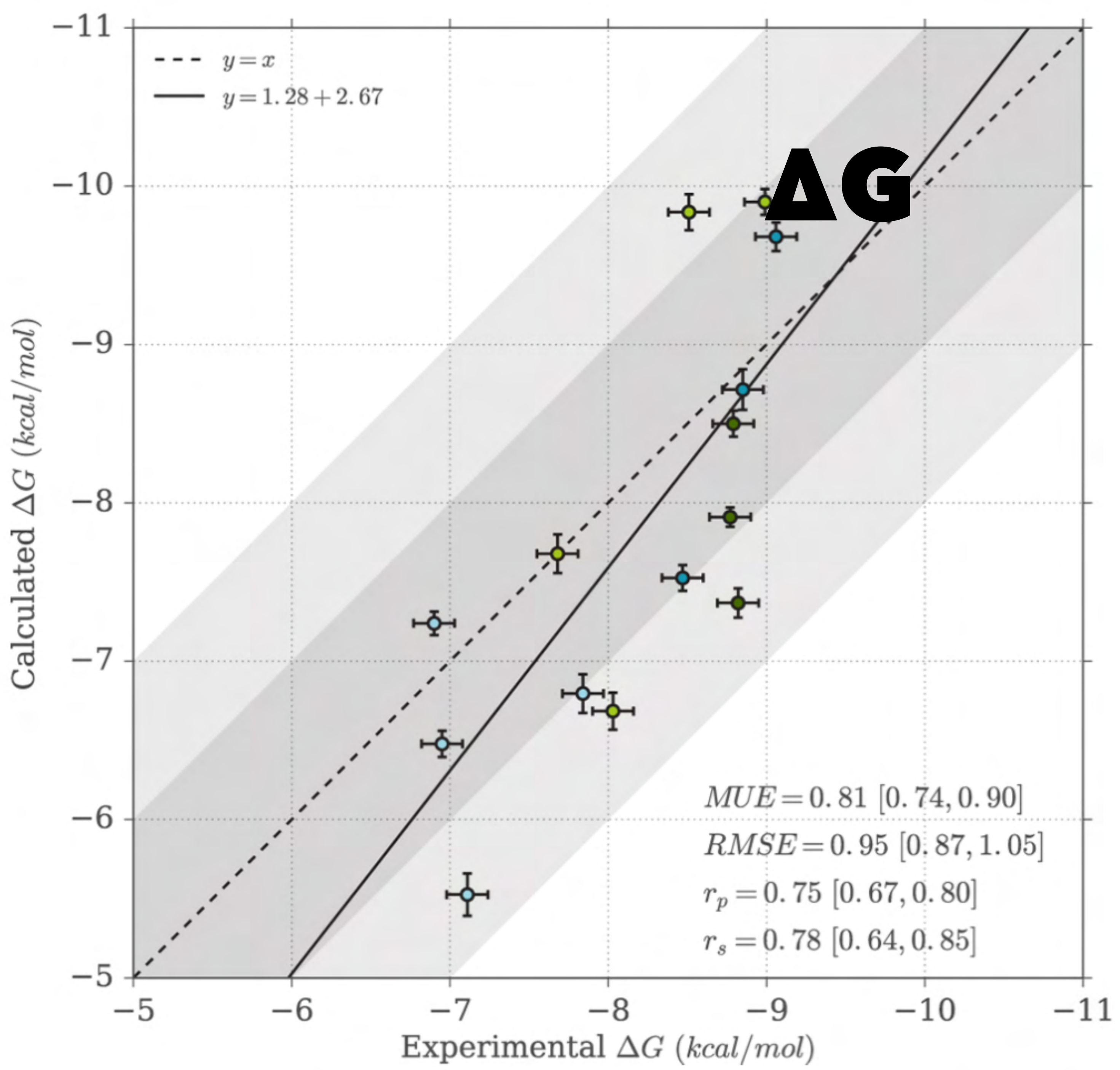
3-5x reduction in molecules synthesized



***best-case scenarios!**

ABSOLUTE



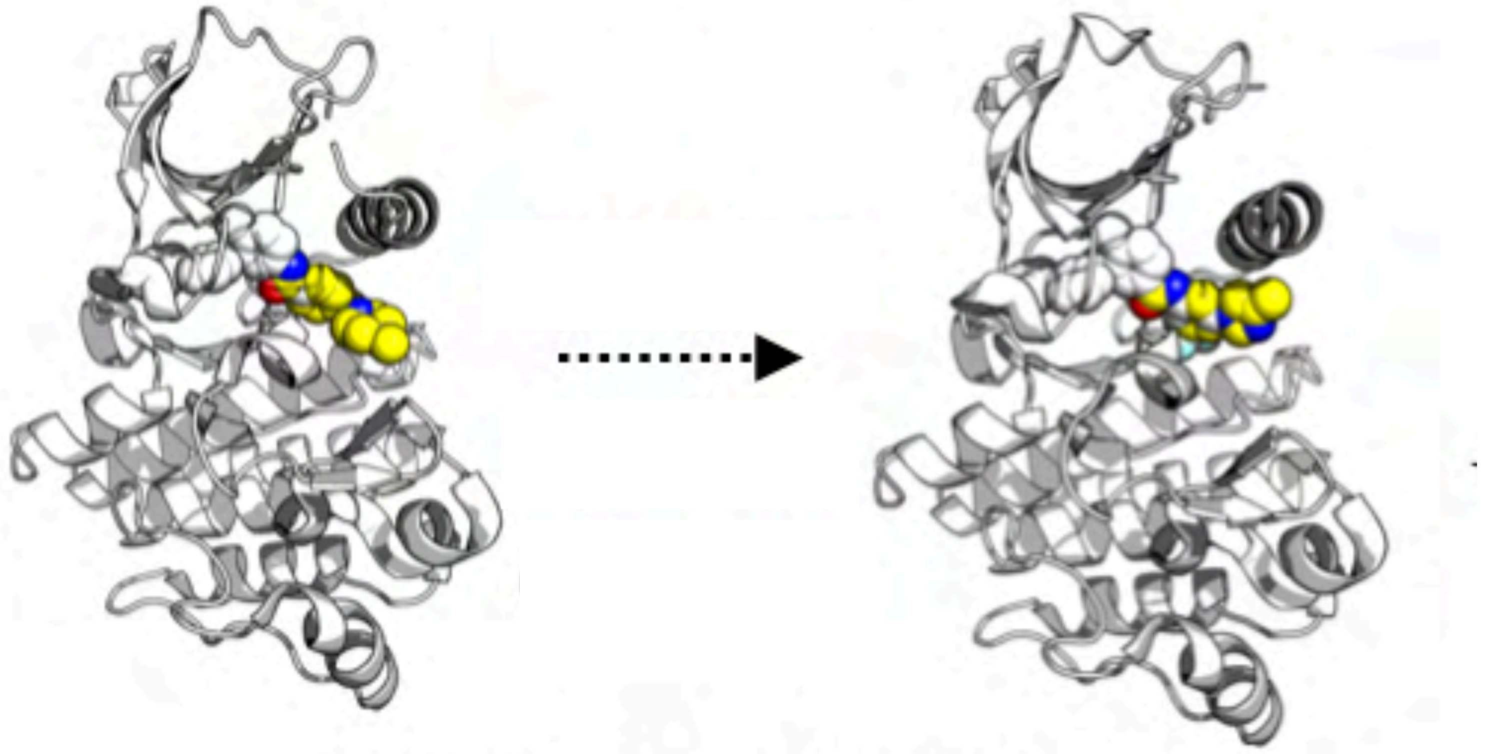


Aldeghi et al. JACS 139:946, 2017.
https://doi.org/10.1021/jacs.6b11467

ALCHEMICAL FREE ENERGY CALCULATIONS CAN BE USED TO COMPUTE MULTIPLE PROPERTIES OF INTEREST

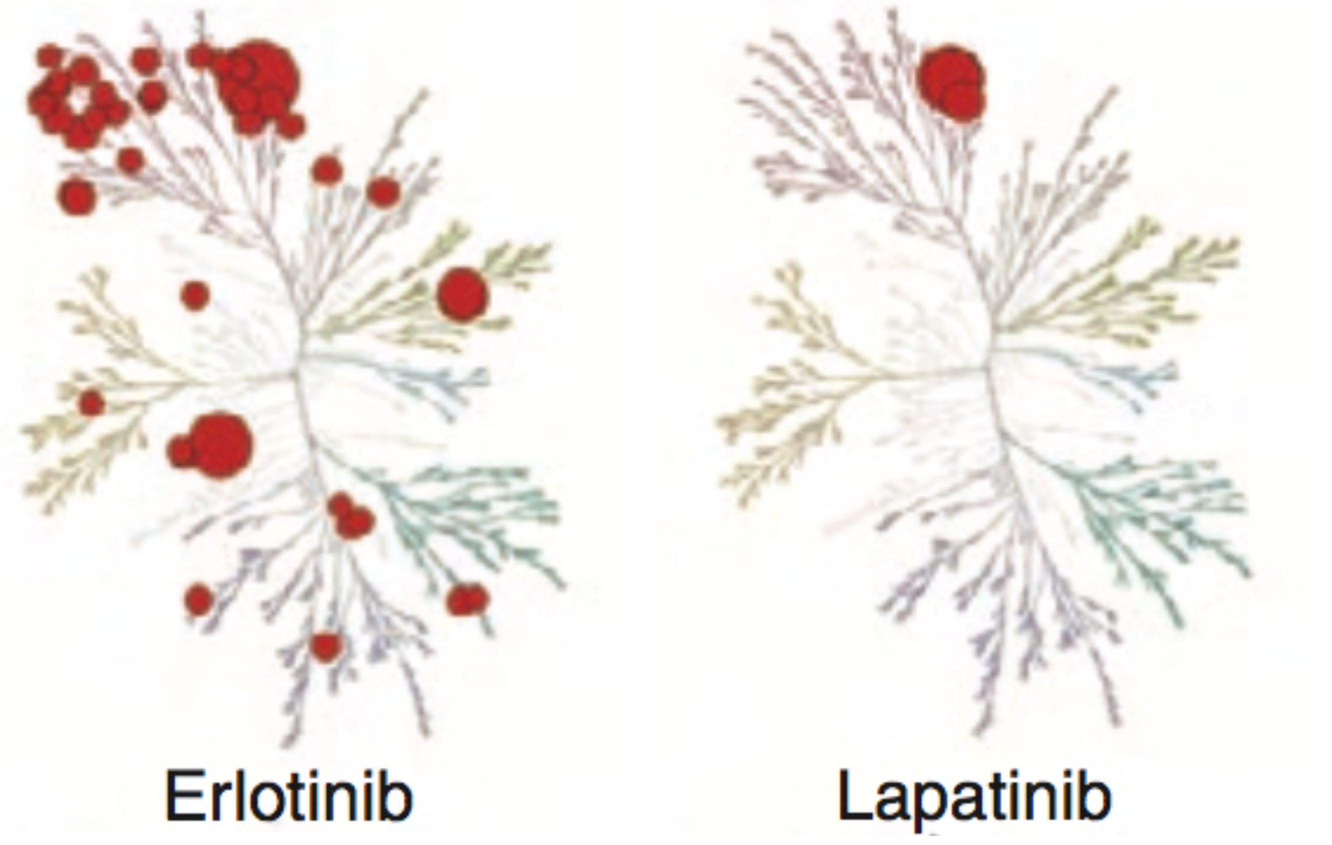
driving affinity / potency

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



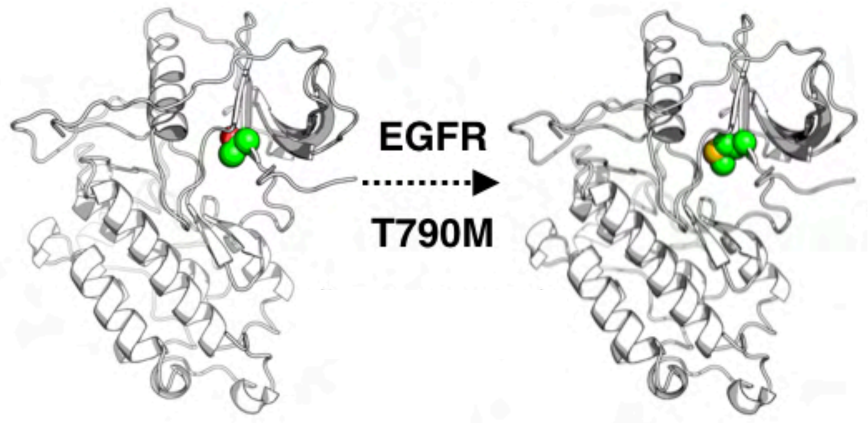
driving selectivity

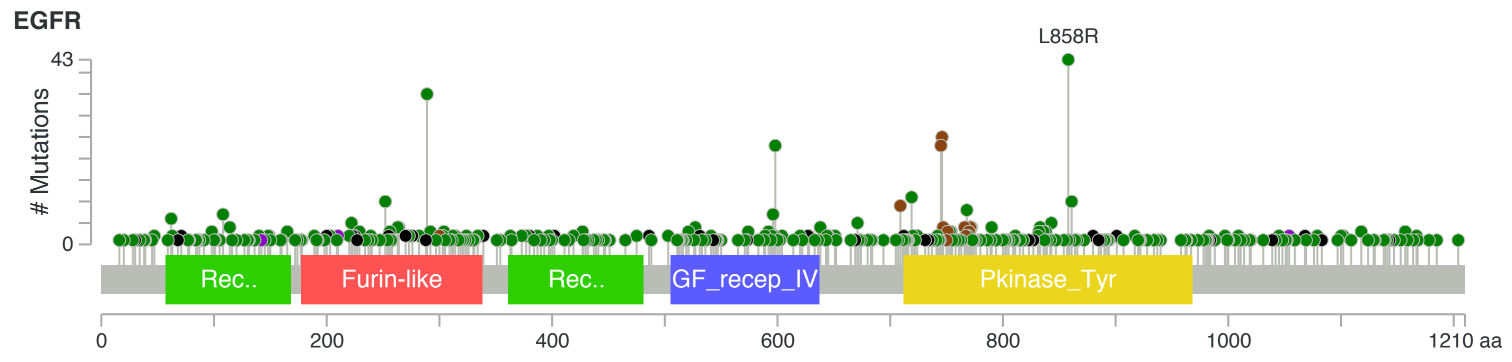
Moraca, Negri, de Olivera, Abel JCIM 2019
https://doi.org/10.1021/acs.jcim.9b00106
Aldeghi et al. JACS 139:946, 2017.
https://doi.org/10.1021/jacs.6b11467



predicting clinical drug resistance/sensitivity

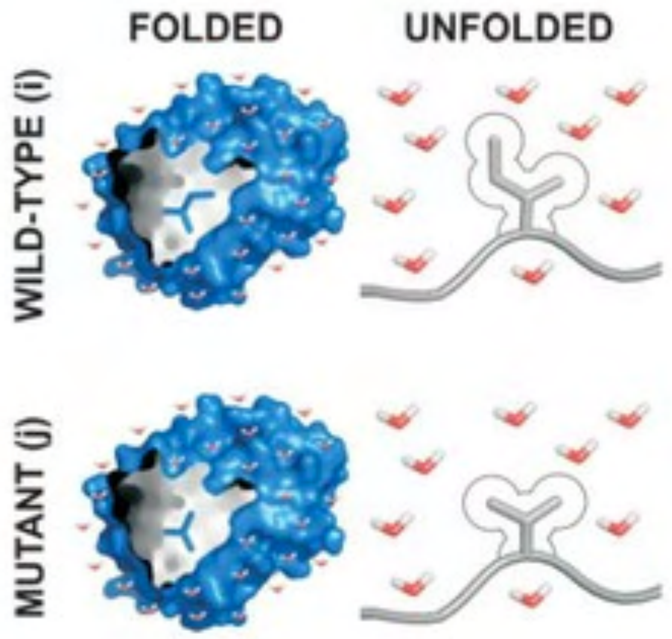
Hauser, Negron, Albanese, Ray, Steinbrecher, Abel, Chodera, Wang. Communications Biology 1:70, 2018
https://doi.org/10.1038/s42003-018-0075-x
Aldeghi, Gapsys, de Groot. ACS Central Science 4:1708, 2018
https://doi.org/10.1021/acscentsci.8b00717

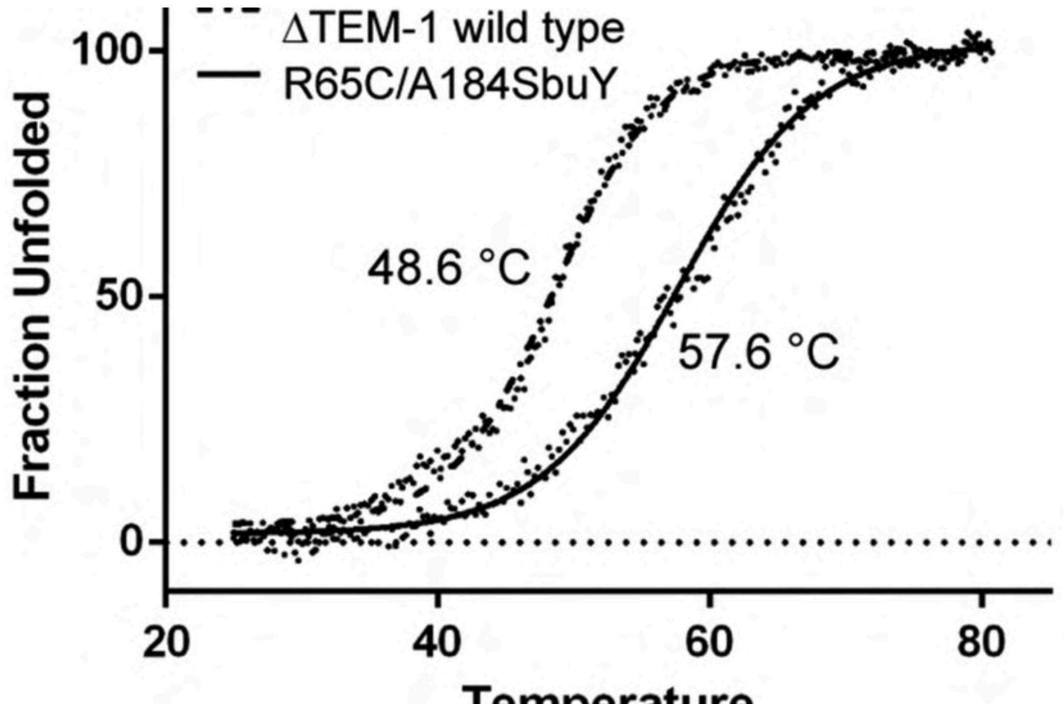




optimizing thermostability

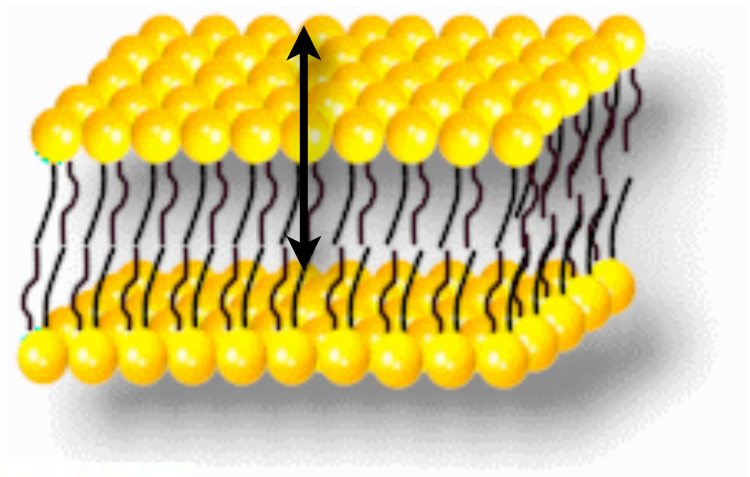
Gapsys, Michielssens, Seeliger, and de Groot. Angew Chem 55:7364, 2016
https://doi.org/10.1002/anie.201510054



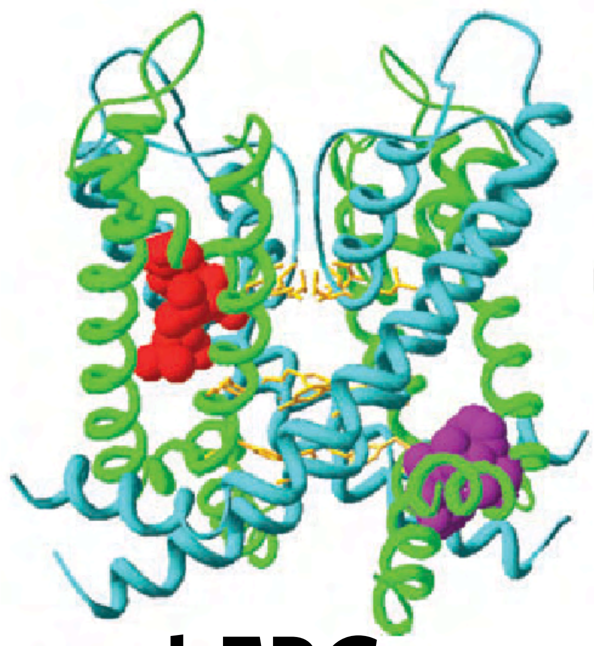


…AND HOLD THE POTENTIAL FOR COMPUTING MANY MORE USEFUL OBJECTIVES FOR DISCOVERY PROGRAMS

partition coefficients (logP, logD) and permeabilities



structure-enabled ADME/Tox targets

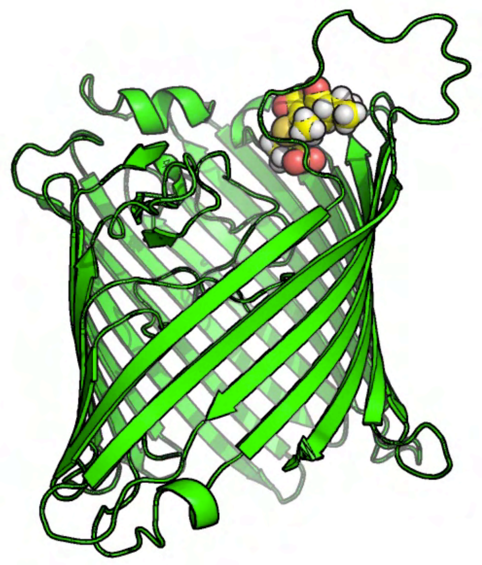


hERG

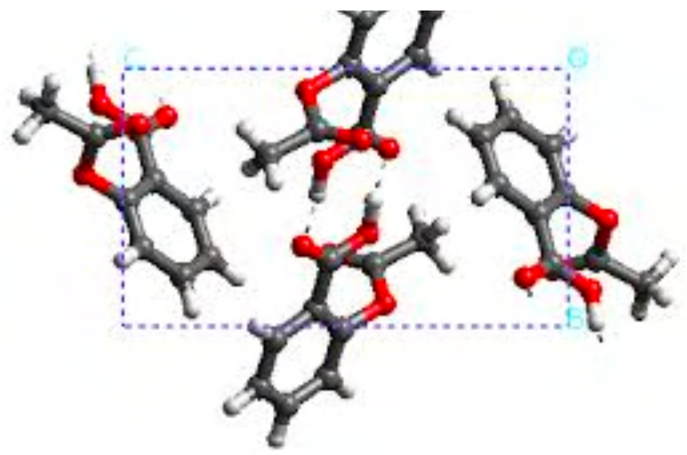
G
G'
distal side
F'
F
E
D
N-term.
C
I
C-term.
K
J
A
heme

CYP3A4

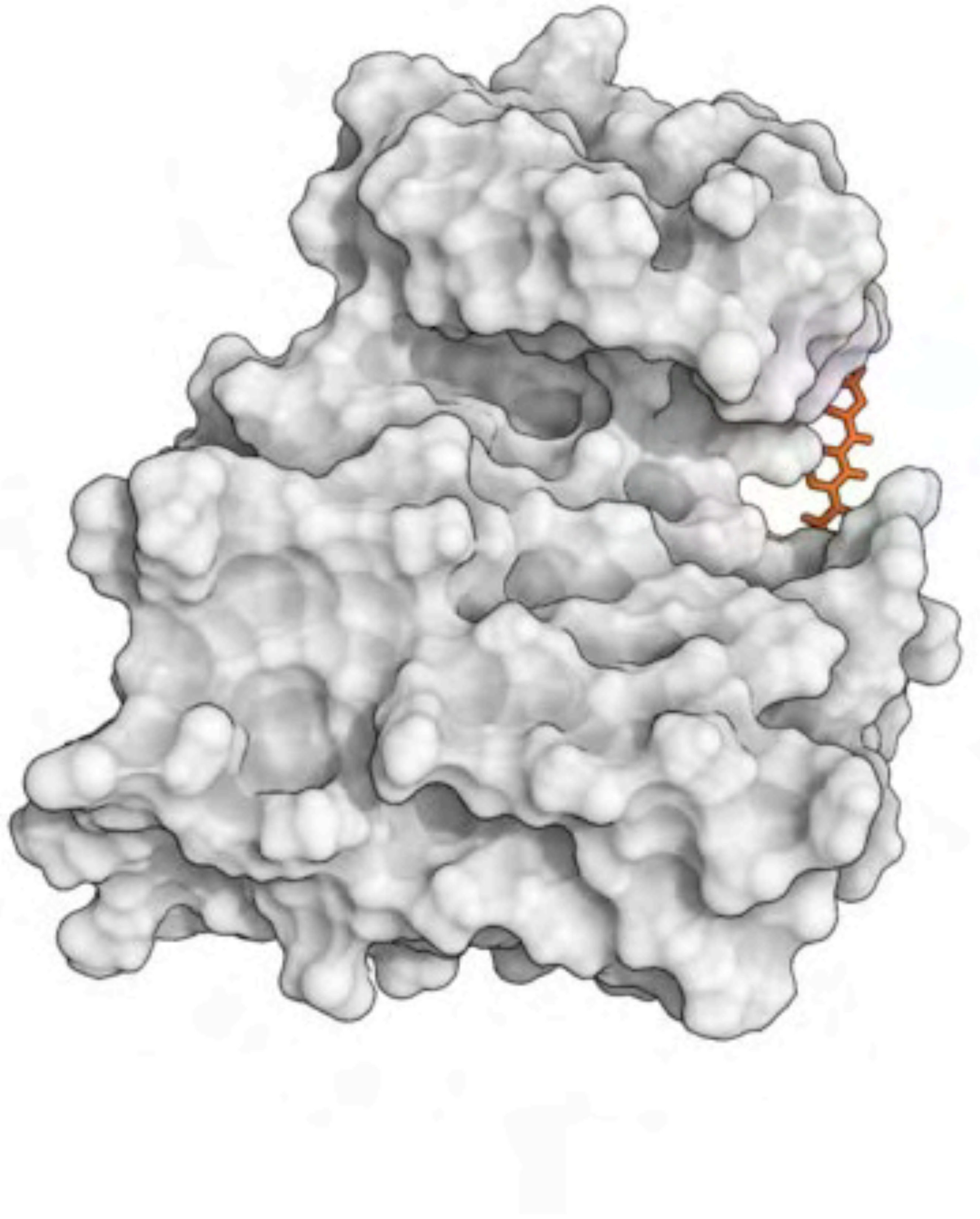
porin permeation



crystal polymorphs, etc.

