Torsion drive discussion slides

Lee-Ping Wang

Open Force Field Initiative
Advisory board meeting, 15 January 2020

Acknowledgments: Xavier Lucas

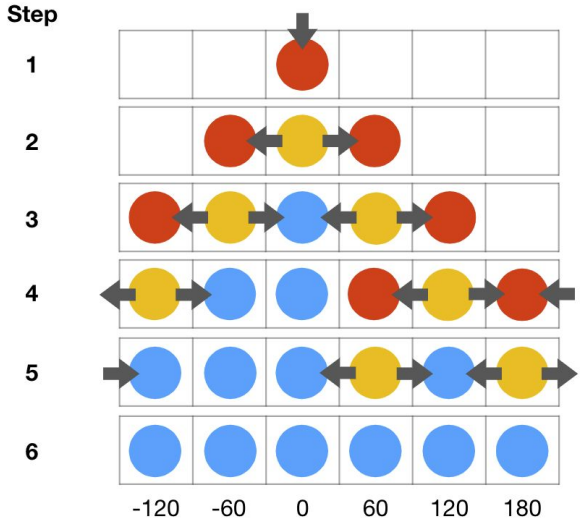
#torsions on Slack

Background

- A major portion of OpenFF fitting data comes from QM *torsion drives*
  - Not doing any fragmentation at the moment; molecules are small (<30 heavy atoms).
- Paper describing torsion drive method is nearly finished
  - Authors: Yudong Qiu (UC Davis), Daniel Smith (MolSSI), Chaya Stern (MSKCC), Mudong Feng (UC San Diego), Lee-Ping Wang (UC Davis)
  - Project was started in early 2018 but major applications and paper writing were completed using OpenFF funding
  - Will be circulated for industry partner feedback soon

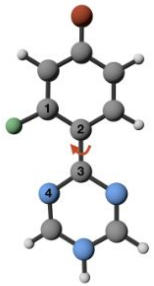
Torsion drive procedure

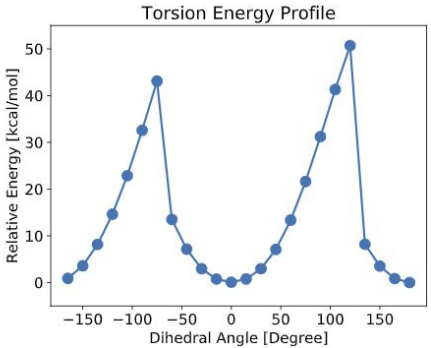
- This is a recursive procedure designed to avoid hysteresis, discontinuities and getting stuck in high-energy local minima.
    - 1D torsion drive cost is ~2x of “traditional” sequential relaxed scan
    - Given parallel resources and multiple initial structures, wall time can be much lower than sequential scan

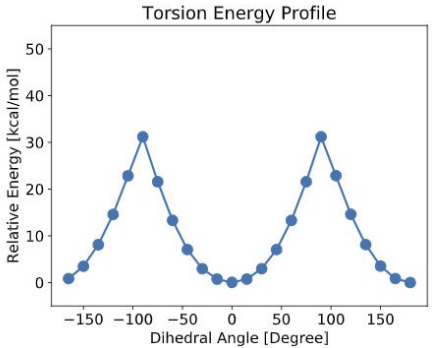


Red = active point; Orange = active point from previous step; Blue = inactive point
A “step” involves running constrained optimizations started from the previous iteration’s active points.

