open
forcefield

@openforcefield

www.openforcefield.org

Sage, the Open Force Field 2.0.0 and the road map forward

AIChE, Development of Intermolecular Potential Models

November 17, 2022

Michael R. Shirts, Department of Chemical and Biological Engineering

University of Colorado Boulder

Acknowledgements: lots of people to thank in a huge effort!

open forcefield consortium

NSF

NIH

QM fitting
Chaya Stern (MSKCC)
Daniel Smith (MolSSI)
Yudong Qiu (UC Davis)
Victoria Lim (UC Irvine)
Hyesu Jang (UC Davis)
Lee-Ping Wang (UC Davis)

Software Infrastructure
Jeff Wagner (UC San Diego / UC Irvine)
Matt Thompson (CU Boulder)
Simon Boothroyd (MSKCC / CU Boulder)
David Slochower (UC San Diego)

Physical Properties
Simon Boothroyd (MSKCC / CU Boulder)
David Hahn (Janssen)
Hannah Bruce McDonald (MSKCC)
Dominic Rufa (MSKCC)
Owen Madin (CU Boulder)
Vytas Gapsys (Max Planck Institute)


John Chodera (MSKCC)

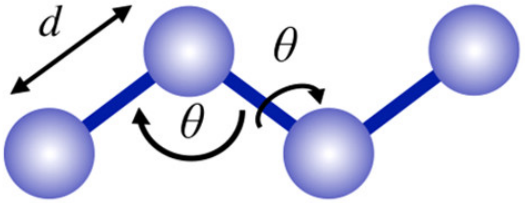

Michael Gilson (UC SD)


David Mobley (UC Irvine)


Michael Shirts (CU Boulder)

https://openforcefield.org/about/team/

There are still lots of opportunities to improve small molecule force fields even at low levels of physics



- Simple and effective functional forms have been in use for 30+ years
- We can still improve these force fields with **better chemical perception**, **better optimization** techniques, and **better training datasets**
- This is a large undertaking!
- Various groups have been working on improving these force fields for decades

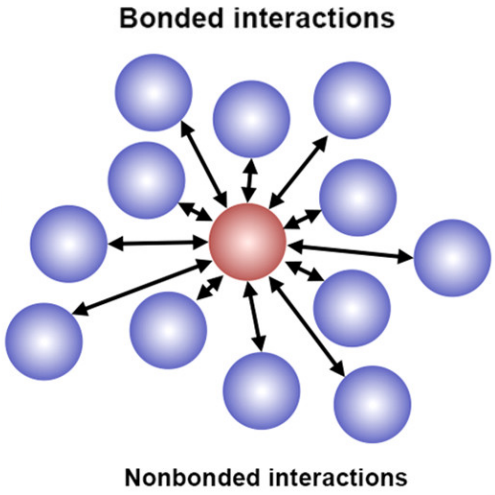


Figure adapted from Riniker S *J. Chem. Inf. Model.* 2018, 58, 3, 565–578

OPEN SOFTWARE - OPEN DATA - OPEN SCIENCE:
Rapidly facilitating force field science!

OPEN SOFTWARE

Automated infrastructure enables rapid experimentation with minimum human intervention

OPEN DATA

Access to large, high quality experimental and quantum chemical data facilities easy curation of balanced train / test sets

OPEN SCIENCE

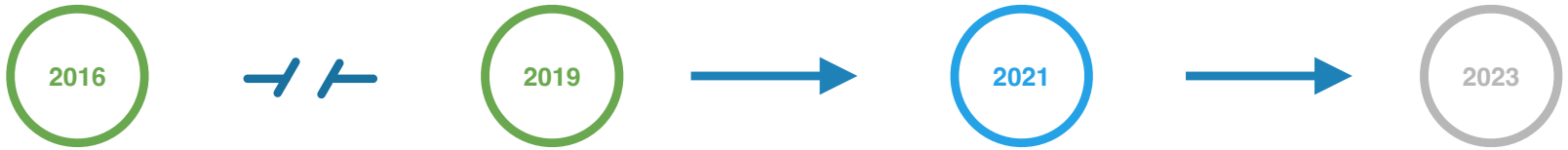
Exploring new force field science:

hypothesis - build software - train - test - iterate

is now almost routine

We're generating a series of force fields

- OpenFF force field progression since the Initiative's inception



SMIRNOFF99Frosst

Initial SMIRNOFF port of the parm99Frosst force field

Parsley

Retrained **valence** parameters against a redesigned QC data set

Sage

Retrained **vdW** parameters against physical property data + retrained **valence** parameters

Rosemary

Self consistent biopolymer + small molecule force field

Additional physics, still being finalized

What makes the Open Force Field Initiative different?

