Parameterization perspective I:
Parameter optimization methodology

Lee-Ping Wang, Yudong Qiu, Simon Boothroyd, Daniel Smith

Aug 2019 | **#release-1, #forcebalance, #torsions, #valence, #propertyestimator** on Slack

As of this week, our first major round of parameter optimization is complete

- Current version of optimized force field available online is our first release candidate
- Numerical versioning scheme: X.Y.Z

  X = major release

  Y = minor release

  Z = bugfix
- If we make it to four major releases, they could be called "parsley, sage, rosemary & thyme"

1.Are you go-ing to Scar - bo-rough Fair?
Pars-ley, sage, rose - mar - y and thyme.

OpenFF “parsley” 1.0.0-RC1: Optimized parameters

At a glance: Which parameters were optimized?

Bond stretching: Equilibrium lengths and force constants	172	parameters
Angle bending: Equilibrium angles and force constants	74	parameters
Torsions: Barrier heights (phases not optimized)	254	parameters
Lennard-Jones σ and ϵ parameters	30	parameters
Total:	530	parameters

At a glance: What data did we use?

QM data generation: Yudong Qiu & Daniel Smith
Experimental data curation: Simon Boothroyd

- Valence (bond & angle): QM optimized geometries and calculated vibrational frequencies
  1785 optimized structures
  895 sets of frequencies
- Torsion: QM torsion drives (Energy vs. torsion profiles of constrained optimized QM geometries)
  1086 torsion drives (15° resolution)
- Lennard-Jones: Density and ΔH_{vap} of molecular liquids
  39 liquid density measurements
  19 ΔH_{vap} measurements

OpenFF “parsley” 1.0.0-RC1: Optimized parameters

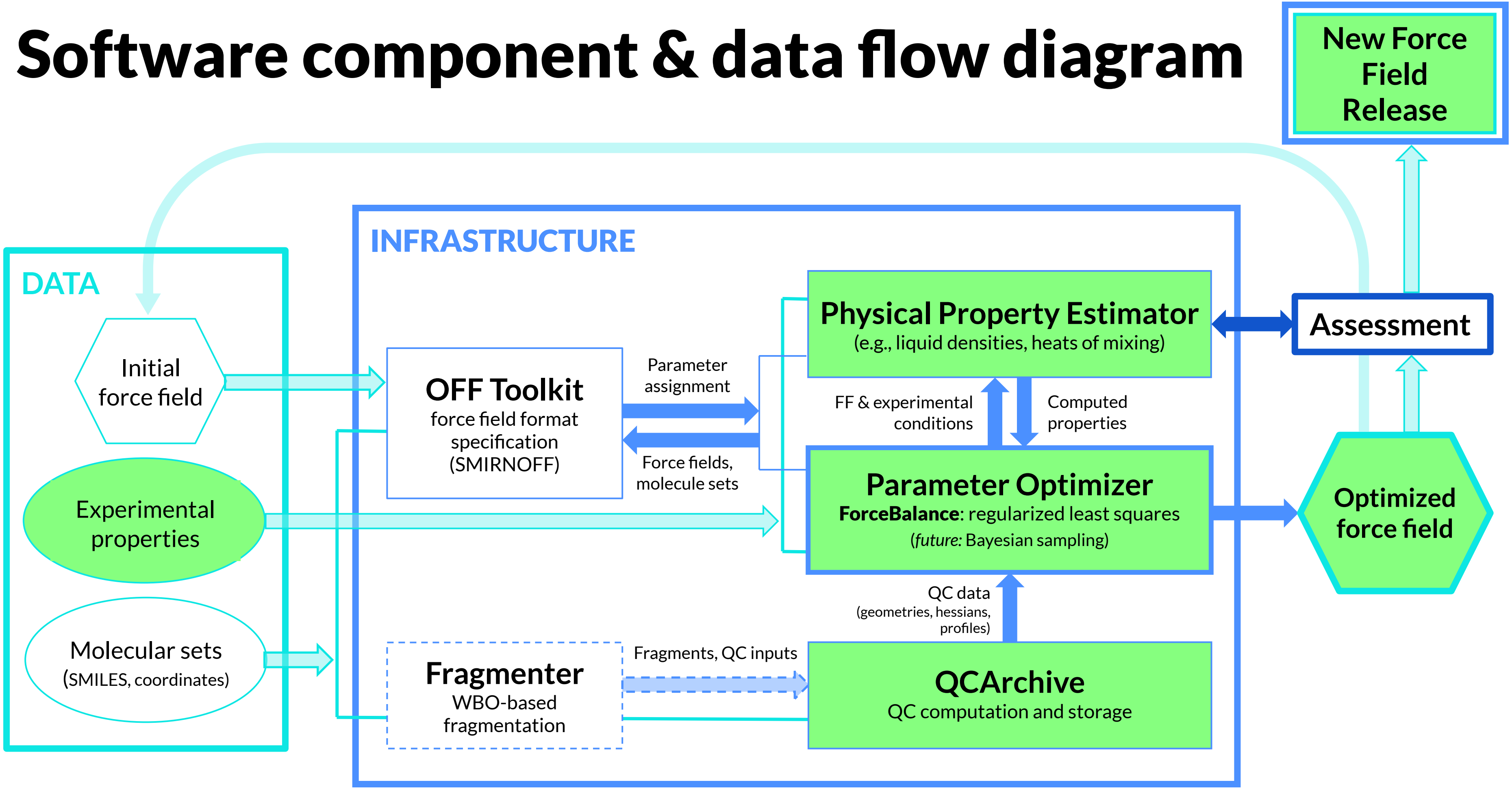
At a glance: How were the parameters optimized?

- Start from the SMIRNOFF99Frosst parameter set adopted from AMBER99 and parm@Frosst
- Regularized, nonlinear least-squares optimization as implemented in *ForceBalance* software
- Parameters were optimized in three major stages:
  1) Fitting valence and torsion parameters to QM calculations (Yudong Qiu)
  2) Keeping (1) params frozen, fit LJ parameters to thermodynamic properties (Simon Boothroyd)
  3) Keeping (2) LJ params frozen, refit valence and torsion parameters to QM calculations

Optimized parameters, fitting data and optimization output is currently located at:
https://github.com/lpwgroup/forcebalance-qcarchive/releases

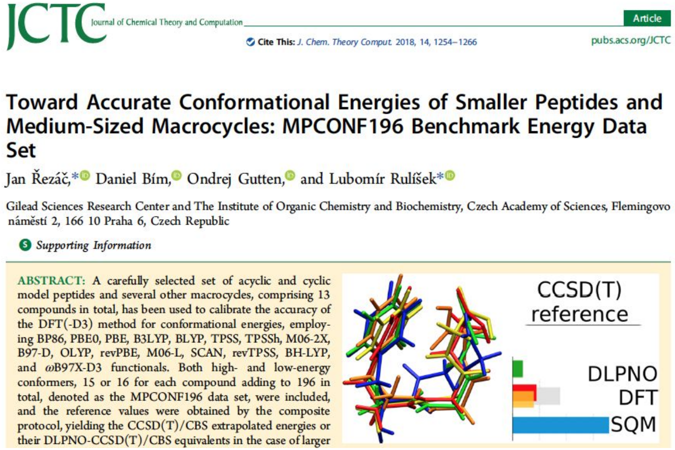
Force field provided in .offxml format, ready for simulations
Repository includes detailed release notes for each parameterization run
Downloadable files includes plots and analysis of fine optimization details

Software component & data flow diagram



New Force Field Release
DATA
Initial force field
Experimental properties
Molecular sets
(SMILES, coordinates)
INFRASTRUCTURE
OFF Toolkit
force field format specification (SMIRNOFF)
Parameter assignment
Force fields, molecule sets
Physical Property Estimator
(e.g., liquid densities, heats of mixing)
FF & experimental conditions
Computed properties
Parameter Optimizer
ForceBalance: regularized least squares
(*future:* Bayesian sampling)
QC data
(geometries, hessians, profiles)
Fragmenter
WBO-based fragmentation
Fragments, QC inputs
QCArchive
QC computation and storage
Assessment
Optimized force field

Selection of QM level of theory



JCTC Journal of Chemical Theory and Computation — Article

Cite This: *J. Chem. Theory Comput.* 2018, 14, 1254−1266 — pubs.acs.org/JCTC

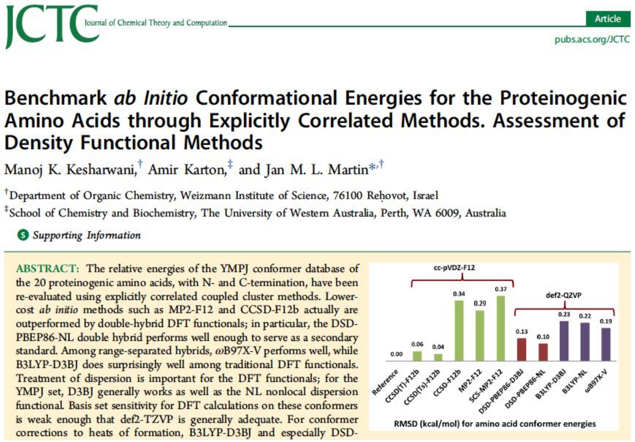
Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set

Jan Řezáč,* Daniel Bím, Ondrej Gutten, and Lubomír Rulíšek*

Gilead Sciences Research Center and The Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Flemingovo náměstí 2, 166 10 Praha 6, Czech Republic

