fDAG-ACDC: Uncertainty-Adaptive Factor-Count Selection for High-Dimensional Causal Discovery

Anonymous authors
Paper under double-blind review

Abstract

Causal representation learning (CRL) methods aim to discover low-dimensional representations whose components correspond to the mechanisms that generate high-dimensional observations. Changing the number of latent components affects the interpretation of the inferred mechanisms, yet the correct number of mechanisms is rarely known. Hence there is a need for model-selection criteria to determine the number of components. However, existing criteria are generally inapplicable to CRL models because adding latent structure can improve global fit by absorbing misspecification rather than capturing distinct mechanisms. In this work, we address this gap, focusing on factor-directed acyclic graphs (f-DAGs) as a test case. We introduce fDAG-ACDC, an uncertainty-adaptive extension of the accumulated cutoff discrepancy criterion (ACDC), which compares observed and posterior-predictive distributions of each node and its estimated parents within each perturbation environment. Theoretically, we prove model selection consistency of fDAG-ACDC under intuitive conditions. In simulations, fDAG-ACDC's mean absolute selection error is just 0.5, while validation likelihood and oracle graph-recovery baselines over-select by more than 3 on average. On Perturb-seq and organoid data, fDAG-ACDC selects fewer factors than validation likelihood while matching or nearly matching larger models in predicting biological summaries, and its Perturb-seq factors are enriched for known gene programs.

1 Introduction

Causal representation learning (CRL) seeks low-dimensional representations whose components correspond to the mechanisms that generate high-dimensional observations (Schölkopf et al., 2021; Moran & Aragam, 2026). Such representations cannot be identified from independent and identically distributed observations alone (Locatello et al., 2019), so CRL methods rely on additional information or assumptions such as auxiliary variables or weak supervision (Khemakhem et al., 2020; Locatello et al., 2020; Brehmer et al., 2022), interventional data (Ahuja et al., 2023; Squires et al., 2023; von Kügelgen et al., 2023; Zhang et al., 2023; Varıcı et al., 2025), sparsity of the latent mechanisms or of the decoder (Lachapelle et al., 2022; Moran et al., 2022), or a grouping of the observed variables (Morioka & Hyvärinen, 2024). For example single-cell experiments measure expression levels of thousands of genes, including across hundreds or thousands of genetic perturbations (Dixit et al., 2016; Adamson et al., 2016; Replogle et al., 2022). To use this data to gain insights into the underlying biological mechanisms, CRL methods have been developed that assume the existence of gene programs and regulatory structure that provide interpretability and leverage perturbation information when it is available (Kotliar et al., 2019; Lopez et al., 2022; 2023; Gu et al., 2026).

Whatever particular structure a CRL method relies on, it must specify how many latent mechanisms the representation contains. Identifiability results either assume this number is known (von Kügelgen et al., 2023; Brehmer et al., 2022) or identify it under a correctly specified and usually noiseless model (Zhang et al., 2023; Buchholz et al., 2023; Jiang & Aragam, 2023). In cases with interventions, existing guarantees require at least one intervention per latent variable (Squires et al., 2023; Varıcı et al., 2025). In applications, the number of latent mechanisms is rarely known in advance (Moran & Aragam, 2026), and it is typically fixed without a stated criterion (Lopez et al.,

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.

2023; Bereket & Karaletsos, 2023; Gu et al., 2026). Choosing the correct number of latent mechanisms is important because the choice alters the interpretation of every component: with too few, distinct mechanisms are merged and some dependencies are omitted, while with too many, mechanisms are split, or spurious components appear that absorb misspecification that the model cannot otherwise explain.

Model selection and the effects of misspecification both remain underexplored for CRL (Moran & Aragam, 2026). Information criteria such as AIC and BIC require an effective parameter count (Akaike, 1974; Schwarz, 1978), which is not well defined for the overparameterized neural models used in CRL. Predictive scores such as WAIC, leave-one-out cross-validation, and validation likelihood avoid this requirement (Watanabe, 2010; Vehtari et al., 2017). However, they target predictive risk rather than the correct number of mechanisms. When the model is misspecified, as is often the case in practice, added latent structure can keep improving predictive fit by absorbing the misspecification rather than capturing distinct mechanisms, so predictive criteria overestimate the number of latent components as the sample size grows (Cai et al., 2021; Miller & Dunson, 2019; Li et al., 2026). Bayesian alternatives that place a prior on the number of components (Miller & Harrison, 2018) inherit the same behavior under misspecification (Cai et al., 2021), and the same concern applies to overfitted factor models that rely on shrinkage priors to remove superfluous factors (Bhattacharya & Dunson, 2011; Mauri & Dunson, 2026). Classical factor-number criteria assume an approximate linear factor model (Bai & Ng, 2002). Graph-recovery scores such as the structural Hamming distance target the causal structure directly, but they require the true causal graph, which is unavailable outside simulations.

Posterior predictive checks quantify how well a fitted model reproduces chosen aspects of the data (Gelman et al., 1996). The accumulated cutoff discrepancy criterion (ACDC) turns this model-criticism principle into a selection rule for models built from interpretable processes, such as mixture components or matrix-factorization factors, and it is consistent under misspecification when the discrepancies are defined at the level of individual processes (Li & Huggins, 2024; Li et al., 2026). The existing ACDC method is not applicable to CRL models because the latent components interact through a causal graph and it odes not account for heterogeneous interventional environments.

In this work, we address the need for a principled component-selection criterion that is applicable to CRL. As a concrete test case, we focus on factor-directed acyclic graphs (f-DAGS; Lopez et al., 2022). Continuous-optimization methods have made causal discovery practical for increasingly large graphs (Zheng et al., 2018; Brouillard et al., 2020; Bello et al., 2022; Cundy et al., 2021; Annadani et al., 2023), but they estimate a graph on a prespecified set of variables, and a graph with one node for every measured gene is difficult to estimate and interpret. An f-DAG instead introduces a small number of latent factor nodes and permits edges only between observed variables and factors, so that observed-variable effects are routed through the factors and the causal graph becomes more interpretable. Methods for fitting f-DAGs, from differentiable causal discovery (Lopez et al., 2022) to amortized Bayesian inference (Gu et al., 2026), condition on a chosen number of factors m, so the user must set m before fitting the model.

Approach. We propose fDAG-ACDC, an uncertainty-adaptive extension of ACDC to fitted f-DAGs. Each candidate graph defines one local structural unit for every observed node, consisting of the node and its estimated observed parents. fDAG-ACDC compares the observed and posterior-predictive distributions of each unit within each perturbation environment, then averages the local discrepancies to construct a factor count versus fit curve. We quantify the uncertainty across fitted curves to identify the candidates that are statistically competitive with the empirical minimum, then use the smallest count in the contiguous competitive block that contains that minimum. The procedure applies to f-DAGs fitted by any method that provides posteriors over the latent factors, and it does not require knowing either the intervention targets or the true graph. Thus, our construction adapts the component-level checks of the original ACDC criterion and replaces its fixed, problem-specific cutoff with a cutoff estimated from the noisy criterion curves.

Theoretical guarantees. We express the selection target through the *minimum-sufficient factor count* m_0, which is the smallest candidate that attains the minimum of a population criterion measuring local distributional mismatch. Several larger candidates may attain the same value, so the definition favors the least complex f-DAG with optimal local fit. Under uniform convergence of the fitted criterion, a valid uncertainty estimate, separation of candidates below m_0, and an appro-

priate cutoff value depending on the criterion-estimation rate, the pointwise selector recovers m_0 with probability tending to one, and the contiguous selector does so under an additional interval condition (Theorems 1 and 2). Our guarantee concerns the criterion-defined count m_0; identifying a unique generating factor count requires an additional identifiability condition, which we discuss in Section 2.3.

Numerical results. We complement our theoretical results with experiments on six simulation settings and two single-cell perturbation datasets (Section 4). In the simulations, the mean absolute selection error $\mathbb{E}[|\widehat{m} - m_0|]$ of the fDAG-ACDC estimator $\widehat{m}$ is 0.5, whereas validation likelihood and oracle graph-recovery baselines overestimate m_0 by 3 on average. On the two real datasets, fDAG-ACDC selects fewer factors than validation likelihood, which chooses the largest candidate, while matching or nearly matching larger models in predicting independent biological summaries, and the selected Perturb-seq factors are enriched for known unfolded-protein-response programs.

Contributions. In summary, we make the following contributions:

1. *A model-selection criterion for CRL.* We extend the local-adequacy principle of ACDC from mixture and matrix factorization-type models to models where latent components interact through a causal graph and are fit across heterogeneous interventional environments.
2. *Selection without ground truth or a parameter count.* Unlike information criteria, predictive scores, and graph-recovery metrics, our approach targets local structural adequacy, so additional factors are rewarded only when they improve the reproduction of local causal structure. This approach does not require any kind of effective parameter count or known intervention targets.
3. *An uncertainty-calibrated selector with guarantees.* We replace the fixed, problem-specific cutoff of the original ACDC criterion – which requires heuristics or additional information to choose – with a rule that accounts for the uncertainty in the estimated criterion. We prove that this rule provides consistent recovery of the minimum-sufficient factor count m_0.
4. *Validation on f-DAGs.* We instantiate the criterion for f-DAGs and show on simulations that it selects close to the true factor count where predictive and oracle graph-recovery scores over-select, and on two single-cell perturbation screens that it selects fewer factors than validation likelihood without reducing agreement with independent biological summaries.

2 Background and Problem Formulation

2.1 Interventional data and causal structure

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ denote d observed random variables. We observe samples from a finite collection of environments $\mathcal{R}$, so the data are $\mathcal{D}_N = \{(\boldsymbol{x}_i, r_i)\}_{i=1}^N$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ and with $r_i \in \mathcal{R}$ identifying the environment that generated sample i. The collection $\mathcal{R}$ may include an observational environment as well as interventions. We assume that the environment label is observed, but the target of an intervention need not be known. We write P_r for the distribution of $\boldsymbol{X}$ in environment r and n_r for the corresponding sample size.

A causal model for these data consists of a directed acyclic graph (DAG) G on $\mathcal{V} = [d]$ and a mechanism for each node given its parents. Under the causal Markov assumption, the density in environment r factorizes as

$$p_r(\boldsymbol{x}) = \prod_{j=1}^{d} p_{r,j}\big(x_j \mid \boldsymbol{x}_{\mathrm{pa}_G(j)}\big)\,, \tag{1}$$

where $\mathrm{pa}_G(j)$ is the parent set of node j in G, and $p_{r,j}$ is the same in every environment that does not target node j (Schölkopf et al., 2021).

When d is large, estimating and interpreting G directly is difficult: the number of DAGs grows super-exponentially in d, the conditions under which constraint-based methods recover G consistently become more restrictive as the graph grows (Uhler et al., 2013), and an unrestricted graph over a large number of variables is hard to interpret (Segal et al., 2005). In many systems, however, large groups of variables share a small number of mechanisms. Gene expression, for example, is

organized into programs of co-regulated genes (Segal et al., 2005; Kotliar et al., 2019; Heimberg et al., 2016), and perturbations of different genes converge on the same programs (Replogle et al., 2022). In the screen of Adamson et al. (2016), which we analyze in Section 4.2, the responses to about 100 perturbations decompose into components that track the three branches of the unfolded protein response. A causal representation should describe such a system through a small number of latent factors, each of which summarizes a mechanism shared by many observed variables (Moran & Aragam, 2026). Routing causal effects through these factors reduces the number of parameters and the size of the search space, and it yields a structure that can be interpreted in terms of mechanisms rather than individual variables (Segal et al., 2005; Lopez et al., 2022).

