slides: http://choderalab.org/news

TEACHING FREE ENERGY CALCULATIONS TO LEARN

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

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

15 July 2022 - MLSMM - Barcelona

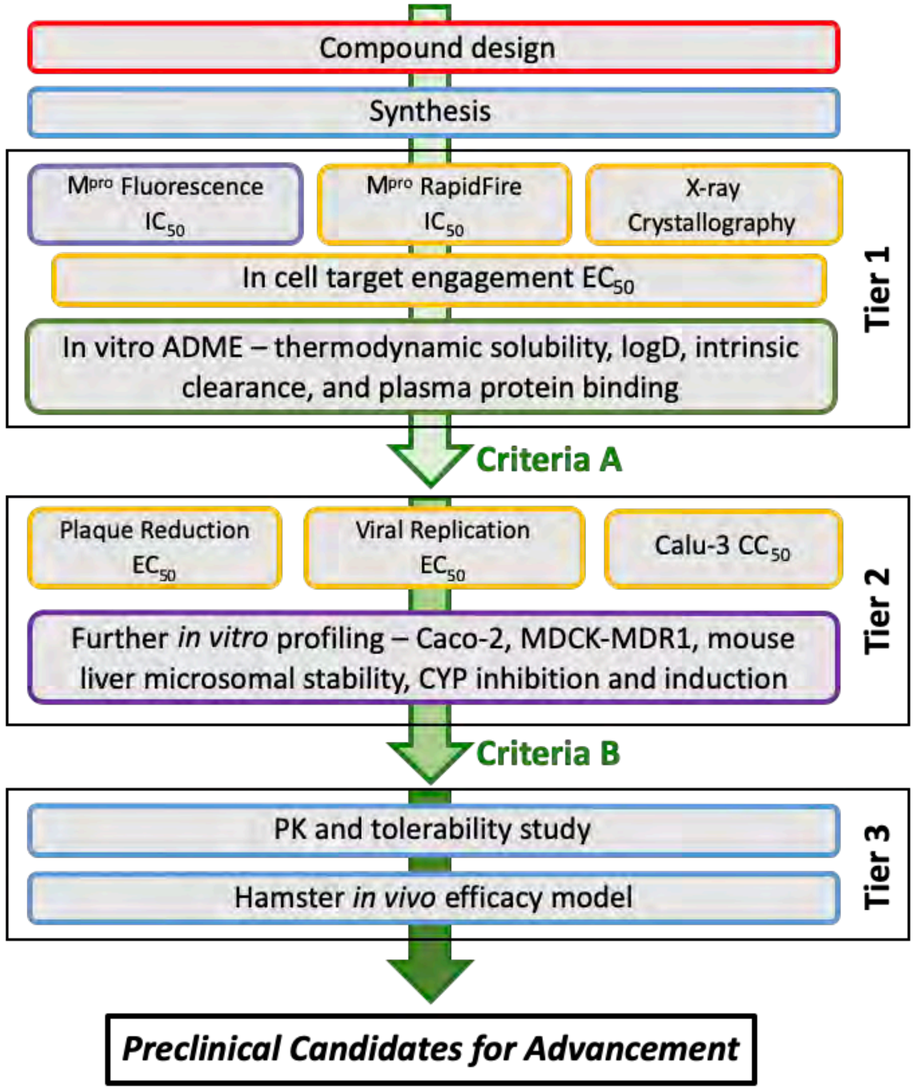
DESIGNING REAL DRUG CANDIDATES IS CHALLENGING

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

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



assay purpose

Does it inhibit the target? How does it bind?

Does it work in cells?

Does it have a chance of working in humans?

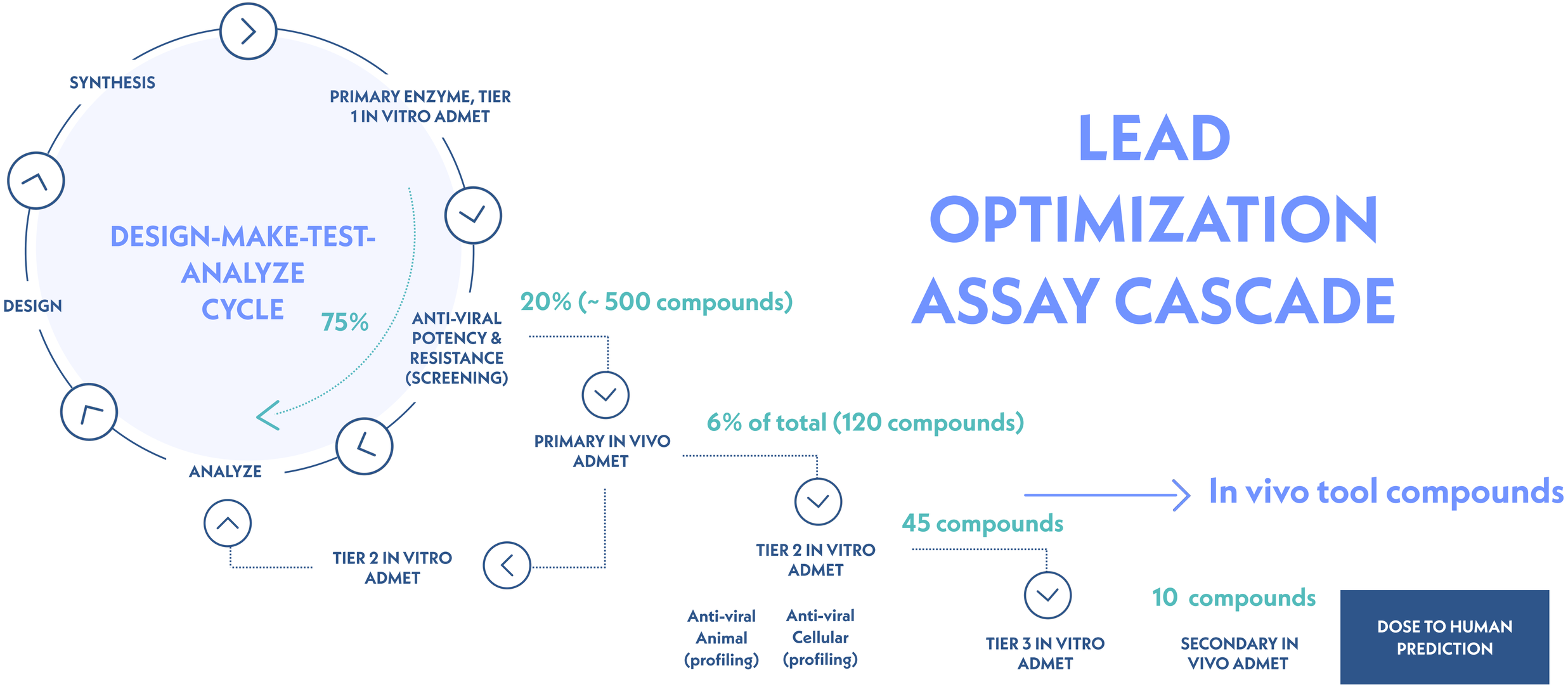
Does it kill the virus in cells?

Could it cause bad side effects?

Can oral dosing deliver sufficient drug?

Does it actually work against the disease?

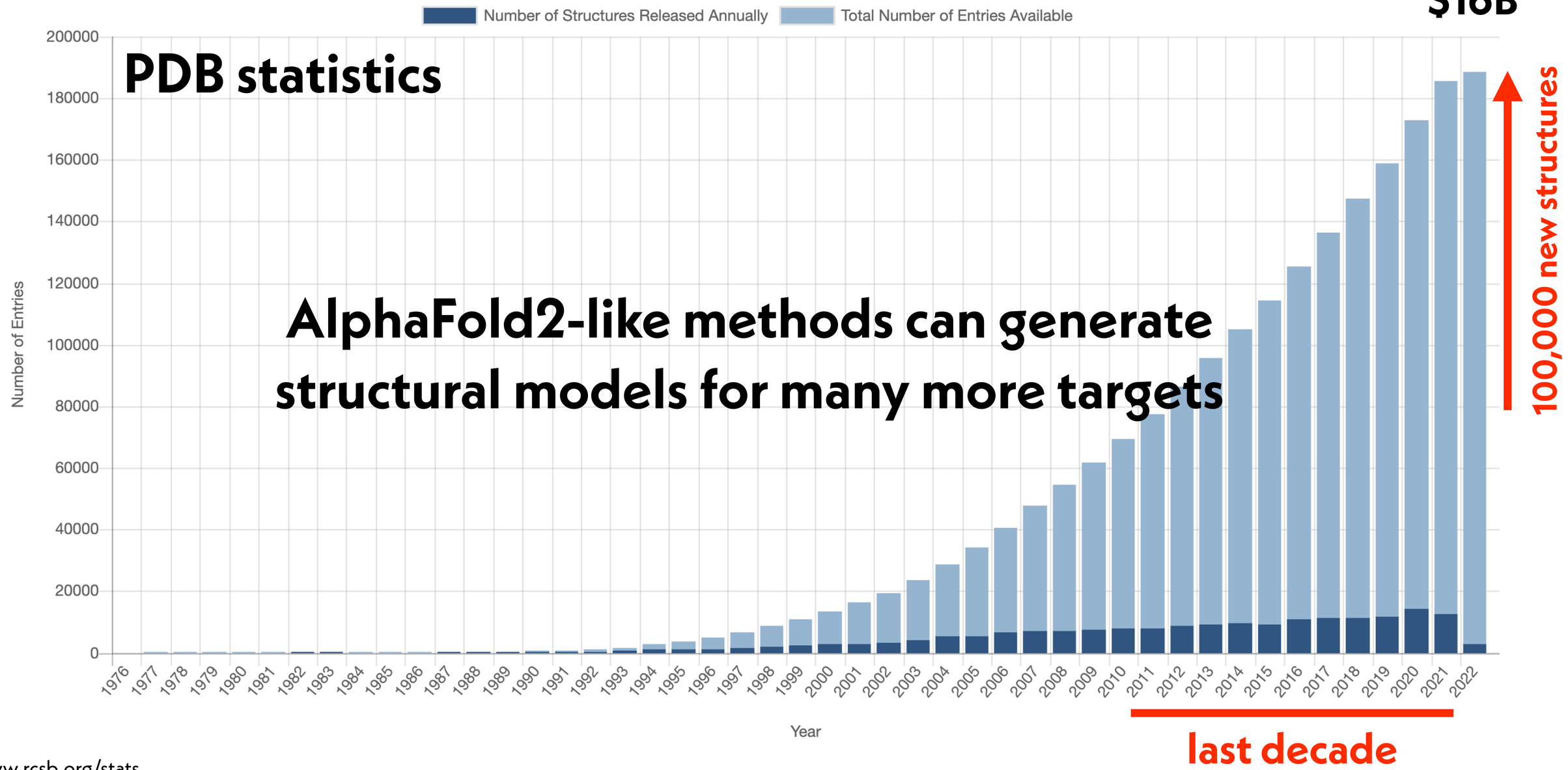
MUCH OF THE TIME IS SPENT IN PREDICTING COMPOUNDS THAT WILL IMPROVE OR MAINTAIN POTENCY



LEAD OPTIMIZATION ASSAY CASCADE

STRUCTURAL DATA IS NOW AN ABUNDANT RESOURCE FOR DRUG DISCOVERY

$16B

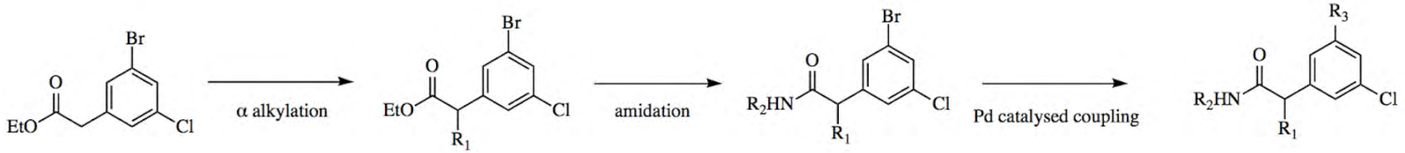


AlphaFold2-like methods can generate structural models for many more targets

http://www.rcsb.org/stats

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

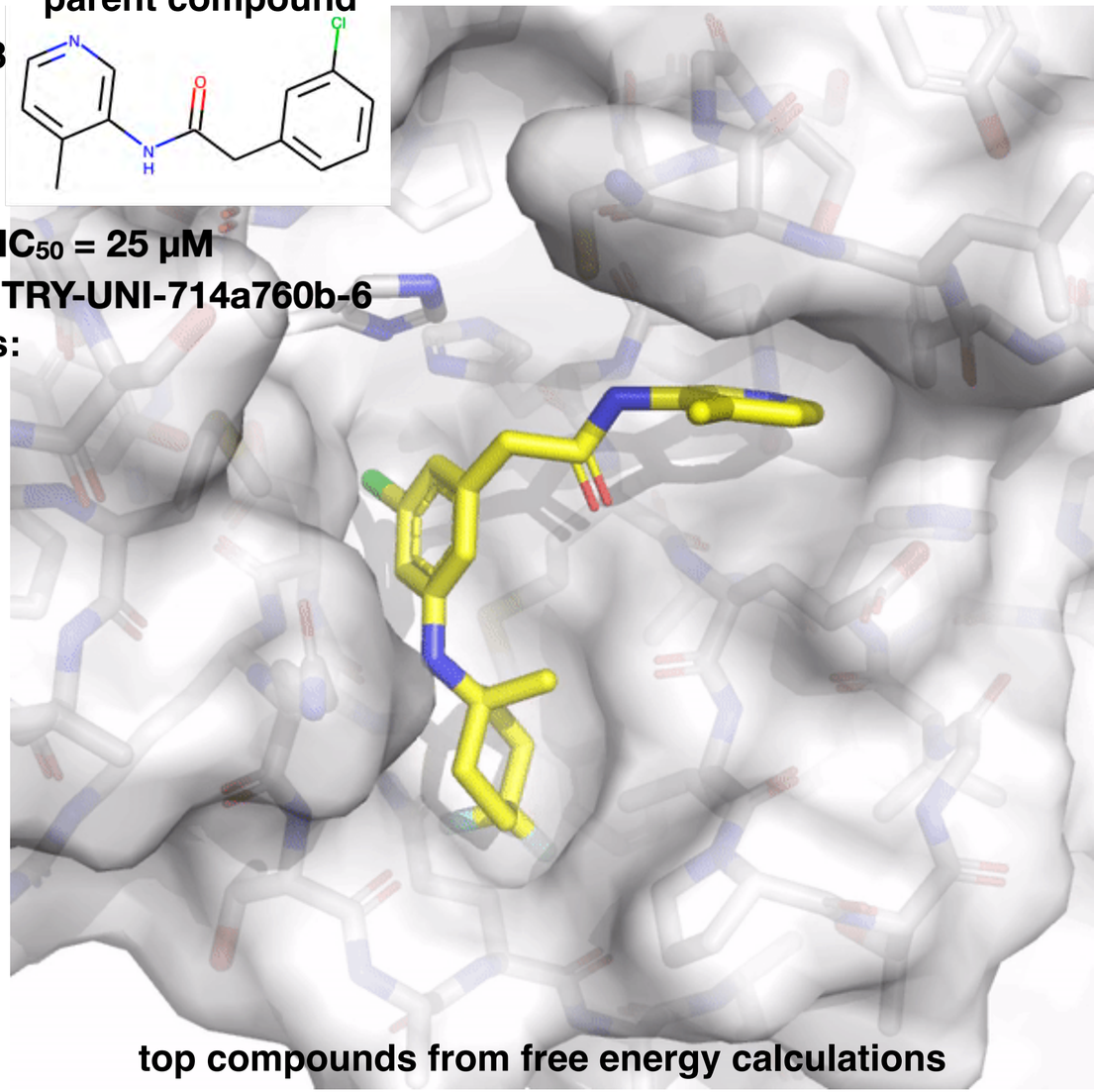
Can we engage S4 from this 5,000-compound virtual synthetic library varying R3



parent compound

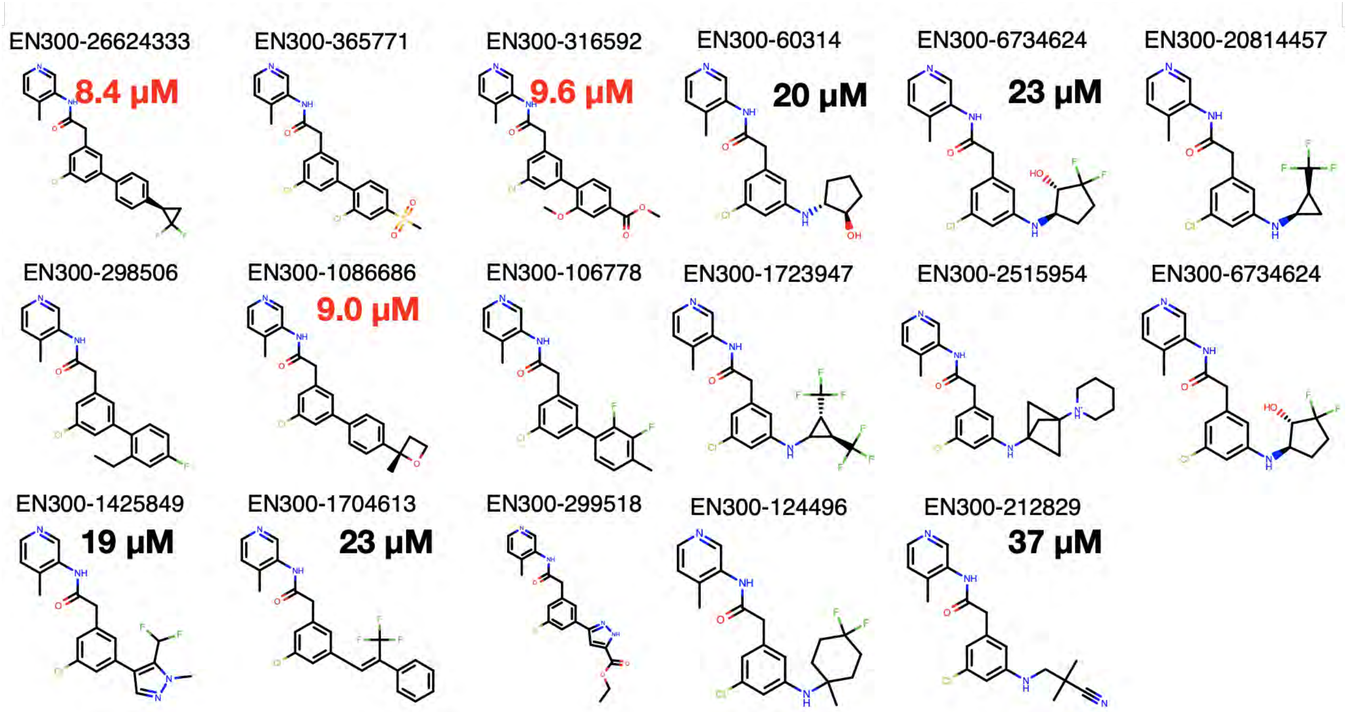
IC_{50} = 25 μM

TRY-UNI-714a760b-6



top compounds from free energy calculations

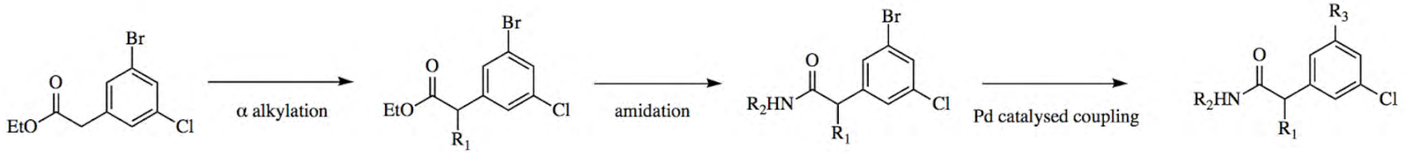
Top free energy calculation compounds and experimental affinity measurements:



