Hypothesis Testing using Causal and Causal Variational Generative Models

*

Anonymous Author(s)
Affiliation
Address
email

Abstract

Hypothesis testing and the usage of expert knowledge, or causal priors, has not been well explored in the context of generative models. We propose a novel set of generative architectures, Causal Gen and Causal Variational Gen, that can utilize nonparametric structural causal knowledge combined with a deep learning functional approximation. We show how, using a deliberate (non-random) split of training and testing data, these models can generalize better to similar, but out-of-distribution data points, than non-causal generative models and prediction models such as Variational autoencoders and Fully Connected Neural Networks. We explore using this generalization error as a proxy for causal model hypothesis testing. We further show how dropout can be used to learn functional relationships of structural models that are difficult to learn with traditional methods. We validate our methods on a synthetic pendulum dataset, as well as a trauma surgery ground level fall dataset.

1 Introduction

In fields such as medicine and economics, an explainable model, in particular a causal model makes diligent use of prior knowledge. This is usually in a structural causal model (SCM) form that instantiates unidirectional relationships between variables using a Directed Acyclic Graph (DAG) [1]. The confidence in a causal model needs to be much higher than in a statistical model, as its beliefs are invariant and exist outside the domain of the data. When ground truth is unavailable, it is often the subject of domain experts to hypothesize and test causal models using experiments or observational data.

Generative models have been crucial to solving many problems in modern machine learning [2], and generating useful synthetic datasets. Causal generative models learn or use causal information for generating data and interventions, producing more interpretable results, and tackling biased datasets [3–5]. Generative Causal Models also have the ability to put constraints on the solution space of a machine learning problem and significantly decrease the training time and allow optimizers to find a better solution.

Recently, [6] introduces a *Causal Layer*, which allows for direct interventions to generate images outside the distribution of the training dataset in its CausalVAE framework. Another method, Causal Counterfactual Generative Modeling (CCGM), in which partial priors are included on the diagonals of the structural model matrix, further extends the counterfactual modeling capabilities by allowing the testing of new models with alternative structure to potentially "de-bias" datasets or to view data outside of the training distribution [7].

*

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Do not distribute.

CausalVAE and CCGM focus on causal discovery, but in many real-world applications, a causal model is available or readily hypothesized. It is often of interest to test various causal model hypotheses for generalization to out-distributions (from the training set) instead of prediction of samples within the distribution of the training set. Thus we propose CGEN and CVGEN, which are generative and variational generative models that forgo causal discovery for causal hypothesis testing. Combined with non-random dataset splits to test generalization to non-overlapping distributions across dimensions, we allow for a systematic way to test structural causal model hypotheses and use those models to generate synthetic data outside training distributions.

2 Background

2.1 Causality and Model Hypothesis Testing

Causality literature has explored the benefits of interventions and counterfactual modeling once a causal model is known [1]. We explore use of the causal layer as a hypothesis testing space. Both CGEN and CVGEN architectures accept non-parametric (structural only, no functional-form) causal priors as a binary Structural Causal Model (SCM) and use deep learning to approximate the functional relationships that minimize a means-squared reconstruction error (MSE).

Structural causal priors are mainly about the ordering and absence of connections between variables. It is this absence of certain edge that prevents information flow, reducing the likelihood that spurious connections are learned within the training dataset distribution. Thus, when comparing our architecture to traditional deep learning prediction and generative models, we show how hypothesized causal models might perform worse when testing within the same distribution as the training data, but drastically improve performance when splitting the test and train distributions to have no overlap across a dimension. This effect is seen the most in small datasets where non-causal methods can "memorize" spurious patterns in the data and vastly overfit the training distribution [8].

3 Causal Gen and Causal Variational Gen

3.1 Causal Hypothesis Testing with

Our model CGEN, uses a similar causal layer as in both CCGM and CausalVAE [6, 7]. The causal layer consists of a binary adjacency matrix $\mathbf{S}$ followed by non-linear functions defined by MLPs. We solidify the structure of $\mathbf{S} \in \{0,1\}^{d \times d}$ as the structural model matrix paths, where $\mathbf{S}$ can be split into a DAG term and a diagonal term:

$$\mathbf{S} = \underbrace{\mathbf{A}}_{\text{DAG}} + \underbrace{\mathbf{D}}_{\text{diag.}} \tag{1}$$

$\mathbf{A}$ represents a DAG adjacency matrix and $\mathbf{D}$ has 1 on the diagonal for exogenous variables and 0 if endogenous. Then, given tabular inputs $\boldsymbol{x} \in \mathbb{R}^d$, $\mathbf{S}_{ij}$ determines whether variable i is a parent of variable j.

