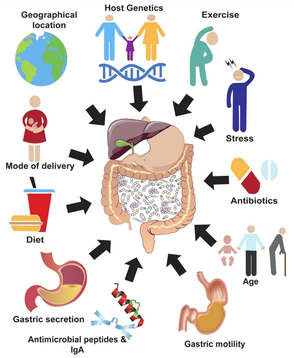
# Explainable Machine Learning to Identify CLL Based on Gut Microbiome Data

Tereza Faitová [1], Ramtin Zargari Marandi[2], Carsten Utoft Niemann[1,3]

PDF poster | LinkedIn

## Introduction




**Gut microbiome**
- is an ecosystem in our guts formed by microbial species
- controls the digestion of food, immune system and CNS

**Gut microbiome composition**
- is mostly comprised of bacteria
- is highly dynamic
- is affected by many factors (see picture)
- rich and diverse microbiome is favorable & health promoting

**Gut microbiome in CLL**
- is less diverse than in healthy individuals
- can affect disease development (shown in TCL1 mice)
- is depleted of short-chain fatty acid producing bacteria

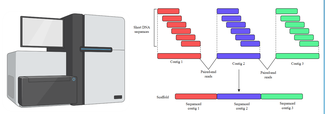
Hypothesis: Gut microbiome of CLL patients is different from microbiomes of other patients

## Materials and Methods

1) Patient and healthy cohorts

2) Feces samples collection

3) Shotgun metagenomic sequencing



4) Cohorts overview, Machine learning, Data analysis

87 CLL | 59 healthy | 165 baseline HSCT | 9 kidney donors | 91 thorax surgery patients

674 gut microbial species abundance, Age & Sex

**Classification model**

Development set
69 (21%) CLL & 259 (79%) no-CLL

Validation set
18 (22%) CLL & 65 (78%) no-CLL

5-fold cross validation & model development

Validation, best model selection & interpretation

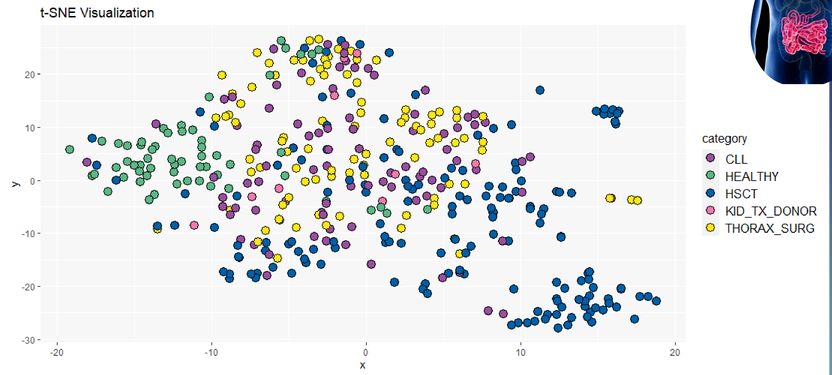
**Statistical analyses**

Alpha & beta diversities

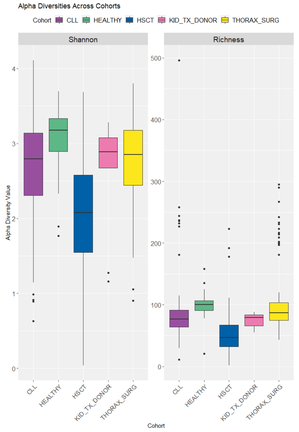
Dimensional reduction

Relative abundance analyses

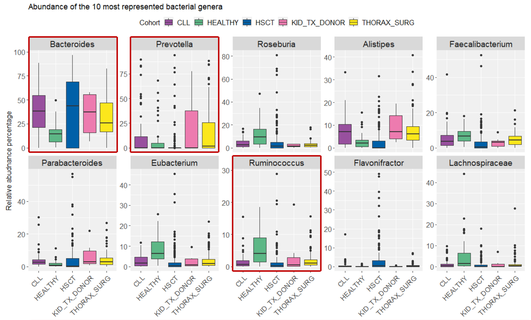
## Results




**t-SNE Visualization**
High dimensional metagenomic data in 2D - t-SNE models pairwise similarities between samples and maps them to 2D space.
Each point = one sample. Color code = cohort. Samples with similar microbial community compositions are closer together.
Underlying patterns revealed:
- no clear clusters, but similarities **within HSCT** and HEALTHY cohorts
- CLL, THORAX_SURG, and KIDNEY DONORS seem to have heterogeneous microbiome compositions

