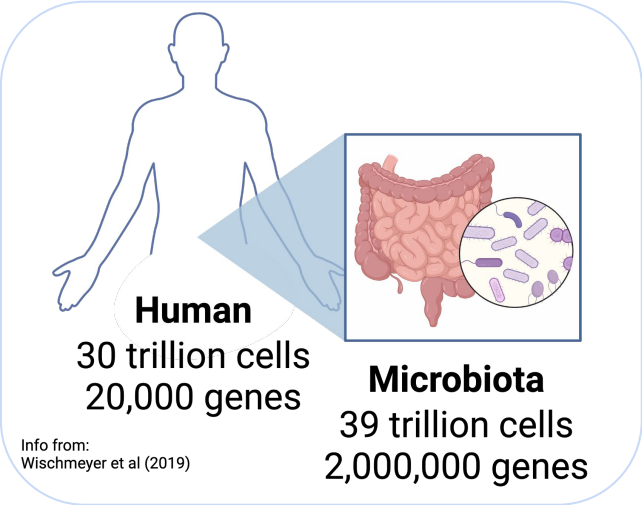
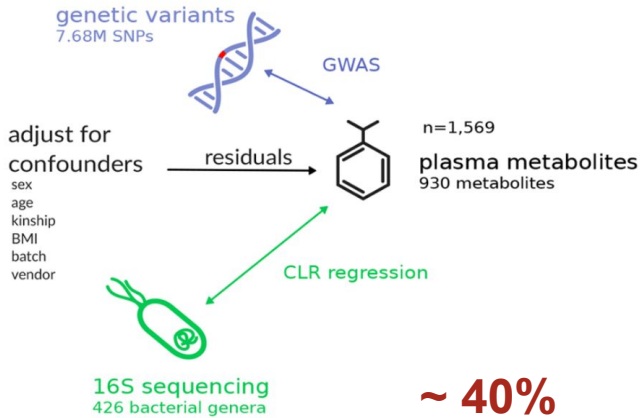
DietxMeds

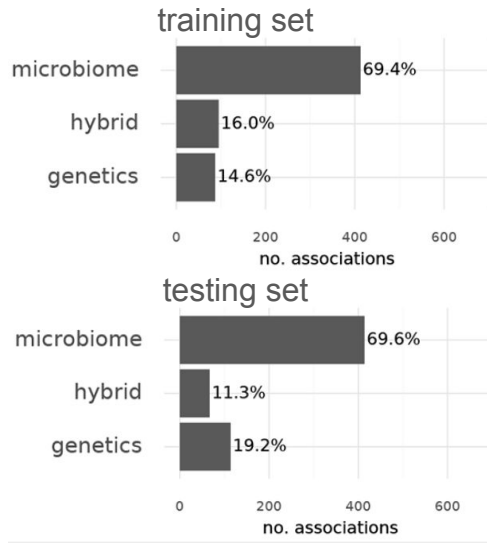
Anika O’Brian, Mia Giallorenzi, Melissa Hopkins,
Avery Yang, Crystal Perez

Can we identify foods that affect drug efficiency using metabolites, microbes, and client information from human stools?

What is the human gut microbiome?





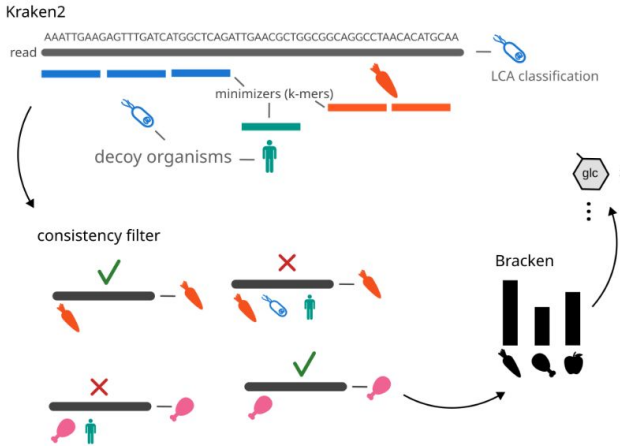


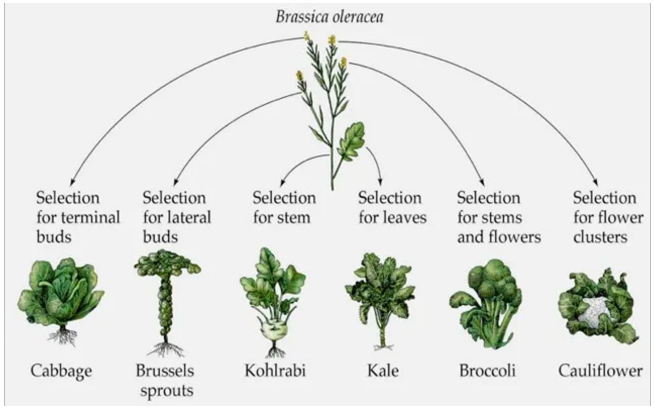
Diener et al 2023

diet → microbes → metabolome

Metagenomic Estimation of Dietary Intake (MEDI)

- Identify **food-derived DNA** genomes in stool sequencing data
- For each sample, find the LCA of all reads and reject inconsistencies
- **Estimate the distribution of species** and cultivars
- Convert species and abundance to nutrient content



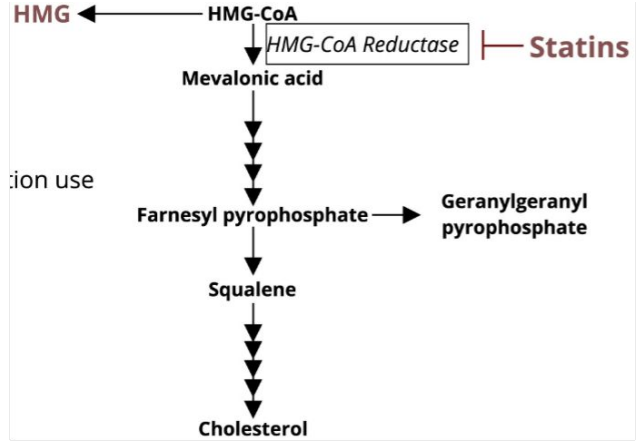


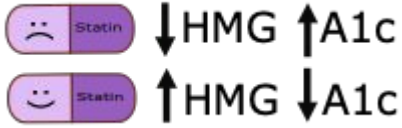
1. Figure 2A from Diener, C et al. (*Nat. Metab.*, 2025)
2. https://platform.vox.com/wp-content/uploads/sites/2/chorus/uploads/chorus_asset/file/3395076/brassica-oleracea.0.jpg?quality=90&strip=all&crop=0,0,100,100

