EcoSynth - Understanding and Enhancing Soil Health and Microbial Biodiversity Using Artificial Intelligence.

Built By: **Alfaxad Eyembe** , **Vijaya Sekhar Gullapalli**

Introduction

Soils, intricate ecosystems supporting the very essence of life for plants, animals, and humans, play a pivotal role in shaping our planet's sustainability. The vitality of soils extends beyond mere ground beneath our feet; it is intertwined with the health of our climate, the flourishing of biodiversity, the integrity of our lands, and the security of our food sources. As global awareness of the profound impact of land use changes on soil health grows, the need for comprehensive monitoring becomes increasingly apparent.

While strides have been made in characterizing the chemical and biological factors influencing soil health, questions persist about the nuanced effects of various land use changes across diverse ecosystems worldwide. Notably, the stark contrast in carbon storage between intensively managed agricultural land and natural soil underscores the critical importance of understanding the extent of soil degradation in different contexts. Yet, the intricacies of how different types of land use alterations influence the magnitude of soil degradation and the role played by environmental and biological factors remain largely enigmatic.

In this dynamic landscape, the abundance of publicly available data on land use and diverse soil characteristics, including biotic features such as species abundance, opens a realm of possibilities.

This is why we built **EcoSynth**, a groundbreaking initiative harnessing the power of Artificial Intelligence (AI) and Machine Learning (ML) to unravel the complexities surrounding soil health and biodiversity. By delving into vast datasets across European landscapes, **EcoSynth** seeks to provide novel insights that transcend traditional boundaries, offering a holistic understanding of the intricate interplay between land use, soil characteristics, and ecosystem dynamics.

We have leveraged cutting-edge AI technologies to illuminate the intricate tapestry of soils, unveiling unprecedented insights that hold the key to a sustainable and resilient future. EcoSynth stands at the forefront of this exploration, paving the way for a new era of understanding and safeguarding our planet's vital ecosystems.

Platform Description

EcoSynth stands as a pioneering platform that seamlessly integrates advanced technologies to revolutionize our understanding of soil health and biodiversity. Our holistic approach, driven by innovative AI and machine learning models, empowers users with unparalleled insights into European topsoil characteristics.

1. Visualization Tool - Mapping European Soil Characteristics:

EcoSynth's cutting-edge visualization tool is a gateway to unraveling the complexity of European topsoil. Seamlessly integrating with the Land Use / Cover Area frame statistical Survey Soil (LUCAS Soil) dataset, this tool transforms intricate soil data into visually compelling maps. Originating from the European Union, the LUCAS Soil dataset offers a comprehensive survey of topsoil characteristics across diverse landscapes. Our visualization tool provides an intuitive and insightful interface, allowing users to explore the nuanced features of surveyed areas dynamically. By mapping European soil characteristics, EcoSynth facilitates a deeper understanding of the intricate relationships within these vital ecosystems.

