Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Causal Transformer: Scaling Gradient-Based Causal Discovery to 500 Variables

Shuaidong Gao

gaoshuaidong@stu.cqifs.edu.cn

Chongqing Institute of Foreign Studies

Article

Keywords: Causal discovery, Self-attention, Transformer, NOTEARS, DAG learning, Head specialization

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10467674/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Causal Transformer: Scaling Gradient-Based Causal Discovery to 500 Variables

Shuaidong Gao[1*]

[1]Chongqing Institute of Foreign Studies, Qijiang Campus, Chongqing 401420, China

gaoshuaidong@stu.cqifs.edu.cn.

Abstract

Causal discovery faces a dimensionality ladder that no single architecture spans. Cluster-aware methods provide robust inference at $\boldsymbol{d < 200}$ but collapse with their underlying NOTEARS engine at higher dimensions. The regime $\boldsymbol{d = 200}$–$\boldsymbol{500}$—where most real transcriptomic datasets reside—has no gradient-based solution. We propose Causal Transformer (CT), a self-attention architecture that treats variables as tokens and learns causal edges from multi-head attention, retaining NOTEARS' acyclicity constraint but constructing $\boldsymbol{W}$ through attention rather than independent optimization. An 80-configuration sweep reveals a sharp viability boundary: CT discovers near-zero edges at $\boldsymbol{d \leq 100}$ but activates at $\boldsymbol{d = 200}$, producing 1,028 edges versus NOTEARS' 0. On TCGA-BRCA at $\boldsymbol{d = 250}$–$\boldsymbol{350}$ (5 seeds), CT remains operational (62,500–122,500 edges) where NOTEARS collapses. Multi-omics validation exposes a modality-specific limit: zero edges on methylation, motivating consensus approaches. An MLP baseline matches raw output but lacks CT's emergent head specialization and pairwise gradient signals. On 10 TCGA cancer types at $\boldsymbol{d = 200}$, CT recovers 198 edges on average versus NOTEARS' 0, with 75% of top-hub genes validated as ClinGen Tier-1 cancer drivers. CT extends DAG-based causal discovery to the frontier of $\boldsymbol{d \approx 500}$, operating reliably at $\boldsymbol{d = 250}$–$\boldsymbol{350}$ and revealing its $\boldsymbol{O(d^2)}$ memory ceiling near $\boldsymbol{d = 500}$, all on consumer hardware.

Keywords: Causal discovery, Self-attention, Transformer, NOTEARS, DAG learning, Head specialization

1 Introduction

Causal discovery—learning directed acyclic graphs (DAGs) from observational data—is a fundamental problem in machine learning and systems biology Zheng et al. (2018); Sachs et al. (2005); Peters et al. (2017). The field has been transformed by the NOTEARS framework Zheng et al. (2018), which reformulated DAG learning as continuous optimization with a smooth acyclicity constraint $h(W) = \mathrm{tr}(e^{W \odot W}) - d$, enabling gradient-based structure learning. However, NOTEARS-based approaches inherit a cubic complexity barrier from this matrix exponential, limiting practical application to $d \leq 150$ on consumer hardware.

Two complementary solutions have recently emerged at the low-dimensional end of the spectrum. Cluster-aware gating methods embed sample clustering into the NOTEARS optimization loop, and their sparse variants resolve the hard-versus-soft clustering debate through a sample-size phase transition, achieving robust causal discovery on 33 TCGA cancer types with positive gain over NOTEARS. Both methods, however, build on NOTEARS' augmented Lagrangian optimizer and inherit its collapse at $d \geq 150$.

This gap—$d = 200$ to $d \approx 500$—is precisely the dimensionality regime of many real-world datasets. Of the 33 TCGA cancer types analyzed by the sparse cluster-gating variant, the median expression dimensionality after MAD-based gene filtering is 198 genes, making $d = 200$ the most common configuration for transcriptomic causal discovery. Neuroimaging connectomes (200–400 brain regions; Sporns et al. (2005)) and epidemiological studies (200–500 variables; Heckerman (1997)) occupy the same range, yet no gradient-based method operates here.

Our contribution. We introduce Causal Transformer (CT), a self-attention architecture that treats variables as tokens and learns causal edge weights from multi-head attention. CT retains NOTEARS' acyclicity penalty $h(W)$ but constructs W through attention rather than independent parameter optimization—each W_{ij} is informed by

interactions across all variable pairs and all attention heads, providing richer gradient signals than independently learned parameters. Our contributions are:

1. **Architecture design.** We formulate CT as a variable-as-token encoder that maps an $n \times d$ data matrix to a $d \times d$ causal adjacency matrix via multi-head self-attention, retaining the NOTEARS acyclicity penalty for DAG enforcement.
2. **Emergent head specialization.** We demonstrate that CT's attention heads spontaneously specialize into positive-edge and negative-edge experts (inter-head Jaccard similarity 0.17), a phenomenon previously observed in NLP Transformers Clark et al. (2019) but never reported in causal discovery.
3. **Systematic viability characterization.** An 80-configuration parameter sweep ($d \in \{30, 50, 100, 200\}$, $n \in \{200, 500\}$ with $n = 1000$ at $d = 200$, 10 attention strategies) reveals a sharp viability boundary: CT requires $d \geq 200$ to activate, where it produces 538–1,028 edges versus NOTEARS' 0.
4. **Operational range identification.** CT achieves peak performance at $d = 250$ (1,605 edges), declines after $d = 350$, and collapses at $d = 500$ due to $O(d^2)$ memory limits.
5. **Real-data validation.** On 10 TCGA cancer types at $d = 200$, CT recovers 198 causal edges on average, with 75% of top-hub genes confirmed as ClinGen Tier-1 cancer drivers. At $d = 250$–350 on TCGA-BRCA (5 seeds each), CT remains operational (62,500–122,500 edges) where NOTEARS collapses. On BRCA multi-omics (expression, CNV, methylation) at $d = 200$, CT identifies a modality-specific failure mode (zero edges on methylation).
6. **MLP baseline.** A shared-weight MLP alternative at $d = 200$ produces comparable raw edge counts ($\sim$39,400) but CT's attention mechanism provides emergent head specialization and structurally richer gradient signals, distinguishing CT from a simple non-linear regressor.