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

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

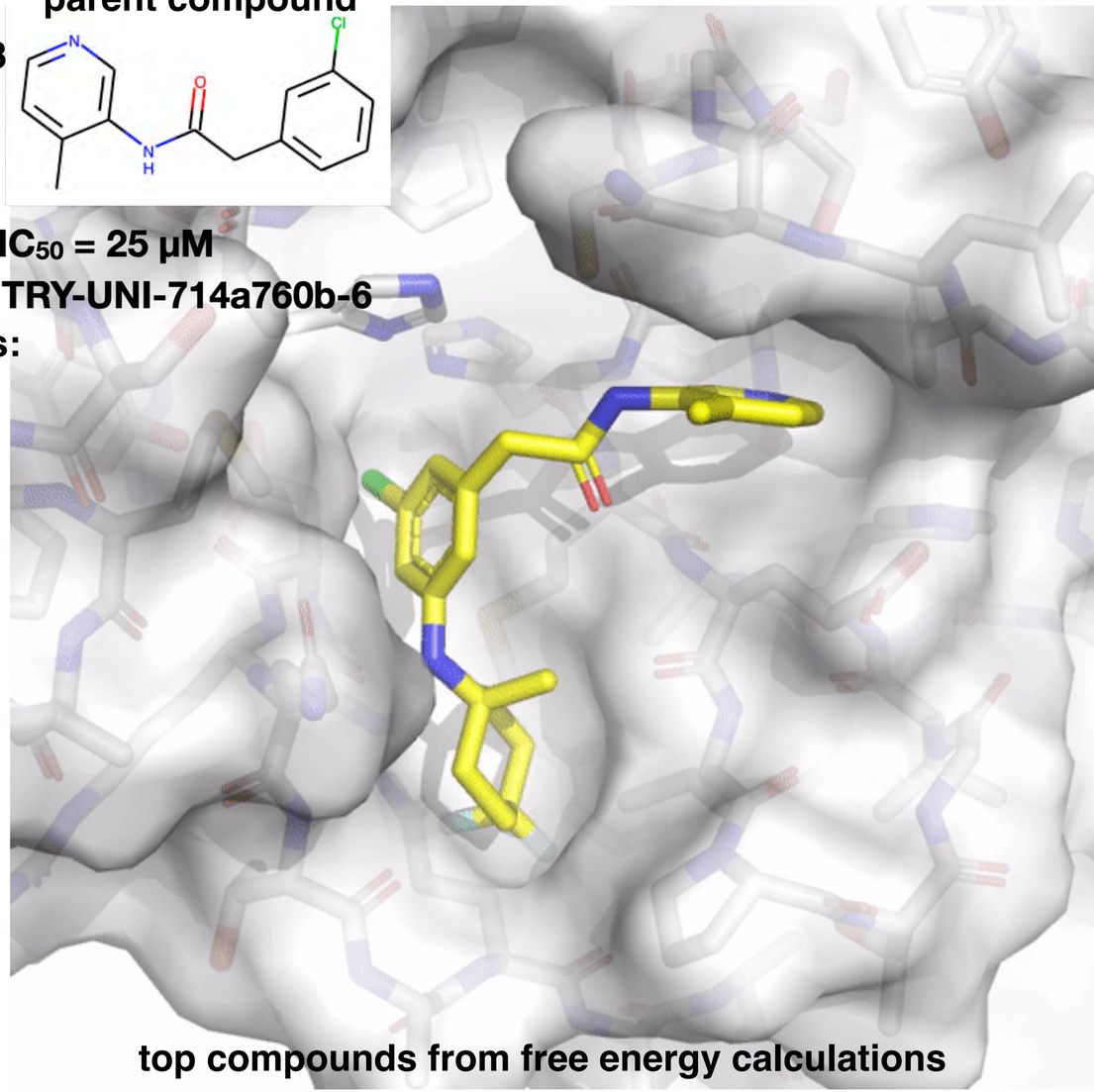
Can we engage S4 from this 5,000-compound virtual synthetic library varying R3



parent compound

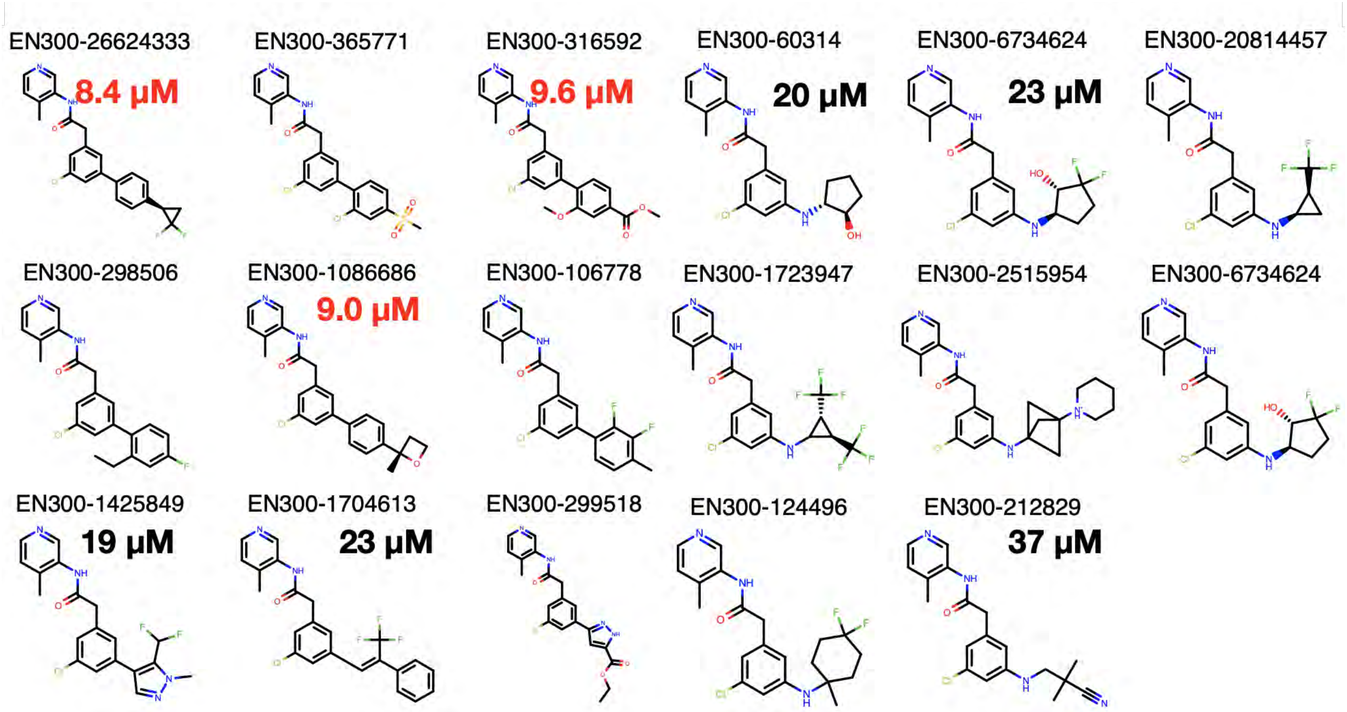
IC_{50} = 25 μM

TRY-UNI-714a760b-6



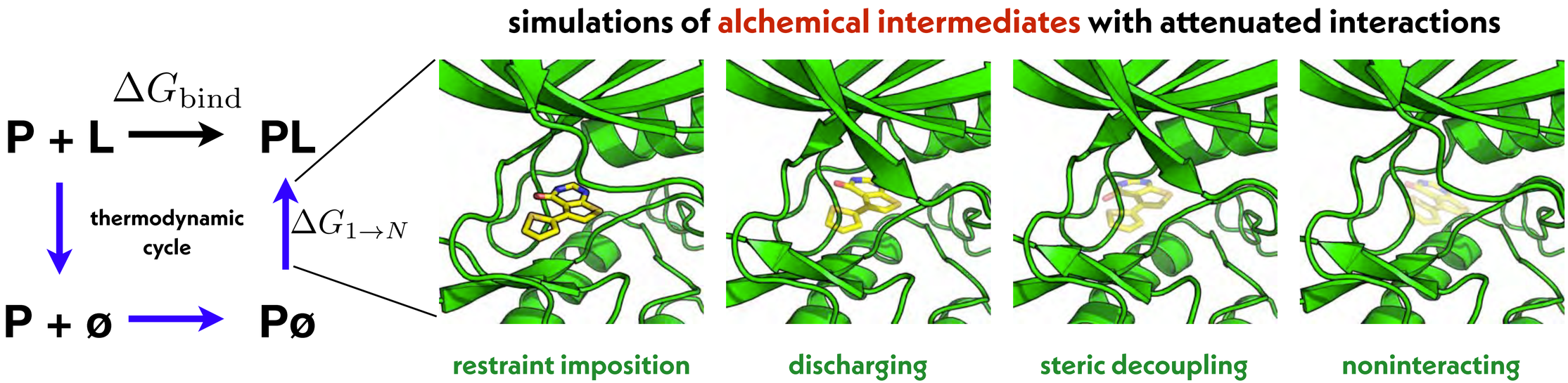
top compounds from free energy calculations

Top free energy calculation compounds and experimental affinity measurements:



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

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



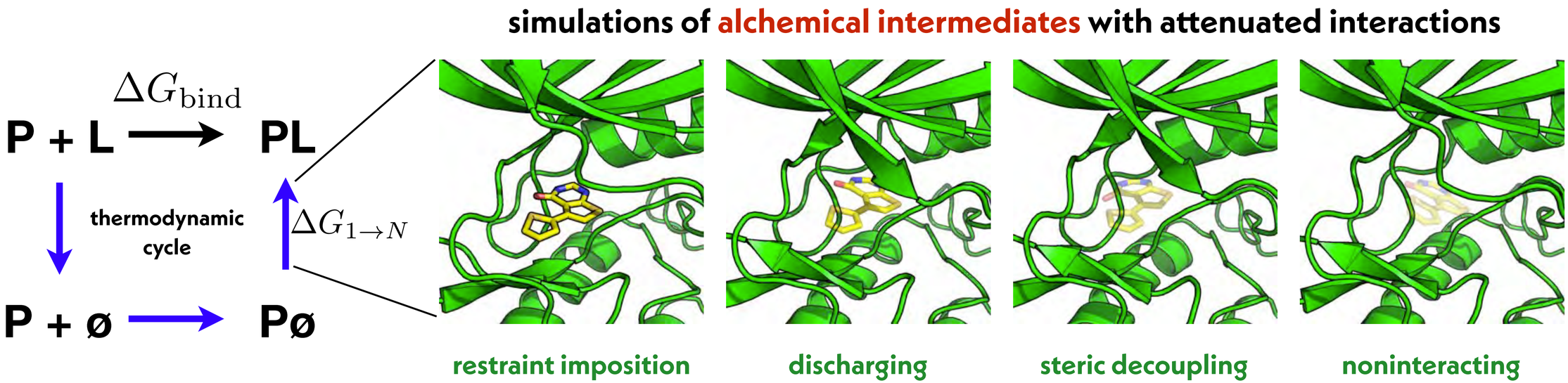
Includes all contributions from enthalpy and entropy of binding to a flexible receptor

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

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

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

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



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

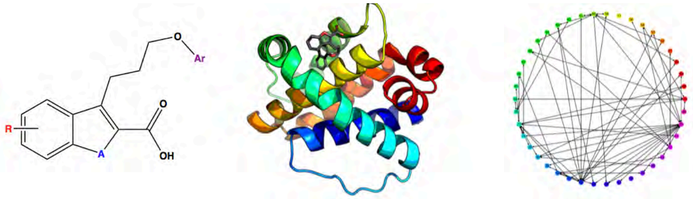
$$\Delta G_{0\to 1} = -k_B T \ln \frac{Z_1}{Z_0} = -k_B T \ln \frac{Z_{\lambda_2}}{Z_{\lambda_1}} \frac{Z_{\lambda_3}}{Z_{\lambda_2}} \cdots \frac{Z_{\lambda_N}}{Z_{\lambda_{N-1}}}$$

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

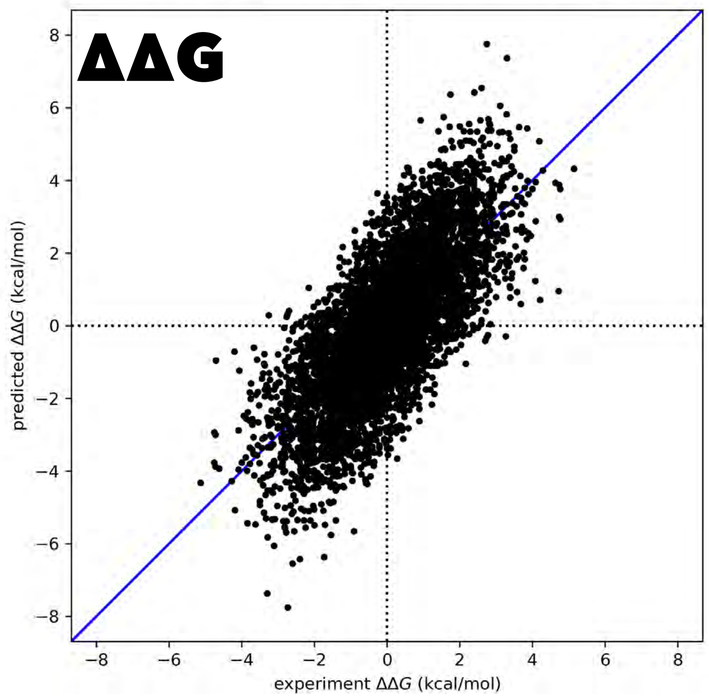
Pioneering work from many: McCammon, van Gunsteren, Kollman, Jorgensen, Chipot, Roux, Boresch, Fujitani, Pande, Shirts, Swope, Christ, Mobley, Schrödinger, and many more

CURRENT ACCURACIES ARE SUFFICIENT TO ACCELERATE DISCOVERY, BUT HOW CAN WE GO FURTHER?

RELATIVE



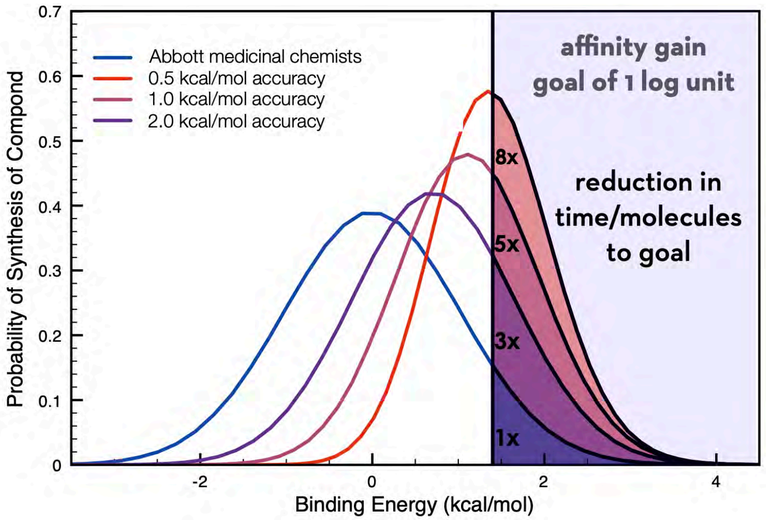
all within-target pairs ΔΔG (N = 5620)
RMSE: OPLS 1.37 [95%: 1.34, 1.39] kcal/mol
MUE : OPLS 1.09 [95%: 1.07, 1.11] kcal/mol



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

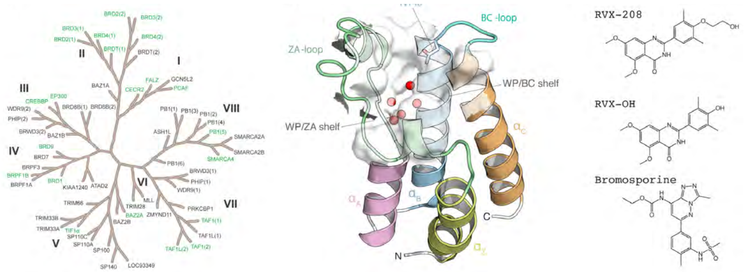
ΔΔG RMSE ~ 1.4 kcal/mol for well-behaved* proteins/chemistries:

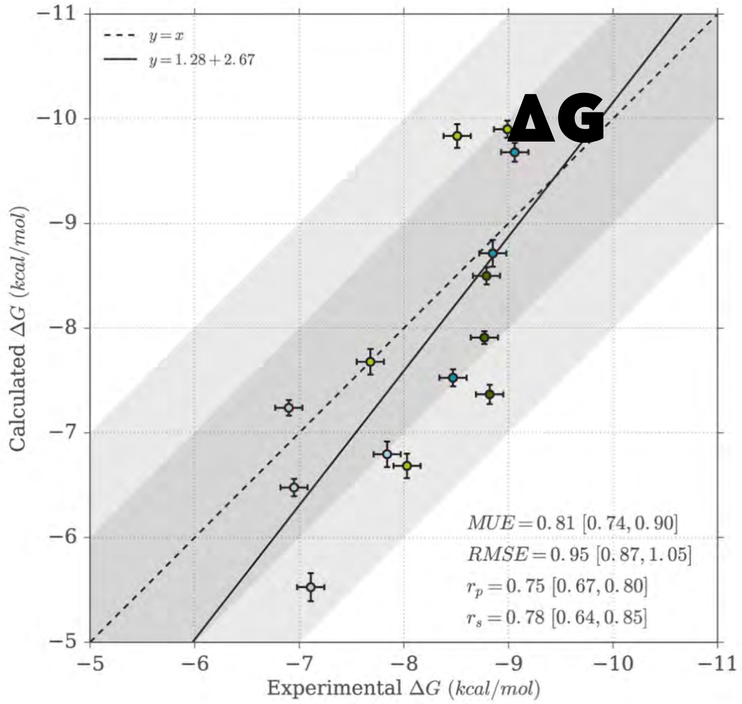
3-5x reduction in molecules synthesized



***best-case scenarios!**

ABSOLUTE



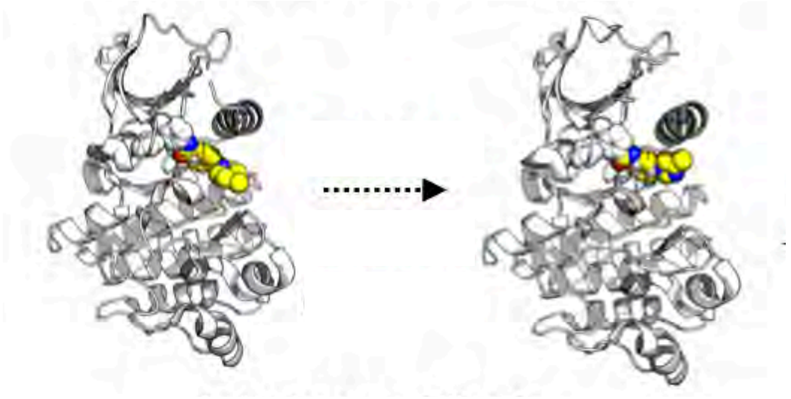


Aldeghi et al. JACS 139:946, 2017.
https://doi.org/10.1021/jacs.6b11467

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

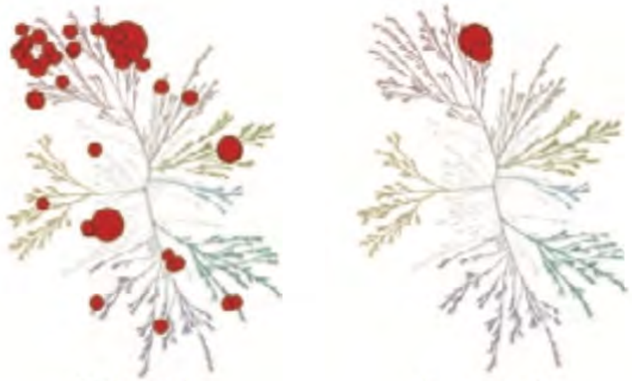
driving affinity / potency

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



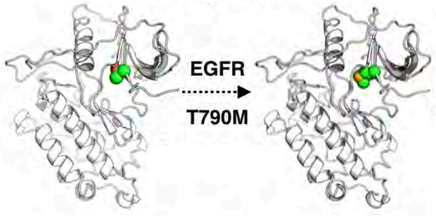
driving selectivity

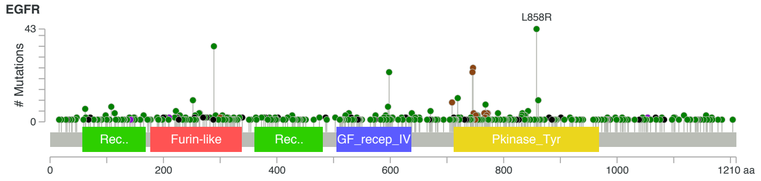
Moraca, Negri, de Olivera, Abel JCIM 2019
https://doi.org/10.1021/acs.jcim.9b00106
Aldeghi et al. JACS 139:946, 2017.
https://doi.org/10.1021/jacs.6b11467



predicting clinical drug resistance/sensitivity

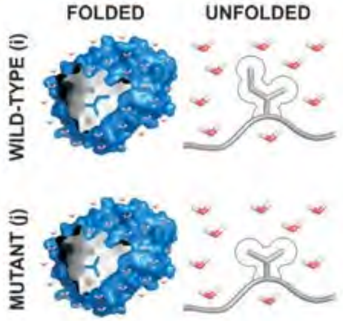
Hauser, Negron, Albanese, Ray, Steinbrecher, Abel, Chodera, Wang.
Communications Biology 1:70, 2018
https://doi.org/10.1038/s42003-018-0075-x
Aldeghi, Gapsys, de Groot. ACS Central Science 4:1708, 2018
https://doi.org/10.1021/acscentsci.8b00717

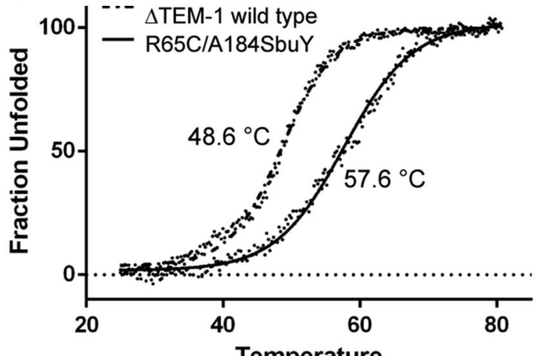




optimizing thermostability

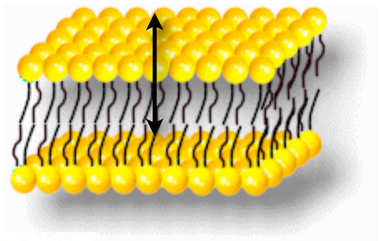
Gapsys, Michielssens, Seeliger, and de Groot. Angew Chem 55:7364, 2016
https://doi.org/10.1002/anie.201510054



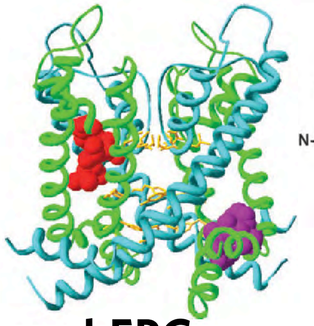


...AND HOLD THE POTENTIAL FOR EVEN BROADER APPLICABILITY AS MORE STRUCTURAL DATA EMERGES

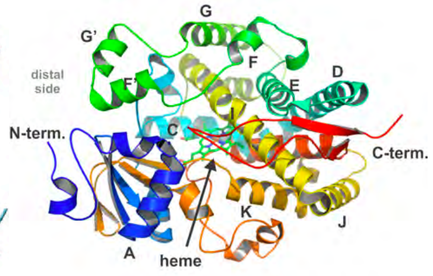
partition coefficients (logP, logD) and permeabilities



structure-enabled ADME/Tox targets

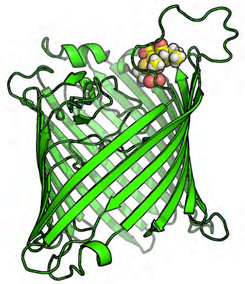


hERG

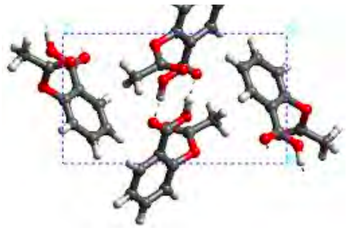


CYP3A4

porin permeation

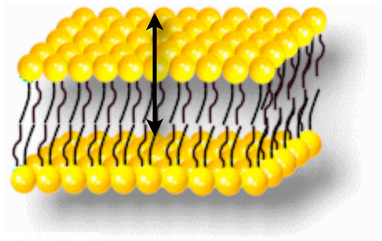


crystal polymorphs, etc.

