

Master's degree in computer science

Overview - Report

Contents

[1- Introduction 4](#_Toc189570227)

[2- Background 5](#_Toc189570228)

[2.1 – Protein-Protein Interactions 5](#_Toc189570229)

[2.2 – The Gut Microbiome 5](#_Toc189570230)

[2.3 – Connecting PPIs and the Gut Microbiome 6](#_Toc189570231)

[3- Problem 6](#_Toc189570232)

[3.1 – Understanding the Challenge 6](#_Toc189570233)

[3.2 – Research Hypothesis 7](#_Toc189570234)

[3.3 – Research Goals 7](#_Toc189570235)

[4 - Methodology 7](#_Toc189570236)

[4.1 - OTU Data Experiment 8](#_Toc189570237)

[Objective 8](#_Toc189570238)

[Methodology 8](#_Toc189570239)

[Key Findings 8](#_Toc189570240)

[Challenges 9](#_Toc189570241)

[4.2 - PPI Dataset Experiment 9](#_Toc189570242)

[Objective 9](#_Toc189570243)

[Methodology 9](#_Toc189570244)

[Key Findings 9](#_Toc189570245)

[Challenges 10](#_Toc189570246)

[4.3 - Gut Microbiome Proteome Exploration 10](#_Toc189570247)

[Objective 10](#_Toc189570248)

[Methodology 10](#_Toc189570249)

[Key Findings 10](#_Toc189570250)

[Challenges 11](#_Toc189570251)

[4.4 - Augmented OTU Dataset with Metadata 11](#_Toc189570252)

[Objective 11](#_Toc189570253)

[Key Findings 12](#_Toc189570254)

[Challenges 12](#_Toc189570255)

[5- Results 13](#_Toc189570256)

[5.1 – Results by Project 13](#_Toc189570257)

[5.1.1 – OTU Data Experiment 13](#_Toc189570258)

[5.1.2 – PPI Dataset Experiment 13](#_Toc189570259)

[5.1.3 – Gut Microbiome Proteome Exploration 14](#_Toc189570260)

[5.1.4 – Augmented OTU Dataset with Metadata 14](#_Toc189570261)

[5.2 – Summary of Key Results 14](#_Toc189570262)

[5.3 – Summary of Key Results 15](#_Toc189570263)

[6- Future Directions 16](#_Toc189570264)

[6.1 – Acquiring Real-World PPI and Microbiome Datasets 16](#_Toc189570265)

[6.2 – Enhancing Machine Learning Methodologies 16](#_Toc189570266)

[6.3 – Refining Data Preprocessing and PPI Validation Techniques 17](#_Toc189570267)

[6.4 – Bridging Computational and Biological Expertise 17](#_Toc189570268)

[6.5 – Expanding Disease Applications Beyond Type 2 Diabetes 17](#_Toc189570269)

[6.6 – Final Roadmap: Key Next Steps 18](#_Toc189570270)

[7- Conclusion 18](#_Toc189570271)

[7.1 – Key Findings 19](#_Toc189570272)

[7.2 – Limitations 19](#_Toc189570273)

[7.3 – Future Directions 20](#_Toc189570274)

[7.4 – Final Thoughts 21](#_Toc189570275)

# Introduction

Understanding the relationship between the gut microbiome and human disease is a growing field in both biological and computational research. One promising avenue of study is **Protein-Protein Interactions (PPIs)**—the physical or functional interactions between proteins that govern most biological processes. Disruptions in these interactions are associated with diseases such as **cancer, neurodegenerative disorders, and metabolic conditions** like Type 2 Diabetes (T2D).

This research explores the potential of **PPIs as a predictive tool for disease classification**, using the gut microbiome as a model system. The gut microbiome, which consists of trillions of microorganisms living in the digestive tract, plays a crucial role in immunity, metabolism, and inflammation. Since recent studies have linked microbiome composition to disease states as well as the importance of the gut microbiome as representative of general health, analyzing PPIs within this environment could provide new insights into disease mechanisms and predictive modeling.

To develop and refine a methodology for microbiome-based disease classification, this report details **four preparatory projects** conducted as foundational steps in this research:

1. **OTU Data Experiment** – Investigated the feasibility of using AI for microbiome-based disease classification by generating and analyzing synthetic Operational Taxonomic Unit (OTU) data.
2. **PPI Dataset Experiment** – Applied machine learning models to an existing PPI dataset to evaluate their effectiveness in identifying meaningful protein interactions.
3. **Gut Microbiome Proteome Exploration** – Explored the proteomic landscape of the gut microbiome to identify disease-relevant proteins and potential interactions with the human proteome.
4. **Augmented OTU Dataset with Metadata** – Enhanced synthetic OTU datasets with biological metadata to assess the impact of additional context on predictive accuracy.

