MT-BN: Multi-Scale Topological Bayesian Networks for Tractable, Interpretable Structure Learning in $p \gg n$

Aaryan Arora
Acton-Boxborough Regional High School, Massachusetts, United States of America
MIT PRIMES

Supervised by:
Dr. Gil Alterovitz, Biomedical Cybernetics Laboratory, Harvard Medical School, Massachusetts, United States of America

January 2, 2026

MT-BN: Multi-Scale Topological Bayesian Networks for Tractable, Interpretable Structure Learning in $p \gg n$

Aaryan Arora

January 2, 2026

Abstract

High-dimensional datasets with far more variables than samples ($p \gg n$) overwhelm classical flat Bayesian-network learners: their search space over directed acyclic graphs (DAGs) grows super-exponentially and, by treating all nodes at one level, they ignore the modular, multi-scale organization that real systems exhibit, hurting both computational tractability and interpretability. We introduce MT-BN (Multi-Scale Topological Bayesian Networks), a Bayesian structure-learning framework that infers an adaptive hierarchy of modules and learns directed influence networks at multiple resolutions, while defining within-resolution directionality on resolution-specific innovations rather than on inherited shared signal. Concretely, MT-BN learns a nested partition of the p variables (via a truncated nCRP prior), associates each module at each level with latent states, decomposes each state into inherited signal plus a level-specific innovation, and learns within-level DAGs exclusively on innovation latents to avoid spurious dependencies induced by common ancestry. Connectivity is regularized by topology-aware priors that encourage sparsity, hierarchy-consistent proximity structure, and heterogeneous hub degree profiles, and a hybrid inference pipeline combines variational inference for continuous latents with stochastic structure search over hierarchies and DAGs to yield scalable computation and edge posterior summaries. Under a minimum module size constraint, MT-BN replaces flat structure search over p observed nodes with multi-resolution structure search over module graphs whose node counts are bounded by the number of admissible modules at each level; equivalently, MT-BN searches over a product of within-level DAG spaces $\prod_{\ell=1}^{L} \mathsf{DAG}(M_\ell)$ rather than $\mathsf{DAG}(p)$, and reduces the candidate edge universe from $p(p-1)$ to $\sum_{\ell=1}^{L} M_\ell(M_\ell-1)$. On the DREAM5 Network Inference Challenge, MT-BN outperforms flat Bayesian-network baselines, achieving higher edge-recovery accuracy (e.g., Net1 AUPR 0.325 vs. 0.218, +49%; Net3 AUPR 0.071 vs. 0.043, +65%) and consistent AUROC gains (Net1 0.733 vs. 0.683; Net3 0.597 vs. 0.559). We further demonstrate MT-BN on tuberculosis gene-expression cohorts ($p = 6503$ genes, $n = 2722$ samples), where MT-BN recovers interpretable immune programs (including interferon/ISG and upstream JAK–STAT control) supported by external pathway evidence and yields stable hub- and edge-based candidates for biomarker nomination. These results highlight MT-BN as a practical multi-scale alternative to flat BN learning for structure discovery in the $p \gg n$ regime.

Keywords

Bayesian networks, structure learning, multi-scale modeling, hierarchical latent variables, causal discovery, variational inference, gene regulatory networks, $p \gg n$

Contents

1 Introduction

Bayesian networks (BNs) provide a principled framework for representing multivariate dependence through directed acyclic graphs (DAGs).[1] They are attractive in scientific settings because a learned structure can be interpreted as a compact hypothesis about directional influence, and because Bayesian scoring naturally regularizes against overfitting by integrating uncertainty in parameters.[2] However, standard BN structure learning becomes brittle in the high-dimensional regime $p \gg n$: the number of DAGs grows super-exponentially in p, statistical evidence is limited for orienting edges, and common algorithms exhibit unstable behavior under small perturbations of the data or hyperparameters.[2, 3] These difficulties are pronounced in genomics and other biological systems where p may be in the thousands with only tens to hundreds of samples [4, 5], and similarly arise in financial dependence modeling when one seeks directed relationships among large collections of assets or latent factors.[6]

A second difficulty is that dependency structure in real systems is rarely arbitrary. Empirical networks often exhibit modular organization, sparse long-range connectivity across modules, and heterogeneous degree profiles in which a small number of hubs mediate substantial directed influence.[7] Flat BN learning procedures, whether score-based, constraint-based, or hybrid, typically treat the full variable set as the graph node set, forcing the method to decide simultaneously among an enormous space of possible edges while ignoring multi-resolution organization.[8, 9, 10] Practical workarounds often impose a two-stage pipeline: variables are clustered into modules and a module-level graph is learned afterwards, or a latent factor model is fit and a graph is learned on inferred factors.[11, 12, 13, 14] While these approaches can improve stability, they decouple representation learning from structure learning and can confound within-module shared signal with genuine directed relationships.

This paper introduces MT-BN, a multi-scale topological Bayesian network framework for tractable, interpretable structure learning in high dimensions. MT-BN replaces a single flat structure learning problem with a coupled model that simultaneously (i) learns a hierarchy of nested modules that partitions the p observed variables at multiple resolutions, (ii) associates each module with latent trajectories across samples that summarize module activity, and (iii) learns directed acyclic graphs among modules at each resolution. The hierarchy captures vertical organization, while within-level DAGs capture horizontal directed dependencies among modules at the same resolution. By moving graph learning to the module level, MT-BN reduces the effective search dimension from p to the number of modules at each resolution, enabling localized structure updates and substantially shrinking the search space.

A key modeling principle that distinguishes MT-BN from existing hierarchical or modular graphical models is that within-level directed structure at a given resolution is defined on what is newly expressed at that resolution rather than on signal inherited from coarser scales. Concretely, each module state at level ℓ decomposes into an inherited component determined by its parent module at level $\ell - 1$ and an innovation component capturing residual variation newly expressed at level ℓ. MT-BN places directed edges exclusively among innovation latents rather than among full module states. This choice prevents sibling modules from appearing spuriously dependent due solely to shared inheritance and yields a clean interpretation: directed edges at resolution ℓ represent relationships among innovations conditional on higher-level structure and shared signal. Under standard structural equation model assumptions, these directed edges admit a causal interpretation at the corresponding resolution; otherwise they should be interpreted as directed predictive dependencies induced by the model class.

MT-BN further regularizes and accelerates learning through topology-aware priors on within-level DAGs. These priors encode global sparsity, hierarchy-induced block structure favoring denser connectivity among nearby modules in the hierarchy, and heterogeneous degree profiles consistent with hub structure. The priors reduce posterior mass over implausible graphs and enable efficient local score updates under single-edge proposals. Because exact posterior inference over hierarchies and DAGs is intractable in the targeted regime, MT-BN uses a blocked hybrid inference strategy: variational inference for continuous latent variables and parameters conditional on a fixed discrete structure, coupled with Metropolis-style stochastic search over discrete structures guided by a variational structure score. The result is a scalable procedure for locating high-scoring multi-scale structures and producing stability-based uncertainty summaries over recurrent edges and module assignments.

We validate MT-BN on three case studies spanning benchmark and real-world settings. First, we evaluate directed edge recovery on DREAM5 network inference tasks with available ground truth. Second, we apply MT-BN to tuberculosis gene expression data to identify modular structure and candidate regulatory drivers at multiple resolutions. Across these studies, MT-BN yields interpretable module hierarchies, directed dependencies that are stable under the induced exploration distribution, and competitive performance relative to baselines that do not incorporate multi-resolution organization.

The contributions of this work are as follows. We propose a unified probabilistic framework that couples hierarchical partition learning, latent representation learning, and multi-scale directed structure learning within one Bayesian model. We introduce the innovation-based causality principle, placing within-level directed structure on resolution-specific innovations rather than inherited module states. We develop topology-aware graph priors that en-

code sparsity, hierarchy-induced proximity, and hub structure, and that admit localized updates suitable for efficient discrete search. Finally, we present a scalable hybrid inference procedure and demonstrate the utility of MT-BN across biological and financial datasets in the high-dimensional regime.

Organization. Section 2 reviews related work. Section 3 summarizes the BN and variational inference preliminaries used later. Section 4 formalizes the learning problem. Section 5.1–Section 5.8 present the MT-BN framework, inference, and outputs. Section 6 formalizes computational efficiency gains. Section 7 reports experimental results, and Section 8 discusses limitations and future directions.

2 Related Work

This section positions MT-BN relative to prior work in Bayesian network (BN) structure learning, modular and hierarchical network models, latent-variable graphical modeling, and topology-aware priors. The central distinction is that MT-BN is a joint model of (i) a multi-level hierarchy over variables, (ii) latent module representations, and (iii) within-level directed structure defined on resolution-specific innovations rather than on inherited shared signal.

2.1 Bayesian network structure learning in high dimensions

Classical BN structure learning methods are typically organized around a scoring criterion (e.g., marginal likelihood or a large-sample approximation such as BIC) combined with a discrete search procedure over DAG space.[15, 16] Greedy equivalence search (GES) is a canonical example of score-based search over Markov equivalence classes, with asymptotic guarantees under standard assumptions but a search space that becomes difficult to explore as p grows.[17] Constraint-based methods such as the PC algorithm instead infer a CPDAG from conditional-independence tests and then orient edges where possible, but are well known to become statistically brittle when reliable high-order conditional independences cannot be estimated (a common situation in $p \gg n$ regimes).[18]

More recently, continuous relaxations formulate DAG learning as optimization over a weighted adjacency matrix by imposing a smooth acyclicity constraint. NOTEARS is a representative approach that replaces combinatorial search by a continuous objective coupled to an exact differentiable characterization of acyclicity.[19] These methods can be highly effective for medium-scale problems, but they still ultimately target a single flat DAG over observed variables (or a single layer of latent variables) and therefore do not, by themselves, resolve the statistical and computational issues that arise when dependencies are multi-scale, modular, and hub-dominated in very high dimensions.

MT-BN is complementary to these lines. Rather than proposing a new flat-DAG solver, MT-BN changes the object being learned: it replaces a single p-node structure problem with a hierarchy of smaller within-level module graphs coupled through a vertical inheritance model. This shifts difficulty from searching an enormous global DAG space to jointly inferring a hierarchy plus multiple smaller DAGs on innovation latents, with topology-aware priors further shrinking the effective search space.

2.2 Module-based and clustered representations for networks

A common strategy in genomics and other domains is to reduce dimensionality by clustering variables into modules and then learning relationships between modules. The module networks framework is a prominent example: it groups genes into co-regulated modules and associates each module with regulators and condition-specific dependencies, yielding a probabilistic description that is more interpretable than a fully gene-level BN in many settings.[20] More broadly, many pipelines follow a two-stage pattern in which module discovery (clustering, community detection, or factor extraction) is performed first, and graph learning is performed second on aggregated module summaries.

The key limitation for MT-BN's target setting is not modularity itself, but how modularity is used. In two-stage pipelines, the representation is typically treated as fixed (or only weakly coupled) when learning directed structure. This decoupling makes it hard to prevent inherited shared signal from creating spurious within-level dependencies: if sibling modules share a strong common parent-level component, any graph learned directly on the module summaries can be dominated by that shared inheritance rather than by resolution-specific interactions.[11, 21] MT-BN addresses this by defining within-level directed structure on innovation latents $U^{(\ell)}$ that isolate what is newly expressed at level ℓ, conditional on higher-level variation. This is a modeling choice about what counts as within-level dependence and what is treated as inherited context, and it is enforced explicitly through the vertical inheritance construction in Section 5.5.

2.3 Latent-variable graphical models and factor-analytic structure

A second major line of work explains high-dimensional dependence via latent variables. Factor models represent covariance structure through low-rank latent factors, while latent-variable graphical models combine sparse conditional dependence among observed variables with a small number of latent confounders. A representative formulation decomposes the observed precision matrix into a sparse component plus a low-rank component corresponding to latent effects, and can be fit via convex optimization under suitable conditions.[21] These approaches are powerful for distinguishing direct from confounded associations in undirected settings, and they provide an important conceptual baseline: shared latent variation can induce apparent dependencies that are not direct interactions.

MT-BN differs in both target and semantics. First, MT-BN is explicitly multi-resolution: it introduces a hierarchy of latent summaries, not a single latent layer, and it uses the hierarchy to define what is inherited versus what is new at each resolution. Second, MT-BN's within-level objects are directed and are learned on innovations rather than on full latent states. In standard latent-variable graphical models, the latent component typically represents shared variation that is integrated into the model primarily to improve recovery of direct relationships among observables. In MT-BN, shared variation is elevated to a structured multi-level hierarchy and then explicitly removed (via conditioning on ancestors) before within-level directionality is inferred on the residual innovation signal.

2.4 Hierarchical network models and multi-scale community structure

There is extensive work on hierarchical structure in networks, particularly in the form of nested community models and hierarchical stochastic block models (SBMs). The nested SBM formalism provides a principled Bayesian approach to representing networks at multiple resolutions by recursively grouping nodes into blocks, often yielding strong compression and interpretability for large graphs.[22] However, this literature typically assumes an observed network is given and seeks a hierarchical description of its connectivity patterns, often in undirected or non-causal directed settings. The hierarchy is a model of block structure in an observed adjacency matrix rather than a model that jointly generates data matrices X through latent module trajectories and within-level structural equations.

MT-BN instead treats the hierarchy as a prior over nested partitions of variables that is learned jointly with directed within-level structure and latent representations from the underlying data. This distinction matters for novelty and identifiability: MT-BN's hierarchy is not merely a descriptive compression of an observed graph, but a component of a generative model in which inheritance across levels explains shared signal and within-level edges are restricted to innovations.

2.5 Topology-aware priors, sparsity, and hub structure

Bayesian approaches to BN learning often incorporate sparsity through edge-count penalties or Beta–Bernoulli constructions, reflecting the empirical belief that real networks are sparse.[15] Separately, empirical studies of biological and technological networks motivate priors that allow heterogeneous degree profiles and hub-like behavior. Most existing priors emphasize either global sparsity or degree heterogeneity, and when modularity is included it is often encoded through block constraints or partition-dependent edge probabilities.

MT-BN combines these preferences in a multi-resolution setting via a product-of-experts construction: a global sparsity component, a hierarchy-induced proximity component, and a hub-structure component applied at each level (Section 5.4). The intent is not to claim that any one of these ingredients is new in isolation, but that their joint use inside a multi-resolution latent-innovation framework yields a search space reduction mechanism that aligns with common empirical regularities while preserving a clear causal target at each resolution.

2.6 Benchmarks and gene network inference context

Gene regulatory network inference has a long history of benchmarking methodologies under partially observed ground truth. The DREAM5 Network Inference Challenge remains a widely used reference point for comparing network inference methods across in silico and in vivo settings, while also highlighting the difficulty of evaluation when real-network ground truth is incomplete.[23] MT-BN's case studies use this benchmarking tradition as motivation for evaluating not only edge-recovery accuracy where ground truth is available, but also interpretability and stability of inferred multi-scale structure.

Summary of the gap MT-BN targets. Across these literatures, there exist effective methods for (i) learning a single DAG, (ii) clustering into modules, (iii) introducing latent variables to explain shared dependence, and (iv) describing hierarchical block structure in an observed network. MT-BN targets the intersection that is less directly addressed: joint learning of a multi-level hierarchy and directed within-level structure where causality at

each resolution is defined on what is newly expressed at that resolution (innovations), not on inherited shared signal. This design choice is the mechanism by which MT-BN aims to avoid spurious within-level edges induced by hierarchical inheritance while retaining a coherent multi-scale directed interpretation.

3 Preliminaries

This section summarizes the Bayesian network and inference concepts needed to formalize MT-BN. The goal is to fix notation and state standard results that will be used implicitly in the model construction and inference procedure.

