Predictive models of fecal microbial biomarkers for obesity trajectories in preschool children

Mentor: Prof., Dr. Anita Kozyrskyj | Mentee: Xin (David) Zhao

CHILD COHORT STUDY

SYMBIOTA SYNERGY IN MICROBIOTA RESEARCH

UNIVERSITY OF TORONTO

UNIVERSITY OF ALBERTA

UNIVERSITY OF ALBERTA

Facts and innovation

1. Children in Canada were often referred to the weight management services as late as teen or tween.

❖ **Predictive model: see invisible signs**

2. Existing models mostly predict the static status of obesity at single time point.

❖ **BMI trajectory: age-of-onset, intensity, duration**

3. Data availability increases of (infant) gut microbiota.

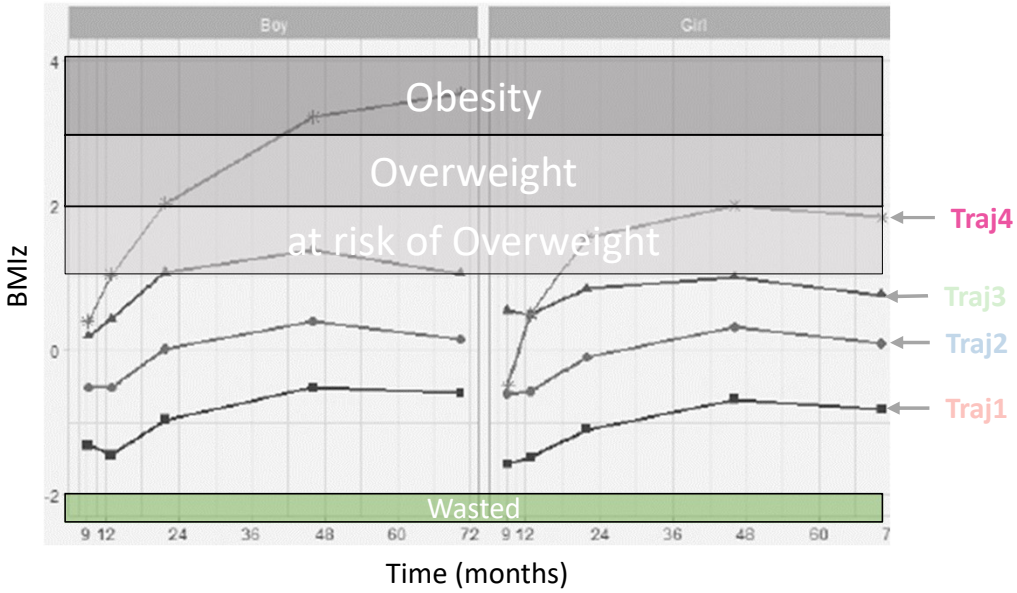
❖ **Fecal biomarkers: diagnosis and therapeutics**

Can fecal microbial features improve the prediction of childhood BMI trajectory?

Precision Medicine

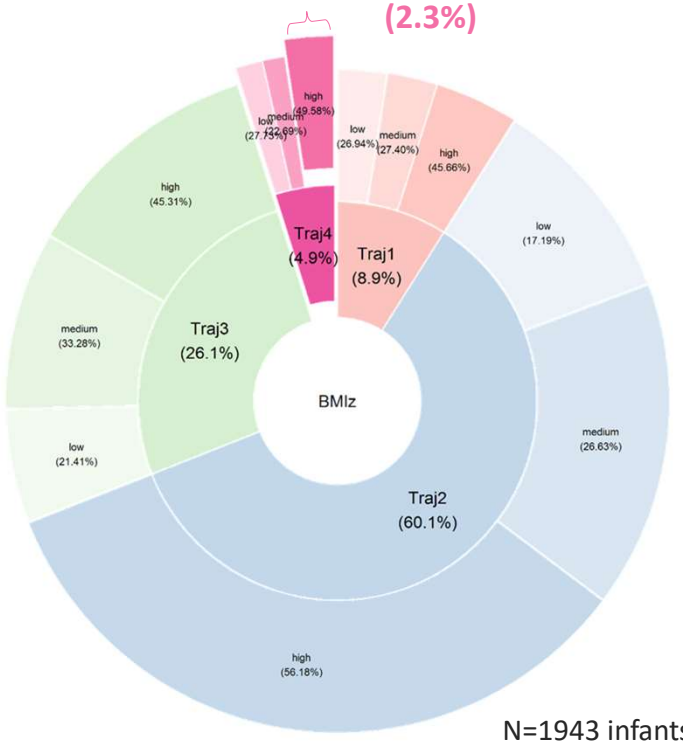
Outcome variable

Latent BMIz trajectory (Credit to Myrtha)



low-risk (97.7%)

high-risk* (2.3%)