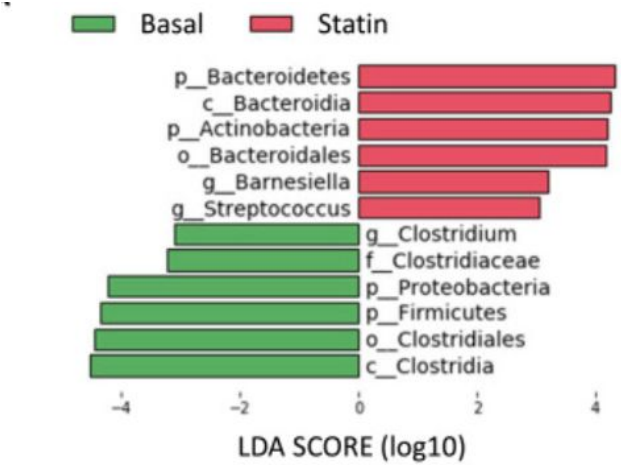
Statins ↓ LDL cholesterol and preventing heart disease

Why MaxStat?

- Understanding factors that enhance or inhibit their effectiveness can lead to **better patient outcomes**
- Explore **discrepancy in patient outcomes**: why do some patients benefit more than others?
- **Minimize side effects**: insulin resistance/diabetes, muscle soreness
- Apply to other medications for **microbiome-informed drug therapies**



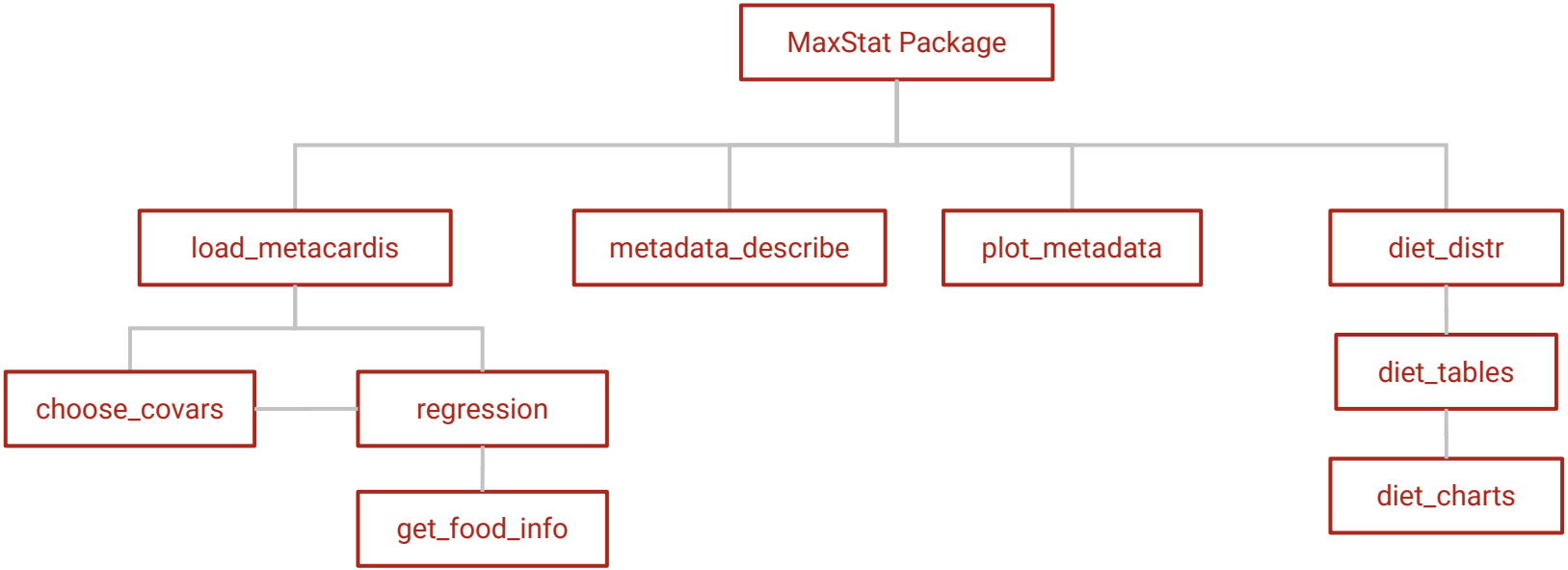




Source: [Wilmanski et al. 2019]

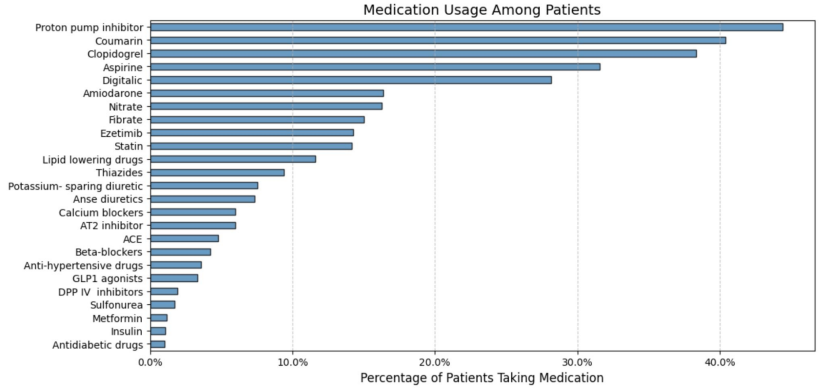
Design Overview

Develop a package to easily identify and interpret relationships between MEDI data and other datasets, e.g., medication and microbiome data

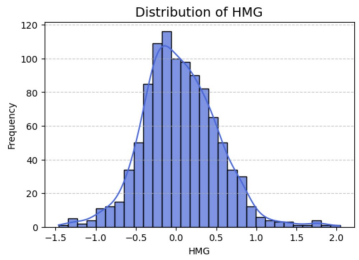


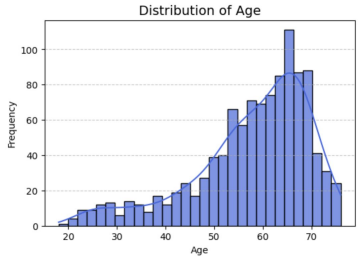
Meet Shelly, our potential user!

