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

Introduction

What is metabolomics?

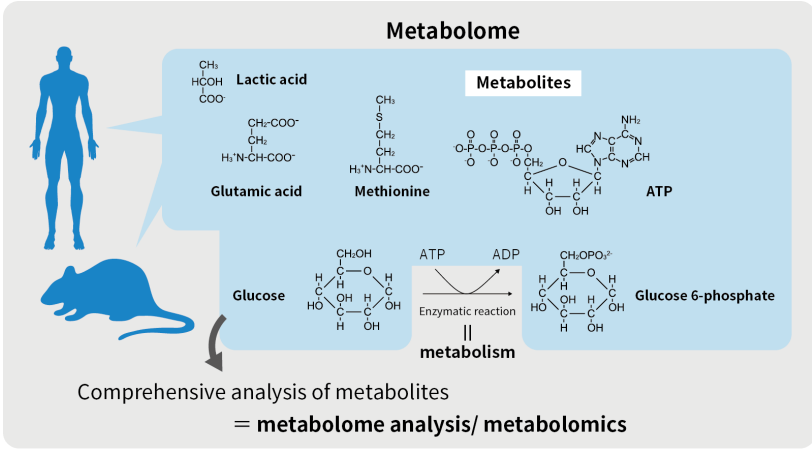
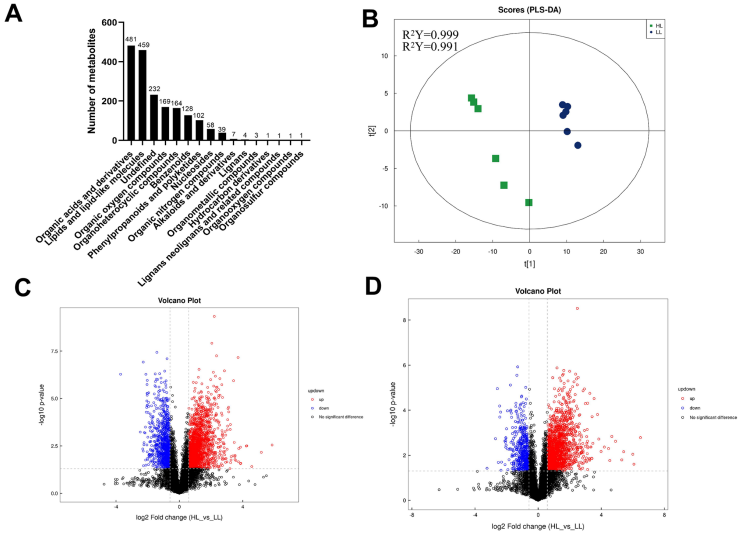


Figure 1: From Human Metabolome Technologies

Effect size drives biological insight



Traditional testing lacks power

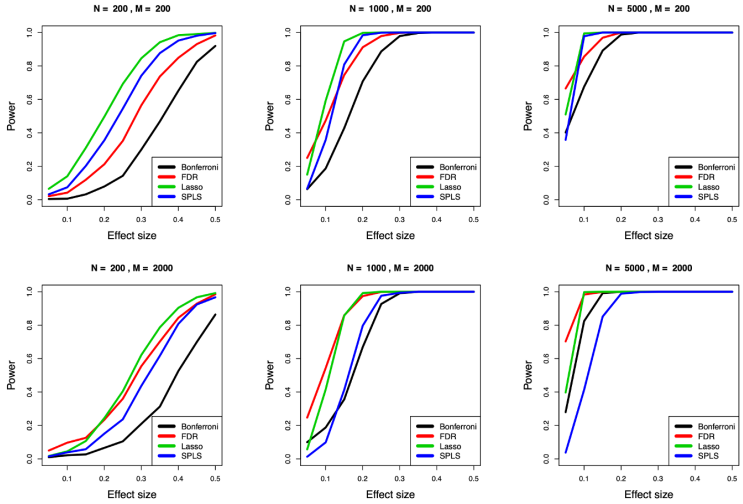


Figure 2: Hanglin, M. et al. (2022). "Quantitative Comparison of Statistical Methods for

