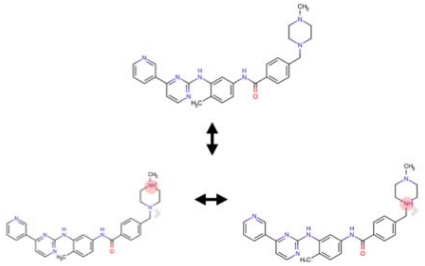
# Torsion fitting and small molecule fragmentation

**Open Force Field Consortium**
January 2019 Workshop, San Diego

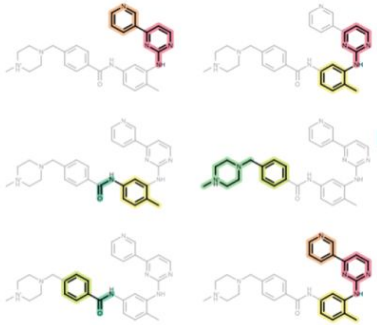
**John Chodera** (MSKCC), **Lee-Ping Wang** (UC Davis), **Daniel Smith** (MolSSI), **Chaya Stern** (MSKCC), **Yudong Qiu** (UC Davis)

**#torsions** on Slack

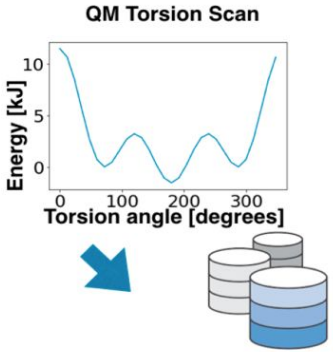
# Overview of the torsion fitting pipeline



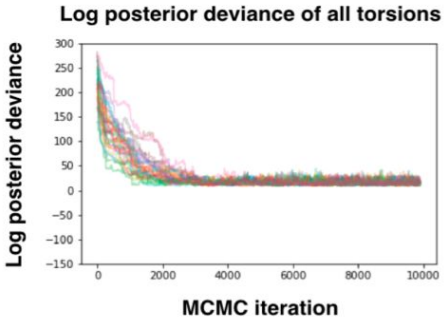
Enumerate reasonable ionization, protonation and tautomers



Fragment molecule without destroying important chemistry




Run multi-dimensional QM Torsion scan and deposit Into database




Fit QM torsion profiles using Bayesian inference and MCMC to avoid minima

**Related code:**

**Geometry optimization: https://github.com/leeping/geomeTRIC**

**Multi dimensional torsion drives: https://github.com/lpwgroup/torsiondrive**

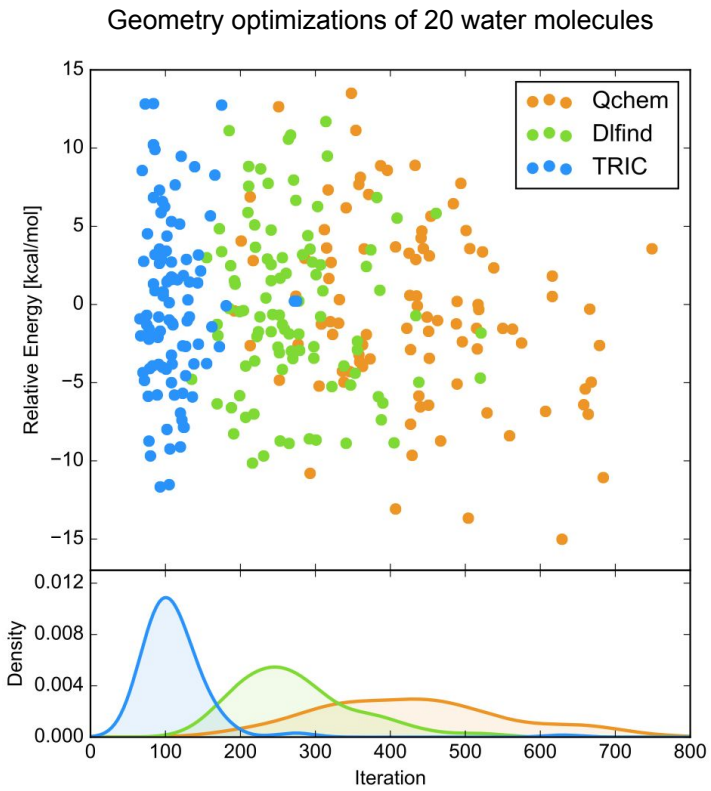
**Automated QC parallelization: https://github.com/MolSSI/QCFractal**

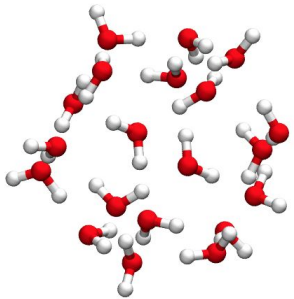
**Bayesian torsion fitting: https://github.com/choderalab/torsionfit**

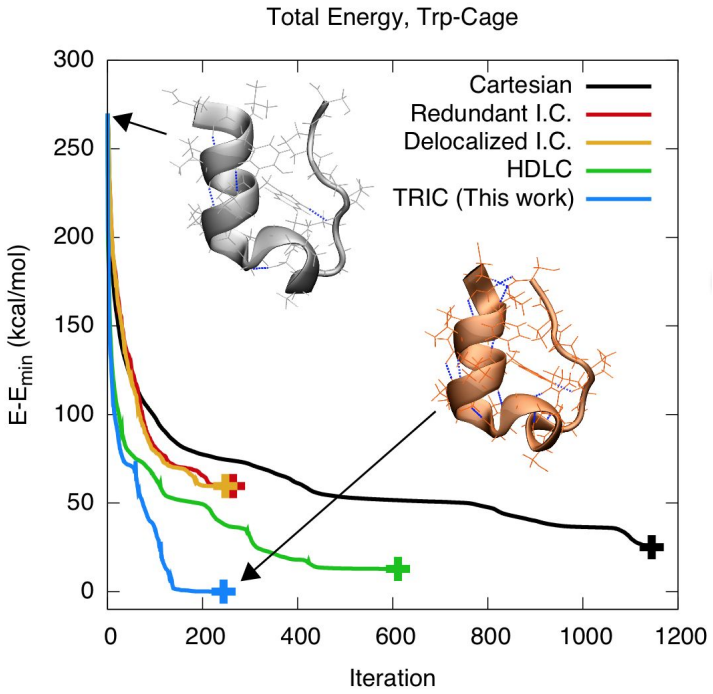
# Progress in 2018: Driving of selected torsions

- “Given a molecule, initial conformation, QM calculation specification and *N* selected torsions to drive, produce a *N*-dimensional potential energy profile where torsions are constrained along a specified grid and orthogonal degrees of freedom are minimized to the fullest extent possible.”
- Achieved through four open-source tools: torsiondrive, geomeTRIC, Psi4 and QCArchive

# geomeTRIC: Open-source geometry optimization



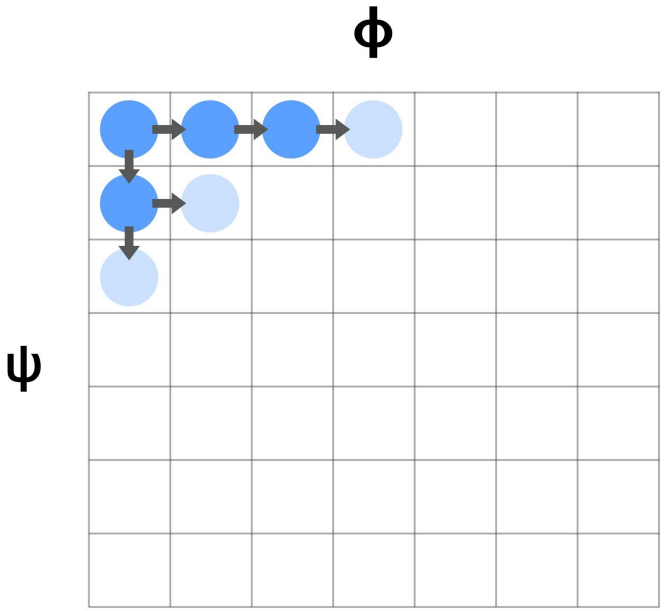





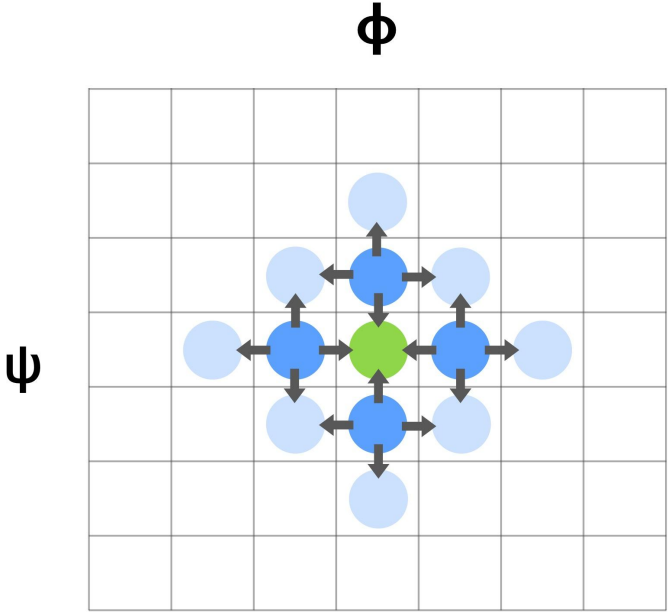

- Efficient geometry optimizer interfacing with QC software
- New *translation-rotation internal coordinates* (TRIC) improves performance for intermolecular degrees of freedom

