**2022 Winter Cambridge AI+ Programme**

**Project Proposal**

Structure your Project Proposal based on the headings provided as follow, please use a clear and legible font. For further format information, please refer to the FAQ section.

|  |  |  |  |
| --- | --- | --- | --- |
| **Programme Pathway** | Biotechnology Engineering & Healthcare Technology | | |
| **Course Group** | 22CamAI+ BIOB-G1 | **Group Name** | 42 |
| **Group Leader** | Yixin Guo | | |
| **Group Members** | Yufeng Wen, Bohao Peng, Ruihan Hou, Haohong Qiu | | |

1. Project title

Provide a short descriptive title of no more than 50 words

|  |
| --- |
| Predicting Soil Microbial Diversity Based on Soil Properties by AI Approach |

1. Project summary

In no more than 800 words, intelligible to a reader who is not a specialist in this field, summarize the aims, significance and expected outcomes of your proposed project.

|  |
| --- |
| Soil microbes play a significant role in the ecosystem and are involved in many environmental mediation processes, for instance, nutrient cycling and plant growth, which are essential to environmental preservation and agricultural development etc. In the past, there had been tons of research looking into the relationship between soil microbes and environmental factors.  In the Yellow River Basin area of China, perennial water and soil erosion have led to dramatic changes in the microbial community in this region, and the microbes inhabited in the soil are the most abundant and active organisms on the land surface. Due to its special geological condition and the data accessibility, we select the Yellow River Basin area as our sampling site.  Samples were collected from three river sections in the upper, middle, and lower parts of the gully region, and the microbes in each sample were characterized by 16sRNA (for bacteria) and ITS (for fungus) high-throughput sequencing by the company as described by Hou et al. With a series of bioinformatic analyses, the particle composition and nutrient content of the soil sediment were deduced to be the main factors affecting the changes of microbial communities in the gully region Wangdonggou. 16S rRNA sequence reads can be clustered into OTUs at a 97% similarity level. Each OTU is characterized by a representative DNA sequence and can be assigned a taxonomic lineage by comparing it with a known bacterial 16S rRNA database. Using the known database, OTUs can be assigned to families, orders, classes, or phyla. The raw data and pre-processed files used in this project were obtained from Hou et al., too, including OTU information of both bacteria and fungus from the sampling sites.  On the other hand, soil properties, which include the particle size information, the relative abundance of C and N, and the absolute content of P, were previously analyzed by Hou’s team and we requested the results and the original data from them.  For the first part of work in this project, we plan to investigate into the complex relationship between soil microbes and environmental factors and distinguish certain keystone soil microbes at phylum level with the given soil properties by machine learning methods such as PCoA, statistical means as network analyses.  Due to the high-dimensionality of sequence data (high number of variables representing all amplicons of phylogenetic markers or partial genes in metagenome sequencing), we need to apply some machine learning methods which help us reduce its dimension. By performing PCoA on the given OTU dataset, we reduce its dimension and bring convenience on further analysis(we may only check on relation between principal coordinate and related properties).  Since focusing on all soil microbes is far too much complicated and inefficient, we plan to classify both bacteria and fungus at the level of phylum in this project. Network analysis can help disentangle the perplexing microbial co-abundance interrelations and provide a comprehensive insight into the microbial community structure and assembly patterns. One of the most useful features of network analysis is the ‘hubs’ (also termed keystone operational taxonomic units (OTUs)), which are commonly used for microbial classification. These hubs have been proposed as keystone taxa, as their removal has been computationally shown to cause a drastic shift in the composition and function of a microbiome. Thus, network analysis can be a powerful tool to infer the keystone taxa from microbial communities.  Based on the soil microbial taxonomy and other information, we plan to use R, Origin and probably some other online tools, such as DataVizCatalogue and From data to Viz, to present these information above by pictures. ComplexHeatmap(https://jokergoo.github.io/ComplexHeatmap-reference/book/) and ggplot2(https://ggplot2-book.org/) are supposed to be the main R packages to use in our project.  For the second part of work, we intend to adopt the same dataset to further construct an artificial neural network (ANN) to establish a deep learning model that can predict the microbial diversity in the soil given certain soil properties. Apart from constructing this neural network, we also want to interpretate its meaning through structural equation model. |

1. Significance/Contribution to the discipline

In no more than 600 words. This section will normally justify the project from a review of the literature on the topic.