2. User-Friendly APIs - Robust Soil Prediction Models:

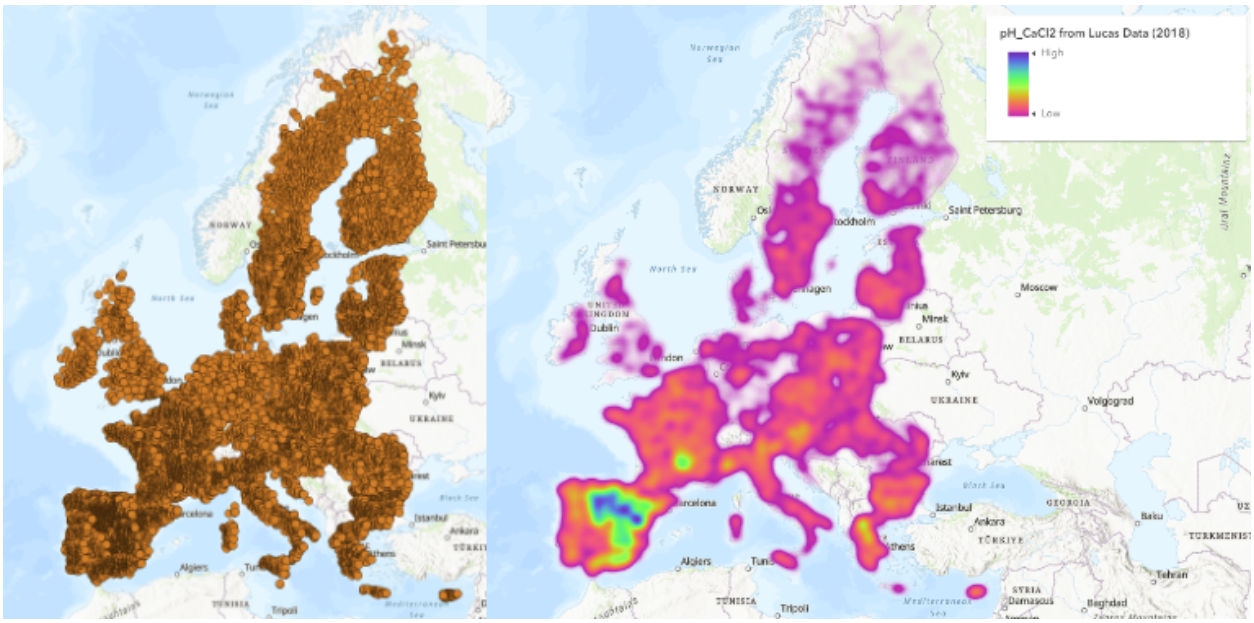
EcoSynth's functionality is anchored in user-friendly APIs housing a robust family of models designed to redefine soil analysis. These models exhibit a remarkable accuracy in predicting and classifying soil conditions, particularly in distinguishing between managed and unmanaged land. Beyond conventional analyses, our models leverage an array of physical, chemical, and biological features to offer a holistic understanding of soil health. They excel in detailed land use classifications, providing users with unprecedented insights. Notably, the models showcase an advanced capability to organize and cluster microbial species, capturing variations that serve as crucial indicators of soil health. EcoSynth's APIs empower users to navigate the intricate landscape of soil dynamics with unparalleled precision.

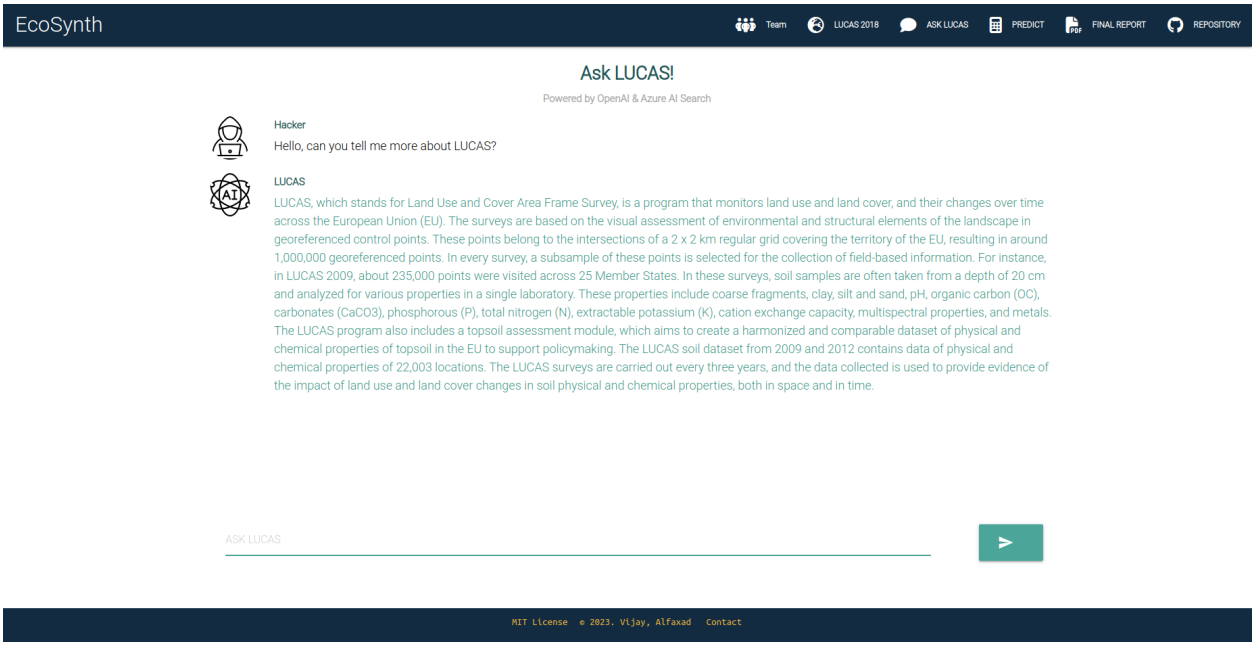
3. Semantic Search Platform - Unveiling Insights with AI-Powered Search:

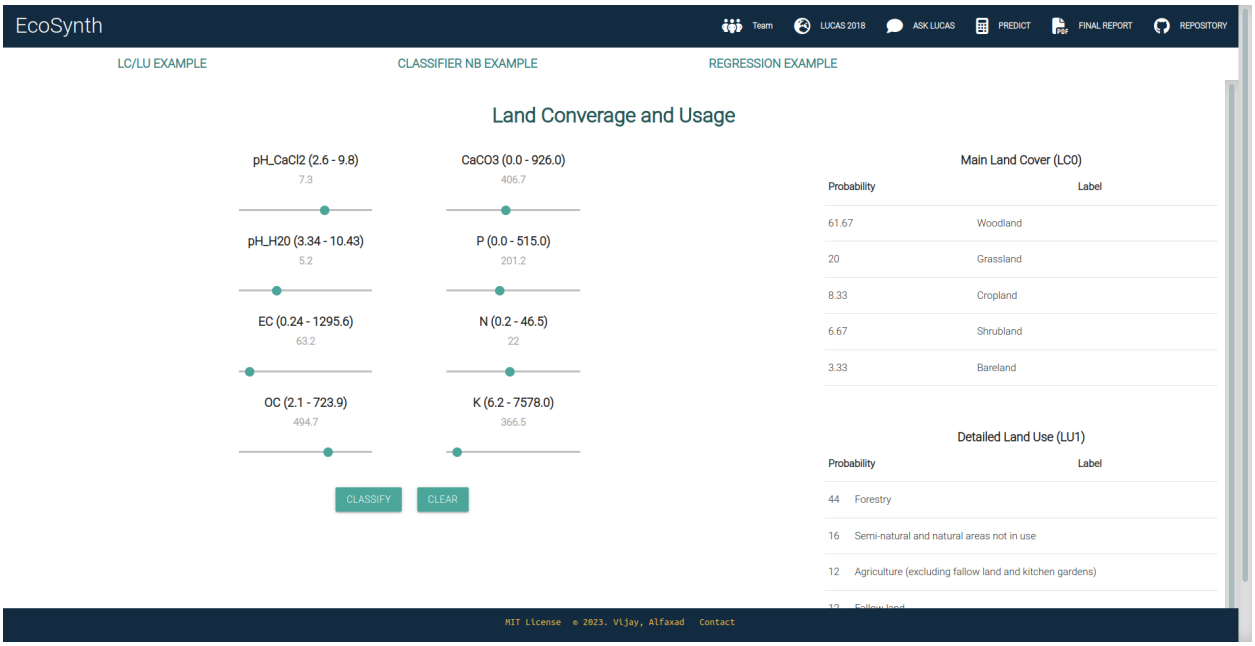
EcoSynth introduces a powerful semantic search platform, synergizing the capabilities of ChatGPT and Azure AI Search. This transformative tool enables users to embark on a journey of discovery within the vast expanse of the LUCAS Soil dataset. Fueled by state-of-the-art language models, the semantic search platform empowers users to extract valuable insights, uncover correlations, and gain a comprehensive understanding of the surveyed soil characteristics. Whether exploring specific features or seeking broader trends, our platform delivers tailored and informative results. By leveraging the capabilities of AI-powered search, EcoSynth opens new avenues for users to navigate and comprehend the intricate details of soil health.

