Latent Unknown Clustering with Integrated Multi-Omics Data (LUCID)

Yinqi Zhao, Nikolaos Stratakis, David V. Conti

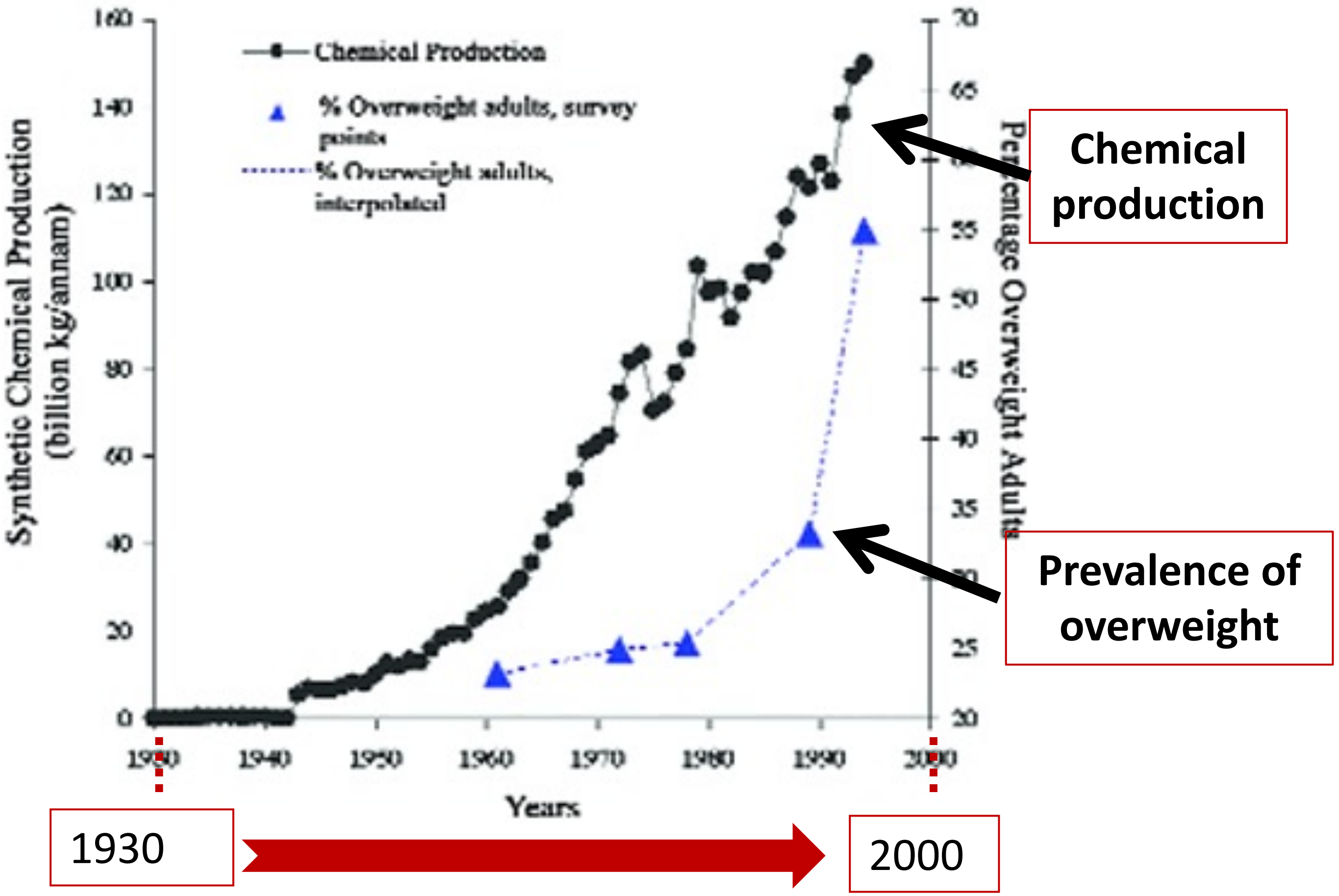
Exposome Data Challenge Event

28th April 2021

Research Question

Does early-life organochlorine exposure increase obesity risk, and if yes, what are the underlying mechanistic pathways?

Environmental obesogen hypothesis



Chemical production

Prevalence of overweight

1930

2000

Organochlorines (OCs)

A major class of suspected obesogens

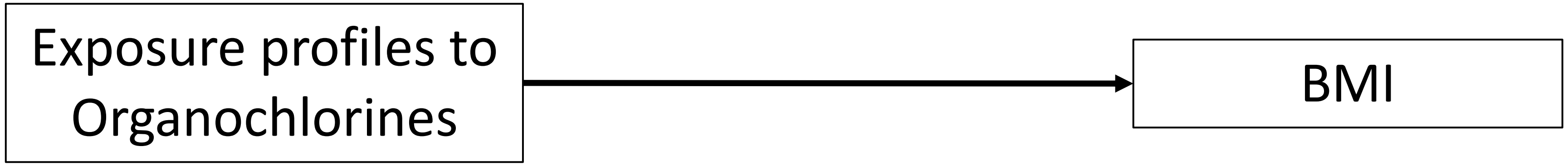
Compounds widely used in the past as pesticides or in industrial processes

Pregnancy and childhood are critical periods of increased susceptibility to chemical effects

Baillie-Hamilton. J Altern Complement Med. 2002

Guided Hypothesis

Conventional study



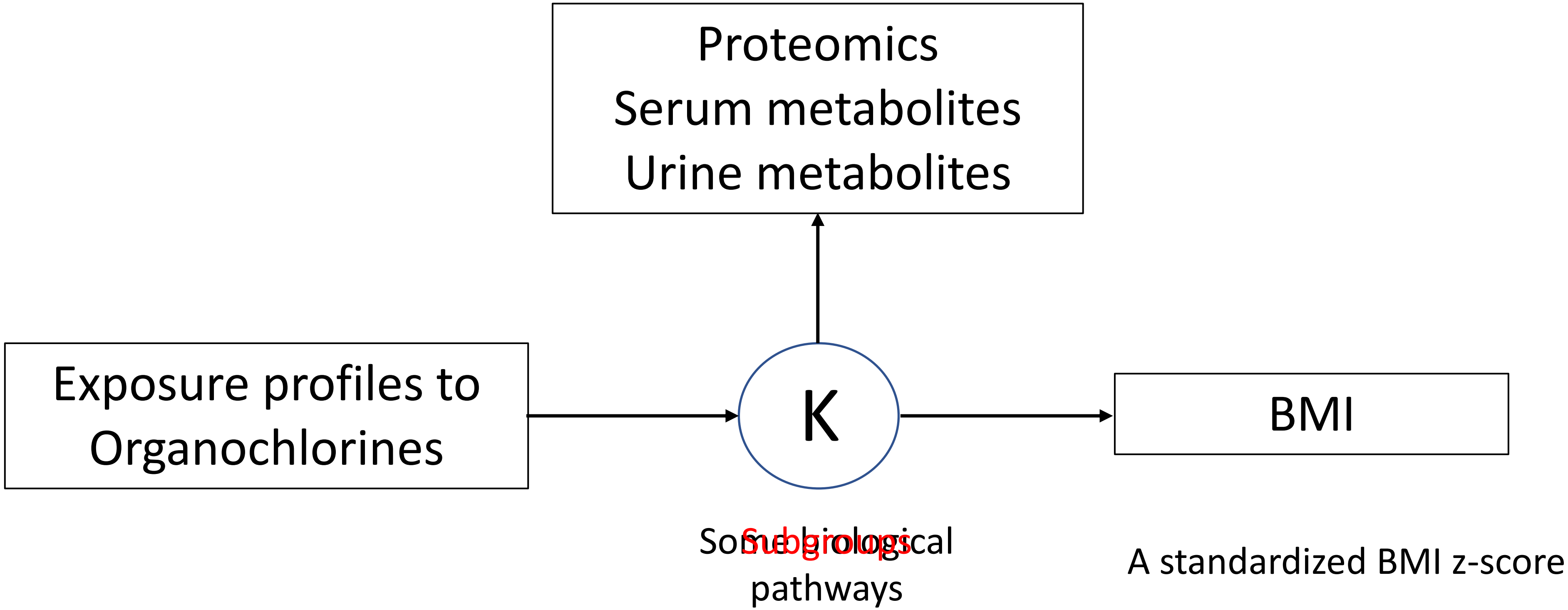
Exposure profiles to Organochlorines

BMI

A standardized BMI z-score

Unanswered question: What is the underlying biological process between E and Y?

