MolSSI Community Guidelines for Computational Molecular Sciences

Sina Mostafanejad
January 2022
Founded in August 2016
NSF grant #OAC-1547580

Serve and enhance software development efforts in computational molecular sciences
Highlights

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

- QCArchive
- SEAMM
- MDI
- MOPAC
- MMIC
- BSE
- COVID-19 Hub
- libXC
Guidelines & Best Practices

Recommendations

Best Practices

Guidelines
Topics

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

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

HPC
- Formatting ML Datasets

AI & ML
- Publishing Datasets

Data Science
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 consists of the following set of documents

- 1. MolSSI Guidelines for Software Publications on Zenodo Platform
   - 1.1. MolSSI Publishing Guidelines on Zenodo Platform
  - 2.1. MolSSI Formatting Guidelines for Machine Learning Products
- 3. High-Performance Computing Guidelines
  - 3.1. MolSSI Guidelines on APOD Cyclic Parallelization Strategy
- 4. References
- 5. Acknowledgements

Indices and tables

- Index
- Module Index
- Search Page
1.1. MolSSI Formatting Guidelines for Machine Learning Products

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
MolSSI Formatting Guidelines for Machine Learning Products

Mohmmad Mottanaejad, Paul Save

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-2138142 grants.

I. Requirements and Policies

A. 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.

B. Dataset File Formats

Datasets are tables of data hosting variables (chemical species such as atoms, molecules, macromolecules, etc.) in rows and features or descriptions 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 mining the most useful features:

1. Structured Features
2. Unstructured Features

These categories can be further subdivided into:

- Structured Features:
  - Numerical Features
  - Categorical Features

- Unstructured Features:
  - Text Features
  - Image Features
  - Audio Features

In this document, we will focus on the structured features, as they are more commonly used in machine learning applications.

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.

Files

- MLC Formatting Guidelines.odt

21.0 kB

Download
SEAMM: Simulation Environment for Atomistic and Molecular Modeling

Nash, Jessica, Mann-Arnold, Eifor, Save, Paul.

Also cleaned up the printing.

Uploaded on December 21, 2021.

7 more version(s) exist for this record.

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

Save, 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.

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

Save, 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 of the resulting model is specified by giving two independent parameters such.

Uploaded on November 27, 2021.

2 more version(s) exist for this record.
Reference Handler

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

• 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**