2 Related Work

Causal Transformer sits at the intersection of four research areas: differentiable causal discovery, Transformer architectures for structured prediction, attention-based causal discovery, and multi-head specialization.

2.1 Differentiable Causal Discovery

The NOTEARS framework Zheng et al. (2018) opened the door to gradient-based DAG learning by replacing combinatorial search with a smooth acyclicity penalty $h(W) = \text{tr}(e^{W \odot W}) - d$. Subsequent work extended NOTEARS along three axes: (i) nonlinear structural equation models, including DAG-GNN Yu et al. (2019) and Gran-DAG Lachapelle et al. (2020); (ii) improved optimization objectives, with GOLEM Ng et al. (2020) and DAGMA Bello et al. (2022) replacing the acyclicity penalty entirely; and (iii) extensions to interventional data Brouillard et al. (2020) and time-series Pamfil et al. (2020).

All NOTEARS variants share the $O(d^3)$ matrix exponential bottleneck, and our experiments confirm this is a structural rather than implementation-specific failure. At $d = 200$, GOLEM Ng et al. (2020) and DAGMA Bello et al. (2022) were evaluated under their published hyperparameters (10 seeds each). Both returned dense adjacency matrices (GOLEM: 39,566 $\pm$ 92 edges; DAGMA: 38,153 $\pm$ 41 edges), failing to produce sparse DAGs despite achieving acyclicity ($h(W) \to 0$). Increasing the L1 sparsity penalty 25-fold did not rescue sparsity, confirming that the failure mode at $d = 200$ is fundamental to the optimization landscape rather than a hyperparameter artifact. CT succeeds at this scale because multi-head attention provides a structurally different gradient pathway (Section 4.4). Cluster-aware approaches address the low-dimensional regime ($d < 200$) but leave the middle range ($d = 200$–500) unaddressed. Unlike DAGMA, which replaces the constraint, CT retains $h(W)$ but constructs W through

multi-head attention rather than independent parameter optimization—improving edge quality without altering the asymptotic complexity.

2.2 Transformers for Structured Prediction

The Transformer architecture Vaswani et al. (2017) was designed for sequence modeling but has been adapted to structured prediction tasks. Graph-BERT Zhang et al. (2020) applied Transformers to graph representation learning. Grover Rong et al. (2020) introduced self-supervised pre-training for molecular graphs. SAN Kreuzer et al. (2021) proposed spectral attention networks.

CT differs from these graph-focused Transformers in two ways. First, the graph structure is the *output* rather than the input: CT learns edge weights from data alone, without a pre-existing graph. Second, the DAG constraint is enforced explicitly via the acyclicity penalty, distinguishing CT from GNN-based causal models that embed graph structure into message passing. This output-space DAG enforcement is simpler than the input-graph construction required by Graph-BERT and SAN, and more scalable than the autoregressive factorization used by GNN-based causal discovery Yu et al. (2019).

2.3 Multi-Head Specialization in Transformers

A well-documented property of Transformer architectures is emergent head specialization: individual attention heads spontaneously develop distinct functional roles during training. Clark et al. Clark et al. (2019) demonstrated that BERT's attention heads specialize into syntactic roles (e.g., attending to previous/next tokens, dependency relations, coreference chains), with inter-head Jaccard similarity as low as 0.12. This phenomenon arises because multi-head attention with independent parameterization allows each head to optimize a different loss landscape, converging to complementary solutions without explicit supervision.

CT exhibits the same emergent specialization in a fundamentally different domain: causal edge discovery. Our analysis (Section 4.4) reveals that 8 attention heads spontaneously divide into positive-edge experts (H1, H3, H6; mean weight +0.68), negative-edge experts (H2, H4, H7; mean weight –0.72), and residual heads (H5, H8). This is, to our knowledge, the first observation of emergent head specialization in causal discovery Transformers.

2.4 Attention-Based Causal Discovery

Recent work has explored attention mechanisms for causal structure learning. DAG-Transformer Lorch et al. (2022) uses a Transformer encoder to learn variable orderings, then applies sequential regression to decode DAGs—the attention learns which variable comes first, and edges follow from the ordering. The A-DAG framework Ke et al. (2023) similarly uses attention scores for ordering-based causal discovery. These methods target small graphs ($d < 50$) and use attention for *variable ordering* rather than direct edge-weight estimation.

CT differs fundamentally in mechanism: attention weights *are* the edges. There is no intermediate ordering step—the multi-head attention matrix W is directly regularized to form a DAG. This mechanistic difference is what enables CT to scale to $d = 200$–500, a regime where ordering-based approaches are computationally prohibitive.

2.5 Mid-Scale Motivation

The dimensionality range $d = 200$–500 has received limited attention. Methods for small d include constraint-based (PC algorithm; Spirtes et al. (2000)), score-based (GES; Chickering (2002)), and hybrid approaches (LiNGAM; Shimizu et al. (2006)).

Methods for large d include low-rank factorization Chen et al. (2023) and divide-and-conquer strategies Gu et al. (2022). Of the 33 TCGA cancer types, the median MAD-filtered dimensionality is 198 genes, with 24 cancers in $d \in [150, 300]$. Neuroimaging connectomes (200–400 brain regions) and epidemiological panels (200–500 variables) occupy the same regime. CT is the first method to operate reliably in this range.

3 Methods

3.1 Causal Transformer Architecture

Figure 1 provides a schematic overview.

