Joint Progression Modeling (JPM): A Probabilistic Framework for Mixed-Pathology Progression

Hongtao Hao HHAO9@WISC.EDU
University of Wisconsin–Madison, USA

Joseph L. Austerweil JOSEPH.AUSTERWEIL@GMAIL.COM
Chiba Institute of Technology, Japan & University of Wisconsin-Madison, USA

for the Alzheimer's Disease Neuroimaging Initiative*

Abstract

Event-based models (EBMs) infer disease progression from cross-sectional data, and standard EBMs assume a single underlying disease per individual. In contrast, mixed pathologies are common in neurodegeneration. We introduce the Joint Progression Model (JPM), a probabilistic framework that treats single-disease trajectories as partial rankings and builds a prior over joint progressions. We study several JPM variants (Pairwise, Bradley–Terry, Plackett–Luce, and Mallows) and analyze three properties: (i) *calibration*—whether lower model energy predicts smaller distance to the ground truth ordering; (ii) *separation*—the degree to which sampled rankings are distinguishable from random permutations; and (iii) *sharpness*—the stability of sampled aggregate rankings. All variants are calibrated, and all achieve near-perfect separation; sharpness varies by variant and is well-predicted by simple features of the input partial rankings (number and length of rankings, *conflict*, and *overlap*). In synthetic experiments, JPM improves ordering accuracy by roughly 21% over a strong EBM baseline (SA-EBM) that treats the joint disease as a single condition. Finally, using NACC, we find that the Mallows variant of JPM and the baseline model (SA-EBM) have results that are more consistent with prior literature on the possible disease progression of the mixed pathology of AD and VaD.

Keywords: Neurodegenerative diseases, Disease progression, Event-based model, Rank aggregations, Bayesian inference, Mixed Pathology

Data and Code Availability We used both synthetic and real-world datasets (NACC and ADNI). NACC data can be requested through https://naccdata.org, and ADNI via https://adni.loni.usc.edu. We developed a package for JPM, which can be installed via `pip install pyjpm`. Package source codes can be found at https://github.com/jpcca/pyjpm. Reproducible codes for the experiments in this study are available at https://github.com/hongtaoh/jpm.

Institutional Review Board (IRB) The IRB at University of Wisconsin-Madison has reviewed and approved the research (#2025-1254).

1. Introduction

Modern neuroscience often models the progression of neurodegenerative diseases (NDDs) with frameworks like the Event-Based Model (EBM, Fonteijn et al., 2012; Young et al., 2024). These models have provided critical insights into the sequential evolution of biomarkers from cross-sectional data (Young et al., 2014, 2018; Archetti et al., 2019; Venkatraghavan et al., 2019). They have been successfully applied across multiple diseases, such as Alzheimer's (AD), frontotemporal dementia, Huntington's, and Parkinson's (Young et al., 2018; Fonteijn et al., 2012; Oxtoby et al., 2021; Wijeratne et al., 2023; Firth et al., 2020). However, they are built on a simplifying, yet unrealistic, assumption: Each patient suffers from a single, isolated pathology. This assumption is inconsistent with decades of neuropathological

*. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found in Appendix A.

© 2025 H. Hao, J.L.A. & f.t.A.D.N. Initiative*.

Table 1: Participant biomarker measurements

ID	Impacted	ABETA	ADAS13	CDRSB	Entorhinal	FAQ	FDG	Fusiform	Hippocampus	LDELTOTAL	MMSE	MOCA	MidTemp	PTAU	RAVLT	TAU	TRABSCOR	WholeBrain
1	Yes	846.90	13.47	1.64	0.00	-0.17	1.13	0.01	0.01	0.83	30.90	18.41	0.01	11.53	44.87	378.36	124.99	0.74
2	Yes	1187.20	19.92	0.04	0.00	-0.66	1.26	0.02	0.01	5.22	29.12	22.72	0.01	12.32	53.85	113.02	112.70	0.68
3	No	923.78	13.40	-0.10	0.00	2.07	1.30	0.02	0.00	7.18	30.05	22.65	0.01	21.54	43.63	228.85	38.96	0.70
4	Yes	762.25	7.75	0.79	0.00	-0.37	1.22	0.02	0.01	2.66	28.85	19.61	0.01	-3.13	21.16	541.09	78.48	0.68
5	No	506.22	20.85	0.24	0.00	0.48	1.37	0.01	0.00	3.98	28.66	26.52	0.01	14.58	24.91	185.07	114.84	0.68

Note: Biomarker names are abbreviated for brevity. Values of 0.0 result from rounding to two decimal places. Detailed biomarker descriptions are provided in Table 5 (Appendix F).

evidence. The clinical reality is that patients often have multiple diseases (mixed pathology, or comorbidity), especially in the case of NDDs (Rahimi and Kovacs, 2014; Forrest and Kovacs, 2023; Buchman et al., 2021; Chu et al., 2023; Jørgensen et al., 2020; Stirland et al., 2025; Jellinger, 2021). Autopsy studies show that 45.8% of clinically probable AD patients have mixed pathologies such as infarcts, and Lewy bodies (Schneider et al., 2009). The rate of mixed pathologies is even higher in Progressive Supranuclear (70%; Popli et al., 2025) and Parkinson's (84%; Buchman et al., 2021). How can we build models of how diseases jointly progress based on individual disease progression?

To illustrate the challenge of modeling mixed pathology progression, consider Table 1. It contains data of 5 participants (from a larger data set): 3 impacted and 2 healthy. Impacted participants have all triple pathologies: AD, Vascular Dementia (VaD), and Lewy Body Dementia (LBD). Healthy participants are free from all of them. Biomarkers involved in each of the three pathologies are available at Appendix B, and their measurements are continuous. Now the question: what is the ordinal progression of this comorbidity? We define a disease progression as the ordinal order in which biomarkers get pathological. For example, ABETA $\rightarrow$ TAU $\rightarrow$ MOCA $\cdots$. Note that we assume all individuals share a common progression for a specific disease.

A seemingly straightforward solution would be to treat comorbidity as a unique disease and apply existing models such as EBM from scratch. This approach, however, is both data-inefficient and practically infeasible. It discards the wealth of information available in large single-disease cohorts and imposes unrealistic data requirements. The scale of this data-sparsity problem is starkly illustrated by the National Alzheimer's Coordinating Center (NACC) dataset, one of the world's most comprehensive clinical collections for AD and related dementias. While it contains over 13,400 participants with AD alone, only 50 participants with a mixed diagnosis of AD and VaD, and also have the cerebrospinal fluid (CSF) biomarkers essential for modern AD diagnosis and prognosis. (Participants in studies focusing on VaD do not always collect CSF as it can be costly and painful to do so). It is statistically untenable to build a robust progression model from such a small or incomplete sample.

In this study, we introduce the Joint Progression Model (JPM), a comprehensive probabilistic framework designed to model mixed pathologies by integrating knowledge from single-disease cohorts. Critically, the JPM is designed for both generative and inferential tasks. This enables not only the generation of synthetic data to probe disease mechanisms and validate models, but also the robust inference of progression and patients' disease stages. The framework's core connects joint progression to the rich theoretical literature on probabilistic ranking, e.g., Mallows (1957) and Plackett (1975), using these established models to build a prior within a flexible Bayesian framework that can be paired with any disease progression likelihood function, such as the EBM.

In this paper, we make the following contributions. Firstly, we develop the JPM as a theoretical framework for mixed-pathology modeling using the EBM as its likelihood. Secondly, we formally analyze the statistical properties of the JPM, explore how it connects to real-world scenarios, and identify the conditions required for accurate and reliable inference. Lastly, we apply the JPM to both synthetic and real-world datasets, demonstrating up to a 21% improvement in progression estimation accuracy compared to the state-of-the-art single-disease EBM (Hao et al., 2025).

2. Related work

The EBM (Fonteijn et al., 2012) was a seminal contribution, enabling the inference of biomarker event sequences from cross-sectional data. A family of variants followed, extending the EBM to handle patient

heterogeneity (Venkatraghavan et al., 2019), discover disease subtypes (Young et al., 2018; Hao and Austerweil, 2025), and scale to high-dimensional data (Tandon et al., 2023; Wijeratne and Alexander, 2024). While alternative frameworks using deep learning on longitudinal data exist for diagnosis predictions (Lee et al., 2019; Yang et al., 2021), they do not recover event sequences. Crucially, all prior work assumes a single pathology.

This single-disease focus is a major limitation, as co-neuropathologies are highly prevalent and clinically significant (Schneider et al., 2009; Forrest and Kovacs, 2023). To date, modeling mixed pathology progression remains challenging. (Young et al., 2024). The common practices are to exclude comorbid patients from studies, e.g., in Jørgensen et al. (2020), or predict patient state conversion, e.g., healthy to diseased (Oflaz et al., 2023; Maag et al., 2021). Both practices do not allow us to understand how diseases interact as they progress jointly.

We propose to fill this gap by applying probabilistic ranking theory (Tang, 2019). There exist several possible probabilistic models over rankings or orderings, such as the Mallows (Mallows, 1957), Bradley-Terry (Bradley and Terry, 1952), and Plackett-Luce models (Luce et al., 1959; Plackett, 1975). While widely used in other fields such as election (Gormley and Murphy, 2009), consumer preference (Feng and Tang, 2022), and information retrieval (Liu et al., 2009), these models have not yet been applied to joint disease progression. Below, we demonstrate how we integrate these models into a generative and inference framework to tackle the mixed-pathology problem.

3. JPM theory and algorithms

3.1. Problem statement

We define a disease progression as the ordinal sequence in which relevant biomarkers become pathological. We call the progression of an individual disease as a **partial ranking**, and the progression of the comorbidity as an **aggregate ranking**. Consider two diseases with their respective progressions:

- Disease 1: Aβ $\rightarrow$ Tau $\rightarrow$ cognitive decline $\rightarrow$ hippocampus atrophy
- Disease 2: white matter lesions (WML) $\rightarrow$ hippocampus atrophy $\rightarrow$ cognitive decline

The challenge is to determine the most probable aggregate ranking for patients with both diseases. WML $\rightarrow$ Aβ $\rightarrow$ Tau $\rightarrow$ hippocampus atrophy $\rightarrow$ cognitive decline is probably more likely than cognitive decline $\rightarrow$ atrophy $\rightarrow$ WML $\rightarrow$ Tau $\rightarrow$ Aβ. But by how much? Is it also more probable than WML $\rightarrow$ cognitive decline $\rightarrow$ Aβ $\rightarrow$ Tau $\rightarrow$ hippocampus atrophy? A sophisticated model is needed to address these questions when we have multiple partial rankings with biomarkers whose order may conflict across them.

Formally, let $\mathcal{X}^{(k)}$ denote the set of biomarkers associated with disease k, with $k \in \{1, \ldots, K\}$. Let $\mathcal{S} = \left(\sigma^{(1)}, \sigma^{(2)}, ..., \sigma^{(K)}\right)$ denote the K partial rankings, each a total ordering encoding the corresponding (individual) disease's progression over its biomarkers. Note that this is a partial ranking from the perspective of the comorbidity. Let

$$\mathcal{X} = \bigcup_{k=1}^{K} \mathcal{X}^{(k)}$$

be the set of all unique biomarkers, with $m = |\mathcal{X}|$. Any aggregate ranking σ is a permutation of $\mathcal{X}$.

JPM aims to obtain the posterior of aggregate rankings (σ) given the observed mixed-pathology data (D) and partial rankings ($\mathcal{S}$):

$$P(\sigma \mid D, \mathcal{S}) \propto P(D \mid \sigma, \mathcal{S})\, P(\sigma \mid \mathcal{S})$$

Because $\mathcal{S}$ is fixed ground truth, we can drop it from the data likelihood:

$$P(\sigma \mid D, \mathcal{S}) \propto P(D \mid \sigma) \cdot P(\sigma \mid \mathcal{S})$$

We adopt an energy-based formulation for the prior $P(\sigma \mid \mathcal{S})$:

$$P(\sigma \mid \mathcal{S}) \propto \exp\left(-E(\sigma \mid \mathcal{S})\right)$$

where $E(\sigma \mid \mathcal{S})$ is the **energy function** that assigns a low value to more probable rankings. This exponential form ensures positivity and allows the model to concentrate probability mass around low-energy (i.e., high-likelihood) rankings. Below we describe four principled choices for this energy function

3.2. Variant 1: Pairwise Preferences (PP)

The Pairwise Preferences model generalizes majority voting across partial rankings. It treats the energy of aggregate ranking as the outcome of weighted pairwise comparisons between biomarkers.

$$E(\sigma \mid \mathcal{S}) = -\sum_{i,j \in \mathcal{X}, i \neq j} w_{ij} \mathbf{1}[i <_\sigma j]$$

where

- $\mathbf{1}[i <_\sigma j] = 1$ if i precedes j in ranking σ (i.e., $i \prec j$); otherwise 0.
- w_{ij} is the weight of preference for i preceding j.

We derive w_{ij} from partial rankings:

$$w_{ij} = \sum_{k=1}^{K} \lambda_k \cdot r_k(i,j), \quad \forall i,j \in \mathcal{X}, i \neq j$$

with:

- $\lambda_k \geq 0$ is the importance of $\sigma^{(k)}$ (default $\lambda_k = 1$).
- If both i and j are present in $\sigma^{(k)}$: $r_k(i,j) = 1$ when $i \prec j$, and $r_k(i,j) = -1$ when $j \prec i$.
- $r_k(i,j) = 0$ otherwise.

3.3. Variant 2: Generalized Bradley–Terry (BT)

The Bradley-Terry model (Bradley and Terry, 1952) interpreted rankings as arising from latent strength parameters assigned to each item. Each item $i \in \mathcal{X}$ is associated with a parameter θ_i, forming a vector of $\boldsymbol{\theta}$. The probability of i preceding j is:

$$P(i \prec j) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}$$

Let $C_{i,j}$ be the number of times of $i \prec j$ across all partial rankings. If i and j are never jointly observed, or i never precedes j, then $C_{i,j} = 0$. For a partial ranking $\sigma^{(k)} = \left(\sigma_1^{(k)}, \sigma_2^{(k)}, \ldots, \sigma_{n_k}^{(k)}\right)$, its likelihood is:

$$p(\sigma^{(k)} \mid \boldsymbol{\theta}) = \prod_{i,j \in T^{(k)}, i \prec j} P(i \prec j)^{C_{i,j}}$$

where $T^{(k)}$ is the set of all pairs of items in $\sigma^{(k)}$. The log-likelihood is:

$$\log p(\sigma^{(k)} \mid \boldsymbol{\theta}) = \sum_{i,j \in T^{(k)}, i \prec j} C_{i,j} \cdot \left(\theta_i - \log(e^{\theta_i} + e^{\theta_j})\right)$$

Maximizing the total log-likelihood yields parameter estimates:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \sum_{k=1}^{K} \log P(\sigma^{(k)} | \boldsymbol{\theta})$$

The corresponding energy for an aggregate ranking σ is:

$$E(\sigma \mid \mathcal{S}) = E(\sigma \mid \hat{\boldsymbol{\theta}}) = -\log P(\sigma^{(k)} \mid \hat{\boldsymbol{\theta}})$$

3.4. Variant 3: Plackett–Luce (PL)

The Plackett-Luce model (Plackett, 1975; Luce et al., 1959) generates rankings sequentially, choosing one item at a time with probability proportional to an associated score. Each item $i \in \mathcal{X}$ has a parameter α_i. For a partial ranking $\sigma^{(k)} = [x_1, x_2, \ldots, x_{n_k}]$, where $n_k \leq m$ and $\{x_1, x_2, \ldots, x_{n_k}\} \subseteq \mathcal{X}$:

$$P(\sigma^{(k)} \mid \boldsymbol{\alpha}) = \prod_{t=1}^{n_k} \frac{e^{\alpha_{x_t}}}{\sum_{j=t}^{n_k} e^{\alpha_{x_j}}}$$

The log-likelihood is:

$$\log P(\sigma^{(k)} \mid \boldsymbol{\alpha}) = \sum_{t=1}^{n_k} \left(\alpha_{x_t} - \log \sum_{j=t}^{n_k} e^{\alpha_{x_j}} \right)$$

Parameters are estimated via maximum likelihood:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\alpha} \sum_{k=1}^{K} \log P(\sigma^{(k)} | \boldsymbol{\alpha})$$

For an aggregate ranking $\sigma = [y_1, y_2, \ldots, y_m]$, the energy is:

$$E(\sigma \mid \mathcal{S}) = E(\sigma \mid \hat{\boldsymbol{\alpha}}) = -\log P(\sigma^{(k)} \mid \hat{\boldsymbol{\alpha}})$$

3.5. Variant 4: BT-informed Mallow's Model

The Mallows model (Mallows, 1957) places probability mass around a central ranking σ_0:

$$P(\sigma) \propto \exp(-\theta \cdot d(\sigma, \sigma_0))$$

where $d(\cdot,\cdot)$ is a distance metric (e.g., Kendall's tau or Reverse Major Index (RMJ; Feng and Tang, 2022)). Dispersion θ controls how concentrated the distribution is. The challenge is inferring θ and σ_0 from a small number of heterogeneous partial rankings. Existing methods, for instance, BayesMallows (Vitelli et al., 2018; Sørensen et al., 2019), require richer data and are unstable when only 2-4 partial rankings are available—typical in mixed-pathology settings. In our preliminary experiments, estimated central rankings given a fixed set of partial rankings varied greatly

across runs (Kendall's W ≈ 0.1), while BT produced stable estimates (> 0.9). We decide to use the inference method from BT to provide an estimate of σ_0 and set θ manually, which we call *BT-informed Mallows model*. The energy for the BT-informed Mallows model is:

$$E(\sigma \mid \mathcal{S}) = \theta \cdot d(\sigma, \sigma_0)$$

where the dispersion θ provides a flexible knob for synthetic data generation: larger θ yields a more peaked distribution on the central ordering σ_0, and smaller θ yields a more uniform distribution over σ.

3.6. Generative and inference algorithms

JPM can be used for both generative and inferential purposes. With a defined JPM, we can define algorithms for generating aggregate rankings based on input partial rankings, which is crucial for creating synthetic data. We can also infer the aggregate rankings from a data set (Algorithm 1 in Appendix C). In terms of generating aggregate rankings, PL sampling is straightforward, as rankings are generated sequentially (See Algorithm 2). Other models (PP, BT, Mallows) do not admit direct sampling. We employ Metropolis-Hastings MCMC (Metropolis et al., 1953; Hastings, 1970), starting from a random permutation and proposing swaps between positions (See Algorithm 3). Note that MCMC can be applied to PL as well. Throughout the paper, we used Kendall's τ as the distance metric.

4. Theoretical analysis

We explored five key questions regarding the JPM framework:

Q1: Is JPM valid as an inference model?

Q2: Is JPM valid as a generative model?

Q3: Which JPM variants are suitable for synthetic data generation?

Q4: For a given dataset, should JPM be preferred over a single-disease model for inference (e.g., SA-EBM (Hao et al., 2025))?

Q5: If JPM is appropriate for inference, which variant should be used?

This section addresses the first three questions. Q4 and Q5 will be addressed in the Discussion.

4.1. Q1: Model calibration

JPM as an inference algorithm aims to correctly assign energies to different aggregate rankings. We define a "valid inference model" as one that assigns low energies to aggregate rankings with smaller distances from the ground truth aggregate ranking (σ_{gt}). Formally, for any two candidate rankings σ_A and σ_B, a well-behaved model should satisfy the condition:

$$E_{\text{JPM}}(\sigma_A) < E_{\text{JPM}}(\sigma_B) \Longrightarrow d(\sigma_A, \sigma_{\text{gt}}) < d(\sigma_B, \sigma_{\text{gt}})$$

We refer to the degree to which a model satisfies this property as its **calibration**. We used the 4,050 aggregate rankings generated in Sec. 5.1 to study model calibration. We treat each aggregate ranking as an experiment. In each experiment, the given aggregate ranking served as the ground truth (σ_{gt}), and we also generated 1,000 random aggregate rankings based on the original associated input partial rankings which generated this σ_{gt}. As discussed in Sec. 3.6, we can use different JPM variants to generate an aggregate ranking from a set of partial rankings. Therefore, in each experiment, we tried each of the four inference variants: BT, PL, PP and Mallows ($\theta = 1$) when generating the 1,000 random aggregate rankings.

We measured *calibration* by calculating the Spearman's ρ correlation between (1) the energies of 1,000 randomly generated aggregate rankings, and (2) the average Kendall's τ distance between those random aggregate rankings and σ_{gt}.

Table 2: Calibration of JPM variants (mean with 95% CI). See Fig. 1 in Appendix E for details.

Generative variant	BT	Mallows ($\theta = 1$)	PL	PP
BT	0.802 ± 0.009	0.977 ± 0.001	0.605 ± 0.007	0.884 ± 0.003
Mallows ($\theta = 1$)	0.783 ± 0.010	0.954 ± 0.002	0.590 ± 0.007	0.866 ± 0.003
Mallows ($\theta = 10$)	0.780 ± 0.010	0.950 ± 0.002	0.579 ± 0.007	0.862 ± 0.003
PL	0.849 ± 0.007	0.913 ± 0.004	0.660 ± 0.006	0.840 ± 0.005
PP	0.791 ± 0.009	0.944 ± 0.002	0.590 ± 0.007	0.913 ± 0.002

As shown in Table 2, all JPM variants achieve a *calibration* greater than 0.58, indicating a positive correlation. The Mallows variant scores the highest, which is expected as its energy function is explicitly defined by Kendall's τ. **In sum, Table 2 indicates that JPM variants may all be valid inference algorithms**, because they can theoretically identify rankings closer to the ground truth.

