DISCRETE SCORE MATCHING ENABLES CAUSAL DISCOVERY FROM COUNT DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Count data pose a challenge for score-matching-based causal discovery: derivatives are unavailable, and simply replacing them with finite differences does not generally suffice for causal discovery. We generalize SCORE's constant-curvature criterion (Rolland et al., 2022) by conditioning on the node's value, yielding the conditional curvature score (CCS) for ordering. We also extend curvature-based parent recovery through the off-diagonal curvature score (OCS), enabling directed acyclic graph (DAG) recovery with both scores constructed from score functions for continuous data and concrete scores for counts. In the bivariate setting, zero CCS exactly characterizes a semiparametric generalized linear model (GLM) conditional form in which the conditional family need not be specified in advance, unlike in classical GLMs. For bivariate semiparametric GLM DAGs under our regularity condition, canonical-parameter nonlinearity is necessary and sufficient for identifiability. In multivariate DAGs, this nonlinearity enables DAG recovery through CCS and OCS. Our framework identifies a new class of semiparametric GLM DAGs that strictly contains the nonlinear Gaussian ANM class identified by SCORE. We introduce DISCO (*DI*screte *SCO*re), a count-DAG recovery algorithm that estimates CCS and OCS using discrete diffusion. Experiments demonstrate accurate DAG recovery across Poisson, negative binomial, binomial, and mixed-family settings, as well as scalability to 1,000-node DAGs on a single GPU.

1 INTRODUCTION

Many causal-discovery problems are naturally count-valued: gene-regulatory analysis from RNA-seq counts (Love et al., 2014), community structure from species abundances (Stoklosa et al., 2022), and topic structure from word counts (Blei et al., 2003). Such counts rarely follow a single distributional family: sequencing studies model expression counts as negative binomial (NB) but allele counts as binomial (Love et al., 2014; Robinson et al., 2010; Castel et al., 2015), and ecological surveys mix Poisson and overdispersed NB (Stoklosa et al., 2022).

Order-based methods provide a scalable alternative to DAG search by estimating a causal order and then selecting parents (Ghoshal and Honorio, 2018; Gao et al., 2020; Peters et al., 2014). Score-matching-based methods follow this approach using log-density derivatives (Rolland et al., 2022; Montagna et al., 2023a;b; Sanchez et al., 2023). Concrete score matching extends score estimation to discrete data (Meng et al., 2022), but finite-difference curvature need not satisfy SCORE's sink criterion: even at a sink, the node's conditional density can contribute curvature that varies with its value. We call this contribution *own-factor curvature*.

Existing count-DAG methods obtain identifiability by exploiting properties of specified conditional families, such as Poisson or known quadratic variance function (QVF) families (Park and Raskutti, 2015; 2018). More flexible conditionally parametric models allow general parent effects but still specify the family itself (Bodik and Chavez-Demoulin, 2025). Their guarantees consequently depend on choosing appropriate conditional families. Extending score-matching-based causal discovery to counts without prespecified families therefore requires an ordering criterion that accommodates unknown own-factor curvature.

We address these limitations by generalizing SCORE's constant-curvature criterion through the CCS and extending curvature-based parent recovery through the OCS. Our *conditional curvature score* (CCS) removes the own-factor contribution to curvature variance while retaining variation induced

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.

by children. The CCS is obtained from derivatives of the continuous score or finite differences of log concrete-score entries.

We establish identifiability for a new class of semiparametric GLM DAGs. In the bivariate setting, zero CCS exactly characterizes a semiparametric GLM conditional form without prespecifying the conditional family. Within this class, canonical-parameter nonlinearity is necessary and sufficient for identifiability. In multivariate settings, this nonlinearity enables the CCS to recover a causal order, after which the *off-diagonal curvature score* (OCS) recovers parents. On continuous data, the identifiable class strictly contains SCORE's nonlinear Gaussian additive noise models (ANMs).

We introduce DISCO (*DI*screte *SCO*re), a scalable count-DAG recovery algorithm that estimates the CCS and OCS. We first estimate joint concrete scores using score-entropy discrete diffusion, then project them to the marginal concrete scores needed to compute the CCS and OCS. With OCS-based parent selection, a rank transform makes the recovered DAG invariant to strictly increasing coordinate-wise transformations.

Our contributions are as follows.

- **Discrete score matching for count-DAG discovery.** We generalize SCORE's constant-curvature principle to count data through the conditional curvature score (CCS), which identifies sinks despite node-specific own-factor curvature. We also extend curvature-based parent recovery through the OCS.
- **A new identifiable class of semiparametric GLM DAGs.** We derive the semiparametric GLM conditional form from the bivariate zero-CCS condition without prespecifying the family. Canonical-parameter nonlinearity is necessary and sufficient for bivariate identifiability. This nonlinearity also yields multivariate DAG identifiability. In the continuous setting, the resulting identifiable class strictly contains the nonlinear Gaussian ANM class identified by SCORE.
- **Scalable count-DAG recovery with discrete diffusion.** We develop DISCO, which uses score-entropy discrete diffusion to estimate joint concrete scores and projects them to the marginal concrete scores needed to compute the CCS and OCS. DISCO scales to 1,000-node count DAGs on a single GPU.

2 Background

2.1 Related Work

Score-matching-based causal discovery. For a continuous density p, the score function is $s(x) = \nabla_x \log p(x)$ (Hyvärinen, 2005). SCORE identifies sinks in nonlinear Gaussian ANMs through the constant-curvature criterion, $\mathrm{Var}[\partial_j s_j(X)] = 0$ (Rolland et al., 2022). Subsequent work allows non-Gaussian noise (NoGAM; Montagna et al., 2023b), improves Hessian estimation (DiffAN and SciNO; Sanchez et al., 2023; Kang et al., 2025), and uses off-diagonal Hessian components for parent recovery (DAS; Montagna et al., 2023a).

For discrete causal discovery, Vo et al. (2026) use a marginalization-based generalized score (Lyu, 2009) to recover a causal order under a non-decreasing conditional-randomness assumption. In contrast, we use concrete scores (Meng et al., 2022), estimated through discrete diffusion with SEDD's score-entropy objective (Lou et al., 2024), and establish DAG identifiability within our semiparametric GLM class under the stated edgewise nonlinearity and regularity conditions, without requiring a cross-node ordering of conditional randomness.

Count-DAG identifiability. Count-DAG identifiability typically relies on distribution-specific conditional structure: Poisson and QVF DAGs assume known conditional variance structure (Park and Raskutti, 2015; Park and Park, 2019b; Park and Raskutti, 2018), while conditionally parametric causal models still prespecify the conditional family (Bodik and Chavez-Demoulin, 2025). Our semiparametric GLM DAG class leaves conditional families, links, and canonical-parameter functions unspecified.

2.2 NOTATION

Let $G = (V, E)$ be a directed acyclic graph with $V = [d] := \{1, \dots, d\}$. Let $\mathrm{Pa}_G(j)$ and $\mathrm{Ch}_G(j)$ denote the parents and children of node j in G. A *source* is a node j with $\mathrm{Pa}_G(j) = \emptyset$, a *non-source* is a node j with $\mathrm{Pa}_G(j) \neq \emptyset$, and a *sink* is a node j with $\mathrm{Ch}_G(j) = \emptyset$. For $R \subseteq V$, G_R denotes the induced subgraph of G on R. A *causal order* is a permutation $\pi : V \to [d]$ with $\pi(j) < \pi(i)$ for every edge $j \to i$; we also call it an ordering of G.

Let P be the distribution of $X = (X_1, \dots, X_d)$. For each $j \in V$, the coordinate X_j takes values in $\mathcal{X}_j \subseteq \mathbb{R}$ and has a dominating measure ν_j (counting or Lebesgue). With $\nu = \bigotimes_{j=1}^{d} \nu_j$, write $p = dP/d\nu$ for the density; throughout, this term includes probability mass functions. The set $\mathcal{X}_j$ is the declared effective support. We call X_j a *count coordinate* when $\mathcal{X}_j$ is a finite or countably infinite set of consecutive integers. Define $\mathring{\mathcal{X}}_j = \mathrm{int}(\mathcal{X}_j)$ for a continuous coordinate and $\mathring{\mathcal{X}}_j = \mathcal{X}_j$ for a count coordinate.

For $A \subseteq V$, write $X_A = (X_i)_{i \in A}$, $\mathcal{X}_A = \prod_{i \in A} \mathcal{X}_i$, and $\mathring{\mathcal{X}}_A = \prod_{i \in A} \mathring{\mathcal{X}}_i$. For $j \in A$, $x_A \in \mathcal{X}_A$, and $y \in \mathcal{X}_j$, let $x_A^{j \to y}$ be the vector obtained by replacing coordinate j with y. For $R \subseteq V$, let P_R denote the marginal distribution of X_R and p_R its density.

We use a single coordinate operator $\mathcal{D}_j$ for both continuous and count coordinates. With e_j the jth unit vector,

$$\big(\mathcal{D}_j f,\ \mathcal{D}_j^2 f\big)(x) = \begin{cases} \big(\partial_{x_j} f(x),\ \partial_{x_j}^2 f(x)\big), & X_j \text{ continuous}, \\ \big(f(x{+}e_j) - f(x),\ f(x{+}2e_j) - 2f(x{+}e_j) + f(x)\big), & X_j \text{ count}. \end{cases}$$

On a count coordinate these are *partial* operators: every expectation or conditional variance involving a finite-difference curvature is restricted to states where all required shifted values remain in the support. Appendix A gives the formal domains.

3 CURVATURE SCORES

We introduce two curvature scores: the *conditional curvature score* (CCS) for ordering and the *off-diagonal curvature score* (OCS) for parent recovery. Define the diagonal and off-diagonal log-density curvatures $H_j := \mathcal{D}_j^2 \log p$ and $H_{ij} := \mathcal{D}_i \mathcal{D}_j \log p$ for $i \neq j$. For the marginal distribution P_R, we denote the corresponding curvatures by $H_{j,R}$ and $H_{ij,R}$.

For node j, we ask whether its log-density curvature still varies with the other variables after fixing X_j. The conditional curvature score measures this remaining variation.

Definition 1 (Conditional curvature score)**.** For node j with $H_j \in L^2$,

$$S_j = \mathbb{E}\big[\mathrm{Var}\{\, H_j(X) \mid X_j \,\}\big]. \tag{1}$$

For the marginal distribution P_R, write $S_j(R) = \mathbb{E}_{P_R}[\mathrm{Var}_{P_R}\{H_{j,R}(X_R) \mid X_j\}]$, so $S_j = S_j(V)$, with the same admissible-domain convention. The CCS vanishes exactly when $H_j(X)$ is almost surely a function of X_j alone. By contrast, the *constant-curvature criterion* requires $H_j(X)$ to be almost surely constant, equivalently $\mathrm{Var}[H_j(X)] = 0$. For continuous data, $H_j = \partial_j s_j$ is a diagonal component of the log-density Hessian, and this is SCORE's sink criterion. The CCS therefore allows curvature to vary across values of X_j. Section 5 establishes when the CCS separates sinks from nonsinks.

To define the OCS, fix a causal order π of G. For each node i, define its predecessors by $\mathrm{Pred}(i) = \{j : \pi(j) < \pi(i)\}$ and let $R_i = \{i\} \cup \mathrm{Pred}(i)$. The OCS is computed under the marginal distribution P_{R_i}.

Definition 2 (Off-diagonal curvature score)**.** For a node i and a predecessor $j \in \mathrm{Pred}(i)$ with $H_{ij,R_i} \in L^1$,

$$\mathrm{OCS}_{j \to i} = \mathbb{E}\big[|H_{ij,R_i}(X_{R_i})|\big]. \tag{2}$$

For count coordinates, H_{ij,R_i} records how the log concrete-score entry for a shift in coordinate j changes when X_i steps by one level. The OCS averages its absolute magnitude under P_{R_i}.

4 SEMIPARAMETRIC GLM DAGS

A classical generalized linear model (GLM) specifies a conditional family and a link function that maps the conditional mean to a linear predictor (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). Semiparametric GLMs relax the requirement to prespecify the conditional family (Rathouz and Gao, 2009; Yang et al., 2018). This conditional form naturally arises from the zero-CCS condition in Theorem 5, motivating the DAG class introduced below.

Definition 3 (Semiparametric GLM DAG)**.** A distribution P follows a *semiparametric GLM DAG* on G if its density factorizes as $p(x) = \prod_{j=1}^{d} p_j(x_j \mid x_{\mathrm{Pa}_G(j)})$. Source marginal densities p_j on $\mathcal{X}_j$ are arbitrary and need not belong to an exponential family. For every non-source node j, its conditional density has the form

$$p_j(x_j \mid x_{\mathrm{Pa}_G(j)}) = b_j(x_j) \exp\{x_j\, \eta_j(x_{\mathrm{Pa}_G(j)}) - A_j(\eta_j(x_{\mathrm{Pa}_G(j)}))\}, \tag{3}$$

where η_j is an unspecified canonical-parameter function and $A_j(t) := \log \int_{\mathcal{X}_j} b_j(u) e^{ut}\, d\nu_j(u)$. The support $\mathcal{X}_j$ and base function b_j are unspecified but do not depend on the parent values. Additionally, each η_j depends nontrivially on every parent.

The support $\mathcal{X}_j$ and base function b_j determine the conditional family, while η_j captures all parent dependence. The canonical-parameter function η_j can be represented through a strictly increasing link g_j and an index f_j:

$$g_j\big(\mathbb{E}[X_j \mid X_{\mathrm{Pa}_G(j)}]\big) = f_j(X_{\mathrm{Pa}_G(j)}), \qquad f_j := g_j \circ A_j' \circ \eta_j.$$

This link–index pair is nonunique, so we state all structural assumptions directly in terms of η_j.

Definition 4 (Nonlinear semiparametric GLM DAG)**.** A semiparametric GLM DAG is *nonlinear* if, for every edge $k \to i$, there are fixed values z of the other parents in the product-support interior such that $\eta_i(\cdot, z)$ is nonlinear on $\mathring{\mathcal{X}}_k$. Equivalently, $\mathcal{D}_k^2 \eta_i \not\equiv 0$ on the corresponding derivative or finite-difference domain.

For Gaussian conditionals with fixed variance, canonical-parameter nonlinearity coincides with the mechanism nonlinearity used by SCORE (Rolland et al., 2022). Definition 4 extends this restriction to other conditional families. Parent-dependent nuisance parameters lie outside the class, for example a negative binomial whose size depends on the parents.

Table 1: Example conditional distributions, with index $f = f_j(x_{\mathrm{Pa}_G(j)})$ and logistic function sigmoid. Noncanonical links can yield nonlinear η even for affine f. NB and Gamma use the mean as their second parameter, with fixed size r and shape α, respectively.

Conditional $X_j \mid x_{\mathrm{Pa}_G(j)}$	Conditional family	Support	Link g
$\mathrm{Poisson}(\exp f)$	Poisson	$\mathbb{N}_0$	log
$\mathrm{Poisson}(\mathrm{softplus}\, f)$	Poisson	$\mathbb{N}_0$	$\mathrm{softplus}^{-1}$
$\mathrm{NB}(r, \exp f)$	NB(r)	$\mathbb{N}_0$	log
$\mathrm{Binomial}(M, \mathrm{sigmoid}(f))$	Binomial(M)	$\{0, \dots, M\}$	$\mu \mapsto \mathrm{logit}(\mu/M)$
$\mathcal{N}(f, \tau^2)$	Gaussian	$\mathbb{R}$	identity
$\mathrm{Gamma}(\alpha, \exp f)$	Gamma(α)	$(0, \infty)$	log

Regularity conditions. We assume strictly positive densities on product-support interiors and at least three support values for count coordinates. The results in Section 5 are stated under the corresponding regularity conditions in Appendix B.

5 CURVATURE CHARACTERIZATION AND DAG IDENTIFIABILITY

5.1 BIVARIATE CHARACTERIZATION AND IDENTIFIABILITY

For both continuous and count variables, we first characterize the conditional distributions with zero CCS. We then establish when the causal direction is identifiable within the resulting model class.

Theorem 5 (Bivariate characterization by the CCS)**.** *$S_Y = 0$ if and only if $p(y \mid x)$ has the semiparametric GLM conditional form in* (3)*.*

If $S_Y = 0$, the log-density curvature in y depends only on y. Integrating derivatives or summing finite differences gives $\log p(x, y) = a(x) + \log b_Y(y) + y\eta(x)$, which yields this conditional form after normalization.

Theorem 6 (Bivariate identifiability characterization)**.** *Let P have a semiparametric GLM representation with edge $X \to Y$, with Condition R1 (affine reversibility; Appendix B.2) imposed in the affine case. Within this model class,*

$$X \to Y \text{ is identifiable} \quad \iff \quad \eta \text{ is nonlinear on } \mathring{\mathcal{X}}_X.$$

To relate this characterization to the CCS, consider $X \to Y$. The child conditional satisfies

$$\log p(y \mid x) = \underbrace{\log b_Y(y)}_{\text{family-specific term}} + \underbrace{y\eta(x) - A_Y(\eta(x))}_{\text{parent contribution, affine in } y}.$$

Taking a second derivative or second difference in y removes the affine parent contribution, leaving $H_Y = \kappa_Y(Y)$, where $\kappa_Y(y) := \mathcal{D}_Y^2 \log b_Y(y)$ is the curvature contributed by the child's conditional density. This curvature need not be constant: for a negative binomial conditional with fixed size $r > 0$, $\kappa_r(y) = \log \frac{(y+r+1)(y+1)}{(y+r)(y+2)}$, which is zero for $r = 1$ and nonconstant otherwise. Thus, simply replacing derivatives with finite differences in SCORE does not ensure constant curvature at a sink. Conditioning on Y, however, gives $S_Y = 0$.