3.1 Directed acyclic graphs and Bayesian networks

Let $G = (V, E)$ be a directed acyclic graph (DAG) on node set $V = \{1, \dots, d\}$. For a node $v \in V$, write $\mathrm{Pa}_G(v) \subseteq V \backslash \{v\}$ for its parent set. A Bayesian network associated with G is a joint distribution on random variables $X_1, \dots, X_d$ that satisfies the directed Markov property: each variable is conditionally independent of its non-descendants given its parents.[16] Equivalently, the joint density factorizes as

$$p(x_1, \dots, x_d \mid G, \theta) = \prod_{v=1}^{d} p\big(x_v \mid x_{\mathrm{Pa}_G(v)}, \theta_v\big), \tag{1}$$

where $\theta = \{\theta_v\}$ denotes local parameters of the conditional distributions.

Given i.i.d. samples $x^{(1)}, \dots, x^{(n)}$ collected into a data matrix $X \in \mathbb{R}^{n \times d}$, the likelihood under a BN factorizes across nodes:

$$p(X \mid G, \theta) = \prod_{t=1}^{n} \prod_{v=1}^{d} p\big(x_v^{(t)} \mid x_{\mathrm{Pa}_G(v)}^{(t)}, \theta_v\big) = \prod_{v=1}^{d} p\big(X_{:,v} \mid X_{:,\mathrm{Pa}_G(v)}, \theta_v\big). \tag{2}$$

This decomposability is the basis for efficient score updates under local modifications of parent sets in score-based structure learning.

Two DAGs can encode the same set of conditional independences, forming a Markov equivalence class. In purely observational settings, without additional assumptions beyond the directed Markov property and faithfulness, the data identify the equivalence class rather than a unique DAG. Methods that output a single directed graph therefore either target an arbitrary representative of the class or rely on additional modeling assumptions that break equivalence.

3.2 Linear Gaussian structural equation models

A common parametric BN family is the linear Gaussian structural equation model (SEM).[24] Let $X \in \mathbb{R}^{n \times d}$ be the data matrix with columns $X_{:,v}$. Under a DAG G, a linear Gaussian SEM specifies that for each node v,

$$X_{:,v} = \sum_{u \in \mathrm{Pa}_G(v)} X_{:,u}\, \beta_{u \to v} + \varepsilon_v, \qquad \varepsilon_v \sim \mathcal{N}(0, \sigma_v^2 I_n), \tag{3}$$

with independent noise across nodes. This induces a BN with Gaussian conditionals and hence a joint Gaussian distribution whose precision matrix respects the DAG. The SEM view is useful because it separates the directed structure G from the regression parameters $\{\beta_{u \to v}\}$ and noise variances $\{\sigma_v^2\}$, and because it clarifies how directed edges correspond to predictive relations conditional on parents.

Causal interpretation of the arrows in G generally requires assumptions beyond observational factorization. In MT-BN we will adopt an SEM-based generative model for latent innovations at each resolution and interpret directed edges as causal at that resolution only under the corresponding SEM assumptions (appropriate noise structure and absence of additional unmodeled confounding after conditioning on higher-level latents). Without these assumptions, the learned DAGs should be read as directed dependencies within the chosen model class.

3.3 Bayesian scoring and decomposability

Bayesian structure learning places a prior $p(G)$ over DAGs and a prior $p(\theta \mid G)$ over parameters.[15] The posterior over structures is

$$p(G \mid X) \propto p(G)\, p(X \mid G), \qquad p(X \mid G) = \int p(X \mid G, \theta)\, p(\theta \mid G)\, d\theta, \tag{4}$$

where $p(X \mid G)$ is the marginal likelihood. The log posterior score is therefore

$$\log p(G \mid X) = \log p(G) + \log p(X \mid G) + \text{const.} \tag{5}$$

When the model is conjugate and node-conditional likelihoods depend only on $\mathrm{Pa}_G(v)$, the marginal likelihood factorizes across nodes:

$$p(X \mid G) = \prod_{v=1}^{d} p\big(X_{:,v} \mid X_{:,\mathrm{Pa}_G(v)}\big), \tag{6}$$

where each factor is obtained by integrating out θ_v in the local regression problem induced by $\mathrm{Pa}_G(v)$. This implies that local changes to a graph affecting only one node's parent set modify only a constant number of terms in $\log p(G) + \log p(X \mid G)$. Practical score-based algorithms exploit this property to compute score differences under add/delete/reverse moves without recomputing global quantities.

Bayesian learning principles of this kind are also used in biological inference settings, including proteomics applications that explicitly review Bayesian methodology through Bayesian networks.[25]

In high dimensions, structure priors play an essential role because many distinct graphs can achieve similar likelihood. Priors that penalize edge counts, restrict in-degree, or encode plausible topology can dominate posterior concentration and stabilize discrete search.

3.4 Regularization via edge-inclusion priors

A basic and widely used sparsity prior treats edges as exchangeable conditional on an unknown inclusion probability.[26] Let E denote the edge set and let $N = d(d-1)$ be the number of possible directed edges excluding self-loops. If conditional on θ edges are independent Bernoulli and $\theta \sim \mathrm{Beta}(a, b)$, then integrating out θ yields the Beta–Bernoulli marginal

$$p(G) \propto \frac{\mathrm{B}(a + |E|,\ b + N - |E|)}{\mathrm{B}(a, b)}. \tag{7}$$

This provides a calibrated global penalty on $|E|$ without fixing a single inclusion probability a priori. More structured priors can condition edge probabilities on covariates or on hierarchical relationships among nodes, and degree-based preferences can encode hub-like behavior. MT-BN will use a product of sparsity, proximity, and hub components as an unnormalized potential on DAGs at each resolution.

3.5 Variational inference and the evidence lower bound

Let Y denote latent variables and parameters in a probabilistic model with joint density $p(X, Y)$. Exact posterior inference for $p(Y \mid X)$ is often intractable. Variational inference approximates the posterior by a tractable family $\mathcal{Q}$ and solves [27]

$$\max_{q \in \mathcal{Q}} \mathcal{L}(q) = \mathbb{E}_q[\log p(X, Y)] - \mathbb{E}_q[\log q(Y)]\,. \tag{8}$$

The quantity $\mathcal{L}(q)$ is the evidence lower bound (ELBO), and it satisfies

$$\log p(X) = \mathcal{L}(q) + \mathrm{KL}(q(Y) \,\|\, p(Y \mid X))\,, \tag{9}$$

so maximizing $\mathcal{L}(q)$ is equivalent to minimizing the KL divergence from q to the true posterior.

In a mean-field family $q(Y) = \prod_i q_i(Y_i)$, coordinate ascent updates satisfy the standard optimality condition

$$\log q_i^\star(y_i) = \mathbb{E}_{q_{-i}}[\log p(X, Y)] + \text{const}, \tag{10}$$

where $q_{-i} = \prod_{j \neq i} q_j$. When the model is conditionally conjugate, the right-hand side corresponds to a familiar Bayesian update in the exponential family, yielding closed-form updates for q_i. MT-BN will use this machinery conditional on a fixed discrete structure.

3.6 Metropolis–Hastings on discrete structures

Let S be a discrete structure (such as a graph or hierarchy) and let $\pi(S)$ be a target distribution on structures. Metropolis–Hastings constructs a Markov chain with stationary distribution π by proposing $S' \sim Q(S \to \cdot)$ and accepting with probability [28]

$$\alpha(S, S') = \min\left\{1,\ \frac{\pi(S')}{\pi(S)} \cdot \frac{Q(S' \to S)}{Q(S \to S')}\right\}. \tag{11}$$

If $\pi(S) \propto \exp(s(S))$ for a score function $s(S)$, then the acceptance ratio depends only on the score difference $s(S') - s(S)$ and the proposal ratio. In Bayesian structure learning, a natural choice is $s(S) = \log p(S) + \log p(X \mid S)$, but computing $\log p(X \mid S)$ exactly is often intractable when S couples to latent variables. A common scalable alternative is to use a variational surrogate score in place of the exact marginal likelihood. MT-BN follows this approach by using an ELBO-based structure score, yielding a practical exploration procedure designed for MAP discovery and stability summaries in regimes where exact posterior sampling is not computationally feasible.

4 Problem Statement

We are given an observational dataset

$$X \in \mathbb{R}^{n \times p},$$

with n samples and p measured variables, in the high-dimensional regime $p \gg n$. Our goal is to infer directed dependence structure at multiple resolutions while simultaneously learning a hierarchy of modules that organizes variables into nested groups. We emphasize that, absent interventions, any causal interpretation requires additional modeling assumptions; accordingly, we state the target in terms of a structured probabilistic model class and the corresponding posterior over model structures.

4.1 Multi-resolution modular structure

Fix a user-chosen maximum depth $L \geq 1$. A hierarchical modular structure is represented by a rooted tree $\mathcal{T}$ that induces, for each level $\ell \in \{1, \ldots, L\}$, a partition of the p observed variables into M_ℓ modules

$$\{C_1^{(\ell)}, \ldots, C_{M_\ell}^{(\ell)}\}, \qquad \bigcup_{m=1}^{M_\ell} C_m^{(\ell)} = \{1, \ldots, p\}, \qquad C_m^{(\ell)} \cap C_{m'}^{(\ell)} = \emptyset \;\; (m \neq m').$$

Nestedness means that partitions refine with depth: for any $\ell < L$ and any variable i, the level-$(\ell+1)$ module containing i is a subset of the level-ℓ module containing i. Equivalently, each observed variable i has a unique path

$$\pi(i) = \big(z_i^{(1)}, \ldots, z_i^{(L)}\big),$$

where $z_i^{(\ell)} \in \{1, \ldots, M_\ell\}$ indexes its module at level ℓ. The hierarchy $\mathcal{T}$ is fully determined by the collection of paths $\{\pi(i)\}_{i=1}^p$ together with the parent-child relations between modules across levels.

4.2 Innovation-based directed structure at each resolution

At each level ℓ, MT-BN associates each module $m \in \{1, \ldots, M_\ell\}$ with a latent trajectory across samples. We distinguish (i) a full module state $Z_m^{(\ell)} \in \mathbb{R}^{n \times r_\ell}$ and (ii) an innovation trajectory $U_m^{(\ell)} \in \mathbb{R}^{n \times r_\ell}$, where r_ℓ is a user-chosen latent dimension. The modeling principle is that within-level directed structure is defined on innovations $U^{(\ell)}$ rather than on full states $Z^{(\ell)}$. Intuitively, $Z_m^{(\ell)}$ contains signal inherited from coarser resolutions, whereas $U_m^{(\ell)}$ isolates the variation newly expressed at resolution ℓ after conditioning on higher-level latents.

Accordingly, for each level ℓ we seek a directed acyclic graph (DAG)

$$G^{(\ell)} = (V^{(\ell)}, E^{(\ell)}), \qquad V^{(\ell)} = \{1, \ldots, M_\ell\},$$

that parameterizes a within-level structural equation model for innovations at that resolution. The collection $G = \{G^{(\ell)}\}_{\ell=1}^L$ defines a multi-scale directed dependency representation.

4.3 Desired outputs

The outputs of the learning problem are:

1. A hierarchy $\mathcal{T}$ of nested partitions of $\{1, \ldots, p\}$ with depth L (or with effective depth up to L under truncation).
2. A set of within-level DAGs $\{G^{(\ell)}\}_{\ell=1}^L$ defined on modules at each resolution, interpreted as directed structure on innovations at that resolution.
3. Latent module trajectories $\{Z_m^{(\ell)}, U_m^{(\ell)}\}$ and associated parameters Θ sufficient to (approximately) explain the observed data via a hierarchical latent-variable model.

In addition to point estimates, we aim to provide uncertainty summaries over discrete structures and latent quantities induced by the model and inference procedure.

4.4 Model-based objective and inferential target

We formalize learning as inference in a joint probabilistic model over

$$(\mathcal{T}, G, Z, U, \Theta),$$

where $\mathcal{T}$ is the hierarchy, G are the within-level DAGs, (Z, U) are latent module states and innovations, and Θ are continuous parameters. Given a prior $p(\mathcal{T}, G, \Theta)$ and a likelihood $p(X \mid Z, U, \Theta, \mathcal{T})$ induced by the hierarchical latent construction, the inferential target is the posterior

$$p(\mathcal{T}, G, Z, U, \Theta \mid X).$$

Because exact inference is intractable in the regime of interest, we seek scalable approximate inference procedures that (i) locate high-scoring discrete structures $(\mathcal{T}, G)$ and (ii) produce stable summaries of multi-scale directed dependencies and hierarchical module structure.

4.5 Evaluation criteria

We evaluate solutions along three axes. First, statistical accuracy: when ground truth is available (e.g., benchmark networks), we evaluate recovery of directed edges and modular structure. Second, interpretability: inferred modules should be coherent and the multi-scale graphs should yield intelligible drivers and pathways. Third, scalability: runtime and memory should remain practical in high-dimensional settings by exploiting modularization and localized score updates.

5 MT-BN Framework

5.1 Method Overview and Intuition

Bayesian network structure learning becomes brittle in the high-dimensional regime $p \gg n$ because the space of directed acyclic graphs (DAGs) grows super-exponentially in p while the amount of statistical signal available to validate and orient edges remains limited. In many domains, however, dependency structure is not arbitrary. Empirical networks often exhibit modular organization, sparse long-range connectivity across modules, and heterogeneous degree patterns in which a small number of hubs mediate a large fraction of directed influence. MT-BN is designed to exploit these regularities by replacing a single flat graph-learning problem with a multi-level representation that is both statistically efficient and computationally tractable, with tractability driven by operating on module-level latent variables and by topology-aware priors that shrink the effective search space.

MT-BN jointly learns three coupled objects. First, it learns a multi-level hierarchy of nested modules that partitions the p observed variables at multiple resolutions. Second, it associates each module at each resolution with a latent state that summarizes the module's activity across samples. Third, it learns a within-level DAG among modules at each resolution, yielding a directed influence network at multiple scales. The hierarchy captures vertical organization, while the within-level DAGs capture horizontal directed dependencies among modules at the same resolution. In contrast to two-stage pipelines that cluster variables and subsequently learn separate graphs, MT-BN couples modularization, representation learning, and within-level structure inference within a single probabilistic model. Figure 1 provides a schematic contrast between flat structure learning and MT-BN's hierarchical, multi-resolution representation.

A key modeling principle is that within-level directed structure at a given resolution is defined on what is newly expressed at that resolution rather than on signal inherited from coarser scales. For each level ℓ, each module m has a latent state $Z_m^{(\ell)}$ that decomposes into a component inherited from its parent module and a level-specific innovation $U_m^{(\ell)}$. MT-BN does not place directed edges on the full latent states $Z^{(\ell)}$, which combine inherited and level-specific variation. Instead, within-level DAGs are defined on innovation latents $U^{(\ell)}$, which isolate residual variation after accounting for inherited signal. If edges were learned directly on $Z^{(\ell)}$, sibling modules could appear strongly dependent purely due to shared inheritance from their parent, leading to spurious within-level edges and potentially misleading directions. By defining within-level structure exclusively on innovations, MT-BN targets directed relationships among modules at level ℓ conditional on higher-level variation.

