MolSSI Education: Empowering the Next Generation of Computational Molecular Scientists.

Jessica A. Nash[1], Mohammad Mostafanejad[1], T. Daniel Crawford[1], and Ashley Ringer McDonald[1]

[1]Affiliation not available

April 20, 2022

Abstract

The Molecular Sciences Software Institute (MolSSI) is a research and education center that supports software development in the computational molecular sciences (CMSs). One of MolSSI's core objectives is to provide education and training for the next generation of computational researchers. MolSSI Education targets various career stages and skill levels through its live workshops, online resources, and software fellowship program. MolSSI Education focuses its efforts within four areas: programming and software development, high-performance computing (HPC) and artificial intelligence (AI), faculty and curriculum development, and the MolSSI software fellowship program. This article delineates educational efforts at the MolSSI, overall goals, and resources that can be useful to researchers in the computational molecular sciences.

Introduction

Computational Molecular Science is a broad term that describes the application of computational resources to chemical theory and systems. It is part of many fields, including chemistry, physics, materials science, and chemical engineering, and encompasses methods such as molecular mechanics, electronic structure, and machine learning (ML). Today, computational methods are widely used in industry and academia and are indispensable parts of many scientific breakthroughs.

Traditionally, computational practitioners have been the ones responsible for building the tools of the trade, namely, software. Over the past few decades, computational researchers have developed dozens of software packages that are used by thousands of scientists worldwide. Often, software has been seen as a by-product of scientific research. However, as scientific problems become more complex, it is necessary for scientists to focus on software engineering, a trend that is reflected in the emergence of the Research Software Engineer. For scientific software developers, deep knowledge of programming languages can be vital for creating performant and usable software. For researchers, knowing how to program can help improve research efficiency and increase the speed of data analysis. Thus, computational scientists often need both mastery of their scientific field and competency in computing skills such as programming and software development.

The Molecular Sciences Software Institute (MolSSI) aims to enable new science by supporting software development efforts in computational molecular science (CMS). MolSSI was founded in 2016 with a grant from the US National Science Foundation (NSF) and is a collaborative, multi-institute center. In 2021, the NSF renewed MolSSI's funding for an additional five years. MolSSI provides software infrastructure, standards, community engagement, and training and education. MolSSI's open-source software projects address several needs in the CMS community(*MolSSI Software Projects*, n.d.). MolSSI's training programs include workshops, summer schools, and a fellowship program for graduate students and post-doctoral researchers.

Since its establishment, MolSSI has funded 93 Software Fellow projects, hosted more than 25 community-led workshops and reached more than 1500 students through educational events.

Our education program focuses on four areas: (i) programming and software development, (ii) high-performance computing, (iii) faculty and curriculum development, and (iv) a software fellowship program. By working in these areas, we hope to reach CMS researchers across a broad range of skill levels and career stages. A full list of MolSSI Education resources can be accessed on the MolSSI Education website.(*MolSSI Education*, n.d.) For all of our training materials, we directly engage with the CMS community. We deliver these resources through synchronous workshops and asynchronously through the MolSSI Education website or YouTube channel.

MolSSI Education Components

Training in Programming and Software Development

The MolSSI's programming and software development resources are designed for beginner to intermediate programmers and introduce fundamental principles using examples relevant to computational molecular scientists. MolSSI's resources in this area focus on establishing programming skills and best practices for scientific software development using specific examples. Our curricula in programming and software development is created by MolSSI Software Scientists and associates.

We lead our programming and software development workshops using a live-coding style. Live-coding is a method popularized by the Software Carpentry organization and has been found to be successful in training novice programmers. In the live-coding approach to teaching, an instructor will share their screen and type code into their programming environment while explaining the thought process and reasoning behind their actions. We intersperse live coding with small challenges or exercises to allow students to apply concepts they have just learned.

