Sina Mostafanejad

January 2022

THE MOLECULAR SCIENCES SOFTWARE INSTITUTE

MolSSI Community Guidelines for Computational Molecular Sciences

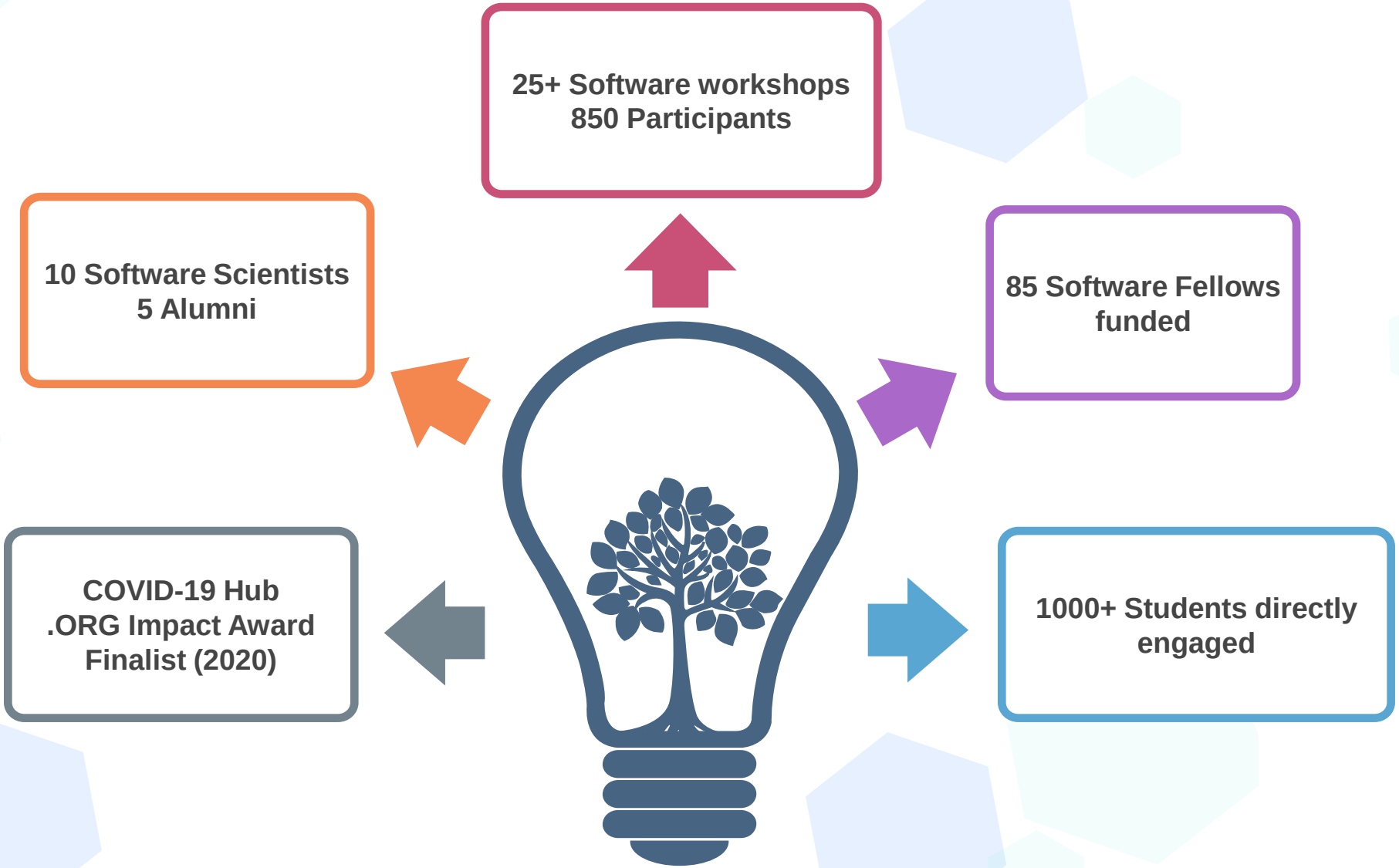