2.2 Causal representation learning with latent factors

Causal representation learning (CRL) augments the observed variables with latent factors and places the causal structure on the augmented system (Schölkopf et al., 2021; Moran & Aragam, 2026). Following Moran & Aragam (2026, Section 4), a CRL model with m latent factors $\boldsymbol{Z} = (Z_1, \dots, Z_m)$ is specified by a DAG B_m on $\mathcal{V} \cup \mathcal{F}_m$, where $\mathcal{F}_m = [m]$ indexes the factors, and by a mechanism for each node. Writing $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{z})$, the model for environment r is

$$p_r(\boldsymbol{x}, \boldsymbol{z}) = \prod_{u \in \mathcal{V} \cup \mathcal{F}_m} p_{r,u}\Big(w_u \mid \boldsymbol{w}_{\mathrm{pa}_{B_m}(u)}\Big), \qquad p_r(\boldsymbol{x}) = \int p_r(\boldsymbol{x}, \boldsymbol{z})\, \mathrm{d}\boldsymbol{z}, \tag{2}$$

and, as in equation 1, an intervention changes only the mechanisms of its targets, which may now include factors. The factor count m sets the resolution of the representation, and the model classes are nested in m because a factor without edges has no effect.

Two restrictions on B_m cover the settings of interest. (i) Under a *factor structure*, B_m has no edges between observed nodes and none from observed to latent nodes (Moran & Aragam, 2026, Definition 1). It then consists of a causal DAG over the factors and a bipartite measurement graph from the factors to the observed variables, which gives the nonlinear factor model $\boldsymbol{X} = f(\boldsymbol{Z}) + \boldsymbol{\varepsilon}$ with causally related factors, and interventions act on the factors. The identifiability results cited in Section 1 use this class, mostly with a dense mixing function f, and several works make the measurement graph explicit (Silva et al., 2006; Kivva et al., 2021; Jiang & Aragam, 2023; Moran et al., 2022). (ii) Under a *factor-mediated structure*, B_m has edges only between observed and factor nodes, so every causal effect between observed variables passes through a factor (Lopez et al., 2022). Each factor then summarizes a mechanism through which one group of observed variables affects another, such as a regulatory program acting on gene expression. In case (ii), the paths through the factors define a DAG over the observed variables, and our criterion in Section 3 is built from this DAG. Our approach can be extended to models of class (i) by building the local units from the measurement graph.

A factor-directed acyclic graph (f-DAG) is a DAG of class (ii) (Lopez et al., 2022). It is represented by a node-to-factor matrix $U_m \in \{0,1\}^{d \times m}$ and a factor-to-node matrix $V_m \in \{0,1\}^{m \times d}$, and its induced observed-node, or half-square, graph has adjacency matrix $A_m = U_m \circ V_m$, where $\circ$ denotes Boolean matrix multiplication. Thus, $(A_m)_{ij} = 1$ if and only if at least one factor lies on a directed two-edge path from observed node i to observed node j. The half-square graphs of f-DAGs with m factors are the adjacency matrices of Boolean rank at most m (Lopez et al., 2022), so an f-DAG with $m \ll d$ is a compact representation of shared causal structure.

Our criterion does not depend on how the f-DAG is estimated. For each factor count m, it requires only a fitted model $\widehat{\Theta}_m = (\widehat{G}_m, \widehat{q}_m, \widehat{p}_m)$ with three parts: a fitted graph $\widehat{G}_m$, which may summarize a posterior distribution over graphs, for example by its mode; an exact or approximate posterior $\widehat{q}_m(\boldsymbol{z} \mid \boldsymbol{x}, r)$ over the factors of an observation in environment r; and the fitted distribution $\widehat{p}_m(\boldsymbol{x} \mid \boldsymbol{z}, r)$ of the observations given the factors. The graph $\widehat{G}_m$ induces the observed parent set $\widehat{\mathrm{pa}}_m(j)$ for every node j, and a posterior-predictive replicate of an observed sample $(\boldsymbol{x}_i, r_i)$ is generated under the same environment by

$$\boldsymbol{z}_i^{\mathrm{rep}} \sim \widehat{q}_m(\boldsymbol{z} \mid \boldsymbol{x}_i, r_i), \qquad \boldsymbol{x}_i^{\mathrm{rep}} \sim \widehat{p}_m(\boldsymbol{x} \mid \boldsymbol{z}_i^{\mathrm{rep}}, r_i). \tag{3}$$

Variational methods supply $\widehat{q}_m$ and $\widehat{p}_m$ as their encoder and decoder, as in ABCDEFG, which fits f-DAGs with a masked variational autoencoder and can learn how interventions with unknown targets act on the factors (Gu et al., 2026). Methods whose factors are deterministic functions of their inputs, such as DCD-FG (Lopez et al., 2022), correspond to a point-mass posterior approximation.

2.3 SELECTING THE NUMBER OF COMPONENTS WITH ACDC

Let $\mathcal{M} = \{m_{\min}, \ldots, m_{\max}\}$ be a finite, ordered set of candidate numbers of latent components for nested model classes. Rather than assessing global fit, the accumulated cutoff discrepancy criterion (ACDC) asks whether each fitted component is an adequate approximation to a distinct process (Li & Huggins, 2024; Li et al., 2026). For a model with m components fitted to the data, ACDC assigns each component $k \in [m]$ a discrepancy $D_{m,k}$ between the part of the data attributed to that component and the component's fitted distribution. For example, the attributed data are the observations weighted by their posterior probabilities of membership in a mixture model, and the inferred component-level contributions in a matrix factorization model. Given a tolerance $\rho \geq 0$ for the misspecification of each component, the population criterion and its target are

$$C_\rho(m) = \operatorname*{agg}_{k\in[m]} \left(D_{m,k} - \rho\right)_+, \qquad \mathcal{A}_\rho = \operatorname*{arg\,min}_{m\in\mathcal{M}} C_\rho(m), \qquad m_0 = \min \mathcal{A}_\rho, \tag{4}$$

where agg is either the maximum, which requires every component to be adequate, or the sum used by ACDC, which is less sensitive to noisy estimates of individual discrepancies. Both vanish if and only if every component is within tolerance, so they define the same target whenever some candidate attains zero. The plug-in estimator $\widehat{m}$ is the smallest minimizer of the estimated criterion.

We call m_0 the *minimum-sufficient* number of components: larger candidates may attain the same minimum, but no smaller candidate does. It coincides with a data-generating number of components only under an additional identifiability condition, but it remains well defined as the most parsimonious adequate count when several representations fit the data equally well. For mixture and matrix factorization models, Li et al. (2026) show that ACDC recovers m_0 with probability tending to one when ρ equals the largest component discrepancy of the m_0-component model. The guarantee requires every model with fewer components to be separated from the data-generating distribution by a margin that depends on ρ. Crucially, however, ρ is problem-specific. Li et al. (2026) calibrate it on related datasets with known structure, which perturbation screens rarely provide, or read it from the stability of the selection over a range of tolerances, which requires a user-chosen stability width.

3 METHODOLOGY

Fitting an f-DAG separately for each $m \in \mathcal{M}$ produces the fitted models $\{\widehat{\Theta}_m : m \in \mathcal{M}\}$ of Section 2.2, and applying ACDC to them raises four challenges. (a) ACDC attributes each observation to separable contributions of the components (Li et al., 2026), whereas the nonlinear structure of an f-DAG means the factor-level contributions to an observation cannot be directly disentangled. (b) Interventions change mechanisms, so the data come from environments with different distributions. (c) The estimated parent sets, and hence the dependence structure, changes with m. (d) The tolerance ρ is unknown (Section 2.3).

3.1 LOCAL ADEQUACY CRITERION

We instantiate the population adequacy criterion $C(m)$ of Section 2.3 by comparing observed and posterior-predictive local distributions. To address challenge (a), we replace the components of ACDC with local structural units defined by the observed-node graph. Fix a candidate factor count m and let $G^\star_m$ denote the corresponding population observed-node graph. For node j, define its local structural unit as

$$K_{jm} = \{j\} \cup \mathrm{pa}_{G^\star_m}(j), \qquad k_{jm} = |K_{jm}|. \tag{5}$$

This neighborhood contains the child and its estimated observed parents. A fitted model is locally adequate at node j only if it can reproduce the joint behavior of these variables within each perturbation environment.

Let P_{jmr} and Q_{jmr} be the observed and posterior-predictive distributions of $\boldsymbol{X}_{K_{jm}}$ conditional on environment $R = r$. Let $\mathcal{R}_{jm}$ be the set of scorable environments for this local unit and let n_{jmr} be its number of valid observations in environment r. We define the population-level discrepancy as the weighted, rescaled average of the Kullback–Leibler (KL) divergences between the observed and posterior-predictive distributions:

$$D_j(m) = \frac{1}{k_{jm}} \sum_{r\in\mathcal{R}_{jm}} w_{jmr} \mathrm{KL}[P_{jmr} \,\|\, Q_{jmr}], \qquad w_{jmr} = \frac{n_{jmr}}{\sum_{u\in\mathcal{R}_{jm}} n_{jmu}}. \tag{6}$$

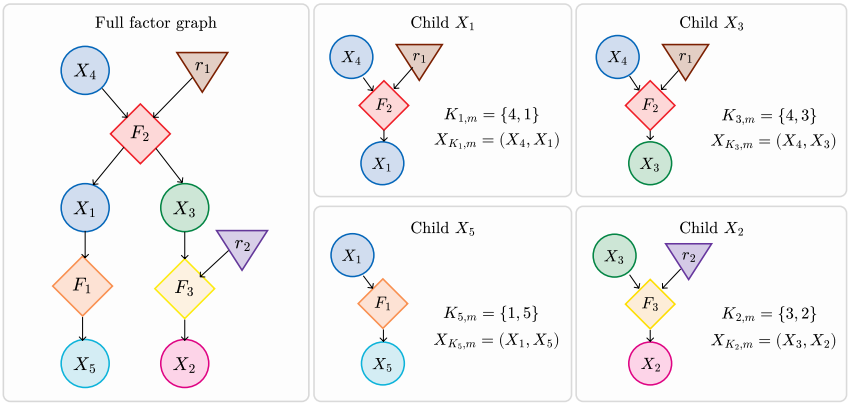


Figure 1: **Local structural units used by the ACDC criterion.** For each observed child node X_j, directed paths through intermediate factor nodes induce an observed parent set, defining $K_{jm} = \{j\} \cup \mathrm{pa}_{G^\star_m}(j)$. The vector $\boldsymbol{X}_{K_{jm}}$ contains the child and its observed parents, whose local distributions are compared conditional on environment $R = r$.

Conditioning on the environment addresses challenge (b) since the weights accumulate mismatch across environments without pooling different perturbation distributions. The normalization factor k_{jm}^{-1} addresses challenge (c) by converting the joint local KL into a per-variable discrepancy, which prevents larger local neighborhoods from being penalized solely because more variables are compared.