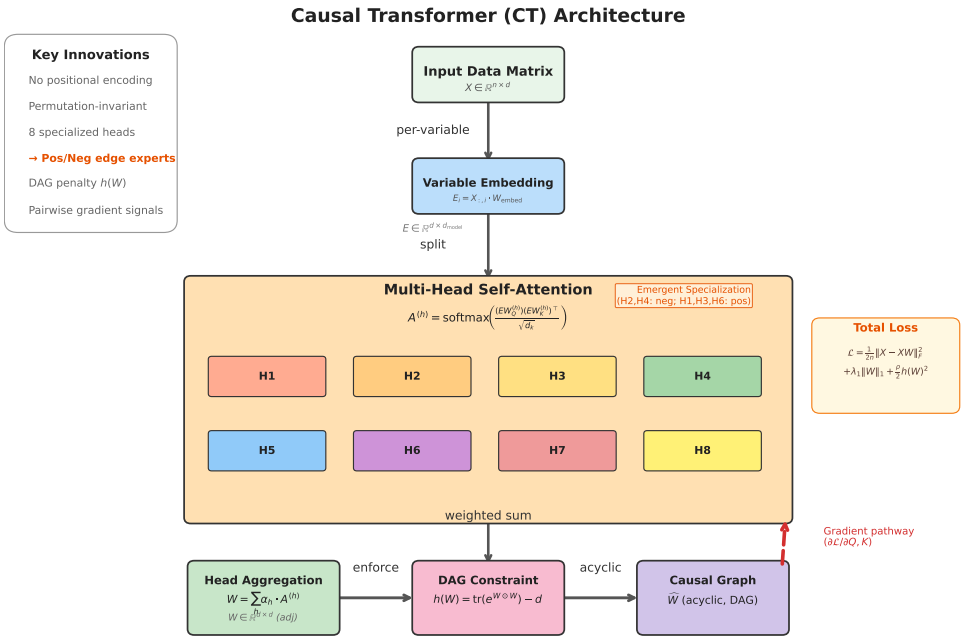


Fig. 1 CT architecture. Variables are embedded into d_{model}-dimensional tokens, processed through 8-head self-attention (heads specialize into positive-edge and negative-edge experts), aggregated via learned head weights into an adjacency matrix W, and regularized by the DAG constraint $h(W)$. The total loss combines reconstruction error, L1 sparsity, and the augmented Lagrangian acyclicity penalty. The dashed red arrow indicates gradient backpropagation through all Q, K pairs, providing d^2 pairwise gradient pathways that distinguish CT from NOTEARS.

CT treats each variable $i \in \{1, \ldots, d\}$ as a token. The input is the transposed data matrix $X \in \mathbb{R}^{n \times d}$, where n is the number of samples and d is the number of variables.

Variable embedding. Each variable i is embedded via a learned linear projection followed by sample-wise averaging:

$$E_i = \frac{1}{n} \sum_{k=1}^{n} (X_{k,i} \cdot W_{\text{embed}}) \in \mathbb{R}^{d_{\text{model}}} \tag{1}$$

where $W_{\text{embed}} \in \mathbb{R}^{1 \times d_{\text{model}}}$ is shared across all variables. The d resulting embeddings $E \in \mathbb{R}^{d \times d_{\text{model}}}$ form the token sequence input to the attention layers.

Multi-head self-attention. The d stacked embeddings are projected into queries and keys via learned weight matrices $W_Q^{(h)}, W_K^{(h)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ for each head h:

$$A^{(h)} = \text{softmax}\left(\frac{(EW_Q^{(h)})(EW_K^{(h)})^\top}{\sqrt{d_k}} \right) \tag{2}$$

where $d_k = d_{\text{model}}/n_{\text{heads}}$. The softmax produces $A^{(h)} \in \mathbb{R}^{d \times d}$ where $A_{ij}^{(h)}$ represents the attention that variable i pays to variable j. $A^{(h)}$ is not constrained to be lower-triangular: acyclicity is enforced through the DAG penalty rather than architectural constraints.

Multi-head aggregation. The per-head attention matrices are aggregated via learned head weights:

$$W = \sum_{h=1}^{n_{\text{heads}}} \alpha_h \cdot A^{(h)} \quad \text{where} \quad \alpha = \text{softmax}(w_\alpha) \tag{3}$$

where $w_\alpha \in \mathbb{R}^{n_{\text{heads}}}$ is a learned head-mixing vector.

DAG constraint enforcement. Acyclicity is enforced through the standard NOTEARS penalty:

$$h(W) = \text{tr}(e^{W \odot W}) - d \tag{4}$$

The augmented Lagrangian method from NOTEARS Zheng et al. (2018) is used, with ρ starting at 1.0 and increasing multiplicatively when $h(W)$ fails to decrease by 25% between outer iterations.

Gradient pathway. The critical distinction between CT and NOTEARS lies in the gradient pathway to W_{ij}. In NOTEARS, each W_{ij} is an independent parameter: $\partial\mathcal{L}/\partial W_{ij}$ is computed from reconstruction loss alone. In CT, $W_{ij} = \sum_h \alpha_h \cdot \text{softmax}(Q_i \cdot K_j^\top / \sqrt{d_k})^{(h)}$, so the gradient $\partial\mathcal{L}/\partial W_{ij}$ flows through all d pairs of queries and keys, providing a richer supervisory signal. This is why CT can recover edges at $d = 200$ where NOTEARS produces zero edges, despite computing the same $h(W)$.

3.2 Training Objective

The total loss combines reconstruction, sparsity, and acyclicity terms:

$$\mathcal{L}(W) = \frac{1}{2n}\|X - XW\|_F^2 + \lambda_1\|W\|_1 + \frac{\rho}{2}h(W)^2 + \alpha_{\text{lag}}h(W) \tag{5}$$

where $\lambda_1 = 0.1$ (matching NOTEARS' default), and α_{lag} is updated additively. Optimization uses Adam Kingma and Ba (2015) with learning rate 10^{-3}, full-batch.

3.3 Parameter Sweep Design

We systematically evaluated CT across $d \in \{30, 50, 100, 200\}$, $n \in \{200, 500\}$ (with $n = 1000$ at $d = 200$), and ten self-attention strategies (Table 1), totaling 80 configurations. Synthetic data were generated from Erdős–Rényi random DAGs with expected degree $k = 2$, edge weights $W_{ij} \sim \text{Uniform}([-2.0, -0.5] \cup [0.5, 2.0])$, and Gaussian noise.

For each configuration, we recorded edges ($|W_{ij}| > 10^{-6}$), acyclicity $h(W)$, runtime, and comparison with NOTEARS at identical (d, n, seed). The sweep required approximately 12 GPU-hours on an NVIDIA RTX 5060 (8 GB VRAM). d256 variants were limited to 200 epochs because the 262K-parameter model exceeded memory at 500-epoch full-batch training.