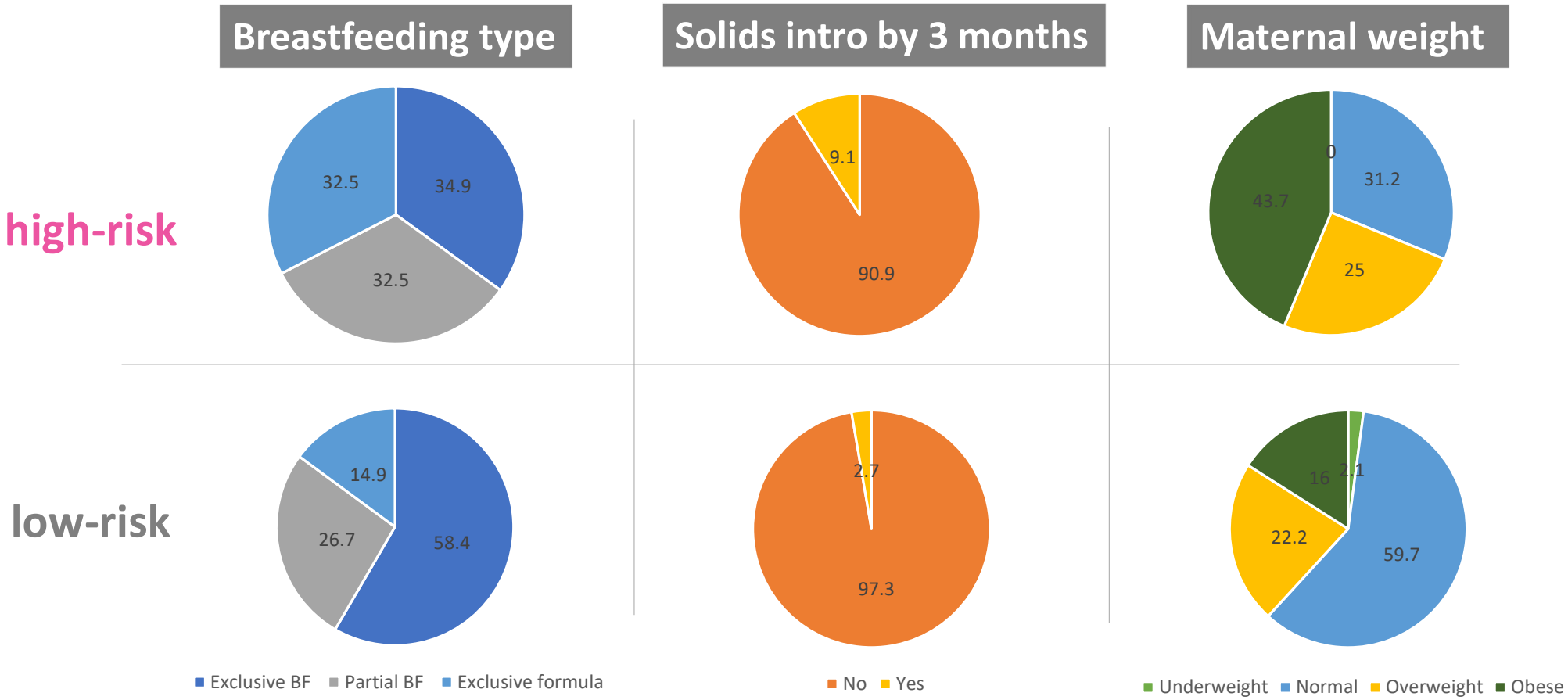
N=1943 infants

high-risk*= Early onset and sustained obesity trajectory

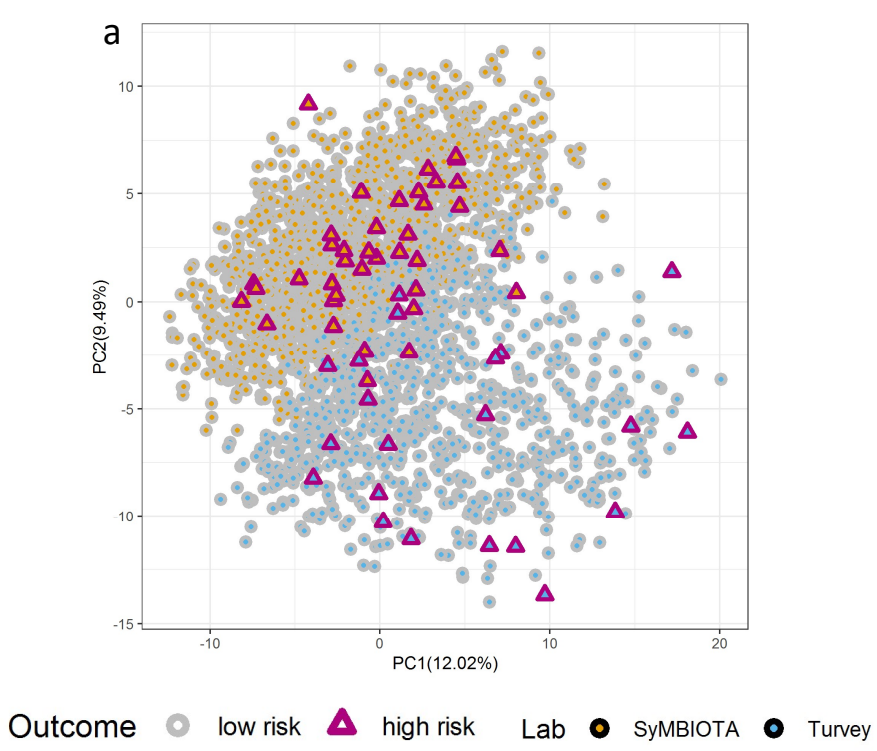
low 0-65%; medium 65-85%; high 85-100% posterior probability

