Multi-Channel Causal Variational Autoencoder for multimodal biomedical causal disentanglement

Safaa Al-Ali, Irene Balelli

Multi-Channel Causal Variational Autoencoder for multimodal biomedical causal disentanglement

Safaa Al-Ali[1,*], Irene Balelli[1] and for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]*Inria Center at Université Côte d'Azur - Epione Team, 2004 Rte des Lucioles, 06902, Valbonne, France*

Abstract

The multimodal nature of clinical assessment and decision-making, and the high rate of healthcare data generation, motivate the need to develop approaches specifically tailored to the analysis of these complex and potentially high-dimensional multimodal datasets. This poses both technical and conceptual challenges: how can such heterogeneous data be analyzed jointly? How can modality-specific information be identified from shared information? Variational autoencoders (VAEs) offer a robust framework for learning latent representations of complex data distributions, while being flexible enough to adapt to different data types and structures, and having already been successfully applied for latent disentanglement of multimodal (multi-channel) data. We aim at tackling multi-channel disentanglement from a causal perspective, and seek at identifying causal relationships between channels, beyond simple statistical associations. To do that, we propose Multi-Channel Causal VAE (MC^{2}VAE), a novel causal disentanglement approach for multi-channel data, whose objective is to jointly learn modality-specific latent representations from a multi-channel dataset, and identify a causal structure between the latent channels. Each channel is projected into its own latent space, where a causal discovery step is integrated to learn the hidden causal graph. Finally, the decoder takes into account the discovered graph to predict the data. Covariate of interest can be integrated as well when available, and accounted in the causal graph structure. Extensive experiments on synthetically generated multi-channel datasets demonstrate the ability of MC^{2}VAE in effectively uncovering the underlying latent causal structures across multiple channels, hence making it a strong candidate for real-world multi-channel causal disentanglement. Application to multi-channel data on neurodegeneration extracted from the Alzheimer's Disease Neuroimaging Initiative highlights the existence of a biologically meaningful latent causal structure, whose pertinence is supported by multiple previous experimental and modelling work, and provides actionable insight for disease progression.

Keywords

Multi-channel biomedical data, Causal disentanglement, Variational Autoencoder, Alzheimer's Disease

1. Introduction

In the healthcare domain, clinical decision making - whether for diagnosis, prognosis or therapeutics - relies on integrating as much patient-specific information as possible. This information often comes from diverse sources such as clinical assessments, multimodal medical imaging, or

† Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

safaa.al-ali@inria.fr (S. Al-Ali); irene.balelli@inria.fr (I. Balelli)

0000-0003-1864-8190 (S. Al-Ali); 0000-0002-4593-8217 (I. Balelli)

genomics profiles, enabling a comprehensive view of the patient, its pathology and response to a treatment. This statement, together with the increasing availability of healthcare data, motivates the development of methods specifically tailored to the joint analysis of data coming from different sources, which can play a crucial role towards personalized medicine [1]: we will refer to such data with the term *multi-channel*. Multi-channel data are obtained by gathering together observations generated from multiple sources: each source can capture a specific portion of information about the phenomenon under study, and can contribute to its overall understanding.

Unfortunately, multi-channel data analysis is far from trivial, and comes with several challenges, due to the intrinsic heterogeneity of such data, the possible high dimensionality of some channels, and the potential correlations between channels, each one being a specific piece of a same puzzle. Finally, an effective integration of multi-channel data should be able to preserve each channel-specific information and highlight the cross-channels one.

Variational Autoencoders (VAEs) have gained significant attention for their ability to learn complex data distributions in an unsupervised manner [2]. VAEs are Bayesian generative models, consisting of an encoder, whose role is to project the input data into a lower-dimensional latent space, and a decoder which reconstructs the original data back from its latent representation. The VAE inference process is efficiently carried out using amortized inference [3], where the posterior moments are parameterized with neural networks. The flexibility of VAEs makes them particularly suitable for dealing with different types of data, hence they appear as good candidates to perform multi-channel analysis. Nevertheless, VAEs in their classical formulation and specifically when applied to the multi-channel scenario, focus on correlation-type relationships, and are not able to highlight directed causal patterns.

Causal learning is a very active and evolving area of research that aims to identify causal relationships between observations, going beyond simple statistical association, and ultimately improving interpretability, explainability and deployability. Classical methods developed to discover the underlined causal structure from a set of observations are typically well suited to deal with relatively low-dimensional data. More recently, attempts to couple causal discovery and machine learning techniques to cope with higher dimensional and complex datasets, have shown promising results [4]. Nevertheless, scaling up to multi-channel data still appears as an extremely challenging task.

We propose a novel model called Multi-Channel Causal Variational Autoencoder (MC^{2}VAE), for the joint analysis of multi-channel data, with the objective of identifying meaningful causal relationships between each modality, enhancing our comprehension of the overall picture each channel is contributing to. To do so, MC^{2}VAE will rely on VAEs, which ensure the identification of a latent projection for each channel, and causal discovery techniques to identify the causal relationships between channels through their lower-dimensional representations: this knowledge will be valuable for helping the VAEs to reconstruct the data. Moreover, MC^{2}VAE is able to account for covariates of interest, when available, and study their causal impact on the latent system. To the best of our knowledge, this is the first approach investigating the discovery of inter-modalities causal relationships.

The rest of the paper is organised as follows. In Section 2, we summarise the state-of-the-art on multi-channel data analysis, with a specific focus on VAEs, and causal disentanglement. In Section 3, we provide a formal definition of MC^{2}VAE, and describe the optimisation strategy.

Section 4 shows results on synthetically generated data, while Section 5 focuses on the real-case application. Section 6 concludes the paper.

Table 1
Statement of significance

Problem or Issue	The multimodal nature of clinical assessment and the rapid growth of healthcare data motivate the need for methods tailored to the joint analysis of such complex and high-dimensional datasets.
What is Already Known	Variational autoencoders (VAEs) can learn latent representations of complex multimodal data but do not explicitly capture causal inter-modalities relationships.
What This Paper Adds	This paper proposes Multi-Channel Causal VAE (MC^{2}VAE), a causal disentanglement approach that jointly learns modality-specific latent spaces and infers causal relationships between channels, accounting for additional latent covariate causal effects. Experiments on synthetic datasets confirm its ability to uncover latent causal structures, and application to multimodal data on Alzheimer's Disease reveals biologically plausible causal patterns aligned with prior research, offering insight into disease progression.
Who would benefit from the new knowledge	Mulitmodal data analysis is fundamental for precision medicine. Our findings highlight the utility of our model in uncovering meaningful multi-channel causal pathways for biomedical data with a potential strong implication for actionable healthcare.

2. Related works

2.1. Multi-channel Representation Learning with VAEs

Representation learning aims to disentangle the observed variables by projecting them into latent lower-dimensional independent features, which correspond to distinct generative factors, providing a compact representation of the data and its variability [5, 6]. To solve this task, several models have been proposed, including Recurrent Neural Networks (RNNs) [7], Generative Adversarial Networks (GANs) [8], Deep Reinforcement Learning (DRL) [9], and VAEs [10].

In the multi-channel context, VAEs have already been successfully deployed. For instance, Antelmi *et al.* [11] proposed a sparse VAE that jointly learns latent relationships across multiple channels, under the assumption that the latent space is shared across all channels. On the other hand, in [12] the authors study the spatio-temporal dynamics of disease evolution through a multi-channel VAE, where each channel is projected separately in its latent space, and a latent neural dynamical system describes the time evolution of the latent variables. A supervised multi-channel variational autoencoder (MVAE) based on conditional VAE (CVAE) has also been proposed [13], and applied to analyze time signals, while a generalized multi-channel conditional variational autoencoder [14] has been used for multi-channel audio source separation under underdetermined conditions. In [15], Bayesian Networks in conjunction with sparse autoencoders are used to incorporate arbitrary multi-scale, multi-channel data without making

specific distribution assumptions. A conditional generative modeling (CGM) approach for unsupervised disentangled learning based on variational autoencoder (VAE) was also proposed [16]: it employs a categorical conditional prior distribution in the latent space to learn global uncertainty in the data. Finally, to enforce multi-channel coherence, Wesego *et al.* [17] proposed to learn the correlation among the latent variables of unimodal VAEs using score-based models.

Despite the interest of the above-described approaches (and other variants, developed in the past 2 decades) for the joint analysis of multi-channel data, none of them explicitly include a notion of causal relationships between the involved channels, which are mixed on the basis of correlation.

2.2. Causal Disentangled Representation Learning

Causal discovery [18, 19] is a branch of causal research that aims to unveil the causal relationships between observed variables. Causal links are defined as intrinsically asymmetric relationships, whose cause-and-effect actions can be typically represented through Directed Acyclic Graphs (DAGs). Causal discovery, which has already proven relevant in diverse research fields, including medicine [20, 21], biology [22], physics [23], and economics [24], is a powerful tool to understand the underlying mechanism driving data generation, especially when prior expert knowledge is not already fully established. Further, it stands as a necessary preliminary step to perform causal inference which is the basis for counterfactual reasoning (*e.g.*, [25]).

Classical approaches to causal discovery (*e.g.* [26, 27, 28, 29]) strongly suffer from the curse of dimensionality, and are not well adapted to deal with complex and high-dimensional datasets. Therefore, Causal Disentangled Representation Learning (CDRL) [30] has recently emerged, and seeks to solve this bottleneck by relying on machine and deep learning methods for feature extraction. Unlike conventional representation learning, which may merge multiple causal factors into a single representation, causal disentanglement strives to identify and separate causal latent factors that contribute to the generation of the observed variables, hence enabling a more comprehensive understanding of the underlying data-generating process.

Among CDRL approaches, CausalVAE [31] proposes a supervised VAE-based structure. Given some external high-level information about data - the *concepts*, with an established causal structuring - the authors include a Causal Layer in their pipeline, and learn a causal graph in the latent space that mirrors the known causal relationships between the *concepts*. The model shows good results on both synthetic and benchmark datasets. However, by leveraging the a priori known graph on the *concepts* of interest, they enforce the dimension of the latent space and the causal structure therein. An extension of CausalVAE was later proposed by Komandouri *et al.* [32] to relax the linear assumption of causal relationships. Other methods proposed in the literature include CausalGan [33], which is based on GAN and is designed to perform causal inference on images, but requires a prior causal graph. causalPIMA [34] has been recently proposed for causal representation learning, integrating physics-based constraints, and a product of experts. Of note, the authors use the term *multimodal* to refer to the multivariate nature of the causal latent space, such as colors or geometric shapes as the driving generating features for an image. This definition differs from the one used in this paper. Finally, DAG-GNN [35] leverages graph neural networks (GNN) to learn a DAG from (unimodal) data, eventually on a lower-dimensional latent space, and without requiring supervision.

Despite the significant achievements reached, we noticed that to the best of our knowledge, the CDRL models available in the literature are not designed to deal with the multi-channel scenario we wish to address here. Moreover, several methods require injecting additional knowledge to drive the causal discovery in the latent space, while we aim at solving the discovery task in an unsupervised manner, due to the realistic unavailability of an established prior latent causal structure across channels, especially in healthcare. In [36] we have proposed a simplified version of MC^{2}VAE to investigate drug-induced Torsades de Pointes, a life-threatening arrhythmia. Here, we substantially extend [36] by introducing non-linear causation, covariates' effect, and performing extensive experiments.

