On the Generalization and Adaption Performance of Causal Models

Nino Scherrer [1 2] **Anirudh Goyal** [1] **Stefan Bauer** [3] **Yoshua Bengio** [1] **Nan Rosemary Ke** [4]

Abstract

Learning models that offer robust out-of-distribution generalization and fast adaptation is a key challenge in modern machine learning. Modelling causal structure into neural networks holds the promise to accomplish robust zero and few-shot adaptation. Recent advances in differentiable causal discovery have proposed to factorize the data generating process into a set of modules, i.e. one module for the conditional distribution of every variable where only causal parents are used as predictors. Such a modular decomposition of knowledge enables adaptation to distributions shifts by only updating a subset of parameters. In this work, we systematically study the generalization and adaption performance of such modular neural causal models by comparing it to monolithic models and structured models where the set of predictors is not constrained to causal parents. Our analysis shows that the modular neural causal models outperform other models on both zero and few-shot adaptation in low data regimes and offer robust generalization. We also found that the effects are more significant for sparser graphs as compared to denser graphs.

1. Introduction

Deep Learning models have demonstrated remarkable capabilities when the test distribution matches the training distribution, but their performance significantly degrades as the test distribution diverges from the training distribution (Lake et al., 2017; Rosenfeld et al., 2018; Bahdanau et al., 2018; Packer et al., 2018; Nichol et al., 2018b; Cobbe et al., 2019; Taori et al., 2020; Djolonga et al., 2021; Koh et al., 2021). However, such distribution shifts are inevitable in the real world and can occur in various settings, e.g. across hospitals in healthcare or across locations in agriculture (Wang et al., 2021). This sensitivity to distribution shift inherently limits the robust and safe deployment in the wild. At the same time, deep learning systems constructed with a multi-layered monolithic architecture tend to co-adapt different components of the network. Due to such a monolithic structure, when the distribution changes, a majority of the components of the network are likely to adapt in response to these changes, potentially leading to poor performance on out-of-distribution samples (Ke et al., 2021; Goyal et al., 2019) and interference between subtasks or subdistributions. Endowing neural networks with the ability to capture the underlying causal structure holds the promise to accomplish much out-of-distribution adaptation and generalization by properly factorizing the knowledge that is stationary (causal mechanisms) from the knowledge that isn't (the state of the random variables and interventions that change the distribution).

Given the underlying causal graph G, every causal mechanism represents a conditional probability distribution $p(X_i|X_{pa(i,G)})$ of a given variable X_i where only causal parents $X_{pa(i)}$ are used as predictors. In such a causal framework, distribution shifts can be interpreted as interventions (i.e. perturbations) that affect certain mechanisms locally (Magliacane et al., 2018; Zhang et al., 2013; Schölkopf et al., 2021). As usually not the complete environment and its structure changes at once, adapting to a distribution shift in such a framework is therefore equivalent to adapting the intervened mechanisms.

The promising opportunities of causal models in machine learning have led to a flurry of work and accompanying advances along various research axes (e.g. causal discovery (Zheng et al., 2018; Yu et al., 2019; Bengio et al., 2019; Ke et al., 2019; Lachapelle et al., 2019; Brouillard et al., 2020; Scherrer et al., 2021; Annadani et al., 2021; Lorch et al., 2021; Lippe et al., 2021; Geffner et al., 2022; Ke et al., 2022), domain adaptation (Zhang et al., 2013; Peters et al., 2016; Rojas-Carulla et al., 2018; Magliacane et al., 2018; Bengio et al., 2019; Le Priol et al., 2021), robustness of neural networks (Zhang et al., 2020; Kyono et al., 2020), causal models in reinforcement learning (RL) (De Haan et al., 2019; Dasgupta et al., 2019; Nair et al., 2019; Goyal et al., 2019; Rezende et al., 2020; Ke et al., 2021), etc. While most of these works either analyze a problem-specific objective, such as structure discovery or a success rate on a

[1]Mila, Université de Montréal [2]ETH Zurich [3]KTH Stockholm [4]DeepMind. Correspondence to: Nino Scherrer <nino.scherrer@gmail.com>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

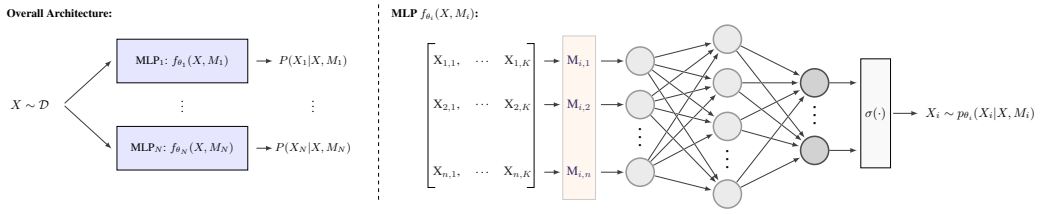


Figure 1: Architectural Setup: We employ the same model architecture consisting of a stack of MLPs across all considered models. Every MLP $f_{\theta_i}(X, M_i)$ consisting of a masking layer M_i and one hidden layer, takes a data sample $X \sim \mathcal{D}$ as input, applies a input mask M_i to the input and outputs the logits of a categorical distribution on its output layer. The logits are finally normalized through a softmax-activation function $\sigma(\cdot)$ which results in the CPD $p(X_i|X, M_i)$. Between the considered models, we vary the learning objective to learn the parameters θ and the inputs masks M. While *monolithic models* and *structured models* with expert knowledge have a fixed mask M, the causal discovery model optimizes M using an additional set of parameters.

task, little attention has been paid to a *systematic analysis* of the generalization and adaptation capabilities of causal models. Previous work including speed of adaptation analysis of Bengio et al. (2019); Le Priol et al. (2021) is limited to causal and anti-causal models in a bivariate setting. The work of Ke et al. (2021) analyzes generalization and adaptation performance of models with different inductive biases in different high-dimensional RL environments, where the underlying causal structure as well as causal variables are not given and need to be learned directly from high dimensional visual input.

In this work, we *systematically* investigate zero and few-shot adaptation capabilities of *monolithic models* and *structured models* where causal variables are explicitly given. As an evaluation setting, we consider the task of predicting missing values (e.g. given all the other variables of the sample) under unseen distribution shifts (see **??**). In order to investigate the effect of the different inductive biases, we employ the same model architecture (i.e. one MLP per conditional distribution) across all considered models (see Figure 1). Hence, all models have the same expressive abilities and only vary in their training objectives and pre-existing domain knowledge. Within the class of *structured models*, we distinguish between expert knowledge models where we provide certain structure upfront of training (e.g. causal graph, anti-causal graph) and models where causal structure is learned from data. We train monolithic models with different training objectives including a pseudo-likelihood objective as well as a meta-learning objective which explicitly optimizes the parameters of the monolithic models to adapt quickly to changes in distribution, hence confounding many different problems. This setup allows us to uncover generalization and adaptation discrepancies between different models and analyze if and where models are prone to fail.

Contributions. (i) We show that generalization capabilities of different models vary significantly with the amount of available training samples. (ii) We demonstrate that *structured models* significantly outperform *monolithic models* in low-data regimes. (iii) We show that a general evaluation metric is prone to drawing erroneous conclusions with respect to robustness and show how a general evaluation metric can be dissected into refined metrics to investigate if and how specific models fail. (iv) We show that non-causal models can fail drastically in settings where the underlying causal structure is sparse. (v) We evaluate few-shot adaptation in various settings and show that causal models are the fastest and most robust to adapt. (vi) We show how models adapt in parameter space and relate this to the speed of adaptation and robustness. (vii) We propose and investigate a new adaptation objective for causal models which enables an efficient adaptation in low training and adaptation-data regimes.

2. Problem Setting

In our work, we consider setting where high-level causal variables are observed and given (i.e. they do not need to be inferred from high-dimensional input). We limit the number of variables $N \in \{10, 20\}$ where causal variables $X = \{X_1, \ldots, X_n\}$ are directly observed. We generate synthetic observational and interventional data $\mathcal{D} = (\mathcal{D}_{obs}, \mathcal{D}_{int})$ on causal acyclic graphs and fit different models to data in order to learn conditional probability distributions $p(X_i|\cdot)$ for all variables X_i. During test time (with and without adaptation), we predict all variables X_j for $j \in \{1, \ldots, N\}$ of unseen interventional distributions $p(X|do(X_k))$ using all other variables X_i $(i \neq j)$ of the sample.

3. Model Architecture and Training Setup

Model Architectures In order to disentangle architectural effects from the training objectives, we employ the same architecture across all evaluated models. We follow the setup of (Ke et al., 2019; Scherrer et al., 2021) and choose a stack of N independent MLPs for a setting with N observed variables. Hence for every variable X_i, there exists an MLP parametrized by θ_i that represents the conditional probability distribution (CPD) $p(X_i|X, M_i)$ where $M_i \in \{0, 1\}^N$ denotes a input mask. Specifically, every MLP has an input layer of size $N \times K$, one hidden layer of 64 neurons with Leaky ReLU activations of slope 0.1 and a linear output layer of size K. The output layer represents

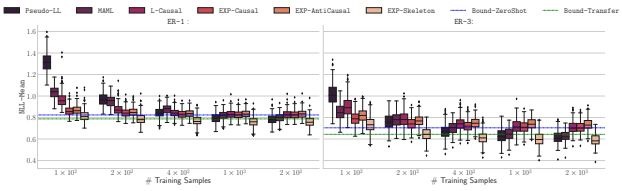


Figure 2: OOD Generalization with Varying Amounts of Training Data. We report `NLL-Mean` over a set of unseen interventions across `ER`-graphs of varying density ($N = 20$ and 10 graphs per setting) and training sets of increasing size (up to 2k samples). *Structured models* outperform *monolithic models* consistently on low data regimes on the `NLL-mean` metric. `EXP-Skeleton` outperforms all other models on all settings. The causal models attain `Bound-ZeroShot` (green) as expected with sufficient amount of samples.

the unnormalized log-probabilities of each possible category that are finally mapped to valid CPDs through a softmax activation function $\sigma(\cdot)$.

Training Objectives In our study, we evaluate six different learning paradigms on a fixed architecture consisting of a stack of N MLPs (see Section 3 for a detailed description) with the goal to learn the underlying generative mechanisms $p(X_i|\cdot)$. To this end, we consider two techniques to learn a *monolithic model* (i.e. `Pseudo-LL` and `MAML`) and four different techniques to learn a *structured model* (i.e. `EXP-Causal`, `EXP-AntiCausal`, `EXP-Skeleton` and `L-Causal`). Within the *structured models*, we construct three models using expert knowledge (`EXP`) and a model where the causal structure is learned from data (i.e. `L-Causal`) without any supervision. The exact training objectives are described in the appendix § C.2

4. Analysis of Generalization Performance

We start by analyzing OOD generalization (zero-shot) performance for the different models discussed in Section 3. In particular, our experiments seek to answer the following questions: (a) How different models generalize under different circumstances. (b) Analyzing how different parts of our models contribute to the failure or success for OOD generalization.

