| Dates | Activities | Notes |
| --- | --- | --- |
| 2024/11/20 | Met Prof. Roy and Dr Yun in our group meeting. We discussed to select research topics.  Classification (phenotyping)   * Glycopeptide-resistant enterococci (GRE) set (12 trains, 35 isolates) * Bacillus & Brevibacillus set (7 species & 34(33) strains)   Regression (TVC quantification)   * Milk spoilage (0 – 168h) * Meat spoilage (0-78h) | I chose Glycopeptide-resistant enterococci (GRE) set (12 trains, 35 isolates) as I have more experience in clinical microbiology. This topic will be easier for me to understand what and why I’m doing this. |
| 2024/11/21 | Dr Yun sent 3 relevant papers to read. He told me to wait for the raw data that he needs to check first. |  |
| 2024/12/03 | I sent my presentation slides about my project that I will present for LIFE748 Assessment to Prof. Roy and Dr Yun for feedback. | Dr Yun corrected what I misunderstood about what to do and some text errors. Prof. Roy suggested giving full technical terms instead of just abbreviation alone. |
| 2025/01/08 | Dr Yun sent raw data for the project. He suggested me to align and trim the data to find the appropriate range of m/z that cover all the samples, then try to do pre-processing as described in the paper. | Due to my assignments, I struggled with them and feared failing the modules, so I was focusing on my coursework more than my project. I also re-learn Python to help me with coding by looking back at what I have learnt from LIFE733 and watched some Python courses on YouTube to help my coding skill for my project. |
| 2025/04/08 | Started to write some codes for pooling text files from 560 samples to store data in one csv file, and trimming. | As the data is very big, my codes ran slowly. It took a lot of time to see any results, especially when there are errors and I have to run the code again after fixing them. |
| 2025/04/28 | I tried to do baseline correction and normalization. | I was trying to use Matlab and R more as Dr Yun is more familiar with them than Python. The baseline correction results looked good, so I put the plots on my poster just to be on time to submit, but I think I should go back to Python as I understand it more than Matlab, which will take longer time for me to write any codes as I’m not familiar with it. |
| 2025/05/22 | In a group meeting, Dr Yun said my csv file is not what he expected to see and needed to be rearranged from the beginning. | The lipid dataset isn’t aligned well with their m/z ratio, so I can’t do anything with that further. |
| 2025/05/23 | Dr Yun confirmed that the lipid data is not okay, so I can work with only protein data by trimming with the same m/z range as in previous study (1000 – 18000) and put into a proper matrix. | The lipid data were from analysis on 4 four different days, so they have different set of m/z ratio and can’t do well with my simple codes. |
| 2025/05/25 | Sent a trimmed protein matrix file to Dr Yun |  |
| 2025/05/27 | Dr Yun checked my file and suggested that I can trim more to 10000 as there are still a lot of blanks. He also suggested to do spectral alignment, which they didn’t do it in the previous study, to see if it will give better results than before. | I used Icoshift as Dr Yun suggested. At first it failed and I didn’t know what’s happening. Then I copied the errors to asked ChatGPT and it showed me how to solve the errors. It turned out that it was not just my codes, but also the Python core package itself that need to be fixed. |
| 2025/05/30 | I sent the aligned data with plots to show differences between before and after to Dr Yun | Dr Yun commented that the alignment was successful. He also suggested that I could do baseline correction from arPLS algorithm. Then I could try peak picking to get the real peak data by using a package from Scipy. After that I can try pair normalisation and scaling methods and plot PCA to see which one is the best, then continue to supervised classification like PLS-DA. |
| 2025/06/02 | I sent what I think it is baseline corrected result to Dr Yun, but it turned out that it was just baseline, not baseline corrected result. So I subtracted my baseline from my data to get the result. I also asked about lambda parameter. I have tried with different lambda from 10, 50, and 100. Dr Yun suggested it to be 105 to 106. I edited it and sent the results on the same day. | I also asked about the process. I thought it could be normalisation --> scaling --> peak picking --> PCA --> PLS-DA. Dr Yun said that peak picking should come before normalisation. He also corrected my understanding about my project that it’s not classification between resistant vs susceptible. It’s between different strains of resistant ones. |
| 2025/06/03 | Dr Yun commented that my baseline data is a little bit conservative, but it’s better than having over-fitted baselines.  I then continued peak picking code with find\_peak from Scipy package. | I have tried to pick peaks with  prominence=50  threshold=0.1  I rounded m/z decimal points to 2 for binning detected peaks and made a matrix |
| 2025/06/06 | I sent my results to Dr Yun, and he said threshold=0.1 is too low. He recommended to do smoothing by Savitzky-Golay filtering before peak picking to reduce noises. |  |
