
Disentangling Latent Risk Pathways via Bayesian Hypergraph Inference

Anonymous Authors¹

Abstract

Electronic health records (EHR) pose large-scale multi-disease modeling problems in which many outcomes are rare and strongly influenced by shared risk factors. While modern approaches achieve strong predictive performance, they often treat diseases independently or rely on black-box architectures, offering limited insight into how risk factors organize disease risk and little principled uncertainty quantification. We introduce a Bayesian hypergraph inference framework that reframes multi-disease modeling around **latent, risk-factor-modulated disease pathways**. Risk factors act on hyperedges, latent disease subsets with shared risk patterns, allowing diseases to participate in multiple distinct pathways and enabling interpretable, higher-order structure beyond pairwise associations. A repulsion prior encourages parsimonious and identifiable structure, while posterior inference provides calibrated uncertainty over both disease groupings and risk-factor influence. To enable scalable inference on large EHR datasets, we develop a structured variational inference algorithm that preserves logical dependencies among hyperedge existence, disease membership, and pathway-level effects. Experiments on simulated data and the UK Biobank demonstrate stable and interpretable disease pathway structure, well-calibrated uncertainty, improved estimation for rare diseases, and competitive predictive performance.

1. Introduction

Electronic health records (EHR) enable large-scale modeling of disease risk across populations and many outcomes simultaneously. In this regime, individuals are often susceptible to multiple conditions, disease prevalence varies

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

widely from common chronic disorders to rare diseases with few observed cases, and shared biological pathways, behavioral exposures, and social determinants induce complex dependencies across diseases (Goh et al., 2007; Skou et al., 2022). Understanding how disease risk is organized across populations is therefore central to epidemiology, precision medicine, and population health.

Crucially, these dependencies among diseases are **not uniform across risk factors**: different risk drivers organize diseases in different ways. For example, age may jointly increase risk for cardiovascular and metabolic diseases, while smoking predominantly affects respiratory and oncological conditions (Murray et al., 2020). These disease groupings are overlapping, inherently uncertain, and specific to the risk factor under consideration, reflecting distinct etiological patterns through which different exposures are associated with disease risk (Menche et al., 2015).

Modeling disease risk in this multi-disease setting poses two fundamental challenges. First, disease prevalence exhibits a pronounced long-tail structure: while some chronic conditions are common, most diseases are rare, providing insufficient data for stable independent estimation. Second, latent etiological structure is complex and higher-order: diseases can participate in multiple, risk-factor-specific pathways, requiring models to share information across outcomes without collapsing distinct latent mechanisms. A successful solution must therefore balance statistical efficiency with structured inductive bias, enabling interpretability through the attribution of risk while explicitly representing uncertainty arising from limited and heterogeneous data. The core challenge is not merely prediction, but **learning risk-factor-specific, overlapping latent structure over disease outcomes with uncertainty**.

Existing approaches address aspects of multi-disease modeling, but each falls short of this goal in critical ways. Independent disease-specific models, such as logistic regression, are transparent but treat diseases in isolation, ignoring shared structure and yielding poorly calibrated uncertainty for rare outcomes (King & Zeng, 2001; Heinze & Schemper, 2002). Conversely, multitask learning and joint modeling approaches improve robustness through information sharing (Caruana, 1997; Zhang & Yang, 2022), but typically rely on black-box architectures that entangle all risk fac-

tors into a single latent space. This obscures the specific pathways through which different risk drivers exert their effects and provides limited insight into uncertainty over the learned structure. Finally, more structured probabilistic models capture outcome-level dependence (Blumm et al., 2009; Menche et al., 2015), but often focus on marginal correlations aggregated across all factors rather than disentangling risk-factor-specific disease organization, and can be difficult to scale to modern EHR datasets. As a result, existing methods cannot answer a central epidemiological question: **through which shared disease pathways does a given risk factor exert its effects, and how uncertain are we about that structure?**

To address this gap, we introduce a **Bayesian Hypergraph Pathway Inference (BHPI)** framework that reframes multi-disease modeling as the discovery of latent, risk-factor-modulated disease pathways (Figure 1). Our key insight is representational: while standard graphs and multitask models capture pairwise correlations or entangled shared effects, etiological pathways are inherently higher-order, involving groups of diseases rather than isolated pairs. We therefore model disease pathways as hyperedges in a latent hypergraph, where nodes represent diseases, and each hyperedge captures a shared pattern of risk-factor influence. Risk factors act directly on hyperedges to induce structured effects across groups of diseases. This decouples risk-factor influence from individual outcomes and naturally supports overlapping, factor-specific disease organization. Structurally, we introduce a repulsion prior that promotes parsimonious pathway discovery by discouraging redundant or highly overlapping hyperedges associated with the same risk factor. This prior mitigates latent structure collapse and plays a critical role in stabilizing inference in the long-tail regime, where weak signals from rare diseases would otherwise lead to degenerate solutions. Posterior inference in this model is challenging due to strong logical dependencies among hyperedge existence, disease membership, and hyperedge-level effects. Standard mean-field variational inference (VI) breaks these dependencies, leading to poorly calibrated uncertainty. We therefore develop a scalable structured VI algorithm that explicitly preserves these couplings, enabling uncertainty-aware learning of disease pathways and risk propagation in large EHR datasets.

We summarize our contributions as follows:

- We introduce a Bayesian hypergraph pathway inference framework with a repulsion prior, enabling parsimonious learning of risk-factor-specific, higher-order disease structure.
- We develop a scalable structured VI algorithm that preserves the logical dependencies (existence \rightarrow membership \rightarrow effect), overcoming known limitations of standard mean-field approximations.

- We validate the framework on synthetic data and UK Biobank, demonstrating stable and interpretable pathway recovery, improved rare-disease risk estimation, and well-calibrated uncertainty.

2. Related Work

Independent Multi-Disease Models A common baseline for multi-disease risk modeling is to train separate models for each outcome, typically using regularized logistic regression (Hoerl & Kennard, 1970; Tibshirani, 1996; Zou & Hastie, 2005). Specific penalized (Firth, 1993; King & Zeng, 2001; Heinze & Schemper, 2002) and Bayesian (Gelman et al., 2008) extensions further improve numerical stability in low-prevalence settings. But these approaches fundamentally treat diseases as independent prediction tasks, therefore failing to borrow statistical strength for rare outcomes.

Multi-task Learning and Representation-based Models Multi-task learning (MTL) methods share information via common representations across related disease prediction tasks, using shared latent factors (Caruana, 1997; Evgeniou & Pontil, 2004; Argyriou et al., 2008; Standley et al., 2020), hierarchical priors (Bonilla et al., 2007), or neural network architectures (Miotto et al., 2016; Choi et al., 2016; Rajkumar et al., 2018; Harutyunyan et al., 2019; Rasmy et al., 2021) to jointly model multiple disease risks. While these approaches can implicitly capture commonalities across outcomes, the learned representations are typically implicit and black-box, obscuring the specific grouping of diseases and offering limited uncertainty quantification over the structure itself.

Structured Outcome Learning A complementary line of work focuses on modeling statistical dependence among diseases through disease networks or comorbidity structures. Such models characterize patterns of disease co-occurrence in EHR data and have been used for descriptive analysis, risk stratification, and hypothesis generation (Blumm et al., 2009; Barabási et al., 2011; Jensen et al., 2014; Menche et al., 2015). However, these static networks aggregate all risk factors into a single correlation structure, failing to disentangle how different inputs drive different dependencies. Probabilistic approaches like multi-label learning (Tsoumakas & Vlahavas, 2007; Read et al., 2011) or correlated topic models (Blei et al., 2006; Li & McCallum, 2006) capture outcome dependence but often lack the feature-specific modulation required to understand etiology.

Hypergraph Representation Learning Hypergraphs have emerged as a powerful tool for modeling higher-order correlations in data (Zhou et al., 2006; Feng et al., 2019; Yadati et al., 2019; Antelmi et al., 2023). Recent work in biomedical ML has applied hypergraphs to patient-concept

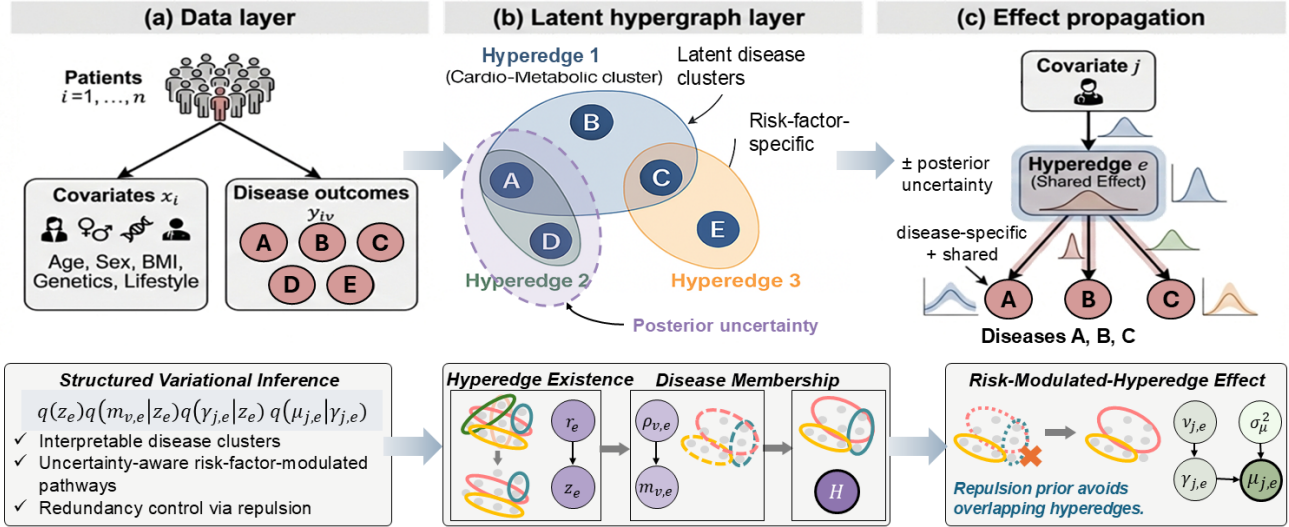


Figure 1. Bayesian Hypergraph Pathway Inference Workflow. (a) Input data consist of patient covariates and multiple disease outcomes. (b) A latent disease hypergraph models higher-order structure via hyperedges, allowing diseases to participate in multiple pathways with uncertain existence and membership. (c) Risk factors act on hyperedges to induce structured effects across diseases, yielding disentangled attribution. Shaded regions denote posterior uncertainty. Bottom panels: Structured Variational Inference (left) preserves logical dependencies to learn the latent topology including hyperedge existence and membership (middle), while a repulsion prior (right) discourages redundant pathways to ensure a parsimonious and interpretable latent structure.

interactions (Xu et al., 2023; Zhang et al., 2024). However, existing hypergraph neural networks generally assume the hypergraph structure is observed or inferred via heuristic similarity metrics. Our work differs by treating the hypergraph topology as a latent random variable to be inferred within a fully Bayesian generative framework, specifically designed to disentangle risk-factor modulation.

3. Bayesian Hypergraph Pathway Model

We propose a Bayesian hypergraph pathway model for learning **latent, higher-order output structure** in multi-disease prediction. The core object is a latent disease hypergraph whose hyperedges encode overlapping disease pathways, each selectively modulated by input features. Rather than modeling outcomes independently or through a single shared representation, our model learns which diseases are linked through which pathways, and which inputs activate each pathway, with calibrated uncertainty over both the latent structure and its effect propagation on outcomes.