MolSSI Education offers an introduction to Python programming in its flagship undergraduate workshop, Python Data and Scripting for Computational Molecular Scientists. In this workshop, students are introduced to Python syntax, working with text files, visualization, and running command-line programs. We offer this workshop synchronously once or twice a year. Registration is free and open to the public, though we typically partner with organizations that focus on undergraduate researchers such as the MERCURY Consortium(*MERCURY Consortium – Molecular Education and Research Consortium in Undergraduate Computational Chemistry*, n.d.; Mcdonald et al., 2020).

For students who have more experience programming or who are planning to work on software development projects, MolSSI offers the Python Package Best Practices workshop. The "Best Practices" covered in this workshop are topics recommended and often used in scientific and open source software development, presented in a cohesive, hands-on format. Topics include version control, collaboration workflows, testing, documentation, and project structure. These are all practical skills widely used in software projects but currently rarely taught formally. Concepts are covered at a high level first, then we demonstrate with hands-on material. For example, when introducing software testing, we discuss the benefits and motivations of testing and also show specific examples of how one might test Python code. These workshops can be requested for groups or universities by contacting the MolSSI Education team. We typically offer this workshop at least twice a year either to the public or in partnership with an academic research group.

MolSSI's other efforts in this area include a workshop on data visualization, and a Python scripting workshop aimed at biochemists . Additionally, MolSSI offers workshop materials on object oriented programming and design patterns. A full list of resources is given in Table 1.

Training in High-Performance Computing and Artificial Intelligence

MolSSI Education's newest initiative is in high-performance computing (HPC), artificial intelligence (AI) and ML. Our HPC and AI Education Programs consist of five major divisions: online educational resources,(*MolSSI Education*, n.d.) certified university curricula, industrial training programs, instructor-led hands-on workshops, and community guidelines and best practices for the CMSs(*MolSSI Guidelines, Checklist, and Best Practices*, n.d.). MolSSI Education launched this initiative in 2021 in recognition of the key focus areas of the Exascale Computing Project (ECP), top national priorities and strategic plans.

Table 1: Online resources from the MolSSI Education Program. [a] indicates that resources are available on the MolSSI Education website at `education.molssi.org/resources`. [b] are under development.

Topics	Resources
Programming and Software Development[a]	Python Data and Scripting Python Data and Scripting for Biochemistry Scientific Data Visualization using Python Best Practices in Software Development Object Oriented Programming and Design Patterns
Molecular Science[a]	Quantum Mechanics Tools Molecular Mechanics Tools
Fundamentals of HPC[b]	Principles of Scientific HPC RAJA and Kokkos models for abstraction and portability Slurm and Moab workload management systems and schedulers
Homogeneous Parallel Programming[b]	Message Passing Interface Shared-Memory Parallel Programming with OpenMP
Heterogeneous Parallel Programming	Fundamentals of Heterogeneous Parallel Programming with CUDA C/C++[a] A Systematic Approach to CUDA C/C++ Parallel Programming[b] Applications of Heterogeneous Programming in Computational Molecular Scie OneAPI: A Unified Approach to Heterogeneous Parallel Programming[b] Modern Platforms for Heterogeneous Parallel Programming[b]

We base our HPC online educational resources on the open-source industry standards, vendors' expert recommendations, and community guidelines and best practices. Table 1 provides a high-level view of our online educational resources in HPC designed for a variety of user backgrounds and skill levels. For example, the Fundamentals of HPC program would be most beneficial to beginner and intermediate-level users while the Homogeneous Parallel Programming resources are designed for intermediate and advanced HPC users. The Heterogeneous Parallel Programming section, on the other hand, provides online resources for all background levels.

A key component of the MolSSI Education program in HPC is industry partnerships. In collaboration with NVIDIA Deep Learning Institute (DLI) and Intel, we offer a series of certified instructor-led hands-on workshops. Our collaboration with NVIDIA DLI through the Certified Instructor and University Ambassador Programs allows the members of the CMS community to have free access to certified training programs that otherwise would involve a registration fee. These programs are divided into four major specialization areas: Data Science, Deep Learning, Accelerated Computing and Conversational AI. Intel, on the other hand, enabled us to offer a variety of training modules within two main focus areas: Essentials of Data Parallel C++ (DPCPP) and Basics of OpenMP Offload. For our certified instructor-led hands-on workshops, university courses and industrial training programs, we provide the registered users with access to a cloud virtual machine armed with a variety of accelerator architectures such as CPUs, GPUs and FPGAs. The users

can access all training resources using either the command line or JupyterLab's user-friendly notebooks.