To estimate the population-level discrepancy and the estimation uncertainty, we consider replicate fits, where replicates can include different seeds for stochastic fitting procedures and bootstrapped datasets to account for sampling uncertainty. For replicate s, let K_{jsm} be the neighborhood induced by the fitted graph $\widehat{G}_{sm}$, let $k_{jsm} = |K_{jsm}|$, and let $\mathcal{R}_{jsm}$ be its scorable environments. The empirical distributions $\widehat{P}_{jsmr}$ and $\widehat{Q}_{jsmr}$ use the observed and posterior-predictive samples, respectively. If n_{jsmr} is the corresponding valid sample count, set $\widehat{w}_{jsmr} = n_{jsmr} / \sum_{u \in \mathcal{R}_{jsm}} n_{j,s,m,u}$. The empirical local discrepancy is

$$\widehat{D}_{js}(m) = \frac{1}{k_{jsm}} \sum_{r \in \mathcal{R}_{jsm}} \widehat{w}_{jsmr} \mathrm{KL}\left[\widehat{P}_{jsmr} \,\|\, \widehat{Q}_{jsmr}\right], \tag{7}$$

and the replicate-level criterion averages over the d observed nodes. We estimate the KL divergences using Gaussian local approximations with oracle-approximating shrinkage (OAS) covariance estimates (Chen et al., 2010); implementation details are in Appendix A. Because d does not depend on m, this average is comparable across candidates:

$$C_s(m) = \frac{1}{d} \sum_{j=1}^{d} \widehat{D}_{js}(m), \qquad C(m) = \frac{1}{d} \sum_{j=1}^{d} D_j(m). \tag{8}$$

3.2 Uncertainty-adaptive selection

To address challenge (d), we replace the fixed tolerance ρ with a cutoff estimated from the variability of the fitted criterion. Suppose S paired replicate curves $\{C_s(m) : m \in \mathcal{M}\}$ are available. At each candidate m, we estimate the criterion value with the robust median (Harrell & Davis, 1982),

$$\widehat{C}_S(m) = \sum_{i=1}^{S} a_{i,S} C_{(i)}(m), \quad a_{i,S} = \Pr\left(\frac{i-1}{S} < U_S \le \frac{i}{S}\right), \quad U_S \sim \mathrm{Beta}\left(\frac{S+1}{2}, \frac{S+1}{2}\right). \tag{9}$$

Let $\mathcal{M}_{\mathrm{sel}}$ be the eligible selection set and define

$$\widehat{b}_S \in \arg\min_{m \in \mathcal{M}_{\mathrm{sel}}} \widehat{C}_S(m), \qquad \widehat{\Delta}_S(m) = \widehat{C}_S(m) - \widehat{C}_S(\widehat{b}_S). \tag{10}$$

We estimate uncertainty in each gap by a nonparametric bootstrap over complete replicate curves (Efron & Tibshirani, 1994). If $\widehat{C}^*_{S,b}$ denotes the Harrell–Davis curve after bootstrap resampling, define the bootstrap standard deviation (SD) by

$$\widehat{s}_S(m,\ell) = \mathrm{SD}_b\left\{\widehat{C}^*_{S,b}(m) - \widehat{C}^*_{S,b}(\ell)\right\}. \tag{11}$$

Using the same resampled replicate indices for all candidates preserves the covariance between points on the same fitted curve.

For a multiplier κ_S, define the statistically competitive set

$$\widehat{\mathcal{A}}_S = \left\{m \in \mathcal{M}_{\mathrm{sel}} : \widehat{\Delta}_S(m) \le \kappa_S \widehat{s}_S(m, \widehat{b}_S)\right\}. \tag{12}$$

The pointwise selector is

$$\widehat{m}_S = \min \widehat{\mathcal{A}}_S. \tag{13}$$

Our default selector instead returns the left endpoint of the contiguous competitive block containing the empirical optimum:

$$\widehat{m}^{\mathrm{con}}_S = \min\left\{m \in \mathcal{M}_{\mathrm{sel}} : m \le \widehat{b}_S \text{ and } [m, \widehat{b}_S]_{\mathcal{M}} \subseteq \widehat{\mathcal{A}}_S\right\}. \tag{14}$$

This rule implements the minimum-sufficient principle while avoiding selection by an isolated finite-sample acceptance. We use $\kappa_S = \sqrt{\log S}$, which grows slowly enough that the effective cutoff vanishes under a root-S criterion-estimation rate. The complete post-fitting algorithm is given in Appendix A.

3.3 Selection consistency

ACDC aims to select the smallest structurally adequate factorization, not merely the empirical minimizer. We therefore ask whether its data-dependent cutoff recovers the leftmost population optimum as S grows. This requires stable criterion and paired uncertainty estimates, a positive gap for underfitted candidates, and, for the contiguous rule, an interval-valued optimum.

For this subsection, all minima and maxima are over the finite set $\mathcal{M}_{\mathrm{sel}}$. Let

$$\mathcal{A} = \arg\min_{m \in \mathcal{M}_{\mathrm{sel}}} C(m), \qquad m_0 = \min \mathcal{A}. \tag{15}$$

Assumption 1 (Stable fitted criterion)**.** *For a deterministic sequence $r_S \to 0$,*

$$\max_{m \in \mathcal{M}_{\mathrm{sel}}} |\widehat{C}_S(m) - C(m)| = O_{\mathbb{P}}(r_S), \tag{16}$$

where $\widehat{C}_S$ is the empirical central criterion used by the selector.

Assumption 2 (Valid paired uncertainty)**.** *Uniformly over $m, \ell \in \mathcal{M}_{\mathrm{sel}}$,*

$$\widehat{s}_S(m,\ell) = O_{\mathbb{P}}(r_S) \tag{17}$$

and paired comparisons among population minimizers are nondegenerate at this rate, or degenerate with both estimated gap and paired standard error equal to zero with probability tending to one. The formal version is stated in Appendix D.1.

Assumption 3 (Separated underfitting and adaptive cutoff)**.** *If $m_0 > \min \mathcal{M}_{\mathrm{sel}}$, then*

$$\delta_- := \min_{m < m_0} \{C(m) - C(m_0)\} > 0, \tag{18}$$

and the multiplier satisfies

$$\kappa_S \to \infty, \qquad \kappa_S r_S \to 0. \tag{19}$$

Assumption 4 (Interval-valued population optimum)**.** *The population-optimal set is order-convex:*

$$m, \ell \in \mathcal{A}, \quad m \le u \le \ell, \quad u \in \mathcal{M}_{\mathrm{sel}} \quad \Longrightarrow \quad u \in \mathcal{A}. \tag{20}$$

Theorem 1 (Selection consistency)**.** *Under Assumptions 1–3,*

$$\Pr(\widehat{m}_S = m_0) \longrightarrow 1. \tag{21}$$

Theorem 2 (Consistency of contiguous selection)**.** *Under Assumptions 1–4,*

$$\Pr(\widehat{m}^{\mathrm{con}}_S = m_0) \longrightarrow 1. \tag{22}$$

The first theorem establishes consistency of the pointwise selector; the second adds the interval condition needed by the contiguous rule. Thus, the adaptive cutoff asymptotically separates underfitting from statistically equivalent adequacy without a problem-specific tolerance. Proofs are in Appendix D.

Table 1: Simulation selections. TH/TS = targeted hard/soft; UH/US = untargeted hard/soft. F1 and SHD are oracles; MAE averages six settings.

Regime	Setting	m_0	Val. NLL	F1	SHD	ACDC
High	TH	10	13	13	13	10
High	TS	10	14	14	14	10
High	UH	10	13	14	14	9
High	US	10	12	13	13	12
Low	TH	5	9	8	8	5
Low	TS	5	8	9	8	5
		MAE	3.17	3.50	3.33	0.50

Table 2: Post-fitting ablation across the six simulation settings.

Variant	MAE
Full ACDC	0.50
No size adjustment	1.00
Non-contiguous	0.50
No paired uncertainty	3.00

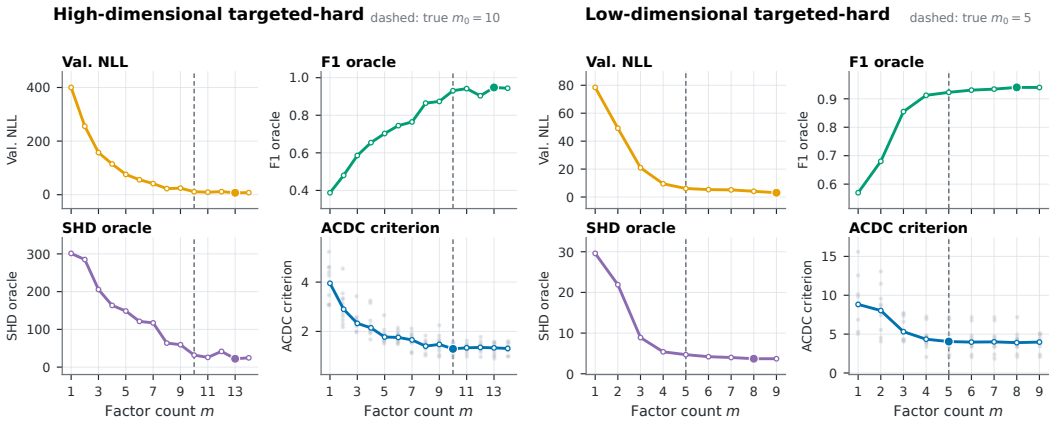


Figure 2: Targeted-hard selection curves for $(d, m_0) = (100, 10)$ (left) and $(20, 5)$ (right). Dashed lines mark m_0.

4 Empirical Experiments

We evaluate ACDC on synthetic f-DAGs with known factor counts and two single-cell perturbation datasets without causal ground truth. Each candidate is fitted with ABCDEFG (Gu et al., 2026); full data-generation, training, and validation details are given in Appendices B and C.

4.1 Simulation Study

The six settings vary dimension, target information, and intervention type. High- dimensional settings use $(d, m_0) = (100, 10)$ and $m = 1{:}15$; low-dimensional settings use $(d, m_0) = (20, 5)$ and $m = 1{:}10$. Each contains ten paired data-and-fitting replicates of 100,000 observations. Targeted settings provide the perturbed variable, whereas untargeted settings provide only the environment; hard interventions replace the target mechanism, while soft interventions modify it.

We compare ACDC with validation negative log-likelihood (Val. NLL), a practical predictive baseline, and graph F1 and structural Hamming distance (SHD), simulation-only oracles that use the true graph. Table 1 shows that all three over-select. ACDC is exact in four settings, within one factor in five, and attains mean absolute error (MAE) 0.50, versus 3.17, 3.50, and 3.33 for Val. NLL, F1, and SHD.

Figure 2 illustrates the difference in the targeted-hard settings: predictive and graph-recovery scores improve beyond m_0, whereas ACDC enters a statistically competitive low-discrepancy region at the minimum-sufficient count.

Ablation. Table 2 changes one component while holding all fitted models fixed, with $\mathrm{MAE} = 6^{-1} \sum_{q=1}^{6} |\widehat{m}_q - m_{0,q}|$. Removing size adjustment doubles MAE, and removing paired uncertainty causes the largest degradation. The pointwise and contiguous rules agree on the full ten-replicate

curves; we retain the latter because it matches the interval-valued population-optimum condition in Theorem 2. Definitions and subset-stability results are in Appendix B.

4.2 Real-data applications

Without a true causal graph, we test whether the selected factors explain independently defined biology: unfolded protein response (UPR) branch scores in Adamson Perturb-seq (Adamson et al., 2016) and published cell-type composition in the Amin organoid screen (Amin et al., 2024). We use gene set enrichment analysis (GSEA) and cross-validated R^2, following latent-factor variance-explained analyses (Argelaguet et al., 2018). Macro R^2 averages equally over three UPR branches or 11 cell subclasses; Amin's variance-weighted R^2 weights subclasses by total outcome variance. Definitions are in Appendix C.

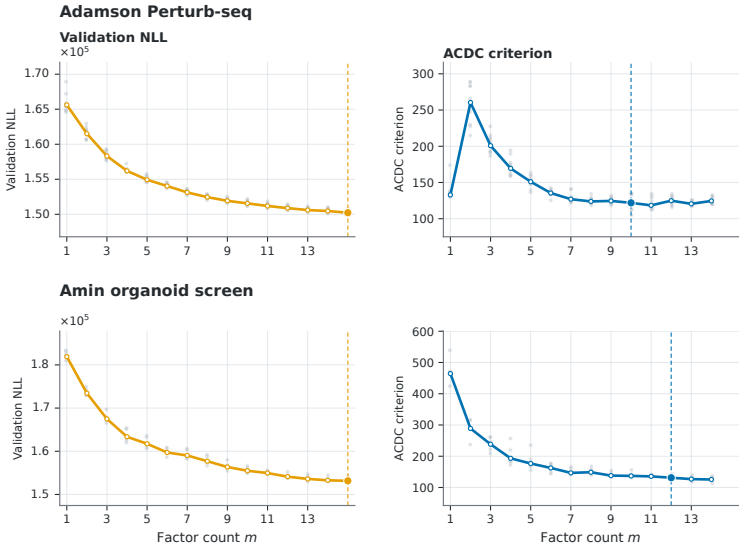


Figure 3: Validation NLL (left) and ACDC (right) for Adamson (top) and Amin (bottom). Gray points are runs; dashed lines and filled circles mark selections.

