The Impact of Psi4 and QC* on Open Force Field Development

Pavan Behara
Project Scientist, UCI Center for Neurotherapeutics
Openforcefield@David Mobley's Lab
Dec 8 2023

OpenFF partners with industry to build accurate, pre-competitive force fields for general use

INDUSTRY

AbbVie
Amgen
AstraZeneca
Bayer
Cresset
Eli Lilly
Janssen
OpenEye
Pfizer
Redesign
Roche
Ventus
Vertex
... and others

Open Molecular Software Foundation

PROJECT STAFF

Jeff Wagner
Technical Lead

Lily Wang
Science Lead

ACADEMIC

John Chodera (MSKCC)

Michael Gilson (UC San Diego)

David Mobley (UC Irvine)

Michael Shirts (CU Boulder)

Plus affiliates:

- Danny Cole (Newcastle)
- Lee-Ping Wang (UCD)
- Dennis Della Corte (BYU)
- MolSSI (Virginia Tech)

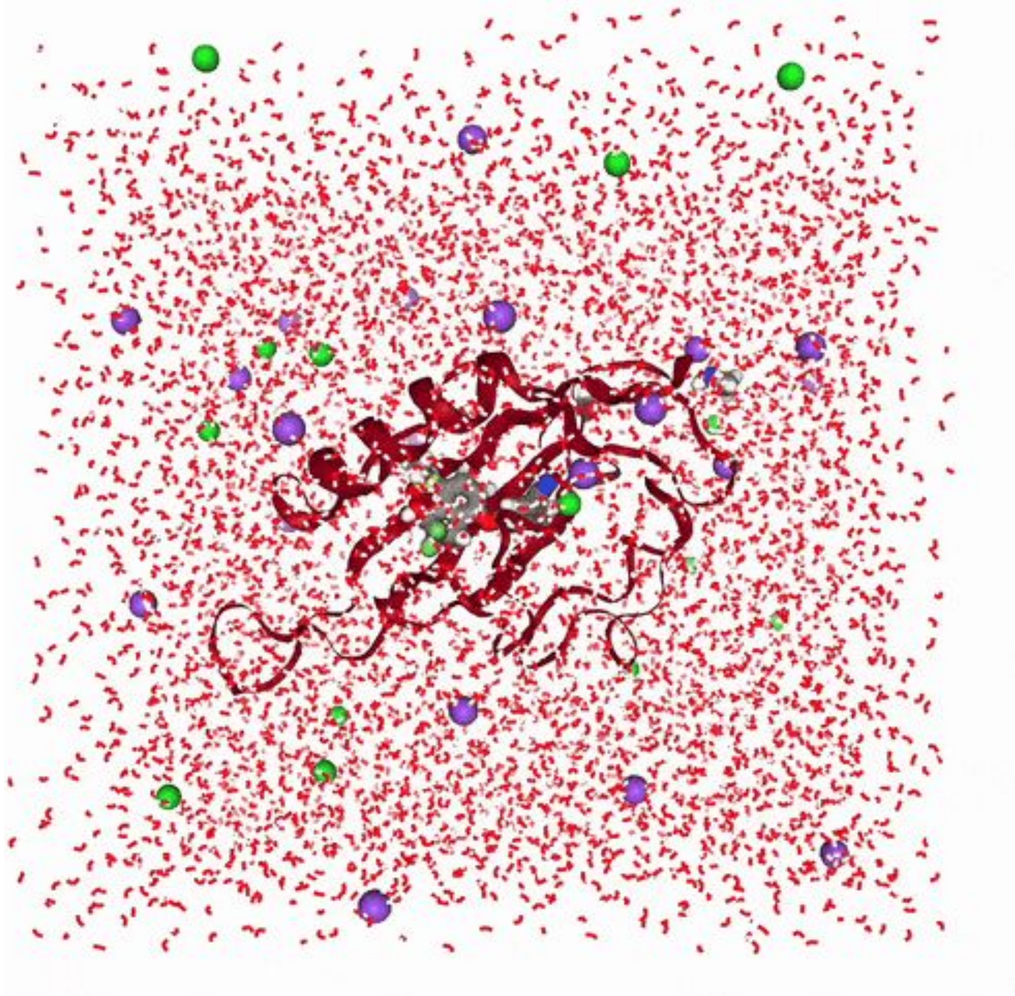
Open Force Field makes it easy to run MM simulations

OpenFF Code

```
from openff.toolkit import Molecule, Topology, ForceField
ligand = Molecule.from_file('inputs/PT2385.sdf')
top = Topology.from_pdb('inputs/solvated_complex.pdb',
                        unique_molecules=[ligand]
                        )
ff = ForceField("openff-2.0.0.offxml", "ff14sb_off_impropers_0.0.3.offxml")
sys = ff.create_openmm_system(top)
```

