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. Terrestrial sites 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 biosignature detection and organic preservation at these sites on Earth could provide valuable information on the potential for life on early Mars, if it ever arose.

The main objective of this ML 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 biosignature 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. To date, GC-MS have been the main method for detecting organic matter on astrobiology space missions, hence it is a key focus on this project (Scheller et al., 2022; Williams et al., 2021). Further analysis will use KMeans clustering to group life development preference based on geological settings. As a result, we will examine if the ML 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) personal research funded by NASA sub 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. Since the dataset is small, results can vary when running the code. This paper tried to show the average results that were obtained.

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 (gram [g] of TOC per 100 g 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 organic carbon content supports development of life.
- Fatty acid methyl ester (FAME) abundance detection [$C_{10:0}$ through $C_{18:0}$]
    - Includes individual FAME 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 sediment 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
        If the total was <10 pg FAME/mg sample, it could be background, ancient, or abiotic origin.
    1 = suggests microbial presence
        If it is between10–100 pg FAME/mg sample, it suggests microbial residues, possibly viable.
    2 = strong evidence of recent or active microbial life
        If >100 pg FAME/mg sample, this is a strong indicator of recent or active microbial life.
- Does bacterial FAME abundance baseline support bacterial community presence? [Bacteria_presence]
    0 = low to no microbial input
        If the total was <10 pg FAME/mg sample, it could be background, ancient, or abiotic origin.
    1 = suggests microbial presence
        If it is between10–100 pg FAME/mg sample, it suggests microbial residues, possibly viable.

2 = strong evidence of recent or active microbial life

If >100 pg FAME/mg sample, this is a strong indicator of recent or active microbial life.

3. Results and Discussion

3.1 Geological Setting Variable Grouping

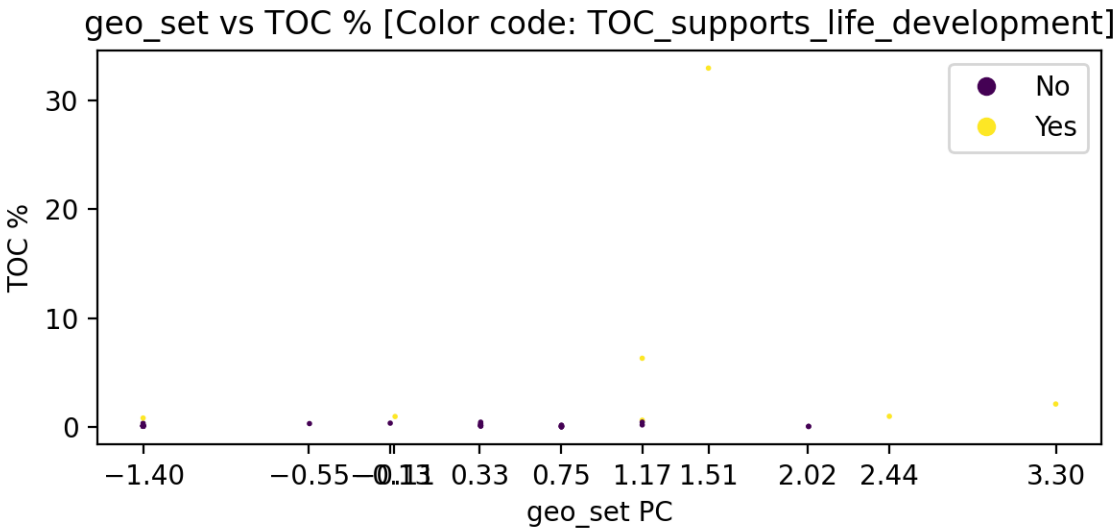
To create a single geological setting feature, the Environment and Sample_type variables were combined with PCA to create "Geological Setting (geo_set)". This allowed for easier comparison of geological settings to other variables, such as TOC. Apparent_dominant_ion was intended to be grouped into the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

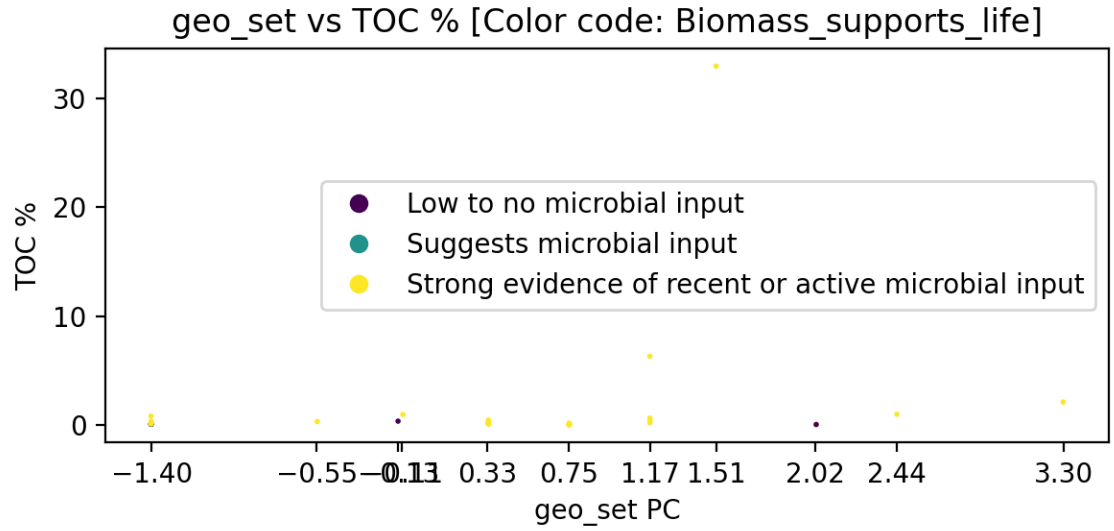
Environment	Sample_type	geo_set
Lake	Lakebed	-1.403482
Fjord	Inactive Hydrothermal Vent	-0.130875
Fjord	Active Hydrothermal Vent	-0.546755
Man-Made	Active Hydrothermal Vent	-0.105908
Hot-Spring System	Active Hydrothermal Vent	0.334939
Hot-Spring System	Inactive Hydrothermal Vent	0.750819
Hot-Spring System	Relict Hydrothermal Vent	1.166699
Lake	Shale	1.507676
Mountain	Gossan	2.023426
Mountain	Precipitate	2.439306
Island	Ooid Sand	3.296033

