Who Guards the Guardians?
The Challenges of Evaluating Identifiability of Learned Representations

Shruti Joshi[1] **Théo Saulus**[1] **Wieland Brendel**[2] **Philippe Brouillard**[1] **Dhanya Sridhar**[1] **Patrik Reizinger**[2]

[1]Mila - Québec AI Institute & Université de Montréal
[2]Max-Planck-Institute for Intelligent Systems, ELLIS Institute Tübingen, University of Tübingen

Abstract

Identifiability in representation learning is commonly evaluated using standard metrics (e.g., MCC, R^2, DCI) on synthetic benchmarks with known ground-truth factors. These metrics are assumed to reflect recovery up to the equivalence class guaranteed by identifiability theory. We show that this assumption holds only under specific structural conditions: each metric implicitly encodes assumptions about both the data-generating process (DGP) and the encoder. When these assumptions are violated, metrics become misspecified and can produce systematic false positives and false negatives. Such failures occur both within classical identifiability regimes and in post-hoc settings where identifiability is most needed. We introduce a taxonomy separating DGP assumptions from encoder geometry, use it to characterize the validity domains of existing metrics, and release an evaluation suite for reproducible stress testing and comparison.

1 INTRODUCTION

Learning representations that are interpretable, modular, and controllable is a long-standing goal across machine learning. Identifiability formalises this objective: a representation achieves these properties when it recovers the ground-truth generative factors uniquely, up to a specified equivalence class (Comon, 1994; Hyvärinen and Pajunen, 1999). Strong identifiability guarantees now exist for nonlinear representation learners under auxiliary information (Hyvärinen et al., 2019; Khemakhem et al., 2020a), temporal structure (Hyvärinen and Morioka, 2016), mechanism sparsity (Lachapelle et al., 2022), or for restricted classes of models (Khemakhem et al., 2020c; Marconato et al., 2024). Causal representation learning (CRL) (Schölkopf et al., 2021) builds on these foundations by additionally requiring that the identified factors admit a causal semantics—typically as variables in a structural causal model with predictable responses under interventions and distribution shifts (Arjovsky et al., 2020; Peters et al., 2016). These results have wide-reaching implications and are increasingly adopted in fields such as mechanistic interpretability (Elhage et al., 2022), where identifiability of learned features is now recognised as a prerequisite for reliable interpretation (Song et al., 2025; Joshi et al., 2025), and in the analysis of pretrained representations more broadly (Roeder et al., 2021).

In practice, these theoretical guarantees are validated empirically. Given ground-truth factors $\mathbf{z} \sim p(\mathbf{z}) \in \mathbb{R}^d$ and learned representation codes $\hat{\mathbf{z}} \in \mathbb{R}^m$, a metric $\mathcal{M}(\mathbf{z}, \hat{\mathbf{z}}) \to [0, 1]$ returns a scalar interpreted as the degree of identifiability. The standard protocol is to compute $\mathcal{M}$ on a synthetic benchmark with known $\mathbf{z}$, and interpret a high score as evidence that the encoder has recovered the true factors up to a specified equivalence class, e.g., permutation and rescaling.

However, this puts all faith into the metrics—*"Who guards the guardians?"*[1] Each metric encodes structural assumptions about the latent factor distribution $p(\mathbf{z})$, the relationship between ground-truth and learned representation dimensionalities (d and m), the sample size n, and the equivalence class targeted. Yet these assumptions are typically left implicit: papers routinely report a single metric score—MCC (Khemakhem et al., 2020c), R^2, or DCI-D (Eastwood and Williams, 2018)—as evidence of identifiability, without verifying if the evaluation setting is consistent with the metric's validity domain. Prior work has observed that metrics can disagree on method rankings and are sensitive to factors such as nonlinearity strength and hyperparameter choice (Sepliarskaia et al., 2019; Carbonneau et al., 2022), and that specific metrics produce false positives when latent factors are statistically related (Yao et al., 2025). However, these remain empirical observations tied to particular settings; no prior work characterises *when* and *why* failures arise, nor whether they reflect systematic misspecification predict-

[1]"Quis custodiet ipsos custodes?"—Juvenal, *Satires* VI.

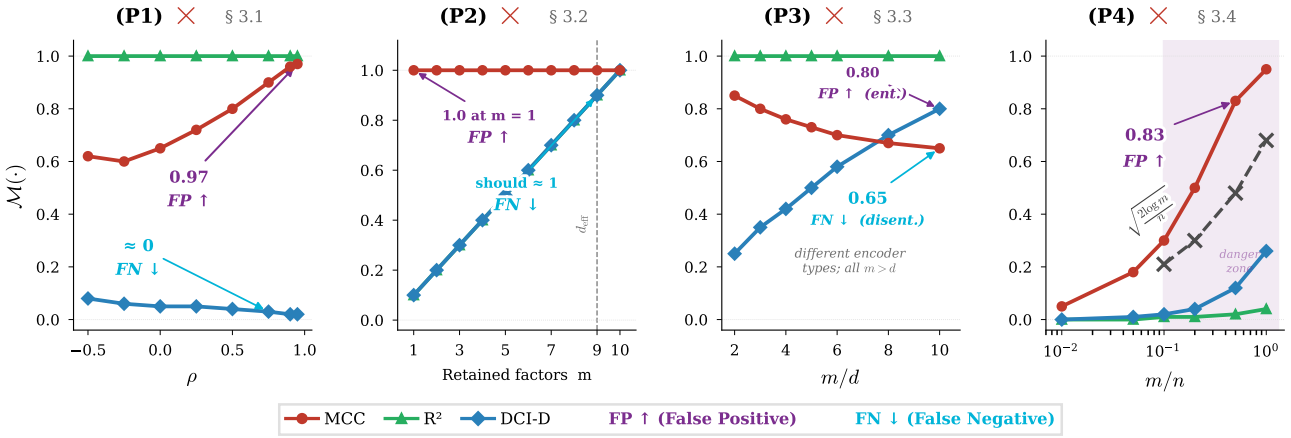

Figure 1: **Every identifiability metric fails under at least one common evaluation setting.** We test four desiderata (Properties 1 to 4) using controlled synthetic encoders that isolate metric behaviour from optimisation artefacts. **(P1)** Latent correlation: MCC conflates correlation with identifiability (FP↑ to 0.97); DCI-D penalises it (FN↓). **(P2)** Factor dropping: DCI-D reports perfect disentanglement even when 9 of 10 factors are lost. **(P3)** Overcompleteness: MCC inflates for entangled encoders; DCI-D deflates for disentangled ones. **(P4)** Null encoder: all metrics inflate as m/n grows, with MCC scaling in the order of $\sqrt{2\log m/n}$. Only R^2 is robust across (P1), (P3), and (P4), but shares the (P2) limitation. No single metric is trustworthy across all settings.

able from each metric's design. A theorem may guarantee recovery despite correlated factors or only up to an affine transform, whereas, e.g., using MCC targets axis-aligned recovery of independent factors—a strictly stronger assumption whose violation produces systematically wrong scores due to a structural mismatch between what the metric measures and what the experiment intends to measure. This leads to the question:

> *Can the structural conditions under which a metric faithfully measures identifiability be characterised, and can these conditions be used to predict when false positives and false negatives will arise?*

We show that the answer is yes: each metric's failure modes follow predictably from its encoded assumptions.

Structural misspecification. When a metric's encoded assumptions do not match the latent factor structure ($p(\mathbf{z})$) or the properties of the encoder producing $\hat{\mathbf{z}}$, we say the metric is *misspecified* for that evaluation setting. Unlike finite-sample noise, misspecification is a population-level property that would persist even when the number of samples $n \to \infty$, producing *false positives* (high scores despite lack of identifiability) or *false negatives* (low scores despite identifiability up to the desired equivalence class). To predict when and how misspecification arises, we organise assumptions along two orthogonal axes: (i) *latent factor structure*—whether ground-truth factors are independent, correlated, or linked by functional constraints that reduce the effective dimensionality below d; and (ii) *encoder properties*—the equivalence class, the dimensionality ratio m/d, and if factor information is distributed across coordinates.

Main contributions We introduce a two-axis taxonomy (§ 2) separating assumptions about latent factor structure from encoder properties, with formal desiderata for identifiability metrics (Properties 1 to 4). Through controlled synthetic experiments that isolate metric behaviour from optimisation artefacts, we show that no existing metric satisfies all desiderata and characterise precisely how each fails. We derive closed-form analyses showing that (i) MCC approaches 1 when latent factors are highly correlated, even when the encoder remains entangled (§ 3.1), and (ii) the expected MCC under an encoder producing random representations independent of the ground truth is governed by the representation-to-sample ratio (m/n) (§ 3.4). DCI-D is similarly inflated for entangled encoders when $m > d$ (§ 3.3). We also find that a fundamental limitation of all metrics is that they cannot distinguish lossless compression from lossy omission of latent factors when there exist multi-factor dependencies among them (§ 3.2). Detailed discussion of related work appears in § B.

2 A TAXONOMY FOR METRIC (MIS)SPECIFICATION

Identifiable representation learning posits a two-step data generating process.

Formal setup. Ground truth factors $\mathbf{z}$ are sampled first, and an observation $\mathbf{x} := g(\mathbf{z})$ is then generated via an unknown map $g : \mathbb{R}^d \to \mathbb{R}^n$ (Hyvärinen and Pajunen, 1999). A learned encoder $f : \mathbb{R}^n \to \mathbb{R}^m$ produces $\hat{\mathbf{z}} := f(\mathbf{x})$. It *identifies*

the generative factors up to a restricted equivalence class, typically axis-aligned transformations such as permutation and componentwise rescaling, under which the representation $\hat{\mathbf{z}}$ is identified (also called disentangled) (Schmidhuber, 1992; DiCarlo and Cox, 2007; Bengio et al., 2013; Higgins et al., 2018). We adopt the standard notion of identifiability (Hyvärinen and Pajunen, 1999).

Definition 1 (Identifiability up to $\mathcal{G}$)**.** *For $\mathcal{G}$, a class of transformations acting on $\mathbb{R}^d$ and some $h \in \mathcal{G}$, where $h : \mathbb{R}^d \to \mathbb{R}^d$, the encoder f* identifies *the latent factors up to $\mathcal{G}$ if $f \circ g = h$.*

Three standard equivalence classes are: (i) *Permutation and rescaling* ($\mathcal{G}_{\text{perm}}$): $h(\mathbf{z}) = \mathbf{PDz}$ where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}$ a diagonal scaling matrix, (ii) *Affine* ($\mathcal{G}_{\text{aff}}$): $h(\mathbf{z}) = \mathbf{Az} + \mathbf{b}$ with $\mathbf{A}$ invertible, and (iii) *Elementwise nonlinear* ($\mathcal{G}_{\text{nl}}$): $h(\mathbf{z}) = (h_1(z_{\pi(1)}), \ldots, h_d(z_{\pi(d)}))$ where each h_j is a smooth invertible function. All three assume $m = d$. When $m \neq d$ (Eastwood et al., 2023; Chen et al., 2025), Defn. 1 does not apply directly, which we extend to partial and overcomplete recovery in § 2.2.

Each metric $\mathcal{M}$ implicitly targets one of these three equivalence classes, and using a metric outside its target class produces systematically wrong scores. Based on a systematic review of the causal representation learning and nonlinear ICA literature (§ C), we study the three most commonly used metrics: MCC (in two variants: MCC-P based on Pearson correlation, and MCC-S based on Spearman rank correlation), R^2, and DCI-D (the disentanglement component). MCC computes an optimal one-to-one matching of codes to factors via pairwise correlations, hence targeting elementwise identifiability: (i) in $\mathcal{G}_{\text{perm}}$ (both MCC-P and MCC-S), (ii) and in $\mathcal{G}_{\text{nl}}$ (MCC-S). $R^2$is often used by training a linear probe from $\hat{\mathbf{z}}$ to $\mathbf{z}$ and measures explained variance, hence used for evaluating linear identifiability under $\mathcal{G}_{\text{aff}}$—it cannot distinguish between $\mathcal{G}_{\text{aff}}$ and $\mathcal{G}_{\text{perm}}$ DCI-D trains a probe (linear or nonlinear, e.g., gradient boosted trees (GBT) (Natekin and Knoll, 2013)) to predict each ground-truth factor from the learned codes, then measures how concentrated the resulting feature importances are: a score of 1 means each code is important for predicting at most one factor. Unlike MCC, DCI-D does not require one-to-one code–factor alignment and can handle $m \neq d$, but it remains sensitive to how the probe distributes importance across coefficients—correlated or entangled codes spread importance across multiple factors, deflating the score even when all information is preserved (§ 3).

To predict metric failures, we consider two orthogonal axes: the *latent factor structure*—whether factors are independent, correlated, or linked by deterministic constraints—and the *encoder geometry*—the equivalence class, dimension ratio m/d, and how factor information is distributed across codes. We define each axis in turn. We use a simple physical system as a running example throughout this section to illustrate how the DGP types and encoder geometries defined below arise naturally in practice.

RUNNING EXAMPLE: A CIRCUIT WITH A RESISTOR.

Setting. Current flows through a resistor, causing it to dissipate heat and exchange energy with its environment. Sensors (ammeter, voltmeter, thermometer, thermal camera) record the circuit state, producing observations $\mathbf{x}$.

Factors. Four physical quantities influence measurements: T (ambient temperature), R (resistance), I (current), and V (voltage). All four are independently measurable, but not independently variable. Two physical laws constrain them: $R = R_0\big(1 + \alpha\,(T - T_0)\big)$ and $V = IR$, where R_0 is the resistance at reference temperature T_0 and α is the material's temperature coefficient of resistance. So, the factor set (T, R, I, V) has four entries but two degrees of freedom.

Why this matters. An unsupervised learner has no access to Ohm's law. If it discovers a feature tracking V, that feature is correct even though V is, in principle, determined by I and R. Recovering all four factors reflects the DGP at a particular level of description; a different granularity, say, retaining only (T, I), is equally valid. Which level is appropriate depends on the downstream task and cannot be determined from the representation alone.

2.1 FACTOR DEPENDENCIES REDUCE EFFECTIVE DIMENSIONALITY

Standard disentanglement benchmarks sample each latent factor independently ($\mathbf{D}_\perp$) (Matthey et al., 2017; Burgess and Kim, 2018; Gondal et al., 2019). However, identifiability theorems do not always assume this (Lachapelle et al., 2022; Hyvärinen et al., 2019; Morioka and Hyvärinen, 2023; Khemakhem et al., 2020a,c; Ahuja et al., 2022) and permit statistical dependence ($\mathbf{D}_\rho$), e.g., through confounding or noisy causal mechanisms ($z_2 = f(z_1) + \varepsilon, \varepsilon \not\equiv 0$). In both cases, every factor retains a unique degree of freedom, so $d_{\text{eff}} = d$. We argue that a third, orthogonal generalisation is equally important: factors may be linked by *deterministic functional constraints* that reduce the effective dimensionality of the factor set below d. Such constraints arise naturally from definitional redundancies (e.g., encoding position on both linear and logarithmic scales) and physical laws (see the running example box). This is generic in unsupervised settings where the target factors are not known a priori.

Setup. Two DGP types standard in the literature that define the regime where identifiability theory operates.

- $\mathbf{D}_\perp$ **— Independent factors.** Factors vary independently; each contributes a unique degree of freedom. This is the implicit assumption behind most metrics.
  - ▹ T and I are set by independent exogenous sources.

- $\mathbf{D}_\rho$ — **Correlated factors.** Factors are statistically dependent but each retains a unique degree of freedom; no factor is a deterministic function of the others.
  - A thermostat induces a correlation between T and I.

Extended setup for unknown abstraction level. The standard settings above assume that each factor contributes independent information. In practice, however, factors may be linked by deterministic relationships that reduce the effective dimensionality below d. This can happen both when we know the ground truth latent factor set, and when they are not known a priori. In such a case, as in the circuit example, a learner with no knowledge of Ohm's law might reasonably include both resistance R and temperature T as separate factors, unaware that $R = R_0(1 + \alpha T)$. In either case, some factors carry no independent information, and metrics that treat every factor as a free degree of freedom will be misspecified.

Definition 2 (Effective dimensionality)**.** *For latent factors $\mathbf{z} \in \mathbb{R}^d$ subject to k independent smooth constraints $c_1(\mathbf{z}) = 0, \ldots, c_k(\mathbf{z}) = 0$, the* effective dimensionality *is $d_{\text{eff}} = d - k$, i.e., the number of factors that can vary freely. Under $\mathbf{D}_\perp$ and $\mathbf{D}_\rho$, $d_{\text{eff}} = d$. Under $\mathbf{D}_f$/$\mathbf{D}_F$, $d_{\text{eff}} < d$.*

- $\mathbf{D}_f$ — **Single-factor constraint.** One factor is a deterministic function of exactly one other, $d_{\text{eff}} = 1$.
  - $R = R_0(1 + \alpha T)$.
- $\mathbf{D}_F$ — **Multi-factor constraint.** A deterministic relationship involves multiple factors, reducing d_{eff} further.
  - $V = IR$: voltage is determined jointly by current and resistance, so (T, R, I, V) has $d_{\text{eff}} = 2$.

$\mathbf{D}_\perp$ and $\mathbf{D}_\rho$ address statistical relationships among independently varying factors ($d_{\text{eff}} = d$). $\mathbf{D}_f$ and $\mathbf{D}_F$ address settings where deterministic constraints reduce $d_{\text{eff}} < d$, also addressing what happens when the abstraction level is unknown. Although single- and multi-factor dependencies might not seem qualitatively different, metrics behave differently (c.f. § 3.2). Formal description in § D.1.

2.2 ENCODER STRUCTURE AND DIMENSION MISMATCH

Identifiability theory typically assumes $m = d$: the encoder's output dimension matches the number of latent factors. In practice—particularly when disentangling representations from pretrained models where d is unknown—the common regime is $m > d$, often $m \gg d$. We organise encoders along two axes: the *equivalence class* up to which factors are identified, and the *dimension ratio* m/d.

Matched dimension. The equivalence classes from Defn. 1 define three encoder types at $m = d$.

- **E1** — **Elementwise linear.** Recovery up to $\mathcal{G}_{\text{perm}}$ (Defn. 1 (i)). This is the strongest form of identifiability that can be guaranteed. Ideally, every metric should score 1 here; any that does not has an intrinsic calibration defect.
  - $(2T,\ -R,\ 0.5I,\ 3V)$.
- **E2** — **Elementwise nonlinear.** Recovery up to $\mathcal{G}_{\text{nl}}$ (Defn. 1 (iii)). Each code is a smooth invertible function of exactly one factor. The parameter α (Tab. 1) controls the degree of nonlinearity; $\alpha = 0$ reduces to **E1**.
  - $(\tanh T,\ R^3,\ \sqrt[3]{I},\ \sinh V)$.
- **E3** — **Linearly entangled.** Recovery up to $\mathcal{G}_{\text{aff}}$ (Defn. 1 (ii)). All factor information is preserved, but distributed across coordinates via rotation or shearing. The degree of entanglement is controlled by the condition number κ of $\mathbf{A}$ (Tab. 1): $\kappa = 1$ reduces to **E1**.
  - $A\,(T, R, I, V)^\top$: every code mixes all factors.

Dimension mismatch breaks coordinate-wise evaluation. In practice, m may differ from d, motivating a more general notion of identifiability.

Definition 3 (Identifiability under dimension mismatch)**.** *Let $S \subseteq \{1, \ldots, d\}$ and let $\mathcal{G}$ follow from Defn. 1. The encoder f identifies the factors S up to $\mathcal{G}$ if there exist $T \subseteq \{1, \ldots, m\}$ with $|T| = |S|$ and $h \in \mathcal{G}$ such that,*

$$\pi_T \circ f \circ g \;=\; \pi_S \circ h,$$

where $\pi_T \colon \mathbb{R}^m \to \mathbb{R}^{|T|}$ and $\pi_S \colon \mathbb{R}^d \to \mathbb{R}^{|S|}$ select the coordinates indexed by T and S resp. Since h may permute coordinates, S identifies which factors are recovered but not which codes carry them. When $m = d$ and $T = S = \{1, \ldots, d\}$, both projections become identity, and this reduces to Defn. 1.

In words: among the m learned codes in $\hat{\mathbf{z}}$, there exist $|S|$ of them (indexed by T) that together recover the factors in S up to the allowed transformation class $\mathcal{G}$; the remaining $m - |S|$ codes are ignored. The choice of h in Defn. 3 inherits from Defn. 1: for identifiability up to $\mathcal{G}_{\text{perm}}$, the composition $\pi_T \circ f \circ g$ recovers each factor up to permutation and rescaling; for $\mathcal{G}_{\text{nl}}$, up to a smooth monotonic nonlinear function; and under $\mathcal{G}_{\text{aff}}$, $\pi_T \circ f \circ g$ may return an invertible linear mix of the factors in S rather than individual factors. Next, we can define dimensionality-mismatched encoders.

- *E4* —**Undercomplete**. The encoder outputs fewer dimensions than there are ground-truth factors ($m < d$), so $|S| < d$: some factors are unrecoverable regardless of π_T. While this is lossy in the standard sense, it may be a valid lossless compression of information in the case of redundant ground truth latent factors. E.g., under $\mathbf{D}_f$ or $\mathbf{D}_F$, the ground-truth factors contain deterministic redundancies, so an encoder that recovers all d_{eff} independently varying factors already captures the full information of $\mathbf{z}$ (Defn. 2). Defn. 3 reports only *which* factors appear in S; judging whether $|S| \geq d_{\text{eff}}$ constitutes lossless recovery requires additionally knowing the constraint structure of

the DGP. No current metric makes this distinction: all treat $|S| < d$ uniformly, whether the omitted factors are redundant or independently informative.

▷ $(T,\ I)$: $|S| = 2 = d_{\text{eff}}$.

- ***E8*** — **Distributed**. A type of *overcomplete* ($m > d$) code. Ground-truth factors are recoverable only through a many-to-one map, multiple codes jointly encode a single factor, and r must aggregate across them. Coordinate-wise metrics implicitly assume each factor is encoded by a single code.

  ▷ (a_1, a_2, T, I, V) where $V = \sqrt{a_1^2 + a_2^2}$ is fully determined by (a_1, a_2), yet neither alone predicts V.

Additional overcomplete geometries—linear duplication (**E5**), nonlinear duplication (**E6**), and linear superposition (**E7**)—are constructed and evaluated in § 3.

E9 — Control baseline. $\hat{\mathbf{z}} \sim \text{Uniform}([0,1]^m)$, independent of $\mathbf{x}$. Every metric should return ≈ 0. With these definitions in hand, we can formally characterise metric failures. Formal constructions in § D.2.

