INTERPRETABLE INTERACTION DICTIONARIES FOR MIXTURE RECOVERY IN DISCRETE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised representation learning typically maps observations into spaces where latent structure becomes easier to recover, but the resulting coordinates are often difficult to relate to observable properties of the data distribution. We study an alternative for discrete data in which representation coordinates are explicit binary interaction moments. Rather than seeking only a set of statistics sufficient for a downstream recovery task, we aim to construct a structured and human-interpretable representation whose higher-order coordinates remain traceable to simpler dependencies in the data. We instantiate this principle with ODIN, which uses an information-geometric criterion to select a nested hierarchy of interaction atoms, and apply the resulting representation to mixtures of product-Bernoulli distributions. Mixture parameters are estimated by constrained moment matching in the learned moment space. For any fixed dictionary, we show that full column rank of the lifted-moment Jacobian provides a local identifiability certificate, and that enriching the dictionary cannot decrease its rank. This separates the statistical role of dictionary construction from the geometric conditions governing recovery. We evaluate how the learned interaction representation affects parameter recovery and Jacobian geometry on synthetic mixtures, and assess its usefulness on real-world binary clustering benchmarks.

1 INTRODUCTION

A large part of unsupervised representation learning is built on a latent-generative assumption: observations are produced by hidden variables through an unknown process, and the goal is to recover useful information about this hidden structure from data alone. This viewpoint appears in causal representation learning (Zhang et al., 2024) and nonlinear independent component analysis (Hyvarinen et al., 2019; Lachapelle et al., 2022), where a central question is *identifiability*: whether latent variables or generative parameters can be recovered from the observed distribution. A common strategy is to map observations into a representation space in which latent structure becomes easier to estimate. Yet a representation can support recovery while remaining difficult to interpret: its coordinates need not correspond to observable properties of the data, and successful inversion alone does not explain which structures of the observed distribution are being used.

Dictionary learning offers one route toward more interpretable representations by decomposing observations or activations into elementary units, or *atoms*. This idea has attracted renewed attention in mechanistic interpretability (Lindsey et al., 2024; Ameisen et al., 2025), but interpretability is not automatically obtained from reconstruction. The review of Sharkey et al. (2025) highlights limitations of the dominant reconstructive paradigm, including its reliance on linear representation hypotheses and the absence of guarantees that learned atoms correspond to meaningful structure in the underlying generative process. Empirical studies similarly show that reconstruction-based dictionaries need not recover canonical latent concepts (Leask et al., 2025; Zhu et al., 2026), and that functional importance alone does not close this gap (Braun et al., 2024). These observations motivate a different question: *can we construct a representation that is useful for latent-variable recovery while retaining an explicit and human-readable relation to statistical structure in the observed data?*

This question is particularly concrete for discrete data, as encountered in electronic health records (Zheng et al., 2025; Pham et al., 2026), behavioral surveys (Sorgente et al., 2025), and tabular learning (Hollmann et al., 2025; Jiang et al., 2025; Gorishniy et al., 2021). For binary obser-

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

vations, interactions among variables provide natural observable coordinates of the data law (Amari, 2001; Sugiyama et al., 2016a; Enouen & Sugiyama, 2026). For a subset S of variables, the interaction atom$\phi_S(x) = \prod_{i \in S} x_i$ indicates their simultaneous activation, and its expectation $\eta_S = \mathbb{E}[\phi_S(X)]$ is exactly the probability of this co-activation event. We therefore depart from reconstruction-based binary and categorical dictionary learning (Henniges et al., 2010; Exarchakis & Lücke, 2017) and use interaction moments themselves as representation coordinates. Each selected coordinate has an explicit statistical meaning, while higher-order interactions can be organized hierarchically from simpler ones. This is important because our objective is not merely to obtain any collection of moments sufficient for a downstream inverse problem. Random or conditioning-driven moment subsets may also provide enough independent equations for recovery, but they need not provide a structured account of which dependencies in the data were selected and how higher-order coordinates relate to simpler ones. We instead seek a representation that remains inspectable independently of the downstream recovery algorithm.

We instantiate this principle for mixtures of product-Bernoulli (pB) distributions (Section 2), where the latent variable is the mixture component and the parameters of interest are the component weights and Bernoulli means. Our framework separates two questions that are often conflated: *which observable interaction moments should constitute the representation?* and *when does a given representation contain enough independent information to locally recover the latent parameters?* The resulting pipeline, summarized in Figure 1, has three stages. *(i) Dictionary construction (Section 3).* ODIN selects a nested hierarchy of interaction atoms from the empirical distribution using an information-geometric criterion inspired by Enouen & Sugiyama (2026). The resulting dictionary provides a finite, structured, and interpretable collection of co-activation statistics. *(ii) Lifting (Section 4).* Each observation is mapped to the coordinates defined by these atoms. Under a pB component, the expectation of every lifted coordinate is a monomial in the Bernoulli means, yielding a finite nonlinear moment representation of the mixture. *(iii) Mixture inversion (Section 5).* Mixture parameters are estimated by constrained moment matching. For any fixed dictionary, we characterize local recoverability through the Jacobian of the lifted moment map: full column rank provides a local identifiability certificate, and dictionary enrichment cannot decrease this rank. Importantly, this condition characterizes the information available in a selected representation; it is not assumed to be guaranteed by the ODIN selection criterion itself.

Our contribution is therefore not a claim that hierarchical interaction selection is necessary for mixture recovery. Rather, we study mixture recovery under the additional requirement that the representation remain statistically meaningful and human-interpretable. This separation also enables us to ask empirically whether information-guided dictionaries differ from alternative moment selections not only in rank, but also in conditioning, finite-sample recovery, and structural interpretability. Section 6 evaluates these questions on synthetic mixtures, where latent parameters are known, and on real-world binary clustering benchmarks.

2 Problem Setting and Previous Work

We begin from the general latent-generative hypothesis underlying many unsupervised learning problems. Observations are assumed to be generated by an unknown process governed by latent parameters, which we write abstractly as $X \sim g(\eta)$, where g denotes the generative process and η denotes the latent generative parameters. The goal is to recover information about η from observations of X, up to the natural non-identifiabilities of the model.

In this paper, we make this assumption concrete in the discrete setting. We take $X = (X_1, \ldots, X_d) \in \{0,1\}^d$ and choose $g(\eta)$ to be a mixture of m pB distributions. The latent variable is a component label $H \in [m]$, with mixing weights $\pi \in \Delta_{m-1}$. Conditionally on $H = h$, the coordinates of X are independent and satisfy $X_i \mid H = h \sim \text{Bernoulli}(\mu_{h,i})$, where $\mu_h = (\mu_{h,1}, \ldots, \mu_{h,d}) \in (0,1)^d$. Writing $\eta = ((\pi_h, \mu_h))_{h=1}^m$, the induced marginal law is

$$p_\eta(x) = \sum_{h=1}^{m} \pi_h \prod_{i=1}^{d} \mu_{h,i}^{x_i}(1-\mu_{h,i})^{1-x_i}, \qquad x \in \{0,1\}^d. \tag{1}$$

Thus, in the general notation above, the chosen generative process is $g(\eta) = p_\eta$. Appendix E discusses the scope of the product-Bernoulli assumption and interpretation under model misspecification.

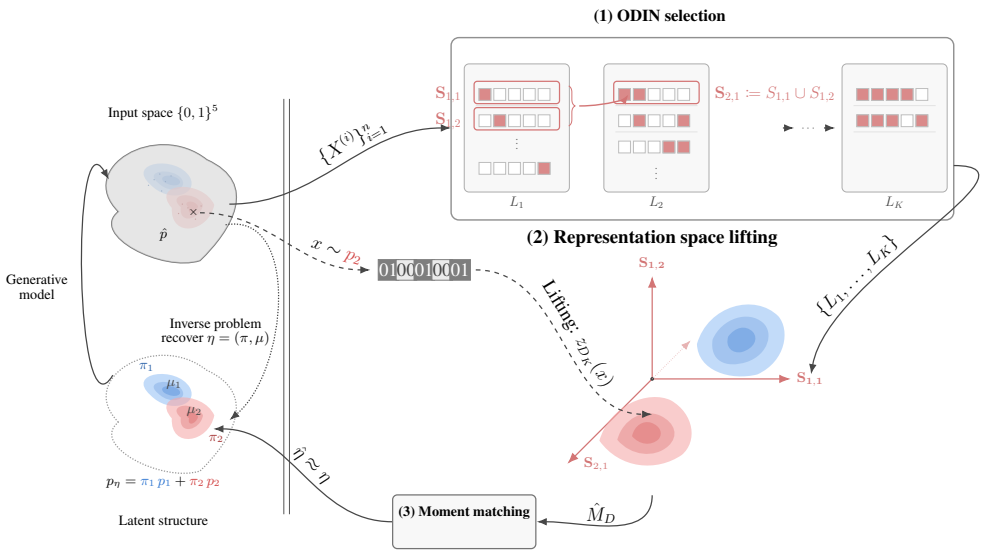


Figure 1: **ODIN as a learned finite lifting for mixture recovery.** The observations live in the discrete input space $\{0,1\}^d$ and are assumed to be generated from an unknown product-Bernoulli mixture law p_η, with latent parameters $\eta = (\pi, \mu)$. From samples $\{X^{(i)}\}_{i=1}^n$, we form the empirical distribution $\hat{p}$. ODIN uses $\hat{p}$ to perform nonparametric atom selection: it constructs a nested hierarchy of interaction dictionaries $D^{(1)} \subset \cdots \subset D^{(K)}$, where L_k denotes the layer of atoms added at level k. Each atom $S \subseteq [d]$ is an explicit co-activation pattern represented by the monomial $\phi_S(x) = \prod_{j \in S} x_j$; higher-order atoms are generated hierarchically by fusing previously selected atoms, as illustrated by $S_{1,1}, S_{1,2} \mapsto S_{2,1}$. The final dictionary $D = D^{(K)}$ defines the lifted representation $z_D(x) = (\phi_S(x))_{S \in D} \in \mathbb{R}^{|D|}$. From the lifted samples, we compute the empirical moment vector $\hat{M}_D = n^{-1} \sum_{i=1}^n z_D(X^{(i)})$, which is matched to the pB mixture moment map $\Xi_{D,m}(\pi, \mu) = \sum_{h=1}^m \pi_h \upsilon_D(\mu_h)$, with $\upsilon_D(\mu_h) = (\prod_{j \in S} \mu_{h,j})_{S \in D}$. The constrained moment-matching step estimates $\hat{\eta} = (\hat{\pi}, \hat{\mu})$. Thus ODIN does not recover the latent parameters directly; it selects the observable interaction coordinates on which the subsequent moment-matching problem estimates the mixture.

Given i.i.d. observations $X^{(1)}, \ldots, X^{(n)} \sim p_\eta$, the goal is to recover η from the observed distribution, or from its empirical approximation, subject to the unavoidable label ambiguity of mixture models. Namely, two parameter values $\eta = ((\pi_h, \mu_h))_{h=1}^m$ and $\eta' = ((\pi'_h, \mu'_h))_{h=1}^m$ are regarded as equivalent if there exists a permutation $\sigma \in \mathfrak{S}_m$ such that $\pi'_h = \pi_{\sigma(h)}$ and $\mu'_h = \mu_{\sigma(h)}$ for all $h \in [m]$. Thus, identifiability in this work means identifiability of the mixture parameters up to permutation of the component labels.

Interaction selection and information geometry. The need to choose informative interaction moments connects our problem to feature-interaction and mode-interaction selection. Classical feature selection removes irrelevant variables, whereas feature-interaction selection asks which subsets of variables should be allowed to interact in the model (Lyu et al., 2023). For distribution learning over discrete variables, this becomes a structure learning problem for hierarchical log-linear models: instead of selecting edges in a pairwise graph, one selects higher-order interaction sets in a hypergraph. This perspective is developed explicitly in recent information-geometric works (Amari, 2001; Sugiyama et al., 2016b; 2018; Ghalamkari et al., 2023). Building on this line, Enouen & Sugiyama (2026) introduce refined information, a nonnegative decomposition of KL error along a hierarchy of mode interactions, and propose MAHGenTa as an algorithm for selecting higher-order interactions in visible log-linear distribution learning. In MAHGenTa, the selected interactions define the final visible energy-based model. In this paper, by contrast, the selected interactions define observable coordinates for a downstream inverse problem. We use the dictionary-learning version ODIN (**O**rder-wise **D**ictionary learning by **I**nformation decompositio**N**), a nonparametric hierarchical

procedure that shares the same information-geometric principle of selecting interactions by their KL contribution, but uses the selected atoms as a finite dictionary of moments rather than as the final density model. In this way, ODIN determines which interaction moments are made available for product-Bernoulli mixture recovery.

Previous work on finite mixture recovery. Product-Bernoulli mixtures are the binary locally independent case of latent class analysis (LCA), a latent-variable mixture framework in which an unobserved categorical variable partitions a heterogeneous population into latent classes, and observed indicators are conditionally independent within each class (Sorgente et al., 2025). Finite mixture recovery is usually studied through likelihood methods, such as EM (Dempster et al., 1977), or through observable moments. Tensor and moment-based approaches give generic identifiability and spectral recovery guarantees under conditional-independence and rank assumptions (Allman et al., 2009; Anandkumar et al., 2014). These guarantees depend on the chosen observables: in discrete settings, classical linear-independence conditions may fail for conditional i.i.d. models and Bernoulli mixtures, motivating alternative diversity-based criteria (Lyu & Yang, 2025). Recoverability therefore depends not only on the mixture family, but also on the observable representation. Our framework takes this observation as a starting point: rather than fixing a small set of low-order moments, we let an information-geometric dictionary select which interaction moments enter the observable representation, and tie local identifiability to the rank of the resulting Jacobian.

Lifted observable representations. The dependence on observables motivates a lift-then-recover perspective. Rather than recovering parameters from raw coordinates or fixed low-order moments, one may transform observations through a feature map z and study the induced map $\eta \mapsto \mathbb{E}_\eta[z(X)]$. The choice of z determines which aspects of the distribution are visible to the inverse problem. Kernel mean embeddings provide one formal instance: characteristic kernels embed probability measures injectively into an RKHS (Sriperumbudur et al., 2011), and Song et al. (2014) use such embeddings for nonparametric spectral recovery in multi-view latent-variable models. More broadly, modern neural methods also rely on learned embeddings or latent spaces in which inference becomes more tractable (Wang & Wang, 2024; Varici et al., 2024; Lu et al., 2024). These liftings are typically infinite-dimensional, implicit, or learned end-to-end, which makes them effective for recovery but opaque as representations: one cannot read off which structures of the data law support identifiability, nor inspect the coordinates as statistics. Our work follows the same lifting principle, but replaces these feature spaces by a finite, explicit dictionary of observable binary interactions, so that each lifted coordinate is itself an interpretable moment of the data law and identifiability can be tied to a finite, analyzable Jacobian.

3 ODIN AS HIERARCHICAL INTERACTION DICTIONARY LEARNING

How should the finite interaction dictionary D be chosen? The full interaction lattice $2^{[d]}$ is exponentially large; taking all monomials up to a fixed order is typically infeasible and introduces many irrelevant moments; and a fixed or random subset ignores the dependency structure of the data. Since the selected atoms will serve as the observable moment coordinates through which the mixture is recovered, this choice is central to the method.

We use ODIN as a data-driven construction. For binary observations, ODIN represents an interaction pattern $S \subseteq [d]$ by the monomial atom $\phi_S(x) = \prod_{i \in S} x_i$, an explicit observable event: the simultaneous activation of the variables indexed by S. ODIN builds D hierarchically, adding interactions that explain residual dependence left by the atoms already chosen.

Given an empirical distribution $\hat{p}$, a dictionary $D \subseteq 2^{[d]}$ induces the exponential-family model

$$\mathcal{M}_D = \left\{ q(\cdot;\theta,D) : q(x;\theta,D) = \exp\left(\sum_{S \in D} \theta_S \phi_S(x) - \psi(\theta,D) \right) \right\}.$$

The residual dependency left unexplained by D is $\mathcal{L}(D) := \inf_{q \in \mathcal{M}_D} \mathrm{KL}(\hat{p} \| q)$. For a candidate atom $S \notin D$, ODIN scores the gain by the refined information (Enouen & Sugiyama, 2026) $\mathrm{RI}_{D,S}(\hat{p}) := \mathcal{L}(D) - \mathcal{L}(D \cup \{S\})$, namely the decrease in residual KL obtained by adding S.

The hierarchy is indexed by interaction order. ODIN starts from the unary layer $L_1 := \{\{1\}, \ldots, \{d\}\}$ with $D^{(1)} := L_1$. For $k \geq 2$, order-k candidates are formed by combining overlapping atoms from

the previous layer: $\mathcal{C}_k(L_{k-1}) := \Big\{ S_1 \cup S_2 : S_1, S_2 \in L_{k-1},\ |S_1 \cup S_2| = k,\ |S_1 \cap S_2| = k-2 \Big\}$. This rule enforces a heredity constraint: an order-k atom is reachable only if it can be assembled from selected order-$(k-1)$ atoms. After empirical support and information-based pruning, the remaining admissible set at D is $\mathcal{A}_k(D) \subseteq \mathcal{C}_k(L_{k-1})$. At each order $k \geq 2$, ODIN greedily adds the admissible atom with largest refined information, stopping when no admissible atom remains or when the best gain falls below a threshold τ. If L_k denotes the atoms selected at order k, this procedure yields a nested hierarchy $D^{(1)} \subseteq D^{(2)} \subseteq \cdots \subseteq D^{(K)}$, where K is the maximum interaction order. Algorithmic details, including the precise updates, are given in Appendix A. Here ODIN is not the final density model: it selects the finite monomial dictionary used for lifted moment matching. Algorithmic details and diagnostics are given in Appendices A and H.

4 ODIN INDUCES A FINITE LIFTED MOMENT GEOMETRY

Once ODIN has selected a dictionary D, its atoms define the coordinates of a finite representation space. Because the goal is interpretability, these coordinates are not random directions or black-box features, but explicit observable quantities. We therefore use the monomial lifting $z_D(x) := (\phi_S(x))_{S \in D} = \big(\prod_{i \in S} x_i\big)_{S \in D} \in \{0,1\}^{|D|}$. Each coordinate records the activation of a selected interaction event. The representation space makes these interactions explicit coordinates, rather than leaving them as implicit nonlinear functions of the raw binary input.

Conditional on $H = h$, the raw variables are independent Bernoulli variables with mean vector $\mu_h \in [0,1]^d$. Then, for every $S \in D$, $\mathbb{E}[\phi_S(X) \mid H = h] = \mathbb{E}\left[\prod_{i \in S} X_i \,\middle|\, H = h\right] = \prod_{i \in S} \mu_{h,i}$. This motivates the component embedding $v_D(\mu) := \big(\prod_{i \in S} \mu_i\big)_{S \in D} \in [0,1]^{|D|}$. Consequently, $\mathbb{E}[z_D(X) \mid H = h] = v_D(\mu_h)$. Since $z_D(X) \in \{0,1\}^{|D|}$, the conditional law $\mathbb{P}(z_D(X) \in \cdot \mid H = h)$ is a discrete distribution on the hypercube with first moment $v_D(\mu_h)$; this center lies in the convex hull $[0,1]^{|D|}$. Under the lifting, each component induces its own conditional law on $\{0,1\}^{|D|}$, centered at its own embedding: a draw from component h fluctuates around $v_D(\mu_h)$. The unobservable latent component $\{H = h\}_{h=1}^m$ thus becomes a geometric organization in $[0,1]^{|D|}$ with centers $\{v_D(\mu_h)\}_{h=1}^m$, each surrounded by a cloud of lifted observations from its own component. The marginal first moment of the lifted observation is the convex combination $\mathbb{E}[z_D(X)] = \sum_{h=1}^m \pi_h\, v_D(\mu_h), \qquad \pi_h := \mathbb{P}(H = h)$. Identifying the components from observable moments thus reduces to recovering the centers $\{v_D(\mu_h)\}_{h=1}^m \subseteq [0,1]^{|D|}$ and the weights π from $\mathbb{E}[z_D(X)]$. Whether this is possible depends on the geometry of the centers, which the next proposition decomposes atom by atom.

