

A Architecture

This section provides the definitions of the architectures described in the main paper.

Transformers [62] can be viewed as general set functions [34], making them ideally suited for NPs, which must ingest datasets. We begin by briefly overviewing transformers, defining the attention operations and how we construct a transformer layer, followed by how we integrate transformers into the MACE-TNP architecture.

A.1 Transformers

MHSA and MHCA Throughout this work we make use of two operations: multi-head self-attention (MHSA) and multi-head cross-attention (MHCA). Let $\mathbf{Z} \in \mathbb{R}^{N \times D_z}$ be a set of N tokens of dimensionality D_z . Then, for $\forall n = 1, \dots, N$, the MHSA operation updates this set of tokens as follows

$$\mathbf{z}_n \leftarrow \text{cat} \left(\left\{ \sum_{m=1}^N \alpha_h(\mathbf{z}_n, \mathbf{z}_m) \mathbf{z}_m^T \mathbf{W}_{V,h} \right\}_{h=1}^H \right) \mathbf{W}_O, \quad (5)$$

where $\mathbf{W}_{V,h} \in \mathbb{R}^{D_z \times D_V}$ and $\mathbf{W}_O \in \mathbb{R}^{HD_V \times D_z}$ are the value and projection weight matrices, H denotes the number of heads, and α_h is the attention mechanism. We opt for the most widely used softmax formulation

$$\alpha_h(\mathbf{z}_n, \mathbf{z}_m) = \text{softmax}(\{\mathbf{z}_n^T \mathbf{W}_{Q,h} \mathbf{W}_{K,h}^T \mathbf{z}_m\}_{m=1}^N)_m, \quad (6)$$

where $\mathbf{W}_{Q,h} \in \mathbb{R}^{D_z \times D_{QK}}$ and $\mathbf{W}_{K,h} \in \mathbb{R}^{D_z \times D_{QK}}$ are the query and key matrices.

The MHCA operation performs attention between two *different* sets of tokens $\mathbf{Z}_1 \in \mathbb{R}^{N_1 \times D_z}$ and $\mathbf{Z}_2 \in \mathbb{R}^{N_2 \times D_z}$. For $\forall n = 1, \dots, N_1$, the following update on $\mathbf{z}_{1,n}$ is performed:

$$\mathbf{z}_{1,n} \leftarrow \text{cat} \left(\left\{ \sum_{m=1}^{N_2} \alpha_h(\mathbf{z}_{1,n}, \mathbf{z}_{2,m}) \mathbf{z}_{2,m}^T \mathbf{W}_{V,h} \right\}_{h=1}^H \right) \mathbf{W}_O. \quad (7)$$

In order to obtain the attention blocks used within the transformer, these operations are typically combined with residual connections, layer-isations and point-wise MLPs.

More specifically, we define the MHSA operation as follows:

$$\begin{aligned} \tilde{\mathbf{Z}} &\leftarrow \mathbf{Z} + \text{MHSA}(\text{layer-norm}_1(\mathbf{Z})) \\ \mathbf{Z} &\leftarrow \tilde{\mathbf{Z}} + \text{MLP}(\text{layer-norm}_2(\tilde{\mathbf{Z}})). \end{aligned} \quad (8)$$

Similarly, the MHCA operation is defined as:

$$\begin{aligned} \tilde{\mathbf{Z}}_1 &\leftarrow \mathbf{Z}_1 + \text{MHCA}(\text{layer-norm}_1(\mathbf{Z}_1), \text{layer-norm}_1(\mathbf{Z}_2)) \\ \mathbf{Z}_1 &\leftarrow \tilde{\mathbf{Z}}_1 + \text{MLP}(\text{layer-norm}_2(\tilde{\mathbf{Z}}_1)). \end{aligned} \quad (9)$$

Masked-MHSA Consider the general case in which we want to update N token $\mathbf{Z} \in \mathbb{R}^{N \times D_z}$. There might be some situations where we want to make the update of a certain token $\mathbf{z}_n \in \mathbf{Z}$ independent of some other tokens. In that case, we can specify a set $M_n \subseteq \mathbb{N}_{\leq N}^+$ containing the indices of the tokens we want to make the update of \mathbf{z}_n independent of. Then, we can modify the pre-softmax activations within the attention mechanism $\tilde{\alpha}_h(\mathbf{z}_n, \mathbf{z}_m)$, where $\alpha_h(\mathbf{z}_n, \mathbf{z}_m) = \text{softmax}(\tilde{\alpha}_h(\mathbf{z}_n, \mathbf{z}_m))$ as follows:

$$\tilde{\alpha}_h(\mathbf{z}_n, \mathbf{z}_m) = \begin{cases} -\infty & \text{if } m \in M_n \\ \mathbf{z}_n^T \mathbf{W}_{Q,h} \mathbf{W}_{K,h}^T \mathbf{z}_m & \text{otherwise} \end{cases} \quad (10)$$

From the indices of M_n we can construct a binary masking matrix $\mathbf{M} \in \{0, 1\}^{N \times N}$:

$$\mathbf{M}_{n,m} = \begin{cases} 0 & \text{if } m \in M_n \\ 1 & \text{otherwise} \end{cases}$$

When used in the context of MHSA, we refer to this operation as masked-MHSA and represent it as $\mathbf{Z} = \text{masked-MHSA}(\mathbf{Z}, \mathbf{M})$.

919 A.2 Model-Averaged Causal Estimation Transformer Neural Processes (MACE-TNPs)

920 We refer to Nguyen and Grover [47], Ashman et al. [2] for a complete description of standard
 921 TNP architectures, and focus on describing the architecture of the MACE-TNP in more detail. Our
 922 proposed architecture is conceptually similar to the standard TNP architectures, but incorporates
 923 specific design choices and inductive biases that make it suitable for causal estimation.

924 We assume we have access to N_{obs} observational samples and want to predict the distribution of
 925 N_{int} interventional samples. The inputs to the MACE-TNP are: the observational dataset $\mathcal{D}_{\text{obs}} \in$
 926 $\mathbb{R}^{N_{\text{obs}} \times D \times d_{\text{data}}}$, the values of the node we intervene upon $\mathbf{x}_j \in \mathbb{R}^{N_{\text{int}}}$ (implying we intervene on node
 927 j), and the outcome node index i . Let $\mathcal{D}_{\text{obs},i} \in \mathbb{R}^{N_{\text{obs}} \times d_{\text{data}}}$ denote the observational data at node i . We
 928 omit the batch dimension for notational convenience.

929 **Data pre-processing** The first step involves pre-processing the interventional dataset by masking
 930 out all nodes except the one being intervened upon. Specifically, for each intervention, we zero out
 931 the values of the remaining $D - 1$ nodes. The resulting input to the model is an interventional matrix
 932 $\mathcal{D}_{\text{int}} \in \mathbb{R}^{N_{\text{int}} \times D \times d_{\text{data}}}$. Let $\mathcal{D}_{\text{int},i} \in \mathbb{R}^{N_{\text{int}} \times d_{\text{data}}}$ denote the interventional data at node i .

933 **Embedding** To differentiate between the different type of variables, we employ six different types
 934 of encodings, depending on the source of the data (observational (obs) or interventional (int)), and the
 935 type of the node (node we intervene upon (j), outcome node (i), or node we marginalise over). These
 936 are all performed using 2-layer MLPs of dimension d_{embed} . In the following we use $\mathcal{D}_{\text{obs},\{k \in [D] \setminus \{i,j\}\}}$
 937 to denote nodes in the observational dataset that are being marginalised over.

$$\begin{aligned}
 \text{observational, intervention node: } \mathbf{Z}_{\text{obs},j} &= \text{MLP}_{\text{obs},j}(\mathcal{D}_{\text{obs},j}) & (11) \\
 \text{observational, outcome node: } \mathbf{Z}_{\text{obs},i} &= \text{MLP}_{\text{obs},i}(\mathcal{D}_{\text{obs},i}) \\
 \text{observational, marginal nodes: } \mathbf{Z}_{\text{obs},\{k \in [D] \setminus \{i,j\}\}} &= \text{MLP}_{\text{obs}}(\mathcal{D}_{\text{obs},\{k \in [D] \setminus \{i,j\}\}}) \\
 \text{interventional, intervention node: } \mathbf{Z}_{\text{int},j} &= \text{MLP}_{\text{int},j}(\mathcal{D}_{\text{int},j}) \\
 \text{interventional, outcome node: } \mathbf{Z}_{\text{int},i} &= \text{MLP}_{\text{int},i}(\mathcal{D}_{\text{int},i}) \\
 \text{interventional, marginal nodes: } \mathbf{Z}_{\text{int},\{k \in [D] \setminus \{i,j\}\}} &= \text{MLP}_{\text{int}}(\mathcal{D}_{\text{int},\{k \in [D] \setminus \{i,j\}\}}),
 \end{aligned}$$

938 where $\{k \in [D] \setminus \{i,j\}\}$ represents the set of indices from $\{1, \dots, D\}$ excluding i and j . The
 939 representations are then concatenated back together in the original node order:

$$\mathbf{Z}_{\text{obs}} = \text{concat}([\mathbf{Z}_{\text{obs},k}]_{k \in [D]}), \quad \text{where } \mathcal{D}_k = \begin{cases} \mathbf{Z}_{\text{obs},i} & \text{if } k = i \\ \mathbf{Z}_{\text{obs},j} & \text{if } k = j \\ \mathbf{Z}_{\text{obs},k} & \text{otherwise} \end{cases}$$

$$\mathbf{Z}_{\text{int}} = \text{concat}([\mathbf{Z}_{\text{int},k}]_{k \in [D]}), \quad \text{where } \mathbf{Z}_k = \begin{cases} \mathbf{Z}_{\text{int},i} & \text{if } k = i \\ \mathbf{Z}_{\text{int},j} & \text{if } k = j \\ \mathbf{Z}_{\text{int},k} & \text{otherwise} \end{cases}$$

940 After the embedding stage, we obtain the representation of the observational dataset $\mathbf{Z}_{\text{obs}} \in$
 941 $\mathbb{R}^{N_{\text{obs}} \times D \times d_{\text{embed}}}$, and the representation of the interventional one $\mathbf{Z}_{\text{int}} \in \mathbb{R}^{N_{\text{int}} \times D \times d_{\text{embed}}}$.