Develop infrastructure to **rapidly evaluate the effect of parameters** on physical properties of interest

Creating workflows for fragmenting and generating **QM data for torsions** and other **valence terms**

Using **chemical perception** (SMARTS/ SMIRKS) to assign force field parameters

Distinguishing scientific elements of the Open Force Field Initiative

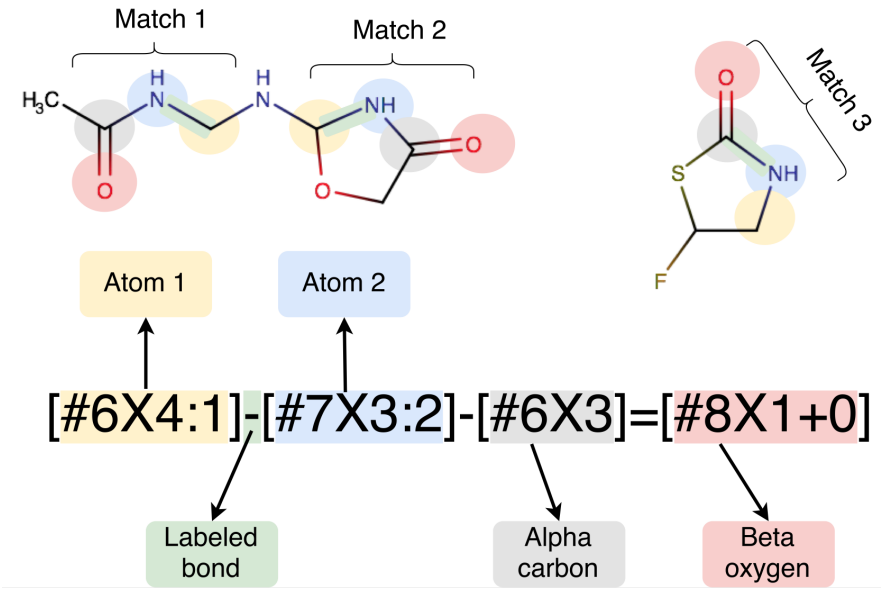
Enabling the use of **Bayesian inference** as a framework for making force field science decisions

Curate and generate **experimental and QM datasets** for force field parameterization and assessment

The SMIRKS Native Open Force Field (SMIRNOFF)

SMIRNOFF avoids atom typing and simplifies parameter assignment!

match bonds directly:



Use of industry-standard SMARTS/SMIRKS chemical perception greatly simplifies tooling for parameter assignment while solving issues with extensibility and flexibility.

Ditching “atom types” for SMIRKS (“parameter types”) allows parameter simplification

For example, GAFF2 has 16 vdW types for carbon

c	1.8606	0.0988
cs	1.8606	0.0988
ca	1.8606	0.0988
cc	1.8606	0.0988
cd	1.8606	0.0988
ce	1.8606	0.0988
cf	1.8606	0.0988
cp	1.8606	0.0988
cq	1.8606	0.0988
cz	1.8606	0.0988
cu	1.8606	0.0988
cv	1.8606	0.0988
cg	1.9525	0.1596
ch	1.9525	0.1596
cx	1.9069	0.1078
cy	1.9069	0.1078

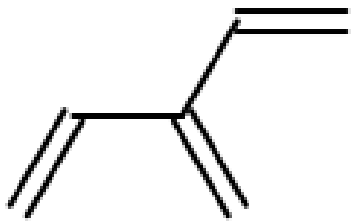
But this should be three SMIRKS strings

[#6:1]	1.8606	0.0988
[#6X1:1]	1.9525	0.1596
[#6X3r3,#6X3r4:1]	1.9069	0.1078

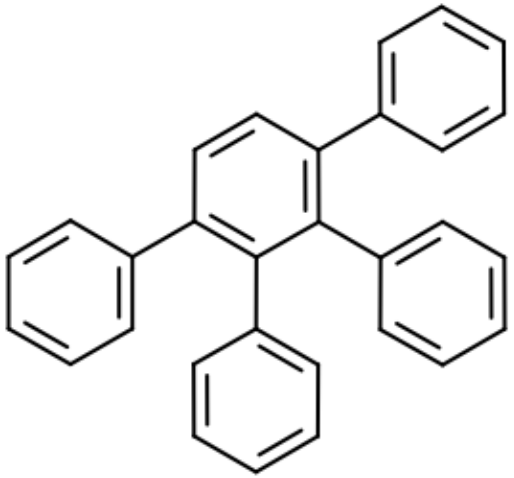
Very relevant when attempting to automatically fit parameters — are there 32 parameters here, or 6?