Founded in August 2016

NSF grant #OAC-1547580

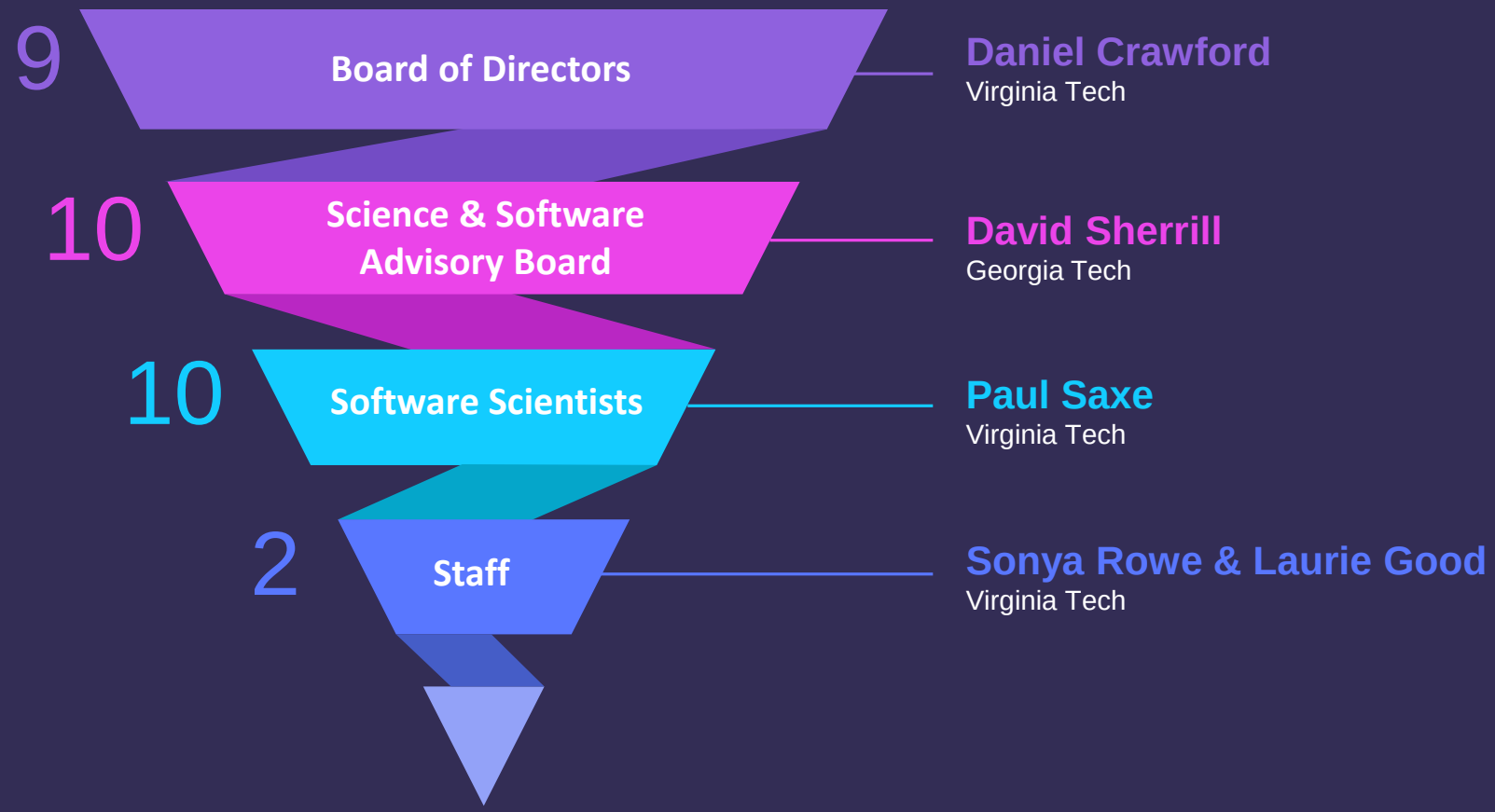
Serve and enhance software development efforts in computational molecular sciences

