Logbook Science Fair 2025-2026

Table of Contents

General Format:
Date: [Day]
People Involved: [Ivan, Matthew, etc.]
Title: [If you had to give it a title or to summarize it, what would it be?]
Topic/Section: [...]
Accomplished and Goals: [Don't have to do this for each one, just if there is something to add]
What We Did On This Day (WWDOTD): [More in-depth notes style. Diagrams, examples, annotations, quotes, etc. I usually have ½ a page to a full page for this]

Copy and Paste:

Date:
People Involved:
Title:
Topic/Section:
Goals:
WWDOTD:
WWHC:
Notes:

September:

Science Fair Prep, Brainstorming, and High-Level Planning

Date: Sep 24, 2025
People Involved: Ivan, Matthew, Cole
Title: Science Fair Meeting #2
WWHC: Finished Logbook
BRIEF TODO:
Solidify our idea for our project
Topic/Section: Planning
Accomplished and Goals:

- Figure out more specifically abt what we are going to do
- Create Project Proposal

WWDOTD:

- Got to meet with Cole and talked generally about the project plan

October:

Date: Oct 1, 2025
People Involved: Ivan, Matthew
Title: Discussed plans
WWHC: Got a better understanding of the project we might have done
BRIEF TODO:
Redo project proposal with new research, solidify the idea to incorporate CS better
Topic/Section: Planning
Accomplished and Goals: Redo project proposal and do more research
WWDOTD:

- Discussed plans revolving around the original idea that Ivan planned
    - Came up with a list of general issues that could impact the project significantly
        - Parental consent is required for the survey
        - A large amount of data is required
        - Bacteria distribution from the diet is unreliable

    - Possible fixes
        - Remove depression as one of the Responding Variables
        - Only test for Bifidobacterium
- Ivan will do another project proposal tomorrow

Works Used:

1. https://www.canada.ca/en/health-canada/services/science-research/science-advice-decision-making/research-ethics-board/consent-process.html

Date: Oct 1, 2025 : 11:30pm

Planning and Drafting (Project Proposal, Methodology, Background Research, etc.)

People Involved: Ivan
Title: Solidifying Scope and Early Draft of Project Proposal
Topic/Section: Planning and Drafting
Accomplished and Goals: Started Background of PP, need to get completed
WWDOTD:

- Added TODO List, made title page, and started Background for PP
    - Solidified the scope of the project and title:
    - No survey anymore, similar to last year
    - Look at previous data -> Compare to adult (do research ~10 papers) -> SVM

WWHC: Completed ½ of the first draft of the Background of PP

Date: Oct 2, 2025
People Involved: Ivan, Matthew
Title: Science Fair Meeting
WWHC: Obtained more information on the Science Fair process, Got better idea of what to do on the project
Topic/Section: Planning
Goals: Obtain public data on the topic, sign up for the CYSF portal, and submit ethics
WWDOTD:

- Discussed plan with Science Fair mentor
    - Submit ethics soon

        - Create another project proposal
            - Where to obtain data (public data sources)
            - Data analysis process
            - AI building process
    - Ivan will finishthe project proposal

Works Used:

Date: Oct 10, 2025
People Involved: Ivan, Matthew, Cole
Title: Did Background Research and Signed Up for CYSF
Topic/Section: Planning, Research
Accomplished and Goals: Read all the linked papers in Sources To Read
What We Did On This Day (WWDOTD):

- Added many sources to the sources to read document
- Signed up for CYSF

Date: Oct 12, 2025
People Involved: Matthew
Title: Updated Project Proposal
Topic/Section: Planning, Research
Accomplished and Goals: Finish Project Proposal
What We Did On This Day (WWDOTD):

- Update project proposal draft
    - Reworded some sections
    - Added a section on our methods and goals
    - Added Works Cited

Date: Oct 13, 2025
People Involved: Ivan, Matthew
Title: Completed CYSF Registration
Topic/Section: Report
Accomplished and Goals: Wait for Ethics Form to return

What We Did On This Day (WWDOTD):
- Wrote our basic project details and ethics form
    - Submitted and now waiting for feedback

Date: Oct 15, 2025
People Involved: Matthew, Ivan
Title: Transferred Data
Topic/Section: Data
Accomplished and Goals: Process the data
What We Did On This Day (WWDOTD):
- Moved the data from Ivan's computer to Matthew's computer due to Matthew's computer having more storage and processing power

Date: Oct 19, 2025
People Involved: Ivan
Title: QUIME2 Data Processing & Project Proposal
Topic/Section: Data, Planning
Accomplished and Goals: Work on DADA2 data processing
What We Did On This Day (WWDOTD):
- Completed the QUIME2 process and moved on to DADA2
    - Updated the files in the shared drive
    - Created a backup on Matthew's computer
- Updated the Project Proposal
    - Added "Expected Results" section
    - Added "Project Significance" section

Date: Oct 20, 2025
People Involved: Ivan
Title: DADA2
Topic/Section: Data
Accomplished and Goals: Work on DADA2 data processing
What We Did On This Day (WWDOTD):
- Started DADA2 processing

Date: Oct 21, 2025
People Involved: Ivan, Matthew
Title: Science Fair Meeting
Topic/Section: Planning
Accomplished and Goals: Read all papers in Sources to Read

What We Did On This Day (WWDOTD):
- Attended Science Fair Meeting
    - Must do background research and read all papers linked in sources to read
    - Reread methodology (No room for interpretation)
        - Explain and define ALL terms
    - Cite Datasets and sources used
    - Good news: finished and finalized rough methodology and plan for this project. See Methodology for more details (below)

November:

Date: November 4, 2025
People Involved: Matthew
Title: Researched papers
Topic/Section: Research
What We Did On This Day (WWDOTD):
- Researched 6 papers and added notes in the “sources to read” document
- Thought about methodology

Date: November 4-5, 2025
People Involved: Ivan
Title: Finished Methodology
Topic/Section: Writing
What We Did On This Day (WWDOTD):
- Reviewed the methodology process I did and wrote down quick jot notes to outline this to myself.
- Then, using Cole’s template, I constructed the methodology to submit to him later today, on Nov. 5
- Link: https://docs.google.com/document/d/1okoGi5FCG7hB42lzkT6OrWJf4_jDZ6aw3iJ7YtUEviY/edit?tab=t.0

Date: November 10-16, 2025
People Involved: Ivan
Title: Processed Data
Topic/Section: Data
What We Did On This Day (WWDOTD):

- Prepared to process data
    - All that's left to do is to run the processing (2 days)
- Will use Matthew's spare computer to process data

Date: November 18, 2025
People Involved: Matthew, Ivan
Title: Science Fair Meeting
Topic/Section: Planning
What We Did On This Day (WWDOTD):

- Did a progress check and divided tasks for the next week

Date: November 25, 2025
People Involved: Matthew
Title: Set up the computer
Topic/Section: SVM
What We Did On This Day (WWDOTD):

- Set up a spare Macintosh computer and cleared much of the storage
- Prepared to train SVM on that computer and process another data set for the project if time permits

Data Processing and Quality Control (Using QIIME2 and DADA2)

Date: November 25-26, 2025
People Involved: Ivan
Title: DADA2 and processing
Topic/Section: QC and Data Analysis
What We Did On This Day (WWDOTD):

- Imported all raw sequencing reads into QIIME 2 using a finalized paired-end manifest file.

- Performed quality control and visualization of read quality using demux summarize to guide trimming strategy.
- Executed the full DADA2 denoising pipeline to remove noise, correct sequencing errors, and construct amplicon sequence variants (ASVs).
    - Had to pivot from paired DADA2 to only the front DADA2, as there was not enough data and overlap to use both the front and the reverse reads. The only noticeable difference is that some of the classifications may be off; however, overall, it should not make too much of a difference and will lead to a better **quality** of results.
- Generated and reviewed three core DADA2 outputs: table.qza (feature table), rep-seqs.qza (representative sequences), and denoising-stats.qza (pipeline performance).
- Converted these .qza files into viewer-friendly .qzv visualizations and assessed read retention, sequence counts, and error-filtering efficiency.
- Standardized and cleaned the metadata file, collapsing all sample identifiers into two analytic groups: MDD and Control.
- Downloaded and added the appropriate SILVA 138 99% Naive Bayes classifier (human stool weighted), preparing the system for taxonomy assignment.
- Ran taxonomic classification to create taxonomy.qza and visualized taxon-level composition with an interactive taxa barplot (taxa-barplot.qzv).
- This was then collapsed for only the genera, as we cared specifically for the genera of each sample.
- Verified correct mapping between sample IDs and metadata group labels across all files, ensuring analytical consistency.
- Organized the project directory and established the foundation for all downstream analyses,s including alpha/beta diversity, group comparisons, and machine-learning model development.