Why all these atom types? The larger issue is more fundamental

Atom typing discards bond order, but we *need* bond order for parameter assignments



- With knowledge of which bonds are single, double, aromatic, have formal charges, etc., parameter assignment is straightforward
- Without it, atom types must “carry” bond order information, which is almost impossible to do in general
- SMIRKS force fields handle this seamlessly



SMIRNOFF allowed significant compression of smirnoff99Frosst, our AMBER-lineage starting point

Description	Force Field	Lines of parameters
Basic Amber FF:	parm99	720
Merck Frosst small mol:	parm@Frosst	2893
Total:		3613

smirnoff99Frosst	332

Chemical Space Coverage

Database	smirnoff 99Frosst	parm @Frosst
DrugBank	99.7%	60%
ZINC	99.8%	52%
eMolecules	99.5%	--

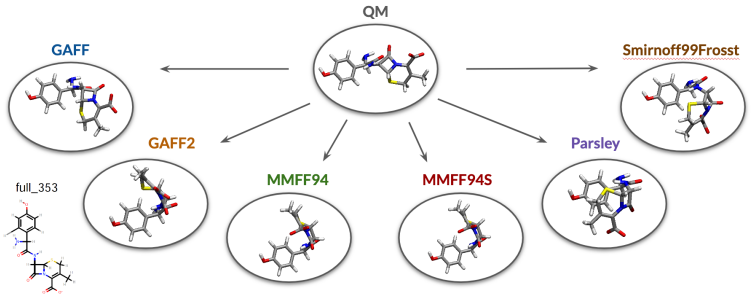
- Less than 1/10 the size of the original force field
- Removes redundancy
- Almost completely covers pharmaceutical chemical space

Chris Bayly

Caitlin Bannan

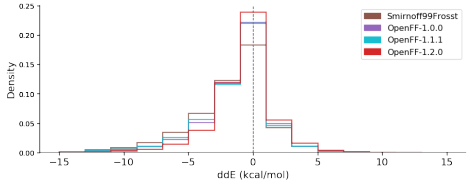
BENCHMARK ASSESSMENT OF MOLECULAR ENERGIES AND GEOMETRIES WITH RESPECT TO QM DATA[1]

1. **Geometry optimization:** Molecular geometries are optimized with various FFs starting from the same QM geometry.

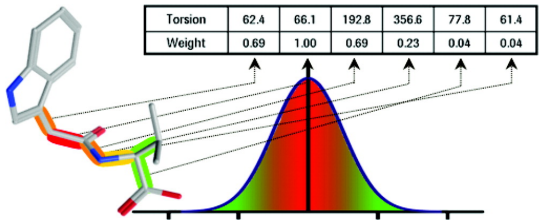


2. **Energy comparison:** The *ddE* for some conformer *i* is calculated relative to the conformer with the lowest QM energy (0) for different force fields:

$$ddE = dE_{FF,i} - dE_{QM,i} = (E_{FF,i} - E_{FF,0}) - (E_{QM,i} - E_{QM,0})$$

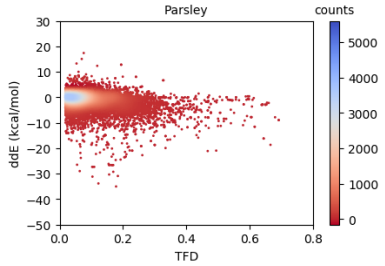


3. **Geometry comparison:** Geometry is evaluated through root-mean-square deviation (RMSD) and torsion fingerprint deviation (TFD)[2].



TFD uses Gaussian-weighted differences of torsion angles between two conformations and may be more independent of molecular size for structure comparison purposes.

4. **ddE vs TFD plots:** Compare high density regions of energy vs. geometry data. A perfect match between FF and QM would result with all points at (0, 0).