942 **MACE Transformer Encoder** We utilise a transformer-based architecture composed of L layers,
 943 where we alternate between attention among samples, followed by attention among nodes. This
 944 choice preserves 1) permutation-invariance with respect to the observational samples, 2) permutation-
 945 equivariance with respect to the interventional samples, 3) permutation-invariance with respect to
 946 the nodes we marginalise over, and 4) permutation-equivariance with respect to the outcome and
 947 interventional nodes. Although we generally omit the batch dimension for convenience, we include it
 948 in this subsection to accurately reflect our implementation. Thus, the input to the MACE transformer
 949 encoder are the observational data representation $\mathbf{Z}_{\text{obs}} \in \mathbb{R}^{B \times N_{\text{obs}} \times D \times d_{\text{embed}}}$ and interventional data
 950 representation $\mathbf{Z}_{\text{int}} \in \mathbb{R}^{B \times N_{\text{int}} \times D \times d_{\text{embed}}}$, with B the batch size.

Attention among samples We propose two variants to perform attention among samples. We use the less costly MHSA + MHCA variant for the experiments in the main paper and show that it performs better in appendix [C.2.2](#)

1. **Masked-MHSA** among the observational and interventional samples: At each layer l , we first move the node dimension to the batch dimension for efficient batched attention: $\mathbf{Z}_{\text{obs}}^l \in \mathbb{R}^{B \times N_{\text{obs}} \times D \times d_{\text{embed}}} \rightarrow \mathbb{R}^{(B \times D) \times N_{\text{obs}} \times d_{\text{embed}}}$ and $\mathbf{Z}_{\text{int}}^l \in \mathbb{R}^{B \times N_{\text{int}} \times D \times d_{\text{embed}}} \rightarrow \mathbb{R}^{(B \times D) \times N_{\text{int}} \times d_{\text{embed}}}$. We then concatenate the two representations $\mathbf{Z}^l \in \mathbb{R}^{(B \times D) \times (N_{\text{obs}} + N_{\text{int}}) \times d_{\text{embed}}} = [\mathbf{Z}_{\text{obs}}^l, \mathbf{Z}_{\text{int}}^l]$, and construct a mask $\mathbf{M} \in \mathbb{R}^{N_{\text{obs}} + N_{\text{int}}}$ that only allows interventional tokens to attend to observational ones.

$$\mathbf{M}_{n,m} = \begin{cases} 1 & \text{if } m < N_{\text{obs}} \\ 0 & \text{otherwise} \end{cases}$$

We then perform masked-MHSA: $\mathbf{Z}^l = \text{masked-MHSA}(\mathbf{Z}^l, \mathbf{M})$. This strategy has a computational complexity $\mathcal{O}((N_{\text{obs}} + N_{\text{int}})^2)$.

2. **MHSA + MHCA**: An alternative, less costly strategy, is to perform MHSA on the observational data, followed by MHCA between the interventional and observational data. More specifically, as in the previous case we move the node dimension to the batch dimension and then perform:

$$\begin{aligned} \mathbf{Z}_{\text{obs}}^l &= \text{MHSA}(\mathbf{Z}_{\text{obs}}^l) \\ \mathbf{Z}_{\text{int}}^l &= \text{MHCA}(\mathbf{Z}_{\text{int}}^l, \mathbf{Z}_{\text{obs}}^l). \end{aligned}$$

We then concatenate the two representations into $\mathbf{Z}^l \in \mathbb{R}^{(B \times D) \times (N_{\text{obs}} + N_{\text{int}}) \times d_{\text{embed}}} = [\mathbf{Z}_{\text{obs}}^l, \mathbf{Z}_{\text{int}}^l]$. This strategy has a reduced computational cost of $\mathcal{O}(N_{\text{obs}}^2 + N_{\text{obs}}N_{\text{int}})$ and is the strategy we use for the results in the main paper.

Attention among nodes The output of the attention among samples at layer l $\mathbf{Z}^l \in \mathbb{R}^{(B \times D) \times (N_{\text{obs}} + N_{\text{int}}) \times d_{\text{embed}}}$ is then fed into the next stage: attention among nodes. We first reshape the data $\mathbf{Z}^l \in \mathbb{R}^{(B \times D) \times (N_{\text{obs}} + N_{\text{int}}) \times d_{\text{embed}}} \rightarrow \mathbf{Z}'^l \in \mathbb{R}^{(B \times (N_{\text{obs}} + N_{\text{int}})) \times D \times d_{\text{embed}}}$, and then perform MHSA between the nodes:

$$\mathbf{Z}^{l+1} = \text{MHSA}(\mathbf{Z}'^l)$$

This is then reshaped back into $\mathbf{Z}^{l+1} \in \mathbb{R}^{B \times (N_{\text{obs}} + N_{\text{int}}) \times D \times d_{\text{embed}}}$, and then split into the observational and interventional data representations that are fed into layer $l + 1$: $\mathbf{Z}_{\text{obs}}^{l+1} \in \mathbb{R}^{B \times N_{\text{obs}} \times D \times d_{\text{embed}}}$ and $\mathbf{Z}_{\text{int}}^{l+1} \in \mathbb{R}^{B \times N_{\text{int}} \times D \times d_{\text{embed}}}$.

MACE Decoder We parameterise the output distribution of the NP as a Mixture of Gaussians (MoG) with N_{comp} components. The NP outputs the mean, standard deviation and weight corresponding to each component for each interventional query $\{x_j^n\}_{n=1}^{N_{\text{int}}}$: $\{\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{w}\}(x_j^n) := \{\mu_k(x_j^n), \sigma_k(x_j^n), w_k(x_j^n)\}_{k=1}^{N_{\text{comp}}}$. These are computed based on the outcome interventional representation from the final layer of the MACE Transformer Encoder. More specifically, the input to the decoder is $\mathbf{Z}_{\text{int},i}^L \in \mathbb{R}^{N_{\text{int}} \times d_{\text{embed}}}$. This is then passed through a two-layer MLP of hidden size d_{emb} , followed by an activation function

$$\mathbf{z}_{\text{out}} = \text{activation}(\text{MLP}(\mathbf{Z}_{\text{int},i}^L))$$

Finally, we use linear layers to project the embedding $\mathbf{z}_{\text{out}} \in \mathbb{R}^{N_{\text{int}} \times d_{\text{embed}}}$ to the parameters of a mixture of N_{comp} Gaussian components:

$$\begin{aligned} \boldsymbol{\mu} &= \text{Linear}_{\text{mean}}(\mathbf{z}_{\text{out}}) \in \mathbb{R}^{N_{\text{int}} \times N_{\text{comp}}} \\ \text{pre-}\boldsymbol{\sigma} &= \text{Linear}_{\text{std}}(\mathbf{z}_{\text{out}}) \in \mathbb{R}^{N_{\text{int}} \times N_{\text{comp}}} \\ \text{pre-}\mathbf{w} &= \text{Linear}_{\text{weight}}(\mathbf{z}_{\text{out}}) \in \mathbb{R}^{N_{\text{int}} \times N_{\text{comp}}}. \end{aligned}$$

We then apply element-wise transforms to obtain valid parameters:

$$\boldsymbol{\sigma} = \text{softplus}(\text{pre-}\boldsymbol{\sigma}) \quad \mathbf{w} = \text{softmax}(\text{pre-}\mathbf{w}),$$

with the softmax being applied along the component dimension.

986 **Loss** The output parameters are then used to evaluate the per-dataset loss of the MACE-TNP, which,
 987 as shown in section 4 requires the evaluation of the log-posterior interventional distribution of the
 988 MoG. We restate the equation of the loss presented in section 4 for completeness:

$$\mathcal{L}_\theta(\mathbf{x}_i, \{\mu, \sigma, \mathbf{w}\}(\mathbf{x}_j)) = \sum_{n=1}^{N_{\text{int}}} \log p_\theta(x_i^n | \text{do}(x_j^n), \mathcal{D}) = \sum_{n=1}^{N_{\text{int}}} \log \left(\sum_{k=1}^{N_{\text{comp}}} w_k(x_j^n) \cdot \mathcal{N}(x_i^n | \mu_k(x_j^n), \sigma_k^2(x_j^n)) \right) \quad (12)$$

989 where $\mathcal{N}(x | \mu, \sigma)$ represents the Gaussian distribution with mean μ and standard deviation σ .

990 B Data Generation

991 We provide in fig. 4 a diagram showing how we sample training data from a specified Bayesian
 992 Causal Model to infer its posterior interventional distribution (see discussion in section 4).