From the structural prior $\mathbf{S}$, each of the input variables are “selected" to be parents of output variables through a Hadamard product with the features $\boldsymbol{x}$. For each output variable, its parents are passed through a non-linear η fully connected network-network. The η networks are trained as general function approximators, learning the relationships between parent and child nodes.

$$\hat{\boldsymbol{x}}_i = \eta_i(\mathbf{S}_i \circ \boldsymbol{x}) \tag{2}$$

where $\mathbf{S}_i$ represents the i-th column vector of $\mathbf{A}$, and $\hat{\boldsymbol{x}}_i$ is the i-th reconstructed output [9]. In the case of exogenous variable $\boldsymbol{x}_i$, a corresponding 1 at $\mathbf{D}_{ii}$, ‘leaks’ the variable through, encouraging η to learn the identity function. The end-to-end structure is trained on a reconstruction loss, defined by $\ell(\boldsymbol{x}, \hat{\boldsymbol{x}})$. We use the L2 loss (Mean Squared Error):

$$\ell_{\text{CGEN}} = ||\boldsymbol{x} - \eta_i(\mathbf{S}_i \circ \boldsymbol{x})||_2 \tag{3}$$

3.2 Variational Hypothesis testing and Data Generation with

We extend CGEN to a variational model CVGEN, that includes sampling functionality like a VAE [2]. Thus CVGEN can generate new data points that are not deterministic on the inputs, allowing for synthetic data generation. CVGEN consists of an encoder, a CGEN causal layer and a decoder.

These encoder and decoder networks do not compress the data, but enable a transformation of the inputs to a normally distributed space, enabling sampling without preventing the relevance of the causal priors given by $\mathbf{S}_i$. Thus, CVGEN is an extension of CGEN with

$$\boldsymbol{z}_i = f_{enc}(\boldsymbol{x}_i), \quad \hat{\boldsymbol{z}}_i = \eta_i(\mathbf{S}_i \circ \boldsymbol{z}), \quad \hat{\boldsymbol{x}}_i = f_{dec}(\hat{\boldsymbol{z}}_i) \tag{4}$$

The loss function for CVGEN includes a weighted Kullback–Leibler (KL) divergence loss to normalize the latent space on top of the reconstruction loss as in CGEN. We also add a weighted latent reconstruction loss for the embedded CGEN which enforces separation of the encoder and decoders as transformations, and the η networks as the functional approximators on these transformations.

$$\ell_{KL} = KL(\boldsymbol{z}_i || \mathcal{N}(0,1)) \tag{5}$$

$$\ell_{latent} = \ell(\boldsymbol{z}, \hat{\boldsymbol{z}}) \tag{6}$$

$$\ell_{MSE} = ||\boldsymbol{x} - \eta_i(\mathbf{S}_i \circ \boldsymbol{x})||_2 \tag{7}$$

$$\ell_{\text{CGEN}} = \ell_{MSE} + \lambda_{KL} * \ell_{KL} + \lambda_{latent} * \ell_{latent} \tag{8}$$

3.3 Using Dropout for Improved Functional Forms

The causal structural prior, $\mathbf{S}$, can enforce a variety of causal models, so there are cases where variables may be a function of themselves as well as another variable to account for potential confounders. In these cases, we need to prevent the model from learning a simple identity function of the variable as if it was exogenous, and allow it to learn a relationship with the other input variables as well. We do this using a dropout layer on the inputs into the η_i networks. Dropout allows some parent edges to be blocked with a specified probability each iteration of training, forcing the η functions to learn causal effects for multiple subsets of parent inputs. Thus, given a dropout mask ζ_p with dropout probability p, the new CGEN model becomes:

$$\hat{\boldsymbol{x}}_i = \eta_i(\zeta_p(\mathbf{S}_i \circ \boldsymbol{x})) \tag{9}$$

with an equivalent change occurring in the CVGEN model architecture.

4 Problem Setting

4.1 Sun Pendulum Image Dataset

A synthetic pendulum image dataset is introduced in [6]. This dataset is generated by sweeping sun positions (x_{sun}) and pendulum angles (θ) to produce realistic shadow width (w_{shadow}) and shadow locations (x_{shadow}) from deterministic non-linear functions. Figure 1 shows the true DAG and a couple close, but incorrect, DAGs for this model and an example generated image. Here, the sun and pendulum variables are exogenous, and the shadow variables are endogenous. This methodology provides a physics-based dataset where the causal, ground truth model is known exactly. We take these values $\mathbf{u} = [\theta, x_{sun}, w_{shadow}, x_{shadow}]^T \in \mathbb{R}^d$, where $d = 4$ and compile a tabular dataset to test out the CGEN and CVGEN. Our results later can be visualized in an image form so inference and synthetic data is more interpretable.

