TECHNISCHE UNIVERSITÄT DARMSTADT

Graph Structure Inference with BAM: Neural Dependency Processing via Bilinear Attention

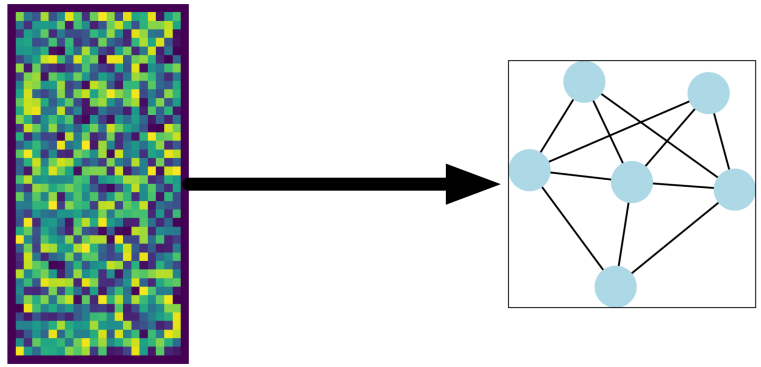
Philipp Froehlich, Heinz Koeppl

Department of Electrical Engineering and Information Technology

Technische Universität Darmstadt

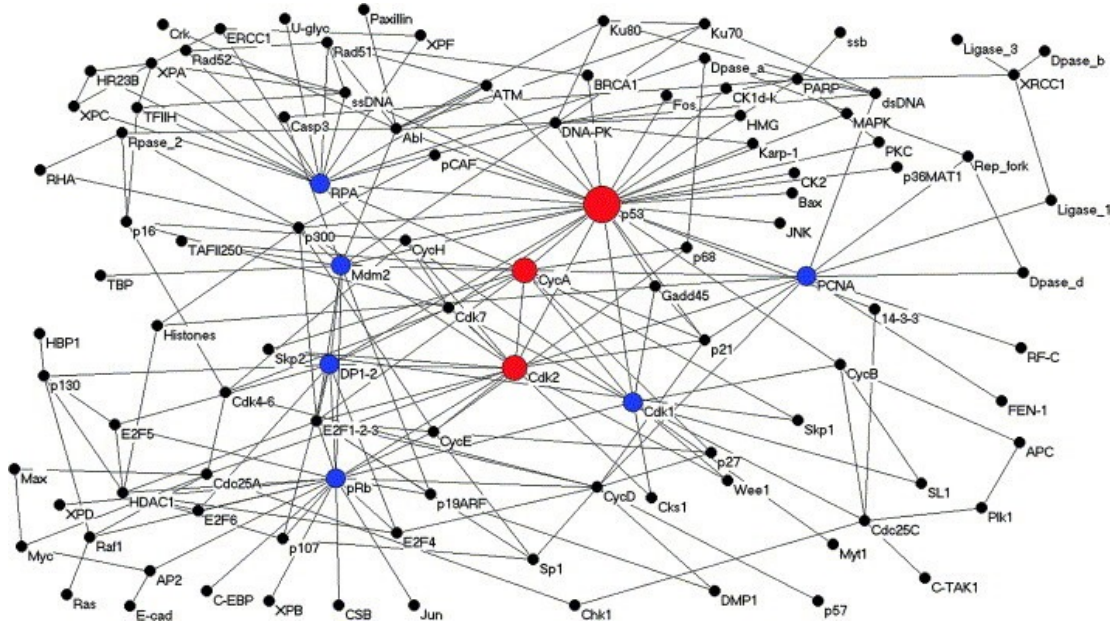
38^{th} Conference on Neural Information Processing Systems

Goal: Detecting dependencies from data



Fundamental to scientific inquiry across diverse disciplines:

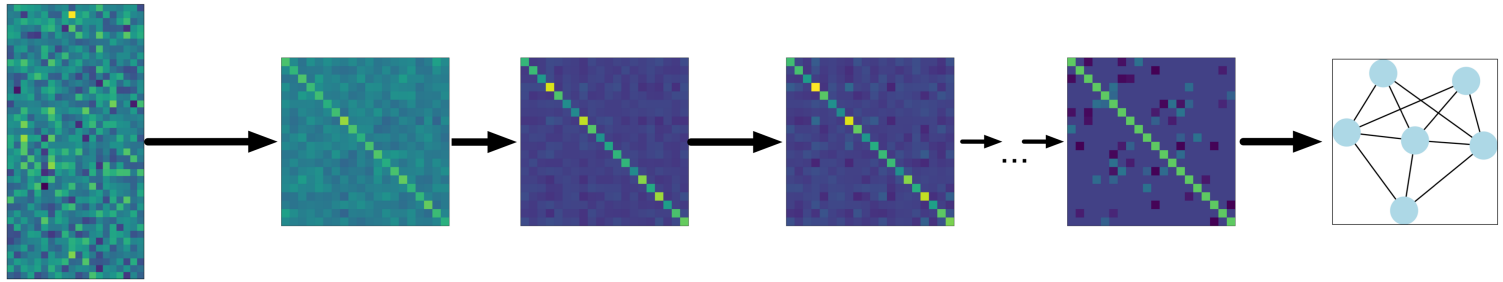
- **Biology / gene regulatory networks** (Bühlmann et. al., *Annual Review of Statistics and Its Application*, 2014).
- **Climate science** (Nowack et. al., *Nat. commun.*, 2020).
- **Cognitive science** (Gerstenberg et. al., *Psychological Review*, 2021).
- **Economics** (Barfuss et. al., *Physical Review E*, 2016).



p53 network: >200 interacting genes & proteins regulating cell survival.

(Dartnell et. al., *FEBS letters*, 2005)

Motivation: GLASSO-inspired deep learning



Data matrix $\boldsymbol{X}$

Empirical covariance

Pre-defined coordinate descent matrix operations in the SPD-space

Graph estimation $\widehat{\boldsymbol{G}}$

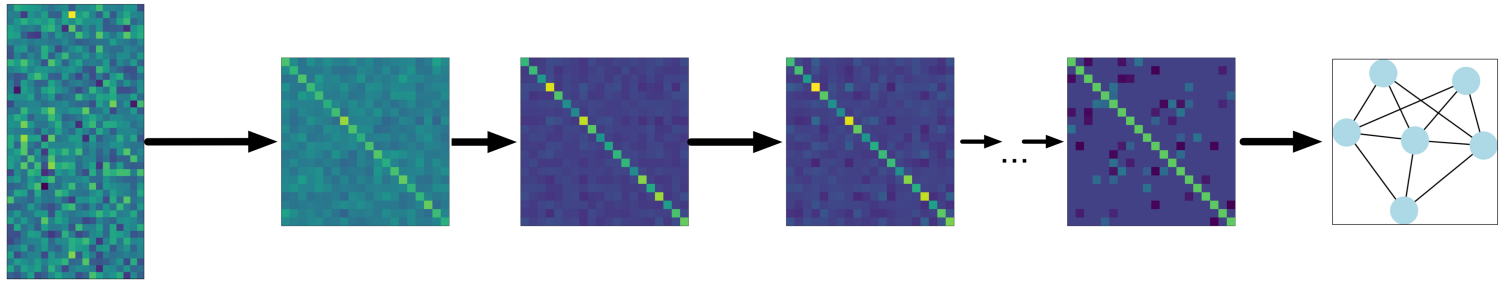
Graphical lasso algorithms:

Optimization problem within the positive semidefinite cone:

$$\widehat{\boldsymbol{G}}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\Sigma} \succ 0} F(\boldsymbol{\Sigma}, \boldsymbol{X}).$$

(Yuan & Lin, *Biometrika*, 2007), (Banerjee et. al., *JMLR*, 2008).

Motivation: GLASSO-inspired deep learning



Data matrix $\boldsymbol{X}$

Empirical covariance

Pre-defined coordinate descent matrix operations in the SPD-space

Graph estimation $\widehat{\boldsymbol{G}}$

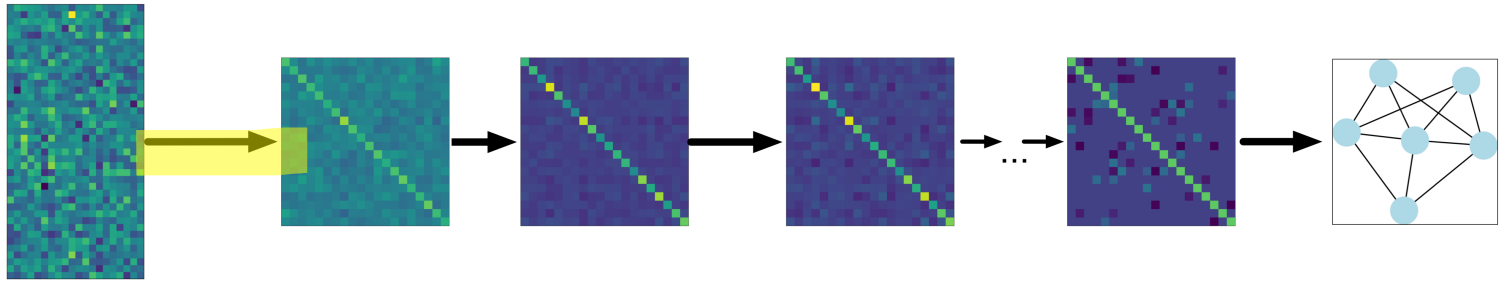
Graphical lasso algorithms:

Optimization problem within the positive semidefinite cone:

$$\widehat{\boldsymbol{G}}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\Sigma} \succ 0} F(\boldsymbol{\Sigma}, \boldsymbol{X}).$$

(Yuan & Lin, *Biometrika*, 2007), (Banerjee et. al., *JMLR*, 2008).

