FROM MEASUREMENTS TO CONCEPTS: LLM-GUIDED LATENT-AWARE CAUSAL DISCOVERY FOR MEDICAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery provides a structural basis for intervention and counterfactual analysis, making it valuable for understanding complex disease processes from observational medical data. However, medical datasets often contain multiple measurements reflecting shared but unobserved clinical states. Existing latent-variable causal discovery methods primarily search for hidden structure within predefined statistical model classes, but do not explicitly determine the appropriate level of causal abstraction. We formulate this problem as *causal representation selection*, coupling causal representation selection with causal graph learning. We propose LLCD, a framework explicitly addresses latent-aware causal discovery from this representation-selection perspective. LLCD uses LLM-derived semantics to propose candidate latent concepts, learns them with a missing-aware variational model, and selects among alternative latent–observed representations using held-out structural adequacy. We evaluate LLCD on semi-synthetic benchmarks with known latent causal structure and real-world ADNI and PPMI cohorts, assessing latent concept recovery, causal structure recovery, and causal interventional analysis. On the semi-synthetic benchmarks, LLCD improves directed F1 by 66.0% and 63.5% over the strongest baselines on ADNI and PPMI, respectively. LLCD also reduces causal-effect NMAE by approximately 20% on both datasets. The code has been made public [1].

1 INTRODUCTION

Causal discovery aims to recover the underlying causal structure of a system from observational data, providing a basis for downstream intervention analysis and counterfactual reasoning. In medical applications, however, clinically meaningful states are often not directly observed but are reflected through multiple correlated measurements. Cognitive status, for example, may be characterized by several neuropsychological assessments. Treating every measurement as an independent causal entity can obscure the higher-level organization of the disease process, making the choice of latent causal representation particularly important for medical causal discovery.

Existing latent-variable causal discovery methods recover hidden structure using conditional-independence, covariance-rank, measurement-model, or score-based assumptions Spirtes et al. (2001); Silva et al. (2006); Kummerfeld & Ramsey (2016); Dong et al. (2024; 2026). These methods are statistically principled, but typically search within a predefined latent-variable model class. In biomedical settings, the candidate abstractions themselves are often suggested by variable semantics: several measurements may be known to quantify a common pathological or functional state. Large language models (LLMs) can exploit such metadata, but semantic plausibility alone does not establish that a proposed latent construct is supported by the observed data. Moreover, clinical measurements are frequently incomplete, making latent representation learning and subsequent causal discovery more difficult. A practical latent-aware discovery method must therefore address three coupled questions: which latent concepts should be considered, how they should be represented from incomplete measurements, and whether replacing their indicators by latent variables yields a better causal representation.

[1] https://anonymous.4open.science/r/research-artifact-20260923-b36143-773B

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We formulate this problem as *causal representation selection*. Rather than accepting a fixed observed or latent node space, we consider alternative representations in which each semantically proposed measurement group can either remain observed or be replaced by a latent clinical construct. The resulting candidate representations have different node spaces and therefore cannot be reliably compared by their raw graph scores alone. The goal is to use semantic knowledge to define plausible hypotheses while letting observational data determine which abstractions should be retained and what causal structure should be learned over them.

To this end, we introduce **LLCD**, an LLM-guided framework for latent-aware causal discovery, designed for settings where candidate causal abstractions have meaningful domain semantics, rather than arbitrary unidentified latent factors. LLCD first uses variable metadata to propose candidate latent constructs, their measurement indicators, auxiliary context, and coarse causal layers. For each candidate representation, it learns selected constructs with a mask-aware, context-conditioned variational model, allowing partially observed indicators and context variables to contribute without global imputation. It then performs layer-constrained score-based causal discovery in the resulting mixed latent–observed space. LLCD evaluates each learned representation on held-out data using residual dependence implied by the discovered graph, retains representations within a tolerance of the best structural adequacy score, and selects the simplest among them. This provides a quantitative, data-driven criterion for deciding whether a proposed latent abstraction is useful for causal modeling.

We evaluate LLCD on semi-synthetic ADNI and PPMI benchmarks with known latent causal structure and on the corresponding real-world cohorts. We study whether LLCD recovers meaningful latent concepts, improves latent-aware causal structure recovery, and yields better structural causal models for causal analysis. On the semi-synthetic benchmarks, LLCD improves directed F1 by 66.0% and 63.5% over the strongest baselines on ADNI and PPMI, respectively, and reduces intervention-effect NMAE by about 20% on both datasets. On the real cohorts, the selected representations also achieve the best held-out SCM fitting.

2 Related Work

2.1 LLM-assisted Causal Discovery

Classical causal discovery infers structure primarily from observational statistics, including constraint-based PC and FCI Spirtes et al. (2001), score-based GES Chickering (2002), and methods based on functional or optimization assumptions such as DirectLiNGAM and NOTEARS Shimizu et al. (2006); Zheng et al. (2018). These approaches have limited access to variable semantics, which can help distinguish statistically compatible structures with different domain interpretations.

Recent work therefore uses large language models as semantic priors for causal discovery Kıcıman et al. (2023). LLMs have been queried for pairwise causal relations or used in structured graph construction Long et al. (2023); Jiralerspong et al. (2024), while Causal Modelling Agents combine metadata reasoning with data-driven structural causal modeling Abdulaal et al. (2024). More recent methods use LLM-derived priors to constrain conventional discovery procedures Du et al. (2025); Chen et al. (2026); Vashishtha et al. (2025); Darvariu et al. (2024). These studies motivate combining semantic knowledge with observational evidence.

2.2 Causal Discovery with Latent Variables

Latent-variable causal discovery spans measurement-model recovery and more general structure learning. BPC and FOFC identify measurement clusters using covariance or tetrad constraints Silva et al. (2006); Kummerfeld & Ramsey (2016), TSC exploits trek-separation rank constraints Sullivant et al. (2010), and MIMBuild estimates relations among recovered latent variables Silva et al. (2006). RLCD uses covariance-rank constraints to recover measured–latent structures Dong et al. (2024), while LGES performs score-based search under generalized latent factor models Dong et al. (2026).

Recent work further broadens this setting. Prashant et al. Prashant et al. (2025) develop differentiable discovery for nonlinear hierarchical latent causal models with identifiability guarantees, while Li and Liu Li & Liu (2025) study complex observed–latent relations under explicit structural assumptions. Chen et al. Chen et al. (2024) use tensor-rank conditions for discrete latent structures, and Dai et al. Dai et al. (2025) extend rank-based discovery to selection bias. Sun et al. Sun et al. (2025) learn

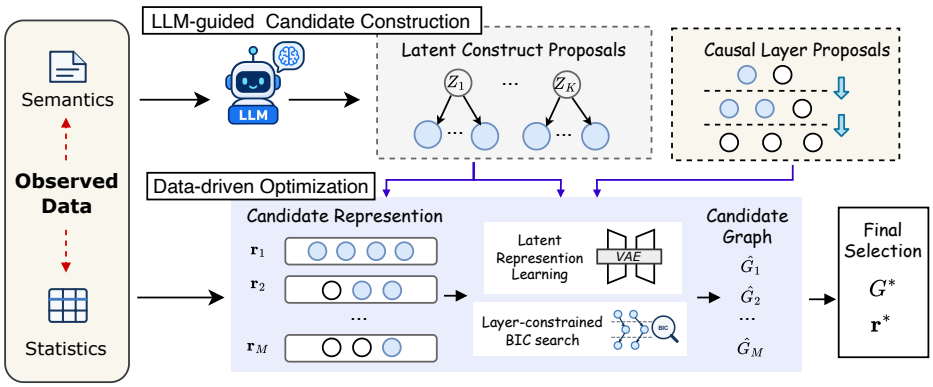


Figure 1: Overview of LLCD. Candidate latent constructs define alternative latent-aware representations, which are evaluated to select the final representation $\mathbf{r}^*$ and causal graph G^*.

identifiable latent causal representations from multimodal biomedical observations. In contrast, LLCD uses medical semantics to propose clinically meaningful latent concepts and lets data select the final representation and causal graph.

3 Problem Setup

Let $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ denote an observational dataset over recorded variables $\mathcal{X} = \{X_1, \ldots, X_p\}$. We consider settings in which some groups of observed variables may jointly reflect underlying clinical constructs that are not directly recorded.

Suppose we are given a set of candidate latent constructs

$$\mathcal{C} = \{c_k = (Z_k, \mathcal{I}_k)\}_{k=1}^{K}, \tag{1}$$

where Z_k denotes a candidate latent construct and $\mathcal{I}_k \subseteq \mathcal{X}$ is its associated set of measurement indicators, and let $\mathcal{Z} = \{Z_k\}_{k=1}^{K}$ denote the corresponding set of candidate latent variables. These candidates are not assumed to belong to the true causal system a priori; rather, they define alternative latent-aware representations to be evaluated.

For a candidate Z_k, its indicators are assumed to admit a measurement interpretation of the form $X_{kj} = g_{kj}(Z_k, \eta_{kj})$ for $X_{kj} \in \mathcal{I}_k$, where η_{kj} captures measurement-specific variation.

A latent-aware representation is indexed by $\mathbf{r} = (r_1, \ldots, r_K) \in \{0,1\}^K$. If $r_k = 1$, candidate Z_k is instantiated and its indicators are represented through the latent construct; if $r_k = 0$, the corresponding indicators remain as observed structural variables. Let $\mathcal{Z}_\mathbf{r} = \{Z_k : r_k = 1\}$ and $\mathcal{M}_\mathbf{r} = \bigcup_{k:r_k=1} \mathcal{I}_k$. The structural node space associated with representation $\mathbf{r}$ is therefore

$$\mathcal{V}_\mathbf{r} = (\mathcal{X} \setminus \mathcal{M}_\mathbf{r}) \cup \mathcal{Z}_\mathbf{r}. \tag{2}$$

For each representation, we distinguish between measurement relations and structural causal relations. Selected constructs introduce measurement edges $E_\mathbf{r}^{\text{meas}} = \{Z_k \to X_{kj} : r_k = 1,\ X_{kj} \in \mathcal{I}_k\}$, while structural edges $E_\mathbf{r}^{\text{str}}$ are defined over $\mathcal{V}_\mathbf{r}$. The complete latent-aware graph is $G_\mathbf{r} = (\mathcal{X} \cup \mathcal{Z}_\mathbf{r}, E_\mathbf{r}^{\text{str}} \cup E_\mathbf{r}^{\text{meas}})$.

Let $\mathcal{R}$ denote the set of admissible representations induced by the candidate constructs. The problem is to identify a representation $\mathbf{r}^* \in \mathcal{R}$ and its corresponding causal graph $G^* = G_{\mathbf{r}^*}$ that provide an appropriate latent-aware description of the observed system.