Implementation details. We keep the training and evaluation setup between different models as similar as possible. All models have the same number of parameters to represent the CPDs and are trained for the same number of steps, we train models on the same range of learning rates and pick the best performing one for each model individually. All experiments are run with 10 random seeds, we report the mean and standard deviation of the results. In particular, we use the Adam optimizer (Kingma & Ba, 2014) across all models and have evaluated learning rates from the set $\{1e{-}2, 1e{-}3, 1e{-}4\}$, weight decay from the set $\{1e{-}4, 1e{-}5, 0\}$ and train all models for 1000 iterations, except the `L-Causal` model which was trained for multiple rounds with iterative optimization. For the detailed setup with all model-specific hyperparameters, we refer to appendix §C.3.1.

Performance Bounds on Causal Models. To assess the performance of the causal models, we compute two upper bounds on the performance of the causal model by accessing the data-generating causal model. (a) `Bound-ZeroShot` shows the maximal zero-shot adaptation performance and (b) `Bound-Adaptation` show the maximal adaptation performance on the transfer distribution. Note that the causal model `EXP-Causal` that relies on expert knowledge should naturally attain this bound faster than the model `L-Causal` that learns the causal structure from data.

4.1. Generalization performance.

We analyze the generalization performance of different models under different settings. There are three aspects of the settings that we consider. First is the amount of training data $\mathcal{D}$, then we look at the density of the underlying SEM that generated the data; at last, we look at the size of the graph. To be specific, we vary the amount of training data between 10^2 and 10^4, the graph density of the underlying SEM between ER-1 and ER-3 and the size of the graph between 10 and 20.

The training data $\mathcal{D}^T$ consist of both observational $\mathcal{D}^T_{obs}$ and interventional data $\mathcal{D}^T_{int}$. We keep the number of observational data $\mathcal{D}^T_{obs}$ the same as the number of interventional data $\mathcal{D}^T_{int}$ in training. Test data $\mathcal{D}^t$ is kept fixed across experiments. We report the `NLL-Mean` of the test data on the model, this is average NLL scores across all variables (including the intervened-on variable).

Summary. Results for comparisons between different amount of data and graph sparsity are found in Figure 2, results for comparisons between data with different number of nodes are in Figure 4. Refer to §*F*.3 in the appendix for a complete set of results. We found that *causal models* outperform *monolitic models* (`Pseudo-LL` and `MAML`) when the amount of training data is low (Figure 2). We also found that the performance gap widens as the density of the graphs decreases (Figure 2). The performance gap also widens as the size of the SEM increases (Figure 4). These results suggest that models with the correct structure (such as causal models) generalize better compared to models with no structure, especially when trained on a small amount of data coming from sparse and large graphs. Furthermore, `EXP-Causal` models outperforms `EXP-AntiCausal` models under all settings, and `EXP-AntiCausal` can perform worse than monolitic models (`Pseudo-LL and MAML`) under some settings. This suggests that having the correct structure is important and having the wrong structure can hurt performance.

We observe slower convergence of `causal` models to `Bound-ZeroShot` on dense graphs than on sparse graph,

as the identification of the causal structure is more challenging in such settings (Ke et al., 2019; Scherrer et al., 2021). An interesting observation is that `EXP-Skeleton` models can outperform `EXP-Causal` models, this is because that when the intervention is not on the children of the predicted node, then `EXP-Skeleton` models can use the value of the child and the the value of the parents to predict the value of the node, whereas, the `EXP-Causal` models only uses the parents to predict the value of the node. However, the `EXP-Skeleton` models will fail catastrophically when the intervention is on the children of the predicted node, we will see more analysis about this in Section 4.2.

4.2. Dissecting generalization performance

The analysis in Section 4.1 reports an average evaluation metric across all nodes, which may not help us to understand the model's performance in detail, as noted at the end of the previous section. Hence, in this section, we aim to dissect the `NLL-Mean` metric into sub-metrics, which could help to systematically identify if and where models are prune to failure. Such an analysis helps us to assess the the individual model robustness to distribution shifts and uncover failure settings which are hidden in the `NLL-Mean` metric. To this end, we consider a data regime (1k training samples) where all models perform similar with respect to the general `NLL-Mean` score.

We dissect `NLL-Mean` systematically into: (a) `NLL-Intervention`: NLL on intervened variable X_i, (b) `NLL-Root`: Mean-NLL on root variables in G, (c) `NLL-Parents`: Mean-NLL on parent variables $X_{pa(i,G)}$ for a intervention on X_i (excluding root variables) and (d) `NLL-Remainder`: Mean-NLL of all variables except root and intervention variables.

Note that, causal models are not expected to yield comparable performance on `NLL-Intervention` and `NLL-Root`, as causal models only use the parents of each variables to make its prediction. As this is an empty set for root variables and hard intervened variables, causal models estimate such variables from the marginal distribution of such variables. In contrast, monolithic models also rely on (potentially unstable) correlated predictors and anti-causal predictors and hence benefit from stronger performance as long none of these predictors were affected by an intervention.

Summary. Building upon our introduced sub-metrics, we find that important failure and robustness insights are hidden in a general evaluation scores such as `NLL-Mean` (see Figure 3). We observe different performance trends across all sub-metrics. While *causal models* maintain a robust performance as they only rely on the inferred causal predictors, *monolithic models* and *structured models* that rely on anti-causal predictors (i.e. `EXP-AntiCausal` and `EXP-Skeleton`) show a significant deterioration in performance with large standard deviations on `NLL-Parents` as certain predictors got unstable due to the present distribution shift (i.e. modelled by a single-target intervention). We further observe stronger effects on sparse graphs where less stable predictors are available (see §F.3). Overall, this experiment confirms that all models expect *causal models* show bigger risk for catastrophical failures by relying on unstable predictors.

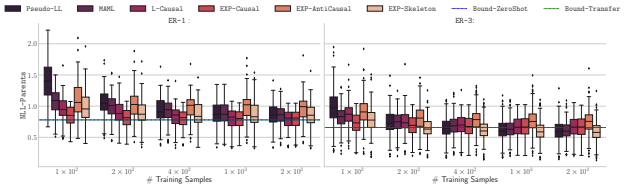


Figure 3: `NLL-Parents` metric across various sizes of training data.. A dissection of the general `NLL-Mean` metric into more fine-graded sub-metrics reveals important failure scenarios. In particular, as shown here on the `NLL-Parents` metric, *non-causal models* can catastrophically fail to predict the parent variables of an intervened variables. Note that effect is stronger on sparse graphs where less stable predictors are available.

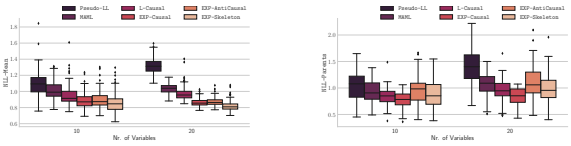


Figure 4: Metric across increasing number of variables (fixed setting `ER-1` with 100 training samples). We observe across all our experiments that the general generalization performance gaps with respect to `NLL-Mean` metric between *monolithic models* and *structured models* increases with increasing number of variables. On the same time, we observe a similar behavior on `NLL-Parents` where *causal models* maintain robust generalization performance while all *non-causal models* show big risks for catastrophic failure.

5. Analysis of Adaptation Performance

We evaluate the adaption performance by evaluating how fast (speed of adaptation) and how well (how much overfitting) different models adapt to changes in distribution. We use two different adaptation techniques: (i) *unconstrained adaptation* and (ii) *sparse adaptation* (explained in Appendix C.3). Throughout this analysis, we pay particular attention to the `NLL-Parents` metrics and see how fast the affected models recover.

Speed of adaptation. We fix the training data size of all models to be 10^3 samples, as all models have converged on the generalization performance by then (Figure 2). We analyze the speed of adaption of different models by evaluating their adaption performance when finetuned using different amounts of data. Results shown in Figure 5.

We observe that the *structured models* adapt considerably faster than the *monolithic models* across all settings. We

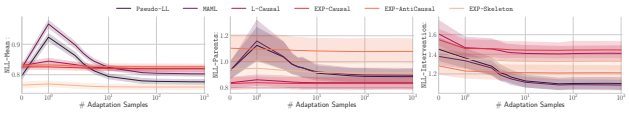


Figure 5: Speed of Adaptation in terms of different metrics. *Structured models* adapt considerably faster than *monolithic models* across all settings and metrics. *Monolithic models* show a sensitivity to overfitting if only low amounts of adaptation samples are available. We observe slightly faster adaptation for *causal model* than for the other *structured models*.

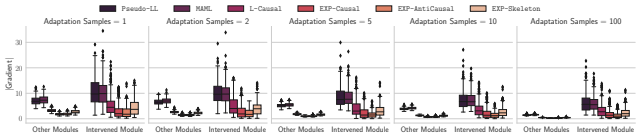


Figure 6: Parameter Space Analysis. While adapting to a shift in distribution, *monolithic models* update most modules that were not affected by the intervention quite heavily compared to *structured models*. *Causal models* show remarkable adaptation behaviour in parameter space with localized updates on intervened modules.

observe slightly faster adaptation for *causal model* than for the other *structured models*. Within the *monolithic models*, we observe that models that are trained with an adaptation objective (i.e. with MAML (Finn et al., 2017)) adapt faster with respect to the intervened module than models trained on a naive pseudo-likelihood objective.

Risk of overfitting. Ideally, one aims to adapt to a transfer distribution using a less number of adaptation samples. In such a setting, a single update (i.e. gradient step with respect to the samples) using the available samples may be not sufficient to exploit all available information. Hence, it would be desirable to perform multiple updates on a small number of adaptation samples to extract all relevant information but without overfitting to the adaptation samples. To this end, we investigate the risk of overfitting when performing one or multiple updates on a fixed amount of data.

Across all our experiments, *monolithic models* show strong overfitting effect when the number of adaptation samples are less, even on a single update step (see Figure 5). In contrast, *structured models* show reduced overfitting effects over multiple gradient steps, especially *causal models*. For less number of adaptation samples, the speed of adaptation of *causal models* can be further improved by employing a sparse adaptation objective. Overall, the adaptation landscape of the *causal models* is significantly *different* from all other models, and hence allows to continuously improve the adaptation performance over multiple update steps.

How does adaptation affect the parameter space θ? Based on the results from the previous analyses, we aim to further investigate the adaptation performance of models. We compare the effect on the parameters space θ between different models by employing the unconstrained adaptation objective, the most general adaptation objective.