Motivation: GLASSO-inspired deep learning



Data matrix $\boldsymbol{X}$

Empirical covariance

Pre-defined coordinate descent matrix operations in the SPD-space

Graph estimation $\widehat{\boldsymbol{G}}$

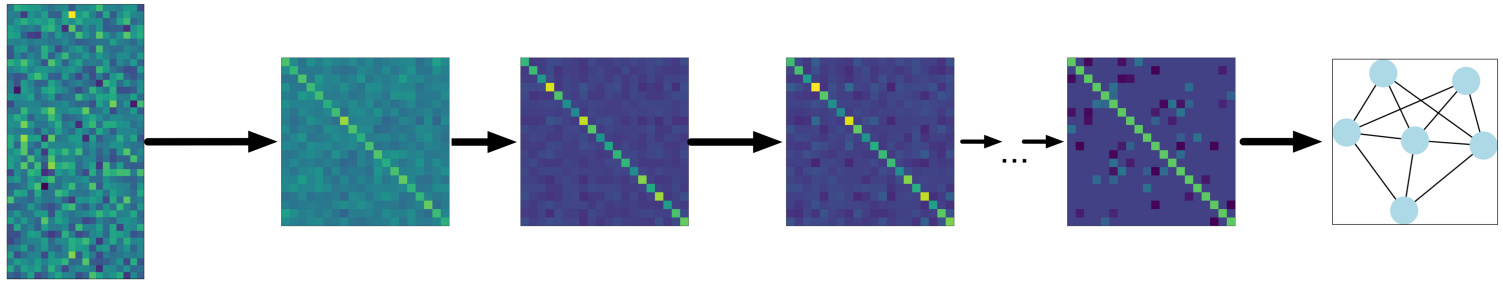
Graphical lasso algorithms:

Optimization problem within the positive semidefinite cone:

$$\widehat{\boldsymbol{G}}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\Sigma} \succ 0} F(\boldsymbol{\Sigma}, \boldsymbol{X}).$$

(Yuan & Lin, *Biometrika*, 2007), (Banerjee et. al., *JMLR*, 2008).

Motivation: GLASSO-inspired deep learning



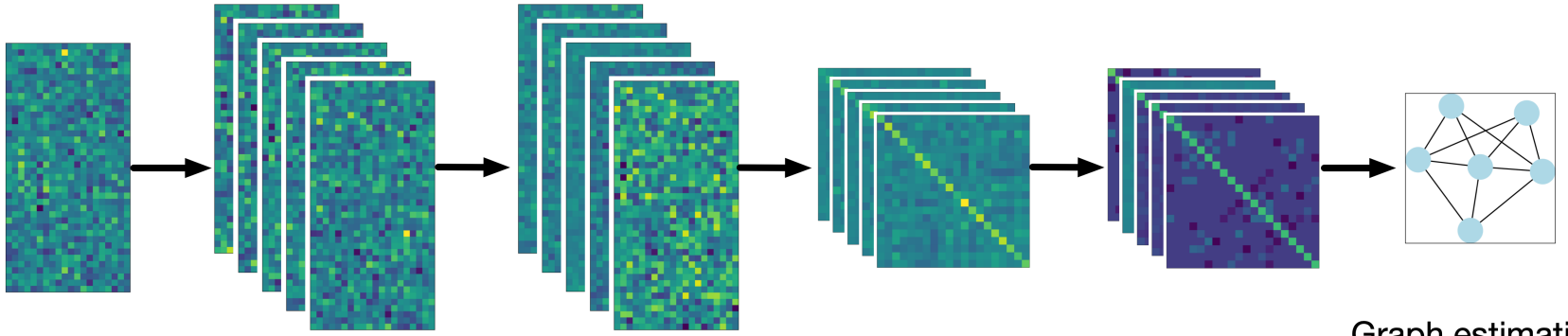
Data matrix $\boldsymbol{X}$ | Empirical covariance | Pre-defined coordinate descent matrix operations in the SPD-space | Graph estimation $\widehat{\boldsymbol{G}}$

Graphical lasso algorithms:

Optimization problem within the positive semidefinite cone:

$$\widehat{\boldsymbol{G}}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\Sigma} \succ 0} F(\boldsymbol{\Sigma}, \boldsymbol{X}).$$

(Yuan & Lin, *Biometrika*, 2007), (Banerjee et. al., *JMLR*, 2008).



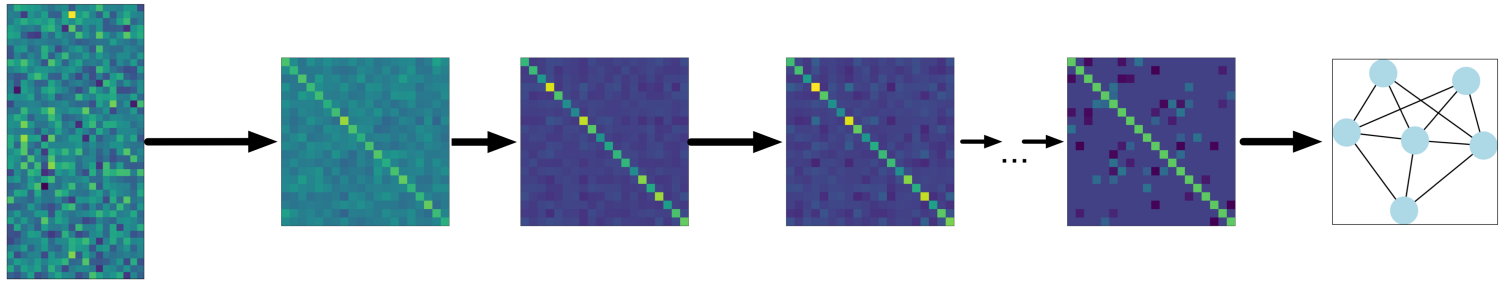
Data matrix $\boldsymbol{X}$ | Learned Euklidean representation | Learned representation in SPD space | Graph estimation $\widehat{\boldsymbol{G}}$

Our approach:

Learn shape-invariant matrix-operations via deep learning to approximate the solution mapping

$$f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \widehat{\boldsymbol{G}}.$$

Motivation: GLASSO-inspired deep learning



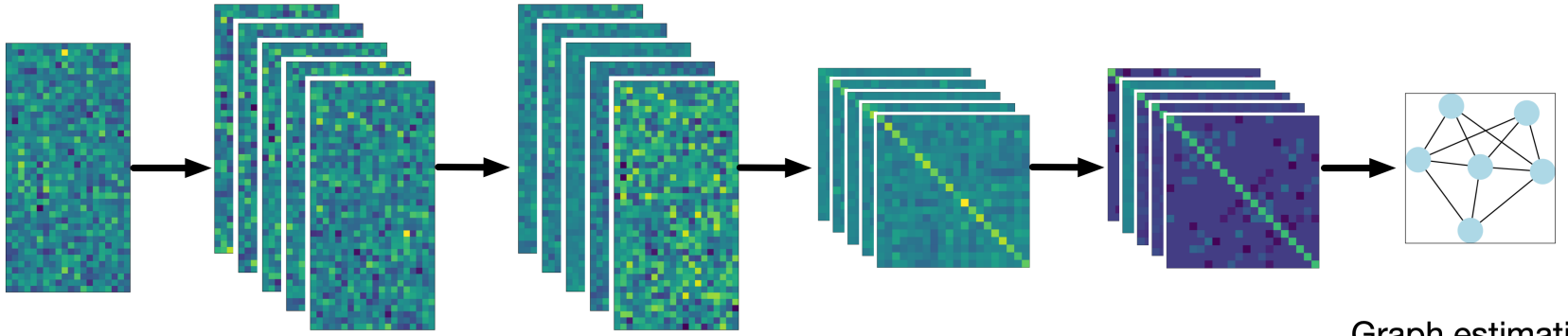
Data matrix $\boldsymbol{X}$ | Empirical covariance | Pre-defined coordinate descent matrix operations in the SPD-space | Graph estimation $\widehat{\boldsymbol{G}}$

Graphical lasso algorithms:

Optimization problem within the positive semidefinite cone:

$$\widehat{\boldsymbol{G}}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\Sigma} \succ 0} F(\boldsymbol{\Sigma}, \boldsymbol{X}).$$

(Yuan & Lin, *Biometrika*, 2007), (Banerjee et. al., *JMLR*, 2008).