4 Method

4.1 Overview

Figure 1 illustrates LLCD. Given variable metadata $\mathcal{S}$, an LLM proposes candidate latent constructs $\mathcal{Z}$, their measurement indicators $\mathcal{I}_k$, auxiliary context variables, and coarse causal layers ℓ. These

Algorithm 1 LLCD: LLM-guided latent concept causal discovery

Require: Training data $\mathcal{D}_{\mathrm{tr}}$, validation data $\mathcal{D}_{\mathrm{val}}$, metadata $\mathcal{S}$
Ensure: Selected representation $\mathbf{r}^*$ and graph G^*
Obtain candidate latent concepts $\mathcal{C}$ and causal layers ℓ from the LLM
Enumerate valid candidate representations $\mathcal{R}$
for all $\mathbf{r} \in \mathcal{R}$ **do**
  for all selected Z_k in context-topological order **do**
    Resolve auxiliary context under $\mathbf{r}$
    Learn the context-conditioned latent representation of Z_k
    Replace $\mathcal{I}_k$ by posterior mean $\widehat{\mathbf{z}}_k$
  end for
  Learn $\widehat{G}^{\mathrm{str}}_{\mathbf{r}}$ by layer-constrained BIC search
  Compute held-out structural score $v_{\mathrm{CI}}(\mathbf{r})$
  Compute $C(\mathbf{r}) = |\widehat{E}^{\mathrm{str}}_{\mathbf{r}}| + \lambda_Z |\mathcal{Z}_{\mathbf{r}}|$
end for
$v_{\min} \leftarrow \min_{\mathbf{r}\in\mathcal{R}} v_{\mathrm{CI}}(\mathbf{r})$
$\mathcal{R}_\epsilon \leftarrow \{\mathbf{r} : v_{\mathrm{CI}}(\mathbf{r}) \le v_{\min} + \epsilon\}$
$\mathbf{r}^* \leftarrow \arg\min_{\mathbf{r}\in\mathcal{R}_\epsilon} C(\mathbf{r})$
return $\mathbf{r}^*, G^*$

proposals define alternative latent-aware representations. For each representation, LLCD learns the selected latent constructs, performs layer-constrained causal discovery over the resulting mixed latent–observed space, and evaluates the learned structure on held-out data. The final representation is selected by balancing structural adequacy and complexity.

4.2 Semantic Latent Concept Proposal

We provide the LLM only with variable-level metadata $\mathcal{S}$, including variable names, clinical descriptions, and temporal annotations when available. It returns

$$\{(Z_k, \mathcal{I}_k, \mathcal{U}_k)\}_{k=1}^{K}, \tag{3}$$

where $\mathcal{I}_k$ is the proposed measurement indicator set and

$$\mathcal{U}_k \subseteq (\mathcal{X} \setminus \mathcal{I}_k) \cup (\mathcal{Z} \setminus \{Z_k\}) \tag{4}$$

is an additional set of auxiliary context variables used only for latent representation learning. Membership in $\mathcal{U}_k$ does not imply a structural edge into Z_k.

The LLM also provides a coarse causal layer assignment ℓ, used to restrict implausible edge directions during subsequent graph search. The candidate latent constructs induce a set of admissible representations

$$\mathcal{R} = \left\{\mathbf{r} \in \{0,1\}^K\right\}, \tag{5}$$

where each $\mathbf{r}$ specifies a compatible subset of latent constructs to instantiate. In longitudinal settings, baseline and follow-up versions of the same construct are additionally constrained to be selected jointly.

4.3 Context-Conditioned Latent Representation

Because measurement indicators and context variables may contain missing values, we define

$$f(\mathbf{x}_A, \mathbf{m}_A) = [\mathbf{x}_A \odot \mathbf{m}_A, \mathbf{m}_A], \tag{6}$$

where $\mathbf{m}_A$ is the observation mask. For representation $\mathbf{r}$, each auxiliary context variable $U \in \mathcal{U}_k$ is resolved as

$$\rho_{\mathbf{r}}(U) = \begin{cases} \widehat{\mathbf{z}}_j, & U = Z_j,\ r_j = 1, \\ f(\mathbf{x}_{\mathcal{I}_j}, \mathbf{m}_{\mathcal{I}_j}), & U = Z_j,\ r_j = 0, \\ f(\mathbf{x}_U, \mathbf{m}_U), & U \in \mathcal{X}, \end{cases} \tag{7}$$

with

$$\mathbf{c}_k^{(\mathbf{r})} = \mathrm{Concat}_{U \in \mathcal{U}_k} \rho_{\mathbf{r}}(U). \tag{8}$$

When a selected latent construct is used as context for another selected construct, its representation must be learned first so that it is available when constructing the downstream context. We therefore define a context-dependency graph

$$H_{\mathrm{ctx}} = (\mathcal{Z}, E_{\mathrm{ctx}}), \qquad E_{\mathrm{ctx}} = \{Z_j \to Z_k : Z_j \in \mathcal{U}_k \cap \mathcal{Z}\}, \tag{9}$$

where an edge $Z_j \to Z_k$ indicates that the learned representation of Z_j is required when constructing the context of Z_k. For a representation $\mathbf{r}$, selected constructs are learned in a topological order of the subgraph induced by $\mathcal{Z}_{\mathbf{r}}$. Representations that induce cycles in this context-dependency graph are discarded.

For each selected Z_k, we learn

$$p_{\psi_k}(\mathbf{z}_k \mid \mathbf{c}_k^{(\mathbf{r})}) = \mathcal{N}\Big(\boldsymbol{\mu}_{\psi_k}^p, \mathrm{diag}[(\boldsymbol{\sigma}_{\psi_k}^p)^2]\Big), \tag{10}$$

$$q_{\phi_k}(\mathbf{z}_k \mid \mathbf{x}_{\mathcal{I}_k}, \mathbf{m}_{\mathcal{I}_k}, \mathbf{c}_k^{(\mathbf{r})}) = \mathcal{N}\Big(\boldsymbol{\mu}_{\phi_k}^q, \mathrm{diag}[(\boldsymbol{\sigma}_{\phi_k}^q)^2]\Big), \tag{11}$$

where the distribution parameters are predicted from the corresponding context and mask-aware measurement inputs. The model is optimized using

$$\mathcal{L}_k = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{ctx}}\mathcal{L}_{\mathrm{ctx}} + \beta D_{\mathrm{KL}}(q_{\phi_k} \| p_{\psi_k}) + \lambda_{\mathrm{scale}}\mathcal{L}_{\mathrm{scale}}, \tag{12}$$

where, with $\|\mathbf{a}\|_{\mathbf{m}}^2 = \sum_i m_i a_i^2 / (\sum_i m_i + \epsilon)$,

$$\begin{aligned} \mathcal{L}_{\mathrm{rec}} &= \|d_{\theta_k}(\mathbf{z}_k) - \mathbf{x}_{\mathcal{I}_k}\|_{\mathbf{m}_{\mathcal{I}_k}}^2, \quad \mathbf{z}_k \sim q_{\phi_k}, \\ \mathcal{L}_{\mathrm{ctx}} &= \|d_{\theta_k}(\boldsymbol{\mu}_{\psi_k}^p) - \mathbf{x}_{\mathcal{I}_k}\|_{\mathbf{m}_{\mathcal{I}_k}}^2, \\ \mathcal{L}_{\mathrm{scale}} &= \mathrm{Mean}(\boldsymbol{\mu}_{\phi_k}^q)^2 + \big(\mathrm{Std}(\boldsymbol{\mu}_{\phi_k}^q) - 1\big)^2. \end{aligned} \tag{13}$$

Here, $\mathcal{L}_{\mathrm{rec}}$ reconstructs the observed indicators from the posterior latent representation, while $\mathcal{L}_{\mathrm{ctx}}$ requires the context-conditioned prior to retain predictive information about the same indicators. The KL term aligns the posterior with this prior, and $\mathcal{L}_{\mathrm{scale}}$ regularizes the latent representation toward zero mean and unit variance.

For causal discovery, we use the posterior mean

$$\widehat{\mathbf{z}}_k = \boldsymbol{\mu}_{\phi_k}^q \left(f(\mathbf{x}_{\mathcal{I}_k}, \mathbf{m}_{\mathcal{I}_k}), \mathbf{c}_k^{(\mathbf{r})}\right). \tag{14}$$

4.4 Causal Structure Learning

For each candidate observed or latent variable $V \in \mathcal{X} \cup \mathcal{Z}$, the LLM assigns a coarse causal layer $\ell(V) \in \{0, \ldots, L\}$, where a smaller layer index indicates an earlier position in the causal ordering. For each representation $\mathbf{r}$, the transformed dataset $\mathcal{D}_{\mathrm{tr}}^{(\mathbf{r})}$ replaces the indicators of each selected construct by $\widehat{\mathbf{z}}_k$ and retains all remaining observed variables. The layer assignment restricts the admissible graph space to

$$\mathcal{G}_\ell^{(\mathbf{r})} = \{G = (\mathcal{V}_{\mathbf{r}}, E) : G \text{ is a DAG}, \; V_i \to V_j \in E \Rightarrow \ell(V_i) \le \ell(V_j)\}. \tag{15}$$

Thus, edges across layers can only proceed from earlier to later layers, while same-layer edges are allowed and their orientations are determined from data.

Each latent construct Z_k is treated as a single grouped causal node, although its learned representation $\mathbf{z}_k \in \mathbb{R}^{d_k}$ may be multidimensional. For a node V_j with parents $\mathrm{Pa}_G(V_j)$, we concatenate the numerical representations of its parent nodes into $\mathbf{h}_j$ and use the local model

$$\mathbf{v}_j = \mathbf{B}_j \mathbf{h}_j + \mathbf{b}_j + \boldsymbol{\varepsilon}_j, \qquad \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_j). \tag{16}$$

Thus, scalar observed nodes use standard linear-Gaussian regression, whereas multidimensional latent nodes use multivariate Gaussian regression.

The graph is estimated by

$$\widehat{G}_{\mathbf{r}}^{\mathrm{str}} = \arg\max_{G \in \mathcal{G}_\ell^{(\mathbf{r})}} \mathrm{BIC}\left(G; \mathcal{D}_{\mathrm{tr}}^{(\mathbf{r})}\right), \tag{17}$$

using greedy add, delete, and reverse operations. The BIC is the sum of local maximized Gaussian log-likelihoods with the corresponding parameter penalty. Hence, the LLM constrains the admissible search space but does not determine individual structural edges.

4.5 Global Causal Representation Selection

Because BIC scores are not directly comparable across representations with different node spaces, we use held-out structural adequacy for representation selection. For each $\widehat{G}^{\mathrm{str}}_{\mathbf{r}}$, auxiliary linear-Gaussian mechanisms are fitted on the training split and evaluated on the validation split. For a node V_j, let $\mathrm{Pa}_{\mathbf{r}}(V_j)$ and $\mathrm{De}_{\mathbf{r}}(V_j)$ denote its parents and descendants, respectively, in $\widehat{G}^{\mathrm{str}}_{\mathbf{r}}$. We define

$$\mathcal{A}_j = \mathcal{V}_{\mathbf{r}} \setminus (\{V_j\} \cup \mathrm{Pa}_{\mathbf{r}}(V_j) \cup \mathrm{De}_{\mathbf{r}}(V_j)), \tag{18}$$

which contains the nodes that should not retain systematic dependence with the residual of V_j under the learned structure.

Let $\widehat{\varepsilon}_{j,a}$ denote the a-th dimension of the validation residual of V_j, and $V_{l,b}$ the b-th numerical dimension of another node V_l. We measure residual dependence by

$$v_{\mathrm{CI}}(\mathbf{r}) = \frac{1}{|\mathcal{T}_{\mathbf{r}}|} \sum_{(j,a,l,b)\in\mathcal{T}_{\mathbf{r}}} |\mathrm{Corr}(\widehat{\varepsilon}_{j,a}, V_{l,b})|, \tag{19}$$

where $\mathcal{T}_{\mathbf{r}}$ contains all valid residual–variable dimension pairs with $V_l \in \mathcal{A}_j$. Lower values indicate less unexplained dependence on held-out data.

