DOCUMENTATION TO ACCOMPANY MATERIALS FINGERPRINTING MANUSCRIPT

ADAM SPANNAUS[1], KODY J. H. LAW[2], PIOTR LUSZCZEK[3], FARZANA NASRIN[4], CASSIE PUTMAN MICUCCI[5], PETER K. LIAW[6], LOUIS J. SANTODONATO[7], DAVID J. KEFFER[6], *, AND VASILEIOS MAROULAS[8,*]

This document provides descriptions of the python source files and data to accompany the Materials Fingerprinting manuscript of Spannaus et al. This software depends on the following libraries and has been tested with the versions listed:

- python - 3.8,
- numpy - 1.18 and 1.19,
- sklearn - 0.23,
- scipy - 1.4,
- matplotlib - 3.2, and 3.3,
- ripser - 0.4.1 and 0.5.2.

The software has been tested with the listed versions of these libraries on linux and Mac OSX. To install any missing dependencies, all may be installed via 'pip', eg, `$ pip install Ripser`, from the command line.

To get started after installing the required libraries, download the repository and save the directory to the Desktop. All commands are to be run from the terminal in the Materials-Fingerprinting directory. The program is set to compute the distances in parallel and cannot be run from within an `ipython` session. To run the binary classification, type `$ python tda_classify2.py`, and the three-way classification, enter `$ python tda_classify3.py`, from the command line in the fingerprinting directory, which will run the classification with the default settings of added noise ($\sigma = 0.25$) and percent missing (33%). Furthermore the code is set to run the classification method on a subset of the data presented in the manuscript. The number of structures, along with the added noise and sparsity may be changed in the code as described below.

We give a description of each file.

- `tda_classify2.py:` These files run the main classification routines for binary scenario. The default setting is for an even split between 500 BCC and FCC structures. If you want to investigate a different proportion, that value may be set on line 181, and the number of structures is set on line 104. To change the amount of noise or percent missing, you may specify these values in line 179 and 180 in `tda_classify2.py`.
- `tda_classify3.py:` These files run the main classification routines for the multi-class classification setting. A multi-class scenario, the number of BCC and FCC structures to use may be set on line 104 of the file `tda_classify3.py` and the number of HCP structures may be set on line 105 of the same file. If you want to investigate a different proportion, that value may be set on line 189. To change the amount of noise or percent missing, you may specify these values in line 187 and 188 in the file `tda_classify3.py`.
- `classify_utils.py:` This file contains the functions to load the data and create persistence diagrams. The path to load the data is presently set to use the Desktop folder. If the code is installed in a different directory, from the Desktop, the path must be specified on line 114. If this path is incorrect, the program will prompt for the correct path to the data.

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable,world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

- `dist.py:` This file contains the functions and classes for constructing the feature matrix for the classification algorithm. It is set to process all the distances in parallel, and must be run from the command line. It will not work correctly from an interactive python session, such as `ipython` within Spyder.
- `distances.py:` Contains the d_p^c, equation 3.1 in the manscript, and Wasserstein persistence diagram distance function definitions.

(1) Oak Ridge National Laboratory, Oak Ridge, TN 37830

(2) School of Mathematics, University of Manchester, Manchester, UK

(3) Innovative Computing Laboratory, University of Tennessee, Knoxville, TN 37996

(4) Department of Mathematics, University of Hawaii at Manoa, Honolulu, HI 96822

(5) Eastman Chemical Company, Kingsport, TN 37662

(6) Department of Materials Science and Engineering, University of Tennessee, Knoxville, TN 37996

(7) Advanced Research Systems, Inc., Macungie, PA 18062

(6) Department of Materials Science and Engineering, University of Tennessee, Knoxville, TN 37996
Email address, Corresponding author: `dkeffer@utk.edu`

(8) Department of Mathematics, University of Tennessee, Knoxville, TN 37996
Email address, Corresponding author: `vmaroula@utk.edu`