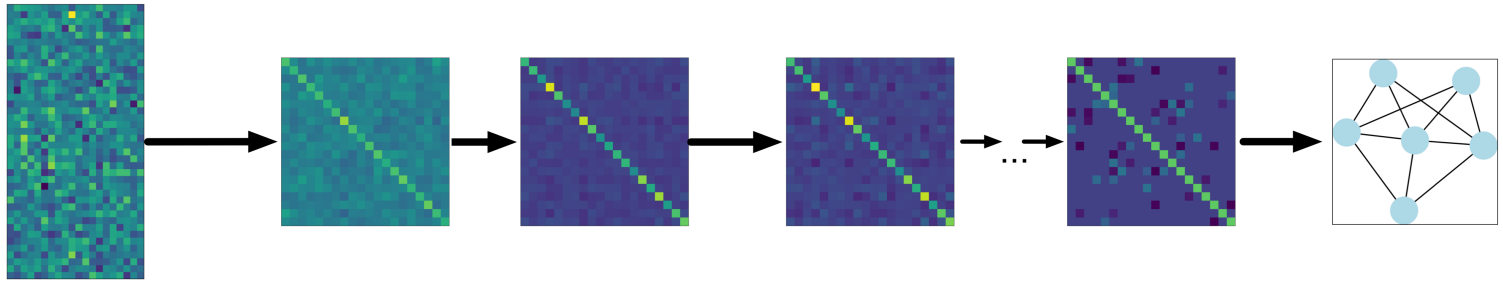
Data matrix $\boldsymbol{X}$ | Learned Euklidean representation | Learned representation in SPD space | Graph estimation $\widehat{\boldsymbol{G}}$

Our approach:

Learn shape-invariant matrix-operations via deep learning to approximate the solution mapping

$$f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \widehat{\boldsymbol{G}}.$$

Motivation: GLASSO-inspired deep learning



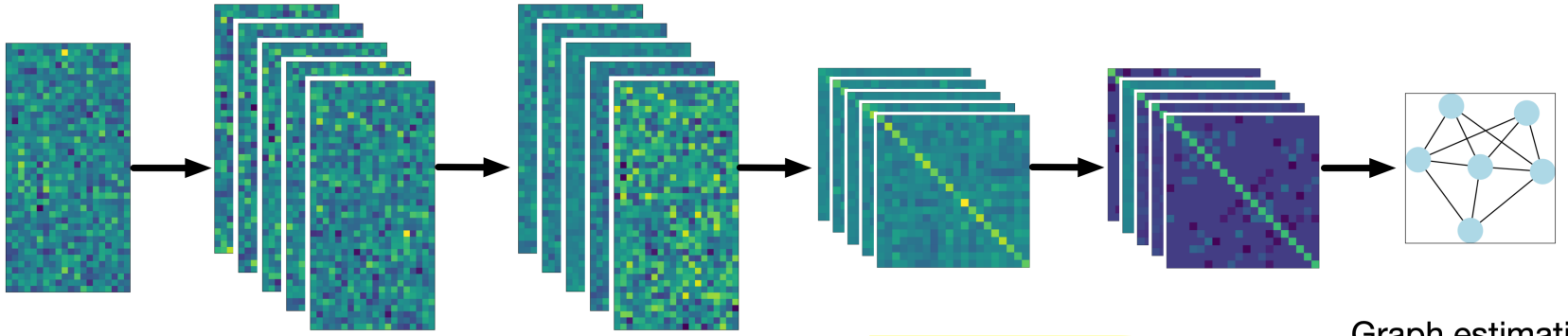
Data matrix $\boldsymbol{X}$ | Empirical covariance | Pre-defined coordinate descent matrix operations in the SPD-space | Graph estimation $\widehat{\boldsymbol{G}}$

Graphical lasso algorithms:

Optimization problem within the positive semidefinite cone:

$$\widehat{\boldsymbol{G}}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\Sigma} \succ 0} F(\boldsymbol{\Sigma}, \boldsymbol{X}).$$

(Yuan & Lin, *Biometrika*, 2007), (Banerjee et. al., *JMLR*, 2008).



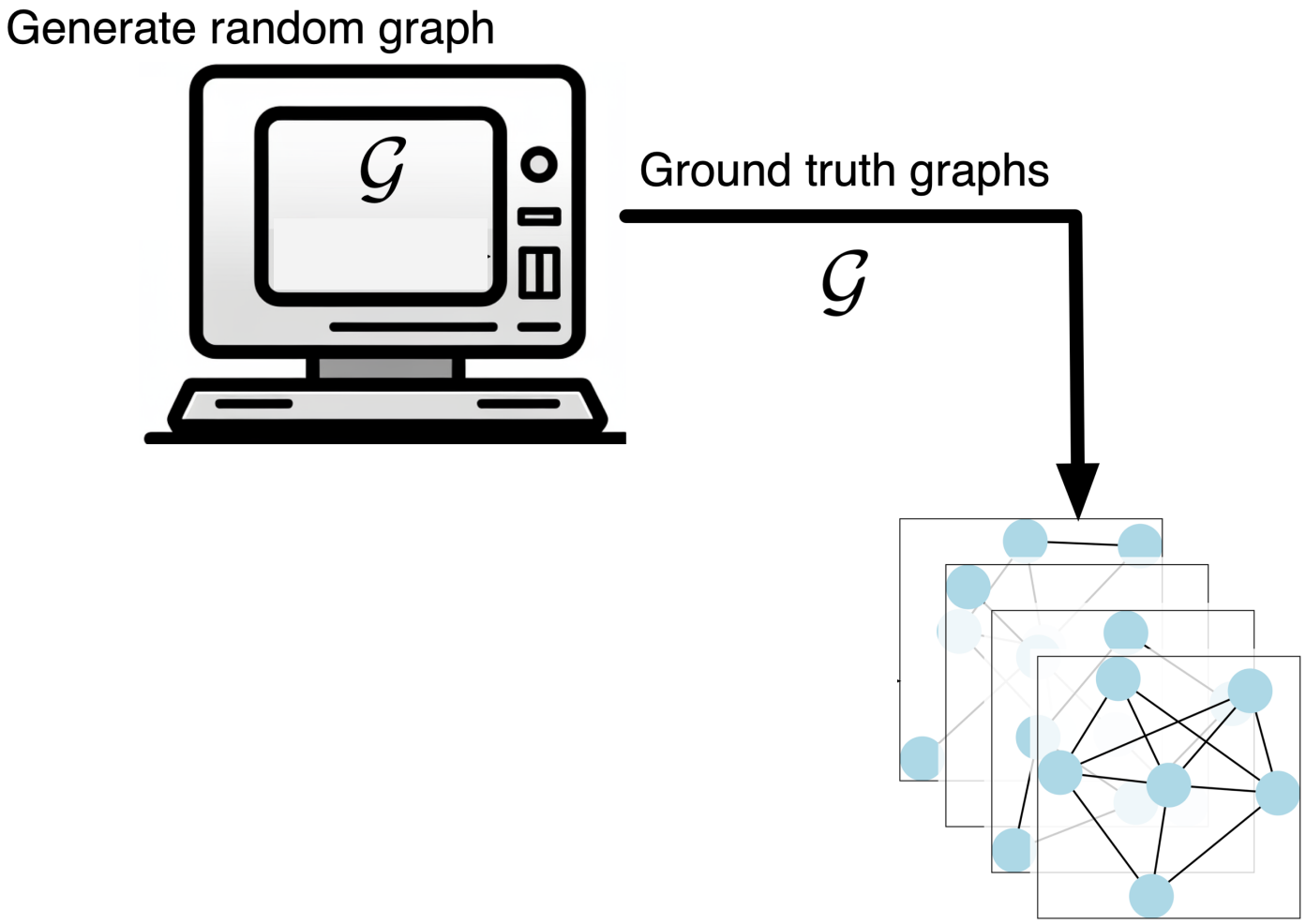
Data matrix $\boldsymbol{X}$ | Learned Euklidean representation | Learned representation in SPD space | Graph estimation $\widehat{\boldsymbol{G}}$

Our approach:

Learn shape-invariant matrix-operations via deep learning to approximate the solution mapping

$$f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \widehat{\boldsymbol{G}}.$$