3 METRICS AS MEASUREMENT INSTRUMENTS

We study the structural sensitivity of identifiability metrics through controlled synthetic experiments. In each experiment, we sample ground-truth factors $\mathbf{z} \in \mathbb{R}^d$ according to a DGP type ($\mathbf{D}_\perp$–$\mathbf{D}_F$) and construct representations $\hat{\mathbf{z}} = T(\mathbf{z})$ via a transformation matching the encoder type (**E1**–**E10**). *The representation encoder is not learned.* This design isolates metric misspecification from optimisation artefacts: every failure we observe is a property of the metric, not of training. Unless otherwise noted, we report results for $n = 1000$ samples, $d = 5$ ground-truth factors, and average over 5 seeds; confidence bands show 95% intervals. For metrics requiring a trained predictor (DCI, R^2), data is split into (80/20) training and test sets. Full experimental details and parameter definitions are in § G[2].

Metrics evaluated. We evaluate the metrics introduced in § 2, grouped into four families: *correlation-based* (MCC-P/S, MCC-RDC (Lopez-Paz et al., 2013)), regression-based (DCI-D, R^2), and (mutual information) MI-based (MIG (Chen et al., 2018), InfoMEC (Hsu et al., 2023)), and conditional independence testing based (T-MEX (Yao et al., 2025)). In main text, we focus on the commonly used metrics spanning the first two families: MCC, DCI-D, and R^2.

Sanity checks. We first ask: *do metric scores remain stable when the encoder perfectly recovers each factor (**E1**), but the DGP varies from independent to correlated to functionally redundant?* Any metric faithful to the equivalence class should return ≈ 1 across $\mathbf{D}_\perp$–$\mathbf{D}_F$ under **E1**, since the encoder–factor relationship is identical in all cases. We find that MCC-P, MCC-S, and R^2 have outputs ≈ 1, while DCI-D exhibits a systematic dip under $\mathbf{D}_f$, particularly at small d (Fig. 7; the dip diminishes as d grows from 5 to 20 but does not vanish). The dip arises because the redundant factor creates collinearity in the regression probe, inflating the importance mass assigned to the dependent factor and reducing the disentanglement score. This persists as n is increased (Fig. 11). A second test is to assess sensitivity to *encoder* nonlinearity rather than DGP structure; Fig. 9 shows that flat curve with MCC-S and DCI-D, as expected.

[2]We will release a unified implementation of all metrics with improved robustness, and our metric evaluation suite upon acceptance.

Table 1: m/d **(overcompleteness ratio),** d/n **(sample ratio), and** m/n **(representation-to-sample ratio).**

Sym.	Meaning	Range
SCALING PARAMETERS		
n	# i.i.d. paired samples	50–10k
m	Dim. of $\hat{\mathbf{z}} \in \mathbb{R}^m$	1–200
d	# ground-truth factors $\mathbf{z} \in \mathbb{R}^d$	2–20
COMPLEXITY PARAMETERS		
ρ	Pairwise correlation ($\mathbf{D}_\rho$); off-diagonal entries of Σ	$(-1, 1)$
α	Nonlinearity strength (**E2**); α=0: linear, α=1: fully nonlinear $\hat{z}_j = (1-\alpha) s_j z_{\pi(j)} + \alpha\, h_j(z_{\pi(j)})$	$[0, 1]$
κ	Condition # of mixing matrix (**E2**, **E7**); κ=1: orthogonal, κ=50: ill-cond. $A = U\text{diag}(\text{linspace}(1, \kappa^{-1}, d))V^\top$	1–50

3.1 CORRELATED AND ENTANGLED LATENT FACTORS LEAD TO BOTH FALSE POSITIVES AND FALSE NEGATIVES

We first study how latent-factor correlation ($\mathbf{D}_\rho$) interacts with metric scores under encoders **E1** (perfectly disentangled) and **E3** (linearly entangled). The encoder is held fixed; only the pairwise correlation $\rho \in (-1, 1)$ among ground-truth factors varies. Any change in the metric score is therefore a pure artifact of the latent covariance structure.

Property 1 (Invariance to latent correlation)**.** *For $\mathbf{z} \in \mathbb{R}^d$ with pairwise correlations $\text{Corr}(Z_i, Z_j) = \rho_{ij}$, fix encoder f. A metric $\mathcal{M}$ is* invariant to the latent correlation structure *if, for every encoder f, $\mathcal{M}(\hat{\mathbf{z}}, \mathbf{z})$ does not depend on $(\rho_{ij})_{i \neq j}$ and only depends on f.*

Violation of Property 1 means $\mathcal{M}$ conflates representation quality with the covariance structure of the DGP.

Setup: $\mathbf{D}_\rho$ + E1/E3. Consider d ground-truth factors with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \Sigma), \Sigma_{ii} = 1, \Sigma_{ij} = \rho$ for $i \neq j$. Note that the equicorrelation matrices are positive semidefinite only for $\rho \geq -1/(d-1)$; at $d = 10$ this gives $\rho \gtrsim -0.11$,

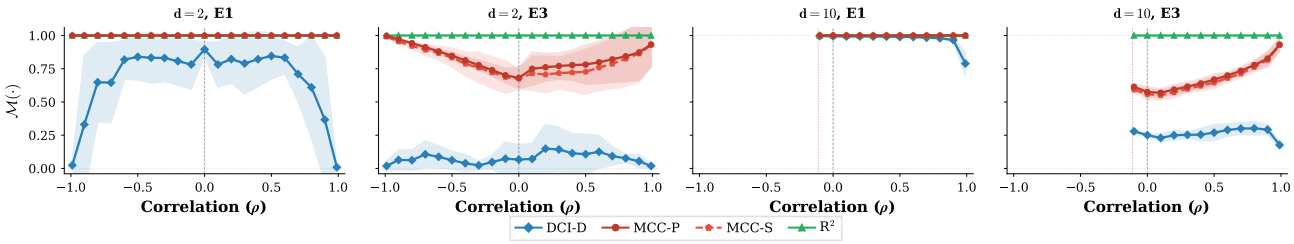

Figure 2: MCC **conflates correlation with identifiability.** Under **E3**, MCC increases with ρ and approaches the score of the perfectly disentangled encoder **E1** at high correlation, despite the encoder remaining entangled. DCI better separates **E1** from **E3** but collapses to near-zero scores making it hard to distinguish from a non-identifiable encoder. The bias sharpens with increasing d. See Fig. 14 for all metrics.

so strongly negative correlations are infeasible at moderate d. **E1** is realised as $\hat{z}_j = s_j z_j$ with $s_j > 0$. For **E3**, the encoder is a full-rank linear map $\hat{\mathbf{z}} = \mathbf{A}\mathbf{z} + \mathbf{b}$ with $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\mathrm{linspace}(1, \kappa^{-1}, d))\,\mathbf{V}^\top$, where $\mathbf{U}, \mathbf{V}$ are random orthogonal matrices and $\kappa \geq 1$ controls the condition number (degree of entanglement).

Theoretical analysis. We derive a closed-form expression for MCC-P under $\mathbf{D}_\rho$ + **E3** (§ F.1), yielding:

Proposition 1 (MCC produces false positives under correlation)**.** *Under $\mathbf{D}_\rho$ + **E3**, MCC-P depends explicitly on ρ, violating Property 1. Moreover, at both extremes $\rho \to +1$ and $\rho \to -1$, $\mathrm{MCC}(\rho) \to 1$, despite an entangled encoder.*

Prop. 1 predicts not merely a sensitivity issue w.r.t. ρ, but a failure where the metric saturates at 1 even for an entangled encoder identified only up to $\mathcal{G}_{\text{aff}}$. Whereas MCC is designed to distinguish such encoders from ones identified up to $\mathcal{G}_{\text{perm}}$, making an entangled representation indistinguishable from a disentangled one. Under correlated factors and non-axis-aligned encoders, MCC systematically overestimates identifiability, the gap between **E1** and **E3** narrows as ρ increases (Fig. 2). We observe that the gap and the bias sharpens with growing d. Fig. 13 studies the interaction between ρ and κ at $d = 10$, confirming variation of each metric's values with ρ rather than κ.

Takeaway. MCC cannot reliably compare representations learned from correlated data, scoring near 1 at high ρ (false positive) for an entangled encoder. DCI-D is overly sensitive to κ, scoring near zero for any non-trivial entanglement(false negative; Fig. 13).

3.2 METRICS CANNOT DETECT MULTI-FACTOR REDUNDANCY

We now study what happens when the encoder outputs fewer dimensions than the number of ground-truth factors ($m < d$). We construct **E4** by selecting m factors and applying elementwise rescaling, so the retained factors are *perfectly* identified. The central question is: *can metrics distinguish an encoder that drops a redundant factor (lossless) from one that drops an informative factor (lossy)?*

Property 2 (Faithfulness to effective dimensionality)**.** *Let $\mathbf{z} \in \mathbb{R}^d$ have effective dimensionality $d_{\text{eff}} \leq d$ (Defn. 2). $\mathcal{M}$ is* faithful to the effective dimensionality *if $\mathcal{M} = 1$ whenever the encoder recovers all d_{eff} independently varying factors (even if $m < d$), and $\mathcal{M} < 1$ whenever the encoder fails to recover at least one independently varying factor.*

Setup: $\mathbf{D}_\perp$/$\mathbf{D}_f$ + E4. Under $\mathbf{D}_\perp$, all d factors are independent, so every omission is lossy. Under $\mathbf{D}_f$, one factor is a deterministic function of another ($z_2 = z_1^3$, so $d_{\text{eff}} = d - 1$); dropping z_2 is lossless. Under $\mathbf{D}_F$, one factor depends on two others ($z_k = g(z_i, z_j)$, $d_{\text{eff}} = 9$); dropping z_k is again lossless. In all cases: $\hat{z}_j = s_j z_j$ for $j \in S$, $|S| = m$.

Fig. 3 reveals a split between metric families. MCC-P/S perform optimal one-to-one matching and score only matched pairs, yielding 1.0 for any $m \geq 1$ regardless of whether omitted factors are redundant or informative. R^2 and DCI-D train a probe to predict *all* d factors from the representation. Under $\mathbf{D}_\perp$ (left), unrepresented factors are unpredictable and $R^2 \approx m/d$. Under $\mathbf{D}_f$ (middle), the redundant factor $z_2 = z_1^3$ is predictable from the retained z_1, so R^2 and DCI-D stay near 1.0 at $m = d_{\text{eff}}$, thus correctly satisfying Property 2. Under $\mathbf{D}_\rho$ (correlated factors), $R^2 > m/d$ because the probe partially predicts dropped factors from correlated retained ones; we defer this to Fig. 17.

Under $\mathbf{D}_F$, the redundant factor $z_k = g(z_i, z_j)$ depends jointly on two other factors. Although $d_{\text{eff}} = 9$ (same as $\mathbf{D}_f$), the nonlinear probe fails to detect the relationship Fig. 3 shows a false negative: a lossless encoder is penalised as though it were lossy.

Takeaway. Regression-based metrics (R^2, DCI-D) detect single-factor redundancy ($\mathbf{D}_f$), but no current metric detects multi-factor redundancy ($\mathbf{D}_F$).

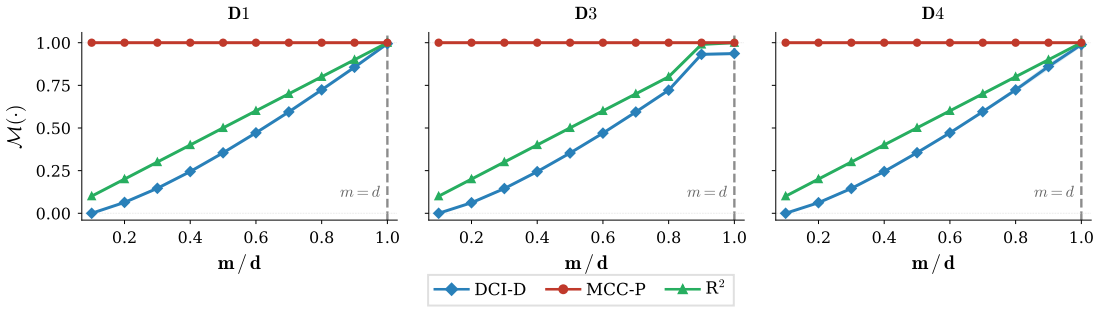

Figure 3: **Regression-based metrics detect single-factor redundancy.** *Left* ($\mathbf{D}_\perp$): every dropped factor is informative; R^2 follows m/d and DCI-D declines steadily. MCC-P/S report 1.0 even at m=1 (false positive). *Right* ($\mathbf{D}_f$): at m=9d_{eff}, the dropped factor is redundant ($z_2 = z_1^3$); R^2 and DCI-D plateau near 1.0, correctly recognising ossless compression, then decline as informative factors are removed. MCC-P/S remain at 1.0 throughout both panels and cannot distinguish the two.

3.3 METRICS CANNOT COMPARE OVERPARAMETRISED ENCODERS

When $m > d$, the encoder outputs more codes than there are factors. We first formalise the desired property we want a metric to exhibit.

Property 3 (Invariance to overcompleteness.)**.** *Let f be an encoder with $m = d$ that identifies factors up to equivalence class $\mathcal{G}$ (Defn. 3), and let f' be an overcomplete encoder ($m > d$) that identifies the same factors up to the same $\mathcal{G}$. A metric $\mathcal{M}$ is invariant to the overcomplete dimension if $|\mathcal{M}(f') - \mathcal{M}(f)| \leq \epsilon(n)$, where $\epsilon(n) \to 0$ as $n \to \infty$.*

Violation of Property 3 implies that the metric either spuriously rewards extra codes that add no per-factor information, or that it penalises an encoder that has not lost any factors but merely represents them using multiple codes. In either case, this would represent a metric conflating dimensionality with identifiability.

Setup. We compare four overcomplete geometries (E5–E8) against the matched-dimension entangled baseline **E3**, under $\mathbf{D}_\perp$. We first fix $m/d = 2$ (d=20, n=1600) and then sweep $m/d \in \{1, 1.5, 2, 3\}$ (d=5, n=1000) to test whether the results are stable as overcompleteness increases. At moderate overcompleteness (m/d=2), all metrics correctly separate entangled from disentangled encoders (Fig. 19). Fig. 4 tests whether this holds as m/d increases.

Fig. 4 shows that increasing m/d does not uniformly increase or decrease scores. Instead, it amplifies the mismatch between each metric's implicit equivalence class and the encoder's geometry. Two cases are particularly informative.

MCC **cannot be used for distributed codes (E8).** Each factor is encoded as k codes (e.g., $\sin z_j, \cos z_j$ for k=2); no single code suffices to recover the factor. MCC pairs each factor with exactly one code, so the best match (say $\sin z_j$) has correlation strictly less than 1 with z_j. As k grows, per-code information thins and MCC-P drops from ∼0.85 at m/d=2 to ∼0.65 at m/d=10, even though the factors are fully recoverable from their code subsets. This is a structural failure and it worsens monotonically with m/d. DCI-D does not exhibit this failure as the nonlinear probe can fit all m codes, selecting the k codes in each disjoint subset. Since each selected code predicts only one factor, $D_i = 1$ and DCI-D stays near 1.0 at all tested m/d.

Linear entanglement at high m/d increase DCI-D (E7). However, DCI-D increases substantially even for the linearly entangled encoder, from ∼0.42 at m/d=1.5 to ∼0.80 at m/d=10. This produces a false positive, that could mislead model comparison.

Only E5 (elementwise linear duplication) satisfies Property 3 across all metrics and all tested m/d values.

Takeaway. No metric satisfies Property 3 across all encoder types; overcomplete representations require multi-metric evaluation or matched-dimension controls.

3.4 HIGH REPRESENTATION-TO-SAMPLE RATIO INCREASES RISK OF FALSE POSITIVES

A metric should assign ≈ 0 to a random encoder that carries no information about $\mathbf{z}$. Unlike the population-level misspecification studied in § 3.1 to 3.3, the false-positive inflation in this section is a finite-sample phenomenon: the bias vanishes as $n \to \infty$. We include it because the sample regimes encountered in practice—particularly in mechanistic interpretability, where m/n routinely exceeds 1—are far from this asymptotic limit, making the finite-sample floor operationally indistinguishable from structural misspecification.

Property 4 (Insensitivity to uninformative encoders)**.** *For any encoder f independent of $\mathbf{z}$, a metric $\mathcal{M}$ should satisfy $\mathcal{M} \approx 0$ regardless of the dimensionality ratio m/d and sample size n.*

Setup. We construct a null encoder E9 ($\hat{\mathbf{z}} \sim$

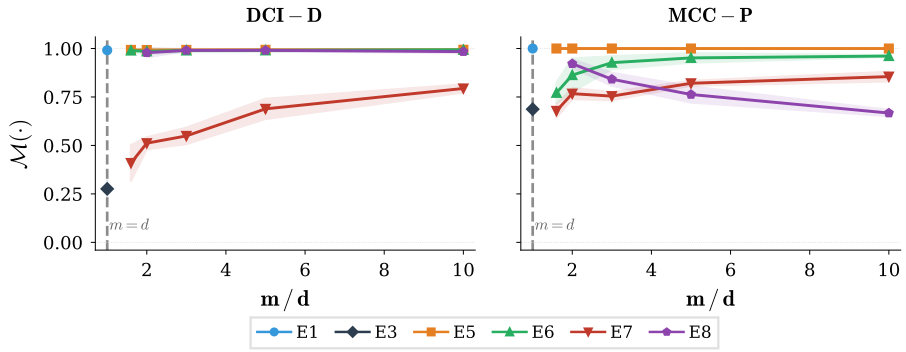

Figure 4: **Sweeping m/d reveals encoder-specific violations of Property 3.** E5 (elementwise linear duplication) is the only encoder for which all metrics remain stable. DCI-D increases for entangled E7 as m/d grows. R^2 collapses for nonlinear E6. MCC-P decreases for disjoint E8. d=5, n=1000.

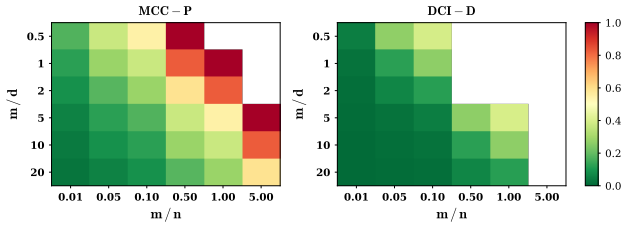

Figure 5: **Null-encoder scores reveal that m/n (columns), not m/d (rows), governs false-positive inflation.** Each cell shows the metric score of a random encoder E9 that carries no information about $\mathbf{z}$; any score above 0 is a false positive. See § F.3 for the theoretical analysis and Fig. 18 for the Gaussian null (nearly identical).

Uniform$([0,1]^m)$ and sweep over both m/d and m/n.

MCC violates Property 4 whenever $m/n \geq 0.1$. Reading along any row of Fig. 5 (fixed m/d, varying m/n), scores increase steadily; reading along any column (fixed m/n, varying m/d), scores are approximately constant. The false-positive rate is therefore governed by m/n, not m/d. At $m/n = 0.5$ and $m/d = 1$, MCC-P reports 0.83 for a representation that is pure noise. DCI-D satisfies Property 4 in the large-sample regime, but shows moderate inflation at higher estimation ratios, particularly when m/d is small. R^2 satisfies the property across the entire $(m/d, m/n)$ grid.

Theoretical analysis. We derive this behaviour in § F.3. Under the null, each entry of the $m \times d$ sample correlation matrix has mean zero and standard deviation $\approx 1/\sqrt{n}$ by the Central Limit Theorem (CLT). Hungarian matching picks the best one-to-one assignment from m candidates per column; the expected maximum of m draws from $\mathcal{N}(0, 1/n)$ scales as $\sqrt{2\log m/n}$ (Cai and Jiang, 2011), giving $\mathbb{E}[\text{MCC-P}] \gtrsim \sqrt{2\log m/n}$ (up to a constant). This depends on m and n but not on d, explaining the column-varying, but constant across rows pattern in Fig. 5.

Practical implications. The $m/n \gtrsim 0.1$ threshold is routinely exceeded: evaluating a pretrained LLM such as Llama-3.2-8B ($m = 4096$) with a few hundred samples gives $m/n \in [0.5, 10]$; even standard disentanglement benchmarks with $m = 64$ and $n = 500$ labelled samples yield $m/n > 0.1$. DCI-D requires more samples and exhibits the same inflation. R^2 is the most robust to false positives, but requires $n \gtrsim 500$ under nonlinear encoders (Fig. 11).

Takeaway. MCC is unreliable whenever $m/n \gtrsim 0.1$. Always verify $m \ll n$ and report null-encoder baselines alongside metric scores.

4 CONCLUSION

All existing identifiability metrics can be deceptive (Fig. 1). We provide a taxonomy (§ 2) and theoretical and empirical analyses to characterise these failure modes, then propose four properties (Properties 1 to 4) for future metric design. We distil our findings into a practitioner checklist (§ A) and a metric selection lookup table (Tab. 3). Our results have direct consequences for any pipeline that uses identifiabilility metrics to make downstream predictions.

Limitations. Our analysis uses synthetic encoders by design, to isolate metric misspecification from optimisation artefacts. The taxonomy does not cover stochastic encoders or discrete factors, all of which arise in practice. Lastly, a systematic study of how metric failures manifest across different families of learned encoders (rather than constructed ones) would be a complementary direction.

References

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022. Cited on page 3.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893. Cited on page 1.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. Cited on page 3.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011. Cited on page 24.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018. Cited on page 3.

T Tony Cai and Tiefeng Jiang. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39(3):1496–1525, 2011. Cited on pages 8 and 24.

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE transactions on neural networks and learning systems*, 35(7):8747–8761, 2022. Cited on pages 1 and 12.

Guangyi Chen, Yunlong Deng, Peiyuan Zhu, Yan Li, Yifan Shen, Zijian Li, and Kun Zhang. Causalverse: Benchmarking causal representation learning with configurable high-fidelity simulations. *arXiv preprint arXiv:2510.14049*, 2025. Cited on page 3.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. Cited on pages 5 and 12.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. Cited on pages 1 and 12.

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8): 333–341, 2007. Cited on page 3.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations*, 2018. Cited on pages 1, 12, and 17.

Cian Eastwood, Andrei Liviu Nicolicioiu, Julius von Kügelgen, Armin Kekić, Frederik Träuble, Andrea Dittadi, and Bernhard Schölkopf. Dci-es: An extended disentanglement framework with connections to identifiability, 2023. URL https://arxiv.org/abs/2210.00364. Cited on pages 3 and 13.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. Cited on pages 1 and 12.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922. Cited on page 23.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019. Cited on page 3.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. Cited on page 3.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pages 308–334. Springer, 1992. Cited on page 23.

