On Domain Generalization Datasets as Proxy Benchmarks for Causal Representation Learning

Olawale Salaudeen*
Department of Computer Science
University of Illinois at Urbana-Champaign
oes2@illinois.edu

Nicole Chiou
Department of Computer Science
Stanford University
nicchiou@stanford.edu

Sanmi Koyejo
Department of Computer Science
Stanford University
sanmi@cs.stanford

Abstract

Benchmarking causal representation learning for real-world high-dimensional settings where most relevant causal variables are not directly observed remains a challenge. Notably, one promise of causal representations is their robustness to interventions, enabling models to generalize effectively under distribution shifts—domain generalization. Given this connection, we ask to what extent domain generalization performance can serve as a reliable proxy task/benchmark for causal representation learning in such complex datasets. In this work, we provide theoretical evidence that one condition that identifies reliable domain generalization tasks that are reliable proxies is when non-causal correlations with labels/outcomes In-Distribution are reversed or have sufficiently reduced signal-to-noise ratio Out-Of-Distribution. Additionally, we demonstrate that benchmarks with this reversal do not have strong positive correlations between in-distribution (ID) and out-of-distribution (OOD) accuracy, commonly called "accuracy on the line." Finally, we characterize our derived conditions on state-of-the-art domain generalization benchmarks to identify effective proxy tasks for causal representation learning.

1 Introduction

Many of machine learning's successes have been driven by large-scale pattern recognition on training data assumed to be independently and identically distributed (i.i.d.) with the testing or deployment data. However, this i.i.d. assumption is often unrealistic, as model developers typically have limited control over the distribution of real-world data that the model will eventually encounter. For example, in computer vision, real-world scenarios may introduce various interventions to the data, such as camera blurring, noise, compression artifacts, or shifts in brightness, contrast, and background [23, 7]. These interventions can be particularly problematic when they alter spurious statistical relationships that the model has learned to exploit as shortcuts for its predictions [16, 25, 36].

Models that incorporate or learn structural knowledge of the domains they are applied to have been shown to be more efficient and generalize better across different settings [44, 56, 17, 18, 8, 6, 42, 70, 68]. An example of such a structure is the principle of *independent causal mechanisms* [21, 3, 24, 45, 58, 26, 46, 59], which posits that the generative process of a system's variables consists of autonomous components, or mechanisms, that operate independently and do not inform one another.

*Corresponding author. Email: oes2@illinois.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

This implies that the conditional distribution of each variable, given its causes (mechanisms), is independent of the other variables and mechanisms [47]. Learning causal representations is an active area of research [59]. Datasets for causal representation learning are primarily (semi)parametric where (some) causal variables are known and potentially intervenable [65, 31, 32, 1, 35]. Then, success is assessed by how well learned disentangled representations (mechanisms) explain outcome variance, using R^2 or MCC (Mathew's Correlation Coefficient) [37]. However, the task of causal representation learning with complex datasets with limited knowledge or control over generative mechanisms remains a challenge, especially without requiring most (or at least some) relevant causal variables to be directly observed [38]—we identify that benchmarking causal representation learning in this setting is also challenging.

Causal representation learning is closely tied to domain generalization, which aims to learn representations from multiple observed domains that give predictors whose performance is invariant to new domains (new data distributions). Many works in domain generalization [4, 55, 54, 41, 33, 39, 11, 15]. have been motivated by the principle of independent causal predictors, which aims to identify causal predictors from observational data by searching for feature sets that maintain stable (invariant) predictive accuracy across interventional distributions [46, 22, 4]. Additionally, more recent work motivates learning causal representations from multiple datasets arising from unknown interventions [66].

Thus, one may naively consider domain generalization as a proxy task to benchmark causal representation learning in more complex settings. This work studies when performance on a domain generalization task is informative of the causal representation learning task. Specifically, when benchmarking a set of models, including a disentangled causal model, when does the causal model transfer best out-of-domain, i.e., domain generalization?

1.1 Our Contributions

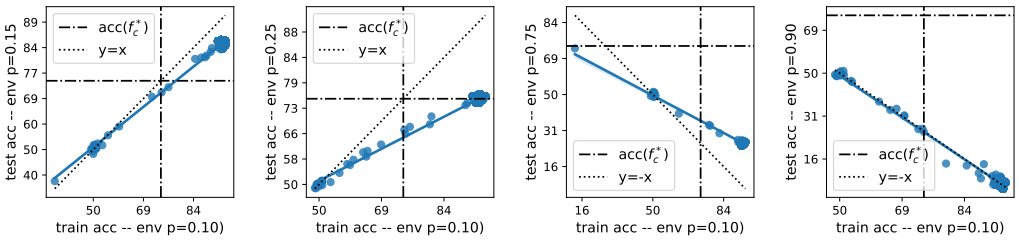
- We show that models without causal representations can often achieve better transfer accuracy than models with causal representations.
- We give conditions to reliably benchmark causal representation learning with domain generalization as a proxy task. These conditions motivate benchmarks with adversarial shifts in (spurious) correlations between non-causal features and labels. We also show that these shifts result in *accuracy on the inverse line*.
- Finally, we empirically analyze and categorize the utility of state-of-the-art real-world domain generalization benchmarks for causal representation learning.
- Our findings apply to benchmarking minimax formulations of domain generalization [51]. Additionally, we find empirically that *accuracy on the line* for ERM models may not imply the same for models from state-of-the-art domain generalization algorithms.

2 Motivation

The ColoredMNIST dataset [4] illustrates the challenge of inferring causal representations from improved domain generalization. ColoredMNIST modifies the grayscale MNIST dataset by adding color as a spurious correlation. Digits are red or green based on binary observed labels of 'digit ≥ 5' with 25% label noise. Color matches observed labels with probability p_e, inducing a *spurious correlation* or *shortcut*. Each domain is defined by a different p_e. Additionally, the observed labels are noisy versions of the true labels, so the color is potentially more correlated with the observed labels than the digit itself. More on the ColoredMNIST generative mechanism is in Appendix B.

For example, consider a training domain where $p_e = P(Y = 1 \mid \text{color} = \text{green}) = 0.9$. A color-based predictor would achieve 90% accuracy in-domain. Under a shift where $q_e = Q(Y = 1 \mid \text{color} = \text{green}) > 0.75$, the color-based model will still outperform the causal model in transfer accuracy. With this example, we aim to emphasize that for a domain generalization benchmark to be informative about causal representation learning, the non-causal correlations from training to test domains must change enough for the causal model to achieve the highest transfer accuracy.

To demonstrate this, we test Empirical Risk Minimizers (ERM) on a ColoredMNIST training domain where $P(Y = 1 \mid \text{color} = \text{green}) = 0.1$, across various test domains with $Q(Y = 1 \mid \text{color} = \text{green}) = q_e$. We train convolutional neural networks with varying hyperparameters and evaluate them on different test domains—more in Appendix A. Figure 1 demonstrates that causal models need



(a) m = 0.79
Pearson R = 0.97

(b) m = 0.50
Pearson R = 0.97

(c) m = -0.52
Pearson R = -0.96

(d) m = -0.98
Pearson R = -0.98

Figure 1: Correlations between model performance In-Distristribution vs. Out-of-Distribution on ColoredMNIST variations. m is the slope of the line, and R is the Pearson correlation coefficient. The axis-parallel dashed lines denote the maximum within-domain accuracy of 75%, and $y = x$ represents invariant performance across training and test (target) domains. Models achieving above 75% accuracy use color as a predictor. Figures 1a and 1b represent shifts where color-based predictors achieve the highest transfer accuracy—above 75% accuracy. Without domain knowledge, one might conclude that the best ERM solution is the most domain-general and, therefore, learns causal representations. However, Figures 1c and 1d show that these models are not causal; some features that improve ID accuracy hurt OOD performance.

not transfer the best OOD. This observation underscores this work's key question: which ID-OOD shifts allow us to infer causal representations from domain generalizability? We formalize and analyze this question theoretically in the next section.

3 Theoretical Analysis

Notation. Let $\mathcal{X}$ denote an input feature space, and let $\mathcal{Y}$ denotes an output space. $\mathcal{X}$ is composed of the union of subspaces $\mathcal{Z}_\text{c}$ and $\mathcal{Z}_\text{e}$, such that $X = [Z_\text{c}, Z_\text{e}] \in \mathcal{X}$. P represents a probability distribution over the triple $Z_\text{c}, Z_\text{e}, Y$, or equivalently, the pair X, Y. The function $\ell(\cdot, \cdot) \to \mathbb{R}$ denotes a loss function, and $\mathcal{R}^e(f)$ defines the expected loss $\mathbb{E}_{P_e}[\ell(Y, f(X))]$ w.r.t. distribution P_e for function $f \in \mathcal{F}$, where $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$. Within $\mathcal{F}$, $\mathcal{F}_\text{c} \subset \mathcal{F}$ comprises functions f_c defined by $f_c(X) = f_c([Z_\text{c}, 0])$. We denote $f_X \in \mathcal{F} \backslash \mathcal{F}_\text{c}$.

Definition 1 (Reliable Domain Generalization Proxy Task for Causal Representation Learning)**.** *Consider a domain generalization task with P_{ID} and P_{OOD} on which models are trained and then tested, respectively. Given a set of trained models $F = \{f_X : f_X \in \mathcal{F} \backslash \mathcal{F}_c\}$ and $f_c^* \in \mathcal{F}_c$,*

$$\max_{f \in F} acc_{OOD}(f_X) < acc_{OOD}(f_c^*), \tag{1}$$

where acc_{OOD} is the accuracy achieved on P_{OOD} and f_c^* is a causal model.

We consider the following structural equation model widely used in the distribution shift literature [67, 50, 49, 43, 54]:

$$\mu(M, \Lambda) = \begin{cases} \eta \sim \text{Bern}(q); C \sim \text{Bern}(p) \\ Y = \begin{cases} C & \text{if } \eta = 1 \\ \neg C & \text{if } \eta = 0 \end{cases} \\ Z_\text{c} \sim \mathcal{N}\big(C \cdot \mu_c, \Sigma_\text{c}\big) \\ Z_\text{e} \sim \mathcal{N}\big(Y \cdot M\mu_e, \Lambda\Sigma_\text{e}\big) \\ X = Z_\text{c} \oplus Z_\text{e}, \end{cases} \tag{2}$$

where η is a noise term, $Z_\text{c} \in \mathbb{R}^m$, $Z_\text{e} \in \mathbb{R}^k$, and $Y \in \{0, 1\}$. We also define interventions (shift-sources), M, Λ that parameterize a domain, where $M \in \mathbb{R}^{k \times k}$ and $\Lambda \succeq 0 \in \mathbb{R}^{k \times k}$. We use this model to be consistent with the literature. While the mechanism $Z_\text{c} \to Y$ is not causal, it is sufficient for our purposes and enables convenient analysis. Specifically, we want to analyze the effect of including features whose correlations with labels shift across domains on transfer accuracy.

Remark 1 (Partially Informative Causal Features (PICF)). *$Z_e \not\perp\!\!\!\perp Y \mid Z_c$. This property describes a setting where spurious correlations are not redundant to causal correlations, e.g., color in ColoredMNIST can be used to improve accuracy than just the digit. Previous work demonstrates settings where not having this property makes causal vs. non-causal solutions indistinguishable via transfer under some conditions [2], i.e., the Fully Informative Causal Features (FICF) setting.*

Remark 2 (Causality and Invariance). *For causal features Z_c, $\mathbb{E}[Y \mid \mathbf{do}(Z_c = z_c)] = \mathbb{E}[Y \mid Z_c = z_c]$, where $\mathbf{do}(\cdot)$ is the $\mathbf{do}$-operator [45], i.e., intervention or distribution shift. Fixing $P(Z_c)$ across domains in Equation 2 maintains this property and avoids additional complexity related to $\mathbf{do}(Z_c)$.*

Definition 2 (Optimal ID Causal Predictor f_c^*).

$$f_c^* = \underset{f \in \mathcal{F}_c}{\operatorname{argmin}} \mathcal{R}^{ID}(f) \tag{3}$$

Definition 3 (Optimal ID Predictor f_X^*).

$$f_X^* = \underset{f \in \mathcal{F} \setminus \mathcal{F}_c}{\operatorname{argmin}} \mathcal{R}^{ID}(f) \tag{4}$$

By definition, $f_c^* \in \mathcal{F}_\text{c}$ does not use spurious features, and $f_X^* \in \mathcal{F}$ uses spurious features.

Theorem 1. *WLOG, let $P_{ID} = \mu(I, I)$, generated by Equation 2 and denote $P_{OOD} = \mu(M, \Lambda)$ as an arbitrary target domain, parameterized by interventions M, Λ, where $M \in \mathbb{R}^{k \times k}$ and $\Lambda \in \mathbb{R}^{k \times k} \succ 0$ and Λ is symmetric. Let $\mathcal{E}_{train} = \{P_{ID}\}$ and $\mathcal{E}_{test} = \{P_{OOD}\}$. Let $\mathcal{F}$ be the class of linear classifiers of the form $Z_c \cdot \beta_c + Z_e \cdot \beta_e$. We then consider two models $f_X \in \mathcal{F} \setminus \mathcal{F}_c$ and $f_c^* \in \mathcal{F}_c$, Definition 2-3.*

$$\max_{f \in \mathcal{F} \setminus \mathcal{F}_c} acc_{OOD}(f_X) < acc_{OOD}(f_c^*) \tag{5}$$

if and only if

$$\frac{p(\mu_c^T \Sigma_c^{-1} \mu_c) + \alpha(\Sigma_e^{-1}\mu_e)^T M \mu_e}{\sqrt{(\mu_c^T \Sigma_c^{-1} \mu_c) + p(1-p)(\mu_c^T \Sigma_c^{-1} \mu_c)^2 + (\mu_e^T \Sigma_e^{-1} \Lambda \mu_e) + \alpha(1-\alpha)((\Sigma_e^{-1}\mu_e)^T M \mu_e)^2}} < \frac{p(\mu_c^T \Sigma_c^{-1} \mu_c)}{\sqrt{(\mu_c^T \Sigma_c^{-1} \mu_c) + p(1-p)(\mu_c^T \Sigma_c^{-1} \mu_c)^2}} \tag{6}$$

where $\alpha = pq + (1-p)(1-q) > 0$. All variables besides M and Λ are fixed for a given setting.

Two conditions for Equation 15 to hold are:

1. Spurious Correlation Reversal.
$$(\Sigma_e^{-1}\mu_e)^T M \mu_e < 0 \tag{7}$$
2. Sufficient Decrease in Signal-to-Noise Ratio. *Specifically referring to $\alpha(\Sigma_e^{-1}\mu_e)^T M \mu_e$ and $(\mu_e^T \Sigma_e^{-1} \Lambda \mu_e) + \alpha(1-\alpha)((\Sigma_e^{-1}\mu_e)^T M \mu_e)^2$, respectively.*

The proof for Theorem 1 is provided in Appendix C.1.

Remark 3. *The SNR condition is more intuitive with some simplifications. Let $Z_c = Z_e = \Lambda = I$, $\|\mu_c\| = 1$, and $p = 0.5$.*

$$\max_{f \in \mathcal{F} \setminus \mathcal{F}_c} acc_{OOD}(f_X) < acc_{OOD}(f_c^*) \text{ if and only if}$$

$$\textit{Spurious Correlation Reversal:} \qquad \mu_e^T M \mu_e < 0 \tag{8}$$

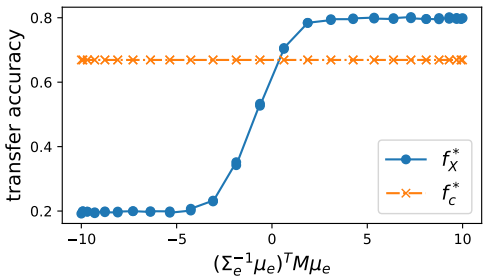
$$\textit{Sufficient decrease in signal-to-noise ratio:} \qquad (\mu_e^T M \mu_e)^2 + 2.5(\mu_e^T M \mu_e) < \|\mu_e\|^2 \tag{9}$$

For $\lambda_{\max}$ optimistically and $\lambda_{\min}$ pessimistically, one needs the following for the eigenvalues of M:

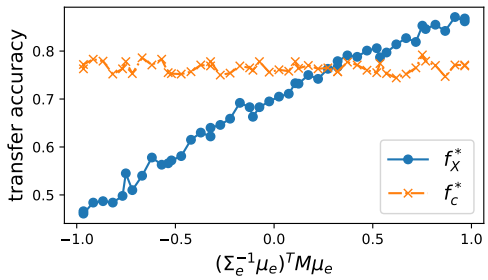
$$\|\mu_e\|^2 < \frac{1 - 2.5\lambda_{\min}}{\lambda_{\min}^2} \qquad \|\mu_e\|^2 < \frac{1 - 2.5\lambda_{\max}}{\lambda_{\max}^2}. \tag{10}$$

If $M \succ 0$, i.e., no spurious correlation reversal, Equation 10 does not hold if $\lambda > \frac{1}{2.5}$ for $\mu_e \neq 0$.

1. Spurious Correlation Reversal This condition suggests that the intervention effect of M should invert the direction of the spurious correlation, i.e., $\langle \Sigma_\text{e}^{-1} \mu_e, M \mu_e \rangle < 0$.