Supporting Information

ABSTRACT: A carefully selected set of acyclic and cyclic model peptides and several other macrocycles, comprising 13 compounds in total, has been used to calibrate the accuracy of the DFT(-D3) method for conformational energies, employing BP86, PBE0, PBE, B3LYP, BLYP, TPSS, TPSSh, M06-2X, B97-D, OLYP, revPBE, M06-L, SCAN, revTPSS, BH-LYP, and ωB97X-D3 functionals. Both high- and low-energy conformers, 15 or 16 for each compound adding to 196 in total, denoted as the MPCONF196 data set, were included, and the reference values were obtained by the composite protocol, yielding the CCSD(T)/CBS extrapolated energies or their DLPNO-CCSD(T)/CBS equivalents in the case of larger



JCTC Journal of Chemical Theory and Computation — Article

pubs.acs.org/JCTC

Benchmark *ab Initio* Conformational Energies for the Proteinogenic Amino Acids through Explicitly Correlated Methods. Assessment of Density Functional Methods

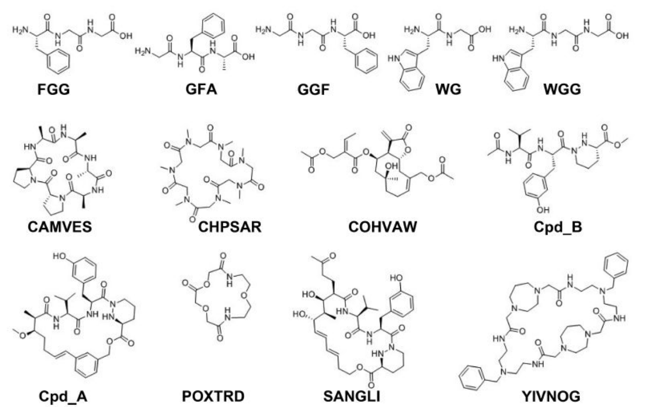
Manoj K. Kesharwani,[†] Amir Karton,[‡] and Jan M. L. Martin*[†]

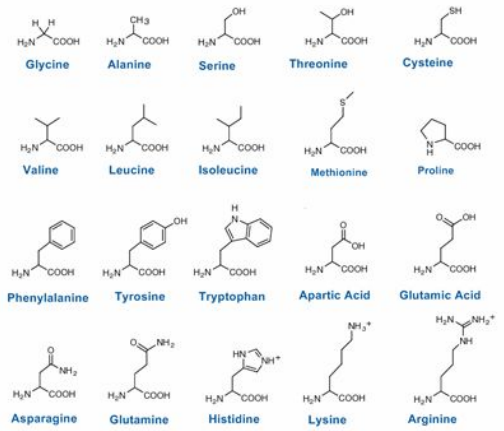
[†]Department of Organic Chemistry, Weizmann Institute of Science, 76100 Reḥovot, Israel
[‡]School of Chemistry and Biochemistry, The University of Western Australia, Perth, WA 6009, Australia

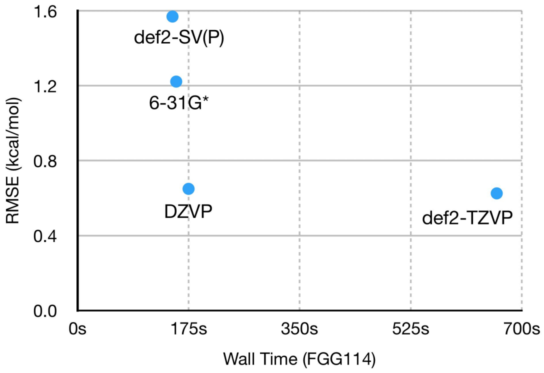
Supporting Information

ABSTRACT: The relative energies of the YMPJ conformer database of the 20 proteinogenic amino acids, with N- and C-termination, have been re-evaluated using explicitly correlated coupled cluster methods. Lower-cost *ab initio* methods such as MP2-F12 and CCSD-F12b actually are outperformed by double-hybrid DFT functionals; in particular, the DSD-PBEP86-NL double hybrid performs well enough to serve as a secondary standard. Among range-separated hybrids, ωB97X-V performs well, while B3LYP-D3BJ does surprisingly well among traditional DFT functionals. Treatment of dispersion is important for the DFT functionals; for the YMPJ set, D3BJ generally works as well as the NL nonlocal dispersion functional. Basis set sensitivity for DFT calculations on these conformers is weak enough that def2-TZVP is generally adequate. For conformer corrections to heats of formation, B3LYP-D3BJ and especially DSD-

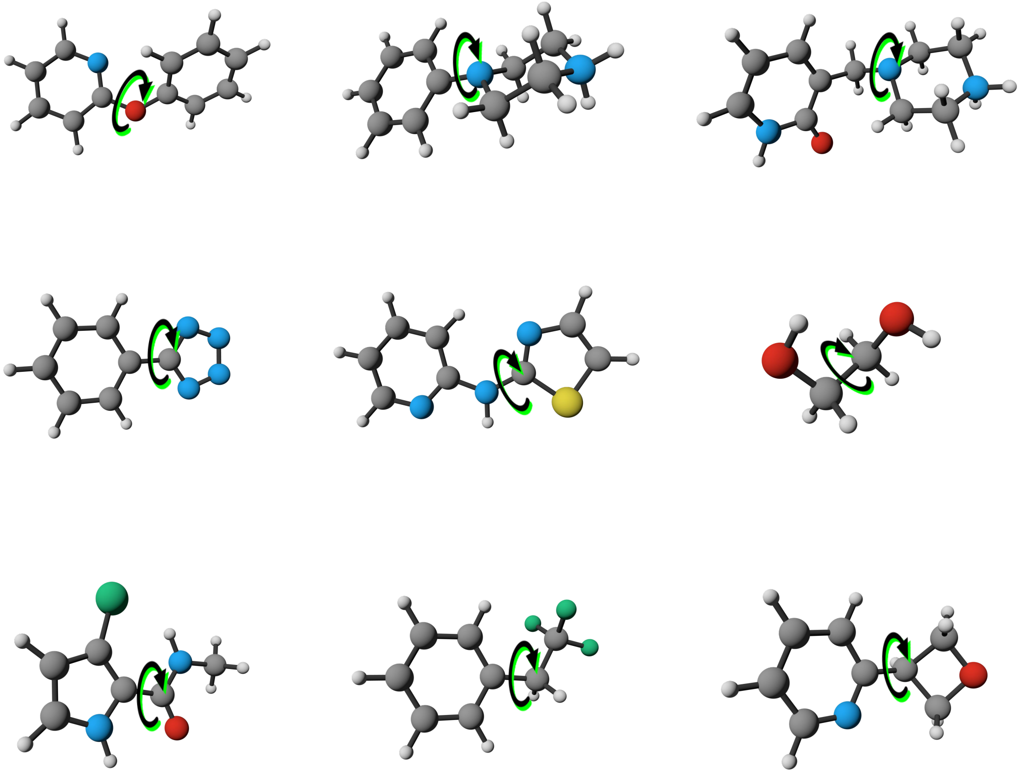
- Based on published benchmark studies of conformational energies.
- B3LYP-D3(BJ) functional and Salahub's DZVP basis gives good compromise between accuracy & cost







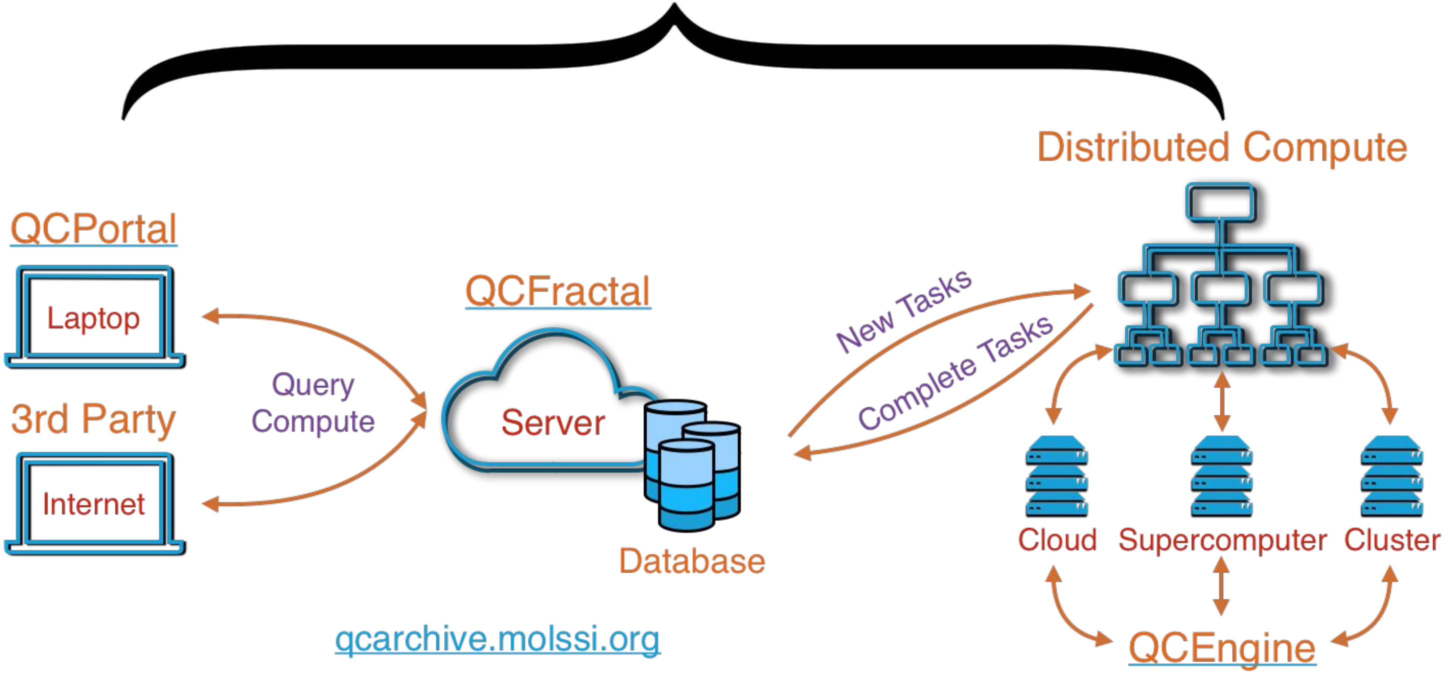
Molecules used in QC calculations



- We started with a set of 468 small molecules provided by Roche
- 820 rotatable bonds involving 4 heavy atoms not in rings; after filtering for intramolecular hydrogen bonds, ended up with 669
- A “coverage set” of molecules was created (David Mobley) to ensure full coverage of the SMIRNOFF parameters, leading to 417 more torsion drives.
- A total of 1,785 optimized conformers were generated from these two sets of molecules; freq. calculations @ lowest minimum.