While the *calibration* results in Table 2 are promising, JPM does not guarantee the accuracy of the fi-

nal inferred result, as the search process during inference may not explore the entire space of possible rankings. Consider this concrete example: JPM explores permutations that are all far from the ground truth. However, JPM correctly finds that permutations that are further away from the ground truth have larger distances, i.e., good *calibration* scores. This way, valid inference algorithms could be created from a calibrated JPM, but a specific inference algorithm may still fail to find good results.

4.2. Q2: Characterizing the Generation of Aggregate Rankings

In Q1 experiments, we generated aggregate rankings based on partial rankings, using JPM as a generative framework. We can conclude that in the real-world, if the generative process can be accurately represented by a JPM, then JPM may be a valid inference model (because of the good *calibration* scores). This leads to the next question: **How well can JPM's generative process emulates the formation of real-world joint disease progressions?** We analyze this from two perspectives: **separation** and **sharpness**.

Separation assesses whether a model's energy landscape clearly distinguishes the true aggregate ranking (σ_{gt}) from random noise. In other words, the ground truth aggregate ranking must be more consistent with the evidence from the input partial rankings ($\mathcal{S}$) than a random ranking is. Formally:

$$E(\sigma_{\text{gt}} \mid \mathcal{S}) \ll E(\sigma_{\text{random}} \mid \mathcal{S})$$

We define *separation* using the Area Under the Receiver Operating Characteristic Curve (AUROC), which measures the probability that a ranking sampled from the JPM has a lower energy than a randomly generated one:

$$\text{separation} = P(E_{\text{JPM}} < E_{\text{rand}}) + \frac{1}{2}P(E_{\text{JPM}} = E_{\text{rand}}),$$

where E_{JPM} represents the energies of rankings sampled by JPM, and E_{rand} represents the energies of random permutations of the biomarkers in $\mathcal{X}$.

Sharpness addresses a related question: Given a set of input partial rankings ($\mathcal{S}$), how stable is the generative output? A model with high *sharpness* will consistently produce similar aggregate rankings across repeated runs. We define *sharpness* as the Kendall's W of the sampled aggregate rankings.

We analyzed *separation* and *sharpness* also using the 4,050 aggregate rankings generated in Sec. 5.1, but in a different way than we did in the study of *calibration*. To study *separation*, for each of the five generative variants (as in Table 3), in each of the 4,050 "experiments", we generated 1,000 aggregate rankings using the variant, and 1,000 random permutations of the biomarkers in $\mathcal{X}$.

As shown in Table 3, all JPM models exhibit high *separation*, but they differ significantly in their *sharpness*. Notably, the Mallows variant's *sharpness* approaches 1.0 as its parameter θ increases. This result carries two important implications. First, JPM is best suited for generative tasks where the underlying progression is meaningfully different from a random ordering (i.e., high *separation*). Second, assuming the high-separation condition is met, the JPM framework offers the flexibility to generate synthetic data with a wide range of *sharpness*, allowing it to simulate diverse real-world scenarios.

Table 3: Mean separation and sharpness by data generation framework (mean ± 95% CI). See Fig. 2 in Appendix E for details.

Generative variant	Separation	Sharpness
BT	1.00 ± 0.00	0.96 ± 0.00
Mallows ($\theta = 1$)	1.00 ± 0.00	0.47 ± 0.00
Mallows ($\theta = 10$)	1.00 ± 0.00	0.61 ± 0.00
PL	1.00 ± 0.00	0.74 ± 0.01
Pairwise	1.00 ± 0.00	0.86 ± 0.00

4.3. Q3: Which Generative Variant to Choose?

Our analysis of the simulation data reveals that for every variant except the PP model, the *sharpness* of the generated output is highly predictable from the characteristics of the input partial rankings (see Table 4 in Appendix E). We identify four key predictive characteristics: the number of partial rankings (K), the average length of the partial rankings ($\bar{\ell}$), and two measures we call **conflict** and **overlap**.

We define *conflict* as the average pairwise normalized Kendall's τ distance (See Algorithm 4 in Appendix D) among the input partial rankings, restricted to their common items:

$$\text{Conflict} = \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} d_\tau\big(\sigma^{(i)}, \sigma^{(j)}\big)$$

We define *overlap* as the proportion of biomarkers that appear in at least two partial rankings relative to the total number of unique biomarkers, m:

$$\text{Overlap} = \frac{\left| \bigcup_{1 \le i < j \le K} \left(\sigma^{(i)} \cap \sigma^{(j)} \right) \right|}{m}.$$

This provides a clear guide for model selection. A user can first determine a target *sharpness* that suits their research needs. Then, by calculating these four characteristics from their input data and plugging them into regression models (as illustrated in Table 4), they can select the JPM variant that will produce an output with the desired sharpness. For a detailed breakdown of how each characteristic influences *sharpness*, see Figures 3, 4, 5, 6 and 7 in Appendix E.

5. Synthetic experiments & results

To evaluate the JPM framework, we designed a series of experiments using synthetic data. The pipeline for these experiments involved two main stages: (1) generating realistic synthetic datasets under various conditions, and (2) applying and evaluating both the JPM and the SA-EBM (for comparison, Hao et al., 2025) on these datasets.

5.1. Synthetic data generation

Our synthetic data generation process was designed to simulate plausible clinical research scenarios. We selected 18 biomarkers from Alzheimer's Disease Neuroimaging Initiative (ADNI, Mueller et al., 2005): 12 are commonly used in AD progression modeling studies (Young et al., 2014; Archetti et al., 2019), and an additional 6 to ensure pathological diversity. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). In Appendix F, we detailed the list of these 18 biomarkers and how we obtained their distribution parameters.

To generate one synthetic dataset, we first generated 2-4 randomized partial rankings, each with 6-12 biomarkers randomly chosen from the 18 biomarkers mentioned above. These ranges were chosen to reflect the practical constraints observed in the real-world NACC data. Based on this set of partial rankings, we used JPM variants (BT, PL, PP, and Mallows with $\theta = 1, 10$) to generate a ground truth aggregate ranking. If MCMC was used (Algorithm 3), we ran it for 500 iterations. We then generated the final biomarker data using two different generative models: (1) Event-Based Model (EBM; Fonteijn et al., 2012); (2) Sigmoid Model (See details in Appendix H), which was inspired by a hypothetical AD cascade model (Jack et al., 2010) and adapted from prior work (Venkatraghavan et al., 2019; Young et al., 2015).

We varied the following in data generation. (1) Participant size (J): 50, 100, or 200 participants. (2) Healthy ratio (R), i.e., the percent of control participants: 0.25, 0.5, or 0.75. (3) Experimental configurations (See Appendix G for the full list). These include variations in disease stage type (ordinal vs. continuous) and distributions (e.g., Dirichlet-Multinomial vs. Uniform), and biomarker data distributions (Normal vs. Non-Normal).

For each of the $3 \times 3 = 9$ combinations of participant size (J) and healthy ratio (R), we generated 10 random datasets. In total, considering 5 JPM variants (Mallows with two temperature settings) and 9 experimental conditions, this resulted in $5 \times 9 \times 90 = 4050$ datasets for the aggregate disease scenarios. In parallel, we generated corresponding datasets for each individual disease. The configuration for these datasets matched the aggregate one, but the participant size was set to be **four times larger**, a design choice also inspired by properties of the NACC data.

5.2. Experiment setup and evaluation

For each synthetic aggregate dataset, we first ran SA-EBM (Hao et al., 2025) on the corresponding individual disease datasets to estimate their respective partial rankings. These estimated rankings were then used to construct a JPM, whose result served as an informed prior (as in Algorithm 1) for a final SA-EBM run on the aggregate, i.e., mixed-pathology dataset. As a benchmark, we ran SA-EBM alone without the JPM-resultant prior. We chose SA-EBM because it has the state-of-the-art performance on reconstructing ordinal disease progressions from cross-sectional datasets (Hao et al., 2025). For all SA-EBM and

JPM inference methods, we used 20,000 iterations for MCMC.

We evaluated performance on two tasks. (1) Ordering: The agreement between the estimated ranking and the ground truth, measured by the normalized Kendall's τ distance (See Algorithm 4 in Appendix D). A lower value indicates better performance; (2) Staging: The accuracy of participant staging, measured by the Mean Absolute Error (MAE).

5.3. Results

The results are grouped by the model used to generate the synthetic data. If the Mallows model variant was used for inference, i.e., reconstructing the mixed-pathology progression and assigning participants disease stages, we used a fixed $\theta = 1$.

Data Generated by PP When data was generated using the PP model (See Figures 8 and 9 in Appendix I), JPM demonstrated a clear and significant advantage over the single-disease EBM. Error is reported as 95% confidence intervals throughout the paper. The JPM-EBM achieved the best ordering performance with Kendall's τ of 0.15 ± 0.01, outperforming SA-EBM (0.19 ± 0.01). The PL and BT priors also performed well (0.17 ± 0.01). This advantage was most pronounced for smaller sample sizes ($J = 50, 100$), where a prior is most impactful. In the staging task, JPM also outperformed SA-EBM (0.79 ± 0.07 MAE), with the PP achieving the best result (0.75 ± 0.07 MAE).

Data Generated by BT JPM also showed a significant advantage when the data were generated using the BT model. The BT, PL, and PP variants achieved the best ordering results (0.15 ± 0.01). The Mallows model performed comparably to SA-EBM (0.19 ± 0.01). Staging results were similar to those from the PP-generated data, with BT having the best performance. See Figures 10 and 11 in Appendix I.

Data Generated by PL When data were generated with the PL model, the performance advantage of JPM on the ordering task was smaller but still present. The PL variant achieved the best outcome ($\tau = 0.17 \pm 0.01$), slightly better than the standard EBM (0.20 ± 0.01). The margin of advantage in the staging task was similar to the above two experiments (0.07 MAE). See Figures 12 and 13 in Appendix I.

Data Generated by the Mallows Variant When data were generated using Mallows ($\theta = 1.0$), on the ordering task, the Mallows variant performed similarly to SA-EBM (0.19 ± 0.01), only slightly better than PL (0.20 ± 0.01) and PP (0.21 ± 0.01). The BT variant performed the worst in this condition (0.23 ± 0.01). A similar pattern was observed when data was generated with a higher concentration ($\theta = 10.0$). Staging results on Mallows-generated data reflected the patterns seen in the ordering task, i.e., SA-EBM and the Mallows model were the top performers. See Figures 14, 15, 16 and 17 in Appendix I for more details.

6. NACC experiment & results

We obtained data from the National Alzheimer's Coordinating Center (NACC) and the Alzheimer's Disease Sequencing Project (ADSP) (Mukherjee et al., 2023). We used Uniform Data Set (UDS) visits conducted between September 2005 and June 2025. In Appendix J, we detailed what data were used, how we processed the data, and the information about biomarkers that we included for AD, VaD and the mixed-pathology of AD and VaD.

The resulting dataset of AD contains 188 healthy participants and 59 AD patients. VaD dataset contains 3,732 healthy participants and 525 VaD patients. The mixed-pathology dataset contains 188 healthy participants and 37 patients with both AD and VaD.

To get the progression of AD, we ran SA-EBM (Hao et al., 2025) ten times; each time with a random seed. We obtained the final result with the seed associated with the highest log data likelihood. The VaD dataset is too large to run with multiple seeds. We ran with two seeds and obtained the result which seems more plausible. For mixed-pathology, we first ran with SA-EBM with ten random seeds and obtained the result using the highest-likelihood seed. We also tried four variants of the JPM. For the Mallows model, we used $\theta = 1$. For each variant, we also tried 10 random seeds and obtained the result using the seed associated with the highest data log likelihood. In all runs, we used 20,000 MCMC iterations with 500 burn in and no thinning. Detailed results, including the progressions and the trace plots, are presented in Appendix K.

The timelines of the baseline SA-EBM model and the JPM-Mallows model align better with the current understanding of how vascular disease interacts with Alzheimer's pathology. They positioned vascular damage (WMHnorm) as an event that occurred

after the initial molecular pathology was present and cognitive symptoms were just beginning, but before the cascade of major, widespread brain volume loss. This sequence supports a synergistic role for vascular disease, where it exacerbates the ongoing Alzheimer's pathology, accelerating the subsequent neurodegeneration and brain atrophy (Attems and Jellinger, 2014; Snyder et al., 2015; Lee et al., 2016). Between the two results, the one by JPM-Mallows seemed a little better because decades of research have shown that the buildup of amyloid and tau proteins begins years, or even decades, before significant cognitive symptoms or brain atrophy are detectable (Jack et al., 2010; Sperling et al., 2011; Bateman et al., 2012).

7. Discussion

In this paper, we presented a framework to model the joint progression of multiple diseases. We explained how partial rankings from individual disease progressions can be used to construct a prior distribution for the aggregate rankings.

We analyzed some theoretical properties of different JPM models. *Calibration* was positive for all variants, with Mallows performing best, reflecting its distance-based energy. *Separation* was near-perfect across the board, indicating that sampled rankings fall well below random permutations in energy. *Sharpness* differed by variant and can be tuned (e.g., Mallows via θ); importantly, *sharpness* is strongly predicted by four observable properties of the input partial rankings—number of rankings, average length, *conflict*, and *overlap*—enabling practical variant selection before data generation or inference.

In synthetic benchmarks, JPM delivered 21% improvements in ordering accuracy over SA-EBM (Hao et al., 2025) which treats the joint disease as a single trajectory, highlighting the advantage of incorporating multi-disease structure. In NACC, we compared all results and concluded that the results by the baseline and the Mallows variant best reflect what past literature indicated.

More specifically, the progression of "AD only", obtained using SA-EBM, is similar to the finding by Hao et al. (2025) which was based on the ADNI dataset: CSF biomarkers first, followed by cognitive decline and atrophy in the brain. There are differences in the two results. For example, in this study, cognitive decline seems to occur earlier than CSF biomarker changes. These differences do not alter the overall trend, though.

Reflecting on our results, should JPM be preferred over a single-disease model (Q4) and which variant of JPM should be used for inference (Q5)? The answer is clear based on our synthetic experimental results: if data is generated by a model that has high sharpness, i.e., ground truth of the aggregate ranking is stable in repeated attempts with the same set of partial rankings, then we can expect a larger advantage of JPM over a single-disease EBM inference model. Even if such an assumption does not hold, JPM's results are not too far away from EBM, and are still reliable. Based on our synthetic experimental results, it is desirable if the inference variant matches the variant that underlies the generative process of the aggregate rankings. What this infers is that we can first estimate the *sharpness* (refer to Sec. 4.3) of the underlying generative model and then choose the JPM variant with a similar *sharpness* for inferential tasks.

The NACC results show that the Mallows model and the SA-EBM were the top performers. This indicates that the underlying generative process of the real-world AD and VaD mixed pathology might not have a high *sharpness*. Reconciling the synthetic results and real-world application is an important direction for future work.

Although our results demonstrate that the approach has promise, there are some limitations that should be explored in future work. Firstly, we assume all patients share the same joint disease progression. Future work should explore how to incorporate subtypes such as in SuStaIn (Young et al., 2018) into the JPM framework. Secondly, JPM's performance struggles when event sequences are continuous. Integrating the approach taken by the Temporal EBM (Wijeratne et al., 2023) could help with this issue. Lastly, some disease interactions might result in a disease progression that is completely different from the individual disease progressions. In these cases, the JPM would be inappropriate to apply. Domain knowledge of the underlying mechanisms for how the individual diseases will interact is essential for knowing when JPM is appropriate to apply and which variant to choose.

Acknowledgments

We thank the CHTC at the University of Wisconsin-Madison for computing support. JLA was funded by the Japan Probabilistic Computing Consortium Association.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The ADSP Phenotype Harmonization Consortium (ADSP-PHC) is funded by NIA (U24 AG074855, U01 AG068057 and R01 AG059716).

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI David Holtzman, MD), P30 AG066518 (PI Lisa Silbert, MD, MCR), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI Julie A. Schneider, MD, MS), P30 AG072978 (PI Ann McKee, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Jessica Langbaum, PhD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Glenn Smith, PhD, ABPP), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P30 AG086401 (PI Erik Roberson, MD, PhD), P30 AG086404 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

The NACC database is funded by NIA/NIH Grant U24 AG072122. SCAN is a multi-institutional project that was funded as a U24 grant (AG067418) by the National Institute on Aging in May 2020. Data collected by SCAN and shared by NACC are contributed by the NIA-funded ADRCs as follows:

Arizona Alzheimer's Center - P30 AG072980 (PI: Eric Reiman, MD); R01 AG069453 (PI: Eric Reiman (contact), MD); P30 AG019610 (PI: Eric Reiman, MD); and the State of Arizona which provided additional funding supporting our center; Boston University - P30 AG013846 (PI Neil Kowall MD); Cleveland ADRC - P30 AG062428 (James Leverenz, MD); Cleveland Clinic, Las Vegas – P20AG068053; Columbia - P50 AG008702 (PI Scott Small MD); Duke/UNC ADRC – P30 AG072958; Emory University - P30AG066511 (PI Levey Allan, MD, PhD); Indiana University - R01 AG19771 (PI Andrew Saykin, PsyD); P30 AG10133 (PI Andrew Saykin, PsyD); P30 AG072976 (PI Andrew Saykin, PsyD); R01 AG061788 (PI Shannon Risacher, PhD); R01 AG053993 (PI Yu-Chien Wu, MD, PhD); U01 AG057195 (PI Liana Apostolova, MD); U19 AG063911 (PI Bradley Boeve, MD); and the Indiana University Department of Radiology and Imaging Sciences; Johns Hopkins - P30 AG066507 (PI Marilyn Albert, Phd.); Mayo Clinic - P50 AG016574 (PI Ronald Petersen MD PhD); Mount Sinai - P30 AG066514 (PI Mary Sano, PhD); R01 AG054110 (PI Trey Hedden, PhD); R01 AG053509 (PI Trey Hedden, PhD); New York University - P30AG066512-

01S2 (PI Thomas Wisniewski, MD); R01AG056031 (PI Ricardo Osorio, MD); R01AG056531 (PIs Ricardo Osorio, MD; Girardin Jean-Louis, PhD); Northwestern University - P30 AG013854 (PI Robert Vassar PhD); R01 AG045571 (PI Emily Rogalski, PhD); R56 AG045571, (PI Emily Rogalski, PhD); R01 AG067781, (PI Emily Rogalski, PhD); U19 AG073153, (PI Emily Rogalski, PhD); R01 DC008552, (M.-Marsel Mesulam, MD); R01 AG077444, (PIs M.-Marsel Mesulam, MD, Emily Rogalski, PhD); R01 NS075075 (PI Emily Rogalski, PhD); R01 AG056258 (PI Emily Rogalski, PhD); Oregon Health & Science University - P30 AG066518 (PI Lisa Silbert, MD, MCR); Rush University - P30 AG010161 (PI David Bennett MD); Stanford – P30AG066515; P50 AG047366 (PI Victor Henderson MD MS); University of Alabama, Birmingham – P20; University of California, Davis - P30 AG10129 (PI Charles DeCarli, MD); P30 AG072972 (PI Charles DeCarli, MD); University of California, Irvine - P50 AG016573 (PI Frank LaFerla PhD); University of California, San Diego - P30AG062429 (PI James Brewer, MD, PhD); University of California, San Francisco - P30 AG062422 (Rabinovici, Gil D., MD); University of Kansas - P30 AG035982 (Russell Swerdlow, MD); University of Kentucky - P30 AG028283-15S1 (PIs Linda Van Eldik, PhD and Brian Gold, PhD); University of Michigan ADRC - P30AG053760 (PI Henry Paulson, MD, PhD) P30AG072931 (PI Henry Paulson, MD, PhD) Cure Alzheimer's Fund 200775 - (PI Henry Paulson, MD, PhD) U19 NS120384 (PI Charles DeCarli, MD, University of Michigan Site PI Henry Paulson, MD, PhD) R01 AG068338 (MPI Bruno Giordani, PhD, Carol Persad, PhD, Yi Murphey, PhD) S10OD026738-01 (PI Douglas Noll, PhD) R01 AG058724 (PI Benjamin Hampstead, PhD) R35 AG072262 (PI Benjamin Hampstead, PhD) W81XWH2110743 (PI Benjamin Hampstead, PhD) R01 AG073235 (PI Nancy Chiaravalloti, University of Michigan Site PI Benjamin Hampstead, PhD) 1I01RX001534 (PI Benjamin Hampstead, PhD) IRX001381 (PI Benjamin Hampstead, PhD); University of New Mexico - P20 AG068077 (Gary Rosenberg, MD); University of Pennsylvania - State of PA project 2019NF4100087335 (PI David Wolk, MD); Rooney Family Research Fund (PI David Wolk, MD); R01 AG055005 (PI David Wolk, MD); University of Pittsburgh - P50 AG005133 (PI Oscar Lopez MD); University of Southern California - P50 AG005142 (PI Helena Chui MD); University of Washington - P50 AG005136 (PI Thomas Grabowski MD); University of Wisconsin - P50 AG033514 (PI Sanjay Asthana MD FRCP); Vanderbilt University – P20 AG068082; Wake Forest - P30AG072947 (PI Suzanne Craft, PhD); Washington University, St. Louis - P01 AG03991 (PI John Morris MD); P01 AG026276 (PI John Morris MD); P20 MH071616 (PI Dan Marcus); P30 AG066444 (PI John Morris MD); P30 NS098577 (PI Dan Marcus); R01 AG021910 (PI Randy Buckner); R01 AG043434 (PI Catherine Roe); R01 EB009352 (PI Dan Marcus); UL1 TR000448 (PI Brad Evanoff); U24 RR021382 (PI Bruce Rosen); Avid Radiopharmaceuticals / Eli Lilly; Yale - P50 AG047270 (PI Stephen Strittmatter MD PhD); R01AG052560 (MPI: Christopher van Dyck, MD; Richard Carson, PhD); R01AG062276 (PI: Christopher van Dyck, MD); 1Florida - P30AG066506-03 (PI Glenn Smith, PhD); P50 AG047266 (PI Todd Golde MD PhD)

References

D. Archetti, S. Ingala, V. Venkatraghavan, V. Wottschel, A. L. Young, F. Barkhof, D. C. Alexander, N. P. Oxtoby, G. B. Frisoni, A. Redolfi, for the Alzheimer's Disease Neuroimaging Initiative, and for EuroPOND Consortium. Multi-study validation of data-driven disease progression models to characterize evolution of biomarkers in alzheimer's disease. *NeuroImage: Clinical*, 24: 101954, 2019.