In Figure 3, each run's minimum validation NLL is smallest at the largest candidate, $m = 15$, in both datasets, whereas ACDC selects $m = 10$ for Adamson and $m = 12$ for Amin. For Adamson, every curated UPR and endoplasmic-reticulum (ER)-stress program is enriched in at least one selected factor. The selected model attains macro $R^2 = 0.494$, 90% of the grid-best 0.548 at $m = 14$ and comparable to the 0.481 at $m = 15$. For Amin, the selected model has the highest mean cell-composition score on the grid (macro $R^2 = 0.262$; variance-weighted $R^2 = 0.321$), although the scores for $m \geq 8$ differ by about one run-to-run standard deviation or less. Its marker-gene enrichment is spread across many factors rather than specific to individual cell types. Details are in Appendices C and B.4.

5 Conclusion

Overall, the experiments, together with our consistency theory, support fDAG-ACDC as a practical method for selecting compact factorized causal representations that balances structural adequacy and parsimony. In simulations, it recovers the minimum-sufficient factor count substantially more accurately than predictive and oracle graph-recovery criteria. Furthermore, our ablations suggest that the size adjustment and the paired uncertainty improve accuracy. On two perturbation screens, it selects fewer factors than validation likelihood without reducing agreement with independent biological summaries.

Several limitations remain. First, our guarantee concerns the criterion-defined count m_0 as the number of fitting replicates grows, identifies a data-generating count only under additional identifiability conditions, and requires refitting all candidate models for every replicate. Second, the Gaussian approximations used to estimate the KL divergences depend only on the first and second moments. Third, the real-data evidence is indirect, as it is difficult to obtain ground truth in most CRL settings. Finally, our experiments use one fitting method and only f-DAGs, so other fitting methods and CRL models with a factor structure (Section 2.2) remain to be tested.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and language polishing, literature retrieval and discovery, research ideation and execution, drafting parts of the manuscript, and supporting the development of mathematical arguments and proofs. AI-generated outputs were used as suggestions and were reviewed by the authors before being incorporated into the paper.

REFERENCES

Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, Ryan A. Pak, Andrew N. Gray, Carol A. Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan S. Weissman. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, 2016.

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

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Neal D. Amin, Kevin W. Kelley, Konstantin Kaganovsky, Massimo Onesto, Jin Hao, Yuki Miura, James P. McQueen, Noah Reis, Genta Narazaki, Tommy Li, Shravanti Kulkarni, Sergey Pavlov, and Sergiu P. Paşca. Generating human neural diversity with a multiplexed morphogen screen in organoids. *Cell Stem Cell*, 31(12):1831–1846.e9, 2024.

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. BayesDAG: Gradient-based posterior inference for causal discovery. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1738–1763, 2023.

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis: A framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, 2018.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8226–8239, 2022.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Anirban Bhattacharya and David B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21865–21877, 2020.

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

Diana Cai, Trevor Campbell, and Tamara Broderick. Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1158–1169, 2021.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Chris Cundy, Aditya Grover, and Stefano Ermon. BCD Nets: Scalable variational approaches for bayesian causal discovery. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7095–7110, 2021.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2022.

Yichen Gu, Yuxuan Song, Weizhou Qian, Yixin Wang, and Joshua D. Welch. Amortized bayesian causal discovery of extended factor graphs. *arXiv preprint arXiv:2607.22934*, 2026.

Frank E. Harrell and C. E. Davis. A new distribution-free quantile estimator. *Biometrika*, 69(3): 635–640, 1982.

Graham Heimberg, Rajat Bhatnagar, Hana El-Samad, and Matt Thomson. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Systems*, 2(4):239–250, 2016.

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

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

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217, 2020.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Dylan Kotliar, Adrian Veres, M. Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A. Melton, and Pardis C. Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife*, 8:e43803, 2019.

Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Megan G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie E. Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 428–484, 2022.

Jiawei Li and Jonathan H. Huggins. Structurally aware robust model selection for mixtures. *arXiv preprint arXiv:2403.00687*, 2024.

Jiawei Li, Nguyen Nguyen, Meng Lai, Ioannis Ch. Paschalidis, and Jonathan H. Huggins. Robust model selection for discovery of latent mechanistic processes. *arXiv preprint arXiv:2602.22062*, 2026.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359, 2020.

Romain Lopez, Jan-Christian Huetter, Jonathan K. Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Romain Lopez, Nataša Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan K. Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 662–691, 2023.

Lorenzo Mauri and David B. Dunson. Overfitted high-dimensional matrix factorizations via adaptive spectral shrinkage. *arXiv preprint arXiv:2606.19540*, 2026.

Jeffrey W. Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

Jeffrey W. Miller and Matthew T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

Gemma E. Moran and Bryon Aragam. Toward interpretable deep generative models via causal representation learning. *Journal of the American Statistical Association*, 121(553):259–275, 2026.

Gemma E. Moran, Dhanya Sridhar, Yixin Wang, and David M. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Eran Segal, Dana Pe'er, Aviv Regev, Daphne Koller, and Nir Friedman. Learning module networks. *Journal of Machine Learning Research*, 6:557–588, 2005.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

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

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

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

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

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

Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

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

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

A Additional Methodological Details

A.1 Gaussian local discrepancy

For replicate s, candidate factor count m, local structural unit j, and environment r, the empirical implementation estimates the observed and posterior-predictive distributions by multivariate Gaussian approximations:

$$\widehat{P}_{j,s,m,r} = \mathcal{N}\Big(\widehat{\boldsymbol{\mu}}^{\mathrm{obs}}_{j,s,m,r}, \widehat{\Sigma}^{\mathrm{obs}}_{j,s,m,r}\Big)\,, \tag{23}$$

$$\widehat{Q}_{j,s,m,r} = \mathcal{N}\Big(\widehat{\boldsymbol{\mu}}^{\mathrm{rep}}_{j,s,m,r}, \widehat{\Sigma}^{\mathrm{rep}}_{j,s,m,r}\Big)\,. \tag{24}$$

The superscripts obs and rep identify moment estimates computed from the observed local samples and their row-aligned posterior-predictive replicates, respectively. For k-dimensional Gaussian distributions $P = \mathcal{N}(\boldsymbol{\mu}_P, \Sigma_P)$ and $Q = \mathcal{N}(\boldsymbol{\mu}_Q, \Sigma_Q)$, we use

$$\mathrm{KL}(P\|Q) = \frac{1}{2}\left\{\mathrm{tr}(\Sigma_Q^{-1}\Sigma_P) + (\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)^\mathsf{T}\Sigma_Q^{-1}(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P) - k + \log\frac{\det\Sigma_Q}{\det\Sigma_P}\right\}. \tag{25}$$

A.2 Size adjustment

For independent coordinates with diagonal covariance matrices, Equation equation 25 decomposes exactly into a sum of univariate discrepancies,

$$\mathrm{KL}(P\|Q) = \sum_{a=1}^{k}\mathrm{KL}\big[\mathcal{N}(\mu_{P,a}, \sigma_{P,a}^2)\,\|\,\mathcal{N}(\mu_{Q,a}, \sigma_{Q,a}^2)\big]. \tag{26}$$

Thus, when the mismatch contributed by each coordinate remains of comparable order, the joint KL grows linearly with k even if model quality does not deteriorate.

The same leading-order behavior extends to correlated Gaussian distributions. Let

$$A = \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}, \qquad \boldsymbol{\delta} = \Sigma_Q^{-1/2}(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P), \tag{27}$$

and denote the eigenvalues of A by $\lambda_1, \ldots, \lambda_k$. Then Equation equation 25 can be written as

$$\mathrm{KL}(P\|Q) = \frac{1}{2}\left\{\sum_{a=1}^{k}\big(\lambda_a - \log\lambda_a - 1\big) + \|\boldsymbol{\delta}\|_2^2\right\}. \tag{28}$$

More explicitly, suppose that for constants $0 < c < C < \infty$ and $M < \infty$,

$$c \le \lambda_a \le C \quad (a = 1, \ldots, k), \qquad \|\boldsymbol{\delta}\|_2^2 \le Mk. \tag{29}$$

Since $f(\lambda) = \lambda - \log\lambda - 1$ is bounded on $[c, C]$, Equation equation 28 yields

$$\mathrm{KL}(P\|Q) \le \frac{k}{2}\left\{\sup_{\lambda\in[c,C]} f(\lambda) + M\right\} = O(k). \tag{30}$$

The linear scaling arises because covariance and mean mismatch accumulate across k spectral directions. Dividing by $k_{j,s,m}$ in the empirical criterion (and by $k_{j,m}$ in its population analogue) therefore estimates an average discrepancy per local variable and removes this leading size effect.

A.3 OAS covariance estimator

We use the commonly implemented simplified form of the oracle-approximating shrinkage (OAS) estimator for both covariance matrices in Equation equation 24 (Chen et al., 2010). Let $S_{j,s,m,r}$ denote either corresponding maximum-likelihood sample covariance based on $n_{j,s,m,r}$ observations, and let $k = k_{j,s,m}$. Define

$$\mu_{j,s,m,r} = \frac{\mathrm{tr}(S_{j,s,m,r})}{k}, \qquad \alpha_{j,s,m,r} = \frac{\mathrm{tr}(S_{j,s,m,r}^2)}{k^2}. \tag{31}$$

For $k > 1$ and a positive denominator, the shrinkage coefficient is

$$\widehat{\rho}_{j,s,m,r} = \min\left\{1,\ \frac{\alpha_{j,s,m,r} + \mu_{j,s,m,r}^2}{(n_{j,s,m,r} + 1)\left(\alpha_{j,s,m,r} - \mu_{j,s,m,r}^2/k\right)}\right\}. \tag{32}$$

If the denominator is nonpositive because of degeneracy or finite-precision rounding, we set $\widehat{\rho}_{j,s,m,r} = 1$. For $k = 1$, shrinkage does not change the covariance estimate, and we set $\widehat{\rho}_{j,s,m,r} = 0$. The resulting covariance estimate is

$$\widehat{\Sigma}_{j,s,m,r}^{\mathrm{OAS}} = (1 - \widehat{\rho}_{j,s,m,r})S_{j,s,m,r} + \widehat{\rho}_{j,s,m,r}\mu_{j,s,m,r}I_k + \varepsilon I_k, \tag{33}$$

where the final term is a small numerical safeguard. OAS is applied separately to the observed and replicated samples for every replicate, candidate factor count, environment, and local structural unit.

A.4 POST-FITTING ACDC ALGORITHM

Algorithm 1 Uncertainty-adaptive ACDC factor-count selection

Require: Ordered candidates $\mathcal{M}$, selection set $\mathcal{M}_{\text{sel}}$, S paired fitted models and datasets, environment labels, bootstrap size B
Ensure: Selected factor count $\widehat{m}_S^{\text{con}}$

for $s = 1, \dots, S$ **do**
  for $m \in \mathcal{M}$ **do**
    Generate a row-aligned posterior-predictive replicate under $\widehat{\Theta}_{s,m}$.
    for $j = 1, \dots, d$ **do**
      Set $K_{j,s,m} = \{j\} \cup \text{pa}_{\widehat{G}_{s,m}}(j)$ and $k_{j,s,m} = |K_{j,s,m}|$.
      Compute $\widehat{D}_{j,s}(m)$ by Equation equation 7, using separate within-environment OAS covariance estimates.
    end for
    Set $C_s(m) = d^{-1} \sum_{j=1}^{d} \widehat{D}_{j,s}(m)$.
  end for
end for
Compute the Harrell–Davis center $\widehat{C}_S(m)$ for every m.
$\widehat{b}_S \leftarrow \arg\min_{m \in \mathcal{M}_{\text{sel}}} \widehat{C}_S(m)$.
for $b = 1, \dots, B$ **do**
  Draw one size-S index vector and use it for every m to compute $\widehat{C}^*_{S,b}(m) - \widehat{C}^*_{S,b}(\widehat{b}_S)$.
end for
Estimate $\widehat{s}_S(m, \widehat{b}_S)$ from the bootstrap gaps and form $\widehat{\mathcal{A}}_S$ using $\kappa_S = \sqrt{\log S}$.
$q \leftarrow \widehat{b}_S$.
while q has a predecessor q^- in $\mathcal{M}_{\text{sel}}$ and $q^- \in \widehat{\mathcal{A}}_S$ **do**
  $q \leftarrow q^-$.
end while
return q.