For the parent X, the curvature separates as

$$H_X = \kappa_X(X) + \underbrace{\mathcal{D}_X^2\{Y\eta(X) - A_Y(\eta(X))\}}_{\psi_X(X,Y)\text{: residual curvature}}.$$

Here $\kappa_X(x) := \mathcal{D}_X^2 \log p_X(x)$ is the source's own-factor curvature. Given X, both $\kappa_X(X)$ and the log-partition term are fixed, leaving only $Y\mathcal{D}_X^2\eta(X)$ to vary. Hence $S_X = \mathbb{E}[\{\mathcal{D}_X^2\eta(X)\}^2 \operatorname{Var}(Y \mid X)]$. Since $\operatorname{Var}(Y \mid X) > 0$, $S_Y = 0 < S_X$ exactly when η is nonlinear.

Remark 7 (Translation to link and index). Identifiability is determined by nonlinearity of the canonical-parameter function η, not by the link or index in isolation. Appendix D.2 gives the link–index characterization.

5.2 Multivariate Identifiability

We extend the CCS argument to recover a causal order in multivariate DAGs and use the OCS to recover parents. A remaining set R obtained by repeatedly removing sinks is *ancestral*. Its marginal P_R retains the semiparametric GLM factorization on G_R and edgewise nonlinearity.

Theorem 8 (DAG recovery by the CCS and OCS)**.** *Let P follow a semiparametric GLM DAG on G.*

(i) If P follows a nonlinear semiparametric GLM DAG, then, for every ancestral remaining set R and $j \in R$, j is a sink of G_R if and only if $S_j(R) = 0$.

(ii) Given a causal order and $j \in \operatorname{Pred}(i)$, $j \in \operatorname{Pa}_G(i)$ if and only if $\operatorname{OCS}_{j\to i} > 0$.

On an ancestral remaining set R, the bivariate decomposition extends to $H_{j,R}(x_R) = \kappa_j(x_j) + \psi_{j,R}(x_R)$, where $\kappa_j(x_j) := \mathcal{D}_j^2 \log p_j(x_j \mid x_{\operatorname{Pa}_G(j)})$ is the own-factor curvature and $\psi_{j,R}$ is the residual curvature from the child contributions. Conditioning on X_j makes $\kappa_j(X_j)$ fixed, so the CCS measures the conditional variation of $\psi_{j,R}$. This residual vanishes at a sink; at a nonsink, edgewise nonlinearity makes its averaged conditional variance positive.

For parent recovery, node i is a sink of G_{R_i}. Under P_{R_i}, $H_{ij,R_i} = \mathcal{D}_j\eta_i$ if $j \in \operatorname{Pa}_G(i)$ and $H_{ij,R_i} = 0$ otherwise. Thus a nonconstant parent effect suffices for OCS; nonlinearity is not required.

Corollary 9 (DAG identifiability)**.** *Let P follow a nonlinear semiparametric GLM DAG on G. Then G is identifiable from P within this nonlinear semiparametric GLM class.*

This class strictly contains SCORE's nonlinear Gaussian ANMs. For example, the Gamma model in Appendix D.1 is identifiable despite its nonconstant sink curvature $H_Y = -(\alpha - 1)/Y^2$, while $S_Y = 0$.

6 DISCO ALGORITHM

DISCO consists of concrete-score estimation, CCS-based ordering, and parent selection given the estimated order. After applying the rank transform, we estimate joint concrete scores using score-entropy discrete diffusion and project them to the marginal concrete scores needed for CCS-based ordering and OCS-based parent selection. Parent recovery can use CAM pruning (Bühlmann et al., 2014), PCM-GAM (Lundborg et al., 2024), or the OCS-based rule. The OCS-based rule reuses the fitted score networks, providing a scalable alternative to regression- and testing-based parent selection. Appendix I.2 gives detailed comparisons of these methods. Appendix H.2.8 presents a continuous-data version and a comparison with DiffAN (Sanchez et al., 2023).

6.1 CONCRETE-SCORE ESTIMATION

We first apply the rank transform, mapping the distinct training values of each count variable j to the empirical rank grid $\{0, \dots, K_j\}$, where $K_j + 1$ is the number of distinct training values. We then estimate concrete scores on the transformed data using discrete diffusion.

The rank transform makes neighboring differences invariant to strictly increasing transformations on the support of each variable. For estimation, X, $\mathcal{X}_j$, and p_R refer to the transformed data, rank grid, and corresponding marginal density.

For $j \in R \subseteq V$ and $y \in \mathcal{X}_j$, the concrete-score entry is the probability ratio $s^R_{j,y}(x_R) := p_R(x_R^{j \to y})/p_R(x_R)$. On the transformed data, we train a joint concrete-score network $s_\theta(x; \sigma)$ to estimate the concrete score of the noise-corrupted joint distribution by minimizing a neighbor-restricted denoising score-entropy objective (Lou et al., 2024). Here σ denotes the diffusion noise level. As $\sigma \to 0$, the concrete score of the noise-corrupted distribution converges to that of the data distribution.

Marginal concrete-score projection. Ordering requires concrete scores of successively smaller marginals. These scores cannot generally be obtained by simply restricting the joint concrete score to the remaining variables. The following identity estimates them via conditional-mean regression.

Proposition 10 (Marginalization of the concrete score)**.** *Let $j \in R \subseteq V$ with count-valued X_j. For each neighboring replacement $y \in \mathcal{X}_j$, the concrete-score entries satisfy*

$$s^R_{j,y}(x_R) = \mathbb{E}\big[s^V_{j,y}(X_V) \mid X_R = x_R\big] .$$

Applying Proposition 10 to the noise-corrupted distribution gives the same conditional-mean identity at each noise level σ. Using predictions from $s_\theta(x; \sigma)$ as targets, we learn a single mask-conditioned projection network with remaining set R as an input. The resulting estimator $\widehat{s}^R_{\theta,\phi}(x_R; \sigma)$, with projection parameters ϕ, provides marginal concrete scores for different sets R without refitting. Appendices F and G provide the theoretical analysis and implementation details, respectively.

6.2 CCS AND OCS COMPUTATION

We compute the CCS and OCS at near-zero noise levels on held-out folds, using estimators fitted on the other folds. Fold indices on fitted quantities are suppressed.

Let $\widehat{\ell}_{j,R}(x_R; \sigma)$ denote the projection network's predicted log probability ratio for increasing x_j by one. Where all required neighbors lie in the grid, it equals $\log\big[\widehat{s}^R_{\theta,\phi}(x_R; \sigma)\big]_{j,\, x_j+1}$, and for $i, j \in R$ we estimate curvature by

$$\widehat{H}_{ij,R}(x_R; \sigma) := \widehat{\ell}_{j,R}(x_R + e_i; \sigma) - \widehat{\ell}_{j,R}(x_R; \sigma). \tag{4}$$

For the CCS, we first average diagonal curvature estimates over a finite set of noise levels $\mathcal{G}_\sigma$, without taking absolute values: $\widehat{H}_{j,R} := |\mathcal{G}_\sigma|^{-1} \sum_{\sigma \in \mathcal{G}_\sigma} \widehat{H}_{jj,R}(\cdot; \sigma)$.

To estimate the CCS, we group evaluation observations by their rank value of X_j, compute the sample variance of $\widehat{H}_{j,R}$ within each group, and average these variances with sample-size weights.

Let $\mathcal{J}_{j,v} = \{a : X_{aj} = v\}$ and $n_v = |\mathcal{J}_{j,v}|$. Keeping the levels $\mathcal{V}_j = \{v : n_v \geq n_{\min}\}$ gives

$$\widehat{S}_j(R) = \frac{1}{\sum_{v\in\mathcal{V}_j} n_v} \sum_{v\in\mathcal{V}_j} n_v \, \widehat{\mathrm{Var}}\Big(\{\widehat{H}_{j,R}(X_a) : a \in \mathcal{J}_{j,v}\}\Big). \tag{5}$$

Here $\widehat{\mathrm{Var}}$ is the sample variance.

We compute the OCS separately on each held-out fold. Given an estimated causal order $\widehat{\pi}$, define $\mathrm{Pred}_{\widehat{\pi}}(i) = \{j : \widehat{\pi}(j) < \widehat{\pi}(i)\}$ and $\widehat{R}_i = \{i\} \cup \mathrm{Pred}_{\widehat{\pi}}(i)$. Let $\mathcal{I}_1, \ldots, \mathcal{I}_B$ denote these folds. For candidate edge $j \to i$, let $\mathcal{I}_b^{ij}$ index the evaluation observations used for that edge. For nonempty $\mathcal{I}_b^{ij}$, we take absolute values before averaging over noise levels and observations:

$$\widehat{\mathrm{OCS}}_{j\to i}^{(b)} = \frac{1}{|\mathcal{I}_b^{ij}|\,|\mathcal{G}_\sigma|} \sum_{a\in\mathcal{I}_b^{ij}} \sum_{\sigma\in\mathcal{G}_\sigma} \big|\widehat{H}_{ij,\widehat{R}_i}(X_{a,\widehat{R}_i};\sigma)\big|. \tag{6}$$

6.3 DAG Recovery

Starting from $R = V$, DISCO repeatedly removes a minimizer of $\widehat{S}_j(R)$ and assigns that node the last available position. Given this order, parents can be selected using CAM, PCM-GAM, or OCS. Algorithm 1 presents the OCS-based implementation.

For scalable parent selection, we use the OCS estimates without fitting additional regressions or performing conditional-independence tests. Estimated absolute curvatures can be positive even for nonparents. To account for this noise baseline, we standardize the OCS estimates for each target node and fold, obtaining $Z_{j\to i}^{(b)}$. We retain edges whose selection frequency $\Pi_{j\to i}(\tau_Z) := B^{-1}\sum_{b=1}^{B} \mathbf{1}\{Z_{j\to i}^{(b)} > \tau_Z\}$ is at least $\pi_{\min}$.

Algorithm 1 DISCO with OCS-based parent selection.

Input: marginal concrete-score estimator $\widehat{s}_{\theta,\phi}^{R}$ and evaluation folds $\{\mathcal{I}_b\}_{b=1}^{B}$

Ordering

$R \leftarrow V$
while $|R| > 1$ **do**
  Compute $\widehat{S}_j(R)$ for all $j \in R$
  $\widehat{j}_R \in \arg\min_{j\in R} \widehat{S}_j(R)$
  $\widehat{\pi}(\widehat{j}_R) \leftarrow |R|$
  $R \leftarrow R \setminus \{\widehat{j}_R\}$
end while
Assign the sole remaining node to position 1

Parent selection given $\widehat{\pi}$

for each node $i \in V$ **do**
  $\widehat{R}_i \leftarrow \{i\} \cup \mathrm{Pred}_{\widehat{\pi}}(i)$
  Compute the foldwise $\widehat{\mathrm{OCS}}_{j\to i}^{(b)}$
  Standardize and form $\Pi_{j\to i}(\tau_Z)$
  Select $j \to i$ if $\Pi_{j\to i}(\tau_Z) \geq \pi_{\min}$
end for
Output: DAG $\widehat{G}$

For the consistency result, we consider a variant of Algorithm 1 that uses consistent curvature estimates and selects parents by thresholding the raw OCS at a deterministic level $\lambda_n > 0$.

Theorem 11 (Plug-in Consistency of DISCO with raw OCS thresholding)**.** *Under the assumptions of Theorem 8 and Assumption 15, this variant recovers a causal order of G with probability tending to one as $n \to \infty$. If $\lambda_n \to 0$ and the curvature-estimation error ε_n from Assumption 15 satisfies $\varepsilon_n = o_p(\lambda_n)$, then* $\Pr\{\widehat{G} = G\} \longrightarrow 1$.

Transformation invariance. For the population result on count data, let T_j map each value in the ordered support of coordinate j to its support rank. The fitted map $\widehat{T}_j$ is its sample counterpart.

Proposition 12 (Invariance to strictly increasing transformations)**.** DISCO *with OCS-based parent selection returns the same graph under strictly increasing transformations of each variable when using the same seed.*

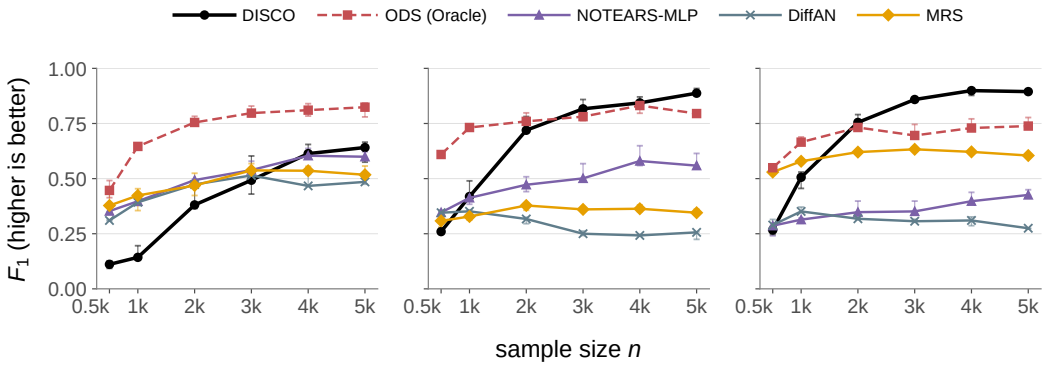
Proofs of the formal results in this section are provided in Appendix E.

Computational complexity. Following DiffAN (Sanchez et al., 2023), we fix the training epochs and evaluation sample size, and count each score evaluation at unit cost. With fixed hidden widths and mask draws per observation, the rank transform and training take $O(nd)$ for n training samples on consecutive counts. Ordering requires $O(d^2)$ shifted-state evaluations in $O(d)$ sequential batches. OCS-based parent selection uses $O(d)$ batches and $O(d^2)$ candidate-edge operations. These counts give $O(nd + d^2)$ for the OCS-based implementation; CAM and PCM-GAM require additional regressions or tests.

7 Numerical Experiments

We evaluate whether conditional curvature restores the ordering signal when nonconstant own-factor curvature invalidates the constant-curvature criterion, whether DISCO achieves family-agnostic count-DAG recovery within the semiparametric GLM DAG class, and whether it scales to large DAGs. We compare DISCO with DiffAN (Sanchez et al., 2023), NOTEARS-MLP (Zheng et al., 2020), ODS (Park and Raskutti, 2018), and MRS (Park and Park, 2019a). ODS (Oracle) receives the true conditional families and QVF parameters, whereas DISCO receives no family labels. For parent selection, we compare CAM's additive-regression pruning (Bühlmann et al., 2014) and PCM-GAM's conditional-mean independence testing (Lundborg et al., 2024). Unless otherwise stated, DISCO uses OCS-based parent selection. Experimental settings and evaluation metrics are given in Appendix H.1. Additional experiments and algorithm diagnostics appear in Appendices H and I.

DAG recovery versus sample size. Figure 1 compares DISCO with the baselines on DAGs with $d = 100$ as $n \in \{500, 1000, 2000, 3000, 4000, 5000\}$ increases: NB with $r = 1$ in panel (a), NB with $r = 6$ in panel (b), and the mixed NB $\to$ Bin $\to$ Poi configuration in panel (c). DISCO attains the highest median F_1 at larger sample sizes under $r = 6$ and mixed families, whereas ODS (Oracle) remains strongest under $r = 1$. As illustrated by the negative binomial example in Section 5.1, own-factor curvature vanishes when $r = 1$, so the constant-curvature criterion is already sufficient and estimating the more general conditional-curvature criterion can incur an additional finite-sample efficiency cost. Nevertheless, DISCO shows improved recovery at larger sample sizes. DiffAN shows no consistent gains with more data under $r = 6$ and mixed families, where the constant-curvature criterion fails because sink curvature is nonconstant.



(a) NB ($r = 1$) (b) NB ($r = 6$) (c) NB $\to$ Bin $\to$ Poi

Figure 1: End-to-end F_1 versus sample size on DAGs with $d = 100$: (a) NB with $r = 1$; (b) NB with $r = 6$; and (c) the mixed NB $\to$ Bin $\to$ Poi configuration. Points and error bars show the median and interquartile range. ODS (Oracle) uses the true conditional family and QVF parameters.

Conditional curvature ablation. We compare CCS and the constant-curvature criterion on the same sets of remaining nodes using the same fitted networks ($d = 50$, $n = 5000$). Table 2(a) shows that CCS improves the accuracy of identifying sinks, with larger gains for Poisson and negative binomial data than for binomial data. This improvement does not always increase full-DAG recovery accuracy: for binomial data, mean F_1 is 0.877 with CCS and 0.884 with the constant curvature.