Throughout, MT-BN interprets within-level directionality under the modeling assumptions made explicit later: innovations follow a within-level structural equation model consistent with a DAG, and higher-level latents act as conditioning variables that explain shared inherited signal across descendants. Under standard causal assumptions

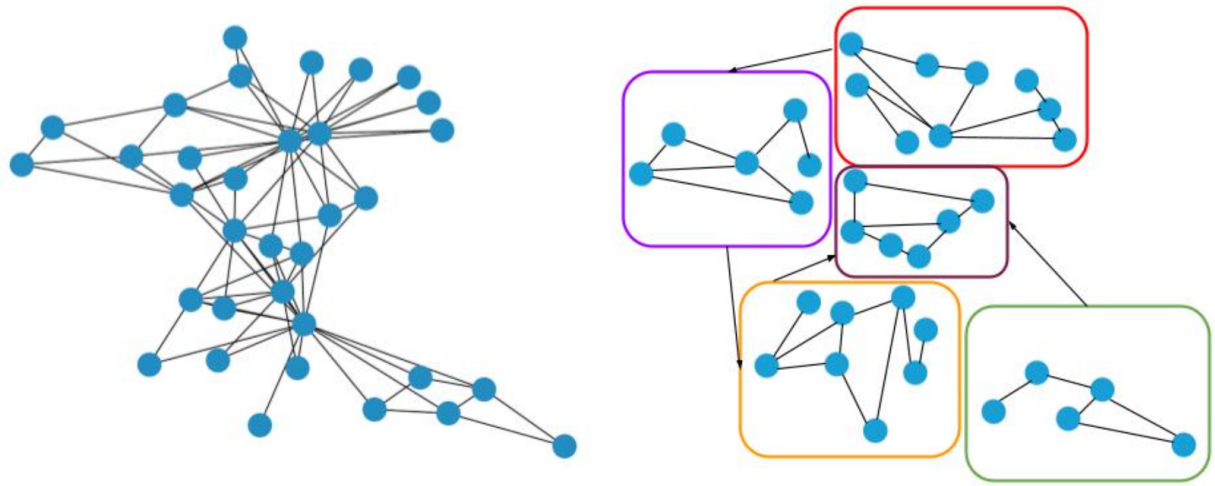

Figure 1: Schematic comparison of flat structure learning (left) versus MT-BN's multi-resolution modular representation (right). MT-BN learns a hierarchy of modules and within-level directed structure at each resolution on innovation latents rather than on inherited shared signal.

for SEM-based graph learning (including appropriate noise structure and the absence of additional unmodeled confounding at the corresponding resolution after conditioning on higher-level latents), directed edges on innovations admit a causal interpretation at that resolution; without these assumptions, they should be read as directed predictive dependencies induced by the model class.

MT-BN further regularizes and accelerates structure learning by incorporating topology-aware priors. These priors encode sparsity, heterogeneous degree patterns consistent with hub structure, and hierarchy-induced block structure favoring denser connectivity among nearby modules in the hierarchy and sparser connectivity across distant branches. These priors shrink posterior mass over implausible graphs, thereby reducing the effective search space and improving interpretability. The following subsections formalize the MT-BN model, specify priors, and present a hybrid inference algorithm for approximate posterior inference and MAP estimation, along with computational considerations.

5.2 Setup and Notation

We observe a data matrix $X \in \mathbb{R}^{n \times p}$ with n samples and p observed variables. Rows correspond to samples and columns correspond to variables. MT-BN represents multi-scale organization through a hierarchy of maximum depth L. The hierarchy $\mathcal{T}$ induces, at each level $\ell \in \{1, \dots, L\}$, a partition of the p variables into M_ℓ modules, where M_ℓ is the realized number of modules at level ℓ under $\mathcal{T}$ (and is therefore random a priori under the nCRP prior in Section 5.3). We write $V^{(\ell)} = \{1, \dots, M_\ell\}$ for the set of level-ℓ module indices.

Each module $m \in V^{(\ell)}$ has a latent trajectory across samples

$$Z_m^{(\ell)} \in \mathbb{R}^{n \times r_\ell},$$

where r_ℓ is a user-chosen latent dimension for level ℓ. We interpret the t-th row $Z_{m,t}^{(\ell)} \in \mathbb{R}^{r_\ell}$ as the level-ℓ latent state of module m in sample t. MT-BN separates inherited signal from resolution-specific signal by introducing an innovation latent

$$U_m^{(\ell)} \in \mathbb{R}^{n \times r_\ell},$$

whose rows $U_{m,t}^{(\ell)}$ capture variation newly expressed at level ℓ after conditioning on higher levels. Linear maps act on latent coordinates (columns) via right-multiplication. For example, if $U_j^{(\ell)} \in \mathbb{R}^{n \times r_\ell}$ and $A \in \mathbb{R}^{r_\ell \times r_\ell}$, then $U_j^{(\ell)} A \in \mathbb{R}^{n \times r_\ell}$.

The hierarchy induces nested partitions of the observed variables. Each observed variable $i \in \{1, \dots, p\}$ is assigned a path through the hierarchy,

$$\pi(i) = \left(z_i^{(1)}, z_i^{(2)}, \dots, z_i^{(L)}\right),$$

where $z_i^{(\ell)} \in V^{(\ell)}$ denotes the level-ℓ module containing i. The induced level-ℓ cluster is $C_m^{(\ell)} = \{i : z_i^{(\ell)} = m\}$, so $\{C_1^{(\ell)}, \ldots, C_{M_\ell}^{(\ell)}\}$ partitions $\{1, \ldots, p\}$. Nestedness means membership refines down the hierarchy: for each i, $C_{z_i^{(\ell+1)}}^{(\ell+1)} \subseteq C_{z_i^{(\ell)}}^{(\ell)}$. Each module m at level $\ell \geq 2$ has a unique parent module, denoted $\mathrm{pa}(m, \ell) \in V^{(\ell-1)}$. For notational convenience, we treat level $\ell = 1$ as having an implicit root ancestor and later impose the convention $Z_m^{(1)} := U_m^{(1)}$.

At each level ℓ, MT-BN also learns a within-level directed acyclic graph $G^{(\ell)}$ on the M_ℓ modules. We write $\mathrm{Pa}_{G^{(\ell)}}(m) \subseteq V^{(\ell)} \setminus \{m\}$ for the parent set of node m in $G^{(\ell)}$. These graphs encode within-level directed structure on innovations $\{U_m^{(\ell)}\}$ rather than on full states $\{Z_m^{(\ell)}\}$, as formalized in Section 5.5.

We use standard notation I_k for the $k \times k$ identity matrix, $\mathrm{vec}(\cdot)$ for column-stacking vectorization, and $\|\cdot\|_F$ for the Frobenius norm. We also fix distributional conventions used throughout. The Beta function is $\mathrm{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$. For $x \sim \mathrm{InvGamma}(a, b)$ we use the shape–scale convention with density proportional to $x^{-a-1} \exp(-b/x)$ on $x > 0$. Finally, $\mathcal{MN}(M, \Sigma_r, \Sigma_c)$ denotes the matrix-normal distribution on $\mathbb{R}^{n \times r}$ satisfying $\mathrm{vec}(Y) \sim \mathcal{N}(\mathrm{vec}(M), \Sigma_c \otimes \Sigma_r)$.

5.3 Hierarchical Partitions and the nCRP Prior

MT-BN places a nested Chinese Restaurant Process (nCRP) prior on the hierarchical assignments of observed variables. Concretely, each observed variable $i \in \{1, \ldots, p\}$ is treated as an exchangeable object that selects a depth-L path

$$\pi(i) = \big(z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(L)}\big),$$

thereby inducing a hierarchy $\mathcal{T}$ of nested partitions $\{C_m^{(\ell)}\}_{m=1}^{M_\ell}$ at each level ℓ. We use level-specific concentration parameters $\alpha_\ell > 0$, $\ell = 1, \ldots, L$, and we fix the maximum depth L as a user-chosen model parameter.

For notational convenience, define a single implicit root at level 0 and set $z_i^{(0)} = 1$ for all i. Fix a level $\ell \in \{1, \ldots, L\}$ and a parent module a at level $\ell - 1$. Let

$$S_a^{(\ell)} = \{i : z_i^{(\ell-1)} = a\}$$

be the set of variables whose paths pass through parent a. Within each parent a, the assignments $\{z_i^{(\ell)}\}_{i \in S_a^{(\ell)}}$ follow a Chinese Restaurant Process with concentration α_ℓ. Let $n_{a,k}^{(\ell)}$ be the number of variables in $S_a^{(\ell)}$ currently assigned to an existing child k under a, excluding i. Then the conditional assignment probabilities are

$$\Pr\Big(z_i^{(\ell)} = k \,\Big|\, \{z_{i'}^{(\ell)}\}_{i' \in S_a^{(\ell)} \setminus \{i\}}\Big) = \frac{n_{a,k}^{(\ell)}}{\sum_{k'} n_{a,k'}^{(\ell)} + \alpha_\ell}, \tag{12}$$

and the probability of instantiating a new child module under a is

$$\Pr\Big(z_i^{(\ell)} = \text{new} \,\Big|\, \{z_{i'}^{(\ell)}\}_{i' \in S_a^{(\ell)} \setminus \{i\}}\Big) = \frac{\alpha_\ell}{\sum_{k'} n_{a,k'}^{(\ell)} + \alpha_\ell}. \tag{13}$$

When a new child is instantiated, it receives the next unused index among the children of a, which updates the number of modules M_ℓ implied by the hierarchy.

To prevent degenerate fragmentation and to ensure well-posed local structure learning, we enforce a minimum module size $m_{\min}$. Formally, we use a truncated nCRP prior supported only on hierarchies satisfying

$$|C_m^{(\ell)}| \geq m_{\min} \qquad \text{for all instantiated modules } m \text{ and levels } \ell, \tag{14}$$

that is,

$$p(\mathcal{T}) \propto p_{\mathrm{nCRP}}(\mathcal{T}) \prod_{\ell=1}^{L} \prod_{m=1}^{M_\ell} \mathbf{1}\Big\{|C_m^{(\ell)}| \geq m_{\min}\Big\}. \tag{15}$$

In inference, this truncation is implemented by restricting reassignment proposals to moves that preserve (14).

Unless otherwise stated, $\{\alpha_\ell\}_{\ell=1}^{L}$ are treated as fixed hyperparameters. A Gamma hyperprior $\alpha_\ell \sim \mathrm{Gamma}(\nu_\alpha, \omega_\alpha)$ can be incorporated with standard auxiliary-variable updates, but is not required for the formulation used here.

5.4 Topology-Aware Priors for Multi-Scale DAG Structure

Conditioned on the hierarchy $\mathcal{T}$, MT-BN places topology-aware priors on the within-level directed acyclic graphs $\{G^{(\ell)}\}_{\ell=1}^{L}$. Each $G^{(\ell)}$ is a DAG on the M_ℓ level-ℓ modules and parameterizes the within-level structural equation model on innovations in (23). The role of the graph prior is twofold: it regularizes structure learning toward sparse, modular, hub-dominated networks commonly observed in practice, and it yields localized score updates under single-edge proposals.

Fix a level ℓ. Let $V^{(\ell)} = \{1, \dots, M_\ell\}$ and let $E^{(\ell)} \subseteq V^{(\ell)} \times V^{(\ell)}$ denote a directed edge set with no self-loops. We restrict the support to acyclic graphs and optionally enforce a maximum in-degree constraint $d_{\max}$. Define the admissible set

$$\mathcal{G}_\ell = \left\{ G^{(\ell)} = (V^{(\ell)}, E^{(\ell)}) : G^{(\ell)} \in \mathsf{DAG}(V^{(\ell)}),\ \max_{v \in V^{(\ell)}} d_{\text{in}}^{(\ell)}(v) \le d_{\max} \right\}.$$

We specify the multi-level graph prior through an unnormalized potential and use it only through ratios inside the variational structure score. Concretely, we define the joint prior on discrete structure by

$$p\big(\mathcal{T}, \{G^{(\ell)}\}_{\ell=1}^{L}\big) \propto p(\mathcal{T}) \prod_{\ell=1}^{L} \widetilde{p}\big(G^{(\ell)} \mid \mathcal{T}\big)\, \mathbf{1}\Big\{G^{(\ell)} \in \mathcal{G}_\ell\Big\}, \tag{16}$$

where $\widetilde{p}(G^{(\ell)} \mid \mathcal{T})$ is given in (17). Since the space of admissible DAGs on M_ℓ nodes is finite, a normalizing constant exists, but it is not required for scoring or inference.

Factorization as a product of structural preferences. We specify $\widetilde{p}(G^{(\ell)} \mid \mathcal{T})$ as a product of three interpretable components:

$$\widetilde{p}\big(G^{(\ell)} \mid \mathcal{T}\big) = p_{\text{sp}}\big(G^{(\ell)}\big)\ p_{\text{prox}}\big(G^{(\ell)} \mid \mathcal{T}\big)\ p_{\text{hub}}\big(G^{(\ell)}\big). \tag{17}$$

This is a product-of-experts prior: p_{sp} favors global sparsity, p_{prox} favors hierarchy-induced block structure, and p_{hub} favors heterogeneous degree profiles. Because all three terms are ultimately functions of $(V^{(\ell)}, E^{(\ell)})$, the product defines a proper prior on the finite admissible set $\mathcal{G}_\ell$ up to a normalizing constant that cancels in score differences.

Global sparsity prior. Let $N_\ell = M_\ell(M_\ell - 1)$ denote the number of possible directed edges excluding self-loops. We use the Beta–Bernoulli marginal likelihood over edges:

$$p_{\text{sp}}\big(G^{(\ell)}\big) = \frac{\mathrm{B}\big(a_\ell + |E^{(\ell)}|,\ b_\ell + N_\ell - |E^{(\ell)}|\big)}{\mathrm{B}(a_\ell, b_\ell)}. \tag{18}$$

Equivalently, this is obtained by integrating out a global edge-inclusion probability $\theta_\ell \sim \mathrm{Beta}(a_\ell, b_\ell)$ under independent Bernoulli edge indicators conditional on θ_ℓ. This term supplies a calibrated global penalty on $|E^{(\ell)}|$ without fixing a single edge probability a priori. For the top resolution $\ell = 1$, proximity is not defined because modules have no parents in $\mathcal{T}$. We therefore set

$$p_{\text{prox}}\big(G^{(1)} \mid \mathcal{T}\big) := 1,$$

and define the proximity construction below only for $\ell \ge 2$.

Hierarchy-induced proximity prior. MT-BN encourages denser connectivity among modules that are nearby in the hierarchy and sparser connectivity between distant branches. For two level-ℓ modules $u, v \in V^{(\ell)}$, let $\mathrm{par}_\ell(u)$ denote their parent at level $\ell - 1$, and let $\mathrm{gpar}_\ell(u)$ denote their grandparent at level $\ell - 2$ (defined only for $\ell \ge 3$). We define a coarse proximity class map $\kappa_\ell(u, v) \in \{1, 2, 3\}$ by

$$\kappa_\ell(u, v) = \begin{cases} 1, & \text{if } \mathrm{par}_\ell(u) = \mathrm{par}_\ell(v), \\ 2, & \text{if } \ell \ge 3,\ \mathrm{gpar}_\ell(u) = \mathrm{gpar}_\ell(v) \text{ and } \mathrm{par}_\ell(u) \ne \mathrm{par}_\ell(v), \\ 3, & \text{otherwise.} \end{cases} \tag{19}$$

For each bin $b \in \{1, 2, 3\}$, define

$$N_{\ell,b} = \Big|\{(u, v) \in V^{(\ell)} \times V^{(\ell)} : u \ne v,\ \kappa_\ell(u, v) = b\}\Big|, \qquad E_{\ell,b} = \Big|\{(u, v) \in E^{(\ell)} : \kappa_\ell(u, v) = b\}\Big|.$$

We introduce bin-specific edge probabilities $\theta_{\ell,b} \sim \mathrm{Beta}(\eta^{(1)}_{\ell,b}, \eta^{(0)}_{\ell,b})$ and conditionally independent Bernoulli edges within each bin given $\theta_{\ell,b}$. Integrating out $\{\theta_{\ell,b}\}$ yields

$$p_{\mathrm{prox}}\big(G^{(\ell)} \mid \mathcal{T}\big) = \prod_{b=1}^{3} \frac{\mathrm{B}\Big(\eta^{(1)}_{\ell,b} + E_{\ell,b},\ \eta^{(0)}_{\ell,b} + N_{\ell,b} - E_{\ell,b}\Big)}{\mathrm{B}\Big(\eta^{(1)}_{\ell,b}, \eta^{(0)}_{\ell,b}\Big)}. \tag{20}$$

This prior is sensitive only to counts by proximity class and therefore admits constant-time updates under single-edge proposals.