(a) Transfer accuracy where we fix the scaling effects of M, $\Lambda = I$, and vary the rotation effect of M on μ_e in the target domain, i.e., $M = \text{rotation\_matrix}(\theta)$ where $\theta \in [0, 2\pi]$.



(b) Transfer accuracy where we fix the rotation effects of M, $\Lambda = I$ on μ_e in the target domain, and vary the scaling effect of M, i.e., $M = aI$ where $a \in \mathbb{R}$.

Figure 2: Models are logistic regression trained/evaluated on 1000 disjoint samples. $Z_c, Z_e \in \mathbb{R}^2$ and are concatenated to get X. $\mu_c = \mu_e = [1, 1]$. In-Distribution is defined by $\mu(I, I)$, and new domains are defined by $\mu(M, \Lambda)$.

2. Sufficient decrease in signal-to-noise ratio. Alternatively, M, Λ should sufficiently decrease the signal-to-noise ratio of Z_e—via increased non-signal variance or change in covariance structure.

Figure 2 demonstrates the conditions in Theorem 1 empirically with simulated data of Equation 2. Additional details on the simulation can be found in Appendix A. Figure 2a illustrates the *alignment effect* of M. Recall that $u^T v = \|u\|\|v\| \cos(\theta)$. For $(\Sigma_e^{-1}\mu_e)^T M\mu_e$, we fix the magnitudes $\|M\mu_e\|$ and $\|\Sigma_e^{-1}\mu_e\|$ to have unit norm. We then vary θ, the angle between the training $\Sigma_e^{-1}\mu_e$ and test $M\mu_e$. We select 50 θ's split between $[0, 2\pi]$. Figure 2b shows the *magnitude effect of* M. We fix θ with (i) positive M_+ and (ii) negative signal contribution M_-. We vary M's *scaling* by sampling $a \in [-1, 1]$ where $M = aM_\pm$. When M's scaling inverts the spurious correlation, the causal model transfers better than the non-causal model. Also, when the scaling reduces the spurious signal-to-noise ratio in the target domain ($\|M\| \ll 1$), condition 2 of Theorem 1 holds, and the causal model still outperforms non-causal models.

3.1 Accuracy on the Inverse Line

From our discussion, suppose one includes non-causal features with an independent view of the label, e.g., anticausal features [58]. One would expect that the non-causal features can be used to increase ID accuracy but at the cost of a decrease in OOD accuracy. However, [43, 61] show that for many distribution shift benchmarks, an increase in ID accuracy strongly implies an increase in OOD accuracy, i.e., a (log)linear relationship between ID and OOD accuracy, *accuracy on the line*. This is particularly true for benchmarks they refer to as *natural*, i.e., non-(semi)synthetic,

Our following results suggest that this observation may be a sign of benchmarks where causal models do not outperform non-causal models in transfer accuracy, i.e., one cannot benchmark causal representation learning with the domain generalization task. We examine this relationship below.

Definition 4 (Correlation Property [43]; Accuracy on the Line)**.**

$$\left|\Phi^{-1}\left(acc_{P_{ID}}(f)\right) - c \cdot \Phi^{-1}\left(acc_{P_{OOD}}(f)\right)\right| \leq \alpha \forall\, f \tag{11}$$

where $c \in \mathbb{R}$, $\alpha \geq 0$ and Φ is the Gaussian CDF.

Theorem 2 (Accuracy on the line)**.** *Let $P_{ID} = \mu(M_{ID}, \Lambda_{ID})$ and $P_{OOD} = \mu(M_{OOD}, \Lambda_{OOD})$.*

The correlation property, Definition 4, holds if and only if for any arbitrary classifiers, $[w_c, w_e]$,

$$\left| \frac{pw_c^T\mu_c + \alpha w_e^T M_{ID}\mu_e}{\sqrt{(w_c^T\Sigma_c^{-1}w_c) + p(1-p)(\mu_c^T w_c)^2 + (w_e^T\Lambda_{ID}\Sigma_e^{-1}w_e) + \alpha(1-\alpha)(w_e^T M_{ID}\mu_e)^2}} - c \cdot \frac{pw_c^T\mu_c + \alpha w_e^T M_{OOD}\mu_e}{\sqrt{(w_c^T\Sigma_c^{-1}w_c) + p(1-p)(\mu_c^T w_c)^2 + (w_e^T\Lambda_{OOD}\Sigma_e^{-1}w_e) + \alpha(1-\alpha)(w_e^T M_{OOD}\mu_e)^2}} \right| < \epsilon \tag{12}$$

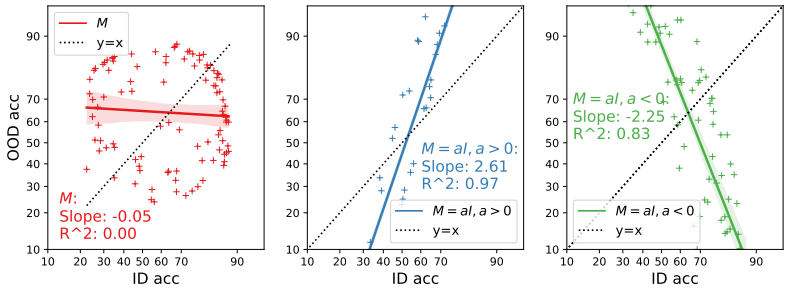


Figure 3: ID vs. OOD accuracy on probit scale. When M's are arbitrary and satisfy the SNR condition in Theorem 1, the accuracy on the line phenomenon does not occur. For $M_{\text{ID}} = I$ and $M_{\text{OOD}} = aI$, where a is allowed to vary, we observe accuracy on the line for $|a| \approx 1$. When $a < 0$, we have the spurious correlation reversal condition and have accuracy on the inverse line.

where $c \in \mathbb{R}$, $\epsilon \geq 0$. *Proof in Appendix C.2.*

Theorem 2 builds on Theorem 1, allowing us to examine the link between *accuracy on the line* and the utility of domain generalization benchmarks for causal representation learning. It shows that accuracy on the line generally does not occur for Equation 2 or under strong distribution shifts. Specifically, Equation 12 holds only holds for $\epsilon \approx 0$ when $M_{\text{ID}} \approx M_{\text{OOD}}$ and $\Lambda_{\text{ID}} \approx \Lambda_{\text{OOD}}$. In this case, datasets with accuracy on the line would not realistically meet the conditions of Theorem 1. Similarly, [43] show that accuracy on the line can be achieved asymptotically w.r.t. feature size, i.e., $m + k \to \infty$. However, the setting where this holds also does not satisfy Theorem 1's conditions, making it unsuitable for benchmarking causal representation learning—Appendix C.2.1.

On the other hand, if one observes accuracy on the inverse line, where WLOG $\text{sign}(w^e M_{\text{ID}} \mu_e) \neq \text{sign}(w^e M_{\text{OOD}} \mu_e)$, then the spurious correlation reversal condition from Theorem 1 is satisfied. This allows such a domain generalization task to be a reliable proxy benchmark for causal representation learning. We discuss accuracy on the inverse line more in Appendix C.2.

Finally, we extensively evaluate accuracy on the line on popular real-world domain generalization datasets. We also evaluate accuracy on the line for state-of-the-art domain generalization algorithms on these datasets [20, 27]. Many state-of-the-art domain generalization datasets exhibit properties that our results suggest would make them ineffective for benchmarking causal representation learning under current domain generalization benchmarking practices—Table 1 and Appendix B.

Table 1: ID acc vs. OOD acc. We train on a set of ID distributions and test on a left-out OOD distribution. Accuracies are probit transformed before comparison. We find that only one dataset configuration gives accuracy on the inverse line—ColoredMNIST with Env 2 (-90) as OOD. However, there is notable variance in the strength of correlations across datasets and their configurations. Additional datasets and analysis are provided in Appendix B.

ColoredMNIST [4]

OOD	slope	intercept	Pearson R
Env 0	0.55	0.11	0.40
Env 1	0.92	0.02	0.98
Env 2	-0.69	-0.38	-0.59

Camelyon [5, 27]

OOD	slope	intercept	Pearson R
Env 0	0.61	0.39	0.56
Env 1	0.47	0.35	0.78
Env 2	0.42	0.60	0.46
Env 3	0.43	0.81	0.63
Env 4	0.75	-0.04	0.67

TerraIncognita [27]

OOD	slope	intercept	Pearson R
Env 0	0.55	-0.64	0.69
Env 1	0.47	-0.73	0.62
Env 2	0.47	-0.45	0.79
Env 3	0.23	-0.59	0.59

PACS [29]

OOD	slope	intercept	Pearson R
Env 0	0.73	-0.38	0.92
Env 1	0.57	-0.19	0.89
Env 2	1.04	0.00	0.91
Env 3	0.64	-0.41	0.88

4 Empirical Results and Discussion

Table 1 illustrates ambiguity about the ability of many distribution shift datasets to serve as proxy benchmarks for causal representation learning. Some datasets, however, clearly do not satisfy the criteria our results suggest, i.e., there is a strong ID-OOD accuracy correlation, e.g., PACS. We further discuss our findings in Appendix B and provide an empirical evaluation of additional datasets, where we also find that accuracy on the line for ERM models does not imply accuracy on the line for models generated by state-of-the-art domain generalization algorithms.

Our findings also highlight potential incompatibilities of common machine learning practices with the goal of causal representation learning—particularly, model selection and averaging benchmarking results across dataset configurations and different datasets. Our results suggest that model selection based on ID or held-out-domain validation accuracy may be biased towards non-causal models with higher domain-specific accuracies. Furthermore, when averaging over multiple datasets/configurations to evaluate domain generalization, including some datasets/configurations that do not satisfy our conditions may result in misleading results—more in Appendix B.1.

4.1 Related Work

Previous work has studied the accuracy of the line phenomenon. Some study theoretical conditions for the phenomenon to occur [57], and some identify real-world datasets with weak or negative linear correlations [34, 62]. Our work uniquely connects this phenomenon to the utility of the domain generalization task in benchmarking causal representation learning. Additionally, we empirically examine this property for new state-of-the-art domain generalization benchmarks.

Conditions for domain generalization benchmarks to be reliable proxies for the causal representation learning task are largely missing in the literature, despite the close tie between these two areas in previous work [46, 22, 4]. While using domain generalization as a proxy task is not currently commonplace in causal representation learning, it offers one potential solution to benchmarking challenges in high-dimensional and observational real-world data [38, 66]. This work uniquely characterizes prescriptive conditions for identifying datasets that are appropriate proxy benchmarks for causal representation learning. Like ImageNet [13] for object recognition, a well-specified real-world benchmark has the potential to yield remarkably rapid and externally valid progress [63, 53, 14]. This work applies to the principles and practices of identifying and constructing such benchmarks.

5 Conclusion

To evaluate causal representation learning in *natural* datasets, it is crucial to identify metrics that can identify models with causal representations in an arbitrary set of models. In this work, we provide evidence that to use domain generalization as a proxy task to evaluate causal representations, out-of-domain samples should either (i) have a reversal of spurious (non-causal) correlations or (ii) spurious features should have a sufficient decrease in signal-to-noise ratio. Furthermore, we identify a lack of accuracy on the line as a signature of datasets with these desired properties. Future work includes extending our theoretical analysis beyond our assumed model of distribution shifts, characterizing the implications of this extension on benchmarking, and curating a reliable set of benchmarks for benchmarking causal representation learning via domain generalization. *Finally, insofar as the goals of causal representation learning and domain generalization are aligned, the same questions and concerns about benchmarking causal representation learning equally apply to benchmarking domain generalization directly.*

Acknowledgments

OS was partly supported by the UIUC Beckman Institute Graduate Research Fellowship, NSF-NRT 1735252, GEM Associate Fellowship, and the Alfred P. Sloan MPhD Program. SK acknowledges support by NSF 2046795 and 2205329, NIFA award 2020-67021-32799, the Alfred P. Sloan Foundation, and Google Inc.

References

[1] Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

[2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34: 3438–3450, 2021.

[3] John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[6] Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly Stachenfeld, Pushmeet Kohli, Peter Battaglia, and Jessica Hamrick. Structured agents for physical construction. In *International conference on machine learning*, pages 464–474. PMLR, 2019.

[7] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

[8] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 29, 2016.

[9] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[10] Annie S Chen, Yoonho Lee, Amrith Setlur, Sergey Levine, and Chelsea Finn. Confidence-based model selection: When to take shortcuts for subpopulation shifts. *arXiv preprint arXiv:2306.11120*, 2023.

[11] Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35: 22131–22148, 2022.

[12] Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] David Donoho. Data science at the singularity. *Harvard Data Science Review*, 6(1), 2024.

[15] Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022.

[16] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[17] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

[19] Simon Guiroy, Christopher Pal, Gonçalo Mordido, and Sarath Chandar. Improving meta-learning generalization with activation-based early-stopping. In *Conference on lifelong learning agents*, pages 213–230. PMLR, 2022.

[20] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[21] Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115, 1944.

[22] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[24] Kevin D Hoover. The logic of causal inference: Econometrics and the conditional analysis of causation. *Economics & Philosophy*, 6(2):207–234, 1990.

[25] Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, volume 126, pages 750–782. PMLR, 2020.

[26] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[27] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[28] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[30] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[31] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. icitris: Causal representation learning for instantaneous temporal effects. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[32] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[33] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170, 2021.

[34] Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. In *Conference on Causal Learning and Reasoning*, pages 553–573. PMLR, 2023.

[36] Adian Liusie, Vatsal Raina, Vyas Raina, and Mark John Francis Gales. Analyzing biases to spurious correlations in text classification tasks. In *AACL*, 2022. URL `https://api.semanticscholar.org/CorpusID:253762052`.

[37] Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pages 662–691. PMLR, 2023.

[38] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6979–6987, 2017.

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

[40] Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.

[41] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pages 7313–7324. PMLR, 2021.

[42] Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 5 2022. URL `http://arxiv.org/abs/2105.06422`.

[43] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[44] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

[45] J Pearl. *Causality*. Cambridge university press, 2009.

[46] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[47] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[48] Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with xrm. *arXiv preprint arXiv:2309.16748*, 2023.

[49] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[50] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

[51] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2641–2657. PMLR, 2022.

[52] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[53] Olawale Salaudeen and Moritz Hardt. Imagenot: A contrast with imagenet preserves model rankings. *arXiv preprint arXiv:2404.02112*, 2024.

[54] Olawale Salaudeen and Sanmi Koyejo. Causally inspired regularization enables domain general representations. In *International Conference on Artificial Intelligence and Statistics*, pages 3124–3132. PMLR, 2024.

[55] Olawale Elijah Salaudeen and Oluwasanmi O Koyejo. Exploiting causal chains for domain generalization. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

[56] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.

[57] Amartya Sanyal, Yaxi Hu, Yaodong Yu, Yian Ma, Yixin Wang, and Bernhard Schölkopf. Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation. *arXiv preprint arXiv:2406.19049*, 2024.

[58] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[59] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[60] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

[61] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[62] Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[63] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[64] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[65] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34: 16451–16467, 2021.

[66] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

[67] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[68] Jiaxuan Wang, Sarah Jabbour, Maggie Makar, Michael Sjoding, and Jenna Wiens. Learning concept credible models for mitigating shortcuts. In *Advances in Neural Information Processing Systems*, volume 35, pages 33343–33356, 12 2022.

[69] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

[70] Jiayun Zheng and Maggie Makar. Causally motivated multi-shortcut identification and removal. In *Advances in Neural Information Processing Systems*, volume 35, pages 12800–12812, 12 2022.

A Semi-Synthetic Experiments

$$\mu(M,\Lambda) = \begin{cases} \eta \sim \text{Bern}(q); C \sim \text{Bern}(p) \\ Y = \begin{cases} C & \text{if } \eta = 1 \\ \neg C & \text{if } \eta = 0 \end{cases} \\ Z_{\text{c}} \sim \mathcal{N}\big(C \cdot \mu_c, \Sigma_{\text{c}}\big) \\ Z_{\text{e}} \sim \mathcal{N}\big(Y \cdot M\mu_e, \Lambda\Sigma_{\text{e}}\big) \\ X = Z_{\text{c}} \oplus Z_{\text{e}}, \end{cases} \tag{13}$$

where η is a noise term, $Z_{\text{c}} \in \mathbb{R}^m$, $Z_{\text{e}} \in \mathbb{R}^k$, and $Y \in \{0, 1\}$. Furthermore, we define interventions (shifts), or context variables, M, Λ that parameterize a domain, where $M \in \mathbb{R}^{k\times k}$ and $\Lambda \succeq 0 \in \mathbb{R}^{k\times k}$.

Parameters. We use the following parameters across our experiments: $p = 0.5$, $q =$, $\mu = [1, 1]$, $\Sigma_{\text{c}} = \text{diag}([1, 1])$, $\mu_e = [1, 1]$. We expect our results to hold independent of these parameters. We chose these parameters for the ease of intuition of the results on the simulated dataset. We use a samples size of 1000 for each domain.

Accuracy on the line models. We use Logistic Regression classifiers to evaluate our accuracy on the line results. To generate arbitrary classifiers, we randomly sample all parameters of Equation 13 50 times, independent and identically distributed (iid), and evaluate classifiers learned from samples generated by these random parameters on P_{ID} and P_{OOD}.

B Real-World Datasets

We evaluated benchmarks in *DomainBed*, which features object recognition benchmarks for domain generalization [20]. We also evaluated benchmarks in *Wilds*, which aims to be more representative of real-world shifts, showcasing a variety of real-world applications.