3. Multi-channel Causal Variational Autoencoder (MC^{2}VAE)

To deal with multi-channel data and the need for recovering their mutual causal relationships, we propose Multi-Channel Causal Variational Autoencoder, MC^{2}VAE. Our approach aims to uncover a causal graph between channels' latent variables in a fully unsupervised manner, *i.e.* without requiring any additional information beyond the data itself. In this way, we aim to obtain an interpretable and richer representation of such complex data.

The overall architecture of MC^{2}VAE is summarized in Figure 1. We denote by $\mathbf{X} = (X_i)_{i=1,\dots,N}$ the dataset of observed variables, where N is the total number of subjects, and $X_i := (\mathbf{x}_i^m)_{m=1,\dots,M}$ is the dataset for the i^{th} subject consisting of observations from a total of M channels, $\mathbf{x}_i^m$. Each $\mathbf{x}_i^m$ lies in a specific d_m-dimensional space. Some covariates of interest which can have a causal effect over the different channels may be observed for each subject: for the sake of simplicity, we consider here only a single binary covariate, $v = (v_i \in \{0,1\})_{i=1,\dots,N}$; the extension to multiple covariates is straightforward. The latent vector for subject i following the encoding operation is denoted by $\mathbf{z}_i := (z_i^m)_{m=1,\dots,M} \in \mathbb{R}^M$, where each z_i^m corresponds to the latent representation of $\mathbf{x}_i^m$. After passing through the causal layer, where the covariate can be eventually integrated, the transformed channel-specific latent variables, $\mathbf{z}_i^c$, will be feed to the decoders.

The latent structural causal model and the generative model. MC^{2}VAE hypothesizes the existence of a latent structural causal model (SCM) relating the channel-specific latent variables, $\mathbf{z}^c = \{z^{c,m}\}_{m=1,\dots,M}$. Let A be a weighted adjacency matrix, strictly upper triangular up to some permutations of its columns and rows, which fully describes the causal graph structure across $\mathbf{z}^c$, *i.e.* the ij-th element of A, a_{ij}, provides the strength of the causal relationship of z_i^c (the parent variable) to z_j^c (the children) ($a_{ij} = 0$ denotes no causal relationships between z_i^c and z_j^c). Under a linear SCM assumption, one would get:

$$\mathbf{z}^c = (\mathbb{I}_M - A^\top)^{-1}(\mathbf{z} + \beta v), \tag{1}$$

where $\mathbf{z} = (z^m)_{m=1,\dots,M}$ provides the independent noise term, $\mathbb{I}_M$ denotes the M-dimensional identity matrix, and $\beta \in \mathbb{R}^M$ is the vector of v's causal weights over the modalities *i.e.* $\beta_i \neq 0$ represents the strength of the causal impact of the covariate v on the modality i. The

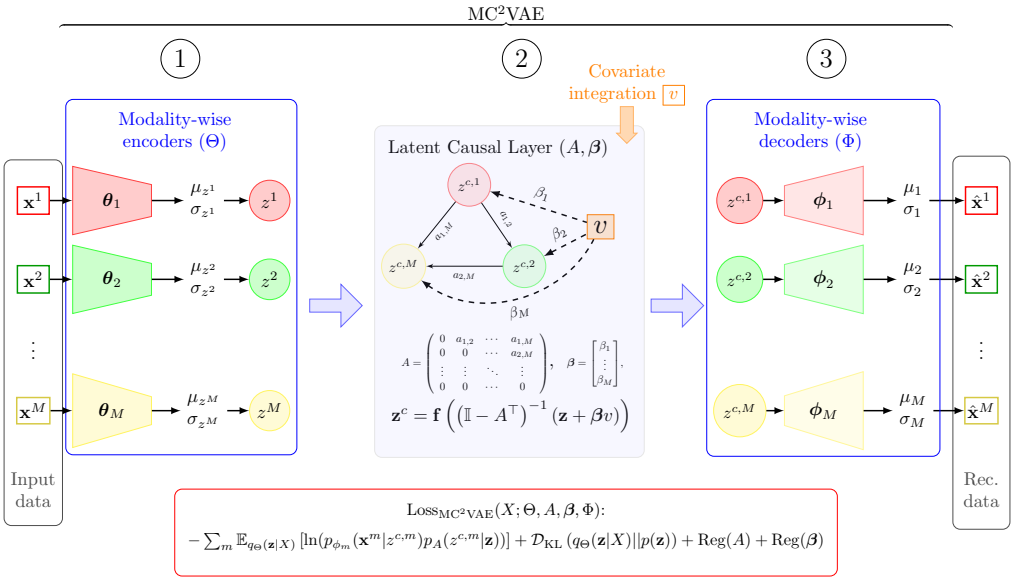

Figure 1: Workflow of MC^{2}VAE. 1) Each channel m is encoded to a one-dimensional latent variable z^m through its channel-specific encoder. 2) The latent variables $\mathbf{z} = (z^m)_{m=1,\dots,M}$ pass into the Latent Causal Layer, where the structural causal model is learned. When available, the covariate v is integrated as well. 3) The decoder takes the transformed causal variables $\mathbf{z}^c = (z^{c,m})_{m=1,\dots,M}$ to reconstruct the observations for each channel, through a channel-specific decoding.

superscript $\top$ denotes the matrix transpose.

Of note, thanks to the channel-specific encoding and decoding operations, our VAE-based structure allows to smoothly account for a generalized version [35] of the linear SCM given in Equation (1): $\mathbf{z}^c = \mathbf{f}_\Phi\left((\mathbb{I}_M - A^\top)^{-1}(\mathbf{g}_\Theta(\mathbf{z}) + \beta v)\right)$, where $\mathbf{f}_\Phi$ and $\mathbf{g}_\Theta$ are (possibly non-linear) functions, which respectively affect the nature of the causal relationships (but not the underlying causal structure) and the noise density.

We consider the following generative model, parameterized by (A, β, Φ):

$$\begin{cases} \mathbf{z} \sim p(\mathbf{z}) \\ p_{(A,\beta)}\left(\mathbf{z}^c|\mathbf{z}, v\right) = \prod_{m=1}^M p_{(A,\beta)}(z^{c,m}|\mathbf{z}, v) \\ p_\Phi\left(\mathbf{X}|\mathbf{z}^c\right) = \prod_{m=1}^M F_m\left(\mathbf{x}^m|f_{m,\phi_m}(z^{c,m})\right), \end{cases} \tag{2}$$

where:

- $p(\mathbf{z})$ is the prior distribution over the latent exogenous variables, supposed here to be a product of M independent standard Gaussians;

- $p_{(A,\beta)}\left(\mathbf{z}^c|\mathbf{z}, v\right)$ is the Markov factorization of the joint distribution of the latent endogenous variables;
- $(F_m(\cdot|\boldsymbol{\eta}_m))_{\boldsymbol{\eta}\in H_m}$ are parametric families of distributions for each m, in our case all Gaussians, since we will always consider continuous variables throughout the paper;
- $\forall m, f_{m,\boldsymbol{\phi}_m} : \mathbb{R} \to H_m$ is the decoder function for the m-th channel, a neural network parameterized by $\boldsymbol{\phi}_m \in \Phi$. According to the above, for each m, the output of $f_{m,\boldsymbol{\phi}_m}$ will be a d_m-dimensional mean, μ_m^{D}, and a variance-covariance matrix, which we assume to be diagonal to ease the derivations, $\Sigma_m^{\mathrm{D}} = \sigma_m^{\mathrm{D}^2}\mathbb{I}_{d_m}$.

Formulation of the loss function. In order to optimise the parameters of MC^{2}VAE, we need to maximize the marginal log likelihood, $\mathcal{L}$:

$$\mathcal{L} = \sum_{i=1}^{N} \log\left[p_{(A,\beta,\Phi)}(X_i)\right] = \sum_{i=1}^{N} \log\left[\int p_\Phi\left(X_i|\mathbf{z}_i^c\right) p_{(A,\beta)}\left(\mathbf{z}_i^c|\mathbf{z}_i, v_i\right) p(\mathbf{z}_i) d\mathbf{z}_i\right]$$

Due to the intractability of the integrals involved in the above expression, we apply variational Bayes and introduce a tractable posterior $q_\Theta(\mathbf{z}|\mathbf{X})$ to approximate the true one. In particular we assume what follows:

$$q_\Theta(\mathbf{z}|\mathbf{X}) = \prod_{m=1}^{M} G_m\left(z^{c,m}|g_{m,\boldsymbol{\theta}_m}(\mathbf{x}^m)\right), \tag{3}$$

where

- $(G_m(\cdot|\boldsymbol{\gamma}_m))_{\boldsymbol{\gamma}\in L_m}$ are parametric families of distributions for each m, in our case all Gaussians;
- $\forall m, g_{m,\boldsymbol{\theta}_m} : \mathbb{R}^{d_m} \to L_m$ is the encoder function for the m-th channel, a neural network parameterized by $\boldsymbol{\theta}_m \in \Theta$. For each m, the output of $g_{m,\boldsymbol{\theta}_m}$ will be a 1-dimensional mean, μ_m^{E}, and a standard deviation, σ_m^{E}.

According to the detailed modeling assumptions, and applying Jensen's inequality, we get to the following tractable Evidence Lower Bound (ELBO) for MC^{2}VAE, which we denote by $\mathcal{E}_{(A,\beta,\Phi,\Theta)}$:

$$\mathcal{L} \geq \mathcal{E}_{(A,\beta,\Phi,\Theta)} = \mathbb{E}_{q_\Theta}\left\{\log\left[p_\Phi\left(\mathbf{X}|\mathbf{z}^c\right)\right]\right\} + \mathbb{E}_{q_\Theta}\left\{\log\left[p_{(A,\beta)}\left(\mathbf{z}^c|\mathbf{z}, v\right)\right]\right\} - \mathcal{D}_{\mathrm{KL}}\left(q_\Theta(\mathbf{z}|\mathbf{X})||p(\mathbf{z})\right), \tag{4}$$

where $\mathcal{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler divergence.

Using the generative model (2), the inference one (3), and the previously defined latent SCM, each term of Equation (4) can be explicitly derived, leading us to:

$$\mathcal{E}_{(A,\beta,\Phi,\Theta)} = -\frac{1}{2}\sum_{i=1}^{N}\left\{\sum_{m=1}^{M}\left[d_m \log\left(2\pi\sigma_m^{\mathrm{D}^2}\right) + \frac{||\mathbf{x}^m - \mu_m^{\mathrm{D}}||^2}{\sigma_m^{\mathrm{D}^2}} - \log\left(\frac{\sigma_m^{\mathrm{E}^2}}{2\pi}\right) + \sigma_m^{\mathrm{E}^2} + \mu_m^{\mathrm{E}^2} - 1\right] + I_i\right\}, \tag{5}$$

where

$$I_i = \det(B) + tr\left(B^{-1}\Sigma^{\mathrm{E}}\right) + \left(\boldsymbol{\mu}^E - \boldsymbol{\xi}(A,\beta)\right)^\top B^{-1}\left(\boldsymbol{\mu}^E - \boldsymbol{\xi}(A,\beta)\right), \tag{6}$$

and

- $B = \left(\mathbb{I}_M - A^\top\right)^{-1}\left(\left(\mathbb{I}_M - A^\top\right)^{-1}\right)^\top$;
- $\boldsymbol{\mu}^E \in \mathbb{R}^M$ is obtained by concatenation of all $\mu_m^E, m = 1, \ldots, M$;
- $\Sigma^E := diag(\sigma_1^{\mathrm{E}^2}(\mathbf{x}^1, \theta_{1,2}), \ldots, \sigma_M^{\mathrm{E}^2}(\mathbf{x}^M, \theta_{M,2})) \in \mathbb{R}^{M\times M}$;
- $\boldsymbol{\xi}(A,\beta) = \left(\mathbb{I}_M - A^\top\right)^{-1}\beta v$;
- tr is the trace operator.