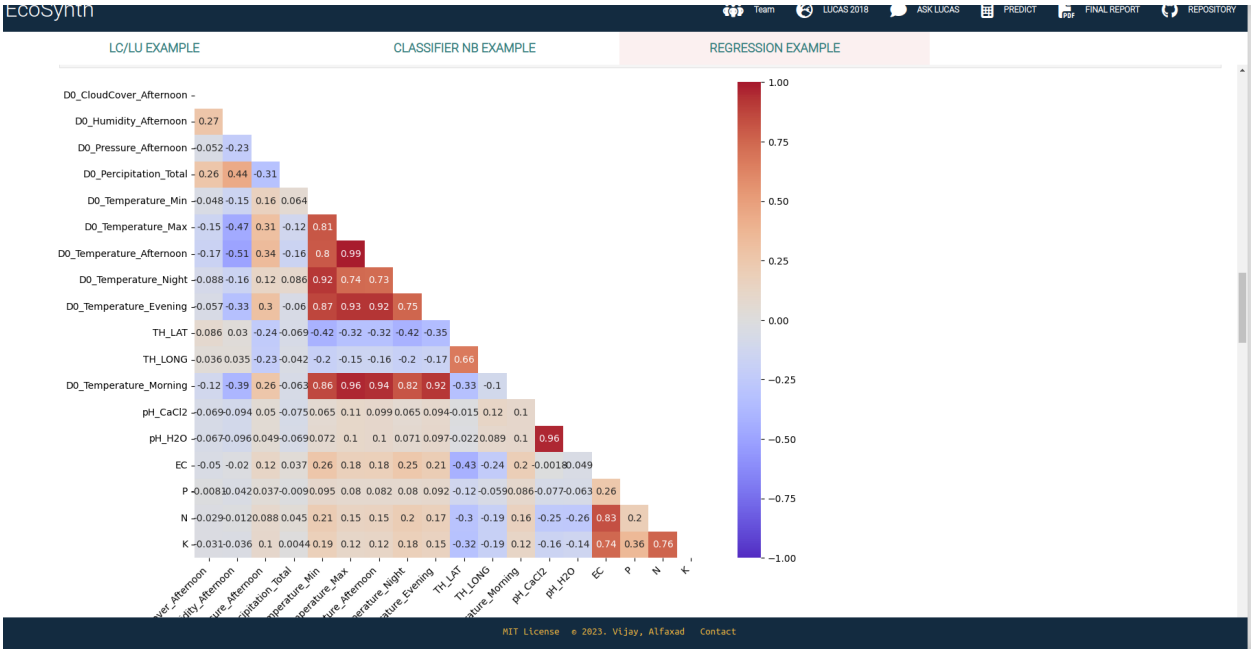
Open-Sourced Models for Collaborative Advancements:

At the heart of EcoSynth's commitment to transparency and collaboration lies our open-sourced models. By inviting public actors to assess the efficiency and accuracy of our soil prediction models, we foster a culture of innovation and collective expertise. This open approach not only encourages collaborative efforts but also ensures the continual improvement of our models. Through shared knowledge and collective exploration, EcoSynth's open-sourced models become a catalyst for advancements in the understanding of soil health and biodiversity. Join us in this collaborative endeavor, as we collectively strive for a sustainable and resilient future.


pH_CaCl2 from Lucas Data (2018)
High
Low
NORWAY
Saint Petersburg
Moscow
BELARUS
UKRAINE
Kyiv
Minsk
Volgograd
Istanbul
Ankara
Athens
Algiers
Tunis
Tripoli
Tehran
Baku
Baghdad
Damascus


EcoSynth
Team
LUCAS 2018
ASK LUCAS
PREDICT
FINAL REPORT
REPOSITORY
Ask LUCAS!
Powered by OpenAI & Azure AI Search
Hacker
Hello, can you tell me more about LUCAS?
LUCAS
LUCAS, which stands for Land Use and Cover Area Frame Survey, is a program that monitors land use and land cover, and their changes over time across the European Union (EU). The surveys are based on the visual assessment of environmental and structural elements of the landscape in georeferenced control points. These points belong to the intersections of a 2 x 2 km regular grid covering the territory of the EU, resulting in around 1,000,000 georeferenced points. In every survey, a subsample of these points is selected for the collection of field-based information. For instance, in LUCAS 2009, about 235,000 points were visited across 25 Member States. In these surveys, soil samples are often taken from a depth of 20 cm and analyzed for various properties in a single laboratory. These properties include coarse fragments, clay, silt and sand, pH, organic carbon (OC), carbonates (CaCO3), phosphorous (P), total nitrogen (N), extractable potassium (K), cation exchange capacity, multispectral properties, and metals. The LUCAS program also includes a topsoil assessment module, which aims to create a harmonized and comparable dataset of physical and chemical properties of topsoil in the EU to support policymaking. The LUCAS soil dataset from 2009 and 2012 contains data of physical and chemical properties of 22,003 locations. The LUCAS surveys are carried out every three years, and the data collected is used to provide evidence of the impact of land use and land cover changes in soil physical and chemical properties, both in space and in time.
ASK LUCAS
MIT License © 2023. Vijay, Alfaxad Contact





Models Insights and Performance & Results:

This section provides an in-depth exploration of EcoSynth's AI models, outlining their architecture, methodologies, and the impactful results they yield. From predictive accuracy to innovative insights, discover how our models redefine soil analysis and contribute to a deeper understanding of soil health and biodiversity.

