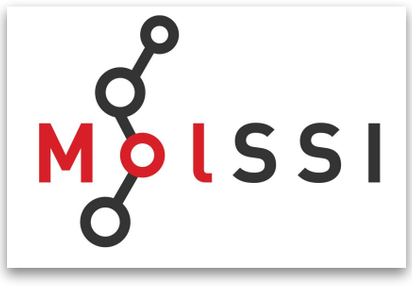
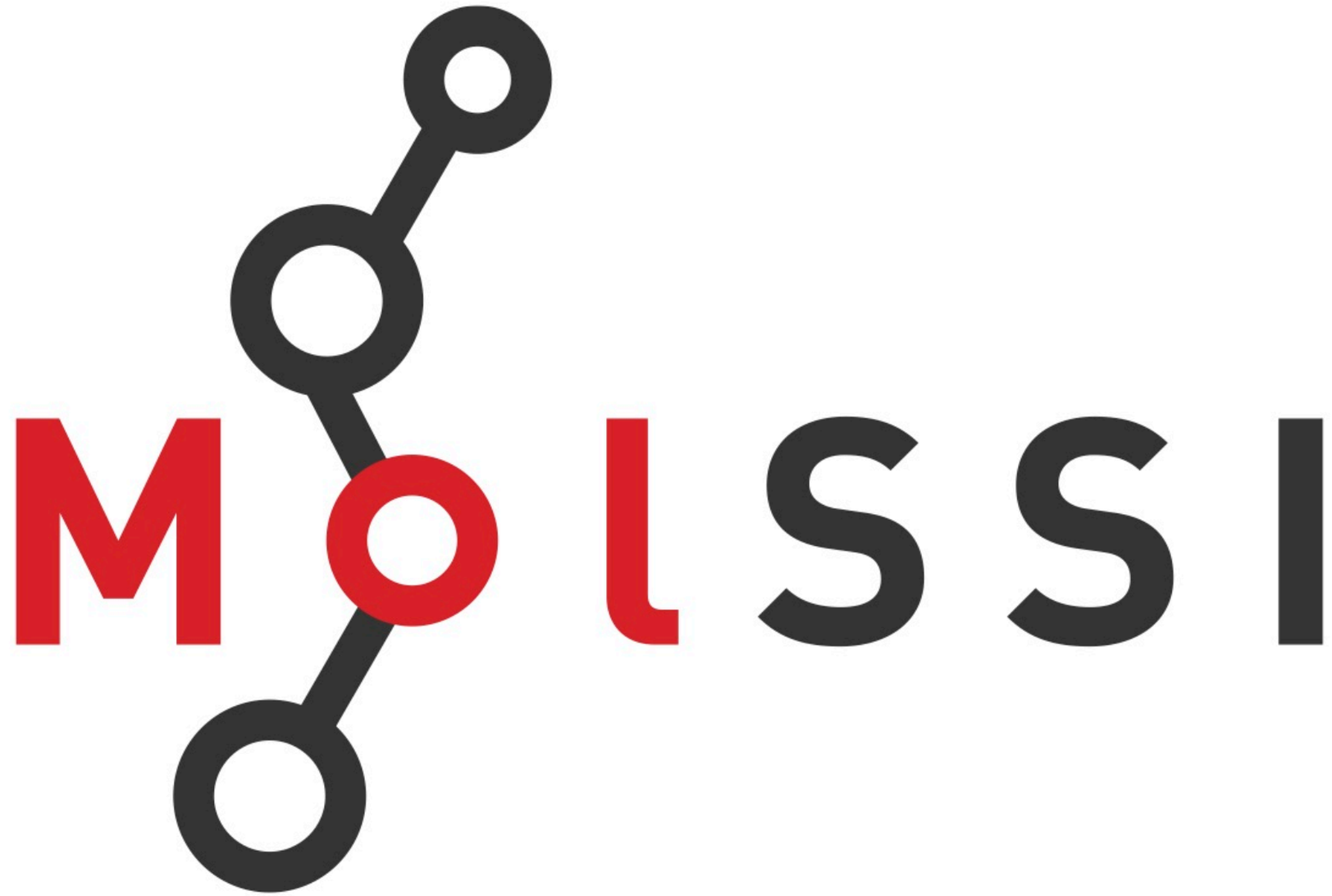
The MolSSI, standards and interoperability. A fair beginning.

Paul Saxe, psaxe@vt.edu

8 July 2019

Berlin





The Molecular Sciences Software Institute

... a nexus for science, education, and cooperation for the global computational molecular sciences community.

What is the MolSSI?

- Launched August 1st, 2016, funded by the National Science Foundation.
- Collaborative effort by Virginia Tech (TDC), Rice U. (C. Clementi), Stony Brook U. (R. Harrison), U.C. Berkeley (T. Head-Gordon), Rutgers U. (S. Jha), U. Southern California (A. Krylov), and Iowa State U (T. Windus).
- Part of the NSF's commitment to the White House's National Strategic Computing Initiative (NSCI).
- Total budget of $19.42M for five years, potentially renewable to ten years.
- Joint support from numerous NSF divisions: Advanced Cyberinfrastructure (ACI), Chemistry (CHE), and Division of Materials Research (DMR)
- Designed to **serve** and **enhance** the software development efforts of the broad field of computational molecular science (CMS) – a broad domain that includes quantum chemistry, computational materials science, and biomolecular simulation.

MolSSI

Who is the MolSSI?

- Software Scientists: A team software engineering experts, drawn both from newly minted Ph.D.s and established researchers in molecular sciences, computer science, and applied mathematics.
- Software Fellows: A cohort of ~24 graduate students and postdocs supported simultaneously and selected from research groups across the U.S. by the MolSSI's Science and Software Advisory Board.
- Board of Directors: Seven co-PIs who oversee the MolSSI's activities and provide guidance and expertise.
- Science and Software Advisory Board: Representatives from academia, industry, national laboratories, and international facilities who advise the MolSSI on the most important software priorities for the community.
- Community-Code Partners: Approximately 40 computational molecular science software packages whose developers work with the MolSSI on standards, training, and infrastructure.