3.1. Problem Setup

For each subject $i = 1, \dots, N$, we observe a feature vector $\mathbf{x}_i \in \mathbb{R}^P$ and binary outcome $Y_{i,v} \in \{0, 1\}$ over a fixed set of V diseases. Outcomes are defined only when the subject is at risk for disease v , indicated by $R_{i,v} = 1$, and all modeling and inference are restricted to valid subject-disease pairs. Our objectives are twofold: (i) to predict disease risk across many correlated outcomes, including low-

prevalence conditions; and (ii) to recover latent, overlapping disease pathways that explain how input features jointly influence multiple outcomes.

3.2. Observation Model

We link latent pathway structure to observed outcomes through a multivariate logistic observation model:

$$y_{i,v} \mid \tilde{\eta}_{i,v} \sim \text{Bernoulli}(\sigma(\tilde{\eta}_{i,v})), \quad \tilde{\eta}_{i,v} = \alpha_v + \mathbf{x}_i^\top \beta_v, \quad (1)$$

where $\alpha_v \sim \mathcal{N}(0, \varrho_v^2)$ is a disease-specific baseline risk and $\beta_v \in \mathbb{R}^P$ captures risk-factor influences. Equation (1) serves as a measurement model linking latent structure to observed outcomes. Rather than estimating disease-specific effects β_v independently, we impose structured sharing of feature effects across diseases via a latent hypergraph, enabling information borrowing and interpretable higher-order representation over the output space.

3.3. Latent Hypergraph Representation

We represent latent disease pathways using a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V} = \{1, \dots, V\}$ correspond to diseases and each hyperedge $e \in \mathcal{E}$ represents shared response patterns to input features across subsets of outcomes. Hyperedges may overlap, allowing diseases to participate in multiple pathways. The hypergraph is encoded by an incidence matrix $H \in \{0, 1\}^{V \times E}$, where $H_{v,e} = \mathbb{I}\{\text{disease } v \text{ belongs to pathway } e\}$, and E is a fixed upper bound on the number of hyperedges. This representation captures a higher-order output structure that

cannot be expressed using pairwise graphs or global shared representations.

3.4. Hypergraph-Induced Feature Effects

Disease-specific feature effects are induced by hyperedge-level effects:

$$\beta_{j,v} = d_v^{-1} \cdot \sum_{e=1}^E H_{v,e} \mu_{j,e}, \quad (2)$$

where $\mu_{j,e}$ denotes the effect of risk factor j on hyperedge e . The normalizing constant $d_v = E^{-1/2}$ stabilizes the variance of induced effects and ensures risk magnitudes remain consistent as E increases. This construction induces input-modulated, higher-order dependencies among outputs: a disease may be influenced by different features through distinct pathways, while a single feature may act on multiple disease subsets.

3.5. Hypergraph Structure Learning

Learning the hypergraph requires a combinatorial search over both pathway existence and disease membership. We adopt a hierarchical sparsity model in which each hyperedge e has a binary existence indicator z_e , and conditional on existence, disease-level membership indicators $m_{v,e}$. The incidence matrix is defined as $H_{v,e} = z_e \cdot m_{v,e}$, with structured priors:

$$\begin{aligned} z_e | r_e &\sim \text{Bernoulli}(r_e), & r_e &\sim \text{Beta}(a_r, b_r), \\ m_{v,e} | z_e, \rho_{v,e} &\sim (1 - z_e) \cdot \delta_0 + z_e \cdot \text{Bernoulli}(\rho_{v,e}), \\ \rho_{v,e} &\sim \text{Beta}(a_\rho, b_\rho). \end{aligned}$$

This hierarchy separates global hyperedge discovery from within-hyperedge disease composition, yielding parsimonious and interpretable latent structure.

3.6. Sparse Features-Modulated Hyperedge Effects

Not all features act on all hyperedges. For each feature j , we introduce a spike-and-slab prior over hyperedge-level effects:

$$\begin{aligned} \mu_{j,e} | \gamma_{j,e}, \sigma_\mu^2 &\sim (1 - \gamma_{j,e})\delta_0 + \gamma_{j,e} \cdot \mathcal{N}(0, \sigma_\mu^2), \\ \sigma_\mu^2 &\sim \text{Inv-Gamma}(a_\mu, b_\mu), \end{aligned}$$

where $\gamma_{j,e}$ selects whether feature j influences hyperedge e . To avoid degenerate solutions in which multiple hyperedges provide redundant explanations for the same feature, we introduce a repulsion prior that penalizes selecting highly overlapping hyperedges for a given feature j :

$$\mathcal{R}_{\text{rep}}(\gamma_j | H) \propto \exp \left\{ -\lambda \sum_{e_1 < e_2} O(S_{e_1}, S_{e_2}) \cdot \gamma_{j,e_1} \gamma_{j,e_2} \right\},$$

$$O(S_{e_1}, S_{e_2}) = \frac{|S_{e_1} \cap S_{e_2}|}{\min(|S_{e_1}|, |S_{e_2}|)},$$

where $S_e = \{v : H_{v,e} = 1\}$, λ controls repulsion strength, and $O(S_{e_1}, S_{e_2}) \in [0, 1]$ is the overlap coefficient between hyperedges e_1 and e_2 , in which $O(S_{e_1}, S_{e_2}) = 1$ means complete overlaps. This encourages a disentangled and identifiable pathway structure within a single feature while allowing overlap hyperedges across different features. Moreover, we enforce the hierarchy $\gamma_{j,e} = 0$ whenever $z_e = 0$, ensuring that risk factors only select globally active hyperedges. Together, the joint prior on the risk-factor-modulated hyperedge effect selector is:

$$\begin{aligned} \gamma_j | \nu_j, H &\sim \prod_{e=1}^E \left\{ z_e \cdot \text{Bernoulli}(\gamma_{j,e}; \nu_{j,e}) \cdot \mathcal{R}_{\text{rep}}(\gamma_j | H) \right. \\ &\quad \left. + (1 - z_e) \delta_0(\gamma_{j,e}) \right\}, \end{aligned}$$

$$\nu_{j,e} \sim \text{Beta}(a_\nu, b_\nu).$$

This hierarchy enforces the logical constraints: $z_e = 0 \Rightarrow m_{v,e} = \gamma_{j,e} = \mu_{j,e} = 0$, and $\gamma_{j,e} = 0 \Rightarrow \mu_{j,e} = 0$, ensuring coherent structure learning and improving identifiability without sacrificing predictive performance.

4. Scalable Structured Variational Inference

Posterior inference is challenging due to the combination of a non-conjugate likelihood and a combinatorial latent hypergraph with hard logical constraints linking discrete structure variables and continuous effects. These constraints induce strong posterior dependencies that violate the independence assumptions underlying standard mean-field factorization (e.g., $q(z_e) \prod_{v=1}^V q(m_{v,e})$), which assigns non-zero probability to configurations like $\{z_e = 0, m_{v,e} = 1\}$, a logical impossibility in our generative process. Preserving these dependencies is essential not only for uncertainty calibration but also for enforcing the repulsion-based identifiability constraints during inference.

4.1. Structured Variational Family

We propose a structured VI scheme that explicitly preserves the logical hierarchy while retaining computational tractability. We define the variational family as:

$$\begin{aligned} \mathcal{Q} = & \prod_v q(\alpha_v) \prod_e q(z_e) q(r_e) \prod_{v,e} q(m_{v,e} | z_e) q(\rho_{v,e}) \\ & \prod_{j,e} q(\mu_{j,e} | \gamma_{j,e}) q(\gamma_{j,e} | z_e) q(\nu_{j,e}) \prod_{i,v} q(\omega_{i,v}) q(\sigma_\mu^2). \end{aligned}$$

Key structural properties: (1) **Conditional Dependence**: the posteriors for disease membership $m_{v,e}$ and feature effects $\gamma_{j,e}, \mu_{j,e}$ are conditioned on the hyperedge existence z_e ; (2) **Zero-Preservation**: if $q(z_e)$ concentrates on 0, the conditional factors $q(m_{v,e} | z_e = 0)$, $q(\gamma_{j,e} | z_e = 0)$ and $q(\mu_{j,e} | z_e = 0)$ collapse to Dirac deltas at 0, correctly removing inactive pathways consistently from the model. Overall,

this structured variational formulation respects the hierarchical design of the generative model while remaining scalable to large EHR cohorts.

4.2. Optimization via Latent Variable Augmentation

We approximate the posterior by minimizing the Kullback-Leibler (KL) divergence between the variational family and the true posterior, equivalently maximizing the evidence lower bound (ELBO). To handle the non-conjugate logistic likelihood, we utilize Pólya–Gamma (PG) augmentation (Polson et al., 2013), introducing variables $\omega_{i,v}$ that render the likelihood conditionally Gaussian. Given the structural dependencies in BHPI, we employ Coordinate Ascent Variational Inference (CAVI, (Bishop & Nasrabadi, 2006)). While the variational family admits a product form, the optimal update for a factor $q_j(\theta_j)$ must preserve conditional dependencies within the prior. The general update rule is:

$$q^*(\theta_j) \propto \exp \left[\mathbb{E}_{q(\Theta \setminus \theta_j)} \{ \log p(\Theta, \mathcal{Y}) \} \right].$$

For factors with internal logical dependencies, we use structured updates that account for the conditional entropy within the variational family; see Appendix A.3 for the full derivation. The variational posterior factors for BHPI take the following forms:

$$\begin{aligned} q(z_e) &= \text{Bernoulli}(r_e^*), \\ q(m_{v,e} | z_e) &= (1 - z_e)\delta_0(m_{v,e}) + z_e \text{Bernoulli}(\rho_{v,e}^*), \\ q(\gamma_{j,e} | z_e) &= (1 - z_e)\delta_0(\gamma_{j,e}) + z_e \text{Bernoulli}(\nu_{j,e}^*), \\ q(\mu_{j,e} | \gamma_{j,e}) &= (1 - \gamma_{j,e})\delta_0(\mu_{j,e}) + \gamma_{j,e} \mathcal{N}(\mu_{j,e}^*, \sigma_{j,e}^{2*}), \\ q(\alpha_v) &= \mathcal{N}(\alpha_v^*, \varrho_v^{2*}), \quad q(\omega_{i,v}) = \text{PG}(1, \tilde{\eta}_{i,v}^*), \\ q(r_e) &= \text{Beta}(a_e^{*(r)}, b_e^{*(r)}), \quad q(\rho_{v,e}) = \text{Beta}(a_{v,e}^{*(\rho)}, b_{v,e}^{*(\rho)}), \\ q(\nu_{j,e}) &= \text{Beta}(a_{j,e}^{*(\nu)}, b_{j,e}^{*(\nu)}), \quad q(\sigma_\mu^2) = \text{Inv-Gamma}(a_\mu^*, b_\mu^*). \end{aligned}$$

The quantities $r_e^*, \rho_{v,e}^*, \nu_{j,e}^*, \mu_{j,e}^*, \sigma_{j,e}^{2*}, \alpha_v^*, \varrho_v^{2*}, \tilde{\eta}_{i,v}^*, a_e^{*(r)}, b_e^{*(r)}, a_{v,e}^{*(\rho)}, b_{v,e}^{*(\rho)}, a_{j,e}^{*(\nu)}, b_{j,e}^{*(\nu)}, a_\mu^*, b_\mu^*$ are the variational parameters optimized to approximate the true posterior distribution. The optimized variational parameters aggregate evidence across the hierarchy. For instance, the update for $q(z_e)$ serves as a “global switch” that prunes redundant hyperedges by consolidating evidence from downstream disease-membership and feature-sparsity terms.