We find that in addition to the dataset sets already shown to have accuracy on the line by previous works [43, 61], many of the benchmarks we evaluated also exhibit a strong correlation between ID and OOD accuracy. These findings highlight a gap in desired domain generalization datasets as proxy benchmarks for causal representation learning.

We provide two figures for each domain generalization benchmark: (i) ERM accuracy on the line- we evaluate the correlation between the ID and OOD accuracy of naive ERM models under the aforementioned varying conditions. Given the multisource settings of this work, we evaluate each training sub-ID-domain against the corresponding OOD domain. (ii) Model-specific accuracy on the line – we evaluate a set of domain generalization algorithms, enumerated below and including ERM, on the domain generalization task. We treat the training domains as one concatenated ID distribution against the corresponding OOD domain. Each of these figures is also accompanied by a table of the slope and intercept of a regression of OOD accuracy on ID accuracy and the corresponding Pearson R, p-value, and standard error.

B.1 Discussion.

A few important points can be gleaned from our empirical analysis of naive ERM:

1. We only observe accuracy on the inverse line for one configuration of ColoredMNIST out of all benchmarks and their configuration.

2. For some benchmarks, e.g., Spawrious-O2O-Easy, some configurations exhibit a low Pearson R, so the ID-OOD correlation is not strong, relatively. However, for other configurations of the same datasets, there is a high Pearson R. Thus, some configurations of a dataset may be more reliable than others. The average over configurations of a dataset is often reported [20]. However, when only one configuration satisfies the properties for the causal model to transfer best, averaging over these configurations also yields an ineffective benchmark.

   Moreover, averages over many datasets are also often used to demonstrate superior methods for domain generalization – this suffers from the same limitation at a grander scale [20].

   In the subpopulation shift literature, where the distribution shift is w.r.t the mixture of samples from a fixed set of groups (subpopulations), worst-group transfer accuracy is the standard evaluation metric [27]. A similar norm for the broader domain generalization task evaluation is beneficial.

3. A similar concern exists for the practice of cross-validation, where models are selected based on accuracy on some held-out test set. This could be a standard IID hold-out test set or a held-out domain as in Gulrajani and Lopez-Paz [20]. As demonstrated in this work, selecting models based on highest accuracy on a held-out set, IID or a fixed OOD set, may lead to selecting models that are overfit to the additionally informative spurious correlations in the held-out set. Previous work has proposed alternative criteria for model selection such as implied conditional independencies [54], cross-risk minimization [48], early stopping [19], and confidence-based aggregation with ensembles of models [10].

4. Benchmarks with arbitrarily selected *natural datasets* [61], e.g., tumor films from different hospitals, are not guaranteed to illuminate causal properties or robustness to interventions/distribution shifts from OOD accuracy.

B.2 Architectures

Models We use the experimental setup of DomainBed for the following results [20] — https://github.com/facebookresearch/DomainBed.

Table 2: MNIST ConvNet architecture.

#	Layer
1	Conv2D (in=d, out=64)
2	ReLU
3	GroupNorm (groups=8)
4	Conv2D (in=64, out=128, stride=2)
5	ReLU
6	GroupNorm (groups=8)
7	Conv2D (in=128, out=128)
8	ReLU
9	GroupNorm (groups=8)
10	Conv2D (in=128, out=128)
11	ReLU
12	GroupNorm (8 groups)
13	Global average-pooling

We leverage a ConvNet architecture for the ColoredMNIST dataset (Table 2); we vary hyperparameters enumerated in [20] in addition to randomly varying the number of convolutional layers between 3 and 4. For other methods, we use the following architectures as featurizers with varying hyperparameters and seeds: ResNet (18, 50). We also vary the number of hidden layers with ReLU activations on top of these featurizers – we vary between 0-3 layers. We also randomly either fine-tune ImageNet weights or train from randomly initialized weights. We also randomly select whether or not there is data augmentation. We included models trained with different epochs as in previous accuracy on the line work [62].

Like in previous work, we evaluate the accuracy of the line phenomenon in the state-of-the-art distribution shift benchmarks. We focus on benchmarks not sufficiently characterized in previous work [49, 43, 61, 62, 34]. Like in previous work, we evaluate accuracy on the line with models trained on naive ERM. In later sections, we will discuss our findings for models trained with other algorithms, i.e., not naive ERM. Furthermore, unlike some previous work, we perform our analysis in each setting of the common leave-one-domain-out configurations common in the literature [20], i.e., each domain/distribution is evaluated as OOD w.r.t to the other domain/distributions in the dataset as ID – $\binom{n}{1}$ evaluations where n is the number of domain/distributions in the dataset.

B.3 Domain Generalization Algorithms.

- **Adaptive Risk Minimization (ARM) [69].** trains models to adapt at test time using unlabeled data from shifted domains. It directly optimizes for effective adaptation by learning to adjust during training across multiple domains, rather than focusing on invariant features or robustness.
- **Empirical Quantile Risk Minimization (EQRM) [15].** minimizes the α-quantile of risk across domains, linking training and test domains via a shared meta-distribution to adapt to likely shifts.
- **Invariant Risk Minimization (IRM) [4]** aims to learn predictors that generalize across different training environments by finding invariant representations. IRM encourages models to learn features that have a consistent causal relationship with the target variable, regardless of domain-specific spurious correlations. This is achieved by seeking a predictor that remains optimal across all training environments.
- **Correlation Alignment (CORAL) [60]** aims to minimize the discrepancy between the feature distributions of different domains by aligning second-order statistics (covariances); this is achieved by reducing the distance between the covariance matrices of the source and target domains.
- **Empirical Risk Minimization (ERM) [64]** is a foundational approach in machine learning that focuses on minimizing the average loss (risk) over the training data. ERM assumes that the training and test data are drawn from the same distribution, and it aims to find a predictor that performs well by directly optimizing the empirical loss on the observed training examples.
- **Maximum Mean Discrepancy (MMD) [30]** aims to minimize this discrepancy, ensuring that the feature distributions across domains are aligned, facilitating better generalization to unseen domains. This is achieved by comparing the feature distributions of source and target domains by mapping data points into a reproducing kernel Hilbert space (RKHS) and calculating the distance between their means.
- **Causal Invariant Representation Learning (CausIRL_(CORAL/MMD)) [12]** is designed to improve domain generalization by combining causal invariant representation learning with established distribution alignment techniques like CORAL and MMD. In CausIRL, the aim is to learn representations invariant to domain-specific features (spurious correlations) by focusing on causal mechanisms.
- **Group Distributionally Robust Optimization (GroupDRO) [52].** is a method that focuses on improving robustness to distribution shifts by ensuring good performance across all groups within the data. Instead of optimizing for average performance, GroupDRO minimizes the worst-case loss over predefined groups. This approach assigns higher weights to underperforming groups during training, allowing the model to improve its worst-case accuracy and generalize better across unseen domains.
- **Variance Risk Extrapolation (VREx) [28].** aims to improve robustness by minimizing the variance of risks across different training domains. Instead of just focusing on minimizing the average loss, VREX ensures that the model's risk is evenly distributed across domains by reducing the variability in risk.
- **Information Bottleneck Invariant Risk Minimization (IB_IRM) [2].** combines the principles of the Information Bottleneck (IB) and Invariant Risk Minimization (IRM). IB-IRM aims to balance predictive performance and robustness by learning representations that capture the essential, invariant causal features while ignoring spurious correlations.

B.3.1 Discussion

We find substantial variance in the correlation between ID and OOD accuracy for models given by state-of-the-art domain generalization algorithms.

For the ColoredMNIST configuration that has accuracy on the inverse line for ERM, we observe that state-of-the-art domain generalization algorithms also exhibit accuracy on the inverse line. In other configurations, some algorithms, like IRM, IB_IRM, CausIRL_CORAL, and MMD, do not give models that have significant linear correlations ID-OOD, i.e., their p-values are far from significant.

B.4 DomainBed Results

ColoredMNIST [4]. The ColoredMNIST dataset is a variation of the MNIST dataset where the primary goal is still to predict a binary label assigned to each image based on the digit. Whereas the MNIST images are grayscale, the ColoredMNIST digits are colored either red or green to spuriously correlate with the binary label. Additionally, the observed label associated with each image is a noisy version of the true binary label, making the true label more correlated with the color than the digit. The ColoredMNIST generative mechanism proposed by [4] is described as follows: (1) a preliminary binary label $\widetilde{y}$ is assigned to each image according to the digit (digit ≥ 5), (2) the observed label is obtained by flipping $\widetilde{y}$ with probability 0.25, (3) the color of the digit is sampled by flipping y with probability p_e. For the results presented in Figure 1, we consider $p_e \in \{0.10, 0.15, 0.25, 0.75, 0.90\}$ to define five distinct domains, using $p_e = 0.10$ as the training domain and evaluating trained models on the remaining test domains.

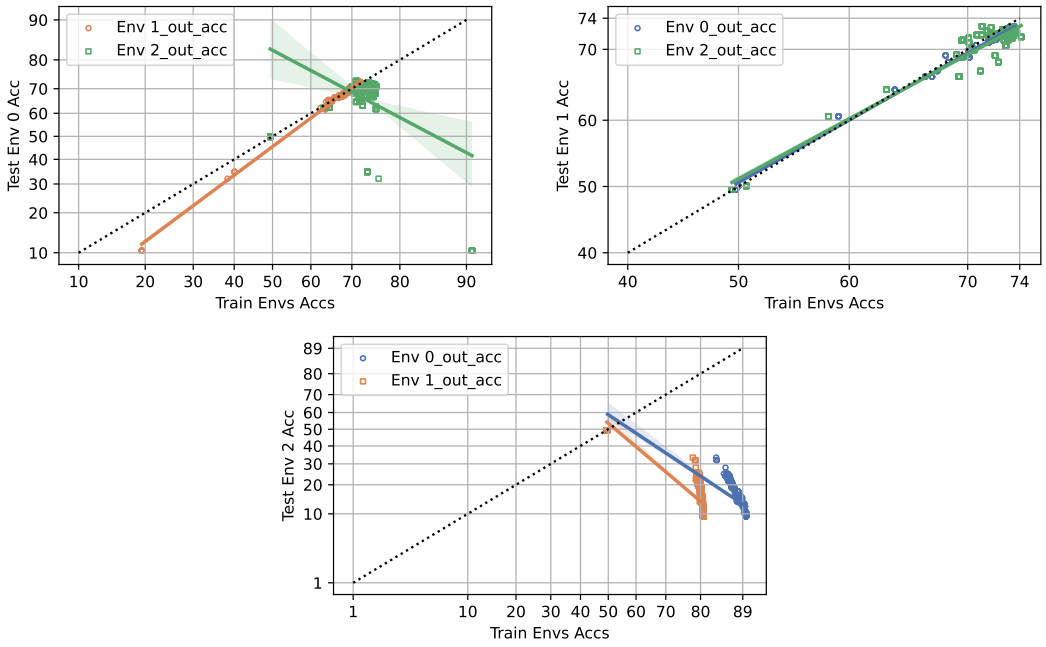


Figure 4: ColoredMNIST correlations between in-distribution vs. out-of-distribution model accuracy for different training domains.

Table 3: ColoredMNIST ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 1 acc	Env 0 acc	1.24	-0.11	1.00	0.00	0.01
Env 2 acc	Env 0 acc	-0.88	0.95	-0.43	0.00	0.09
Env 0 acc	Env 1 acc	0.94	0.02	1.00	0.00	0.00
Env 2 acc	Env 1 acc	0.91	0.03	0.96	0.00	0.01
Env 0 acc	Env 2 acc	-1.10	0.22	-0.84	0.00	0.03
Env 1 acc	Env 2 acc	-1.41	0.09	-0.74	0.00	0.06

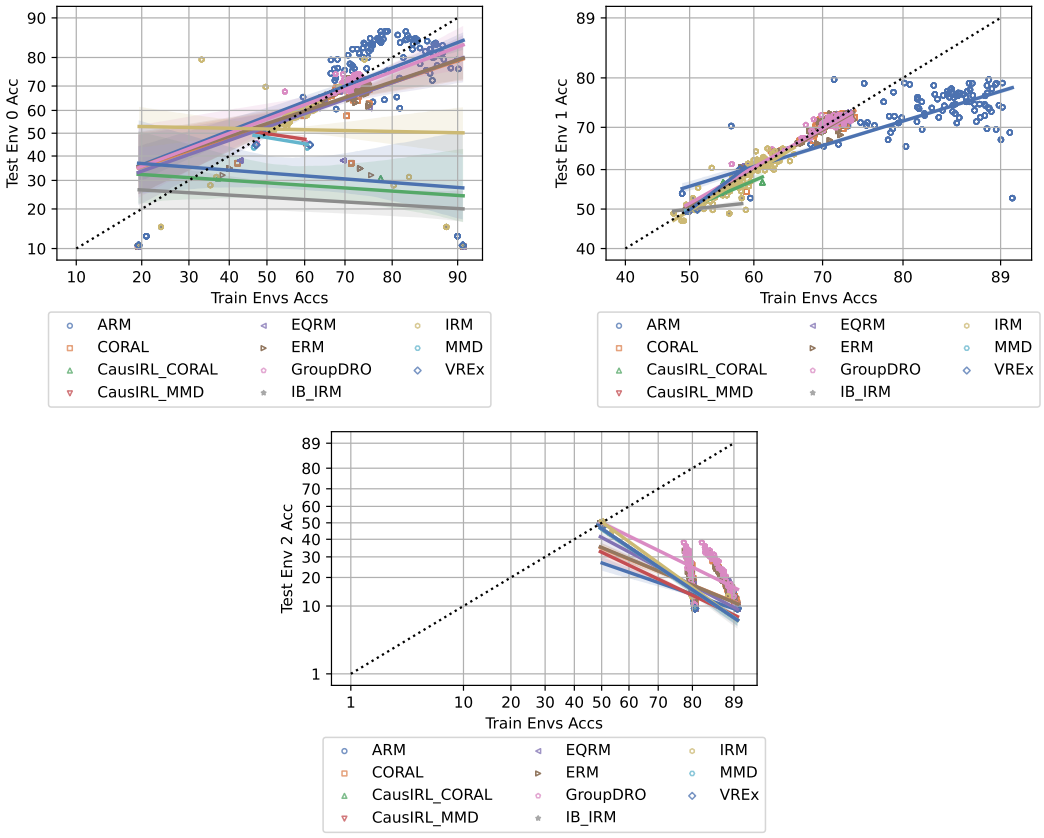


Figure 5: ColoredMNIST correlations between in-distribution vs. out-of-distribution model accuracy for different domain generalization algorithms.

Table 4: ColoredMNIST ID vs. OOD properties.

Algorithm	OOD	slope	intercept	Pearson R	p-value	standard error
ARM	Env 0 acc	0.64	0.19	0.47	0.00	0.04
CORAL	Env 0 acc	0.55	0.09	0.40	0.00	0.05
CausIRL_CORAL	Env 0 acc	-0.11	-0.55	-0.14	0.42	0.13
CausIRL_MMD	Env 0 acc	-0.24	-0.00	-0.39	0.00	0.03
EQRM	Env 0 acc	0.58	0.07	0.42	0.00	0.06
ERM	Env 0 acc	0.55	0.11	0.40	0.00	0.04
GroupDRO	Env 0 acc	0.62	0.16	0.44	0.00	0.04
IB_IRM	Env 0 acc	-0.10	-0.71	-0.14	0.30	0.09
IRM	Env 0 acc	-0.03	0.05	-0.03	0.42	0.04
MMD	Env 0 acc	-0.28	-0.05	-0.44	0.00	0.07
VREx	Env 0 acc	-0.13	-0.44	-0.14	0.36	0.14
ARM	Env 1 acc	0.50	0.15	0.77	0.00	0.01
CORAL	Env 1 acc	0.96	0.01	0.98	0.00	0.01
CausIRL_CORAL	Env 1 acc	0.69	0.01	0.91	0.00	0.04
CausIRL_MMD	Env 1 acc	0.95	-0.00	0.92	0.00	0.03
EQRM	Env 1 acc	0.95	0.02	0.99	0.00	0.00
ERM	Env 1 acc	0.92	0.02	0.98	0.00	0.01
GroupDRO	Env 1 acc	0.91	0.04	0.98	0.00	0.01
IB_IRM	Env 1 acc	0.17	-0.00	0.40	0.00	0.06
IRM	Env 1 acc	0.98	-0.01	0.90	0.00	0.02
MMD	Env 1 acc	-0.00	-0.01	-0.00	1.00	0.00
VREx	Env 1 acc	1.23	-0.01	0.99	0.00	0.03
ARM	Env 2 acc	-0.58	-0.62	-0.64	0.00	0.02
CORAL	Env 2 acc	-0.68	-0.39	-0.60	0.00	0.03
CausIRL_CORAL	Env 2 acc	-1.11	-0.10	-0.96	0.00	0.04
CausIRL_MMD	Env 2 acc	-0.79	-0.45	-0.80	0.00	0.03
EQRM	Env 2 acc	-0.88	-0.23	-0.83	0.00	0.04
ERM	Env 2 acc	-0.69	-0.38	-0.59	0.00	0.03
GroupDRO	Env 2 acc	-0.81	0.00	-0.55	0.00	0.04
IB_IRM	Env 2 acc	-1.12	-0.09	-0.96	0.00	0.06
IRM	Env 2 acc	-1.20	0.01	-0.98	0.00	0.01
MMD	Env 2 acc	-1.11	-0.10	-0.96	0.00	0.05
VREx	Env 2 acc	-1.12	-0.09	-0.96	0.00	0.04

Spawrious [40]. The Spawrious image classification benchmark suite consists of six different datasets, including one-to-one (O2O) spurious correlations, where a single spurious attribute correlates with a binary label, and many-to-many (M2M) spurious correlations across multiple classes and spurious attributes. Each benchmark task is proposed with three difficulty levels: Easy, Medium, and Hard. The dataset contains images of four dog breeds $c \in \{bulldog, dachshund, labrador, corgi\}$ found in six backgrounds $b \in \{beach, desert, dirt, jungle, mountain, sand\}$. Images are generated using text-to-image models and filtered using an image-to-text model for quality control. This benchmark suite consists of 152,064 images of dimensions (3, 224, 224).

