The MolSSI Quantum Chemistry Archive Project

Daniel G. A. Smith, Levi N. Naden, Doaa Altarawy, and Matt Welborne

The Molecular Sciences Software Institute

@dga_smith

qcarchive.molssi.org

QCArchive Overview

MolSSI

A central source to compile, aggregate, query, and share quantum chemistry data.

QCArchive

A MolSSI Project

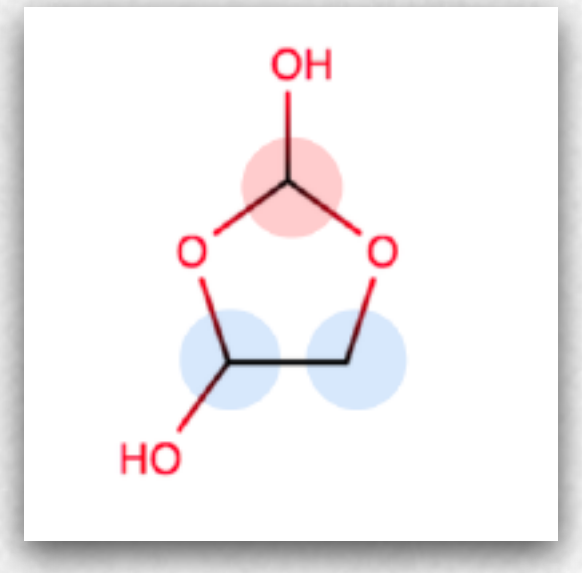
Design Goals

- Quantum chemistry data for all of the computational molecular sciences community.
- Analysis, visualization, and quick start data guides.
- Access data via Python, Jupyter notebook integration, REST API, and web apps.
- Evaluate on many community resources simultaneously.
- Store billions of quantum chemistry results.
- Prevent duplicate computation.
- Removing “the middle man”.

qcarchive.molssi.org

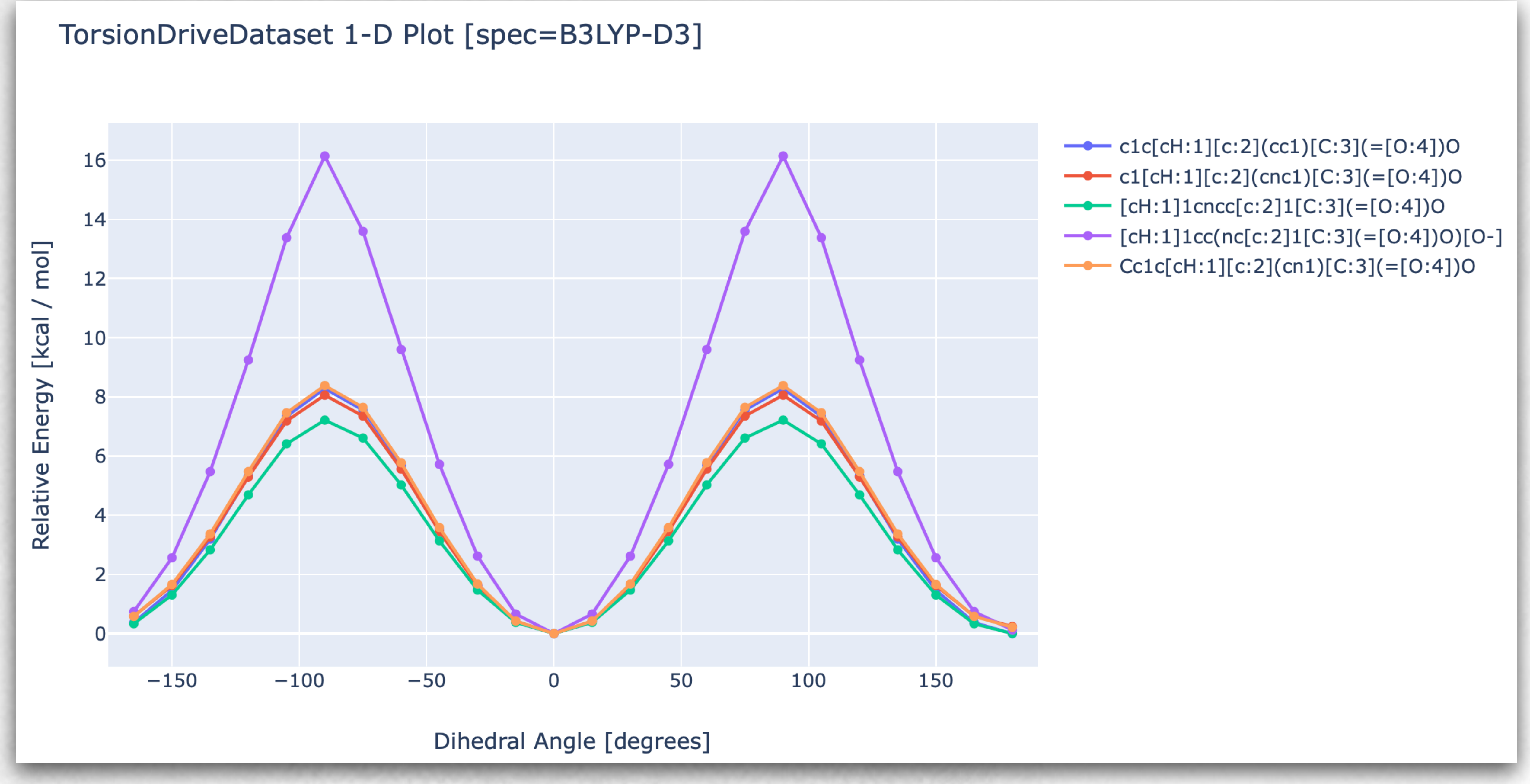
Open Force Field

https://openforcefield.org



Requirements

- Goal: Aggregate and compute open data for force field fitting, machine learning, and education.
- Store: Constrained geometry optimization, Torsion Drives, Hessians, partial charges, ESPs, and more!
- Search: SMILES, InChI, etc
- Compute: Multiple campus clusters, burst at XSEDE/DOE