Date: November 30, 2025
People Involved: Matthew
Title: Methodology Second Draft
Topic/Section: Methodology
What We Did On This Day (WWDOTD):

- Reread the methodology section and rewrote it into the second draft
    - Reworded for clarity of ideas and fixed many grammatical errors
    - Rewrote many of the paragraphs in the document
    - Extended some paragraphs and ideas

December:

Date: December 6, 2025
People Involved: Ivan, Matthew
Title: Data report
Topic/Section: Data
What We Did On This Day (WWDOTD):

- Analyzed all the data and created multiple graphs
- Summarized all the data and created tables to show the correlation between the varying levels of certain genus and the prevalence of MDD

Results, Statistical Tests, Tabulating, and Graphing

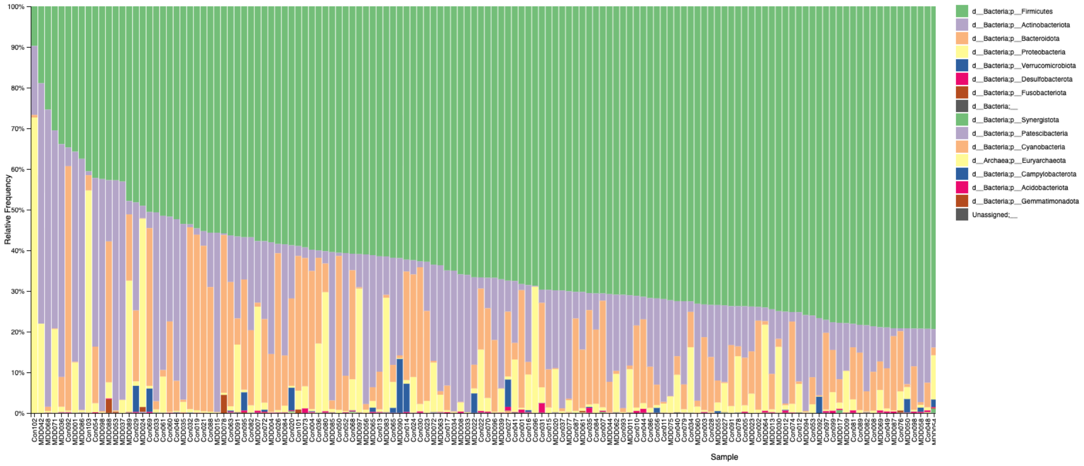
Date: December 12, 2025
People Involved: Ivan, Matthew
Title: Science Fair Meeting
Topic/Section: Planning
What We Did On This Day (WWDOTD):

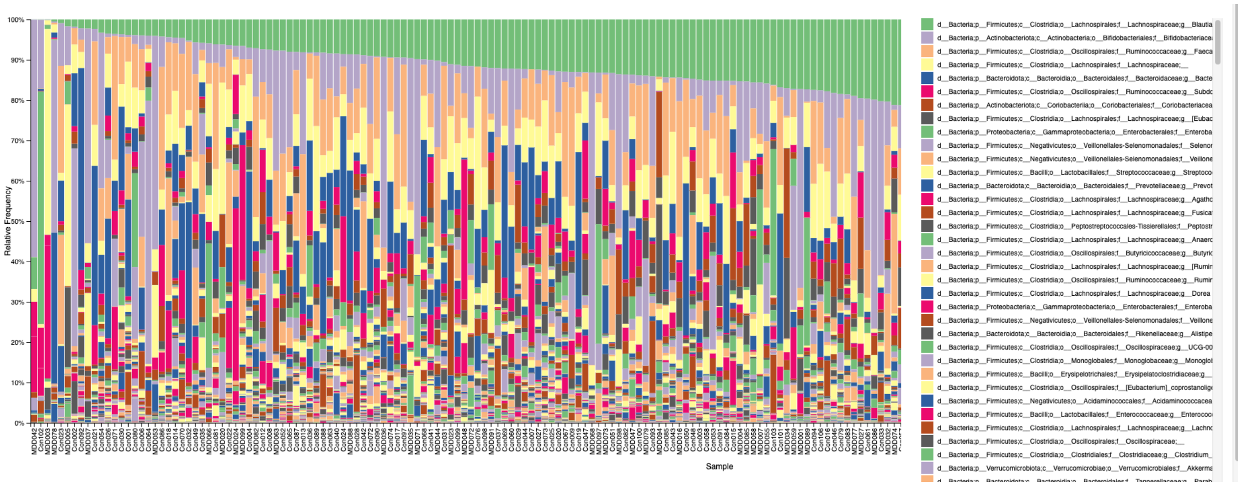
- Set goals for the following winter break
  - Matthew:
    - Finish Methodology final draft
    - Finish Background Research first draft
    - Start thinking about the presentation
  - Ivan:
    - Create the SVM
    - Start thinking about the presentation
- Showed Cole the graphs created from the data sets processed and received confirmation to start creating the SVM
- Thought about the background research and planned who will do what and where
- Updated TODO list

Date: December 15, 2025
People Involved: Ivan
Title: Graphing
Topic/Section: Analysis
What We Did On This Day (WWDOTD):

- Using the built-in methods in Terminal (Specifically ITerm2), barplots were generated. This gave the proportions of different kingdoms, phyla, classes, orders, families, genera, species, etc. for each sample.

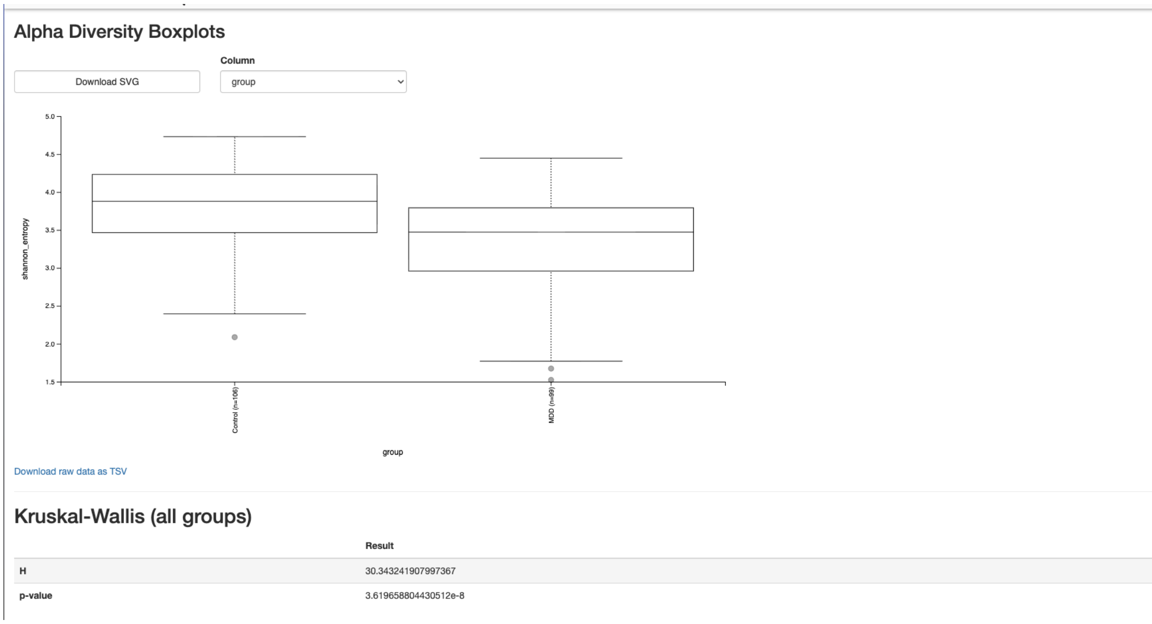
- The genus level was chosen for this project because it is commonly used and provides a strong balance between accuracy and specificity. Species-level classification tends to have lower reliability and can result in many samples remaining unclassified if they do not closely match any reference instance. Genus-level classification is broad enough to minimize unclassified cases while remaining biologically meaningful for concluding.
    - Phlum and Genera barplots respectively:





    -
    -
- This data was exported using built-in methods to CSV, which was processed in Python to give the change in relative abundance between MDD and Control for each genus. These graphs were in the form of barplots.
- Relevant biological statistics were also calculated today, such as:
    - Alpha diversity (Within group)
        - Result: p-value << 0.05, strong indicator that the diversity of each group is fundamentally different.
        - 1. Shannon index was done for each to determine diversity for each group (MDD and Control)

    - 2. Kruskal-Wallis test gives a p-value to quantify the difference in diversity (Shannon index) between both groups



Alpha Diversity Boxplots

Kruskal-Wallis (all groups)

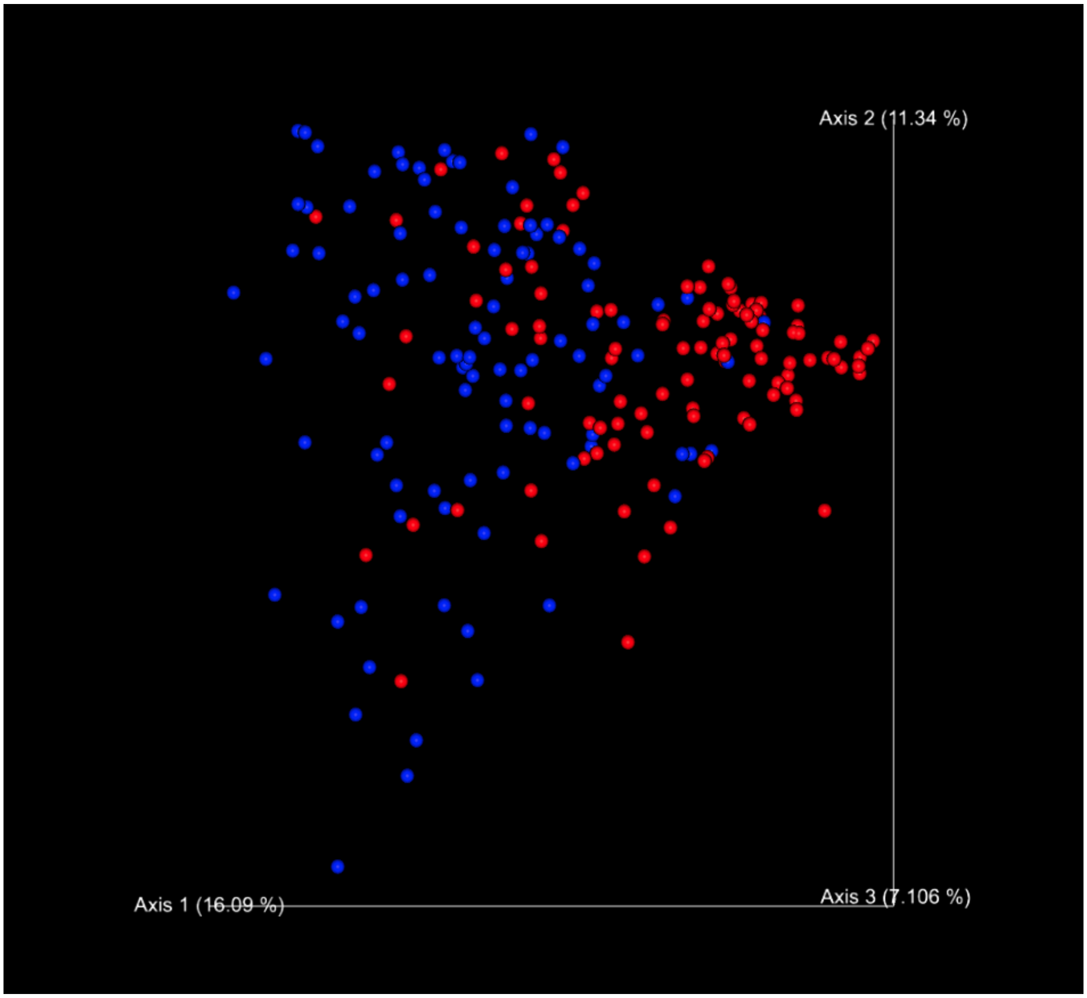
	Result
H	30.343241907997367
p-value	3.619658804430512e-8

    -

- Beta diversity (between-group)
    - Done with Bray-Curtis (sees difference based on **abundance**). Evolutionary was not taken into account because we collapsed it into a genus table earlier, and we would need a phylogenetic tree.
    - Example Bray-Curtis distance matrix:

Sample	**A**	**B**	**C**	**D**
A	0	0.3	0.7	0.6
B	0.3	0	0.8	0.5
C	0.7	0.8	0	0.2

- Each number represents the ‘distance’ or how different two samples are. 0 is the same, 1 is completely different.
- PCoA Emperor Plots are a way to graph the separation of the different classes (each dot is a sample, and how close they are is the similarity).