For the O2O task, the class (dog breed) and background combinations are sampled such that $\mu\%$ of the images per class contain a spurious background b^{sp} and $(100-\mu)\%$ contain a generic background b^{ge}. While the generic background is held constant for each class, each spurious background is observed in only one class ($p_{train}(b_i^{sp} \mid c_j) = 1$ if $i = j$ and 0 if $i \neq j$). Two separate training domains are defined by varying the value of μ. These induced spurious correlations are reverted to yield a test domain with a single background for each class ($p_{test}(b_i \mid c_i) = 1$).

For the M2M task, disjoint class and background groups are constructed $\mathcal{B}_1, \mathcal{B}_2, \mathcal{C}_1, \mathcal{C}_2$, each with two elements. To introduce the training domains, class-background combinations (c, b) are selected with $c \in \mathcal{C}_i$ and $b \in \mathcal{B}_i$. Each training domain consists of a single background per class such that $p^e_{train}(b_k \mid c_k) = e$, with domain index $e \in \{0, 1\}$, $b_k \in \mathcal{B}_i$, $c_k \in \mathcal{C}_i$. In contrast, the test domain is generated by selecting combinations from $c \in \mathcal{C}_i$ and $b \in \mathcal{B}_j$ with $i \neq j$ and sampling backgrounds such that $p_{test}(b_1 \mid c_k) = p_{test}(b_2 \mid c_k) = 0.5$ for $c_k \in \mathcal{C}_i$, $\{b_1, b_2\} = \mathcal{B}_j$.

The difficulty level (Easy, Medium, Hard) differs due to the splits in the available class-background combinations. These splits were empirically determined, and the full details of the final data combinations are found in Table 2 of Lynch et al. [40].

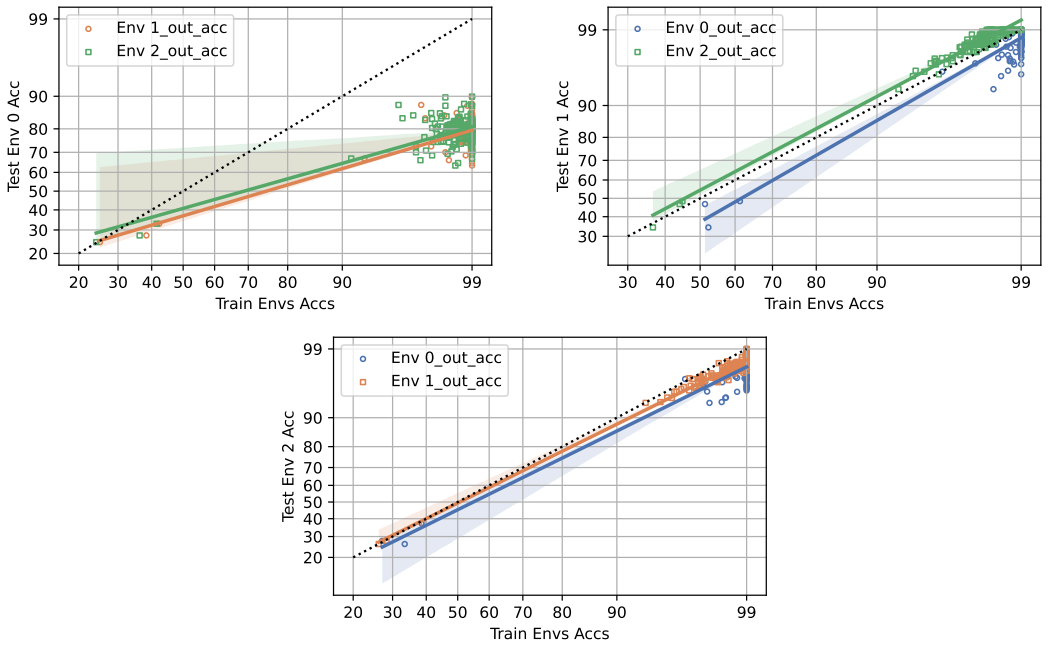


Figure 6: Spawrious-O2O-Easy correlations between in-distribution vs. out-of-distribution model accuracy for different training domains.

Table 5: SpawriousO2O_easy ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 1 acc	Env 0 acc	0.50	-0.33	0.75	0.00	0.04
Env 2 acc	Env 0 acc	0.47	-0.23	0.72	0.00	0.04
Env 0 acc	Env 1 acc	1.09	-0.33	0.93	0.00	0.04
Env 2 acc	Env 1 acc	1.01	0.11	0.98	0.00	0.02
Env 0 acc	Env 2 acc	0.93	-0.12	0.94	0.00	0.03
Env 1 acc	Env 2 acc	0.94	-0.02	0.98	0.00	0.01

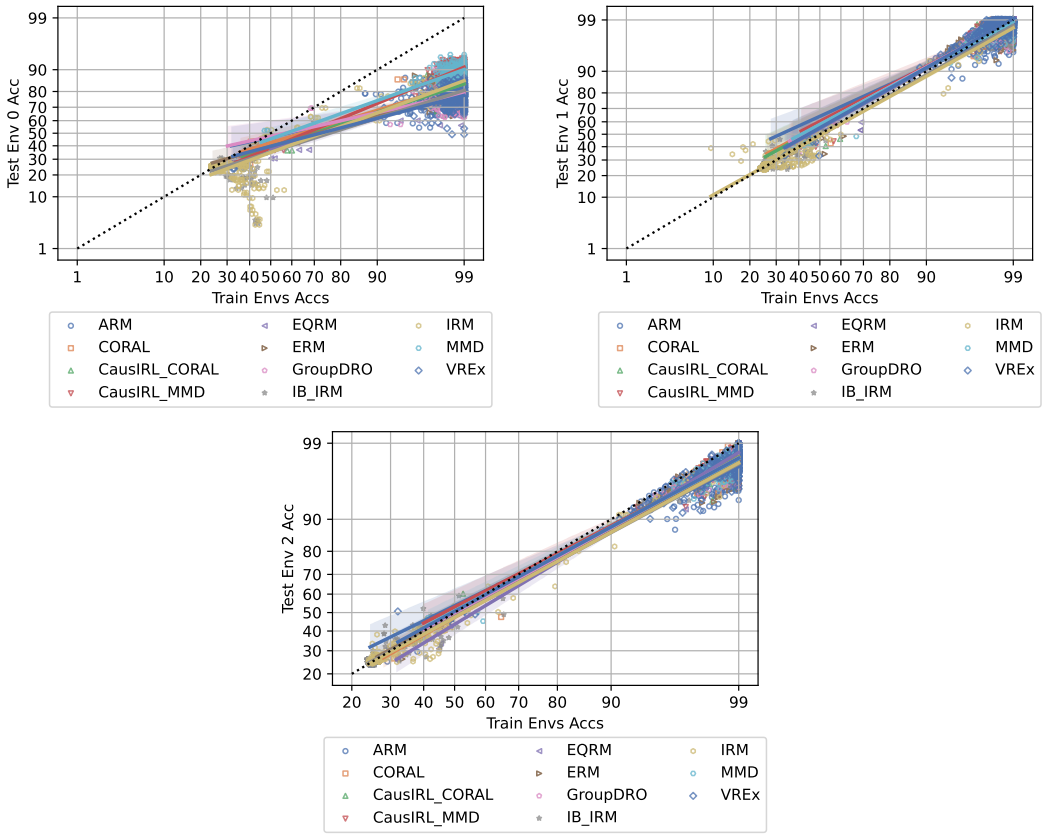


Figure 7: Spawrious-O2O-Easy correlations between in-distribution vs. out-of-distribution model accuracy for different domain generalization algorithms.

Table 6: SpawriousO2O_easy ID vs. OOD properties.

Algorithm	OOD	slope	intercept	Pearson R	p-value	standard error
ARM	Env 0 acc	0.38	-0.08	0.57	0.00	0.03
CORAL	Env 0 acc	0.51	-0.19	0.74	0.00	0.03
CausIRL_CORAL	Env 0 acc	0.55	-0.28	0.70	0.00	0.03
CausIRL_MMD	Env 0 acc	0.70	-0.30	0.87	0.00	0.02
EQRM	Env 0 acc	0.49	-0.32	0.66	0.00	0.03
ERM	Env 0 acc	0.48	-0.26	0.73	0.00	0.03
GroupDRO	Env 0 acc	0.37	-0.06	0.49	0.00	0.04
IB_IRM	Env 0 acc	0.53	-0.38	0.89	0.00	0.02
IRM	Env 0 acc	0.62	-0.39	0.86	0.00	0.02
MMD	Env 0 acc	0.60	-0.11	0.76	0.00	0.03
VREx	Env 0 acc	0.44	-0.26	0.65	0.00	0.03
ARM	Env 1 acc	0.77	0.37	0.86	0.00	0.03
CORAL	Env 1 acc	0.92	0.19	0.93	0.00	0.02
CausIRL_CORAL	Env 1 acc	0.93	0.15	0.92	0.00	0.02
CausIRL_MMD	Env 1 acc	0.87	0.26	0.89	0.00	0.03
EQRM	Env 1 acc	0.95	0.12	0.92	0.00	0.02
ERM	Env 1 acc	0.92	0.18	0.89	0.00	0.03
GroupDRO	Env 1 acc	0.92	0.19	0.91	0.00	0.02
IB_IRM	Env 1 acc	0.96	-0.02	0.99	0.00	0.01
IRM	Env 1 acc	0.94	-0.03	0.98	0.00	0.01
MMD	Env 1 acc	0.90	0.20	0.91	0.00	0.02
VREx	Env 1 acc	0.95	0.13	0.93	0.00	0.02
ARM	Env 2 acc	0.82	0.09	0.92	0.00	0.02
CORAL	Env 2 acc	0.95	-0.07	0.95	0.00	0.02
CausIRL_CORAL	Env 2 acc	0.91	0.02	0.94	0.00	0.02
CausIRL_MMD	Env 2 acc	0.87	0.08	0.94	0.00	0.02
EQRM	Env 2 acc	1.01	-0.16	0.94	0.00	0.02
ERM	Env 2 acc	0.92	-0.03	0.95	0.00	0.02
GroupDRO	Env 2 acc	0.92	-0.01	0.95	0.00	0.02
IB_IRM	Env 2 acc	0.90	-0.06	0.99	0.00	0.01
IRM	Env 2 acc	0.91	-0.07	0.99	0.00	0.01
MMD	Env 2 acc	0.90	0.02	0.95	0.00	0.02
VREx	Env 2 acc	0.91	0.02	0.95	0.00	0.02

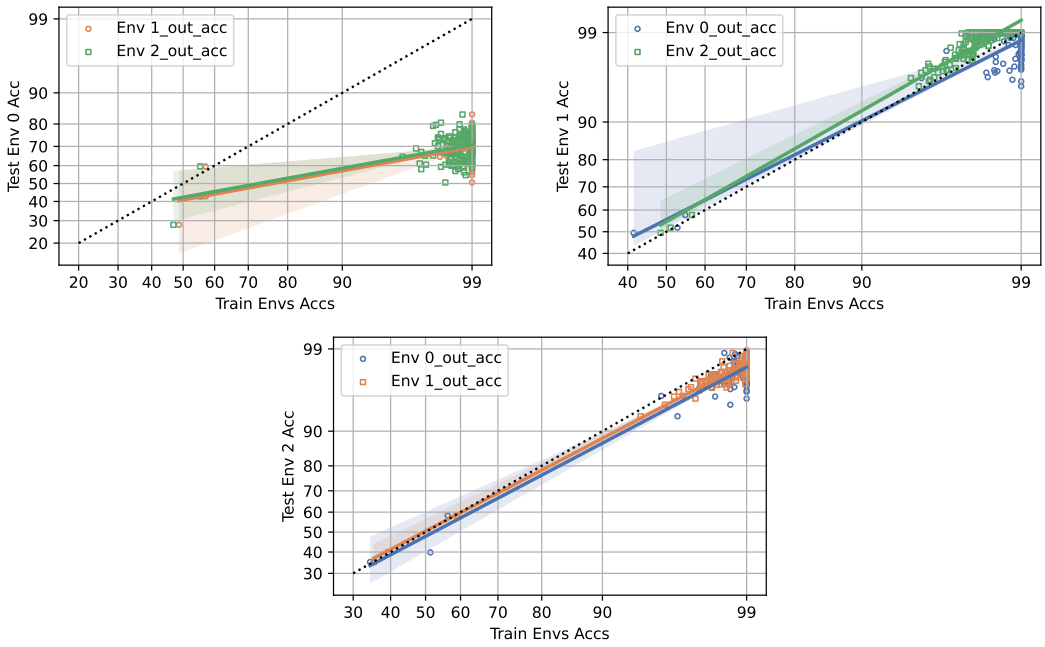


Figure 8: Spawrious-O2O-Hard correlations between in-distribution vs. out-of-distribution model accuracy for different training domains.

Table 7: SpawriousO2O_hard ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 1 acc	Env 0 acc	0.32	-0.23	0.49	0.00	0.05
Env 2 acc	Env 0 acc	0.32	-0.19	0.50	0.00	0.04
Env 0 acc	Env 1 acc	0.90	0.14	0.89	0.00	0.04
Env 2 acc	Env 1 acc	1.01	0.12	0.97	0.00	0.02
Env 0 acc	Env 2 acc	0.92	-0.05	0.93	0.00	0.03
Env 1 acc	Env 2 acc	0.92	0.01	0.97	0.00	0.02

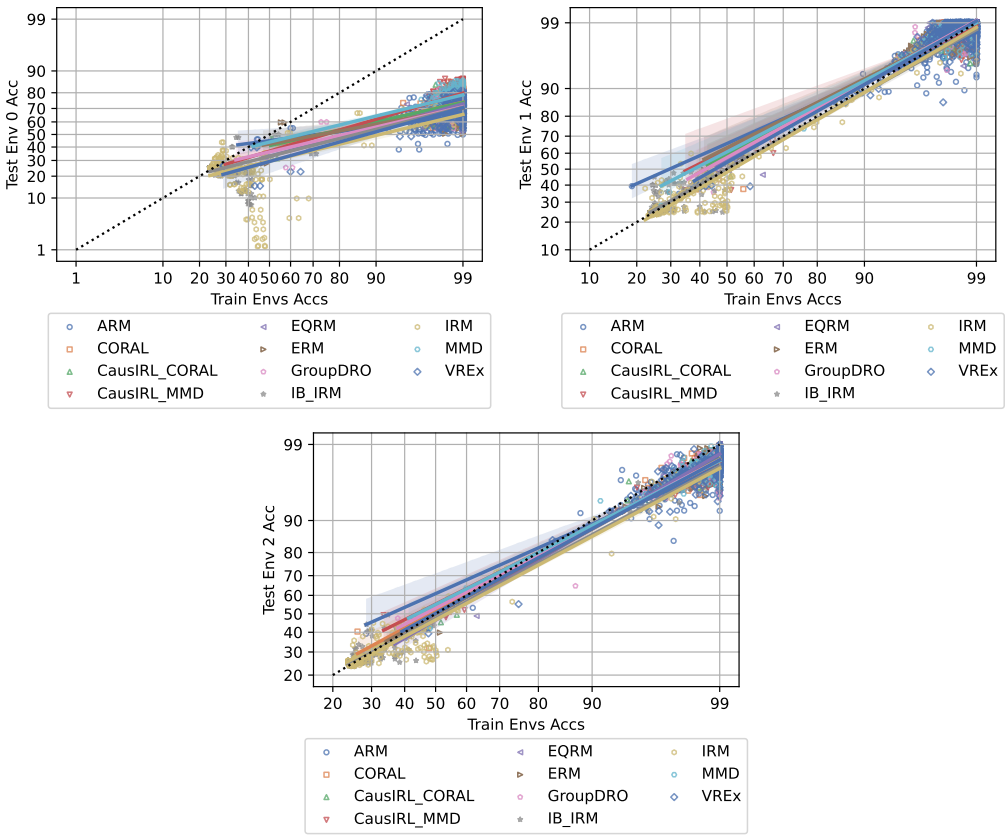


Figure 9: Spawrious-O2O-Hard correlations between in-distribution vs. out-of-distribution model accuracy for different domain generalization algorithms.

Table 8: SpawriousO2O_hard ID vs. OOD properties.