Johannes Attems and Kurt A Jellinger. The overlap between vascular disease and alzheimer's disease-lessons from pathology. *BMC medicine*, 12(1):206, 2014.

Randall J Bateman, Chengjie Xiong, Tammie LS Benzinger, Anne M Fagan, Alison Goate, Nick C Fox, Daniel S Marcus, Nigel J Cairns, Xianyun Xie, Tyler M Blazey, et al. Clinical and biomarker changes in dominantly inherited alzheimer's disease. *New England Journal of Medicine*, 367(9): 795–804, 2012.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Aron S Buchman, Lei Yu, Shahram Oveisgharan, Jose M Farfel, Julie A Schneider, and David A Bennett. Person-specific contributions of brain

pathologies to progressive parkinsonism in older adults. *The Journals of Gerontology: Series A*, 76(4):615–621, 2021.

Yaping Chu, Warren D Hirst, and Jeffrey H Kordower. Mixed pathology as a rule, not exception: Time to reconsider disease nosology. *Handbook of Clinical Neurology*, 192:57–71, 2023.

Yifan Feng and Yuxuan Tang. On a mallows-type model for (ranked) choices. *Advances in Neural Information Processing Systems*, 35:3052–3065, 2022.

Nicholas C Firth, Silvia Primativo, Emilie Brotherhood, Alexandra L Young, Keir XX Yong, Sebastian J Crutch, Daniel C Alexander, and Neil P Oxtoby. Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression. *Alzheimer's & Dementia*, 16(7): 965–973, 2020.

Hubert M Fonteijn, Marc Modat, Matthew J Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z Hobbs, Rachael I Scahill, Sarah J Tabrizi, Sebastien Ourselin, Nick C Fox, et al. An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012.

Shelley L Forrest and Gabor G Kovacs. Current concepts of mixed pathologies in neurodegenerative diseases. *Canadian Journal of Neurological Sciences*, 50(3):329–345, 2023.

Isobel Claire Gormley and Thomas Brendan Murphy. A grade of membership model for rank data. 2009.

Hongtao Hao and L. Austerweil, Joseph. Bayesian event-based model for disease subtype and stage inference. In *Proceedings of the Machine Learning for Health Conference 2025*, 2025.

Hongtao Hao, Vivek Prabhakaran, Veena A Nair, Nagesh Adluru, and Joseph L. Austerweil. Stage-aware event-based modeling (SA-EBM) for disease progression. In Monica Agrawal, Kaivalya Deshpande, Matthew Engelhard, Shalmali Joshi, Shengpu Tang, and Iñigo Urteaga, editors, *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298 of *Proceedings of Machine Learning Research*. PMLR, 15–16 Aug 2025. URL https://proceedings.mlr.press/v298/hao25a.html.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

Kurt A Jellinger. Pathobiological subtypes of alzheimer disease. *Dementia and Geriatric Cognitive Disorders*, 49(4):321–333, 2021.

Isabella Friis Jørgensen, Alejandro Aguayo-Orozco, Mette Lademann, and Søren Brunak. Age-stratified longitudinal study of alzheimer's and vascular dementia patients. *Alzheimer's & Dementia*, 16(6):908–917, 2020.

Garam Lee, Kwangsik Nho, Byungkon Kang, Kyung-Ah Sohn, and Dokyoon Kim. Predicting alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, 9(1): 1952, 2019.

Seonjoo Lee, Fawad Viqar, Molly E Zimmerman, Atul Narkhede, Giuseppe Tosto, Tammie LS Benzinger, Daniel S Marcus, Anne M Fagan, Alison Goate, Nick C Fox, et al. White matter hyperintensities are a core feature of alzheimer's disease: evidence from the dominantly inherited alzheimer network. *Annals of neurology*, 79(6):929–939, 2016.

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Basil Maag, Stefan Feuerriegel, Mathias Kraus, Maytal Saar-Tsechansky, and Thomas Züger. Modeling longitudinal dynamics of comorbidities. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 222–235, 2021.

Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

Shubhabrata Mukherjee, Seo-Eun Choi, Michael L Lee, Phoebe Scollard, Emily H Trittschuh, Jesse Mez, Andrew J Saykin, Laura E Gibbons, R Elizabeth Sanders, Andrew F Zaman, et al. Cognitive domain harmonization and cocalibration in studies of older adults. *Neuropsychology*, 37(4):409, 2023.

Z. Oflaz, C. Yozgatligil, and A. S. Selcuk-Kestel. Modeling comorbidity of chronic diseases using coupled hidden Markov model with bivariate discrete copula. *Statistical Methods in Medical Research*, 32(4):829–849, 2023.

Neil P Oxtoby, Louise-Ann Leyland, Leon M Aksman, George EC Thomas, Emma L Bunting, Peter A Wijeratne, Alexandra L Young, Angelika Zarkali, Manuela MX Tan, Fion D Bremner, et al. Sequence of clinical and neurodegeneration events in parkinson's disease progression. *Brain*, 144(3): 975–988, 2021.

Sebastian Palmqvist, Henrik Zetterberg, Niklas Mattsson, Per Johansson, Alzheimer's Disease Neuroimaging Initiative, Lennart Minthon, Kaj Blennow, Mattias Olsson, Swedish BioFINDER Study Group, Oskar Hansson, et al. Detailed comparison of amyloid pet and csf biomarkers for identifying early alzheimer disease. *Neurology*, 85(14):1240–1249, 2015.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Tanav Popli, Subhamoy Pal, Allyson Gregoire, Jonathan M Reader, Alexis Passamani, Kelly M Bakulski, Arijit Bhaumik, Emile Pinarbasi, Henry Paulson, and Amanda Cook Maher. High rates of diagnostic discordance and co-pathology: Insights into psp from the nacc dataset. *Alzheimer's & Dementia*, 21(5):e70248, 2025.

Jasmin Rahimi and Gabor G Kovacs. Prevalence of mixed pathologies in the aging brain. *Alzheimer's research & therapy*, 6(9):82, 2014.

Julie A Schneider, Zoe Arvanitakis, Sue E Leurgans, and David A Bennett. The neuropathology of probable alzheimer disease and mild cognitive impairment. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 66(2):200–208, 2009.

Heather M Snyder, Roderick A Corriveau, Suzanne Craft, James E Faber, Steven M Greenberg, David Knopman, Bruce T Lamb, Thomas J Montine, Maiken Nedergaard, Chris B Schaffer, et al. Vascular contributions to cognitive impairment and dementia including alzheimer's disease. *Alzheimer's & Dementia*, 11(6):710–717, 2015.

Øystein Sørensen, Marta Crispino, Qinghua Liu, and Valeria Vitelli. Bayesmallows: an r package for the bayesian mallows model. *arXiv preprint arXiv:1902.08432*, 2019.

Reisa A Sperling, Paul S Aisen, Laurel A Beckett, David A Bennett, Suzanne Craft, Anne M Fagan, Takeshi Iwatsubo, Clifford R Jack Jr, Jeffrey Kaye, Thomas J Montine, et al. Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia*, 7(3):280–292, 2011.

Lucy E Stirland, Radmila Choate, Preeti Pushpalata Zanwar, Panpan Zhang, Tamlyn J Watermeyer, Martina Valletta, Mario Torso, Stefano Tamburin, Usman Saeed, Gerard R Ridgway, et al. Multimorbidity in dementia: Current perspectives and future challenges. *Alzheimer's & Dementia*, 21(8): e70546, 2025.

Raghav Tandon, Anna Kirkpatrick, and Cassie S Mitchell. sEBM: Scaling event based models to predict disease progression via implicit biomarker selection and clustering. In *International Conference on Information Processing in Medical Imaging*, pages 208–221. Springer, 2023.

Wenpin Tang. Mallows ranking models: maximum likelihood estimate and regeneration. In *International Conference on Machine Learning*, pages 6125–6134. PMLR, 2019.

Vikram Venkatraghavan, Esther E Bron, Wiro J Niessen, Stefan Klein, Alzheimer's Disease Neuroimaging Initiative, et al. Disease progression timeline estimation for Alzheimer's disease using

discriminative event based modeling. *NeuroImage*, 186:518–532, 2019.

Valeria Vitelli, Øystein Sørensen, Marta Crispino, Arnoldo Frigessi, and Elja Arjas. Probabilistic preference learning with the mallows rank model. *Journal of Machine Learning Research*, 18(158):1–49, 2018.

Peter Wijeratne and Daniel Alexander. Unscrambling disease progression at scale: fast inference of event permutations with optimal transport. *Advances in Neural Information Processing Systems*, 37:63316–63341, 2024.

Peter A Wijeratne, Arman Eshaghi, William J Scotton, Maitrei Kohli, Leon Aksman, Neil P Oxtoby, Dorian Pustina, John H Warner, Jane S Paulsen, Rachael I Scahill, et al. The temporal event-based model: Learning event timelines in progressive diseases. *Imaging Neuroscience*, 1:1–19, 2023.

Liuqing Yang, Xifeng Wang, Qi Guo, Scott Gladstein, Dustin Wooten, Tengfei Li, Weining Z Robieson, Yan Sun, Xin Huang, and Alzheimer's Disease Neuroimaging Initiative. Deep learning based multimodal progression modeling for alzheimer's disease. *Statistics in Biopharmaceutical Research*, 13(3):337–343, 2021.

Alexandra L Young, Neil P Oxtoby, Pankaj Daga, David M Cash, Nick C Fox, Sebastien Ourselin, Jonathan M Schott, and Daniel C Alexander. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, 137(9):2564–2577, 2014.

Alexandra L Young, Neil P Oxtoby, Sebastien Ourselin, Jonathan M Schott, Daniel C Alexander, Alzheimer's Disease Neuroimaging Initiative, et al. A simulation system for biomarker evolution in neurodegenerative disease. *Medical image analysis*, 26(1):47–56, 2015.

Alexandra L Young, Razvan V Marinescu, Neil P Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C Firth, David M Cash, David L Thomas, Katrina M Dick, Jorge Cardoso, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nature communications*, 9(1):4273, 2018.

Alexandra L Young, Neil P Oxtoby, Sara Garbarino, Nick C Fox, Frederik Barkhof, Jonathan M Schott, and Daniel C Alexander. Data-driven modelling of neurodegenerative disease progression: thinking outside the black box. *Nature Reviews Neuroscience*, 25(2):111–130, 2024.

Appendix A. ADNI Information

A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Appendix B. Biomarkers Involved in each disease

- AD: ABETA, PTAU, TAU, Entorhinal, Hippocampus, Fusiform, MidTemp, WholeBrain, FDG, ADAS13, CDRSB, LDELTOTAL, RAVLT, MMSE, MOCA, TRABSCOR, FAQ
- Vascular dementia: FDG, WholeBrain, TRABSCOR, MMSE, MOCA, FAQ, ADAS13, CDRSB
- LBD: FDG, Hippocampus, Entorhinal, ABETA, TAU, PTAU, MMSE, MOCA, TRABSCOR, FAQ

Appendix C. Sampling Algorithms

Algorithm 1: JPM Inference Algorithm

Input: Individual diseases $\mathcal{D} = \{D^{(1)}, \ldots, D^{(K)}\}$, mixed-pathology data D_{m-p}, energy function $E(\sigma)$, data likelihood function $P(D_{m-p} \mid \sigma)$
Output: Best aggregate ranking $\sigma^\star$

```
// Step 0: Estimate partial rankings from individual diseases
```

Estimate partial rankings $\mathcal{S}$ from $\mathcal{D}$;

```
// Init
```

$\sigma \leftarrow$ random permutation of $\mathcal{X}$;
$\ell \leftarrow -\infty$;
$\sigma^\star \leftarrow \sigma$, $\ell^\star \leftarrow \ell$;
for $t = 1$ **to** T **do**

```
// Step 1: Proposal
```

Propose $\sigma' \leftarrow$ swap two random items in σ;

```
// Step 2: Compute data likelihood
```

$\ell' \leftarrow \log P(D_c \mid \sigma') - E(\sigma')$;

```
// Step 3: Acceptance probability
```

$p \leftarrow \min(1, \exp(\ell' - \ell))$;
Draw $u \sim \text{Uniform}(0, 1)$;

```
// Step 4: Accept/Reject
```

if $u < p$ **then**
$\sigma \leftarrow \sigma'$, $\ell \leftarrow \ell'$;
end

```
// Step 5: Track best result
```

if $\ell > \ell^\star$ **then**
$\sigma^\star \leftarrow \sigma$, $\ell^\star \leftarrow \ell$;
end
end
return $\sigma^\star$;

Explanations:

- The energy function is one of the required inputs. This means we need to decide on the JPM variant.

- We use EBM as the data likelihood function, i.e., $P(D_{m-p} \mid \sigma)$ and $\log P(D_{m-p} \mid \sigma)$. More specifically, we used SA-EBM (Hao et al., 2025). Note that EBM is just one choice for the likelihood function. Future work can explore other possible likelihoods.
- In the estimation of partial rankings from individual diseases' data, we also used SA-EBM (Hao et al., 2025), which has state-of-the-art performance on estimating ordinal disease progressions from cross-sectional datasets.

Algorithm 2: Plackett–Luce Sampling

Input: Parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$
Output: Aggregate ranking $\sigma = [i_1, i_2, \ldots, i_m]$
for $t = 1$ **to** m **do**
  `// Sample next item proportional to its weight`
  `// 𝒳 is the list of biomarkers in the aggregate ranking, whose length is m`
  Sample i_t with probability $\frac{e^{\alpha_i}}{\sum_{q \in \mathcal{X}} e^{\alpha_q}}$;
  `// Remove item from available set`
  $U \leftarrow \mathcal{X} \setminus \{i_t\}$;
end
return $\sigma = [i_1, i_2, \ldots, i_m]$;

Algorithm 3: JMP Generative Algorithm (Metropolis–Hastings MCMC)

Input: Set $\mathcal{X}$, Energy function $E(\sigma)$
Output: Best aggregate ranking $\sigma^\star$
`// Init`
$\sigma \leftarrow$ random permutation of $\mathcal{X}$;
$\ell \leftarrow E(\sigma)$;
$\sigma^\star \leftarrow \sigma$, $\ell^\star \leftarrow \ell$;
for $t = 1$ **to** T **do**
  `// Step 1: Proposal`
  Propose $\sigma' \leftarrow$ swap two random items in σ;
  `// Step 2: Compute energy`
  $\ell' \leftarrow E(\sigma')$;
  `// Step 3: Metropolis-Hastings acceptance probability`
  $p \leftarrow \min(1, \exp(\ell - \ell'))$;
  Draw $u \sim \text{Uniform}(0, 1)$;
  `// Step 4: Accept/Reject`
  if $u < p$ **then**
    $\sigma \leftarrow \sigma'$, $\ell \leftarrow \ell'$;
  end
  `// Step 5: Track minimum energy state`
  if $\ell < \ell^\star$ **then**
    $\sigma^\star \leftarrow \sigma$, $\ell^\star \leftarrow \ell$;
  end
end
return $\sigma^\star$;

Appendix D. Normalized Kendall's τ distance

Algorithm 4: Normalized Kendall's Tau Distance

Input: Two rankings r_1, r_2 of length n. The two rankings are two arrays of indices of the same set of items.

Output: Normalized Kendall's tau distance $d \in [0, 1]$

```
concordant ← 0, discordant ← 0;
for p = 1 to n − 1 do
    for q = p + 1 to n do
        // Compare relative order of items p and q in both rankings
        if (r1[p] − r1[q]) · (r2[p] − r2[q]) > 0 then
            concordant ← concordant + 1;
        end
        if (r1[p] − r1[q]) · (r2[p] − r2[q]) < 0 then
            discordant ← discordant + 1;
        end
    end
end
total ← concordant + discordant;
if total = 0 then
    return 0;
end
else
    return discordant / total;
end
```

Appendix E. Model Analysis

Table 4: Regression Models Predicting Model Sharpness

	PL (1)	BT (2)	Pairwise (3)	Mallows $\theta = 1.0$ (4)	Mallows $\theta = 10.0$ (5)
const	1.142***	1.013***	0.942***	1.082***	1.240***
	(0.015)	(0.002)	(0.010)	(0.005)	(0.005)
n_{pr}	-0.003	-0.000	0.001	-0.113***	-0.121***
	(0.002)	(0.000)	(0.001)	(0.001)	(0.001)
mean_len	0.004***	0.001***	0.002**	-0.047***	-0.048***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
conflict	-0.257***	-0.030***	-0.116***	-0.001	0.002
	(0.008)	(0.001)	(0.006)	(0.002)	(0.003)
overlap_rate	-0.567***	-0.073***	-0.095***	0.283***	0.292***
	(0.008)	(0.001)	(0.006)	(0.005)	(0.005)
R^2	0.651	0.630	0.193	0.858	0.852
Adj. R^2	0.650	0.629	0.192	0.858	0.851
N	3240	3240	3240	3240	3240

Standard errors in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$

In Table 4, n_{pr} refers to the number of partial rankings, and mean_len indicates the average number of partial rankings.