    - Ours showed a partially separate group, distinct but overlapping, expected in stool samples.
- PERMANOVA (ANOVA test but for Bray-Curtis plots) statistics: (Gives meaningful statistical metrics for what we are seeing in the plot):

Metric	**Value**

pseudo-F	**17.363**
p-value	**0.001**
Groups	Control (n=106), MDD (n=99)
Permutations	999

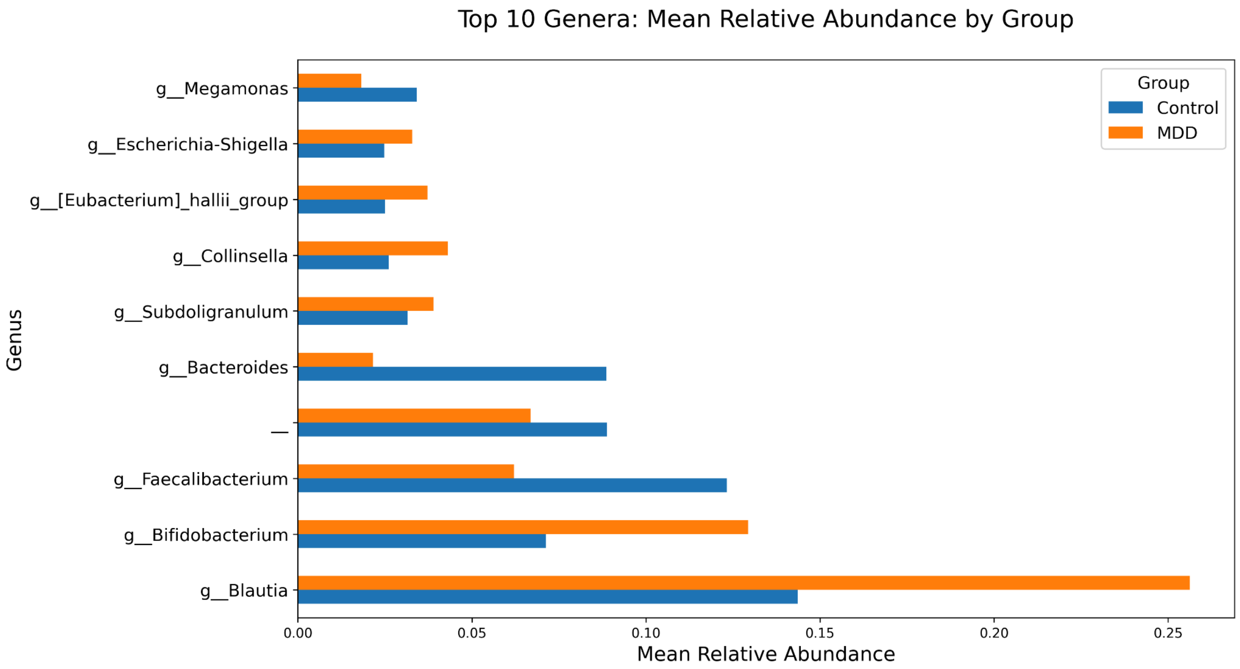
        - F: Gives separation; a higher value indicates greater separation, lower values indicate lower separation. Analogous in meaning to the F-statistic in ANOVA tests.
        - P-value: Gives the “confidence” of this separation. Smaller p-values indicate that the separation likely didn’t occur by chance, and can be concluded as “statistically significant”. The reverse is also true. This statistic is also analogous in meaning to the ANOVA test.
        - Permutations, at a high level, give the “depth” of calculation for the p-value. 999 is a robust and reliable depth.

    - P-value was 0.001, another strong indicator that, comparatively, they have different diversities.
- P-value for different genera across Control to MDD
    - Most major differences (smallest p-values)

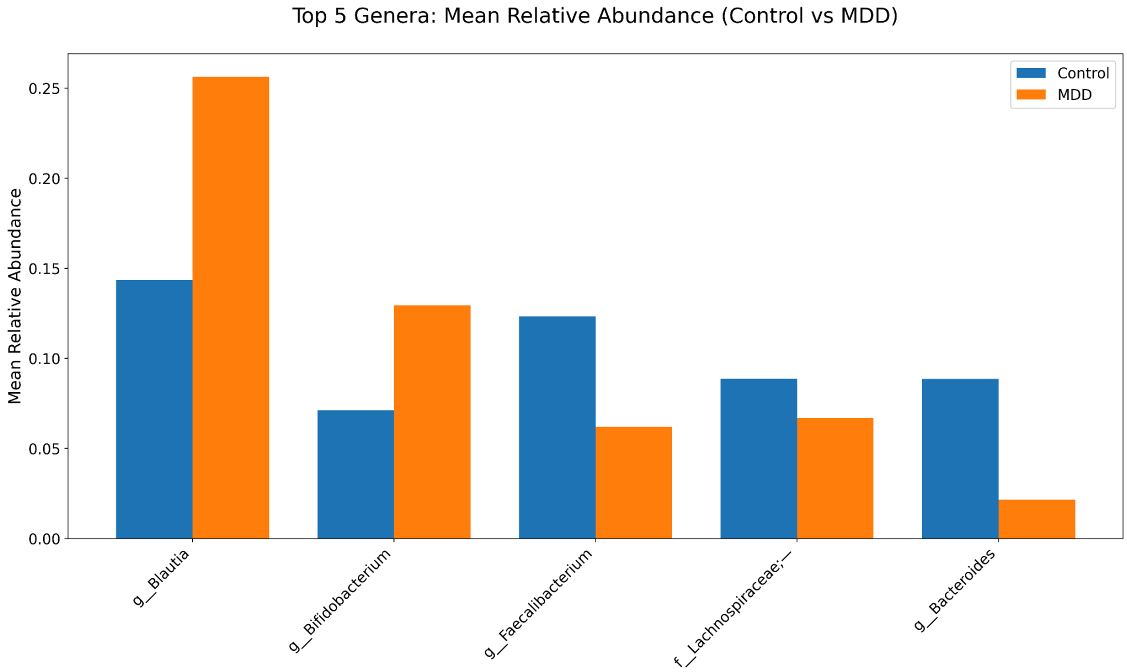
Bacteroides	9.07×10^{-14}
Faecalibacterium	1.39×10^{-7}
Parabacteroides	1.30×10^{-7}

Blautia	1.13×10^{-5}
Alistipes	1.45×10^{-5}

    - The above, highlighted, have the greatest relative abundance and smallest p-value. These make the strongest individual biomarkers as they are easy to detect, and their dysregulation is also detectable.
- Relative abundance for different genera
    - The top 10 and top 5 genera’ relative abundance variation was graphed in Pytho.n
    - Top 10:



- Top 5:



SVM Coding

All Code can be found at github repository:
Date: December 19, 2025
People Involved: Ivan
Title: SVM Coding
Topic/Section: Analysis
What We Did On This Day (WWDOTD):

- On this day, I converted the relevant data into csv using QIIME2's built-in csv export
- This was then fed into scikit-learn’s built-in SVC class (from scikit.svm import SVC)
  - By first splitting the data into training (X_train, y_train) and test (X_test, y_test)
  - I used an SVC (Support Vector Class) with these settings (how it will learn)
    - kernel = “rbf” ‘exports’ the data into higher dimensional space to draw the “line” to separate the classes, which would be nonlinear in the regular space.
    - C = 1 balances misclassifying (too simple boundary which gets many misclassified) vs. overfitting (way too complicated boundary in order ot find a pattern) (this is the default)

    - Gamma = “scale”, how much individual points influence the final boundary w.r.t. Distance. Too high leads to overfitting, too low leads to underfitting. “Scale” lets the gamma adapt in order to avoid extremes.
    - Probability = True (Allows the SVM, when giving the predicted class, to give the probability of it being correct)
    - Random_state = 42. SVM training is random; by specifically choosing a specific instance of “training” (by convention 42), other scientists can get the same model by letting *their* random_state be 42 as well.
- Once trained with X_train, y_train using svm.fit(), it is saved in a .joblib so it doesn’t have to be trained each time it needs to be used.

