1. Introduction and Background

Astrobiology is a relatively new field that has only gained traction in the past couple of decades. Even though research first started in the 1970s with NASA's Viking missions to Mars, this type of research did not gain popularity until much later (Rench, 2025). Due to the young age of astrobiology, not much time has been spent developing machine learning (ML) and artificial intelligence (AI) algorithms for life detection and biosignatures. To date, only 8 papers have been published that discuss the use of ML-AI for this topic (NASA, 2025). Therefore, the development of new ML-AI algorithms is crucial for the advancement of astrobiology. Rovers have been on Mars for decades and provide valuable data to start developing ML-AI algorithms.

Two important astrobiological sites to study are lakebed sediments and hydrothermal vents. These sites on Earth host large microbial communities and provide a high preservation potential for organic matter (Eigenbrode et al., 2018; Freissinet et al., 2015; Michalski et al., 2017; Summons et al., 2011). Similar sites have been found on Mars, with Fe/Mg-smectite rich lacustrine and fluviodeltaic sediments found at lakebeds Gale crater and Oxia Planum, and at putative hydrothermal systems at Eridania Basin (Eigenbrode et al., 2018; Freissinet et al., 2015; Michalski et al., 2017; Quantin-Nataf et al., 2021). Understanding microbial diversity and organic preservation at these sites on Earth could provide valuable information on the potential for life on early Mars, if it ever arose. GC-MS is also the main instrument deployed in astrobiology space missions and has been the most used approach to date for organic matter detection, hence it is a key focus on this project (Scheller et al., 2022; Williams et al., 2021).

The main objective of this proposal is to start developing the ability to link the habitability of Mars' environments to life development preference at geological sites using terrestrial analogs. Initial analysis will use fatty acid methyl esters (FAMEs) detection from gas chromatography mass spectroscopy (GC-MS) and total organic carbon (TOC) data for principal component analysis (PCA), determining geological site preference for life development. Further analysis will use clustering to group life development preference based on geological settings. As a result, we will examine if the algorithm is able to accurately predict life development preference at geological sites. This analysis could be used for current Mars' rover missions, allowing for more specific and focused fatty acid and TOC analysis to occur at sites with known geological settings.

2. Data Information and Legend

The test data used was gathered between three different sources: 1) my own personal research funded by NASA awarded HabMars grant (80NSSC24K0076), 2) Williams A. J. et al. (2019) Astrobiology,19, 522–546., and 3) Williams A. J. et al. (2021) Astrobiology, 21, 60-82. However, while multiple studies are included, this is still a small dataset. Precautions were taken to minimize imbalance or false trends; however, data analyzed is still considered preliminary results until a larger dataset is available for analysis.

The dataset includes the following information:

- Sample name [Sample_name]
- Environment [Environment]
    - 0 = Lake
    - 1 = Fjord
    - 2 = Man-Made
    - 3 = Hot-Spring System
    - 4 = Mountain
    - 5 = Island
- Sample Type [Sample_type]
    - 0 = Lakebed
    - 1 = Active Hydrothermal Vent
    - 2 = Inactive Hydrothermal Vent
    - 3 = Relict Hydrothermal Vent
    - 4 = Gossan
    - 5 = Precipitate
    - 6 = Ooid Sand
    - 7 = Shale
- Apparent dominant ion [Apparent_dominant_ion]
    - 0 = Sulfur
    - 1 = Iron
    - 2 = Magnesium
    - 3 = Unknown
- Total organic carbon (TOC) percentage (milligram [mg] of TOC per 100 mg of TOC) [TOC]
- Does TOC support life development? [TOC_supports_life_development]
    - 0 = No
    - 1 = Yes

    - <0.5% was determined to be the cutoff point for TOC that does not support life development. This is a general proxy and should not be used as a direct comparison for life development. Assimilable organic carbon (AOC) is better for direct comparison of if sediment supports development of life.
- Fatty acid methyl ester (FAME) abundance detection [C10 through C18]
    - Includes data from $C_{10:0}$ to $C_{18:0}$. Anything less than $C_{10:0}$ was not included since those FAMEs do not help indicate recent or active life in the sediment. FAME concentrations are given in picogram (pg) of FAME per mg of sample.
- Sum of FAME abundance from $C_{10:0}$ to $C_{18:0}$ (pg of FAME per mg of sample) [Sum_C10_C18]
- Biomass FAME abundance baseline (pg of FAME per mg of sample) [Biomass_FAME_baseline]
    - The sum of $C_{16:0}$ and $C_{18:0}$ FAME abundances, as these FAMEs provide a proxy for biomass in the samples.
- Bacterial FAME abundance baseline (pg of FAME per mg of sample) [Bacteria_FAME_baseline]
    - The sum of $C_{16:1}$, $C_{18:1}$, iso-$C_{15:0}$, anteiso-$C_{15:0}$, and iso-$C_{17:0}$ FAME abundances, as these FAMEs provide a proxy for bacterial presence in the samples.
- Does biomass FAME abundance baseline support life? [Biomass_supports_life]
        0 = low to no microbial input
        1 = suggests microbial presence
        2 = strong evidence of recent or active microbial life
    - If the total was <10 pg FAME/mg sample, it could be background, ancient, or abiotic origin (0). If it is between10–100 pg FAME/mg sample, it suggests microbial residues, possibly viable (1). If >100 pg FAME/mg sample, this is a strong indicator of recent or active microbial life (2).
- Does bacterial FAME abundance baseline support bacterial community presence? [Bacteria_presence]
        0 = low to no microbial input
        1 = suggests microbial presence
        2 = strong evidence of recent or active microbial life
    - If the total was <10 pg FAME/mg sample, it could be background, ancient, or abiotic origin (0). If it is between10–100 pg FAME/mg sample, it suggests microbial residues, possibly viable (1). If >100 pg FAME/mg sample, this is a strong indicator of recent or active microbial life (2).