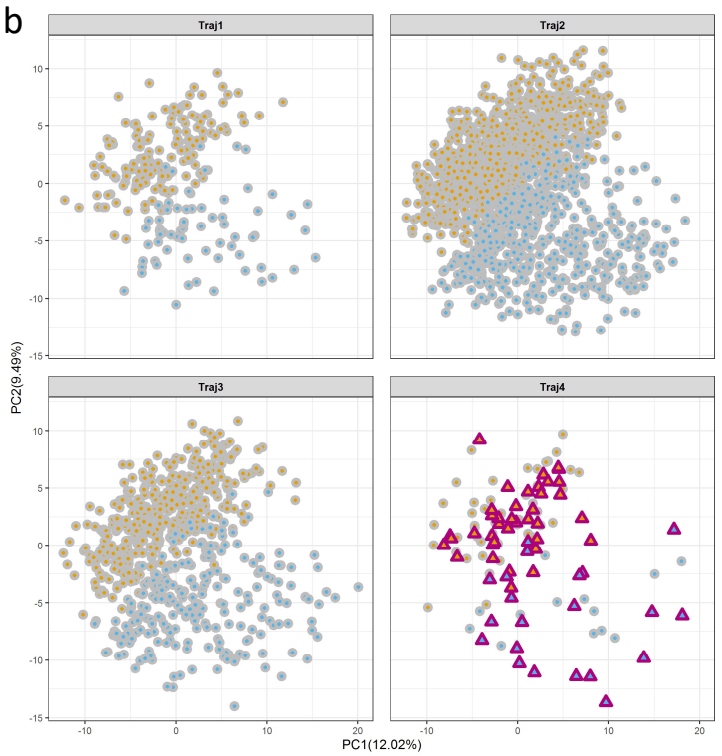
UNIVERSITY OF ALBERTA

Covariables associated with the high-risk group



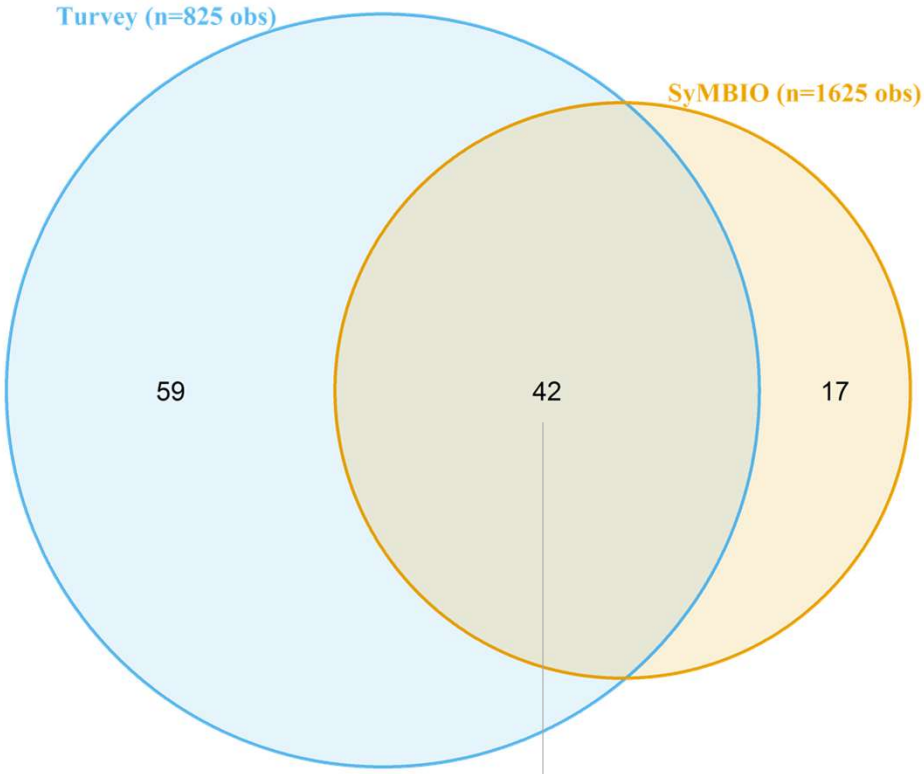
∑ = 2450 stool samples (507 repeat measurements) = 1625 SyMBIOTA + 825 Turvey