Date: December 20, 2025
People Involved: Ivan
Title: SVM Coding
Topic/Section: Analysis
What We Did On This Day (WWDOTD):

- On this day, I converted the relevant data into csv using QIIME2's built-in csv export
- This was then fed into scikit-learn’s built-in SVC class (from scikit.svm import SVC)
  - By first splitting the data into training (X_train, y_train) and test (X_test, y_test)
  - I used an SVC (Support Vector Class) with these settings (how it will learn)
    - kernel = “rbf” ‘exports’ the data into higher dimensional space to draw the “line” to separate the classes, which would be nonlinear in the regular space.
    - C = 1 balances misclassifying (too simple boundary which gets many misclassified) vs. overfitting (way too complicated boundary in order ot find a pattern) (this is the default)
    - Gamma = “scale”, how much individual points influence the final boundary w.r.t. Distance. Too high leads to overfitting, too low leads to underfitting. “Scale” lets the gamma adapt in order to avoid extremes.
    - Probability = True (Allows the SVM, when giving the predicted class, to give the probability of it being correct)
    - Random_state = 42. SVM training is random; by specifically choosing a specific instance of “training” (by convention 42), other scientists can get the same model by letting *their* random_state be 42 as well.
    -

  - X, y = load_data(DATASETS[mode])
  -
  - X_train, X_test, y_train, y_test = train_test_split(
  - X, y,
  - test_size=0.25,
  - stratify=y,
  - random_state=42
  -)
  -
  - scaler = StandardScaler()
  - X_train_scaled = scaler.fit_transform(X_train)
  - X_test_scaled = scaler.transform(X_test)
  -
  - svm = SVC(
  - kernel="rbf",
  - C=1.0,
  - gamma="scale",
  - probability=True,
  - class_weight="balanced",
  - random_state=42
  -)
  -
  - svm.fit(X_train_scaled, y_train)
  -
  - joblib.dump(svm, MODELS[mode]["model"])
  - joblib.dump(scaler, MODELS[mode]["scaler"])
  -
  - evaluate_model(svm, scaler, X_test, y_test, mode)# This is a function defined in the same file
  - “””
    - Prints metrics using sklearn.metrics
  - “””
  -

- Once trained with X_train, y_train using svm.fit(), it is saved in a .joblib so it doesn’t have to be trained each time it needs to be used.
- Added a way for users to interact with the model using if __name__ == “__main__”:
  Using Standard Scaler on the input (after formatting it correctly, key outliers

        are done usingbuilt-inn methods from the SVC:
        decision = svm.decision_function(user_scaled)[0]
    - prob = svm.predict_proba(user_scaled)[0][1]
    - pred = svm.predict(user_scaled)[0]

Date: December 20, 2025 - January 2, 2026
People Involved: Matthew
Title: Background Research
Topic/Section: Research
What We Did On This Day (WWDOTD):

- Completed almost all of the background research over the course of multiple days
- Formalised the notes we took on a scientific paper and formalised it into a full background research paper
- Sent the paper to Ivan to do a second reread before I review the final draft

January

Website Coding

All Code can be found at this github links:

-

Date: January 03, 2026
People Involved: Ivan
Title: Started website
Topic/Section: Website
What We Did On This Day (WWDOTD):

- Worked on the basic elements and content sections of the homepage
    - Large Content
        - Hero section (not a stereotypical hero, but it serves as an “introduction” to the page. Format:

            - [div (light-gray, translucent)]
            - [Title (Bold Lato]
            - [Text (Nunito spaced)]
            - [/div]
        - Gradient sections
            - Background image of an appropriate gradient (see colour scheme), with white title and text, according to the font scheme, and an image
- Colour scheme decided (Red-yellow, purple-pink, blue-green)
- Fonts decided (Lato bold for titles, Nunito spaced for text)
- Work done on VSCode, this current dummy file structure
    - SVM_Website
        - Svm_backend (currently empty)
        - Svm_frontend
            - Svm_home.html
            - Styles.css (currently empty)
- All CSS is done in the fileand will be transferred once finalized.
- Most of the time was spent perfecting the header; in this way, more pages have been planned
    - Svm_try (Users can try the model here)
    - Svm_methods (Brief, infographic-heavy methodology for users)
    - Svm_limits (Discussing important limits, guidelines, use, and other important cautions and tips).
    - Svm_results (The statistics and results of the SVM, also infographic-heavy)

Date: January 05, 2026
People Involved: Ivan
Title: Started website
Topic/Section: Website
What We Did On This Day (WWDOTD):

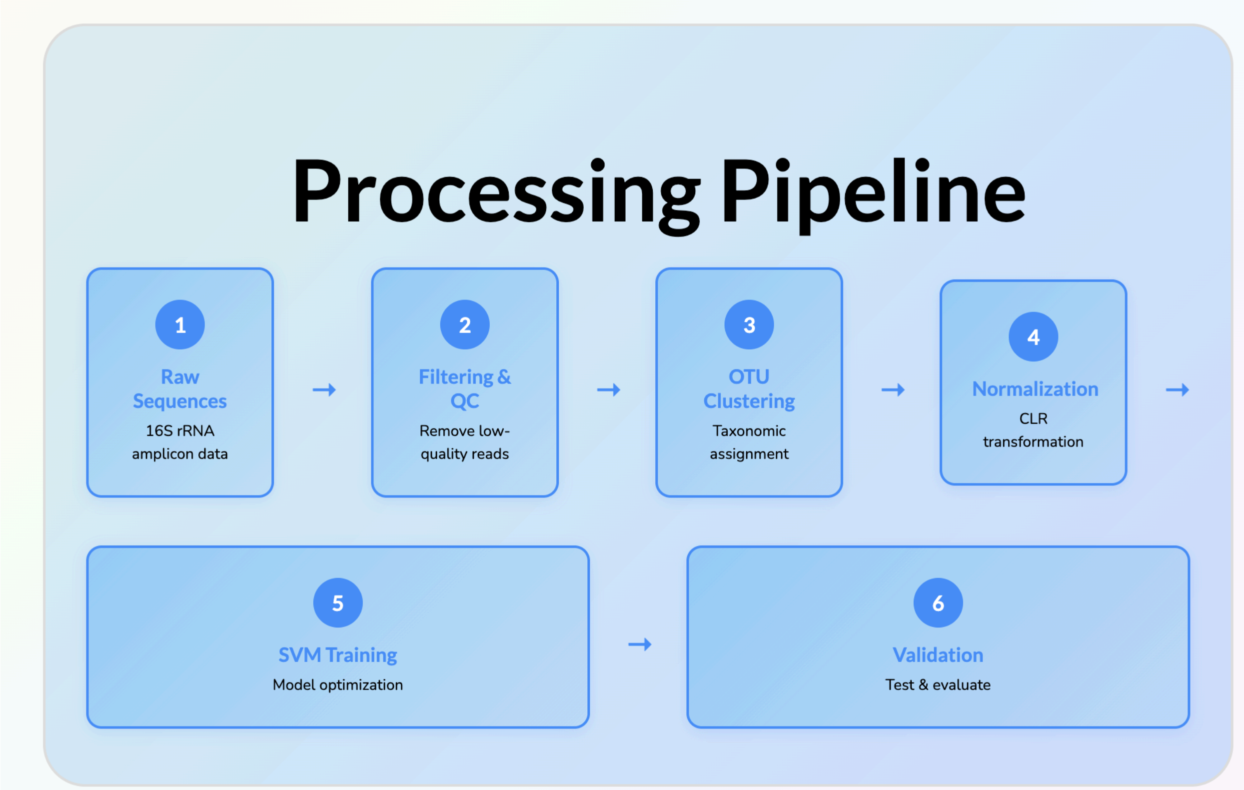
- Finished svm_home's hero section and 3 coloured sections (test_bg, test_ry, and test_pp)
- Added In-file CSS to styles.css
- Created dummy files for all frontend
- Decided to use Netlify's free service to host the frontend and Render to host the backend
- Decided to have sections ("cards" will be how they will be referred to), to aesthetically display information
- Pseudo-coded the cards
    - Display: flex

    - Border: var(--xxx) 3px solid;
    - Background: rgb(xxx)
    - Etc.
- Finalized card style
    - [div with above characteristics such as semi-transparent background and smooth animations]
        - [Emoji or Number in icon-circle, tilt on hover]
        - [Title]
        - [Text]
    - [/div]

Date: January 06, 2026
People Involved: Ivan
Title: Started website
Topic/Section: Website
What We Did On This Day (WWDOTD):

- Converted the Pseudo-code for the cards into actual in-file CSS for svm_limits, svm_methods, and svm_results
- Using variations and additions to this baseline, many other similar styles were able to be created relatively quickly
    - Cards for each gradient of the colour scheme [USED]
    - Hero-section cards [UN-USED, overcomplicates]
    - Timeline Cards [USED]
    - Nested Cards [SEMI-USED, cards are nested in sections]
- These were used to colour the webpage and give information aesthetically to the user.
- Added cards to svm_home for a quick-view on the purpose and what the website has to offer
- Added a slight fade-in animation on load for the majority of elements, and a rotate animation for the icon (primarily emoji, however, numbers too for timeline cards), on hover.
- Various examples of how cards are used are added below [UPDATED WITH MARGIN AND RESIZING CHANGES]:

- Pipeline Cards [UPDATE: Added animation to arrows]:



Processing Pipeline

1 **Raw Sequences** 16S rRNA amplicon data →

2 **Filtering & QC** Remove low-quality reads →

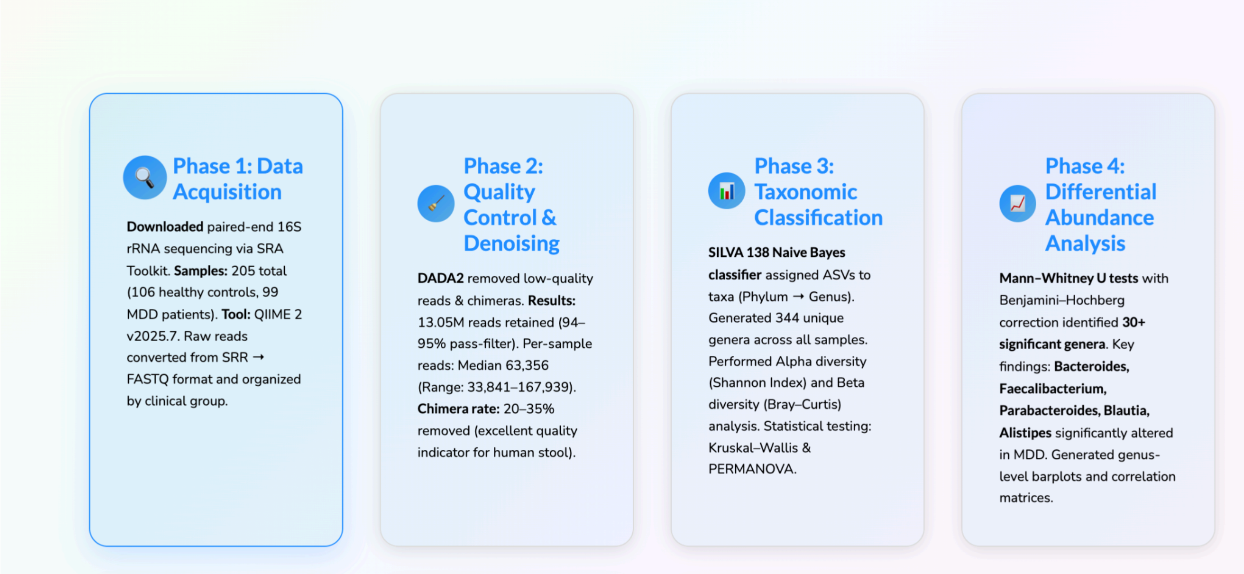
3 **OTU Clustering** Taxonomic assignment →

4 **Normalization** CLR transformation →

5 **SVM Training** Model optimization →

6 **Validation** Test & evaluate

- Regular Cards:



🔍 Phase 1: Data Acquisition

Downloaded paired-end 16S rRNA sequencing via SRA Toolkit. **Samples:** 205 total (106 healthy controls, 99 MDD patients). **Tool:** QIIME 2 v2025.7. Raw reads converted from SRR → FASTQ format and organized by clinical group.