Guided Hypothesis



Proteomics
Serum metabolites
Urine metabolites

Exposure profiles to Organochlorines

K

BMI

Some biological pathways

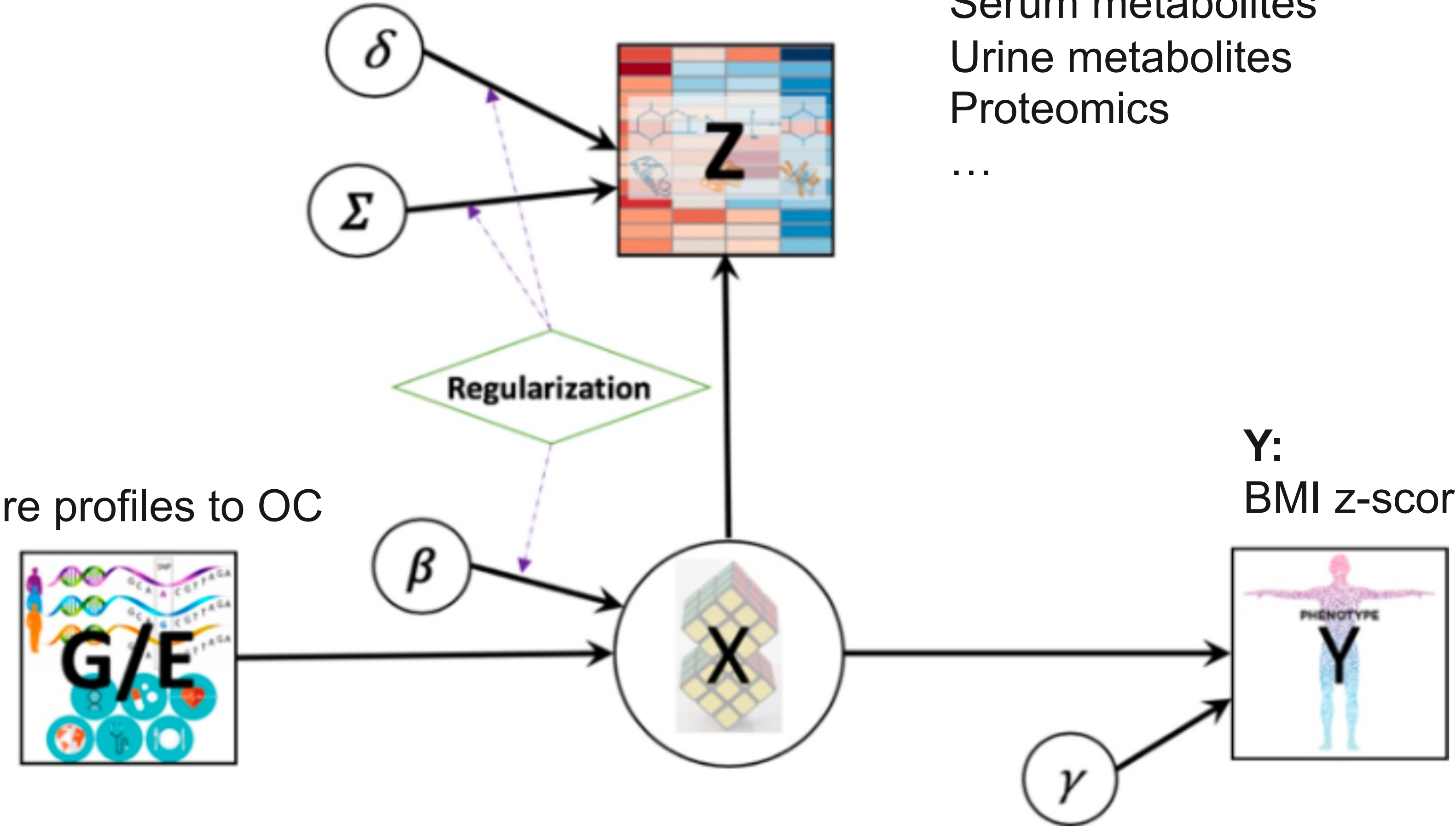
Subgroups

A standardized BMI z-score

Goal: estimate the subgroup characterized by certain exposure and omics profiles

An overview of LUCID model

Prof. David Conti



Z :
Serum metabolites
Urine metabolites
Proteomics
…

E: Exposure profiles to OC

Y:
BMI z-score

Peng et al., 2019

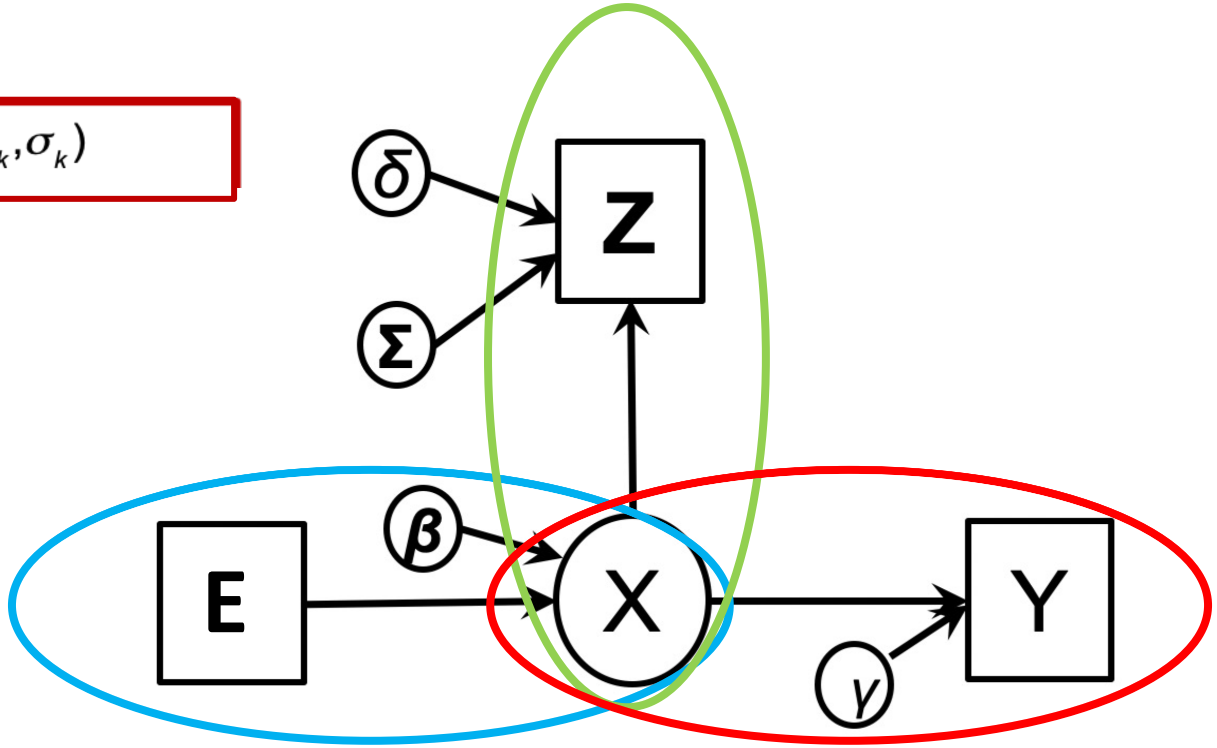
Statistical Modeling

$$f(\mathbf{Z} \mid X = k) = \phi(\mathbf{Z}; \delta_k, \Sigma_k)$$

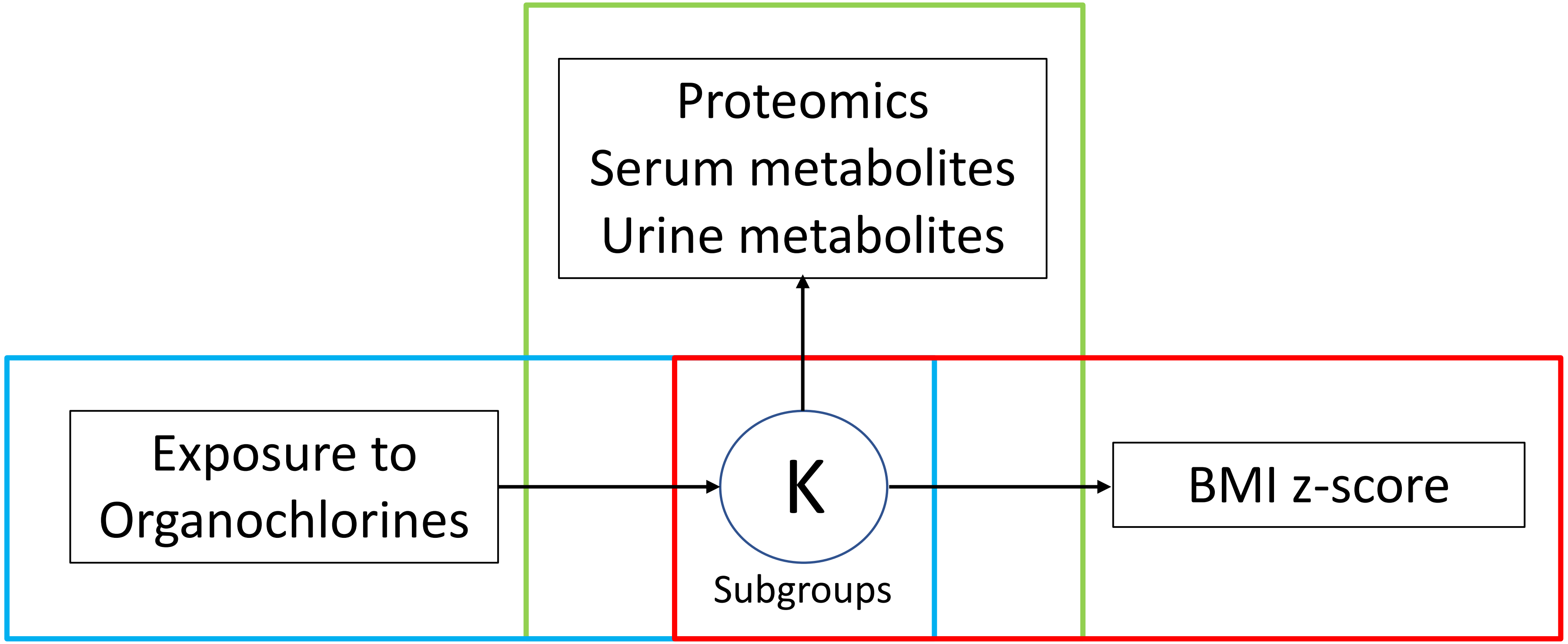
$$f(X = k \mid \mathbf{E}) = S(\beta, k, \mathbf{E}) = \frac{\exp(\beta_k \mathbf{E})}{\sum_{k'} \exp(\beta_{k'} \mathbf{E})}$$

$$f(Y \mid X = k) = \phi(Y; \gamma_k, \sigma_k)$$

$$
\begin{aligned}
&Q(\boldsymbol{\Theta}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} P(X_i = j \mid \boldsymbol{\mathcal{D}}; \boldsymbol{\Theta}) \log \frac{S(\beta, j, \boldsymbol{E}_{\mathrm{i}})}{P(X_i = j \mid \boldsymbol{\mathcal{D}}; \boldsymbol{\Theta})} \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{k} P(X_i = j \mid \boldsymbol{\mathcal{D}}; \boldsymbol{\Theta}) \log \frac{\phi\left(\mathbf{Z}_i; \delta_j, \Sigma_j\right)}{P(X_i = j \mid \boldsymbol{\mathcal{D}}; \boldsymbol{\Theta})} \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{k} P(X_i = j \mid \boldsymbol{\mathcal{D}}; \boldsymbol{\Theta}) \log \frac{f\left(Y_i; \gamma_j, \sigma_j\right)}{P(X_i = j \mid \boldsymbol{\mathcal{D}}; \boldsymbol{\Theta})}
\end{aligned}
$$



