# Research Design Overview

**Refined Research Question**

Can a hierarchical Bayesian Markov Chain Monte Carlo (MCMC) framework, employing spike-and-slab priors for automatic feature selection among stock-level predictors derived from Jensen, Kelly, and Pedersen (2023), enhance out-of-sample predictive accuracy of monthly dividend-adjusted returns for S&P 500 stocks compared to traditional interpretable machine learning models?

**Data Overview**

The data comprises monthly returns (including dividends) for the constituents of the S&P 500 index, covering the period from 1965 to 2024. These returns have been sourced from CRSP and include unique stock identifiers (PERMNO) and corresponding dates to facilitate merging. Additionally, monthly stock-level factor data, spanning the same timeframe (1965–2024), has been obtained, encompassing 153 stock-level characteristics organized into 13 economically interpretable themes as defined by Jensen, Kelly, and Pedersen (2023). These themes include Accruals, Debt Issuance, Investment, Low Leverage, Low Risk, Momentum, Profit Growth, Profitability, Quality, Seasonality, Size, Short-Term Reversal, and Value.

**Bayesian MCMC Model Overview**

I will employ a hierarchical Bayesian regression framework utilizing Markov Chain Monte Carlo (MCMC) techniques to forecast monthly returns. The model integrates spike-and-slab priors at the individual predictor (factor) level for automatic variable selection. Each factor's regression coefficient (beta) will be assigned a spike-and-slab prior:

* **Spike:** Concentrated at zero, representing an assumption that most predictors have negligible predictive power.
* **Slab:** Broader distribution allowing significant deviation from zero, representing predictors with substantial predictive impact.

Factors will be hierarchically structured based on the 13 economic themes from Jensen, Kelly, and Pedersen (2023), enabling partial pooling of information and improving statistical efficiency and interpretability.

**Baseline ML Models Overview**

To benchmark the performance of the Bayesian model, two traditional interpretable machine learning approaches will be employed:

* **LASSO (Least Absolute Shrinkage and Selection Operator):** Regularization approach enforcing sparsity and identifying key predictive features by shrinking insignificant predictors towards zero.
* **Decision Trees:** Non-parametric method providing clear interpretability through intuitive, tree-based decision rules, facilitating the identification of critical stock-level predictors and their thresholds.

**Comparison Plan**

The Bayesian MCMC model and baseline ML models output different prediction types—Bayesian models produce posterior predictive distributions, whereas ML models generate single-point forecasts. To effectively compare forecasting accuracy, I will:

* Convert Bayesian posterior predictive distributions into comparable point predictions by utilizing posterior predictive means.
* Evaluate predictive accuracy using out-of-sample metrics, specifically Mean Squared Error (MSE) and predictive R².
* Additionally, assess probabilistic forecasting performance of the Bayesian model using metrics such as Continuous Ranked Probability Score (CRPS) to fully leverage the Bayesian output's distributional information.
* Feature selection comparison will be based on factors selected by spike-and-slab posterior inclusion probabilities versus those identified by LASSO and decision trees.
* Economic interpretability comparison will involve evaluating the consistency and economic rationale behind selected predictors across all models.

**Expected Contributions**

* Methodological Contribution: Demonstrate the efficacy of hierarchical Bayesian MCMC models with spike-and-slab priors for simultaneous feature selection and forecasting accuracy improvement in financial return prediction contexts.
* Empirical Insights: Provide comprehensive evidence regarding the predictive power of stock-level factors organized into economically meaningful themes, enhancing understanding of market dynamics.
* Academic Relevance: Bridge gaps between advanced Bayesian statistical modeling and practical applications in asset pricing and financial forecasting, enriching the literature by juxtaposing traditional ML benchmarks with Bayesian inference techniques.