Learning towards Robustness in Causally-Invariant Predictors

Xiangyu Zheng∗2, Mingzhou Liu∗1, Xinwei Sun†3, Fang Fang4, Yizhou Wang1,5
1 Computer Science Department, Peking University, 100871, Beijing, China
2 Department of Statistics, Guanghua School of Management, Peking University, 100871, Beijing, China
3 School of Data Science, Fudan University, 200433, Shanghai, China
4 School of Psychological and Cognitive sciences, Peking University, 100871, Beijing, China
5 DeepWise AI Lab, 100080, Beijing, China

Abstract

We propose to learn an invariant causal predictor that is robust to distributional shifts, in the supervised regression scenario. Based on a disentangled causal factorization that describes the underlying data generating process, we attribute the distributional shifts to mutation of generating factors, which covers a wide range of cases of distributional shifts as we do not make prior specifications on the causal structure or the source of mutation. Under this causal framework, we identify a set of invariant predictors based on the do-operator. We provide a sufficient and necessary condition for a predictor to be min-max optimal, i.e., minimizes the worst-case quadratic loss among all domains. This condition is justifiable under the Markovian and faithfulness assumptions, thus inspiring a practical algorithm to identify the optimal predictor. For empirical estimation, we propose a permutation-regeneration scheme guided by a local causal discovery procedure. The utility and effectiveness of our method are demonstrated in simulation data and two real-world applications: Alzheimer’s disease diagnosis and gene function prediction.

∗Equal Contribution
†Corresponding Author
1 Introduction

Standard supervised learning methods heavily rely on the i.i.d assumption. Performance of these models can degrade a lot when this assumption does not hold, which usually happens in real-world scenarios. This can cause serious issues, especially in safe-critic tasks such as healthcare [1][2]. To resolve this problem, it is desired for a predictor to be stably transferable and robust to samples from unseen distributions. As such robustness is often measured by the worst-case performance [3][4][5][6], our goal is to find a stable predictor with minimal worst-case error (a.k.a. min-max optimum).

Recently, there is increasing attention on leveraging causality into learning for robustness, since the causation is believed to be more stable than correlation. Inspired by this, existing works assumed the existence of causal features (or invariant representations) that induced invariant predictors for transfer. For example, the ICP [7][8] assumed the response \( Y \) is invariantly distributed conditioning on its causal features. The IRM and its variants [9][10][11] assumed such causal features as a data representation that induces the invariant predictor. Similarly, the [12] attributed the distributional shifts to exogenous variables and modified the least square loss towards robustness. Further, the [6][13] could utilize the invariance beyond causal features by specifying the variable set with mutable causal mechanisms. They then proposed the “graph surgery estimator” to remove the dependency on these mutable variables, which was provably to be min-max optimal under some conditions. Although the proposed invariant form was not restricted to conditional distribution, they still relied on the pre-specification of the mutable variables set; besides, their analysis is not complete, i.e., the estimator is not min-max optimal under some cases.

To resolve these problems, we first identify the invariant predictors by decomposing the joint distribution following Causal Markov Condition [14]. Accordingly, the joint distribution is decomposed into several autonomous generating factors (i.e., distribution of each variable given its cause), as changing one of them will not affect others [14]. Due to this autonomy, it is necessary to have varied factors to account for the distributional shifts. We call the set of these variables the mutable set that is not assumed to be known, and that of others the stable set. Then we could obtain a set of invariant predictors, which are conditioned on the intervened mutable variables and any stable subset. Among these predictors, the [6][13] proved that the one with the whole stable set is min-max optimal if the intervened distribution can degenerate to the conditional one. Although this conclusion seems natural since this predictor can utilize all information in the stable set, it is not always true when the above degeneration condition does not hold. Such a counter-intuitive finding, together with the identification of invariant predictors, is shown in the following example for illustration.

**Example 1** (Alzheimer’s Disease (AD)). In AD, the patients suffer from memory loss, since the disease \( Y \) can cause atrophy in memory-related brain regions [15] such as the hippocampus \( X_8 \) and the medial temporal lobe \( X_2 \) by releasing toxic proteins [16]. Along anatomical pathways, these proteins further spread from one brain region to another (e.g. from the hippocampus to the medial temporal lobe) [17][18]. The degree of atrophy varies across ages, resulting in AD’s heterogeneity, as supported by existing findings [19][20] and reflected by our learned causal graph in Fig. 1(a) over all brain regions \( \{X_i\}_{i=1}^9 \) where the hippocampus \( X_8 \) is detected to be mutable across ages. We then intervene on these mutable brain region volumes (as marked in blue) to obtain the set of invariant predictors. We empirically find in Fig. 1(b) that instead of the whole stable set (marked by \( \triangle \)), only keeping \( \{X_3, X_5\} \) in the stable set achieves the minimum worst-case mean square error (MSE). Besides, it is interesting to note that the inclusion of \( X_2 \) can cause a sudden jump in terms of MSE, since the \( v \)-structure of \( X_8 \rightarrow X_2 \leftarrow Y \) can prevent the intervention distribution from degenerating to the conditional distribution.

To resolve this problem and learn the min-max predictor in any case, we first translate the “degeneration condition” in [13][6] into a graphical criterion, which is justifiable with the assistance of domain index variable [21]. When this condition is determined to not hold, we transform the worst-case quadratic loss into an optimization problem over all subsets of the stable set, the minimizer of which is sufficient and necessary for the predictor to be min-max optimal. For estimation, we mimic the \( do \)-operator to propose a permutation-regeneration scheme, to generate data from the intervened distribution. Theoretical correctness and the utility of our method are demonstrated in simulation data and two real-world applications, Alzheimer’s disease diagnosis and gene function prediction.
According to the Causal Markov Condition [14], we have the causal factorization of the joint distribution without any adjustment. We denote the invariant predictor set as \((\text{Causal Model})\) Assumption 2.1 \(\text{Markovian and Faithfulness}\)

Our goal is to find an optimal invariant predictor \(f^* \in \mathcal{F}^i\) using data from \(\mathcal{E}_{T^i}\) such that \(f^*\) is robust to distributional shifts across all domains. A natural way to measure this robustness is to investigate the predictor’s worst-case performance, which provides a safeguard for prediction in unseen domains. That is, we want \(f^*\) to have the following min-max optimal property:

\[
f^*(x) = \arg\min_{f \in \mathcal{F}^i} \max_{P \in \mathcal{P}} \mathbb{E}_{P_\mathcal{E}}[(Y, f(x))^2].
\]

**Notations.** In a causal directed acyclic graph (DAG) \(G\), we denote the parents, children, descendants, and Markovian blanket of the node set \(V\) as \(\text{PA}(V), \text{Ch}(V), \text{Do}(V), \text{Blanket}(V)\), respectively. We denote d-separation and independence by \(\perp_G\) and \(\perp\), respectively.

Next, we introduce some basic assumptions: causal model, Markovian and faithfulness.

**Assumption 2.1 (Causal Model).** We assume the joint distribution \(P^\rho(X, Y)\) is entailed by a DAG \(G := (V, E)\) for all \(\rho \in \mathcal{E}\), where \(V := X \cup Y\) denotes the node set and \(E\) denotes the edge set. Each variable \(V_i \in V\) is generated from a structural equation \(g_i: V_i = g_i(\text{PA}(V_i), U_i)\), where \(U_i\) denotes the exogenous variable. Each edge in \(E\) represents a direct causal relationship [22].

The causal DAG model is commonly assumed in the literature of causal inference [14][23]. Equipped with the causal model especially structural equations, we give definitions of the intervention and do-operator. The intervention \(\text{do}(V_i = v_i)\) means lifting \(V_i\) from its original generating mechanism \(V_i = g_i(\text{PA}(V_i), U_i)\) and set it to a constant value \(v_i\) [14]. Under the following Markovian assumption, the intervention distribution \(p(v|\text{do}(v_i)) := p(v)/\prod_{i \in v_i} p(v_i|\text{pa}(v_i))\).

**Assumption 2.2 (Markovian and Faithfulness).** The Markovian means \(\{U_i\}\) are mutually independent. The faithfulness means \(\forall\) disjoint sets \(V_i, V_j, V_k: V_i \perp_{V_j|V_k} \Rightarrow V_i \perp_{\perp_{G}} V_j|V_k\).

According to the Causal Markov Condition [14], we have the causal factorization of the joint probability \(p(v) = \prod_i p(v_i|\text{pa}(v_i))\), where each term \(p(v_i|\text{pa}(v_i))\) represents the generating of \(V_i\).
given its parents $PA(V_i)$. This factorization allows us to find the reason behind distributional shifts as the variation of some generating factors. Besides, the Markovian and faithfulness together imply $V_i \perp V_j | V_k \iff V_i \perp_G V_j | V_k$.

3 Methodology

In this section, we introduce our method to learn the min-max optimal predictor $f^*$ in Eq. (1). Firstly, in Sec. 3.1 we introduce a causal framework to formulate the distributional shifts across domains and identify a set of invariant predictors. Then, in Sec. 3.2 we provide an algorithm to identify the $f^*$ among the set of invariant predictors, which is inspired by a comprehensive analysis of the min-max property. Finally, in Sec. 3.3 we propose an efficient method to implement the algorithm and estimate the identified $f^*$.

3.1 Causal Framework and Invariant Predictors

As is known, extrapolation to unseen domains is generally unachievable when the distribution can change arbitrarily. Certain restrictions must be used to build connections between the seen and unseen domains. To make this connection, we view the joint distribution $P^e(X, Y)$ from a causal perspective and factorize it into products of autonomous causal factors. Specifically, according to assumptions 2.1 and 2.2 the $P^e(X, Y)$ is entailed by a DAG and admits the following causal factorization:

$$p^e(x, y) = p^e(y|pa(y)) \prod_{i=1}^{P} p^e(x_i|pa(x_i)), \text{ for } e \in \mathcal{E},$$

(2)

where each factor describes the causal mechanism that generates the variable given its cause. A nice property of causal factorization over other types of factorization is its autonomous configuration, that is, each cause-effect relationship represents an autonomous physical mechanism that is independent of the others. As a result, a change in $P^e(X, Y)$ across domains can be essentially attributed to the variation of at least one generating factor in (2). This naturally leads to the following framework that describes the distributional shifts.

Causal Framework for Distributional Shifts. We consider $\mathcal{E}$ contains any $e$ such that $P^e$ admits the following factorization:

$$p^e(x, y) = p(y|pa(y)) \prod_{i \in S} p(x_i|pa(x_i)) \prod_{i \in M} p^e(x_i|pa(x_i)),$$

(3)

where $S$ and $M$ respectively denote the stable and mutable sets such that $X_S := \{X_i | \forall e \in \mathcal{E}, p^e(x_i|pa(x_i)) = p(x_i|pa(x_i))\}$, $X_M := \{X_i | \exists e_1, e_2 \in \mathcal{E}, p^e_1(x_i|pa(x_i)) \neq p^e_2(x_i|pa(x_i))\}$.

Different from [6], we do not pre-specify the mutable variables set $X_M$.

This framework is closely related to the concept of “soft intervention” [24] and “mechanism change” [25] in the literature of causal discovery, which refer to the change of generating factors under certain known experimental interventions. In this regard, it is equivalent to define $P = \{P|P$ admits factorization in Eq. (3)$\}$, when $p^e$ is allowed to vary arbitrarily across $\mathcal{E}$.

Equipped with such a framework, we can identify a set of invariant predictors via the $do$-operator. To ensure the existence of such predictors, we assume the $p(y|pa(y))$ is invariant across domains, which has been widely adopted in the existing literature [9, 7, 4, 26]. Specifically, we consider a set of predictors by implementing the $do$-operator on $X_M$, i.e., $\{f_{S} := \mathbb{E}_{P^e}[Y|x_{S\perp}, do(x_M)]|S \subset S\}$. This $do$-operation removes the effect of mutable generation of $X_M$ given their parents. Therefore, according to the definition of intervention distribution, we have the following invariance property:

Proposition 3.1 (Invariance Property). For each $S\subset S$, we have

$$p^e(y, x_{S\perp}|do(x_M)) = p(y|pa(y)) \prod_{i \in S\perp} p(x_i|pa(x_i)),$$

which is invariant across $e \in \mathcal{E}$. Therefore, $f_{S\perp} := \mathbb{E}_{P^e}[Y|x_{S\perp}, do(x_M)]$ is invariant across $\mathcal{P}$.

Remark 3.2. Due to its invariance, we omit the superscript $e$ in $p^e(y, x_{S\perp}|do(x_M))$ and denote $f_{S\perp} := \mathbb{E}_{P^e}[Y|x_{S\perp}, do(x_M)]$ hereafter.

This invariance property is of great importance, as it suggests that the predictor $f_{S\perp}$ learned from the training domains can be stably transferred to unseen test domains. Then, among such a set of
invariant predictors, one is tempted to ask: which predictor is the most robust one? Formally speaking, consider $\mathcal{F}^I = \{f_{S_-}|S_- \subset S\}$ in the regression scenario, our goal is to learn the predictor $f_{S_-}$ such that $f_{S_-} = f^* := \arg\min_{f_{S_-} \in \mathcal{F}^I} \max_{P \in \mathcal{P}} \mathbb{E}_{P}[(Y - f_{S_-}(x))^2]$. At a first glance, it seems that $f_{S}$ with the whole set $S$ is the optimal $f^*$ since it uses information of the full stable set $X_S$, as also claimed in \cite{13} under some conditions. However, as we show later, this conclusion is not always true.

### 3.2 Identification with Min-max Property

In this section, we analyze the theoretical properties for identifying the min-max predictor $f^*$, which is equivalent to finding the optimal subset in $S$. Our analysis is roughly composed of two main results: Thm. 3.3 and Thm. 3.5. Specifically, we in Thm. 3.3 provide a justifiable graphical criterion that is sufficient for the whole set $S$ to be optimal, which degenerates to the result in \cite{13}. When this condition fails, we in Thm. 3.5 provide a sufficient and necessary condition for a subset to induce the min-max optimal predictor. Particularly, we will introduce a counter-example to show that such an optimal subset does not have to be the whole set.

