IV-CRL: INSTRUMENTAL VARIABLE APPROACH FOR CAUSAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study causal representation via linear latent DAGs with jointly observed exogenous instruments W and endogenous measurements X, unknown instrument targets, and cross-loadings without pure measurements. Proportional IV total effects identify the latent total-effect matrix up to permutation and scaling. Reverse-topological residual peeling then recovers the latent DAG and sparse loadings. We give necessary-and-sufficient conditions for population recovery and an example where $P(X)$ is nonidentifying but $P(X, W)$ is identifying. IV-CRL-PEEL combines IV total-effect clustering with trimmed-regression peeling. It improves graph recovery over observational and IV-subspace baselines; with population total effects, it recovers every main-benchmark graph. A masked OneK1K illustration recovers the cis-eQTL groups and empirical latent DAG exactly, with partial agreement between estimated loadings and empirical reference links.

1 INTRODUCTION

Causal representation learning (CRL) aims to recover causally meaningful latent variables and their relations from a high-dimensional observed vector X. Observational factorization may identify a latent subspace while leaving its basis arbitrary, and a rotated basis generally destroys the causal graph. Existing work resolves this ambiguity using interventions, multiple environments, auxiliary variation, non-Gaussianity, or structural restrictions on the measurement model (Ahuja et al., 2023; Squires et al., 2023; von Kügelgen et al., 2023; Buchholz et al., 2023; Varici et al., 2025; Xie et al., 2020; Adams et al., 2021). We study a different regime motivated by genetics and other proxy-rich sciences. Each sample contains the observed variables X and exogenous variables W that act as potential instruments for the latent causal variables with unknown targets, while every observed variable may load on several latent variables and no pure measurement is required.

The closest interventional CRL methods learn from hard or soft changes to latent mechanisms, and several already allow the intervention targets to be unknown (Zhang et al., 2023; von Kügelgen et al., 2023; Buchholz et al., 2023; Varici et al., 2024b; Jin & Syrgkanis, 2024). Our distinction is therefore not unknown targets alone. Although conditioning on W can be viewed as indexing a continuum of additive soft interventions, existing identifiability results rely on conditions for variation across interventional environments that we do not impose; Section 6 gives the precise comparison. We instead ask whether the joint law $P(X, W)$ identifies the latent representation under additive-IV and sparse-loading assumptions. A complementary latent-variable causal-discovery literature primarily uses observational distributions together with non-Gaussianity, heterogeneity, independent-noise constraints, or graphical measurement assumptions; some extensions additionally use unknown interventions (Xie et al., 2020; Adams et al., 2021; Xie et al., 2022; Jiang & Aragam, 2023; Lee & Gu, 2026).

Consider

$$Z = BZ + \Gamma W + \varepsilon, \qquad X = LZ + \eta, \tag{1}$$

where $Z \in \mathbb{R}^d$ contains latent causal variables, $X \in \mathbb{R}^p$ contains observed variables, $W \in \mathbb{R}^K$ contains observed instruments, B is the (acyclic) direct causal effect matrix, and L is the loading matrix. The moment-exogeneity conditions $\mathrm{Cov}(\varepsilon, W) = 0$ and $\mathrm{Cov}(\eta, W) = 0$ give the *IV-to-observed total-effect matrix* (i.e., total effects of W on X)

$$\Theta := \mathrm{Cov}(X, W)\mathrm{Cov}(W)^{-1} = L(I - B)^{-1}\Gamma := M\Gamma. \tag{2}$$

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.

Its kth column θ_k is the total effect on X of a unit change in W_k, after partialling out the other instruments. If W_k additively targets only Z_j, then $\theta_k = \gamma_k m_j$ where γ_k is a scalar (effect of W_k on Z_j) and m_j is the jth column of M. An *IV total-effect direction* (or ray) is the one-dimensional span of a nonzero θ_k, ignoring its sign and magnitude. Full column rank makes two IV total effects proportional exactly when their instruments share a latent target. If every latent variable is targeted by at least one instrument, the proportionality groups identify the *latent-to-observed total-effect matrix* (i.e., total effects of Z on X) , $M = L(I - B)^{-1}$, up to permutation and componentwise scaling, allowing the IV-to-latent assignment Γ to be learned.

Identifying the latent-to-observed total-effect matrix M is only half of the problem. A perturbation to Z_j propagates through its descendants, and hence m_j is not the loading column ℓ_j. Instead,

$$m_j = M_{\mathrm{ch}(j)} b_j + \ell_j, \qquad b_j = B_{\mathrm{ch}(j),j}, \tag{3}$$

where $\mathrm{ch}(j)$ denotes the children of node j in the latent DAG. Throughout, the latent DAG is the graph encoded by B; relationships with instruments and observed variables are encoded by Γ and L. Assume each latent node has at most q children in the latent DAG and each column of L has at most s nonzeros, while allowing each observed variable to load on multiple latent variables. For a sink j, $m_j = \ell_j$ is sparse. Once all sinks are resolved, their directions can be used to subtract descendant propagation from their parents; iterating this operation peels the DAG in reverse-topological batches. We call this operation *causal depropagation*.

We characterize this batch procedure exactly through three requirements. Before a node's true children are available, no candidate set of already resolved directions may produce an s-sparse residual (P1). At the correct round, the true child set must be the unique smallest feasible set (P2), and its regression coefficients must be unique (P3). These conditions are necessary and sufficient for unique population recovery by batch peeling.

IV total-effect directions and sparse loading columns provide complementary identifying information. We exhibit inequivalent sparse latent-DAG models that generate the same observational law $P(X)$, and hence observational sparse factorization cannot choose between them. The IV total effects in the joint law $P(X, W)$ identify the data-generating latent-to-observed total-effect matrix, after which sparse peeling recovers its latent DAG and loading matrix. Arbitrarily correlated measurement errors are allowed as long as they are uncorrelated with the instruments.

Our contributions are:

- **A distinct IV supervision regime.** Sample-level additive instruments identify the latent-to-observed total-effect matrix from a single joint law $P(X, W)$. Their latent targets need not be labeled: proportional IV total effects recover the IV-to-latent partition. The additive shifts leave the latent mechanisms and disturbance law invariant, so our identification uses IV total-effect collinearity rather than variation induced by mechanism-changing interventions as in existing work.
- **Causal depropagation.** Reverse-topological batch peeling converts latent total-effects into the latent DAG and loading matrix; P1–P3 exactly characterize unique population recovery by the algorithm.
- **Strict observational–IV separation and a theorem-aligned estimator.** An explicit pair of observationally indistinguishable sparse models demonstrates that $P(X)$ alone is insufficient, while the IV total effects in $P(X, W)$ select the data-generating model. IV-CRL-PEEL implements the identification argument using IV total-effect clustering, normalized-column SVD, cardinality-first trimmed regressions, and batch updates.
- **Diagnostic empirical evidence.** The complete pipeline outperforms observational and IV-subspace baselines in simulations. Population-M peeling recovers every main-benchmark graph, pointing to estimation of M as a finite-sample bottleneck. A masked-gene OneK1K illustration recovers the known cis target partition and empirical latent DAG exactly, with partial agreement between estimated loadings and empirical reference links.

2 PROBLEM SETUP

We observe i.i.d. samples of (X, W) generated by equation 1. The latent graph is a DAG encoded in B_{kj}: $j \to k$ if $B_{kj} \neq 0$. Let $[d] = \{1, \ldots, d\}$ for any positive integer d, and let $e_j \in \mathbb{R}^d$ denote the jth canonical basis vector (i.e., a unit vector with the jth element being 1). In addition, we impose:

(A1) *Rank conditions:* $L \in \mathbb{R}^{p\times d}$ is full column rank and $\mathrm{Cov}(W)$ is nonsingular.
(A2) *Moment exogeneity:* $\mathrm{Cov}(\varepsilon, W) = 0$ and $\mathrm{Cov}(\eta, W) = 0$.
(A3) *Coverage:* every latent variable is targeted by at least one instrument, i.e., every row of Γ is nonzero.
(A4) *Single-target instruments:* every instrument targets at most one latent variable, i.e., every nonzero column $\Gamma_{\cdot k}$ equals $\gamma_k e_{t(k)}$ for a scalar $\gamma_k \neq 0$, where $t(k) \in [d]$ indicates the latent variable targeted by k.
(A5) *Sparse loadings and sparse DAG:* $\|L_{\cdot j}\|_0 \leq s < p$ and $|\,\mathrm{ch}(j)| \leq q < d$ for every j.