|  |
| --- |
| 1.Many research before focus on getting to know statistical correlation between physiochemical properties and soil microbial community and conclude what may be the main factor influencing the latter. Beside concluding the crucial factor, our research seeks for predicting relative abundance of keystone taxa in microbial community in Wangdonggou.  2. In previous studies, most of them only considered the impact of total carbon content on microbial communities. Our project no longer simply considered the total carbon content and focused on the content of specific functional groups. Soil organic C (SOC）is composed of different functional groups (C−H, -COO, C=O,C=C, etc). The chemical composition of SOC determines the quality and stability of SOC and the microbial community composition in soils. Generally, the functional group of C−H and -COO are closely related to labile SOC, and the C=O and C=C functional group influence recalcitrant SOC.  3.We apply an up-to-date statistic mean called structural equation model(SEM) in microbial community. In SEM, path analysis is useful in explaining the causal relationships among variables. A common function of path analysis is mediation, which assumes that a variable can influence an outcome directly and indirectly through another variable. SEM brings researchers closer to the causal understanding by fitting data to models representing causal hypotheses and by determining goodness-of-fit as well as comparing the fit among models representing alternative, mechanistic hypotheses. This project will also combine the physical and chemical property indexes such as soil SOC, TN, Olsen-P, soil texture to build a SEM which somehow explains the ANN,  4. This project will use ANN (Artificial Neural Network) as a deep learning tool to further predict the microbial content in soil, so as to help predict the microbial content in various environmental soils. And through long-term training and multiple data sets, the AI will gradually strong, continues to collect a variety of different soil conditions at the same time, the effects of microbial analysis can also be combined with scientific research personnel, analysis the way of land use and soil characteristics, the relationship between calculated the relative abundance of the soil microorganisms, to analyze the stand or fall of soil utilization way. It can also carry out the backward calculation and calculate the soil characteristic structure required by the correct microbial community structure so that researchers can have relatively correct direction of thought when carrying out field research and trying to restore the good terrestrial microbial system that has been destroyed. |

1. Project Timeline (Gantt Chart) & allocation of tasks

Please illustrate the timeline of your projects on a Gantt Chart with a clear explanation of the roles & tasks of each member in your group.

**Part I Data Processing**

Ruihan Hou: refer to previous researches and seek to improve the first part of our project, which mainly includes the statistical analysis of the obtained microbial information and the realization of data visualization.

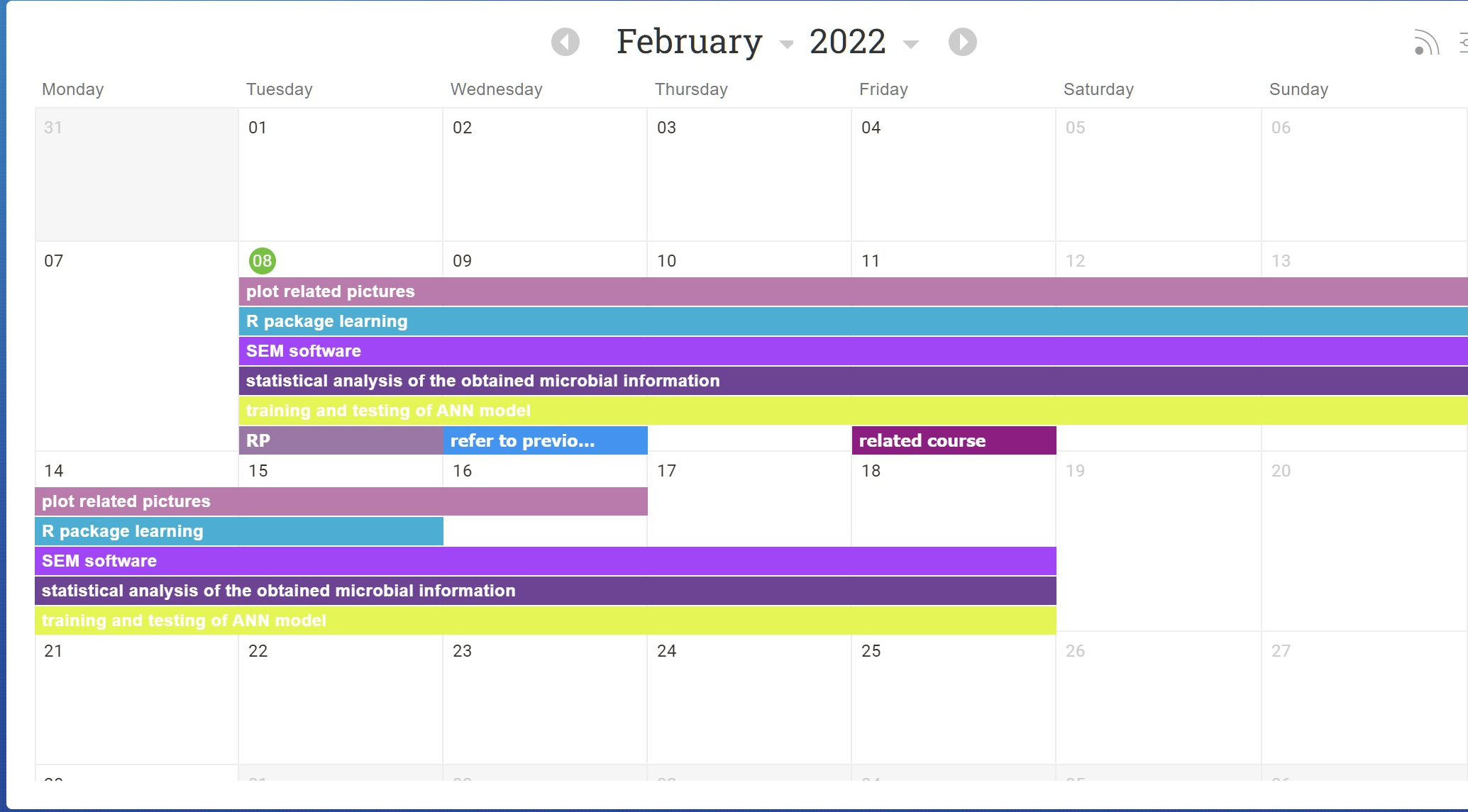
Bohao Peng: work with Hou to sort out the data we have and visualize them by R language