Hub-structure prior on degree heterogeneity. Real networks frequently exhibit heavy-tailed out-degree profiles. To encourage this behavior, we place a prior preference on the out-degrees $d^{(\ell)}_{\mathrm{out}}(u)$. Because degrees are discrete, we define p_{hub} as a discretized log-normal potential on $d^{(\ell)}_{\mathrm{out}}(u) + 1$:

$$p_{\mathrm{hub}}\big(G^{(\ell)}\big) = \prod_{u=1}^{M_\ell} \frac{\varphi_{\mu_\ell,\sigma_\ell}\big(\log(d^{(\ell)}_{\mathrm{out}}(u)+1)\big)}{\sum_{k=0}^{M_\ell-1} \varphi_{\mu_\ell,\sigma_\ell}\big(\log(k+1)\big)}, \tag{21}$$

where $\varphi_{\mu,\sigma}(x) = (\sigma\sqrt{2\pi})^{-1}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)$ is the normal density. The denominator in (21) converts the log-normal preference into a proper discrete distribution on $k \in \{0,\dots,M_\ell - 1\}$ for each node. In practice, this term is still used as a local structural preference within $\widetilde{p}$, and its contribution to acceptance ratios depends only on the degrees of the nodes whose out-degrees change under the proposal.

Optionally, we place hyperpriors on $(\mu_\ell, \sigma^2_\ell)$ to adapt the degree profile to the domain:

$$\mu_\ell \sim \mathcal{N}(\mu_0, \tau_0^2), \qquad \sigma^2_\ell \sim \mathrm{InvGamma}(a_\sigma, b_\sigma), \tag{22}$$

though fixed (μ_ℓ, σ_ℓ) is sufficient for the formulation used here.

Independence across resolutions. Conditioned on $\mathcal{T}$, MT-BN assumes within-level graph potentials contribute multiplicatively across levels through (17). Equivalently, the joint discrete prior in (16) treats the level-wise graph terms as independent components given $\mathcal{T}$ up to an overall normalizing constant on the finite admissible space $\prod_{\ell=1}^{L} \mathcal{G}_\ell$. This reflects the modeling choice that horizontal causality is learned separately at each resolution, while cross-resolution dependence is mediated by the hierarchy and the vertical inheritance model in (25).

5.5 Joint Distribution and Posterior Formulation

This subsection specifies the complete MT-BN probabilistic model and the posterior inference target. The model combines (i) an nCRP prior over the hierarchy $\mathcal{T}$, (ii) topology-aware priors over within-level DAGs $\{G^{(\ell)}\}_{\ell=1}^{L}$, (iii) a multi-level latent inheritance construction separating inherited signal from resolution-specific innovations, and (iv) a linear measurement model mapping bottom-level latent states to observed variables.

Matrix-normal convention. For an $n \times r$ random matrix Y, we write

$$Y \sim \mathcal{MN}_{n,r}(M, \Sigma, \Omega)$$

to mean $\mathrm{vec}(Y) \sim \mathcal{N}(\mathrm{vec}(M), \Omega \otimes \Sigma)$. In particular, $\mathcal{MN}_{n,r}(M, I_n, \sigma^2 I_r)$ corresponds to independent rows with isotropic covariance $\sigma^2 I_r$ in latent dimensions.

Horizontal structural equations on innovations. Fix a level ℓ. Let $G^{(\ell)}$ be a DAG on $V^{(\ell)} = \{1,\dots,M_\ell\}$ with parent sets $\mathrm{Pa}_{G^{(\ell)}}(m)$. Conditional on $G^{(\ell)}$, the innovation latents $\{U^{(\ell)}_m\}_{m=1}^{M_\ell}$ are generated by a linear Gaussian SEM with independent sample rows:

$$U^{(\ell)}_m \,\Big|\, \{U^{(\ell)}_j\}_{j\in\mathrm{Pa}_{G^{(\ell)}}(m)}, \{A^{(\ell)}_{j\to m}\}, \sigma^2_{\eta,\ell} \sim \mathcal{MN}_{n,r_\ell}\left(\sum_{j\in\mathrm{Pa}_{G^{(\ell)}}(m)} U^{(\ell)}_j A^{(\ell)}_{j\to m},\ I_n,\ \sigma^2_{\eta,\ell} I_{r_\ell}\right). \tag{23}$$

Because $G^{(\ell)}$ is acyclic, there exists a topological ordering $m_1, \ldots, m_{M_\ell}$ such that each node's parents appear earlier in the order. Consequently, the within-level joint density of $U^{(\ell)} = \{U_m^{(\ell)}\}_{m=1}^{M_\ell}$ is well-defined by the DAG factorization

$$p\big(U^{(\ell)} \mid G^{(\ell)}, A^{(\ell)}, \sigma^2_{\eta,\ell}\big) = \prod_{m=1}^{M_\ell} p\Big(U_m^{(\ell)} \,\Big|\, \{U_j^{(\ell)}\}_{j\in \mathrm{Pa}_{G^{(\ell)}}(m)}, A^{(\ell)}, \sigma^2_{\eta,\ell}\Big), \tag{24}$$

where each conditional factor is given by (23). MT-BN defines within-level causality exclusively on the innovation variables $U^{(\ell)}$, not on the inherited full states $Z^{(\ell)}$.

Vertical inheritance model. For each level $\ell \geq 2$ and module $m \in \{1, \ldots, M_\ell\}$, the full latent state $Z_m^{(\ell)} \in \mathbb{R}^{n\times r_\ell}$ is generated as inherited signal from its parent state plus a level-specific innovation:

$$Z_m^{(\ell)} \,\Big|\, Z^{(\ell-1)}_{\mathrm{pa}(m,\ell)}, U_m^{(\ell)}, B^{(\ell)}_{\mathrm{pa}(m,\ell)\to m}, \sigma^2_{\xi,\ell} \sim \mathcal{MN}_{n,r_\ell}\Big(Z^{(\ell-1)}_{\mathrm{pa}(m,\ell)} B^{(\ell)}_{\mathrm{pa}(m,\ell)\to m} + U_m^{(\ell)},\ I_n,\ \sigma^2_{\xi,\ell} I_{r_\ell}\Big), \tag{25}$$

where $B^{(\ell)}_{\mathrm{pa}(m,\ell)\to m} \in \mathbb{R}^{r_{\ell-1}\times r_\ell}$. At the top level, we impose the deterministic identification

$$Z_m^{(1)} := U_m^{(1)} \qquad \text{for all } m \in \{1, \ldots, M_1\}, \tag{26}$$

which is treated as a definitional constraint (equivalently, a Dirac measure) rather than as a separate noisy regression.

Measurement model. Each observed variable loads on the bottom-level module state. Let $m(i) = z_i^{(L)}$ denote the bottom-level module containing observed variable i. Then

$$X_{:,i} \,\Big|\, Z^{(L)}_{m(i)}, w_i, \psi_i \sim \mathcal{N}\Big(Z^{(L)}_{m(i)} w_i,\ \psi_i I_n\Big), \qquad w_i \in \mathbb{R}^{r_L},\ \psi_i > 0, \tag{27}$$

with conditional independence across $i \in \{1, \ldots, p\}$ given $\{Z_m^{(L)}\}$ and parameters.

Priors on discrete structure. The hierarchy $\mathcal{T}$ is induced by the nCRP paths $\{\pi(i)\}_{i=1}^p$ with the minimum-size constraint in (14). The discrete structure prior is specified jointly as in (16): an nCRP prior on $\mathcal{T}$ together with, for each level ℓ, an unnormalized topology-aware potential $\widetilde{p}(G^{(\ell)} \mid \mathcal{T})$ restricted to admissible DAGs $G^{(\ell)} \in \mathcal{G}_\ell$. This construction is used only through log-ratios and therefore does not require evaluating normalizing constants.

Priors on continuous parameters. Let Θ denote the collection of continuous parameters

$$\Theta = \Big\{\{A^{(\ell)}_{j\to m}\},\ \{B^{(\ell)}_{\mathrm{pa}\to m}\},\ \{w_i\},\ \{\sigma^2_{\eta,\ell}\},\ \{\sigma^2_{\xi,\ell}\},\ \{\psi_i\}\Big\}.$$

We place independent zero-mean Gaussian priors on linear maps and loadings:

$$\mathrm{vec}\Big(A^{(\ell)}_{j\to m}\Big) \sim \mathcal{N}\Big(0, \lambda^{-1}_{A,\ell} I\Big), \qquad \mathrm{vec}\Big(B^{(\ell)}_{\mathrm{pa}\to m}\Big) \sim \mathcal{N}\Big(0, \lambda^{-1}_{B,\ell} I\Big), \qquad w_i \sim \mathcal{N}\big(0, \lambda_w^{-1} I_{r_L}\big), \tag{28}$$

and inverse-gamma priors on variances:

$$\sigma^2_{\eta,\ell} \sim \mathrm{InvGamma}(a_{\eta,\ell}, b_{\eta,\ell}), \qquad \sigma^2_{\xi,\ell} \sim \mathrm{InvGamma}(a_{\xi,\ell}, b_{\xi,\ell}), \qquad \psi_i \sim \mathrm{InvGamma}(a_\psi, b_\psi). \tag{29}$$

We write $p(\Theta)$ for the product of these parameter priors. We interpret $A^{(\ell)}_{j\to m}$ as a free parameter only when $(j \to m) \in E^{(\ell)}$; absent edges correspond to the constraint $A^{(\ell)}_{j\to m} \equiv 0$ and introduce no free parameter.

Full joint distribution with named components. Let $G = \{G^{(\ell)}\}_{\ell=1}^{L}$, $U = \{U_m^{(\ell)}\}_{\ell,m}$, and $Z = \{Z_m^{(\ell)}\}_{\ell,m}$. The MT-BN model defines the joint density

$$
\begin{aligned}
&p\big(X, Z, U, G, \mathcal{T}, \Theta\big) \\
&= \underbrace{p(\mathcal{T})}_{\substack{\text{nCRP hierarchy prior with}\\ \text{minimum-size constraint}}} \underbrace{\prod_{\ell=1}^{L} \widetilde{p}\big(G^{(\ell)} \mid \mathcal{T}\big)\, \mathbf{1}\Big\{G^{(\ell)} \in \mathcal{G}_\ell\Big\}}_{\substack{\text{topology-aware within-level DAG prior potentials}\\ \text{and DAG constraints}}} \underbrace{p(\Theta)}_{\text{parameter priors}} \\
&\quad \times \underbrace{\prod_{\ell=1}^{L} p\big(U^{(\ell)} \mid G^{(\ell)}, A^{(\ell)}, \sigma_{\eta,\ell}^2\big)}_{\text{within-level SEM on innovations}} \underbrace{\prod_{\ell=2}^{L} \prod_{m=1}^{M_\ell} p\big(Z_m^{(\ell)} \mid Z_{\mathrm{pa}(m,\ell)}^{(\ell-1)}, U_m^{(\ell)}, B^{(\ell)}, \sigma_{\xi,\ell}^2\big)}_{\text{vertical inheritance model}} \\
&\quad \times \underbrace{\prod_{i=1}^{p} p\big(X_{:,i} \mid Z_{m(i)}^{(L)}, w_i, \psi_i\big)}_{\text{measurement model}} \underbrace{\prod_{m=1}^{M_1} \delta\Big(Z_m^{(1)} - U_m^{(1)}\Big)}_{\text{top-level identification}},
\end{aligned}
\tag{30}
$$

where $p(U^{(\ell)} \mid \cdot)$ factorizes according to (24) with conditionals given by (23), the inheritance conditionals are given by (25), and the measurement conditionals are given by (27). The Dirac factors in the final product encode the deterministic constraint (26).

Posterior distribution and estimation targets. The posterior is

$$
p\big(\mathcal{T}, G, Z, U, \Theta \mid X\big) = \frac{p\big(X, Z, U, G, \mathcal{T}, \Theta\big)}{p(X)}, \qquad p(X) = \sum_{\mathcal{T}} \sum_{G} \int p\big(X, Z, U, G, \mathcal{T}, \Theta\big)\, d(Z, U, \Theta), \tag{31}
$$

where the sums are over admissible hierarchies and DAGs under the constraints of Sections 5.3 and 5.4. MT-BN supports both approximate uncertainty quantification (posterior edge probabilities and co-assignment probabilities) and point estimation (MAP hierarchies and graphs), as described in Section 5.6.

5.6 Inference and Optimization

The posterior over MT-BN parameters couples discrete objects (the hierarchy $\mathcal{T}$ and within-level DAGs $\{G^{(\ell)}\}_{\ell=1}^{L}$) with high-dimensional continuous latent variables and parameters (Z, U, Θ). Exact inference is intractable in the regimes of interest. MT-BN therefore uses a blocked hybrid strategy: (i) a variational approximation for continuous variables conditional on discrete structure, and (ii) Metropolis-style stochastic search over discrete structures guided by a variational score. This procedure is designed to yield scalable MAP structure estimates and calibrated uncertainty summaries under a variationally defined surrogate distribution, rather than exact posterior samples under the true marginal likelihood.

Variational objective for fixed structure. Fix a discrete structure $S := (\mathcal{T}, G)$, where $G = \{G^{(\ell)}\}_{\ell=1}^{L}$. Let $Y := (Z, U, \Theta)$ denote all continuous variables. We approximate the conditional $p(Y \mid X, S)$ by a tractable family $q(Y)$ and maximize the evidence lower bound (ELBO)

$$
\mathcal{L}(q; S) = \mathbb{E}_q[\log p(X, Y \mid S)] - \mathbb{E}_q[\log q(Y)]\,, \tag{32}
$$

where $p(X, Y \mid S)$ is the joint density of observations and continuous variables induced by the full model in Section 5.5 after conditioning on fixed discrete structure $S = (\mathcal{T}, G)$; equivalently, $\log p(X, Y \mid S) = \log p(X, Y, S) - \log p(S)$, and $\log p(S)$ is handled separately in (35).

Variational family. We use a mean-field family

$$
\begin{aligned}
q(Y) = &\left(\prod_{\ell=1}^{L}\prod_{m=1}^{M_\ell} q\big(U_m^{(\ell)}\big)\right)\left(\prod_{\ell=2}^{L}\prod_{m=1}^{M_\ell} q\big(Z_m^{(\ell)}\big)\right)\\
&\times\left(\prod_{\ell=1}^{L}\prod_{(j\to m)\in E^{(\ell)}} q\big(A_{j\to m}^{(\ell)}\big)\right)\left(\prod_{\ell=2}^{L}\prod_{m=1}^{M_\ell} q\big(B_{\mathrm{pa}(m,\ell)\to m}^{(\ell)}\big)\right)\\
&\times\left(\prod_{i=1}^{p} q(w_i)\,q(\psi_i)\right)\left(\prod_{\ell=1}^{L} q(\sigma_{\eta,\ell}^2)\,q(\sigma_{\xi,\ell}^2)\right).
\end{aligned}
\tag{33}
$$

The top-level states $Z_m^{(1)}$ are not assigned variational factors because they are deterministically identified with $U_m^{(1)}$ by the model definition. We take $q(U_m^{(\ell)})$ and $q(Z_m^{(\ell)})$ to be matrix-normal, $q(A)$, $q(B)$, and $q(w_i)$ to be Gaussian, and $q(\psi_i)$, $q(\sigma_{\eta,\ell}^2)$, $q(\sigma_{\xi,\ell}^2)$ to be inverse-gamma.