MolSSI

MolSSI Highlights So Far

- Hired **twelve Software Scientists** – the full contingent as originally planned, but we are considering hiring an additional team member.
- **17 software workshops with more than 500 participants** so far; at least another eight to be held in 2019.
- New software components currently under development including an **open QM database (QCArchive)**, a **general QM/MM driver**, a **new basis set exchange**, a **reference integral** implementation, and more.
- **Community-driven working groups** established in forcefield interoperability, quantum chemistry data exchange, and tensor algebra interfaces.
- Held one software summer school (second one coming in July at TACC), five "Best Practices" workshops, three Software Fellowship bootcamps, and two undergraduate programming schools; many more educational workshops and schools coming in 2019.
- **24 Software Fellows currently supported**, plus **11 new Fellows** starting July 1, for a **total of 50 Fellows** funded overall.

Watch molssi.org for the latest information!

MolSSI

MolSSI Goal #1

To Provide Software Expertise and Infrastructure...

- MolSSI works with CMS research groups nationwide and internationally to design, develop, test, deploy, and maintain key code infrastructure and frameworks for the entire community.
- MolSSI interacts with partners in industry, NSF supercomputing centers, national laboratories, and international facilities to identify and act on emerging hardware trends, access leading-edge computing architectures, further educational goals, set software priorities, and identify future workforce career paths.

MolSSI

MolSSI Goal #2

To Provide Education and Training...

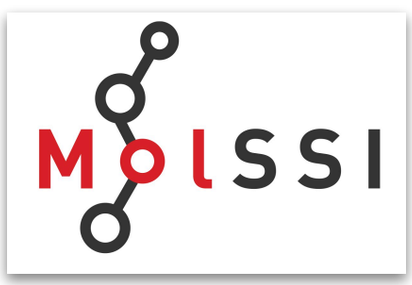
- MolSSI serves as an education and outreach nexus for the worldwide CMS community.
- MolSSI organizes summer schools, targeted workshops, high-school and undergraduate training programs, and on-line resources and classes to provide current and future CMS students with a modern and complete set of programming skills.
- MolSSI reaches beyond the traditional student cohort to computer scientists and mathematicians seeking interdisciplinary applications.

MolSSI

MolSSI Goal #3

To Provide Community Engagement and Leadership...

- MolSSI will enable the CMS community to establish its own standards for interoperability, best practices, and curation tools.
- Through a **"grass roots"** approach, MolSSI engages the community broadly using interoperability workshops and focus groups to catalyze the consensus needed for standardization of data structures, APIs, and frameworks for the entire CMS software ecosystem.



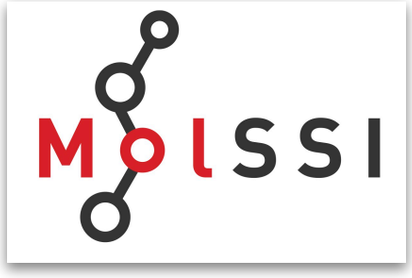
MolSSI Community Code Partners So Far…

- ACESIII
- ADF
- Amber
- APBS
- BOSS
- CFOUR
- CHARMM
- Columbus
- Dalton
- Tiger-CI
- Dirac
- DL_POLY
- ELSI
- FHI-aims
- GAMESS
- Gaussian
- Gromacs
- LAMMPS
- Molcas
- Turbomole
- Molpro
- MPQC
- MRChem
- NAMD
- NWChem
- NWChemEx
- ONETEP
- OpenMM
- Orca
- VASP
- PARSEC
- PCMSolver
- PLUMED
- PQS
- PSI4
- Q-Chem
- QBox
- QMworks
- Quantum ESPRESSO
- Schrödinger

We encourage all community codes in the computational molecular sciences to work with us!

Computed Data

- Obviously trivial!
  - We know the structure down to the atom
  - We know the methods and results
- What could go wrong?



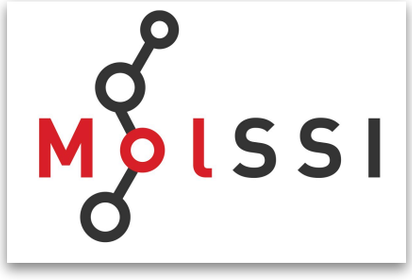
Computed Data

- Experiment has real problems:
  - What is the sample. Exactly?
  - What was actually measured?

(This is this theorist's simplified view! The grass on the other side of the fence is always greener.)

- Computation doesn't have these problems
  - Yet we have real problems describing our data

Are we lazy?
Is it not important to us?
or
Are our "instruments" too complex?