Harold Hotelling and Margaret Richards Pabst. Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics*, 7(1): 29–43, 1936. Cited on page 24.

Kyle Hsu, William Dorrell, James Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization. *Advances in Neural Information Processing Systems*, 36: 45463–45488, 2023. Cited on pages 5, 12, and 17.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, 2016. Cited on pages 1 and 12.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. Cited on pages 1, 2, 3, and 12.

Aapo Hyvärinen, Hiroshi Sasaki, and Richard E. Turner. Nonlinear ica using auxiliary variables and generalized

contrastive learning. *Journal of Machine Learning Research*, 2019. Cited on pages 1, 3, and 12.

Shruti Joshi, Andrea Dittadi, Sébastien Lachapelle, and Dhanya Sridhar. Identifiable steering via sparse autoencoding of multi-concept shifts, 2025. URL https://arxiv.org/abs/2502.12179. Cited on pages 1 and 12.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2020a. Cited on pages 1, 3, and 12.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33: 12768–12778, 2020b. Cited on page 12.

Ilyes Khemakhem, Ricardo Pio Monti, Diederik P. Kingma, and Aapo Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica, 2020c. URL https://arxiv.org/abs/2002.11537. Cited on pages 1 and 3.

Sébastien Lachapelle, Pau Rodriguez, 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*, pages 428–484. PMLR, 2022. Cited on pages 1, 3, and 12.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019. Cited on page 13.

David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. *Advances in neural information processing systems*, 26, 2013. Cited on pages 5 and 17.

Emanuele Marconato, Sébastien Lachapelle, Sebastian Weichwald, and Luigi Gresele. All or none: Identifiable linear properties of next-token predictors in language modeling. *arXiv preprint arXiv:2410.23501*, 2024. Cited on pages 1 and 12.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. Cited on page 3.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. *arXiv preprint arXiv:2310.15709*, 2023. Cited on page 3.

Aaron Mueller, Andrew Lee, Shruti Joshi, Ekdeep Singh Lubana, Dhanya Sridhar, and Patrik Reizinger. From isolation to entanglement: When do interpretability methods identify and disentangle known concepts? *arXiv preprint arXiv:2512.15134*, 2025. Cited on page 12.

Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:63623, 2013. Cited on page 3.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):947–1012, 2016. Cited on page 1.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021. Cited on page 1.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural computation*, 4(6):863–879, 1992. Cited on page 3.

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. Cited on pages 1 and 12.

Anna Sepliarskaia, Julia Kiseleva, and Maarten de Rijke. How to not measure disentanglement. *arXiv preprint arXiv:1910.05587*, 2019. Cited on pages 1 and 12.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T Diab, Virginia Smith, and Kun Zhang. Position: Mechanistic interpretability should prioritize feature consistency in saes. *arXiv preprint arXiv:2505.20254*, 2025. Cited on pages 1 and 12.

Dingling Yao, Shimeng Huang, Riccardo Cadei, Kun Zhang, and Francesco Locatello. The third pillar of causal analysis? a measurement perspective on causal representations. *arXiv preprint arXiv:2505.17708*, 2025. Cited on pages 1, 5, 12, and 17.

CONTENTS

A PRACTITIONER CHECKLIST

A metric score is interpretable only if two conditions hold: (1) the (DGP, encoder) pair lies in a structurally valid region for that metric, and (2) the sample size n is large enough relative to the relevant dimension to ensure estimation stability. Before reporting scores, verify the following conditions.

Before evaluation.

1. **Check the overparametrisation ratio m/n.** If $m/n > 0.1$, MCC scores are unreliable: the expected score under a null encoder exceeds $\sqrt{2 \log m/n}$ (§ 3.4). Increase n or reduce m before interpreting results.
2. **Report a null-encoder baseline.** Compute every metric on a random or constant encoder with the same (m, n, d). Without this baseline, false positives are indistinguishable from genuine identifiability (§ 3.4).
3. **Know your DGP assumptions.** Determine whether latent factors are independent ($\mathbf{D}_\perp$) or correlated ($\mathbf{D}_\rho$), and whether the representation is matched ($m = d$), overcomplete ($m > d$), or undercomplete ($m < d$).

Choosing a metric.

4. **Matched dimension, independent factors** ($m = d$, $\mathbf{D}_\perp$): all three metrics (MCC, DCI-D, R^2) are reliable.
5. **Correlated factors** ($\mathbf{D}_\rho$): prefer R^2. MCC conflates correlation with identifiability (Prop. 1); DCI-D collapses under moderate entanglement (§ 3.1).
6. **Overcomplete representations** ($m > d$): no single metric is reliable across all encoder geometries. Use multiple metrics and compare against matched-dimension controls (§ 3.3).
7. Consult Tab. 3 for a full lookup table.

Interpreting scores.

8. **A high** MCC **does not imply identifiability** when m/n is large or factors are correlated.
9. **A high** DCI**-D does not imply disentanglement** when the encoder is overcomplete and linearly entangled.
10. **No pairwise metric detects multi-factor redundancy** ($\mathbf{D}_F$); higher-order statistics are needed (§ 3.2).

B RELATED WORK

Identifiability theory. Nonlinear ICA (Comon, 1994; Hyvärinen and Pajunen, 1999)establishes sufficient conditions under which latent factors can be recovered up to well-defined equivalence classes. Identifiability guarantees often leverage auxiliary variables (Hyvärinen et al., 2019; Khemakhem et al., 2020a), temporal structure (Hyvärinen and Morioka, 2016), mechanism sparsity (Lachapelle et al., 2022), and restricted model classes (Khemakhem et al., 2020b; Marconato et al., 2024). Causal representation learning extends these results by additionally requiring that identified factors admit causal semantics with predictable behaviour under interventions (Schölkopf et al., 2021). These works establish when identifiability holds in theory. We study whether the metrics used to verify these guarantees empirically are faithful to the equivalence classes the theorems provide. Our results indicate that even with correlations between latent factors, reliability on metrics drops § 3.1.

Identifiability and disentanglement metrics. A substantial body of work has proposed metrics for evaluating learned representations against ground-truth factors, including DCI (Eastwood and Williams, 2018), MIG (Chen et al., 2018), MCC (Khemakhem et al., 2020b), InfoMEC (Hsu et al., 2023), and T-MEX (Yao et al., 2025). Sepliarskaia et al. (2019) showed that several metrics disagree on comparing methods and cautioned against relying on a single score. Carbonneau et al. (2022) surveyed metrics and noted the lack of a unified framework connecting metric assumptions to evaluation validity. Our work differs from both. Instead of comparing metric rankings aross methods, we identify the structural conditions on the DGP and encoder geometry under which each metric's score is interpretable, and show that the resulting failure modes are misspecification, not optimisation failures.

Overcomplete representations and mechanistic interpretability. Recent work in mechanistic interpretability uses sparse autoencoders to extract interpretable features from pretrained models (Elhage et al., 2022), and identifiability of these features is increasingly recognised as necessary for reliable interpretation (Song et al., 2025; Joshi et al., 2025; Mueller et al., 2025). These settings are inherently overcomplete $m >> d$ and sample-constrained $m >> n$. We show that current metrics are not reliable under overcompleteness: MCC does not work for overcomplete distributed codes, DCI-D may spuriousy reward a

linearly entangled representation (§ 3.3), and the high m/n ratios typical of these evaluations may push the metrics into the regime where they can score high even with a random representation (§ 3.4).

Relationship to prior evaluation studies. Locatello et al. (2019) demonstrated that unsupervised disentanglement learning requires inductive biases, studying how *learning algorithms* behave under different model and data assumptions. Our work is complementary: we study how *evaluation metrics* behave under different structural regimes, holding the encoder fixed. Their finding that unsupervised disentanglement is impossible without inductive biases is orthogonal to our finding that even supervised metrics are structurally misspecified under conditions the underlying identifiability theorems explicitly permit. Eastwood et al. (2023) extended DCI to handle dimension mismatch; our Defn. 3 generalises this to arbitrary equivalence classes and connects it to the full DGP taxonomy, revealing failure modes beyond what dimension-mismatch alone predicts.

C METRIC USAGE REVIEW

We conducted a systematic review of evaluation metrics used in causal representation learning (CRL) and nonlinear independent component analysis (ICA). Using the Semantic Scholar API, we retrieved papers published between 2020 and 2025 at major ML conferences (NeurIPS, ICLR, ICML, AISTATS, UAI, AAAI, CLeaR, JMLR) based on the terms 'causal representation learning' and 'nonlinear ICA'. **Among the 62 papers identified, most relied on** MCC **(25), followed by** R^2 **(9) and** DCI **(2).** None employed more recent metrics such as MIG or T-MEX. Finally, several papers did not use standard metrics at all, instead reporting performance in terms of objective optimization or relying on qualitative assessments.

Also, nonlinear ICA papers use MCC (61%) **more often than CRL** (29%)**.**

D TAXONOMY

D.1 DATA CLASSES

Let $\mathbf{z} = (Z_1, \dots, Z_d)^\top \in \mathbb{R}^d$ denote the ground-truth latent factors with joint density $p(\mathbf{z})$. We classify the factor distribution along two axes: *statistical dependence* (mutual information) and *functional dependence* (deterministic constraints). Classes $\mathbf{D}_\perp$–$\mathbf{D}_\rho$ operate within the standard CRL setting ($d_{\text{eff}} = d$); Classes $\mathbf{D}_f$–$\mathbf{D}_F$ extend it to settings where functional constraints reduce the effective dimensionality ($d_{\text{eff}} < d$; cf. Defn. 2).

$\mathbf{D}_\perp$ — Independent factors. The factors are mutually independent and non-redundant:

$$p(z_1, \dots, z_d) = \prod_{j=1}^{d} p(z_j), \qquad I(Z_i; Z_j) = 0 \quad \forall\, i \neq j.$$

No statistical, functional, or structural dependence exists among factors. In particular, $\text{Corr}(Z_i, Z_j) = 0$ for all $i \neq j$, and $d_{\text{eff}} = d$.

Canonical example: $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$.

$\mathbf{D}_\rho$ — Correlated (statistically dependent) factors. The factors share information but each retains a unique degree of freedom; no factor is a deterministic function of any subset of the others:

$$\exists\, i \neq j :\ I(Z_i; Z_j) > 0, \qquad H(Z_j \mid Z_{\setminus j}) > 0 \quad \forall\, j,$$

where $Z_{\setminus j} := (Z_1, \dots, Z_{j-1}, Z_{j+1}, \dots, Z_d)$. The first condition asserts statistical dependence; the second asserts non-redundancy: no factor is determined by the rest. Hence $d_{\text{eff}} = d$.

Canonical example: $(Z_1, Z_2, Z_3)^\top \sim \mathcal{N}(0, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$ with $\rho \in (0, 1)$.

Remark. $\mathbf{D}_\rho$ subsumes both causal dependence ($Z_j = f(Z_i) + \varepsilon$, $\varepsilon \not\equiv 0$) and confounded dependence (shared latent common cause), as well as nonlinear dependence invisible to linear measures. For instance, $Z_1 \sim \mathcal{N}(0, 1)$, $Z_2 = Z_1^2 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ satisfies $\text{Corr}(Z_1, Z_2) = 0$ yet $I(Z_1; Z_2) > 0$: the dependence is real but purely nonlinear. As long as $H(\varepsilon) > 0$, the relationship is non-deterministic and falls under $\mathbf{D}_\rho$. **In this paper, we'll focus on linear non-deterministic dependence only.**

$\mathbf{D}_f$ — Single-factor functional constraint. At least one factor is a deterministic function of exactly one other factor, reducing the effective dimensionality:

$$\exists\, i \neq j,\ \exists\, f : \mathbb{R} \to \mathbb{R} : \quad Z_j = f(Z_i) \quad \text{a.s.}, \qquad H(Z_j \mid Z_i) = 0.$$

Two structurally distinct subcases arise:

$\mathbf{D}_f$A — Invertible (information-preserving). The map f is injective, so f^{-1} exists. Then $H(Z_j) = H(Z_i)$ and $I(Z_i;\, Z_j) = H(Z_i)$. The intrinsic dimension of (Z_i, Z_j) is 1, but no information is lost. *Canonical example:* $Z_2 = Z_1^3$.

$\mathbf{D}_f$B — Non-invertible (collapsed). The map f is many-to-one, so f^{-1} does not exist. Then $H(Z_j) < H(Z_i)$ and $I(Z_i;\, Z_j) = H(Z_j) < H(Z_i)$: information is destroyed. *Canonical example:* $Z_2 = \text{sign}(Z_1)$.

In both subcases, $d_{\text{eff}} \leq d - 1$ (one constraint removes one degree of freedom).

- $R = R_0(1 + \alpha\, T)$: resistance is an invertible function of temperature ($\mathbf{D}_f$A).

In this paper, $\mathbf{D}_f$A will be of interest to us.

$\mathbf{D}_F$ — Multi-factor functional constraint (synergistic). At least one factor is a deterministic function of two or more other factors, but not of any single one:

$$\exists\, k,\ \exists\, S \subset \{1, \ldots, d\} \text{ with } |S| \geq 2 : \quad Z_k = g(Z_S) \quad \text{a.s.},$$

where $Z_S := (Z_j)_{j \in S}$, and no function of a strict subset of Z_S determines Z_k. Formally:

$$H(Z_k \mid Z_S) = 0, \qquad H(Z_k \mid Z_T) > 0 \quad \forall\, T \subsetneq S.$$

The constraint cannot be decomposed into single-variable contributions: dependence is deterministic but *synergistic*. All pairwise linear correlations may vanish ($\text{Corr}(Z_j, Z_k) = 0$ for each $j \in S$) even though (Z_S) jointly determines Z_k.

Canonical example: $Z_1 = Z_2\, Z_3$, $Z_2 \perp Z_3$.

- $V = IR$: voltage is jointly determined by current and resistance, so (T, R, I, V) has $d_{\text{eff}} = 2$.

Summary. Under $\mathbf{D}_\perp$–$\mathbf{D}_\rho$, $d_{\text{eff}} = d$ (no functional constraints). Under $\mathbf{D}_f$–$\mathbf{D}_F$, $d_{\text{eff}} < d$ (deterministic constraints reduce the number of free degrees of freedom; cf. Defn. 2). The horizontal axis of the validity-domain map (Tab. 3) captures this progression.

D.2 ENCODER TAXONOMY

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ denote the learned encoder, producing $\hat{\mathbf{z}} := f(\mathbf{x}) = f(g(\mathbf{z})) \in \mathbb{R}^m$. We classify encoders by (i) the equivalence class $\mathcal{G}$ up to which factors are identified (Defn. 1), and (ii) the dimension ratio m/d. Throughout, S_d denotes the symmetric group on $\{1, \ldots, d\}$.

Matched dimension ($m = d$).

E1 — Elementwise linear (permutation & rescaling). The encoder identifies each factor up to $\mathcal{G}_{\text{perm}}$:

$$\exists\, \pi \in S_d,\ \exists\, a_j \neq 0 : \quad \hat{Z}_j = a_j\, Z_{\pi(j)}, \qquad j = 1, \ldots, d.$$

No cross-factor mixing or nonlinear reparameterisation is present beyond scaling and permutation. This is the strongest form of identifiability and every metric should score 1.

Canonical example: $\hat{Z}_1 = 2Z_3$, $\hat{Z}_2 = -Z_1$, $\hat{Z}_3 = 0.5\, Z_2$ (for $d = 3$).

E2 — Elementwise nonlinear (invertible componentwise). The encoder identifies each factor up to $\mathcal{G}_{\text{nl}}$:

$$\exists\, \pi \in S_d : \quad \hat{Z}_j = g_j(Z_{\pi(j)}), \qquad j = 1, \ldots, d,$$

where each $g_j : \mathbb{R} \to \mathbb{R}$ is a smooth, invertible scalar function. Information is preserved factor-wise, but linear correlation between $\hat{Z}_j$ and $Z_{\pi(j)}$ may be misleading. The parameter α (Tab. 1) controls the degree of nonlinearity; $\alpha = 0$ reduces to **E1**.

Canonical example: $\hat{Z}_j = Z_{\pi(j)}^3$.

E3 — Linearly entangled. The encoder identifies factors up to $\mathcal{G}_{\text{aff}}$:

$$\hat{\mathbf{z}} = \mathbf{A}\,\mathbf{z}, \qquad \mathbf{A} \in \mathbb{R}^{d\times d},\ \det(\mathbf{A}) \neq 0,$$

with $\mathbf{A}$ not a signed permutation matrix (i.e., at least one row has two or more nonzero entries). All factor information is preserved globally, but individual factors are distributed across coordinates. The condition number $\kappa(\mathbf{A})$ controls the degree of entanglement; $\kappa = 1$ reduces to **E1**.

Canonical example: $\hat{Z}_2 = a\,Z_2 + b\,Z_3$ with $ab \neq 0$.

Dimension mismatch ($m \neq d$).

The standard definition of identifiability (Defn. 1) assumes $m = d$. When $m \neq d$, we use the generalised notion of Defn. 3: the encoder identifies a subset $S \subseteq \{1, \ldots, d\}$ of factors via a readout $r : \mathbb{R}^m \to \mathbb{R}^{|S|}$.

E4 — Undercomplete ($m < d$). The encoder outputs fewer dimensions than there are ground-truth factors, so $|S| < d$: some factors are unrecoverable regardless of the readout r. Each retained factor is encoded elementwise:

$$\hat{Z}_j = a_j\,Z_{i(j)}, \qquad j = 1, \ldots, m,$$

with all $i(j)$ distinct and $a_j \neq 0$.

Under $\mathbf{D}_f$–$\mathbf{D}_F$, this need not be lossy in the information-theoretic sense: if the encoder recovers all d_{eff} independently varying factors, it captures the full information of $\mathbf{z}$ (Defn. 2). No current metric distinguishes omission of a redundant factor from omission of an informative one.

▸ $(T,\ I)$: $|S| = 2 = d_{\text{eff}}$.

Overcomplete encoders ($m > d$).

We now define four overcomplete encoder types ($m > d$), each corresponding to a distinct code–factor geometry.

E5 — Overcomplete elementwise linear. Each output coordinate is a scaled copy of exactly one ground-truth factor. Let $\sigma : \{1, \ldots, m\} \to \{1, \ldots, d\}$ be a surjective assignment (every factor is represented at least once; some are duplicated), and let $a_j \neq 0$. Then:

$$\hat{Z}_j = a_j\,Z_{\sigma(j)}, \qquad j = 1, \ldots, m, \qquad m > d.$$

The surjectivity of σ ensures no factor is lost; factors assigned to multiple indices appear as independently scaled copies. The readout r must aggregate (many-to-one) across codes that share a source factor.

Example (d=2, m=4): $\hat{Z}_1 = 1.3\,Z_1,\ \hat{Z}_2 = -0.7\,Z_2,\ \hat{Z}_3 = 0.9\,Z_1,\ \hat{Z}_4 = -1.5\,Z_2$.

E6 — Overcomplete, multiple codes per factor. The first d output coordinates are elementwise nonlinear transforms of individual factors (one per factor); the remaining $m - d$ coordinates are nonlinear functions that may depend on multiple factors simultaneously:

$$\hat{Z}_j = \begin{cases} g_j(Z_{\pi(j)}), & j = 1, \ldots, d, \\ \phi_j(Z_1, \ldots, Z_d), & j = d{+}1, \ldots, m, \end{cases}$$

where each $g_j : \mathbb{R} \to \mathbb{R}$ is an invertible scalar function, $\pi \in S_d$ is a permutation, and each $\phi_j : \mathbb{R}^d \to \mathbb{R}$ is a (possibly non-invertible) nonlinear map. The first d coordinates preserve factor-wise information up to $\mathcal{G}_{\text{nl}}$; the additional $m - d$ coordinates introduce cross-factor codes that carry redundant or mixed information. Recovery requires a many-to-one readout that can select or aggregate across both single-factor and multi-factor codes.

Example (d=2, m=3): $\hat{Z}_1 = \tanh(Z_1),\ \hat{Z}_2 = Z_2^3,\ \hat{Z}_3 = Z_1 \cdot Z_2$.

E7 — Overcomplete, linearly entangled. The encoder is a dense linear map with $m > d$:

$$\hat{\mathbf{z}} = \mathbf{A}\,\mathbf{z}, \qquad \mathbf{A} \in \mathbb{R}^{m\times d},\ \operatorname{rank}(\mathbf{A}) = d,$$

where at least one row of $\mathbf{A}$ has two or more nonzero entries, so each coordinate of $\hat{\mathbf{z}}$ mixes several factors (one-to-many). The matrix $\mathbf{A}$ is constructed via its singular value decomposition $\mathbf{A} = \mathbf{U}\,\operatorname{diag}(s_1, \ldots, s_d)\,\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{m\times m}$, $\mathbf{V} \in \mathbb{R}^{d\times d}$

orthogonal, and $s_1 \geq \cdots \geq s_d > 0$. The condition number $\kappa(\mathbf{A}) = s_1/s_d$ controls the degree of entanglement. Since $\text{rank}(\mathbf{A}) = d$, the factor information is globally preserved; recovery requires r to unmix the linear superposition.

Example (d=2, m=4): every $\hat{Z}_j$ is a distinct linear combination of Z_1 and Z_2.

E8 — Overcomplete, nonlinear disjoint subsets. Let $k \geq 2$ be an integer and set $m = k \cdot d$. There exist pairwise-disjoint index sets $S_1, \ldots, S_d \subset \{1, \ldots, m\}$ with $|S_i| = k$, $\bigsqcup_{i=1}^{d} S_i = \{1, \ldots, m\}$, such that each ground-truth factor Z_i is encoded *only* in the coordinates indexed by S_i (no cross-factor mixing):

$$\forall\, j \in S_i : \quad \hat{Z}_j = h_j(Z_{\pi(i)}),$$

where $\pi \in S_d$ is a permutation and each $h_j : \mathbb{R} \to \mathbb{R}$ is a scalar nonlinear function (not necessarily invertible individually). The factor is recoverable from its own subset via a decoder $f_i : \mathbb{R}^k \to \mathbb{R}$:

$$Z_{\pi(i)} = f_i\big(\hat{Z}_j : j \in S_i\big).$$

For $k = 2$, the canonical implementation uses $\hat{Z}_{2i} = \sin(Z_{\pi(i)})$, $\hat{Z}_{2i+1} = \cos(Z_{\pi(i)})$, with perfect reconstruction via $Z_{\pi(i)} = \text{atan2}(\hat{Z}_{2i}, \hat{Z}_{2i+1})$. For $k > 2$, an interval-based encoding partitions the range of each factor into k bins; exactly one code per factor is active for each sample. Coordinate-wise metrics fail because the readout r must aggregate (many-to-one) across the k codes in each S_i; no single $\hat{Z}_j$ suffices to recover Z_i.