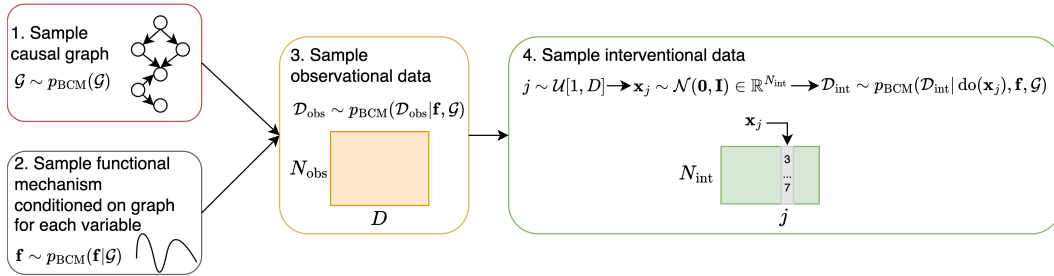


Figure 4: Overview of the data generation process. We first sample a graph \mathcal{G} , and a functional mechanism (conditioned on the sampled graph) for each of the D nodes in the dataset. These are then used to draw N_{obs} observational samples. To construct the interventional dataset, we first randomly sample a node to intervene upon j , draw N_{int} intervention values $\mathbf{x}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and set the values of node j to be \mathbf{x}_j . We then draw N_{int} samples of each node to form an interventional dataset \mathcal{D}_{int} .

993 B.1 Two-node Linear Gaussian Models

994 The data generation details for the two-node linear Gaussian experiments from section 5.1 and the
 995 derivations of the posterior interventional distribution are explained in this section.

996 We examine the basic scenario involving n independent and identically distributed (i.i.d.) random
 997 vectors, each consisting of two components, defined as $X^i := [X_1^i, X_2^i]^T$ for $i \in \{1, 2, \dots, n\}$.
 998 Let the observed dataset be denoted by $\mathcal{D}_{\text{obs}} := \{X^1, X^2, \dots, X^n\}$. For the sake of notational
 999 simplicity, we drop the subscript BCM from p_{BCM} in eq. (2) throughout the subsequent proofs. In
 1000 this setting, where the random vectors are composed of only two nodes (X_1, X_2), there exist three
 1001 distinct possible structural SCMs:

$$\mathcal{G}_1 := \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} : X_1 = wX_2 + U_1 \text{ and } X_2 = U_2 \quad (13)$$

$$\mathcal{G}_2 := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} : X_1 = U_1 \text{ and } X_2 = wX_1 + U_2 \quad (14)$$

$$\mathcal{G}_3 := \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} : X_1 = U_1 \text{ and } X_2 = U_2 \quad (15)$$

1002 We consider two models, one where the causal graph is identifiable (appendix B.1.1) and one where
 1003 it is not identifiable (appendix B.1.2).

1004 B.1.1 Identifiable Case

1005 We begin with the case where the error terms U_1 and U_2 are Gaussian distributed and the noise
 1006 variances of U_1 and U_2 are equal and known—a setting shown to be identifiable in Peters and
 1007 Bühlmann [52]. Fixing $\sigma^2, \sigma_w^2 \in \mathbb{R}^+$, we consider the following hierarchical model:

$$\begin{aligned} \mathcal{G} &\sim \mathcal{U}\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}, \quad U_i \sim \mathcal{N}(0, \sigma^2) \quad \text{for } i = 1, 2 \\ w &\sim \mathcal{N}(0, \sigma_w^2) \quad \text{if } \mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2\}. \end{aligned}$$

1008 which induces the following joint distribution:

$$p(X, w, \mathcal{G}) = p(X|w, \mathcal{G})p(w|\mathcal{G})p(\mathcal{G}) \text{ if } \mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2\} \text{ or} \quad (16)$$

$$p(X, \mathcal{G}_3) = p(X|\mathcal{G}_3)p(\mathcal{G}_3), \quad \text{otherwise.} \quad (17)$$

1009 We show below that the above models can be identified by the posterior.

1010 **Theorem B.1.** *Let $\mathcal{D}_{obs} := \{X^{(1)}, X^{(2)}, \dots, X^{(n)}\}$ be i.i.d. observations generated by*
 1011 *one of the simple models described in eqs. (13) to (15). The posterior over the graphs*
 1012 *$[p(\mathcal{G}_1|\mathcal{D}_{obs}), p(\mathcal{G}_2|\mathcal{D}_{obs}), p(\mathcal{G}_3|\mathcal{D}_{obs})]$ is*

$$\frac{1}{c} \left[\frac{\sigma}{\sqrt{(\sigma_w^2 S_2 + \sigma^2)}} \exp\left(\frac{\sigma_w^2}{2\sigma^2} \frac{S_{12}^2}{\sigma_w^2 S_2 + \sigma^2}\right), \frac{\sigma}{\sqrt{(\sigma_w^2 S_1 + \sigma^2)}} \exp\left(\frac{\sigma_w^2}{2\sigma^2} \frac{S_{12}^2}{\sigma_w^2 S_1 + \sigma^2}\right), 1 \right],$$

1013 where c is a constant of normalisation and

$$S_{12} := \sum_{i=1}^n X_1^i X_2^i, \quad S_1 := \sum_{i=1}^n X_1^{i2}, \quad S_2 := \sum_{i=1}^n X_2^{i2}. \quad (18)$$

1014 The posterior interventional distribution is a mixture of 2 Gaussian distributions

$$p(X_1 = y | \text{do}(X_2 = x), \mathcal{D}_{obs}) = p(\mathcal{G}_1 | \mathcal{D}_{obs}) \mathcal{N}(y | \mu_{1-2}(x), \sigma_{1-2}^2(x)) + (1 - p(\mathcal{G}_1 | \mathcal{D}_{obs})) \mathcal{N}(y | 0, \sigma^2),$$

1015 with

$$\mu_{1-2}(x) := \frac{\sigma_w^2 S_{12}}{\sigma_w^2 S_2 + \sigma^2} \cdot x \quad \text{and} \quad \sigma_{1-2}^2(x) := \sigma^2 \left(1 + \frac{\sigma_w^2 x^2}{\sigma_w^2 S_2 + \sigma^2} \right). \quad (19)$$

1016 **Proof:** First, we find the full conditional distribution, $p(w|\mathcal{G}, \mathcal{D}_{obs})$ and the posterior distribution over
 1017 the DAG models, $p(\mathcal{G}|\mathcal{D}_{obs})$. Following Bayes' rule we have

$$\begin{aligned} p(w|\mathcal{G}_1, \mathcal{D}_{obs}) &\propto p(w|\mathcal{G}_1)p(\mathcal{G}_1) \prod_{i=1}^n p(X_i|\mathcal{G}_1, w) \propto p(w|\mathcal{G}_1) \prod_{i=1}^n p(X_i|\mathcal{G}_1, w) \\ &\propto \exp\left(-\frac{w^2}{2\sigma_w^2} - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_1^i - wX_2^i)^2\right) \end{aligned}$$

1018 and by completing the square we obtain

$$p(w|\mathcal{G}_1, \mathcal{D}_{obs}) = \mathcal{N}\left(w \left| \frac{\sigma^2 S_{12}}{\sigma_w^2 S_2 + \sigma_1^2}, \frac{\sigma_1^2 \sigma_w^2}{\sigma_w^2 S_2 + \sigma_1^2} \right.\right), \quad (20)$$

1019 with S_{12} and S_2 being defined in eq. (18).

1020 Similarly, conditioned on the \mathcal{G}_2 model, the full conditional distribution is again Gaussian

$$p(w|\mathcal{G}_2, \mathcal{D}_{\text{obs}}) = \mathcal{N}\left(w \left| \frac{\sigma^2 S_{12}}{\sigma_w^2 S_1 + \sigma^2}, \frac{\sigma^2 \sigma_w^2}{\sigma_w^2 S_1 + \sigma^2} \right. \right). \quad (21)$$

1021 Using eq. (20) and eq. (21), the posterior for the \mathcal{G} is

$$\begin{aligned} p(\mathcal{G}_1|\mathcal{D}_{\text{obs}}) &\propto \int p(w)p(\mathcal{G}_1) \prod_{i=1}^n p(X_i|\mathcal{G}_1, w)dw \\ &\propto \exp\left(-\frac{S_1 + S_2}{2\sigma^2}\right) \frac{\sigma}{\sqrt{(\sigma_w^2 S_2 + \sigma^2)}} \exp\left(\frac{\sigma_w^2}{2\sigma^2} \frac{S_{12}^2}{\sigma_w^2 S_2 + \sigma^2}\right) \end{aligned} \quad (22)$$

$$\begin{aligned} p(\mathcal{G}_2|\mathcal{D}_{\text{obs}}) &\propto \int p(w)p(\mathcal{G}_2) \prod_{i=1}^n p(X_i|\mathcal{G}_2, w)dw \\ &\propto \exp\left(-\frac{S_1 + S_2}{2\sigma^2}\right) \frac{\sigma}{\sqrt{(\sigma_w^2 S_1 + \sigma^2)}} \exp\left(\frac{\sigma_w^2}{2\sigma^2} \frac{S_{12}^2}{\sigma_w^2 S_1 + \sigma^2}\right) \end{aligned} \quad (23)$$

$$p(\mathcal{G}_3|\mathcal{D}_{\text{obs}}) \propto p(\mathcal{G}_3) \prod_{i=1}^n p(X_i|\mathcal{G}_3) \propto \exp\left(-\frac{S_1 + S_2}{2\sigma^2}\right). \quad (24)$$