Of note, the generative part of (4) can be computed through Monte Carlo approximation by considering multiple samples of $\mathbf{z} \sim q_\Theta(\mathbf{z}|\mathbf{X})$. Some details of the proof are provided in Appendix A.1.

Up to now, non-linearities have been accounted through the encoding and decoding functions, and sampling. Nevertheless, one may want to explicitly model a deterministic non-linear latent SCM, according to a given vector-valued function $\mathbf{f}_{\text{det}} : \mathbb{R}^M \to \mathbb{R}^M$,

$$\mathbf{z}^c = \mathbf{f}_{\text{det}}\left((\mathbb{I}_M - A^\top)^{-1}(\mathbf{z} + \beta v)\right) = \mathbf{f}_{\text{det}}(\hat{\mathbf{z}}^c) \tag{7}$$

To do so, one strategy is to linearize back Equation (7) by using a first-order Taylor expansion, under some mild assumptions over $\mathbf{f}_{\text{det}}$, with the great advantage of preserving Gaussianity:

$$\mathbf{f}_{\text{det}}(\hat{\mathbf{z}}^c) \approx \mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta)) + \mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))\left((\mathbb{I}_M - A^\top)^{-1}\mathbf{z}\right), \tag{8}$$

where $\mathbf{J}_{\mathbf{f}_{\text{det}}}$ is the Jacobian of $\mathbf{f}_{\text{det}}$, and the expansion is realized around $\boldsymbol{\xi}(A,\beta)$, the mean of $\hat{\mathbf{z}}^c$ under a centered Gaussian prior for $\mathbf{z}$. Injecting (8) into the previously defined model, the new ELBO differs from Equation (5) only in the definition of the I_i term (Eq. (6)), which now writes:

$$I_i^{\mathbf{f}_{\text{det}}} = \det(\tilde{B}) + tr\left(\tilde{B}^{-1}\Sigma^{\mathrm{E}}\right) + \left(\boldsymbol{\mu}^E - \mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta))\right)^\top \tilde{B}^{-1}\left(\boldsymbol{\mu}^E - \mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta))\right), \tag{9}$$

where $\tilde{B} := \mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))\left(\mathbb{I}_M - A^\top\right)^{-1}\left[\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))\left(\mathbb{I}_M - A^\top\right)^{-1}\right]^\top \in \mathbb{R}^{M\times M}$.

For more details about the derivation of $\mathbb{E}_{q_\Theta}\left\{\log\left[p_{(A,\beta)}\left(\mathbf{z}^c|\mathbf{z}, v\right)\right]\right\}$, see Appendix A.2.

Constraints to enforce acyclicity. We recall that our objective is to learn a directed acyclic graph relating the channel-specific latent variables. The acyclicity is encoded in the adjacency matrix A, which we expect to be strictly upper triangular up to a permutation. To enforce this requirement, we add a penalization term to our ELBO inspired by Zheng *et al.* [37], where they show the following result:

Theorem 1. *A matrix $A \in \mathbb{R}^{M \times M}$ represents a directed acyclic graph if and only if*

$$\mathcal{H}(A) := tr\left(e^{A \circ A}\right) - M = 0, \tag{10}$$

where $\circ$ is the Hadamard product, and e^* is the matrix exponential. Moreover, the smooth differentiable function $\mathcal{H}$ has the simple associated gradient:

$$\nabla \mathcal{H}(A) = \left(e^{A \circ A}\right)^\top \circ 2A. \tag{11}$$

Additionally, to enforce sparsity among the channel-specific latent variables [38], we introduce a L_1 penalty on the adjacency matrix, *i.e.*, $L_1(A) = ||A||_1$. The same penalty is coherently added also to induce sparsity on the covariate causal impact: $L_1(\beta) = ||\beta||_1$.

Finally the objective function to be minimized writes:

$$Loss_{\text{MC}^2\text{VAE}}(X; \Theta, A, \beta, \Phi) := -\mathcal{E}_{(A,\beta,\Phi,\Theta)} + \eta_1 \mathcal{H}(A) + \eta_2 L_1(A) + \eta_3 L_1(\beta), \tag{12}$$

where $\{\eta_i\}_{i=1,2,3}$ are the weights of the regularisation terms.

Identifiability. The identifiability of our causal latent models is tightly related to the underlined non-linear causal patterns and the independence of the exogenous latent variables. Under these assumptions, Zhang and Hyvärinen [39] have provided a systematical characterization of the identifiability of the post-nonlinear causal model.

Optimization. We have implemented MC^{2}VAE using Python 3.10.9 and Pytorch 2.2.1. Algorithm 1 briefly outlines the steps performed at each training epoch. Optimization is efficiently carried out through minibatch stochastic gradient descent using the Adam optimizer [40]. The experiments detailed in Section 4 were conducted on a Dell PC. The system was powered by an Intel Core$^{\text{TM}}$ i7-10850H processor (6 cores at 2.70 GHz), with 15.2 GiB of RAM and 1 TB of storage. The environment ran on Fedora Linux 42 (kernel 6.16.8-200.fc42.x86_64). On average, a full 10-fold cross-validation training of MC^{2}VAE is completed within approximately 90 minutes in the case of five modalities (see Section 4.3.1). The code used to run MC^{2}VAE (and to generate the synthetic datasets illustrated in Section 4) is made publicly available on Gitlab.

4. Experimental results

4.1. Architectures

We will consider four distinct architectures for MC^{2}VAE, described in Table 2, varying the encoder-decoder structures, and eventually adding a deterministic non-linear function $\mathbf{f}_{\text{det}}$ in the latent causal layer.

The four architectures were selected to study the contributions of two core modeling components: (i) the projection performed by the encoders and decoders (linear vs. non-linear), and (ii) the impact of the causal model in the latent space (linear, non-linear, or deterministic). Architectures 1 and 4 correspond to the fully unsupervised scenario in which no information

Algorithm 1 MC^{2}VAE

Require: Multi-modal data $(\mathbf{x}_m)_{m=1,\dots,M}$; regularisation hyper-parameters $\{\eta_i\}_{i=1,2,3}$; E epochs; batch size; optimiser hyper-parameters

Optional: Covariates v; M custom encoder-decoder architectures (default to one linear layer); $\mathbf{f}_{\text{det}}$ non-linear function (default to $\mathbb{I}_M$);

for $e = 1, \dots, E$, and each batch **do**
  for $m = 1, \dots, M$ **do**
    Encode $\mathbf{x}^m$, and sample $z^m \sim q_{\boldsymbol{\theta}_m}(z^m|\mathbf{x}^m)$ (reparameterization trick)
  end for
  $\mathbf{z} \leftarrow concat(z^m)_{m=1,\dots,M}$
  Causal mixing: $\mathbf{z}^c = \mathbf{f}_{\text{det}}\left((\mathbb{I}_M - A^\top)^{-1}(\mathbf{z} + \beta v)\right)$ (Latent Causal Layer)
  for $m = 1, \dots, M$ **do**
    Decode $z^{c,m}$, and sample $\mathbf{x}^m \sim p_{\boldsymbol{\phi}_m}(\mathbf{x}^m|z^{c,m})$
  end for
  Compute the ELBO (5) (using (9) if $\mathbf{f}_{\text{det}}$, (6) otherwise)
  Return $Loss_{\text{MC}^2\text{VAE}}(X; \Theta, A, \beta, \Phi)$ (Eq.(12))
  Backpropagate
end for

	Encoder/Decoder	$\mathbf{f}_{\text{det}}$	ELBO $\mathcal{E}_{(A,\beta,\Phi,\Theta)}$
Arch. 1	linear	$\mathbf{f}_{\text{det}} = \mathbb{I}$	(6)
Arch. 2	linear	Same $\mathbf{f}_{\text{det}}$ used for data generation	(9)
Arch. 3	linear	$\mathbf{f}_{\text{det}} = \text{PReLU}$, with learnable negative slope α	(9)
Arch. 4	non-linear (MLP with hyperbolic tangent activation)	$\mathbf{f}_{\text{det}} = \mathbb{I}$	(6)

Table 2
MC^{2}VAE's architectures considered for the experimental part. Arch. 2 will be considered only for synthetically generated data (Section 4.3).

about the true causal mechanism is available, with either linear or non-linear encoder–decoder mappings. Architectures 2 and 3 serve as controlled intermediate variants: Architecture 2 incorporates the ground-truth non-linear causal function (available in synthetic simulations), while Architecture 3 uses a flexible parametric PReLU approximation for a more generalized causal mechanism. These intermediate architectures enable a principled assessment of how the availability and complexity of the causal mechanism affect MC^{2}VAE's performance.

4.2. Graph structure metrics

We consider four standard metrics to evaluate the accuracy of the discovered latent graph in the synthetic scenario, when the ground truth is known, treating the causal discovery task as a classification task: *False Discovery Rate* (FDR), *True Positive Rate* (TPR), *False Positive Rate* (FPR),

and *Structural Hamming Distance* (SHD). Their definitions are summarized in Table 3 where:

- FP is the number of false positives (edges predicted but not present in the true graph),
- TP is the number of true positives (correctly predicted edges).
- FN is the number of false negatives (edges that are present in the true graph but not in the discovered one).
- TN is the number of true negatives (non-edges correctly predicted as non-edges).
- Reverse counts edges with flipped direction in the predicted graph.

Metric	Definition	Interpretation
FDR ↓	$\frac{FP}{TP+FP}$	Proportion of incorrectly predicted edges among all predicted ones
TPR ↑	$\frac{TP}{TP+FN}$	Proportion of correctly identified edges
FPR ↓	$\frac{FP}{FP+TN}$	Proportion of incorrectly predicted non-edges
SHD ↓	$FP + FN +$ Reverse	Total number of edge errors (missing, extra, reversed)

Table 3
Graph structure metrics used to evaluate predicted latent DAGs against ground-truth. True edges refer to edges that are present in the true graph.

Note that TPR and FPR should be considered jointly: a fully connected (resp. an empty graph) can trivially maximize (resp. minimize) these metrics.