4.2 Medical Trauma Dataset

We also analyze our model on a real-world dataset of brain-trauma ground level fall patients that includes multiple health factors along with the focus on predicting a decision to proceed with surgery or not. We used an initial SHAP analysis to select for three variables of high prediction impact: Glasgow Coma Scale/Score for head trauma severity (GCS), Diastolic Blood Pressure (DBP), the presence of any Co-Morbidities (Co-Morb), one demographic variable Age, along with the Surgery outcome of interest. Without any ground truth, we test two plausible structural models shown in 2 based on knowledge of the selected variables and how they might interact to inform the surgery decision.

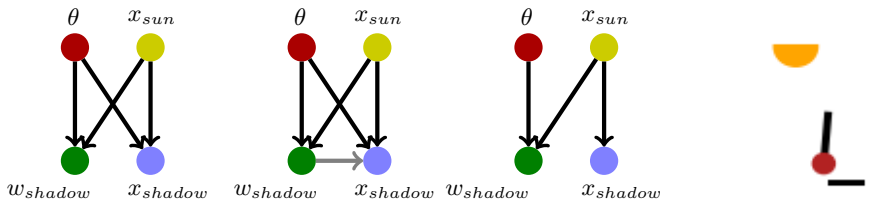


Figure 1: Three hypothesized causal models are tested with the CGEN. (a) is the true causal model. (b) is a causal model with an extra link, introducing leakage. (c) is the causal model with a missing link, so it is under-represented. We pass each model through CGEN. (d) is an example visual representation of the pendulum toy example

Figure 2: Two hypothesized structural causal priors for a medical dataset on trauma patients and decision to perform surgery, H1 and H2.

4.3 Train/Test Data Splits

In order to test generalization error, we use a deliberate non-random split of our datasets (as well as a baseline random split). This is done on a single feature column of the tabular data, splitting the data on that column at either the 25% or 75% quantile, with the larger side (either the upper or lower 75%) becoming the training data. An example of this train test split is visualized for both datasets in 3.

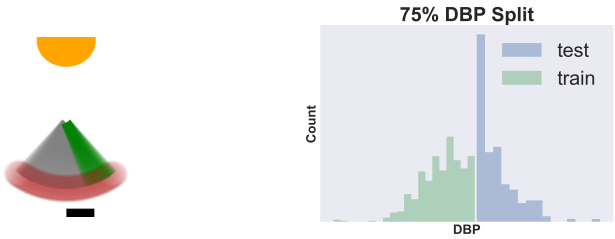


Figure 3: a) A 75% datasplit on the pendulum angle feature (grey is training angle, green is testing angles b) A 75% datasplit on the Diastolic Blood Pressure data

5 Experiments

We test CGEN and CVGEN on the pendulum and ground level fall dataset and compare to a Fully Connected Neural Network and a VAE respectively on random and deliberately split data. The resulting reconstruction losses are listed below

5.1 Results

The CGEN and CVGEN models consistently generalize better than their respective non-causal comparisons on both the pendulum dataset where the ground truth causal prior is known, and on the medical dataset where two medically-informed causal structural priors are hypothesized and tested.

It is clear in 4 both the CGEN and CVGEN better capture the pendulum angle and shadow size of this out-of-distribution test sample than the VAE or NN.

	Pendulum Comparison			
	Random		Split	
Method	Train	Test	Train	Test
NN	0.02	0.02	0.04	10.27
VAE	0.11	0.06	16.97	89.4
CGEN	0.03	0.03	0.02	0.26
CVGEN	0.064	0.51	19.81	38.62

Table 1: Comparison of Traditional Deep Learning Techniques on a random and deliberate dataset split with CGEN and CVGEN when the ground truth causal structural information is known. Caveat for the variational models, the training losses include the other terms as well.

