Method	Total Nodes	LIN **IN**	RFF **IN**	LIN **OUT**	RFF **OUT**
DoWhy	10	0.03 (0.03)	0.13 (0.03)	0.0 (0.0)	0.04 (0.01)
DECI	10	0.1 (0.02)	0.2 (0.03)	0.04 (0.01)	0.11 (0.02)
FiP	10	0.03 (0.01)	0.09 (0.02)	0.02 (0.0)	0.03 (0.01)
Cond-FiP	10	0.09 (0.03)	0.21 (0.03)	0.05 (0.01)	0.11 (0.01)
DoWhy	20	0.01 (0.0)	0.12 (0.03)	0.0 (0.0)	0.13 (0.02)
DECI	20	0.06 (0.01)	0.15 (0.03)	0.07 (0.03)	0.15 (0.02)
FiP	20	0.03 (0.01)	0.1 (0.03)	0.06 (0.04)	0.09 (0.02)
Cond-FiP	20	0.09 (0.02)	0.26 (0.05)	0.13 (0.02)	0.3 (0.03)
DoWhy	50	0.0 (0.0)	0.09 (0.02)	0.0 (0.0)	0.17 (0.04)
DECI	50	0.04 (0.01)	0.11 (0.02)	0.03 (0.01)	0.18 (0.04)
FiP	50	0.03 (0.01)	0.08 (0.02)	0.03 (0.01)	0.14 (0.04)
Cond-FiP	50	0.1 (0.02)	0.26 (0.04)	0.1 (0.01)	0.46 (0.06)
DoWhy	100	0.0 (0.0)	0.08 (0.02)	0.0 (0.0)	0.2 (0.05)
DECI	100	0.02 (0.01)	0.1 (0.02)	0.02 (0.01)	0.22 (0.05)
FiP	100	0.01 (0.01)	0.07 (0.02)	0.02 (0.01)	0.19 (0.05)
Cond-FiP	100	0.09 (0.02)	0.29 (0.06)	0.13 (0.02)	0.56 (0.08)

Table 30: **Results for Counterfactual Generation.** We compare Cond-FiP against the baselines for the task of generating counterfactual data from the input noise variables. Each cell reports the mean (standard error) RMSE over the multiple test datasets for each scenario. Shaded rows denote the case where the graph size is larger than the train graph sizes ($d = 20$) for Cond-FiP. We observe that Cond-FiP is worse than the baselines, and leave the improvement of Cond-FiP for counterfactual generation as future work.