We define the set of structurally adequate representations as

$$\mathcal{R}_\epsilon = \left\{ \mathbf{r} \in \mathcal{R} : v_{\mathrm{CI}}(\mathbf{r}) \le \min_{\mathbf{r}'\in\mathcal{R}} v_{\mathrm{CI}}(\mathbf{r}') + \epsilon \right\}, \tag{20}$$

where ϵ controls the tolerance from the best held-out structural score. Among these representations, we favor simpler structures using

$$C(\mathbf{r}) = |\widehat{E}^{\mathrm{str}}_{\mathbf{r}}| + \lambda_Z |\mathcal{Z}_{\mathbf{r}}|, \tag{21}$$

where $|\widehat{E}^{\mathrm{str}}_{\mathbf{r}}|$ is the number of learned structural edges and λ_Z controls the penalty on instantiated latent constructs.

The final representation and graph are

$$\mathbf{r}^* = \arg\min_{\mathbf{r}\in\mathcal{R}_\epsilon} C(\mathbf{r}), \qquad G^* = \left(\mathcal{X} \cup \mathcal{Z}_{\mathbf{r}^*}, E^{\mathrm{meas}}_{\mathbf{r}^*} \cup \widehat{E}^{\mathrm{str}}_{\mathbf{r}^*}\right). \tag{22}$$

5 Experiments

5.1 Experimental Setup

Datasets We evaluate LLCD on two real-world medical datasets, ADNI Petersen et al. (2010) and PPMI Marek et al. (2018). ADNI contains 32 observed variables, including 13 longitudinal measurements available at baseline and follow-up, while PPMI contains 38 variables with 13 longitudinal measurements. Continuous variables are standardized using training-set statistics, and disjoint training, validation, and test splits are used for model fitting, representation selection, and final evaluation. Detailed variable definitions and cohort statistics are provided in Appendix A.1.1.

Semi-synthetic benchmark We additionally construct semi-synthetic versions of ADNI and PPMI with predefined latent measurement structure and causal graphs. ADNI contains four longitudinal latent construct types—amyloid, tau, neurodegeneration, and cognition—while PPMI contains motor severity, non-motor burden, and cognition. Synthetic data are generated from SCMs fitted according to these predefined latent-aware graphs, providing ground truth for latent measurement groups, causal structure, and interventional effects. Full graph specifications, measurement models, and data-generation procedures are provided in Appendix A.1.2.

Implementation details We use latent dimensions of 2 for ADNI and 4 for PPMI. Candidate representations are selected on the validation set with $\epsilon = 0.005$ and $\lambda_Z = 1$. For latent-construct proposal, we compare GPT-5.6 Sol, Gemini 3.8 Flash, and DeepSeek V4 Pro on the semi-synthetic benchmarks. As their results are similar, GPT-5.6 Sol is used in subsequent experiments unless otherwise specified. Complete optimization settings, hyperparameters, and prompt templates are provided in Appendix A.2.1 and A.2.4.

Baselines We compare LLCD against both observed-variable and latent-variable causal discovery methods. Observed-variable baselines include PC and FCI Spirtes et al. (2001), GES Chickering (2002), NOTEARS Zheng et al. (2018), DirectLiNGAM Shimizu et al. (2006), and the LLM-assisted methods CMA Abdulaal et al. (2024), CausalOrder Vashishtha et al. (2025), LLMprior Darvariu et al. (2024). Latent-variable baselines include BPC Silva et al. (2006)+MIMBuild, TSC Sullivant et al. (2010)+MIMBuild, FOFC Kummerfeld & Ramsey (2016)+MIMBuild, SALAD-CS Ng et al. (2024), and LGES Dong et al. (2026), covering measurement-model, rank-based, tetrad-based, and score-based latent structure discovery.

5.2 RQ1: Can LLCD Identify Plausible Latent Concepts?

Table 1 shows that different LLMs (GPT-5.6 Sol, Gemini 3.8 Flash, and DeepSeek V4 Pro) identify highly consistent measurement-related latent concepts. On ADNI, all three LLMs agree on all latent concepts. On PPMI, there is only minor disagreement for the broader non-motor symptom construct. This suggests that LLMs can provide stable semantic priors for clinically meaningful latent-variable grouping. We further asked a clinical expert to independently evaluate each unique concept–indicator combination on a five-point scale in terms of *construction plausibility* and *naming accuracy*. All proposed concepts received scores of at least 4 on both criteria, which demonstrates semantic latent structures proposed by the LLMs are medically plausible.

We further evaluate latent structure recovery on the semi-synthetic benchmarks in Table 2. The classical latent-variable baselines operate only on latent constructs and their indicators and are evaluated using the corresponding measurement-related variables, while LLCD performs discovery in the full variable space; all reported metrics are computed on the same latent–measurement subgraph. Predicted and ground-truth latent variables are matched one-to-one by maximizing the total Jaccard similarity between their indicator sets.

ΔK measures the error in the number of latent constructs, C-Jac. denotes the mean Jaccard similarity between matched indicator sets, Meas. F1 measures the recovery of latent–indicator assignments, Skel. F1 evaluates the recovery of skeleton edges in the latent–measurementgraph, and SHD measures the structural discrepancy between the estimated and ground-truth graphs. LLCD achieves a Skel. F1 of 0.949 and SHD of 1.6 on ADNI, and perfect structural recovery on PPMI. Compared with PureLLM, the main gain is in structural recovery, supporting the combination of LLM-based latent concept proposal with data-driven graph estimation.

Table 1: Latent clinical concepts identified by different LLMs on ADNI and PPMI, together with clinical expert ratings. * indicates concepts instantiated at both baseline and follow-up.

Dataset	Latent concept	Measurement indicators	LLM agreement	Construct plaus. ↑	Naming acc. ↑
ADNI	Amyloid pathology*	ABETA, Amyloid centiloids	3/3	**5**	**5**
	Tau pathology*	TAU, PTAU	3/3	4	**5**
	Neurodegeneration*	Hippocampus, Entorhinal, MidTemp, Fusiform, Ventricles	3/3	**5**	4
	Cognitive impairment*	ADAS13, MMSE	3/3	4	**5**
PPMI	Dopaminergic integrity	DAT caudate mean, DAT putamen mean	3/3	4	4
	Motor impairment*	NP2PTOT, NP3TOT, NHY	3/3	4	**5**
	Cognitive function*	MCATOT, SDMTOTAL, HVLT recall, semantic fluency, LNS	3/3	4	4
	Non-motor symptom burden*	NP1RTOT, NP1PTOT, SCOPA-AUT	2/3	**5**	**5**
	Non-motor symptom burden*	NP1RTOT, NP1PTOT	1/3	4	4

5.3 RQ2: Does LLCD Recover Better Latent-Aware Structures?

We next evaluate whether latent-aware representations improve causal structure recovery. On real-world data, where ground-truth graphs are unavailable, we report graph density and out-of-sample SCM fitting; on synthetic data, we additionally evaluate projected directed and skeleton F1 against the known graph. Specifically, because different methods may operate at the indicator, LLM-proposed latent, or ground-truth latent level, their estimated graphs are first projected into a common ground-truth semantic space: measurement indicators and matched latent concepts are mapped to their corresponding ground-truth latent variables, after which intra-concept and duplicate edges are

Table 2: Latent structure recovery on the semi-synthetic ADNI and PPMI benchmarks. Results are reported as mean ± standard deviation over five random seeds except for the deterministic methods.

	ADNI					PPMI				
Method	$\Delta K \downarrow$	C-Jac. ↑	Meas. F1 ↑	Skel. F1 ↑	SHD ↓	$\Delta K \downarrow$	C-Jac. ↑	Meas. F1 ↑	Skel. F1 ↑	SHD ↓
BPC+MIMBuild	4.0	0.385	0.683	0.500	9.0	2.0	0.600	0.778	0.571	6.0
FOFC+MIMBuild	4.0	0.479	0.791	0.500	9.0	2.0	0.500	0.703	0.714	4.0
TSC+MIMBuild	4.0	0.500	0.818	0.500	9.0	**0.0**	**1.000**	**1.000**	0.900	6.0
LGES	2.5±0.7	0.490±0.162	0.659±0.146	0.421±0.057	17.5±2.1	4.8±0.4	0.100±0.041	0.237±0.083	0.040±0.089	8.8±0.4
SALAD-CS	**0.0±0.0**	0.348±0.048	0.485±0.044	0.206±0.100	12.8±0.8	1.0±0.0	0.359±0.064	0.464±0.067	0.292±0.124	8.8±1.8
PureLLM	**0.0±0.0**	**0.875±0.000**	**0.917±0.000**	0.909±0.000	2.0±0.0	1.0±0.0	**1.000±0.000**	0.957±0.000	0.429±0.000	8.0±0.0
LLCD	**0.0±0.0**	**0.875±0.000**	**0.917±0.000**	**0.949±0.018**	**1.6±1.3**	**0.0±0.0**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**0.0±0.0**

Table 3: Performance comparison on real-world and synthetic datasets.

Method	Real-world						Synthetic					
	ADNI			PPMI			ADNI			PPMI		
	Density	R^2 ↑	NLL ↓	Density	R^2 ↑	NLL ↓	Proj. Dir. F1 ↑	Proj. Skel. F1 ↑	R^2 ↑	Proj. Dir. F1 ↑	Proj. Skel. F1 ↑	R^2 ↑
PC	0.111	0.310	1.123	0.129	0.165	1.278	0.359	0.791	0.236	<u>0.600</u>	<u>0.903</u>	0.331
FCI	0.103	0.166	1.254	–	–	–	0.143	0.791	0.372	0.270	<u>0.903</u>	0.337
GES	0.246	<u>0.344</u>	<u>1.095</u>	–	–	–	0.242	0.627	0.494	0.420	0.706	0.348
DirectLiNGAM	0.101	0.337	1.105	0.061	0.178	1.271	0.211	0.667	0.460	0.185	0.667	0.299
NOTEARS	0.061	0.332	1.110	0.052	<u>0.183</u>	<u>1.269</u>	0.218	0.655	<u>0.540</u>	0.125	0.688	<u>0.358</u>
CMA	0.201	0.225	1.248	0.152	0.095	1.327	0.517	0.565	0.331	0.390	0.422	0.142
CausalOrder	0.387	0.317	1.123	0.338	0.107	1.320	<u>0.526</u>	0.542	0.408	0.335	0.333	0.187
LLMprior	0.065	0.322	1.114	0.054	0.166	1.279	0.320	<u>0.792</u>	0.423	0.121	0.576	0.230
LLCD (Ours)	0.047	**0.418**	**1.074**	0.072	**0.507**	**0.793**	**0.873**	**0.918**	**0.557**	**0.981**	**0.981**	**0.567**

removed. Structural metrics are then computed in this common projected space. All methods use the same downstream linear-Gaussian SCM for fitting evaluation. Full projection and evaluation details are provided in Appendix A.2.2.

As shown in Table 3, LLCD achieves the best SCM fitting on both real-world datasets while retaining sparse graphs. More importantly, LLCD substantially improves projected directed structure recovery on synthetic ADNI and PPMI, achieving F1 scores of 0.873 and 0.981, compared with the best baseline scores of 0.526 and 0.600, respectively. It also achieves the highest downstream R^2 on both synthetic datasets.

5.4 RQ3: Do the Learned Structures Yield Better SCMs for Causal Analysis?

We further evaluate whether the learned structures support accurate interventional reasoning on synthetic data with known ground-truth SCMs. For ADNI, we intervene on APOE4 from 0 to 1 and evaluate 24 downstream outcomes; for PPMI, `INTERVAL_MONTHS` is changed from the 25th to 75th percentile (14 to 61 months), with 22 outcomes. Effects are estimated from $10{,}000$ samples per arm, and errors are normalized by the ground-truth factual standard deviation and macro-averaged across constructs to obtain EffectNMAE. For partially oriented graphs, we sample up to 100 consistent DAG extensions using randomized Dor–Tarsi elimination and report the EffectNMAE range. Further details are provided in Appendix A.2.3.

