Causal Discovery in Gut Microbes for PCOS

Mariana Paco Mendivil
mpacomendivil@ucsd.edu

Candus Shi
c6shi@ucsd.edu

Nicole Zhang
nwzhang@ucsd.edu

Biwei Huang
bih007@ucsd.edu

Jelena Bradic
jbradic@ucsd.edu

Abstract

The human gut microbiome has become a significant factor in understanding metabolic health, influencing conditions such as type 2 diabetes (T2D) and polycystic ovary syndrome (PCOS). Despite its recognized impact, much of the current research on the human gut microbiome and diseases remain limited to associative and correlational studies, leaving gaps in understanding the underlying causal relationships. This study addresses these gaps by utilizing causal discovery algorithms and causal inference methods and comparing them with prediction models to investigate microbial contributions to T2D and PCOS. First, we graph the microbe-microbe interaction networks on the genus level for healthy and diseased cohorts using a version of the Peter-Clark (PC) algorithm altered to reduce the multiple testing burden. Then, we graph the microbe-disease interaction network on the genus level for a disease using the constraint-based causal discovery from heterogeneous/nonstationary data algorithm (CD-NOD) and compare the microbes directly linked to disease with microbes from a variational autoencoder (VAE) prediction model. Our results show that there are microbes causal to T2D and microbes causal to PCOS (expand). This work aims to provide a framework for investigating causal relationships between the gut microbiome and other diseases as well as guide further research and wet-lab experiments and develop a stronger understanding of the role of the gut microbiome in precision medicine.

Code:
https://github.com/nzhang20/Causal-Discovery-for-Biomedical-Applications

1 Introduction 3
2 Methods 7
3 Results 8
4 Discussion 8
5 Conclusion 8
References 9

Appendices A1

1 Introduction

The human gut microbiome has gained significant attention in recent years for its important role in metabolic health. While there has been extensive research that links the microbiome to health disorders such as type 2 diabetes (T2D) (Zhou et al. 2019) and polycystic ovary syndrome (PCOS) (Yang et al. 2024), the majority of these studies remain correlational, leaving causal relationships undiscovered. Understanding these relationships is essential to improving and personalizing medical treatments for such diseases. This study builds on recent advancements in causal discovery algorithms to investigate how microbial taxa influence metabolic disorders. Our goal is to find patterns that conventional association-based methods might miss by leveraging the marginal and conditional independencies found in the data as well as network theory to assess where a causal relationship might occur and if possible, its causal direction. Given the high-dimensional nature of this type of data, we also explore different feature pruning techniques to reduce the multiple testing burden and for ease of interpretation.

We focus on two aspects of causal discovery and causal inference in the gut microbiome. First, we are curious to see how the microbe-microbe interaction networks may differ between the two outcome groups. Due to the high number of features compared to the number of samples in gut microbiome abundance data, we first reduce the number of edges between microbes using a sparse correlation method, SparCC (Weiss et al. 2016; Friedman and Alm 2012), and a sparse precision matrix estimator, graphical lasso (Friedman, Hastie and Tibshirani 2008). Then, we graph the two networks for the corresponding cohorts using a constraint-based causal discovery algorithm similar to the Peter-Clark (PC) algorithm (Glymour, Zhang and Spirtes 2019), but with a smaller depth and without a direction orientation step to reduce the multiple testing burden.

Second, we are also interested in graphing the microbe-disease interaction network, where we are particularly interested in the microbes directly linked to disease status. Instead of reducing the number of edges, we reduce the number of features using logistic lasso regression to account for the outcome variable. Finally, we graph the network using the features that survive lasso and a disease status node using the constraint-based causal discovery from heterogeneous/nonstationary data algorithm (CD-NOD) (Huang et al. 2019) to identify microbes directly linked to disease status. Given the predictive nature of graphing a microbe-disease interaction network, we are also interested in developing a prediction model for disease using a causal representation learning technique xxx and comparing the results from CD-NOD to the microbes used in the model.