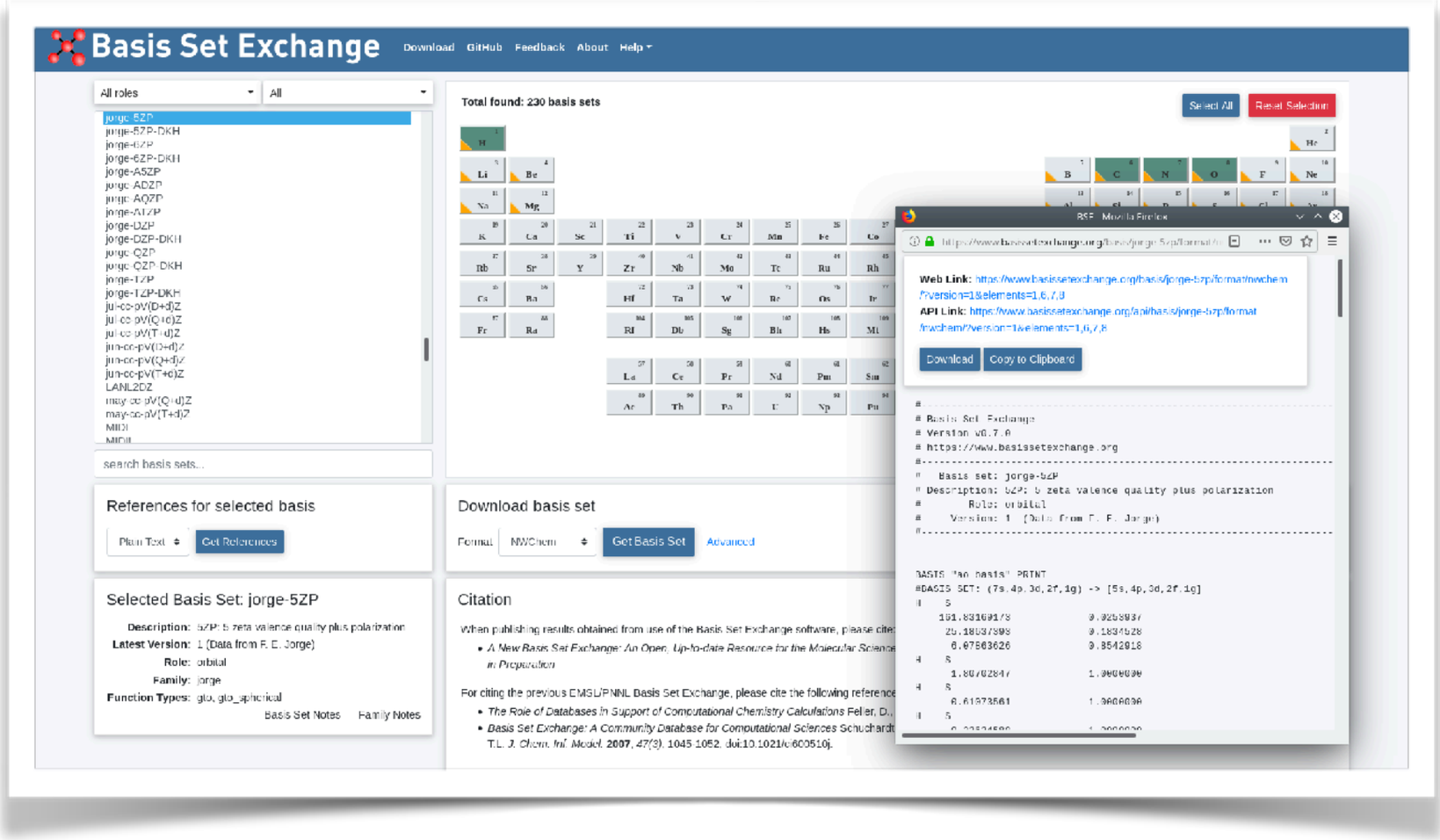
Projects of Interest

- Database
  - Basis Set Exchange
- Standards
  - QCSchema
  - MMSchema
- Environments
  - QCArchive
  - Simulation Environment for Atomistic and Molecular Modeling (SEAMM)
- Interface Libraries
  - MolSSI Driver Interface (MDI)

A New and Improved Basis Set Exchange (BSE)

https://www.basissetexchange.org

- Database of quantum chemistry basis sets (Gaussians)
- Collaboration with original BSE developers (PNNL/EMSL)
- Highly-utilized by both developers and end users
- Improved provenance and reproducibility of calculations via unique identifiers
- Improved curation/reliability of the raw data
- New UX features

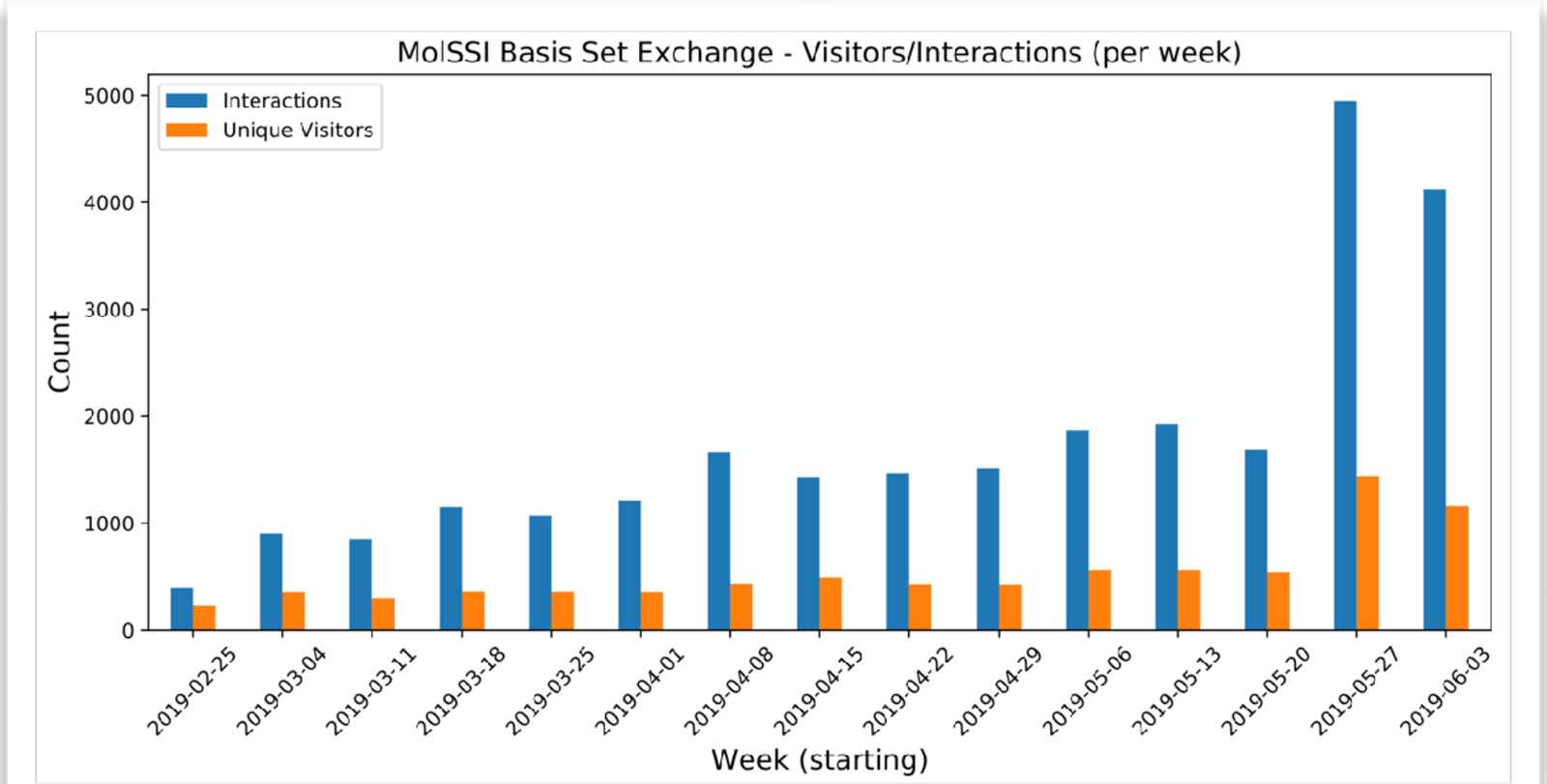


Software Scientists: Ben Pritchard and Doaa Altarawy

A New and Improved Basis Set Exchange (BSE)

https://www.basissetexchange.org

- Usage is increasing since March 1st launch
- Old site retired on May 31st
- Many users return, performing multiple actions
- As of May 2019
  - More than 300 basis sets
  - 2211 Unique Users
  - 8343 Actions
  - 37% Visitors are from US