Proposition 1 (Atom-wise decomposition of lifted component separation)**.** *Let $D \subseteq 2^{[d]}$ be a finite interaction dictionary. Then for every pair of component parameters $\mu_h, \mu_{h'} \in [0,1]^d$,*

$$\|v_D(\mu_h) - v_D(\mu_{h'})\|_2^2 = \sum_{S \in D} \Delta_S(\mu_h, \mu_{h'}), \qquad \Delta_S(\mu_h, \mu_{h'}) := \Big(\prod_{i \in S} \mu_{h,i} - \prod_{i \in S} \mu_{h',i}\Big)^2.$$

Consequently, if $D \subseteq D'$, the gain in lifted separation obtained by adding the atoms $D' \setminus D$ is exactly

$$\|v_{D'}(\mu_h) - v_{D'}(\mu_{h'})\|_2^2 - \|v_D(\mu_h) - v_D(\mu_{h'})\|_2^2 = \sum_{S \in D' \setminus D} \Delta_S(\mu_h, \mu_{h'}).$$

The proof is given in Appendix B. Proposition 1 yields a clean separation criterion: atom S contributes to the lifted separation of a component pair $(\mu_h, \mu_{h'})$ if and only if $\prod_{i \in S} \mu_{h,i} \neq \prod_{i \in S} \mu_{h',i}$, and atoms with large $\Delta_S(\mu_h, \mu_{h'})$ identify the interactions whose joint activation probabilities most distinguish the two components. The decomposition is also monotone in D: adding atoms can only enlarge the lifted separation, by exactly $\sum_{S \in D' \setminus D} \Delta_S(\mu_h, \mu_{h'})$.

Observable data give only the convex combination $\mathbb{E}_\eta[z_D(X)] = \sum_{h=1}^m \pi_h\, v_D(\mu_h)$, not the individual embeddings $v_D(\mu_h)$, and distinct mixture parameters η, η' can produce the same observable moment even when their components are pairwise separated. Component separation is therefore necessary but not sufficient for recovering η from observable moments; the missing ingredient is the conditioning of the lifted moment map.

Accordingly, define the lifted moment map $\Xi_{D,m}(\eta) := \sum_{h=1}^{m} \pi_h \, v_D(\mu_h)$, so that mixture recovery is the inversion of $\eta \mapsto \Xi_{D,m}(\eta)$. Empirically, with $\hat{M}_D := n^{-1} \sum_{r=1}^{n} z_D(X^{(r)})$, recovery seeks η such that $\Xi_{D,m}(\eta) \approx \hat{M}_D$. ODIN's role in the pipeline is now precisely specified: it selects the observable interaction moments through which this inversion is attempted. Solving this inversion is the subject of the next section.

5 ODIN-TO-MIXTURE INVERSION AS A CONSTRAINED OPTIMIZATION PROBLEM

Given observations $X^{(1)}, \ldots, X^{(n)}$, let $\hat{M}_D := n^{-1} \sum_{t=1}^{n} z_D(X^{(t)})$ be the empirical lifted moment. With $\Theta_m := \{(\pi, \mu) : \pi \in \Delta_{m-1},\ \mu_{h,i} \in [0,1]\}$ and $F_D(\eta) := \|\hat{M}_D - \Xi_{D,m}(\eta)\|_2^2$, the ODIN-to-mixture estimator is any solution of

$$\hat{\eta} \in \arg\min_{\eta \in \Theta_m} F_D(\eta). \tag{2}$$

Compactness of Θ_m and continuity of F_D ensure existence. The coordinate expansion and connection to classical moment matching are given in Appendix C.1. In the well-specified case, the data are generated by some $\eta^\star \in \Theta_m^\circ := \{(\pi, \mu) : \pi_h > 0,\ \sum_h \pi_h = 1,\ \mu_{h,i} \in (0,1)\}$, and $M_D := \mathbb{E}_{\eta^\star}[z_D(X)] = \Xi_{D,m}(\eta^\star)$. For finite n, $\hat{M}_D$ need not lie in $\Xi_{D,m}(\Theta_m)$, so $\hat{\eta}$ is best read as a finite-sample least-squares approximation to the population inverse problem. Boundary parameters are excluded because they are structurally degenerate: if $\pi_h = 0$, then μ_h is inactive and cannot be identified. The objective is non-convex, and full-column-rank local identifiability does not imply that numerical optimization reaches the corresponding solution. In practice, we use multiple random initializations and retain the solution with smallest moment-matching loss; this is an optimization heuristic rather than a global recovery guarantee.

Local identifiability via the Jacobian. Eliminate the simplex constraint via $\pi_m = 1 - \sum_{h<m} \pi_h$, and let $\mathcal{J}_D(\eta^\star) \in \mathbb{R}^{|D| \times ((m-1)+md)}$ denote the Jacobian of $\Xi_{D,m}$ at $\eta^\star$ in the free coordinates $(\pi_1, \ldots, \pi_{m-1}, \mu_1, \ldots, \mu_m)$. The columns are $v_h^\star - v_m^\star$ for $h < m$ in the weight directions, and $\pi_h^\star J v_D(\mu_h^\star)$ for $h \in [m]$ in the component directions, with $v_h^\star := v_D(\mu_h^\star)$. At a population exact solution, the residual vanishes and the Hessian of the least-squares objective reduces to $2\,\mathcal{J}_D(\eta^\star)^\top \mathcal{J}_D(\eta^\star)$. Thus local isolation is controlled by injectivity of the lifted Jacobian $\mathcal{J}_D(\eta^\star)$.

Proposition 2 (Local identifiability from the lifted Jacobian)**.** *Let $\eta^\star \in \Theta_m^\circ$ satisfy $\Xi_{D,m}(\eta^\star) = M_D$. If*

$$\operatorname{rank} \mathcal{J}_D(\eta^\star) = (m-1) + md,$$

then $\eta^\star$ is locally identifiable from the exact lifted moments, up to permutation of mixture labels.

The proof is given in Appendix C.

Role of the dictionary. Each atom $S \in D$ contributes one lifted moment coordinate and therefore one row of $\mathcal{J}_D(\eta^\star)$. Hence D determines the differential information available for local inversion. Three consequences follow. First, a necessary dimension condition is $|D| \geq (m-1)+md$. Singleton moments alone provide only d rows and cannot satisfy the rank condition for $m \geq 2$, so higher-order atoms are needed not only for interpretability but also to supply the missing observable equations.

Second, as shown in Appendix C.4, the smallest singular value $\sigma_{\min}(\mathcal{J}_D(\eta^\star))$ controls error propagation: for $\hat{\eta}$ close to $\eta^\star$,

$$\|\hat{\eta} - \eta^\star\|_2 \lesssim \sigma_{\min}(\mathcal{J}_D(\eta^\star))^{-1} \|\hat{M}_D - M_D\|_2.$$

Since $\|\hat{M}_D - M_D\|_2 = O_P(\sqrt{|D|/n})$ for bounded lifted coordinates, D affects recovery through both the number of selected moments and the conditioning of the lifted Jacobian (Appendix C.4). Third, dictionary enrichment only adds rows to the Jacobian, so rank cannot decrease and full rank persists under enlargement of D (Appendix D). A concrete example showing how dictionary choice changes the rank condition is given in Appendix C.5.

6 EXPERIMENTS

We evaluate ODIN on synthetic product-Bernoulli mixtures with known components and on real-world binary datasets, where recovery is assessed through its clustering interpretation.

6.1 SYNTHETIC EXPERIMENTS

Setup. We generate four independent datasets at each sample size from the product-Bernoulli mixture in equation 1, using $m = 4$ components in dimension $d = 6$. The parameters form two nearby pairs, (μ_1, μ_2) and (μ_3, μ_4), making component recovery deliberately ill-conditioned. Although $d < 2m - 1$, this particular population mixture is identifiable up to label permutation through a full-rank three-view grouping; see Appendix J.2.

We compare five estimators on the same samples: **Song-δ** and **Song-RBF**, the multiview kernel spectral method of Song et al. (2014) with delta and RBF kernels; **VaDE** (Jiang et al., 2016); **EM** for product-Bernoulli mixtures; and **ODIN**, which learns a dictionary and solves the lifted moment-matching problem in equation 2. Dataset and implementation details are given in Appendix F; optimization diagnostics and the multiview ablation are reported in Appendices J.1 and J.4.



Figure 2: **Lifted geometry and layer-wise recovery. Left:** separation of the nearby component pairs in cumulative ODIN representations. **Right:** layer-wise recovery error; the dashed line marks the first layer at which the lifted Jacobian has full column rank.

Recovery. Following Song et al. (2014), we measure component recovery by the permutation-aligned weighted error $\mathrm{Err} = \min_{\sigma \in \mathfrak{S}_m} \sum_{h=1}^{m} \pi_h^\star \left\| p_h^\star - p_{\sigma(h)}^{\mathrm{est}} \right\|_2^2$. Figure 5 in Appendix F.1 reports this error across sample sizes. ODIN improves steadily with n and remains competitive with EM, whereas Song-δ, Song-RBF, and VaDE have larger errors over the tested sample sizes. In particular, ODIN outperforms VaDE, indicating that a finite interaction lift is better adapted to this discrete mixture-recovery problem than a generic deep latent embedding.

Lifted geometry and rank. For each cumulative dictionary D_k, we compute $\Delta_{h,\ell}^{(k)} = \|v_{D_k}(\mu_h) - v_{D_k}(\mu_\ell)\|_2$. The left panels of Figure 2 show that the nearby pairs remain the hardest to distinguish, while higher-order atoms amplify their observable interaction differences.

The parameter dimension is $(m-1) + md = 27$. Consequently, D_1 and D_2, with 6 and 21 moments respectively, cannot satisfy the local rank condition. At layer 3, the ODIN dictionary first reaches full column rank, coinciding with the observed decrease in recovery error shown in the right panel. This is consistent with Proposition 2, which characterizes local recoverability once a full-rank dictionary has been obtained; it does not imply that the ODIN selection criterion is guaranteed to produce such a dictionary.

6.2 REAL-WORLD DATA: CLUSTERING APPLICATIONS

Setup. On real datasets, the true mixture parameters and latent components are unknown, so the synthetic recovery error is unavailable. We therefore evaluate the fitted mixture through its clustering interpretation: after estimating the product-Bernoulli mixture from ODIN-selected moments, each observation is assigned to the component with largest posterior probability. External labels are used only for evaluation. We use three UCI binary-label benchmarks and set $m = 2$ in the main

comparison: Congressional Voting Records (*vote*) (con, 1987), Mushroom (mus, 1981), and SPECT Heart (*spect*) (Cios et al., 2001). The *vote* dataset contains 16 binary congressional votes and has a clear two-party structure. The *mushroom* dataset is categorical and is converted to a 116-dimensional one-hot binary representation; see Appendix F.3. The *spect* dataset contains 22 binary image-derived cardiac features. For *mushroom* and *spect*, the binary labels may collapse several latent subgroups, so label agreement is not the only meaningful notion of recovery. See full description in Appendix F.

Baselines and metrics. We compare ODIN with EM, NeuNcut (He et al., 2025), DipEncoder (Leiber et al., 2022), and VaDE (Jiang et al., 2017). We report ACC, NMI, and ARI with respect to the external labels. We also report a label-free post-hoc KL score: for each method, the induced partition is converted into an input-space product-Bernoulli mixture using cluster proportions and within-cluster Bernoulli means, and we compute $\mathrm{KL}(\hat{p}\|p^{\mathrm{est}})$, where $\hat{p}$ is the empirical distribution of the binary observations. This measures how well the induced mixture explains the observed binary distribution under a common product-Bernoulli parametrization. Full preprocessing, hyperparameters, and KL details are given in Appendix G.

6.2.1 Numerical clustering results

Clustering performance. Table 1 reports the main real-data results. ODIN ties the best ACC on *vote* and obtains the best NMI. On *mushroom*, it achieves the strongest agreement with the edible/poisonous labels. On *spect*, ODIN obtains the best ACC, whereas EM gives the best NMI and KL. These differences show that agreement with the external labels and distributional fit can rank methods differently. Full results with ARI and standard deviations are reported in Appendix J.

Representation geometry. Figure 3 gives a qualitative view of the learned representations. On *vote*, the ODIN layers progressively organize the data in a way that visually aligns with the external binary labels. On *mushroom* and *spect*, separated regions are visible but need not coincide with the labels, consistent with the possibility of finer latent subgroups. This motivates reporting both label-based metrics and the label-free KL score. Additional representation diagnostics are given in Appendix I.

Alternative structure and interpretability. We also fit ODIN mixtures with larger values of m to examine finer partitions of the same lifted representation. Figure 4 shows the *spect* representation for $m \in \{2, 4, 10\}$. Because ODIN coordinates are explicit co-activation atoms, recovered components can be inspected through enriched interaction patterns. The callouts illustrate representative patterns over named SPECT regions. These are statistical descriptions of the fitted components, not clinical claims; atom-enrichment scores are detailed in Appendix K.

7 Discussion

We studied mixture recovery through a finite representation whose coordinates remain explicitly tied to observable structure in the data. ODIN constructs this representation from binary interaction moments rather than from opaque latent features: each selected atom is a co-activation event, and the hierarchical construction records how higher-order interactions are built from simpler ones. The role of ODIN is therefore representation construction, while recoverability is analyzed separately for the resulting fixed dictionary through the lifted moment map and its Jacobian. This distinction is important. Full column rank of the lifted Jacobian provides a local identifiability certificate, but the ODIN selection rule is not guaranteed to produce such a dictionary. Dictionary enrichment can only increase rank, and the synthetic experiments show that higher-order interaction moments can supply the additional coordinates needed to reach full rank. However, matched-budget random dictionaries may also reach full rank. The value of ODIN is therefore not that hierarchical selection is uniquely necessary for recovery, but that it restricts an exponentially large interaction search space to a structured, data-dependent, and human-interpretable family of moments. In our ablation, ODIN also lies in a comparatively well-conditioned region of the matched random ensemble, although information-based selection is not explicitly optimized for Jacobian conditioning or parameter recovery. The recovery stage has its own limitations. The moment-matching objective is non-convex, and local identifiability does not imply that numerical optimization reaches the corresponding solution. We therefore rely on multiple random restarts and select the minimum-loss solution; restart diagnostics show that distinct estimates can arise and that moment loss is not a perfect proxy for

Method	*vote*			*mushroom*			*spect*		
	ACC ↑	NMI ↑	KL ↓	ACC ↑	NMI ↑	KL ↓	ACC ↑	NMI ↑	KL ↓
NeuNcut	0.858	0.409	2.702	0.608	0.035	**16.040**	0.764	0.102	7.314
DipEncoder	**0.892**	0.519	2.646	0.610	0.052	17.619	0.693	0.020	7.834
EM	0.884	0.511	**2.635**	0.731	0.309	21.747	0.633	**0.119**	**6.836**
VaDE	0.846	0.396	2.761	0.592	0.027	24.194	0.618	0.062	7.542
ODIN	**0.892**	**0.526**	2.683	**0.837**	**0.391**	20.098	**0.794**	0.072	7.005

Table 1: Real-data clustering performance with $m = 2$, matching the number of external binary labels. ACC and NMI measure agreement with the provided labels, while KL is a label-free post-hoc product-Bernoulli fit score; lower KL is better. Best values are bolded.

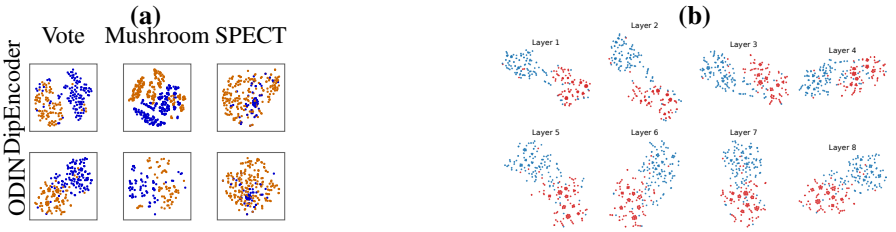


Figure 3: Qualitative t-SNE projections of learned representations. **(a)** DipEncoder and ODIN embeddings on *vote*, *mushroom*, and *spect*, colored by external binary labels. **(b)** ODIN embeddings on *vote* across dictionary layers, also colored by external labels.

parameter error. Likewise, the practical ODIN implementation uses the tractable score $|J(S)|$ rather than exact refined information. This proxy tracks exact RI well at second order but degrades at third order, precisely where higher-order selection becomes most important. These observations motivate inversion-aware and better-conditioned selection criteria as an important direction for future work. Scalability is another limitation. The full interaction family grows exponentially with dimension, and our runtime analysis shows that dictionary construction, rather than mixture inversion, is the main computational bottleneck. Hierarchical fusion and support pruning control this growth, but they also impose heredity constraints: a useful high-order atom may never become reachable if its lower-order ancestors are discarded. More flexible candidate-generation rules, improved RI approximations, and recovery-aware screening are therefore natural extensions.On real data, the main advantage of the interaction representation is interpretability. The recovered components are not only points in an embedding space; they can be described through the same explicit interaction atoms used for inversion. On SPECT, for example, recovered groups can be summarized by enriched co-activation patterns over named cardiac regions. These descriptions should be interpreted as statistical characterizations of the fitted components rather than domain-level causal or clinical conclusions.Overall, the results support a narrower but more defensible claim than universal recovery guarantees: ODIN provides a finite, inspectable interaction representation on which latent mixture recovery can be attempted and diagnostically analyzed.

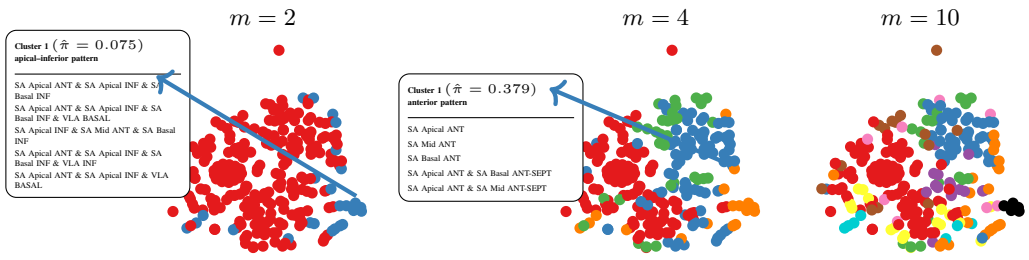


Figure 4: t-SNE embeddings of the ODIN representation on *spect* for mixtures with different values of m. Colors indicate recovered mixture components, not ground-truth labels. The callouts show representative ODIN atoms enriched in selected components, illustrating how recovered clusters can be inspected through explicit interaction patterns.

AI USE STATEMENT

Generative AI tools were used during the development of this work as research and writing assistants. In particular, they were used to assist with the implementation and debugging of experimental code. They were also used for literature search and organization, drafting and editing parts of the manuscript, improving readability, and assisting with LaTeX formatting and scientific figure preparation.

Generative AI tools were not used to generate experimental observations or to replace the execution of the reported experiments. All numerical results reported in the paper were obtained by executing the corresponding experimental pipelines. Relevant literature and citations were manually verified, and AI-assisted code was inspected and tested before being used to produce reported results. The authors reviewed all AI-assisted material and take full responsibility for the final content of this work, including its text, mathematical claims, code, experiments, and conclusions.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce the theoretical and empirical results throughout the paper and appendices. Appendix A gives the complete ODIN dictionary-construction procedure. Appendices B–D provide proofs and additional details for the lifted representation, moment-matching formulation, local identifiability, stability, and dictionary-enrichment results. Appendix F describes the synthetic and real-data experimental settings and implementation details, while Appendix G specifies the preprocessing, hyperparameters, and evaluation procedures used for the real-data experiments. Additional optimization diagnostics, ablations, representation analyses, and interpretability experiments are reported in the remaining experimental appendices. All reported numerical results are obtained from executable implementations of the corresponding procedures.

ETHICS STATEMENT

This work develops a general statistical learning methodology and evaluates it on synthetic data and publicly available benchmark datasets. We do not make clinical or causal claims from the recovered components. In particular, the interpretations reported for the SPECT dataset are intended only as statistical descriptions of patterns in the benchmark data and should not be interpreted as medical conclusions or diagnostic recommendations.

REFERENCES

Mushroom. UCI Machine Learning Repository, 1981. DOI: https://doi.org/10.24432/C5959T.

Congressional Voting Records. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5C01P.

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009. doi: 10.1214/09-AOS689.

Shun-ichi Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2. doi: 10.1007/978-0-387-45528-0.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. *Advances in Neural Information Processing Systems*, 37:107286–107325, 2024.