Supervised causal learning approach

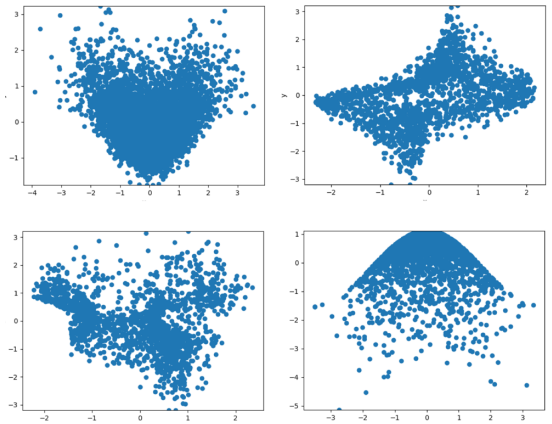


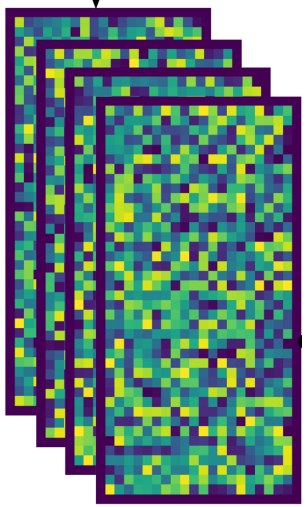
Supervised causal learning approach

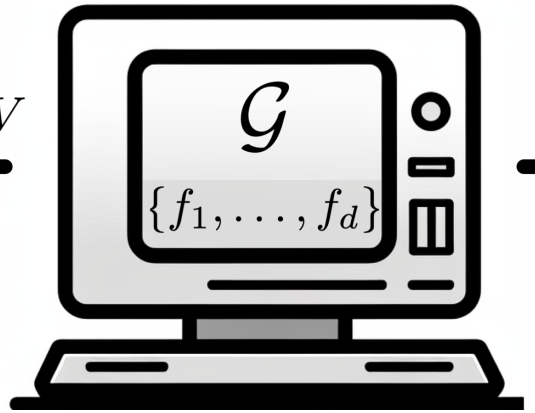
Generate random graph and coupling

SEM

$$X_v = f_v(X_{\mathrm{pa}_\mathcal{G}(v)}, \epsilon_v) \quad \forall v \in V$$

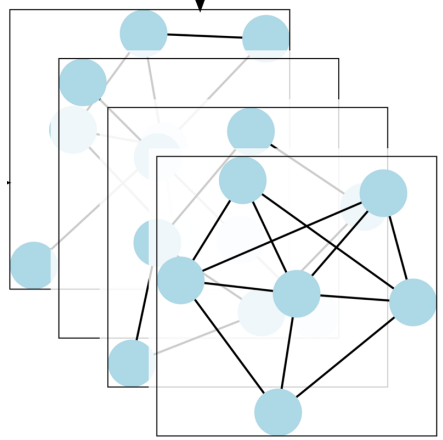






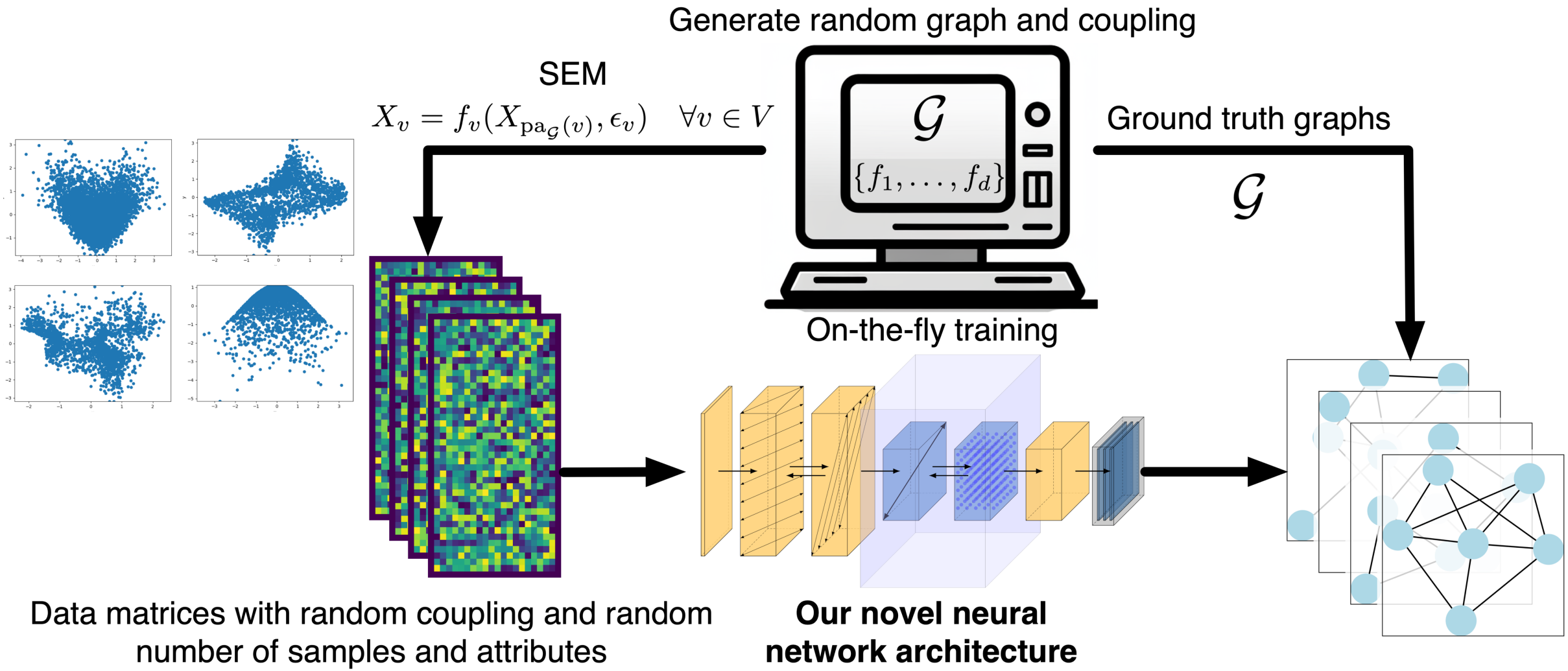
Ground truth graphs

$\mathcal{G}$

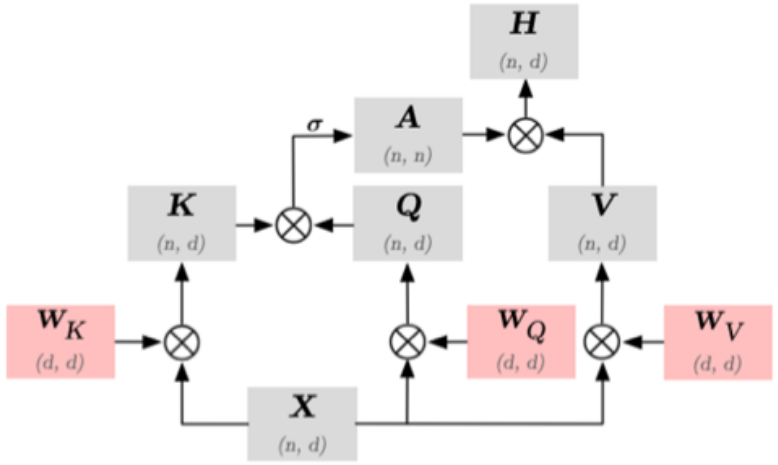


Data matrices with random coupling and random number of samples and attributes

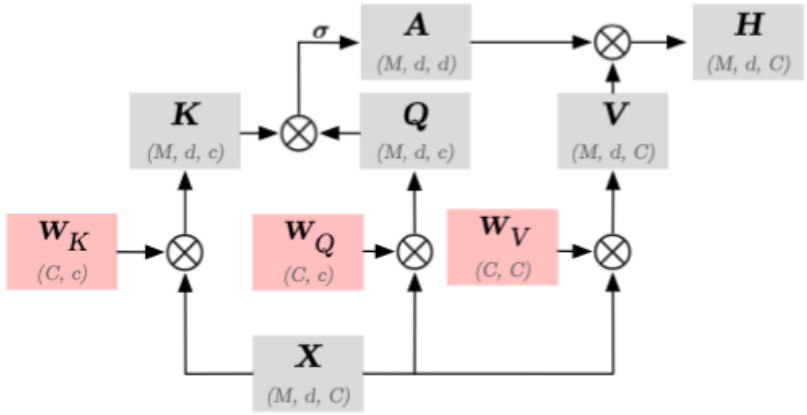
Supervised causal learning approach



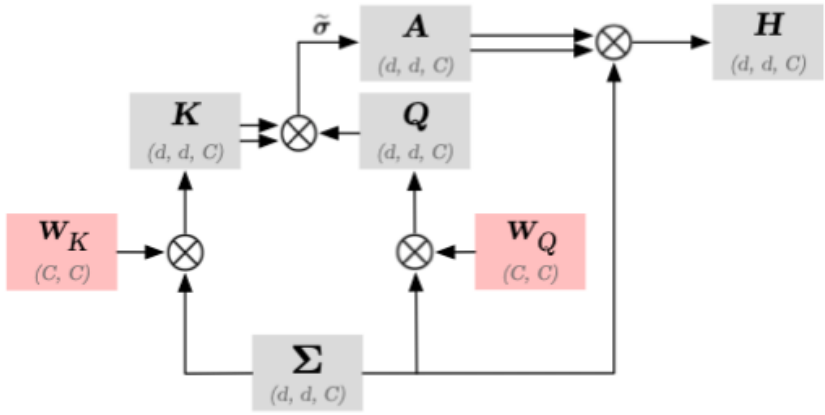
Novel bilinear attention mechanism



Traditional attention for sequences (Bahdanau et. al., *In ICLR*, 2015).



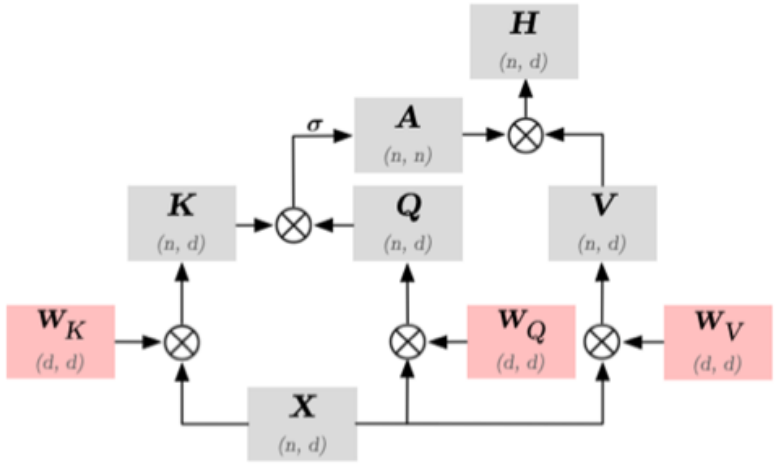
Observational attention on 3D tensors (Kossen et. al., *In NeurIPS,* 2021).



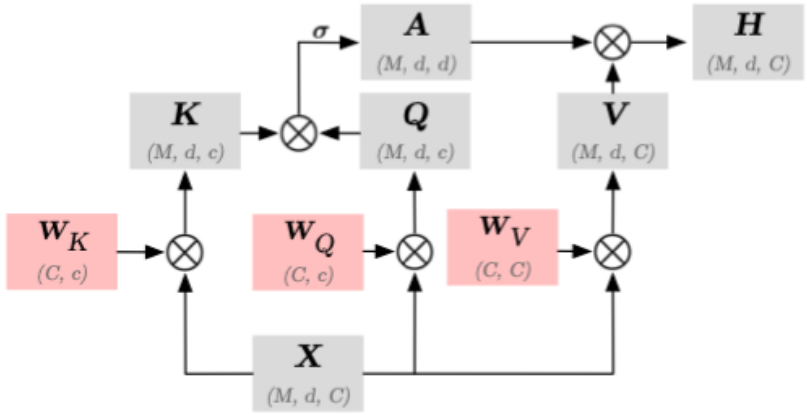
Bilinear attention, performing bilinear operations instead of matrix multiplications.