Parent recovery given an estimated order. We compare OCS, CAM, and PCM-GAM given the same estimated DISCO order at $d = 50$ and $n = 5000$, using disjoint ordering and parent-recovery samples. CAM improves mean F_1 over OCS for NB and Bin, whereas OCS performs best for Poi (Table 2(b)). Appendix I.2 gives the test settings, SHD, and true-order comparisons.

Table 2: Ordering and parent-selection comparisons ($d = 50$, $n = 5000$). (a) Sink-selection accuracy: Constant denotes the constant-curvature criterion; gain is in percentage points. (b) Mean F_1 given the same estimated DISCO order.

(a) Sink selection (%)

Family	Constant	CCS	Gain
Poi	68.3	**81.7**	+13.4
NB	73.3	**93.3**	+20.0
Bin	80.0	**83.3**	+3.3

(b) Parent recovery (F_1)

Selector	Poi	NB	Bin
OCS	**0.858**	0.751	0.828
CAM	0.850	**0.823**	**0.898**
PCM-GAM	0.712	0.703	0.738

Scalability with graph size. Table 3 reports F_1 and total wall-clock time as d grows on Poisson DAGs with $n = 5000$. At $d = 1000$, DISCO attains median $F_1 = 0.7688$ in 57.4 minutes on a single GPU, whereas all baselines exceed the 240-minute time limit. This demonstrates practical scalability; the runtimes are not intended as a hardware-matched speed comparison.

Table 3: Scalability on Poisson DAGs ($n = 5000$; numerical entries are medians). *Timeout*: runtime exceeds 240 minutes.

	$d = 100$		$d = 200$		$d = 500$		$d = 1000$	
Method	F_1	Total time (min)	F_1	Total time (min)	F_1	Total time (min)	F_1	Total time (min)
DISCO	**0.915**	2.7	**0.835**	3.5	**0.769**	10.8	**0.769**	57.4
ODS (Oracle)	0.725	12.5	0.713	37.0	0.729	204.7	–	*Timeout*
MRS	0.786	3.5	0.763	10.8	0.742	96.3	–	*Timeout*
NOTEARS-MLP	0.467	174.1	–	*Timeout*	–	*Timeout*	–	*Timeout*
DiffAN	–	*Timeout*	–	*Timeout*	–	*Timeout*	–	*Timeout*

Family-agnostic DAG learning on real MLB data. We revisit the Lahman batting data previously analyzed for count-DAG learning by Park and Park (2019b), using $d = 17$ count variables from $n = 1{,}400$ player–seasons during 2019–2025. All methods in this diagnostic use the same q4-binned representation, chosen to increase observations per level for CCS estimation. We evaluate ordering against seven domain-informed accounting and containment relations among standard baseball statistics (Major League Baseball, n.d.). ODS requires a conditional-family specification, and its mean ordering agreement A_{top} with these references varies from 0.333 to 0.429 across family choices. In contrast, DISCO requires no family or link input and attains mean $A_{\text{top}} = 0.905$. Although the partial references do not define a ground-truth DAG, DISCO shows higher agreement with these domain-informed references than the three fixed-family ODS variants, without requiring a family choice. Appendix H.3 provides the complete diagnostics, including pairwise family-specification comparisons.

8 Conclusion

DISCO enables score-matching-based causal discovery from count data without prespecifying nodewise conditional families, relaxing a central assumption of existing identifiable count-DAG methods. The CCS and OCS provide ordering and parent-recovery criteria for nonlinear semiparametric GLM DAGs under the stated assumptions. A joint concrete-score network and marginal projection allow DISCO to estimate both criteria, with OCS-based recovery scaling to 1,000-node DAGs in our experiments. The main remaining theoretical challenge is to establish end-to-end consistency of DISCO, from finite-noise discrete-diffusion score estimation and marginal projection to curvature estimation and DAG recovery.

AI USE STATEMENT

Generative AI tools were used to assist with aspects of the proofs and experimental work. The authors reviewed the AI-assisted material and take responsibility for the final manuscript and results.

REPRODUCIBILITY STATEMENT

Implementation details, including preprocessing, training objectives, mask sampling, and parent selection, are provided in Appendix G. Appendix H specifies the data-generating mechanisms, baseline configurations, evaluation metrics, and replication protocols. Appendix I describes the algorithm diagnostics, and Appendix H.3 gives the Lahman cohort construction, paired cross-validation design, and reference sets. The appendix also contains the assumptions and proofs of the theoretical results.

REFERENCES

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

J. Bodik and V. Chavez-Demoulin. Identifiability of causal graphs under non-additive conditionally parametric causal models. *Journal of Machine Learning Research*, 26(264):1–55, 2025.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

S. E. Castel, A. Levy-Moonshine, P. Mohammadi, E. Banks, and T. Lappalainen. Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, 16:195, 2015. doi:10.1186/s13059-015-0762-6.

M. Gao, Y. Ding, and B. Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 11599–11611, 2020.

A. Ghoshal and J. Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, 2018.

A. Huang. Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, 17(6):359–380, 2017. doi:10.1177/1471082X17697749.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

J. Kang, S. Kim, C. Lee, D. Hwang, J. Chung, Y. Ko, S. Lee, S. Kim, and S. Lim. Score-informed neural operator for enhancing ordering-based causal discovery. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pages 113109–113151, 2025.

S. Lahman. Lahman baseball database. Society for American Baseball Research. `https://sabr.org/lahman-database/`, accessed September 22, 2026.

A. Lou, C. Meng, and S. Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 32819–32848, 2024.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

A. R. Lundborg, I. Kim, R. D. Shah, and R. J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6):2851–2878, 2024. doi:10.1214/24-AOS2447.

S. Lyu. Interpretation and generalization of score matching. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 359–366, 2009.

Major League Baseball. Glossary of standard statistics: At-bat, Hit, and Intentional walk. MLB.com, n.d. Accessed September 21, 2026.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 2nd edition, 1989.

C. Meng, K. Choi, J. Song, and S. Ermon. Concrete score matching: Generalized score matching for discrete data. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 34532–34545, 2022.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Scalable causal discovery with score matching. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning (CLeaR)*, volume 213 of *Proceedings of Machine Learning Research*, pages 752–771, 2023a.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning (CLeaR)*, volume 213 of *Proceedings of Machine Learning Research*, pages 726–751, 2023b.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

G. Park and H. Park. Identifiability of generalized hypergeometric distribution (GHD) directed acyclic graphical models. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 158–166, 2019a.

G. Park and S. Park. High-dimensional Poisson structural equation model learning via ℓ_1-regularized regression. *Journal of Machine Learning Research*, 20(95):1–41, 2019b.

G. Park and G. Raskutti. Learning large-scale Poisson DAG models based on overdispersion scoring. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pages 631–639, 2015.

G. Park and G. Raskutti. Learning quadratic variance function (QVF) DAG models via overdispersion scoring (ODS). *Journal of Machine Learning Research*, 18(224):1–44, 2018.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

P. J. Rathouz and L. Gao. Generalized linear models with unspecified reference distribution. *Biostatistics*, 10(2):205–218, 2009.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 18741–18753, 2022.

P. Sanchez, X. Liu, A. Q. O'Neil, and S. A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *International Conference on Learning Representations (ICLR)*, 2023.

J. Stoklosa, R. V. Blakey, and F. K. C. Hui. An overview of modern applications of negative binomial modelling in ecology and biodiversity. *Diversity*, 14(5):320, 2022.

V. Vo, T. Le, H. Zhao, E. V. Bonilla, and D. Phung. Ordering-based causal discovery via generalized score matching. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD)*, pages 4706–4717, 2026. doi:10.1145/3770855.3817672.

Z. Yang, Y. Ning, and H. Liu. On semiparametric exponential family graphical models. *Journal of Machine Learning Research*, 19(57):1–59, 2018.

K. Zhang, S. Zhu, M. Kalander, I. Ng, J. Ye, Z. Chen, and L. Pan. gCastle: A Python toolbox for causal discovery. *arXiv preprint arXiv:2111.15155*, 2021.

X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. P. Xing. Learning sparse nonparametric DAGs. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425, 2020.

APPENDIX CONTENTS

A Notation **14**
B Regularity Conditions **15**
B.1 Common Regularity 15
B.2 Semiparametric GLM Representations 15
B.3 Moment Conditions 15
C Proofs for Section 5 . . . . . **16**
C.1 Proof of Theorem 5 16
C.2 Proof of Theorem 6 16
C.3 Proof of Theorem 8 17
C.4 Proof of Corollary 9 18
D Additional Identifiability Results **18**
D.1 Strict Extension of SCORE 18
D.2 Link–Index Characterization 19
E Proofs for Section 6 . . . . . **19**
E.1 Proof of Proposition 10 19
E.2 Proof of Theorem 11 20
E.3 Proof of Proposition 12 22
F Theoretical Properties of Concrete-Score Estimation **22**
F.1 Denoising Score-Entropy Objective 22
F.2 Vanishing-Noise Limit 23
G DISCO Implementation **23**
G.1 Rank Transform 23
G.2 Joint Concrete-Score Training 23
G.3 Marginal Concrete-Score Projection 24
G.4 Curvature Estimation 24
G.5 Parent Selection 25
H Experimental Details and Additional Results **25**
H.1 Experimental Setup 25
H.2 Additional Benchmarks 26
H.3 Real Data: Lahman Batting Counts 31
I Algorithm Diagnostics **33**
I.1 Marginal Projection Methods 33
I.2 Parent Selection 34

A NOTATION

Table 4: Recurring symbols.

Symbol	Meaning (defined in)
$G = (V, E), \lvert V\rvert = d$	DAG and number of nodes (Sec. 2.2)
$\widehat{G}$	estimated DAG (Sec. 6.3)
$\mathrm{Pa}_G(j),\ \mathrm{Ch}_G(j)$	parents and children of j (a sink has $\mathrm{Ch}_G(j) = \emptyset$)
source node	a node j with $\mathrm{Pa}_G(j) = \emptyset$
$R \subseteq V$	remaining node set, ancestral when obtained by recursive sink removal
$G_R,\ P_R,\ p_R$	induced subgraph, marginal distribution of X_R, and its density
$p_{R,\sigma}$	density of the noise-corrupted R-marginal (App. F)
$\pi,\ \mathrm{Pred}(i),\ R_i$	causal order, predecessors $\{j : \pi(j) < \pi(i)\}$, and $R_i = \{i\} \cup \mathrm{Pred}(i)$ (Sec. 3)
$\widehat{\pi},\ \mathrm{Pred}_{\widehat{\pi}}(i),\ \widehat{R}_i$	estimated counterparts of the preceding row (Sec. 6.2)
$x_A^{j\to y}$	x_A with coordinate j reset to y (Sec. 2.2)
$\mathcal{D}_j,\ \mathcal{D}_j^2$	unit coordinate (second) difference, with ∂_j, ∂_j^2 if continuous
b_j	unspecified base function of a non-source conditional (Def. 3)
$g_j,\ f_j$	link and its corresponding index
η_j	canonical-parameter function encoding parent dependence
A_j	log-partition of a non-source conditional's natural exponential family
$\kappa_j(x_j)$	own-factor curvature of node j (Sec. 5.2)
$\psi_X,\ \psi_{j,R}$	residual curvature in the bivariate or marginal decomposition (Secs. 5.1–5.2)
$H_{j,R} = \mathcal{D}_j^2 \log p_R$	diagonal curvature of the marginal (Sec. 3)
$H_{ij,R} = \mathcal{D}_i\mathcal{D}_j \log p_R$	off-diagonal curvature (parent recovery)
$\mathcal{A}_j^{(2)},\ \mathcal{A}_{ij}^{(1,1)}$	admissible-difference events for $\mathcal{D}_j^2$, $\mathcal{D}_i\mathcal{D}_j$
$S_j(R)$	conditional curvature score (CCS), (1)
$\mathrm{OCS}_{j\to i}$	off-diagonal curvature score, (2)
$O_{j\to i}(R)$	OCS on a general admissible marginal in the consistency proof
T_j	population rank transform (Sec. 6.3)
$\widehat{T}_j$	fitted rank transform from the score-training split (Sec. 6.1)
$s_{j,y}^R$	concrete-score entry, neighboring when $y = x_j \pm 1$
s_θ	joint concrete-score network
$\widehat{s}_{\theta,\phi}^R$	R-marginal concrete-score estimate
$\widehat{\ell}_{j,R}(\cdot;\sigma)$	log-ratio prediction for increasing x_j by one (Sec. 6.2)
$\widehat{S}_j(R),\ \widehat{\mathrm{OCS}}_{j\to i}^{(b)}$	estimated CCS and foldwise estimated OCS (Sec. 6.2)

Admissible domains. A finite difference is defined only where all required shifted values remain in the support; we call this its *admissible domain*. For $r \in \{1, 2\}$, define

$$\mathcal{X}_j^{\circ,r} = \begin{cases} \{x_j \in \mathcal{X}_j : x_j, x_j + 1, \ldots, x_j + r \in \mathcal{X}_j\}, & X_j \text{ count}, \\ \mathring{\mathcal{X}}_j, & X_j \text{ continuous}. \end{cases}$$

Thus $\mathcal{D}_j^r$ requires $x_j \in \mathcal{X}_j^{\circ,r}$, and, for $i \neq j$, $\mathcal{D}_i\mathcal{D}_j$ requires $x_i \in \mathcal{X}_i^{\circ,1}$ and $x_j \in \mathcal{X}_j^{\circ,1}$. The corresponding curvature events are

$$\mathcal{A}_j^{(2)} = \{X_j \in \mathcal{X}_j^{\circ,2}\}, \qquad \mathcal{A}_{ij}^{(1,1)} = \{X_i \in \mathcal{X}_i^{\circ,1},\ X_j \in \mathcal{X}_j^{\circ,1}\}, \quad i \neq j.$$

Continuous coordinates impose only the support-interior restriction, which has probability one.

For these curvatures, population moments and almost-sure statements use the marginal distribution conditioned on the corresponding event. For the CCS, both the outer expectation and the conditional

variance use $P_R(\cdot \mid \mathcal{A}_j^{(2)})$; the OCS uses $P_{R_i}(\cdot \mid \mathcal{A}_{ij}^{(1,1)})$. If the event is determined by the conditioning variables, this restriction leaves the conditional distribution of the other variables unchanged.

B Regularity Conditions

B.1 Common Regularity

We assume finite, strictly positive densities on product-support interiors. Each coordinate support is a connected interval with nonempty interior or a consecutive integer set with at least three levels. For the distribution-level characterization, $\log p$ is jointly C^2 in the continuous coordinates for each fixed choice of count coordinates.

B.2 Semiparametric GLM Representations

For the semiparametric GLM representations in Definition 3, count supports are additionally bounded below. Source densities and non-source base functions are finite and strictly positive on their support interiors, and their logarithms are C^2 in continuous coordinates. Each η_j is finite and jointly C^2 in its continuous parent coordinates for each fixed choice of count parents. Required finite differences are finite on their admissible domains, and the canonical parameters lie in the interior of the canonical-parameter domain:

$$\eta_j(\mathring{\mathcal{X}}_{\mathrm{Pa}_G(j)}) \subseteq \operatorname{int} \Xi_j, \qquad \Xi_j := \{t \in \mathbb{R} : A_j(t) < \infty\}.$$

This interiority justifies differentiation under the integral, so A_j is smooth and $A_j''(t) = \mathrm{Var}_t(X_j) \in (0, \infty)$ on $\operatorname{int} \Xi_j$.

Condition R1 (Affine reversibility)**.** In the affine branch of Theorem 6, write $\eta(x) = \beta_0 + \beta_1 x$ with $\beta_1 \neq 0$, and define

$$\widetilde{b}_X(x) = p_X(x) e^{-A_Y(\beta_0+\beta_1 x)}, \qquad \widetilde{\Xi}_X = \left\{ t : \int_{\mathcal{X}_X} \widetilde{b}_X(x) e^{tx}\, d\nu_X(x) < \infty \right\}.$$

For count Y, require $\beta_1 e \in \operatorname{int} \widetilde{\Xi}_X$ at each support endpoint e. There are two endpoints for finite support and one for one-sided infinite support. No additional restriction is imposed for continuous Y.

Integrating out a sink removes its conditional density from the DAG factorization. Therefore an ancestral marginal is finite and strictly positive and retains the remaining conditional densities and parent sets. Edgewise nonlinearity is also preserved when present. The own-factor curvature is $\kappa_j = \mathcal{D}_j^2 \log b_j$ for a non-source and $\kappa_j = \mathcal{D}_j^2 \log p_j$ for a source, consistent with Section 5.2. Its curvature decomposition is

$$\begin{aligned} H_{j,R} &= \kappa_j(x_j) + \psi_{j,R}(x_R), \\ \psi_{j,R} &= \sum_{c \in \mathrm{Ch}_{G_R}(j)} \mathcal{D}_j^2 \Big\{ x_c \eta_c(x_{\mathrm{Pa}_{G_R}(c)}) - A_c\big(\eta_c(x_{\mathrm{Pa}_{G_R}(c)})\big) \Big\}. \end{aligned} \tag{7}$$

The child base functions do not depend on x_j and disappear under $\mathcal{D}_j^2$.