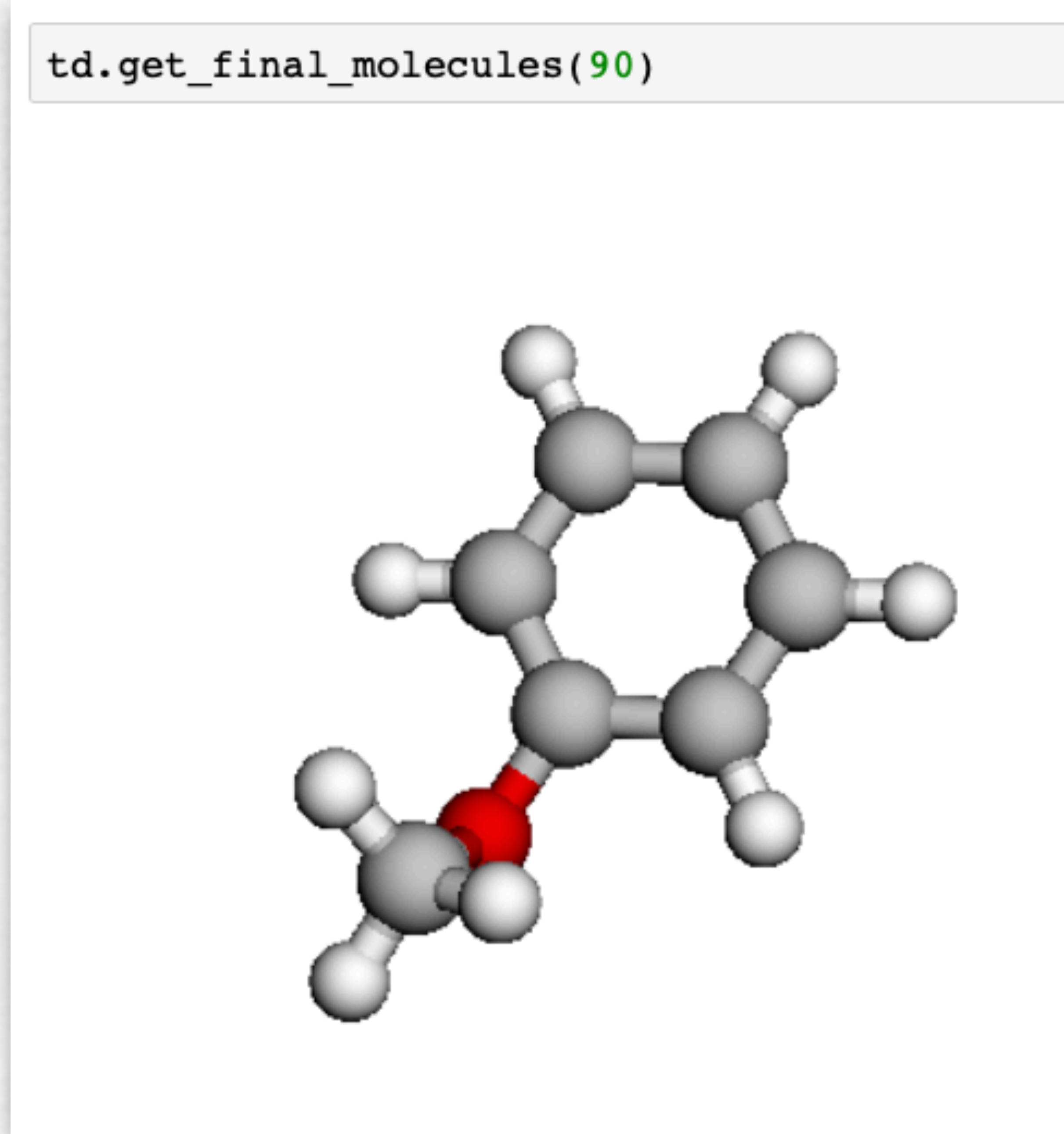
Computed (4 months)

- 2,400 torsion drives
- 170,000 geometry optimizations
- 50,000 Hessians
- ~60,000 geometry optimizations/ week [limited by core time]

QCArchive Sponsor

- Sponsoring features within QCArchive
- All open-source code, available to the community

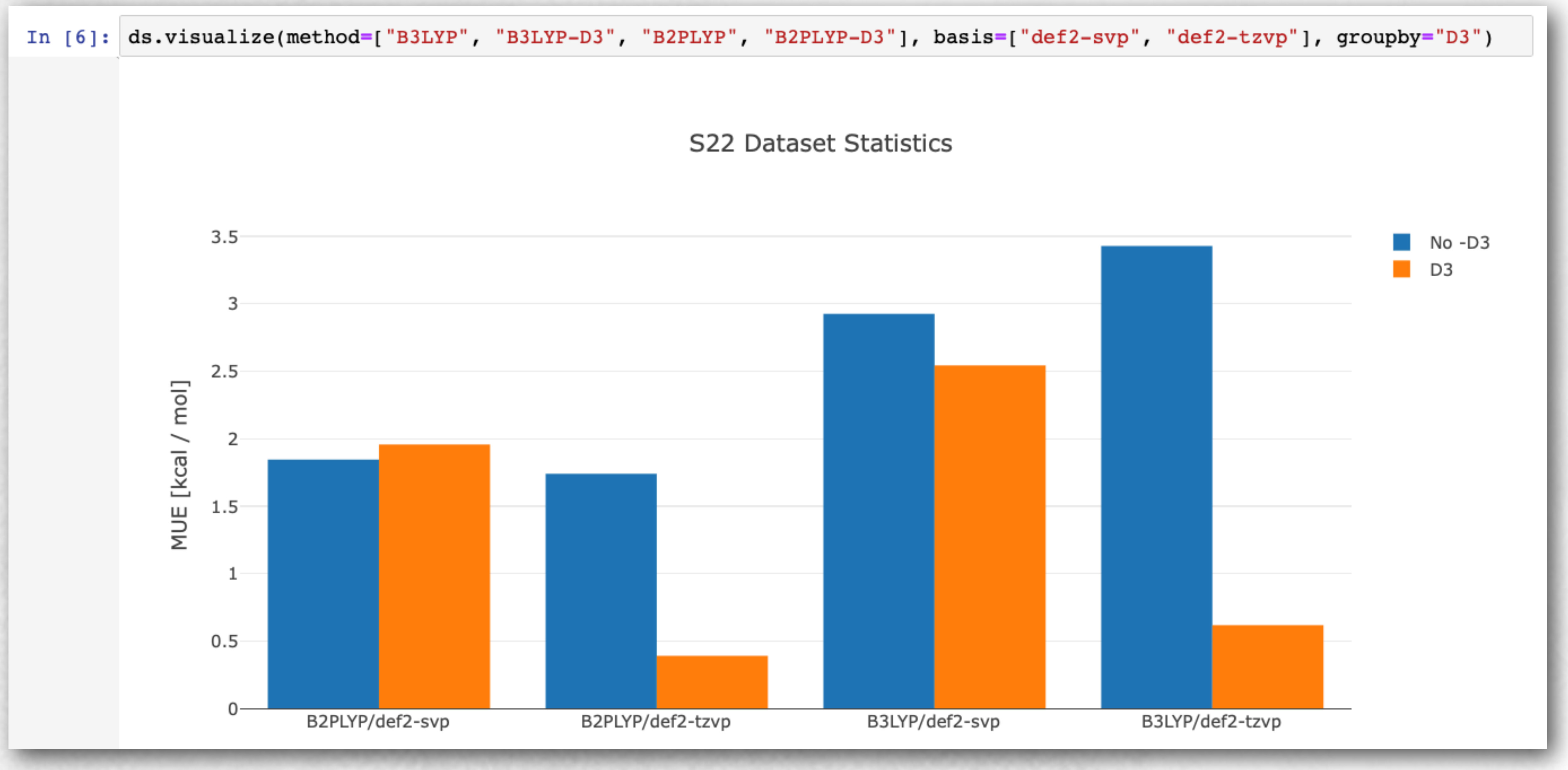
Interactive and Gateway Sessions



```
td.get_final_molecules(90)
```

Jupyter-Notebook Integration

- Molecular visualization, statistics, trajectories, etc
- Utilizing community-built and industry standard tools
- Interactive sessions to explore and compute new data
- Leveraging the greater Jupyter community of tools



```
In [6]: ds.visualize(method=["B3LYP", "B3LYP-D3", "B2PLYP", "B2PLYP-D3"], basis=["def2-svp", "def2-tzvp"], groupby="D3")
```

Web Apps

- Working in partnership with the Science Gateways Community Institute.
- Web-based statistics and visualization
- Targeting at CMS researchers and undergraduate educational initiatives
- Data-driven initiatives:
    - What is the best method for X
    - How long will X take?

SGCI
Science Gateways
Community Institute

Engage with QCArchive

View our data

- Browse our current 5.5M+ results and growing!
- Contribute “cookbook” use cases for interesting data applications.

Compute Open Data

- Work with us to compute additional open data.
- Expand our use cases for a web app framework.