3. Results and Discussion

3.1 Geological Setting Variable Grouping

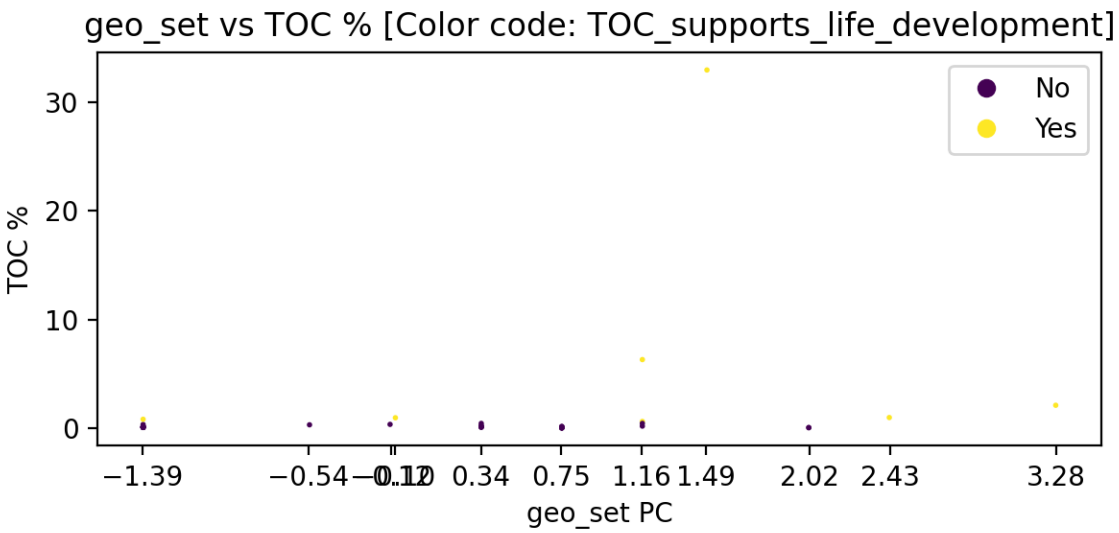
To create a single geological setting feature, the Environment and Sample_type variables were combined with PCA to create “Geological Setting (geo_set)”. This allowed for easier comparison of geological settings to other variables, such as TOC. Apparent_dominant_ion was intended to be grouped into the geo_set variable, but not enough information for each sample was available. This is something that would be interesting to include in future analysis, once more data is available.

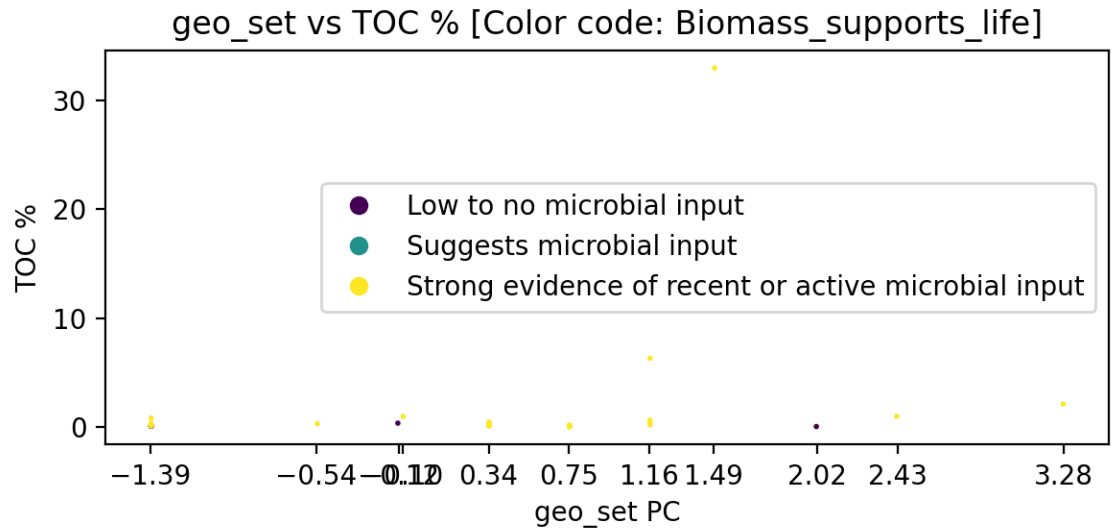
Environment	Sample_type	geo_set
Lake	Lakebed	-1.385524
Fjord	Inactive Hydrothermal Vent	-0.123878
Fjord	Active Hydrothermal Vent	-0.535365
Man-Made	Active Hydrothermal Vent	-0.096693
Hot-Spring System	Active Hydrothermal Vent	0.341978
Hot-Spring System	Inactive Hydrothermal Vent	0.753465
Hot-Spring System	Relict Hydrothermal Vent	1.164952
Lake	Shale	1.494886
Mountain	Gossan	2.015111
Mountain	Precipitate	2.426598
Island	Ooid Sand	3.276757

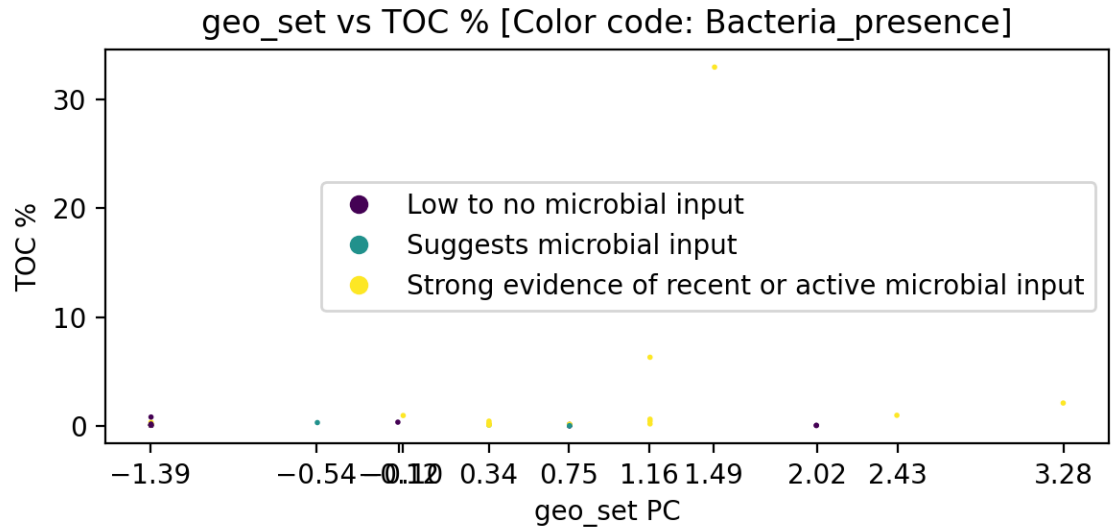
3.2 Geological Setting Comparison to Variables:

3.2.1 TOC Percentage

The geo_set variable was compared to TOC percentage with the following color coding to determine if any trends were found: TOC_supports_life_development, Biomass_supports_life, and Bacteria_presence.

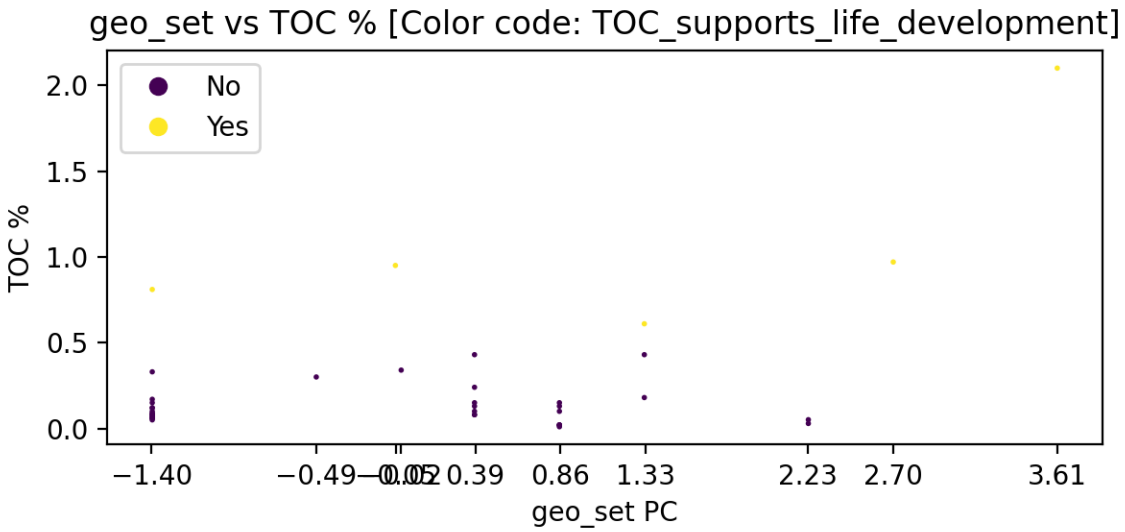

geo_set vs TOC % [Color code: TOC_supports_life_development]
No
Yes
TOC %
0
10
20
30
−1.39
−0.54
0.34
0.75
1.16
1.49
2.02
2.43
3.28
geo_set PC

