LEARNING BIOMOLECULAR POTENTIALS FOR DRUG DISCOVERY

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

DISCLOSURES:
Scientific Advisor, OpenEye Scientific and Foresite Labs
All opinions/views are my own.

NeurIPS - recorded 25 Nov 2020

WE'RE BUILDING TOOLS TO ENABLE AUTONOMOUS MOLECULAR DESIGN

WE'RE BUILDING TOOLS TO ENABLE AUTONOMOUS MOLECULAR DESIGN

Ed Griffen
Medchemica

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

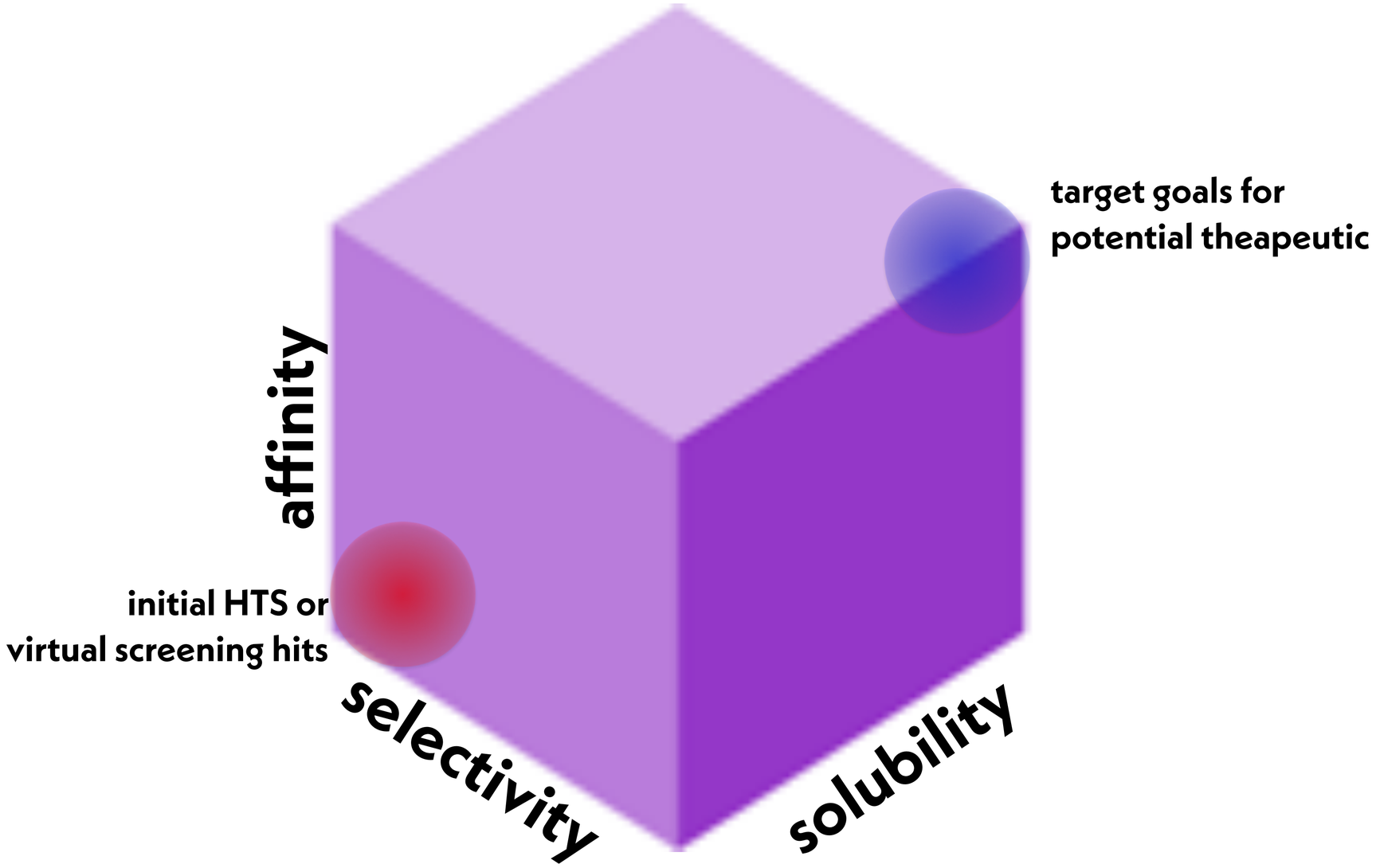
Property	Target range	Rationale
protease assay	IC_{50} < 10 nM	Extrapolation from other anti-viral programs
viral replication assay	EC_{50} < 5 μM	Suppression of virus at achievable blood levels
plaque reduction assay	EC_{50} < 5 μM	Suppression of virus at achievable blood levels
route of administration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

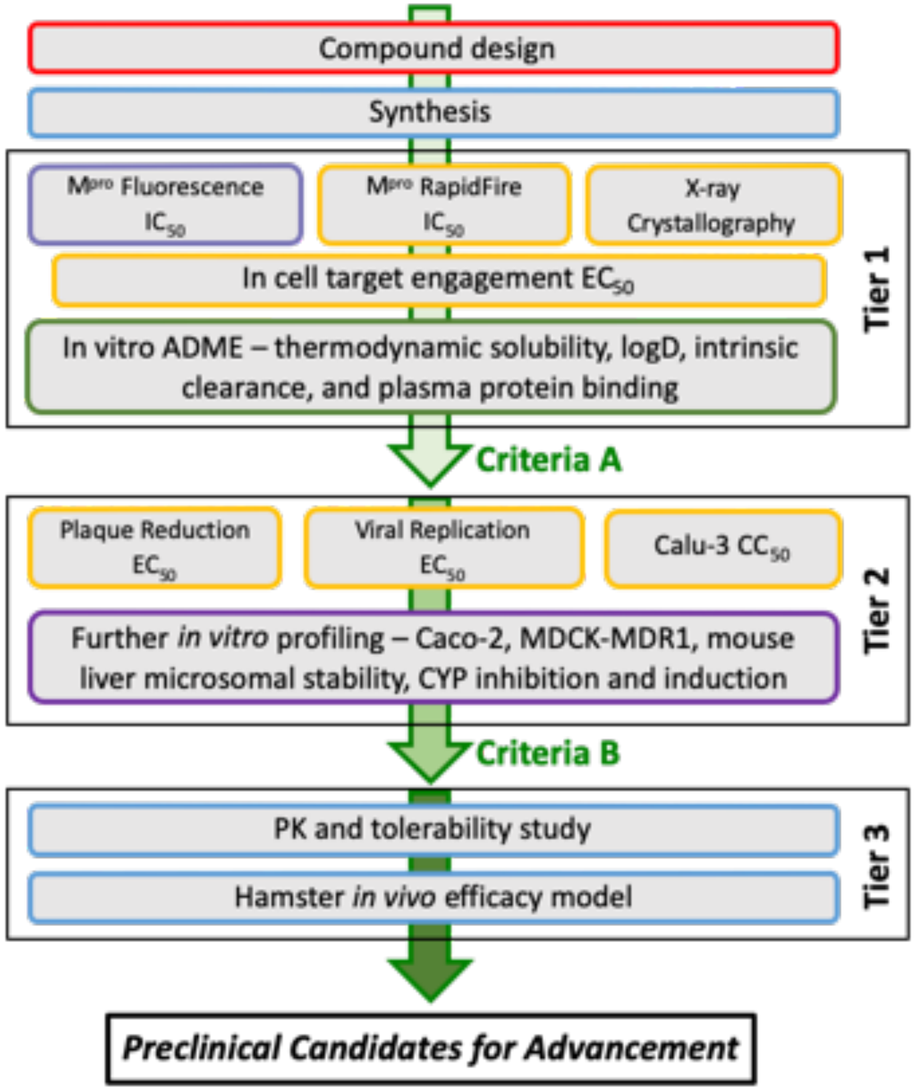
An international effort to
DISCOVER A COVID ANTIVIRAL

https://covid.postera.ai/covid

DRUG DISCOVERY INVOLVES COMPLEX DESIGN OBJECTIVES



TO GET THERE, DRUG DESIGN INVOLVES MAKING A LOT OF DECISIONS ABOUT WHICH MOLECULES TO MAKE AND ASSAYS TO RUN



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?

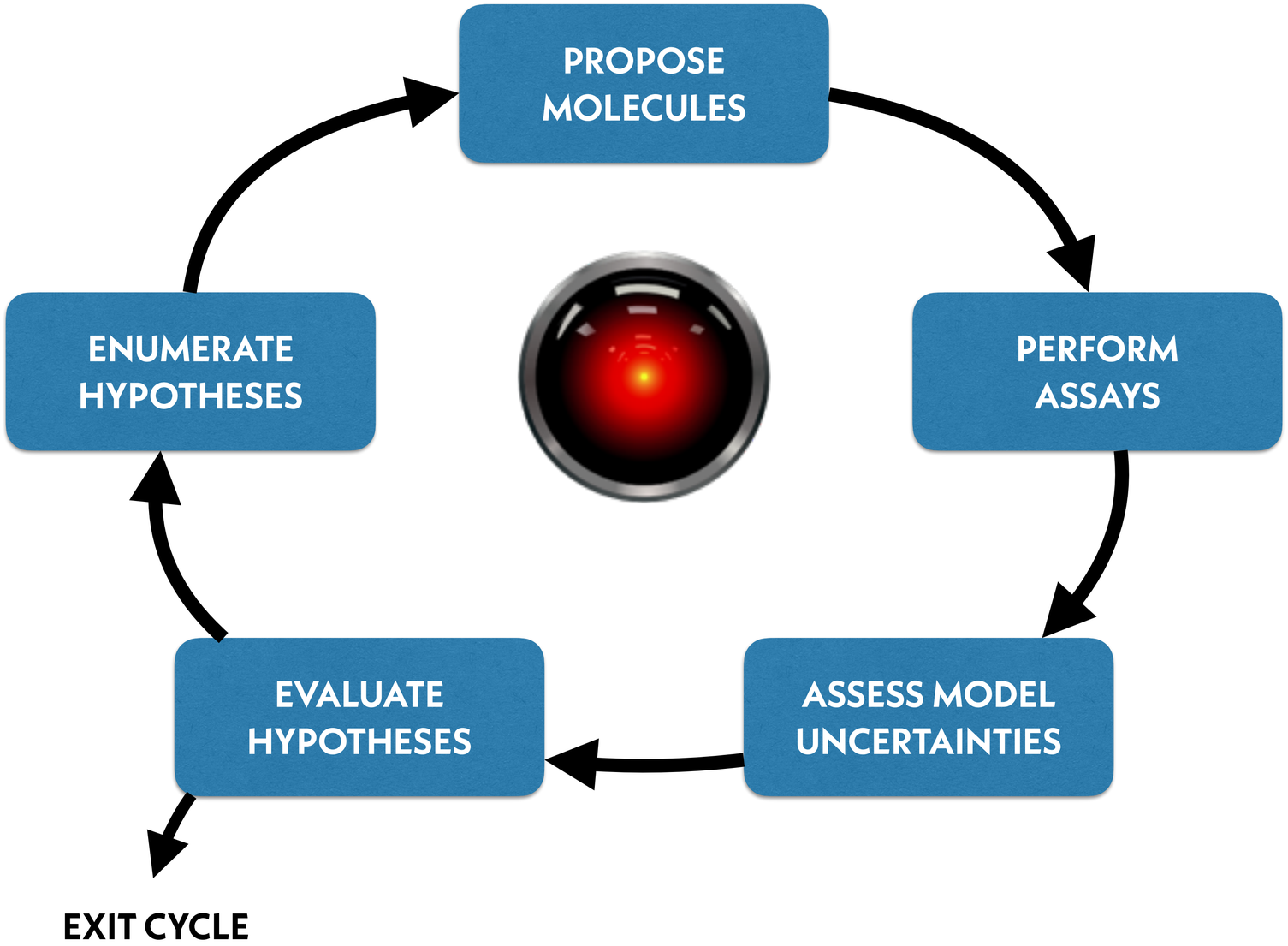
Does it kill the virus in cells?

Could it cause bad side effects?

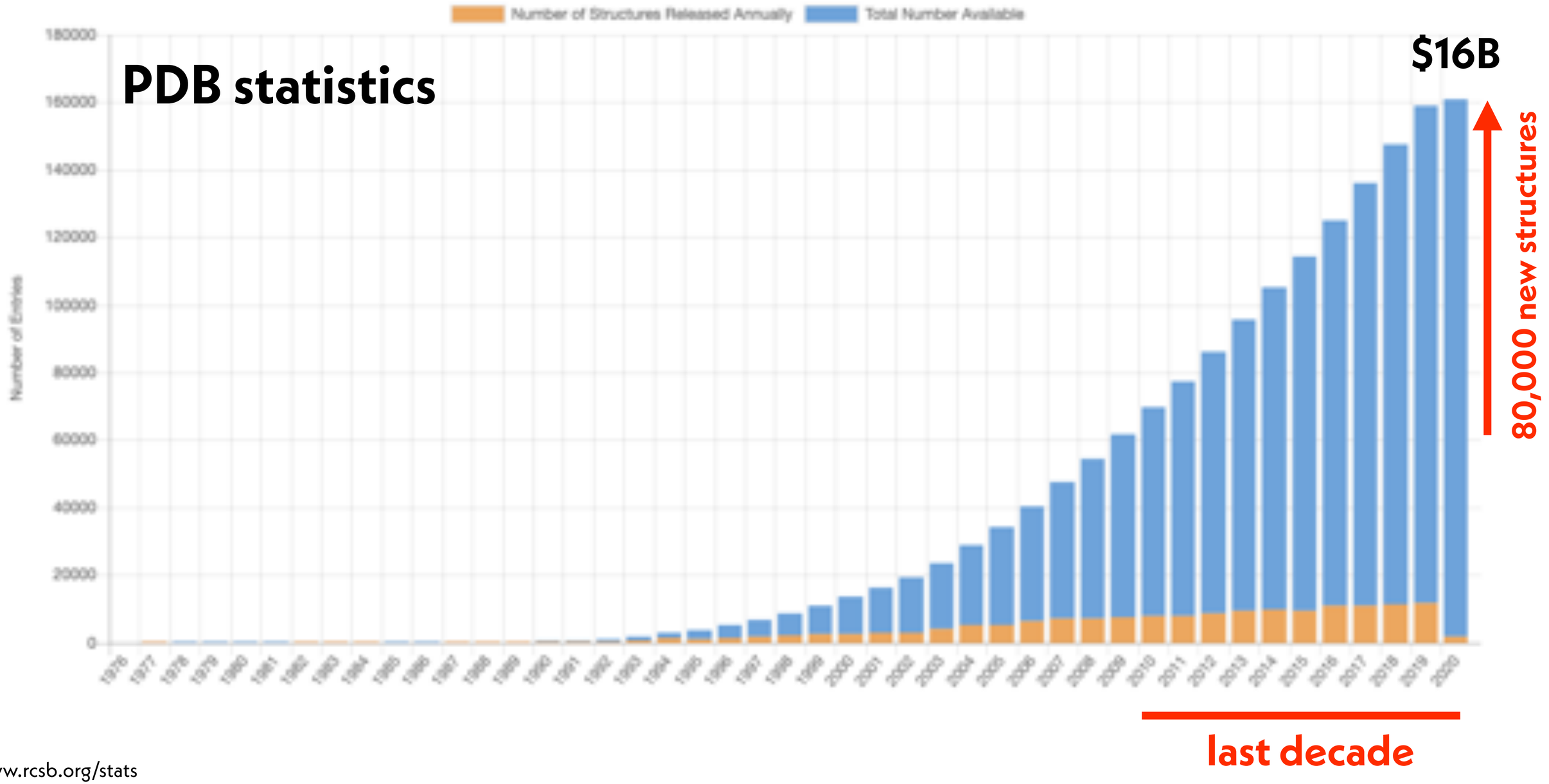
Can oral dosing deliver sufficient drug?

Does it actually work against the disease?