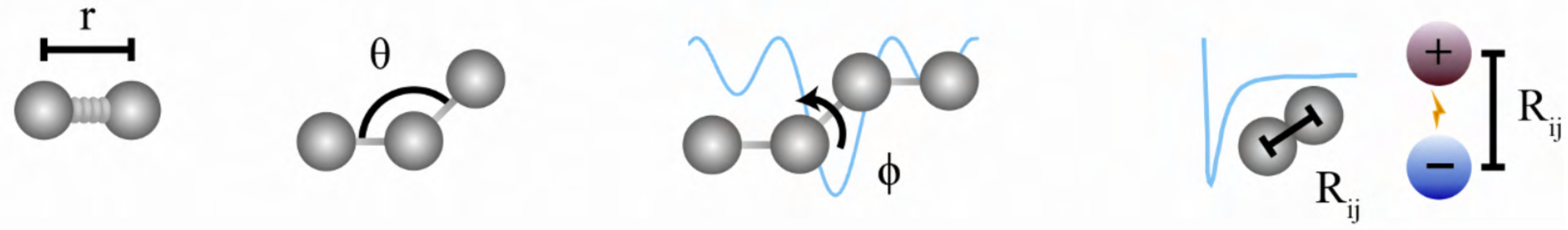


FREE ENERGY CALCULATIONS (AND MUCH OF COMP CHEM) FUNDAMENTALLY RELIES ON MOLECULAR MECHANICS FORCE FIELDS



typical class I molecular mechanics force field

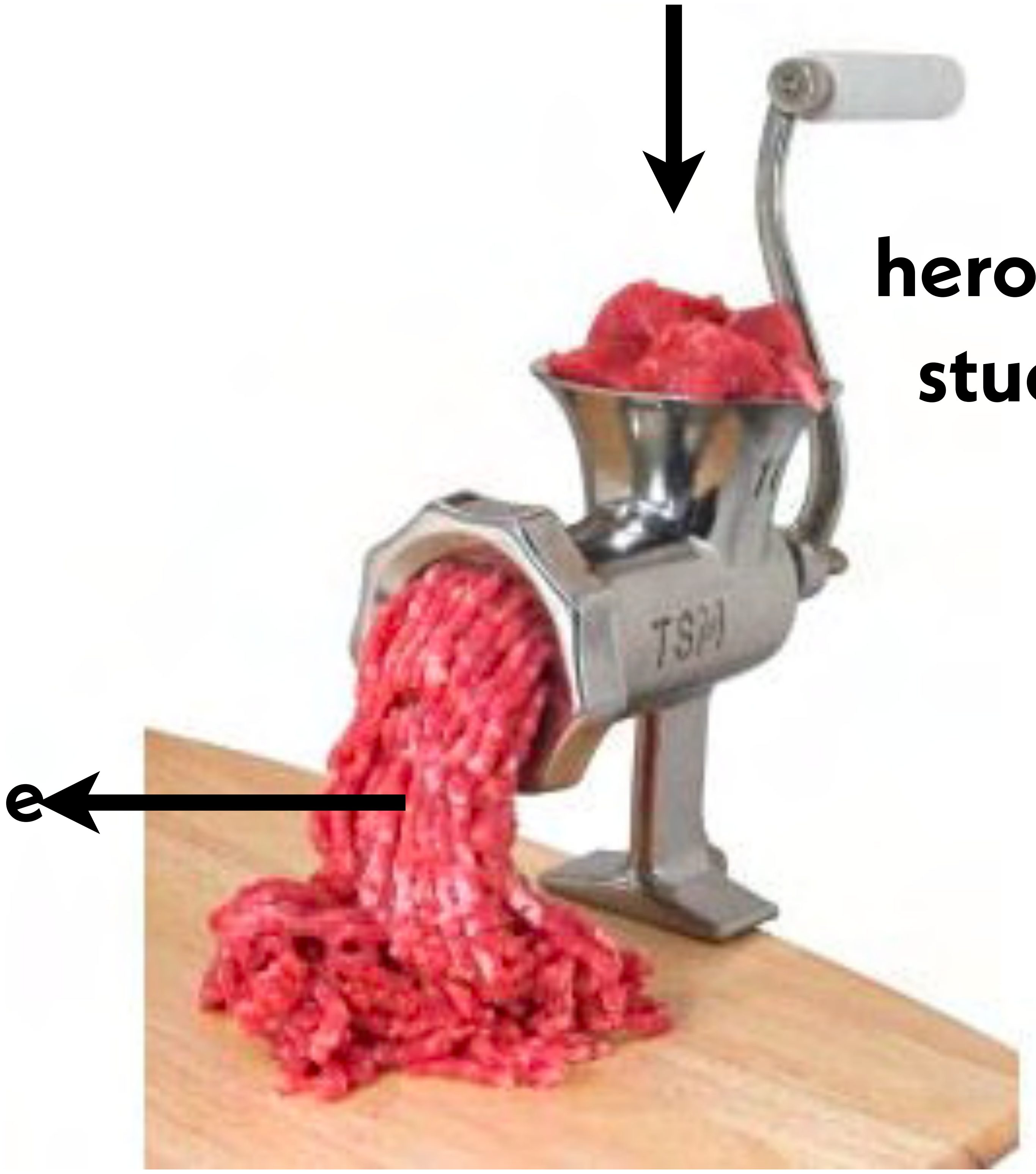
$$E_{total} = \overbrace{\sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]}^{\textbf{Bonded}} + \overbrace{\sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}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



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

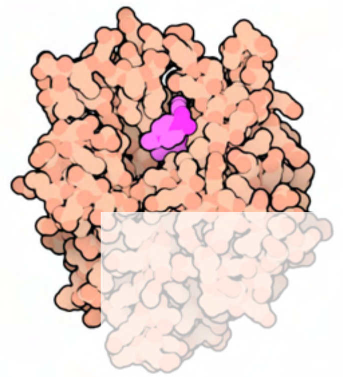
experimental data
quantum chemistry
keen chemical intuition



heroic effort by graduate students and postdocs

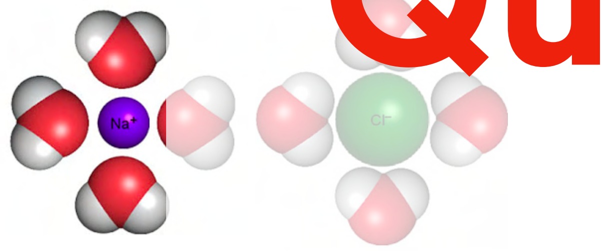
a parameter set we desperately hope someone actually uses

FORCE FIELDS HAVE TRADITIONALLY BEEN HEROIC PRODUCTS OF HUMAN EFFORT



proteins

post-translational modifications



water

ions

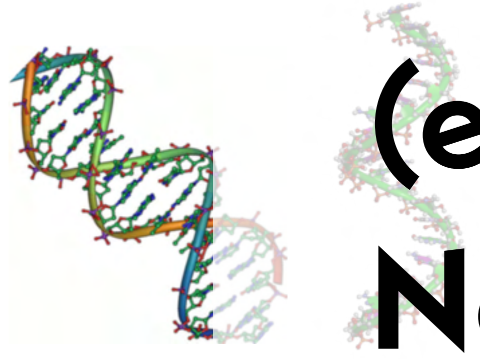
Quickly adds up to >100 human-years

small molecules

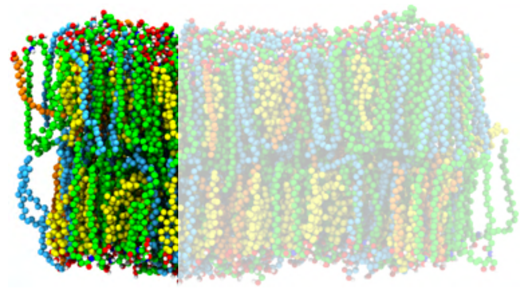
Intended to be compatible, but not co-parameterized

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

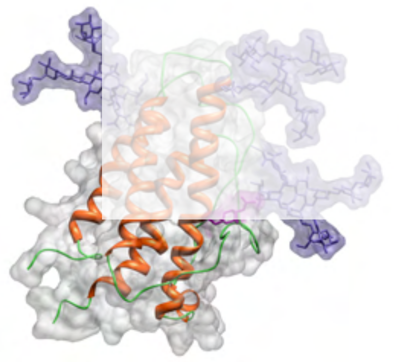


nucleic acids



lipids

How can we bring this problem into the modern era?



carbohydrates

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 [illegible] 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 [illegible] ations 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*, [illegible]

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*, [illegible]

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) was parameterized with this chemical universe:

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)
(c)

GAFF 1 was finished in 1999

Extension to new chemical space is nontrivial

Parameter fitting code was never released

Atom types cause numerous complications

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

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.

http://openforcefield.org

THE OPEN FORCE FIELD INITIATIVE AIMS TO BUILD A MODERN INFRASTRUCTURE FOR FORCE FIELD SCIENCE

Open source Python Toolkit: use the parameters in most simulation packages

Open curated QM / physical property datasets: build your own force fields

Open source infrastructure: for improving force fields with in-house data

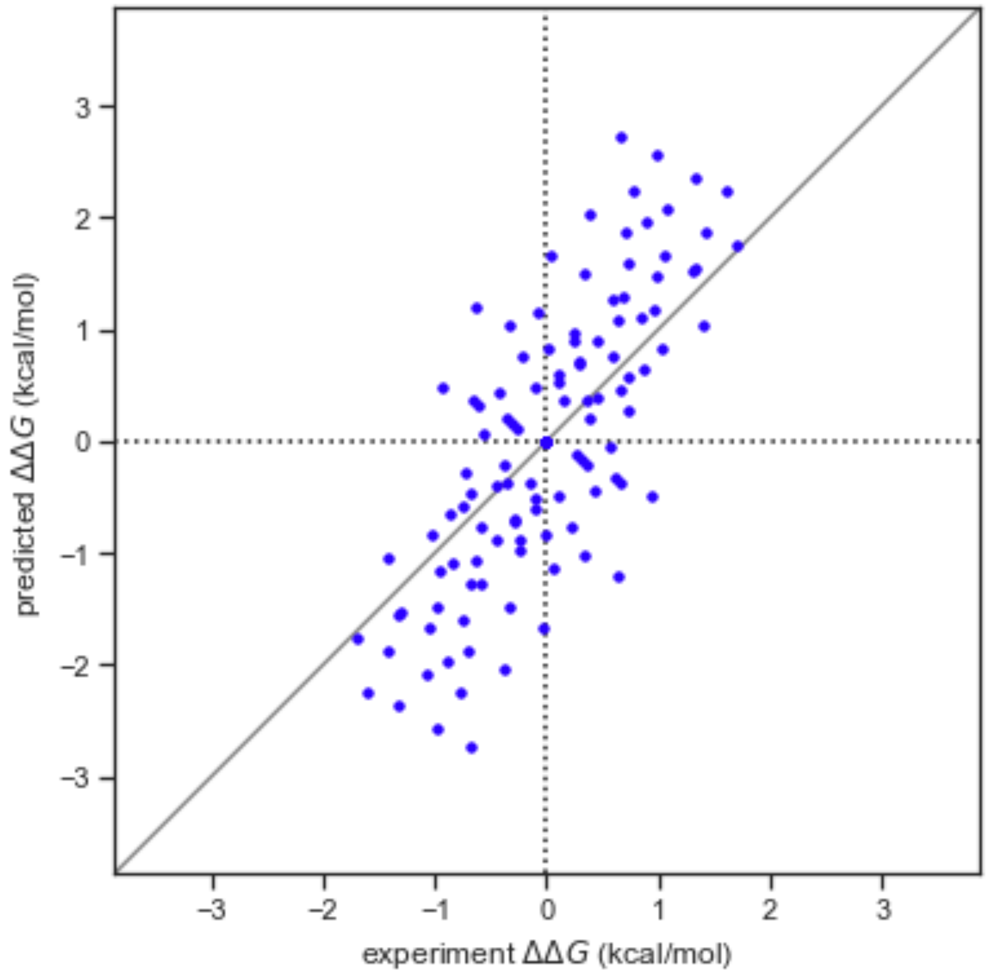
Open science: everything we do is free, permissively licensed, and online

http://openforcefield.org

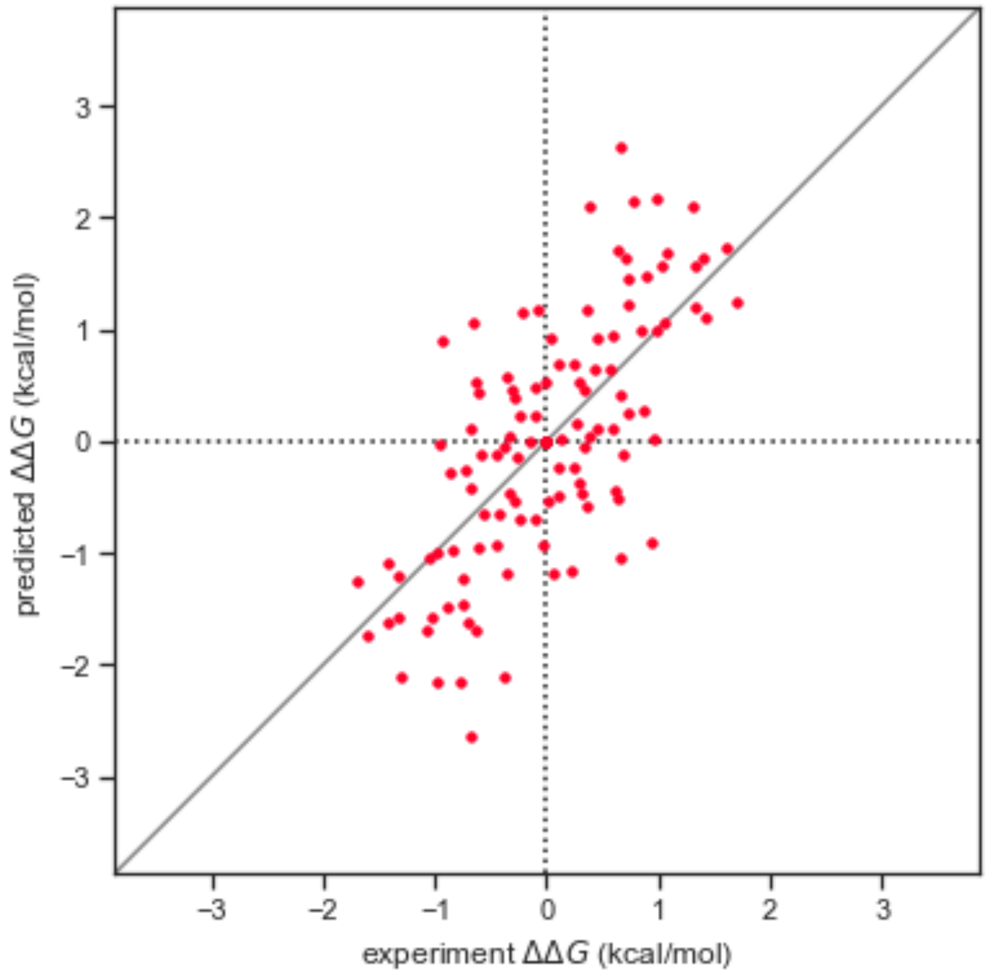
WE'VE MADE RAPID AND SIGNIFICANT PROGRESS

Open Force Field Initiative →

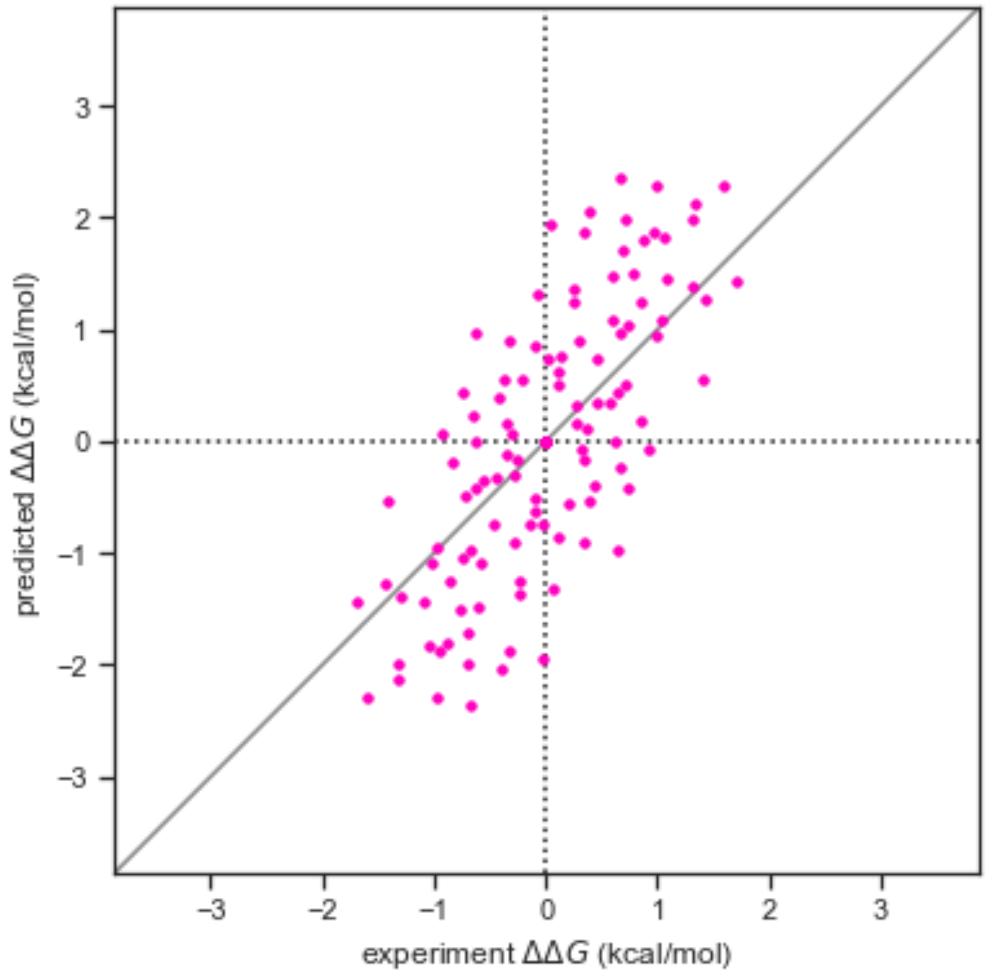
GAFF 1 (1999)



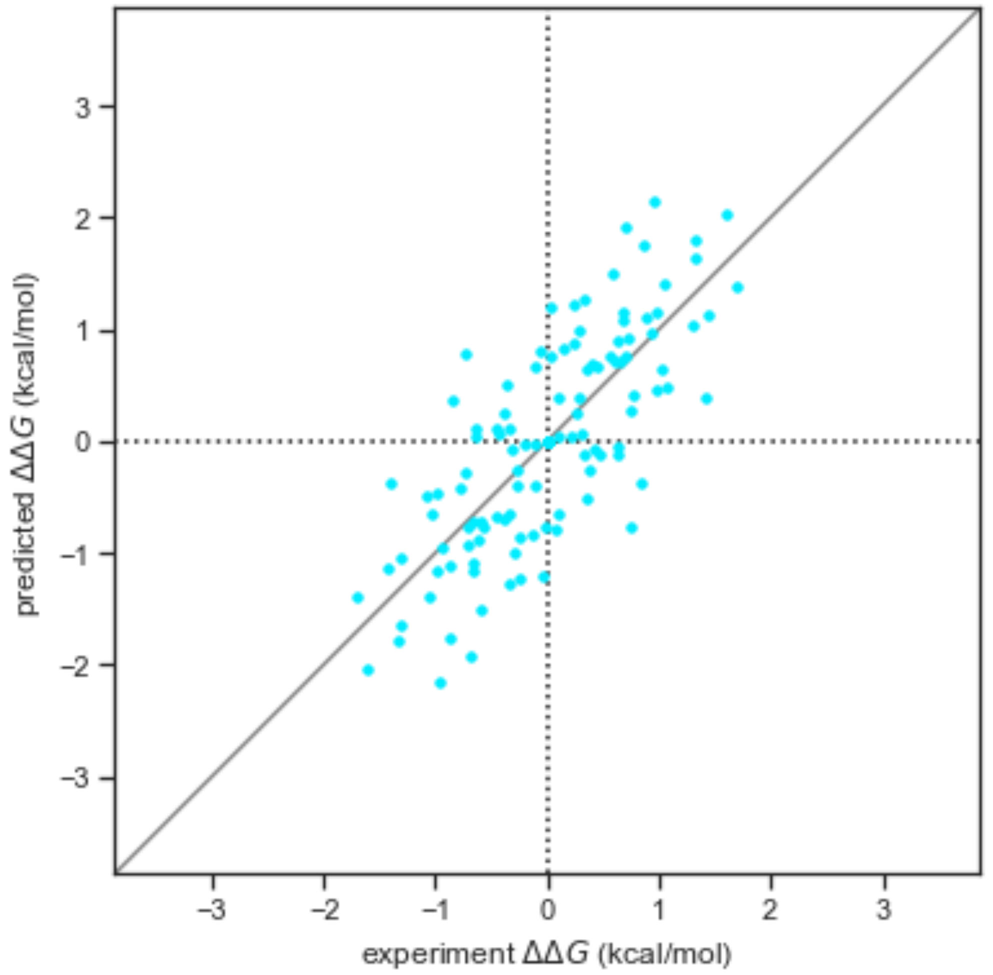
OPLS2.1 (2015)



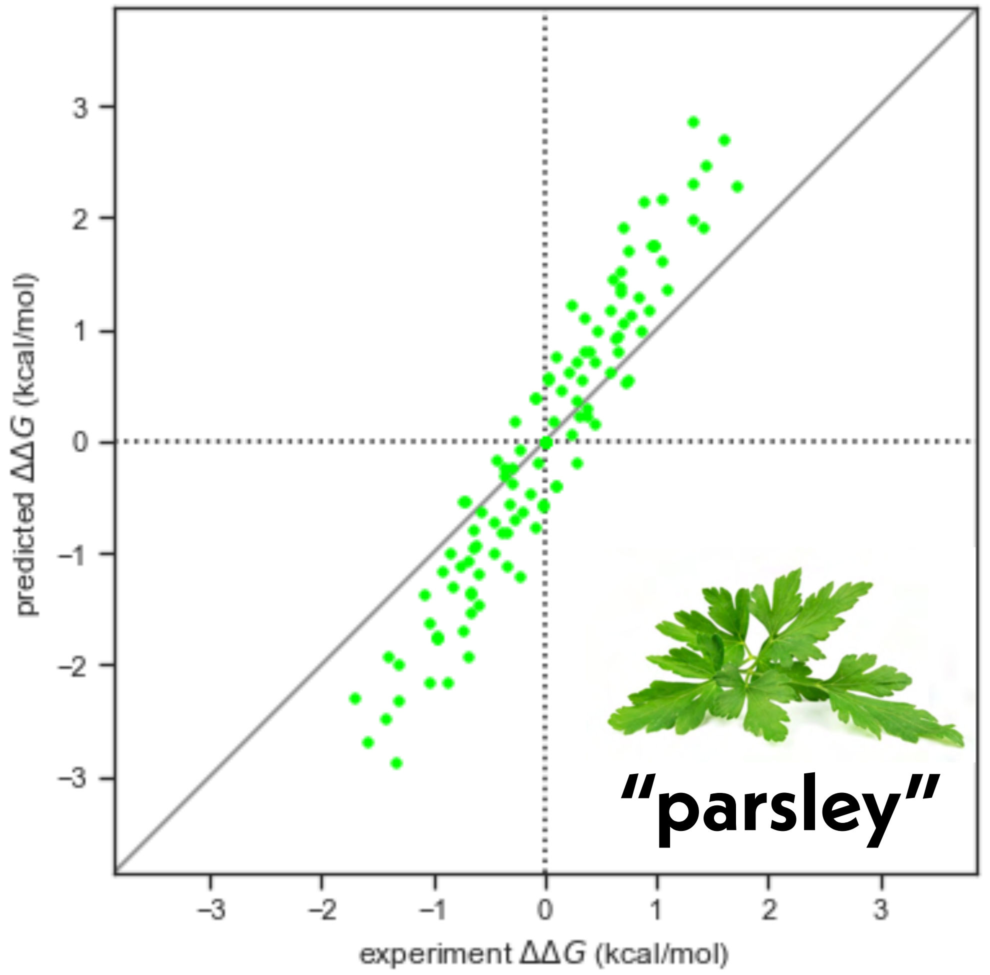
GAFF 2 (2016)

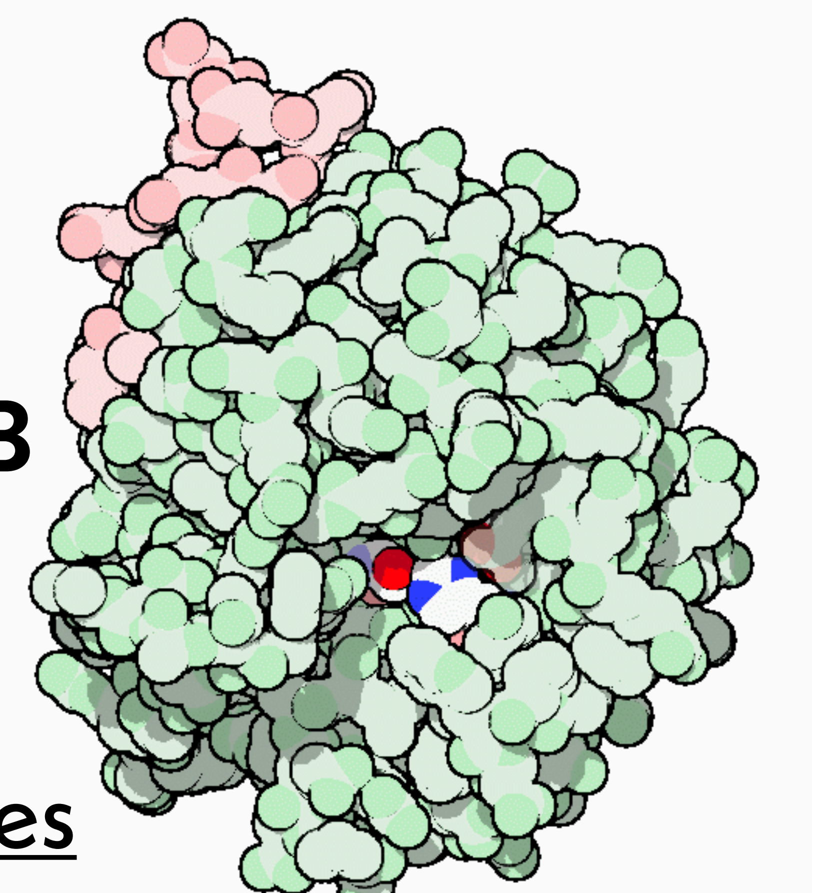


smirnoff99Frosst (2018)



openff 1.0 (2019)