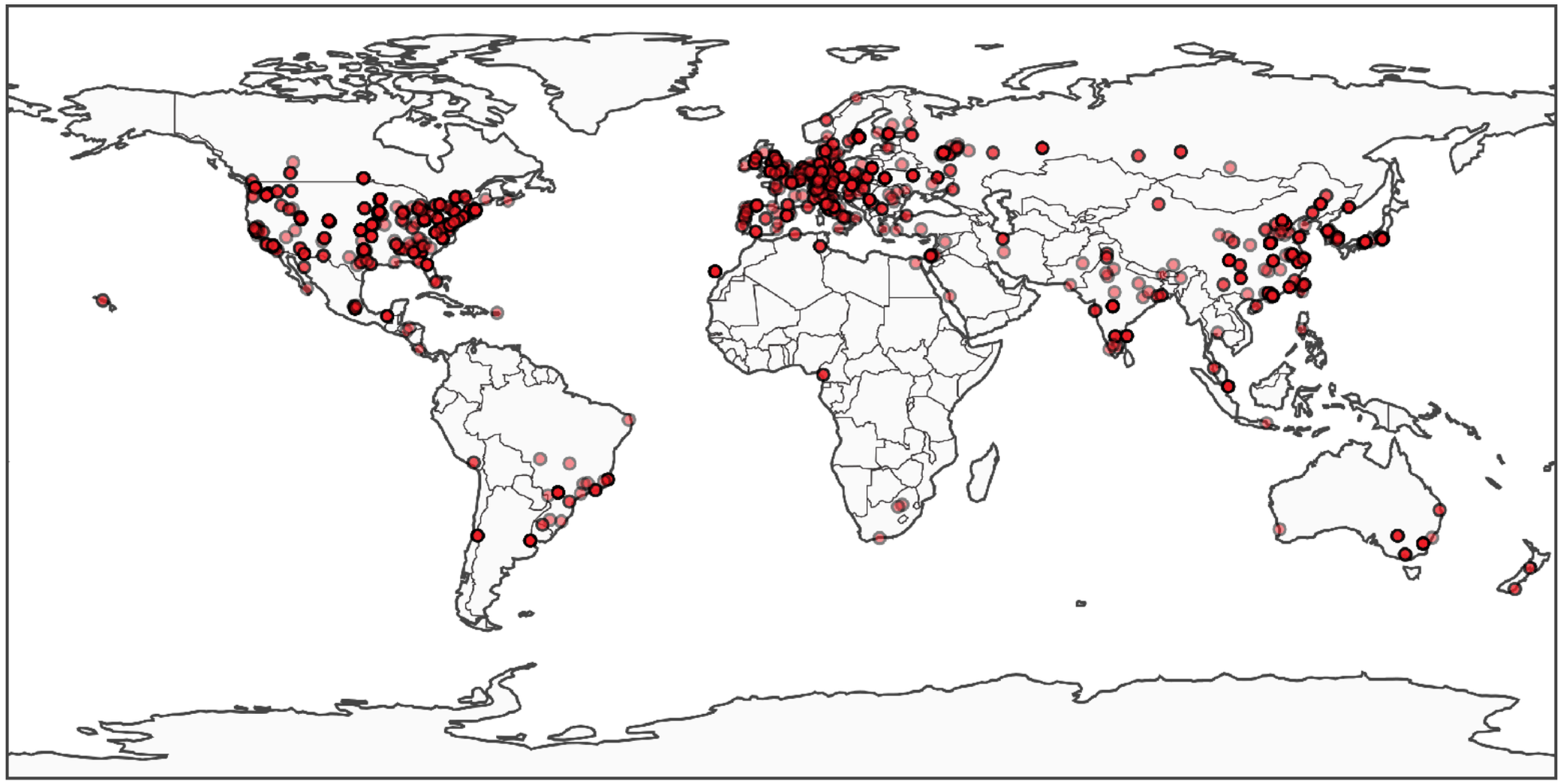


MolSSI

Software Scientists: Ben Pritchard and Doaa Altarawy

A New and Improved Basis Set Exchange (BSE)

https://www.basissetexchange.org



MolSSI

Software Scientists: Ben Pritchard and Doaa Altarawy

https://github.com/MolSSI/QCSchema

QCSchema

- Communication channel between all pieces 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
- Input
- Output
- Optimization Trajectory

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

```
{
    …Input
    "provenance": {
        "creator": "My QM Program",
        "version": “1.1rc1",
         …
    },
    "properties": {
        “scf_n_iterations”: 2.0,
        "scf_total_energy”: -5.4331918814433 23,
        "nuclear_repulsion_energy”: 2.11670883436,
        “one_electron_energy”: -11.67399006298957,
    …
    },
    "error": "",
    "success": true,
    "raw_output": "Output storing was not requested."
}
```

MMSchema

- Just starting (last week!)
- Define molecular mechanics/dynamics data
- Work with and gain acceptance of the MD community