https://www.github.com/leeping/geomeTRIC

# TorsionDrive: N-D wavefront propagation




Regular 2-D scan



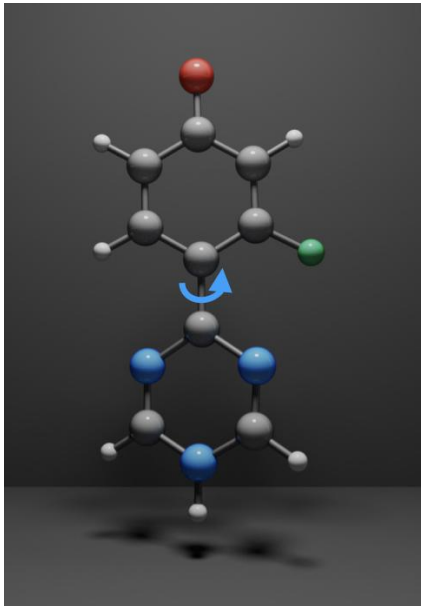

TorsionDrive 2-D scan

- ✓ Lower Energy
- ✓ More Robust
- ✓ Symmetric
- ✓ Fewer rounds of optimization (if distributed)

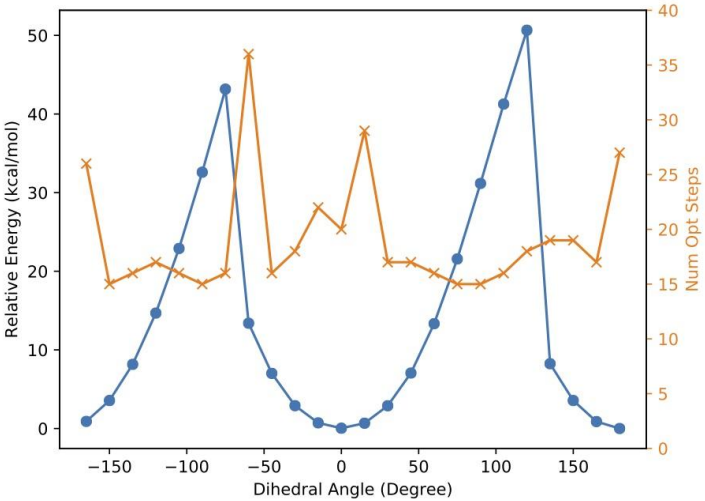
N(optimizations) = 2 * dimensionality * N(grid points)

https://www.github.com/lpwgroup/torsiondrive

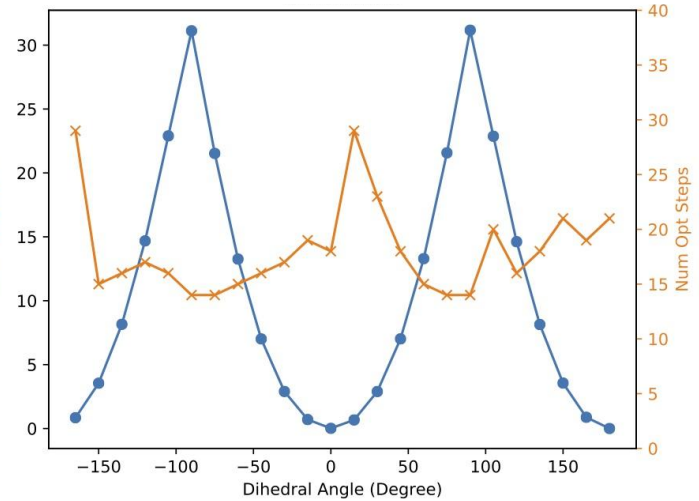
# TorsionDrive: 1-D Comparison



Structure



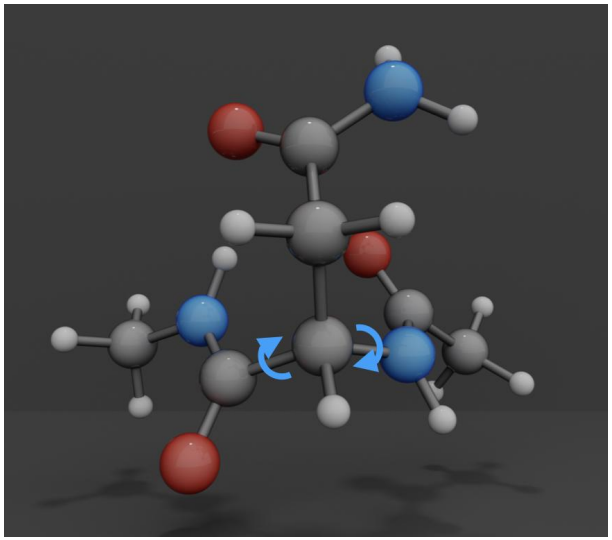

Geometric 1-D scan



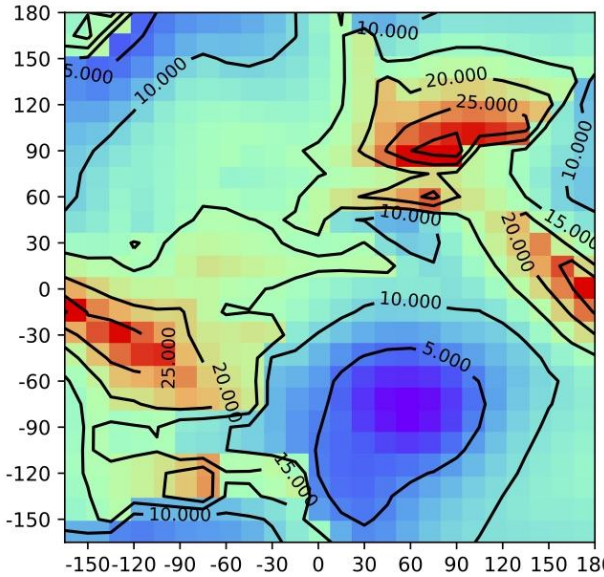

TorsionDrive 1-D scan

- Unidirectional torsion drive using geomeTRIC alone results in structures getting “stuck” in higher-energy local minima, and results dependent on scan direction.
- Wavefront propagation using torsiondrive + geomeTRIC recovers symmetric energy profile.