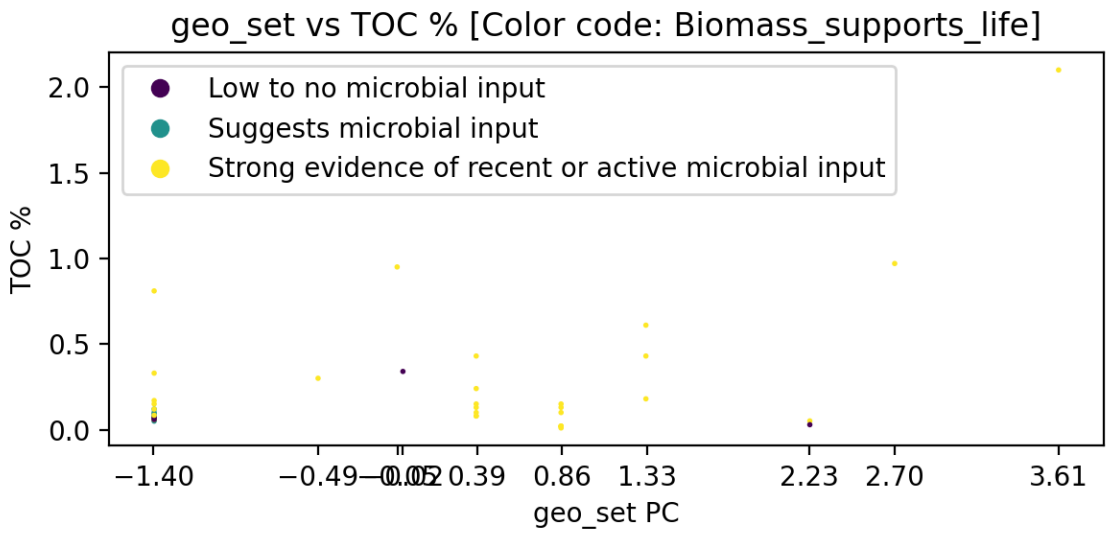

geo_set vs TOC % [Color code: Biomass_supports_life]
Low to no microbial input
Suggests microbial input
Strong evidence of recent or active microbial input
TOC %
0
10
20
30
−1.39
−0.54
0.34
0.75
1.16
1.49
2.02
2.43
3.28
geo_set PC

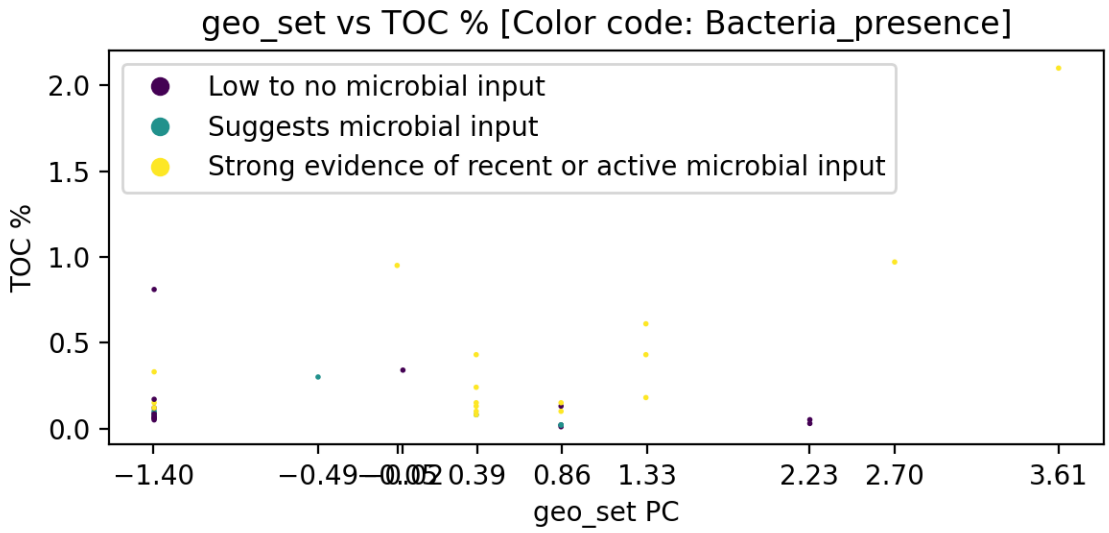

geo_set vs TOC % [Color code: Bacteria_presence]
Low to no microbial input
Suggests microbial input
Strong evidence of recent or active microbial input
TOC %
0
10
20
30
−1.39
−0.54
0.34
0.75
1.16
1.49
2.02
2.43
3.28
geo_set PC

As seen in the images, there is difficulty finding trends with the variables. However, two samples with 33% TOC and 6% TOC are causing a large spread in the charts. Those samples were removed to help determine if there were any finer trends that could be found within the other samples. Removal of the samples required a new normalization and PCA analysis.

Environment	Sample_type	geo_set
Lake	Lakebed	-1.396262
Fjord	Inactive Hydrothermal Vent	-0.019579
Fjord	Active Hydrothermal Vent	-0.489139
Man-Made	Active Hydrothermal Vent	-0.051577
Hot-Spring System	Active Hydrothermal Vent	0.385986
Hot-Spring System	Inactive Hydrothermal Vent	0.855546
Hot-Spring System	Relict Hydrothermal Vent	1.325106
Mountain	Gossan	2.232228
Mountain	Precipitate	2.701788
Island	Ooid Sand	3.608911