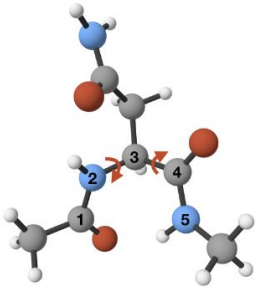
Torsion drive: the basic results

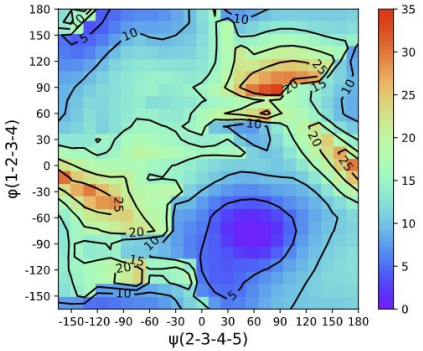


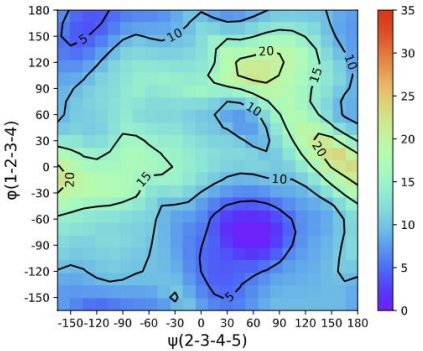




Sequential 1-D scan (middle) vs. TorsionDrive (right) results for molecule pictured at left. Fluorine (F) shown in green.

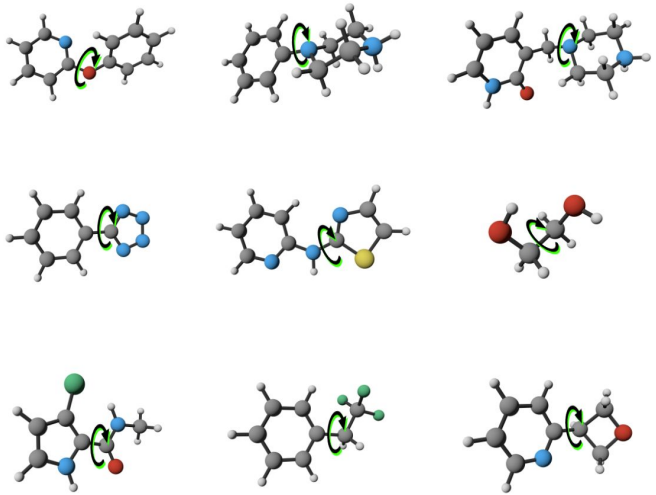


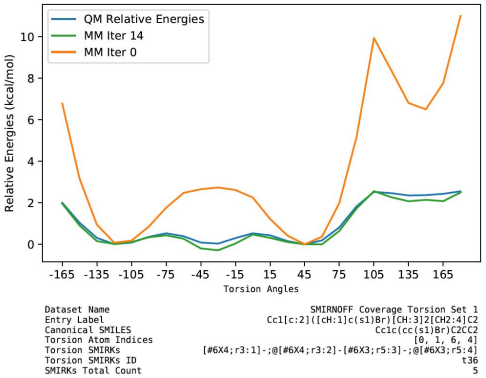




Sequential 2-D scan (middle) vs. TorsionDrive (right) results for glutamine dipeptide pictured at left.