Stool samples were collected at 3 months of age

Merge OTU tables from two labs-1



shared OTU | Lab-exclusive OTU

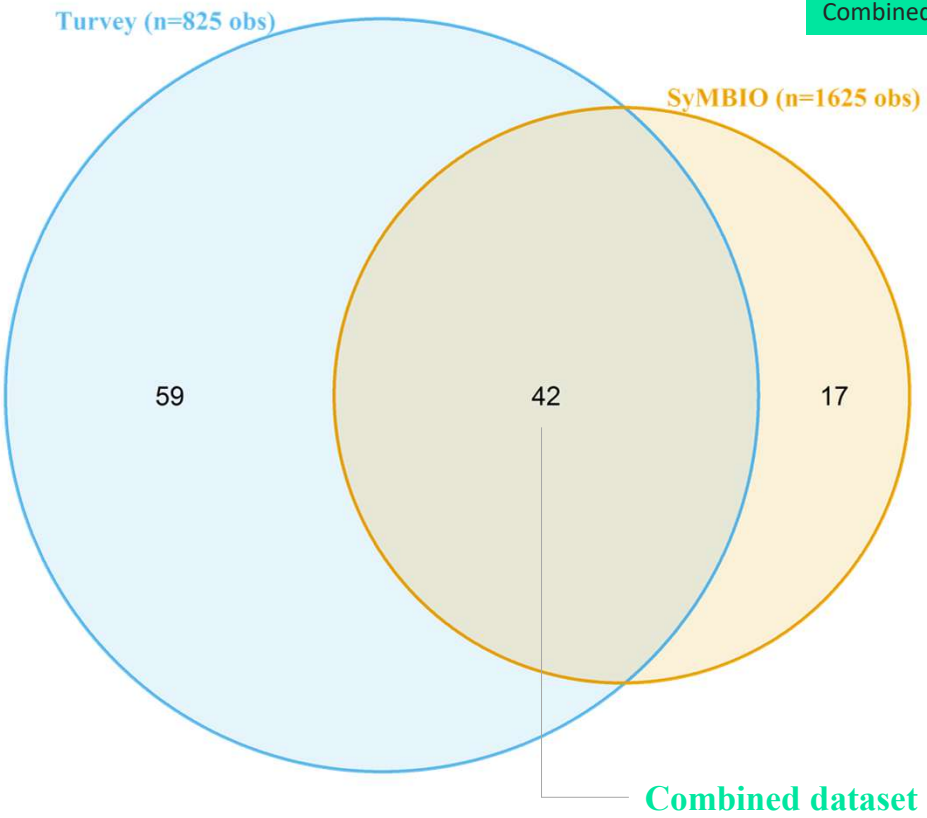
Turvey

ID	OTU1	…	OTU42	OTU_a	OTU_b	OTU_c	OTU_d
001							
…							
825							

SyMBIO

ID	OTU1	…	OTU42	OTU_A	OTU_B
001					
002					
003					
…					
1625					