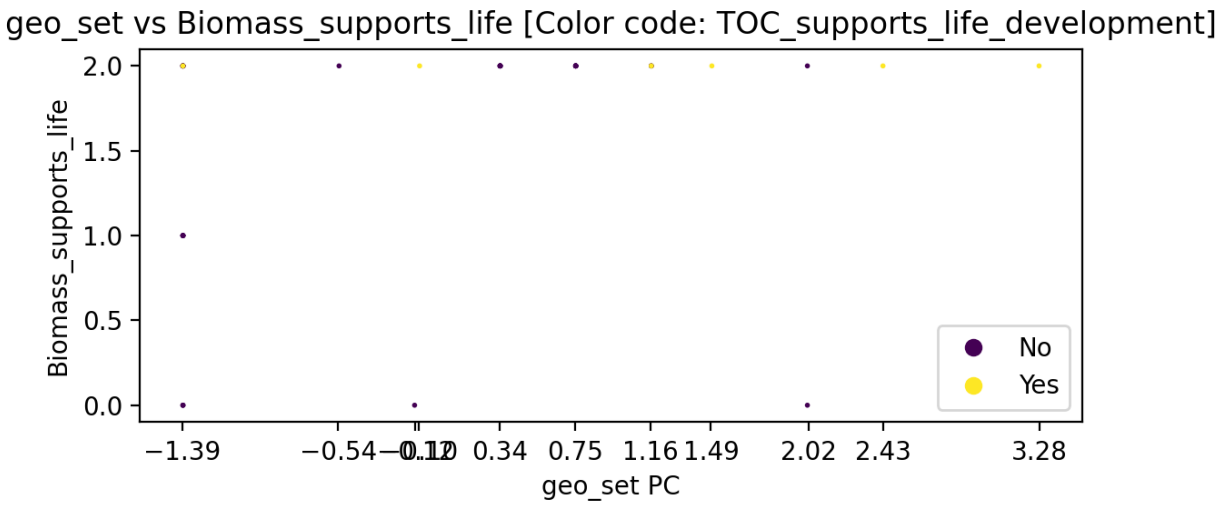


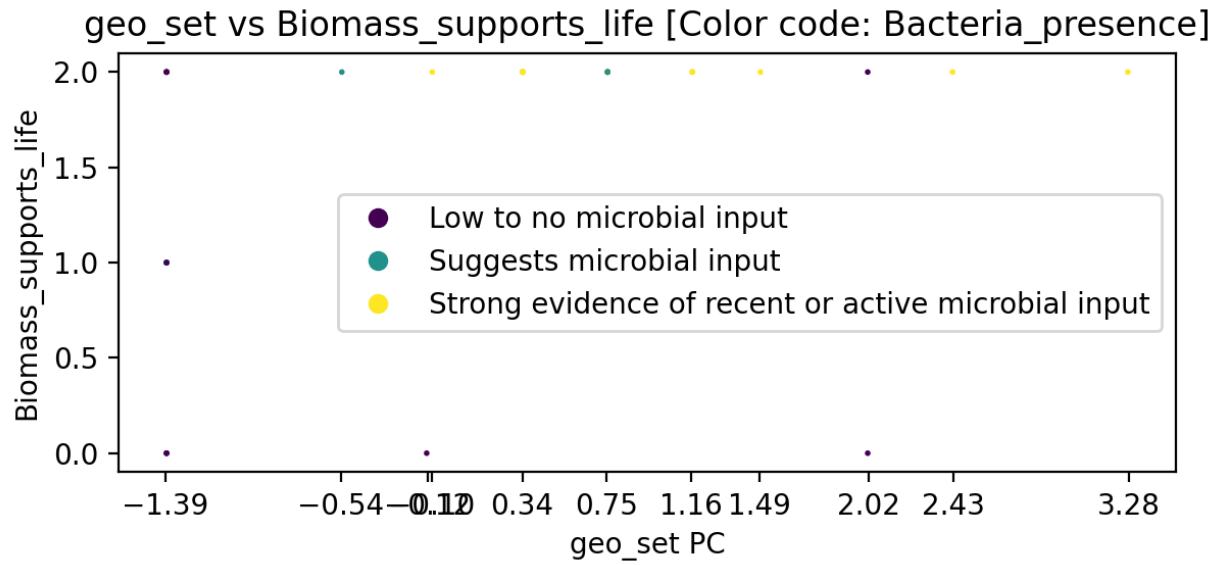




3.2.2 Biomass Supports Life Development

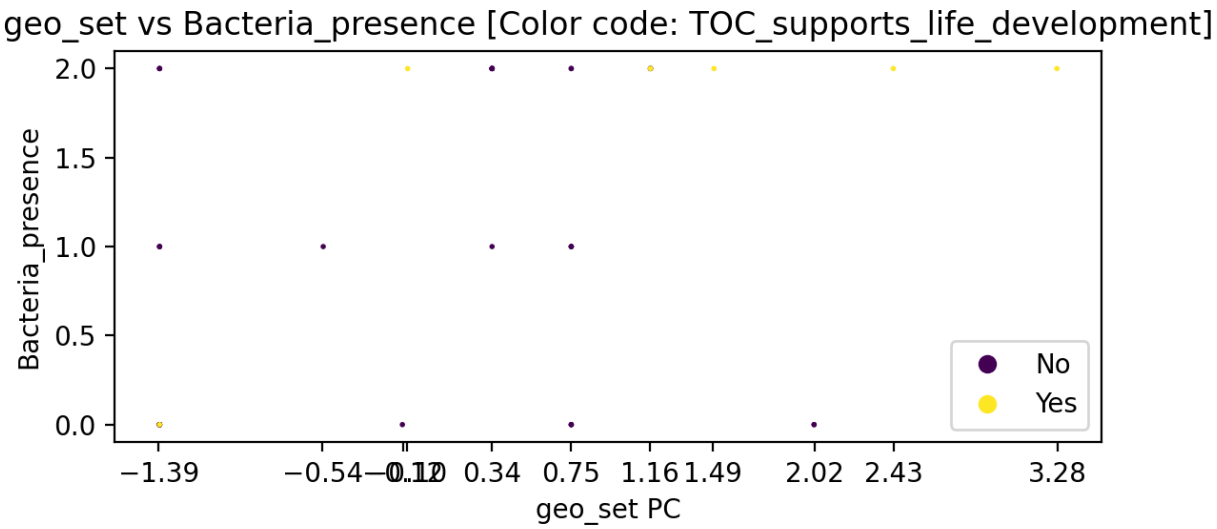
The geo_set variable was compared to Biomass_supports_life with the following color coding to determine if any trends were found: TOC_supports_life_development and Bacteria_presence.