AUTONOMOUS REASONING ENGINES REQUIRE MODELS THAT CAN LEARN

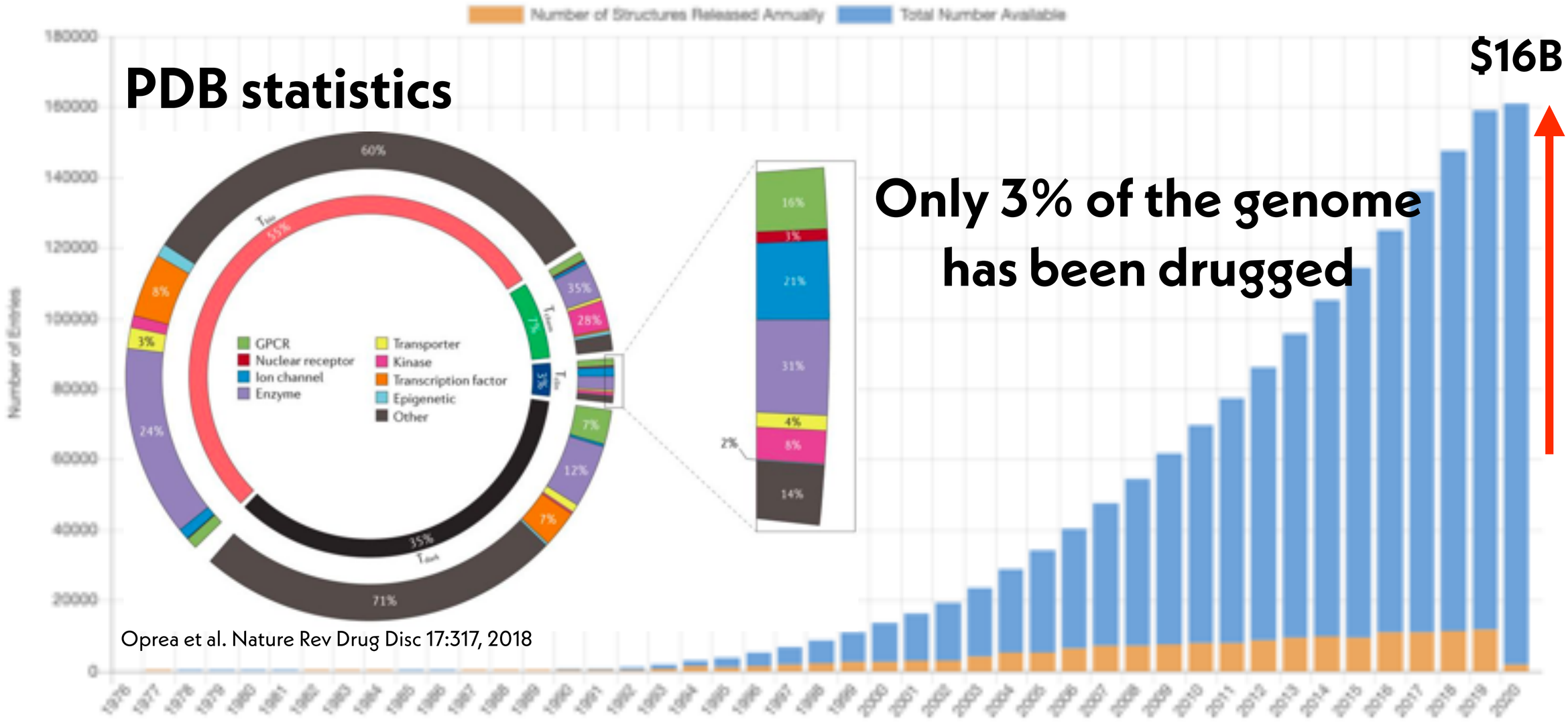


THE LAST DECADE HAS PRODUCED AN ENORMOUS NUMBER OF BIOMOLECULAR STRUCTURES



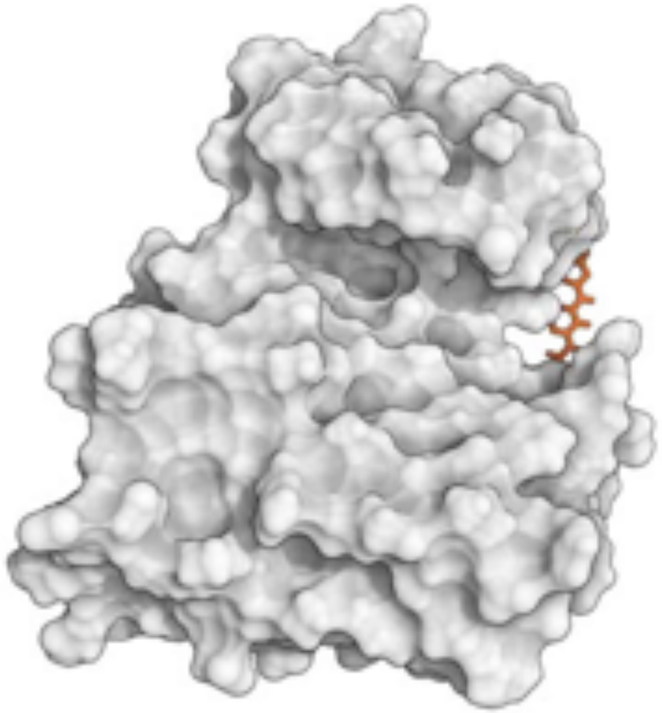
http://www.rcsb.org/stats

THE LAST DECADE HAS PRODUCED AN ENORMOUS NUMBER OF BIOMOLECULAR STRUCTURES

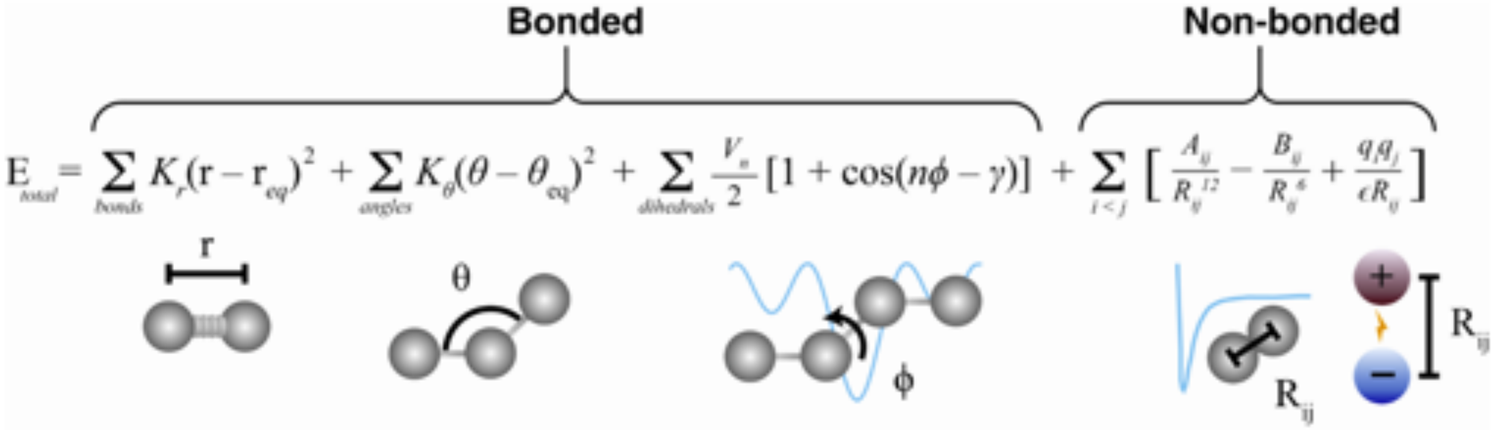


http://www.rcsb.org/stats

BIOMOLECULAR SIMULATIONS CAN PREDICT USEFUL PROPERTIES LIKE BINDING AFFINITIES, BUT THEY CAN'T LEARN FROM DATA



typical molecular mechanics force field

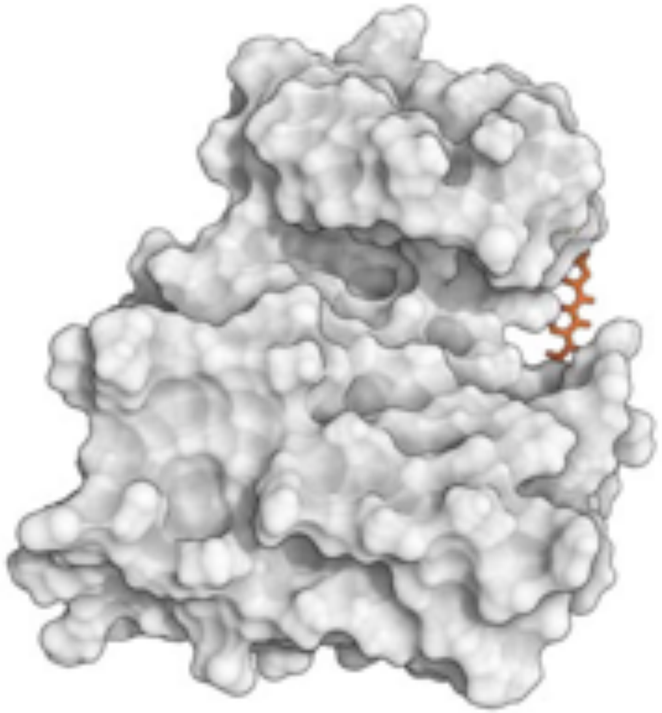


...OR CAN THEY?

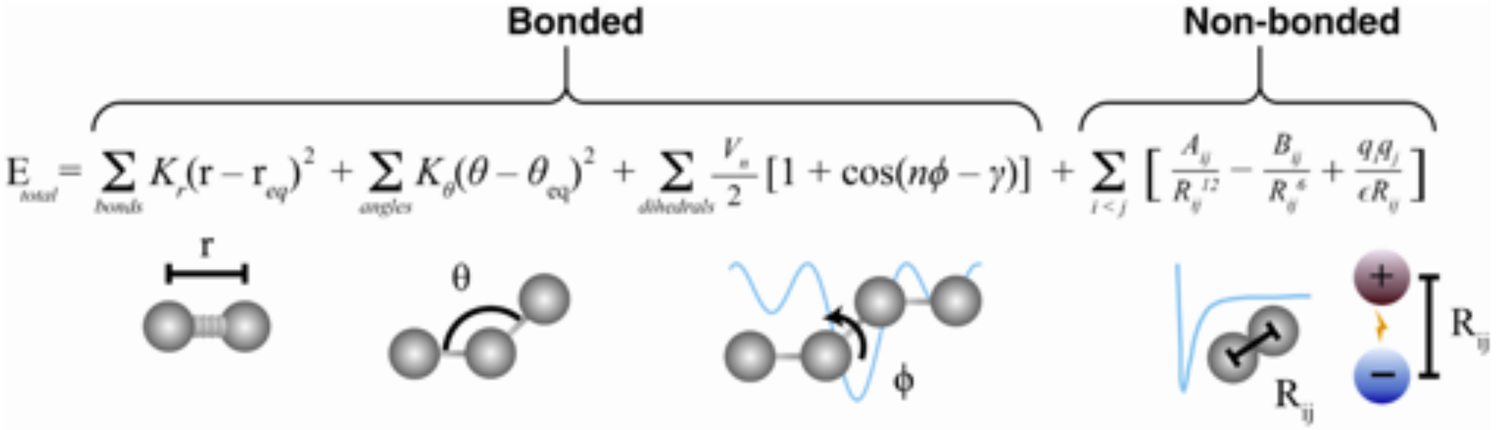
Shan, Kim, Eastwood, Dror, Seeliger, Shaw. JACS 133:9181, 2011

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

BIOMOLECULAR SIMULATIONS CAN PREDICT USEFUL PROPERTIES LIKE BINDING AFFINITIES, BUT THEY CAN'T LEARN FROM DATA



typical molecular mechanics force field



...OR CAN THEY?

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

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

HOW ARE FORCEFIELDS MADE?

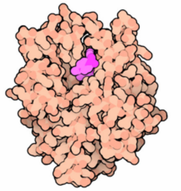
HOW ARE FORCEFIELDS MADE?

experimental data
quantum chemistry
keen chemical intuition

heroic effort by graduate students and postdocs

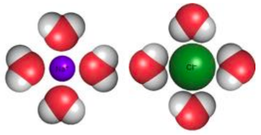
a parameter set we desperately hope someone actually uses

FORCE FIELD CONSTRUCTION TRADITIONALLY REQUIRES HEROIC EFFORT



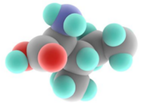
proteins

post-translational modifications

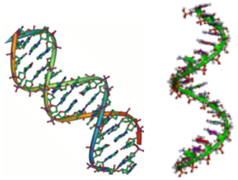


water

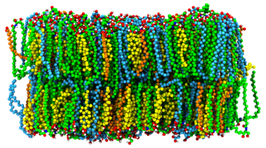
ions



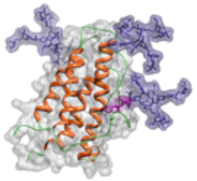
small molecules



nucleic acids



lipids



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 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.

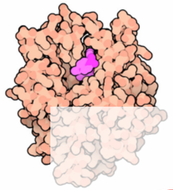
M. Zgarbova; M. Otyepka; J. Sponer; A. Mladek; P. Banas; T. E. Cheatham; P. Jurecka. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.*, **2011**, *7*, 2886–2902.

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.

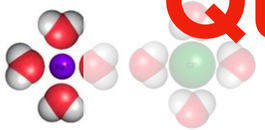
K. N. Kirschner; A. B. Yongye; S. M. Tschampel; J. González-Outeiriño; C. R. Daniels; B. L. Foley; R. J. Woods. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.*, **2008**, *29*, 622–655.

FORCE FIELD CONSTRUCTION TRADITIONALLY REQUIRES HEROIC EFFORT



proteins

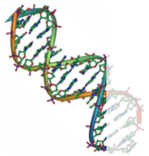
post-translational modifications



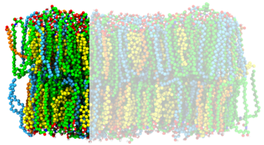
water

ions

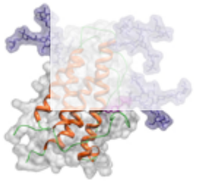
non-standard residues



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; [illegible] Kollman. A second generation force field for the simulation of proteins, nucleic [illegible] *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 [illegible] specific ion parameters. *J. Phys. Chem. B*, 2009, *113*, 13279–13290.

P. Li; B. P. Roberts; D. K. Chakraworty; K. M. Merz, Jr. Rational Design of Particle Mesh Ewald Compatible [illegible] Simulations 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 [illegible] *J. Comput. Chem.*, **2004**, *25*, 1157–1174.

R. Galindo-Murillo; J. C. Robertson; M. Zgarbovic; J. Sponer; M. Otyepka; P. Jureska; T. E. Cheatham. [illegible] DNA. *J. Chem. Theory Comput.*, **2016**, [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 [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.