	Hypothesis 1						**Hypothesis 2**					
	Random		DBP 75%		GCS 25%		Random		DBP 75%		GCS 25%	
Method	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
NN	0.03	0.03	0.04	7.16	0.04	0.84	0.03	0.04	0.03	1.02	0	0
VAE	1.39	0.91	1.43	11.19	0.63	2.01	1.23	0.86	1.16	8.78	0	0
CGEN	0.02	0.02	0.02	0.04	0.02	0.03	0.02	0.02	0.02	0.03	0	0
CVGEN	7.09	8.21	0.74	0.23	0.39	1.08	7.1	8.17	0.74	0.17	0	0

Table 2: H1 and H2 results for the medical dataset across one random and two deliberate dataset splits. See appendix for variances.

For the medical dataset, the reconstruction of the surgery probabilities in 5 shows CGEN and CVGEN both produced slightly better balanced distributions while the NN and VAE are biased slightly towards either outcome.

5.2 Causal Hypothesis Testing

We demonstrate how CGEN can be used to detect incorrect causal hypotheses. Here, we consider three causal structures, shown in Figure 1.

Pendulum Results		
Hypothesis	Train	Test
True (a)	0.01	0.01
Leaked (b)	0.02	0.02
Under (c)	0.52	19.28

Medical Results		
Hypothesis	Train	Test
H1	0.024	0.035
H2	0.017	0.025

Table 3: Pendulum and Medical results of CGEN for each of the hypothesized causal models.

For the pendulum dataset, we clearly see in 3 how the ground truth model achieves the lowest train and test error, indicating it would be the natural choice after hypothesizing and testing the three models. The leaked causal model cannot generalize as well to the out-of-distribution points despite having more connections. The under-determined causal model performs the worst, lacking necessary information to generalize or even minimize training loss. In the medical dataset, it is the second hypothesis from 2, which includes a path from Age to No-Comorbidities that generalizes better indicating a likely better causal model.

Figure 4: The plots offer a visual example of a specific instance of of each model's reconstruction in the following order: a) ground truth b) CGEN c) CVGEN d) NN C) VAE.

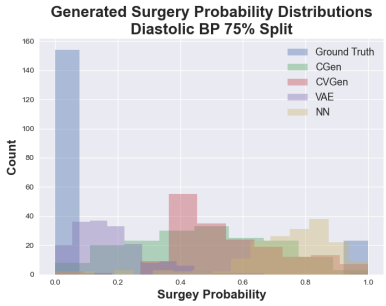


Figure 5: The test dataset reconstruction distributions for all models using medical H1 on the Diastolic BP 75% data.

6 Conclusion

In this paper, we demonstrate the value of CGEN and CVGEN as causal model hypothesis testing spaces and generative models. We verify their performance with ground truth and incorrect causal priors on a physics dataset while comparing to standard deep learning prediction and generative models. We show how these architectures can be used to test causal hypotheses using a real-world medical dataset with ground level fall, trauma surgery decisions. CGEN and CVGEN consistently outperform baselines and offer a novel architecture, along with a deliberate data split methodology, that can empower practitioners and domain experts to improve causally informed modeling and deep learning. We also hope to extend both CGEN and CVGEN to more flexible architecture which can combine recent progress with differential causal inference and binary sampling to better automate full or partial causal discovery. The results are a promising start to much further deep learning causal modeling and discovery research.

References

[1] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, no. none, pp. 96 – 146, 2009. [Online]. Available: https://doi.org/10.1214/09-SS057

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. [Online]. Available: https://arxiv.org/abs/1312.6114

[3] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath, "Causalgan: Learning causal implicit generative models with adversarial training," 2017. [Online]. Available: https://arxiv.org/abs/1709.02023

[4] Y. Yu, J. Chen, T. Gao, and M. Yu, "Dag-gnn: Dag structure learning with graph neural networks," 2019. [Online]. Available: https://arxiv.org/abs/1904.10098

[5] B. van Breugel, T. Kyono, J. Berrevoets, and M. van der Schaar, "Decaf: Generating fair synthetic data using causally-aware generative networks," 2021. [Online]. Available: https://arxiv.org/abs/2110.12884

[6] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "Causalvae: Disentangled representation learning via neural structural causal models." [Online]. Available: https://arxiv.org/abs/2004.08697

[7] S. Bhat, J. Jiang, O. Pooladzandi, and G. Pottie, "De-biasing generative models using counterfactual methods," 2022. [Online]. Available: https://arxiv.org/abs/2207.01575

[8] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," 2017. [Online]. Available: https://arxiv.org/abs/1706.05394

[9] I. Ng, S. Zhu, Z. Fang, H. Li, Z. Chen, and J. Wang, "Masked gradient-based causal structure learning," 2019. [Online]. Available: https://arxiv.org/abs/1910.08527