	Attention	Bilinear Attention (BAM)
Multiplication	Matrix multiplication	Bilinear tensor operation
Softmax	$\exp(\boldsymbol{S})\boldsymbol{\Lambda}(\boldsymbol{S})$	$\sqrt{\boldsymbol{\Lambda}(\boldsymbol{S})}\exp(\boldsymbol{S})\sqrt{\boldsymbol{\Lambda}(\boldsymbol{S})}$

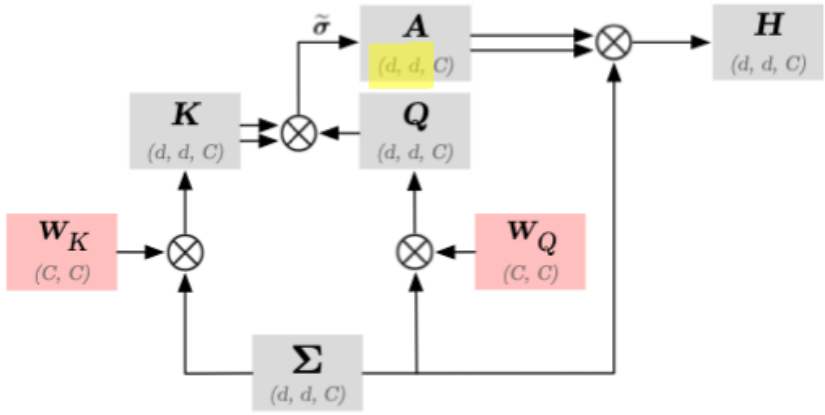
Novel bilinear attention mechanism



Traditional attention for sequences (Bahdanau et. al., *In ICLR*, 2015).



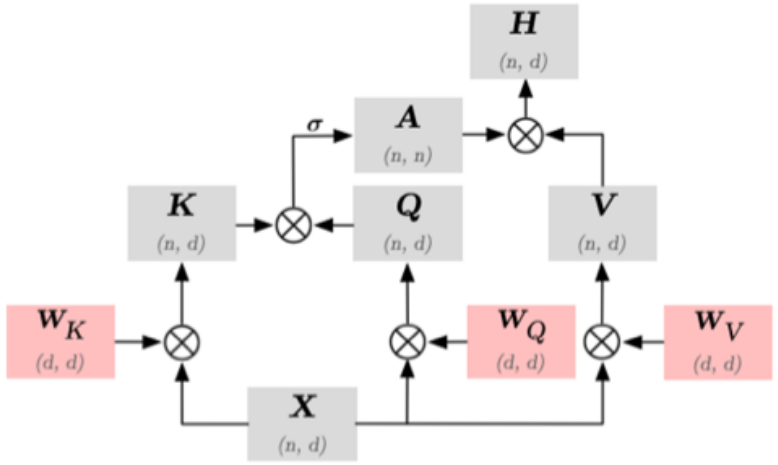
Observational attention on 3D tensors (Kossen et. al., *In NeurIPS,* 2021).



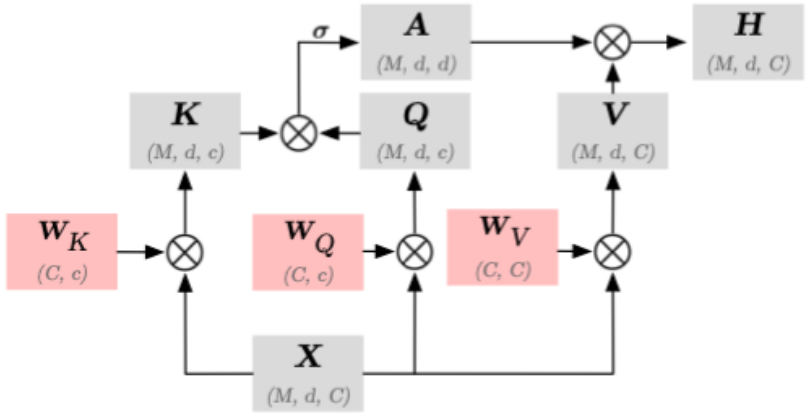
Bilinear attention, performing bilinear operations instead of matrix multiplications.

	Attention	Bilinear Attention (BAM)
Multiplication	Matrix multiplication	Bilinear tensor operation
Softmax	$\exp(\boldsymbol{S})\boldsymbol{\Lambda}(\boldsymbol{S})$	$\sqrt{\boldsymbol{\Lambda}(\boldsymbol{S})}\exp(\boldsymbol{S})\sqrt{\boldsymbol{\Lambda}(\boldsymbol{S})}$

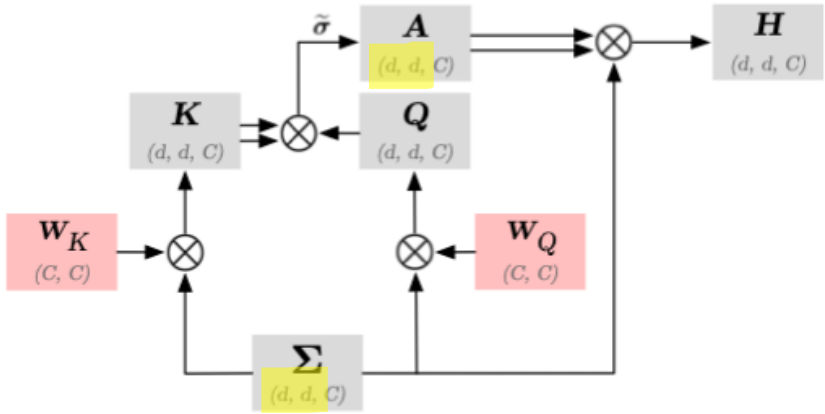
Novel bilinear attention mechanism



Traditional attention for sequences (Bahdanau et. al., *In ICLR*, 2015).



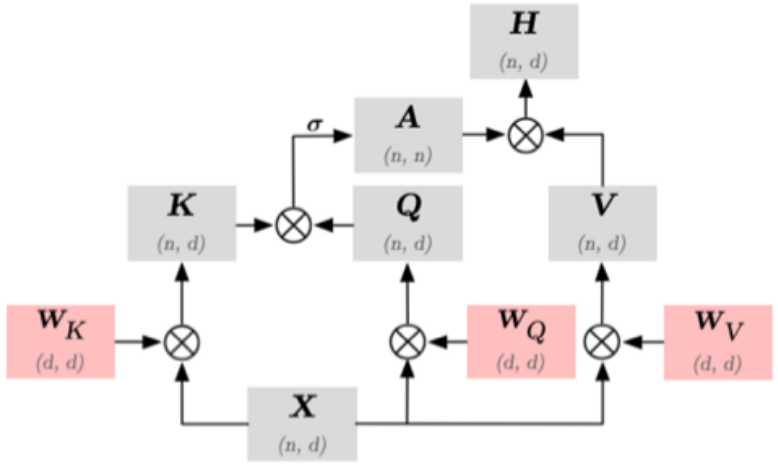
Observational attention on 3D tensors (Kossen et. al., *In NeurIPS,* 2021).



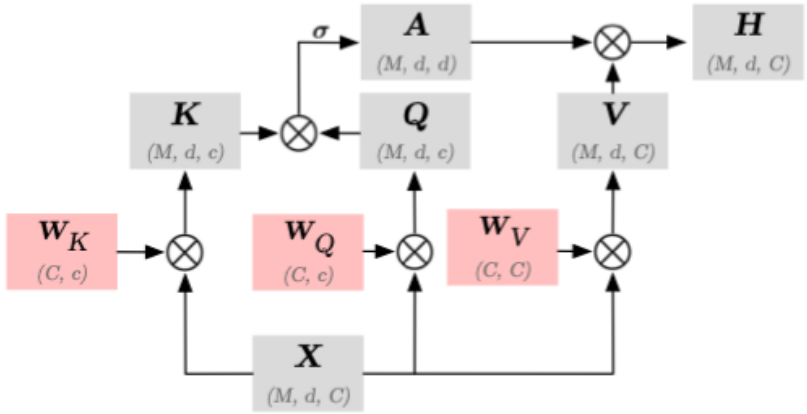
Bilinear attention, performing bilinear operations instead of matrix multiplications.

	Attention	Bilinear Attention (BAM)
Multiplication	Matrix multiplication	Bilinear tensor operation
Softmax	$\exp(\boldsymbol{S})\boldsymbol{\Lambda}(\boldsymbol{S})$	$\sqrt{\boldsymbol{\Lambda}(\boldsymbol{S})}\exp(\boldsymbol{S})\sqrt{\boldsymbol{\Lambda}(\boldsymbol{S})}$

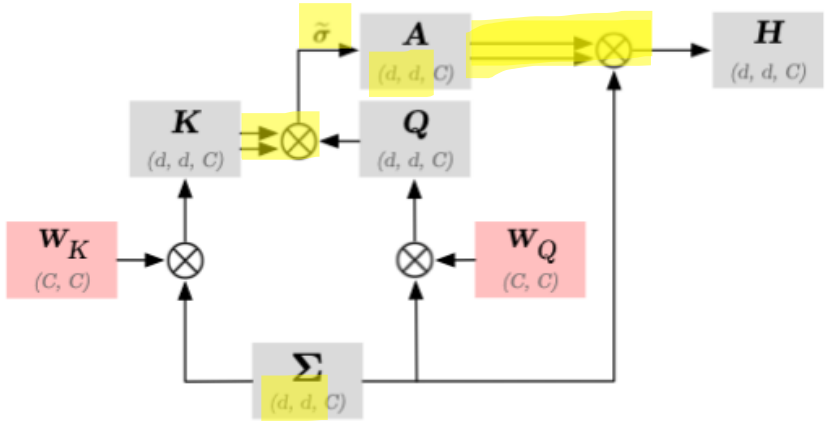
Novel bilinear attention mechanism



Traditional attention for sequences (Bahdanau et. al., *In ICLR*, 2015).



