The Molecular Sciences Software Institute

...a nexus for science, education, and cooperation for the global computational molecular sciences community.
Computational Molecular Sciences (CMS)

- A broad field including quantum chemistry, computational materials science, and biomolecular simulation
- The history of CMS reaches back decades to the genesis of computational science.
- CMS is now a “full partner with experiment”.
- For an impressive array of chemical, biochemical, and materials challenges, our community has developed simulations and models that directly impact:
  - Development of new chiral drugs;
  - Elucidation of the functionalities of biological macromolecules;
  - Development of more advanced materials for solar-energy storage, technology for CO₂ sequestration, etc.
CMS Codes Are Developed and Used Globally

Gaussian

DELSE

GAMESS

LAMMPS

Dalton

ORCA

\langle cc | cc \rangle

Q-CHEM

Psi4

TURBOMOLE

MOLPRO

MOLSS1
CMS Codes Are Developed and Used Globally

These codes represent decades of development by thousands of programmers, and are used by hundreds of thousands of scientists worldwide.
Code Complexity and Historical Legacy

- CMS programs contain millions of lines of hand-written code and require hundreds of programmers to develop and maintain.

- Incredible language diversity: F77, F90, F95, HPF, C, C++, C++11, C++14, C++17, Python, perl, Javascript, etc.

- Incredible algorithmic diversity: structured and unstructured grids, dense and sparse linear algebra, graph traversal, fast Fourier transforms, MapReduce, and more.

- The packages have evolved in an ad hoc manner over decades because of the intricacy of the scientific problems they are designed to solve.
Rapidly Evolving Computing Hardware

- **Multi- and many-core architectures** are the norm, but many CMS codes are developed with limited view to parallel task management.

- Reduced-power solutions will also require improved **error recovery and checkpointing** at the software level – capabilities absent in nearly all CMS codes.

- Anticipated architectural innovations will yield **even greater hardware complexity** – more advanced accelerators, specialized computing cores, reconfigurable logic...

- Many CMS codes (especially for quantum chemistry) are limited to shared-memory paradigms and cannot yet take advantage of GPUs or **large-scale distributed-memory systems**.
Inertia in the Scientific Education Culture

- Undergraduate programs in chemistry and physics typically require no training in software development or programming.
- Graduate programs in these areas require minimal coursework between the bachelor and Ph.D.
- Most computer science students lack the underlying knowledge of the scientific domains to help develop creative software solutions.
- Due credit for software development is elusive due to a culture that judges productivity based on citations of peer-reviewed papers.
- Thus, a “just get the physics working” approach pervades much of CMS software development.
What is the MolSSI?

- Launched August 1st, 2016, funded by the National Science Foundation.
- 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: Office of Advanced Cyberinfrastructure (OAC), Chemistry (CHE), and Division of Materials Research (DMR)
- Designed to serve and enhance the software development efforts of the field of computational molecular science – a broad domain that includes quantum chemistry, computational materials science, and biomolecular simulation.
Who is the MolSSI?

- **Software Scientists**: A team of 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 ~35 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**: Eight 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 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 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 online 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 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 Software Scientists

• A team of ~12 software engineering experts, drawn both from newly minted Ph.D.s and established researchers in molecular sciences, computer science, and applied mathematics.

• Dedicated to multiple responsibilities:
  • Developing software infrastructure and frameworks;
  • Interacting with CMS research groups and community code developers;
  • Providing forums for standards development and resource curation;
  • Serving as mentors to MolSSI Software Fellows;
  • Working with industrial, national laboratory, and international partners;

**Approximately 50% of the Institute’s resources directly support the MolSSI Software Scientists.**
MolSSI Software Scientists

Paul Saxe
Ph.D., U.C. Berkeley

Andrew Abi-Mansour
Ph.D., Indiana Univ.

Doaa Altarawy
Ph.D., Virginia Tech

Samuel Ellis
Ph.D., Iowa State Univ.

Taylor Barnes
Ph.D., Caltech

Eliseo Marin-Rimoldi
Ph.D., Univ. Notre Dame

Jonathan Moussa
Ph.D., U. C. Berkeley

Levi Naden
Ph.D., U. Virginia

Susi Lehtola
Ph.D., U. Helsinki.

Jessica Nash
Ph.D., N.C. State Univ.