Repulsion-Aware Variational Updates A unique challenge arises in updating the feature–hyperedge activation variables $\{\gamma_{j,e}\}$. Although $q(\gamma_{j,e} | z_e)$ is Bernoulli, its parameter $\nu_{j,e}^*$ is coupled across hyperedges by the repulsion prior $\mathcal{R}_{\text{rep}}(\gamma_j | H)$. This transforms the update into a competitive selection process where hyperedges associated with the same risk factor j compete for activation. Under CAVI, the optimal log-odds for $\nu_{j,e}^*$ incorporates the expected log-prior contribution from all competing hyperedges $\{e' \neq e\}$.

Algorithm 1 Repulsion-Aware Coordinate Ascent (BHPI)

Input: Data \mathcal{Y} , max hyperedges E , repulsion strength λ .
Initialize: Variational parameters.
while ELBO not converged **do**
 // 1. Augmentation & Baseline Updates
 Update Polya-Gamma variables $\tilde{\eta}_{i,v}^*$ and baseline intercepts $\alpha_v^*, \varrho_v^{2*}$ (Equations (9) and (10))
 // 2. Update Latent Risk-Hypergraph Structure
 for each hyperedge $e = 1, \dots, E$ **do**
 Update risk-factor effects $\mu_{j,e}^*, \sigma_{j,e}^{2*}$ when $\gamma_{j,e} = 1$ (Equations (11) and (12))
 Compute expected overlap $\mathbb{E}_q [O(S_e, S_{e'})]$
 Update effect activation $\nu_{j,e}^*$ (Equation (13))
 Update disease memberships $\rho_{v,e}^*$ (Equation (14))
 Update hyperedge existence r_e^* (Equation (15))
 end for
 // 3. Hyperparameter Update
 Update $a_e^{*(r)}, b_e^{*(r)}, a_{v,e}^{*(\rho)}, b_{v,e}^{*(\rho)}, a_{j,e}^{*(\nu)}, b_{j,e}^{*(\nu)}, a_\mu^*, b_\mu^*$
end while
return Optimized variational posterior distributions.

Concretely, we replace the discrete combinatorial overlap with its variational expectation $\mathbb{E}_q [O(S_e, S_{e'})]$. This term acts as a repulsive penalty, suppressing the simultaneous activation of multiple hyperedges that cover redundant disease pathways for the same feature, which ensures the discovery of a parsimonious and identifiable latent structure. The complete coordinate ascent procedure is summarized in Algorithm 1, with full mathematical derivations for each update rule provided in Appendix A.4. The per-iteration complexity of BHPI is approximately $\mathcal{O}(N \cdot E \cdot (P + V))$, ensuring that the computational cost scales linearly with the number of samples and the dimensions of the latent hypergraph.

5. Experiments

We evaluate the proposed BHPI framework to assess its ability to recover latent disease hypergraph structure modulated by shared risk factors, while maintaining accurate disease risk prediction. While predictive performance can be evaluated on held-out outcomes, assessment of latent hypergraph structure requires access to ground truth, which is unavailable in real-world EHR data. We therefore rely on simulations to evaluate both predictive accuracy and structural recovery under controlled conditions.

5.1. Simulation Design and Baselines

We generate a latent disease hypergraph with $V = 30$ diseases and $E = 5$ hyperedges, where each hyperedge corresponds to a latent disease pathway driven by shared risk-factor influences. Diseases may participate in multiple hyperedges, reflecting overlapping risk mechanisms.

Table 1. Predictive performance on simulated data. Mean (standard deviation) AUC on held-out test sets ($\times 100$). BHPI achieves competitive disease risk prediction while simultaneously inferring latent disease pathways.

MODEL	$N = 2000$	$N = 5000$
BHPI	75.00 (0.83)	74.63 (0.46)
OPTIMAL LOGISTIC	73.86 (0.83)	74.15 (0.49)
LGBM	68.50 (0.94)	69.68 (0.56)
BINARY RELEVANCE	71.50 (0.94)	72.61 (0.64)
CLASSIFIER CHAIN	71.25 (1.03)	72.63 (0.63)
RAkELD	70.99 (0.95)	72.50 (0.60)

Table 2. Recovery of latent risk structure on simulated data ($N = 2000$). Metrics include agreement between true and inferred disease-level risk contributions β , AUC for recovering hypergraph structure H , and risk-factor-hyperedge associations γ ($\times 100$).

MODEL	$\text{COR}(\beta, \hat{\beta})$	H -AUC	γ -AUC
BHPI	97.85	96.96	97.85
OPTIMAL LOGISTIC ¹	95.65	-	-

Risk-factor influences on latent hypergraphs $\gamma \in \mathbb{R}^{P \times E}$ are sparse: the first predictor affects multiple hyperedges, while each remaining predictor influences a single hyperedge. Nonzero influences are drawn from $\mathcal{N}(\mu, 0.5^2)$, with $\mu \in \{1, 1.5, 2\}$. Risk factors are independently sampled from a standard normal distribution. Disease outcomes are generated from the hypergraph-induced logistic model (Equations 1 and 2). We evaluate sample sizes $N = 2000$ and $N = 5000$. For each setting, we perform 50 independent replicates using a 60/20/20 split for training, validation, and testing. Hyperparameters are selected on the validation set, and all reported predictive metrics are computed on held-out test data. We compare BHPI against (1) optimally tuned logistic regression (ridge, lasso, elastic-net), (2) LightGBM (LGBM), and (3) standard multi-label learning baselines: Binary Relevance, Classifier Chains, and RAkELD. These baselines represent independent, autoregressive, and higher-order label-interaction modeling strategies, respectively. While they are strong predictive competitors, they do not infer latent disease pathways or quantify uncertainty over shared mechanisms.

5.2. Predictive Performance and Structural Recovery

Table 1 shows that BHPI achieves predictive performance comparable to optimally tuned logistic regression and consistently outperforms tree-based and multi-label baselines. Despite its structured inductive bias, BHPI does not sacrifice predictive accuracy. The inferred disease-level risk contributions closely match the ground truth, with correlations exceeding 97%. Beyond prediction, BHPI accu-

¹ H and γ recovery are not defined for logistic regression.

Table 3. Role of the repulsion prior in hypergraph identifiability ($N = 2000$). Predictive performance remains stable, while repulsion reduces redundancy and improves stability of inferred hyperedges, as measured by overlap coefficients (Ov.) across replicates, effective hyperedge (Eff.#/RF) per risk factor (RF), and hyperedge overlap per RF (Ov./RF).

λ	AUC	Ov.	EFF.#/RF	Ov./RF
$= 0$	74.69	0.800	1.992	0.097
> 0	75.00	0.812	1.874	0.082

rately recovers latent hypergraph structure. The inferred disease-hyperedge incidence matrices closely match the ground truth, with AUCs exceeding 95% (Table 2). Risk-factor-hyperedge associations are also recovered with high fidelity, indicating that BHPI disentangles overlapping disease pathways rather than collapsing them into pairwise associations. In contrast, independent logistic regression cannot recover latent hypergraph structure by construction. A visualization of the true and inferred incidence matrices for a representative replicate is provided in Appendix B.1. This comparison demonstrates that the inferred hypergraph closely matches the true block structure with limited redundancy, identifying risk-factor-modulated disease pathways rather than collapsing them into pairwise associations.

5.3. Ablation of the Repulsion Prior

We examine the role of the repulsion prior by setting $\lambda = 0$, which allows a single risk factor to select highly overlapping hyperedges. Removing the repulsion prior has negligible impact on predictive accuracy, indicating that repulsion is not required for strong prediction. However, it substantially increases redundancy and instability in the inferred hypergraph, as reflected by reduced overlap coefficients across replicates, greater hyperedge overlap, and a larger effective number of active hyperedges per risk factor (Table 3). These results demonstrate that the repulsion prior primarily improves identifiability and interpretability of latent disease pathways, rather than acting as a predictive regularizer.

5.4. Sensitivity Analysis

We evaluate the robustness of BHPI to the repulsion strength λ , and the assumed maximum number of latent hyperedges E . Across a wide range of λ values, predictive performance remains stable, indicating that BHPI does not rely on fine-tuned regularization. Increasing λ sharpens posterior inclusion probabilities (PIPs) and reduces redundancy between inferred hyperedges, leading to more stable latent structures across random splits. We further examine robustness to over-specification of model capacity by varying E , which shows that even when E is over-specified, the effective number of active hyperedges remains stable, demonstrating effective self-regularization. Detailed sensitivity curves and

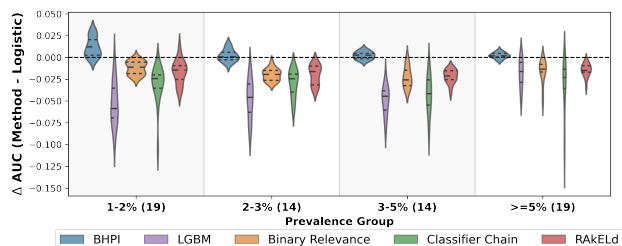


Figure 2. Predictive stability via structured information sharing. Distribution of per-disease ΔAUC scores (relative to tuned Logistic Regression) stratified by disease prevalence. While baselines like LGBM and Classifier Chains exhibit high variance and instability for rare diseases, BHPI maintains robust performance by borrowing strength across shared latent pathways.

structural diagnostics are provided in Appendix B.3.

5.5. UK Biobank: Disentangling Real-World Disease Pathways

We evaluate BHPI on the UK Biobank (UKB), a large prospective cohort study linked to electronic health records (Bycroft et al., 2018). UKB serves as a challenging benchmark for multi-disease modeling due to its high-dimensional feature space, extreme class imbalance, and the complex dependency structure among chronic conditions. After quality control, the final analysis cohort comprises $N \approx 277,291$ participants. Additional dataset characteristics are summarized in Table 5. We consider $P = 71$ baseline risk factors spanning demographics, biomarkers, lifestyle exposures, and socioeconomic indicators. Disease outcomes comprise $V = 66$ chronic conditions derived from ICD-10 codes and mapped to Phecodes to improve phenotypic coherence and reduce coding noise (Wu et al., 2019; Wei et al., 2017). Disease prevalence ranges from common to extremely rare, placing the analysis firmly in the rare-disease regime. We set the maximum number of latent hyperedges to $E = 60$ to allow sufficient flexibility for capturing heterogeneous and overlapping disease pathways. Model performance is evaluated using stratified 60/20/20 train-validation-test splits, with results averaged over 100 independent random splits.

Predictive Stability in the Rare-Disease Regime We first assess whether BHPI improves risk estimation without sacrificing predictive accuracy. Figure 2 reports per-disease ΔAUC relative to optimally tuned logistic regression, stratified by disease prevalence. While performance on common diseases is largely saturated across all methods, a clear divergence emerges for low-prevalence conditions. Discriminative baselines, such as LightGBM and Classifier Chains, exhibit high-variance failure modes as evidenced by long lower tails in the ΔAUC distribution. In contrast, BHPI consistently matches or exceeds the baseline, suggesting that BHPI stabilizes estimation for rare diseases by linking them to related conditions through shared latent structure.

Disentangling Risk-Modulated Pathways with Posterior Uncertainty To understand the structural mechanisms underlying the observed predictive stability, we first examine how BHPI disentangles latent disease pathways while enforcing parsimony. Figure 3(left) visualizes posterior uncertainty over the learned latent structure. The disease-hyperedge matrix exhibits a sparse, approximately block-structured pattern with overlapping memberships, indicating that diseases cluster into localized latent pathways while remaining allowed to participate in multiple contexts. In contrast, the risk-factor-hyperedge matrix reveals selective, pathway-specific activation patterns, highlighting that different predictors modulate distinct subsets of latent pathways rather than acting globally. Together, these patterns demonstrate that BHPI disentangles risk-factor-specific disease organization while explicitly quantifying uncertainty over both pathway composition and modulation.