Brian Buckley, Adrian O'Hagan, and Marie Galligan. Variational bayes latent class analysis for EHR-based phenotyping with large real-world data. *Frontiers in Applied Mathematics and Statistics*, 10: 1302825, 2024. doi: 10.3389/fams.2024.1302825.

Miguel A. Carreira-Perpiñán and Steve Renals. Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12(1):141–152, 2000. doi: 10.1162/089976600300015925.

Krzysztof Cios, Lukasz Kurgan, and Lucy Goodenday. Spect heart. https://doi.org/10.24432/C5P304, 2001.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22, 1977.

James Enouen and Mahito Sugiyama. A complete decomposition of KL error using refined information and mode interaction selection. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=DxUXI15C38.

Georgios Exarchakis and Jörg Lücke. Discrete sparse coding. *Neural Computation*, 29(11):2979–3013, 11 2017. ISSN 0899-7667. doi: 10.1162/neco_a_01015. URL https://doi.org/10.1162/neco_a_01015.

Kazu Ghalamkari, Mahito Sugiyama, and Yoshinobu Kawahara. Many-body approximation for non-negative tensors. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Jiří Grim. EM cluster analysis for categorical data. In Dit-Yan Yeung, James T. Kwok, Ana Fred, Fabio Roli, and Dick de Ridder (eds.), *Structural, Syntactic, and Statistical Pattern Recognition*, volume 4109 of *Lecture Notes in Computer Science*, pp. 640–648, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-37236-3. doi: 10.1007/11815921_70.

Mats Gyllenberg, Timo Koski, Edwin Reilink, and Martin Verlaan. Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31(2):542–548, 1994. doi: 10.2307/3215044.

Wei He, Shangzhi Zhang, Chun-Guang Li, Xianbiao Qi, Rong Xiao, and Jun Guo. Neural normalized cut: A differential and generalizable approach for spectral clustering. *Pattern Recognition*, 164: 111545, 2025. doi: 10.1016/j.patcog.2025.111545.

Marc Henniges, Gervasio Puertas, Jörg Bornschein, Julian Eggert, and Jörg Lücke. Binary sparse coding. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 450–457. Springer, 2010.

Noah Hollmann, Samuel Müller, Lennart Purucker, et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

Rebecca A. Hubbard, Jing Huang, Joanna Harton, Arman Oganisian, Grace Choi, Levon Utidjian, Ihuoma Eneli, L. Charles Bailey, and Yong Chen. A bayesian latent class approach for EHR-based phenotyping. *Statistics in Medicine*, 38(1):74–87, 2019. doi: 10.1002/sim.7953.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. PMLR, 2019.

Jun-Peng Jiang, Si-Yang Liu, Hao-Run Cai, Qile Zhou, and Han-Jia Ye. Representation learning for tabular data: A comprehensive survey. *arXiv preprint arXiv:2504.16109*, 2025.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *International Joint Conference on Artificial Intelligence*, 2016. URL https://api.semanticscholar.org/CorpusID:2546662.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 1965–1972, 2017.

A. Juan and E. Vidal. Bernoulli mixture models for binary images . In *Pattern Recognition, International Conference on*, volume 4, pp. 367–370 Vol.3, Los Alamitos, CA, USA, August 2004. IEEE Computer Society. doi: 10.1109/ICPR.2004.1334543. URL https://doi.ieeecomputersociety.org/10.1109/ICPR.2004.1334543.

Lukasz A Kurgan, Krzysztof J Cios, Ryszard Tadeusiewicz, Marek Ogiela, and Lucy S Goodenday. Knowledge discovery approach to automated cardiac spect diagnosis. *Artificial intelligence in medicine*, 23(2):149–169, 2001.

Sebastien 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 Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 428–484. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/lachapelle22a.html.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Collin Leiber, Lena G. M. Bauer, Michael Neumayr, Claudia Plant, and Christian Böhm. The dipencoder: Enforcing multimodality in autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 846–856, 2022. doi: 10.1145/3534678.3539407.

Collin Leiber, Lukas Miklautz, Claudia Plant, and Christian Böhm. Benchmarking deep clustering algorithms with clustpy. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 625–632. IEEE, 2023. doi: 10.1109/ICDMW60847.2023.00087.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits*, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html. * Equal contribution.

Yiding Lu, Haobin Li, Yunfan Li, Yijie Lin, and Xi Peng. A survey on deep clustering: from the prior perspective. *Vicinagearth*, 1(1):4, 2024.

Fuyuan Lyu, Xing Tang, Dugang Liu, Chen Ma, Weihong Luo, Liang Chen, xiuqiang He, and Xue (Steve) Liu. Towards hybrid-grained feature interaction selection for deep sparse network. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49325–49340. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9ab8da29b1eb3bec912a06e0879065cd-Paper-Conference.pdf.

Yichen Lyu and Pengkun Yang. Identifiability and estimation in high-dimensional nonparametric latent structure models. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3879–3880. PMLR, 2025. URL https://proceedings.mlr.press/v291/lyu25a.html.

Patryk Marszał ek, Tomasz Kuśmierczyk, Witold Wydmański, Jacek Tabor, and Marek Śmieja. Zeus: Zero-shot embeddings for unsupervised separation of tabular data. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 86464–86501. Curran Associates, Inc., 2025. doi: 10.52202/085713-2896. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7d3298e48220b289318b533a848ea069-Paper-Conference.pdf.

João M Pereira, Joe Kileel, and Tamara G Kolda. Tensor moments of gaussian mixture models: Theory and applications. *arXiv preprint arXiv:2202.06930*, 2022.

Minh-Khoi Pham et al. Retrieval-aligned tabular foundation models enable robust clinical risk prediction in electronic health records under real-world constraints. *arXiv preprint arXiv:2604.01841*, 2026.

Ekaterina Poslavskaya and Alexey Korolev. Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding? *CoRR*, abs/2312.16930, 2023. doi: 10.48550/ARXIV.2312.16930. URL https://doi.org/10.48550/arXiv.2312.16930.

Donya Rahmani, Mahesan Niranjan, Damien Fay, Akiko Takeda, and Jacek Brodzki. Estimation of gaussian mixture models via tensor moments with application to online learning. *Pattern Recognition Letters*, 131:285–292, 2020.

Hafiz Tayyab Rauf, André Freitas, and Norman William Paton. Tabledc: Deep clustering for tabular data. *Proc. ACM Manag. Data*, 3(3), June 2025. doi: 10.1145/3725366. URL https://doi.org/10.1145/3725366.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*,

2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94. Survey Certification.

Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric estimation of multi-view latent variable models. In *International Conference on Machine Learning*, pp. 640–648. PMLR, 2014.

Angela Sorgente, Rossella Caliciuri, Matteo Robba, Margherita Lanz, and Bruno D. Zumbo. A systematic review of latent class analysis in psychology: Examining the gap between guidelines and research practice. *Behavior Research Methods*, 57(11):301, 2025. doi: 10.3758/s13428-025-02812-1.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Information decomposition on structured space. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 575–579, 2016a.

Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Information decomposition on structured space. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 575–579. IEEE, 2016b. doi: 10.1109/isit.2016.7541364. URL http://dx.doi.org/10.1109/ISIT.2016.7541364.

Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Legendre decomposition for tensors. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Boris Van Breugel and Mihaela Van Der Schaar. Position: why tabular foundation models should be a research priority. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, 2024.

Tian Wang and Chuang Wang. Latent neural operator for solving forward and inverse pde problems. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 33085–33107. Curran Associates, Inc., 2024. doi: 10.52202/079017-1042. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/39f6d5c2e310a5a629dcfc4d517aa0d1-Paper-Conference.pdf.

Gongjun Xu. Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45(2):675–707, 2017. doi: 10.1214/16-AOS1464.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrlic, et al. A scoping review of self-supervised representation learning for clinical decision making using ehr categorical data. *npj Digital Medicine*, 2025. DOI:10.1038/s41746-025-01692-1.

Zhenyu Zhu, Marco Fumero, Francesco Locatello, and Volkan Cevher. Diagnosing and fixing latent recovery in sparse autoencoders. In *ICLR 2026 Workshop on Unifying Concept Representation Learning*, 2026. URL https://openreview.net/forum?id=Hl3rEn7S4P.

APPENDIX ORGANIZATION

Appendix A gives the full ODIN dictionary-learning procedure and the computation of refined information. Appendix B proves the atom-wise separation result. Appendix C details the lifted moment-matching formulation, the local identifiability proof, and the perturbation bound. Appendix D records the effect of dictionary enrichment on the lifted Jacobian. Finally, Appendices F and G give dataset preprocessing, hyperparameters, and additional experimental results. The code is provided as anonymized supplementary material and will be released publicly upon acceptance.

A ODIN ALGORITHM

We describe the practical procedure implementing ODIN's hierarchical learning framework introduced above. The full procedure is summarized in Algorithm 1. The algorithm constructs a dictionary incrementally by interaction order, enforcing hierarchical admissibility and guaranteeing a non-increasing KL projection error. At a given order k, candidate interactions are generated exclusively from atoms selected at order $k-1$, according to the admissibility condition.

Candidate interactions are first filtered by empirical frequency, removing rare patterns according to the threshold (Sugiyama et al., 2016b):

$$\hat{p}(S) \geq \sigma. \tag{3}$$

Here

$$\hat{p}(S) := \frac{1}{n}\sum_{r=1}^{n} \phi_S(X^{(r)}) = \frac{1}{n}\sum_{r=1}^{n}\prod_{i\in S} X_i^{(r)}$$

is the empirical frequency of the simultaneous activation of all variables in S.

Once $\mathrm{RI}_{D,S}$ is computed, we also apply an information-pruning step. Candidate atoms whose refined information is below a threshold τ are discarded:

$$\mathrm{RI}_{D,S}(\hat{p}) \geq \tau. \tag{4}$$

The remaining candidates are selected in decreasing order of refined information and added sequentially to the dictionary.

Numerical scoring of candidate interactions. The interaction-selection principle underlying ODIN is refined information, defined as the reduction in KL projection error obtained by enriching the current dictionary D with a candidate atom S:

$$\mathrm{RI}_{D,S}(\hat{p}) = \mathcal{L}(D) - \mathcal{L}(D\cup\{S\}), \qquad \mathcal{L}(D) = \inf_{q\in\mathcal{M}_D} \mathrm{KL}(\hat{p}\|q).$$

Computing this quantity exactly requires solving an additional log-linear projection problem for each candidate S. Since the hierarchical search may evaluate thousands of candidate interactions, repeatedly computing these projections is prohibitively expensive in practice.

Accordingly, the ODIN dictionaries used in our experiments are constructed using the tractable interaction score J_S introduced by Enouen & Sugiyama (2026). For $S \subseteq [d]$, we define

$$J_S := \sum_{T\subseteq S} (-1)^{|S|-|T|}\mathrm{KL}(\hat{p}_T\|u_T),$$

where $\hat{p}_T$ denotes the empirical marginal distribution on T, and u_T is the uniform distribution on $\{0,1\}^{|T|}$. Equivalently,

$$J_S = \sum_{T\subseteq S} (-1)^{|S|-|T|}\big(|T|\log 2 - H_{\hat{p}}(T)\big).$$

In practice, candidates are ranked according to

$$\mathrm{score}(S) = |J_S|.$$

Unlike exact refined information, $\mathrm{RI}_{D,S}$, the quantity J_S depends only on the empirical marginal distributions associated with subsets of S and does not require fitting a new log-linear model for each candidate. Its cost therefore depends on the order $|S|$, through the $2^{|S|}$ subsets involved in the inclusion–exclusion sum, rather than on enumerating the complete binary state space $\{0,1\}^d$.

The distinction between these two quantities is important. Exact refined information is the information-geometric criterion motivating the selection procedure, whereas J_S is the computational score used by ODIN in the reported experiments. We do not assume that $|J_S|$ is equal to $\mathrm{RI}_{D,S}$, nor that it preserves its candidate ranking at all interaction orders. Appendix H.4 therefore evaluates their empirical agreement in low-dimensional settings where exact refined information can be computed, and quantifies the corresponding computational trade-off.

Algorithm 1: Hierarchical Information-Based Dictionary Learning

Input: Binary data $X \in \{0,1\}^{N \times d}$; maximum order K; thresholds σ, τ
Output: Order-wise hierarchical dictionary $\{D_k\}_{k=0}^{K}$
Initialize $D_0 \leftarrow \emptyset$; $L_1 \leftarrow \{\{i\} : i \in [d],\ \hat{p}(\{i\}) > 0\}$; `// unary interactions`
Set $D_1 \leftarrow L_1$;
for $k \leftarrow 2$ **to** K **do**
  `// Initialize the current order-`k` layer and dictionary state`
  $L_k \leftarrow \emptyset$; $D_{k,0} \leftarrow D_{k-1}$;
  `// Generate admissible order-`k` candidates`

$$\mathcal{A}_k \leftarrow \{S_1 \cup S_2 : S_1, S_2 \in L_{k-1},\ |S_1 \cup S_2| = k,\ |S_1 \cap S_2| = k-2\}.$$

  `// Prune candidates with insufficient empirical support`
  `equation 3`

$$\mathcal{A}_k \leftarrow \{S \in \mathcal{A}_k : \hat{p}(S) \geq \sigma\}.$$

  `// Sequential expansion by maximal refined information`
  $t \leftarrow 1$;
  while $\mathcal{A}_k \neq \emptyset$ **do**
    `// Compute refined information at the current dictionary`
    For each $S \in \mathcal{A}_k$, compute

$$\mathrm{RI}_{D_{k,t-1},S}(\hat{p}) = \mathcal{L}(D_{k,t-1}) - \mathcal{L}(D_{k,t-1} \cup \{S\}).$$

    Select

$$S_{k,t} \in \arg\max_{S \in \mathcal{A}_k} \mathrm{RI}_{D_{k,t-1},S}(\hat{p}).$$

    if

$$\mathrm{RI}_{D_{k,t-1},S_{k,t}}(\hat{p}) < \tau$$

    then
      break
    Update

$$D_{k,t} \leftarrow D_{k,t-1} \cup \{S_{k,t}\}.$$

    $L_k \leftarrow L_k \cup \{S_{k,t}\}$;
    $\mathcal{A}_k \leftarrow \mathcal{A}_k \setminus \{S_{k,t}\}$;
    $t \leftarrow t + 1$;
  $D_k \leftarrow D_{k-1} \cup L_k$;
  `// Stop if no interaction is selected at this order`
  if $L_k = \emptyset$ **then**
    break

End-to-end runtime. Per-candidate scoring cost does not by itself determine the practical cost of the complete pipeline, because ODIN may evaluate many candidate atoms before constructing the final dictionary. We therefore also report end-to-end wall-clock runtimes, separating dictionary construction from mixture inversion. All methods were timed on the same hardware; the compute environment is described in Table 2.

The runtime decomposition shows that ODIN's computational bottleneck is dictionary construction rather than mixture inversion. Across these settings, the inversion stage remains between 6.75 and 8.00 seconds, whereas construction time increases sharply with the number of candidate atoms scored. Candidate count is therefore a more informative predictor of computational cost than the ambient dimension d alone. For example, *vote* has larger dimension than the synthetic $d = 15$ setting, yet its construction time is substantially smaller because hierarchical pruning leaves fewer candidate interactions to score. These results locate ODIN's principal scalability limitation in candidate generation and scoring, rather than in the subsequent moment-matching inversion.

Method	$d = 5$	$d = 15$	$d = 20$	*vote* ($d = 16$)	*spect* ($d = 22$)
ODIN: atoms scored	23	5,436	31,598	2,438	79,168
ODIN: construction	0.16 s	65.50 s	578.15 s	15.72 s	1,451.88 s
ODIN: inversion	6.75 s	6.85 s	7.05 s	7.51 s	8.00 s
ODIN: total	**6.92 s**	**72.64 s**	**586.71 s**	**23.37 s**	**1,463.65 s**
VaDE	56.69 s	56.52 s	56.06 s	7.81 s	11.06 s
ACeDeC	301.60 s	297.65 s	298.58 s	197.80 s	287.59 s
DipEncoder	153.62 s	75.77 s	75.55 s	10.30 s	11.38 s

Table 2: End-to-end wall-clock runtime comparison. For ODIN, dictionary construction and mixture inversion are reported separately, together with the total runtime and the number of candidate atoms scored during dictionary construction. All methods were evaluated on the same hardware.

Remark: Alternative proxies. Other tractable proxies could be used to approximate refined information. We do not study them here; throughout the experiments, we use the score J_S proposed by Enouen & Sugiyama (2026).

B Details on the ODIN-induced lifting

This appendix proves Proposition 1 and records the interpretation of the atom-wise quantities appearing in the lifted geometry.

For $S \in D$, the coordinate $[v_D(\mu)]_S = \prod_{i \in S} \mu_i$ is the probability, under a product-Bernoulli component with parameter μ, that all variables in S are simultaneously active. Therefore, for two components μ, ν, the quantity

$$\Delta_S(\mu, \nu) := \left(\prod_{i \in S} \mu_i - \prod_{i \in S} \nu_i \right)^2$$

is the contribution of atom S to their lifted squared separation. In particular, S contributes nothing to separating this pair precisely when

$$\prod_{i \in S} \mu_i = \prod_{i \in S} \nu_i.$$

Thus an atom may be present in the representation but geometrically irrelevant for a given pair of components. Conversely, atoms with large $\Delta_S(\mu, \nu)$ are those whose simultaneous activation probabilities most distinguish the components.

Proof of Proposition 1. By definition,

$$v_D(\mu) - v_D(\nu) = \left(\prod_{i \in S} \mu_i - \prod_{i \in S} \nu_i \right)_{S \in D}.$$

Taking the squared Euclidean norm gives

$$\|v_D(\mu) - v_D(\nu)\|_2^2 = \sum_{S \in D} \left(\prod_{i \in S} \mu_i - \prod_{i \in S} \nu_i \right)^2 = \sum_{S \in D} \Delta_S(\mu, \nu).$$

If $D \subseteq D'$, applying the same identity to D' and subtracting the identity for D yields

$$\|v_{D'}(\mu) - v_{D'}(\nu)\|_2^2 - \|v_D(\mu) - v_D(\nu)\|_2^2 = \sum_{S \in D' \setminus D} \Delta_S(\mu, \nu).$$

This proves the claim. □

The proposition is component-level: it explains how ODIN atoms decompose separation between individual product-Bernoulli components. Mixture recovery, however, is governed by the observable lifted moment

$$\mathbb{E}_\eta[z_D(X)] = \sum_{h=1}^{m} \pi_h v_D(\mu_h) = \Xi_{D,m}(\eta).$$

Thus component-level separation is an interpretable diagnostic, while the inverse problem itself is controlled by the finite lifted moment map $\Xi_{D,m}$.

C Details on ODIN-to-mixture inversion

This appendix gives the details behind Proposition 2. The goal is to make explicit how the ODIN dictionary D enters the local inverse problem through the Jacobian of the lifted moment map. We also separate the population exact-fit problem from the finite-sample least-squares problem.

C.1 Moment matching in the lifted ODIN coordinates

Once ODIN has selected a dictionary $D \subseteq 2^{[d]}$, the component embedding is

$$v_D(\mu) = \left(\prod_{i \in S} \mu_i \right)_{S \in D},$$

and the lifted mixture moment map is

$$\Xi_{D,m}(\eta) := \sum_{h=1}^{m} \pi_h v_D(\mu_h), \qquad \eta = ((\pi_h, \mu_h))_{h=1}^{m}.$$

Given observations $X^{(1)}, \dots, X^{(n)}$, the empirical lifted moment vector is

$$\hat{M}_D := \frac{1}{n} \sum_{t=1}^{n} z_D(X^{(t)}) = (\hat{M}_S)_{S \in D}, \qquad \hat{M}_S = \frac{1}{n} \sum_{t=1}^{n} \phi_S(X^{(t)}).$$

Thus ODIN-to-mixture inversion is the problem of matching $\hat{M}_D$ to the image of $\Xi_{D,m}$.

