Benchmarking QM theory for drug-like molecules to train force fields

Pavan K. Behara[1], Hyesu Jang[2,3], Joshua T. Horton[4], David L. Dotson[5,7], Simon Boothroyd[6,7], Chapin E. Cavender[8], Vytautas Gapsys[9], Trevor Gokey[1], David F. Hahn[10], Jessica Maat[1], Owen Madin[13], Ivan J. Pulido[11], Matthew W. Thompson[7], Jeffrey Wagner[7], Lily Wang[1,12], John D. Chodera[11,7], Daniel J. Cole[4,7], Michael K. Gilson[8,7], Michael R. Shirts[13,7], Chris Bayly[3], Lee-Ping Wang[2,7], David L. Mobley[1,7]

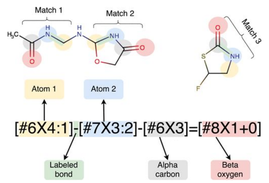
[1]Pharmaceutical Sciences, University of California Irvine, Irvine, CA, USA, [2]Department of Chemistry, University of California Davis, Davis, CA, USA, [3]OpenEye Scientific Software, Santa Fe, NM, USA, [4]School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1, 7RU, United Kingdom, [5]Datryllic LLC, Phoenix, AZ, USA, [6]Boothroyd Scientific Consulting Ltd, London, United Kingdom, [7]Open Force Field Consortium, Open Molecular Software Foundation, Davis, CA, USA, [8]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA, [9]Computational Biomolecular Dynamics Group, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany, [10]Computational Chemistry, Jannsen Research & Development, Turnhoutseweg 30, Beerse B-2340, Belgium, [11]Computational & Systems Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA, [12]Research school of Chemistry, Australian National University, Canberra, ACT 0200, AU, [13]Chemical & Biological Engineering, University of Colorado Boulder, Boulder, CO, USA

open forcefield

@openforcefield
www.openforcefield.org

Open Force Field (OpenFF) Initiative

- The Open Force Field Initiative is a partnership between academic and industry researchers to develop open, reproducible force fields for atomistic simulations.
- SMIRKS-native Open Force Field (SMIRNOFF): Parameters built on direct chemical perception, using substructure queries.



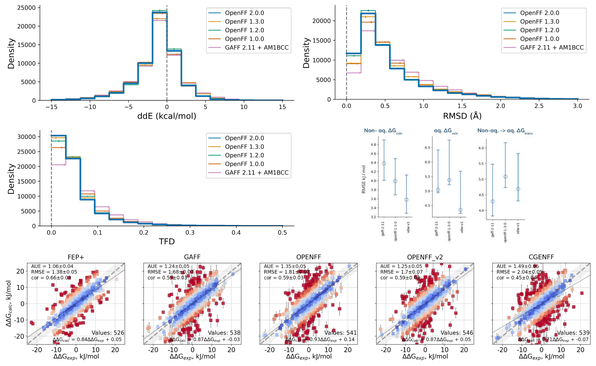
Open Force Fields

PARSLEY (OPENFF 1.0, 2019)
Fit valence parameters against quantum chemical data

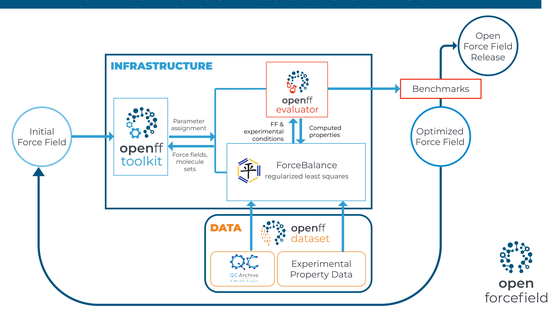
SAGE (OPENFF 2.0, 2021)
Fit lennard-jones parameters against physical properties of binary mixtures

ROSEMARY (OPENFF 3.0, 202x)
Self-consistent force field for biopolymers and small molecules

- Benchmarks against QM data in conformer energetics (ddE), RMSD of geometries, Torsion Fingerprint Deviations (TFD), solvation free energies, relative binding free energies on 546 protein-ligand systems show excellent performance.

