Automated Prior Elicitation for Bayesian Metabolomics Analysis

JSM 2025 | Flexible Prior Elicitation for Bayesian Analysis

Chiraag Gohel

The Rahnavard Lab, The George Washington University

2025-08-06

Table of contents I

Introduction

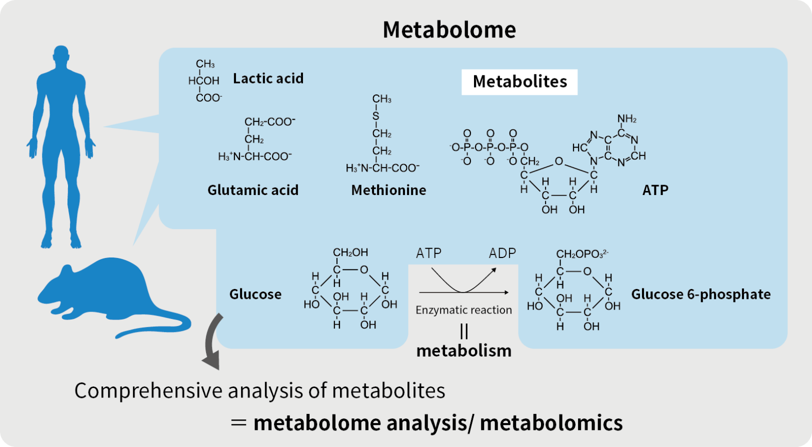
Common issues in statistical testing

Simulation Study

Conclusion

Introduction

What is metabolomics?



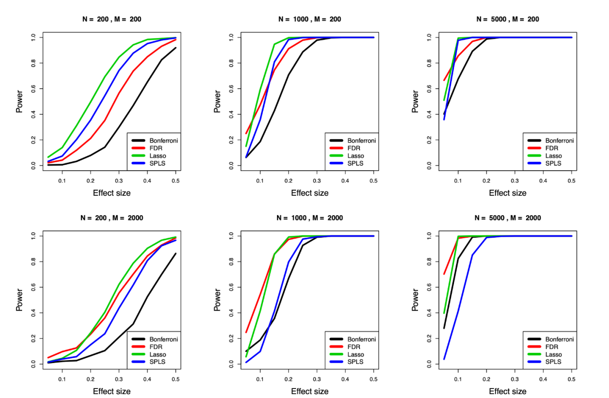
Comprehensive analysis of metabolites

= **metabolome analysis/ metabolomics**

Figure 1: From Human Metabolome Technologies

Common issues in statistical testing

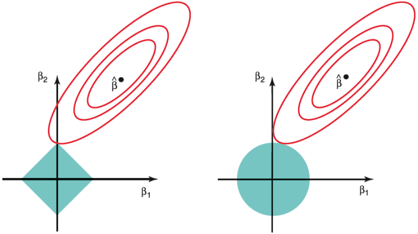
Univariate testing may lack power



- **Peluso et al. 2021**: "...the complex non-normal structure of metabolic profiles and outcomes may bias the permutation results leading to overly conservative threshold estimates..."
- **Henglin et al. 2022**: "We observed that when the number of metabolites was similar to or exceeded the number of study subjects, as is common with nontargeted metabolomics performed in small cohorts, sparse multivariate models demonstrated the most consistent results and the most statistical power."