We work over the compact feasible set

$$\Theta_m = \Big\{ (\pi, \mu) : \ \pi \in \Delta_{m-1}, \ \mu_{h,i} \in [0,1] \Big\}.$$

The empirical lifted moment-matching objective is

$$F_D(\eta) := \left\| \hat{M}_D - \Xi_{D,m}(\eta) \right\|_2^2, \qquad \eta \in \Theta_m.$$

Equivalently,

$$\hat{\eta} \in \arg\min_{\eta \in \Theta_m} F_D(\eta).$$

Expanded in coordinates, this is

$$(\hat{\pi}, \hat{\mu}) \in \arg\min_{\pi, \mu} \sum_{S \in D} \left(\hat{M}_S - \sum_{h=1}^{m} \pi_h \prod_{i \in S} \mu_{h,i} \right)^2,$$

subject to $\pi \in \Delta_{m-1}$ and $\mu_{h,i} \in [0,1]$.

Connection to moment methods. This constrained minimization formulation is the recovery principle induced by the ODIN lift. It recasts ODIN-to-mixture inversion as a numerical moment-matching problem, in the spirit of optimization-based method-of-moments approaches (Rahmani et al., 2020; Pereira et al., 2022). Classical tensor-decomposition methods arise as a more structured special case: when the lifted observables can be organized into conditionally independent views, the associated cross-moment tensor has a low-rank decomposition that can be exploited for latent-variable recovery (Anandkumar et al., 2014; Song et al., 2014). ODIN follows the same moment-based principle, but replaces the full dense higher-order tensor dictionary by a learned finite dictionary D of interaction atoms $\phi_S(x) = \prod_{i \in S} x_i$.

C.2 Exact fit and empirical approximation

The empirical objective

$$F_D(\eta) := \left\| \Xi_{D,m}(\eta) - \widehat{M}_D \right\|_2^2$$

always admits a minimizer on Θ_m, since Θ_m is compact and F_D is continuous. In general, however, this minimizer only provides a best approximation to the empirical lifted moment vector $\widehat{M}_D$.

At the population level, suppose that there exists $\eta^\star \in \Theta_m^\circ$ such that

$$M_D = \Xi_{D,m}(\eta^\star).$$

Then $\eta^\star$ is an exact interior solution of the population lifted moment equations.

For finite n, an exact empirical solution need not exist:

$$\widehat{M}_D \notin \Xi_{D,m}(\Theta_m)$$

in general, especially when

$$|D| > (m-1) + md.$$

What holds instead is convergence of the empirical lifted moments. For a fixed finite dictionary D, the strong law of large numbers gives

$$\widehat{M}_D^{(n)} \longrightarrow M_D \qquad \text{almost surely as } n \to \infty.$$

Therefore, the empirical estimator should be interpreted as a local least-squares approximation to the population inverse problem, rather than as an exact solution of

$$\Xi_{D,m}(\eta) = \widehat{M}_D.$$

C.3 Proof of the local identifiability criterion

We now prove Proposition 2. Assume that there exists an interior population exact solution

$$\eta^\star \in \Theta_m^\circ, \qquad \Xi_{D,m}(\eta^\star) = M_D,$$

where

$$\Theta_m^\circ := \left\{ (\pi, \mu) : \pi_h > 0,\ \sum_{h=1}^m \pi_h = 1,\ \mu_{h,i} \in (0,1) \right\}.$$

The restriction to the interior removes boundary degeneracies. In particular, if $\pi_h = 0$, then component h is inactive and its parameter μ_h cannot be identified from the observed law.

Consider the population least-squares objective

$$F_D^{\mathrm{pop}}(\eta) := \frac{1}{2} \left\| \Xi_{D,m}(\eta) - M_D \right\|_2^2.$$

Since each coordinate of $\Xi_{D,m}$ is polynomial in (π, μ), the residual map

$$r_D(\eta) := \Xi_{D,m}(\eta) - M_D$$

is smooth. Its Hessian satisfies

$$\nabla^2 F_D^{\mathrm{pop}}(\eta) = J_{r_D}(\eta)^\top J_{r_D}(\eta) + \sum_{S \in D} r_{D,S}(\eta)\, \nabla^2 r_{D,S}(\eta).$$

At the exact population solution, $r_D(\eta^\star) = 0$, and therefore

$$\nabla^2 F_D^{\mathrm{pop}}(\eta^\star) = J_{r_D}(\eta^\star)^\top J_{r_D}(\eta^\star).$$

Thus the Gauss–Newton Hessian is exact at the population exact solution. In finite-sample or misspecified settings, the second term need not vanish.

Because the mixture weights satisfy the simplex constraint $\sum_{h=1}^m \pi_h = 1$, we work in the free coordinates

$$\bar{\eta} = (\pi_1, \ldots, \pi_{m-1}, \mu_1, \ldots, \mu_m),$$

with

$$\pi_m = 1 - \sum_{h=1}^{m-1} \pi_h.$$

Let $\widetilde{J}_{r_D}(\eta^\star)$ denote the Jacobian of the residual map with respect to these free coordinates. For any feasible tangent direction δ,

$$\delta^\top \nabla^2 F_D^{\mathrm{pop}}(\eta^\star)\delta = \left\| \widetilde{J}_{r_D}(\eta^\star)\delta \right\|_2^2.$$

Thus first-order injectivity of the lifted moment map on the feasible tangent space implies second-order isolation of the exact population minimizer.

For $h = 1, \dots, m$, define

$$v_h^\star := \left(\prod_{i \in S} \mu_{h,i}^\star \right)_{S \in D} \in \mathbb{R}^{|D|}$$

and let

$$G_h^\star := \left[\frac{\partial}{\partial \mu_{h,i}} \prod_{j \in S} \mu_{h,j} \Big|_{\mu_h = \mu_h^\star} \right]_{\substack{S \in D \\ i = 1, \dots, d}} \in \mathbb{R}^{|D| \times d}.$$

The derivatives of the lifted moment map with respect to the free mixture weights are

$$\frac{\partial \Xi_{D,m}}{\partial \pi_h}(\eta^\star) = v_h^\star - v_m^\star, \qquad h = 1, \dots, m-1,$$

whereas the derivatives with respect to the component parameters are

$$\frac{\partial \Xi_{D,m}}{\partial \mu_h}(\eta^\star) = \pi_h^\star G_h^\star, \qquad h = 1, \dots, m.$$

Hence the reduced Jacobian of the lifted moment map in free coordinates is

$$\mathcal{J}_D(\eta^\star) := \left[\; v_1^\star - v_m^\star \; \cdots \; v_{m-1}^\star - v_m^\star \; \middle| \; \pi_1^\star G_1^\star \; \middle| \; \cdots \; \middle| \; \pi_m^\star G_m^\star \; \right]. \tag{5}$$

This matrix has size

$$|D| \times \big((m-1) + md \big).$$

If $\mathcal{J}_D(\eta^\star)$ has full column rank, then for every nonzero free-coordinate perturbation $\bar{\delta}$,

$$\mathcal{J}_D(\eta^\star)\bar{\delta} \neq 0.$$

Consequently,

$$\delta^\top \nabla^2 F_D^{\mathrm{pop}}(\eta^\star)\delta > 0 \qquad \forall\, \delta \in T_{\eta^\star}\Theta_m \setminus \{0\}.$$

Thus $\eta^\star$ is an isolated exact minimizer of the population least-squares criterion in the labelled parameterization.

It remains to connect this differential condition to local identifiability. Since $\mathcal{J}_D(\eta^\star)$ has full column rank, there exists a set of $(m-1)+md$ coordinates of $\Xi_{D,m}$ whose square sub-Jacobian is invertible at $\eta^\star$. By the inverse function theorem applied to this coordinate projection, the lifted moment map $\Xi_{D,m}$ is locally injective at $\eta^\star$ in the labelled free coordinates. Therefore, there exists a neighborhood U of $\eta^\star$ such that

$$\Xi_{D,m}(\eta) = \Xi_{D,m}(\eta^\star), \qquad \eta \in U,$$

implies

$$\eta = \eta^\star$$

in the labelled parameterization.

For an unlabelled mixture, the remaining ambiguity is the usual permutation of mixture components. Hence the locally equivalent parameterizations are

$$\left\{ \left((\pi_{\sigma(h)}^\star, \mu_{\sigma(h)}^\star) \right)_{h=1}^m : \sigma \in \mathfrak{S}_m \right\}.$$

This proves Proposition 2.

C.4 Conditioning and a perturbation bound

The rank condition guarantees local identifiability, while its quantitative stability is controlled by the smallest singular value of the reduced lifted Jacobian. Work in the free coordinates

$$(\pi_1, \dots, \pi_{m-1}, \mu_1, \dots, \mu_m)$$

and write

$$\Phi(\eta) := \Xi_{D,m}(\eta).$$

Assume that

$$\sigma^\star := \sigma_{\min}\big(\mathcal{J}_D(\eta^\star)\big) > 0.$$

Since Φ is continuously differentiable and

$$D\Phi(\eta^\star) = \mathcal{J}_D(\eta^\star),$$

there exists a neighborhood U of $\eta^\star$ such that, for every $\eta \in U$,

$$\|\Phi(\eta) - \Phi(\eta^\star)\|_2 \geq \frac{\sigma^\star}{2} \|\eta - \eta^\star\|_2 .$$

Indeed, by continuity of $D\Phi$, the neighborhood U can be chosen such that

$$\|D\Phi(\eta) - D\Phi(\eta^\star)\|_{\mathrm{op}} \leq \frac{\sigma^\star}{2} \qquad \forall\, \eta \in U.$$

Using the integral mean-value representation,

$$\Phi(\eta) - \Phi(\eta^\star) = \int_0^1 D\Phi\big(\eta^\star + t(\eta - \eta^\star)\big)(\eta - \eta^\star)\, dt,$$

the preceding lower bound follows.

Let $\widehat{\eta}$ be a local minimizer of the empirical least-squares criterion lying in U. Since $\eta^\star$ is feasible,

$$\left\|\Phi(\widehat{\eta}) - \widehat{M}_D\right\|_2 \leq \left\|\Phi(\eta^\star) - \widehat{M}_D\right\|_2 = \left\|M_D - \widehat{M}_D\right\|_2 .$$

Therefore,

$$\begin{aligned}\|\Phi(\widehat{\eta}) - M_D\|_2 &\leq \left\|\Phi(\widehat{\eta}) - \widehat{M}_D\right\|_2 + \left\|\widehat{M}_D - M_D\right\|_2 \\ &\leq 2\left\|\widehat{M}_D - M_D\right\|_2 .\end{aligned}$$

Combining this estimate with the local lower bound yields

$$\|\widehat{\eta} - \eta^\star\|_2 \leq \frac{4}{\sigma^\star}\left\|\widehat{M}_D - M_D\right\|_2 .$$

Since every lifted coordinate is a bounded Bernoulli monomial, standard concentration gives

$$\left\|\widehat{M}_D - M_D\right\|_2 = O_P\left(\sqrt{\frac{|D|}{n}}\right),$$

and hence

$$\|\widehat{\eta} - \eta^\star\|_2 = O_P\left(\frac{1}{\sigma^\star}\sqrt{\frac{|D|}{n}}\right).$$

This bound describes the local rank–conditioning trade-off. Adding atoms increases $|D|$ and can improve rank or increase $\sigma^\star$, but it also increases the dimension of the empirical moment vector. Thus dictionary enrichment is statistically useful when the added atoms improve the conditioning of the lifted inverse problem sufficiently to compensate for the additional empirical moment noise.

C.5 Example: dictionary choice changes the rank condition

Consider a two-component product-Bernoulli mixture with $d = 4$. The number of free parameters is

$$(m-1) + md = 1 + 2 \cdot 4 = 9.$$

First consider the singleton dictionary

$$D_{\text{sing}} = \big\{\{1\}, \{2\}, \{3\}, \{4\}\big\}.$$

The corresponding reduced Jacobian has only four rows, and therefore

$$\operatorname{rank} \mathcal{J}_{D_{\text{sing}}}(\eta) \le 4 < 9$$

for every interior parameter η. Hence singleton moments alone cannot satisfy the full-rank criterion.

Now consider the richer dictionary

$$D = \big\{\{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{1,2,3\}\big\},$$

which contains nine atoms. At the parameter

$$\pi_1^\star = \pi_2^\star = \frac{1}{2},$$

with

$$\mu_1^\star = \left(\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}\right), \qquad \mu_2^\star = \left(\frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}\right),$$

the matrix $\mathcal{J}_D(\eta^\star)$ is 9×9, and direct calculation gives

$$\det \mathcal{J}_D(\eta^\star) = \frac{5}{663552} \neq 0.$$

Therefore,

$$\operatorname{rank} \mathcal{J}_D(\eta^\star) = 9,$$

so the local identifiability condition holds at this parameter point.

D Dictionary enrichment and the lifted Jacobian

The previous section shows that local identifiability is controlled by the rank of $\mathcal{J}_D(\eta^\star)$. We now record how this Jacobian changes when the dictionary is enlarged.

D.1 Monotonicity under dictionary enrichment

Let $D \subseteq D' \subseteq 2^{[d]}$, and fix $\eta \in \Theta_m^\circ$. Order the atoms of D' so that the atoms of D appear first. Then the lifted Jacobian for D' is obtained by stacking the old rows with the rows contributed by the added atoms:

$$\mathcal{J}_{D'}(\eta) = \begin{bmatrix} \mathcal{J}_D(\eta) \\ R_{D' \setminus D}(\eta) \end{bmatrix}.$$

Consequently,

$$\mathcal{J}_{D'}(\eta)^\top \mathcal{J}_{D'}(\eta) = \mathcal{J}_D(\eta)^\top \mathcal{J}_D(\eta) + R_{D' \setminus D}(\eta)^\top R_{D' \setminus D}(\eta),$$

and therefore

$$\mathcal{J}_{D'}(\eta)^\top \mathcal{J}_{D'}(\eta) \succeq \mathcal{J}_D(\eta)^\top \mathcal{J}_D(\eta).$$

It follows that

$$\operatorname{rank} \mathcal{J}_{D'}(\eta) \ge \operatorname{rank} \mathcal{J}_D(\eta).$$

Moreover, if $\mathcal{J}_D(\eta)$ has full column rank, then

$$\sigma_{\min}(\mathcal{J}_{D'}(\eta)) \ge \sigma_{\min}(\mathcal{J}_D(\eta)).$$

Thus dictionary enrichment cannot destroy an already full-rank Jacobian.

D.2 Persistence of full rank

As an immediate consequence, if $D \subseteq D'$ and

$$\operatorname{rank} \mathcal{J}_D(\eta) = (m-1) + md,$$

then

$$\operatorname{rank} \mathcal{J}_{D'}(\eta) = (m-1) + md.$$

Therefore, once the full-rank condition holds for a dictionary D, it continues to hold for every enrichment $D' \supseteq D$.

D.3 When a candidate atom strictly improves rank

Let $S \notin D$, and order the atoms of $D \cup \{S\}$ so that S appears last. Then

$$\mathcal{J}_{D\cup\{S\}}(\eta) = \begin{bmatrix} \mathcal{J}_D(\eta) \\ r_S(\eta) \end{bmatrix},$$

where $r_S(\eta)$ is the row contributed by atom S. Therefore,

$$\operatorname{rank} \mathcal{J}_{D\cup\{S\}}(\eta) > \operatorname{rank} \mathcal{J}_D(\eta)$$

if and only if

$$r_S(\eta) \notin \operatorname{RowSpan}(\mathcal{J}_D(\eta)).$$

Thus a candidate atom strictly improves the first-order inverse geometry exactly when its Jacobian row contributes a new independent sensitivity direction.

E Interpretation and scope of the product-Bernoulli assumption

Latent-class interpretation and applications. The mixture-inversion stage of ODIN is instantiated using an m-component mixture of product-Bernoulli distributions,

$$p_\eta(x) = \sum_{h=1}^{m} \pi_h \prod_{i=1}^{d} \mu_{h,i}^{x_i}(1-\mu_{h,i})^{1-x_i}, \qquad x \in \{0,1\}^d.$$

Equivalently, one first samples a latent component $H \sim \operatorname{Categorical}(\pi)$ and then, conditionally on $H = h$, samples

$$X_i \mid H = h \sim \operatorname{Bernoulli}(\mu_{h,i}), \qquad i \in [d],$$

independently across coordinates. Thus, mixtures of product-Bernoulli distributions correspond to binary latent-class models: the latent variable represents heterogeneous subpopulations, while the observed binary variables are conditionally independent within each subpopulation. Importantly, the model does not assume that the observed variables are marginally independent; their dependence is induced by variation of the Bernoulli parameters across latent classes.

This model is used in applications involving heterogeneous binary populations. Examples include probabilistic identification of bacteria (Gyllenberg et al., 1994), restricted binary latent-class models for cognitive diagnosis (Xu, 2017), and binary image recognition and clustering of binarized handwritten digits (Juan & Vidal, 2004; Bishop, 2006). Closely related mixed-data latent-class extensions have also been used for phenotyping electronic health records (Hubbard et al., 2019; Buckley et al., 2024). More generally, product-Bernoulli mixtures constitute the binary, locally independent case of latent-class analysis (Sorgente et al., 2025).

Representational capacity and modeling restriction. The product-Bernoulli family has broad representational capacity when the number of components is unrestricted. Indeed, any probability distribution p on $\{0,1\}^d$ can be written as a mixture of at most 2^d product-Bernoulli distributions. For every $a \in \{0,1\}^d$, consider the degenerate product component with $\mu_i = a_i$. This component assigns probability one to a, and assigning it mixture weight $p(a)$ reconstructs the original distribution exactly. This observation is sometimes described as the universal expressiveness of Bernoulli-product mixtures (Grim, 2006).

This expressiveness result should not be interpreted as saying that the assumption is nonrestrictive in the setting considered here. The exact construction may require exponentially many components and uses boundary parameters $\mu_{h,i} \in \{0, 1\}$, whereas our local-identifiability analysis is formulated for interior parameters and a fixed, typically small, number of components. Choosing m therefore remains a substantive modeling decision: the model assumes that the observed heterogeneity can be approximated by a small number of conditionally independent binary subpopulations.

Identifiability of product-Bernoulli mixtures. The product-Bernoulli family is also a meaningful test case because its parameter-identifiability problem is nontrivial. Finite mixtures of multivariate Bernoulli distributions are not strictly identifiable over their entire parameter space: distinct parameter values can induce the same observed distribution (Gyllenberg et al., 1994; Carreira-Perpiñán & Renals, 2000). Nevertheless, with a known number of components and sufficiently many observed variables, the model is generically identifiable up to permutation of the component labels (Allman et al., 2009). Generic identifiability is an almost-everywhere statement and does not by itself certify a particular numerical parameter or fitted solution.

Moreover, standard component-wise linear-independence conditions cannot be applied directly when $m \geq 3$, because univariate Bernoulli distributions belong to a two-dimensional vector space, and hence no collection of three or more such component distributions can be linearly independent. Alternative identifiability conditions therefore rely on grouped variables, parameter separation, or the joint observable representation (Allman et al., 2009; Lyu & Yang, 2025).

Our local result addresses this issue from the perspective of the selected observable moments. For a fixed dictionary D, Proposition 2 gives a pointwise population criterion: if the reduced Jacobian of the lifted moment map has full column rank at an interior exact solution, then that solution is locally identifiable up to label permutation. The condition must be evaluated at the parameter of interest and does not establish strict or global identifiability of the entire product-Bernoulli family.

Scope of dictionary construction and mixture inversion. The ODIN pipeline contains two blocks whose assumptions should be distinguished. First, ODIN dictionary construction uses empirical frequencies of binary co-activation events and does not assume that the observations were generated by a product-Bernoulli mixture. The interaction atoms

$$\phi_S(x) = \prod_{i \in S} x_i$$

and their empirical expectations are well defined for any distribution on $\{0, 1\}^d$. Consequently, the dictionary-learning and lifting stages can be applied to arbitrary binary data. The same counting principle can also be extended to finite categorical observations by introducing category-indicator functions.

Second, mixture inversion uses the product-Bernoulli assumption to map the selected interaction moments back to latent parameters. Under this assumption,

$$\mathbb{E}_\eta[\phi_S(X)] = \sum_{h=1}^{m} \pi_h \prod_{i \in S} \mu_{h,i},$$

which defines the lifted moment map $\Xi_{D,m}$ and its corresponding Jacobian. Extending the inversion theory to another component family would require deriving the associated moment map, parameter constraints, and Jacobian-rank condition; the dictionary construction itself would not need to be replaced.