Haohong Qiu: refer to previous works related to our topic and some PPTs, summarize the general analyzing procedure, and help to collect useful information for our project

**Part II ANN & SEM**

Yixin Guo: learn about the background theory of structural equation modelling (SEM) and which software to use, and assist Wen’s work to realize the construction of ANN

Yufeng Wen: learn about the background theory of Artificial Neural Network (ANN) and realize the construction (training and testing) of ANN model



1. References cited/Bibliography of planned reading

Please provide a bibliography of planned reading

|  |
| --- |
| **References**  <https://www.dataquest.io/blog/jupyter-notebook-tutorial/>  <https://www.dataquest.io/blog/advanced-jupyter-notebooks-tutorial/>  <https://www.coursera.org/learn/neural-networks-deep-learning/lecture/Cuf2f/welcome>  <https://betterprogramming.pub/build-your-first-neural-network-from-scratch-c5d5490a3f76>  <https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/>  **Bibliography**  WANG R., SALMAN A., GAO X., GUO S. L. 2019. Variations of soil microbial communities along a valley bottom of the loess plateau and the influencing factors. *China Environmental Science*, 39, 4350-4359  LARSEN, P. E., FIELD, D. & GILBERT, J. A. 2012. Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods,* 9**,** 621-+.  WU, X. H., WANG R., GAO C. Q., DU L. L., KHAN A., BARMON M., GUO S. L. 2021. Variations of soil properties effect on microbial community structure and functional structure under land uses. *Acta Ecologica Sinica*, 41, 7989-8002  LESNIK, K. L. & LIU, H. 2017. Predicting Microbial Fuel Cell Biofilm Communities and Bioreactor Performance using Artificial Neural Networks. *Environ Sci Technol,* 51**,** 10881-10892.  KUANG, J., HUANG, L., HE, Z., CHEN, L., HUA, Z., JIA, P., LI, S., LIU, J., LI, J., ZHOU, J. & SHU, W. 2016. Predicting taxonomic and functional structure of microbial communities in acid mine drainage. *ISME J,* 10**,** 1527-39.  ZHOU, J., DENG, Y., LUO, F., HE, Z., TU, Q. & ZHI, X. 2010. Functional molecular ecological networks. *mBio,* 1.  MISHRA, U., GAUTAM, S., RILEY, W. J. & HOFFMAN, F. M. 2020. Ensemble Machine Learning Approach Improves Predicted Spatial Variation of Surface Soil Organic Carbon Stocks in Data-Limited Northern Circumpolar Region. *Front Big Data,* 3**,** 528441.  KNIGHT, R., VRBANAC, A., TAYLOR, B. C., AKSENOV, A., CALLEWAERT, C., DEBELIUS, J., GONZALEZ, A., KOSCIOLEK, T., MCCALL, L. I., MCDONALD, D., MELNIK, A. V., MORTON, J. T., NAVAS, J., QUINN, R. A., SANDERS, J. G., SWAFFORD, A. D., THOMPSON, L. R., TRIPATHI, A., XU, Z. Z., ZANEVELD, J. R., ZHU, Q., CAPORASO, J. G. & DORRESTEIN, P. C. 2018. Best practices for analysing microbiomes. *Nat Rev Microbiol*, 16, 410-422.  WANG R., HU Y. X., KHAN A., DU L. L., WANG Y., HOU F. B., GUO S. L. 2021. Soil prokaryotic community structure and co-occurrence patterns on the fragmented Chinese Loess Plateau: effects of topographic units of a soil eroding catena. *Catena*, 198, 105035  MARCIO F. A. L., EIKO E. K., You must choose, but choose wisely: Model-based approaches for microbial community analysis, *Soil Biology and Biochemistry*, Volume 151, 108042 |

**FAQ**

**What are the format requirements for the Project Proposal?**

The format for a project proposal varies between different disciplines, but all good proposals must be both informative and persuasive. Please follow the following general format requirements:

* + File Format: .doc or .docx
  + Word Count: 2,000 (Max); 1500 (Min)
  + Font: Times New Roman
  + Font Size: 15 (Max); 9 (Min)
  + Line Spacing: 2 (Max); 0.8 (Min)
  + Citation Format: Harvard Citing and Referencing Style ([click here](http://www.imperial.ac.uk/admin-services/library/learning-support/reference-management/harvard-style/)for more information)

**Statement on Plagiarism**

Make sure that you acknowledge the authors of ALL publications you use to write your proposal. Failure to do so will be considered as plagiarism. Do not copy word for word what an author has said. You may think that the original author has presented the information using the best possible words in the best format. However, it is best to analyse the information presented and re-write it in your own words. If you absolutely have to quote an author ad verbatim, then make sure that you use quotation marks and italics to indicate it.