B.3 Moment Conditions

Score moments. Where the CCS is used, we assume $H_{j,R} \in L^2\Big(P_R(\cdot \mid \mathcal{A}_j^{(2)})\Big)$ for every evaluated ancestral set R and $j \in R$. In the bivariate setting, this means $H_Z \in L^2\Big(P(\cdot \mid \mathcal{A}_Z^{(2)})\Big)$ for $Z \in \{X, Y\}$. Where the OCS is used, we assume $H_{ij,R_i} \in L^1\Big(P_{R_i}(\cdot \mid \mathcal{A}_{ij}^{(1,1)})\Big)$ for every evaluated predecessor marginal R_i and $j \in \mathrm{Pred}(i)$.

Score-entropy regression. For each positive active target U, assume $\mathbb{E}[U \log^+ U] < \infty$, where $\log^+ u = \max\{\log u, 0\}$. This moment condition holds on finite population supports, but a finite empirical rank grid alone does not ensure it.

C PROOFS FOR SECTION 5

All derivatives, finite differences, and conditional moments below use the notation and admissible-event convention in Appendix A.

C.1 PROOF OF THEOREM 5

Proof. Write $F(x, y) = \log p(x, y)$. By Definition 1, $S_Y = 0$ exactly when $\mathcal{D}_Y^2 F(X, Y)$ is almost surely a function of Y alone. Support positivity and smoothness extend this equality to the admissible domain, so $\mathcal{D}_Y^2 F(x, y)$ is independent of x.

Fix $x_0 \in \mathring{\mathcal{X}}_X$ and let $r_x(y) = F(x, y) - F(x_0, y)$. Since $\mathcal{D}_Y^2 r_x = 0$, a connected continuous support or a consecutive count support implies that r_x is affine. Thus there are functions $a(x)$ and $\eta(x)$ such that $r_x(y) = a(x) + y\eta(x)$. Define $b_Y(y) := \exp\{F(x_0, y)\}$. Then

$$\log p(x, y) = a(x) + \log b_Y(y) + y\eta(x). \tag{8}$$

For any fixed distinct y_0, y_1 in the support interior, $\eta(x) = \{r_x(y_1) - r_x(y_0)\}/(y_1 - y_0)$. Thus a and η are finite, and they are C^2 when X is continuous.

Define the log-partition function $A_Y(t) = \log \int_{\mathcal{X}_Y} b_Y(u) e^{ut}\, d\nu_Y(u)$. Exponentiating (8) and integrating over y gives

$$p_X(x) = e^{a(x)} \int_{\mathcal{X}_Y} b_Y(u) e^{u\eta(x)}\, d\nu_Y(u) = \exp\{a(x) + A_Y(\eta(x))\}.$$

The marginal density $p_X(x)$ is finite and positive for P_X-almost every x. For each such x, the integral is therefore finite, and division by $p_X(x)$ yields

$$p(y \mid x) = \frac{e^{a(x)} b_Y(y) e^{y\eta(x)}}{e^{a(x)+A_Y(\eta(x))}} = b_Y(y) \exp\{y\eta(x) - A_Y(\eta(x))\}.$$

This gives (3). Finiteness of $A_Y(\eta(x))$ alone does not imply that $\eta(x)$ lies in the interior of the canonical-parameter domain.

Conversely, suppose the conditional density has this form. In $\log p(x, y) = \log p_X(x) + \log b_Y(y) + y\eta(x) - A_Y(\eta(x))$, all terms other than $\log b_Y(y)$ are constant or affine in y. Hence $\mathcal{D}_Y^2 \log p(x, y) = \mathcal{D}_Y^2 \log b_Y(y)$, which depends only on y, and $S_Y = 0$. □

C.2 PROOF OF THEOREM 6

Proof. In the given $X \to Y$ representation, η is nonconstant and $A_Y'' > 0$, so the conditional mean $\mathbb{E}[Y \mid X = x] = A_Y'(\eta(x))$ is nonconstant. Thus X and Y are dependent, leaving only the two edge directions to compare.

First suppose $\eta(x) = \beta_0 + \beta_1 x$, where $\beta_1 \neq 0$. Use $\widetilde{b}_X$ and $\widetilde{\Xi}_X$ from Condition R1, and write $\widetilde{A}_X(t) = \log \int_{\mathcal{X}_X} \widetilde{b}_X(x) e^{tx}\, d\nu_X(x)$. The joint density can be written as $p(x, y) = b_Y(y) e^{\beta_0 y} \widetilde{b}_X(x) e^{x\beta_1 y}$. Integrating over x gives $p_Y(y) = b_Y(y) \exp\{\beta_0 y + \widetilde{A}_X(\beta_1 y)\}$. Dividing the joint density by this marginal density therefore gives $p(x \mid y) = \widetilde{b}_X(x) \exp\{x\beta_1 y - \widetilde{A}_X(\beta_1 y)\}$.

To show that the constructed $Y \to X$ representation belongs to the same model class, we check its regularity. First consider canonical-parameter interiority. If Y is continuous, the integral defining p_Y is finite for almost every y. Around any interior support point y, choose one such finite point on each side. Hölder's inequality makes the finiteness domain $\widetilde{\Xi}_X$ convex, so $\beta_1 y$ lies in its interior. For count Y, the integral is finite at every support value because $p_Y(y) \leq 1$. Every level with neighbors on both sides therefore gives an interior parameter. Condition R1 covers the endpoints.

The reverse parameter $\widetilde{\eta}(y) = \beta_1 y$ is nonconstant because $\beta_1 \neq 0$. To check smoothness, observe that

$$\log \widetilde{b}_X(x) = \log p_X(x) - A_Y(\beta_0 + \beta_1 x),$$
$$\log p_Y(y) = \log b_Y(y) + \beta_0 y + \widetilde{A}_X(\beta_1 y).$$

Canonical-parameter interiority and the stated smoothness make these expressions C^2 in continuous variables and finite at count values. Finally, its affine-reversibility base is $p_Y(y)e^{-\widetilde{A}_X(\beta_1 y)} = b_Y(y)e^{\beta_0 y}$, with parameter domain $\Xi_Y - \beta_0$. If X is count, then at each support endpoint e, forward canonical-parameter interiority gives $\beta_0 + \beta_1 e \in \operatorname{int} \Xi_Y$, or equivalently $\beta_1 e \in \operatorname{int}(\Xi_Y - \beta_0)$. Thus the reverse representation also satisfies Condition R1. Taking P_Y as the source therefore gives a reverse representation in the same model class. The direction is consequently not identifiable when η is affine.

Conversely, suppose a reverse semiparametric GLM representation exists, with canonical-parameter function $\widetilde{\eta}$. Choose distinct y_0, y_1 for which both conditional representations hold. Positivity permits any two support values for count Y. When Y is continuous, Fubini's theorem gives a full-measure set of valid values, from which two distinct points can be chosen. The forward representation and Bayes' rule give

$$\log p(x \mid y_1) - \log p(x \mid y_0) = (y_1 - y_0)\eta(x) + C,$$

where C is independent of x. The reverse representation expresses the same difference as $x\{\widetilde{\eta}(y_1) - \widetilde{\eta}(y_0)\} + C'$, where C' is also independent of x. Equating the two expressions and dividing by $y_1 - y_0 \neq 0$ shows that $\eta(x)$ is affine almost everywhere. Support positivity and smoothness extend this equality to the support interior. Therefore a reverse representation exists exactly when η is affine, proving the result. □

C.3 Proof of Theorem 8

Proof. (I) Sink characterization. Fix an ancestral remaining set $R \subseteq V$ and a node $j \in R$. Write $P_R^{(j)} = P_R(\cdot \mid \mathcal{A}_j^{(2)})$, which equals P_R when X_j is continuous. All probabilities, expectations, and variances in this part use $P_R^{(j)}$, with the convention in Appendix A. We first establish

$$j \text{ is a sink of } G_R \quad \Longleftrightarrow \quad H_{j,R}(X_R) = h_{j,R}(X_j) \quad P_R^{(j)}\text{-a.s. for some measurable } h_{j,R}. \tag{9}$$

If j is a sink of G_R, (7) gives $H_{j,R} = \kappa_j(X_j)$, proving the forward implication in (9) with $h_{j,R} = \kappa_j$.

For the reverse implication, we prove the contrapositive. Suppose j is a nonsink, and choose a sink $c^\star$ in the subgraph induced by $\mathrm{Ch}_{G_R}(j)$. Such a node exists because this subgraph is a nonempty finite DAG. Then $c^\star$ is not a parent of any other child of j, so only its own child contribution in (7) depends on $X_{c^\star}$. Let $U = X_{R\setminus\{c^\star\}}$. All parents of $c^\star$, including j, belong to $R \setminus \{c^\star\}$. Define the function $a(u) := (\mathcal{D}_j^2 \eta_{c^\star})\big(u_{\mathrm{Pa}_{G_R}(c^\star)}\big)$. Thus $\eta_{c^\star}(X_{\mathrm{Pa}_{G_R}(c^\star)})$ and its log-partition term are functions of U alone. Since $\mathcal{D}_j^2$ leaves $X_{c^\star}$ fixed, linearity gives $H_{j,R} = a(U)X_{c^\star} + B(U)$, where $B(U)$ collects all terms independent of $X_{c^\star}$. Definition 4 gives an admissible u with $a(u) \neq 0$. Support positivity and smoothness then give $\mathbb{P}\{a(U) \neq 0\} > 0$. Support positivity also makes $X_{c^\star} \mid U$ nondegenerate, including under the admissible-event restriction. Hence $H_{j,R} \mid U$ is nondegenerate on $\{a(U) \neq 0\}$. If $H_{j,R}$ were a function of X_j alone, it would be fixed given U, since U contains X_j. This contradiction proves the reverse implication in (9).

For a nonsink, $H_{j,R} \in L^2$ and the identity $X_{c^\star} = \{H_{j,R} - B(U)\}/a(U)$ give a finite conditional second moment for $X_{c^\star}$ given U on $\{a(U) \neq 0\}$. The conditional nondegeneracy above therefore implies $\mathrm{Var}(H_{j,R} \mid U) = a(U)^2 \mathrm{Var}(X_{c^\star} \mid U) > 0$ on this event. Since U includes X_j, the conditional law of total variance gives

$$0 < \mathbb{E}\{\mathrm{Var}(H_{j,R} \mid U)\} \leq \mathbb{E}\{\mathrm{Var}(H_{j,R} \mid X_j)\} = S_j(R) < \infty.$$

Thus every nonsink has positive CCS, and every sink has zero CCS.

(II) Parent recovery. Fix a causal order, a node i, and a predecessor $j \in \mathrm{Pred}(i)$. Write $R_i = \{i\} \cup \mathrm{Pred}(i)$. Expectations below use the admissible-event convention for $\mathcal{A}_{ij}^{(1,1)}$.

The predecessor set contains all parents of i and no descendants of i. The local Markov property therefore gives $p(x_i \mid x_{\mathrm{Pred}(i)}) = p_i(x_i \mid x_{\mathrm{Pa}_G(i)})$. Since $\log p_{R_i} = \log p(x_i \mid x_{\mathrm{Pred}(i)}) + \log p_{\mathrm{Pred}(i)}$ and the second term is independent of x_i, we have $H_{ij,R_i} = \mathcal{D}_i\mathcal{D}_j \log p_i(x_i \mid x_{\mathrm{Pa}_G(i)})$. If i is a source, its conditional density equals its marginal density $p_i(x_i)$, so this off-diagonal curvature is zero for every predecessor, as required.

For a non-source, applying $\mathcal{D}_i$ to the log conditional density in (3) gives $\mathcal{D}_i \log b_i(x_i) + \eta_i(x_{\mathrm{Pa}_G(i)})$. The mixed operators commute on the admissible domain. Applying $\mathcal{D}_j$ therefore removes the first term and yields

$$H_{ij,R_i} = \mathcal{D}_j \eta_i, \qquad \mathrm{OCS}_{j\to i} = \mathbb{E}\big[|\mathcal{D}_j \eta_i(X_{\mathrm{Pa}_G(i)})|\big].$$

If j is not a parent, η_i is independent of x_j. For a parent, the nonconstant dependence in Definition 3 gives a point where $\mathcal{D}_j \eta_i \neq 0$; support positivity and smoothness give positive probability to this event. Hence

$$j \in \mathrm{Pa}_G(i) \quad \Longleftrightarrow \quad \mathbb{P}\Big(|H_{ij,R_i}| > 0 \,\Big|\, \mathcal{A}_{ij}^{(1,1)}\Big) > 0. \tag{10}$$

Since $H_{ij,R_i} \in L^1$, the OCS is finite and positive exactly for parents. For example, $\eta_i(x_j) = a + bx_j$ with $b \neq 0$ gives $\mathrm{OCS}_{j\to i} = |b| > 0$, so nonlinearity is not needed in this part. □

C.4 Proof of Corollary 9

Proof. Suppose that P has nonlinear semiparametric GLM DAG representations on both G and $\widetilde{G}$. For any remaining set R that is ancestral for both representations, the right-hand side of (9) depends only on P_R. It therefore identifies the same sink set in G_R and $\widetilde{G}_R$. Starting from $R = V$, choose a common sink using a fixed tie-breaking rule and remove it. The new set remains ancestral for both graphs, so induction gives a single ordering π compatible with both DAG representations.

For each node i, $R_i = \{i\} \cup \mathrm{Pred}(i)$ is ancestral for both representations, so (10) applies to both graphs. Its right-hand side depends only on P_{R_i} and hence recovers the same parent set for i. Thus $G = \widetilde{G}$. Both (9) and (10) were established without moment conditions. □

D Additional Identifiability Results

D.1 Strict Extension of SCORE

Corollary 13 (Strict extension of SCORE)**.** *SCORE's identifiable nonlinear Gaussian ANMs form a proper subclass of the continuous nonlinear semiparametric GLM DAGs identified by Corollary 9. For a nonlinear semiparametric GLM DAG with constant own-factor curvature at every node, including these Gaussian ANMs, every ancestral remaining set R and every $j \in R$ satisfy*

$$\mathrm{Var}(H_{j,R}) = 0 \quad \Longleftrightarrow \quad S_j(R) = 0 \quad \Longleftrightarrow \quad j \text{ is a sink of } G_R.$$

Proof. For every non-source node in a Gaussian ANM, choose

$$b_i(x) = (2\pi\tau_i^2)^{-1/2} e^{-x^2/(2\tau_i^2)}, \qquad A_i(t) = \frac{\tau_i^2 t^2}{2}.$$

Then $\eta_i = f_i/\tau_i^2$, so positive scaling preserves edgewise nonlinearity and Corollary 9 applies under the regularity conditions in Appendix B.2. Gaussian source marginals are also allowed, and every node has constant own-factor curvature $\kappa_i = -1/\tau_i^2$.

With constant own-factor curvature, a sink has $H_{j,R} = \kappa_j$, so both scores are zero. At a nonsink, Theorem 8(i) and the law of total variance give $\mathrm{Var}(H_{j,R}) \geq S_j(R) > 0$.

To see that the inclusion is strict, let $X_1 \sim \mathrm{Exp}(1)$ and

$$X_2 \mid X_1 = x \sim \mathrm{Gamma}(\alpha, \text{scale} = 1 + x), \qquad \alpha > 4.$$

This conditional has

$$b_2(y) = \frac{y^{\alpha-1}}{\Gamma(\alpha)}, \qquad \eta_2(x) = -\frac{1}{1+x}, \qquad A_2(t) = -\alpha \log(-t).$$

Here η_2 is nonlinear, while $\kappa_2(y) = -(\alpha - 1)/y^2$ is nonconstant. Hence $S_2 = 0$ but $\mathrm{Var}(H_2) > 0$; $\alpha > 4$ ensures a finite second curvature moment. The CCS therefore identifies a model for which SCORE's constant-curvature criterion fails. □

D.2 Link–Index Characterization

For a conditional family–link–index parameterization of $Y \mid X$, let $g_{Y,c} := (A'_Y)^{-1}$ be the canonical link, set $h := g_{Y,c} \circ g_Y^{-1}$, and write $I_f := f(\mathring{\mathcal{X}}_X)$. Then the canonical-parameter function is $\eta = h \circ f$. For any link g, its affine class is $[g] := \{\alpha \circ g : \alpha \text{ is increasing affine}\}$. Thus a canonical-class link is a member of $[g_{Y,c}]$.

Corollary 14 (Link–index characterization)**.** *In the setting of Theorem 6, its characterization becomes*

$$X \to Y \text{ identifiable} \quad \Longleftrightarrow \quad h \circ f \text{ is nonlinear on } \mathring{\mathcal{X}}_X.$$

(i) If $g_Y \in [g_{Y,c}]$, then h is affine and $\eta = \lambda f + \mu$ for some $\lambda \neq 0$. Hence $X \to Y$ is identifiable if and only if f is nonlinear on $\mathring{\mathcal{X}}_X$.

(ii) If $g_Y \notin [g_{Y,c}]$, then h is nonlinear on its full domain, but its restriction to I_f may be affine. For a nonconstant affine index $f(x) = \alpha + \beta x$, the direction is identifiable exactly when h is nonlinear on $I_f = \alpha + \beta\, \mathring{\mathcal{X}}_X$. For a nonlinear index, the direction fails to be identifiable exactly when $f = h^{-1} \circ a$ on $\mathring{\mathcal{X}}_X$ for an affine function a whose values lie in $h(I_f)$.