QCArchive

A MolSSI Project

Extend our datasets

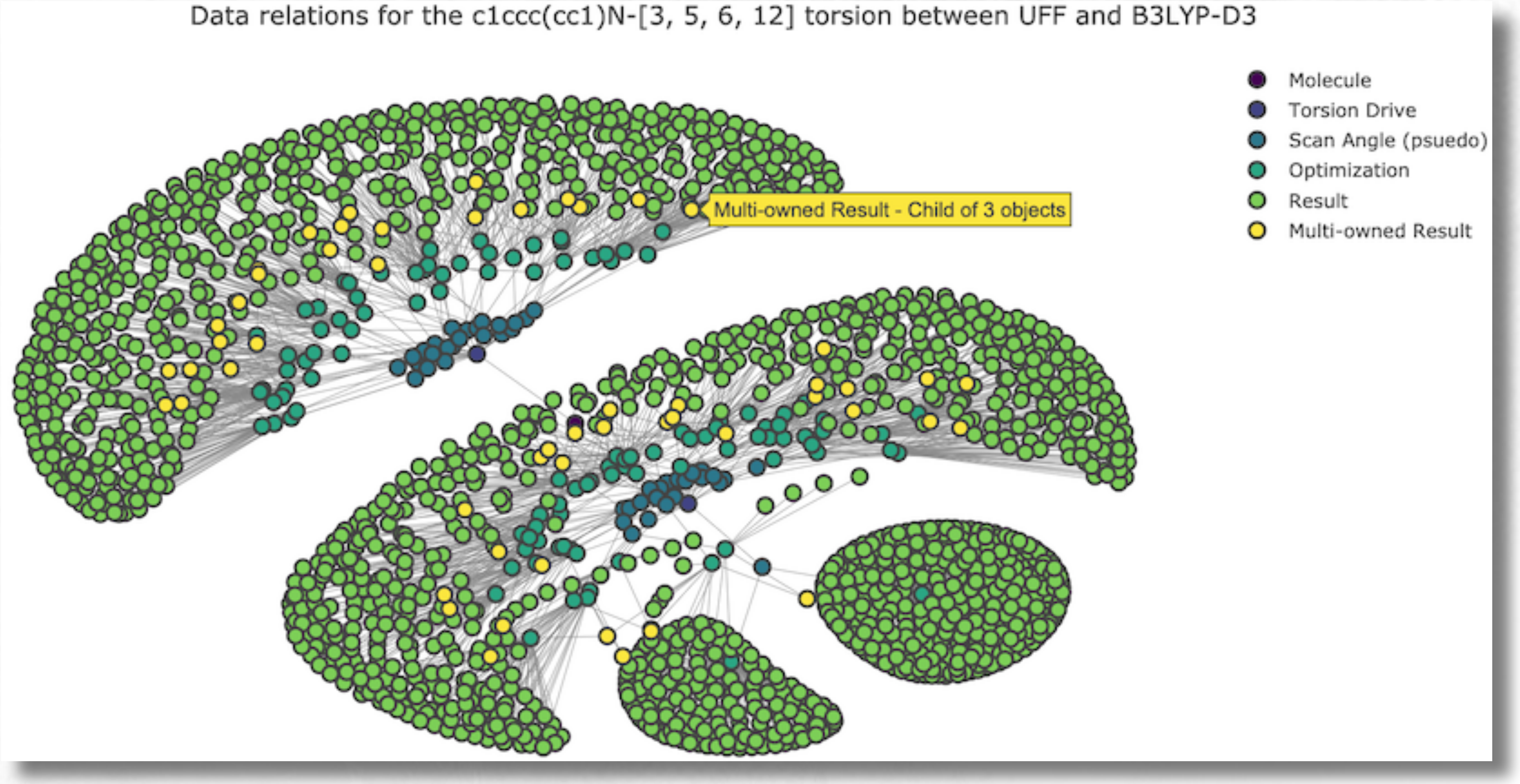
- Get in touch and help compute additional methods for our benchmark datasets.
- Add additional benchmark datasets to the Archive.

Talk to us

- Tell us about science applications that we are missing.
- Chat about how to refine current presentations and ideas.

QCArchive Infrastructure

- Quantum chemistry software projects for all CMS developers
- Composable building blocks
- Developed openly on GitHub (github.com/MolSSI)
- Used by dozens of downstream programs



QCSchema

- Standardized IO for quantum chemistry

QCElemental

- Units
- QCSchema Models
- Molecule Parsing
- Visualization

QCEngine

- Consume and produce QCSchema for many programs
- Not just quantum chemistry

QCFractal

- High-throughput quantum chemistry
- Common pipelines
- Data Organization
- Visualization

QCSchema

https://github.com/MolSSI/QCSchema

- Communication channel between all piece of the ecosystem.
- *Community* project useful for many aspects of quantum chemistry.
- Not only JSON, but any key/value/array language (BSON/HDF5/XML/YAML/msgpack/parquet)

- Molecule
- QC Input/Output
- Optimization Structures
- Wavefunction Quantities

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

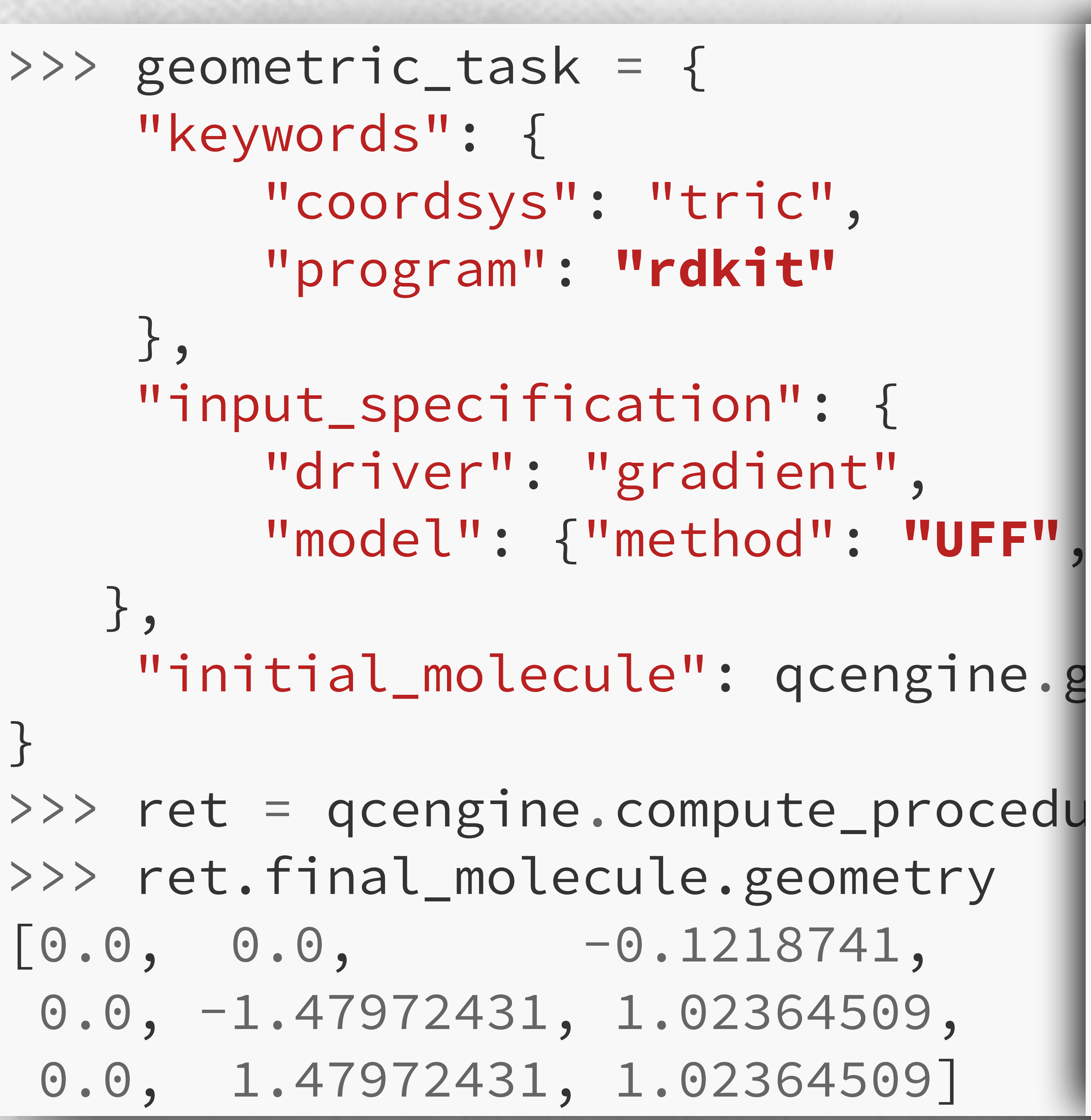
```
{
    …Input
    "provenance": {
        "creator": "My QM Program",
        "version": "1.1rc1",
    },
    "properties": {
        "scf_n_iterations": 2.0,
        "scf_total_energy": -5.433191881443323
        "nuclear_repulsion_energy”: 2.11670883
        “one_electron_energy": -11.67399062298
    …
    },
    "error": "",
    "success": true,
```