Algorithm	OOD	slope	intercept	Pearson R	p-value	standard error
ARM	Env 0 acc	0.22	-0.13	0.39	0.00	0.03
CORAL	Env 0 acc	0.38	-0.23	0.57	0.00	0.03
CausIRL_CORAL	Env 0 acc	0.37	-0.22	0.62	0.00	0.03
CausIRL_MMD	Env 0 acc	0.50	-0.34	0.76	0.00	0.02
EQRM	Env 0 acc	0.45	-0.44	0.63	0.00	0.03
ERM	Env 0 acc	0.31	-0.19	0.49	0.00	0.03
GroupDRO	Env 0 acc	0.39	-0.35	0.59	0.00	0.03
IB_IRM	Env 0 acc	0.43	-0.44	0.87	0.00	0.01
IRM	Env 0 acc	0.40	-0.54	0.65	0.00	0.03
MMD	Env 0 acc	0.41	-0.15	0.64	0.00	0.03
VREx	Env 0 acc	0.47	-0.55	0.74	0.00	0.02
ARM	Env 1 acc	0.76	0.41	0.83	0.00	0.03
CORAL	Env 1 acc	0.93	0.19	0.91	0.00	0.02
CausIRL_CORAL	Env 1 acc	0.91	0.24	0.92	0.00	0.02
CausIRL_MMD	Env 1 acc	0.85	0.30	0.87	0.00	0.03
EQRM	Env 1 acc	0.94	0.14	0.93	0.00	0.02
ERM	Env 1 acc	0.85	0.34	0.88	0.00	0.03
GroupDRO	Env 1 acc	0.94	0.19	0.92	0.00	0.02
IB_IRM	Env 1 acc	0.95	-0.00	0.98	0.00	0.01
IRM	Env 1 acc	0.99	-0.05	0.98	0.00	0.01
MMD	Env 1 acc	0.88	0.28	0.91	0.00	0.02
VREx	Env 1 acc	0.97	0.11	0.92	0.00	0.02
ARM	Env 2 acc	0.75	0.28	0.88	0.00	0.02
CORAL	Env 2 acc	0.90	0.03	0.95	0.00	0.02
CausIRL_CORAL	Env 2 acc	0.91	0.00	0.95	0.00	0.02
CausIRL_MMD	Env 2 acc	0.84	0.13	0.94	0.00	0.02
EQRM	Env 2 acc	0.98	-0.09	0.94	0.00	0.02
ERM	Env 2 acc	0.91	-0.00	0.95	0.00	0.02
GroupDRO	Env 2 acc	0.88	0.08	0.91	0.00	0.02
IB_IRM	Env 2 acc	0.90	-0.06	0.99	0.00	0.01
IRM	Env 2 acc	0.89	-0.08	0.98	0.00	0.01
MMD	Env 2 acc	0.85	0.13	0.94	0.00	0.02
VREx	Env 2 acc	0.91	0.01	0.93	0.00	0.02

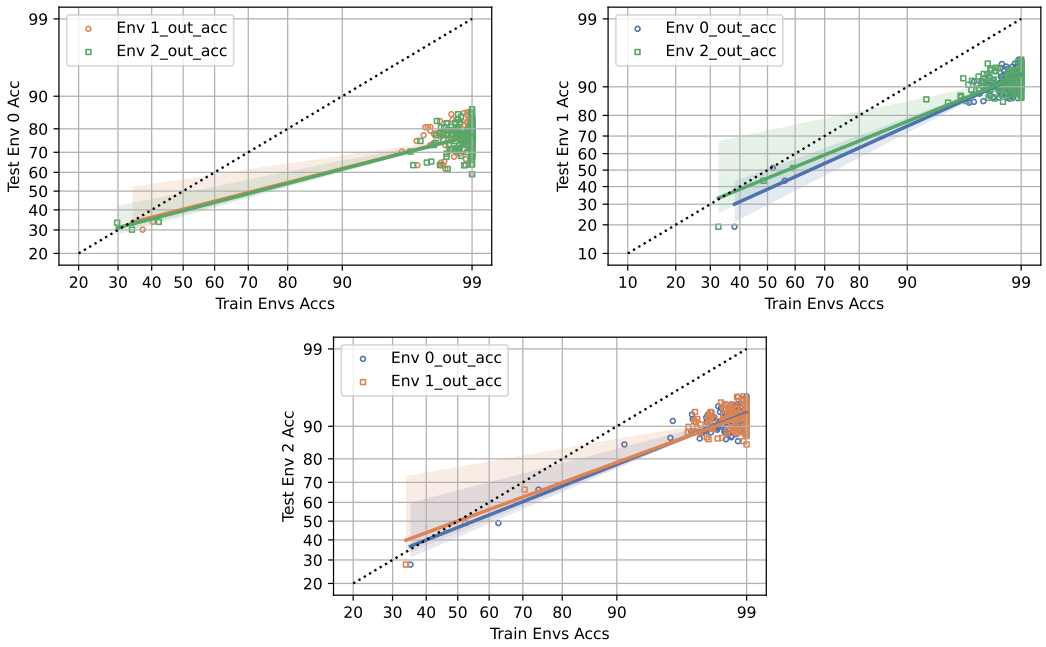


Figure 10: Spawrious-M2M-Easy correlations between in-distribution vs. out-of-distribution model accuracy for different training domains.

Table 9: SpawriousM2M_easy ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 1 acc	Env 0 acc	0.43	-0.24	0.69	0.00	0.04
Env 2 acc	Env 0 acc	0.44	-0.26	0.71	0.00	0.04
Env 0 acc	Env 1 acc	0.76	-0.29	0.89	0.00	0.03
Env 2 acc	Env 1 acc	0.68	-0.12	0.85	0.00	0.03
Env 0 acc	Env 2 acc	0.67	-0.09	0.86	0.00	0.03
Env 1 acc	Env 2 acc	0.62	0.00	0.83	0.00	0.03

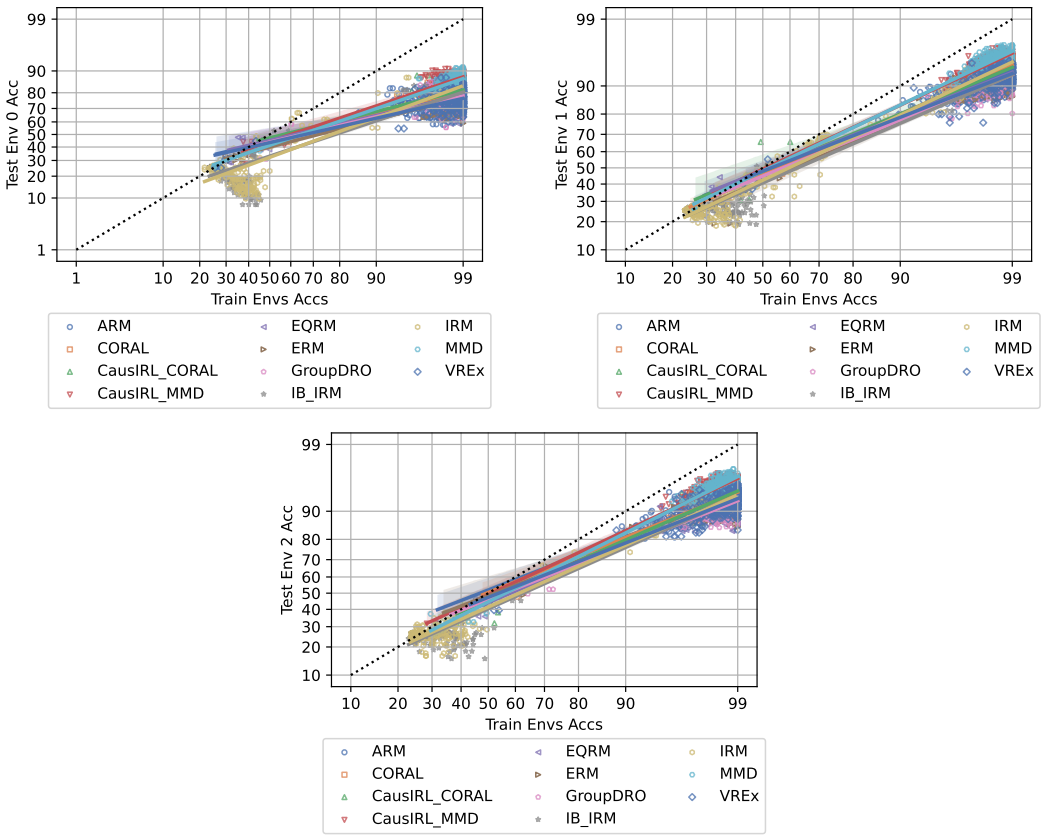


Figure 11: Spawrious-M2M-Easy correlations between in-distribution vs. out-of-distribution model accuracy for different domain generalization algorithms.

Table 10: SpawriousM2M_easy ID vs. OOD properties.

Algorithm	OOD	slope	intercept	Pearson R	p-value	standard error
ARM	Env 0 acc	0.40	-0.14	0.61	0.00	0.03
CORAL	Env 0 acc	0.44	-0.14	0.72	0.00	0.02
CausIRL_CORAL	Env 0 acc	0.41	-0.07	0.69	0.00	0.02
CausIRL_MMD	Env 0 acc	0.56	-0.14	0.85	0.00	0.02
EQRM	Env 0 acc	0.39	-0.12	0.72	0.00	0.02
ERM	Env 0 acc	0.43	-0.25	0.70	0.00	0.03
GroupDRO	Env 0 acc	0.40	-0.15	0.62	0.00	0.03
IB_IRM	Env 0 acc	0.55	-0.44	0.87	0.00	0.02
IRM	Env 0 acc	0.62	-0.46	0.88	0.00	0.02
MMD	Env 0 acc	0.58	-0.22	0.90	0.00	0.02
VREx	Env 0 acc	0.38	-0.16	0.66	0.00	0.03
ARM	Env 1 acc	0.68	-0.06	0.89	0.00	0.02
CORAL	Env 1 acc	0.74	-0.11	0.94	0.00	0.02
CausIRL_CORAL	Env 1 acc	0.70	-0.05	0.92	0.00	0.02
CausIRL_MMD	Env 1 acc	0.78	-0.02	0.95	0.00	0.01
EQRM	Env 1 acc	0.66	-0.04	0.92	0.00	0.02
ERM	Env 1 acc	0.72	-0.20	0.87	0.00	0.02
GroupDRO	Env 1 acc	0.68	-0.13	0.88	0.00	0.02
IB_IRM	Env 1 acc	0.73	-0.24	0.97	0.00	0.01
IRM	Env 1 acc	0.79	-0.19	0.97	0.00	0.01
MMD	Env 1 acc	0.81	-0.05	0.96	0.00	0.01
VREx	Env 1 acc	0.66	-0.05	0.86	0.00	0.02
ARM	Env 2 acc	0.65	0.05	0.89	0.00	0.02
CORAL	Env 2 acc	0.67	0.01	0.90	0.00	0.02
CausIRL_CORAL	Env 2 acc	0.73	-0.11	0.91	0.00	0.02
CausIRL_MMD	Env 2 acc	0.78	-0.02	0.93	0.00	0.02
EQRM	Env 2 acc	0.70	-0.15	0.90	0.00	0.02
ERM	Env 2 acc	0.64	-0.04	0.84	0.00	0.02
GroupDRO	Env 2 acc	0.66	-0.10	0.82	0.00	0.03
IB_IRM	Env 2 acc	0.75	-0.25	0.97	0.00	0.01
IRM	Env 2 acc	0.75	-0.22	0.97	0.00	0.01
MMD	Env 2 acc	0.84	-0.14	0.95	0.00	0.02
VREx	Env 2 acc	0.66	-0.07	0.86	0.00	0.02

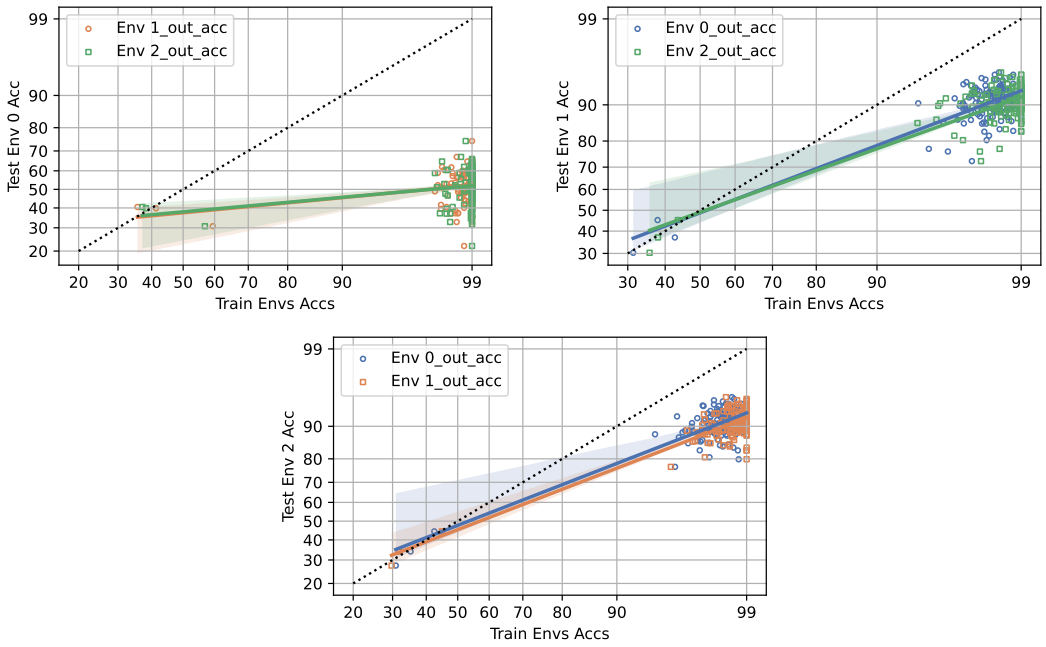


Figure 12: Spawrious-M2M-Hard correlations between in-distribution vs. out-of-distribution model accuracy for different training domains.

Table 11: SpawriousM2M_hard ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 1 acc	Env 0 acc	0.16	-0.32	0.24	0.00	0.05
Env 2 acc	Env 0 acc	0.15	-0.31	0.24	0.00	0.05
Env 0 acc	Env 1 acc	0.64	-0.03	0.80	0.00	0.04
Env 2 acc	Env 1 acc	0.60	-0.02	0.78	0.00	0.04
Env 0 acc	Env 2 acc	0.65	-0.06	0.83	0.00	0.04
Env 1 acc	Env 2 acc	0.65	-0.12	0.85	0.00	0.03

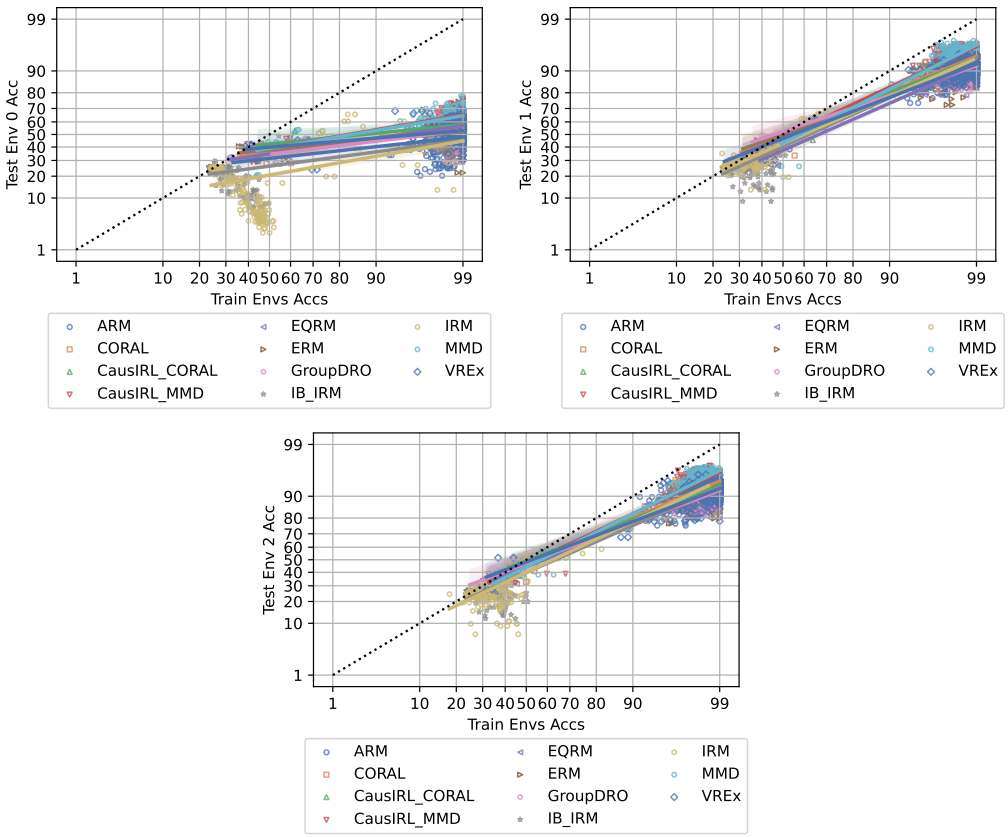


Figure 13: Spawrious-M2M-Hard correlations between in-distribution vs. out-of-distribution model accuracy for different domain generalization algorithms.

Table 12: SpawriousM2M_hard ID vs. OOD properties.