Every unidentifiable case also has a canonical-link representation with an affine index.

Proof. Theorem 6 reduces the problem to determining when $\eta = h \circ f$ is affine on $\mathring{\mathcal{X}}_X$. If $f = h^{-1} \circ a$ for an affine function a, then $\eta = a$ is affine. Conversely, suppose that $h \circ f = a$ for an affine a. The map h is strictly increasing on I_f, so its inverse is well defined on $h(I_f)$. Applying this inverse pointwise yields $f = h^{-1} \circ a$ on $\mathring{\mathcal{X}}_X$.

For clause (i), $g_Y \in [g_{Y,c}]$ holds exactly when h is affine. Therefore, $\eta = \lambda f + \mu$ with $\lambda \neq 0$, and η is nonlinear exactly when f is nonlinear.

For clause (ii), first consider a nonconstant affine index $f(x) = \alpha + \beta x$ with $\beta \neq 0$. It maps $\mathring{\mathcal{X}}_X$ bijectively and affinely onto $I_f = \alpha + \beta \mathring{\mathcal{X}}_X$. Consequently, $h \circ f$ is affine on $\mathring{\mathcal{X}}_X$ exactly when h is affine on I_f. For a nonlinear index, the preceding inverse argument shows that the exceptions are exactly the functions $f = h^{-1} \circ a$ with a affine. Condition R1 supplies the reverse representation in the affine case. Taking $g_Y = g_{Y,c}$ then makes $f = \eta$ affine as well, which proves the final statement. □

E Proofs for Section 6

E.1 Proof of Proposition 10

Proof. Let $S = V \backslash R$ and choose a regular conditional distribution of X_S given X_R. All conditional-expectation identities below use their P_R-almost-sure versions. When continuous coordinates are present, the calculation uses density ratios rather than conditioning on a positive-probability atom. For P_R-almost every x_R with $p_R(x_R) > 0$, Bayes' rule gives

$$\begin{aligned}\mathbb{E}[s^V_{j,y}(X_V) \mid X_R = x_R] &= \int \frac{p_V(x_R^{j\to y}, x_S)}{p_V(x_R, x_S)} \frac{p_V(x_R, x_S)}{p_R(x_R)}\, d\nu_S(x_S) \\ &= \frac{p_R(x_R^{j\to y})}{p_R(x_R)} = s^R_{j,y}(x_R).\end{aligned}$$

Here y is a neighboring value in $\mathcal{X}_j$, and ν_S is the product of the dominating measures on the removed coordinates. This proves the stated identity.

It remains to verify integrability for the neighboring entries used by DISCO. Fix $\delta \in \{-1, +1\}$. For $A \in \{R, V\}$, define the zero-extended entry

$$\bar{s}_j^{A,\delta}(x_A) = \begin{cases} s^A_{j,\, x_j+\delta}(x_A), & x_j + \delta \in \mathcal{X}_j, \\ 0, & x_j + \delta \notin \mathcal{X}_j. \end{cases}$$

The support event depends only on x_j, so the preceding identity implies $\bar{s}_j^{R,\delta}(X_R) = \mathbb{E}[\bar{s}_j^{V,\delta}(X_V) \mid X_R]$. Moreover,

$$\mathbb{E}[\bar{s}_j^{V,\delta}(X_V)] = \sum_{x_j:\, x_j+\delta\in\mathcal{X}_j} \int p_V(x_j+\delta, x_{-j})\, d\nu_{-j}(x_{-j}) \le 1.$$

Thus these zero-extended neighboring entries belong to L^1.

For an arbitrary fixed replacement value y, the same Bayes calculation gives $s_{j,y}^R(X_R) = \mathbb{E}[s_{j,y}^V(X_V) \mid X_R]$ as an identity of nonnegative conditional expectations. This is an ordinary integrable conditional-expectation identity whenever $s_{j,y}^V \in L^1(P_V)$. Integrability need not hold for a fixed replacement value. On a countably infinite support with positive mass at y, $\mathbb{E}[s_{j,y}^V(X_V)] = \infty$. □

E.2 Proof of Theorem 11

Proof. For count-valued P, let $\mathcal{A}(G)$ be the nonempty ancestral remaining sets of G, and write $H_{jj,R} := H_{j,R}$. For $R \in \mathcal{A}(G)$ and $i, j \in R$, use the events from Appendix A: $A_{jj,R} := \mathcal{A}_j^{(2)}$ and $A_{ij,R} := \mathcal{A}_{ij}^{(1,1)}$ for $i \neq j$. Set $Q_{ij,R} := P_R(\cdot \mid A_{ij,R})$; the support assumptions give $P_R(A_{ij,R}) > 0$.

Training and evaluation samples. Let $\mathcal{F}_n$ contain the training sample of size n, the fitting randomness, and a finite nonempty noise grid $\mathcal{G}_{\sigma,n} \subset (0,\infty)$. For every $\sigma \in \mathcal{G}_{\sigma,n}$, the noise-specific estimated curvatures $\widehat{H}_{ij,R}^{(n)}(\cdot;\sigma)$ are $\mathcal{F}_n$-measurable. They use ordinary, unclipped shifts and are evaluated only on $A_{ij,R}$. The evaluation observations $X^{(1)}, \dots, X^{(m_n)}$ are iid from P, independent of $\mathcal{F}_n$, with $m_n \to \infty$ as $n \to \infty$.

Estimators and graph threshold. For parent selection, let

$$\mathcal{T}(G) := \{(R, i, j) : R \in \mathcal{A}(G),\ i \text{ is a sink of } G_R,\ j \in R \setminus \{i\}\},$$

and write $O_{j\to i}(R) := \mathbb{E}_{Q_{ij,R}}|H_{ij,R}|$. For each triple, some causal order places $R \setminus \{i\}$ first and i next, so Theorem 8(ii) applies. We use the L^2 and L^1 moment conditions of Appendix B.3 for all diagonal entries with $R \in \mathcal{A}(G)$, $j \in R$, and all triples in $\mathcal{T}(G)$, respectively.

Write $\overline{H}_{ij,R}^{(n)} := |\mathcal{G}_{\sigma,n}|^{-1} \sum_{\sigma\in\mathcal{G}_{\sigma,n}} \widehat{H}_{ij,R}^{(n)}(\cdot;\sigma)$. Let $\widehat{S}_{j,n}(R)$ be the grouped estimator (5) applied to $\overline{H}_{jj,R}^{(n)}$ on $A_{jj,R}$, with fixed $n_{\min} = k \ge 2$ and unbiased within-group sample variances. Let $\widehat{O}_{j\to i,n}(R)$ use the averaging rule in (6) for marginal R and noise grid $\mathcal{G}_{\sigma,n}$, restricting observations to $A_{ij,R}$ and using the evaluation sample as one fold; absolute values are taken before noise averaging. Empty evaluation averages and maxima are zero. Given $\widehat{\pi}_n$, write $\widehat{R}_{i,n} := \{i\} \cup \mathrm{Pred}_{\widehat{\pi}_n}(i)$. The graph $\widehat{G}$ includes $j \to i$ exactly when $j \in \mathrm{Pred}_{\widehat{\pi}_n}(i)$ and $\widehat{O}_{j\to i,n}(\widehat{R}_{i,n}) > \lambda_n$. The procedure may be completed arbitrarily outside $\mathcal{A}(G)$.

Assumption 15 (Marginal curvature estimation)**.** Under the setup above, the noise-specific fitted marginal curvatures converge uniformly over the evaluation noise grid to the clean curvatures:

$$\varepsilon_n := \max_{\substack{R\in\mathcal{A}(G),\ i,j\in R\\ \sigma\in\mathcal{G}_{\sigma,n}}} \left\|\widehat{H}_{ij,R}^{(n)}(\cdot;\sigma) - H_{ij,R}\right\|_{L^2(Q_{ij,R})} \xrightarrow{p} 0, \tag{11}$$

with the error norms finite almost surely.

We first show

$$\max_{R\in\mathcal{A}(G)} \max_{j\in R} |\widehat{S}_{j,n}(R) - S_j(R)| \xrightarrow{p} 0, \tag{12}$$

$$\max_{(R,i,j)\in\mathcal{T}(G)} |\widehat{O}_{j\to i,n}(R) - O_{j\to i}(R)| \xrightarrow{p} 0. \tag{13}$$

Step 1: Consistency of the CCS estimator. For the CCS, fix R, j and write $Q = Q_{jj,R}$, $Z = X_j$, $f = H_{jj,R} \in L^2(Q)$, and $\mathsf{S}_N(g)$ for the grouped estimator (5) using curvature g on the N admissible observations. Here $N \to \infty$ in probability, and conditional on N these observations are iid from Q. Let N_v be the count at level v and $D_N = \sum_v N_v \mathbf{1}\{N_v \geq k\}$. For any finite set F of positive-probability levels, each level is eventually retained, giving $\liminf_N D_N/N \geq Q(Z \in F)$ almost surely. Taking F to the countable support yields $D_N/N \to 1$; this also holds in probability at the random admissible count.

The contribution of levels in F to $\mathsf{S}_N(f)$ converges to $\sum_{v \in F} Q(Z = v) \operatorname{Var}_Q(f \mid Z = v)$. On $D_N > 0$, the remaining contribution is bounded by

$$\frac{c_k}{D_N} \sum_{a=1}^{N} f(X_R^{(a)})^2 \mathbf{1}\{Z_a \notin F\}, \qquad c_k := \frac{k}{k-1} \leq 2.$$

By the law of large numbers and $D_N/N \to 1$, this bound converges to $c_k \mathbb{E}_Q[f^2 \mathbf{1}\{Z \notin F\}]$, which tends to zero as F increases. The population tail has the same bound, so $\mathsf{S}_N(f) \to_p \mathbb{E}_Q[\operatorname{Var}_Q(f \mid Z)] = S_j(R)$.

Now put $f_n = \overline{H}_{jj,R}^{(n)}$ and $h_n = f_n - f$. The triangle inequality and Assumption 15 give

$$\begin{aligned} \|h_n\|_{L^2(Q)} &\leq \frac{1}{|\mathcal{G}_{\sigma,n}|} \sum_{\sigma \in \mathcal{G}_{\sigma,n}} \big\|\widehat{H}_{jj,R}^{(n)}(\cdot;\sigma) - H_{jj,R}\big\|_{L^2(Q)} \\ &\leq \varepsilon_n. \end{aligned}$$

Since $\sqrt{\mathsf{S}_N(\cdot)}$ is a seminorm, on $D_N > 0$,

$$\big|\sqrt{\mathsf{S}_N(f_n)} - \sqrt{\mathsf{S}_N(f)}\big|^2 \leq \mathsf{S}_N(h_n) \leq c_k \frac{N}{D_N} T_n, \qquad T_n := \frac{1}{N} \sum_{a=1}^{N} h_n(X_R^{(a)})^2,$$

with $T_n = 0$ if $N = 0$. Independence gives $\mathbb{E}[T_n \mid \mathcal{F}_n] \leq \varepsilon_n^2$. For every $t > 0$, conditional Markov's inequality gives

$$\Pr(T_n > t) \leq \mathbb{E}[\min\{1, \varepsilon_n^2/t\}] \longrightarrow 0,$$

because the bounded integrand converges to zero in probability. Together with $D_N/N \to 1$, this proves $\widehat{S}_{j,n}(R) \to_p S_j(R)$ and, by finiteness of the index set, (12).

Step 2: Consistency of the OCS estimator. For the OCS, fix $(R, i, j) \in \mathcal{T}(G)$ and let $\widetilde{O}_{j \to i,n}(R)$ average $|H_{ij,R}|$ on the same admissible observations. The $L^1(Q_{ij,R})$ moment and the law of large numbers give $\widetilde{O}_{j \to i,n}(R) \to_p O_{j \to i}(R)$, while

$$\begin{aligned} &\mathbb{E}\Big[|\widehat{O}_{j \to i,n}(R) - \widetilde{O}_{j \to i,n}(R)| \mid \mathcal{F}_n\Big] \\ &\quad \leq \frac{1}{|\mathcal{G}_{\sigma,n}|} \sum_{\sigma \in \mathcal{G}_{\sigma,n}} \big\|\widehat{H}_{ij,R}^{(n)}(\cdot;\sigma) - H_{ij,R}\big\|_{L^1(Q_{ij,R})} \\ &\quad \leq \varepsilon_n. \end{aligned}$$

The first inequality uses $\big||a| - |b|\big| \leq |a - b|$ before averaging over noise levels; the second uses the L^2 error bound. The same conditional Markov argument and finiteness of $\mathcal{T}(G)$ yield (13).

Step 3: Recovery of a causal order. By Theorem 8(i), sinks have zero CCS and nonsinks have positive CCS on every ancestral set. If nonsinks exist, their scores have a positive minimum Δ over this finite collection. When all CCS errors are below $\Delta/3$, every empirical minimizer is a sink. Starting from V, each removal therefore preserves ancestrality and yields a causal order. By (12), this event has probability tending to one. If no nonsinks exist, every order is valid.

Step 4: Recovery of the DAG. For a nonparent triple in $\mathcal{T}(G)$, Theorem 8(ii) gives $H_{ij,R} = 0$ $Q_{ij,R}$-almost surely, so $\widetilde{O}_{j \to i,n}(R) = 0$. Hence

$$\Pr\{\widehat{O}_{j \to i,n}(R) > \lambda_n\} \leq \mathbb{E}[\min\{1, \varepsilon_n/\lambda_n\}] \longrightarrow 0.$$

For a true parent, $O_{j\to i}(R) > 0$, so (13) and $\lambda_n \to 0$ give selection with probability tending to one. A union bound over $\mathcal{T}(G)$, together with the correct-order event, proves $\Pr\{\widehat{G} = G\} \to 1$. □

Consistency of common-grid clipping and median/IQR standardization is not established here. Applying the theorem to fitted grids requires control of rank and truncation errors, and Assumption 15 requires any evaluation-noise bias to vanish.

E.3 Proof of Proposition 12

Proof. Let $Y_j = \phi_j(X_j)$ with ϕ_j strictly increasing, and write the ordered support as $v_{j,0} < v_{j,1} < \cdots$. Strict monotonicity gives $T_j^Y(\phi_j(v_{j,k})) = k = T_j^X(v_{j,k})$ for every support value. Thus $T^Y(Y) = T^X(X)$ almost surely, so their marginal CCS and OCS values, and hence their causal orders and parent sets, coincide. The fitted transform also depends only on this order, including its clipping and between-value rules. It therefore produces identical transformed training and evaluation inputs, proving the fixed-seed claim. □

F Theoretical Properties of Concrete-Score Estimation

Fix a remaining set R and a noise level $\sigma > 0$. Starting from $x_0 \sim p_R$, corrupt the coordinates independently with kernels $q_\sigma^j(b \mid a) = [\exp(\sigma Q^j)]_{ab}$, where Q^j is a continuous-time Markov-chain rate matrix. We assume $q_\sigma^j(b \mid a) > 0$ for every pair of states a, b in the coordinate state space and every positive noise level used below. The finite-grid reflecting birth–death kernel used in DISCO satisfies this condition. The resulting transition probability and noisy marginal are

$$p_{\sigma|0}(x_R \mid x_0) = \prod_{j\in R} q_\sigma^j(x_j \mid x_{0j}), \qquad p_{R,\sigma}(x_R) = \sum_{x_0} p_{\sigma|0}(x_R \mid x_0) p_R(x_0).$$

The corresponding noisy concrete score is $s_{j,y}^R(x_R;\sigma) = p_{R,\sigma}(x_R^{j\to y})/p_{R,\sigma}(x_R)$.

F.1 Denoising Score-Entropy Objective

For $u, v > 0$, the score-entropy loss is

$$\ell_{\mathrm{SE}}(u, v) = u - v + v\log(v/u). \tag{14}$$

Denoising objective. Using the corruption and noisy concrete-score notation above, the count-structured loss below is the birth–death restriction of the denoising score-entropy objective of Lou et al. (2024). Each Q^j has nonnegative off-diagonal entries and zero row sums. We estimate each non-self entry $y \neq x_j$ by $\big[s_\theta(x_R;\sigma)\big]_{j,y}$, fix the self entry to 1, and minimize

$$\begin{aligned}\mathcal{L}_{\mathrm{DSE}}(\theta) = \mathbb{E}_t\, \mathbb{E}_{x_0\sim p_R}\, \mathbb{E}_{x_R\sim p_{\sigma|0}(\cdot|x_0)} \sum_{j\in R}\sum_{y\neq x_j} w_{x_j y} \\ \times\, \dot\sigma(t)\, \ell_{\mathrm{SE}}\Big(\big[s_\theta(x_R;\sigma)\big]_{j,y}, \rho_{j,y}^{x_0}\Big),\end{aligned} \tag{15}$$

where ℓ_{SE} is defined in (14), $t \sim \mathrm{Unif}(\epsilon_t, 1)$ with $\epsilon_t \in (0,1)$, and $\sigma = \sigma(t) > 0$ with $0 < \dot\sigma(t) < \infty$ on the sampled interval. We set $\rho_{j,y}^{x_0} = q_\sigma^j(y \mid x_{0j})/q_\sigma^j(x_j \mid x_{0j})$ and $w_{x_j y} = Q^j(x_j, y)$, matching the training specification in Appendix G.2.