Interpretation under model misspecification. When the data are not generated by an m-component product-Bernoulli mixture, the recovered parameters should not be interpreted as true generative parameters. Instead, they define a minimum-distance approximation in the selected moment coordinates:

$$\eta_D^\dagger \in \arg\min_{\eta \in \Theta_m} \|M_D - \Xi_{D,m}(\eta)\|_2^2,$$

where M_D contains the population moments selected by ODIN. The finite-sample estimator analogously minimizes the distance between the empirical selected moments and those generated by the fitted mixture. Therefore, under misspecification, the fitted parameters describe the product-Bernoulli

mixture that best reproduces the selected interactions, rather than the unknown data-generating mechanism.

The one-hot representation of the Mushroom dataset provides a concrete example. Indicators associated with categories of the same original variable are mutually exclusive and therefore cannot be conditionally independent Bernoulli variables. In this case, the fitted product-Bernoulli mixture should be interpreted as a relaxed approximation to a product-categorical latent-class model. The corresponding preprocessing and approximation are discussed in Appendix F.3.

F Datasets and Preprocessing

This appendix summarizes the datasets used in the real-data benchmark and the preprocessing applied before running ODIN and the competing baselines. Table 5 reports the main dataset characteristics.

F.1 Synthetic toy dataset

The synthetic recovery benchmark uses a four-component mixture of product Bernoulli distributions in dimension $d = 6$. Each observation $X \in \{0,1\}^6$ is generated by first sampling a latent component $H \sim \mathrm{Categorical}(\pi)$, and then sampling the coordinates conditionally independently according to

$$X_i \mid H = h \sim \mathrm{Bernoulli}(\mu_{h,i}), \qquad i = 1, \dots, 6.$$

Thus the population distribution is

$$p(x) = \sum_{h=1}^{4} \pi_h \prod_{i=1}^{6} \mu_{h,i}^{x_i} (1 - \mu_{h,i})^{1-x_i}.$$

The parameters are chosen to make the recovery problem deliberately nontrivial. Components 1 and 2 form a nearby pair, and components 3 and 4 form a second nearby pair. Consequently, first-order moments alone provide weak separation between components inside each pair. The purpose of this construction is to test whether higher-order ODIN atoms supply additional observable interaction moments that improve mixture recovery.

For this setting, the number of free mixture parameters is

$$(m - 1) + md = (4 - 1) + 4 \cdot 6 = 27.$$

Therefore, any lifted moment system with fewer than 27 selected atoms is dimensionally incapable of having a full-column-rank Jacobian. This makes the example useful for illustrating the transition from underdetermined early dictionary layers to richer lifted representations in which local identifiability can become possible.

Recovery benchmark. Figure 5 reports the permutation-aligned component recovery error across sample sizes for the five estimators considered in Section 6.1.

F.2 Real-word Datasets

Congressional Voting Records We use the *Congressional Voting Records* dataset from OpenML. The dataset contains 16 categorical voting attributes and two classes, *democrat* and *republican*. Missing entries, encoded as "?", are removed at the sample level: we discard every sample containing at least one missing value. After filtering, the dataset contains 232 samples. The remaining categorical responses are binarized by mapping "y" to 1 and the other observed response to 0, yielding a binary matrix of size 232×16.

SPECT Heart We use the *SPECT Heart* dataset from OpenML. The dataset already consists of binary attributes, so no feature transformation is required beyond conversion to integer-valued binary variables. Samples with missing values are removed, although in practice the retained benchmark contains 267 samples and 22 binary features. The dataset has two classes, encoded as 0 and 1. The repository metadata gives the variables only as F1–F22 "partial diagnoses," while the anatomical ROI names come from the original SPECT paper's 5-slice ROI layout: 3 short-axis slices (apical, mid, basal), 1 horizontal long-axis slice, and 1 vertical long-axis slice, totaling 22 ROIs Kurgan et al. (2001).

Component	π_h	μ_h
1	0.20	(0.30, 0.30, 0.70, 0.70, 0.40, 0.40)
2	0.30	(0.32, 0.28, 0.72, 0.68, 0.42, 0.38)
3	0.25	(0.60, 0.60, 0.20, 0.20, 0.80, 0.80)
4	0.25	(0.62, 0.58, 0.22, 0.18, 0.78, 0.82)

Table 3: Synthetic product-Bernoulli mixture parameters used in the recovery benchmark. The construction contains two nearby component pairs, $(1, 2)$ and $(3, 4)$, so that low-order moments provide limited separation and higher-order interaction moments are needed for stable recovery.

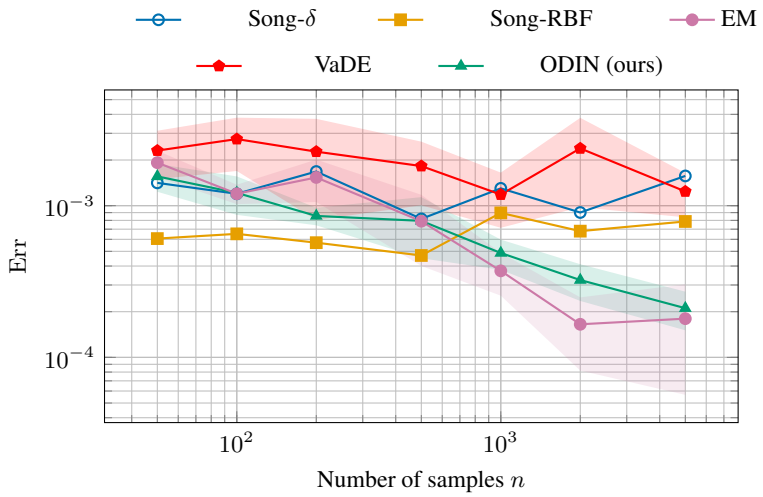


Figure 5: Recovery on the synthetic product-Bernoulli mixture.

Mushroom We use the *Mushroom* dataset from OpenML. All input variables are categorical, and the target has two classes, *e* (edible) and *p* (poisonous). We transform the dataset through a full one-hot encoding, without dropping any category. This produces a binary representation with 8124 samples and 116 features. In addition, we keep track of the one-hot groups associated with each original categorical variable, since these groups are useful when interpreting the resulting representation.

Summary Across all three benchmarks, the final input to every method is binary. This common representation is especially natural for ODIN, whose dictionary is built from binary interaction atoms, but it also ensures a uniform evaluation setting across all baselines.

F.3 One-hot encoding and the product-Bernoulli approximation

Modern tabular data are heterogeneous: features may be Boolean, categorical, ordinal, integer-valued, or continuous, and recent tabular foundation-model work has emphasized the importance of learning across such mixed feature types (Hollmann et al., 2025; Van Breugel & Van Der Schaar, 2024). For categorical variables, a standard preprocessing step is to replace each category by binary indicator variables through dummy or one-hot encoding; Boolean variables are already in this form, and other discrete variables can be binarized when an event-based representation is desired (Poslavskaya & Korolev, 2023). Thus a broad class of discrete observations can be represented in $\{0, 1\}^d$.

The *mushroom* dataset requires a categorical-to-binary transformation. For each original categorical variable C_g, with category set $\mathcal{A}_g$, we introduce one binary indicator

$$X_{g,a} = \mathbf{1}\{C_g = a\}, \qquad a \in \mathcal{A}_g.$$

The resulting indicators satisfy the structural one-hot constraint

$$\sum_{a \in \mathcal{A}_g} X_{g,a} = 1$$

Feature	ROI name
F1	Short Axis Apical — ANT
F2	Short Axis Apical — LAT
F3	Short Axis Apical — INF
F4	Short Axis Apical — SEPTAL
F5	Short Axis Mid — ANT
F6	Short Axis Mid — ANT-LAT
F7	Short Axis Mid — INF-LAT
F8	Short Axis Mid — INF
F9	Short Axis Mid — ANT-SEPTAL
F10	Short Axis Basal — ANT
F11	Short Axis Basal — ANT-LAT
F12	Short Axis Basal — INF-LAT
F13	Short Axis Basal — INF
F14	Short Axis Basal — ANT-SEPTAL
F15	Horizontal Long Axis — APICAL
F16	Horizontal Long Axis — LAT
F17	Horizontal Long Axis — BASAL
F18	Horizontal Long Axis — SEPTAL
F19	Vertical Long Axis — ANT
F20	Vertical Long Axis — APICAL
F21	Vertical Long Axis — INF
F22	Vertical Long Axis — BASAL

Table 4: Anatomical meaning of the 22 binary SPECT Heart features used in our analysis. The UCI/OpenML dataset exposes these variables as partial diagnoses $F1, \ldots, F22$; the anatomical ROI names follow the slice-and-ROI definition in Fig. 3 of Kurgan et al. (2001).

Dataset	Samples	Features	Classes
Congressional Voting Records	232	16	2
SPECT Heart	267	22	2
Mushroom	8124	116	2

Table 5: Main characteristics of the real-data benchmark after preprocessing. Class labels are *democrat*/*republican* for Congressional Voting Records, $0/1$ for SPECT Heart, and *edible*/*poisonous* for Mushroom.

for every sample and every original variable g. In particular, two different indicators from the same categorical group cannot be simultaneously active:

$$X_{g,a}X_{g,b} = 0, \qquad a \neq b.$$

This constraint is important when interpreting the product-Bernoulli mixture model. A product-Bernoulli component treats the binary coordinates as conditionally independent given the latent component. If applied literally to a one-hot group, it would assign positive probability to impossible states in which two categories of the same original variable are active at the same time. Thus, for one-hot categorical data, the product-Bernoulli mixture should be viewed as a relaxed binary approximation to an underlying product-categorical mixture, rather than as a correctly specified model over the full binary hypercube.

This misspecification affects only the mixture-inversion stage, not the ODIN dictionary construction itself. ODIN operates on empirical co-activation events $\phi_S(x) = \prod_{i \in S} x_i$. Atoms containing two mutually exclusive categories from the same one-hot group have zero empirical support and are therefore removed by the support filter. Consequently, the selected ODIN atoms mainly represent meaningful cross-variable category co-occurrences, such as the joint presence of one cap-shape category and one odor category, rather than impossible within-variable co-activations.

The moment-matching step then fits a product-Bernoulli mixture to the selected binary interaction moments. This is not identical to fitting a product-categorical mixture: the latter would enforce exactly one active category per original variable inside each component. Nevertheless, the selected

ODIN moments avoid the most direct within-group contradictions because unsupported one-hot violations are not included as informative atoms. We therefore interpret the *mushroom* experiment as testing whether ODIN interactions provide a useful finite binary lifting for categorical data after one-hot encoding, while acknowledging that the fitted product-Bernoulli mixture is an approximation to the categorical generative structure.

G EXPERIMENTAL DETAILS

Post-hoc KL evaluation. For the label-free KL score, all methods are evaluated through the same input-space product-Bernoulli mixture, independently of their internal parametrization. Given predicted assignments $\hat{c}_i$, we estimate $\hat{\pi}_h = n_h/n$ and $\hat{\mu}_{h,j} = n_h^{-1} \sum_{i:\hat{c}_i=h} X_{ij}$, with smoothing $\varepsilon = 10^{-12}$ to avoid probabilities equal to 0 or 1. This defines $p^{\mathrm{est}}(x) = \sum_{h=1}^{m} \hat{\pi}_h \prod_{j=1}^{d} \hat{\mu}_{h,j}^{x_j}(1 - \hat{\mu}_{h,j})^{1-x_j}$. We compute $\mathrm{KL}(\hat{p}\|p^{\mathrm{est}})$ on the empirical support of the observed data, namely $\mathrm{KL}(\hat{p}\|p^{\mathrm{est}}) = -\sum_x \hat{p}(x) \log p^{\mathrm{est}}(x) - H(\hat{p})$. This common post-hoc protocol makes the distributional comparison independent of whether a method is probabilistic, graph-based, or representation-based.

Cluster numbers. Although the external labels are binary, we also evaluate over-clustering settings because *mushroom* and *spect* may contain latent structure not aligned with the provided binary labels. For the deep and graph-based baselines, we use $m \in \{4, 6, 10, 12\}$ clusters unless stated otherwise. The post-hoc KL is always computed using the clusters actually returned by the method; if a method returns fewer nonempty clusters, the post-hoc mixture uses those nonempty clusters.

VaDE. We use the **VaDE** implementation from `clustpy.deep` Leiber et al. (2023). VaDE learns a variational autoencoder together with a Gaussian mixture prior in latent space. In the real-data experiments, we use embedding dimension 2, batch size 128, 60 autoencoder pretraining epochs, and 200 clustering epochs. The pretraining and clustering learning rates are 10^{-3} and 10^{-4}, respectively. The latent Gaussian mixture is initialized with a diagonal-covariance Gaussian mixture model using 20 initializations and covariance regularization 10^{-2}. For evaluation, we use the hard VaDE assignments and apply the common post-hoc Bernoulli KL protocol in the original binary input space, rather than using VaDE's latent Gaussian parameters.

ACeDeC. We use the **ACeDeC** implementation from `clustpy.deep` Leiber et al. (2023). ACeDeC learns a deep embedded representation designed to isolate cluster-relevant information from shared variation. We use embedding dimension 50, batch size 128, 100 pretraining epochs, and 300 clustering epochs. The pretraining and clustering learning rates are 10^{-3} and 10^{-4}, respectively. We use the `acedec` initialization and enable final reclustering. Predicted labels are saved for each run and evaluated using ACC, NMI, ARI, and the common post-hoc Bernoulli KL score.

DipEncoder. We use the **DipEncoder** implementation from `clustpy.deep`Leiber et al. (2023). DipEncoder encourages multimodal latent embeddings using a clustering loss based on Hartigan's Dip test. We use embedding dimension 50, batch size 128, 100 pretraining epochs, and 150 clustering epochs. The pretraining and clustering learning rates are 10^{-3} and 10^{-4}, respectively, and the clustering loss weight is set to 1.0. We save the predicted labels, learned embedding, projection axes, and method-specific thresholds when available. As for the other methods, the KL score is computed only from the hard predicted clusters through the common post-hoc Bernoulli mixture.

NeuNcut. We use the **NeuNcut** implementation based on a neural normalized-cut objective. The model is a multilayer perceptron with hidden dimensions $[512, 512]$, trained for 300 epochs with batch size 256. We use Adam with learning rate 5×10^{-3}, weight decay 10^{-4}, and cosine annealing. Inputs are normalized before training. For each minibatch, we build a full affinity matrix with bandwidth $\sigma = 1.0$, compute soft cluster assignments through the network, and optimize the normalized-cut loss together with an orthogonality regularization term weighted by $\gamma = 80$. Final clusters are obtained by taking the largest soft assignment. The resulting assignments are evaluated using the same ACC, NMI, ARI, and post-hoc KL protocol as the other baselines.

Method	Emb. dim.	Batch	Pretrain	Cluster epochs	LR pretrain	LR cluster
VaDE	2	128	60	200	10^{-3}	10^{-4}
ACeDeC	50	128	100	300	10^{-3}	10^{-4}
DipEncoder	50	128	100	150	10^{-3}	10^{-4}

Table 6: Hyperparameters used for the deep clustering baselines.

Method	Hidden dims.	Epochs	Batch	LR	Additional parameters
NeuNcut	$[512, 512]$	300	256	$5 \cdot 10^{-3}$	$\sigma = 1.0,\ \gamma = 80,\ \mathrm{wd} = 10^{-4}$

Table 7: Hyperparameters used for the NeuNcut baseline.

Reproducibility. For each method, dataset, and run, we save the predicted labels, encoded ground-truth labels, configuration file, fitted model or model weights, and evaluation metrics. For deep clustering baselines, we additionally save the learned embedding when available. For the post-hoc KL computation, we save the estimated mixture weights $\hat{\pi}_h$, component means $\hat{\mu}_h$, negative log-likelihood, empirical entropy, number of observed binary states, and final KL value.

H Dictionary-learning diagnostics

H.1 KL drop across ODIN layers

ODIN constructs the dictionary progressively by adding interaction atoms of increasing order. At order k, the unrestricted number of possible monomial interactions among d binary variables is $\binom{d}{k}$. A direct search over all such interactions is therefore combinatorial. ODIN avoids this full enumeration by growing the dictionary hierarchically: candidate order-k atoms are generated from atoms retained at order $k-1$, and are then filtered by empirical support and information gain.

To diagnose this construction, we track the residual KL error after each layer. Let $D^{(k)}$ denote the cumulative dictionary after completing layer k. The residual error is

$$L(D^{(k)}) = \inf_{q \in \mathcal{M}_{D^{(k)}}} \mathrm{KL}(\hat{p} \,\|\, q),$$

where $\hat{p}$ is the empirical distribution and $\mathcal{M}_{D^{(k)}}$ is the log-linear model supported on the selected atoms. Since $D^{(k)} \subseteq D^{(k+1)}$, the model class only expands with k, hence $L(D^{(k)})$ is non-increasing across layers.

Figure 6 shows this residual KL trajectory for the Vote and SPECT datasets. In both cases, the first layers exhibit a plateau: unary atoms and the earliest low-order interactions do not yet explain the main dependence structure of the data. The sharp decrease occurs only once ODIN adds higher-order interactions. For Vote, most of the KL reduction happens around the fourth layer, after which the residual error becomes nearly zero. For SPECT, the drop is more gradual, with substantial improvement around layers four and five, followed again by a near-saturation regime.

This behavior is important for the interpretation of the dictionary. The selected atoms are not merely numerous; they are added at the layers where they produce a measurable reduction in residual dependence. The flattening of the curve at later layers indicates that additional high-order atoms contribute only marginal information. Thus, the KL trajectory provides an empirical stopping diagnostic: once the residual KL has stabilized, further expansion of the dictionary mainly increases complexity without substantially improving the log-linear approximation.

The KL trajectory also gives a practical stopping diagnostic. Although the maximum order can be set to $K_{\max} = d$, this value is only an upper bound on the hierarchy. The effective depth of the dictionary is data-dependent. At each layer, ODIN continues only if there exist admissible atoms with sufficient empirical support and non-negligible information gain. Thus the procedure stops early when no new atoms survive the support and score filters.

This is reflected in Figure 6. The residual KL of the dictionary-supported log-linear model decreases across layers and becomes nearly zero at the final retained layers. Hence the selected dictionary

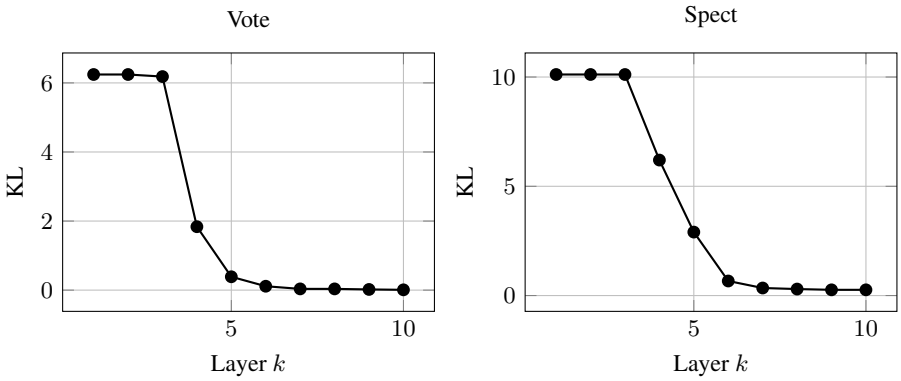


Figure 6: Evolution of the residual KL divergence across ODIN layers for the Vote and SPECT datasets. The decrease measures the amount of empirical dependence explained by the cumulative dictionary $D^{(k)}$.

already provides an accurate approximation of the empirical distribution before the full combinatorial hierarchy is explored. This criterion is consistent with the structural diagnostics in Figures 7 and 8: the number of selected atoms decreases in the higher layers, and the remaining supports and proxy scores become small. In practice, we therefore set $K_{\max} = d$, but the algorithm terminates earlier when either no atoms are selected at a new layer or the decrease in residual KL becomes negligible.

H.2 Growth and statistics of the selected dictionary

A central difficulty in using interaction atoms is the combinatorial growth of the candidate space. For d binary variables, the full order-k layer contains $\binom{d}{k}$ possible monomial atoms. Thus, even for the Vote dataset, where $d = 16$, the number of possible interactions grows rapidly and peaks at the middle orders. Figure 7 compares this combinatorial envelope with the number $|L_k|$ of atoms retained by ODIN at each layer.