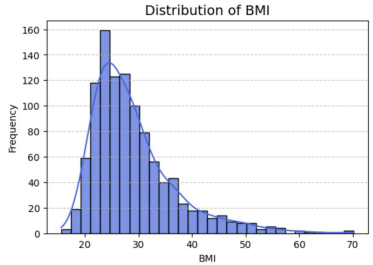
Metadata_describe and plot_metadata

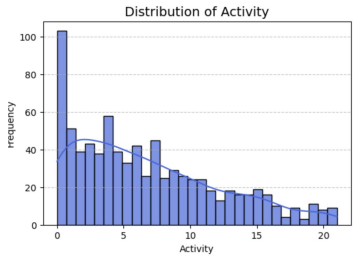


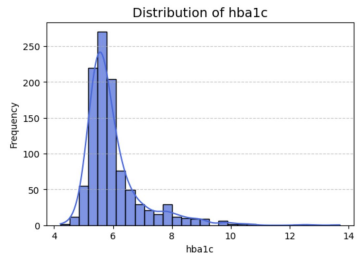
	hmg	age	bmi	activity	hba1c	seq_depth	microbial_load	id	status	gender	...	statin
count	1029.0	1087.0	1052.0	815.0	1037.0	1015.0	1001.0	1087.0	1087.0	1066.0	...	1087.0
mean	0.051626	57.732291	28.972378	6.738229	6.005477	22157304.602956	1192504782 87.519485	NaN	NaN	NaN	...	NaN
std	0.46543	12.064403	7.754655	5.399989	1.00902	3967365.632905	57887649559.728127	NaN	NaN	NaN	...	NaN
min	-1.46283	18.0	15.60574	0.0	4.21	18422001.0	9241706161.0	NaN	NaN	NaN	...	NaN
25%	-0.24986	52.0	23.668425	2.255208	5.4	20436709.5	74637681159.0	NaN	NaN	NaN	...	NaN
50%	0.01854	60.0	27.146335	5.6875	5.7	20968630.0	114139344262.0	NaN	NaN	NaN	...	NaN
75%	0.348438	66.0	32.087075	10.125	6.2	22152640.0	158121827411.0	NaN	NaN	NaN	...	NaN
max	2.042377	76.0	70.08356	21.0	13.7	63260733.0	310069444444.0	NaN	NaN	NaN	...	NaN
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1087.0	4.0	2.0	...	NaN
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	x10MCx1135	IHD372	Male	...	NaN
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	372.0	655.0	...	NaN
mean (proportion)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.383625
sum (total 1s)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	417.0