Algorithm	OOD	slope	intercept	Pearson R	p-value	standard error
ARM	Env 0 acc	0.18	-0.48	0.27	0.00	0.04
CORAL	Env 0 acc	0.23	-0.37	0.43	0.00	0.03
CausIRL_CORAL	Env 0 acc	0.17	-0.19	0.26	0.00	0.04
CausIRL_MMD	Env 0 acc	0.29	-0.27	0.59	0.00	0.02
EQRM	Env 0 acc	0.25	-0.39	0.46	0.00	0.03
ERM	Env 0 acc	0.16	-0.31	0.24	0.00	0.04
GroupDRO	Env 0 acc	0.21	-0.40	0.29	0.00	0.04
IB_IRM	Env 0 acc	0.22	-0.64	0.56	0.00	0.02
IRM	Env 0 acc	0.30	-0.82	0.48	0.00	0.03
MMD	Env 0 acc	0.28	-0.28	0.52	0.00	0.03
VREx	Env 0 acc	0.14	-0.26	0.26	0.00	0.03
ARM	Env 1 acc	0.66	-0.08	0.84	0.00	0.02
CORAL	Env 1 acc	0.69	-0.02	0.88	0.00	0.02
CausIRL_CORAL	Env 1 acc	0.67	-0.02	0.85	0.00	0.02
CausIRL_MMD	Env 1 acc	0.78	-0.05	0.91	0.00	0.02
EQRM	Env 1 acc	0.74	-0.32	0.90	0.00	0.02
ERM	Env 1 acc	0.61	0.00	0.78	0.00	0.03
GroupDRO	Env 1 acc	0.57	0.06	0.78	0.00	0.03
IB_IRM	Env 1 acc	0.77	-0.24	0.95	0.00	0.01
IRM	Env 1 acc	0.74	-0.18	0.97	0.00	0.01
MMD	Env 1 acc	0.85	-0.17	0.90	0.00	0.02
VREx	Env 1 acc	0.64	-0.08	0.85	0.00	0.02
ARM	Env 2 acc	0.66	-0.04	0.82	0.00	0.03
CORAL	Env 2 acc	0.71	-0.07	0.87	0.00	0.02
CausIRL_CORAL	Env 2 acc	0.66	-0.03	0.81	0.00	0.03
CausIRL_MMD	Env 2 acc	0.80	-0.10	0.89	0.00	0.02
EQRM	Env 2 acc	0.66	-0.14	0.88	0.00	0.02
ERM	Env 2 acc	0.64	-0.06	0.83	0.00	0.02
GroupDRO	Env 2 acc	0.63	-0.07	0.81	0.00	0.03
IB_IRM	Env 2 acc	0.74	-0.26	0.95	0.00	0.01
IRM	Env 2 acc	0.80	-0.25	0.93	0.00	0.02
MMD	Env 2 acc	0.84	-0.15	0.90	0.00	0.02
VREx	Env 2 acc	0.64	-0.04	0.81	0.00	0.03

PACS [29]. The PACS domain generalization benchmark consists of 9,991 images of dimensions (3, 224, 224) and seven classes $c \in \{dog, elephant, giraffe, guitar, horse, house, person\}$. This dataset is comprised of four domains $d \in \{art, cartoons, photos, sketches\}$ and is evaluated in a leave-one-domain-out fashion.

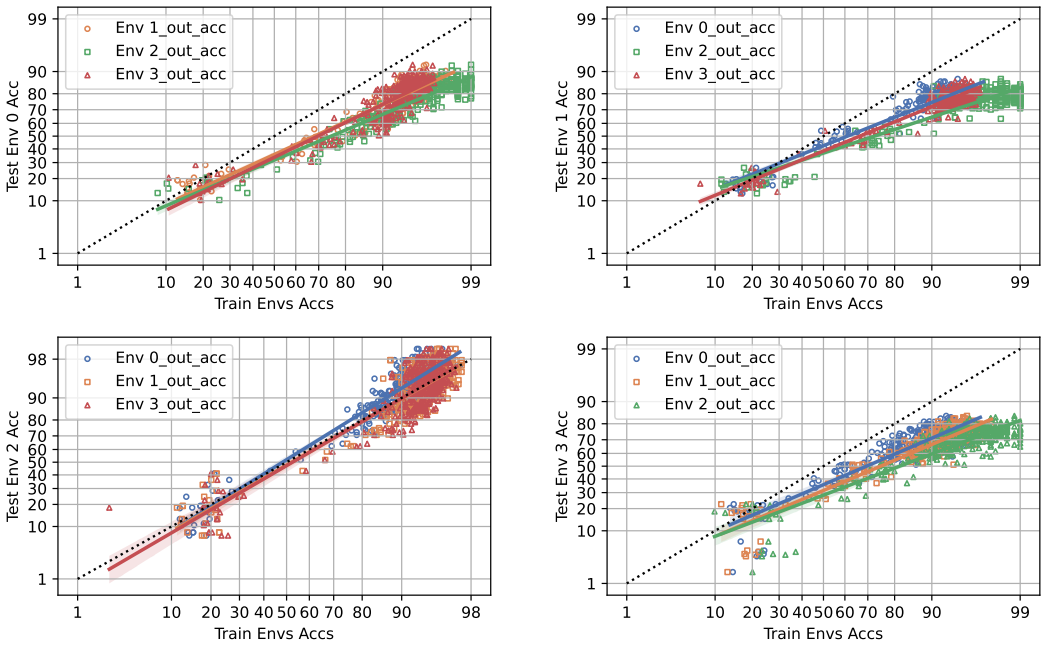


Figure 14: PACS correlations between in-distribution vs. out-of-distribution model accuracy for different training domains.

Table 13: PACS ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 1 acc	Env 0 acc	0.77	-0.36	0.94	0.00	0.02
Env 2 acc	Env 0 acc	0.71	-0.47	0.95	0.00	0.01
Env 3 acc	Env 0 acc	0.82	-0.42	0.92	0.00	0.02
Env 0 acc	Env 1 acc	0.67	-0.19	0.95	0.00	0.01
Env 2 acc	Env 1 acc	0.56	-0.34	0.92	0.00	0.01
Env 3 acc	Env 1 acc	0.69	-0.29	0.96	0.00	0.01
Env 0 acc	Env 2 acc	1.11	0.05	0.94	0.00	0.02
Env 1 acc	Env 2 acc	1.03	-0.06	0.91	0.00	0.02
Env 3 acc	Env 2 acc	1.05	-0.07	0.89	0.00	0.03
Env 0 acc	Env 3 acc	0.72	-0.37	0.91	0.00	0.02
Env 1 acc	Env 3 acc	0.71	-0.48	0.93	0.00	0.01
Env 2 acc	Env 3 acc	0.64	-0.57	0.90	0.00	0.02

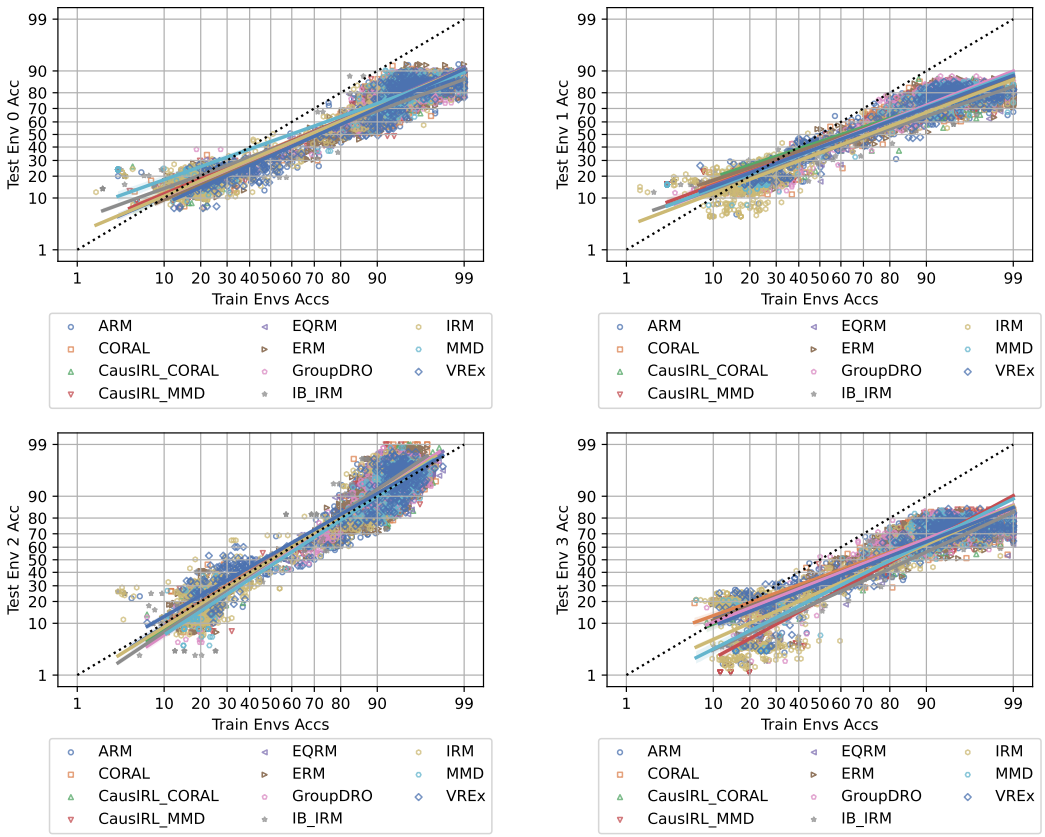


Figure 15: PACS correlations between in-distribution vs. out-of-distribution model accuracy for different domain generalization algorithms.

Table 14: PACS ID vs. OOD properties.

Algorithm	OOD	slope	intercept	Pearson R	p-value	standard error
ARM	Env 0 acc	0.75	-0.38	0.90	0.00	0.01
CORAL	Env 0 acc	0.70	-0.30	0.90	0.00	0.01
CausIRL_CORAL	Env 0 acc	0.71	-0.32	0.92	0.00	0.01
CausIRL_MMD	Env 0 acc	0.69	-0.31	0.96	0.00	0.01
EQRM	Env 0 acc	0.73	-0.34	0.94	0.00	0.01
ERM	Env 0 acc	0.73	-0.38	0.92	0.00	0.01
GroupDRO	Env 0 acc	0.70	-0.33	0.90	0.00	0.01
IB_IRM	Env 0 acc	0.61	-0.31	0.93	0.00	0.01
IRM	Env 0 acc	0.71	-0.33	0.97	0.00	0.01
MMD	Env 0 acc	0.60	-0.14	0.93	0.00	0.01
VREx	Env 0 acc	0.75	-0.43	0.96	0.00	0.01
ARM	Env 1 acc	0.60	-0.22	0.90	0.00	0.01
CORAL	Env 1 acc	0.60	-0.22	0.90	0.00	0.01
CausIRL_CORAL	Env 1 acc	0.56	-0.15	0.86	0.00	0.02
CausIRL_MMD	Env 1 acc	0.62	-0.23	0.94	0.00	0.01
EQRM	Env 1 acc	0.61	-0.25	0.92	0.00	0.01
ERM	Env 1 acc	0.57	-0.19	0.89	0.00	0.01
GroupDRO	Env 1 acc	0.65	-0.25	0.91	0.00	0.01
IB_IRM	Env 1 acc	0.58	-0.37	0.95	0.00	0.01
IRM	Env 1 acc	0.63	-0.37	0.94	0.00	0.01
MMD	Env 1 acc	0.64	-0.26	0.95	0.00	0.01
VREx	Env 1 acc	0.62	-0.26	0.95	0.00	0.01
ARM	Env 2 acc	1.06	-0.02	0.93	0.00	0.01
CORAL	Env 2 acc	1.05	0.03	0.93	0.00	0.01
CausIRL_CORAL	Env 2 acc	1.07	-0.03	0.91	0.00	0.02
CausIRL_MMD	Env 2 acc	1.09	-0.09	0.97	0.00	0.01
EQRM	Env 2 acc	1.05	0.01	0.94	0.00	0.02
ERM	Env 2 acc	1.04	0.00	0.91	0.00	0.02
GroupDRO	Env 2 acc	1.12	-0.08	0.93	0.00	0.01
IB_IRM	Env 2 acc	1.12	-0.02	0.96	0.00	0.02
IRM	Env 2 acc	1.05	-0.01	0.95	0.00	0.01
MMD	Env 2 acc	1.09	-0.12	0.97	0.00	0.01
VREx	Env 2 acc	0.97	0.10	0.96	0.00	0.01
ARM	Env 3 acc	0.62	-0.38	0.87	0.00	0.01
CORAL	Env 3 acc	0.60	-0.35	0.89	0.00	0.01
CausIRL_CORAL	Env 3 acc	0.66	-0.45	0.90	0.00	0.01
CausIRL_MMD	Env 3 acc	0.91	-0.82	0.94	0.00	0.01
EQRM	Env 3 acc	0.65	-0.46	0.89	0.00	0.02
ERM	Env 3 acc	0.64	-0.41	0.88	0.00	0.01
GroupDRO	Env 3 acc	0.66	-0.42	0.88	0.00	0.01
IB_IRM	Env 3 acc	0.77	-0.83	0.85	0.00	0.02
IRM	Env 3 acc	0.74	-0.65	0.88	0.00	0.01
MMD	Env 3 acc	0.84	-0.72	0.93	0.00	0.01
VREx	Env 3 acc	0.66	-0.48	0.93	0.00	0.01

B.5 WILDS Results

Terra Incognita [9, 27]. The Terra Incognita dataset contains photographs of wild animals taken by camera traps at locations $d \in \{L100, L38, L43, L46\}$. This dataset contains 24,788 examples of dimensions (3, 224, 224) and 10 classes for the image classification task.

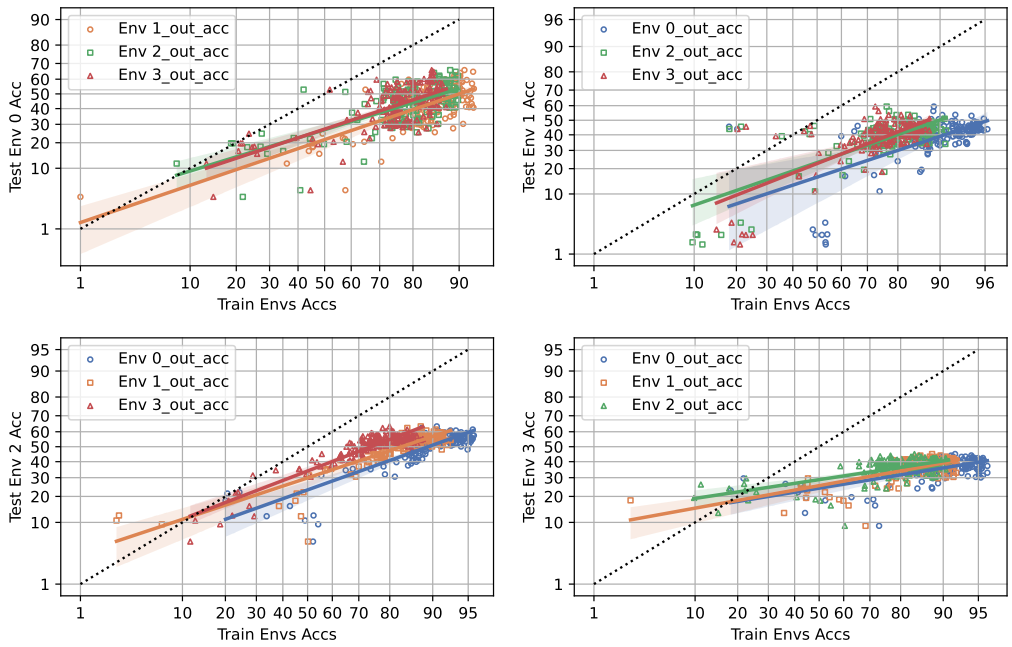


Figure 16: Terra Incognita correlations between in-distribution vs. out-of-distribution model accuracy for different training domains.