🧹 Phase 2: Quality Control & Denoising

DADA2 removed low-quality reads & chimeras. **Results:** 13.05M reads retained (94–95% pass-filter). Per-sample reads: Median 63,356 (Range: 33,841–167,939). **Chimera rate:** 20–35% removed (excellent quality indicator for human stool).

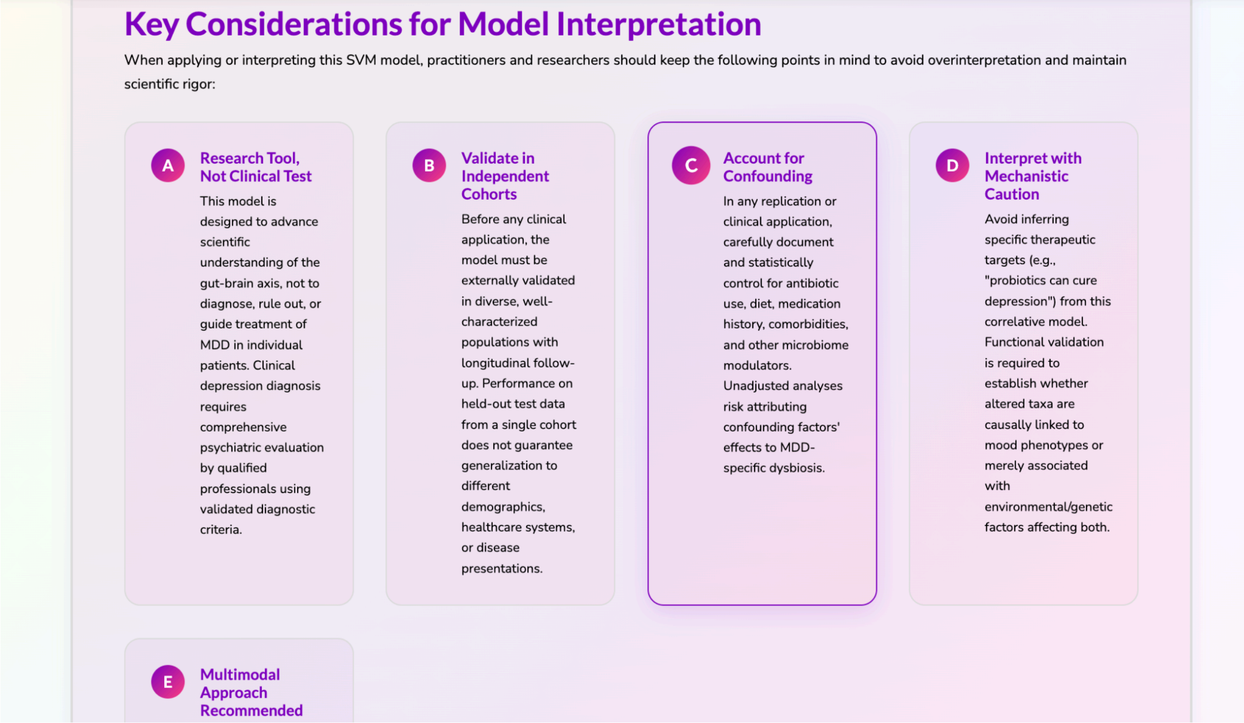
📊 Phase 3: Taxonomic Classification

SILVA 138 Naive Bayes classifier assigned ASVs to taxa (Phylum → Genus). Generated 344 unique genera across all samples. Performed Alpha diversity (Shannon Index) and Beta diversity (Bray–Curtis) analysis. Statistical testing: Kruskal–Wallis & PERMANOVA.

📈 Phase 4: Differential Abundance Analysis

Mann–Whitney U tests with Benjamini–Hochberg correction identified **30+ significant genera**. Key findings: **Bacteroides, Faecalibacterium, Parabacteroides, Blautia, Alistipes** significantly altered in MDD. Generated genus-level barplots and correlation matrices.

- Nested cards:



Key Considerations for Model Interpretation

When applying or interpreting this SVM model, practitioners and researchers should keep the following points in mind to avoid overinterpretation and maintain scientific rigor:

A **Research Tool, Not Clinical Test**

This model is designed to advance scientific understanding of the gut-brain axis, not to diagnose, rule out, or guide treatment of MDD in individual patients. Clinical depression diagnosis requires comprehensive psychiatric evaluation by qualified professionals using validated diagnostic criteria.

B **Validate in Independent Cohorts**

Before any clinical application, the model must be externally validated in diverse, well-characterized populations with longitudinal follow-up. Performance on held-out test data from a single cohort does not guarantee generalization to different demographics, healthcare systems, or disease presentations.

C **Account for Confounding**

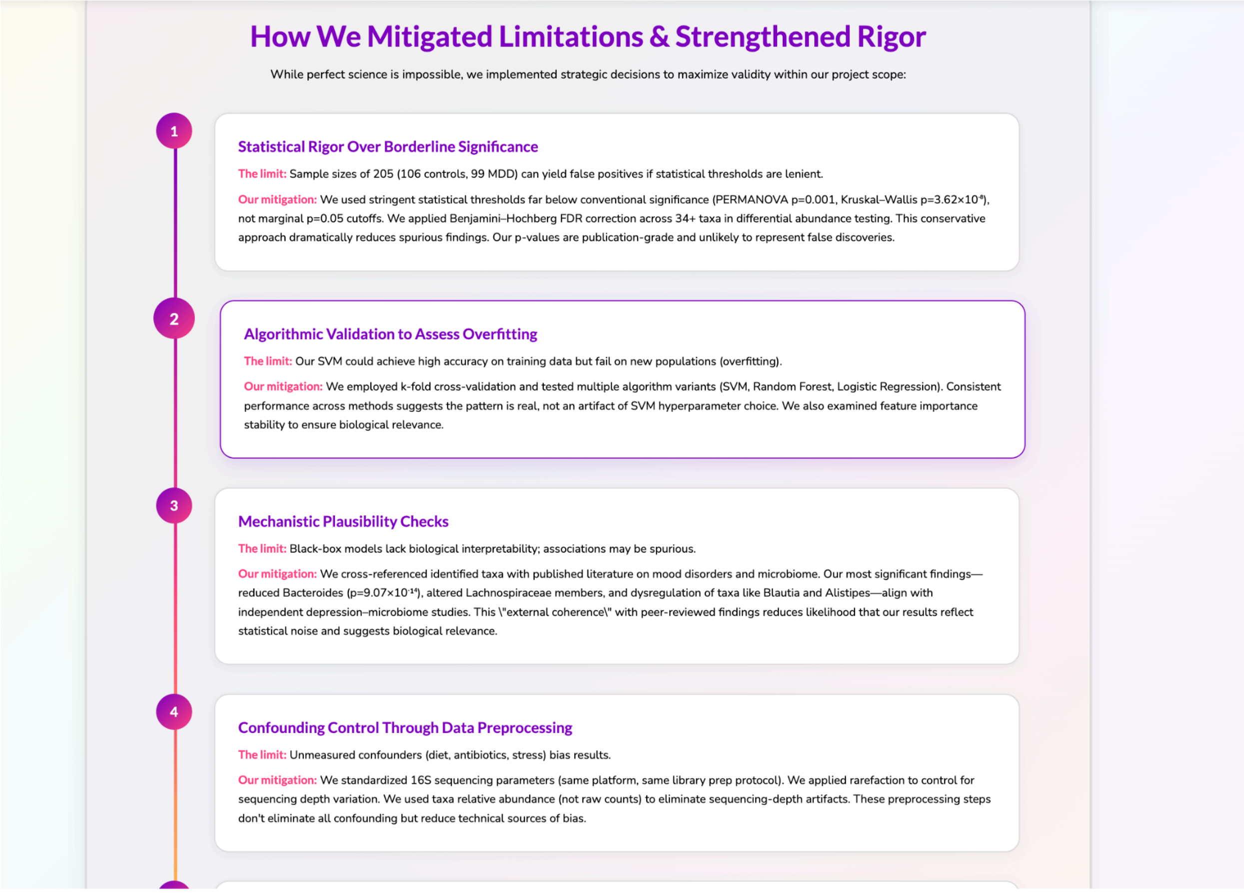
In any replication or clinical application, carefully document and statistically control for antibiotic use, diet, medication history, comorbidities, and other microbiome modulators. Unadjusted analyses risk attributing confounding factors' effects to MDD-specific dysbiosis.

D **Interpret with Mechanistic Caution**

Avoid inferring specific therapeutic targets (e.g., "probiotics can cure depression") from this correlative model. Functional validation is required to establish whether altered taxa are causally linked to mood phenotypes or merely associated with environmental/genetic factors affecting both.