thrombin
PDB101: 1PPB

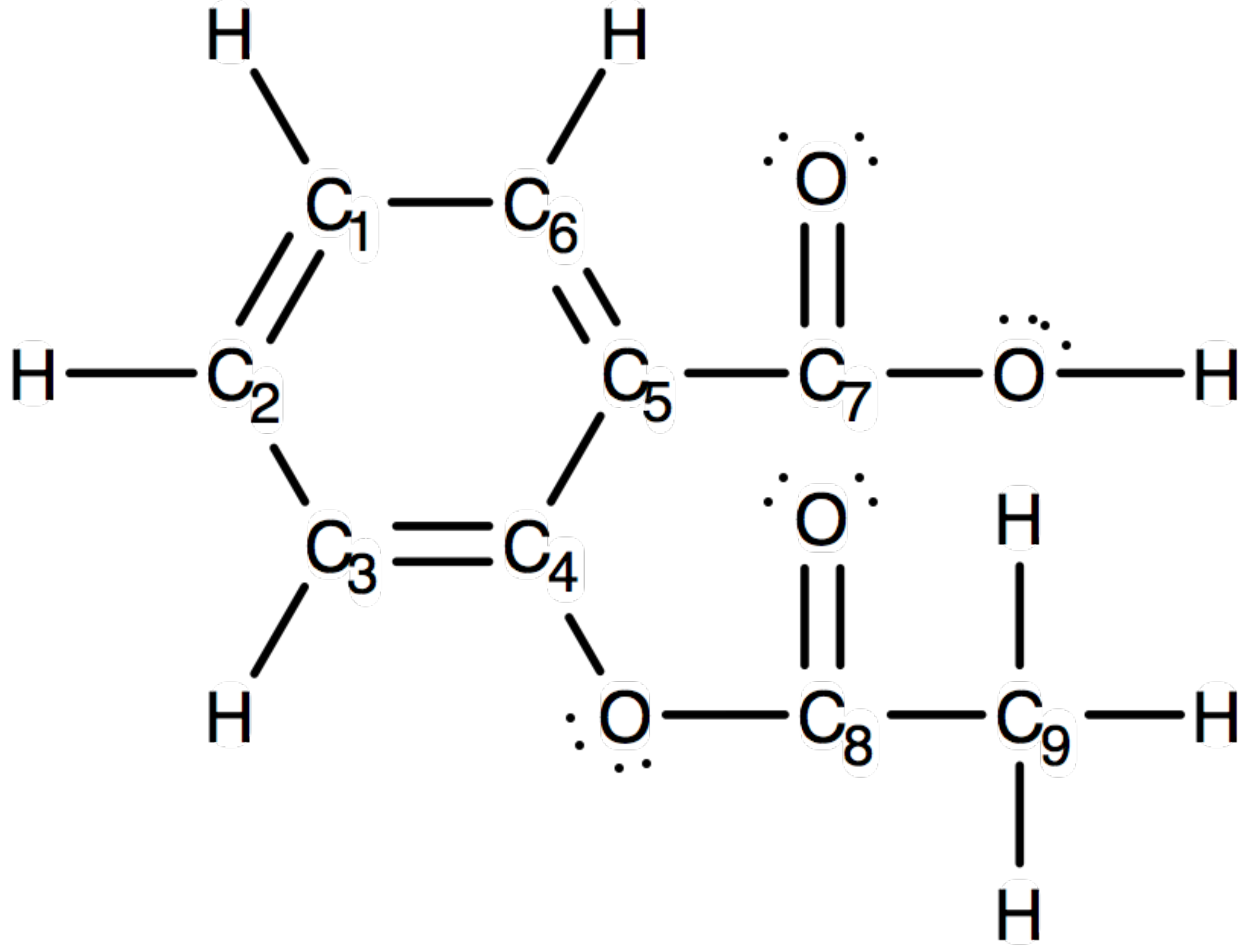
HANNAH BRUCE MACDONALD
MSKCC

http://github.com/choderalab/perses

DOMINIC RUFA

FUNDAMENTALLY, FORCE FIELD PARAMETERIZATION IS DIFFICULT BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

input molecular graph

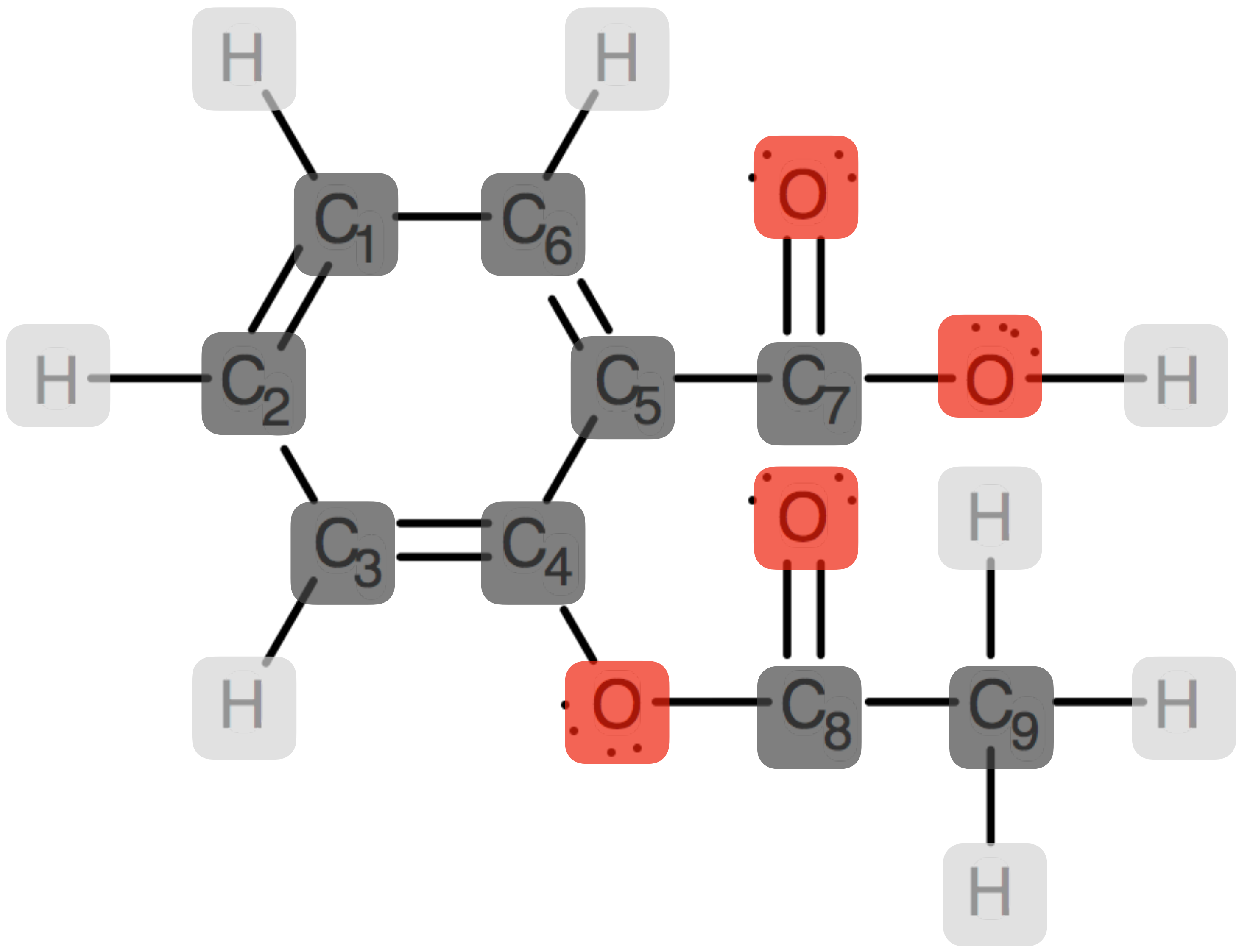


aspirin

JOSH FASS

FUNDAMENTALLY, FORCE FIELD PARAMETERIZATION IS DIFFICULT BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

"atom-typed" molecule



aspirin

3 atom-types

⬜ *hydrogen*

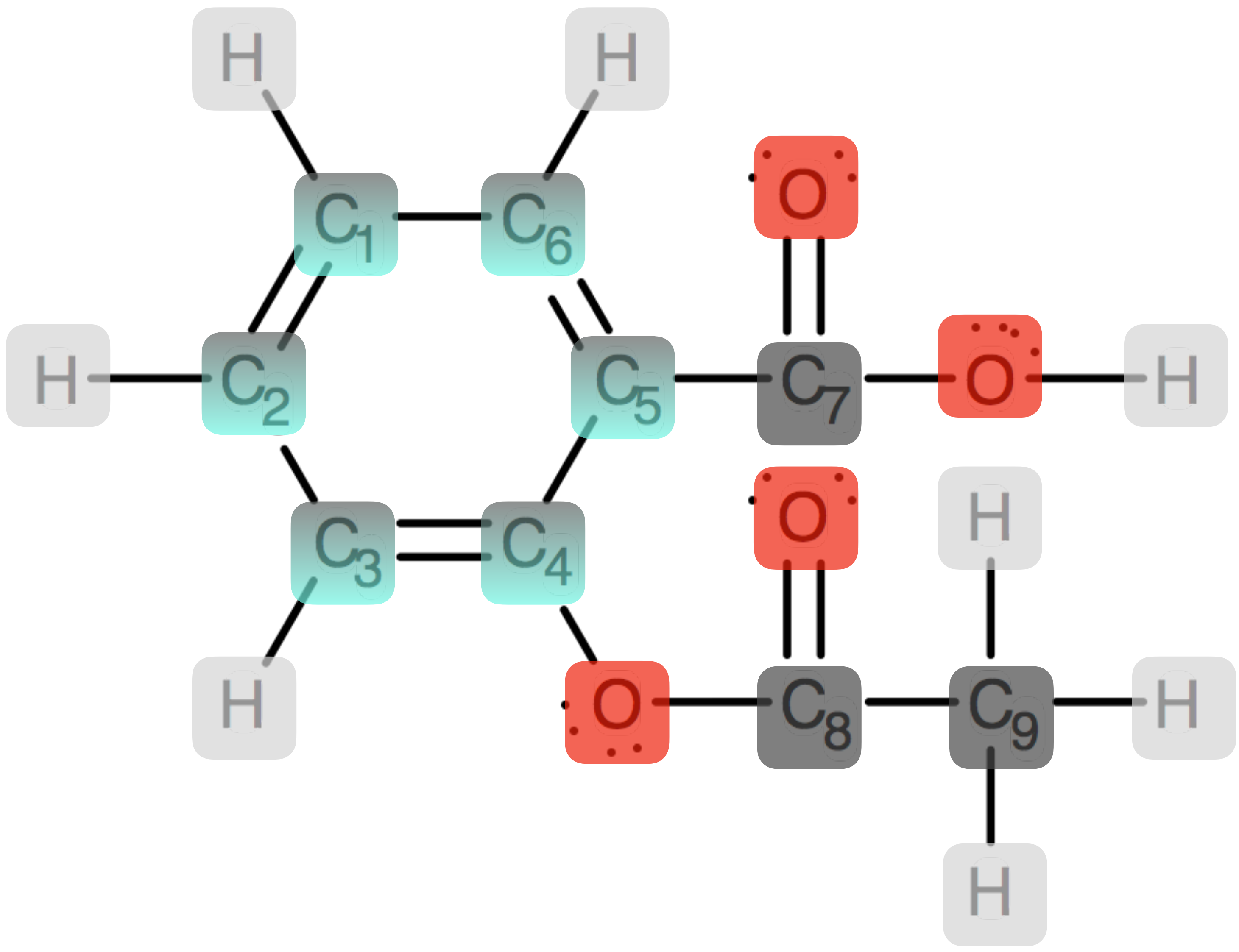
⬛ *carbon*

🟥 *oxygen*

JOSH FASS

FUNDAMENTALLY, FORCE FIELD PARAMETERIZATION IS DIFFICULT BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

"atom-typed" molecule



aspirin

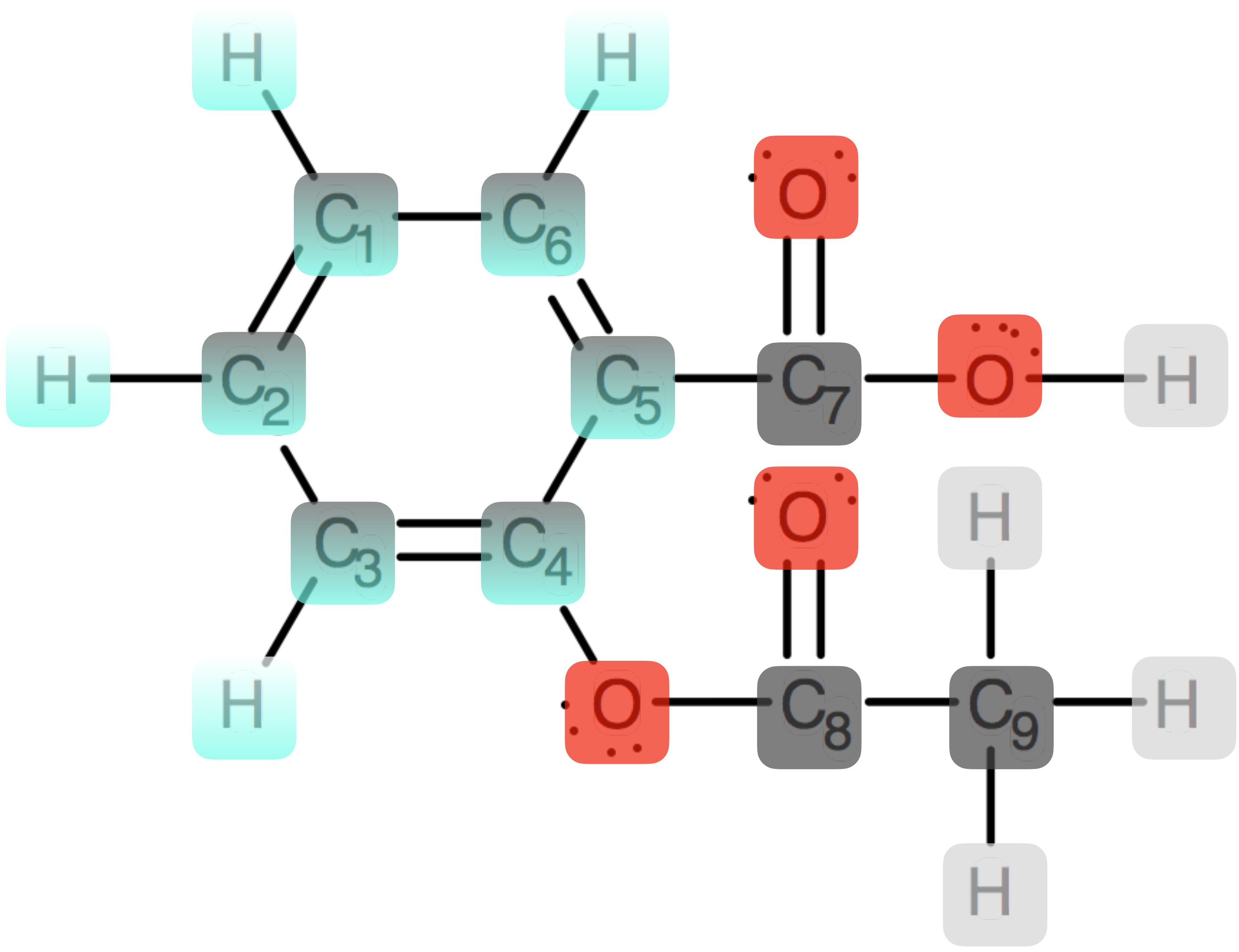
4 atom-types

- ⬜ *hydrogen*
- ⬛ *carbon*
- 🟦 *carbon in an aromatic ring*
- 🟥 *oxygen*

JOSH FASS

FUNDAMENTALLY, FORCE FIELD PARAMETERIZATION IS DIFFICULT BECAUSE IT’S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

“atom-typed” molecule



aspirin

5 atom-types

⬜ *hydrogen*

🟦 *hydrogen bound to a carbon in an aromatic ring*

⬛ *carbon*

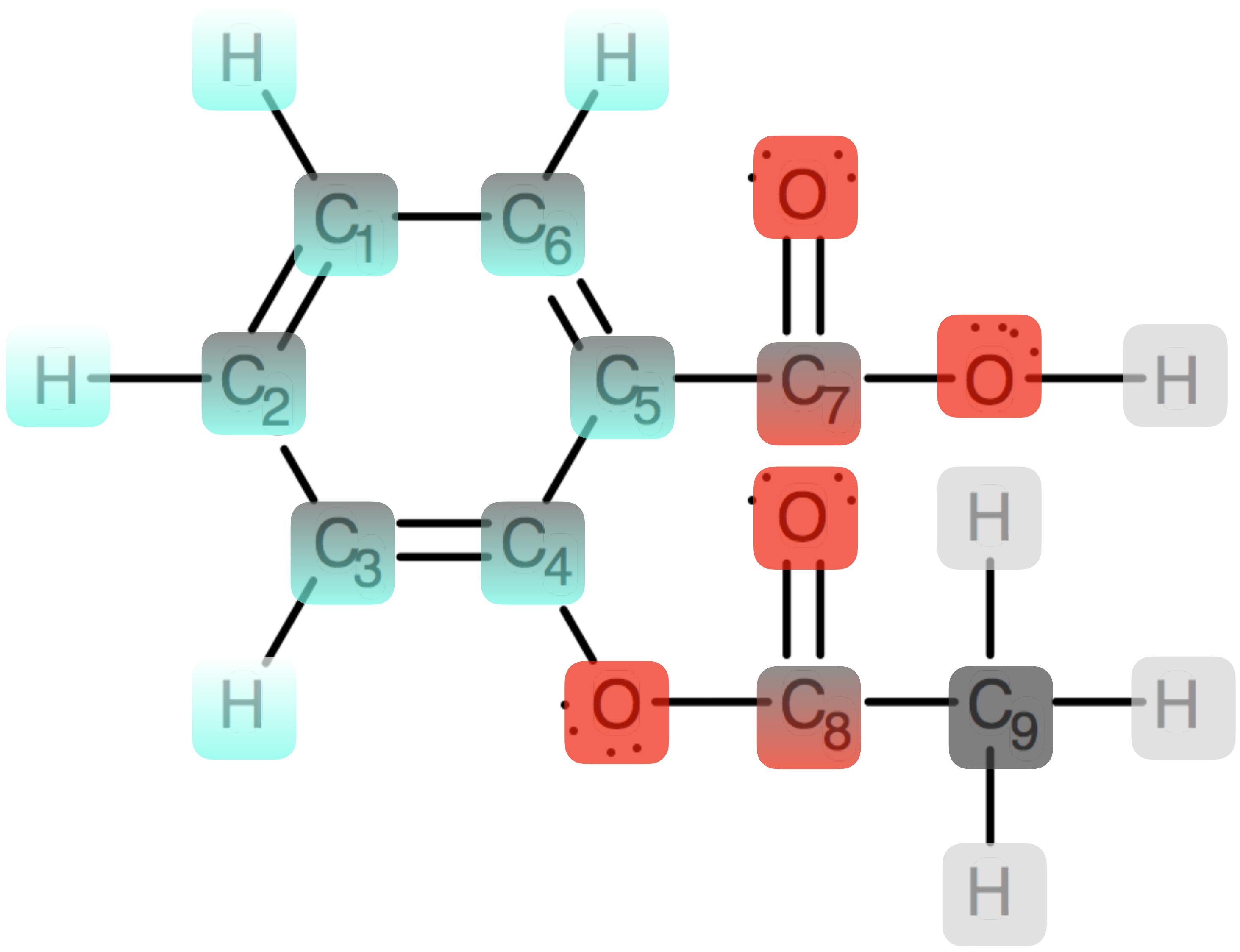
▨ *carbon in an aromatic ring*

🟥 *oxygen*

JOSH FASS

FUNDAMENTALLY, FORCE FIELD PARAMETERIZATION IS DIFFICULT BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

“atom-typed” molecule



aspirin

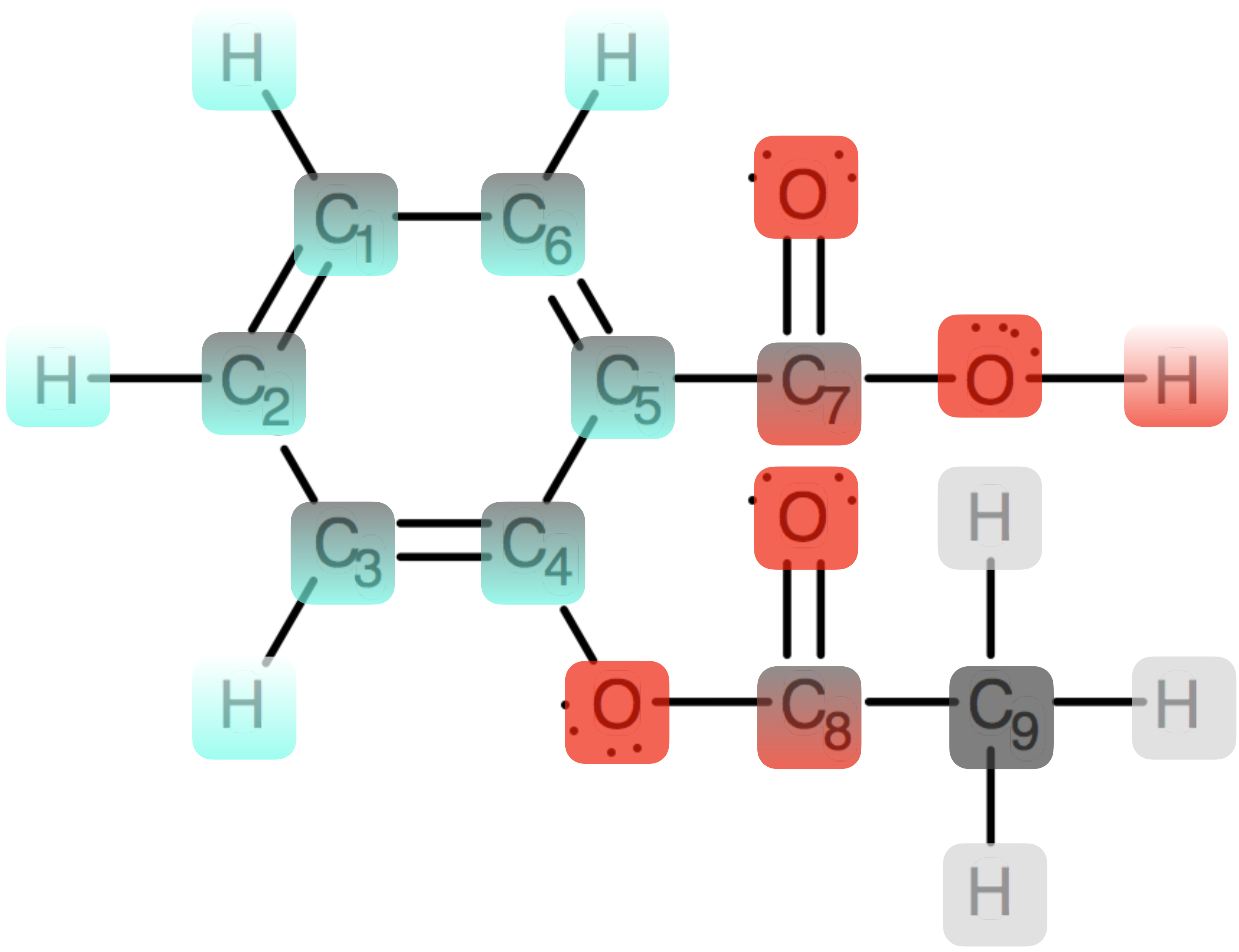
6 atom-types

- *hydrogen*
  - *hydrogen bound to a carbon in an aromatic ring*
- *carbon*
  - *carbon in an aromatic ring*
  - *carbon bound to oxygen*
- *oxygen*

JOSH FASS

FUNDAMENTALLY, FORCE FIELD PARAMETERIZATION IS DIFFICULT BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

“atom-typed” molecule



aspirin

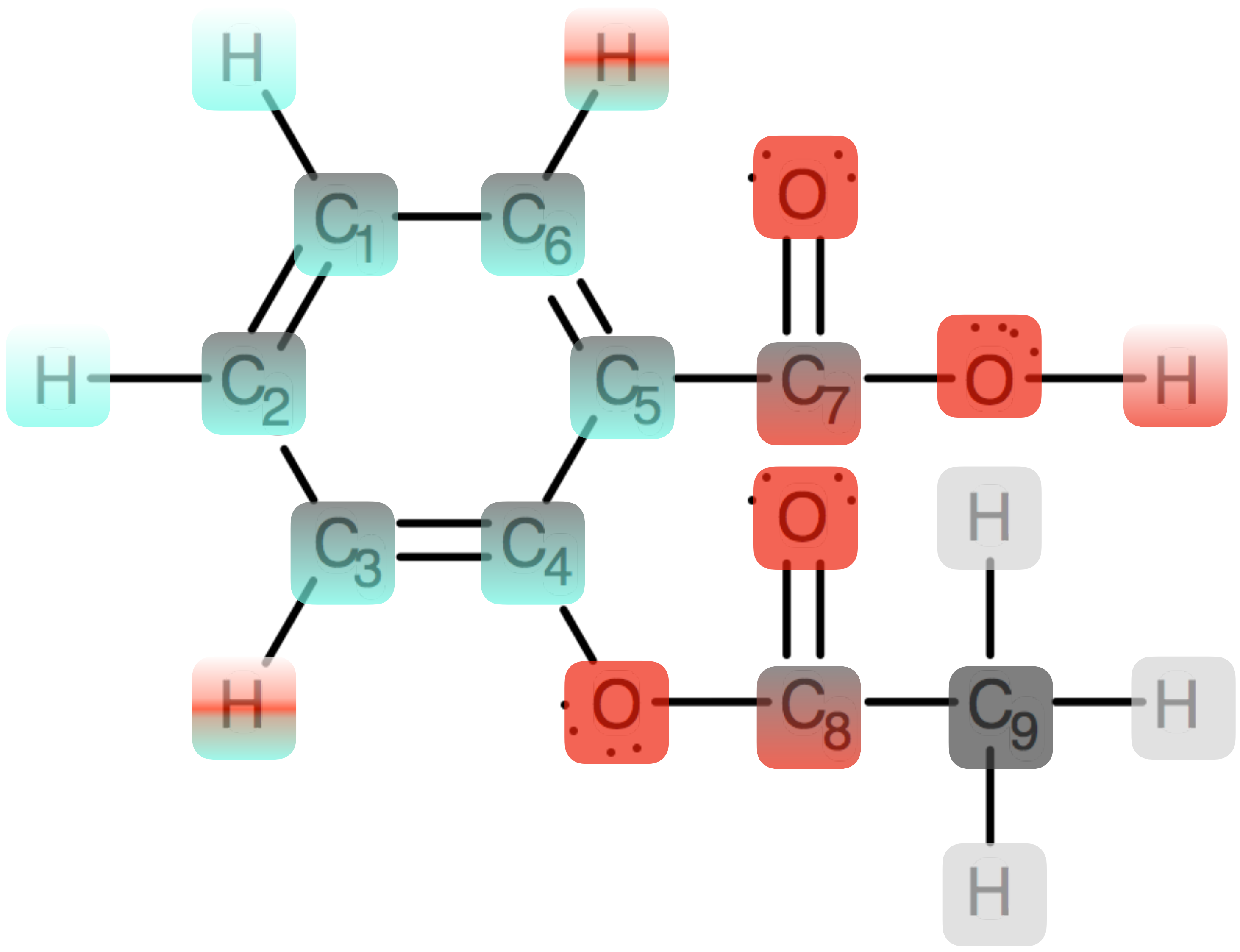
7 atom-types

- ⬜ *hydrogen*
  - 🟦 *hydrogen bound to a carbon in an aromatic ring*
  - 🟥 *hydrogen bound to an oxygen*
- ⬛ *carbon*
  - 🟦 *carbon in an aromatic ring*
  - 🟥 *carbon bound to an oxygen*
- 🟥 *oxygen*

JOSH FASS

FUNDAMENTALLY, FORCE FIELD PARAMETERIZATION IS DIFFICULT BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

"atom-typed" molecule



aspirin

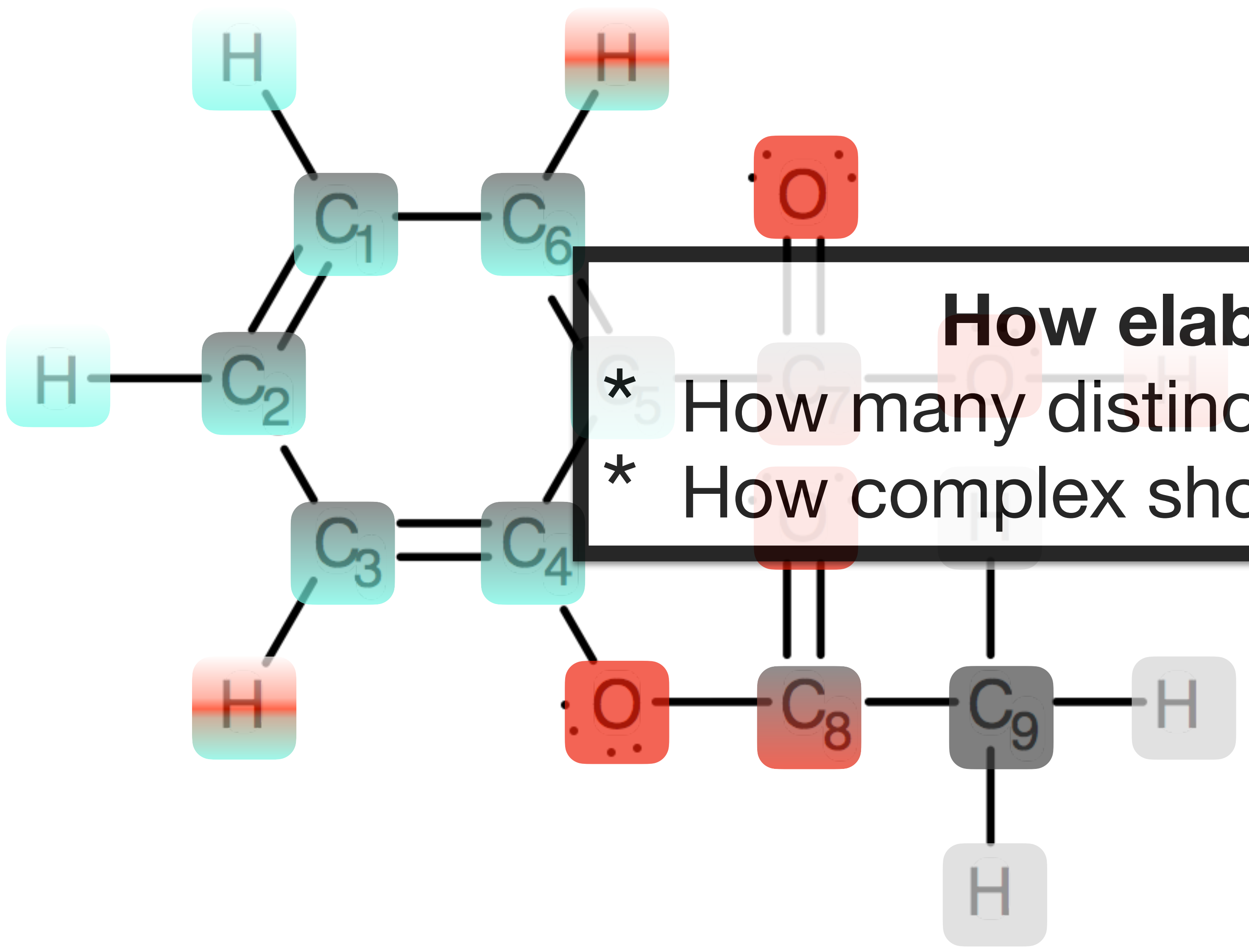
8 atom-types

- ⬜ *hydrogen*
  - 🟦 *hydrogen bound to a carbon in an aromatic ring*
    - ▨ *hydrogen bound to a carbon in an aromatic ring, and 3 bonds away from an oxygen*
  - 🟥 *hydrogen bound to an oxygen*
- ⬛ *carbon*
  - ▨ *carbon in an aromatic ring*
  - ▨ *carbon bound to an oxygen*
- 🟥 *oxygen*

JOSH FASS

FUNDAMENTALLY, FORCE FIELD PARAMETERIZATION IS DIFFICULT BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

"atom-typed" molecule



How elaborate should we go?

* How many distinct atom types are justified?

* How complex should their definitions be?

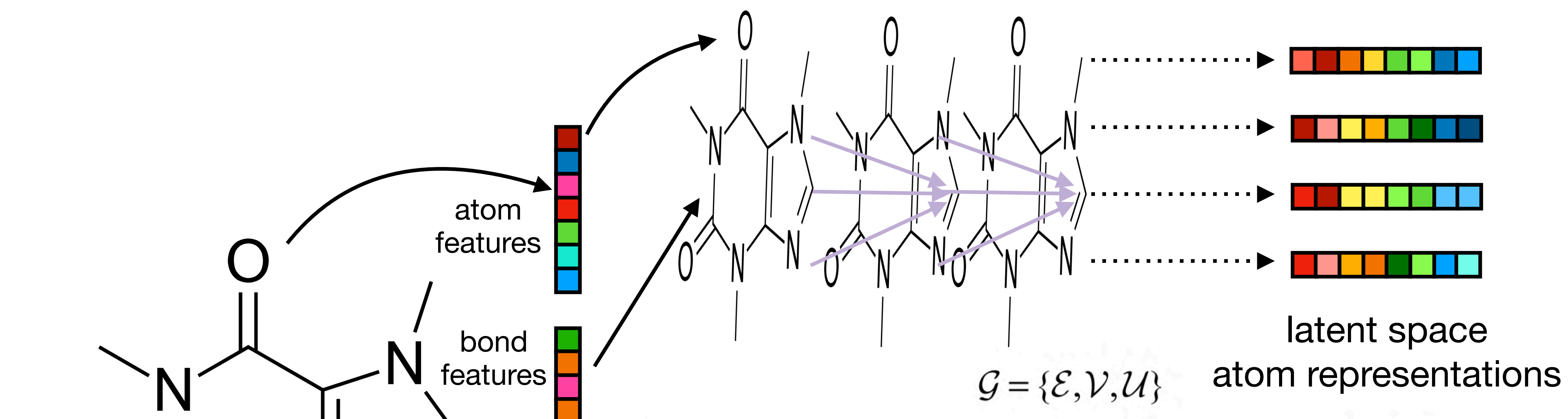
aspirin

8 atom-types

- ⬜ *hydrogen*
  - 🟦 *hydrogen bound to a carbon in an aromatic ring*
  - 🟥 *hydrogen bound to a carbon in an aromatic ring, and 3 bonds away from an oxygen*
  - 🟥 *hydrogen bound to an oxygen*
- ⬛ *carbon*
  - 🟦 *carbon in an aromatic ring*
  - 🟥 *carbon bound to an oxygen*
- 🟥 *oxygen*

JOSH FASS

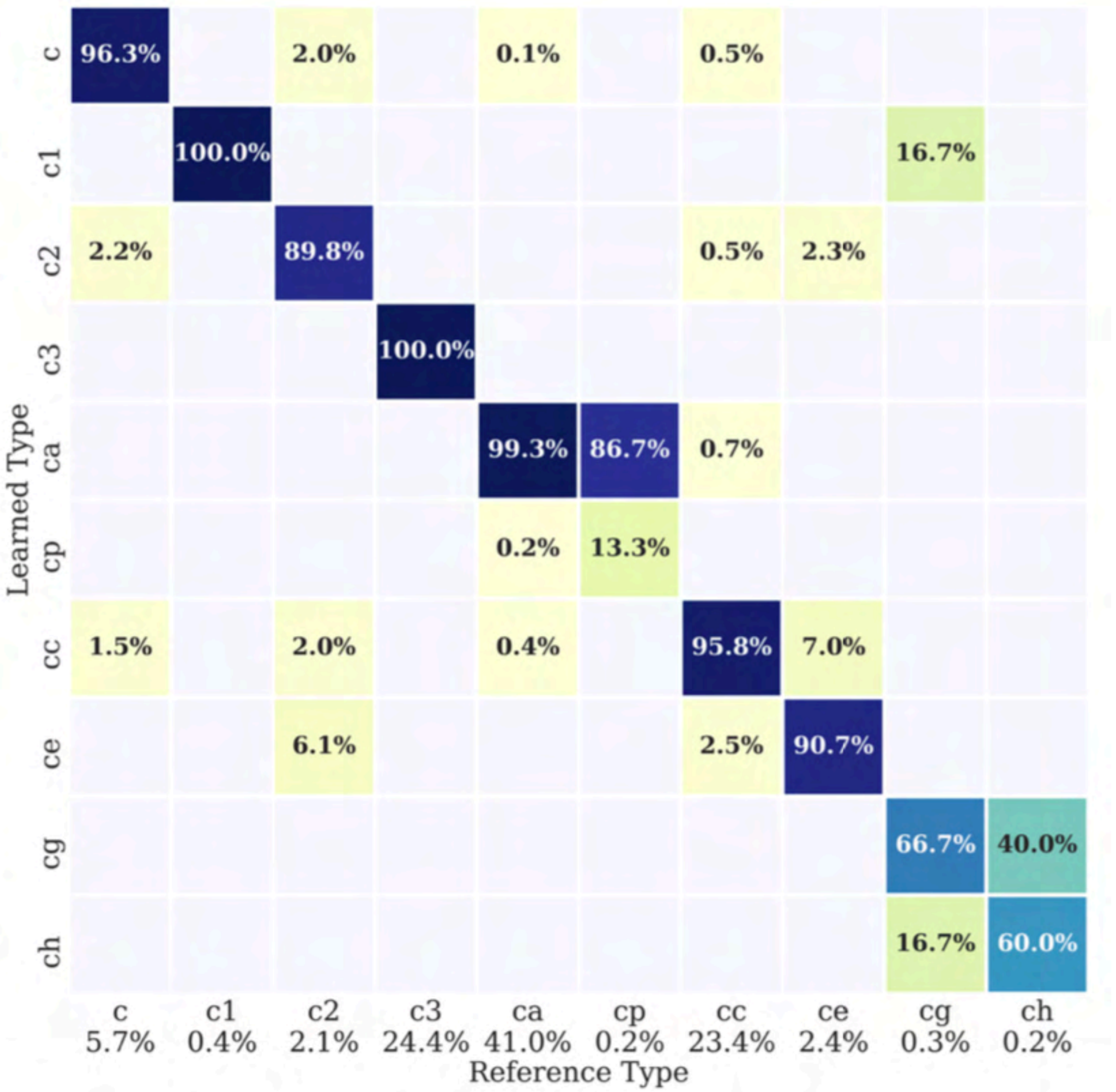
GRAPH CONVOLUTIONAL NETWORKS CAN LEARN CHEMICAL ENVIRONMENTS WITHOUT REQUIRING DISCRETE ATOM TYPES


O
N
N
O
N
N
atom features
bond features
latent space atom representations
$\mathcal{G}=\{\mathcal{E},\mathcal{V},\mathcal{U}\}$

$$\mathbf{e}'_k = \phi^e(\mathbf{e}_k, \mathbf{v}_{rk}, \mathbf{v}_{sk}, \mathbf{u}) \qquad \bar{e}'_i = \rho^{e\to v}(E'_i)$$
$$\mathbf{v}'_i = \phi^v(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u}) \qquad \bar{e}' = \rho^{e\to u}(E')$$
$$\mathbf{u}' = \phi^u(\bar{\mathbf{e}}', \bar{\mathbf{v}}', \mathbf{u}) \qquad \bar{\mathrm{v}}' = \rho^{v\to u}(V')$$

chemical graph of molecule to be parameterized

Graph convolutional network (message-passing network)