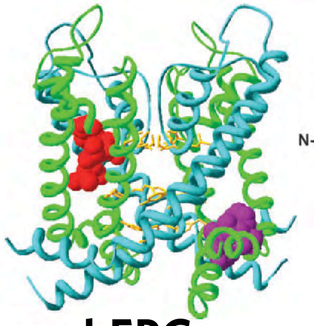


...AND HOLD THE POTENTIAL FOR EVEN BROADER APPLICABILITY AS MORE STRUCTURAL DATA EMERGES

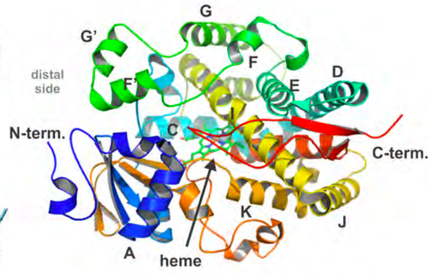
partition coefficients (logP, logD) and permeabilities



structure-enabled ADME/Tox targets

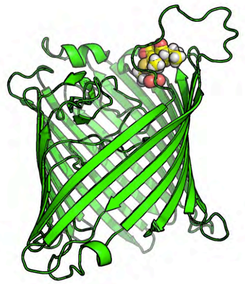


hERG

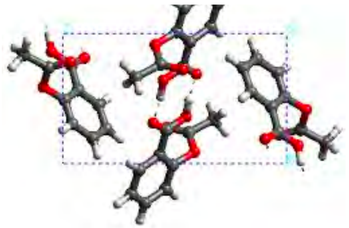


CYP3A4

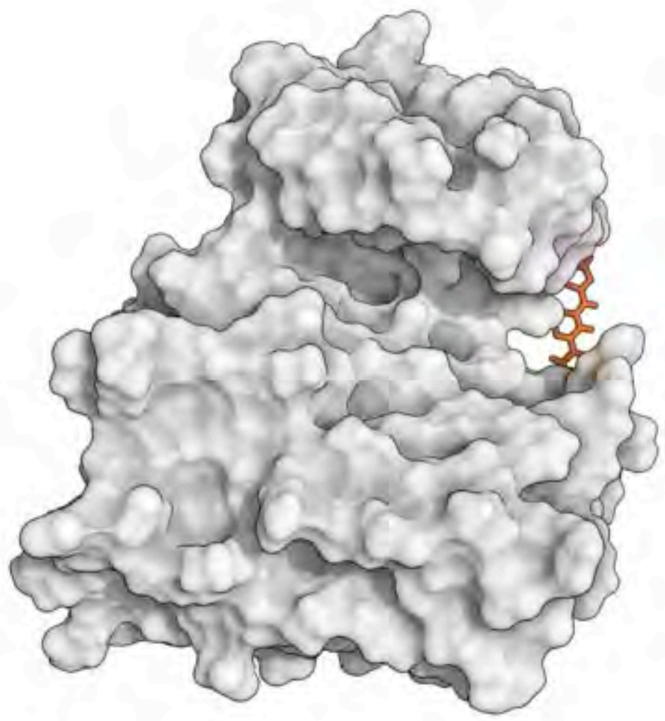
porin permeation



crystal polymorphs, etc.

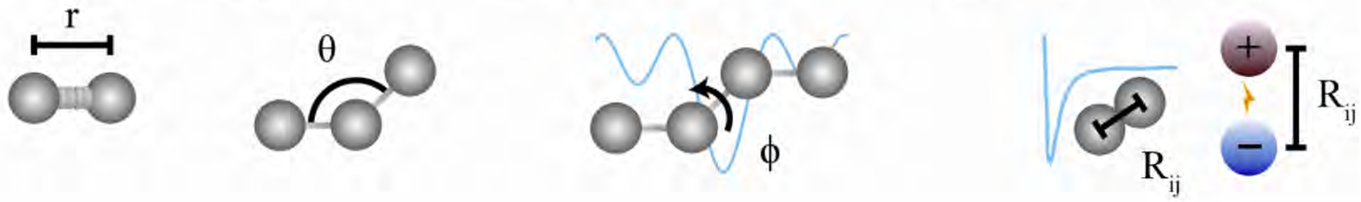


FREE ENERGY CALCULATIONS (AND MUCH OF COMP CHEM) CURRENTLY RELIES ON MOLECULAR MECHANICS FORCE FIELDS



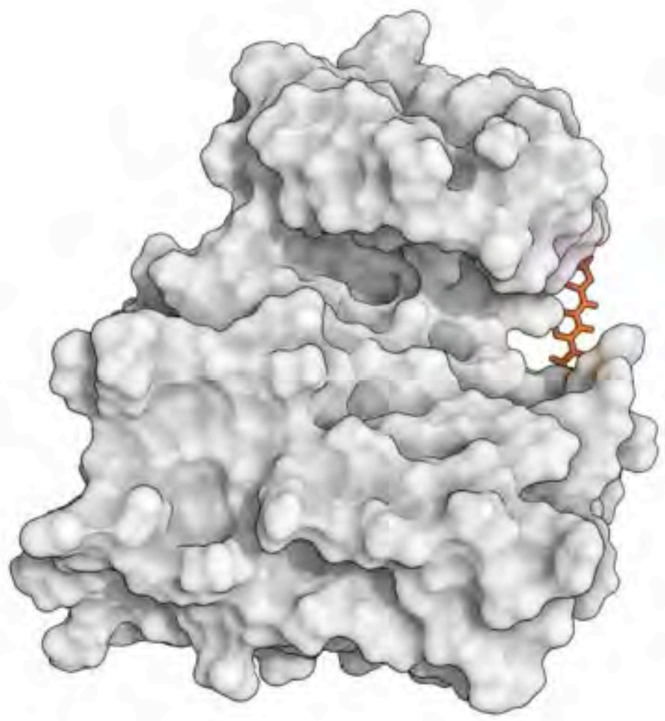
typical class I molecular mechanics force field

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



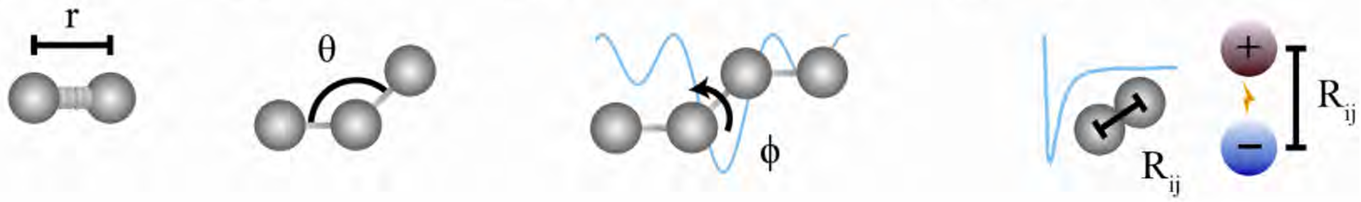
Shan, Kim, Eastwood, Dror, Seeliger, Shaw. JACS 133:9181, 2011
Durrant, McCammon. Molecular dynamics simulations and drug discovery. BMC Biology, 2011

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



typical class I molecular mechanics force field

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



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

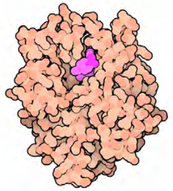
FORCE FIELDS HAVE TRADITIONALLY BEEN HEROIC PRODUCTS OF HUMAN EFFORT

experimental data
quantum chemistry
keen chemical intuition

heroic effort by graduate students and postdocs

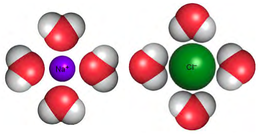
a parameter set we desperately hope someone actually uses

FORCE FIELDS HAVE TRADITIONALLY BEEN HEROIC PRODUCTS OF HUMAN EFFORT



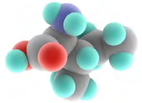
proteins

post-translational modifications

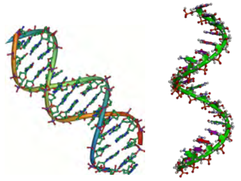


water

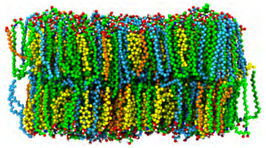
ions



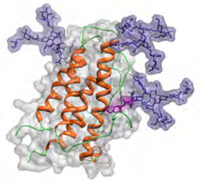
small molecules



nucleic acids



lipids



carbohydrates

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.

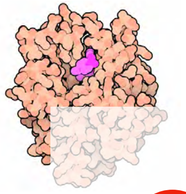
M. Zgarbova; M. Otyepka; J. Sponer; A. Mladek; P. Banas; T. E. Cheatham; P. Jurecka. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.*, **2011**, *7*, 2886–2902.

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.

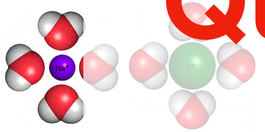
FORCE FIELDS HAVE TRADITIONALLY BEEN HEROIC PRODUCTS OF HUMAN EFFORT



proteins

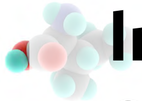
post-translational modifications

Quickly adds up to >100 human-years



water

ions



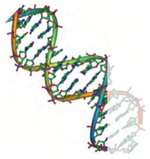
non-natural amino acids

Intended to be compatible, but not co-parameterized

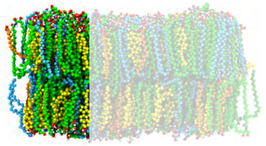
Significant effort is required to extend to new areas

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

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

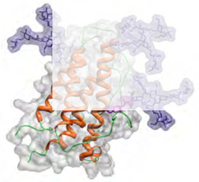


nucleic acids



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; [illegible] 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-mediated 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 Divalent 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. Comp. Chem.*, **2004**, *25*, 1157–1174.

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

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

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

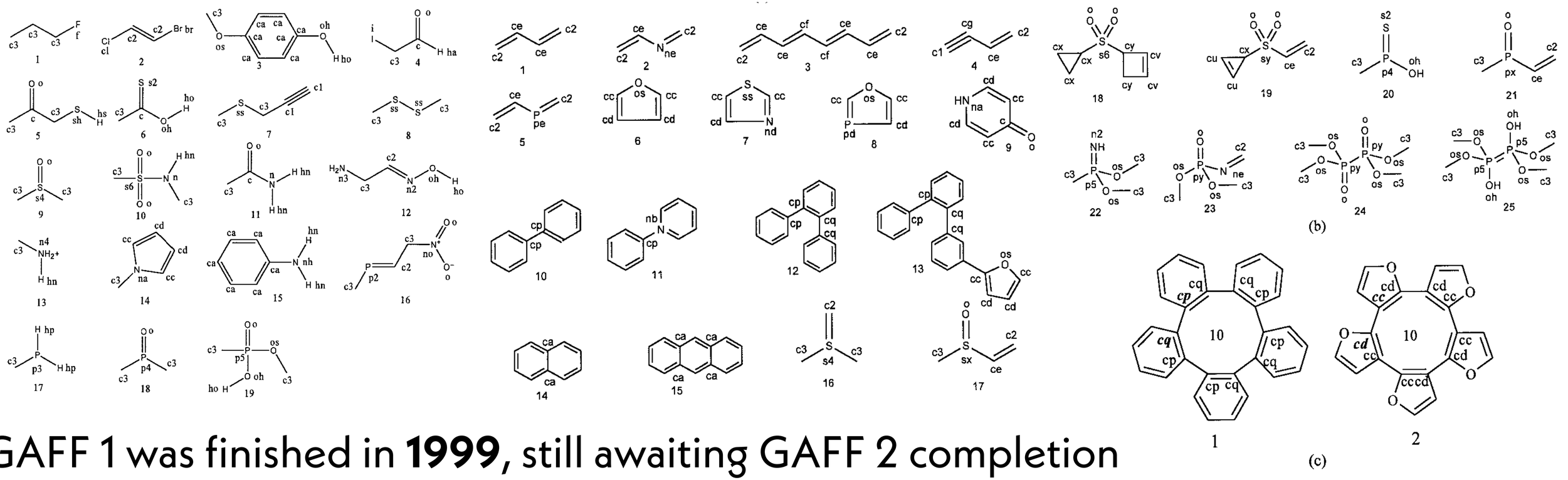
Å. Skjevik; B. D. Madej; R. C. Walker; K. Teigen. Lipid11: A modular framework for lipid simulations using amber. *J. Phys. Chem. B*, **2012**, *116*, 11124–11136.

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

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

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

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



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

Extension to new chemical space is **nontrivial**

Parameter fitting code was **never released**

Atom types have introduced numerous **errors**

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

CAN WE MAKE BUILDING BIMOLECULAR FORCE FIELDS AS EASY AS TRAINING A MACHINE LEARNING MODEL?

training a neural network

```
import tensorflow as tf
mnist = tf.keras.datasets.mnist

(x_train, y_train),(x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([
  tf.keras.layers.Flatten(input_shape=(28, 28)),
  tf.keras.layers.Dense(128, activation='relu'),
  tf.keras.layers.Dropout(0.2),
  tf.keras.layers.Dense(10, activation='softmax')
])

model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

model.fit(x_train, y_train, epochs=5)
model.evaluate(x_test, y_test)
```

Run code now Try in Google's interactive notebook

https://www.tensorflow.org/overview

import your tools

grab a standard, curated dataset

define a novel model architecture

declare your objectives in training it

fit it

use it

CAN WE MAKE BUILDING BIMOLECULAR FORCE FIELDS AS EASY AS TRAINING A MACHINE LEARNING MODEL?

training a neural network

```python
import tensorflow as tf
mnist = tf.keras.datasets.mnist

(x_train, y_train),(x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0

model = tf.keras.models.Sequential([
  tf.keras.layers.Flatten(input_shape=(28, 28)),
  tf.keras.layers.Dense(128, activation='relu'),
  tf.keras.layers.Dropout(0.2),
  tf.keras.layers.Dense(10, activation='softmax')
])

model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

model.fit(x_train, y_train, epochs=5)
model.evaluate(x_test, y_test)
```

Run code now Try in Google's interactive notebook

https://www.tensorflow.org/overview

fitting a force field

```python
import openforcefield as off
training_data, benchmark_data = off.datasets.load('2019-Q1')

force_field_model = off.models.ForceFieldModel([
    off.models.forces.HarmonicBond(),
    off.models.forces.HarmonicAngle(),
    off.models.forces.PeriodicTorsion(max_order=6),
    off.models.forces.LennardJones(),
    off.models.forces.BondChargeCorrections(),
])

model.compile(optimizer='L-BFGS',
     loss='error-weighted',
     metrics=['accuracy'])

model.fit(training_data)

model.evaluate(test_data)
```

Run code now Try in Google's interactive notebook

open
forcefield

An open and collaborative approach to better force fields

OPEN SOURCE

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

OPEN SCIENCE

Scientific reports as blog posts, webinars and preprints

OPEN DATA

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

NEWS | TUTORIALS | ROADMAP

http://openforcefield.org

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
MolSSI QCArchive quantum chemical data: http://qcarchive.molssi.org

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

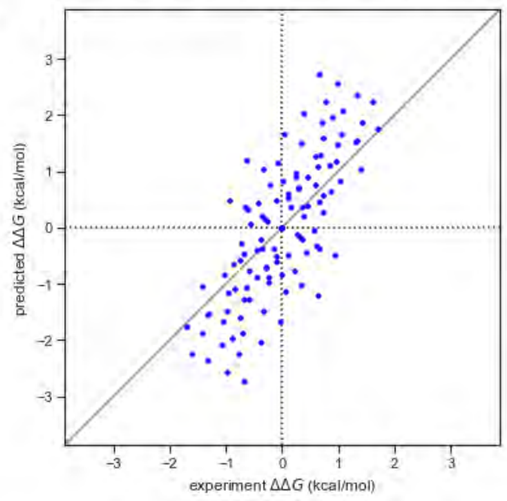
Open science: everything we do is free, permissively licensed, and online

http://openforcefield.org

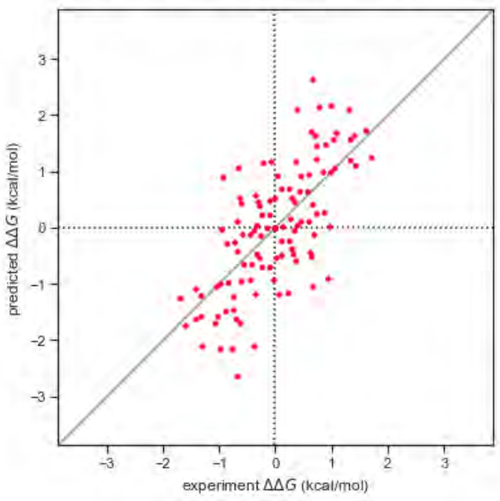
WE'VE MADE RAPID AND SIGNIFICANT PROGRESS IN ACCURACY, BUT WE'RE STILL STICK WITH SLOW GENERATIONS

Open Force Field Initiative →

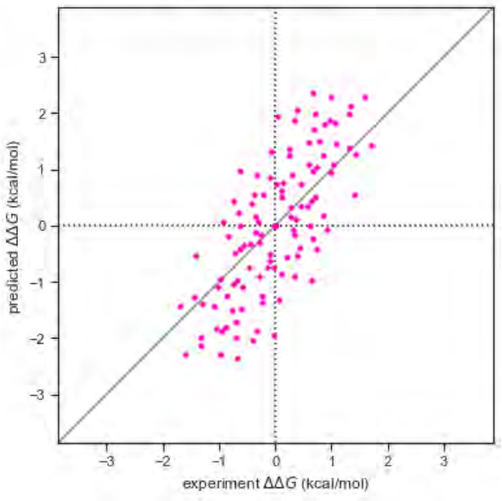
GAFF 1 (1999)



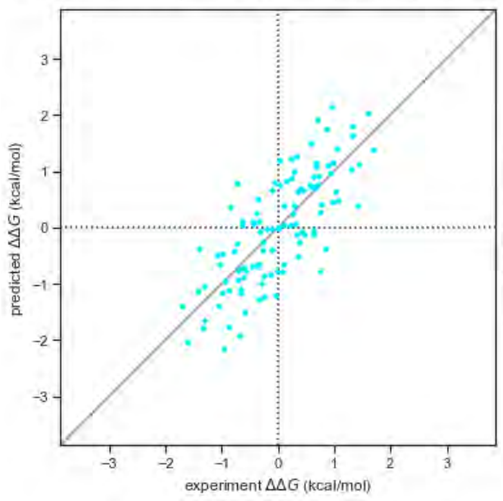
OPLS2.1 (2015)