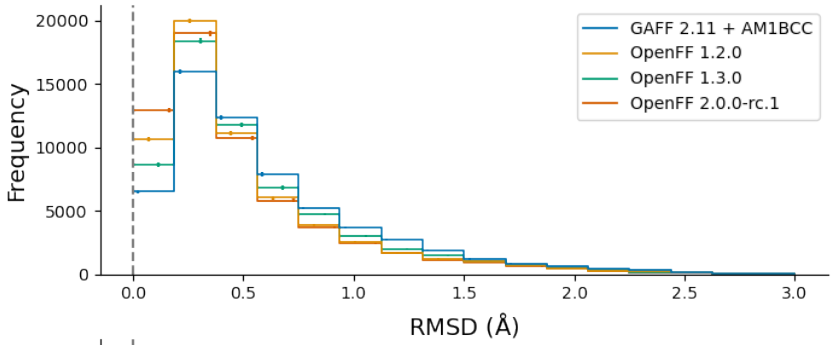


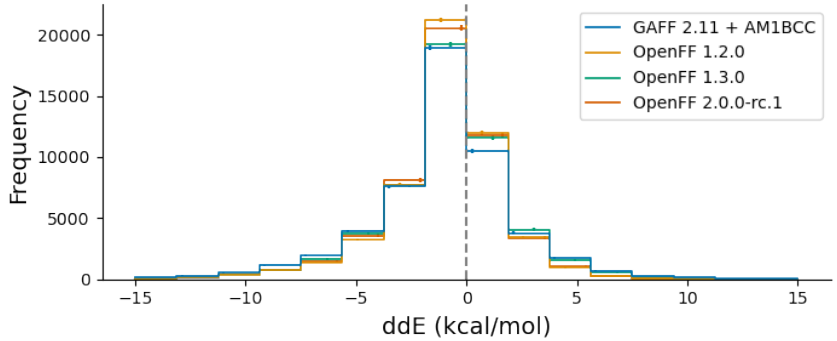
There is not a direct relationship between the accuracy in geometry and the accuracy in relative energies.

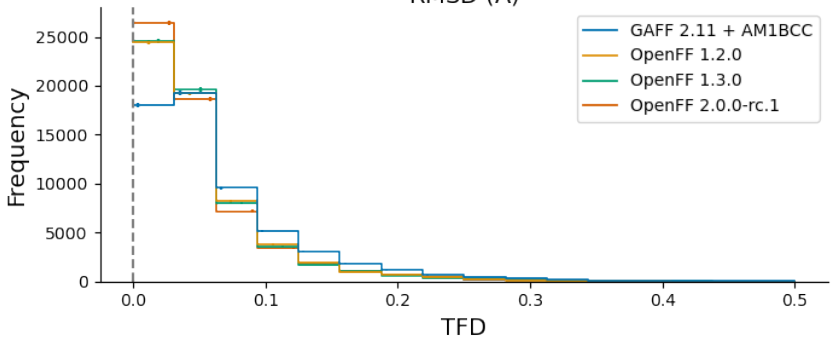
1. https://doi.org/10.26434/chemrxiv.12551867.v2 2. https://pubs.acs.org/doi/full/10.1021/ci2002318 3. https://openforcefield.org/news/benchmark_small_molecules/ 4. Victoria Lim https://youtu.be/RZcAL-SSmZU

We've been doing automated benchmarking in collaboration with our industry partners

- **OpenFF 2.0.0** showed excellent performance when benchmarked against the **Public OpenFF Industry Benchmark Season 1** versus 1.x



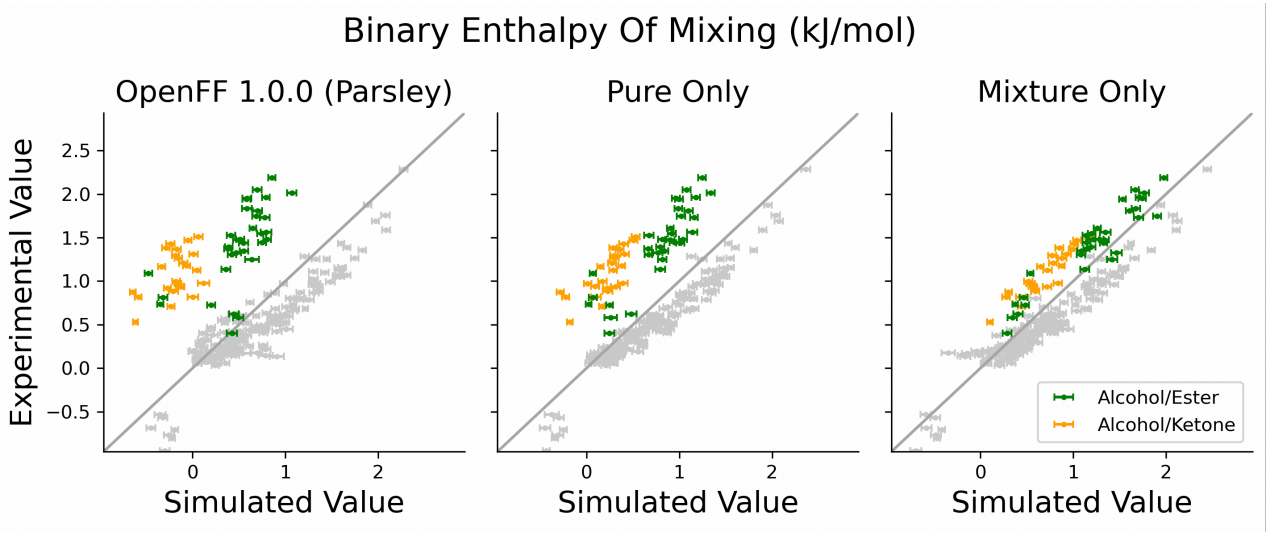


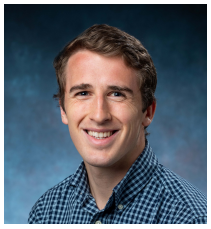


Energies at B3LYP-D3B(BJ) / DZVP basis set

Our work showed that we should train OpenFF 2.0.0 “Sage” LJ parameters against liquid mixtures