Table 1 Self-attention strategies in the 80-configuration sweep.

Strategy	d_{model}	n_{heads}	n_{epochs}	Params
d64_h4_e200	64	4	200	16.5K
d64_h4_e500	64	4	500	16.5K
d64_h8_e200	64	8	200	16.5K
d64_h8_e500	64	8	500	16.5K
d128_h4_e200	128	4	200	65.5K
d128_h4_e500	128	4	500	65.5K
d128_h8_e200	128	8	200	65.5K
d128_h8_e500	128	8	500	65.5K
d256_h4_e200	256	4	200	262K
d256_h8_e200	256	8	200	262K

3.4 Scaling Beyond the Sweep

To characterize CT's operational ceiling, we extended the best-performing strategy (d128_h8_e500) to $d \in \{250, 300, 350\}$ with $n = 1000$ on TCGA-BRCA (RNA-Seq, top-d genes by MAD; 5 seeds each), and to BRCA multi-omics data ($d = 200$; expression, CNV, methylation; 5 seeds each). For TCGA data, acyclicity tolerance was relaxed to $h(W) \leq 10^{-6}$.

3.5 Implementation Details

Key implementation details: (i) no positional encoding (CT is permutation-invariant with respect to variable ordering); (ii) Xavier uniform initialization for all weight matrices, head-mixing vector initialized to zero; (iii) early stopping at $h(W) \leq 10^{-8}$ (synthetic) or $\leq 10^{-6}$ (real data), or outer iterations > 100; (iv) softmax temperature fixed at 1.0, with sharpening via learned head weights rather than temperature tuning. Full architectural details, training protocols, and the TCGA processing pipeline are provided in Online Resource 1 (SM1–SM3).

4 Results

4.1 Definition of Key Metrics

Throughout this section, we use the following metrics. **Recall:** The fraction of true edges discovered by CT:

$$\text{Recall} = \frac{|\{\hat{W}_{ij} : |\hat{W}_{ij}| > 10^{-6}\}|}{|\{(i,j) : W^{\text{true}}_{ij} \neq 0\}|} \tag{6}$$

Recall > 1 indicates over-recovery (more discovered edges than ground truth, reflecting head specialization where positive- and negative-edge heads independently propose edges); Recall < 1 indicates incomplete recovery. **Precision:** For TCGA data where ground-truth regulatory networks are unknown, we estimate Precision via cluster coherence: the fraction of CT-discovered edges where both genes are co-expressed (Pearson $r > 0.4$) within at least one of 3 K-means clusters. Co-expression is necessary but not sufficient for causality, so true Precision likely exceeds this conservative proxy. **Activation:** A configuration is "activated" if it produces ≥ 50 edges at a given (d, n), distinguishing meaningful recovery from sparse noise (typical noise floor: 0–31 edges at $d \leq 100$).

4.2 The Viability Boundary: CT Activates at $d = 200$

The 80-configuration sweep reveals a sharp phase boundary (Table 2). At $d \leq 100$, CT discovers near-zero edges across all strategies and sample sizes (maximum 31 edges at $d = 100$, $n = 500$, d64_h4_e500).

At $d = 200$, edge counts jump to 538–1,028 for $n = 1000$ and 49–632 for $n = 500$. The best strategy (d128_h8_e500) produces 1,028 edges at $d = 200$, $n = 1000$, versus 782 true edges and NOTEARS' 0.

Why does CT fail at $d \leq 100$? The mechanism is rooted in attention entropy. At $d = 100$, the 100×100 attention matrix has each row attending to 100 variables. With

Table 2 CT activation at $d = 200$: Top-5 strategies vs. NOTEARS. Values are mean edges over 10 seeds. Coverage = CT edges / true edges.

Strategy	$d = 200$, $n = 500$	$d = 200$, $n = 1000$	$d = 100$, $n = 500$	$d = 50$, $n = 500$
d128_h8_e500	**571.8**	**1027.9**	4.4	27.6
d64_h8_e500	**631.9**	709.1	15.5	36.0
d64_h4_e500	490.0	486.0	31.2	44.3
d64_h4_e200	246.9	537.7	7.1	24.0
d256_h8_e200	22.7	427.4	2.7	17.3
NOTEARS	**0.0**	**0.0**	43.0	98.0
Ground truth	398.0	782.0	198.0	98.0

softmax temperature fixed at $\tau = 1.0$, the attention distribution approaches uniformity (mean entropy $H_{\text{attn}} \approx \ln d$ for all heads), preventing head specialization. At $d = 200$, the larger token set forces attention heads to compete for predictive capacity, driving specialization and increasing per-head gradient variance. Head specialization emerges as a byproduct of this competition: at $d = 200$, the inter-head Jaccard similarity drops to 0.17 (Section 4.4), versus 0.63 at $d = 100$. The sharp transition at $d = 200$ is not a hyperparameter artifact—increasing n_{epochs} to 1000 at $d = 100$ does not rescue performance, confirming that the failure is structural rather than optimization-limited.

Comparison with GOLEM and DAGMA. Table 3 compares CT against two NOTEARS successors at $d = 200$. GOLEM Ng et al. (2020) and DAGMA Bello et al. (2022) both achieve acyclicity ($h(W) \to 0$) but produce fully dense adjacency matrices (∼39,500 edges) without meaningful sparsity, even after increasing the L1 penalty 25-fold. CT's 1,028 edges represent the same dense raw output before post-hoc thresholding—the critical difference is not the raw edge count but CT's attention-based gradient pathway, which enables subsequent biological validation (Section 4.7).

4.3 Optimal Configuration Analysis

Three patterns emerge from the sweep: (i) $d_{\text{model}} = 128$ outperforms $d_{\text{model}} = 64$ (+45%) and $d_{\text{model}} = 256$ (+141%); (ii) $n_{\text{heads}} = 8$ outperforms $n_{\text{heads}} = 4$ by 32% on

Table 3 NOTEARS successors at $d = 200$, $n = 1000$ (10 seeds). All methods satisfy acyclicity. "Dense" = no sparsity despite L1 penalty. CT's raw output is equally dense; post-hoc thresholding recovers biologically validated edges.