Merge OTU tables from two labs-2

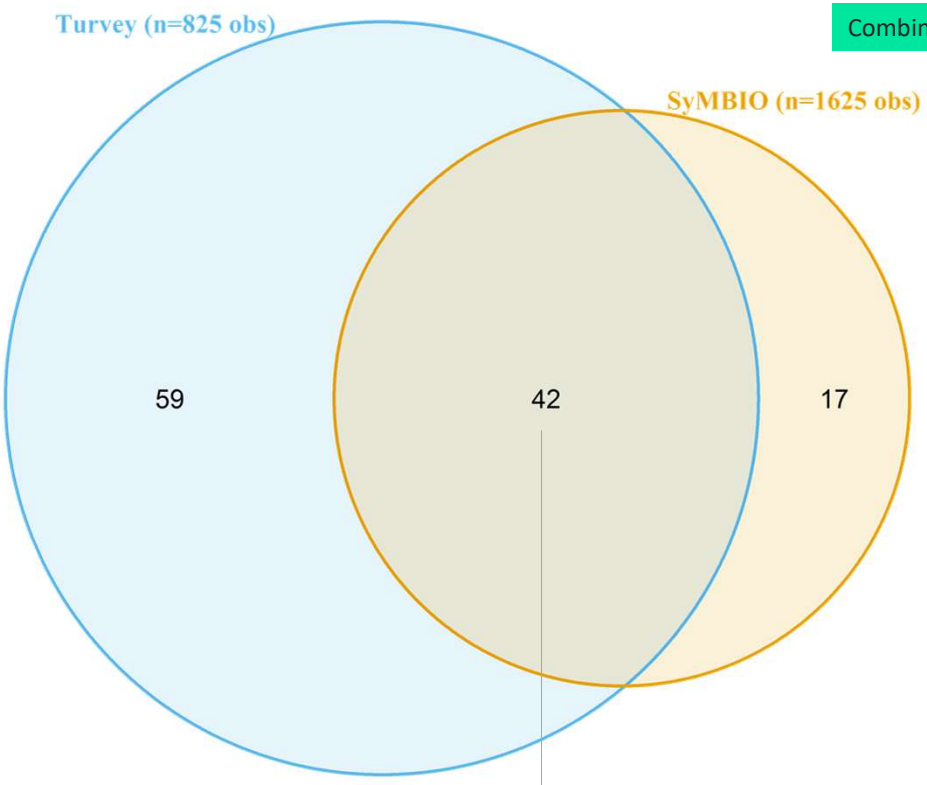


Combined

shared OTU | Lab-exclusive OTU

ID	OTU1	…	OTU42	OTU_a	OTU_b	OTU_c	OTU_d	OTU_A	OTU_B
001								NA	NA
…								NA	NA
825								NA	NA
826				NA	NA	NA	NA		
827				NA	NA	NA	NA		
828				NA	NA	NA	NA		
…				NA	NA	NA	NA		
2450				NA	NA	NA	NA		