B SIMULATION DETAILS AND REPRODUCIBILITY

This appendix records the data-generation, training, and post-fitting settings used in Section 4.1. All settings were fixed before applying any factor-count selector. Tables 3– 7 provide a compact reproducibility checklist.

B.1 SIMULATION DESIGN

Table 3 lists the six setting-specific designs. The number of environments includes the observational environment, and $|E_0|$ is the number of directed edges in the projected observed-variable graph.

Table 3: Setting-specific simulation parameters.

Regime	Intervention	d	m_0	Candidate m	Envs.	Targets	$\lvert E_0 \rvert$
High	Targeted hard	100	10	1:15	11	Yes	223
High	Targeted soft	100	10	1:15	11	Yes	223
High	Untargeted hard	100	10	1:15	12	No	274
High	Untargeted soft	100	10	1:15	12	No	274
Low	Targeted hard	20	5	1:10	11	Yes	30
Low	Targeted soft	20	5	1:10	11	Yes	30

For the bipartite factorization, let U denote the observed-to-factor adjacency, V the factor-to-observed adjacency, and W the intervention-to-factor adjacency in the untargeted settings. The hard and soft variants within each row of Table 4 use exactly the same graph and structural coefficients.

The reported nonzero counts were read from the saved simulation arrays, rather than replaced by their generator targets.

Table 4: Properties of the three fixed data-generating graph families. The degree target gives the requested mean factor parent/child degrees; realized bipartite and projected edge counts are reported separately.

Graph family	Shared by	Degree target	$\|U\|_0$	$\|V\|_0$	$\|W\|_0$	$\|E_0\|$
$d = 100$, targeted	Hard and soft	$5/5$	48	53	–	223
$d = 100$, untargeted	Hard and soft	$5/5$	47	60	10	274
$d = 20$, targeted	Hard and soft	$2/2$	10	13	–	30

The simulator evaluates observed and factor vertices in topological order. For an unintervened vertex v, the structural equation is

$$Y_v = \sum_{u \in \mathrm{pa}(v)} \beta_{uv} Y_u + \mathbf{1}\{v \text{ is observed}\}\epsilon_v, \qquad \epsilon_v \sim \mathcal{N}(0, 0.1^2). \tag{34}$$

Table 5 gives the remaining common generator settings and the precise intervention constructions.

Table 5: Common data-generation parameters.

Component	Setting
Graph generator	Enhanced acyclic bipartite generator; $p_{\text{vertex}} = p_{\text{factor}} = 0$; at least one unique observed parent and child per factor; graph dropout 0; no post-generation rescaling.
Graph seed	44 for every setting.
Data-and-fitting replicates	Ten seeds, $s = 0, \ldots, 9$; the graph and coefficients are fixed within a setting, while exogenous noise, intervention realizations, the data split, initialization, and stochastic optimization are regenerated.
Sample size	100,000 observations per seed, allocated as evenly as possible across environments (9,090–9,091 per environment when there are 11 environments; 8,333–8,334 when there are 12).
Data split	Random 60/20/20 training, validation, and test split, generated with the corresponding data seed; all candidate counts within a seed use the same observations and split.
Structural functions	Linear SEM in Equation equation 34; no hidden nonlinear layer ($n_{\text{hidden}} = 0$).
Nonzero coefficients	Independent Rademacher signs and magnitudes drawn uniformly from $[0.8, 1]$.
Exogenous noise	Independent $\mathcal{N}(0, 0.1^2)$ noise at observed vertices; factor vertices are noiseless.
Targeted environments	One observational and ten single-observed-target environments. Targets and environment labels are supplied to the fitted model.
Targeted hard intervention	The target equation is replaced by an independent $\mathcal{N}(0, 1)$ draw.
Targeted soft intervention	The target is evaluated using $B^{(r)} = B + E^{(r)}$, where entries of $E^{(r)}$ have independent signs and magnitudes drawn uniformly from $[1, 2]$; only vertices already evaluated in the causal order contribute.
Untargeted environments	One observational, ten single-factor, and one combined-factor environment. Environment labels are supplied, but the intervention-to-factor mapping is not.
Untargeted hard intervention	The affected factor equation is replaced by an independent $\mathcal{N}(0, 1)$ draw.
Untargeted soft intervention	The ordinary factor mechanism is retained and receives an intervention-parent dose drawn from a Gamma distribution with shape 0.1 and scale 0.1.

B.2 Model fitting and computation

We fit every candidate f-DAG with ABCDEFG (Gu et al., 2026), whose masked variational autoencoder is shown in Figure 4.

Table 6 reports the configuration used for every candidate fit. Across the six settings, this corresponds to 800 fitted models: four high-dimensional settings with 10×15 fits each and two low-dimensional settings with 10×10 fits each.

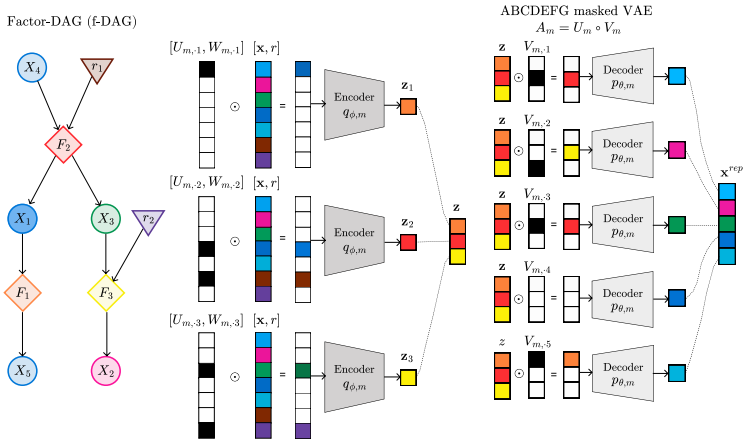


Figure 4: **ABCDEFG masked VAE architecture for an extended f-DAG.** The trainable f-DAG model generates binary edge matrices U_m, V_m, and W_m, where W_m maps interventions to factors. Columns of these matrices are applied as masks inside the encoder and decoder branches, determining both the direction of data flow and the learned causal connections.

Table 6: ABCDEFG training and computing configuration for the simulation study.

Component	Setting
Model	BasicFG version of ABCDEFG with a linear generative mechanism, hard sampled graph masks, temperature $\tau = 1$, and hidden width 1,000.
Training seed	Equal to the data seed; the same seed is used for every candidate count paired within a replicate.
Epochs	500 epochs.
Batch and precision	Batch size 128; 32-bit floating-point training; zero data-loader workers.
Optimizers	Two AdamW optimizers: learning rate 5×10^{-4} and weight decay 10^{-3} for neural parameters; learning rate 5×10^{-3} and no weight decay for factor-graph parameters.
Variational coefficients	Latent-state KL coefficient 10^{-8} and factor-graph KL coefficient 10^{-8}.
Graph sparsity penalties	Bipartite penalty $\|U\|_1 + \|V\|_1$ with coefficient 1.
Intervention sparsity penalty	Coefficient 0 for targeted fits and 10 for untargeted fits.
Coefficient schedule	Sigmoid schedule from 0 to 1 with midpoint epoch 500 and scale 125.
Within-fit checkpoint	Maximum validation directed-edge graph F1, as defined below.
Accelerator	NVIDIA A40 GPU; two CPU cores and 24 GB host memory.
Software environment	Python 3.11 (Anaconda 2024.02) with PyTorch Lightning; GCC 13.2.0 toolchain.

For simulation checkpointing, let E_0 be the directed edge set of the data-generating observed graph and let $\widehat{E}$ be the directed edge set estimated at a given training epoch. Treating directed edges as the positive class, we compute

$$\mathrm{Prec} = \frac{|\widehat{E} \cap E_0|}{|\widehat{E}|}, \qquad \mathrm{Rec} = \frac{|\widehat{E} \cap E_0|}{|E_0|}, \qquad F_1 = \frac{2\,\mathrm{Prec}\,\mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}. \tag{35}$$

Here Prec and Rec denote directed-edge precision and recall, with a score set to zero when its denominator vanishes. Reversing an edge therefore produces one false positive and one false negative. The retained checkpoint is the epoch with the largest validation graph F_1. The known graph is used

for this purpose only in the simulation study; all factor-count selectors are subsequently applied to the same retained candidate checkpoints.

B.3 ACDC EVALUATION AND ABLATIONS

The settings in Table 7 are applied after model fitting. Consequently, ACDC and all ablation variants reuse the same checkpoints and validation observations.

Table 7: Post-fitting ACDC evaluation parameters.

Component	Setting
Empirical sample	Validation split (20,000 rows per seed and candidate).
Predictive sample	One row-aligned posterior-predictive replicate per validation observation; posterior evaluation batch size 512.
Evaluated graph	Fixed hard graph from the fitted BasicFG masks; projected edge $i \to j$ is present when $(UV)_{ij} > 0$.
Local environments	Gaussian moments are estimated separately within each environment. When the direct observed target is known, that node is not scored in its directly targeted environment and its remaining environment weights are renormalized.
Covariance estimator	OAS estimated separately for observed and replicated local samples; numerical ridge $\varepsilon = 10^{-5}$.
Local size adjustment	Joint local KL divided by $k_{j,s,m} = 1 + \lvert \mathrm{pa}_{\widehat{G}_{s,m}}(j) \rvert$.
Node aggregation	Arithmetic mean over all observed nodes.
Replicate center	Harrell–Davis estimator of the median across the ten paired replicate curves.
Uncertainty	50,000 paired bootstrap resamples of complete replicate curves; multiplier $\kappa_S = \sqrt{\log S}$.
Selection rule	Left endpoint of the contiguous competitive block containing the empirical reference candidate.
Eligible candidates	$m = 1, \ldots, 14$ when $d = 100$ and $m = 1, \ldots, 9$ when $d = 20$; the largest fitted candidate is retained only as a search-range diagnostic.

Table 8 specifies the three post-fitting ablations reported in Table 2. Only the named component is changed in each row.

Table 8: Definitions of the ACDC component ablations.

Variant	Change from full ACDC
No size adjustment	Average the raw local joint KL values without dividing by $k_{j,s,m}$.
Pointwise instead of contiguous	Return the smallest individually competitive candidate without requiring a competitive path to the empirical reference.
No paired uncertainty	Return the empirical minimum of the central criterion curve.