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

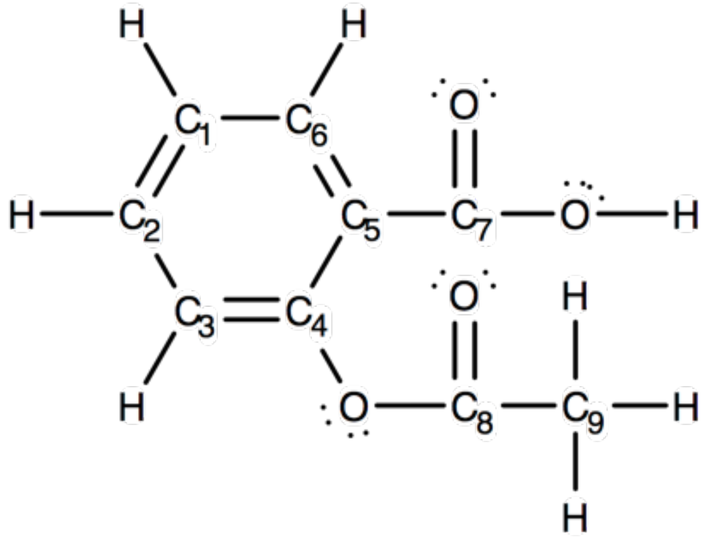
Open source infrastructure: for improving force fields with in-house data

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

http://openforcefield.org

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

input molecular graph

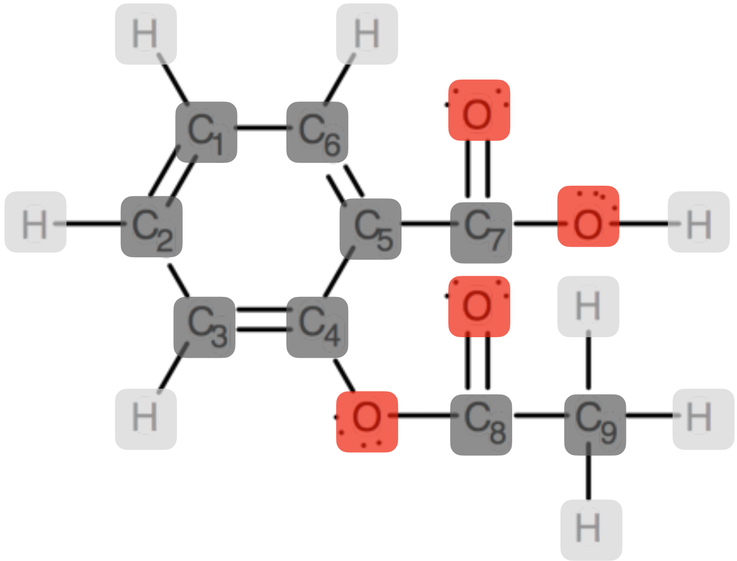


aspirin

JOSH FASS

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

"atom-typed" molecule



aspirin

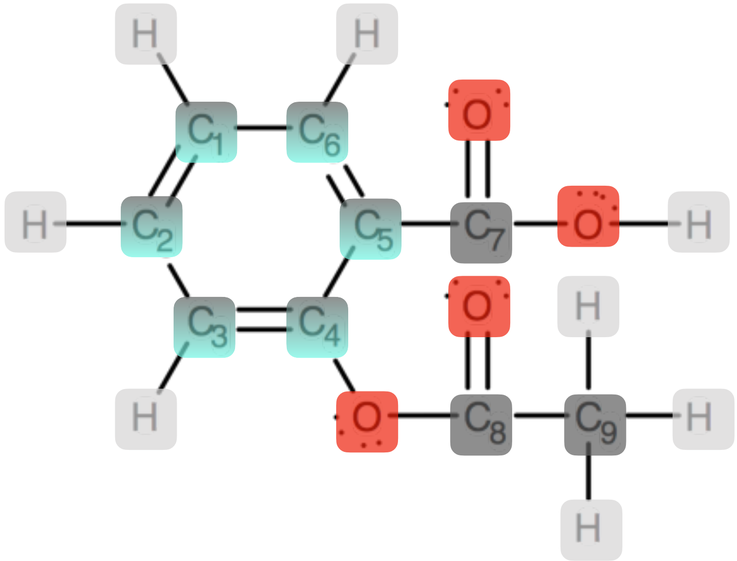
3 atom-types

- *hydrogen*
- *carbon*
- *oxygen*

JOSH FASS

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

"atom-typed" molecule



aspirin

4 atom-types

hydrogen

carbon

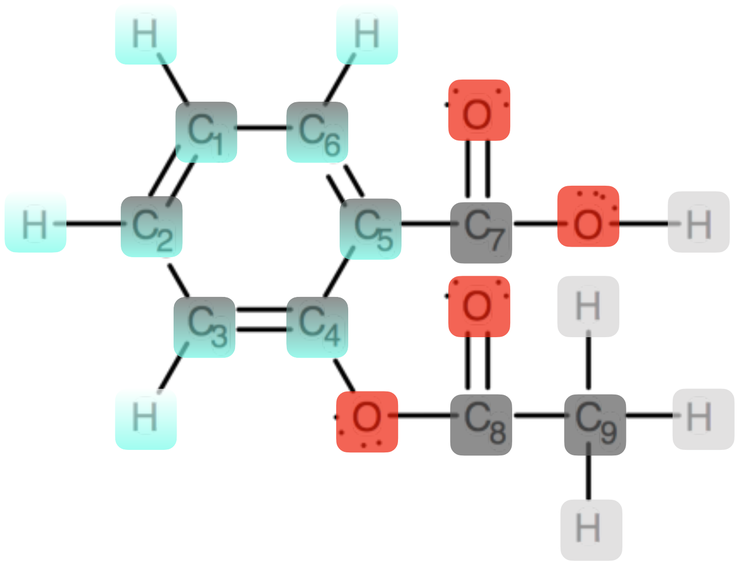
carbon in an aromatic ring

oxygen

JOSH FASS

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

"atom-typed" molecule



aspirin

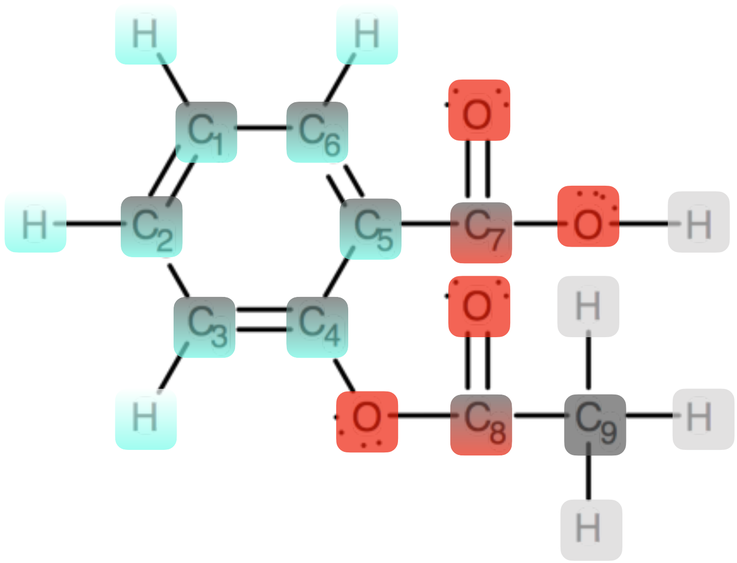
5 atom-types

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

JOSH FASS

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

"atom-typed" molecule



aspirin

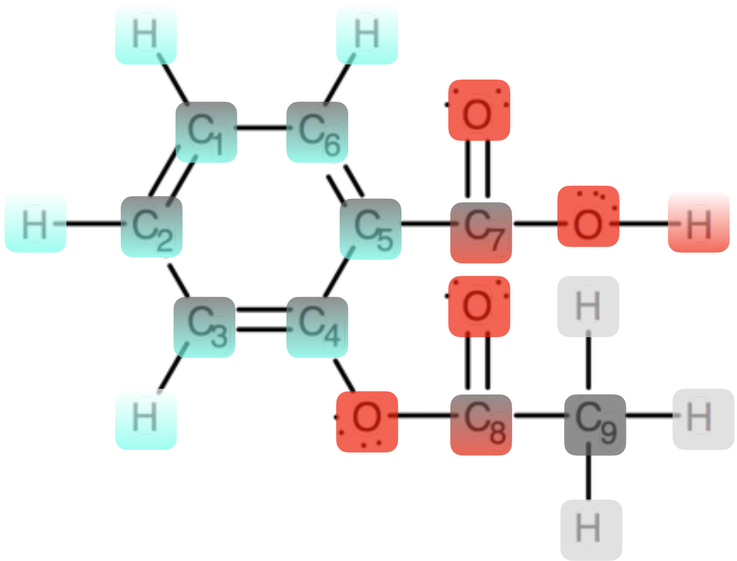
6 atom-types

- *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 PARAMETERIZASTION IS HARD BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

"atom-typed" molecule



aspirin

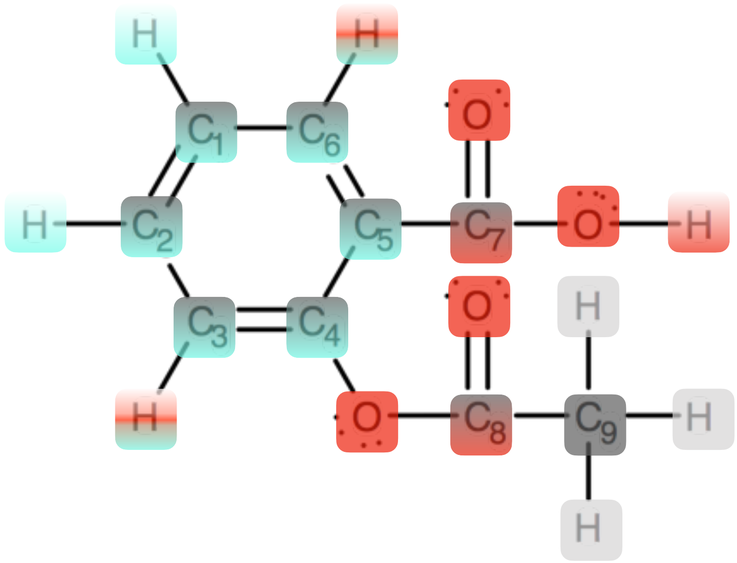
7 atom-types

- *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 PARAMETERIZASTION IS HARD BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

"atom-typed" molecule



aspirin

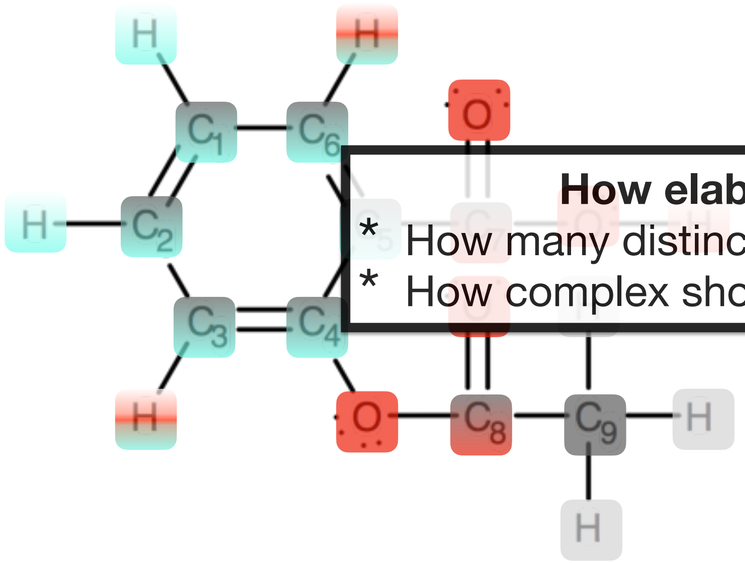
8 atom-types

- *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 PARAMETERIZASTION IS HARD BECAUSE IT'S A MIXED DISCRETE-CONTINUOUS OPTIMIZATION PROBLEM

"atom-typed" molecule



aspirin

How elaborate should we go?

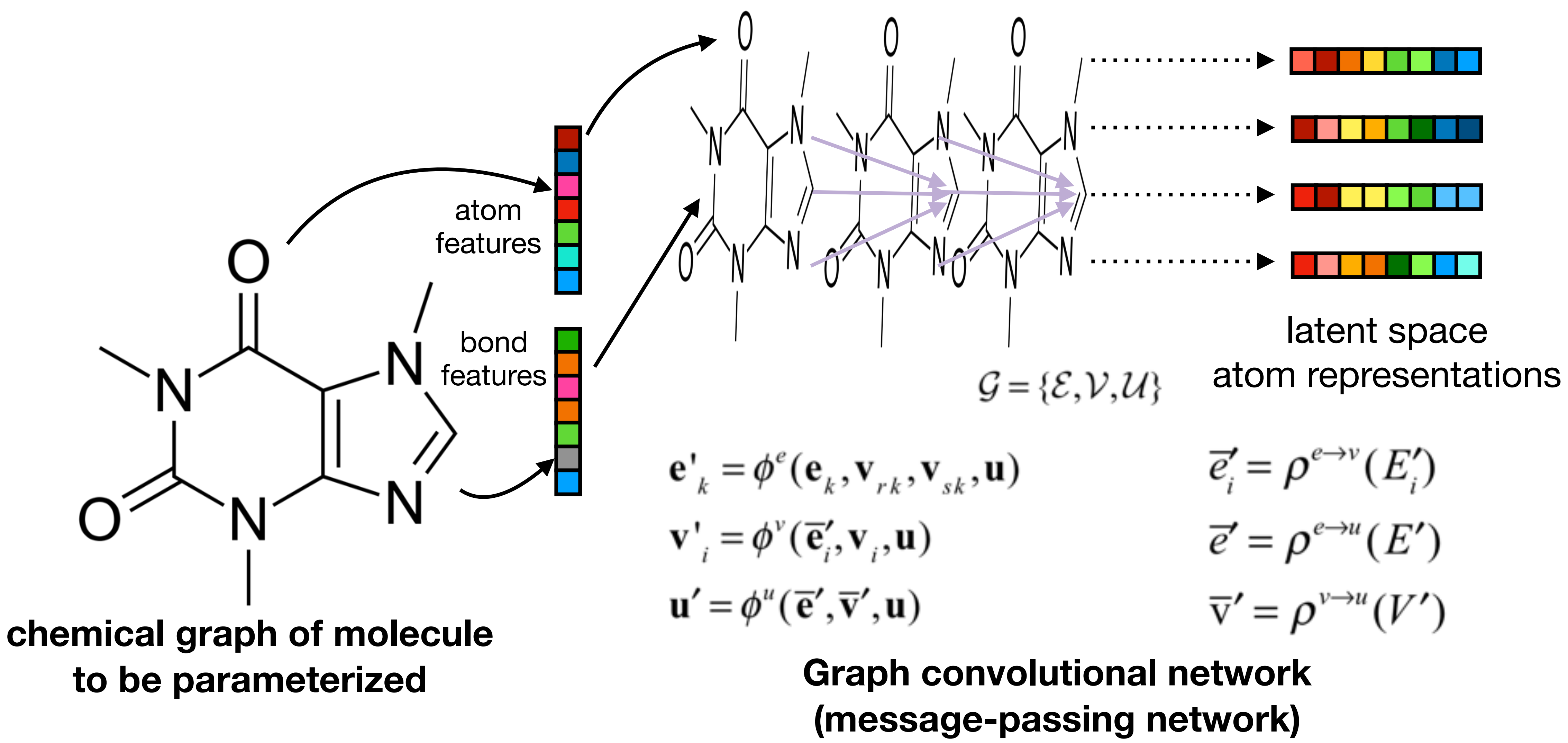
* How many distinct atom types are justified?
* How complex should their definitions be?