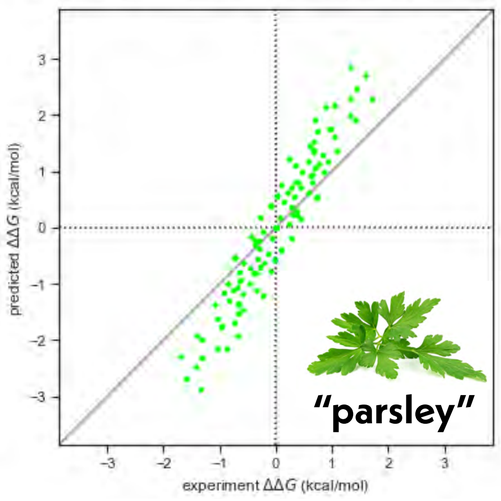
GAFF 2 (2016)



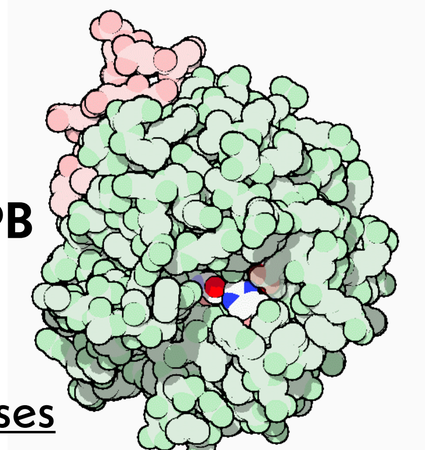
smirnoff99Frosst (2018)



openff 1.0 (2019)



thrombin
PDB101: 1PPB



HANNAH BRUCE MACDONALD
MSKCC

http://github.com/choderalab/perses

DOMINIC RUFA

NEW GENERATIONS OF MACHINE LEARNING MODELS ARE PARTICULARLY WELL-SUITED TO CHEMISTRY

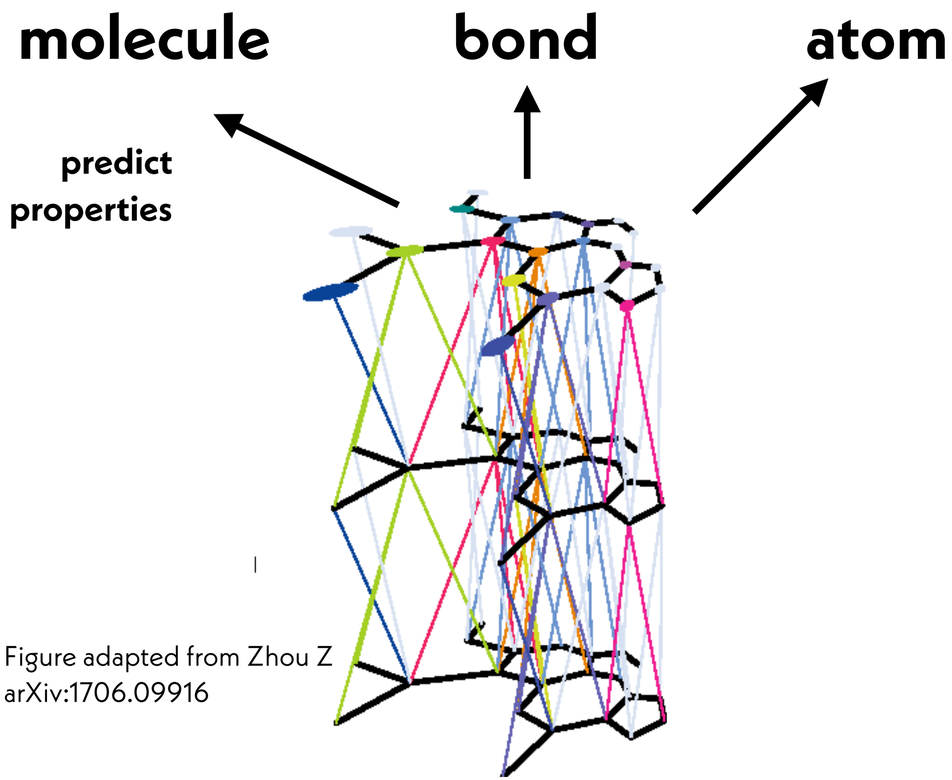


Figure adapted from Zhou Z arXiv:1706.09916

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

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

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

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

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

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

NEW GENERATIONS OF MACHINE LEARNING MODELS ARE PARTICULARLY WELL-SUITED TO CHEMISTRY

molecule **bond** **atom**

predict properties

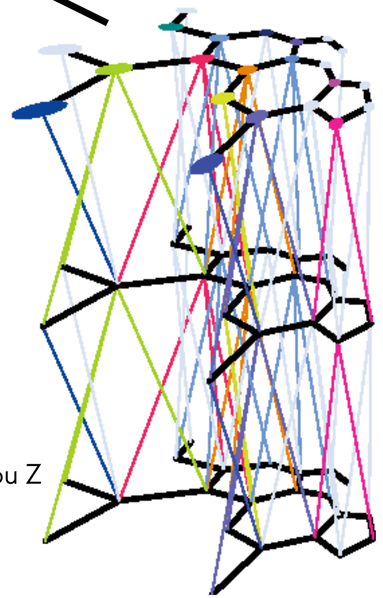
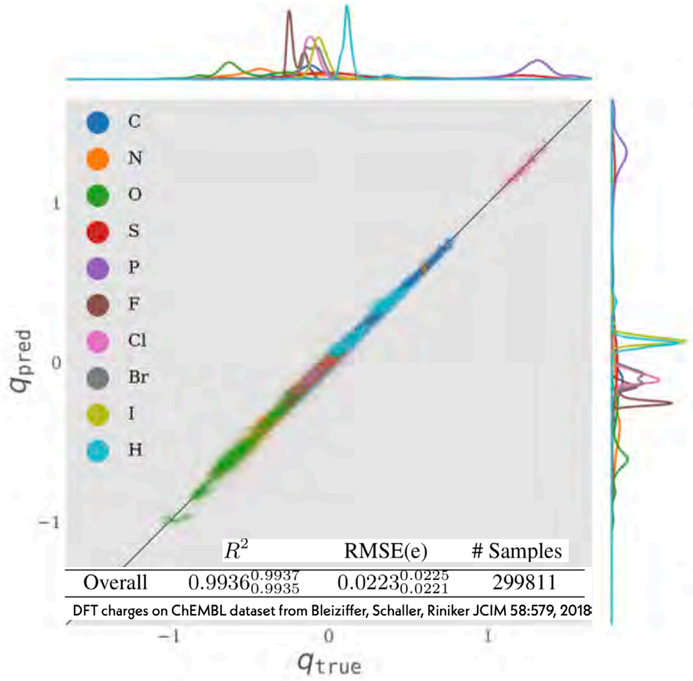


Figure adapted from Zhou Z arXiv:1706.09916

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

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

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



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

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

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

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

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

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

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

Gimlet

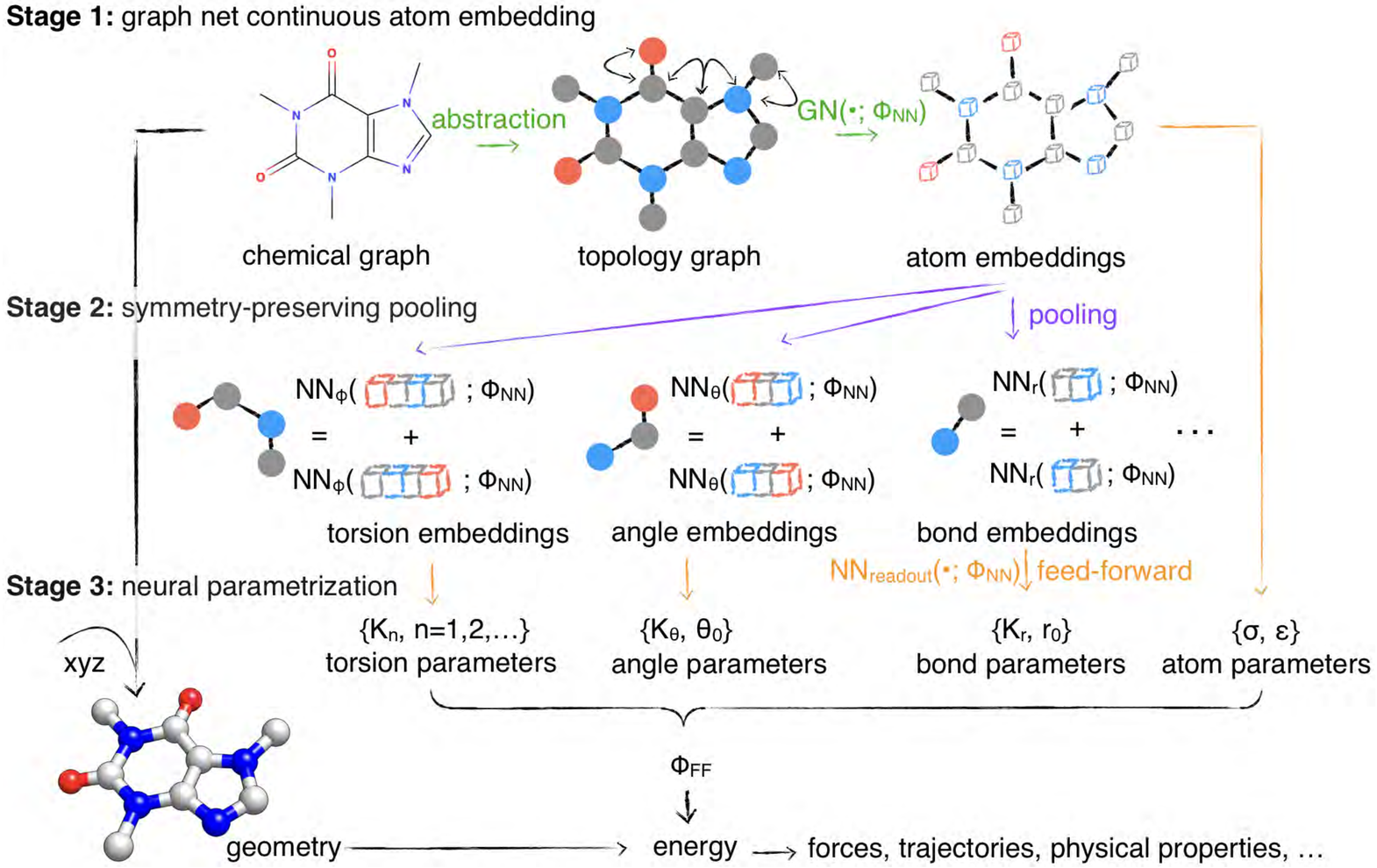
Graph Inference on MoLEcular Topology

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

YUANQING WANG

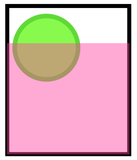
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm

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



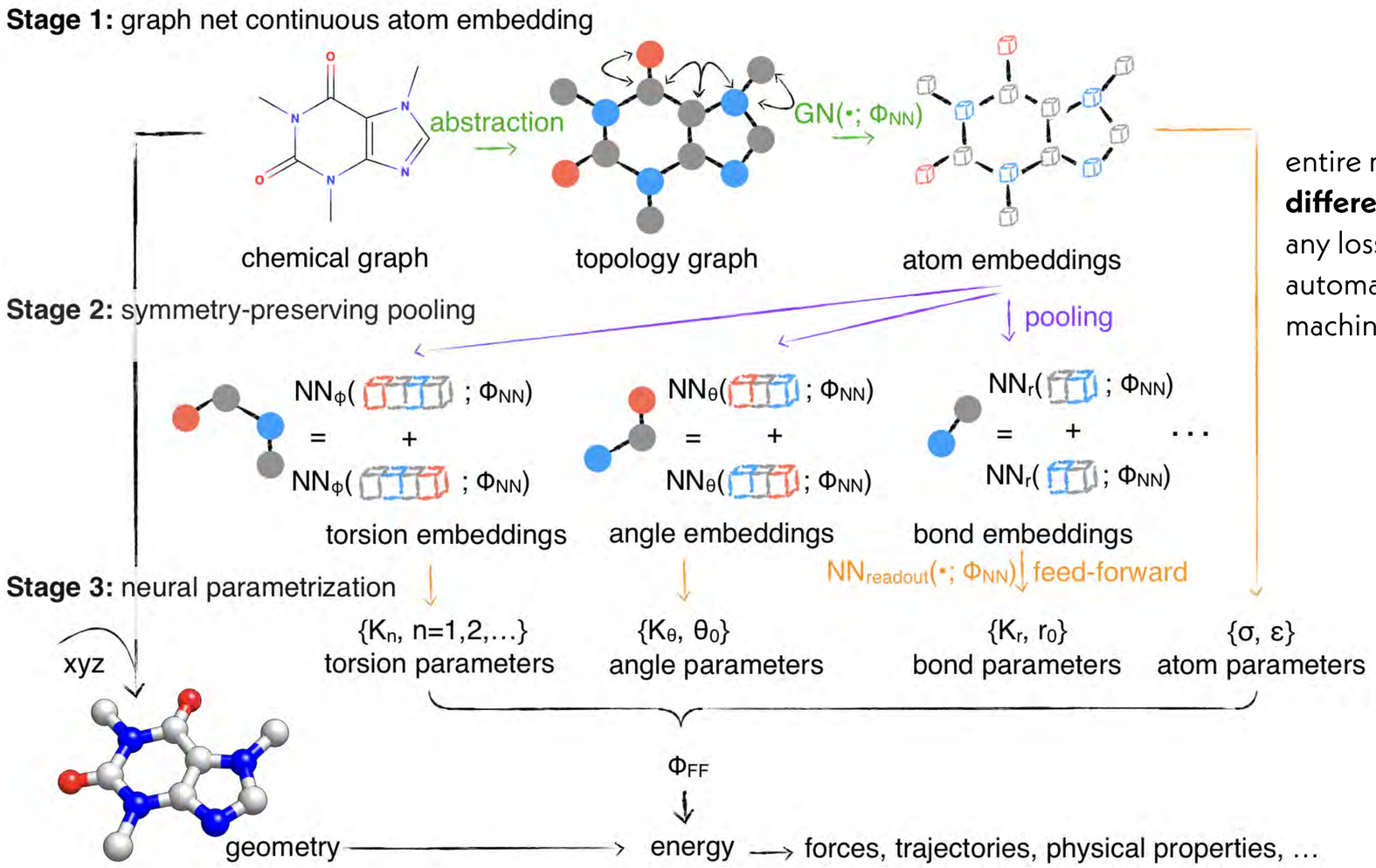
JOSH FASS

YUANQING WANG



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

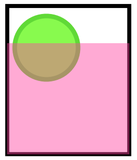
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



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

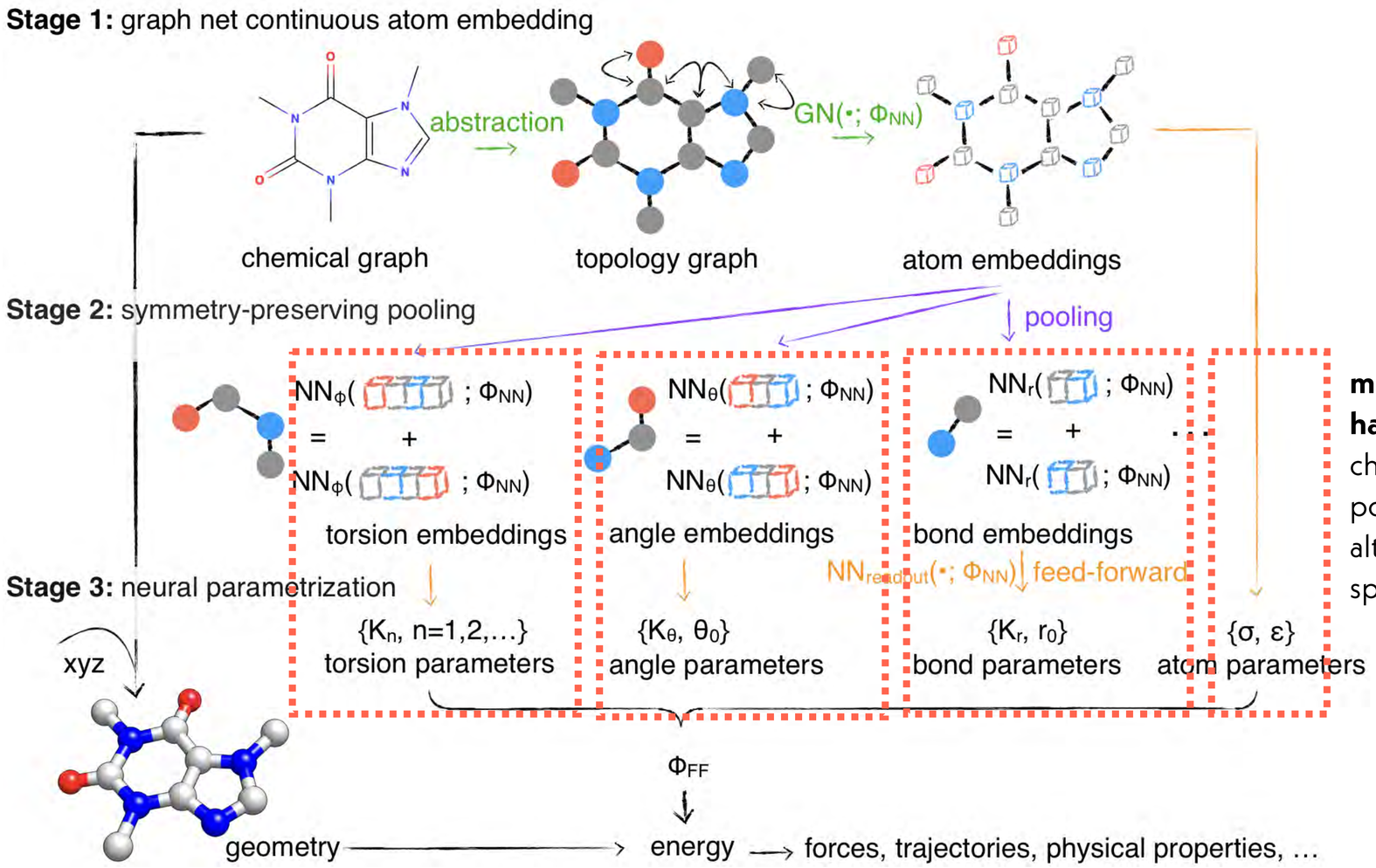
JOSH FASS

YUANQING WANG



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

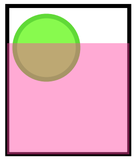
espaloma: **e**xtensible **s**urrogate **p**otential of ***a**b initio* **l**earned and **o**ptimized by **m**essage-passing **a**lgorithm