Each project contributed to refining machine learning methodologies, dataset preprocessing, and biological validation strategies, ultimately shaping a more robust framework for microbiome-based disease prediction. However, challenges such as **data reliability, overfitting in machine learning models, and the need for real-world validation** emerged as critical factors influencing the success of the study.

This report provides a detailed account of these projects, their methodologies, results, and insights, paving the way for future research aimed at integrating biological and computational approaches for disease prediction.

# Background

To understand the rationale behind this research, it is essential to explore two key concepts: **Protein-Protein Interactions (PPIs) and the Gut Microbiome**. These biological systems form the foundation for using computational approaches to predict disease states.

## 2.1 – Protein-Protein Interactions

Proteins do not function in isolation; they interact with each other to regulate nearly all cellular processes, including **metabolism, immune response, and disease progression**. PPIs refer to **the physical or functional associations between proteins**, which can be either **direct** (e.g., forming protein complexes) or **indirect** (e.g., signaling pathways).

Understanding PPIs is critical in disease research because disruptions in these interactions often contribute to pathological conditions such as:

* **Cancer:** Mutations in key proteins can lead to uncontrolled cell growth.
* **Neurodegenerative Diseases:** Abnormal protein aggregation is linked to conditions like Alzheimer's and Parkinson’s disease.
* **Metabolic Disorders:** Altered protein interactions in the gut microbiome can impact insulin regulation, potentially influencing Type 2 Diabetes (T2D).

By identifying patterns in protein interactions, researchers can:

* **Detect disease-specific biomarkers** that indicate the presence or progression of a disease.
* **Uncover key pathways disrupted in disease states**, improving our understanding of disease mechanisms.
* **Develop computational models to predict risk of diseases**, using machine learning to analyze large datasets of PPIs.

## 2.2 – The Gut Microbiome

The **gut microbiome** is a vast and diverse collection of microorganisms residing in the human digestive tract. These microbes interact with each other and the human host, playing crucial roles in:

1. **Metabolism:** Assisting in nutrient absorption and digestion.
2. **Immune Regulation:** Helping the body recognize and respond to harmful pathogens.
3. **Inflammation and Disease Development:** Imbalances in microbiome composition (dysbiosis) are associated with conditions like inflammatory bowel disease, obesity, and T2D.

Recent studies have established strong links between the gut microbiome and **chronic diseases, including diabetes and cardiovascular disease**. Since microbial proteins interact with human proteins through complex biochemical pathways, studying PPIs within the microbiome could provide new insights into disease prediction and treatment strategies.

## 2.3 – Connecting PPIs and the Gut Microbiome

The intersection of **PPIs and the gut microbiome** represents an emerging field of research. Since the gut microbiome reflects overall health, analyzing **protein interactions within microbiome-associated organisms** could serve as a proxy for broader disease mechanisms.

This project aims to explore whether **PPIs among gut microbiome proteins can be leveraged as predictive features in disease classification models**. The hypothesis is that by mapping these interactions, machine learning models can be trained to identify patterns linked to disease states.

# Problem

The primary objective of this research titled “IA Generativa para Classificação do Microbioma Humano” is to explore whether **Protein-Protein Interactions (PPIs) in the gut microbiome can be used to predict disease states**. This goal presents several challenges that arise from both biological complexity and computational constraints.

## 3.1 – Understanding the Challenge

Protein-Protein Interactions (PPIs) are fundamental to cellular processes, but not all proteins interact. Their interactions are governed by factors such as **molecular structure, charge distribution, and biochemical properties**. While PPI disruptions are well-documented in diseases like cancer and neurodegenerative disorders, their role in microbiome-related diseases is less understood.

A major question in this research is:  
**Can microbiome-associated PPIs provide meaningful insights into disease prediction?**

To answer this, several **key challenges** must be addressed:

1. Availability of Reliable PPI Data

* While large-scale PPI databases exist, they often lack comprehensive microbiome-specific data, limiting their utility for gut-related disease research.
* Many microbiome protein interactions remain uncharacterized or unverified in existing repositories.

1. Synthetic vs. Real-World Data Issues
   * Early experiments relied on synthetic microbiome datasets, generated using AI tools like ChatGPT, which raised concerns about biological realism and data validity.
   * Real-world microbiome datasets are limited, fragmented, restricted and require extensive preprocessing, adding complexity to machine learning applications.
2. Bridging Biology and Machine Learning

* The interdisciplinary nature of this research requires expertise in both biological preprocessing and machine learning methodologies.
* Overfitting risks arise when working with small or synthetic datasets, leading to misleadingly high accuracy in predictive models.