Blocked coordinate ascent for continuous variables. For fixed S, coordinate ascent updates satisfy the standard mean-field optimality condition

$$
\log q^\star(y) = \mathbb{E}_{q(Y\setminus y)}[\log p(X, Y \mid S)] + \text{const.} \tag{34}
$$

Because all continuous conditionals in Section 5.5 are conjugate Gaussian or inverse-gamma, these updates reduce to closed-form Bayesian multivariate linear regression updates. In implementation we perform a small number of sweeps per outer iteration and exploit locality so that after a discrete structure move only factors in the affected Markov blanket are refreshed.

A variational score for discrete structure. Define the structure score

$$
\mathcal{F}(S; q) = \mathcal{L}(q; S) + \log p(\mathcal{T}) + \sum_{\ell=1}^{L} \log \widetilde{p}\big(G^{(\ell)} \mid \mathcal{T}\big), \tag{35}
$$

which is a lower bound on $\log p(X, S) = \log p(S) + \log p(X \mid S)$ because $\mathcal{L}(q; S) \le \log p(X \mid S)$. We restrict scoring and proposals to admissible structures: $\mathcal{T}$ must satisfy (14) and each $G^{(\ell)}$ must lie in $\mathcal{G}_\ell$. Equivalently, $\mathcal{F}(S; q) = -\infty$ for inadmissible S, matching the indicator factors in (30). Let

$$
\mathcal{F}^\star(S) := \max_q \mathcal{F}(S; q),
$$

and write $\widehat{q}_S$ for the approximate maximizer obtained by running a bounded number of variational sweeps. MT-BN uses $\mathcal{F}(S; \widehat{q}_S)$ as a scalable surrogate for comparing structures. Here $\widetilde{p}(G^{(\ell)} \mid \mathcal{T})$ denotes the unnormalized graph potential in (17), consistent with the joint discrete prior in (16).

Metropolis-style stochastic search over structure. We explore the discrete space $S = (\mathcal{T}, G)$ using local proposals with proposal kernel $Q(S \to S')$. Given a current state $(S, \widehat{q}_S)$, we propose $S' \sim Q(S \to \cdot)$, run a short local variational refresh to obtain $\widehat{q}_{S'}$, and accept S' with probability

$$
\alpha = \min\left\{1,\ \exp\Big(\mathcal{F}(S'; \widehat{q}_{S'}) - \mathcal{F}(S; \widehat{q}_S)\Big) \cdot \frac{Q(S' \to S)}{Q(S \to S')}\right\}. \tag{36}
$$

Equation (36) defines a Markov chain targeting the surrogate distribution proportional to $\exp(\mathcal{F}^\star(S))$ in the idealized limit where $\widehat{q}_S$ achieves the exact maximizer and scores are computed exactly. In practice, because $\widehat{q}_S$ is obtained by truncated variational optimization and $\mathcal{F}$ uses an ELBO rather than the exact marginal likelihood, the chain is an approximate exploration procedure. We therefore use it primarily for (i) locating high-scoring structures (MAP estimation under $\mathcal{F}$) and (ii) producing uncertainty summaries with respect to this variationally defined surrogate.

Within-level DAG updates

For each level ℓ, we update $G^{(\ell)}$ using single-edge add/delete/reverse proposals subject to acyclicity and the in-degree cap $d_{\max}$. Each proposal changes only one node's parent set in the within-level SEM, hence it modifies only the corresponding conditional factors in the innovation model and the degree-dependent components of the graph prior. We exploit this locality by refreshing only the affected variational factors in $\widehat{q}_{S'}$ (the Markov blanket of the modified node in level ℓ) before evaluating $\mathcal{F}(S'; \widehat{q}_{S'})$.

Hierarchy updates

We update $\mathcal{T}$ through local reassignment moves on nCRP paths. For an observed variable i, we propose modifying one level assignment $z_i^{(\ell)}$ at a time conditional on its parent assignment $z_i^{(\ell-1)}$, preserving nestedness and enforcing the minimum-size constraint. The proposal distribution is proportional to the nCRP predictive probabilities restricted to admissible moves, and the score change is evaluated by recomputing only the likelihood terms and proximity-bin counts impacted along the affected ancestor chain.

Overall schedule and outputs

MT-BN alternates between batches of within-level graph proposals across $\ell = 1, \ldots, L$, batches of hierarchy proposals, and a bounded number of global variational refinement sweeps. The primary point estimate reported by the method is the highest-scoring structure encountered,

$$\widehat{S} = \arg \max_{S \in \mathcal{S}_{\text{visited}}} \mathcal{F}(S; \widehat{q}_S),$$

together with its corresponding variational approximation $\widehat{q}_{\widehat{S}}$. For uncertainty summaries, we retain a sequence of visited structures and report edge and co-assignment frequencies with respect to the surrogate exploration distribution induced by (36), as formalized in Section 5.8.

5.7 Implementation Complexity and Local Update Costs

This subsection analyzes the computational cost of the blocked inference procedure in Section 5.6 and specifies implementation choices that make MT-BN practical at large p. Throughout, n denotes the number of samples, p the number of observed variables, L the hierarchy depth, M_ℓ the number of modules at level ℓ, r_ℓ the latent dimension at level ℓ, and $d_{\max}$ the maximum in-degree enforced for each within-level DAG $G^{(\ell)}$. We write $\bar{s}_\ell \approx p/M_\ell$ for the average number of observed variables associated with a level-ℓ module. Complexity is stated for one outer iteration of the procedure in Section 5.6, which alternates (i) discrete proposals in $(\mathcal{T}, G)$ and (ii) bounded variational refinement to evaluate the variational structure score $\mathcal{F}(S; \widehat{q}_S)$.

Unit of computation and update schedule. One outer iteration consists of a batch of within-level graph proposals, a batch of hierarchy proposals, and a bounded number of variational coordinate-ascent sweeps. Concretely, we perform S_G single-edge proposals for each level-ℓ graph $G^{(\ell)}$, S_T local hierarchy proposals (single-variable, single-level reassignment moves) distributed over variables, and S_q variational sweeps. Importantly, during the discrete proposal phases we do not rerun full variational optimization from scratch; instead we perform a localized refresh to obtain $\widehat{q}_{S'}$ sufficient to stably evaluate score differences in (36). Full sweeps are performed only in the variational refinement block.

Locality of structure moves and what is recomputed. Both graph and hierarchy proposals are implemented so that score updates and subsequent variational refreshes involve only localized recomputation. A single edge move in $G^{(\ell)}$ changes only one node's parent set in the within-level SEM (23) and only the corresponding terms in the graph priors (sparsity/proximity counts and degrees). A single hierarchy move for one observed variable changes only a constant-size neighborhood in $\mathcal{T}$ (the affected child/parent modules along the variable's ancestor chain), and only measurement terms for the variable together with vertical-inheritance factors for affected modules. In both cases, we refresh only the variational factors in the Markov blanket of the modified components before re-evaluating $\mathcal{F}(S; \widehat{q}_S)$.

Acyclicity enforcement and admissible edge moves

Each $G^{(\ell)}$ is maintained as a DAG under add/delete/reverse proposals subject to the in-degree cap $d_{\max}$. Because M_ℓ is typically far smaller than p, exact cycle checks are feasible. We use a reachability-based check for edge additions and reversals. For each node $a \in V^{(\ell)}$, maintain a bitset $\text{reach}_\ell(a) \subseteq V^{(\ell)}$ representing nodes reachable from a. A proposed edge addition $u \to v$ is admissible if and only if $u \neq v$ and $u \notin \text{reach}_\ell(v)$. This check is $O(1)$ in bitset lookup time. If admissible, we update reachability using bitset unions. In the worst case, an incremental closure update can cost $O(M_\ell^2/w)$ bit operations per accepted addition, where w is the machine-word size, although in the sparse regime it is typically much smaller.

Edge deletions (and reversals implemented as delete-plus-add) are more delicate because reachability can decrease. We therefore handle deletions by postponing exact reachability maintenance and periodically recomputing $\{\text{reach}_\ell(a)\}_a$ from the current adjacency structure after every R accepted proposals at level ℓ. A full recomputation

using bitset-based transitive closure costs $O(M_\ell^3/w)$, which is negligible for the module-graph sizes encountered in typical MT-BN settings. Proposals that violate $\max_m d_{\text{in}}^{(\ell)}(m) \le d_{\max}$ are rejected immediately.

Caching variational sufficient statistics for fast local scoring

Score evaluation under $\mathcal{F}(S; \widehat{q}_S)$ requires ELBO terms $\mathcal{L}(\widehat{q}_S; S)$ and log-prior terms. Graph-prior differences are computable in constant time per proposal because the sparsity and proximity priors depend only on counts and the hub prior depends only on degrees. The dominant remaining work is the local variational refresh needed to obtain $\widehat{q}_{S'}$ after a proposal.

The local refreshes reduce to Bayesian multivariate linear regressions induced by (23), (25), and (27). To compute these updates efficiently, we cache expectations under the current variational factors. For a matrix-normal factor $q(U_m^{(\ell)}) = \mathcal{MN}(\mu_{U,m}^{(\ell)}, I_n, \Sigma_{U,m}^{(\ell)})$, the required second moment satisfies

$$\mathbb{E}_q\left[U_m^{(\ell)\top} U_m^{(\ell)}\right] = \mu_{U,m}^{(\ell)\top} \mu_{U,m}^{(\ell)} + n\, \Sigma_{U,m}^{(\ell)}. \tag{37}$$

For distinct modules $j \neq m$, mean-field independence implies

$$\mathbb{E}_q\left[U_j^{(\ell)\top} U_m^{(\ell)}\right] = \mu_{U,j}^{(\ell)\top} \mu_{U,m}^{(\ell)}. \tag{38}$$

Analogous cached moments are maintained for each $q(Z_m^{(\ell)})$. These cached moments are sufficient to form the expected Gram matrices and cross-products appearing in coordinate updates without repeatedly scanning all n samples.

Cost of local variational refreshes

We summarize the dominant computational cost of the local regressions induced by each model component.

For the within-level SEM (23), updating the factors associated with a single child module m at level ℓ involves regressions of $U_m^{(\ell)} \in \mathbb{R}^{n \times r_\ell}$ on its parents $\{U_j^{(\ell)}\}_{j \in \mathrm{Pa}(m)}$. With $|\mathrm{Pa}(m)| \le d_{\max}$, the effective predictor dimension is $d_{\max} r_\ell$. Using cached moments (37)–(38), the dominant per-refresh cost is solving a linear system of size $d_{\max} r_\ell$, together with forming the corresponding expected cross-products:

$$O\big(n(d_{\max} r_\ell)^2 + (d_{\max} r_\ell)^3\big), \tag{39}$$

where the $n(d_{\max} r_\ell)^2$ term reflects forming cross-products when cached moments are not yet available or must be updated, and the cubic term reflects solving the small system. In practice, most proposals update cached moments incrementally, so the per-proposal constant factors are small.

For vertical inheritance (25), updating a module factor $q(Z_m^{(\ell)})$ and its associated $q(B_{\text{pa}\to m}^{(\ell)})$ is a regression of $Z_m^{(\ell)}$ on $(Z_{\text{pa}(m,\ell)}^{(\ell-1)}, U_m^{(\ell)})$, with predictor dimension $(r_{\ell-1} + r_\ell)$. The corresponding local refresh cost is

$$O\big(n(r_{\ell-1} + r_\ell)^2 + (r_{\ell-1} + r_\ell)^3\big). \tag{40}$$

The top level uses the deterministic identification $Z_m^{(1)} := U_m^{(1)}$ and therefore does not introduce an additional regression factor.

For the measurement model (27), each observed variable $X_{:,i}$ updates $(q(w_i), q(\psi_i))$ via regression on the assigned bottom-level state $Z_{m(i)}^{(L)}$. Given cached module moments $\mathbb{E}_q[Z_m^{(L)\top} Z_m^{(L)}]$, the per-variable cost is

$$O\big(n r_L^2 + r_L^3\big), \tag{41}$$

and computing the module-level moment $\mathbb{E}_q[Z_m^{(L)\top} Z_m^{(L)}]$ costs $O(n r_L^2)$ per module. Since many variables share the same bottom module, we compute module moments once per module and reuse them across the $\bar{s}_L$ variables in that module.

Per-iteration complexity

We now state the dominant costs per outer iteration under bounded $d_{\max}$ and fixed latent dimensions $\{r_\ell\}$. The total cost decomposes into (i) discrete-structure proposals with local variational refresh and (ii) global variational refinement sweeps.

For graph proposals, we perform S_G proposals per level. Each proposal requires an acyclicity check and a local refresh affecting only a constant number of module factors at that level (typically the modified child module and its incident parameter factors). Thus the graph-proposal cost per iteration satisfies

$$O\left(\sum_{\ell=1}^{L} S_G\Big(\mathrm{acyc}_\ell + n(d_{\max} r_\ell)^2 + (d_{\max} r_\ell)^3\Big)\right), \tag{42}$$

where acyc_ℓ denotes the amortized cost of maintaining the acyclicity data structure at level ℓ, bounded by $O(M_\ell^2/w)$ per accepted addition in the worst case and $O(M_\ell^3/wR)$ per proposal for periodic recomputation of reachability.

For hierarchy proposals, we perform S_T local reassignment moves. Each move affects a constant number of modules along the ancestor chain and triggers a constant number of local vertical-inheritance refreshes (40) plus measurement updates (41) for the moved variable and any newly created or removed bottom-module bookkeeping. A conservative bound is

$$O\left(S_T\Big(\sum_{\ell\in\mathcal{A}(i)} \big(n(r_{\ell-1}+r_\ell)^2 + (r_{\ell-1}+r_\ell)^3\big) + nr_L^2 + r_L^3\Big)\right), \tag{43}$$

where $\mathcal{A}(i)$ denotes the set of affected levels on the ancestor chain for the moved variable and is bounded by L. In typical usage, only a small number of levels are modified per proposal, so the effective constant is modest.

For variational refinement, one global sweep updates all SEM regressions across all levels and modules, all vertical-inheritance regressions across all non-root modules, and all measurement regressions. Under bounded $d_{\max}$, the cost of one sweep is

$$O\left(\sum_{\ell=1}^{L} M_\ell\Big(n(d_{\max} r_\ell)^2 + (d_{\max} r_\ell)^3\Big) + \sum_{\ell=2}^{L} M_\ell\Big(n(r_{\ell-1}+r_\ell)^2 + (r_{\ell-1}+r_\ell)^3\Big) + M_L n r_L^2 + p r_L^2\right), \tag{44}$$

and S_q sweeps multiply (44) by S_q. When $d_{\max}$ and r_ℓ are treated as small constants, (44) is approximately linear in $\sum_\ell M_\ell$ and in p, with dependence on n entering through cross-product formation.