The models were trained on LUCAS 2018 dataset and microbiomes sequence data based on LUCAS datasets.

1. **Ensemble Tree Models:**
   Ensemble tree models especially random forests were used to predict Land coverage and Land Usage. We are pleased to share the results of our models on various tasks along with main predictive features that influenced the model's predictions and performances.

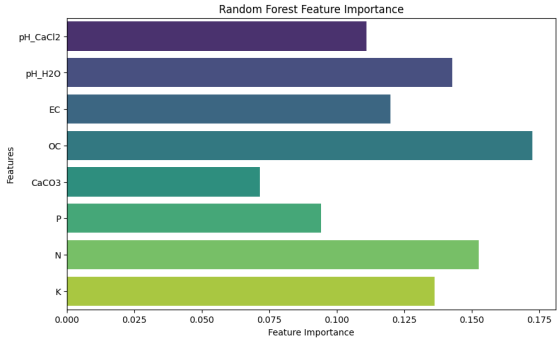
   Land Coverage(Main) Prediction/LC0: we used a random forest classifier and 8 features to classify the type of land coverage of an area. The following are diagrams of classification report and feature importance. Our model performed with a 94.8% accuracy.

```
Accuracy: 0.9484996628455833

Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      1474
         1.0       0.99      1.00      0.99      1520
         2.0       0.87      0.84      0.85      1531
         3.0       0.81      0.87      0.84      1453
         4.0       0.98      1.00      0.99      1492
         5.0       1.00      1.00      1.00      1465
         6.0       1.00      1.00      1.00      1505
         7.0       0.94      0.88      0.91      1424

    accuracy                           0.95     11864
   macro avg       0.95      0.95      0.95     11864
weighted avg       0.95      0.95      0.95     11864
```

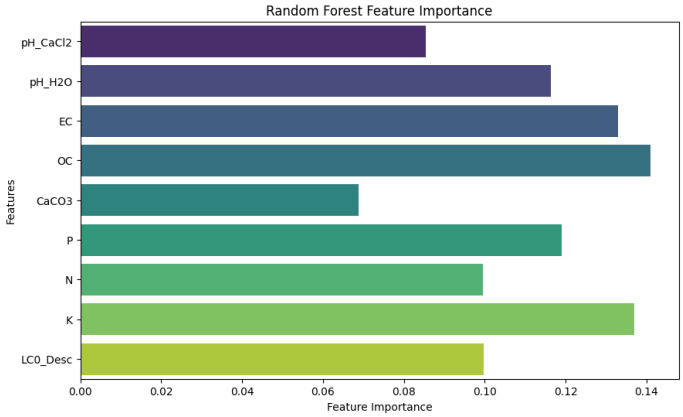


   Land Coverage (Further) Prediction/LC1: we used a random forest classifier and 9 features to classify the type of land coverage of an area. The following are diagrams of classification report and feature importance. Our model performed with a 98.66% accuracy.

```
Accuracy: 0.9866126091807026

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       538
           1       1.00      1.00      1.00       530
           2       0.92      0.92      0.92       566
           3       0.94      0.84      0.89       573
           4       1.00      1.00      1.00       592
           5       1.00      1.00      1.00       604
           6       0.94      0.80      0.87       557
           7       1.00      1.00      1.00       569
           8       1.00      1.00      1.00       532
           9       0.99      1.00      1.00       580
          10       1.00      1.00      1.00       574
          11       0.95      1.00      0.97       553
          12       0.97      0.82      0.89       532
          13       1.00      1.00      1.00       549
```

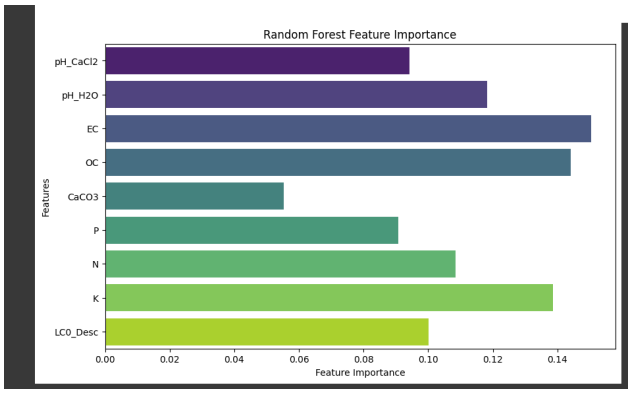


Land Usage Prediction(LU1):we used a random forest classifier and 9 features to classify the type of land coverage of an area. The following are diagrams of classification report and feature importance. Our model performed with a 99.8% accuracy.