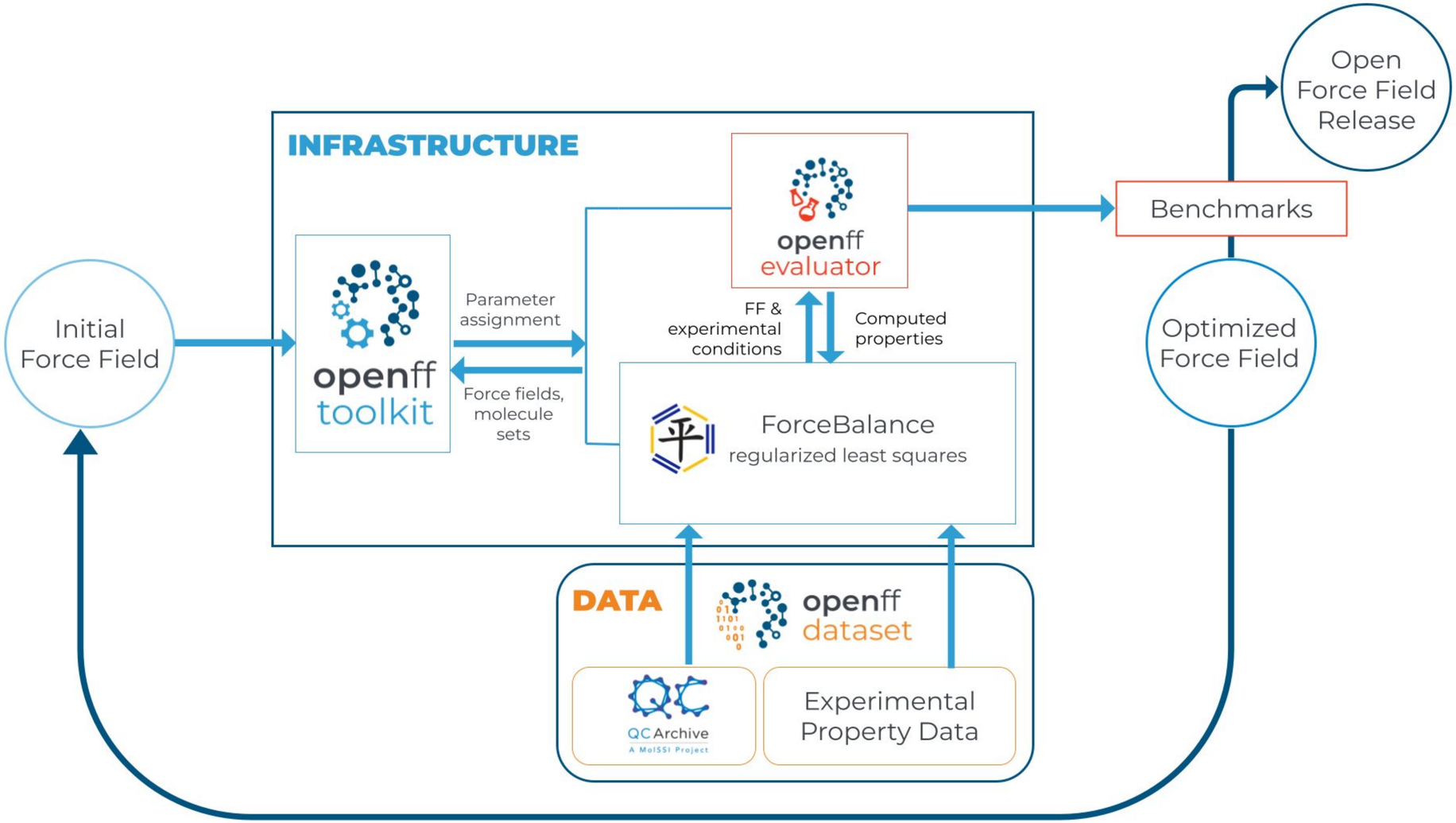
OpenMM Code

```
import openmm
from openff.units import Quantity, unit
from openmm import unit as openmm_unit

# Construct and configure a Langevin integrator at 300 K with an appropriate friction consta
integrator = openmm.LangevinIntegrator(
    300 * openmm_unit.kelvin,
    1 / openmm_unit.picosecond,
    0.002 * openmm_unit.picoseconds,
)

# Combine the topology, system, integrator and initial positions into a simulation
simulation = openmm.app.Simulation(top.to_openmm(), sys, integrator)
simulation.context.setPositions(top.get_positions().to_openmm())

# Add a reporter to record the structure every 10 steps
dcd_reporter = openmm.app.DCDReporter("trajectory.dcd", 250)
simulation.reporters.append(dcd_reporter)
simulation.context.setVelocitiesToTemperature(300 * openmm_unit.kelvin)
simulation.runForClockTime(1.5 * openmm_unit.minute)
```