Population minimizer. Restrict the objective to modeled entries with $w_{x_j y} > 0$ and omit entries with zero generator rate. At a fixed sampled time t, for a corrupted state x_R and modeled entry (j, y) with $p_{R,\sigma}(x_R) > 0$ and $p_{R,\sigma}(x_R^{j\to y}) > 0$, set $\bar\rho_{j,y}(x_R) = \mathbb{E}[\rho_{j,y}^{X_0} \mid X_R = x_R]$. Expanding the loss gives, almost surely, the finite conditional risk $u - \bar\rho_{j,y}(x_R)\log u + C$ at prediction $u > 0$, where C is finite and does not depend on u. Its derivative $1 - \bar\rho_{j,y}(x_R)/u$ shows that the unique minimizer is $u = \bar\rho_{j,y}(x_R)$. The strictly positive generator and time weights do not alter this conditional minimizer.

Kernel positivity justifies cancellation of the transition probabilities, and the product form of the corruption kernel gives

$$\bar{\rho}_{j,y}(x_R) = \sum_{x_0} \frac{p_R(x_0)p_{\sigma|0}(x_R \mid x_0)}{p_{R,\sigma}(x_R)} \frac{q^j_\sigma(y \mid x_{0j})}{q^j_\sigma(x_j \mid x_{0j})}$$
$$= \frac{1}{p_{R,\sigma}(x_R)} \sum_{x_0} p_R(x_0)p_{\sigma|0}(x_R^{j\to y} \mid x_0) = \frac{p_{R,\sigma}(x_R^{j\to y})}{p_{R,\sigma}(x_R)} = s^R_{j,y}(x_R;\sigma).$$

On the finite common corruption-model grid, both the generator weights and the number of modeled entries are bounded. These entrywise minima therefore make the corrupted concrete score the unique population minimizer of (15) on modeled entries.

F.2 Vanishing-Noise Limit

Proposition 16 (Vanishing-noise limit)**.** *Suppose q^j_σ tends to the identity as $\sigma \downarrow 0$, which holds for the finite-grid reflecting birth–death kernel used in* DISCO. *Then $p_{R,\sigma} \to p_R$ pointwise, and for every x_R with $p_R(x_R) > 0$ and every $y \in \mathcal{X}_j$,*

$$s^R_{j,y}(x_R;\sigma) \xrightarrow[\sigma\downarrow 0]{} s^R_{j,y}(x_R).$$

Proof. On the countable support, $p_{R,\sigma}(x_R) = \sum_{x_0} p_{\sigma|0}(x_R \mid x_0)p_R(x_0)$. For each fixed x_0, the summand converges to $\mathbf{1}\{x_R = x_0\}p_R(x_0)$ as $\sigma \downarrow 0$. It is bounded by the summable function $p_R(x_0)$ because $0 \le p_{\sigma|0}(x_R \mid x_0) \le 1$. Dominated convergence therefore gives $p_{R,\sigma}(x_R) \to p_R(x_R)$ for every x_R. If $p_R(x_R) > 0$, continuity of the ratio yields

$$s^R_{j,y}(x_R;\sigma) = \frac{p_{R,\sigma}(x_R^{j\to y})}{p_{R,\sigma}(x_R)} \longrightarrow \frac{p_R(x_R^{j\to y})}{p_R(x_R)} = s^R_{j,y}(x_R).$$

□

G DISCO Implementation

Cross-fitting. We partition the observations into evaluation folds $\mathcal{I}_1, \dots, \mathcal{I}_B$. For each fold, we fit the rank transform, joint concrete-score network, and projection network on the other folds. We then apply the fitted transform and networks to the held-out fold to compute the CCS and OCS.

G.1 Rank Transform

For coordinate j, let $v_{j,0} < \cdots < v_{j,K_j}$ be the distinct values observed in the score-training split and set $\widehat{T}_j(v_{j,k}) = k$. Apply this fitted transform to the held-out evaluation split without using its empirical distribution. A held-out value below $v_{j,0}$ is assigned rank 0, a value above v_{j,K_j} is assigned rank K_j, and a value strictly between $v_{j,k}$ and $v_{j,k+1}$ is assigned rank $\max(k, 1)$. The first interior gap is assigned rank 1 to keep the observed minimum as a separate state. Thus the evaluation values are clipped or coarsened onto the rank grid according to order rather than Euclidean distance.

G.2 Joint Concrete-Score Training

Training objective. We train the joint concrete-score network once on the score-training split, using the denoising score-entropy objective (15) with $R = V$: each minibatch draws $t \sim \mathrm{Unif}(\epsilon_t, 1)$, perturbs $x_0 \sim p_V$ to $x \sim p_{\sigma(t)|0}(\cdot \mid x_0)$, and weights the loss by $\dot{\sigma}(t)$. We use the geometric schedule $\sigma(t) = \sigma_{\min}^{1-t}\sigma_{\max}^t$ with $(\epsilon_t, \sigma_{\min}, \sigma_{\max}) = (10^{-3}, 10^{-3}, 3)$. The training objective is restricted to the neighboring replacements $y = x_j \pm 1$ and masks replacements outside the common corruption-model grid $\{0, \dots, K\}$, where $K = \max_j K_j$. The birth–death corruption uses this same grid with reflecting boundaries for every coordinate. The common upper bound K may exceed a coordinate's largest training rank K_j, so this grid is not the coordinate's population support.

Network architecture. Writing $K_j^\vee = \max(K_j, 1)$ and $K^\vee = (K_1^\vee, \dots, K_d^\vee)$, the network is a multilayer perceptron with coordinate-wise normalized inputs $[\, x/K^\vee,\ \log(1+x)/\log(1+K^\vee),\ \log\sigma\,]$ and output reshaped to $\mathbb{R}^{d\times 2}$, with the last axis indexing the two directions.

G.3 Marginal Concrete-Score Projection

G.3.1 Mask Distribution

The default score-entropy estimator draws a remaining-set size uniformly from $\{2, \dots, d\}$ and then a subset uniformly conditional on that size, using one mask per training observation and epoch. Independently with probability 0.2, it replaces the sampled mask by the full set V. For $d \geq 2$, the resulting distribution is

$$\mathcal{Q}(R) = \frac{0.8}{(d-1)\binom{d}{|R|}} + 0.2\,\mathbf{1}\{R = V\}, \qquad 2 \leq |R| \leq d.$$

Thus every subset of size at least two has positive sampling probability; for $d > 2$, neither subsets nor final mask sizes are uniformly distributed. Positive support does not mean that every queried mask is observed during finite training or that prediction error is uniformly controlled.

G.3.2 Score-Entropy Training

Each projection update draws $X_0 \sim p_V$ and corrupts the full observation to $X_\sigma \sim p_{\sigma|0}(\cdot \mid X_0)$. The joint network is held fixed and evaluated at this same X_σ; the projection network receives $(X_{\sigma,R}, R, \sigma)$ with removed coordinates zero-masked, and the score-entropy loss is restricted to retained outputs $j \in R$. Masks are sampled independently of the observation. In practice, the projection target is the conditional mean of the joint network's clipped predictions, which may differ from the exact marginal concrete score because of approximation error and clipping. Evaluation at finite near-zero noise levels approximates the clean target as motivated by Proposition 16; it does not give a global consistency guarantee.

G.4 Curvature Estimation

Evaluation. We use five geometrically spaced noise levels from 10^{-3} to 5×10^{-3} in $\mathcal{G}_\sigma$ and direct log-ratio outputs from the projection network.

We extend (4) to grid boundaries by replacing $x_R + e_i$ with the clipped shift $C_{i,K}(x_R) := x_R^{\,i\to\min(x_i+1,K)}$, using the corresponding projection-network outputs at boundary points as well. We do not impose the upper-bound filter $x_j + 2 \leq K$. For diagonal curvature ($i = j$), when $x_j = K$, the clipped shift leaves the input unchanged, so the two projection-network predictions coincide and the estimate is zero. This convention is separate from the population-admissible finite differences used in the theory.

For the grouped conditional variance (5), we retain rank values with at least $n_{\min} = 2$ evaluation observations and weight each by its sample size. A candidate with no retained level has an unavailable score. The implementation averages available fold scores and selects a finite minimizer; if every candidate score is unavailable, it selects the first remaining node.

Symmetric off-diagonal curvature evaluation. For parent selection, we use the commuting-difference identity

$$H_{ij,\widehat{R}_i} = \mathcal{D}_i\mathcal{D}_j \log p_{\widehat{R}_i} = \mathcal{D}_j\mathcal{D}_i \log p_{\widehat{R}_i} = H_{ji,\widehat{R}_i}. \tag{16}$$

For each node i, we obtain the $\widehat{R}_i$-marginal concrete score for $\widehat{R}_i = \{i\} \cup \mathrm{Pred}_{\widehat{\pi}}(i)$, evaluate the projection network on the unshifted batch and the batch shifted in coordinate i, $C_{i,K}(x_{\widehat{R}_i})$, and extract the entries for all predecessor coordinates $j \in \mathrm{Pred}_{\widehat{\pi}}(i)$ from those two outputs. Thus the off-diagonal curvatures in (4) require only one shifted batch per node. We compute these curvatures for all predecessors simultaneously, averaging their absolute values over noise levels and observations for each node. All candidate parents use the same unshifted score predictions and remaining-set mask.

G.5 Parent Selection

We estimate OCS with $B = 3$ cross-fitting folds. For every candidate edge, $\mathcal{I}_b^{ij} = \mathcal{I}_b$ in (6), so the estimate uses all observations in the held-out fold. For a node i with at least two candidate parents, we robustly standardize these estimates within each fold,

$$Z_{j\to i}^{(b)} = \frac{\widehat{\mathrm{OCS}}_{j\to i}^{(b)} - \mathrm{median}_{k\in \mathrm{Pred}_{\widehat{\pi}}(i)} \widehat{\mathrm{OCS}}_{k\to i}^{(b)}}{s_i^{(b)}}.$$

Here $s_i^{(b)}$ is the IQR of the candidate OCS estimates, falling back successively to the scaled median absolute deviation, standard deviation, and unit scale when degenerate. We apply the selection-frequency rule in Section 6.3.

We use $(\tau_Z, \pi_{\min}) = (2, 1)$ unless otherwise stated. For a single candidate, we omit centering and use the median robust scale across children with at least two candidates in the same fold. If none exist, we use the robust scale of all candidate scores in that fold. The same threshold and selection-frequency rule still apply. Candidate sets of size two or three remain unable to pass the default threshold in exact arithmetic. This practical rule does not provide a finite-sample error-control guarantee. Appendix I.2 evaluates threshold sensitivity and compares parent-selection methods.

H Experimental Details and Additional Results

H.1 Experimental Setup

Graph and sampling protocol. For synthetic experiments, ERk denotes an Erdős–Rényi DAG obtained from a random permutation used as the causal order, with each admissible earlier-to-later edge included independently with probability $2k/d$ (expected edge count $k(d-1)$). Unless stated otherwise, we use ER3 DAGs. Within each setting and seed, all methods receive the same graph and observations. For sample-size curves, smaller samples are row prefixes of the largest sample.

Replication and pairing. Unless stated otherwise, we use 10 paired seeds for each synthetic experiment. The Lahman comparisons use three runs on each fixed cohort; these are computational repetitions, not independent sampled cohorts. Ablations also share learned scores whenever applicable; graph seeds are the independent replication units.

Evaluation metrics. We use $A_{\mathrm{top}} = |\mathcal{E}|^{-1} \sum_{(u,v)\in\mathcal{E}} \mathbf{1}\{\widehat{\pi}(u) < \widehat{\pi}(v)\}$ for ordering agreement. The nonempty evaluation set $\mathcal{E}$ is the true DAG edge set E for synthetic data and the partial directed reference set $\mathcal{R}$ for real data. DAG recovery comparisons report F_1, precision, recall, and structural Hamming distance (SHD), defined as FP + FN, so reversing a true edge counts twice. Parent-selection comparisons hold the order fixed. For partial-reference edge evaluations, *Direct* and *Rev.* count reference relations recovered in the same and opposite directions, respectively. An estimated edge absent from an incomplete reference is not treated as a confirmed false edge.

Recovery curves show medians and interquartile ranges. Synthetic benchmark tables report medians, whereas the ordering/projection ablations and additional scale-free, COM–Poisson, and continuous-data comparisons report means and sample standard deviations. Parent-selection tables and the Lahman study report means. Conditional family abbreviations are Poi, NB, and Bin, listed in that order in matched comparisons.

Default estimator. Unless stated otherwise, we train the joint network for up to 800 epochs on count data and select it by validation loss. We train the projection network for 800 epochs with the score-entropy loss. The COM–Poisson experiment (Table 10), parent-selection experiments (Tables 16 and 17), and 2019–2025 Lahman DISCO analysis use 400 projection epochs. Masks follow Appendix G.3, and target and parameter exponential moving averages are disabled. The parent-selection rule is specified in Appendix G.5. Separate continuous-data settings are given in Appendix H.2.8.

Data generation. For non-source nodes, write the index as $z_j = b_0 + \sum_{k \in \mathrm{Pa}_G(j)} \beta_{kj} \phi(X_k)$. Table 5 gives the six benchmark mechanisms; μ_j denotes the conditional mean and q_j the binomial success probability. The intercept is fixed within each mechanism. We do not standardize the generated counts, normalize by in-degree, or modify observations after generation.

Table 5: The six single-family benchmark mechanisms. Coefficients are drawn independently and uniformly from the stated intervals.

Conditional family	Conditional response	$\phi(x)$	b_0	Coefficients
Poi	$\mu_j = 0.2 + \mathrm{softplus}(z_j)$	x	0.5	$[0.30, 0.45]$
	$\mu_j = \exp(\widetilde{z}_j)$	$\log(1+x)$	0.4	$[0.15, 0.30]$
NB	$\mu_j = 0.2 + \mathrm{softplus}(z_j)$	x	0.5	$[0.30, 0.45]$
	$\mu_j = \exp(\widetilde{z}_j)$	$\log(1+x)$	0.4	$[0.15, 0.30]$
Bin	$q_j = \Phi(z_j)$	$x/50$	-1.28	$[0.50, 1.00]$
	$q_j = \mathrm{sigmoid}(z_j)$	$\sqrt{x/50}$	-2.197	$[0.60, 0.90]$

Here $\widetilde{z}_j = \mathrm{clip}(z_j, -3, 3.5)$ limits the exponential response, and Φ is the standard normal CDF. Source means are 2 for the softplus mechanisms and 1.5 for the exponential mechanisms. Negative binomial nodes have size $r = 6$. Binomial nodes have size $M = 50$ and source probability 0.10; generated probabilities are clipped to $[10^{-8}, 1 - 10^{-8}]$. Mixed-family benchmarks use the softplus/affine rows for Poisson and negative binomial children and the sigmoid/square-root row for binomial children, with the feature map determined by the child's conditional family.

CCS comparison. Table 2(a) uses the mechanisms in Table 5 at $d = 50$ and $n = 5000$. CCS and the constant-curvature criterion use the same fitted networks and 60 matched nontrivial ancestral remaining sets per conditional family. Accuracy is the fraction of sets for which the selected node is a true sink; sets within a graph seed are not independent replicates.

H.1.1 Baseline Methods

ODS. ODS (Poi) follows the QVF overdispersion ordering criterion of Park and Raskutti (2018). The multivariate experiments use regression estimates of the conditional mean and variance because exact-match conditioning cells collapse at the graph densities considered here. The fixed-family ODS variants (Poisson, negative binomial, and binomial) and ODS (Oracle) differ only in the supplied QVF coefficient and family-specific regression link.

NOTEARS-MLP. We use the gCastle nonlinear MLP (Zhang et al., 2021) with squared loss on standardized $\log(1 + x)$ counts.

DiffAN. We use the DiffAN implementation of Sanchez et al. (2023). Runtime includes diffusion ordering and parent pruning. Runs exceeding the time limit are marked Timeout.

MRS. We use the MRS implementation (Park and Park, 2019a) with GES skeleton estimation. Mixed-family and scalability experiments use its Poisson setting. Single-family experiments use the Poisson and binomial settings for the corresponding families, and a hyper-Poisson proxy for negative binomial data.

H.2 Additional Benchmarks

H.2.1 Invariance to Strictly Increasing Transformations

Table 6 compares DISCO with DISCO-NoRank after applying $\log(1 + x)$, Anscombe's $2\sqrt{x + 3/8}$ transform, and $x^2 + x$ to the same Poisson data. With the rank transform, median

$A_{\text{top}} = 0.9217$ and $F_1 = 0.8941$ remain unchanged across transformations. Without it, median F_1 falls from 0.8991 on raw counts to 0.3013 under $x^2 + x$.

Table 6: Invariance to strictly increasing transformations on Poisson DAGs ($d = 50$, $n = 5000$). Entries are medians. DISCO-NoRank omits the rank transform.