THE MOLECULAR SCIENCES SOFTWARE INSTITUTE

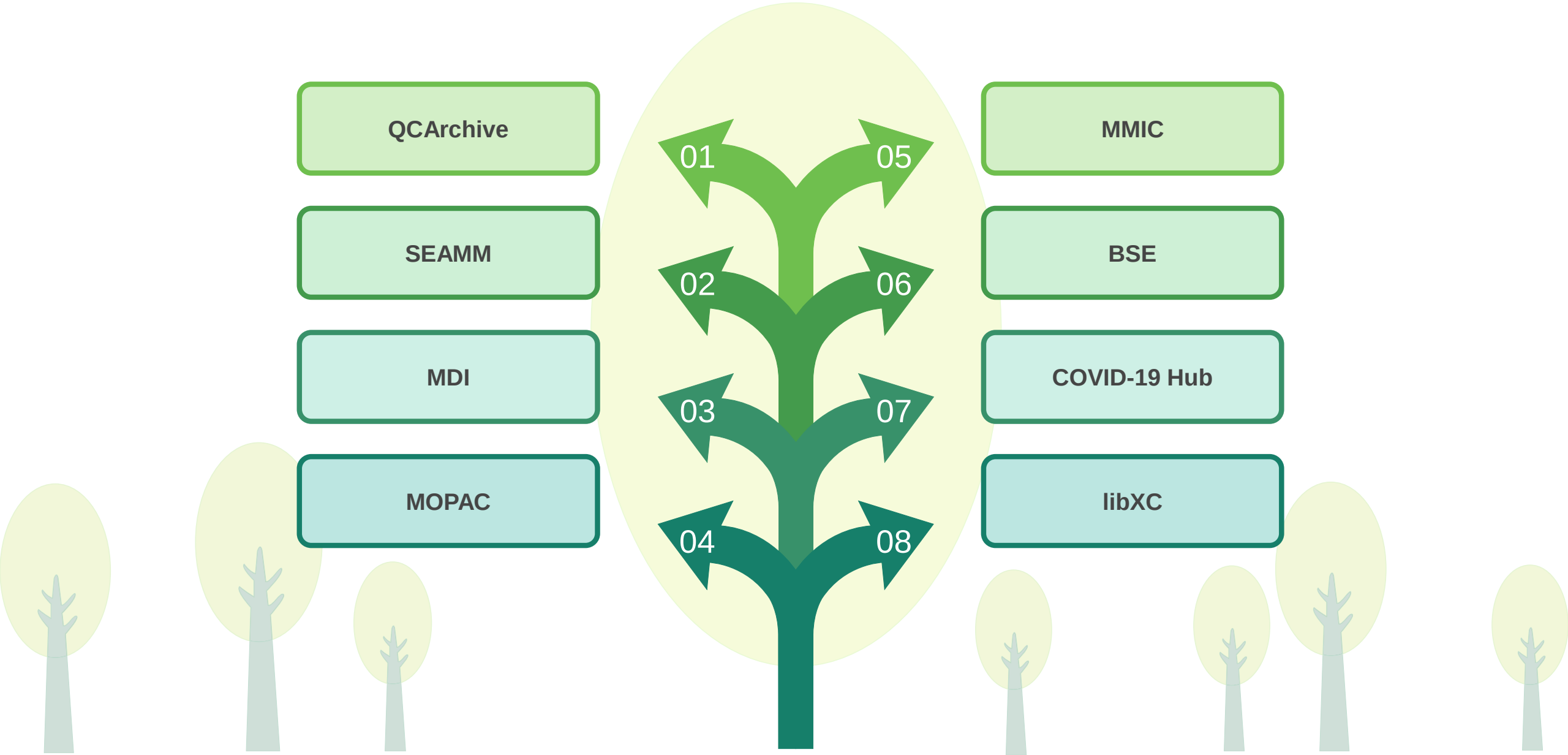
Highlights


25+ Software workshops
850 Participants
10 Software Scientists
5 Alumni
85 Software Fellows
funded
COVID-19 Hub
.ORG Impact Award
Finalist (2020)
1000+ Students directly
engaged