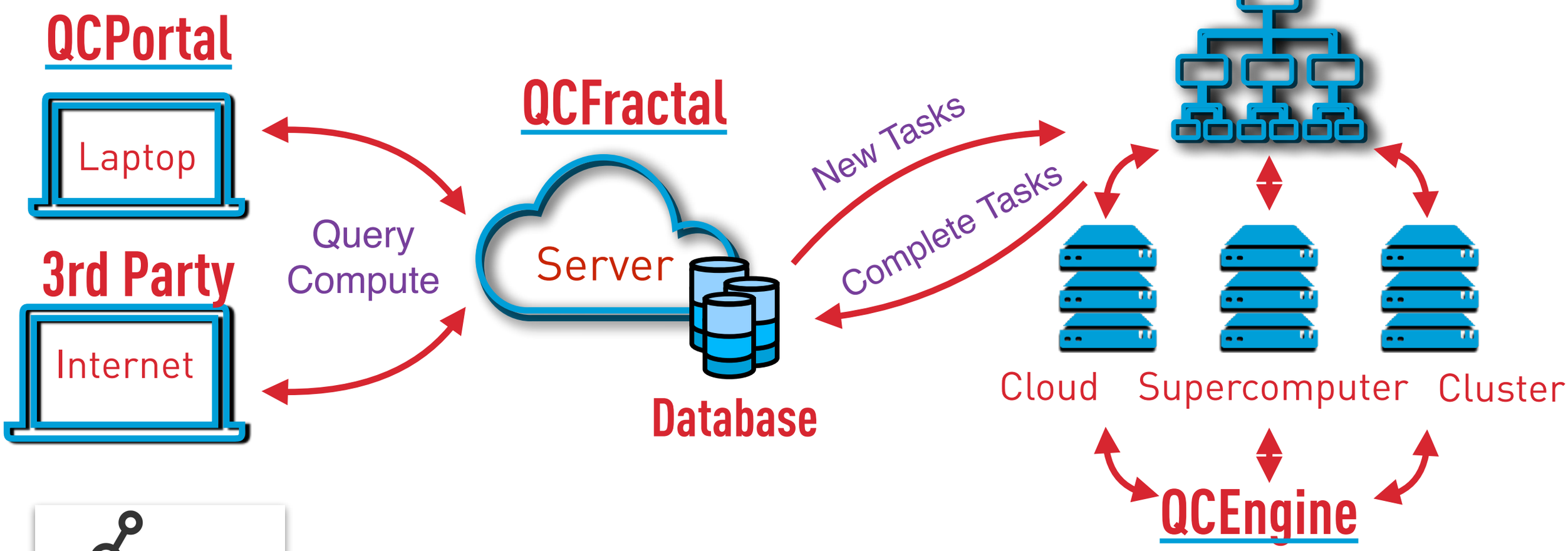
MolSSI

QCArchive Overview

QC Archive
A MolSSI Project

Goals:

- High-throughput quantum chemistry on multi-physical site compute
- Laptop to campaign-scale compute orchestration
- Procedures run with a variety of different programs
- Common abstraction and organization layers
- Share and collaborate structured data



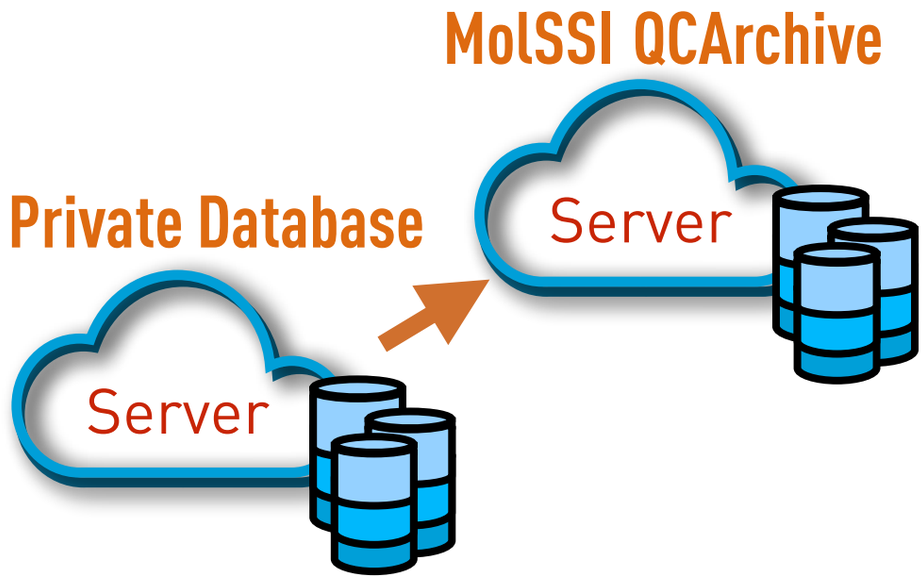
MolSSI

Software Scientists: Daniel Smith, Levi Naden, and Doaa Altarawy

MolSSI and Self-Hosted Databases

QC Archive
A MolSSI Project

- 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
- ~2M current results
- ~40 community datasets
- ~200 unique IPs per week
- Can host ~500M results with current hardware, looking to expand!

MolSSI

https://qcarchive.molssi.org

Reproducible Procedures and Data

QC Archive
A MolSSI Project

Procedures

- Procedures = small reproducible series of computations
- Exact input of pipeline and version data available
- Geometry optimizations, torsion drives, finite difference computations, spectral computations sets, 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'}
```

```
optimization.energies
```

```
[-287.88654576113106,
 -287.885448441852
```

```
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

Data Organization

- Many single computations or procedure grouped together known as Collections
- Reproducible and tweakable
- Exportable sets of data in canonical forms
- Working with IJQC to formalize and distribute datasets
- Data organization for ML, methodology assessment, Force Field Optimizations, etc