We first introduce Thm. 3.3 that provides a sufficient condition for $f_{S}$ to be optimal. Indeed, this result has been established in \cite{13} when $p(y|x_S, do(x_M))$ can degenerate to conditional distribution. However, this condition is too general and can not be justified. Equipped with the structural causal model, we provide a graphical criterion that is equivalent to the above condition. With such a graphical language, this condition is justifiable via causal discovery.

**Theorem 3.3 (Graphical Criterion for $f^* = f_{S}$).** Suppose assumptions 2.1, 2.2 hold. Denote $X_M^d := X_M \cap \text{Ch}(Y)$ as mutable variables in $Y$’s children, and $K := \text{Do}(X_M^d \setminus X_M^d)$ as descendants of $X_M^d$. Then, $p(y|x_S, do(x_M))$ can degenerate to conditional distribution if and only if $Y$ does not point to any member of $K$. Further, under either of the two conditions, we have $f^* = f_{S}$.

**Proof Sketch:** If $Y$ does not point to any member of $K$, we can use the rules \cite{14} of do-calcus to remove the “do” in $p(y|x_S, do(x_M))$ and degenerate it into $p(y|x_M \cup \text{do}(x_M))$. For example, in Fig. 2(a), we have $p(y|x_{s_1}, x_{s_2}, do(x_m)) = p(y|x_{s_2}, do(x_m)) = p(y|x_{s_2})$. Such a degeneration (to conditional distribution) permits us to construct a probability distribution in $\mathcal{P}$ for any predictor $f$, that has larger or equal quadratic loss than $f_{S}$, thus proving the optimality of $f_{S}$. On the contrary, if $Y$ points to $K$, then there exists a collider in $X \setminus X_M^d$ in some paths between $Y$ and $X_M^d$, making it incapable to remove the “do”.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Illustration of the graphical criterion in Thm. 3.3 which respectively holds and not holds in the graph (a) and (b).}
\end{figure}

Compared with \cite{13,0}, the Thm. 3.3 translates the unjustifiable intervention degeneration condition into a graphical notion, which can be justified by the data.

**Proposition 3.4.** Under assumptions 2.1, 2.2 the $K$ is identifiable; besides, we can determine whether $Y \rightarrow K$ from the joint distribution over training domains.

Thm. 3.3 reminds us the sufficient condition for $f^* = f_{S}$ may not always hold. When the condition fails, the invariant predictor with the whole stable set is not necessarily min-max optimal. Indeed, we provide a counter-example in Fig. 2(b) (taken from the causal graph of Alzheimer’s disease) that, the $f_{I}(X_S)$ has a larger error than $f_{S}$ under some conditions on $P(Y)$ and $P(X_S|Y)$ \cite{13}. In the following, we show that when the condition in Thm. 3.3 fails, we can still identify the min-max predictor by considering a distribution family $\{P_J\}_{J}$, where $J$ is a definite function from $PA(X_M)$ to $X_M$ and

---

\footnote{Please refer to the supplementary for details.}
We introduce methods to implement Alg. 1. Specifically, it involves whether Alg. 1 is composed of two steps, respectively inspired by Thm. 3.3 and Thm. 3.5. Specifically, if the condition in Thm. 3.3 holds, then we directly have \( f^* = f_S \) with the whole stable set \( S \). Otherwise, we have to search over \( \{ f_S \mid S \in \text{POWER}(S) \} \) to find \( f^* \). It’s worthful to mention that such a searching cost, as also required in existing works [13, 6], can be reduced since some subsets have the same \( h^* \) (please see the supplementary for details). Practically, we can parameterize \( J_0 \) as the neural network, due to its ability to approximate any definite function [27].

Algorithm 1 Identification of the min-max optimal predictor \( f^* \)

**INPUT:** The training data \( \{D_n\mid e \in E_T\} \)

**OUTPUT:** Min-max predictor \( f^* \)

1: Identify \( K \) and the skeleton of DAG.
2: if \( Y \) does not point to any member of \( K \) then
3: \( S^* = S \) according to Thm. 3.3
else
4: set \( h_{\text{min}} = \infty, S^* = \emptyset \).
5: for \( S_- \in \text{POWER}(S) \)
6: estimate \( f_{S_-} \) and \( h^*(S_-) := \max_{J_0} h(S_-, J_0) \).
7: \( h_{\text{min}} = h^*(S_-), S^* = S_- \) according to Thm. 3.5
8: if \( h^*(S_-) < h_{\text{min}} \) then
9: \( h_{\text{min}} = h^*(S_-), S^* = S_- \) according to Thm. 3.5
10: end if
11: Estimate \( f_{S^*}(X) \) and return.

Alg. 1 is composed of two steps, respectively inspired by Thm. 3.3 and Thm. 3.5. Specifically, if the condition in Thm. 3.3 holds, then we directly have \( f^* = f_S \) with the whole stable set \( S \). Otherwise, we have to search over \( \{ f_S \mid S \in \text{POWER}(S) \} \) to find \( f^* \). It’s worthful to mention that such a searching cost, as also required in existing works [13, 6], can be reduced since some subsets have the same \( h^* \) (please see the supplementary for details). Practically, we can parameterize \( J_0 \) as the neural network, due to its ability to approximate any definite function [27].

### 3.3 Local Causal Discovery and Empirical Estimation

We introduce methods to implement Alg. 1. Specifically, it involves i) detecting \( K \) and determining whether \( Y \to K \); ii) estimating \( f_{S_-} \); and iii) estimating \( h(S_-, J_0) \) for optimization over \( \theta \). Due to space limit, we only introduce the main ideas of our method, with details left in the supplementary.

**Detect \( K \) and whether \( Y \to K \).** According to the definition of \( X^0_M \) and \( K \), we first detect the undirected skeleton of DAG via the PC algorithm [23], followed by detecting \( X_M \) with the assistance of domain index variable \( E \). Specifically, we have \( X_i \in X_M \) iff \( E \perp X_i \) conditioned on any subset, according to [21]. In a similar way, we can identify \( \text{PA}(X_i), \text{Ch}(X_i) \) for \( i \in M \), which is sufficient to detect \( X^0_M := \{X_i \mid i \in M, Y \in \text{PA}(X_i)\} \). Applying this method iteratively, we can detect \( \text{De}(X_M) \) and \( \text{PA}(X_i) \) for \( X_i \in \text{De}(X_M) \), which is sufficient to identify \( K := \text{De}(X'_M) \setminus X^0_M \) and have \( Y \to K \) iff \( \exists X_i \in K \) such that \( Y \in \text{PA}(X_i) \).
We consider three settings of \( P \). We then regenerate data for \( X_M \) in a sample-wise manner to generate data from \( P(X_M) \). We then regenerate data for \( X_M \)'s descendents in the intervened graph via estimating structural equations. This can be achieved because \( \text{De}(X_M) \) and their parents are identifiable, as mentioned earlier.

Estimate \( h(S, J_\theta) \). As \( \text{PA}(X_M) \) is identifiable, we regenerate data for \( X_M \) from \( J(\text{PA}(X_M)) \). Then similar to the estimation of \( S \), we regenerate data for \( \text{De}(X_M) \) in the intervened graph.

Indeed, we find it is sufficient to regenerate data for a subset of \( \text{De}(X_M) \), which can reduce the incurred approximation errors. Besides, we find searching over some subsets is redundant because they have the same/larger worst-case quadratic loss.

4 Experiment

In this section, we evaluate our method in simulation data and two real-world applications, i.e., Alzheimer’s disease diagnosis and gene function prediction.

Compared Baselines. We compare the proposed algorithm with the following methods: (i) ICP \(^7\) that only used \( Y \)’s parent nodes for prediction by assuming \( P(Y|PA(Y)) \) is invariant; (ii) IC \(^8\) assumed the invariance of \( P(Y|X_H) \) for some \( X_H \) that is not restricted to \( PA(Y) \); (iii) Anchor regression \(^9\) that interpolated between ordinary least square (LS) and causal minimax LS, which constrained the residue in the anchor (sources of distributional change) subspace to be small; (iv) IRM \(^10\) that learned a representation on which the prediction mechanism is invariant; (v) HRM \(^11\) that extended the IRM to the case when the domain labels are unknown by exploring the heterogeneity in data via clustering; and (vi) IB-IRM \(^11\) that proposed an information bottleneck constraints to supplement the invariance principle in IRM.

Implementation Details. We left implementation details in the supplementary due to space limit.

4.1 Simulation

Data Generation. We follow the causal graph in Fig. 3 to generate \( X_S := \{X_1, X_2, X_3\} \) and \( X_M := \{X_4\} \) via the following structural equations: \( x_3 \leftarrow u_3 \) with \( u_3 \sim N(-2, 1) \); \( x_2 \leftarrow \alpha_2 x_3 + u_2 \) with \( \alpha_2 = 0.5 \) and \( u_2 \sim N(0, 1) \); \( y \leftarrow g_y(x_2) + u_y \) with \( u_y \sim N(0, 1) \); \( x_4 \leftarrow \beta_4 y + u_4 \) with \( \beta_4 = e - 5 \) varied in different domains, \( u_4 \sim N(0, 1) \); \( x_1 \leftarrow g_1(x_4, y) + u_1 \) with \( u_1 \sim N(0, 0.2) \).

We consider three settings of \( g_y(\cdot) \) and \( g_1(\cdot) \): Setting-1, \( g_y(x_2) = -1.5x_2 \) and \( g_1(x_4, y) = x_4 \); Setting-2, \( g_y(x_2) = -1.5x_2, g_1(x_4, y) = x_4 + 2.5y \); Setting-3, \( g_y(x_2) = \text{Sigmoid}(x_2 + 1), g_1(x_4, y) = x_4 + 2.5y \). For each setting, we generate 10 domains with \( e \) ranging from 1 to 10, where \( n_e = 200 \) samples are generated for each domain. To remove the effect of randomness, we repeat for 10 times, in each we randomly pick five domains for training and the rest for test. In setting-1, the condition in Thm. \(^3\) holds \( (Y \) does not point to \( X_1) \) and we expect \( S \) to be the optimal set. In setting-2/3, the condition is violated. Thus, we need to compare \( h^*(S) \) for each \( S \) to find the optimal subset, as indicated by Thm. \(^3\).

Results Analysis. We report the maximal mean squared error (max MSE) over the test sets and the estimated \( h^*(S) \) from training sets for each \( S \) in Tab. 1. As shown, in setting-1, \( f_S \) has the lowest max MSE over test sets, which agrees with Thm. \(^3\) in setting-2/3, the \( S = \{X_1\} \) in setting-2, \( \{X_1, X_2\} \) in setting-3 with minimal \( h^* \) also have the minimal max MSE over test sets for \( f_S \). Besides, notice that the estimated \( h^*(S) \) are close to the maximal test MSE for most subsets, e.g., \( h^*(\{X_1, X_3\}) = 0.0079, 0.0082 \) for max MSE in setting-3, \( h^*(\{X_2, X_3\}) = 1.232, 1.136 \) for max MSE in setting-2. These results verify the correctness and more importantly, the effectiveness of Thm. \(^3\) and Thm. \(^3\) in selecting the most robust predictor. Besides, it is interesting to note that the whole set \( S \) is not the optimal set in setting-2/3 when the condition in Thm. \(^3\) is violated.
4.2 Alzheimer’s Disease Diagnosis

In this section, we evaluate our method in the diagnosis of Alzheimer’s disease, which is the most common type of dementia among elder people.

ADNI Dataset and Preprocessing. We consider the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset [http://adni.loni.ucla.edu], in which the imaging data are acquired from structural Magnetic Resonance Imaging (sMRI) scan. After data-preprocessing via Dartel VBM [29] and Statistical Parametric Mapping (SPM) for segmentation, we implement the Automatic Anatomical Labeling (AAL) atlas [30] and region index provided in [31] to partition the whole brain into 9 brain regions: frontal lobe $X_1$, medial temporal lobe $X_2$, parietal lobe $X_3$, occipital lobe $X_4$, cingulum $X_5$, insula $X_6$, amygdala $X_7$, hippocampus $X_8$, and pallidum $X_9$. In addition to brain region volumes, we also include demographics (age, gender $X_{10}$) and genetic information (number of APOE-4 alleles $X_{11}$). With these covariates, we predict the Functional Activities Questionnaire (FAQ) score $Y$ for each patient, which was commonly used to measure the AD severity [32]. In total, we use $n = 757$ samples enrolled in ADNI-GO/1/2 periods. We use the age variable to split the dataset into seven domains (age $<$ 60, 60-65, 65-70, 70-75, 75-80, 80-85, $>$ 85), which each contains $n_e = 27, 59, 90, 240, 182, 117, 42$ samples, respectively. We repeat 15 times, with each time randomly taking four domains for training and the rest for test.

Causal Discovery for Detecting $X_M$. We first implement the PC algorithm [29] to learn the undirected skeleton, followed by our algorithm to learn directions to determine $X_M$, $PA(X_M)$, $De(X_M)$, etc. As shown in Fig. 1, we have $Y \rightarrow X_2, X_8$, in which the $X_2$ (media temporal lobe) and $X_8$ (hippocampus) are early degenerated regions [33, 34]. Besides, we also observe that $X_8 \in X_M$, which echoes existing studies [19, 20] that detected distributional changes over brain region volumes across ages. Our findings provide a possible explanation that can be supported by existing studies [20]: this distributional change of atrophy is partially sourced from the hippocampus and propagates to other brain regions. To detect the optimal predictor, we compare $h^*(S_-)$ for each subset $S_-$ since the condition in Thm. 3.3 is violated, as $Y$ points to $X_2 \in K$.