As shown in Table 4, LLCD achieves the lowest EffectNMAE on both datasets, with 0.0202 on ADNI and 0.0101 on PPMI. Compared with the second-best CausalOrder results of 0.0247 and 0.0130, LLCD reduces the effect estimation error by approximately 20% on both datasets. These results suggest that the latent-aware structures better preserve the propagation of interventions through the learned SCM.

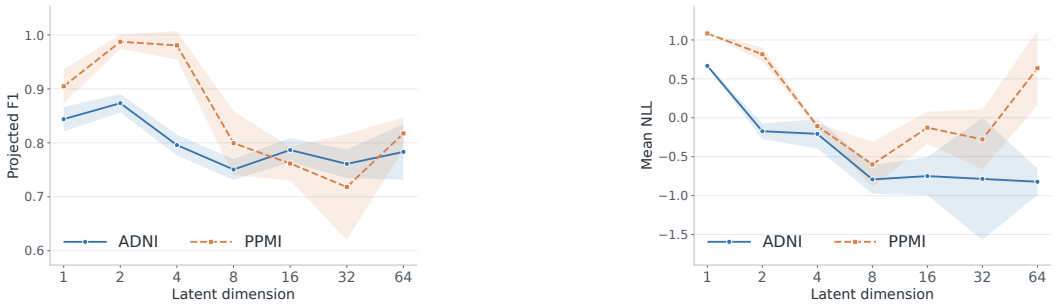
5.5 Ablation Study

Effect of latent dimension We study the effect of latent dimension in Fig. 2. ADNI achieves the best Projected F1 at $d = 2$, while PPMI shows similar structural recovery at $d = 2$ and $d = 4$ but better SCM fitting at $d = 4$. Overall, larger latent dimensions improve model capacity but do not consistently benefit structure recovery, revealing a trade-off between representation capacity and

Table 4: Effect estimation NMAE (↓).

Dataset	PC	FCI	GES	DirectLiNGAM	NOTEARS	CMA	CausalOrder	LLMprior	**LLCD (Ours)**
ADNI	0.1515–0.5366	0.5366	0.5366	0.5153	0.5366	0.0863	0.0247	0.2956	**0.0202**
PPMI	0.0760	0.1172	0.0166	0.1215	0.0764	0.0360	0.0130	0.0914	**0.0101**

structural identifiability. We therefore use $d = 2$ for ADNI and $d = 4$ for PPMI. Further analysis is provided in Appendix A.2.5.



(a) Projected Directed F1.

(b) Mean NLL.

Figure 2: Effect of latent dimension on causal structure recovery and downstream SCM fitting. Results are averaged over five random seeds, with error bars indicating one standard deviation.

Component ablation Table 5 evaluates the contributions of latent modeling and the LLM-based causal layer prior. Removing latent modeling degrades both downstream fitting and structure recovery: on real data, R^2 drops from 0.418 to 0.303 on ADNI and from 0.507 to 0.192 on PPMI, while semi-synthetic Projected F1 decreases from 0.873 to 0.835 and from 0.981 to 0.834, respectively. Removing the causal layer prior further reduces Projected F1 to 0.710 on ADNI and 0.699 on PPMI. These results show that latent modeling improves data representation, while the causal layer prior provides complementary structural guidance.

Table 5: Ablation study of latent modeling and the LLM-based causal layer prior ℓ. Results are reported as mean $\pm$ standard deviation over five random seeds.

Method	ADNI				PPMI			
	Real		Semi-synthetic		Real		Semi-synthetic	
	R^2 ↑	NLL↓	Proj. Dir. F1↑	NLL↓	R^2 ↑	NLL↓	Proj. Dir. F1↑	NLL↓
LLCD	**0.418±0.003**	**1.074±0.004**	**0.873±0.017**	**-0.173±0.101**	**0.507±0.017**	**0.793±0.030**	**0.981±0.027**	**-0.108±0.035**
w/o latent	0.303±0.000	1.131±0.000	0.835±0.019	1.042±0.004	0.192±0.000	1.263±0.000	0.834±0.052	1.243±0.004
w/o latent & ℓ	0.143±0.000	1.310±0.000	0.710±0.014	1.257±0.009	0.106±0.000	1.318±0.000	0.699±0.014	1.340±0.005

6 Conclusion

We presented LLCD, an LLM-guided framework that uses medical semantics to propose latent clinical concepts and observational data to validate their representations and causal relations. Experiments on semi-synthetic and real-world ADNI and PPMI data show improved latent-aware structure recovery and downstream SCM fitting. LLCD does not resolve unrestricted latent causal discovery from observational data. Its goal is narrower: when domain semantics suggest a set of plausible clinical abstractions, LLCD provides a data-driven mechanism for deciding which of them should participate in causal modeling. Thus LLCD is limited by its dependence on the candidate concepts and ordering priors, while quantitative causal evaluation remains restricted to semi-synthetic settings. Future work will explore more scalable representation search and broader validation with richer latent structures and real-world causal evidence.

REFERENCES

Ahmed Abdulaal, Adamos Hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel Castro, and Daniel Alexander. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 57559–57610, 2024.

Lyuzhou Chen, Xiangyu Wang, Taiyu Ban, Derui Lyu, Qinrui Zhu, Xin Wang, and Huanhuan Chen. Autonomous causal discovery: Evaluating llms' priors and constraint strategies for reliability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zhengming Chen, Ruichu Cai, Feng Xie, Jie Qiao, Anpeng Wu, Zijian Li, Zhifeng Hao, and Kun Zhang. Learning discrete latent variable structures with tensor rank conditions. *Advances in Neural Information Processing Systems*, 37:17398–17427, 2024.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Haoyue Dai, Yiwen Qiu, Ignavier Ng, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Latent variable causal discovery under selection bias. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12161–12178. PMLR, 13–19 Jul 2025.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. Large language models are effective priors for causal graph discovery. *arXiv preprint arXiv:2405.13551*, 2024.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *International Conference on Learning Representations*, volume 2024, pp. 43084–43118, 2024.

Xinshuai Dong, Ignavier Ng, Haoyue Dai, Jiaqi Sun, Xiangchen Song, Peter Spirtes, and Kun Zhang. Score-based greedy search for structure identification of partially observed causal models. In *International Conference on Learning Representations*, volume 2026, pp. 9900–9926, 2026.

Huaming Du, Yujia Zheng, Baoyu Jing, Yu Zhao, Gang Kou, Guisong Liu, Tao Gu, Weimin Li, and Carl Yang. Causal discovery through synergizing large language model and data-driven reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 543–554, 2025.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1655–1664, 2016.

Xiuchuan Li and Tongliang Liu. Efficient and trustworthy causal discovery with latent variables and complex relations. In *International Conference on Learning Representations*, volume 2025, pp. 34349–34383, 2025.

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*, 2023.

Kenneth Marek, Sohini Chowdhury, Andrew Siderowf, Shirley Lasch, Christopher S. Coffey, Chelsea Caspell-Garcia, Tanya Simuni, Danna Jennings, Caroline M. Tanner, John Q. Trojanowski, Leslie M. Shaw, John Seibyl, Norbert Schuff, Andrew Singleton, Karl Kieburtz, Arthur W. Toga, Brit Mollenhauer, Douglas Galasko, Lana M. Chahine, Daniel Weintraub, Tatiana Foroud, Duygu Tosun-Turgut, Kathleen Poston, Vanessa Arnedo, Mark Frasier, Todd Sherer, and the Parkinson Progression Marker Initiative. The parkinson progression marker initiative (ppmi) – establishing a pd biomarker cohort. *Annals of Clinical and Translational Neurology*, 5 (12):1460–1477, 2018. doi: 10.1002/acn3.644.

Ignavier Ng, Xinshuai Dong, Haoyue Dai, Biwei Huang, Peter Spirtes, and Kun Zhang. Score-based causal discovery of latent variable causal models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37531–37552. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ng24a.html.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, CR Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3): 201–209, 2010.

Parjanya Prashant, Ignavier Ng, Kun Zhang, and Biwei Huang. Differentiable causal discovery for latent hierarchical causal models. In *International Conference on Learning Representations*, volume 2025, pp. 31850–31875, 2025.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7 (2), 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. The MIT press, 2001.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for gaussian graphical models. 2010.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, Eran Segal, Eric P Xing, and Kun Zhang. Causal representation learning from multimodal biomedical observations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 18536–18568, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2ec91b31ac4aa578d99309d4f00c81ad-Paper-Conference.pdf.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *International Conference on Learning Representations*, volume 2025, pp. 95330–95371, 2025.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A Appendix

A.1 Dataset

A.1.1 Source Dataset

We evaluate our method on two longitudinal medical cohorts, the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Parkinson's Progression Markers Initiative (PPMI). For both datasets, we construct longitudinal samples by pairing two temporally ordered visits from the same subject, with the earlier visit treated as baseline and the later visit as follow-up. All train/validation/test splits are performed at the subject level to prevent information leakage across splits.

ADNI The ADNI cohort is used to characterize longitudinal progression of Alzheimer's disease. After preprocessing, the dataset contains 2,196 subjects and 28,617 longitudinal visit pairs, with at most 20 pairs retained per subject. We split the data into 1,542/323/331 subjects for training, validation, and testing, corresponding to 19,968/4,222/4,427 visit pairs, respectively. The median interval between paired visits is 1.566 years. The variables cover demographic and risk factors, Alzheimer's disease biomarkers, structural MRI measurements, and cognitive and diagnostic assessments. Specifically, we use five baseline covariates, 13 longitudinal measurements retained at both baseline and follow-up, and the follow-up interval, resulting in 32 observed variables. These measurements include amyloid and tau biomarkers, regional structural MRI volumes, and cognitive and clinical assessments.

PPMI The PPMI cohort is used to characterize longitudinal progression of Parkinson's disease. The processed dataset contains 1,440 subjects and 90,689 longitudinal visit pairs. We split the data into 1,008/216/216 subjects for training, validation, and testing, corresponding to 63,154/13,578/13,957 visit pairs, respectively. The median interval between paired visits is 31.967 months. The 38 observed variables cover demographic and genetic risk factors, DaTScan-derived dopaminergic imaging measurements, medication-related variables, motor and non-motor symptoms, and cognitive assessments. Measurements from the earlier and later visits are retained separately to represent baseline and follow-up disease states.

Table 6 and Table 7 summarize the observed variables used in the real-data experiments. Variables marked with † are available only at baseline, while ‡ denotes the follow-up interval. For longitudinal measurements, baseline and follow-up versions are listed in the same row. No manually defined variable grouping or latent-concept annotation is provided to the LLM.

Missing Data and Preprocessing. Both cohorts contain naturally missing clinical measurements, particularly for imaging and biomarker variables. We preserve the original missingness rather than applying global imputation. All variables are standardized using the mean and standard deviation estimated from non-missing observations in the training set, while missing entries remain unobserved. During score-based causal structure learning, each local model for a target variable and its candidate parent set is fitted using only samples in which all involved variables are observed. Candidate local structures with fewer than 50 valid samples are excluded. To reduce score-scale differences caused by varying numbers of available observations, we normalize the linear-Gaussian BIC local score by the corresponding number of valid samples. The same preprocessing and missing-data handling strategy is used throughout the real-data experiments.

