MATSML: a machine-learning toolkit for materials science

Huan Tran
Georgia Institute of Technology
huan.tran@mse.gatech.edu

August 7, 2021

Contents

1 Overview

MATSML is a python-based machine-learning (ML) toolkit for some generic problems in materials science. Being initiated to support some of my (Huan Tran) non-polymer ML works, MATSML was designed to be *portable and self-contained*, providing necessary tools to follow the *entire* workflows reported, i.e., obtaining raw data, featurizing them, training models, and making predictions using the models developed. Given this objective, the materials used for my works can be found in some specific examples while other examples are for tutorial purposes and beyond. MATSML is freely available at `https://github.com/huantd/matsml.git`.

All computed data used in this toolkit are from my works. In cases that involve experimental data, when copyright and ownership prevent the sharing, suitable freely available alternative data are provided for demonstration purpose. Questions, requests, and comments on any parts of MATSML are welcome at `huan.tran@mse.gatech.edu` or `huantd@gmail.com`.

2 Technical details

2.1 Code and installation

MATSML is freely available at `https://github.com/huantd/matsml.git`. Dependencies needed for MATSML are

1. `numpy`, version
2. `scikit-learn`, version
3. `keras`
4. `tensorflow`
5. `tensorflow_probability`
6. `matplotlib`

Users are suggested to (1) create a devoted environment for this toolkit, e.g., using conda with python3, (2) install the dependencies, and (3) then install the source code of MATSML by

```
python3 setup.py install
```

2.2 Functionalities

In a typical workflow of materials informatics, one needs to (1) prepare/generate data, (2) featurize (or fingerprint) the data, (3) learning the featurized data to make models, and (4) using the developed models to make prediction on the featurized data of new cases. This toolkit supports all of these steps with an exception of (1). Because preparing or generating data is generally very time- and labor-consuming, some raw datasets are provided. More details on the data can be found in Sec. 3 of this manual.

3 Data

3.1 Sources

Computed data used in this toolkit come from the following papers

1. Vu Ngoc Tuoc and Tran Doan Huan, *Predicted binary compounds of tin and sulfur*, J. Phys. Chem. C **122**, 17067 (2018),
2. Tran Doan Huan, *Pressure-stabilized binary compounds of magnesium and silicon*, Phys. Rev. Mater. **2**, 023803 (2018),
3. Chiho Kim, Tran Doan Huan, Sridevi Krishnan, and Rampi Ramprasad, *A hybrid organic-inorganic perovskite dataset*, Sci. Data **4**, 170057 (2017).

For the purpose of using in this toolkit, these datasets were collected, formated, and provided as specified in the corresponding examples.

3.2 Specifications

Having raw data, the next step is to