Speaker Notes

Automated Prior Elicitation for Bayesian Metabolomics Analysis

Chiraag Gohel

2025-08-06

What is metabolomics?

Introduction: Metabolomics represents the comprehensive quantitative measurement of small-molecule metabolites in biological systems.

- Comprehensive profiling of small-molecule metabolites in cells, tissues, or biofluids to capture the biochemical phenotype in real time.
- Links upstream genomic and proteomic variation to downstream physiological phenotypes and disease states.
- Detects pathway-level perturbations in energy metabolism, lipid metabolism, and amino acid metabolism underlying exposures, interventions, or pathological processes.
- Enables discovery and validation of biomarkers for diagnosis, prognosis, and treatment response.

Scientific Context: Metabolomics provides the most proximal readout of biological activity, reflecting the integrated response of cellular processes to genetic, environmental, and pathological influences.

Effect size drives biological insight

Statistical Challenge: High-dimensional metabolomics data requires careful consideration of effect size estimation for meaningful biological interpretation.

- **Multiple Testing Context**: With thousands of metabolites tested simultaneously, p-values alone provide insufficient discrimination—effect sizes (fold-change, standardized mean differences) reveal biologically meaningful signal.
- **Multiple Testing Control**: Ranking by effect size magnitude prior to FDR adjustment reduces spurious discoveries and improves reproducibility.
- **Power Analysis**: Effect size estimates inform sample size calculations for follow-up studies and clinical validation experiments.

- **Pathway Analysis**: Effect sizes aggregate naturally in gene-set enrichment analysis and hierarchical models to quantify pathway-level perturbations.
- **Meta-Analysis**: Confidence intervals around effect sizes enable systematic combination of results across laboratories, instruments, and analytical platforms.

Statistical Priority: Effect size estimation addresses both the magnitude and practical significance of observed metabolic changes.

Traditional testing lacks power

Statistical Problem: Standard hypothesis testing approaches exhibit suboptimal performance in the high-dimensional, small-sample regime characteristic of metabolomics studies.

- **Dimensionality Challenge**: Small sample sizes (n=10-50) combined with high dimensionality (p=1000-10000 metabolites) result in severely reduced statistical power.
- **Detection Limitations**: Traditional methods suffer from high false negative rates, failing to detect many true biological effects.
- **Methodological Evidence**: Peluso et al. (2021) demonstrated that "the complex non-normal structure of metabolic profiles and outcomes may bias the permutation results leading to overly conservative threshold estimates."
- **Empirical Validation**: Henglin et al. (2022) showed 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."

Bayesian Solution: Bayesian methods offer superior performance in this regime through principled incorporation of prior information, but require effective prior elicitation strategies.

Available Resources: Modern biological databases (HMDB, PubChem) and large language models provide unprecedented opportunities for automated prior specification.

Prior work

Methodological Foundation: This work builds upon recent advances in large language model-assisted statistical modeling and automated prior elicitation.

- **LLM-Lasso Framework (Zhang et al. 2025)**: Established the use of large language models for feature selection in high-dimensional regression problems, demonstrating that LLMs can effectively identify relevant predictors through domain knowledge integration.
- **AutoElicit Methodology (Capstick et al. 2025)**: Developed automated expert prior elicitation using LLMs for predictive modeling, providing theoretical and empirical evidence that LLMs can systematically translate qualitative domain expertise into quantitative prior distributions.

Research Contribution: We extend these methodological advances to metabolomics applications, where extensive biological knowledge is available and essential for meaningful statistical inference.

Prior elicitation framework overview

System Architecture: We present a comprehensive automated prior elicitation framework integrating biological knowledge bases with large language models for Bayesian metabolomics analysis.