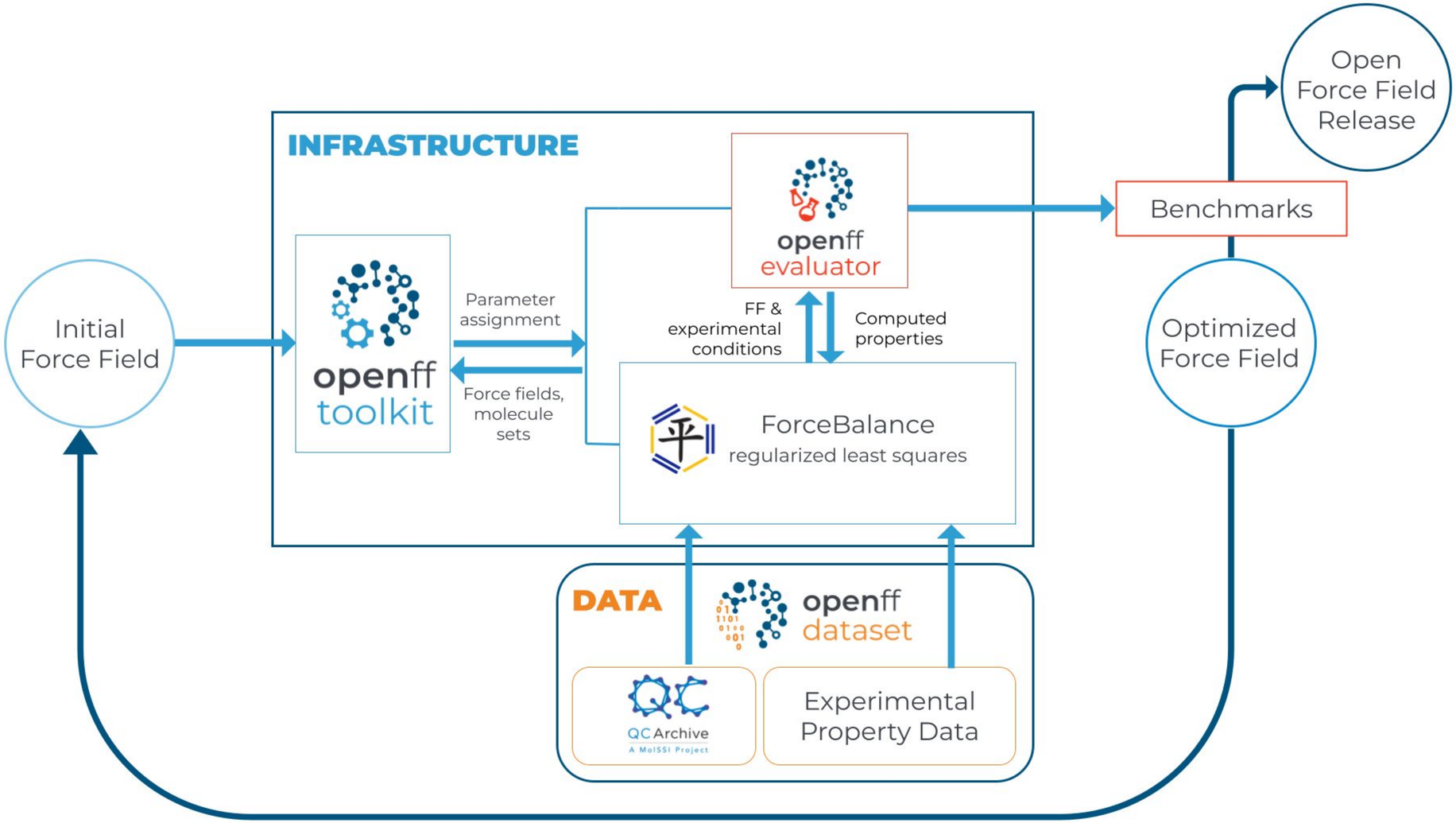


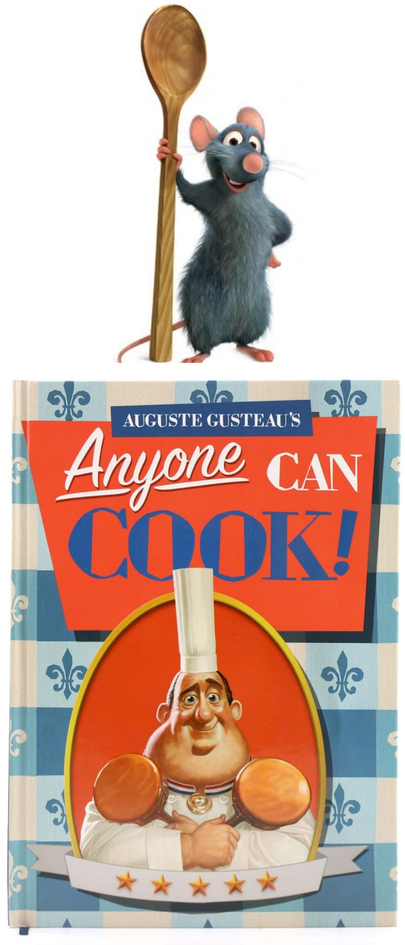
https://github.com/openforcefield/2023-workshop-vignettes/blob/main/G-PDB-to-simulation.ipynb