- Mixture properties (binary densities, enthalpies of mixing) have advantage over traditional training sets (density, ΔH_{vap})
  - Improved data availability
  - Better for capturing diverse interactions
  - Limited changes in molecular polarization
- Performed a pilot study over a subset of molecules

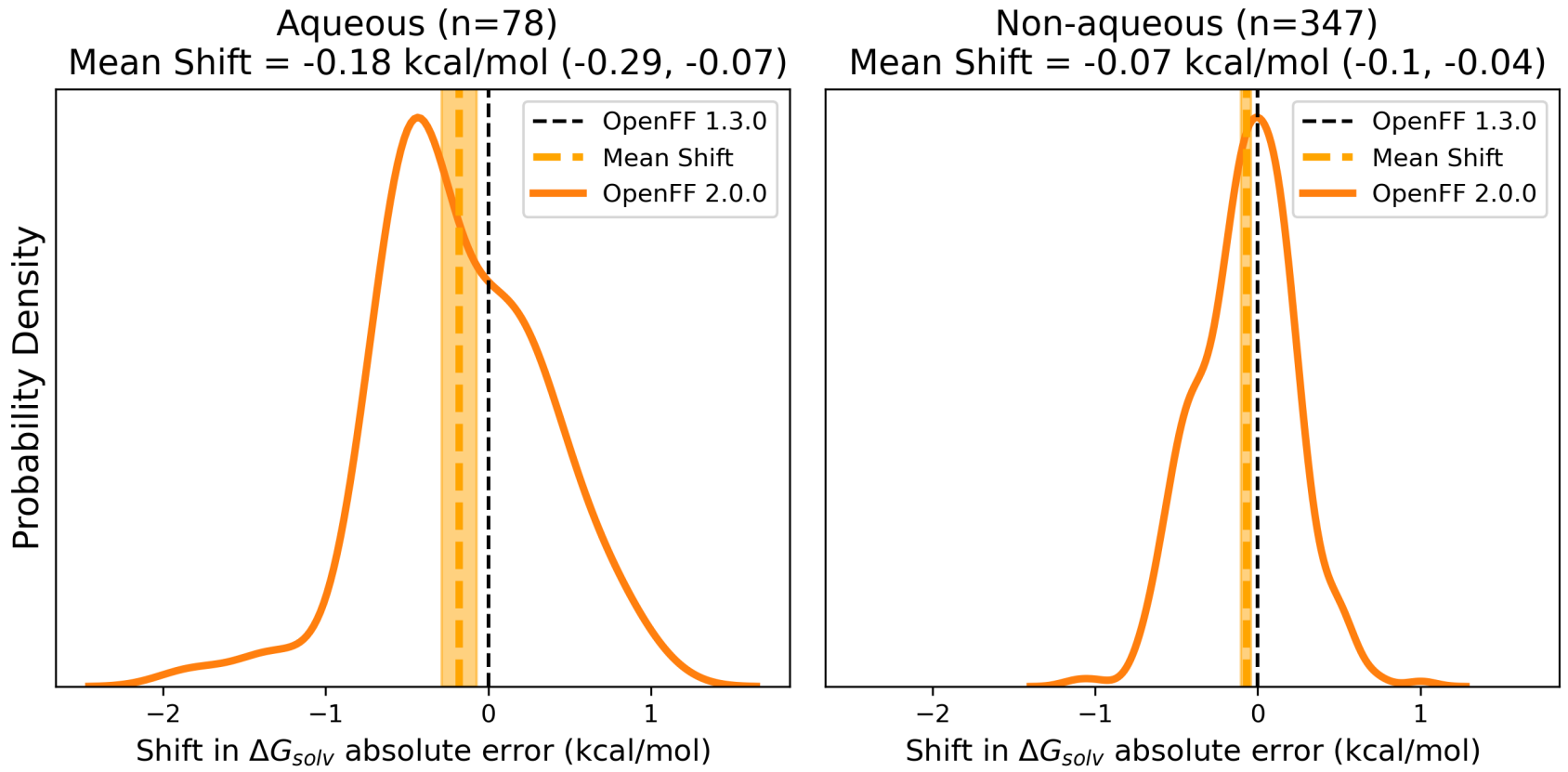




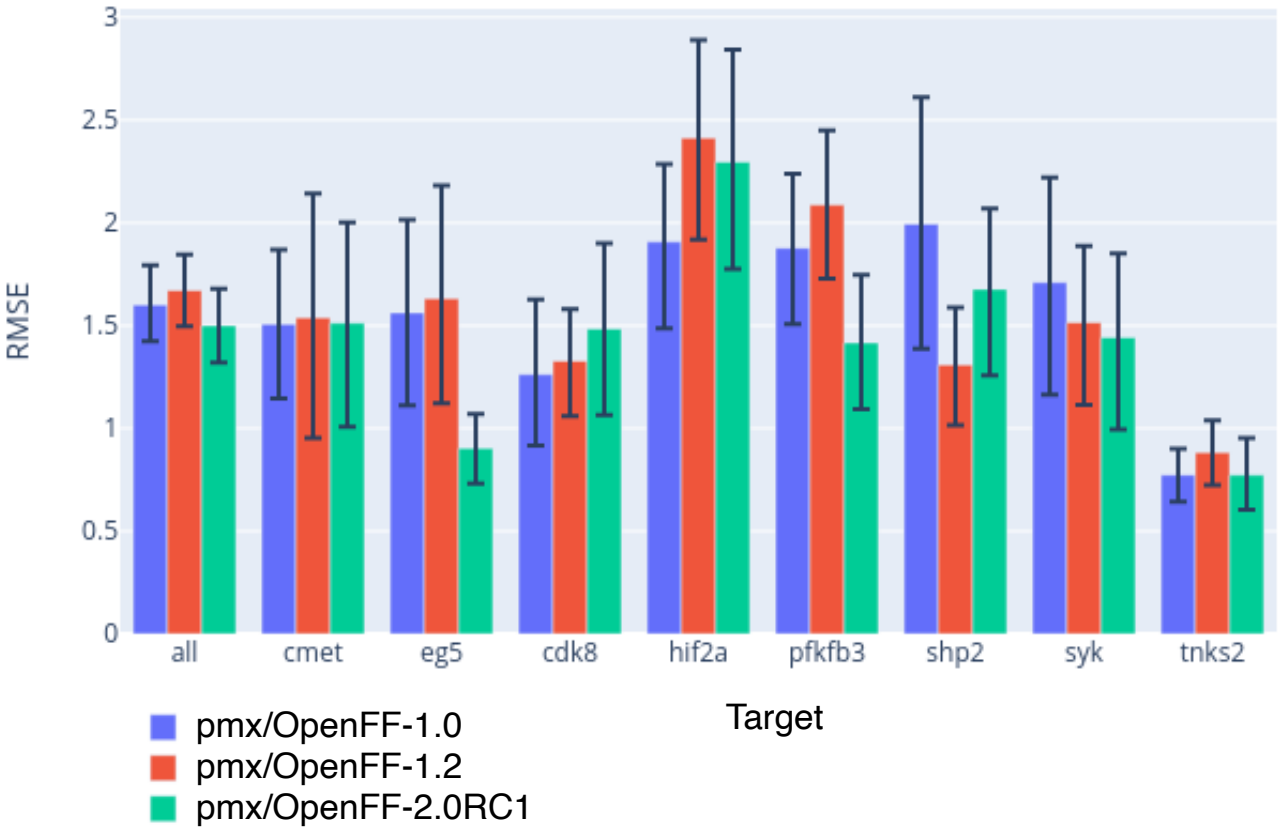
Owen Madin

Simon Boothroyd

Validated improved Lennard-Jones parameters versus solvation free energies.



OpenFF 2.0.0 (Sage) slightly improved results over OpenFF 1.0.0 (Parsley) for protein ligand binding free energies



- RMSE based on ΔΔG in kcal/mol
- Error bars are 95% CI
- OpenFF 2.0.0 (Sage) is generally slightly, but non-significantly better