## 3.2 – Research Hypothesis

This study hypothesizes that:

1. **PPIs within the gut microbiome contribute to disease prediction** and can serve as meaningful biological markers.
2. **Machine learning models can effectively classify disease states** based on microbiome-derived PPI patterns, provided that data quality and preprocessing challenges are addressed.

## 3.3 – Research Goals

To test this hypothesis, the following steps will be taken:

* **Obtain and refine gut microbiome datasets**, ensuring they contain relevant protein interaction data.
* **Develop machine learning models** capable of identifying predictive patterns in PPI data.

By addressing these challenges, this research aims to contribute to the growing field of computational microbiome analysis and pave the way for more accurate, data-driven disease prediction models.

# 4 - Methodology

The methodology followed in this research was designed to gradually refine the approach to **microbiome-based disease classification**. A step-by-step pipeline was adopted, transitioning from **broad exploratory analysis** to **more focused machine learning applications**. The research progressed through **four preparatory projects**, each addressing specific challenges related to **data acquisition, preprocessing, model training, and validation**.

The **core methodology** for the final research goal can be summarized as follows:

1. **Obtain a gut microbiome dataset** containing microbial profiles of healthy and diseased individuals (Type 2 Diabetes – T2D).
2. **Identify key microorganisms** most relevant to T2D.
3. **Extract the proteome** (all proteins) associated with these key microorganisms.
4. **Investigate Protein-Protein Interactions (PPIs)** between microbial proteins and those in the human gut microbiome.
5. **Curate the dataset** by cross-referencing known PPI databases.
6. **Apply machine learning techniques** to predict disease status based on these interactions.

Since this pipeline is complex, **preliminary projects** were conducted to refine and validate key aspects before applying them to real-world datasets.

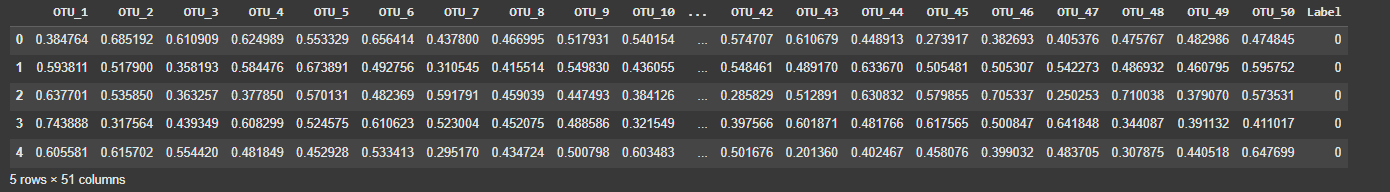
## 4.1 - OTU Data Experiment

### Objective

To assess the feasibility of **synthetically generated Operational Taxonomic Unit (OTU) data** for microbiome-based disease classification.

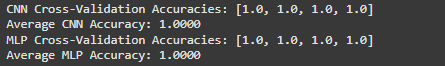
### Methodology

* **Synthetic OTU data** was generated using ChatGPT, with microbial profiles labeled as either **Healthy** or **T2D**.
* A dataset containing **50 OTUs per sample** was created, and **machine learning models** were trained to classify disease presence.



* The dataset was split **(80% training, 20% testing)**, normalized, and processed using **a neural network with two layers (64 and 32 neurons)**.
* **Cross-validation (5 folds)** and **dropout layers (50%)** were applied to counteract potential overfitting.

### Key Findings

* **Unexpected 100% accuracy** was achieved in the initial model, suggesting possible overfitting, lack of data complexity or synthetic bias (implemented cross-fold validation and dropout layers to mitigate this instance).
* Increasing dataset size **dropped accuracy to ~50%**, raising concerns about the reliability of synthetic data.
* Applying **Convolutional Neural Networks (CNNs)** failed to improve classification performance.
* Data augmentation (adding **random noise**) increased dataset size and restored **100% accuracy**, further indicating synthetic bias.

### Challenges

* **Over-synthesized data** led to unrealistically high model performance.
* Lack of **biological variability** in the dataset reduced its real-world applicability.

## 4.2 - PPI Dataset Experiment

### Objective

To evaluate various **machine learning models** on an **existing public dataset of Protein-Protein Interactions (PPIs)** and optimize classification accuracy.