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

JOSH FASS

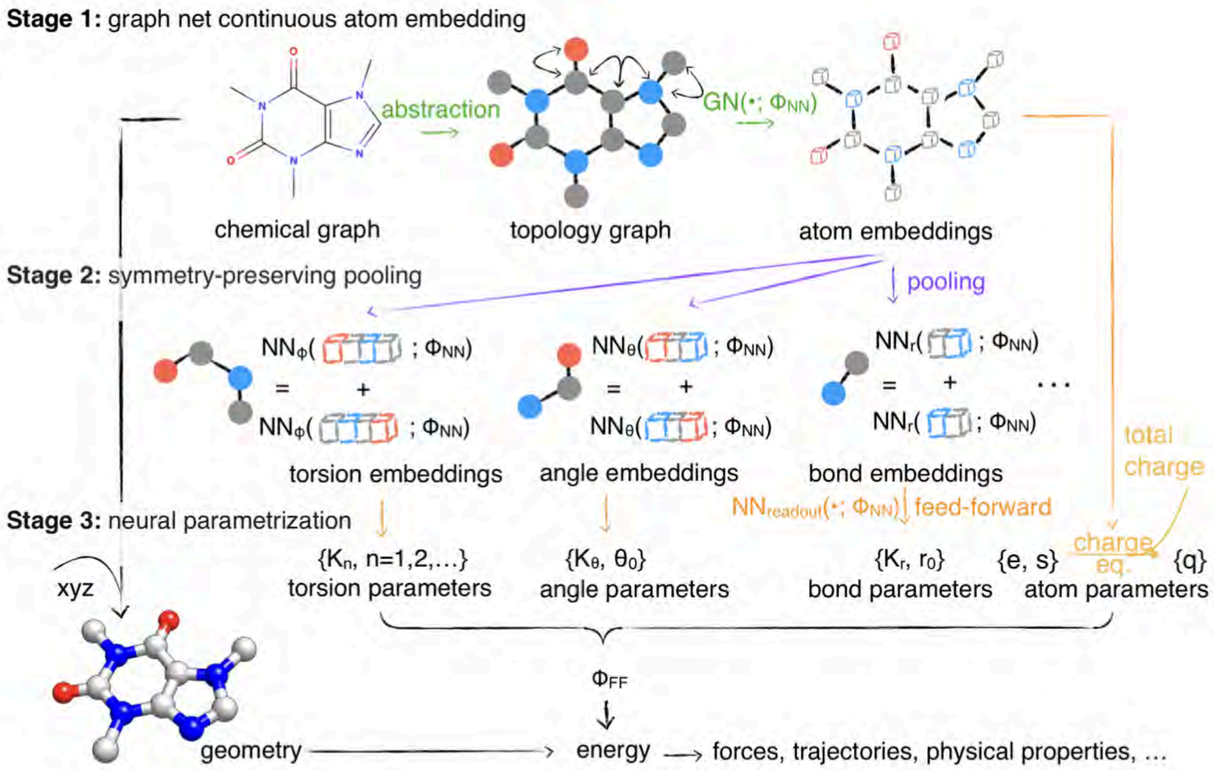
YUANQING WANG



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

ESPALOMA MAKES BUILDING A NEW FORCE FIELD EASY

espaloma architecture



(implemented in pytorch)