We automated much of the process of Force field development

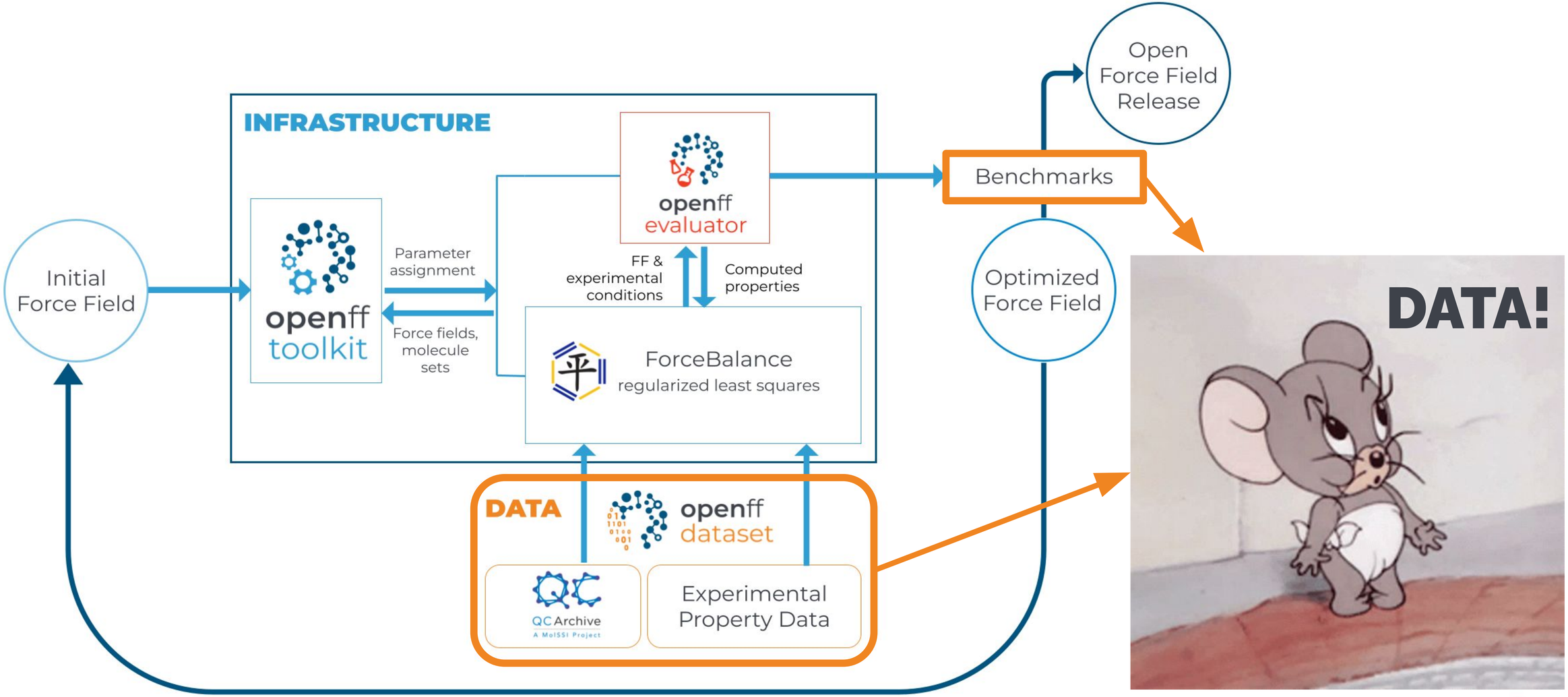


Our tools are permissively licensed, industry standard, and community driven





Force field fitting and benchmarking needs a lot of QM data



There will be data!

Psi4 is our calculation engine for QM

- QM datasets we generate in our work
  - » Geometry optimizations of conformers
  - » Torsion scans (1D, 2D)
  - » Single point energies and gradients
  - » Hessians
  - » Electrostatic potential surfaces