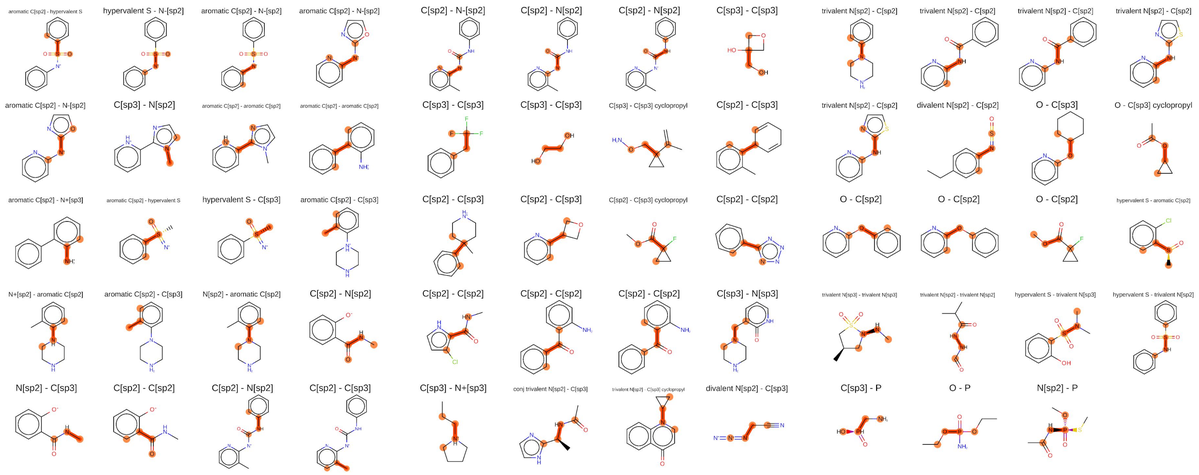


How our force fields are trained?



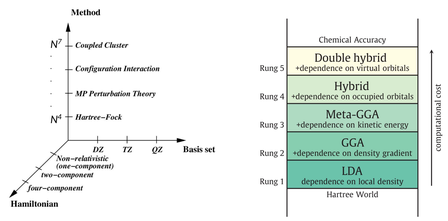
What to train a force field on?

- Valence parameters in a force field describing the angles, bonds, and torsions are trained on QM data.
- Many benchmarks[1-7] viz., Database-2015B, GMTKN55, MGCDB84, are intended for improving the QM method, and use larger basis sets (QZVP or higher).
- In this study 59 drug-like molecules with
  - molecules with non-zero formal charges,
  - with strong internal interactions,
  - with central bond conjugated(< 10 kcal/mol rotational barrier) or
  - with halogen
  - charged molecules with different functional groups
    - -1 charged functional groups: c[O-], C(=O)[N-], c[N-]c, S(=O)(=O)[N-], S(=[N-])(=O)
    - +1 charged functional groups: [NH+,nH+](=,:[C,c])[C,c], [NH+]([*])[*], [NH2+]([*])[*], [NH3+][*]
- Selected one molecule per each group (by picking a center molecule using MACCS keys fingerprint).
- All the calculations are done using Psi4 quantum chemistry package, and data is stored on MolSSI's QCArchive repository.



2D depictions of molecules used in this study and the atoms involved in the torsion are highlighted.

Why is it difficult to pick a theory level?

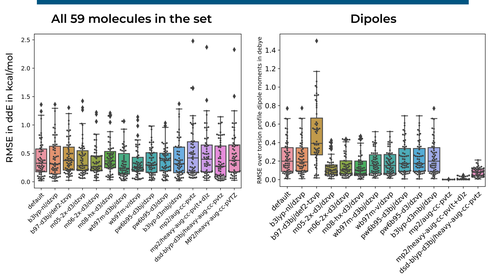


The three axes along which a WFT method may be improved: Basis set, Hamiltonian, and the treatment of electron correlation.
Courtesy: Timo Fleig, urn:nbn:de:hbz:061-20070312-091913-8

Jacob's ladder of density functional approx. for XC energy
- John Perdew's metaphor
Courtesy: 10.1016/j.ccr.2015.03.019

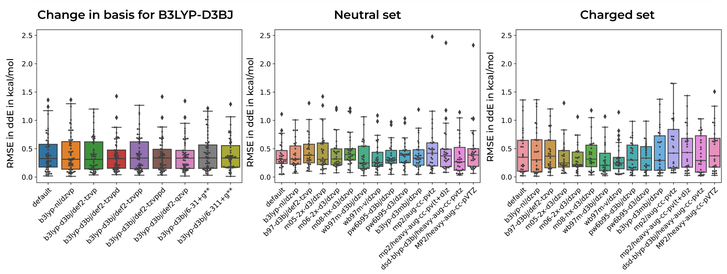
- Earlier study by Yudong Qiu, et al., found B3LYP-D3BJ/DZVP to be optimal for conformer energetics based on MPCONF196 (smaller peptides and medium macrocycles), and YMPJ16(20 amino acids) datasets.