The figure shows that ODIN does not enumerate or retain the full order-k interaction layer. Instead, the dictionary grows through a hierarchical construction: order-k atoms are proposed only from previously retained lower-order atoms, and are then filtered by empirical support and proxy information score. Consequently, the effective layer sizes remain much smaller than $\binom{d}{k}$, especially in the middle orders where the unrestricted combinatorial class is largest. This does not mean that the combinatorial nature of the problem disappears: ODIN still has to proceed layer by layer. Rather, the point is that the empirical hierarchy prunes most branches of the interaction lattice before they can generate high-order candidates.

To understand which atoms survive this pruning, we also examine their empirical support and their proxy scores. For an atom S, its empirical support is

$$\widehat{\mathrm{supp}}(S) = \frac{1}{n}\sum_{t=1}^{n} \phi_S(X^{(t)}),$$

namely the empirical frequency with which the corresponding co-activation event occurs. Since $\phi_S(x) = 1$ requires all variables in S to be active simultaneously, support naturally decreases with the order of the atom. This is visible in Figure 8: the largest supports occur among the low-order atoms, especially the singleton and pairwise layers, whereas higher-order atoms are much rarer.

However, high support is not the same as high information. The proxy score measures the residual gain obtained by adding an atom to the current dictionary; it is therefore a conditional quantity, evaluated after the effects already captured by the existing dictionary have been taken into account. An atom may be frequent but redundant, while a less frequent interaction may still explain a residual dependency not captured by lower-order atoms. The score plot confirms this distinction: proxy scores are concentrated on a relatively small subset of atoms, and most later atoms have small but nonzero scores. Thus ODIN is not selecting atoms merely because they are common events; it selects atoms that survive both an empirical support filter and an information-based relevance filter.

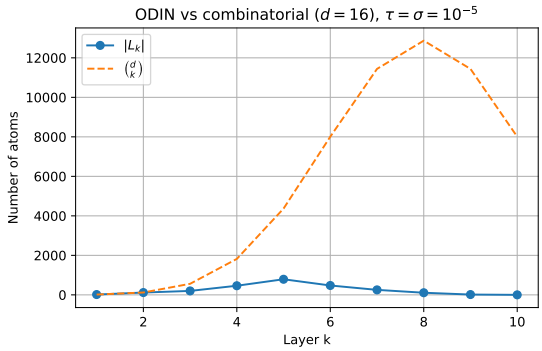
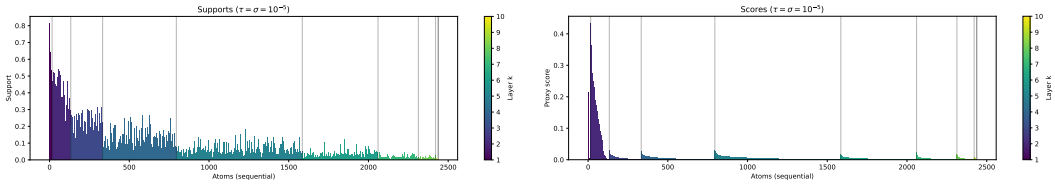


Figure 7: Layerwise dictionary size on the Vote dataset. The dashed curve is the full number $\binom{16}{k}$ of order-k interactions, while the solid curve is the number $|L_k|$ of atoms retained by ODIN after hierarchical generation and pruning.



(a) Empirical support

(b) Proxy score

Figure 8: Atom-level diagnostics for the ODIN dictionary on the Vote dataset, with atoms ordered sequentially by the construction procedure and colored by layer. Low-order atoms have the largest empirical support, but support alone does not determine informativeness. The proxy scores are sparse, indicating that only a subset of the supported atoms contributes substantial residual information.

The two thresholds play complementary roles. The support threshold removes atoms whose empirical moments are too rare to be reliable, while the score threshold removes atoms that occur in the data but add negligible residual information. The sensitivity plots in Figure 9 show that, over the tested range of threshold values, the qualitative layer profile remains stable: the dictionary grows up to intermediate orders and then decreases as higher-order co-activations become too rare or insufficiently informative. This supports the use of $\tau = \sigma = 10^{-5}$ as a working choice in the Vote experiment: it is small enough to retain meaningful interactions, but still removes negligible supports and near-zero information gains.

H.3 Proxy and Computational Cost of Interaction Scoring

H.4 Accuracy and Computational Cost of Proxy Scoring

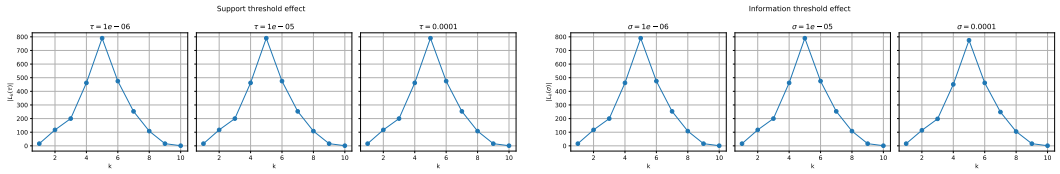
As described in Section 3, the reported ODIN dictionaries are constructed using $|J(S)|$, while $\mathrm{RI}_{D,S}$ provides the corresponding information-geometric reference criterion. Here we quantify the approximation induced by this computational substitution.

The distinction is important because the two scores contain different information. Exact refined information,

$$\mathrm{RI}_{D,S} = \mathcal{L}(D) - \mathcal{L}(D \cup \{S\}),$$

depends on the current dictionary D, whereas $J(S)$ depends only on the marginals indexed by subsets of S. Consequently, the proxy need not preserve the exact candidate ranking, particularly at higher interaction orders.

Empirical agreement between exact and proxy rankings. Enouen & Sugiyama (2026, Appendix B.6) report correlations between $J(S)$ and marginal refined information of 0.8741, 0.7563, and 0.6376 for dimensions $d = 3, 4, 5$, respectively. These results establish an empirical relationship



(a) Support-threshold sensitivity (b) Score-threshold sensitivity

Figure 9: Effect of the support and score thresholds on the number of atoms retained per layer. Across the tested values, the same qualitative pattern persists: ODIN concentrates atoms in the intermediate layers and prunes the high-order layers strongly.

between the two scores, but they do not directly determine whether the proxy selects the same leading candidates as exact refined information during greedy dictionary construction.

To assess ranking agreement more directly, we generate 100 random third-order log-linear distributions for each $d \in \{4, 5, 6\}$:

$$p_\theta(x) = \exp\left(\sum_{\substack{T \subseteq [d] \\ 1 \le |T| \le 3}} \theta_T \phi_T(x) - \psi(\theta)\right), \qquad \theta_T \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(-1, 1).$$

For every generated distribution, we compare the candidate rankings produced by optimization-based exact refined information and by the proxy score $|J(S)|$. We report the Pearson correlation between the scores, the Spearman correlation between their rankings, the proportion of trials in which the highest-ranked candidate agrees, and the mean overlap between the two top-three candidate sets.

For pairwise atoms, the proxy closely tracks exact refined information: Pearson correlations range from 0.9608 to 0.9833, and the leading candidate agrees in 93% to 100% of the trials. The agreement is substantially weaker for third-order atoms. Their Pearson correlations range from 0.3962 to 0.4636, while Top-1 agreement decreases from 48% at $d = 4$ to 27% at $d = 6$. The mean Top-3 overlap similarly decreases from 0.807 to 0.413.

We therefore present $|J(S)|$ as a tractable screening heuristic, rather than as a theoretically or empirically exact replacement for $\mathrm{RI}_{D,S}$.

Layerwise computational profile. Figure 10 shows the computational profile of this procedure on the Vote dataset. The left panel reports the total proxy-scoring complexity per layer, measured in FLOPs, together with the number of scored atoms. The right panel reports the empirical runtime per layer, again together with the number of scored atoms. Both curves peak at intermediate orders. This matches the structure of the interaction lattice: early layers contain few possible atoms, while very high-order atoms are strongly pruned by support and score thresholds. The main computational burden therefore appears in the middle layers, where the admissible candidate set is largest.

The difference between exact and proxy scoring is substantial. In our benchmark, exact refined-information scoring costs approximately 354.31 seconds per atom and 4.49×10^8 FLOPs per atom. In contrast, proxy scoring costs approximately 3.89×10^{-4} seconds per atom and 3.28×10^3 FLOPs per atom. Thus the proxy is approximately 9.1×10^5 times faster in wall-clock time and 1.37×10^5 times cheaper in FLOPs per atom.

Consequently, exact refined information supplies the information-geometric target, while $|J(S)|$ supplies a tractable ordering heuristic.

Compute environment. All experiments were run on a compute node of an institutional computing cluster. The node used Ubuntu 22.04.5 LTS with Linux kernel 5.15.0. It was equipped with two Intel Xeon Gold 6240R CPUs at 2.40GHz, providing 48 physical cores and 96 hardware threads in total, and 187GiB of system memory. GPU experiments used a single NVIDIA Quadro RTX 8000 GPU with 48GB of VRAM. The installed NVIDIA driver was 595.58.03. The PyTorch environment used torch 1.9.0 with CUDA 10.2 runtime support; no local `nvcc` compiler was required. Intermediate outputs, checkpoints, CSV result files, and generated figures were stored on the cluster filesystem.

Order	d	Pearson	Spearman	Top-1	Top-3 overlap
2	4	0.9833	0.9446	1.000	0.940
2	5	0.9753	0.9137	0.980	0.957
2	6	0.9608	0.8869	0.930	0.940
3	4	0.4636	0.3880	0.480	0.807
3	5	0.4349	0.3773	0.360	0.547
3	6	0.3962	0.3697	0.270	0.413

Table 8: Empirical agreement between optimization-based exact refined-information scoring and the proxy score $|J(S)|$. Results are computed over 100 randomly generated third-order log-linear distributions for each dimension. Pearson measures score correlation, Spearman measures rank correlation, Top-1 is the proportion of trials in which the leading candidate agrees, and Top-3 overlap is the mean fraction of shared candidates between the two top-three sets.

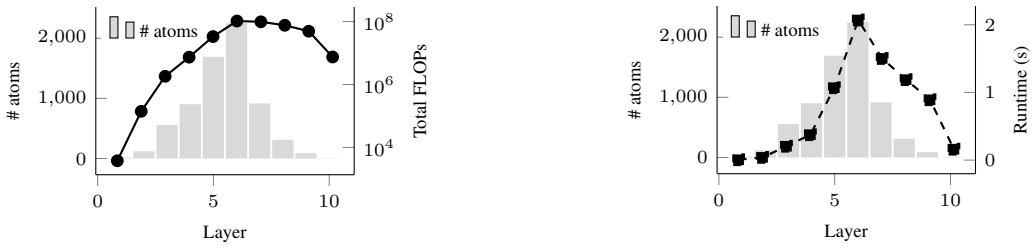


Figure 10: **Computational complexity across ODIN layers on the Vote dataset.** Left: total proxy-scoring complexity per layer, measured in FLOPs, together with the number of scored atoms. Right: empirical runtime per layer together with the number of scored atoms. The cost peaks at intermediate orders, where the number of admissible higher-order candidate interactions is largest.

I LIFTED MOMENT REPRESENTATION DIAGNOSTICS

We complement the quantitative clustering results with two-dimensional visual diagnostics of the learned representations. These plots are intended only as qualitative diagnostics: t-SNE preserves local neighborhoods rather than global geometry, so distances across panels should not be interpreted quantitatively. Nevertheless, the visualizations help illustrate how different methods organize the data internally, and whether the resulting representation is aligned with the external binary labels or with the clusters inferred by the method.

Figure 11 colors the embeddings by the available external labels. On *vote*, the external partition is visibly well aligned with the learned representations for several methods, reflecting the strong two-party structure of the dataset. The picture is less direct on *mushroom* and *spect*. In these datasets, the external binary label is only one possible coarse summary of the data, and the learned latent structure need not be organized along that single annotation. This is especially important for *spect*, where the classes remain substantially mixed in the two-dimensional projections.

Figure 12 instead colors the same embeddings by the cluster assignments produced by each method. Comparing Figures 11 and 12 separates two questions: whether a representation visually separates the external labels, and whether the method itself produces a stable partition of its latent space. The difference is important because agreement with the external label is not the only relevant criterion on real data: the inferred mixture components may capture substructure that is not identical to the available binary annotation.

Finally, Figures 13 show how the ODIN representation evolves as higher-order dictionary layers are added. These diagnostics illustrate that ODIN does not define a fixed black-box embedding. Instead, the representation is progressively enriched by adding selected interaction atoms. Later layers therefore correspond to moment coordinates with increasingly high-order co-activation structure. The visual evolution should be read together with the rank and recovery diagnostics in the main text: the role of the additional layers is not necessarily to improve two-dimensional label separation, but to provide additional observable moments for the lifted inverse problem.

Method	Avg. time per atom (s)	Avg. FLOPs per atom
ODIN-Exact	354.31	4.49×10^8
ODIN-Proxy	3.89×10^{-4}	3.28×10^3

Table 9: Average per-atom scoring cost for exact refined-information scoring and proxy scoring. Exact scoring is information-geometrically principled but computationally infeasible as a default layerwise scoring rule. Proxy scoring makes the hierarchical search practical.

J Recovery and optimization diagnostics

J.1 Optimization protocol and restart diagnostics

The ODIN-to-mixture inversion step solves a finite, non-convex moment-matching problem. For a fixed ODIN dictionary D and a fixed number of mixture components m, we minimize

$$\mathcal{L}_{D,m}(\eta) = \sum_{S \in D} \left(\widehat{M}_S - \sum_{h=1}^{m} \pi_h \prod_{i \in S} \mu_{h,i} \right)^2,$$

where

$$\widehat{M}_S = \frac{1}{n} \sum_{t=1}^{n} \phi_S(X^{(t)})$$

is the empirical lifted moment associated with atom S. The optimization is performed in unconstrained coordinates

$$\pi_h = \frac{\exp(\alpha_h)}{\sum_{\ell=1}^{m} \exp(\alpha_\ell)}, \qquad \mu_{h,i} = \sigma(\beta_{h,i}),$$

which enforces the simplex constraint on π and the box constraint $\mu_{h,i} \in (0,1)$ throughout optimization.

Because $\mathcal{L}_{D,m}$ is non-convex, the result can depend on initialization. For every dataset and every value of m, we therefore run Adam from ten independent random initializations and retain the run with the smallest final moment-matching loss. The initial parameters are sampled as

$$\pi^{(0)} \sim \mathrm{Dirichlet}(1, \ldots, 1), \qquad \mu_{h,i}^{(0)} \sim \mathrm{Unif}(0.1, 0.9).$$

The selected minimum-loss run is then used for posterior assignments, clustering metrics, and the embedding visualizations reported in the appendix.

The optimization settings used in the real-data experiments are:

Dataset	m values	dictionary	steps	learning rate
vote	$\{2, 4, 6, 8, 10, 12\}$	$D^{(10)}$	500	0.03
mushroom	$\{2, 4, 6, 8, 10, 12\}$	$D^{(9)}$	2000	0.10
spect	$\{2, 4, 6, 8, 10, 12\}$	$D^{(14)}$	2000	0.03

Here $D^{(k)}$ denotes the cumulative ODIN dictionary through layer k, while D denotes the final selected dictionary when no truncation is used. The larger number of optimization steps for *mushroom* and *spect* reflects the larger lifted moment systems used for these datasets.

Figure 15 reports the optimization trajectories on *spect* across several values of m. Each curve corresponds to one random restart. The loss decreases rapidly in the early iterations and then enters a slower refinement regime. Across the tested values of m, the best final losses are comparable. Thus, for this dictionary and dataset, increasing m beyond a moderate value does not substantially reduce the fitted lifted-moment error. This behavior is consistent with the non-convexity of the inversion problem: different initializations may converge to distinct stationary points, but the best attainable losses remain similar over the tested range of component numbers.

Figure 14 complements this restart diagnostic by showing the loss and gradient norm for representative runs on the *vote* dataset. The loss decreases quickly at the beginning of optimization, while the gradient norm also decays and eventually stabilizes at a small value. This provides a direct optimization check that the selected solutions are close to first-order stationary points of the lifted moment-matching objective.

Method	*vote*				*mushroom*				*spect*			
	ACC ↑	NMI ↑	ARI ↑	KL ↓	ACC ↑	NMI ↑	ARI ↑	KL ↓	ACC ↑	NMI ↑	ARI ↑	KL ↓
NeuNcut	0.858 ± 0.000	0.409 ± 0.000	0.510 ± 0.000	2.702 ± 0.000	0.608 ± 0.000	0.035 ± 0.000	0.046 ± 0.000	$\mathbf{16.040 \pm 2.622}$	0.764 ± 0.000	0.102 ± 0.000	$\mathbf{0.216 \pm 0.000}$	7.314 ± 0.000
ACeDeC	0.885 ± 0.008	0.509 ± 0.034	0.591 ± 0.025	$\mathbf{2.635 \pm 0.000}$	0.712 ± 0.086	0.168 ± 0.074	0.206 ± 0.091	20.290 ± 4.728	0.572 ± 0.018	0.070 ± 0.012	-0.083 ± 0.000	7.069 ± 0.000
DipEncoder	$\mathbf{0.892 \pm 0.002}$	0.519 ± 0.002	0.612 ± 0.005	2.646 ± 0.000	0.610 ± 0.047	0.052 ± 0.023	0.056 ± 0.025	17.619 ± 3.095	0.693 ± 0.000	0.020 ± 0.037	-0.015 ± 0.067	7.834 ± 0.000
EM	0.884 ± 0.000	0.511 ± 0.000	0.587 ± 0.000	$\mathbf{2.635 \pm 0.000}$	0.731 ± 0.000	0.309 ± 0.000	0.212 ± 0.000	21.747 ± 0.237	0.633 ± 0.000	$\mathbf{0.119 \pm 0.000}$	0.068 ± 0.000	$\mathbf{6.836 \pm 0.000}$
VaDE	0.846 ± 0.003	0.396 ± 0.003	0.477 ± 0.009	2.761 ± 0.101	0.592 ± 0.010	0.027 ± 0.003	0.034 ± 0.007	24.194 ± 0.094	0.618 ± 0.000	0.062 ± 0.002	0.053 ± 0.000	7.542 ± 0.536
ODIN	$\mathbf{0.892 \pm 0.000}$	$\mathbf{0.526 \pm 0.000}$	$\mathbf{0.614 \pm 0.000}$	2.683 ± 0.006	$\mathbf{0.837 \pm 0.035}$	$\mathbf{0.391 \pm 0.097}$	$\mathbf{0.460 \pm 0.091}$	20.098 ± 0.990	$\mathbf{0.794 \pm 0.000}$	0.072 ± 0.012	-0.096 ± 0.002	7.005 ± 0.046

Table 10: Extended real-data clustering results for $m = 2$, matching the number of external binary labels. Values are reported as mean $\pm$ standard deviation. Higher is better for ACC, NMI, and ARI; lower is better for KL. KL is computed as a label-free post-hoc product-Bernoulli fit score: hard assignments are converted into cluster proportions and within-cluster Bernoulli means, and $\mathrm{KL}(\hat{p} \| p^{\mathrm{est}})$ is evaluated on the empirical support of the observed binary data. Best values in each column are bolded.