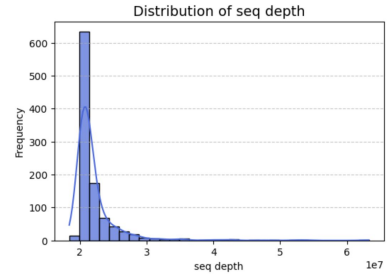


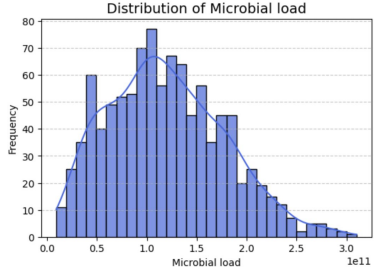


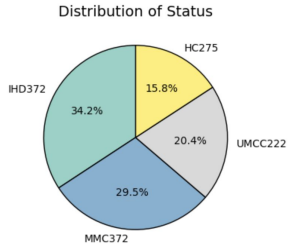


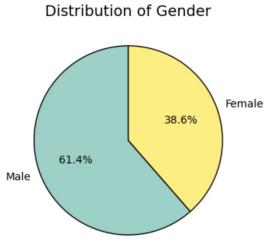


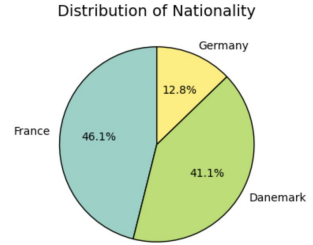




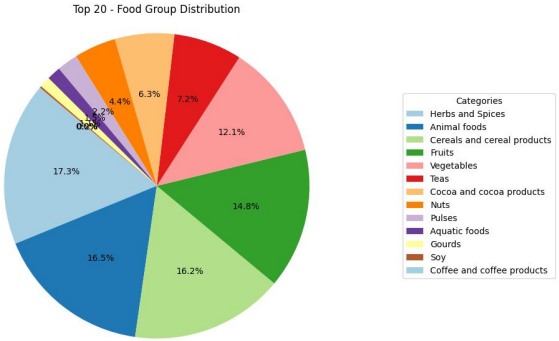




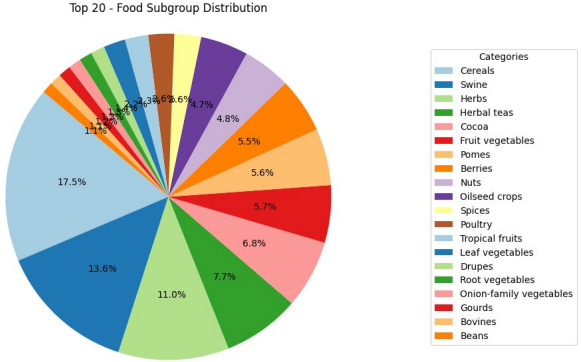




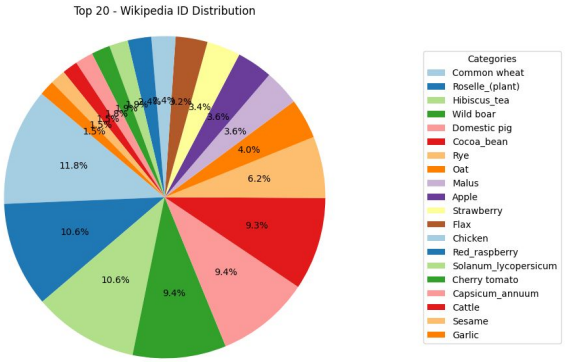
Diet_tables and Diet_charts



	food_group	count	percentage_of_total
0	Herbs and Spices	1281	17.292117
1	Animal foods	1226	16.549676
2	Cereals and cereal products	1203	16.239201
3	Fruits	1099	14.835313
4	Vegetables	897	12.108531
5	Teas	536	7.235421
6	Cocoa and cocoa products	465	6.276998
7	Nuts	329	4.441145
8	Pulses	163	2.200324
9	Aquatic foods	109	1.471382
10	Gourds	85	1.147408
11	Soy	14	0.188985
12	Coffee and coffee products	1	0.013499

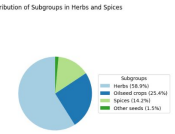


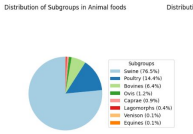
	food_subgroup	count	percentage_of_total
0	Cereals	1203	16.239201
1	Swine	938	12.661987
2	Herbs	755	10.191685
3	Herbal teas	527	7.113931
4	Cocoa	465	6.276998
5	Fruit vegetables	391	5.278078
6	Pomes	387	5.224082
7	Berries	376	5.075594
8	Nuts	329	4.441145
9	Oilseed crops	325	4.387149
10	Spices	182	2.456803
11	Poultry	177	2.389309
12	Tropical fruits	159	2.146328
13	Leaf vegetables	149	2.011339
14	Drupes	95	1.282397
15	Root vegetables	89	1.201404
16	Onion-family vegetables	86	1.160907
17	Gourds	85	1.147408
18	Bovines	78	1.052916
19	Beans	77	1.039417

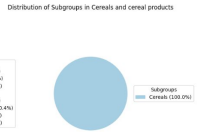


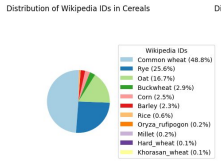
	wikipedia_id	count	percentage_of_total
0	Common wheat	587	8.002727
1	Roselle_(plant)	527	7.184731
2	Hibiscus_tea	527	7.184731
3	Wild boar	469	6.394001
4	Domestic pig	469	6.394001
5	Cocoa_bean	465	6.339468
6	Rye	308	4.199046
7	Oat	201	2.740286
8	Malus	179	2.440354
9	Apple	179	2.440354
10	Strawberry	169	2.304022
11	Flax	160	2.181322
12	Chicken	121	1.649625
13	Red_raspberry	118	1.608725
14	Solanum_lycopersicum	94	1.281527
15	Cherry tomato	94	1.281527
16	Capsicum_annuum	90	1.226994
17	Cattle	77	1.049761
18	Sesame	77	1.049761
19	Garlic	76	1.036128

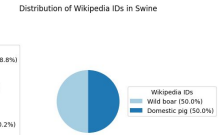
Diet_charts Distribution of Food Subgroups and Wiki IDs