Example (d=2, k=2, m=4): $(\sin Z_1,\ \cos Z_1,\ \sin Z_2,\ \cos Z_2)$.

Control baselines.

E9 — Random (independent of data). $\hat{\mathbf{z}} \sim \text{Uniform}([0,1]^m)$, independent of $\mathbf{x}$. Every metric should return ≈ 0; any nonzero score is a false positive.

Summary of code–factor geometry.

- *One-to-one*: r selects one code per factor; each code represents exactly one factor. Applies to **E1**, **E2**.
- *Many-to-one*: multiple codes carry information about the same factor; r must aggregate. Applies to **E5**, **E6**, **E8**.
- *One-to-many*: each code entangles multiple factors; r must unmix. Applies to **E3**, **E7**.

EXAMPLES OF ENCODERS.

E1 $(2T, -R, 3V, 0.5I)$
E3 $\mathbf{A}\,(T, R, I, V)^\top + \mathbf{b}$
E4 $(T,\ I)$
E8 $(a_1, a_2, T, I, V),\ R = \sqrt{a_1^2 + a_2^2}$
E9 $\hat{\mathbf{z}} \sim \mathcal{N}(0, I_5)$

E1 and **E2** preserve factor-wise information up to invertible reparameterisations; **E3** preserves global information but mixes factors; **E4** loses information; **E5** duplicates information via elementwise linear copies; **E6** combines elementwise nonlinear codes with cross-factor codes; **E7** mixes all factors linearly in an overcomplete space; **E8** distributes each factor across disjoint nonlinear codes.

E METRICS

See § F for a more detailed description of each metric.

F EXPECTED METRICS' BEHAVIOUR: THEORY AND DERIVATIONS

F.1 MCC: CORRELATED LATENT FACTORS AND LINEAR ENTANGLEMENT

We consider three ground-truth latent variables

$$\mathbf{z} = (Z_1, Z_2, Z_3)^\top,$$

Metric	Short description
DCI	Measures *Disentanglement* (D), *Completeness* (C), and *Informativeness* (I) by assessing how well latent dimensions predict ground-truth factors using supervised regressors (Eastwood and Williams, 2018).
MCC	Evaluates alignment between learned and ground-truth latent variables via an optimal one-to-one matching that maximizes pairwise correlations (P=pearson, S=spearman, RDC=Randomized Dependence Coefficient (Lopez-Paz et al., 2013)).
R^2	Quantifies the proportion of variance in ground-truth factors explained by the learned representation through linear regression.
T-MEX	Assesses disentanglement by measuring how selectively latent variables respond to interventions on ground-truth factors (Yao et al., 2025).
MIG	Computes the gap between the top two mutual information scores between a factor and latent variables [CITE].
InfoMEC	Measures equivalence classes of representations by evaluating how much information about the ground-truth factors is preserved under invertible transformations (Hsu et al., 2023) (M=modularity, E=explicitness, C=compactness).

Table 2: Common evaluation metrics for causal representation learning with ground-truth factors.

Table 3: **No single metric satisfies all four properties.** Each cell shows whether a metric satisfies (✓), partially satisfies (∼), or violates (✗) the corresponding property. Superscripts reference the relevant subsection. See § G for MI-based metrics.

Metric	P1	P2	P3	P4
MCC-P	✗	✗	✗	✗
MCC-S	✗	✗	✗	✗
R^2	✓	∼	✗	✓
DCI-D	∼	∼	✗	∼
MIG	✗	✗	✗	∼
T-MEX	∼	✗	✗	∼

P1: ρ-invariance[3.1] P2: d_{eff}-sensitivity[3.2]
P3: OC-invariance[3.3] P4: Uninformative-sensitivity[3.4]

with the following second-order structure:

$$\mathbb{E}[Z_i] = 0 \quad \text{for all } i, \tag{1}$$
$$\sigma^2(Z_1) = 1, \tag{2}$$
$$\sigma^2(Z_2) = \sigma^2(Z_3) = 1, \tag{3}$$
$$\text{Corr}(Z_2, Z_3) = \rho, \quad |\rho| < 1, \tag{4}$$

and Z_1 uncorrelated with (Z_2, Z_3).

Consider a learned representation with linear mixing of the form:

$$\hat{Z}_1 = sZ_1, \tag{5}$$
$$\hat{Z}_2 = aZ_2 + bZ_3, \tag{6}$$
$$\hat{Z}_3 = cZ_2 + dZ_3, \tag{7}$$

with $s, a, b, c, d \neq 0$. Our goal is to compute the Mean Correlation Coefficient (MCC) between $\mathbf{Z}$ and $\hat{\mathbf{Z}} := (\hat{Z}_1, \hat{Z}_2, \hat{Z}_3)$ as a function of the latent correlation ρ.

We start by computing the covariances between the true latents and the learned coordinates. For Z_1 we immediately have

$$\mathrm{Cov}(Z_1, \hat{Z}_1) = s\sigma^2(Z_1) = s, \tag{8}$$
$$\mathrm{Cov}(Z_1, \hat{Z}_2) = 0, \tag{9}$$
$$\mathrm{Cov}(Z_1, \hat{Z}_3) = 0, \tag{10}$$

since Z_1 is uncorrelated with Z_2 and Z_3, and $\sigma^2(Z_1) = 1$.

For Z_2 and $\hat{Z}_2 = aZ_2 + bZ_3$,

$$\begin{aligned}
\mathrm{Cov}(Z_2, \hat{Z}_2) &= \mathrm{Cov}\big(Z_2, aZ_2 + bZ_3\big) && (11)\\
&= a\,\mathrm{Cov}(Z_2, Z_2) + b\,\mathrm{Cov}(Z_2, Z_3) && (12)\\
&= a\sigma^2(Z_2) + b\rho\sqrt{\sigma^2(Z_2)\sigma^2(Z_3)} && (13)\\
&= a + b\rho, && (14)
\end{aligned}$$

using $\sigma^2(Z_2) = \sigma^2(Z_3) = 1$ and $\mathrm{Cov}(Z_2, Z_3) = \rho$.

Similarly, the variance of $\hat{Z}_2$ is

$$\begin{aligned}
\sigma^2(\hat{Z}_2) &= \sigma^2(aZ_2 + bZ_3) && (15)\\
&= a^2\sigma^2(Z_2) + b^2\sigma^2(Z_3) + 2ab\,\mathrm{Cov}(Z_2, Z_3) && (16)\\
&= a^2 + b^2 + 2ab\rho. && (17)
\end{aligned}$$

Therefore

$$\mathrm{Corr}(Z_2, \hat{Z}_2) = \frac{\mathrm{Cov}(Z_2, \hat{Z}_2)}{\sqrt{\sigma^2(Z_2)\sigma^2(\hat{Z}_2)}} = \frac{a + b\rho}{\sqrt{a^2 + b^2 + 2ab\rho}}. \tag{18}$$

Analogously,

$$\mathrm{Cov}(Z_3, \hat{Z}_2) = a\,\mathrm{Cov}(Z_3, Z_2) + b\,\mathrm{Cov}(Z_3, Z_3) = a\rho + b, \tag{19}$$
$$\mathrm{Corr}(Z_3, \hat{Z}_2) = \frac{b + a\rho}{\sqrt{a^2 + b^2 + 2ab\rho}}. \tag{20}$$

Repeating the same computation for $\hat{Z}_3 = cZ_2 + dZ_3$ yields

$$\mathrm{Corr}(Z_2, \hat{Z}_3) = \frac{c + d\rho}{\sqrt{c^2 + d^2 + 2cd\rho}}, \tag{21}$$
$$\mathrm{Corr}(Z_3, \hat{Z}_3) = \frac{d + c\rho}{\sqrt{c^2 + d^2 + 2cd\rho}}. \tag{22}$$

Finally, for Z_1 we have

$$\mathrm{Corr}(Z_1, \hat{Z}_1) = \mathrm{sgn}(s), \qquad \mathrm{Corr}(Z_1, \hat{Z}_2) = \mathrm{Corr}(Z_1, \hat{Z}_3) = 0. \tag{23}$$

F.1.1 Correlation matrix and MCC

Collecting the correlations into a matrix $C(\rho)$, we obtain

$$\begin{aligned}
C(\rho) &= \begin{pmatrix} \mathrm{Corr}(Z_1, \hat{Z}_1) & \mathrm{Corr}(Z_1, \hat{Z}_2) & \mathrm{Corr}(Z_1, \hat{Z}_3) \\ \mathrm{Corr}(Z_2, \hat{Z}_1) & \mathrm{Corr}(Z_2, \hat{Z}_2) & \mathrm{Corr}(Z_2, \hat{Z}_3) \\ \mathrm{Corr}(Z_3, \hat{Z}_1) & \mathrm{Corr}(Z_3, \hat{Z}_2) & \mathrm{Corr}(Z_3, \hat{Z}_3) \end{pmatrix} && (24)\\
&= \begin{pmatrix} \pm 1 & 0 & 0 \\ 0 & r_{22}(\rho) & r_{23}(\rho) \\ 0 & r_{32}(\rho) & r_{33}(\rho) \end{pmatrix}, && (25)
\end{aligned}$$

where $r_{22}, r_{23}, r_{32}, r_{33}$ are given by (18)–(22).

The Mean Correlation Coefficient (MCC) between $\mathbf{z}$ and $\hat{\mathbf{z}}$ for the 3×3 case is defined as

$$\mathrm{MCC}(\rho) = \frac{1}{3} \max_{\pi \in S_3} \sum_{i=1}^{3} \left|\mathrm{Corr}(Z_i, \hat{Z}_{\pi(i)})\right|, \tag{26}$$

where S_3 is the set of all permutations of $\{1, 2, 3\}$.

Since $\mathrm{Corr}(Z_1, \hat{Z}_1) = \pm 1$ and $\mathrm{Corr}(Z_1, \hat{Z}_2) = \mathrm{Corr}(Z_1, \hat{Z}_3) = 0$, the optimal permutation always pairs Z_1 with $\hat{Z}_1$, contributing 1 to the sum. The remaining degrees of freedom are in how we pair (Z_2, Z_3) with $(\hat{Z}_2, \hat{Z}_3)$.

There are two relevant pairings:

1. "Diagonal" pairing: $Z_2 \leftrightarrow \hat{Z}_2$ and $Z_3 \leftrightarrow \hat{Z}_3$, giving a sum

$$S_{\mathrm{diag}}(\rho) = |r_{22}(\rho)| + |r_{33}(\rho)|.$$

2. "Swapped" pairing: $Z_2 \leftrightarrow \hat{Z}_3$ and $Z_3 \leftrightarrow \hat{Z}_2$, giving

$$S_{\mathrm{swap}}(\rho) = |r_{23}(\rho)| + |r_{32}(\rho)|.$$

Thus

$$\mathrm{MCC}(\rho) = \frac{1}{3}\Big(1 + \max\{S_{\mathrm{diag}}(\rho),\ S_{\mathrm{swap}}(\rho)\}\Big). \tag{27}$$

The dependence of MCC on the latent correlation ρ is therefore entirely through the functions $r_{22}(\rho), \ldots, r_{33}(\rho)$.

Effect of the sign of ρ in a symmetric example. To see how the sign of ρ affects $\mathrm{MCC}(\rho)$, consider a symmetric mixing:

$$\hat{Z}_2 = Z_2 + \varepsilon Z_3, \tag{28}$$
$$\hat{Z}_3 = \varepsilon Z_2 + Z_3, \tag{29}$$

with $0 < \varepsilon < 1$. In this case

$$a = d = 1, \qquad b = c = \varepsilon,$$

and substituting into (18)–(22) yields

$$r_{22}(\rho) = \frac{1 + \varepsilon\rho}{\sqrt{1 + \varepsilon^2 + 2\varepsilon\rho}}, \tag{30}$$
$$r_{33}(\rho) = \frac{1 + \varepsilon\rho}{\sqrt{1 + \varepsilon^2 + 2\varepsilon\rho}} = r_{22}(\rho), \tag{31}$$

and the off-diagonal entries are

$$r_{32}(\rho) = \frac{\varepsilon + \rho}{\sqrt{1 + \varepsilon^2 + 2\varepsilon\rho}}, \tag{32}$$
$$r_{23}(\rho) = \frac{\varepsilon + \rho}{\sqrt{\varepsilon^2 + 1 + 2\varepsilon\rho}} = r_{32}(\rho). \tag{33}$$

To verify: from (22) with $d = 1$, $c = \varepsilon$, the numerator is $d + c\rho = 1 + \varepsilon\rho$, matching r_{22}. From (20) with $a = 1$, $b = \varepsilon$, the numerator is $b + a\rho = \varepsilon + \rho$. From (21) with $c = \varepsilon$, $d = 1$, the numerator is $c + d\rho = \varepsilon + \rho$, matching r_{32}. All four denominators equal $\sqrt{1 + \varepsilon^2 + 2\varepsilon\rho}$.

The two pairings therefore give

$$S_{\mathrm{diag}}(\rho) = |r_{22}(\rho)| + |r_{33}(\rho)| = 2\,|r_{22}(\rho)|, \tag{34}$$
$$S_{\mathrm{swap}}(\rho) = |r_{23}(\rho)| + |r_{32}(\rho)| = 2\,|r_{32}(\rho)|. \tag{35}$$

These are *not* equal in general. We now show that the diagonal pairing always dominates. Consider the difference of the numerators:

$$(1 + \varepsilon\rho) - (\varepsilon + \rho) = (1 - \varepsilon)(1 - \rho) \geq 0, \tag{36}$$

$$(1 + \varepsilon\rho) + (\varepsilon + \rho) = (1 + \varepsilon)(1 + \rho) > 0, \tag{37}$$

for all $\rho \in (-1, 1)$ and $\varepsilon \in (0, 1)$. Since both share the same (positive) denominator, we have $|r_{22}(\rho)| \geq |r_{32}(\rho)|$ with equality only at the boundary $\rho = 1$ or $\varepsilon = 1$. Hence $S_{\text{diag}}(\rho) \geq S_{\text{swap}}(\rho)$, and

$$\text{MCC}(\rho) = \frac{1}{3}\Big(1 + 2\,|r_{22}(\rho)|\Big) = \frac{1}{3}\left(1 + \frac{2\,|1 + \varepsilon\rho|}{\sqrt{1 + \varepsilon^2 + 2\varepsilon\rho}}\right). \tag{38}$$

Since $1 + \varepsilon\rho \geq 1 - \varepsilon > 0$ for all $|\rho| < 1$, the absolute value is redundant and we may write

$$\text{MCC}(\rho) = \frac{1}{3}\left(1 + \frac{2(1 + \varepsilon\rho)}{\sqrt{1 + \varepsilon^2 + 2\varepsilon\rho}}\right). \tag{39}$$

Derivative of $r_{22}(\rho)$. We now compute the derivative of $r_{22}(\rho)$ with respect to ρ. Write $N(\rho) := 1 + \varepsilon\rho$ and $D(\rho) := 1 + \varepsilon^2 + 2\varepsilon\rho$, so that $r_{22} = N/\sqrt{D}$. Then

$$\frac{\partial r_{22}}{\partial \rho} = \frac{N'\sqrt{D} - N \cdot \frac{D'}{2\sqrt{D}}}{D} = \frac{N'\,D - N \cdot \frac{D'}{2}}{D^{3/2}}. \tag{40}$$

We have $N' = \varepsilon$ and $D' = 2\varepsilon$, so the numerator is

$$\begin{aligned} N'\,D - N \cdot \tfrac{D'}{2} &= \varepsilon\,(1 + \varepsilon^2 + 2\varepsilon\rho) - (1 + \varepsilon\rho) \cdot \varepsilon \\ &= \varepsilon\big[(1 + \varepsilon^2 + 2\varepsilon\rho) - (1 + \varepsilon\rho)\big] \\ &= \varepsilon\big(\varepsilon^2 + \varepsilon\rho\big) = \varepsilon^2(\varepsilon + \rho). \end{aligned} \tag{41}$$

Therefore

$$\frac{\partial r_{22}}{\partial \rho} = \frac{\varepsilon^2(\varepsilon + \rho)}{\big(1 + \varepsilon^2 + 2\varepsilon\rho\big)^{3/2}}. \tag{42}$$

The denominator in (42) is strictly positive for all $\rho \in (-1, 1)$ and $0 < \varepsilon < 1$, since

$$1 + \varepsilon^2 + 2\varepsilon\rho \geq 1 + \varepsilon^2 - 2\varepsilon = (1 - \varepsilon)^2 > 0.$$

Non-monotonicity of $r_{22}(\rho)$. The numerator in (42) changes sign at $\rho^* = -\varepsilon$:

$$\frac{\partial r_{22}}{\partial \rho} \begin{cases} < 0 & \text{if } \rho < -\varepsilon, \\ = 0 & \text{if } \rho = -\varepsilon, \\ > 0 & \text{if } \rho > -\varepsilon. \end{cases} \tag{43}$$

Thus $r_{22}(\rho)$ is *not* monotonically increasing on $(-1, 1)$. It attains its minimum at $\rho^* = -\varepsilon$, where

$$r_{22}(-\varepsilon) = \frac{1 - \varepsilon^2}{\sqrt{1 + \varepsilon^2 - 2\varepsilon^2}} = \frac{1 - \varepsilon^2}{\sqrt{1 - \varepsilon^2}} = \sqrt{1 - \varepsilon^2}. \tag{44}$$

Monotonicity of MCC. Since $1 + \varepsilon\rho > 0$ on $(-1, 1)$, we have $|r_{22}| = r_{22}$, and by (39) the MCC inherits the same monotonicity structure: decreasing on $(-1, -\varepsilon)$ and increasing on $(-\varepsilon, 1)$, with minimum

$$\text{MCC}(-\varepsilon) = \frac{1}{3}\big(1 + 2\sqrt{1 - \varepsilon^2}\,\big). \tag{45}$$

At the boundary values:

$$\lim_{\rho \to 1} r_{22}(\rho) = \frac{1 + \varepsilon}{\sqrt{(1 + \varepsilon)^2}} = 1, \tag{46}$$

$$\lim_{\rho \to -1^+} r_{22}(\rho) = \frac{1 - \varepsilon}{\sqrt{(1 - \varepsilon)^2}} = 1. \tag{47}$$

Hence $r_{22}(\rho) \to 1$ at *both* extremes $\rho \to \pm 1$, and $\text{MCC}(\rho) \to 1$ in both limits. The minimum of MCC is in the interior, at $\rho = -\varepsilon$.

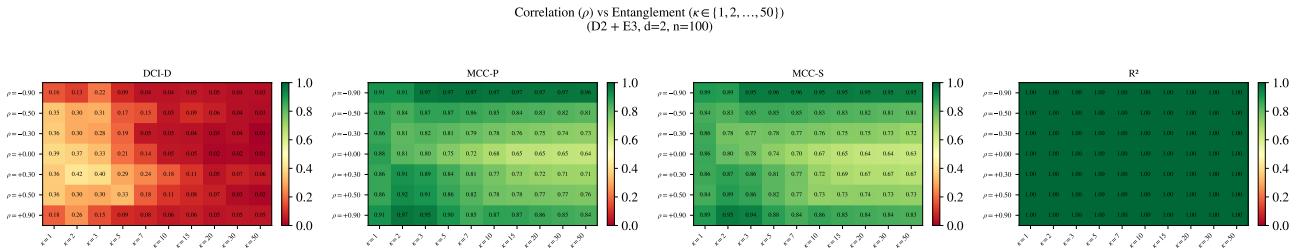

Figure 6: MCC overestimates identifiability as ρ increases.

Implication. Even though $\hat{Z}_2$ and $\hat{Z}_3$ remain linearly entangled mixtures of Z_2 and Z_3 for all $\varepsilon > 0$, the MCC score varies with the correlation ρ between the ground-truth factors, despite the underlying entanglement structure of the learned representation being unchanged.

For the practically relevant regime $\rho > 0$, positive correlations inflate MCC monotonically: the MCC is strictly increasing on $(0, 1)$ since $0 > -\varepsilon = \rho^*$. For negative correlations, the MCC first decreases (reaching its minimum at $\rho = -\varepsilon$) and then increases again toward 1 as $\rho \to -1$.

The non-trivial dependence on ρ—including the fact that the minimum is in the interior and that the MCC approaches 1 at *both* boundary values $\rho \to \pm 1$—demonstrates that MCC conflates representation quality with the covariance structure of the ground-truth factors.

To illustrate this further, consider an even simpler (degenerate) example:

$$a = b = c = d = 1, \tag{48}$$

so that

$$\hat{Z}_2 = Z_2 + Z_3, \qquad \hat{Z}_3 = Z_2 + Z_3. \tag{49}$$

In other words, $\hat{Z}_2$ and $\hat{Z}_3$ are identical, fully redundant, and both are symmetric mixtures of Z_2 and Z_3.

Using the same covariance structure as before, with

$$\sigma^2(Z_2) = \sigma^2(Z_3) = 1, \qquad \mathrm{Cov}(Z_2, Z_3) = \rho,$$

we compute

$$\begin{aligned} \mathrm{Cov}(Z_2, \hat{Z}_2) &= \mathrm{Cov}\big(Z_2, Z_2 + Z_3\big) = \sigma^2(Z_2) + \mathrm{Cov}(Z_2, Z_3) = 1 + \rho, \\ \sigma^2(\hat{Z}_2) &= \sigma^2(Z_2 + Z_3) = \sigma^2(Z_2) + \sigma^2(Z_3) + 2\,\mathrm{Cov}(Z_2, Z_3) = 2 + 2\rho. \end{aligned}$$

Hence, for $\rho > -1$,

$$\mathrm{Corr}(Z_2, \hat{Z}_2) = \frac{1+\rho}{\sqrt{2+2\rho}} = \sqrt{\frac{1+\rho}{2}} =: r(\rho). \tag{50}$$

Verification via the general formula. Setting $\varepsilon = 1$ in (30) gives $r_{22} = (1+\rho)/\sqrt{2+2\rho}$, matching (50).

By symmetry we also have

$$\mathrm{Corr}(Z_3, \hat{Z}_2) = \mathrm{Corr}(Z_2, \hat{Z}_2) = \mathrm{Corr}(Z_2, \hat{Z}_3) = \mathrm{Corr}(Z_3, \hat{Z}_3) = r(\rho),$$

and $\mathrm{Corr}(Z_1, \hat{Z}_1) = \pm 1$, $\mathrm{Corr}(Z_1, \hat{Z}_2) = \mathrm{Corr}(Z_1, \hat{Z}_3) = 0$ as before.