B.4 BASELINE SELECTORS AND STABILITY DIAGNOSTICS

Validation negative log-likelihood. For run s, candidate factor count m, and training epoch t, let $\widehat{\mu}_{ij,s,m,t}$ and $\widehat{\sigma}_{ij,s,m,t}$ denote the decoder mean and standard deviation for variable j of validation observation i. Let M_{ij} be the corresponding likelihood mask, equal to zero when that variable's likelihood contribution is excluded by the intervention protocol and one otherwise. The validation negative log-likelihood (Val. NLL) is

$$L^{\mathrm{val}}_{s,m,t} = \frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \sum_{j=1}^{d} M_{ij} \left[\frac{(x^{\mathrm{val}}_{ij} - \widehat{\mu}_{ij,s,m,t})^2}{2\widehat{\sigma}^{,2}_{ij,s,m,t}} + \frac{1}{2} \log\Bigl(2\pi \widehat{\sigma}^{,2}_{ij,s,m,t}\Bigr) \right]. \tag{36}$$

Thus, Val. NLL averages over validation observations and sums over the observed variables retained by the likelihood mask. For each run and candidate, we retain the epoch

$$\widehat{t}_{s,m} \in \arg\min_{t} L^{\mathrm{val}}_{s,m,t}, \qquad \widehat{m}_{\mathrm{NLL}} \in \arg\min_{m \in \mathcal{M}_{\mathrm{sel}}} \frac{1}{S} \sum_{s=1}^{S} L^{\mathrm{val}}_{s,m,\widehat{t}_{s,m}}. \tag{37}$$

We use only the held-out reconstruction term and exclude latent-variable and graph regularization penalties. This puts candidates on a common predictive-fit scale: because larger m introduces more latent and graph parameters, the absolute value of an aggregate penalty can increase mechanically with model size even when its per-parameter behavior is unchanged.

Graph F1 selects the maximizer of mean directed-edge F1 against the true simulation graph, and SHD selects the minimizer of mean structural Hamming distance, with an edge reversal counted as one edit. The latter two are simulation-only oracles because they require the true graph.

For subset stability, we enumerate all $\binom{10}{S}$ subsets of the ten paired replicate curves for each $S \in \{6, \ldots, 10\}$ and reapply the full ACDC selector. Figure 5 shows that selections concentrate around the full-replicate values as the number of paired replicates increases.

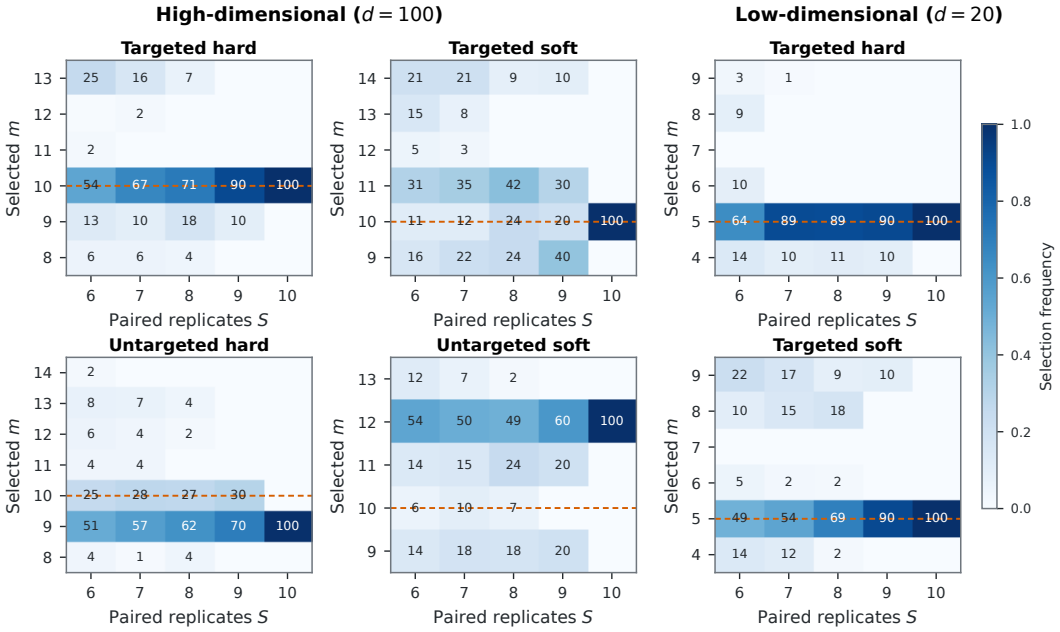


Figure 5: No-replacement subset stability of the ACDC selector. At each S, cell values are percentages over all $\binom{10}{S}$ subsets of the ten paired replicate curves. Dashed horizontal lines mark the data-generating factor count m_0.

C Real-Data Biological Validation

The analyses in this appendix are post-selection diagnostics: no biological annotation or validation score enters the ACDC criterion. They test whether the selected latent representations contain recognizable biology and predict an independently published phenotype.

C.1 Consensus factor–gene rankings

For every run, we reuse the exact checkpoint evaluated by ACDC. Here g indexes genes, $f \in \{1, \ldots, m\}$ indexes factors, s indexes fitting runs, and m is the candidate factor count. Let $\pi^{(s,m)}_{gf}$ be the fitted soft probability of the factor-to-gene edge $f \to g$. For the two-layer decoder used in the

real-data fits, we calculate the nonnegative local factor–gene sensitivity

$$I_{gf}^{(s,m)} = \pi_{gf}^{(s,m)} \left| \sum_h W_{gh}^{(2)} \mathbf{1}\{b_{gh}^{(1)} > 0\} W_{ghf}^{(1)} \right|. \tag{38}$$

where h indexes the hidden units in the gene-specific decoder, $W_{ghf}^{(1)}$ is the first-layer weight from factor f to hidden unit h for gene g, $b_{gh}^{(1)}$ is its corresponding bias, and $W_{gh}^{(2)}$ is the second-layer weight from that hidden unit to the output for gene g. The run and candidate superscripts (s, m) on these decoder parameters are suppressed for readability, and $\mathbf{1}\{b_{gh}^{(1)} > 0\}$ identifies ReLU units active at the latent prior mean. This is the magnitude of the decoder Jacobian $|\partial \widehat{x}_g / \partial z_f|$ at the latent prior mean, multiplied by the soft graph mask. To remove gene-wide decoder responsiveness shared by all factors, we divide each $I_{gf}^{(s,m)}$ by its mean over f and convert the resulting values to centered within-factor percentile ranks.

Factor labels are exchangeable across runs. For the selected m, one run is chosen as the medoid according to mean pairwise matched cosine similarity between its rank profiles and those of the other runs. Factors in every other run are aligned to the medoid by the Hungarian assignment maximizing total cosine similarity. Averaging the aligned rank profiles produces the consensus gene ranking for each displayed factor.

C.2 Preranked gene set enrichment analysis

We apply the weighted preranked GSEA statistic of Subramanian et al. (2005) to each consensus ranking. If S_p is the set of genes in program p, the running sum moves up by the absolute ranking statistic, normalized over genes in S_p, and moves down uniformly for genes outside S_p. The enrichment score is the signed excursion of largest absolute magnitude. We estimate its null distribution using 5,000 random gene sets of the same size, normalize positive and negative scores separately to obtain the normalized enrichment score (NES), and compute the pooled, sign-specific GSEA false-discovery rate (FDR). Thus, the cell colors in Figures 6 and 7 are NES values and the cell labels are the corresponding GSEA FDR q-values.

The Adamson analysis uses eight programs fixed before examining the consensus profiles: the Molecular Signatures Database (MSigDB) Hallmark unfolded-protein response and Reactome definitions of the ATF6, IRE1, PERK, and ATF4 responses, ER-associated degradation, N-linked glycosylation, and ER protein targeting (Liberzon et al., 2015; Gillespie et al., 2022). The source definitions were retrieved through Enrichr (Kuleshov et al., 2016) and fixed before analysis. The analysis uses only genes retained in the fitted data; program overlaps range from 6 to 34 genes.

For Amin, marker definitions come directly from Tables S3a and S3b of Amin et al. (2024). Cluster markers with positive average log-fold change and adjusted $p < 0.05$ are grouped according to the authors' 11 subclass annotations. We retain at most the 100 strongest available markers per subclass; after intersecting with the fitted genes, the sets contain 51–100 genes. This construction uses the published cell annotations rather than biological categories devised for the present analysis.

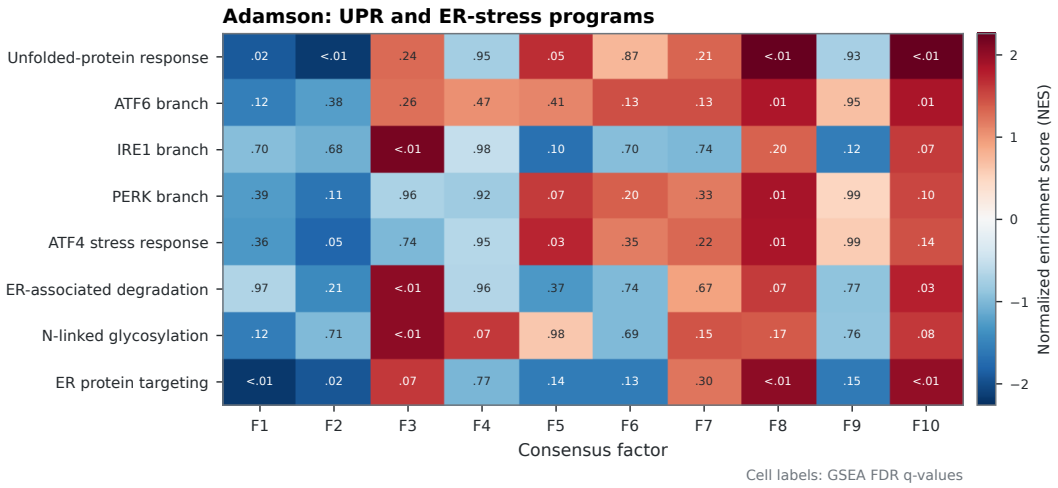


Figure 6: Preranked GSEA of Adamson consensus factors against UPR and ER-stress gene programs. Cell colors show normalized enrichment scores and labels show GSEA FDR q-values.

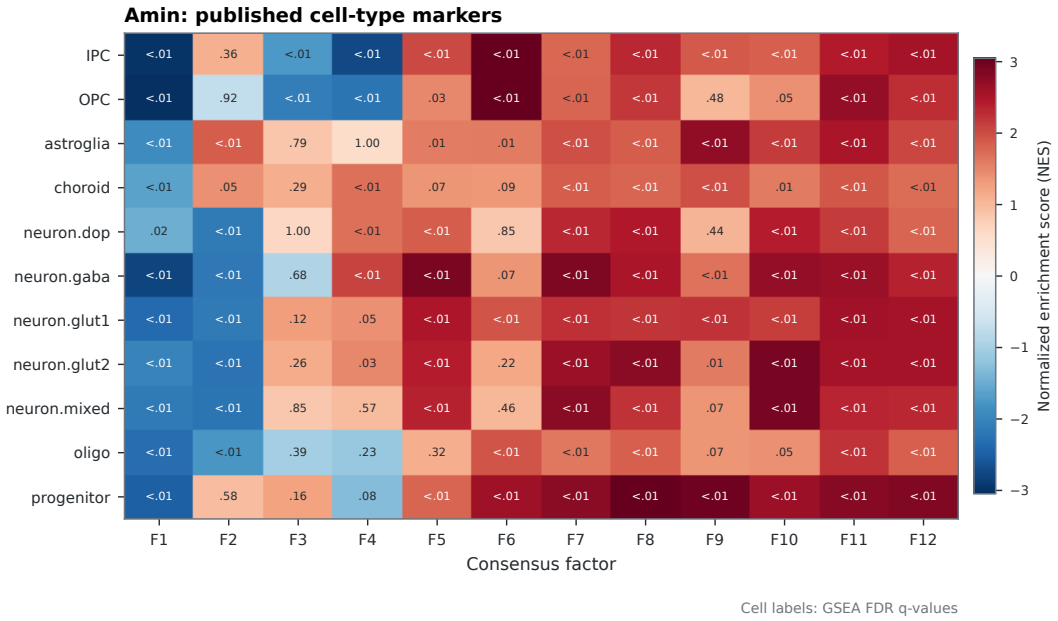


Figure 7: Preranked GSEA of Amin consensus factors against published cell-type marker families. Cell colors show normalized enrichment scores and labels show GSEA FDR q-values.