Benjamin Pritchard
Ph.D., Univ. Buffalo

Sina Mostafanejad
Ph.D., Florida State Univ.
A Sample of MolSSI Software Infrastructure Projects

- **MolSSI/EMSL Basis-Set Exchange** – an overhaul and expansion of the well-known basis-set exchange originally developed by Pacific Northwest National Labs with more than 40,000 unique visitors per month;

- **MolSSI Integral Reference Project (MIRP)** – provides both reference data and reference implementations of common integrals found in computational chemistry;

- **MolSSI QCArchive** – open, community-driven, multi-use quantum chemistry databases of benchmarks, force-field-related datasets, and other information for data mining and other machine learning initiatives;

- **MolSSI QC Schema** – a JSON-based standard for common data to enable more complex workflows among quantum chemistry codes;
MolSSI SEAMM (Simulation Environment for Atomistic and Molecular Modeling) – a user-friendly computational environment for creating and running simulations of molecular, crystalline, amorphous, and fluid systems that are described at the level of atoms;

MolSSI Driver Interface – provides a standardized API for fast, on-the-fly communication between computational chemistry codes;

MolSSI COVID-19 Molecular Structure and Therapeutics Hub – a community-driven data repository and curation service for molecular structures, models, therapeutics, and simulations related to computational research related to therapeutic opportunities for fighting the COVID-19 pandemic.
The Basis Set Exchange (BSE):
basissetexchange.org

- Original launch: June 1, 2019
- Website & library continuing to be actively maintained and expanded
  - ~600 basis sets
  - 21 formats
- General approval of new features
  - Direct links to basis sets is a common use case
  - Conversion between formats
- Very popular with users!

Software Scientists: Benjamin Pritchard and Doaa Altarawy
Users of the Basis Set Exchange Website

June 2020

Software Scientists: Benjamin Pritchard and Doaa Altarawy
The Basis Set Exchange (BSE):
basissetexchange.org

- Usage is steady for the past year
- Users perform multiple actions
- Month of June 2020:
  - 10,700 Unique Users
  - 49,000 Actions
  - 21% Visitors are from US

Software Scientists: Benjamin Pritchard and Doaa Altarawy
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 via structured data
QCArchive: Software Ecosystem

- Quantum chemistry software projects for all CMS developers
- Developed openly on GitHub (312 summed Github stars)
- Major external contributors:
  - Amir Kimiyaie (Virginia Tech, GRA)
  - Lori Burns (Georgia Tech)
  - David Dotson (OpenFF)
  - Trevor Gokey (OpenFF)
  - John Gentle (TACC)

QCSchema
- Standardized I/O for quantum chemistry

QCElemental
- Units
- QCSchema Models
- Molecule Parsing
- Visualization

QCEngine
- Runs external CMS software
- Consumes and produces QCSchema
- Not just quantum chemistry

QCFractal
- Structured data storage and retrieval
- High-throughput calculations
- Python interface
- Visualization

qcarchive.molssi.org
MolSSI COVID-19 Molecular Structure and Therapeutics Hub
covid.molssi.org

- Major open-access data and tools repository spun up in a matter of weeks
- Partnership between MolSSI and BioExcel
- Contributions are rapidly reviewed and vetted by community expert volunteers
- Submissions can be tagged with DOIs upon request
- Large-scale contributions from dozens of around the world;
- Dozens of organizations contributing or in talks: Folding @ Home, AWS, Microsoft AI for Good, Human Brain Initiative, JEDI, D.E. Shaw, Open Science Grid, PRACE, IRIS-HEP, Zenodo, CERN, etc.
- Led by MolSSI co-Director Teresa Head-Gordon

From DE Shaw

Software Scientists: Levi Naden, Andrew Abi-Mansour, Sam Ellis
MolSSI Software Fellows

- A cohort of ~35 Fellows supported simultaneously – graduate students and postdocs selected by the Science and Software Advisory Board from research groups across the U.S.
- Fellows work directly with the Software Scientists, thus providing a conduit between the Institute and the CMS community itself.
- Fellows work on their own projects, as well as contribute to the MolSSI development efforts where appropriate, and they engage in outreach and education activities under the Institute guidance.
- Funding for MolSSI Software Fellows follows a flexible, two-component structure, providing up to two years of support.

Approximately 30% of the Institute’s resources directly support the MolSSI Software Fellows.
MolSSI Software Fellowship Program: Impact

• We currently support 35 Fellows simultaneously for a total of 75 funded over the lifetime of the project.
• MolSSI Software Fellows report at least 43 publications, 72 presentations, and 46 source code repositories arising from their projects so far.
• MolSSI Software Fellow Internships, Employment, and Awards:
  • Assistant Professorship at Emory University
  • NVidia – GPU-Accelerated Data Analytics Software Internship
  • Schrödinger, Inc. – Machine Learning Summer Internship
  • Corning, Inc. – NSF Internship Program
  • DeepCure, Inc. – Computational Chemist
  • Google, Inc. – Software Engineering Internship
  • Q-Chem, Inc. – Internship
  • 2019 Michigan Institute for Computational Discovery and Engineering Fellow
  • Peter A. Rock Price in Physical Chemistry, U.C. Davis
  • Best Poster Award at the 2019 Conference of the International Society of Theoretical Chemical Physics
  • Junior Research Award from the Institute for Advanced Computational Science
  • American Physical Society Travel Award
  • Fellows Society, Florida State University

https://molssi.org/molssi-software-fellows/
MolSSI Community Code Partners So Far...