Accuracy: 0.9980605804153884

Classification Report:

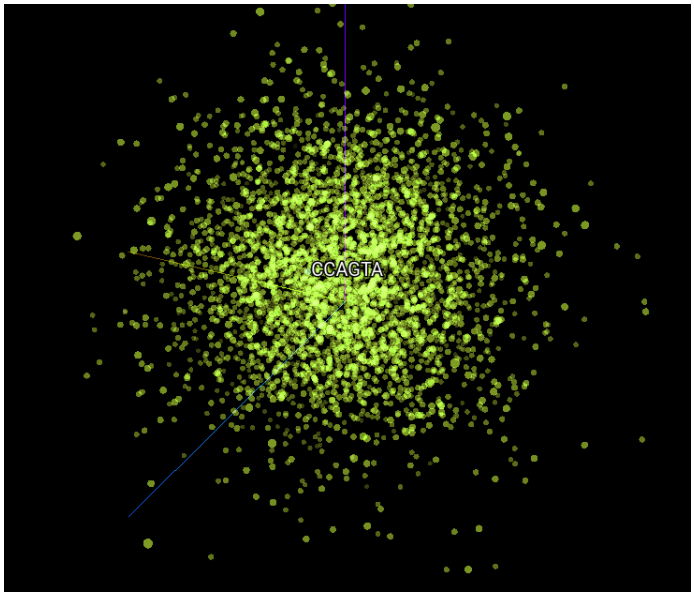
	precision	recall	f1-score	support
0	1.00	1.00	1.00	1110
1	1.00	1.00	1.00	1126
2	1.00	1.00	1.00	1082
3	1.00	0.97	0.98	1079
4	1.00	1.00	1.00	1071
5	1.00	1.00	1.00	1101
6	1.00	1.00	1.00	1080
7	1.00	1.00	1.00	1063
8	1.00	1.00	1.00	1133
9	1.00	1.00	1.00	1093
10	0.99	1.00	0.99	1124
11	1.00	1.00	1.00	1090
12	0.99	0.98	0.99	1099
13	1.00	1.00	1.00	1050
14	1.00	1.00	1.00	1093
15	1.00	1.00	1.00	1063
16	1.00	1.00	1.00	1105
17	1.00	1.00	1.00	1111
18	1.00	1.00	1.00	1069
19	1.00	1.00	1.00	1033
20	1.00	1.00	1.00	1079
21	1.00	1.00	1.00	1112
22	0.97	1.00	0.99	1114
23	1.00	1.00	1.00	1107
24	1.00	1.00	1.00	1095
25	1.00	1.00	1.00	1077
accuracy			1.00	28359
macro avg	1.00	1.00	1.00	28359
weighted avg	1.00	1.00	1.00	28359



2. **Nucleotide-transformer-v2-50m-multi-species fine-tuned model.**

To understand the significance of microbiome sequence data to soil health and ecosystem biodiversity, we fine tuned a foundational model named Nucleotide transformer(50M parameters version) that was released by InstaDeepAI that was trained sequences of various species. We fine tune it and apply it to the sequence dataset so that we can obtain “sequence embeddings” that we can use on downstream task such as clustering and classification visualization and sequence extraction. This was done successfully for both Prokaryotes and

Eukaryotes, however it was done on < 10% of the whole data due to compute and size constraints that we faced. Regardless, clustering, visualization and useful sequence extraction was successful for the given subset of the dataset.