C.3 Cross-validated UPR branch-score prediction

For Adamson, we construct a perturbation-level branch score for each of the ATF6, IRE1, and PERK programs. Within each retained gene, expression is standardized across perturbation environments and centered by the control environment. The branch score is then the mean standardized response over the frozen Reactome genes retained in the fitted expression matrix. This gives a three-dimensional external biological response for each non-control perturbation.

For each fitting run and candidate $m = 1, \ldots, 15$, the predictors are the saved signed latent perturbation effects. We use the same nested five-fold ridge-regression readout as in the Amin composition

analysis below. For branch b, the coefficient of determination is

$$R_b^2 = 1 - \frac{\sum_r (y_{rb} - \widehat{y}_{rb})^2}{\sum_r (y_{rb} - \bar{y}_b)^2}, \qquad R^2_{\text{macro}} = \frac{1}{3} \sum_b R_b^2, \tag{39}$$

where $b \in \{\text{ATF6}, \text{IRE1}, \text{PERK}\}$ indexes UPR branches and r indexes perturbations held out by the outer cross-validation. Here y_{rb} is the observed score for branch b under perturbation r, $\widehat{y}_{rb}$ is its cross-validated prediction from the learned factor activities, and $\bar{y}_b$ is the observed mean of that branch score across perturbations. Thus, R^2_{macro} gives the three branches equal weight. This diagnostic is a lightweight external readout of UPR branch information in the learned factors; it is not the independent epistasis random-forest branch classifier used by Adamson et al. (2016).

At the ACDC-selected $m = 10$, the across-run mean macro R^2 is 0.494. The branch-specific means are 0.360 for ATF6, 0.581 for IRE1, and 0.541 for PERK. The best candidate by this diagnostic is $m = 14$ with mean macro $R^2 = 0.548$, so $m = 10$ reaches about 90% of the best observed UPR branch-prediction performance.

Table 9: Cross-validated Adamson UPR branch-score prediction across candidate factor counts. Macro R^2 is reported as mean $\pm$ standard deviation over ten runs; branch columns report across-run mean R^2. The ACDC-selected factor count is bolded.

m	Macro R^2	ATF6	IRE1	PERK
1	-0.032 ± 0.123	-0.040	0.004	-0.060
2	0.052 ± 0.141	-0.057	0.106	0.107
3	0.176 ± 0.108	-0.007	0.270	0.265
4	0.176 ± 0.130	-0.028	0.266	0.291
5	0.245 ± 0.130	0.024	0.356	0.355
6	0.315 ± 0.134	0.083	0.430	0.431
7	0.251 ± 0.224	-0.018	0.380	0.391
8	0.385 ± 0.112	0.168	0.516	0.472
9	0.426 ± 0.222	0.198	0.543	0.537
10	$\mathbf{0.494 \pm 0.117}$	**0.360**	**0.581**	**0.541**
11	0.473 ± 0.193	0.390	0.559	0.469
12	0.500 ± 0.199	0.424	0.516	0.561
13	0.540 ± 0.086	0.459	0.583	0.578
14	0.548 ± 0.112	0.452	0.612	0.580
15	0.481 ± 0.167	0.407	0.518	0.519

C.4 Cross-validated cell-composition prediction

Table S4 of Amin et al. (2024) reports cell counts for 47 clusters in each of 46 treatment conditions. We aggregate the counts using the 11 subclass labels in Table S3a and divide by the total count in each condition. For each fitting run and candidate $m = 1, \ldots, 15$, the predictors are the fitted condition-level normalized latent activities and the multivariate response is the vector of published subclass proportions.

We use nested five-fold cross-validation with identical outer folds for every m and run. Within an outer training fold, predictors and responses are standardized using training data only. The ridge penalty is selected by five-fold inner cross-validation from $10^{-4}, 10^{-3.5}, \ldots, 10^{4}$, after which predictions are made for the held-out conditions. Concatenating the five held-out folds gives exactly one out-of-sample prediction for every condition. For subclass c, we calculate the standard coefficient of determination

$$R_c^2 = 1 - \frac{\sum_{r=1}^{46} (y_{rc} - \widehat{y}_{rc})^2}{\sum_{r=1}^{46} (y_{rc} - \bar{y}_c)^2}, \qquad R^2_{\text{macro}} = \frac{1}{11} \sum_{c=1}^{11} R_c^2. \tag{40}$$

Here $r = 1, \ldots, 46$ indexes treatment conditions and $c = 1, \ldots, 11$ indexes the published cell subclasses. The quantities y_{rc} and $\widehat{y}_{rc}$ are the observed and cross-validated predicted proportions

of subclass c in condition r, and $\bar{y}_c$ is the observed mean across conditions. The macro score gives every subclass equal weight. We also report the variance-weighted coefficient of determination

$$R^2_{\mathrm{vw}} = \frac{\sum_{c=1}^{11} v_c R_c^2}{\sum_{c=1}^{11} v_c} = 1 - \frac{\sum_{c=1}^{11}\sum_{r=1}^{46}(y_{rc} - \widehat{y}_{rc})^2}{\sum_{c=1}^{11} v_c}, \qquad v_c = \sum_{r=1}^{46}(y_{rc} - \bar{y}_c)^2. \tag{41}$$

Thus, subclasses with greater observed variation across treatment conditions receive greater weight. This is the variance-weighted multi-output R^2 used in the reported analysis. Ridge regression is used only as a regularized readout of the learned representation (Hoerl & Kennard, 1970); it does not refit or alter the causal model.

At $m = 12$, the across-run mean macro R^2 is 0.262 (standard deviation 0.066). The subclass-specific results in Table 11 show that the representation is especially predictive for glutamatergic neurons, intermediate progenitors, GABAergic neurons, and oligodendrocyte progenitors. The composition of dopaminergic neurons is not predicted above the cross-validation baseline. Reporting this variation avoids interpreting the aggregate R^2 as uniformly strong across all cell types.

Table 10: Cross-validated Amin cell-composition prediction across candidate factor counts. Entries are mean ± standard deviation over five runs. The ACDC-selected factor count is bolded.

m	Macro R^2	Variance-weighted R^2
1	-0.046 ± 0.000	-0.033 ± 0.000
2	-0.017 ± 0.006	0.032 ± 0.010
3	0.051 ± 0.044	0.090 ± 0.061
4	0.114 ± 0.023	0.169 ± 0.041
5	0.206 ± 0.059	0.254 ± 0.078
6	0.216 ± 0.043	0.289 ± 0.064
7	0.139 ± 0.060	0.155 ± 0.091
8	0.243 ± 0.071	0.291 ± 0.094
9	0.240 ± 0.070	0.317 ± 0.081
10	0.203 ± 0.055	0.251 ± 0.022
11	0.254 ± 0.049	0.305 ± 0.073
12	$\mathbf{0.262 \pm 0.066}$	$\mathbf{0.321 \pm 0.084}$
13	0.251 ± 0.048	0.257 ± 0.073
14	0.215 ± 0.110	0.272 ± 0.119
15	0.246 ± 0.081	0.264 ± 0.119

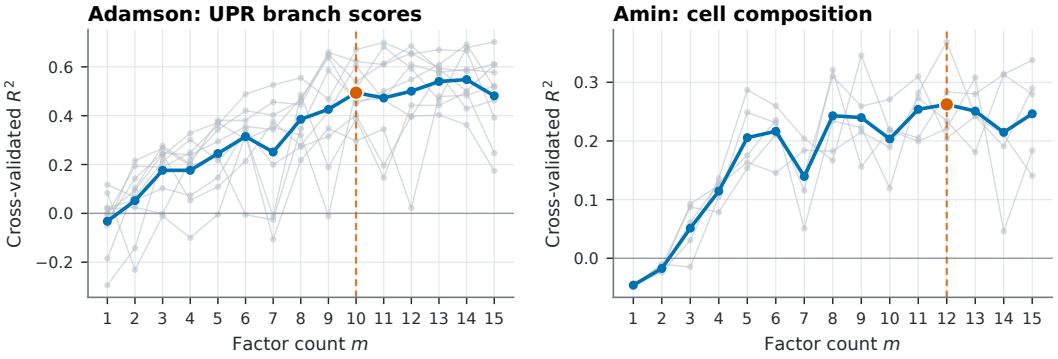


Figure 8: Cross-validated biological readouts across candidate factor counts: Adamson UPR branch-score prediction (left) and Amin cell-composition prediction (right).

Table 11: Five-fold cross-validated cell-composition prediction at the ACDC-selected $m = 12$. Entries are the mean and standard deviation of subclass R^2 across the five fitting runs.

Published subclass	R^2
Glutamatergic neuron 1	0.566 ± 0.083
Intermediate progenitor (IPC)	0.497 ± 0.053
GABAergic neuron	0.426 ± 0.121
Oligodendrocyte progenitor (OPC)	0.298 ± 0.116
Glutamatergic neuron 2	0.272 ± 0.189
Mixed neuron	0.271 ± 0.248
Oligodendrocyte	0.262 ± 0.073
Progenitor	0.161 ± 0.208
Astroglia	0.076 ± 0.153
Choroid	0.062 ± 0.050
Dopaminergic neuron	-0.006 ± 0.055

D Complete Proof of Selection Consistency

This appendix gives the complete proofs of Theorems 1 and 2. Throughout, write $\mathcal{M}_0 = \mathcal{M}_{\text{sel}}$ for the finite ordered selection set,

$$C_{\min} = \min_{m \in \mathcal{M}_0} C(m), \qquad \mathcal{A} = \{m \in \mathcal{M}_0 : C(m) = C_{\min}\}, \qquad m_0 = \min \mathcal{A}, \tag{42}$$

and define the uniform estimation error

$$E_S = \max_{m \in \mathcal{M}_0} |\widehat{C}_S(m) - C(m)|. \tag{43}$$

D.1 Formal assumptions

This subsection expands Assumptions 1– 4 in the notation used by the proof.

Assumption 1: stable fitted criterion. The empirical central criterion satisfies

$$E_S = O_{\mathbb{P}}(r_S), \qquad r_S \longrightarrow 0. \tag{44}$$

Assumption 2: valid paired uncertainty. The paired standard errors satisfy, uniformly over the finite set $\mathcal{M}_0 \times \mathcal{M}_0$,

$$\widehat{s}_S(m, \ell) = O_{\mathbb{P}}(r_S). \tag{45}$$

To state the within-optimum condition without an undefined $0/0$ ratio, let

$$\mathcal{P}_{\mathcal{A}} = \big\{\{m, \ell\} : m, \ell \in \mathcal{A},\ m \neq \ell\big\} \tag{46}$$

be the finite set of distinct unordered pairs of population minimizers. Assume there is a deterministic partition

$$\mathcal{P}_{\mathcal{A}} = \mathcal{P}_0 \cup \mathcal{P}_+ \tag{47}$$

with the following properties. For the degenerate pairs,

$$\Pr\Big\{\widehat{C}_S(m) = \widehat{C}_S(\ell) \text{ and } \widehat{s}_S(m, \ell) = 0 \text{ for every } \{m, \ell\} \in \mathcal{P}_0\Big\} \longrightarrow 1. \tag{48}$$

For the nondegenerate pairs,

$$\max_{\{m,\ell\} \in \mathcal{P}_+} \frac{|\widehat{C}_S(m) - \widehat{C}_S(\ell)|}{r_S} = O_{\mathbb{P}}(1), \tag{49}$$

$$\min_{\{m,\ell\} \in \mathcal{P}_+} \frac{\widehat{s}_S(m, \ell)}{r_S} \xrightarrow{p} \sigma_{\min} > 0. \tag{50}$$

If either part of the partition is empty, its corresponding conditions are vacuous. Because $\mathcal{A}$ is finite, pointwise versions of these conditions imply the stated uniform versions by taking a finite maximum or minimum.