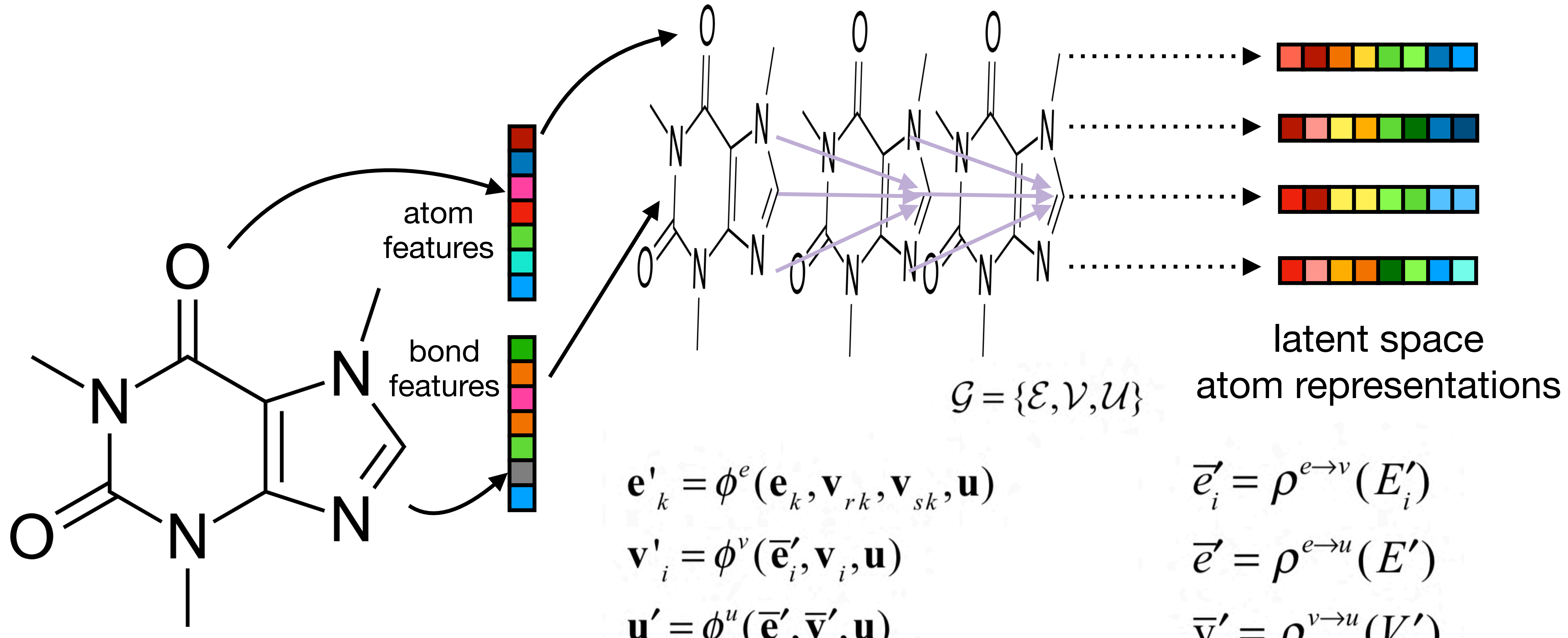
Learned Type \ Reference Type	c	c1	c2	c3	ca	cp	cc	ce	cg	ch
c	96.3%		2.0%		0.1%		0.5%			
c1		100.0%							16.7%	
c2	2.2%		89.8%				0.5%	2.3%		
c3				100.0%						
ca					99.3%	86.7%	0.7%			
cp					0.2%	13.3%				
cc	1.5%		2.0%		0.4%		95.8%	7.0%		
ce			6.1%				2.5%	90.7%		
cg									66.7%	40.0%
ch									16.7%	60.0%
	5.7%	0.4%	2.1%	24.4%	41.0%	0.2%	23.4%	2.4%	0.3%	0.2%

GAFF 1.81 atom types predicted with 98.31% [95% CI: 97.94, 98.63] accuracy

JOSH FASS

YUANQING WANG

GRAPH CONVOLUTIONAL NETWORKS CAN LEARN CHEMICAL ENVIRONMENTS WITHOUT REQUIRING DISCRETE ATOM TYPES



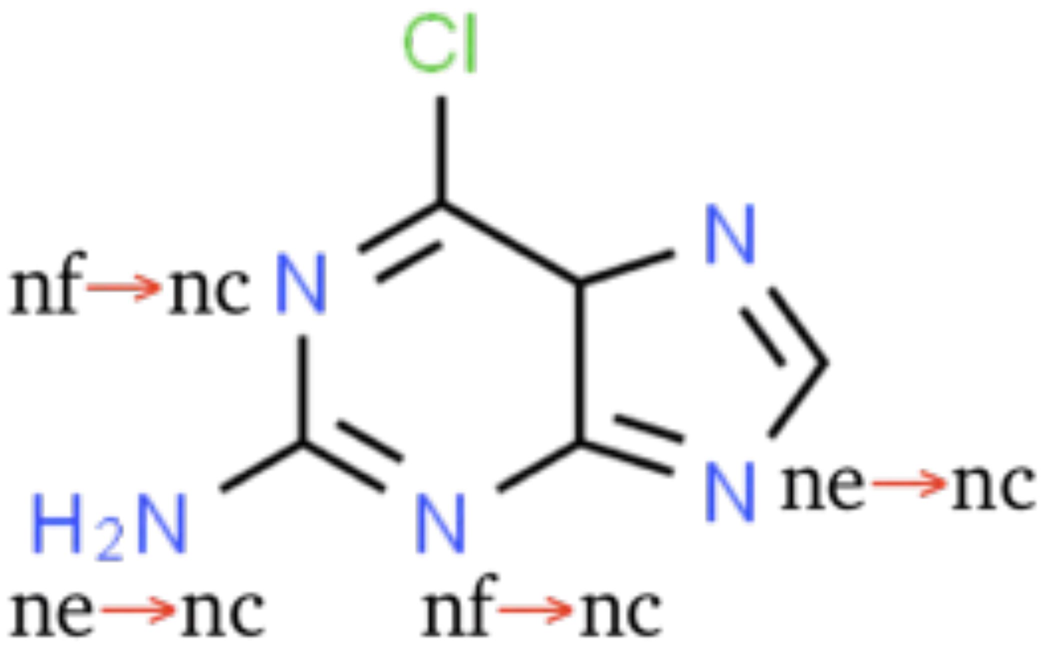
$$\mathbf{e}'_k = \phi^e(\mathbf{e}_k, \mathbf{v}_{rk}, \mathbf{v}_{sk}, \mathbf{u}) \qquad \bar{e}'_i = \rho^{e\to v}(E'_i)$$

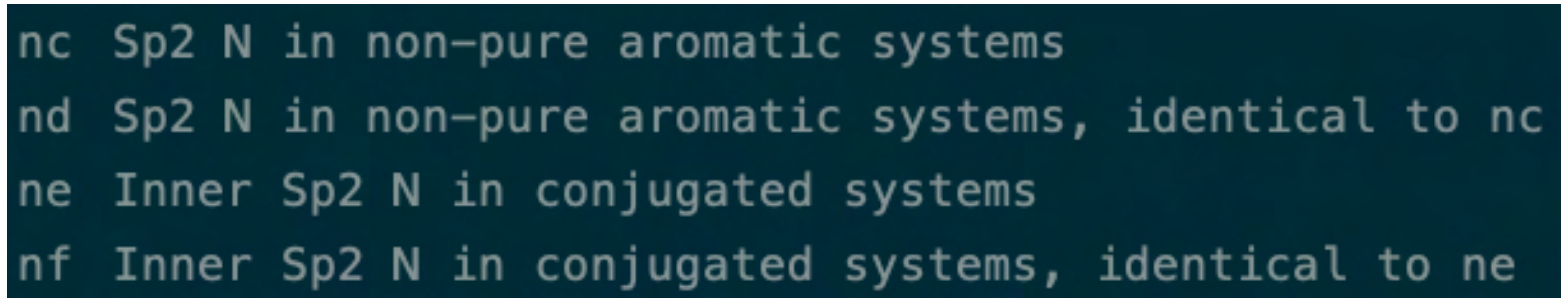
$$\mathbf{v}'_i = \phi^v(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u}) \qquad \bar{e}' = \rho^{e\to u}(E')$$

$$\mathbf{u}' = \phi^u(\bar{\mathbf{e}}', \bar{\mathbf{v}}', \mathbf{u}) \qquad \bar{v}' = \rho^{v\to u}(V')$$

chemical graph of molecule to be parameterized

Graph convolutional network (message-passing network)





```
nc  Sp2 N in non-pure aromatic systems
nd  Sp2 N in non-pure aromatic systems, identical to nc
ne  Inner Sp2 N in conjugated systems
nf  Inner Sp2 N in conjugated systems, identical to ne
```

JOSH FASS

YUANQING WANG

GRAPH CONVOLUTIONAL NETWORKS ARE PARTICULARLY WELL-SUITED TO CHEMISTRY

molecule **bond** **atom**

predict properties

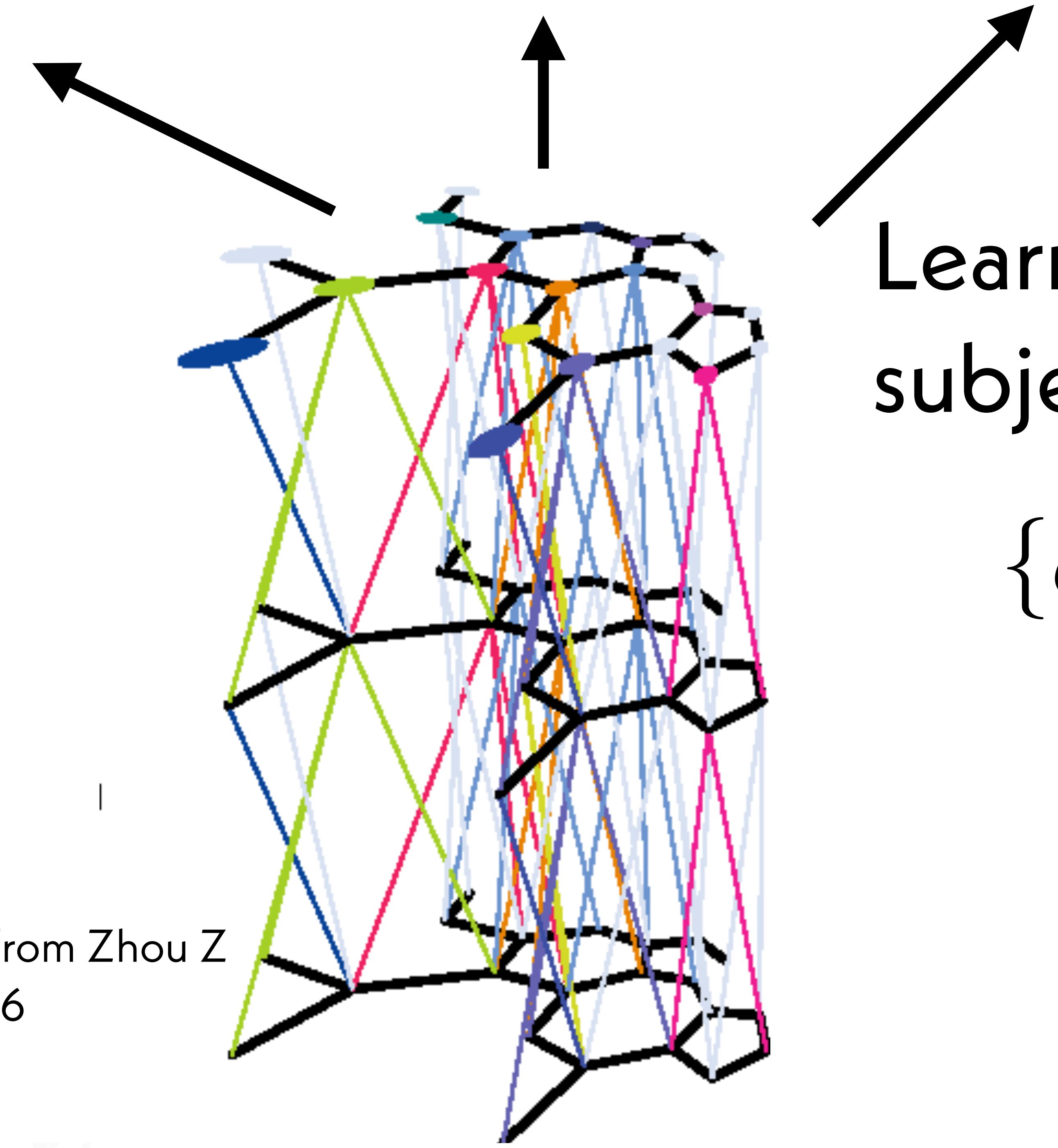
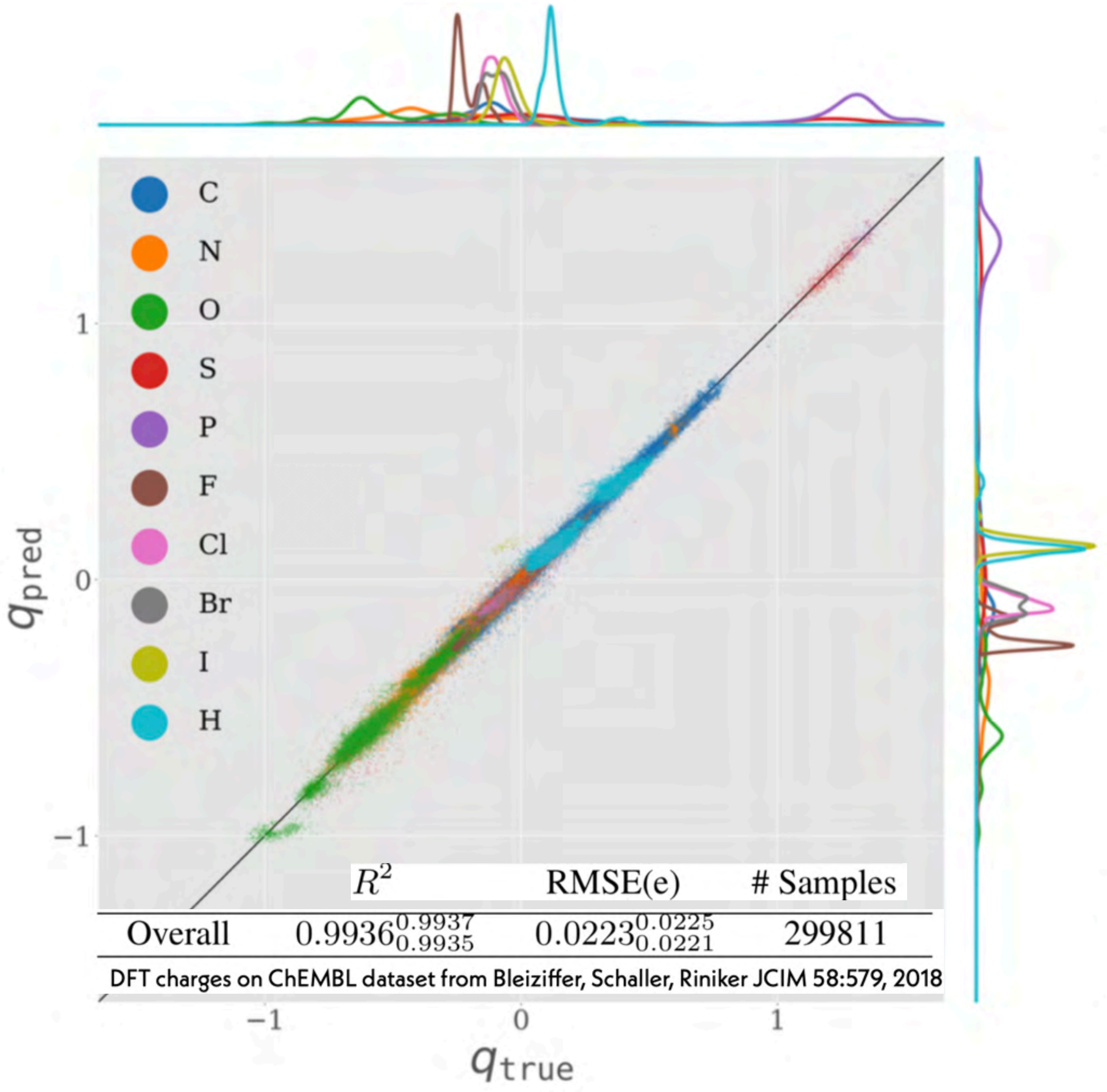


Figure adapted from Zhou Z arXiv:1706.09916

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

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

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



	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

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

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

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

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

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

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

gimlet

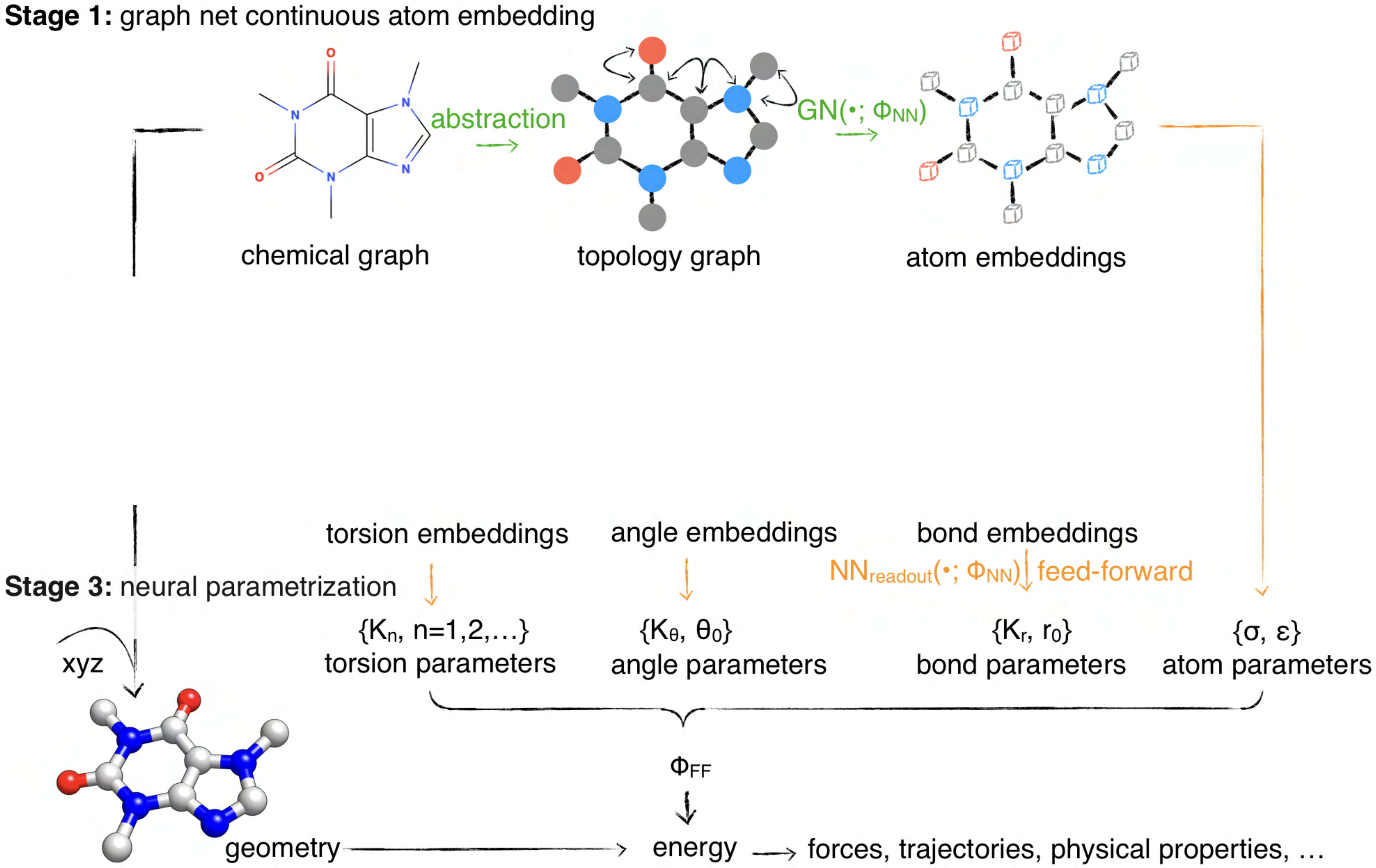
Graph Inference on MoLEcular Topology

preprint: https://arxiv.org/abs/1909.07903

code: http://github.com/choderalab/gimlet

YUANQING WANG

espaloma: **e**xtensible **s**urrogate **p**otential of ***ab*** *initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



Stage 1: graph net continuous atom embedding

chemical graph $\xrightarrow{\text{abstraction}}$ topology graph $\xrightarrow{GN(\cdot;\,\Phi_{NN})}$ atom embeddings

torsion embeddings

angle embeddings

bond embeddings

$NN_{\text{readout}}(\cdot;\,\Phi_{NN})$ feed-forward

Stage 3: neural parametrization

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

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

$\{K_r, r_0\}$ bond parameters

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

Φ_{FF}

xyz

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

JOSH FASS

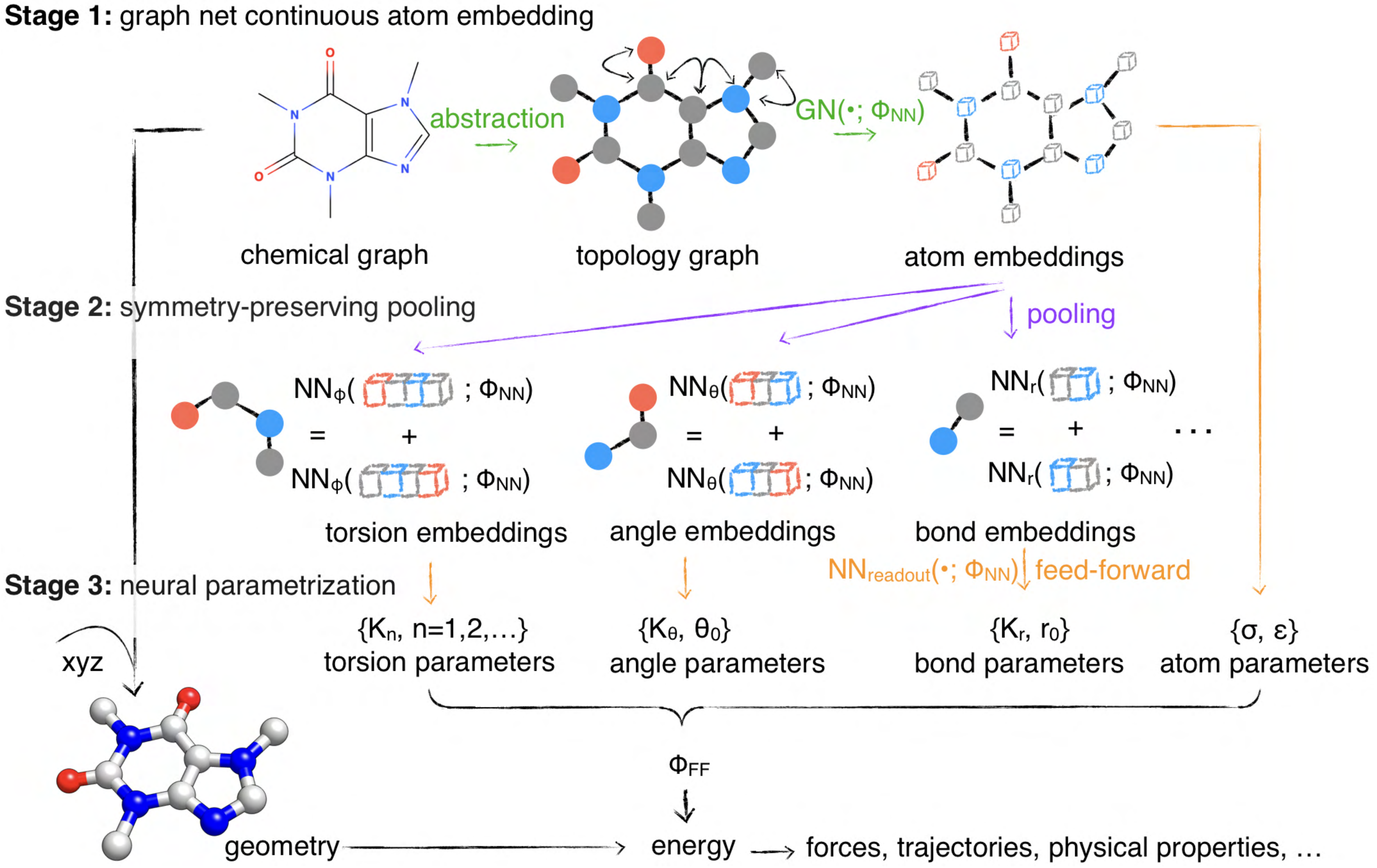
YUANQING WANG

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

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

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

use of only **chemical graph** means that model can generate parameters for small molecules, proteins, nucleic acids, covalent ligands, carbohydrates, etc.



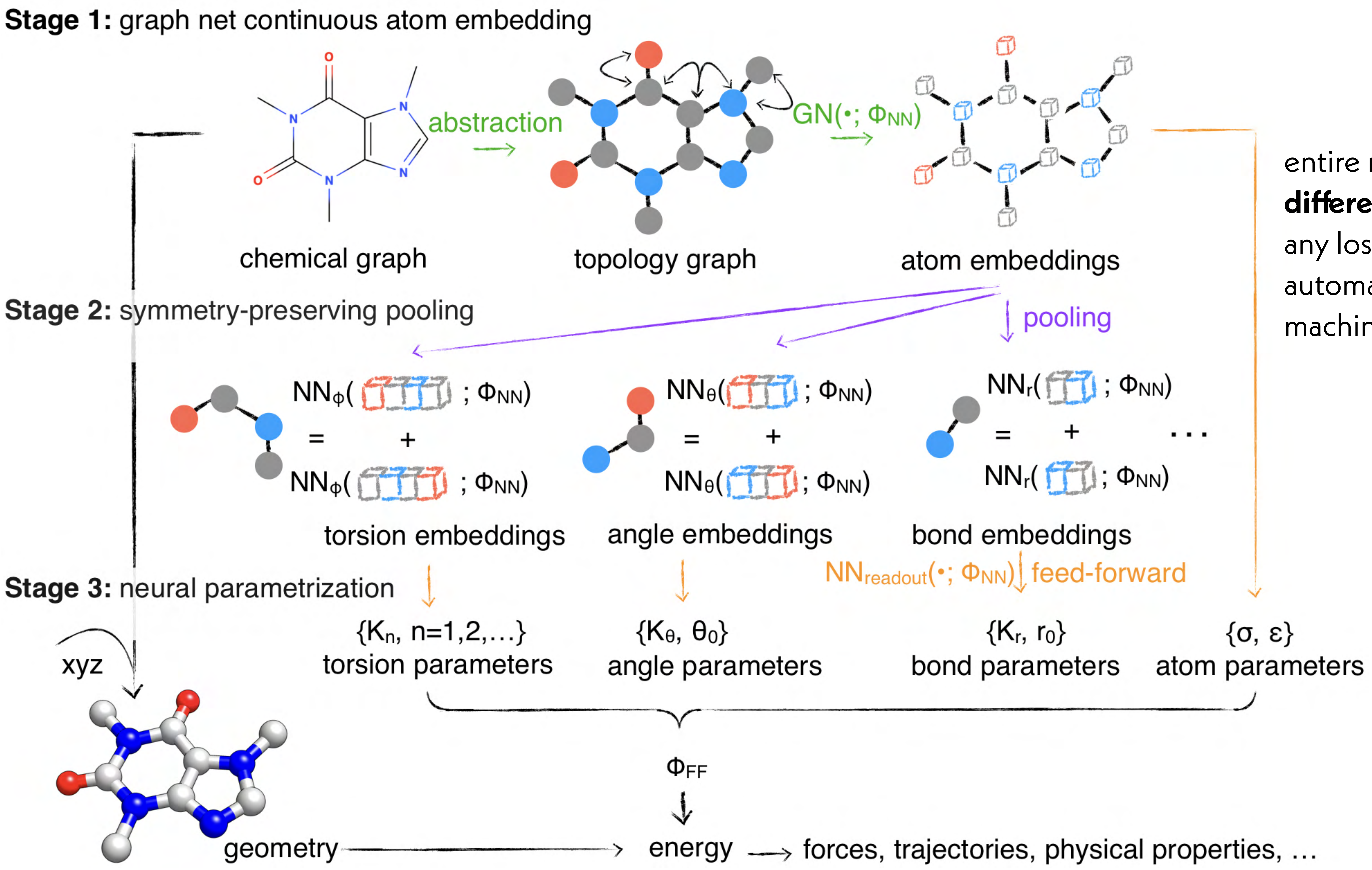
JOSH FASS

YUANQING WANG

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

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

espaloma: **e**xtensible **s**urrogate **p**otential of ***ab*** *initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



entire model is **end-to-end differentiable** so can be fit to any loss function by standard automatic differentiation machine learning frameworks

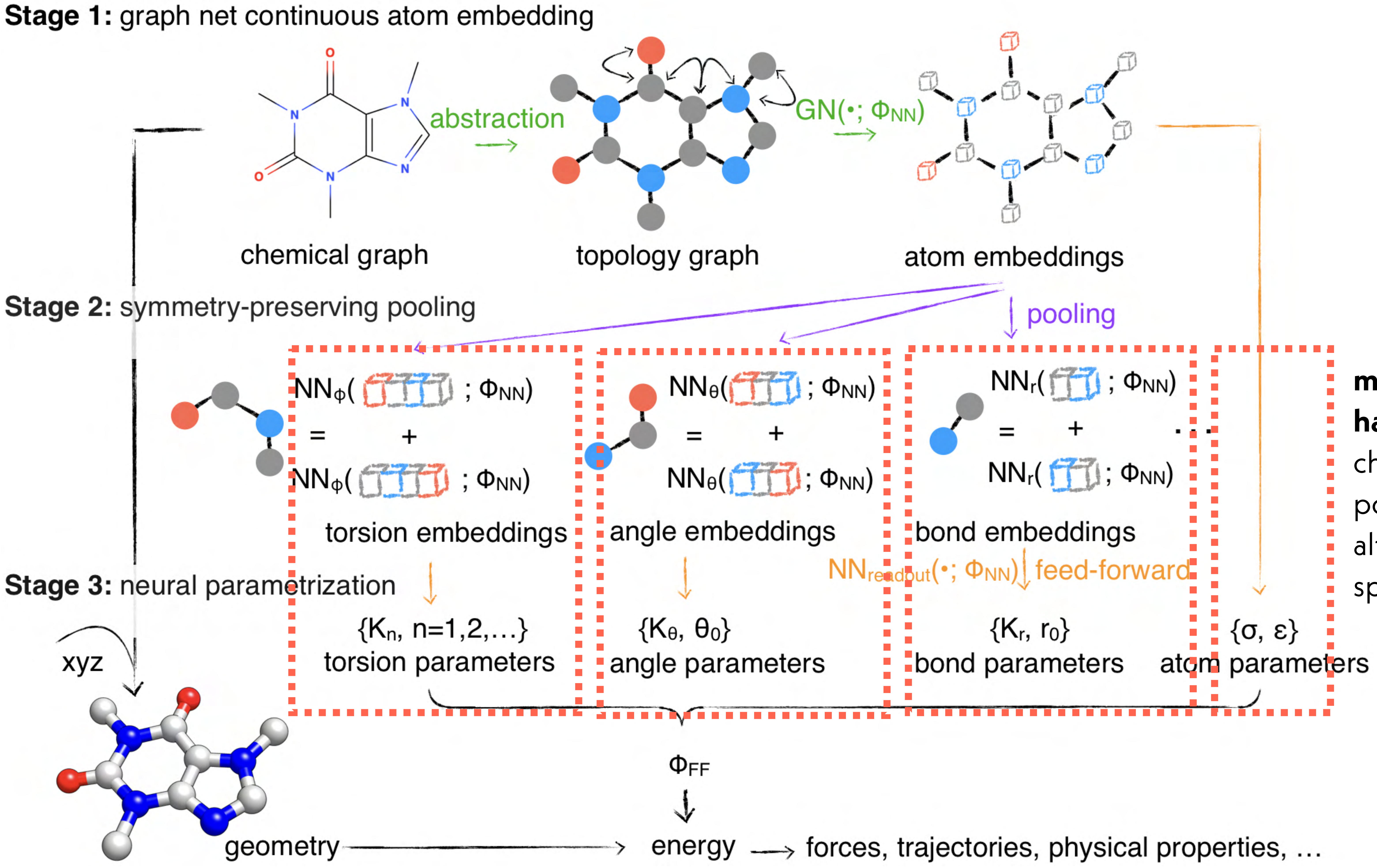
JOSH FASS

YUANQING WANG

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

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

espaloma: **e**xtensible **s**urrogate **p**otential of ***ab*** *initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



Stage 1: graph net continuous atom embedding

chemical graph $\xrightarrow{\text{abstraction}}$ topology graph $\xrightarrow{GN(\cdot;\,\Phi_{NN})}$ atom embeddings

Stage 2: symmetry-preserving pooling

pooling

$NN_\phi(\ ;\ \Phi_{NN}) = + NN_\phi(\ ;\ \Phi_{NN})$ torsion embeddings

$NN_\theta(\ ;\ \Phi_{NN}) = + NN_\theta(\ ;\ \Phi_{NN})$ angle embeddings

$NN_r(\ ;\ \Phi_{NN}) = + NN_r(\ ;\ \Phi_{NN})$ bond embeddings

...