The first four assumptions identify the latent total-effect matrix M, while (A5) is used to identify its decomposition into the latent DAG B and loading matrix L; Appendix Table 1 summarizes these roles.

We do not impose row-sparsity or pure-measurement conditions, and hence an observed variable may load directly on several latent variables. We also allow $\Sigma_\eta := \mathrm{Cov}(\eta)$ to be non-diagonal: residual measurement errors across the coordinates of X may remain correlated after subtracting LZ. Unlike traditional IV approaches where the target of each IV is typically assumed to be known, our method does not make such an assumption because latent variables are themselves unknown *a priori*.

Our default estimation procedure takes d, the number of latent variables, as a user-supplied hyper-parameter, as is customary in latent factor models. This keeps finite-sample rank selection separate from the paper's main problem of recovering the causal structure within a chosen latent dimension. At the population level, however, Assumptions (A3)&(A4) imply that Γ has row rank d. Since $M = L(I - B)^{-1}$ also has rank d, $\mathrm{rank}(\Theta) = d$, so the latent dimension is identified from the population reduced form. For completeness, Section 5 briefly explores finite-sample spectrum rules for estimating d and tests them empirically.

The unavoidable equivalence is relabeling and rescaling. For a permutation P and invertible diagonal D, the change $Z' = D^{-1}P^{\mathsf{T}}Z$ maps

$$(B, L, M) \mapsto (D^{-1}P^{\mathsf{T}}BPD,\ LPD,\ MPD). \tag{4}$$

We identify the graph and parameters modulo this trivial equivalence.

3 Identification

3.1 Instruments identify the latent total-effect matrix

For nonzero vectors $u, v \in \mathbb{R}^p$, write $u \sim v$ when $u = cv$ for some scalar $c \neq 0$. A *ray class* is an equivalence class under $\sim$; it records a direction while discarding its sign and magnitude. Under (A1)–(A4), all instruments assigned to the same latent variable produce IV total-effect columns in the same ray class.

Proposition 1 (IV total-effect direction identification)**.** *Under (A1)–(A4), two nonzero columns θ_k, θ_ℓ of Θ are proportional iff $t(k) = t(\ell)$. Their d ray classes thus identify M up to column permutation and nonzero scaling.*

The proof is provided in Appendix A.1. This identifies M, not B and L individually. Sparsity is essential to separate M into (B, L): with unrestricted L, every acyclic B' induces the same M through $L' = M(I - B')$. Instruments alone therefore do not identify the DAG. The identity $L = M(I - B)$ always gives the recursion in equation 3; assumption (A5) makes its residual ℓ_j sparse and limits the candidate child sets that must be searched.

3.2 Exact characterization of batch peeling

Now that M is identified, the next step is to resolve the latent DAG B and loadings L. We propose a peeling algorithm, which proceeds in reverse-topological batches. The full algorithm of the population version is shown in Figure 1. Suppose that a set T is proposed as the children of node j in the latent DAG. Regressing the latent total-effect direction m_j on the directions M_T subtracts the propagation attributed to those candidate children. We measure how sparse the unexplained component can become by

$$d_j(T) := \min_\beta \|m_j - M_T\beta\|_0. \tag{5}$$

For the true child set, equation 3 gives $m_j - M_{\mathrm{ch}(j)} b_j = \ell_j$, and therefore $d_j(\mathrm{ch}(j)) \le s$ by (A5). Thus $d_j(T) \le s$ is the population signature of a candidate set whose subtraction leaves an s-sparse loading vector. An incorrect set may also have this property through accidental cancellation: a linear combination of incorrect columns M_T may agree with m_j on all but s coordinates. The order in which candidate directions become available is therefore part of identification.

Let reverse depth be $h(j) = 0$ for a sink and $h(j) = 1 + \max_{k \in \mathrm{ch}(j)} h(k)$ otherwise. We index rounds from zero: sinks are resolved in round zero, nodes of depth r are resolved in round r, and those nodes become available as regressors in round $r + 1$. Accordingly, a nonempty set T becomes available in round $a(T) = 1 + \max_{k \in T} h(k)$ where $a(\varnothing) = 0$.

Population batch peeling. Input M, s, q.

1. Initialize the resolved set $R \leftarrow \varnothing$ and the round index $r \leftarrow 0$.
2. Freeze the available set $R_0 \leftarrow R$ and initialize the new batch $A \leftarrow \varnothing$.
3. For each $j \notin R_0$, find the smallest $k \le q$ for which some $T \subseteq R_0$, $|T| = k$, satisfies $d_j(T) \le s$. If no such set exists, leave j unresolved in this round.
4. If such a minimum-cardinality T exists and both it and a coefficient β satisfying $\|m_j - M_T \beta\|_0 \le s$ are unique, store T and β and add j to A; otherwise stop and report non-uniqueness.
5. After evaluating every unresolved node against the same R_0, update $R \leftarrow R_0 \cup A$ and $r \leftarrow r + 1$.
6. If $R = [d]$, proceed to Step 7. If $A = \varnothing$, stop and report incomplete recovery; otherwise repeat from Step 2.
7. Place each stored coefficient vector in its selected column of B and return B and $L = M(I - B)$.

Figure 1: Population batch-peeling procedure. Freezing the resolved set prevents a node found early in a round from being used as a regressor for another node at the same reverse depth. Under P1–P3, every feasible minimum-cardinality candidate is unique.

The following are truth-relative population conditions: M, $\mathrm{ch}(j)$, b_j, $h(j)$, and $a(T)$ are defined by the underlying causal representation. They characterize when the population procedure recovers that representation.

P1 (no premature fit)
For any candidate T, with its availability time $a(T)$ computed from the true reverse depths, $a(T) < h(j)$ and $|T| \le q$ imply $d_j(T) > s$.

P2 (correct minimal fit)
For $\mathcal{F}_j = \{T : a(T) \le h(j),\ |T| \le q,\ d_j(T) \le s\}$, the true child set $\mathrm{ch}(j)$ is the unique minimum-cardinality member.

P3 (coefficient uniqueness)
For the true child set, $\{\beta : \|m_j - M_{\mathrm{ch}(j)}\beta\|_0 \le s\} = \{b_j\}$.

P1 is a stage-specific no-cancellation condition, analogous to faithfulness: it rules out premature sparse residual representations created by special alignments of latent total-effect directions and loadings. P2 is a stage-specific minimality condition: once the true children are available, they form the unique smallest set whose removal leaves an s-sparse residual. It rules out equally or more parsimonious alternative explanations of the node's propagated effects. P3 is a within-support coefficient-identifiability condition: conditional on the true child set, there is a unique coefficient vector whose removal leaves an s-sparse loading vector. It rules out ambiguity between the strengths of latent edges and the loadings.

Theorem 2 (Exact algorithmic characterization)**.** *Given M, s, q, population batch peeling uniquely recovers B and L iff P1–P3 hold for every node.*

The proof is an induction on reverse depth. P1 prevents an early declaration; when round $h(j)$ arrives, $\mathrm{ch}(j)$ is available and feasible because its residual is ℓ_j; P2 selects $\mathrm{ch}(j)$ and P3 selects b_j. Necessity follows by reversing these three failure modes. Appendix A.2 gives the full proof.

For fixed s, q, let $\mathcal{C}_{s,q}$ be the representations satisfying (A1)–(A5) and P1–P3, with P1–P3 evaluated relative to each representation's own M, graph, and reverse depths.