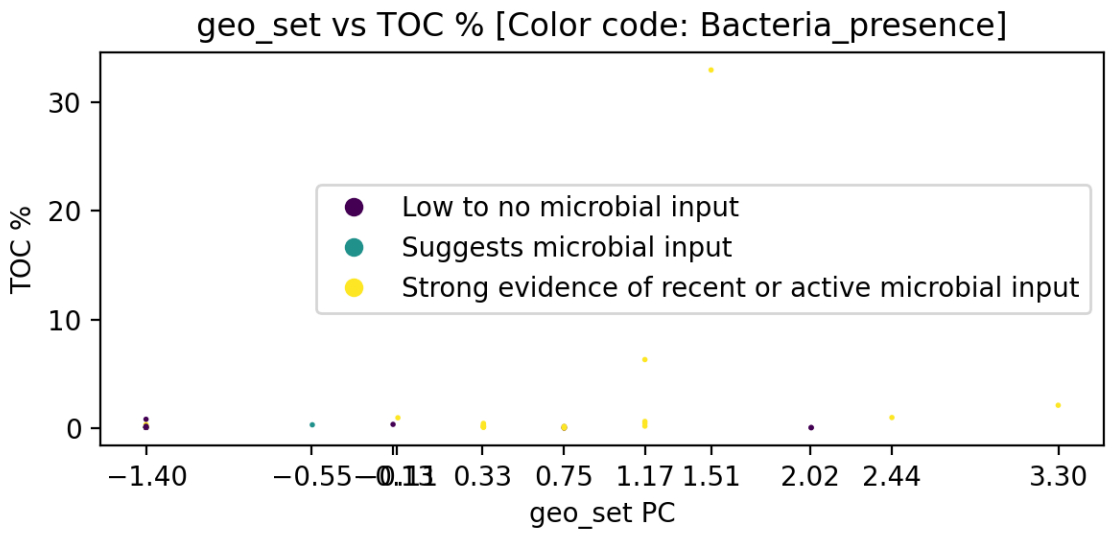
3.2 Geological Setting Comparison to Variables:

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.