	Raw counts		$\log(1+x)$		Anscombe		$x^2 + x$	
Method	A_{top}	F_1	A_{top}	F_1	A_{top}	F_1	A_{top}	F_1
DISCO	0.9217	0.8941	0.9217	0.8941	0.9217	0.8941	0.9217	0.8941
DISCO-NoRank	0.9365	0.8991	0.8971	0.7261	0.8900	0.8596	0.8341	0.3013

H.2.2 Single-Family DAG Recovery

The binomial–probit panel uses ODS (Poi) with a Poisson QVF, a non-oracle specification for binomial data.

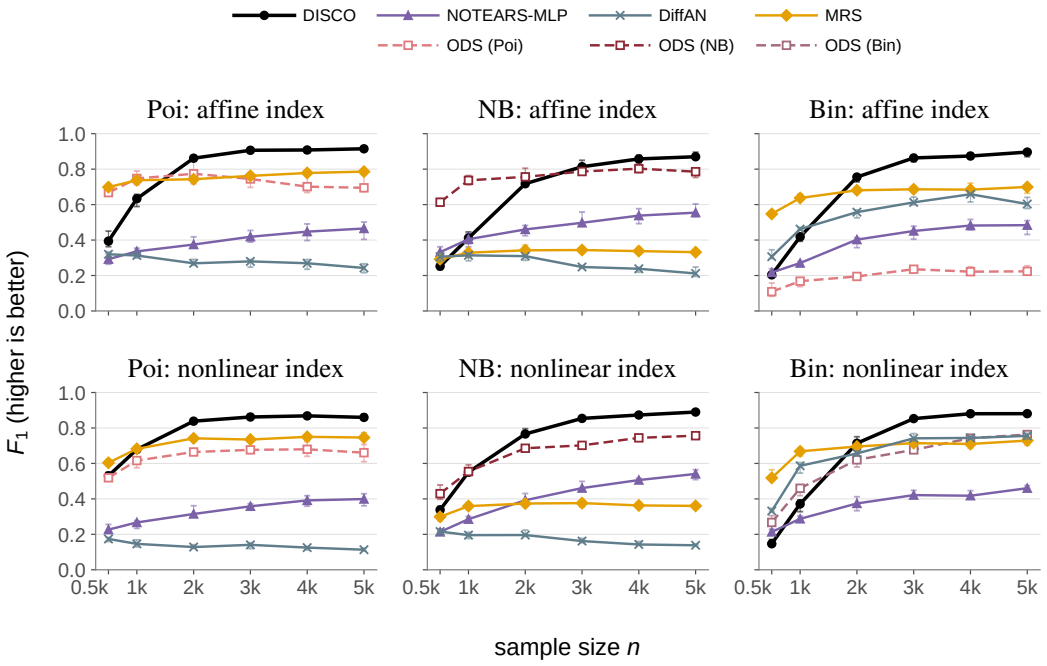


Figure 2: Extended single-family comparison across all six mechanisms and six sample sizes. Columns correspond to the Poi, NB, and Bin conditional families; the top and bottom rows use the first and second mechanisms for each family in Table 5, respectively.

H.2.3 Mixed-Family DAG Recovery

We split a causal order into three consecutive groups of nearly equal size and assign Poisson, negative binomial, and binomial mechanisms to these groups in all six permutations. The arrows in the panel titles indicate this assignment order. Each node uses the mechanism for its assigned conditional family in Table 5. Figure 1(c) reports the NB $\to$ Bin $\to$ Poi configuration. To examine sensitivity to conditional-family specification, Figure 3 compares DISCO with ODS (Oracle), ODS (Poi), ODS (NB), and ODS (Bin) across all six configurations.

H.2.4 Scalability with Graph Size

Table 7 reports A_{top} for the Poisson scalability experiments in Table 3 under the same 240-minute time limit.

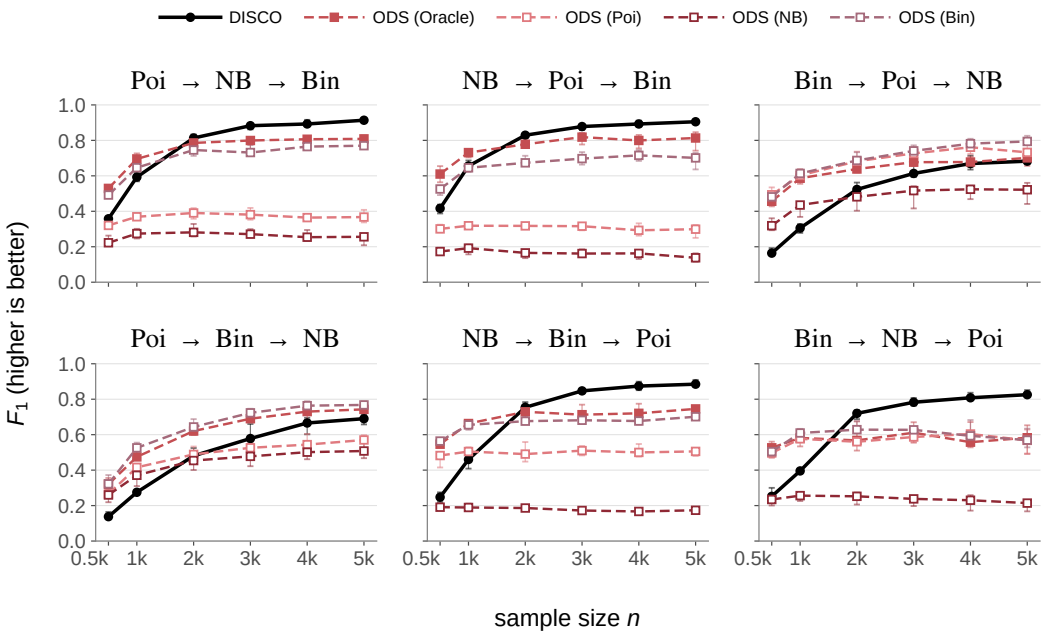


Figure 3: Mixed-family DAG recovery across six conditional-family configurations. Each panel compares DISCO with ODS (Oracle), ODS (Poi), ODS (NB), and ODS (Bin). ODS (Oracle) uses the nodewise true conditional family and QVF parameters. Points and error bars show the median and interquartile range.

Table 7: Ordering accuracy in the scalability experiments at $n = 5000$. Entries are medians; configurations marked Timeout in Table 3 are omitted.

Method	d	A_{top}
DISCO	100	0.9369
	200	0.9021
	500	0.9143
	1000	0.9077
ODS (Oracle)	100	0.8186
	200	0.8164
	500	0.8113
MRS	100	0.7971
	200	0.7791
	500	0.7895
NOTEARS-MLP	100	0.5414

Total time in Table 3 covers the complete DAG recovery procedure, including parent selection. MRS, NOTEARS-MLP, and DiffAN were first evaluated with a four-hour limit. Configurations finishing within this limit were evaluated on the remaining seeds. For DiffAN at $d = 100$ and $d = 200$, the time limit was reached during parent pruning after ordering had finished.

CPU baselines use one thread, while DISCO uses GPU training and CPU post-processing. The runtime comparison is therefore descriptive and does not imply matched hardware.

H.2.5 SPARSE DAG RECOVERY

We repeat the single-family and mixed-family designs with ER1 graphs, holding the mechanisms, coefficient distributions, and sample-size grid fixed. Table 8 reports medians for DISCO at $n = 5000$.

Table 8: Single-family and mixed-family DAG recovery on ER1 graphs ($d = 100$, $n = 5000$). Entries are medians for DISCO.

Setting	Conditional family	Index/configuration	A_{top}	F_1	SHD
Single-family	Poi	Affine	0.816	0.750	49.0
		Nonlinear	0.897	0.873	25.0
	NB	Affine	0.841	0.788	42.5
		Nonlinear	0.907	0.884	22.0
	Bin	Affine	0.886	0.860	28.5
		Nonlinear	0.904	0.886	23.0
Mixed-family	Poi → NB → Bin		0.891	0.845	29.5
	Poi → Bin → NB		0.845	0.796	44.0
	NB → Poi → Bin		0.810	0.777	44.0
	NB → Bin → Poi		0.766	0.705	58.0
	Bin → Poi → NB		0.739	0.685	67.0
	Bin → NB → Poi		0.723	0.633	75.0

H.2.6 SCALE-FREE DAG RECOVERY

We replace the ER graph generator with the Barabási–Albert preferential-attachment generator, keeping $d = 50$ and $n = 5000$ for each of Poisson, negative binomial, and binomial. SF1 and SF3 attach each new node to one and three existing nodes, respectively, yielding 49 and 141 edges. A random node ordering orients the skeleton acyclically; attachment time is not supplied as a causal order. The conditional mechanisms and coefficient ranges are unchanged from the parent-selection study in Appendix I.2. Seed labels identify independent repetitions within a graph type, not identical observations across ER and SF graphs.

Table 9: Scale-free count-DAG recovery ($d = 50$, $n = 5000$). Entries are means $\pm$ sample standard deviations.

Graph	Conditional family	A_{top}	F_1	SHD
SF1	Poi	0.784 ± 0.065	0.705 ± 0.091	28.6 ± 8.5
	NB	0.822 ± 0.054	0.741 ± 0.094	24.8 ± 8.9
	Bin	0.778 ± 0.060	0.782 ± 0.043	20.1 ± 4.0
SF3	Poi	0.918 ± 0.019	0.719 ± 0.074	65.3 ± 14.4
	NB	0.904 ± 0.035	0.620 ± 0.114	82.4 ± 19.7
	Bin	0.769 ± 0.050	0.693 ± 0.077	71.2 ± 14.8

Recovery varies with both graph structure and conditional family. In particular, the higher A_{top} on SF3 for Poisson and negative binomial does not imply a corresponding increase in F_1. These results extend the empirical evaluation beyond ER graphs; they do not establish performance independent of graph structure or conditional family.

H.2.7 DAG Recovery Beyond Standard Count Families

We test recovery beyond the Poisson, negative binomial, and binomial benchmark conditional families using mean-parametrized Conway–Maxwell–Poisson (COM–Poisson) distributions (Huang, 2017). For a parent-independent shape $\nu > 0$,

$$p(y \mid x_{\mathrm{Pa}_G(j)}) = \exp\{y\eta(x_{\mathrm{Pa}_G(j)}) - \nu \log(y!) - A_\nu(\eta(x_{\mathrm{Pa}_G(j)}))\}, \qquad y \in \mathbb{N}_0.$$

We use $\nu \in \{0.5, 1, 2\}$ on DAGs with $d = 50$ and $n = 5000$; $\nu = 1$ reduces to the Poisson distribution. Graphs and coefficients are paired by seed. At fixed parent values, all three settings share source mean 2 and non-source conditional mean $0.2 + \mathrm{softplus}(0.5 + \sum_{k \in \mathrm{Pa}_G(j)} \beta_{kj} X_k)$, with $\beta_{kj} \sim \mathrm{U}(0.30, 0.45)$. We solve for the canonical parameter to match this mean, rather than treating the usual COM–Poisson rate as its expectation. Hence conditional means are matched, while the corresponding canonical parameters and samples may differ. Sampling approximates the unbounded distribution through adaptive support expansion, with bounds on the omitted probability mass and its first-moment contribution. DISCO receives neither the conditional family nor ν; the same configuration is used for all three ν values.

Table 10: Recovery beyond the three benchmark conditional families ($d = 50$, $n = 5000$). Entries are means $\pm$ sample standard deviations. The $\nu = 1$ row uses results from the corresponding Poisson experiment; graph metrics include DISCO parent selection.

Conditional family	A_{top}	F_1	SHD
COM–Poisson ($\nu = 0.5$)	0.934 ± 0.018	0.853 ± 0.033	37.8 ± 8.5
COM–Poisson ($\nu = 1$)	0.928 ± 0.020	0.877 ± 0.043	32.4 ± 10.2
COM–Poisson ($\nu = 2$)	0.920 ± 0.019	0.873 ± 0.026	33.6 ± 6.5

Ordering remains accurate in both non-Poisson settings, while DAG recovery varies: $\nu = 0.5$ has lower mean F_1 than $\nu = 1$. This supports recovery on additional fixed-carrier one-parameter families, not arbitrary count conditional distributions or parent-dependent shape parameters.

H.2.8 Continuous-Data Ordering Recovery

Algorithm. Starting from the DiffAN MLP implementation (Sanchez et al., 2023), DISCO-Cont replaces its ordering rule with CCS and adds marginal projection. We train an unmasked joint diffusion-score network for 1000 epochs and a width-256 mask-conditioned projection network for 400 epochs, using squared error against noise predictions from the fixed joint network on corrupted inputs. Both stages use a learning rate of 10^{-3} and a batch size of 1000; neither network is refitted during ordering. Inputs are standardized for network training, but CCS uses original-unit curvature at $t = 0.05$ and held-out squared residuals from five-fold cross-fitted cubic splines of X_j (six quantile knots; ridge penalty 10^{-3}).

The DiffAN comparator uses masking without residue updates. For Gaussian data, it is trained separately with a 3000-epoch cap and the implementation's early-stopping and voting rules. For Gamma data, both ordering procedures share the same learned joint network, with DiffAN evaluated at $t = 0.05$ without retraining. This shared-network comparison changes marginal-score approximation as well as sink selection, so it is not a CCS-only ablation.

The Gamma joint network omits the first hidden-layer normalization to retain amplitude information relevant to score derivatives. Antithetic noise pairs at the same input and noise level balance opposite corruption directions without changing the denoising target. Half the noise levels come from $t \sim \mathrm{Uniform}(0, 0.5)$ and half from noise standard deviations drawn log-uniformly on $[0.05, 0.8]$. This mixture broadens noise-scale coverage while retaining the original low-noise draws, since low denoising loss alone need not imply accurate score derivatives. Input standardization is unchanged, and projection retains $t \sim \mathrm{Uniform}(0, 0.5)$.

Experimental setup. We compare Gaussian and Gamma conditional families at $d = 30$ and $n = 5000$ for both ER1 and ER3. A random node permutation defines the possible forward

edges, of which exactly kd are sampled for ERk. Each mechanism f_j is drawn independently from a zero-mean Gaussian process with a multivariate RBF kernel of unit variance and unit length scale. In the Gaussian model, sources are standard normal and non-sources follow $X_j = f_j(X_{\mathrm{Pa}_G(j)}) + \varepsilon_j$ with independent $\varepsilon_j \sim \mathcal{N}(0,1)$. Gamma conditionals have fixed shape 6 and mean $2\exp\{0.5 f_j(X_{\mathrm{Pa}_G(j)})\}$; sources have the same shape and mean 2. Within each family and seed, both methods use the same graph and observations.

Table 11: Continuous-data ordering recovery ($d = 30$, $n = 5000$). Entries report A_{top} (higher is better) as means $\pm$ sample standard deviations.

Graph	Method	Gaussian	Gamma
ER1	DISCO-CONT	0.877 ± 0.063	0.640 ± 0.126
	DiffAN	0.850 ± 0.072	0.540 ± 0.137
ER3	DISCO-CONT	0.866 ± 0.041	0.718 ± 0.021
	DiffAN	0.882 ± 0.049	0.657 ± 0.057

Results. Table 11 shows higher mean A_{top} for DISCO-CONT on Gaussian ER1 and for DiffAN on Gaussian ER3. With the shared joint network on Gamma data, DISCO-CONT has higher mean A_{top} at both graph densities, although ordering is less accurate on ER1 than on ER3.

H.3 REAL DATA: LAHMAN BATTING COUNTS

Data and preprocessing. We use the Lahman batting table through 2025 (Lahman, n.d.). For each season from 2019 to 2025, we retain complete stints, select the 200 players with the most at-bats, aggregate retained stints by player, and pool the seven seasons, yielding $n = 1{,}400$ player–season observations from 490 players on $d = 17$ batting-count variables. The main analysis uses a four-level quantile representation (q4) to increase the number of observations per count level for CCS estimation. Each pooled column is cut at its empirical quartiles, with no additional support-rank recoding after binning. This support compression is a finite-sample preprocessing choice and need not preserve the conditional-independence structure of the raw counts; we therefore examine alternative support resolutions and raw counts below.

Order evaluation. Standard batting definitions (Major League Baseball, n.d.) motivate the seven domain-informed directions

$$\mathcal{R}_7 = \{\mathrm{AB}\to\mathrm{H}, \mathrm{AB}\to\mathrm{SO}, \mathrm{AB}\to\mathrm{GIDP}, \mathrm{H}\to\mathrm{2B}, \mathrm{H}\to\mathrm{3B}, \mathrm{H}\to\mathrm{HR}, \mathrm{BB}\to\mathrm{IBB}\}.$$

These accounting or containment relations are used only for post-fit diagnostics and are not supplied during fitting. Because several imply parent-dependent feasible ranges, they need not satisfy the parent-independent-support assumptions of Definition 3. We therefore interpret A_{top} as agreement with a partial directional reference, not as full-DAG causal accuracy or as evidence that the Lahman distribution belongs exactly to the proposed model class.

