STATISTICAL CLASSIFICATION OF BIOSIGNATURE INFORMATION: ADVANCING LIFE DETECTION THROUGH MACHINE LEARNING. Casey Benderoth[1,2], Abdullah Shahid[3], Tao Sheng[4], Sunanda Sharma[5], Diana Gentry[6], [1]NASA Ames Research Center OSTEM Internship Program (cbenderoth17@gmail.com), [2]University of Vermont, Burlington, VT, [3]North Carolina State University, Raleigh, NC, [4]University of Pittsburgh, Pittsburgh, PA, [5]Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA, [6]NASA Ames Research Center, Mountain View, CA.

Introduction: The search for extraterrestrial life is among the most compelling pursuits in planetary science, offering novel understandings of requirements for life, planetary exploration, and habitat identification. Central to this endeavor is the identification of biosignatures—measurable indicators of past or present life that are not formed by abiotic processes [1]. The Statistical Classification of Biosignature Information (SCOBI) project is building and analyzing a multi-modal terrestrial analog dataset to create a binary classifier (a technique of machine learning) to accurately classify potential biosignature observations as *indicative* or *non-indicative* of life. This paper focuses on the creation of the standardized dataset for classifier training and evaluating feature importance across data types.

Our dataset includes 1,975 unique observations derived from various instrumentation techniques of 21 representative types of samples. These observations are drawn from terrestrial systems, small solar system bodies, the Moon, and Mars and are curated to represent biogenic, abiogenic, and mixed (biogenic with non-biogenic material) material contexts relevant to exploration of similar worlds. Data were sourced from public repositories, peer-reviewed studies, and measurements at NASA's Ames Research Center (ARC) and Jet Propulsion Laboratory (JPL), with an emphasis on ensuring proper labeling and standardization. This dataset forms the foundation for a supervised binary classifier that assesses the potential biogenicity of samples based on their compositional signatures. Initial analyses demonstrate the effectiveness of this dataset to train a classifier and highlight the value of systematic data curation with multiple data types to enhance classifier performance.

Dataset Development and Organization: To sufficiently represent the diverse set of data that could be generated from a multi-instrument payload in future life detection missions, we selected four data types to provide information about a sample: elemental abundance (from X-ray diffraction and laser induced breakdown spectroscopy), stable isotopic fractionation (from isotopic water analysis, isotope ratio mass spectrometry, and secondary ion mass spectrometry), Raman spectroscopy, and reflectance spectroscopy. These four data types were selected because they have extensive coverage for both biogenic and abiogenic materials in the literature and because they represent plausible near-term planetary exploration techniques.

Data Type	Scope	# Observations
Elemental	>300 ppm	169
Isotopic	$\delta^{13}C$, $\delta^{18}O$, $\delta^{2}H$	219
Raman	200–1800 cm^{-1}	453
Reflectance	400-2100 nm	1,135
	Total Observations	1,975

Table 1: List of observations for each data type and instrumentation scope.

We systematically organized the dataset by populating individual observations for each of the four data types, representative sample types, and classes. For each of the 21 sample types, each data type record consists of a raw data file (comma-separated values) with one or more observations, a metadata text file (Markdown), and a JupyterLab IPython notebook for data standardization and summarization, resulting in 12 files per sample type. This structure, which makes use of open and machine-readable files and tools, facilitates increased accessibility and analysis of the dataset, ensuring that each sample's compiled data is systematically categorized and stored across different file formats for public and internal usage.

Data Features: The elemental features used for the representative systems are H, Ti, Cl, C, P, Mn, Si, Al, Fe, S, Ca, Mg, K, Na, and O, after removing elements without data for at least 30% of observations. A lower threshold (equivalent to a common limit of detection) of 300 ppm was imposed across all measurements to mitigate possible data source bias. These features cover the essential building blocks of Earth-like biochemistry (C, H, N, O, P, S), important trace elements (Fe, Ca, Mg, Na, K), and many of the major components of non-biogenic materials (Si, Al, Ti, Mg).

The stable isotopic fractionation data features are $\delta^{13}C_{VPDB}$, $\delta^{18}O_{VSMOW}$, and $\delta^{2}H_{VSMOW}$. Like the elemental features, these were primarily chosen based on sufficient data availability in the literature for both biogenic and abiogenic materials. Analyses were run both with and without $\delta^{2}H$, as the presence of data for this isotope is strongly related to whether the material contains liquid water. A lower threshold of 0.05 per mille across all measurements was imposed.