Table 15: TerraIncognita ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 1 acc	Env 0 acc	0.61	-0.79	0.72	0.00	0.05
Env 2 acc	Env 0 acc	0.56	-0.62	0.71	0.00	0.04
Env 3 acc	Env 0 acc	0.60	-0.61	0.71	0.00	0.05
Env 0 acc	Env 1 acc	0.55	-0.99	0.61	0.00	0.06
Env 2 acc	Env 1 acc	0.58	-0.74	0.75	0.00	0.04
Env 3 acc	Env 1 acc	0.65	-0.75	0.73	0.00	0.05
Env 0 acc	Env 2 acc	0.59	-0.73	0.85	0.00	0.03
Env 1 acc	Env 2 acc	0.55	-0.53	0.88	0.00	0.02
Env 3 acc	Env 2 acc	0.63	-0.41	0.92	0.00	0.02
Env 0 acc	Env 3 acc	0.27	-0.70	0.64	0.00	0.03
Env 1 acc	Env 3 acc	0.30	-0.66	0.66	0.00	0.03
Env 2 acc	Env 3 acc	0.25	-0.56	0.63	0.00	0.02

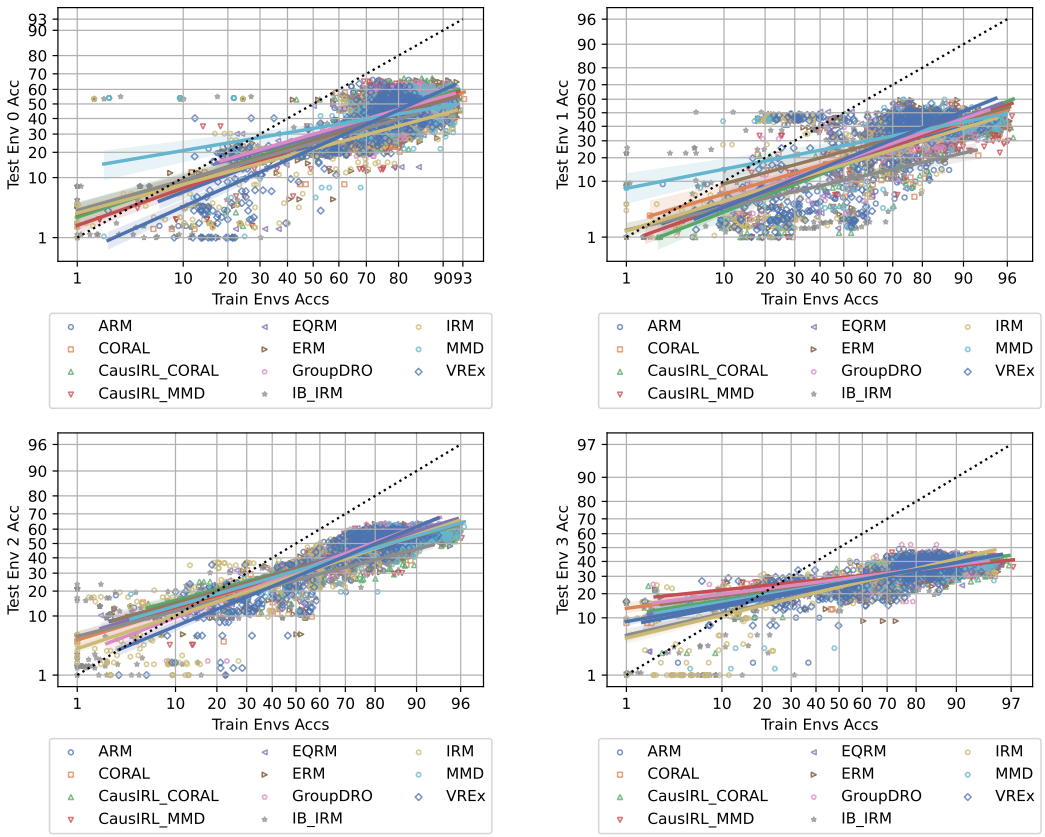


Figure 17: Terra Incognita correlations between in-distribution vs. out-of-distribution model accuracy for different domain generalization algorithms.

Table 16: TerraIncognita ID vs. OOD properties.

Algorithm	OOD	slope	intercept	Pearson R	p-value	standard error
ARM	Env 0 acc	0.68	-0.66	0.74	0.00	0.03
CORAL	Env 0 acc	0.56	-0.63	0.73	0.00	0.02
CausIRL_CORAL	Env 0 acc	0.59	-0.60	0.73	0.00	0.03
CausIRL_MMD	Env 0 acc	0.62	-0.67	0.83	0.00	0.02
EQRM	Env 0 acc	0.49	-0.57	0.68	0.00	0.02
ERM	Env 0 acc	0.55	-0.64	0.69	0.00	0.03
GroupDRO	Env 0 acc	0.51	-0.53	0.67	0.00	0.03
IB_IRM	Env 0 acc	0.50	-0.69	0.63	0.00	0.03
IRM	Env 0 acc	0.48	-0.79	0.70	0.00	0.02
MMD	Env 0 acc	0.30	-0.43	0.51	0.00	0.02
VREx	Env 0 acc	0.79	-0.77	0.85	0.00	0.02
ARM	Env 1 acc	0.59	-0.93	0.62	0.00	0.03
CORAL	Env 1 acc	0.52	-0.85	0.68	0.00	0.03
CausIRL_CORAL	Env 1 acc	0.68	-0.98	0.76	0.00	0.03
CausIRL_MMD	Env 1 acc	0.63	-0.95	0.78	0.00	0.02
EQRM	Env 1 acc	0.65	-0.92	0.70	0.00	0.03
ERM	Env 1 acc	0.47	-0.73	0.62	0.00	0.03
GroupDRO	Env 1 acc	0.61	-0.89	0.70	0.00	0.03
IB_IRM	Env 1 acc	0.40	-1.27	0.42	0.00	0.04
IRM	Env 1 acc	0.53	-0.97	0.52	0.00	0.04
MMD	Env 1 acc	0.34	-0.61	0.49	0.00	0.03
VREx	Env 1 acc	0.70	-0.87	0.72	0.00	0.03
ARM	Env 2 acc	0.46	-0.47	0.79	0.00	0.02
CORAL	Env 2 acc	0.52	-0.51	0.84	0.00	0.01
CausIRL_CORAL	Env 2 acc	0.44	-0.44	0.78	0.00	0.02
CausIRL_MMD	Env 2 acc	0.49	-0.50	0.88	0.00	0.01
EQRM	Env 2 acc	0.51	-0.44	0.88	0.00	0.01
ERM	Env 2 acc	0.47	-0.45	0.79	0.00	0.02
GroupDRO	Env 2 acc	0.63	-0.51	0.86	0.00	0.02
IB_IRM	Env 2 acc	0.43	-0.65	0.75	0.00	0.02
IRM	Env 2 acc	0.56	-0.56	0.85	0.00	0.02
MMD	Env 2 acc	0.48	-0.48	0.87	0.00	0.01
VREx	Env 2 acc	0.69	-0.58	0.89	0.00	0.02
ARM	Env 3 acc	0.32	-0.67	0.66	0.00	0.02
CORAL	Env 3 acc	0.22	-0.59	0.69	0.00	0.01
CausIRL_CORAL	Env 3 acc	0.26	-0.63	0.69	0.00	0.01
CausIRL_MMD	Env 3 acc	0.17	-0.55	0.76	0.00	0.01
EQRM	Env 3 acc	0.28	-0.63	0.78	0.00	0.01
ERM	Env 3 acc	0.23	-0.59	0.59	0.00	0.01
GroupDRO	Env 3 acc	0.21	-0.58	0.56	0.00	0.01
IB_IRM	Env 3 acc	0.38	-0.72	0.69	0.00	0.02
IRM	Env 3 acc	0.40	-0.72	0.74	0.00	0.02
MMD	Env 3 acc	0.28	-0.64	0.74	0.00	0.01
VREx	Env 3 acc	0.30	-0.66	0.77	0.00	0.01

Camyleon [5, 27]. The Camelyon17-wilds dataset contains histopathological images of lymph node tissue collected from two hospitals, denoted as Hospital A, Hospital B. This dataset contains 327,680 examples of dimensions (3, 96, 96) and 2 classes (tumor, non-tumor).

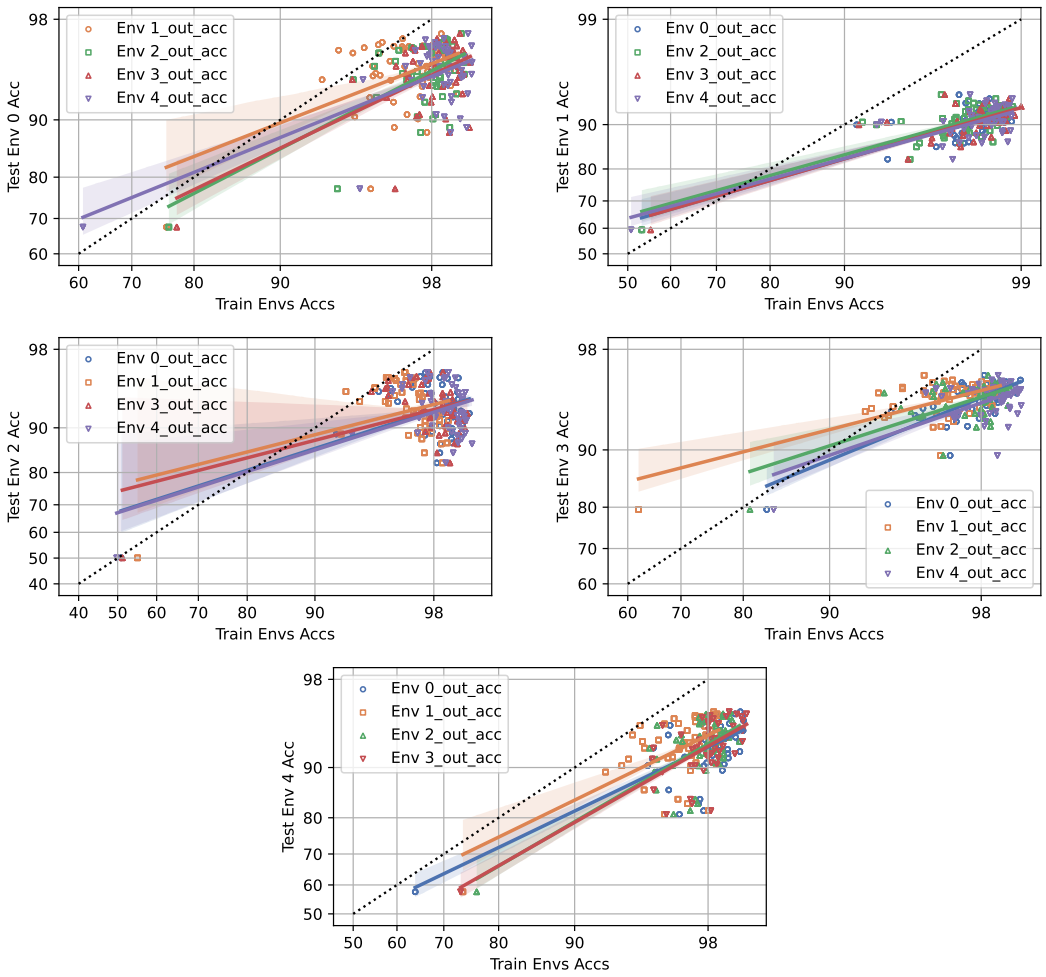


Figure 18: Camelyon correlations between in-distribution vs. out-of-distribution model accuracy for different training domains.

Table 17: WILDSCamelyon ID vs. OOD properties.

ID	OOD	slope	intercept	Pearson R	p-value	standard error
Env 1 acc	Env 0 acc	0.59	0.51	0.49	0.00	0.07
Env 2 acc	Env 0 acc	0.78	0.06	0.64	0.00	0.06
Env 3 acc	Env 0 acc	0.73	0.14	0.58	0.00	0.06
Env 4 acc	Env 0 acc	0.61	0.37	0.61	0.00	0.05
Env 0 acc	Env 1 acc	0.49	0.32	0.79	0.00	0.02
Env 2 acc	Env 1 acc	0.47	0.38	0.76	0.00	0.03
Env 3 acc	Env 1 acc	0.49	0.32	0.79	0.00	0.02
Env 4 acc	Env 1 acc	0.46	0.35	0.76	0.00	0.02
Env 0 acc	Env 2 acc	0.49	0.45	0.53	0.00	0.05
Env 1 acc	Env 2 acc	0.39	0.72	0.37	0.00	0.06
Env 3 acc	Env 2 acc	0.39	0.66	0.42	0.00	0.05
Env 4 acc	Env 2 acc	0.48	0.45	0.52	0.00	0.05
Env 0 acc	Env 3 acc	0.62	0.41	0.77	0.00	0.03
Env 1 acc	Env 3 acc	0.39	0.94	0.65	0.00	0.03
Env 2 acc	Env 3 acc	0.49	0.69	0.64	0.00	0.04
Env 4 acc	Env 3 acc	0.52	0.57	0.61	0.00	0.04
Env 0 acc	Env 4 acc	0.73	-0.03	0.71	0.00	0.05
Env 1 acc	Env 4 acc	0.74	0.05	0.62	0.00	0.06
Env 2 acc	Env 4 acc	0.88	-0.32	0.70	0.00	0.06
Env 3 acc	Env 4 acc	0.86	-0.30	0.74	0.00	0.05

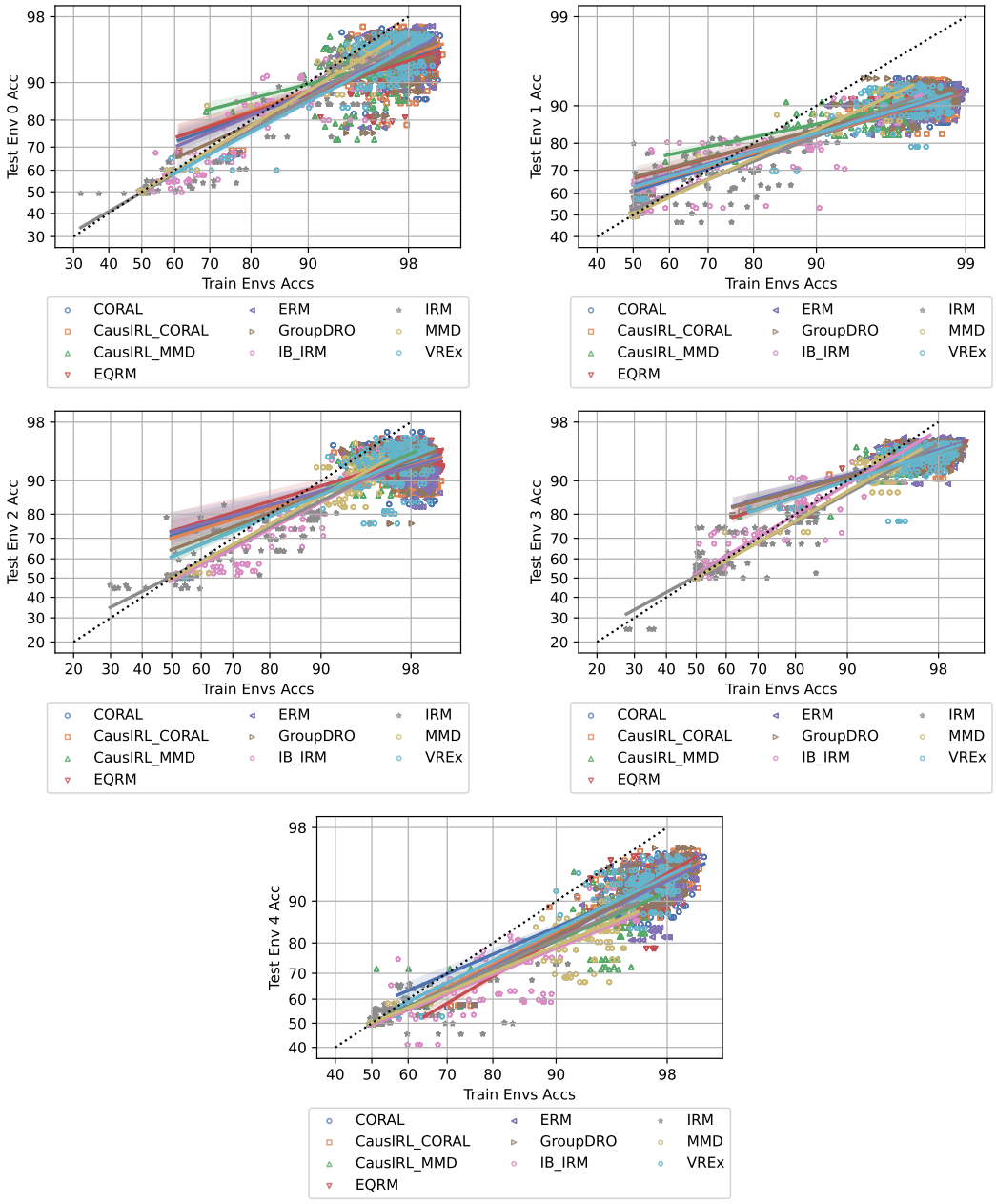


Figure 19: Camelyon correlations between in-distribution vs. out-of-distribution model accuracy for different domain generalization algorithms.

Table 18: WILDSCamelyon ID vs. OOD properties.