Vytas Gapsys

Next generation of OpenFF force fields

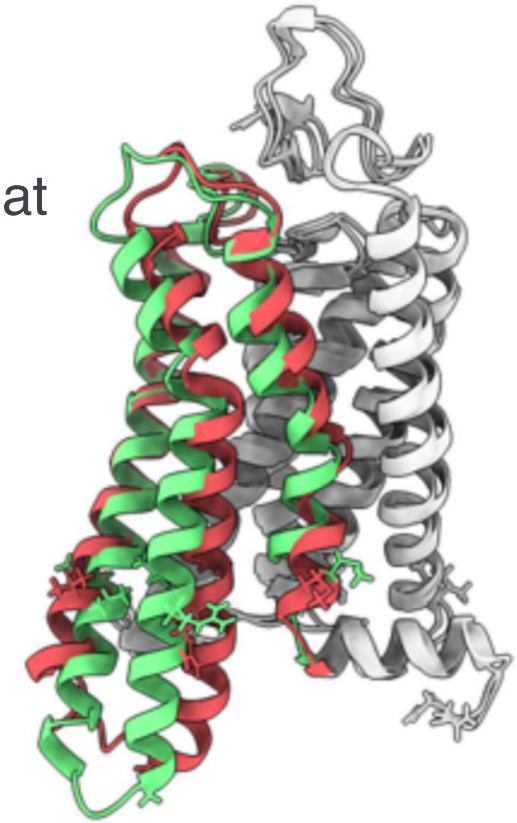
Rosemary (OpenFF 3.x) series:

- In early 2023

- First support of biopolymers (proteins)
- A number of other science goals:
  - Refit electrostatics
  - Off-atom charge sites

COMING UP: Open FF biomolecules for fully consistent small molecule / biomolecule force fields

- If we have great **small molecule** force fields, it should be possible to construct great **biomolecule** force fields.
- May require additional evaluation of torsional potentials that are perhaps overly general
- MORE importantly: How do we know if we have a good protein force field?
- Establishing community benchmarks and experimental datasets:
  - NMR spin relaxation, chemical shifts, NOEs
  - X-ray data from protein crystal simulations



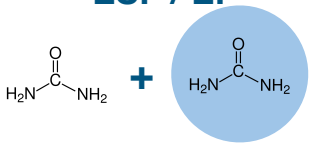
COMING UP: Refit charge models - AM1BCC charge model currently being re-trained against QC and experimental data

33 Double-bonded oxygen in a lactone or lactam

[#8X1$(*=[#6r]@[#7r,#8r]):1]

openff evaluator

ESP / EF



RESP2 δ=0.6

Mixture enthalpies + densities

Lily Wang

Simon Boothroyd

AM1BCC Ported to SMIRNOFF

A majority of the original AM1-BCC parameters have been ported to SMIRNOFF

Integrate Into Fitting Infrastructure

ForceBalance and the OpenFF Evaluator extended to support co-optimising against QC and exp. data

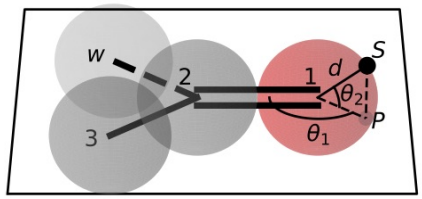
Training

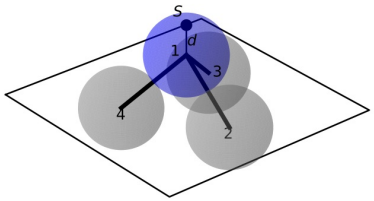
Test fits being performed against a combination of QC ESP / EF data and mixture exp. data

Testing

Benchmarking test fits against experimental solvation / transfer free energy data

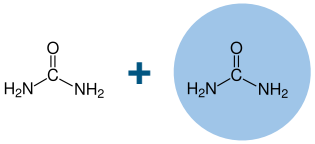
COMING UP: Virtual sites - OpenFF infrastructure implementation enables new science