Questions Addressed in Data Challenge



Proteomics
Serum metabolites
Urine metabolites

Exposure to Organochlorines

K

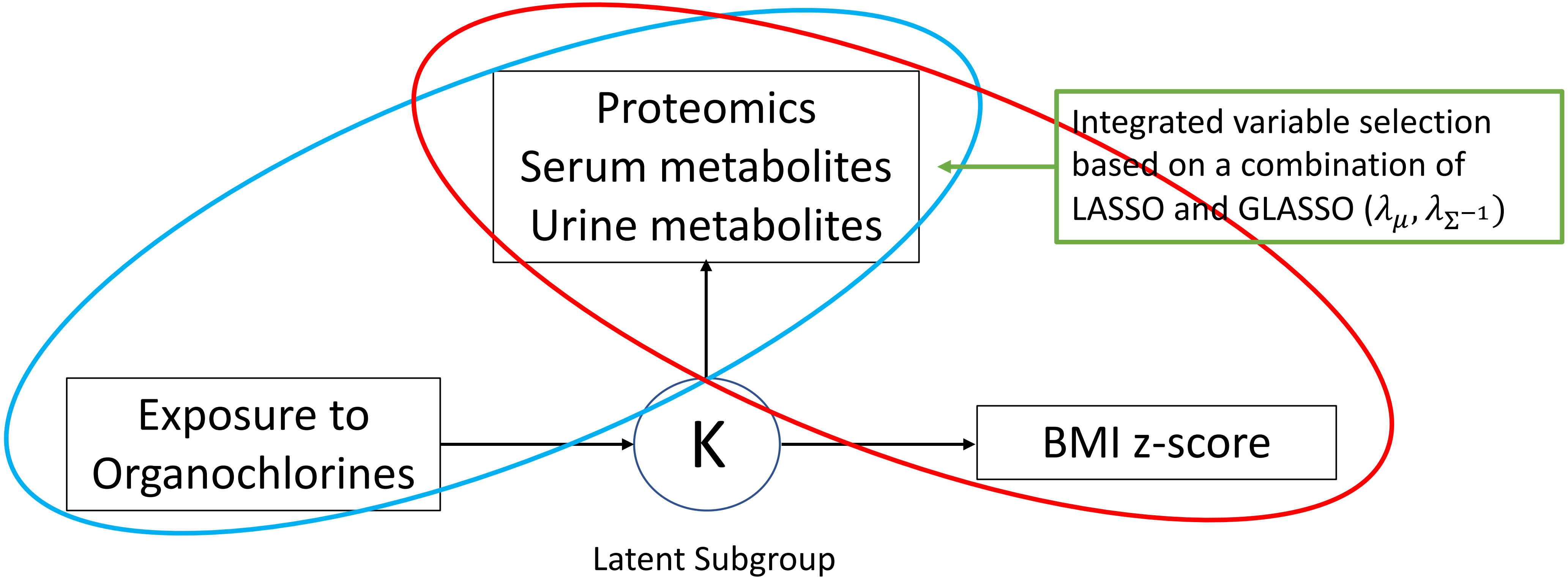
Subgroups

BMI z-score

Goal:

1. Identify the subgroups with high BMI z-score (obesity)
2. Characterize the subgroups by distinguished omics profiles
3. Estimate the association between exposure and subgroups

Variable Selection for Metabolites



Proteomics
Serum metabolites
Urine metabolites

Integrated variable selection based on a combination of LASSO and GLASSO $(\lambda_{\mu}, \lambda_{\Sigma^{-1}})$

Exposure to Organochlorines

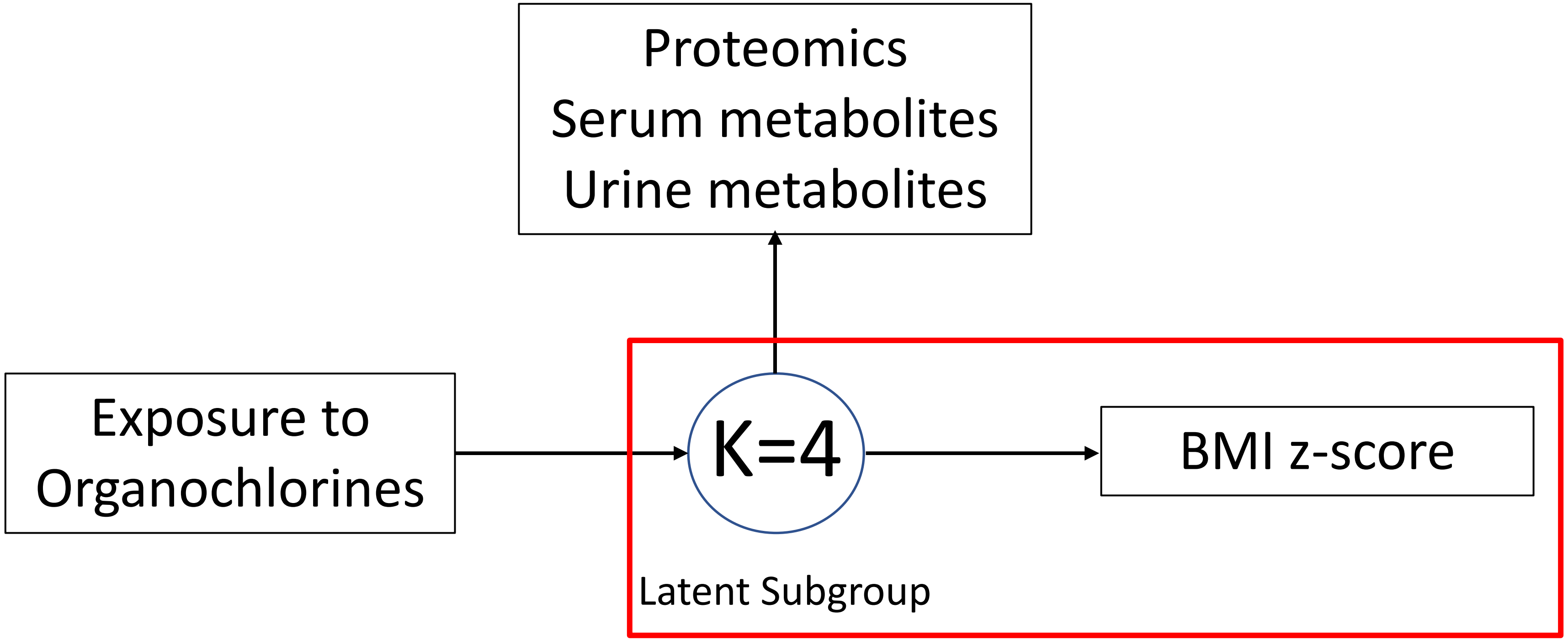
K

BMI z-score

Latent Subgroup

1. Preliminary screening (45 omics signatures identified)
2. Integrated variable selection(14 omics signatures)

Results



Proteomics
Serum metabolites
Urine metabolites

Exposure to Organochlorines

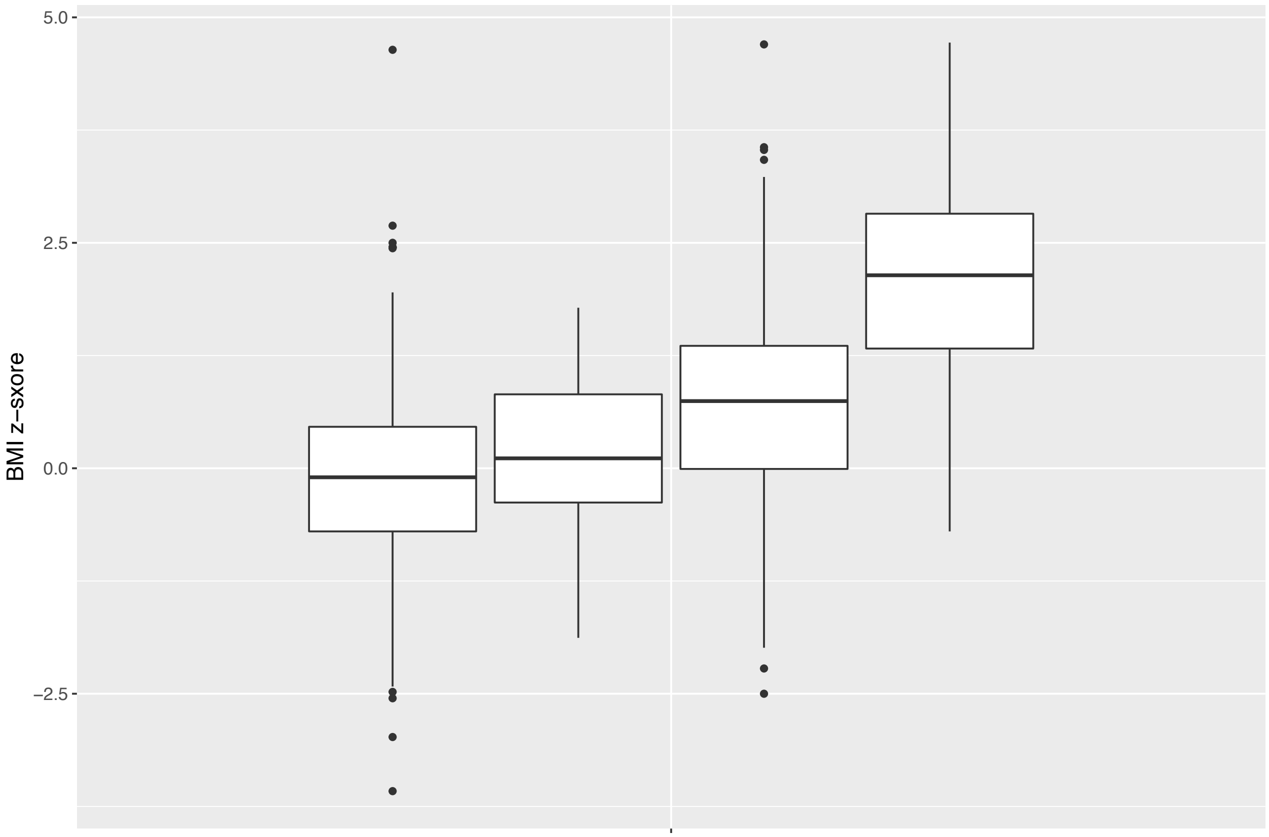
K=4

Latent Subgroup

BMI z-score

1. Association between BMI z-score and each subgroup

Distribution of BMI z-score for each subgroup



Subgroup 1 (593)