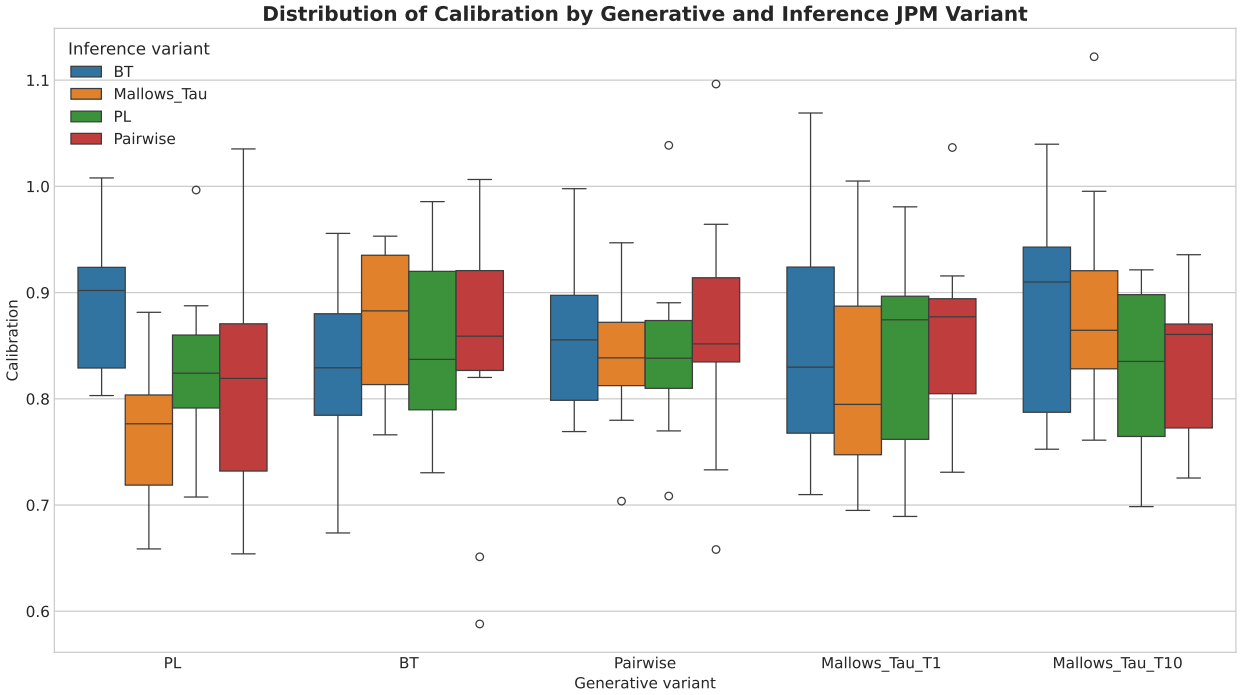

Figure 1: Distribution of calibration by generative and inference JPM variant

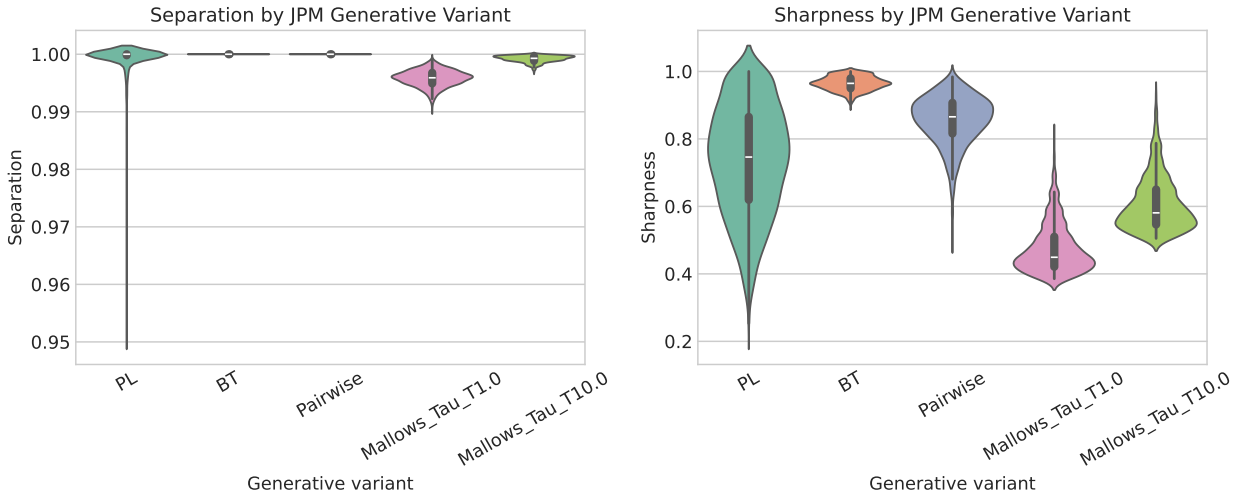

Figure 2: Separation and sharpness of different JPM variants

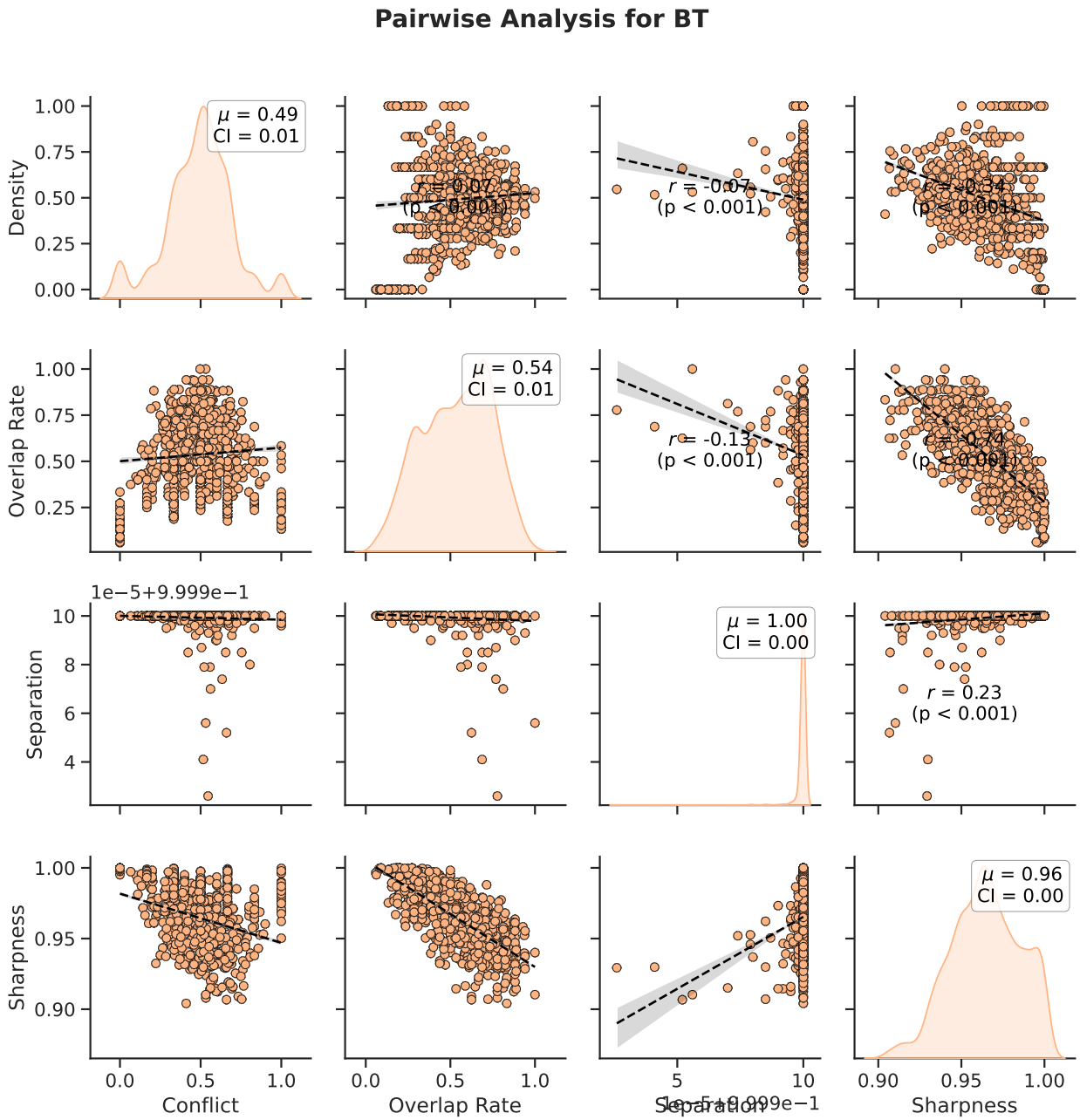

Figure 3: Correlation between sharpness and input partial ranking characteristics when the aggregate rankings are generated with the BT variant

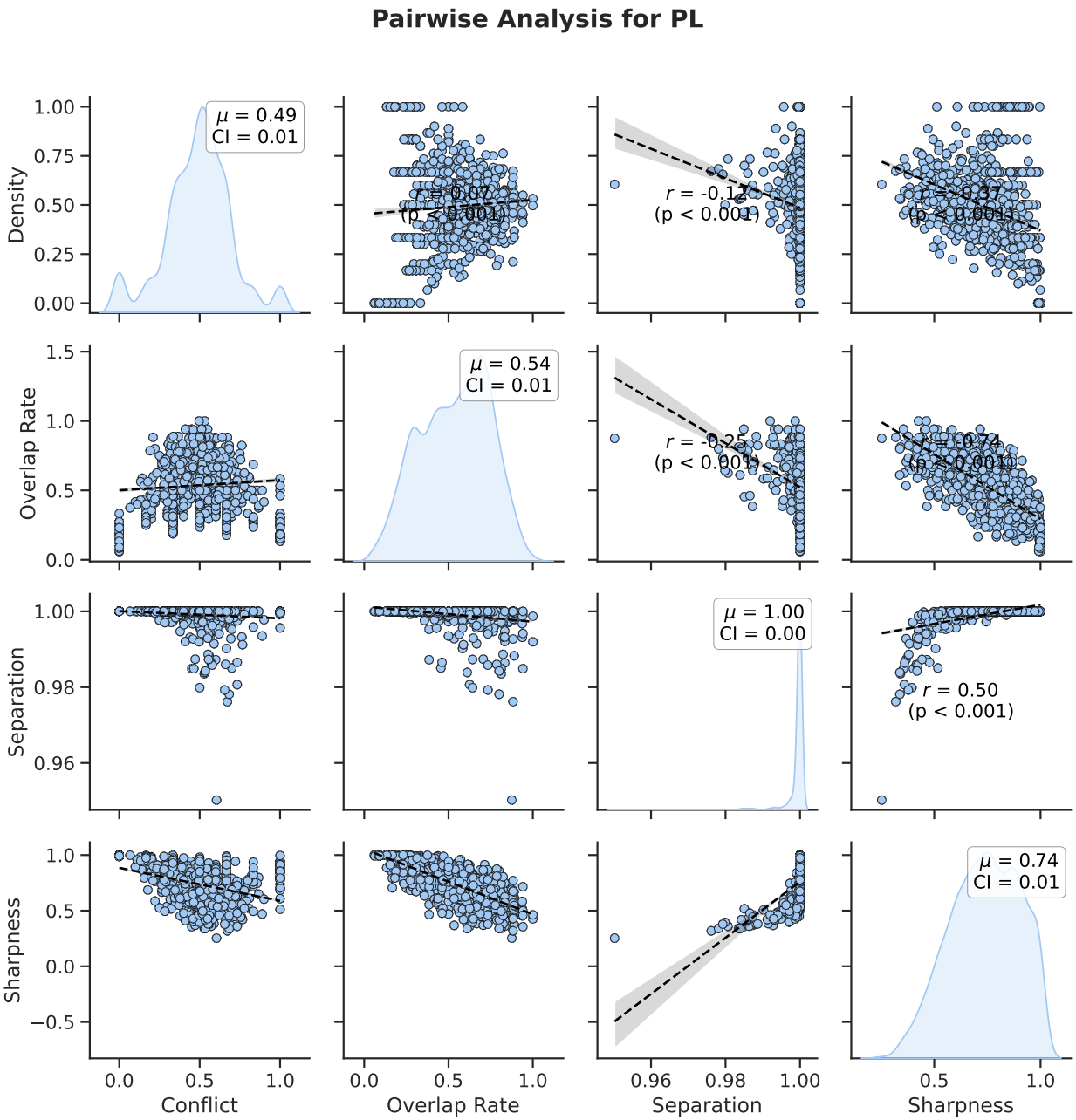

Figure 4: Correlation between sharpness and input partial ranking characteristics when the aggregate rankings are generated with the PL variant

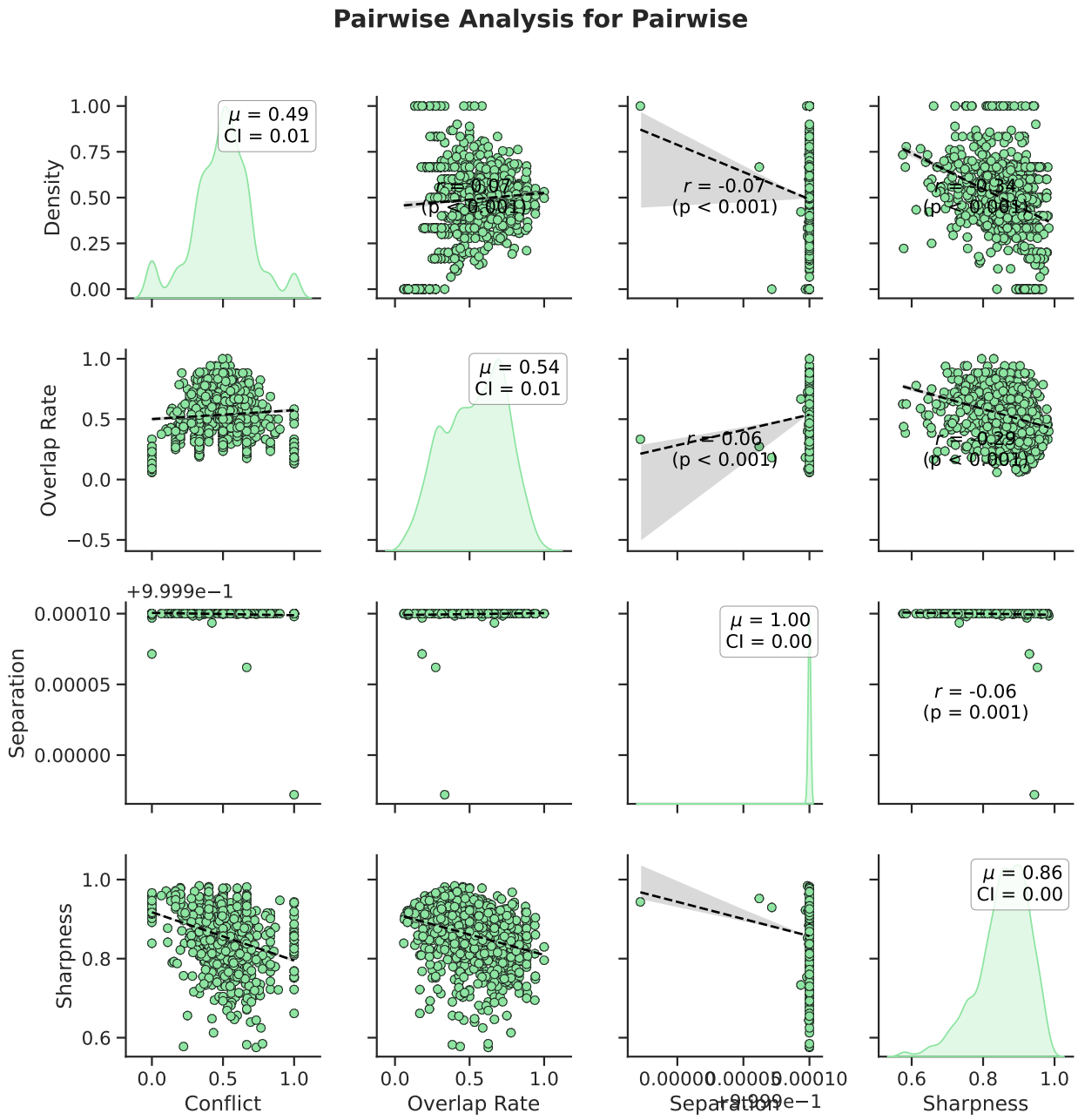

Figure 5: Correlation between sharpness and input partial ranking characteristics when the aggregate rankings are generated with the PP variant

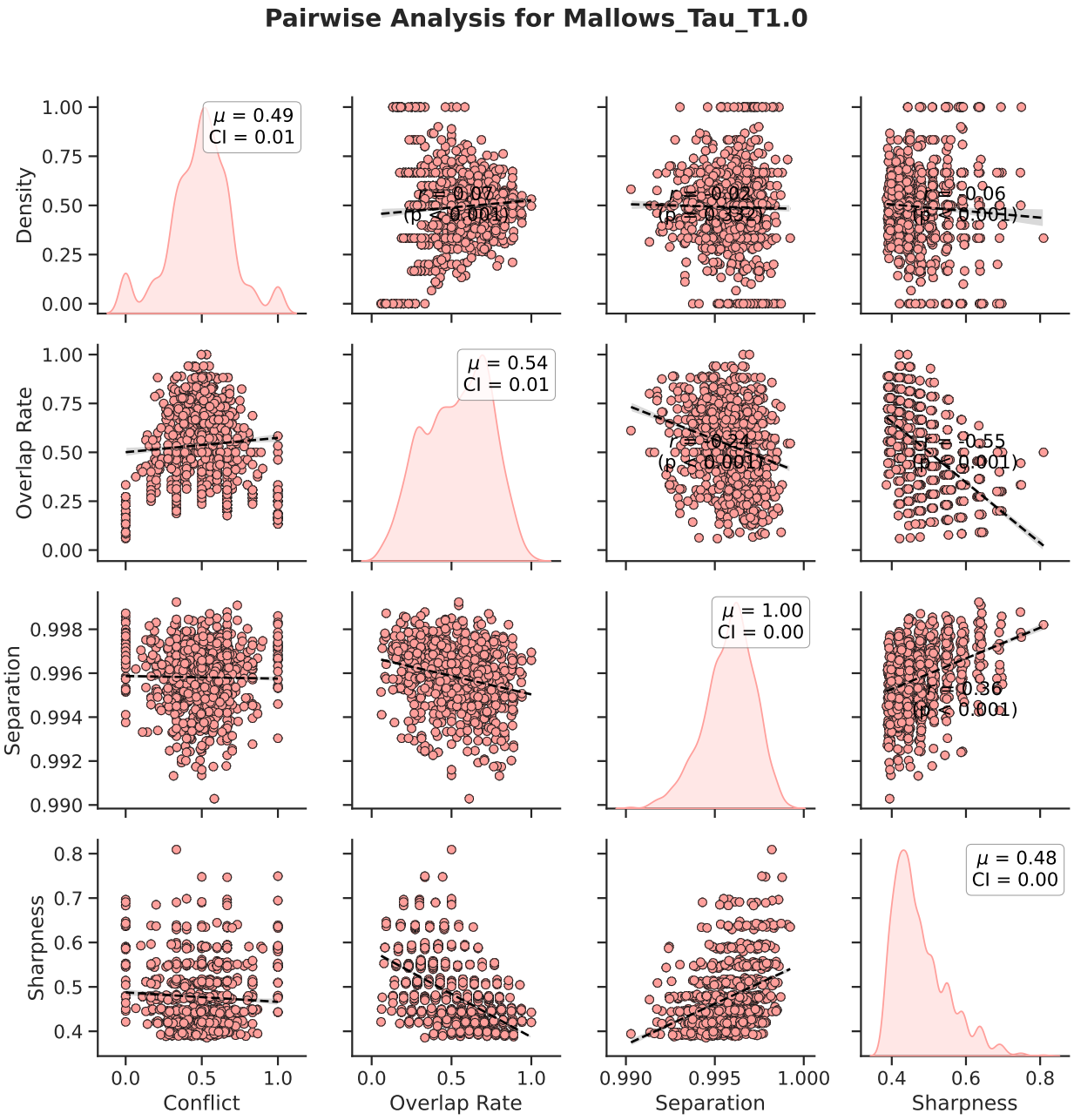

Figure 6: Correlation between sharpness and input partial ranking characteristics when the aggregate rankings are generated with the Mallows ($\theta = 1$) variant

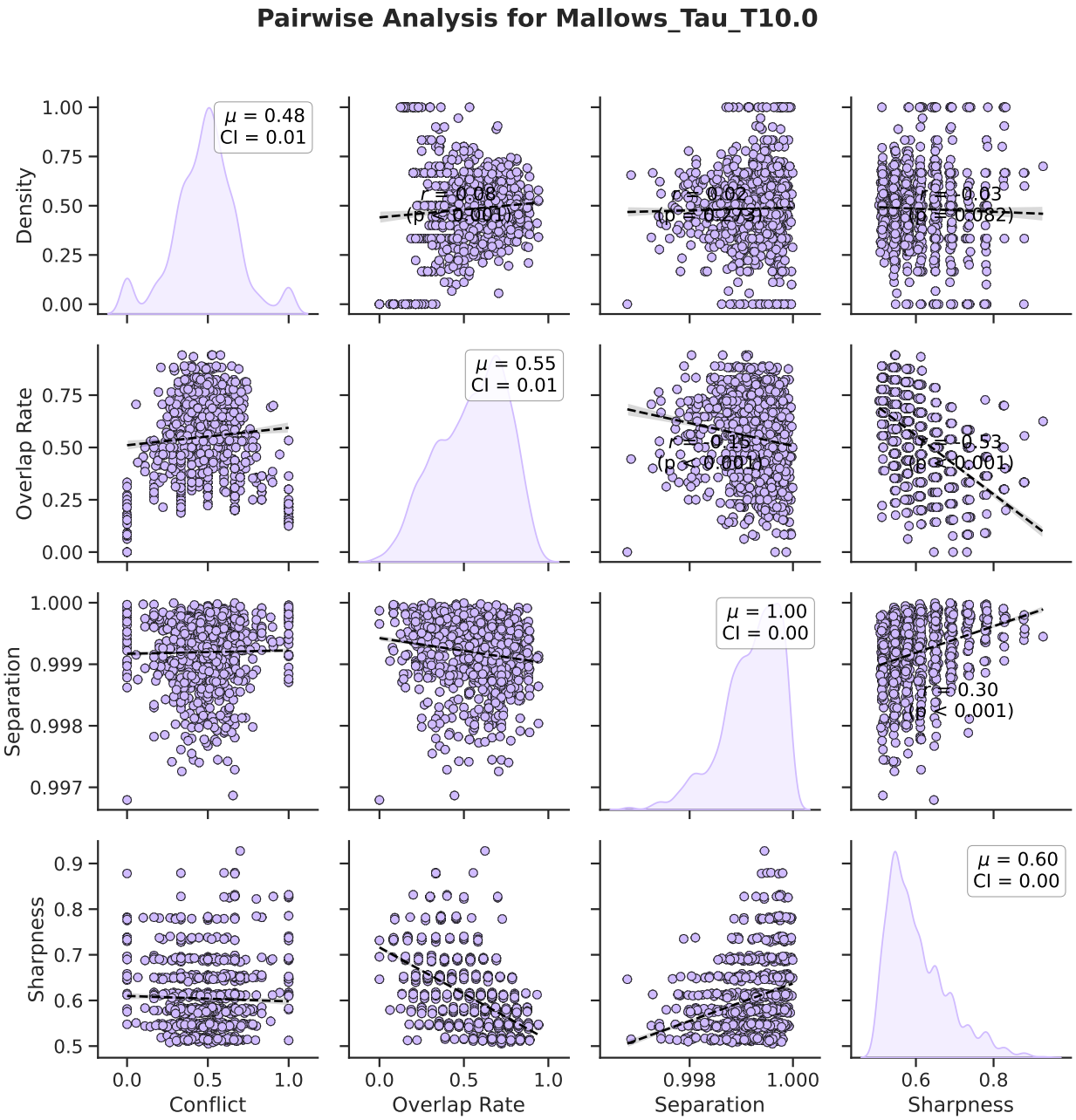

Figure 7: Correlation between sharpness and input partial ranking characteristics when the aggregate rankings are generated with the Mallows ($\theta = 10$) variant

Appendix F. Biomarkers for synthetic experiments

This is how we processed the ADNI data and obtained the parameters reported in Table 6. We obtained the `adnimerge` table (Version: September 7, 2023) from the Alzheimer's Disease Cooperative Study data system. We only included the baseline visits from participants with a diagnosis of CN, Early and Late MCI, or AD. For brain region MRI values, we applied intracranial volume (ICV) normalization because people have different brain sizes. We excluded participants with any missing values from any of the 18 biomarkers reported in Table 5. After de-duplicating the final dataset, we had 420 participants with 350 (83.3%) from the protocol of ADNI2 and 70 (18.3%) from ADNIGO. The distribution of diagnosis of these 420 participants is:

- AD: 67 (16.0%)
- EMCI: 175 (41.7%)
- LMCI: 101 (24.0%)
- CN: 77 (18.3%)

We then applied SA-EBM (Hao et al., 2025) (the Conjugate Priors variant) to obtain the biomarker distribution parameters reported in Table 6.