![Figure 4: Results on ADNI dataset. (a) the $f_{S_-}$ in the ascending order of $h^*(x)$ (x-axis) and their worst-case errors. (b) Comparison with baselines.](image-url)
Results Analysis. Fig. 4(a) and (b) respectively show the relations between $h^*$ and maximal test MSE, and comparisons between our methods and baselines. Specifically, we sort $f_{S_\cdot}$ according to the ascending order of $h^*$ on the x-axis and show their maximal test MSEs on the y-axis in Fig. 4(a). We observe that (i) the $S^* := \{X_3, X_5\}$ that with the minimal $h^*$ also has the minimal test MSEs among all subsets; (ii) the maximal test MSE shows an approximately increasing trend with respect to $h^*(S_\cdot)$ however with oscillations. These observations suggest the utility of Thm. 3.5 in finding the optimal predictor especially when the whole set is not necessarily the optimal subset. It is also interesting to note that parietal lobe $X_8$, cingulum $X_5$, and hippocampus $X_8 \in X_{M}$ used by $f^*$, are common AD biomarkers [33, 35, 36]. While for oscillations, it may be due to the limited number of test domains, and the worst case among which may not fully reflect the worst case in the whole distribution family $P$. Another interesting phenomenon in Fig. 4(a) is the nearly piece-wise constant curves especially in the first half. Besides, it is interesting to observe the sudden jump in MSE from 1.2 to 2.8 once the $X_2$ is used, which may be due to $Y \rightarrow X_2 \in K$. In addition, we find that the whole set $S$ (marked by $\triangle$) is not the optimal set.

After obtaining $S^*$, it is shown in Fig. 4(b) that our method can outperform others. Specifically, the advantages over ICP, IC, IRM, and their extensions, i.e., HRM, IB-IRM can be attributed to the proper use of invariance beyond causal features; while the advantage over Anchor regression may be due to the non-parametric form of causal min-max objective that is more aligned with the goal of robustness, especially when the extent of distributional shifts is allowed to vary wildly.

4.3 Gene Function Prediction

In this section, we evaluate our method in gene function prediction, which can potentially help better understand the human-disease progress [37].

Dataset. We consider the International Mouse Phenotyping Consortium (IMPC) dataset (http://www.crm.umontreal.ca/2016/Genetics16/competition_e.php) that was originally published in a causal inference challenge [38] and later used as a benchmark for domain generalization [39]. The IMPC contains the hematology phenotype of both wild-type mice and mutant mice with 13 kinds of single-gene knockout. To predict the gene function, we knock out this gene and assess the cell counts of monocyte $Y$, with cell counts of neutrophil $X_1$, lymphocyte $X_2$, eosinophil $X_3$, basophil $X_4$, and large unstained cell (LUC) $X_5$ as covariates. We use the kind of knocked-out gene to divide domains. The training domains contain the wild-type mice and five randomly picked gene knockouts. The test domains contain the rest nine kinds of gene knockouts. This random train-test split is repeated 45 times.

Causal Discovery For Detecting $X_{M}$. We implement our algorithm to detect $X_{M}$, PA($X_{M}$), De($X_{M}$), etc. We show the detected causal graph in Fig. 5 where we have $Y \rightarrow X_2, X_3$, and $X_2 \rightarrow X_4$, which respectively echoes the existing studies that monocyte can activate the lymphocyte [40] and increase the number of LUC [41], and that the lymphocyte participates in activating basophil [42]. Since $X_5$ is mutable and $X_2 \in De(X_5)$ is pointed by $Y$, the condition in Thm. 3.3 is violated. We then compare $h^*(S_\cdot)$ to find the optimal predictor.

Results Analysis. Fig. 6(a) presents the maximal MSE over the test sets for each predictor $f_{S_\cdot}$. Similarly, we observe that the maximal test MSE increases with $h^*$ and the subset (i.e., \{X_3\}) with minimal $h^*$ also performs the best. The ability of $X_3$ in predicting $Y$ can be explained by the directed activated path from monocyte $Y$ to basophil $X_4$. After obtaining $f^*$, we compared it with other methods. As shown in Fig. 6(b), ours can outperform others by a significant margin, which together with the Alzheimer’s disease experiment, demonstrates the utility of our method in learning robust predictors.
5 Conclusion and Discussion

In this paper, we propose to learn the min-max predictor for robustness to distributional shifts, in the supervised regression scenario. We first attribute the distributional shifts to the mutable variables, in the framework of causal factorization. We then identify a set of invariant predictors. Among these predictors, we provide a sufficient and necessary condition for a predictor to be min-max optimal. Benefit from the justifiability of this condition, we propose an algorithm that incorporates a permutation-regeneration scheme for practical estimation. Our method can outperform others in Alzheimer’s disease diagnosis and gene function prediction. This can be explained by the effectiveness of our method in detecting the min-max predictor. In both experiments, it is also interesting to find that the whole set is not min-max optimal.

Limitations and Future Directions. Our algorithm has to search over most subsets in the stable set, which limits its application in high-dimensional scenarios. To resolve this problem, we will investigate the causal structure in the parametric model, to pursue a delicate theoretical solution. In addition, we are interested to extend our study to scenarios where the DAG is also allowed to vary across domains, which may happen when the number of domains is large.
## Appendix

### Contents

1. **Introduction**
2. **Problem Setup & Basic Assumptions**
3. **Methodology**
   3.1 Causal Framework and Invariant Predictors
   3.2 Identification with Min-max Property
   3.3 Local Causal Discovery and Empirical Estimation
4. **Experiment**
   4.1 Simulation
   4.2 Alzheimer's Disease Diagnosis
   4.3 Gene Function Prediction
5. **Conclusion and Discussion**

**A** Appendix for Sec. 3.1: Causal Framework and Invariant Predictors
   A.1 Proof for Prop. 3.1: Invariance Property

**B** Appendix for Sec. 3.2: Identification with Min-max Property
   B.1 Proof for Thm. 3.3: Graphical Criterion for \( f^* = f_S \)
   B.2 Proof for Prop. 3.4: Justifiability of Thm. 3.3
   B.3 Counter Example of \( f^* \neq f_S \)
   B.4 Proof for Thm. 3.5: Min-max Property
   B.5 Proof for Prop. 3.6: Identifiability in Thm. 3.5

**C** Appendix for Sec. 3.3: Local Causal Discovery and Empirical Estimation
   C.1 Details of Local Causal Discovery
   C.2 Details of Estimation of \( f_{S_+} \)
   C.3 Equivalence of Subsets for Computational Efficiency

**D** Appendix for Sec. 4: Experiment
   D.1 Simulation
   D.2 Alzheimer's Disease Diagnosis
   D.3 Gene Function Prediction
A Appendix for Sec. 3.1: Causal Framework and Invariant Predictors

A.1 Proof for Prop. 3.1: Invariance Property

Proposition 3.1 (Invariance Property). Suppose \( \mathcal{E} \) contains any \( e \) such that \( P^e \) admits the following factorization: \( p^e(x, y) = p(y|pa(y)) \prod_{i \in S} p(x_i|pa(x_i)) \prod_{i \in M} p^e(x_i|pa(x_i)) \). Let \( \mathcal{P} \) denote the distribution set \( \{ P^e | e \in \mathcal{E} \} \). Then, for each \( S_- \subset S \), we have \( f_{S_-} := E_{P^e}[Y|X_{S_-}, do(x_M)] \) is invariant across \( \mathcal{P} \).

Proof. By definition,
\[
p^e(y, x_{S_-}|do(x_M)) = p(y|pa(y)) \prod_{i \in S_-} p(x_i|pa(x_i))
\]
is invariant for \( P^e \in \mathcal{P} \). As
\[
p^e(y|x_{S_-}, do(x_M)) = \int_{x_S \setminus x_{S_-}} p^e(y, x_S|do(x_M)) dy dx_S \setminus x_{S_-},
\]
we have \( p^e(y|x_{S_-}, do(x_M)) \) is also invariant for \( P^e \in \mathcal{P} \). According to
\[
E_{P^e}[Y|X_{S_-}, do(x_M)] = \int_{x_S \setminus x_{S_-}} yp^e(y|x_{S_-}, do(x_M)) dy,
\]
we have the invariance of \( E_{P^e}[Y|X_{S_-}, do(x_M)] \), thus the invariance of the predictor \( f_{S_-} \). \( \Box \)

B Appendix for Sec. 3.2: Identification with Min-max Property

B.1 Proof for Thm. 3.3: Graphical Criterion for \( f^* = f_S \)

Theorem 3.3 (Graphical Criterion for \( f^* = f_S \)). Suppose assumptions 2.1.2.2 hold. Denote \( X_{2, y} := X_M \cap \text{Ch}(Y) \) as mutable variables in \( Y \)'s children, and \( K := \text{De}(X_{2, y}^0) \setminus X_{2, y}^0 \) as descendants of \( X_{2, y}^0 \). Then, \( p(y|x_S, do(x_M)) \) can degenerate to conditional distribution if and only if \( Y \) does not point to any member of \( K \). Further, under either of the two conditions, we have \( f^* = f_S \).

Proof. Denote the causal DAG as \( G \), the intervened graph that removes all arrowheads into \( V \) as \( G\overline{\overline{V}} \). Define \( X_{1, M} := X_M \setminus X_{2, y}^0, K_2 := (X \setminus X_{2, y}^0) \setminus \text{De}(X_{2, y}^0) \).

We firstly prove the equivalence of the following conditions (1), (2), and (3):

1. \( Y \perp_{G\overline{\overline{V}}} K[K_2]; \)
2. \( Y \) and \( K \) are not adjacent in \( G\overline{\overline{V}} \)
3. \( p(y|x_S, do(x_M)) \) can degenerate to conditional distribution.

(1)\(\Rightarrow\)(2) If \( Y \) and \( K \) are adjacent, they are also adjacent in \( G\overline{\overline{V}} \), because \( K \cap X_{2, y}^0 = \emptyset \), so \( Y \) and \( K \) cannot be d-separated by any variable in \( G\overline{\overline{V}} \), which contradicts with (1).

(2)\(\Rightarrow\)(3) Define
\[
I := p(y|k, k_2, do(x^0_M)) = \frac{p(y|pa(y)) \prod_{X \in K} p(x_j|pa(x_j)) \prod_{X \in K_2} p(x_i|pa(x_i))}{\int_{Y} p(y|pa(y)) \prod_{X \in K} p(x_j|pa(x_j)) \prod_{X \in K_2} p(x_i|pa(x_i)) dy}.
\]
Since \( \text{PA}(Y) \cap \{X_M^0, K\} = \emptyset \) and \( \forall X_i \in K_2, \text{PA}(X_i) \cap \{X_M^0, K\} = \emptyset \), we have:
\[
I = \frac{p(y, k_2) \prod_{X \in K} p(x_j|pa(x_j))}{\int_{Y} p(y, k_2) \prod_{X \in K} p(x_j|pa(x_j)) dy}.
\]
If \( Y \) and \( K \) are not adjacent, then \( \forall X_j \in K, Y \notin \text{PA}(X_j) \). Therefore, \( I = \frac{p(y, k_2)}{\int_{Y} p(y, k_2) dy} = p(y|k_2). \)

Note the edge between \( Y \) and \( X \in K \) can only be \( Y \rightarrow X \).
We firstly show that $p^c(y|x_S, do(x_M))$ can not degenerate to any conditional distribution, i.e., (3) does not hold, then $p^c(y|x_S, do(x_M))$ can not degenerate to any conditional distribution.

We firstly show $Y \not\in G_{X_M^0}$, i.e., $p^c(y|x_S, do(x_M)) \neq p^c(y|k_2, do(x_M^0))$, then show $p^c(y|x_S, do(x_M)) \neq p^c(y|k_2, do(x_M^0)) \Rightarrow p^c(y|x_S, do(x_M))$ can not degenerate to any conditional distribution.

Since $Y \not\in PA(X_M^1)$, we have:

$$p^c(y|x_S, x_M^0, do(x_M^0)) = \frac{\int p^c(y|x_S, x_M^0|do(x_M^0)) dy}{\int p^c(y|x_S, x_M^0|do(x_M^0)) dy}$$

Since $K \cup K_2 = X_S \cup X_M^1$, we have $p^c(y|x_S, do(x_M)) = p^c(y|k, k_2, do(x_M^0))$. Thus, we can prove: $Y \not\in G_{X_M^0}$, $K|K_2 \Rightarrow p^c(y|x_S, do(x_M)) = p^c(y|k, k_2, do(x_M^0)) \neq p^c(y|k_2, do(x_M^0))$.

Next, we prove $p^c(y|x_S, do(x_M)) \neq p^c(y|k_2, do(x_M^0)) \Rightarrow p^c(y|x_S, do(x_M))$ can not degenerate to any conditional distribution.

Suppose $p^c(y|x_S, do(x_M)) = p^c(y|k', k_2, do(x_M^0))$. We will show if $k' \neq \emptyset$, then the do-operator can not be removed with either Rule 2 (action to observation) or Rule 3 (deletion of action). To express $do(x_M^0)$ explicitly, denote $X_M^1 = \{X_M^0, i=1\}$ and $p^c(y|k', k_2, do(x_M^0)) = p^c(y|k', k_2, do(x_M^0))$.

- Rule 2 can not remove the do-operator of any $X_{M,i}^0 \in X_M^0$.

Recall Rule 2 states that “$p(y|do(x), do(z), w) = p(y|do(x), z, w)$ if $Y \not\in G_{X_M^0}$, $Z|X$, $W$ for any disjoint subsets of variables $X$, $Y$, $Z$, and $W$ ”.

If Rule 2 can remove the do-operator of $X_{M,i}^0 \in X_M^0$, then

$$Y \perp G_{X_M^0 \{X_{M,i}^0\} X_{M,i}^0} X_{M,i}^0|K'|, K_2, X_M^0 \setminus \{X_{M,i}^0\}.$$  \hspace{1cm} (4)

As we have $Z = \{X_{M,i}^0\}, X = X_M^0 \setminus X_{M,i}^0, W = K' \cup K_2$ in the notations of Rule 2.