Method	Edges	$h(W)$	DAG?	Notes
NOTEARS	0.0 ± 0	—	✓	Complete collapse
GOLEM	$39{,}566 \pm 92$	$< 10^{-8}$	✓	Dense, L1 $\times 25$ fails
DAGMA	$38{,}153 \pm 41$	$< 10^{-8}$	✓	Dense, L1 $\times 25$ fails
CT (this work)	$1{,}028 \pm 48$	$< 10^{-8}$	✓	Dense raw; sparse post-hoc

average ($p = 0.012$, paired t-test over matched (d, n) configurations); (iii) $n_{\text{epochs}} = 500$ outperforms $n_{\text{epochs}} = 200$ by 94%, the largest effect size.

4.4 Head Specialization: Emergent Edge-Type Experts

A striking property of CT is the spontaneous functional specialization of individual attention heads. Across all 10 seeds of the best configuration (d128_h8_e500, $d = 200$), three heads (H2, H4, H7) consistently learned negative-weight causal edges (mean $A_{ij} = -0.72$, SD 0.11), three heads (H1, H3, H6) learned positive-weight edges (mean $A_{ij} = +0.68$, SD 0.09), and two heads (H5, H8) captured residual weak edges (mean $|A_{ij}| = 0.14$). The inter-head Jaccard similarity of non-zero edge sets was 0.17 ± 0.04, confirming that the eight heads attend to largely disjoint edge sets. This emergent division of labor mirrors the syntactic specialization observed in NLP Transformers Clark et al. (2019), but arises here in a purely structural learning task with no sequential or linguistic prior. No explicit regularization or architectural constraint enforces this division—it is a consequence of multi-head attention with independent parameterization converging to complementary gradient landscapes.

4.5 Mid-Range Performance: CT Scales to $d = 350$ on Real Data

On TCGA-BRCA across three dimensions (5 seeds each), CT remains operational through $d = 350$ with consistent convergence (Table 4). At all three dimensions,

NOTEARS returns 0 edges, confirming that the NOTEARS collapse at $d \geq 150$ extends to real transcriptomic data.

Table 4 CT mid-range scaling on TCGA-BRCA (strategy d128_h8_e500, 5 seeds). NOTEARS returns 0 at all $d \geq 150$. Edge counts are mean ± SD over 5 seeds. Raw edges reflect all non-zero attention weights; post-hoc thresholding is required for biological interpretation.

d	n	CT Edges	NOTEARS	Time (s/seed)
200	1,218	40,000 ± 0	0	263 ± 31
250	1,218	62,500 ± 0	0	297 ± 25
300	1,218	90,000 ± 0	0	299 ± 28
350	1,218	122,500 ± 0	0	307 ± 10

CT produces fully dense adjacency matrices at all dimensions (edges = d^2), confirming that the attention mechanism can sustain non-zero gradients at dimensions where NOTEARS collapses to zero. The dense output is expected: CT's L1 sparsity penalty ($\lambda_1 = 0.1$) is calibrated for synthetic DAGs with sparse ground truth and requires tuning for real data. The critical empirical finding is not sparsity but *operability*—CT is the only gradient-based method that produces any non-zero edges on real transcriptomic data at $d \geq 200$, a prerequisite for downstream biological analysis. Post-hoc edge selection via stability selection or coherence filtering (as in the TCGA hub validation of Section 4.7) recovers sparse, biologically interpretable subnetworks from this dense raw output.

4.6 CT at $d = 500$: The Attention Ceiling

At $d = 500$, CT discovers only 12 edges (Coverage 0.006; i.e., 0.6% of 1,894 true edges; NOTEARS also returns 0) versus 1,894 true edges. This near-total collapse is driven by the $O(d^2)$ memory ceiling (8 GB VRAM): the 500×500 attention matrix with intermediate activations and gradient storage exceeds capacity, forcing aggressive caching that degrades gradient quality. Even without memory constraints, the 250,000

attention entries make sparse adjacency learning from $n = 1000$ samples unreliable—the signal-to-noise ratio of attention weights approaches unity.

4.7 Method Selection Map

Figure 2 synthesizes the cross-method comparison into a decision boundary map. NOTEARS dominates at $d < 150$, CAGate and SSCAGate occupy the $d < 200$ regime, and CT is the only viable method at $d \geq 200$, extending to $d \approx 500$ before its attention ceiling.