Table 5: Glossary of Biomarkers with Source, Units, and Interpretation

Abbrev.	**Full Name**	**Source**	**Unit / Scale**	**Higher Values Indicate**
ABETA	Amyloid Beta ($A\beta_{1-42}$)	CSF	pg/mL	Less pathology (less amyloid)
ADAS13	ADAS-Cog (13-item)	Cognitive test	0–85	More pathology (worse cognition)
CDRSB	CDR Sum of Boxes	Clinical rating	0–18	More pathology (greater severity)
EntorhinalNorm	Entorhinal Volume (norm.)	MRI	Ratio	Less pathology (greater volume)
FAQ	Functional Activities Q.	Informant test	0–30	More pathology (worse function)
FDG	FDG-PET (PCC/precuneus)	PET	SUVR	Less pathology (better metabolism)
FusiformNorm	Fusiform Volume (norm.)	MRI	Ratio	Less pathology (greater volume)
HippocampusNorm	Hippocampal Volume (norm.)	MRI	Ratio	Less pathology (greater volume)
LDELTOTAL	Logical Memory II (Delayed)	Cognitive test	0–25	Less pathology (better recall)
MMSE	Mini-Mental State Exam	Cognitive test	0–30	Less pathology (better cognition)
MOCA	Montreal Cognitive Assessment	Cognitive test	0–30	Less pathology (better cognition)
MidTempNorm	Middle Temporal Volume (norm.)	MRI	Ratio	Less pathology (greater volume)
PTAU	Phosphorylated Tau_{181}	CSF	pg/mL	More pathology (tangle burden)
RAVLT-immediate	Rey Auditory Verbal Learning (Immediate)	Cognitive test	0–75	Less pathology (better memory)
TAU	Total Tau	CSF	pg/mL	More pathology (axonal injury)
TRABSCOR	Trail Making Test B	Cognitive test	Seconds	More pathology (slower exec. func.)
VentricleNorm	Ventricular Volume (norm.)	MRI	Ratio	More pathology (ventricular enlarge.)
WholeBrainNorm	Whole Brain Volume (norm.)	MRI	Ratio	Less pathology (greater volume)

Table 6: Biomarker Parameterization and Non-Normal Sampling Distributions

Biomarker	θ_{mean}	θ_{std}	ϕ_{mean}	ϕ_{std}	**Non-normal Distribution (Per Implementation)**
Cognitive / Functional Measures					
MMSE	24.74	2.26	28.75	1.30	Triangular$(\mu-2\sigma, \mu-1.5\sigma, \mu)$; $\mathcal{N}(\mu+\sigma, (0.3\sigma)^2)$; Exp$(0.7\sigma) + (\mu - 0.5\sigma)$ (equal mixture).
ADAS13	26.09	7.77	11.22	4.83	Same mixture structure as MMSE (triangular + Gaussian + exponential).
RAVLT_immediate	25.00	6.39	41.70	10.42	Same mixture structure as MMSE.
CDRSB	2.12	1.66	0.04	0.16	Same mixture structure as MMSE.
FAQ	8.12	6.54	0.52	0.90	Same mixture structure as MMSE.
LDELTOTAL	2.95	2.82	9.77	3.53	Same mixture structure as MMSE.
MOCA	18.85	3.52	24.46	2.62	Same mixture structure as MMSE.
TRABSCOR	276.62	32.45	97.77	43.00	Same mixture structure as MMSE.
CSF Biomarkers & FDG					
ABETA	712.35	240.58	1077.89	353.53	Pareto(1.5)$\cdot\sigma + (\mu - 2\sigma)$; $\mathcal{U}(\mu - 1.5\sigma, \mu + 1.5\sigma)$; Logistic$(\mu, \sigma)$ (equal mixture).
TAU	350.36	148.44	211.02	78.13	Same mixture structure as ABETA.
PTAU	34.99	16.15	19.46	8.24	Same mixture structure as ABETA.
FDG	1.09	0.12	1.29	0.12	Same mixture structure as ABETA.
Subcortical Volumes					
VentricleNorm	0.03256	0.01228	0.02064	0.00929	Beta(0.5,0.5)$\cdot 4\sigma + (\mu - 2\sigma)$; Exp$(0.4\sigma)$ with $\pm$ sign; $\mathcal{N}(\mu, (0.5\sigma)^2) + \{0, 2\sigma\}$ spike.
HippocampusNorm	0.00394	0.00053	0.00502	0.00059	Same mixture structure as VentricleNorm.
Cortical Atrophy Measures					
WholeBrainNorm	0.66854	0.03340	0.71070	0.03502	Gamma$(2, 0.5\sigma)+(\mu-\sigma)$; Weibull$(1.0)\cdot\sigma+(\mu-\sigma)$; $\mathcal{N}(\mu, (0.5\sigma)^2) \pm \sigma$.
EntorhinalNorm	0.001997	0.000401	0.002576	0.000377	Same mixture structure as WholeBrainNorm.
FusiformNorm	0.01095	0.00127	0.01255	0.00131	Standard Cauchy$(\mu, \sigma) + \mathcal{N}(0, (0.2\sigma)^2)$, clipped to $[\mu - 4\sigma, \mu + 4\sigma]$.
MidTempNorm	0.01205	0.00145	0.01397	0.00129	10% $\mathcal{N}(\mu, 0.2\sigma)$ spike + 90% Logistic$(\mu + \sigma, 2\sigma)$.

Implementation Notes: After sampling, all values are perturbed by additional noise $\mathcal{N}(0, (0.2\sigma)^2)$ and clipped to $[\mu - 5\sigma, \mu + 5\sigma]$ for stability.

Appendix G. Experimental setup

Table 7: Summary of nine simulation experiments. S denotes the event sequence, k_j the participant stage distribution, and X the biomarker distribution.

Exp.	**Event Sequence (S)**	**Stage Distribution (k_j)**	**Biomarker Distribution (X)**
1	Ordinal	Dirichlet-Multinomial (bell-shaped)	Normal
2	Ordinal	Dirichlet-Multinomial (bell-shaped)	Non-Normal
3	Ordinal	Uniform (discrete)	Normal
4	Ordinal	Uniform (discrete)	Non-Normal
5	Ordinal	Continuous Uniform	Normal
6	Ordinal	Continuous Uniform	Non-Normal
7	Ordinal	Continuous Skewed (Beta, $\alpha = 5,\ \beta = 2$)	Non-Normal
8	Ordinal	Continuous Uniform	Sigmoid for diseased; Normal for healthy
9	Ordinal	Continuous Skewed (Beta, $\alpha = 5,\ \beta = 2$)	Sigmoid for diseased; Normal for healthy

Appendix H. Sigmoid Model

The **Sigmoid model** assumes that biomarker values for control individuals are normally distributed:

$$x \sim \mathcal{N}(\mu_\phi, \sigma_\phi^2),$$

where μ_ϕ and σ_ϕ^2 denote the mean and variance in the healthy state, respectively.

For participants affected by disease, the biomarker distribution is shifted away from this baseline according to a smooth, sigmoidal transition:

$$x \sim \mathcal{N}(\mu_{n,\phi}, \sigma_{n,\phi}^2) \; + \; \frac{(-1)^{I_n} R_n}{1 + \exp(-\rho_n(k_j - \xi_n))},$$

where $I_n \sim \text{Bernoulli}(0.5)$ randomly determines the direction of the shift, $R_n = \mu_{n,\theta} - \mu_{n,\phi}$ specifies the magnitude of separation between disease and control means, and the slope parameter is given by

$$\rho_n = \max\left(1, \; \frac{|R_n|}{\sqrt{\sigma_{n,\theta}^2 + \sigma_{n,\phi}^2}}\right).$$

Appendix I. Synthetic experiment results

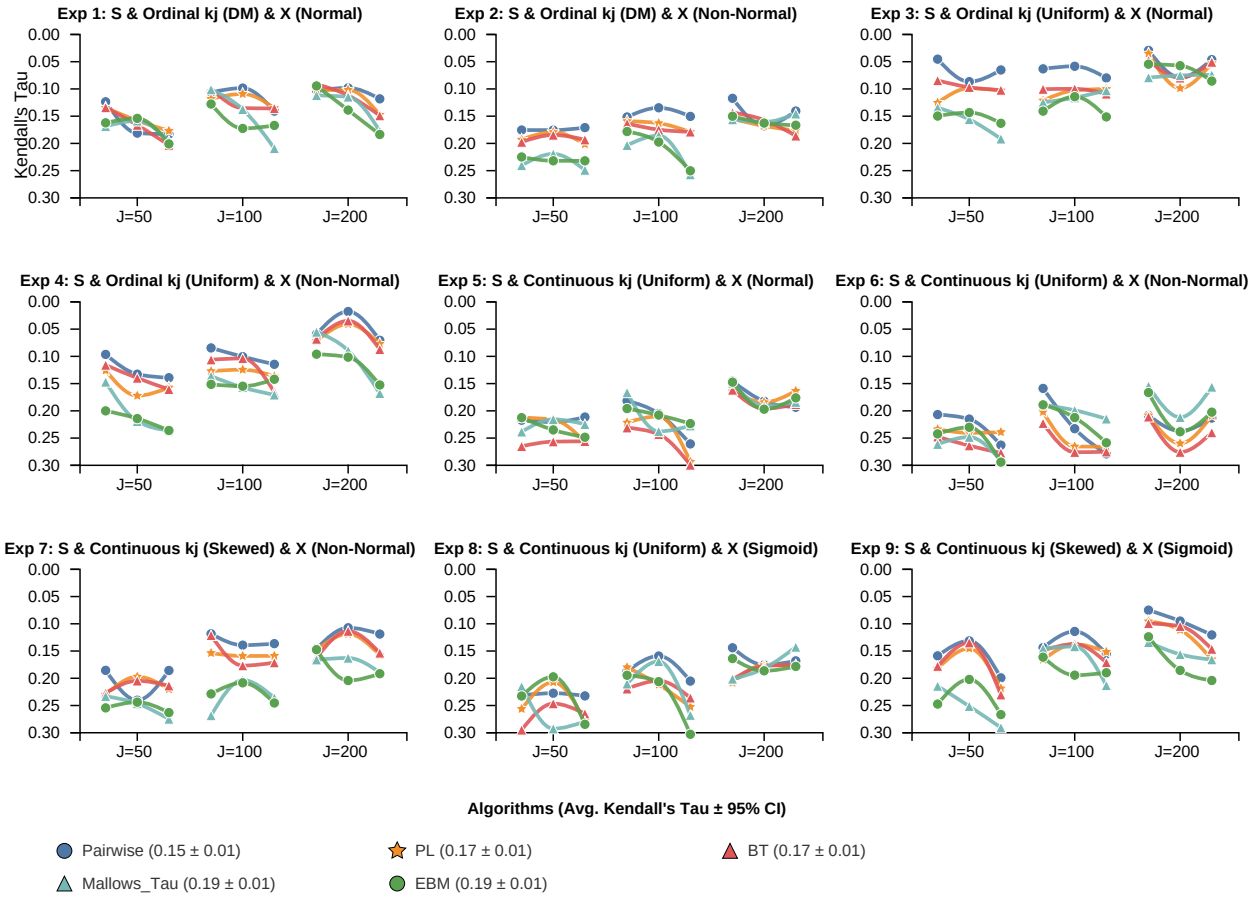

Figure 8: **Average normalized Kendall's Tau (±95% CI) for data generated by PP**. Each panel displays a different experiment with varying configurations represented by the panel title. The X-axis of each subplot shows participant sizes ($J = 50, 100, 200$). For each participant size, there are three different healthy ratios ($0.25, 0.5, 0.75$, from left to right). The Y-axis is the normalized Kendall's tau distance. A lower tau indicates the estimation is closer to the ground truth. Data points represent mean performance across 10 variants of the same experimental configurations, participant size, and healthy ratio. JPM (except for Mallows with $\theta = 1.0$) outperforms the single-disease EBM model, achieving a 21% increase in the performance of the ordering task. This advantage is consistent across experiments, participant sizes, and healthy ratios, but more pronounced when J is small.