In the following, we explain why Eq. 4 can not be true. Note $X_{M,i}^0 \in Ch(Y)$ and the direct edge $Y \rightarrow X_{M,i}^0$ is reserved in the intervened graph $G_{X_M^0 \{X_{M,i}^0\} X_{M,i}^0}$, which means that $Y$ and $X_{M,i}^0$ can not be d-separated by any set of variables in the intervened graph. Thus, Eq. 4 can not be true.

- Rule 3 can not remove the do-operator of all $X_{M,i}^0 \in X_M^0$.

Recall Rule 3 states that “$p(y|do(x), do(z), w) = p(y|do(x), w)$ if $Y \not\in G_{X_M^0}$, $Z|X$, $W$ for any disjoint subsets of variables $X$, $Y$, $Z$, and $W$ ”. If Rule 3 can remove the do-operator of $X_M^0$, then:

$$Y \perp G_{X_M^0(K' \cup K_2)} X_M^0|K' \cup K_2.$$  \hspace{1cm} (5)

---

\footnote{\text{\(Z(W)\) is the set of \(Z\)-nodes that are not ancestors of any \(W\)-node in \(G_X\).}}
because the notations in Rule 3 mean \( X = \emptyset, Z = X_M^0, W = K' \cup K_2 \). In the following, we will show that when \( K' \neq \emptyset \), Eq. [3] cannot hold. When \( K' \neq \emptyset \), note by definition \( K' \subset \text{De}(X_M^0) \), so \( \text{An}(K') \cap X_M^0 \neq \emptyset \). Therefore, \( X_M^0 (K' \cup K_2) = X_M^0 \setminus \{ \text{An}(K') \cup K_2 \} \neq X_M^0 \). That is \( X_M^0 \setminus X_M^0 (K' \cup K_2) \neq \emptyset \). Suppose \( X_M^0 \setminus X_M^0 (K' \cup K_2) \), then the edge \( Y \rightarrow X_M^0 \) is in the intervened graph \( G_{X_M^0 (K' \cup K_2)} \), so \( Y \) and \( X_M^0 \) can not be d-separated by any variable set. So Eq. [3] does not hold.

In summary, we have proved that when \( K' \neq \emptyset \), the \( do \)-operator on \( X_M^0 \) can not be removed entirely by Rule 2 and 3.

Besides, according to Corollary 3.4.2 in [14], the inference rules are complete in the sense that if the intervention probability (with \( do \)) can be reduced to a probability expression (without \( do \)), the "reduction" can be realized by a sequence of transformations, each conforming to one of the Inference Rules 1-3. Note that only Rule 2 and 3 are related to the disappearance of \( do \)-operator, so it is sufficient to prove that Rule 2 and 3 can not remove the \( do \)-operator on \( X_M^0 \).

We then prove under either of conditions (1), (2), or (3), \( f^* = f_S \).

Given any one of the three conditions (1), (2), or (3), \( f^*(x) = E_{P^e}[Y|x_S, do(x_M)] \) satisfies the following min-max property:

\[
f^*(x) = \arg\min_{f:X \rightarrow Y} \max_{P \in P} E_P[Y - f(x)]^2.
\]

Under any one of the conditions (1)-(3), we have \( p^e(y|x_S, do(x_M)) = p^e(y|k_2) \) for \( P^e \in P \).

For \( P^e \in P \), let \( p^e(x_M^0) = \sum_{X_i \in V \setminus X_M^0} p^e(v) \) be the marginal distribution of \( X_M^0 \). Define \( \tilde{p}^e \) as:

\[
\tilde{p}^e(v) = p(y|pa(y)) \prod_{X_i \in K} p^e(x_i|pa(x_i)) \prod_{X_i \in K_2} p^e(x_i|pa(x_i)) p^e(x_M^0),
\]

by replacing the term \( \prod_{X_i \in X_M^0} p(x_i|pa(x_i)) \) in \( p^e(v) \) with \( p^e(x_M^0) \).

(i) By the definition of \( P \), \( \tilde{p}^e \in P \) and \( \tilde{p}^e(y|x) = \tilde{p}^e(y|x_S, x_M^0) \).

(ii) In the following, we will show \( \tilde{p}^e(y, k_2) = p^e(y, k_2) \).

First, note that \( K \subset \text{De}(X_M^0) \) and \( X_M^0 \subset \text{Ch}(Y) \), we have \( K \cup X_M^0 \subset \text{De}(Y) \). Thus, \( PA(Y) \cap \{ K \cup X_M^0 \} = \emptyset \) because otherwise there would be a cycle.

Second, \( PA(K_2) \cap \{ K \cup X_M^0 \} = \emptyset \) because if there exist \( X_i \in K \cup X_M^0 \) and also \( X_i \in PA(K_2) \), then \( K_2 \cap \text{De}(X_M^0) \subset \emptyset \), which contradicts with the definition that \( K_2 := (X \setminus X_M^0) \setminus \text{De}(X_M^0) \).

In summary, we have \( PA(K_2 \cup Y) \cap \{ K \cup X_M^0 \} = \emptyset \), which leads to

\[
p^e(k_2, y) = \int p(y|pa(y)) \prod_{X_i \in K_2} p^e(x_i|pa(x_i)) \prod_{X_i \in K} p^e(x_i|pa(x_i)) \prod_{X_i \in X_M^0} p^e(x_i|pa(x_i)) d\theta_M dK
\]

\[
= p(y|pa(y)) \prod_{X_i \in K_2} p^e(x_i|pa(x_i)) \int \prod_{X_i \in K} p^e(x_i|pa(x_i)) \prod_{X_i \in X_M^0} p^e(x_i|pa(x_i)) d\theta_M dK
\]

\[
= p(y|pa(y)) \prod_{X_i \in K_2} p^e(x_i|pa(x_i))
\]

and

\[
\tilde{p}^e(k_2, y) = \int p(y|pa(y)) \prod_{X_i \in K_2} p^e(x_i|pa(x_i)) \prod_{X_i \in K} p^e(x_i|pa(x_i)) p^e(x_M^0) d\theta_M dK
\]

\[
= p(y|pa(y)) \prod_{X_i \in K_2} p^e(x_i|pa(x_i)) \int \prod_{X_i \in K} p^e(x_i|pa(x_i)) p^e(x_M^0) d\theta_M dK
\]

\[
= p(y|pa(y)) \prod_{X_i \in K_2} p^e(x_i|pa(x_i))
\]

Therefore, we have \( \tilde{p}^e(k_2, y) = p^e(k_2, y) \).
Note that \( K_2 \subset X \), we have

\[
\text{Var}_{P^e}(Y|K_2) = E_{P^e} [ \text{Var}_{P^e}(Y|X)|K_2] + \text{Var}_{P^e} [E_{P^e}(Y|X)|K_2],
\]

therefore

\[
E_{P^e} [\text{Var}_{P^e}(Y|K_2)] = E_{P^e} [\text{Var}_{P^e}(Y|X)] + E_{P^e} [\text{Var}_{P^e} E(Y|X)|K_2],
\]

and hence \( E_{P^e} [\text{Var}_{P^e}(Y|K_2)] \geq E_{P^e} [\text{Var}_{P^e}(Y|X)] \).

(iii) Because \( \bar{p}^e(k_2, y) = p^e(k_2, y) E_{\bar{P}^e} [\text{Var}_{P^e}(Y|K_2)] = E_{P^e} [E_{P^e}(Y|X)] \), we have \( E_{P^e} [\text{Var}_{P^e}(Y|K_2)] \geq E_{P^e} [\text{Var}_{P^e}(Y|X)] \). Besides, since \( \bar{p}^e(y|x) = \bar{p}^e(y|k_2) \),

\[
E_{\bar{P}^e} [\text{Var}_{\bar{P}^e}(Y|K_2)] = E_{P^e} [\text{Var}_{P^e}(Y|X)].
\]

(iv) In summary, for each \( P^e \in \mathcal{P} \), we may construct \( \bar{P}^e \) such that

\[
E_{P^e} [\text{Var}\bar{P}^e(Y|X)] \geq E_{P^e} [\text{Var}_{P^e}(Y|X)].
\]

Denote \( \bar{P} := \{ \bar{P}^e \in P \} \) and \( P^* := \text{argmax}_{P^e \in \mathcal{P}} E_{P^e} [\text{Var}_{P^e}(Y|X)] \), then \( P^* \in \bar{P} \). Besides, note that for any \( \bar{P}^e \in \bar{P} \), \( \bar{P}^e \) is identifiable; besides, we can determine whether \( Y \rightarrow K \) from the joint distribution over training domains.

\[ \square \]

B.2 Proof for Prop. 3.4: Justifiability of Thm. 3.3

**Proposition 3.4.** Denote \( X_0^M := X_M \cap \text{Ch}(Y) \) as mutable variables in \( Y \)'s children, and \( K := \text{De}(X_0^M) \setminus X_0^M \) as descendants of \( X_0^M \). Under assumptions 2.1, 2.2, the \( K \) is identifiable; besides, we can determine whether \( Y \rightarrow K \) from the joint distribution over training domains.

**Proof.** We firstly show \( K \) is identifiable. Since all variables in \( K \) are descendants of \( Y \), we have \( Y \rightarrow X_i, X_i \in K \) iff \( X_i \) is adjacent to \( Y \) in the skeleton of DAG (which is identifiable under assumption 2.2). Thus, we can determine whether \( Y \rightarrow K \).

Note that \( K = (X_0^M \setminus \text{De}(X_0^M)) \cap \text{De}(X_M) = (X_0^M \setminus \text{De}(X_0^M)) \cap \{ \text{De}(X_M^0) \cup X_M^0 \} \). So it suffices to prove the identifiability of \( X_0^M \setminus \text{De}(X_0^M) \), where \( X_M^0 := X_M \cap \text{Ch}(Y) \). This can be accomplished by three steps: (i) identification of \( X_M^0 \), (ii) identification of \( X_M^0 \), and (iii) identification of \( X_M^0 \).

The following algorithm shows step (i), which is the same as [21].

**Algorithm 2 Detection of \( X_M^0 \) and construct the causal skeleton of \( G \)

1. Start with \( X_M = \emptyset \). For \( V_i \in V \), test if \( V_i \perp E \) or if there exist a subset \( C_{v_i, e} \subseteq V \) such that \( V_i \perp E|C_{v_i, e} \). If \( V_i \perp E \) and there exists no such \( C_{v_i, e} \), then include \( V_i \) in \( X_M = X_M \cup V_i \).

2. Start with an undirected graph \( G \) including edges for any two variables in \( V \) and the arrows \( E \rightarrow V_i \) for \( V_i \in X_M \). For each pair of \( (V_i, V_j) \). If \( V_i \perp V_j \) or there exists a subset \( C_{v_i, v_j} \subseteq V \) such that \( V_i \perp V_j|C_{v_i, v_j} \), we delete the edge \( V_i - V_j \) from \( G \).

The following Alg. [3] shows the steps (ii) and (iii), which basically relies on the faithful assumption (conditional independence in probability \( \Rightarrow \) d-separation in graph).

Explanations for 2.2:

- Line [7] in 2.2: The set \( A \) is the final output. The set \( B \) only plays a part as an instrumental set that starts with \( X_M \cap \text{Ch}(Y) \) and ends with \( \emptyset \). During the process, \( B \) stores the nodes in \( X_M \cap \text{Ch}(Y) \) that has not been searched for the children. Once \( X_j \in B \) is searched, it is excluded from the set \( B \) (Line [18]) and the children of \( X_j \) are added to \( B \) if it has not been visited (Line [8] and [14]), which is essentially a breadth-first-search algorithm.

15
Algorithm 3 Detection of $X_{M}^{M}$ and $X_{M}^{D} \cup \text{De}(X_{M}^{M})$

2.1 Detect $X_{M}^{M} := X_{M} \cap \text{Ch}(Y)$:
1: for $X_i \in X_M$ and adjacent to $Y$ do
2: if $Y \not\perp E[C_{y,e} \cup \{X_i\}]$
3: then $X_i \in X_M \cap \text{Ch}(Y)$
4: end for

2.2 Detect $(\text{Ch}(Y) \cap X_M) \cup \text{De}(\text{Ch}(Y) \cap X_M)$
1: Start with $A = B = \text{Ch}(Y) \cap X_M$ and visited($X_i$) = FALSE
2: while $B \neq \emptyset$ do
3: for $X_j \in B$ do
4: if $X_j \in \text{Adj}(X_j)$ do
5: if $X_j \not\in X_M$ and $X_j \perp E[C_{x,i} \cup \{X_j\} \setminus D_{x} \cdot \{X_j\}]$
6: $A = A \cup \{X_i\}$
7: if visited($X_i$) = FALSE then
8: $B = B \cup \{X_i\}$
9: end if
10: end if
11: if $X_j \in X_M$ and $X_j \not\in \text{Adj}(Y)$ and $X_j \perp Y[C_{x,i} \cup \{X_j\} \setminus D_{y} \cdot \{X_j\}]$ then
12: $A = A \cup \{X_i\}$
13: if visited($X_i$) = FALSE then
14: $B = B \cup \{X_i\}$
15: end if
16: end if
17: end for
18: Let $B = B \setminus \{X_j\}$
19: end for
20: end while

* Line 5 to 10 (the case when $X_i \not\in X_M$): The fact $X_i \not\in X_M$ means $E$ and $X_i$ are not adjacent. Besides, note that $X_i \in X_M \cap \text{Ch}(Y)$, there is a structure in the form $E \rightarrow \cdots \rightarrow X_j \rightarrow X_i$ where $X_i$ and $E$ are not adjacent. In the notation $X_i \perp E[C_{i,j} \cup \{X_i\} \setminus D_{i,j}]$, $C_{i,j}$ denotes a separating set such that $X_i \perp E[C_{i,j}]$ and $D_{i,j}$ denotes the set of variables along the directed path between $E \rightarrow \cdots \rightarrow X_j$. The existence of $C_{i,j}$ is guaranteed since $X_i$ and $E$ are not adjacent, so a separating set has been found when constructing the skeleton. The set $D_{i,j}$ is also clear as it is determined in the breadth-first-search process.