As posterior mass concentrates, this uncertainty resolves into a compact global structure. Although initialized with $E = 60$ candidate hyperedges, the posterior concentrates on a parsimonious solution, with an expected 27.2 active hyperedges (28 with $\text{PIP} > 0.5$). As summarized in Figure 3(right), most hyperedges involve 2–12 diseases, while individual diseases participate in multiple pathways, reflecting heterogeneous risk contexts. On the predictor side, individual risk factors act on a limited number of pathways. This structured sparsity distinguishes BHPI from dense latent factor models that conflate unrelated mechanisms and hinder interpretability. Figure 4(left) provides a global view of how risk factors propagate through latent pathways to affect multiple diseases. The resulting structure is modular, overlapping and sparse, offering an interpretable summary of heterogeneous disease etiology learned from EHR data.

Pathway-Level Illustration To concretely demonstrate how our framework translates global posterior structure into interpretable mechanisms, we visualize representative latent pathways and their modulating risk factors in Figure 4(right). This case study highlights three core properties of BHPI: *sparse disease membership*, *selective risk-factor activation*, and *graded posterior uncertainty* encoded directly through learned PIPs.

(a) Divergent risk pathways (mechanism disentanglement): A single risk factor may influence multiple, non-redundant disease pathways. For example, **Smoking** is inferred as a modulator for multiple non-redundant hyperedges. BHPI disentangles its impact into distinct biological contexts: a respiratory injury pathway grouping *Emphysema*, *Asthma*, *Chronic airway obstruction* and *Chronic Bronchitis*, and a separate mucosal damage pathway linking *Ulcers*. The repulsion prior prevents these pathways from collapsing into a single dense factor, avoiding the entanglement typical of standard multitask models.

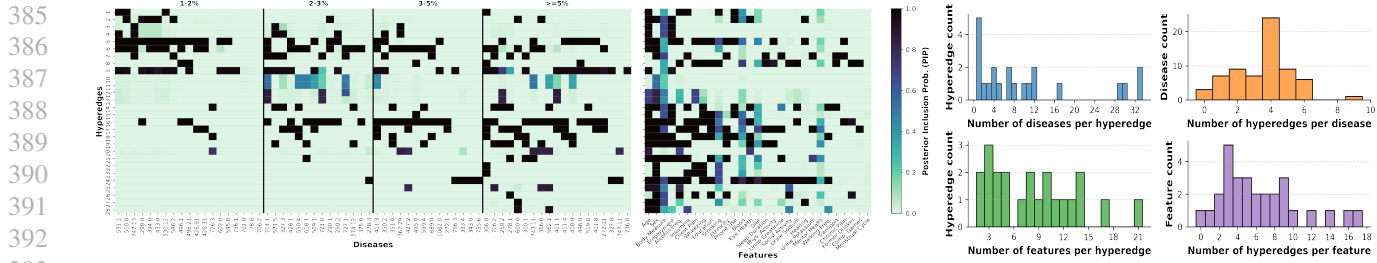


Figure 3. Disentanglement and parsimony of risk-modulated pathways. (Left) PIPs for disease–pathway membership and risk-factor activation, revealing sparse, overlapping disease organization and pathway-specific modulation with quantified uncertainty. (Right) Summary of the thresholded structure ($\text{PIP}(z_e = 1) > 0.5$), showing compact pathway sizes, limited pathway participation per disease and risk factor, and overall structural parsimony.

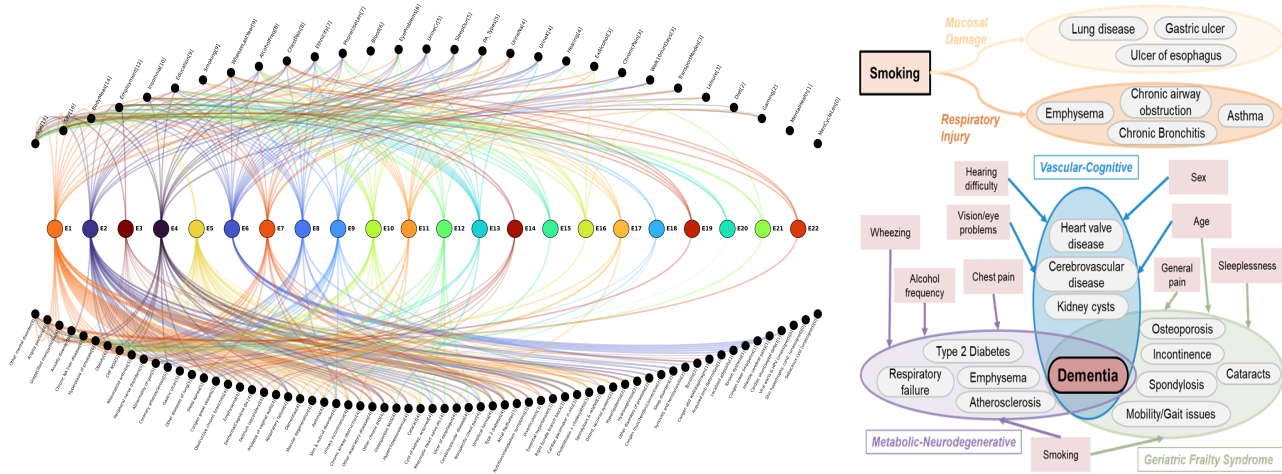


Figure 4. Global visualization of the learned risk-factor-to-disease pathway structure (hyperedges with $\text{PIP}(z_e = 1) > 0.5$) (Left). Risk factors (top) connect to latent pathways (middle), which in turn connect to diseases (bottom). Edge widths correspond to PIPs, illustrating uncertainty, overlap, and modular risk propagation. Node annotations report the number of incident pathways. Pathway-level illustration (Right). Divergent risk pathways for smoking (top) and convergent disease pathways involving Dementia (bottom).

(b) Convergent disease contexts (heterogeneity discovery): Conversely, BHPI captures disease heterogeneity by allowing a single disease to participate in multiple pathways. For instance, **Dementia** belongs simultaneously to: (i) Vascular-cognitive pathways: shared with *Cerebrovascular disease* and *Heart valve disease*, modulated by *Age*, *Sex*, and sensory impairment, indicating vascular etiology; (ii) Geriatric frailty pathway: grouped with *Osteoporosis*, *Incontinence*, and *Mobility issues*, driven by *Age*, *General pain*, reflecting systemic decline; (iii) Metabolic-neurodegenerative pathway: linked with *Type 2 Diabetes* and *Atherosclerosis*, modulated by *Alcohol frequency* and *Smoking*, consistent with metabolic exhaustion mechanisms. These overlapping memberships demonstrate how BHPI represents heterogeneous disease etiology without enforcing mutually exclusive clustering.

6. Conclusion and Discussion

We introduced **Bayesian Hypergraph Pathway Inference (BHPI)**, a structured framework for multi-disease modeling

that discovers latent, risk-factor–modulated disease pathways. By representing pathways as hyperedges and allowing risk factors to act directly on this structure, BHPI captures higher-order and overlapping disease organization that is inaccessible to pairwise or independent outcome models. Across simulations and the UK Biobank, BHPI demonstrates stable prediction in the rare-disease regime, identifies parsimonious and non-redundant pathways via a repulsion prior, and provides calibrated posterior uncertainty over both pathway composition and risk-factor modulation through structured variational inference, offering interpretability beyond black-box architectures.

Several extensions warrant future investigation. Incorporating longitudinal EHR data would allow BHPI to model disease trajectories and time-varying pathway activation. Scaling inference to high-dimensional genetic features may require stochastic or amortized variants of the current framework. Finally, while BHPI focuses on associative structure learning, integrating causal constraints or Mendelian randomization could further strengthen causal interpretability.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Antelmi, A., Cordasco, G., Polato, M., Scarano, V., Spagnuolo, C., and Yang, D. A survey on hypergraph representation learning. *ACM Computing Surveys*, 56(1):1–38, 2023. doi: 10.1145/3605776.

Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008. doi: 10.1007/s10994-007-5040-8.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011. doi: 10.1038/nrg2918.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Blei, D., Lafferty, J., et al. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Blumm, N., Barabasi, A.-L., and Christakis, N. A dynamic network approach for the study of human. *PLoS computational biology*, 5:e1000353, 05 2009. doi: 10.1371/journal.pcbi.1000353.

Bonilla, E. V., Chai, K., and Williams, C. Multi-task gaussian process prediction. *Advances in neural information processing systems*, 20, 2007.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Caruana, R. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pp. 301–318. PMLR, 2016.

Evgeniou, T. and Pontil, M. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 109–117, 2004.

Feng, Y., You, H., Zhang, Z., Ji, R., and Gao, Y. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3558–3565, 2019.

Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2:1360–1383, Dec 2008. doi: 10.1214/08-AOAS191.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007. doi: 10.1073/pnas.0701361104.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Heinze, G. and Schemper, M. A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419, 2002.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J., and Brunak, S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications*, 5(1):4022, 2014.

King, G. and Zeng, L. Logistic regression in rare events data. *Political Analysis*, 9:137 – 163, 2001.

Li, W. and McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pp. 577–584, 2006.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. Deep patient: an unsupervised representation to predict the future

- of patients from the electronic health records. *Scientific reports*, 6(1):26094, 2016.
- Murray, C. J., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet*, 396(10258):1223–1249, 2020. doi: 10.1016/S0140-6736(20)30752-2.
- Polson, N. G., Scott, J. G., and Windle, J. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.
- Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. MedBERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.
- Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011. doi: 10.1016/j.eswa.2025.128048.
- Skou, S. T., Mair, F. S., Fortin, M., Guthrie, B., Nunes, B. P., Miranda, J. J., Boyd, C. M., Pati, S., Mtenga, S., and Smith, S. M. Multimorbidity. *Nature Reviews Disease Primers*, 8(1):48, 2022. doi: 10.1038/s41572-022-00376-4.
- Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., and Savarese, S. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pp. 9120–9132. PMLR, 2020.
- Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.
- Tsoumakas, G. and Vlahavas, I. Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning*, pp. 406–417. Springer, 2007.
- Wei, W.-Q., Bastarache, L. A., Carroll, R. J., Marlo, J. E., Osterman, T. J., Gamazon, E. R., et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE*, 12(7):e0175508, 2017. doi: 10.1371/journal.pone.0175508.
- Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J. C., Theodoratou, E., and Wei, W.-Q. Mapping ICD-10 and ICD-10-CM codes to Phecodes: Workflow development and initial evaluation. *JMIR Medical Informatics*, 7(4):e14325, 2019. doi: 10.2196/14325.
- Xu, R., Ali, M. K., Ho, J. C., and Yang, C. Hypergraph transformers for EHR-based clinical predictions. *AMIA Joint Summits on Translational Science Proceedings*, 2023: 582–591, 2023.
- Yadati, N., Nimishakavi, M., Yadav, P., Nitin, V., Louis, A., and Talukdar, P. Hypergen: A new method for training graph convolutional networks on hypergraphs. *Advances in neural information processing systems*, 32, 2019.
- Zhang, Y. and Yang, Q. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022. doi: 10.1109/TKDE.2021.3070203.
- Zhang, Z., Cui, H., Xu, R., Xie, Y., Ho, J. C., and Yang, C. Tacco: Task-guided co-clustering of clinical concepts and patient visits for disease subtyping based on ehr data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6324–6334. Association for Computing Machinery, 2024. doi: 10.1145/3637528.3671594.
- Zhou, D., Huang, J., and Schölkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in neural information processing systems*, 19, 2006.
- Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

A. Variational Inference Details

In this section, we provide the full derivation of the Coordinate Ascent Variational Inference (CAVI) updates summarized in the main text. We begin by defining the complete log-joint probability of the generative model and the variational family.