Algorithm	OOD	slope	intercept	Pearson R	p-value	standard error
CORAL	Env 0 acc	0.54	0.45	0.49	0.00	0.03
CausIRL_CORAL	Env 0 acc	0.54	0.48	0.49	0.00	0.03
CausIRL_MMD	Env 0 acc	0.39	0.76	0.30	0.00	0.04
EQRM	Env 0 acc	0.51	0.48	0.57	0.00	0.02
ERM	Env 0 acc	0.61	0.39	0.56	0.00	0.03
GroupDRO	Env 0 acc	0.68	0.22	0.57	0.00	0.03
IB_IRM	Env 0 acc	0.86	-0.00	0.96	0.00	0.01
IRM	Env 0 acc	0.87	-0.01	0.97	0.00	0.01
MMD	Env 0 acc	0.92	0.00	0.99	0.00	0.00
VREx	Env 0 acc	0.81	0.01	0.78	0.00	0.02
CORAL	Env 1 acc	0.50	0.28	0.77	0.00	0.01
CausIRL_CORAL	Env 1 acc	0.48	0.32	0.76	0.00	0.01
CausIRL_MMD	Env 1 acc	0.35	0.62	0.58	0.00	0.02
EQRM	Env 1 acc	0.45	0.42	0.84	0.00	0.01
ERM	Env 1 acc	0.47	0.35	0.78	0.00	0.01
GroupDRO	Env 1 acc	0.43	0.46	0.70	0.00	0.01
IB_IRM	Env 1 acc	0.66	0.07	0.90	0.00	0.01
IRM	Env 1 acc	0.67	0.07	0.89	0.00	0.01
MMD	Env 1 acc	0.77	0.01	0.99	0.00	0.00
VREx	Env 1 acc	0.50	0.32	0.72	0.00	0.01
CORAL	Env 2 acc	0.45	0.57	0.50	0.00	0.02
CausIRL_CORAL	Env 2 acc	0.46	0.53	0.56	0.00	0.02
CausIRL_MMD	Env 2 acc	0.66	0.28	0.73	0.00	0.02
EQRM	Env 2 acc	0.46	0.62	0.63	0.00	0.02
ERM	Env 2 acc	0.42	0.60	0.46	0.00	0.03
GroupDRO	Env 2 acc	0.57	0.38	0.65	0.00	0.02
IB_IRM	Env 2 acc	0.80	-0.03	0.95	0.00	0.01
IRM	Env 2 acc	0.77	0.02	0.93	0.00	0.01
MMD	Env 2 acc	0.84	-0.00	0.98	0.00	0.00
VREx	Env 2 acc	0.63	0.29	0.64	0.00	0.02
CORAL	Env 3 acc	0.41	0.83	0.69	0.00	0.01
CausIRL_CORAL	Env 3 acc	0.43	0.78	0.67	0.00	0.01
CausIRL_MMD	Env 3 acc	0.50	0.65	0.65	0.00	0.02
EQRM	Env 3 acc	0.50	0.66	0.83	0.00	0.01
ERM	Env 3 acc	0.43	0.81	0.63	0.00	0.02
GroupDRO	Env 3 acc	0.42	0.79	0.63	0.00	0.02
IB_IRM	Env 3 acc	0.92	0.06	0.96	0.00	0.01
IRM	Env 3 acc	0.84	0.03	0.94	0.00	0.01
MMD	Env 3 acc	0.89	0.01	0.99	0.00	0.00
VREx	Env 3 acc	0.49	0.65	0.62	0.00	0.02
CORAL	Env 4 acc	0.64	0.18	0.74	0.00	0.02
CausIRL_CORAL	Env 4 acc	0.75	-0.01	0.81	0.00	0.02
CausIRL_MMD	Env 4 acc	0.67	0.01	0.94	0.00	0.01
EQRM	Env 4 acc	0.89	-0.26	0.82	0.00	0.02
ERM	Env 4 acc	0.75	-0.04	0.67	0.00	0.03
GroupDRO	Env 4 acc	0.80	-0.11	0.80	0.00	0.02
IB_IRM	Env 4 acc	0.63	-0.02	0.90	0.00	0.01
IRM	Env 4 acc	0.68	-0.00	0.92	0.00	0.01
MMD	Env 4 acc	0.63	0.00	0.94	0.00	0.01
VREx	Env 4 acc	0.73	0.04	0.82	0.00	0.02

C Theoretical Results

C.1 Proof of Theorem 1

Theorem 3. *WLOG, let $P_{ID} = \mu(I, I)$, generated by Equation 2 and denote $P_{OOD} = \mu(M, \Lambda)$ as an arbitrary target domain, parameterized by interventions M, Λ, where $M \in \mathbb{R}^{k\times k}$ and $\Lambda \in \mathbb{R}^{k\times k} \succ 0$ and Λ is symmetric. Let $\mathcal{E}_{train} = \{P_{ID}\}$ and $\mathcal{E}_{test} = \{P_{OOD}\}$. Let $\mathcal{F}$ be the class of linear classifiers of the form $Z_c \cdot \beta_c + Z_e \cdot \beta_e$. We then consider two models $f_X \in \mathcal{F}\backslash\mathcal{F}_c$ and $f_c^* \in \mathcal{F}_c$, Definition 2-3.*

$$\max_{f\in\mathcal{F}\backslash\mathcal{F}_c} acc_{OOD}(f_X) < acc_{OOD}(f_c^*) \tag{14}$$

if and only if

$$\frac{p(\mu_c^T\Sigma_c^{-1}\mu_c) + \alpha(\Sigma_e^{-1}\mu_e)^T M\mu_e}{\sqrt{(\mu_c^T\Sigma_c^{-1}\mu_c) + p(1-p)(\mu_c^T\Sigma_c^{-1}\mu_c)^2 + (\mu_e^T\Sigma_e^{-1}\Lambda\mu_e) + \alpha(1-\alpha)((\Sigma_e^{-1}\mu_e)^T M\mu_e)^2}} < \tag{15}$$

$$\frac{p(\mu_c^T\Sigma_c^{-1}\mu_c)}{\sqrt{(\mu_c^T\Sigma_c^{-1}\mu_c) + p(1-p)(\mu_c^T\Sigma_c^{-1}\mu_c)^2}}$$

where $\alpha = pq + (1-p)(1-q) > 0$. All variables besides M and Λ are fixed for a given setting.

Two conditions for Equation 15 to hold are:

1. Spurious Correlation Reversal.
$$(\Sigma_e^{-1}\mu_e)^T M\mu_e < 0 \tag{16}$$
2. Sufficient Decrease in Signal-to-Noise Ratio. *Specifically referring to $\alpha(\Sigma_e^{-1}\mu_e)^T M\mu_e$ and $(\mu_e^T\Sigma_e^{-1}\Lambda\mu_e) + \alpha(1-\alpha)((\Sigma_e^{-1}\mu_e)^T M\mu_e)^2$, respectively.*

Proof. Let

$$\alpha = pq + (1-p)(1-q),$$

The distribution of Z_{c} is:

$$Z_{\mathrm{c}} \sim p\cdot\mathcal{N}(\mu_c, \Sigma_{\mathrm{c}}) + (1-p)\cdot\mathcal{N}(0, \Sigma_{\mathrm{c}})$$

The distribution of Z_{e} is:

$$Z_{\mathrm{e}} \sim (2pq + 1 - p - q)\cdot\mathcal{N}(M\mu_e, \Lambda\Sigma_{\mathrm{e}}) + (p + q - 2pq)\cdot\mathcal{N}(0, \Lambda\Sigma_{\mathrm{e}})$$

Let $\mathrm{w_c} = \Sigma_{\mathrm{c}}^{-1}\mu_c$ and $\mathrm{w_e} = \Sigma_{\mathrm{e}}^{-1}\mu_e$. For the linear combinations, we have:

$$\mathrm{w_c}^T Z_{\mathrm{c}} = (\Sigma_{\mathrm{c}}^{-1}\mu_c)^T Z_{\mathrm{c}}$$
$$\mathrm{w_e}^T Z_{\mathrm{e}} = (\Sigma_{\mathrm{e}}^{-1}\mu_e)^T Z_{\mathrm{e}}$$

The distribution of $\mathrm{w_c}^T Z_{\mathrm{c}}$ is:

$$\mathrm{w_c}^T Z_{\mathrm{c}} \sim \mathcal{N}\left(p\mu_c^T\Sigma_{\mathrm{c}}^{-1}\mu_c, (\mu_c^T\Sigma_{\mathrm{c}}^{-1}\mu_c) + p(1-p)(\mu_c^T\Sigma_{\mathrm{c}}^{-1}\mu_c)^2\right)$$

The distribution of $\mathrm{w_e}^T Z_{\mathrm{e}}$ is:

$$\mathrm{w_e}^T Z_{\mathrm{e}} \sim \mathcal{N}\left(\alpha(\Sigma_{\mathrm{e}}^{-1}\mu_e)^T M\mu_e, (\mu_e^T\Sigma_{\mathrm{e}}^{-1}\Lambda\mu_e) + \alpha(1-\alpha)((\Sigma_{\mathrm{e}}^{-1}\mu_e)^T M\mu_e)^2\right)$$

Combining the two linear combinations, we get:

$$(\Sigma_{\mathrm{c}}^{-1}\mu_c)^T Z_{\mathrm{c}} + (\Sigma_{\mathrm{e}}^{-1}\mu_e)^T Z_{\mathrm{e}}$$

The mean and variance are:

$$\mathbb{E}[(\Sigma_{\text{c}}^{-1}\mu_c)^T Z_{\text{c}} + (\Sigma_{\text{e}}^{-1}\mu_e)^T Z_{\text{e}}] = p(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c) + \alpha(\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e$$

$$\text{Var}[(\Sigma_{\text{c}}^{-1}\mu_c)^T Z_{\text{c}} + (\Sigma_{\text{e}}^{-1}\mu_e)^T Z_{\text{e}}] = (\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c) + p(1-p)(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c)^2 + \tag{17}$$

$$(\mu_e^T \Sigma_{\text{e}}^{-1}\Lambda\mu_e) + \alpha(1-\alpha)((\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e)^2 \tag{18}$$

Let

$$N_c = p(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c),$$

$$N_e = \alpha(\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e,$$

$$D_c = (\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c) + p(1-p)(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c)^2$$

$$D_e = (\mu_e^T \Sigma_{\text{e}}^{-1}\Lambda\mu_e) + \alpha(1-\alpha)((\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e)^2$$

The expected accuracies are:

$$\text{acc}_{P_{\text{ID}}}(f_c^*) = \Phi\left(\frac{N_c}{\sqrt{D_c}}\right)$$

and

$$\text{acc}_{P_{\text{ID}}}(f_X^*) = \Phi\left(\frac{N_c + N_e}{\sqrt{D_c + D_e}}\right) \tag{19}$$

To compare the accuracies, we have:

$$\frac{N_c + N_e)}{\sqrt{D_c + D_e}} < \frac{N_c}{\sqrt{D_c}},$$

which, when written out, is:

$$\frac{p(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c) + \alpha(\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e}{\sqrt{(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c) + p(1-p)(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c)^2 + (\mu_e^T \Sigma_{\text{e}}^{-1}\Lambda\mu_e) + \alpha(1-\alpha)((\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e)^2}} <$$

$$\frac{p(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c)}{\sqrt{(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c) + p(1-p)(\mu_c^T \Sigma_{\text{c}}^{-1}\mu_c)^2}},$$

where $\alpha = pq + (1-p)(1-q) > 0$. All variables besides M and Λ are fixed for a given setting.

Final Conditions

1. *Effective Signal Contribution.*

   $$(\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e < 0$$

   or

2. *Sufficient Decrease in Signal-to-Noise Ratio.* Specifically referring to $\alpha(\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e$ and $(\mu_e^T \Sigma_{\text{e}}^{-1}\Lambda\mu_e) + \alpha(1-\alpha)((\Sigma_{\text{e}}^{-1}\mu_e)^T M\mu_e)^2$, respectively.

□

C.2 Proof of Theorem 2

Theorem 4 (Accuracy on the line)**.** *Let $P_{ID} = \mu(M_{ID}, \Lambda_{ID})$ and $P_{OOD} = \mu(M_{OOD}, \Lambda_{OOD})$.*

The correlation property, Definition 4, holds if and only if for any arbitrary classifiers, $[w_c, w_e]$,

$$\left| \frac{p w_c^T \mu_c + \alpha w_e^T M_{ID} \mu_e}{\sqrt{(w_c^T \Sigma_c^{-1} w_c) + p(1-p)(\mu_c^T w_c)^2 + (w_e^T \Lambda_{ID} \Sigma_e^{-1} w_e) + \alpha(1-\alpha)(w_e^T M_{ID} \mu_e)^2}} - c \cdot \frac{p w_c^T \mu_c + \alpha w_e^T M_{OOD} \mu_e}{\sqrt{(w_c^T \Sigma_c^{-1} w_c) + p(1-p)(\mu_c^T w_c)^2 + (w_e^T \Lambda_{OOD} \Sigma_e^{-1} w_e) + \alpha(1-\alpha)(w_e^T M_{OOD} \mu_e)^2}} \right| < \epsilon \quad (20)$$

where $c \in \mathbb{R}$, $\epsilon \geq 0$.

Proof. Define arbitrary $\mathrm{w_c}$ and $\mathrm{w_e}$.

Let

$$N_c = p \mathrm{w_c}^T \mu_c \quad (21)$$
$$N_e^{\mathrm{ID}} = \alpha \mathrm{w_e}^T M_{\mathrm{ID}} \mu_e \quad (22)$$
$$N_e^{\mathrm{OOD}} = \alpha \mathrm{w_e}^T M_{\mathrm{OOD}} \mu_e \quad (23)$$
$$D_c = (\mathrm{w_c}^T \Sigma_{\mathrm{c}}^{-1} \mathrm{w_c}) + p(1-p)(\mu_c^T \mathrm{w_c})^2 \quad (24)$$
$$D_e^{\mathrm{ID}} = (\mathrm{w_e}^T \Lambda_{\mathrm{ID}} \Sigma_{\mathrm{e}}^{-1} \mathrm{w_e}) + \alpha(1-\alpha)(\mathrm{w_e}^T M_{\mathrm{ID}} \mu_e)^2 \quad (25)$$
$$D_e^{\mathrm{OOD}} = (\mathrm{w_e}^T \Lambda_{\mathrm{OOD}} \Sigma_{\mathrm{e}}^{-1} \mathrm{w_e}) + \alpha(1-\alpha)(\mathrm{w_e}^T M_{\mathrm{OOD}} \mu_e)^2 \quad (26)$$

$$\mathrm{acc}_{P_{\mathrm{OOD}}}(f_X) = \Phi\left(\frac{N_c + N_e^{\mathrm{ID}}}{\sqrt{D_c + D_e^{\mathrm{ID}}}}\right) \qquad \mathrm{acc}_{P_{\mathrm{OOD}}}(f_X) = \Phi\left(\frac{N_c + N_e^{\mathrm{OOD}}}{\sqrt{D_c + D_e^{\mathrm{OOD}}}}\right)$$

$$\left| \Phi^{-1}\left(\mathrm{acc}_{P_{\mathrm{OOD}}}(f_X)\right) - c \cdot \Phi^{-1}\left(\mathrm{acc}_{P_{\mathrm{OOD}}}(f_X)\right) \right| < \epsilon \quad \forall f_X \iff$$
$$\left| \frac{N_c + N_e^{\mathrm{ID}}}{\sqrt{D_c + D_e^{\mathrm{ID}}}} - c \cdot \frac{N_c + N_e^{\mathrm{OOD}}}{\sqrt{D_c + D_e^{\mathrm{OOD}}}} \right| < \epsilon \quad \forall \mathrm{w_c}, \mathrm{w_e}$$

Thus, we have

$$\left| \frac{p \mathrm{w_c}^T \mu_c + \alpha \mathrm{w_e}^T M_{\mathrm{ID}} \mu_e}{\sqrt{(\mathrm{w_c}^T \Sigma_{\mathrm{c}}^{-1} \mathrm{w_c}) + p(1-p)(\mu_c^T \mathrm{w_c})^2 + (\mathrm{w_e}^T \Lambda_{\mathrm{ID}} \Sigma_{\mathrm{e}}^{-1} \mathrm{w_e}) + \alpha(1-\alpha)(\mathrm{w_e}^T M_{\mathrm{ID}} \mu_e)^2}} - c \cdot \frac{p \mathrm{w_c}^T \mu_c + \alpha \mathrm{w_e}^T M_{\mathrm{OOD}} \mu_e}{\sqrt{(\mathrm{w_c}^T \Sigma_{\mathrm{c}}^{-1} \mathrm{w_c}) + p(1-p)(\mu_c^T \mathrm{w_c})^2 + (\mathrm{w_e}^T \Lambda_{\mathrm{OOD}} \Sigma_{\mathrm{e}}^{-1} \mathrm{w_e}) + \alpha(1-\alpha)(\mathrm{w_e}^T M_{\mathrm{OOD}} \mu_e)^2}} \right| < \epsilon \quad (27)$$

where $c \in \mathbb{R}$, $\epsilon \geq 0$. □

C.2.1 [43]'s model

[43] consider a binary classification problem where the label y is distributed uniformly on $\{-1, 1\}$ both in the original distribution D and shifted distribution D_1. Conditional on y, we assume D such that $x \in \mathbb{R}^d$ is an isotropic Gaussian, i.e.,

$$x \mid y \sim \mathcal{N}(\mu \cdot y, \sigma^2 I_{d \times d}),$$

where $\mu \in \mathbb{R}^d$ is the mean vector and $\sigma^2 > 0$ is the variance. The distribution shift is modeled as a change in μ and σ, such that the shifted distribution D_1 corresponds to shifted parameters

$$\mu_1 = \alpha \cdot \mu + \beta \cdot \eta \quad \text{and} \quad \sigma_1 = \gamma \cdot \sigma,$$

where $\alpha, \beta, \gamma > 0$ are fixed scalars and η is uniformly distributed on the sphere in $\mathbb{R}^d$.

Theorem 5 ([43] Theorem 1)**.** *In the setting described above where η is independent of θ, let $\delta \in (0, 1)$. With probability at least $1 - \delta$, we have*

$$\left| \Phi^{-1}(acc_{D_1}(\theta)) - \frac{\alpha}{\gamma} \Phi^{-1}(acc_D(\theta)) \right| \leq \frac{\beta}{\gamma\sigma} \cdot \sqrt{\frac{2\log(2/\delta)}{d}},$$

where Φ is the standard normal cumulative distribution function and Φ^{-1} is its inverse (the probit transform), $acc_D(\theta)$ is the accuracy on the original distribution, and $acc_{D_1}(\theta)$ is the accuracy on the shifted distribution. The deviation from linearity is of order $d^{-1/2}$ and vanishes in high dimensions.