Psi4

geomeTRIC

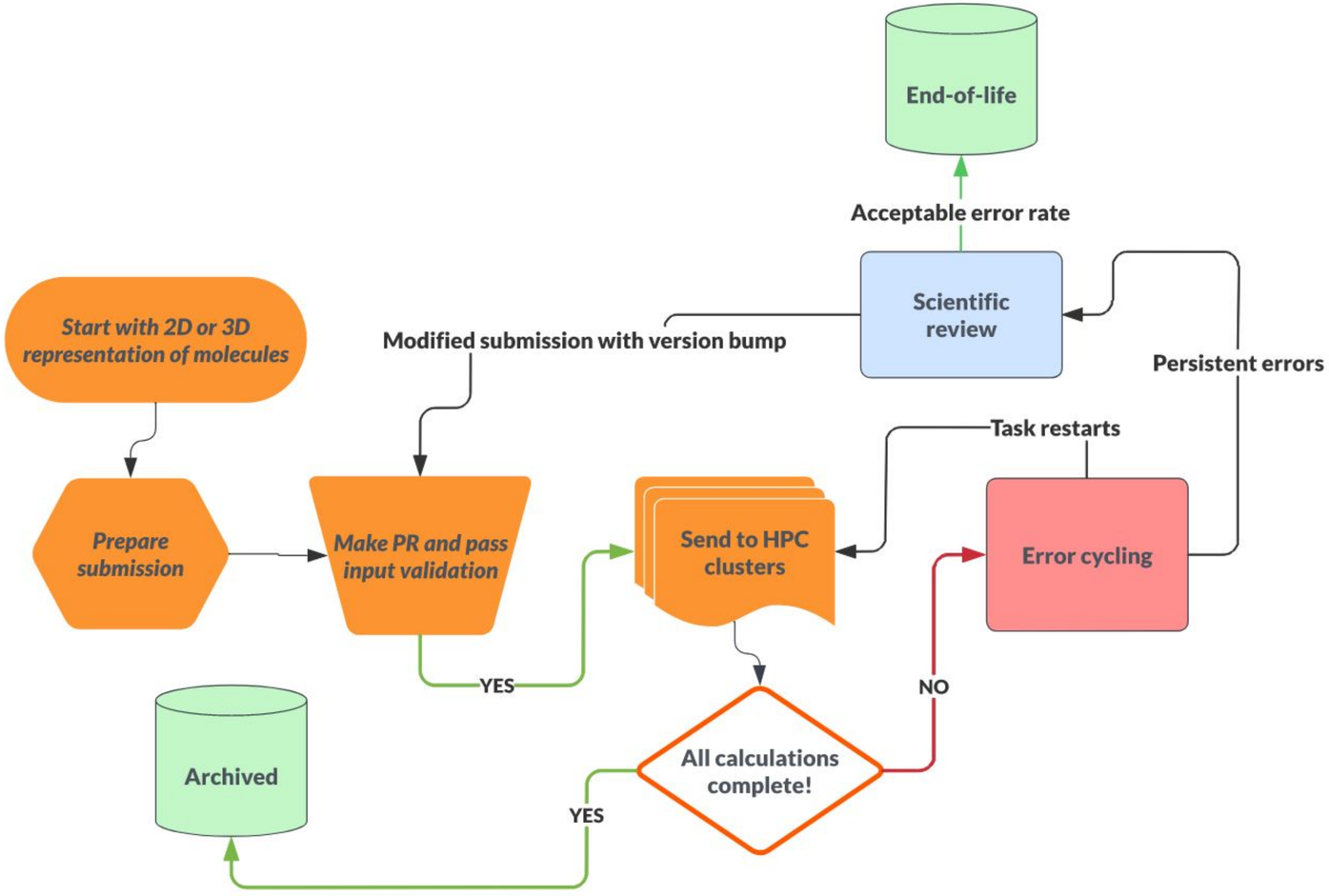
Torsiondrive

Lee-Ping Wang's Lab

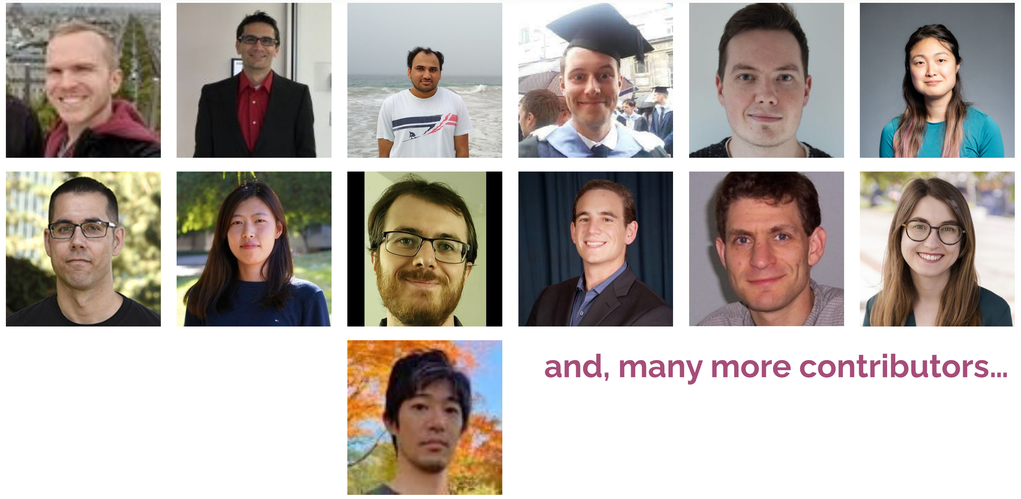
MolSSI

QC Archive
A MolSSI Project

We generate a large amount of QC data with substantial automation



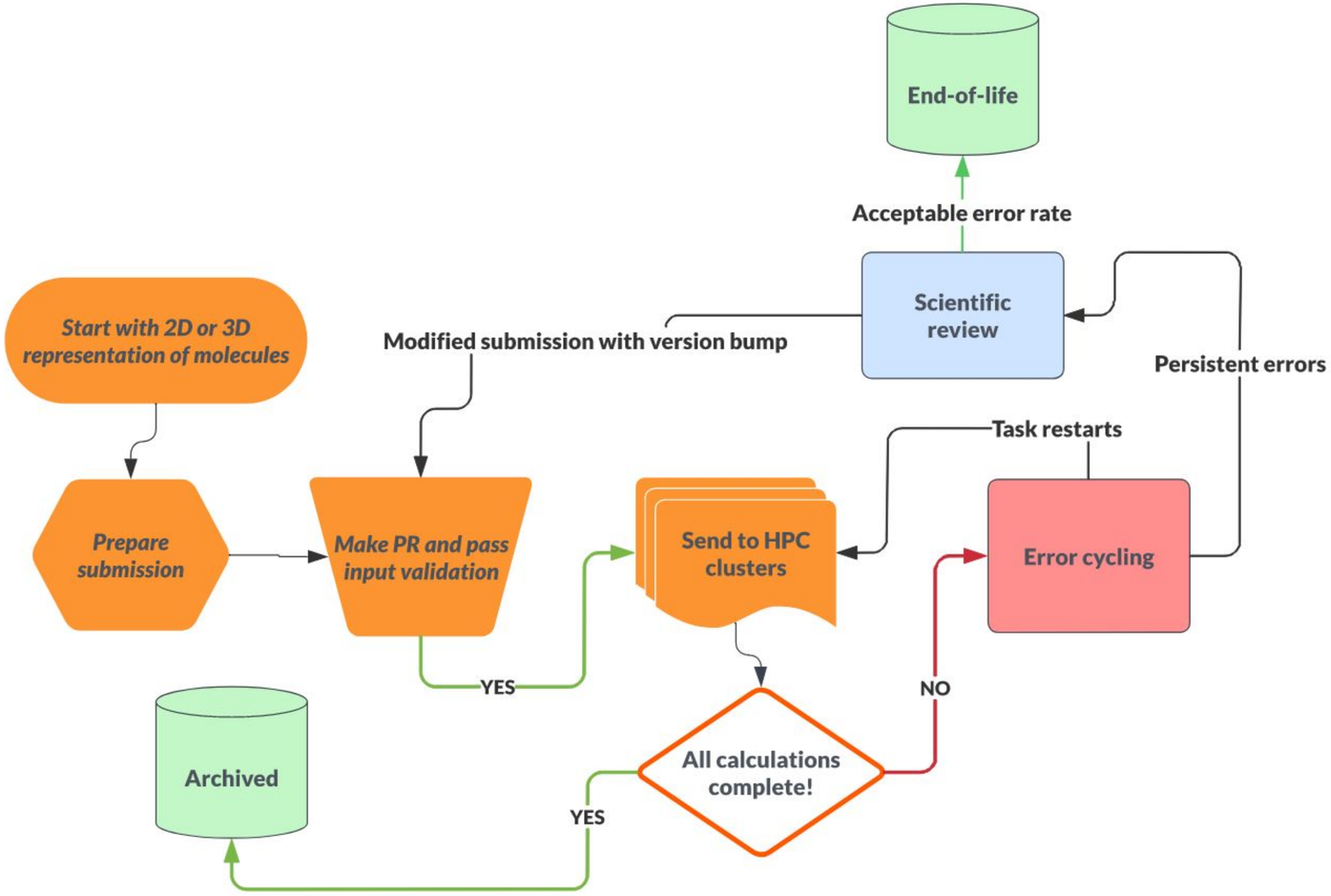
https://github.com/openforcefield/qca-dataset-submission