E **Multimodal Approach Recommended**

- Timeline cards:



How We Mitigated Limitations & Strengthened Rigor

While perfect science is impossible, we implemented strategic decisions to maximize validity within our project scope:

1 **Statistical Rigor Over Borderline Significance**

The limit: Sample sizes of 205 (106 controls, 99 MDD) can yield false positives if statistical thresholds are lenient.

Our mitigation: We used stringent statistical thresholds far below conventional significance (PERMANOVA p=0.001, Kruskal–Wallis p=3.62×10^{-8}), not marginal p=0.05 cutoffs. We applied Benjamini–Hochberg FDR correction across 34+ taxa in differential abundance testing. This conservative approach dramatically reduces spurious findings. Our p-values are publication-grade and unlikely to represent false discoveries.

2 **Algorithmic Validation to Assess Overfitting**

The limit: Our SVM could achieve high accuracy on training data but fail on new populations (overfitting).

Our mitigation: We employed k-fold cross-validation and tested multiple algorithm variants (SVM, Random Forest, Logistic Regression). Consistent performance across methods suggests the pattern is real, not an artifact of SVM hyperparameter choice. We also examined feature importance stability to ensure biological relevance.

3 **Mechanistic Plausibility Checks**

The limit: Black-box models lack biological interpretability; associations may be spurious.

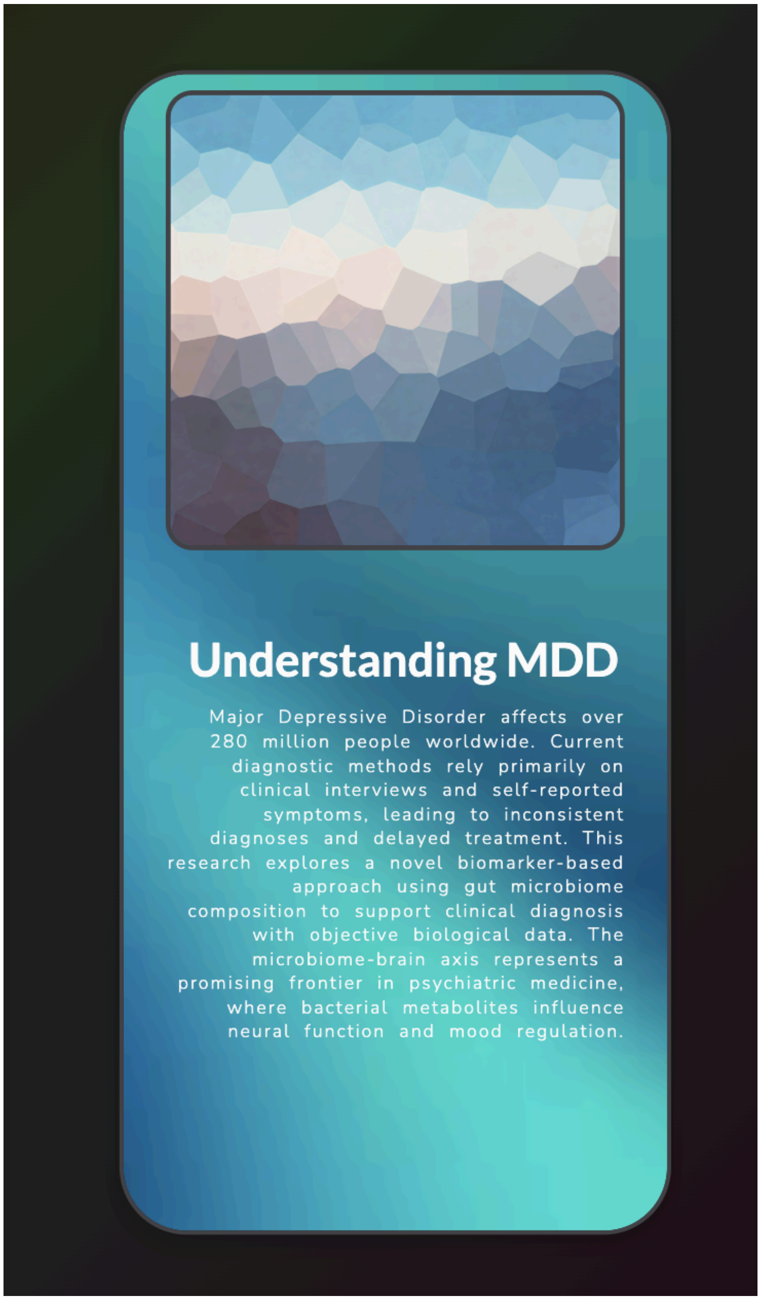
Our mitigation: We cross-referenced identified taxa with published literature on mood disorders and microbiome. Our most significant findings—reduced Bacteroides (p=9.07×10^{-14}), altered Lachnospiraceae members, and dysregulation of taxa like Blautia and Alistipes—align with independent depression–microbiome studies. This \"external coherence\" with peer-reviewed findings reduces likelihood that our results reflect statistical noise and suggests biological relevance.

4 **Confounding Control Through Data Preprocessing**

The limit: Unmeasured confounders (diet, antibiotics, stress) bias results.

Our mitigation: We standardized 16S sequencing parameters (same platform, same library prep protocol). We applied rarefaction to control for sequencing depth variation. We used taxa relative abundance (not raw counts) to eliminate sequencing-depth artifacts. These preprocessing steps don't eliminate all confounding but reduce technical sources of bias.

- Added a prototype Pros/Cons section, which will be added to svm_limits in the future.
- Added a prototype large timeline section for svm_methods (not to be confused with the timeline cards
    - Bug:
        - Goal: Sections appear alternately on either side of the middle line of the timeline
        - Bug: Only appearing on one side of the line (primarily the right)
- Attempted adding @media atrule for screen size to make the website fluid for different screen sizes
    - Bug:
        - Goal: For test_xx, on sufficiently small screen sizes, the accompanying image should appear above the text, as so:

- 

Understanding MDD

Major Depressive Disorder affects over 280 million people worldwide. Current diagnostic methods rely primarily on clinical interviews and self-reported symptoms, leading to inconsistent diagnoses and delayed treatment. This research explores a novel biomarker-based approach using gut microbiome composition to support clinical diagnosis with objective biological data. The microbiome-brain axis represents a promising frontier in psychiatric medicine, where bacterial metabolites influence neural function and mood regulation.

Date: January 09, 2026
People Involved: Ivan
Title: Started website
Topic/Section: Website
What We Did On This Day (WWDOTD):

- Fixed the timeline error and added the prototype timeline with dummy text to the website. Unfortunately, I do not have records of the prototype timeline; however, the finished timeline has been added instead. The main difference (aside from margin realignment) is the text being filled in.

- 

0

Goal & Data Origin

Goal: Compare MDD vs Control microbiomes, identify differentially abundant taxa, build SVM classifier, deploy interactive tool. **Data:** Public NCBI SRA — paired-end 16S rRNA sequencing, 205 samples (106 Control, 99 MDD).

1

SRA Acquisition & Conversion

Tool: SRA Toolkit. **Tasks:** Download SRR files → Convert to FASTQ paired-end (_R1, _R2) → Organize into fastq_MDD/ and fastq_Control/ directories. **Result:** ~100K reads per sample, ready for QC.

- ● Download all SRR files from SRA
- ● Fastq-dump: SRR → paired FASTQs

1a

Metadata Curation

Create manifest (sample-id, group) and metadata files. Verify SRR accessions match clinical data. Label: MDD or Control for each sample.

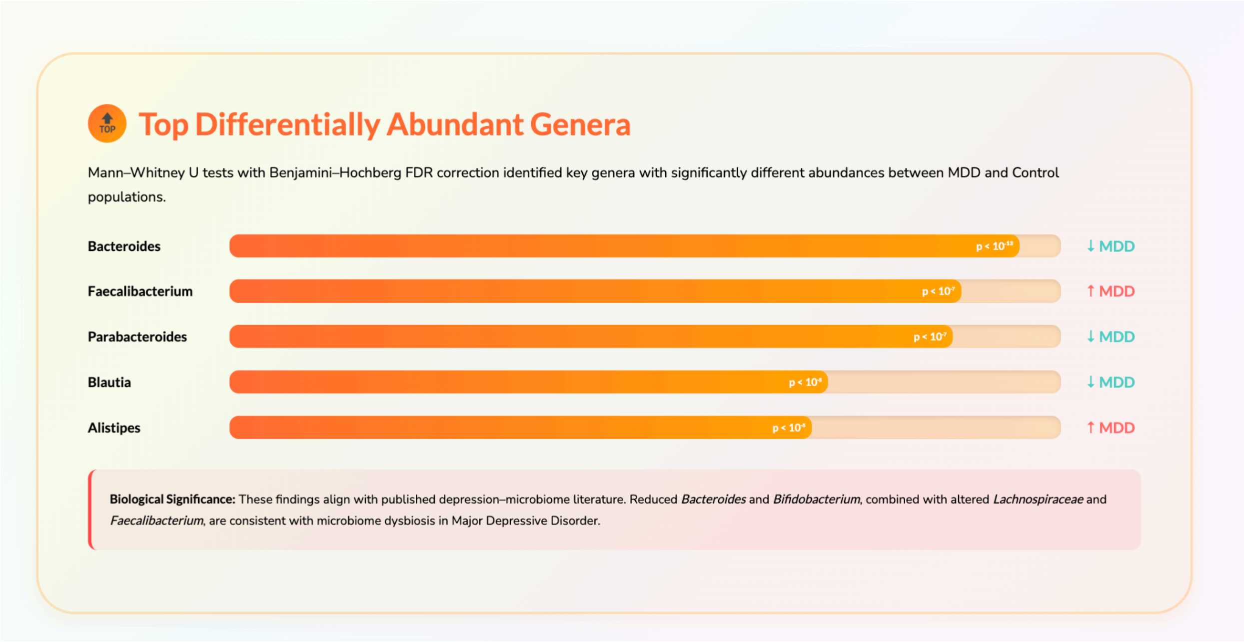
1b

Data Organization

Sort FASTQ files into directories by

- Added a light-rainbow background to fit the **gradient-and-emoji-heavy, modern theme.**
- Added template for gradient bar graphs, used in methods and results, using flexbox curved rectangles and clever positioning.
    - In hindsight, just using an image would have saved more time, and perhaps could have looked better.