A.1. Polya-Gamma Augmentation

To handle the non-conjugate logistic likelihood, we introduce auxiliary variables $\omega_{i,v} \mid \tilde{\eta}_{i,v} \sim \text{Polya-Gamma}(1, \tilde{\eta}_{i,v})$. This transforms the conditional posterior of the log-odds into a Gaussian form.

A.2. Full Joint Posterior

Let $\Theta = \left\{ \mathbf{z}(\mathbb{R}^E), \mathbf{r}(\mathbb{R}^E), \mathbf{m}(\mathbb{R}^{V \times E}), \boldsymbol{\rho}(\mathbb{R}^{V \times E}), \boldsymbol{\mu}(\mathbb{R}^{p \times E}), \boldsymbol{\gamma}(\mathbb{R}^{p \times E}), \boldsymbol{\nu}(\mathbb{R}^{p \times E}), \sigma_\mu^2, \boldsymbol{\alpha} \right\}$ denote the full set of model parameters, and the observed data is \mathcal{Y} . For each subject i , the diagnosis status for disease v is $y_{i,v} \in \{0, 1\}$ if the disease is observed after the baseline, otherwise $y_{i,v} = \text{NA}$. To eliminate the effect of prior-baseline diagnosis, we define a risk indicator $R_{i,v} = \mathbb{I}(\text{no diagnosis of } v \text{ before baseline})$. Then the disease outcome is only defined when $R_{i,v} = 1$. Conditional on ω where $\omega_{i,v} \sim \text{Polya-Gamma}(1, 0)$, the log-likelihood becomes

$$\log p(\mathcal{Y} \mid \Theta, \omega) \propto \sum_{i,v} R_{i,v} \left[\kappa_{i,v} \tilde{\eta}_{i,v} - \frac{1}{2} \omega_{i,v} \tilde{\eta}_{i,v}^2 \right], \quad \text{where } \kappa_{i,v} = y_{i,v} - 1/2. \quad (3)$$

The posterior distribution of the model parameters is expressed as

$$p(\Theta \mid \mathcal{Y}, \omega) \propto p(\mathcal{Y} \mid \Theta, \omega) p(\Theta),$$

where $p(\Theta)$ is the prior distribution over all parameters. Thus, the posterior kernel is given by

$$\begin{aligned} p(\Theta \mid \mathcal{Y}, \omega) \propto & p(\mathcal{Y} \mid \Theta, \omega) \cdot \prod_{e=1}^E \prod_{v=1}^V p(m_{v,e} \mid \rho_{v,e}, z_e) p(\rho_{v,e}) \cdot \prod_{e=1}^E p(z_e \mid r_e) p(r_e) \\ & \cdot \prod_{j=1}^p \prod_{e=1}^E p(\mu_{j,e} \mid \gamma_{j,e}, \sigma_\mu^2) \cdot \prod_{j=1}^p p(\gamma_j \mid \mathbf{z}, \mathbf{m}, \boldsymbol{\nu}_j) \cdot \prod_{j=1}^p \prod_{e=1}^E p(\nu_{j,e}) \cdot p(\sigma_\mu^2). \end{aligned} \quad (4)$$

The log posterior is

$$\begin{aligned} \log p(\Theta \mid \mathcal{Y}, \omega) \propto & \log p(\mathcal{Y} \mid \Theta, \omega) + \sum_{e=1}^E \sum_{v=1}^V \log p(m_{v,e} \mid \rho_{v,e}, z_e) \\ & + \sum_{e=1}^E \log p(z_e \mid r_e) + \sum_{j=1}^p \sum_{e=1}^E \log p(\mu_{j,e} \mid \gamma_{j,e}, \sigma_\mu^2) + \sum_{j=1}^p \log p(\gamma_j \mid \mathbf{z}, \mathbf{m}, \boldsymbol{\nu}_j) \\ & + \sum_{e=1}^E \sum_{v=1}^V \log p(\rho_{v,e}) + \sum_{e=1}^E \log p(r_e) + \sum_{j=1}^p \sum_{e=1}^E \log p(\nu_{j,e}) + \log p(\sigma_\mu^2), \end{aligned} \quad (5)$$

$$\begin{aligned} \text{with } \log p(\gamma_j \mid \mathbf{z}, \mathbf{m}, \boldsymbol{\nu}_j) = & \sum_{e=1}^E z_e [\gamma_{j,e} \log \nu_{j,e} + (1 - \gamma_{j,e}) \log(1 - \nu_{j,e})] \\ & - \omega_\gamma \sum_{e_1 < e_2} O(H_{e_1}, H_{e_2}) \cdot z_{e_1} \gamma_{j,e_1} \cdot z_{e_2} \gamma_{j,e_2}. \end{aligned}$$

A.3. Optimal Updating Rule under Coordinate Ascent

Since the posterior in (4) is computationally intractable, we can use the variational inference (VI) framework (Blei et al., 2017), which approximates the posterior distribution of the model parameters by a simpler distribution. To develop an efficient VI algorithm, we first approximate the posterior by minimizing the Kullback-Leibler (KL) divergence between a

variational family $q(\Theta)$ and the true posterior, i.e.,

$$q^*(\Theta) = \arg \min_{q(\Theta) \in \mathcal{Q}} \text{KL}(q(\Theta) \| p(\Theta | \mathcal{Y})),$$

$$\begin{aligned} \text{where } \text{KL}(q(\Theta) \| p(\Theta | \mathcal{Y})) &= \mathbb{E}_{q(\Theta)} [\log q(\Theta)] - \mathbb{E}_{q(\Theta)} [\log p(\Theta | \mathcal{Y})] \\ &= \mathbb{E}_{q(\Theta)} [\log q(\Theta)] - \mathbb{E}_{q(\Theta)} [\log p(\Theta, \mathcal{Y})] + \log p(\mathcal{Y}). \end{aligned}$$

Recall that the evidence lower bound (ELBO) is defined as

$$\mathcal{L}(q) = \mathbb{E}_{q(\Theta)} [\log p(\Theta, \mathcal{Y})] - \mathbb{E}_{q(\Theta)} [\log q(\Theta)] = \log p(\mathcal{Y}) - \text{KL}(q(\Theta) \| p(\Theta | \mathcal{Y})), \quad (6)$$

so that minimizing this KL divergence is equivalent to maximizing the ELBO. Furthermore, we use the mean-field variational inference (MFVI) (Blei et al., 2017), which assumes mutual independence across each parameter, to express the variational distribution $q(\Theta)$ in a factorized form, i.e., $q(\Theta) = \prod_j q_j(\theta_j)$. Therefore, we can find $q^*(\Theta)$ using a coordinate ascent variational inference (CAVI) (Bishop & Nasrabadi, 2006) to iteratively optimize each factor of the mean-field variational distribution while holding the others fixed, until a convergence criterion has met. That is, the optimal $q_j(\theta_j)$ is given by maximizing the ELBO in Equation (6) in terms of the factor θ_j , i.e., $q_j^*(\theta_j) = \arg \max_{q_j(\theta_j)} \mathcal{L}(q_j)$, where

$$\begin{aligned} \mathcal{L}(q_j) &= \mathbb{E}_{q_j(\theta_j)} [\mathbb{E}_{q_{-j}(\Theta_{-j})} \{\log p(\theta_j, \Theta_{-j}, \mathcal{Y})\}] - \mathbb{E}_{q_j(\theta_j)} [\log q_j(\theta_j)] \\ &= \mathbb{E}_{q_j(\theta_j)} [\log (\exp [\mathbb{E}_{q_{-j}(\Theta_{-j})} \{\log p(\theta_j, \Theta_{-j}, \mathcal{Y})\}])] - \mathbb{E}_{q_j(\theta_j)} [\log q_j(\theta_j)] \\ &= \mathbb{E}_{q_j(\theta_j)} [\log \tilde{p}(\theta_j | \Theta_{-j}, \mathcal{Y})] - \mathbb{E}_{q_j(\theta_j)} [\log q_j(\theta_j)] + \log(\text{const}) \\ &= -\text{KL}(q_j(\theta_j) \| \tilde{p}(\theta_j | \Theta_{-j}, \mathcal{Y})) + \log(\text{const}), \\ \tilde{p}(\theta_j | \Theta_{-j}, \mathcal{Y}) &= \frac{\exp [\mathbb{E}_{q_{-j}(\Theta_{-j})} \{\log p(\theta_j, \Theta_{-j}, \mathcal{Y})\}]}{\text{const}}, \end{aligned}$$

Θ_{-j} denotes all variables in Θ except θ_j , $q_{-j}(\Theta_{-j}) = \prod_{k \neq j} q_k(\theta_k)$, and const is a normalization term irrelevant to $q_j(\theta_j)$. So the optimal update for $q_j(\theta_j)$ is obtained by minimizing the KL divergence between $q_j(\theta_j)$ and $\tilde{p}(\theta_j | \Theta_{-j}, \mathcal{Y})$, that is,

$$q_j^*(\theta_j) = \arg \min_{q_j(\theta_j)} \text{KL}(q_j(\theta_j) \| \tilde{p}(\theta_j | \Theta_{-j}, \mathcal{Y})) = \tilde{p}(\theta_j | \Theta_{-j}, \mathcal{Y}) \propto \exp [\mathbb{E}_{q_{-j}(\Theta_{-j})} \{\log p(\Theta, \mathcal{Y})\}]. \quad (7)$$

By matching $q_j(\theta_j)$ to the conditional distribution of θ_j given the observed data \mathcal{Y} and the other factors, this update rule (7) makes $q_j^*(\theta_j)$ as close to the true posterior as possible within the distribution family.

In our structured variational family, certain variables are modeled via conditional distributions (e.g., $q(m_{v,e}, z_e) = q(m_{v,e} | z_e)q(z_e)$). To derive the optimal update for the marginal factor $q(\theta_j)$, we isolate the terms in the ELBO that depend on θ_j . Unlike the standard mean-field case, the entropy term $\mathbb{E}_q[\log q]$ now contributes a conditional entropy component, leading to the update rule:

$$q^*(\theta_j) \propto \exp [\mathbb{E}_{q(\Theta \setminus \theta_j)} \{\log p(\Theta, \mathcal{Y}) - \log q(\Theta_{-j} | \theta_j)\}], \quad (8)$$

A.4. Coordinate Ascent Updates for BHPI

Polya-Gamma augmented variables

$$q(\omega_{i,v}) = \text{Polya-Gamma}(1, \tilde{\eta}_{i,v}^*), \quad \text{where } \tilde{\eta}_{i,v}^* = \sqrt{\mathbb{E}_q [\tilde{\eta}_{i,v}^2]}. \quad (9)$$

Disease-specific intercepts $q^*(\alpha_v) = \mathcal{N}(\alpha_v^*, \varrho_v^{2*})$, where

$$\varrho_v^{2*} = \left[\sum_{i=1}^N R_{i,v} \mathbb{E}_q [\omega_{i,v}] + \frac{1}{\sigma_\alpha^2} \right]^{-1}, \quad \alpha_v^* = \varrho_v^{2*} \left\{ \sum_{i=1}^N R_{i,v} (\kappa_{i,v} - \mathbb{E}_q [\omega_{i,v}] \mathbb{E}_q [\eta_{i,v}]) \right\}. \quad (10)$$