QCEngine

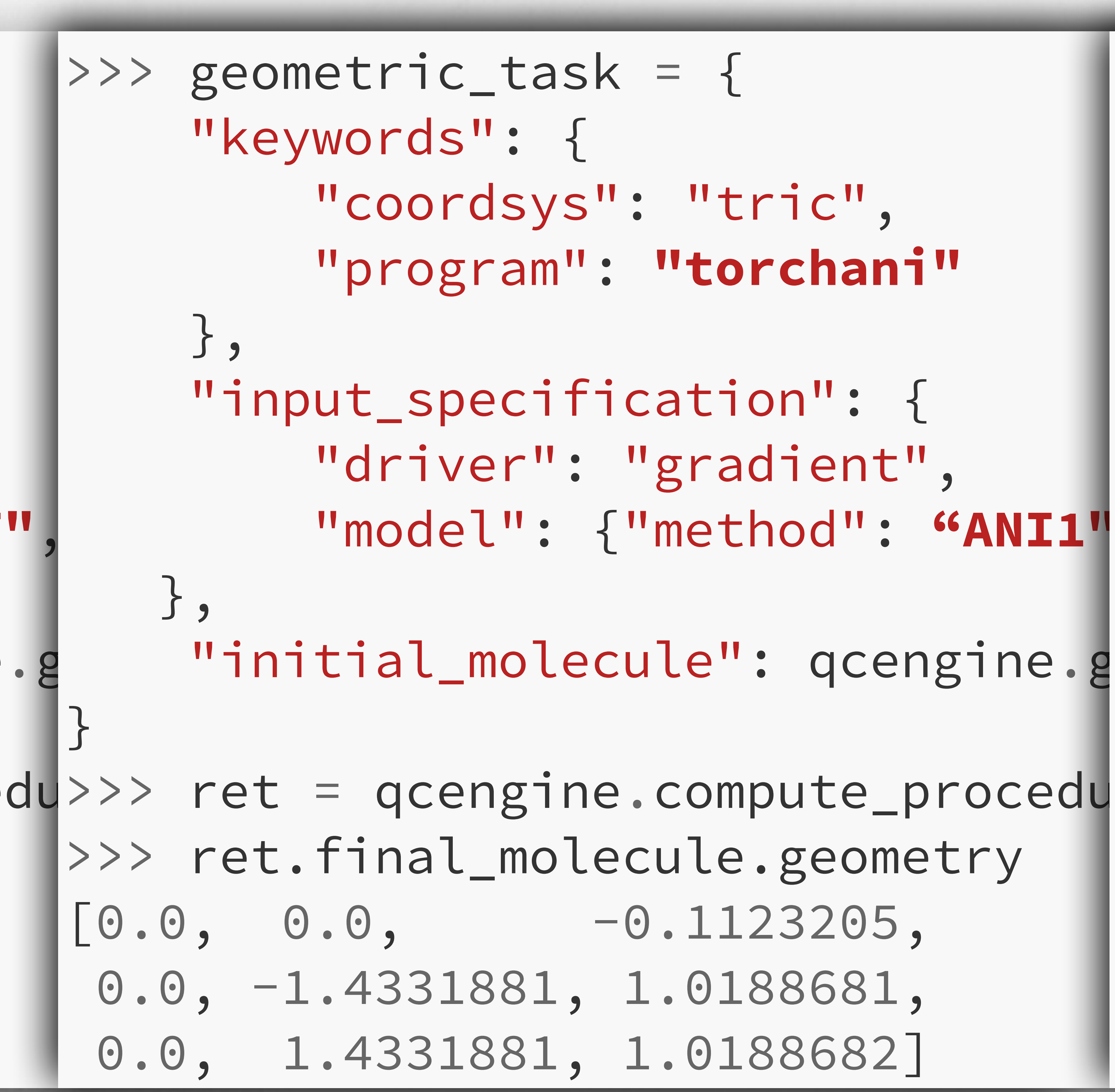
https://github.com/MolSSI/QCEngine

- Quantum chemistry, semiempirical, AI energy evaluator, and force field agnostic backend to produce/consume Schema. Effectively our compute abstraction layer.