1.1 Literature Review

T2D

T2D is a metabolic disease where individuals have chronic high blood sugar, otherwise known as hyperglycaemia. This is a result of insulin resistance where the pancreas produces insulin, but the cells do not respond to it, leading the pancreas to try to produce more. The pancreas eventually fails to keep producing insulin leading to low insulin levels and high blood sugar, and this can lead to increased risks of developing other diseases such as heart disease and kidney disease (ADA 2025). T2D affects millions of people, and many studies have been conducted to investigate its underlying cause, its common precursor coined as "prediabetes", and other factors that can contribute or affect the development of T2D (Tabák et al. 2012; Qin et al. 2012; Mehta et al. 2000).

Given the impactful role of the gut microbiome on human health, numerous studies have also investigated the relationship between gut microbiota and T2D. For example, (Zhou et al. 2019) conducted a longitudinal study of multi-omic data on healthy individuals vs individuals with prediabetes (an early stage of T2D) to determine how microbes behave differently between the two cohorts. They found that variation in microbes between and within individuals of each cohort differed, that each cohort responded to infections and immunizations differently, and through associations, that host-microbe interactions differed between the two cohorts. In particular, they found that "the genus *Holdemania* was significantly associated with *Clostridium XIVb* and *Phascolarctobacterium* in insulin-sensitive participants, but significantly correlated with *Clostridium XIVa*, *Clostridium XVII*, *Collinsella*, *Lachnospiracea incertae sedis*, and *unclassified Lachnospiraceae* in insulin-resistant participants". (Baars et al. 2024) also found common results from various studies investigating this relationship: there appears to be "a reduction of butyrate-producing bacteria such as *Faecalibacterium*, *Clostridium*, and *Akkermansia* in individuals with T2D".

These analyses demonstrate that there are significant differences in microbe interactions between healthy and prediabetic individuals, and furthermore, that we can discover the causal graph from their data and use causality to determine which host-microbe interactions this study found through associations are not spurious, but causal.

PCOS

PCOS is a complex endocrine disorder linked to metabolic diseases such as obesity and T2D. It affects 6-13% of women of reproductive age, and 70% of affected women remain undiagnosed as the causes of PCOS largely remains a mystery (WHO 2025). In fact, it was not until quite recently that the scientific community has peaked interest in studying PCOS and its causes. Current diagnostic methods use hormone and metabolic biomarkers, but these techniques are insufficient to differentiate between different PCOS subtypes, such as those characterized by hyperandrogenism. Due to inconsistent study findings, regional differences, and heterogeneity in studies, the association between PCOS and gut microbiota is not well-defined. (Yang et al. 2024) conduct an individual participant data meta-analysis and systematic review to see if gut microbiota characteristics between healthy individuals and PCOS patients, between different subtypes of PCOS, and regional differences can be identified using data from a variety of clinical trials.

Using Wilcoxon tests with Benjamini-Hochberg corrected p-values, they found differential bacteria between the healthy and PCOS groups: PCOS patients had slightly lower levels of *Bacillota* and higher levels of *Actinobacteriota*; PCOS patients in China had lower alpha diversity than healthy controls, whereas PCOS patients in Europe had higher diversity; PCOS patients with high testosterone (HT) had different microbial patterns compared to those with low testosterone (LT), including lower levels of *Faecalibacterium* and higher levels of *Prevotella*.

With biomarkers like *Faecalibacterium* and *Prevotella*, PCOS subtypes have distinct gut microbiota compositions that are impacted by geography and testosterone levels. These results highlight the possibility of personalized treatments based on microbiota. However, to handle population variety and improve strain-level assessments, extensive, global research is required. Given this complexity, we are interested in identifying potential biomarkers for all types of PCOS, i.e. disregarding the hyperandrogenism subtypes.