Stage 3: neural parametrization

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

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

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

$\{K_r,\ r_0\}$ bond parameters

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

xyz

Φ_{FF}

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

modular and extensible handling of potential terms:
charge model parameters,
point polarizabilities,
alternative vdW forms,
special 1-4 parameters, etc.

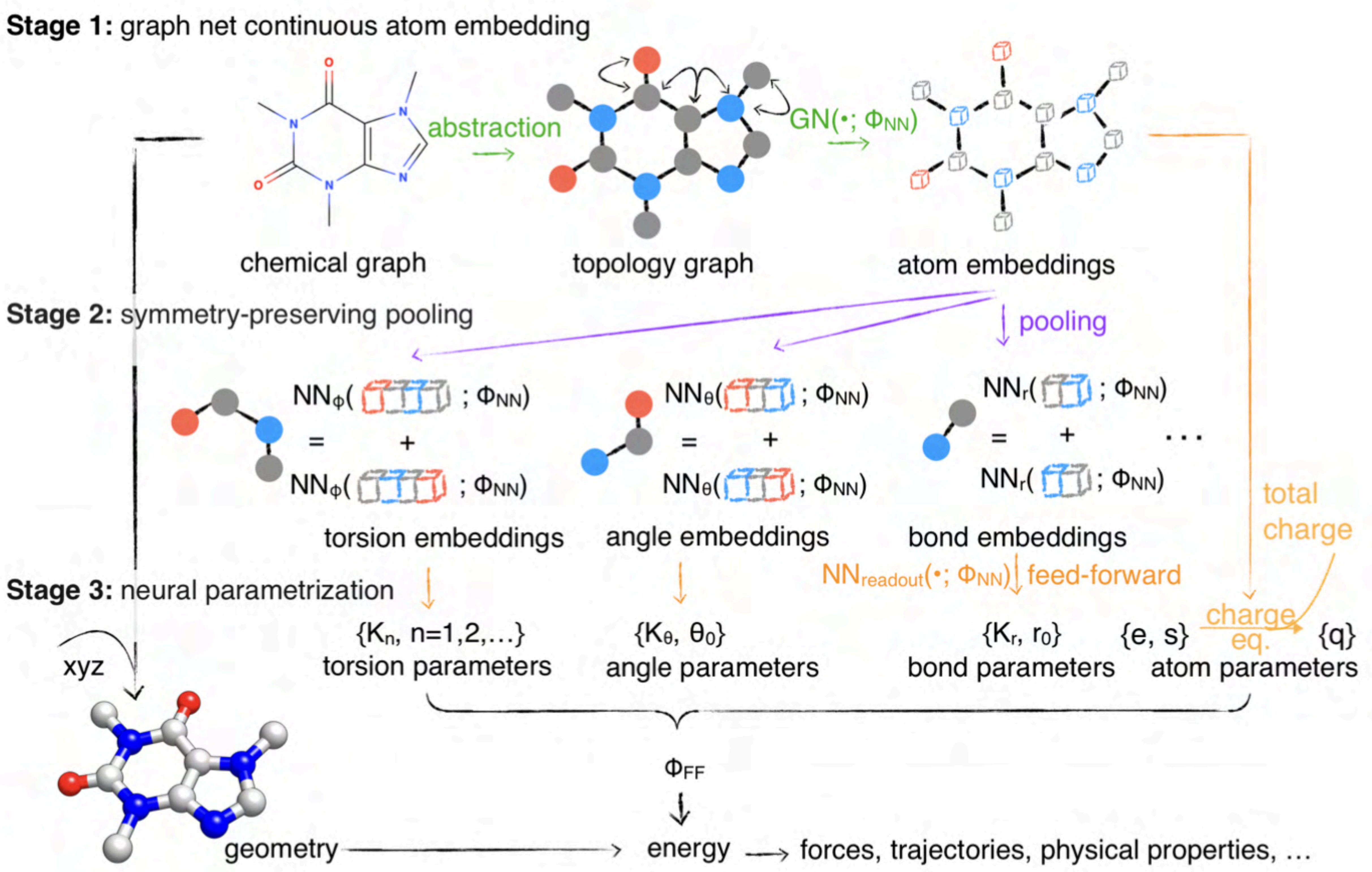
JOSH FASS

YUANQING WANG

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

ESPALOMA MAKES BUILDING A NEW FORCE FIELD EASY

espaloma architecture

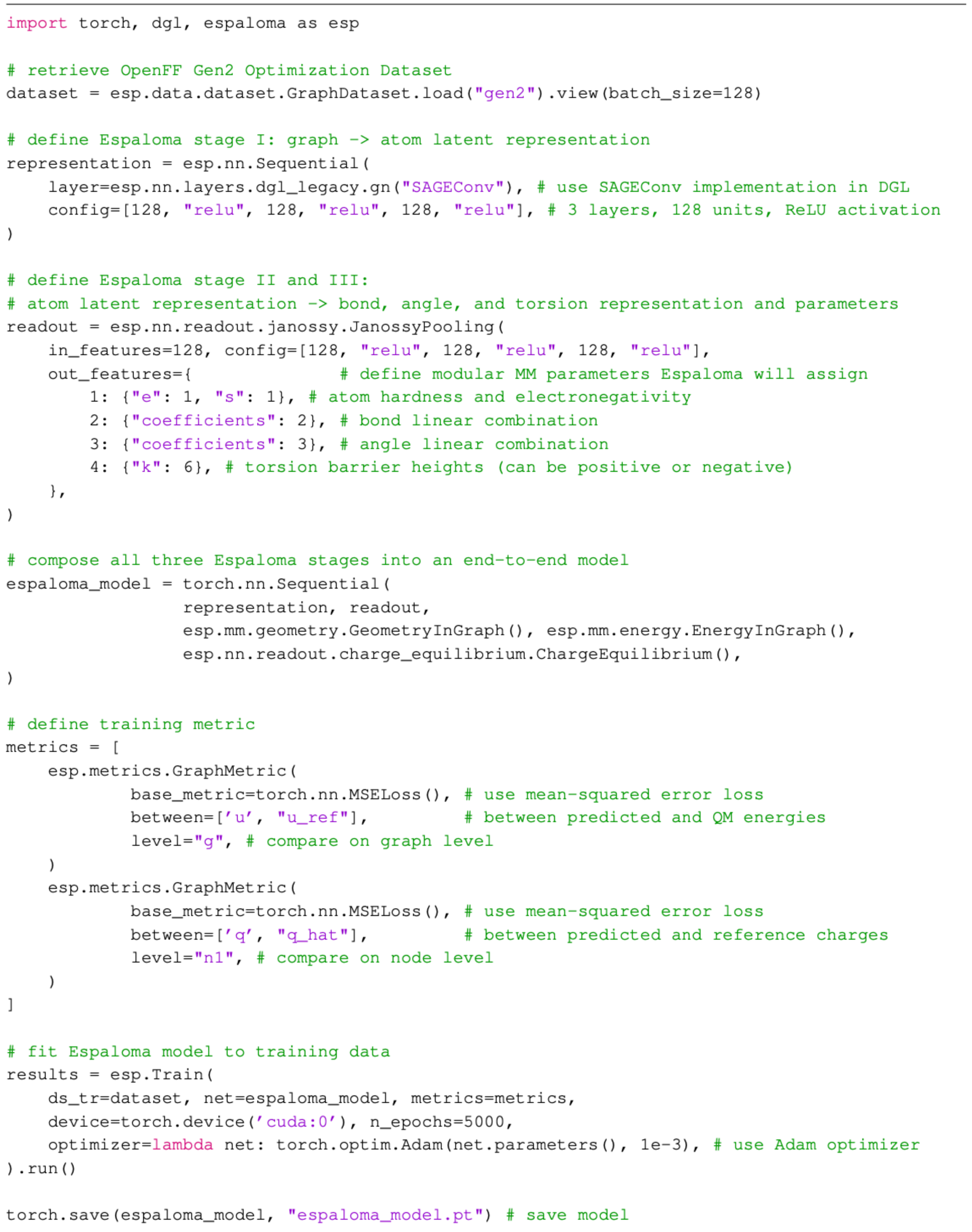

Stage 1: graph net continuous atom embedding
abstraction
GN(•; Φ_NN)
chemical graph
topology graph
atom embeddings
Stage 2: symmetry-preserving pooling
pooling
NN_φ(; Φ_NN) = + NN_φ(; Φ_NN)
NN_θ(; Φ_NN) = + NN_θ(; Φ_NN)
NN_r(; Φ_NN) = + NN_r(; Φ_NN) . . .
torsion embeddings
angle embeddings
bond embeddings
total charge
Stage 3: neural parametrization
NN_readout(•; Φ_NN) feed-forward
{K_n, n=1,2,...}
torsion parameters
{K_θ, θ_0}
angle parameters
{K_r, r_0}
bond parameters
{e, s}
charge eq.
{q}
atom parameters
xyz
Φ_FF
geometry
energy
forces, trajectories, physical properties, ...

(implemented in pytorch)

http://github.com/choderalab/espaloma

YUANQING WANG

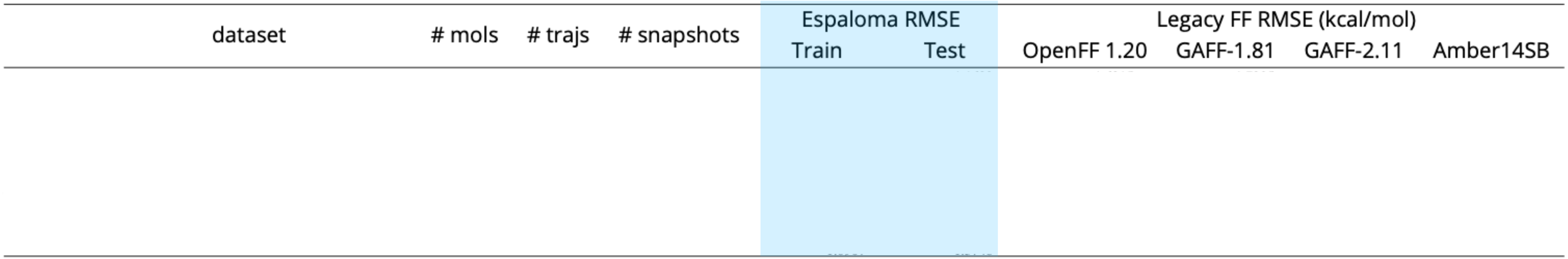
building a new force field



```
import torch, dgl, espaloma as esp

# retrieve OpenFF Gen2 Optimization Dataset
dataset = esp.data.dataset.GraphDataset.load("gen2").view(batch_size=128)

# define Espaloma stage I: graph -> atom latent representation
representation = esp.nn.Sequential(
    layer=esp.nn.layers.dgl_legacy.gn("SAGEConv"), # use SAGEConv implementation in DGL
    config=[128, "relu", 128, "relu", 128, "relu"], # 3 layers, 128 units, ReLU activation
)

# define Espaloma stage II and III:
# atom latent representation -> bond, angle, and torsion representation and parameters
readout = esp.nn.readout.janossy.JanossyPooling(
    in_features=128, config=[128, "relu", 128, "relu", 128, "relu"],
    out_features={              # define modular MM parameters Espaloma will assign
        1: {"e": 1, "s": 1}, # atom hardness and electronegativity
        2: {"coefficients": 2}, # bond linear combination
        3: {"coefficients": 3}, # angle linear combination
        4: {"k": 6}, # torsion barrier heights (can be positive or negative)
    },
)

# compose all three Espaloma stages into an end-to-end model
espaloma_model = torch.nn.Sequential(
                 representation, readout,
                 esp.mm.geometry.GeometryInGraph(), esp.mm.energy.EnergyInGraph(),
                 esp.nn.readout.charge_equilibrium.ChargeEquilibrium(),
)

# define training metric
metrics = [
    esp.metrics.GraphMetric(
            base_metric=torch.nn.MSELoss(), # use mean-squared error loss
            between=['u', "u_ref"],         # between predicted and QM energies
            level="g", # compare on graph level
    )
    esp.metrics.GraphMetric(
            base_metric=torch.nn.MSELoss(), # use mean-squared error loss
            between=['q', "q_hat"],         # between predicted and reference charges
            level="n1", # compare on node level
    )
]

# fit Espaloma model to training data
results = esp.Train(
    ds_tr=dataset, net=espaloma_model, metrics=metrics,
    device=torch.device('cuda:0'), n_epochs=5000,
    optimizer=lambda net: torch.optim.Adam(net.parameters(), 1e-3), # use Adam optimizer
).run()

torch.save(espaloma_model, "espaloma_model.pt") # save model
```

Listing 1. Defining and training a modular Espaloma model.

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



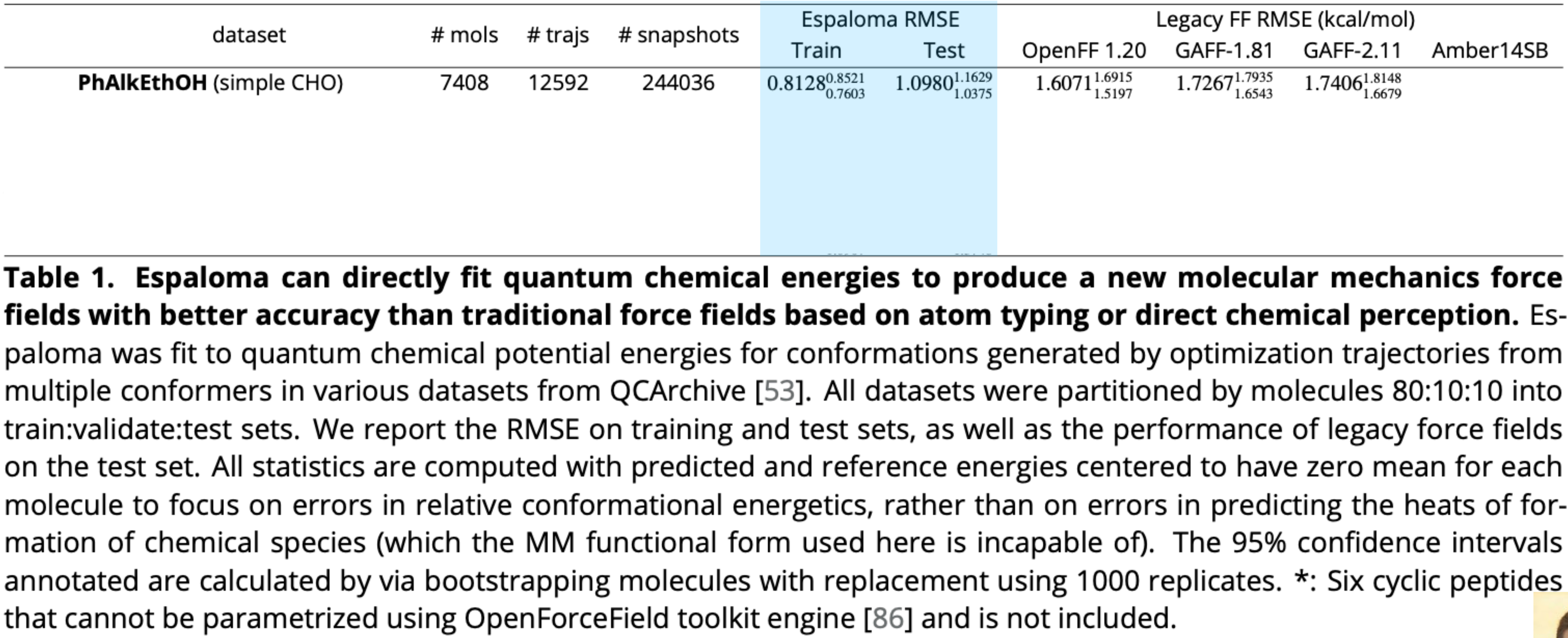
dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol)			
				Train	Test	OpenFF 1.20	GAFF-1.81	GAFF-2.11	Amber14SB

Table 1. Espaloma can directly fit quantum chemical energies to produce a new molecular mechanics force fields with better accuracy than traditional force fields based on atom typing or direct chemical perception. Espaloma was fit to quantum chemical potential energies for conformations generated by optimization trajectories from multiple conformers in various datasets from QCArchive [53]. All datasets were partitioned by molecules 80:10:10 into train:validate:test sets. We report the RMSE on training and test sets, as well as the performance of legacy force fields on the test set. All statistics are computed with predicted and reference energies centered to have zero mean for each molecule to focus on errors in relative conformational energetics, rather than on errors in predicting the heats of formation of chemical species (which the MM functional form used here is incapable of). The 95% confidence intervals annotated are calculated by via bootstrapping molecules with replacement using 1000 replicates. *: Six cyclic peptides that cannot be parametrized using OpenForceField toolkit engine [86] and is not included.

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

YUANQING WANG

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



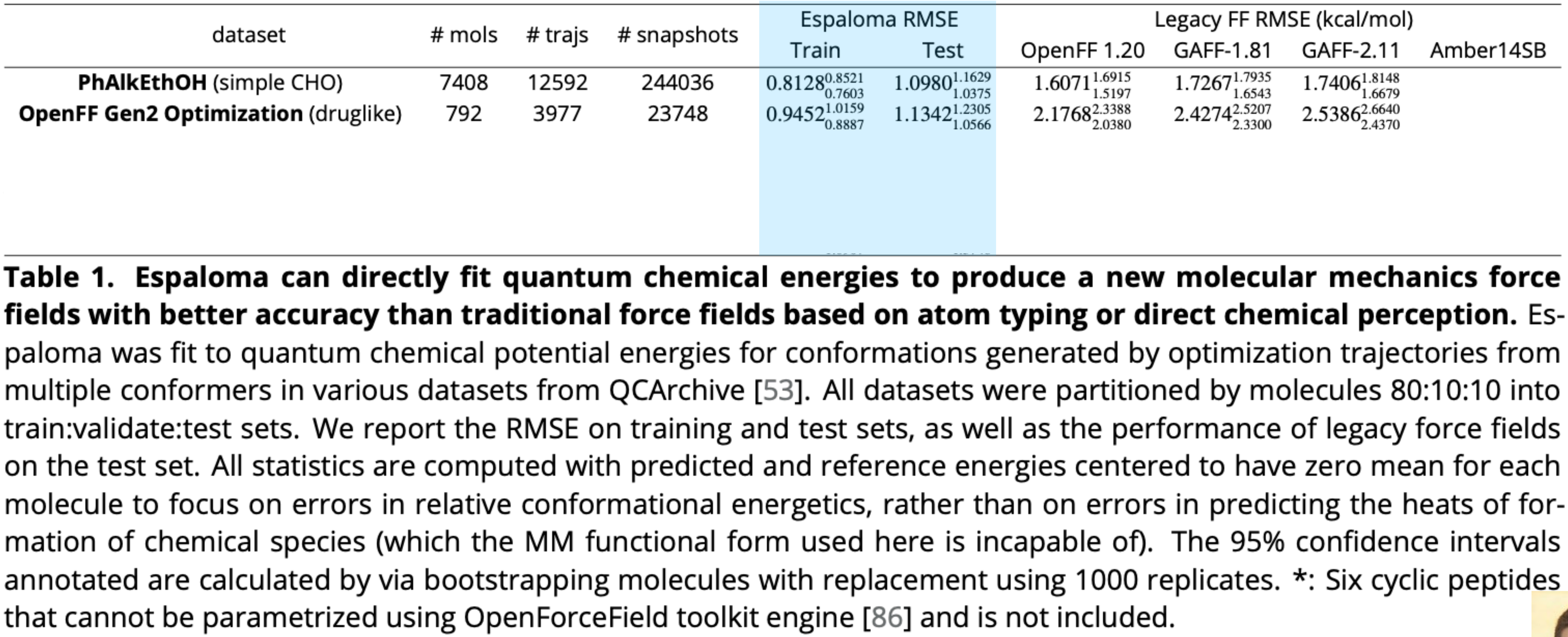
dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol)			
				Train	Test	OpenFF 1.20	GAFF-1.81	GAFF-2.11	Amber14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8128^{0.8521}_{0.7603}$	$1.0980^{1.1629}_{1.0375}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	

Table 1. Espaloma can directly fit quantum chemical energies to produce a new molecular mechanics force fields with better accuracy than traditional force fields based on atom typing or direct chemical perception. Espaloma was fit to quantum chemical potential energies for conformations generated by optimization trajectories from multiple conformers in various datasets from QCArchive [53]. All datasets were partitioned by molecules 80:10:10 into train:validate:test sets. We report the RMSE on training and test sets, as well as the performance of legacy force fields on the test set. All statistics are computed with predicted and reference energies centered to have zero mean for each molecule to focus on errors in relative conformational energetics, rather than on errors in predicting the heats of formation of chemical species (which the MM functional form used here is incapable of). The 95% confidence intervals annotated are calculated by via bootstrapping molecules with replacement using 1000 replicates. *: Six cyclic peptides that cannot be parametrized using OpenForceField toolkit engine [86] and is not included.

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

YUANQING WANG

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



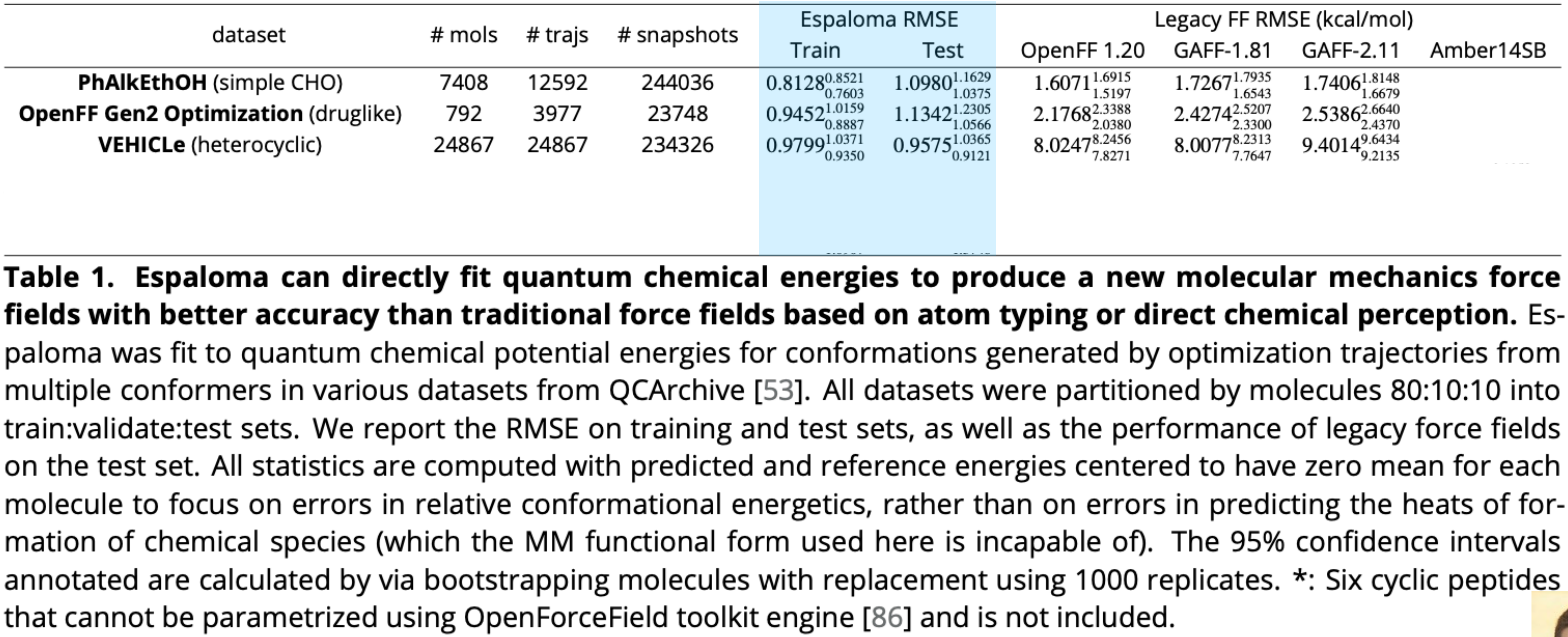
dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol)			
				Train	Test	OpenFF 1.20	GAFF-1.81	GAFF-2.11	Amber14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8128^{0.8521}_{0.7603}$	$1.0980^{1.1629}_{1.0375}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.9452^{1.0159}_{0.8887}$	$1.1342^{1.2305}_{1.0566}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	

Table 1. Espaloma can directly fit quantum chemical energies to produce a new molecular mechanics force fields with better accuracy than traditional force fields based on atom typing or direct chemical perception. Espaloma was fit to quantum chemical potential energies for conformations generated by optimization trajectories from multiple conformers in various datasets from QCArchive [53]. All datasets were partitioned by molecules 80:10:10 into train:validate:test sets. We report the RMSE on training and test sets, as well as the performance of legacy force fields on the test set. All statistics are computed with predicted and reference energies centered to have zero mean for each molecule to focus on errors in relative conformational energetics, rather than on errors in predicting the heats of formation of chemical species (which the MM functional form used here is incapable of). The 95% confidence intervals annotated are calculated by via bootstrapping molecules with replacement using 1000 replicates. *: Six cyclic peptides that cannot be parametrized using OpenForceField toolkit engine [86] and is not included.

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

YUANQING WANG

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



dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol)			
				Train	Test	OpenFF 1.20	GAFF-1.81	GAFF-2.11	Amber14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8128^{0.8521}_{0.7603}$	$1.0980^{1.1629}_{1.0375}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.9452^{1.0159}_{0.8887}$	$1.1342^{1.2305}_{1.0566}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)	24867	24867	234326	$0.9799^{1.0371}_{0.9350}$	$0.9575^{1.0365}_{0.9121}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	

Table 1. Espaloma can directly fit quantum chemical energies to produce a new molecular mechanics force fields with better accuracy than traditional force fields based on atom typing or direct chemical perception. Espaloma was fit to quantum chemical potential energies for conformations generated by optimization trajectories from multiple conformers in various datasets from QCArchive [53]. All datasets were partitioned by molecules 80:10:10 into train:validate:test sets. We report the RMSE on training and test sets, as well as the performance of legacy force fields on the test set. All statistics are computed with predicted and reference energies centered to have zero mean for each molecule to focus on errors in relative conformational energetics, rather than on errors in predicting the heats of formation of chemical species (which the MM functional form used here is incapable of). The 95% confidence intervals annotated are calculated by via bootstrapping molecules with replacement using 1000 replicates. *: Six cyclic peptides that cannot be parametrized using OpenForceField toolkit engine [86] and is not included.

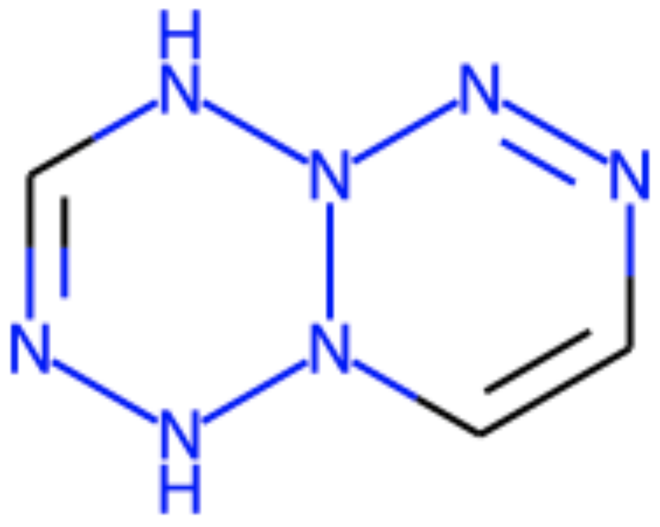
preprint: https://arxiv.org/abs/2010.01196

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

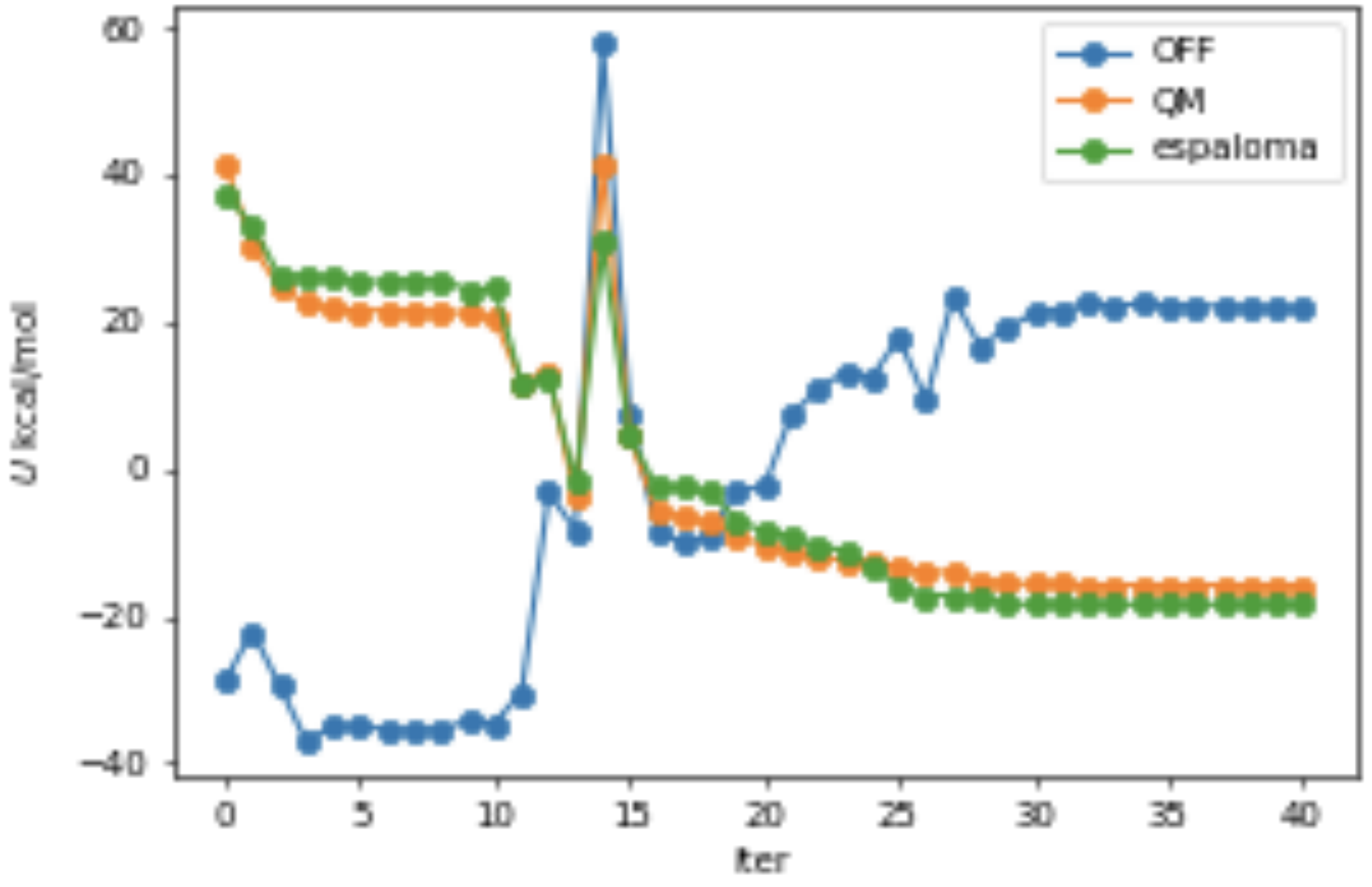
YUANQING WANG

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

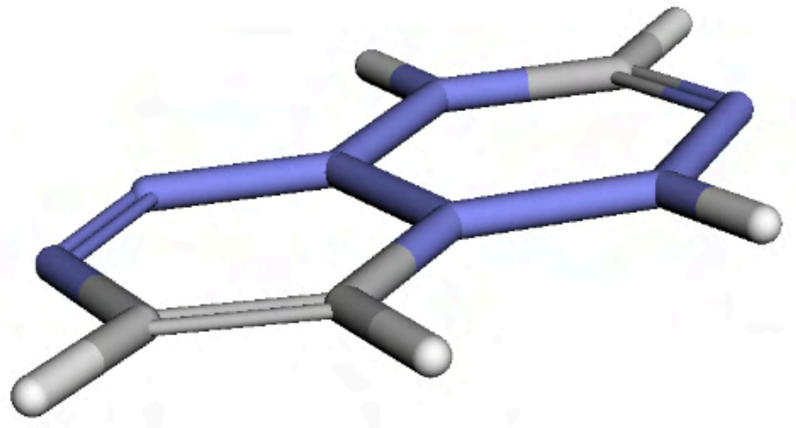
dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol)			
				Train	Test	OpenFF 1.20	GAFF-1.81	GAFF-2.11	Amber14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8128^{0.8521}_{0.7603}$	$1.0980^{1.1629}_{1.0375}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.9452^{1.0159}_{0.8887}$	$1.1342^{1.2305}_{1.0566}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)	24867	24867	234326	$0.9799^{1.0371}_{0.9350}$	$0.9575^{1.0365}_{0.9121}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	