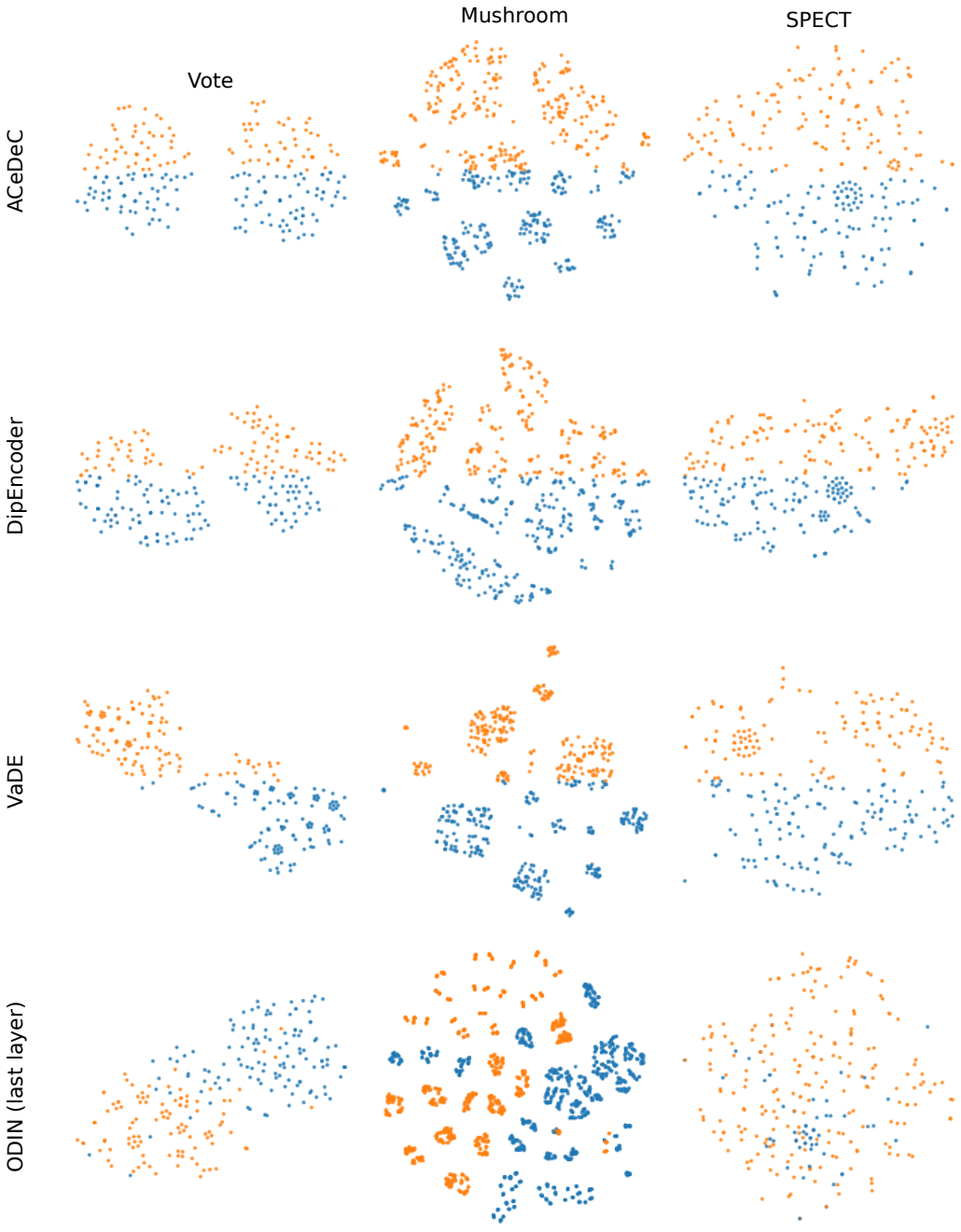


Figure 11: t-SNE projections of the learned representations on the three real datasets, colored by the available external binary labels. Columns correspond to *vote*, *mushroom*, and *spect*; rows correspond to ACeDeC, DipEncoder, VaDE, and ODIN. These plots evaluate visual alignment between each learned representation and the external annotation. The strong separation on *vote* contrasts with the weaker alignment on *mushroom* and especially *spect*, where the external label need not coincide with the dominant latent mixture structure.

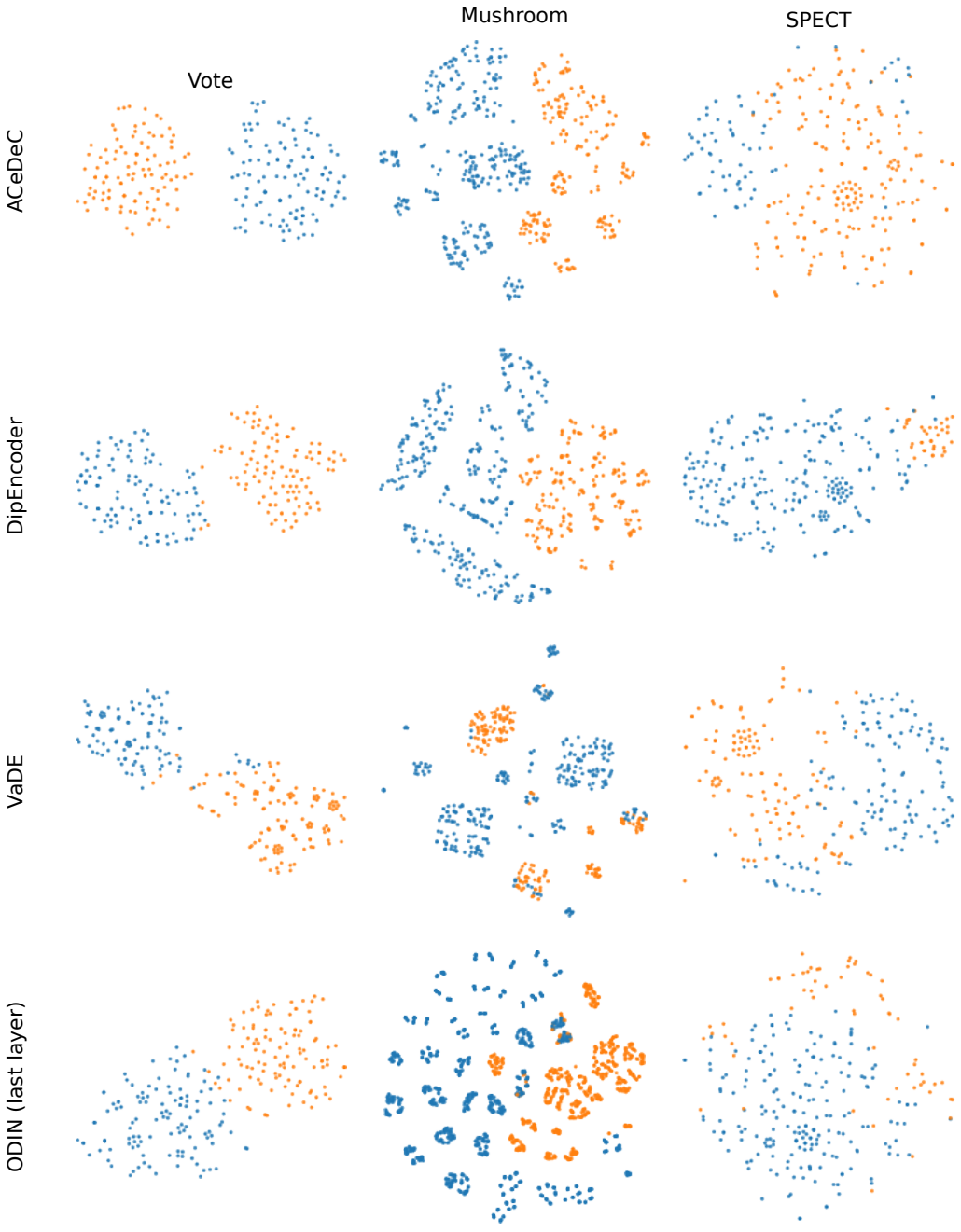
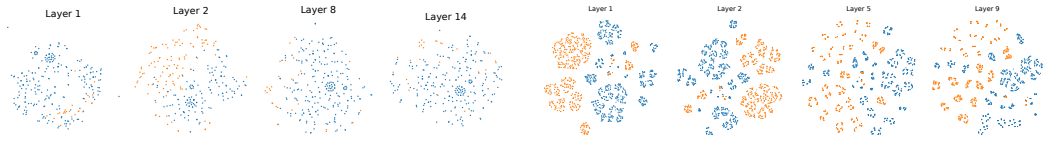


Figure 12: t-SNE projections of the same learned representations, colored by the cluster assignments produced by each method. Columns correspond to *vote*, *mushroom*, and *spect*; rows correspond to ACeDeC, DipEncoder, VaDE, and ODIN. Unlike Figure 11, this figure shows the partition induced by the method itself. The comparison highlights cases where the inferred clusters form coherent regions in the learned representation even when they do not exactly match the external binary labels.

J.2 Dimensional precondition for local identifiability

The local identifiability theory in Section 5 is governed by the rank of the lifted moment Jacobian. For a mixture with m components in d binary dimensions, the number of free parameters is

$$p_{m,d} = (m-1) + md,$$



(a) *SPECT* (b) *Mushroom*

Figure 13: Layerwise evolution of the ODIN representation through t-SNE projections of cumulative dictionaries at selected layers. Points are colored by the external dataset labels. As ODIN adds higher-order interaction atoms, the geometry of the lifted representation changes substantially. On *SPECT*, the external binary labels remain relatively mixed across layers, suggesting that the selected interaction structure is not fully aligned with the provided annotation. On *Mushroom*, the progressive enrichment of the dictionary produces a much clearer organization of the data in the lifted space.

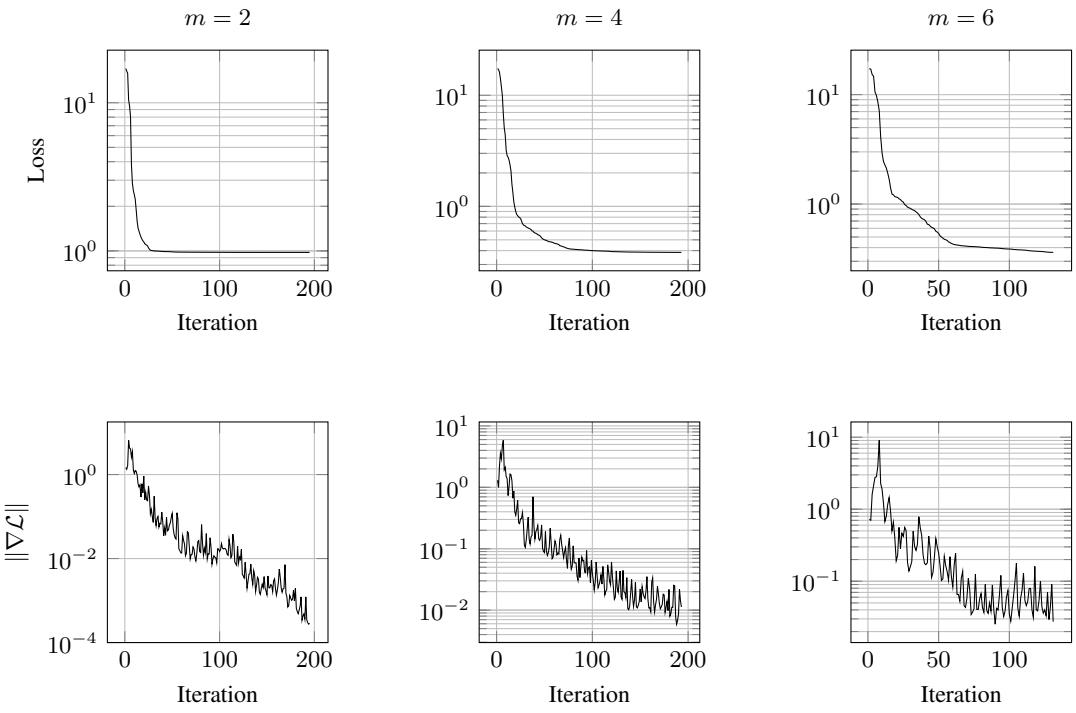


Figure 14: **Optimization diagnostics for ODIN moment matching on the *vote* dataset.** Columns correspond to different numbers of mixture components $m \in \{2, 4, 6\}$. Top row: value of the lifted moment-matching objective as a function of the iteration number. Bottom row: corresponding gradient norm. All y-axes are shown on a logarithmic scale. The trajectories exhibit a rapid initial decrease, followed by a slower convergence phase and a final plateau where the gradient norm is small. This is consistent with convergence to a first-order stationary point of the non-convex ODIN moment-matching objective.

where $m - 1$ accounts for the simplex-constrained mixture weights and md for the component Bernoulli parameters. For a dictionary D, the reduced Jacobian $\mathcal{J}_D(\eta)$ has $|D|$ rows and $p_{m,d}$ columns. Therefore a necessary condition for full column rank is

$$|D| \geq p_{m,d} = (m - 1) + md.$$

This condition is not sufficient: even when the number of selected atoms exceeds the number of free parameters, the rows of the Jacobian may still be linearly dependent at the fitted solution. Nevertheless, it provides a simple experimental sanity check. If $|D| < (m - 1) + md$, then the lifted moment system is dimensionally underdetermined and local identifiability cannot hold. If

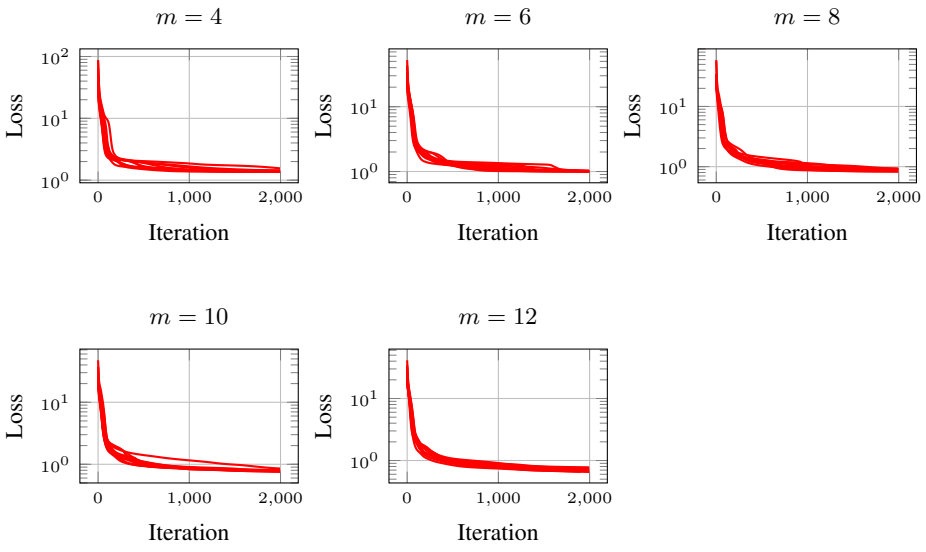


Figure 15: **Optimization convergence on SPECT.** Loss trajectories across ten random restarts for different numbers of mixture components m. Each curve corresponds to one restart.

$|D| \geq (m-1) + md$, then local identifiability is at least dimensionally possible, although it still depends on the numerical rank of $\mathcal{J}_D(\hat{\eta})$.

Dataset	m	$p_{m,d} = (m-1) + md$	first feasible layer $k_{\min}$	$\lvert D^{(k_{\min})}\rvert$
vote	2	33	2	133
vote	4	67	2	133
vote	6	101	2	133
vote	8	135	3	333
vote	10	169	3	333
vote	12	203	3	333
spect	2	45	2	248
spect	4	91	2	248
spect	6	137	2	248
spect	8	183	2	248
spect	10	229	2	248
spect	12	275	3	985
mushroom	2	233	2	3183
mushroom	4	467	2	3183
mushroom	6	701	2	3183
mushroom	8	935	2	3183
mushroom	10	1169	2	3183
mushroom	12	1403	2	3183

Table 11: Dimensional precondition for local identifiability across real datasets and numbers of mixture components. Here $p_{m,d} = (m-1)+md$ is the number of free parameters in an m-component product-Bernoulli mixture in d binary dimensions, and $k_{\min}$ is the first ODIN layer satisfying $|D^{(k)}| \geq p_{m,d}$. The unary layer is dimensionally insufficient in all three datasets, while early ODIN interaction layers provide enough lifted coordinates to make local identifiability dimensionally possible. This condition is necessary but not sufficient; full local identifiability still requires a full-column-rank Jacobian at the fitted solution.

The table 11 shows that the unary dictionary is dimensionally insufficient in all three real datasets. This is expected: for $m \geq 2$, the number of unary moments is only d, whereas the number of free mixture parameters is $(m-1) + md > d$. Thus first-order statistics alone cannot yield a locally identifiable product-Bernoulli mixture through moment matching. In contrast, the first ODIN

interaction layers rapidly expand the lifted moment system past this necessary threshold. For example, on *vote*, layer 2 is already sufficient for $m \leq 6$, while layer 3 is sufficient for all tested values up to $m = 12$. On *spect*, layer 2 is sufficient up to $m = 10$, and layer 3 is sufficient for $m = 12$. On *mushroom*, layer 2 already exceeds the dimensional threshold for all tested values of m.

This diagnostic should not be interpreted as proving local identifiability. The inequality $|D^{(k)}| \geq (m-1) + md$ only states that full column rank is not ruled out by dimension. The actual rank condition depends on the reduced Jacobian $\mathcal{J}_D(\hat{\eta})$. Nevertheless, the dimensional diagnostic provides a useful sanity check: it confirms that the ODIN hierarchy supplies the additional interaction moments needed before local inversion can even be possible.

J.3 Ablation: information-based selection versus random dictionaries

The full interaction family contains $2^d - 1$ nonempty atoms, so searching directly for a recovery-optimal dictionary is combinatorial. ODIN restricts this search through its hierarchical heredity rule and information-based screening. We therefore ask whether this structured restriction still produces useful moment systems compared with randomly selecting the same number of higher-order atoms.

On the synthetic mixture with $d = 6$ and $m = 4$, all methods contain the same 6 singleton and 15 pairwise moments. We then add $b \in \{6, 8, 10, 15, 20\}$ third-order moments, selected either by ODIN or uniformly at random from the available triples. Thus the comparison controls for dictionary size and changes only the selection mechanism. We evaluate Jacobian rank, smallest singular value $\sigma_{\min}(\mathcal{J}_D)$, and parameter-recovery error.

These results clarify the role of ODIN. Random interaction dictionaries can also provide enough independent equations for local recovery, so information selection is not required for full rank. Rather, ODIN provides a structured way to navigate an exponentially large interaction space while retaining an interpretable hierarchy of data-dependent moments. Its dictionary is also comparatively well conditioned within the matched-budget random ensemble, although refined-information selection is not explicitly optimized for mixture recovery.

J.4 Ablation: direct moment matching versus multiview spectral inversion

We compare two inversion procedures that use ODIN-selected interaction moments on the synthetic Bernoulli-mixture benchmark. The first is the direct procedure used in the main experiments: after selecting a dictionary D, we fit the mixture parameters by minimizing the lifted moment-matching objective

$$\min_{\eta} \sum_{S \in D} \left(\widehat{M}_S - \sum_{h=1}^{m} \pi_h \prod_{i \in S} \mu_{h,i} \right)^2 .$$

This procedure uses the selected ODIN atoms as a single lifted moment system and enforces the mixture constraints directly through the parameterization of π and μ.

The multiview variant instead splits the observed variables into three views, runs ODIN separately on each view, and constructs lifted view variables

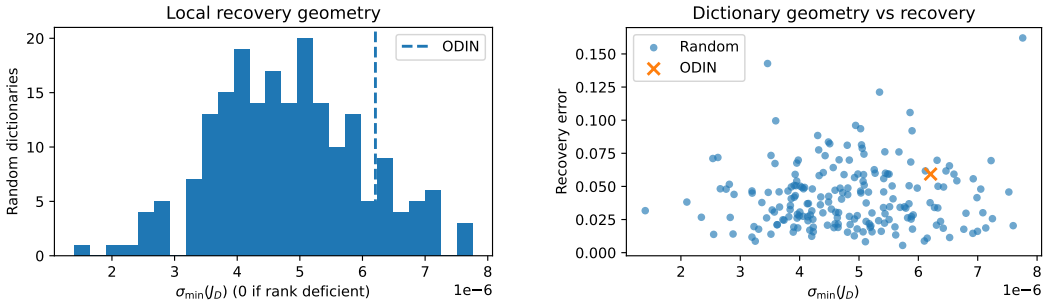
$$Z_a = z_{D_a}(X^{(a)}), \qquad a \in \{1, 2, 3\}.$$

It then estimates the empirical cross-moments

$$\widehat{M}_{ab} = \frac{1}{n} \sum_{t=1}^{n} Z_a^{(t)} (Z_b^{(t)})^\top, \qquad \widehat{T}_{123} = \frac{1}{n} \sum_{t=1}^{n} Z_1^{(t)} \otimes Z_2^{(t)} \otimes Z_3^{(t)}.$$

The pairwise moments are used to whiten the lifted views, and the whitened third-order moment is decomposed by a tensor power method. The resulting components are mapped back to lifted component means, and the unary coordinates are used to recover the Bernoulli parameters. Thus, unlike the direct method, the multiview variant relies on a spectral whitening and tensor-decomposition step after ODIN feature selection.

Figure 17 shows that direct ODIN inversion achieves lower recovery error across all tested sample sizes. The difference is consistent with the finite-sample conditioning of the two procedures. Direct optimization fits the same lifted moment equations used to define the inverse problem and can



(a) Distribution of $\sigma_{\min}(\mathcal{J}_D)$ for matched random dictionaries; the dashed line indicates ODIN.