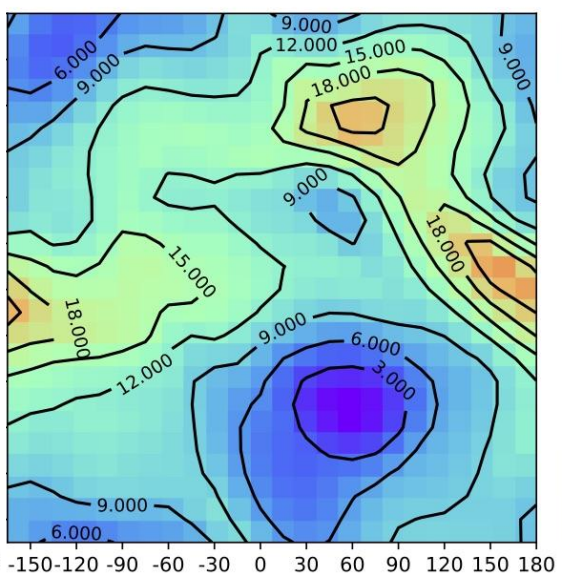
# TorsionDrive: 2-D Comparison



Structure



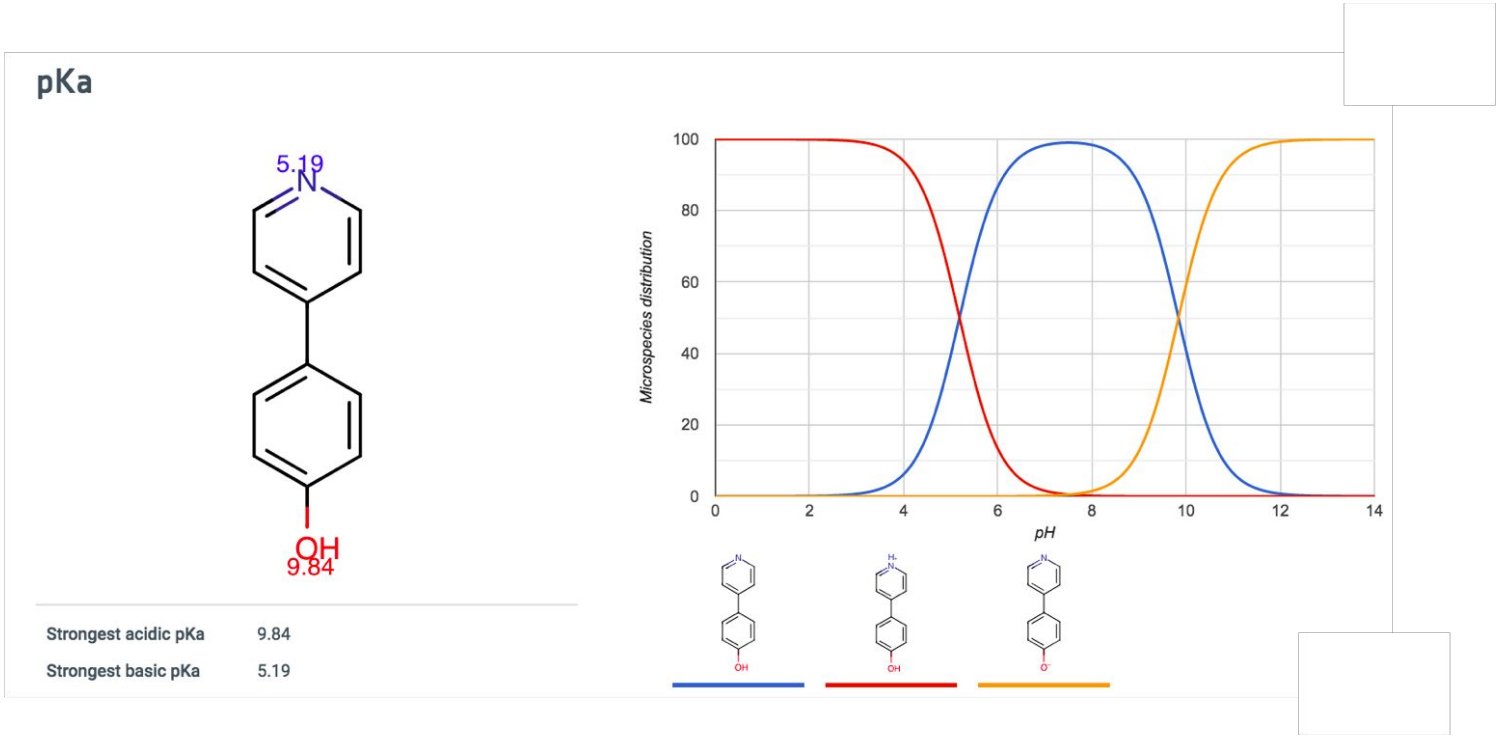

Geometric 2-D scan




TorsionDrive 2-D scan

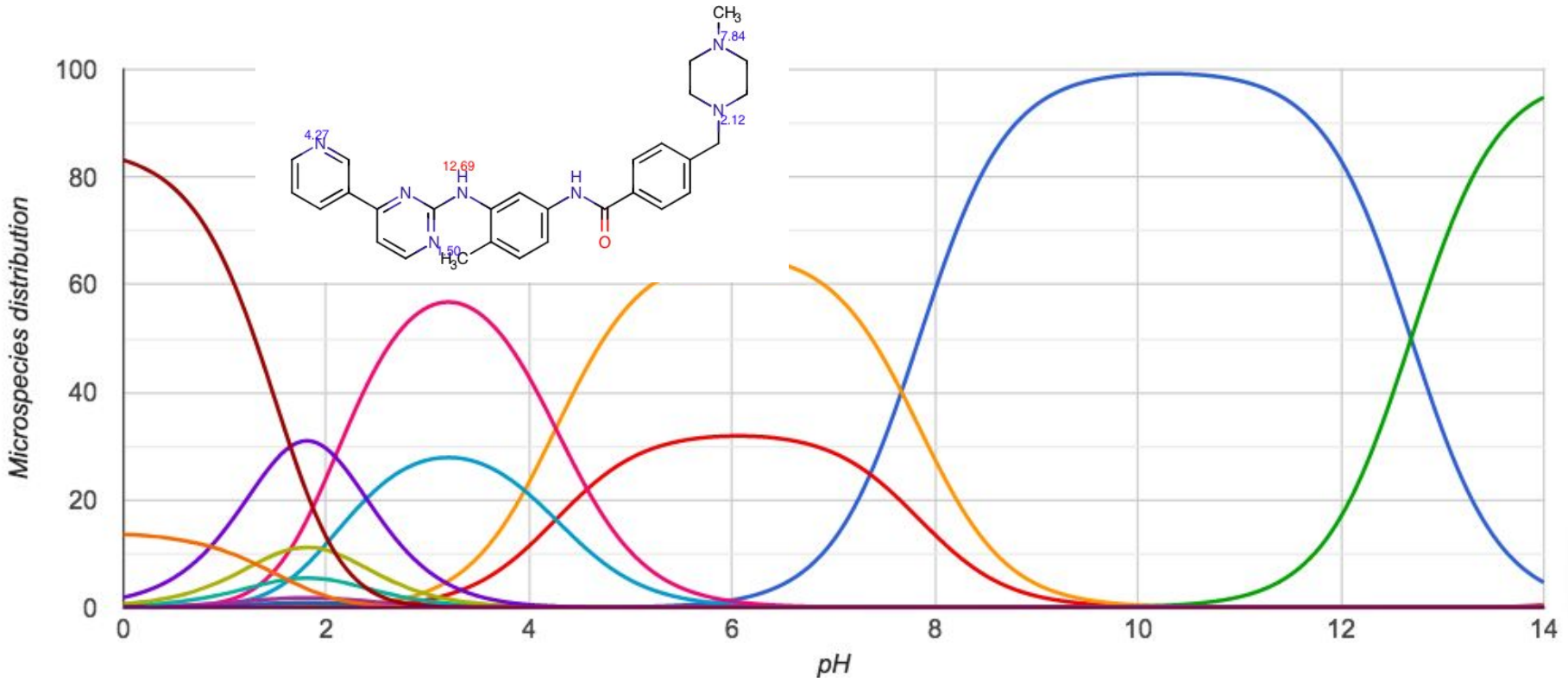
- Two-dimensional scan often provides richer information than a pair of one-dimensional scans.
- Local minima in “unidirectional” scan leads to energy discontinuities in potential surface, spurious high-energy regions, and dependence on scan direction / choice of leading and trailing dimension.
- Wavefront propagation using torsiondrive + geomeTRIC uncovers local minima and smooth features of minimum-energy surface (though discontinuities may exist in structural degrees of freedom).