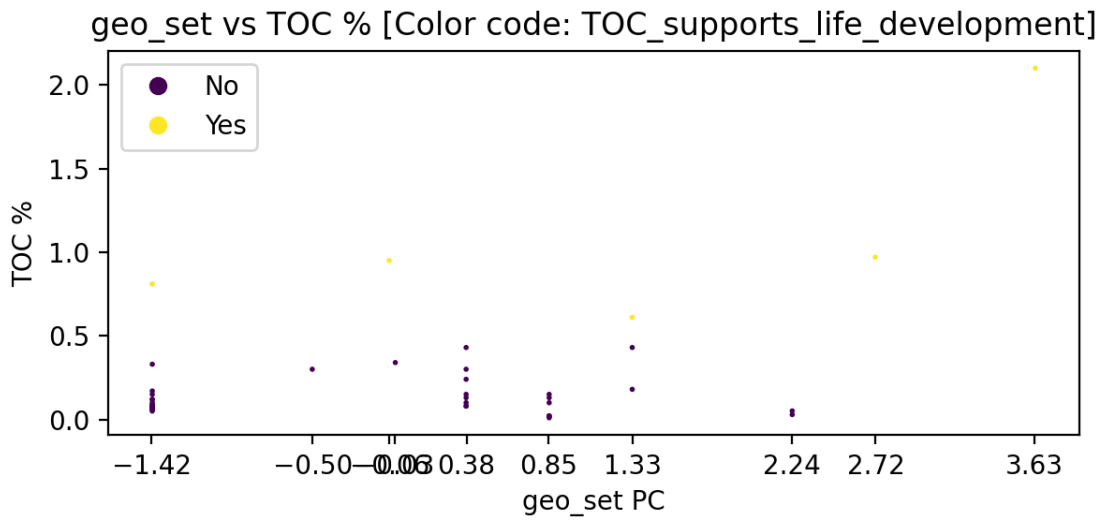


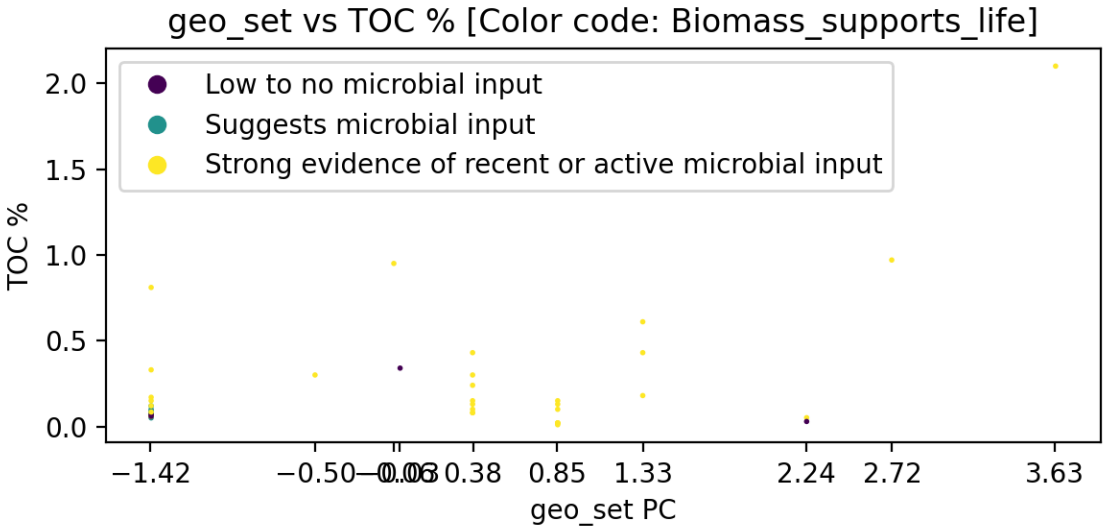


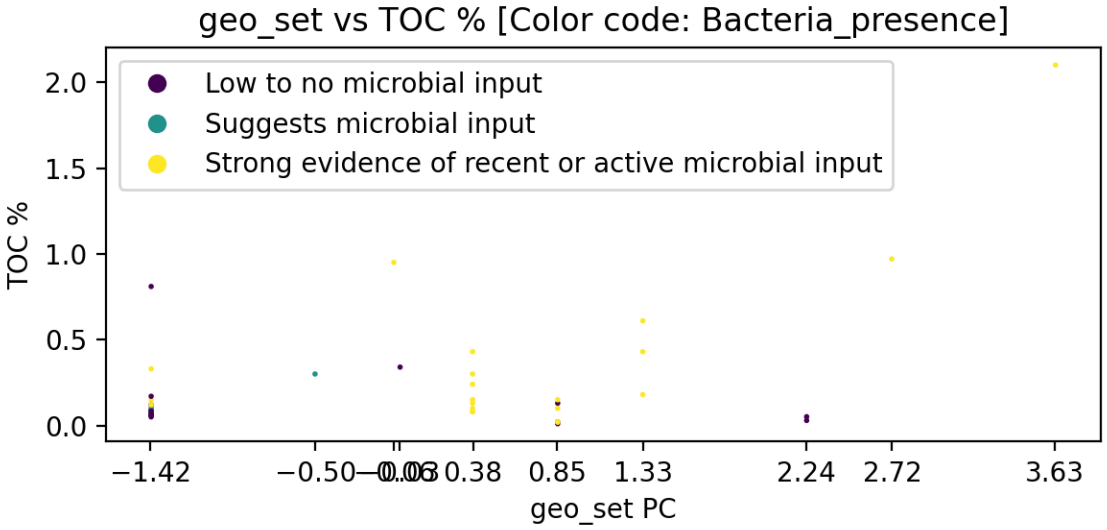


As seen in the images, there is difficulty finding trends with the variables. However, two samples with 33% TOC and 6% TOC are causing a large spread in the charts. Those samples were removed to help determine if there were any finer trends that could be found within the other samples. Removal of the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

Environment	Sample_type	geo_set
Lake	Lakebed	-1.416612
Fjord	Inactive Hydrothermal Vent	-0.026668
Fjord	Active Hydrothermal Vent	-0.501681
Man-Made	Active Hydrothermal Vent	-0.061763
Hot-Spring System	Active Hydrothermal Vent	0.378154
Hot-Spring System	Inactive Hydrothermal Vent	0.853167
Hot-Spring System	Relict Hydrothermal Vent	1.328180
Lake	Shale	2.243111
Mountain	Gossan	2.718123
Mountain	Precipitate	3.633054

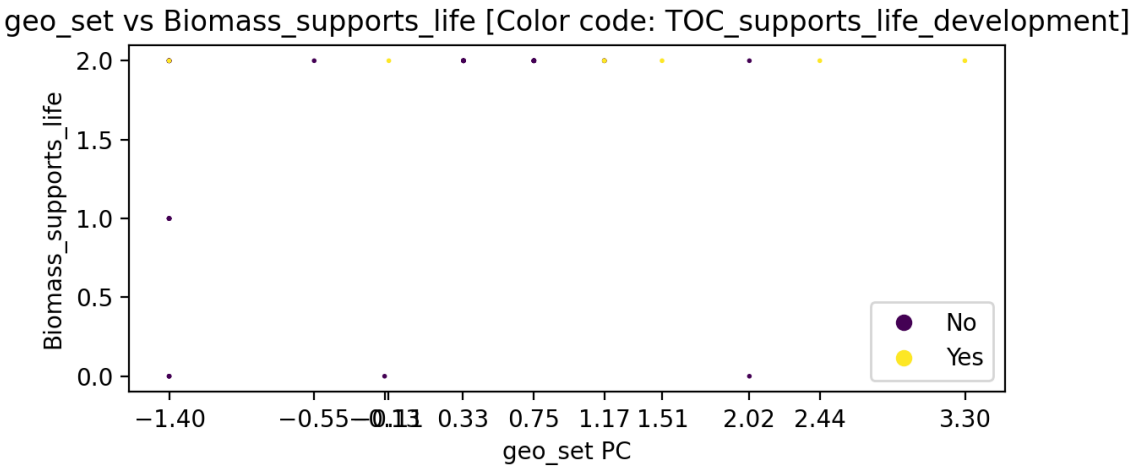

geo_set vs TOC % [Color code: TOC_supports_life_development]
No
Yes
TOC %
2.0
1.5
1.0
0.5
0.0
−1.42
−0.50
0.38
0.85
1.33
2.24
2.72
3.63
geo_set PC