While *monolithic models* adapt the parameters of many modules (i.e. independent MLPs) heavily, measured with respect to the gradient magnitude, the adaptation of *structured models* results in smaller updates, especially on the non-intervened modules (see Figure 6). *Causal models* show remarkable adaptation behaviour in parameter space with *localized updates on certain modules*, and significantly reduced gradient magnitudes. If a *causal model* is trained on enough training samples and has access to enough adaptation adaptation samples, the unconstrained adaptation without knowledge about the intervention target yields nearly the same update as if we enforce the sparse update on the known intervened module (see §G.2).

Efficient Adaptation of Causal Model Due to space constraints, we refer to the appendix §E for an analysis of the proposed adaptation objective.

6. Conclusion

In this work, we systematically analyzed the generalization and adaptation performance of different models ranging from *monolithic models* that have no inbuilt structure to *structured models* that are either provided with structural expert knowledge upfront or learn structure from data. Our experiments show that the *causal models* significantly outperform *non-causal* models in low-data regimes and offer robust generalization across all settings. In a further analysis, we evaluated few-shot adaptation in various settings and show that *causal models* offer fast and robust adaptation with only less number of adaptation samples. Based on these results, we analyze how the adaptation performance relates to changes in the parameter space and then proposed a new adaptation objective that dynamically modulates the degree of adaptation and hence allows more sample efficient adaptation. In this work, we considered relatively low-dimensional settings where causal variables are explicitly given. Translating our systematic evaluation and score dissection analysis to high-dimensional evaluation setups such as (Ke et al., 2021) would be an interesting direction for future work.

Acknowledgements

The authors would like to acknowledge the support of the following agencies for research funding and computing support: Compute Canada, the Canada Research Chairs, CIFAR. Yoshua Bengio is a CIFAR Senior Fellow and Stefan Bauer is a CIFAR Azrieli Global.

References

Alet, F., Lozano-Pérez, T., and Kaelbling, L. P. Modular meta-learning. In *Conference on Robot Learning*, pp. 856–868. PMLR, 2018. (Cited on 8)

Annadani, Y., Rothfuss, J., Lacoste, A., Scherrer, N., Goyal,

A., Bengio, Y., and Bauer, S. Variational causal networks: Approximate bayesian inference over causal structures. *arXiv preprint arXiv:2106.07635*, 2021. (Cited on 1, 8)

Bahdanau, D., Murty, S., Noukhovitch, M., Nguyen, T. H., de Vries, H., and Courville, A. Systematic generalization: what is required and can it be learned? *arXiv preprint arXiv:1811.12889*, 2018. (Cited on 1)

Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. (Cited on 8)

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019. (Cited on 1, 2, 8)

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21865–21877. Curran Associates, Inc., 2020. (Cited on 1, 8)

Chen, Y., Friesen, A. L., Behbahani, F., Doucet, A., Budden, D., Hoffman, M., and de Freitas, N. Modular meta-learning with shrinkage. *Advances in Neural Information Processing Systems*, 33:2858–2869, 2020. (Cited on 8)

Cobbe, K., Klimov, O., Hesse, C., Kim, T., and Schulman, J. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pp. 1282–1289. PMLR, 2019. (Cited on 1)

Cundy, C., Grover, A., and Ermon, S. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on 8)

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., and Kurth-Nelson, Z. Causal Reasoning from Meta-reinforcement Learning. *arXiv preprint arXiv:1901.08162*, 2019. (Cited on 1)

De Haan, P., Jayaraman, D., and Levine, S. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32, 2019. (Cited on 1)

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16458–16468, 2021. (Cited on 1)

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017. (Cited on 5, 8, 9)

Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., Sharma, A., Lamb, A., Kukla, M., Pawlowski, N., et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022. (Cited on 1, 8)

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019. (Cited on 1)

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C., and Bengio, Y. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019. (Cited on 1, 2, 4, 8, 9, 10, 12)

Ke, N. R., Didolkar, A., Mittal, S., Goyal, A., Lajoie, G., Bauer, S., Rezende, D., Bengio, Y., Mozer, M., and Pal, C. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*, 2021. (Cited on 1, 2, 5, 8)

Ke, N. R., Chiappa, S., Wang, J., Bornschein, J., Weber, T., Goyal, A., Botvinic, M., Mozer, M., and Rezende, D. J. Learning to induce causal structure. *arXiv preprint arXiv:2204.04875*, 2022. (Cited on 1, 8)

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on 3, 10, 11)

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021. (Cited on 1)

Kyono, T., Zhang, Y., and van der Schaar, M. Castle: Regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33: 1501–1512, 2020. (Cited on 1, 8)

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019. (Cited on 1, 8)

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. (Cited on 1)

Le Priol, R., Babanezhad, R., Bengio, Y., and Lacoste-Julien, S. An analysis of the adaptation speed of causal models. In *International Conference on Artificial Intelligence and Statistics*, pp. 775–783. PMLR, 2021. (Cited on 1, 2, 8)

Lippe, P., Cohen, T., and Gavves, E. Efficient neural causal discovery without acyclicity constraints. *arXiv preprint arXiv:2107.10483*, 2021. (Cited on 1, 8, 9, 10, 11)

Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. Dibs: Differentiable bayesian structure learning. *arXiv preprint arXiv:2105.11839*, 2021. (Cited on 1, 8)

Madan, K., Ke, N. R., Goyal, A., Schölkopf, B., and Bengio, Y. Fast and slow learning of recurrent independent mechanisms. *arXiv preprint arXiv:2105.08710*, 2021. (Cited on 8, 9)

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pp. 10869–10879, 2018. (Cited on 1, 8)

Nair, S., Zhu, Y., Savarese, S., and Fei-Fei, L. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019. (Cited on 1)

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018a. (Cited on 10)

Nichol, A., Pfau, V., Hesse, C., Klimov, O., and Schulman, J. Gotta learn fast: A new benchmark for generalization in rl. *arXiv preprint arXiv:1804.03720*, 2018b. (Cited on 1)

Packer, C., Gao, K., Kos, J., Krähenbühl, P., Koltun, V., and Song, D. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018. (Cited on 1)

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. (Cited on 8)

Peters, J., Bühlmann, P., and Meinshausen, N. 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. (Cited on 1, 8)

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. (Cited on 8)

Rezende, D. J., Danihelka, I., Papamakarios, G., Ke, N. R., Jiang, R., Weber, T., Gregor, K., Merzic, H., Viola, F., Wang, J., et al. Causally correct partial models for reinforcement learning. *arXiv preprint arXiv:2002.02836*, 2020. (Cited on 1)

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018. (Cited on 1, 8)

Rosenfeld, A., Zemel, R., and Tsotsos, J. K. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. (Cited on 1)

Scherrer, N., Bilaniuk, O., Annadani, Y., Goyal, A., Schwab, P., Schölkopf, B., Mozer, M. C., Bengio, Y., Bauer, S., and Ke, N. R. Learning neural causal models with active interventions. *arXiv preprint arXiv:2109.02429*, 2021. (Cited on 1, 2, 4, 9, 10, 12)

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021. (Cited on 1, 8)

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. (Cited on 1)

Wang, J., Lan, C., Liu, C., Ouyang, Y., Zeng, W., and Qin, T. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021. (Cited on 1)

Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019. (Cited on 1, 8)

Zhang, C., Zhang, K., and Li, Y. A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems*, 33:289–301, 2020. (Cited on 1, 8)

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827. PMLR, 2013. (Cited on 1, 8)

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9472–9483, 2018. (Cited on 1, 8)

A. Background

Causal Graph. A causal graph is commonly represented by a directed acyclic graph (DAG) $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ with $|\mathbf{V}| = N$ and $|\mathbf{E}| = M$. Such a graph is defined over finite set of vertices $\mathbf{V}$ associated with a finite set of random variables (or observables) $\mathbf{X} = \{X_1, \dots, X_N\}$, where directed edges in the causal graph $\mathbf{G}$ point from causes to effects. For convenience, the set of X_i's parents in $\mathbf{G}$ is usually denoted as $X_{pa(i,\mathbf{G})}$ and the set of X_i's children in $\mathbf{G}$ by $X_{ch(i,\mathbf{G})}$.

Adjacency Matrix. The connectivity between vertices $\mathbf{V}$ in a graph is commonly represented by an *adjacency matrix* $\mathbf{A} \in \{0,1\}^{N \times N}$ such that $\mathbf{A}_{i,j} = 1$ if node j is a parent of node i.

Structural Causal Model (SCM). An SCM (Pearl, 1995; Peters et al., 2017), also known as structural equation model (SEM), is defined by a causal graph $\mathbf{G}$ over a set of random variables (or observables) $\mathbf{X} = \{X_1, \dots, X_N\}$ and a set of associated structural equations. The structural equations express the functional relationships among the causal variables through functions f_i and jointly independent noise variables U_i as $X_i = f_i(X_{pa(i,G)}, U_i) \forall i \in \{1, \dots N\}$. The noise variables U_i ensure that the set of structural equations can represent general conditional probability distributions $P(X_i | X_{pa(i,G)})$. The joint distribution entailed by the variables $\mathbf{X} = \{X_1, \dots, X_N\}$ can be factorized such that each variable is conditionally independent of other variables given its parents in the graph $\mathbf{G}$:

$$P(X_1, \dots, X_N) = \prod_{i=1}^{N} P(X_i | X_{pa(i,\mathbf{G})}) \tag{1}$$

In the causality literature, this factorization is also known as the causal factorization (Schölkopf et al., 2021).

B. Related Work

Differentiable Causal Discovery. Recent advances in differentiable causal discovery focused on building new algorithms for causal discovery from observational data (Zheng et al., 2018; Yu et al., 2019; Lachapelle et al., 2019; Annadani et al., 2021; Lorch et al., 2021; Cundy et al., 2021; Geffner et al., 2022; Ke et al., 2022) or fused data (observational and interventional data) (Bengio et al., 2019; Ke et al., 2019; Brouillard et al., 2020; Lippe et al., 2021) using advances in deep learning. Such methods are primarily concerned to identify the underlying causal structure from data, and not evaluate the zero and few-shot capabilities of the learned models.

Speed of Adaptation While Bengio et al. (2019); Ke et al. (2019) included an analysis for adaptation or generalization speed of causal models compared to monolithic models, these analyses are focused on a specific setting, such as a specific number of variables. The work of Le Priol et al. (2021) analyzes the speed of adaption of causal and anti-causal models, however, the analysis is only limited to the bivariate settings. Schölkopf et al. (2021) discussed the generalization and adaption performance of causal models against monolithic models in a high-dimensional setting, however, no experimental analysis is included. Ke et al. (2021) proposed a novel suite of RL enviromments and tasks for analyzing causal discovery in a high-dimensional RL setting, and the work analyzed generalization and adaption performance of models with different inductive biases. In our work, we perform a systematic analysis of generalization and adaption performance of causal models against monolithic models, as well as models that are explicitly optimized using meta-learning objectives such as MAML (Finn et al., 2017) on settings where the causal variables are explicitly given.