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

We encourage all community codes in the computational molecular sciences to work with us!
MolSSI Education Milestones 2019

- 17 events held in 2019 including biennial MolSSI Software Summer School and 5 day camp for undergraduates at Rice University
- Over 450 students participated in person in MolSSI Education events last year
- Large expansion and development of workshops and online resources

Locations of 2019 and early 2020 MolSSI Education Events
2 week, 50 student summer school at the Texas Advanced Computing Center (TACC) in Austin, TX.

10 day curriculum - Python, C++, Parallel Programming, MolSSI “Best Practices”

9 out of 11 software scientists participated as instructors.

Three software fellow instructors.
Responding to COVID-19

- MolSSI Education usually operates on a personal, face to face basis
- In response to covid-19, we had to rethink all of our 2020 plans
- MolSSI Education reacted quickly - announcing our first webinar series on March 24, 2020

- Have held three webinar series - April, May, and June. All have had 50+ students and hit registration caps in less than 24 hours.
Expansion of Online Materials

- **Python Package Development Best Practices** - material rewritten to more accurately reflect student needs.
- **Python Data Analysis** - Lessons made to follow Python Scripting Workshop to give more information about Python libraries numpy and pandas.
- **Parallel Programming** - Parallel Programming with MPI and OpenMP. Lessons expanded by Taylor Barnes.
- **QM Tools** - Introduces several types of quantum chemistry calculations a student might use, including geometry optimizations, inter- and intra-molecular potential energy scans, and energy calculations.
- **MM Tools** - Introduces molecular dynamics simulations using the software OpenMM, and analysis of simulation results using MDTraj.
- **ab initio MD** - hands-on introduction to concepts in computational molecular science through the development of a simple molecular dynamics program using forces derived from ab initio calculations.

education.molssi.org
MolSSI Webinars

- April 2020 - Python Data and Scripting Webinar Series (55 students registered)
- May 2020 - Python Package Development Best Practices (61 students registered)
- June 2020 - Python Data and Scripting Fast Track Series (for summer research students) (105 students registered)**

**For the June webinar, we focused on undergraduates - special registration period for MERCURY and MU3C
MolSSI Webinar Instructors

Dr. Jessica Nash
MolSSI

Prof. Ashley Ringer McDonald
California Polytechnic State University

Prof. Jay Foley
William Paterson University

Dr. Dominic Sirianni
University of Richmond

Samragni Banerjee
MolSSI Software Fellow

Victor Chavez
MolSSI Software Fellow

Dr. João Rodrigues
MolSSI Software Fellow

education.molssi.org
MolSSI Webinars Reach a Global Audience

Student institution locations - April, May, June webinars

education.molssi.org
MolSSI Board of Directors

Cecilia Clementi, Rice U., Co-Director for Biomolecular Simulation and International Engagement

T. Daniel Crawford, Virginia Tech, Director

Robert J. Harrison, Stony Brook U., Co-Director for Parallel Computing and Emerging Technologies

Teresa Head-Gordon, U.C. Berkeley, Co-Director for Laboratory, Industrial, and Academic Outreach and Education

Shantenu Jha, Rutgers U., Co-Director for Software Engineering Process, Middleware, and Infrastructure

Anna Krylov, U. Southern California, Co-Director for Quantum Chemistry

Theresa Windus, Iowa State U., Deputy Director and Co-Director for Code and Data Interoperability

Dominika Zgid, U. Michigan, Co-Director for Materials Science
MolSSI Highlights So Far

- **Twelve Software Scientists** on board, and we are continuing to add to our team.

- **25 software workshops with more than 850 participants** so far; many more virtual workshops planned during the pandemic.

- New software components currently under development including an **open QM database (QCArchive)**, a **general QM/MM driver**, a **new basis set exchange**, a **user-friendly environment** for complex workflows, and more.

- **Hundreds of students** directly engaged so far in two Software Summer Schools, “Best Practices” workshops, Software Fellowship bootcamps, undergraduate programming schools, and webinars. Many of our educational curricula are now available online.

- **35 Software Fellows currently supported** for a **total of 75 Fellows** funded overall.

- Our COVID-19 molecular structure and therapeutics hub is engaging the global biomolecular simulation community.
Acknowledgments

- The many dozens of members of the CMS community who helped to develop the vision for – and reality of – the Institute over the last seven years;
- NSF OAC-1547580.

Watch molssi.org for the latest information!