- **Biological Context Integration**: HMDB database information including metabolic pathways, biological functions, and disease associations provides structured biological context for LLM analysis.
- **Multi-Model LLM Infrastructure**: Support for multiple LLM providers (OpenAI GPT-4o and O3 models, Google Gemini 2.0/2.5) with intelligent caching, error handling, and fallback mechanisms ensures robust operation.
- **Qualitative-to-Numerical Mapping**: Novel mapping functions translate qualitative LLM predictions (direction, magnitude, confidence) into calibrated numerical prior distributions using conservative and moderate strength parameterizations.
- **Hierarchical Bayesian Architecture**: LLM-informed metabolite grouping enables intelligent pooling while avoiding excessive shrinkage through carefully calibrated hierarchical structures.
- **Validation Framework**: Rigorous evaluation using Monte Carlo subsampling with bias-variance decomposition provides comprehensive performance assessment against empirical ground truth.

Implementation Status: This framework represents a production-ready system suitable for deployment in metabolomics research applications.

LLM prior elicitation process

Methodological Workflow: The prior elicitation process consists of three sequential steps transforming biological knowledge into statistical priors.

Step 1: Qualitative Analysis

- Input: Metabolite identifier, experimental condition specification (e.g., “Type 2 diabetes vs control”), and optional structured biological context from HMDB database.
- Output: Structured qualitative predictions including direction (increase/decrease/unchanged), effect magnitude (small/moderate/large), confidence level (continuous scale 0-1), and textual reasoning.

Step 2: Quantitative Mapping

- Prior mean specification: _j^LLM determined by magnitude-direction interaction using calibrated effect size mappings.
- Prior variance specification: (_j$^{\text{LLM})}$2 determined by confidence-dependent uncertainty parameterization.
- Mapping calibration: Parameters derived from empirical analysis of metabolomics effect size distributions in literature.

Step 3: Bayesian Integration

- Prior specification: LLM-derived parameters define Normal prior distributions for log fold change parameters _j in generalized linear model framework.
- Model estimation: Standard Bayesian inference proceeds using informative priors within PyMC implementation.

Critical Innovation: Step 2 mapping functions represent the key methodological contribution, requiring careful calibration to ensure biological plausibility and statistical validity.

Priors

Prior Specification Framework: We implement multiple prior parameterizations to evaluate the impact of different strength assumptions on statistical inference.

Mapping Parameterizations:

- **Conservative Mapping**: Effect sizes (0.08, 0.15, 0.25) for (small, moderate, large) magnitudes with uncertainty parameters (0.5, 0.7, 0.9) for (high, medium, low) confidence levels.
- **Moderate Mapping**: Effect sizes (0.12, 0.22, 0.35) for (small, moderate, large) magnitudes with uncertainty parameters (0.3, 0.5, 0.7) for (high, medium, low) confidence levels.
- **Calibration Basis**: Parameters derived from empirical analysis of effect size distributions in published metabolomics studies, ensuring biological plausibility.

Benchmark Priors:

- **Oracle Prior**: _j ~ N(_j^true, 0.25) representing perfect biological knowledge with realistic measurement uncertainty, establishing theoretical performance upper bound.
- **Weakly Informative Prior**: _j ~ N(0, 2) representing standard uninformative Bayesian baseline for comparison.

Parameter Justification: All prior parameters are data-driven, derived from systematic analysis of effect size distributions in metabolomics literature to ensure biological and statistical validity.

Modeling

Statistical Model: All methods employ identical log-link generalized linear model specifications, differing only in prior parameterizations to ensure fair comparison.

- **Log-Link Structure**: Appropriate for positive abundance data, providing directly interpretable log fold change parameters _j.
- **Numerical Stability**: Epsilon regularization term prevents log(0) computational issues inherent in metabolomics abundance measurements.
- **Likelihood Equivalence**: Identical likelihood specifications across all methods ensure that performance differences reflect solely the influence of prior information.

Comparative Framework: This design enables direct assessment of prior specification impact on statistical inference quality without confounding from different likelihood assumptions.