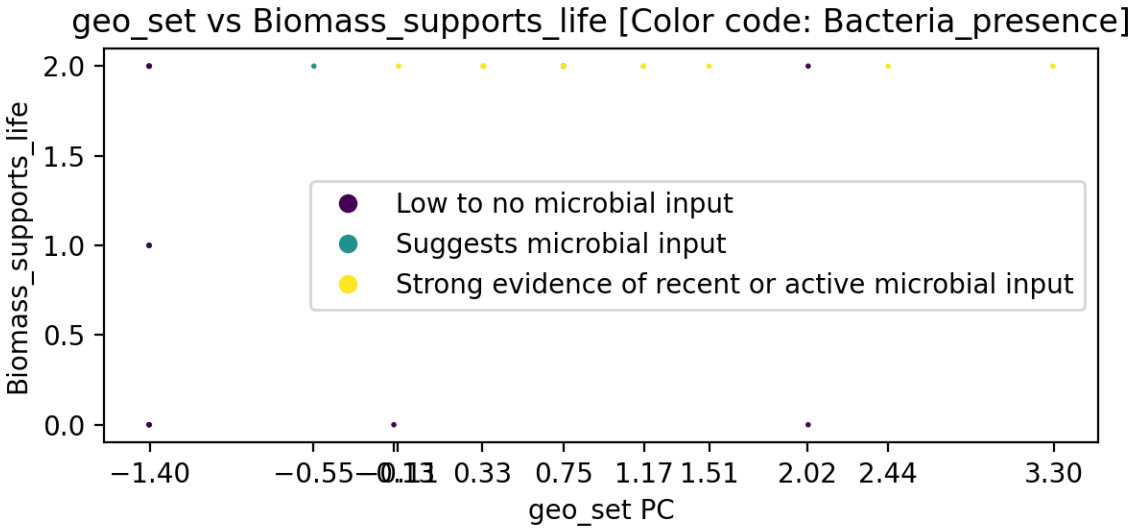

geo_set vs TOC % [Color code: Biomass_supports_life]
Low to no microbial input
Suggests microbial input
Strong evidence of recent or active microbial input
TOC %
2.0
1.5
1.0
0.5
0.0
−1.42
−0.50
0.38
0.85
1.33
2.24
2.72
3.63
geo_set PC


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 %
2.0
1.5
1.0
0.5
0.0
−1.42
−0.50
0.38
0.85
1.33
2.24
2.72
3.63
geo_set PC

3.2.2 Biomass Supports Life Development

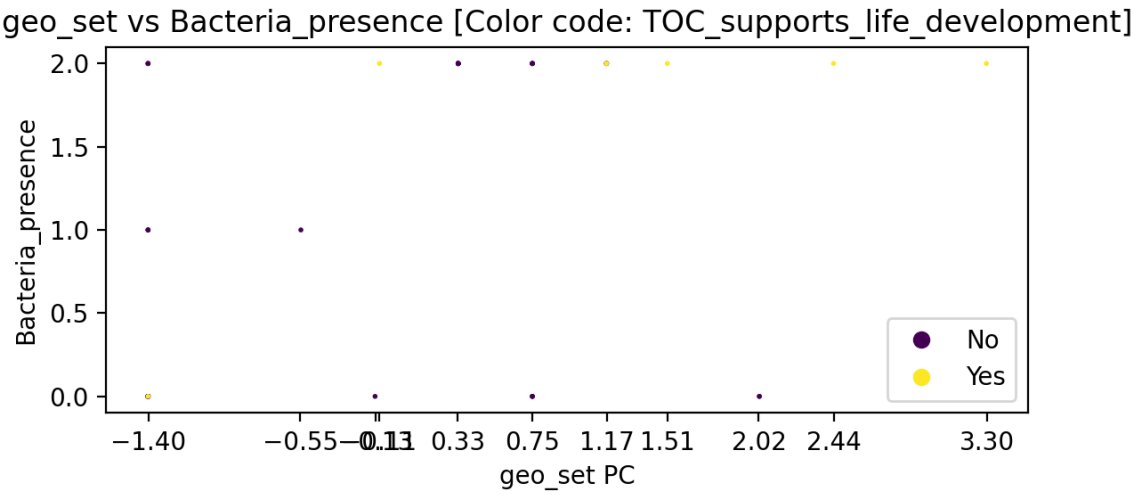
The geo_set variable was compared to Biomass_supports_life with the following color coding to determine if any trends were found: TOC_supports_life_development and Bacteria_presence.