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

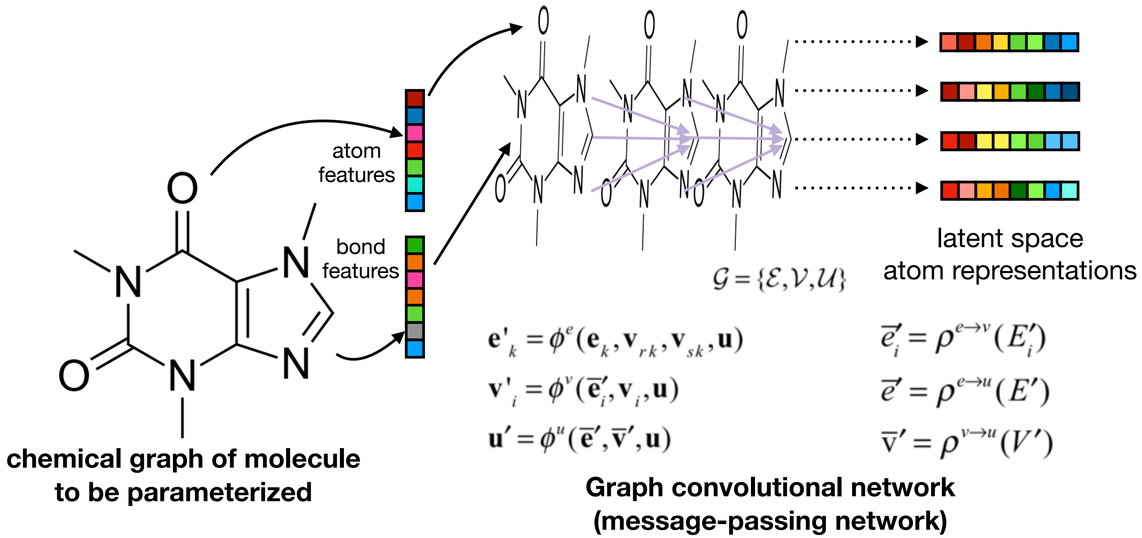
GRAPH CONVOLUTIONAL NETWORKS CAN LEARN CHEMICAL ENVIRONMENTS WITHOUT REQUIRING DISCRETE ATOM TYPES

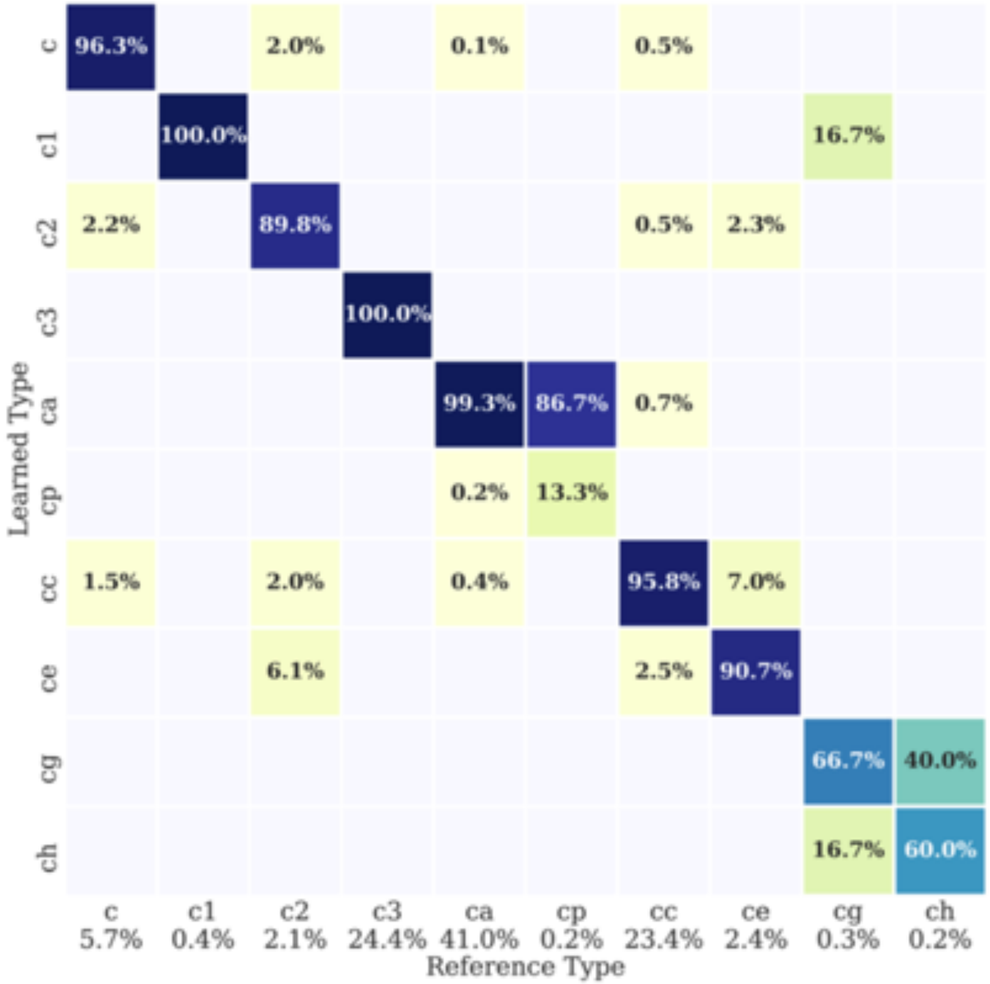


JOSH FASS

YUANQING WANG

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

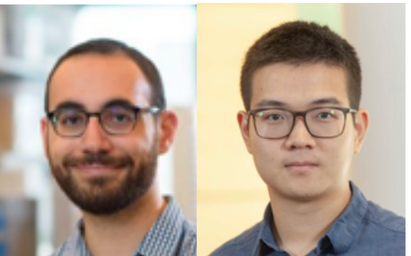




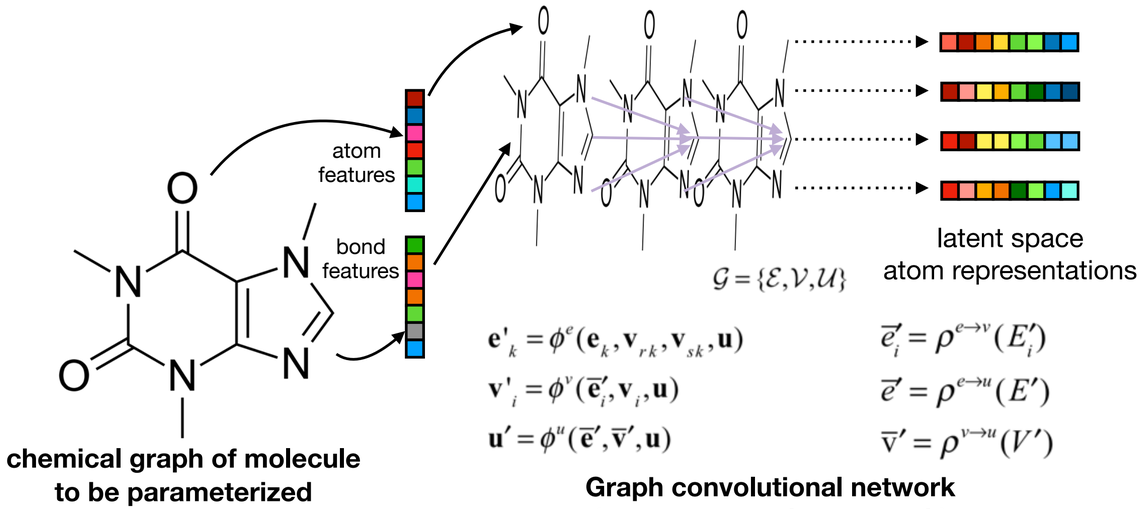
GAFF 1.81 atom types predicted with 98.31% [95% CI: 97.94, 98.63] accuracy

JOSH FASS

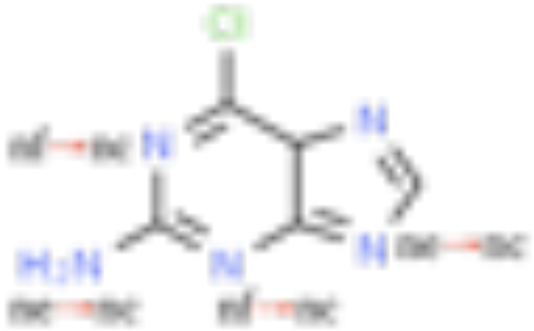
YUANQING WANG



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



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

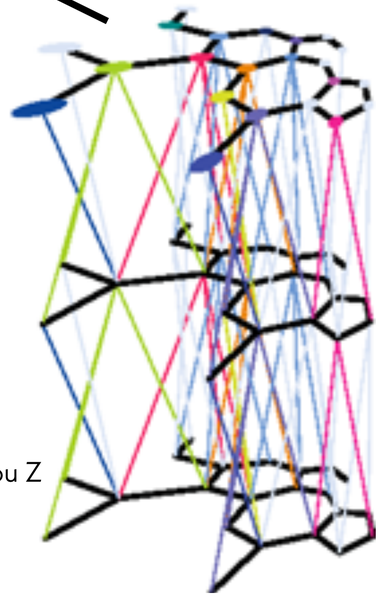


Figure adapted from Zhou Z arXiv:1706.09916

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

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

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

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

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

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

GRAPH CONVOLUTIONAL NETWORKS ARE PARTICULARLY WELL-SUITED TO CHEMISTRY

molecule **bond** **atom**

predict properties

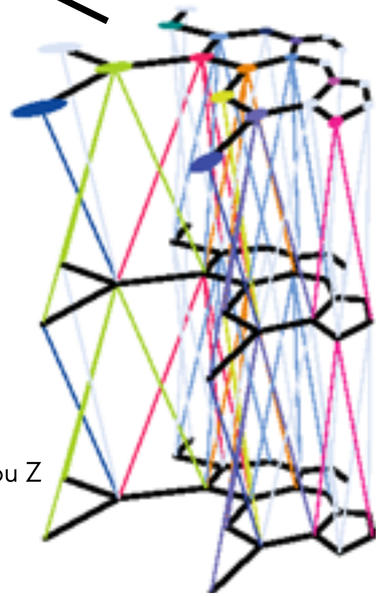
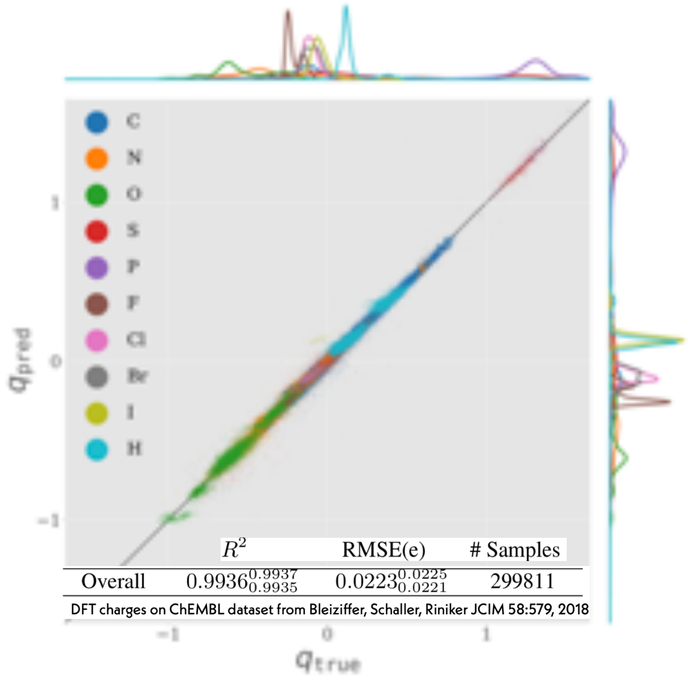


Figure adapted from Zhou Z arXiv:1706.09916

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

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

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



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

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

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

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

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

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

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

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

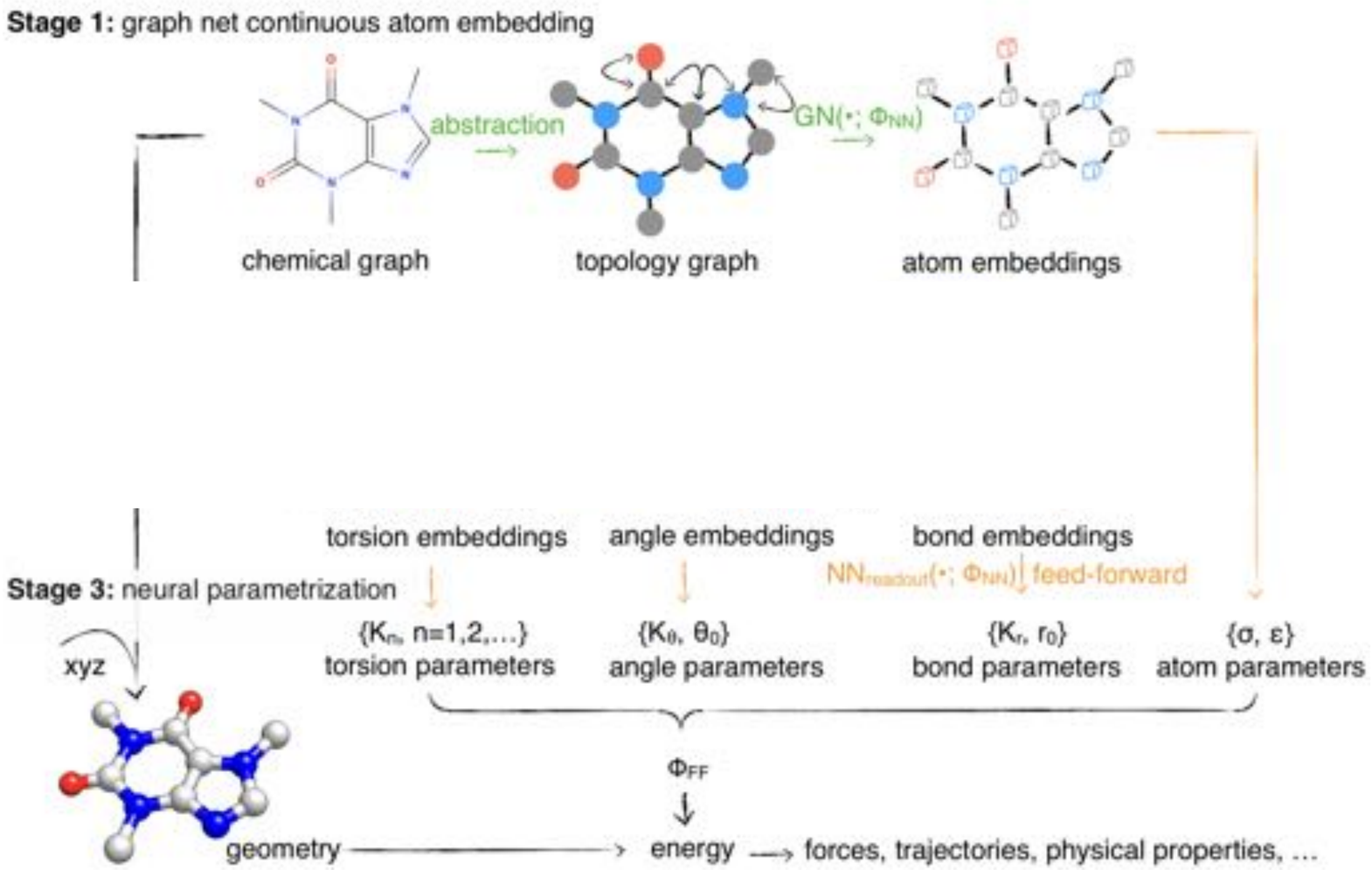
Gimlet

Graph Inference on MoLEcular Topology

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

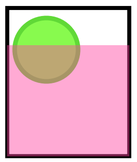
YUANQING WANG

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



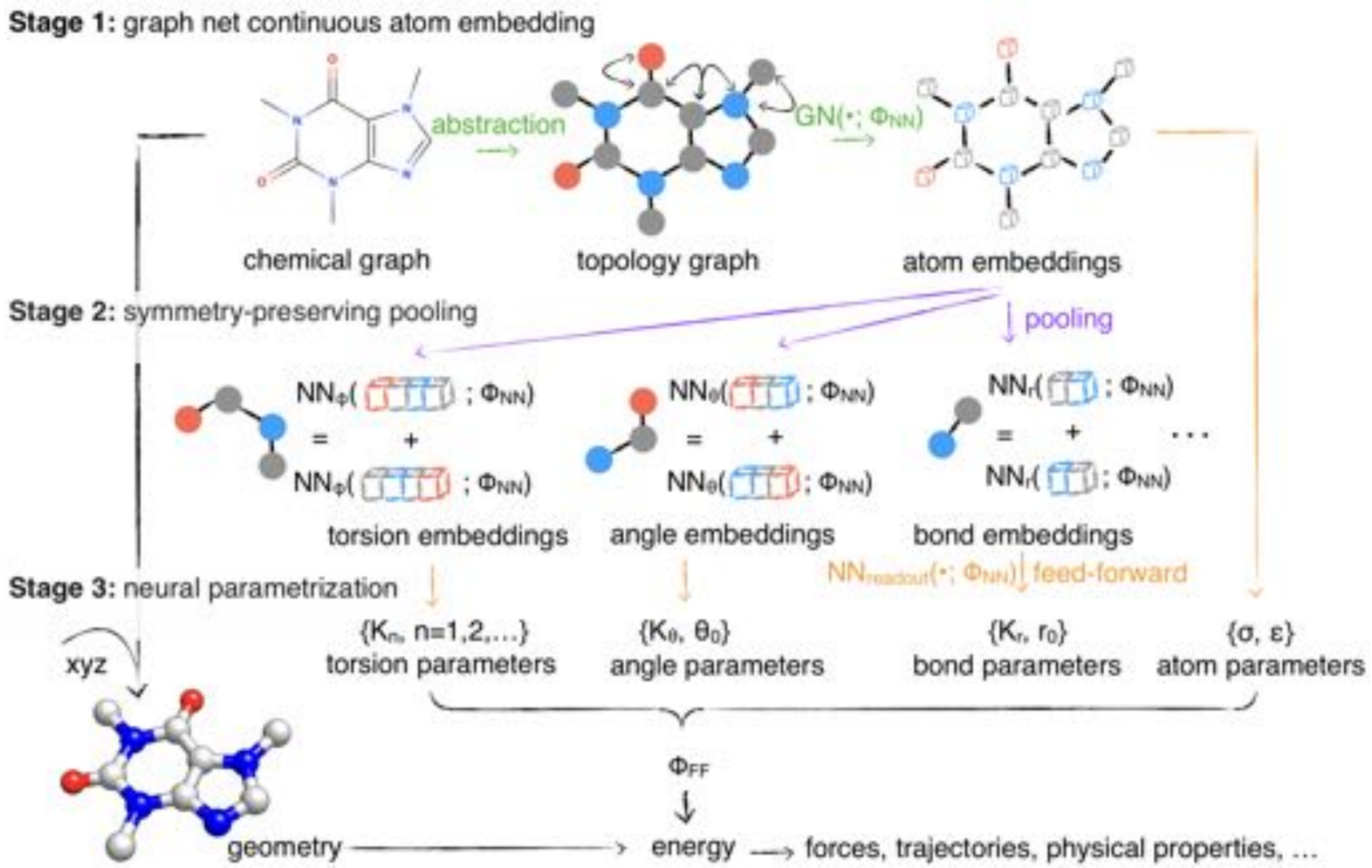
JOSH FASS

YUANQING WANG



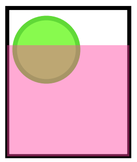
preprint: https://arxiv.org/abs/2010.01196
code: https://github.com/choderalab/espaloma