1022 Next, the posterior interventional distribution is

$$\begin{aligned} p(X_1|\text{do}(X_2 = x), \mathcal{D}_{\text{obs}}) &= p(X_1, \mathcal{G}_3|\text{do}(x), \mathcal{D}_{\text{obs}}) + \sum_{\mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2\}} \int p(X_1, \mathcal{G}, w|\text{do}(x), \mathcal{D}_{\text{obs}})dw \\ &= p(X_1|\text{do}(x), \mathcal{G}_3, \mathcal{D}_{\text{obs}})p(\mathcal{G}_3|\mathcal{D}_{\text{obs}}) + \sum_{\mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2\}} p(\mathcal{G}|\mathcal{D}_{\text{obs}}) \int p(X_1|\text{do}(x), \mathcal{G}, w)p(w|\mathcal{G}, \mathcal{D}_{\text{obs}})dw. \end{aligned}$$

1023 Conditioned on the model graphs, the interventional distributions are

$$\begin{aligned} p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_1, w) &= \mathcal{N}(y|wx, \sigma^2), \quad p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_2, w) = \mathcal{N}(y|0, \sigma^2) \\ p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_3) &= \mathcal{N}(y|0, \sigma^2). \end{aligned}$$

1024 Then, we have

$$\int p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_2, w)p(w|\mathcal{G}_2, \mathcal{D}_{\text{obs}})dw = \int \mathcal{N}(y|0, \sigma^2)p(w|\mathcal{G}_2, \mathcal{D}_{\text{obs}})dw = \mathcal{N}(y|0, \sigma^2),$$

$$\int p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_3, w)p(w|\mathcal{G}_3, \mathcal{D}_{\text{obs}})dw = \int \mathcal{N}(y|0, \sigma^2)p(w|\mathcal{G}_3, \mathcal{D}_{\text{obs}})dw = \mathcal{N}(y|0, \sigma^2),$$

1025 and

$$\begin{aligned} \int p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_1, w)p(w|\mathcal{G}_1, \mathcal{D}_{\text{obs}})dw &= \int \mathcal{N}(y|wx, \sigma^2)p(w|\mathcal{G}_1, \mathcal{D}_{\text{obs}})dw \\ &= \mathcal{N}\left(y \left| \frac{\sigma_w^2 S_{12}}{\sigma_w^2 S_2 + \sigma^2} \cdot x, \sigma^2 \left(1 + \frac{\sigma_w^2 x^2}{\sigma_w^2 S_2 + \sigma^2}\right) \right. \right). \end{aligned}$$

1026 Hence, the interventional distribution is simply a mixture of 2 Gaussian distributions

$$p(X_1 = y|\text{do}(X_2 = x), \mathcal{D}_{\text{obs}}) = p(\mathcal{G}_1|\mathcal{D}_{\text{obs}})\mathcal{N}\left(y|\mu_{1-2}(x), \sigma_{1-2}^2(x)\right) + (1 - p(\mathcal{G}_1|\mathcal{D}_{\text{obs}}))\mathcal{N}(y|0, \sigma^2),$$

1027 with $p(\mathcal{G}_1|\mathcal{D}_{\text{obs}})$ calculated in eqs. (22) to (24) and $\mu_{1-2}(x), \sigma_{1-2}^2(x)$ defined in eq. (34).

1028 Similarly, the next result easily follows

$$p(X_2 = y | \text{do}(X_1 = x), \mathcal{D}_{\text{obs}}) = p(\mathcal{G}_2|\mathcal{D}_{\text{obs}})\mathcal{N}(y|\mu_{2-1}(x), \sigma_{2-1}^2(x)) + (1 - p(\mathcal{G}_2|\mathcal{D}_{\text{obs}}))\mathcal{N}(y|0, \sigma^2),$$

1029 with $p(\mathcal{G}_2|\mathcal{D}_{\text{obs}})$ calculated in eq. (23) and

$$\mu_{2-1}(x) := \frac{\sigma_w^2 S_{12}}{\sigma_w^2 S_1 + \sigma^2} \cdot x \quad \text{and} \quad \sigma_{2-1}^2(x) := \sigma^2 \left(1 + \frac{\sigma_w^2 x^2}{\sigma_w^2 S_1 + \sigma^2} \right).$$

1030

□

1031 **Remark:** It can be shown that if \mathcal{D}_{obs} is generated by one of the models presented in eqs. (13)
 1032 to (15), then the posterior distribution $p(\mathcal{G} | \mathcal{D}_{\text{obs}})$ asymptotically concentrates around the true
 1033 data-generating structure \mathcal{G}^* [12, 65, 14]. Consequently, in the infinite data limit, the posterior
 1034 interventional distribution converges to a Gaussian distribution whose mean and variance depend on
 1035 the intervened node and the true underlying causal mechanism \mathcal{G}^* .

1036 B.1.2 Non-identifiable Case

1037 Second, we consider the errors' variances to be unknown while keeping the same SCMs described in
 1038 eqs. (13) to (15). Therefore, we place priors on these extra parameters as well chosen such that the
 1039 model is not identifiable [22]. We propose the following hierarchical model for fixed $\alpha > \frac{1}{2}$, and
 1040 $\beta, \eta > 0$

$$\mathcal{G} \sim \mathcal{U}\{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\}$$

$$\text{If } \mathcal{G} = \mathcal{G}_1, \text{ then } \tau_1^2 \sim \text{InvGamma}(\alpha, \beta), \tau_2^2 \sim \text{InvGamma}(\alpha - \frac{1}{2}, \beta), w \sim \mathcal{N}(0, \eta\tau_1^2)$$

$$\text{If } \mathcal{G} = \mathcal{G}_2, \text{ then } \tau_1^2 \sim \text{InvGamma}(\alpha - \frac{1}{2}, \beta), \tau_2^2 \sim \text{InvGamma}(\alpha, \beta), w \sim \mathcal{N}(0, \eta\tau_2^2)$$

$$\text{If } \mathcal{G} = \mathcal{G}_3, \text{ then } \tau_1^2 \sim \text{InvGamma}(\alpha - \frac{1}{2}, \beta), \tau_2^2 \sim \text{InvGamma}(\alpha, \beta).$$

1041 Then, this hierarchical model introduces the following joint distributions

$$\text{If } \mathcal{G} = \mathcal{G}_1, \text{ then } p(X, w, \tau_1^2, \tau_2^2, \mathcal{G}_1) = p(X|w, \tau_1^2, \tau_2^2, \mathcal{G}_1)p(w|\mathcal{G}_1, \tau_1^2)p(\tau_2^2)p(\mathcal{G}_1)$$

$$\text{If } \mathcal{G} = \mathcal{G}_2, \text{ then } p(X, w, \tau_1^2, \tau_2^2, \mathcal{G}_2) = p(X|w, \tau_1^2, \tau_2^2, \mathcal{G}_2)p(w|\mathcal{G}_2, \tau_2^2)p(\tau_1^2)p(\mathcal{G}_2)$$

$$\text{If } \mathcal{G} = \mathcal{G}_3, \text{ then } p(X, \tau_1^2, \tau_2^2, \mathcal{G}_3) = p(X|\tau_1^2, \tau_2^2, \mathcal{G}_3)p(\tau_1^2)p(\tau_2^2)p(\mathcal{G}_3).$$

1042 For completeness, we show below that the above priors result in the same posterior for graphs in the
 1043 same Markov equivalence class — \mathcal{G}_1 and \mathcal{G}_2 . We begin by recalling a simple result before stating
 1044 the main theorem of this subsection.

Lemma 1. For any $\nu > 0$ and $A, B, C \in \mathbb{R}$ such that $CA^2 > B$ we have

$$\int_{-\infty}^{\infty} \frac{dx}{(A^2x^2 - 2Bx + C)^{\frac{\nu+1}{2}}} = \sqrt{\pi} \frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})} \times \frac{(A^2)^{\frac{\nu-1}{2}}}{(CA^2 - B^2)^{\frac{\nu}{2}}},$$

1045 where $\Gamma(\cdot)$ is the usual Gamma-function.