The last component of MolSSI'S HPC and AI Education initiative involves establishing best practices for data management. Due to the dire need of the scientific software community for improving the quality of the scientific data management plans according to FAIR principles(Wilkinson et al., 2016) and ensuring the reproducibility, interoperability, and replicability of the computational research products, we have developed a public platform(*MolSSI Guidelines, Checklist, and Best Practices*, n.d.) for publishing community guidelines and best practices for all domains of CMS. We have founded this platform upon our years of experience in serving the CMS and the open-source software communities, hundreds of interviews and surveys gathered from our discovery project, expert recommendations and best practices documents provided by the major HPC vendors and our collaborations with the NSF-funded projects such as the Xpert Network.(Barakhshan & Eigenmann, 2021)

Faculty and Curriculum Development

MolSSI's efforts in faculty and curriculum development focus on helping faculty members to incorporate MolSSI resources into their classes and promoting the use of programming in the chemistry curriculum. Though MolSSI Education resources were originally designed to enable students to more effectively participate in research experiences, there have been a notable number of faculty using the MolSSI Education materials as a starting point to develop curricular resources to incorporate programming into their courses. Examples of these types of curricular innovations are highlighted in MolSSI's recent ACS Symposium Series Book, *Teaching Programming Across the Chemistry Curriculum* (McDonald & Nash, 2021). The examples in this book incorporate programming in a variety of different classes, ranging from general chemistry to graduate-level courses. Programming is used to analyze data, make visualizations, demonstrate physical phenomena that are otherwise hard to describe, solve chemistry problems numerically, and more.

The MolSSI has partnered with faculty professional development groups such as Enhancing Science Courses by Integrating Python (ESCIP) (https://escip.github.io/) and Psi4Education (https://psicode.org/posts/psi4education/) to work with faculty to develop curricular resources. ESCIP is a group of faculty from the Cottrell Scholars Program, sponsored by Research Corporation for Science Advancement, who develop and share curriculum that utilizes Python programming for chemistry, physics, and math courses. Psi4Education is the education and outreach program of the quantum chemistry software package Psi4 that uses Psi4's Python interface, Psi4Numpy to create lab activities for use across all levels of the chemistry curriculum. MolSSI supports these curricular development efforts by meeting with these faculty development groups to advise them on strategies and best practices for teaching programming. MolSSI also sponsors symposia and workshops for faculty at conferences, such as the Biennial Conference on Chemical Education, and hosts instructor training workshops to help faculty upskill so they can better teach best programming practices to their students.

Additionally, MolSSI is involved in the University Ambassador Program with the NVIDIA DLI. Through this partnership, MolSSI has access to teaching kits designed for university courses, which include syllabi, lecture notes, curricular resources, and programming activities that cover major topics in HPC, deep learning, and robotics. MolSSI can support faculty in implementing these courses at their institutions and help them customize the contents of each course in the CMS domain to meet their program goals and requirements.

MolSSI Software Fellowship Program

MolSSI provides direct support to software development efforts of early-career researchers through its Software Fellowship program. MolSSI Software Fellowships are highly selective awards that fund graduate students and post-docs who develop software infrastructure for CMS. MolSSI prioritizes projects of broad interest with potential high impact to the CMS communities. As of 2022, the MolSSI Software Fellowships are year-long awards, with calls for applications starting in February and awards starting in July.

Each cohort of MolSSI Software Fellows receives special training in a week-long Software Fellow Bootcamp. The Bootcamp curriculum includes topics in software development best practices, software design and distribution, and special topics related to the projects and interests of the software fellow cohort. The Bootcamp material draws heavily from the MolSSI Education resources featured in this paper, particularly the Python Package Best Practices and Software Design workshops.