Domain Adaptation. Multiple approaches have been proposed that exploit the causal structure of the data generating process in order to address the problem of domain adaptation (Zhang et al., 2013; Peters et al., 2016; Bareinboim & Pearl, 2016; Rojas-Carulla et al., 2018; Magliacane et al., 2018). While these works analyze specific instances of domain adaptation problems with varying assumptions, our work is concerned to investigate zero- and few-shot adaptation abilities of various *monolithic* and *structured models*.

Improving Robustness through Causal Structure. Zhang et al. (2020) showed a connection between the vulnerability / robustness of neural neural networks and their lack of causal reasoning. Kyono et al. (2020) showed that learning causal structure as an auxiliary tasks improves the in-distribution generalization capabilities of overparameterized feed-forward neural networks. However, the work only investigates out-of-sample generalization within the same distribution, and does not consider out-of-distribution settings.

Transfer through Modular Knowledge Decomposition. Recent work has shown that architectural inductive biases which promote modular decomposition of knowledge can provide a useful basis for transfer of knowledge from one task to another task (Alet et al., 2018; Chen et al., 2020; Madan et al., 2021). Such architectures employ a meta-learning approach to update

different subset of parameters of the network over different timescales and show such an approach leads to improvements in sample efficiency as compared to training all the parameters at once (Madan et al., 2021). Such methods learn directly from low dimensional pixel data and don't explicitly learn causal variables.

B.1. Limitations

In the present work, we have only experimented with one specific class of a neural causal discovery framework (Ke et al., 2019; Lippe et al., 2021) to learn causal structure from data. Hence, the performance may vary with other classes of neural causal discovery frameworks. However, we have introduced a causal model where the true causal structure is presented upfront and hence represents an performance upper-bound given a certain amount of training data. Furthermore, we have only conducted experiments on datasets where the causal variables are explicitly observed and the underlying causal graph is acyclic.

C. Model and Training Setup

C.1. Data Generation

In order to systematically investigate the effect of different training objectives on generalization and adaptation performance, we employ a synthetic data generation setup. We generate observational and interventional data $\mathcal{D} = (\mathcal{D}_{obs}, \mathcal{D}_{int})$ of discrete and non-linear nature governed by causal graphs.

Graph Generation. We distinguish between *structured* and *random* graphs in order to represent a wide diversity of possible graphs. For structured graphs, we follow the setup of (Ke et al., 2019) and generated various DAGs with acyclic and cyclic skeletons. In order to generate random graphs, we follow the Erdós–Rényi (`ER`) model with varying edge densities (i.e. `ER-1`, `ER-2` and `ER-3`) as in (Scherrer et al., 2021).

Conditional Probabilty Distributions (CPD) In order to generate discrete, observational data $\mathcal{D}_{obs}$ given a causal DAG G, we perform ancestral sampling based on the topological order of the causal DAG G as proposed in (Ke et al., 2019). Similar to (Ke et al., 2019; Lippe et al., 2021), we model the CPDs $p(X_i|X_{pa(i,G)})$ using randomly initalized one-hidden layer MLPs (weights orthogonally in the range $[-2.5, 2.5]$ and biases uniformly in the range $[-1.1, 1.1]$) with a hidden dimensionality of 48 where all inputs except the parents $X_{pa(i,G)}$ are masked to 0 during the sampling process. We perform point interventions on a single node modelled by an Uniform distribution $U[1, K]$, where K is the number of possible categorical assignments.

C.2. Training Objectives

C.2.1. Monolithic Models

Pseudo-Loglikelihood (`Pseudo-LL`): As a simple monolithic model, we train a model using maximum likelihood on an unconstrained input mask M on observational data $\mathcal{D}_{obs}$. We minimize:

$$\theta_i^* = \arg\min_{\theta_i} \mathop{\mathbb{E}}_{X \sim \mathcal{D}_{obs}} [-log(f(X, M_i; \theta_i))] \tag{2}$$

for every MLP$_i$ independently, where M_i denotes the i'th column of the input mask and θ_i the corresponding MLP parameters. M is a unconstrained input mask with ones everywhere except zeroes on the diagonal. This prevents from learning an identity mapping $X_i = X_i$. Hence, the `Pseudo-LL` model learn CPD's of the form $p(X_i|X \setminus X_i)$ on data from $\mathcal{D}_{obs}$.

Model-Agnostic Meta-Learning (`MAML`): Motivated by the adaptation capabilities of meta-learned models, we use the model-agnostic meta learning algorithm (MAML) (Finn et al., 2017) in order to train a monolithic model on a variety of interventional distributions $\{\mathcal{D}_{int(l)}\}_{l=1}^{L}$. To this end, we employ the following meta-optimization formulation:

$$\begin{aligned} \theta_i^* &= \arg\min_{\theta_i} \sum_{l \sim p(Int)} \mathop{\mathbb{E}}_{X \sim \mathcal{D}_{int(l)}} \big[-log\big(f(X, M_i; \hat{\theta}_{i,l})\big)\big] \\ \text{where: } \hat{\theta}_{i,l} &= \theta_i - \alpha \nabla_{\theta_i} \mathop{\mathbb{E}}_{X \sim \mathcal{D}_{int(l)}} \big[-log(f(X, M_i; \theta_i\big] \end{aligned} \tag{3}$$

where $p(Int) = \mathcal{U}[1, \dots, L]$ denotes a uniform distribution over the available interventional distributions, M_i is the input mask, α is the step-size parameter of the inner update and $\hat{\theta}_i, l$ represents the updated model parameters. The above

meta-optimization objective is optimized using Adam (Kingma & Ba, 2014) and the inner updated is done using stochastic gradient descent (SGD). Note that the meta-optimization is performed over the model parameters θ_i, whereas the objective is computed using the updated model parameters $\hat{\theta}_i$. For the input mask M_i, we follow same setup as for `Pseudo-LL`. Hence, the `MAML` models learns CPD's of the form $p(X_i|X \setminus X_i)$ on different interventional distributions $\mathcal{D}_{int(l)} \sim \mathcal{D}_{int}$. In our experiment, we rely on the first-order approximation of MAML (Nichol et al., 2018a).

C.2.2. Structured Models

Learning with Expert Knowledge (`EXP-Causal`, `EXP-AntiCausal` and `EXP-Skeleton`)**:** Given the adjacency matrix A of the ground-truth causal structure G, the anti-causal adjacency matrix (i.e. the transpose of the causal adjacency) or the adjacency matrix of the undirected skeleton, we inject the the provided expert knowledge by setting the input mask to $M = A$ and train the models `EXP-Causal`, `EXP-AntiCausal` and `EXP-Skeleton` using maximum likelihood training on observational data $\mathcal{D}_{obs}$. To this end, we minimize Equation (2) for every MLP independently. Hence, `EXP-Causal` learns CPD's of the form $p(X_i|X_{pa(i,G)})$ (i.e. only causal predictors), `EXP-AntiCausal` learns CPD's of the form $p(X_i|X_{ch(i,G)})$ (i.e. only anti-causal predictors) and `EXP-Skeleton` learns CPD's of the form $p(X_i|X_{pa(i,G)}), X_{ch(i,G)})$ (i.e. causal and anti-causal predictors).

Learning Causal Structure (`L-Causal`): We use a causal discovery framework to learn a structural causal model (SCM) from data. To this end, we follow the setup of Lippe et al. (2021) and introduce an additional set of parameters $\gamma = (u, v)$ with $u \in \mathbb{R}^{N\times N}$ and $v \in \mathbb{R}^{N\times N}$ which define a continuous relaxation of an adjacency matrix $\gamma = \sigma(u) \cdot \sigma(v)$. Such a soft-adjacency matrix can be conveniently used to sample input masks M. In order to train the parameters θ of the MLPs and the learnable input mask γ, the framework relies on a optimization formulation using two alternating phases of optimization (Ke et al., 2019; Lippe et al., 2021; Scherrer et al., 2021). These are performed until convergence in an iterative manner. Under freezed mask parameters γ, we train during phase 1 (called "Distribution Fitting") the parameters θ_i of each MLP on observational data $\mathcal{D}_{obs}$ using a similar maximum likelihood objective as in Equation (2):

$$\theta_i^* = \arg\min_{\theta_i} \mathop{\mathbb{E}}_{X\sim\mathcal{D}_{obs}} \mathop{\mathbb{E}}_{M\sim p(M;u,v)} [-log(f(X, M_i; \theta_i))] \tag{4}$$

where we sample a set of input masks M from $p(M; u, v)$ instead of relying on a fixed mask M, where $M_{ij} \sim Ber(\sigma(u_{ij}) \cdot \sigma(v_{ij}))$. During phase 2 (called "Graph Fitting"), we freeze the previously trained MLP parameters θ and optimize the mask parameters γ using different sets of interventional data $\mathcal{D}_{int(l)} \sim \mathcal{D}_{int}$. To this end, we employ the optimization formulation of (Lippe et al., 2021):

$$\begin{aligned}\gamma^* = (u^*, v^*) = \arg\min_{(u^*,v^*)} = \mathbb{E}_{l\sim p_I(I)} \mathbb{E}_{X\sim\mathcal{D}_{int(l)}} \mathbb{E}_{M\sim p(M;u,v)} \left[\sum_{i=1}^{N} -log f(X, M_i; \theta_i)\right] \\ + \underbrace{\lambda_{\text{sparse}} \sum_{i=1}^{N}\sum_{j=1}^{N} \sigma(u_{ij}) \cdot \sigma(v_{ij})}_{:=\text{Regularizer}}\end{aligned} \tag{5}$$

where $p_I(I)$ denotes a distribution over interventions (uniform in our case) and $X \sim \mathcal{D}_{int(l)}$ refers to a a set of data drawn from the interventional dataset $\mathcal{D}_{int(l)}$. As in phase 1 (i.e. distribution fitting), masks M are sampled from $p(M; u, v)$ which represents a distribution over adjacency matrices. For a detailed optimization formulation and gradient derivations, we refer to Lippe et al. (2021).

C.3. Adaptation Techniques

During test time, we aim to adapt pretrained models to interventional distributions that were not presented during training. Given a set of adaptation samples $\mathcal{D}_{int}^A$, we consider:

(a) Unconstrained Adaptation. Finetune all MLPs using $\mathcal{D}_{int}^A$ for multiple gradient steps.

(b) Sparse Adaptation. We only finetune the module that was affected by the intervention. The affected module is either known or predicted in the setting of unknown interventions.