and, many more contributors...

Dataset Name	Purpose
OpenFF Gen 1, 2, 3	For FF fitting
OpenFF Industry Benchmark	For FF testing
OpenFF Protein	For biomolecular FF fitting
OpenFF ESP	Charge models, v-sites, polarized FFs
SPICE sets, RNA datasets	ML potentials
Many other...	

Interfacing with QCArchive via QCF & QCEngine



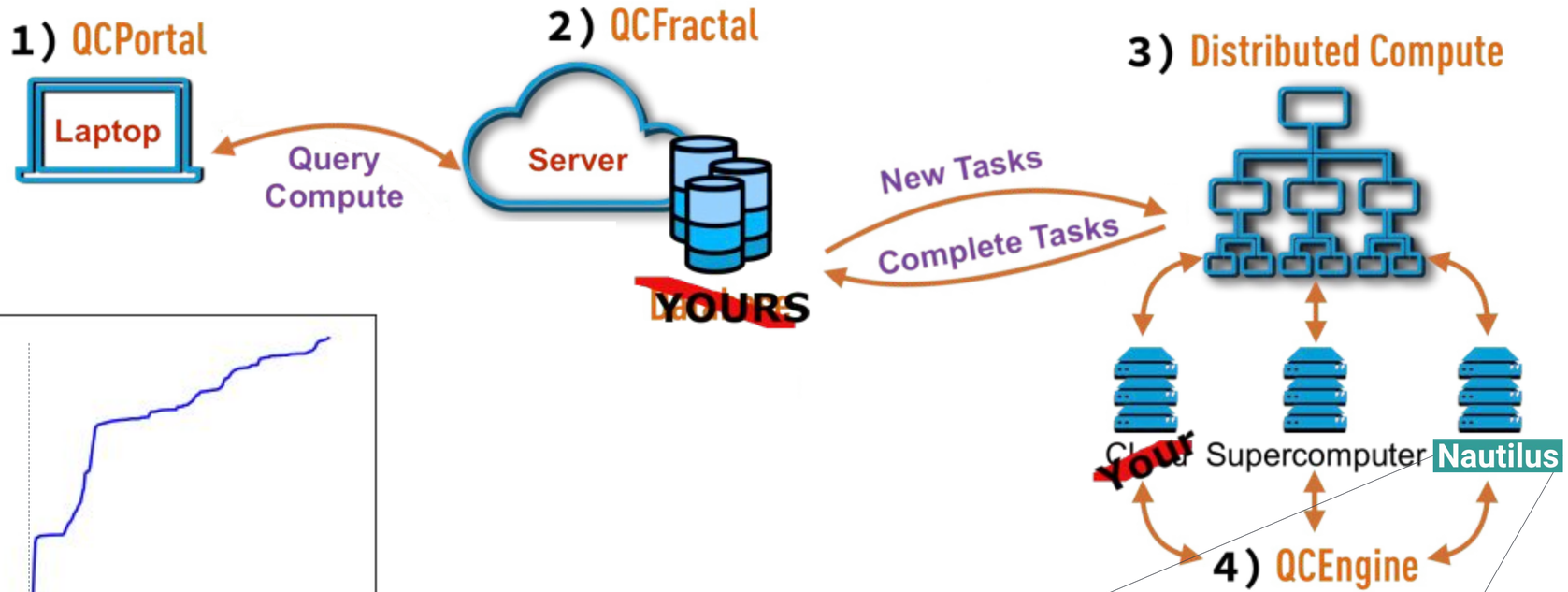
We care about

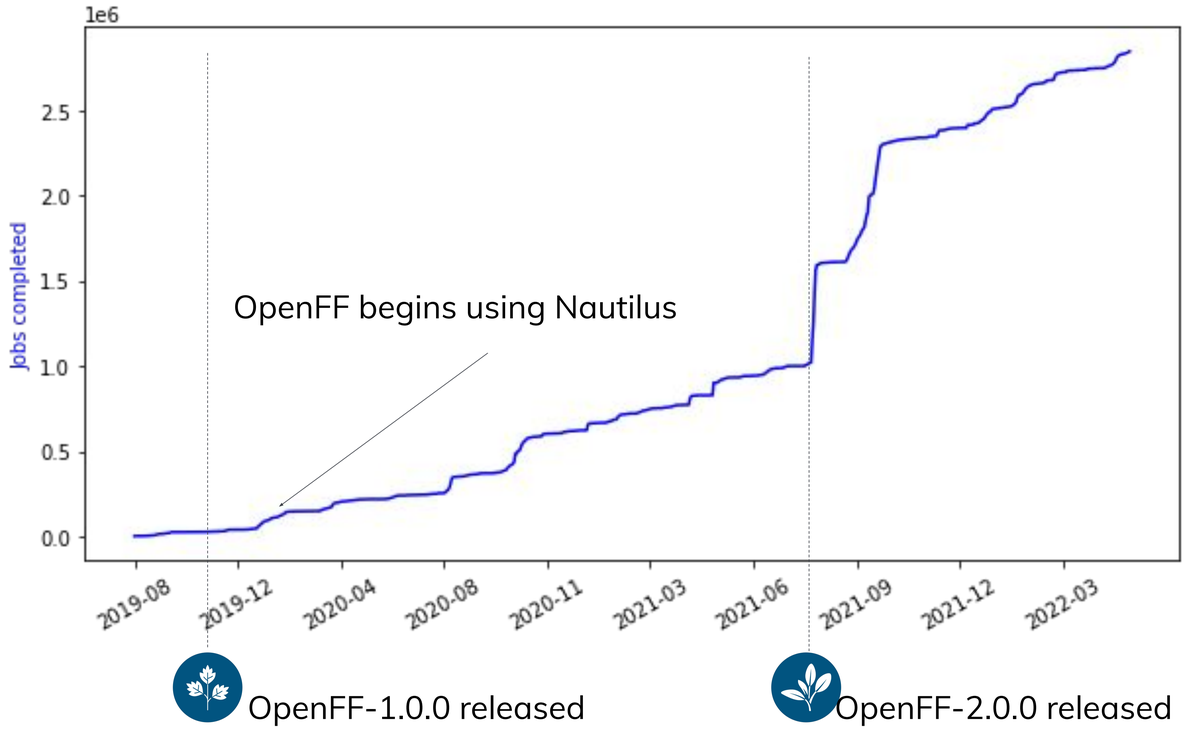
- » Origin of data
- » Adding metadata
- » Input validation
- » Error cycling
- » Reproducibility
- » Recalculate w/ another method

https://github.com/openforcefield/qca-dataset-submission

The National Research Platform is capable of running enormous quantum chemistry workloads

QC Archive
A MolSSI Project





NRP NATIONAL RESEARCH PLATFORM

You can easily access our data with a few lines of code

```
mamba create -n qcf-env -c conda-forge qcfractal openff-qcsubmit

conda activate qcf-env
```

```
> from qcportal import PortalClient

# Connecting to the server

> client = PortalClient(

                "https://api.qcarchive.molssi.org/")

# requesting dataset

> ds = client.get_dataset("optimization",

                        "OpenFF Gen 2 Opt Set 1 Roche")
```

QCArchive

QCPortal

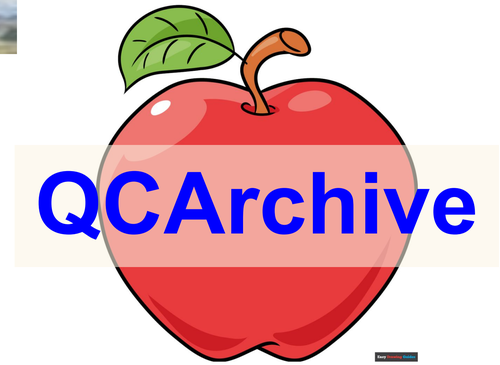
You can easily access our data with a few lines of code

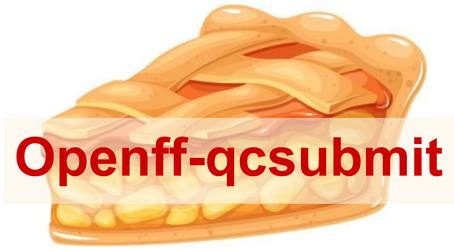