# Enumerate protonation and tautomeric states

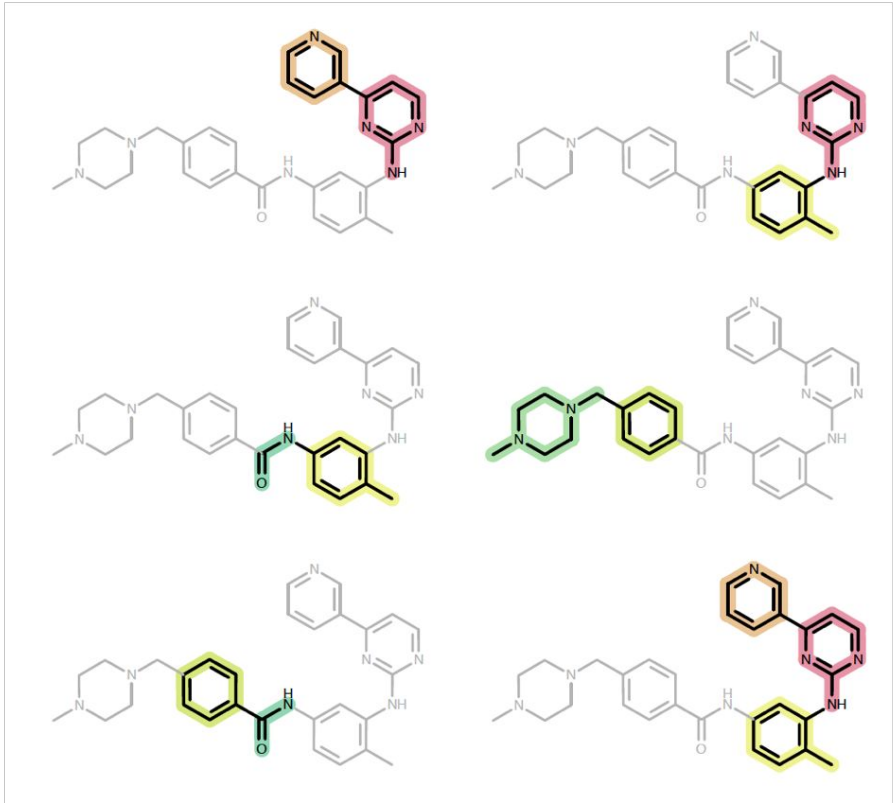


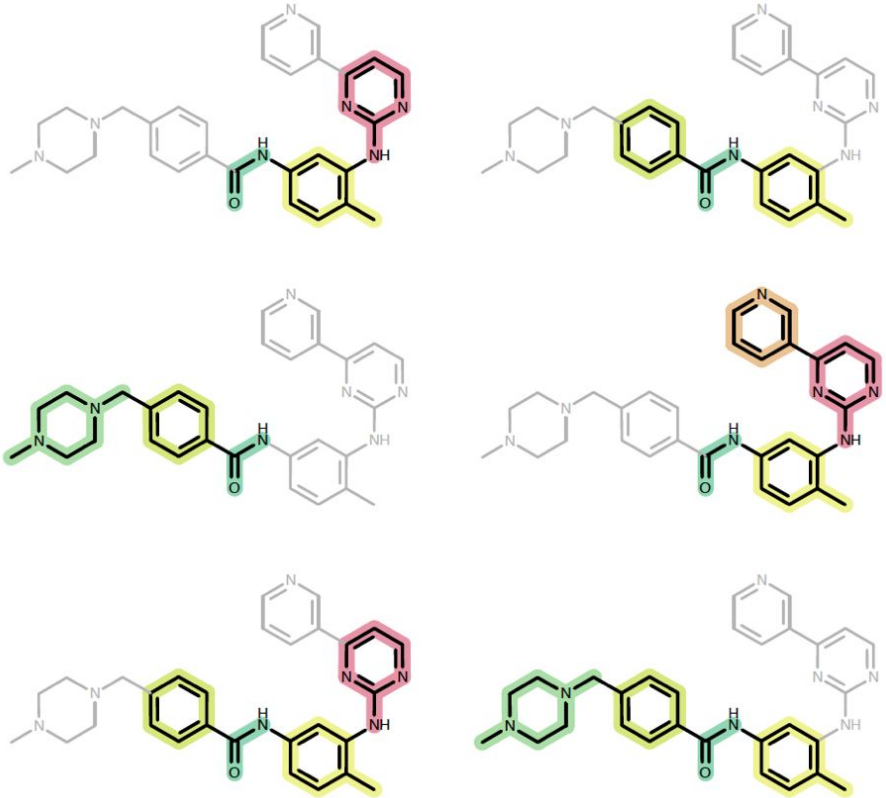

**Currently, enumeration generates unreasonable states. Unreasonable states get filtered out when AM1 fails.**

# Imatinib


Microspecies distribution
pH
100
80
60
40
20
0
0
2
4
6
8
10
12
14
CH3
7.84
2.12
4.27
12.69
1.50
H3C
H
N
O

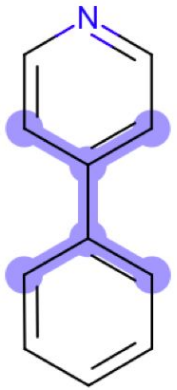
# Molecules are fragmented to reduce computational cost and conformational distribution

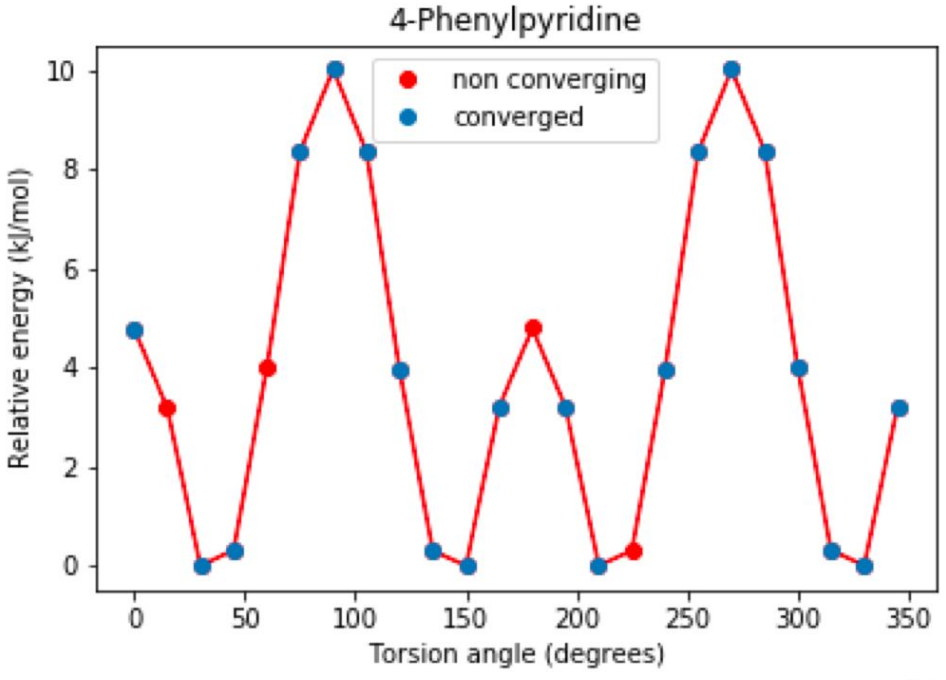




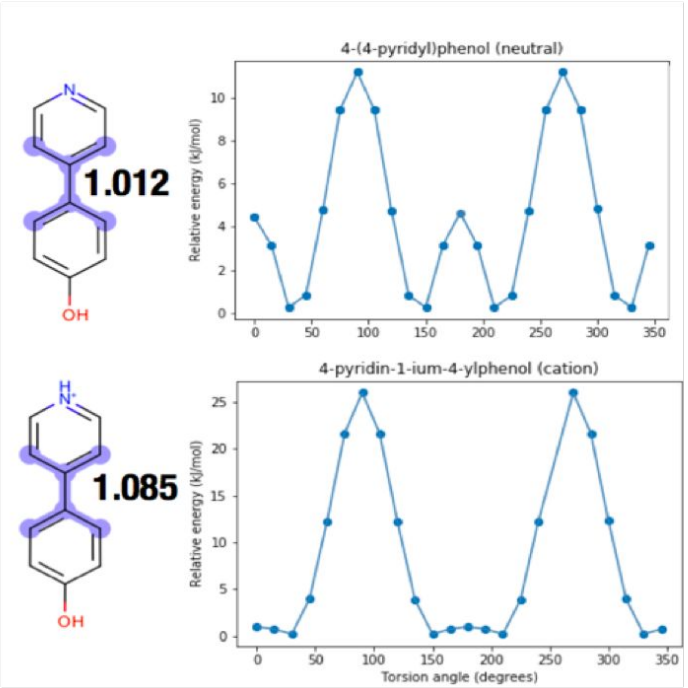
# What are the pitfalls when fragmenting molecules for QM torsion scans?