Merge OTU tables from two labs-3

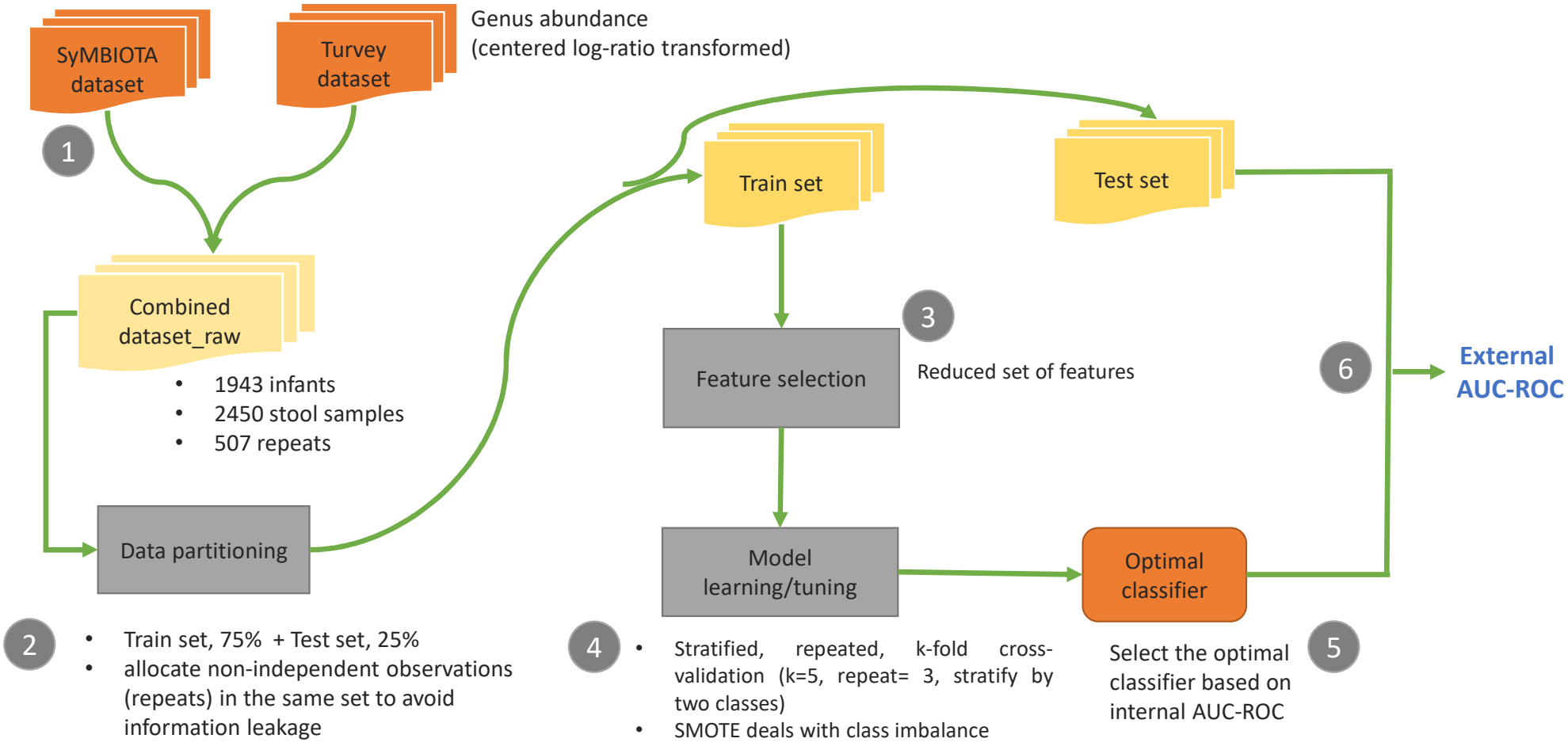


Combined

shared OTU

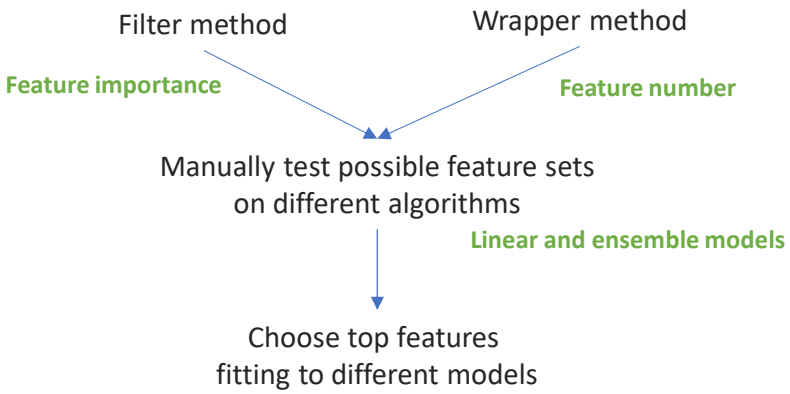
ID	OTU1	…	OTU42
001			
…			
825			
826			
827			
828			
…			
2450			

Pipeline for machine learning modeling



Feature selection was performed on the train set.

- **Filter method (univariate association)**
  - pros: features importance (eg. effect size)
  - cons: subjective determination of threshold (features number)
- **Wrapper method (Recursive feature elimination, RFE)**
  - pros: optimal feature number and combination
  - cons: algorithm-dependent