### Methodology

* The dataset consisted of **two files**: one containing **interacting protein sequences**, the other **non-interacting sequences**.
* **Machine learning models tested:**
  + Random Forest
  + XGBoost
  + Support Vector Classifier (SVC)
  + Gradient Boosting
  + Neural Network (MLP)
  + Convolutional Neural Networks (CNNs)

|  |
| --- |
| **Results & Model Performance** |
| | **Model** | **Accuracy (%)** | **After Hyperparameter Tuning (%)** | | --- | --- | --- | | SVC | 59% | - | | Gradient Boosting | 78% | - | | XGBoost | 96% | **98.6%** | | Random Forest | 96% | **98%** | | Neural Network (MLP) | 55% | **77%** | | CNN | 81% | - | |

### Key Findings

* **Tree-based models (XGBoost, Random Forest)** performed the best, aligning with previous research findings.
* **Deep learning models (Neural Networks, CNNs)** underperformed on structured data, possibly due to the dataset format.
* **Hyperparameter tuning** improved accuracy, but results remained **slightly lower than prior studies** using the same dataset.

### Challenges

* **Limited ability to innovate in data preprocessing**, as the pipeline followed a predefined structure from previous research and lacked the biology knowledge necessary to perform further improvements.
* Deep learning models did **not outperform traditional ML models**, highlighting the importance of dataset format and feature engineering.

## 4.3 - Gut Microbiome Proteome Exploration

### Objective

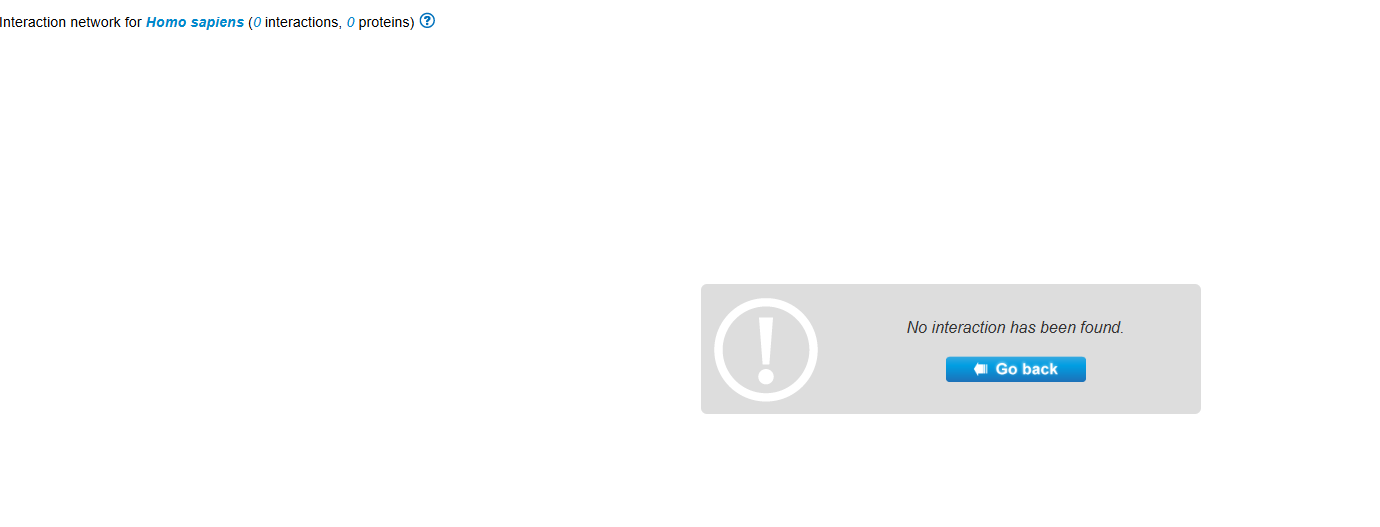
To identify **proteins relevant to Type 2 Diabetes** within the gut microbiome and investigate **potential interactions** with the human proteome.

### Methodology

* Identified microorganisms linked to T2D using ChatGPT:
  + Beneficial:
    - Akkermansia muciniphila;
    - Faecalibacterium prausnitzii;
    - Bifidobacterium spp.
  + Potentially Harmful:
    - *Ruminococcus spp;*
    - *Prevotella spp;*
    - *Desulfovibrio spp;*
    - *Proteobacteria (E. coli pathogenic strains);*
  + Other Microorganisms of Interest:
    - **Clostridium spp;**
* Downloaded proteomes for these microorganisms from UniProt.
* **Filtered proteins** based on their Protein Existence (PE) level:
* **Excluded PE=4 & PE=5 proteins** (predicted or uncertain) thus not being certain of their existence/ their existence being extremely unlikely.
* **Ruminococcus spp. data removed** due to database inconsistencies.
* **Filtered proteins** based on their Protein Existence (PE) level:

### Key Findings

* **No interactions were found** between gut microbiome proteins and human proteins.