A.1.2 SYNTHETIC DATASET

To quantitatively evaluate latent-variable recovery and causal structure discovery against known ground truth, we construct ADNI-like and PPMI-like synthetic datasets with predefined latent variables and causal graphs. The synthetic generator is deliberately separated from the VAE-based representation model evaluated in our method: latent ground truth is generated from a classical measurement model and a predefined structural model, avoiding circular evaluation in which the same latent model is used for both data generation and recovery.

Synthetic Data Generation For each latent measurement block K, we first estimate a one-dimensional classical measurement model using only the training split of the corresponding real dataset. Let X_{Kj} denote the j-th observed indicator and let $\widetilde{X}_{Kj}$ denote its standardized value using training-set statistics. We estimate a one-dimensional factor direction $\widehat{L}_K$ by SVD and define the loading of each indicator as

$$\lambda_{Kj} = \mathrm{Corr}\left(\widehat{L}_K, \widetilde{X}_{Kj}\right). \tag{23}$$

The corresponding measurement model is

$$\widetilde{X}_{Kj} = \lambda_{Kj} L_K + s_{Kj}\epsilon_{Kj}, \qquad \epsilon_{Kj} \sim \mathcal{N}(0,1), \tag{24}$$

where

$$s_{Kj} = \sqrt{\max\left(1 - \lambda_{Kj}^2,\, 0.03\right)}. \tag{25}$$

Table 6: Observed variables used for ADNI. Suffixes _bl and _fu denote baseline and follow-up measurements, respectively. † indicates a baseline-only variable and ‡ denotes the follow-up interval.

Variable	Time point	Description
Age†	Baseline	Age at the baseline visit.
Sex†	Baseline	Biological sex.
PTEDUCAT†	Baseline	Years of education.
APOE4†	Baseline	APOE $\varepsilon 4$ carrier/count indicator.
Hypertension†	Baseline	Hypertension status.
ABETA_bl / ABETA_fu	Baseline / follow-up	Cerebrospinal-fluid amyloid-β level.
Amyloid_centiloids_bl / _fu	Baseline / follow-up	Amyloid PET burden on the Centiloid scale.
TAU_bl / TAU_fu	Baseline / follow-up	Cerebrospinal-fluid total tau level.
PTAU_bl / PTAU_fu	Baseline / follow-up	Cerebrospinal-fluid phosphorylated tau level.
Ventricles_bl / _fu	Baseline / follow-up	Ventricular volume derived from structural MRI.
Hippocampus_bl / _fu	Baseline / follow-up	Hippocampal volume derived from structural MRI.
Entorhinal_bl / _fu	Baseline / follow-up	Entorhinal cortical volume derived from structural MRI.
Fusiform_bl / _fu	Baseline / follow-up	Fusiform volume derived from structural MRI.
MidTemp_bl / _fu	Baseline / follow-up	Middle temporal volume derived from structural MRI.
ICV_bl / ICV_fu	Baseline / follow-up	Estimated intracranial volume.
ADAS13_bl / ADAS13_fu	Baseline / follow-up	ADAS-Cog 13 score; higher values generally indicate greater cognitive impairment.
MMSE_bl / MMSE_fu	Baseline / follow-up	Mini-Mental State Examination score; lower values generally indicate greater cognitive impairment.
DX_bl / DX_fu	Baseline / follow-up	Clinical diagnostic category.
interval_years‡	Interval	Time between the baseline and follow-up visits in years.

For PPMI, the measurement equation of NP3TOT additionally includes the observed medication ON/OFF state as a measurement-context variable.

We reuse demographic, genetic, treatment, and follow-up-interval variables from the real longitudinal pairs as exogenous templates, with missing template values median-imputed before simulation. Given a predefined ground-truth DAG G^*, each latent or structural variable is generated as a noisy linear function of its parents,

$$L_j = \sum_{i \in \mathrm{Pa}_{G^*}(j)} a_{ij} Z_i + \eta_j, \qquad \eta_j \sim \mathcal{N}(0, \sigma_j^2), \tag{26}$$

followed by standardization. The latent variables are then projected to their observed indicators using Eq. 24.

To obtain the final synthetic benchmark, we further fit an SCM on the complete generated data using the known latent variables and the known GT graph. All graph variables are normalized using training-set statistics,

$$z_j = \frac{x_j - \mu_{j,\mathrm{tr}}}{\sigma_{j,\mathrm{tr}}}, \tag{27}$$

and the joint distribution is factorized according to the GT DAG:

$$p(\mathbf{z}) = \prod_j p\left(z_j \mid \mathbf{z}_{\mathrm{Pa}_{G^*}(j)}\right). \tag{28}$$

Table 7: Observed variables used for PPMI. Suffixes _t and _fu denote the earlier and later visits, respectively. † indicates a baseline-only variable and ‡ denotes the follow-up interval.

Variable	Time point	Description
AGE_AT_VISIT_t†	Baseline	Age at the earlier visit.
SEX†	Baseline	Biological sex.
EDUCYRS†	Baseline	Years of education.
GBA_CARRIER†	Baseline	Indicator of a GBA genetic variant.
LRRK2_CARRIER†	Baseline	Indicator of an LRRK2 genetic variant.
OTHER_PD_GENE_CARRIER†	Baseline	Indicator of other Parkinson's disease-related genetic variants.
APOE4_CARRIER†	Baseline	APOE $\varepsilon 4$ carrier indicator.
DAT_CAUDATE_MEAN†	Baseline	Mean DaTScan striatal binding ratio in the caudate.
DAT_PUTAMEN_MEAN†	Baseline	Mean DaTScan striatal binding ratio in the putamen.
DAT_PUTAMEN_ASYMMETRY†	Baseline	Asymmetry of putamen DaTScan binding.
LEDD_DOSE_t†	Baseline	Levodopa equivalent daily dose at baseline.
NP3TOT_ON_OFF_t / _fu	Baseline / follow-up	Medication ON/OFF state during the MDS-UPDRS III assessment.
NP2PTOT_t / NP2PTOT_fu	Baseline / follow-up	MDS-UPDRS Part II score for motor experiences of daily living.
NP3TOT_t / NP3TOT_fu	Baseline / follow-up	MDS-UPDRS Part III motor examination score.
NHY_t / NHY_fu	Baseline / follow-up	Hoehn–Yahr disease stage.
NP4TOT_t / NP4TOT_fu	Baseline / follow-up	MDS-UPDRS Part IV motor-complication score.
NP1RTOT_t / NP1RTOT_fu	Baseline / follow-up	Rater-assessed MDS-UPDRS Part I non-motor score.
NP1PTOT_t / NP1PTOT_fu	Baseline / follow-up	Patient-reported MDS-UPDRS Part I non-motor score.
SCOPA_AUT_TOTAL_t / _fu	Baseline / follow-up	SCOPA-AUT total score for autonomic symptoms.
MCATOT_t / MCATOT_fu	Baseline / follow-up	Montreal Cognitive Assessment (MoCA) total score.
SDMTOTAL_t / SDMTOTAL_fu	Baseline / follow-up	Symbol Digit Modalities Test score reflecting processing speed and attention.
HVLT_RECALL_t / _fu	Baseline / follow-up	Hopkins Verbal Learning Test immediate-recall score.
SEM_FLUENCY_t / _fu	Baseline / follow-up	Semantic fluency score based on animal naming.
LNS_t / LNS_fu	Baseline / follow-up	Letter–Number Sequencing score reflecting working memory.
INTERVAL_MONTHS‡	Interval	Time between the earlier and later visits in months.

For each non-root node, we fit a conditional affine-Gaussian model

$$z_j \mid \mathbf{z}_{\mathrm{Pa}(j)} \sim \mathcal{N}\left(\beta_{j0} + \boldsymbol{\beta}_j^\top \mathbf{z}_{\mathrm{Pa}(j)}, \sigma_j^2\right), \tag{29}$$

where the conditional mean is estimated using ridge regression with $\lambda = 10^{-4}$. The final dataset is generated by ancestral sampling in the topological order of G^*:

$$z_j = \widehat{\beta}_{j0} + \widehat{\boldsymbol{\beta}}_j^\top \mathbf{z}_{\mathrm{Pa}(j)} + \widehat{\sigma}_j \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0,1). \tag{30}$$

This GT-SCM fitting step is used only to construct the synthetic benchmark and does not perform latent-variable or causal-structure discovery.

Ground-Truth Causal Structures The two synthetic datasets use disease-specific latent measurement blocks and ground-truth causal structures. Table 8 summarizes the predefined latent constructs and their observed indicators. Tables 9 and 10 list the complete ground-truth edges used for ADNI

Table 8: Ground-truth latent measurement blocks used in the synthetic datasets. The same measurement block is instantiated separately at baseline and follow-up.

Dataset	Latent construct	Observed indicators
ADNI	Amyloid	ABETA, Amyloid_centiloids
	Tau	TAU, PTAU
	Neurodegeneration	Ventricles, Hippocampus, Entorhinal, Fusiform, MidTemp, ICV
	Cognition	ADAS13, MMSE, DX
PPMI	Motor severity	NP2PTOT, NP3TOT, NHY
	Non-motor burden	NP1RTOT, NP1PTOT, SCOPA_AUT_TOTAL
	Cognition	MCATOT, SDMTOTAL, HVLT_RECALL, SEM_FLUENCY, LNS

Table 9: Ground-truth edges used in the ADNI synthetic dataset. Braces indicate multiple child nodes sharing the same parent.

Edge type	Ground-truth edges
Baseline structure	$\mathrm{APOE4} \to \mathrm{abeta}_{\mathrm{bl}}$;
	$\mathrm{abeta}_{\mathrm{bl}} \to \mathrm{tau}_{\mathrm{bl}}$;
	$\mathrm{tau}_{\mathrm{bl}} \to \mathrm{neurodegeneration}_{\mathrm{bl}}$;
	$\mathrm{neurodegeneration}_{\mathrm{bl}} \to \mathrm{cognition}_{\mathrm{bl}}$
Temporal persistence	$\mathrm{abeta}_{\mathrm{bl}} \to \mathrm{abeta}_{\mathrm{fu}}$;
	$\mathrm{tau}_{\mathrm{bl}} \to \mathrm{tau}_{\mathrm{fu}}$;
	$\mathrm{neurodegeneration}_{\mathrm{bl}} \to \mathrm{neurodegeneration}_{\mathrm{fu}}$;
	$\mathrm{cognition}_{\mathrm{bl}} \to \mathrm{cognition}_{\mathrm{fu}}$
Follow-up structure	$\mathrm{abeta}_{\mathrm{fu}} \to \mathrm{tau}_{\mathrm{fu}}$;
	$\mathrm{tau}_{\mathrm{fu}} \to \mathrm{neurodegeneration}_{\mathrm{fu}}$;
	$\mathrm{neurodegeneration}_{\mathrm{fu}} \to \mathrm{cognition}_{\mathrm{fu}}$
Interval effects	$\mathrm{interval_years} \to \mathrm{abeta}_{\mathrm{fu}}$;
	$\mathrm{interval_years} \to \mathrm{tau}_{\mathrm{fu}}$;
	$\mathrm{interval_years} \to \mathrm{neurodegeneration}_{\mathrm{fu}}$;
	$\mathrm{interval_years} \to \mathrm{cognition}_{\mathrm{fu}}$
Cross-time effects	$\mathrm{abeta}_{\mathrm{bl}} \to \mathrm{neurodegeneration}_{\mathrm{fu}}$;
	$\mathrm{tau}_{\mathrm{bl}} \to \mathrm{neurodegeneration}_{\mathrm{fu}}$
Hypertension effects	$\mathrm{Hypertension} \to \mathrm{neurodegeneration}_{\mathrm{bl}}$;
	$\mathrm{Hypertension} \to \mathrm{neurodegeneration}_{\mathrm{fu}}$;
	$\mathrm{Hypertension} \to \mathrm{cognition}_{\mathrm{bl}}$;
	$\mathrm{Hypertension} \to \mathrm{cognition}_{\mathrm{fu}}$
Measurement	$\mathrm{abeta}_s \to \{\mathrm{ABETA}_s, \mathrm{Amyloid_centiloids}_s\}$;
	$\mathrm{tau}_s \to \{\mathrm{TAU}_s, \mathrm{PTAU}_s\}$;
	$\mathrm{neurodegeneration}_s \to \{\mathrm{Ventricles}_s, \mathrm{Hippocampus}_s, \mathrm{Entorhinal}_s\}$;
	$\mathrm{neurodegeneration}_s \to \{\mathrm{Fusiform}_s, \mathrm{MidTemp}_s, \mathrm{ICV}_s\}$;
	$\mathrm{cognition}_s \to \{\mathrm{ADAS13}_s, \mathrm{MMSE}_s, \mathrm{DX}_s\}, \quad s \in \{\mathrm{bl}, \mathrm{fu}\}$