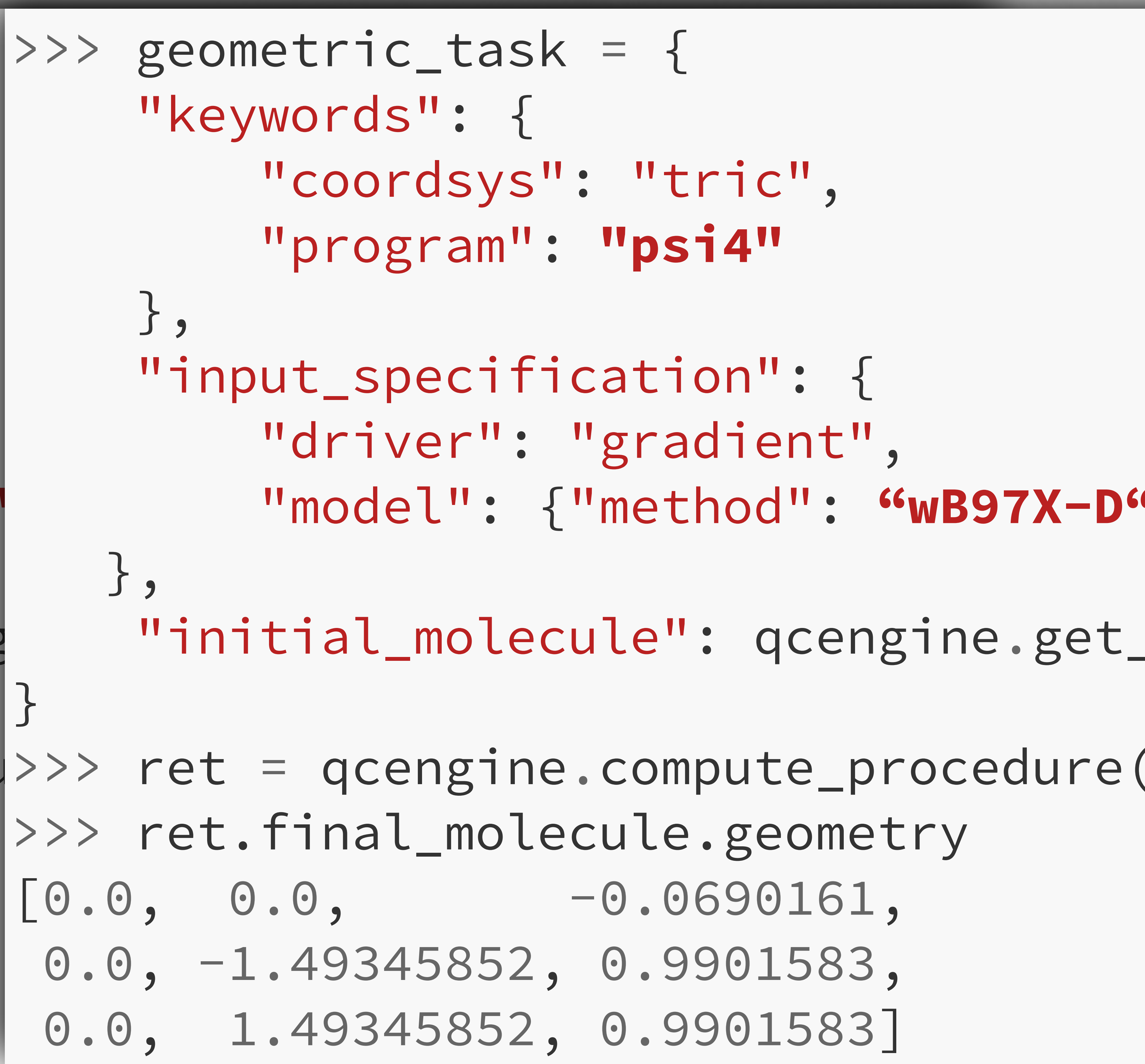
- Modular building block approach



```
>>> geometric_task = {
    "keywords": {
        "coordsys": "tric",
        "program": "rdkit"
    },
    "input_specification": {
        "driver": "gradient",
        "model": {"method": "UFF",
   },
    "initial_molecule": qcengine.g
}
>>> ret = qcengine.compute_procedu
>>> ret.final_molecule.geometry
[0.0,  0.0,        -0.1218741,
 0.0, -1.47972431, 1.02364509,
 0.0,  1.47972431, 1.02364509]
```



```
>>> geometric_task = {
    "keywords": {
        "coordsys": "tric",
        "program": "torchani"
    },
    "input_specification": {
        "driver": "gradient",
        "model": {"method": “ANI1"
   },
    "initial_molecule": qcengine.g
}
>>> ret = qcengine.compute_procedu
>>> ret.final_molecule.geometry
[0.0,  0.0,       -0.1123205,
 0.0, -1.4331881, 1.0188681,
 0.0,  1.4331881, 1.0188682]
```



```
>>> geometric_task = {
    "keywords": {
        "coordsys": "tric",
        "program": "psi4"
    },
    "input_specification": {
        "driver": "gradient",
        "model": {"method": “wB97X-D“
   },
    "initial_molecule": qcengine.get_
}
>>> ret = qcengine.compute_procedure(
>>> ret.final_molecule.geometry
[0.0,  0.0,        -0.0690161,
 0.0, -1.49345852, 0.9901583,
 0.0,  1.49345852, 0.9901583]
```

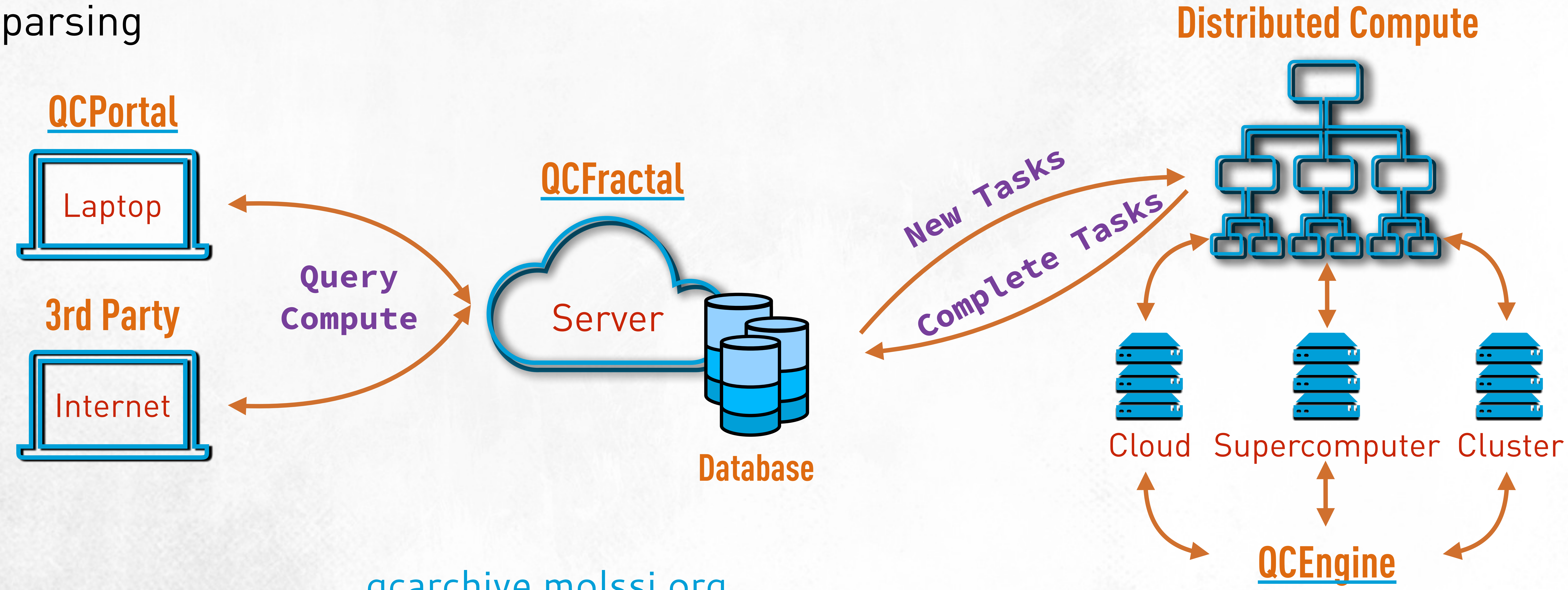
QCFractal

https://github.com/MolSSI/QCFractal

QC Archive

A MolSSI Project

Goals:
- High-throughput quantum chemistry
- Laptop to campaign-scale compute orchestration
- Procedures run with a variety of different programs
- Organize data with common abstraction and collection layers
- Share and collaborate structured data
- Ease of use, less data parsing