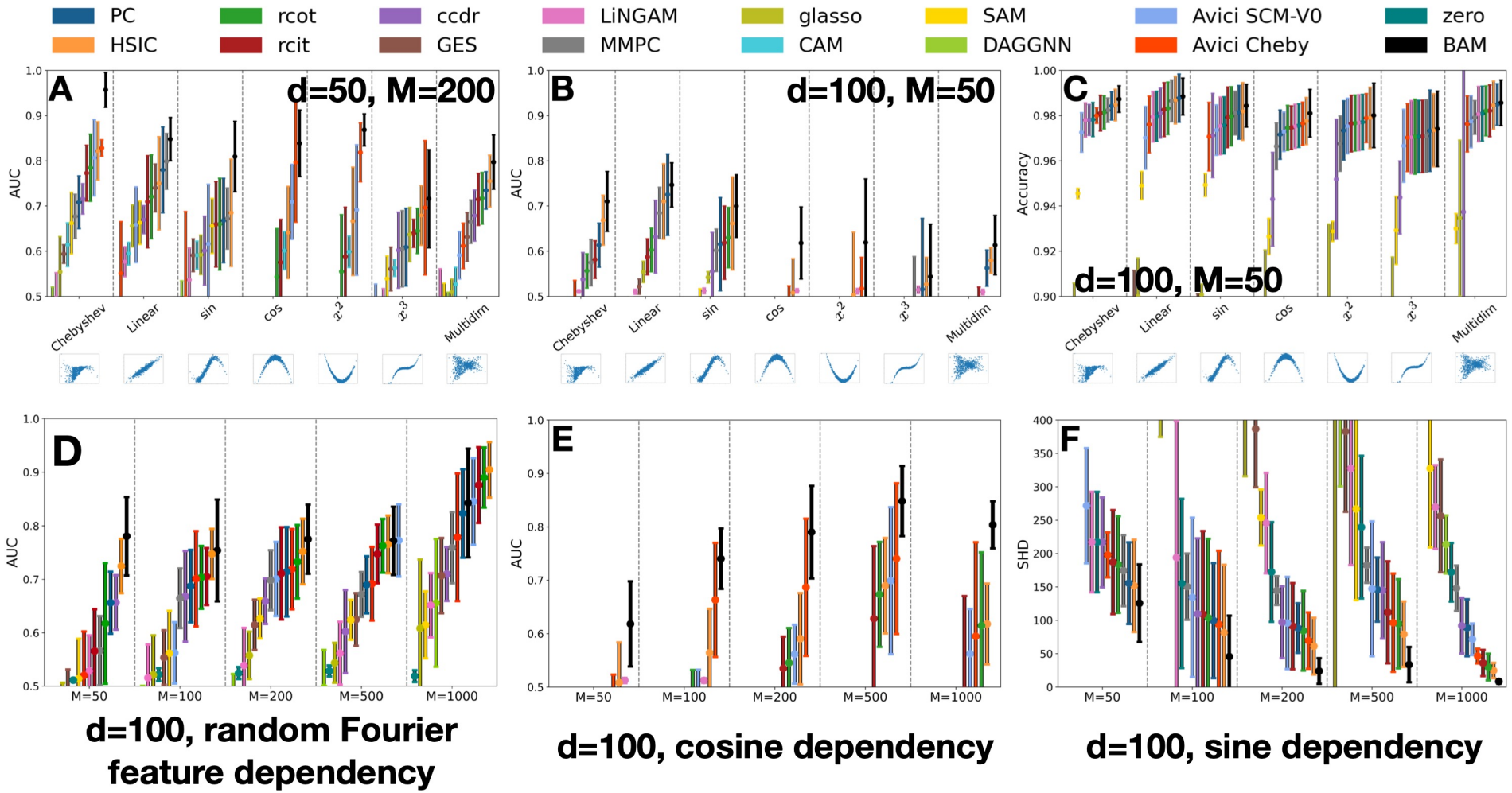
Observational attention on 3D tensors (Kossen et. al., *In NeurIPS,* 2021).



Bilinear attention, performing bilinear operations instead of matrix multiplications.

	Attention	Bilinear Attention (BAM)
Multiplication	Matrix multiplication	Bilinear tensor operation
Softmax	$\exp(\boldsymbol{S})\boldsymbol{\Lambda}(\boldsymbol{S})$	$\sqrt{\boldsymbol{\Lambda}(\boldsymbol{S})}\exp(\boldsymbol{S})\sqrt{\boldsymbol{\Lambda}(\boldsymbol{S})}$

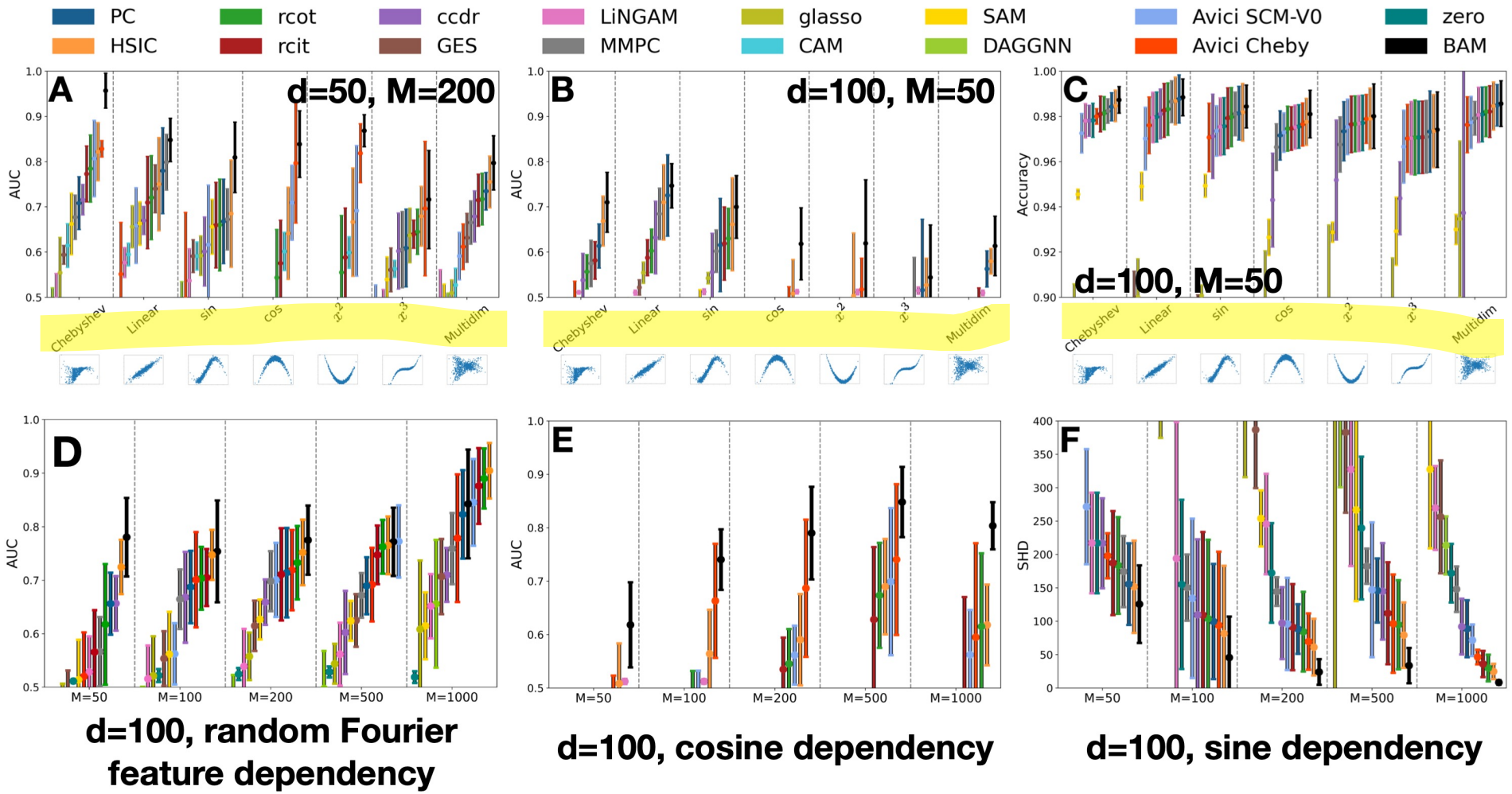
Results for undirected graph inference



Generalizes beyond training: From Chebyshev to linear, sine, and complex dependencies.

Detects dependencies despite zero-gradients at data center and oscillations.