* Potential reasons for this:
  + The **Interactome Database may lack relevant PPI data**.
  + **Excluding PE=4 and PE=5 proteins** may have removed key interactions.
  + The dataset used may not have fully represented **biologically validated PPIs**.

### Challenges

* **Database limitations** may have prevented detection of key interactions.
* **Exclusion criteria** could have removed relevant proteins.
* **Incomplete microbiome PPI knowledge** remains a barrier to accurate prediction.

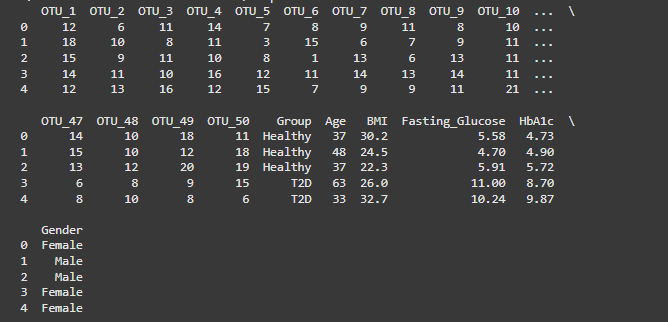
## 4.4 - Augmented OTU Dataset with Metadata

### Objective

To assess whether adding **biological metadata** to synthetic OTU datasets improves **disease classification performance**.

**Methodology**

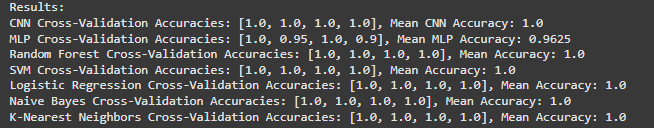
* ChatGPT was used to generate **OTU data with additional features**:
  + **Age, gender, BMI, and other metabolic indicators** relevant to T2D.



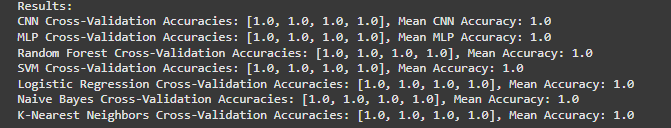
* Various **machine learning models** were trained on:
  + **Original OTU dataset**
  + **Augmented dataset (OTUs + metadata)**
* To test robustness, the dataset was **scaled from 100 to 500, then to 1000 entries** using:
  + **SMOTE (Synthetic Minority Over-sampling Technique)**
  + Additional synthetic data from ChatGPT.

### Key Findings

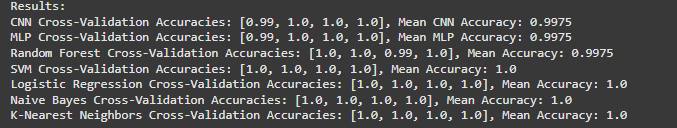
* All models achieved **100% accuracy**, except for **MLP**.

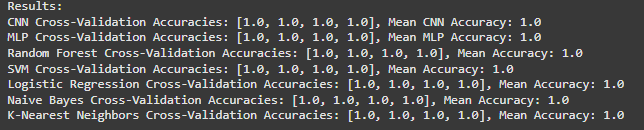


* After increasing dataset size, performance remained unchanged.



* Comparing ChatGPT-generated **500-entry and 1000-entry datasets** showed, respectively, similar accuracy.





### Challenges

* **Overfitting persisted** despite data augmentation techniques.
* **Lack of biological variability** in synthetic metadata limited real-world applicability.

# Results

The results of this research vary significantly depending on the dataset and experimental approach used. While some models achieved **high accuracy**, key limitations such as **overfitting, reliance on synthetic data, and incomplete PPI datasets** impacted the biological relevance of findings. This section summarizes the key results obtained across the four preparatory projects and discusses their implications.

## 5.1 – Results by Project

### 5.1.1 – OTU Data Experiment

**Key Findings**

* Initial machine learning models trained on **synthetic OTU data** achieved **100% accuracy**, which raised concerns about **overfitting**.
* Expanding the dataset **significantly reduced accuracy (to ~50%)**, confirming that synthetic data lacked variability.
* Using **data augmentation (adding noise to synthetic samples)** restored **100% accuracy**, reinforcing the concern that synthetic data may be inherently biased.

**Limitations**

* Results are likely **not generalizable to real-world microbiome classification** due to synthetic nature of data.
* Model performance on **real, diverse datasets remains unknown**.

### 5.1.2 – PPI Dataset Experiment

**Key Findings**

* XGBoost and Random Forest performed the best, with accuracy improving from 96% to 98.6% and 98%, respectively, after hyperparameter tuning.
* Deep learning models (MLP, CNNs) performed worse than expected, showing that structured PPI datasets favor traditional ML models over deep learning architectures.
* Tuning hyperparameters helped close the gap between the study's performance and prior research but did not exceed existing benchmarks.