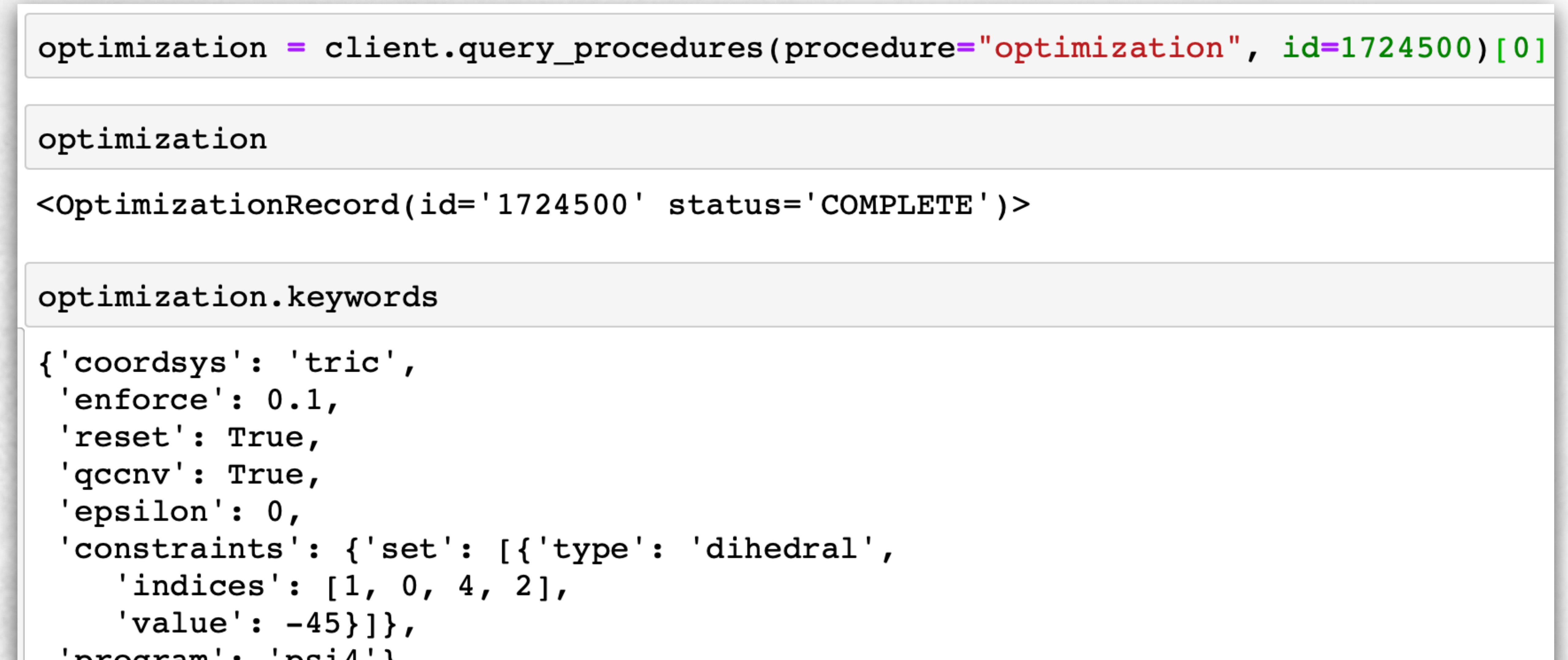


qcarchive.molssi.org

Reproducible Procedures and Workflows

Procedures

- Procedures = small reproducible series of computations
- Exact input of pipeline and version data available
- Geometry optimizations, torsion evaluations, finite difference computations, spectral computations, etc



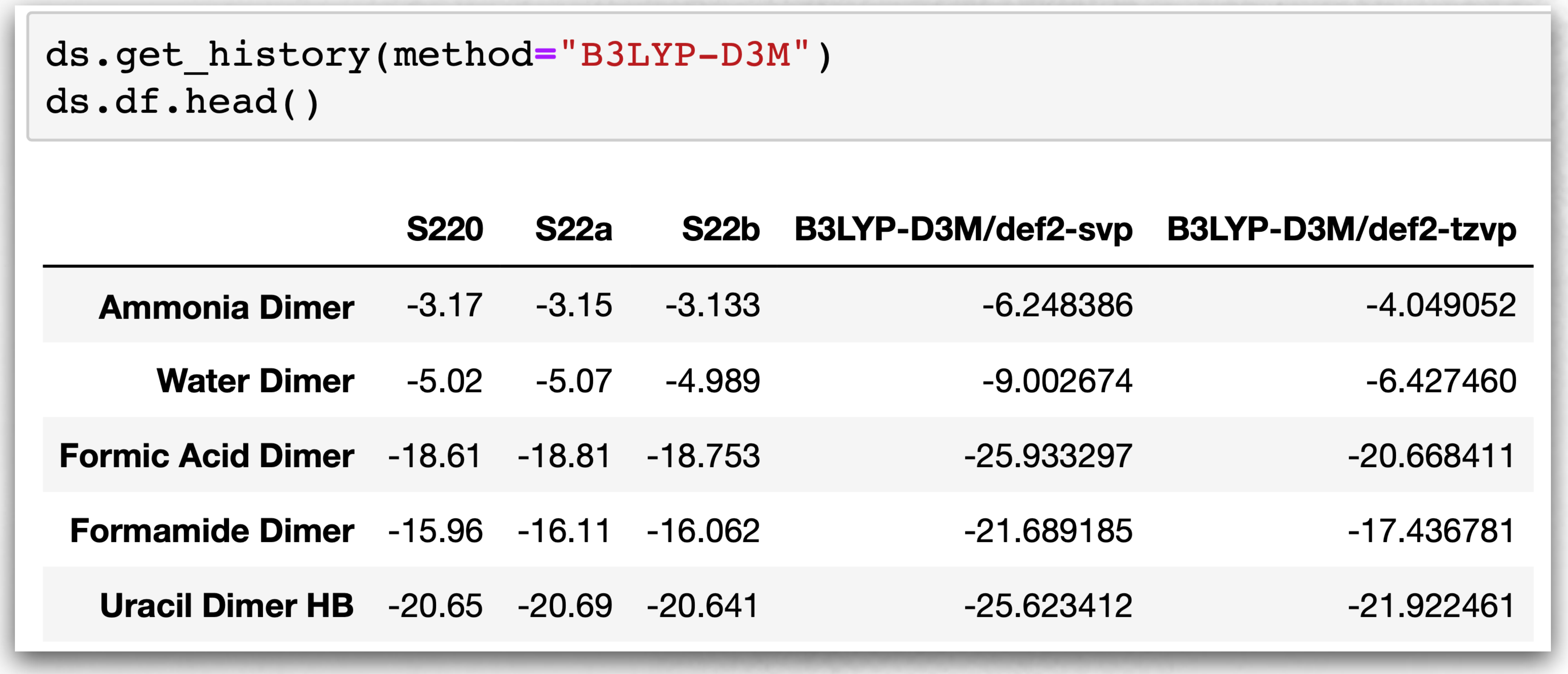
```
optimization = client.query_procedures(procedure="optimization", id=1724500)[0]
```

```
optimization
```

```
<OptimizationRecord(id='1724500' status='COMPLETE')>
```

```
optimization.keywords
```

```
{'coordsys': 'tric',
 'enforce': 0.1,
 'reset': True,
 'qccnv': True,
 'epsilon': 0,
 'constraints': {'set': [{'type': 'dihedral',
    'indices': [1, 0, 4, 2],
    'value': -45}]},
 'program': 'psi4'}
```



```
ds.get_history(method="B3LYP-D3M")
ds.df.head()
```

	S220	S22a	S22b	B3LYP-D3M/def2-svp	B3LYP-D3M/def2-tzvp
Ammonia Dimer	-3.17	-3.15	-3.133	-6.248386	-4.049052
Water Dimer	-5.02	-5.07	-4.989	-9.002674	-6.427460
Formic Acid Dimer	-18.61	-18.81	-18.753	-25.933297	-20.668411
Formamide Dimer	-15.96	-16.11	-16.062	-21.689185	-17.436781
Uracil Dimer HB	-20.65	-20.69	-20.641	-25.623412	-21.922461