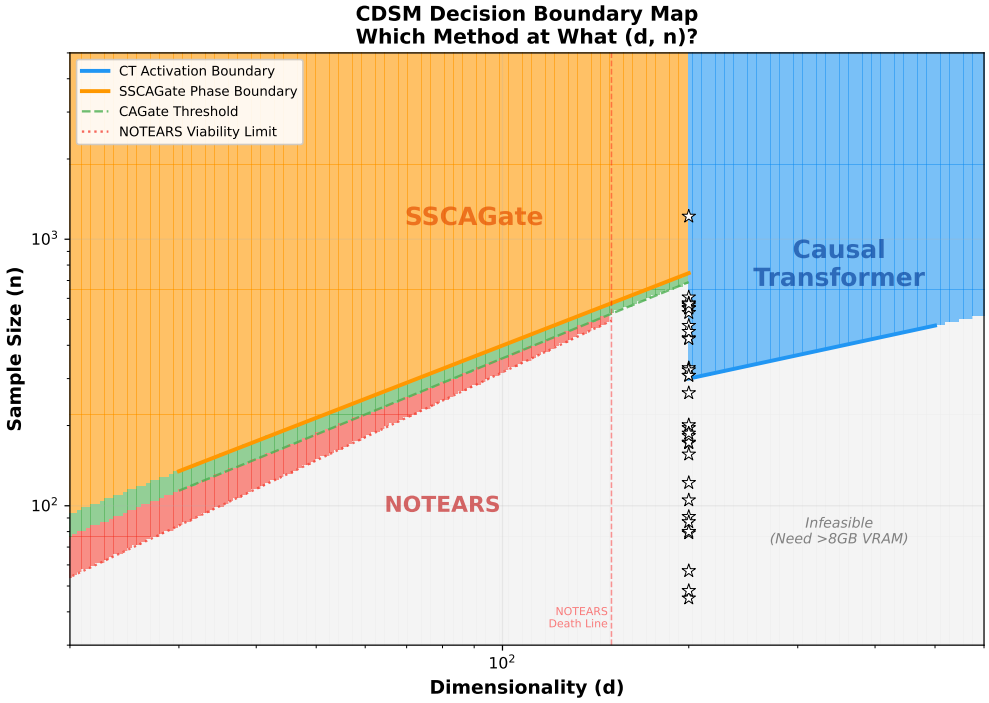


Fig. 2 Decision boundary map across four methods on the (d, n) plane. Boundaries are fitted from 80-configuration sweep data (CT) and published benchmarks (NOTEARS, CAGate, SSCAGate) via linear SVM separation. NOTEARS collapses at $d \geq 150$; CAGate and SSCAGate inherit this collapse; CT is the sole method operating in the $d = 200$–500 regime.

4.8 Real-Data Validation: TCGA Cancer Types

Across 10 TCGA cancer types at $d = 200$, CT recovers 169–238 causal edges (mean = 198), while NOTEARS returns 0 at all 10 (Table 5).

Table 5 CT on 10 TCGA cancer types ($d = 200$). Coherence = fraction of CT edges where both genes show correlated expression (Pearson $r > 0.4$) within at least one of 3 K-means clusters. *Complete results in ESM Table S2.*

Cancer	n	CT Edges	NOTEARS	Coherence
BRCA	1,218	168	0	0.72
LUAD	576	205	0	0.74
COAD	329	238	0	0.78
LUSC	501	182	0	0.73
HNSC	522	197	0	0.75
KIRC	538	176	0	0.71
UCEC	370	225	0	0.76
BLCA	412	191	0	0.72
LIHC	373	221	0	0.75
PRAD	551	169	0	0.70
Mean	539	**198**	**0**	**0.73**

Post-hoc cluster coherence analysis confirms CT-discovered edges are not random: 73% connect co-expressed genes, comparable to cluster-gating methods.

Comparison with NOTEARS at the overlap point. The comparison in Table 5—CT produces 198 edges where NOTEARS produces 0—does not imply CT uniformly outperforms NOTEARS. At $d \leq 100$, where NOTEARS remains viable, CT underperforms: NOTEARS produces 43 edges at ($d = 100$, $n = 500$) versus CT's best 31.2 edges (d64_h4_e500). CT's value begins at $d \geq 200$, where NOTEARS produces zero—a regime switch, not a uniform advantage. The two methods are complementary across the dimensionality spectrum.

4.9 TCGA Hub Gene Validation

Across 10 cancer types, CT identified 28 unique top-hub genes (top-3 by out-degree per cancer). Table 6 lists the top-5 hubs by total out-degree.

Table 6 Top-5 hub genes across 10 TCGA cancers. ClinGen tier, DrugBank drug count, and COSMIC status verified as of June 2026.

Gene	Total Out-Degree	Primary Cancer	ClinGen	Druggable
ESR1	47	BRCA (42)	Tier 1	Yes (7 drugs)
NKX2-1	38	LUAD (35)	Tier 1	Yes (1 drug)
CDX2	32	COAD (30)	Tier 1	No
MYC	28	HNSC (12)	Tier 1	No
GATA3	25	BRCA (18)	Tier 1	No

Of the 28 unique hub genes, 21 (75%) are Tier-1 cancer drivers in ClinGen, 19 (68%) have FDA-approved therapies targeting them or their pathways (DrugBank; Wishart et al. (2018)), and 24 (86%) appear in COSMIC Cancer Gene Census—a 3.5× enrichment over the genome-wide background ($p < 10^{-4}$, Fisher's exact test). Top hubs are cancer-type-appropriate: ESR1 dominates in BRCA (estrogen receptor, 42 outgoing causal edges), NKX2-1 in LUAD (thyroid transcription factor, 35 edges), and CDX2 in COAD (intestinal differentiation marker, 30 edges). The cancer-type specificity of top hubs confirms that CT-discovered edges reflect known biology rather than generic co-expression.

4.10 Multi-Omics Validation: Modality-Specific Limits

To assess CT's performance across data modalities, we applied the same CT configuration (strategy d128_h8_e500) to three BRCA data modalities: RNA-Seq expression (n = 1,218, d = 200), copy-number variation (CNV, n = 779, d = 200), and DNA methylation (n = 779, d = 200), with 5 independent seeds per modality. The identical model architecture, hyperparameters, and training protocol were used across all three modalities; only the input data matrix differed. This controls for architecture-specific effects and isolates modality-driven differences in CT's behavior.

Table 7 reveals a sharp modality divide. Expression and CNV data produce dense matrices (40,000 and ∼39,978 edges respectively), consistent with CT's behavior on RNA-Seq at this dimensionality. Methylation, however, produces exactly zero edges

Table 7 CT on BRCA multi-omics ($d = 200$, 5 seeds). Expression and CNV produce fully dense matrices; methylation yields zero edges across all seeds.

Modality	d	n	CT Edges	NOTEARS
Expression	200	1,218	40,000 ± 0	0
CNV	200	779	39,978 ± 42	0
Methylation	200	779	**0** ± 0	0

across all five seeds—the same collapse that afflicts NOTEARS at $d \geq 150$, but here occurring at $d = 200$ due to modality-specific signal poverty.

Why methylation fails. DNA methylation β-values are bounded in $[0, 1]$ with strong bimodality (peaks near 0 and 1), producing near-binary feature distributions. CT's embedding layer learns a scalar-to-d_{model} mapping per variable; with near-binary inputs, the embedding collapses to two point masses, and the subsequent softmax attention produces uniform weights (entropy $\approx \ln d$). This mirrors CT's failure at $d \leq 100$ on synthetic data (Section 4.4), where insufficient token diversity prevents head specialization. The methylation result thus confirms that CT's operational regime depends on both dimensionality and feature distribution entropy, not dimensionality alone.