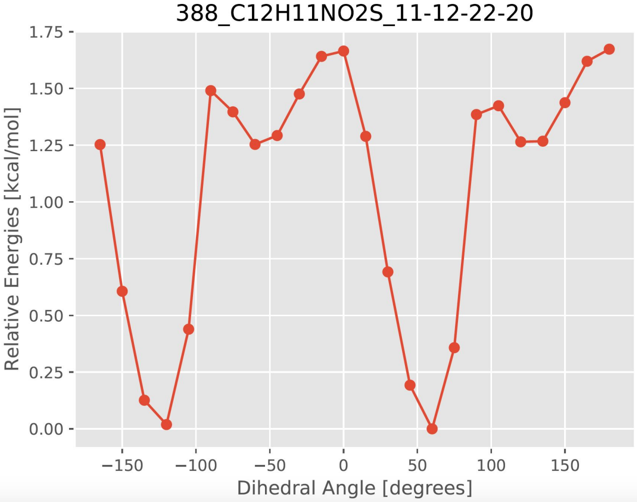
QCArchive

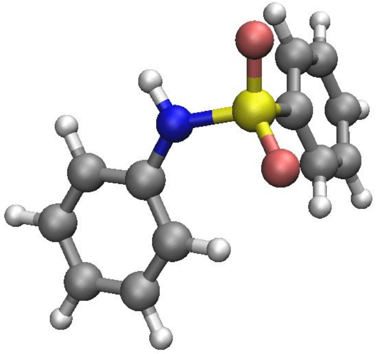
A MolSSI Project



qcarchive.molssi.org

Quantum chemistry on QCArchive

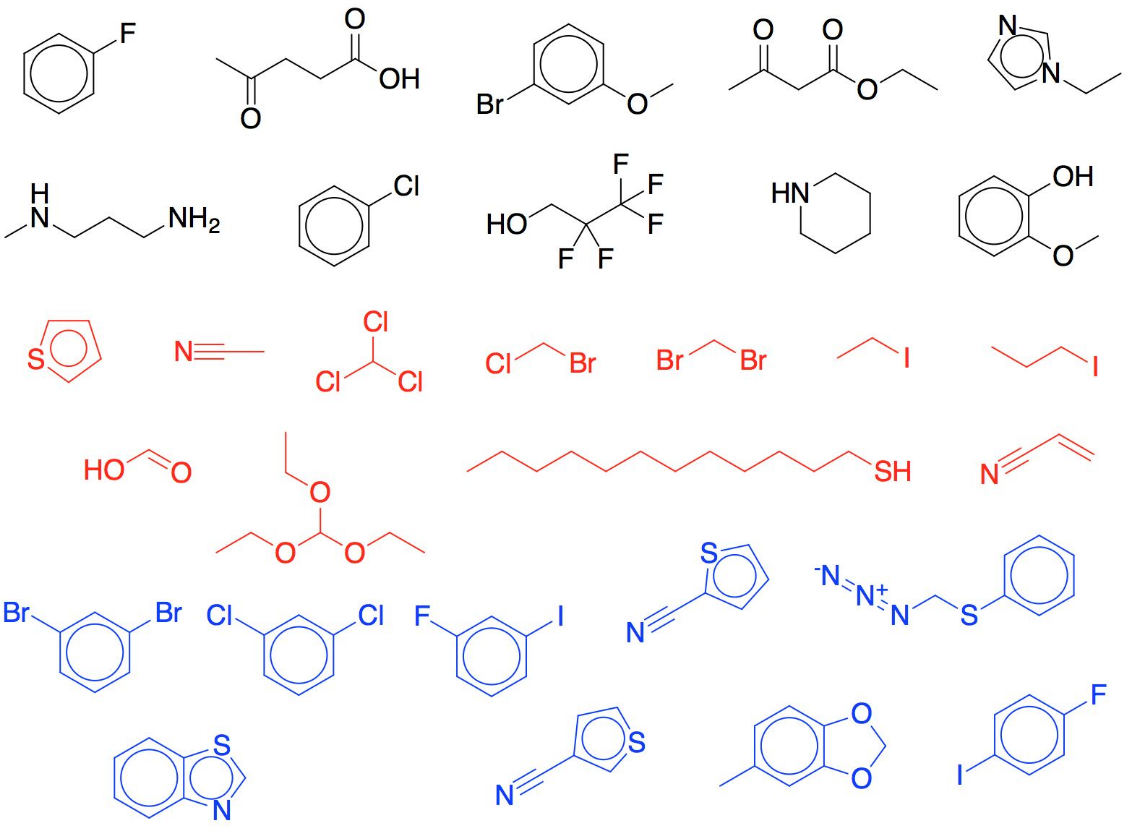




- All QM calculations were done inside *QCArchive* ecosystem
- *QCArchive* executes *torsiondrive* service which performs wavefront propagation of constrained optimizations
- Constrained optimizations carried out by *geomeTRIC* package calling *Psi4* for energies and gradients; >250k optimizations run, zero optimization convergence errors
- *QCArchive* also implements unconstrained geometry optimization and Hessian calculations, used to inform valence parameters
- Completed calculations were downloaded from *QCArchive* and converted into *ForceBalance* readable formats

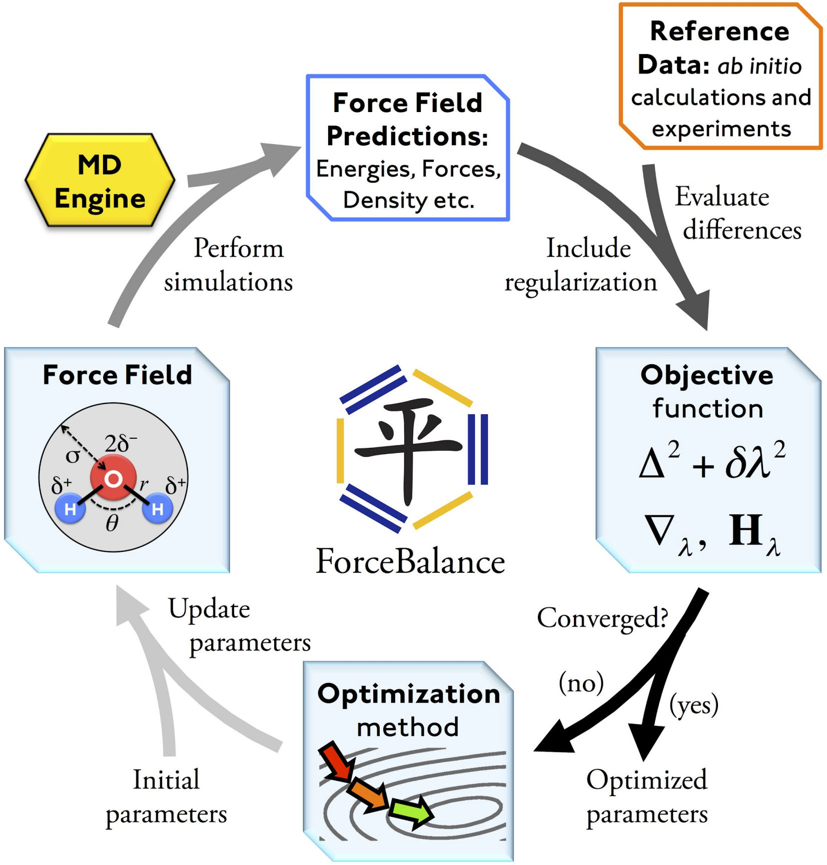
Selection of experimental data

Density only; ΔH_{vap} only; has both



- Molecules with experimental thermodynamic property data identified from *ThermoML* database (covers data published in last 10 years)
- Focus on small compounds with good parameter coverage and availability of density and ΔH_{vap}
- Selected set of molecules covers 15 out of 35 SMIRNOFF nonbonded types (5 H, 2 C, 3 O, 1 N, 1 F, 1 S, 1 Cl, 1 Br)

ForceBalance is a force field optimization tool



- Python toolkit with a main executable ForceBalance for carrying out optimizations.
- Designed for flexibility, FB allows the user to optimize force fields using a wide range of:
  (1) Functional forms
  (2) Reference data (QM or expt.)
  (3) MM simulation software
- Designed for reproducibility, FB enables systematic improvement of models by adding data or physical detail to previous runs.
- Freely available for commercial use via 3-clause BSD license.
- Software distribution comes with 20+ example calculations plus all data sets (including those used in published work)

Theory of force field parameter updates

Force field parameters vary across many orders of magnitude and may obey complex functional relationships and constraints.