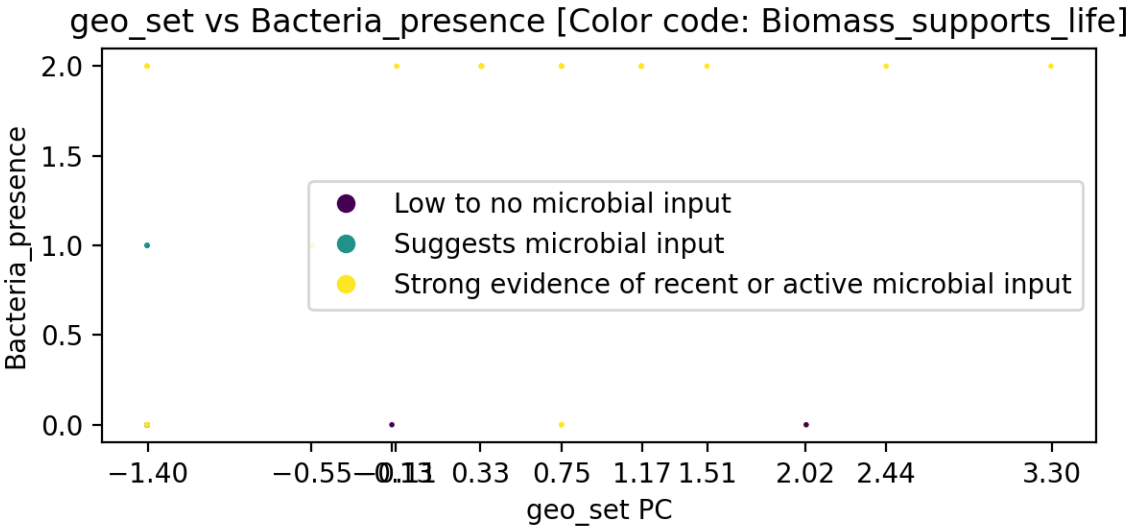




3.2.3 Bacteria Presence

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

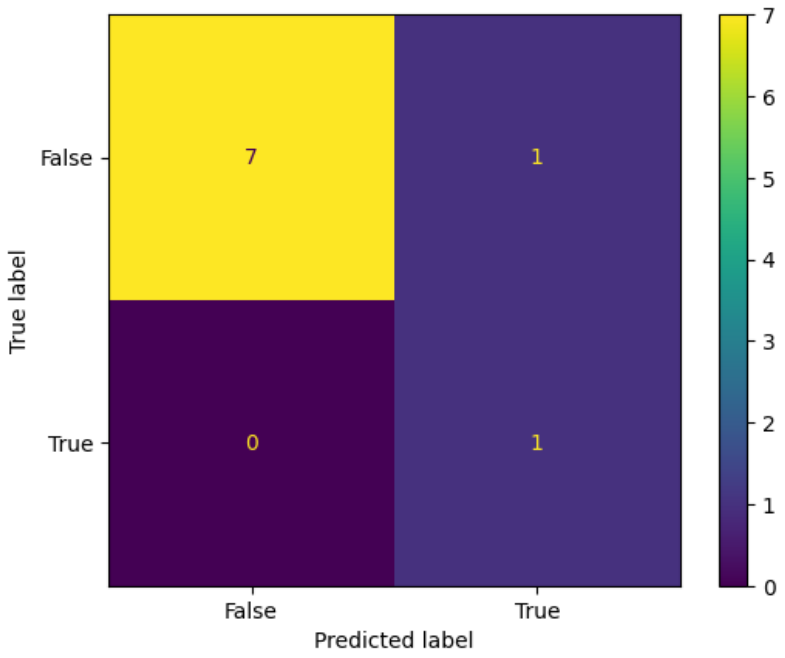
1. TOC is not an accurate provider of if life can develop in a system, at least not set at the cutoff point that is implemented in this study. In multiple samples, TOC did not support life, but the FAME abundances supported biomass and bacteria presence.
2. Biomass having the ability to support life 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 compared: 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 (91.62%). TOC class imbalance was also checked, with 7 data points supporting life development, and 38 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 0.001). 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% testing and 80% training, resulting in a cross-validated accuracy score of 88.89% ± 12.17%. 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	0.635021	0.392074	0.467153	0.624359
Sample_type	0.635021	1.000000	0.225594	0.319060	0.908069
Biomass_supports_life	0.392074	0.225594	1.000000	0.544010	0.548708
Bacteria_presence	0.467153	0.319060	0.544010	1.000000	0.621867
TOC Life Development Prediction	0.624359	0.908069	0.548708	0.621867	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. However, these variables did undergo KMeans clustering and can be found in sections 3.4.

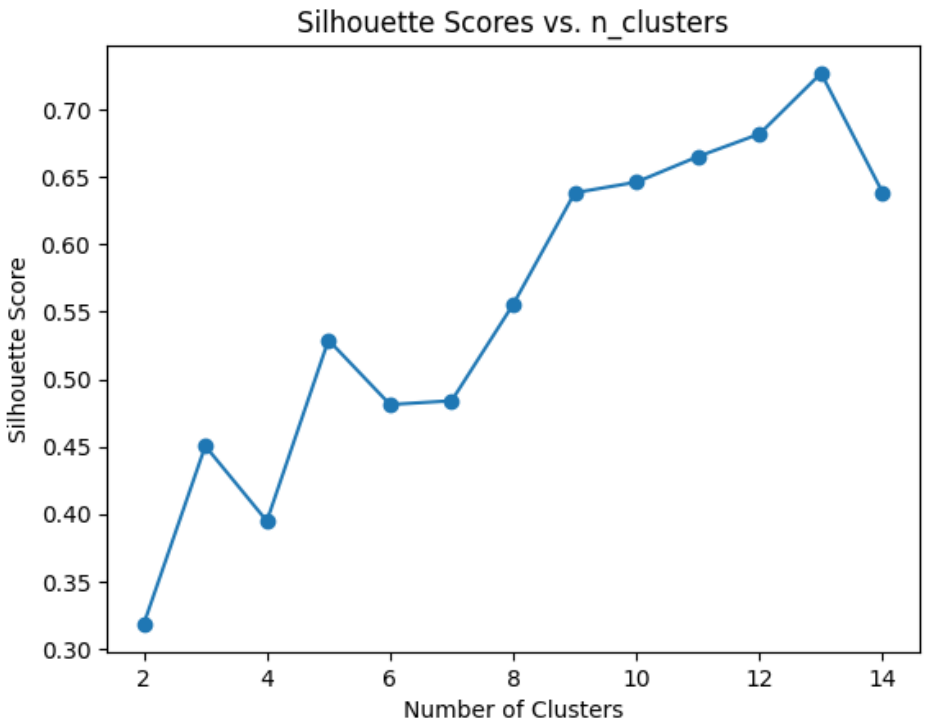
3.4 KMeans Clustering

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.

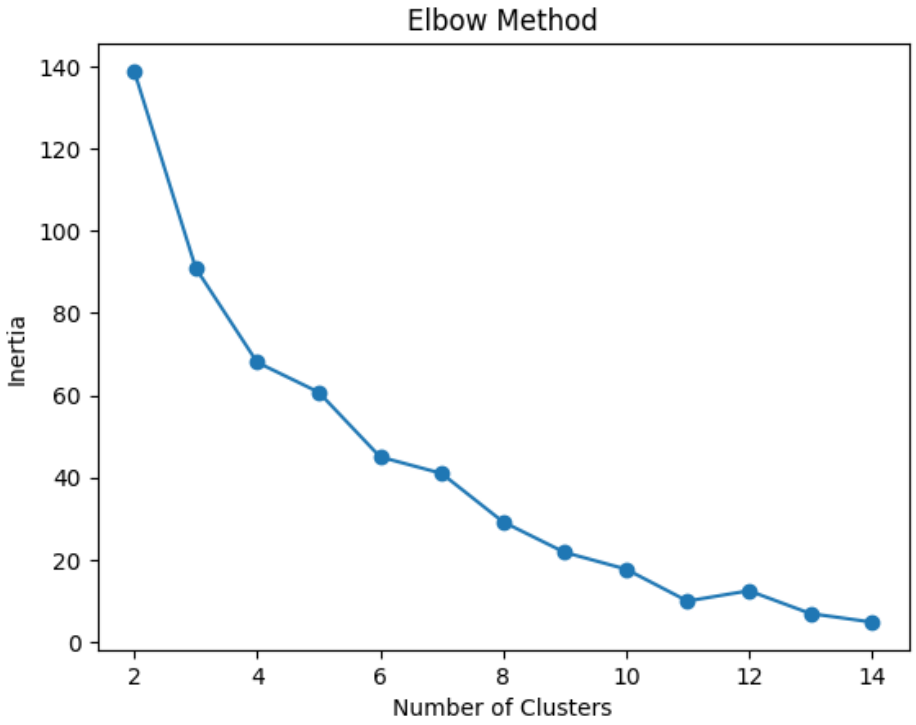
The dataframe was copied from the TOC logistic regression analysis. The dataframe underwent normalization and PCA analysis. A total of 5 PCs were created, with 4 PCs making up greater than 90% of the variance (95.72%). The ideal n_clusters value was determined through:

1. Plotting silhouette scores vs. n_clusters



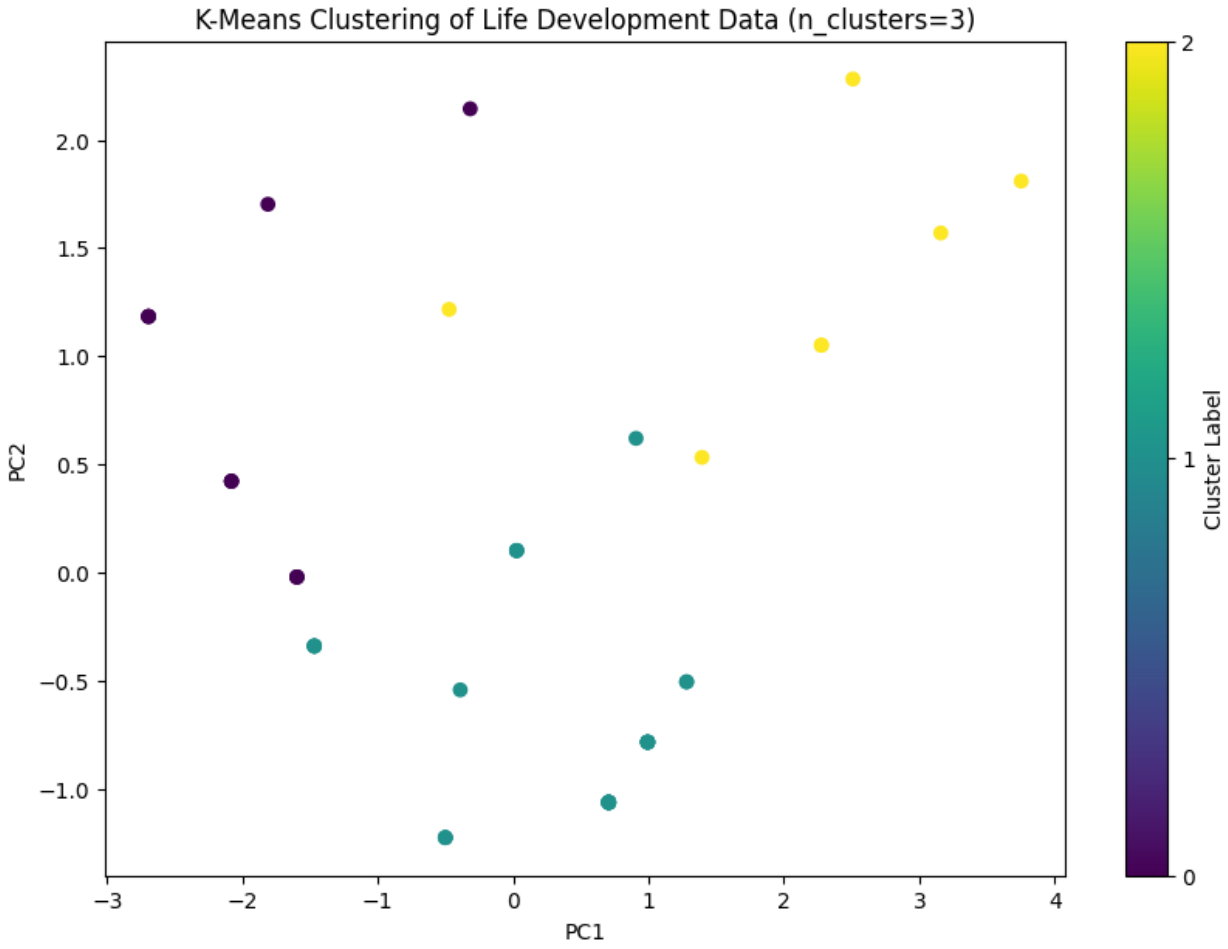
This silhouette scores vs. n_clusters showed that anything after n=6 diminishes the cluster analysis value. You can see a sharp increase after n=6 with a continual increase in silhouette score with the range that was tested.