Theorem 3 (Identification of the causal representation)**.** *Within $\mathcal{C}_{s,q}$, $P(X, W)$ identifies the latent dimension, unknown IV target partition, latent DAG, structural coefficients, and loading matrix, up to the permutation/scaling equivalence in equation 4.*

Proof sketch. The rank of Θ gives d, and Proposition 1 gives a normalized M. Two members of $\mathcal{C}_{s,q}$ with the same $P(X, W)$ therefore yield the same M after ray alignment. Applied to each, Theorem 2 makes both equal the deterministic peeling output for fixed (M, s, q), so (B, L) coincide in that normalization. Relabeling or rescaling the rays gives exactly equation 4; measurement noise never enters Θ. □

3.3 Worked example: observational ambiguity resolved by instruments

Consider the chain $1 \to 2 \to 3$ with $d = 3, p = 6, s = 4, q = 1$ and

$$L = \begin{pmatrix} 1 & 4/5 & 0 \\ 7/10 & 11/10 & 0 \\ 0 & 1 & 9/10 \\ 0 & 3/5 & 6/5 \\ 11/10 & 0 & 7/10 \\ 4/5 & 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 & 0 \\ 2/5 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}. \tag{6}$$

Every row of L has two nonzeros and every column is four-sparse, so there are no pure measurements. With independent Gaussian structural and measurement errors, let $M = L(I - B)^{-1}$ and $\Sigma_X = MM^\mathsf{T} + I_6$. Appendix A.3 constructs an inequivalent four-sparse loading matrix and latent DAG with the same Σ_X. Both loadings have twelve nonzeros and both measurement-noise covariances are I_6, so $P(X)$ remains ambiguous under equally sparse loadings and independent measurement noise.

Now add one unit-strength valid instrument per latent node. Then $\Theta = M$, so the IV total-effect matrix identifies the data-generating latent total-effect matrix. Sparse peeling uses

$$m_3 = \ell_3, \qquad m_2 - \tfrac{1}{2}m_3 = \ell_2, \qquad m_1 - \tfrac{2}{5}m_2 = \ell_1,$$

thereby resolving node 3, then node 2, and finally node 1, and recovering the chain $1 \to 2 \to 3$. Although $m_2 - \frac{5}{2}m_1 = -\frac{5}{2}\ell_1$ is also four-sparse, it uses node 1, which is unavailable when node 2 is peeled.

4 IV-CRL-Peel: Sample Batch-Peeling Procedure

The sample procedure replaces population quantities by estimates and exact residual sparsity by a trimmed-loss threshold. For centered X and W, estimate $\widehat{\Theta} = X^\mathsf{T}W(W^\mathsf{T}W)^{-1}$, screen negligible columns, and normalize retained columns as $\widehat{u}_k = \widehat{\theta}_k/\|\widehat{\theta}_k\|_2$. Multi-start projective k-means partitions them into d groups using the sign-invariant loss $1 - (\widehat{u}_k^\mathsf{T}v)^2$ for a unit cluster direction v. Each group is represented by the first left singular vector of its normalized columns, rescaled to norm $\sqrt{p}$, giving $\widehat{M}$. This fixes ray scale and keeps the coordinate-averaged trimmed loss comparable across p. Fewer than d usable IV columns triggers an insufficient-coverage error. We supply d by default; Appendix B studies spectral estimation.

For an unresolved node j and candidate child set T, define

$$\widehat{\mathcal{L}}_s(j, T) = \min_{\beta} \frac{1}{p-s} \sum_{a=1}^{p-s} r_{(a)}(\beta)^2, \quad r(\beta) = \widehat{m}_j - \widehat{M}_T\beta, \tag{7}$$

where $r_{(1)}^2 \le \cdots \le r_{(p)}^2$. The loss discards the s largest squared residuals; it is zero at the true child set in the population because the remainder is the s-sparse loading vector ℓ_j. We can view equation 7 as the least-squares analogue of equation 5: $(p-s)\widehat{\mathcal{L}}_s(j, T) = \min_{\beta, e:\, \|e\|_0 \le s} \|\widehat{m}_j - \widehat{M}_T\beta - e\|_2^2$. In implementation, multi-start iterations alternate retaining the $p - s$ smallest residuals and least-squares refitting, which leads to an approximate minima.

With these estimates, the full algorithm is shown in Figure 2. The output consists of $(\widehat{B}, \widehat{L})$ and a completion indicator, so the method can abstain rather than force unsupported edges. Each round requires at most $O(d^{q+1})$ candidate trimmed-regression fits in the worst case; cardinality-first stopping often evaluates fewer.

Finite-sample batch peeling. Input $\widehat{M}, s, q, \tau$.

1. Initialize the resolved set $R \leftarrow \varnothing$ and the round index $r \leftarrow 0$.
2. Freeze the available set $R_0 \leftarrow R$ and initialize the new batch $A \leftarrow \varnothing$.
3. For each $j \notin R_0$, find the smallest $k \leq q$ for which some $T \subseteq R_0$, $|T| = k$, satisfies $\widehat{\mathcal{L}}_s(j, T) \leq \tau$. If no such set exists, leave j unresolved in this round.
4. Among the qualifying sets at that cardinality, store a minimum-loss set and refit its coefficients on the $p - s$ retained coordinates; add j to A.
5. After evaluating every unresolved node against the same R_0, update $R \leftarrow R_0 \cup A$ and $r \leftarrow r + 1$.
6. If $R = [d]$ or $A = \varnothing$, proceed to Step 7; otherwise repeat from Step 2.
7. Form $\widehat{B}$ from the stored coefficients, with zero columns for unresolved nodes. Return $\widehat{B}$, $\widehat{L} = \widehat{M}(I - \widehat{B})$, the unresolved nodes, and completion status $\mathbf{1}\{R = [d]\}$.

Figure 2: Finite-sample counterpart of Figure 1. Exact residual sparsity is replaced by the trimmed-loss threshold τ; the resolved set remains frozen within each round.

5 Experiments

We organize evaluation around four questions: **(Q1)** does the pipeline improve on observational and IV-informed subspace baselines; **(Q2)** which stage limits finite-sample recovery; **(Q3)** how do sample size, IV strength, threshold, dimension, and controlled stresses affect reliability; and **(Q4)** can the full pipeline recover a deliberately hidden, genetically anchored low-dimensional system in real high-dimensional data? Appendices B&C contain full designs, threshold sensitivity, additional diagnostics, and masked-gene details. GES denotes the greedy equivalence search (Chickering, 2002).

5.1 Q1-Q3: Simulations

Factorial design and comparison logic. The main benchmark contains 2,400 data sets: 100 repetitions over $n \in \{250, 500, 1000\}$, IV strength $\{0.3, 0.7\}$, one or two IVs per node, and no measurement noise or SNR 5. We use $d = 5, p = 50, q = 1, s = 10$. Each possible forward edge in a random topological ordering is included with probability .3 subject to the out-degree cap; nonzero DAG coefficients have random signs and magnitudes in $[0.3, 0.7]$. Each loading column has exactly s randomly located nonzeros with random signs and magnitudes in $[0.5, 1.5]$, conditional on full column rank.

All learned-representation baselines feed their estimated scores to the same Gaussian GES learner. *PCA+GES* is an observational subspace baseline. *Sparse-varimax+GES* rotates PCA loadings and retains the s largest entries per component. *IV-SVD+GES* uses the leading left singular vectors of $\widehat{\Theta}$, an IV-informed subspace basis. External interventional CRL methods are not benchmarked because their experimental inputs and identification assumptions differ from our paired additive-IV design. *Oracle-M peeling* supplies population total effects; *oracle-Z+GES* supplies realized latent scores.

Our graph metrics are directed precision and recall, which assess oriented edges, and skeleton Matthews correlation coefficient (MCC), which also credits correctly recovered nonedges. Appendix B specifies node matching and GES orientation conventions. Completion is reported separately as the fraction of data sets for which peeling resolves every node. In incomplete outputs, unresolved nodes retain zero outgoing columns in $\widehat{B}$; returned edges are scored normally and missing true edges as false negatives. Figure 3 compares methods by sample size and instrument strength; Figure 4 displays the performance of IV-CRL for every factorial condition.