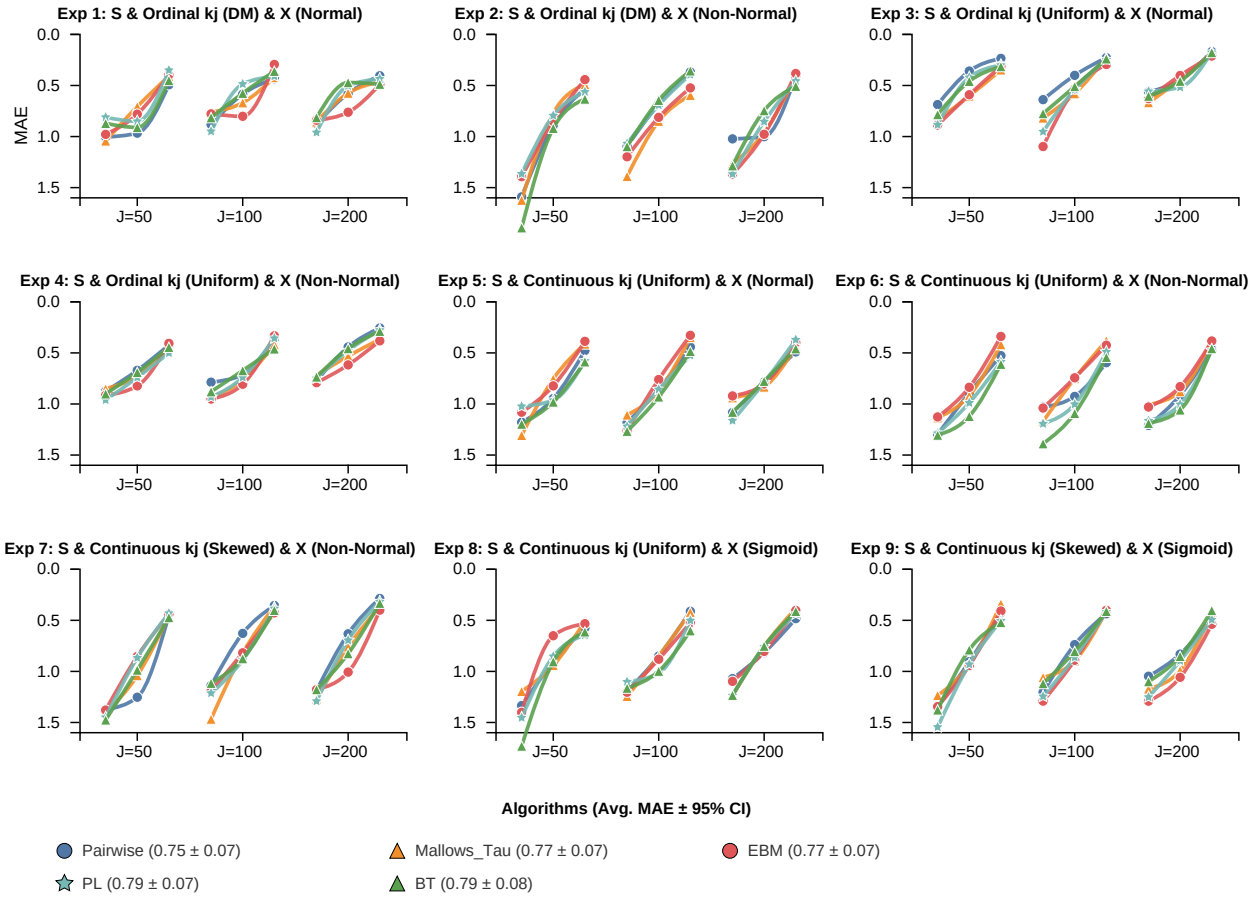

Figure 9: MAE (staging) for PP generated data

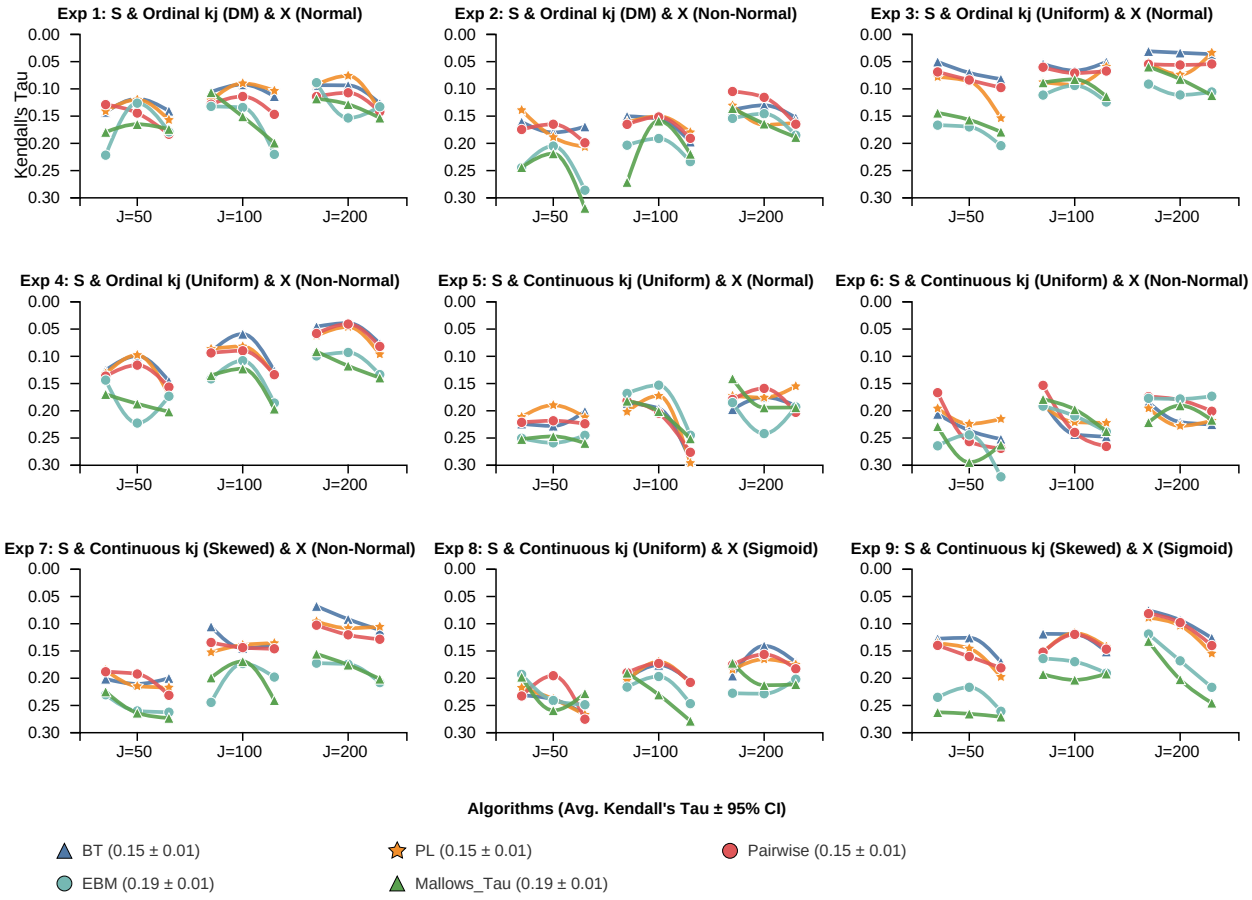

Figure 10: Kendall's tau distance (ordering) for BT generated data

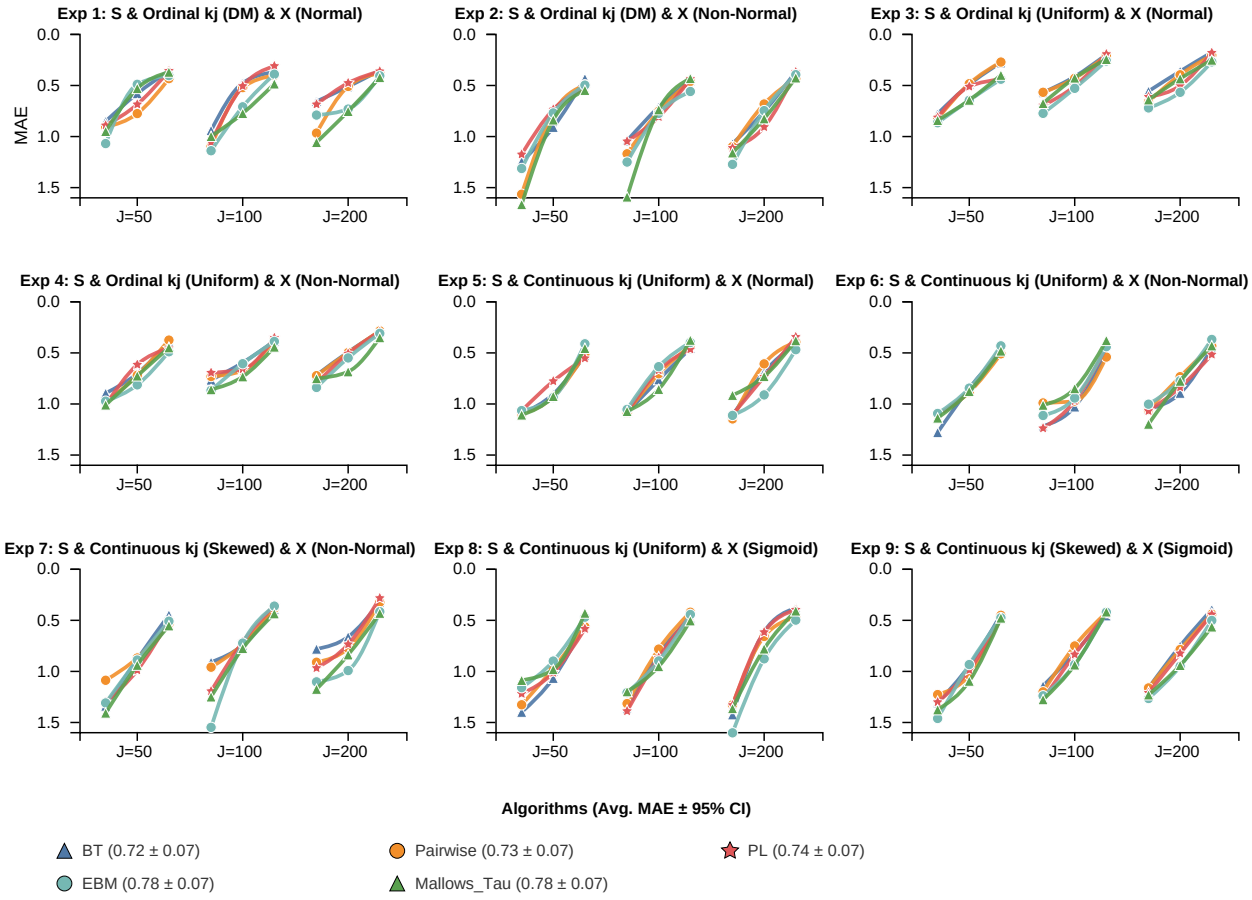

Figure 11: MAE (staging) for BT generated data

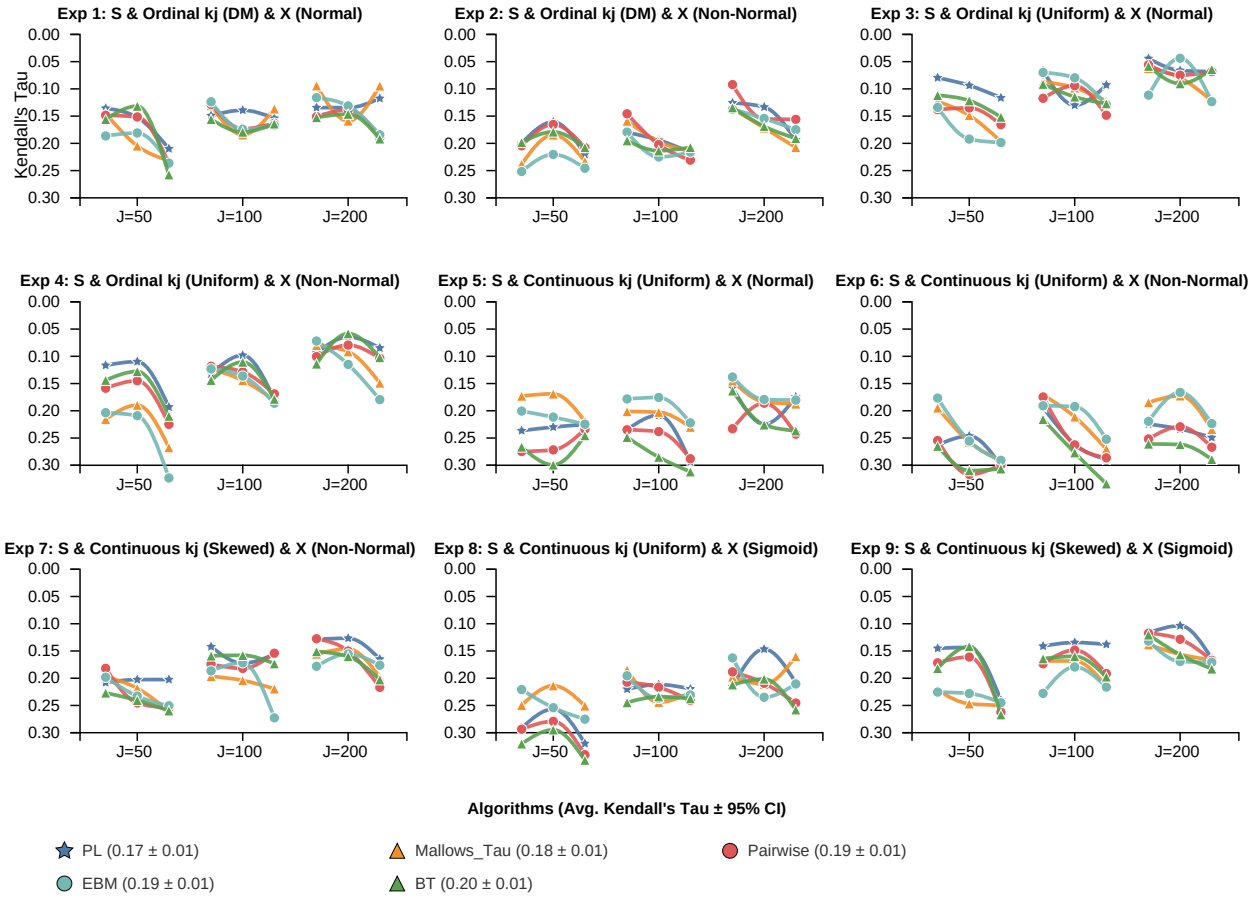

Figure 12: Kendall's tau distance (ordering) for PL generated data

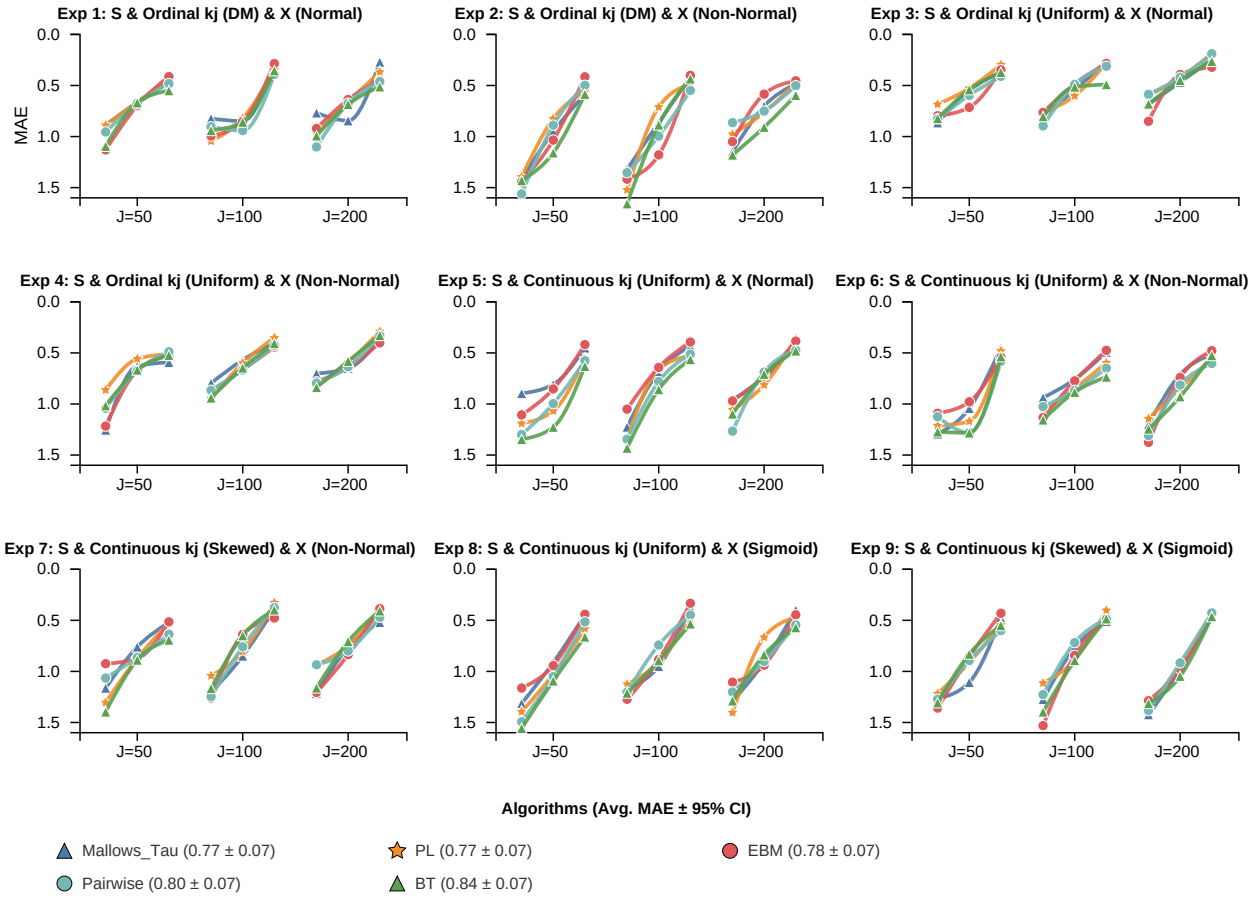

Figure 13: MAE (staging) for PL generated data

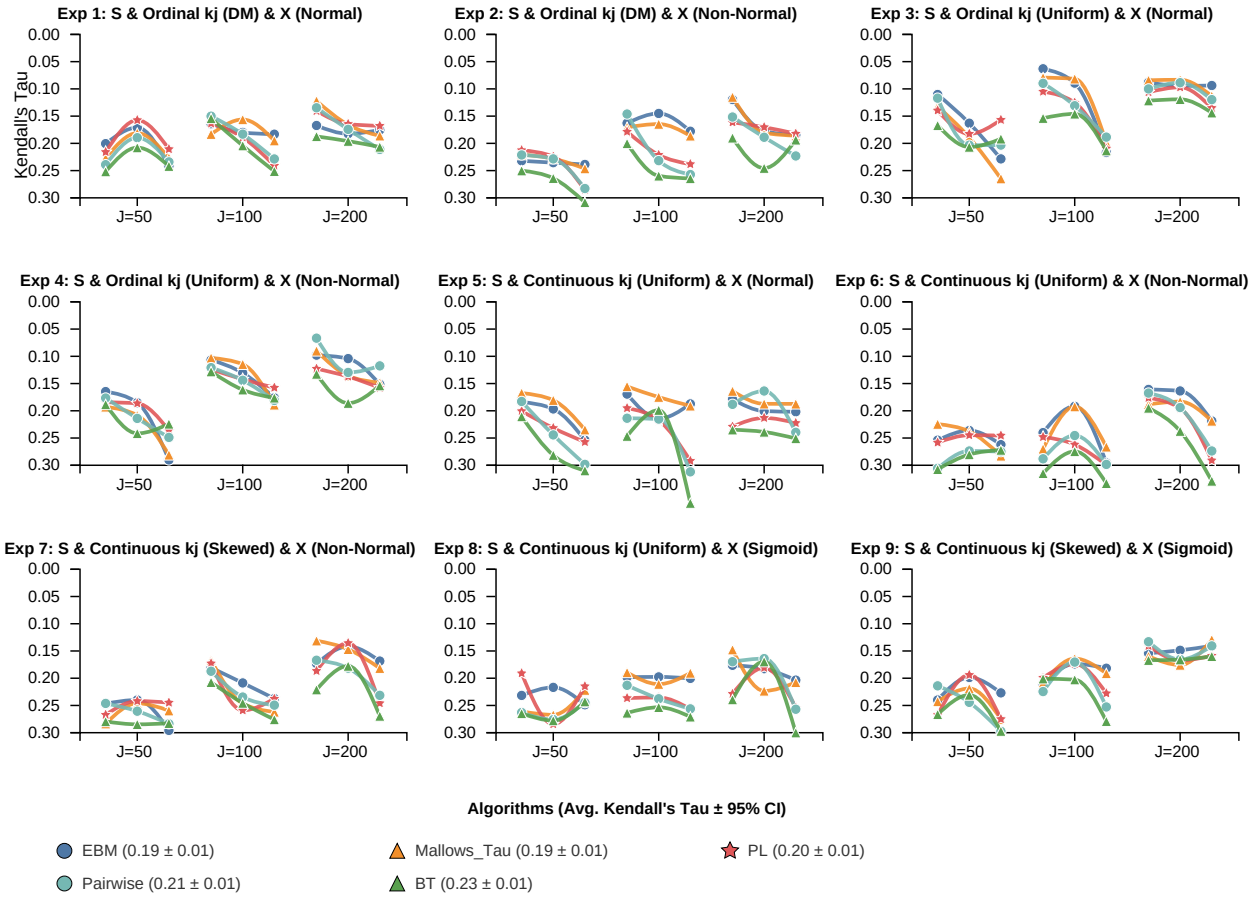

Figure 14: Kendall's tau distance (ordering) for Mallows (dispersion = 1.0) generated data

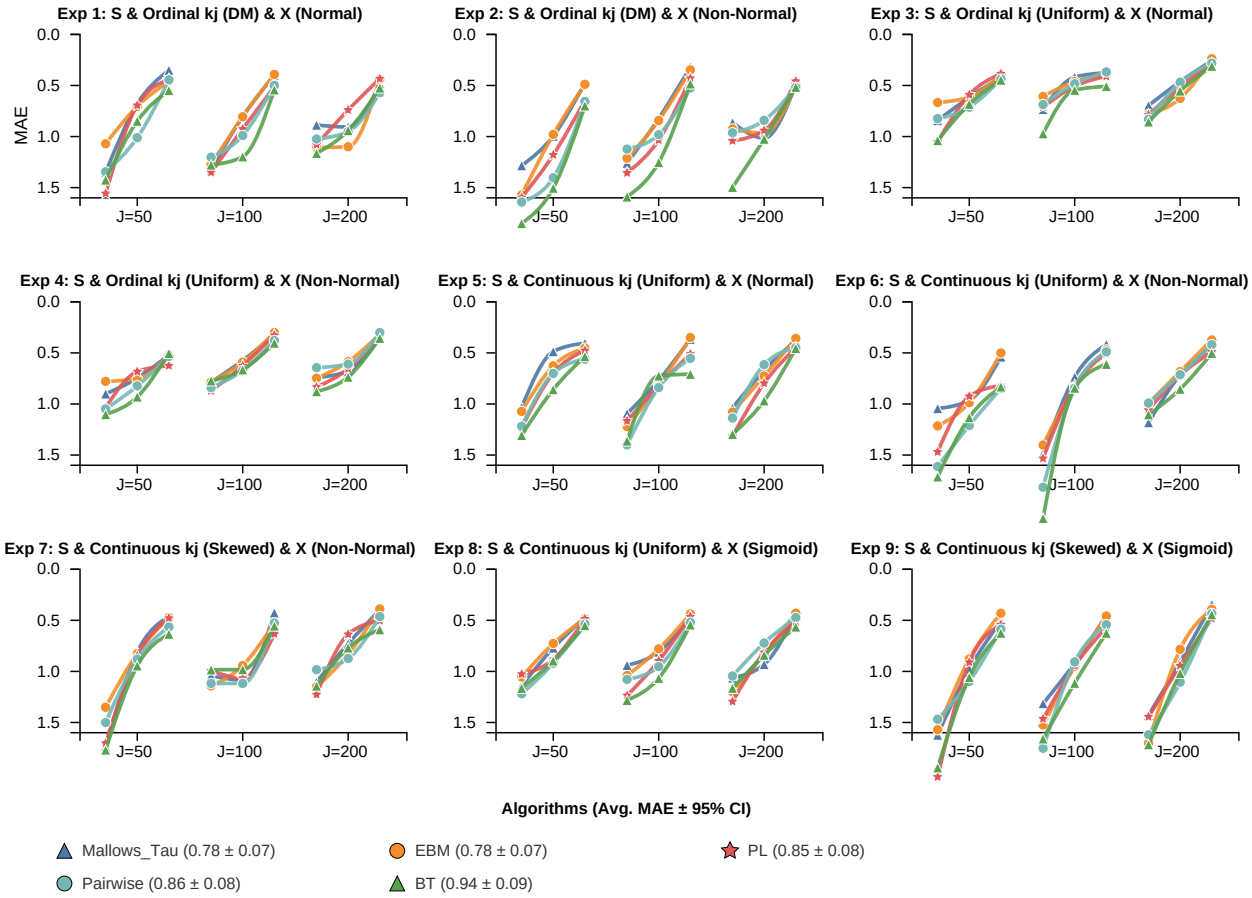

Figure 15: MAE (staging) for Mallows (dispersion = 1.0) generated data

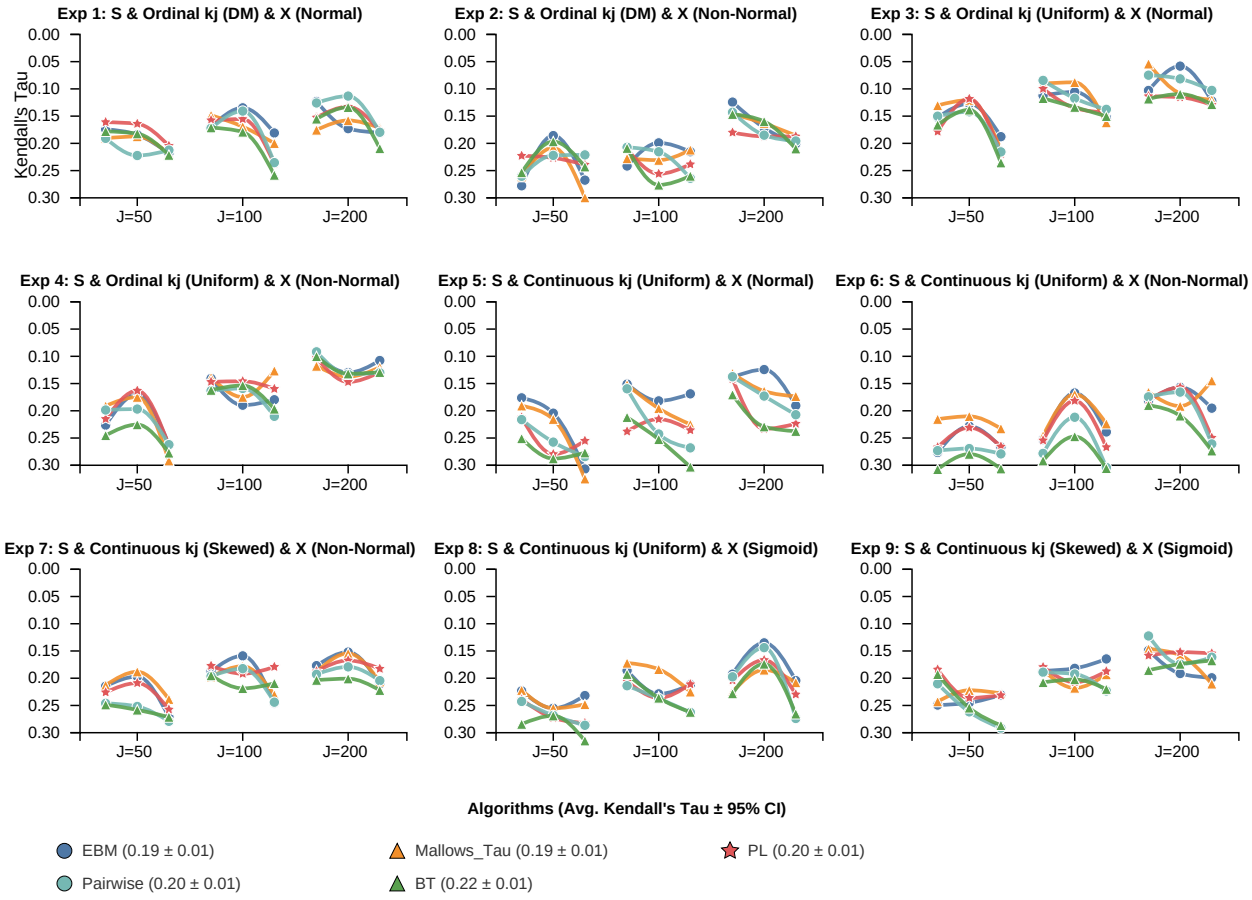

Figure 16: Kendall's tau distance (ordering) for Mallows ($\theta = 10.0$) generated data