1046 **PROOF:** Completing the square in the dominator we have

$$\int_{-\infty}^{\infty} \frac{dx}{(A^2x^2 - 2Bx + C)^{\frac{\nu+1}{2}}} = \left(\frac{A^2}{CA^2 - B^2} \right)^{\frac{\nu+1}{2}} \int_{-\infty}^{\infty} \frac{dx}{\left(\frac{A^4}{CA^2 - B^2} \left(x - \frac{B}{A^2} \right)^2 + 1 \right)^{\frac{\nu+1}{2}}}.$$

1047 Next, we recall the probability density function (pdf) of a shifted and scaled version of the standard
 1048 student-t distribution (i.e. $Z = \mu + \sigma T$, with $T \sim t(\nu)$)

$$f_Z(z) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi}} \left(1 + \frac{1}{\nu} \left(\frac{z - \mu}{\sigma^2}\right)^2\right)^{-\frac{\nu+1}{2}}.$$

1049 Then, matching the terms gives the desired result. \square

1050 **Theorem B.2.** Let $\mathcal{D}_{obs} := \{X^{(1)}, X^{(2)}, \dots, X^{(n)}\}$ be i.i.d. observations generated by
 1051 one of the simple models described in eqs. (13) to (15). The posterior over the graphs
 1052 $[p(\mathcal{G}_1|\mathcal{D}_{obs}), p(\mathcal{G}_2|\mathcal{D}_{obs}), p(\mathcal{G}_3|\mathcal{D}_{obs})]$ is

$$\frac{1}{c} \left[\frac{(S_2^\eta)^{(\nu-1)/2}}{(S_2^\beta)^{(\nu-1)/2} [S_1^\beta S_2^\eta - S_{12}^2]^\nu}, \frac{(S_1^\eta)^{(\nu-1)/2}}{(S_1^\beta)^{(\nu-1)/2} [S_2^\beta S_1^\eta - S_{12}^2]^\nu}, \frac{1}{(S_1^\beta)^{(\nu-1)/2} (S_2^\beta)^{(\nu/2)}} \right],$$

1053 where c is the constant of normalisation, $\nu := 2\alpha + n$ and

$$S_i^\eta := S_i + \frac{1}{\eta}, \quad S_i^\beta := S_i + 2\beta \text{ for } i \in \{1, 2\}, \quad (25)$$

1054 with S_1, S_2 and S_{12} defined in eq. (18). The posterior interventional distribution is a mixture of 2
 1055 shifted and scaled Student-t distributions

$$p(X_1 = y | \text{do}(X_2 = x), \mathcal{D}_{obs}) = p(\mathcal{G}_1 | \mathcal{D}_{obs}) t(\nu, \mu_{1-2}(x), \sigma_{1-2}(x)) + (1 - p(\mathcal{G}_1 | \mathcal{D}_{obs})) t\left(\nu, 0, \sqrt{\frac{S_1^\beta}{\nu - 1}}\right)$$

1056 with

$$\mu_{1-2}(x) := \frac{x S_{12}}{S_2^\eta} \quad \text{and} \quad \sigma_{1-2}^2(x) := \frac{S_2^\eta (S_1^\beta S_2^\eta - S_{12}^2) + x^2 (S_1^\beta S_2^\eta - S_{12}^2)}{(S_2^\eta)^2 \nu}. \quad (26)$$

1057 **Proof:** Similar to the case presented in Appendix B.1.1, we start by deriving the posterior over the
 1058 three models described above and use the same definitions from eq. (18).

$$\begin{aligned} p(\mathcal{G}_1 | \mathcal{D}_{obs}) &\propto \int p(\mathcal{D}_{obs} | \mathcal{G}_1, w, \tau_1^2, \tau_2^2) p(w | \tau_1^2) p(\tau_1^2) p(\tau_2^2) d(\tau_1^2) d(\tau_2^2) dw \\ &= \int \frac{1}{(\sqrt{2\pi\tau_1^2})^n} \exp\left(-\frac{1}{2\tau_1^2} \sum_{i=1}^n (X_1^i - w X_2^i)^2\right) \frac{1}{(\sqrt{2\pi\tau_2^2})^n} \exp\left(-\frac{1}{2\tau_2^2} S_2\right) \frac{1}{\sqrt{2\pi\eta\tau_1^2}} \\ &\times \exp\left(-\frac{w^2}{2\eta\tau_1^2}\right) \frac{\beta^{2\alpha-\frac{1}{2}}}{\Gamma(\alpha)\Gamma(\alpha-\frac{1}{2})} (\tau_1^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\tau_1^2}\right) (\tau_2^2)^{-\alpha-1/2} \exp\left(-\frac{\beta}{\tau_2^2}\right) d(\tau_1^2) d(\tau_2^2) dw \\ &= \frac{1}{\sqrt{\eta}} \frac{1}{(2\pi)^{n+1/2}} \frac{\beta^{2\alpha-1/2}}{\Gamma(\alpha)\Gamma(\alpha-\frac{1}{2})} \frac{\Gamma(\alpha+\frac{n-1}{2})}{\left(\beta+\frac{S_2}{2}\right)^{(\nu-1)/2}} \int_{-\infty}^{\infty} \frac{\Gamma(\alpha+\frac{n+1}{2})}{\left(\beta+\frac{\sum_{i=1}^n (X_1^i - w X_2^i)^2}{2} + \frac{w^2}{2\eta}\right)^{\alpha+\frac{n+1}{2}}} dw \\ &= \frac{(2\beta)^{2\alpha-\frac{1}{2}} \Gamma(\alpha+\frac{n-1}{2}) \Gamma(\alpha+\frac{n}{2})}{\sqrt{\eta} \pi^n \Gamma(\alpha) \Gamma(\alpha-\frac{1}{2})} \cdot \frac{(1/\eta + S_2)^{(\nu-1)/2}}{(2\beta + S_2)^{(\nu-1)/2}} \cdot \frac{1}{[(2\beta + S_1)(1/\eta + S_2) - S_{12}^2]^{\alpha+n/2}} \\ &= \frac{(2\beta)^{2\alpha-\frac{1}{2}} \Gamma(\alpha+\frac{n-1}{2}) \Gamma(\alpha+\frac{n}{2})}{\sqrt{\eta} \pi^n \Gamma(\alpha) \Gamma(\alpha-\frac{1}{2})} \cdot \left(\frac{S_2^\eta}{S_2^\beta}\right)^{(\nu-1)/2} \cdot \frac{1}{(S_1^\beta S_2^\eta - S_{12}^2)^\nu}, \end{aligned} \quad (27)$$

1059 where in the last step we used the result of Lemma 1. We note that the conditions in the lemma are
 1060 fulfilled as

$$(2\beta + S_1)(1/\eta + S_2) > S_1 S_2 \geq (S_{12})^2, \quad (28)$$

1061 where we used the fact that $\beta, \eta > 0$ and the Cauchy-Schwartz inequality in the second step.

1062 Similarly, we find the posterior for the second model, \mathcal{G}_2

$$p(\mathcal{G}_2|\mathcal{D}_{\text{obs}}) \propto \frac{(2\beta)^{2\alpha-\frac{1}{2}}\Gamma(\alpha+\frac{n-1}{2})\Gamma(\alpha+\frac{n}{2})}{\sqrt{\eta}\pi^n\Gamma(\alpha)\Gamma(\alpha-\frac{1}{2})} \cdot \left(\frac{S_1^\eta}{S_1^\beta}\right)^{(\nu-1)/2} \cdot \frac{1}{(S_2^\beta S_1^\eta - S_{12}^2)^\nu}$$

1063 and for the third model, \mathcal{G}_3

$$\begin{aligned} p(\mathcal{G}_3|\mathcal{D}_{\text{obs}}) &\propto \int p(\mathcal{D}_{\text{obs}}|\mathcal{G}_3, \tau_1^2, \tau_2^2) p(\tau_1^2) p(\tau_2^2) d(\tau_1^2) d(\tau_2^2) \\ &= \int \frac{1}{\left(\sqrt{2\pi\tau_1^2}\right)^n} \exp\left(-\frac{1}{2\tau_1^2}S_1\right) \frac{1}{\left(\sqrt{2\pi\tau_2^2}\right)^n} \exp\left(-\frac{1}{2\tau_2^2}S_2\right) \\ &\quad \times \frac{\beta^{2\alpha-1/2}}{\Gamma(\alpha)\Gamma(\alpha-\frac{1}{2})} (\tau_1^2)^{-\alpha-1/2} \exp\left(-\frac{\beta}{\tau_1^2}\right) (\tau_2^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\tau_2^2}\right) d(\tau_1^2) d(\tau_2^2) \quad (29) \\ &= \frac{(2\beta)^{2\alpha-1/2}\Gamma(\alpha+\frac{n-1}{2})\Gamma(\alpha+\frac{n}{2})}{\pi^n\Gamma(\alpha)\Gamma(\alpha-\frac{1}{2})} \cdot \frac{1}{(S_1^\beta)^{(\nu-1)/2}(S_2^\beta)^\nu}. \quad (30) \end{aligned}$$

1064 Conditioned on the models, the interventional distributions are

$$\begin{aligned} p(X_1 = y|\text{do}(x), \mathcal{G}_1, w, \tau_1^2) &:= p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_1, w, \tau_1^2) = \mathcal{N}(y|wx, \tau_1^2), \\ p(X_1 = y|\text{do}(x), \mathcal{G}_2, w, \tau_1^2) &= p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_2, w, \tau_1^2) = \mathcal{N}(y|0, \tau_1^2), \\ p(X_1 = y|\text{do}(x), \mathcal{G}_3, \tau_1^2) &:= p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_3, \tau_1^2) = \mathcal{N}(y|0, \tau_1^2). \end{aligned}$$

1065 Then, the posterior interventional distribution, $p(X_1|\text{do}(x), \mathcal{D}_{\text{obs}})$, is

$$\begin{aligned} \sum_{\mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2\}} p(\mathcal{G}|\mathcal{D}_{\text{obs}}) \int p(X_1 = y|\text{do}(x), \mathcal{G}, w, \tau_1^2) p(w, \tau_1^2, \tau_2^2|\mathcal{G}, \mathcal{D}_{\text{obs}}) d(\tau_1^2) d(\tau_2^2) dw \\ + p(\mathcal{G}_3|\mathcal{D}_{\text{obs}}) \int p(X_1 = y|\text{do}(x), \mathcal{G}_3, \tau_1^2) p(\tau_1^2, \tau_2^2|\mathcal{G}_3, \mathcal{D}_{\text{obs}}) d(\tau_1^2) d(\tau_2^2). \end{aligned}$$

1066 First, we find the last term

$$\begin{aligned} \int p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_3, \tau_1^2) p(\tau_1^2, \tau_2^2|\mathcal{G}_3, \mathcal{D}_{\text{obs}}) d(\tau_1^2) d(\tau_2^2) &\propto \\ \int \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left(-\frac{y^2}{2\tau_1^2}\right) p(\mathcal{D}_{\text{obs}}|\mathcal{G}_3, \tau_1^2, \tau_2^2) p(\tau_1^2) p(\tau_2^2) d(\tau_1^2) d(\tau_2^2) &\propto \\ \propto \frac{1}{(2\beta + S_1 + y^2)^{\alpha+\frac{n-1}{2}}} \propto \frac{1}{\left(1 + \frac{y^2}{S_1^\beta}\right)^{\alpha+\frac{n-1}{2}}}, \quad (31) \end{aligned}$$

1067 which is a scaled Student-t distribution with $\nu = 2\alpha + n$ degrees of freedom and $\sigma^2 = \frac{S_1^\beta}{\nu-1}$.