4.3. Application to synthetic data

4.3.1. Data generation

In order to validate our model, we have applied it to several multi-channel synthetic datasets, obtained by varying the number of latent nodes, the density of the latent causal graphs, the channels' dimensions, the applied non-linearities, and the final sample size. We uniquely denote each generated dataset as $\mathbf{D}_M^\kappa$, where M is the number of channels and κ is the number of existing latent directed edges. Further details about the configuration of each $\mathbf{D}_M^\kappa$ are provided in Table 4. Synthetic data generation is performed by using the generative model of MC^{2}VAE, given by Equation (2). Hereafter, we detail and motivate the steps followed for data generation. Firstly, we randomly generated Directed Acyclic Graphs (DAGs) on M nodes by initializing an upper-triangular $M \times M$ matrix A, with entries uniformly drawn from the interval $[-2, 2]$, then masked by a Bernoulli matrix with a chosen success probability, to induce sparsity. A random permutation of the rows and columns of A has sometimes been applied to shuffle the channels' ordering (see Supplementary Figure B.1 for a few examples or randomly generated latent causal DAGs). Next, latent exogenous variables are sampled independently from a standard normal distribution. When accounting for the effect of a covariate, a binary variable was also randomly generated, as well as the associated M-dimensional (sparse) vector β which modulates the causal influence of the covariate over each channel. We further performed causal mixing by using the SCM given in (7) for several possible choices of $\mathbf{f}_{\text{det}}$ (see Table 4), hence apply linear decoding to obtain the final multi-channel dataset with the required number of features per

channel. Of note, by imposing a latent causal structure between modalities and covariates we are synthetically establishing the underlying model of explained variation of observed modalities, upon decoding. Indeed, referring to the notation from Section 3, a latent causal relationship from z^{c,m_1} towards z^{c,m_2} implies that a variation in z^{c,m_1} induces a variation in z^{c,m_2}, which is subtly reflected by sampling observations in the higher dimensional spaces where $\mathbf{x}^{m_1}$ and $\mathbf{x}^{m_2}$ respectively lies. We finally added to each channel some additive Gaussian noise to further mimic the existence of unexplained random noise in the observation space. Classical standardisation is applied to each dataset before being fed to MC^{2}VAE.

Table 4
Configurations of the generated datasets $\mathbf{D}_M^\kappa$. The overline on the number of edges (e.g., $\overline{10}$) indicates that a permutation of the adjacency matrix A has been applied.

M	κ	# features/channel	# subjects	$\mathbf{f}_{\text{det}}$	Covariate
3	6	$[10, 5, 7]$	2000	Tanh	Yes
4	$\{8, 10, \overline{10}\}$	$[10, 30, 35, 40]$	500	Polynomial	Yes
5	$\{10, 12\}$	$[10, 5, 7, 15, 20]$	5000	Cosine	Yes
8	24	$[5, 10, 15, 20, 25, 30, 35, 40]$	8000	Tanh	No

4.3.2. Performance of graph discovery

We study the ability of MC^{2}VAE to learn the latent cross-channels causal graphs, and evaluate the impact of the different proposed architectures as well as the complexity of the ground truth causal model in terms of the number of channels and edges. The hyperparameters used to train MC^{2}VAE for each scenario are provided in Supplementary Table B.1. For each dataset we perform a 10-fold cross validation strategy. As a consequence, for every experiment we end up with 10 final estimated adjacency matrices, which may underline some slightly different causal structures. We propose two possible strategies to combine such matrices into a consensus one. The first strategy consists in retaining the estimated A corresponding to the fold in which the reconstruction error (either on the train dataset - Min(MSE Train) - or on the test one - Min(MSE Test)), evaluated through mean squared error, is minimized. The second strategy proposes to keep those edges that appear in a minimum number of folds f over the 10 folds (Freq.=f): as thresholds for the frequency of occurrence we consider 3, 4 or 5 out of 10.

In Figure 2 (above), we evaluate the graph discovery performance across the four considered architectures using the above-mentioned strategies for all generated datasets. We consider two criteria, TPR and SHD, and report the mean values, and the corresponding standard deviations. Since SHD is typically expected to increase with the graph dimension, to ease the comparison Figure 2 (above) , right panel, shows for each architecture the SHD score normalized by the number of channels. In both plots we also report the scores obtained in the best fold, averaged across all datasets and for each architecture, as a reference (Best(fold)). Figure 2 (below) further isolates the results obtained using Min(MSE Test) and Freq.=3 on TPR and SHD respectively,

to better appreciate their overall performance. The reported results highlight the ability of MC^{2}VAE to accurately recover the latent causal graph, across the different generated datasets. Concerning the choice of the architecture, the results reported in Figure 2 do not provide us with a strong rationale to privilege Architectures 2 or 3 where the non-linear function $\mathbf{f}_{\text{det}}$ is explicitly considered during training: therefore, in what follows, we will only focus on Architectures 1 and 4.

In Figures 3 and 4, we show the results of the causal graph discovery with respect to all metrics discussed in Section 4.2 (*i.e.*, FDR, TPR, FPR and SHD), and for 6 synthetically generated data considering Architectures 1 and 4 respectively. Equivalent results for Architectures 2 and 3 are left to Supplementary Figures B.2 and B.3. MC^{2}VAE demonstrates robust performance in edge discovery, regardless of the number of modalities (or channels) and features, for both

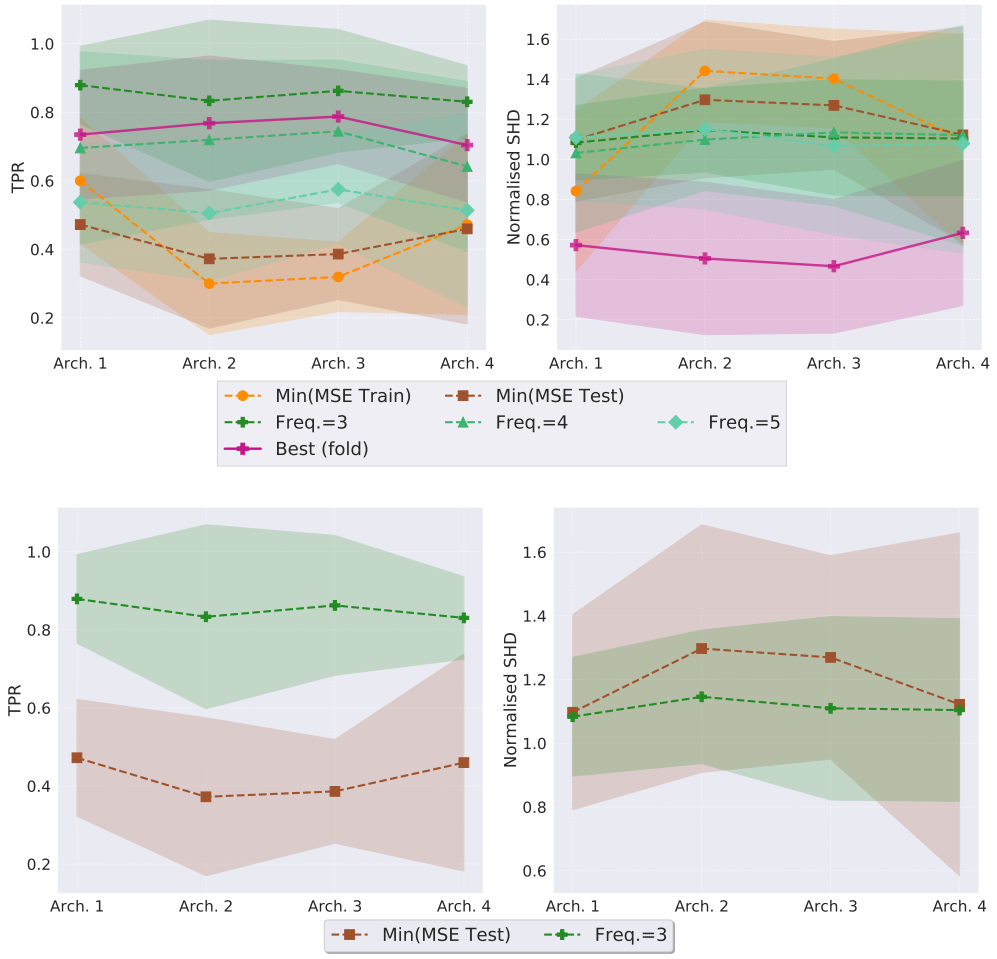

Figure 2: Above: TPR and normalized SHD over all the datasets comparing all architectures. Below: TPR and normalized SHD per architecture over all the datasets using the criteria MSE(test) and minimum frequency of occurrences 3. The dotted lines represent the mean values over all the datasets.

architectures, with the best fold results reaching in some cases a perfect reconstruction of the latent causal graph. Additionally, we notice that the graph selection strategies based on Min(MSE Test) and on minimum edge occurrence Freq.=3 or 4 seem particularly effective, ensuring good results on all tested datasets.

Overall, our method shows strong adaptability across heterogeneous datasets and complex high-dimensional settings, allowing us to recover the latent causal relationships across multiple channels, making it a promising approach for real-world multi-channel causal disentanglement.

4.3.3. Prior knowledge integration

Prior knowledge on the causal relationships across the channels can be smoothly injected into MC^{2}VAE during training to drive the latent graph discovery task. To showcase the impact of injecting a prior over the causal structure, we consider here as an example dataset $\mathbf{D}_5^{10}$(see Table 4). We consider Architecture 4, and choose to use as a soft prior the latent graph structure which satisfied the criteria Freq.=5, rather than using the ground truth: this is motivated by the fact that one can easily put this 2-step strategy in place in a real-case scenario and in the absence of a strong prior. The prior is provided to MC^{2}VAE in the form of a $M \times M$ binary matrix A^p, where an entry $a_{ij}^p = 1$ indicates that an edge is expected from the i-th latent channel towards the

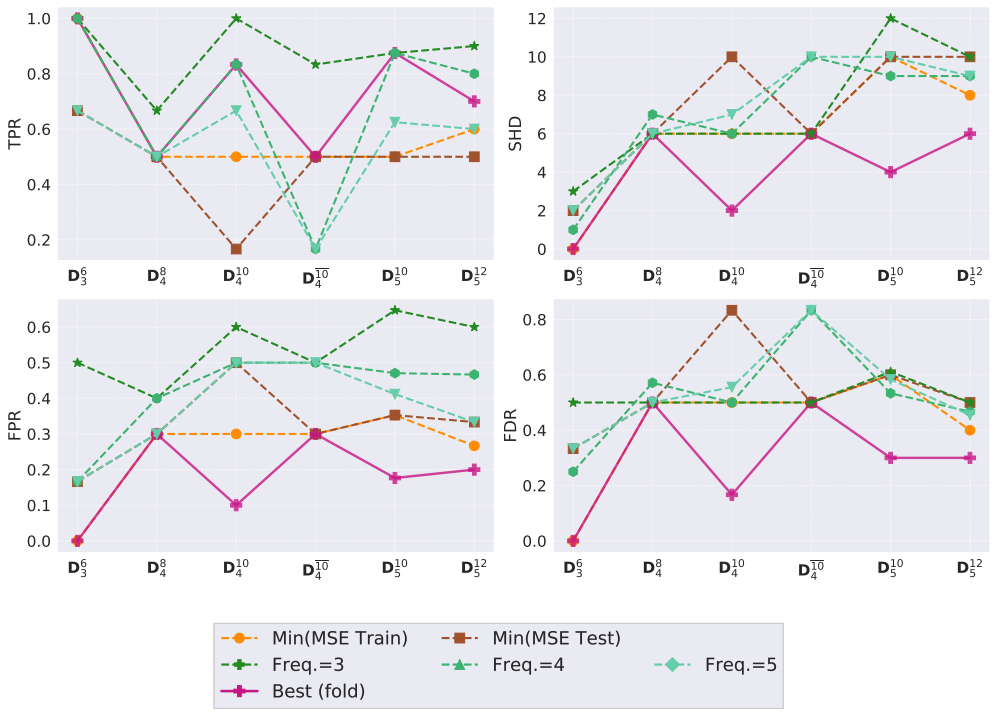

Figure 3: Architecture 1. Performance of graph discovery in terms of TPR, SHD, FPR and FDR across six synthetic datasets. The magenta curves, *Best(fold)*, represent the best results achieved by MC^{2}VAE in each panel across the different folds.