* Line 11 to 19 (the case when $X_i \in X_M$): Firstly, we explain why it is unnecessary to consider the case when $X_i \in X_M$ and $X_i \in \text{Adj}(Y)$. If $X_i \in \text{PA}(Y)$, $X_i$ cannot be in $\text{De}(\text{Ch}(Y) \cap X_M) \cup \{\text{Ch}(Y) \cap X_M\}$ as it would induce a cycle in this way. If $X_i \in \text{Ch}(Y)$, it means $X_i \in \text{Ch}(Y) \cap X_M$ and has been identified in 2.1 and included in set $A$ in the beginning.
So the remaining case is when $X_i \in X_M$ and $X_i \not\in \text{Adj}(Y)$. Note in this case $X_j \in \text{Ch}(Y)$ or $X_j \in \text{De}(Y)$, there exists a structure $Y \rightarrow \cdots \rightarrow X_j \rightarrow X_i$, which is the same as $E \rightarrow \cdots \rightarrow X_j \rightarrow X_i$ in Line 5 to 10.

B.3 Counter Example of $f^* \neq f_S$

Consider the DAG in Fig. 7, in which we denote $Y, X_s, X_m$ are binary variables. We will show that in this scenario, there exists $P(Y), P(X_s | X_m, Y)$ such that $f^*_S := \mathbb{E}[Y | x_s, do(x_m)]$ is not min-max optimal. We show this by proving that:

$$\mathbb{E} [Y - \mathbb{E}[Y | x_s, do(x_m)]]^2 > \mathbb{E} [Y - \mathbb{E}[Y | do(x_m)]]^2.$$  \hspace{1cm} (6)

Since we have that

$$\mathbb{E} [Y - \mathbb{E}[Y | x_s, do(x_m)]]^2 = \mathbb{E}[Y^2] + \mathbb{E}[\mathbb{E}^2 | Y | x_s, do(x_m))] - 2\mathbb{E}[Y \cdot \mathbb{E}[Y | x_s, do(x_m)]],$$

16
Besides, we have
\[ E \] 
\[ a \] 
\[ \text{Substituting Eq. (10) into Eq. (8), (9), we have} \]
\[ \text{E} \left[ E^2[Y_\cdot | x_s, do(x_m)] \right] > 2 \text{E} \left[ Y \cdot \text{E}[Y | x_s, do(x_m)] \right] - E^2[Y]. \]  
(7)

Since we have \( p(y|x_s, do(x_m)) = \frac{p(y)p(x_m|y)}{\sum_y p(y)p(x_m|y)} \), we have
\[ \text{E}[Y | x_s, do(x_m)] = \frac{p(y = 1)p(x_m|y = 1)}{\sum_y p(y)p(x_m|y)}. \]  
(10)

Substituting Eq. (10) into Eq. (8), (9), we have
\[ \text{E} \left[ E^2[Y | X_s, do(X_m)] \right] = \sum_{x_s \cdot x_m} \left[ \sum_y p(x_s|x_m, y)p(x_m|y)p(y) \right] \cdot \left[ \frac{p(y = 1)p(x_s|x_m, y = 1)}{\sum_y p(y)p(x_m|y)} \right]^2, \]
\[ \text{E} \left[ Y \cdot \text{E}[Y | X_s, do(X_m)] \right] = \sum_{x_s \cdot x_m} \left[ \sum_y p(x_s|x_m, y)p(x_m|y)p(y) \right] \cdot \left[ \frac{p(y = 1)p(x_s|x_m, y = 1)}{\sum_y p(y)p(x_m|y)} \right] \]
\[ = \sum_{x_s \cdot x_m} \left[ \sum_y p(x_s|x_m, y = 1)p(x_m|y = 1)p(y = 1) \right] \cdot \left[ \frac{p(y = 1)p(x_s|x_m, y = 1)}{\sum_y p(y)p(x_m|y)} \right]. \]

Denote \( a_y := p(y = 1), p(x_m = 1|y) := a_{my}, p(x_s = 1|x_m, y) = a_{smy}, \) then the left hand side in Eq. (7) has
\[ \text{E} \left[ E^2[Y | X_s, do(X_m)] \right] = \sum_{x_s = x_m} \left[ \sum_y p(x_s|x_m, y)p(x_m|y)p(y) \right] \cdot \left[ \frac{p(y = 1)p(x_s|x_m, y = 1)}{\sum_y p(y)p(x_m|y)} \right]^2 + \]
\[ = \sum_{x_s \cdot x_m} \left[ \sum_y p(x_s|x_m, y)p(x_m|y)p(y) \right] \cdot \left[ \frac{p(y = 1)p(x_s|x_m, y = 1)}{\sum_y p(y)p(x_m|y)} \right]^2 + \]
\[ \sum_{x_s \cdot x_m} \left[ \sum_y p(x_s|x_m, y)p(x_m|y)p(y) \right] \cdot \left[ \frac{p(y = 1)p(x_s|x_m, y = 1)}{\sum_y p(y)p(x_m|y)} \right]^2 + \]
\[ \sum_{x_s \cdot x_m} \left[ \sum_y p(x_s|x_m, y)p(x_m|y)p(y) \right] \cdot \left[ \frac{p(y = 1)p(x_s|x_m, y = 1)}{\sum_y p(y)p(x_m|y)} \right]^2. \]
The right-hand side has
\[ 2\mathbb{E} [Y | \mathbb{E} [Y | x, \text{do}(x_m)]] - \mathbb{E} [Y^2] = 2 \left( (x_s = 1, x_m = 1) \frac{a_y a_{s11} a_{m1}}{a_y a_{s11} + (1 - a_y) a_{s10}} + \right. \\
\left. (s = 1, m = 0) \frac{a_y a_{s01} (1 - a_{m1})}{a_y a_{s01} + (1 - a_y) a_{s00}} + \right. \\
\left. (x_s = 0, x_m = 1) \frac{a_y (1 - a_{s1})}{a_y (1 - a_{s1}) + (1 - a_y) a_{s01}} + \right. \\
\left. (x_s = 0, x_m = 0) \frac{a_y (1 - a_{s01})(1 - a_{m1})}{a_y (1 - a_{s01}) + (1 - a_y)(1 - a_{s00})} \right) - a_y^2. \]

When \( a_y \neq 0 \), the term \( a_y^2 \) can be removed. Then let \( a_y \to 0 \), the left-hand side approximates to:
\[ \frac{a_{s11} a_{m0}}{a_{s10}} + \frac{a_{s01}^2 (1 - a_{m0})}{a_{s00}} + \frac{(1 - a_{s11})^2 a_{m0}}{(1 - a_{s10})} + \frac{(1 - a_{s01})^2 (1 - a_{m0})}{(1 - a_{s00})}; \]

and the right hand side approximates to:
\[ 2 \left( \frac{a_{s11} a_{m1}}{a_{s10}} + \frac{a_{s01}^2 (1 - a_{m1})}{a_{s00}} + \frac{(1 - a_{s11})^2 a_{m1}}{(1 - a_{s10})} + \frac{(1 - a_{s01})^2 (1 - a_{m1})}{(1 - a_{s00})} \right) - 1 \]

Then the Eq. (2) is equivalent to:
\[ \frac{a_{s11} (a_{m0} - 2a_{m1})}{a_{s10}} + \frac{a_{s01}^2 (2a_{m0} - a_{m1})}{a_{s00}} + \frac{(1 - a_{s11})^2 (a_{m0} - 2a_{m1})}{(1 - a_{s10})} + \frac{(1 - a_{s01})^2 (2a_{m1} - a_{m0} - 1)}{(1 - a_{s00})} > -1. \]

Let \( a_{s10} \to 0, a_{s11} \to 1, a_{s01} = a_{s10} = 0.5 \) and \( a_{m0} - 2a_{m1} > 0 \), the above inequality holds.

B.4 Proof for Thm. 3.5: Min-max Property

**Theorem 3.5** (Min-max Property). Denote \( h^*(S_-) := \max_{f \in \mathcal{F}} \mathbb{E}_{P_x} [(Y - f_{S_-}(x))^2] \) as the maximal expected loss over \( J \) for \( S_- \). Then, we have \( h^*(S_-) = \max_{S_- \subseteq S} \mathbb{E}_{P_x} [(Y - f_{S_-}(x))^2] \). In this regard, the optimal subset \( S^* \) for \( f^* = f_{S^*} \) can be attained via \( S^* := \arg \min_{S_- \subseteq S} h^*(S_-) \).

**Proof.** We show the maximum loss is attained when \( X_M \) is a definite function of \( \text{PA}(X_M) \).

Let \( f_{S_-}(x_{S_-}, x_M) := E[Y | x_{S_-}, \text{do}(x_M)] \) be an invariant predictor. Then
\[ L^*_{P_x}(f_{S_-}) = \int \int (y - f_{S_-}(x_{S_-}, x_M))^2 p(y/pa(y)) \prod_{i \in S} p(x_i|pa(x_i)) \prod_{i \in M} p^*(x_i|pa(x_i)) dy dx. \]

And the maximum loss
\[ L^*_{S_-} = \arg \max_{f_{S_-}} L^*_{P_x}(f_{S_-}) = \arg \max_{\{p^*(x_i|pa(x_i)) \}_{i \in M}} L^*_{P_x}(f_{S_-}). \]

Let \( X' := X \setminus \{X_M \setminus \text{PA}(X_M)\} \) and \( h(x_M, pa(x_M)) := \int_{X'} (y - f_{S_-}(x))^2 \prod_{i \in X'} p(x_i|pa(x_i)) dx' \), which does not rely on the mutable distribution \( \{p^*(x_i|pa(x_i)) \}_{i \in M} \). Let \( m^*(pa(x_M)) := \arg \min_{x_M} h(x_M, pa(x_M)) \).

Firstly, consider the case of \( X_M = \{X_M\} \). Then
\[ \max_{P_x} L^*_{P_x}(f_{S_-}) = \max_{P} \left( \int h(x_M, pa(x_M)) p^*(x_M|pa(x_M)) dx_M \right) \prod_{X_i \in \text{PA}(X_M)} p(x_i|pa(x_i)) dpa(x_M) \]
\[ = \int \left( \max_{P_x} \left( \int h(x_M, pa(x_M)) p^*(x_M|pa(x_M)) dx_M \right) \right) \prod_{X_i \in \text{PA}(X_M)} p(x_i|pa(x_i)) dpa(x_M) \]
\[ = \int h(m^*(pa(x_M)), pa(x_M)) \prod_{X_i \in \text{PA}(X_M)} p(x_i|pa(x_i)) dpa(x_M). \]
When $X_M$ is multivariate, we consider the maximization sequentially by the topological order $\{X_{M,1}, X_{M,2}, \cdots, X_{M,j}\}$, where $X_{M,j}$ is a node that is not a parent of any other nodes in $\{X_{M,i}; i \geq j\}$ in the sub-graph over $X_M$. That is, we firstly consider 
\[
\max_{p_e} \left( \prod_{i=1}^{j-1} \max_{p_e} (x_{M,i}|pa(x_{M,i})) \right)
\]
and factorize $\max_{p_e} \left\{ \cdots \right\}$ as 
\[
\max_{p_e} (x_{M,1}|pa(x_{M,1})) \cdots \max_{p_e} (x_{M,2}|pa(x_{M,2})) \max_{p_e} (x_{M,1}|pa(x_{M,1})) \left\{ \cdots \right\}.
\]

Note that the sub-graph on $X_M$ is always a DAG, so such a topological order always exists.

\section*{B.5 Proof for Prop. 3.6: Identifiability in Thm. 3.5}

**Proposition 3.6.** Denote $P_J := \Pr(y, x_S | do(X_M = J(pa(x_M))))$, and $f_{S,-} := \mathcal{E}[Y | x_{S,-}, do(x_M)]$. Under assumptions 2.1 and 2.2, we have $PA(X_M), P_J,$ and $f_{S,-}$ are identifiable.

**Proof.** To generate data distributed as $P_J$, we need to use $J(pa(x_M))$ to regenerate $X_M$, then regenerate $De(X_M)$ with structural equations. To estimate $f_{S,-}$, we need to intervene on $X_M$, then regenerate $De(X_M)$ with structural equations. So, it’s suffice to show $X_M, De(X_M), PA(X_M), PA(De(X_M))$ are identifiable. Identification of $X_M$ has been shown in Alg. \[3\] We first show the identification of $De(X_M)$.

**Algorithm 4** Detection of $De(X_M) \cup X_M$

1: Start with $A = B = X_M$ and visited($X_i$) = FALSE
2: while $B \neq \emptyset$ do
3: for $X_i \in B$ do
4: for $X_j \in Adj(X_i)$ do
5: if $X_i \notin X_M$ and $X_i \perp E|C_{e,x_i} \cup \{X_j\} \setminus D_{X_i,e}$ then
6: $A = A \cup \{X_i\}$
7: if visited($X_i$) = FALSE then
8: $B = B \cup \{X_i\}$
9: end if
10: end if
11: if $X_i \in X_M$ and $\hat{\Delta}_{X_j \rightarrow X_i} < \hat{\Delta}_{X_i \rightarrow X_j}$ then
12: $A = A \cup \{X_i\}$
13: if visited($X_i$) = FALSE then
14: $B = B \cup \{X_i\}$
15: end if
16: end if
17: end for
18: Let $B = B \setminus \{X_j\}$
19: end for
20: end while

• Line 5 to 10: this case is the same as Algorithm \[3\] which is based on (i) the structure $E \rightarrow \cdots \rightarrow X_j \rightarrow X_i$ and (ii) $X_i$ and $E$ are not adjacent.