2. The elbow method

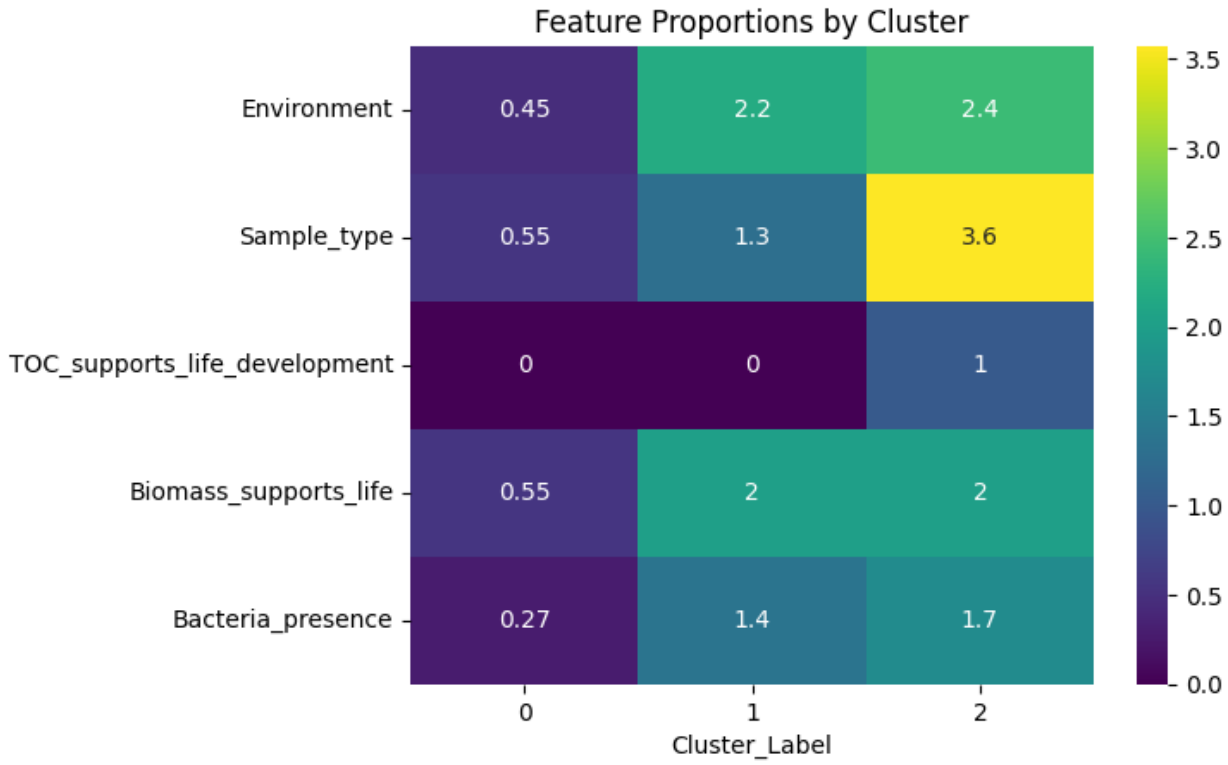


The “elbow” in the plot is where the line sharpy decreases then levels off. The elbow represents the best trade-off between model complexity and explained variance. As with the silhouette scores vs. n_clusters plot, further n_clusters beyond the “elbow” diminishes the cluster analysis value. In this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

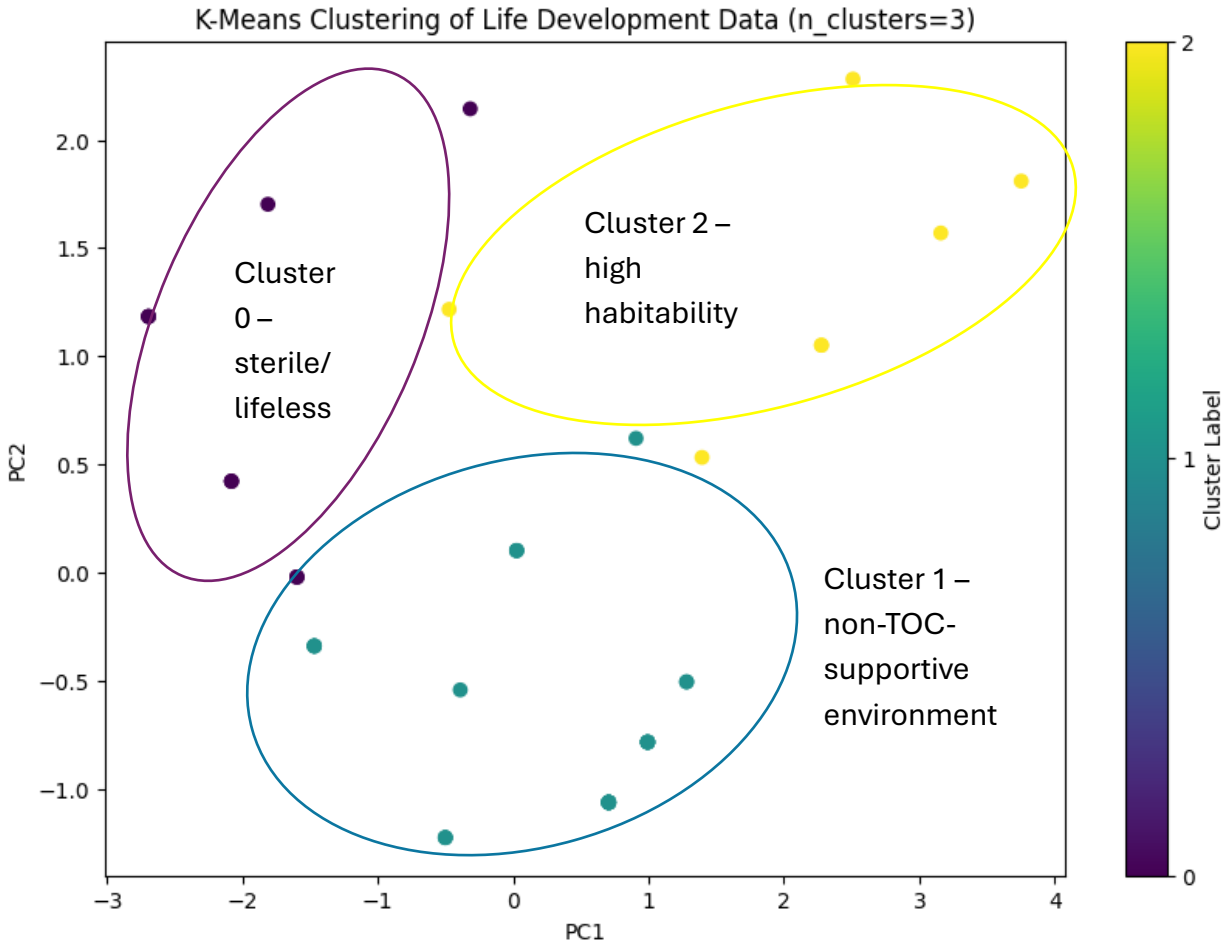


You can see three somewhat distinctive clusters. A cluster summary was created to help make the plot more interpretable. The distributions of each key feature were compared to each cluster. A heatmap was used for better visualization.