    - More research could have been done on available iframes as well

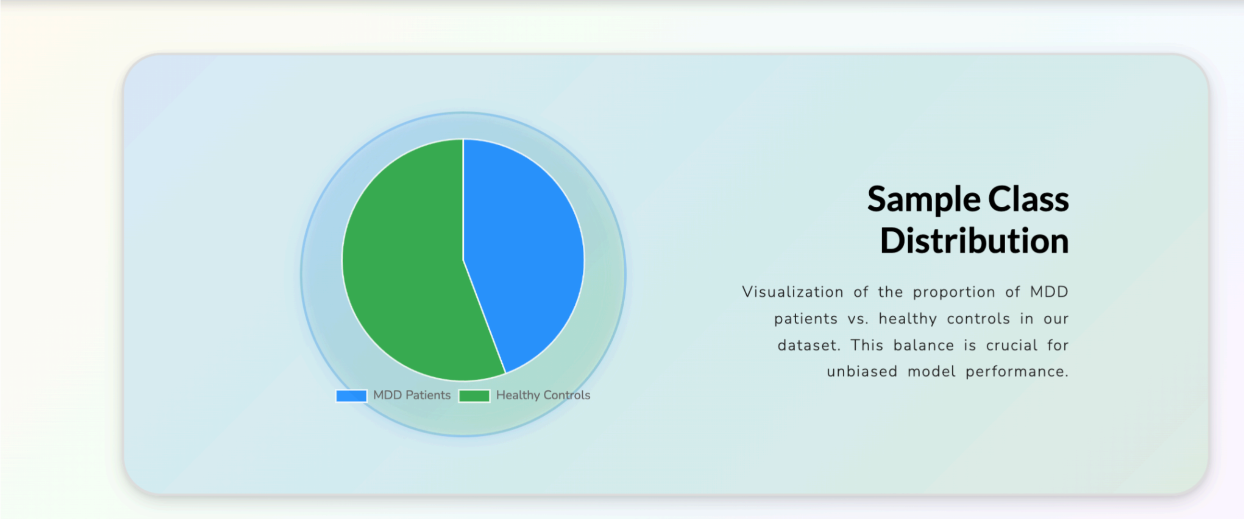


Top Differentially Abundant Genera

Mann–Whitney U tests with Benjamini–Hochberg FDR correction identified key genera with significantly different abundances between MDD and Control populations.

Biological Significance: These findings align with published depression–microbiome literature. Reduced *Bacteroides* and *Bifidobacterium*, combined with altered *Lachnospiraceae* and *Faecalibacterium*, are consistent with microbiome dysbiosis in Major Depressive Disorder.

- Added template for pie graphs for svm_methods
    - In hindsight, researching for and using an available iframe would have made the job much easier.
    - Images wouldn't have worked, as some interactivity would have been lost.



Sample Class Distribution

Visualization of the proportion of MDD patients vs. healthy controls in our dataset. This balance is crucial for unbiased model performance.

    -

Date: January 11-17, 2026
People Involved: Ivan
Title: Started website
Topic/Section: Website
What We Did On This Day (WWDOTD):

- Added more varieties of cards
    - Non-uniform borders

        - Interactive
        - Non-interactive
- Implemented timeline for Methodology (above) and Pros/Cons for Limits
- For all varieties of cards and elements, the placeholder Lorem Ipsum text has been replaced by actual informational text using my own words, Methodology, Background Research, and statistics from the Progress Report
- Fixed animation and other small bugs
- Uploaded to GitHub as a repository, 'Svm frontend.'
    - I then hooked that toNetlifyy (a free service where you can host frontend), creating a dummy site until the backend is hooked up.
- I downloaded the VS Code extension 'Prettier' to clean up indent issues and messy code.
- I had bugs with the 'Live Share' and 'Live Server' extension not working and giving errors. I restarted VS Code and my computer, which fixed the issue.
- Added the front-end JavaScript and a call to the backend in svm_try.html
    - This step took the majority of the above-listed time, as I had to learn more complicated JavaScript and HTML elements
    - I learned how to create scrollable checkbox elements (for general checkboxing)
    - Luckily, many functionality was similar to Python (ex. .join(""), try {} catch () {}, .length), something I am very well versed in, functions such as .map() I had to dig deep into for my needs
    - I learned more about .addEventListener, and its inputs such as "input", "change",
    - I learned more about async and await, especially when calling backendURL (currently a placeholder) using fetch(), a function I had never used before.
    - After searching documentation for a long while, I figured out about .querySelectorAll() to help me create the selected Array.
    - **Key Things I Learned**
        - Promises, async, and await
            - Promise - a placeholder, something that will come in the future (takes time). Ex. When you place an order at a busy restaurant in real life, you get a tticket That tells you that you will get a seat eventually. This is the placeholder.
            - An async function (defined asynch function functionName() handles the Promise, allowing for functions to "pause" and wait until a Promise is fulfilled (a seat gets available). It allows a function to break, ie, become asynchronous.

        - Await is the keyword inside the function that lets you wait. Once the Promise is fulfilled, the variable you are trying to set, for example, gets stored with that value, and the function continues (ex. const seat = await seatAvaiable();, here seatAvaiable is waiting until a seat is available).
    - .trim() is a relatively simple element, which allows you to trim whitespace from a string on both sides. It allowed me to simplify my code a ton!
    - .map() allows me to transform an array by applying a function to each element, returning a **new** array (I thought it modified, which created a since-fixed bug). This was way more efficient than the nested loop nonsense I was initially trying to accomplish with the selected Array
    - .addEventListener is fundamental to the script, allowing me to trigger a function when an event occurs (allowing me to dynamically update the Hints for the number of genera selected and the number of reads inputted.
    - .querySelectorAll (a variation of .querySelector) gives all HTML elements of a given tag (ex. <p>), child (ex. div > p), id, or class. I used this in an Array in order to get an array of selected values for the multiple option section.

Date: January 19 - 24, 2026
People Involved: Ivan
Title: Integrated and Created Website Backend
Topic/Section: Website
What We Did On This Day (WWDOTD):

- I learned that to host my model, I need to use something called FastAPI to connect the model with Render (Host's backend)
    - I needed requirements.txt (all Python libraries needed to run main.py (which handles and unboxes the .joblibs, handles the POST request by fetch(), and connects it back to the Netlify website))

```
scikit-learn

pandas

joblib

fastapi

pydantic

numpy

uvicorn
```

    - The only library I had to add later (everything else is pretty standard for this type of job) is uvicorn, allowing me to test the model locally, as a sanity check, before publishing to Render
  - main.py (explained above)
    - Key components and what I learned
      - @app.post("/predict") handles the '/predict' POST call.
      - The rest of the code is completed similarly to the if __name__ == "__main__": in the training file.
  - Genera_order.joblib and svm_model_all.joblib (the former makes sure the feature vector sent matches the input the model needs, while the latter holds the model itself.)
- After FastAPI was done, I had to create a GitHub repository and link that into Render.
  - Similar standard steps as with creating the frontendGitHubb repo
  - Created, added, pushed, and committed. Then it was ready to be linked.
- Next, using Render's free plan, I hosted the GitHub repository, with these presets (This step was much simpler than I had expected, taking a good 30 minutes for the entire Render setup to be completed).
  - Name: svm_backend
  - Branch: Main
  - Environment: Python 3
  - Build Command: pip install -r requirements.txt
- Finally, I linked the render service link (.onrender.com) to the const backendURL in the JS in svm_try, and after some minor troubleshooting, it linked cleanly.

Date: January 19 - 24, 2026
People Involved: Ivan
Title: The Final Step
Topic/Section: Website
What We Did On This Day (WWDOTD):

- I edited grammar mistakes manually and used a grammar checker extension in VS Code. I used the Progress Report to fact-check all added statistics to prevent minor errors.
- I deleted a card element in svm_home.html, which initially gave information on each page in the website, and its purpose. I felt that was unnecessary, instead replacing it with buttons to access each website at the bottom, a footer of sorts.
- 🎉🥳Published and checked out the website (svm-mdd.netlify.app),

Presentation and Poster

Date: January 29, 2026
People Involved: Ivan, Matthew, Cole
Title: Presentation Planning
Topic/Section: Presentation, Poster
What We Did On This Day (WWDOTD):

- Discussed presentation plan
  - 4 Minutes: Background Information, Methodology
  - 6 Minutes: Results, Limitations, and Discussion
  - The final presentation may have an additional one-minute presentation
- Planned out poster
  - Created a document on Canva
    - Plan to print the poster at a store instead of a trifold
    - Planned out the layout