Raw Raman spectra (514, 532 nm excitation) were truncated to a range of 200–1800 cm^{-1}, scaled, and smoothed using a Hampel filter for outlier removal followed by Savitzky-Golay filtering. The wavenumber range was chosen based on data availability; the exclusion of the >3000 cm^{-1} range

frequently used for mineralogical measurements is due to the lack of comparable measurements for many biological materials. Basic summary parameters, including the highest value, lowest value, value range, mean value, maximum and minimum first and second derivatives, number of peaks, number of troughs, broadest peak, mean peak width, highest peak, and highest peak wavenumber, were extracted from the data using tools from the open SciPy and scikit-learn libraries to serve as features for the Raman spectroscopy data. These features are a preliminary set intended to be relatively agnostic to assumptions about the type of material or potential life under observation.

Raw reflectance spectra were similarly processed, though without scaling. The common wavelength range imposed was 400-2100 nm, corresponding to common visible and near-IR (VNIR) instruments. The reflectance spectroscopy data features selected were the highest value, lowest value, value range, mean intensity value, maximum and minimum first and second derivatives, number of peaks, number of troughs, broadest peak, mean peak width, highest peak, lowest trough, broadest trough, and mean trough width. As with the Raman data, this feature set is preliminary.

Classifier Analysis and Results: To evaluate the ability of the classifier to distinguish between biogenic and abiogenic samples, we employed both supervised and unsupervised classification algorithms. The classifier performance results were assessed across the four data types to identify the relative and combined importance of each data type and individual feature in determining whether a potential biosignature is an indication of life.

Summary parameters were used to train several supervised machine learning models, including *k*-nearest neighbors (KNN), logistic regression with L2 regularization (LR), Random Forest (RF), support vector machines (SVM-PC), and Gaussian Naïve Bayes (GNB). Additionally, Principal Component Analysis (PCA) was employed as an unsupervised learning method to complement insights from the supervised models. Classification performance was assessed using Monte Carlo simulations with 2,000 iterations of 50% train-test splits. A voting classifier combining all five models was further evaluated by selectively removing each data type at a time to analyze its impact. To analyze the impact of data features, N-1 testing was conducted, systematically removing individual features to measure their contributions to classification accuracy.

The Receiver Operating Characteristic Area Under the Curve (ROC AUC) of the combined voting classifier plots the true positive rate against the false positive rate across various classification thresholds. According to the ROC AUC results, the individual features with the highest contribution to accurate biosignature classification are H, Fe, Mg, Ti, C, P, and max second derivative (reflectance).

When considering each data type as a whole, elemental abundance data demonstrated the highest predictive power (AUC=0.89) among the data types included. However, composite biosignatures, consisting of features from multiple measurement types, proved more beneficial in certain contexts – particularly mixed sample types. In particular, the addition of isotopic fractionation data to other data types frequently resulted in a strongly complementary feature set with enhanced overall classifier accuracy, despite the individual performance of isotopic data being below the mean (AUC=0.80). This increase in combined classification accuracy indicates that the classifier is able to make use of information within the higher dimensional data.

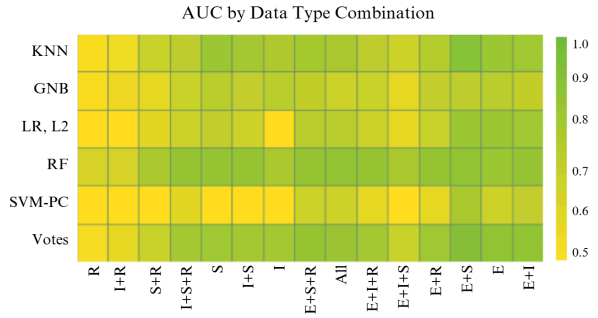


Figure 1: AUC values by combination of elemental (E), isotopic (I), Raman (R), and reflectance (S) data types for supervised machine learning models.

Significance and Future Work: This study demonstrates the novel application of machine learning to improve biosignature detection by integrating diverse data types. The compiled dataset and binary classifier developed in this project is foundational in its approach toward advancing the accuracy and reliability of biosignature interpretation and can be applied to support upcoming planetary missions, offering a standardized framework for biosignature analysis and contributing to the design of instrument payloads and data returns optimized for life detection [2].

Future efforts will aim to broaden the dataset by incorporating additional data to improve classifier training and augment its predictive capabilities. The dataset and associated notebook files will be publicized and uploaded to the Open Data Repository.

Acknowledgments: C.B. was funded by the Vermont Space Grant Consortium through the NASA OSTEM Internships program. A.S. received support through the NASA OSTEM Internships program. T.S. received support from the NASA Space Life Sciences Training Program.

References: [1] Grenfell J. L. (2017) doi: 10.1016/j.physrep.2017.08.003. [2] Warren-Rhodes K. et al. (2023) doi:10.1038/s41550-022-01882-x.