```
mamba create -n qcf-env -c conda-forge qcfractal openff-qcsubmit
conda                                   activate                                qcf-
```






```
> from qcportal import PortalClient
> client = PortalClient("https://api.qcarchive.molssi.org/")
> from openff.qcsubmit.results import OptimizationResultCollection
> optimization_result_collection =
        OptimizationResultCollection.from_server(
                client=client,
                datasets=["OpenFF Gen 2 Opt Set 1 Roche"],
                spec_name="default")
```





OpenFF Protein datasets: A crown jewel of QCA

- Rosemary - a fully consistent small molecule + protein force field is on the horizon.

Rosemary training and testing sets, generated at B3LYP-D3(BJ)/DZVP	
OpenFF-Protein-Dipeptide-2D-TorsionDrive-v2.1	**Two-dimensional TorsionDrives** on phi and psi for dipeptides of the 20 canonical amino acids and 6 alternate protomers/ tautomers.
OpenFF-Protein-Capped-1-mer-Sidechains-v1.3	**Two-dimensional TorsionDrives** on chi1 and chi2 for capped 1-mers of amino acids with a rotatable bond in the sidechain.
OpenFF-Protein-Capped-3-mer-Backbones-v1.0	**Two-dimensional TorsionDrives** on phi and psi for capped 3-mers Ace-Y-X-Y-Nme with Y = {Ala, Val}.
OpenFF-Protein-Capped-3-mer-Omega-v1.0	TorsionDrives on omega for capped 3-mers Ace-Ala-X-Ala-Nme.