QCArchive torsion drives and ForceBalance target preparation

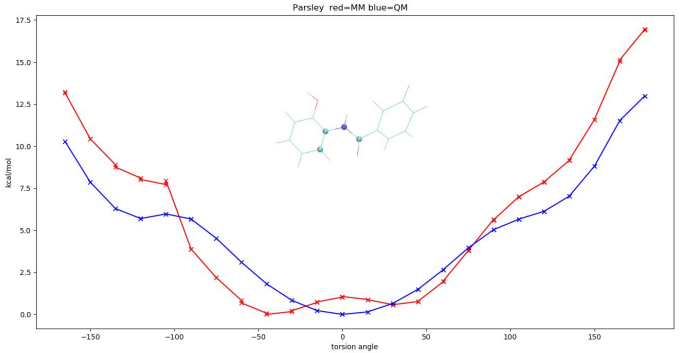


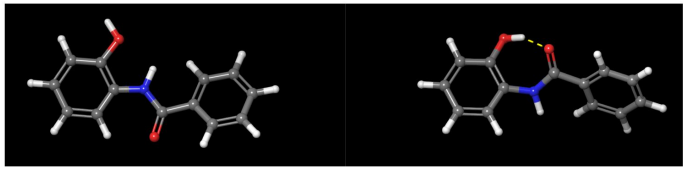


- B3LYP-D3(BJ)/DZVP(DFT) level of theory was selected
- Constrained geometry optimizations used open source geomeTRIC (for optimization) and Psi4 (for energies/gradients)
- Calculations were carried out using MolSSI's "QCArchive" distributed infrastructure
- For fitting Parsley parameters, a total of 1085 torsion drives were included in the training data set (669 from "Roche set" and 417 from "coverage set")
- QCArchive torsion drives were filtered by confirming no changes in topology, and *removing any that contained intramolecular hydrogen bonds*
    - The rationale is to avoid building strong intramolecular electrostatics into the energy profiles, which we thought would adversely affect the quality of fitted potentials.

Top = example molecules; bottom = example of improved MM torsion profile

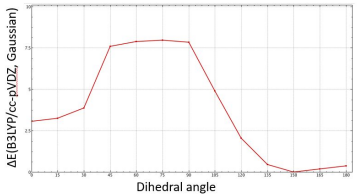
A question concerning intramolecular H-bonding interactions





Dihedral angle: 0 deg
ΔE(m06-2X-D3/aug-cc-pVDZ, Jaguar) = +2.3 kcal/mol
ΔE(B3LYP/cc-pVDZ, Gaussian) = +3.1 kcal/mol

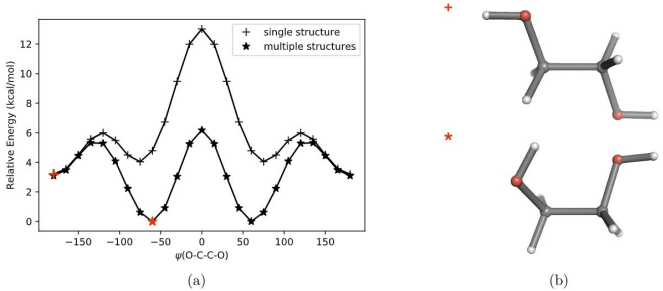
Dihedral angle: -141 deg

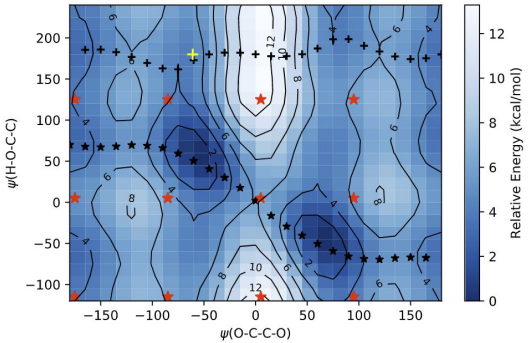


- Xavier Lucas pointed out that some QCArchive torsion drives did not contain intramolecular H-bonds that could lower the energy
- For this case, gas phase calculations did not significantly overbind the intramolecular H-bonded conformation
  (ΔE=1.7 kcal/mol calculated from differences in optimized geometries using M06-2X/aug-cc-pVDZ with and without Poisson Boltzmann (PBF) implicit water)
- We plan to create two torsion drive datasets, one containing intramolecular H-bonds and one without, and assess the impact of including the former set in fitting

Top = QM vs. MM torsion profiles from Parsley training data set; Bottom = Calculation from Xavier Lucas including lower-energy minimum

TorsionDrive can start from multiple initial structures





- There is no guarantee TorsionDrive can find the global minimum, but it can naturally “stitch together” results from multiple starting structures
- Multiple starting structures from “known” local minima (e.g. from conformational search) may be provided
- This can also be used to decrease calculation walltime (given sufficient parallel resources)
- One current question is whether and how to use the constrained optimizations that are higher in energy than the lowest energy at each grid point; these are not part of the surface but we have this data.
- I have an early idea about how to optimize FFs in the presence of implicit solvent models (core idea is to use PCM/COSMO implicit solvent for both QM and MM)

Top = 1-D TorsionDrive results for single initial structure (“+”) or multiple initial structures (“*”)
Bottom = 2-D TorsionDrive results showing initial conformations of 1D torsion drives