In FB, the optimization algorithm sees an array of ***mathematical parameters*** that are well-behaved, i.e. are fully unconstrained and are on order 1.

Physical parameters are related to mathematical parameters by shifting and scaling:

Physical parameters = Initial values + Rescaling matrix × Mathematical parameters

$$\mathbf{k}_{\mathrm{phys}} = \mathbf{k}_{\mathrm{phys}}^{(0)} + \mathbf{T}\mathbf{k}_{\mathrm{math}}$$

$$\mathbf{T} = \begin{pmatrix} 0.01\ \mathrm{nm} & 0 & 0 & 0 \\ 0 & 10^5 \dfrac{\mathrm{kJ}}{\mathrm{mol\ nm}} & 0 & 0 \\ 0 & 0 & 10^\circ & 0 \\ 0 & 0 & 0 & 10^2 \dfrac{\mathrm{kJ}}{\mathrm{mol\ rad}} \end{pmatrix}$$

Rescaling matrix consists of ***prior widths*** on diagonal representing the size of expected changes over the optimization (or over parameters of the same type).

Typically, one prior width should be specified for each parameter type (fewer than 10 independent user-specified values).

Theory of force field parameter updates

The rescaling matrix **T** is almost always diagonal; off-diagonal could be used to constrain net charges on molecules to stay constant.

More generally, ***evaluated parameters*** may be defined as any mathematical function of physical parameters:

Full set of parameters comprises:

Physical parameters,

Evaluated parameters,

and deeper evaluated parameters if any...

$$\mathbf{k}_{\mathrm{full}} = \mathbf{k}_{\mathrm{phys}} \oplus \mathbf{k}^{(0)}_{\mathrm{eval}}\left(\mathbf{k}_{\mathrm{phys}}\right) \oplus \mathbf{k}^{(1)}_{\mathrm{eval}}\left(\mathbf{k}_{\mathrm{phys}};\mathbf{k}^{(0)}_{\mathrm{eval}}\right) \oplus \cdots$$

“Physical” & “evaluated” parameters may be used as scratch variables not to be read by the MM software. Thus, parameters that are actually used by the MM software can be defined such that they obey almost any desired mathematical relationship - such as summing up to a constant, restricted to within a range, or obeying a geometric / trigonometric relationship.

Theory of objective function

The objective function is a weighted sum of least-squares contributions called *targets* plus regularization:

Objective function

Weighted sum over targets

Regularization

user-specified *w*, unity usually sufficient

$$L_{\text{tot}}\left(\mathbf{k}_{\text{math}}\right)=\sum_{i\in\text{targets}} w_i L_i\left(\mathbf{k}_{\text{math}}\right)+w_{reg}\left|\mathbf{k}_{\text{math}}\right|^2$$

Each target is a weighted sum of least-squares contributions for one or more properties:

User-specified weights for properties (unity usually sufficient)

$$L_i\left(\mathbf{k}_{\text{math}}\right)=\sum_{j\in\text{properties}} w_{ij} L_{ij}\left(\mathbf{k}_{\text{math}}\right)$$

Each property is a weighted and normalized sum over individual data points:

Overall normalization to remove units

Weighted, normalized sum over data points (uniform or automatic weights)

$$L_{ij}\left(\mathbf{k}_{\text{math}}\right)=\frac{1}{d_{ij}^2}\frac{\sum_{p\in\text{points}} w_{ijp}\left|y_{ijp}\left(\mathbf{k}_{\text{math}}\right)-y_{ijp}^{(\text{ref})}\right|^2}{\sum_{p\in\text{points}} w_{ijp}}$$

Theory of optimization algorithm

The matrix of second derivatives (*Hessian*) of a least-squares objective function can be estimated if the first derivatives of residuals are known (Gauss-Newton approximation):

$$H_{pq}\left[\mathbf{k}_{\text{math}}\right]=\frac{\partial^2}{\partial k_p\,\partial k_q}\left|y\left(\mathbf{k}_{\text{math}}\right)-y^{(ref)}\right|^2=2\frac{\partial y}{\partial k_q}\frac{\partial y}{\partial k_p}+\ 2\underbrace{\frac{\partial^2 y}{\partial k_p\,\partial k_q}}_{\text{approximate as zero}}$$

This enables highly efficient quasi-Newton optimization algorithms to be used.

$$\mathbf{k}_{\text{math}}^{(i+1)}=\mathbf{k}_{\text{math}}+\left(\mathbf{H}_{\text{approx}}\left[\mathbf{k}_{\text{math}}\right]+\lambda\mathbf{I}\right)^{-1}$$

The λ parameter is used to restrict the optimization step to lie within a trust radius (which is adjusted on-the-fly based on step quality), or it can be used in line-search minimization to determine the next step.

FB implements BFGS Hessian updating algorithm as an alternative, less efficient approach.

New targets in ForceBalance for QM fitting

New optimized geometry target

- Involves matching MM optimized structure to QM optimized structure
- Geometries are expressed as internal coordinates, then (MM – QM) differences are scaled: (bond 0.05 Å, angle 8°, improper torsion 20°)
- Each molecule contributes 1–10 to the total objective function

New torsion profile target

- Involves matching MM energy to QM energy along torsion profile
- MM structures are minimized with 1 kcal/mol/Å harmonic restraints (torsion atoms frozen) prior to comparison with QM
- MM energies are referenced to lowest energy structure in QM

OpenMM implementation of vibrational frequency target

- Uses OpenMM to compute vibrational frequencies and vibrational modes
- Matching of vibrational modes was not performed due to numerous pathological cases
- Use of internal coordinate Hessian is planned for a future update (Cartesian to IC Hessian conversion codes have been implemented)

PropertyEstimator for physical properties

PropertyEstimator (Simon Boothroyd)

- Software toolkit for simulation of physical properties
- Drop-in replacement for ForceBalance native property calculation codes
- A platform for improved performance and improved methods for rapid & robust estimation

FB & PropertyEstimator interface (Yudong Qiu & Simon Boothroyd)

- FB asks PropertyEstimator for thermodynamic properties and gradients (computed using thermodynamic fluctuation formulas)
- PE-provided quantities are used to build FB objective function for optimization and associated gradients and approximate Hessian
- As part of this implementation, FB now uses the OpenFF toolkit to set parameters using the API.

Basic ForceBalance workflow

1) Add special comments to force field XML file indicating parameters to be optimized:

```
<Bond smirks="[#6X3:1]-[#6X3:2]" length="1.45 * angstrom" k="820.0 * angstrom**-2 *
mole**-1 * kilocalorie" id="b4" parameterize="k,length"/>
```

2) Create parameterization target folders containing theoretical and/or experimental data

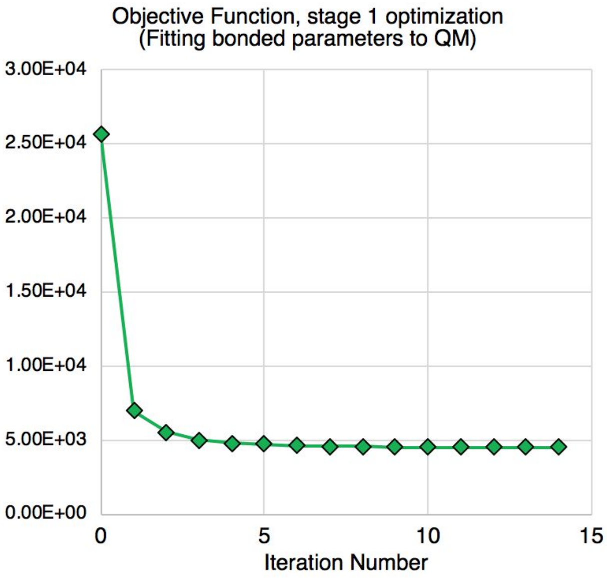
```
$ ls targets/
optimized_geometry_1  optimized_geometry_2  vibrational_frequency_1  properties_1 ...
```

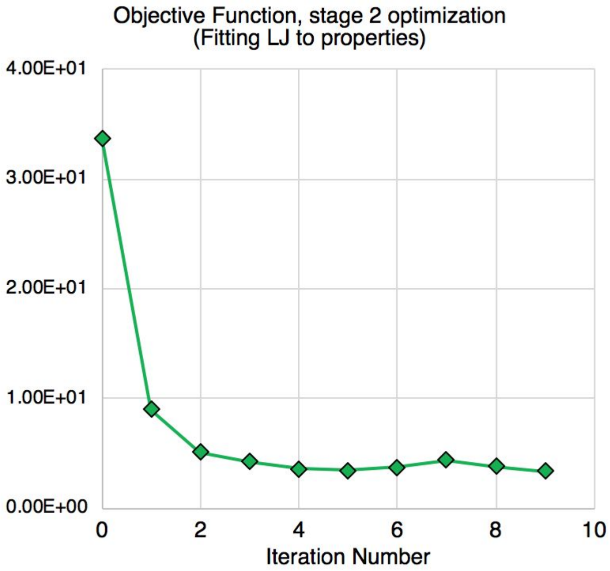
3) Specify calculation settings using input file

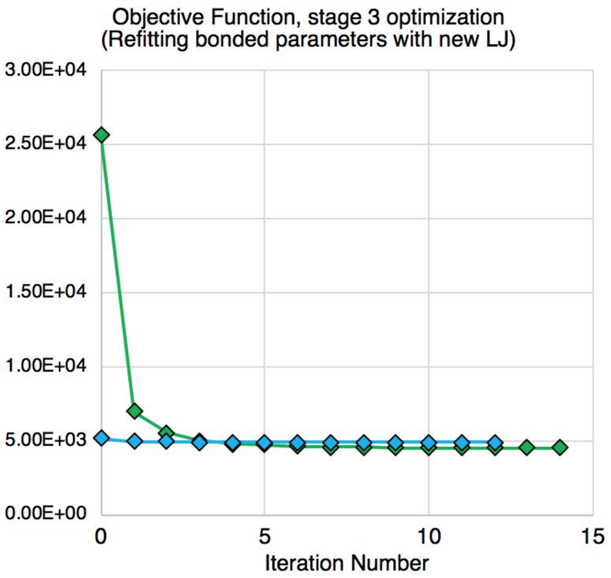
```
$options
jobtype optimize
forcefield fit_bonds_angles.offxml
...
```

4) Run and wait for results. For large jobs, distributed computing is supported.