1068 Computing the next integral follows the same pattern presented in eq. (30)

$$\begin{aligned} \int p(X_1 = y|\text{do}(X_2 = x), \mathcal{G}_2, w, \tau_1^2) p(w, \tau_1^2, \tau_2^2|\mathcal{G}_2, \mathcal{D}_{\text{obs}}) d(\tau_1^2) d(\tau_2^2) dw &\propto \\ \int \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left(-\frac{y^2}{2\tau_1^2}\right) p(\mathcal{D}_{\text{obs}}|w, \tau_1^2, \tau_2^2) p(w|\tau_2^2) p(\tau_1^2) p(\tau_2^2) d(\tau_1^2) d(\tau_2^2) dw &\propto \\ \frac{(S_1 + \frac{1}{\eta})^{\alpha+\frac{n-1}{2}}}{(2\beta + S_1 + y^2)^{\alpha+\frac{n-1}{2}}} \cdot \frac{1}{\left[(2\beta + S_2)(S_1 + \frac{1}{\eta}) - (S_{12})^2\right]} \propto \frac{1}{\left(1 + \frac{y^2}{S_1^\beta}\right)^{\alpha+\frac{n-1}{2}}}, \quad (32) \end{aligned}$$

1069 which is again a scaled Student-t distribution with $\sigma^2 = \frac{S_1^\beta}{\nu-1}$ and ν degrees of freedom. Finally,
 1070 using similar steps as in eq. (27), the last term is

$$\begin{aligned} & \int p(X_1 = y | \text{do}(X_2 = x), \mathcal{G}_1, w, \tau_1^2) p(w, \tau_1^2, \tau_2^2 | \mathcal{G}_1, \mathcal{D}_{\text{obs}}) d(\tau_1^2) d(\tau_2^2) dw \propto \\ & \int \frac{1}{\sqrt{2\pi\tau_1^2}} \exp\left(-\frac{(y-wx)^2}{2\tau_1^2}\right) p(\mathcal{D}_{\text{obs}} | w, \tau_1^2, \tau_2^2) p(w | \tau_1^2) p(\tau_1^2) f p \tau_2^2 d(\tau_1^2) d(\tau_2^2) dw \propto \\ & \frac{1}{\left[(S_2 + x^2 + \frac{1}{\eta})(2\beta + S_1 + y^2) - (S_{12} + xy)^2\right]^{\alpha + \frac{n}{2}}}, \end{aligned} \quad (33)$$

1071 where the denominator is always bounded away from 0 as in eq. (28). By completing the square,
 1072 we obtain another scaled and shifted student-t distribution with $\nu = 2\alpha + n$, $\mu_{1-2}(x)$ and $\sigma_{1-2}^2(x)$
 1073 defined in eq. (34). Then, combining eqs. (27) to (33) we obtain the result.

1074 Similarly, we show

$$p(X_2 = y | \text{do}(x), \mathcal{D}_{\text{obs}}) = p(\mathcal{G}_2 | \mathcal{D}_{\text{obs}}) t(\nu, \mu_{2-1}(x), \sigma_{2-1}^2(x)) + (1 - p(\mathcal{G}_2 | \mathcal{D}_{\text{obs}})) t\left(\nu, 0, \sqrt{\frac{S_2^\beta}{\nu-1}}\right)$$

1075 with

$$\mu_{2-1}(x) := \frac{xS_{12}}{S_1^\eta} \quad \text{and} \quad \sigma_{2-1}^2(x) := \frac{S_1^\eta(S_2^\beta S_1^\eta - S_{12}^2) + x^2(S_2^\beta S_1^\eta - S_{12}^2)}{(S_1^\eta)^2 \nu}. \quad (34)$$

1076

□

1077 **Remark:** We employ asymmetric priors for τ_1^2 and τ_2^2 to ensure that the posterior assigns equal
 1078 probability to the models \mathcal{G}_1 and \mathcal{G}_2 , which belong to the same Markov equivalence class. In particular,
 1079 setting $2\beta = \eta$ results in $S_1^\eta = S_1^\beta$ and $S_2^\eta = S_2^\beta$, which implies that the posterior distribution over
 1080 graphs,

$$[p(\mathcal{G}_1 | \mathcal{D}_{\text{obs}}), p(\mathcal{G}_2 | \mathcal{D}_{\text{obs}}), p(\mathcal{G}_3 | \mathcal{D}_{\text{obs}})],$$

1081 takes the form:

$$\frac{1}{c} \left[\frac{1}{(S_1^\beta S_2^\beta - S_{12}^2)^{\nu/2}}, \frac{1}{(S_2^\beta S_1^\beta - S_{12}^2)^{\nu/2}}, \frac{1}{(S_1^\beta)^{(\nu-1)/2} (S_2^\beta)^{\nu/2}} \right], \quad (35)$$

1082 where c is a normalising constant.

1083 This setup corresponds to the prior structure proposed by Geiger and Heckerman [22, Equation 12],
 1084 obtained by setting the precision matrix T in the Wishart distribution to the identity. As noted in their
 1085 Geiger and Heckerman [22, Section 4], a change of variables transforms the Wishart prior on the
 1086 covariance of X into the prior used here for the weights w and error variances τ_1^2 and τ_2^2 .

1087 Assuming a true data-generating mechanism, the posterior concentrates on its Markov equivalence
 1088 class. If the true graph is \mathcal{G}_1 or \mathcal{G}_2 , then $p(\mathcal{G}_1 | \mathcal{D}_{\text{obs}}) = p(\mathcal{G}_2 | \mathcal{D}_{\text{obs}}) \rightarrow 1/2$. Conversely, if \mathcal{G}_3 is the
 1089 true graph, then $p(\mathcal{G}_3 | \mathcal{D}_{\text{obs}}) \rightarrow 1$ [12, 65, 14].

1090 The degrees of freedom, $\nu = \alpha + n$, grow with the sample size n , so the corresponding Student- t
 1091 distributions converge to Gaussians in the large-sample regime.

1092 B.2 Three-node Experiments

1093 In the three-node experiments (section 5.2) we use two datasets with two different functional mecha-
 1094 nisms $f_i(\cdot)$ as defined in eq. (1): one sampled from a GP prior, and one based on neural networks.
 1095 In both cases, we sample Erdős-Rényi graphs with graph degree chosen uniformly from $\{1, 2, 3\}$.
 1096 Following Ormaniec et al. [49], we standardise all variables upon generation.

1097 **GP functional mechanism** To model $f_i(\cdot)$ we use a GP with a squared exponential kernel, with a
 1098 randomly sampled lengthscale for each parent set PA_i of size $|\text{PA}_i|$. More specifically, we sample
 1099 the lengthscale from a log- distribution $\{\lambda_p\}_{p=1}^{|\text{PA}_i|} \sim \text{Log}(-1, 1)$, followed by clipping between
 1100 $\lambda_p = \text{clip}(\lambda_p, 0.1, 5)$ to ensure that a too long lengthscale does not result in independence of the
 1101 variable from a parent. This defines the kernel matrix between the n -th and m -th samples as:

$$\mathbf{K}_{nm} = \exp(-(\text{PA}_i^n - \text{PA}_i^m)^T \mathbf{\Lambda}^{-1} (\text{PA}_i^n - \text{PA}_i^m)),$$

1102 with $\mathbf{\Lambda} := \text{Diag}(\lambda_1, \dots, \lambda_{|\text{PA}_i|})$. We then add noise with variance $\sigma^2 \sim \text{Gamma}(1, 5)$ and sample
 1103 the variables as follows

$$X_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I})$$

1104 **Neural network-based functional mechanism** We sample each variable as follows

$$\begin{aligned} \eta_i &\sim \mathcal{N}(0, 1) \\ X_i^n &\sim \text{ResNet}_\theta([\text{PA}_i^n, \eta_i]) + \sigma\epsilon, \end{aligned}$$

1105 where $\sigma^2 \sim \text{Gamma}(1, 10)$, $\epsilon \sim \mathcal{N}(0, 1)$. ResNet_θ is a residual neural network with a randomly
 1106 sampled number of blocks $N_{\text{blocks}} \sim \mathcal{U}\{1, \dots, 8\}$ and randomly sampled hidden dimension $d_{\text{hidden}} \sim$
 1107 $\mathcal{U}\{2^5, 2^6, 2^7, 2^8\}$. We use the GELU [27] activation function.