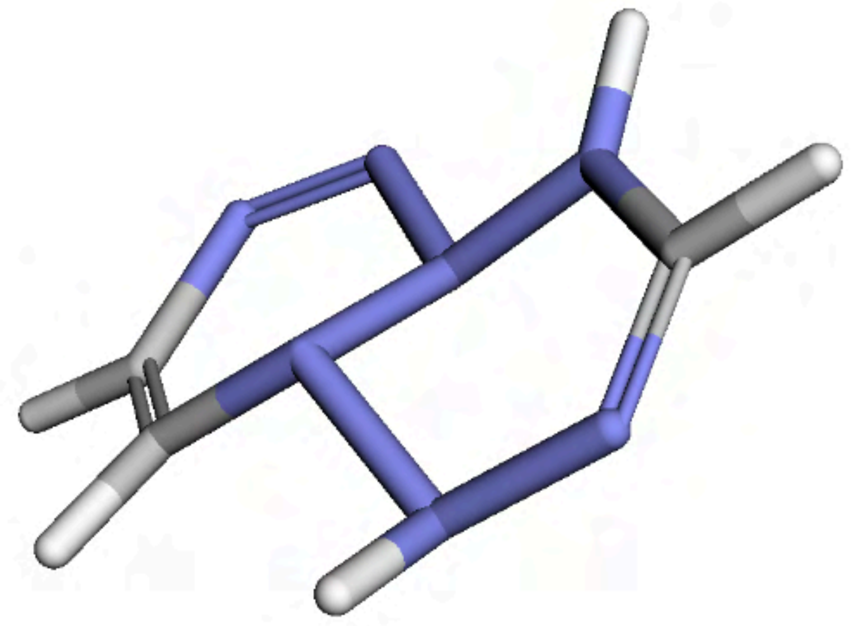
Comparison with QCArchive data



initial



QM minimized

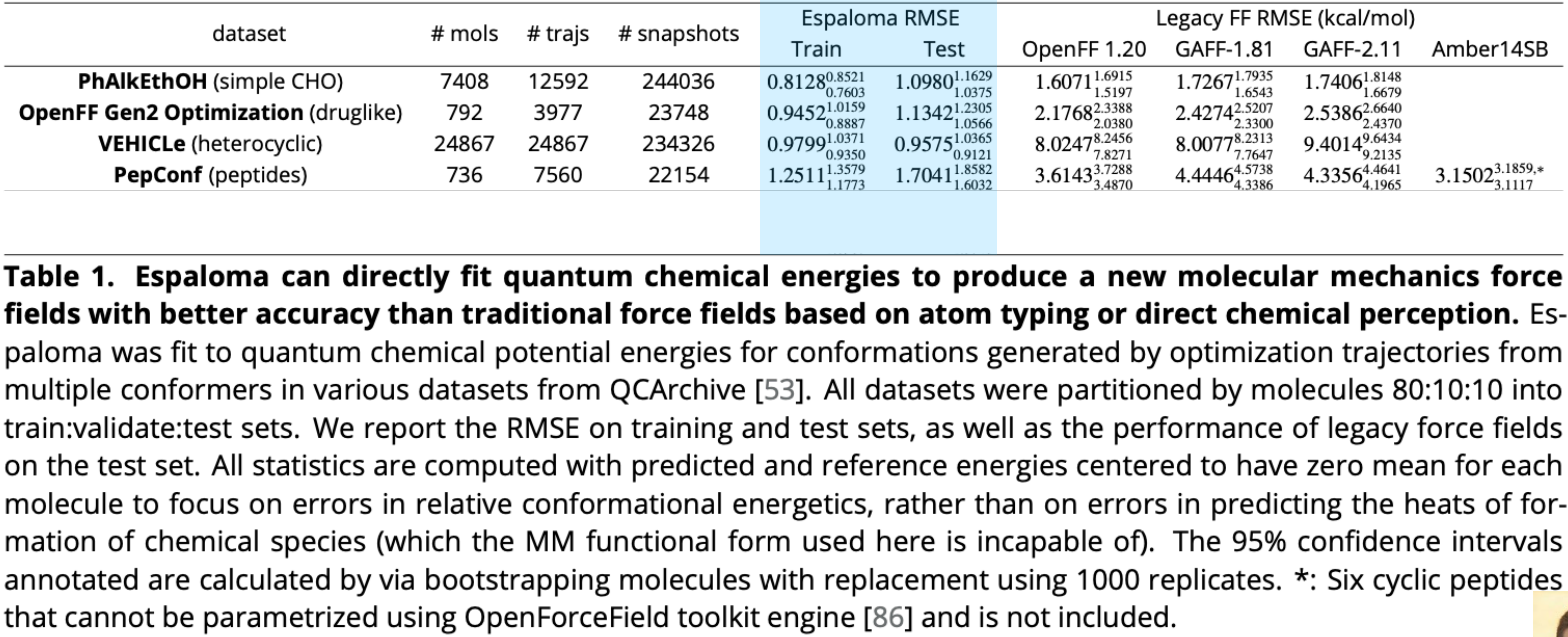


DFT B3LYP-D3(BJ) / DZVP

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

YUANQING WANG

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



dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol)			
				Train	Test	OpenFF 1.20	GAFF-1.81	GAFF-2.11	Amber14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8128^{0.8521}_{0.7603}$	$1.0980^{1.1629}_{1.0375}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.9452^{1.0159}_{0.8887}$	$1.1342^{1.2305}_{1.0566}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)	24867	24867	234326	$0.9799^{1.0371}_{0.9350}$	$0.9575^{1.0365}_{0.9121}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	
PepConf (peptides)	736	7560	22154	$1.2511^{1.3579}_{1.1773}$	$1.7041^{1.8582}_{1.6032}$	$3.6143^{3.7288}_{3.4870}$	$4.4446^{4.5738}_{4.3386}$	$4.3356^{4.4641}_{4.1965}$	$3.1502^{3.1859,*}_{3.1117}$

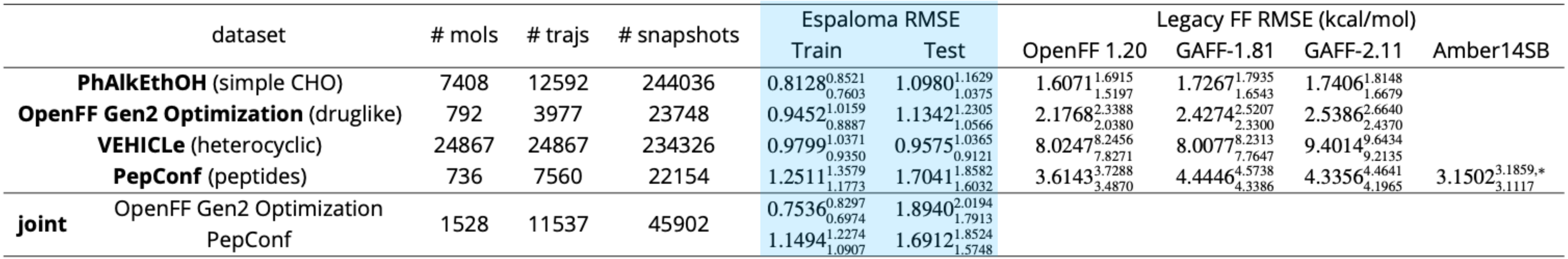
Table 1. Espaloma can directly fit quantum chemical energies to produce a new molecular mechanics force fields with better accuracy than traditional force fields based on atom typing or direct chemical perception. Espaloma was fit to quantum chemical potential energies for conformations generated by optimization trajectories from multiple conformers in various datasets from QCArchive [53]. All datasets were partitioned by molecules 80:10:10 into train:validate:test sets. We report the RMSE on training and test sets, as well as the performance of legacy force fields on the test set. All statistics are computed with predicted and reference energies centered to have zero mean for each molecule to focus on errors in relative conformational energetics, rather than on errors in predicting the heats of formation of chemical species (which the MM functional form used here is incapable of). The 95% confidence intervals annotated are calculated by via bootstrapping molecules with replacement using 1000 replicates. *: Six cyclic peptides that cannot be parametrized using OpenForceField toolkit engine [86] and is not included.

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

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

YUANQING WANG

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



	dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol)			
					Train	Test	OpenFF 1.20	GAFF-1.81	GAFF-2.11	Amber14SB
	PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8128_{0.7603}^{0.8521}$	$1.0980_{1.0375}^{1.1629}$	$1.6071_{1.5197}^{1.6915}$	$1.7267_{1.6543}^{1.7935}$	$1.7406_{1.6679}^{1.8148}$	
	OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.9452_{0.8887}^{1.0159}$	$1.1342_{1.0566}^{1.2305}$	$2.1768_{2.0380}^{2.3388}$	$2.4274_{2.3300}^{2.5207}$	$2.5386_{2.4370}^{2.6640}$	
	VEHICLe (heterocyclic)	24867	24867	234326	$0.9799_{0.9350}^{1.0371}$	$0.9575_{0.9121}^{1.0365}$	$8.0247_{7.8271}^{8.2456}$	$8.0077_{7.7647}^{8.2313}$	$9.4014_{9.2135}^{9.6434}$	
	PepConf (peptides)	736	7560	22154	$1.2511_{1.1773}^{1.3579}$	$1.7041_{1.6032}^{1.8582}$	$3.6143_{3.4870}^{3.7288}$	$4.4446_{4.3386}^{4.5738}$	$4.3356_{4.1965}^{4.4641}$	$3.1502_{3.1117}^{3.1859,*}$
joint	OpenFF Gen2 Optimization	1528	11537	45902	$0.7536_{0.6974}^{0.8297}$	$1.8940_{1.7913}^{2.0194}$				
	PepConf				$1.1494_{1.0907}^{1.2274}$	$1.6912_{1.5748}^{1.8524}$				

Table 1. Espaloma can directly fit quantum chemical energies to produce a new molecular mechanics force fields with better accuracy than traditional force fields based on atom typing or direct chemical perception. Espaloma was fit to quantum chemical potential energies for conformations generated by optimization trajectories from multiple conformers in various datasets from QCArchive [53]. All datasets were partitioned by molecules 80:10:10 into train:validate:test sets. We report the RMSE on training and test sets, as well as the performance of legacy force fields on the test set. All statistics are computed with predicted and reference energies centered to have zero mean for each molecule to focus on errors in relative conformational energetics, rather than on errors in predicting the heats of formation of chemical species (which the MM functional form used here is incapable of). The 95% confidence intervals annotated are calculated by via bootstrapping molecules with replacement using 1000 replicates. *: Six cyclic peptides that cannot be parametrized using OpenForceField toolkit engine [86] and is not included.

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

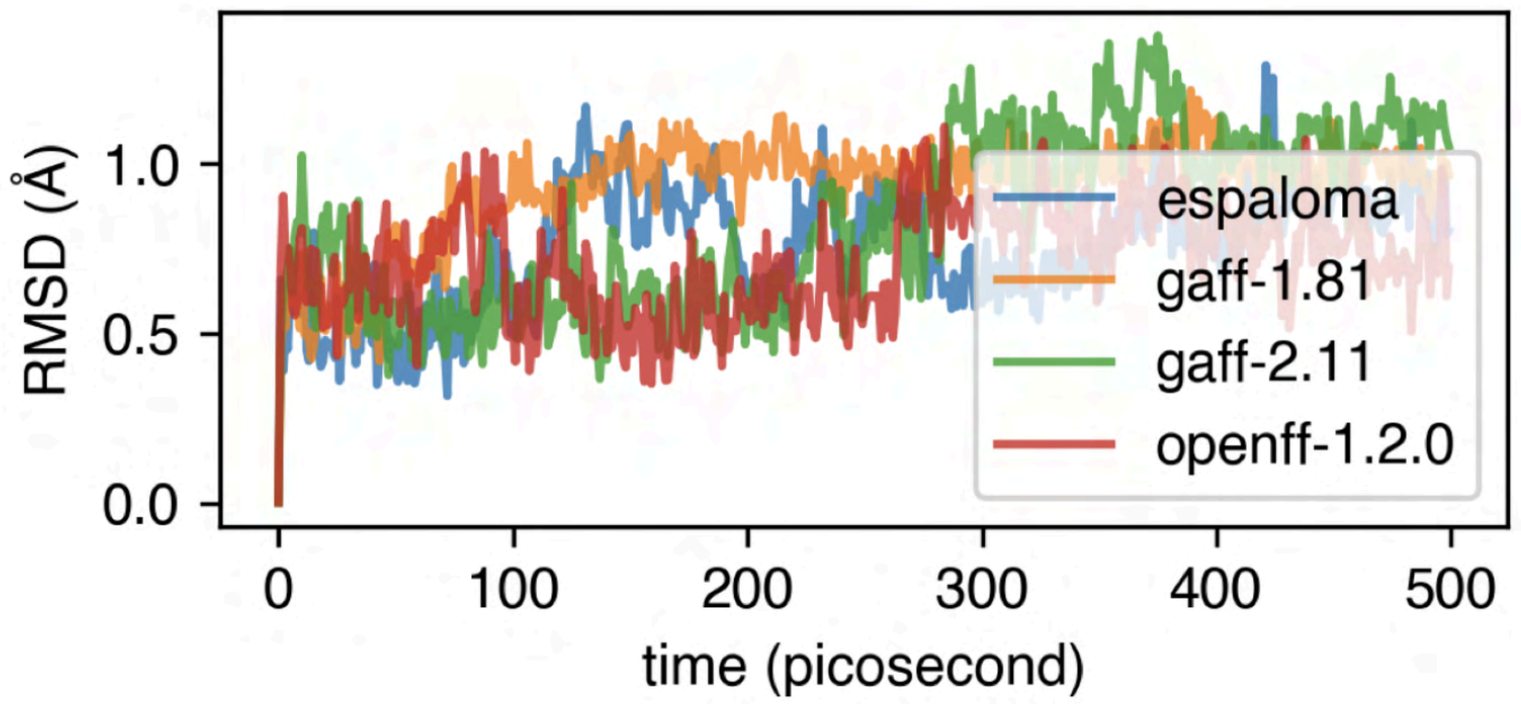
YUANQING WANG

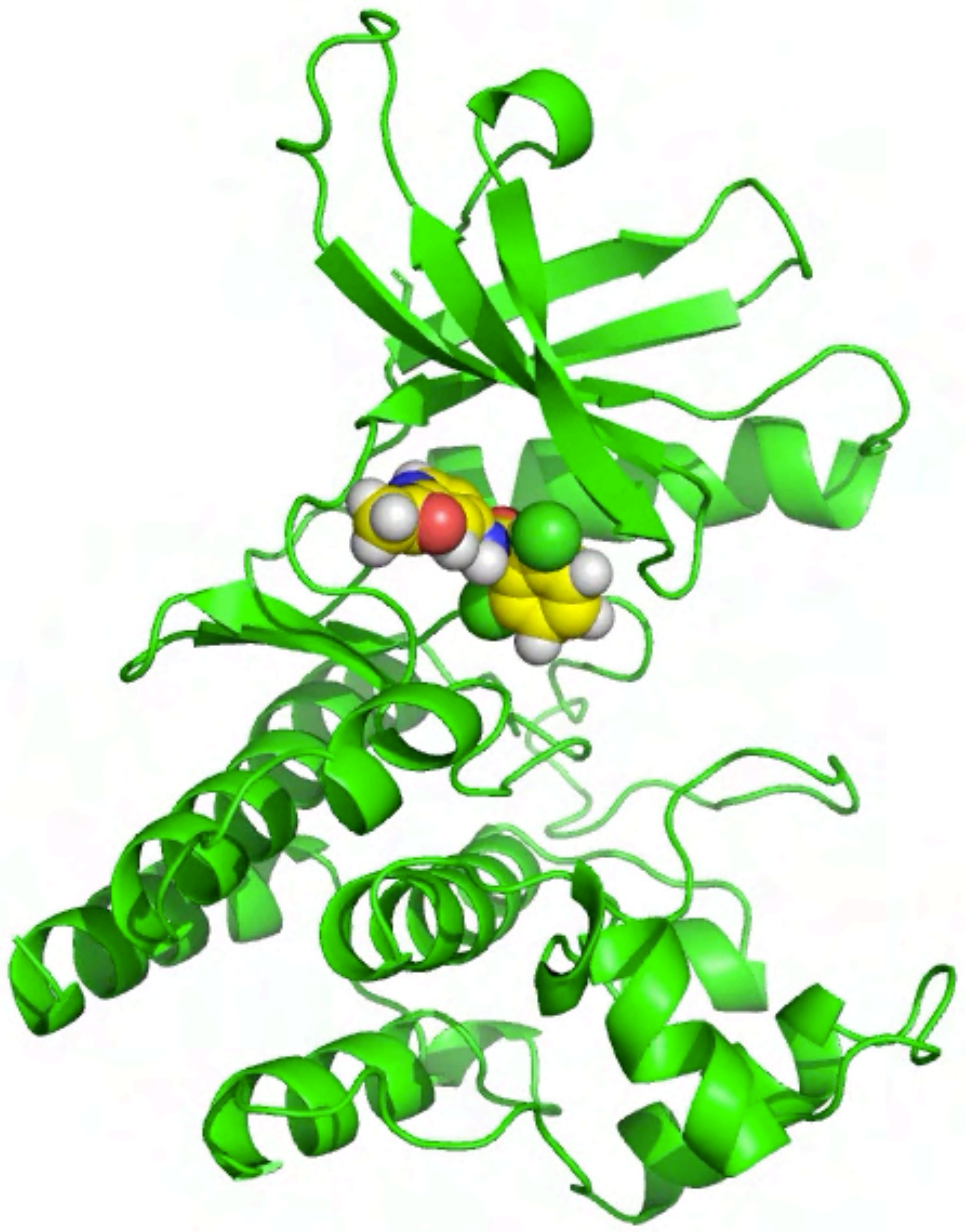
ESPALOMA OUTPERFORMS CURRENT FORCE FIELDS IN QM ACCURACY AND CAN BE EASILY TRAINED FOR HETEROGENEOUS SYSTEMS

espaloma can produce a complete protein+ligand force field suitable for simulation

joint	OpenFF Gen2 Optimization	1528	11537	45902	$0.7536^{0.8297}_{0.6974}$	$1.8940^{2.0194}_{1.7913}$
	PepConf				$1.1494^{1.2274}_{1.0907}$	$1.6912^{1.8524}_{1.5748}$

ligand heavy atom RMSD (after aligning protein)



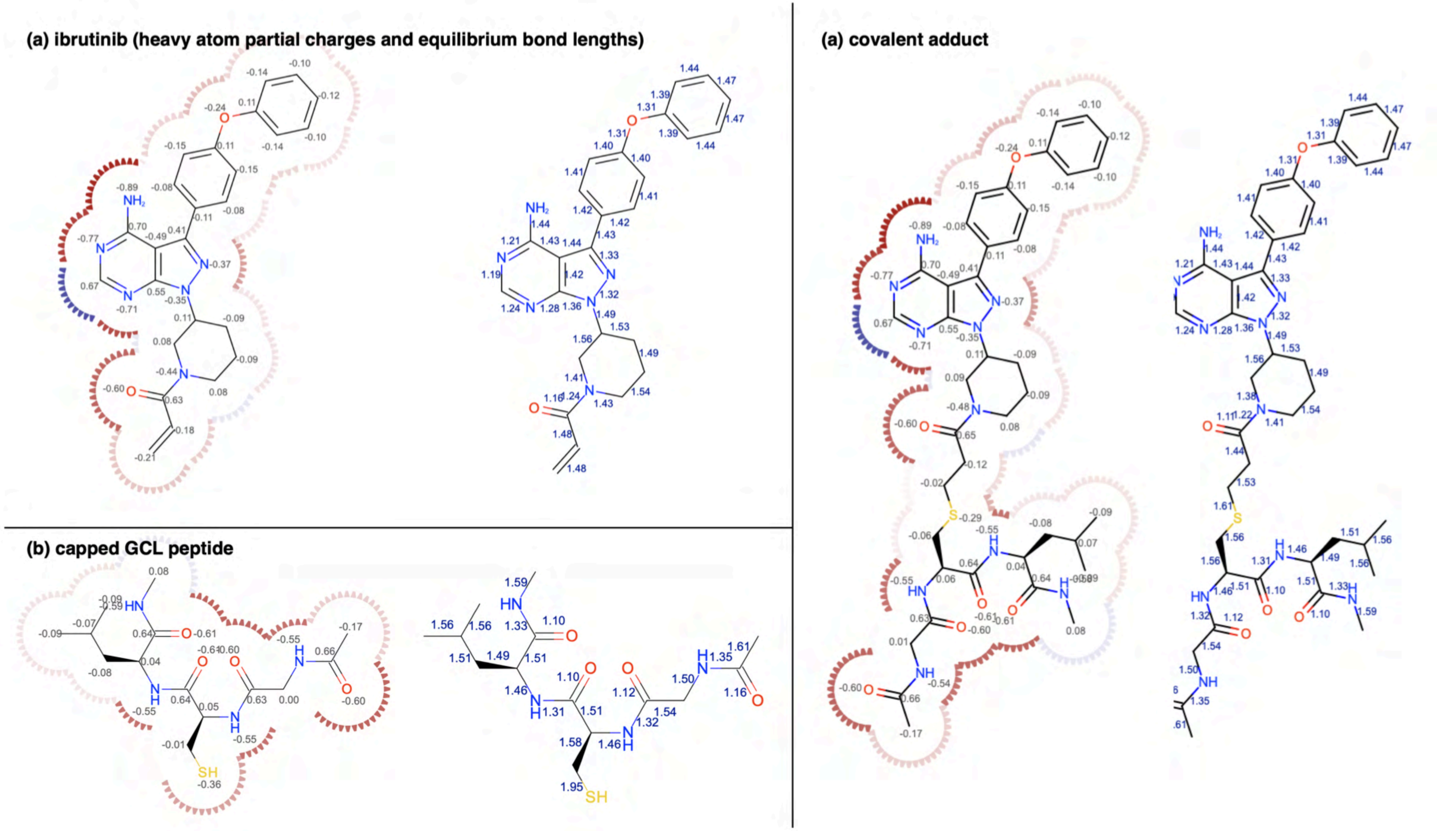


Tyk2 from OpenFF benchmark set
espaloma force field (protein/ligand)
+ TIP3P water
https://arxiv.org/abs/2105.06222

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

YUANQING WANG

ESPALOMA SELF-CONSISTENTLY TREATS BIOPOLYMERS, SMALL MOLECULES, AND COVALENT LIGANDS/MODIFICATIONS



(a) ibrutinib (heavy atom partial charges and equilibrium bond lengths)

(b) capped GCL peptide

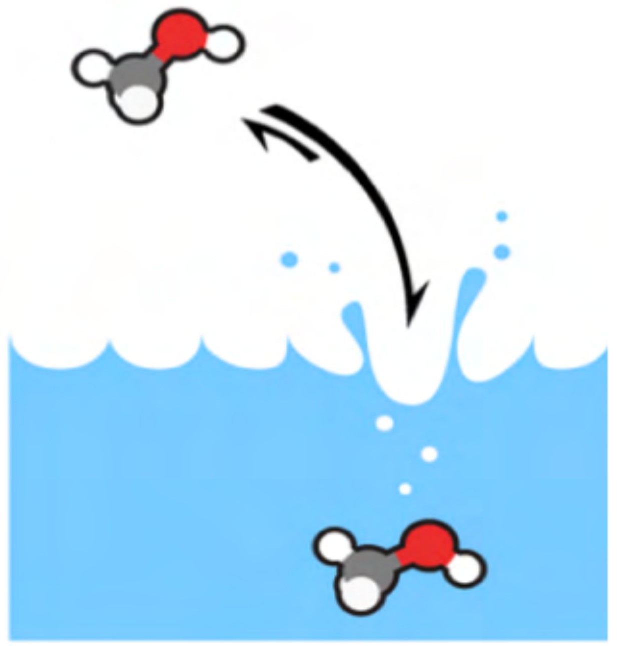
(a) covalent adduct

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

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

YUANQING WANG

ESPALOMA CAN EASILY FIT BOTH QUANTUM CHEMICAL AND EXPERIMENTAL FREE ENERGIES

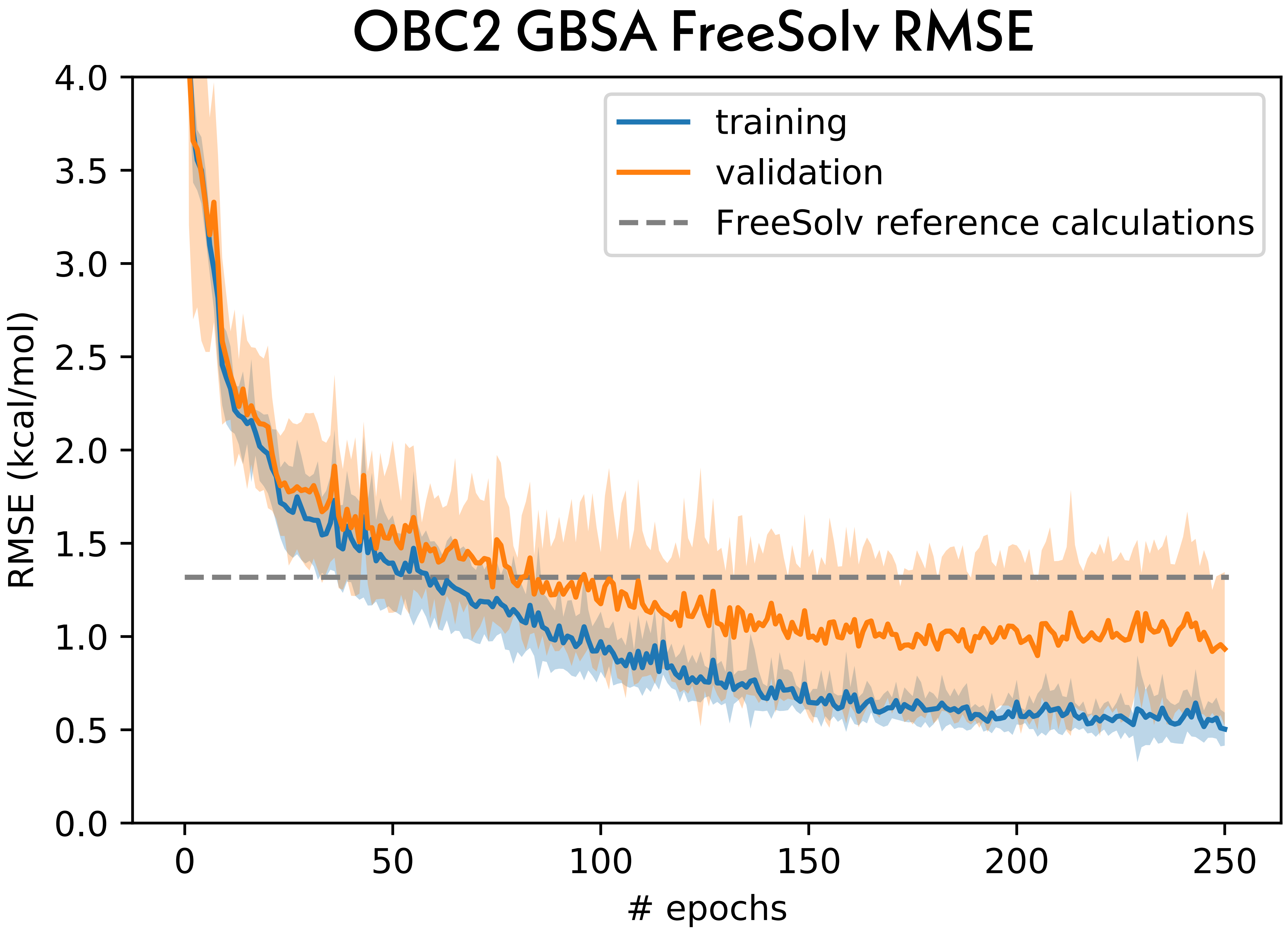


experimental hydration
free energies from **FreeSolv**
https://github.com/MobleyLab/FreeSolv

loss function:

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

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



JOSH FASS

YUANQING WANG

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

CLASS II FORCE FIELDS MAY PROVIDE SUBSTANTIALLY INCREASED ACCURACY WITH RESPECT TO QUANTUM CHEMISTRY AT MM SPEEDS

$$
\begin{aligned}
E = {} & \sum_{b}[{}^{2}K_{b}(b-b_{0})^{2} + {}^{3}K_{b}(b-b_{0})^{3} + {}^{4}K_{b}(b-b_{0})^{4}] \\
& + \sum_{\theta}[{}^{2}K_{\theta}(\theta-\theta_{0})^{2} + {}^{3}K_{\theta}(\theta-\theta_{0})^{3} + {}^{4}K_{\theta}(\theta-\theta_{0})^{4}] \\
& + \sum_{\phi}[{}^{1}K_{\phi}(1-\cos\phi) + {}^{2}K_{\phi}(1-\cos 2\phi) + {}^{3}K_{\phi}(1-\cos 3\phi)] \\
& + \sum_{\chi}K_{\chi}\chi^{2} + \sum_{i>j}\frac{q_{i}q_{j}}{r_{ij}} + \sum_{i>j}\epsilon\left[2\left(\frac{r^{*}}{r_{ij}}\right)^{9} - 3\left(\frac{r^{*}}{r_{ij}}\right)^{6}\right] \\
& + \sum_{b}\sum_{b'}K_{bb'}(b-b_{0})(b'-b'_{0}) + \sum_{\theta}\sum_{\theta'}K_{\theta\theta'}(\theta-\theta_{0}) \times \\
& \quad (\theta'-\theta'_{0}) \\
& + \sum_{b}\sum_{\theta}K_{b\theta}(b-b_{0})(\theta-\theta_{0}) \\
& + \sum_{\phi}\sum_{b}(b-b_{0})[{}^{1}K_{\phi b}\cos\phi + {}^{2}K_{\phi b}\cos 2\phi + {}^{3}K_{\phi b}\cos 3\phi] \\
& + \sum_{\phi}\sum_{b'}(b'-b'_{0})[{}^{1}K_{\phi b'}\cos\phi + {}^{2}K_{\phi b'}\cos 2\phi + \\
& \quad {}^{3}K_{\phi b'}\cos 3\phi] \\
& + \sum_{\phi}\sum_{\theta}(\theta-\theta_{0})[{}^{1}K_{\phi\theta}\cos\phi + {}^{2}K_{\phi\theta}\cos 2\phi + {}^{3}K_{\phi\theta}\cos 3\phi] \\
& + \sum_{\phi}\sum_{\theta}\sum_{\theta'}K_{\phi\theta\theta'}(\theta-\theta_{0})(\theta'-\theta'_{0})\cos\phi
\end{aligned}
\tag{1}
$$

bond-bond: angle node

angle-angle: torsion node

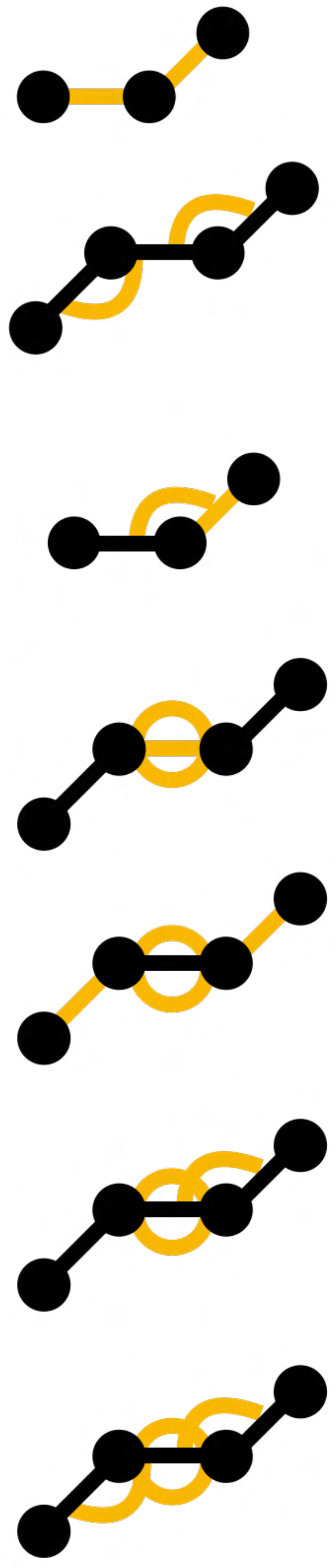
bond-angle: angle node

torsion-(center) bond: torsion

torsion-(side) bond: torsion

torsion-angle: torsion

torsion-angle-angle: torsion



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

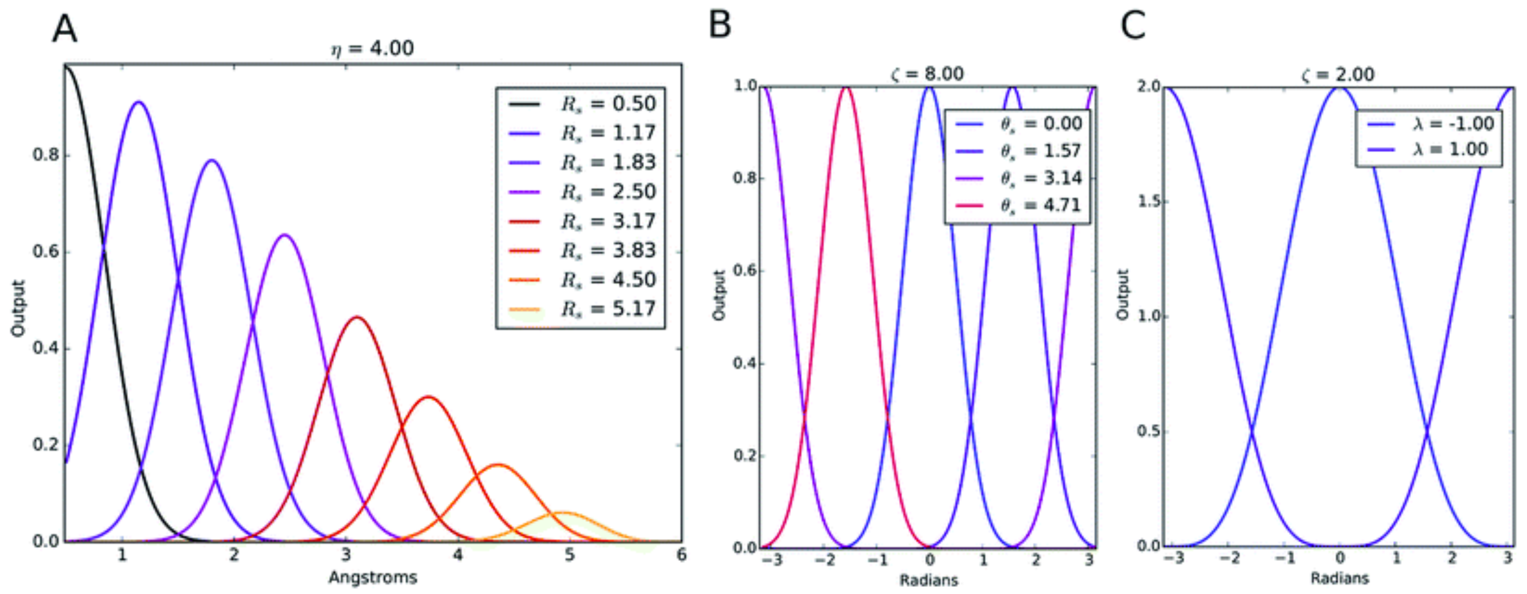
ANI family of quantum machine learning (QML) potentials

radial and **angular** features

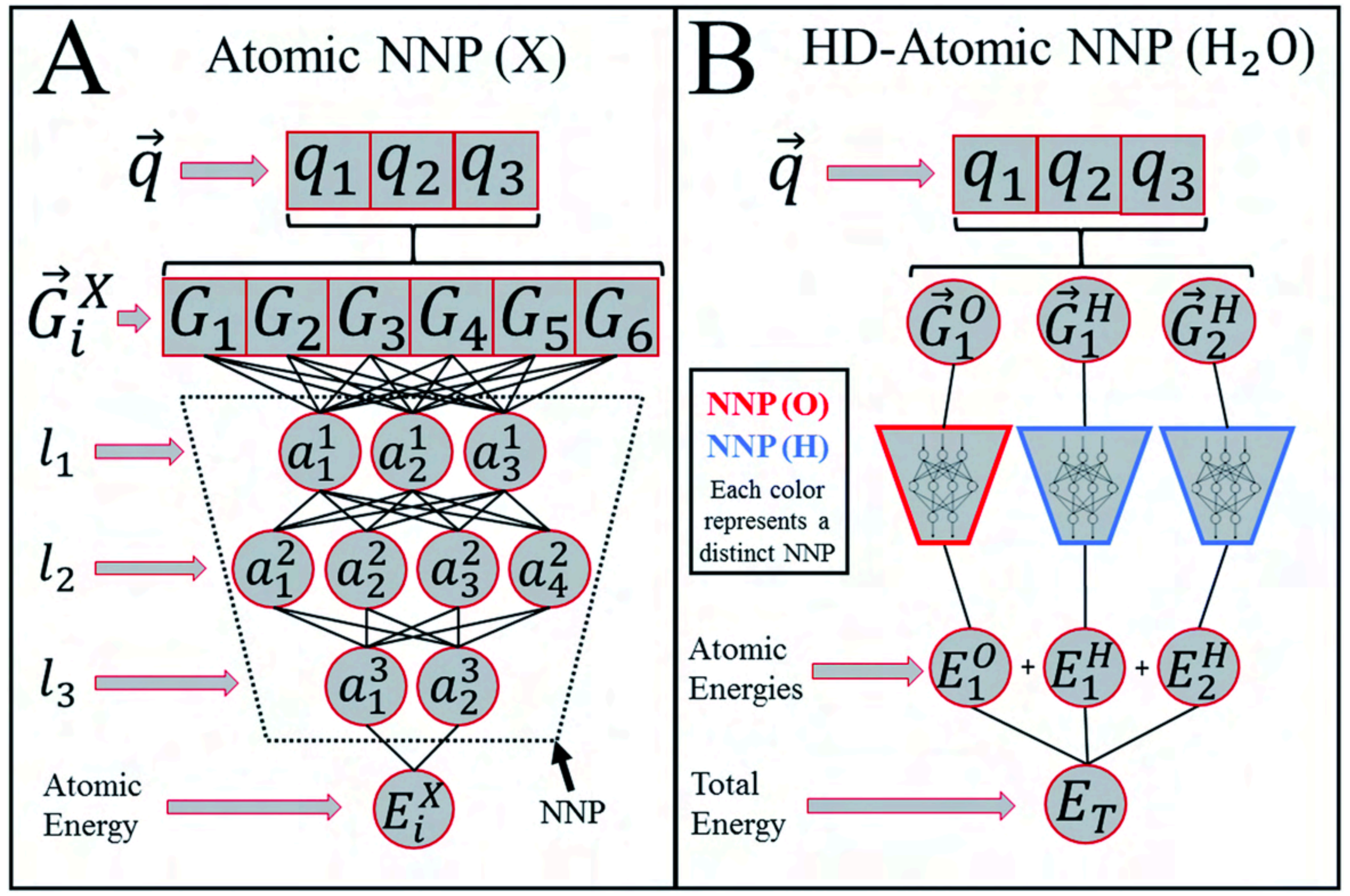
$$f_C(R_{ij}) = \begin{cases} 0.5 \times \cos\left(\frac{\pi R_{ij}}{R_C}\right) + 0.5 & \text{for } R_{ij} \leq R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases}$$

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

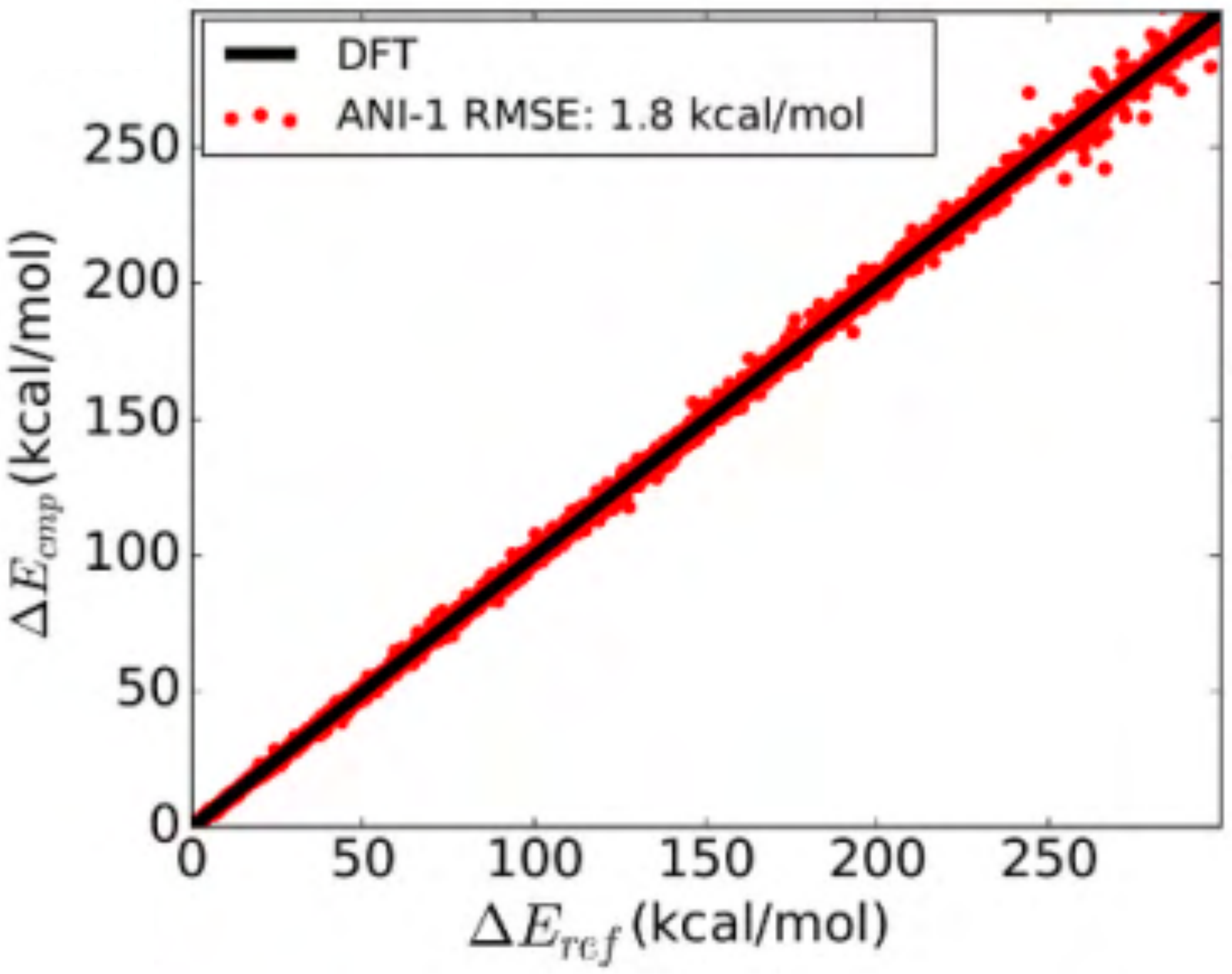
$$G_m^{A_{\text{mod}}} = 2^{1-\zeta} \sum_{j,k \neq i}^{\text{all atoms}} \left(1+\cos(\theta_{ijk}-\theta_s)\right)^{\zeta} \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2} - R_s\right)^2\right] f_C(R_{ij}) f_C(R_{ik})$$



deep neural network for each atom



excellent agreement with DFT

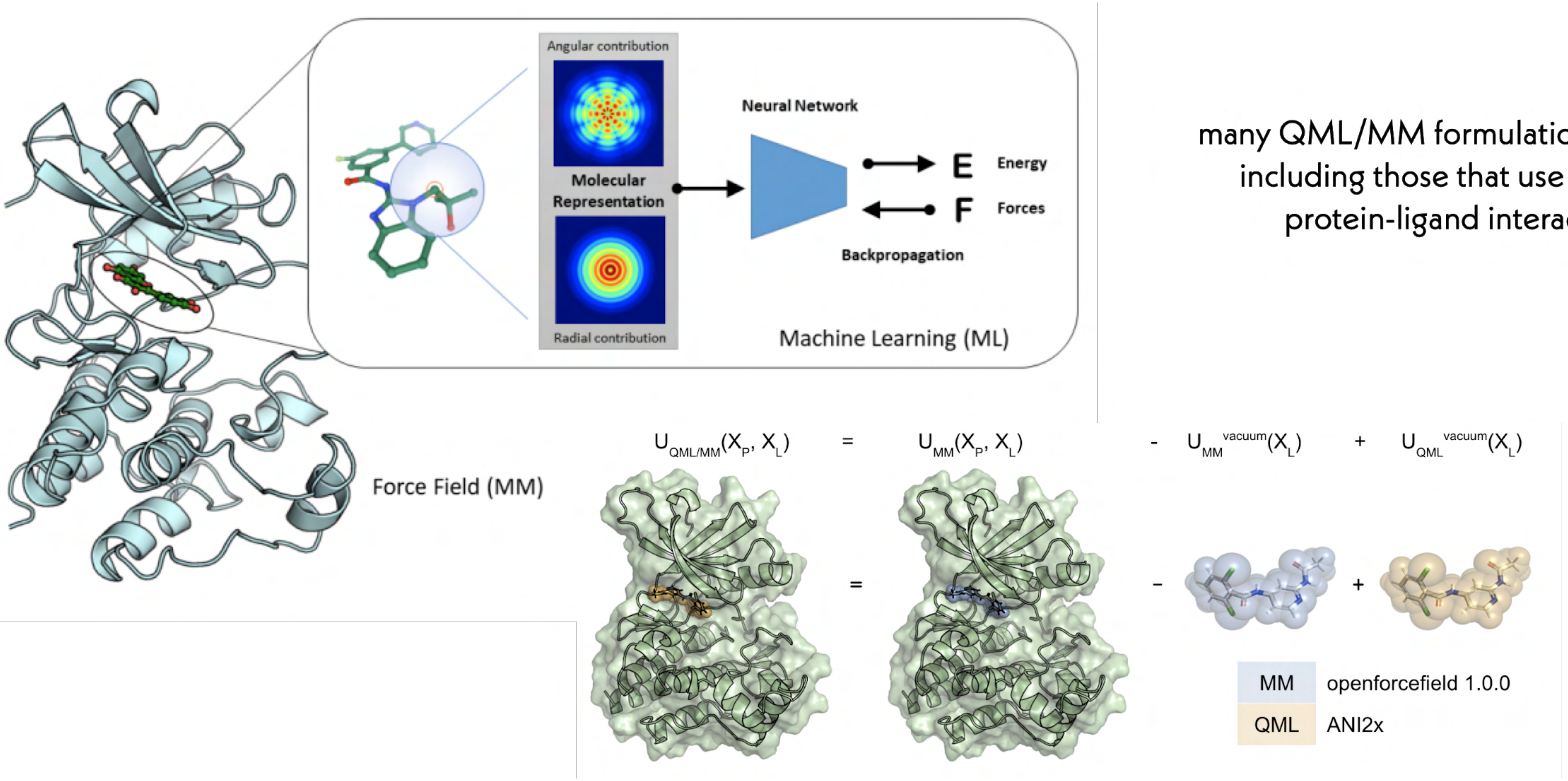


OLEXANDR ISAYEV

ADRIAN ROITBERG

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

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



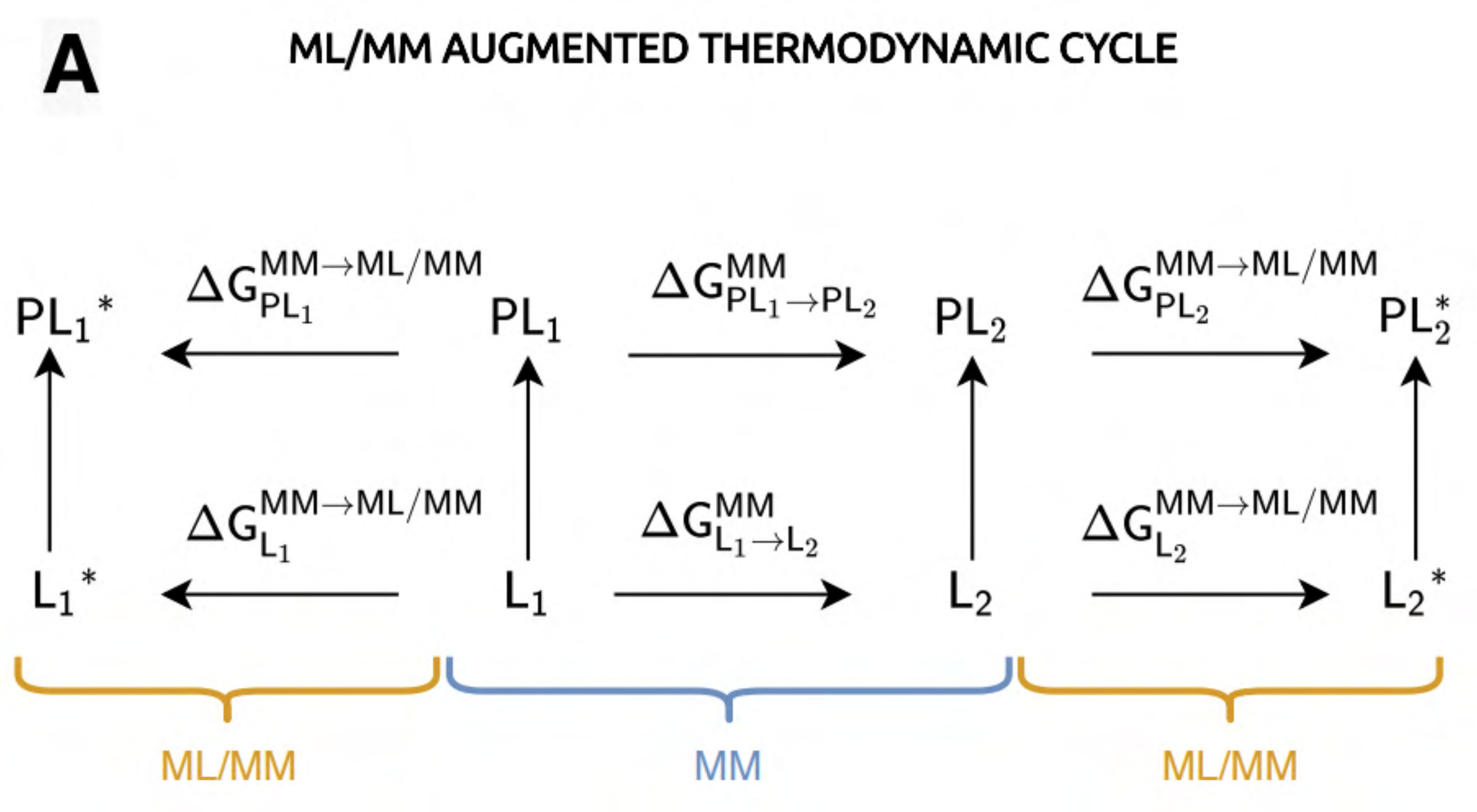
many QML/MM formulations possible, including those that use QML for protein-ligand interactions

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



A

ML/MM AUGMENTED THERMODYNAMIC CYCLE

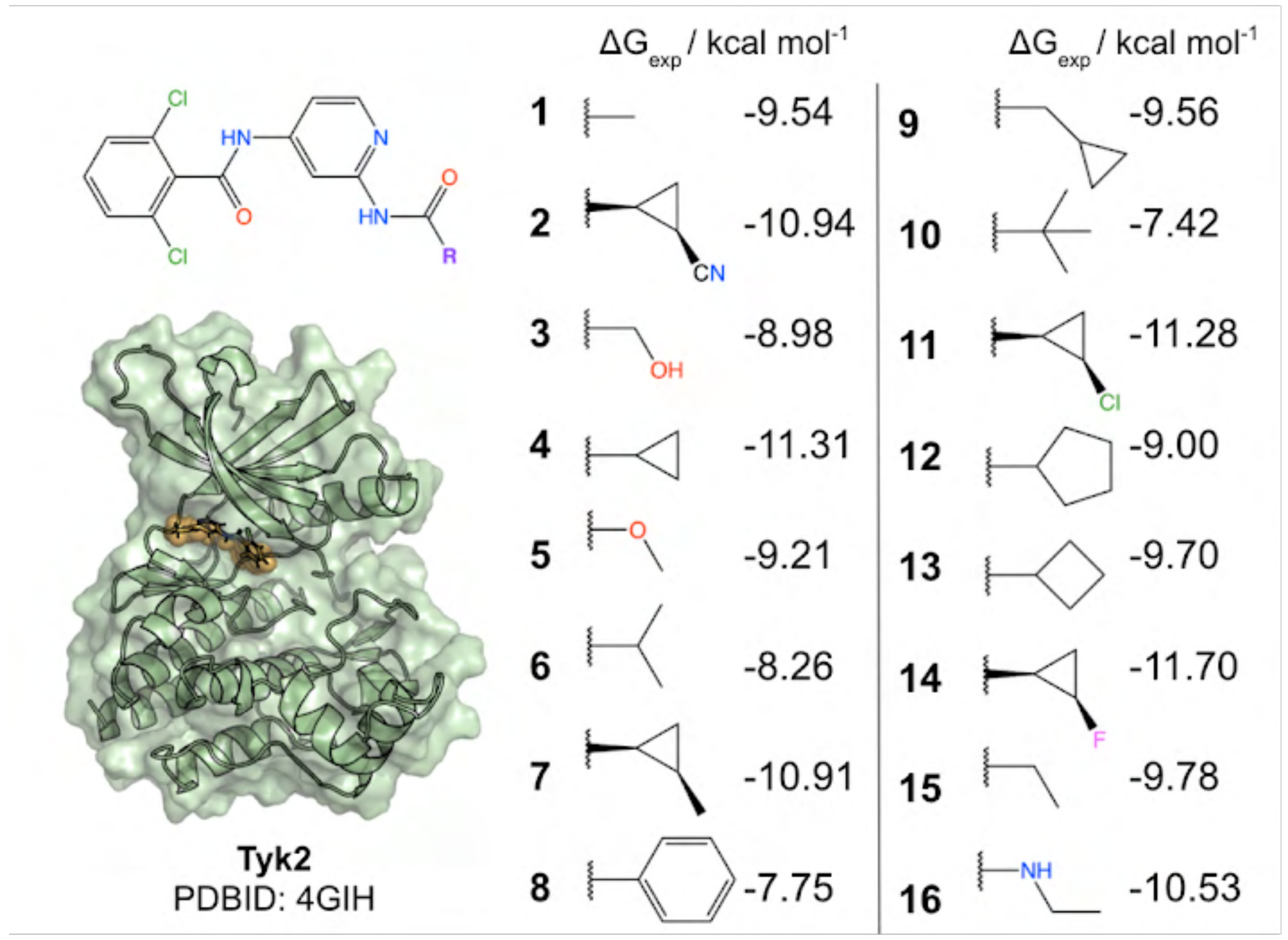
$$\begin{tikzcd}
PL_1{}^{*} & PL_1 \arrow[l, "\Delta G^{MM\to ML/MM}_{PL_1}"'] \arrow[r, "\Delta G^{MM}_{PL_1\to PL_2}"] & PL_2 \arrow[r, "\Delta G^{MM\to ML/MM}_{PL_2}"] & PL_2^{*} \\
L_1{}^{*} \arrow[u] & L_1 \arrow[u] \arrow[l, "\Delta G^{MM\to ML/MM}_{L_1}"'] \arrow[r, "\Delta G^{MM}_{L_1\to L_2}"]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

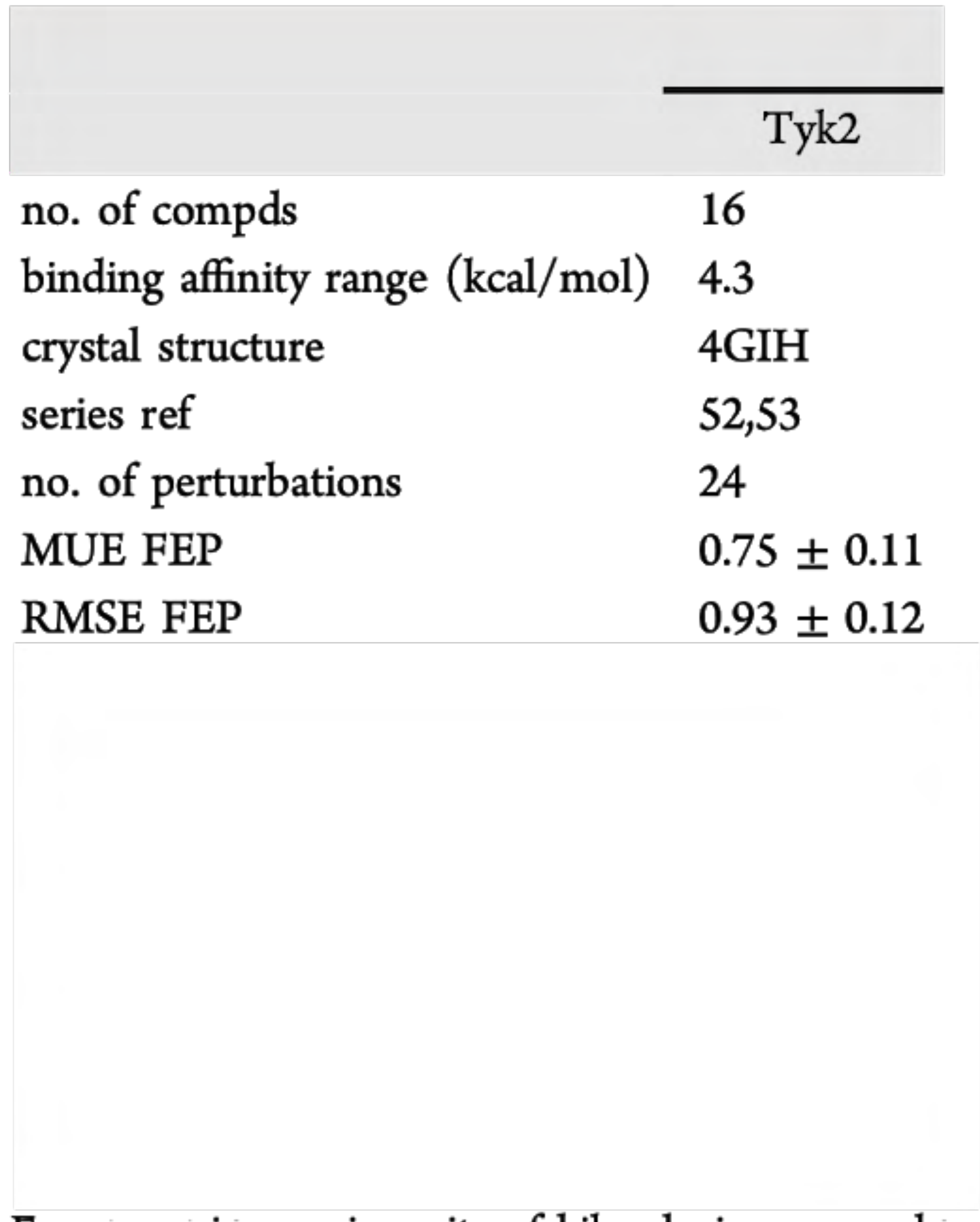
HYBRID QUANTUM MACHINE LEARNING / MOLECULAR MECHANICS (QML/MM) FREE ENERGY CALCULATIONS CUT ERROR IN HALF

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

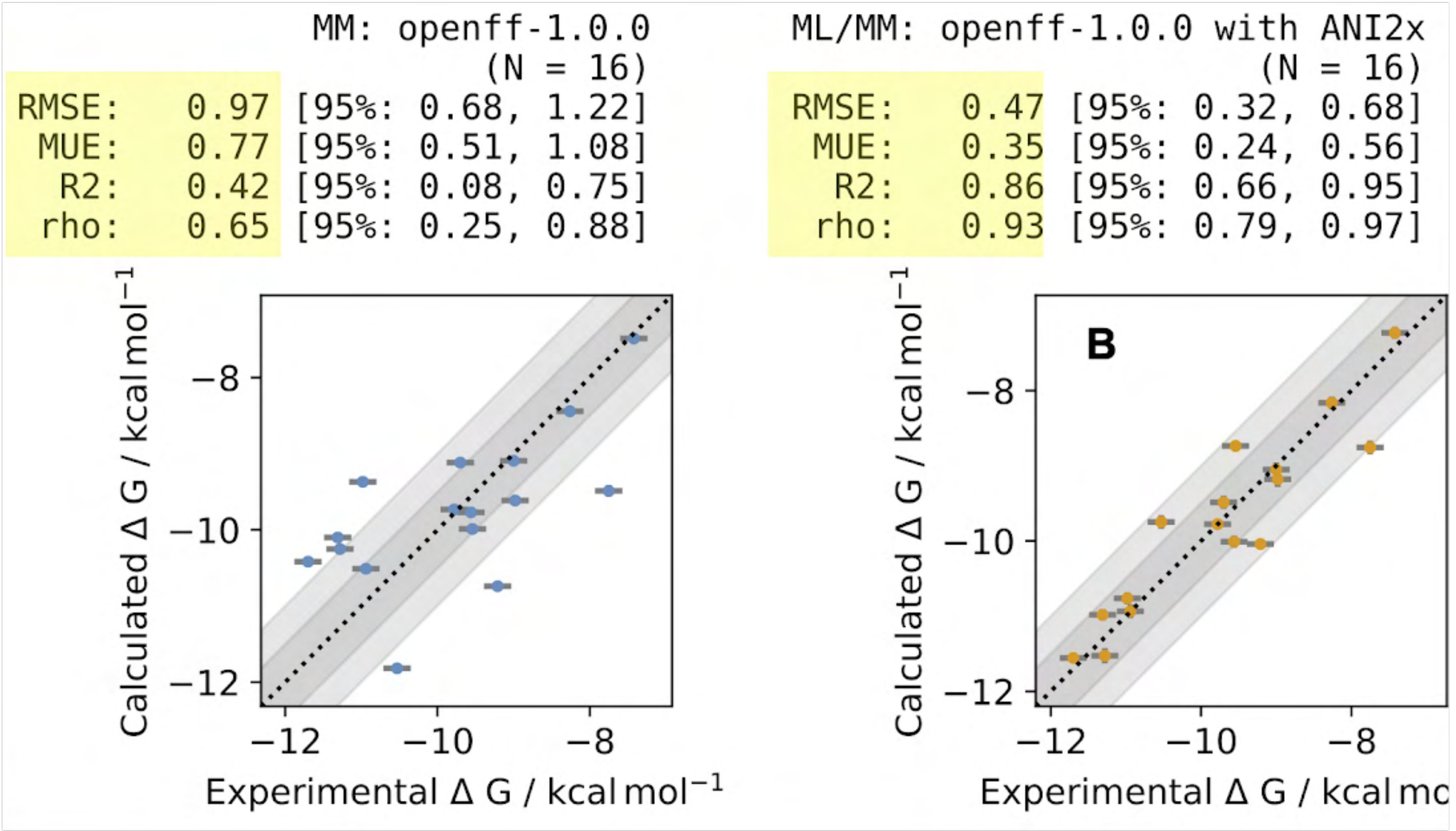




	Tyk2
no. of compds	16
binding affinity range (kcal/mol)	4.3
crystal structure	4GIH
series ref	52,53
no. of perturbations	24
MUE FEP	0.75 ± 0.11
RMSE FEP	0.93 ± 0.12