Results: Overall convergence behavior



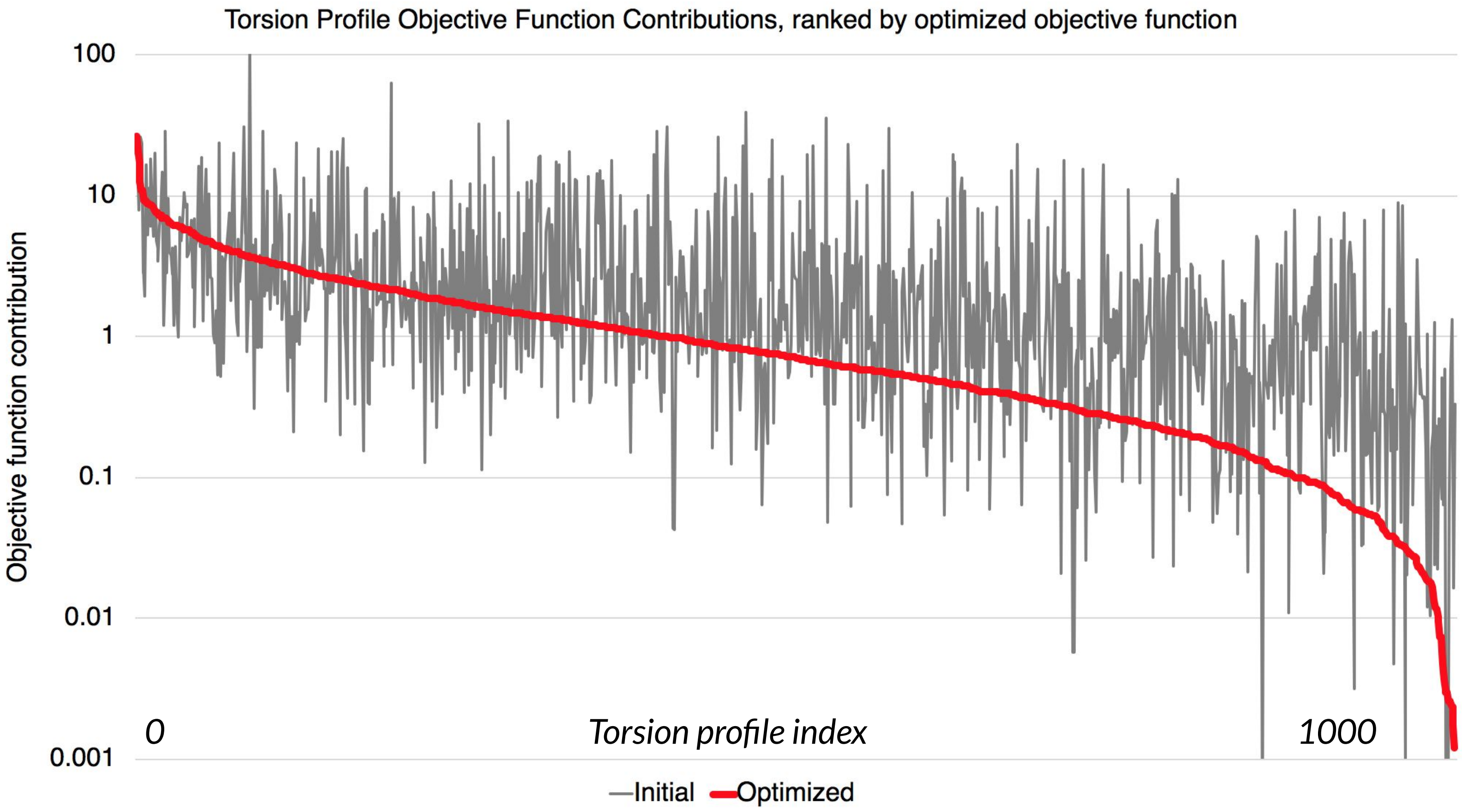




Overall optimization characteristics:

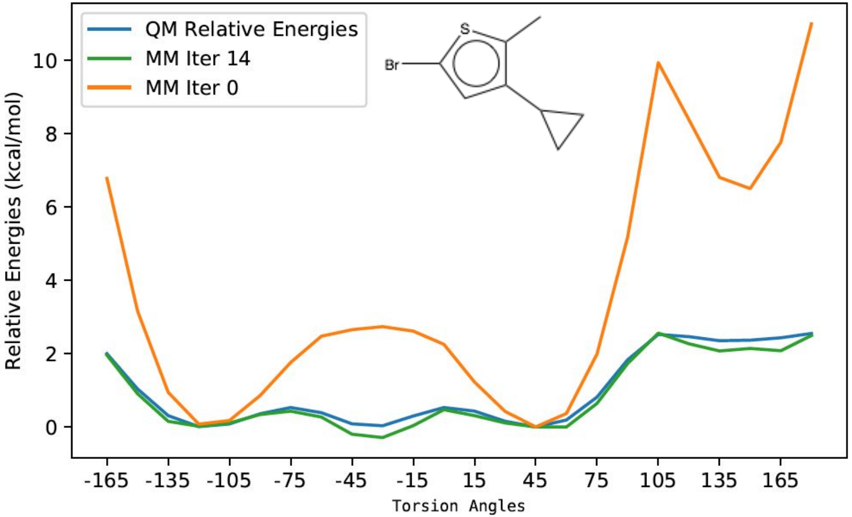
- Optimization “converges” within 10-15 nonlinear cycles when fitting to QM data
- Fluctuations in thermodynamic properties prevent tight convergence in stage 2 (manually stopped)
- Stage 3 optimization (with optimized LJ) has slightly higher final objective function than stage 2

Results: Fitting of torsion profiles



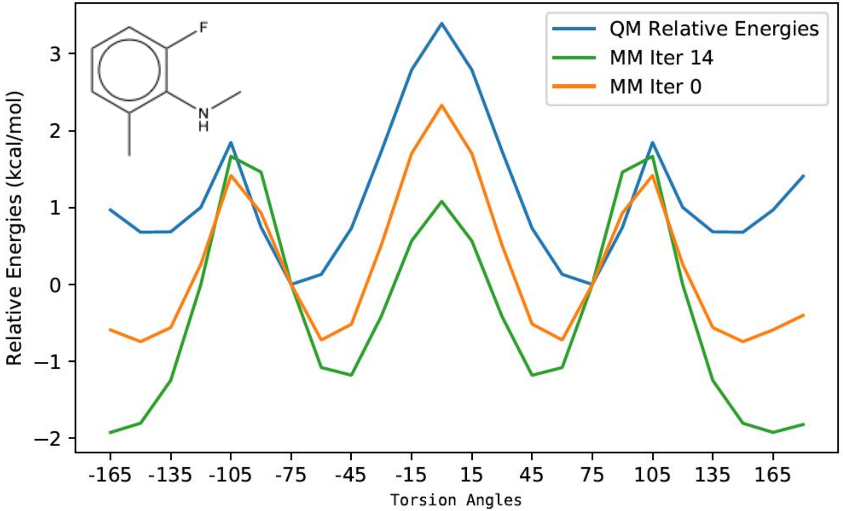
Results: Fitting of torsion profiles

Major improvement after optimization:



Dataset Name	SMIRNOFF Coverage Torsion Set 1
Entry Label	Cc1[c:2]([cH:1]c(s1)Br)[CH:3]2[CH2:4]C2
Canonical SMILES	Cc1c(cc(s1)Br)C2CC2
Torsion Atom Indices	[0, 1, 6, 4]
Torsion SMIRKs	[#6X4;r3:1]-;@[#6X4;r3:2]-[#6X3;r5:3]-;@[#6X3;r5:4]
Torsion SMIRKs ID	t36
SMIRKs Total Count	5