Risk-factor-hyperedge effects $q^*(\mu_{j,e} | \gamma_{j,e}) = (1 - \gamma_{j,e}) \cdot \delta_0 + \gamma_{j,e} \cdot \mathcal{N}(\mu_{j,e}^*, \sigma_{j,e}^{2*})$, where

$$\sigma_{j,e}^{2*} = \left(\mathbb{E}_q [\sigma_\mu^{-2}] + \sum_{i=1}^N \sum_{v=1}^V R_{i,v} x_{i,j}^2 \mathbb{E}_q [\omega_{i,v}] \mathbb{E}_q [s_{v,e}^2 | z_e = 1] \right)^{-1}, \quad s_{v,e} = d_v^{-1} H_{v,e}, \quad a_{i,v}^{\setminus(j,e)} = \eta_{i,v} - x_{i,j} s_{v,e} \mu_{j,e}, \quad (11)$$

$$\mu_{j,e}^* = \sigma_{j,e}^{2*} \left\{ \sum_{i=1}^N \sum_{v=1}^V R_{i,v} x_{i,j} \left(\kappa_{i,v} - \mathbb{E}_q [\omega_{i,v}] \left\{ \mathbb{E}_q [a_{i,v}^{\setminus(j,e)} | z_e = 1] + \alpha_v^* \right\} \right) \mathbb{E}_q [s_{v,e} | z_e = 1] \right\}. \quad (12)$$

Risk-factor-hyperedge effect selector $q(\gamma_{j,e} | z_e) = (1 - z_e) \cdot \delta_0 + z_e \cdot \text{Bernoulli}(\nu_{j,e}^*)$, where

$$\text{logit}(\nu_{j,e}^*) = \frac{1}{2} \left\{ \frac{(\mu_{j,e}^*)^2}{\sigma_{j,e}^{2*}} + \log(\sigma_{j,e}^{2*}) - \mathbb{E}_q [\log \sigma_\mu^2] \right\} + \mathbb{E}_q \left[\log \frac{\nu_{j,e}}{1 - \nu_{j,e}} \right] - \lambda \sum_{e' \neq e} \mathbb{E}_q [O(S_e, S_{e'})] r_{e'} \nu_{j,e'}. \quad (13)$$

Node-hyperedge incidence $q(m_{v,e} | z_e) = (1 - z_e) \cdot \delta_0 + z_e \cdot \text{Bern}(\rho_{v,e}^*)$, where

$$\begin{aligned} \text{logit}(\rho_{v,e}^*) &= \mathbb{E}_q \left[\log \frac{\rho_{v,e}}{1 - \rho_{v,e}} \right] + d_v^{-1} \mathbb{E}_q [\zeta_{v,e} | z_e = 1] - \frac{1}{2} d_v^{-2} \mathbb{E}_q [\xi_{v,e} | z_e = 1] \\ &\quad - \lambda \sum_{j=1}^P \nu_{j,e} \sum_{e' \neq e} \mathbb{E}_q [O(S_e, S_{e'} | m_{v,e} = 1) - O(S_e, S_{e'} | m_{v,e} = 0)] r_{e'} \nu_{j,e'}, \end{aligned} \quad (14)$$

$$\zeta_{v,e} = \sum_{i=1}^N R_{i,v} \left(\kappa_{i,v} - \omega_{i,v} [a_{i,v}^{(-e)} + \alpha_v] \right) b_{i,e}, \quad \xi_{v,e} = \sum_{i=1}^N R_{i,v} \omega_{i,v} b_{i,e}^2, \quad b_{i,e} = \mathbf{x}_i^\top \boldsymbol{\mu}_e, \quad a_{i,v}^{(-e)} = \eta_{i,v} - s_{v,e} b_{i,e}.$$

Hyperedge existence $q(z_e) = \text{Bernoulli}(r_e^*)$, where

$$\begin{aligned} \text{logit}(r_e^*) &= \sum_{v=1}^V \rho_{v,e}^* \left(d_v^{-1} \mathbb{E}_q [\zeta_{v,e} | z_e = 1] - \frac{1}{2} d_v^{-2} \mathbb{E}_q [\xi_{v,e} | z_e = 1] \right) + \mathbb{E}_q \left[\log \frac{r_e}{1 - r_e} \right] \\ &\quad + \sum_{v=1}^V \left\{ \rho_{v,e}^* [\mathbb{E}_q [\log \rho_{v,e}] - \log \rho_{v,e}^*] + (1 - \rho_{v,e}^*) [\mathbb{E}_q [\log(1 - \rho_{v,e})] - \log(1 - \rho_{v,e}^*)] \right\} \\ &\quad + \sum_{j=1}^P \left\{ \nu_{j,e}^* [\mathbb{E}_q [\log \nu_{j,e}] - \log \nu_{j,e}^*] + (1 - \nu_{j,e}^*) [\mathbb{E}_q [\log(1 - \nu_{j,e})] - \log(1 - \nu_{j,e}^*)] \right\} \\ &\quad - \lambda \sum_{j=1}^P \nu_{j,e} \sum_{e' \neq e} \mathbb{E}_q [O(S_e, S_{e'})] r_{e'} \nu_{j,e'}. \end{aligned} \quad (15)$$

Slab variance $q(\sigma_\mu^2) = \text{Inverse-Gamma}(a_\mu^*, b_\mu^*)$, where $a_\mu^* = a_\mu + \frac{1}{2} \sum_{j=1}^P \sum_{e=1}^E r_e^* \nu_{j,e}^*$, and $b_\mu^* = b_\mu + \frac{1}{2} \sum_{j=1}^P \sum_{e=1}^E r_e^* \nu_{j,e}^* \left[(\mu_{j,e}^*)^2 + \sigma_{j,e}^{2*} \right]$.

Prior inclusion probabilities

$$\begin{aligned} q(\nu_{j,e}) &= \text{Beta}(a_{j,e}^{*(\nu)}, b_{j,e}^{*(\nu)}), & \text{where } a_{j,e}^{*(\nu)} &= a_\nu + r_e^* \nu_{j,e}^*, & b_{j,e}^{*(\nu)} &= b_\nu + 1 - r_e^* \nu_{j,e}^*. \\ q(\rho_{v,e}) &= \text{Beta}(a_{v,e}^{*(\rho)}, b_{v,e}^{*(\rho)}), & \text{where } a_{v,e}^{*(\rho)} &= a_\rho + r_e^* \rho_{v,e}^*, & b_{v,e}^{*(\rho)} &= b_\rho + 1 - r_e^* \rho_{v,e}^*. \\ q(r_e) &= \text{Beta}(a_e^{*(r)}, b_e^{*(r)}), & \text{where } a_e^{*(r)} &= a_r + r_e^*, & b_e^{*(r)} &= b_r + 1 - r_e^*. \end{aligned}$$

A.4.1. EXPECTATIONS

$$\mathbb{E}_q [\log \rho_{v,e}] = \psi(a_{v,e}^{*(\rho)}) - \psi(a_{v,e}^{*(\rho)} + b_{v,e}^{*(\rho)}), \quad \mathbb{E}_q [\log(1 - \rho_{v,e})] = \psi(b_{v,e}^{*(\rho)}) - \psi(a_{v,e}^{*(\rho)} + b_{v,e}^{*(\rho)})$$

$$\mathbb{E}_q [\log \nu_{j,e}] = \psi(a_{j,e}^{*(\nu)}) - \psi(a_{j,e}^{*(\nu)} + b_{j,e}^{*(\nu)}), \quad \mathbb{E}_q [\log(1 - \nu_{j,e})] = \psi(b_{j,e}^{*(\nu)}) - \psi(a_{j,e}^{*(\nu)} + b_{j,e}^{*(\nu)})$$

$$\mathbb{E}_q [\log r_e] = \psi(a_e^{*(r)}) - \psi(a_e^{*(r)} + b_e^{*(r)}), \quad \mathbb{E}_q [\log(1 - r_e)] = \psi(b_e^{*(r)}) - \psi(a_e^{*(r)} + b_e^{*(r)})$$

$$\mathbb{E}_q [\sigma_\mu^{-2}] = \frac{a_\mu^*}{b_\mu^*}, \quad \mathbb{E}_q [\log \sigma_\mu^2] = \log b_\mu^* - \psi(a_\mu^*), \quad \mathbb{E}_q [\omega_{i,v}] = \frac{1}{2\eta_{i,v}^*} \tanh(\eta_{i,v}^*/2)$$

$$\mathbb{E}_q [O(S_e, S_{e'})] \approx \frac{[\boldsymbol{\rho}_{\cdot,e}^*]^\top \boldsymbol{\rho}_{\cdot,e'}^*}{\min\left(\sum_{v=1}^V \rho_{v,e}^*, \sum_{v=1}^V \rho_{v,e'}^*\right)}$$

$$\mathbf{h}_e^{(v=1)} = (\rho_{1,e}^*, \rho_{2,e}^*, \dots, \rho_{v-1,e}^*, 1, \rho_{v+1,e}^*, \dots, \rho_{V,e}^*)^\top, \quad \mathbf{h}_e^{(v=0)} = (\rho_{1,e}^*, \rho_{2,e}^*, \dots, \rho_{v-1,e}^*, 0, \rho_{v+1,e}^*, \dots, \rho_{V,e}^*)^\top,$$

$$\mathbb{E}_q [O(S_e, S_{e'} \mid m_{v,e} = 1) - O(S_e, S_{e'} \mid m_{v,e} = 0)] = \frac{[\mathbf{h}_e^{(v=1)}]^\top \boldsymbol{\rho}_{\cdot,e'}^*}{\min\left(\|\mathbf{h}_e^{(v=1)}\|_1, \sum_{v=1}^V \rho_{v,e'}^*\right)} - \frac{[\mathbf{h}_e^{(v=0)}]^\top \boldsymbol{\rho}_{\cdot,e'}^*}{\min\left(\|\mathbf{h}_e^{(v=0)}\|_1, \sum_{v=1}^V \rho_{v,e'}^*\right)},$$

$$\mathbb{E}_q [\eta_{i,v}] = d_v^{-1} \sum_{e=1}^E r_e^* \rho_{v,e}^* \left[\sum_{j=1}^P x_{i,j} \nu_{j,e}^* \mu_{j,e}^* \right], \quad \mathbb{E}_q [\tilde{\eta}_{i,v}] = \mathbb{E}_q [\eta_{i,v}] + \alpha_v^*$$

$$\mathbb{E}_q [b_{i,e}^2 \mid z_e = 1] = \sum_{j=1}^P x_{i,j}^2 \nu_{j,e}^* \left[(\mu_{j,e}^*)^2 + \sigma_{j,e}^{2*} \right] + 2 \sum_{j < k} x_{i,j} x_{i,k} \nu_{j,e}^* \mu_{j,e}^* \nu_{k,e}^* \mu_{k,e}^*$$

$$\mathbb{E}_q [\eta_{i,v}^2] = d_v^{-2} \sum_e r_e^* \rho_{v,e}^* \mathbb{E}_q [b_{i,e}^2 \mid z_e = 1] + 2d_v^{-2} \sum_{e < e'} r_e^* r_{e'}^* \rho_{v,e}^* \rho_{v,e'}^* \left[\sum_{j=1}^P x_{i,j} \nu_{j,e}^* \mu_{j,e}^* \right] \left[\sum_{j=1}^P x_{i,j} \nu_{j,e'}^* \mu_{j,e'}^* \right]$$

$$\mathbb{E}_q [\tilde{\eta}_{i,v}^2] = \mathbb{E}_q [\eta_{i,v}^2] + [(\alpha_v^*)^2 + \varrho_v^{2*}] + 2\alpha_v^* \mathbb{E}_q [\eta_{i,v}]$$

$$\mathbb{E}_q [\xi_{v,e} \mid z_e = 1] = \sum_{i=1}^N R_{i,v} \mathbb{E}_q [\omega_{i,v}] \mathbb{E}_q [b_{i,e}^2 \mid z_e = 1]$$

$$\mathbb{E}_q [a_{i,v}^{(-e)}] = \mathbb{E}_q [\eta_{i,v}] - d_v^{-1} r_e^* \rho_{v,e}^* \left[\sum_{j=1}^P x_{i,j} \nu_{j,e}^* \mu_{j,e}^* \right]$$

$$\mathbb{E}_q [a_{i,v}^{\setminus(j,e)} \mid z_e = 1] = \mathbb{E}_q [a_{i,v}^{(-e)}] + d_v^{-1} \rho_{v,e}^* \left[\sum_{j=1}^P x_{i,j} \nu_{j,e}^* \mu_{j,e}^* \right] - d_v^{-1} \rho_{v,e}^* x_{i,j} \nu_{j,e}^* \mu_{j,e}^*$$

$$\mathbb{E}_q [\zeta_{v,e} \mid z_e = 1] = \sum_{i=1}^N R_{i,v} \left(\kappa_{i,v} - \mathbb{E}_q [\omega_{i,v}] \left\{ \mathbb{E}_q [a_{i,v}^{(-e)}] + \alpha_v^* \right\} \right) \left[\sum_{j=1}^P x_{i,j} \nu_{j,e}^* \mu_{j,e}^* \right]$$