Q1–Q2: recovery improves when the IV signal is estimable. Figure 3 shows a pronounced strength-by-sample-size interaction. With weak IVs, IV-CRL-Peel rises from precision/recall/MCC $.074/.061/.047$ at $n = 250$ to $.654/.705/.610$ at $n = 1000$; at $n \leq 500$, its MCC remains below sparse varimax. With strong IVs, it already attains $.712/.738/.669$ at $n = 250$, improves to $.917/.888/.884$ at $n = 500$, and reaches $.969/.910/.922$ at $n = 1000$. In contrast, IV-SVD and PCA remain near zero across n, while sparse-varimax MCC stays near .22–.26. These comparisons support the complete identification-and-peeling pipeline over subspace estimation followed by GES. Oracle-M peeling attains unit recovery throughout, pointing to estimation of M as a finite-sample

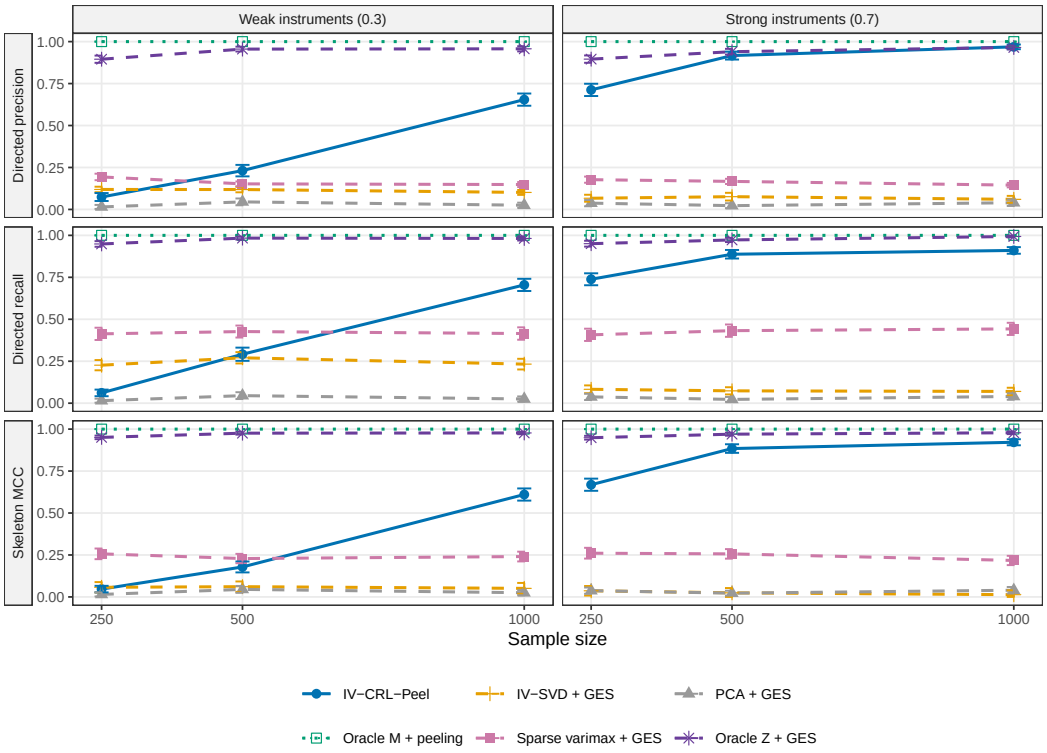


Figure 3: Directed precision, directed recall, and skeleton MCC versus sample size and IV strength, averaged over measurement-noise and instrument-count conditions; bars are 95% Monte Carlo intervals. The complete IV-CRL-PEEL pipeline ($\tau = .06$) outperforms IV-SVD in the strong-IV regime and increasingly with n in the weak-IV regime. Oracle-M peeling attains unit recovery; oracle-Z GES is an observational benchmark

bottleneck; with strong IVs, the learned pipeline also approaches the oracle-Z GES benchmark as n grows.

Q3a: noise, instrument multiplicity, and threshold sensitivity. As shown in Figure 4, under weak IVs, useful recovery generally requires $n = 1000$, and both measurement noise and splitting a fixed total strength across two instruments slow the improvement. Under strong IVs, performance is already substantial at $n = 250$ and remains high by $n = 500$–1000 across both noise and multiplicity conditions; measurement noise is most visible at $n = 250$. The two-instrument curves otherwise track the one-instrument curves closely. Increasing τ from .06 to .08 improves all four metrics averaged across factorial conditions (Appendix Table 2), motivating principled tuning as future work.

Q3b: dimension and controlled stresses. A separate 2,500-data-set suite has six scaling configurations varying $d \in \{5, 10, 20\}$ with $p/d = 10$ and $n/d \in \{50, 100\}$, plus 19 one-feature stress configurations. The stress reference is $(n, d, p, s) = (1000, 5, 50, 10)$ with two IVs per node, strength .7, and SNR 5. The graph-complexity block instead uses $(n, d, p, s) = (1000, 10, 100, 20)$ so that $q = 2$ produces a substantive branching contrast. Appendix Table 4 and Figure 6 show that $n/d = 50$ is insufficient as d grows, whereas results remain strong at $n/d = 100$. Approximate loadings, correlated measurement errors, and higher out-degree have modest effects in the tested ranges; off-target IV effects, correlated instruments, and unequal IV strengths reduce recovery (Appendix Figure 7). The latter two stresses are permitted by our assumptions, not violations.

Latent dimension estimation. At the population level, the assumptions imply $d = \mathrm{rank}(\Theta)$, but finite-sample rank selection is distinct from graph recovery within a chosen dimension. We therefore take d as supplied by default and study two spectral selection approaches as an optional preliminary

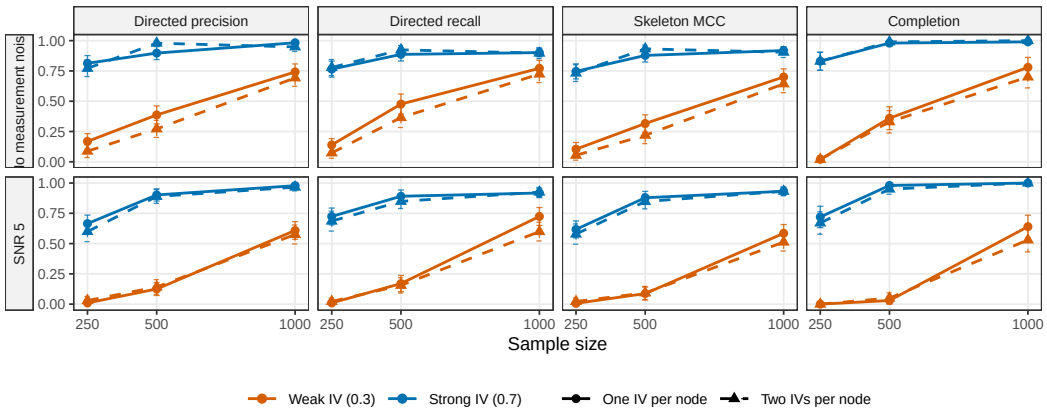


Figure 4: Factorial IV-CRL-PEEL sensitivity at $\tau = .06$ with 95% Monte Carlo intervals (100 repetitions per point). Rows vary noise; columns show metrics; color denotes IV strength; line type and shape denote IV count.

step. We test them using 600 additional simulated data sets. As reported in Appendix Table 3, an eigengap approach selects d exactly in 96.3% of cases (mean absolute error .097), whereas a permutation-threshold approach is exact in 71.7% and tends to overestimate. These results suggest that d is estimable in the tested regimes.

5.2 Q4: ONEK1K MASKED-GENE ILLUSTRATION

We use the OneK1K single-cell eQTL study (Yazar et al., 2022) to illustrate the pipeline in genomics. Figure 5 (left) combines a fitted latent-gene DAG with empirical response links; Appendix C details their construction. The three latent variables Z are donor-level expression profiles of *ENO1*, *CYB5A*, and *ANXA6*, observed during reference construction and then hidden from our algorithm. Instruments W are their cis-eQTL genotypes, and X contains seven other expression profiles selected as stable responses. We fit IV-CRL-PEEL with $d = 3$, $s = 4$, $q = 2$, and $\tau = .06$ using only the six genotypes and seven measured expressions.

As shown in Figure 5 (right), IV total-effect clustering recovers all cis targets and peeling recovers both latent edges without extras. For $Z \to X$, cutting at the largest adjacent magnitude gap among the top s entries recovers four of seven reference links: *ENO1→FKBP2*, *ENO1→CD53*, *CYB5A→CTD-2054N24.2*, and *ANXA6→TMED8*; two additional entries are selected.

6 RELATED WORK

Interventional causal representation learning. Informative interventions can identify latent coordinates under linear or nonlinear mixing (Ahuja et al., 2023; Squires et al., 2023; von Kügelgen et al., 2023; Buchholz et al., 2023), including with unknown targets (Varici et al., 2024a;b; 2025; Bing et al., 2024; Jin & Syrgkanis, 2024). Existing guarantees use generalized faithfulness under soft mechanism changes, genericity across perfect-intervention domains, precision- or score-function changes, or sufficiently diverse environments. Although $P(X \mid W = w)$ can be viewed as a continuum of environments, here W creates additive shifts while B and the disturbance law remain invariant; identification instead uses IV total-effect collinearity and sparse depropagation.