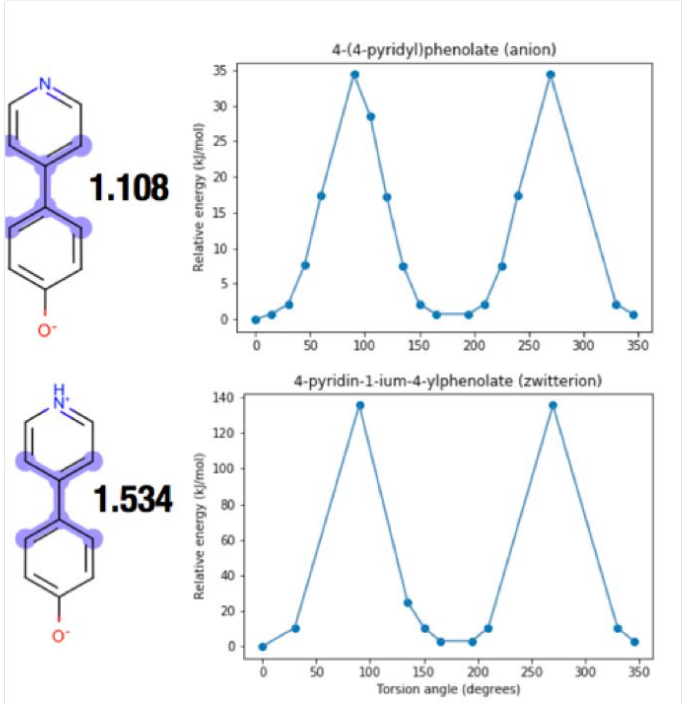
4-phenylpyridine



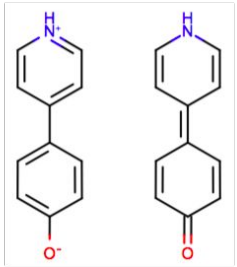



# Different protonation states and seemingly small changes far from the bond may result in different torsion profiles

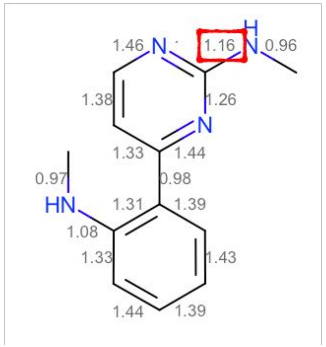


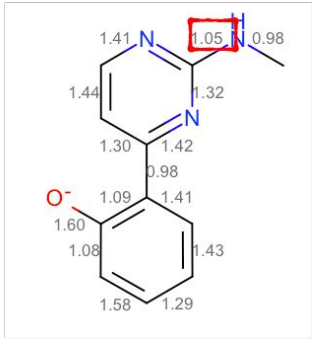



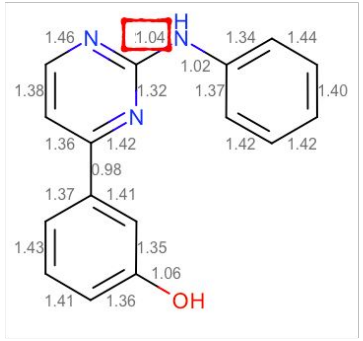

**The central bond in the zwitterion is part of the extended π system.**

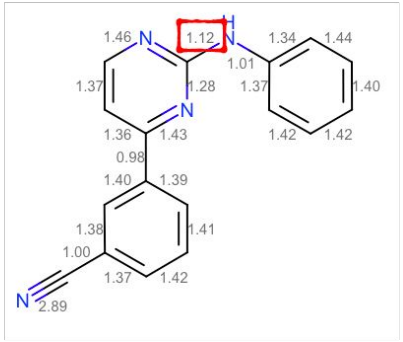


# Changes far away from bond may change conjugation.

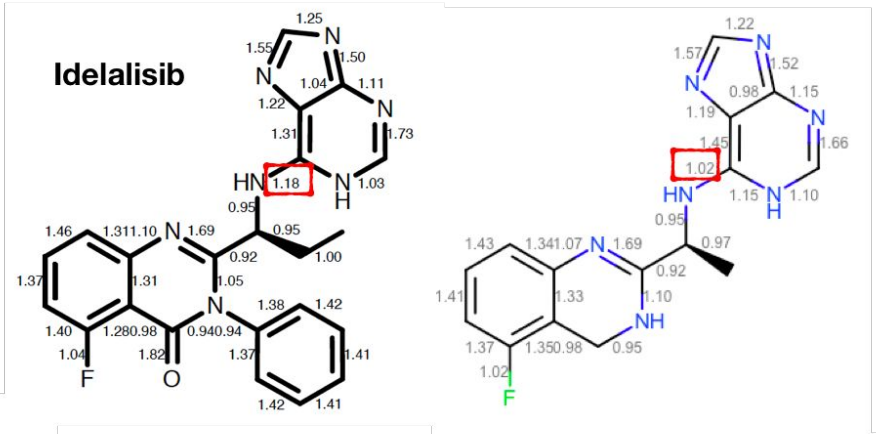


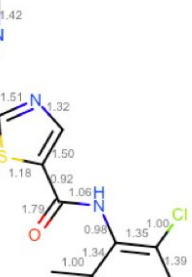



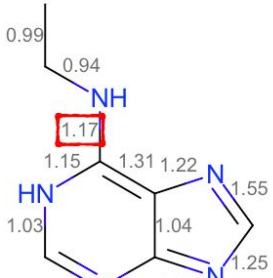
# m Kinase inhibitor set





The smaller fragment has WBO that is closer to full molecule



# We want to find the fragments that retain the correct chemical environment yet are reasonably small for QM torsion scan.

---

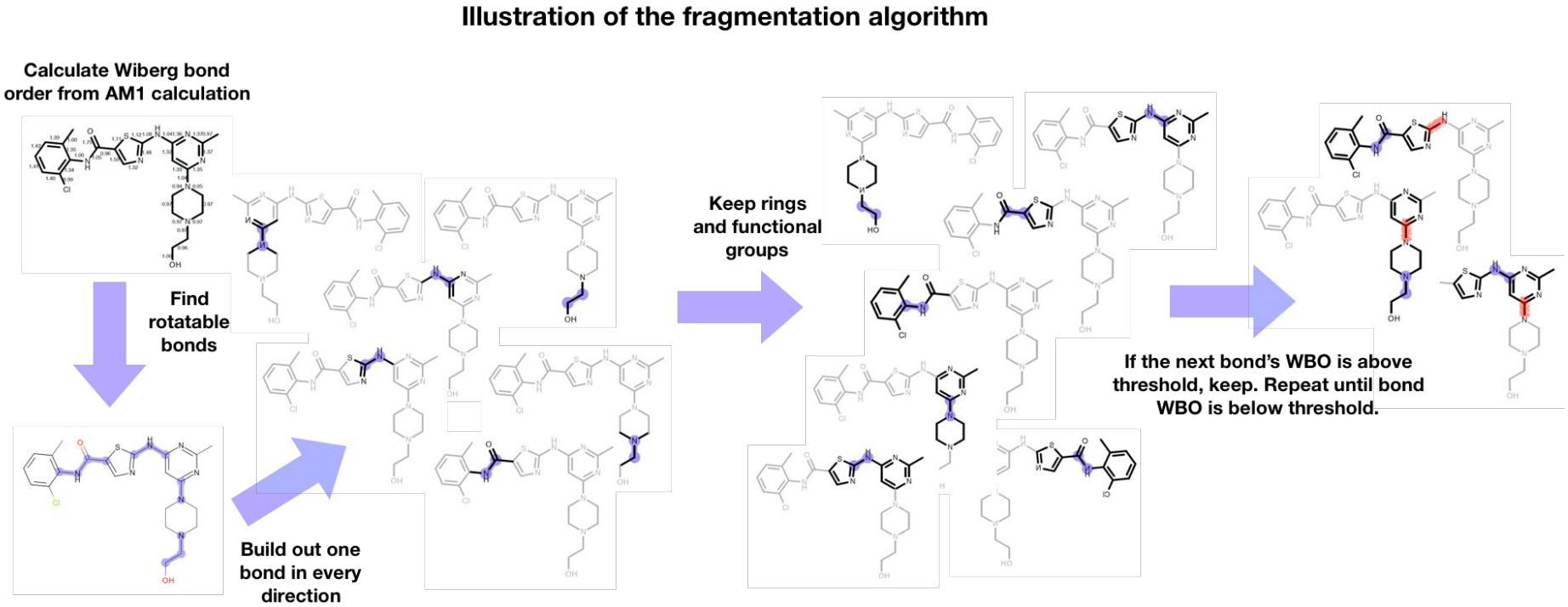
## In each fragment we want:

- A central rotatable bond
- All substituents of the central rotatable bonds
- Correct resonance structure
- 1 – 2 rotatable bonds for computational feasibility

## What we don't want to fragment:

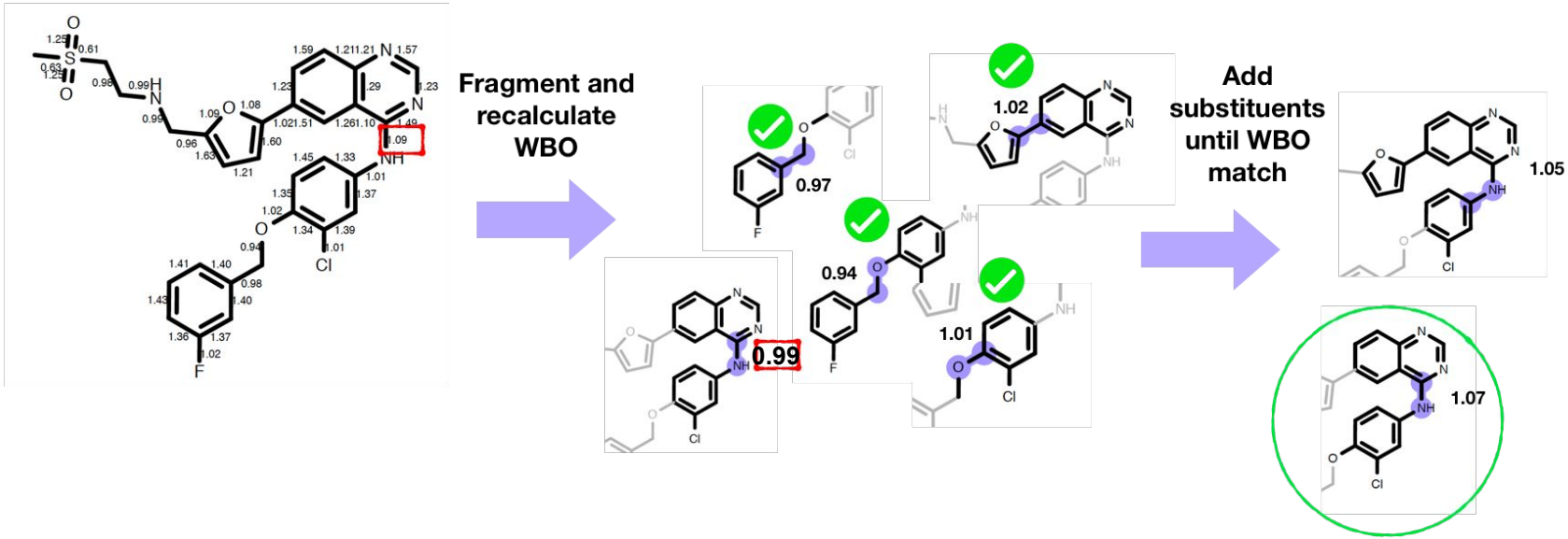
- Ring systems
- Extended pi electron systems (Wiberg bond order can help here)
- Some functional groups
- Non-rotatable ring substituents and/or ortho groups