MolSSI

https://qcarchive.molssi.org

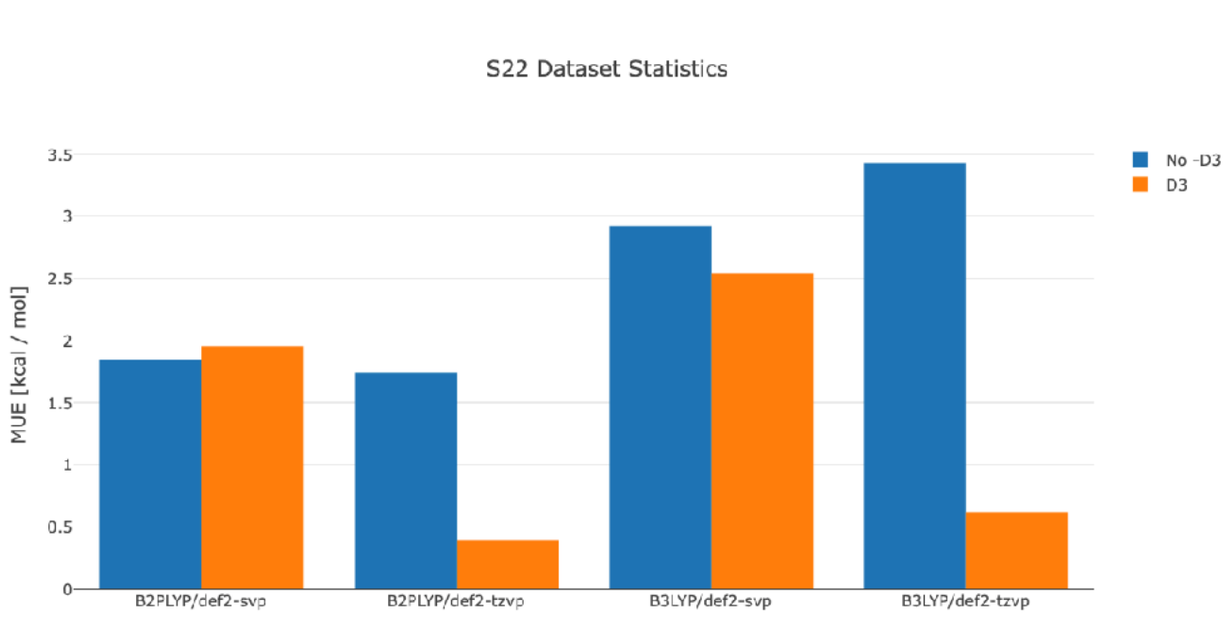
Interactive Sessions and Gateway

QC Archive
A MolSSI Project

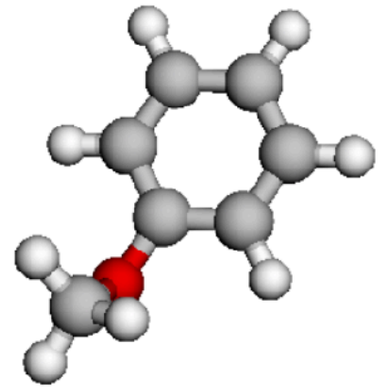
Jupyter-Notebook Integration

- 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")
```




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



Science Gateway

- Working with SGCI
- Web-based statistics and visualization
- Aiming at non-CMS researchers and undergraduate educational initiatives
- Data-driven initiates:
  - What is the best method for X
  - How long will X take?

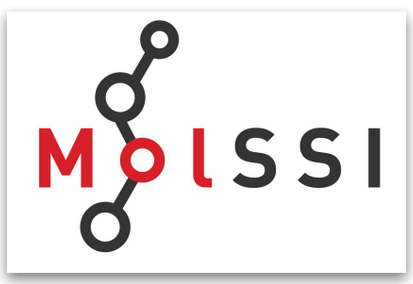
MolSSI

https://qcarchive.molssi.org

Reproducibility and Replicability

Reproducibility is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis. This definition is synonymous with "computational reproducibility," and the terms are used interchangeably in this report.

Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.



National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25303

Replicability

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Round Robin Study: Molecular Simulation of Thermodynamic Properties from Models with Internal Degrees of Freedom

Michael Schappals,[†] Andreas Mecklenfeld,[‡] Leif Kröger,[§] Vitalie Botan,[§] Andreas Köster,[∥] Simon Stephan,[†] Edder J. García,[†] Gabor Rutkai,[∥] Gabriele Raabe,[‡] Peter Klein,[⊥] Kai Leonhard,[§] Colin W. Glass,[#] Johannes Lenhard,[▽] Jadran Vrabec,[∥] and Hans Hasse*[,†]

J. Chem. Theory Comput., **2017**, *13* (9), pp 4270–4280
DOI: 10.1021/acs.jctc.7b00489

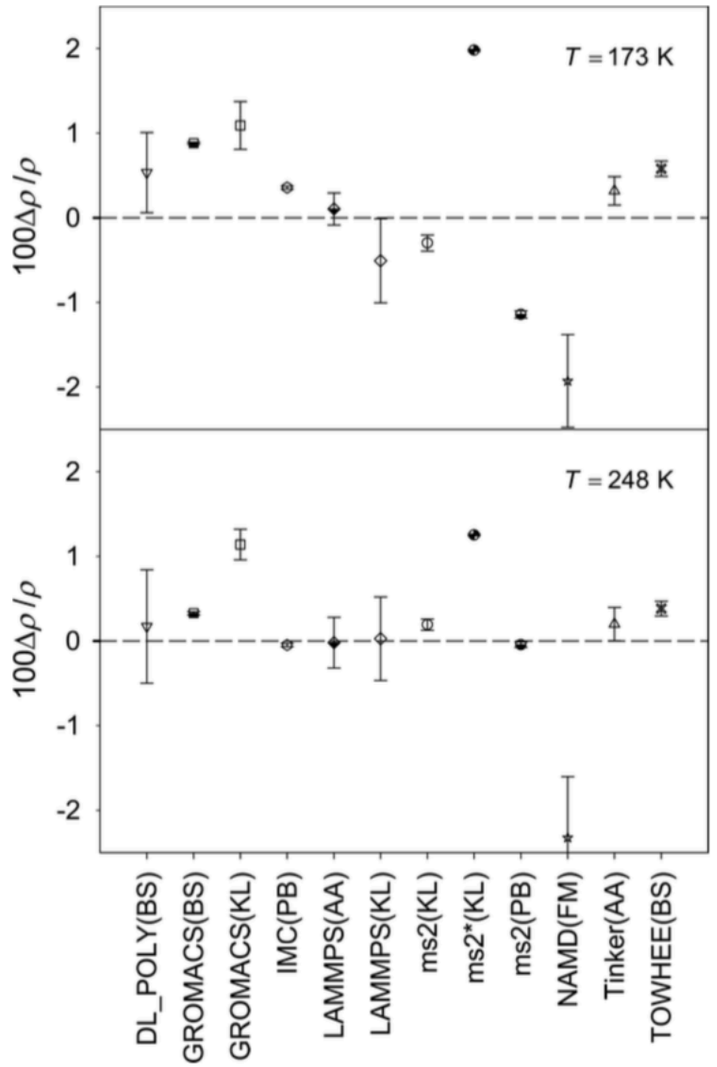


Figure 10. Statistical uncertainty of the data obtained for the density of *n*-butane at 41 MPa and 173 (top) and 248 K (bottom) from the OPLSAMBER force field. Symbols: mean values with error bars determined from block averages of the production phase. Dashed line: arithmetic mean of all results.

MolSSI

