A Tuned Hyperparameters

For all the tasks, we tune the batch size, learning rate, contribution penalty (λ) and weight decay (μ). On the CauseMe datasets we tune the number of hidden units, whereas on the DREAM-3 dataset we tune the number of hidden layers. For computational efficiency, hyperparameters are tuned using a Tree-structured Parzen Estimator [5]. Tuned hyperparameters are provided in Tables 3-6.

Table 3. Tuned Hyperparameters of NAVAR (MLP) on the CauseMe Datasets. We Indicate the Different Variations of the "Nonlinear VAR" Dataset by the Number of Variables N and Number of Time Steps T. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

	K	**Hidden Units**	**Layers**	**Batch Size**	**Learning Rate**	$\boldsymbol{\lambda}$	$\boldsymbol{\mu}$
Tuning range	-	[8, 128]	-	[16, 256]	[5e-5, 5e-3]	[0, 0.5]	[1e-7, 0.5]
Nonlinear VAR							
N=3, T=300	5	32	1	64	0.00005	0.1344	2.903e-3
N=5, T=300	5	16	1	64	0.0001	0.1596	2.420e-3
N=10, T=300	5	128	1	64	0.0005	0.2014	8.557e-3
N=20, T=300	5	32	1	64	0.0002	0.2434	4.508e-3
Climate	2	32	1	16	0.0002	0.3924	4.322e-3
Weather	5	32	1	64	0.0001	0.0560	4.903e-3
River	5	8	1	256	0.0001	0.1708	5.092e-4

Table 4. Tuned Hyperparameters of NAVAR (LSTM) on the CauseMe Datasets. We Indicate the Different Variations of the "Nonlinear VAR" Dataset by the Number of Variables N and Number of Time Steps T. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

	K	**Hidden Units**	**Layers**	**Batch Size**	**Learning Rate**	$\boldsymbol{\lambda}$	$\boldsymbol{\mu}$
Tuning range	-	[8, 128]	-	[16, 256]	[5e-5, 5e-3]	[0, 0.5]	[1e-7, 0.5]
Nonlinear VAR							
N=3, T=300	5	16	1	64	0.0001	0.1370	8.952e-4
N=5, T=300	5	32	1	32	0.00005	0.2445	2.6756e-4
N=10, T=300	5	64	1	128	0.0001	0.0784	7.1237e-4
N=20, T=300	5	128	1	64	0.00005	0.3512	1.901e-6
Climate	2	64	1	128	0.0002	0.2334	6.231e-4
Weather	5	8	1	256	0.0005	0.0172	1.687e-3
River	5	128	1	128	0.001	0.0544	4.465e-4

Table 5. Tuned Hyperparameters of NAVAR (MLP) on the DREAM-3 Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

	K	**Hidden Units**	**Layers**	**Batch Size**	**Learning Rate**	$\boldsymbol{\lambda}$	$\boldsymbol{\mu}$
Tuning range	-	-	[1, 4]	[16, 256]	[5e-5, 5e-3]	[0, 0.5]	[1e-7, 0.5]
Ecoli1	2	10	1	128	0.0005	0.1883	1.114e-4
Ecoli2	2	10	1	32	0.001	0.2011	1.710e-4
Yeast1	2	10	2	16	0.002	0.2697	1.424e-4
Yeast2	2	10	1	256	0.0002	0.1563	2.013e-4
Yeast3	2	10	1	16	0.0002	0.1559	1.644e-4

Table 6. Tuned Hyperparameters of NAVAR (LSTM) on the DREAM-3 Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

	K	**Hidden Units**	**Layers**	**Batch Size**	**Learning Rate**	$\boldsymbol{\lambda}$	$\boldsymbol{\mu}$
Tuning range	-	-	-	-	[5e-5, 5e-3]	[0, 0.5]	[1e-7, 0.5]
Ecoli1	21	10	1	46	0.002	0.2208	1.094-5
Ecoli2	21	10	1	46	0.002	0.1958	3.233e-6
Yeast1	21	10	1	46	0.002	0.2343	5.309e-5
Yeast2	21	10	1	46	0.002	0.2189	1.987-5
Yeast3	21	10	1	46	0.002	0.2128	1.049e-5

B ROC Curves

The receiver operating characteristics (ROC) of the different methods are compared in Figure 5. Here, an ROC curve represents the trade-off between the true-positive rate (TPR) and the false-positive rate (FPR) achieved by a given method while inferring the underlying pairwise causal relationships.

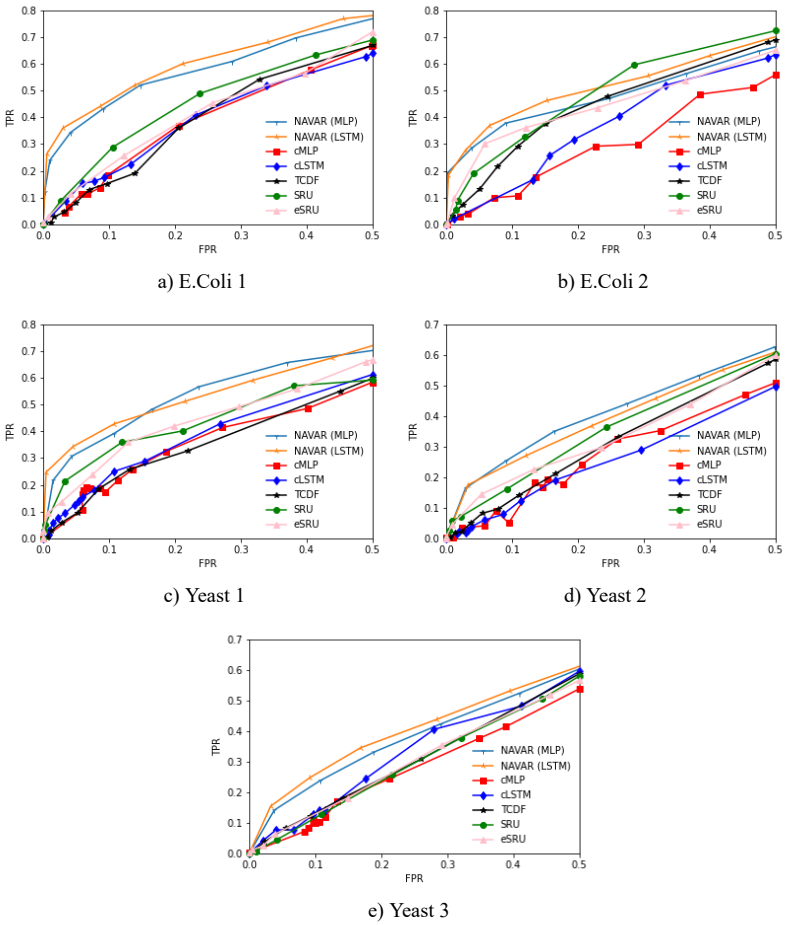

Figure 5. ROC Curves for Neural Methods on the DREAM-3 Datasets. Curves for the Methods other than NAVAR are Extracted from [15, Figure 7]