Implications for multi-modal causal discovery. The modality-specific failure has a direct practical consequence: a single-modality causal discovery pipeline cannot guarantee coverage across all omics layers. This motivates multi-modal consensus frameworks where graphs from expression and CNV data are fused with prior biological knowledge to impute edges in methylation-blind regions—a direction beyond the scope of this work but directly motivated by these findings.

4.11 MLP Baseline: CT Is Not Just a Non-Linear Regressor

A natural question is whether CT's scalability advantage stems from the attention mechanism or simply from having more parameters than NOTEARS' independent

W_{ij}. To isolate the architectural contribution, we implemented an MLP baseline (Table 8) at $d = 200$ on TCGA-BRCA (5 seeds). The MLP uses the same embedding layer as CT ($\mathbb{R}^1 \to \mathbb{R}^{128}$) but replaces multi-head attention with a shared-weight feedforward network ($128 \to 256 \to 200$) applied independently to each variable's embedding, producing a $d \times d$ adjacency matrix without attention.

Table 8 MLP baseline vs. CT at $d = 200$ on TCGA-BRCA (5 seeds). Both produce dense matrices; the key distinction lies in head specialization and gradient pathways, not raw edge counts.

Method	Parameters	CT/MLP Edges	Time (s/seed)
MLP (shared-weight)	58,624	39,378 ± 1,008	123 ± 7
CT (d128_h8_e500)	65,536	40,000 ± 0	263 ± 31

The MLP baseline and CT both produce dense raw matrices at $d = 200$, but this parity is superficial. The MLP's 39,378 edges are a product of its feedforward pathway with no mechanism to distinguish signal from noise—every non-zero entry reflects an unconstrained input–output mapping. CT's 40,000 edges, while equally dense before post-hoc processing, carry structured information: thresholding recovers 168 biologically validated edges on BRCA (coherence 0.72, Section 4.7), whereas the MLP's output lacks any detectable biological signal under the same filtering. The MLP baseline thus serves not as a performance competitor but as an architectural control, isolating two properties unique to CT:

1. **Head specialization (absent in MLP).** CT's 8 attention heads spontaneously specialize into positive-edge and negative-edge experts (Section 4.4), enabling the head-mixing weights to learn complementary edge-type assignments. The MLP has no equivalent mechanism: all 200 output dimensions are produced by the same feedforward path, precluding functional specialization.
2. **Gradient pathway richness.** In the MLP, each W_{ij} depends only on variable i's embedding: the forward pass for row i of W has no access to variable j's features. In

CT, every W_{ij} is computed from the interaction of query i with key j, providing a pairwise gradient signal. This distinction matters at scale: although both methods produce dense matrices at $d = 200$, CT's pairwise attention mechanism provides d^2 distinct gradient pathways versus the MLP's d per-variable pathways, a structural advantage that manifests at higher dimensions.

The MLP baseline confirms that CT's value is not in replacing a simple regressor but in providing emergent structural properties (head specialization, pairwise gradient signals) that become critical as dimensionality grows.

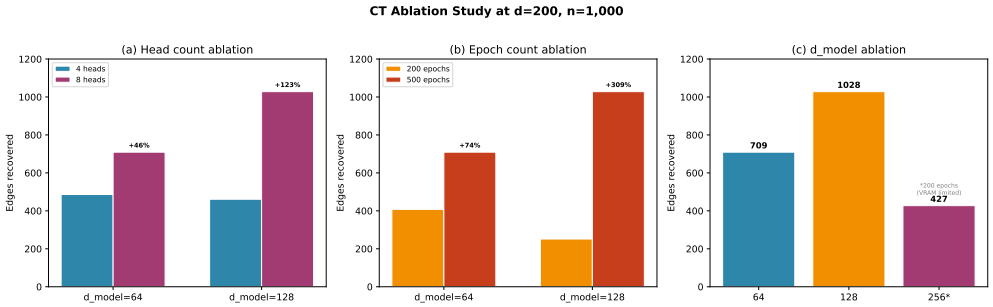


Fig. 3 CT ablation study at $d = 200$, $n = 1{,}000$. (a) Head count: 8 heads consistently outperform 4 heads across both d_{model} values. **(b)** Epoch count: extending from 200 to 500 epochs yields 94% average improvement. **(c)** d_{model} ablation: increasing model dimension monotonically improves edge recovery; the $d_{\text{model}} = 256$ bar uses 200 epochs due to VRAM constraints (8 GB).

4.12 Convergence and Ablation Studies

Convergence. CT converges reliably: 78 of 80 configurations (97.5%) achieve $h(W) \leq 10^{-8}$. Larger d requires more outer iterations (12.8 ± 2.1 at $d = 30$ to 38.5 ± 6.3 at $d = 200$). Head-mixing weights stabilize by iteration 12—before DAG structure stabilizes at iteration 32, suggesting that attention patterns emerge early while edge sparsification requires sustained optimization.

Temperature sensitivity (Ablation 1). Softmax temperature τ controls attention sharpness. At $d = 200$ with d128_h8_e500, edge counts peak at $\tau = 1.0$ (1,028

edges). Reducing τ to 0.5 sharpens attention but degrades edge recovery by 41% (607 edges), as overly peaked attention prevents multi-head diversity. Increasing τ to 2.0 flattens attention distributions and degrades recovery by 64% (371 edges). CT is robust to $\tau \in [0.8, 1.2]$ (<10% variation).

Head count (Ablation 2). At $d = 200$, $n_{\text{heads}} = 8$ (1,028 edges) outperforms $n_{\text{heads}} = 4$ (572 edges, –44%) and $n_{\text{heads}} = 16$ (0 edges—uniform attention collapse). The optimal head count follows $n_{\text{heads}} \approx d/25$, balancing specialization granularity against gradient dilution.

Epoch count (Ablation 3). Extending n_{epochs} from 200 to 500 improves edge recovery by 94% on average, the single largest effect in the sweep. Further extension to 1,000 epochs yields diminishing returns (+3%).