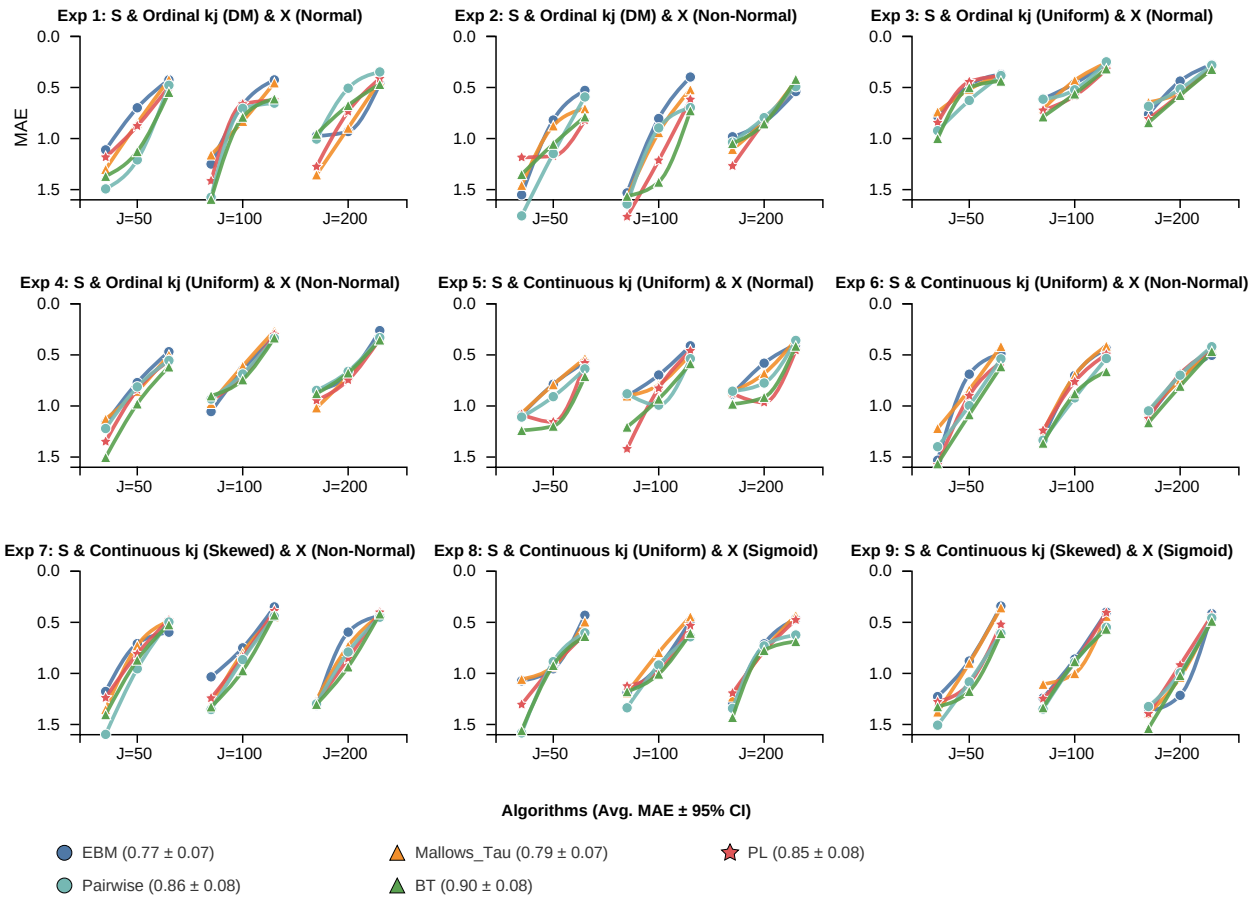

Figure 17: MAE (staging) for Mallows ($\theta = 10.0$) generated data

Appendix J. NACC Data Preprocessing Pipeline

This appendix provides a comprehensive description of the data processing pipeline used to select and prepare the National Alzheimer's Coordinating Center (NACC) dataset for our real-world experiments. The process involved three main stages: cohort definition, biomarker data integration, and final data transformation.

The following data are needed:

- UDS with NP and Genetics plus FTLD and LBD modules: `ftldlbd_nacc70.csv`
- CSF: `fcsf_nacc70.csv`
- Phenotype Harmonization Consortium (PHC) scores with NACCID and visit number: `ADSP-PHC_Cognition_NACC_DS_Freeze1.csv`
- Mixed protocol (non-SCAN compliant) MRI: `investigator_mri_nacc70.csv`

J.1. Initial Data Loading and Cohort Definition

The first stage focused on identifying distinct diagnostic cohorts from the primary NACC dataset.

Data Loading and Subject Uniqueness We began with the NACC dataset file `ftldlbd_nacc70.csv`, which contained longitudinal data for **54,631 unique participants**. To create a cross-sectional dataset for our analysis, we filtered these records to retain only the earliest visit for each participant, identified by `NACCVNUM = 1`.

Pathological Cohort Identification We then defined four primary pathological cohorts by creating Boolean flags based on a combination of clinical diagnosis and derived NACC variables. A participant was considered positive for a pathology if they met any of the criteria listed below:

- **Alzheimer's Disease (AD):** `PROBAD=1`, `POSSAD=1`, `NACCALZP=1`, or `NACCALZD=1`.
- **Lewy Body Dementia (LBD):** `NACCLBDP=1` or `NACCLBDE=1`.
- **Frontotemporal Lobar Degeneration (FTLD):** `NACCFTD=1`, `FTD=1`, `FTLDNOS=1`, or `FTLDMO=1`.
- **Vascular Dementia (VaD):** `NACCVASC=1`, `VASC=1`, or `CVD=1`.

J.2. Biomarker Data Integration and Filtering

The second stage involved merging and filtering data from three separate biomarker files to enrich the primary dataset.

Cerebrospinal Fluid (CSF) Data We loaded CSF biomarker data from `nacc_csf.csv`. To ensure data quality and relevance, we applied two key filters. First, we selected only baseline visit records. Second, because CSF sample dates do not always align perfectly with clinical visit dates, we implemented a temporal window, associating a CSF sample with a visit only if the sample was taken within **180 days** of the visit date. The resulting data was then merged with our main participant dataframe.

Phenotype Harmonization Consortium (PHC) Data Again, we selected only baseline visit records. This dataset was then merged directly with the main dataframe using the participant ID as the key.

MRI Volumetric Data Structural MRI data was loaded from `mri_vol.csv`. We filtered this dataset to include only baseline MRI scans before merging it with the main dataframe on the participant ID.

J.3. Final Cohort Analysis and Data Availability

After successfully merging all data sources, we finalized the cohorts and quantified the availability of biomarker data across them.

Creation of Mutually Exclusive Cohorts Using the pathological flags defined earlier, we created a set of mutually exclusive cohorts. This allowed us to distinguish between participants with a single pathology (e.g., `AD only`) and those with mixed pathologies (e.g., `AD VaD`).

Layered Availability Analysis We generated boolean flags for each participant to indicate the presence of valid data for each biomarker type (`has_csf`, `has_phc`, `has_mri`). We then systematically counted the number of participants within each mutually exclusive cohort that had each possible combination of these data types. The results of this analysis are presented in Table 8, which provides a detailed view of data completeness across the study population.

J.4. Data Transformation for Modeling

As a final preprocessing step, the fully integrated and filtered dataframe, which was in a wide format (one column per biomarker), was transformed into a long format using the `pandas.melt` function. This reshaping is a standard and necessary step to structure the data appropriately for subsequent statistical modeling and analysis.

The final overall data availability after our processing is displayed in Table 8. The biomarkers chosen for AD and VaD, with detailed information about each biomarker, are available in Table 9.

Table 8: Layered Availability of NACC Data Across Cohorts. This table shows the total number of participants in each diagnostic cohort and the count of participants for whom specific data types (MRI, CSF, PHC) are available, both individually and in combination.

Cohort	Total	+MRI	+CSF	+PHC	+MRI+PHC	+CSF+MRI	+CSF+PHC	+CSF+MRI+PHC
healthy	26355	4593	1094	18119	3746	234	990	205
AD_only	13400	1656	463	10023	1474	87	413	82
AD_VaD	4449	630	276	3448	578	53	255	50
VaD_only	3903	581	246	3171	532	40	236	36
FTLD_only	2379	146	27	1024	93	6	17	5
LBD_only	1140	89	30	835	74	5	24	5
FTLD_VaD	817	50	10	505	35	1	4	1
AD_FTLD	682	39	19	333	27	4	14	3
AD_LBD	516	48	12	358	43	4	11	3
LBD_VaD	422	64	20	341	58	2	16	2
AD_LBD_VaD	278	27	10	211	24	1	10	1
AD_FTLD_VaD	227	12	8	150	11	0	5	0
FTLD_LBD	26	0	0	12	0	0	0	0
AD_FTLD_LBD	15	0	0	7	0	0	0	0
FTLD_LBD_VaD	13	0	0	6	0	0	0	0
AD_FTLD_LBD_VaD	9	0	0	5	0	0	0	0

Table 9: Glossary of NACC Biomarkers with Source, Units, and Interpretation

Abbrev.	Full Name	Source	Unit / Scale	Higher Values Indicate	AD	VaD
ptau_abeta	Phospho-Tau$_{181}$/Aβ_{1-42} ratio	CSF	Ratio	More pathology (higher tau-to-amyloid)	True	False
ttau_abeta	Total Tau/Aβ_{1-42} ratio	CSF	Ratio	More pathology (higher tau-to-amyloid)	True	False
CSFABETA	Amyloid Beta (Aβ_{1-42})	CSF	pg/mL	Less pathology (lower deposition)	True	False
CSFPTAU	Phosphorylated Tau$_{181}$	CSF	pg/mL	More pathology (tangle burden)	True	False
CSFTTAU	Total Tau	CSF	pg/mL	More pathology (neurodegeneration)	True	False
HippVolNorm	Hippocampal Volume (normalized)	MRI	Ratio	Less pathology (larger volume)	True	False
EntorhinalNorm	Entorhinal Volume (normalized)	MRI	Ratio	Less pathology (larger volume)	True	False
MidTempNorm	Middle Temporal Volume (normalized)	MRI	Ratio	Less pathology (larger volume)	True	False
FusiformNorm	Fusiform Volume (normalized)	MRI	Ratio	Less pathology (larger volume)	True	False
WholeBrainNorm	Whole Brain Volume (normalized)	MRI	Ratio	Less pathology (larger volume)	True	True
VentricleNorm	Ventricular Volume (normalized)	MRI	Ratio	More pathology (enlargement)	True	True
WMHnorm	White Matter Hyperintensity (normalized)	MRI	Ratio	More pathology (greater burden)	False	True
PHC_MEM	Harmonized Composite Memory Score	Cognitive composite	z-score	Better cognition	True	True
PHC_EXF	Harmonized Composite Executive Function	Cognitive composite	z-score	Better cognition	True	True
PHC_LAN	Harmonized Composite Language Score	Cognitive composite	z-score	Better cognition	True	True

We included the two ratios, i.e., `ptau_abeta` and `ttau_abeta` because Palmqvist et al. (2015) found that these ratios have higher diagnostic power for Alzheimer's disease than raw Abeta values.