1108 B.3 Higher-dimensional experiments

1109 For the higher dimension experiments in section 5.3 we generate the training data for MACE-TNP
 1110 as follows:

- 1111 • We sample number of variables $D \sim \mathcal{U}[5, 40]$.
- 1112 • We sample a type of graph, either an Erdős–Rényi graph or a scale-free graph [3].
- 1113 • The density of the graph (number of edges) is sampled from $\mathcal{U}[\frac{D}{2}, 6D]$.
- 1114 • For each node, we sample a functional mechanism randomly from either a GP with an
 1115 additional latent variable input, or a Neural network with an additional latent variable input:
 - 1116 – GP with latent: We sample a latent $\eta_i \sim \mathcal{N}(0, 1)$, and lengthscales $\{\lambda_p\}_{p=1}^{|\text{PA}_i|+1} \sim$
 1117 $\text{Log}(-0.5, 1)$, where PA_i denotes the set of parents of node index i . Functions are
 1118 sampled from a squared exponential kernel with η_i included as an input and Gaussian
 1119 noise added with variance $\sigma^2 \sim \text{Gamma}(1, 5)$.
 - 1120 – NN with latent:

$$\begin{aligned} \eta_i &\sim \mathcal{N}(0, 1) \\ X_i^n &\sim \text{NN}_\theta([\text{PA}_i^n, \eta_i]) + \sigma\epsilon, \end{aligned}$$

1121 where $\sigma^2 \sim \text{Gamma}(1, 10)$, $\epsilon \sim \mathcal{N}(0, 1)$. NN_θ denotes a randomly initialised neural
 1122 network with 128 hidden dimensions and one hidden layer.

1123 Using a latent as an input ensures that the final distribution is not Gaussian. Following Ormaniec
 1124 et al. [49], we standardise all variables during the data generation process.

1125 For testing for each variable size in table 2, we only generate Erdős–Rényi graphs with density $4D$.
 1126 This is to test the performance of the baselines and our method in the difficult dense graph case. The
 1127 rest of the data generation process is the same as the training data.

1128 C Experimental Details

1129 This section provides additional details and results for the experiments presented in Section 5

1130 C.1 Architecture, training details and hardware

1131 Throughout our experiments we use $H = 8$ attention heads, each of dimension $D_Q = D_{KV} =$
1132 $d_{\text{model}}/8$. The MLPs used in the encoding use two layers and a hidden dimension of $d_{\text{embed}} = d_{\text{model}}$.
1133 Unless otherwise specified, we use a learning rate of 5×10^{-4} with a linear warmup of 2% of the
1134 total iterations, and a batch size of 32.

1135 To train MACE-TNP, we randomise the number of observational samples $N_{\text{obs}} \sim \mathcal{U}\{50, 750\}$, and
1136 set $N_{\text{int}} = 1000 - N_{\text{obs}}$. The training loss is evaluated on these N_{int} samples. For testing, we sample
1137 500 observation points and compute the loss against 500 intervention points.

1138 **Two-node linear Gaussian model** We use $L = 2$ transformer encoder layers, where each trans-
1139 former encoder layer involves the attention over samples, followed by attention over nodes. The
1140 model dimension is $d_{\text{model}} = 128$, and feedforward width $d_{\text{ff}} = 128$. We train the model for 1 epoch
1141 on 50.000 datasets and test on 100 datasets. Training takes roughly 60 minutes on a single NVIDIA
1142 GeForce RTX 2080 Ti GPU 11GB, and testing is performed in less than 5 seconds.

1143 **Three-node experiments** For the experiment in the main paper, we use $L = 2$ transformer encoder
1144 layers, a model dimension $d_{\text{model}} = 128$, and feedforward width $d_{\text{ff}} = 128$. We train the model for 2
1145 epochs on 50.000 datasets for the GP experiment and 100.000 datasets for the NN one, and test on
1146 100 datasets in both cases. When testing the OOD performance, we train on the union of the two
1147 datasets for 2 epochs. Training the models described in the main text required roughly 4 – 6 hours of
1148 GPU time; however, because we ran them on a shared cluster, actual runtimes may vary with cluster
1149 utilization.

1150 **Higher dimensional and Sachs experiments** For the higher dimensional experiments we use
1151 $L = 4$ encoder layers. The model dimension is $d_{\text{model}} = 256$ with feedforward dimension $d_{\text{ff}} = 1024$.
1152 We train the model on data generated as listed in appendix B.3 with 2, 500, 000 datasets in total. The
1153 model was trained on an NVIDIA A100 80GB GPU for 2 epochs which took roughly 20 hours. We
1154 use the model trained for the higher dimensional experiment for the Sachs experiment.

1155 **Hardware** For the two- and three-node experiments, we ran both training and inference on a single
1156 NVIDIA GeForce RTX 2080 Ti (11 GB) with 20 CPU cores on a shared cluster. The only exception
1157 was for our largest three-node GP and NN models (with $d_{\text{model}} = 1024$), where we used a single
1158 NVIDIA RTX 6000 Ada Generation (50 GB) paired with 56 CPU cores; those models required
1159 roughly 25 GB of GPU memory. For the higher-node experiments, we used a single NVIDIA A100
1160 80GB GPU, as well as an RTX 4090 24GB GPU.

1161 C.2 Additional Results

1162 C.2.1 Two-node Linear Gaussian Model

1163 We study the performance of MACE-TNP in both identifiable and non-identifiable causal settings by
1164 generating data according to the models described in appendix B.1.1 and appendix B.1.2, respectively.
1165 For all experiments, we set $\sigma = \sigma_w = 1$ in the identifiable case and $\alpha = 3, \eta = 2\beta = 1$ in the
1166 non-identifiable case. We investigate at different interventional queries, x , how the NP predicted
1167 distributions compare with the analytical ones as a function of the observational sample size. For
1168 simplicity, for the first model B.1.1 we consider an NP which outputs a mixture of 2 components,
1169 while for the second model B.1.2 the NP approximates the true analytical distribution using 3
1170 components.

1171 The flexibility of our architecture also allows for conditional queries, multiple interventions, as well
1172 as easily incorporating interventional data to help identify causal relations. Hence, we investigate here
1173 whether providing a small number $M_{\text{int}} = 5$ of true interventional samples, alongside the observational
1174 data, resolves identifiability challenges in the non-identifiable case. As already shown in fig. 3 (right)
1175 with the green line and discussed in section 5.1 we find that adding extra interventional data

1176 does indeed lower the $\text{KL}(p_{BCM}(x_i|\text{do}(x_j), \mathbf{f}^*, \mathcal{G}^*) \| p_{\theta}(x_i|\text{do}(x_j), \mathcal{D}_{\text{obs}}, \{x_i^n\}_{n=1}^{M_{\text{int}}}))$, suggesting
 1177 that even limited interventions can enhance identifiability. We also test this with an increasing number
 1178 of interventional samples in fig. 5. As soon as the interventional information is rich enough ($M_{\text{int}} \in$
 1179 $\{50, 300\}$), the NP recovers the interventional distribution of the true data-generating mechanism
 1180 even with little to no observational data, as indicated by the near-flat KL curves. We note that the KL
 1181 divergence between two Gaussian mixtures (or between a Student-t mixture and a Gaussian mixture)
 1182 lacks a closed-form expression. Therefore, we approximate it in our experiments by averaging 1000
 1183 Monte Carlo estimates of the log-density ratio.

1184 Then, we show in fig. 6 two examples where the intervention is made at $x = 1$ for the identifiable
 1185 model and at $x = 2$ for the non-identifiable model. A clear distinction is observed between the two
 1186 settings: for the identifiable case, the analytical posterior interventional distribution is a mixture of
 1187 two Gaussian distributions, which, at high observational sample sizes, converges to a single Gaussian
 1188 (i.e. because the observational data gives information regarding the causal structure, the weight
 1189 corresponding to one mode collapses to 0). In contrast, for the non-identifiable case, the posterior
 1190 places equal mass on both \mathcal{G}_1 and \mathcal{G}_2 , and therefore, the mixture structure persists across both regimes.
 1191 In both settings, the NP-predicted distributions closely match the correct interventional distributions,
 1192 with accuracy improving as the number of observational samples increases. This improvement is due
 1193 to two factors. First, larger sample sizes provide the NP with more information about the underlying
 1194 causal model, allowing for enhanced inference. Second, in the non-identifiable case, the posterior
 1195 interventional distribution is a mixture of two Student-t distributions with a number of degrees of
 1196 freedom proportional to the number of observational samples. Thus, in the high sample regime, the
 1197 mixture distribution converges to a mixture of Gaussians, which is the class that parameterises the
 1198 output of the NP model.