openff toolkit

ESP / EF



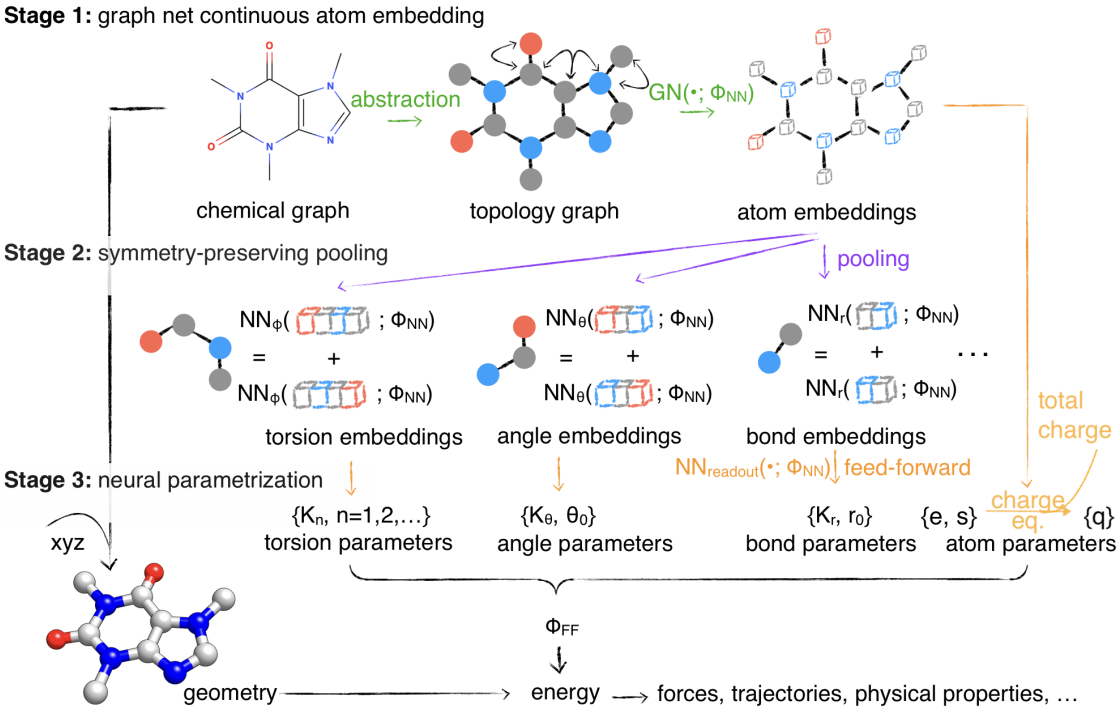
RESP2 δ=0.6

Trevor Gokey

Hypothesis	Software	Training	Testing	Interoperability
The inclusion of off-site charges should improve the accuracy of electrostatic interactions	Virtual site support added to the OpenFF toolkit. Support for training to QC ESP + EF data in progress	Virtual sites will be trained against ESP / EF QC and liquid simulations	Trained parameters benchmarked against experimental and physical property data	Ensuring v-sites are implemented in a way that major simulation packages support

Describing chemical environments using convolutional graph neural networks

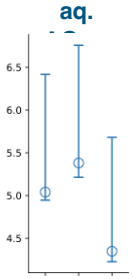
- Chemical environment is more complex than can be described by strings
- Use **convolutional graph neural nets** to describe the environment as vectors
- Then train to energies (or other observables)
- STILL the same functional form, but **continuous** parameters.
- Testing now to fit to AM1-BCC charges
- ESPALOMA proof-of-concept
- *Chem. Sci.*, 2022, **13**, 12016-12033



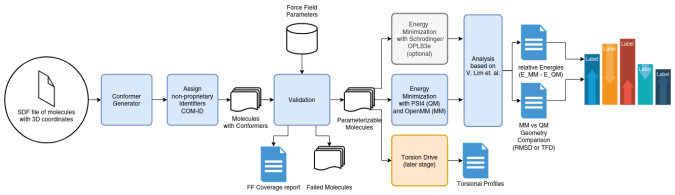
Now Rapidly Approaching:

- Organic polymers that are fully small molecule compatible
- Surrogate modeling for fitting condensed phase properties
- Co-fitted water model
- Making the functional more complex in a data-driven manner
- Bayesian decision-making on complexity of models

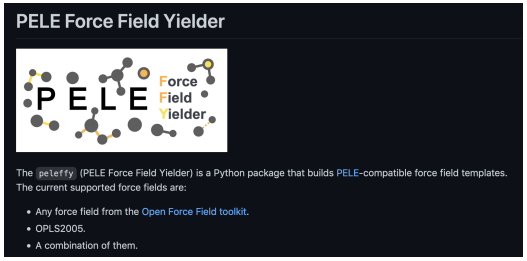
SUMMARY



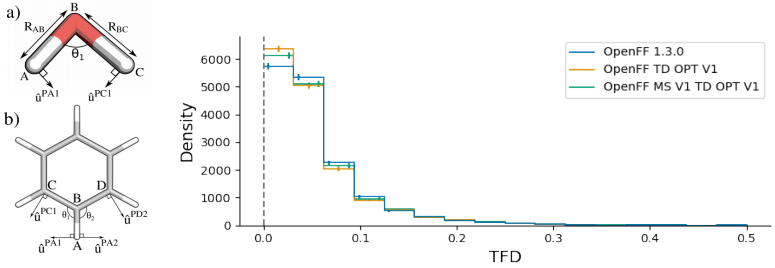
Sage looks even better than we expected!



Automated benchmarking has been a major focus and will point the way forward.



We see community uptake, with and without our help



New technology yields better starting points and better force fields.

New science and infrastructure features coming up!

OpenFF tools are freely available for you to use

- All code, datasets, force fields available online at:

https://github.com/openforcefield

- Keep up to date with our progress and find tutorials:

https://openforcefield.org/

open
forcefield

- You can start fitting your own force fields with our tools **today**!