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



JOSH FASS

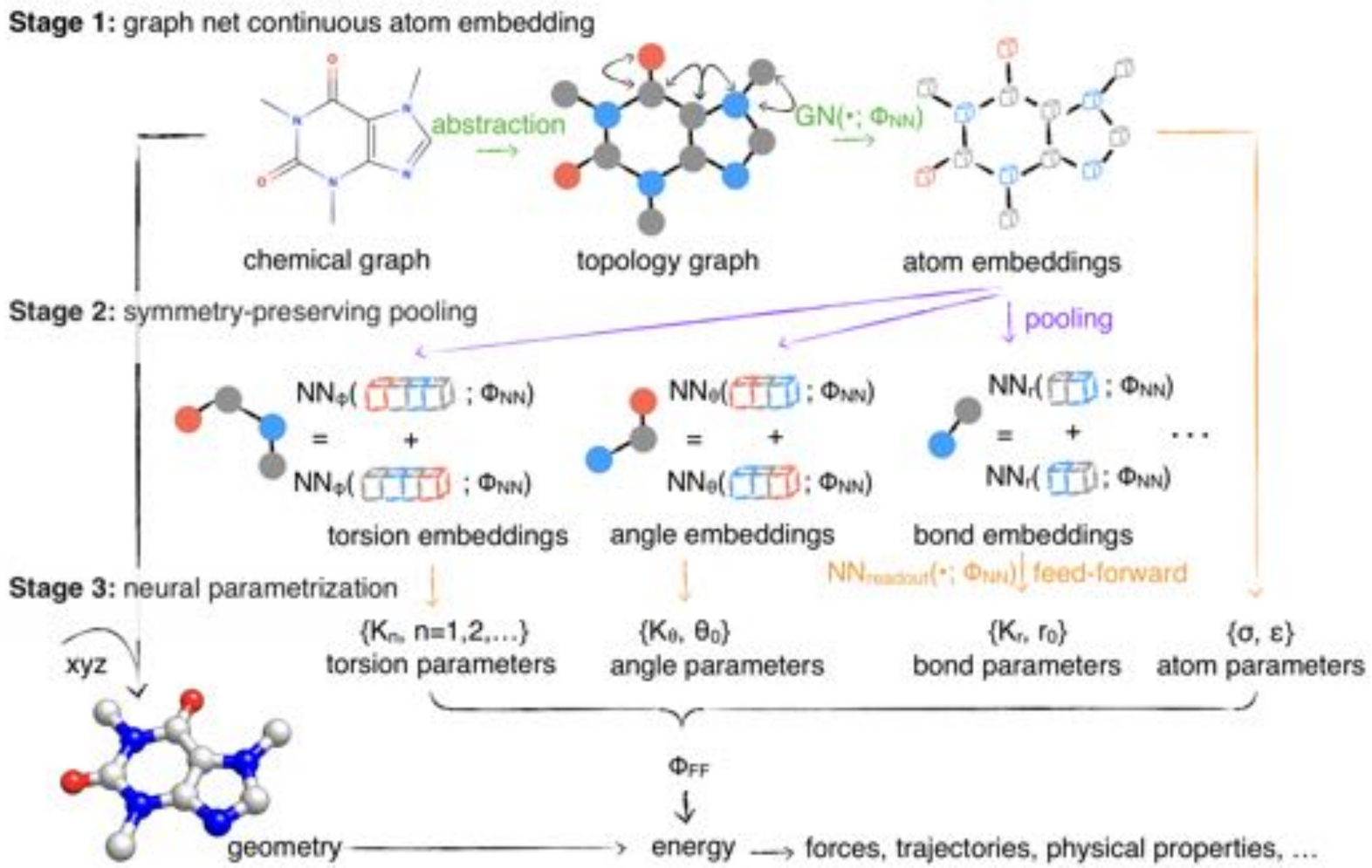
YUANQING WANG



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

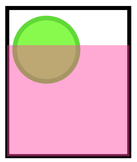
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm

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



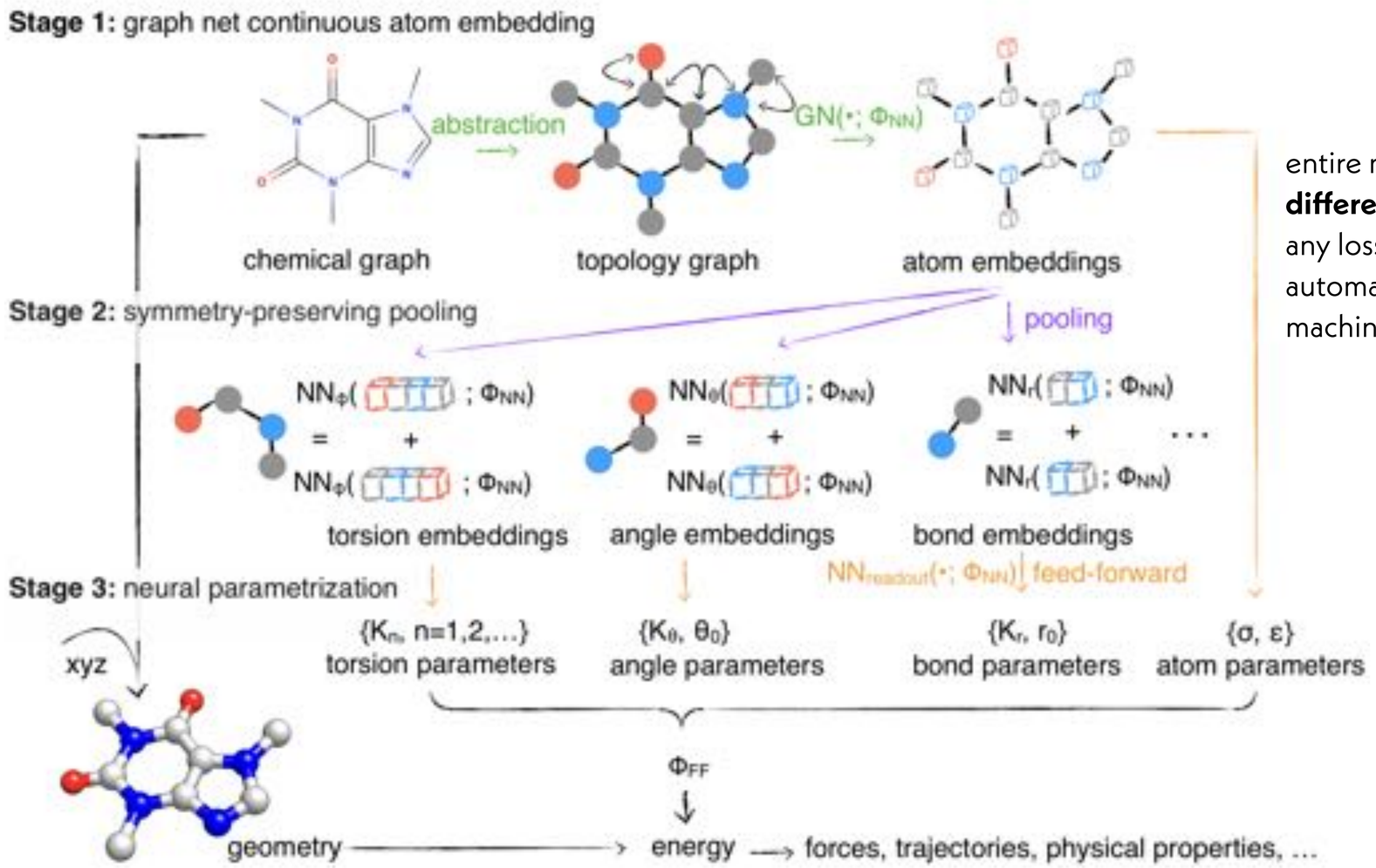
JOSH FASS

YUANQING WANG



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

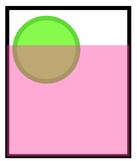
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



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

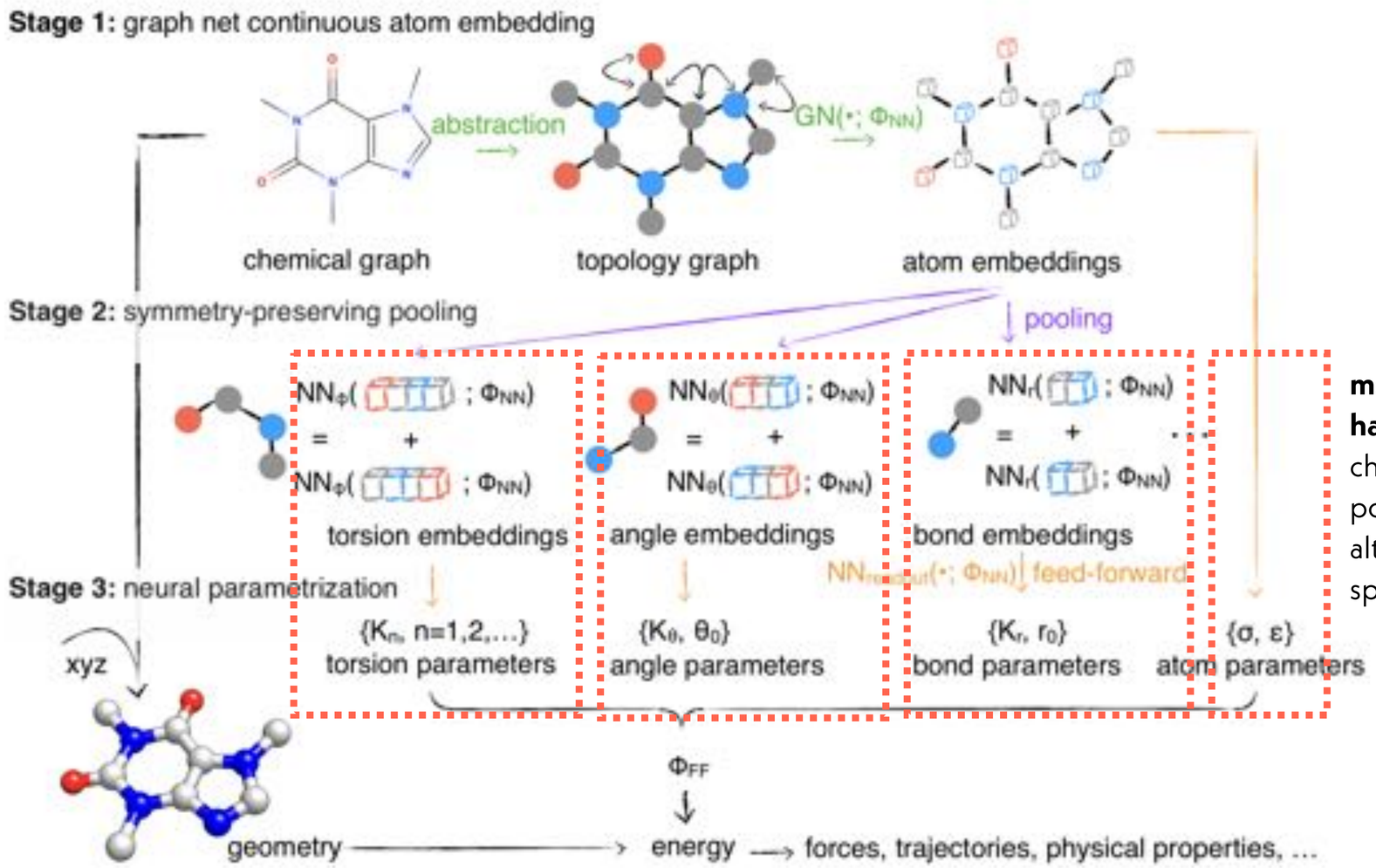
JOSH FASS

YUANQING WANG



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

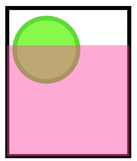
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



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

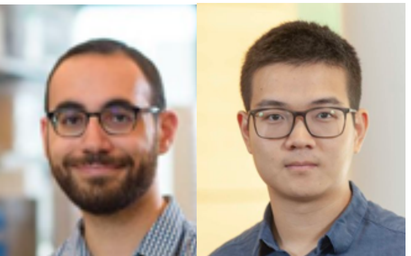
ESPALOMA CAN LEARN EXISTING MOLECULAR MECHANICS FORCE FIELDS

recovery of GAFF 1.81 parameters and energies on AlkEthOH test set

Quantity	Test RMSE	Test MAPE	Test R^2
Harmonic Bond + Angle Energy (kcal/mol)	$0.4392^{0.4392}_{0.4392}$	$0.0157^{0.0162}_{0.0153}$	$0.9958^{0.9961}_{0.9955}$
Bond Force Constant k_r (kcal / (mol * angstrom ** 2))	$35.4048^{50.2660}_{18.0387}$	$0.0180^{0.0215}_{0.0148}$	$0.8619^{0.9653}_{0.7154}$
Equilibrium Bond Length b_r (angstrom)	$0.0127^{0.0200}_{0.0013}$	$0.0015^{0.0021}_{0.0011}$	$0.9956^{1.0000}_{0.9890}$
Angle Force Constant k_θ (kcal / (mol * rad ** 2))	$3.7995^{3.9648}_{3.6293}$	$0.0276^{0.0290}_{0.0264}$	$0.8601^{0.8805}_{0.8361}$
Equilibrium Angle Value b_θ (rad)	$0.0043^{0.0045}_{0.0041}$	$0.0018^{0.0018}_{0.0017}$	$0.9202^{0.9335}_{0.9018}$