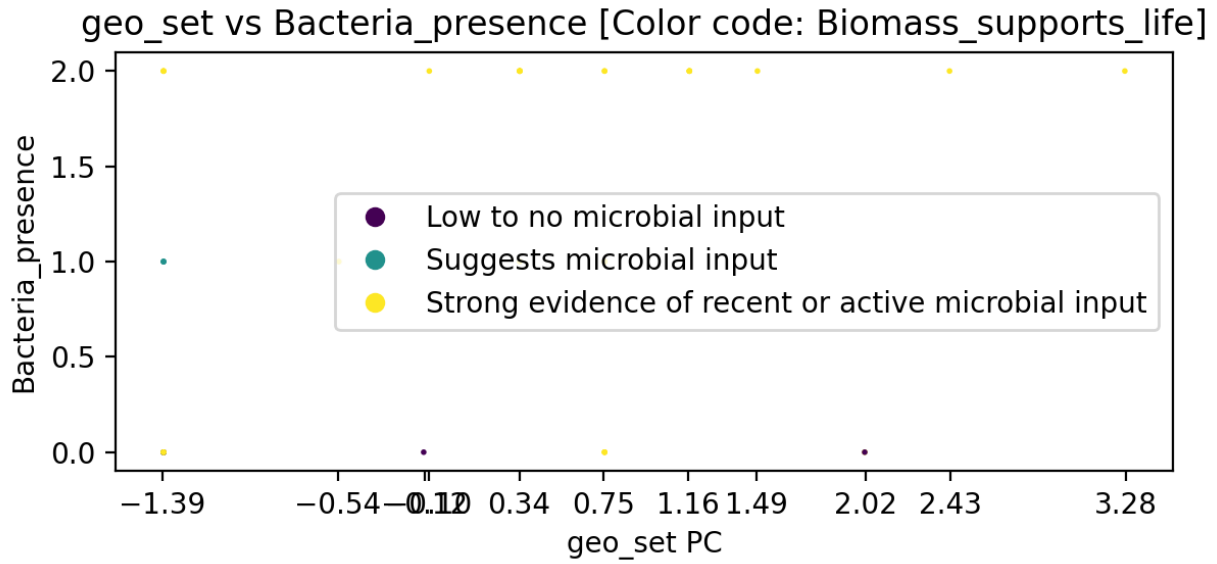




3.2.3 Bacteria Presence

The geo_set variable was compared to Bacteria_presence with the following color coding to determine if any trends were found: TOC_supports_life_development and Biomass_supports_life.





3.2.4 Trends Found

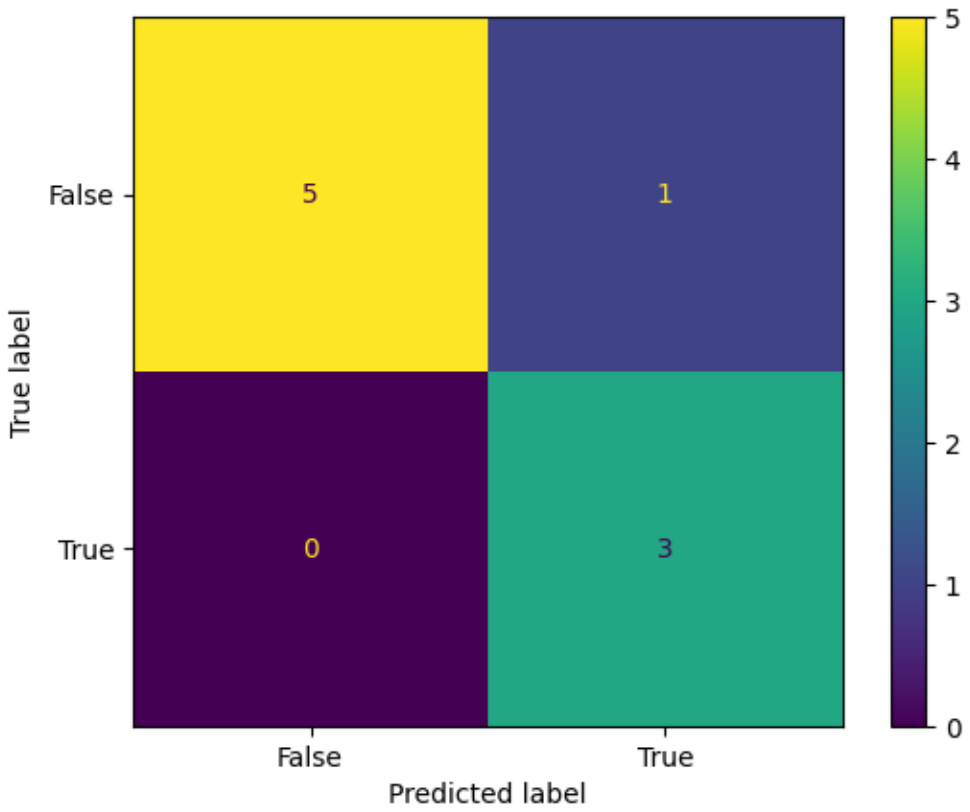
1. TOC is not an accurate provider of if life can develop in a system, at least not set at the cutoff point that is implemented in this study. In multiple samples, TOC did not support life, but the FAME abundances supported biomass and bacteria presence.
2. Biomass having the ability to support life development does not mean that bacterial growth is occurring at that site.

3.3 Logistic Regression

3.3.1 TOC Supports Life Development

Although TOC proved to not be an accurate variable to determine if life can develop in a system at this point, a new dataframe was prepared to complete a logistic regression to determine what variables influence if TOC supports life development. The following variables were used to compare: Environment, Sample_type, Biomass_supports_life, and Baceria_presence.

First the dataframe was cleaned to drop any samples did had NaN. The cleaned dataframe then dropped TOC_supports_life_development to allow normalization and PCA analysis on the remaining variables. A total of 4 PCs were created, with 3 PCs making up greater than 90% of the variance (91.73%). TOC class imbalance was also checked, with 7 data points supporting life development, and 37 not supporting life development. Since there was a heavy class imbalance and small dataset, the TOC features included all PCs and was scaled, with the logistic regression function including class balance and L2 regularization (strength of 0.01). In addition, the threshold for logistic regression was lowered to 0.3 to help increase the weight of TOC that supports life. Ideal regularization strength was checked with the GridSearchCV function. The data was split for 20.5% testing and 79.5% training, resulting in a cross-validated accuracy score of 88.06% ± 13.17%. There were no clear signs of overfitting after adding the L2 regularization, class balance, and lowered threshold.



The logistic regression method proved to show fairly accurate results, even with a small dataset. This could be useful in the future, when more data points are available, seeing if TOC can be used as a proxy to determine if life can develop in a system.

A correlation matrix showed that Sample_type had the strongest correlation to if TOC supported life development. Environment and Bacteria_presence held the next two highest scores, while Biomass_supports_life typically had the lowest scores.

	Environment	Sample_type	Biomass_supports_life	Bacteria_presence	TOC Life Development Prediction
Environment	1.000000	0.644562	0.386722	0.401868	0.743461
Sample_type	0.644562	1.000000	0.229866	0.326478	0.807059
Biomass_supports_life	0.386722	0.229866	1.000000	0.509920	0.690686

	Environment	Sample_type	Biomass_supports_life	Bacteria_presence	TOC Life Development Prediction
Bacteria_presence	0.401868	0.326478	0.509920	1.000000	0.715297
TOC Life Development Prediction	0.743461	0.807059	0.690686	0.715297	1.000000

3.3.2 Biomass Supports Life Development and Bacteria Presence

These two variables were not run with a logistic regression since the data was not 0s and 1s. There was a higher level of complexity to these variables, requiring 0s, 1s, and 2s. However, these variables did undergo KMeans clustering and can be found in sections 3.4.