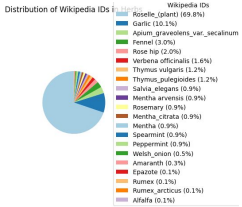


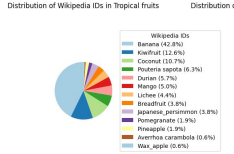


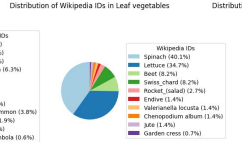


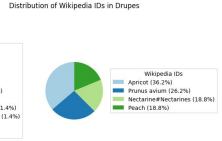


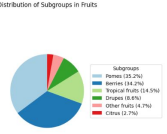


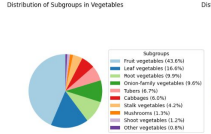


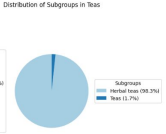


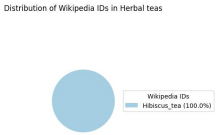


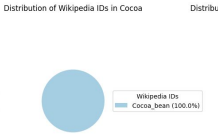


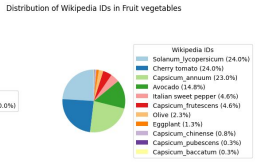


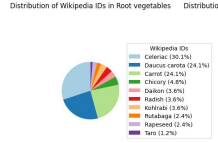


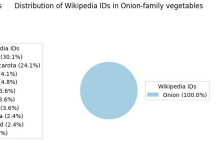


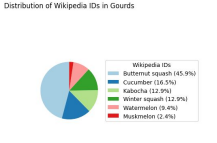


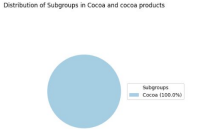


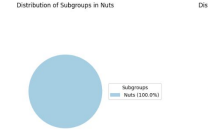


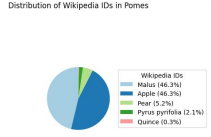


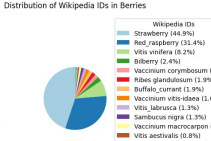


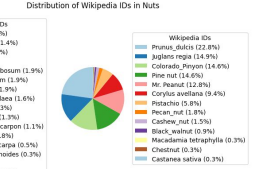


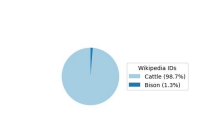


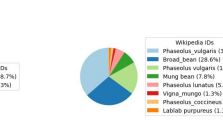


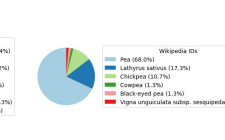