Latent causal graphs and sparse measurement models. Observational latent-graph identification uses non-Gaussianity, heterogeneity, independent-noise constraints, or graphical measurement structure (Shimizu et al., 2006; Xie et al., 2020; Adams et al., 2021; Xie et al., 2022); unknown environment-level interventions provide another route (Jiang & Aragam, 2023), and recent work weakens pure-child conditions through expansion or related graphical assumptions (Adams et al., 2021; Lee & Gu, 2026). Our continuous linear model requires no pure measurement: every observed variable may load directly on several latent variables. Although global sparse-basis uniqueness can

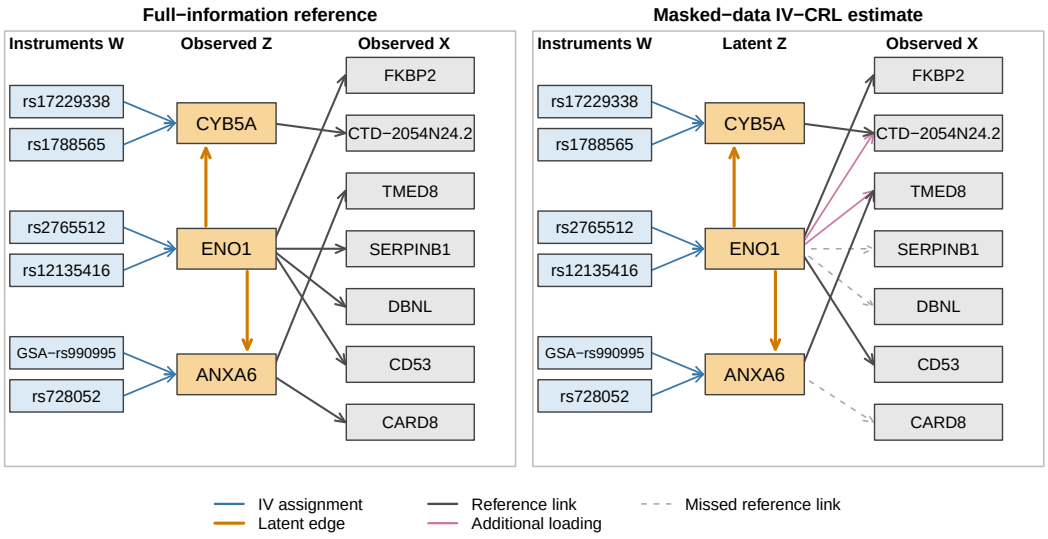


Figure 5: OneK1K masked-gene illustration: full-information empirical reference (left) and estimate after hiding Z (right). Blue arrows denote cis-eQTL assignments and orange arrows latent edges. Gray arrows denote empirical $Z \to X$ reference links and their recovery. Purple arrows are additional selected loadings; dashed gray arrows are missed reference links.

itself identify representations (Spielman et al., 2012; Qu et al., 2014; Sturma et al., 2025; Xu et al., 2024), our worked example (Section 3.3) lies outside that regime. Instruments first identify the latent total-effect matrix M, after which stage-specific sparsity removes propagation only along child directions available at the current peeling round.

Instrumental-variable perspective. Classical IV and proxy methods identify effects or correct measurement error (Hu & Schennach, 2008; Miao et al., 2018), while shift interventions learn observed-variable structure (Rothenhäusler et al., 2015). We instead identify a latent coordinate system and DAG: IV total-effect directions resolve rotations, and sparse peeling converts their propagation into latent edges and loadings.

7 Discussion

Additive instruments can resolve latent rotations and, with sparse depropagation, identify a latent DAG without pure measurements. This separates selection of a causally meaningful latent basis from recovery of directed relationships within that basis, while allowing observed variables to cross-load. At the population level, P1–P3 characterize unique recovery by batch peeling. These guarantees remain conditional on exogeneity, single-target coverage, and the stated sparsity conditions. The OneK1K analysis illustrates recovery relative to an empirical reference.

Finite-sample performance depends on estimating M accurately and separating sparse loading residuals from estimation error. The contrast with population-M peeling motivates error bounds that connect IV strength, separation between total-effect directions, and residual separation to graph recovery. Data-adaptive choices of the trimming budget and acceptance threshold should account for uncertainty in both IV grouping and peeling, rather than optimizing completion alone. Further priorities include sensitivity analysis for multi-target instruments and uncertainty quantification for selected edges. Although spectral rank selection is promising in our simulations, propagating dimension-selection errors through the subsequent stages remains an open problem.

AI USE STATEMENT

We used generative AI tools to assist with the following tasks to varying degrees: generating synthetic data sets, developing theoretical models, formulating and proving mathematical claims, providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, and interpreting results. We have not used generative AI tools for proposing or refining hypotheses. Assisting with translation and supporting qualitative and thematic data analysis are not applicable to this work. The authors verified the mathematical statements and proofs, references, implementation, and experimental results and retain full responsibility for the paper.

ETHICS STATEMENT

The real-data study uses previously released, donor-deidentified OneK1K data and makes no individual-level predictions.

REPRODUCIBILITY STATEMENT

All experiments are generated from the stated structural equations. Appendix B specifies parameter distributions, full factorial designs, baselines, metrics, thresholds, and Monte Carlo counts.

REFERENCES

Jeffrey Adams, Niels Richard Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407, 2023.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Proceedings of the 3rd Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 843–867, 2024.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Yingyao Hu and Susanne M. Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jikai Jin and Vasilis Syrgkanis. Learning linear causal representations from general environments: Identifiability and intrinsic ambiguity. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Seunghyun Lee and Yuqi Gu. Identifiability of latent causal graphical models without pure children. *Transactions on Machine Learning Research*, 2026.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 37.1–37.18, 2012.

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32540–32560, 2023.

Nils Sturma, Moritz Kranzlmüller, Irem Portakal, and Mathias Drton. Matching criterion for identifiability in sparse factor analysis. *arXiv preprint arXiv:2502.02986*, 2025.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, 2024a.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 48603–48638, 2023.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-gaussian latent hierarchical structure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55389–55433, 2024.

Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R. P. Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L. Cook, Chun Jimmie Ye, Kirsten A. Fairfax, Alex W. Hewitt, and Joseph E. Powell. Single-cell eqtl mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, 2022. doi: 10.1126/science.abf3041.

Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Theory, Proofs, and Examples

A.1 IV total-effect algebra and directions

Solving the structural equations gives

$$Z = (I - B)^{-1}(\Gamma W + \varepsilon), \qquad X = M\Gamma W + M\varepsilon + \eta.$$

Table 1: Role of the main assumptions.

Condition	Identifying role
Rank conditions (A1)	make M full rank and the IV regression well-defined
Moment exogeneity (A2)	yields the IV-to-observed total-effect identity $\Theta = M\Gamma$
Coverage (A3) + single-target instruments (A4)	with (A1)–(A2), identify $d = \mathrm{rank}(\Theta)$ and the unknown IV target groups
Sparsity (A5) + P1–P3	uniquely decomposes M into DAG B and loading matrix L

Taking covariance with W and using (A2) yields

$$\mathrm{Cov}(X, W) = M\Gamma\mathrm{Cov}(W),$$

so nonsingularity of $\mathrm{Cov}(W)$ proves equation 2.

Proof of Proposition 1. Under single-target instruments, $\theta_k = \gamma_k m_{t(k)}$. Instruments sharing a target therefore have proportional nonzero IV total effects. Conversely, if $\theta_k = c\theta_\ell$, then $\gamma_k m_{t(k)} = c\gamma_\ell m_{t(\ell)}$. Full column rank of M implies that distinct columns are linearly independent, hence $t(k) = t(\ell)$. Coverage (A3) ensures that every latent variable has at least one nonzero IV total effect, while the single-target condition (A4) ensures that each such effect belongs to only one latent variable. There are therefore exactly d ray classes, one per latent variable. Selecting an arbitrary nonzero representative and normalization in each class recovers M up to permutation and nonzero column scaling. □

The graph is invariant under the equivalence in equation 4. To verify the parameter transformation, use

$$I - D^{-1}P^{\mathsf{T}}BPD = D^{-1}P^{\mathsf{T}}(I - B)PD$$

and thus $L'(I - B')^{-1} = L(I - B)^{-1}PD = MPD$.