(c) Regularized Adaptation. By relaxing the Sparse Mechanisms hypothesis, we aim to compute adaptation weights that control the magnitude of adaptation (lower adaptation where knowledge can be directly reused, more adaptation where

knowledge has changed). To this end, we compute the NLL on the adaptation data for every variable X_i and store it as score s_i, assessing the fit of each MLP given the new transfer data. We use this score to compute the weight of adaptation of a certain mechanism by employing a temperature-scaled softmax over these scores $[s_1, \dots, s_N]$. The temperature t allows to control the magnitude of co-adaptation and interpolate between unconstrained adaptation ($t = \infty$) and sparse adaptation ($t = 0$).

C.3.1. Hyperparameter Setup

Hyperparameters:	`Pseudo-LL`, `EXP-Causal`, `EXP-AntiCausal` and `EXP-Skeleton`
Optimizer:	Adam (Kingma & Ba, 2014)
Learning Rate:	$\{0.1, 0.01, 0.001, \underline{0.0001}\}$
Weight Decay:	$\{0.01, 0.001, \underline{0.0}\}$
Number of iterations:	$\{500, \underline{1000}, 2000\}$

Hyperparameters	`L-Causal`
Number of Alternating Iterations:	30
Distribution Fitting:	
Optimizer:	Adam (Kingma & Ba, 2014)
Learning Rate:	$\{0.1, 0.01, 0.001, \underline{0.0001}\}$
Weight Decay:	$\{0.01, 0.001, \underline{0.0}\}$
Number of iterations:	$\{500, \underline{1000}, 2000\}$
Graph Fitting*:	
Optimizer:	Adam (Kingma & Ba, 2014)
Learning rate u:	0.005
Learning rate v:	0.02
Number of iterations:	100
Number of Graphs:	100

* All hyperparameters were adopted from Lippe et al. (2021) as we relied on their graph fitting formulation.

Hyperparameters	`MAML`
Inner Loop:	
Optimizer:	SGD
Learning Rate:	$\{\underline{0.1}, 0.01, 0.001\}$
Nr. Iterations:	$\{\underline{1}, 2, 5\}$
Outer Loop:	
Optimizer:	Adam (Kingma & Ba, 2014)
Learning Rate:	$\{0.1, 0.01, 0.001, \underline{0.0001}\}$
Weight Decay:	$\{0.01, 0.001, \underline{0.0}\}$
Nr. Iterations:	$\{500, \underline{1000}\}$
Nr. Tasks per Iteration:	$\{10, \underline{20}, 50\}$

Table 1: Hyperparameters - Model Training

Hyperparameters:	Model Adaptation
Optimizer:	SGD
Learning Rate:	$\{\underline{0.1}, 0.05, 0.01\}$

Table 2: Hyperparameters - Model Adaptation

D. Generalization Convergence of Causal Models (`L-Causal` vs. `EXP-Causal`)

In this analsis, we only focus on *causal models* and seek to compare the performance of the `EXP-Causal` which is provided with the true causal structure upfront, and `L-Causal` which aims to learn the causal structure from data.

Findings. We observe that `EXP-Causal` outperforms `L-Causal` on low training regimes as the employed causal discovery framework can only identify the true causal graph with sufficient amounts of samples (see Figure 7). With increasing amounts of samples, the learned causal structure of `L-Causal` gets closer to the ground-truth structure (see bottom row of Figure 7 for Structural Hamming Distance (SHD) between learned and true structure) and hence the generalization performance improves. Both models attain `Bound-ZeroShot` (blue) as expected with sufficient amount of samples. In addition, we observe slower convergence of `L-Causal` to `Bound-ZeroShot` on dense graphs than on sparse graph, as the identification of the causal structure is more challenging in such settings (Ke et al., 2019; Scherrer et al., 2021).

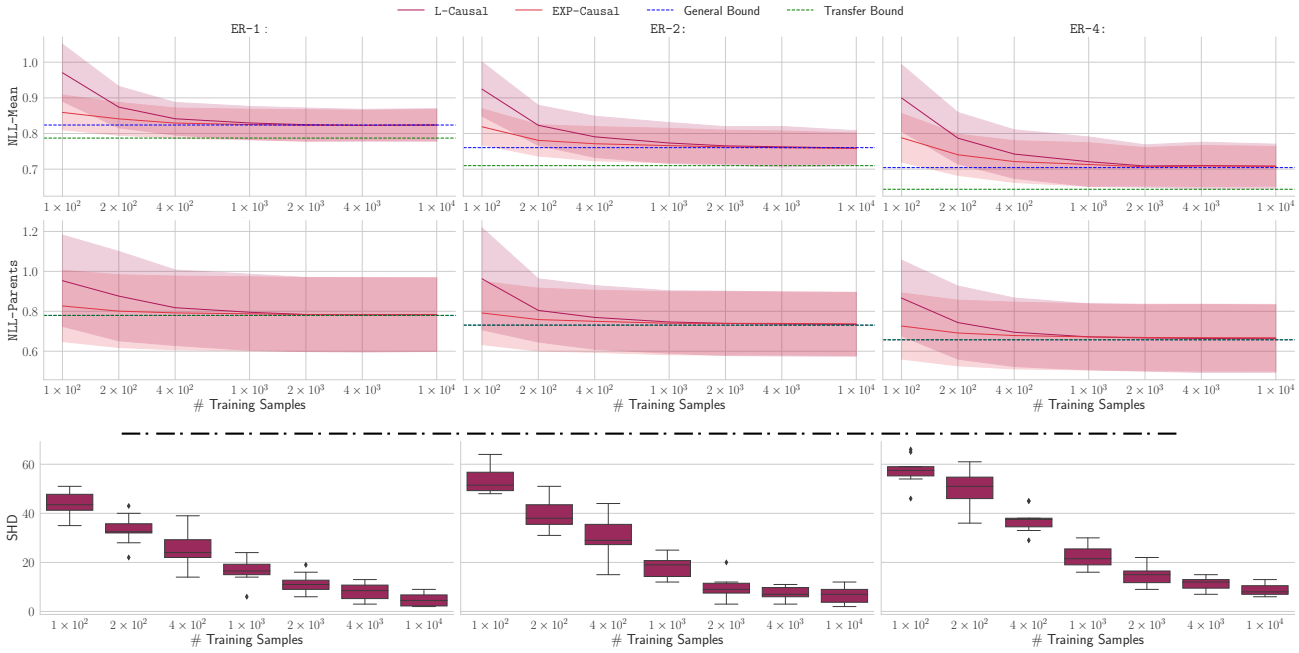


Figure 7: Convergence Behaviour of *Causal Models*. `EXP-Causal` outperforms `L-Causal` on low training regimes as `L-Causal` can only identify the true causal graph with sufficient amounts of samples. With increasing amount of training samples, the structural estimate of `L-Causal` improves (see bottom row) and hence the generalization performance improves and converges to `Bound-ZeroShot`.

E. Efficient Adaptation of Causal Models

Up to this experiment, we have observed the effects of an unconstrained adaptation objective where all modules can be updated, and the sparse adaptation objective where only a specific module is updated (i.e. module is either estimated or known). However, it would be desirable to update the modules in a more efficient manner using less amount of adaptation samples without overfitting. To this end, we employ our proposed regularized adaptation objective and investigate the effect on the speed of adaptation.

Across our experiments, we observed how the proposed adaptation objective further improves the speed of adaptation if only less amount of adaptation samples are available (as shown in Figure 8). It especially improves adaptation if the pretrained model is only trained on few amounts of data. Further, it improves the statistical efficiency compared to the sparse adaptation objective, as available samples are used to update necessary module, if necessary.

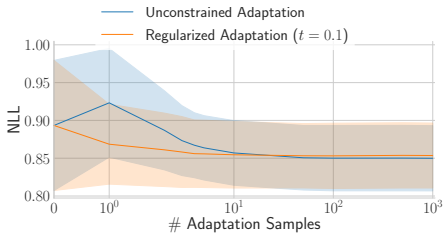


Figure 8: Efficient Adaptation of a causal model. Regularized Adaptation leads to efficient adaptation with low amounts of samples and improves the speed of adaptation compared to unconstrained adaptation.

F. Extended Analysis of Generalization Performance

In this section, we provide an extended generalization analysis of the one presented in Section 4. with additional results and investigations on various settings ($N \in \{10, 20\}$, Graphs: `ER-1`, `ER-2`, `ER-3`). We start with an in-depth analysis of a sparse setting ($N = 20$, `ER-1`, Nr. Training Samples: 1000) in Section F.1 and thereby highlight the importance of the average evaluation metric dissection. In a second step, we analyze in Section F.2 how the generalization performance is affected as the size of SEM increases. As a final step, we provide the complete results on all sub-metrics across all evaluated settings in Section F.3.

F.1. Case Analysis: $N = 10$, `ER-1` Graph

As shown in Section 4.2, an average evaluation metric such as `NLL-Mean` may not not help us to understand the model's performance in detail and does not provide enough insights where models are prone to fail. In order to highlight the importance of the introduced sub-metrics for evaluating the generalization robustness, we fix a sparse evaluation setting (Graph-Type: `ER-1`, $N = 20$, Nr. Training Samples: 1000) where all models perform similar with respect to `NLL-Mean` and dissect the results in detail.

Findings. We observe that *monolithic models* (i.e. `Pseudo-LL` and `MAML`) and `EXP-Skeleton` slightly outperform the two *causal models* (i.e. `L-Causal`, `EXP-Causal`) and the *anti-causal model* on the `NLL-Mean metric` (see Figure 9). However, by looking at the submetrics in more detail, we find that this performance advantage on `NLL-Mean` is due to performance differences on `NLL-Intervention` and `NLL-Root` where causal models are not expected to achieve similar performance as they only rely on causal predictors (i.e. an empty set for root and hard-intervened variables). By excluding all variables where the set of causal predictors is empty (i.e. root and intervention nodes), we arrive at the `NLL-Remainder` metric. On this metric, we observe that all models achieve similar performance ranges except the *anti-causal model* (i.e. `EXP-AntiCausal`) which yields significantly reduced performance. Finally, we focus on the `NLL-Parents` metric, where we only evaluate the ability to predict the parents variables $X_{pa(i,G)}$ of a given intervention target X_i which induced the present distribution shift by a perfect intervention $do(X_i)$. While *non-causal models* can catastrophically fail on this task, we observe that *causal models* maintain their performance and outperform all other models. In summary, all models show difficulties to predict the intervened variables as one expects. On all the remaining variables, *causal models* yield the most robust performance without tendencies for catastrophic failure.

F.2. Generalization Performance Across Graphs of Increasing Size

In this analysis, we seek to investigate how the generalization performance of the considered models changes as the size of the underlying graphs and the corresponding SEM increases. To this end, we fix a sparse class of graphs (i.e. `ER-1`) and analyze the sub-metrics under training datasets $\mathcal{D}^T$ of different size, i.e. $|\mathcal{D}^T| \in \{100, 200, 1000\}$.