JOSH FASS

YUANQING WANG

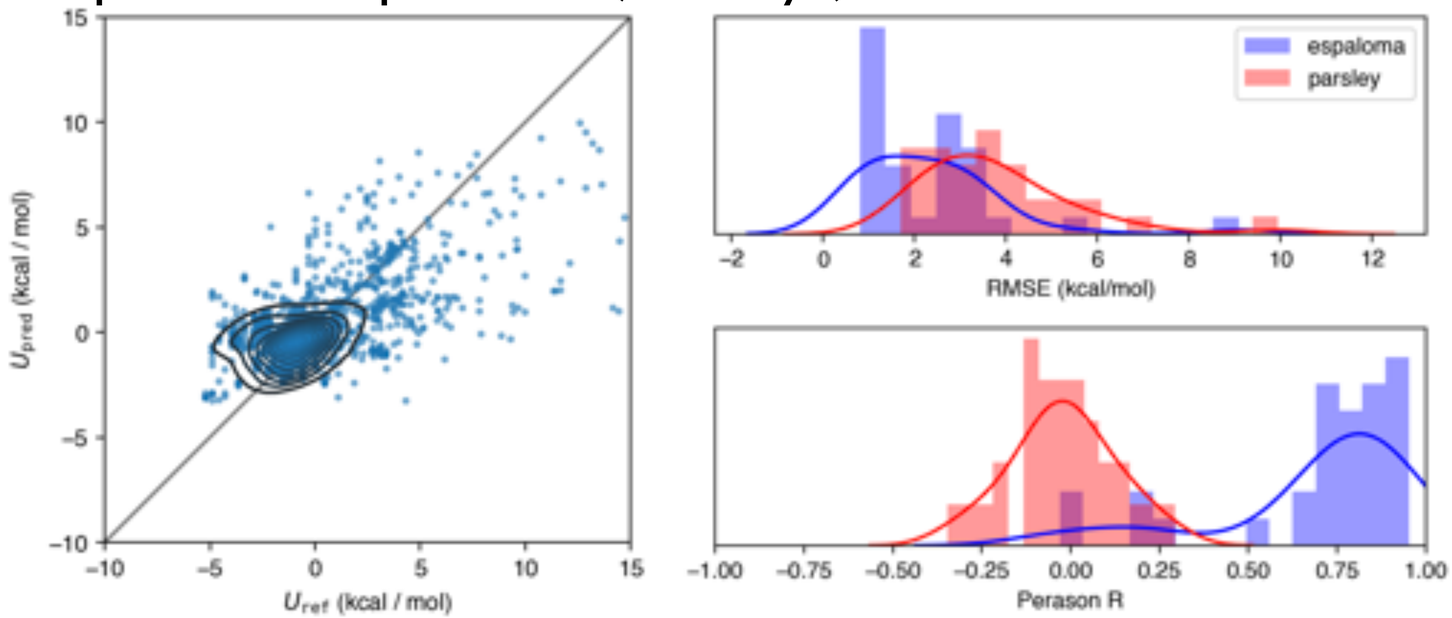


preprint: https://arxiv.org/abs/2010.01196

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

ESPALOMA CAN EASILY FIT BOTH QUANTUM CHEMICAL AND PHYSICAL PROPERTY DATA

energies and gradients from OpenFF OptimizationDataset 1.0 from MolSSI QCArchive
compared with opeff-1.2.0 (“Parsley”) force field fit to same data



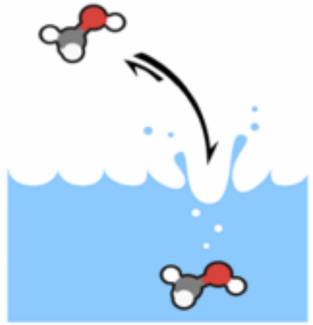
JOSH FASS

YUANQING WANG

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

MolSSI QCArchive: https://qcarchive.molssi.org

ESPALOMA CAN EASILY FIT BOTH QUANTUM CHEMICAL AND PHYSICAL PROPERTY DATA

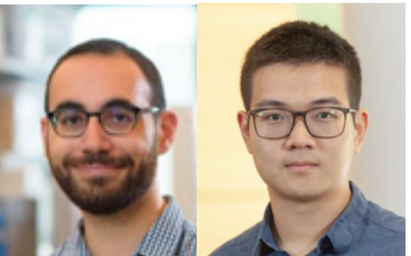


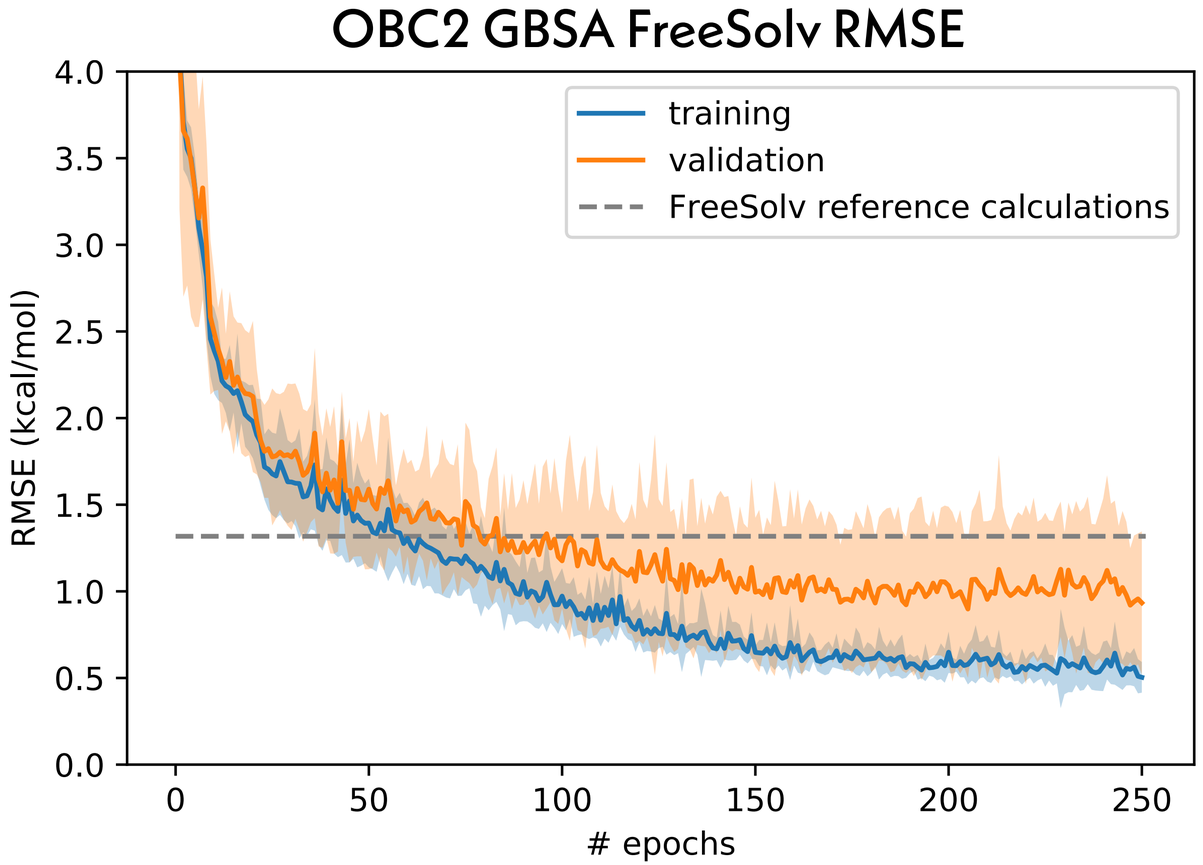
experimental hydration
free energies from **FreeSolv**
https://github.com/MobleyLab/FreeSolv

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





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

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

bond-bond: angle node

angle-angle: torsion node

bond-angle: angle node

torsion-(center) bond: torsion

torsion-(side) bond: torsion

torsion-angle: torsion

torsion-angle-angle: torsion

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

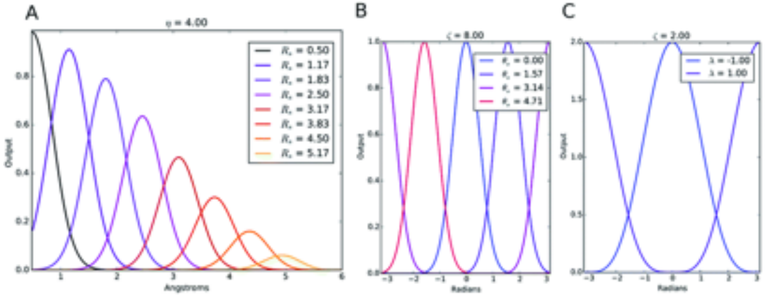
ANI family of quantum machine learning (QML) potentials

radial and **angular** features

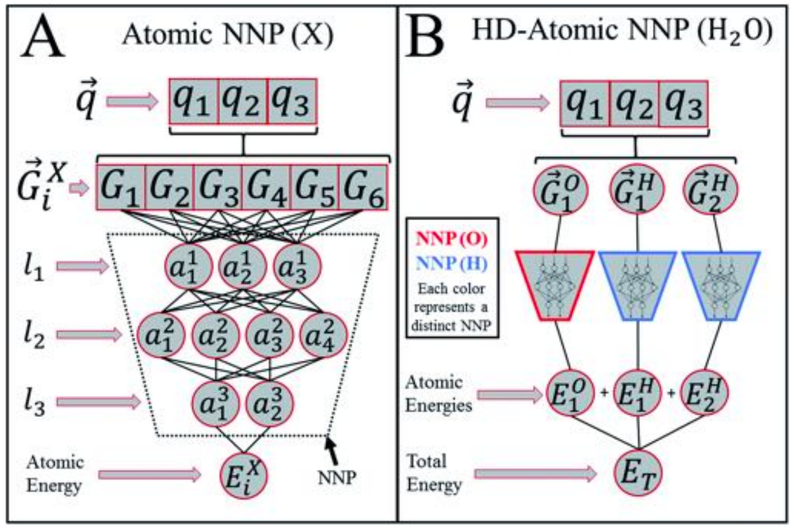
$$f_C(R_{ij}) = \begin{cases} 0.5\times\cos\left(\frac{\pi R_{ij}}{R_C}\right)+0.5 & \text{for } R_{ij} \le R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases}$$

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

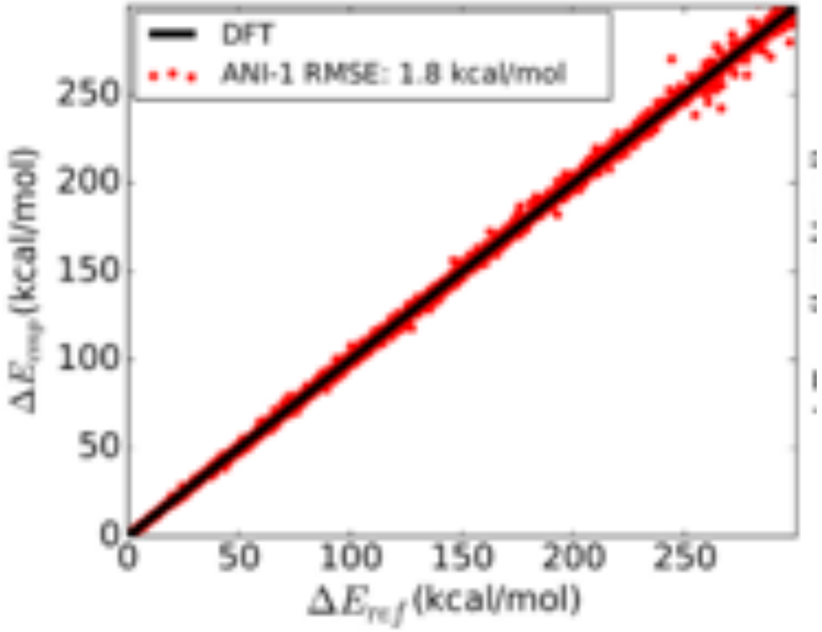
$$G_m^{A_{mod}} = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all atoms}} \left(1+\cos\left(\theta_{ijk}-\theta_s\right)\right)^\zeta \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_s\right)^2\right] f_C(R_{ij}) f_C(R_{ik})$$



deep neural network for each atom



excellent agreement with DFT

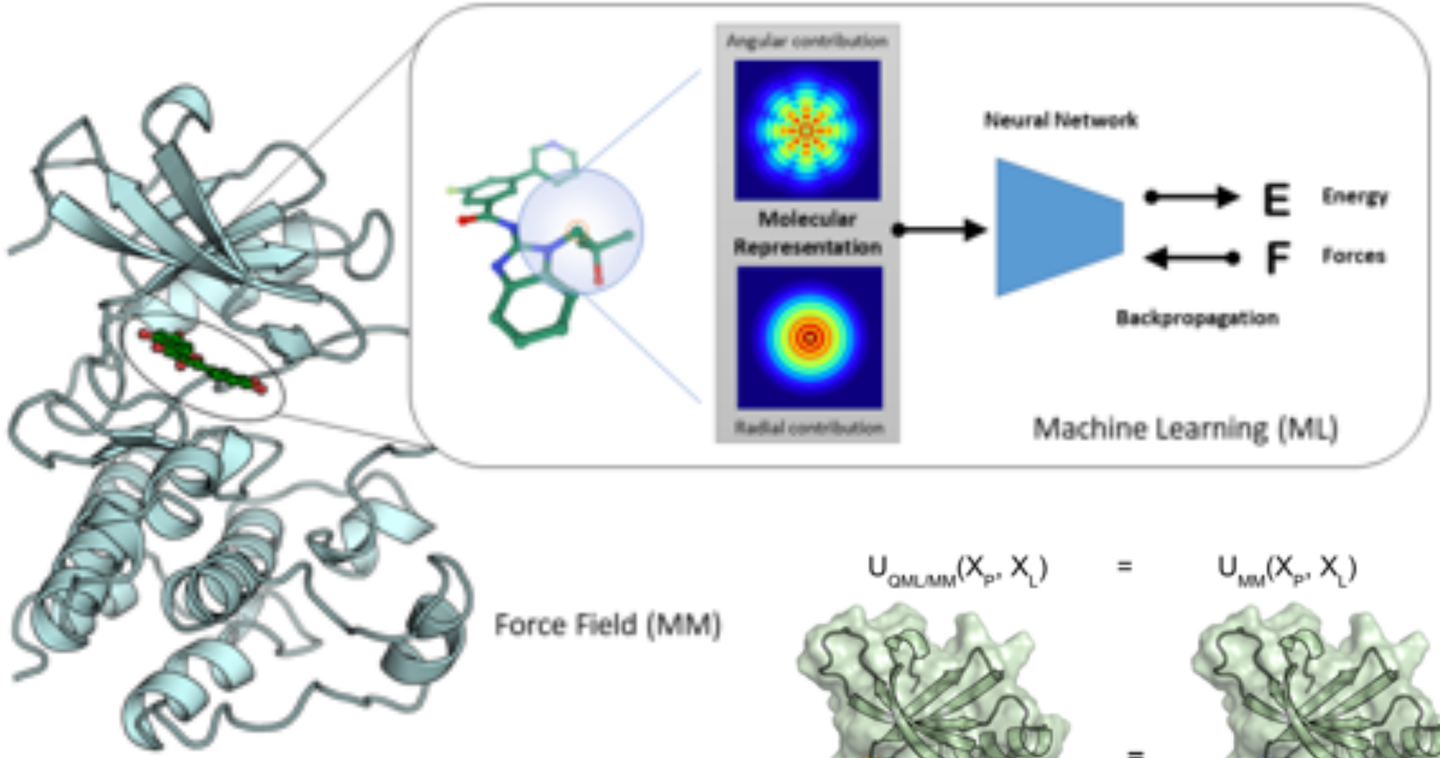


OLEXANDR ISAYEV

ADRIAN ROITBERG

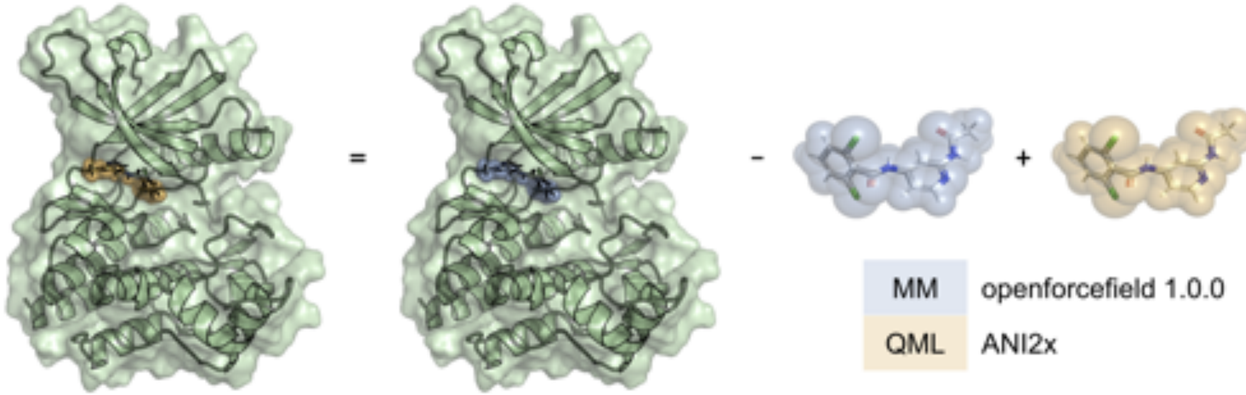
Smith, Isayev, Roitberg. Chemical Science 8:3192, 2017.
http://doi.org/10.1039/c6sc05720a