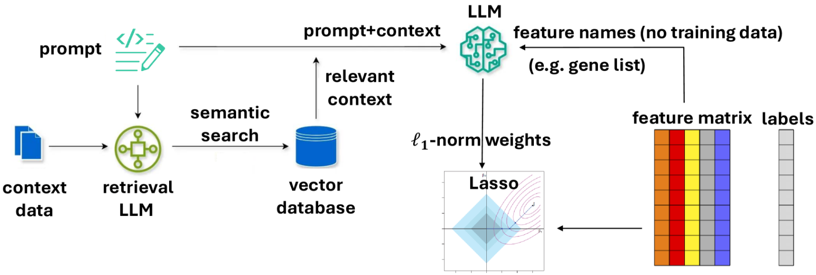
What to do

- High dimensionality $(p >> n)$
- Can lean on assumptions of sparsity
- Prior knowledge from previous studies, literature, and curated databases



Prior Work

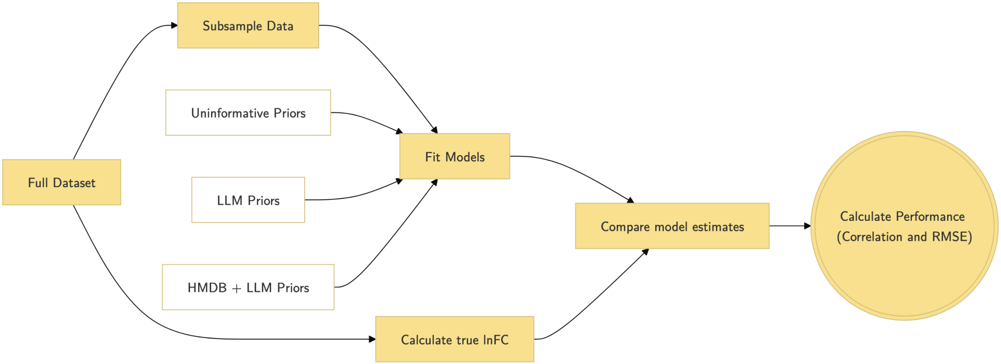
- Current work is inspired from the LLM-Lasso[1]



[1] Zhang, E., Goto, R., Sagan, N., Mutter, J., Phillips, N., Alizadeh, A., Lee, K., Blanchet, J., Pilanci, M., and Tibshirani, R. (2025), "LLM-Lasso: A Robust Framework for Domain-Informed Feature Selection and Regularization," arXiv.

Simulation Study

Empirical Monte-Carlo Subsampling



Full Dataset

Subsample Data

Uninformative Priors

LLM Priors

HMDB + LLM Priors

Fit Models

Calculate true lnFC

Compare model estimates

Calculate Performance
(Correlation and RMSE)

Experimental Design

Ground Truth: Empirical natural log fold change (lnFC) from full MTBLS1 dataset (n=132)

$$\beta_j^{\text{true}} = \log\left(\frac{\bar{y}_j^{\text{case}}}{\bar{y}_j^{\text{control}}}\right)$$

Evaluation: Subsampled data (n=10-40) with cross-validation

Modeling

All Bayesian models use the same log-link GLM structure with different prior specifications:

$$
\begin{aligned}
y_{ij} &\sim \mathcal{N}(\mu_{ij}, \sigma_j^2) && (1)\\
\log(\mu_{ij}) &= \alpha_j + \beta_j \cdot x_i && (2)\\
\alpha_j &\sim \mathcal{N}(\log(\bar{y}_j), 1.0) && (3)\\
\sigma_j &\sim \text{HalfNormal}(0.5) && (4)
\end{aligned}
$$

where y_{ij} is abundance for sample i and metabolite j, $x_i \in \{0, 1\}$ is group indicator, and β_j represents the natural log fold change (lnFC) for metabolite j.

Simulation Study

LLM Prior Elicitation Process

Step 1: LLM analyzes metabolite + study context

$\text{LLM}(\text{metabolite}, \text{condition}) \to \{d_j, m_j, c_j, r_j\}$

Step 2: Map qualitative predictions to numerical priors

$\{d_j, m_j, c_j\} \xrightarrow{\text{mapping}} \{\mu_j^{\text{LLM}}, \sigma_j^{\text{LLM}}\}$

Step 3: Use as informative priors in Bayesian model $\beta_j \sim \mathcal{N}(\mu_j^{\text{LLM}}, \sigma_j^{\text{LLM}})$

Simulation Study

Magnitude-Based Prior Mapping

Magnitude drives effect size, Confidence drives uncertainty. Effect sizes (m_j) are on the natural log scale.