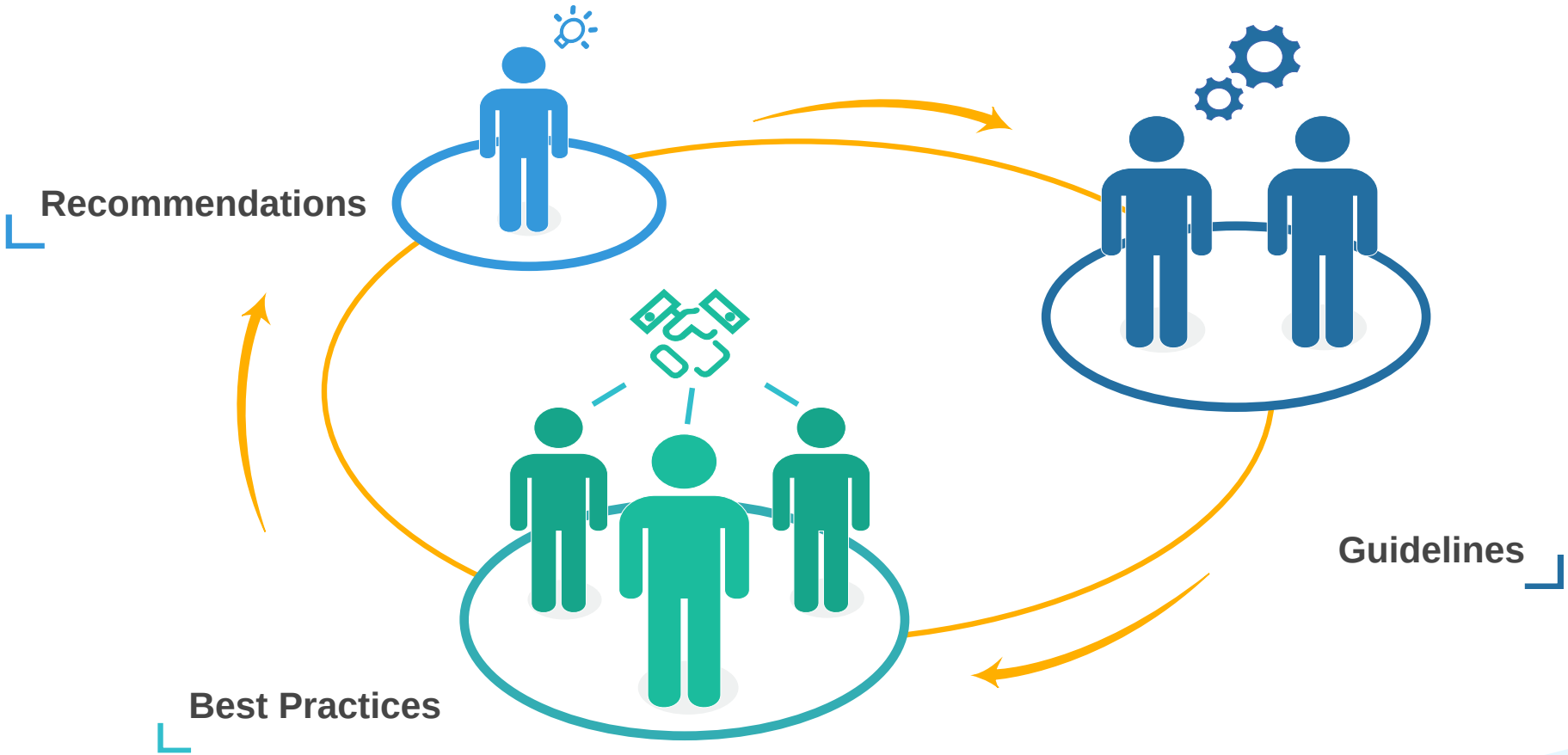
MolSSI Organization



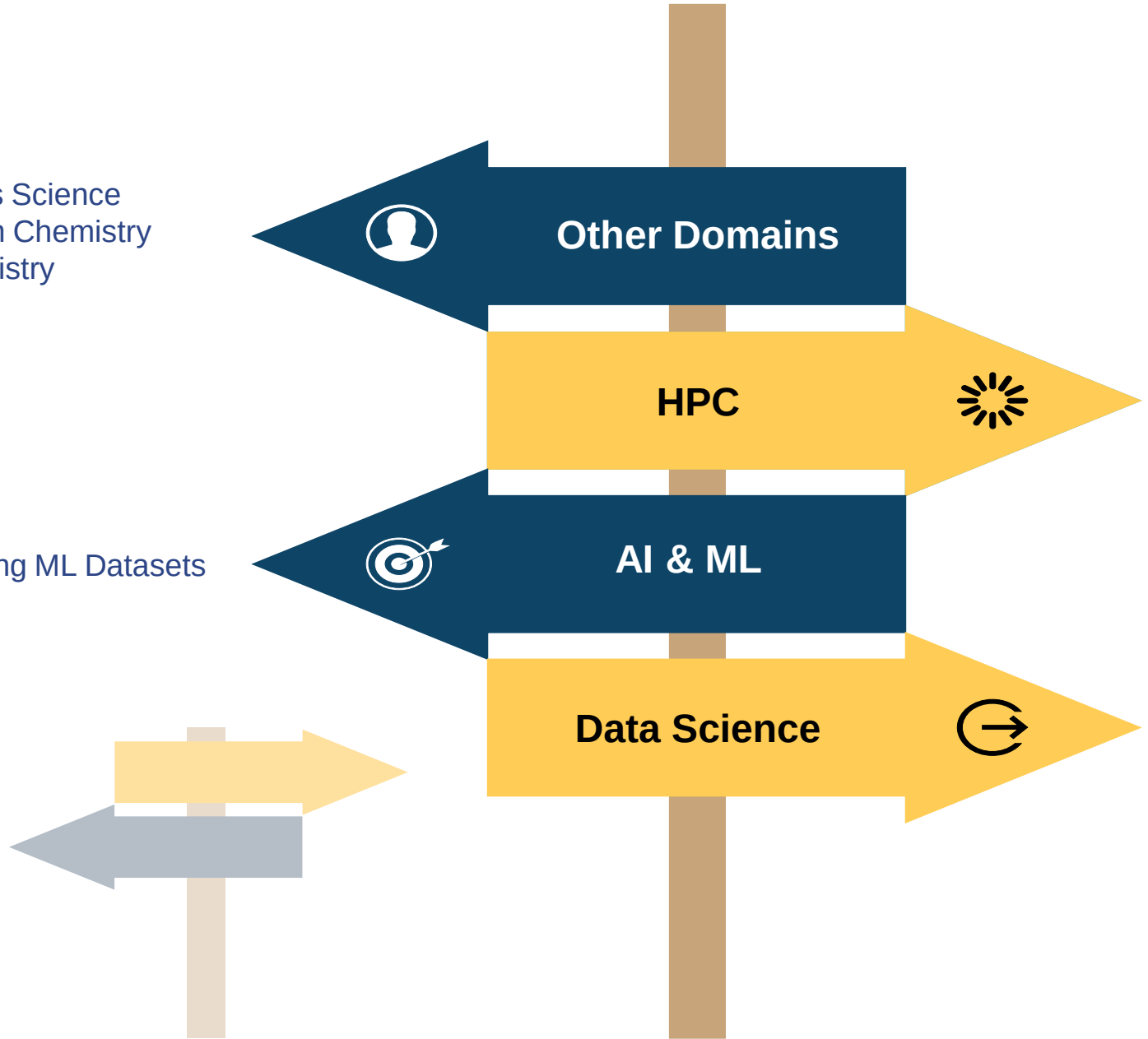
MolSSI Software Projects



Guidelines & Best Practices



Topics



- Materials Science
- Quantum Chemistry
- Biochemistry
- ...

- APOD Approach
- Profiling with Nsight Systems
- Profiling with Nsight Compute

- Formatting ML Datasets

- Publishing Datasets

MolSSI Guidelines, Checklists and Best Practices

The present website hosts the Molecular Sciences Software Institute (MolSSI)'s recommendations and guidelines to promote FAIR data management, and improve OpenSource and appropriate scientific software citation practices across all disciplines within the computer and molecular science (CMS) communities.

Our current list of guidelines cosissts of the following set of documents

Indices and tables

Next

1.1. MolSSI Formatting Guidelines for Machine Learning Products

- Source: DOI 10.5281/zenodo.5389982

This document presents a set of guidelines and best practices for formatting machine learning (ML) products (e.g., datasets, modules, models, *etc.*) before submitting them on the Zenodo platform and tagging them to one or more curated MolSSI community collections.

1.1.1. Requirements and Policies

1.1.1.1. Prerequisites

Before getting started, please take a glance at the MolSSI Publishing Guidelines on Zenodo Platform to familiarize yourself with the basic mechanics and recommended strategies for publishing your software products on Zenodo.