HYBRID QML/MM POTENTIALS ARE A NEAR-TERM PRACTICAL APPROACH TO MORE ACCURATE MODELING FOR DRUG DISCOVERY



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

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



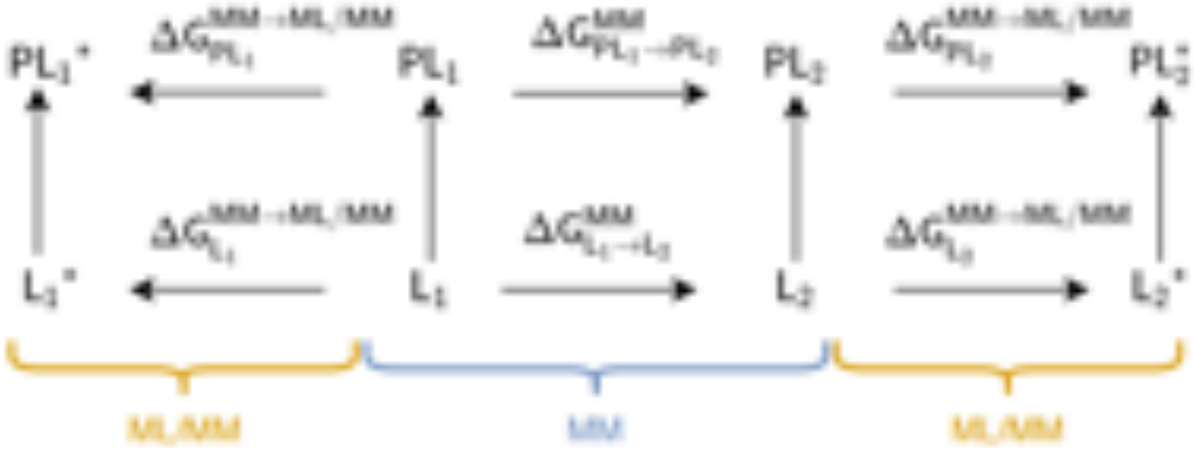
MM	openforcefield 1.0.0
QML	ANI2x

DOMINIC RUFA **HANNAH BRUCE MACDONALD**

OLEXANDR ISAYEV **ADRIAN ROITBERG**

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 PERTURB MM FREE ENERGY CALCULATIONS TO QML/MM WITH AN EFFICIENT NONEQUILIBRIUM SCHEME

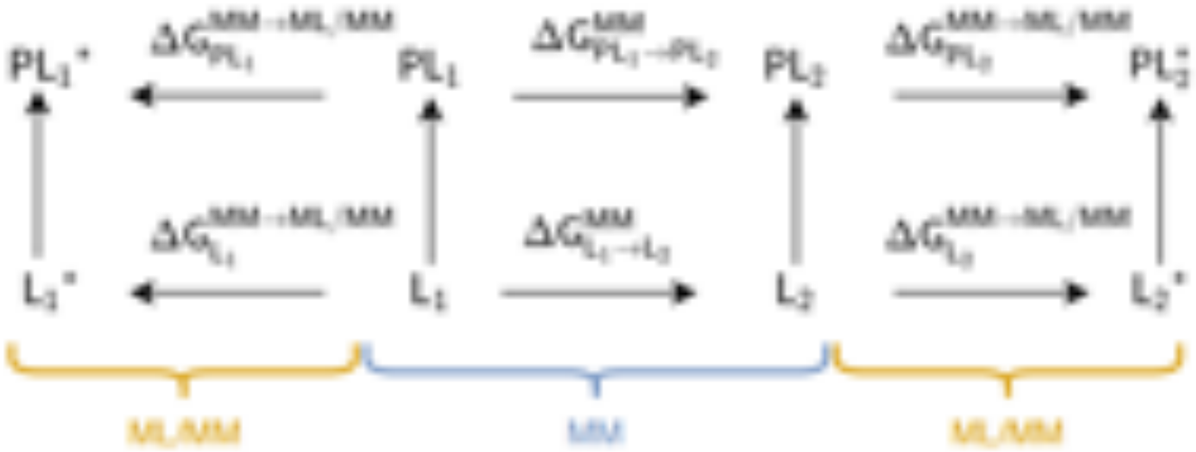


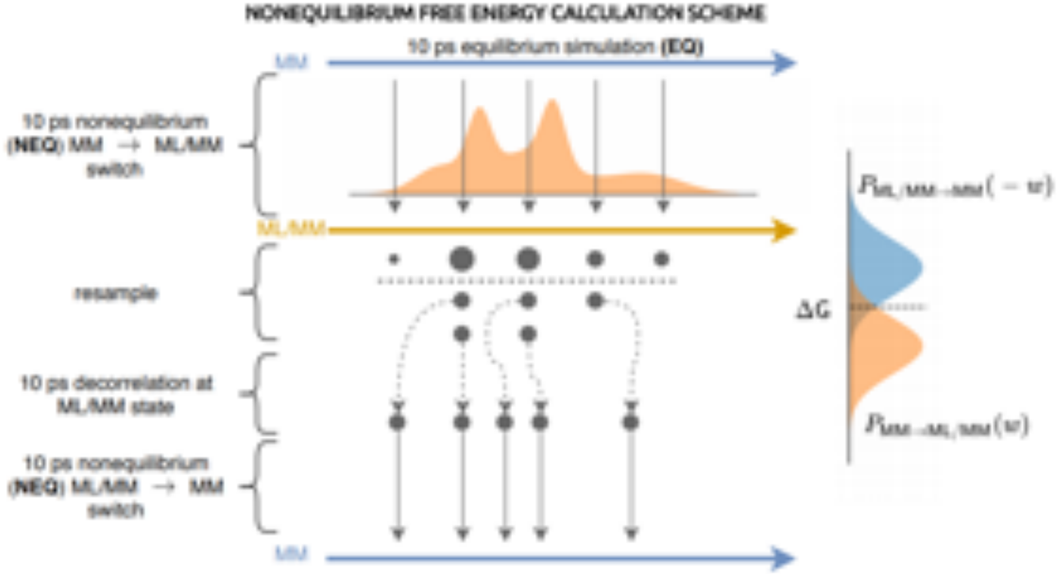
DOMINIC RUFA

HANNAH BRUCE MACDONALD

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 PERTURB MM FREE ENERGY CALCULATIONS TO QML/MM WITH AN EFFICIENT NONEQUILIBRIUM SCHEME





ΔΔG estimated with Bennett acceptance ratio (BAR)

DOMINIC RUFA

HANNAH BRUCE MACDONALD

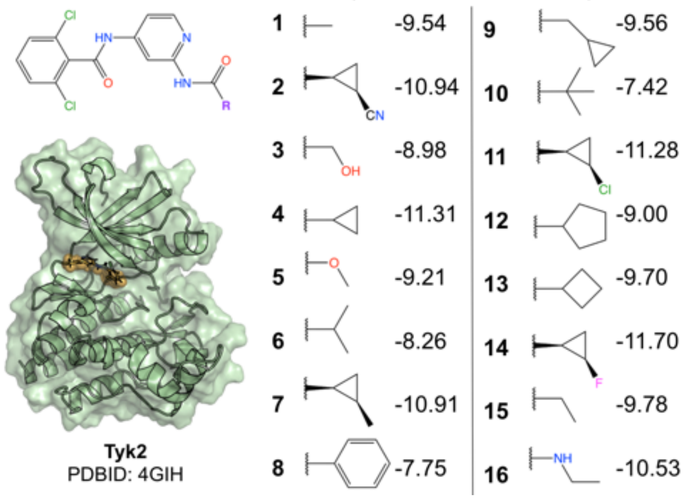
Rufa, Bruce Macdonald, Fass, Wieder, Grinaway, Roitberg, Isayev, and **Chodera**.
preprint: https://doi.org/10.1101/2020.07.29.227959
code: https://github.com/choderalab/qmlify

UNMODIFIED QML/MM CAN CUT THE ERROR IN BINDING AFFINITIES 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

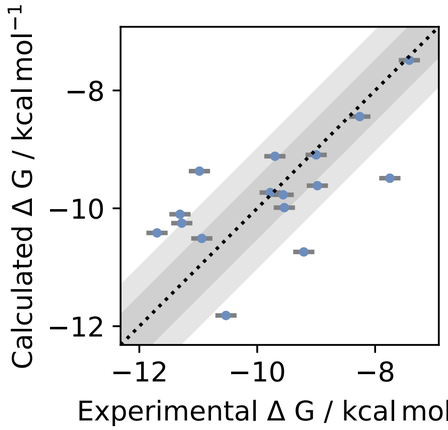
QML/MM (OpenFF 1.0.0 + ANI2x)
AMBER14SB protein force field
TIP3P; Joung and Cheatham ions



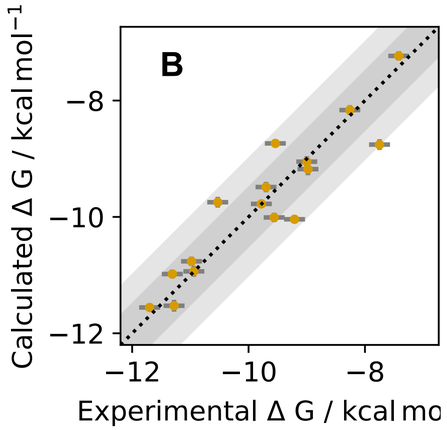
	Tyk2
no. of compds	16
binding affinity range (kcal/mol)	4.3
crystal structure	4GIH
series ref	52,53
no. of perturbations	24
MUE FEP	0.75 ± 0.11
RMSE FEP	0.93 ± 0.12

Free energies are in units of kilocalories per mole.

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




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



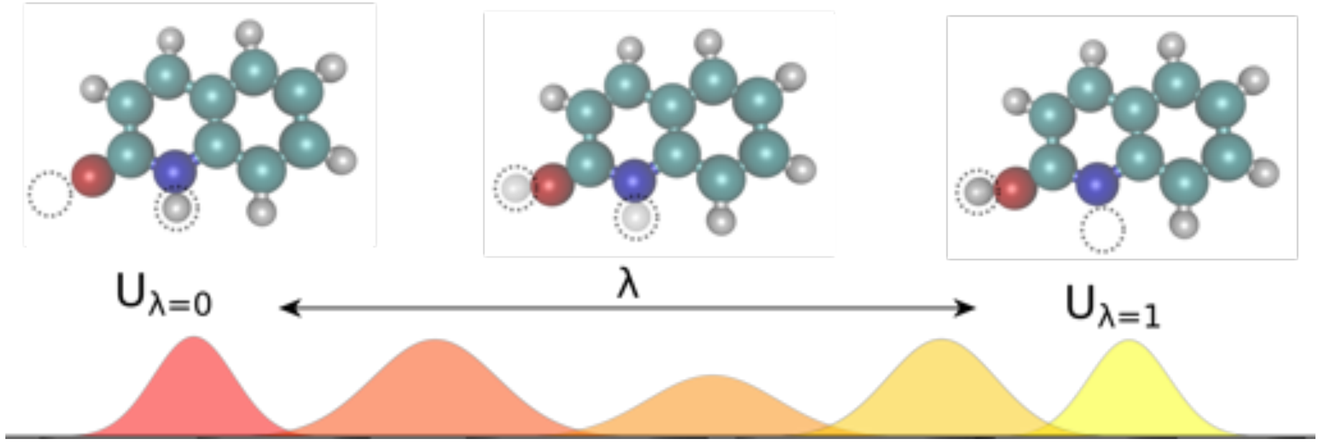
Tyk2 benchmark system from Wang et al. JACS 137:2695, 2015
replica-exchange free energy calculations with solute tempering (FEP/REST)

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

WELL-BEHAVED QML POTENTIALS MAKE ALCHEMICAL FREE ENERGY CALCULATIONS EASY

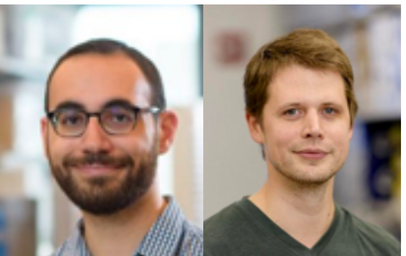
Potentials are free of singularities, so **simple linear alchemical potentials** can robustly compute alchemical free energies

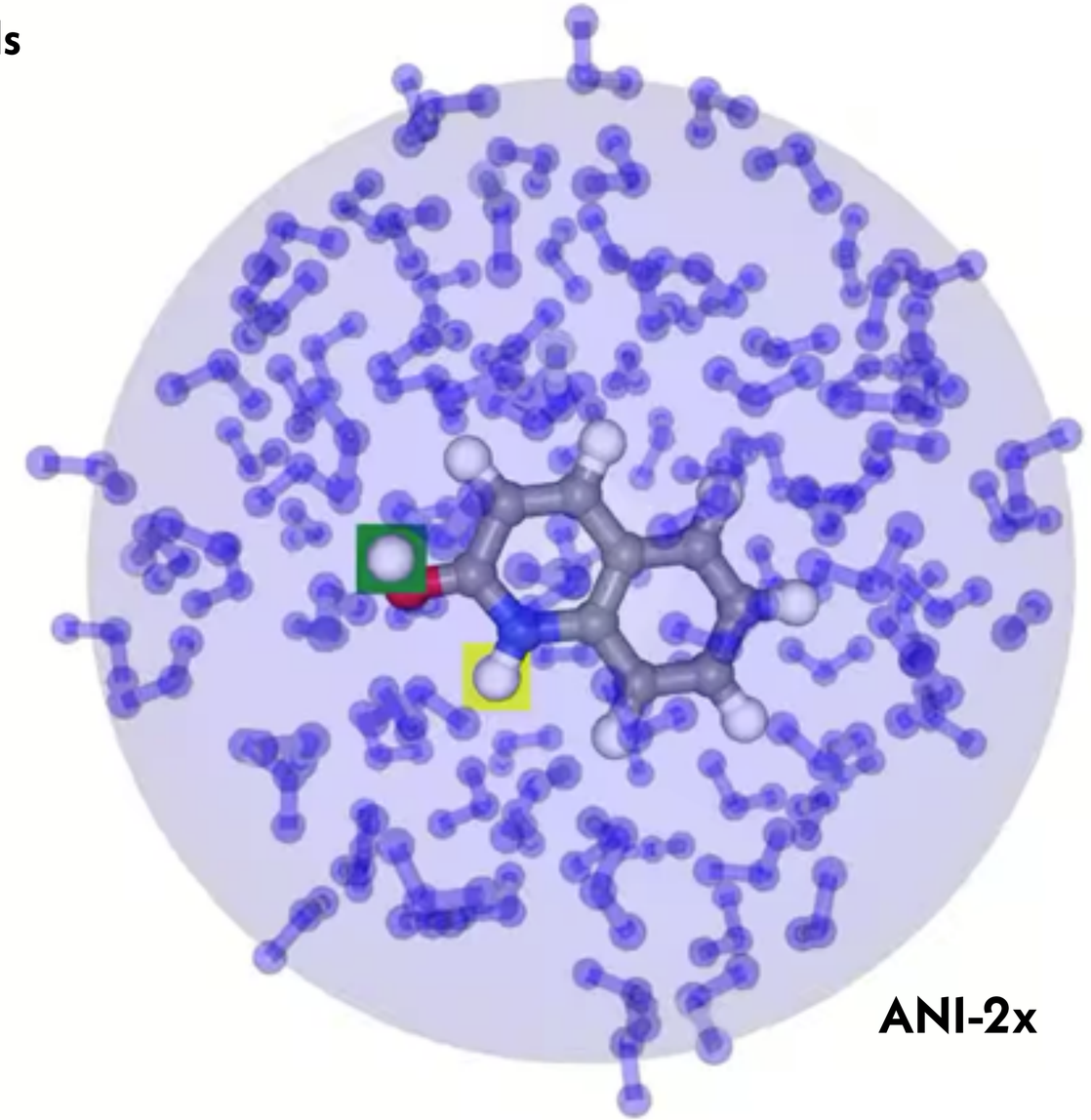
$$U(x;\lambda) = (1-\lambda)U_{\lambda=0}(x) + \lambda U_{\lambda=1}(x)$$