http://github.com/choderalab/espaloma

YUANQING WANG

building a new force field

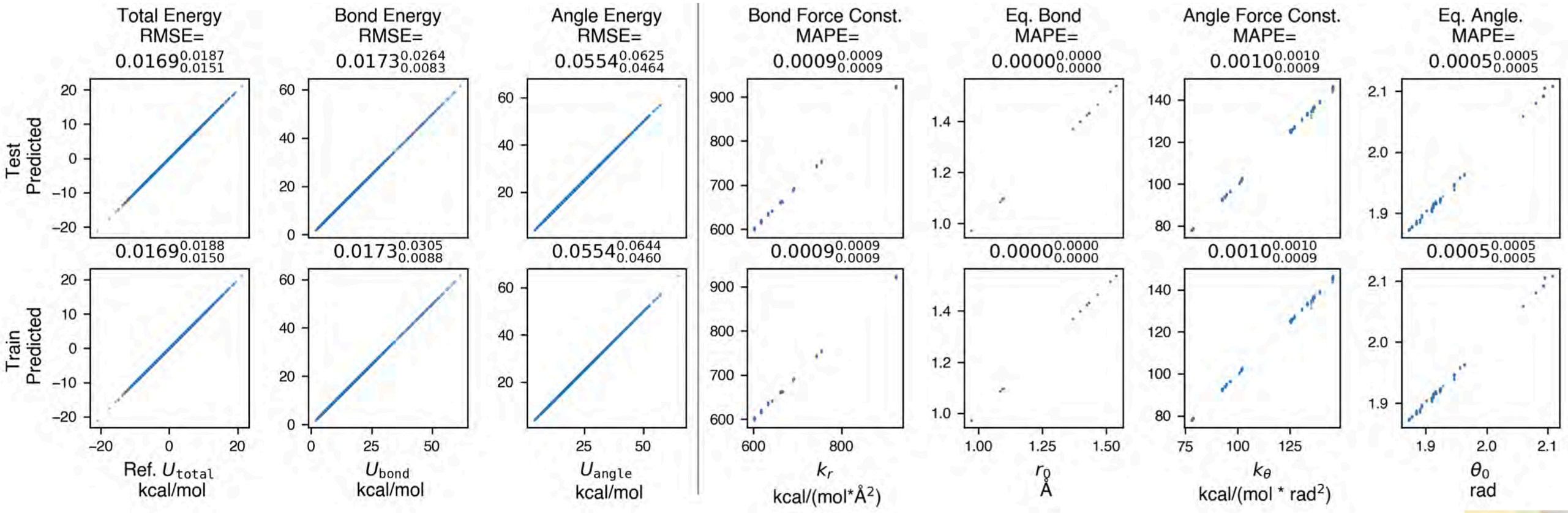
```python
import torch, dgl, espaloma as esp

# 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="nl", # 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 CAN LEARN TO REPRODUCE LEGACY MM FORCE FIELDS WITH LOW RMSE ERROR IN CONFORMATIONAL ENERGIES

conformer energies

force field parameters



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

reference force field: GAFF 1.81 [https://doi.org/10.1002/jcc.20035]
dataset: PhAlkEthOH [https://dx.doi.org/10.1021/acs.jctc.8b00640]

YUANQING WANG

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

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

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

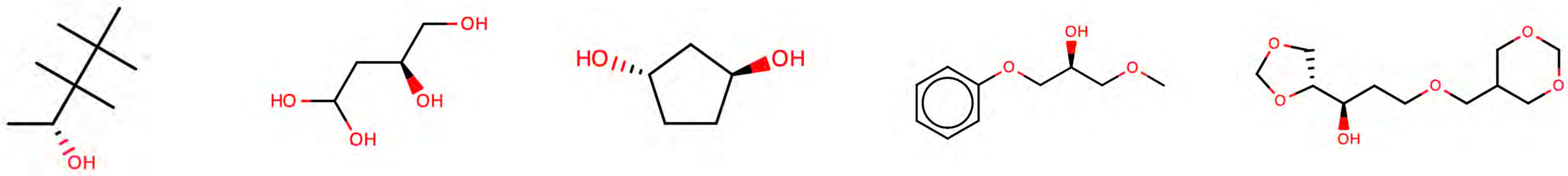
YUANQING WANG

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

(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	

PhAlkEthOh: **Ph**enyls, **Alk**anes, **Eth**ers, and alcohols (**OH**)

(a low-complexity chemical space)



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

PhAlkEthOH doi: https://dx.doi.org/10.1021/acs.jctc.8b00640

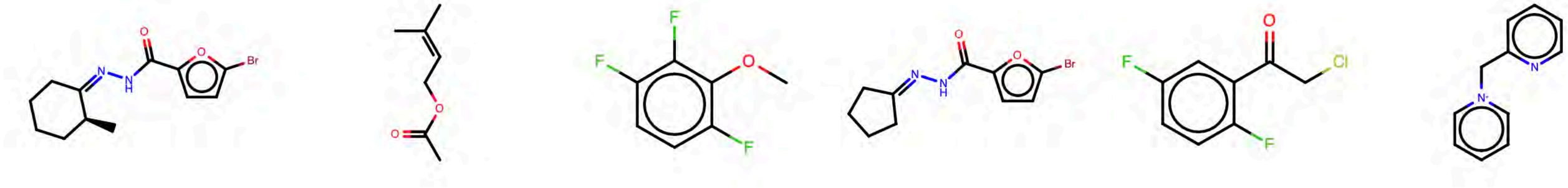
YUANQING WANG

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

(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	

OpenFF Gen2 Optimization set: Diverse druglike fragments challenging for force fields

(a moderate-complexity chemical space)



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

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

OpenFF Gen2 QM data doi: https://doi.org/10.1021/acs.jctc.1c00571

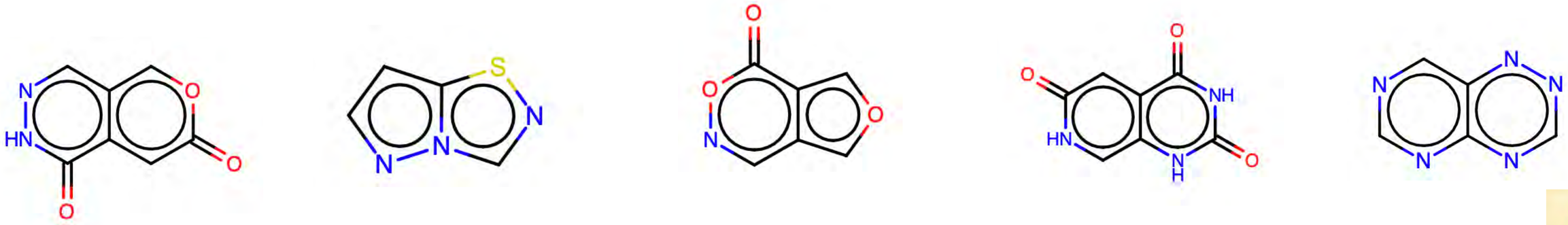
YUANQING WANG

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

(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)	24867	24867	234326	$0.4476^{0.4690}_{0.4273}$	$0.4233^{0.4414}_{0.4053}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	

VEHICLe: Virtual exploratory heterocyclic drug scaffold library

(aromatic bicyclic heterocyclic compounds containing C, N, O, S, H)



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

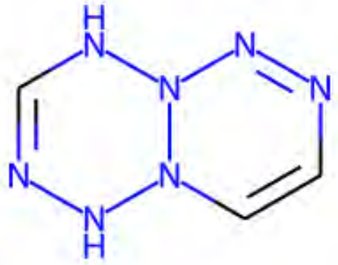
code: http://github.com/choderalab/espaloma

VEHICLe doi: http://doi.org/10.1021/jm801513z

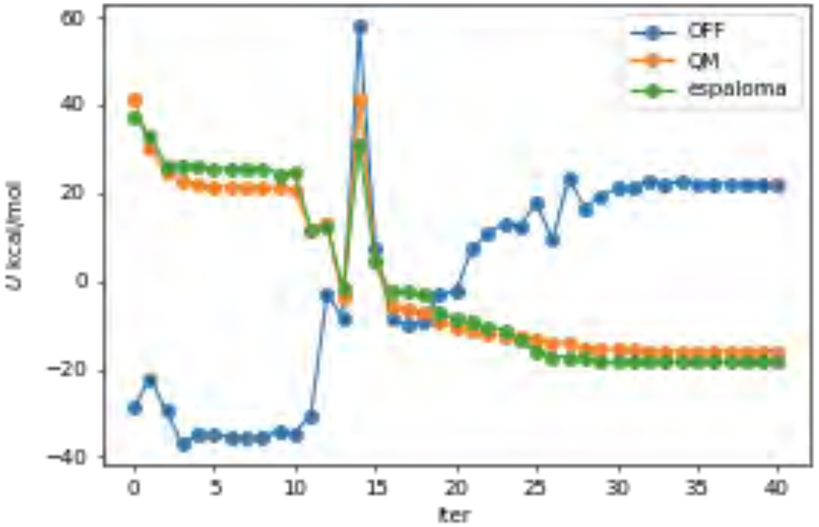
YUANQING WANG

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

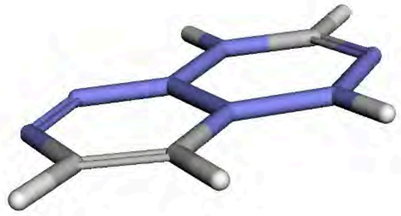
(a) dataset	# mols	# trajs	# snapshots	Espaloma RMSE		Legacy FF RMSE (kcal/mol) (Test molecules)			
				Train	Test	OpenFF 1.2.0	GAFF-1.81	GAFF-2.11	Amber ff14SB
PhAlkEthOH (simple CHO)	7408	12592	244036	$0.8656^{0.9131}_{0.8225}$	$1.1398^{1.2332}_{1.0715}$	$1.6071^{1.6915}_{1.5197}$	$1.7267^{1.7935}_{1.6543}$	$1.7406^{1.8148}_{1.6679}$	
OpenFF Gen2 Optimization (druglike)	792	3977	23748	$0.7413^{0.7920}_{0.6914}$	$0.7600^{0.8805}_{0.6644}$	$2.1768^{2.3388}_{2.0380}$	$2.4274^{2.5207}_{2.3300}$	$2.5386^{2.6640}_{2.4370}$	
VEHICLe (heterocyclic)	24867	24867	234326	$0.4476^{0.4690}_{0.4273}$	$0.4233^{0.4414}_{0.4053}$	$8.0247^{8.2456}_{7.8271}$	$8.0077^{8.2313}_{7.7647}$	$9.4014^{9.6434}_{9.2135}$	



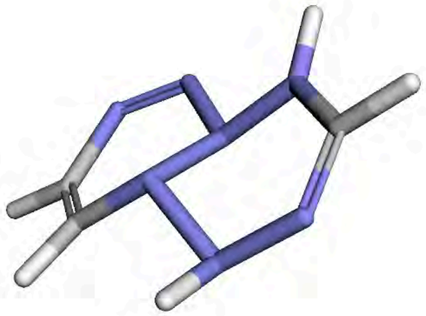
Comparison with QCArchive data



initial



QM minimized



DFT B3LYP-D3(BJ) / DZVP

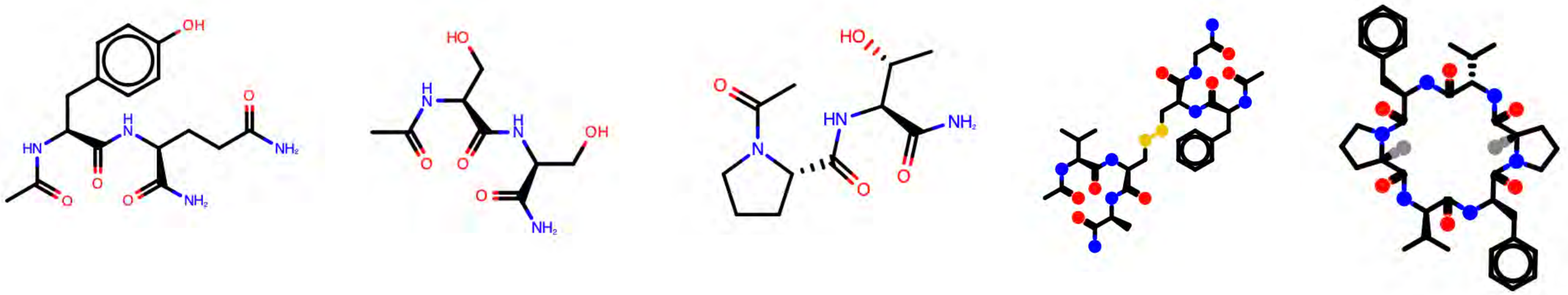
preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

YUANQING WANG

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

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

PepConf: Short peptides, including disulfides and cyclic peptides



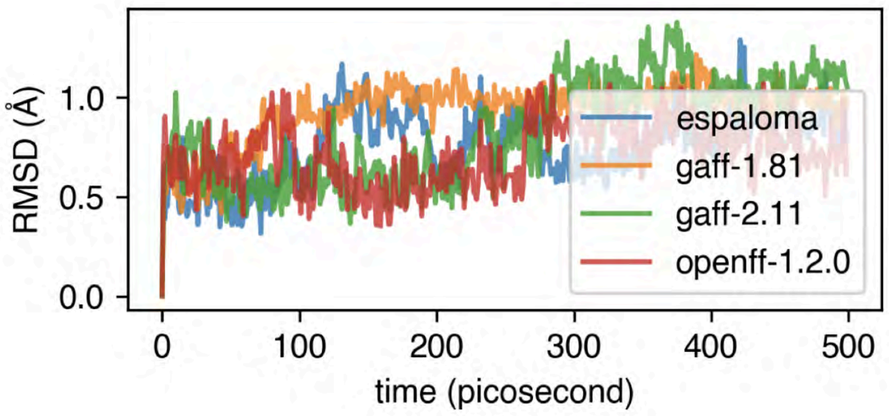
preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

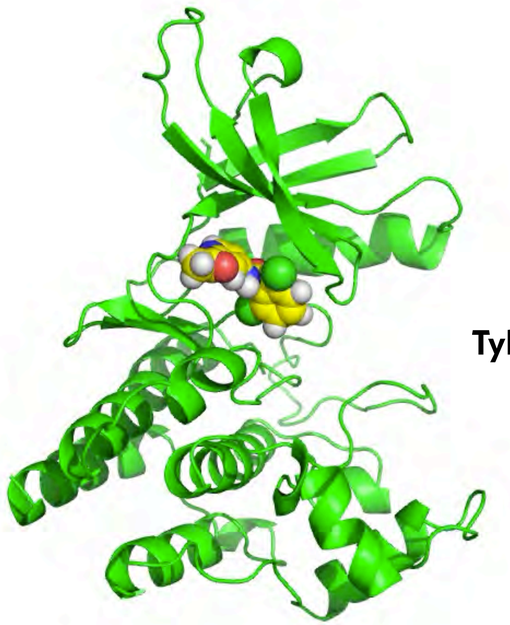
VEHICLe doi: http://doi.org/10.1021/jm801513z

YUANQING WANG

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

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





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

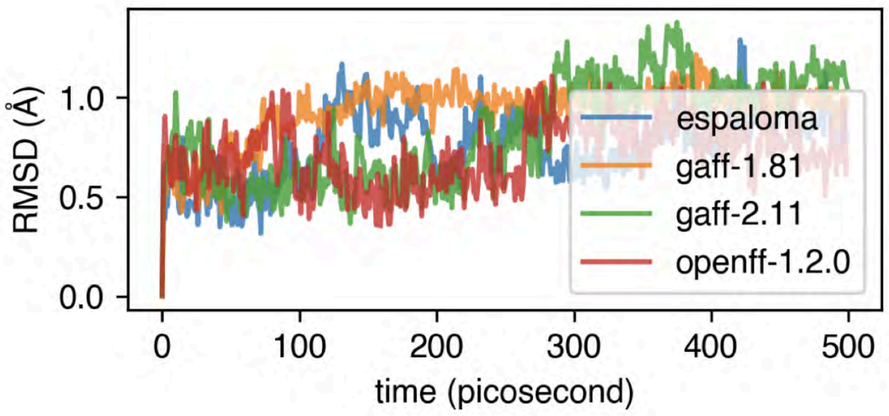
preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

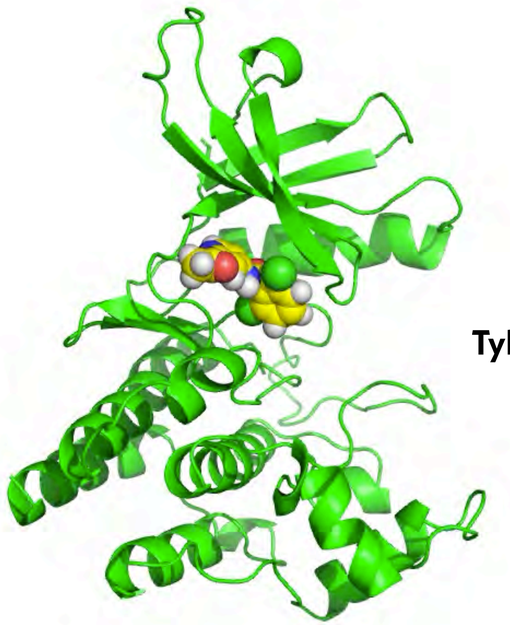
Tyk2 benchmark doi: https://doi.org/10.1021/ja512751q

YUANQING WANG

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

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





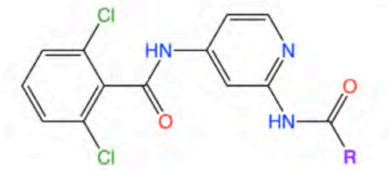
Tyk2 from OpenFF benchmark set
espaloma **joint** model
+ TIP3P water

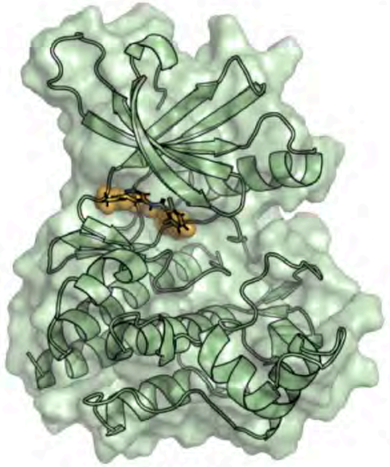
preprint: https://arxiv.org/abs/2010.01196
code: http://github.com/choderalab/espaloma

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

YUANQING WANG

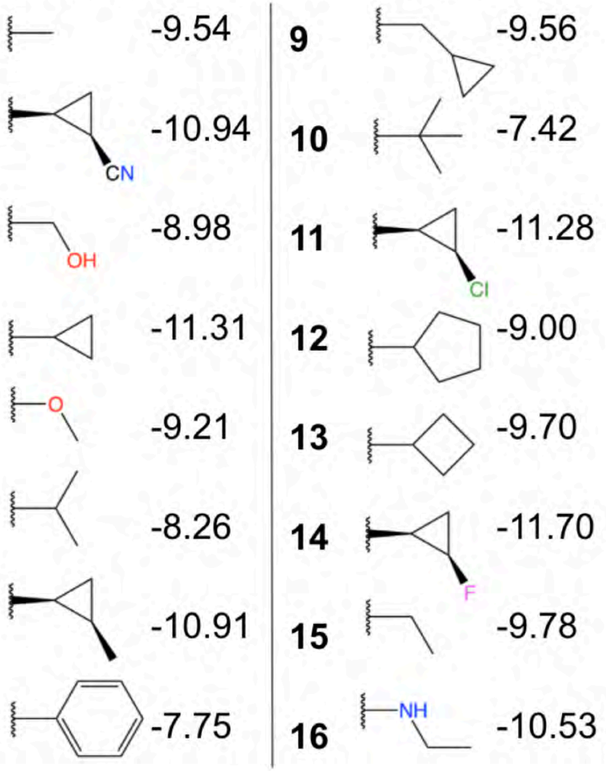
ESPALOMA SMALL MOLECULE PARAMETERS PERFORM AS WELL OR BETTER THAN MODERN BIOMOLECULAR FORCE FIELDS





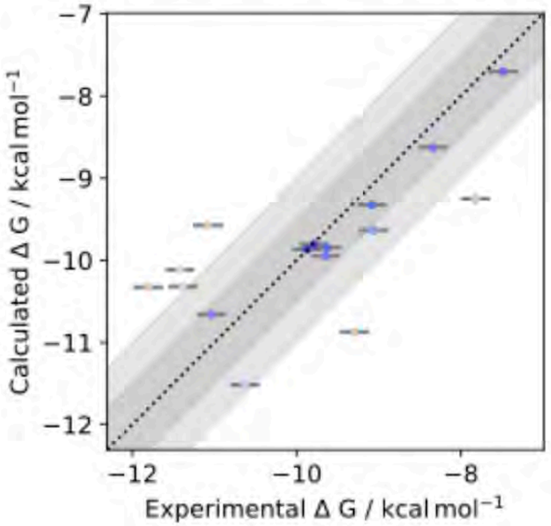
Tyk2
PDBID: 4GIH

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



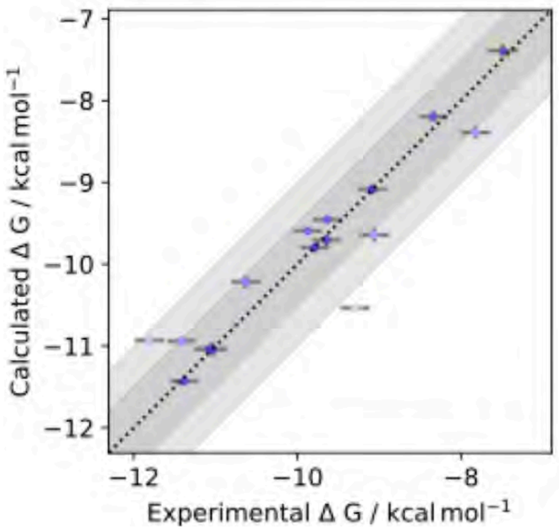
OpenFF 1.2.0 small molecule
Amber ff14SB protein
TIP3P water

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



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

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




MIKE HENRY


IVÁN PULIDO


IVY ZHANG

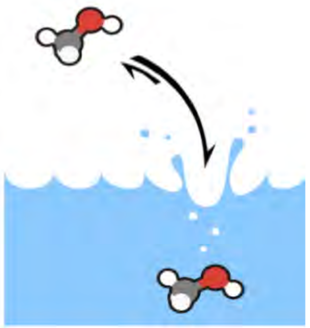

DOMINIC RUFA


HANNAH BRUCE CDONALD


YUANQING WANG

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

ESPALOMA CAN ALSO FIT EXPERIMENTAL FREE ENERGIES

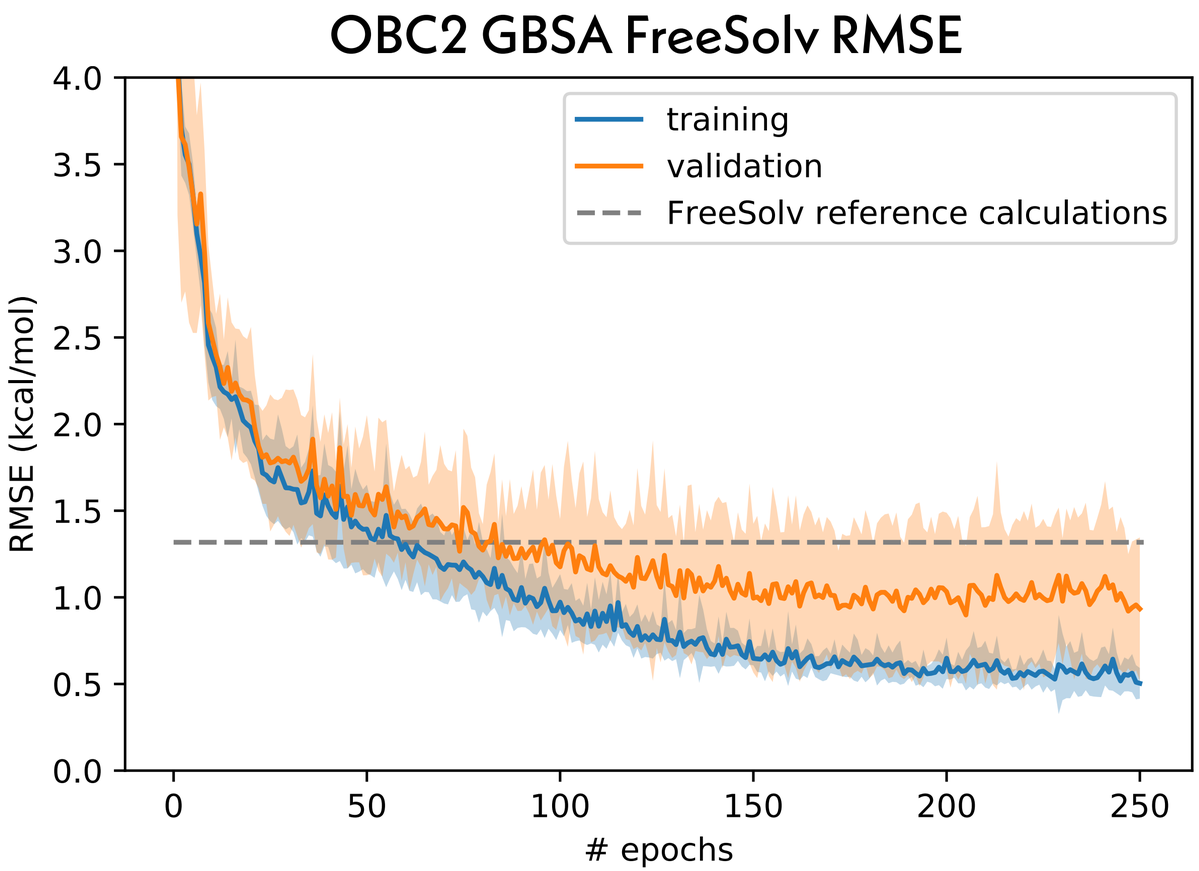


experimental hydration
free energies from **FreeSolv**
https://github.com/MobleyLab/FreeSolv

loss function:

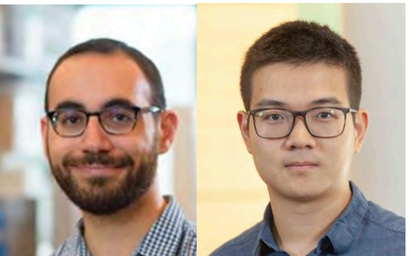
$$L(\Phi_{NN}) = \sum_{n=1}^{N} \frac{[\Delta G_n(\Phi_{NN}) - \Delta G_n^{\exp}]^2}{\sigma_n^2}$$

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

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

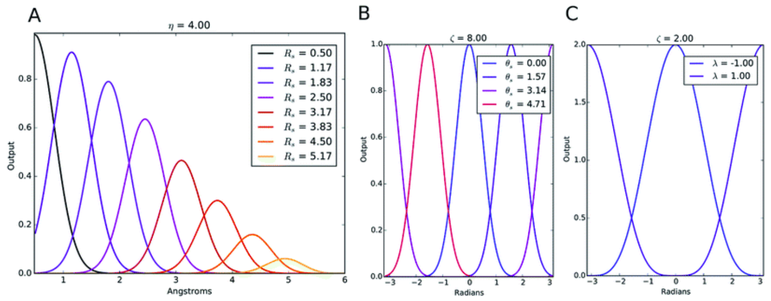
ANI family of quantum machine learning (QML) potentials

radial and **angular** features

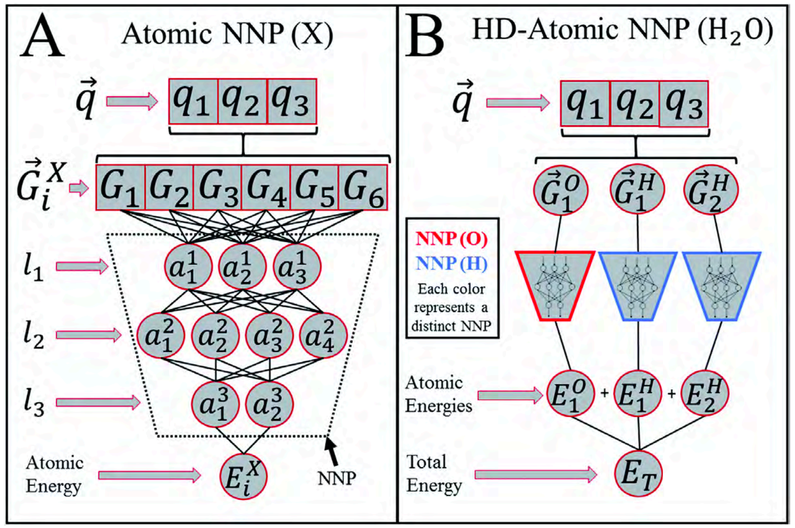
$$f_C(R_{ij}) = \begin{cases} 0.5\times\cos\left(\frac{\pi R_{ij}}{R_C}\right)+0.5 & \text{for } R_{ij} \le R_C \\ 0.0 & \text{for } R_{ij} > R_C \end{cases}$$

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

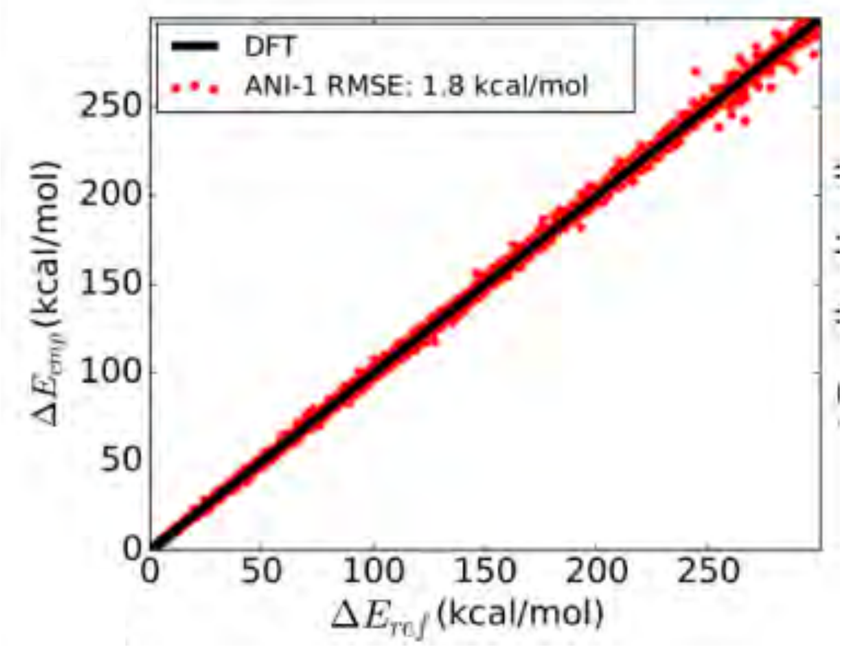
$$G_m^{A_{mod}} = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all atoms}} \left(1+\cos(\theta_{ijk}-\theta_s)\right)^{\zeta} \exp\left[-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_s\right)^2\right] f_C(R_{ij}) f_C(R_{ik})$$



deep neural network for each atom



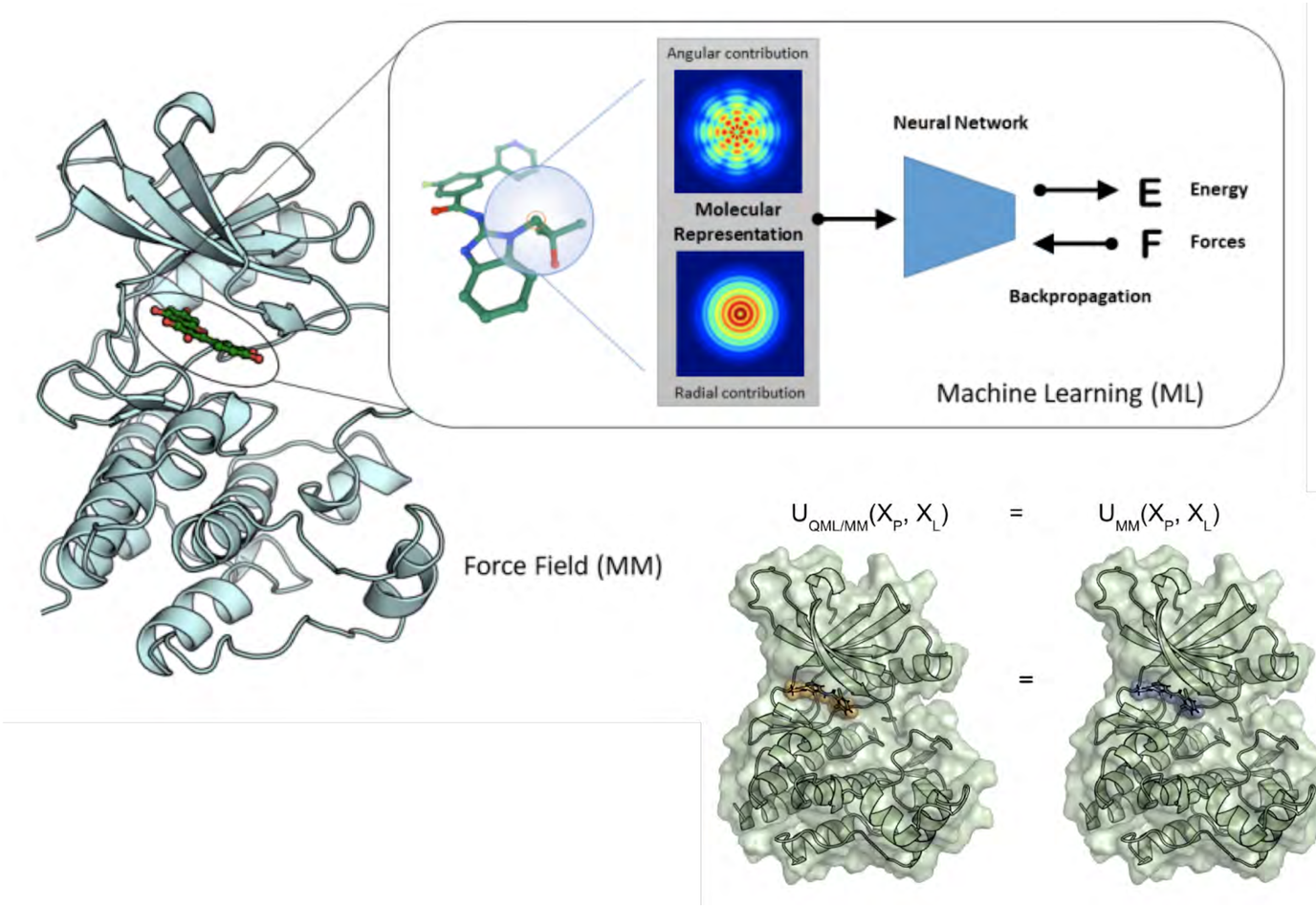
excellent agreement with DFT



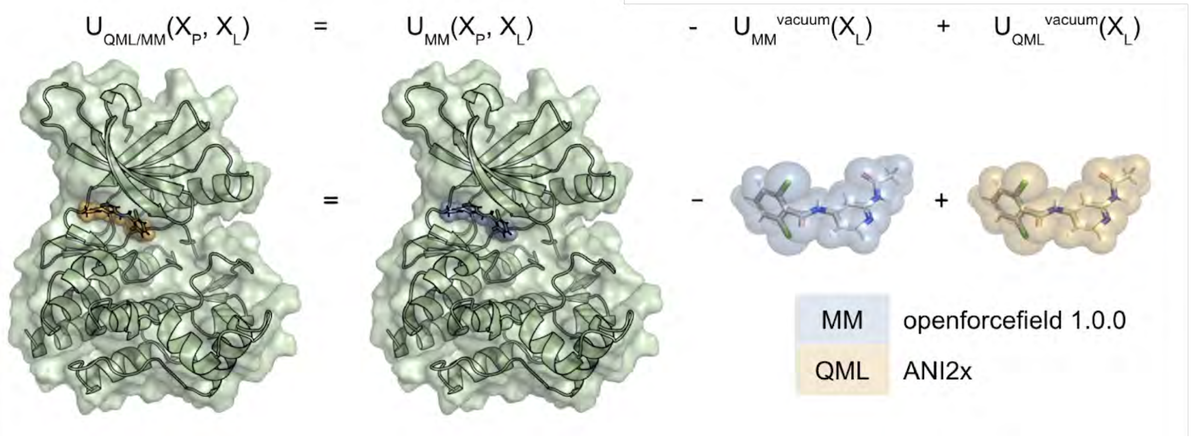
OLEXANDR ISAYEV **ADRIAN ROITBERG**

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

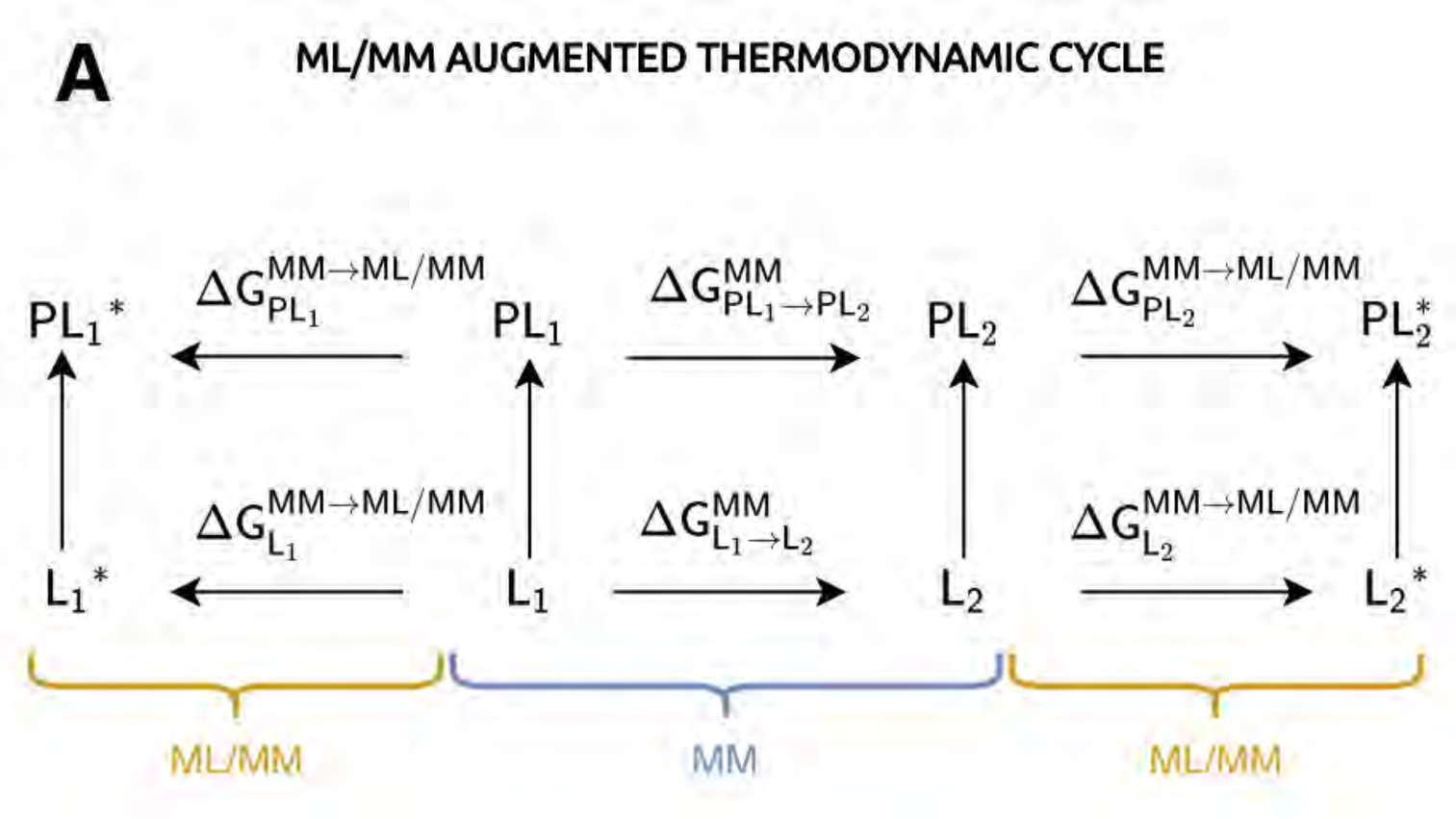


many QML/MM formulations possible, including those that use QML for protein-ligand interactions



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