Findings. In line with our observation that *structured models* are more sample-efficient than *monolithic models* with respect

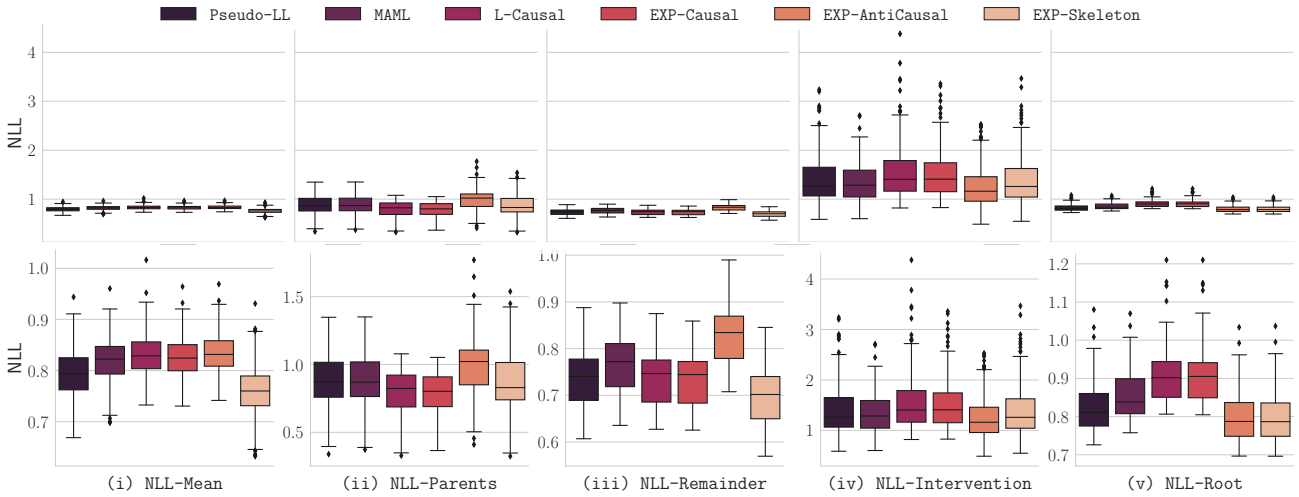


Figure 9: NLL Dissection (Graph-Type: `ER-1`, $N = 20$, Nr. Training Samples: 1000). Reporting the sub-metrics on the same scale (top row) clearly shows that `NLL-Parents` and `NLL-Intervention` are yielding NLL scores on a different scale. Therefore, we zoom in and show all sub-metrics on their own scale (bottom row). While all models achieve comparable results on most metrics, we observe that *non-causal models* can catastrophically fail to predict the parent variables of an intervened variables (i.e. `NLL-Parents`). In contrast, *causal models* maintain their performance and outperform all models on the `NLL-Parents` metric. Furthermore, we observe advantages of *non-causal models* over *causal models* on the `NLL-Intervention` and `NLL-Root` metrics which is in line with our expectation as *non-causal models* make use of non-causal predictors.

to the generalization performance, we find that the performance gap between *structured models* and *monolithic models* widens significantly as the size of the graph increases (i.e. from $N = 10$ to $N = 20$). In particular, we observe a remarkable generalization behaviour of the causal model `EXP-Causal` where the true causal structure is provided upfront. Under a fixed size of the training data, the model maintains robust performance over all metrics when the size of the graph increases. Within the models that are not provided with any domain knowledge upfront, we observe that `L-Causal` and `MAML` clearly outperform `Pseudo-LL` on low-sample regimes. As the number of training samples increases, `L-Causal` is capable of fully identifying the underlying causal structure from data and reaches the same performance as `EXP-Causal`. In general, the performance gap between the models decreases as the amount of training samples is increased.

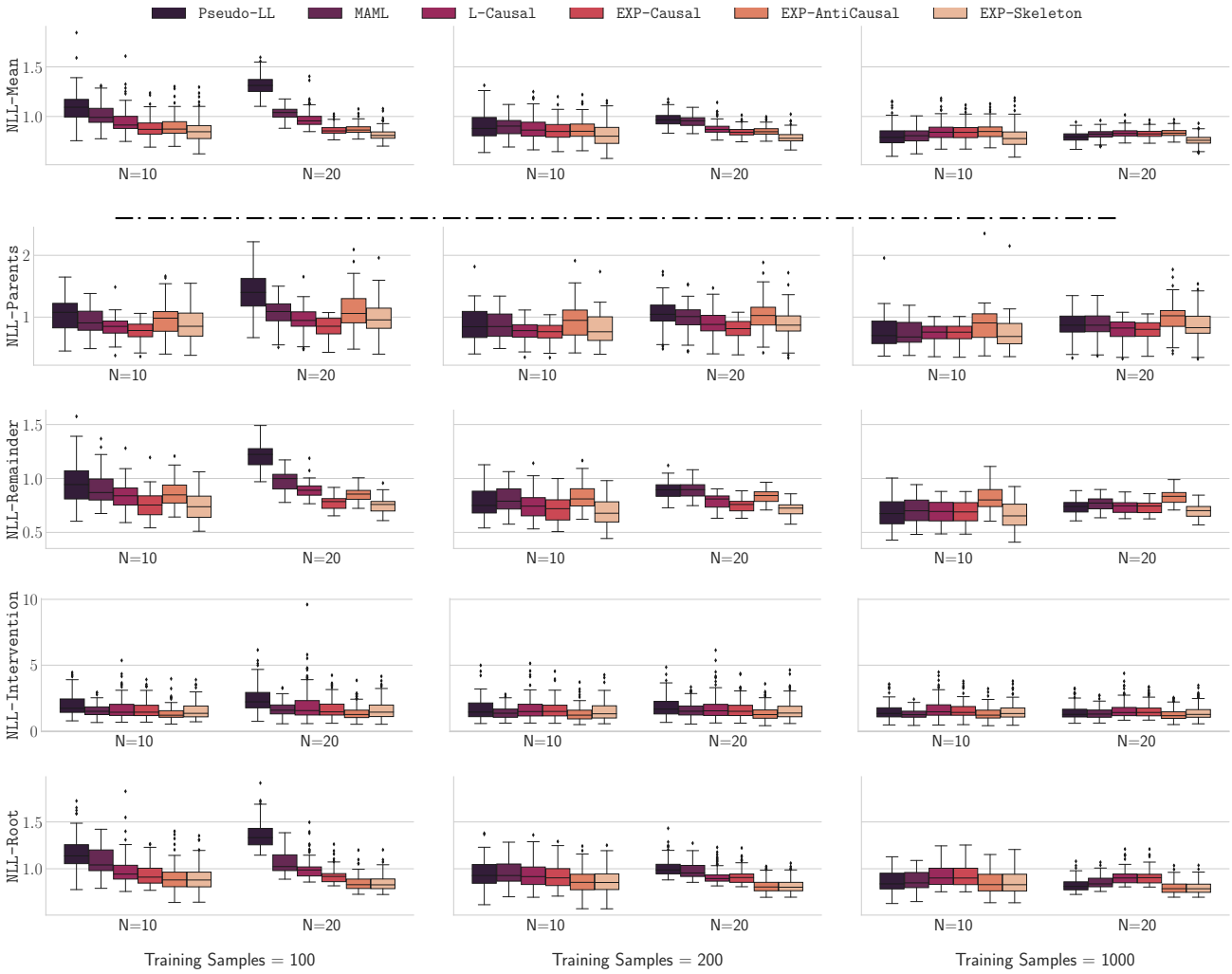


Figure 10: Generalization Performance Across Graphs of Increasing Size. We report the dissected NLL evaluation metrics on ER-1 graphs of size $N \in \{10, 20\}$. We observe that the performance gap between *structured models* and *monolithic models* widens significantly as the size of the graph increases.

F.3. Dissection - Results Across All Settings

In this section, we report all evaluated (sub)-metrics across ER graphs of varying density (i.e. ER-1, ER-2 and ER-3 of size $N \in \{10, 20\}$ on different amount of training samples $|\mathcal{D}^T| \in \{100, 200, 400, 1000, 2000\}$.