[10] ——, "Masked gradient-based causal structure learning," 2019. [Online]. Available: https://arxiv.org/abs/1910.08527

A Appendix

A.1 Background Theory

A.1.1 Constructing a Causal Generative Model

Following the classic VAE model, given inputs $\mathbf{x}$, we encode into a latent space $\mathbf{z}$ with distribution q_ϕ where we have priors given by $p(\cdot)$ [2].

$$\text{ELBO} = \mathbb{E}_{q_\mathcal{X}}\left[\mathbb{E}_{\mathbf{z}\sim q_\phi}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathcal{D}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))\right] \tag{10}$$

In [6], the causal layer is described as a noisy linear SCM:

$$\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} \tag{11}$$

which finds some causal structure of the latent space variables $\mathbf{z}$ with respect to a matrix $\mathbf{A}$. By itself, $\mathbf{A}$ functions as the closest linear approximator for the causal relationships in the latent space of $\mathbf{z}$.

A non-linear mask is applied to the causal layer so that it can more accurately estimate non-linear situations as well. Suppose $\mathbf{S}$ is composed of column vectors $\mathbf{S}_i$. For each latent space concept i, define a non-linear function $g_i : \mathbb{R}^n \to \mathbb{R}$ and modify equation (11) such that

$$\mathbf{z}_i = g_i(\mathbf{S}_i \circ \mathbf{z}) + \boldsymbol{\epsilon} \tag{12}$$

where $\circ$ is the Hadamard product. In this formulation, the view of $\mathbf{A}$ changes from one of function estimation to one of adjacency. That is, if $\mathbf{S}$ is viewed as a binary adjacency matrix, the g_i functions take the responsibility of reconstructing $\mathbf{z}$ given only the the parents, dictated by $\mathbf{S}_i \circ \mathbf{z}$. In the simplest case, if $g_i(\mathbf{v}) = \sum_j v_j$, the summation of all the values of $\mathbf{v}$, then Equation (12) degenerates back to Equation (11) [10].

Including the causal layer introduces many auxiliary loss functions that we mostly adopt [6]. First is a label loss (13), where the adjacency matrix $\mathbf{S}$ should also apply to the labels $\mathbf{u}$. This loss is used in pre-training in its linear form to learn a form of $\mathbf{S}$ prior to learning the encoder and decoders. After pre-training, we apply a nonlinear mask f_i that functions similarly to g_i, but operates on the label space directly, but with the same $\mathbf{S}$.

$$\ell_u = \mathbb{E}_{q_\mathcal{X}}\left[\sum_{i=1}^{n}\|u_i - f_i(\mathbf{A}_i \circ \mathbf{u})\|^2\right] \tag{13}$$

The latent loss tries to enforce the SCM, described by Equation (12).

$$\ell_z = \mathbb{E}_{\mathbf{z}\sim q_\phi}\left[\sum_{i=1}^{n}\|z_i - g_i(\mathbf{A}_i \circ \mathbf{z})\|^2\right] \tag{14}$$

Further enforcing the label spaces, we can define a prior $p(\mathbf{z}|\mathbf{u})$. We use the same conventions as in [6] and say that

$$p(\mathbf{z}|\mathbf{u}) \sim \mathcal{N}(\mathbf{u}_n, \mathbf{I})$$

where $\mathbf{u}_n \in [-1, 1]$ are normalized label values. This translates to an additional KL-loss.

A.2 Final Losses with Sample Variances

A.2.1 Pendulum Model Comparisons (True Causal Model)

Split	Model	Train	Test
Random	CGen	$0.03 \pm 6.83e+01$	$0.03 \pm 7.25e+01$
	CVGen	$22.52 \pm 2.27e-01$	$0.51 \pm 2.88e-01$
	VAE	$43.62 \pm 1.57e+03$	$159.64 \pm 7.44e+03$
	NN	$0.02 \pm 5.10e-05$	$0.02 \pm 2.30e-04$
75% Pend	CGen	$0.02 \pm 2.25e+01$	$0.26 \pm 2.26e+01$
	CVGen	$19.81 \pm 3.00e-03$	$38.62 \pm 6.88e+00$
	VAE	$16.97 \pm 9.03e-02$	$89.4 \pm 7.32e+02$
	NN	$0.04 \pm 6.96e-06$	$10.27 \pm 1.44e+00$