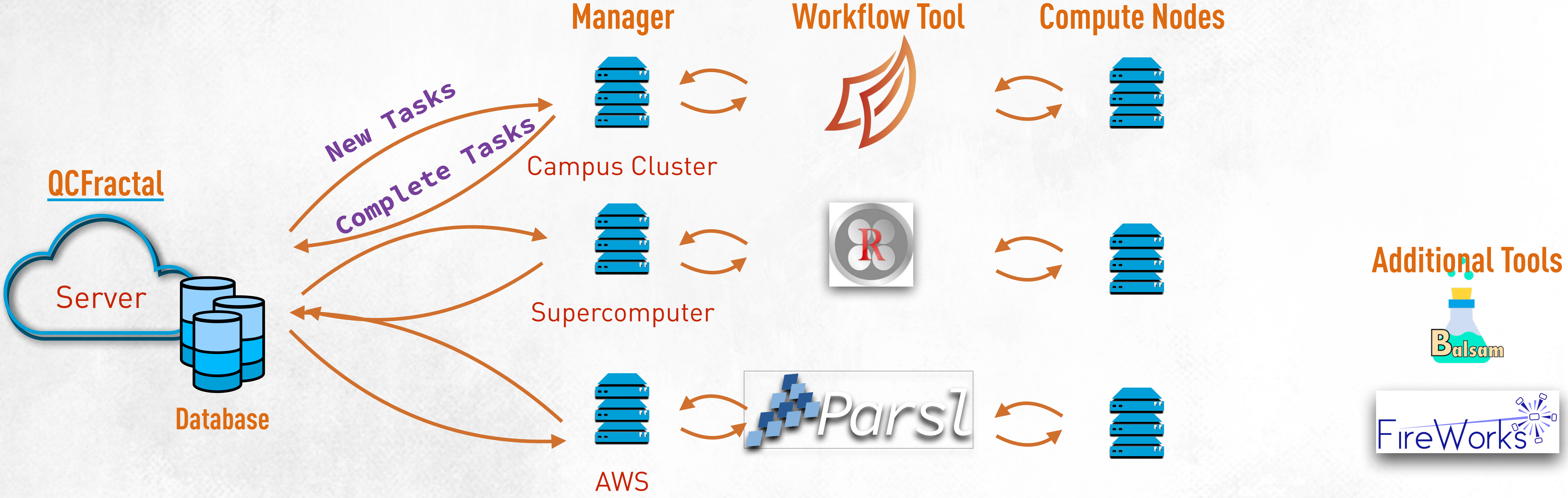
Collections

- After 100+ interviews there seems to be very little common ground on data organization.
- Many single computations or procedure grouped together known as ***Collections***
- Reproducible, recomputable, and tweakable
- Data organization for ML, methodology assessment, forcefield creation, etc

Distributed compute

QCArchive
A MolSSI Project

- Multi-physical-site compute or a single laptop
- Scale up to 500 tasks/second, 300,000 concurrent tasks @ 10 minutes each
- Setup once and walk away
- Managers:
  - Runs on head node or local compute
  - Smart task acquisition
  - ~20 ms per-task overhead



Manager

Workflow Tool

Compute Nodes

New Tasks

Complete Tasks

Campus Cluster

QCFractal

Server

Database

Supercomputer

AWS

Parsl

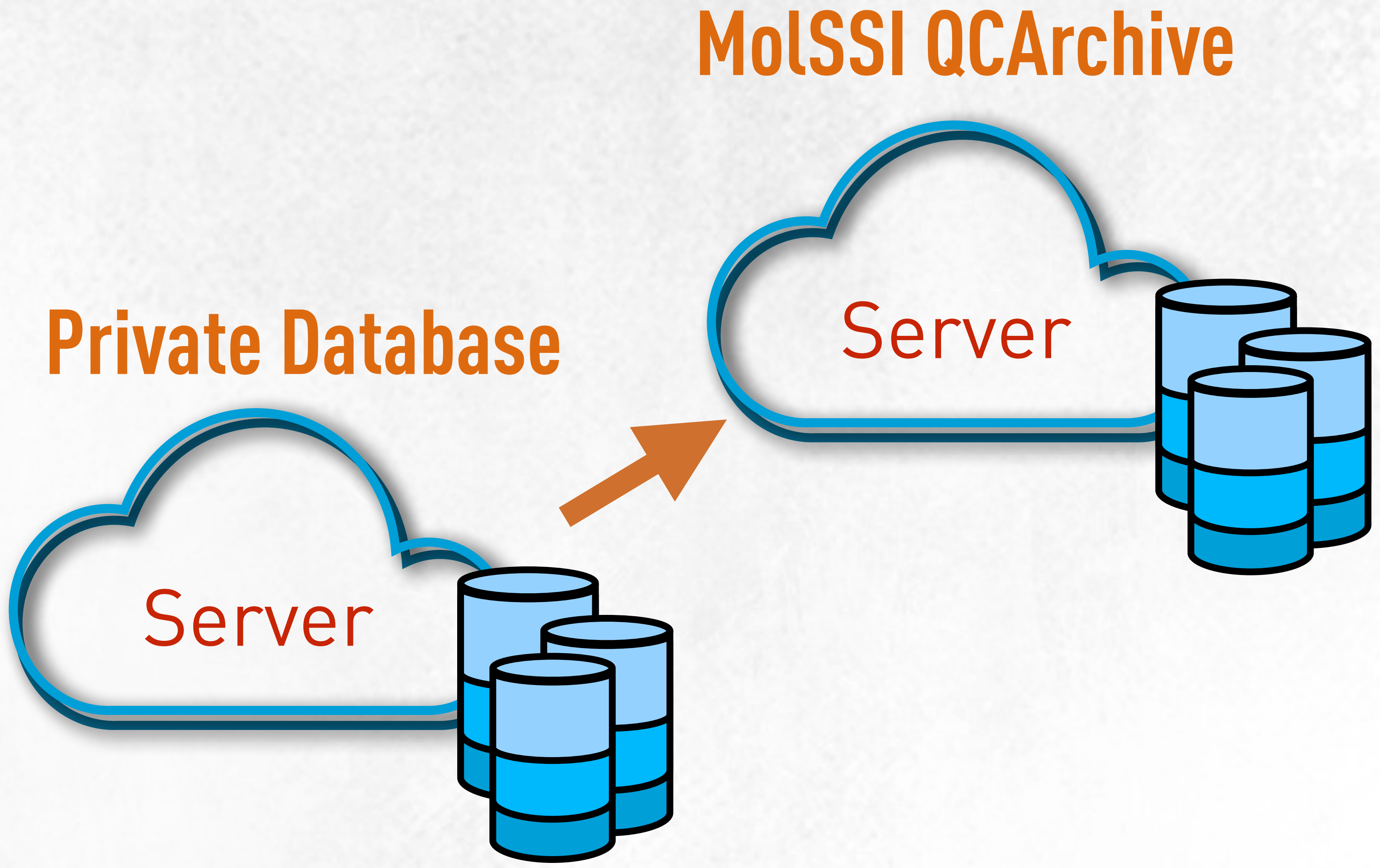
Additional Tools

Balsam

FireWorks

MolSSI and Self-Hosted Databases

- A domain specific SQL database layer
- Generation and computation of new quantum chemistry tasks
- Central MolSSI-hosted server for community data accessed via REST or Python API
- Open-software (QCFractal) used at scale at MolSSI, research groups, and individuals



Self-Hosted

- Long-term private data with access controls
- (or) Quick testing and evaluation environments
- Can migrate data to central MolSSI server after publication
- Identical infrastructure and technology as MolSSI central repository

MolSSI QCArchive

- Open community data
- FAIR Data standards
- ~5.5M current results
- ~60 community datasets
- Can host ~1B results with current hardware, looking to expand!