Conservative Mapping

$$\mu_j^{\text{LLM}} = m_j \cdot \text{sign}(d_j)$$
$$\sigma_j^{\text{LLM}} = f(c_j)$$

where $m_j \in \{0.055, 0.104, 0.173\}$ for magnitude $\in$ {small, moderate, large}

Moderate Mapping

$$\mu_j^{\text{LLM}} = m_j \cdot \text{sign}(d_j)$$
$$\sigma_j^{\text{LLM}} = f(c_j)$$

where $m_j \in \{0.083, 0.152, 0.243\}$ for magnitude $\in$ {small, moderate, large}, and $f(c_j) \in \{0.3, 0.5, 0.7\}$ for confidence $\in$ {high, med, low}

where c_j is LLM confidence, $d_j \in$ {increase, decrease, unchanged} is predicted direction, and m_j is predicted magnitude

Priors

Oracle Prior (Upper Bound)

$$\beta_j \sim \mathcal{N}(\beta_j^{\text{true}}, 0.25)$$

Weakly Informative Prior

$$\beta_j \sim \mathcal{N}(0, 2)$$

LLM-Informed Hierarchical Prior

Group metabolites by LLM predictions and use intelligent pooling:

$$\begin{aligned} \text{Group means}_g &\sim \mathcal{N}(\mu_g^{\text{LLM}}, 3.0) \\ \beta_j &\sim \mathcal{N}(\text{Group means}_{g[j]}, 2.0) \end{aligned}$$

where group g is mapped to μ_g^{LLM} as follows:

$$\mu_g^{\text{LLM}} = \begin{cases} -0.1, & \text{if } g = \text{decrease}, \\ 0.0, & \text{if } g = \text{unchanged}, \\ +0.1, & \text{if } g = \text{increase}. \end{cases}$$

Experimental Design

Ground Truth: Empirical natural log fold change (lnFC) from full MTBLS1 dataset (n=132)

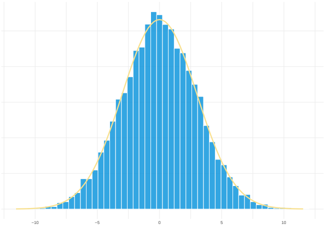
$$\beta_j^{\text{true}} = \log\left(\frac{\bar{y}_j^{\text{case}}}{\bar{y}_j^{\text{control}}}\right)$$

Evaluation: Subsampled data (n=10-40) with cross-validation

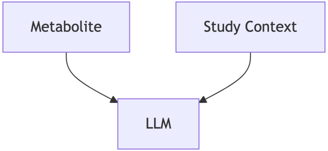
- Oracle provides theoretical upper bound (perfect biological knowledge)
- LLM methods test practical biological knowledge integration
- Classical methods provide statistical baselines

Comparing three different models

Weak Prior



LLM Prior w/no context


Metabolite

Study Context

LLM

LLM Prior w/HMDB context

How well can we recover the truth?

To test the effectiveness of LLM-generated priors, we designed a simulation study.

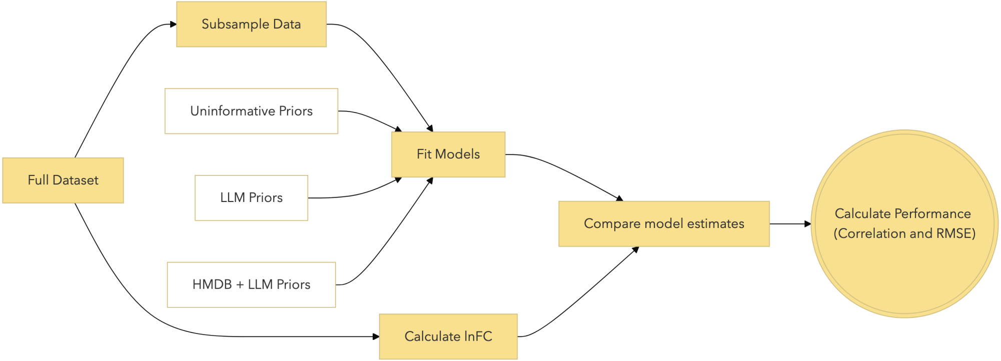
Goal: Recover "ground truth" effect sizes from a small dataset.
Dataset: MTBLS1 (Type 2 Diabetes)
"Ground Truth": Natural log fold changes (lnFC) from the full dataset.