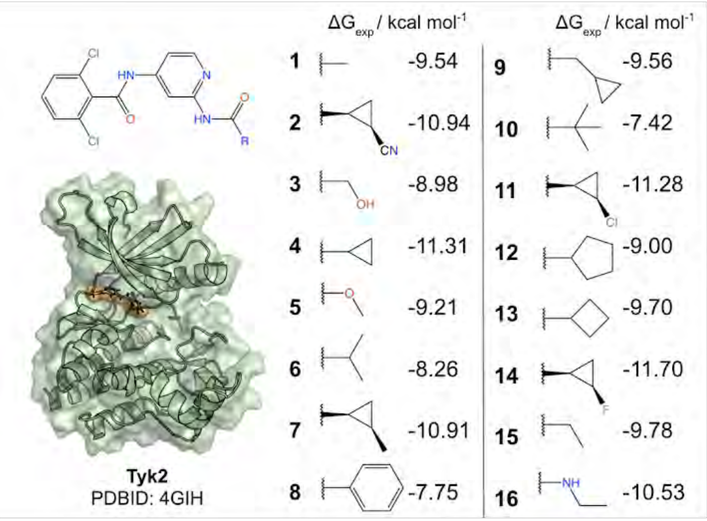


HYBRID QUANTUM MACHINE LEARNING / MOLECULAR MECHANICS (QML/MM) FREE ENERGY CALCULATIONS CUT ERROR IN HALF

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

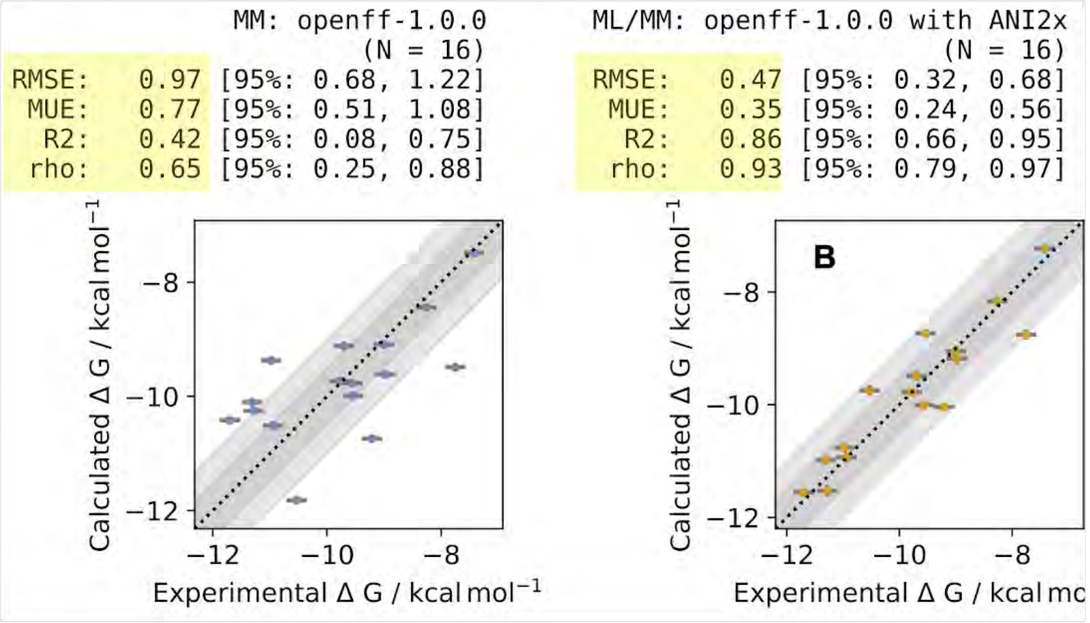
MM (OpenFF 1.0.0 "Parsley")
AMBER14SB protein force field
TIP3P; Joung and Cheatham ions

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

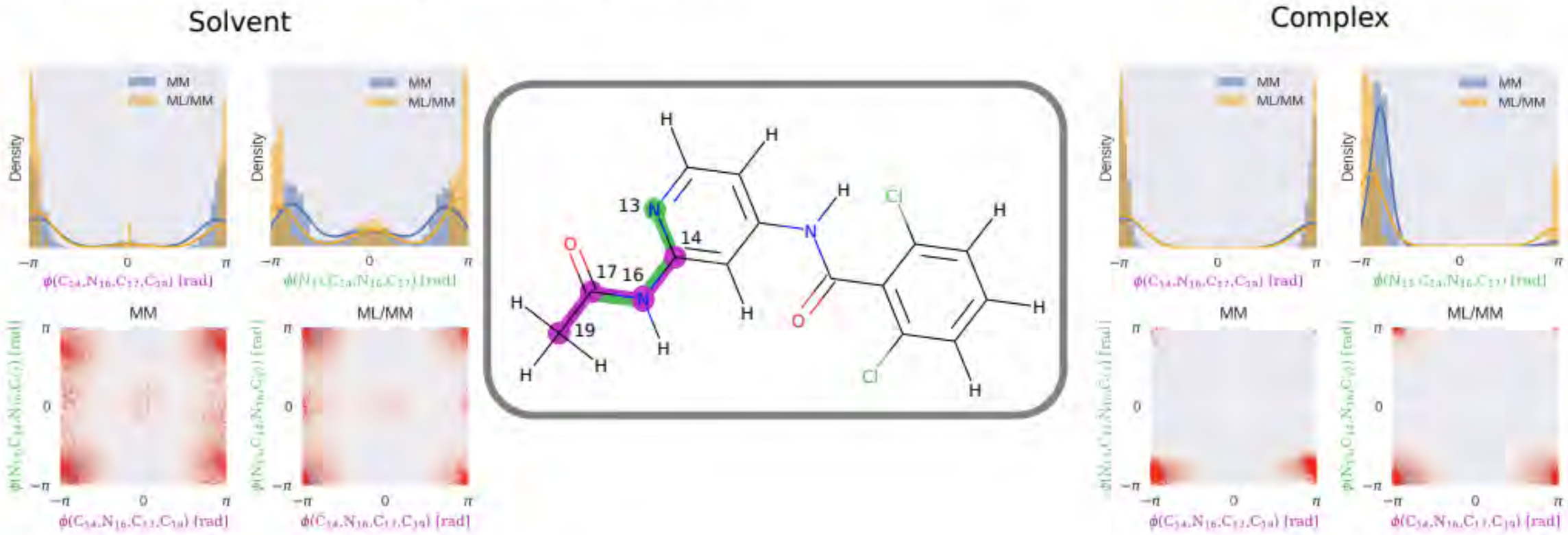
Free energies are in units of kilocalories per mole.



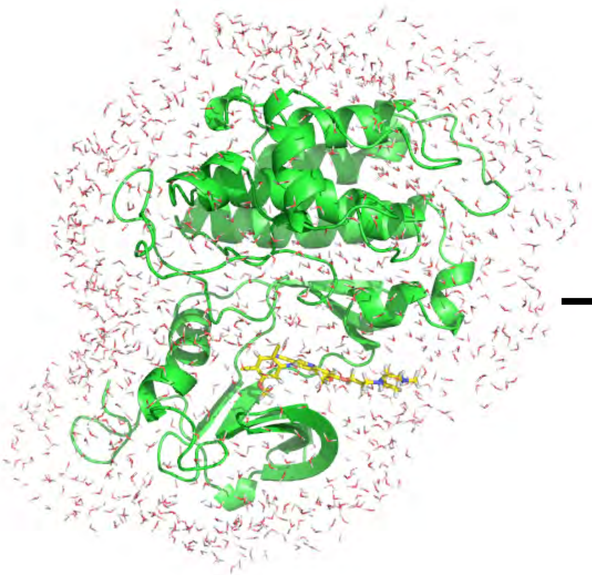
Tyk2 benchmark system from Wang et al. JACS 137:2695, 2015
replica-exchange free energy calculations with solute tempering (FEP/REST)

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

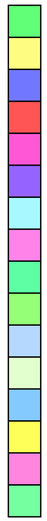
HYBRID QUANTUM MACHINE LEARNING / MOLECULAR MECHANICS (QML/MM) POST-PROCESSING CAN IMPROVE ACCURACY



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



atomic coordinates

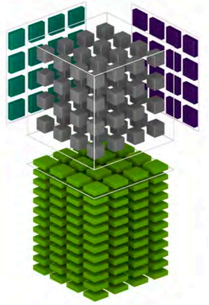


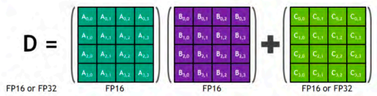
feature computation

NN computation

$$\sum_i$$

energy/force accumulation





tensor cores

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

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

PDB ID	# res	# heavy atoms	OpenMM ns/day (4 fs timestep)	TorchANI QML/MM ns/day (2 fs timestep)	OpenMM QML/MM* ns/day (2 fs timestep)
3BE9	328	48	436	10.4	96.5 / 50.8
2P95	286	50	430	7.93	96.8 / 49.8
1HPO	198	64	547	9.12	101 / 44.6
1AJV	198	75	666	9.19	101 / 40.7

* ANI ensemble size: 1 / 8

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

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

(~2.5x slower than GPU MD right now, but need 2x smaller timestep)
model distillation will become important in building single models that are efficient on hardware

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

OPENMM 8 WILL MAKE QML/MM SIMULATIONS INCREDIBLY EASY

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

OpenMM 8 beta should be out next week!

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

WE NEED A ML MODEL STANDARD AND REPOSITORY TO MAKE THEM EASIER TO DEPLOY AND USE

The OpenMM team has submitted an NIH proposal aiming to define portable standards:

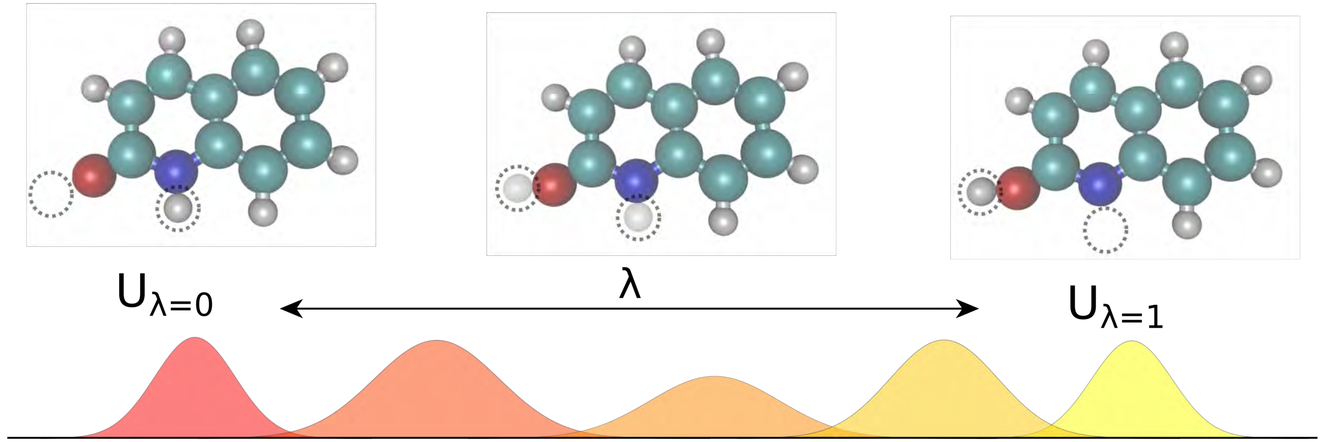
```
from simtk.openmm.app import MLModelRepository
# Grab ANI-1ccx from the ML model repository
model = MLModelRepository('ANI-1ccx')
# or grab a different model by DOI
model = MLModelRepository('10.2084/jctc.2985019')
# Create an OpenMM system from a specified molecular topology
system = model.create_system(topology)
# Simulate it in OpenMM
integrator = openmm.LangevinIntegrator(temperature, collision_rate, timestep)
context = openmm.Context(system, integrator)
context.setPositions(positions)
integrator.step(nsteps)
```

A well-defined portable QML standard would make it easier to build and deliver QML force fields to multiple simulation packages.

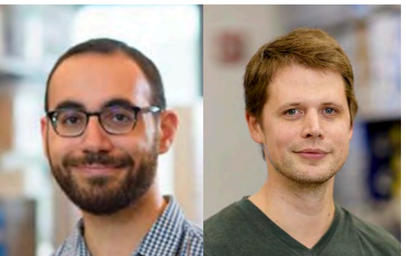
PURE QUANTUM MACHINE LEARNING (QML) POTENTIALS CAN BE USED TO COMPUTE FREE ENERGY DIFFERENCES BETWEEN CHEMICAL SPECIES

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

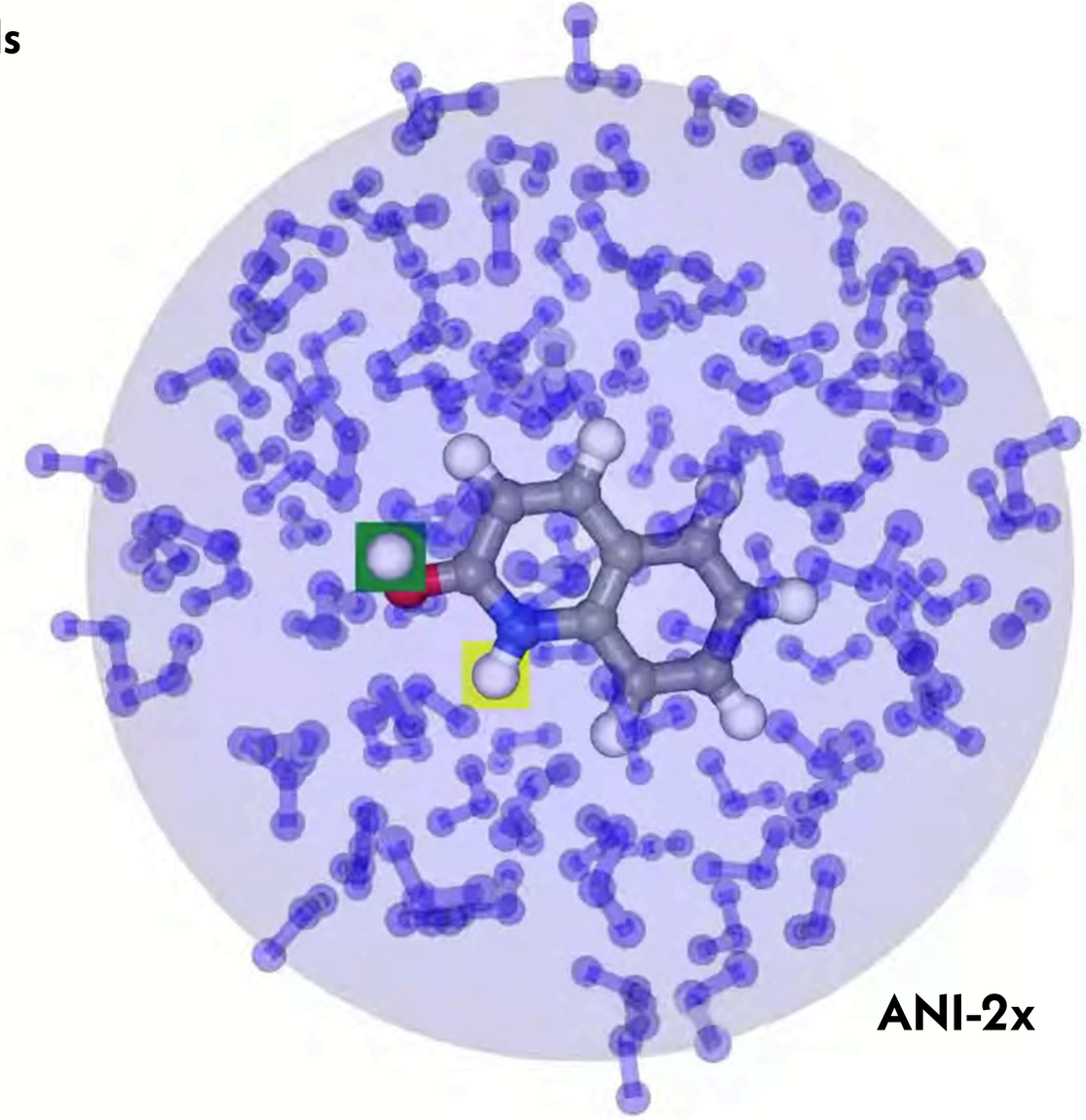
$$U(x;\lambda) = (1-\lambda)U_{\lambda=0}(x) + \lambda U_{\lambda=1}(x)$$



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



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

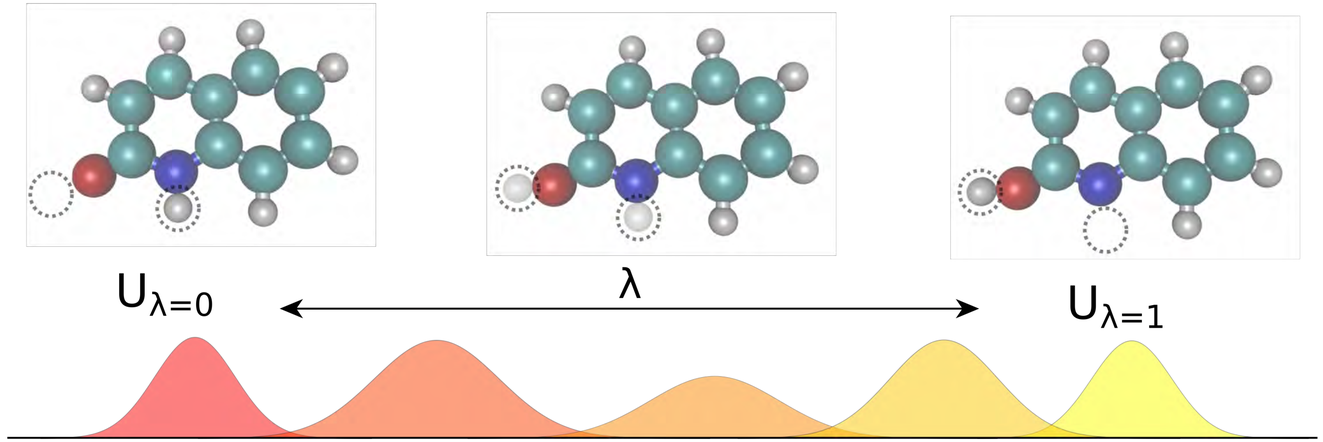


ANI-2x

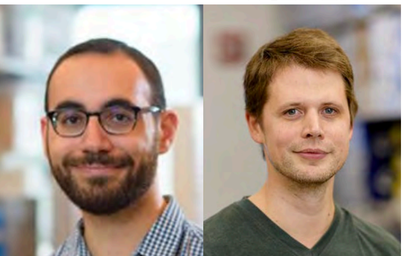
PURE QUANTUM MACHINE LEARNING (QML) POTENTIALS CAN BE USED TO COMPUTE FREE ENERGY DIFFERENCES BETWEEN CHEMICAL SPECIES

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

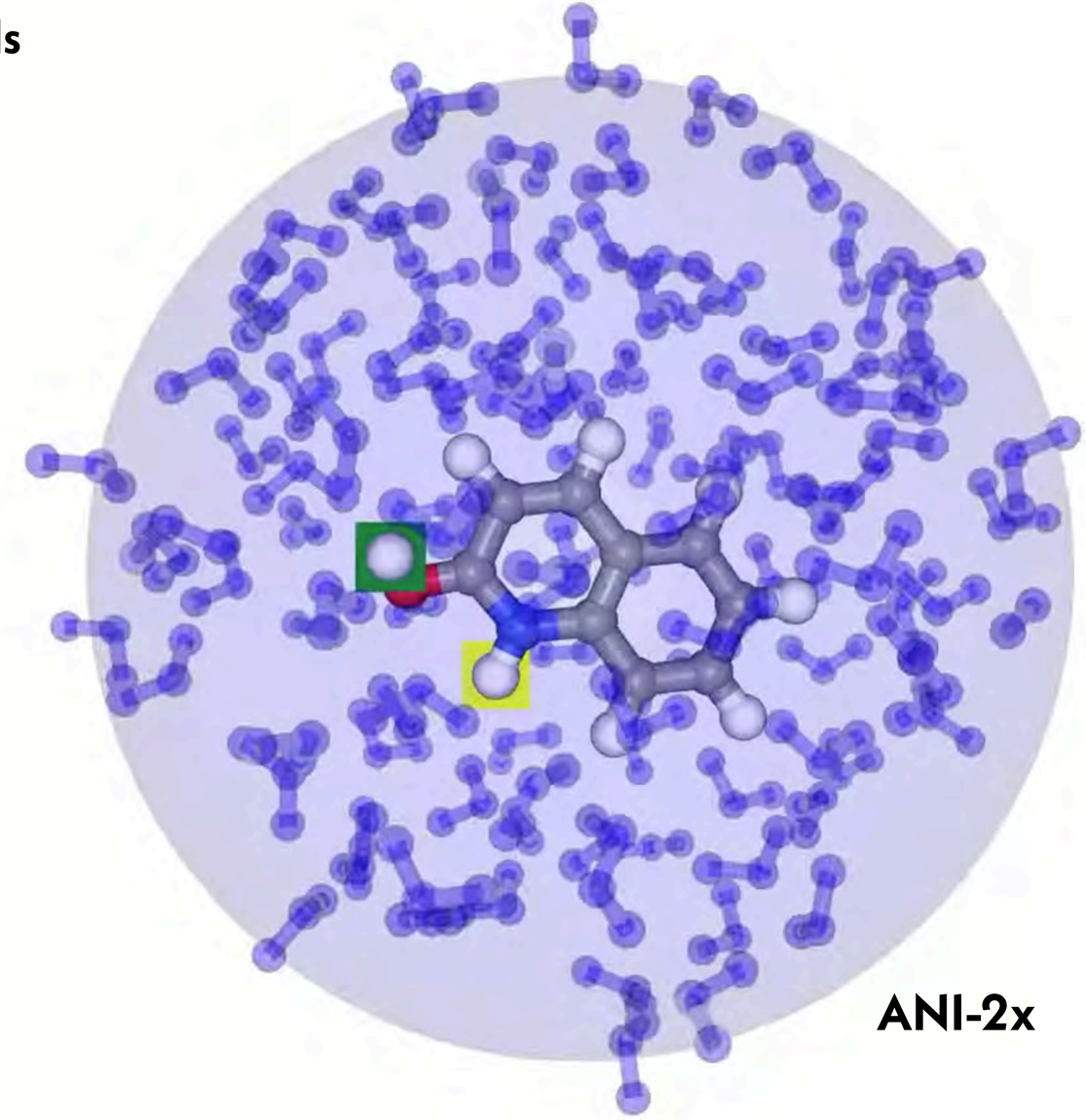
$$U(x;\lambda) = (1-\lambda)U_{\lambda=0}(x) + \lambda U_{\lambda=1}(x)$$



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



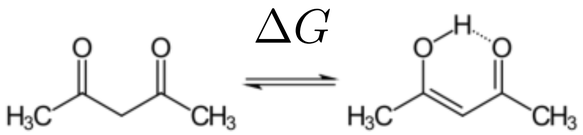
preprint: https://doi.org/10.1101/2020.10.24.353318
code: https://github.com/choderalab/neutromeratio



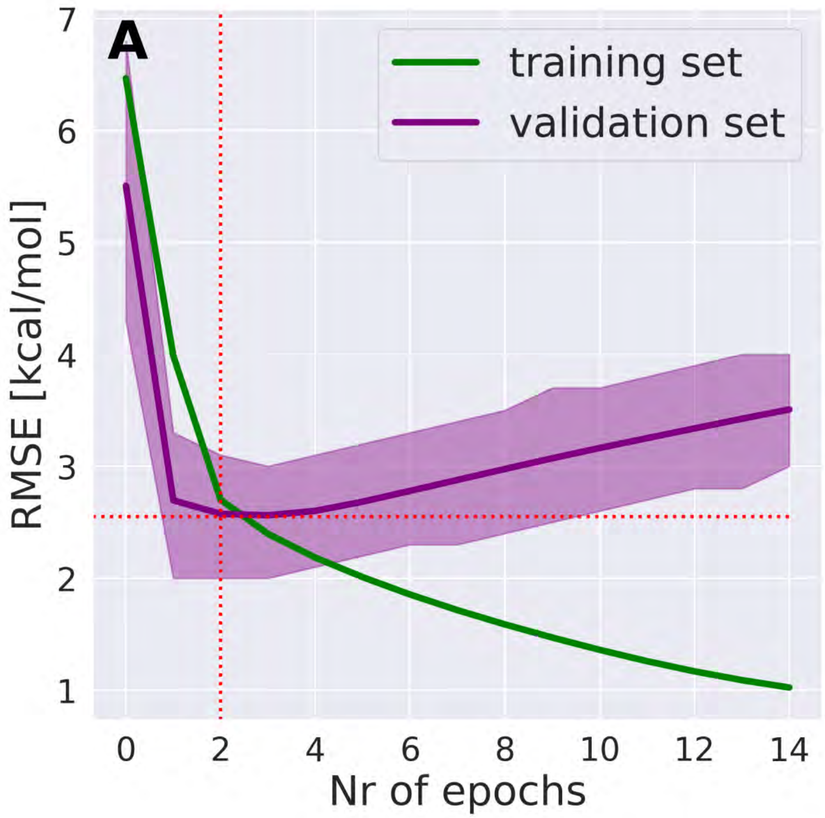
ANI-2x

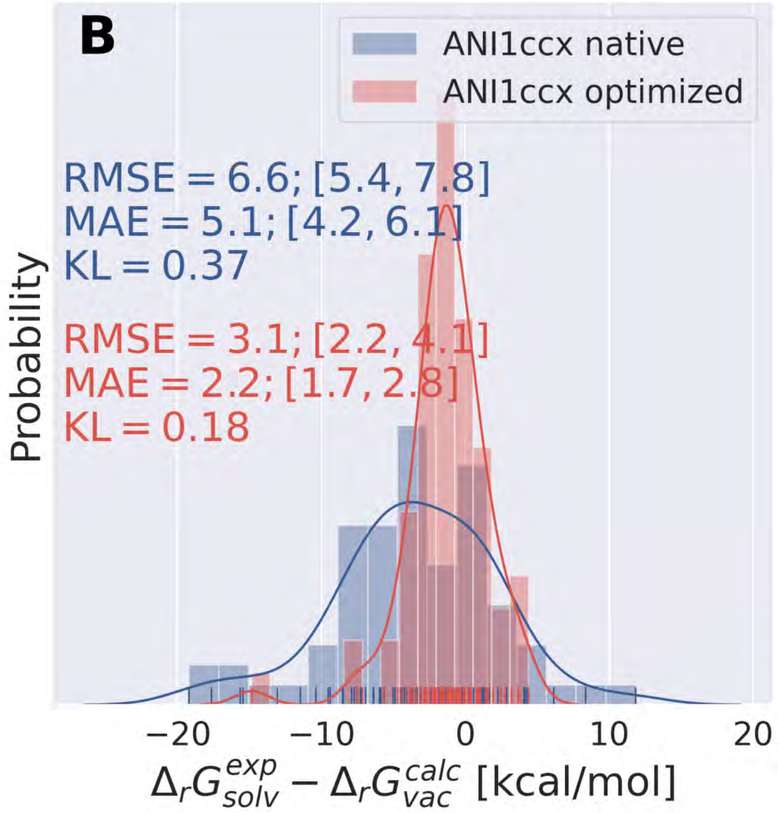
QML POTENTIALS CAN LEARN FROM EXPERIMENTAL DATA TO IMPROVE PHYSICAL MODELS

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

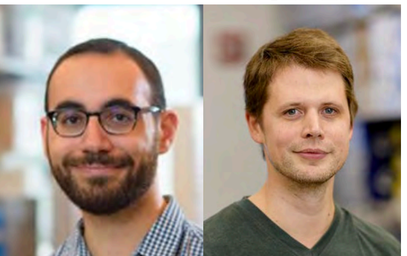


train: 221 tautomer pairs
validate: 57 tautomer pairs
test: 72 tautomer pairs





JOSH FASS
MARCUS WIEDER



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

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

SPICE DES Monomers Single Points Dataset v1.1	2021-11-15-QMDataset-DES-monomers-single-points	Single point energy calculation of DES monomers.	I, C, Br, P, Cl, H, S, O, F, N
SPICE Solvated Amino Acids Single Points Dataset v1.1	2021-11-08-QMDataset-Solvated-Amino-Acids-single-points	Single point energy calculation of solvated amino acids.	N, S, O, C, H
SPICE DES370K Single Points Dataset v1.0	2021-11-08-QMDataset-DES370K-single-points	SPICE single point dataset for ML applications.	'N', 'O', 'Mg', 'H', 'F', 'K', 'Br', 'Na', 'P', 'Cl', 'I', 'Ca', 'S', 'Li', 'C'
SPICE DES370K Single Points Dataset Supplement v1.0	2022-02-18-QMDataset-DES370K-single-points-supplement	SPICE single point dataset for ML applications.	F, H, Cl, S, I, Br, N, Li, O, C, Na
SPICE Dipeptides Single Points Dataset v1.2	2021-11-08-QMDataset-Dipeptide-single-points	SPICE single point dataset for ML applications.	C ,N ,O ,H ,S
SPICE PubChem Set 1 Single Points Dataset v1.2	2021-11-08-QMDataset-pubchem-set1-single-points	SPICE single point dataset for ML applications.	'O', 'Cl', 'N', 'C', 'P', 'Br', 'S', 'F', 'I', 'H'
SPICE PubChem Set 2 Single Points Dataset v1.2	2021-11-09-QMDataset-pubchem-set2-single-points	SPICE single point dataset for ML applications.	'H', 'P', 'C', 'Cl', 'Br', 'N', 'F', 'S', 'O', 'I'
SPICE PubChem Set 3 Single Points Dataset v1.2	2021-11-09-QMDataset-pubchem-set3-single-points	SPICE single point dataset for ML applications.	'N', 'C', 'S', 'Cl', 'Br', 'F', 'P', 'I', 'H', 'O'
SPICE PubChem Set 4 Single Points Dataset v1.2	2021-11-09-QMDataset-pubchem-set4-single-points	SPICE single point dataset for ML applications.	'N', 'S', 'Br', 'O', 'C', 'F', 'H', 'I', 'Cl', 'P'
SPICE PubChem Set 5 Single Points Dataset v1.2	2021-11-09-QMDataset-pubchem-set5-single-points	SPICE single point dataset for ML applications.	'F', 'H', 'S', 'Br', 'Cl', 'N', 'P', 'C', 'I', 'O'
SPICE PubChem Set 6 Single Points Dataset v1.2	2021-11-09-QMDataset-pubchem-set6-single-points	SPICE single point dataset for ML applications.	'Cl', 'O', 'N', 'H', 'C', 'P', 'S', 'F', 'Br', 'I'

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

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

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

WE HAVE AN OPPORTUNITY TO TRANSFORM DRUG DISCOVERY

WE HAVE AN OPPORTUNITY TO TRANSFORM DRUG DISCOVERY

* Quantum machine learning (QML) will replace QM pretty much everywhere, bringing a revolution in accuracy—**if we can make them easy to build, use, and share**

