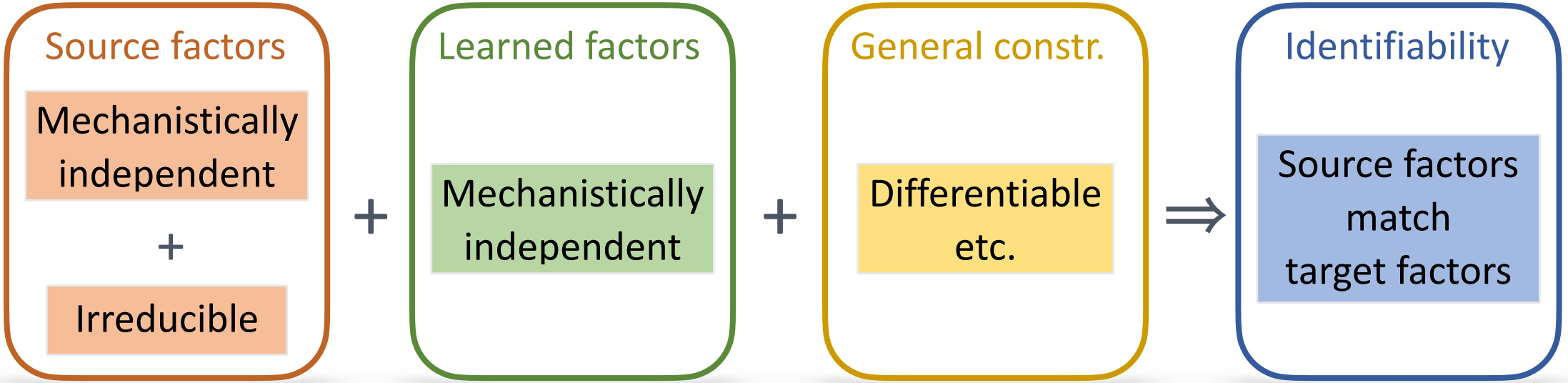
Mechanistic Independence for Identifiable Disentanglement

- Motivation (vision): disentangle objects + their properties
- Disentanglement: latent subspaces correspond to distinct factors of variation
- Identifiability: are those factors uniquely recoverable (up to trivial ambiguities)?
- Difficulties:
  - Factors can be statistically dependent (e.g., “PC“ co-occurs with “desk“)
  - Nonlinear generators need extra structure—independence alone is insufficient for identifiability
- Idea: use structure of the generator (mechanistic independence) to get identifiability

Background: from ICA/ISA to mechanistic independence

- Setup: $x = g(s)$, with latent space $S \subseteq S_1 \times \ldots \times S\_K$ (subspace factors)
- ICA: assumes independent latent variables
- ISA: assumes independent subspaces (dependence allowed within each subspace)
- ISA needs “irreducibility”: true subspaces cannot be split into smaller independent ones
- This paper: reuse the ISA logic, but replace statistical independence by mechanistic independence

Core theorem template

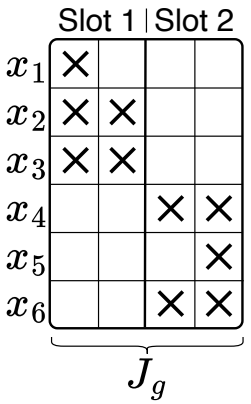


- Irreducibility: subspaces cannot be split into smaller mechanistically independent subspaces
- Identifiable up to: permutation of factor blocks + invertible transforms within each block
- Works with multi-dimensional factors and can tolerate statistically dependent factors

Mechanistic independence criteria

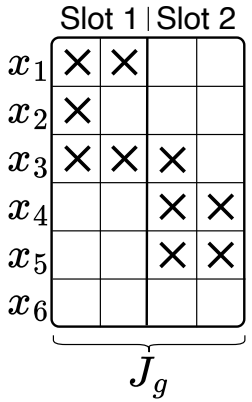
Type D

- Factors affect **disjoint** observation coordinates
- Based on [1]



Type M

- Factors affect **mutual non-inclusive** observation coordinates
- Based on [4]

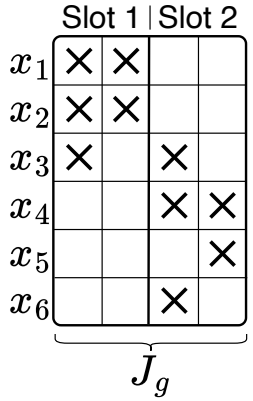


Type H_n

- **Higher-order separability:** mixed derivatives between factors vanish
- Type H_2: factors act additively on observ.
- Based on [2,3]