Performance with functionals



With **"df-CCSD(T)/CBS // MP2/heavy-aug-cc-pVTZ"** as a baseline, distribution of RMSEs for a variety of functionals at the same geometries. Baseline for dipoles is **MP2/heavy-aug-cc-pVTZ**.

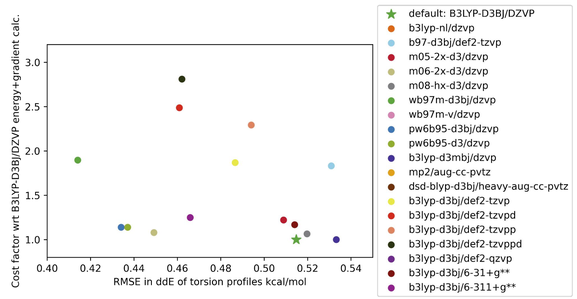
Change in basis, charged/uncharged subsets



Increase in polarization and diffuse functions within B3LYP-D3BJ functional slightly reduces the error. This is in sync with B3LYP-D3BJ/DZVP showing a slightly higher error on charged molecules.

Race of the functionals

- With cost of an energy+gradient calculation as basis, a geometry optimization is ~ 15x expensive, and a torsiondrive ~ 1000x.



Semiempirical methods for Bespoke FF

- Similar benchmark (*without full torsiondrive*) places GFN2-XTB in the first place.
- QM torsion data is a big bottleneck to generate bespoke force fields.
- Josh Horton & Daniel Cole show a promising future where we can build a bespoke force field fast enough based on SQM reference.

SQM	RMSE in ddE kcal/mol
GFN1-XTB	1.53
GFN2-XTB	1.25
ANI2x*	1.60

Conclusions

- **ωB97M-D3BJ/DZVP** is the best among tested functionals with an overall RMSE in ddE of 0.41 kcal/mol in torsion profile energies wrt the baseline.
- RMSE of our current default level, **B3LYP-D3BJ/DZVP**, is 0.51 kcal/mol, which is a great compromise between accuracy and computational cost.

References
1. Folmsbee, Hutchison, DOI: 10.1002/qua.26381, **2.** Lars Goerigk, et al., DOI: 10.1039/C7CP04913G, **3.** Jan Řezáč, et al., DOI: 10.1021/acs.jctc.7b01074, **4.** Narbe Mardirossian, DOI: 10.1080/00268976.2017.1333644, **5.** Manoj Kesharwani, et al., DOI: 10.1021/acs.jctc.5b01066, **6.** Yan Zhao, et al., DOI: 10.1021/ct049851d, **7.** Benjamin Sellers, et al., DOI: 10.1021/acs.jcim.6b00614

Upcoming releases to look out for!!!

- **Software permissively licensed under the MIT License and developed openly on GitHub, https://github.com/openforcefield.**
- All packages are conda installable and many tutorials available, please use and raise any issues/bugs.
- **OpenFF Toolkit v0.11.0 (on the horizon):** This will permit preparation of molecular topologies and parameter assignment for systems containing both small molecules and biopolymers — including those with covalent modifications — and will write to common molecular dynamics formats.
- **OpenFF Interchange (ready for testers):** export from OpenFF Toolkit to several biomolecular simulation formats and vectorized representations via without going through ParmEd.
- **OpenFF Bespokefit (ready for testers):** tool for the generation of bespoke SMIRNOFF format parameters for individual molecules.

Partners in Science

MolSSI · Psi4 Open-Source Quantum Chemistry · OpenMM · RDKit · UCI Brilliant Future · OpenEye Scientific · Open Molecular Software Foundation · NIH National Institutes of Health — Turning Discovery Into Health