Models Compared:

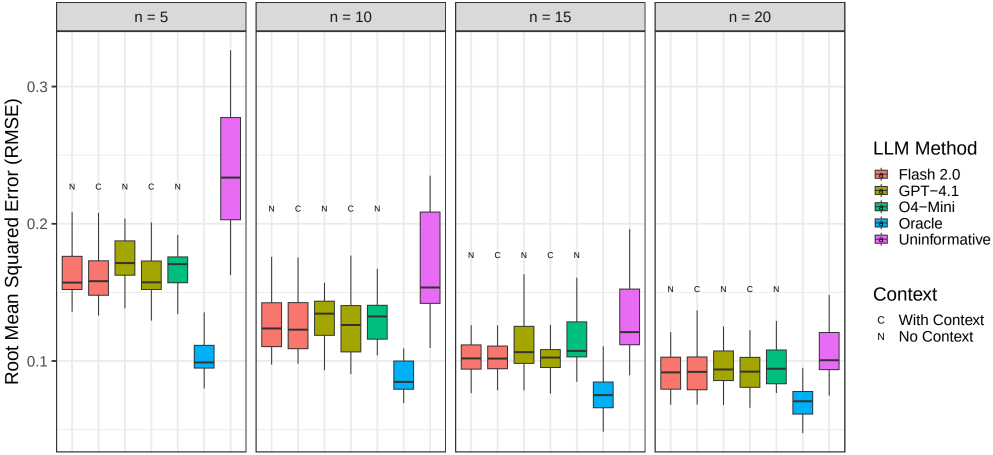
- **Uninformative Bayesian**: A baseline with wide, non-specific priors.
- **LLM Priors**: Priors generated by Gemini using only metabolite names and study context.
- **LLM + Context**: Priors generated by Gemini using metabolite names, study context, AND biological information from the HMDB database.

Overview

For each dataset:



LLM-informed priors improve recovery



Key Findings

- **Oracle Establishes Upper Bound**: Perfect biological knowledge achieves $r = 0.97$, providing theoretical maximum for prior performance.
- **Magnitude-Based Mapping Critical**: Using LLM magnitude predictions (small/moderate/large) for effect sizes significantly improves prior informativeness.
- **Confidence-Calibrated Uncertainty**: High-confidence LLM predictions warrant tighter prior uncertainties, improving statistical efficiency.
- **Empirical Bayes Comparison**: LLM priors compete with James-Stein shrinkage, showing biological knowledge can match statistical methods.
- **Sample Size Effects**: Prior advantage most pronounced at small sample sizes (n=5-10) common in metabolomics studies.

Conclusion

Summary

- **LLM Prior Elicitation Works**: Automated biological knowledge extraction via LLMs produces informative priors for Bayesian metabolomics analysis.
- **Mapping Strategy Matters**: Magnitude-driven effect sizes and confidence-calibrated uncertainties are crucial for translating qualitative LLM insights into effective numerical priors.
- **Practical Impact**: Method particularly valuable for small sample studies (n=5-20) where traditional statistical approaches struggle with high-dimensional metabolomics data.
- **Future Directions**: Integration with structured databases (HMDB) and prompt engineering advances offer paths for further improvement toward oracle-level performance.