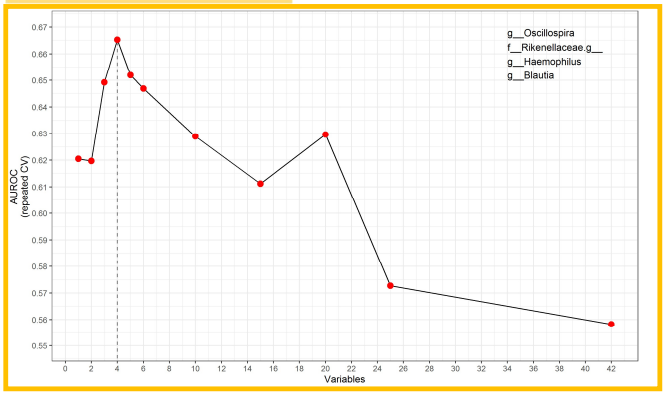


Filter method

Genus taxonomy [effect size] cut-off p<0.05

g__*Oscillospira* [0.08] ✓
g__*Blautia* [0.07] ✓
f__*Rikenellaceae*.g__ [0.07] ✓
g__*Haemophilus* [0.06] ✓
g__*Phascolarctobacterium* [0.06] ✓
g__*Parabacteroides* [0.04]
g__*Staphylococcus* [0.04]

Wrapper method



Feature number: 4-6 (out of 42), depending on algorithms

Regularized logistic regression model-dependent

Menu of predictive models: **Internal validation**

Train set (75%)

Model [1]	Model compartments				
	~microbe[3]	~microbe + lab	~microbe + city	~microbe + gender	~microbe + lab + city + gender
RegLog	0.68 (0.62, 0.74)	0.70 (0.66, 0.73)	0.70 (0.64, 0.76)	0.69 (0.62, 0.75)	0.69 (0.62, 0.76)
GLM	0.68 (0.63, 0.73)	0.68 (0.63, 0.72)	0.69 (0.63, 0.74)	0.68 (0.62, 0.74)	0.69 (0.64, 0.73)
Random Forest	0.58 (0.50, 0.64)	0.62 (0.58, 0.65)	0.62 (0.55, 0.67)	0.61 (0.53, 0.69)	0.62 (0.53, 0.67)
XGBoost	0.63 (0.56, 0.70)	0.64 (0.58, 0.69)	0.67 (0.61, 0.73)	0.65 (0.60, 0.70)	0.65 (0.59, 0.70)
GLMM[2]	0.80 (0.77, 0.83)	0.78 (0.76, 0.80)	0.80 (0.77, 0.82)	0.81 (0.78, 0.84)	0.81 (0.78, 0.84)

[1]The experimental unit for the models, RegLog, GLM, RF and XGBoost was observations (i.e., stool samples); whereas the experimental unit was infants for the GLMM model.

[2]GLMM: generalized linear mixed model. In GLMM, participants are the experimental unit. Lab, city, and participants identify served as (crossed) random-effect variables while the infant gender and genera abundance as fixed-effect variables in the model.

[3]Microbe predictors used in all the models consist of f__*Rikenellaceae*.g__, g__*Blautia*, g__*Oscillospira*, g__*Haemophilus*, and g__*Phascolarctobacterium*.

Menu of predictive models: **External validation**

Test set (25%)

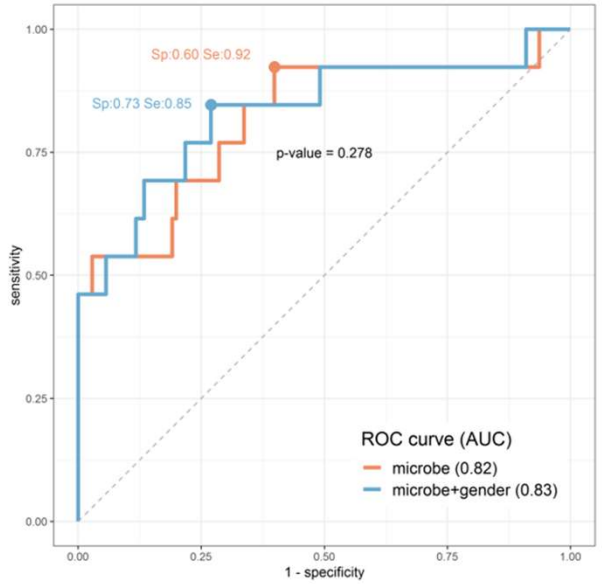
Model [1]	Model compartments				
	~microbe[3]	~microbe + lab	~microbe + city	~microbe + gender	~microbe + lab + city + gender
RegLog	0.69 (0.59, 0.79)	0.67 (0.56, 0.77)	0.70 (0.59, 0.80)	0.66 (0.55, 0.77)	0.68 (0.57, 0.78)
GLM	0.70 (0.60, 0.78)	0.69 (0.58, 0.78)	0.69 (0.58, 0.78)	0.68 (0.57, 0.78)	0.67 (0.55, 0.77)
Random Forest	0.55 (0.43, 0.67)	0.56 (0.43, 0.68)	0.57 (0.44, 0.69)	0.59 (0.45, 0.70)	0.63 (0.51, 0.76)
XGBoost	0.59 (0.44, 0.72)	0.63 (0.49, 0.75)	0.61 (0.48, 0.74)	0.65 (0.50, 0.77)	0.68 (0.55, 0.80)
GLMM[2]	**0.82 (0.70, 0.92)**	**0.75 (0.63, 0.85)**	**0.82 (0.68, 0.92)**	**0.84 (0.70, 0.94)**	**0.84 (0.70, 0.94)**