Throughout the fellowship, the software fellows receive one-on-one mentoring from a MolSSI Software Scientist who provides guidance for the software fellow's project.

Conclusion

The MolSSI Education Team is engaged with, and welcomes comments and collaborations from members of the CMS and scientific software development communities. MolSSI disseminates all new updates and the latest releases to the educational resources through its social media accounts, newsletter and the organizational website.(*The Molecular Sciences Software Institute*, n.d.; *MolSSI Education*, n.d.) All of our educational resources including the hands-on tutorials, instructor-led workshop materials and self-paced online courses are free and available under open source licenses. Members of the scientific community can provide direct feedback, comments, or contributions to all educational resources through opening discussions, issues, and pull requests on the MolSSI Education GitHub repositories.(*MolSSI Education GitHub*, n.d.) We continuously measure the impact of our online educational resources through tracking the user analytics gathered from the hosting websites and repositories.

ACKNOWLEDGMENT

The MolSSI is funded by the National Science Foundation under grant CHE-2136142.

Jessica A. Nash is a Software Scientist and the Education Lead at MolSSI. As Education Lead for the Institute, she develops educational materials for researchers which enhance their capabilities to write and use CMS software. She also currently works as the lead developer for the web interface for MolSSI's SEAMM (Simulation Environment for Atomistic and Molecular Simulation) project.

Mohammad Mostafanejad is the Lead Software Scientist in HPC Education, AI and ML at the Molecular Sciences Software Institute. He is also a Certified Instructor and University Ambassador at NVIDIA Deep Learning Institute and Intel with contributions to major commercial and open-source quantum chemistry software packages such as Q-Chem and Psi4. His most recent efforts involve the development of ML software infrastructure for the Simulation Environment for Atomistic and Molecular Modeling (SEAMM) project.

T. Daniel Crawford is the Director of the Molecular Sciences Software Institute, as well as the University Distinguished Professor and Ethyl Chair of the Department of Chemistry at Virginia Tech. He received his bachelor's degree in 1992 from Duke University and his Ph.D. in 1996 from the University of Georgia. His research focuses on advanced quantum chemical models of molecular spectroscopic properties. He is a Fellow of the American Chemical Society and the winner of 2010 Dirac Medal of the World Association of Theoretical and Computational Chemists.

Ashley Ringer McDonald is an associate professor in the Department of Chemistry and Biochemistry at California Polytechnic State University in San Luis Obispo. She serves on the Board of Directors of The Molecular Sciences Software Institute (MolSSI) as the Co-director for Education, Training, and Faculty Development. In this role, she directs MolSSI's diverse educational programming, including developing educational resources, short courses, and workshops, and oversees MolSSI's engagement with other education efforts and groups.

Prof. Ringer McDonald's research focuses on using multiscale modeling to study molecular interactions in complex chemical contexts and developing software tools in this area. She works significantly in the area of developing curriculum and resources to integrate programming across the chemistry disciplines, and to identify and implement best practices for teaching programming in discipline-specific contexts.

References

https://molssi.org/software-projects/

https://education.molssi.org/

https://mercuryconsortium.org/

Building capacity for undergraduate education and training in computational molecular science: A collaboration between the MERCURY consortium and the Molecular Sciences Software Institute. (2020). https://doi.org/10.1002/qua.26359

https://molssi.github.io/molssi-guidelines/

The FAIR Guiding Principles for scientific data management and stewardship. (2016). *Scientific Data*, *3*, 1–9. https://doi.org/10.1038/sdata.2016.18

Exchanging Best Practices and Tools for Supporting Computational and Data-Intensive Research, The Xpert Network. (2021). *ArXiv*.

Teaching Programming across the Chemistry Curriculum. (2021). American Chemical Society. https://doi.org/10.1021/BK-2021-1387

https://molssi.org/

https://github.com/molssi-education

ESCIP — Enhancing Science Courses by Integrating Python (ESCIP) is funded by the Research Corporation for Scientific Advancement through the Cottrell Scholars Collaborative program.. https://escip.github.io/

Psi4Education: Computational Labs Using Free Software. https://psicode.org/posts/psi4education/