Causal Discovery and the Gut Microbiome

There have been previous attempts to perform causal discovery on the gut microbiome. In particular, (Sazal et al. 2021) attempts to use causal discovery to construct causal networks and implement do-calculus, a causal inference technique developed by (Pearl, Glymour and Jewell 2016) to estimate the causal effects of microbes on other microbes and on outcome variables. For the causal discovery task, they use the PC-stable algorithm (Colombo and Maathuis 2014) which is a variation of PC that removes order-dependence during the estimation of the skeleton of the casual graph. The advantage of PC-stable over PC is that PC may output different results given the order of the conditional independence tests done. After finding the causal graphs, they used do-calculus to quantify the effects of each edge in the graphs which essentially uses the do-operator to intervene on the treatment node, remove all edges pointing towards said node, and to estimate the interventional expectation of the outcome node using a model appropriate for the given data structure like linear regression. They test their pipeline's consistency using simulations and apply their pipeline to real dataset of healthy individuals, individuals with ulcerative colitis (UC), and individuals with Crohn's disease (CD). They used bootstraps to compute confidence intervals for each edge and permutation tests to calculate p-values for the overall network and found bacteria beneficial to UC such as *unclassified Oscillibacter*, *Sutterella wadsworthensis*, and *Bacteroides xylanisolvens*. However, they fail to account for multiple testing issues and covariates in their networks. Since we designed our study before finding this paper, we see a promising role of causal discovery and causal inference in gut microbial data for studying various human diseases.

Additionally, there have been advancement to causal discovery algorithms since the development of the PC and PC-stable algorithms. For example, a variant of the PC algorithm, CD-NOD (Huang et al. 2019), was developed specifically for heterogeneous data, where the heterogeneity of the observed data can help discover the causal structure given certain variables that can change the distribution of the data. This is particularly useful with gut microbiome data where a dataset may contain samples from different studies, hence providing a heterogeneous dataset where the study ID can change the data distribution.

1.2 Data

To answer our research question, we used the NIH Human Microbiome Project (HMP2) dataset (Zhou et al. 2019) for T2D and the aggregated dataset from an individual participant data (IPD) meta analysis and systematic review conducted by (Yang et al. 2024) for PCOS.

The HMP2 dataset (Zhou et al. 2019) followed 106 participants for up to four years, collecting blood, stool, and nasal samples at every self-reported healthy visit and additional visits during periods of respiratory viral infection (RVI), influenza immunization, and other stresses such as antibiotic treatment. Since we are interested in the gut microbes, we look specifically at the visits where gut microbial taxa were profiled using 16S sequencing which provides normalized gut microbe abundance for taxa classified at 6 phyla, 28 classes, 12 orders, 21 families, and 45 genera. As the study authors illustrate, the gut microbiome can fluctuate with the presence of antibiotics and other stressor events such as illness, so we also only look at the visits that were classified as “Healthy”. For each individual, there is information about their race, sex, age, BMI, steady-state plasma glucose (SSPG), and insulin sensitivity classification. For 66 participants, their insulin sensitivity was assessed using an insulin suppression test measured by SSPG: 31 individuals were insulin-sensitive (IS: SSPG $<$ 150 mg/dl), and 35 individuals were insulin-resistant (IR: SSPG $\geq$150 mg/dl). The remaining 40 individuals are classified as unknown due to medical contraindications leading to a lack of insulin suppression tests. Since the dataset is longitudinal but with very few time points per subject, we treated it as a cross-sectional dataset, leaving us with 153 and 178 samples for the IS and IR cohorts respectively.

The IPD meta analysis dataset (Yang et al. 2024) is an aggregation of the 14 studies that were included in the systematic review, but at the individual level. This is different from a meta analysis which analyzes aggregated data or statistics from multiple different studies. Each row of this PCOS dataset represents one sample of gut microbe abundance measurements as well as the sample’s study’s region (Asia or Europe), the sample’s classification as a PCOS patient or a healthy control (HC), and if they were a PCOS patient, whether they had low (LT) or high (HT) testosterone levels. This granularity gives us more data and statistical power behind our results rather than using just one PCOS study. Since the only considerations for confounding their selection criteria specified were no drug interventions, there are other gut microbiome-related confounders that may be present in our data, such as diet, alcohol usage, stress, etc. We examined the study designs of the 14 included studies and found that they varied in external factors including diet, alcohol consumption, the use of antibiotics, and more. Although this is a limitation with the dataset, we chose to continue with this dataset due to its large sample size. This dataset provided us with 1,128 genera and 435 HC & 513 PCOS individuals.