A.2.2 Pendulum Causal Hypotheses

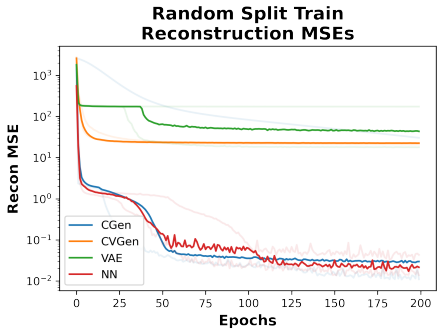
Hypothesis	Train	Test
true	$0.01 \pm 1.03e-05$	$0.01 \pm 6.04e-04$
leaked	$0.016 \pm 5.75e-06$	$0.018 \pm 1.04e-04$
under	$0.522 \pm 1.94e-04$	$19.277 \pm 7.10e+01$

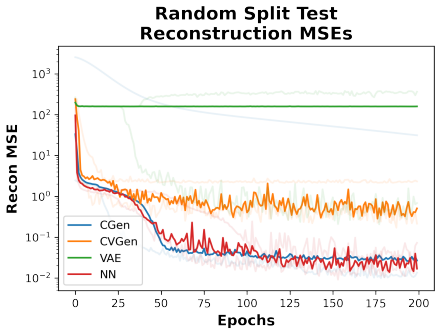
A.2.3 Medical Dataset

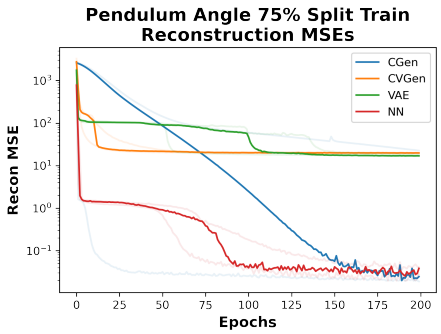
Hypothesis	Split	Model	Train	Test
H1	Random	CGen	$0.02 \pm 2.34e-07$	$0.02 \pm 9.42e-08$
		CVGen	$7.09 \pm 1.71e+00$	$8.21 \pm 2.44e+00$
		VAE	$1.39 \pm 4.71e-03$	$0.91 \pm 5.20e-03$
		NN	$0.03 \pm 3.15e-06$	$0.03 \pm 1.09e-05$
	DBP 75%	CGen	$0.02 \pm 2.00e-06$	$0.04 \pm 3.82e-06$
		CVGen	$0.74 \pm 1.81e-02$	$0.23 \pm 2.32e-02$
		VAE	$1.43 \pm 1.23e+00$	$11.19 \pm 2.65e+00$
		NN	$0.04 \pm 1.04e-06$	$7.16 \pm 3.75e-02$
	GCS 25%	CGen	$0.02 \pm 4.67e-07$	$0.03 \pm 2.67e-06$
		CVGen	$0.39 \pm 4.62e+00$	$1.08 \pm 4.28e+00$
		VAE	$0.63 \pm 7.46e-04$	$2.01 \pm 3.25e-04$
		NN	$0.04 \pm 2.83e-04$	$0.84 \pm 2.53e-03$
H2	Random	CGen	$0.02 \pm 6.01e-08$	$0.02 \pm 3.88e-07$
		CVGen	$7.1 \pm 3.82e+00$	$8.17 \pm 5.36e+00$
		VAE	$1.23 \pm 3.97e-04$	$0.86 \pm 4.24e-05$
		NN	$0.03 \pm 2.99e-03$	$0.04 \pm 1.46e-03$
	DBP 75%	CGen	$0.02 \pm 5.07e-07$	$0.03 \pm 6.59e-08$
		CVGen	$0.74 \pm 3.66e-03$	$0.17 \pm 3.16e+01$
		VAE	$1.16 \pm 1.73e-03$	$8.78 \pm 1.03e+00$
		NN	$0.03 \pm 1.44e-07$	$1.02 \pm 9.80e-03$
	GCS 25%	CGen	$0.02 \pm 1.30e-07$	$0.06 \pm 5.45e-05$
		CVGen	$0.36 \pm 2.44e+00$	$3.83 \pm 3.94e+00$
		VAE	$0.53 \pm 2.87e-05$	$3.44 \pm 5.74e-02$
		NN	$0.09 \pm 2.72e-04$	$1.63 \pm 1.46e-01$

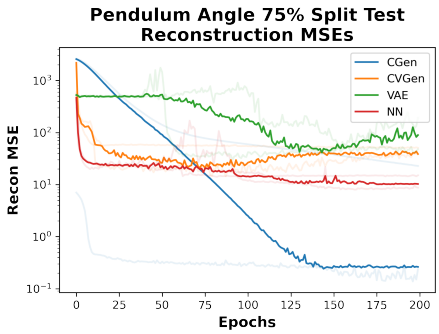
A.3 Train and Test Curves

A.3.1 Pendulum Dataset - H1 (Correct)