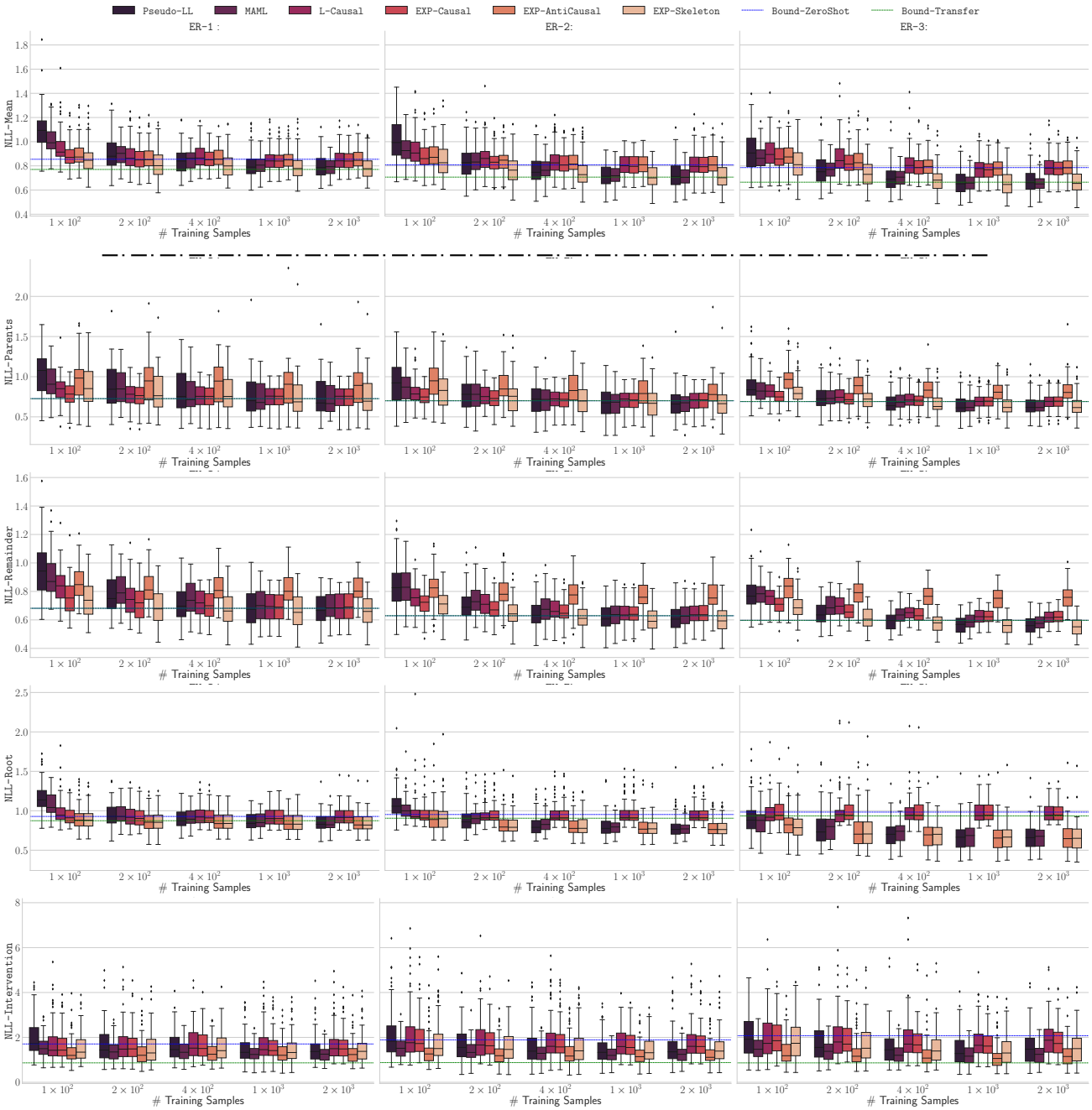


Figure 11: Dissection of OOD Generalization with Varying Amounts of Training Data (N = 10). We report all sub-metrics (i.e. one per row) over various ER graphs (i.e. ER-1, ER-2 and ER-3). The dissection reveals that important failure and robustness insights are hidden in the general evaluation score NLL-Mean (top row). While *non-causal models* yield slightly better performance on NLL-Root and NLL-Intervention, we observe that they can catastrophically fail to predict the parent variables of an intervened variables (i.e. NLL-Parents), especially on sparse graphs. Within the *structured models*, we observe that the EXP-Skeleton model that relies on causal and anti-causal predictors performs best across most settings, but is also prone to fail on NLL-Parents.

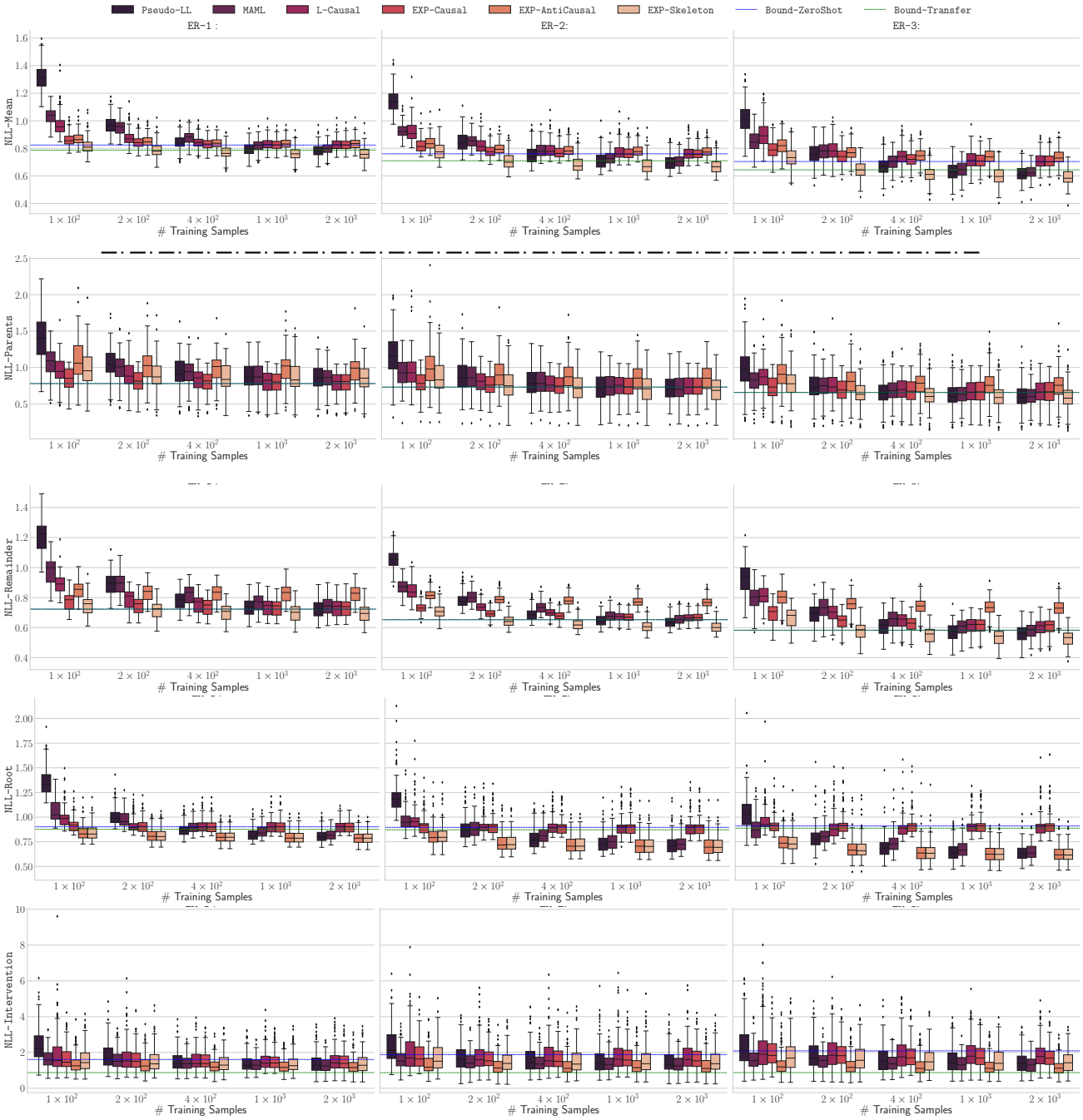


Figure 12: Dissection of OOD Generalization with Varying Amounts of Training Data($\mathbf{N = 20}$). We report all sub-metrics (i.e. one per row) over various `ER` graphs (i.e. `ER-1`, `ER-2` and `ER-3`). The dissection reveals that important failure and robustness insights are hidden in the general evaluation score `NLL-Mean` (top row). While *non-causal models* yield slightly better performance on `NLL-Root` and `NLL-Intervention`, we observe that they can catastrophically fail to predict the parent variables of an intervened variables (i.e. `NLL-Parents`), especially on sparse graphs. Within the *structured models*, we observe that the `EXP-Skeleton` model that relies on causal and anti-causal predictors performs best across most settings, but is also prone to fail on `NLL-Parents`.

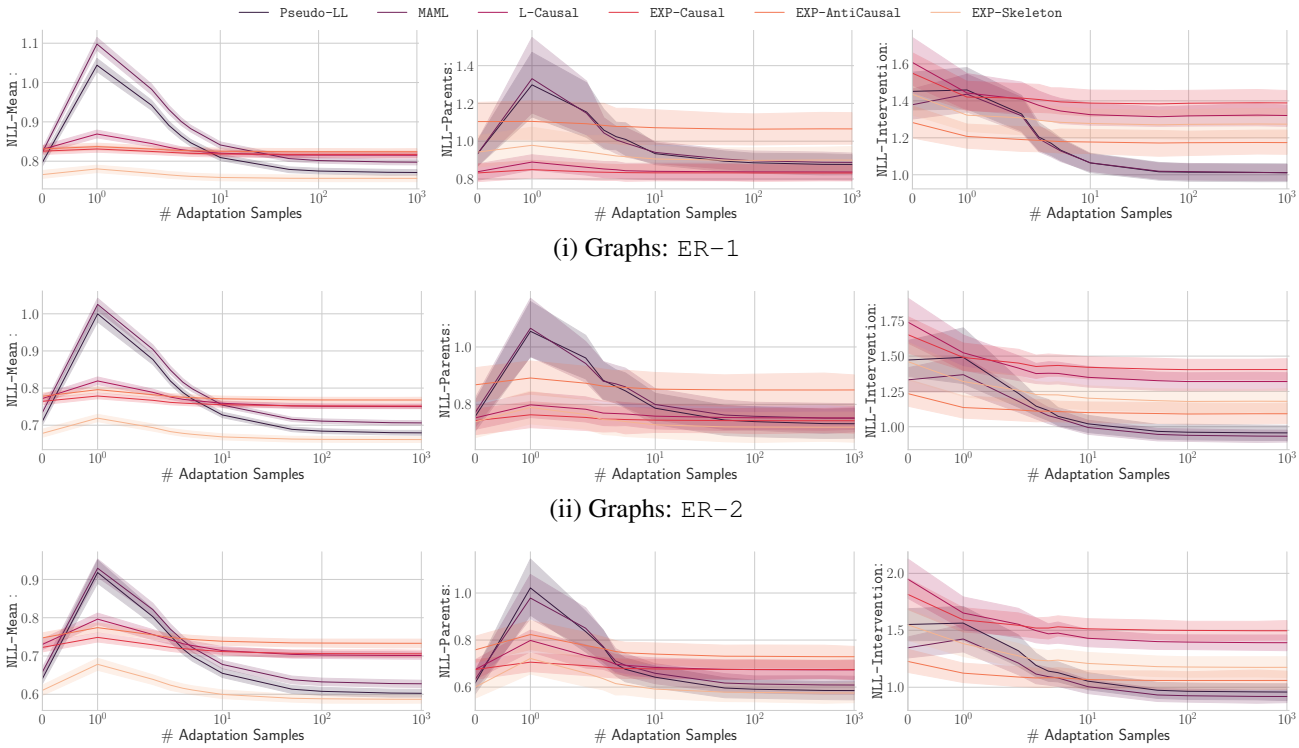
G. Extended Analysis of Adaptation Performance

In this section, we expand our few-shot adaptation analysis from Appendix E with respect to speed of adaptation and overfitting behaviour in Section G.1, and the effect on the parameter space θ in Section G.2. In addition, we provide further results and analysis on the regularized adaptation objective.

G.1. Adaptation Performance

As in Appendix E, we fix the training data size of all models to be 10^3 samples, as all models have converged on the generalization performance by then. We analyze the speed of adaption of different models by evaluating their adaption performance when fine-tuning using different amounts of adaptation data.

Findings. Across all evaluated classes of graphs (e.g. `ER-1`, `ER-2` and `ER-3`), we observe that *structured models* adapt considerable faster than *monolithic models* with respect to the required amount of adaptation samples. When doing a few gradient steps using SGD (i.e. 3 steps with a learning rate of 0.1), we already observe strong overfitting effects for the considered *monolithic models* on all evaluated metrics except `NLL-Intervention` (see Figure 13). By inspecting the `NLL-Parents` metric, we observe the robustness of the two causal models (i.e. `L-Causal` and `EXP-Causal`), especially on sparse graphs.