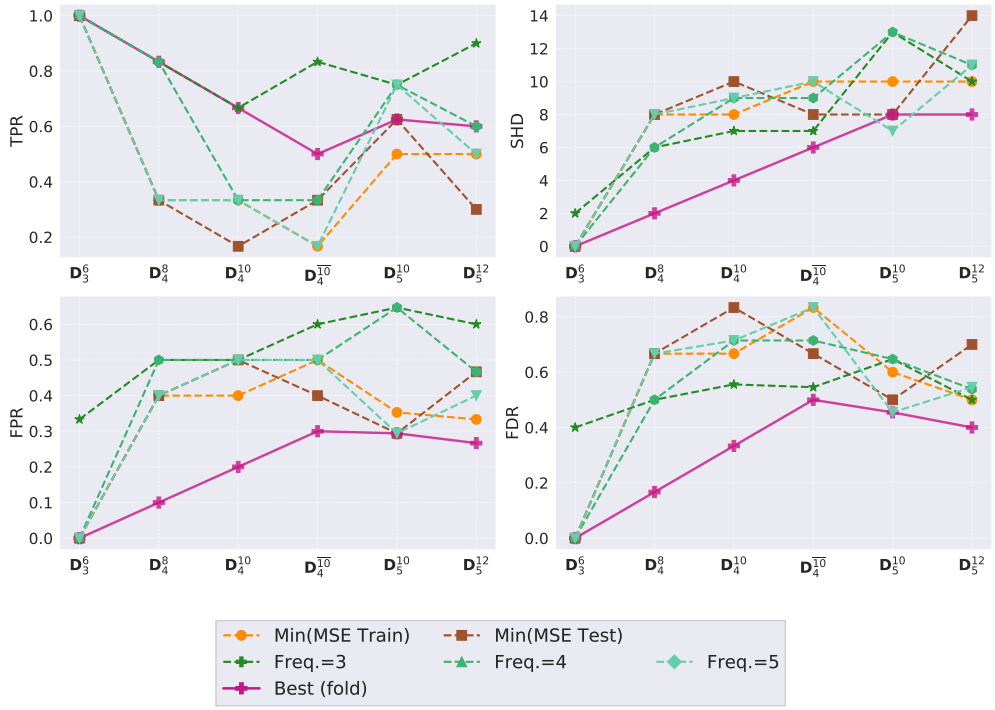

Figure 4: Architecture 4. Performance of graph discovery in terms of TPR, SHD, FPR and FDR across six synthetic datasets. The magenta curves, *Best(fold)*, represent the best results achieved by MC^{2}VAE in each panel across the different folds.

j-th one. The hyperparameters required by MC^{2}VAE for training are kept unchanged, except for the regularisation weights, $\{\eta_i\}_{i=1,2,3}$ which were divided by 10.

Table 5
Performance of reconstruction of matrix A with and without prior.

	w/o prior	**w/ prior**
TPR ↑	0.62	0.75
FPR ↓	0.29	0.25
FDR ↓	0.5	0.11
SHD ↓	8	4

In Figure 5 we show the ground truth matrix used for data generation (left panel), the best matrix obtained by training MC^{2}VAE without any prior (middle panel) and the one we get after injecting some prior knowledge. One can appreciate the benefits of prior injection: almost all latent causal relationships are correctly identified, except for two, which have been inverted. The prior knowledge enhances the accuracy of the discovered adjacency matrix, as reported in Table 5: TPR increases by around 10% and FDR and SHD strongly decrease. These results suggest

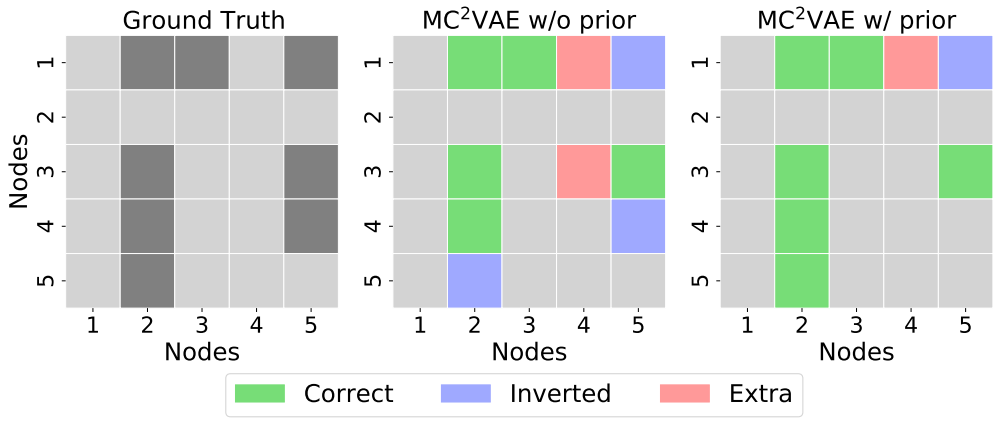

Figure 5: From left to right: Ground truth adjacency matrix A (dark (light) gray boxes indicate the presence (absence) of causal relationships); best matrix A predicted by MC^{2}VAE without prior using Architecture 4; best matrix A predicted by MC^{2}VAE using Architecture 4 and a prior on the graph structure.

that leveraging prior structural knowledge allows to significantly improve the reconstruction accuracy by boosting true positives and reducing false discoveries.

4.3.4. Benchmark

We compare our method with two baselines: MCVAE [11], a VAE-based approach for the joint analysis of multi-channel data, to assess data reconstruction abilities, and DAG-GNN [35], a causal disentanglement method based on graph neural networks, to evaluate the graph structure learning capabilities. For a fair comparison with the chosen baseline methods—which are not equipped to explicitly account for covariates—we conducted both studies using the synthetic dataset $\mathbf{D}_8^{24}$ (see Table 4), generated without any covariate.

We trained MCVAE using the default hyperparameters and set the latent space dimension to 1, as done for MC^{2}VAE and for the data generation process. The number of training epochs was set to 2000, which was enough to ensure convergence. The reconstruction performance comparing MCVAE and MC^{2}VAE, and measured by the mean squared error, is presented in Figure 6, which clearly shows an improved reconstruction ability of our method with respect to MCVAE. This result demonstrates the relevance of integrating the latent causal layer, which allows to model at a finer granularity the mutual causal relationships between channels, while the underlined hypothesis of MCVAE is rather an enforced correlation across channels' latent spaces.

Since DAG-GNN is not inherently designed to handle multi-channel data, we decided to first independently reduce each channel to a single dimension by using Principal Component Analysis (PCA). The resulting one-dimensional projections are then concatenated and used as input to DAG-GNN. As for the previous experiments, we performed again a 10-fold cross-validation strategy. In Figure 7, we compare the ground truth adjacency matrix (left panel) with the adjacency matrices obtained by DAG-GNN (middle panel) and by our proposed method

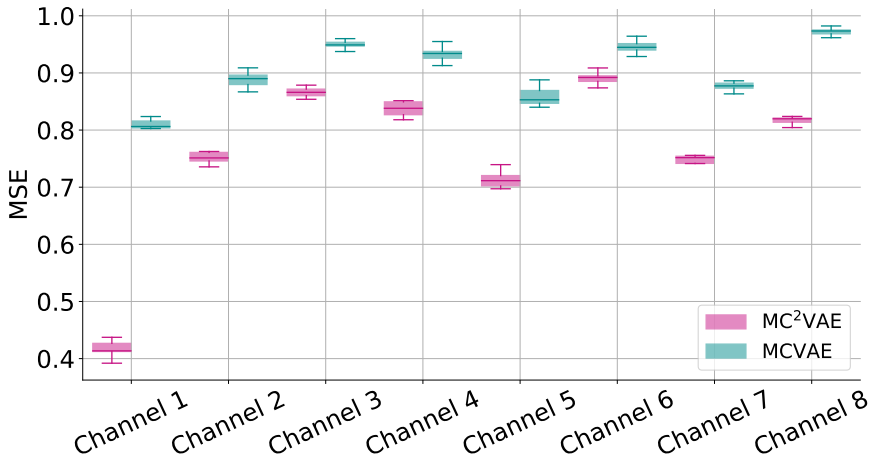

Figure 6: Mean MSE per channel on the testing data for $\mathbf{D}_8^{24}$, obtained by performing a 10-fold cross-validation. Comparison between MCVAE [11] and our model.

(right panel) using Architecture 4. For DAG-GNN, the displayed adjacency matrix corresponds to the fold with the highest True Positive Rate (TPR), and, following the authors' original procedure, we applied a threshold of $1e^{-3}$ to the matrix weights to improve the interpretability of the output. Notably, without thresholding, DAG-GNN produces a fully connected graph, lacking structural specificity. In contrast, for our method, we selected the matrix from the fold that achieved the lowest mean squared error (MSE) on the test set. One can observe that MC^{2}VAE effectively identifies the complex underlying latent causal structure, achieving a TPR of 0.71, showing a clear improvement compared to DAG-GNN. Of course, we recognize that DAG-GNN has not been developed for the causal disentanglement of multi-channel data: this was also the case for the other causal disentanglement methods available in the literature, to the best of our knowledge.

5. Application to real multi-channel data on Alzheimer Disease

We apply MC^{2}VAE to multi-modal medical imaging data and clinical scores extracted from the Alzheimer's Disease Neuroimaging Initiative dataset (ADNI)[2]. We considered a total of 587 observations for 441 participants among cognitively normal (103 subjects), patients diagnosed with Mild Cognitive Impairment (236 subjects) and patients diagnosed with Alzheimer's disease (AD - 102 subjects), at baseline. Among the 441 subjects, 133 were associated with more than one observation over time (up to a maximum of 3 observations per subject). However, since our objective is to build a meaningful latent causal structure across channels, and not to model

[2]The ADNI project was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of early Alzheimer's disease (AD) (see www.adni-info.org for up-to-date information).