QCArchive

A MolSSI Project

MERCURY openstax SGCI Science Gateways Community Institute The Materials Project materialsproject.org International Journal of Quantum Chemistry

Gateways Portal

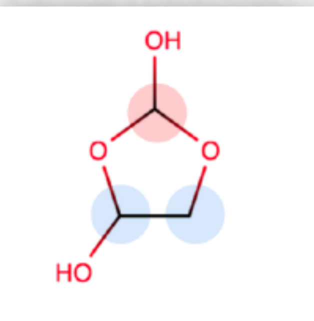
- Reach non-CMS community
- Research-focused web portals
- Educational initiatives

Kitware 3dMolJs pandas jupyter plotly

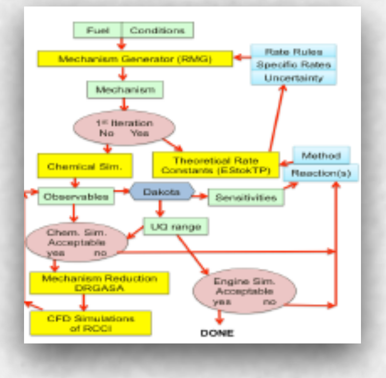
View and Analyze data

- Large-scale analytics
- Valuable community insights

The MolSSI Community Database



The Open Force Field Consortium



PACChem

The Sherrill Group

Private Databases

- Custom ML Datasets
- Methodology assessment
- Modern access to QC computations

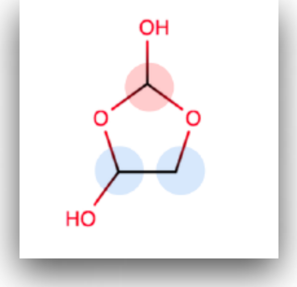
Beta as of August 26th
Monthly substantial releases
Use cases from 100+ research groups
60,000+ downloads
4.5M computations
Rapidly expanding

CMS and Community Software

pandas PSI4 OPEN-SOURCE QUANTUM CHEMISTRY PostgreSQL $\hat{H}(\vec{R})|\Psi\rangle = E|\Psi\rangle$ TORCHANI NWChem Molpro ⟨CC|CC⟩ Q-CHEM

Software Developers

HJK grp @MIT QUBE Kit geomeTRIC PSI4 OptKing



Cyberinfrastructure

Parsl FireWorks Balsam

Thank You!

- Doaa Altarawy (MolSSI)
- Levi Naden (MolSSI)
- Matt Wellborne (MolSSI)
- Lori Burns (Georgia Tech)
- Sam Ellis (MolSSI)
- Jessica Nash (MolSSI)
- Ben Pritchard (MolSSI)
- Chaya Sten (MSKCC)
- Yudong Qiu (UC Davis)

- Fang Liu (MIT)
- Sebastian Lee (Cal Tech)
- David Sherrill (Georgia Tech)
- Daniel Crawford (MolSSI)
- Lee-Ping Wang (UC Davis)
- Jeff Wagner (UCI)
- John D. Chodera (MSKCC)
- Dom Sirianni (Georgia Tech)
- Daniel Nascimento (PNNL)

- Nick Petosa (Microsoft)
- Justin Turney (UGA)
- Bert de Jong (LBNL)
- Theresa Windus (Iowa State)
- Aaron Virshup (Arzeda)
- Marcus Hanwell (Kitware)
- Shantenu Jha (Rutgers)
- Matteo Turilli (Rutgers)
- Kyle Chard (U. Chicago)

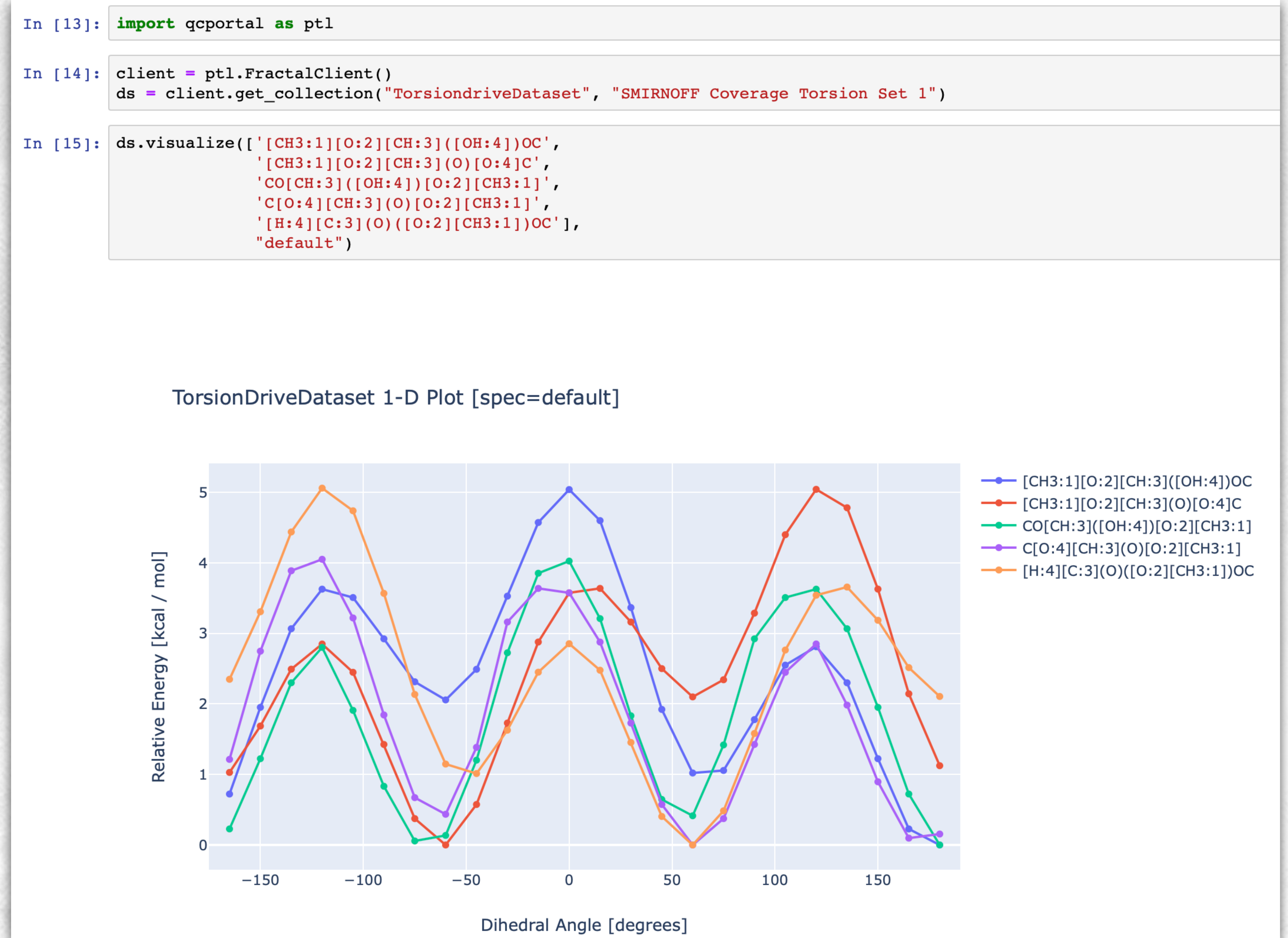
QCArchive

A MolSSI Project

qcarchive.molssi.org

MolSSI

molssi.org



```
In [13]: import qcportal as ptl
```

```
In [14]: client = ptl.FractalClient()
         ds = client.get_collection("TorsiondriveDataset", "SMIRNOFF Coverage Torsion Set 1")
```

```
In [15]: ds.visualize(['[CH3:1][O:2][CH:3]([OH:4])OC',
                       '[CH3:1][O:2][CH:3](O)[O:4]C',
                       'CO[CH:3]([OH:4])[O:2][CH3:1]',
                       'C[O:4][CH:3](O)[O:2][CH3:1]',
                       '[H:4][C:3](O)([O:2][CH3:1])OC'],
                       "default")
```

TorsionDriveDataset 1-D Plot [spec=default]

Relative Energy [kcal / mol]
0
1
2
3
4
5
−150
−100
−50
0
50
100
150
Dihedral Angle [degrees]
[CH3:1][O:2][CH:3]([OH:4])OC
[CH3:1][O:2][CH:3](O)[O:4]C
CO[CH:3]([OH:4])[O:2][CH3:1]
C[O:4][CH:3](O)[O:2][CH3:1]
[H:4][C:3](O)([O:2][CH3:1])OC

Notebook Demonstration

https://docs.qcarchive.molssi.org/en/latest/basic_examples/torsiondrive_datasets.html