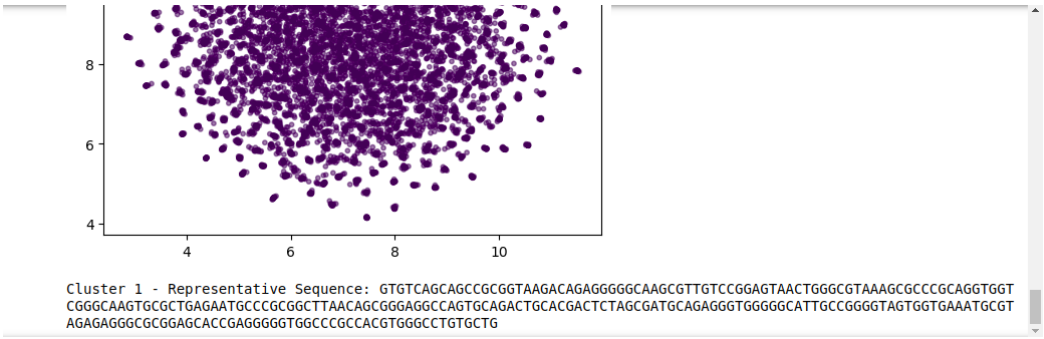
visualized sequence embeddings

,

[10]:
```
print(model.config)
```

```
EsmConfig {
  "_name_or_path": "InstaDeepAI/nucleotide-transformer-v2-50m-multi-species",
  "add_bias_fnn": false,
  "architectures": [
    "EsmForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.0,
  "auto_map": {
    "AutoConfig": "InstaDeepAI/nucleotide-transformer-v2-50m-multi-species--esm_config.EsmConfig",
    "AutoModelForMaskedLM": "InstaDeepAI/nucleotide-transformer-v2-50m-multi-species--modeling_esm.EsmForMaskedLM"
  },
  "emb_layer_norm_before": false,
  "esmfold_config": null,
  "hidden_dropout_prob": 0.0,
  "hidden_size": 512,
  "initializer_range": 0.02,
  "intermediate_size": 2048,
  "is_folding_model": false,
  "layer_norm_eps": 1e-12,
  "mask_token_id": 2,
  "max_position_embeddings": 2050,
  "model_type": "esm",
  "num_attention_heads": 16,
  "num_hidden_layers": 12,
  "pad_token_id": 1,
  "position_embedding_type": "rotary",
  "tie_word_embeddings": false,
  "token_dropout": false,
  "torch_dtype": "float32",
  "transformers_version": "4.36.0",
  "use_cache": false,
  "vocab_list": null,
  "vocab_size": 4107
}
```

fine-tuned model architecture



sample clustering sequence extraction

Limits.

EcoSynth, while achieving remarkable success in soil analysis, encounters inherent limitations primarily rooted in computation constraints. The expansive nature of sequence embeddings and the fine-tuning process of foundational models pose challenges in scaling our work. Computation restrictions hinder our ability to delve deeper into the intricate world of microbial biodiversity. Despite our commitment to pushing the boundaries of soil science, these constraints currently impede the full realization of our vision.

FurtherWork:

In our pursuit of innovation, EcoSynth has taken strides by creatively integrating weather data into our soil analysis framework. Preliminary regression results have compellingly demonstrated the influence of weather on soil health, paving the way for future endeavors to seamlessly integrate weather information into the expansive LUCAS dataset. As we anticipate a significant reduction in memory and compute constraints, our vision extends to extensive experimentation with Sequence embeddings. Scaling their potential holds the key to unlocking deeper insights into microbial biodiversity, propelling EcoSynth to new heights of analytical precision.

Scaling for Impact:

Built with a user-centric philosophy, EcoSynth stands as a testament to our commitment to accessibility and functionality. The successful deployment of our prototype underscores our dedication to providing users with a tangible tool for understanding soil health through AI. Looking ahead, our ambition is to elevate EcoSynth from a prototype to a fully functional tool at scale. With the prospect of expanded resources, we aim to transform EcoSynth into an indispensable asset for researchers, farmers, and environmentalists, facilitating a comprehensive understanding of soil health on a broader scale.

References.

1. https://ai4lifesciences.com/challenge-info/
2. https://huggingface.co/InstaDeepAI/nucleotide-transformer-v2-50m-multi-species#architecture
3. https://www.instadeep.com/2023/07/instadeep-open-sources-the-nucleotide-transformers-its-collection-of-genomics-language-models-to-huggingface/
4. https://esdac.jrc.ec.europa.eu/content/lucas-2018-topsoil-data
5. https://github.com/germs-lab/ref_soil/tree/master
6. https://github.com/sekhargullapalli/aiforlifesciences-hack-2023/tree/main
7. https://towardsdatascience.com/reinforcement-learning-based-energy-optimization-dea8fb687cda

8. https://www.deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40