In this case, the two candidate pairings (diagonal and swapped) give the same sum, and the MCC simplifies to

$$\mathrm{MCC}(\rho) = \frac{1}{3}\big(1 + 2\,r(\rho)\big) = \frac{1}{3}\Big(1 + 2\sqrt{\tfrac{1+\rho}{2}}\Big), \qquad -1 < \rho \le 1. \tag{51}$$

Note that this degenerate case corresponds to $\varepsilon = 1$, where the minimum of r_{22} from (43) occurs at $\rho^* = -1$ (the boundary), consistent with $r(\rho)$ being monotonically increasing on $(-1, 1)$.

This expression makes two important properties explicit.

(i) Asymmetry $\mathrm{MCC}(\rho) \neq \mathrm{MCC}(-\rho)$**.** From (51) we obtain

$$r(\rho) = \sqrt{\frac{1+\rho}{2}}, \quad -1 < \rho \leq 1. \tag{52}$$

Substituting $-\rho$ in place of ρ yields

$$\mathrm{MCC}(-\rho) = \frac{1}{3}\left(1 + 2\sqrt{\frac{1-\rho}{2}}\right). \tag{53}$$

Hence, for any $\rho \in (0, 1)$,

$$\mathrm{MCC}(\rho) = \frac{1}{3}\left(1 + 2\sqrt{\frac{1+\rho}{2}}\right) \neq \frac{1}{3}\left(1 + 2\sqrt{\frac{1-\rho}{2}}\right) = \mathrm{MCC}(-\rho).$$

Even though the entanglement structure for (Z_2, Z_3) is symmetric under the sign flip $\rho \mapsto -\rho$, MCC values differ for positive and negative correlations.

(ii) Faithfulness issues at extreme correlations. The same formula reveals a qualitative difference between the limits $\rho \to 1$ and $\rho \to -1$:

$$\lim_{\rho \to 1} r(\rho) = \sqrt{\frac{1+1}{2}} = 1, \tag{54}$$

$$\lim_{\rho \to -1^+} r(\rho) = \sqrt{\frac{1+(-1)}{2}} = 0. \tag{55}$$

When $\rho \to 1$, Eq. (51) yields $\mathrm{MCC}(\rho) \to (1+2)/3 = 1$. In contrast, when $\rho \to -1$, $\mathrm{MCC}(\rho) \to (1+0)/3 \approx 0.33$.

Contrast with the $\varepsilon < 1$ **case.** In the general symmetric mixing with $\varepsilon < 1$, eqs. (46)–(47) show that $r_{22}(\rho) \to 1$ at *both* $\rho \to +1$ and $\rho \to -1$, so $\mathrm{MCC}(\rho) \to 1$ in both limits. The faithfulness collapse (MCC $\to 0.33$) at $\rho \to -1$ is specific to the degenerate case $\varepsilon = 1$, where $\hat{Z}_2 = \hat{Z}_3 = Z_2 + Z_3 \approx 0$ when $Z_3 \approx -Z_2$. For $\varepsilon < 1$, the representations $\hat{Z}_2 \neq \hat{Z}_3$ remain distinct and their correlations with the true factors recover to 1 as $\rho \to -1$.

These examples show that MCC depends nontrivially on the covariance structure of the ground-truth factors, independently of the underlying entanglement of the learned representation. In particular, for a fixed mixing matrix, MCC can be artificially inflated or deflated by the latent correlation ρ.

F.2 EXPECTED BEHAVIOUR OF DCI

We derive properties of DCI that explain the metric's behaviour in the main-text experiments. We first recall the construction, then state four results organised by failure mode.

Construction. A supervised probe is trained to predict each ground-truth factor z_j from the learned representation $\hat{\mathbf{z}}$, yielding a nonnegative importance matrix $\mathbb{R} \in \mathbb{R}_{\geq 0}^{m \times d}$, where $R_{i,j}$ quantifies the contribution of learned feature $\hat{z}_i$ in predicting z_j. Row $\mathbb{R}_{i,:}$ summarises which factors feature i encodes; column $\mathbb{R}_{:,j}$ summarises which features encode factor j. The DCI scores are computed from $\mathbb{R}$ alone:

Disentanglement. Convert each row to a distribution $p_{j|i} = R_{i,j} / \sum_k R_{i,k}$ and measure concentration:

$$D_i = 1 - \frac{H(p_{\cdot|i})}{\log d}, \qquad D_{\mathrm{DCI}} = \sum_{i=1}^{m} w_i \, D_i, \qquad w_i = \frac{\sum_j R_{i,j}}{\sum_{i',j'} R_{i',j'}}.$$

Completeness. Convert each column to a distribution $\tilde{p}_{i|j} = R_{i,j} / \sum_k R_{k,j}$ and measure concentration:

$$C_j = 1 - \frac{H(\tilde{p}_{\cdot|j})}{\log m}, \qquad C_{\mathrm{DCI}} = \sum_{j=1}^{d} v_j \, C_j, \qquad v_j = \frac{\sum_i R_{i,j}}{\sum_{i',j'} R_{i',j'}}.$$

Informativeness. $I_{\mathrm{DCI}} = \frac{1}{d}\sum_j (1 - L_j)$, where L_j is the normalised prediction loss (e.g., $1 - R_j^2$) of the probe for factor j.

The weights w_i and v_j are proportional to total importance: features or factors with negligible importance contribute negligibly to the global scores. This weighting is the source of the first failure mode.

Proposition 2 (Dropped factors are invisible to DCI)**.** *Under* $\mathbf{D}_\perp$ + ***E4*** *with* $|S| = m < d$ *perfectly identified factors, as* $n \to \infty$*:* $D_{\mathrm{DCI}} \to 1$ *and* $C_{\mathrm{DCI}} \to 1$.

Proof. For retained factors ($j \in S$), the encoder is elementwise, so the probe importance concentrates on a single coordinate: $p_{j|i}$ and $\tilde{p}_{i|j}$ are one-hot for the matched pair, giving $D_i = 1$ and $C_j = 1$.

For discarded factors ($j \notin S$), no learned feature predicts them: $R_{i,j} \approx 0$ for all i. Their weight $v_j \propto \sum_i R_{i,j} \approx 0$ vanishes from C_{DCI}. Likewise, the rows corresponding to codes that encode only retained factors carry all the weight in D_{DCI}.

DCI does not verify that *all* factors are represented; factors that are never encoded produce zero importance, vanish from the weighted averages, and do not penalise the score. □

Implication: This explains the DCI-D false positive in Fig. 3 (left, $\mathbf{D}_\perp$): as factors are dropped, D or C do not penalise omission (only I_{DCI} would drop).

Proposition 3 (Functional dependence decreases D under a perfect encoder)**.** *Under* $\mathbf{D}_f$ *with* $z_2 = f(z_1)$ *(deterministic) and a perfect elementwise encoder* ***E1*** *(*$\hat{z}_j = a_j z_j$*), a nonlinear probe (e.g., gradient boosted trees) yields* $D_{\mathrm{DCI}} < 1$.

Proof. Since $z_2 = f(z_1)$ exactly, $\hat{z}_1 = a_1 z_1$ perfectly determines z_2 via f, so the nonlinear probe assigns $R_{1,2} > 0$ in addition to $R_{1,1} > 0$. Symmetrically, when f is invertible (e.g., $f(z) = z^3$), $\hat{z}_2 = a_2 f(z_1)$ determines z_1 via f^{-1}, so $R_{2,1} > 0$ and $D_2 < 1$. The remaining $d - 2$ codes are independent and achieve $D_i = 1$, but the deflated D_1 and D_2 pull down D_{DCI} through their nonzero weights $w_1, w_2 > 0$. □

With a linear probe (e.g., Lasso), the result can differ. For $z_1 \sim \mathcal{N}(0, 1)$ and $z_2 = z_1^3$, the population normal equations yield zero cross-coefficients—z_1 and z_1^3 are linearly orthogonal under Gaussian moments—so $\mathbb{R}$ is diagonal and $D_{\mathrm{DCI}} = 1$. The deflation under $\mathbf{D}_f$ is therefore *probe-dependent*: it arises only when the probe is expressive enough to detect the functional relationship f.

Implication. This explains the DCI-D dip in the $\mathbf{D}_f$ panels of Fig. 3 (right): even at $m = d$ (dashed line), DCI-D is below 1.0 because the functional constraint between z_1 and z_2 spreads importance across the corresponding codes.

F.3 MCC FALSE-POSITIVE RATE UNDER NULL ENCODERS

We derive the expected behaviour of MCC-P when the learned representation is independent of the ground-truth factors, explaining the inflation observed in Fig. 5.

Setup. Let $\mathbf{z} \in \mathbb{R}^d$ and $\hat{\mathbf{z}} \in \mathbb{R}^m$ be independent random vectors (null encoder), and let $(z^{(1)}, \hat{z}^{(1)}), \ldots, (z^{(n)}, \hat{z}^{(n)})$ be n i.i.d. paired samples. The sample Pearson correlation between $\hat{z}_i$ and z_j is

$$\hat{\rho}_{ij} = \frac{\sum_{t=1}^n (\hat{z}_i^{(t)} - \bar{\hat{z}}_i)(z_j^{(t)} - \bar{z}_j)}{\sqrt{\sum_t (\hat{z}_i^{(t)} - \bar{\hat{z}}_i)^2}\,\sqrt{\sum_t (z_j^{(t)} - \bar{z}_j)^2}}.$$

Since $\hat{\mathbf{z}} \perp \mathbf{z}$, the true correlation is $\rho_{ij} = 0$ for all i, j.

Distribution of sample correlations under the null. For bivariate normal data with $\rho = 0$, the sample correlation satisfies

$$\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim t_{n-2} \tag{56}$$

exactly (Fisher, 1922). For non-Gaussian data, the exact t-distribution does not hold, but the asymptotic result $\sqrt{n}\,\hat{\rho} \xrightarrow{d} \mathcal{N}(0, 1)$ follows from the Central Limit Theorem (CLT) (Hoeffding, 1992). In either case, for large n,

$$\hat{\rho}_{ij} \overset{\text{approx}}{\sim} \mathcal{N}\left(0, \frac{1}{n}\right). \tag{57}$$

Maximum absolute correlation. MCC-P computes the $m \times d$ matrix of absolute sample correlations $|\hat{\rho}_{ij}|$ and applies Hungarian matching to find the optimal one-to-one assignment.

Consider a single column j. The entries $\{|\hat{\rho}_{ij}|\}_{i=1}^{m}$ are approximately half-normal with scale $1/\sqrt{n}$. (They are not exactly independent—they share the $z_j^{(t)}$ samples—but the dependence is weak under the null since the $\hat{z}_i$ are independent across rows; Cai and Jiang (2011) handle this rigorously.) The maximum of m such entries satisfies, by standard extreme value theory for Gaussian maxima,

$$\mathbb{E}\left[\max_{i=1}^{m} |\hat{\rho}_{ij}|\right] \approx \sqrt{\frac{2\log m}{n}}. \tag{58}$$

MCC-P under the null. To lower-bound the Hungarian matching, consider a greedy assignment: assign column 1 its best row, remove that row, assign column 2 its best among the remaining $m-1$ rows, and so on. This produces a valid one-to-one assignment, and column j selects from $m-j+1$ remaining candidates. When $m \gg d$, every column still has $\approx m$ candidates, and the greedy score is close to the average column-wise maximum. Since the Hungarian matching is optimal over all one-to-one assignments, it scores at least as high as the greedy, giving

$$\mathbb{E}[\text{MCC-P}] \gtrsim \frac{1}{d}\sum_{j=1}^{d}\sqrt{\frac{2\log(m-j+1)}{n}} \approx \sqrt{\frac{2\log m}{n}} \qquad \text{when } m \gg d. \tag{59}$$

This bound is non-negligible whenever $\log m/n$ is not small. In practice this inflation is substantial even at moderate ratios:

m/n	$\sqrt{2\log m/n}$ (bound)	MCC-P (observed, m/d=1)
0.1	0.21	~ 0.3
0.2	0.30	~ 0.5
0.5	0.48	~ 0.83
1.0	0.68	~ 0.95

The bound captures the correct scaling: it explains why m/n governs the false-positive rate. But, it underestimates the magnitude at practical sample sizes for two reasons: (i) the extreme value approximation is loose at small m; and (ii) at small n, the exact null distribution of $\hat{\rho}$ follows a scaled t_{n-2} (Eq. (56)), which has heavier tails than the Gaussian, pushing the maximum correlation above the asymptotic prediction.

Extension to MCC-S. Spearman correlation is the Pearson correlation applied to ranks. Under independence, the sample Spearman correlation also satisfies $\hat{\rho}^S_{ij} \approx \mathcal{N}(0, 1/n)$ (Hotelling and Pabst, 1936), so the extreme value argument above applies verbatim: the $\sqrt{2\log m/n}$ floor governs both MCC-P and MCC-S.

Why m/n governs and m/d does not. The bound (58) depends on m (candidates per column) and n (sample size), but not on d (number of columns). Adding more ground-truth factors adds more columns to the matching problem but does not change the distribution of each column's maximum. The MCC averaging divides by d, but since each column contributes approximately the same expected maximum, the average is $\approx \sqrt{2\log m/n}$ regardless of d. This is consistent with the empirical observation in Fig. 5: reading along rows (fixed m/d, varying m/n), scores increase; reading along columns (fixed m/n, varying m/d), scores are approximately constant.

Comparison with R^2 and DCI-D. R^2 uses cross-validated nonlinear regression, which does not exploit the maximum over candidates: it predicts each factor independently and averages the explained variance. Under the null, the cross-validated $R^2_j \approx 0$ for each factor (overfitting is penalised by the held-out evaluation), so $R^2 \approx 0$ regardless of m/n.

DCI-D trains a Lasso probe for each factor. Under the null, the ℓ_1 penalty shrinks most coefficients to zero, but a few features can be spuriously selected— particularly when m is large relative to n (Bühlmann and Van De Geer, 2011). The resulting importance matrix has most entries near zero; the moderate inflation at high m/n and low m/d in Fig. 5 is consistent with a small number of spuriously selected features spreading enough importance mass to inflate the disentanglement score.

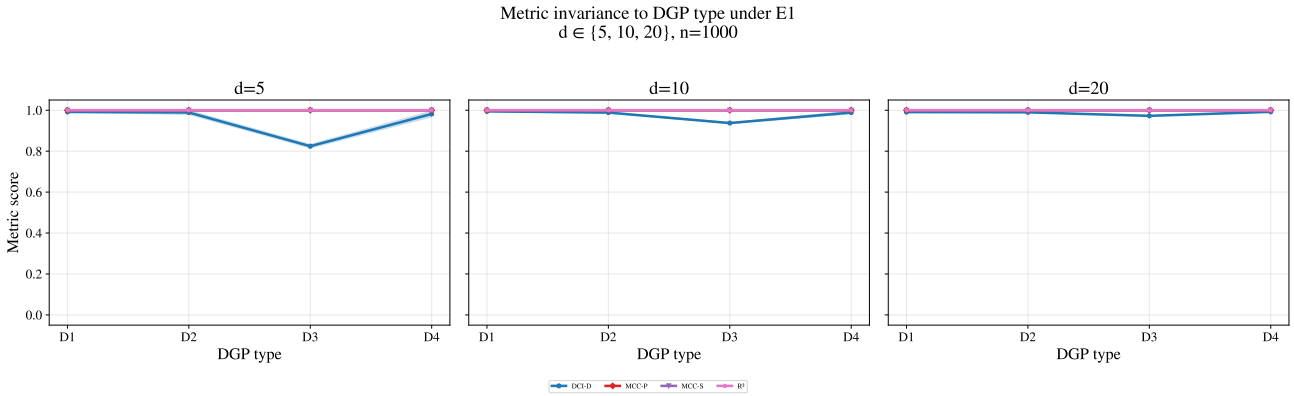

Figure 7: DCI-D is not stable for **D**.

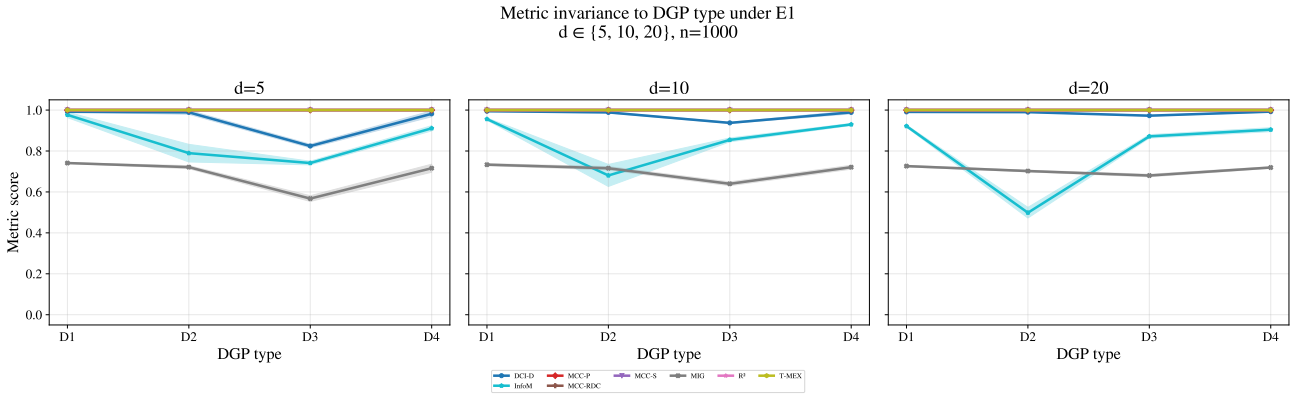

Figure 8: MIG reports < 1 even for the ideal $\mathbf{D}_{\perp}$–**E1** case.

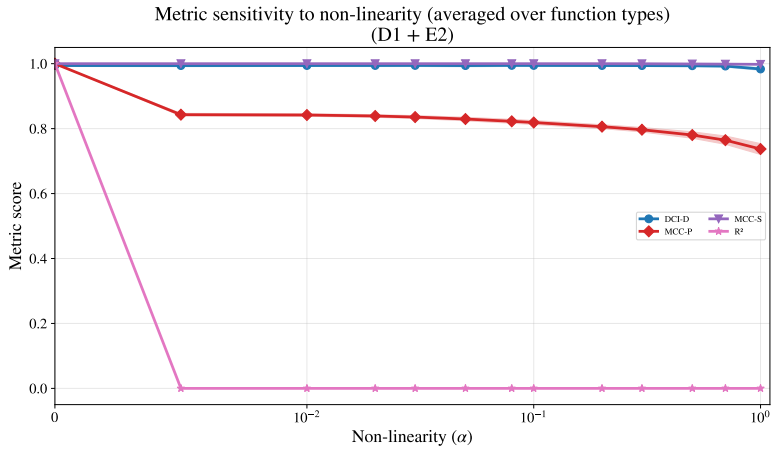

Figure 9: DCI-D and MCC-S are quite stable against increasing non-linearity strength, hence reliable for evaluating **E2**.

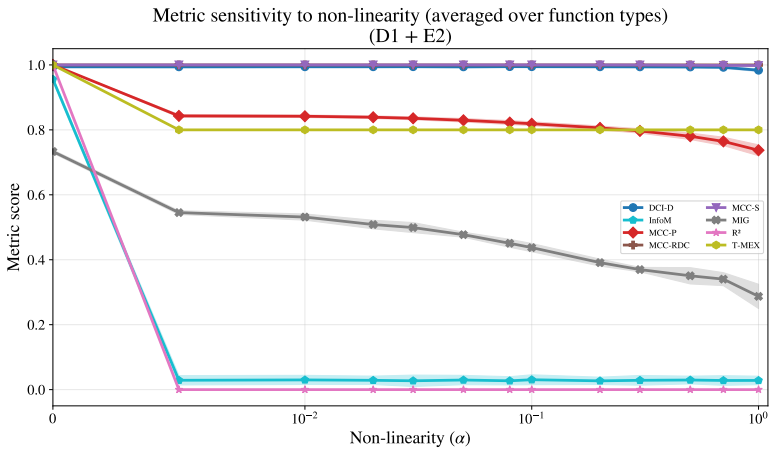

Figure 10: MI-based metrics struggle to evaluate **E2**.

G EXPERIMENTS

G.1 SANITY CHECKS

G.2 CORRELATION AMONG LATENT FACTORS

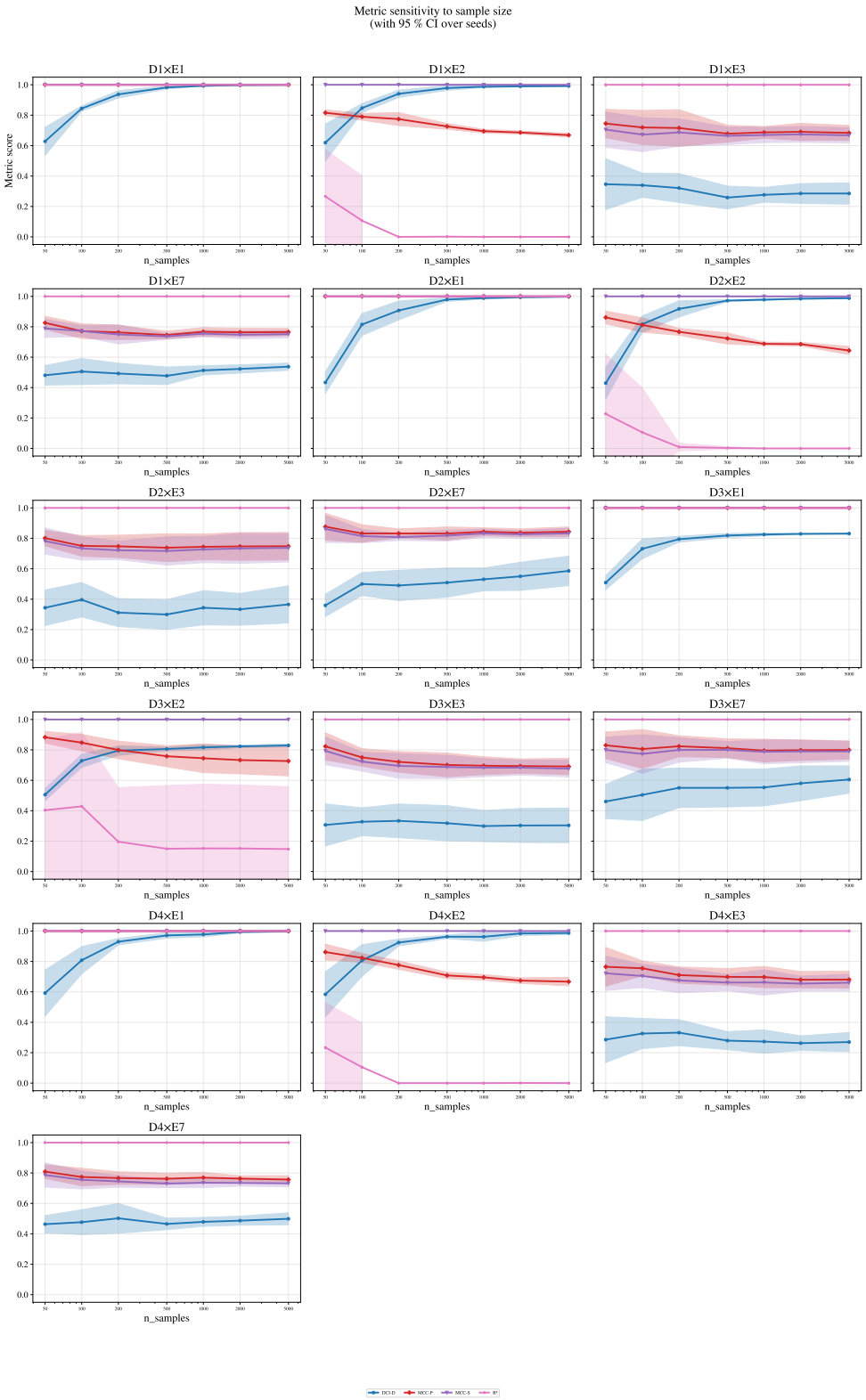

Figure 11: Sample sensitivity of four main metrics (DCI-D, MCC-P, MCC-S, R^2) across all DGP × encoder combinations ($n \in 50, \ldots, 5000$, d=5). MCC-P and MCC-S are sample-efficient: they stabilise by n=100 in most settings. DCI-D requires $n \gtrsim 500$ to converge, with wide confidence intervals at n<200, particularly under entangled encoders (E2, E3). The overcomplete encoder E7 (m=2d=10) inflates DCI-D variance at low n because twice as many codes must be estimated from the same number of samples.