The results reveal the challenges of carrying out molecular simulations. Several iterations were needed to eliminate gross errors. For most simulation tasks, the remaining deviations between the results of the different groups are acceptable from a practical standpoint, but they are often outside of the statistical errors of the individual simulation data. However, there are also cases where the deviations are unacceptable.

Reasonably expert users, calculating the density of liquid alkanes (C_2-C_4), given reference forcefield parameters…

…had gross errors at the beginning?

…could not get the same answer within error bars?

…and have differences of ±0.4%? And ±0.2% with the **same** code!

User Error

2.6. User Error. The input data of the simulation are classified here as follows: (1) specifying the model, (2) specifying the scenario, (3) controlling the algorithms (physical, numerical), (4) controlling the compilation, (5) controlling the actual simulation run, and (6) controlling the evaluation of the simulation results. Any input data are prone to user error.
The importance of user errors has recently been discussed by Wong-ekkabut and Karttunen in a paper entitled: The good, the bad, and the user in soft matter simulations, in which they give many nontrivial examples for user errors and conclude that the user is the "most significant error source" and that "one does not become a theorist by buying chalk, experimentalist by buying a microscope, or a computational scientist by downloading software". On the basis of the experience from the present work, we fully agree and add only that user errors are not a privilege of rookies but regularly happen to experienced users as many examples show.

We can do better.

We must do better!

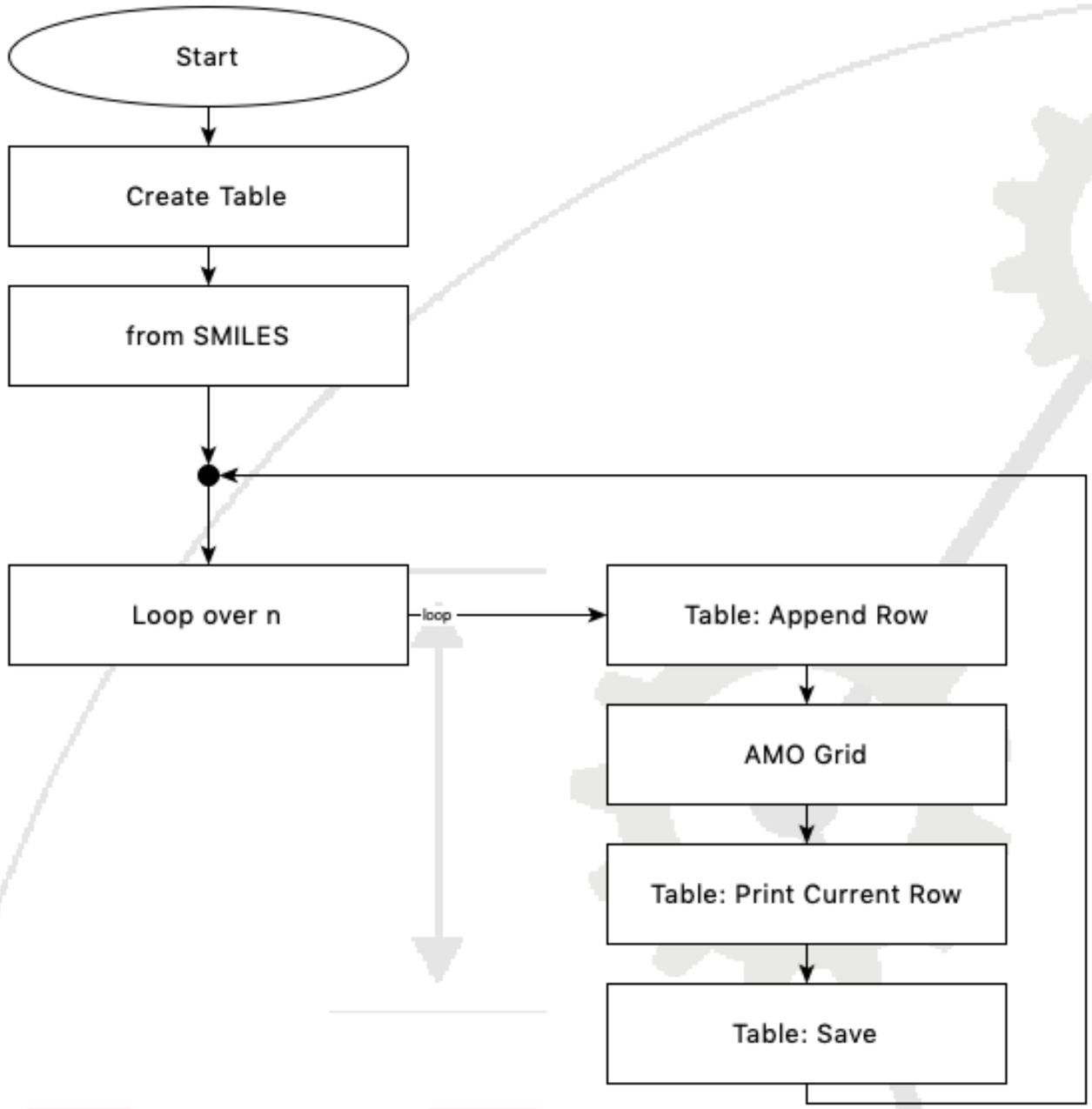
SEAMM

Simulation Environment for Atomistic and Molecular Modeling

- Open environment for molecular and materials modeling and simulation
- Built around reproducible flowcharts
- Provides low-level services for creating and editing flowcharts, handling data, computing, etc.
- All user functionality provide by community-developed plugins

MolSSI

Flowcharts