A.2 Proof of the exact peeling characterization

Proof of Theorem 2. Sufficiency. We induct on peeling round r. Suppose the resolved set at its start is exactly $R_r = \{k : h(k) < r\}$. If $h(j) > r$, every candidate $T \subseteq R_r$ has $a(T) \le r < h(j)$, so P1 makes it infeasible. If $h(j) = r$, every child has smaller reverse depth, hence $\mathrm{ch}(j) \subseteq R_r$. The structural recursion makes $\mathrm{ch}(j)$ feasible because

$$\|m_j - M_{\mathrm{ch}(j)} b_j\|_0 = \|\ell_j\|_0 \le s.$$

By P2, it is the unique feasible set of smallest cardinality; the cardinality-first rule selects it. P3 uniquely selects b_j. Freezing R_r prevents nodes of the same depth from being used within this round. Consequently, the new batch is exactly $\{j : h(j) = r\}$, establishing the induction. After the final round every column of B is recovered and $L = M(I - B)$.

Necessity. Suppose the algorithm uniquely recovers the true child sets and coefficients. A node cannot be resolved before its children; once they are available, their feasible fit makes further delay impossible. Induction therefore gives the true resolution rounds $h(j)$. If P1 failed, an earlier feasible candidate would force an incorrect declaration or a report of non-uniqueness. At round $h(j)$, failure of P2 would similarly force a smaller incorrect set or a report of non-uniqueness among minimum-cardinality sets. Failure of P3 would yield non-unique coefficients on the true child set. Each case contradicts the assumed unique recovery. □

A.3 Strict observational–IV separation

Take the chain $1 \to 2 \to 3$ with

$$L = \begin{pmatrix} 1 & 4/5 & 0 \\ 7/10 & 11/10 & 0 \\ 0 & 1 & 9/10 \\ 0 & 3/5 & 6/5 \\ 11/10 & 0 & 7/10 \\ 4/5 & 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 & 0 \\ 2/5 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}. \tag{8}$$

Each row of L has two nonzeros (no pure measurements), each column has four, and $q = 1$. With structural and measurement noise covariances I_3 and I_6,

$$M = L(I - B)^{-1} = \begin{pmatrix} 33/25 & 4/5 & 0 \\ 57/50 & 11/10 & 0 \\ 29/50 & 29/20 & 9/10 \\ 12/25 & 6/5 & 6/5 \\ 31/25 & 7/20 & 7/10 \\ 1 & 1/2 & 1 \end{pmatrix}, \quad \Sigma_X = MM^{\mathsf{T}} + I_6. \tag{9}$$

Define

$$R = \begin{pmatrix} 0 & 342/725 & -1 \\ 0 & -5643/7250 & 2/5 \\ 1 & 19/20 & 0 \end{pmatrix}, \qquad \widetilde{L} = MR = \begin{pmatrix} 0 & 0 & -1 \\ 0 & -4617/14500 & -7/10 \\ 9/10 & 0 & 0 \\ 6/5 & 627/1450 & 0 \\ 7/10 & 7087/7250 & -11/10 \\ 1 & 3743/3625 & -4/5 \end{pmatrix}.$$

Every column of $\widetilde{L}$ is four-sparse. Take a QR-type factorization $R = QDU$, with Q orthogonal, D positive diagonal, and U unit upper triangular. Numerically,

$$U \approx \begin{pmatrix} 1 & .9500 & 0 \\ 0 & 1 & -.9453 \\ 0 & 0 & 1 \end{pmatrix}.$$

Set $I - \widetilde{B} = U$, $\widetilde{\Omega} = D^{-2}$, and $\widetilde{\Psi} = I_6$. Then $\widetilde{B}$ is acyclic with maximum out-degree one and the alternative latent total-effect matrix is

$$\widetilde{M} = \widetilde{L}(I - \widetilde{B})^{-1} = MRU^{-1} = MQD.$$

Consequently,

$$\widetilde{M}\widetilde{\Omega}\widetilde{M}^{\mathsf{T}} = MQDD^{-2}DQ^{\mathsf{T}}M^{\mathsf{T}} = MM^{\mathsf{T}},$$

and both zero-mean Gaussian models generate the same $P(X)$. Both measurement-noise covariances are I_6, and both L and $\widetilde{L}$ contain exactly twelve nonzeros. Their loading supports and DAGs nevertheless differ, so they are not related by permutation/scaling.

Now add one unit-strength valid instrument per node, so the true $\Theta = M$. Because Q is not a signed permutation, the alternative basis MQD is incompatible with these IV total-effect directions. Peeling is unique. First, m_3 is the only four-sparse column. On rows where m_3 is nonzero, the ratios m_{r2}/m_{r3} are $\{29/18, 1, 1/2, 1/2\}$, so $m_2 - \alpha m_3$ can have two new zeros only at the unique value $\alpha = 1/2$, giving ℓ_2. At this round the ratios $m_{r1}/m_{r3} = \{29/45, 2/5, 62/35, 1\}$ are distinct, so node 1 is not prematurely feasible. Finally, the ratios m_{r1}/m_{r2} repeat only at $2/5$ (rows 3 and 4), uniquely giving $m_1 - \frac{2}{5}m_2 = \ell_1$. Hence P1–P3 hold, completing the verification of the separation example in Section 3.3.

B Supplementary Experiments

B.1 Data generation and evaluation

For each data set, nodes are placed in a random topological order. Each possible forward edge is included with probability .3, subject to the out-degree cap q; nonzero coefficients have independent random signs and magnitudes $\mathrm{Unif}(.3, .7)$. Each loading column has exactly s uniformly selected nonzeros with independent random signs and magnitudes $\mathrm{Unif}(.5, 1.5)$, conditional on full column rank. This produces cross-loading without imposing pure rows. Instruments and structural disturbances are Gaussian. With m instruments per node and fixed total strength g, individual coefficients have magnitude $g/\sqrt{m}$ with random signs. Measurement noise is Gaussian and either absent or scaled so that average signal variance divided by noise variance is 5. Unless a stress test says otherwise, instruments, structural errors, and measurement errors are mutually independent and have independent coordinates.

IV-CRL and its peeling oracles match $\widehat{M}$ to M by a one-to-one assignment maximizing total absolute column-cosine similarity; GES pipelines instead maximize total absolute correlation between estimated scores and realized Z. Directed precision and recall compare aligned edges with the true DAG. GES is scored using one representative DAG of its estimated equivalence class, whose orientation is not generally observationally identifiable. Skeleton MCC ignores orientation and includes nonedges. Completion records whether peeling resolves every node. Runtime is wall-clock time per threshold on the same implementation.

The main benchmark crosses $n \in \{250, 500, 1000\}$, $g \in \{.3, .7\}$, instruments per node $m \in \{1, 2\}$, and measurement noise $\in \{0, \mathrm{SNR}5\}$, with 100 repetitions in every cell (2,400 data sets). The robustness suite uses 100 repetitions in six scaling configurations and 19 controlled-stress configurations (2,500 further data sets). Each robustness data set is evaluated at $\tau \in \{.04, .06, .08\}$; threshold fits are not counted as independent data sets.

Table 2: Threshold sensitivity in the 2,400-data-set main benchmark. Values are averages over all factorial conditions.

τ	directed precision	directed recall	skeleton MCC	completion
.04	.497	.531	.474	.468
.06	.593	.599	.552	.600
.08	.643	.600	.568	.702

Although $\tau = .08$ is best on every grand-average summary, it is not cellwise dominant over $\tau = .06$: directed precision, directed recall, and skeleton MCC decrease in 7, 11, and 9 of the 24 factorial cells, respectively, whereas completion is never lower.

B.2 Optional latent-dimension selection

The primary IV-CRL analysis treats d as supplied. For completeness, this separate rank-selection experiment examines an optional data-driven preliminary step using two instruments per latent node, SNR 5, $p/d = 10$, and 50 repetitions in each combination of $d \in \{5, 10, 20\}$, $n/d \in \{50, 100\}$, and IV strength $g \in \{.3, .7\}$. The eigengap estimate maximizes $\widehat{\sigma}_r/\widehat{\sigma}_{r+1}$ over adjacent singular values of $\widehat{\Theta}$. The permutation estimate counts consecutive ordered singular values exceeding their componentwise 95th-percentile null thresholds after permuting the rows of X relative to W.