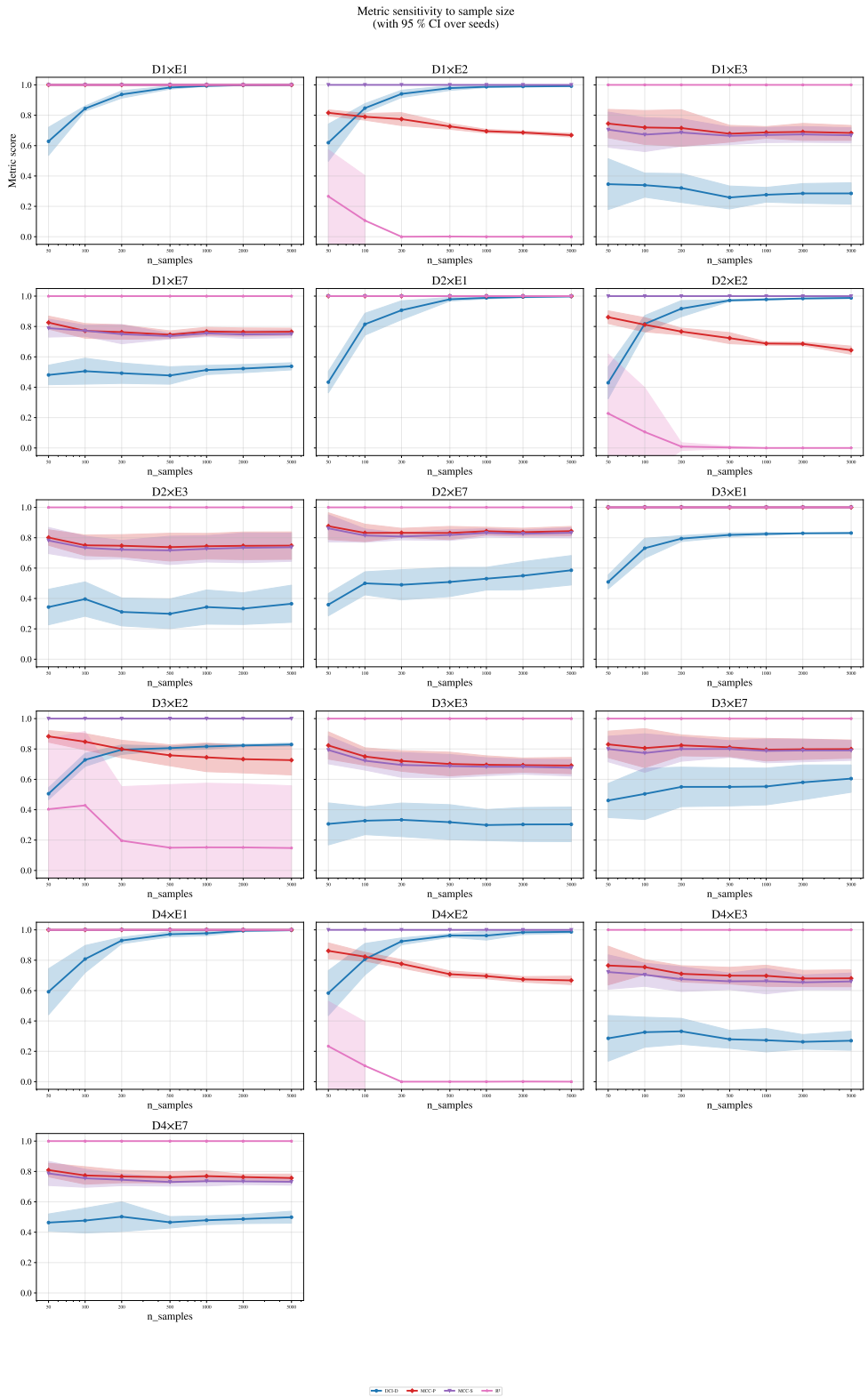

Figure 12: Full metric suite (adding InfoM, MIG, MCC-RDC, T-MEX to the main four). InfoM and MIG produce NaN at n=50 across all DGP × encoder combinations (visible as missing lines), making them unusable below $n \approx 100$. MCC-RDC converges slowly under nonlinear encoders (E2), lagging behind MCC-P and MCC-S until $n \gtrsim 1000$. T-MEX likewise returns NaN universally at small n. The variance heatmap confirms that R^2 under E2 is the single worst-case cell (std ≈ 0.35 at n=50), while correlation-based metrics remain stable (std < 0.01). Overall, metrics split into two reliability tiers: correlation-based measures (MCC-P, MCC-S) are robust across sample sizes, while predictor-based (R^2, DCI-D) and information-theoretic (InfoM, MIG, T-MEX) metrics require $n \gtrsim 500$ for trustworthy estimates.

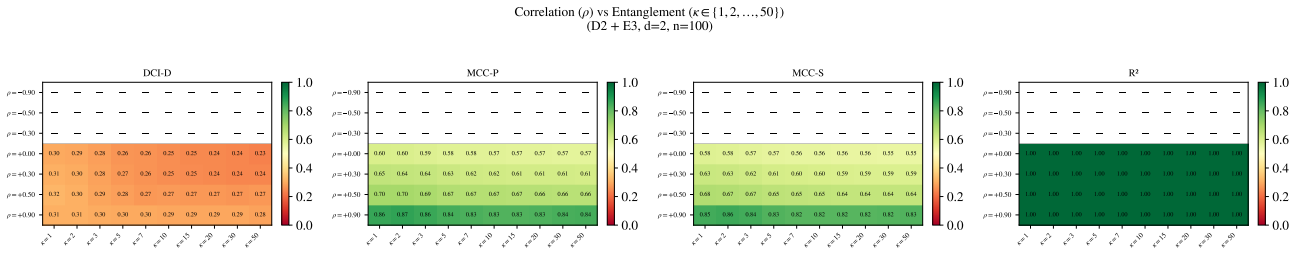

Figure 13: **Disentangling the effects of correlation ρ and entanglement κ (d=10).** An ideal metric would vary only along columns (increasing κ, i.e. worse entanglement) and be constant along rows (changing ρ). MCC-P and MCC-S show clear row-wise gradients, confirming violation of Property 1. DCI-D is more stable but collapses to near-zero even under moderate entanglement (for all $\kappa > 1$).

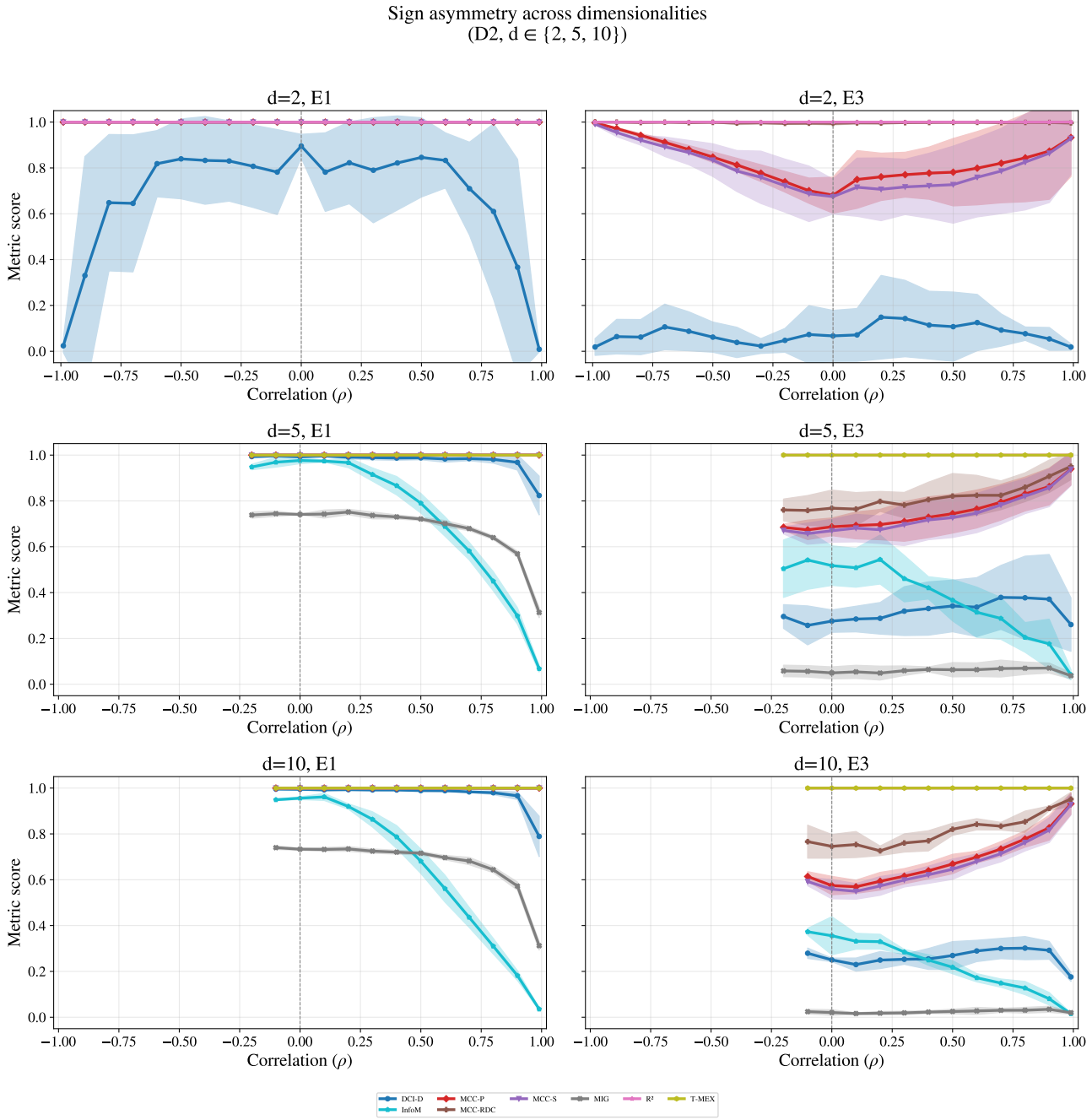

Figure 14: Full metric suite across dimensionalities. The additional metrics reveal two distinct behaviours. MIG (grey) and MCC-RDC (cyan) degrade sharply with increasing $|\rho|$ under E1 at $d \geq 5$: MIG drops from ~0.8 at ρ=0 to ~0.2 at ρ=0.99, while MCC-RDC follows a similar decline, indicating that these metrics conflate inter-factor correlation with non-identifiability. InfoM and T-MEX are absent at d=2 (NaN) but appear at $d \geq 5$; T-MEX (yellow) remains flat near 1.0 under E1 regardless of ρ, showing complete sign- and correlation-invariance. Under E3, the metric spread widens substantially: InfoM (dark blue) tracks DCI-D but at lower absolute values, while MCC-RDC drops more steeply than MCC-P, suggesting that the RDC kernel is sensitive to the interaction between entanglement and inter-factor correlation. The key takeaway is that correlation-based metrics (MCC-P, MCC-S) and T-MEX are robust to the sign of ρ, while tree-based (DCI-D) and MI-based (MIG) metrics are sensitive to it, particularly at low d.

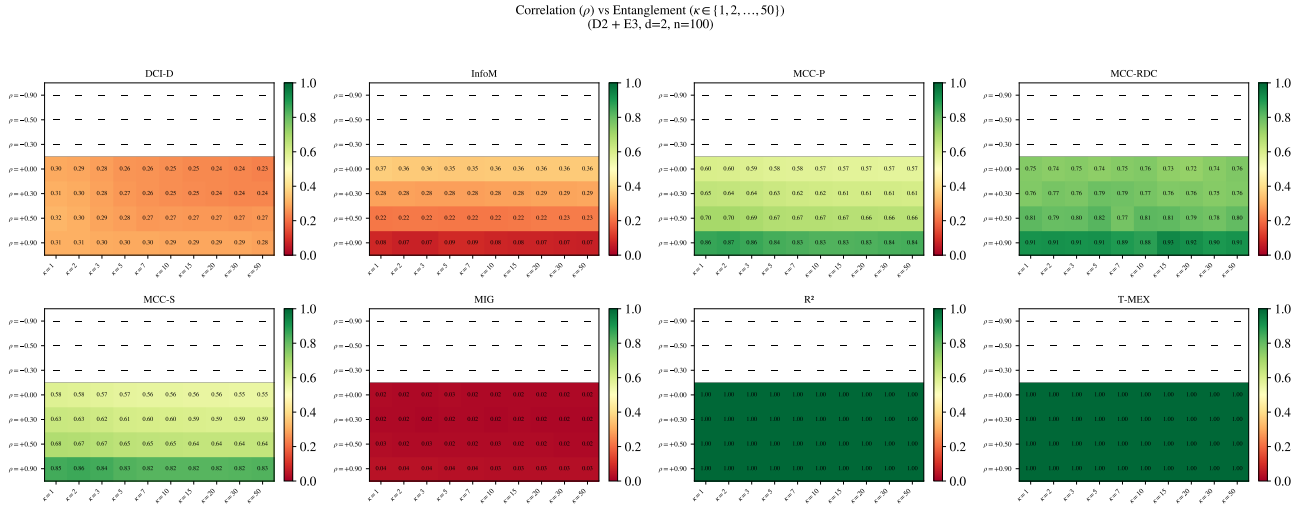

Figure 15: **Full metric suite: ρ–κ heatmaps at d=10.** Extension of Fig. 13 to all metrics. An ideal metric is constant along rows (varying ρ at fixed κ); row-wise gradients indicate spurious sensitivity to inter-factor correlation. d=10, n=1000.

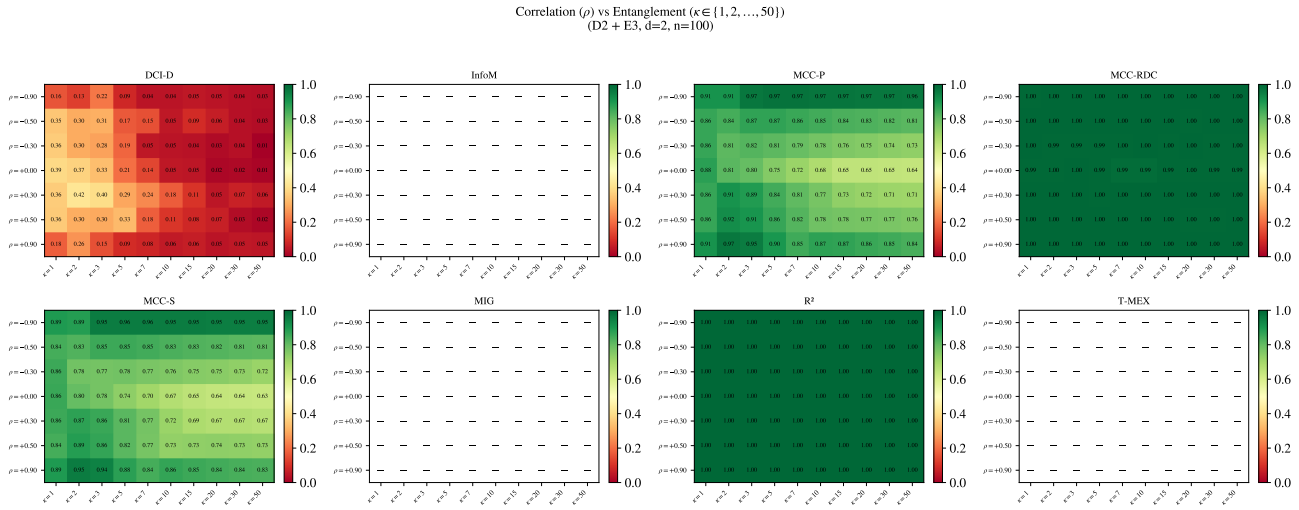

Figure 16: **Full metric suite: ρ–κ heatmaps at d=5.** Extension of Fig. 6 to all metrics. Same layout as Fig. 15; qualitative conclusions carry over from d=10 to d=5. n=1000.

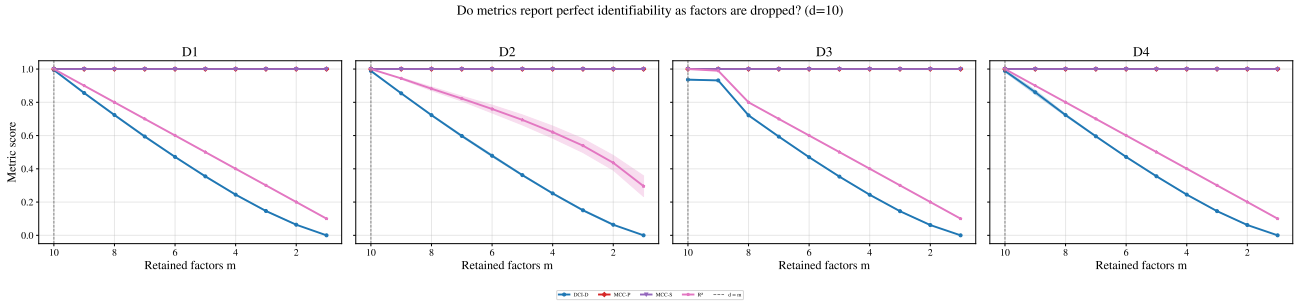

Figure 17: **Metric scores as a function of the number of retained factors across all DGP types.** Extension of Fig. 3 to $\mathbf{D}_\rho$ and $\mathbf{D}_F$. Under $\mathbf{D}_\rho$, R^2 exceeds m/d because the probe partially predicts dropped factors from correlated retained ones. Under $\mathbf{D}_F$ ($z_k = g(z_i, z_j)$, $d_{\text{eff}} = d-1$), metric behaviour is indistinguishable from $\mathbf{D}_\perp$ despite the first omission being lossless: no metric detects the multi-factor redundancy. MCC-P/S remain at 1.0 across all panels. d=10, n=1000.

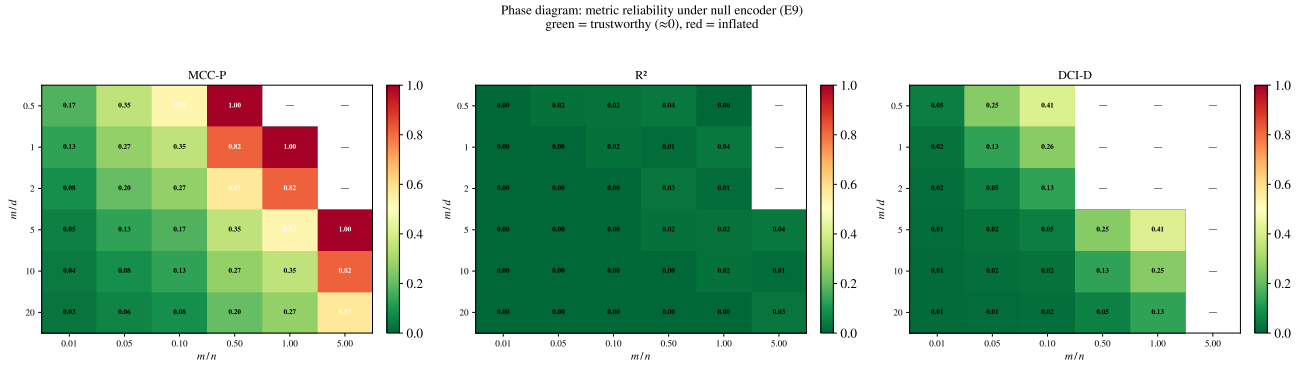

Figure 18: **False positives under a Gaussian null encoder.** Same layout as Fig. 5 (uniform null) but with $\hat{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. The false-positive pattern is nearly identical: MCC-P/S scores are governed by m/n, not m/d, confirming that the inflation is independent of the null distribution.

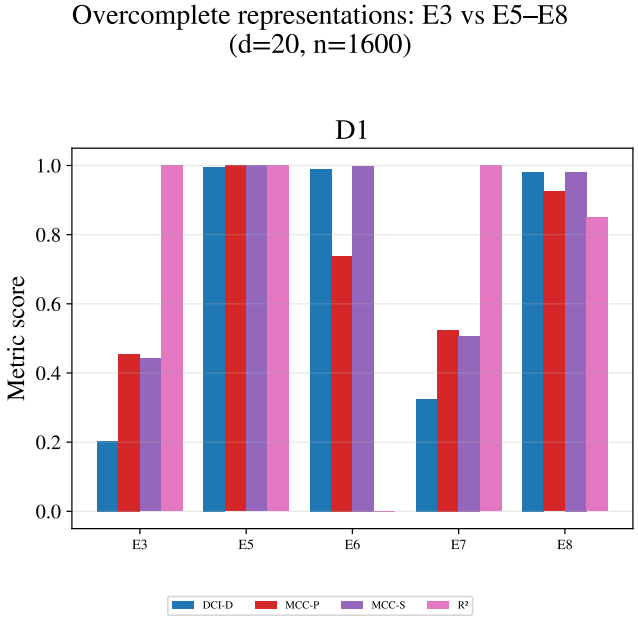

Figure 19: **At moderate overcompleteness (m/d=2),a metrics distinguish entanglement from redundancy.** Overcomplete disentangled encoders (E5–E8) score near 1.0 on DCI-D, MCC, and R^2, whereas the entangled encoder E7 is correctly penalised by DCI-D and MCC. d=20, n=1600. See Fig. 20 for all DGPs.