# Intelligent fragmentation can reduce the misrepresentation of torsions in QM database

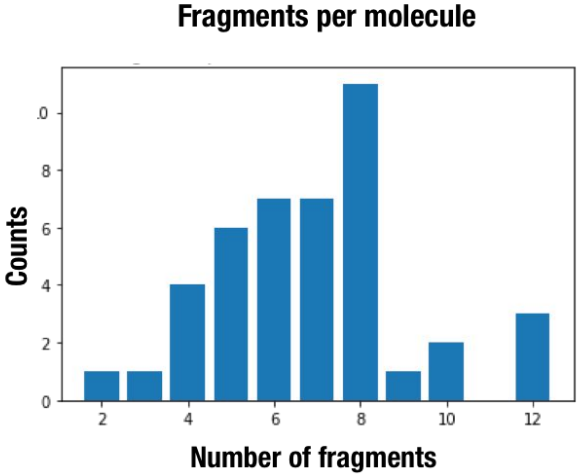


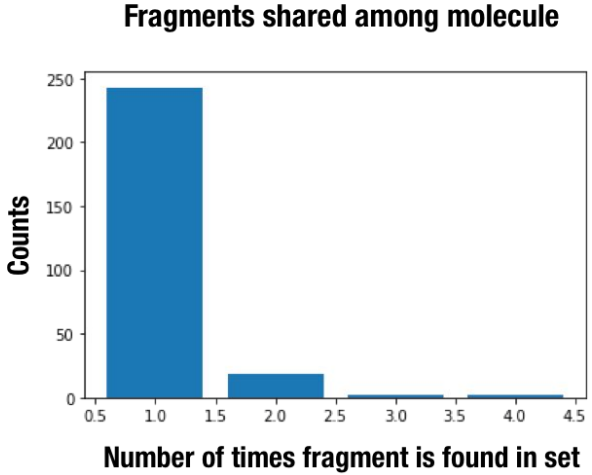
# While initial algorithm will not fragment conjugated bond, it does not ensure that conjugation chemistry is conserved on both sides.

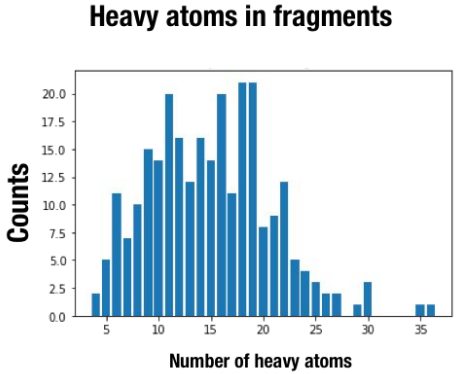
Recalculate WBO on fragment. If it changed, add atoms using MCMC until WBO match (within error) (WIP)

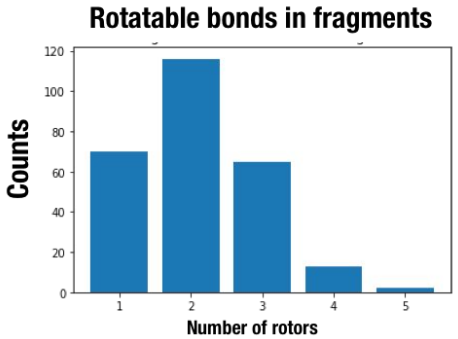


# Fragmenting 43 FDA approved kinase inhibitors generates 295 fragments

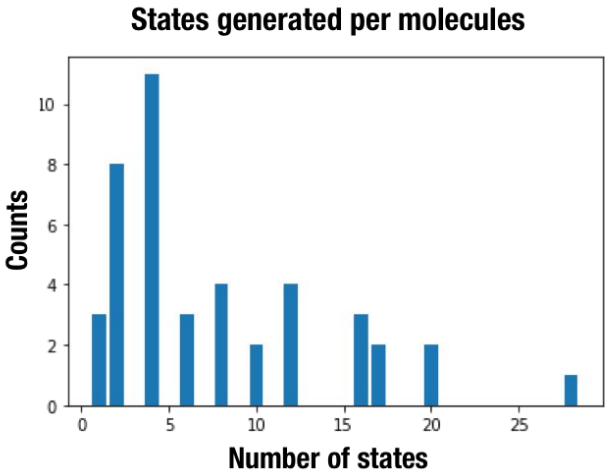


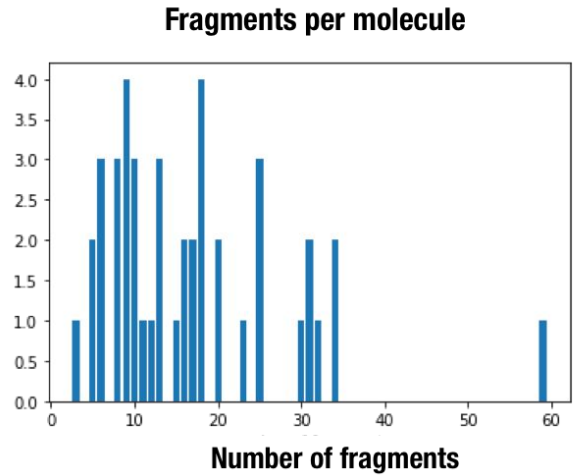



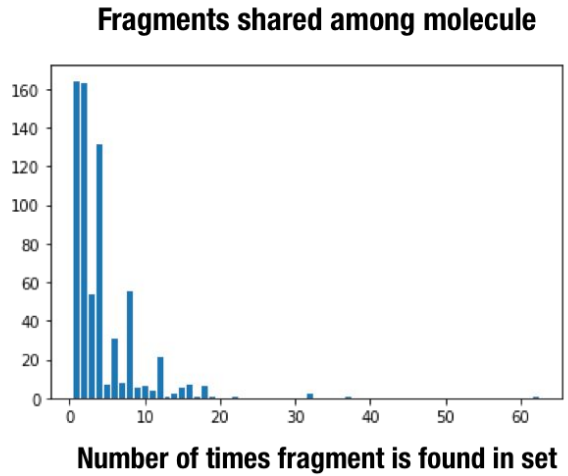



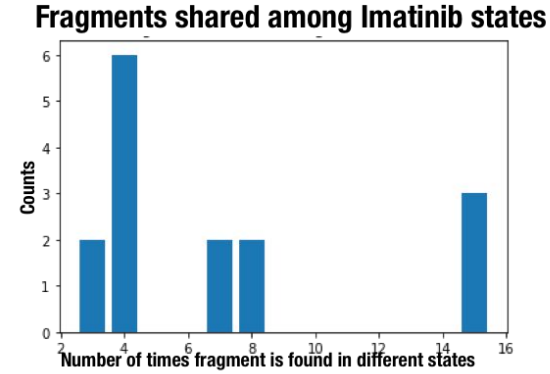
# Expand and Fragment 43 FDA approved kinase inhibitors generates 679 fragments