Table 3: Dimension-selection accuracy averaged over $d \in \{5, 10, 20\}$; each row contains 150 data sets.

n/d	IV strength	eigengap exact	eigengap MAE	permutation exact	permutation MAE
50	.3	.867	.373	.660	.993
100	.3	.987	.013	.773	.513
50	.7	1.000	0	.733	.760
100	.7	1.000	0	.700	.960

The eigengap rule is exact in 96.3% of all 600 data sets. The permutation rule is exact in 71.7% and overestimates in 27.0%, especially when the first trailing singular value lies close to its null threshold.

B.3 Scaling and controlled stresses

The scaling experiment fixes $p/d = 10$, $s = 2d$, $q = 1$, two instruments per node, strength .7, and SNR 5. Table 4 reports the central threshold $\tau = .06$.

The assumption-stress tests change one feature from the $n = 1000, d = 5, p = 50, s = 10$, strong-IV, SNR-5 reference design. Approximate loading contamination with standard deviation .05 gives directed precision/recall and skeleton MCC of $.967/.953/.948$ (reference $.962/.943/.939$), and AR(1) measurement-error correlation .6 gives $.978/.948/.952$ (reference $.987/.961/.967$). Instrument correlation .6, off-target effects with coefficient ratio .2, and IV-strength coefficient of variation 1 reduce skeleton MCC to $.821$, $.858$, and $.667$, respectively. The graph-complexity block instead uses

Table 4: Scaling experiment (100 repetitions per row). Note that completion is an all-or-nothing graph-level indicator; incomplete fits may still recover substantial portions of the graph and are included when calculating precision, recall, and MCC.

d	p	n	n/d	directed precision	directed recall	skeleton MCC	completion	sec.
5	50	250	50	.714	.744	.665	.71	.022
10	100	500	50	.772	.597	.631	.25	.108
20	200	1000	50	.793	.498	.599	.00	.556
5	50	500	100	.951	.905	.902	.97	.020
10	100	1000	100	.981	.955	.962	.89	.118
20	200	2000	100	.998	.993	.995	.97	.733

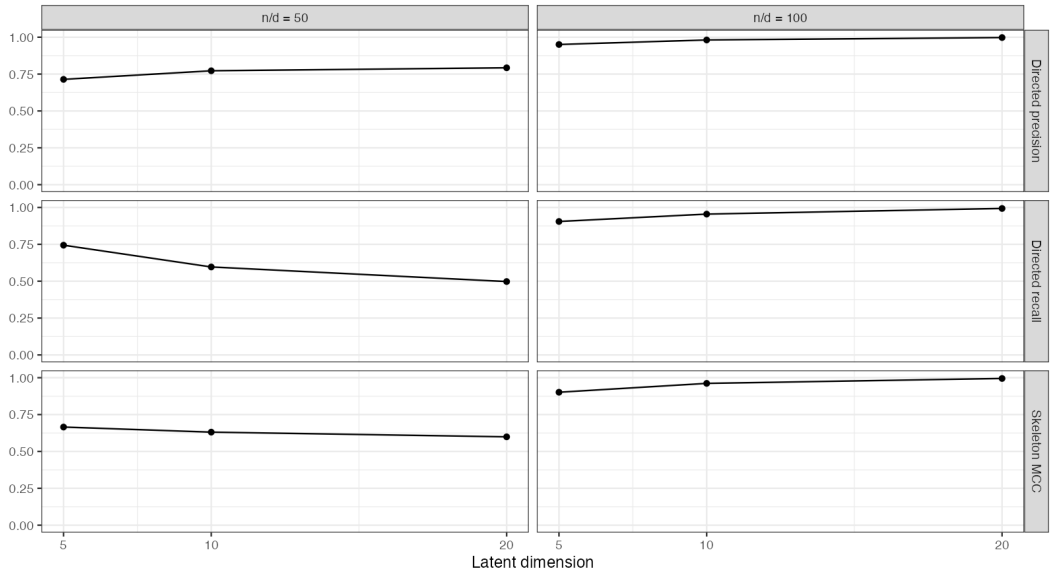


Figure 6: Directed precision, directed recall, and skeleton MCC in the scaling experiment at $\tau = .06$ (100 repetitions per point). When $n/d = 100$, all three remain strong as d grows; when $n/d = 50$, recall declines with dimension even though precision and skeleton MCC show substantial partial recovery.

$d = 10, p = 100, s = 20$ so that the $q = 2$ cap repeatedly permits branching; n, IV strength, and SNR remain at their reference values, and q is varied within this block. Raising q from one to two changes precision/recall/MCC from $.990/.953/.965$ to $.964/.916/.922$, a modest effect.

Across $\tau = .04, .06, .08$, precision/recall/MCC are respectively $.915/.928/.905$, $.942/.909/.906$, and $.923/.823/.839$ in the stress family; $.891/.856/.849$, $.977/.935/.944$, and $.948/.849/.871$ in the graph-complexity family; and $.644/.547/.553$, $.868/.782/.792$, and $.919/.844/.857$ in the scaling family. No single fixed threshold dominates every metric and regime, reflecting the model-specific separation margins of finite-sample peeling.

C Construction of the OneK1K Candidate Gene Set and Empirical Reference Network

This section describes the full-information preprocessing and screening steps used to construct the candidate gene set and the empirical reference network for the OneK1K illustration. All expression analyses use donors as the independent observational units. The construction proceeds from donor-level pseudobulk expression to an eligible gene pool, a cis-eQTL screen, conditional instrument selection, a trans-reach screen, and finally a genetically anchored gene network. The resulting network is an empirical reference.

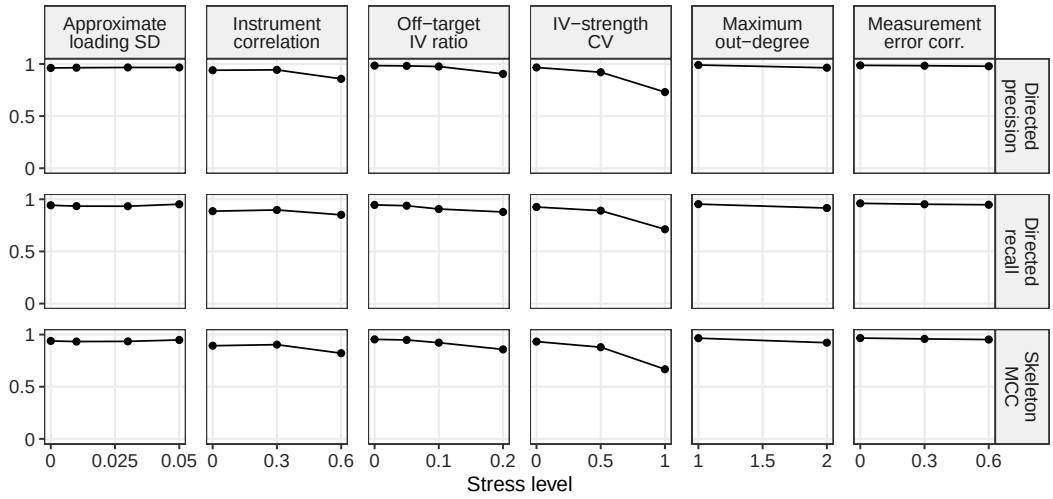


Figure 7: One-factor-at-a-time stress suite at $\tau = .06$, evaluated by directed precision, directed recall, and skeleton MCC. The panels vary the standard deviation of dense contamination added to zero loadings, equicorrelation among instruments, the ratio of an off-target IV effect to its primary-target effect, the coefficient of variation of within-target IV strengths at fixed total strength, the maximum out-degree bound, and AR(1) measurement-error correlation. In the out-degree panel, which uses $d = 10$, the displayed value is both the upper bound used to generate the true DAG and the correctly specified algorithm input q; the realized maximum degree can be smaller. All other features are held at their reference values within each panel.

C.1 Donor-level CD4 expression profiles

The OneK1K cohort provides single-cell RNA-sequencing and genotype data for 982 donors (Yazar et al., 2022). We restrict the expression data to naïve thymus-derived CD4-positive alpha-beta T cells and central-memory CD4-positive alpha-beta T cells. A donor is retained if at least 50 cells belong to these two populations. Raw integer counts are summed across all retained cells from the same donor, producing one pseudobulk expression profile per donor. The resulting data contain 386,774 cells aggregated into $n = 960$ donor-level profiles.