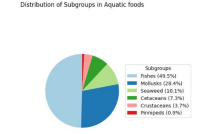


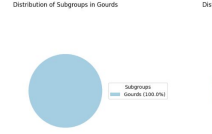


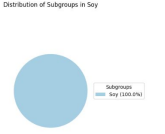


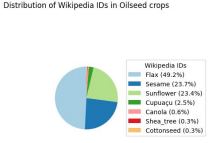


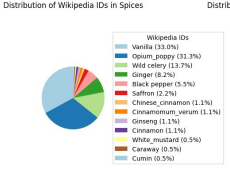


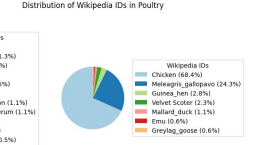


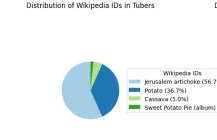


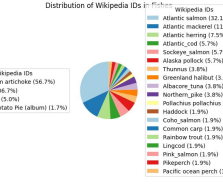


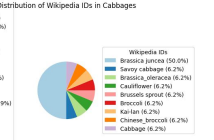


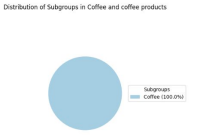




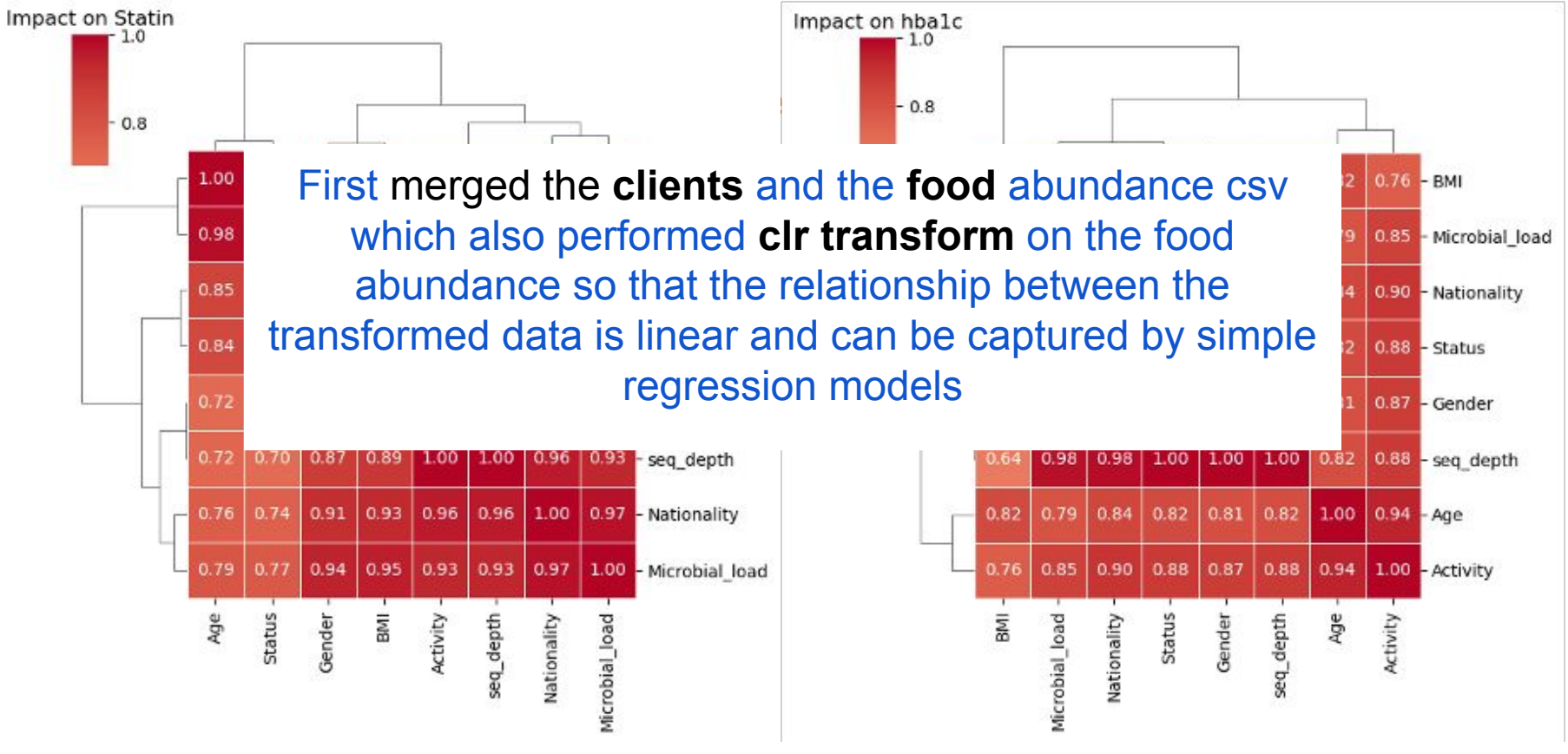






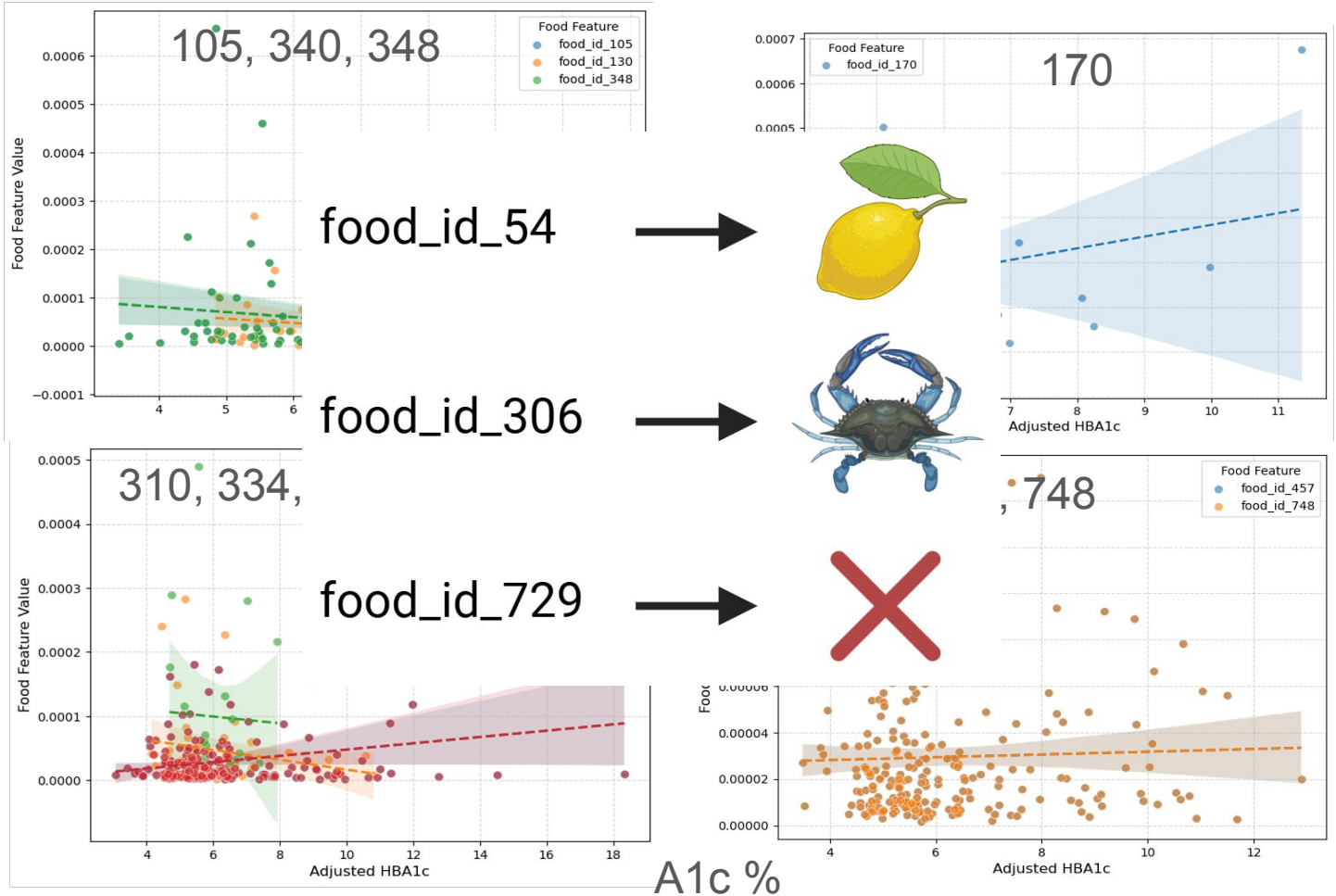


load_metacardis → choose_covars → regression

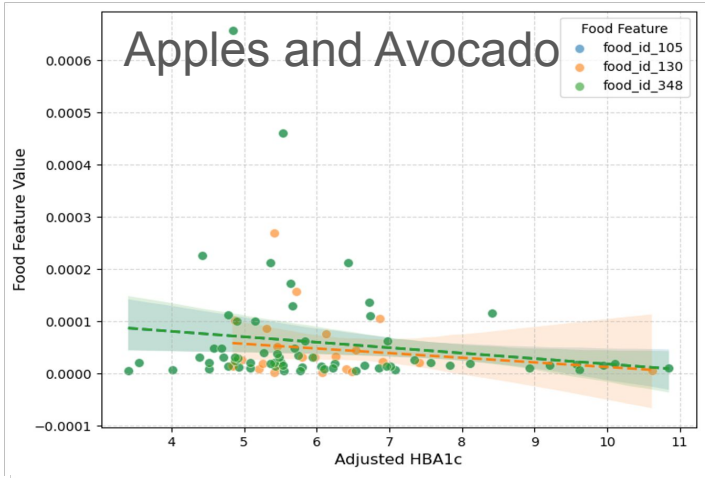


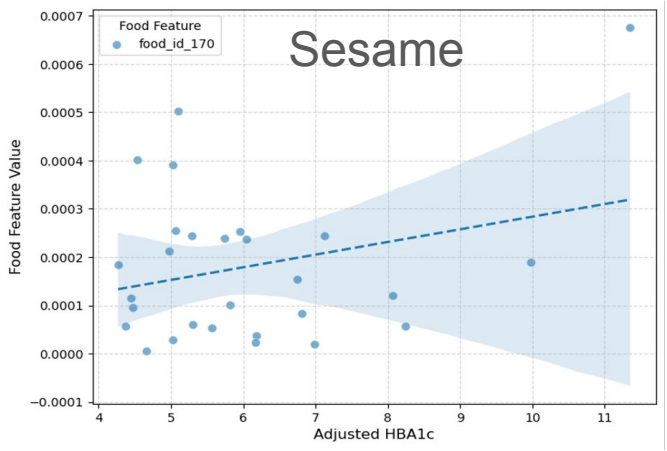
First merged the **clients** and the **food** abundance csv which also performed **clr transform** on the food abundance so that the relationship between the transformed data is linear and can be captured by simple regression models