Simulation Study

Evaluation Design: We implement a rigorous empirical validation framework using Monte Carlo subsampling to assess prior specification performance under realistic constraints.

- **Ground Truth Establishment**: Empirical log fold changes calculated from complete MTBLS1 dataset (n=132, Type 2 diabetes vs control) serve as validation targets.
- **Statistical Challenge**: Methods must recover ground truth parameters using only small subsamples (n=5-20 per group), reflecting typical metabolomics study constraints.
- **Monte Carlo Framework**: Multiple random subsampling iterations provide robust statistical assessment of method performance across different data realizations.

Validation Logic: Informative priors should demonstrate superior parameter recovery compared to uninformative alternatives when data are limited, approaching oracle performance bounds.

Dataset Rationale: MTBLS1 represents a well-characterized metabolomics study with established biological effects, providing realistic validation conditions for method comparison.

LLM-informed priors improve recovery

Performance Results: Empirical validation demonstrates systematic improvement in parameter recovery using LLM-informed priors compared to standard baselines.

Statistical Interpretation:

- Performance metric: Root Mean Square Error (RMSE) measuring deviation from ground truth parameters (lower values indicate superior performance).

- Method comparison: LLM-based approaches compared against uninformative Bayesian baseline and oracle upper bound.
- Context evaluation: C/N annotations indicate with/without HMDB biological context integration.
- Sample size dependency: Results stratified by subsample size to assess performance across data availability scenarios.

Principal Findings:

- **Consistent Superiority**: LLM-informed methods demonstrate statistically significant RMSE reduction compared to uninformative priors across all sample sizes.
- **Small Sample Advantage**: Performance improvement most pronounced at small sample sizes (n=5-10) where prior information provides greatest statistical benefit.
- **Model Robustness**: Similar performance across different LLM architectures (OpenAI, Google) indicates method stability across model choices.
- **Context Analysis**: Biological context integration shows marginal benefit, suggesting metabolite nomenclature contains substantial biological signal.

Statistical Validation: Differences achieve statistical significance ($p<0.05$) across multiple sample sizes and Monte Carlo replicates, with confidence intervals excluding null hypothesis of equal performance.

LLM Informed estimators are finite-sample efficient

Bias-Variance Decomposition: Mean squared error decomposition reveals the statistical mechanisms underlying LLM prior performance advantages.

Analytical Framework:

- Bias component: Systematic deviation of estimator expectation from true parameter values.
- Variance component: Sampling variability of estimator across different data realizations.
- Efficiency criterion: Optimal estimators minimize total mean squared error = bias^2 + variance.
- Graphical interpretation: Lower-left quadrant represents superior bias-variance combinations.

Statistical Findings:

- **Improved Efficiency**: LLM-informed estimators achieve superior bias-variance tradeoffs compared to uninformative approaches, with reduced total MSE.
- **Regularization Effect**: At small sample sizes, LLM priors provide beneficial regularization, substantially reducing estimator variance.

- **Oracle Approximation**: LLM methods approach oracle performance bounds, indicating effective incorporation of biological knowledge.
- **Finite-Sample Optimality**: Classical bias-variance tradeoff principles favor variance reduction over bias avoidance in small-sample regimes.

Theoretical Implication: Results confirm that well-calibrated informative priors can achieve near-optimal finite-sample performance by exploiting the bias-variance tradeoff inherent in Bayesian estimation.

Summary

Principal Findings:

1. **Automated Prior Elicitation Efficacy**: Large language models demonstrate capacity for systematic extraction and quantification of biological domain knowledge, producing statistically informative priors that enhance Bayesian parameter estimation in high-dimensional metabolomics analysis.

2. **Mapping Function Criticality**: The methodological innovation centers on calibrated transformation functions that convert qualitative LLM predictions into numerically appropriate prior distributions through magnitude-dependent effect size specifications and confidence-adjusted uncertainty parameters.