Assumption 3: separated underfitting and adaptive cutoff. The separation condition is

$$\delta_- := \min_{\substack{m \in \mathcal{M}_0 \\ m < m_0}} \{C(m) - C(m_0)\} > 0, \tag{51}$$

and is vacuous if $m_0 = \min \mathcal{M}_0$. Finally, the deterministic multiplier obeys

$$\kappa_S \longrightarrow \infty, \qquad \kappa_S r_S \longrightarrow 0. \tag{52}$$

Assumption 4: interval-valued population optimum. For the contiguous selector, assume additionally that $\mathcal{A}$ is order-convex in $\mathcal{M}_0$:

$$m, \ell \in \mathcal{A}, \quad m \le u \le \ell, \quad u \in \mathcal{M}_0 \quad \Longrightarrow \quad u \in \mathcal{A}. \tag{53}$$

This condition is used only for the contiguous selector.

D.2 The empirical reference is population-optimal

Lemma 1. *Under Equation equation 44,*

$$\Pr(\widehat{b}_S \in \mathcal{A}) \longrightarrow 1. \tag{54}$$

Proof. If $\mathcal{A} = \mathcal{M}_0$, the conclusion is immediate. Otherwise, finiteness of $\mathcal{M}_0$ implies that

$$\delta_{\mathcal{A}} := \min_{m \in \mathcal{M}_0 \setminus \mathcal{A}} \{C(m) - C_{\min}\} > 0. \tag{55}$$

Consider the event $E_S < \delta_{\mathcal{A}}/3$. For every $m \notin \mathcal{A}$,

$$\widehat{C}_S(m) \ge C(m) - E_S > C_{\min} + \frac{2\delta_{\mathcal{A}}}{3}, \tag{56}$$

whereas for every $\ell \in \mathcal{A}$,

$$\widehat{C}_S(\ell) \le C_{\min} + E_S < C_{\min} + \frac{\delta_{\mathcal{A}}}{3}. \tag{57}$$

Thus no candidate outside $\mathcal{A}$ can minimize $\widehat{C}_S$, so $\widehat{b}_S \in \mathcal{A}$. Since $E_S = O_{\mathbb{P}}(r_S) = o_{\mathbb{P}}(1)$, the probability of this event tends to one. □

D.3 Insufficient factor counts are rejected

Lemma 2. *Under the stable fitted criterion, uniform paired standard-error rate, separation, and adaptive multiplier conditions,*

$$\Pr\Big\{ \widehat{\Delta}_S(m) > \kappa_S \widehat{s}_S(m, \widehat{b}_S) \text{ for every } m \in \mathcal{M}_0 \text{ with } m < m_0 \Big\} \longrightarrow 1. \tag{58}$$

Proof. By Lemma 1, $\widehat{b}_S \in \mathcal{A}$ with probability tending to one. On this event, $C(\widehat{b}_S) = C(m_0) = C_{\min}$. Therefore, for every $m < m_0$,

$$\begin{aligned} \widehat{\Delta}_S(m) &= \widehat{C}_S(m) - \widehat{C}_S(\widehat{b}_S) && (59)\\ &\ge C(m) - C(\widehat{b}_S) - 2E_S && (60)\\ &= C(m) - C(m_0) - 2E_S && (61)\\ &\ge \delta_- - 2E_S. && (62) \end{aligned}$$

Because $E_S = o_{\mathbb{P}}(1)$,

$$\Pr\left\{ \min_{\substack{m \in \mathcal{M}_0 \\ m < m_0}} \widehat{\Delta}_S(m) \ge \frac{\delta_-}{2} \right\} \longrightarrow 1. \tag{63}$$

Meanwhile, Equation equation 45, finiteness of $\mathcal{M}_0$, and Equation equation 52 imply

$$\max_{m, \ell \in \mathcal{M}_0} \kappa_S \widehat{s}_S(m, \ell) = O_{\mathbb{P}}(\kappa_S r_S) = o_{\mathbb{P}}(1). \tag{64}$$

Hence, with probability tending to one, every gap below m_0 is bounded away from zero while every corresponding adaptive threshold is smaller than that gap. This proves Equation equation 58. □

D.4 THE MINIMUM-SUFFICIENT FACTOR COUNT IS ACCEPTED

Lemma 3. *Under the stable fitted criterion, within-optimum paired-uncertainty, and adaptive multiplier conditions,*

$$\Pr\Big\{\widehat{\Delta}_S(m_0) \le \kappa_S \widehat{s}_S(m_0, \widehat{b}_S)\Big\} \longrightarrow 1. \tag{65}$$

Proof. By Lemma 1, it is enough to work on the event $\widehat{b}_S \in \mathcal{A}$. If $\widehat{b}_S = m_0$, then $\widehat{\Delta}_S(m_0) = 0$ and $\widehat{s}_S(m_0, m_0) = 0$, so the selection inequality holds.

Suppose instead that $\widehat{b}_S \in \mathcal{A} \setminus \{m_0\}$. If $\{m_0, \widehat{b}_S\} \in \mathcal{P}_0$, Equation equation 48 implies that the estimated gap and its paired standard error are both zero with probability tending to one. The selection inequality again holds.

It remains to consider $\{m_0, \widehat{b}_S\} \in \mathcal{P}_+$. Since $\widehat{b}_S$ minimizes $\widehat{C}_S$, the observed gap is nonnegative, and

$$\frac{\widehat{\Delta}_S(m_0)}{\widehat{s}_S(m_0, \widehat{b}_S)} \le \frac{\max\limits_{\{m,\ell\}\in\mathcal{P}_+} |\widehat{C}_S(m) - \widehat{C}_S(\ell)|/r_S}{\min\limits_{\{m,\ell\}\in\mathcal{P}_+} \widehat{s}_S(m,\ell)/r_S} \tag{66}$$

$$= O_{\mathbb{P}}(1), \tag{67}$$

where the denominator is bounded away from zero with probability tending to one by Equation equation 50. The bound is uniform over every possible random reference $\widehat{b}_S \in \mathcal{A}$ because $\mathcal{A}$ is finite. Since $\kappa_S \to \infty$, tightness of the ratio gives

$$\Pr\left\{ \frac{\widehat{\Delta}_S(m_0)}{\widehat{s}_S(m_0, \widehat{b}_S)} \le \kappa_S \;\middle|\; \{m_0, \widehat{b}_S\} \in \mathcal{P}_+ \right\} \longrightarrow 1. \tag{68}$$

Combining the three cases proves Equation equation 65. □

D.5 THE POPULATION-OPTIMAL BRIDGE IS ACCEPTED

Lemma 4. *Under the stable fitted criterion, within-optimum paired-uncertainty, adaptive multiplier, and order-convexity conditions,*

$$\Pr\Big\{[m_0, \widehat{b}_S]_{\mathcal{M}} \subseteq \widehat{\mathcal{A}}_S\Big\} \longrightarrow 1. \tag{69}$$

Proof. By Lemma 1, it is enough to work on the event $\widehat{b}_S \in \mathcal{A}$. Since $m_0 \le \widehat{b}_S$, order-convexity then gives

$$[m_0, \widehat{b}_S]_{\mathcal{M}} \subseteq \mathcal{A}. \tag{70}$$

The reference candidate itself is accepted because $\widehat{\Delta}_S(\widehat{b}_S) = 0$ and $\widehat{s}_S(\widehat{b}_S, \widehat{b}_S) = 0$.

Consider any distinct $u \in [m_0, \widehat{b}_S]_{\mathcal{M}}$. If $\{u, \widehat{b}_S\} \in \mathcal{P}_0$, the simultaneous event in Equation equation 48 gives $\widehat{\Delta}_S(u) = \widehat{s}_S(u, \widehat{b}_S) = 0$, so u is accepted. For pairs in $\mathcal{P}_+$, define

$$T_S = \frac{\max\limits_{\{m,\ell\}\in\mathcal{P}_+} |\widehat{C}_S(m) - \widehat{C}_S(\ell)|/r_S}{\min\limits_{\{m,\ell\}\in\mathcal{P}_+} \widehat{s}_S(m,\ell)/r_S}. \tag{71}$$

Equations equation 49 and equation 50 imply $T_S = O_{\mathbb{P}}(1)$. Moreover, because $\widehat{b}_S$ minimizes $\widehat{C}_S$,

$$0 \le \frac{\widehat{\Delta}_S(u)}{\widehat{s}_S(u, \widehat{b}_S)} \le T_S \tag{72}$$

whenever $\{u, \widehat{b}_S\} \in \mathcal{P}_+$. Since $\kappa_S \to \infty$ and T_S is tight, $\Pr(T_S \le \kappa_S) \to 1$. On this single event, every such nondegenerate pair satisfies the acceptance inequality simultaneously. Combining the reference, degenerate-pair, and nondegenerate-pair cases with $\Pr(\widehat{b}_S \in \mathcal{A}) \to 1$ proves Equation equation 69. Finiteness of $\mathcal{A}$ ensures that no additional uniformity condition is required. □

D.6 Proof of the main theorem

Proof of Theorem 1. Let A_S be the event that m_0 satisfies the selection inequality in Equation equation 13, and let B_S be the event that every $m < m_0$ fails that inequality. Lemmas 3 and 2 give

$$\Pr(A_S^c) \longrightarrow 0, \qquad \Pr(B_S^c) \longrightarrow 0. \tag{73}$$

On $A_S \cap B_S$, the candidate m_0 is accepted and every smaller candidate is rejected. Because Equation equation 13 selects the smallest accepted candidate, $\widehat{m}_S = m_0$ on this event. The union bound therefore gives

$$\Pr(\widehat{m}_S = m_0) \geq \Pr(A_S \cap B_S) \tag{74}$$

$$\geq 1 - \Pr(A_S^c) - \Pr(B_S^c) \longrightarrow 1. \tag{75}$$

All simultaneous statements used above involve only finitely many candidates or pairs. The same union-bound argument therefore ensures that intersections of their probability-tending-to-one events also have probability tending to one. □

D.7 Proof for the contiguous selector

Proof of Theorem 2. Let G_S be the event $[m_0, \widehat{b}_S]_{\mathcal{M}} \subseteq \widehat{\mathcal{A}}_S$, and let B_S be the event that every $m < m_0$ lies outside $\widehat{\mathcal{A}}_S$. Lemmas 4 and 2 imply

$$\Pr(G_S^c) \longrightarrow 0, \qquad \Pr(B_S^c) \longrightarrow 0. \tag{76}$$

On G_S, m_0 is an admissible left endpoint in Equation equation 14. On B_S, no candidate below m_0 can be an admissible left endpoint, because every interval from such a candidate to $\widehat{b}_S$ contains that rejected candidate itself. Consequently, $\widehat{m}_S^{\mathrm{con}} = m_0$ on $G_S \cap B_S$. The union bound gives

$$\Pr(\widehat{m}_S^{\mathrm{con}} = m_0) \geq 1 - \Pr(G_S^c) - \Pr(B_S^c) \longrightarrow 1. \tag{77}$$

□

D.8 Root-S specialization

Corollary 1. *Suppose all assumptions of Theorem 1 hold with $r_S = S^{-1/2}$. Then the choice*

$$\kappa_S = \sqrt{\log S} \tag{78}$$

satisfies the adaptive multiplier condition, and hence

$$\Pr(\widehat{m}_S = m_0) \longrightarrow 1. \tag{79}$$

If, in addition, $\mathcal{A}$ is order-convex, then

$$\Pr(\widehat{m}_S^{\mathrm{con}} = m_0) \longrightarrow 1. \tag{80}$$

Proof. As $S \to \infty$,

$$\sqrt{\log S} \longrightarrow \infty, \qquad \sqrt{\log S}\, S^{-1/2} = \sqrt{\frac{\log S}{S}} \longrightarrow 0. \tag{81}$$

Thus both conditions in Equation equation 52 hold. The first conclusion follows from Theorem 1; under order-convexity, the second follows from Theorem 2. □

The result concerns the minimum-sufficient target determined by C. It does not by itself prove that this target equals an independently specified data-generating factor count. That interpretation requires the additional identification condition stated in Section 2.3.