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

JOSH FASS **MARCUS WIEDER**





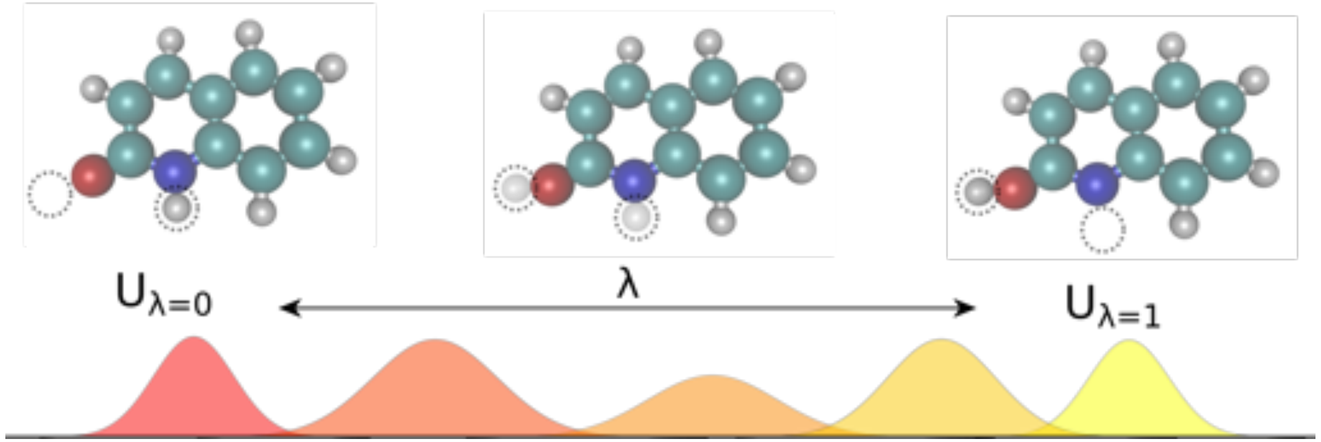
ANI-2x

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

WELL-BEHAVED QML POTENTIALS MAKE ALCHEMICAL FREE ENERGY CALCULATIONS EASY

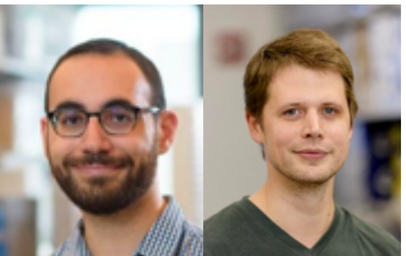
Potentials are free of singularities, so **simple linear alchemical potentials** can robustly compute alchemical free energies

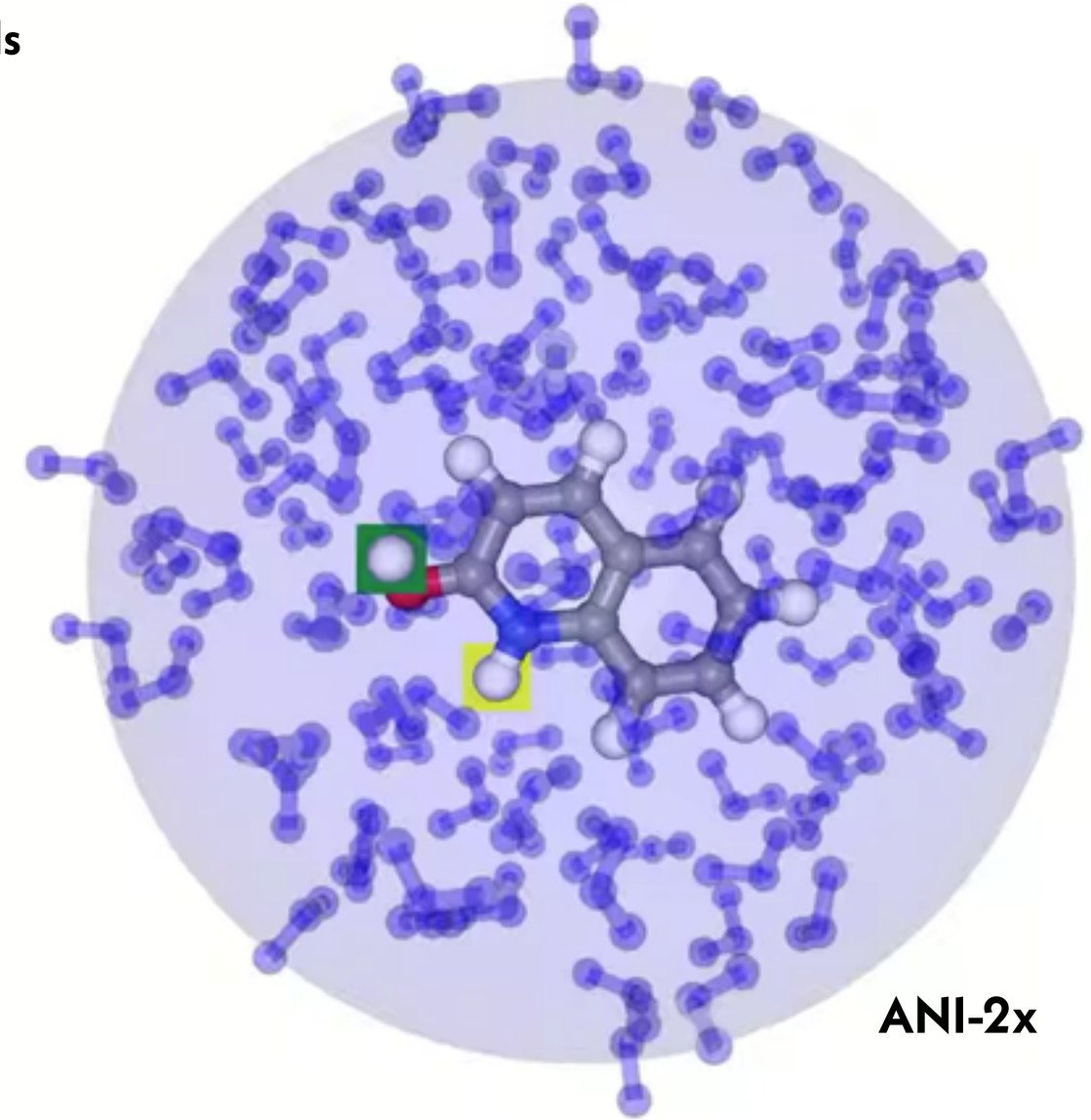
$$U(x;\lambda) = (1-\lambda)U_{\lambda=0}(x) + \lambda U_{\lambda=1}(x)$$



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

JOSH FASS **MARCUS WIEDER**



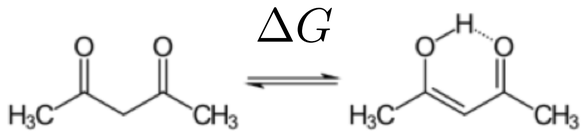


ANI-2x

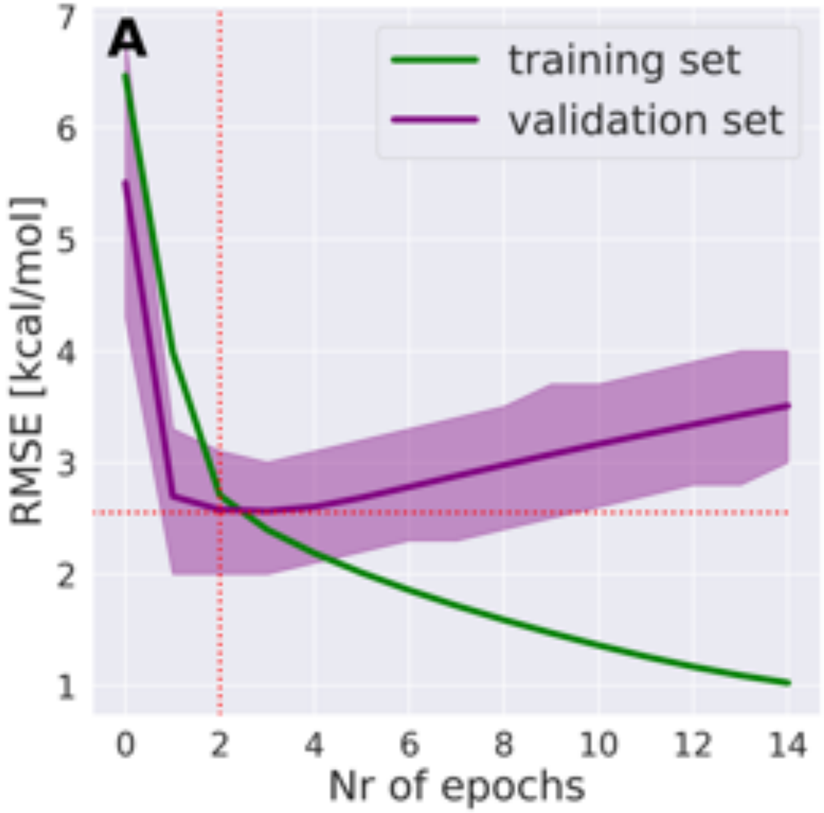
preprint: https://doi.org/10.1101/2020.10.24.353318
code: https://github.com/choderalab/neutromeratio

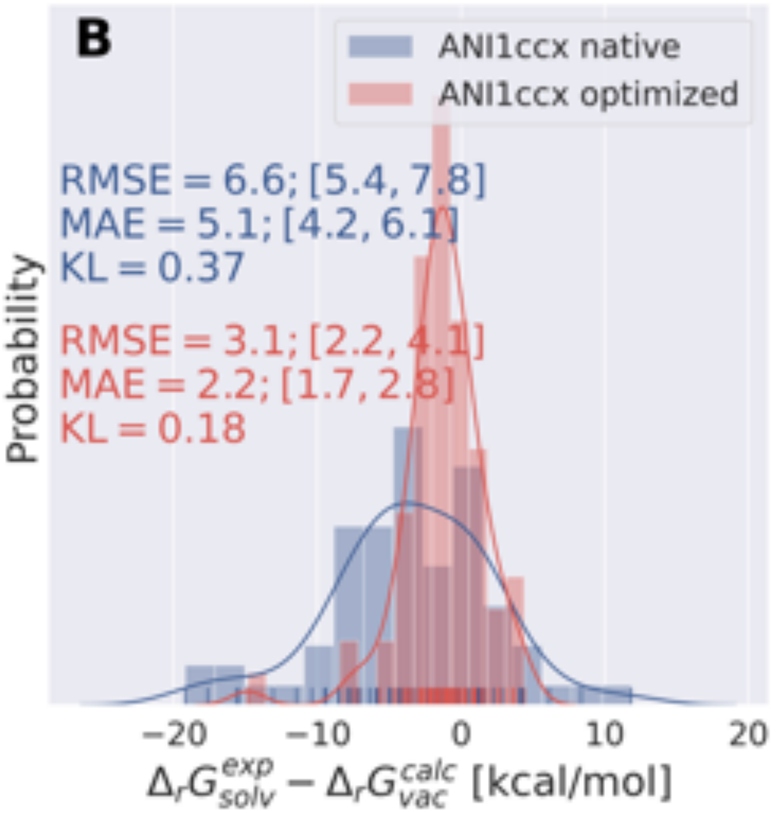
TRAINING ON FREE ENERGY DIFFERENCES FOR FOR A TRAINING SET IMPROVES FREE ENERGY PREDICTIONS ON A TEST SET

tautomer alchemical free energy difference prediction

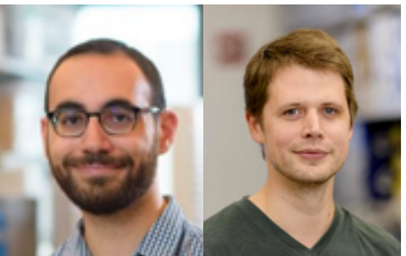


train: 221 tautomer pairs
validate: 57 tautomer pairs
test: 72 tautomer pairs



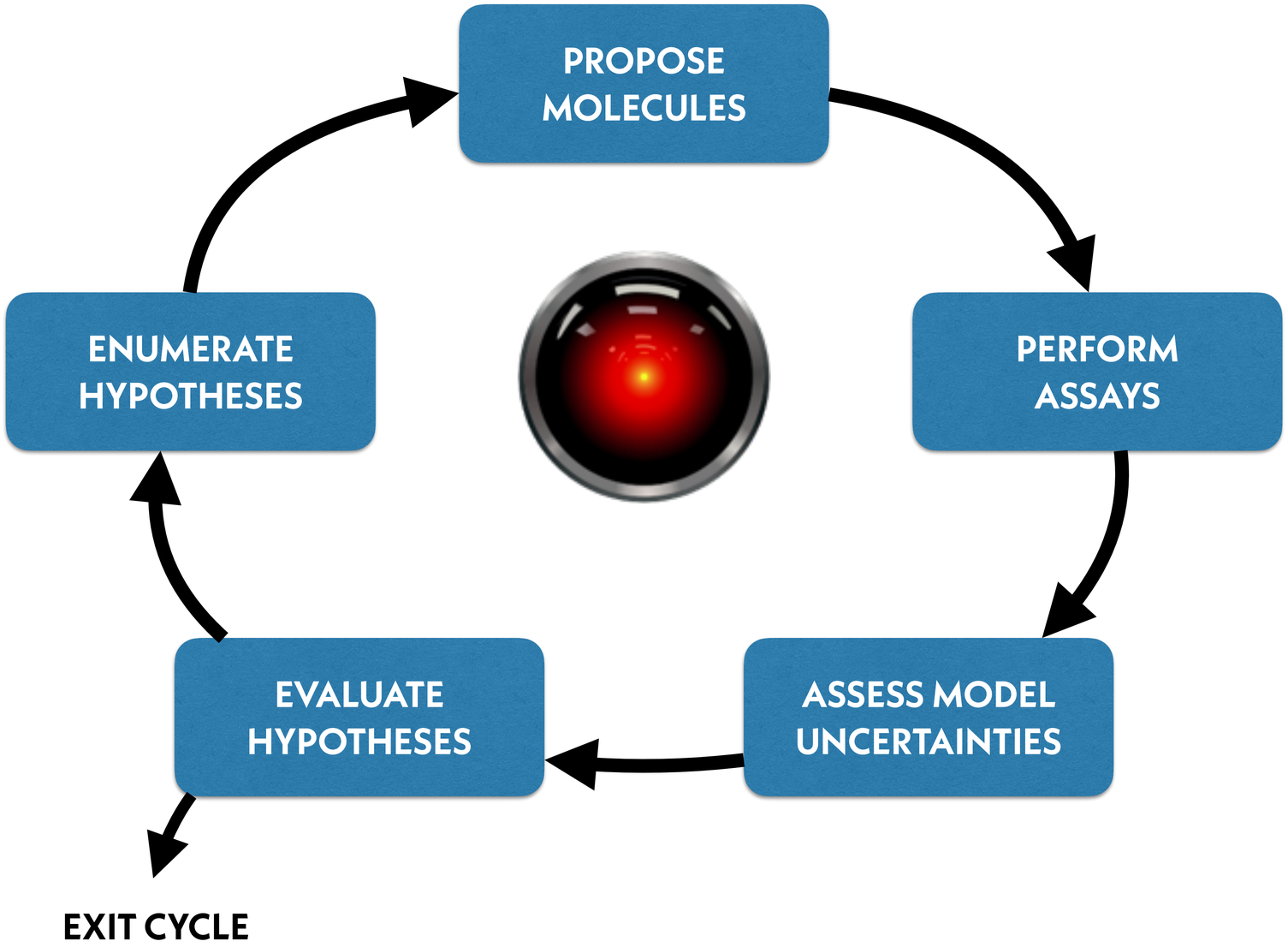


JOSH FASS MARCUS WIEDER



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

PREDICTIVE MODELS THAT LEARN ARE AN ESSENTIAL PART OF AUTONOMOUS DESIGN FRAMEWORKS



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

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