3.4 KMeans Clustering

3.4.1 Data Processing

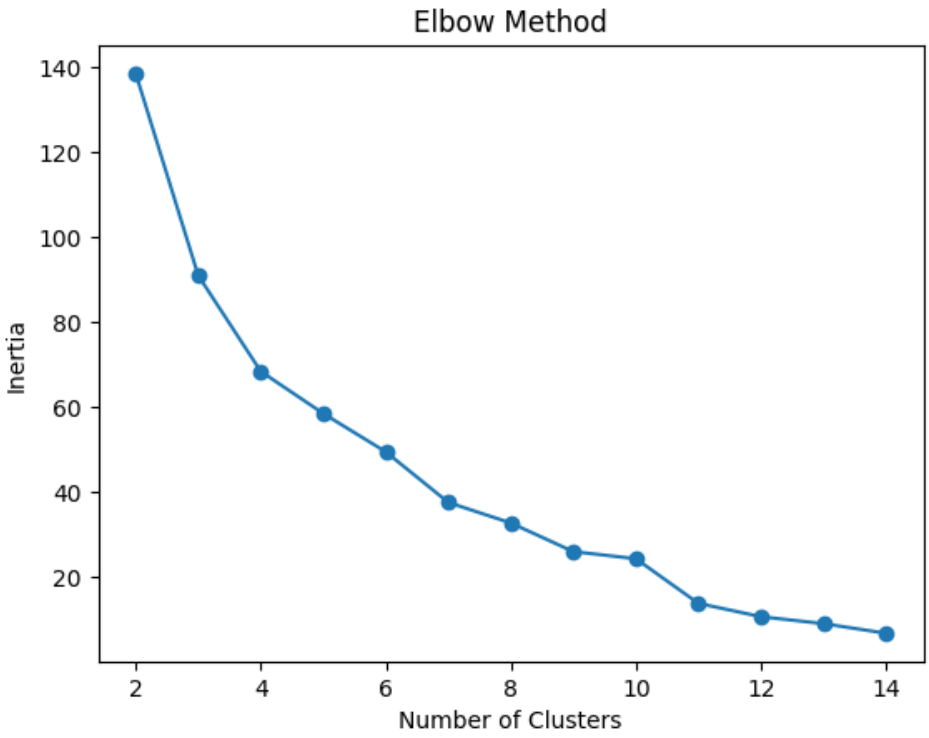
A new dataframe was prepared to complete KMeans clustering to determine how life develop preference groups based off geological settings. The following variables were used for analysis: Environment, Sample_type, Biomass_supports_life, and Baceria_presence.

The dataframe was copied from the TOC logistic regression analysis. The dataframe underwent normalization and PCA analysis. A total of 5 PCs were created, with 4 PCs making up greater than 90% of the variance (95.71%). The ideal n_clusters value was determined through:

1. Plotting silhouette scores vs. n_clusters

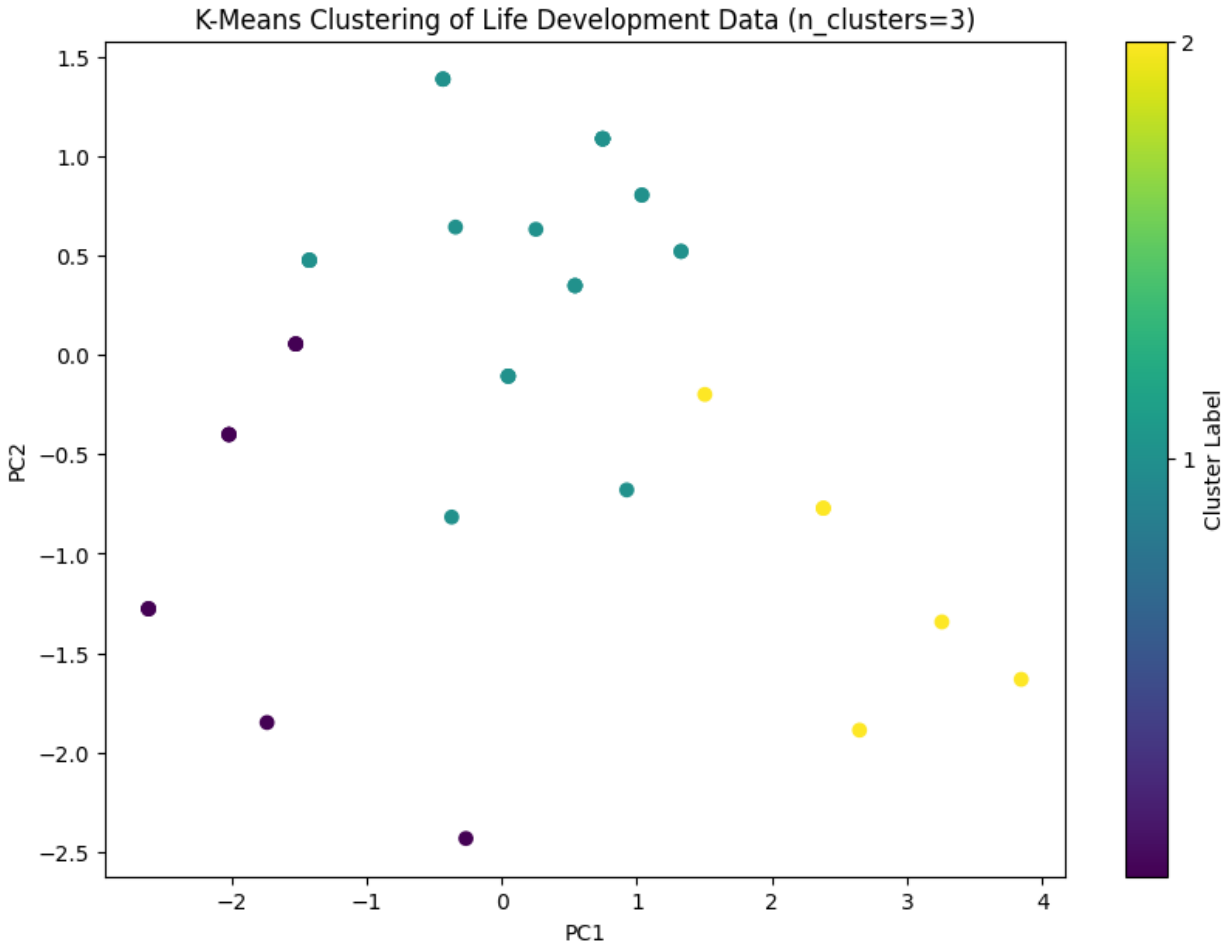
This silhouette scores vs. n_clusters showed that anything after n=6 diminishes the cluster analysis value. You can see a sharp increase after n=6 with a continual increase in silhouette score with the range that was tested.