and PPMI, respectively. For ADNI, we define four latent construct types—amyloid, tau, neurodegeneration, and cognition—each instantiated at both baseline and follow-up, resulting in 8 latent nodes. Its ground-truth graph contains 47 edges, comprising 21 structural edges and 26 measurement edges. For PPMI, three latent construct types—motor severity, non-motor burden, and cognition—are instantiated at both time points, resulting in 6 latent nodes. Its ground-truth graph contains 56 edges, including 32 structural edges, 22 measurement edges, and 2 measurement-context edges.

GT-SCM Fitting Quality Table 11 reports the test-set fitting quality of the GT-SCM used for synthetic sampling. We report statistics over non-root conditional nodes only, since root nodes have no predictors and their R^2 is therefore not informative. Overall, the fitted conditional models closely reproduce the generated GT distributions in both datasets.

Table 10: Ground-truth edges used in the PPMI synthetic dataset. Braces indicate multiple child nodes sharing the same parent.

Edge type	Ground-truth edges
Baseline motor severity	$\mathrm{AGE_AT_VISIT}_t \to \mathrm{motor_severity}_t$; $\mathrm{SEX} \to \mathrm{motor_severity}_t$; $\mathrm{GBA_CARRIER} \to \mathrm{motor_severity}_t$; $\mathrm{LRRK2_CARRIER} \to \mathrm{motor_severity}_t$; $\mathrm{OTHER_PD_GENE_CARRIER} \to \mathrm{motor_severity}_t$; $\mathrm{DAT_CAUDATE_MEAN} \to \mathrm{motor_severity}_t$; $\mathrm{DAT_PUTAMEN_MEAN} \to \mathrm{motor_severity}_t$; $\mathrm{DAT_PUTAMEN_ASYMMETRY} \to \mathrm{motor_severity}_t$
Follow-up motor severity	$\mathrm{motor_severity}_t \to \mathrm{motor_severity}_{\mathrm{fu}}$; $\mathrm{INTERVAL_MONTHS} \to \mathrm{motor_severity}_{\mathrm{fu}}$; $\mathrm{LEDD_DOSE}_t \to \mathrm{motor_severity}_{\mathrm{fu}}$
Baseline non-motor burden	$\mathrm{AGE_AT_VISIT}_t \to \mathrm{nonmotor_burden}_t$; $\mathrm{SEX} \to \mathrm{nonmotor_burden}_t$; $\mathrm{motor_severity}_t \to \mathrm{nonmotor_burden}_t$
Follow-up non-motor burden	$\mathrm{nonmotor_burden}_t \to \mathrm{nonmotor_burden}_{\mathrm{fu}}$; $\mathrm{INTERVAL_MONTHS} \to \mathrm{nonmotor_burden}_{\mathrm{fu}}$; $\mathrm{motor_severity}_t \to \mathrm{nonmotor_burden}_{\mathrm{fu}}$
Baseline cognition	$\mathrm{AGE_AT_VISIT}_t \to \mathrm{cognition}_t$; $\mathrm{EDUCYRS} \to \mathrm{cognition}_t$; $\mathrm{APOE4_CARRIER} \to \mathrm{cognition}_t$; $\mathrm{motor_severity}_t \to \mathrm{cognition}_t$; $\mathrm{nonmotor_burden}_t \to \mathrm{cognition}_t$
Follow-up cognition	$\mathrm{cognition}_t \to \mathrm{cognition}_{\mathrm{fu}}$; $\mathrm{INTERVAL_MONTHS} \to \mathrm{cognition}_{\mathrm{fu}}$; $\mathrm{motor_severity}_t \to \mathrm{cognition}_{\mathrm{fu}}$; $\mathrm{nonmotor_burden}_t \to \mathrm{cognition}_{\mathrm{fu}}$
Motor complications	$\mathrm{AGE_AT_VISIT}_t \to \mathrm{NP4TOT}_t$; $\mathrm{motor_severity}_t \to \mathrm{NP4TOT}_t$; $\mathrm{NP4TOT}_t \to \mathrm{NP4TOT}_{\mathrm{fu}}$; $\mathrm{INTERVAL_MONTHS} \to \mathrm{NP4TOT}_{\mathrm{fu}}$; $\mathrm{LEDD_DOSE}_t \to \mathrm{NP4TOT}_{\mathrm{fu}}$; $\mathrm{motor_severity}_t \to \mathrm{NP4TOT}_{\mathrm{fu}}$
Motor measurements	$\mathrm{motor_severity}_s \to \mathrm{NP2PTOT}_s$; $\mathrm{motor_severity}_s \to \mathrm{NP3TOT}_s$; $\mathrm{motor_severity}_s \to \mathrm{NHY}_s, \quad s \in \{t, \mathrm{fu}\}$
Non-motor measurements	$\mathrm{nonmotor_burden}_s \to \mathrm{NP1RTOT}_s$; $\mathrm{nonmotor_burden}_s \to \mathrm{NP1PTOT}_s$; $\mathrm{nonmotor_burden}_s \to \mathrm{SCOPA_AUT_TOTAL}_s, \quad s \in \{t, \mathrm{fu}\}$
Cognitive measurements	$\mathrm{cognition}_s \to \mathrm{MCATOT}_s$; $\mathrm{cognition}_s \to \mathrm{SDMTOTAL}_s$; $\mathrm{cognition}_s \to \mathrm{HVLT_RECALL}_s$; $\mathrm{cognition}_s \to \mathrm{SEM_FLUENCY}_s$; $\mathrm{cognition}_s \to \mathrm{LNS}_s, \quad s \in \{t, \mathrm{fu}\}$
Measurement context	$\mathrm{NP3TOT_ON_OFF}_t \to \mathrm{NP3TOT}_t$; $\mathrm{NP3TOT_ON_OFF}_{\mathrm{fu}} \to \mathrm{NP3TOT}_{\mathrm{fu}}$

Table 11: Test-set fitting quality of the GT-SCM. Metrics are averaged over non-root conditional nodes. Lower NLL and normalized RMSE are better, while higher R^2 is better.

Dataset	NLL ↓	Norm. RMSE ↓	Mean R^2 ↑	Median R^2 ↑
ADNI	0.703	0.541	0.657	0.731
PPMI	0.971	0.644	0.555	0.588

A.2 EXPERIMENTS

A.2.1 TRAINING SETTING

For the latent representation model, we use dim=2 latent for ADNI and dim=4 latent for PPMI, with two-layer MLPs of hidden width 64. Models are trained for 500 epochs with AdamW, batch size 256, learning rate 10^{-3}, and weight decay 10^{-4}. In Eq. 12, we set $\lambda_{\mathrm{rec}} = 1$, $\lambda_{\mathrm{ctx}} = 0.05$, $\beta = 10^{-3}$, and $\lambda_{\mathrm{scale}} = 10^{-3}$. For causal discovery, we set the minimum accepted BIC improvement to 10^{-3} and require at least 50 valid samples for each local fit. In global representation selection, we use $\epsilon = 0.005$ and $\lambda_Z = 1$. Baseline and follow-up instances of the same longitudinal latent construct are selected jointly.

A.2.2 EVALUATION DETAILS FOR RQ2

Evaluation protocol RQ2 evaluates causal discovery from two complementary perspectives. First, for synthetic datasets with known generating graphs, we directly evaluate whether each method recovers the underlying causal structure. Second, for both real-world and synthetic datasets, we evaluate whether the discovered graph provides an effective structural factorization for unseen samples by fitting the same downstream SCM to every method. This distinction is important because real-world medical datasets do not provide observable ground-truth causal graphs; consequently, graph-recovery quantities such as precision, recall, or SHD are not reported for real data. Instead, real-world evaluation uses graph complexity and test-set SCM fitting quality.

The experiments use longitudinal ADNI and PPMI cohorts together with their corresponding synthetic datasets. Each sample represents a temporally ordered baseline–follow-up pair and contains time-invariant covariates, baseline variables, follow-up variables, and the interval between the two visits. For cross-method comparability, evaluation is restricted to a fixed common observed-node set containing 32 variables for ADNI and 38 variables for PPMI. Subject identifiers are used only for organizing samples and are removed before causal discovery and evaluation. Variables outside the common node set, self-loops, and duplicated relations are also removed from baseline outputs before computing the reported metrics.

Each synthetic dataset contains $5{,}000$ samples. For baselines without a predefined split, we construct training, validation, and test sets using a fixed random seed of 35001 and a $70\%/15\%/15\%$ probabilistic split. Structural recovery is evaluated directly against the same generating graph and is therefore unaffected by this difference in sample partitioning. Downstream SCM metrics are interpreted with the corresponding held-out test split.

Normalization and projection of discovered graphs Because different methods can output DAGs, CPDAGs, PAGs, weighted adjacency matrices, or graphs containing method-specific latent variables, all outputs are normalized before evaluation. We first remove subject identifiers, self-loops, variables outside the common evaluation space, and duplicate relations. Explicitly directed edges and ordinary edges without a special edge-type annotation are treated as directed. Undirected, bidirected, and partially oriented relations are preserved for graph-complexity and skeleton evaluation but are not arbitrarily assigned a causal direction and are therefore not used as parent–child relations in downstream SCM fitting.

LLCD produces both structural edges and measurement edges from latent concepts to their observed indicators. Both types are retained when constructing the downstream SCM representation, whereas synthetic structural-recovery evaluation uses only the structural component. The directed subgraph is additionally checked for acyclicity by topological elimination. If a baseline output contains a directed cycle, its node-wise conditional models can still be evaluated independently, but the resulting collection is not interpreted as a valid joint generative DAG-SCM.

Synthetic graph comparison requires an additional normalization step because different methods may operate at the observed-indicator level, the LLM-proposed latent-concept level, or the generating latent-variable level. We therefore define a semantic projection

$$\pi : V_{\mathrm{pred}} \to V_{\mathrm{GT}}, \tag{31}$$

which maps each measurement indicator to its corresponding ground-truth latent variable. An LLM-proposed latent concept is matched to the ground-truth latent variable having the largest overlap in associated indicators, while ordinary variables without a latent counterpart remain unchanged.