Memory footprint. The dominant stored quantities are the variational means $\mu_{U,m}^{(\ell)} \in \mathbb{R}^{n\times r_\ell}$, $\mu_{Z,m}^{(\ell)} \in \mathbb{R}^{n\times r_\ell}$ (for $\ell \geq 2$), and the column-covariances $\Sigma_{U,m}^{(\ell)} \in \mathbb{R}^{r_\ell\times r_\ell}$, $\Sigma_{Z,m}^{(\ell)} \in \mathbb{R}^{r_\ell\times r_\ell}$. This yields memory $O(n\sum_\ell M_\ell r_\ell)$ for means plus $O\big(\sum_\ell M_\ell r_\ell^2\big)$ for covariances, with the understanding that the top level uses $Z^{(1)} := U^{(1)}$.

Parallelization. Conditional on current parent sets, SEM updates for $\{U_m^{(\ell)}\}_{m=1}^{M_\ell}$ are independent across m within each level and can be parallelized. Vertical-inheritance updates for $\{Z_m^{(\ell)}\}$ are independent across modules given the parent-level moments and are parallelizable within each level after a top-down pass to compute parent expectations. Measurement updates for $\{w_i, \psi_i\}$ are independent across observed variables given module moments and can be parallelized across i. Graph proposals are parallelizable across levels because the graphs are updated independently conditional on $\mathcal{T}$.

Practical defaults and stopping criteria. We initialize $\mathcal{T}$ from the nCRP prior and initialize each $G^{(\ell)}$ as the empty DAG. Continuous variables are warm-started by running a short variational phase with empty graphs and then alternating discrete proposals with local refreshes. We monitor $\mathcal{F}(S;\widehat{q}_S)$ across iterations, along with acceptance rates of discrete proposals. Because the discrete exploration uses the surrogate accept rule (36), convergence is interpreted operationally as stability of the best-scoring structure and stability of edge and co-assignment summaries under the induced exploration distribution.

5.8 Posterior Summaries and Outputs

MT-BN returns (i) a hierarchy $\mathcal{T}$ of nested modules, (ii) within-level DAGs $\{G^{(\ell)}\}_{\ell=1}^{L}$ defined on innovation latents, and (iii) variational summaries of continuous latents and parameters. Because discrete structure is explored using Metropolis–Hastings moves scored by the variational structure objective $\mathcal{F}(S;\widehat{q}_S)$ (Section 5.6), the retained iterates of $(\mathcal{T}, G)$ are not exact posterior samples from $p(\mathcal{T}, G \mid X)$. Accordingly, we report uncertainty as algorithm-induced stability metrics (frequencies under the retained iterates) and as variational uncertainty for continuous quantities conditional on a fixed structure. This subsection defines the concrete outputs reported in experiments.

Retained discrete iterates and point summaries. Let $\{(\mathcal{T}^{(t)}, G^{(t)})\}_{t=1}^{T}$ denote the retained discrete iterates after burn-in and thinning, where $G^{(t)} = \{G^{(\ell,t)}\}_{\ell=1}^{L}$. We emphasize that these iterates are generated by the acceptance rule (36) and therefore reflect exploration of high-scoring structures under $\mathcal{F}$, not draws from the exact posterior. We use them in two ways.

First, we form a best-scoring point estimate

$$(\widehat{\mathcal{T}}, \widehat{G}) = \arg \max_{t \in \{1,\dots,T\}} \mathcal{F}\big(S^{(t)}; \widehat{q}_{S^{(t)}}\big), \tag{45}$$

where $S^{(t)} = (\mathcal{T}^{(t)}, G^{(t)})$. Second, when multiple structures have similar scores, we report stability summaries that quantify how frequently features recur across the retained iterates. These summaries can be interpreted as heuristic confidence measures and are used only as such.

Edge-stability scores across levels. Fix a level ℓ and a reference hierarchy $\widehat{\mathcal{T}}$. Let $V^{(\ell)}(\widehat{\mathcal{T}}) = \{1, \dots, M_\ell(\widehat{\mathcal{T}})\}$ denote its level-ℓ modules. Because module labels and even module identities can vary across iterates when $\mathcal{T}$ changes, edge-stability must be defined with a mapping from each sampled hierarchy $\mathcal{T}^{(t)}$ to the reference $\widehat{\mathcal{T}}$.

We define a module matching map $\phi_t^{(\ell)} : V^{(\ell)}(\mathcal{T}^{(t)}) \to V^{(\ell)}(\widehat{\mathcal{T}})$ by maximum overlap:

$$\phi_t^{(\ell)}(u) = \arg \max_{v \in V^{(\ell)}(\widehat{\mathcal{T}})} \Big| C_u^{(\ell)}(\mathcal{T}^{(t)}) \cap C_v^{(\ell)}(\widehat{\mathcal{T}}) \Big|, \tag{46}$$

breaking ties arbitrarily. Using $\phi_t^{(\ell)}$, we map each iterate-level DAG $G^{(\ell,t)}$ onto the reference node set by pushing edges forward. For any ordered pair $(a, b) \in V^{(\ell)}(\widehat{\mathcal{T}}) \times V^{(\ell)}(\widehat{\mathcal{T}})$, define the edge-stability score

$$\widehat{\pi}_{a\to b}^{(\ell)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\Big\{ \exists\, (u \to v) \in E\big(G^{(\ell,t)}\big) \text{ s.t. } \phi_t^{(\ell)}(u) = a,\ \phi_t^{(\ell)}(v) = b \Big\}. \tag{47}$$

We interpret $\widehat{\pi}_{a\to b}^{(\ell)}$ as the fraction of retained iterates in which an edge consistent with $a \to b$ appears after mapping to $\widehat{\mathcal{T}}$. When $\mathcal{T}$ is fixed (e.g., if one runs MT-BN with a fixed hierarchy), $\phi_t^{(\ell)}$ is the identity and (47) reduces to a simple frequency.

Consensus graph construction. To report a sparse directed summary graph at each level ℓ, we provide two selection rules based on $\widehat{\pi}^{(\ell)}$.

(i) Thresholded consensus graph. For $\tau \in (0, 1)$,

$$\widehat{G}^{(\ell)}(\tau) = \Big(V^{(\ell)}(\widehat{\mathcal{T}}),\ \{a \to b : \widehat{\pi}_{a\to b}^{(\ell)} \geq \tau\} \Big). \tag{48}$$

(ii) Top-K stability graph. Choose K_ℓ and include the K_ℓ edges with largest $\widehat{\pi}_{a\to b}^{(\ell)}$, breaking ties arbitrarily. This avoids tuning a probability threshold when the scale of $\widehat{\pi}$ depends on mixing of the discrete exploration.

In all cases, these graphs are summaries of the retained iterates and do not constitute calibrated posterior credible sets.

Co-assignment stability for hierarchy uncertainty. Uncertainty in the hierarchy is summarized through co-assignment stability of observed variables. For any pair $i, j \in \{1, \dots, p\}$, define the level-ℓ co-assignment score

$$\widehat{\rho}_{ij}^{(\ell)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\Big\{ z_i^{(\ell,t)} = z_j^{(\ell,t)} \Big\}. \tag{49}$$

The matrix $\widehat{\rho}^{(\ell)}$ is a similarity kernel that can be used to form a consensus partition at level ℓ (e.g., by spectral clustering on $\widehat{\rho}^{(\ell)}$). This yields a stable hierarchy summary even when module labels permute across retained iterates.

Variational summaries of latent module states and innovations. For any fixed structure $S = (\mathcal{T}, G)$, variational inference returns $\widehat{q}_S$ over continuous latents and parameters. We report posterior-mean summaries under $\widehat{q}_S$:

$$\widehat{U}_m^{(\ell)} = \mathbb{E}_{\widehat{q}_S}\Big[U_m^{(\ell)}\Big], \qquad \widehat{Z}_m^{(\ell)} = \mathbb{E}_{\widehat{q}_S}\Big[Z_m^{(\ell)}\Big],$$

together with variational covariances (e.g., the column-covariances in the matrix-normal factors). When reporting these quantities alongside the best-scoring discrete estimate $(\widehat{\mathcal{T}}, \widehat{G})$, we take $S = \widehat{S} := (\widehat{\mathcal{T}}, \widehat{G})$.

The innovation decomposition (25) provides a direct measure of resolution-specific variation. For a module m at level ℓ, we report the fraction of state energy attributable to innovation,

$$\kappa_m^{(\ell)} = \frac{\|\widehat{U}_m^{(\ell)}\|_F^2}{\|\widehat{Z}_m^{(\ell)}\|_F^2 + \epsilon_0}, \tag{50}$$

where $\epsilon_0 > 0$ is a fixed numerical constant. Large $\kappa_m^{(\ell)}$ indicates that the module exhibits substantial level-ℓ behavior not explained by inherited signal.

Directed effect magnitudes on innovation latents. Beyond edge existence, MT-BN reports effect magnitudes derived from SEM maps $\{A_{j\to m}^{(\ell)}\}$. For a directed edge $j \to m$ at level ℓ, define the variational strength summary

$$s_{j\to m}^{(\ell)} = \mathbb{E}_{\widehat{q}_S}\left[\|A_{j\to m}^{(\ell)}\|_F\right], \tag{51}$$

and a confidence-weighted score that combines effect magnitude with edge stability under the retained iterates,

$$\widetilde{s}_{j\to m}^{(\ell)} = \widehat{\pi}_{j\to m}^{(\ell)}\, s_{j\to m}^{(\ell)}. \tag{52}$$

When $r_\ell = 1$, these reduce to absolute scalar coefficients and their variational expectations. Because latent coordinates are not identifiable up to general invertible transforms, we interpret (51) as a relative, model-internal magnitude and emphasize invariants such as edge direction and presence in $\widehat{G}^{(\ell)}(\tau)$.

Node-level hub and centrality summaries. For each level ℓ and reference node $a \in V^{(\ell)}(\widehat{\mathcal{T}})$, we report stability-based expected degrees

$$\widehat{d}^{(\ell),\mathrm{out}}(a) = \sum_{b\neq a} \widehat{\pi}_{a\to b}^{(\ell)}, \qquad \widehat{d}^{(\ell),\mathrm{in}}(a) = \sum_{b\neq a} \widehat{\pi}_{b\to a}^{(\ell)},$$

and rank modules accordingly. At the bottom level, these rankings correspond to candidate regulators or drivers (biological or financial) whose influence is repeatedly supported by high-scoring discrete structures and whose effect magnitudes are large under $\widehat{q}_{\widehat{S}}$.

These summaries provide a standardized set of reported outputs: a best-scoring hierarchy and multi-scale directed graphs, stability scores for edges and co-assignments across retained iterates, and variational summaries of latent states and innovations conditional on a fixed structure. Subsequent experiments use these outputs to evaluate accuracy, interpretability, and scalability across domains.

6 Search-Space Reduction and Computational Guarantees

The primary computational bottleneck in Bayesian network structure learning is the combinatorial explosion of candidate directed edges and admissible DAGs as the node count grows. MT-BN mitigates this bottleneck by learning directed structure on module-level innovation variables across multiple resolutions rather than on the full set of p observed variables. This section formalizes (i) reductions in the discrete search space induced by hierarchical modularization and (ii) the locality properties that make score updates and acceptance ratios computable via small, localized refreshes. We also formalize a statistical consequence of the innovation-based design: shared inherited signal can induce strong within-level correlations among full module states, but MT-BN removes this inheritance before learning within-level directed structure, preventing a systematic source of spurious edges.

6.1 Reduction in candidate edges across resolutions

A flat p-node DAG has $p(p-1)$ possible directed edges (excluding self-loops). In MT-BN, directed edges are learned only among modules at each level ℓ, so the relevant edge universe is $\sum_{\ell=1}^{L} M_\ell(M_\ell - 1)$, where M_ℓ is the number of modules at level ℓ.

Proof. (Module-count bound under minimum size) Assume the minimum module size constraint (14) holds at every instantiated module and every level. Then for every $\ell \in \{1, \ldots, L\}$,

$$M_\ell \leq \left\lfloor \frac{p}{m_{\min}} \right\rfloor.$$

□

Proof. At level ℓ, the sets $\{C_m^{(\ell)}\}_{m=1}^{M_\ell}$ form a partition of $\{1,\dots,p\}$ and each has size at least $m_{\min}$. Therefore $p = \sum_{m=1}^{M_\ell} |C_m^{(\ell)}| \geq M_\ell m_{\min}$, which implies $M_\ell \leq p/m_{\min}$. Taking floors yields the claim. □

Proof. (Edge-universe reduction) Under the minimum module size constraint (14), the total number of possible within-level directed edges across all resolutions satisfies

$$\sum_{\ell=1}^{L} M_\ell(M_\ell - 1) \leq L \left\lfloor \frac{p}{m_{\min}} \right\rfloor \left(\left\lfloor \frac{p}{m_{\min}} \right\rfloor - 1 \right).$$

Consequently, for fixed L, the MT-BN within-level edge universe is smaller than the flat p-node edge universe by a factor on the order of $m_{\min}^2$ in the leading p^2 term. ■

Proof. Apply Lemma 6.1 to each M_ℓ and bound termwise. The asymptotic statement follows by comparing the leading p^2 terms: $p(p-1)$ versus $L(p/m_{\min})(p/m_{\min} - 1)$. □

Remark. The reduction in Theorem 6.1 is purely combinatorial and holds before introducing additional constraints such as the in-degree cap $d_{\max}$, which further reduces the set of admissible graphs at each level.

6.2 Reduction in the number of admissible DAG structures

Beyond edge counts, the discrete search space is governed by the number of admissible DAGs. Let $\mathsf{DAG}(d)$ denote the set of labeled DAGs on d nodes. A flat BN structure learner searches over $\mathsf{DAG}(p)$, whereas MT-BN searches over the product space $\prod_{\ell=1}^{L} \mathsf{DAG}(M_\ell)$ (subject to constraints such as in-degree caps and hierarchy-dependent potentials).
Proof. (A counting upper bound for DAGs) For any $d \geq 1$,

$$|\mathsf{DAG}(d)| \leq d!\, 2^{\binom{d}{2}}.$$

□

Proof. Fix a topological ordering σ of the d nodes. Relative to σ, edges may only point forward in the order, so there are at most $2^{\binom{d}{2}}$ edge subsets consistent with σ. There are $d!$ possible orderings, so counting all forward-edge subsets across all orderings yields $d!\, 2^{\binom{d}{2}}$ graphs. This overcounts because a given DAG may admit multiple topological orderings, hence it is an upper bound. □

Proof. (DAG search-space reduction under modularization) Under the minimum module size constraint (14), the number of possible multi-resolution within-level DAG configurations satisfies

$$\prod_{\ell=1}^{L} |\mathsf{DAG}(M_\ell)| \leq \left(\left\lfloor \frac{p}{m_{\min}} \right\rfloor ! \right)^L 2^{L\binom{\lfloor p/m_{\min} \rfloor}{2}}.$$

In particular, the logarithm of the MT-BN DAG configuration count scales at most as

$$O\left(L \left(\frac{p}{m_{\min}} \right)^2 \right)$$

in the leading term, whereas a flat p-node DAG search has a corresponding upper bound scaling as $O(p^2)$ in the leading term. ■

Proof. Apply Lemma 6.2 at each level with $d = M_\ell$, multiply across ℓ, and then apply Lemma 6.1 to bound each M_ℓ by $\lfloor p/m_{\min} \rfloor$. Taking logarithms yields the stated scaling, since $\log(d!) = O(d \log d)$ is lower order than $\binom{d}{2}$ in the leading term. □

Remark. The bound in Theorem 6.2 is conservative. The effective search space is smaller in practice because MT-BN further restricts graphs to $\mathcal{G}_\ell$ via acyclicity, optional in-degree caps, and topology-aware priors (Section 5.4), and because discrete proposals explore only local neighborhoods rather than enumerating the full space.