2. The elbow method

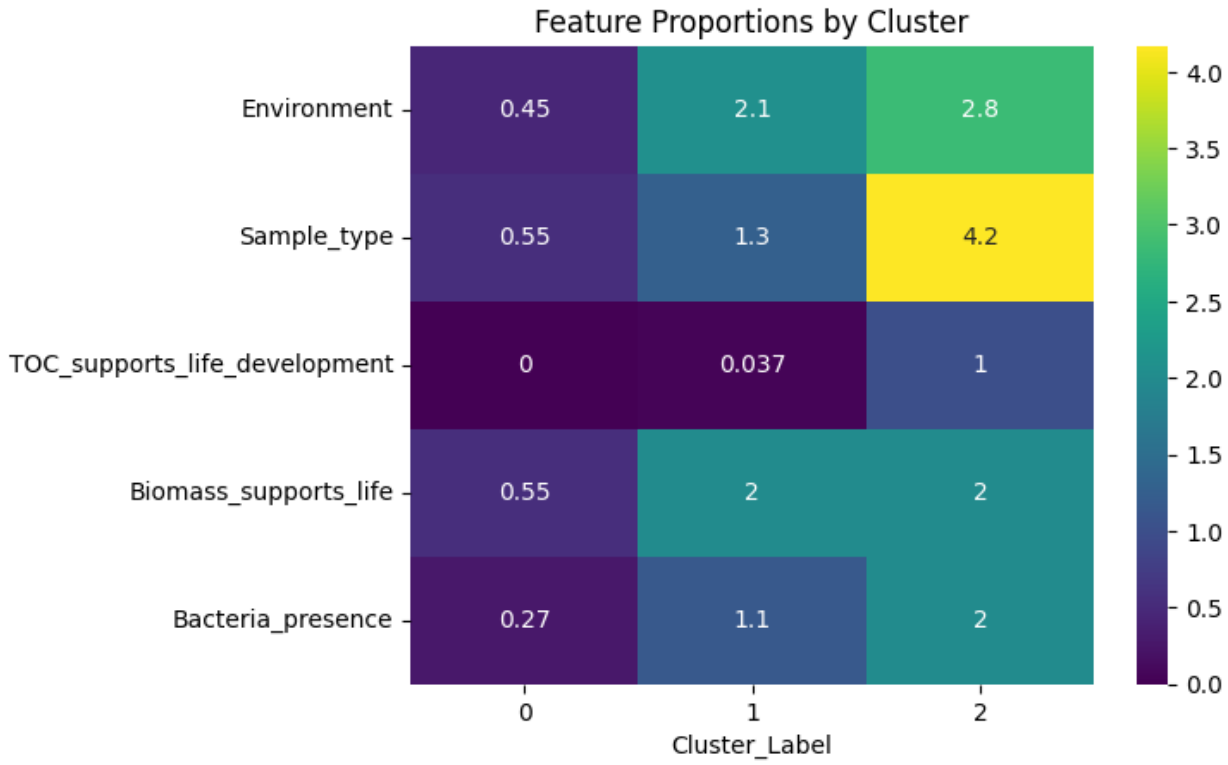


The “elbow” in the plot is where the line sharpy decreases then levels off. The elbow represents the best trade-off between model complexity and explained variance. As with the silhouette scores vs. n_clusters plot, further n_clusters beyond the “elbow” diminishes the cluster analysis value. In this plot, the elbow was around 4 to 5 n_clusters.

Since the dataset is small, it is prone to over-clustering, which is why the silhouette scores vs. n_clusters and elbow methods were required to determine the best n_clusters. It was determined that n=3 would prove the best clustering analysis with this dataset size. KMeans clustering resulted in the following:

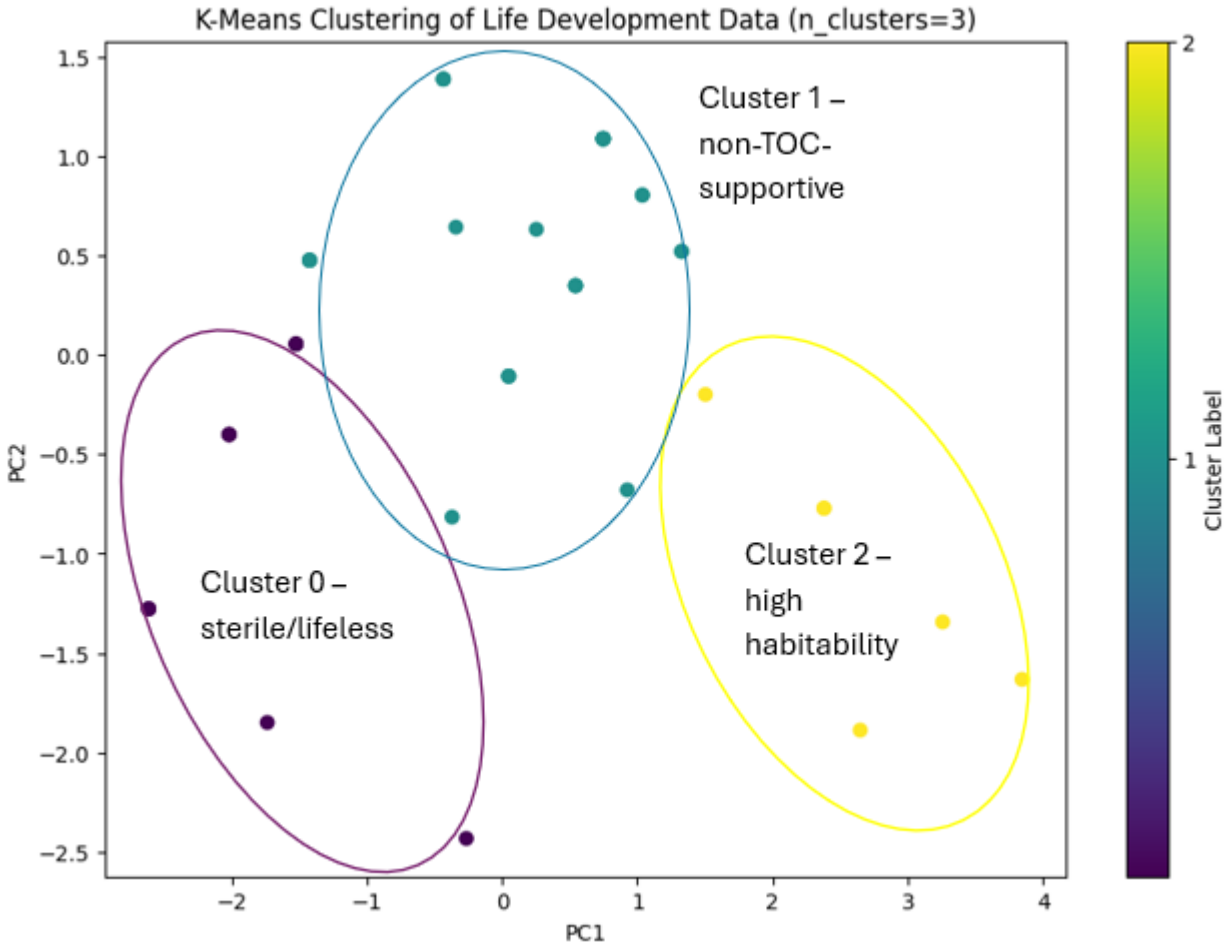


You can see three somewhat distinctive clusters. A cluster summary was created to help make the plot more interpretable. The distributions of each key feature were compared to each cluster. A heatmap was used for better visualization.