Each predicted directed edge $u \to v$ is transformed into $\pi(u) \to \pi(v)$. Edges whose endpoints collapse to the same projected node are removed because they represent within-concept measurement relations rather than concept-level structural relations, and duplicated projected edges are merged. If both $a \to b$ and $b \to a$ are produced after projection, neither orientation is counted as successfully identified. Ground-truth graphs are processed using the same node space and retain only structural edges.

For projected directed edge sets $\widehat{E}$ and E^*, we compute

$$TP = |\widehat{E} \cap E^*|, \qquad FP = |\widehat{E} \setminus E^*|, \qquad FN = |E^* \setminus \widehat{E}|, \tag{32}$$

followed by

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}. \tag{33}$$

Skeleton evaluation ignores edge orientation and instead compares unordered node pairs after applying the same projection. This prevents algorithms that return partially oriented equivalence-class graphs from being unnecessarily penalized when the adjacency itself is correctly recovered. The main text reports directed F1 and skeleton F1 as compact summaries of these two aspects of recovery.

Unified downstream SCM evaluation To isolate the effect of graph structure from differences in downstream model capacity, every discovered graph is evaluated using the same node-wise linear-Gaussian SCM. For an observed target variable X_i, its parent set is formed from the explicitly directed incoming edges of the corresponding method, and the structural equation is

$$X_i = \beta_{i0} + \boldsymbol{\beta}_i^\top \mathbf{X}_{\mathrm{Pa}(i)} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2). \tag{34}$$

A node without parents reduces to a training-set mean model. When LLCD introduces a multidimensional latent parent, all posterior-mean dimensions of that concept are included in the regression design matrix, while evaluation targets remain the observed variables in the common node space.

All preprocessing statistics are estimated exclusively from the training set. Non-numeric values are treated as missing, missing values are imputed using training-set medians, and variables are standardized using the training mean and population standard deviation. Near-zero standard deviations are replaced by one for numerical stability. Ordinary least squares is used to estimate the regression coefficients, while the Gaussian noise variance is estimated from the mean squared training residual with a numerical lower bound of 10^{-8}. The resulting imputation values, normalization statistics, coefficients, and noise variances are then fixed and applied to the held-out test data.

For target node i, test-set predictive fit is measured by

$$R_i^2 = 1 - \frac{\sum_n (x_{ni} - \widehat{x}_{ni})^2}{\sum_n (x_{ni} - \bar{x}_i)^2}, \tag{35}$$

and Gaussian negative log-likelihood is computed as

$$\mathrm{NLL}_i = \frac{1}{2N} \sum_n \left[\log(2\pi\sigma_i^2) + \frac{(x_{ni} - \widehat{x}_{ni})^2}{\sigma_i^2} \right]. \tag{36}$$

Metrics are macro-averaged across target nodes so that each variable contributes equally. Higher R^2 and lower NLL indicate better out-of-sample conditional fit. These metrics evaluate how well the discovered structural factorization supports unseen data; they should not by themselves be interpreted as evidence that every discovered edge corresponds to an identified causal effect.

Graph complexity Graph sparsity is evaluated after normalization to the common node space. Let p denote the number of evaluated nodes, m_s the number of structural relations, and m_m the number of LLCD measurement relations. We compute structural and total graph complexity as

$$\text{Density}_{\text{struct}} = \frac{m_s}{p(p-1)/2}, \qquad \text{Density}_{\text{total}} = \frac{m_s + m_m}{p(p-1)}. \tag{37}$$

The corresponding sparsity is one minus the relevant density. Structural and total density use different normalizing conventions and are therefore not treated as interchangeable quantities. The density reported in the main comparison is used primarily to verify that improvements in downstream fitting are not obtained simply by constructing increasingly dense graphs.

A.2.3 Evaluation Details for RQ3

Intervention queries RQ3 evaluates whether an intervention propagated through a learned SCM reproduces the average intervention effect of the ground-truth SCM. For every method, the intervention variable, intervention values, and evaluated outcomes are fixed in advance and are independent of the learned graph. For ADNI, we evaluate the change from

$$do(\text{APOE4} = 0) \quad \text{to} \quad do(\text{APOE4} = 1), \tag{38}$$

with 24 downstream measurements grouped into eight ground-truth latent constructs. For PPMI, we intervene on `INTERVAL_MONTHS` from its ground-truth factual 25th percentile to its 75th percentile,

$$do(X = Q_{25}) \quad \text{to} \quad do(X = Q_{75}), \tag{39}$$

and evaluate 22 measurements grouped into six constructs. Outcomes are retained even when a learned graph contains no directed path from the intervention to that outcome, in which case the corresponding predicted effect is typically zero.

For an outcome Y, the ground-truth and predicted average intervention effects are defined as

$$\Delta_Y^{\mathrm{GT}} = \mathbb{E}_{\mathcal{M}_{\mathrm{GT}}}[Y \mid do(X = x_H)] - \mathbb{E}_{\mathcal{M}_{\mathrm{GT}}}[Y \mid do(X = x_L)], \tag{40}$$

and

$$\widehat{\Delta}_Y = \mathbb{E}_{\widehat{\mathcal{M}}}[Y \mid do(X = x_H)] - \mathbb{E}_{\widehat{\mathcal{M}}}[Y \mid do(X = x_L)]. \tag{41}$$

Because the ground-truth SCM is known, the same reference effects are used for all methods.

EffectNMAE To compare outcomes with different measurement scales, the absolute intervention-effect error is normalized by the factual standard deviation of the corresponding outcome under the ground-truth system:

$$e_Y = \frac{\left|\widehat{\Delta}_Y - \Delta_Y^{\mathrm{GT}}\right|}{\sigma_Y^{\mathrm{GT}}}. \tag{42}$$

Let $\mathcal{I}_k^{\mathrm{GT}}$ denote the observed indicators belonging to ground-truth construct k. We first average the normalized errors within each construct,

$$\mathrm{EffectNMAE}(k) = \frac{1}{|\mathcal{I}_k^{\mathrm{GT}}|} \sum_{Y \in \mathcal{I}_k^{\mathrm{GT}}} e_Y, \tag{43}$$

and then macro-average across the K constructs:

$$\mathrm{EffectNMAE} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{EffectNMAE}(k). \tag{44}$$

This two-stage averaging prevents constructs containing more observed indicators from dominating the final score. Lower values indicate more accurate recovery of the ground-truth intervention effects.

Executable SCM evaluation For each learned directed graph, we fit the same node-wise linear-Gaussian SCM using the corresponding synthetic training data. Missing values are imputed using training-set statistics and variables are standardized before fitting. The intervention is implemented by fixing the treatment node to the specified value and propagating its effect through the learned directed structure. For linear-Gaussian models, mean intervention effects can be obtained through recursive path propagation; equivalent Monte Carlo propagation is used when required by the latent representation.

For methods that return CPDAGs or PDAGs with unresolved edge orientations, we construct executable DAGs using randomized Dor–Tarsi elimination rather than assigning directions from ground-truth information. At each run, nodes are assigned a random priority and eliminated only when their orientation is consistent with the current partially directed graph; the resulting extension is retained only if it is acyclic and consistent with the original CPDAG/PDAG. Duplicate DAGs

are removed, and a fixed random seed is used for reproducibility. We collect at most 100 unique extensions and stop earlier if no additional unique extensions are found after at most

$$\max(200, 100 \times 50) = 5000 \tag{45}$$

randomized attempts. This procedure samples admissible DAG extensions and does not guarantee exhaustive enumeration of the equivalence class. In the current experiments, 100 unique extensions are obtained for ADNI PC, whereas 4 and 5 unique extensions are obtained for PPMI PC and GES, respectively. The latter sets may cover all consistent extensions in practice, but exhaustive completeness is not assumed.

The resulting EffectNMAE measures recovery of the average intervention effect for the predefined queries rather than correctness of the entire interventional or counterfactual distribution. The evaluation is therefore intended to test whether the learned causal structures yield executable SCMs that preserve downstream effect propagation under controlled synthetic interventions.

A.2.4 LLM PROMPTS

The same prompt templates are used for ADNI and PPMI. Dataset-specific information, including variable definitions, temporal conventions, and candidate nodes, is inserted into the corresponding placeholders.

Latent Construct Proposal The generic system prompt is:

```
You are a biomedical causal discovery assistant. Your task is to infer which latent biomedical
     concepts should be included in a causal graph, based only on dataset metadata and
    observed variable descriptions. Be conservative, do not invent unsupported latents, and
    distinguish baseline/current and future/follow-up measurements when longitudinal
    structure is present. Return valid JSON.
```

The generic user prompt is:

```
We are preparing a latent-variable causal discovery experiment for a
{DATASET}-like {DISEASE} dataset.

Dataset metadata:
{DATASET_METADATA}

Do not assume that a true latent variable is already present in the data.
Your job is to propose latent concepts that should be added to the causal graph
because multiple observed variables appear to be measurements of the same
underlying biomedical construct.

Principles for latent clinical concepts:

A latent clinical concept is an underlying medical, biological, pathological, or
functional state that is not directly observed as a single variable, but is
reflected by multiple observed measurements.

When proposing latent concepts:
1. Group variables only when they plausibly measure or reflect the same
   underlying medical state.
2. Interpret the relation as L -> {X1, X2, ..., Xk}, where the observed
   variables are indicators or manifestations of the latent concept.
3. Do not create latent concepts merely because variables are correlated or
   causally related.
4. Do not merge variables from distinct biological processes or disease stages.
5. Variables without a plausible shared measurement interpretation should
   generally remain separate observed variables.
6. Do not force every variable into a latent construct.
7. Prefer clinically interpretable constructs.
8. A latent concept should simplify the causal representation when multiple
   measurements primarily reflect the same underlying state.

Observed variables and meanings:
{VARIABLE_BLOCK}

Please identify latent concepts that should be included in the causal graph.

Requirements:
1. Use only the observed variables and metadata above.
2. Group observed measurements only when the measurement interpretation is
   biologically and statistically plausible.
3. If the same construct appears at two time points, create separate baseline
```

```
   and future latent concepts.
4. Variables should remain observed unless there is clear evidence for a
   multi-indicator latent construct.
5. For each latent concept, provide:
   - latent_name
   - timepoint: "baseline", "future", or "static"
   - indicator_variables
   - parent_nodes
   - biological_rationale
   - confidence: high / medium / low
6. parent_nodes should list the direct parents used to parameterize the latent
   prior p(latent | parents).
7. Parent nodes may be observed variables or other proposed latent concepts.
8. Parent choices should be conservative and clinically justified.
9. Respect longitudinal order: baseline/current variables can be parents of
   future variables, but future variables should not be parents of
   baseline/current variables.
10. Also provide candidate causal relationships among observed variables,
    latent concepts, and observed outcomes.
11. Return valid JSON only. Do not include markdown fences.

Expected JSON schema:
{
  "latent_concepts": [
    {
      "latent_name": "string",
      "timepoint": "baseline|future|static",
      "indicator_variables": ["string"],
      "parent_nodes": ["string"],
      "biological_rationale": "string",
      "confidence": "high|medium|low"
    }
  ],
  "observed_variables_to_keep": ["string"],
  "suggested_edges": [
    {
      "source": "string",
      "target": "string",
      "rationale": "string",
      "confidence": "high|medium|low"
    }
  ],
  "warnings": ["string"]
}
```