B. Additional Simulation Results

B.1. Visualization of structure discovery

To qualitatively assess the structural recovery performance of BHPI, we visualize the latent disease hypergraph incidence matrix H and the risk-factor-hyperedge effect selector γ_j , along with the effect size μ_j . Figure 5 compares the ground-truth incidence matrix against the posterior mean inferred by BHPI. The model successfully identifies the block-diagonal structure and handles overlapping memberships without introducing significant noise. Furthermore, Figure 6 illustrates that BHPI accurately captures the sparsity and magnitude of feature effects, correctly assigning risk factors to their respective latent pathways.

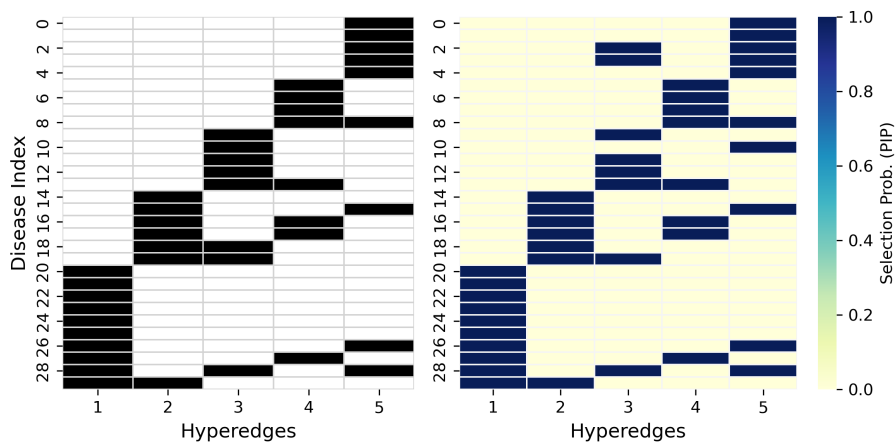


Figure 5. Recovery of latent disease hypergraph. Heatmaps show true (left) and inferred (right) disease hypergraph incidence matrices for a representative simulation replicate. BHPI accurately recovers overlapping hypergraph structure with limited redundancy.

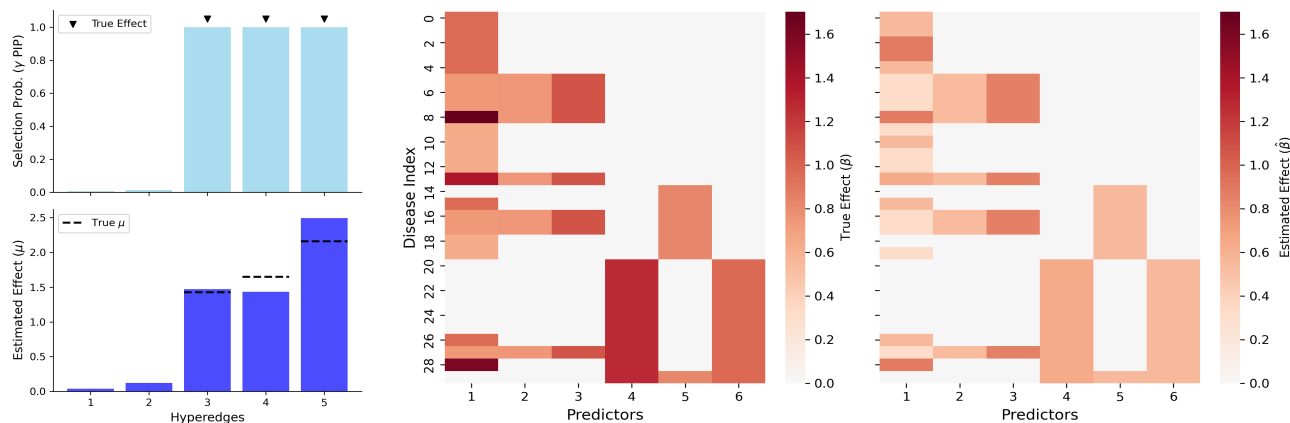


Figure 6. Bar plots show true versus inferred effects of predictor 1 across hyperedges, measured by PIP of γ and magnitude μ .

B.2. Ablation: role of the repulsion prior

The repulsion prior \mathcal{R}_{rep} is designed to enforce identifiability by penalizing redundant latent pathways. In this section, we expand on the results in Table 3, as displayed in Table 4. While predictive AUC is relatively insensitive to the repulsion strength λ , the structural metrics, specifically overlap coefficients, show marked improvement when $\lambda > 0$. Without repulsion, the model often instantiates multiple near-identical hyperedges to represent a single disease pathway, leading to a “collapsed” latent space that is difficult to interpret.

B.3. Sensitivity analysis

We evaluate the robustness of BHPI to the repulsion strength λ , and the assumed maximum number of latent hyperedges E . Regarding λ , as shown in Figure 7, we observe a wide “stable regime” where the balance between sparsity and structure recovery remains optimal, suggesting that BHPI does not require extensive hyperparameter tuning to achieve reliable results. Moreover, from Figure 8, we find that increasing λ sharpens posterior inclusion probabilities (PIPs) and reduces redundancy between inferred hyperedges, leading to more stable latent structures across random splits. We further examine robustness to over-specification of model capacity by varying E as displayed in Figure 9, and find that as E increases beyond the “true” number of pathways, the variational factor $q(z_e)$ effectively prunes the excess components. This shows that even when E is over-specified, the effective number of active hyperedges remains stable, demonstrating effective self-regularization.

Table 4. Role of the repulsion prior in hypergraph identifiability ($N = 2000$ and 5000). Predictive performance remains stable, while repulsion reduces redundancy and improves stability of inferred hyperedges, as measured by overlap coefficients (Ov.) across replicates, effective hyperedge (Eff.#/RF) per risk factor (RF), and hyperedge overlap per RF (Ov./RF).

λ	$N = 2000$				$N = 5000$			
	AUC	Ov.	EFF.#/RF	Ov./RF	AUC	Ov.	EFF.#/RF	Ov./RF
$= 0$	74.69	0.800	1.992	0.097	74.51	0.816	2.392	0.116
> 0	75.00	0.812	1.874	0.082	74.63	0.830	2.151	0.102

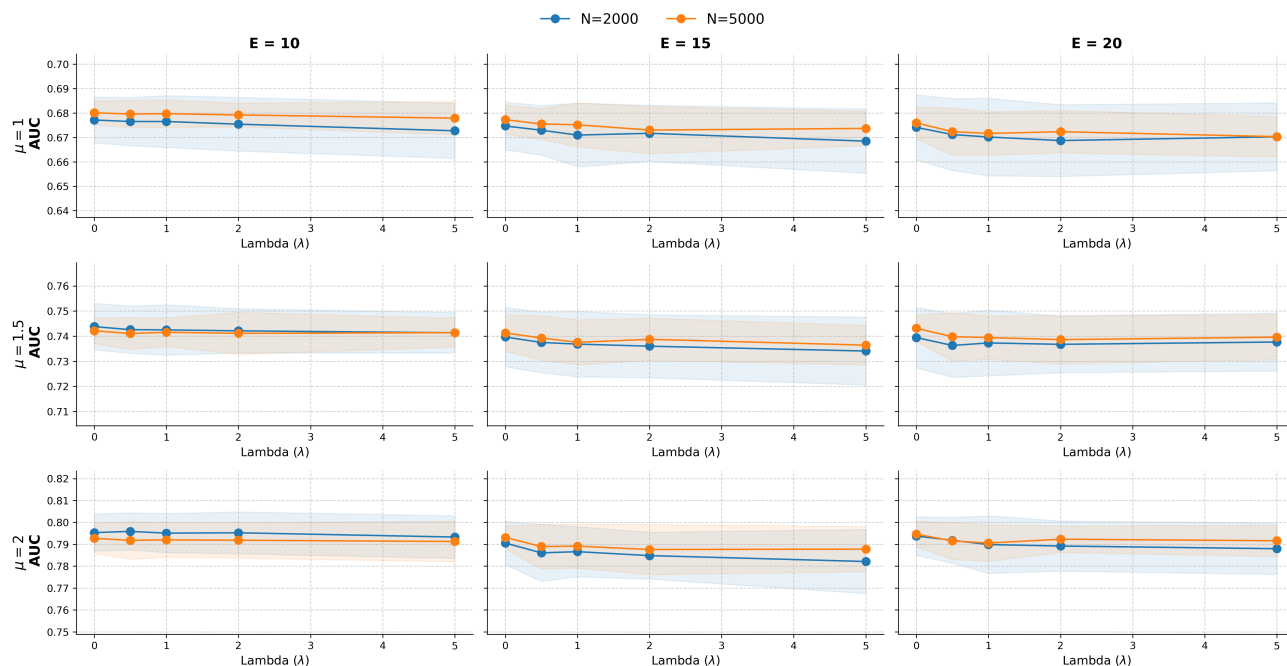


Figure 7. Robustness of predictive performance to repulsion strength across model sizes. Each panel fixes the number of hyperedges E ; predictive AUC remains stable across a wide range of λ and E , indicating robustness to both hyperparameters.

C. Additional Results for UKB Application

In this section, we provide additional information and results for the UK Biobank (UKB) application. This includes detailed dataset characteristics in Table 5, receiver operating characteristic (ROC) analysis for several representative diseases, and an expanded library of qualitative case studies that illustrate the clinical interpretability of the learned latent hypergraph.

The cohort characteristics, summarized in Table 5, reflect the high-dimensional and imbalanced nature of real-world clinical data.

C.1. ROC for representative diseases

To complement the aggregate performance metrics reported in the main text, we visualize the predictive stability of BHPI across diseases with varying prevalence. As shown in Figure 10, BHPI maintains competitive AUCs even for low-prevalence conditions (e.g., prevalence $\approx 1\%$), where baseline models often exhibit higher variance or sensitivity to class imbalance.

C.2. Extended Qualitative Case Studies

While the main text highlights the “Asthma-COPD” and “Smoking” examples, to further validate the clinical utility of BHPI, we provide an extensive analysis of additional latent pathways recovered from the UK Biobank data. These cases illustrate the model’s ability to handle complex geriatric, metabolic, and psychiatric comorbidities.