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)

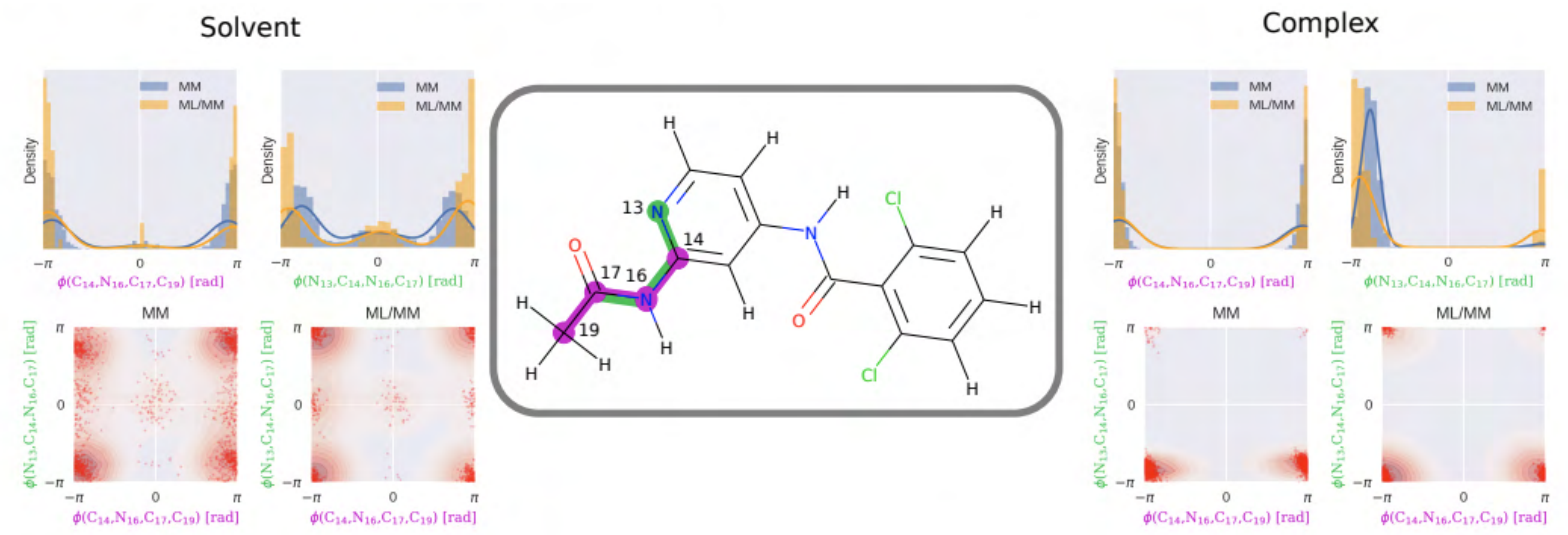


```
MM: openff-1.0.0
             (N = 16)
RMSE:   0.97 [95%: 0.68, 1.22]
 MUE:   0.77 [95%: 0.51, 1.08]
  R2:   0.42 [95%: 0.08, 0.75]
 rho:   0.65 [95%: 0.25, 0.88]
```

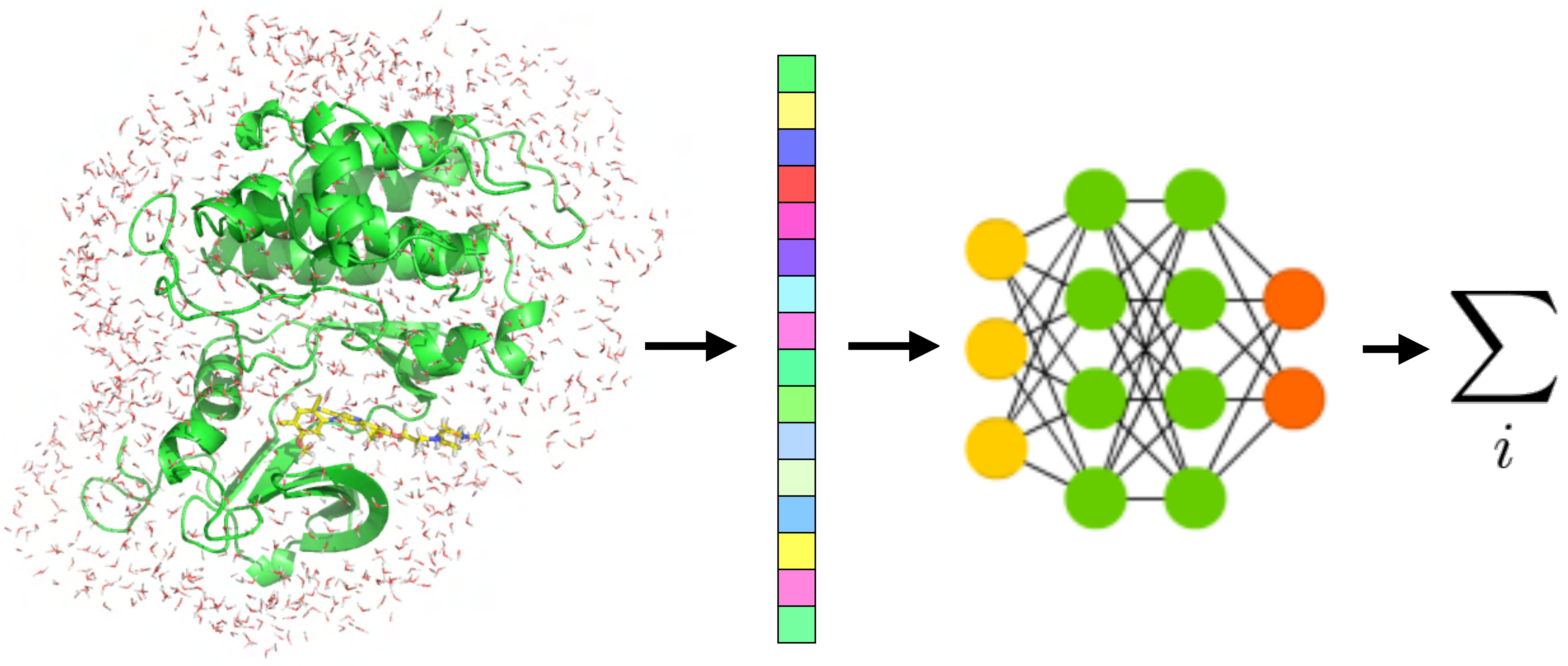
```
ML/MM: openff-1.0.0 with ANI2x
             (N = 16)
RMSE:   0.47 [95%: 0.32, 0.68]
 MUE:   0.35 [95%: 0.24, 0.56]
  R2:   0.86 [95%: 0.66, 0.95]
 rho:   0.93 [95%: 0.79, 0.97]
```

replica-exchange free energy calculations with perses
preprint: https://doi.org/10.1101/2020.07.29.227959
code: https://github.com/choderalab/perses
https://github.com/choderalab/qmlify

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



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



atomic coordinates

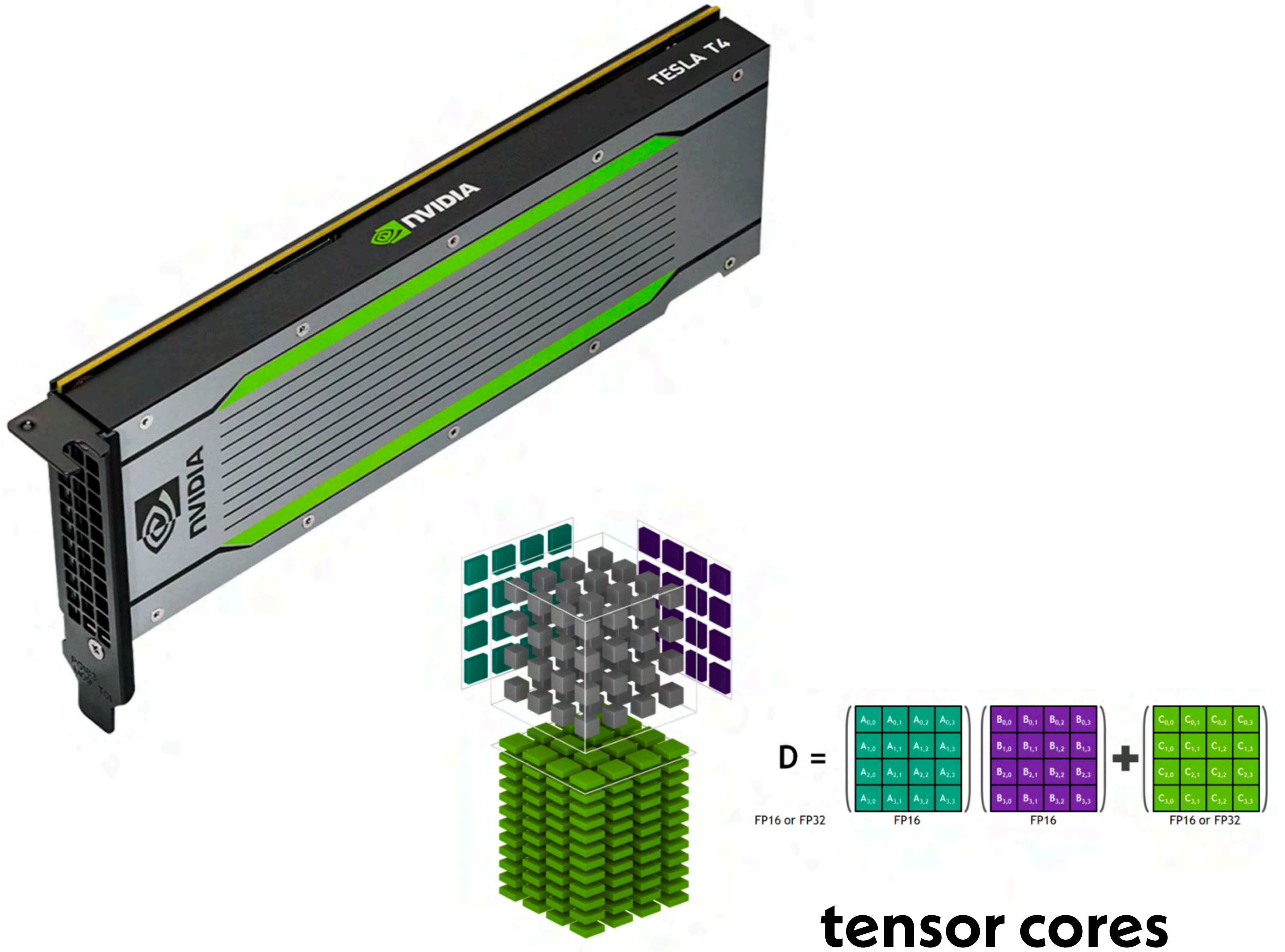
feature computation

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



we need FP32!
(maybe INT32 too fixed-point)

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

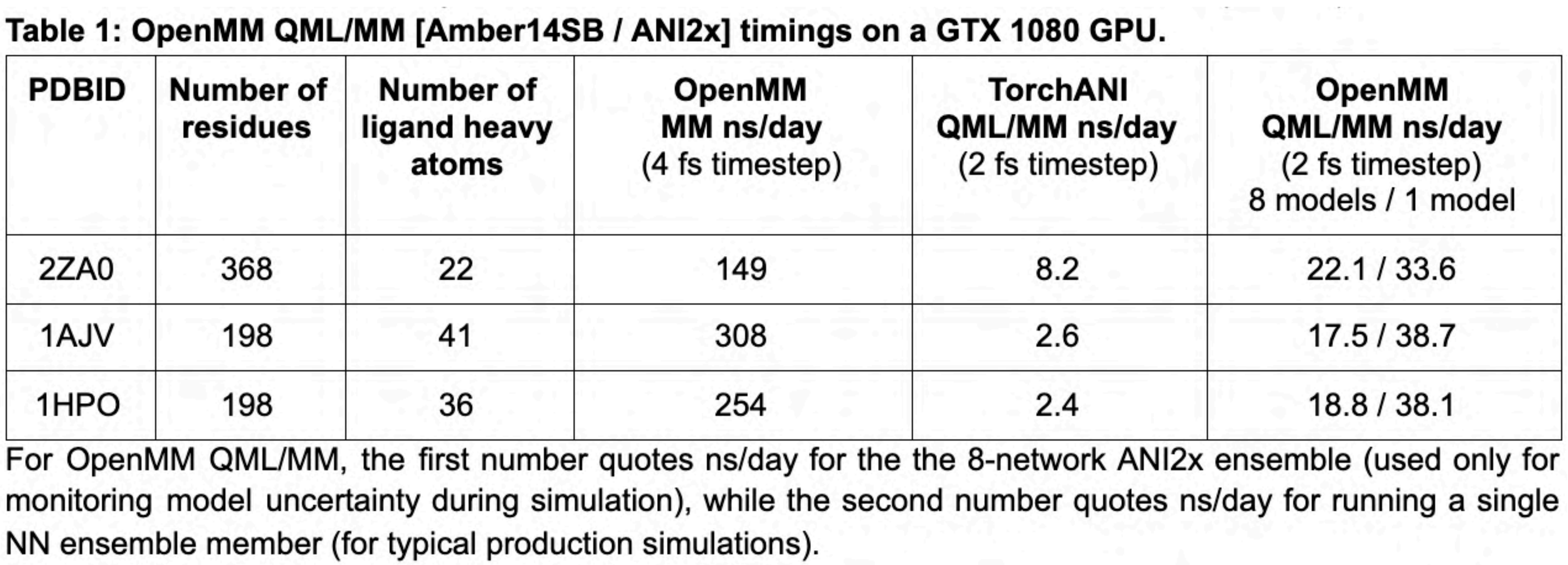


Table 1: OpenMM QML/MM [Amber14SB / ANI2x] timings on a GTX 1080 GPU.

PDBID	Number of residues	Number of ligand heavy atoms	OpenMM MM ns/day (4 fs timestep)	TorchANI QML/MM ns/day (2 fs timestep)	OpenMM QML/MM ns/day (2 fs timestep) 8 models / 1 model
2ZA0	368	22	149	8.2	22.1 / 33.6
1AJV	198	41	308	2.6	17.5 / 38.7
1HPO	198	36	254	2.4	18.8 / 38.1

For OpenMM QML/MM, the first number quotes ns/day for the the 8-network ANI2x ensemble (used only for monitoring model uncertainty during simulation), while the second number quotes ns/day for running a single NN ensemble member (for typical production simulations).

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

(~5x slower than MD right now)

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

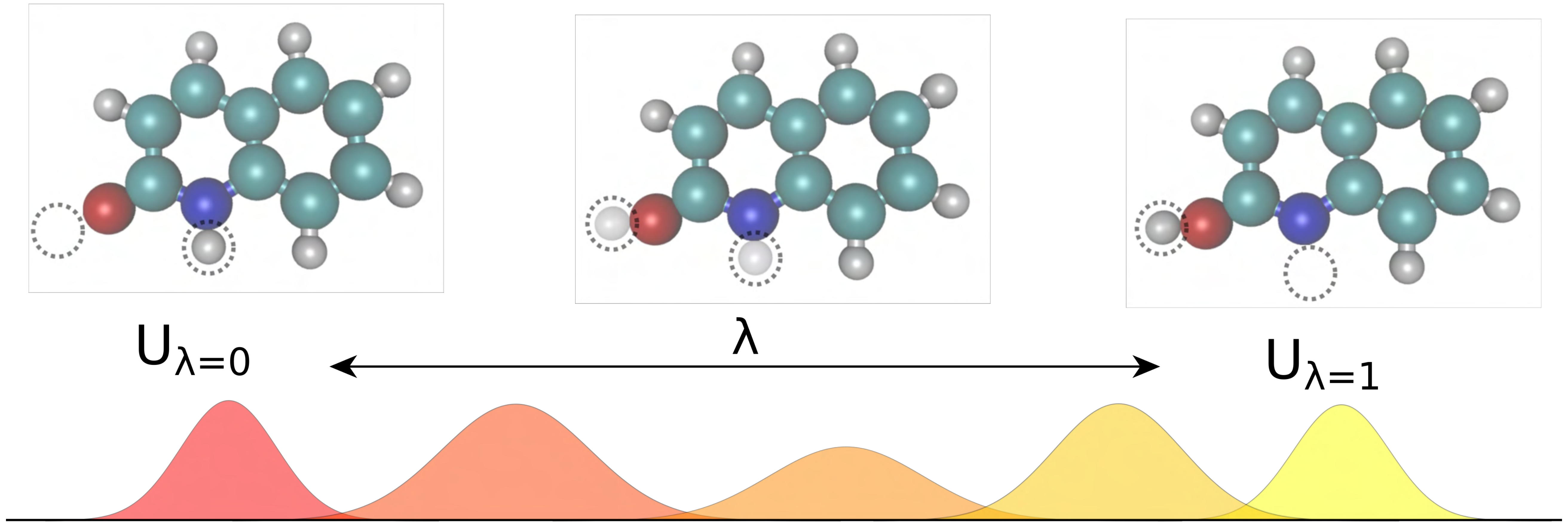
Peter Eastman, Raimondas Galvelis, Gianni de Fabritiis

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

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

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

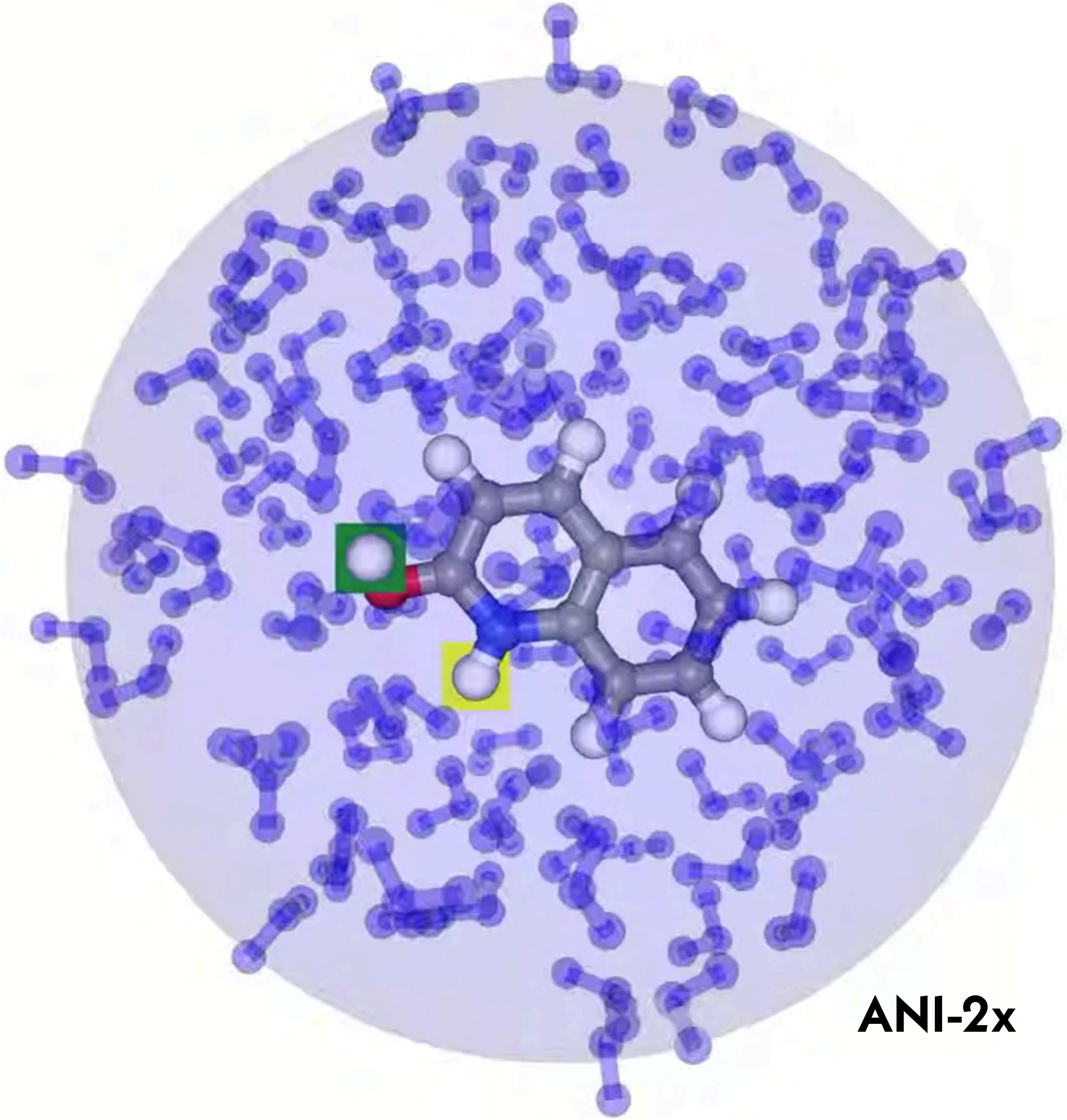


Simple atomic restraints can be used to improve efficiency by preventing atoms from flying away

JOSH FASS

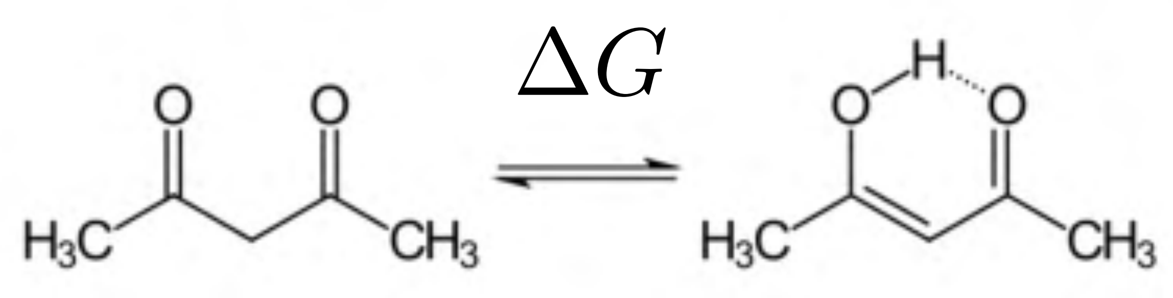
MARCUS WIEDER

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

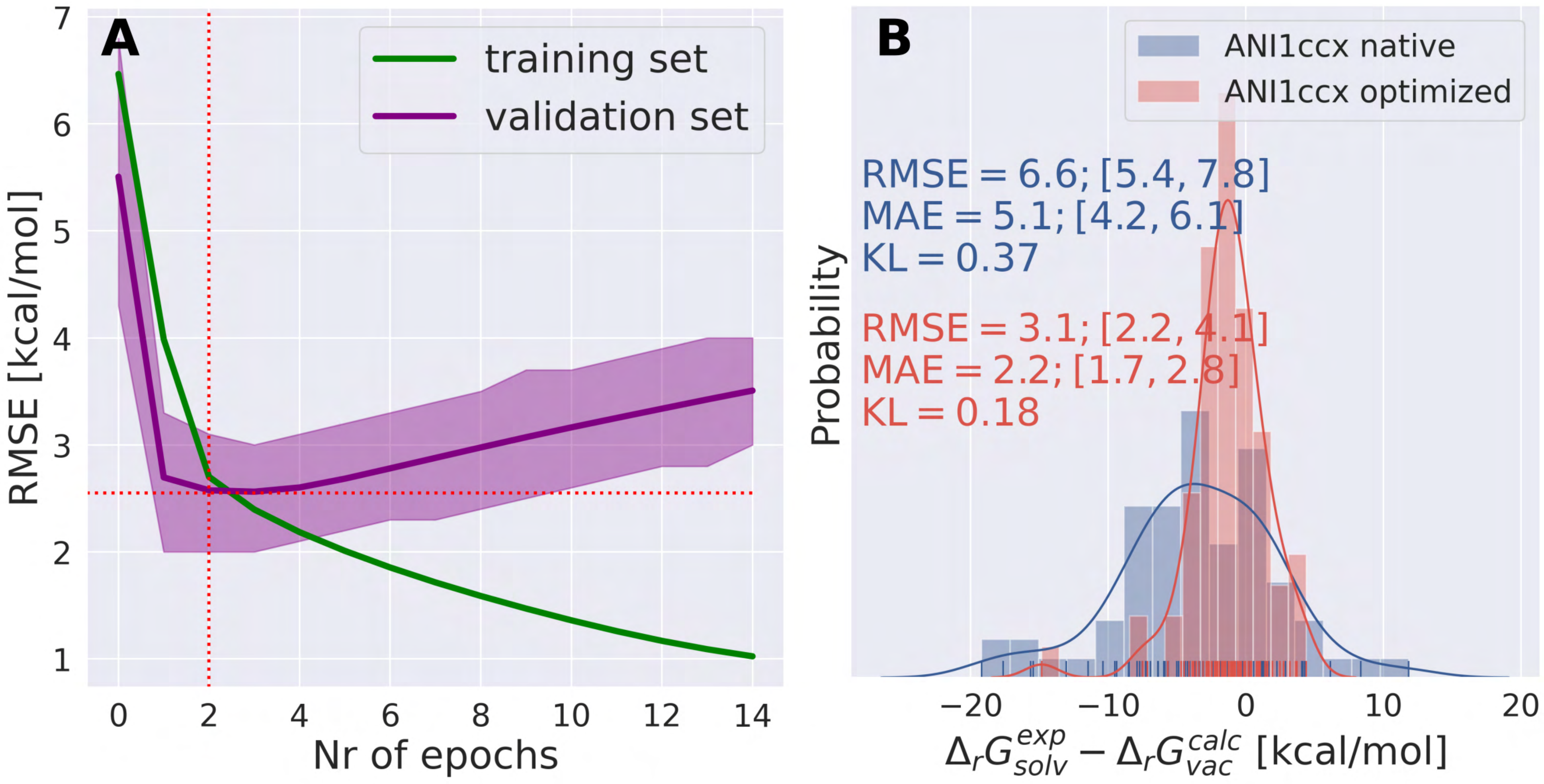


QML POTENTIALS CAN LEARN FROM EXPERIMENTAL DATA TO IMPROVE PHYSICAL MODELS

physical models are data-efficient: retraining on small number of experimental measurements improves accuracy and generalizes well



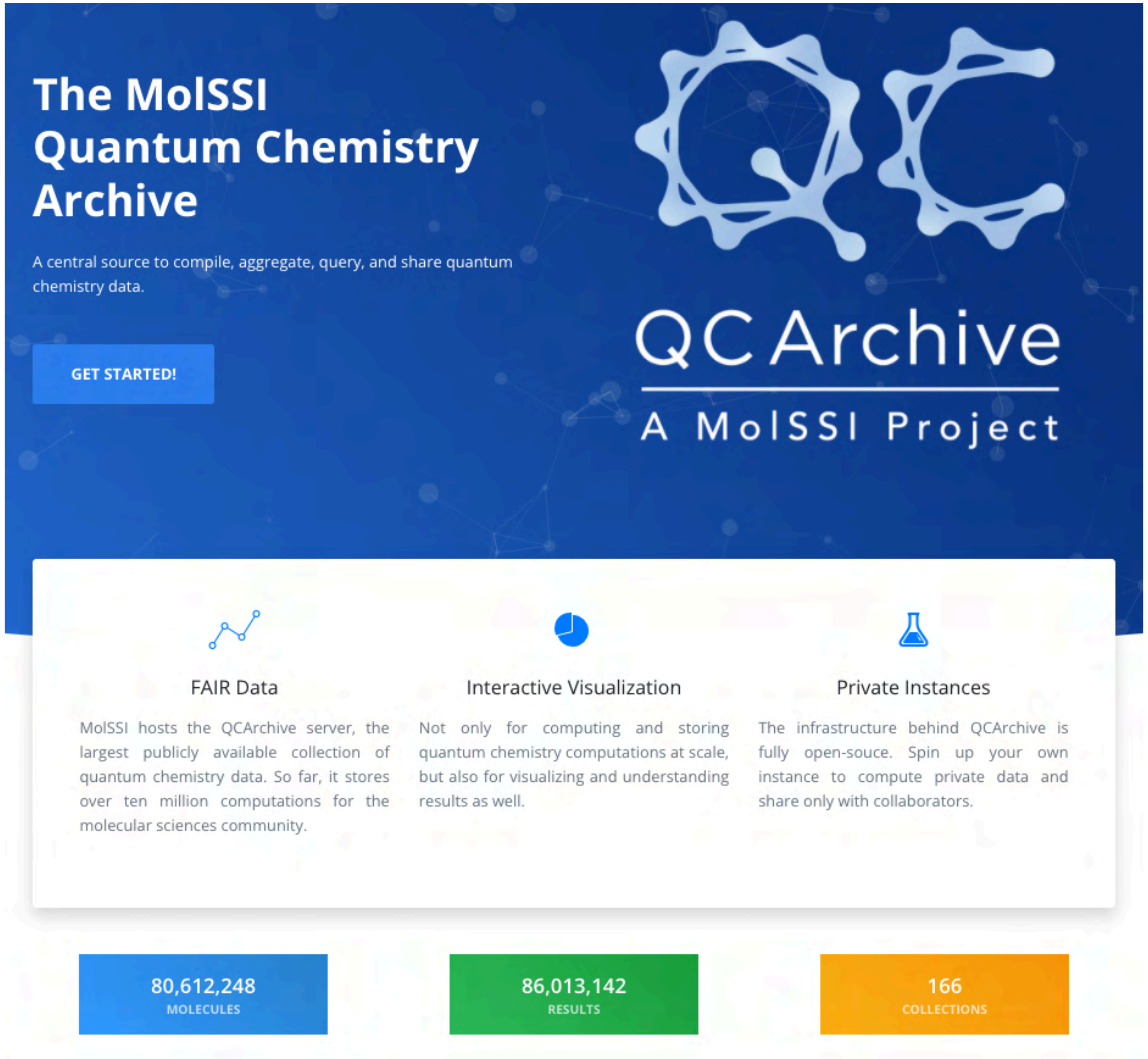
train: 221 tautomer pairs
validate: 57 tautomer pairs
test: 72 tautomer pairs



JOSH FASS

MARCUS WIEDER

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



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.

80,612,248 MOLECULES	86,013,142 RESULTS	166 COLLECTIONS

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

http://qcarchive.molssi.org

INTEGRATING MACHINE LEARNING WILL COMPLETELY CHANGE PRACTICE IN STRUCTURE-ENABLED DRUG DISCOVERY

2021

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

CHODERA LAB

NIH National Institutes of Health

NSF

STIFTUNG CHARITÉ

SCHRÖDINGER.

SILICON Therapeutics

PARKER INSTITUTE for CANCER IMMUNOTHERAPY

Gerstner FAMILY FOUNDATION

BAYER

STARR CANCER CONSORTIUM

OpenEye SCIENTIFIC

open forcefield consortium

XtalPi

VIR

CYCLE FOR SURVIVAL

Scientific Advisor: OpenEye, Foresite Labs
All funding: http://choderalab.org/funding