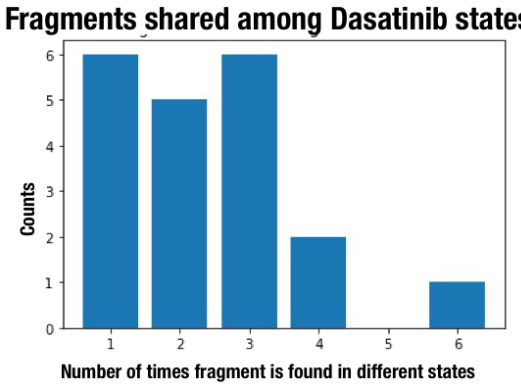


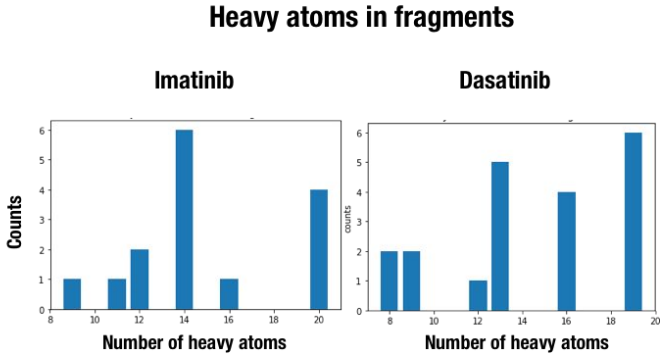



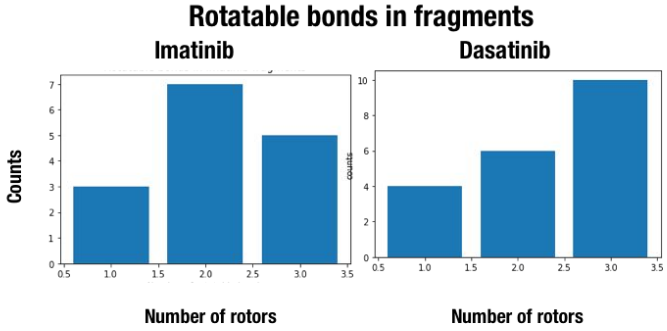
# How many torsion drives does a single drug produce?




| Drug | States | Unique fragments | 1D torsion drives | 2D torsion drives |
|---|---|---|---|---|
| Imatinib | 16 | 15 | 17 | 22 |
| Dasatinib | 6 | 20 | 27 | 36 |

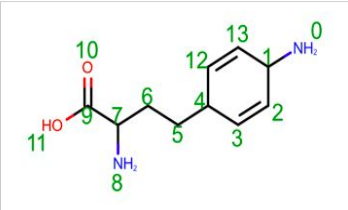


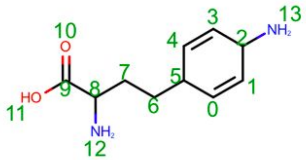



# `cmiles`: indexing molecules for QC database

**Issues `cmiles` addresses:**

**1) Arbitrary node indices in molecular connectivity graphs**




**A SMILES with tags provides a way to recover index order**

```
[C:1]1=[C:3]([C:7]([C:4](=[C:2]([C@:6]1([C:8]([C:9]([C@:10]([N:12]([C:5](=[O:13])[O:14][N:11]
```

**2) Canonical SMILE**

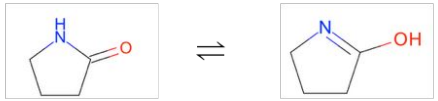
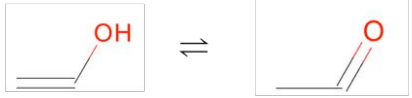
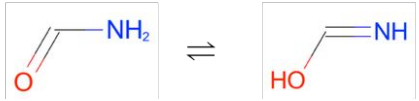
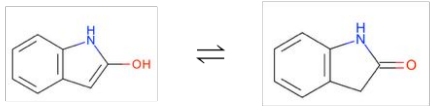
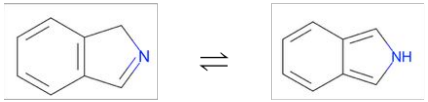
**Canonical SMILES are only canonical with respect to packages and package version.**

**`cmiles` pins toolkit version and will be distributed as Docker image**

**3) Standardized representation of all tautomers**

**`cmiles` also provides InChI and standardized SMILES from openeye and/or rdkit.**

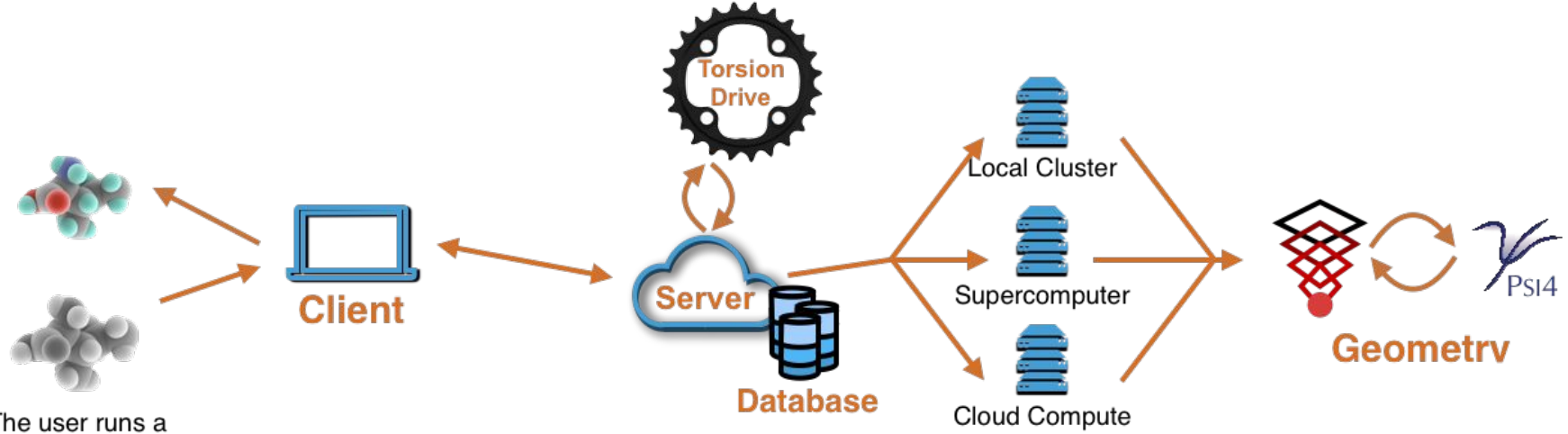
# Limitations of standard tautomer representation

| Tautomer | InChI | RDKit | OpenEye |
|---|---|---|---|
|  | ✓ | ✓ | ✓ |
|  | ✗ | ✓ | ✓ |
|  | ✓ | ✗ | ✓ |
|  | ✗ | ✓ | ✓ |
|  | ✗ | ✗ | ✗ |

QC Archive

A MolSSI Project

# Torsion Use Case

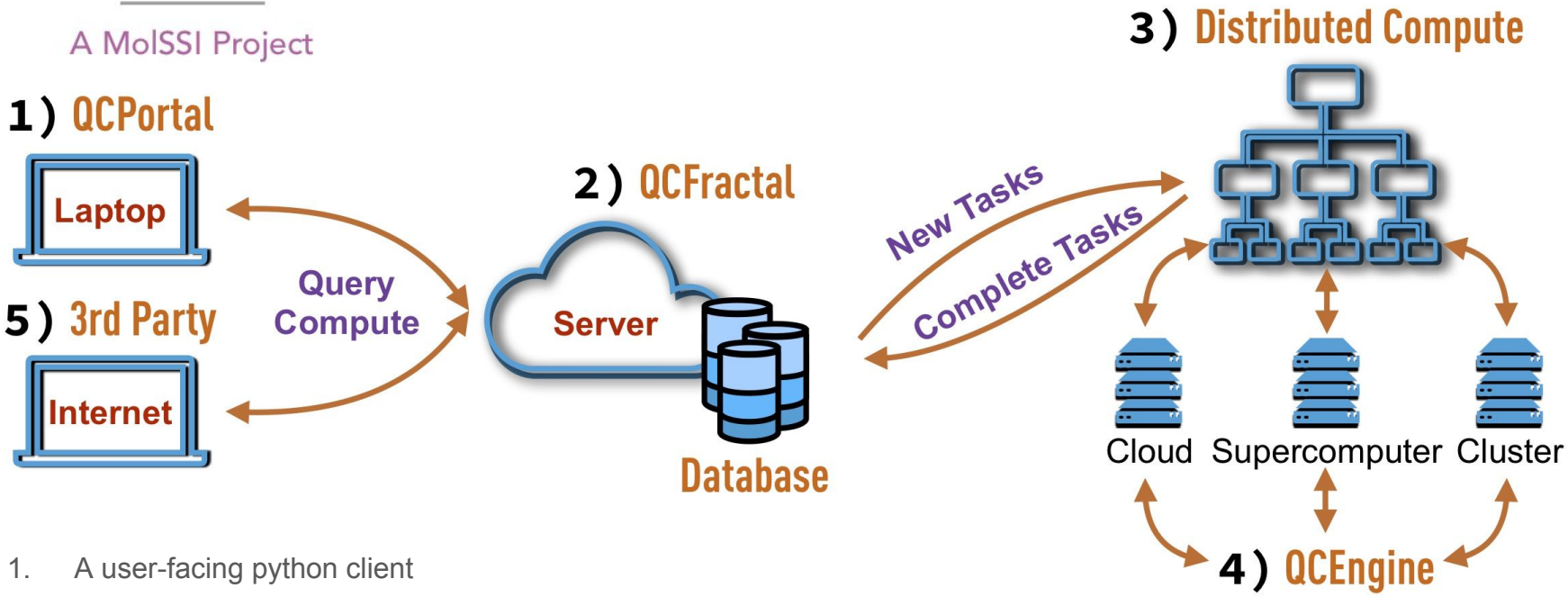



**A.** The user runs a client that requests quantum chemical computations for lists of molecules.

QCArchive

A MolSSI Project

# Software Overview




1. A user-facing python client
2. The primary database and task scheduler
3. A distributed workflow program (Dask, Balsam, RADICAL, AWS Batch, etc)
4. The compute agnostic evaluation engine
5. Other 3rd party REST services