• Line 11 to 19: In this case, we identify the direction between $X_i$ and $X_j$ by the “Independent Causal Mechanism (ICM) Principle” following [21], where $\hat{\Delta}_{X_j \rightarrow X_i}$ and $\hat{\Delta}_{X_i \rightarrow X_j}$ are the estimated HSIC (see Eq. 17 in [21] for the detailed formulation of $\hat{\Delta}$).

The ICM principle means that “the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.”. That is, the changes of $P(X_i|PA(X_i))$ does not influence the other mechanisms $P(X_j|PA(X_j))$ for $j \neq i$. The ICM principle is implied in the definition of “structural causal model” in [14], where each structural equation represents an autonomous physical mechanism.
We then show the identification of \( \text{PA}(X_i) \), \( \text{PA}(\text{De}(X_M)) \).

Algorithm 5 \( \text{PA}(X_i) \) for \( X_i \in X_M \cup \text{De}(X_M) \).

1: for \( X_j \in X_M \cup \text{De}(X_M) \) do
2: for \( X_i \in \text{Adj}(X_j) \) do
3: if \( X_i \notin X_M \) then
4: \( X_i \in \text{PA}(X_i) \) if \( X_i \notin E[\{C_{e,x_i} \setminus D_{e,x_i} \cup \{X_i\}\}] \)
5: else if \( X_j \in X_M \) and \( X_i \notin X_m \) then
6: \( X_i \in \text{PA}(X_i) \) when \( \Delta_{X_i \rightarrow X_j} < \Delta_{X_i \rightarrow X_j} \).
7: else if \( X_j \in X_M \) and \( X_i \notin X_m \) then
8: \( X_i \in \text{PA}(X_i) \) when \( E \notin X_i | C_{e,x_i} \cup \{X_j\} \)
9: end if
10: end for
11: end for

- Line 4: this rule is based on the structure \( E \rightarrow \cdots \rightarrow X_j - X_i \) and \( \{X_i, E\} \) are not adjacent.
- Line 6: this rule is based on the HSIC criterion in [21].
- Line 5: this rule is based on the structure \( E \rightarrow X_i - X_j \) and \( \{E, X_j\} \) are not adjacent.

\( \square \)

C Appendix for Sec. 3.3: Local Causal Discovery and Empirical Estimation

C.1 Details of Local Causal Discovery

In this section, we summarize the identification of \( X_M^0 \), \( \text{De}(X_M^0) \), \( X_M \), \( \text{De}(X_M) \), \( \{\text{PA}(X_i) \}_i \in X_M \cup \text{De}(X_M) \) \( \text{Blanket}(Y) \), \( \text{PC}(Y) := \text{PA}(Y) \cup \text{Ch}(Y) \).

The identification of \( X_M^0 \cup \text{De}(X_M^0) \) is in Alg. 3. To distinguish \( X_M^0 \) and \( \text{De}(X_M^0) \), it suffices to identify the direction of \( X_i - X_j \) in the case when both \( X_i \) and \( X_j \) are in \( X_M^0 \), which can be accomplished by comparing \( \Delta_{X_i \rightarrow X_j} \) and \( \Delta_{X_j \rightarrow X_i} \) (see [21] for details). However, it should be noted that distinguishing \( X_M^0 \) and \( \text{De}(X_M^0) \) for the estimation of \( h(S_-, J) \) and \( f_{S_-} \) is unnecessary. The identification of \( X_M \) and \( \text{PC}(Y) \) is in Alg. 2, where \( \text{PC}(Y) \) can be obtained from the undirected skeleton. \( \text{Blanket}(Y) \) can be identified by [43]. The identification of \( X_M \cup \text{De}(X_M) \) is in Alg. 5, and we can distinguish \( X_M \) from \( \text{De}(X_M) \) using the way as in \( \{X_M^0, \text{De}(X_M^0)\} \). The parents \( \{\text{PA}(X_i) | X_i \in X_M \cup \text{De}(X_M)\} \) can be identified by Alg. 5.

C.2 Details of Estimation of \( f_{S_-} \)

To estimate \( f_{S_-} \), we adopt soft-intervention to replace \( P^r(\mathbf{X}_M | \text{PA}(\mathbf{X}_M)) \) with \( P(\mathbf{X}_M) \) and hence define \( p^r(x, y) = p(y | pa(y)) \Pi_{i \in S} p(x_i | pa(x_i)) p(\mathbf{X}_M) \). Then we have \( f_{S_-} = \mathbb{E} p^r[Y | x_{S_-}, \mathbf{X}_M] \).

To generate data from \( P^r \), we first permute \( \mathbf{X}_M \) in a sample-wise manner to generate data from \( P(\mathbf{X}_M) \). We then regenerate data for \( \mathbf{X}_M \)’s descendants in the intervened graph via estimating structural equations \(^7\) as summarized in Alg. 6.

Indeed, we only need to regenerate \( \text{De}_{\bar{G}_{X_M^0}}(\mathbf{X}_M) \cap \text{Blanket}(Y) \) since \( p'(y | \text{blanket}(y)) = p'(y | x) \). To maximally reduce the approximation error in regeneration, we consider intervene on another variable set \( \mathbf{X}_{do}^* := \mathbf{X}_M^0 \cup (\text{De}(\mathbf{X}_M^0) \setminus \text{Ch}(Y)) \) and regenerate variables in \( \text{De}_{\bar{G}_{X_{do}^*}}(\mathbf{X}_{do}^*) \). We prove \( \text{De}_{\bar{G}_{X_{do}^*}}(\mathbf{X}_{do}^*) \) is the minimum regeneration set in the following proposition.

Proposition C.1. Denote \( \mathbf{X}_{do}^* := \mathbf{X}_M^0 \cup (\text{De}(\mathbf{X}_M^0) \setminus \text{Ch}(Y)) \). Then:

1. For any admissible set \( \mathbf{X}_{do} \), we have \( \text{De}_{\bar{G}_{X_{do}^*}}(\mathbf{X}_{do}) \cap \text{Blanket}(Y) \supset \text{De}_{\bar{G}_{X_{do}^*}}(\mathbf{X}_{do}^*) \);
2. \( \mathbf{X}_{do}^* \), \( \text{De}_{\bar{G}_{X_{do}^*}}(\mathbf{X}_{do}^*) \), and \( \{\text{PA}(X_i) | X_i \in \text{De}_{\bar{G}_{X_{do}^*}}(\mathbf{X}_{do}^*)\} \) are identifiable.

\(^7\)This can be achieved because \( \mathbf{X}_M^0, \text{De}(\mathbf{X}_M) \), and their parents are identifiable, as shown in Alg. 5.
Algorithm 6 Estimation of $f_{S-}$

**INPUT**: training data $\{(x_{(k)}, y_{(k)})\}_{k=1}^{n}, S- \subset S, X_M, \text{De}(X_M), \text{and \{PA}(X_i)\}_{X_i \in \text{De}_G\overline{X_M}(X_M)}$.

**OUTPUT**: Trained $f_{S-}$.

1. Shuffling $\{(X_M)_i\}_{i=1}^{n}$ by randomizing the indices.
2. For $X_i \in \text{De}_G\overline{X_M}(X_M)$ do
3. Regenerate $\{(x_i)_i\}_{i=1}^{n}$ as $\{g_i(pa(x_i))\}_{i=1}^{n}$.
4. Train $f_{S-}$ over the regenerated samples.

**Proof.** (1) Firstly, we prove that a set of variables $X_{do}$ is admissible means $p_{do}(y|x) = p(y|x_S, do(X_M)) \iff \{X_M \cap \text{Ch}(Y)\} \subset X_{do}$ and $\{X_S \cap \text{Ch}(Y)\} \cap X_{do} = \emptyset$.

Note that

$$p(y|x_S, do(X_M)) = \frac{p(y|pa(y)) \prod_{X_i \in X_S \cap \text{Ch}(Y)} p(x_i|pa(x_i))}{\int p(y|pa(y)) \prod_{X_i \in X_S \cap \text{Ch}(Y)} p(x_i|pa(x_i)) dy},$$

$$p_{do}(y|x) = \frac{p(y|pa(y)) \prod_{X_i \in (X \setminus X_{do}) \cap \text{Ch}(Y)} p(x_i|pa(x_i))}{\int p(y|pa(y)) \prod_{X_i \in (X \setminus X_{do}) \cap \text{Ch}(Y)} p(x_i|pa(x_i)) dy}.$$

It can be seen $p_{do}(y|x) = p(y|x_S, do(X_M)) \iff X \setminus X_{do} \cap \text{Ch}(Y) = X_S \cap \text{Ch}(Y)$, which can be rewritten as

$$\{X_M \cap \text{Ch}(Y) \cap X_{do}^C\} \cup \{X_S \cap \text{Ch}(Y) \cap X_{do}^C\} = X_S \cap \text{Ch}(Y).$$

The above equation holds if and only if $\{X_M \cap \text{Ch}(Y)\} \subset X_{do}$ and $\{X_S \cap \text{Ch}(Y)\} \cap X_{do} = \emptyset$. (2) Secondly, we prove that $X_{do}^*$ is an admissible set and $\text{De}_G\overline{X_{do}}(X_{do}^*) = \text{De}(X_M \cap \text{Ch}(Y)) \cap X_S \cap \text{Ch}(Y)$. To simplify the notations, let $X_0 := X_{do}$ and $X_1 := \text{De}_G\overline{X_{do}}(X_{do})$.

The conditions $\{X_M \cap \text{Ch}(Y)\} \subset X_0$ and $\{X_S \cap \text{Ch}(Y)\} \cap X_0 = \emptyset$ hold by definition.

(2.1) show $\text{De}_G\overline{X_0}(X_0) \subset X_1$

Note $X_0 \subset \{X_M \cap \text{Ch}(Y)\} \cup \{\text{De}(X_M \cap \text{Ch}(Y))\}$, we have $\text{De}(X_0) \subset \text{De}(X_M \cap \text{Ch}(Y))$. Besides, since $\text{De}_G\overline{X_0}(X_0) = \text{De}(X_0) \setminus X_0$, Then

$$\text{De}_G\overline{X_0}(X_0) = \text{De}(X_0) \setminus X_0^C = \text{De}(X_0) \cap \{X_M \cap \text{Ch}(Y)\}^C \cap \{\text{De}(X_M \cap \text{Ch}(Y))\}^C \cap \{X_S \cap \text{Ch}(Y)\}^C.$$

(2.2) show $X_1 \subset \text{De}_G\overline{X_0}(X_0)$

Since $X_M \cap \text{Ch}(Y) \subset X_0$, $\text{De}(X_M \cap \text{Ch}(Y)) \subset \text{De}(X_0)$, so $X_1 \subset \text{De}(X_M \cap \text{Ch}(Y)) \subset \text{De}(X_0)$ and hence $X_1 \setminus X_0 \subset \text{De}(X_0) \setminus X_0$. Besides, note that $X_0 \cap X_1 = \emptyset$ such that $X_1 \setminus X_0 = X_0$ and $\text{De}_G\overline{X_0}(X_0) = \text{De}(X_0) \setminus X_1$, we have $X_1 \subset \text{De}_G\overline{X_0}(X_0)$.

(3) given $X_{do}$ satisfying the two conditions, we have

$$X_M \cap \text{Ch}(Y) \subset X_{do} \Rightarrow \text{De}(X_M \cap \text{Ch}(Y)) \subset \text{De}(X_{do});$$

$$X_{do} \subset \{X_S \cap \text{Ch}(Y)\}^C \Rightarrow \{X_S \cap \text{Ch}(Y)\} \subset X_{do}^C.$$

Therefore,
De(\(X_M \cap \text{Ch}(Y)\)) \cap \{X_S \cap \text{Ch}(Y)\} \subset \text{De}(X_{do}) \cap X_{do}^C.

Thus, \(X_1 \subset \text{De}(X_{do})\) for any \(X_{do}\) satisfying \(X_M \cap \text{Ch}(Y) \subset X_{do}\) and \(X_{do} \cap \{X_S \cap \text{Ch}(Y)\} = \emptyset\).

(4) The identification of \(\{\text{PA}(X_i)\}_{i \in \text{De}(X_{do})}^\prime, X_{do}^\prime\) and \(\text{De}(X_{do})\) can be readily obtained in Sec. C.1.

\[\square\]

C.3 Equivalence of Subsets for Computational Efficiency

We find searching over some subsets of \(S\) is redundant because they are expected to have the same/larger worst-case quadratic loss, as shown by the following proposition:

**Proposition C.2.** Denote \(\overline{X} := \text{Blanket}(Y) \cap X_S, \overline{X} := \text{PA}(Y) \cup (\text{Ch}(Y) \setminus \text{De}(X^0_M)) \cap X_S\). Then:

1. for each \(X_{S_-} \supset \overline{X}\), we have \(f_{X_{S_-}} = f_{\overline{X}}\);
2. for each \(X_{S_-} \subset \overline{X}\), we have \(E_{P_{\overline{X}}}[(Y - f_{X_{S_-}}((x)))^2] \geq E_{P_{\overline{X}}}[(Y - f_{\overline{X}}((x)))^2];\)
3. the \(\overline{X}\) and \(\overline{X}\) are identifiable.

In Prop. C.2, 1. 2. respectively mean that \(f_{X_{S_-}}\) is the same as \(f_{\overline{X}}\) when \(X_{S_-} \supset \overline{X}\) and is no better than \(f_{\overline{X}}\) when \(X_{S_-} \subset \overline{X}\). Together with 3., we can identify \(\overline{X}\) and \(\overline{X}\) and skip those \(f_{X_{S_-}}\) during searching.