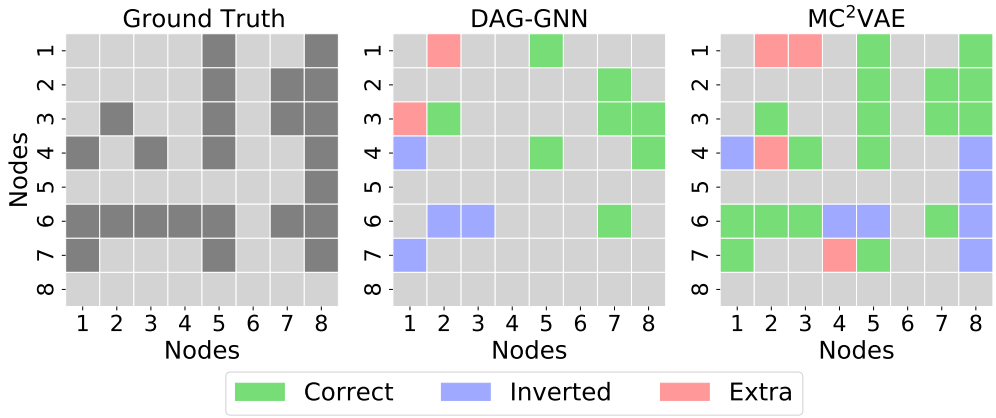

Figure 7: From left to right: the ground truth causal graph of $\mathbf{D}_8^{24}$; the adjacency matrix predicted by DAG-GNN; the adjacency matrix predicted by MC^{2}VAE. Edge colors represent edges classification outcomes: green for correctly identified edges, blue for inverted edges, and pink for false positive (spurious) edges.

the subject-specific temporal dynamics of the disease, each observation is treated here as independent. Of note, the information about the baseline diagnosis has not been provided at any stage to our model to perform the presented analysis.

All participants are associated with data coming from four channels: *Cognitive scores*, *Magnetic resonance imaging* (MRI), *Fluorodeoxyglucose-PET* (FDG-PET) and *AV45-Amyloid PET images* (AV45-PET). Cognitive scores include clinical scores such as Mini-Mental State Examination (MMSE), Clinical Dementia Rating scale (CDR-SB), AD Assessment Scale—Cognitive subtest (ADAS-Cog-11) and Rey Auditory Verbal Learning Test (RAVLT). Before applying MC^{2}VAE, MRI morphometrical biomarkers were extracted as regional volumes using the cross-sectional pipeline of FreeSurfer v6.0[3] and the Desikan-Killiany parcellation [41]. Measurements from AV45-PET and FDG-PET were estimated by co-registering each modality to their respective MRI space, normalizing by the cerebellum uptake and by computing regional amyloid load and glucose hypometabolism using PetSurfer[4] pipeline [42] and the same parcellation. Features were corrected beforehand with respect to intra-cranial volume, sex, and age using a multivariate linear model. The dimensions of the finally obtained 4-channel tabular data are reported in Supplementary Table B.2. Finally, we consider for each subject the recorded information about the presence of $\epsilon 4$ allele of apolipoprotein E (APOE4), the strongest known genetic risk factor for AD, affecting both the disease onset and progression [43]. APOE4 will play here the role of a (binary) covariate.

In Figure 8 we present the weighted latent causal graph relating the four considered channels and APOE4, as inferred by MC^{2}VAE using Architecture 4. Black and blue colored arrows distinguish the causal relationships between channels and the covariate effects, respectively. The training has been realized as described in Section 4.3.3. Specifically, we first performed a 10-fold cross validation, built a soft prior using the frequency criterion, hence reported here the

[3] https://surfer.nmr.mgh.harvard.edu/fswiki
[4] https://surfer.nmr.mgh.harvard.edu/fswiki/PetSurfer

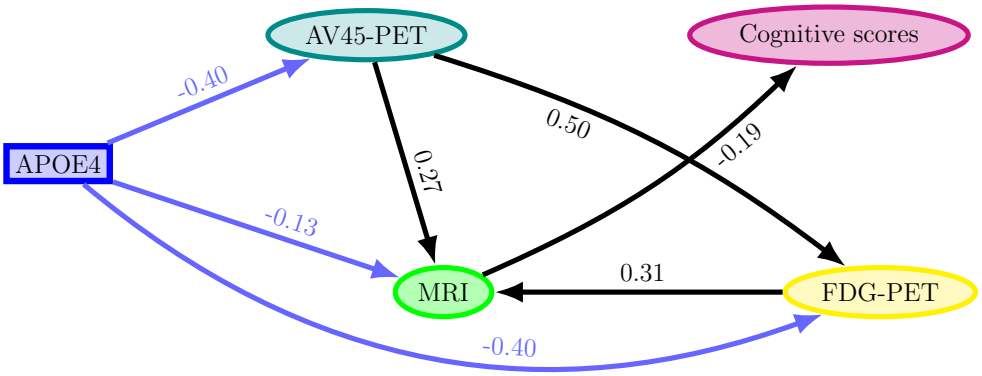

Figure 8: Weighted causal graph estimated by MC^{2}VAE for the ADNI dataset.

final weighted graph that minimizes the MSE over the test data. The directed arrows whose weights were lower than 0.1 were disregarded[5].

Following MC^{2}VAE's definition and global scope, the obtained latent causal structure has to be interpreted as high-level causal associations between the specific information gathered by each modality (and the covariate), all of them contributing to each patient's phenotyping. Very interestingly, the final causal graph estimated by MC^{2}VAE is highly correlated with and provides further evidence for the established hypothetical temporal cascade of the biomarker model of AD's pathophysiology [44, 45, 46]. Indeed, the latest reports an initial brain accumulation of Amyloid-β (Aβ) plaques, which can be identified by amyloid PET, followed by the spreading of tau and synaptic dysfunction, evidenced by FDG-PET, then regional brain atrophy and neuronal injury measured through MRI. Clinical manifestations, which can be assessed by cognitive scores, ultimately follows. Our graph, obtained in a fully unsupervised manner, provides a causal interpretation of the AD's progression model, suggesting that Aβ accumulation plays a central role in determining the course of the disease, and giving further evidence to support the development of amyloid-targeted therapies [47]. Concerning the effect of APOE4, several studies have already underlined the association of APOE4 with an enhanced deposition of Aβ plaques and an increased risk of AD cognitive-functional decline [48, 46]. More recently, Rasi *et al.* [49] demonstrated that FDG-PET can be used to predict the presence of the APOE4 genotype, providing empirical support for the causal relationship from APOE4 towards FDG-PET predicted by MC^{2}VAE.

Finally, we should stress that most of the identified causal relationships are independently supported by multiple studies. For instance, [50] demonstrates that MRI scans can accurately predict clinical scores such as MMSE, CDR-SB, and ADAS-Cog across independent datasets, thus providing evidence for the discovered causal pathway MRI→Cognitive scores. Also, MRI features have been shown to correlate with FDG-PET metabolic indices [51].

Interpreting the weights and signs associated with each predicted arrow is slightly less

[5]This corresponds to edges from APOE4, AV45-PET and FDG towards Cognitive scores, with absolute weights of 0.04, 0.04 and 0.07, respectively.

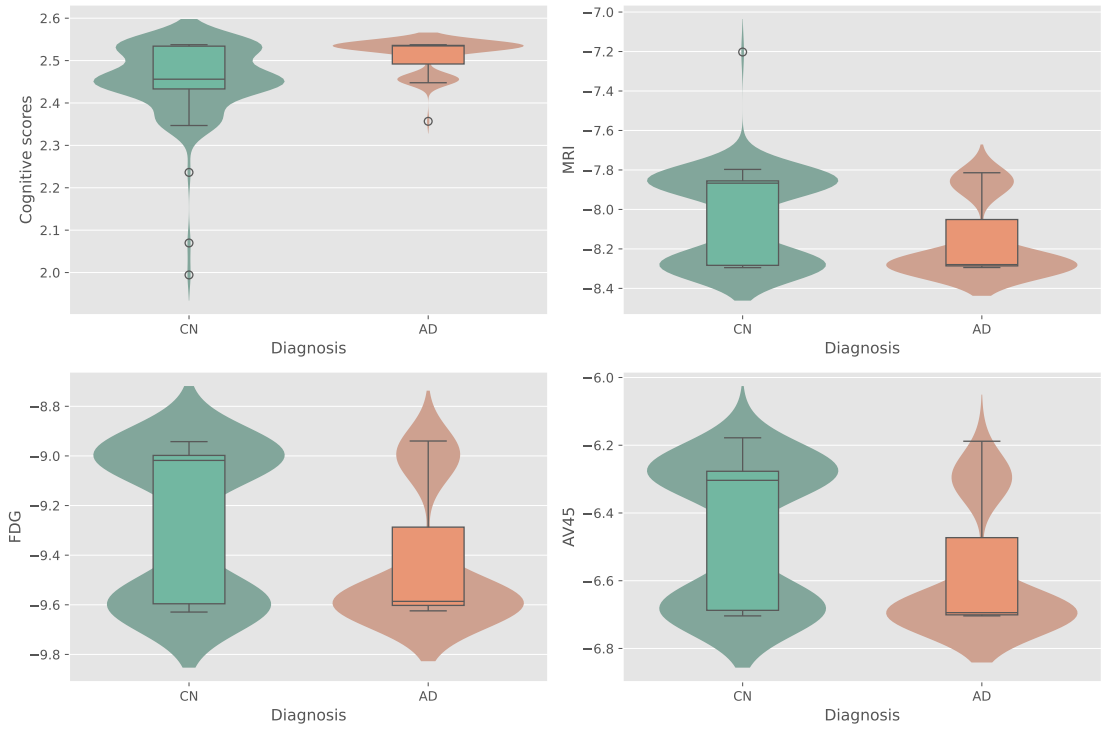

Figure 9: Distribution of each latent modality comparing cognitively normal (CN) and Alzheimer's disease (AD) samples. Both Mann-Whitney U test and the Student's t-test confirm a statistically significant difference between the two groups ($p < 1e-03$).

straightforward than the structure of the graph itself. The absolute values of the weights provide information about hierarchical causal strength. For example, the presence of APOE4 affects both PET modalities more strongly than MRI, in accordance with what already commented above, and a change captured by AV45-PET will induce greater changes in FDG-PET than in MRI. Concerning the sign, it can be partially determined by the encoding-decoding operations, typically rotation invariant. Nevertheless, the sign still provides an indication of the sense of the variation induced by the causal associations (modulo the post-nonlinear model specification, see Section 3). One can have a better idea of what this means in practice for the specific case of the current application on ADNI data by looking at Figure 9, where we plot the distribution of the latent causally related modalities, comparing normal (CN) and diseased (AD) samples. One can observe that the CN population is characterized by lower average *Cognitive scores*, and a higher average MRI, FDG-PET and AV45-PET, elucidating the relationship between the negative sign driving the causal association MRI→*Cognitive scores*, and the positive sign across the imaging modalities. The bimodal shape of each distribution is finally due to the effect of APOE4.

As a concluding remark, our findings provide evidence to support the hypothesized biologically grounded progression of AD from a causal perspective, and highlight the utility of our

model in uncovering meaningful multi-channel causal pathways for biomedical data, with a potential strong implication for actionable healthcare. Indeed, by identifying relevant causal factors rather than merely associative patterns, our framework provides interpretable targets that can inform about patient phenotyping, help design early intervention strategies, and thus potentially guide interventions. Notably, MC^{2}VAE results provide support for research efforts towards Amyloid β-based therapy [52], the prime target for the development of Alzheimer's disease therapy.