# QCSchema

github.com/MolSSI/QCSchema

- Communication channel between all piece of the QC ecosystem.
- Community project useful for many aspects of quantum chemistry.
- Schema for Molecules, Results, Optimization, etc

```
{
   "molecule": {
        "geometry": [0, 0, 0, 0, 0, 1],
        "atoms": ["He", "He"]
    },
    "driver": "energy",
    "model": {
        "method": "SCF",
        "basis": "sto-3g",
    },
}
```

```
{     …Input
    "provenance": {
        "creator": "My QM Program",
        "version": "1.1rc1",
    },
    "properties": {
        "scf_n_iterations": 2.0,
        "scf_total_energy": -5.433191881443323,
        "nuclear_repulsion_energy": 2.11670883436,
        "one_electron_energy": -11.67399006298957,
    …
    },
    "error": "",
    "success": true,
    "raw_output": "Output storing was not requested."
}
```

# QCEngine

github.com/MolSSI/QCEngine

- Consumes and produces QCSchema
- Easily swap different backends (Force Fields, QC Programs, Semiemperical, AI Evaluation, etc)

```
>>> task = {
    "molecule": {
        "geometry": [0, 0, 0, 0, 0, 1],
        "atoms": ["He", "He"]
    },
    "driver": "energy",
    "model": {
        "method": "SCF",
        "basis": "sto-3g",
    },
}
```

```
>>> qcengine.compute(task, "psi4")
{    …Input
    "provenance": {
        "creator": "My QM Program",
        "version": "1.1rc1",
    },
    "properties": {
        "scf_n_iterations": 2.0,
        "scf_total_energy": -5.433191881443323,
        "nuclear_repulsion_energy": 2.11670883436,
        "one_electron_energy": -11.67399006298957,
    …
    },
    "error": "",
    "success": true,
    "raw_output": "Output storing was not requested."
}
```

# Accessing a workflow

Connect to a server

```
import qcportal

client = qcportal.FractalClient("198.82.19.66:12466")
```

Find Collections (workflows)

```
client.list_collections()
```

```
{'openffworkflow': ['chemper1_rdkit',
  'chemper1_psi4',
  'chemper2_rdkit',
  'chemper2_psi4'],
 'dataset': ['Water', 'Water_tmp']}
```

# Adding new compute

Pull a Collection from the server

```
wf = client.get_collection("openffworkflow", "chemper1_psi4")
```

Execute new compute (custom interface for OpenFF)

- Distributed (multi-cluster) computing
- Last run: 134 torsion scans, 2,200 cores on 3 clusters, ~85,000 quantum chemistry tasks

```
descr = {"job_0": {
            "type": "torsiondrive_input",
            "initial_molecule": ethane_molecule,
            "grid_spacing": [120],
            "dihedrals": [[0, 1, 2, 3]],
        }}
wf.add_fragment("[H:3][C:1]([H:4])([H:5])[C:2]([H:6])([H:7])[H:8]",  descr)
```

# Acquire Results

Access final energies, also available:

- Optimization trajectories
- Final geometries
- Individual dipole moments, bond orders, QC quantities
- etc

```
wf.list_final_energies(fragments=["[H:3][C:1]([H:4])([H:5])[C:2]([H:6])([H:7])[H:8]"])
```

```
{'[H:3][C:1]([H:4])([H:5])[C:2]([H:6])([H:7])[H:8]': {'job_0': {(180,): -79.77349108145843,
   (150,): -79.77123026331483,
   (-150,): -79.77123026357674,
   (120,): -79.76872302010649,
   (-120,): -79.76872301846865,
   (90,): -79.77122663113113,
   (-90,): -79.77122663120133,
   (60,): -79.77349259719931,
   (-60,): -79.77349259712148,
   (30,): -79.77122962645942,
   (-30,): -79.77122962640368,
   (0,): -79.76872370548993}}}
```

QC Archive

A MolSSI Project

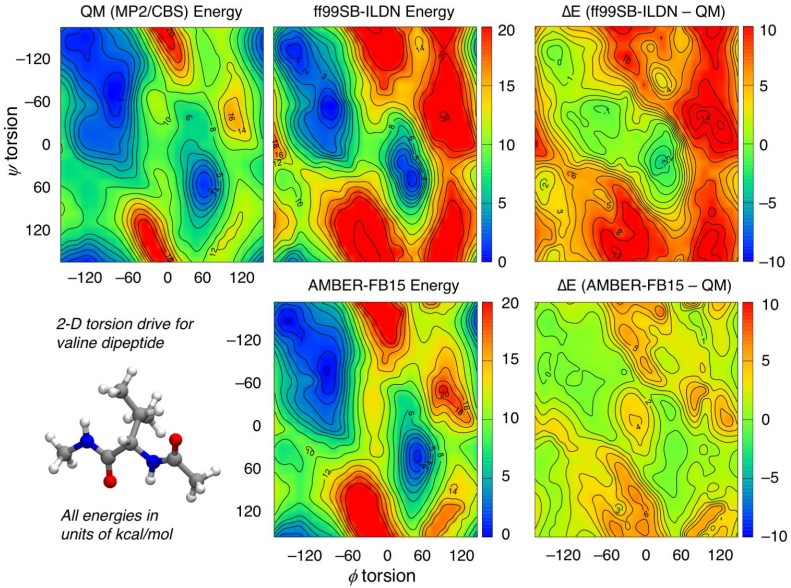
Provide an open, community-wide quantum chemistry database to both facilitate and capture hundreds of millions of hours of computing time to enable large-scale forcefield construction, physical property prediction, new methodology assessment, and machine learning from data that would otherwise end up siloed or inaccessible.

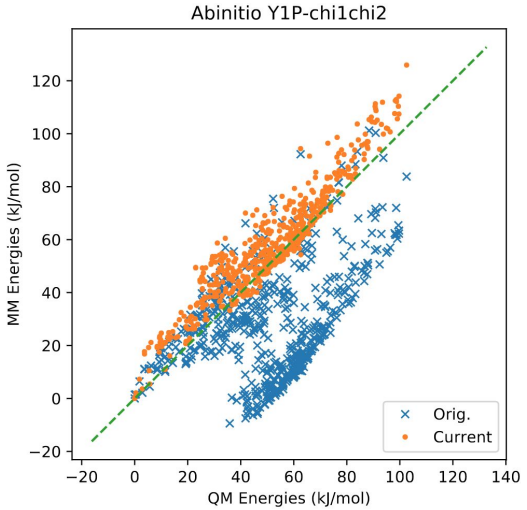
- Bond, Angle, Improper bond data analogs
- Electrostatic Potential (ESP) data
- Hessians and frequency computations next several months

- Capable of execution on hundreds of thousands of cores or a laptop
- MolSSI will host a central database that users can choose to use
- Deployable privately as well with database merge capabilities

# Passing data into ForceBalance

- Torsion drives can be directly translated into ForceBalance `AbInitio` targets.
  - Still need to automate QCArchive → FB pipeline.
- A previous generation of this approach was applied to build AMBER-FB15 force field (left).
- Currently working on adding parameters for phosphorylated AA's (right) in a separate project w/ Paul Nerenberg (Cal State LA) and John Stoppelman (GA Tech).
- Weight attenuates and then zeros out larger QM energies (attenuate at 5 kcal/mol, zero out at 20 kcal/mol); increased weight on conformations where E(MM) < E(QM), referenced to minimum QM energy structure.
- MM minimizations using fitted parameters provide important QM data added as extra targets.
  - This addresses and removes appearance of spurious MM minima far away from QM training data.




Scatter plot from John Stoppelman (GA Tech)

# Bespoke torsion fitting tool

Currently hiring a postdoc to develop a **bespoke torsion fitting tool** that will run fragmentation, torsion drives, quantum chemistry locally (or on a cluster) and refit all torsions in minutes to hours per molecule

Can likely make even more efficient if we process a **congeneric series of molecules** all at once, since many fragments are shared between molecules

# Questions

Would love input on what compound sets we prioritize initially for QM calcs

# Acknowledgements


GitHub:

- github.com/openforcefield/fragmenter
- github.com/openforcefield/cmiles
- https://github.com/MolSSI/QCFractal
- https://github.com/MolSSI/QCPortal
- https://github.com/MolSSI/QCEngine

Slack:

- #torsions
- #database