1.1.1.2. Dataset File Formats

Datasets are tables of data hosting instances (chemical species such as atoms, molecules, macromolecules *etc.*) in rows and features or descriptors in columns. Each label results from an experimental observation or theoretical calculation and corresponds to a feature. As such, labels are stored at the intersection of each row and column.

We recognize two main conceptual categories for featurizing the input data: (i) geometrical data (e.g., coordinates, connectivities, atomic symbols, *etc.*), and (ii) chemical features (e.g., energetics, electronic properties *etc.*).

Geometrical Data

Representation of geometrical data pertinent to individual chemical species such as molecules (or monomers, dimers, polymers, clusters, unit cells, *etc.*) is dependent upon the task and adopted ML algorithm. In general, the raw information on individual molecular structures should be stored as separate files within a subfolder of the root directory called geometries. The recommended file format for storing geometrical data is the Chemical Table Format (*.mol*) which allows for a convenient usage of popular and free chemical data conversion toolkits such as Open Babel. This representation is probably most useful before training each model since the majority of the ML models require a featurized version of these structures into a numerical representation or

September 2, 2021

Technical note Open Access

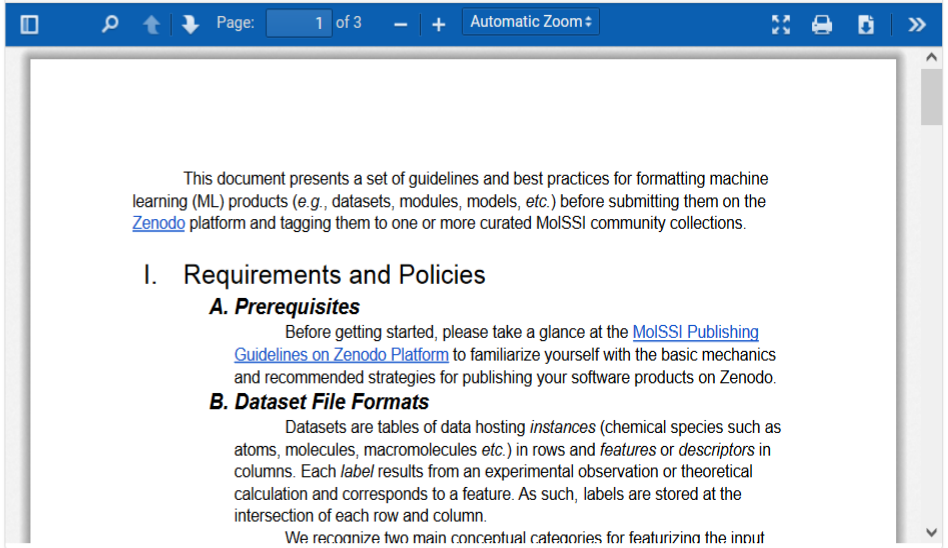
MolSSI Formatting Guidelines for Machine Learning Products

Mostafanejad, Mohammad; Saxe, Paul

This document is a part of MolSSI guidelines and best practices which focuses on formatting machine learning products (*e.g.*, datasets, modules, models, *etc.*) before submitting them on the Zenodo platform and tagging them to one or more curated MolSSI community collections.

This work has been funded by the NSF OAC-1547580 and CHE-2136142 grants.

Preview



Files (113.4 kB)

Name	Size	
ML_Formatting_Guidelines.odt	21.0 kB	Download

md5:32daf9cba263bbce5c4f744bd62f2558

Publication date:
September 2, 2021

DOI:
DOI 10.5281/zenodo.5389982

Keyword(s):
machine learning formatting guidelines MolSSI best practices MolSSI guidelines

Communities:
Molecular Sciences Software Institute
MolSSI Guidelines and Best Practices

Versions

Version 2021.9.2 Sep 2, 2021
10.5281/zenodo.5389982

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.5389981. This DOI represents all versions, and will always resolve to the latest one. Read more.

Molecular Sciences Software Institute

Recent uploads

Search Molecular Sciences Software Institute

December 21, 2021 (2021.12.21) Software Open Access View

SEAMM: Simulation Environment for Atomistic and Molecular Modeling

Nash, Jessica; Marin-Rimoldi, Eliseo; Saxe, Paul;

Also cleaned up the printing.