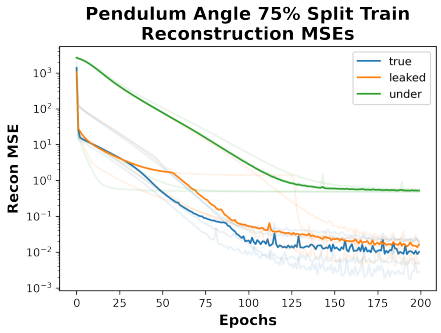


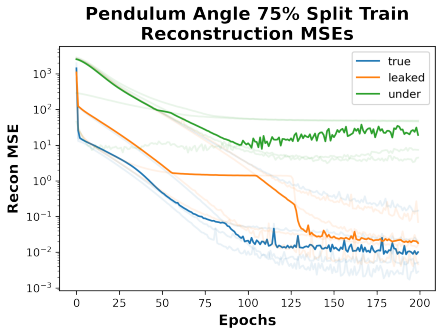




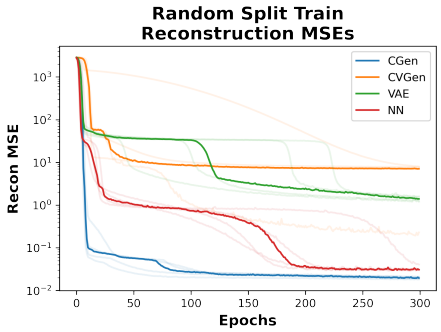


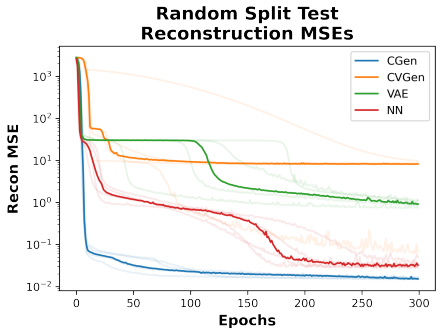
A.3.2 Pendulum Dataset - H2 (Incorrect)