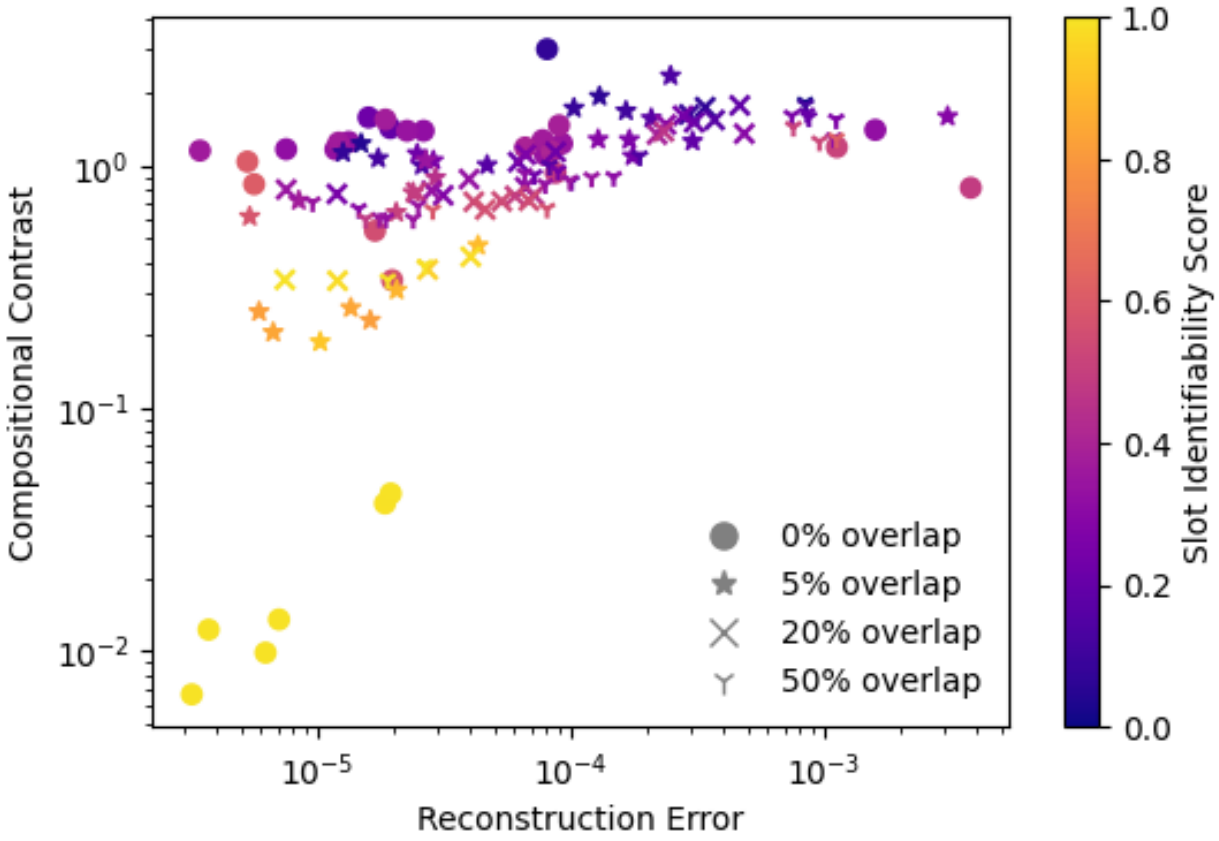
Type S

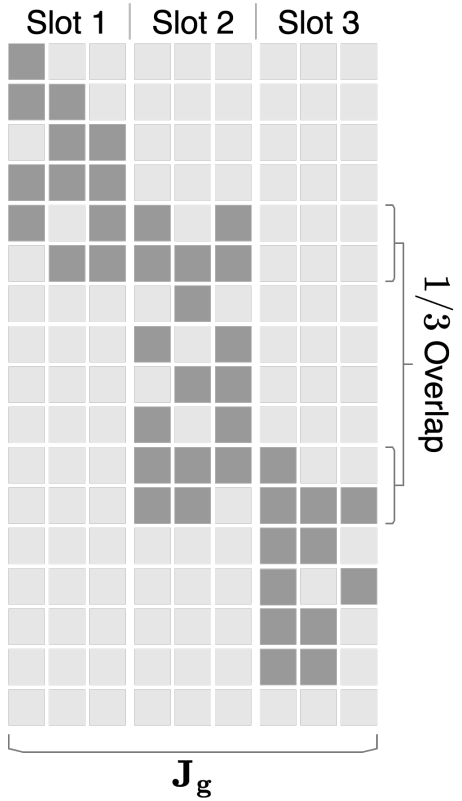
- **Sparsity gap**: Jacobian in factor-aligned basis is **sparser** than for any mixing.



[1] Provably Learning Object-Centric Representations, 2023, Brady et al.
[2] Additive Decoders for Latent Variables Identification and Cartesian-Product Extrapolation, 2023, Lachapelle et al.
[3] Interaction Asymmetry: A General Principle for Learning Composable Abstractions, 2024, Brady et al.
[4] Generalizing Nonlinear ICA Beyond Structural Sparsity, 2023, Zheng et al.

Experiments (synthetic)





Compositional contrast (from [1]): $C_{\text{comp}}(\hat{\boldsymbol{g}}, \boldsymbol{z}) = \sum_{n=1}^{d_x} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \left\| \frac{\partial \hat{g}_n}{\partial \boldsymbol{z}_i}(\boldsymbol{z}) \right\| \left\| \frac{\partial \hat{g}_n}{\partial \boldsymbol{z}_j}(\boldsymbol{z}) \right\|$

Intuition: factors as connected graph components