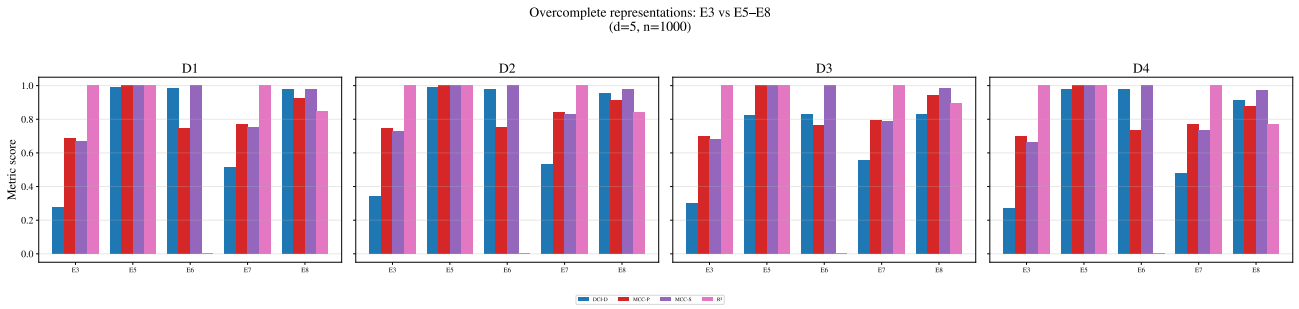

Figure 20: **Overcomplete encoders across all DGP types (d=5, n=1000).** Extension of Fig. 19 from d=20 to d=5. At this smaller d, DCI-D and MCC still separate disentangled overcomplete encoders (E5–E8) from the entangled baseline E7, though the gap is narrower than at d=20.

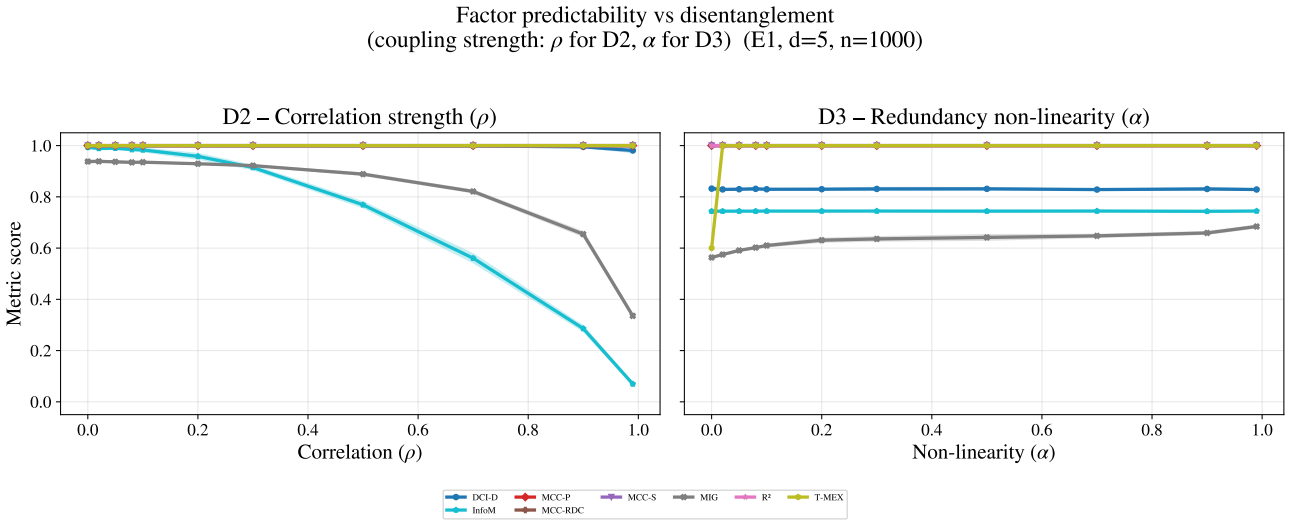

Figure 21: **Metrics conflate factor predictability with disentanglement under functional dependencies.** Full metric suite under **E1** comparing $\mathbf{D}_\rho$ (varying ρ) with $\mathbf{D}_f$ (deterministic constraint $z_2 = f(z_1)$). Under $\mathbf{D}_f$, regression-based metrics (R^2, DCI-D) penalise the encoder when the dependent factor is harder to predict from the retained code, despite perfect elementwise recovery. d=5, n=1000. See Fig. 22 for d=10.

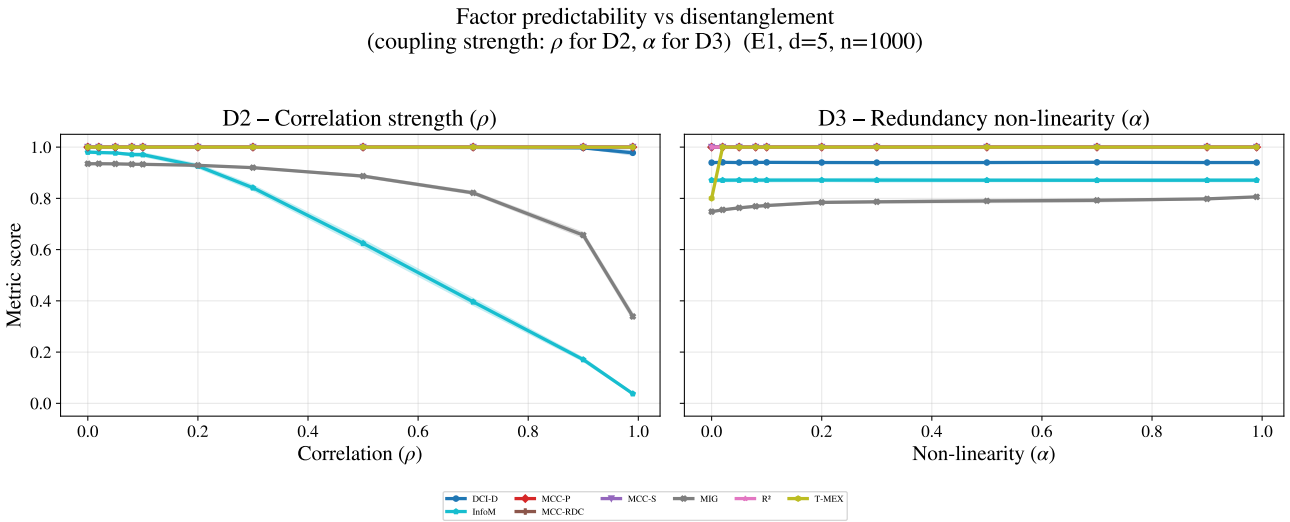

Figure 22: **Predictability vs. disentanglement at d=10.** Same setup as Fig. 21 with d=10 factors. The conflation between factor predictability and measured disentanglement persists at higher dimensionality.

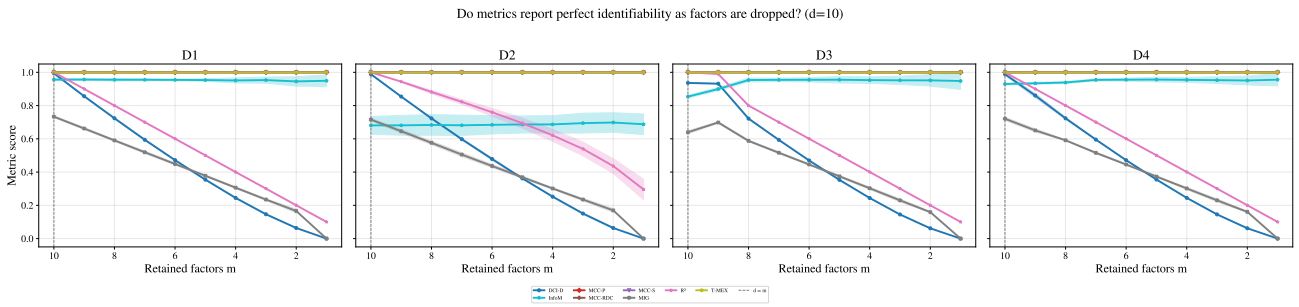

Figure 23: **Full metric suite for the dropped-factor experiment across all DGP types.** Extension of Fig. 17 with additional metrics (MIG, InfoMEC, MCC-RDC, T-MEX). MI-based metrics decline with fewer retained factors even under $\mathbf{D}_f$ at $m = d_{\text{eff}}$, failing to recognise lossless compression of the redundant factor.

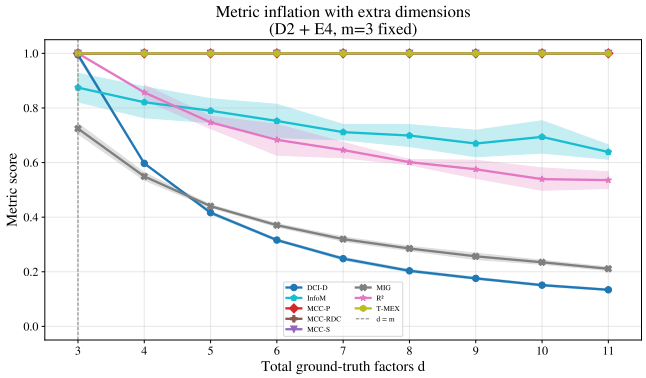

Figure 24: **Effect of inflating the number of ground-truth factors under $\mathbf{D}_\rho$.** The representation dimension is fixed at m=3 while d increases by adding duplicated ground-truth factors. MCC-P/S remain constant as they match only m codes; R^2 and DCI-D decline because the probe must predict an increasing number of factors from the same m codes.

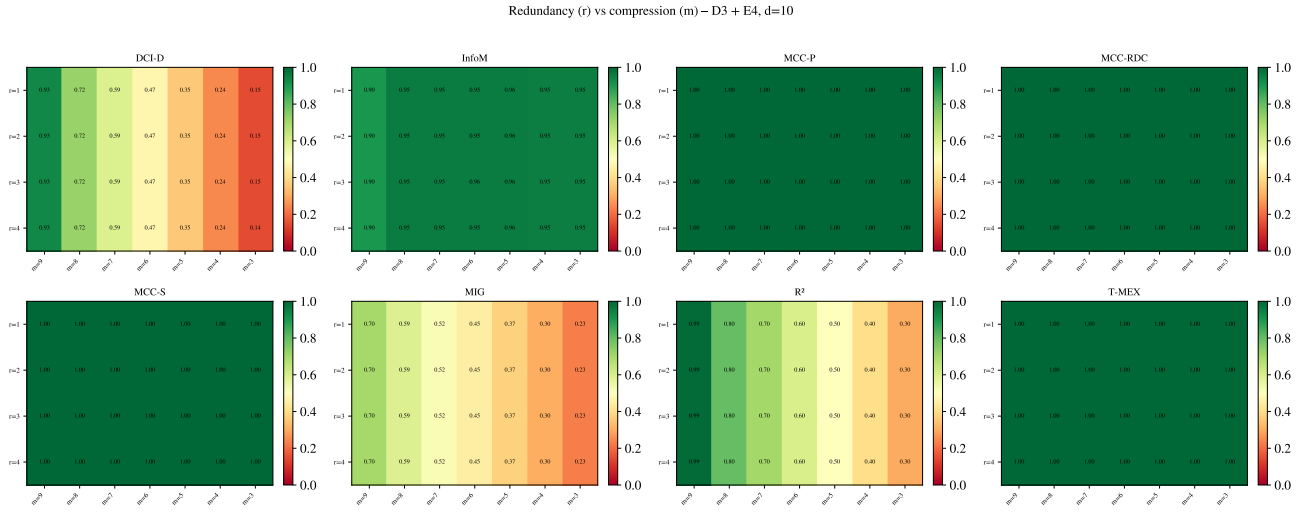

Figure 25: **Redundancy and compression under $\mathbf{D}_f$.** The encoder compresses d factors (with a single-factor constraint $z_2 = f(z_1)$, $d_{\text{eff}} = d-1$) into $m \leq d$ codes. R^2 and DCI-D plateau near 1.0 at $m = d_{\text{eff}}$, correctly recognising that the omitted factor carries no independent information. MCC-P/S report 1.0 at all compression levels, unable to distinguish lossless from lossy omission.

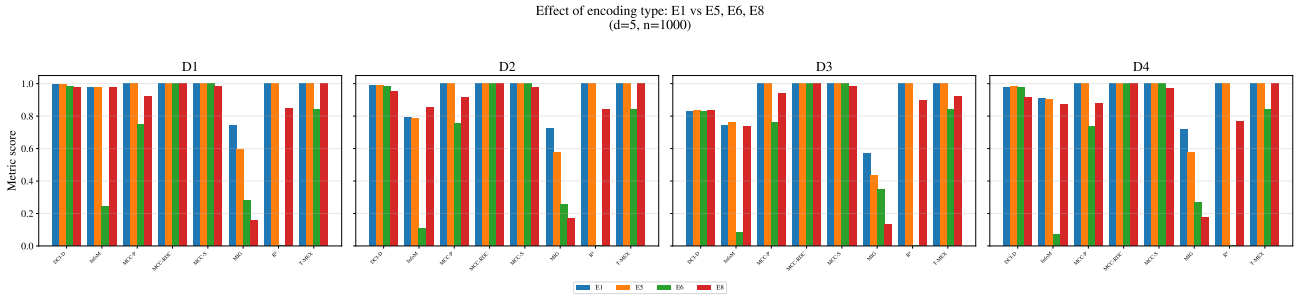

Figure 26: **Full metric suite: scores across encoding types and DGP types.** Each panel compares matched-dimension encoders (**E1**–**E3**) with overcomplete encoders (**E5**–**E8**) under $\mathbf{D}_{\perp}$–$\mathbf{D}_{F}$. MI-based metrics (MIG, InfoMEC) and T-MEX are included alongside the main metrics. d=5, n=1000.

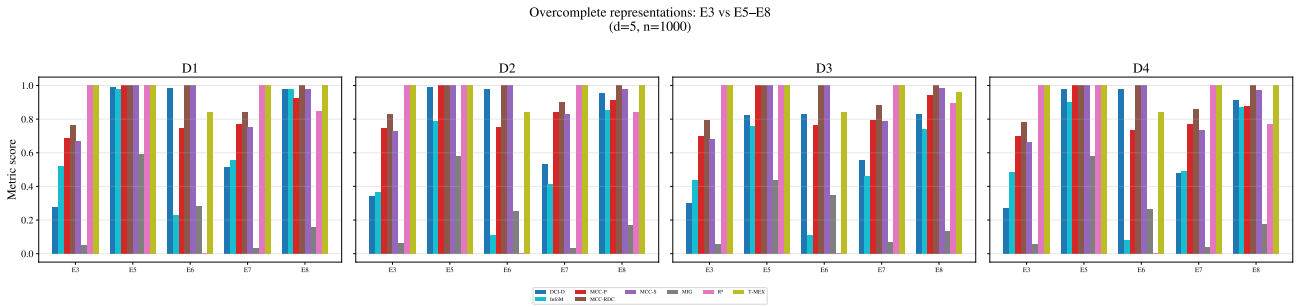

Figure 27: **Full metric suite for the overcomplete m/d sweep.** Extension of Fig. 4 to all metrics (MIG, InfoMEC, MCC-RDC, T-MEX). MI-based metrics decline for distributed codes (**E8**) similarly to MCC, while T-MEX is more robust to overcompleteness. d=5, n=1000.

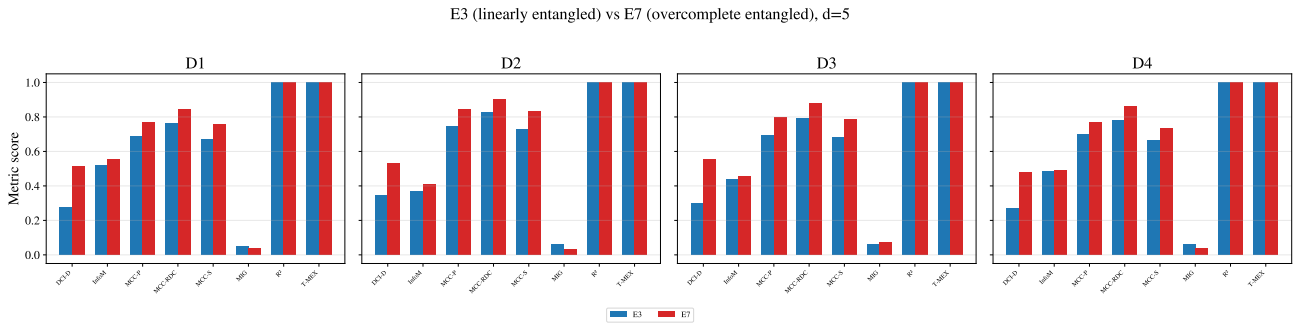

Figure 28: **Matched-dimension entangled (E3) versus overcomplete entangled (E7).** Full metric suite comparing the two entangled geometries as m/d increases. DCI-D inflates for **E7** at high m/d, scoring substantially above the matched-dimension baseline **E3** despite equivalent identifiability status. d=5, n=1000.

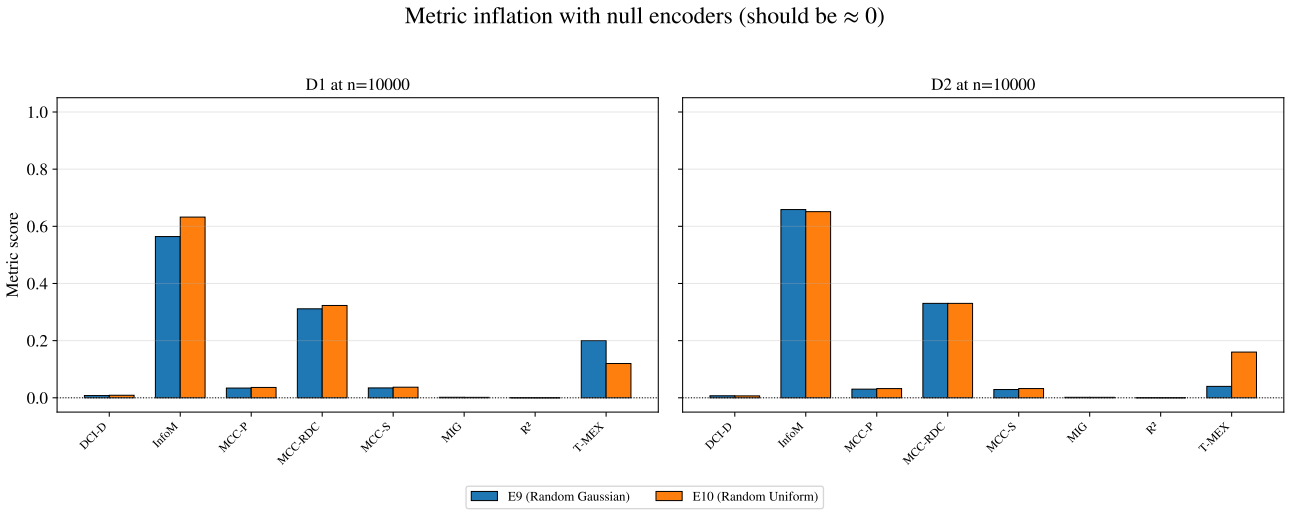

Figure 29: **Metric inflation under null encoders.** Full metric suite showing scores when the representation is independent of **z**. MCC-RDC exhibits persistent inflation that does not vanish with increasing n. MI-based metrics (MIG, InfoMEC) also return non-trivial scores. R^2 remains closest to the expected value of 0. d=5, n=1000.

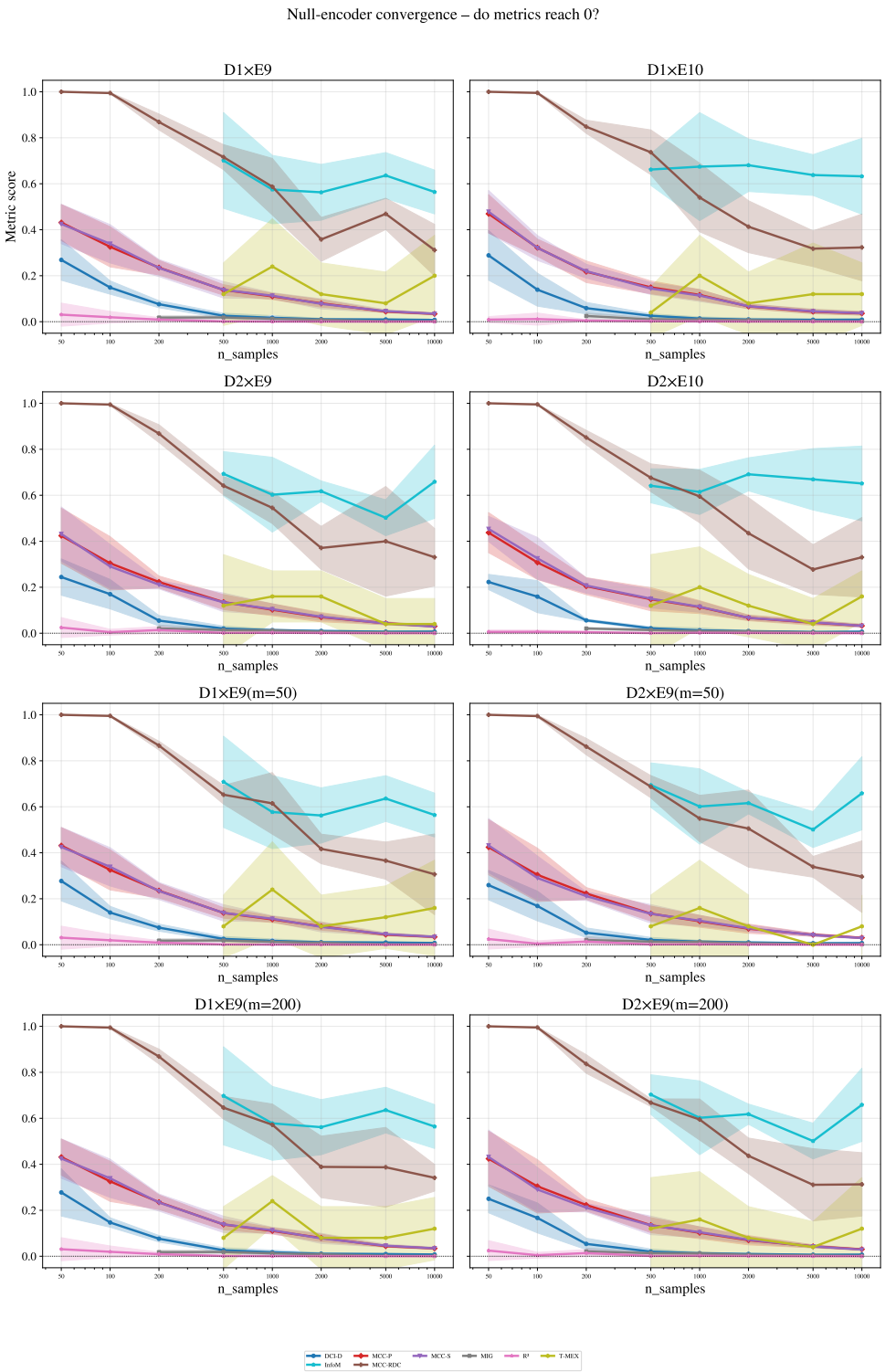

Figure 30: **Convergence of null-encoder scores with increasing n.** Each panel shows one metric under **E9**; scores should converge to 0 as n grows. R^2 converges fastest. MCC-P/S retain elevated scores at large n when m is large, consistent with the $\sqrt{2\log m/n}$ floor (§ F.3). d=5.

