

Curating a Standardized Dataset for Statistical Biosignature Classification

Tao Sheng¹, Sunanda Sharma², Abdullah Shahid³, and Diana Gentry⁴

¹University of Pittsburgh, Pittsburgh, PA (tao_sheng@pitt.edu)

²Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA (susharma@jpl.nasa.gov)

³North Carolina State University, Raleigh, NC

⁴NASA Ames Research Center, Mountain View, CA

Introduction: In recent years, machine learning has been explored as a toolkit for planetary science and operations [Helbert, Azari]. Machine learning has been used to improve our understanding of possible biosignatures and mineral signatures to improve science return on future missions [Warren-Rhodes, Cleaves].

This project, Statistical Classification of Biosignature Information (SCOBI), aims to use the large amount of data from terrestrial biogenic and abiogenic systems to create a classifier for life detection and identify the relative importance of different features. To date, we have included 1277 distinct observations of sixteen representative systems, divided into different categories based on biogenicity. These were drawn from public databases, available literature, or collected using internal instruments. We utilized data from four different analytical techniques and compared the performance of multiple supervised classification algorithms (see Gentry et al. for application of results, Shahid et al. for methodology, this conference). Selection of an appropriate dataset for training and testing is an integral component of this endeavor. Here, we detail the process of selecting and standardizing this diverse dataset.

Dataset Selection: We only included sample data that was properly labeled and standardizable. We performed a wide search for relevant data, drawing from geochemistry, geology, organic chemistry, planetary science, and biology. We excluded “edge cases” of life, such as prions and viruses, and instead focused on more obvious samples that would be clearly indicative or not indicative of life, or mixed systems that include both.

Selection of Techniques. To sufficiently represent the diverse set of data that could be generated from a multi-instrument payload, we selected these techniques: X-Ray diffraction, laser induced breakdown spectroscopy (LIBS), Raman spectroscopy (514.5, 532 nm), and visible-near infrared (VNIR) reflectance spectroscopy. Together, these techniques provide elemental, isotopic fractionation, and chemical composition information about a sample. They are all relatively common laboratory techniques that have been used by rover missions on Mars.

Selection of Data Sources. For each of the techniques, we searched for published datasets and public databases. We used the following databases: RRUFF Project, SOLSA Raman Open Database, Handbook of Minerals Raman Spectra ENS-Lyon, the Caltech Mineral Spectroscopy Server, the PDS Geosciences Spectral Library, the USGS spectral

library, the Biosignatures @ Cornell Astro Catalogue, the ECOSTRESS Spectral Library, the Keck/NASA RELAB, and the NIST Data Repository. We also requested data from several authors who made their data available upon request.

Technique	# Observations
Elemental (XRD, LIBS)	77
Raman (514,532 nm)	447
VNIR Reflectance	717
Isotope	36
Total	1277

Table 1. List of Observations used from Each Technique.

Dataset Standardization: For data from each technique, we followed four major steps: download (.txt/.csv file typically), format (e.g. truncate, interpolate, filter), plot and verify, and feature extraction for input to the algorithm. The formatting step was the most complex, as data collected by different instruments vary in terms of noise, background or artifacts, and signal. With the large number of samples, our aim was to preserve major patterns, rather than all aspects, of a spectrum. We discarded data that were incomplete, poorly labeled or annotated, or acquired with an included technique but using very different parameters or methods.

Results: Using our disparate data sources, we achieved an overall ROC AUC of 0.79 in the mean votes model. Removing elemental abundance data reduced the overall AUC by 7.5%, highlighting how data types have different importance. We also identified certain possible biases in our dataset selection. First, we encountered challenges obtaining data from all techniques for each system. For instance, elemental data on living systems is collected much less often than on mineral samples, which may reflect their relative utility in each field. Second, there were cases in which all or most of the data from a specific technique for a system was from a single source. This opened the possibility that, if a given data type underperformed, this could be due to the acquisition conditions or instrument, rather than the technique. In the future, we aim to expand and finalize our dataset so each system has a similar number of samples, multiple data sources are available for each technique, other techniques such as mass spectrometry can be included.

References:

- Azari, A.R. et al. (2021) *Planetary Science and Astrobiology Decadal Survey White Paper #128*.
- Helbert, J. et al. (2022) *Machine Learning for Planetary Science*. Elsevier Inc.
- Warren-Rhodes, K. et al. (2023) *Nature Astronomy*.
- Cleaves, H.J. et al. (2023) *PNAS*.