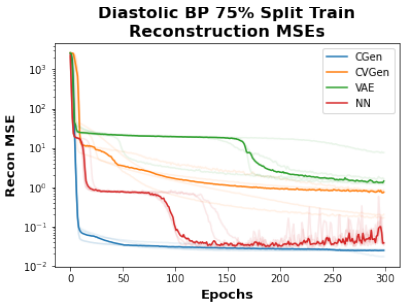


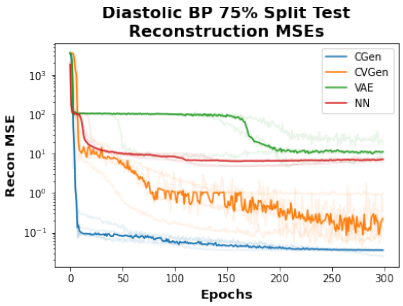


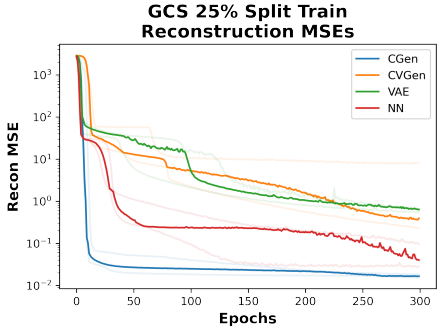
A.3.3 Medical Dataset - H1

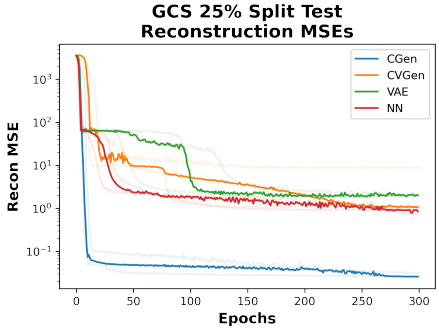




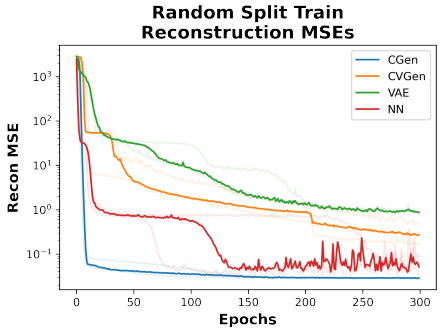


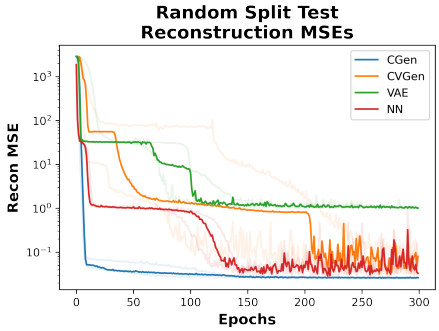


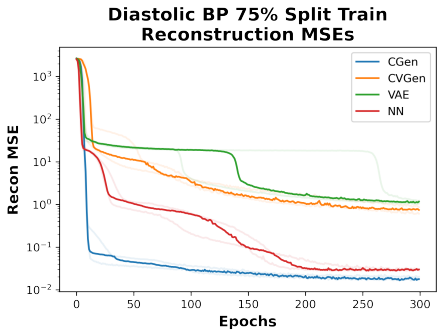


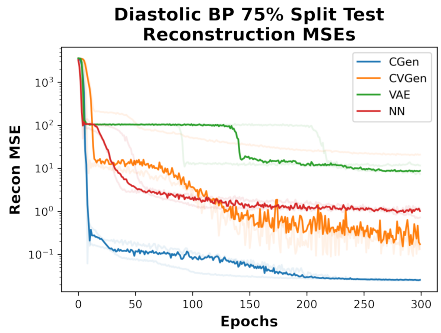


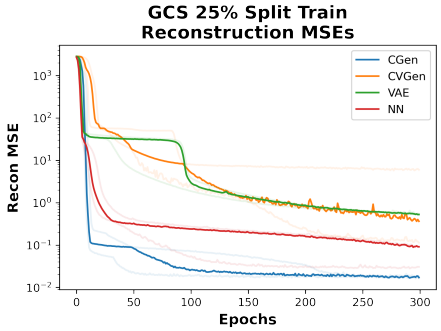
A.3.4 Medical Dataset - H2

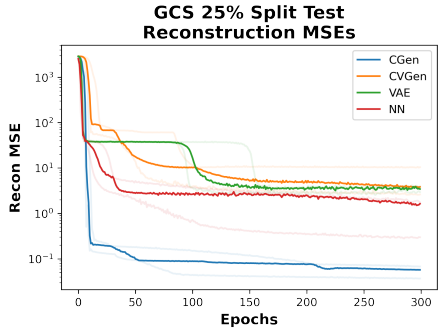






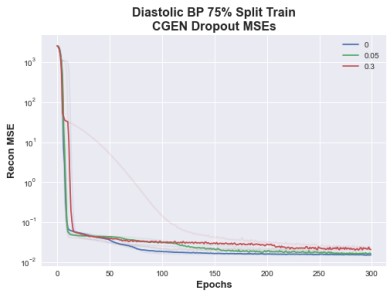


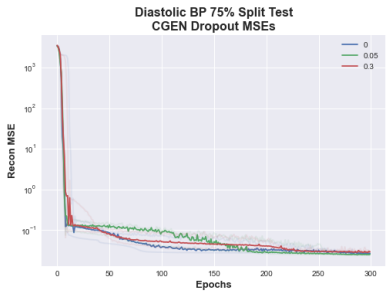




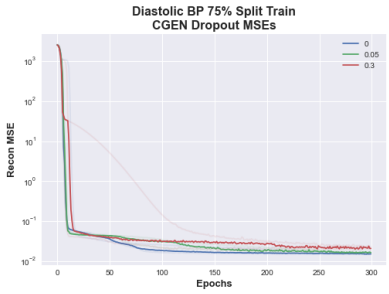
A.4 Dropout Curves Medical

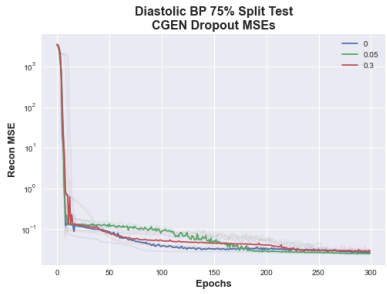
A.4.1 H1





A.4.2 H2


Diastolic BP 75% Split Train
CGEN Dropout MSEs
Recon MSE
Epochs
0
0.05
0.3


Diastolic BP 75% Split Test
CGEN Dropout MSEs
Recon MSE
Epochs
0
0.05
0.3