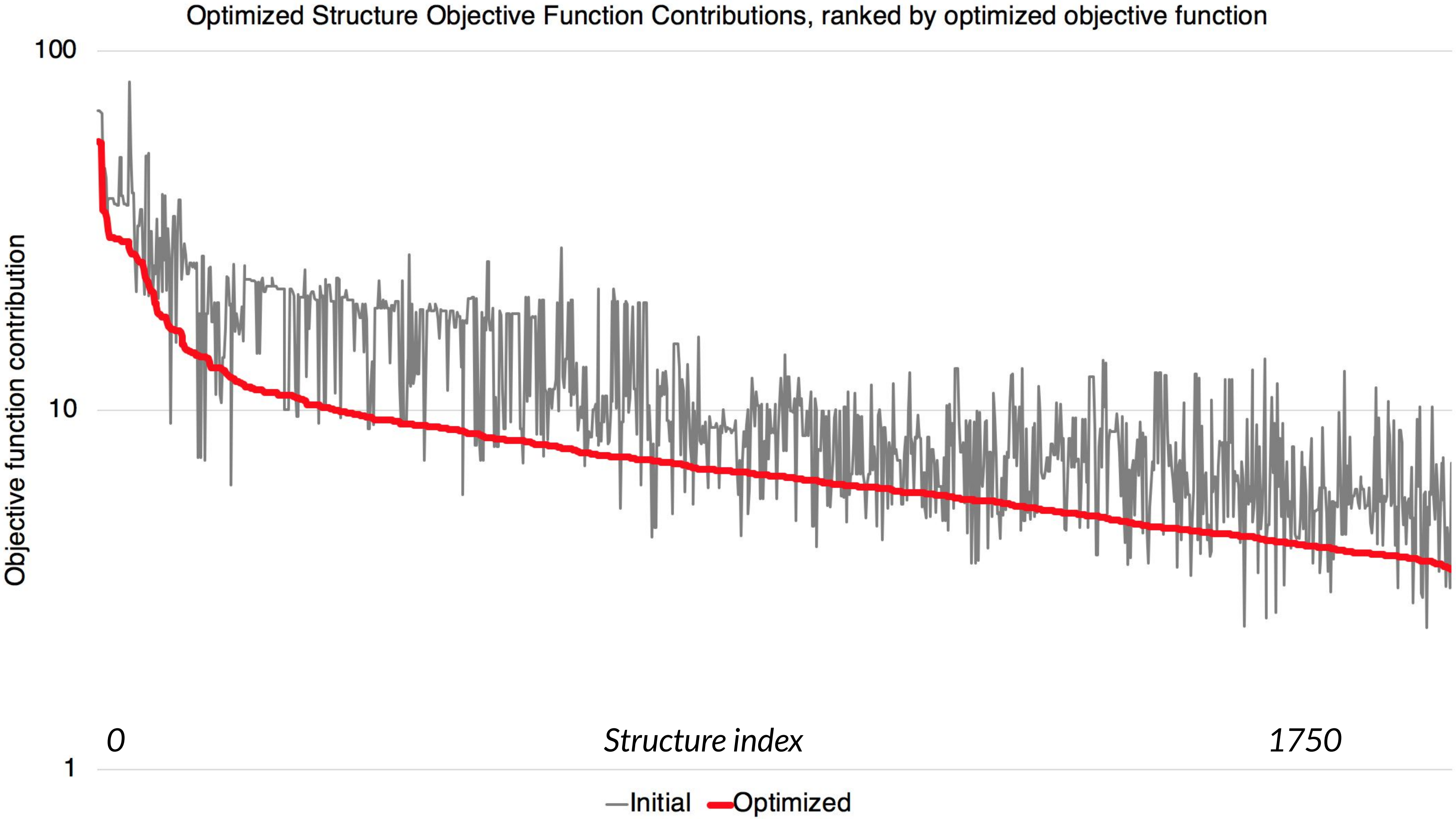
Not looking as good:



Dataset Name	OpenFF Group1 Torsions
Entry Label	C[c:1]1cccc([c:2]1[NH:3][CH3:4])F
Canonical SMILES	Cc1cccc(c1NC)F
Torsion Atom Indices	[3, 4, 8, 7]
Torsion SMIRKs	[*:1]~[#7X3,#7X2-1:2]-!@[#6X3:3]~[*:4]
Torsion SMIRKs ID	t69
SMIRKs Total Count	231

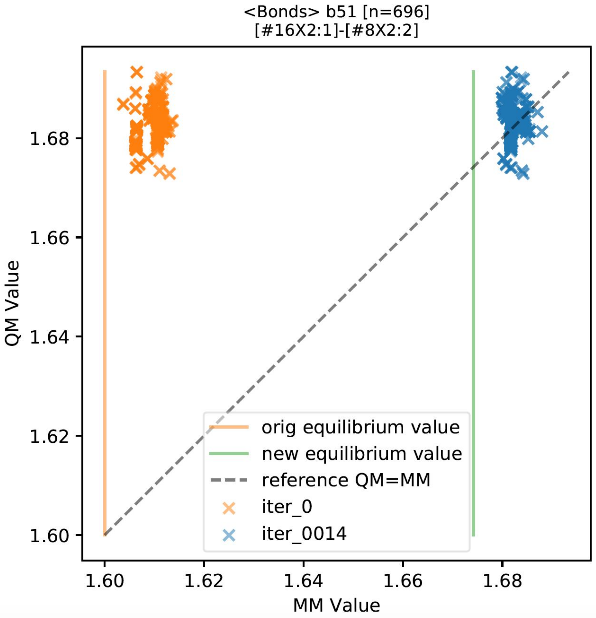
- Most torsion profiles improve agreement with QM significantly
- Some others demonstrate equivocal or slightly worse quality of fit
- Closer examination of individual torsion profiles will inform new parameter types in future releases

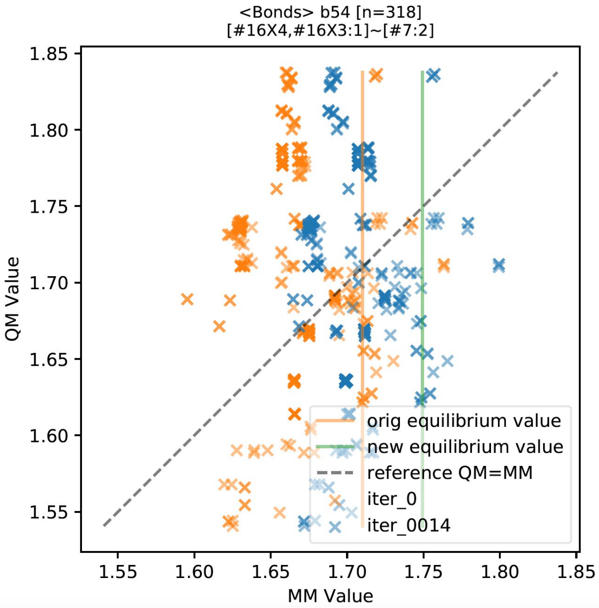
Results: Fitting of optimized geometries

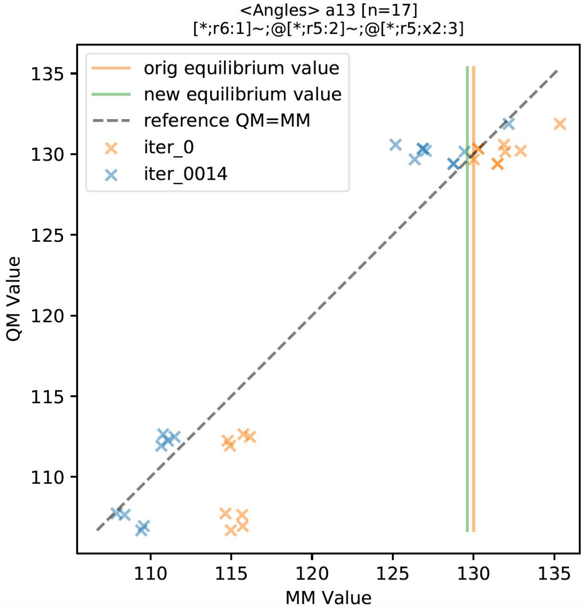


Optimized Structure Objective Function Contributions, ranked by optimized objective function