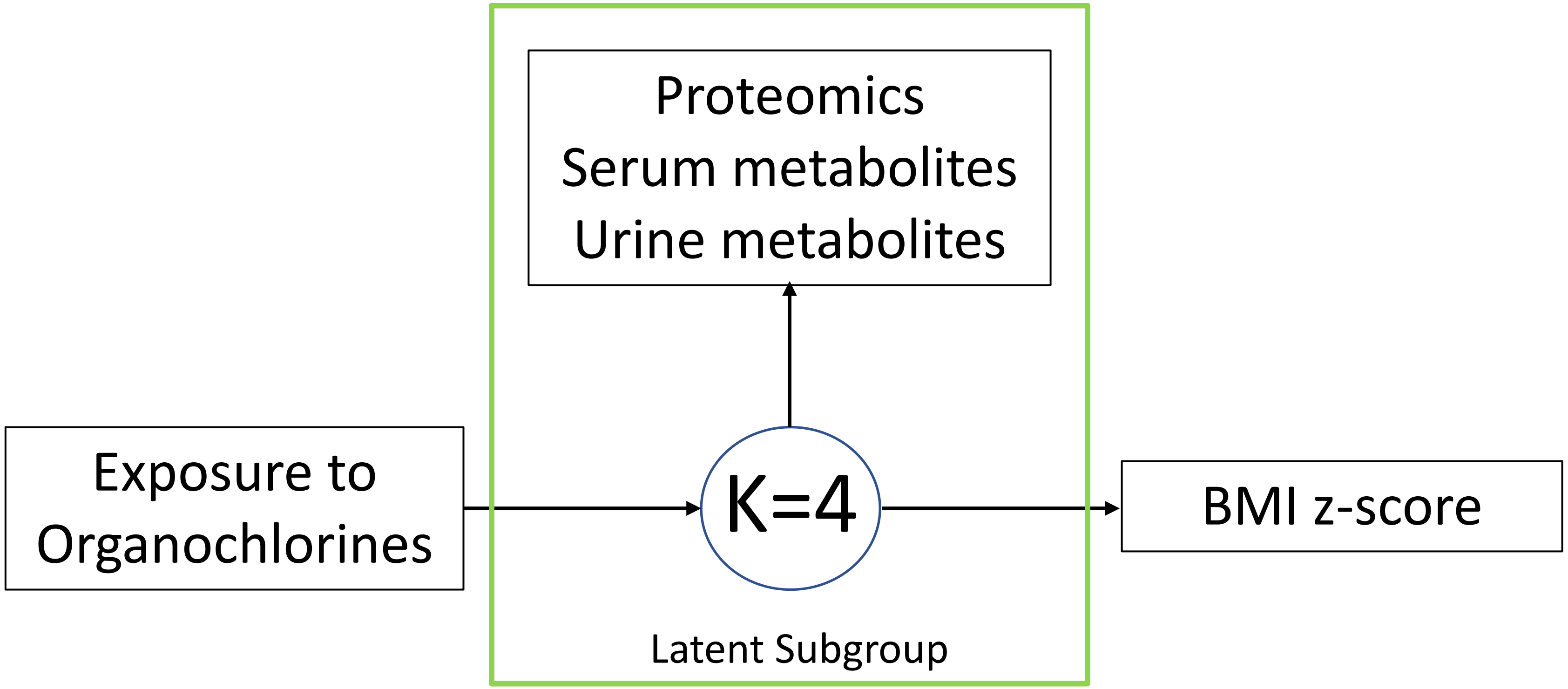
Subgroup 2 (37)

Subgroup 3 (386)

Subgroup 4 (136)

Observations

Results



Exposure to Organochlorines

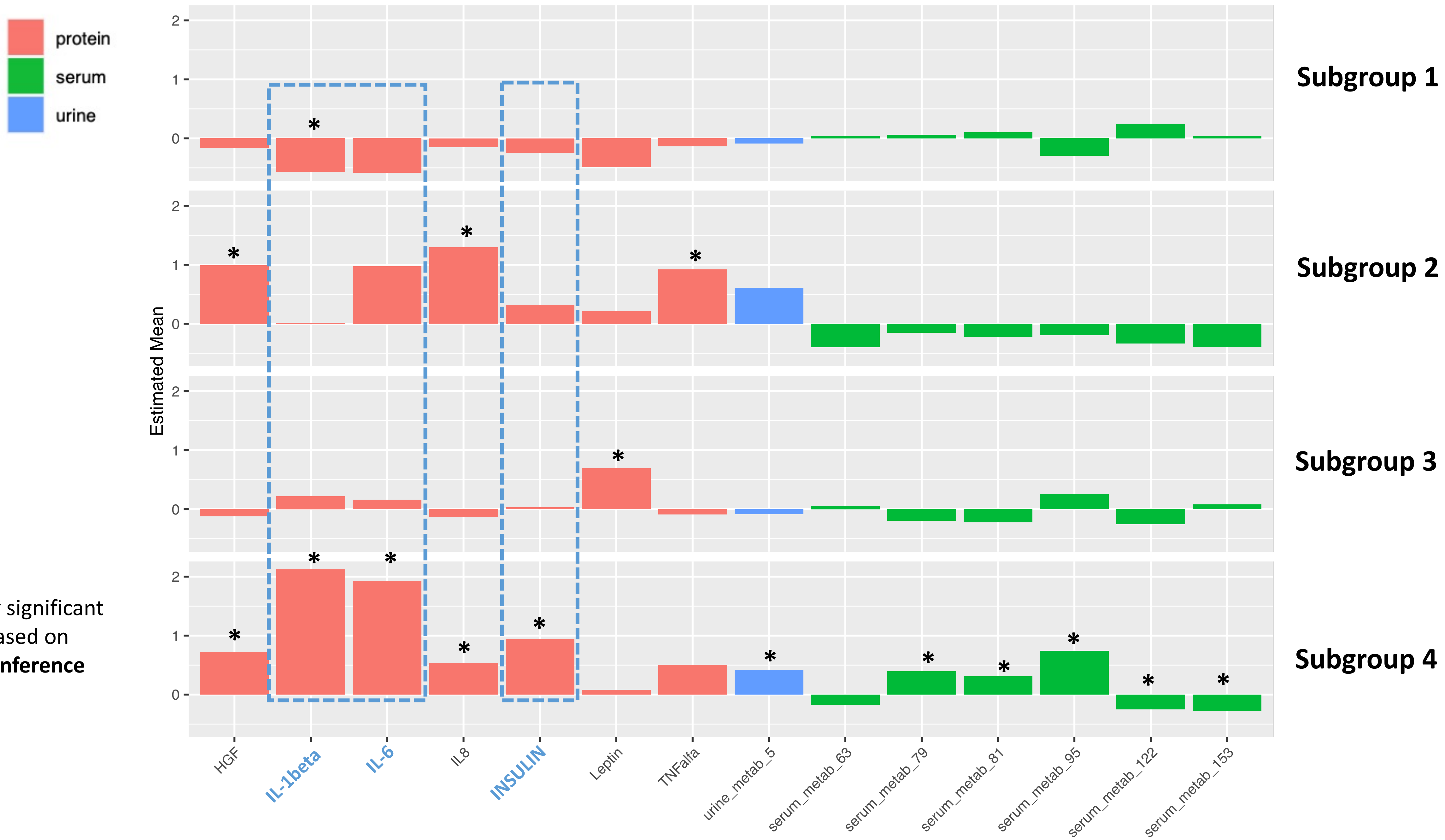
Proteomics
Serum metabolites
Urine metabolites