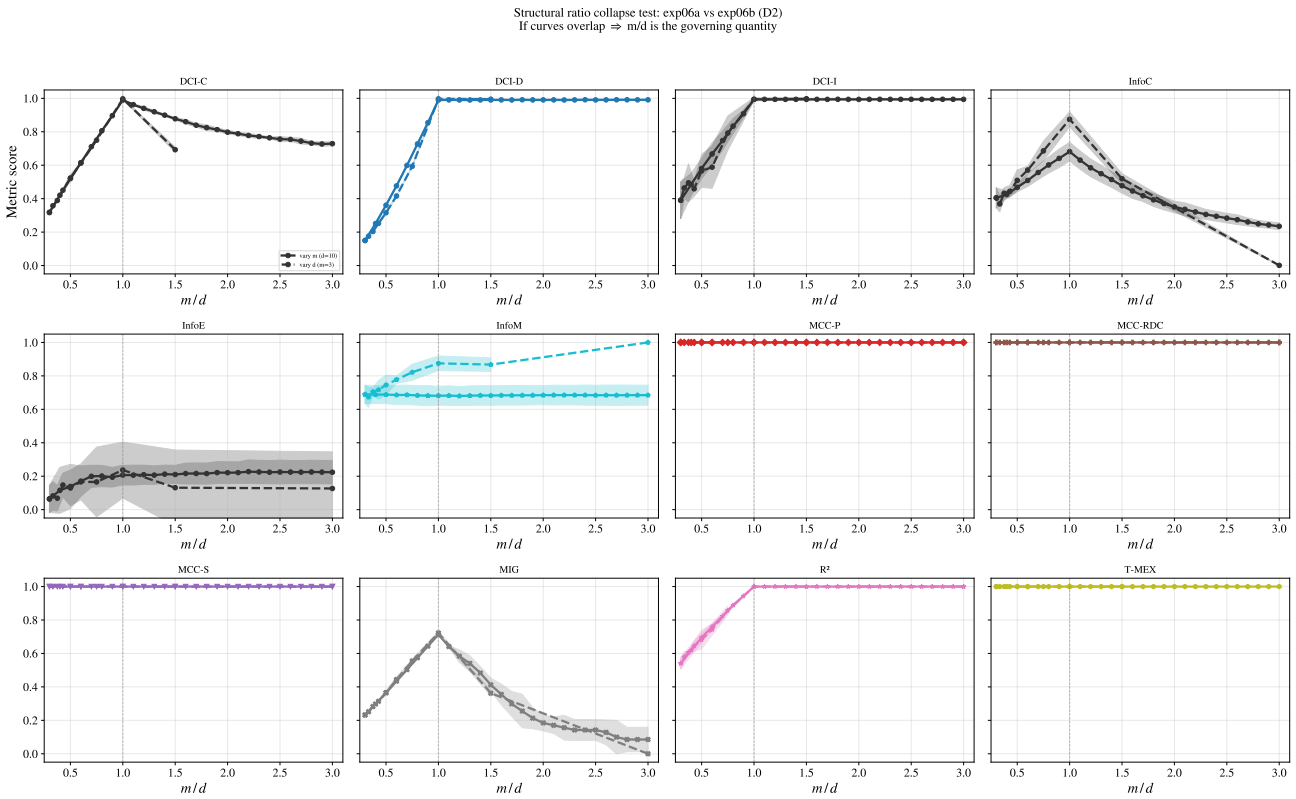

Figure 31: **Metric scores across encoder types under correlated factors ($\mathbf{D}_\rho$).** Overlay of all metrics for varying ρ under $\mathbf{D}_\rho$. MCC-P/S increase with $|\rho|$ under entangled encoders (**E3**), while R^2 and DCI-D are less affected by the correlation structure. d=5, n=1000.

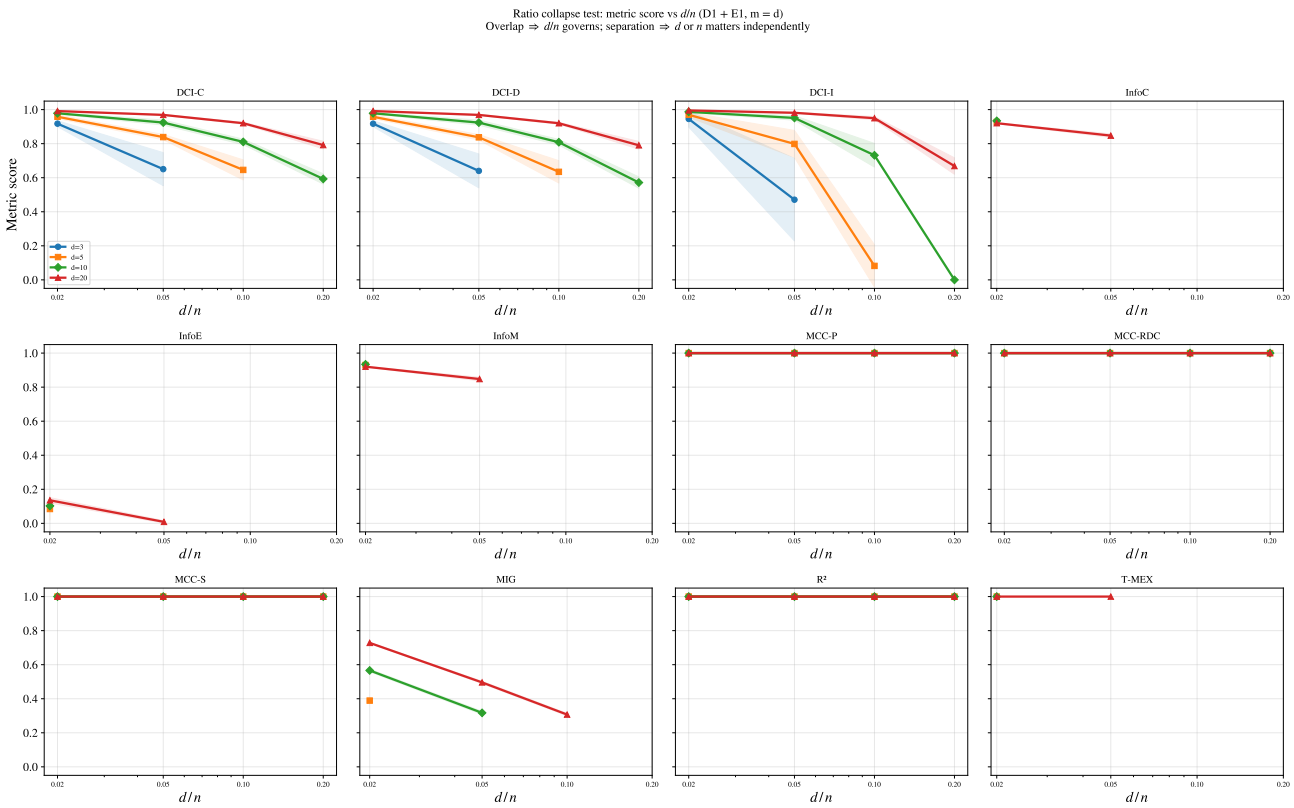

Figure 32: **Metric scores as a function of m/d under different (m, d) configurations.** Each panel shows one metric; overlapping curves from different (m, d) pairs with the same ratio confirm that m/d, not m or d individually, governs metric behaviour in the undercomplete regime. MCC-P/S report 1.0 regardless of m/d; R^2 and DCI-D increase approximately linearly. n=1000.

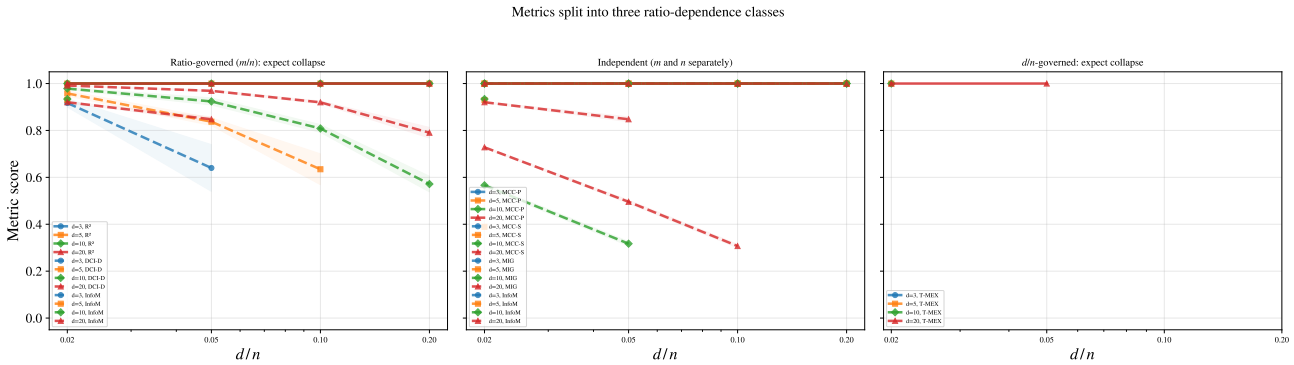

Figure 33: **Ratio-collapse analysis grouped by encoder type.** Same data as Fig. 32, reorganised by encoder geometry. Curves from sweeps over m (fixed d) and over d (fixed m) overlap at matched m/d, confirming the ratio as the governing quantity across encoder types.

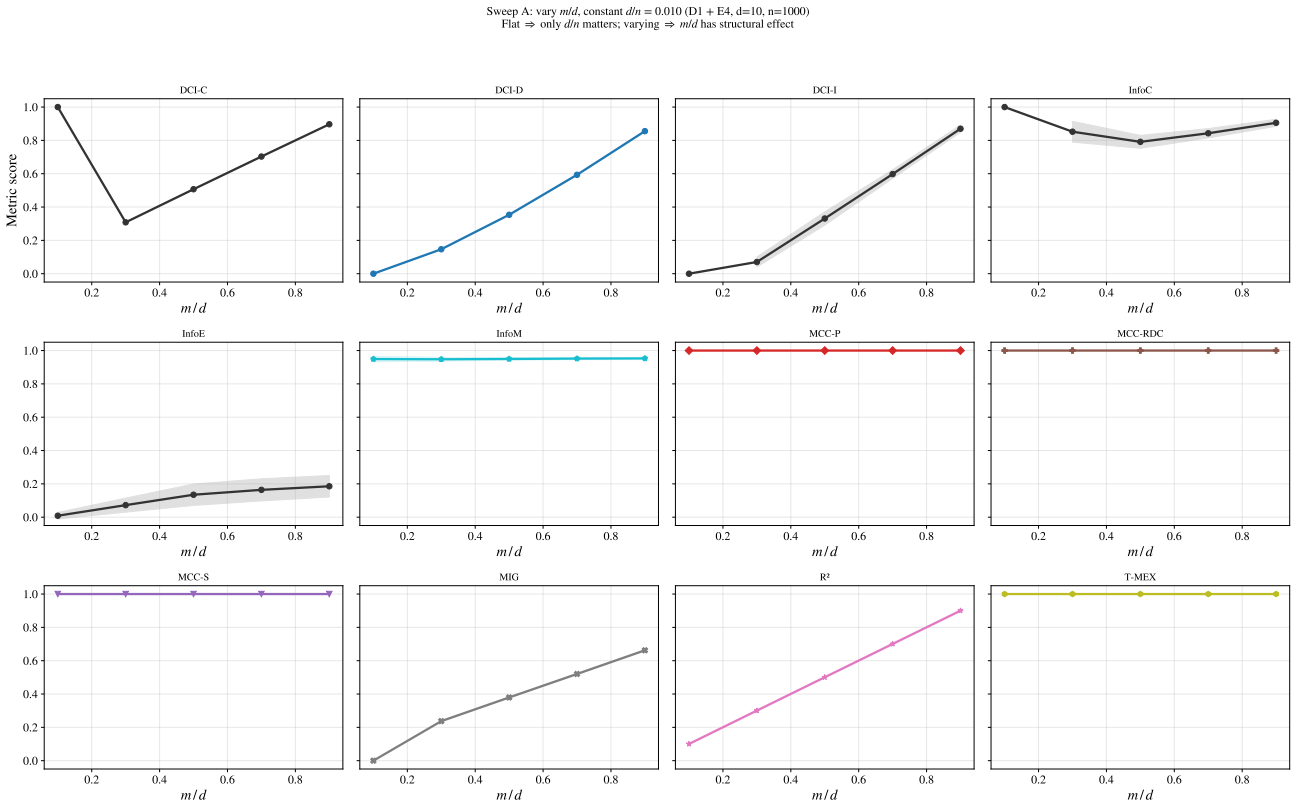

Figure 34: **Comprehensive parameter sweep (part A).** Full metric suite across DGP types and encoder geometries, sweeping scaling and complexity parameters. Extends the targeted analyses of § 3.1 to 3.4 to a broader parameter range. d=5, n=1000.

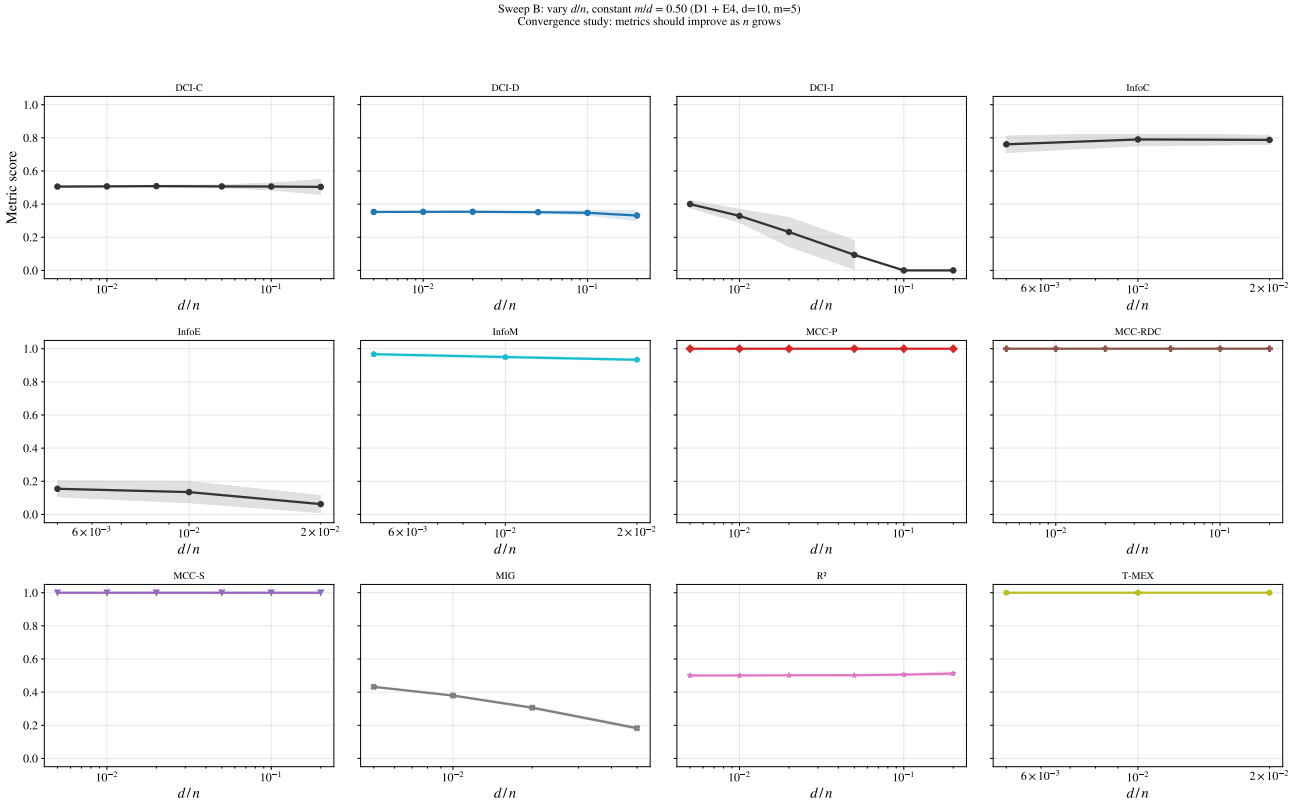

Figure 35: **Comprehensive parameter sweep (part B).** Continuation of Fig. 34 for additional parameter configurations and encoder–DGP combinations.

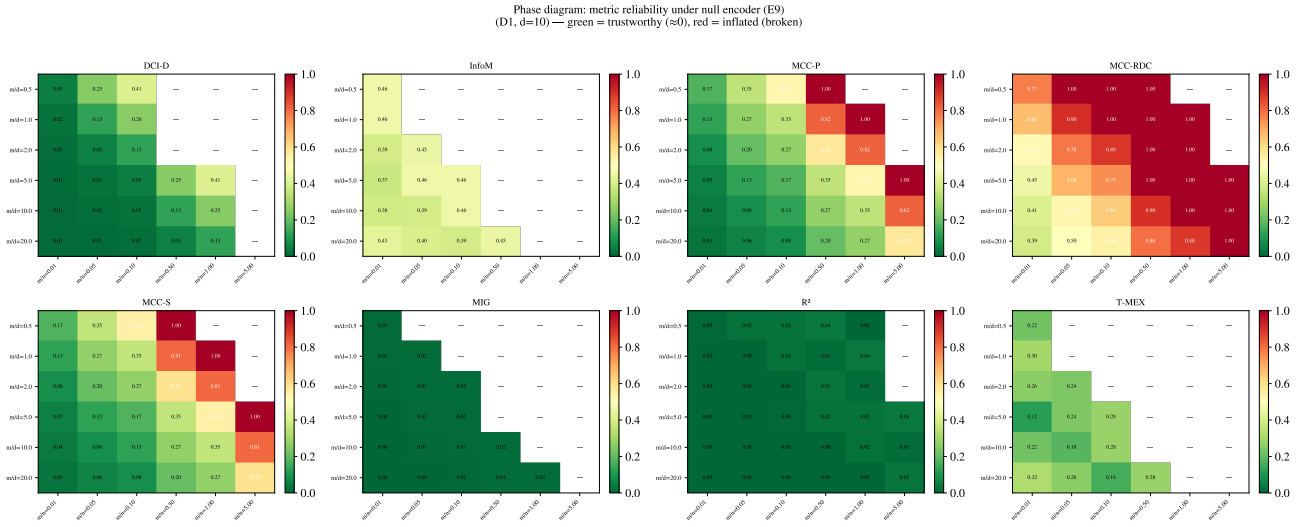

Figure 36: **Full metric suite: false-positive phase diagram under Gaussian null.** Extension of Fig. 18 to all metrics. MCC-RDC shows the highest inflation across the $(m/d,\ m/n)$ grid. MI-based metrics (MIG, InfoMEC) also inflate at moderate m/n.

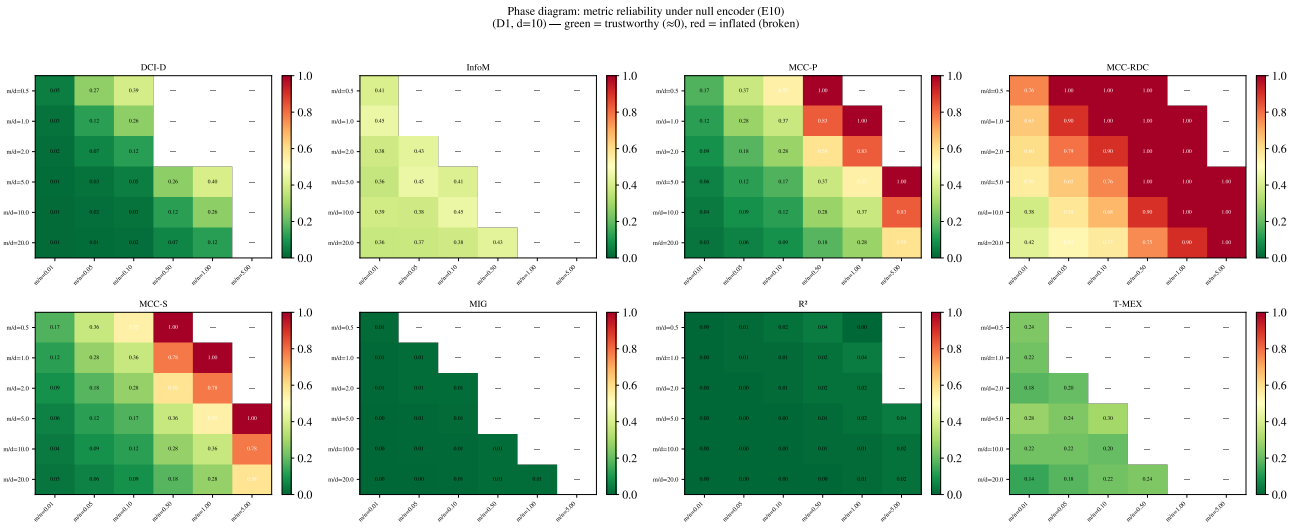

Figure 37: **Full metric suite: false-positive phase diagram under uniform null.** Extension of Fig. 5 to all metrics. The pattern closely mirrors the Gaussian null (Fig. 36), confirming that the false-positive floor is distribution-agnostic and governed by m/n.