Results: Fitting of optimized geometries

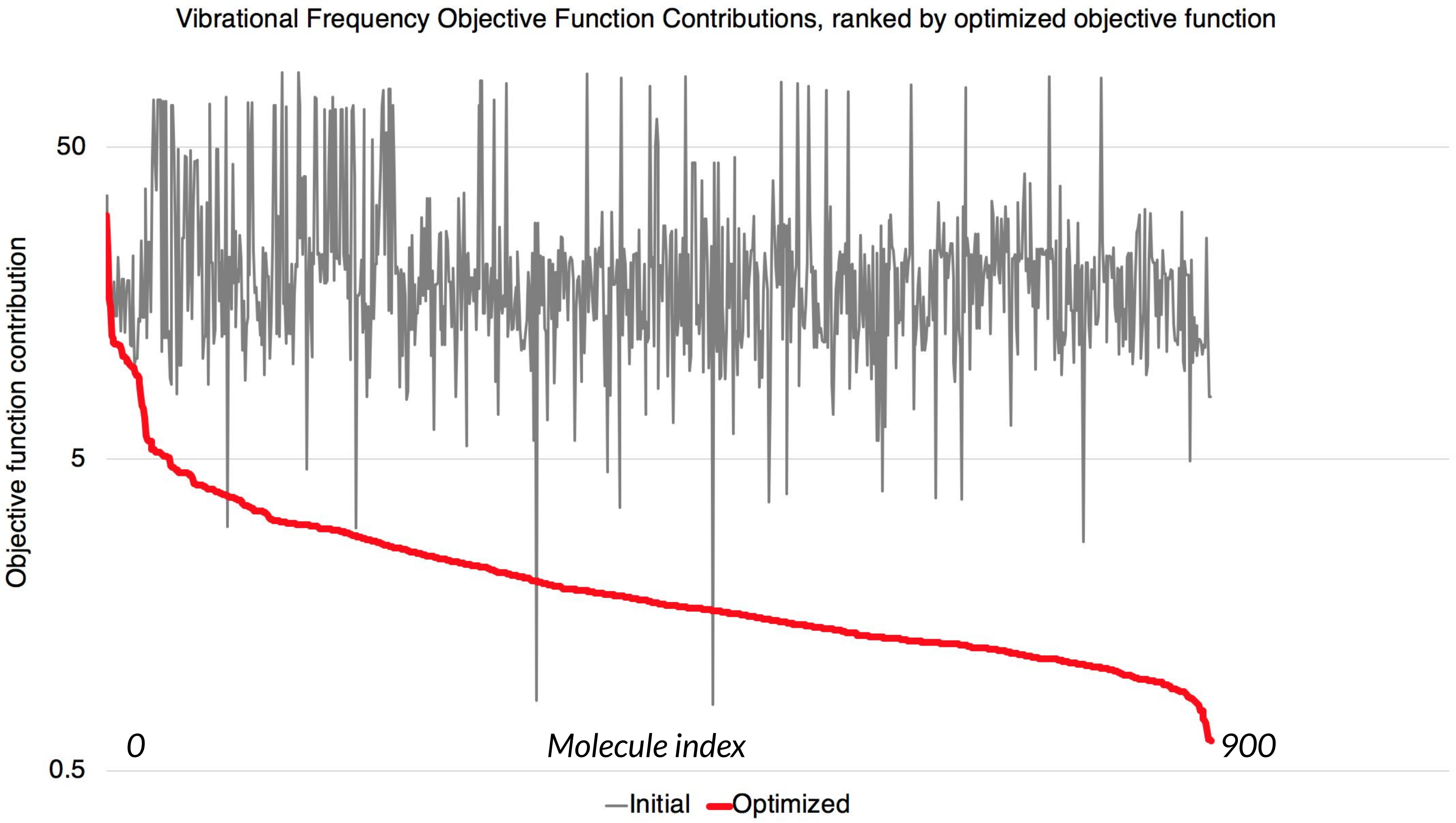




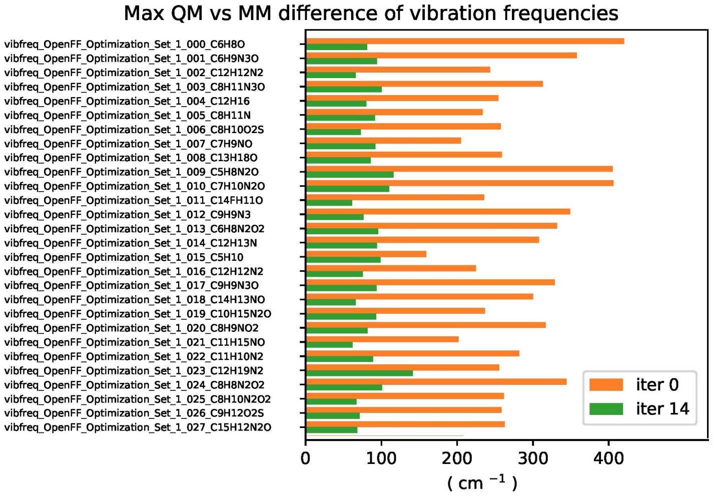


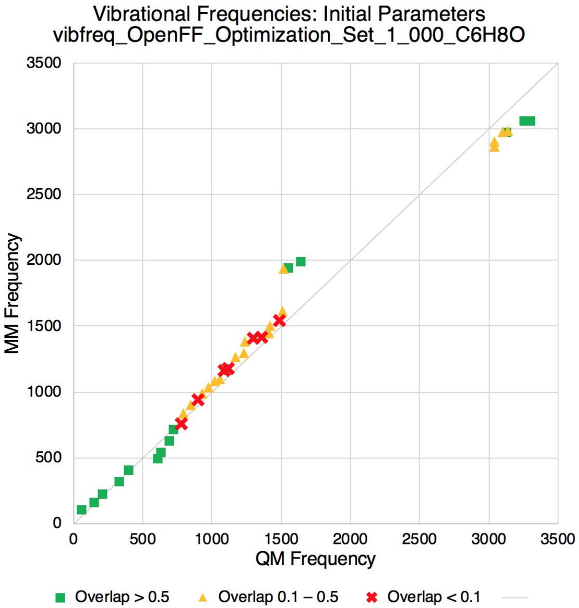
- Within one bond or angle parameter type, MM optimized values more tightly distributed than QM
- Optimization can shift the mean, but cannot easily change distribution shape (middle)
- Energy-minimized angles in constrained systems can be far from equilibrium parameter
- Bimodal or very broad distributions suggest need for some new parameter types, or bond order-based parameter assignment in future releases

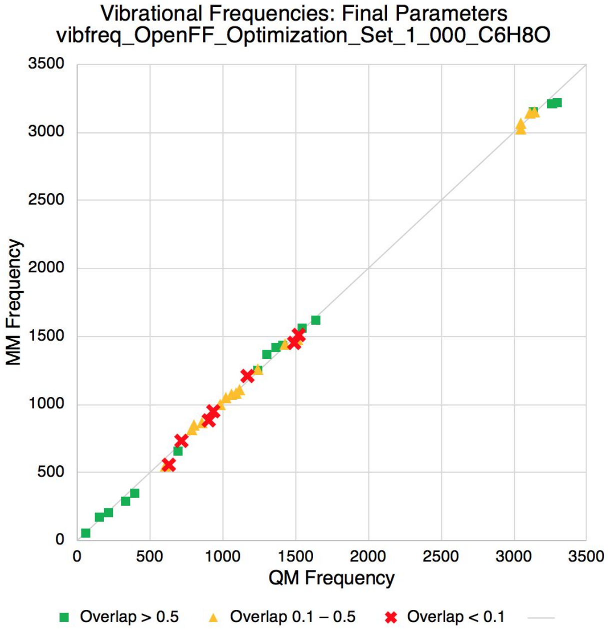
Results: Fitting of vibrational frequencies



Results: Fitting of vibrational frequencies

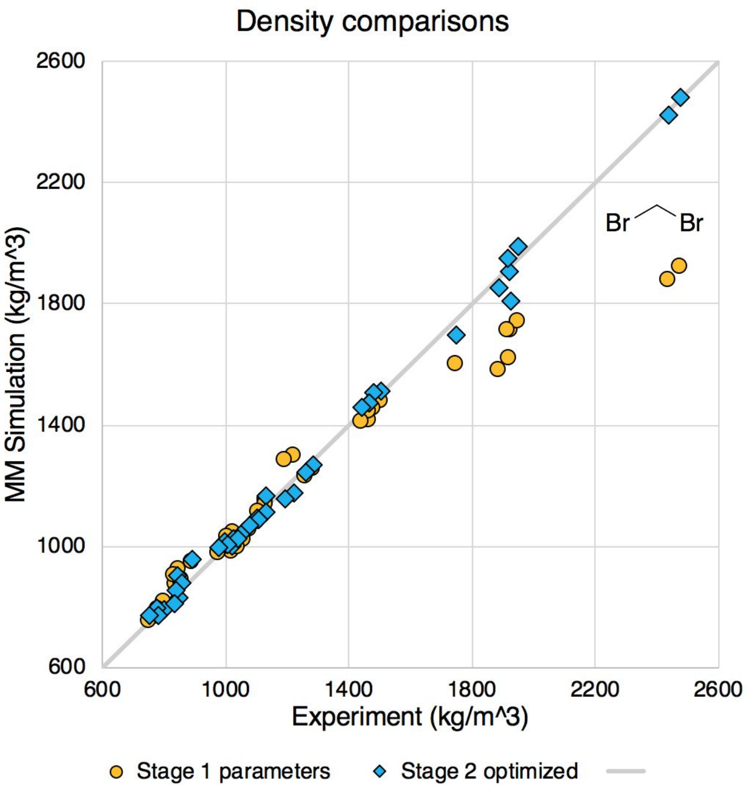


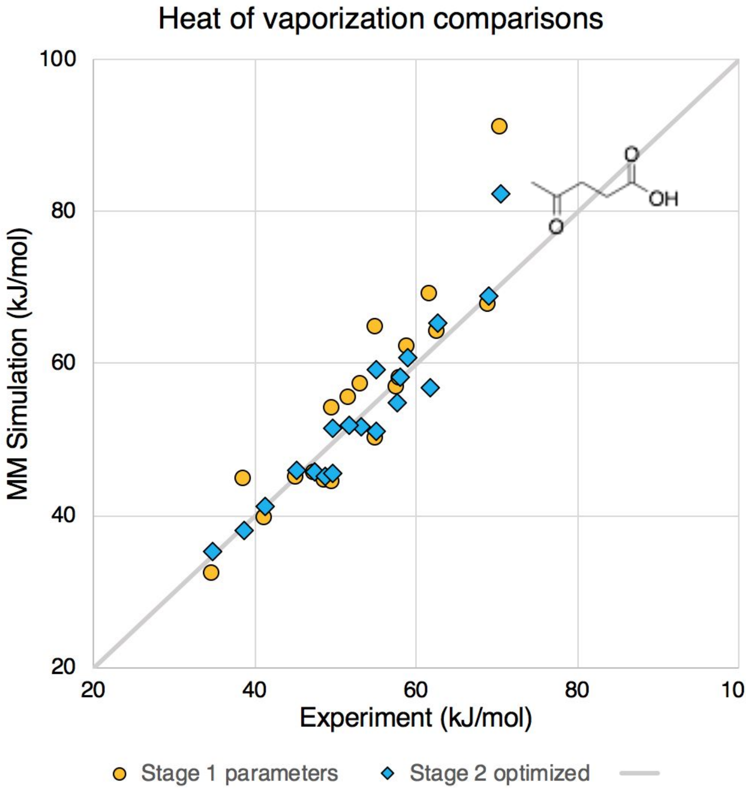




- Most of the contributions come from a few modes with large (>500 cm^{-1}) frequency differences
- Parameter fitting significantly improves correlation without hurting mode alignment
- Future fitting using internal coordinate Hessian would be a more direct approach

