FRIEDRICH-SCHILLER-UNIVERSITÄT JENA
Computer Vision Group Jena
Friedrich Schiller University Jena, Germany

TCD-Arena: Assessing Robustness of Time Series Causal Discovery Methods Against Assumption Violations

Gideon Stein, Niklas Penzel, Tristan Piater, Joachim Denzler

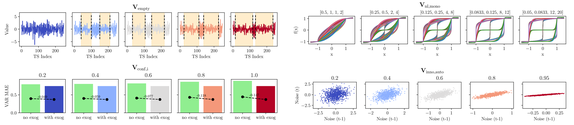
ICLR

The real world is messy.
How robust is **Causal Discovery** outside of idealized (but synthetic) data conditions?

33 disturbances ✓
Time-series ✓
Scalable in intensity ✓
Fully controllable ✓

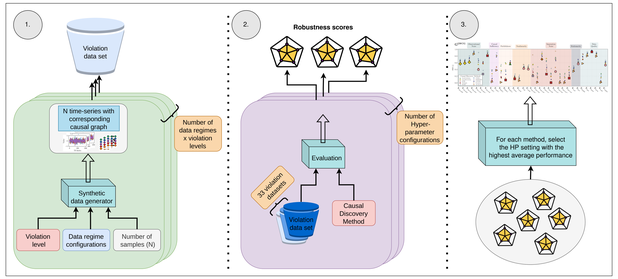
Scan to check out our benchmark!

1. Evaluating Beyond Perfect Data: Causal Discovery methods are often only validated on idealized data that perfectly satisfies their underlying assumptions. It remains largely unknown how these algorithms perform when data distributions violate these assumptions. TCD-Arena addresses this by systematically evaluating methods across 33 distinct violations, to provide a benchmark for algorithmic robustness.

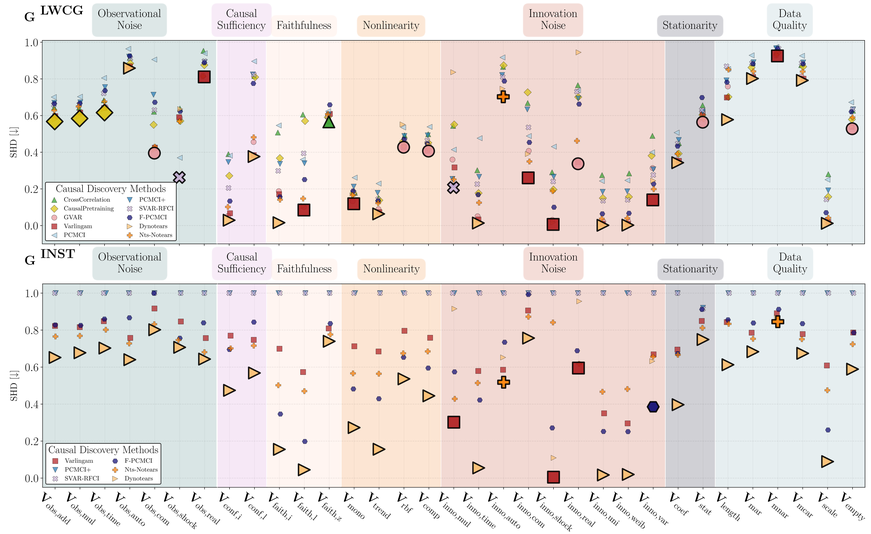


2. Beyond Binary: In the real world, assumption violations are rarely just "present" or "absent". They occur on a spectrum. To reflect this, TCD-arena features scaling the intensity of each violation, allowing to asses tolerance.

Violation	Short Description
$\mathbf{V}_{\text{obs,add}}$	Additive measurement noise.
$\mathbf{V}_{\text{obs,mul}}$	Signal dependent measurement noise.
$\mathbf{V}_{\text{obs,time}}$	Time-varying measurement noise.
$\mathbf{V}_{\text{obs,auto}}$	Autoregressive measurement noise.
$\mathbf{V}_{\text{obs,com}}$	Common source measurement noise.
$\mathbf{V}_{\text{obs,shock}}$	Spike measurement noise.
$\mathbf{V}_{\text{obs,real}}$	Measurement noise extracted from real world TS.
$\mathbf{V}_{\text{conf,inst}}$	Unseen internal common cause.
$\mathbf{V}_{\text{conf,lag}}$	Unseen external common causes.
$\mathbf{V}_{\text{faith,inst}}$	Instantaneous effects cancel out.
$\mathbf{V}_{\text{faith,lag}}$	Lagged effects cancel out.
$\mathbf{V}_{\text{faith,auto}}$	Effects become extremely small.
$\mathbf{V}_{\text{nonsv}}$	Monotonic functions.
$\mathbf{V}_{\text{trend}}$	B-spline functions with a linear trend.
$\mathbf{V}_{\text{rbf}}$	GP-RBF functions.
$\mathbf{V}_{\text{comp}}$	Composite functions.
$\mathbf{V}_{\text{imo,mul}}$	Signal dependent innovation noise.
$\mathbf{V}_{\text{imo,time}}$	Time-dependent innovation noise.
$\mathbf{V}_{\text{imo,auto}}$	Autoregressive innovation noise.
$\mathbf{V}_{\text{imo,com}}$	Common source innovation noise.
$\mathbf{V}_{\text{imo,shock}}$	Spike innovation noise.
$\mathbf{V}_{\text{imo,real}}$	Using a real-world ts as innovation noise.
$\mathbf{V}_{\text{imo,uni}}$	Uniform additive innovation noise.
$\mathbf{V}_{\text{imo,weib}}$	Weibull additive innovation noise.
$\mathbf{V}_{\text{imo,var}}$	Unequal variances in innovation noise.
$\mathbf{V}_{\text{coef}}$	Causal link strengths change over time.
$\mathbf{V}_{\text{stat}}$	Causal relationships are completely resampled over time.
$\mathbf{V}_{\text{length}}$	Reduced time series length.
$\mathbf{V}_{\text{mcar}}$	TS with values missing data completely at random
$\mathbf{V}_{\text{mar}}$	TS with values missing data at random (depending on a real-world TS)
$\mathbf{V}_{\text{mnar}}$	TS with values missing not data at random (depending on TS itself)
$\mathbf{V}_{\text{empty}}$	Temporary complete loss of causal signal.
$\mathbf{V}_{\text{scale}}$	Data standardization.