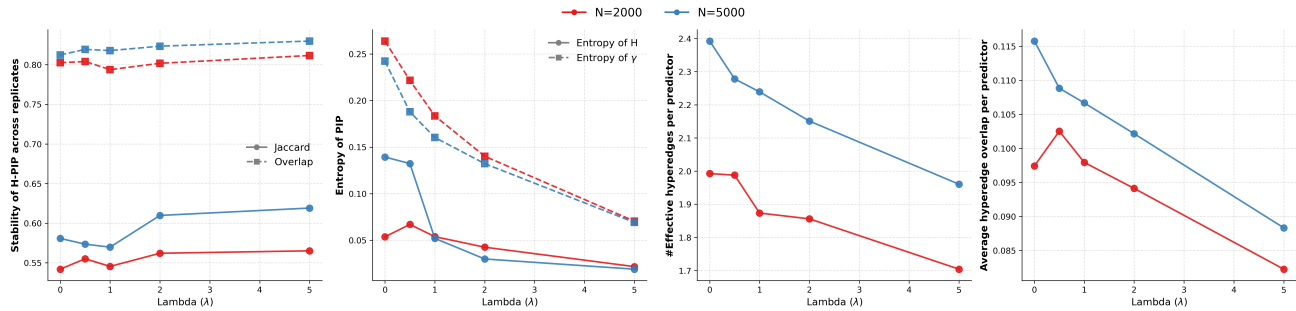


Figure 8. Structural stability, sharpness, and redundancy across repulsion strengths.

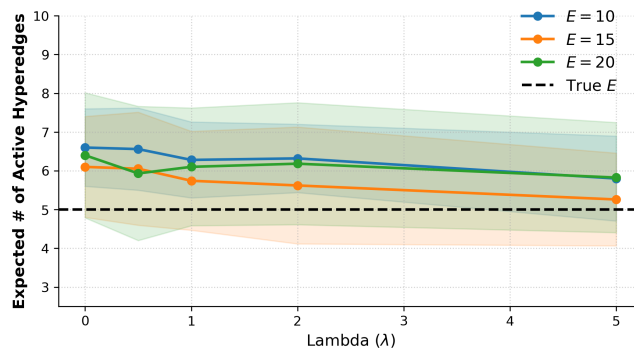


Figure 9. Robustness to over-specification of the maximum number of hyperedges E , measured by the posterior expected number of active hyperedges.

C.2.1. DIVERGENT RISK PATHWAYS

Case A: The Systemic Impact of Sleeplessness. We examine “Sleeplessness / Insomnia,” a risk factor often treated as a generic marker of poor health. BHPI reveals that this feature actively modulates distinct phenotypic pathways: **(i) The Mental Health Pathway (Hyperedge 22):** Sleeplessness is a primary driver of a highly specific cluster ($V = 1$) containing only *Other mental disorder*. Modulated by *Sex* and *Chest pain*, this captures the bidirectional link between sleep and psychiatric instability. **(ii) The Cardio-Metabolic Pathway (Hyperedge 16):** In contrast, Sleeplessness also drives Hyperedge 16 ($V = 17$), co-occurring with *Obesity*, *Type 2 Diabetes*, and *Sleep apnea*. Here, it interacts with lifestyle predictors like *Diet* and *BMI*, reflecting the physiological impact of sleep deprivation on metabolic regulation. **(iii) The Genitourinary Pathway (Hyperedge 27):** Finally, it links to *Hyperplasia of prostate*, capturing the symptom-driven association where nocturia disrupts sleep.

Case B: Alcohol: Organ Toxicity vs. Caloric Impact. **(i) The Liver-Toxicity Pathway (Hyperedge 18):** Alcohol intake predicts Hyperedge 18 ($V = 11$), which includes *Nonalcoholic liver disease* and *Cholelithiasis*. This isolates the direct toxic/inflammatory effects on the hepatic system. **(ii) The Pure Obesity Pathway (Hyperedge 25):** Distinct from organ damage, Alcohol intake also drives Hyperedge 25 ($V = 1$), which contains only *Obesity*. This isolates the caloric contribution of alcohol to body mass separate from pathophysiological organ damage.

C.2.2. CONVERGENT DISEASE PATHWAYS

Case C: Type 2 Diabetes Contexts. **(i) Active Metabolic Syndrome (Hyperedge 16):** Diabetes appears with *Obesity* and *Sleep apnea*, modulated by modifiable lifestyle factors (*Diet*, *Physical activity*). This represents the active, manageable phase of the disease. **(ii) Cardiovascular End-Point (Hyperedge 4):** Diabetes also appears in Hyperedge 4 ($V = 29$), dominated by *Heart failure* and *Atherosclerosis*. Here, predictors are markers of established damage (*Chest pain*, *Wheeze*), modeling Diabetes as a contributor to terminal cardiovascular collapse.

Table 5. UK Biobank dataset characteristics. The cohort exhibits high-dimensional predictors, substantial class imbalance, and a mix of common and rare disease outcomes.

STATISTIC	VALUE
PARTICIPANTS (N)	277,291
RISK FACTORS (P)	71
DISEASE OUTCOMES (V)	66
<i>Disease Prevalence (%)</i>	
MEDIAN	1.8%
RANGE (MIN – MAX)	1.0% – 21.8%
<i>Data Sparsity</i>	
AVG. DISEASES PER PATIENT	4.2
OBSERVATION WINDOW	19 YEARS

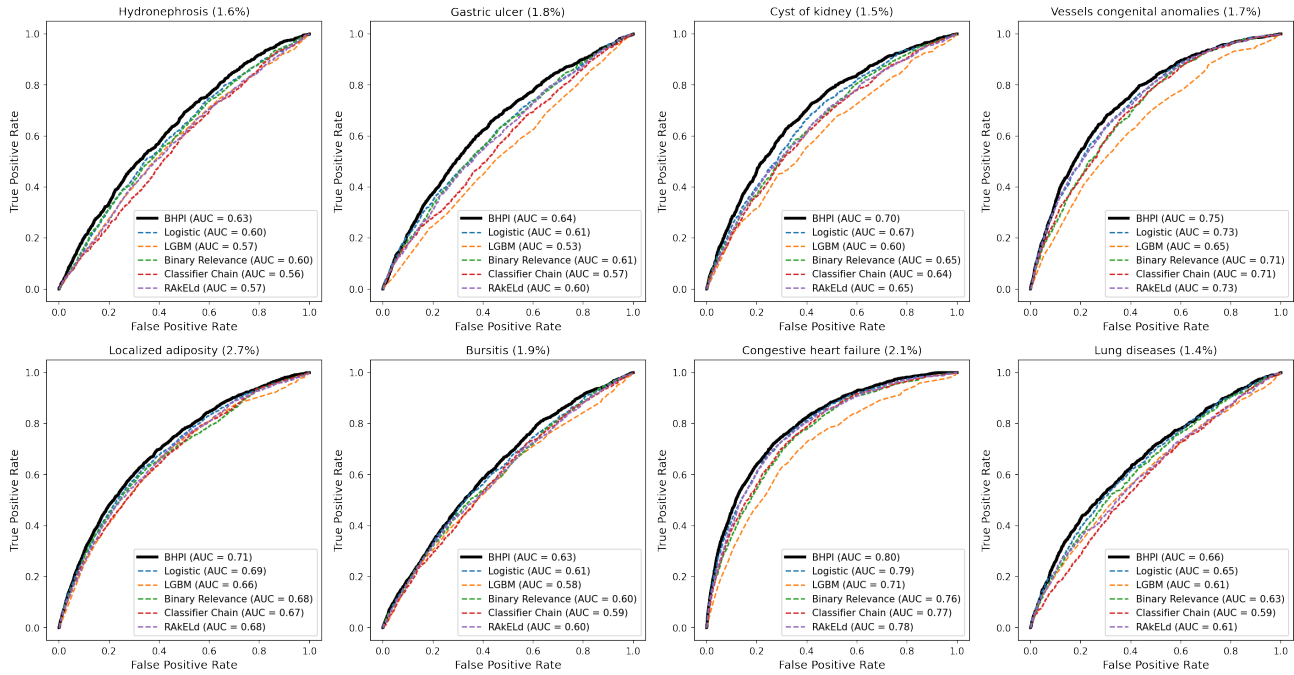


Figure 10. ROC curves of representative diseases for CAVI, logistic regression, and LGBM, with the prevalence indicated in brackets.

Case D: Obesity Heterogeneity. (i) **Frailty & Structural Pathway (Hyperedge 20):** Obesity co-occurs with *Prolapse of vaginal walls* and *Osteoporosis*, modulated by *Age* and *Sex*. (ii) **Metabolic-Digestive Pathway (Hyperedge 13):** Obesity is grouped with *Type 2 Diabetes* and *Liver disease*, driven by *Employment* and *Walking frequency*.

D. Computational Cost

D.1. Theoretical Complexity

As discussed in Section 4.2, the per-iteration computational complexity of the BHPI coordinate ascent algorithm is approximately $\mathcal{O}(N \cdot E \cdot (P + V))$. This linear scaling with respect to the number of samples (N), the number of risk factors (P), and the number of diseases (V) is a critical property that enables the framework to handle biobank-scale data. The primary computational bottleneck lies in the simultaneous update of the hyperedge existence z_e , the disease-membership indicators $m_{v,e}$ and the feature-effect selectors $\gamma_{j,e}$, all of which require aggregating information across all N individuals in the dataset. However, because the variational updates for different hyperedges $e \in \{1, \dots, E\}$ are independent within each iteration, the algorithm can be trivially parallelized to further reduce wall-clock time in high-performance computing environments.

Table 6. Wall-clock runtimes for simulated data and UK Biobank application ($N \approx 277, 291$; $V = 66$). Synthetic tests were conducted on a single-core Intel Xeon 6342 (4 GB RAM); UK Biobank results utilized 4 cores and 60 GB RAM. BHPI maintains inference latency on par with Logistic Regression (< 0.1 ms) while providing structured latent discovery.

Model	Simulated Data (Section 5.1)		UK Biobank			
	Runtime (sec) $N = 2000$	Runtime (sec) $N = 5000$	Model	Memory (GB)	Training (min)	Inference (ms)
Logistic Reg.	0.12	0.23	Logistic Reg.	3.74	5.82	< 0.1
LightGBM	3.56	6.14	LightGBM	3.18	17.41	0.48
Binary Relevance	0.30	0.61	Binary Relevance	1.45	12.24	0.29
Classifier Chains	0.33	0.68	Classifier Chains	1.36	8.23	0.56
RAkEld	0.59	1.35	RAkEld	1.39	3.06	1.56
BHPI ($E = 10/15/20$)	1.6/2.7/3.2	4.1/6.4/9.3	BHPI ($E = 60$)	27.8	73.7	< 0.1

D.2. Empirical Scaling with Model Capacity

To evaluate how the “structural discovery tax” grows with model complexity, we measured the training time on the simulated dataset while varying the maximum number of latent hyperedges E ($E = 10, 15$ and 20 respectively). As shown in the scaling results in Table 6, the training time grows linearly with E . This confirms that our repulsion-aware inference remains stable even as model capacity increases, as the variational “global switch” r_e^* efficiently prunes redundant components without incurring exponential computational overhead.

While the proposed BHPI model incurs a higher computational cost during the training phase compared to independent baseline models, it remains highly scalable and practical for biobank-scale datasets. As summarized in Table 6, the training time for nearly 300,000 samples is approximately 74 minutes, a manageable one-time “structural discovery tax” for learning a complex, interpretable latent hypergraph. Crucially, while BHPI requires more time for one-time structural discovery, its inference latency remains on par with simple Logistic Regression (< 0.1 ms), ensuring the model’s suitability for real-time clinical decision support even at Biobank scale. While BHPI requires more memory (27.8 GB) to maintain its structured variational parameters, this remains well within the capacity of standard modern research servers, proving that the model can effectively scale to the high-dimensional requirements of modern EHR analysis without compromising on predictive speed.