(i) Graphs: `ER-1`

(ii) Graphs: `ER-2`

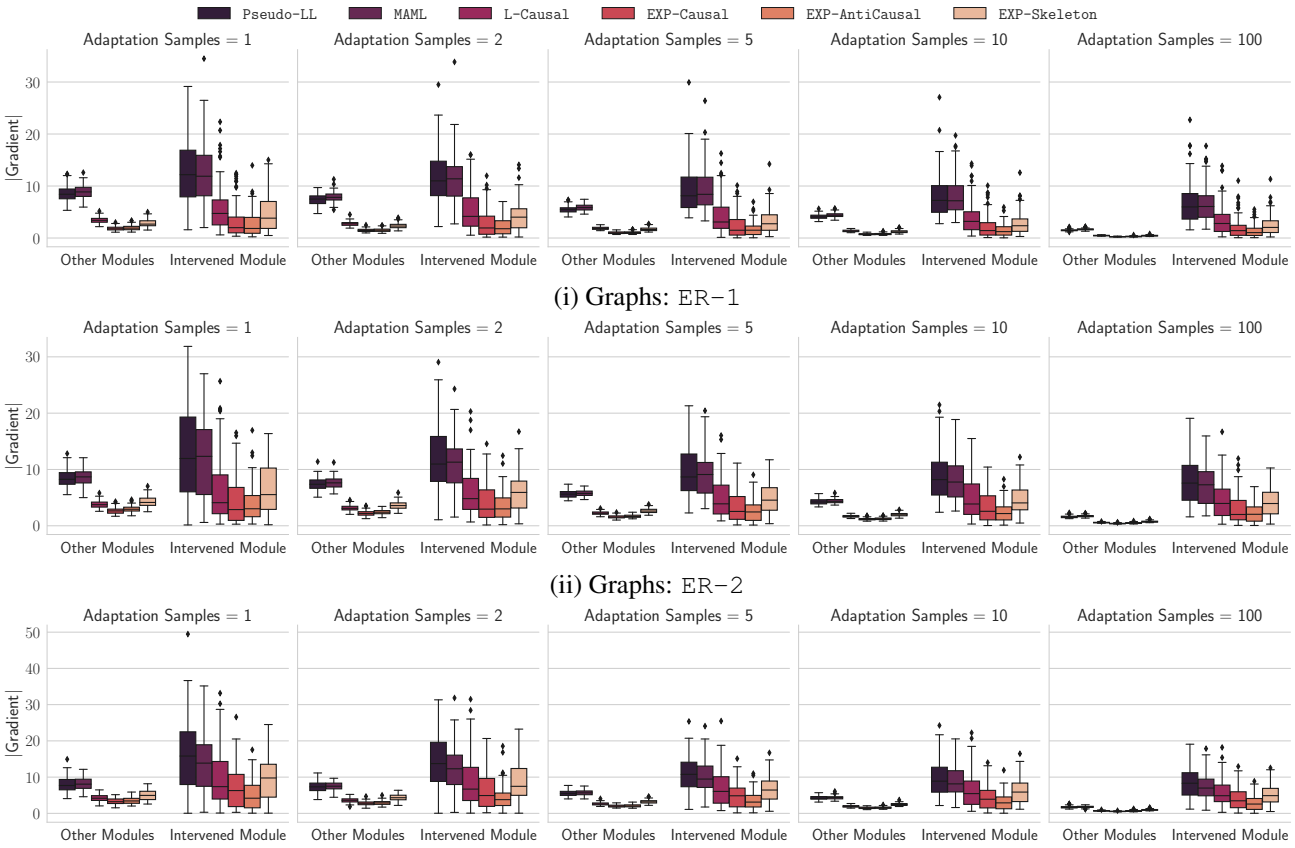
(iii) Graphs: `ER-3`

Figure 13: Speed of Adaptation in terms of different metrics ($N = 20, \mathcal{D}^T = 1000$, 3 **Gradient Steps).** *Structured models* adapt considerably faster than *monolithic models* across all settings and metrics. *Monolithic models* show a sensitivity to overfitting on all classes of graph if only low amounts of adaptation samples are available. In contrast, *structured models* adapt smoothly to the transfer distribution with significantly reduced overfitting effects.

G.2. Parameter Space Analysis

Keeping the adaptation performance of the previous section in mind, we now expand our analysis on the parameter space θ. We seek to answer if the adaptation performance is related to the changes in parameter space θ.

Findings. We find that the overfitting behaviour of *monolithic models* is correlated with the observed updates in parameter space. For the range of adaptation samples where the *monolithic models* are prone to overfit (i.e. 1 to 10 adaptation samples), we observe high gradient magnitudes on the non-intervened modules (referred to as other modules) in *monolithic models* compared to the relatively small updates of *structured models*. As the size of adaptation samples increases (i.e. 100 adaptation samples), we observe significantly reduced gradient magnitudes on non-intervened modules and lower overfitting effects. Within the *structured models* that are built upon structural domain knowledge, we observe that `EXP-Causal` and *EXP-AntiCausal* yield relatively small gradient updates compared to `EXP-Skeleton`. In addition, we observe that the `anti-causal model` yields lower updates on the intervened module as expected as it relies on anti-causal predictors of the intervened variable children, that were not affected by the intervention.



(i) Graphs: `ER-1`

(ii) Graphs: `ER-2`

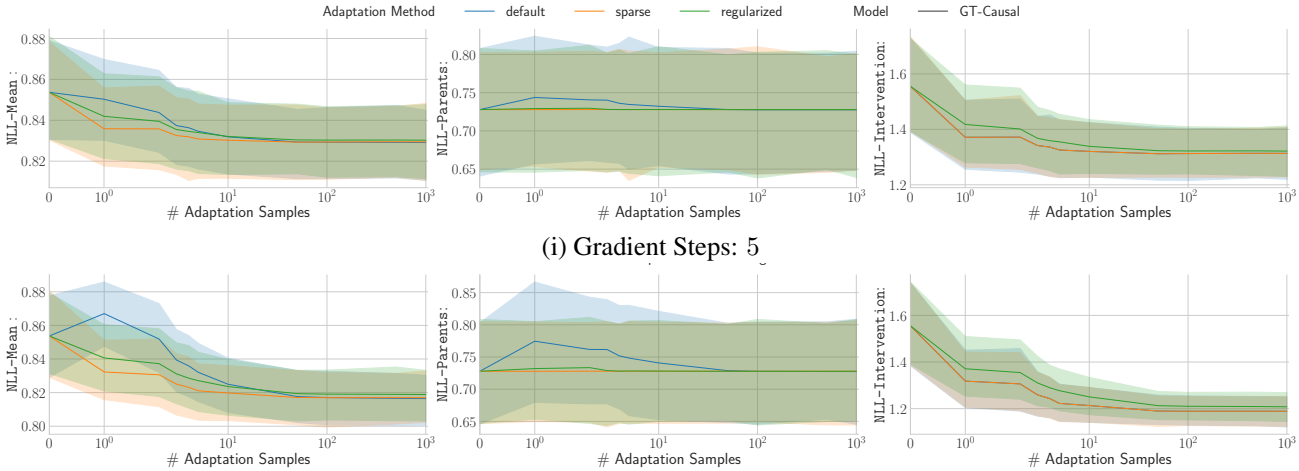
(iii) Graphs: `ER-3`

Figure 14: Parameter Space Analysis ($N = 20, \mathcal{D}^T = 1000$)**.** While adapting to a shift in distribution with an unconstrained adaptation objective using a single gradient step, *monolithic models* update most modules that were not affected by the intervention quite heavily compared to *structured models*. *Causal and anti-causal models* show remarkable adaptation behaviour in parameter space with localized updates on intervened modules.

G.3. Regularized Adaptation

In this section, we provide further insights on the effects of a a regularized adaptation objective on top of `EXP-Causal`. We report speed of adaptation in Figure 15 and analyses on the parameter space in Figure 16.

Findings. We observe that the regularized adaptation objective improves the adaptation performance on low amounts of adaptation samples considerably. Our results indicate that the adaptation objective prevents from overfitting if multiple gradient steps are performed. It is noteable, that the regularized adaptation objective yields nearly the same performance as the sparse adaptation objective, even though the sparse adaptation objective leverages a supervised signal (i.e. knowledge of the intervention location) in the present setting.



(i) Gradient Steps: 5

(ii) Gradient Steps: 10

Figure 15: Regularized Adaptation: Effects on Speed of Adaptation ($N = 10, \mathcal{D}^T = 1000$)**.** For 5 gradients steps using SGD with a learning rate of 0.1, we observe continuously improving adaptation with respect to the `NLL-Mean` metric on all regularization techniques. With 10 gradient steps, we observe an overfitting behaviour of the unconstrained adaptation objective if only low amounts of adaptation samples are available. In contrast, the sparse and regularized adaptation objective still yield continuous improvements, even if only low few adaptation samples are available. It is noteable, that the regularized adaptation objective yields nearly the same performance as the sparse adaptation objective, even though the sparse adaptation objective leverages a supervised signal (i.e. knowledge of the intervention location) in the present setting.

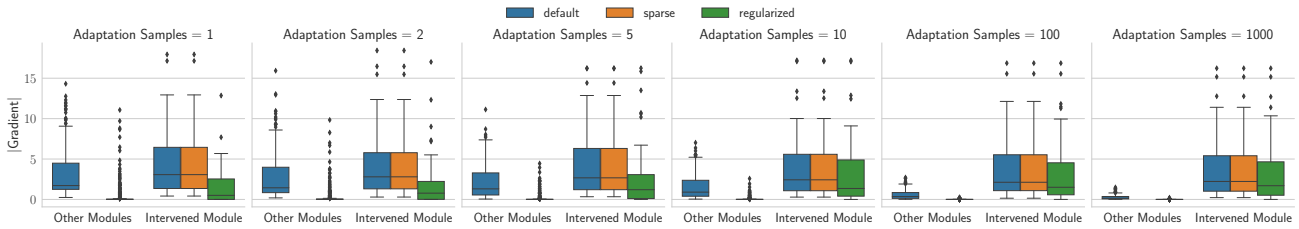


Figure 16: Regularized Adaptation: Effects on Parameter Space θ ($N = 10, \mathcal{D}^T = 1000$)**.** With respect to the parameter space, we observe that the regularized adaptation objective yields smaller updates on the intervened module if only low amounts of adaptation samples are available. In general, the regularized adaptation objective is capable of identifying the intervened mechanisms and only performs updates of low gradient-magnitude on non-intervened modules whereas the unconstrained adaptation objective yields updates of greater magnitude. As the number of adaptation samples increases, the regularized objective yields similar updates on the intervened mechanisms as the other two objectives.