6. Conclusions and future directions

In this paper, we present a novel model for unsupervised Bayesian causal disentanglement from multi-channel data, MC^{2}VAE. Our method leverages variational autoencoders to obtain a compact representation of each channel, but conversely to classical VAEs, it integrates a causal discovery layer to unveil the underlying latent causal relationships across channels. Our approach showed promising results when applied to synthetically generated datasets: MC^{2}VAE is able to detect almost all between-channels causal relationships in all considered scenarios, ensuring good reconstruction abilities. Furthermore, the application to a real multi-channel dataset on neurodegeneration highlights the relevance of MC^{2}VAE in discovering meaningful and actionable causal patterns. This work opens up to several exciting perspectives and extensions, including, among others, the possibility of deploying the learned latent multi-channel causal structure, and modeling interventions on the latent graph, hence expanding to the causal inference domain.

Acknowledgments

This work was supported by the European Union's Horizon 2020 Research and Innovation Program *SimCardioTest* project under grant agreement No. 101016496 and the French government through the National Research Agency (ANR) RHU Talent project (ANR-23-RHUS-0015).
The authors would like to thank *Maxime Sermesant* for his useful comments and suggestions.

References

[1] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, Y. Luo, Multimodal machine learning in precision health: A scoping review, npj Digital Medicine 5 (2022) 171.

[2] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, A. Lerchner, beta-vae: Learning basic visual concepts with a constrained variational framework., ICLR (Poster) 3 (2017).

[3] Y. Kim, S. Wiseman, A. Miller, D. Sontag, A. Rush, Semi-amortized variational autoencoders, in: International Conference on Machine Learning, PMLR, 2018, pp. 2678–2687.

[4] S. Feuerriegel, D. Frauen, V. Melnychuk, J. Schweisthal, K. Hess, A. Curth, S. Bauer, N. Kilbertus, I. S. Kohane, M. van der Schaar, Causal machine learning for predicting treatment outcomes, Nature Medicine 30 (2024) 958–968.

[5] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (2013) 1798–1828.
[6] X. Wang, H. Chen, S. Tang, Z. Wu, W. Zhu, Disentangled representation learning, arXiv preprint arXiv:2211.11695 (2022).
[7] L. R. Medsker, L. Jain, et al., Recurrent neural networks, Design and Applications 5 (2001) 2.
[8] M. Suzuki, K. Nakayama, Y. Matsuo, Joint multimodal learning with deep generative models, arXiv preprint arXiv:1611.01891 (2016).
[9] M. Sewak, Deep reinforcement learning, Springer, 2019.
[10] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
[11] L. Antelmi, N. Ayache, P. Robert, M. Lorenzi, Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data, in: International Conference on Machine Learning, PMLR, 2019, pp. 302–311.
[12] C. Abi Nader, N. Ayache, G. B. Frisoni, P. Robert, M. Lorenzi, A. D. N. Initiative, Simulating the outcome of amyloid treatments in alzheimer's disease from imaging and clinical data, Brain communications 3 (2021) 1–17.
[13] H. Kameoka, L. Li, S. Inoue, S. Makino, Supervised determined source separation with multichannel variational autoencoder, Neural Computation 31 (2019) 1891–1914. doi:10.1162/neco_a_01217.
[14] S. Seki, H. Kameoka, L. Li, T. Toda, K. Takeda, Underdetermined source separation based on generalized multichannel variational autoencoder, IEEE Access 7 (2019) 168104–168115. doi:10.1109/ACCESS.2019.2954120.
[15] M. Sood, A. Sahay, R. Karki, M. A. Emon, H. Vrooman, M. Hofmann-Apitius, H. Fröhlich, Realistic simulation of virtual multi-scale, multi-modal patient trajectories using bayesian networks and sparse auto-encoders, Scientific reports 10 (2020) 10971.
[16] Improving generative modelling in vaes using multimodal prior, IEEE Transactions on Multimedia 23 (2021) 2153–2161. doi:10.1109/TMM.2020.3008053.
[17] D. Wesego, A. Rooshenas, Score-based multimodal autoencoders, arXiv preprint arXiv:2305.15708 (2023).
[18] J. Pearl, Causality, Cambridge university press, 2009.
[19] J. Pearl, D. Mackenzie, The book of why: the new science of cause and effect, Basic books, 2018.
[20] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, S. A. Tsaftaris, Causal machine learning for healthcare and precision medicine, Royal Society Open Science 9 (2022) 220638.
[21] S. Al-Ali, J. Llopis-Lorente, M. T. Mora, M. Sermesant, B. Trenor, I. Balelli, A causal discovery approach to streamline ionic currents selection to improve drug-induced tdp risk assessment, in: 2023 Computing in Cardiology (CinC), volume 50, 2023, pp. 1–4. doi:10.22489/CinC.2023.009.
[22] V. Lagani, S. Triantafillou, G. Ball, J. Tegnér, I. Tsamardinos, Probabilistic computational causal discovery for systems biology, Uncertainty in biology: a computational modeling approach (2016) 33–73.
[23] G. Camps-Valls, A. Gerhardus, U. Ninad, G. Varando, G. Martius, E. Balaguer-Ballester,

R. Vinuesa, E. Diaz, L. Zanna, J. Runge, Discovering causal relations and equations from data, Physics Reports 1044 (2023) 1–68.
[24] B. Huang, K. Zhang, M. Gong, C. Glymour, Causal discovery and forecasting in nonstationary environments with state-space models, in: International conference on machine learning, Pmlr, 2019, pp. 2901–2910.
[25] I. Balelli, S. Al-Ali, E. Dumas, J. Abecassis, Chapter 14 - causality: fundamental principles and tools, in: M. Lorenzi, M. A. Zuluaga (Eds.), Trustworthy AI in Medical Imaging, The MICCAI Society book Series, Academic Press, 2025, pp. 297–314. URL: https://www.sciencedirect.com/science/article/pii/B9780443237614000262. doi:https://doi.org/10.1016/B978-0-44-323761-4.00026-2.
[26] P. Spirtes, C. Glymour, An algorithm for fast recovery of sparse causal graphs, Social science computer review 9 (1991) 62–72.
[27] P. Spirtes, C. N. Glymour, R. Scheines, Causation, prediction, and search, MIT press, 2000.
[28] S. Zhu, I. Ng, Z. Chen, Causal discovery with reinforcement learning, arXiv preprint arXiv:1906.04477 (2019).
[29] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, M. Jordan, A linear non-gaussian acyclic model for causal discovery., Journal of Machine Learning Research 7 (2006).
[30] A. G. Reddy, V. N. Balasubramanian, et al., On causally disentangled representations, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 8089–8097.
[31] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, J. Wang, Causalvae: Disentangled representation learning via neural structural causal models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9593–9602.
[32] A. Komanduri, Y. Wu, W. Huang, F. Chen, X. Wu, Scm-vae: Learning identifiable causal representations via structural knowledge, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 1014–1023.
[33] M. Kocaoglu, C. Snyder, A. G. Dimakis, S. Vishwanath, Causalgan: Learning causal implicit generative models with adversarial training, arXiv preprint arXiv:1709.02023 (2017).
[34] E. Walker, J. A. Actor, C. Martinez, N. Trask, Causal disentanglement of multimodal data, arXiv preprint arXiv:2310.18471 (2023).
[35] Y. Yu, J. Chen, T. Gao, M. Yu, Dag-gnn: Dag structure learning with graph neural networks, in: International conference on machine learning, PMLR, 2019, pp. 7154–7163.
[36] S. Al-Ali, M. T. Mora, M. Sermesant, B. Trénor, I. Balelli, Assessing ionic current blockades and electromechanical biomarkers' interrelations through a novel multi-channel causal variational autoencoder, in: 2024 Computing in Cardiology Conference, volume 51, 2024.
[37] X. Zheng, B. Aragam, P. K. Ravikumar, E. P. Xing, Dags with no tears: Continuous optimization for structure learning, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.
[38] I. Ng, A. Ghassami, K. Zhang, On the role of sparsity and dag constraints for learning linear dags, Advances in Neural Information Processing Systems 33 (2020) 17943–17954.
[39] K. Zhang, A. Hyvärinen, On the identifiability of the post-nonlinear causal model, in: Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI2009), 2009.

[40] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[41] B. Fischl, Freesurfer, Neuroimage 62 (2012) 774–781.
[42] D. N. Greve, C. Svarer, P. M. Fisher, L. Feng, A. E. Hansen, W. Baare, B. Rosen, B. Fischl, G. M. Knudsen, Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain pet data, Neuroimage 92 (2014) 225–236.
[43] J. Kim, J. M. Basak, D. M. Holtzman, The role of apolipoprotein e in alzheimer's disease, Neuron 63 (2009) 287–303.
[44] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, P. M. Thompson, The clinical use of structural mri in alzheimer disease, Nature reviews neurology 6 (2010) 67–77.
[45] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack Jr, J. Kaye, T. J. Montine, et al., Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease, Alzheimer's & dementia 7 (2011) 280–292.
[46] H. Hampel, J. Hardy, K. Blennow, C. Chen, G. Perry, S. H. Kim, V. L. Villemagne, P. Aisen, M. Vendruscolo, T. Iwatsubo, et al., The amyloid-β pathway in alzheimer's disease, Molecular psychiatry 26 (2021) 5481–5503.
[47] P. S. Aisen, B. Vellas, H. Hampel, Moving towards early clinical trials for amyloid-targeted therapy in alzheimer's disease, Nature reviews Drug discovery 12 (2013) 324–324.
[48] P. B. Verghese, J. M. Castellano, D. M. Holtzman, Apolipoprotein e in alzheimer's disease and other neurological disorders, The Lancet Neurology 10 (2011) 241–252.
[49] R. Rasi, A. Guvenis, Radiomics features of fdg pet images predict apoe4, European Journal of Radiology Artificial Intelligence 2 (2025) 100009. URL: https://www.sciencedirect.com/science/article/pii/S3050577125000076. doi:https://doi.org/10.1016/j.ejrai.2025.100009.
[50] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack Jr, J. Ashburner, R. S. Frackowiak, A. D. N. Initiative, et al., Predicting clinical scores from magnetic resonance scans in alzheimer's disease, Neuroimage 51 (2010) 1405–1413.
[51] L. Yan, C. Y. Liu, K.-P. Wong, S.-C. Huang, W. J. Mack, K. Jann, G. Coppola, J. M. Ringman, D. J. Wang, Regional association of pcasl-mri with fdg-pet and pib-pet in people at risk for autosomal dominant alzheimer's disease, NeuroImage: Clinical 17 (2018) 751–760. URL: https://www.sciencedirect.com/science/article/pii/S221315821730308X. doi:https://doi.org/10.1016/j.nicl.2017.12.003.
[52] Y. Zhang, H. Chen, R. Li, K. Sterling, W. Song, Amyloid β-based therapy for alzheimer's disease: challenges, successes and future, Signal transduction and targeted therapy 8 (2023) 248.