Appendix K. NACC results

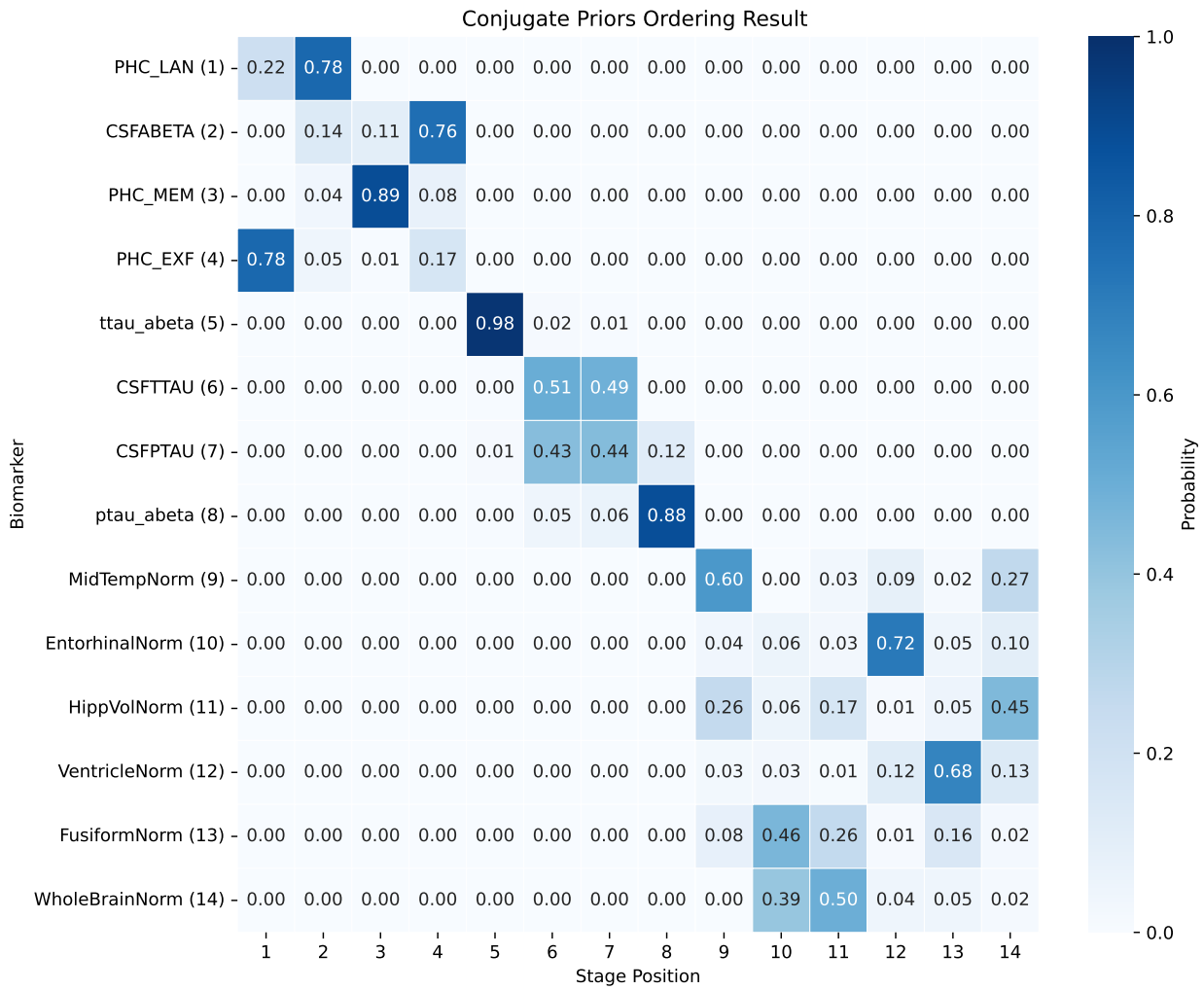

Figure 18: Progression of AD, obtained via SA-EBM

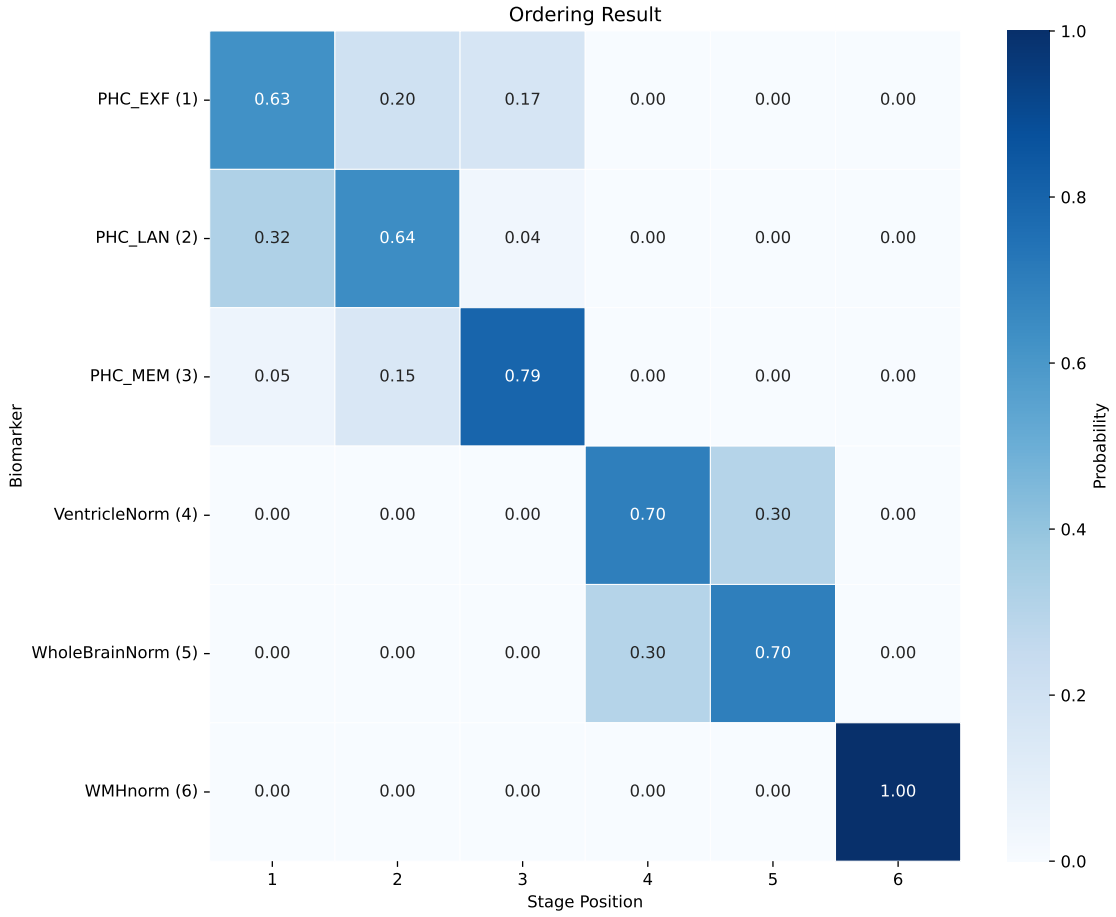

Figure 19: Progression of VaD, obtained via SA-EBM

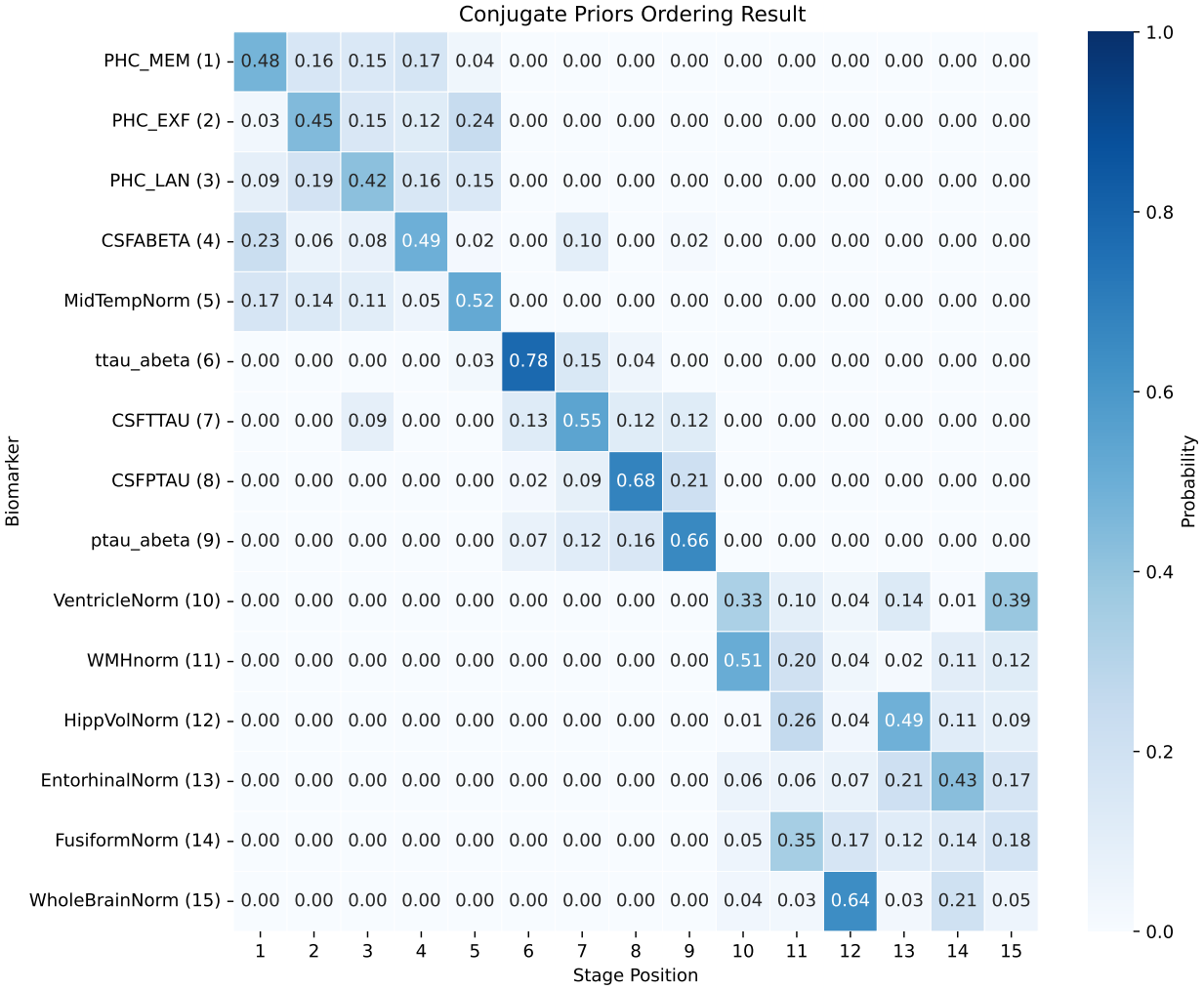

Figure 20: Progression of the mixed pathology of AD and VaD, obtained via SA-EBM

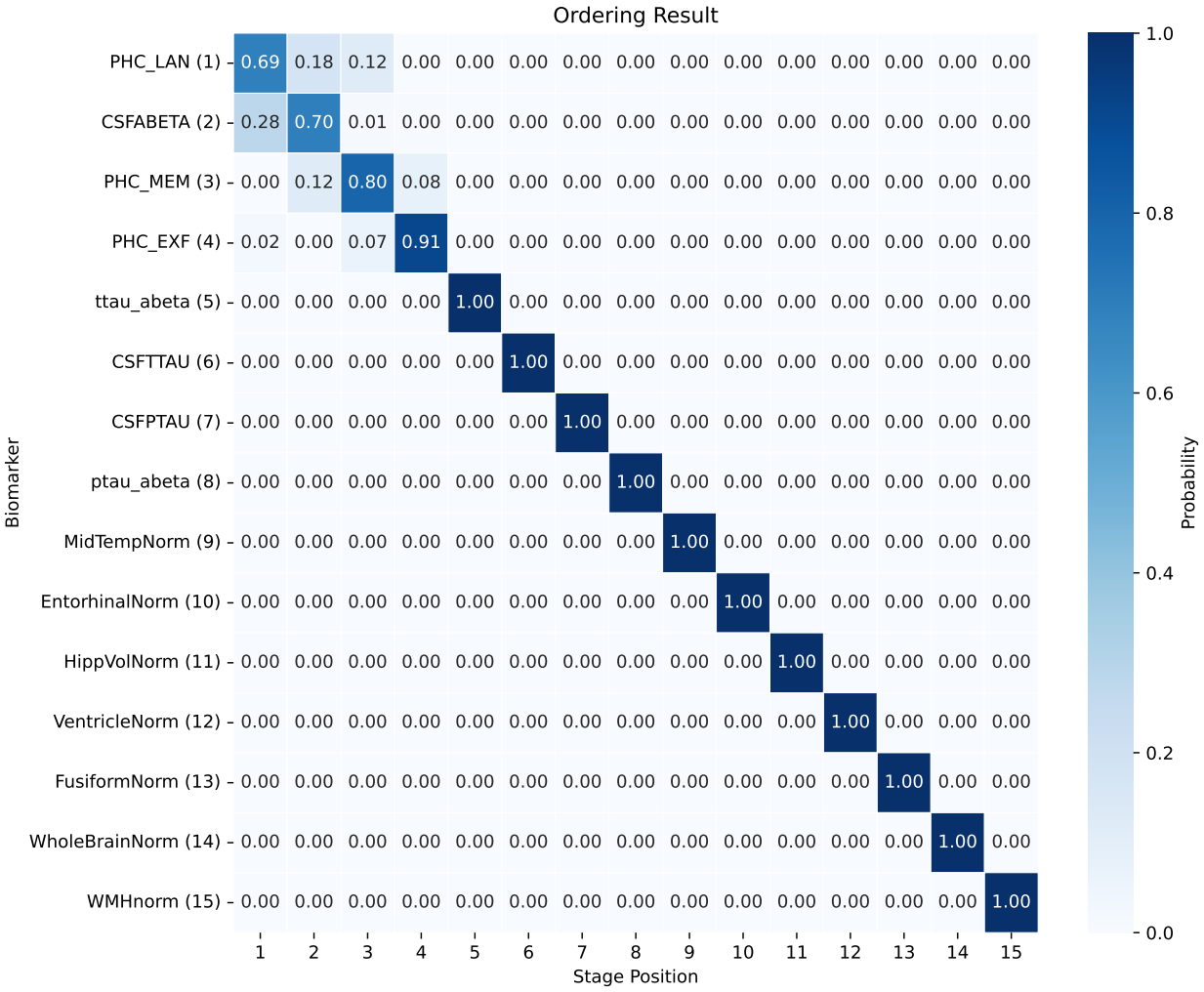

Figure 21: Progression of the mixed pathology of AD and VaD, obtained via JPM (BT)

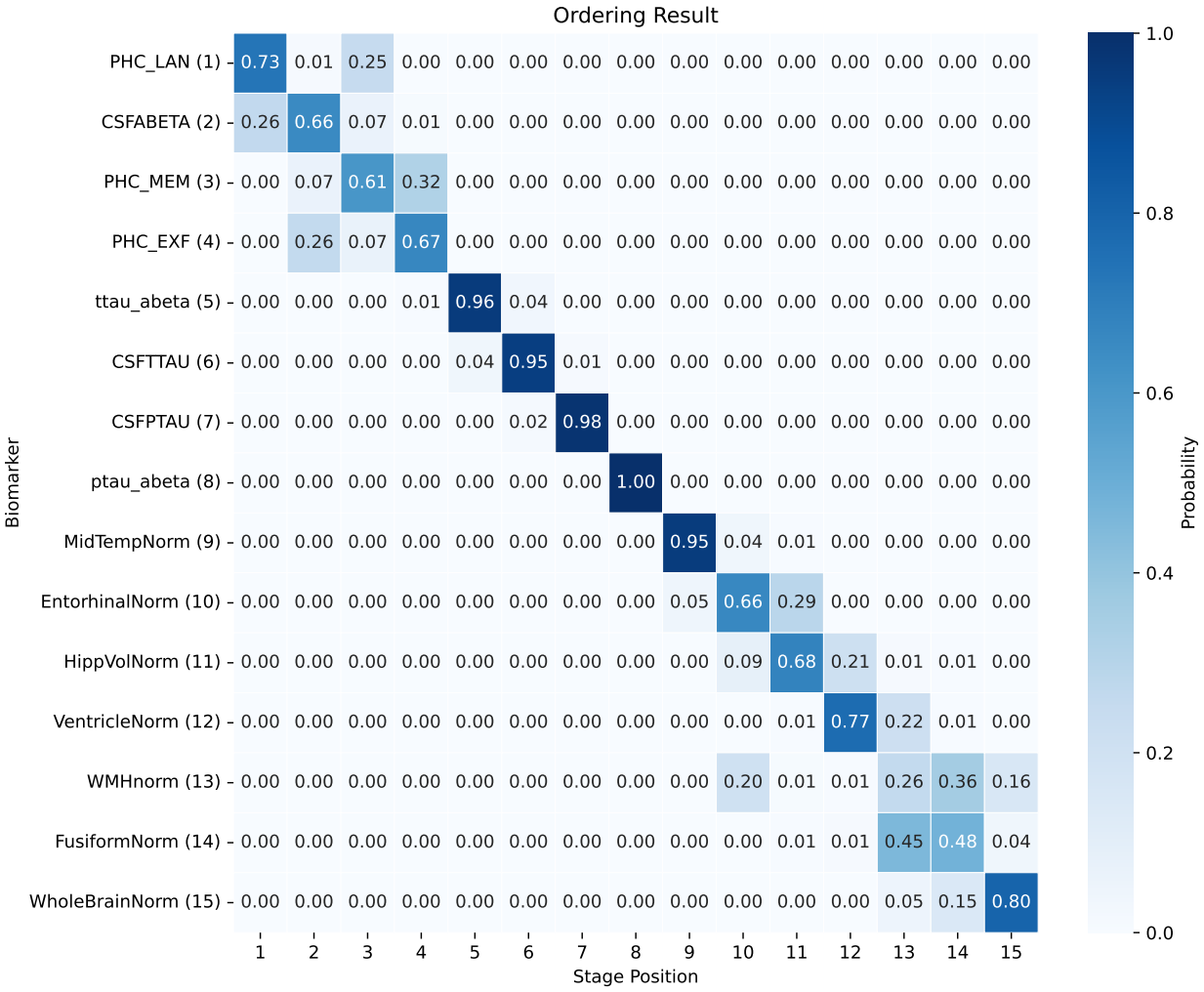

Figure 22: Progression of the mixed pathology of AD and VaD, obtained via JPM (PP)

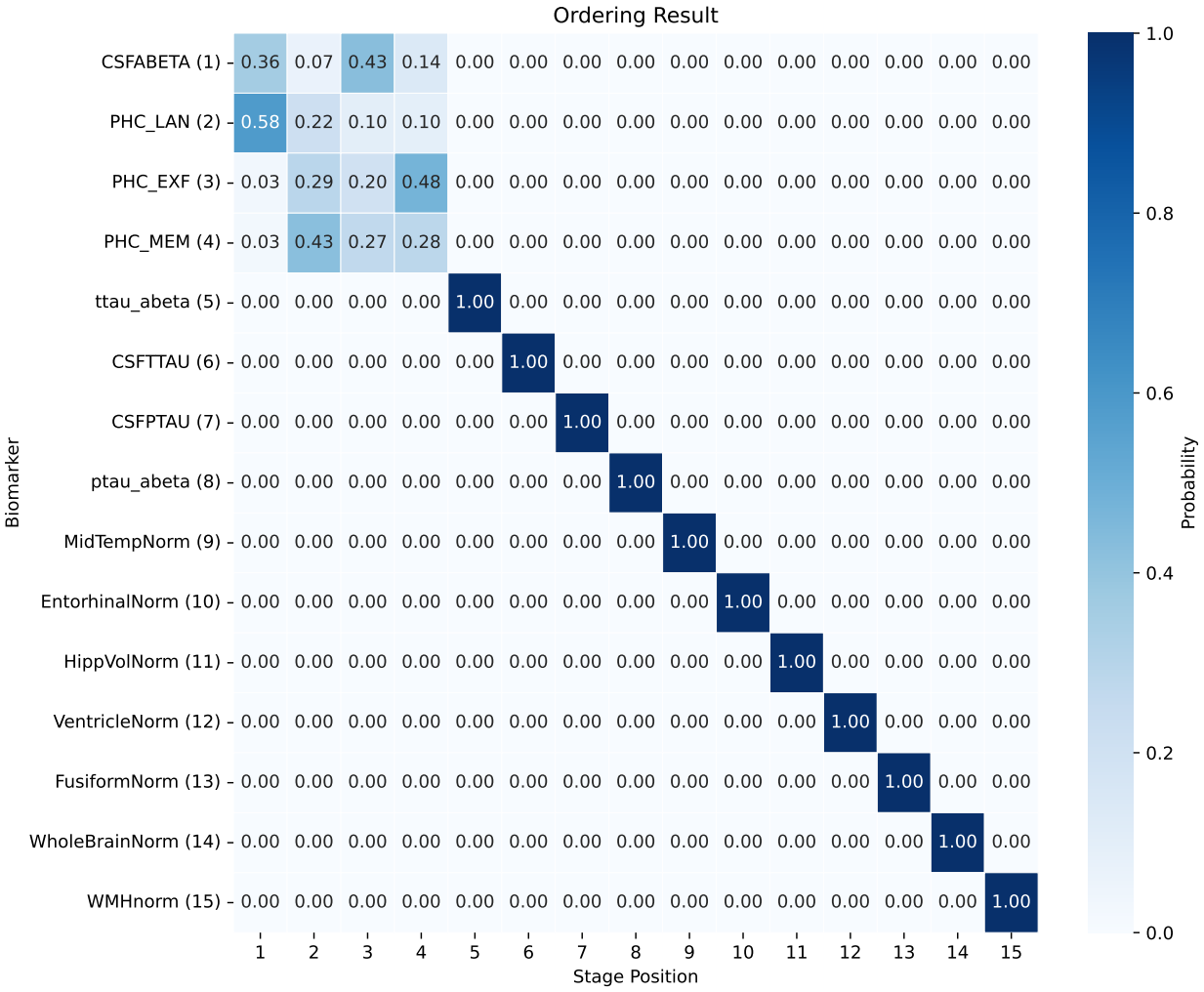

Figure 23: Progression of the mixed pathology of AD and VaD, obtained via JPM (PL)

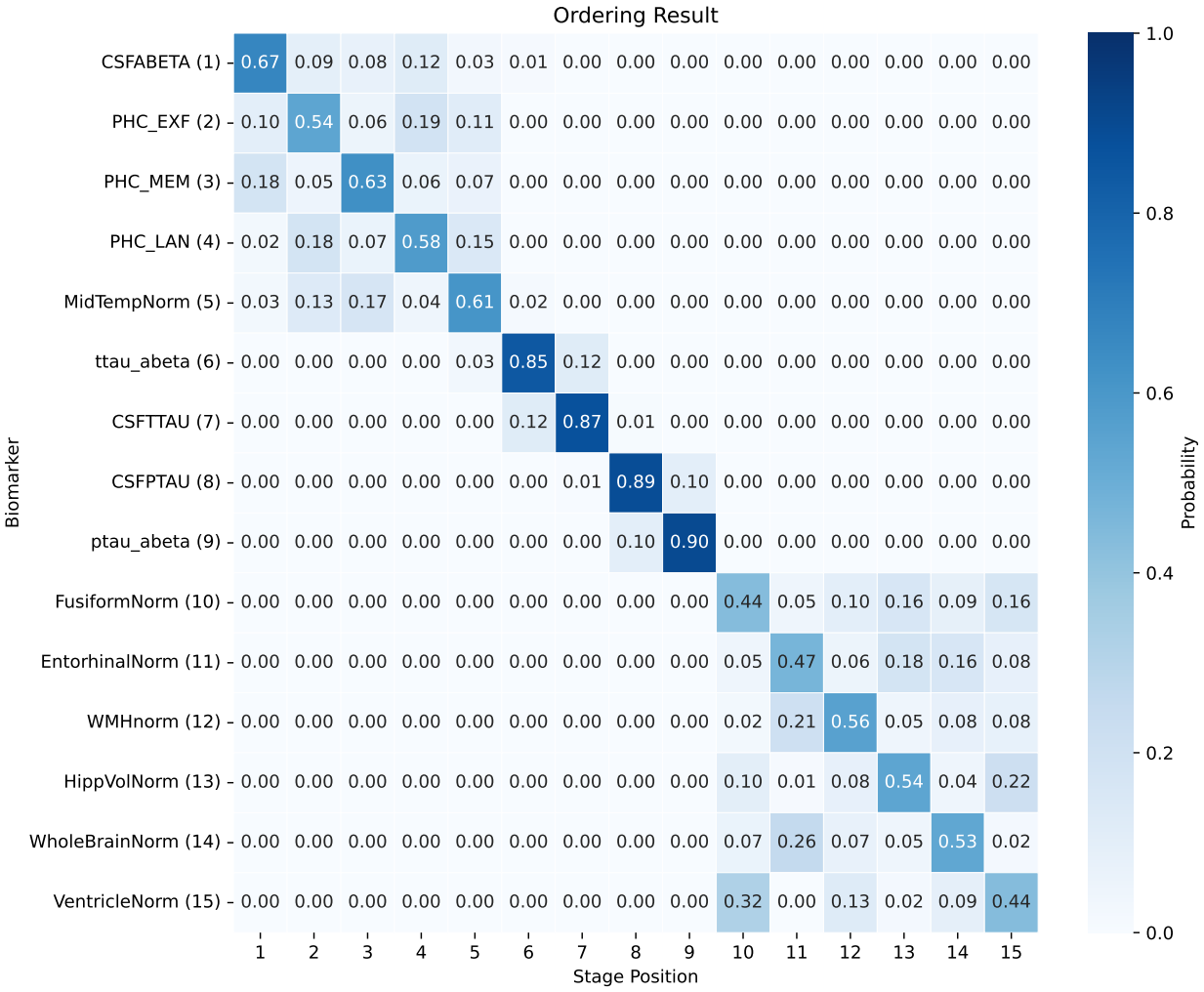

Figure 24: Progression of the mixed pathology of AD and VaD, obtained via JPM (Mallows)

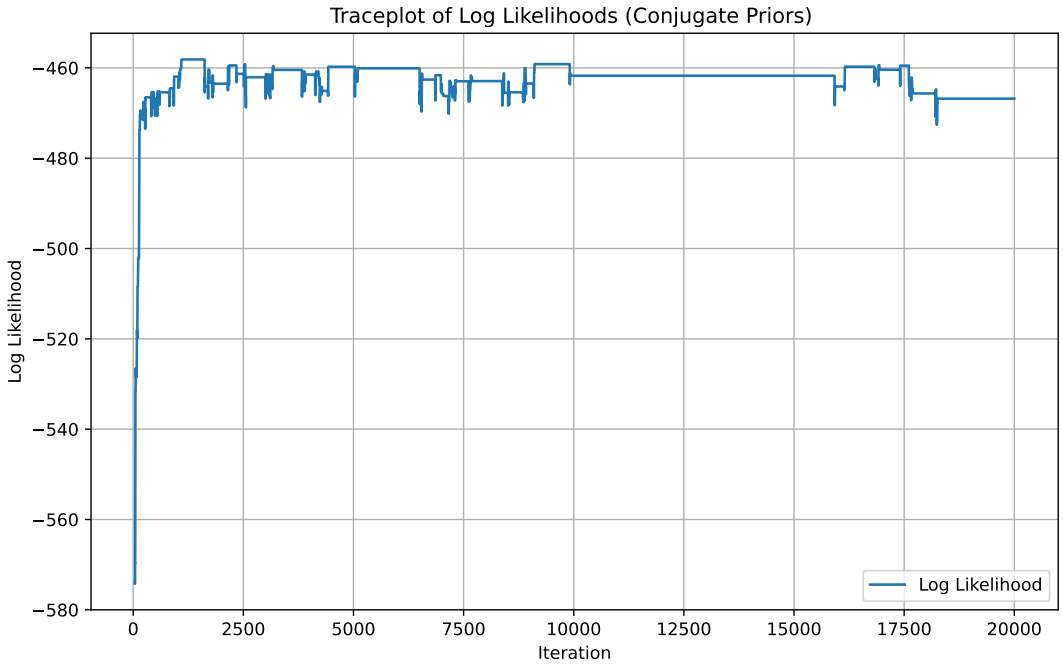

Figure 25: Trace plot of running SA-EBM on the mixed-pathology data

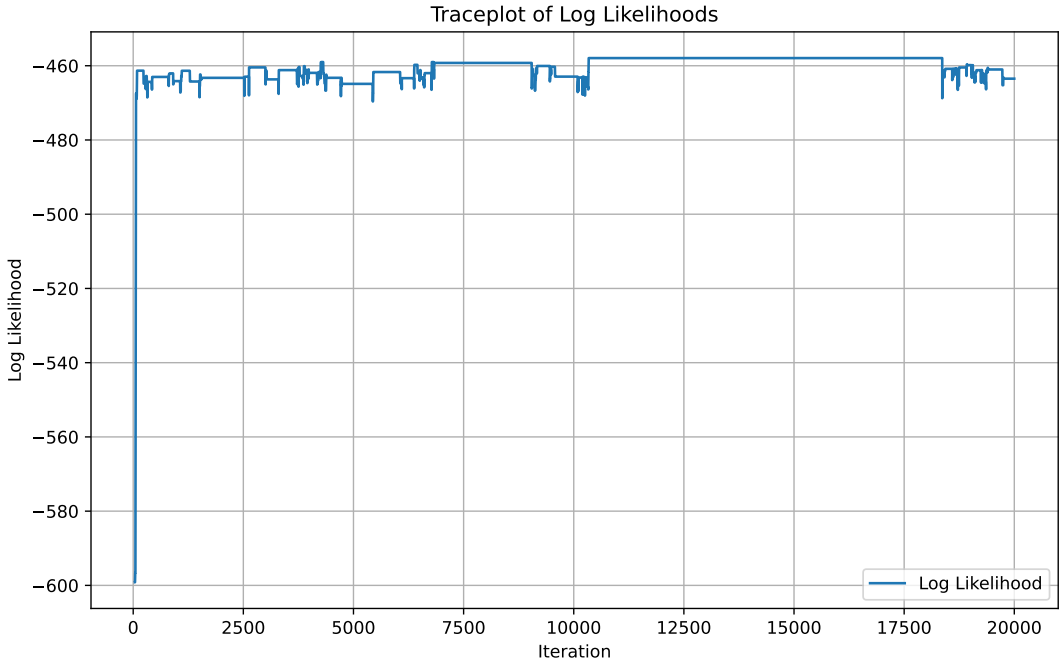

Figure 26: Trace plot of running JPM-Mallows on the mixed-pathology data