2 Methods

In this study, we use causal discovery algorithms and compare them with predictive modeling to explore the causal relationships between the gut microbiome and two diseases: T2D and PCOS. We used datasets that were cross-sectional, meaning they provide a snapshot of the gut microbiome and disease status at a single point in time, which makes it challenging to determine whether changes in the microbiome cause the disease or are a result of it. Rather than recovering this information from experiments that can be expensive, we can use computational methods to discover causality to the best of the data's ability.

Our approach tackles the complexities of working with high-dimensional data (many microbial features) and relatively small sample sizes. We use feature selection and sure screening techniques to reduce the dimensions of these datasets, and we adjust existing causal discovery algorithms to reduce the multiple testing burden. The goal is to build a framework for understanding how gut microbes contribute to disease and to identify potential targets for personalized treatments.

2.1 Data Preprocessing

For the T2D dataset we removed subjects with an unknown insulin resistance status and selecting only the "Healthy" sample visits. We extracted microbial abundance data at the genus level and converted the values to percentages. The dataset was then merged across subject, sample, and microbial abundance files, with categorical variables like disease status (IRIS), gender, and ethnicity encoded numerically.

For the PCOS dataset, we grouped any unclassified microbial data into a single category and numerically encoded binary variables such as region, and disease status. To account for differences in the study sites, we created a study site variable by manually comparing the study sample sizes and regions.

Based on the suggestions provided by (Weiss et al. 2016) on different correlation strategies to use for different structures of a gut microbe dataset, we filtered out rare operational taxonomy units (OTUs), using a rareness threshold of 1%. This helped reduce features substantially for the PCOS dataset from 1,128 genera to 274 genera.

2.2 Feature Selection and Sure Screening

Given the high-dimensional nature of the PCOS dataset, we experimented with different feature selection and sure screening methods to reduce the feature space before running causal discovery algorithms to reduce the multiple testing burden on the causal discovery algorithms. The two tasks at hand call for different methods. For the microbe-microbe interaction network, since the algorithms start with a complete graph, we used SparCC and graphical lasso separately, to reduce the number of edges between pairs of microbes and removed nodes that were disconnected from any other node. For the microbe-disease interaction network, we used logistic lasso regression to remove features that did not contribute to the prediction of disease status.

SparCC xxx

Graphical Lasso

Logistic Lasso Regression

2.3 Causal Discovery Algorithms

After removing edges and features, we proceed with the causal discovery algorithms. For the microbe-microbe interaction network, we perform a series of conditional independence tests for all pairs of microbes that have an edge between them, conditioned on sets of size 1 and 2. Then, we orient the edges as much as possible using Meek's rules. For the microbe-disease interaction network, we apply CD-NOD using the study site and region as the heterogeneity index.

Our algorithm

CD-NOD

2.4 Variational Autoencoder

3 Results

4 Discussion

5 Conclusion

References

ADA. 2025. "Understanding Type 2 Diabetes." [Link]

Baars, Daniel P., Marcos F. Fondevila, Abraham S. Meijnikman, and Max Nieuwdorp. 2024. "The central role of the gut microbiota in the pathophysiology and management of type 2 diabetes." *Cell Host & Microbe* 32 (8): 1280–1300. [Link]

Colombo, Diego, and Marloes H Maathuis. 2014. "Order-Independent Constraint-Based Causal Structure Learning." *Journal of Machine Learning Research* 15. [Link]

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics* 9 (3): 432–441. [Link]

Friedman, Jonathan, and Eric J Alm. 2012. "Inferring Correlation Networks from Genomic Survey Data." *PLoS Comput Biol* 8 (9). [Link]

Glymour, Clark, Kun Zhang, and Peter Spirtes. 2019. "Review of Causal Discovery Methods Based on Graphical Models." *Frontiers in Genetics* 10. [Link]