Prior work

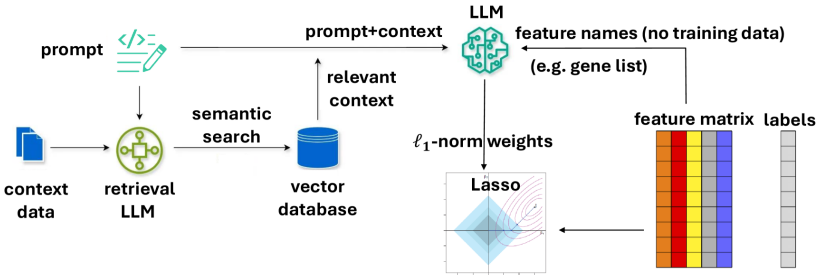
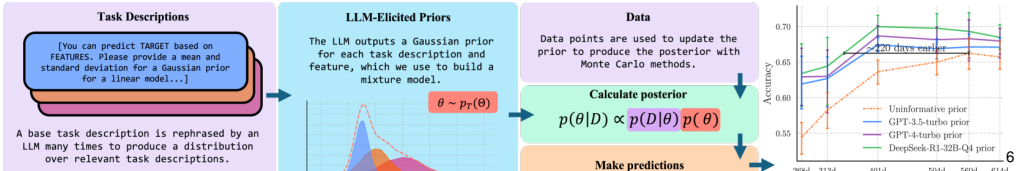
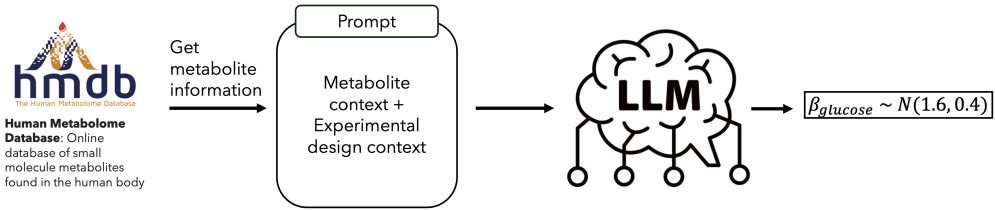


Figure 3: Zhang, E. et al. (2025), "LLM-Lasso: A Robust Framework for Domain-Informed Feature Selection and Regularization," arXiv.



Methods

Prior elicitation framework overview



Metabolite Context Enrichment: HMDB database integration for biological context (pathways, functions, disease associations)

Multi-Model LLM Support: OpenAI (GPT-4o, O3), Google (Gemini 2.0/2.5), with caching and fallback mechanisms

Qualitative-to-Numerical Mapping: Conservative/moderate strength mappings with magnitude-driven effect sizes and confidence-calibrated uncertainties

Hierarchical Bayesian Modeling: LLM-informed metabolite grouping with

LLM prior elicitation process

Step 1: LLM analyzes metabolite + study context

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

where d_j {increase, decrease, unchanged} is predicted direction, m_j {small, moderate, large} is predicted magnitude, c_j $(0, 1)$ is confidence level, and r_j is a string representing the rationale.

Step 2: Map qualitative predictions to numerical priors

$$\overset{\text{LLM}}{j} = f(m_j, d_j)$$
$$\overset{\text{LLM}}{j} = f(c_j)$$

Step 3: Use as informative priors in Bayesian model $_j \;\; N(\overset{\text{LLM}}{j}, \overset{\text{LLM}}{j})$

Priors

LLM Priors: $_j \quad N(^{\text{LLM}}_j, ^{\text{LLM}}_j)$

where $^{\text{LLM}}_j$ and $^{\text{LLM}}_j$ are derived from LLM predictions:

$$^{\text{LLM}}_j = m_j \ \text{sign}(d_j) \tag{1}$$

$$^{\text{LLM}}_j = f(c_j) \tag{2}$$

Conservative Mapping

$m_j \quad \{0.08, 0.15, 0.25\}$ for {small, moderate, large}

$f(c_j) \quad \{0.5, 0.7, 0.9\}$ for {high, med, low} confidence

Moderate Mapping $m_j \quad \{0.12, 0.22, 0.35\}$ for {small, moderate, large}

$f(c_j) \quad \{0.3, 0.5, 0.7\}$ for {high, med, low} confidence

LLM-Informed Hierarchical Prior

Group metabolites by LLM predictions and use intelligent pooling:

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

where group g is mapped to $\overset{\text{LLM}}{g}$ as follows:

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

Modeling

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

$$y_{ij} \quad N(_{ij}, ^2_j) \tag{3}$$

$$\log(_{ij}) = _j + _j \; x_i \tag{4}$$

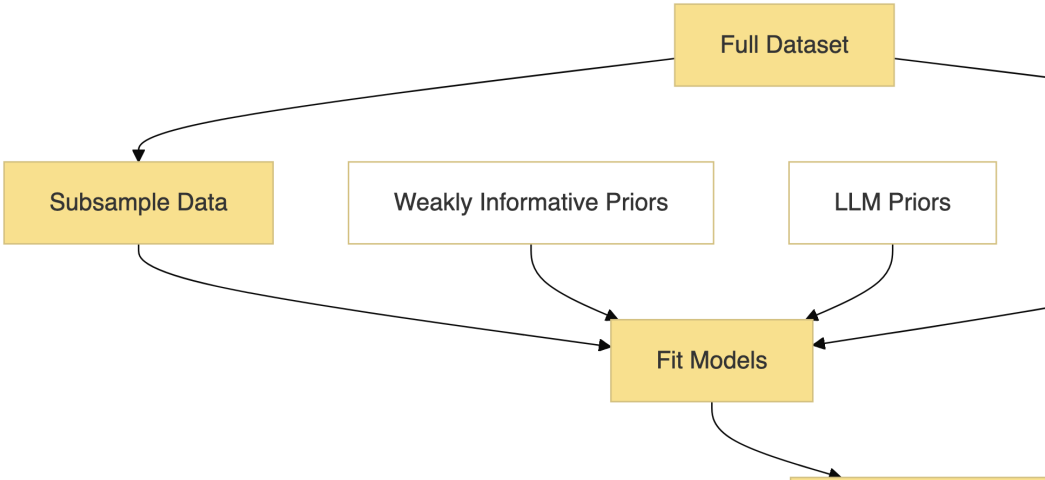
$$_j \quad N(\log(y_j +), 1.0) \tag{5}$$

$$_j \quad \text{HalfNormal}(0.5) \tag{6}$$

where y_{ij} is abundance for sample i and metabolite j, $x_i \;\; \{0, 1\}$ is group indicator, $_j$ represents the natural log fold change (lnFC) for metabolite j, and is a small constant to avoid log(0).

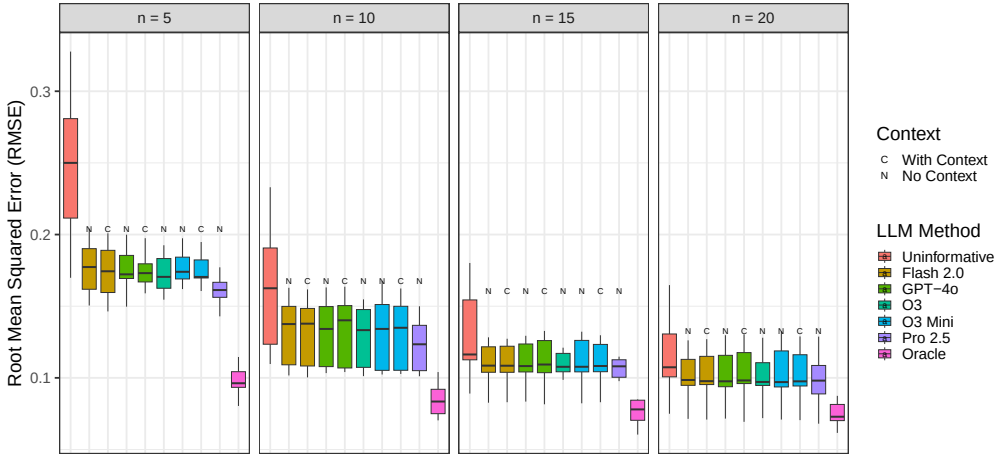
Simulation Study

Empirical Monte-Carlo Subsampling



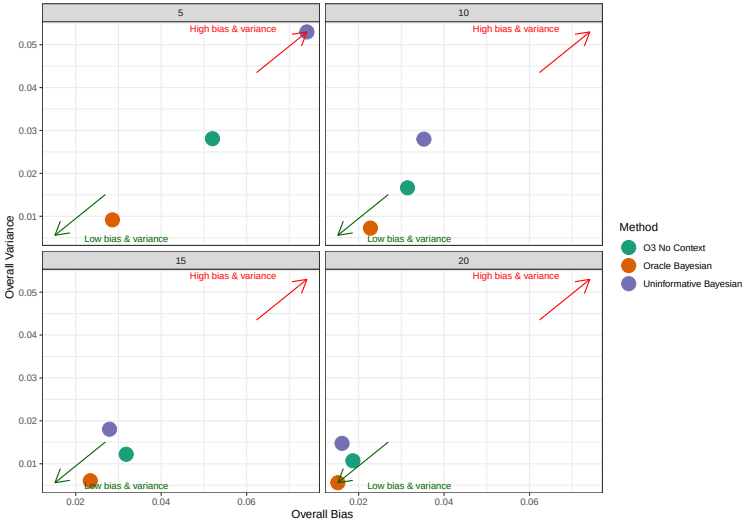
Results

LLM-informed priors improve recovery



$$\text{RMSE} = \frac{1}{n}\sum_{i}(\hat{\theta}_i - \theta_{\text{true},i})^2$$

LLM Informed estimators are finite-sample efficient



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.
Added Context May Not Matter: Including biological context from the HMDB in LLM prompts did not significantly improve prior performance in this study.
Performance is Model Agnostic: Different LLMs (OpenAI, Google) yielded similar results, indicating robustness across models.
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 of other databases, alongside more sophisticated mapping approaches and historical data.