Causal Domain Adaptation: An Information Bottleneck Approach

Extended Abstract

Mohammad Ali Javidian
Appalachian State University
Boone, United States
javidianma@appstate.edu

ABSTRACT

We address a common causal domain adaptation scenario in which the target variable is observed in the source domain but completely unobserved in the target domain. Our goal is to impute the missing target values in the target domain using the remaining observed variables, under various shifts. We cast this problem as learning a compact representation that is stable across mechanisms: retaining information needed to predict the target while filtering out spurious variation. For linear Gaussian causal models, we derive a closed-form Gaussian Information Bottleneck solution that reduces to a canonical-correlation-style projection and can incorporate DAG-aware structure when desired. For nonlinear or non-Gaussian settings, we propose a Variational Information Bottleneck encoder–predictor that scales to high-dimensional data, can be trained on the source domain, and deployed zero-shot in the target domain. Experiments on synthetic and real datasets show that our method consistently produces accurate imputations, enabling practical deployment in high-dimensional causal models and providing a unified, lightweight toolkit for causal domain adaptation.

KEYWORDS

Domain Adaptation; Causality; Information Bottleneck; Directed Acyclic Graphs

ACM Reference Format:
Mohammad Ali Javidian. 2026. Causal Domain Adaptation: An Information Bottleneck Approach: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/UYYQ1151

1 INTRODUCTION

Modern predictive models are seldom deployed in the same environment where they are trained. Changes in demographics, sensing hardware, or operational policies can shift the joint distribution and degrade performance. This motivates *domain adaptation*: transferring knowledge from a labeled *source* domain to an unlabeled or partially labeled *target* domain with a different distribution. When transfer fails, the cause is typically not mere randomness, but a change in the underlying data-generating *mechanisms*.

Causality as a stabilizer. Causal structure offers a principled way to distinguish robust signal from spurious correlation by reasoning about which mechanisms should remain invariant across environments. For instance, *selection diagrams* encode population differences and enable *transportability* analysis via do-calculus [2–5, 11]. *Invariant causal prediction* (ICP) identifies variable sets whose residual behavior is stable across environments [12–14], while *graph surgery* removes unstable components to prevent distribution shifts from propagating into predictions [17, 18]. Related approaches cast adaptation as *graph pruning*, selecting predictors that induce invariant conditionals [6, 10, 15]. Despite strong guarantees, many causal transfer methods rely on causal effect estimation or counterfactual reasoning, can be conservative (often reducing variance at the expense of transfer bias), and may face scalability challenges. More recent developments in *invariant risk minimization* [1] and deep generative methods for causal representation learning [7, 9] aim to improve robustness under broader and potentially unseen shifts.

Our view: compact, mechanism-stable representations. We view adaptation through the lens of learning a *mechanism-stable* summary: a low-dimensional representation that preserves information needed to predict the target while attenuating nuisance variation that is unlikely to transfer. Concretely, we develop a *DAG-aware information bottleneck* that compresses the observed variables X into a bottleneck representation U, explicitly trading compression against predictive sufficiency for T. When causal structure is available, we leverage DAG information (e.g., parents or Markov blanket) to bias the encoder toward stable mechanisms and away from unstable pathways. The result is a compact, causally informed representation that is stable across domains, with formal guarantees in the Gaussian setting and distribution-free motivation in nonlinear or non-Gaussian regimes.

2 PROPOSED METHODOLOGY

Our theory establishes: (i) in the Gaussian case, all optimal bottleneck directions can be expressed using Markov-blanket variables, and the informative spectrum matches the global one; (ii) the population predictor "expected target given the Markov blanket" is identifiable from source data and preserves risk in the target domain under blanket invariance; and (iii) with finite samples, the learned bottleneck predictor concentrates around this population conditional and is robust to shifts outside the blanket. Together, these results justify the Markov blanket as a structurally minimal, transfer-stable interface for bottleneck learning, yielding algorithms that are both efficient (MB–GIB) and expressive (MB–VIB).