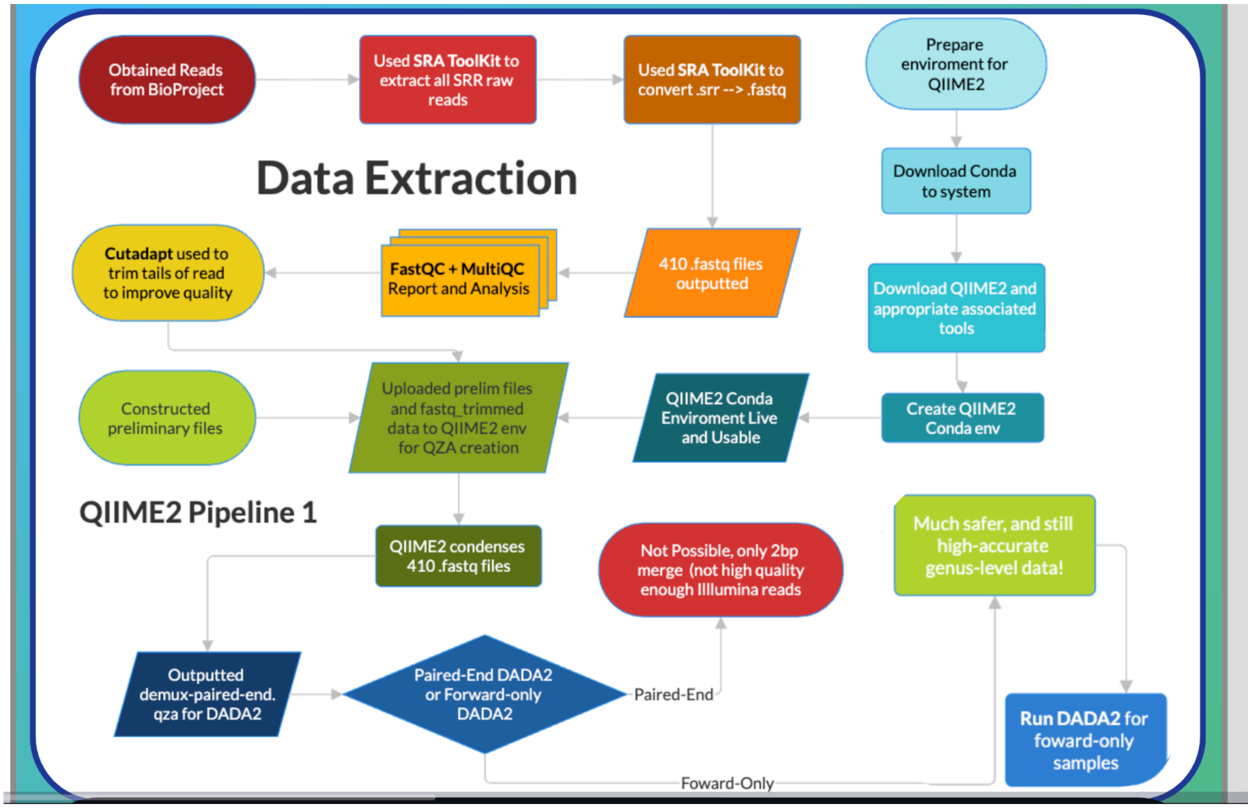
Febuary

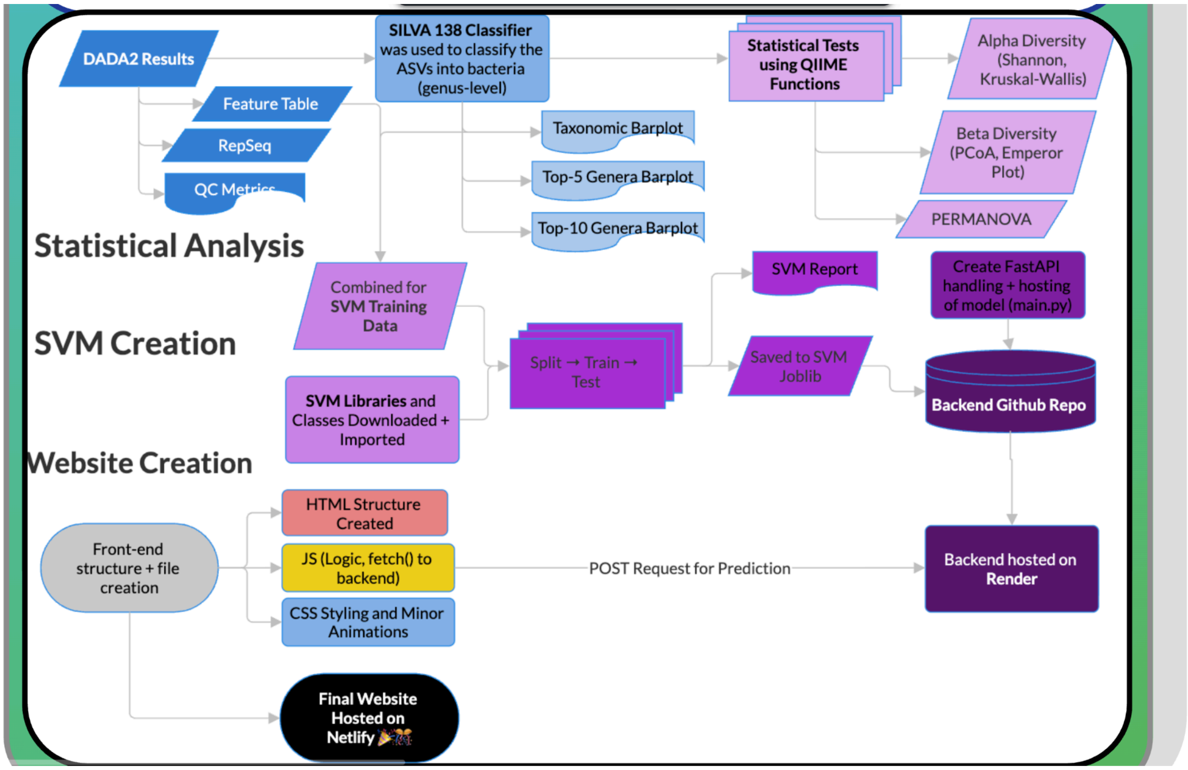
Date: February 6, 2026
People Involved: Ivan, Matthew, Cole
Title: Presentation Planning
Topic/Section: Presentation, Poster
What We Did On This Day (WWDOTD):

- Discussed plans for presentation and gave a lot of tips and ideas for how to present properly at the science fair
- Matthew will be the only one presenting at the regional CYSF
- Updated the poster layout to include colours, titles, and sections
- Divided up tasks and updated the TODO list

Date: Feb 11–13, 2026
People Involved: Ivan, Matthew,
Title: Poster Planning
Topic/Section: Poster
What We Did On This Day (WWDOTD):

- 11: Using Cole's template, adjusted and personalized colours using project style (Clear glass, blue-green gradient, red-yellow gradient, purple-pink gradient)
- 12:
  - Ivan: Added a flowchart for the methodology
  - Matthew: Added Background + Objectives
  - Ivan: Edited Matthew's text

- 

- 

- 13:
    - Ivan: Added graphs and tables to the Results section, finishing it

Date: February 17, 2026

People Involved: Matthew
Title: Script
Topic/Section: Presentation
What We Did On This Day (WWDOTD):

- Began to write the script for the video presentation
- Wrote the sections for background research and problem
    - Divided the speaking points between me and Ivan (With Cole's advice)
        - 30s Intro
        - 1:45 Background
        - 30s Objective
        - 3-4min Methodology
        - 3-4min Results
        - 1min Discussion
        - 1min Conclusion + Impacts
    - Ivan Raizada edited the grammar and conventions

Date: February 19-20, 2026
People Involved: Ivan
Title: Discussion and Limitations
Topic/Section: Poster
What We Did On This Day (WWDOTD):

- 19:
    - Added references to Methodology (link-format)
    - Added acknowledgements to the poster
    - Added the discussion section based on the previous writing
- 20:
    - Added Limitation Section
    - Added Conclusion section
    - Reviewed grammar for added sections
    - Increased margins, spacing, and font size for readability

Date: February 21, 2026
People Involved: Ivan
Title: Final Touches
Topic/Section: Poster
What We Did On This Day (WWDOTD):

- Added a better-formatted Alpha Diversity Shannon Index Box plot (colours adjusted through adjustment in Python)
- Final Touches
    - Made all titles in TitleCase
    - Standardized font size

    - Fixed an issue with borders
    - Added prefix to titles (Figure X: and Table X:)
    - Added links to References (just link so far)
- Poster overall **95%** done

Date: February 23, 2026
People Involved: Ivan
Title: Editing + Printing
Topic/Section: Poster
What We Did On This Day (WWDOTD):

- Edited grammar issues and final colour choices
- Printed using Staples Same-Day printing, 48''x36.''
- Attached to a foam-board base

Script and Portal

Date: February 24, 2026
People Involved: Ivan
Title: Methodology + Results
Topic/Section: Script
What We Did On This Day (WWDOTD):

- Using the existing Methodology section, the Methodology section of the script was added
- The Results section was also added to the script
- Decided on using Prezi for the presentation, on Cole's advice

Date: February 25-26, 2026
People Involved: Ivan
Title: Script Editing + Conclusion
Topic/Section: Poster
What We Did On This Day (WWDOTD):

- Added the Conclusion section of the script
- Used Grammarly to fix grammar errors and weak sentences
- Began working and editing the script based on Cole's added comments

Date: February 27-28, 2026
People Involved: Ivan
Title: Script Editing + Problem
Topic/Section: Poster
What We Did On This Day (WWDOTD):

- Added the Methodology section to the CYSF Portal
- Worked on + Finished the Problem section for the CYSF Portal

        - Revised using Grammarly
    - Finished all of Cole’s comments (having deleted some that I thought would be alright without) by the 28th.

Date: March 1, 2026

People Involved: Ivan

Title: Prezi

Topic/Section: Presentation

What We Did On This Day (WWDOTD):

- Cole recommended using Prezi for the presentation, as it looks professional and unique
- Began work on the Prezi presentation - https://prezi.com/view/AhcJh8YFU7uCJ0BGeDhi/
- Added the banner image for this project:

- Added the 4:3 image for this project:

March

Date: February 27-1, 2026
People Involved: Matthew
Title: Background Research + Citations
Topic/Section: Portal
What We Did On This Day (WWDOTD):

- Began work revising, editing, and finishing my portion of the Background Research
- All that is left is for Ivan to add details on SVM
- Added all the links needed for citations; however, not in AMA format
- Used a citation converter to convert them from APA to AMA format

Date: March 1, 2026
People Involved: Ivan, Matthew
Title: Prezi and Practice
Topic/Section: Presentation
What We Did On This Day (WWDOTD):

- Worked on the Prezi together, finishing up the colour scheme, BR, and Objectives
- Practiced the script together as well

Date: March 2, 2026
People Involved: Ivan, Matthew
Title: School Science Fair
Topic/Section: School Science Fair
What We Did On This Day (WWDOTD):

- Today is the school science fair
- Apart from presenting to the judges, Matthew continued work on Prezi, and Ivan worked on finishing up BR.
- Work was also done in writing the Conclusions and Research sections.