(b) Jacobian conditioning versus recovery error.

Figure 16: **Local recovery geometry under matched-budget selection.** ODIN lies toward the better-conditioned region of the random ensemble, while its recovery error remains competitive rather than uniformly optimal.

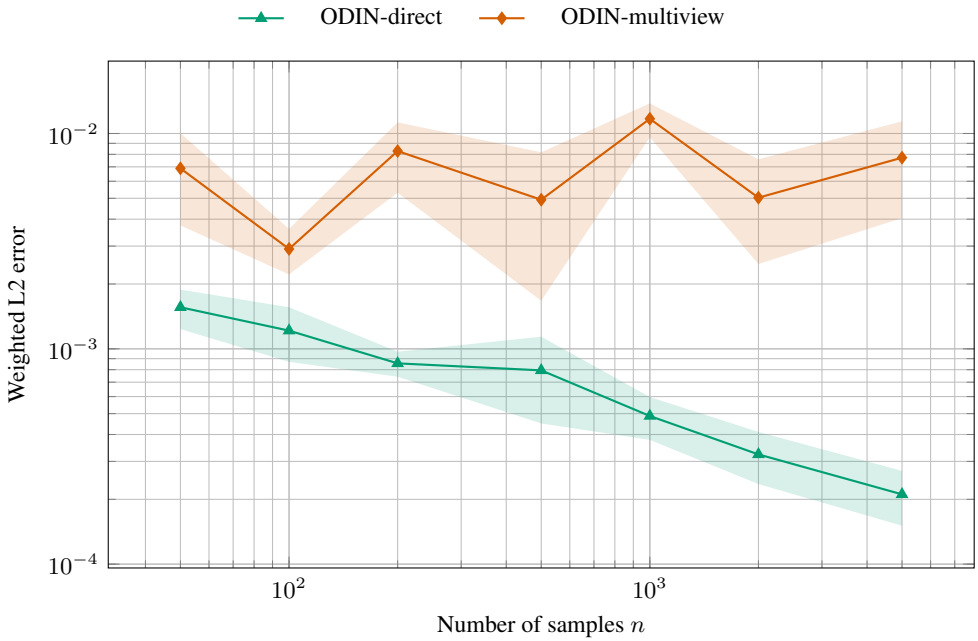


Figure 17: Comparison between the direct ODIN inversion procedure and the multiview variant on the synthetic Bernoulli-mixture benchmark. The plot reports the weighted parameter recovery error as a function of the number of samples. Shaded regions denote one standard deviation across independent runs.

exploit all selected atoms jointly. By contrast, the multiview variant first partitions the variables into separate views, learns separate view dictionaries, and then relies on empirical cross-moments, whitening, and tensor power iterations. Each of these steps can amplify sampling noise, especially when the cross-moment matrices are ill-conditioned or when the view-specific lifted spaces are only approximately aligned.

Therefore, the comparison should be interpreted as an ablation on the inversion step rather than as a rejection of multiview tensor methods in general. In this finite-sample synthetic setting, the ODIN-selected moments are more effectively used by direct constrained moment matching than by the spectral multiview pipeline.

K INTERPRETABILITY ANALYSIS

Non-uniqueness of mixture decompositions. Approximating the observed distribution and recovering its latent parameters are distinct objectives. Even exact knowledge of the distribution need not determine a unique mixture decomposition. For example, consider two product-Bernoulli mixtures with $m = 2$, $d = 2$, and equal weights $\pi = \pi' = (1/2, 1/2)$, whose component means are

$$\mu_1 = (0.2, 0.3), \quad \mu_2 = (0.8, 0.7),$$
$$\mu_1' = (0.1, 0.35), \quad \mu_2' = (0.9, 0.65).$$

These parameterizations differ even up to component permutation, yet both induce the same distribution:

$$p_\eta(x) = p_{\eta'}(x) = \begin{cases} 0.31, & x \in \{(0,0),(1,1)\}, \\ 0.19, & x \in \{(0,1),(1,0)\}. \end{cases}$$

Consequently, $\mathrm{KL}(p_\eta \| p_{\eta'}) = 0$ despite different component parameters. Equality of the observed laws also implies

$$\Xi_{D,m}(\eta) = \Xi_{D,m}(\eta') \qquad \text{for every interaction dictionary } D.$$

Thus the ambiguity persists even when all observable interaction moments are available. Each decomposition nevertheless specifies component activation probabilities that can be inspected and compared.

To formalize the interpretation of alternative fitted decompositions, fix a dictionary D and a number of components m, and define

$$\mathcal{G}_{D,m}(\hat{M}_D) := \arg\min_{\eta \in \Theta_m} F_D(\eta), \qquad F_D(\eta) = \|\hat{M}_D - \Xi_{D,m}(\eta)\|_2^2.$$

Every member of $\mathcal{G}_{D,m}(\hat{M}_D)$ achieves the best approximation to the selected empirical moments within the chosen mixture family. When several parameterizations attain this minimum, the objective alone cannot prefer one over another. Decompositions related only by a permutation of component labels are regarded as the same solution.

If the minimum is zero, every globally optimal decomposition satisfies

$$\hat{M}_S = \sum_{h=1}^{m} \pi_h \prod_{j \in S} \mu_{h,j}, \qquad S \in D.$$

For each atom S, the terms $\pi_h \prod_{j \in S} \mu_{h,j}$ express the contributions of the latent components to the probability of the corresponding co-activation event. Distinct optimal decompositions may therefore reproduce the same selected moments through different component contributions. They provide different latent organizations of the same observable interaction statistics. For an incomplete dictionary, agreement on these moments need not imply agreement on the entire distribution.

If the minimum is positive, global minimizers instead provide equally accurate approximations under the selected least-squares criterion; their fitted moment vectors need not coincide. Global optimality establishes that no admissible parameterization achieves a smaller objective value, while the attained residual measures how closely the selected empirical moments are reproduced. An optimal approximation can therefore remain inadequate when the chosen mixture family cannot represent the relevant data structure.

Subject to these distinctions, each globally optimal decomposition can be regarded as *one possible statistical reading of the data*. Its components describe activation profiles whose weighted contributions explain, or best approximate, the selected interaction moments. ODIN makes this reading inspectable because the coordinates used for estimation are explicit co-activation events. The same atoms can be used to describe which patterns characterize each component, how components differ, and how alternative decompositions organize the observed interactions. Interpretation concerns components with positive weights, since the parameters of a zero-weight component do not affect the fit.

Whether non-uniqueness occurs in the three real datasets studied here has not been established. Product-Bernoulli mixtures can be generically identifiable under appropriate conditions (Allman

et al., 2009), and our local recovery result depends on the rank of the selected moment Jacobian. On real data, however, the true generative parameters and correctness of the assumed mixture family are unknown. The one-hot representation of Mushroom illustrates a modeling mismatch, as discussed in Appendix F.3.

For a prescribed m, ODIN constructs its dictionary and estimates mixture parameters without using external class labels. Its fitted components therefore offer an unsupervised description of the data through selected observable interactions. These descriptions remain conditional on the fitted decomposition and may differ across equally well-fitting solutions. An atom-based interpretation is available for any fitted solution; describing that solution as globally optimal additionally requires establishing its membership in $\mathcal{G}_{D,m}(\hat{M}_D)$. The analysis below characterizes the returned decomposition through its component activation probabilities and enrichment scores.

Interpreting recovered components. Let $\hat{\eta} = ((\hat{\pi}_h, \hat{\mu}_h))_{h=1}^m$ and define the fitted activation

$$\hat{a}_{h,S} := \prod_{j \in S} \hat{\mu}_{h,j}.$$

For a monomial atom S, this quantity is the fitted probability that all binary indicators in S are simultaneously active under component h. In the SPECT experiments, the atoms therefore correspond to explicit anatomical co-activation events among short-axis (SA), horizontal long-axis (HLA), and vertical long-axis (VLA) indicators.

By Proposition 1, the atom-wise contribution of S to the separation of two fitted components h, ℓ is

$$\Delta_{h,\ell}(S) := (\hat{a}_{h,S} - \hat{a}_{\ell,S})^2.$$

Thus $\Delta_{h,\ell}(S)$ ranks atoms by how strongly they distinguish two specific fitted components.

For component-wise interpretation, assuming $0 < \hat{\pi}_h < 1$, we use the signed one-vs-rest enrichment

$$\hat{a}_{-h,S} := \frac{\sum_{\ell \neq h} \hat{\pi}_\ell \hat{a}_{\ell,S}}{1 - \hat{\pi}_h}, \qquad \hat{E}^{\mathrm{ovr}}_{h,S} := \hat{a}_{h,S} - \hat{a}_{-h,S}.$$

Atoms with large positive $\hat{E}^{\mathrm{ovr}}_{h,S}$ have higher joint activation probability in component h than in the remaining fitted mixture. We therefore interpret component h by ranking atoms in decreasing order of $\hat{E}^{\mathrm{ovr}}_{h,S}$.

These component summaries are computed directly from the fitted activation probabilities in the selected dictionary. If an atom such as

$$\{\text{SA Apical ANT}, \text{SA Mid ANT}\}$$

has large positive one-vs-rest enrichment for component h, then the fitted mixture assigns substantially higher probability to this anterior co-activation event in component h than in the rest of the mixture. Similarly, atoms involving apical, septal, inferior, lateral, or long-axis indicators are read as localized anatomical interaction patterns. The component labels below are therefore not external annotations; they are summaries of the fitted moment coordinates themselves.

SPECT component signatures. Figures 18– 20 show the resulting component signatures for SPECT at ODIN layer 14. For $m = 2$, the model separates a large weakly specific component from a small component characterized by high-order apical–inferior and basal-inferior interactions. The small component is defined by atoms such as SA Apical ANT, SA Apical INF, and SA Basal INF, indicating a localized co-activation pattern rather than a diffuse marginal effect.

For $m = 4$, this coarse decomposition is refined into broader anatomical patterns. One component is dominated by anterior short-axis indicators across apical, mid, and basal regions, giving an anterior-wall signature. Another component is characterized by inferolateral and basal indicators. The smaller components are more interaction-driven: for example, one combines anterior and inferolateral atoms, while another is sharply enriched for apical/septal and long-axis apical co-activations.

For $m = 8$, the same structure is further decomposed into finer subtypes. Several components are characterized by high-specificity second- and higher-order atoms, including inferior–basal, septal,

apical–lateral, anterior–inferolateral, and apical/septal long-axis patterns. The smallest components have the sharpest signatures: their top atoms have fitted activation near one inside the component and near zero in the rest of the mixture. This indicates that these components are defined by specific anatomical co-activation events.

The main point is that ODIN controls the representation used for recovery. The mixture is not fitted in an opaque embedding space. It is fitted in a finite dictionary D of observable interaction moments,

$$\widehat{M}_S \approx \sum_{h=1}^{m} \hat{\pi}_h \prod_{j \in S} \hat{\mu}_{h,j}, \qquad S \in D.$$

Consequently, we know which aspects of the data enter the inverse problem: single-region activations and selected higher-order anatomical co-activations. The recovered components can be inspected through the same atoms that define the moment-matching objective. In this sense, the learned partition is interpretable by construction: the model exposes which observable interactions are used to separate the fitted components.

L Additional comparison with recent tabular clustering methods

Motivation. The baselines in Section 6.2 include classical mixture estimation and generic deep clustering methods. To further contextualize ODIN with respect to recent methods developed specifically for tabular data, we add two supplementary baselines: TableDC (Rauf et al., 2025) and ZEUS (Marszał ek et al., 2025).

TableDC is a deep clustering method designed for tabular data and learns a dataset-specific latent representation together with the clustering structure. ZEUS instead follows a zero-shot representation-learning paradigm: it uses a pretrained model to produce embeddings for previously unseen tabular datasets without target-dataset fine-tuning (Marszał ek et al., 2025). These two methods therefore provide complementary modern reference points: TableDC represents a dataset-specific learned tabular embedding, whereas ZEUS represents a pretrained zero-shot embedding.

TableDC and ZEUS construct latent embedding spaces optimized for unsupervised separation, while ODIN constructs an explicit finite dictionary of observable binary interaction moments whose coordinates are subsequently used in the mixture inversion problem. The comparison therefore addresses a narrower empirical question: whether the clustering performance obtained from the interpretable ODIN lifting remains competitive with recent tabular-specific representation-learning approaches.

We restrict this supplementary comparison to *vote* and *spect*, the two datasets in Section 6.2 that are natively binary.

Evaluation protocol. We use the same preprocessed feature matrices as in the main real-data experiments. The *vote* dataset contains $n = 232$ observations and $d = 16$ binary variables, whereas *spect* contains $n = 267$ observations and $d = 22$ binary variables. For all methods, the number of clusters is fixed to two, matching the two external labels and the $m = 2$ setting used in the main experiments.

External labels are not used for fitting the representations or clusters and are introduced only at evaluation time. We report the same three permutation-invariant external clustering measures used elsewhere in the paper: clustering accuracy (ACC), adjusted Rand index (ARI), and normalized mutual information (NMI). Higher values are better for all three metrics.

For TableDC and ZEUS, we use the released implementations and follow their respective inference procedures, setting the requested number of clusters to two. The resulting cluster assignments are evaluated using exactly the same ACC, ARI, and NMI implementations used for ODIN. These methods are evaluated only in the real-data clustering experiment: unlike the pB mixture estimators considered in Section 6.1, they do not estimate the product-Bernoulli parameters (π, μ) required to compute the synthetic component-recovery error.

Results. On *vote*, ODIN and TableDC produce very similar clustering results. TableDC attains ACC 0.8922, while ODIN obtains 0.892; conversely, ODIN obtains slightly higher ARI (0.614 versus



Figure 18: SPECT component-signature atlas at ODIN layer 14 with $m = 2$. Rows correspond to fitted mixture components, sorted by estimated weight. Cells show the highest-ranked ODIN atoms under the one-vs-rest enrichment score $\hat{E}^{\text{ovr}}_{h,S}$. The $m = 2$ solution separates a large weakly specific component from a small component characterized by apical–inferior and basal-inferior co-activation events.

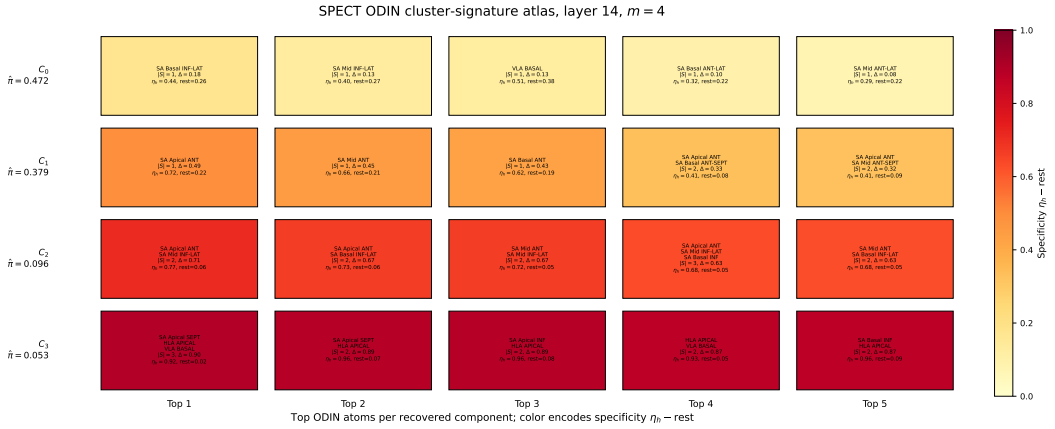


Figure 19: SPECT component-signature atlas at ODIN layer 14 with $m = 4$. The recovered components correspond to broader anatomical signatures, including anterior, inferolateral/basal, anterior–inferolateral, and apical/septal long-axis patterns. Each label is obtained directly by ranking ODIN atoms according to $\hat{E}^{\text{ovr}}_{h,S}$.

	vote			*spect*		
Method	ACC ↑	ARI ↑	NMI ↑	ACC ↑	ARI ↑	NMI ↑
ZEUS (Marszał ek et al., 2025)	0.8707	0.5477	0.4431	0.6592	0.0991	0.1456
TableDC (Rauf et al., 2025)	0.8922	0.6137	0.5199	0.6330	0.0682	0.1100
ODIN	0.892	0.614	0.526	0.794	-0.096	0.072

Table 12: **Additional comparison with recent tabular-specific clustering methods.** All methods are evaluated with two clusters on the same preprocessed *vote* and *spect* datasets. ACC, ARI, and NMI measure agreement with the external binary labels. The ZEUS and TableDC entries correspond to the supplementary runs reported here; ODIN values are the central values from the repeated experiments reported in Appendix J. Because repeated-seed estimates are not yet available for the two supplementary baselines, small numerical differences should not be interpreted as statistically significant.

0.6137) and NMI (0.526 versus 0.5199). ZEUS also recovers substantial agreement with the external labels, although its ACC, ARI, and NMI are lower in this experiment. The small numerical differences between ODIN and TableDC should not be overinterpreted without repeated-seed estimates. More importantly, the result shows that the explicit interaction-moment representation used by ODIN can reach clustering performance comparable to a recent tabular-specific deep clustering representation on this dataset.

The picture is different on *spect*. ODIN obtains the largest ACC (0.794), but this is accompanied by an ARI of -0.096 and an NMI of 0.072. ZEUS instead obtains lower ACC (0.6592) but higher ARI (0.0991) and NMI (0.1456), with TableDC giving intermediate values. Thus the methods induce substantially different partitions of this dataset, and ACC alone is insufficient to characterize their agreement with the external labels. In particular, ODIN's high ACC together with its negative ARI indicates that its two-component partition is strongly affected by the class/cluster-size structure and does not reproduce the external annotation uniformly. This is consistent with the representation diagnostics in Appendix I, where the SPECT labels remain mixed in the learned representation.

Overall, the additional comparison does not indicate uniform dominance of one clustering method across datasets or metrics. Rather, it shows that on the strongly structured *vote* benchmark, ODIN is competitive with recent tabular-specific representation learners despite using an explicit and inspectable interaction dictionary. On the more ambiguous *spect* dataset, the disagreement between ACC, ARI, and NMI reinforces the distinction between recovering a partition of the observed distribution and reproducing a particular external binary annotation. The additional baselines therefore provide a contemporary clustering reference without changing the central role of ODIN in the paper: its contribution is not a black-box embedding optimized solely for label agreement, but a finite observable representation whose coordinates enter directly into the latent mixture-recovery problem.

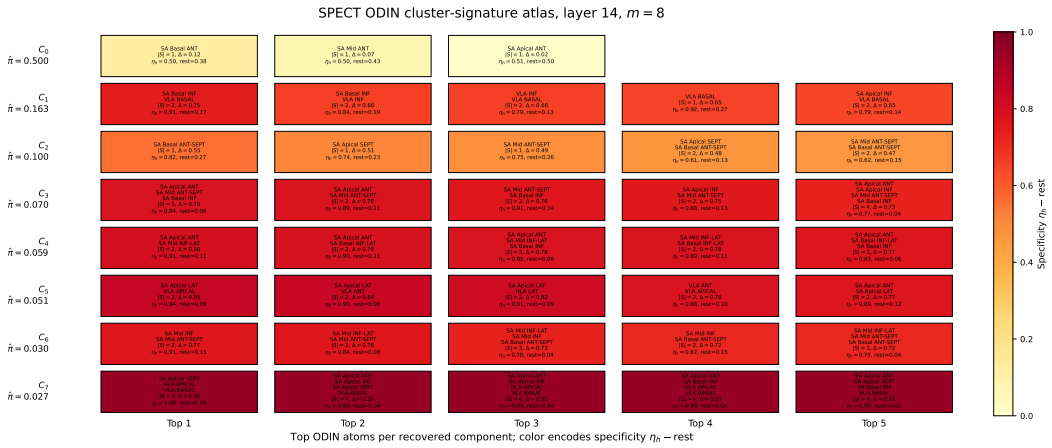


Figure 20: SPECT component-signature atlas at ODIN layer 14 with $m = 8$. Increasing the number of components refines the decomposition into smaller subtypes with sharper higher-order anatomical interaction signatures. These signatures are not post-hoc explanations: they are the selected ODIN moments used by the mixture recovery objective.