K=4

Latent Subgroup

BMI z-score

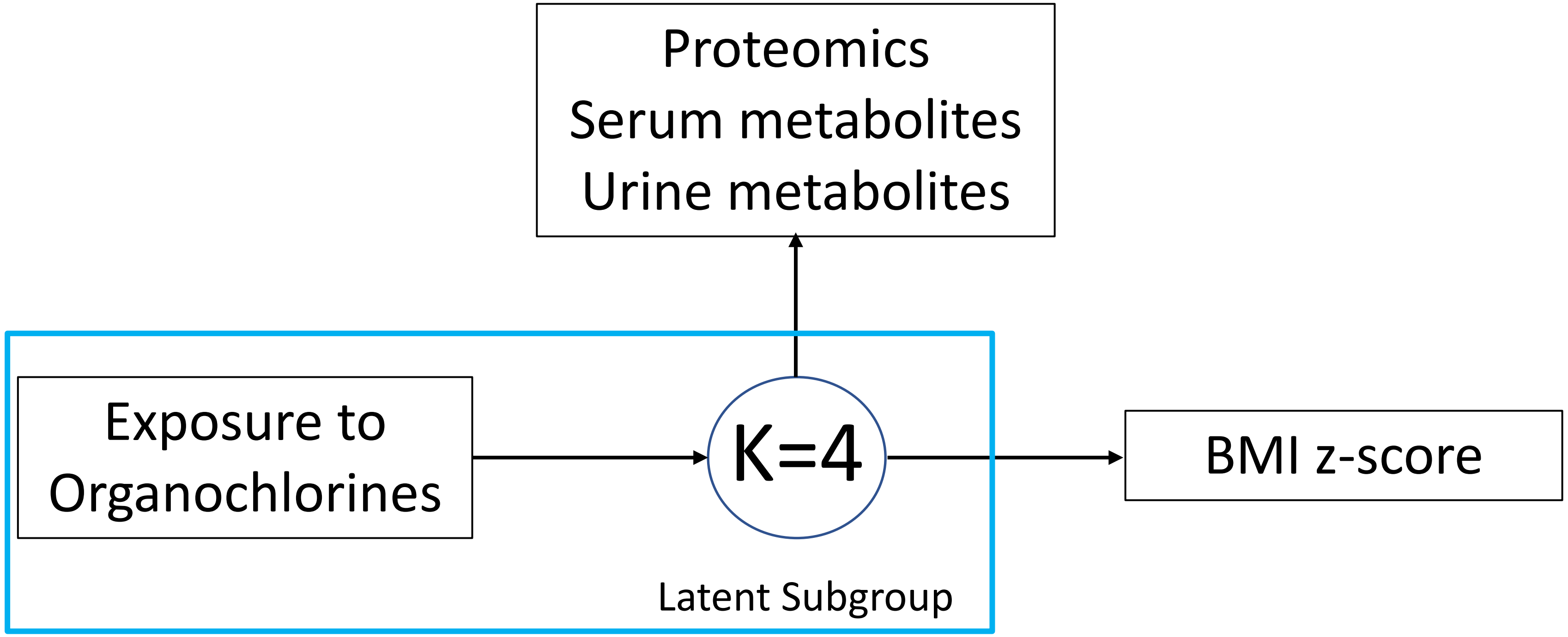
2. Omics profiles for each subgroup



*

Statistically significant estimate based on **bootstrap inference**

Results



Proteomics
Serum metabolites
Urine metabolites

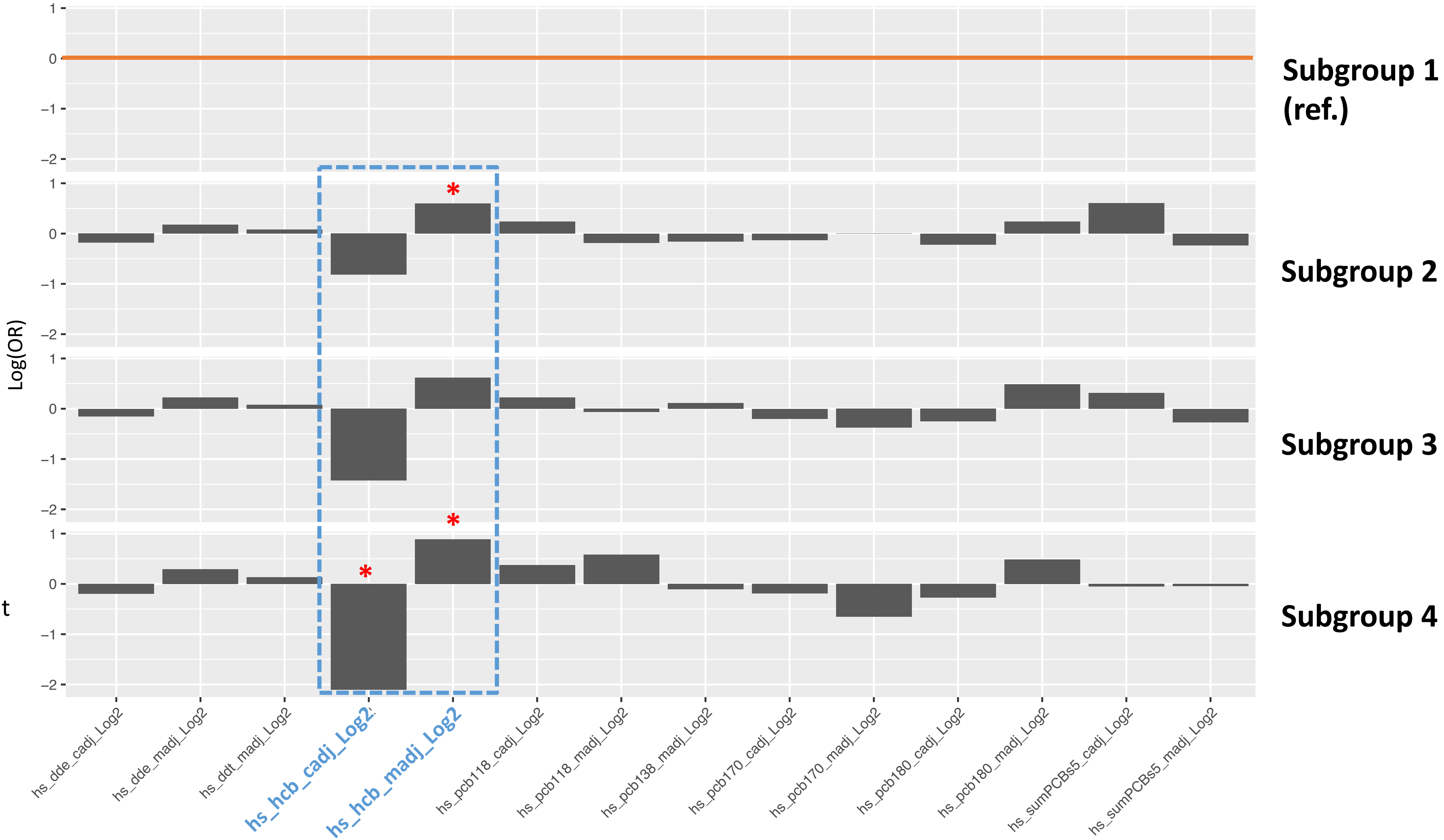
Exposure to Organochlorines

K=4

Latent Subgroup

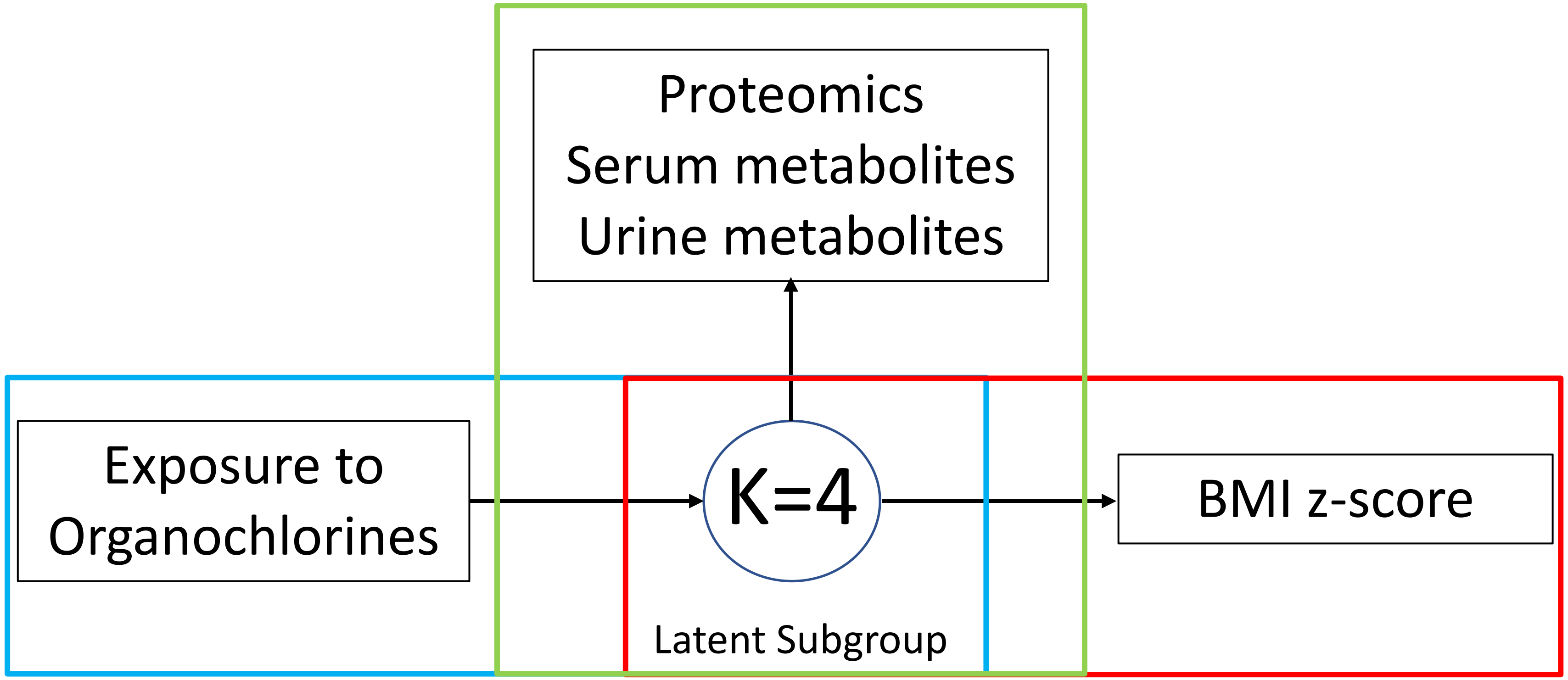
BMI z-score

3. Exposure profiles for each cluster



*
Statistically significant estimate based on **bootstrap inference**

Results



Proteomics
Serum metabolites
Urine metabolites

Exposure to Organochlorines

K=4

Latent Subgroup

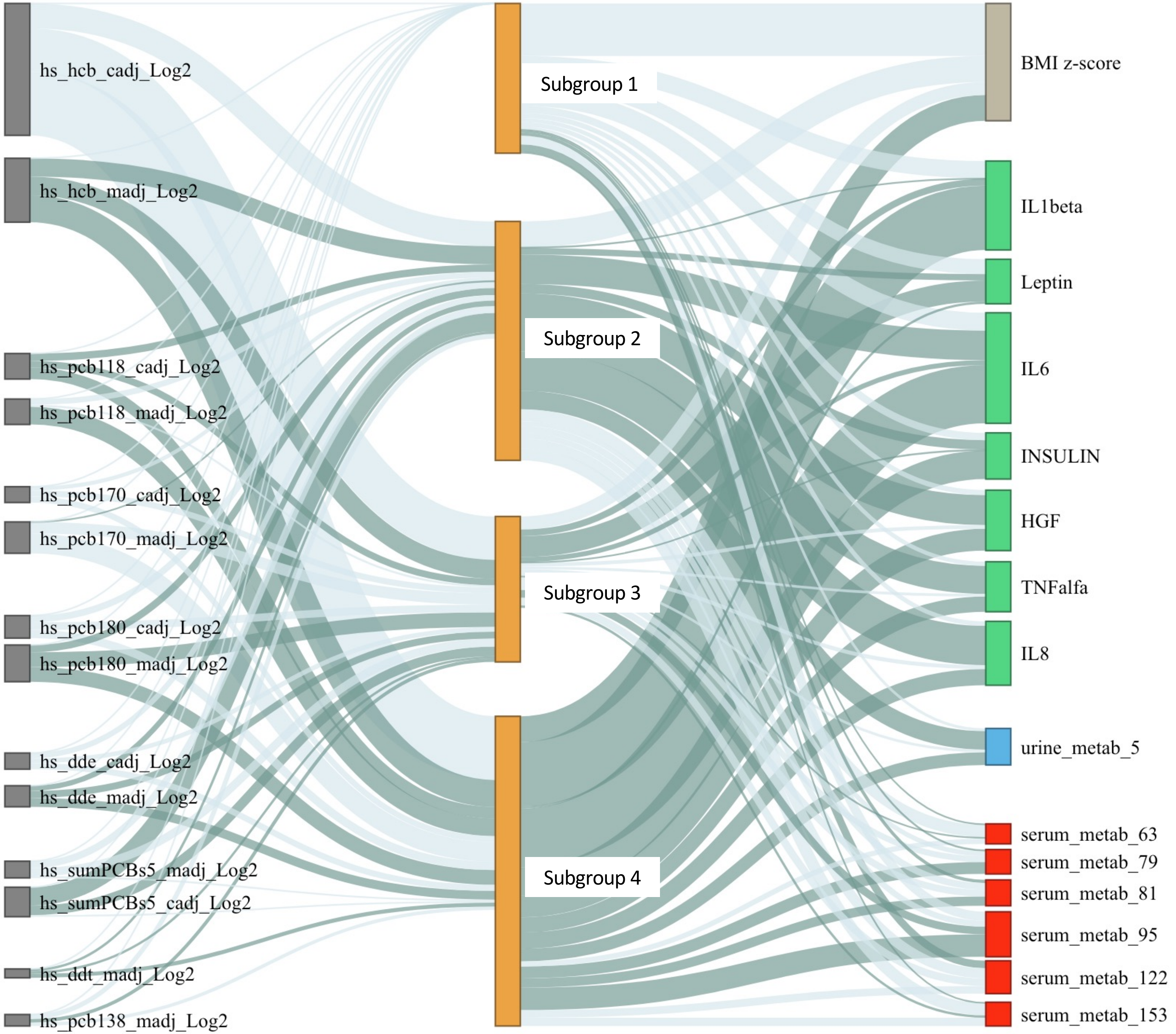
BMI z-score

If we combine all these information together?