**Limitations**

* Data preprocessing was heavily dependent on prior studies, limiting opportunities for innovation.
* Deep learning models underperformed, possibly due to the structured nature of the dataset, requiring further feature engineering.
* The dataset focused on generic PPIs and not microbiome-specific interactions, which is the actual research goal.

### 5.1.3 – Gut Microbiome Proteome Exploration

**Key Findings**

* Proteomes of microorganisms associated with Type 2 Diabetes (T2D) were successfully extracted and filtered for protein existence (PE) levels.
* No interactions were found between gut microbiome proteins and human proteins in the Interactome database.

**Limitations**

* Possible reasons for zero detected interactions:
  + Database limitations – The Interactome might be incomplete or missing microbiome-specific PPIs.
  + Filtering criteria – Removing PE=4 & PE=5 proteins may have excluded relevant but unverified interactions.
  + Microbiome-human PPI knowledge gap – These interactions may exist but are not yet scientifically documented.
* Next steps should involve:
  + Exploring alternative databases that might contain microbiome-human protein interactions.
  + Revisiting filtering criteria to ensure potentially important proteins are not excluded.

### 5.1.4 – Augmented OTU Dataset with Metadata

**Key Findings**

* Adding biological metadata (age, gender, BMI, metabolic indicators) did not significantly impact classification accuracy—all models still achieved 100% accuracy except MLP.
* Expanding the dataset using SMOTE and ChatGPT-generated data resulted in consistent 100% accuracy, further raising concerns about synthetic data overfitting.

**Limitations**

* Synthetic metadata lacked real-world biological variability, making it difficult to assess true predictive power.
* Scalability concerns – Even with augmented datasets, model performance on real microbiome samples is unknown.

## 5.2 – Summary of Key Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Project** | **Objective** | **Key Result** | **Limitation** |
| **OTU Data Experiment** | Assess synthetic microbiome data for disease classification | **100% accuracy, but extreme variability when dataset size changed** | **Overfitting, synthetic data lacks biological relevance** |
| **PPI Dataset Experiment** | Evaluate ML models on public PPI data | **Tree-based models (XGBoost, RF) outperformed deep learning** | **Dataset not microbiome-specific, ML improvements did not exceed prior benchmarks** |
| **Gut Microbiome Proteome Exploration** | Identify microbial proteins linked to T2D and check for PPIs with humans | **No interactions found in the Interactome database** | **Database incompleteness, possible exclusion of relevant proteins** |
| **Augmented OTU Dataset** | Test impact of biological metadata on disease classification | **100% accuracy maintained with added metadata** | **Synthetic nature of dataset limits real-world insights** |

## 5.3 – Summary of Key Results

* **Synthetic Data is a Double-Edged Sword**
  + While synthetic datasets are **useful for testing methodologies**, they can **lead to misleadingly high accuracy** and **overfitting**, making them **less useful for real-world applications**.
* **Machine Learning Performance is Highly Dataset-Dependent**
  + **Tree-based models (XGBoost, Random Forest) outperformed deep learning** on structured datasets.
  + Deep learning approaches require **more advanced feature extraction techniques** to work effectively.
* **PPI Data Availability is a Major Bottleneck**
  + The **absence of gut microbiome-human PPI interactions** in the Interactome database was a **major setback**.
  + **Alternative sources** and **wet-lab validation** may be needed to bridge this gap.
* **Metadata Alone Does Not Enhance Predictive Power**
  + Adding **age, BMI, and other health markers** **did not significantly impact disease classification** when using synthetic OTU data.
  + Real-world studies might yield different results, but synthetic **metadata augmentation was insufficient** in this case.

# Future Directions

While this research demonstrated the potential of Protein-Protein Interactions (PPIs) in microbiome-based disease prediction, it also highlighted several limitations and challenges that must be addressed for meaningful real-world applications. Future research should focus on enhancing data reliability, improving machine learning models, and bridging the gap between computational and biological expertise.

The following directions outline key next steps to overcome current limitations and advance this field.

## 6.1 – Acquiring Real-World PPI and Microbiome Datasets

One of the most critical barriers in this research was the **lack of high-quality, real-world PPI data** for the gut microbiome. Moving forward, efforts should focus on:

* **Accessing experimentally validated PPI datasets** for microbiome-related proteins, as existing public datasets primarily focus on human protein interactions.
* **Collaborating with microbiologists** to obtain curated microbiome proteomics data, ensuring that computational models are trained on biologically relevant interactions.
* **Exploring alternative data sources**, such as:
  + **Mass spectrometry-based proteomics studies** that investigate gut microbiome proteins.
  + **Metagenomic and metaproteomic datasets** that provide insights into microbial community interactions.