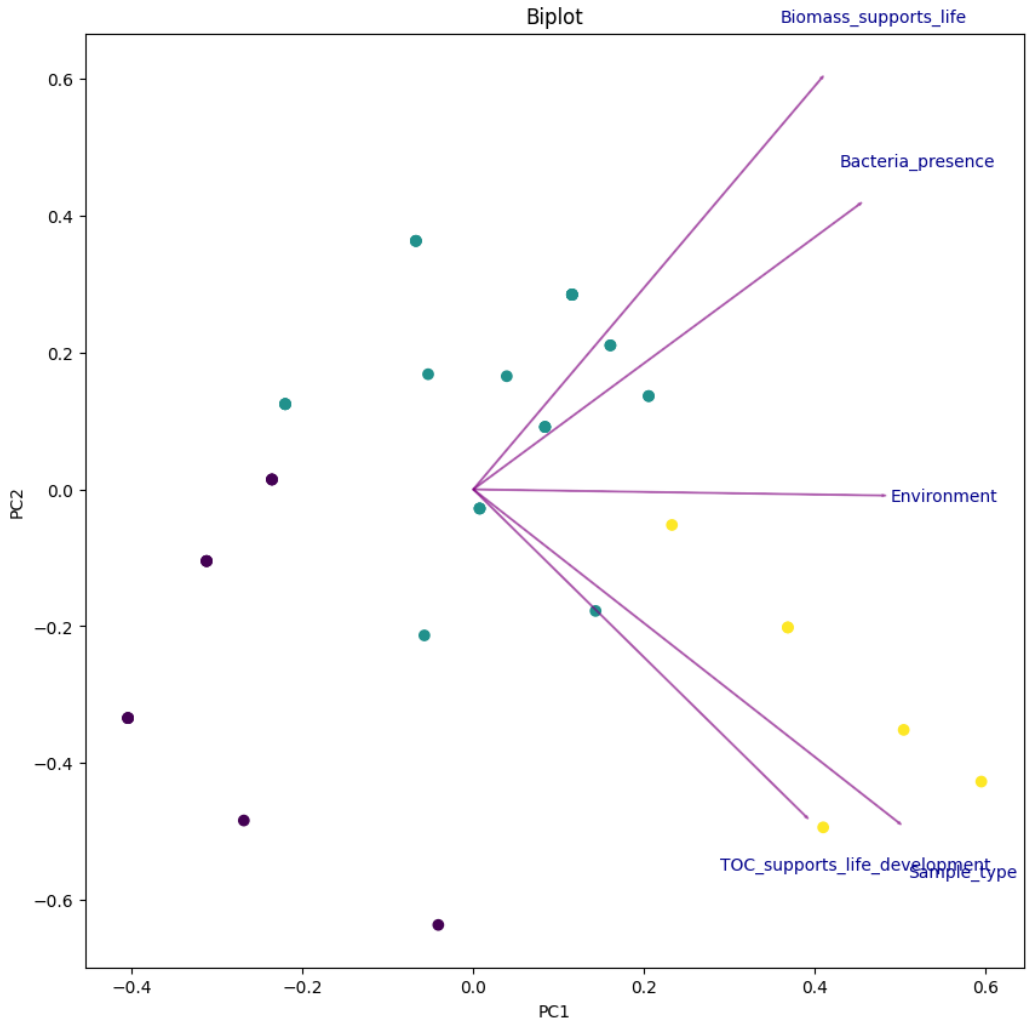


Based on the data, the following information was interpreted:

- Cluster 0
  - This is a sterile or lifeless cluster. Could represent harsh environments or negative samples.
- Cluster 1
  - This is non-TOC-supportive environments, where the environment and sample type influence the biomass FAME abundance and if bacteria are present.
- Cluster 2
  - High priority zones for habitability. High numbers across the board indicate that these environments are strongly linked to life being present.



To see a better visualization of how each variable was contributing to each cluster, a biplot was created:



The following can be analyzed from the biplot:

- Variables most aligned with PC1:
  - Environment
- Variables most aligned with PC2:
  - Biomass_supports_life
  - TOC_supports_life_development
- Variables equally aligned with PC1 and PC2:
  - Bacteria_presence
  - Sample_type
- Variables that are orthogonal to each other:
  - TOC_supports_life_development and Biomass_supports_life
- Variables that have similar loadings:
  - Biomass_supports_life and Bacteria_presence
  - TOC_supports_life_development and Sample_type

3.4.2 Trends

1. There is potential to cluster this data to determine life development preference based on geological settings. Initial cluster trends were analyzed, with three distinctive clusters being generated (lifeless, non-TOC-supportive, and highly habitable).
2. TOC_supports_life_development and Biomass_supports_life do not influence each other, while Biomass_supports_life/Bacteria_presence and TOC_supports_life_development/Sample_type potentially influence each other.
3. Environment is directly influenced by PC1 and influences all other variables to some degree.

4. Citations

1. Eigenbrode, J.L. et al., 2018, Organic matter preserved in 3-billion-year-old mudstones at Gale crater, Mars: Science, v. 360, p. 1096–1101, doi:10.1126/science.aas9185.

2. Freissinet, C. et al., 2015, Organic molecules in the Sheepbed Mudstone, Gale Crater, Mars: Journal of Geophysical Research: Planets, v. 120, p. 495–514, doi:10.1002/2014JE004737.

3. Kivrak, L., Williams, A.J., Buch, A., and He, Y., 2021, TRIMETHYLSULFONIUM HYDROXIDE (TMSH) THERMOCHEMOLYSIS WITH PY-GC-MS AS A METHOD OF ORGANIC BIOSIGNATURE DETECTION: OPTIMIZATION FOR NUCLEOBASE.

4. Michalski, J.R., Dobrea, E.Z.N., Niles, P.B., and Cuadros, J., 2017, Ancient hydrothermal seafloor deposits in Eridania basin on Mars: Nature Communications, v. 8, p. 15978, doi:10.1038/ncomms15978.

5. National Aeronautics and Space Administration (NASA), 2025, AI Astrobiology Life Detection & Biosignatures: https://www.nasa.gov/a-i-astrobiology-life-detection-biosignatures/ (accessed February 2025)

6. O'Reilly, S.S., Mariotti, G., Winter, A.R., Newman, S.A., Matys, E.D., McDermott, F., Pruss, S.B., Bosak, T., Summons, R.E., and Klepac-Ceraj, V., 2017, Molecular biosignatures reveal common benthic microbial sources of organic matter in ooids and grapestones from Pigeon Cay, The Bahamas: Geobiology, v. 15, p. 112–130, doi:10.1111/gbi.12196.

7. Quantin-Nataf, C. et al., 2021, Oxia Planum: The Landing Site for the ExoMars "Rosalind Franklin" Rover Mission: Geological Context and Prelanding Interpretation: Astrobiology, v. 21, p. 345–366, doi:10.1089/ast.2019.2191.
8. Rench, B.M., 2025, Astrobiology Program FAQ: https://astrobiology.nasa.gov/about/faq/ (accessed February 2025)