3. Our benchmarking protocol includes three steps: First, we simulate datasets for each assumption violation. Each dataset varies base parameters (e.g., the number of variables and time-series length) along the severity of the violation. Second, we evaluate every Causal Discovery algorithm on all 33 violation datasets. Third, for our main results, we report the HP configuration with the highest average performance.

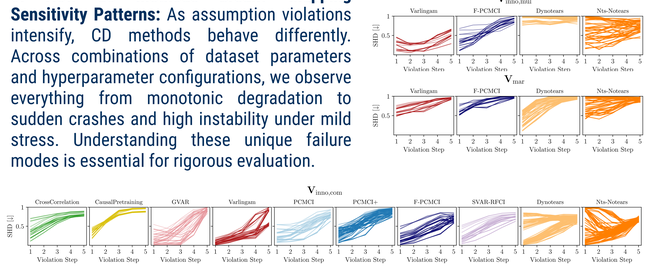


4.Assesing Robustness: We evaluate robustness across 33 violations relying on the average normalized SHD for lagged and instantaneous effects independently. While some methods exhibit better general stability, the ranking depends on the specific violation. This confirms that no single approach is universally robust to all forms of real-world data complexity.

5. From Diversity to Consensus: Since individual methods fail in complementary ways, we investigate ensemble strategies that learn a mapping from multiple method outputs to a final robust prediction. We find that learned ensembles, even as simple as a single linear layer, consistently outperform individual methods, achieving both lower average error (SHD) and higher stability across datasets.

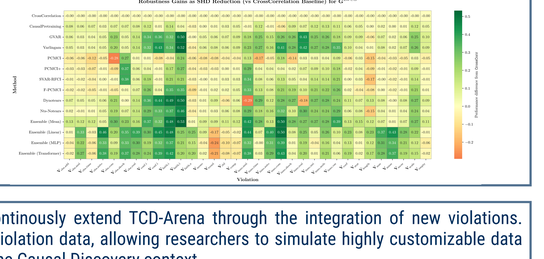
Method	G^{LWCG}	G^{INST}	G^{LSG}
CrossCorrelation	.582±.23	†	.453±.20
CausalPretraining	.530±.25	†	.440±.23
GVAR	.424±.28	†	**.330±.23**
Varlingam	**.408±.29**	.692±.20	.334±.24
PCMCI	.601±.25	†	.447±.22
PCMCI+	.539±.26	.998±.01	.405±.23
SVAR-RFCI	.517±.28	.997±.01	.407±.25
F-PCMCI	.499±.30	.657±.24	.434±.26
Dynotears	.445±.32	**.515±.28**	.365±.27
Nts-Noteras	.445±.28	.674±.13	.358±.24
$\text{Ensemble}_{\text{Avg.}}$	.387±.30	.550±.23	.300±.24
$\text{Ensemble}_{\text{Linear}}$	**.362±.20**	.527±.19	**.281±.16**
$\text{Ensemble}_{\text{MLP}}$	.445±.06	.523±.20	.343±.15
$\text{Ensemble}_{\text{Transformer}}$	.376*±.29	**.470*±.27**	.296*±.24

6. How do methods break? Mapping Sensitivity Patterns: As assumption violations intensify, CD methods behave differently. Across combinations of dataset parameters and hyperparameter configurations, we observe everything from monotonic degradation to sudden crashes and high instability under mild stress. Understanding these unique failure modes is essential for rigorous evaluation.



7. Real-world applications: We find that the robustness of ensembles is maintained for challenging real-world out-of distribution data without retraining. We attribute this to their performance on missingness and autocorrelated violations.

Method	CausalRivers (SHD)
CausalPretraining	0.851
PCMCI	0.978
Nts-Noteras	0.758
PCMCI+	0.835
Varlingam	0.774
Dynotears	0.715
F-PCMCI	0.840
SVAR-RFCI	0.857
GVAR	0.761
CrossCorrelation	0.798
Ensemble (Mean)	0.659
Ensemble (Linear)	0.666
Ensemble (MLP)	0.685
Ensemble (Transformer)	0.666



8.The Road Ahead: We are commited to continously extend TCD-Arena through the integration of new violations. Further, TCD-arena is able to generate multi-violation data, allowing researchers to simulate highly customizable data scenarios which may be useful even outside the Causal Discovery context.