## 6.2 – Enhancing Machine Learning Methodologies

Although machine learning models showed promising results, several improvements can make them more robust and generalizable:

* **Developing hybrid models** that integrate:
  + **Tree-based models (e.g., XGBoost, Random Forest)** for structured data.
  + **Deep learning models (e.g., Transformers, CNNs)** for sequence-based and high-dimensional datasets.
* **Improving feature selection** by incorporating **functional annotations of proteins**, rather than relying solely on sequence data.
* **Addressing overfitting in synthetic datasets** by:
  + Using **transfer learning** with real biological datasets.
  + Testing models on **external validation datasets** to ensure they generalize.

## 6.3 – Refining Data Preprocessing and PPI Validation Techniques

A major challenge in this research was ensuring that **filtered proteins were still biologically relevant**. Future work should:

* **Reevaluate exclusion criteria** for proteins in PPI datasets:
  + Instead of **completely removing PE=4 & PE=5 proteins**, consider weighting them based on confidence scores.
  + Investigate **alternative protein evidence validation techniques** to include potentially relevant, but unverified, interactions.
* **Expand the use of PPI databases** beyond the Interactome by integrating:
  + **STRING database** (a widely used protein interaction network database).
  + **BioGRID and DIP (Database of Interacting Proteins)** for additional PPI sources.
* **Explore experimental validation approaches**:
  + Use **wet-lab experiments** to validate **predicted interactions** in microbiome samples.
  + Collaborate with **biological labs** that conduct PPI studies to cross-check computational findings.

## 6.4 – Bridging Computational and Biological Expertise

This research underscored the **importance of interdisciplinary collaboration** in microbiome-based disease prediction. Future work should:

* **Engage with microbiologists and bioinformaticians** to refine biological interpretations of PPI networks.
* **Develop user-friendly tools** that allow **biological researchers to interact with machine learning models**, providing insights for model refinement.
* **Establish partnerships with experimental labs** to:
  + Validate machine learning predictions with **real-world biological samples**.
  + Improve data curation and feature selection based on **domain expertise**.

## 6.5 – Expanding Disease Applications Beyond Type 2 Diabetes

While this study focused on **Type 2 Diabetes (T2D)**, future research can expand to other diseases with known microbiome associations:

* **Neurodegenerative Diseases (e.g., Alzheimer's, Parkinson’s)**
  + The **gut-brain axis** plays a crucial role in neurological health.
  + Identifying **microbiome PPIs linked to neuroinflammation** could provide new diagnostic biomarkers.
* **Inflammatory Bowel Diseases (IBD)**
  + Conditions like **Crohn’s disease and ulcerative colitis** are directly linked to microbiome imbalances.
  + Analyzing **gut microbiome PPIs** may uncover key inflammatory pathways.
* **Cancer Prediction**
  + Certain microbiome compositions are associated with an **increased risk of colorectal cancer**.
  + Investigating **microbial protein interactions with oncogenic pathways** could improve **early detection strategies**.

## 6.6 – Final Roadmap: Key Next Steps

|  |  |  |
| --- | --- | --- |
| **Future Goal** | **Actionable Steps** | **Expected Outcome** |
| **Obtain Real PPI Data** | Collaborate with biological researchers & explore new databases | More reliable datasets for ML training |
| Improve ML Models | Develop hybrid models & refine feature selection | More accurate disease prediction |
| Expand PPI Validation | Use STRING/BioGRID & experimental wet-lab testing | Higher confidence in detected interactions |
| Enhance Interdisciplinary Collaboration | Work with microbiologists, labs & bioinformaticians | More biologically relevant computational models |
| **Broaden Disease Focus** | Apply methodology to IBD, neurodegenerative diseases, & cancer | Increased medical relevance & impact |

# Conclusion

This research explored the potential of Protein-Protein Interactions (PPIs) in microbiome-based disease prediction, with a focus on Type 2 Diabetes (T2D). Through a series of preparatory projects, various machine learning models, synthetic and public datasets, and microbiome proteomics approaches were tested to evaluate their effectiveness in disease classification.

While significant progress was made in developing a computational pipeline for microbiome-based disease prediction, critical challenges remain—primarily the lack of real-world PPI datasets, over-reliance on synthetic data, and the need for interdisciplinary validation. This conclusion summarizes key findings, highlights limitations, and outlines the path forward for improving microbiome-based disease prediction models.