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/UYYQ1151

2.1 MB–GIB: Closed-Form Solution and Lossless Restriction (Gaussian)

In the linear–Gaussian case, the information bottleneck has an exact solution with a simple interpretation. After standard normalization, the goal is to find a few linear combinations of the allowed inputs that are most strongly related to the target. This is exactly what CCA provides: it ranks directions by how informative they are about the target, and we keep only the leading ones. The encoder is this linear projection; the decoder is the optimal linear predictor of the target from the resulting low-dimensional code. The bottleneck parameter controls how many directions are retained by discarding weak ones. With inputs restricted to the Markov blanket, the retained informative directions match those obtained using all non-target variables, which makes the restriction lossless in this setting.

2.2 MB–VIB: Nonlinear, Non-Gaussian Bottleneck (Practical Variant)

For nonlinear or non-Gaussian data, we use a variational bottleneck restricted to the Markov blanket. A neural encoder outputs a distribution over latent codes, and a probabilistic decoder predicts a distribution for the target given the code. Training on labeled source data balances accurate prediction with an explicit compression penalty that discourages the code from carrying unnecessary input information, thereby suppressing nuisance variation that is brittle under shift. The Markov-blanket restriction further prevents reliance on non-blanket pathways and aligns learning with the mechanism assumed to remain stable across domains.

3 EXPERIMENTAL RESULTS

We evaluate our DAG-aware Information Bottleneck approach for imputation under domain shift in three progressively more realistic settings: (i) a controlled *7-node* structural equation model with a known Markov blanket (useful for transparent diagnostics), (ii) the *64-node MAGIC–IRRI* Gaussian Bayesian network (a benchmark-scale synthetic gene network; Scutari, 2016), and (iii) a real single-cell signaling dataset from *Sachs et al.* [16], which provides a demanding biological transfer test in which the target variable is unobserved at deployment. Across all settings, we compare **MB–GIB** and **MB–VIB** against standard baselines: a Bayesian network (BN), a feedforward deep neural network (DNN), and an IIB-style variant [8]. We consider both covariate shift and generalized target shift, and report MAE, RMSE, and R^2 (mean ± standard error over random seeds), along with runtime. All experiments were run on a Windows workstation with a 12th Gen Intel(R) Core(TM) i9-12900H (2.50 GHz). Code and scripts are available at: https://github.com/majavid/CDA_IB. Due to space limitations, we report results only for the *64-node MAGIC–IRRI* network here; full theoretical and experimental results appear in the arXiv version: https://arxiv.org/abs/2601.04361.

3.1 Simulated Experiments: MAGIC–IRRI Gene Network

We conduct a large-scale stress test on the 64-node *MAGIC–IRRI* Gaussian Bayesian network. To mimic strong perturbations of experimental conditions in a multi-trait genetic setting, we induce substantial marginal shifts in three continuous covariates. Specifically, we change the distribution of **G4156** from $\mathcal{N}(0.7636,\ 0.9721^2)$ to $\mathcal{N}(1.5,\ 2.0^2)$, **G4573** from $\mathcal{N}(0.1196,\ 0.4744^2)$ to $\mathcal{N}(1.0,\ 1.0^2)$, and **G1533** from $\mathcal{N}(0.8004,\ 0.9803^2)$ to $\mathcal{N}(0,\ 3.0^2)$. After applying these shifts in the target domain, we hide the trait of interest HT and evaluate how well each method imputes it from the remaining observed variables. We compare: (1) a source-trained Bayesian network (BN), (2) our bottleneck methods (MB–GIB and MB–VIB), (3) the IIB-style objective, and (4) a pure feedforward DNN. All models are trained on the source domain and deployed zero-shot to the shifted target domain.

Results and analysis. Table 1 shows a clear advantage for **MB–GIB** under simultaneous, large marginal shifts. It achieves the best overall imputation accuracy with MAE = 5.57, RMSE = 7.01, and $R^2 = 0.567$. **MB–VIB** is competitive but less accurate in this setting (MAE = 7.08, RMSE = 10.02, $R^2 = 0.121$), and the **IIB-style** variant performs similarly to MB–VIB. The BN baseline degrades substantially under this out-of-distribution shift (negative $R^2 = -0.096$), suggesting that purely generative propagation without explicit compression is sensitive to large marginal changes. The **DNN** performs worst by a wide margin (MAE = 14.45, RMSE = 17.79, $R^2 = -1.77$), consistent with overfitting to the source support and poor extrapolation to shifted inputs. Overall, these results support the view that *explicit compression with moment control* (as in MB–GIB) offers a stronger bias–variance tradeoff for zero-shot deployment under severe distribution shift in high-dimensional causal networks.