4. Sankey Diagram

Illustrate the association between:

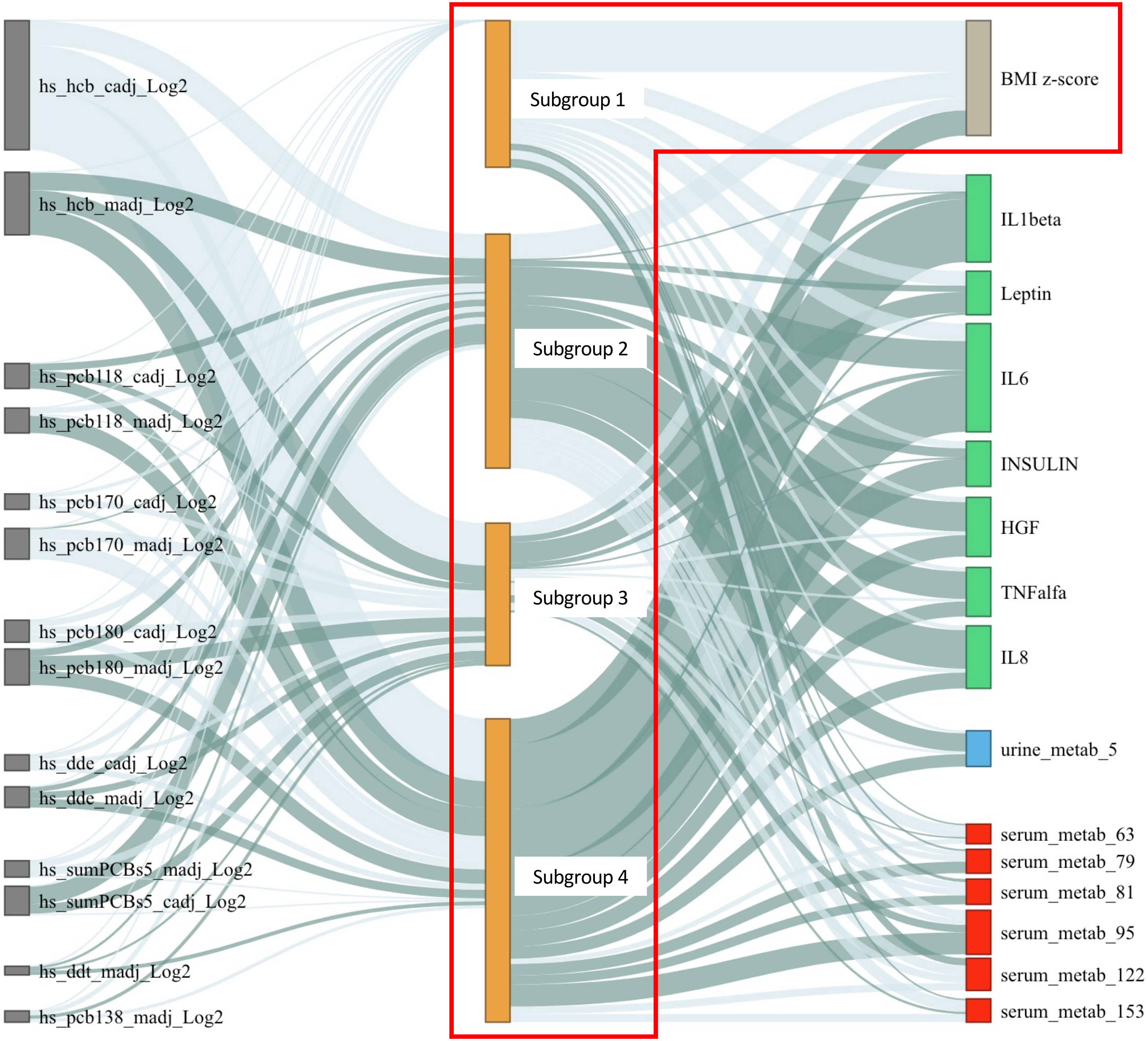
1. Latent subgroup to BMI
2. Latent subgroup to metabolites
3. Exposure to latent subgroup



4. Sankey Diagram

Illustrate the association between:

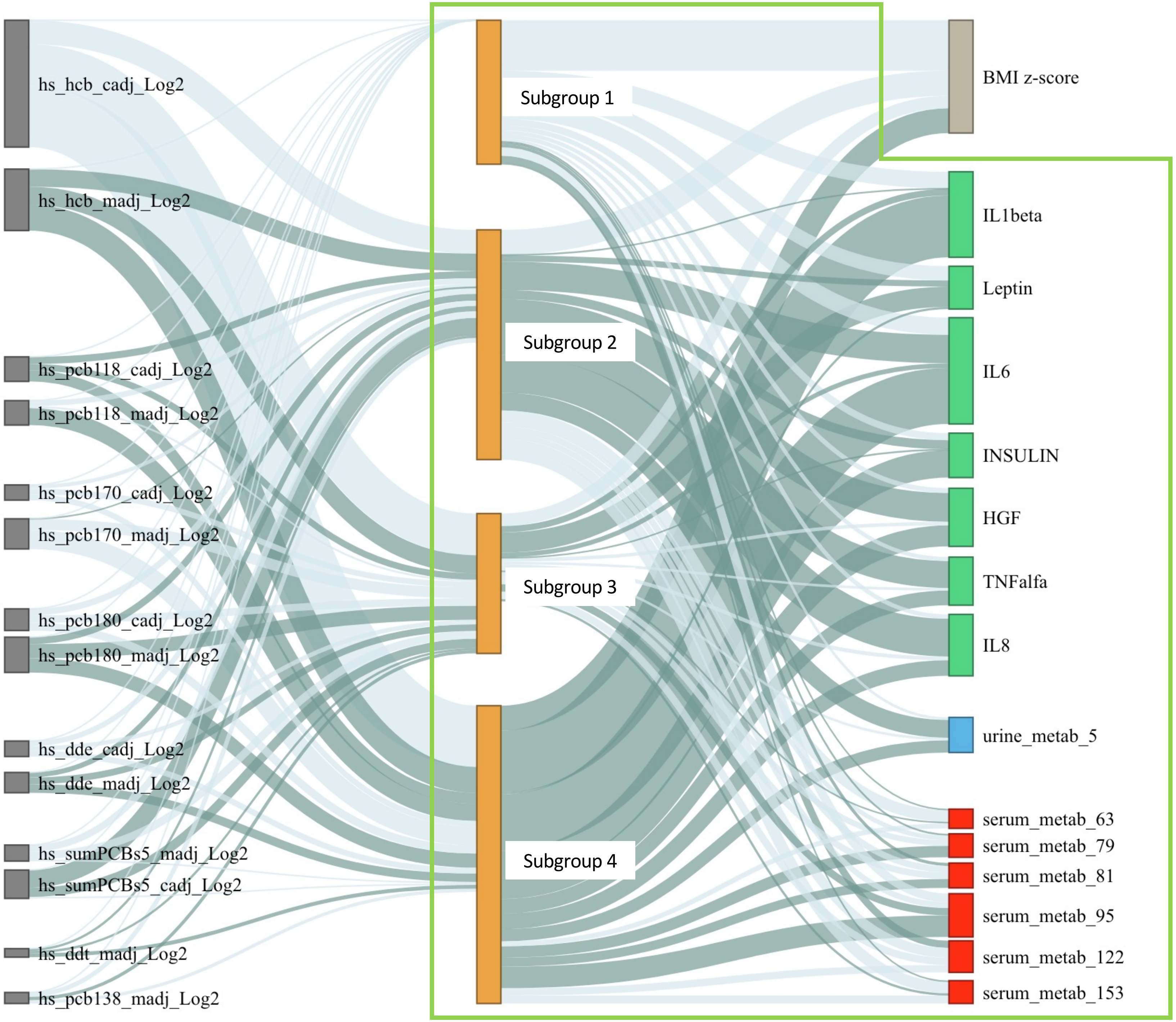
1. Latent subgroup to BMI
2. Latent subgroup to metabolites
3. Exposure to latent subgroup



4. Sankey Diagram

Illustrate the association between:

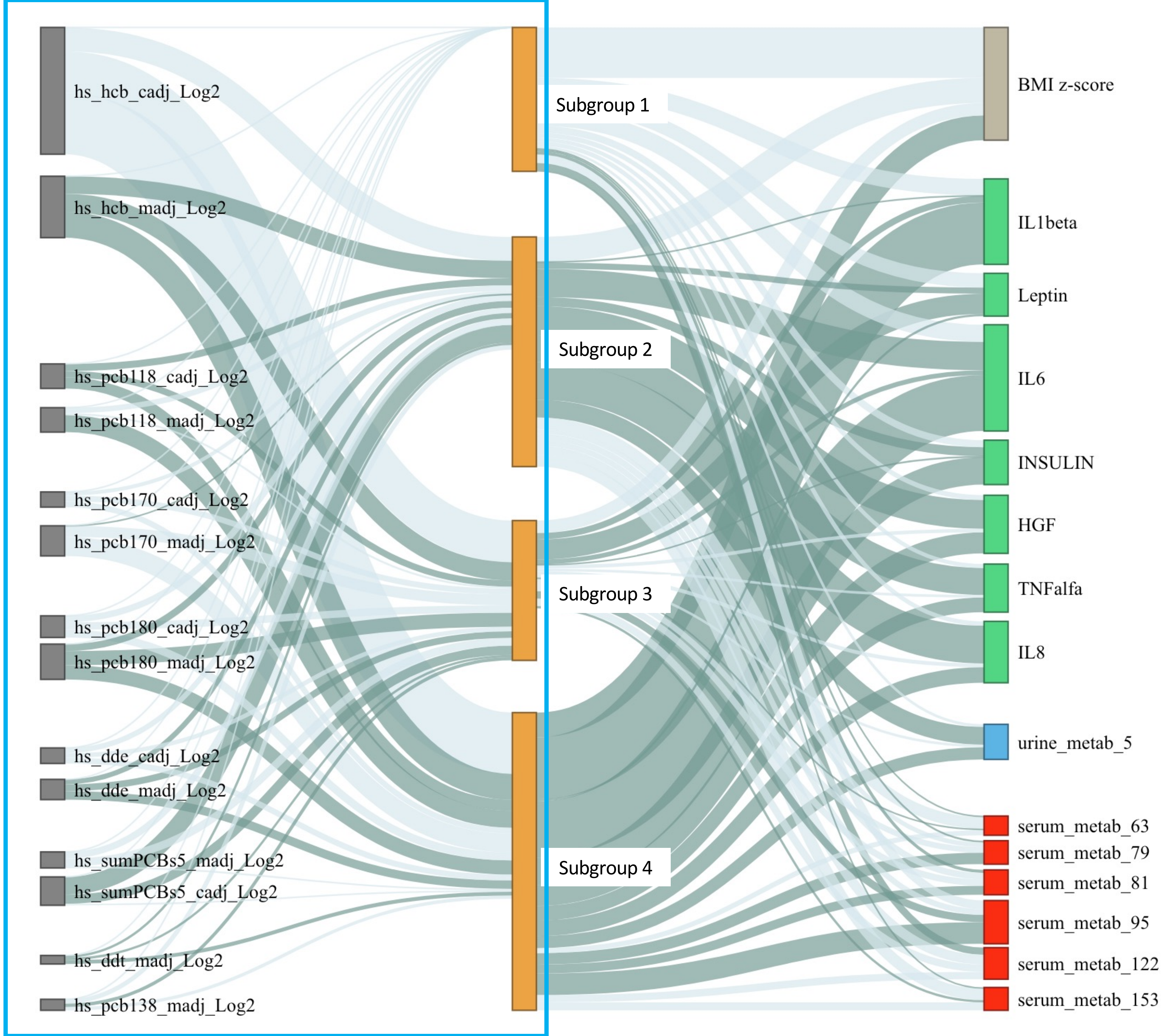
1. Latent subgroup to BMI
2. Latent subgroup to metabolites
3. Exposure to latent subgroup



4. Sankey Diagram

Illustrate the association between:

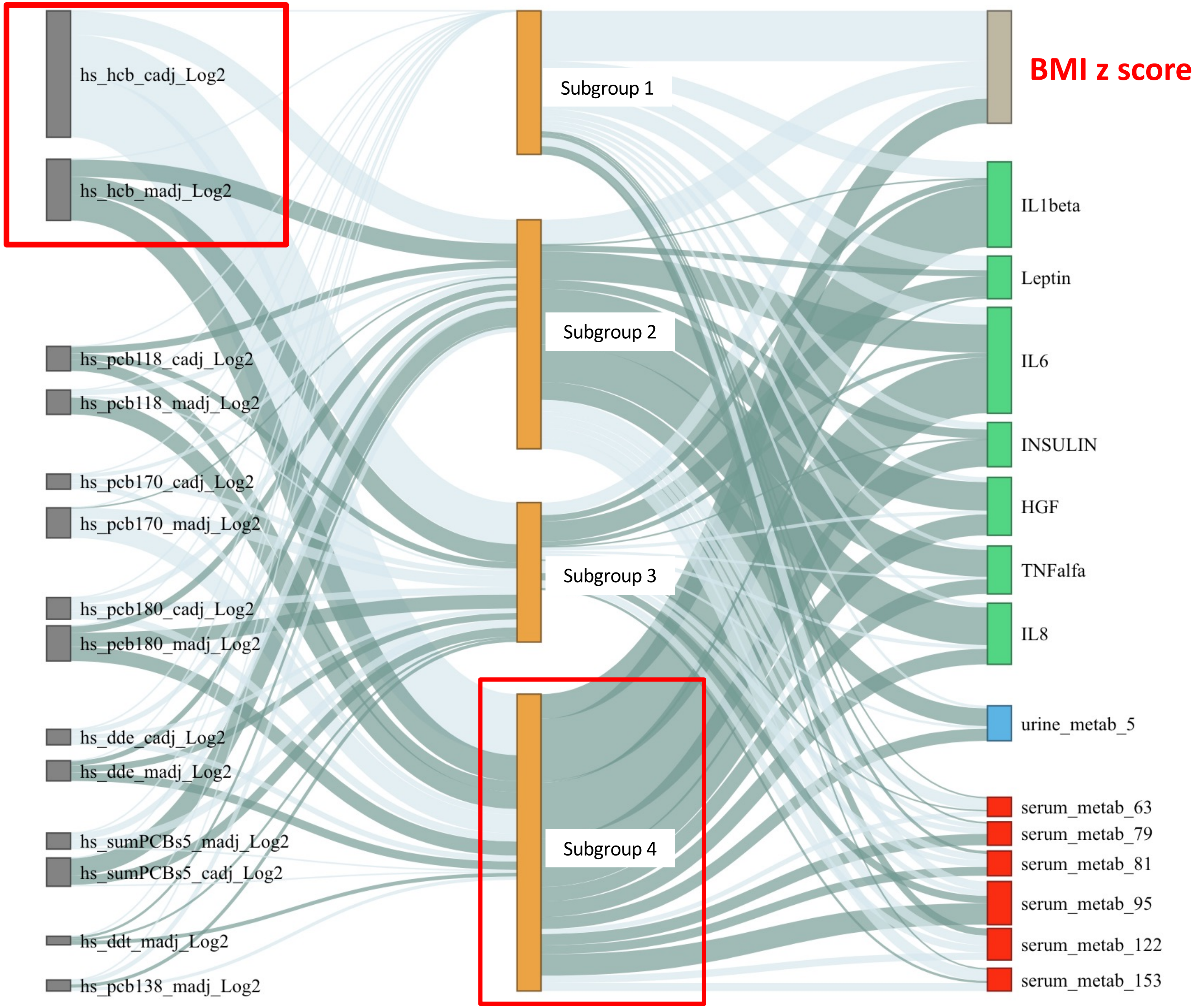
1. Latent subgroup to BMI
2. Latent subgroup to metabolites
3. Exposure to latent subgroup



4. Sankey Diagram

Illustrate the association between:

1. Latent subgroup to BMI
2. Latent subgroup to metabolites
3. Exposure to latent subgroup



Conclusions

- LUCID is a novel method for integrated multi-omics analysis

- We implemented LUCID to discover 4 latent subgroups characterized by distinguished OC exposure and omic profiles and their association with obesity

- We confirmed previous HELIX publication results (Vrijheid EHP 2020) that childhood HCB exposure cross-sectionally associates with reduced childhood BMI-z score

- We identified a new association confirming previous literature that prenatal HCB exposure increases BMI (obesogenic effect)

- Using multi-omic profiling, we identified signatures (IL-1beta, IL-6, insulin) giving insight into underlying mechanistic pathways (eg., systemic inflammation, disturbed glucose metabolism)

Acknowledgments:

Special thanks to Professor David Conti, Dr. Nikos Stratakis for their invaluable suggestions to this project!

References:

1. LUCID method paper:

Peng, C., Wang, J., Asante, I., Louie, S., Jin, R., Chatzi, L., Casey, G., Thomas, D.C., and Conti, D.V. (2019). A Latent Unknown Clustering Integrating Multi-Omics Data (LUCID) with Phenotypic Traits. Bioinformatics.

2. LUCID R package:

Currently available on CRAN (LUCIDus)

An updated version is on Github: USCbiostats/LUCIDus: the new version of LUCID (github.com)

Analysis Pipeline

Step 1: Data Preparation

Exposure: Organochlorines (18)*

Metabolite: Serum metabolites (177)
Urine metabolites (44)
Proteomics (36)

Outcome: BMI z-score

Covariates: Sex, age, Mother's BMI, Mother's education status, cohort

Sample size: 1152

Step 2: Model Building

Preliminary screening on metabolites

↓

Model selection

$K, \lambda_E,, \lambda_\mu, \lambda_{\Sigma^{-1}}$

↓

Fit the model based on selected exposures and metabolites.