Uploaded on December 21, 2021

7 more version(s) exist for this record

November 27, 2021 (2021.11.27) Software Open Access View

Plug-in for SEAMM to allow custom python scripts in flowcharts

Saxe, Paul;

Provides a step in a flowchart for the user to add a Python script which will run in the current environment, giving access to the internal variables of SEAMM as well as the structures.

Uploaded on November 27, 2021

2 more version(s) exist for this record

November 27, 2021 (2021.11.27) Software Open Access View

Plug-in for SEAMM for building fluid systems with PACKMOL.

Saxe, Paul;

Provides a step in a flowchart for packing reasonably small molecules into a periodic cell using PACKMOL. This is used for preparing models of liquids and gases with one or more components given the stoichiometry. The size if the resulting model is specified by giving two independent parameters such

Uploaded on November 27, 2021

2 more version(s) exist for this record

February 10, 2021 (1.0.0) Dataset Open Access View

New upload

Community

Molecular Sciences Software Institute

Zenodo community for the MolSSI.

Curated by:
MolSSI

Curation policy:
Not specified

Created:
August 2, 2021

Harvesting API:
OAI-PMH Interface

Want your upload to appear in this community?

- Click the button above to upload a record directly to this community.
  To add one of your existing records to the community, edit the record, add this community under the "Communities" section, save, and finally publish.

Reference Handler

- **Runtime control**
- **Recommended software citation format**
- **Automatic citation counts**
- **Priority levels**
- **Export to variety of formats**
- **Reference databases**

F1000Research F1000Research 2021, 9:1257 Last updated: 12 JAN 2022

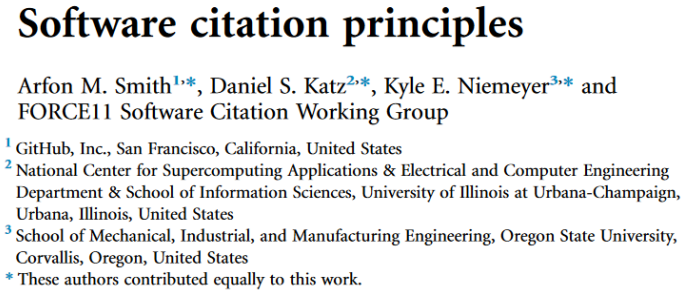
Check for updates

METHOD ARTICLE

REVISED **Recognizing the value of software: a software citation guide [version 2; peer review: 2 approved]**

Previously titled: "The importance of software citation"

Daniel S. Katz[1], Neil P. Chue Hong[2], Tim Clark[3], August Muench[4], Shelley Stall[5], Daina Bouquin[6], Matthew Cannon[7], Scott Edmunds[8], Telli Faez[9], Patricia Feeney[10], Martin Fenner[11], Michael Friedman[12], Gerry Grenier[13], Melissa Harrison[14], Joerg Heber[15], Adam Leary[16], Catriona MacCallum[17], Hollydawn Murray[18], Erika Pastrana[19], Katherine Perry[20], Douglas Schuster[21], Martina Stockhause[22], Jake Yeston[23]



Software citation principles

Arfon M. Smith[1,*], Daniel S. Katz[2,*], Kyle E. Niemeyer[3,*] and FORCE11 Software Citation Working Group

[1] GitHub, Inc., San Francisco, California, United States
[2] National Center for Supercomputing Applications & Electrical and Computer Engineering Department & School of Information Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States
[3] School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, Oregon, United States
* These authors contributed equally to this work.

- https://github.com/MolSSI/reference_handler

References

- MolSSI Website
  - https://molssi.org
- MolSSI Guidelines and Best Practices Platform
  - *https://molssi.github.io/molssi-guidelines*
- MolSSI Communities on Zenodo
  - *https://zenodo.org/communities/molssi*
  - *https://zenodo.org/communities/molssi-guidelines*
- MolSSI ***Reference_Handler*** Software
  - *https://github.com/MolSSI/reference_handler*
- Software Citation References
  - *Katz DS, Chue Hong NP, Clark T et al. "Recognizing the value of software: a software citation guide" F1000Research 2021, 9:1257 (https://doi.org/10.12688/f1000research.26932.2)*
  - Smith AM, Katz DS, Niemeyer KE, FORCE11 Software Citation Working Group. 2016. "Software citation principles" PeerJ Computer Science 2:e86 (https://doi.org/10.7717/peerj-cs.86)