**Proof.** 1. Firstly, note the Markov Blanket of \(Y\) is \(\text{PA}(Y) \cup \text{Ch}(Y) \cup \text{PA}(\text{Ch}(Y))\), so \(Y \perp X \setminus \overline{X}\). We have

\[p\left(y | x^M, \text{do} \left( x^M \right) \right) = p_{G_{X^0_M}} \left(y | x^M \right) = p_{G_{X^0_M}} \left(y | \text{pa}(y), \text{ch}(y) \setminus x^0_M, \text{pa} \left( \text{ch}(y) \setminus x^0_M \right) \right)\]

\[= p\left(y | \text{pa}(y), \text{ch}(y) \setminus x^0_M, \text{pa} \left( \text{ch}(y) \setminus x^0_M \right) \right) \text{do} \left( x^M \right) .\]

The 1st equation is by definition. The 2nd equation is because \(p_{G_{X^0_M}} \left(y | x^M \right) = p_{G_{X^0_M}} \left(y | x^0_M \right)\), where \(x^0_M\) denotes the Markov Blanket of \(X\) in \(G_{X^0_M}\), and \(\text{PA}_{G_{X^0_M}}(Y) = \text{PA}(Y) \cup (\text{Ch}(Y) \setminus \text{De}(X^0_M)) \cup \text{PA}(\text{Ch}(Y) \setminus X^0_M)\), and \(\text{PA}_{G_{X^0_M}} \left( \text{Ch}(Y) \setminus X^0_M \right) = \text{PA} \left( \text{Ch}(Y) \setminus X^0_M \right)\). The 3rd equation is by definition.

Let \(\overline{X} = \text{PA}(Y) \cup (\text{Ch}(Y) \setminus X^0_M) \cup (\text{PA}(\text{Ch}(Y) \setminus X^0_M) \setminus X^0_M)\). Then, the 2nd equation states that for any \(X' \subset X, X^0_M\) and \(\overline{X} \subset X', p\left(y | x', \text{do} \left( x^M \right) \right) = p\left(y | x, \text{do} \left( x^M \right) \right)\).

**Proof of 2.** Secondly, we will find a subset \(\overline{X}\) such that for any \(X' \subset X, p\left(y | x', \text{do} \left( x^M \right) \right) = p\left(y | x' \right)\). Let \(\overline{X} = \text{PA}(Y) \cup (\text{Ch}(Y) \setminus \text{De}(X^0_M))\) where \(\text{De}(X^0_M) := \text{De}(X^0_M) \cup X^0_M\), we have \(\overline{X} \perp_{G_{X^0_M}} X^0_M\). Because for any \(X_i \in \overline{X}\), if there is a path between \(X_i\) and \(X_{M, j} \in X^0_M\) in \(G_{X^0_M}\), it is in the form \(X_i \leftarrow \cdots \leftarrow X_{M, j}\). If \(X_i \in \text{PA}(Y)\), the path can not be directed because \(X_{M, j} \in \text{Ch}(Y)\). If \(X_i \in \text{Ch}(Y) \setminus \text{De}(X^0_M)\), the path can not be directed either because \(X_i \notin \text{De}(X^0_M)\) by definition. So \(X_i \perp_{G_{X^0_M}} X^0_M\), and therefore \(\overline{X} \perp_{G_{X^0_M}} X^0_M\). Similarly, we have \(Y \perp_{G_{X^0_M}} X^0_M\), because \(Y \leftarrow \cdots \leftarrow X_{M, j}\) for any \(X_{M, j} \in X^0_M\) would induce a cycle in \(G\).

In summary, we have \(p\left(y, x | \text{do} \left( x^M \right) \right) = p\left(y, x \right)\), and hence for any \(X' \subset X, p\left(y, x' | \text{do} \left( x^M \right) \right) = p\left(y, x' \right)\). Thus, for any \(X' \subset X, we have p\left(y, x' | \text{do} \left( x^M \right) \right) = p\left(y, x' \right)\) and 2. holds.

**Proof of 3.** The set \(\overline{X}\) can be identified by \([43]\). As for \(\overline{X}\), note that \(\text{PA}(Y) \cap X^0_M = \emptyset\). We have \(\overline{X} = \text{Adj}(Y) \setminus \{\text{De}(X^0_M) \cup X^0_M\}\), where \(\text{Adj}(Y)\) can be obtained from the skeleton, \(\text{De}(X^0_M)\) and \(X^0_M\) can be obtained from Alg. 3 and Alg. 5 respectively.

\[\text{For MSE of conditional expectation, we have } X_1 \subset X_2 \Rightarrow E[Y - E[Y | x_1]]^2 \geq E[Y - E[Y | x_2]]^2\]
D Appendix for Sec. 4: Experiment

Implementation of Baselines. Vanilla uses \( P[Y|X] \) to predict \( Y \) and is implemented by the same neural network as \( f_{S_x} \) (which will be introduced later). Other baselines are implemented by the authors' official codes. Specifically, ICP (https://github.com/juangamella/icp), IC (https://github.com/mrojascarulla/causal_transfer_learning); Anchor regression (https://github.com/rothenhauers/anchor-regression); IRM (https://github.com/facebookresearch/InvariantRiskMinimization); HRM (https://github.com/LJSthu/HRM); IB-IRM (https://github.com/ahujak/IB-IRM).

D.1 Simulation

Implementation Details. In all three settings, the structural equation \( x_1 \leftarrow g_1(x_4, y) + u_1 \) is estimated by a one-layer fully-connected neural network (FC), with training iterations set to 1000, the learning rate set to 0.01. SGD is used for optimization. In Setting-1: \( f_{S_x} \) is parameterized by a one-layer FC, with training iterations set to 2000, the learning rate set to 0.001. \( J_\theta \) is parameterized by the same structure, with training iterations set to 2000, the learning rate set to 0.05. In Setting-2: \( f_{S_x} \) is parameterized by a one-layer FC, with training iterations set to 1000, the learning rate set to 0.001. \( J_\theta \) is parameterized by the same structure, with training iterations set to 5000, the learning rate set to 0.05. In Setting-3: \( f_{S_x} \) is parameterized by a two-layers FC with a sigmoid activation function in the hidden layer, with training iterations set to 5000, the learning rate set to 0.01. \( J_\theta \) is parameterized by the same structure, with training iterations set to 2000, the learning rate set to 0.02. The codes are implemented with PyTorch 1.10 and run on a server with an Intel Xeon E5-2699A v4@2.40GHz CPU.

Additional Results. We report the maximal mean square error (max MSE) over the test sets for our method and baselines in Fig. 8, Fig. 9, and Fig. 10. Besides, we in Tab. 2, Tab. 3, and Tab. 4 report the standard deviation of mean square error (std. of MSE) over the test sets as a measure of transferring stability. As we can see, the maximum and standard deviation of MSE of our method are both low. For example, max MSE is 0.0075, and std. of MSE is 0.0006 in setting-2; max MSE is 0.0082, and std. of MSE is 0.0006 in setting-3. This verifies that our method is both robust and stably transferable to distributional shifts. As for the slight improvements over the baseline, it may be due to the simulation settings being simple enough for the vanilla method to only exploit \( X_2 \) for prediction.

D.2 Alzheimer’s Disease Diagnosis

Implementation Details. The whole brain is partitioned into 9 brain regions according to Tab. 5 and Tab. 6. Data normalization (w.r.t. mean and standard deviation) is used. All structural equations are estimated by a two-layers FC with a sigmoid activation function in the hidden layer. For structural equations generating \( X_2, X_3, \) the training takes 5000 iterations, with the learning rate set to 0.1. For those generating \( X_4, X_5, X_6, X_7, \) the training takes 2000 iterations, with the learning rate set to 0.1. \( f_{S_x} \) is parameterized by a two-layers FC with a sigmoid activation function in the hidden layer, with training iterations set to 20000, the learning rate set to 0.1. \( J_\theta \) is parameterized by the same structure, with training iterations set to 10000, the learning rate set to 0.05. SGD is used for optimization.

We pick four domains with more than 40 patients as the training domains and test on the rest three domains. To remove the effect of randomness, we replicate over all the 15 possible train-test splits.

Additional Results. We firstly report std. of MSE over the test sets for our method and baselines in Tab. 7. As we can see, our method outperforms other baselines by a significant margin. This result demonstrates the utility of our method in learning stably transferable predictors. Then, we compare the performance of \( \bar{X} = \{X_2, X_6, X_7, X_{11}\} \) and \( \{X_{S_x} | X_{S_x} \supset \bar{X}\} \) in Fig. 11. As shown, predictors with \( X_{S_x} \supset \bar{X} \) have approximately the maximal MSE as the one with \( \bar{X} \). This observation verifies the correctness of Prop. 2 and shows searching over some subsets can be skipped. Finally, we show the optimization curve of \( h(S_-, J_\theta) \) and max MSE for 100 randomly picked subsets \( \bar{X} \subset S \), in Fig. 12, Fig. 13, and Fig. 14. As we can see, the optimization over \( J_\theta \) is well converged, and the performance of different subsets is consistent with our expectations. This observation again suggests the utility of Thm. 3.5 in finding the optimal predictor.
D.3 Gene Function Prediction

Implementation Details. Data normalization (w.r.t. mean and standard deviation) is used. Structural equations generating $X_2, X_4$ are estimated by a two-layers FC with a sigmoid activation function in the hidden layer, with training iterations set to 5000, the learning rate set to 0.01. $f_{S_-}$ is parameterized by a two-layers FC with a sigmoid activation function in the hidden layer, with training iterations set to 20000, the learning rate set to 0.01. $J_0$ is parameterized by the same structure, with training iterations set to 120000, the learning rate set to 0.05. SGD is used for optimization.

We use the wide-type mice and three kinds of gene knockouts in the training domains. To remove the effect of randomness, we generate 45 replications, with each trial appending 2 out of the remaining 10 gene knockouts to the training domains and testing on the rest 8 gene knockouts.

Additional Results. Firstly, we report std. of MSE over the test sets for our method and baselines in Tab. 8. Similarly, our method outperforms other baselines by a significant margin, which together with the Alzheimer’s disease experiment, shows the utility of our method in learning stably transferable predictors. Then, we compare the performance of $\mathbf{X} = \{X_1, X_3\}$ and $\{X_{S_-}|X_{S_-} \subset \mathbf{X}\}$, $\mathbf{X} = \{X_1, X_2, X_3\}$ and $\{X_{S_-}|X_{S_-} \supset \mathbf{X}\}$ in Fig. 15. As we can see, subsets $\{X_1\}, \{X_3\} \subset \mathbf{X}$ have worse max MSE than $\mathbf{X}$, the subset $\{X_1, X_2, X_3, X_4\} \supset \mathbf{X}$ has similar max MSE as $\mathbf{X}$. This observation verifies the correctness of Prop. 2. Finally, we show the optimization curve of $h(S_-, J_0)$ in Fig. 16. As we can see, the optimization over $J_0$ well converges. Since the order of $h^*$ in different subsets is basically stable after 50000 iterations, we in the main body report the results obtained at 50000 iterations.

Table 2: Mean (over randomization) of std. (over test domains) of MSE on Simulation Setting-1.

| Method   | Vanilla | ICP | IC | IRM | AncReg | HRM | IB-IRM | Ours |
|----------|---------|-----|----|-----|--------|-----|--------|------|
| MSE      | 0.358   | 0.216 | 0.274 | 0.095 | 0.161 | 0.561 | 0.142 | 0.103 |

Table 3: Mean (over randomization) of std. (over test domains) of MSE on Simulation Setting-2.

| Method   | Vanilla | ICP | IC | IRM | AncReg | HRM | IB-IRM | Ours |
|----------|---------|-----|----|-----|--------|-----|--------|------|
| MSE      | .0057   | .0111 | .0051 | .0034 | .0005 | .0820 | .0142 | .0006 |

Figure 8: Comparison with baselines on Simulation Setting-1.
Figure 9: Comparison with baselines on Simulation Setting-2.

Figure 10: Comparison with baselines on Simulation Setting-3.