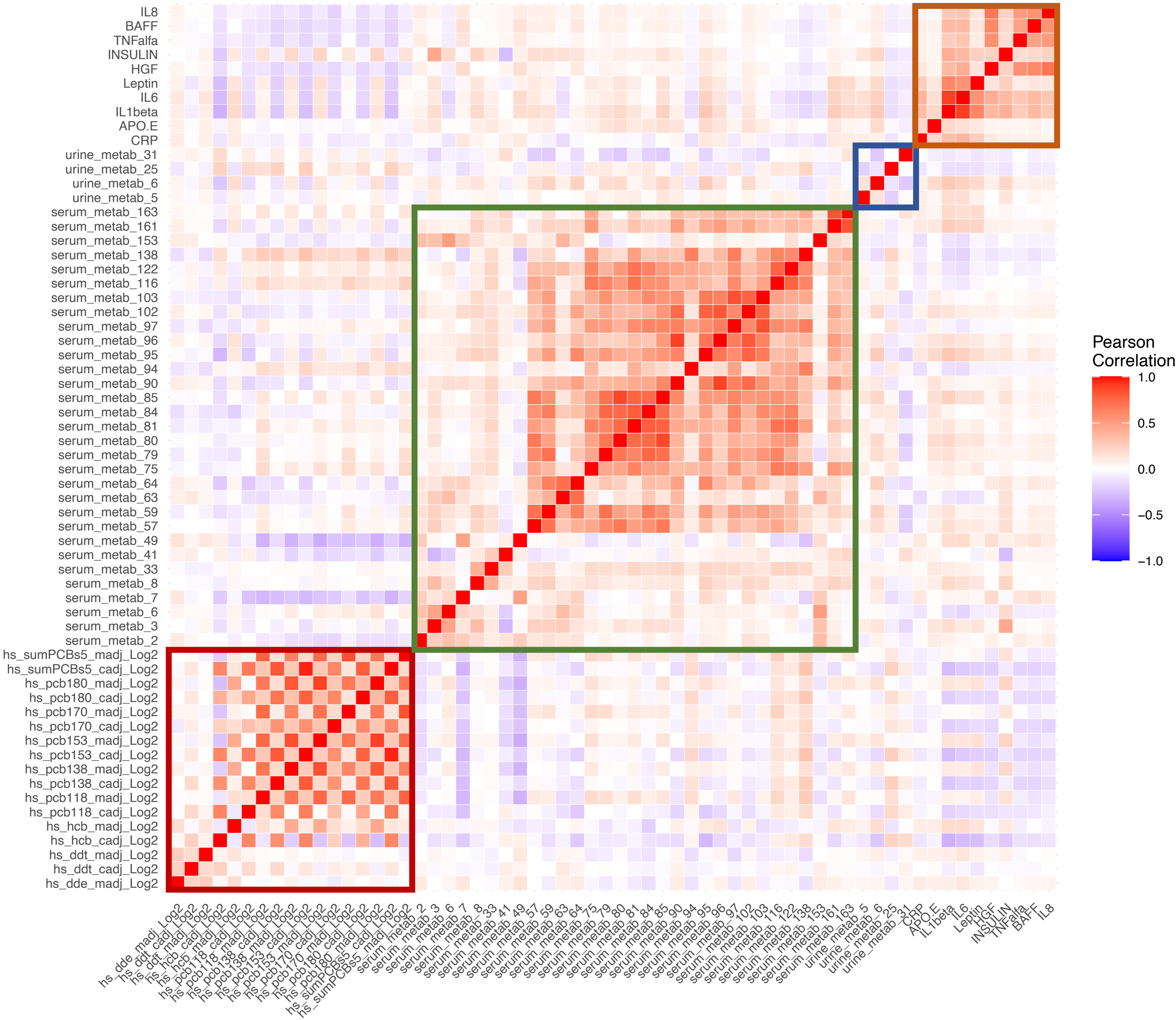
Step 3: Model Interpretation

1. Omics profiles for each subgroup
2. Distribution of BMI for each subgroup
3. Exposure profiles for each cluster
4. Sankey Diagram for association

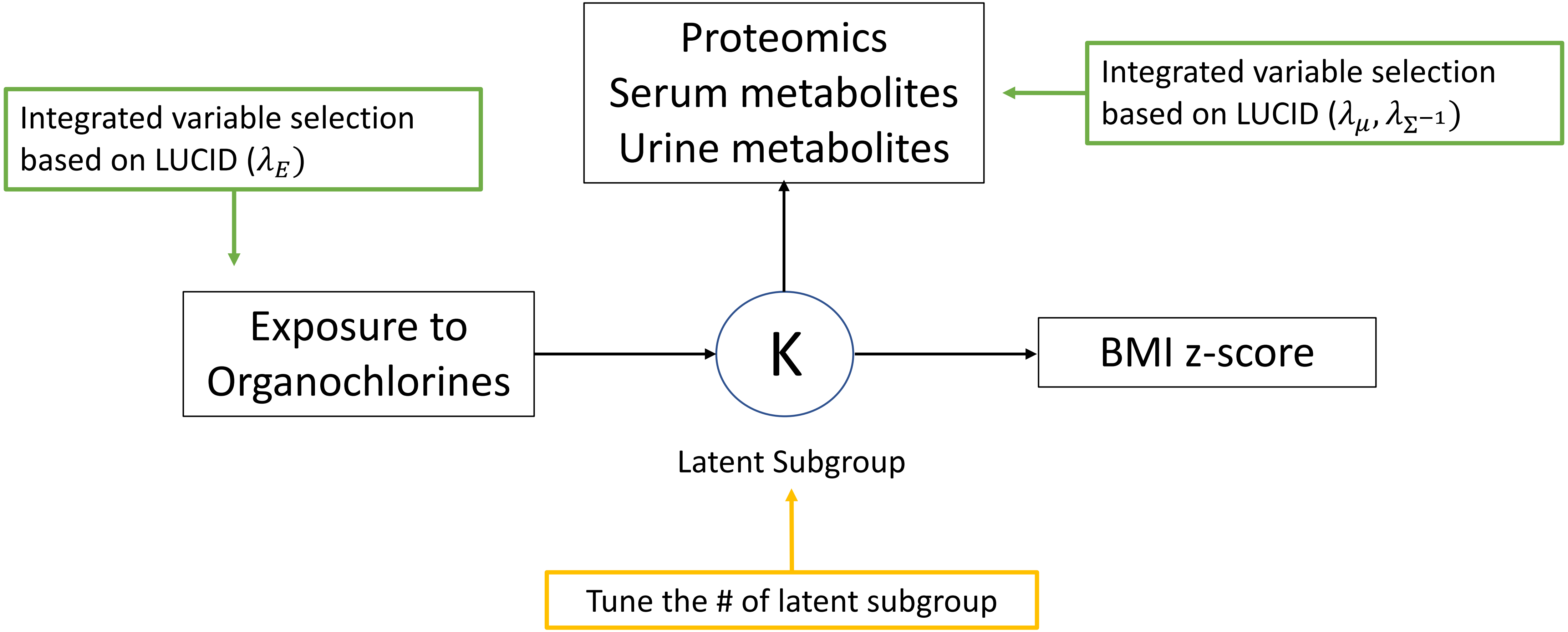
* Concentrations in pregnancy (maternal) and childhood

Preliminary Screening

1. RED: exposure and outcome
2. GREEN: serum metabolomics
3. BLUE: urine metabolomics
4. ORANGE: proteomics



Fitting LUCID Model



Integrated variable selection based on LUCID (λ_E)

Exposure to Organochlorines

Proteomics
Serum metabolites
Urine metabolites

Integrated variable selection based on LUCID ($\lambda_\mu, \lambda_{\Sigma^{-1}}$)

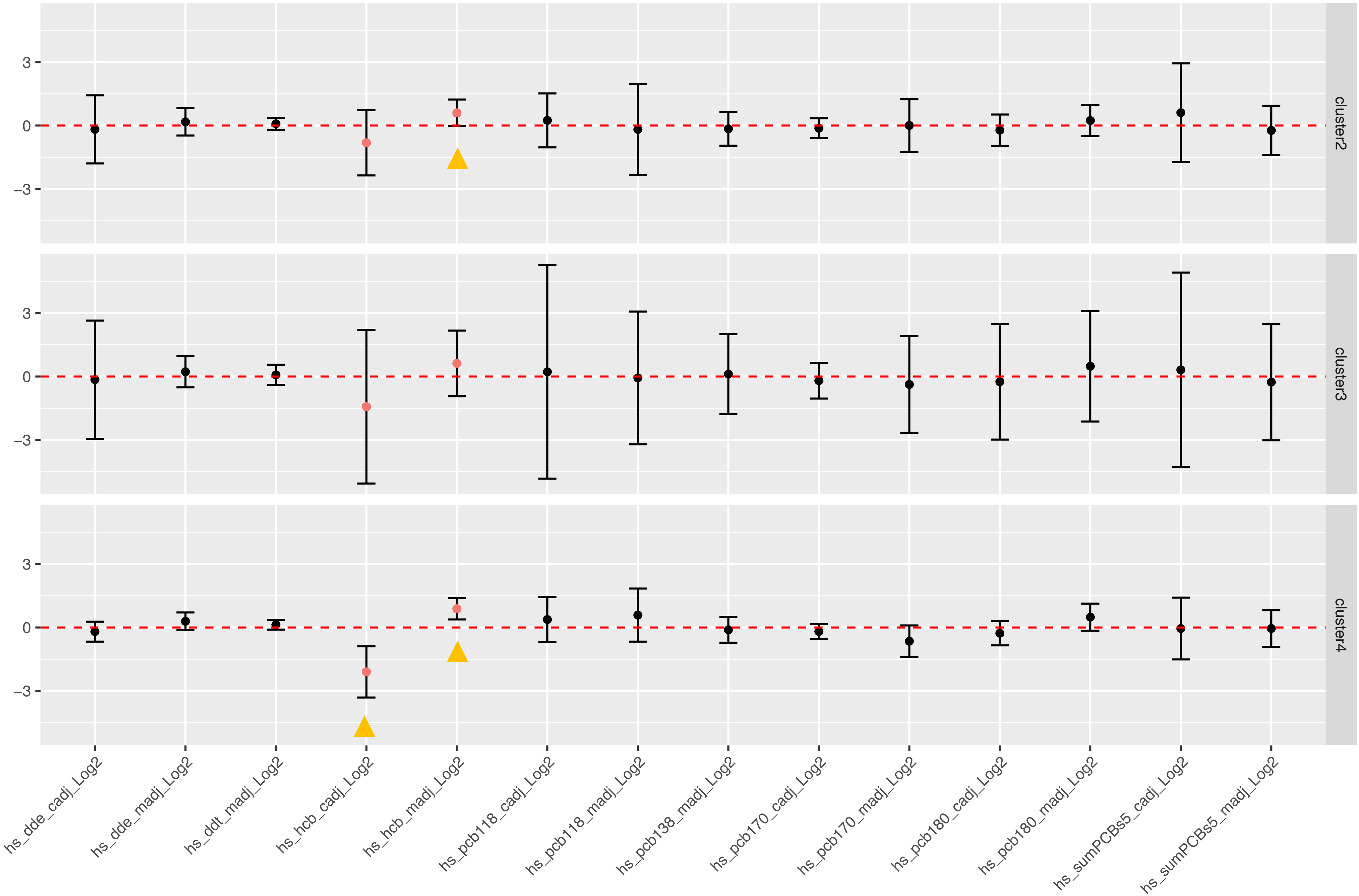
K

Latent Subgroup

BMI z-score

Tune the # of latent subgroup

Point estimate for expsoure effect with 95% CI



Point estimate for metabolite effect with 95% CI