WE HAVE AN OPPORTUNITY TO TRANSFORM DRUG DISCOVERY

* Quantum machine learning (QML) will replace QM pretty much everywhere, bringing a revolution in accuracy—**if we can make them easy to build, use, and share**
* QML/MM hybrid simulations will bring a revolution in the accuracy and utility of structure-based design—**if we can make them fast enough**

WE HAVE AN OPPORTUNITY TO TRANSFORM DRUG DISCOVERY

* Quantum machine learning (QML) will replace QM pretty much everywhere, bringing a revolution in accuracy—**if we can make them easy to build, use, and share**
* QML/MM hybrid simulations will bring a revolution in the accuracy and utility of structure-based design—**if we can make them fast enough**
* QML/MM free energy calculations can learn from project data, enabling biotech to extract much more value from their data—**if we can make them easy to train**

WE HAVE AN OPPORTUNITY TO TRANSFORM DRUG DISCOVERY

* Quantum machine learning (QML) will replace QM pretty much everywhere, bringing a revolution in accuracy—**if we can make them easy to build, use, and share**
* QML/MM hybrid simulations will bring a revolution in the accuracy and utility of structure-based design—**if we can make them fast enough**
* QML/MM free energy calculations can learn from project data, enabling biotech to extract much more value from their data—**if we can make them easy to train**
* Hybrid combinations of ML for short-range and MM for long-range will deliver significant systematic accuracy improvements—**if we can make them practical**

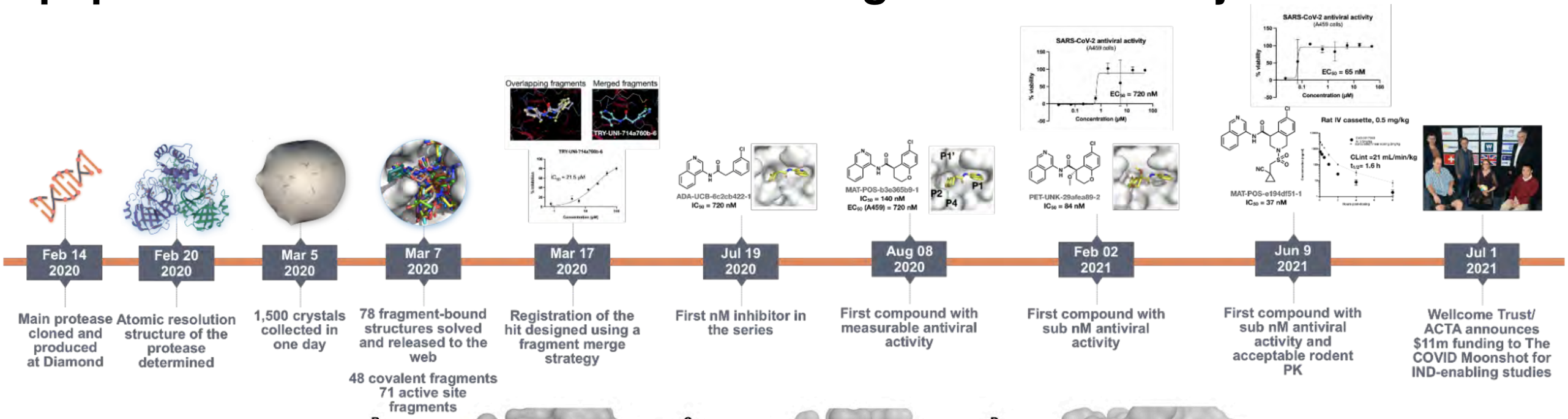
WE HAVE AN OPPORTUNITY TO TRANSFORM DRUG DISCOVERY

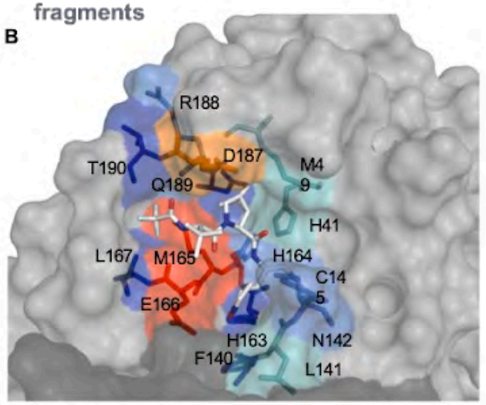
* Quantum machine learning (QML) will replace QM pretty much everywhere, bringing a revolution in accuracy—**if we can make them easy to build, use, and share**
* QML/MM hybrid simulations will bring a revolution in the accuracy and utility of structure-based design—**if we can make them fast enough**
* QML/MM free energy calculations can learn from project data, enabling biotech to extract much more value from their data—**if we can make them easy to train**
* Hybrid combinations of ML for short-range and MM for long-range will deliver significant systematic accuracy improvements—**if we can make them practical**
* ML collective variables will drive a revolution in sampling—**if we can make it easy to go between MD and ML frameworks**

WE HAVE AN OPPORTUNITY TO TRANSFORM DRUG DISCOVERY

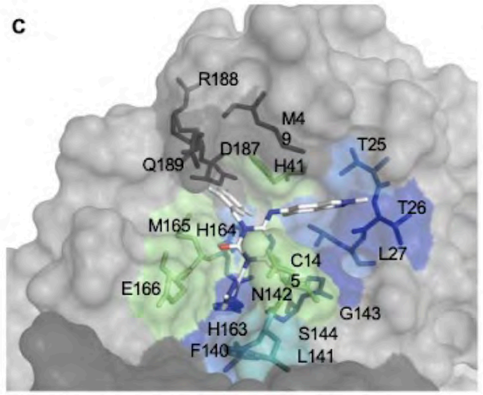
* Quantum machine learning (QML) will replace QM pretty much everywhere, bringing a revolution in accuracy—**if we can make them easy to build, use, and share**
* QML/MM hybrid simulations will bring a revolution in the accuracy and utility of structure-based design—**if we can make them fast enough**
* QML/MM free energy calculations can learn from project data, enabling biotech to extract much more value from their data—**if we can make them easy to train**
* Hybrid combinations of ML for short-range and MM for long-range will deliver significant systematic accuracy improvements—**if we can make them practical**
* ML collective variables will drive a revolution in sampling—**if we can make it easy to go between MD and ML frameworks**
* ML potentials are a solution for multiscale simulations—**if we can facilitate exchange between MD and ML frameworks**

The open science COVID Moonshot produced a novel noncovalent, non-peptidomimetic oral antiviral from a fragment screen in just 18 months

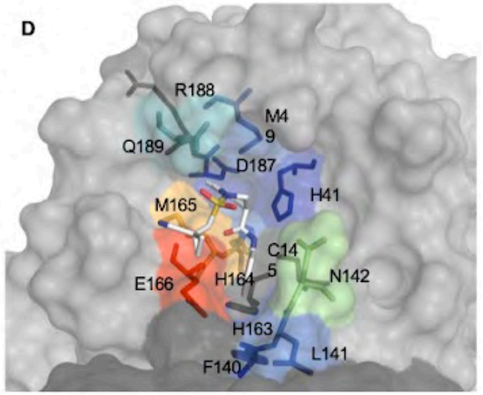




nirmatrelvir (Paxlovid)
Pfizer



S-217622
Shionogi
(modeled)



MAT-POS-e194df51-1
COVID Moonshot

COVID Moonshot structures and data: http://postera.ai/covid
preprint: https://www.biorxiv.org/content/10.1101/2020.10.29.339317v3.abstract
history: https://www.nature.com/articles/d41586-021-01571-1

DNDi
Drugs for Neglected Diseases *initiative*

COVID Moonshot

We are negotiating a straight to generics route with multiple generics manufacturers

We have a path to go “straight to generics” (potentially entirely free of patents)
to enable global, affordable, and accessible access to meet the needs of underserved LMICs

The Moonshot team has been funded as an NIH Antiviral Drug Discovery (AViDD) Center to pursue the same strategy to produce novel antivirals for future pandemics

coronaviruses

flaviviruses

picornaviruses

Warm start targets: nsp13 nsp14 nsp16 2Apro ns2b/3

PLpro nsp3-Mac1 N-protein

Mpro

PROJECT 1 — Antiviral targeting to suppress drug resistance — D A

PROJECT 2 — Target enablement — D S

PROJECT 3 — Fragment-to-lead and target validation

PROJECT 4 — Covalent targeting strategies — D S B A

PROJECT 5 — Lead optimization — D S B A

PROJECT 6 — Preclinical development and translation — D A

OPEN SCIENCE OUTPUT BY YEAR 5:

- **resistance-robust targets**
- **10 target-enabling packages (TEPs)** (data disclosed in real time)
- **9 early lead packages with broad antiviral activity** (data disclosed in real time)
- **6 lead optimization campaigns** (data disclosed in real time)
- **3 INDs** with public **preclinical data packages**

P1: Karla Kirkegaard (Stanford)
Matt Bogyo (Stanford)
Jesse Bloom (Fred Hutch)

P2: Frank von Delft (Diamond Light Source)
Martin Walsh (Diamond Light Source)
Oxford CMD SRF [service facility]

diamond

P3: Alpha Lee (PostEra)
John Chodera (MSKCC)
Frank von Delft (Diamond)
Ed Griffen (Medchemica)
Nir London (Weizmann)
Karla Kirkegaard (Stanford)
Martin Walsh (Diamond)

PostEra

P4: Nir London (Weizmann)
Matt Bogyo (Stanford)

P5: Ed Griffen (Medchemica)
Ben Perry (DNDi)
Alpha Lee (PostEra)
John Chodera (MSKCC)

MedChemica
CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY

P6: Ben Perry (DNDi)
Laurent Fraisse (DNDi)
Annette von Delft (Medchemica)

DNDi

SUPPORTING LETTERS

Pfizer · Takeda · NOVARTIS · grupoINSUD

ADMINISTRATIVE CORE

John Chodera (MSKCC)
Ben Perry (DNDi)
Alpha Lee (PostEra)
Administrative Director
Project Coordinator

PIs

D — DATA INFRASTRUCTURE CORE
Alpha Lee (PostEra)
Matthew Robinson (PostEra)
Frank von Delft (Diamond)
John Chodera (MSKCC)

S — STRUCTURAL BIOLOGY CORE
Frank von Delft (Diamond Light Source)
Daren Fearon (Diamond Light Source)
Martin Walsh (Diamond Light Source)

B — BIOCHEMICAL ASSAY CORE
Nir London (Weizmann)
Haim Barr (Weizmann)

A — ANTIVIRAL EFFICACY AND RESISTANCE CORE
Kris White (Mount Sinai)
Adolfo García-Sastre (Mount Sinai)
Randy Albrecht (Mount Sinai)
Johan Neyts (Leuven) [service facility]

ASAP Discovery website: http:/asapdiscovery.org

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

STIFTUNG CHARITÉ

SCHRÖDINGER.

Scientific Advisor: OpenEye, Foresite Labs

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

STARR CANCER CONSORTIUM

open forcefield consortium