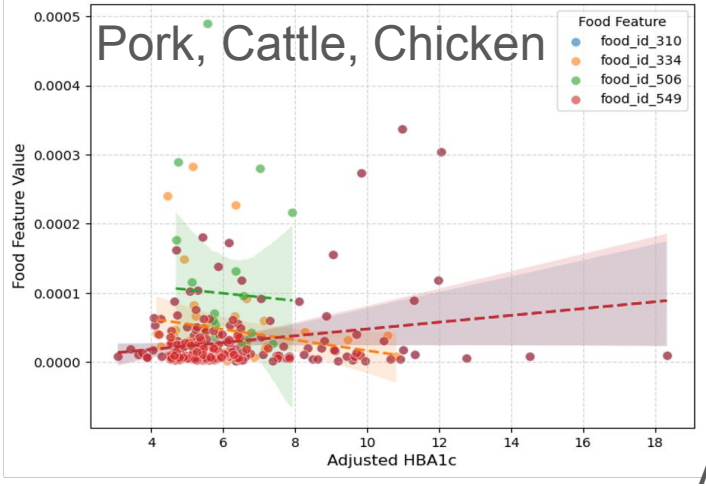
Results

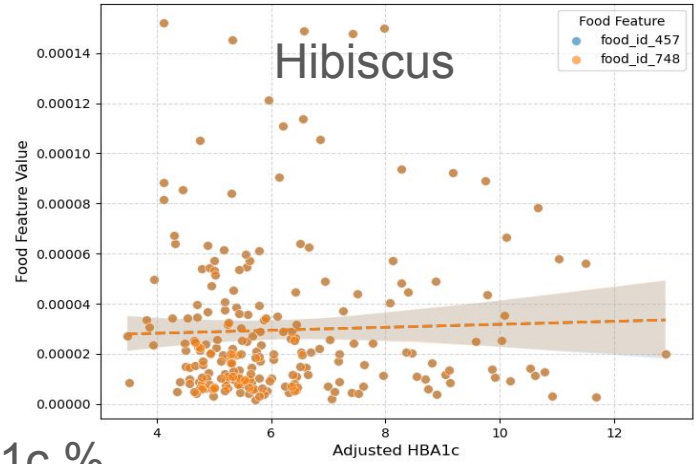

Food Abundance
105, 340, 348
Food Feature
food_id_105
food_id_130
food_id_348
Food Feature Value
food_id_54
170
Food Feature
food_id_170
Adjusted HBA1c
food_id_306
310, 334,
748
food_id_729
Food Feature
food_id_457
food_id_748
Adjusted HBA1c
A1c %

Results









Food Abundance

A1c %

Packages utilized in our components

- Data wrangling: pandas
- Data visualization: matplotlib and seaborn
- Linear regression: statsmodels

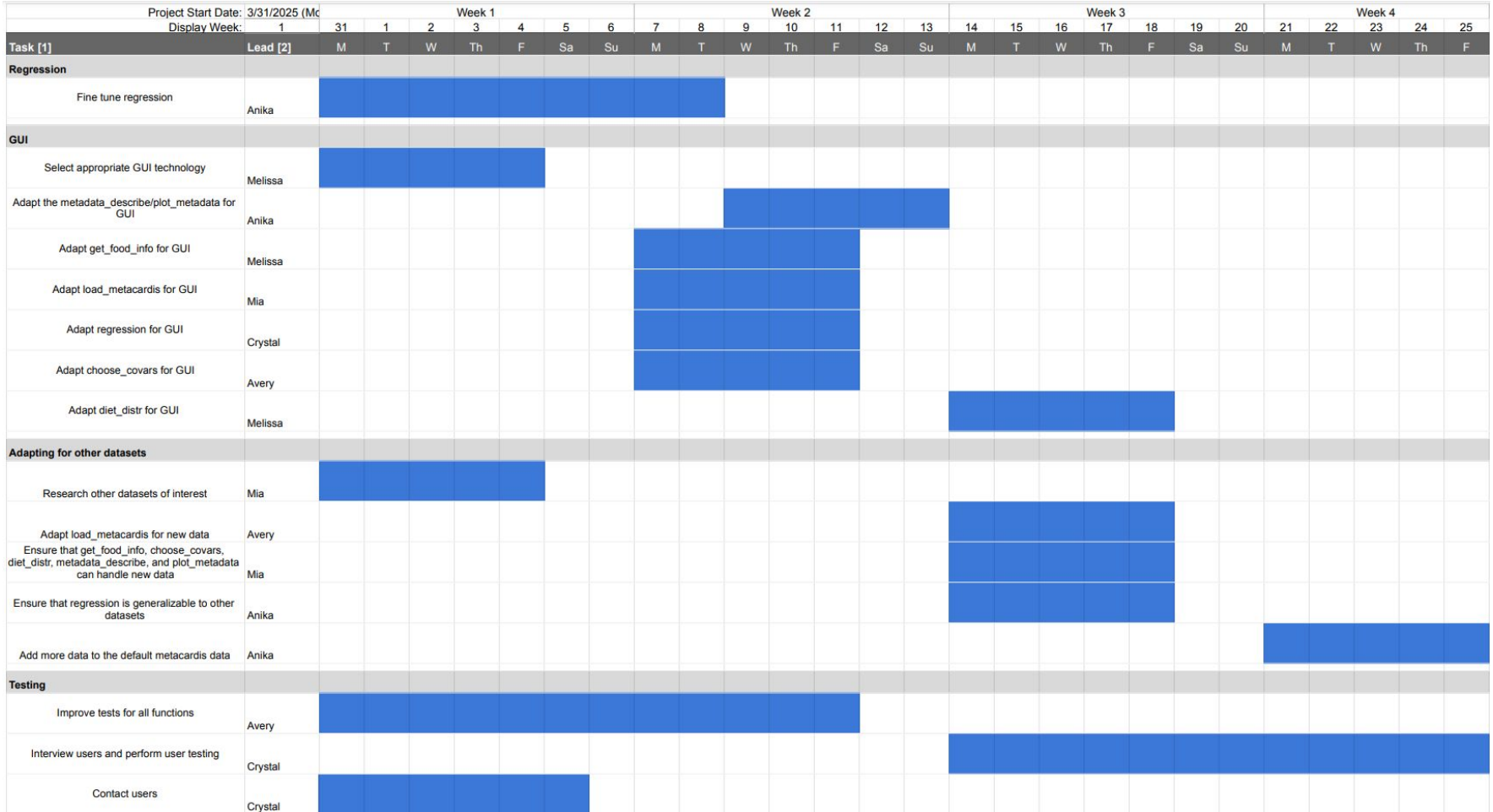
Statsmodels: ols

- More flexible in terms of model construction
- Statistical tests and FDR correction performed easily
- Good for inference and interpretation
- Intuitive