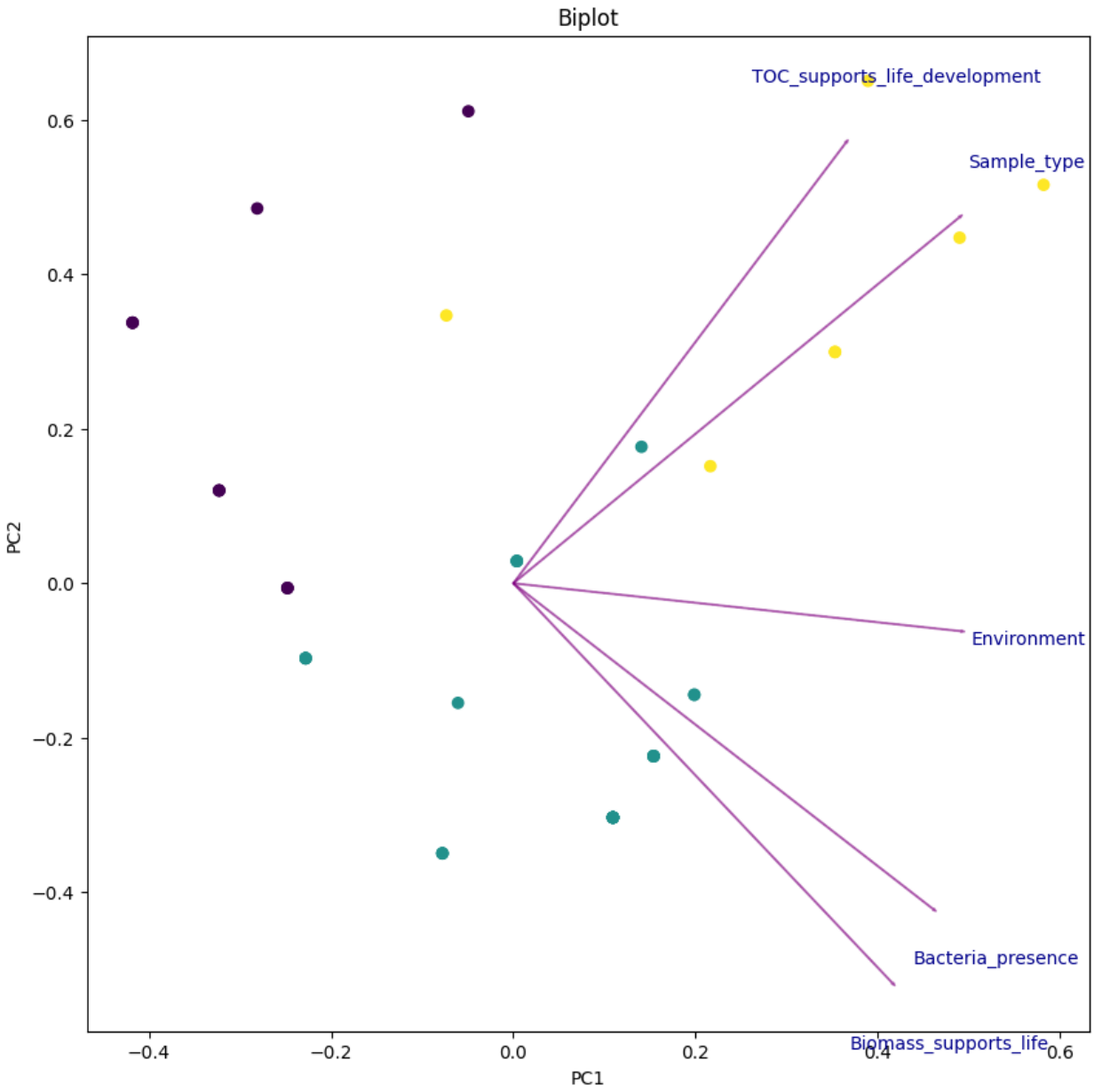


Based on the data, the following information was interpreted:

- 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 directly influences PC1 and influences all other variables to some degree.

4. Citations

1. Eigenbrode JL, Summons RE, Steele A, et al. Organic Matter Preserved in 3-Billion-Year-Old Mudstones at Gale Crater, Mars. Science 2018;360(6393):1096–1101; doi: 10.1126/science.aas9185.
2. Freissinet C, Glavin DP, Mahaffy PR, et al. Organic Molecules in the Sheepbed Mudstone, Gale Crater, Mars. JGR Planets 2015;120(3):495–514; doi: 10.1002/2014JE004737.
3. Michalski JR, Dobrea EZN, Niles PB, et al. Ancient Hydrothermal Seafloor Deposits in Eridania Basin on Mars. Nat Commun 2017;8(1):15978; doi: 10.1038/ncomms15978.
4. National Aeronautics and Space Administration (NASA). AI Astrobiology Life Detection & Biosignatures. 2025. Available from: https://www.nasa.gov/a-i-astrobiology-life-detection-biosignatures/ (Last accessed: February 2025)
5. Quantin-Nataf C, Carter J, Mandon L, et al. Oxia Planum: The Landing Site for the ExoMars "Rosalind Franklin" Rover Mission: Geological Context and Prelanding Interpretation. Astrobiology 2021;21(3):345–366; doi: 10.1089/ast.2019.2191.
6. Rench, B.M. Astrobiology Program FAQ. 2025. Available from: https://astrobiology.nasa.gov/about/faq/ (Last accessed: February 2025)
7. Scheller EL. Aqueous Alteration Processes in Jezero Crater, Mars—Implications for Organic Geochemistry. n.d.; doi: 10.1126/science.abo5204.

8. Summons RE, Amend JP, Bish D, et al. Preservation of Martian Organic and Environmental Records: Final Report of the Mars Biosignature Working Group. Astrobiology 2011;11(2):157–181; doi: 10.1089/ast.2010.0506.
9. Williams AJ, Craft KL, Millan M, et al. Fatty Acid Preservation in Modern and Relict Hot-Spring Deposits in Iceland, with Implications for Organics Detection on Mars. Astrobiology 2021;21(1):60–82; doi: 10.1089/ast.2019.2115.
10. Williams AJ, Eigenbrode J, Floyd M, et al. Recovery of Fatty Acids from Mineralogic Mars Analogs by TMAH Thermochemolysis for the Sample Analysis at Mars Wet Chemistry Experiment on the Curiosity Rover. Astrobiology 2019;19(4):522–546; doi: 10.1089/ast.2018.1819.