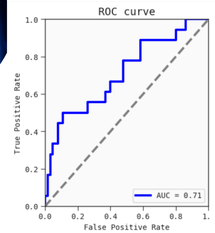


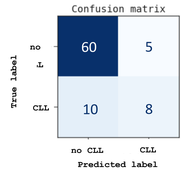

**Richness**:
number of different species present in the community
**Shannon Index**:
considers both species richness and evenness

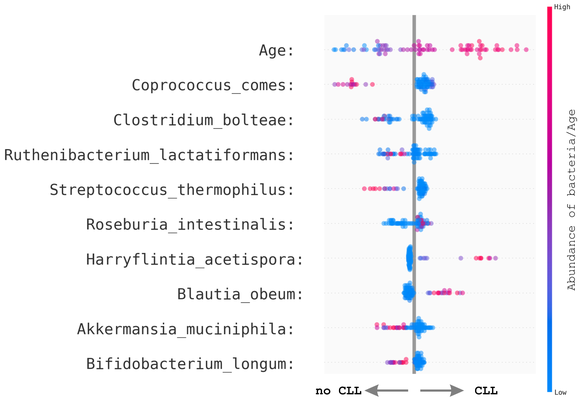



**Alpha diversity** (left)
Shannon diversity: HSCT < CLL < THORAX_SURG < KIDNEY DONORS < HEALTHY
HSCT and HEALTHY cohorts show lowest and highest diversity across metrics, r.
No significant difference in alpha diversity among CLL, THORAX_SURG and KIDNEY DONORS

**Relative abundance of the most abundant genera** (right)
HSCT and CLL are most enriched in *Bacteroides* (enterotype 1), often associated with disease states
HSCT is also significantly enriched in *Flavonifractor*, yet unexplored bacterial genus
HEALTHY are most enriched in *Ruminococcus* (enterotype 2), often described as beneficial
are also enriched in beneficial *Faecalibacterium*, *Lachnospiraceae* and *Eubacterium*
No cohorts were significantly enriched in *Prevotella* (enterotype 3)




Machine learning outcome (LightGBM model):
**ROC curve:** used to evaluate the performance of a binary classification model
**Confusion matrix:** represents how well the model predictions match the actual outcomes
**SHAP plot:**
- illustrates the top contributing features to the identification of CLL - bacterial species and age in our case
- red = higher value, blue = lower value
- interpretation: The higher the age, the higher predicted probability of CLL – positive correlation
  The higher abundance of *Coprococcus comes*, the lower predicted probability of CLL - negative correlation

**Model evaluation:**
- AUC = 0.71; specificity = 0.92, and sensitivity = 0.44 on the validation set
- High specificity = model can classify no-CLL with very high precision (5/65 misclassifications)
- Low sensitivity = model is performing poorly in CLL classification (10/18 misclassifications)

## Conclusion

- CLL microbiomes are different from HEALTHY and HSCT
- CLL microbiomes are less diverse than HEALTHY but more diverse than HSCT.
- CLL and HSCT microbiomes are enriched in *Bacteroides*.
- CLL and HSCT microbiomes are depleted of several beneficial bacteria.
- Our machine learning model is better in classifying **no-CLL** than CLL.

## Acknowledgements

I would like to sincerely thank all patients for providing samples for the study, CHIP Centre of Excellence for providing infrastructure, exceptional expertise, and financial support.
The staff at the Hematology Department and PERSIMUNE biobank by Rigshospitalet were essential for this study as they organized and collected the samples for our research.

## Affiliations

1) Department of Hematology, Rigshospitalet, Copenhagen, DK
2) Centre of Excellence for Health, Immunity and Infections (CHIP), Rigshospitalet, Copenhagen, DK
3) Department of Clinical Medicine, University of Copenhagen, DK

REGION H | Rigshospitalet | chip Centre of Excellence for Health, Immunity and Infections | UNIVERSITY OF COPENHAGEN