Sklearn: linear regression

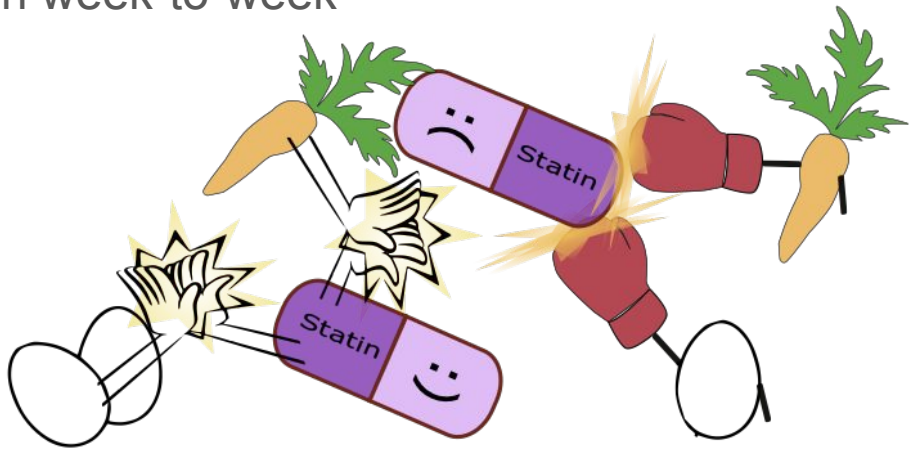
- Not as flexible for model construction (for example must encode categorical variables yourself)
- Good for model selection and prediction
- Lots of emphasis on model performance

Schedule for Future Work



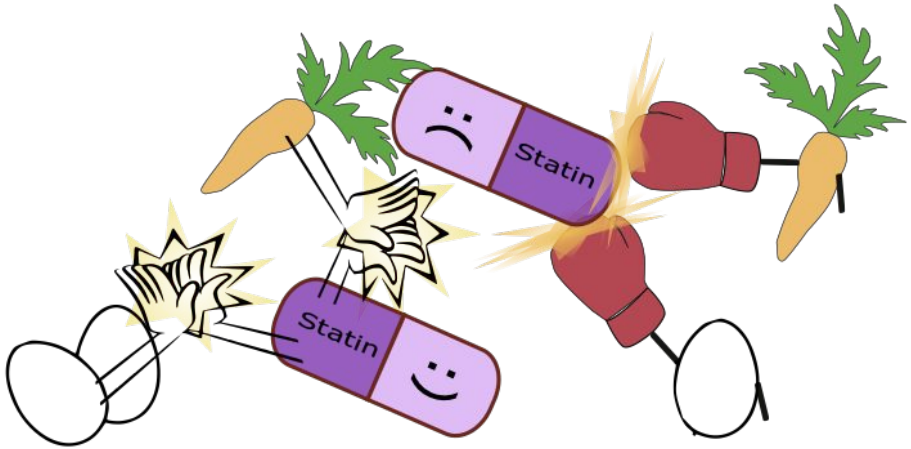
What do we wish we knew earlier?

- Where do you draw the line for a single component (vs multiple)?
- How to work effectively in a group on GitHub
- More practice with GitHub before we had to start using it for the project
- More of a timeline of deliverables for the project/more info about what we were working on week-to-week



Questions?

Thank you (:



Additional Information

Snippet of User Stories

2. Dr. Rodriguez:

Who: Dr. Rodriguez is a clinical researcher working with statin patients

Wants: He wants to identify which foods his patients should eat or avoid to optimize their statin treatment outcomes.

Needs: He needs clear, actionable dietary recommendations based on patient-specific microbiome profiles and statin responses.

Skill: He has basic computer skills and no programming experience.

3. Maya:

Who: Maya is a bioinformatician specializing in microbiome data analysis.

Wants: She wants to integrate new analysis methods into the MEDI tool and customize the food classification algorithms.

Needs: She needs programmatic access to raw data and the ability to modify analysis pipelines.

Skill: Maya is proficient in Python, R, and bioinformatics tools.

4. Tom:

Who: Tom is a new graduate student

Wants: He wants to explore how different gut bacteria influence statin metabolism and insulin resistance.

Needs: He needs well-documented tutorials and example analyses to guide his research.

Skill: Tom has basic programming skills and wants to learn new analytical methods.

Snippet of detailed component specifications

1. **Regression:**
   What it does: Runs a linear regression model with merged data
   Inputs (with type information): list of dependent variables, list of independent variables, dataframe containing all data, and a formula string
   Outputs (with type information): dataframe with significant features, beta coefficients, t-statistics, p-values, and FDR adj. q-values.
   Components used: Merged dataset from load_metacardis, statsmodels ols, and statsmodels multipletests
   Main effects: Runs regression function and creates a new dataframe with test information
   Side effects: Can increase memory usage and computational overhead if dataset is large

2. **diet_dist:**
   What it does: Reshapes the food_abundance dataframe to group together all the foods found for each patient
   Inputs (with type information): food_abundance csv file where each row is a different wiki id food for a patient
   Outputs (with type information): reshaped dataframe in csv format where each patient has one row and multiple columns for the different foods
   Components used: pandas
   Side effects: Might increase memory usage if a large dataframe is used, as it reshapes that and saves it as a new dataframe

5. **metadata_describe():**
   What it does: Outputs tables of statistics about the data (from the metacardis data)
   Inputs (with type information): Data of interest (pandas dataframe)
   Outputs (with type information): Pandas dataframe containing concatenated .describe() functions ran on metadata columns of metacardis data. Some columns will include age, sex (% F), BMI, Nationality, health status, and medication use.
   Components used: Pandas
   Side effects: Can increase memory usage and computational overhead if dataset is large

7. **get_food_info():**
   What it does: Converts food IDs to names and other information
   Inputs (with type information): Food IDs of interest (integers, strings of form “food_id_[int]”, or 1D lists/arrays with any assortment of the 2 data types)
   Outputs (with type information): Food information of interest (as a pandas dataframe)
   Components used: Numpy, Pandas
   Side effects: Temporarily loads the entire food database into a dataframe, which contains several hundred rows.