3. **Biological Context Evaluation**: Contrary to expectations, incorporation of detailed biological context from HMDB database produced minimal performance improvement, suggesting that metabolite nomenclature contains substantial inherent biological signal sufficient for effective prior elicitation.

4. **Cross-Model Robustness**: Consistent performance across multiple LLM architectures (OpenAI GPT-4o/O3, Google Gemini) indicates methodological stability independent of specific model implementation choices.

5. **Small-Sample Regime Optimization**: The approach demonstrates greatest utility in small sample scenarios (n=5-20) characteristic of typical metabolomics studies, where traditional frequentist approaches exhibit reduced statistical power.

6. **Research Extensions**: Future methodological developments include integration of additional biological databases, adaptive mapping function parameterization, and incorporation of historical effect size distributions for enhanced calibration.

Methodological Contribution: This work establishes large language models as effective automated statistical consultants for domain-informed Bayesian analysis, providing a generalizable framework for AI-assisted prior specification across scientific disciplines.

Discussion: Questions regarding methodology, implementation details, or potential applications are welcome.

Backup Slides & Technical Details

Implementation Architecture

Computational Framework: Implementation utilizes PyMC probabilistic programming framework with comprehensive convergence diagnostics including R-hat statistics, effective sample size monitoring, and trace plot analysis to ensure reliable posterior sampling.

Multi-Model Infrastructure: System architecture supports heterogeneous LLM providers through unified API interface with intelligent request routing, exponential backoff retry mechanisms, response caching for computational efficiency, and automatic fallback protocols for enhanced reliability.

Parameter Calibration: Conservative mapping parameters derived from systematic meta-analysis of published metabolomics effect size distributions, ensuring biological plausibility and statistical validity of prior specifications.

Hierarchical Modeling: Careful parameterization of hierarchical structures using minimal pooling strength to prevent excessive shrinkage while maintaining computational stability and biological interpretability.

Statistical Design Choices

Likelihood Specification: Log-link generalized linear model structure selected over log-transformation approaches to preserve appropriate noise characteristics for positive abundance data and provide direct interpretability of log fold change parameters.

Prior Parameterization: Effect size ranges calibrated specifically to diabetes metabolomics literature to ensure biological relevance and appropriate uncertainty quantification for the target application domain.

Validation Design: Sample size ranges (n=5-20) selected to reflect typical constraints in metabolomics studies while enabling rigorous Monte Carlo validation against empirical ground truth parameters.

Cross-Validation Protocol: Stratified random subsampling ensures robustness of performance estimates across different data partitions and reduces dependence on specific train/test configurations.

Methodological Extensions

Uncertainty Integration: Incorporation of prediction uncertainty from LLM confidence scores into prior variance specifications through hierarchical uncertainty propagation methods.

Meta-Analytic Framework: Extension to multi-study settings using LLM-informed random effects models for cross-laboratory and cross-platform metabolomics synthesis.

Adaptive Parameterization: Study-specific mapping function calibration based on experimental design characteristics, sample composition, and analytical platform specifications.

Database Integration: Expansion beyond HMDB to incorporate KEGG, BioCyc, and PubChem knowledge bases for enhanced biological context and pathway-level modeling capabilities.

Statistical Extensions: Enhanced uncertainty quantification through coverage probability analysis, prediction interval calibration, and Bayesian model checking procedures.

Current Limitations

Scope Constraints: Current implementation focuses on univariate differential expression analysis rather than multivariate pathway-level or network-based modeling approaches.

Calibration Dependencies: Method requires empirical ground truth data for mapping function parameterization, limiting immediate applicability to entirely novel biological contexts.

Domain Specificity: Framework optimized for metabolomics applications, though underlying methodology exhibits potential for generalization to other high-dimensional biological domains.

Hierarchical Underutilization: Present implementation does not fully exploit available biological pathway hierarchies and metabolic network structures for enhanced prior specification and improved statistical power.