QC Archive
A MolSSI Project

OpenFF Protein datasets: A crown jewel of QCA

openff-dangerbot commented on Nov 21, 2022 — Member

Lifecycle - Error Cycling Report

Dataset Name	OpenFF Protein Capped 3-mer Backbones v1.0
Dataset Type	TorsionDriveDataset
UTC Datetime	2022-11-21 19:33 UTC

All errored tasks and services will be restarted.
Errored states prior to restart reported below.

`TorsionDriveRecord` current status

specification	COMPLETE	ERROR	INCOMPLETE	RUNNING
default	22	0	0	32

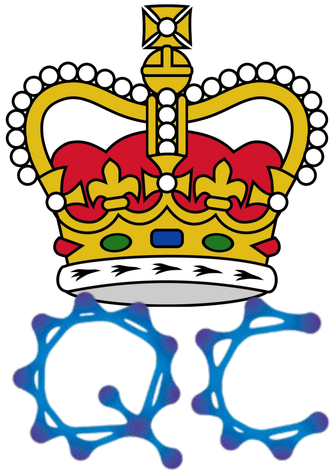
`OptimizationRecord` current status

specification	COMPLETE	ERROR	INCOMPLETE
default	311691	0	2110

QC Archive
A MolSSI Project

SPICE dataset: Another crown jewel of QCA

- https://github.com/openmm/spice-dataset
- Around 1.1 Million Single Point Calculations (energies, forces).
- Covers {H, Li, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I}.
- Accurate QM: ωB97M-D3BJ/def2-TZVPPD.



QC Archive
A MolSSI Project

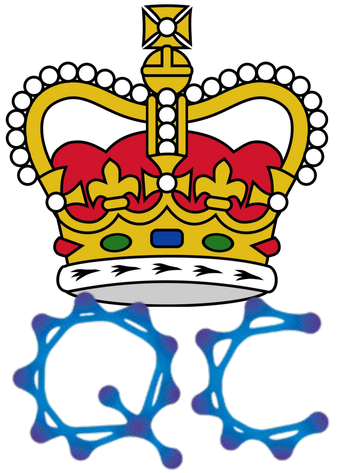
SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials

Peter Eastman ✉, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis & Thomas E. Markland

Scientific Data **10**, Article number: 11 (2023) | Cite this article

SPICE dataset: Another crown jewel of QCA

Subset	Molecules	Conformations	Atoms	Elements
Dipeptides	677	33850	26–60	H, C, N, O, S
Solvated Amino Acids	26	1300	79–96	H, C, N, O, S
DES370K Dimers	3490	345676	2–34	H, Li, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I
DES370K Monomers	374	18700	3–22	H, C, N, O, F, P, S, Cl, Br, I
PubChem	14643	731856	3–50	H, C, N, O, F, P, S, Cl, Br, I
Ion Pairs	28	1426	2	Li, F, Na, Cl, K, Br, I
Total	19238	1132808	2–96	H, Li, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I



QC Archive

A MolSSI Project

SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials

Peter Eastman ✉, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis & Thomas E. Markland

Scientific Data **10**, Article number: 11 (2023) | Cite this article

QCEngine in OpenFF-Bespokefit

- **Bespokefit:** Custom FF builder.
- **On-the-fly parametrization** of the additional torsion terms need QM torsion scans
  - Provided using Torsiondrive procedure in QCEngine
- QCEngine's interface to various QC programs including faster methods like GFN2-XTB is a big advantage.

Bespokefit

Acknowledgement

- **David Mobley (Advisor).**
- **Openforcefield PIs:** Lee-Ping Wang, Michael Shirts, Michael Gilson, Daniel Cole, John Chodera.
- **Openforcefield team (current & past):** Simon Boothroyd, Chapin Cavender, Karmen Condic-Jurkic, Lorenzo D'Amore, David Dotson, Trevor Gokey, David Hahn, Josh Horton, Tobias Hüfner, Hyesu Jang, Victoria Lim, Jessica Maat, Lexie McIsaac, Owen Madin, Josh Mitchell, Benjamin Pritchard (MolSSI), Jaime Rodriguez-Guerra, Jeffrey Setiadi, Chaya Stern, Ken Takaba, Matt Thompson, Jeffrey Wagner, Lily Wang, Willa Wang, Yuanqing Wang, Brent Westbrook, and others.
- Collaborators: Benjamin Pritchard, Christopher Bayly, William Swope, Peter Eastman.
- OpenFF industry and academic collaborators.
- NIH and NSF for funding work that helped pave the way to this effort.
- Consortium and NIH for current funding, plus MolSSI and others for fellowship funding.

open
forcefield

docs.openforcefield.org

Website
https://openforcefield.org/

GitHub
https://github.com/openforcefield/

Zenodo
https://zenodo.org/communities/openforcefield/

Twitter
https://twitter.com/openforcefield

YouTube
https://www.youtube.com/channel/UCh0qJSUm_sYr7nuTzhW806g/videos

LinkedIn
https://www.linkedin.com/company/openforcefield

Email
info@openforcefield.org