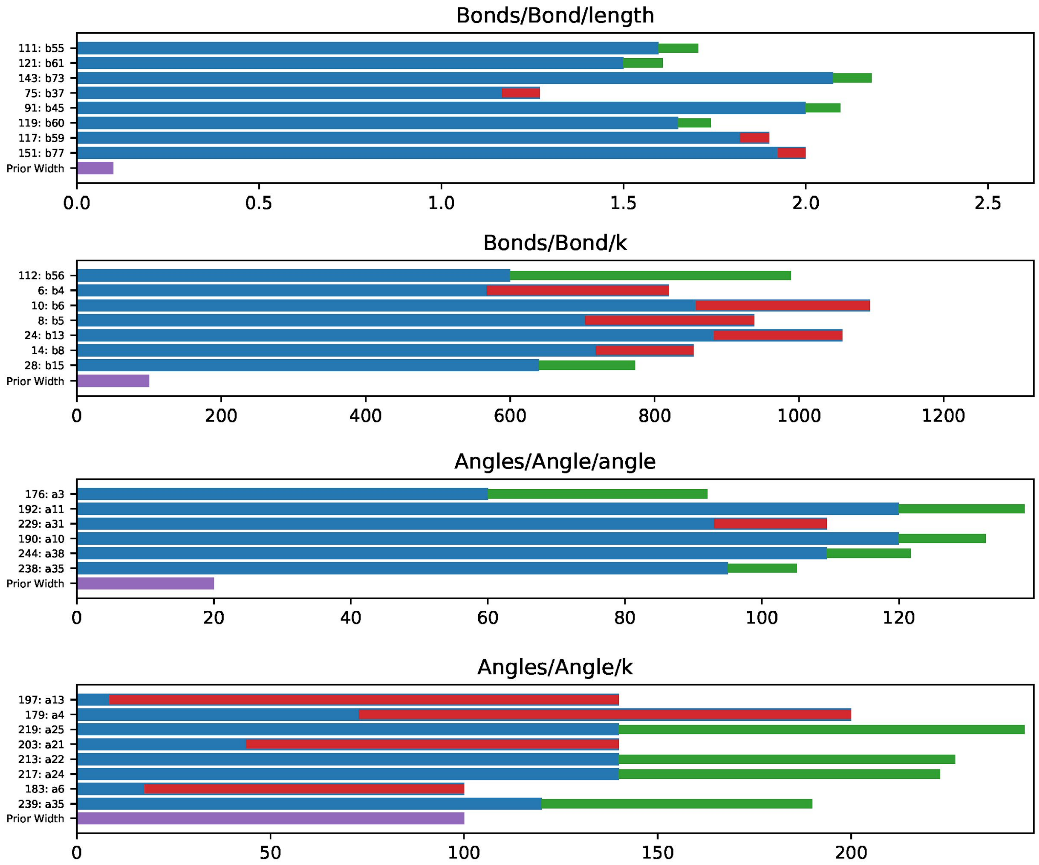
Results: Fitting of thermodynamic properties





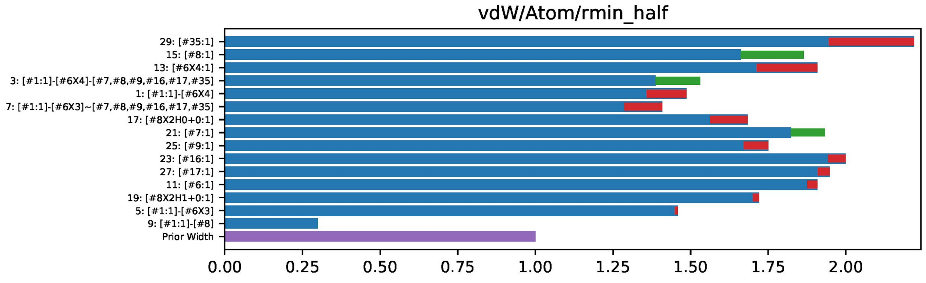
- Notable improvements for both observables, corrected underestimated density for halogens
- Outlier for ΔHvap is a carboxylic acid with possible sampling issues

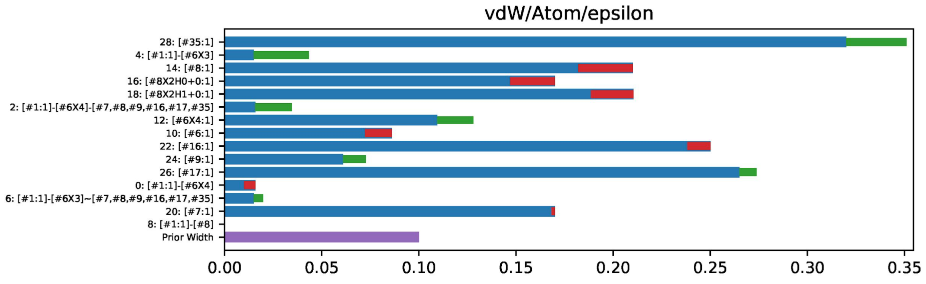
Results: Changes in parameters

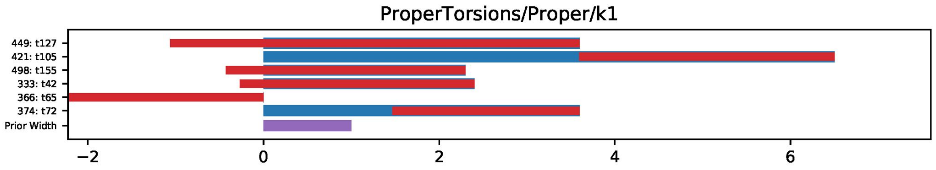


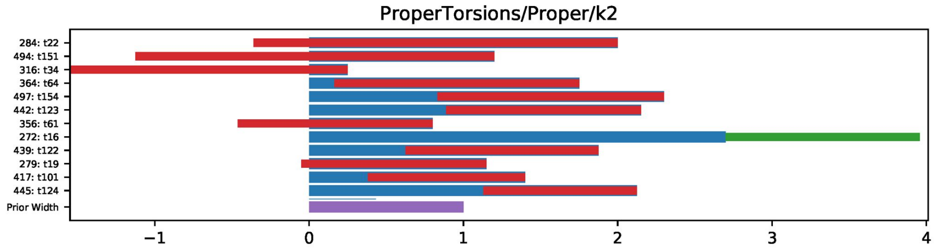
- Plots highlight only the parameters with the largest changes out of several hundred that were optimized.
- Bond lengths have small changes as expected.
- Force constants change by up to 30-40%.
- Equilibrium angle parameters can be far away from energy-minimized geometries (a3 = cyclopropane).
- Some angle force constants are very small (optimized to ~15 kcal/mol/Å^2). Possible that we're seeing slight overfitting due to large prior width for this parameter type.

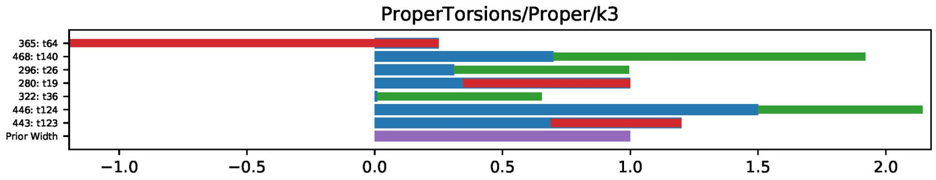
Results: Changes in parameters











- Torsion parameters will often change by relatively large amounts or change sign. Largest changes are ~3 kcal/mol in 1-fold and 2-fold terms.

- vdW parameters change by less than 5% to match thermodynamic properties.

Outlook

Force field is ready for thorough benchmarking and testing

- We look forward to hearing about your results
- We will be running extensive benchmark calculations within our collaboration

Some near-term development goals (next minor release?)

- Replacement of vibrational frequencies by internal coordinate Hessians
- Including torsion drives of flexible rings
- Co-optimization of LJ and bonded parameters using all targets

Longer term goals (not an inclusive list)

- Identify valence degrees of freedom that need to be explicitly scanned
- Incorporate improved charge models (e.g. Schauperl & Gilson's RESP2)
  or optimize charge model parameters to reproduce QM and experimental observables
- Including thermodynamic properties of mixtures in training data set

Thank you!

For help with this release & ongoing collaborations:

Christopher Bayly (all-around wizard)
Jeffrey Wagner, John Chodera (toolkit development & support)
David Mobley, Byron Tjanaka (parameter coverage)
Michael Shirts, Mike Gilson, Owen Madin (experimental data selection)
Xavier Lucas & Roche (providing a great molecule set)
Chaya Stern (conformer generation)
Hyesu Jang (RESP methods & intramolecular H-bonds)
Victoria Lim (QM method benchmarking)
Levi Naden, Andrea Rizzi (ForceBalance Python 3 compatibility)
Trevor Gokey (vibrational analysis)
Jessica Maat (improper torsions)
Michael Schauperl (electrostatic models)
David Slochower (host-guest binding)
Karmen Condic-Jurkic (planning next major release)