Let C_{ig} denote the pseudobulk count for gene g in donor i, and let $N_i = \sum_g C_{ig}$ be the corresponding library size. Donor-level expression is transformed to log counts per million (log-CPM) according to

$$Y_{ig}^{\mathrm{raw}} = \log\left(1 + \frac{10^6 C_{ig}}{N_i}\right).$$

C.2 Eligible gene pool and covariate adjustment

A gene enters the initial expression pool if it is protein coding, lies on an autosome, and is detected in at least 90% of retained donors. We exclude mitochondrial genes, ribosomal protein genes, immunoglobulin genes, T-cell-receptor genes, and genes overlapping the extended major histocompatibility complex region on chromosome 6, defined as 25–34 Mb. These filters leave 8,275 eligible genes.

For each eligible gene, log-CPM expression is regressed on donor age, sex, experimental pool, log pseudobulk library size, and the first five genotype principal components. The residuals are standardized to unit sample standard deviation and form the adjusted expression matrix used throughout the subsequent screens.

C.3 Marginal cis-eQTL screen

Genotype calls are coded as allele dosages 0, 1, and 2. A variant is eligible if it lies on an autosome, has minor-allele frequency at least 5%, and has genotype missingness at most 2%. Missing dosages

are replaced by the variant-specific donor mean. Each eligible dosage is residualized with respect to the same covariate design used for expression and is then standardized. Applying the same adjustment set to expression and genotype is intended to reduce confounding by measured technical factors and ancestry structure.

For an eligible gene g, the cis region comprises all eligible variants between 200 kb upstream of the gene start and 200 kb downstream of the gene end. Every variant in this interval is tested marginally for association with the adjusted expression of g. If K_g variants are tested, the gene-level p-value is

$$p_g^{\mathrm{cis}} = \min\left\{1,\ K_g \min_{1\le k\le K_g} p_{gk}\right\},$$

where p_{gk} is the marginal association p-value for cis variant k. Thus, multiplicity within a gene is addressed by a Bonferroni correction of its minimum variant-level p-value. The values p_g^{cis} are then adjusted across genes using the Benjamini–Hochberg procedure, and genes with false discovery rate at most .05 are retained. This screen identifies 2,350 significant cis-eGenes.

C.4 Conditional selection of cis instruments

Conditionally independent cis variants are selected separately within each of the 2,350 significant cis-eGenes. The procedure begins with the most significant eligible cis variant. At every subsequent step, the gene-expression profile and all remaining candidate dosages are residualized with respect to the variants already selected for that gene. The remaining variant with the smallest conditional p-value is retained only if

$$K_g p_{gk}^{\mathrm{cond}} \le .05,$$

where K_g is the original number of eligible cis variants for gene g. After a variant is selected, any remaining variant having squared dosage correlation $r^2 > .05$ with it is removed from consideration. Selection terminates when no remaining variant satisfies the gene-wide criterion or when three variants have been retained.

A gene is eligible as a candidate latent regulator only if at least two conditionally selected cis variants remain. This requirement reduces the 2,350 cis-eGenes to 514 candidate genes. It does not require exactly two instruments: as many as three are allowed, although each of the three genes ultimately used in the masked illustration has exactly two. At this stage, the selected variants retain their known cis-gene assignments; those assignments are used to construct the full-information empirical reference but are withheld from the IV total-effect clustering step in the subsequent masked analysis.

C.5 Split-replicated trans reach

For each of the 514 candidate genes, its conditionally selected cis SNPs are combined into a single genetic score. Let G_g denote the adjusted donor-by-instrument matrix for gene g, and let Y_g denote its adjusted expression. The first-stage coefficient is estimated by least squares, $\widehat{\alpha}_g = (G_g^\top G_g)^{-1} G_g^\top Y_g$, and the standardized fitted genetic component

$$S_g = G_g \widehat{\alpha}_g$$

is used as the gene-specific score. This score is used only for the trans-reach and full-information network analyses; the final masked analysis retains the individual SNP dosages as separate columns of W.

For each candidate gene g, S_g is tested against every eligible expression gene on a different chromosome from g. Restricting the trans screen to different chromosomes avoids labeling local linkage disequilibrium or extended cis effects as trans associations. The full-sample p-values for the eligible trans genes are adjusted by the Benjamini–Hochberg procedure. A trans association is declared split-replicated if it has full-sample false discovery rate at most .05, has nominal $p < .05$ in each of two sex-stratified donor halves, and has the same effect sign in the full sample and both halves.

The trans-reach screen has two distinct outputs: a set of reach-filtered candidate regulators and a set of split-replicated regulator–response pairs. The regulator is drawn from the 514 genes with at least two conditionally independent cis instruments, whereas its potential response is drawn from the broader pool of 8,275 eligible expression genes.

Output 1: reach-filtered candidate regulators. A candidate regulator is retained if its genetic score has at least one split-replicated trans association. Among the 514 candidates, 60 have at least one full-sample trans association at false discovery rate .05, and 58 have at least one association that also satisfies the two-half replication criterion. These 58 genes form the regulator-node set used in the genetically anchored reference network described below. They will be considered as candidate latent variables in the masked analysis and their cis SNPs will be used as instruments to identify them.

Output 2: split-replicated regulator–response pairs. Separately, we retain each individual pair (g, h) for which the genetic score of candidate regulator g has a split-replicated trans association with the expression of gene h. The response gene h need not belong to the 58-gene regulator set and need not possess qualifying cis instruments. These pairs therefore provide a separate collection of genetically supported trans responses that can be considered as candidate observed measurements in the masked analysis. For example, *CTD-2054N24.2* is retained as a split-replicated trans response of *CYB5A*, although *CTD-2054N24.2* is not itself an eligible regulator node in the reference network.

C.6 Genetically anchored empirical reference network

The full-information reference network is fitted to the 58 reach-filtered regulators while their expression profiles remain observed. For every candidate gene g, the network contains both its adjusted expression Y_g and its SNP-derived score S_g. We fit a Gaussian DAG by BIC-scored hill climbing with 20 random restarts, perturbation size five, and a maximum of eight parents per node. Genetic scores are constrained to be exogenous anchors: the edge $S_g \to Y_g$ is required, all expression-to-score and score-to-score edges are forbidden, and S_g is forbidden from pointing directly to any nonassigned gene. These restrictions encode the cis assignments used to build the empirical reference and exclude modeled direct horizontal pleiotropy by construction.

Network stability is assessed with 100 donor bootstrap samples. Within each bootstrap sample, the constrained hill-climbing fit uses five random restarts and perturbation size three. For each gene-to-gene adjacency, bootstrap strength is the proportion of bootstrap networks containing an edge between the two genes, and conditional direction is the proportion of those appearances having the reported orientation. A directed gene edge is retained when both bootstrap strength and conditional direction are at least .80. The full-data fit contains 187 gene-to-gene edges; stability filtering retains 23 directed edges.

Finally, we select the connected three-gene Z subgraph with strong signal, *CYB5A*←*ENO1*→*ANXA6*. The seven observed measurements X come from two sources. Four are stable directed children in the full-information reference network: *ENO1*→*SERPINB1*, *ENO1*→*DBNL*, *ENO1*→*CD53*, and *ANXA6*→*CARD8*. The other three are split-replicated trans responses from Output 2 of Section C.5: *ENO1*→*FKBP2*, *CYB5A*→*CTD-2054N24.2*, and *ANXA6*→*TMED8*. These latter pairs are genetically supported response proxies rather than edges in the reference network. In the masked analysis, the three regulator genes are hidden, while their six cis SNPs W (shown in Table 5) and the seven response genes X remain observed. The goal is to recover the three latent variables, their causal relationships, and their relationships with the observed responses, using the full-information construction as the empirical benchmark.

Table 5: Conditionally selected cis instruments in the masked OneK1K illustration. Target labels are hidden from IV total-effect clustering and revealed only for evaluation.

Masked gene	cis instrument 1	cis instrument 2
ENO1	rs2765512	rs12135416
CYB5A	rs17229338	rs1788565
ANXA6	GSA-rs990995	rs728052