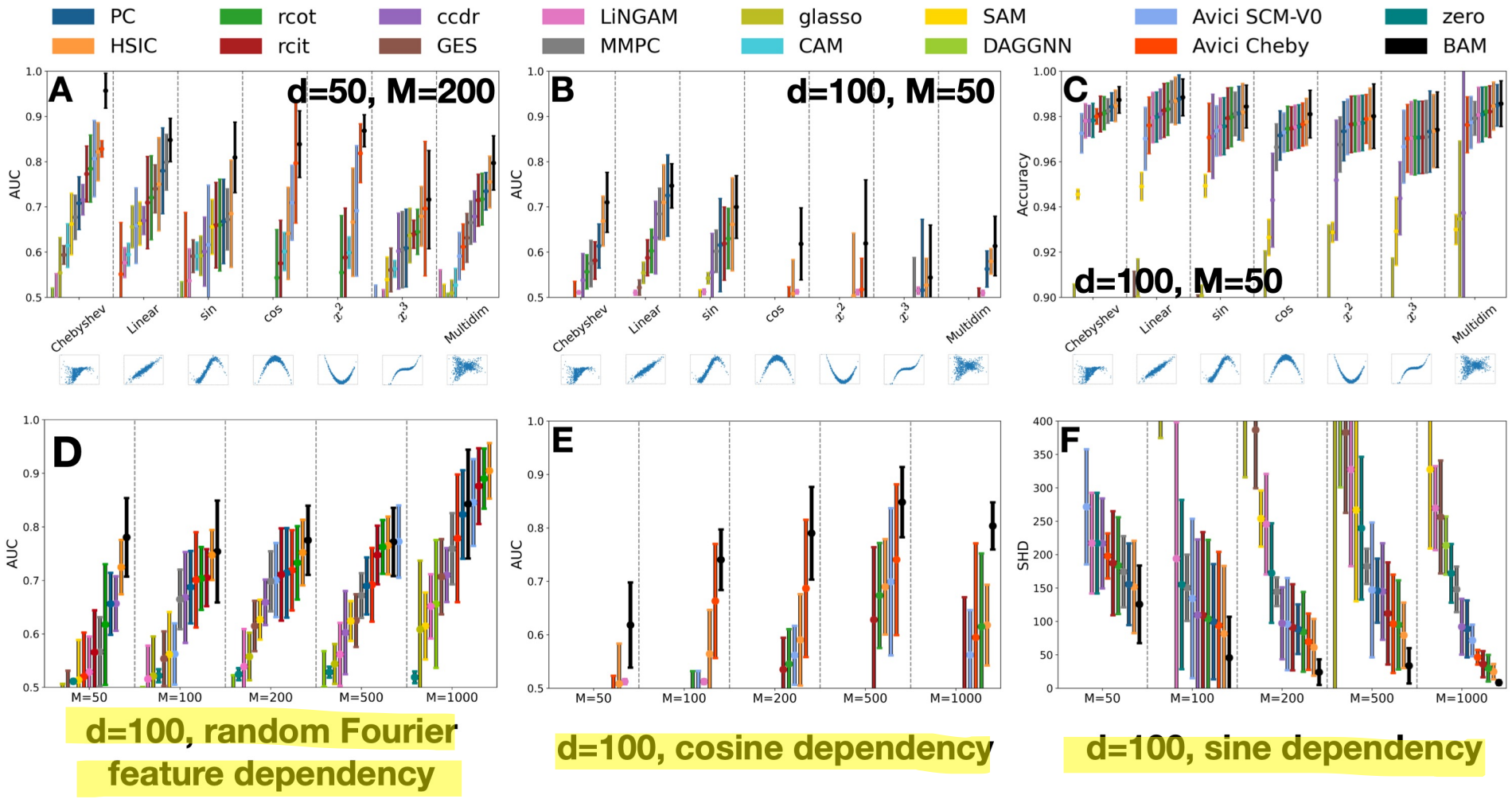
Results for undirected graph inference



Generalizes beyond training: From Chebyshev to linear, sine, and complex dependencies.

Detects dependencies despite zero-gradients at data center and oscillations.

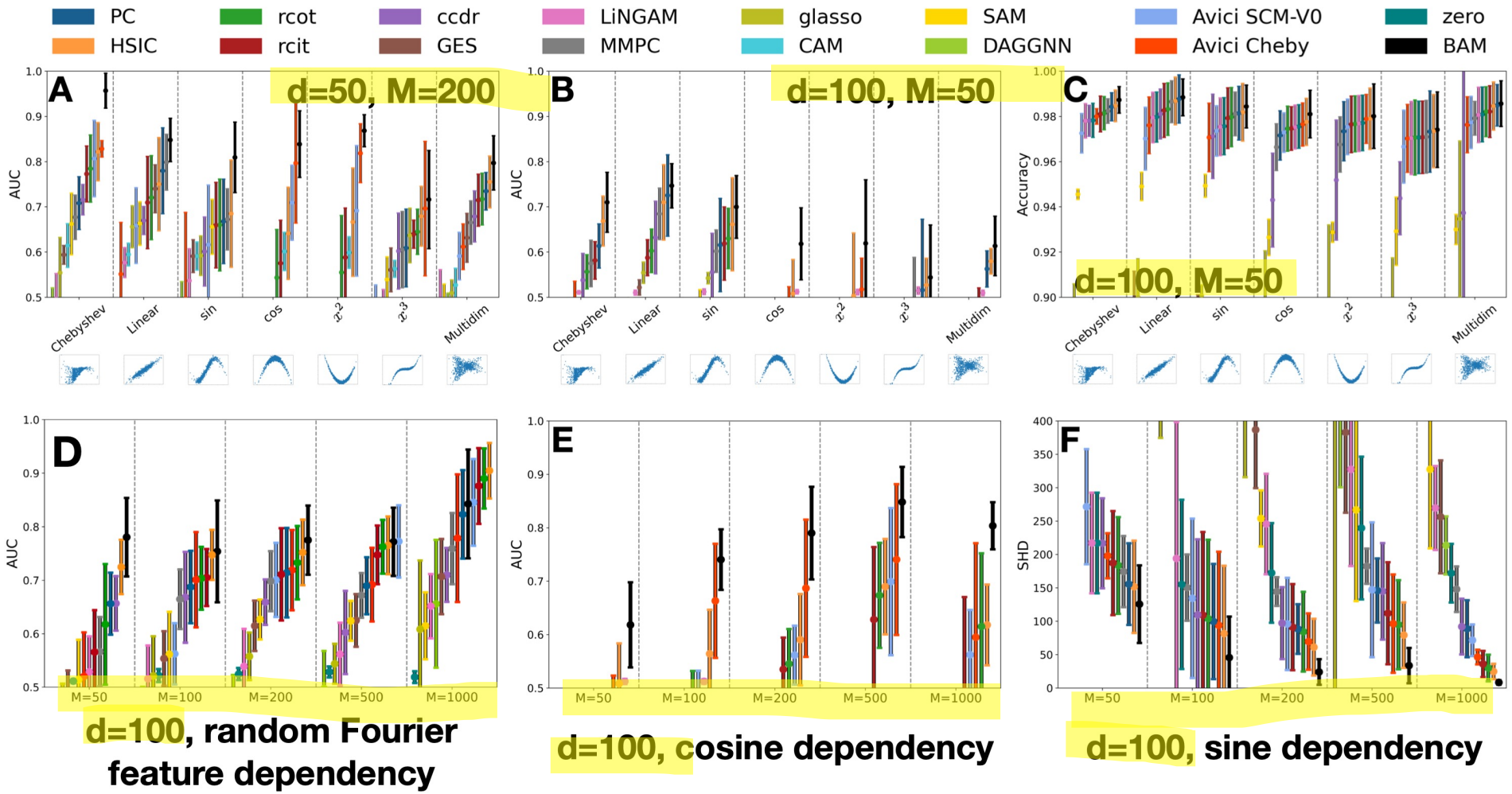
Results for undirected graph inference



Generalizes beyond training: From Chebyshev to linear, sine, and complex dependencies.

Detects dependencies despite zero-gradients at data center and oscillations.

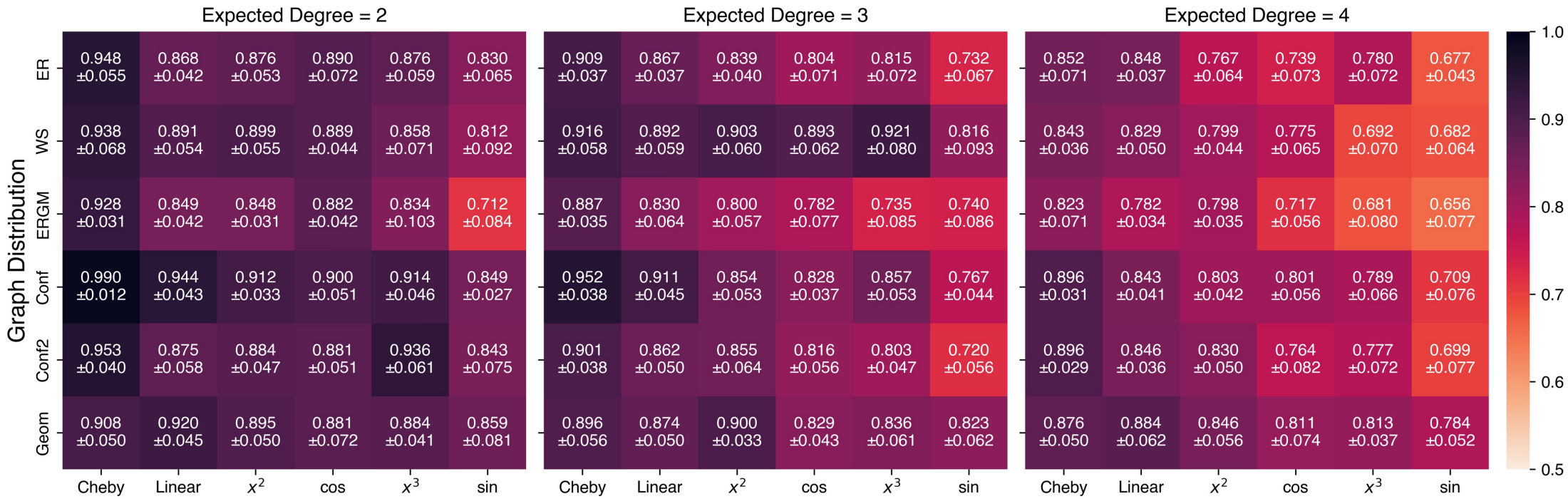
Results for undirected graph inference



Generalizes beyond training: From Chebyshev to linear, sine, and complex dependencies.