| 2025/06/10 | I met Dr Yun in the group meeting. He suggested to try find\_peaks\_cwt as an alternative method for peak picking. | For smoothing parameters, I used:  window\_length=101  polyorder=3  deriv=0  delta=1.0  mode='interp'  I also tried both methods, find\_peaks and find\_peaks\_cwt as Dr Yun recommended. And I compare results between two methods.  For find\_peaks parameters, I used:  height= 400,  threshold= None, (I tried to use threshold, but it ignored many clear peaks, even it's very low like 0.1. So I just use None.)  prominence= 800,  width= 50  For find\_peaks\_cwt parameters, I used:  widths = (100, 400)  nd I also filtered intensities for more than 200 to remove background noises.  I tried to plot first 12 samples for both methods to see peaks from 3 different strains. |
| 2025/06/13 | I sent results to Dr Yun |  |
| 2025/06/17 | Dr Yun commented that the smoothing worked fine. The find\_peak looks good, but find\_peak\_cwt is not good as the peak locations often missed the peak apex. He said I can use find\_peak. After this he suggested that I can try to write a code for peak grouping because there are slight differences in the m/z between different samples even they are the same peak. This will reduce 0s in the matrix. |  |
| 2025/06/23 | I sent the results after peak grouping to Dr Yun. |  |
| 2025/06/24 | Dr Yun said there are too many 0s in my peaks and I need to try adjusting my threshold. I then asked him What should be the limit for threshold that I should not cross to go too far and ruin the data? He said I can try between 4-16 as the data has 4 experimental and 4 biological replicates. | I tried to adjust threshold = 1, 4, 12, 16. I continued with threshold = 12. Then I paired 4 normalisation methods (None, TIC, Max Intensity, Log) with 6 scaling methods (None, Z-score, Pareto, Unit Vector, Max Absolute, Robust) and perform PCA. |
| 2025/07/04 | I had an online meeting with Prof Roy and Dr Yun to talk about my report writing. I also show them my PCA plots and asked them which normalisation and scaling combination is the best. I think log/robust is good and they agreed. I sent the results with email. |  |
| 2025/07/07 | Dr Yun suggested to do PCA without the strains that were most different to others. And try CCA for visualisation. Then do PLS-DA for classification. |  |
| 2025/07/17 | I sent PCA plots, with and without outliers. |  |
| 2025/07/25 | Dr Yun said there is not much clustering trend between the 33 strains in the PCA scores plot. CCA should be used | I have tried CCA for and the plots are still not showing clusters between 33 strains. I'm not sure if I should continue to PLS-DA or go back to try different methods for normalisation and scaling. I prepared my code for PLS-DA just in case I can continue. |
| 2025/07/28 | I sent all CCA plots (1 vs 2, 1 vs 3, and 2 vs 3), PCA plots (35 strains and 33 strains excluding outliers), PCA scores for both 33 and 35 strains, X and Y matrices (to show you if there's something wrong with my processes) with email. Asked Dr Yun if I can continue to PLS-DA or need to try other normalisation/scaling methods or go back further than that. | Dr Yun was out of office for annual leave until 29th August with limited access to his email. He mentioned that he will still check email regularly, so I can send him an email if I have any questions. |
| 2025/08/04 | I sent an email to Dr Yun again. |  |
| 2025/08/11 | I sent my final report draft to Prof Roy and asked if he could help me about my project as Dr Yun was away. Prof Roy asked about how is the PLS-DA. I then sent him PLS-DA plots and CMs between full data and data without outliers. | The report was not completed, only Introduction and Materials & Methods were written as I don’t have enough results. I asked if the introduction is long enough and what should I put into my report. |
| 2025/08/12 | As the results from 35 isolates were not good, Prof Roy suggested to use 12 PFGE for classification. |  |
| 2025/08/19 | I tried PLS-DA, LDA, and RF with 12 PFGE classification. The RF results looked weird. Prof Roy said they are overtrained. He recommended to take 3 of the replicates to generate the model and then to test with the 4th replicate for all models, and increase the number of principal components being fed into the LDA algorithm. | I did what Prof Roy recommended. I increased PCs from 10 to 50 and use my code to find the best PC for each method and data. I also corrected my training/test as Prof Roy said. As there are 2 kind of replicates, biological and experimental, so I try to train my models from both. |
| 2025/08/21 | I sent the results I that I tried as Prof Roy suggested. I found that my RF code has some mistakes, but it didn’t show any error. I think that might cause the overtrain before and I didn’t think I should use RF. After I corrected it, the RF results looked better than before. |  |
| 2025/08/22 | Prof Roy suggested that I should calculate overall CCR (correct classification rate) to see which one is better between PLS-DA, LDA, and RF. He also suggested to discuss the common themes in terms of which isolates are always predicted with high accuracy, and how to improve the analysis in the future. |  |
| 2025/08/26 | I calculated overall CCR for PLS, LDA, RF and re-generated all CMs to be in the same pattern to show training and test CMs and sent them to Prof Roy. I asked if I have enough results for my report. |  |
| 2025/09/01 | I sent my first full report to Prof Roy for feedback. He suggested to add more references and explain more about why I choose those methods. |  |
| 2025/09/02 | I corrected typos and added more references and explanation. I finished my final report. |  |