## 7.1 – Key Findings

Across the four preparatory projects, several important insights emerged:

* Machine Learning Models Can Be Used for Microbiome-Based Disease Prediction
  + Tree-based models (XGBoost, Random Forest) performed better than deep learning approaches on structured datasets.
  + Deep learning models (CNNs, MLPs) struggled with structured data but may still be viable with richer feature extraction techniques.
* Synthetic Data is Useful for Testing but Insufficient for Real-World Applications
  + Machine learning models trained on synthetic OTU and PPI datasets achieved high accuracy, but overfitting was a major concern.
  + Expanding datasets through data augmentation (SMOTE, noise injection) helped, but results still lacked biological variability.
* Current PPI Databases Lack Microbiome-Specific Interactions
  + A key finding was that no interactions were found between gut microbiome proteins and human proteins in the Interactome database.
  + This suggests either a gap in existing biological knowledge or database incompleteness, making it difficult to validate microbiome PPIs computationally.
* Biological Metadata Does Not Necessarily Improve Predictive Accuracy
  + Adding biological factors such as age, BMI, and metabolic markers to OTU datasets did not significantly improve classification accuracy.
  + However, in real-world applications, such metadata may still provide valuable contextual information.
* Interdisciplinary Collaboration is Necessary for Advancing This Research
  + The gap between computational and biological expertise was evident throughout the project.
  + Future progress will require collaborations with microbiologists, bioinformaticians, and biomedical researchers to ensure data quality and biological relevance.

## 7.2 – Limitations

Despite promising advancements, several limitations must be acknowledged:

* Lack of Real-World PPI Data for Microbiome Studies
  + Most PPI research focuses on human protein interactions, with limited microbiome-specific datasets.
  + The inability to detect microbiome-human PPIs in the Interactome suggests that alternative data sources or experimental validation are needed.
* Overfitting in Synthetic Datasets
  + The 100% accuracy observed in some models indicates that the dataset lacked real-world complexity.
  + Over-reliance on AI-generated synthetic data makes it difficult to assess how these models would perform in real biological settings.
* Deep Learning Models Did Not Outperform Traditional ML
  + Despite their success in many biological fields, deep learning models underperformed in this study.
  + This suggests that feature extraction techniques need improvement or that deep learning is not the best approach for structured microbiome data.
* PPI Filtering May Have Removed Relevant Proteins
  + The exclusion of proteins with PE=4 & PE=5 may have inadvertently removed biologically important interactions.
  + Future work should explore alternative filtering strategies to balance data quality with biological completeness.
* Limited Scope of Disease Prediction
  + This study focused primarily on Type 2 Diabetes, but microbiome PPIs could also be relevant to other diseases.
  + Expanding the approach to neurodegenerative diseases, inflammatory bowel disease (IBD), and cancer could provide broader medical applications.

## 7.3 – Future Directions

To overcome these limitations and advance the field, the following **next steps** should be prioritized:

* **Obtain Real-World Microbiome PPI Data**
* Collaborate with **biologists and microbiome researchers** to access validated microbiome PPI datasets.
* Integrate **mass spectrometry-based proteomics and metagenomic studies** into computational pipelines.
* **Improve Machine Learning Feature Selection**
* Explore **graph-based PPI models** that capture **network relationships between proteins**.
* Use **transformer-based architectures** for analyzing **sequential and spatial protein interactions**.
* **Validate Computational Predictions with Experimental Studies**
* Work with **biomedical labs** to validate **predicted microbiome PPIs** using **wet-lab experiments**.
* Test models against **external validation datasets** to assess real-world performance.
* **Broaden Disease Applications Beyond T2D**
* Expand the methodology to investigate microbiome PPIs in **Alzheimer’s, Parkinson’s, IBD, and colorectal cancer**.
* Develop **personalized disease prediction models** based on individual microbiome compositions.
* **Strengthen Interdisciplinary Collaboration**
* Create **automated tools** that allow biologists to easily interpret machine learning predictions.
* Foster **cross-disciplinary partnerships** between **computational researchers, microbiologists, and medical professionals**.

By focusing on these future directions, the research can transition from a **computational proof-of-concept** to a **clinically applicable framework for microbiome-based disease prediction**.

## 7.4 – Final Thoughts

This study laid the groundwork for **machine learning-based microbiome disease prediction** by exploring the potential role of **Protein-Protein Interactions (PPIs)** in classification models. While computational experiments **demonstrated feasibility**, the research also revealed **significant challenges in dataset reliability, model generalizability, and biological validation**.

The **next frontier** in this field will be **integrating real-world PPI data, refining machine learning techniques, and collaborating with domain experts** to ensure that computational findings translate into **meaningful biological and medical applications**.

By addressing these challenges, future research has the potential to **revolutionize disease prediction, precision medicine, and our broader understanding of microbiome-protein interactions in human health**.