Table 4: Mean (over randomization) of std. (over test domains) of MSE on Simulation Setting-3.

|     | Vanilla | ICP | IC | IRM | AncReg | HRM | IB-IRM | Ours |
|-----|---------|-----|----|-----|--------|-----|--------|------|
| MSE | .0007   | .0141 | .0007 | **.0004** | .0005 | .1444 | .0005 | .0005 |
Table 5: Automatic Anatomical Labeling (AAL) indices for brain regions.

| Brain Region          | AAL Index | Brain Region          | AAL Index |
|-----------------------|-----------|-----------------------|-----------|
| Precentral_L          | 1         | Precentral_R          | 2         |
| Frontal_Sup_L         | 3         | Frontal_Sup_R         | 4         |
| Frontal_Sup_Orb_L     | 5         | Frontal_Sup_Orb_R     | 6         |
| Frontal_Mid_L         | 7         | Frontal_Mid_R         | 8         |
| Frontal_Mid_Orb_L     | 9         | Frontal_Mid_Orb_R     | 10        |
| Frontal_Inf_Oper_L    | 11        | Frontal_Inf_Oper_R    | 12        |
| Frontal_Inf_Tri_L     | 13        | Frontal_Inf_Tri_R     | 14        |
| Frontal_Inf_Orb_L     | 15        | Frontal_Inf_Orb_R     | 16        |
| Rolandic_Oper_L       | 17        | Rolandic_Oper_R       | 18        |
| Supp_Motor_Area_L     | 19        | Supp_Motor_Area_R     | 20        |
| Olfactory_L           | 21        | Olfactory_R           | 22        |
| Frontal_Sup_Medial_L  | 23        | Frontal_Sup_Medial_R  | 24        |
| Frontal_Mid_Orb_L     | 25        | Frontal_Mid_Orb_R     | 26        |
| Rectus_L              | 27        | Rectus_R              | 28        |
| Insula_L              | 29        | Insula_R              | 30        |
| Cingulum_Ant_L        | 31        | Cingulum_Ant_R        | 32        |
| Cingulum_Mid_L        | 33        | Cingulum_Mid_R        | 34        |
| Cingulum_Post_L       | 35        | Cingulum_Post_R       | 36        |
| Hippocampus_L         | 37        | Hippocampus_R         | 38        |
| ParaHippocampal_L     | 39        | ParaHippocampal_R     | 40        |
| Amygdala_L            | 41        | Amygdala_R            | 42        |
| Calcarine_L           | 43        | Calcarine_R           | 44        |
| Cuneus_L              | 45        | Cuneus_R              | 46        |
| Lingual_L             | 47        | Lingual_R             | 48        |
| Occipital_Sup_L       | 49        | Occipital_Sup_R       | 50        |
| Occipital_Mid_L       | 51        | Occipital_Mid_R       | 52        |
| Occipital_Inf_L       | 53        | Occipital_Inf_R       | 54        |
| Fusiform_L            | 55        | Fusiform_R            | 56        |
| Postcentral_L         | 57        | Postcentral_R         | 58        |
| Parietal_Sup_L        | 59        | Parietal_Sup_R        | 60        |
| Parietal_Inf_L        | 61        | Parietal_Inf_R        | 62        |
| SupraMarginal_L       | 63        | SupraMarginal_R       | 64        |
| Angular_L             | 65        | Angular_R             | 66        |
| Precuneus_L           | 67        | Precuneus_R           | 68        |
| Paracentral_Lobule_L  | 69        | Paracentral_Lobule_R  | 70        |
| Caudate_L             | 71        | Caudate_R             | 72        |
| Putamen_L             | 73        | Putamen_R             | 74        |
| Pallidum_L            | 75        | Pallidum_R            | 76        |
| Thalamus_L            | 77        | Thalamus_R            | 78        |
| Heschl_L              | 79        | Heschl_R              | 80        |
| Temporal_Sup_L        | 81        | Temporal_Sup_R        | 82        |
| Temporal_Pole_Sup_L   | 83        | Temporal_Pole_Sup_R   | 84        |
| Temporal_Mid_L        | 85        | Temporal_Mid_R        | 86        |
| Temporal_Pole_Mid_L   | 87        | Temporal_Pole_Mid_R   | 88        |
| Temporal_Inf_L        | 89        | Temporal_Inf_R        | 90        |
| Cerebelum_Crus1_L     | 91        | Cerebelum_Crus1_R     | 92        |
| Cerebelum_Crus2_L     | 93        | Cerebelum_Crus2_R     | 94        |
| Cerebelum_3_L         | 95        | Cerebelum_3_R         | 96        |
| Cerebelum_4_5_L       | 97        | Cerebelum_4_5_R       | 98        |
| Cerebelum_6_L         | 99        | Cerebelum_6_R         | 100       |
| Cerebelum_7b_L        | 101       | Cerebelum_7b_R        | 102       |
| Cerebelum_8_L         | 103       | Cerebelum_8_R         | 104       |
| Cerebelum_9_L         | 105       | Cerebelum_9_R         | 106       |
| Cerebelum_10_L        | 107       | Cerebelum_10_R        | 108       |
| Vermis_L_2            | 109       | Vermis_3              | 110       |
| Vermis_4_5            | 111       | Vermis_6              | 112       |
| Vermis_7              | 113       | Vermis_8              | 114       |
| Vermis_9              | 115       | Vermis_10             | 116       |
Table 6: Brain regions partition.

| Brain Region          | AAL Index                  |
|-----------------------|----------------------------|
| Frontal lobe ($X_1$)  | 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 |
| Medial temporal lobe  | $X_2$: 85, 86, 87, 88      |
| Parietal lobe ($X_3$) | 59, 60, 61, 62             |
| Occipital lobe ($X_4$)| 49, 50, 51, 52, 53, 54    |
| Cingulum ($X_5$)      | 31, 32, 33, 34, 35, 36    |
| Insula ($X_6$)        | 29, 30                     |
| Amygdala ($X_7$)      | 41, 42                     |
| Hippocampus ($X_8$)   | 37, 38                     |
| Pallidum ($X_9$)      | 75, 76                     |

Table 7: Mean (over randomization) of std. (over test domains) of MSE on ADNI dataset.

|               | Vanilla | ICP | IC | IRM | AncReg | HRM | IB-IRM | Ours |
|---------------|---------|-----|----|-----|--------|-----|--------|------|
| Mean Test MSE | 0.267   | 0.270 | 0.252 | 0.166 | 0.161 | 0.294 | 0.244 | **0.030** |

Figure 11: Comparison among $\bar{X} = \{X_2, X_6, X_7, X_{11}\}$ and $X_{S_\subset} \supseteq \bar{X}$ on ADNI dataset.

Table 8: Mean (over randomization) of std (over test domains) of MSE on IMPC gene dataset.

|               | Vanilla | ICP | IC | IRM | AncReg | HRM | IB-IRM | Ours |
|---------------|---------|-----|----|-----|--------|-----|--------|------|
| Mean Test MSE | 0.257   | 0.274 | 0.302 | 0.319 | 0.275 | 0.278 | 0.259 | **0.018** |
Figure 12: (a,c,e,g) Optimization curves of $h(S_-, J_\theta)$, (b,d,f,h) worst-case errors of $f_{S_-}$ on ADNI dataset.
Figure 13: (a,c,e) Optimization curves of $h(S_-, J_\theta)$, (b,d,f) worst-case errors of $f_{S_-}$ on ADNI dataset.
Figure 14: (a,c,e) Optimization curves of $h(S_-, J_0)$, (b,d,f) worst-case errors of $f_{S_-}$ on ADNI dataset.
Figure 15: Comparison among $X_{S_-} \subset \overline{X} = \{X_1, X_3\}$ and $X_{S_-} \supset \overline{X} = \{X_1, X_2, X_3\}$ on IMPC gene dataset.

Figure 16: (a) Optimization curves of $h(S_-, J_\theta)$, (b) worst-case errors of $f_{S_-}$ on IMPC gene dataset.
References

[1] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo et al., “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8340–8349.

[2] J. Li, B. Wu, X. Sun, and Y. Wang, “Causal hidden markov model for time series disease forecasting,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 105–12 114.

[3] J. Müller, R. Schmier, L. Ardizzone, C. Rother, and U. Köthe, “Learning robust models using the principle of independent causal mechanisms,” arXiv preprint arXiv:2010.07167, 2020.

[4] X. Sun, B. Wu, X. Zheng, C. Liu, W. Chen, T. Qin, and T.-Y. Liu, “Recovering latent causal factor for generalization to distributional shifts,” Advances in Neural Information Processing Systems, vol. 34, 2021.

[5] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, “Towards a theoretical framework of out-of-distribution generalization,” arXiv preprint arXiv:2106.04496, 2021.

[6] A. Subbaswamy and S. Saria, “I-spec: An end-to-end framework for learning transportable, shift-stable models,” arXiv preprint arXiv:2002.08948, 2020.

[7] J. Peters, P. Bühlmann, and N. Meinshausen, “Causal inference by using invariant prediction: identification and confidence intervals,” Journal of the Royal Statistical Society. Series B (Statistical Methodology), pp. 947–1012, 2016.

[8] P. Bühlmann, “Invariance, causality and robustness,” Statistical Science, vol. 35, no. 3, pp. 404–426, 2020.

[9] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant risk minimization,” arXiv preprint arXiv:1907.02893, 2019.

[10] J. Liu, Z. Hu, P. Cui, B. Li, and Z. Shen, “Heterogeneous risk minimization,” in International Conference on Machine Learning. PMLR, 2021, pp. 6804–6814.

[11] K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish, “Invariance principle meets information bottleneck for out-of-distribution generalization,” Advances in Neural Information Processing Systems, vol. 34, 2021.

[12] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters, “Anchor regression: Heterogeneous data meet causality,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 83, no. 2, pp. 215–246, 2021.

[13] A. Subbaswamy, P. Schulam, and S. Saria, “Preventing failures due to dataset shift: Learning predictive models that transport,” in The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019, pp. 3118–3127.

[14] J. Pearl, Causality. Cambridge University Press, 2009.

[15] P. Vemuri and C. R. Jack, “Role of structural MRI in alzheimer’s disease,” Alzheimer’s research & therapy, vol. 2, no. 4, pp. 1–10, 2010.

[16] G. S. Bloom, “Amyloid-β and Tau: The Trigger and Bullet in Alzheimer Disease Pathogenesis,” JAMA Neurology, vol. 71, no. 4, pp. 505–508, 04 2014.

[17] R. A. Armstrong. “What causes alzheimer’s disease?” Folia Neuropathologica, vol. 51, no. 3, pp. 169–188, 2013.

[18] J. W. Vogel et al., “Spread of pathological tau proteins through communicating neurons in human alzheimer’s disease,” Nature Communications, vol. 11, no. 1, p. 2612, May 2020.

[19] E. Cavedo, M. Pievani, M. Boccardi, S. Galluzzi, M. Bocchetta, M. Bonetti, P. M. Thompson, and G. B. Frisoni, “Medial temporal atrophy in early and late-onset alzheimer’s disease,” Neurobiology of aging, vol. 35, no. 9, pp. 2004–2012, 2014.

[20] C. M. Fiford, G. R. Ridgway, D. M. Cash, M. Modat, J. Nicholas, E. N. Manning, I. B. Malone, G. J. Biessels, S. Ourselin, O. T. Carmichael et al., “Patterns of progressive atrophy vary with age in alzheimer’s disease patients,” Neurobiology of aging, vol. 63, pp. 22–32, 2018.

[21] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, “Causal discovery from heterogeneous/nonstationary data,” Journal of Machine Learning Research, vol. 21, no. 89, pp. 1–53, 2020.

[22] J. Pearl, “Causal diagrams for empirical research,” Biometrika, vol. 82, no. 4, pp. 669–688, 1995.

[23] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, Causation, Prediction, and Search. MIT press, 2000.

[24] F. Eberhardt and R. Scheines, “Interventions and causal inference,” Philosophy of science, vol. 74, no. 5, pp. 981–995, 2007.
[25] J. Tian and J. Pearl, “Causal discovery from changes,” in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 2001, pp. 512–521.

[26] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell, “Representation learning via invariant causal mechanisms,” in *International Conference on Learning Representations (ICLR)*, 2021.

[27] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[28] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, “Invariant models for causal transfer learning,” *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1309–1342, 2018.

[29] J. Ashburner, “A fast diffeomorphic image registration algorithm,” *Neuroimage*, vol. 38, no. 1, pp. 95–113, 2007.

[30] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, “Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single-subject brain,” *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[31] A. L. Young et al., *Nature Communications*, vol. 9, no. 1, p. 4273, Oct 2018.

[32] A. M. Mayo, “Use of the functional activities questionnaire in older adults with dementia,” *Hartford Inst Geriatr Nurs*, vol. 13, p. 2, 2016.

[33] J. Barnes, J. W. Bartlett, L. A. van de Pol, C. T. Loy, R. I. Scahiill, C. Frost, P. Thompson, and N. C. Fox, “A meta-analysis of hippocampal atrophy rates in alzheimer’s disease,” *Neurobiology of aging*, vol. 30, no. 11, pp. 1711–1723, 2009.

[34] R. Duara, D. Loewenstein, E. Potter, J. Appel, M. Greig, R. Urs, Q. Shen, A. Raj, B. Small, W. Barker et al., “Medial temporal lobe atrophy on mri scans and the diagnosis of alzheimer disease,” *Neurology*, vol. 71, no. 24, pp. 1986–1992, 2008.

[35] H. I. Jacobs, M. P. Van Boxtel, J. Jolles, F. R. Verhey, and H. B. Uylings, “Parietal cortex matters in alzheimer’s disease: an overview of structural, functional and metabolic findings,” *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 1, pp. 297–309, 2012.

[36] Y. Zhang, N. Schuff, G.-H. Juhng, W. Bayne, S. Mori, L. Schad, S. Mueller, A.-T. Du, J. Kramer, K. Yaffe et al., “Diffusion tensor imaging of cingulum fibers in mild cognitive impairment and alzheimer disease,” *Neurology*, vol. 68, no. 1, pp. 13–19, 2007.

[37] V. Muñoz-Fuentes, P. Cacheiro, T. F. Meehan, J. A. Aguilar-Pimentel, S. D. Brown, A. M. Flenniken, P. Flicek, A. Galli, H. H. Mashhadi, M. Hrabˇe de Angelis et al., “The international mouse phenotyping consortium (impc): a functional catalogue of the mammalian genome that informs conservation,” *Conservation Genetics*, vol. 19, no. 4, pp. 995–1005, 2018.

[38] CRM workshop, “Causal inference challenge,” Website, 2016, http://www.crm.umontreal.ca/2016/Genetics16/competition_e.php.

[39] S. Magliacane, T. Van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, “Domain adaptation by using causal inference to predict invariant conditional distributions,” *Advances in neural information processing systems*, vol. 31, 2018.

[40] M. W. Carr, S. J. Roth, E. Luther, S. S. Rose, and T. A. Springer, “Monocyte chemoattractant protein 1 acts as a t-lymphocyte chemoattractant,” *Proceedings of the National Academy of Sciences*, vol. 91, no. 9, pp. 3652–3656, 1994.

[41] L. E. Lee, J. Y. Pyo, S. S. Ahn, J. J. Song, Y.-B. Park, and S.-W. Lee, “Clinical significance of large unstained cell count in estimating the current activity of antineutrophil cytoplasmic antibody-associated vasculitis,” *International Journal of Clinical Practice*, vol. 75, no. 10, p. e14512, 2021.

[42] E. Goetzl, D. Foster, and D. Payan, “A basophil-activating factor from human t lymphocytes.” *Immunology*, vol. 53, no. 2, p. 227, 1984.

[43] C. F. Aliferis, I. Tsamardinos, and A. Statnikov, “Hiton: a novel markov blanket algorithm for optimal variable selection,” in *AMIA annual symposium proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 21.