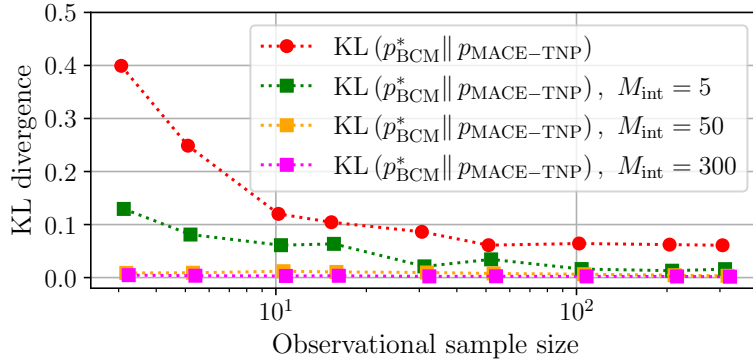


Figure 5: Average KL divergence between the interventional distribution of the true generating mechanism, $\{\mathcal{G}^*, \mathbf{f}^*\}$, and the NP-predicted distribution shown as a function of the observational sample size for the non-identifiable setting. Results are shown for various interventional sample sizes. For simplicity, we only report the medians.

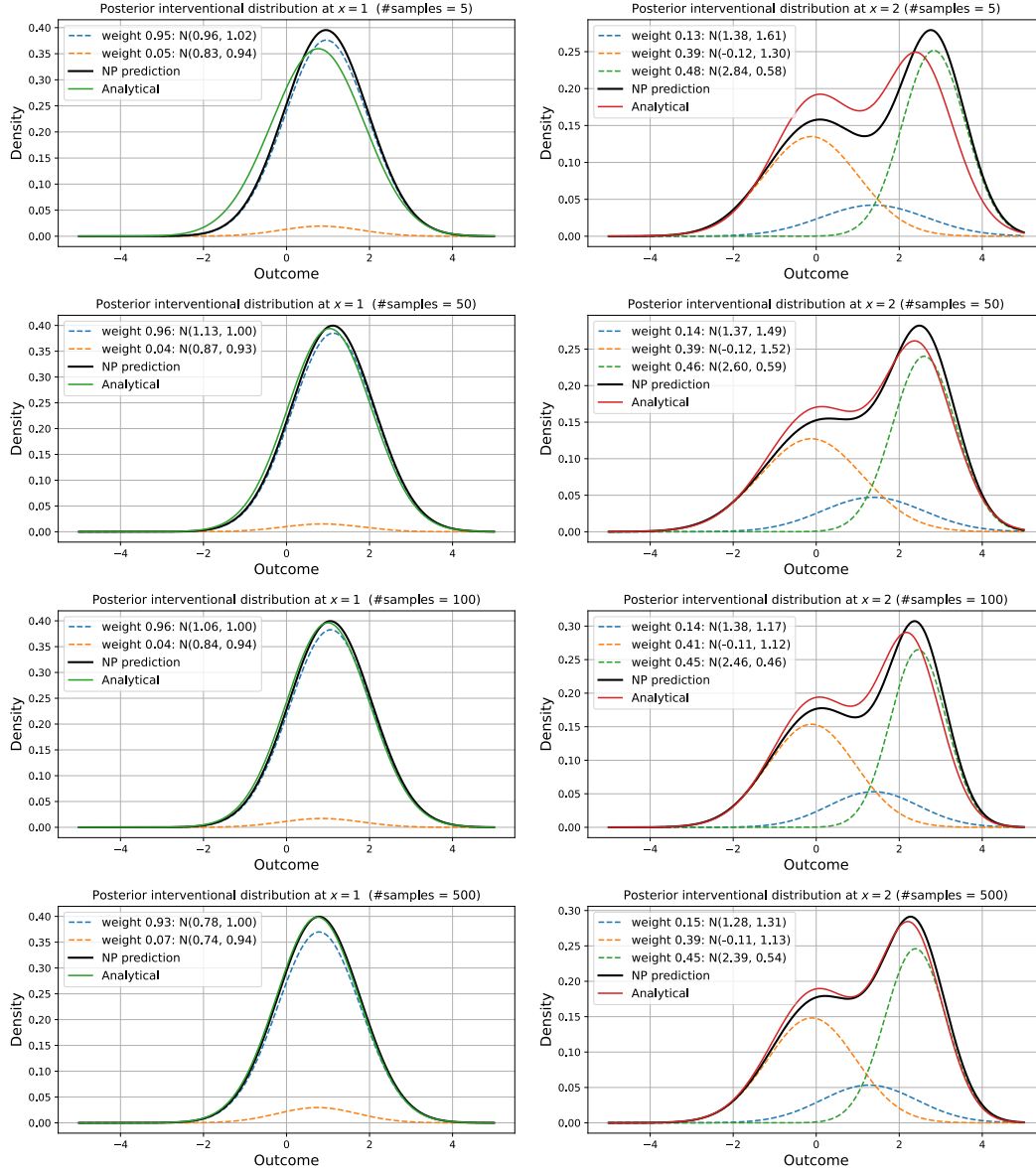


Figure 6: Fitted NP posterior interventional distributions vs. true posterior interventional distributions for identifiable (left) and non-identifiable models (right) at increasing observational sample sizes (5, 50, 100, 500).

1199 C.2.2 Three-node Experiments

1200 In this section we provide additional results on the three-node experiments where we aim to address
 1201 three questions: 1) between the MHSA and MHCA schemes for sample attention introduced in
 1202 appendix [A.2](#) which one performs better? 2) Does increasing the number of MoG components
 1203 improve performance, and 3) How does the model performance vary with the size of the architecture?

1204 Table [3](#) shows the results for the two functional mechanisms used in the three-node experiments: GP
 1205 and NN-based. For each model configuration, we present four sets of results: for a model trained
 1206 on GP and tested on GP (GP / GP), a model trained on NN and tested on NN (NN / NN), and for a
 1207 model trained on the combination between the two datasets and tested on each of them (GP+NN /
 1208 GP and GP+NN / NN). These results allow us to assess whether the influence of model architecture
 1209 is consistent across the functional mechanisms. Notably, models trained on the combined GP+NN

dataset are able to match—within error—the performance of models trained specifically on either GP or NN data. This highlights the strength of the meta-learning approach: even when trained on data generated from diverse functional mechanisms, a single model can generalise effectively across both, achieving performance comparable to specialised models while also benefiting from broader prior coverage. We summarise the findings from table 3:

1. MHSA + MHCA outperforms the masked-MHSA strategy for attention over samples.
2. Increasing the number of MoG components increases the performance of MACE-TNP. There is a larger gap in performance when going from 1 to 3 mixture components, indicating the importance of allowing the model to output non-Gaussian marginal predictions. Increasing the number of components to 10 further improves performance, but the gains are not as significant.
3. Scaling up the model architecture generally leads to decreased NLPID.
4. Training a model on the combination of the two datasets (GP+NN) is able to recover—within error—the performance on both datasets.

Table 3: Results of MACE-TNP under different architectural configurations. M-SA stands for Masked-MHSA, while SA+CA indicates the MHSA+MHCA attention mechanism. For each model, the column name under NLPID indicates the training set / test set (i.e. GP+NN / GP indicates we trained the model on the GP+NN dataset and tested it on the GP one). We report the mean \pm the error of the mean of the NLPID over 100 datasets.

MoG	Attention	d_{model}	d_{ff}	L	NLPID (\downarrow)			
					GP / GP	NN / NN	GP+NN / GP	GP+NN / NN
1	M-SA	128	128	4	629.0 ± 20.0	664.1 ± 16.0	640.1 ± 17.3	668.3 ± 17.2
1	SA+CA	128	128	4	617.4 ± 20.1	664.9 ± 16.4	629.8 ± 17.5	688.0 ± 31.5
3	M-SA	128	128	4	581.8 ± 21.8	538.9 ± 19.1	597.6 ± 19.9	547.1 ± 17.8
3	SA+CA	128	128	4	569.3 ± 23.1	540.5 ± 17.1	582.1 ± 21.6	540.6 ± 18.8
10	M-SA	128	128	4	572.1 ± 21.9	533.2 ± 18.3	599.4 ± 20.3	531.5 ± 19.6
10	SA+CA	128	128	4	563.9 ± 23.4	527.9 ± 19.8	583.9 ± 21.5	531.0 ± 19.4
10	SA+CA	512	256	8	555.7 ± 24.6	527.0 ± 19.1	564.6 ± 23.6	532.1 ± 18.4
10	SA+CA	1024	256	8	558.0 ± 23.9	518.2 ± 19.7	565.6 ± 22.3	521.1 ± 20.8

Finally, table 4 summarises the results for the out-of-distribution (OOD) evaluation for the configuration presented in the main text.

Table 4: Results for the OOD two-node experiment. We show the NLPID (\downarrow) and report the mean \pm the error of the mean over 100 datasets. Each row corresponds to a different functional mechanism used in the test set (GP / NN).

Test \downarrow	Training \rightarrow		
	GP	NN	GP+NN
GP	563.9 ± 23.4	678.0 ± 10.0	583.9 ± 21.5
NN	608.3 ± 17.3	527.9 ± 19.8	531.0 ± 19.4

C.2.3 Sachs Full Results

The full set of results for the Sachs dataset are shown in table 5. The MACE-TNP performs competitively with DECI, and both outperform other methods that use GPs as functional models.

Table 5: Results for the Sachs dataset [55]. We show the NLPID (\downarrow) and report the mean \pm the error of the mean across 5 interventions and across all 10 nodes used as the outcome for each intervention. Each row corresponds to a different baseline.

	Sachs
MACE-TNP	989.8 \pm 100.2
DiBS-GP	1417.5 \pm 186.7
ARCO-GP	1400.7 \pm 208.7
DECI	1000.9 \pm 133.5
NOGAM+GP	1763.7 \pm 297.4