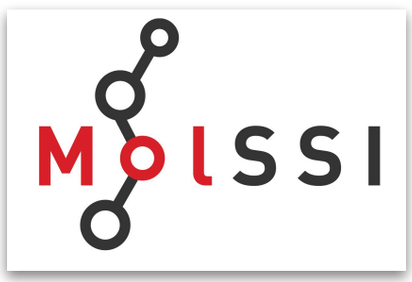

Start
Create Table
from SMILES
Loop over n
loop
Table: Append Row
AMO Grid
Table: Print Current Row
Table: Save

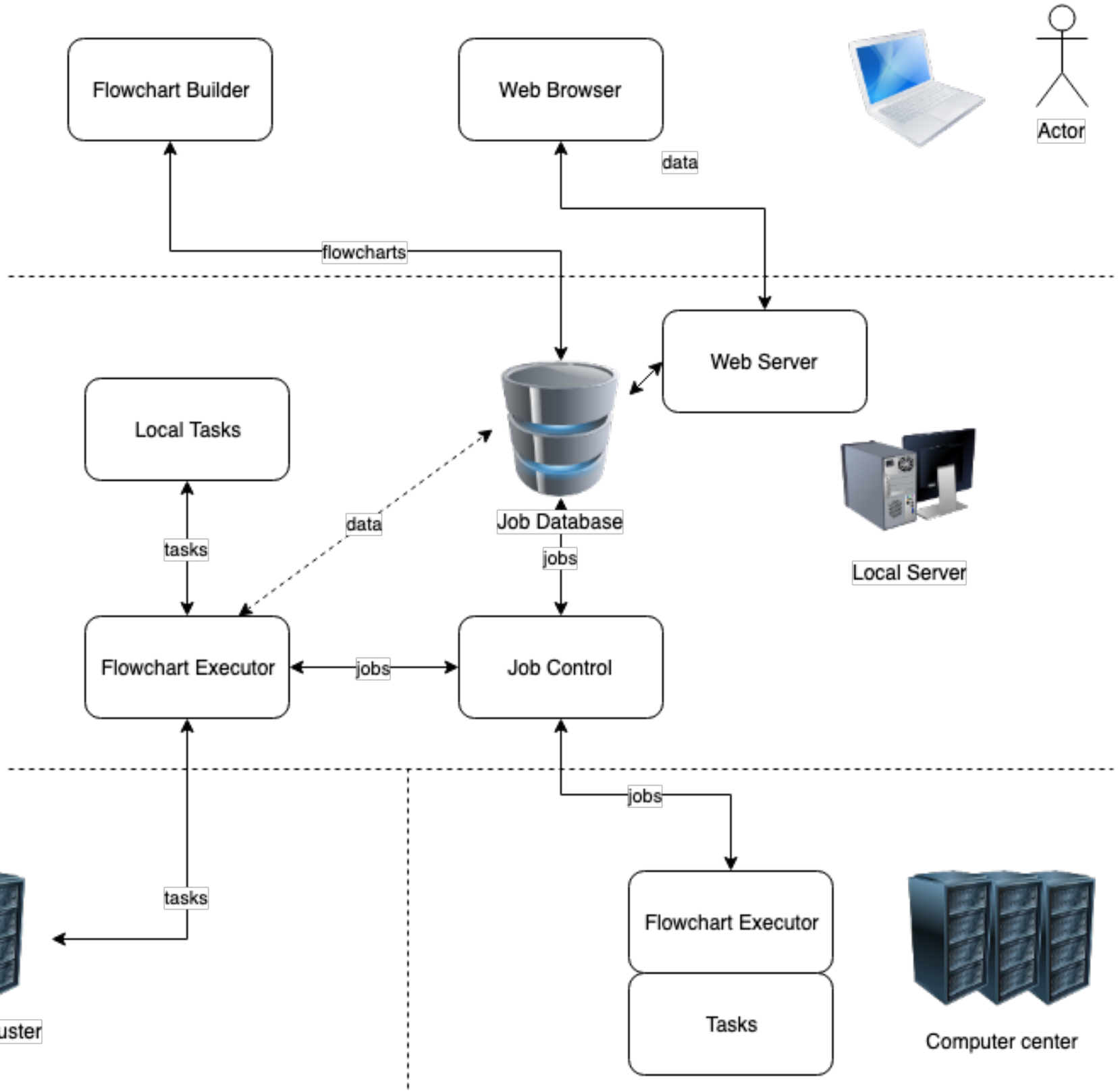
Goals

- Ease-of-use
- Reproducibility

} GUI & Productivity

- A place to “put” codes, particularly smaller helper codes
- Application domain agnostic
- Widely used by subfields and users




Flowchart Builder
Web Browser
Actor
data
flowcharts
Web Server
Local Tasks
Job Database
data
tasks
jobs
Local Server
Flowchart Executor
jobs
Job Control
jobs
tasks
Flowchart Executor
Tasks
Local Cluster
Computer center

QCArchive & SEAMM

- QCArchive is database centric
- SEAMM is workflow centric

The MolSSI Driver Project

What do we need in order to simplify and standardize
the process of interoperating codes on-the-fly?

Driver–Engine Paradigm:

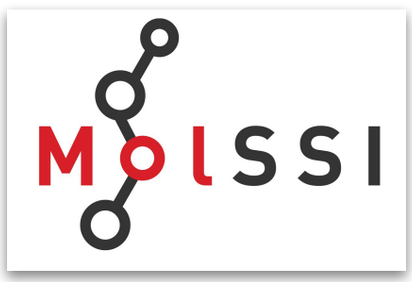
Driver codes orchestrate the high-level program flow of one or more engine codes by sending commands through the MDI.

The MDI Standard:

An API-like definition of a set of commands that can be sent from a driver to an engine, and that cause the engine to respond in a clearly-defined way.

The MDI Library:

A library for inter-code communication that enables easy compliance with the MDI Standard.



Software Scientist: Taylor Barnes

Summary

- The MolSSI is a large project
  - 12 or 13 scientists
  - ~24 Software Fellows at any one time
  - Large educational component
  - Software (mostly) projects spanning many areas
- Five projects related to FAIR standards
  - 1 database, 2 schema standards and 2 environments
  - Still learning!
  - Not yet completely integrated.
- I feel that capturing our workflow — our experiment — is key
  - There is considerable tension and discussion with the MolSSI on this
- Computational materials is underrepresented 😞

Acknowledgments

- The many dozens of members of the CMS community who helped to develop the vision for the Institute over the last five years;
- NSF ACI-1547580.

Watch molssi.org for the latest information!