Detects dependencies despite zero-gradients at data center and oscillations.

Results for undirected graph inference



Robust to simultaneous changes in graph distribution and coupling mechanism.

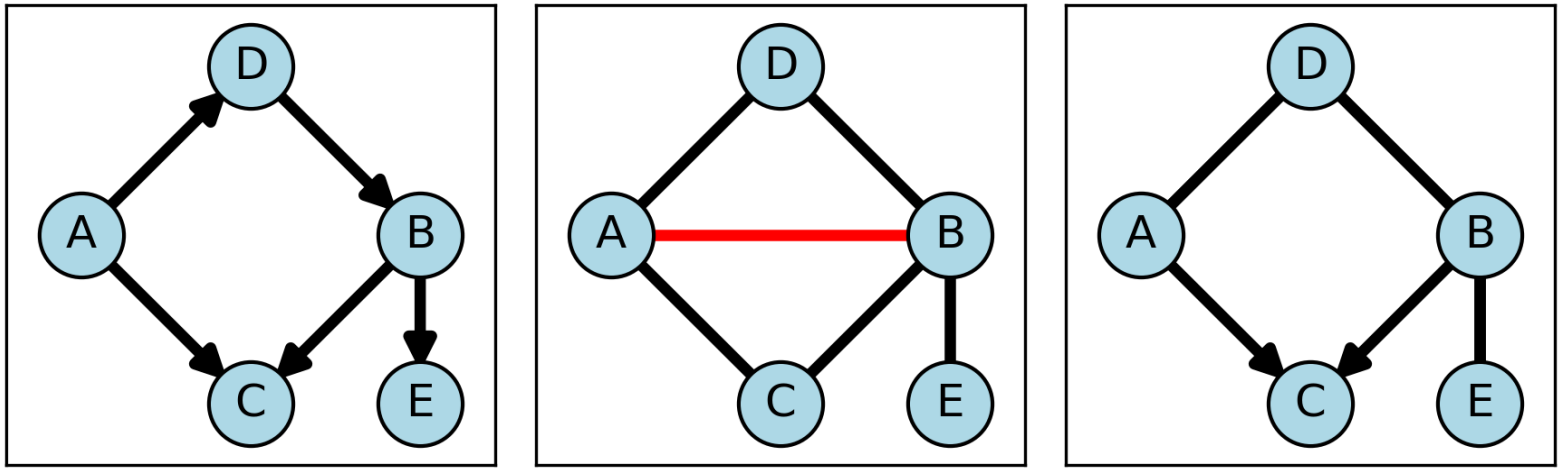
Inferring causal directions

Novel two-step approach to infer causal directions:

First step: Differentiate between:

- **Skeleton edges:** Undirected edges in the DAG.
- **Moralized edges:** Not present in the DAG but emerge due to conditional dependencies among nodes sharing a common child without a connecting edge between the parents.
- **No edge:** Conditionally independent variables.

Second step: Infer v-structures by testing all triplets of possible parents and possible common child nodes.



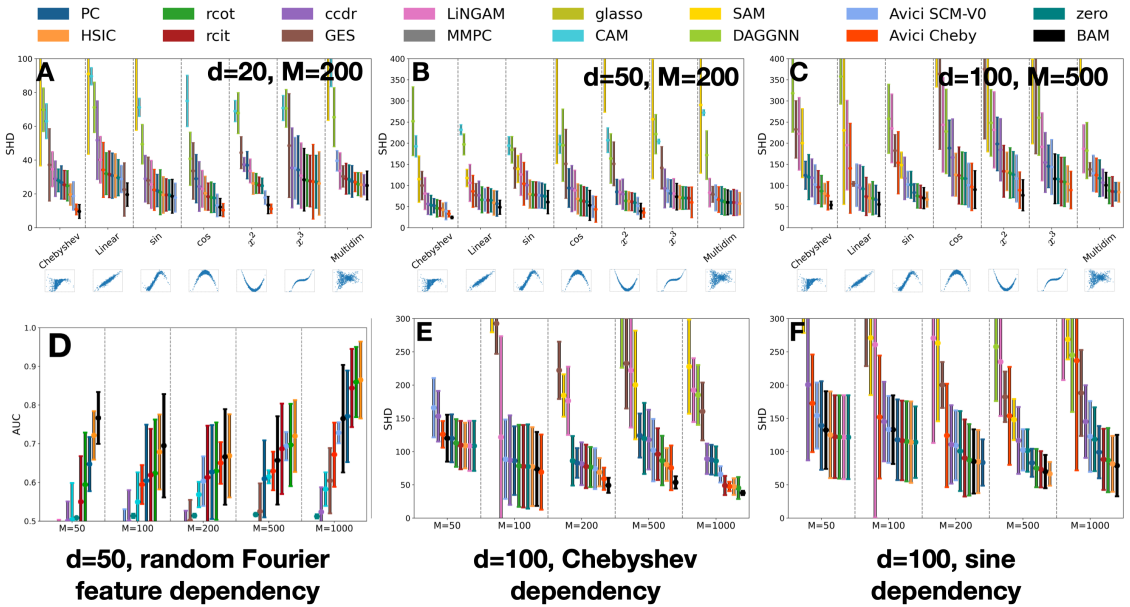
Results for CPDAG inference

Dependency Function	Model	$d = 10$	$d = 20$	$d = 50$	$d = 100$
Chebyshev	BAM	0.79 ± 0.09	0.85 ± 0.05	0.78 ± 0.05	0.76 ± 0.03
	AM	0.73 ± 0.03	0.69 ± 0.02	0.71 ± 0.02	0.70 ± 0.01
Linear	BAM	0.71 ± 0.06	0.72 ± 0.02	0.69 ± 0.06	0.70 ± 0.02
	AM	0.67 ± 0.03	0.63 ± 0.06	0.63 ± 0.04	0.68 ± 0.03
sin	BAM	0.69 ± 0.16	0.68 ± 0.09	0.62 ± 0.07	0.66 ± 0.04
	AM	0.63 ± 0.08	0.62 ± 0.07	0.60 ± 0.06	0.64 ± 0.04
cos	BAM	0.71 ± 0.04	0.74 ± 0.05	0.65 ± 0.11	0.59 ± 0.04
	AM	0.67 ± 0.07	0.64 ± 0.05	0.62 ± 0.07	0.59 ± 0.05
x^2	BAM	0.77 ± 0.05	0.77 ± 0.09	0.71 ± 0.08	0.63 ± 0.11
	AM	0.68 ± 0.06	0.64 ± 0.05	0.64 ± 0.05	0.61 ± 0.05
x^3	BAM	0.56 ± 0.03	0.55 ± 0.17	0.54 ± 0.09	0.59 ± 0.06
	AM	0.58 ± 0.05	0.50 ± 0.10	0.56 ± 0.09	0.57 ± 0.05

Predicting immoralities in first step improved predictions for CPDAG estimation compared to testing all possible immoralities.

Results for CPDAG inference

Dependency Function	Model	$d = 10$	$d = 20$	$d = 50$	$d = 100$
Chebyshev	BAM	0.79 ± 0.09	0.85 ± 0.05	0.78 ± 0.05	0.76 ± 0.03
	AM	0.73 ± 0.03	0.69 ± 0.02	0.71 ± 0.02	0.70 ± 0.01
Linear	BAM	0.71 ± 0.06	0.72 ± 0.02	0.69 ± 0.06	0.70 ± 0.02
	AM	0.67 ± 0.03	0.63 ± 0.06	0.63 ± 0.04	0.68 ± 0.03
sin	BAM	0.69 ± 0.16	0.68 ± 0.09	0.62 ± 0.07	0.66 ± 0.04
	AM	0.63 ± 0.08	0.62 ± 0.07	0.60 ± 0.06	0.64 ± 0.04
cos	BAM	0.71 ± 0.04	0.74 ± 0.05	0.65 ± 0.11	0.59 ± 0.04
	AM	0.67 ± 0.07	0.64 ± 0.05	0.62 ± 0.07	0.59 ± 0.05
x^2	BAM	0.77 ± 0.05	0.77 ± 0.09	0.71 ± 0.08	0.63 ± 0.11
	AM	0.68 ± 0.06	0.64 ± 0.05	0.64 ± 0.05	0.61 ± 0.05
x^3	BAM	0.56 ± 0.03	0.55 ± 0.17	0.54 ± 0.09	0.59 ± 0.06
	AM	0.58 ± 0.05	0.50 ± 0.10	0.56 ± 0.09	0.57 ± 0.05



d=50, random Fourier feature dependency

d=100, Chebyshev dependency

d=100, sine dependency

Predicting immoralities in first step improved predictions for CPDAG estimation compared to testing all possible immoralities.

Maintains robust performance in CPDAG estimation across dimensions and dependencies.

Conservative orientation: high accuracy on directed edges with few false positives.

Conclusion

Key contributions:

- Novel bilinear attention mechanism for processing dependencies.
- Two-step approach for robust CPDAG inference.
- State-of-the-art performance across diverse settings:
  - Generalizes beyond training to diverse dependency types.
  - Detects dependencies despite zero-gradients at data mean.
  - Robust to simultaneous changes in graph distribution and coupling.
  - Conservative yet accurate causal orientation.