For reference-set sensitivity, we additionally use $\mathcal{R}_{17} = \mathcal{R}_7 \cup \mathcal{R}_{\mathrm{PP}}$, where

$$\begin{aligned}\mathcal{R}_{\mathrm{PP}} = \{&\mathrm{G}\to\mathrm{H}, \mathrm{G}\to\mathrm{BB}, \mathrm{G}\to\mathrm{SO}, \mathrm{G}\to\mathrm{RBI}, \mathrm{AB}\to\mathrm{BB}, \mathrm{AB}\to\mathrm{RBI},\\ &\mathrm{H}\to\mathrm{R}, \mathrm{H}\to\mathrm{SO}, \mathrm{HR}\to\mathrm{IBB}, \mathrm{SB}\to\mathrm{CS}\}.\end{aligned}$$

The ten additional directions are extracted from the qualitative relationships discussed in the MLB analysis of Park and Park (2019b); neither $\mathcal{R}_7$ nor $\mathcal{R}_{17}$ is treated as a known causal DAG.

Method settings. General replication, evaluation, and estimator settings follow Appendix H.1, and baseline implementations follow Appendix H. For Lahman, the q4 codes are supplied directly without an additional support-rank transformation. DISCO uses seeds 0–2, while the ODS family-specification analysis uses three paired cross-validation assignments shared across the Poisson, negative-binomial, and binomial fits so that only the supplied family specification changes. All fits use the same fixed cohort, so the reported variation reflects fitting or cross-validation variability rather than sampling uncertainty.

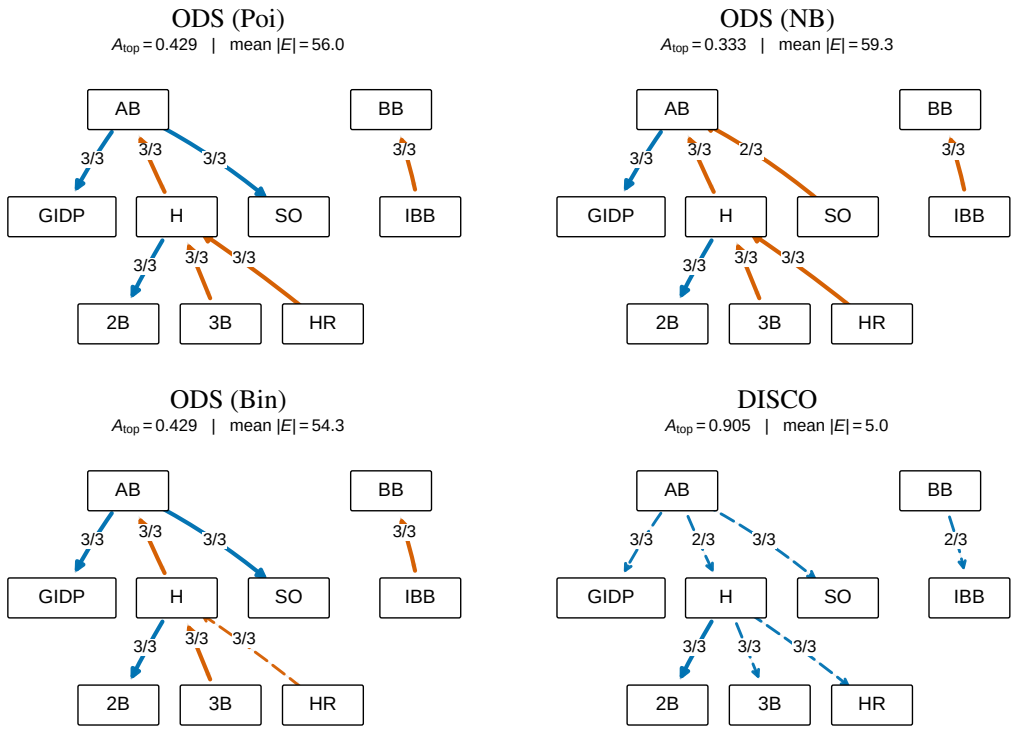


Figure 4: Ordering and parent-edge diagnostics on the Lahman 2019–2025 q4 representation. The top row shows ODS with Poisson and negative-binomial specifications; the bottom row shows ODS with a binomial specification and DISCO. Blue arrows agree with the reference direction and orange arrows reverse it. Arrow labels report ordering support across the three runs. Solid arrows denote parent edges selected in at least two runs, whereas dashed arrows denote majority ordering without majority parent-edge selection. The reference pairs are used only for post-fit evaluation.

Table 12: Pairwise sensitivity of ODS to the supplied conditional family on the Lahman q4 representation. Entries are mean $\pm$ standard deviation across three paired cross-validation assignments. Pairwise SHD compares two fitted directed graphs, not an estimated graph against a causal ground truth.

ODS pair	Order agreement	Edge Jaccard	Pairwise SHD
Poi–NB	0.882 ± 0.007	0.747 ± 0.023	16.67 ± 1.53
Poi–Bin	0.988 ± 0.008	0.924 ± 0.021	4.33 ± 1.15
NB–Bin	0.880 ± 0.004	0.713 ± 0.019	19.00 ± 1.00

Family-specification sensitivity. Under the main q4 setting, the three DISCO orders respect $6/7$, $7/7$, and $6/7$ relations in $\mathcal{R}_7$, yielding mean $A_{\text{top}} = 0.905$. The corresponding mean values for ODS under the Poisson, negative-binomial, and binomial specifications are 0.429, 0.333, and 0.429, respectively. Figure 4 summarizes ordering support and majority parent-edge recovery on these seven reference relations. Because the reference set is incomplete, non-reference edges are not interpreted as false causal edges.

The three ODS fits are compared pairwise in Table 12. Poisson and binomial yield similar estimated orders, whereas comparisons involving the negative-binomial specification show larger changes in both ordering and edge set. Because the q4 codes are compressed counts rather than the original count scale, these results are interpreted as sensitivity to the supplied ODS specification rather than as correctly specified family comparisons.

Table 13: Boundary and support-resolution sensitivity of DISCO on Lahman. The first row reproduces the main q4 analysis; the remaining rows use admissible finite differences with the same DISCO training configuration.

Representation	Boundary	$A_{\text{top}}^{\mathcal{R}_7}$	$A_{\text{top}}^{\mathcal{R}_{17}}$
q4	clipped	0.905	0.843
q4	admissible	0.762	0.804
q6	admissible	0.762	0.765
q8	admissible	0.762	0.784
raw	admissible	0.000	0.137

Table 14: ODS sensitivity on the raw Lahman counts. The binomial row uses empirical nodewise support maxima and is not an oracle binomial specification.

Supplied family	$A_{\text{top}}^{\mathcal{R}_7}$	$A_{\text{top}}^{\mathcal{R}_{17}}$	Mean $\lvert\widehat{E}\rvert$
Poisson	0.333	0.412	73.7
Negative binomial	0.857	0.843	69.3
Binomial	0.286	0.392	79.7

Boundary and support-resolution sensitivity. The main q4 implementation evaluates upper-grid shifts by clipping, whereas the population CCS is defined on the admissible finite-difference domain. As a reproduction check, the original q4 configuration again yields $6/7$, $7/7$, and $6/7$ agreement across the three runs, giving mean $A_{\text{top}}^{\mathcal{R}_7} = 0.905$. We then hold the DISCO training configuration fixed and replace only curvature evaluation by the admissible-domain rule. Under this rule, $A_{\text{top}}^{\mathcal{R}_7} = 0.762$ for each of q4, q6, and q8, while the corresponding $\mathcal{R}_{17}$ values are 0.804, 0.765, and 0.784. Thus the absolute q4 result is sensitive to boundary handling, whereas the admissible-domain ordering agreement is stable across the examined compressed support resolutions.

The identical $\mathcal{R}_7$ agreement across q4–q8 shows that the admissible-domain result is not specific to a single quantile resolution. The substantially lower raw-count agreement indicates that support compression is important in this finite-sample Lahman analysis.

Preprocessing sensitivity. Because quantile compression changes the raw count mean–variance relationship used by QVF-based ODS, we additionally evaluate ODS on the uncompressed Lahman counts. For this sensitivity analysis, the Poisson and negative-binomial variants use log-link regressions, while the binomial variant uses a logit link with empirical nodewise support maxima. On the raw counts, $A_{\text{top}}^{\mathcal{R}_7}$ is 0.333, 0.857, and 0.286 for the Poisson, negative-binomial, and binomial specifications, respectively, with corresponding $\mathcal{R}_{17}$ values 0.412, 0.843, and 0.392. Thus ODS can attain high directional agreement under one supplied family, but the result varies substantially with that specification; these results are therefore treated as preprocessing and family-specification sensitivity rather than as an oracle comparison with DISCO.

Additional robustness checks. For the 2019–2025 q4 analysis, the three DISCO orders agree with $\mathcal{R}_{17}$ on $15/17$, $15/17$, and $13/17$ relations, giving mean $A_{\text{top}} = 0.843$. Repeating the same cohort construction and q4 preprocessing for 2012–2018 gives mean agreement 0.952 on $\mathcal{R}_7$ and 0.922 on $\mathcal{R}_{17}$. These checks show that the reported directional pattern is not confined to one reference set or one analysis period, while remaining descriptive partial-order diagnostics rather than causal validation.

I Algorithm Diagnostics

I.1 Marginal Projection Methods

We compare three projection procedures on DAGs with $d = 50$ and $n = 5000$, using the same mechanisms for the three conditional families as in Appendix I.2. They share the data, three-fold splits, fixed joint networks, the first sink selected from all variables, and initial projection networks.

Additional training settings. We select the joint network by the held-out score loss. Validation uses 10% of each training fold, starts at epoch 300, and is checked every 100 epochs; training stops after two checks without an improvement of at least 10^{-4}. Projection networks use a learning rate of 0.0015, a batch size of 512, and FP32 arithmetic with TF32 disabled.

Amortized projection uses this projection network throughout ordering without further training. After each removal, *Stagewise direct* projection initializes the projection network with its previous parameters and trains it for 100 additional epochs, using predictions from the fixed joint network as regression targets. *Stagewise recursive* projection uses the same training budget but obtains targets from the preceding projection network, whose parameters are held fixed. Both procedures use the joint network for the first reduced-set fit. Thus each stagewise procedure adds 48 fits per fold, for remaining-set sizes 49 through 2. Training targets and projection predictions use the same fully corrupted sample and noise level. No joint concrete-score network is refitted on a reduced set.

Parent selection uses the recovered orders with the same fixed amortized projection networks and default OCS rule, isolating the ordering procedure. Total computation budgets are not matched. Projection time includes fitting the shared initial projection network, the additional training at each removal, and saving network parameters, but excludes joint training, CCS evaluation, and parent selection.

Table 15: Comparison of marginal concrete-score projection methods ($d = 50$, $n = 5000$). Recovery metrics are means $\pm$ sample standard deviations, with seeds paired across methods within each conditional family. Projection time is the mean wall-clock time for the projection stage, in seconds.

Conditional family	Projection	A_{top}	F_1	SHD	Projection time (s)
Poi	Amortized	0.931 ± 0.017	0.892 ± 0.034	28.9 ± 8.7	92.6
	Stagewise direct	0.917 ± 0.016	0.874 ± 0.034	34.0 ± 9.2	511.9
	Stagewise recursive	0.916 ± 0.013	0.869 ± 0.032	35.1 ± 8.2	544.5
NB	Amortized	0.931 ± 0.017	0.792 ± 0.066	51.2 ± 15.4	94.0
	Stagewise direct	0.923 ± 0.019	0.785 ± 0.055	53.3 ± 13.1	515.0
	Stagewise recursive	0.924 ± 0.021	0.788 ± 0.053	52.6 ± 12.9	545.2
Bin	Amortized	0.901 ± 0.027	0.877 ± 0.031	32.3 ± 8.4	94.2
	Stagewise direct	0.901 ± 0.032	0.873 ± 0.036	33.4 ± 9.7	510.3
	Stagewise recursive	0.900 ± 0.025	0.874 ± 0.035	33.1 ± 9.2	542.0

Amortized projection achieves comparable or higher mean A_{top}, higher mean F_1, lower mean SHD, and shorter projection time in each conditional family. These differences are not uniform across seeds.

I.2 Parent Selection

We hold the data, estimated CCS orders, and foldwise OCS estimates fixed and vary only the parent-selection rule on DAGs with $d = 50$ and $n = 5000$. We use the first Poi and NB mechanisms and the second Bin mechanism in Table 5. The held-out folds collectively cover all 5000 observations, and no network is retrained for this comparison.

For the threshold-sensitivity analysis, the median/IQR standardization is unchanged. We vary $\tau_Z \in \{1, 1.5, 2, 2.5, 3\}$ and $\pi_{\min} \in \{2/3, 1\}$: a candidate must satisfy the strict upper-tail test $Z^{(b)}_{j \to i} > \tau_Z$ in at least two or all three folds, respectively.

Table 16: Parent-selection sensitivity for the threshold settings shown, with fixed learned orders and OCS estimates ($d = 50$, $n = 5000$). Entries are arithmetic means over the same seeds. Prec. and Rec. denote precision and recall against the full true DAG, including true edges excluded by the estimated order.

	Poi			NB			Bin		
Rule	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.
$\tau_Z = 2,\ \pi_{\min} = 1$	0.877	0.934	0.827	0.797	0.917	0.708	0.878	0.943	0.821
$\tau_Z = 2,\ \pi_{\min} = 2/3$	0.891	0.898	0.884	0.870	0.898	0.845	0.890	0.895	0.885
$\tau_Z = 1,\ \pi_{\min} = 1$	0.900	0.898	0.901	0.866	0.892	0.843	0.896	0.906	0.885
$\tau_Z = 1,\ \pi_{\min} = 2/3$	0.800	0.709	0.917	0.813	0.735	0.910	0.809	0.732	0.907

Table 16 shows that lowering τ_Z from 2 to 1 while retaining all-fold agreement, or requiring two rather than three folds at $\tau_Z = 2$, increases mean F_1 in all three conditional families. For negative binomial, relaxing fold agreement alone raises F_1 from 0.797 to 0.870. Relaxing both requirements increases false positives and lowers F_1 relative to the one-at-a-time changes. Thus the same OCS estimates can yield different recovery under different selection rules. No threshold is selected or tuned against the true graph; this is not an error-control procedure and does not replace the default $\tau_Z = 2$, $\pi_{\min} = 1$.

Comparison of parent recovery methods. To separate parent selection from ordering, we compare the default DISCO OCS selector with PCM-GAM, a conditional-mean independence test (Lundborg et al., 2024), and CAM pruning, an additive-regression variable-selection procedure (Bühlmann et al., 2014). Under a valid causal order in our semiparametric GLM model, $\mathbb{E}[X_i \mid X_{\mathrm{Pred}(i)}] = A_i'(\eta_i)$. Since $A_i'' > 0$, the nonconstant parent effects in Definition 3, together with the support regularity conditions, make conditional-mean dependence on a predecessor equivalent to parenthood. Thus the population criterion targeted by PCM identifies true parents; this does not guarantee exact recovery by the finite-sample PCM-GAM test.

We use a separate study with $d = 50$ and $n = 5000$, under both the true order and the same estimated DISCO order. Family-specific generators are unchanged from the study above. Each dataset is split into 2500 ordering observations and 2500 parent-recovery observations; PCM further splits the latter into 1250 nuisance-training and 1250 testing observations. This disjoint-sample protocol is distinct from the cross-fitted threshold study above.

PCM-GAM uses the public `comets::pcm` implementation of the projected covariance measure (Lundborg et al., 2024), with GAM mean/projection regressions (basis parameter $k = 4$) and a 500-tree random forest for the variance nuisance. Each candidate is tested at $\alpha = 0.05$ without multiplicity correction. CAM uses the DiffAN CAM pruning routine (Bühlmann et al., 2014; Sanchez et al., 2023) with cutoff 0.001. PCM-GAM and CAM receive raw counts without the true conditional family or link. These are method-specific settings, not tests at a common nominal level; using GAM nuisance estimates does not itself establish count-specific test calibration.

With the true order, CAM has the highest mean F_1 and lowest SHD in every conditional family (Table 17). With the shared estimated order, OCS outperforms PCM-GAM in all three families and has slightly higher F_1 and lower SHD than CAM for Poi; CAM remains strongest for NB and Bin. The OCS gap under the true order reflects limitations of learned scores and the practical selector, rather than the population OCS characterization; this comparison does not separate score-estimation error from thresholding error.

Table 17: Parent recovery under fixed orders ($d = 50$, $n = 5000$). Entries are arithmetic means over seeds shared by all methods and both order conditions. Metrics use the full true DAG, including edges excluded by an estimated order. Estimated denotes the shared DISCO order; OCS denotes the default DISCO parent selector.

Conditional family	Order	Parent selector	F_1	SHD
Poi	True	OCS	0.878	30.8
		PCM-GAM	0.839	54.1
		CAM	0.989	3.2
	Estimated	OCS	0.858	37.2
		PCM-GAM	0.712	105.6
		CAM	0.850	46.3
NB	True	OCS	0.780	50.5
		PCM-GAM	0.832	56.3
		CAM	0.959	12.1
	Estimated	OCS	0.751	59.2
		PCM-GAM	0.703	108.3
		CAM	0.823	54.9
Bin	True	OCS	0.820	43.1
		PCM-GAM	0.827	58.3
		CAM	0.993	1.9
	Estimated	OCS	0.828	42.6
		PCM-GAM	0.738	90.3
		CAM	0.898	28.7