Table 1: Imputation performance on the MAGIC-IRRI DAG under multiple large-shift interventions.

Method	MAE	RMSE	R^2
Bayesian Network	9.3827	11.1872	−0.0957
MB-GIB	**5.5706**	**7.0083**	**0.5670**
MB-VIB	7.0837	10.0190	0.1211
IIB-style	7.8219	10.0456	0.1165
Pure DNN	14.4523	17.7908	-1.7711

3.2 Extensions and Limitations

Extensions. The framework can use *learned or partial* MB(T), switch to *parents-only* inputs when downstream paths are unstable, share a decoder across *multiple environments* with environment-specific encoders, leverage a small labeled target subset for *semi-supervised* fine-tuning, and provide *uncertainty* via VIB predictive uncertainty or MB–GIB's Gaussian posterior.

Limitations. Results depend on *MB-invariance*; if the target-to-blanket mechanism changes (e.g., new parents or altered noise), zero-shot transfer can fail. Mis-specified blankets and severe support shift within MB(T) also hurt performance. MB–VIB requires careful capacity and hyperparameter tuning; practical checks include monitoring target-domain residuals or predictive likelihoods, and mitigating issues by using parents-only inputs, increasing β (more compression), or adding a small amount of target labels. Finite-sample/optimization issues and latent confounding can further violate assumptions.

REFERENCES

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant Risk Minimization. *arXiv preprint arXiv:1907.02893* (2019).

[2] Elias Bareinboim and Judea Pearl. 2011. Transportability of Causal Effects: Completeness Results. *UAI* (2011).

[3] Elias Bareinboim and Judea Pearl. 2012. Causal Transportability with Limited Experiments. *AAAI* (2012).

[4] Elias Bareinboim and Judea Pearl. 2014. External Validity: From Do-Calculus to Transportability Across Populations. In *Journal of Causal Inference*.

[5] Rodrigo Correa and Elias Bareinboim. 2019. Transportability of Experimental Results: A Formal Approach. In *IJCAI*.

[6] William M. Kouw and Marco Loog. 2019. A Review of Domain Adaptation without Target Labels. *IEEE Trans. Pattern Analysis and Machine Intelligence* (2019).

[7] David Krueger, Vicent Cabannes, Ishmael Belghazi, Ben Poole, et al. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *ICML*.

[8] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. 2022. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7399–7407.

[9] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8046–8056.

[10] Luca Magliacane, Thomas Claassen, Karsten Borgwardt, and F. Dániel. 2018. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. In *NeurIPS*.

[11] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[12] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal Inference Using Invariant Prediction: Identification and Confidence Intervals. In *JRSS-B*.

[13] Nina Pfister, Peter Bühlmann, and Jonas Peters. 2019. Invariant Causal Prediction for Sequential Data: What If the Markov Assumption Fails?. In *AISTATS*.

[14] Nina Pfister, Peter Bühlmann, and Jonas Peters. 2019. Stabilizing Causal Structure Learning via Invariant Conditional Distributions. *Biometrika* (2019).

[15] Marta Rojas-Carulla, Bernhard Schölkopf, Richard E. Turner, and Jonas Peters. 2018. Invariant Models for Causal Transfer Learning. In *JMLR Workshop and Conference Proceedings*, Vol. 63. 752–760.

[16] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 5721 (2005), 523–529.

[17] Amit Subbaswamy and Suchi Saria. 2018. Preventing Failure in Exogenous Distribution Shift: A Causal Abstraction Approach. In *NeurIPS*.

[18] Amit Subbaswamy and Suchi Saria. 2019. Preventing Failure Under Distribution Shift Using Risk Extrapolation. In *ICML*.