[1]The experimental unit for the models, RegLog, GLM, RF and XGBoost was observations (i.e., stool samples); whereas the experimental unit was infants for the GLMM model.

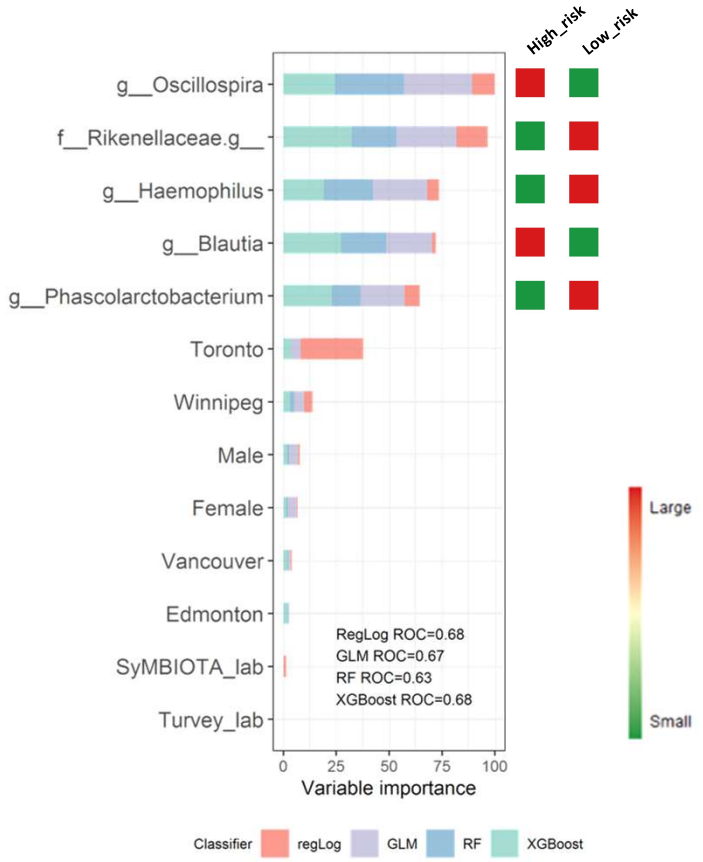
[2]GLMM: generalized linear mixed model. In GLMM, participants are the experimental unit. Lab, city, and participants identify served as (crossed) random-effect variables while the infant gender and genera abundance as fixed-effect variables in the model.

[3]Microbe predictors used in all the models consist of f__*Rikenellaceae*.g__, g__*Blautia*, g__*Oscillospira*, g__*Haemophilus*, and g__*Phascolarctobacterium*.

Interpretation and biomarkers



GLMM base model: AUC 0.82 (0.70, 0.92) on test set



Existing microbiota-based ML models

Outcome	Predictor	AUC-ROC	Algorithm	Sample size	Region of cohort	Year of publication
obesity status	16S-DNA seq of gut microbiota	0.88	regularized logistical regression	101 adult female twins	US	2010
obesity status	functional genes of gut microbiota	0.78	unknown	265 adults	Denmark	2013
obesity status	16S-DNA seq of gut microbiota	0.51-0.65 (CV-AUC)	random forest	varying with datasets	varying with datasets	2016
obesity status	16S rDNA seq gut of microbiota	0.60 (CV-AUC)	random forest, SVM etc.	319 adults	US	2017
obesity status	pathway module of gut microbiota	0.70-0.80 (CV-AUC)	random forest, SVM etc.	136 adults	US	2017
obesity status	16S rDNA seq gut of microbiota	0.70	random forest	212 newborns	Finland	2020

Take-home message

- **Models containing microbial features can predict obesity trajectories for preschoolers with good performance (AUC 0.82).**
- **Machine learning models have identified robust fecal biomarkers for the obesity trajectory.**
- **Microbiome-based ML models we built may be interpretable.**