6.3 Innovation-based structure removes inherited correlation as a driver of within-level edges

The preceding results are computational. MT-BN's innovation design also has a direct structural consequence: it prevents inherited shared signal from being mistaken for within-level directed dependence. The following proposition formalizes the mechanism in the Gaussian case.

Proposition 6.1 (Siblings share inherited dependence through parents, not innovations)**.** *Fix a level $\ell \geq 2$ and let m and m' be two distinct level-ℓ modules with the same parent $a = \mathrm{pa}(m, \ell) = \mathrm{pa}(m', \ell)$. Suppose the vertical inheritance model holds with zero inheritance noise for simplicity,*

$$Z_m^{(\ell)} = Z_a^{(\ell-1)} B_{a\to m}^{(\ell)} + U_m^{(\ell)}, \qquad Z_{m'}^{(\ell)} = Z_a^{(\ell-1)} B_{a\to m'}^{(\ell)} + U_{m'}^{(\ell)},$$

and suppose that conditional on the coarser-scale variables (in particular $Z_a^{(\ell-1)}$), the innovations $U_m^{(\ell)}$ and $U_{m'}^{(\ell)}$ are independent and mean-zero:

$$U_m^{(\ell)} \perp U_{m'}^{(\ell)} \mid Z_a^{(\ell-1)}, \qquad \mathbb{E}[U_m^{(\ell)} \mid Z_a^{(\ell-1)}] = 0, \qquad \mathbb{E}[U_{m'}^{(\ell)} \mid Z_a^{(\ell-1)}] = 0.$$

Then $Z_m^{(\ell)}$ and $Z_{m'}^{(\ell)}$ may be strongly dependent marginally, but their residual dependence after conditioning on the parent is entirely captured by the innovations, in the sense that

$$\mathrm{Cov}\Big(Z_m^{(\ell)}, Z_{m'}^{(\ell)} \mid Z_a^{(\ell-1)}\Big) = 0, \qquad \mathrm{Cov}\Big(Z_m^{(\ell)}, Z_{m'}^{(\ell)}\Big) = B_{a\to m}^{(\ell)\top} \mathrm{Cov}\Big(Z_a^{(\ell-1)}\Big)\, B_{a\to m'}^{(\ell)}.$$

Proof. Condition on $Z_a^{(\ell-1)}$. By the assumed inheritance equations, $Z_m^{(\ell)} - Z_a^{(\ell-1)} B_{a\to m}^{(\ell)} = U_m^{(\ell)}$ and similarly for m'. The conditional covariance is therefore

$$\mathrm{Cov}\Big(Z_m^{(\ell)}, Z_{m'}^{(\ell)} \mid Z_a^{(\ell-1)}\Big) = \mathrm{Cov}\Big(U_m^{(\ell)}, U_{m'}^{(\ell)} \mid Z_a^{(\ell-1)}\Big) = 0$$

by conditional independence. For the marginal covariance, expand using bilinearity and the mean-zero assumptions:

$$\mathrm{Cov}(Z_m^{(\ell)}, Z_{m'}^{(\ell)}) = \mathrm{Cov}\Big(Z_a^{(\ell-1)} B_{a\to m}^{(\ell)}, Z_a^{(\ell-1)} B_{a\to m'}^{(\ell)}\Big) + \mathrm{Cov}(U_m^{(\ell)}, U_{m'}^{(\ell)}),$$

and $\mathrm{Cov}(U_m^{(\ell)}, U_{m'}^{(\ell)}) = 0$ under the unconditional implication of the conditional independence with mean-zero residuals. The remaining term yields the stated expression. □

Remark. Proposition 6.1 formalizes a systematic failure mode of learning within-level edges on full module states $Z^{(\ell)}$: sibling states can appear strongly dependent even when there is no within-level interaction, purely because they inherit from the same parent. MT-BN's modeling choice to define within-level DAGs on innovations $U^{(\ell)}$ targets directed structure after removing inherited shared signal, which both improves interpretability and reduces pressure toward dense within-level graphs.

6.4 Local score updates and computational gains from decomposability

MT-BN evaluates discrete proposals using the variational structure score $\mathcal{F}(S; \widehat{q}_S)$ in (35). The feasibility of exploring $(\mathcal{T}, G)$ hinges on the fact that single-edge and local reassignment moves change only localized terms.

Proof. (Constant-time updates for topology-aware priors) Fix a level ℓ. Consider a single-edge add/delete/reverse proposal that modifies $E^{(\ell)}$ by changing at most one directed edge. Then the log prior potential contribution

$$\log \widetilde{p}\big(G^{(\ell)} \mid \mathcal{T}\big) = \log p_{\mathrm{sp}}\big(G^{(\ell)}\big) + \log p_{\mathrm{prox}}\big(G^{(\ell)} \mid \mathcal{T}\big) + \log p_{\mathrm{hub}}\big(G^{(\ell)}\big)$$

can be updated in $O(1)$ arithmetic time given cached sufficient statistics consisting of $|E^{(\ell)}|$, the proximity-bin counts $\{E_{\ell,b}\}_{b=1}^{3}$, and the out-degrees $\{d_{\mathrm{out}}^{(\ell)}(u)\}_{u=1}^{M_\ell}$. □

Proof. A single-edge modification changes $|E^{(\ell)}|$ by at most one, and therefore changes the Beta function arguments in (18) by constant increments. It changes exactly one proximity bin count $E_{\ell,b}$ (determined by $\kappa_\ell(u, v)$ for the modified edge $u \to v$) by at most one, and therefore changes exactly one factor in (20) by constant increments. Finally, it changes the out-degree of exactly one node (the tail of the modified edge) by ± 1, so only one factor in (21) changes. All updates therefore require a constant number of arithmetic operations when the sufficient statistics are cached. □

Proof. (Per-proposal cost under bounded in-degree) Fix a level ℓ and assume $\max_m d_{\text{in}}^{(\ell)}(m) \le d_{\max}$. Consider a single-edge proposal in $G^{(\ell)}$ together with the local variational refresh described in Section 5.6. If latent dimension r_ℓ is fixed, then the dominant cost of updating the affected innovation-regression block is

$$O\big(n(d_{\max} r_\ell)^2 + (d_{\max} r_\ell)^3\big)\,,$$

plus an amortized acyclicity-maintenance cost that depends on the chosen data structure. The prior contribution updates in $O(1)$ time by Lemma 6.4. ■

Proof. A single-edge change modifies only one node's parent set in the within-level SEM (23). Under mean-field inference, the required local refresh for that node reduces to a Bayesian multivariate regression with predictor dimension at most $d_{\max} r_\ell$. Forming the local normal equations and solving the corresponding linear system yields the stated bound. The prior update claim follows from Lemma 6.4. □

Remark. Theorem 6.4 should be contrasted with flat BN structure learning in which a single-edge move can require recomputing or updating statistics involving p variables and high-order conditional dependencies, often leading to poor scaling when $p \gg n$. In MT-BN, the discrete moves occur on module graphs with $M_\ell \ll p$, and the local refresh costs depend on $d_{\max}$ and latent dimensions rather than directly on p.

7 Case Studies and Empirical Evaluation

This section evaluates MT-BN on (i) a benchmark setting with known ground truth for edges, and (ii) a large real-world transcriptomic dataset without an accepted gold-standard regulatory graph. Across both settings, we emphasize two outputs of MT-BN: multi-resolution modular organization (the learned hierarchy) and within-resolution directed structure defined on innovation latents.

7.1 Benchmark with Ground Truth: DREAM Network Inference

We benchmark MT-BN on the DREAM network inference challenge setting, which provides simulated gene-expression data together with a revealed gold-standard network after prediction submission. Performance is evaluated using area under the precision–recall curve (AUPR) and area under the receiver operating characteristic curve (AUROC), which are standard for gene network inference due to severe class imbalance between true and false edges. The DREAM network inference challenge and its scoring protocol are described in prior work.[23, 29]

Evaluation protocol. For each network instance, MT-BN produces a ranked list of directed edges with associated confidence scores, obtained from the stability-weighted edge summaries described in Section 5.8. We compare these ranked edge lists to the gold-standard adjacency and report AUPR and AUROC. We also compare to a set of flat Bayesian network baselines (denoted BN1–BN6), which correspond to runs submitted in the network inference challenge, holding the evaluation pipeline fixed.

Table 1: DREAM benchmark performance (AUPR/AUROC). MT-BN achieves higher AUPR and AUROC than the flat BN baselines on Networks 1 and 3, and matches or modestly improves performance on Network 4.

Method	Net1 AUPR	Net1 AUROC	Net3 AUPR	Net3 AUROC	Net4 AUPR	Net4 AUROC
MT-BN (ours)	0.325	0.733	0.071	0.597	0.020	0.556
Bayesian networks 1	0.218	0.636	0.041	0.539	0.018	0.501
Bayesian networks 2	0.191	0.647	0.043	0.540	0.018	0.502
Bayesian networks 3	0.042	0.678	0.021	0.549	0.020	0.516
Bayesian networks 4	0.080	0.683	0.021	0.551	0.020	0.516
Bayesian networks 5	0.080	0.683	0.021	0.551	0.020	0.516
Bayesian networks 6	0.043	0.519	0.021	0.559	0.019	0.512

Interpretation. The improvements on Networks 1 and 3 are consistent with the hypothesis that multi-scale modularization and topology-aware graph priors reduce the effective search space and yield better edge ranking under limited samples. On Network 4, MT-BN is competitive with the strongest BN baselines in AUPR and improves AUROC, suggesting that the main gains arise in regimes where flat learners are more sensitive to local optima and spurious dependencies induced by high dimensionality.

7.2 Large-Scale Transcriptomics without Ground Truth: Tuberculosis

We next apply MT-BN to a tuberculosis (TB) gene-expression dataset after batch-effect removal. Unlike DREAM, there is no accepted gold-standard directed regulatory network at this scale. The purpose of this case study is therefore not to compute edge-accuracy metrics, but to test whether MT-BN learns modules and directed inter-module structure that align with well-established host-response biology.

Data and objective. We analyze a matrix

$$X \in \mathbb{R}^{n \times p}, \qquad p = 6503, \quad n = 2722,$$

where columns are genes and rows are samples. MT-BN is run to infer a hierarchy of nested gene modules across multiple resolutions, together with within-level directed graphs on innovation latents as defined in Section 5.5. We summarize results using (i) learned modules, (ii) hub structure and directed module-to-module dependencies, and (iii) external functional enrichment support.

External functional support via STRING. For each learned module, we evaluate enrichment of known functional relationships using STRING, which integrates heterogeneous evidence sources and reports interaction confidence scores on a 0–1 scale.[30, 31] Because STRING confidence scores represent likelihood of association rather than interaction strength, we use them only as independent corroboration that module gene sets correspond to coherent biological programs, not as validation of directionality.

7.2.1 Finding 1: Interferon/ISG program captured as a coherent module

MT-BN identifies a module enriched for canonical interferon-stimulated genes (ISGs) organized around transcription factors including STAT1 and IRF7, with prominent members including ISG15, IFIT1–3, OASL, MX1, MX2, and GBP family genes. This structure matches established blood transcriptional signatures of active TB that emphasize upregulation of type I interferon signaling and downstream ISG programs.[32, 33]

In the MT-BN within-level graph summaries, this module exhibits a hub-dominated pattern in which STAT1/IRF-centered nodes have high outgoing influence scores and large stability-weighted edge strengths, consistent with a cascade-like activation structure within the module under the innovation-based SEM.

7.2.2 Finding 2: Upstream cytokine signaling module consistent with JAK–STAT control

MT-BN also identifies a distinct module centered on cytokine signaling components, including JAK kinases and multiple STAT family transcription factors. At the inter-module scale, the learned directed structure places arrows from this signaling-control module into the ISG module, consistent with the biological pathway logic in which cytokine stimulation activates JAK kinases, which in turn activate STAT transcription factors that drive ISG expression. This aligns with the model's goal of learning within-resolution directionality on innovation latents, where inherited shared signal is conditioned out by higher-level latents.

Implications for biomarker nomination. Although causality claims depend on modeling assumptions, MT-BN yields a practical biomarker-discovery workflow: identify robust modules corresponding to disease-relevant programs, then nominate compact gene panels using stability-weighted hub centrality and edge-strength summaries within those modules. The TB results suggest that MT-BN concentrates candidate biomarkers into interpretable, pathway-coherent subnetworks rather than diffuse edge sets, which is particularly valuable at $p \approx 6500$ where flat graph learners often produce unstable structure.

Limitations of this case study. Because the TB dataset does not provide a gold-standard directed network, we treat these results as biological plausibility evidence supported by external enrichment and prior knowledge, rather than as a definitive validation of directed edges. In future work, these module-derived biomarker candidates should be validated on independent cohorts and, where possible, linked to protein-level assays and prospective outcomes.

8 Discussion and Future Work

MT-BN is intended as a general framework for multi-resolution directed dependence modeling in the high-dimensional regime. The current formulation establishes a coherent probabilistic model, introduces the innovation-based causality

principle, and demonstrates that the resulting outputs align with ground-truth benchmarks and biologically plausible structure in large transcriptomic data. This section outlines several concrete directions that would strengthen MT-BN both methodologically and empirically.

8.1 Information-theoretic objectives for principled partitioning

In the current paper, the hierarchy $\mathcal{T}$ is learned under a truncated nCRP prior coupled to the global variational structure score $\mathcal{F}(S;\widehat{q}_S)$. This couples modularization and structure learning, but the partitioning mechanism is still primarily prior-driven (through the nCRP) and score-driven indirectly through improvements in the ELBO when modules and graphs better explain the data. A natural next step is to incorporate an explicit information-theoretic objective that encourages partitions to be optimal with respect to a measurable tradeoff between compression and predictive structure.

One principled approach is to interpret module latents as a compressed representation of the observed variables and enforce that, at each resolution, modules preserve information relevant to predicting other modules while discarding redundant within-module signal. Concretely, for a partition at level ℓ with module index random variable $M^{(\ell)}$ and module latents $Z^{(\ell)}$, one can introduce a term of the form

$$\mathcal{J}_{\mathrm{IT}}^{(\ell)} = \underbrace{I\Big(X; Z^{(\ell)}\Big)}_{\text{representation fidelity}} - \beta_\ell \underbrace{I\Big(Z^{(\ell)}; Z^{(\ell)}_{\setminus\cdot}\Big)}_{\text{redundancy across modules}} + \gamma_\ell \underbrace{I\Big(U^{(\ell)}; U^{(\ell)}_{\setminus\cdot}\Big)}_{\text{predictive cross-module structure}}, \tag{53}$$

where $I(\cdot;\cdot)$ denotes mutual information, $Z^{(\ell)}_{\setminus\cdot}$ denotes latents of other modules, and $U^{(\ell)}$ are the innovation latents at level ℓ. The first term favors partitions whose module latents retain information about observed variables, the second penalizes redundant representations across modules (encouraging cleaner separation), and the third rewards partitions that preserve the cross-module dependence structure that MT-BN aims to model causally at that resolution.

Operationally, (53) would be implemented through tractable variational bounds on mutual information, yielding additional terms in the ELBO. The tuning parameters β_ℓ and γ_ℓ would control the tradeoff between compressing within-module variation and preserving useful cross-module predictive structure. The expected outcome is a hierarchy that is less sensitive to the nCRP concentration parameters $\{\alpha_\ell\}$ and more directly optimized for the downstream goal of stable directed structure learning.