For ADNI, the dataset-specific metadata and variable block are:

```
Dataset metadata:
- Number of rows/samples: 5000
- Number of observed variables: 32
- Variables are continuous or treated as continuous after normalization.
- Suffix "_bl" denotes baseline.
- Suffix "_fu" denotes future/follow-up.
- Variables without "_bl" or "_fu" are baseline/static covariates.
- "interval_years" is the elapsed follow-up time and is treated as an observed
  root node.

Observed variables and meanings:
- ADAS13_bl: Baseline ADAS-Cog 13 score.
- interval_years: Elapsed time between baseline and follow-up, in years.
- TAU_bl: Baseline CSF total tau.
- Hippocampus_bl: Baseline hippocampal volume.
- ICV_fu: Follow-up intracranial volume.
- PTAU_bl: Baseline CSF phosphorylated tau.
- Entorhinal_fu: Follow-up entorhinal cortex volume.
- ABETA_bl: Baseline CSF amyloid-beta 42.
- ICV_bl: Baseline intracranial volume.
- TAU_fu: Follow-up CSF total tau.
- Entorhinal_bl: Baseline entorhinal cortex volume.
- Amyloid_centiloids_fu: Follow-up amyloid PET Centiloid measurement.
- DX_bl: Baseline diagnosis / disease-stage code.
- Ventricles_bl: Baseline ventricular volume.
- MidTemp_fu: Follow-up middle temporal gyrus volume.
- Ventricles_fu: Follow-up ventricular volume.
- MMSE_fu: Follow-up MMSE score.
- Sex: Participant sex.
- PTAU_fu: Follow-up CSF phosphorylated tau.
- MidTemp_bl: Baseline middle temporal gyrus volume.
- PTEDUCAT: Years of education.
- Fusiform_bl: Baseline fusiform gyrus volume.
```

```
- Hippocampus_fu: Follow-up hippocampal volume.
- APOE4: APOE epsilon-4 carrier/count indicator.
- ADAS13_fu: Follow-up ADAS-Cog 13 score.
- Hypertension: Hypertension status.
- MMSE_bl: Baseline MMSE score.
- DX_fu: Follow-up diagnosis / disease-stage code.
- ABETA_fu: Follow-up CSF amyloid-beta 42.
- Fusiform_fu: Follow-up fusiform gyrus volume.
- Age: Participant age at baseline.
- Amyloid_centiloids_bl: Baseline amyloid PET Centiloid measurement.
```

For PPMI, the corresponding block is:

```
Dataset metadata:
- Number of rows/samples: 5000
- Number of observed variables: 38
- Variables are continuous or treated as continuous after normalization.
- Suffix "_t" denotes the current/baseline time point.
- Suffix "_fu" denotes future/follow-up.
- "INTERVAL_MONTHS" is the elapsed follow-up time and is treated as an observed
  root node.

Observed variables and meanings:
- AGE_AT_VISIT_t: Age at the current/baseline visit.
- SEX: Participant sex.
- EDUCYRS: Years of education.
- GBA_CARRIER: GBA mutation carrier indicator.
- LRRK2_CARRIER: LRRK2 mutation carrier indicator.
- OTHER_PD_GENE_CARRIER: Other Parkinson's disease gene carrier indicator.
- APOE4_CARRIER: APOE epsilon-4 carrier indicator.
- DAT_CAUDATE_MEAN: Mean caudate DaTScan striatal binding ratio.
- DAT_PUTAMEN_MEAN: Mean putamen DaTScan striatal binding ratio.
- DAT_PUTAMEN_ASYMMETRY: Putamen DaTScan binding asymmetry.
- LEDD_DOSE_t: Baseline levodopa equivalent daily dose.
- NP3TOT_ON_OFF_t: Medication state during the baseline MDS-UPDRS III exam.
- NP3TOT_ON_OFF_fu: Medication state during the follow-up MDS-UPDRS III exam.
- INTERVAL_MONTHS: Elapsed time between visits, in months.
- NP2PTOT_t: Baseline MDS-UPDRS Part II score.
- NP3TOT_t: Baseline MDS-UPDRS Part III score.
- NHY_t: Baseline Hoehn--Yahr stage.
- NP4TOT_t: Baseline MDS-UPDRS Part IV score.
- NP2PTOT_fu: Follow-up MDS-UPDRS Part II score.
- NP3TOT_fu: Follow-up MDS-UPDRS Part III score.
- NHY_fu: Follow-up Hoehn--Yahr stage.
- NP4TOT_fu: Follow-up MDS-UPDRS Part IV score.
- NP1RTOT_t: Baseline rater-assessed MDS-UPDRS Part I score.
- NP1PTOT_t: Baseline patient-reported MDS-UPDRS Part I score.
- SCOPA_AUT_TOTAL_t: Baseline SCOPA-AUT total score.
- NP1RTOT_fu: Follow-up rater-assessed MDS-UPDRS Part I score.
- NP1PTOT_fu: Follow-up patient-reported MDS-UPDRS Part I score.
- SCOPA_AUT_TOTAL_fu: Follow-up SCOPA-AUT total score.
- MCATOT_t: Baseline Montreal Cognitive Assessment score.
- SDMTOTAL_t: Baseline Symbol Digit Modalities Test score.
- HVLT_RECALL_t: Baseline Hopkins Verbal Learning Test recall score.
- SEM_FLUENCY_t: Baseline semantic fluency score.
- LNS_t: Baseline Letter--Number Sequencing score.
- MCATOT_fu: Follow-up Montreal Cognitive Assessment score.
- SDMTOTAL_fu: Follow-up Symbol Digit Modalities Test score.
- HVLT_RECALL_fu: Follow-up Hopkins Verbal Learning Test recall score.
- SEM_FLUENCY_fu: Follow-up semantic fluency score.
- LNS_fu: Follow-up Letter--Number Sequencing score.
```

Causal Layer Proposal The second LLM call uses the following generic system prompt:

```
You are a biomedical causal discovery assistant. Your task is to assign causal
layers to a latent-level {DISEASE} causal graph. Return valid JSON only.
```

The generic user prompt is:

```
We are preparing a prior-constrained causal discovery experiment for a
{DATASET}-like {DISEASE} dataset.

The previous LLM step proposed candidate clinical concept nodes and retained
observed nodes. Now assign causal layers for the nodes to be sorted.

Important scope:
- Use only the nodes listed below.
```

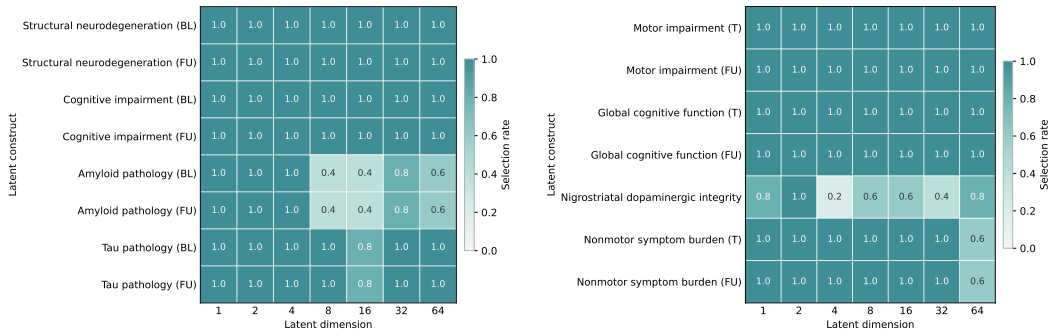
```
Layer convention:
- Layer 0 is the most upstream/root layer.
- Larger layer numbers are more downstream.

Clinical and temporal principles:
1. Static demographic/genetic variables are generally root or near-root.
2. Variables measured later should not be ancestors of variables measured earlier.
3. Future/follow-up nodes should not be ancestors of baseline/current nodes.
4. Use disease-specific clinical knowledge when assigning layers.
5. {DATASET_SPECIFIC_NOTES}
6. The output should be useful as prior information for causal discovery, not
   as a final manually edited graph.

Nodes to be sorted:
{NODE_BLOCK}

Please return valid JSON only, with this schema:
{
  "node_layers": {
    "node_name": 0
  }
}
```

For ADNI, the dataset-specific note is that intracranial volume is treated as a nuisance/scaling variable for MRI measurements rather than as a disease mechanism. The node block contains the proposed baseline and follow-up amyloid, tau, neurodegeneration, and cognition constructs, together with the retained demographic, genetic, vascular, diagnostic, and ICV variables.

```
ADNI node block:
amyloid_pathology_bl
tau_pathology_bl
structural_neurodegeneration_bl
cognitive_impairment_bl
amyloid_pathology_fu
tau_pathology_fu
structural_neurodegeneration_fu
cognitive_impairment_fu
Age
Sex
PTEDUCAT
APOE4
Hypertension
ICV_bl
ICV_fu
DX_bl
DX_fu
```

For PPMI, the node block contains the proposed dopaminergic, motor, non-motor, and cognitive constructs together with the retained demographic, genetic, treatment, medication-context, motor-complication, and imaging variables.

```
PPMI node block:
nigrostriatal_dopaminergic_integrity_static
motor_impairment_t
motor_impairment_fu
nonmotor_symptom_burden_t
nonmotor_symptom_burden_fu
global_cognitive_function_t
global_cognitive_function_fu
AGE_AT_VISIT_t
SEX
EDUCYRS
GBA_CARRIER
LRRK2_CARRIER
OTHER_PD_GENE_CARRIER
APOE4_CARRIER
DAT_PUTAMEN_ASYMMETRY
LEDD_DOSE_t
NP3TOT_ON_OFF_t
NP3TOT_ON_OFF_fu
NP4TOT_t
NP4TOT_fu
```



(a) ADNI latent construct across latent dimensions.

(b) PPMI latent construct across latent dimensions.

Figure 3: Stability of latent concept selection across latent dimensions and random seeds. Each cell reports the fraction of five random seeds in which the corresponding latent construct is selected. A value of 1 indicates that the construct is consistently selected across all seeds.

A.2.5 Latent Concept Selection Stability

In addition to the latent-dimension analysis in Section 5.5, we examine whether the selected latent constructs are stable with respect to latent dimension and random initialization. For each dataset, we consider latent dimensions $d \in \{1, 2, 4, 8, 16, 32, 64\}$ and repeat the representation learning and selection procedure using five random seeds. Figure 3 reports the selection rate of each predefined latent construct, defined as the fraction of random seeds in which the construct is retained by LLCD.

For ADNI, the selected latent structure is highly stable at small latent dimensions. All predefined constructs are consistently retained across all five seeds for $d = 1$, 2, and 4. Variability emerges from $d = 8$ onward and is concentrated primarily in the amyloid pathology constructs, whereas the structural neurodegeneration and cognitive impairment constructs are retained across all dimensions and random seeds. The tau pathology constructs are also stable except for a single run at $d = 16$. These results indicate that the core latent clinical structure recovered on ADNI is robust, while excessively large latent representations introduce greater sensitivity to random initialization.

For PPMI, the selection is more sensitive to both latent dimension and random initialization. At $d = 2$, all predefined constructs are retained across all five seeds. The motor impairment and global cognitive function constructs remain consistently selected across all dimensions, whereas the nigrostriatal dopaminergic integrity construct exhibits substantially greater variation, particularly at $d = 4$ and $d = 32$. Variation in the nonmotor symptom burden constructs appears only at the largest dimension, $d = 64$. Overall, latent concept selection does not change monotonically with increasing latent dimension; instead, larger representations tend to introduce additional variability in which candidate constructs are retained.

Together with the performance-based latent-dimension analysis, these results suggest that increasing latent dimensionality beyond a moderate level does not consistently improve causal structure recovery and may reduce the stability of representation selection.