Huang, Biwei, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. 2019. "Causal Discovery from Heterogeneous/Nonstationary Data with Independent Changes." *CoRR* abs/1903.01672. [Link]

Mehta, Shruti H., Frederick L. Brancati, Mark S. Sulkowski, Steffanie A. Strathdee, Moyses Szklo, and David L. Thomas. 2000. "Prevalence of Type 2 Diabetes Mellitus among Persons with Hepatitis C Virus Infection in the United States." *Annals of Internal Medicine* 133 (8): 592–599. [Link]

Pearl, Judea, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics—A Primer*. John Wiley & Sons Ltd

Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu, Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle LeChatelier, Pierre Renault, Nicolas Pons, Jean-Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, Jian Wang, S. Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jun Wang. 2012. "A metagenome-wide association study of gut microbiota in type 2 diabetes." *Nature* 490 (7418): 55–60. [Link]

Sazal, Musfiqur, Vitalii Stebliankin, Kalai Mathee, Changwon Yoo, and Giri Narasimhan. 2021. "Causal effects in microbiomes using interventional calculus." *Scientific Reports* 11 (1), p. 5724. [Link]

Tabák, Adam G, Christian Herder, Wolfgang Rathmann, Eric J Brunner, and Mika Kivimäki. 2012. "Prediabetes: a high-risk state for diabetes development." *The Lancet* 379 (9833): 2279–2290. [Link]

Weiss, Sophie, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, Amanda Birmingham, Jacob A Cram, Jed A Fuhrman, Jeroen Raes, Fengzhu Sun, Jizhong Zhou, and Rob Knight. 2016. “Correlation detection strategies in microbial data sets vary widely in sensitivity and precision.” *The ISME Journal* 10 (7): 1669–1681. [Link]

WHO. 2025. “Polycystic ovary syndrome.” [Link]

Yang, Yanan, Jiale Cheng, Chongyuan Liu, Xiaopo Zhang, Ning Ma, Zhi Zhou, Weiying Lu, and Chongming Wu. 2024. “Gut microbiota in women with polycystic ovary syndrome: an individual based analysis of publicly available data.” *eClinicalMedicine* 77 . [Link]

Zhou, Wenyu, M. Reza Sailani, Kévin Contrepois, Yanjiao Zhou, Sara Ahadi, Shana R. Leopold, Martin J. Zhang, Varsha Rao, Monika Avina, Tejaswini Mishra, Jethro Johnson, Brittany Lee-McMullen, Songjie Chen, Ahmed A. Metwally, Thi Dong Binh Tran, Hoan Nguyen, Xin Zhou, Brandon Albright, Bo-Young Hong, Lauren Petersen, Eddy Bautista, Blake Hanson, Lei Chen, Daniel Spakowicz, Amir Bahmani, Denis Salins, Benjamin Leopold, Melanie Ashland, Orit Dagan-Rosenfeld, Shannon Rego, Patricia Limcaoco, Elizabeth Colbert, Candice Allister, Dalia Perelman, Colleen Craig, Eric Wei, Hassan Chaib, Daniel Hornburg, Jessilyn Dunn, Liang Liang, Sophia Miryam Schüssler-Fiorenza Rose, Kim Kukurba, Brian Piening, Hannes Rost, David Tse, Tracey McLaughlin, Erica Sodergren, George M. Weinstock, and Michael Snyder. 2019. “Longitudinal multi-omics of host–microbe dynamics in prediabetes.” *Nature* 569 (7758): 663–671. [Link]

Appendices

Please see a copy of our project proposal.

Contributions

MPM, CS, and NZ designed the project. MPM and CS found datasets. CS and NZ performed EDA. CS built the microbe-microbe and microbe-disease interaction networks. MPM built the VAE model. BH proposed the causal discovery algorithm for the microbe-microbe interaction network. JB proposed the feature reduction and sure screening methods. BH and JB provided insightful comments and suggestions to the design. MPM, CS, and NZ interpreted the results and wrote the final report.