8.2 Causal hierarchy learning beyond structural nesting

MT-BN currently enforces a structural hierarchy through nested partitions and a vertical inheritance model. A further direction is to make the hierarchy itself explicitly causal rather than merely structural. The conceptual goal is to treat coarser-resolution latents as causal summaries that mediate shared signal across their descendants, while enforcing that fine-resolution innovations represent genuinely new mechanisms emerging at that resolution. This could be formalized by introducing explicit interventions at higher levels in the generative model and requiring invariance of within-level innovation structure across environments or conditions.

Concretely, one could extend MT-BN to multi-environment data by indexing samples with an environment variable $e \in \{1,\dots,E\}$ and enforcing that the within-level innovation SEMs share structure across e, while inherited latents and noise distributions may vary:

$$G^{(\ell)} \text{ and } \{A^{(\ell)}_{j\to m}\} \text{ are shared across environments,} \qquad p\Big(Z^{(\ell-1)} \mid e\Big) \text{ may vary.} \tag{54}$$

If edges among innovations are stable across environments after conditioning on higher-level latents, the resulting directed structure admits a stronger causal interpretation, and the hierarchy becomes not only a convenient multiscale representation but also an explicit causal abstraction.

8.3 Alternative latent parameterizations and identifiability

The current formulation uses linear Gaussian SEMs on innovation latents and linear inheritance maps across levels. This choice yields conjugacy and tractable variational updates, which is essential for scalability. Nonetheless, several extensions may improve expressiveness while preserving the key modeling principle of innovation-based causality.

First, one can replace the linear SEM in (23) with a nonlinear additive-noise SEM,

$$U_m^{(\ell)} = \sum_{j\in \mathrm{Pa}(m)} f_{j\to m}^{(\ell)}\Big(U_j^{(\ell)}\Big) + \eta_m^{(\ell)},$$

with neural or kernel parameterizations for $f^{(\ell)}_{j\to m}$. Second, one can introduce sparsity-inducing priors directly on $A^{(\ell)}_{j\to m}$ (e.g., spike-and-slab or hierarchical shrinkage) to improve edge selection when $d_{\max}$ is not sufficiently restrictive. Third, one can improve identifiability of latent coordinates by enforcing constraints such as orthonormality of module loadings, whitening of innovation covariances, or anchoring on observed marker variables when available. These additions would strengthen interpretability of effect magnitudes and improve comparability across runs.

8.4 Tuberculosis: biomarker nomination and validation pipeline

The TB case study demonstrates that MT-BN recovers modules corresponding to biologically prominent immune programs, including interferon/ISG signatures and JAK–STAT signaling control. The next stage is to convert these qualitative findings into a rigorous biomarker identification and validation program.

A concrete pipeline is as follows. First, run MT-BN across multiple random seeds and, where possible, across repeated subsamples of the cohort to produce stability-based edge and hub summaries (Section 5.8). For each learned module, compute stability-weighted hub scores $\widetilde{s}^{(\ell)}_{j\to m}$ and degree summaries to identify genes that act as consistent drivers within disease-relevant modules. Second, select compact candidate panels by optimizing a predictive criterion subject to interpretability constraints. For example, one can select a panel $P \subseteq \{1,\dots,p\}$ of size k by maximizing predictive accuracy of TB status or clinically relevant endpoints while regularizing toward MT-BN hubs and module coverage:

$$\max_{|P|=k} \ \mathrm{Perf}(P) \ - \ \lambda_1 \sum_{i\in P} \mathrm{Pen}_{\mathrm{nonhub}}(i) \ - \ \lambda_2 \, \mathrm{Pen}_{\mathrm{redundancy}}(P), \tag{55}$$

where $\mathrm{Perf}(P)$ can be AUPR or AUROC for classification, and penalties enforce that selected genes are stable drivers and not redundant within the same submodule.

Third, validate panels on independent cohorts, including geographically and demographically distinct cohorts where batch effects and confounding differ. Fourth, where longitudinal data exist, evaluate whether MT-BN-derived panels track treatment response dynamics, which would be consistent with the interpretation that the learned modules reflect causal immune programs rather than static correlational signatures. Finally, connect candidate biomarkers to protein-level assays when feasible and evaluate whether the MT-BN-directed structure suggests testable intervention hypotheses, such as perturbing upstream signaling components to modulate downstream ISG responses.

A key advantage of MT-BN for biomarker discovery is that it provides a structured search space. Rather than selecting biomarkers from $p \approx 6500$ genes in an unstructured manner, MT-BN concentrates attention on coherent modules supported by external enrichment and yields a directed, innovation-based notion of control at the module and gene level. This enables biomarker panels that are not only predictive but also mechanistically interpretable and amenable to experimental follow-up.

8.5 Additional empirical directions

Several empirical extensions would further clarify MT-BN's operating regime. First, ablations should isolate the contributions of (i) hierarchical modularization, (ii) innovation-based structure learning, and (iii) topology-aware priors, by removing each component and measuring changes in edge recovery and stability. Second, domain transfer should be evaluated by running MT-BN on additional biological datasets with partial ground truth (e.g., curated pathway databases) and on financial datasets where regime shifts provide natural multi-environment variation for causal invariance testing.

9 Conclusion

This paper presents MT-BN, a multi-resolution Bayesian network framework for learning directed structure in the high-dimensional regime $p \gg n$. The motivating obstacle in this regime is not only estimation error, but the geometry of the hypothesis space: the number of candidate DAGs and edge orientations grows super-exponentially in p, while finite-sample evidence is typically insufficient to distinguish direct effects from indirect correlations. MT-BN responds by changing the object of inference. Rather than attempting to learn a single flat DAG on p variables, MT-BN learns a hierarchy of nested modules and a directed influence network at each resolution, thereby aligning the representation with the empirical modular and hub-dominated structure observed in many scientific domains.

The primary novelty is a new principle for multi-scale directed modeling: within-resolution directed structure should be defined on what is newly expressed at that resolution, not on signal inherited from coarser scales. MT-BN enforces this through an explicit decomposition of each module state into inherited signal and a resolution-specific innovation, and it places the within-level DAG exclusively on innovation latents. This innovation-based formulation

is the mechanism by which MT-BN avoids a fundamental confounding that arises in hierarchical representations: if directed edges are learned on full latent states that share inherited components, sibling modules can appear strongly coupled purely due to common ancestry, producing spurious within-level edges and unstable orientations. By conditioning on higher-level structure and learning directed dependencies only among innovations, MT-BN targets directed relationships among mechanisms that emerge at that scale. This is the conceptual point a reader should take away: MT-BN treats multi-resolution causality as resolution-conditional causality, where each level's arrows describe directed influence among new residual dynamics after accounting for all coarser-scale variation.

A second contribution is that MT-BN turns this principle into a unified probabilistic model that couples three tasks that are often separated in practice: hierarchical modularization, latent representation learning, and directed structure discovery. The hierarchy $\mathcal{T}$ is not a preprocessing artifact; it is a random object under an explicit prior and is learned jointly with within-level graphs and latent variables. The resulting outputs are not a single edge list but a coherent multi-scale object: nested modules, innovation trajectories at every level, and a directed network of within-level influence at every resolution. This coupling is essential both statistically and interpretably: it allows modularization to be guided by structure-learning utility and allows structure learning to be stabilized by modular organization and shared-signal removal.

A third contribution is tractability without abandoning Bayesian structure learning. MT-BN reduces the effective search space by operating on module graphs whose sizes M_ℓ are far smaller than p and by restricting attention to sparse, hub-dominated, hierarchy-consistent structures through topology-aware priors. These priors provide a principled mechanism for shrinking posterior mass away from implausible graphs while retaining flexibility to discover domain-specific structure. Combined with local score decomposability and a blocked hybrid inference strategy, MT-BN supports scalable approximate posterior inference and MAP-style structure estimation, with localized updates under single-edge and local hierarchy proposals.

Empirically, MT-BN improves edge-ranking performance relative to flat Bayesian network baselines on a benchmark setting with ground-truth edges and produces biologically coherent, externally supported organization on large-scale tuberculosis transcriptomics without an accepted gold-standard regulatory DAG. In the TB setting, MT-BN recovers interpretable immune programs, including interferon/ISG structure and upstream cytokine/JAK–STAT control, and it concentrates candidate drivers into pathway-coherent modules suitable for downstream biomarker nomination and experimental follow-up. These results illustrate the intended role of MT-BN: to generate stable, multi-resolution, mechanistically interpretable directed hypotheses in regimes where flat structure learning is brittle.

In summary, MT-BN contributes a new formulation of multi-scale directed learning in which causality at each resolution lives on innovations rather than inherited signal, together with a complete Bayesian framework that jointly learns partitions, latent representations, and within-level DAGs under topology-aware priors. This innovation-based perspective is not a minor modeling choice; it is what makes the multi-resolution directed graphs meaningful rather than artifacts of hierarchical aggregation. By making that principle explicit and implementable, MT-BN provides a practical and conceptually clean path toward scalable directed structure discovery in modern high-dimensional domains.

10 Acknowledgments

I thank Prof. Gil Alterovitz (Biomedical Cybernetics Laboratory, Harvard Medical School) for his mentorship and guidance throughout this project, and Dr. Ning Xie for valuable feedback and support. I am also grateful to Prof. Slava Gerovitch and Dr. Felix Gotti for the opportunity to participate in the MIT PRIMES Program.

References

[1] Gregory F. Cooper and Edward Herskovits. "A Bayesian Method for the Induction of Probabilistic Networks from Data". In: *Machine Learning* 9 (1992), pp. 309–347. DOI: 10.1023/A:1022649401552.

[2] Nir Friedman and Daphne Koller. "Being Bayesian About Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks". In: *Machine Learning* 50.1–2 (2003), pp. 95–125. DOI: 10.1023/A:1020249912095.

[3] Bradley M. Broom, Kim-Anh Do, and Devika Subramanian. "Model averaging strategies for structure learning in Bayesian networks with limited data". In: *BMC Bioinformatics* 13.Suppl 13 (2012), S10. DOI: 10.1186/1471-2105-13-S13-S10.

[4] Nir Friedman et al. "Using Bayesian Networks to Analyze Expression Data". In: *Journal of Computational Biology* 7.3–4 (2000), pp. 601–620. DOI: 10.1089/106652700750050961.

[5] Sung Won Han et al. "Estimation of Directed Acyclic Graphs Through Two-stage Adaptive Lasso for Gene Network Inference". In: *Journal of the American Statistical Association* 111.515 (2016), pp. 1004–1019. DOI: 10.1080/01621459.2016.1142880.

[6] Lupe S. H. Chan, Amanda M. Y. Chu, and Mike K. P. So. "A moving-window bayesian network model for assessing systemic risk in financial markets". In: *PLOS ONE* 18.1 (2023), e0279888. DOI: 10.1371/journal.pone.0279888.

[7] Mark E. J. Newman. "The Structure and Function of Complex Networks". In: *SIAM Review* 45.2 (2003), pp. 167–256. DOI: 10.1137/S003614450342480.

[8] Timo J. T. Koski and John M. Noble. "A Review of Bayesian Networks and Structure Learning". In: *Mathematica Applicanda* 40.1 (2012), pp. 53–103. DOI: 10.14708/ma.v40i1.278.

[9] David Maxwell Chickering. "Learning Bayesian Networks is NP-Complete". In: *Learning from Data: Artificial Intelligence and Statistics V*. Ed. by Doug Fisher and Hans-J. Lenz. Vol. 112. Lecture Notes in Statistics. New York, NY: Springer, 1996, pp. 121–130. DOI: 10.1007/978-1-4612-2404-4_12.

[10] Eyal Segal et al. "Learning Module Networks". In: *Journal of Machine Learning Research* 6.19 (2005), pp. 557–588. URL: https://jmlr.org/papers/v6/segal05a.html.

[11] Peter Langfelder, Bin Zhang, and Steve Horvath. "Eigengene networks for studying the relationships between co-expression modules". In: *BMC Systems Biology* 1.1 (2007), p. 54. DOI: 10.1186/1752-0509-1-54.

[12] Peter Langfelder and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1 (2008), p. 559. DOI: 10.1186/1471-2105-9-559.

[13] Alta de Waal and Keunyoung Yoo. "Latent Variable Bayesian Networks Constructed Using Structural Equation Modelling". In: *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 2018, pp. 688–695. ISBN: 978-0-9964527-7-9. DOI: 10.23919/ICIF.2018.8455240.

[14] James H. Stock and Mark W. Watson. *Implications of Dynamic Factor Models for VAR Analysis*. NBER Working Paper 11467. National Bureau of Economic Research, 2005. DOI: 10.3386/w11467. URL: https://www.nber.org/papers/w11467.

[15] David Heckerman, Dan Geiger, and David M. Chickering. "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data". In: *Machine Learning* 20.3 (1995), pp. 197–243. DOI: 10.1023/A:1022623210503.

[16] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: The MIT Press, 2009. ISBN: 9780262013192.

[17] David Maxwell Chickering. "Optimal Structure Identification with Greedy Search". In: *Journal of Machine Learning Research* 3 (2002), pp. 507–554. URL: http://jmlr.org/papers/v3/chickering02b.html.

[18] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: MIT Press, 2000. ISBN: 9780262194402.

[19] Xun Zheng et al. "DAGs with NO TEARS: Continuous Optimization for Structure Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 31. 2018. URL: https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html.

[20] Eran Segal et al. "Module Networks: Identifying Regulatory Modules and Their Condition-Specific Regulators from Gene Expression Data". In: *Nature Genetics* 34.2 (2003), pp. 166–176. DOI: 10.1038/ng1165.

[21] Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. "Latent Variable Graphical Model Selection via Convex Optimization". In: *The Annals of Statistics* 40.4 (2012), pp. 1935–1967. DOI: 10.1214/11-AOS949.

[22] Tiago P. Peixoto. "Hierarchical Block Structures and High-Resolution Model Selection in Large Networks". In: *Physical Review X* 4.1 (2014), p. 011047. DOI: 10.1103/PhysRevX.4.011047.

[23] Daniel Marbach et al. "Wisdom of Crowds for Robust Gene Network Inference". In: *Nature Methods* 9.8 (2012), pp. 796–804. DOI: 10.1038/nmeth.2016.

[24] Steffen L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford: Clarendon Press, 1996. ISBN: 9780198522195.

[25] Gil Alterovitz et al. "Bayesian Methods for Proteomics". In: *Proteomics* 7.16 (2007), pp. 2843–2855. DOI: 10.1002/pmic.200700422.

[26] Andrew Gelman et al. *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC, 2013. ISBN: 9781439840955. DOI: 10.1201/b16018.

[27] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

[28] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97.

[29] DREAM Challenges. *DREAM5 – Network Inference Challenge*. Challenge page and dataset repository on Synapse; network inference task for gene regulatory networks, part of the DREAM5 systems biology challenges. Synapse / Sage Bionetworks. 2010. URL: https://www.synapse.org/Synapse:syn2787209.

[30] Damian Szklarczyk et al. "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any set of proteins". In: *Nucleic Acids Research* 51.D1 (2023), pp. D638–D646. DOI: 10.1093/nar/gkac1000.

[31] *STRING: Functional Protein Association Networks (Help / Documentation)*. STRING Consortium. URL: https://string-db.org/help/.

[32] Tom H. M. Ottenhoff et al. "Genome-wide expression profiling identifies a type I interferon signature in the blood of tuberculosis patients". In: *PLOS ONE* 7.11 (2012), e49630. DOI: 10.1371/journal.pone.0045839.

[33] Fanli Yi et al. "Transcriptional Profiling of Human Peripheral Blood Mononuclear Cells Stimulated by Mycobacterium tuberculosis PPE57 Identifies Characteristic Genes Associated With Type I Interferon Signaling". In: *Frontiers in Cellular and Infection Microbiology* 11 (2021), p. 716809. DOI: 10.3389/fcimb.2021.716809. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8416891/.