A. Some derivation details

A.1. Computation for $\mathbb{E}_{q_\Theta}\left\{\log\left[p_{(A,\beta)}\left(\mathbf{z}^c|\mathbf{z},v\right)\right]\right\}$ used in Equation (5)

$$
\begin{aligned}
&\mathbb{E}_{q_\Theta}\left\{\log\left[p_{(A,\beta)}\left(\mathbf{z}^c|\mathbf{z},v\right)\right]\right\} = \int q_\Theta(\mathbf{z}|\mathbf{X})\left(-\frac{M}{2}\log(2\pi) - \frac{1}{2}\det\left(B\right)\right.\\
&\left.-\frac{1}{2}\left(\mathbf{z}-\left(\mathbb{I}_M - A^\top\right)^{-1}\beta v\right)^\top B^{-1}\left(\mathbf{z}-\left(\mathbb{I}_M - A^\top\right)^{-1}\beta v\right)\right)d\mathbf{z}\\
=\;& -\frac{M}{2}\log(2\pi) - \frac{1}{2}\det(B)\\
&-\frac{1}{2}\mathbb{E}_{q_\Theta}\left[\left(\mathbf{z}-\left(\mathbb{I}_M - A^\top\right)^{-1}\beta v\right)^\top B^{-1}\left(\mathbf{z}-\left(\mathbb{I}_M - A^\top\right)^{-1}\beta v\right)\right]\\
=\;& -\frac{M}{2}\log(2\pi) - \frac{1}{2}\det\left(B\right) - \frac{1}{2}\left[tr\left(B^{-1}\Sigma^E(X,\boldsymbol{\theta})\right)\right.\\
&\left.+\left(\boldsymbol{\mu}^E(X,\boldsymbol{\theta}) - \left(\mathbb{I}_M - A^\top\right)^{-1}\beta v\right)^\top B^{-1}\left(\boldsymbol{\mu}^E(X,\boldsymbol{\theta}) - \left(\mathbb{I}_M - A^\top\right)^{-1}\beta v\right)\right].
\end{aligned}
$$

A.2. Computation of $\mathbb{E}_{q_\Theta}\left\{\log\left[p_{(A,\beta)}\left(\mathbf{z}^c|\mathbf{z},v\right)\right]\right\}$ used in Equation (9)

$$
\begin{aligned}
\mathbb{E}_{q_\Theta}\left\{\log\left[p_{(A,\beta)}\left(\mathbf{z}^c|\mathbf{z},v\right)\right]\right\} &= \int q_\Theta(\mathbf{z}|\mathbf{X})\Big(-\frac{M}{2}\log(2\pi) - \frac{1}{2}\det\left(\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))B\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))^\top\right)\\
&-\frac{1}{2}(\mathbf{z}-\mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta)))^\top\left(\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))B\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))^\top\right)^{-1}(\mathbf{z}-\mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta)))\Big)d\mathbf{z}\\
&= -\frac{M}{2}\log(2\pi) - \frac{1}{2}\det\left(\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))B\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))^\top\right)\\
&-\frac{1}{2}\mathbb{E}_{q_\Theta}\Big[(\mathbf{z}-\mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta)))^\top\left(\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))B\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))^\top\right)^{-1}(\mathbf{z}-\mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta)))\Big]\\
&= -\frac{M}{2}\log(2\pi) - \frac{1}{2}\det\left(\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))B\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))^\top\right)\\
&-\frac{1}{2}\Big[tr\Big(\left(\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))B\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))^\top\right)^{-1}\Sigma^E(X,\boldsymbol{\theta})\Big)\\
&+\left(\boldsymbol{\mu}^E(X,\boldsymbol{\theta})-\mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta))\right)^\top\left(\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))B\mathbf{J}_{\mathbf{f}_{\text{det}}}(\boldsymbol{\xi}(A,\beta))^\top\right)^{-1}\left(\boldsymbol{\mu}^E(X,\boldsymbol{\theta})-\mathbf{f}_{\text{det}}(\boldsymbol{\xi}(A,\beta))\right)\Big].
\end{aligned}
$$

B. Supplementary figures and tables

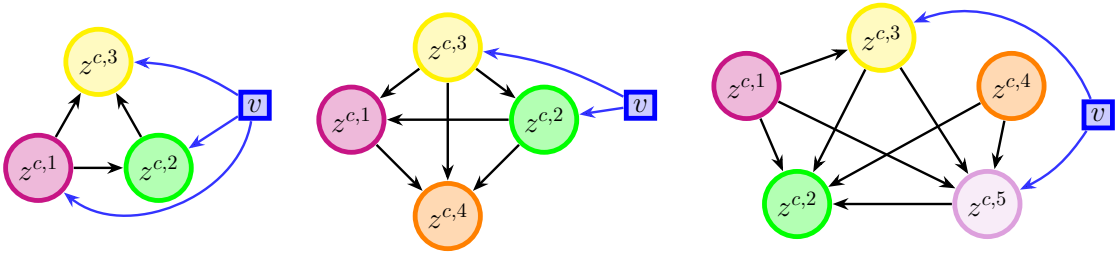

Figure B.1: Examples of causal graphs used to generate datasets $\mathbf{D}_3^6$, $\mathbf{D}_4^8$ and $\mathbf{D}_5^{10}$ presented in Table 4. The covariate is denoted by v.

Table B.1
Hyper-parameters used to train MC^{2}VAE on generated datasets given in Table 4.

# channels	batch size	# epochs
3	24	1500
4	24	1000
5	50	1500
8	50	1000

Table B.2
Number of features for each modality in ADNI dataset used in this paper.

Modality	# of features
Cognitive scores	7
Magnetic resonance imaging	41
Fluorodeoxyglucose-PET	41
AV45-Amyloid PET images	41

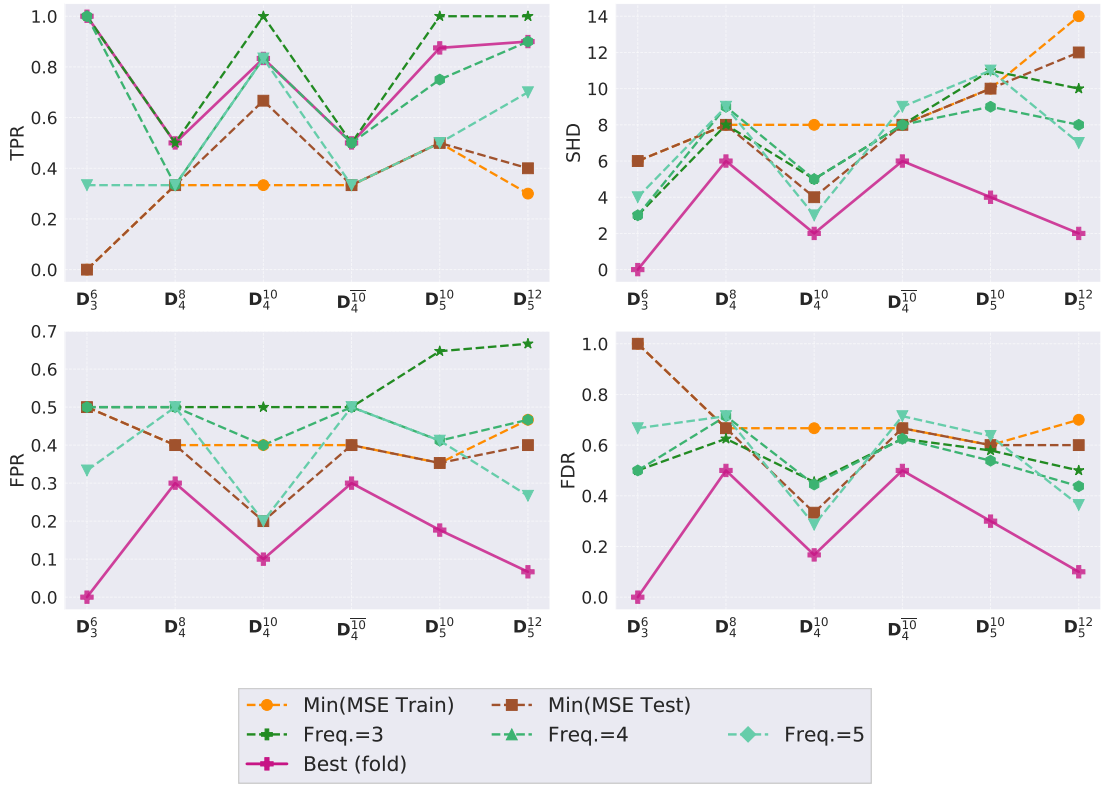

Figure B.2: Architecture 2. Performance of graph discovery in terms of TPR, SHD, FPR and FDR across six synthetic datasets. The magenta curves, *Best(fold)*, represent the best results achieved by MC^{2}VAE in each panel across the different folds.

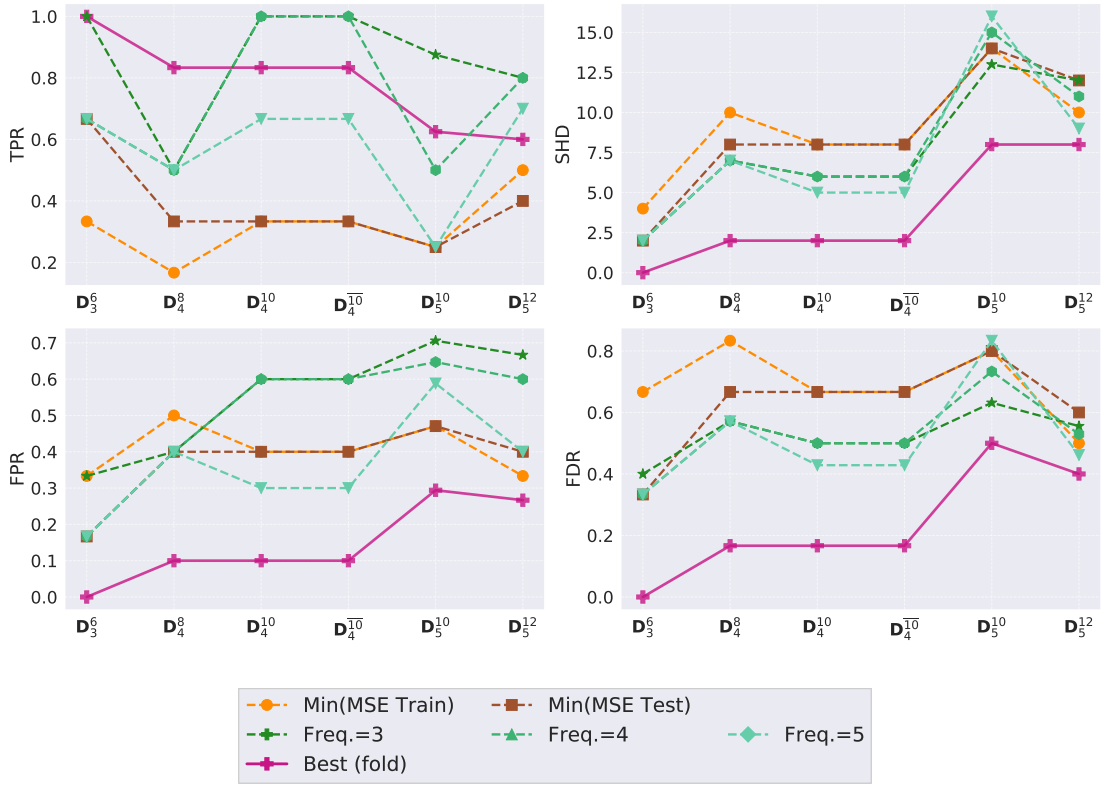

Figure B.3: Architecture 3. Performance of graph discovery in terms of TPR, SHD, FPR and FDR across six synthetic datasets. The magenta curves, *Best(fold)*, represent the best results achieved by MC^{2}VAE in each panel across the different folds.