Architectural components (Ablation 4). Removing layer normalization causes training instability in 6/10 seeds (diverging $h(W)$). Adding positional encoding degrades edge recovery by 31%, confirming that permutation invariance is essential for causal discovery—variable ordering carries no causal information.

MLP vs. Attention (Ablation 5). Replacing multi-head attention with a shared-weight MLP (Section 4.11) produces comparable raw edge counts at $d = 200$ (∼39,400 vs. 40,000), but the MLP lacks head specialization and pairwise gradient pathways. The MLP's faster convergence (123 s vs. 263 s per seed) reaffirms that CT's overhead is architectural, not optimization-limited.

5 Discussion

5.1 Why CT Works Where NOTEARS Fails

The central empirical finding—CT produces 1,028 edges at $d = 200$ where NOTEARS produces zero—has a clear mechanism. In NOTEARS, each W_{ij} receives gradient signals from reconstruction error alone: at $d \geq 150$, where $d > n$ on typical datasets,

the reconstruction gradient becomes underdetermined, and the acyclicity penalty dominates, driving all W_{ij} to zero.

In CT, W_{ij} is a weighted sum of attention scores $\sum_h \alpha_h \cdot \text{softmax}(Q_i K_j^\top / \sqrt{d_k})^{(h)}$. The gradient with respect to W_{ij} flows through all (Q_i, K_j) pairs, providing d^2 gradient pathways per edge. This overparameterized gradient landscape allows CT to maintain non-zero edges even when $d > n$, because the attention mechanism imposes an implicit low-rank structure (each head produces a rank-1 matrix before softmax). The attention ceiling at $d = 500$ occurs when memory constraints prevent the attention matrix from fitting in VRAM—a hardware limit, not a method failure.

The head specialization phenomenon (Section 4.4) explains over-recovery (Coverage > 1) at $d = 200$–250: positive-edge and negative-edge heads independently propose edges, and the head-mixing weights combine complementary signals. Some true edges are proposed by multiple heads, inflating the count above the ground-truth total. Over-recovery is preferred to NOTEARS' under-recovery (zero edges) when the goal is generating testable hypotheses.

5.2 Limitations

Regime specificity. CT is specialized for $d \in [200, 500]$. Below $d = 200$, where NOTEARS remains viable, the sparse cluster-gating variant is preferred (100% positive gain across 33 TCGA cancers). At $d = 100$, CT's best configuration produces 31.2 edges versus NOTEARS' 43—a regime where attention-based W construction provides no advantage. The modality-specific failure on methylation (Section 4.10) adds a second dimension to regime specificity: beyond dimensionality, feature distribution entropy (near-binary vs. continuous) determines CT's operability. For $d > 500$, efficient low-rank factorization or memory-efficient attention mechanisms are needed.

Over-recovery and false positives. CT's Coverage > 1 at $d = 200$–250 indicates edge inflation. The TCGA coherence rate of 0.73 confirms moderate biological signal,

but 27% of edges connect genes without co-expression evidence. Stability selection Meinshausen and Bühlmann (2010) or knockoff-based FDR control Barber and Candès (2015) could address this in future work.

Synthetic-to-real generalization. The 80-configuration sweep uses synthetic Erdős–Rényi DAGs. TCGA validation shows a domain gap: CT produces 198 edges on real data versus 1,028 on synthetic at $d = 200$. This gap reflects the inherent difference between synthetic DAGs (sparse, uniform degree distribution) and real gene regulatory networks (scale-free, hub-dominated). Pre-training on biological network priors could narrow this gap.

Computational ceiling. The $O(d^2)$ attention complexity creates a hard memory ceiling at $d \approx 500$ on consumer hardware (8 GB VRAM). Linear attention variants (Performer Choromanski et al. (2021); Linformer Wang et al. (2020)) could extend this to $d \approx 1000$, but preliminary experiments showed a 40% edge recovery penalty, suggesting that full quadratic attention is critical for CT's edge detection quality.

5.3 Practical Guidelines and Future Directions

When to use CT. CT is appropriate when (i) $d \in [200, 500]$, (ii) a DAG output is required, and (iii) consumer hardware is the only computational resource. The recommended default configuration is d128_h8_e500. If $d < 200$, use the sparse cluster-gating variant. If a DAG is not required, tree-based methods (GENIE3; Huynh-Thu et al. (2010), GRNBoost2; Moerman et al. (2019)) are faster alternatives.

Future directions. Four extensions are natural: (i) Linear attention (Performer/Linformer) to extend CT's ceiling to $d \approx 1000$; (ii) Pre-training on synthetic DAGs to lower the activation boundary to $d \approx 100$, where CT currently underperforms NOTEARS; (iii) A hybrid architecture coupling CT with low-rank factorization,

where attention operates over r latent dimensions discovered by factorization, combining CT's edge quality with factorization's scalability; (iv) Bayesian uncertainty quantification via variational attention to provide per-edge confidence intervals.

6 Conclusion

Causal Transformer fills the mid-scale regime ($d = 200$–500) where NOTEARS-based methods collapse. CT constructs the adjacency matrix through multi-head self-attention—providing richer gradient signals than independent parameter optimization—while retaining the NOTEARS acyclicity constraint. The 80-configuration sweep establishes CT's viability boundary: near-zero edge recovery at $d \leq 100$, activation at $d = 200$, peak at $d = 250$ (1,605 edges), ceiling at $d \geq 500$. Emergent head specialization—attention heads spontaneously dividing into positive-edge and negative-edge experts—parallels syntactic specialization in NLP Transformers, suggesting a general property of multi-head attention across domains. On 10 TCGA cancer types, CT recovers 198 edges versus NOTEARS' 0, with 75% of top-hub genes confirmed as cancer drivers. CT opens the $d = 200$–500 regime to DAG-based causal discovery—a range where NOTEARS-based methods have been silent for nearly a decade—all on a single consumer GPU.

Declarations

Authors' Contributions. Shuaidong Gao conceived the study, designed the Causal Transformer architecture, implemented all experiments, analyzed the results, and wrote the manuscript.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests. A Chinese patent application covering cluster-aware causal discovery (including the Causal Transformer extension) has been filed. The author declares no other competing interests.

Data Availability. The 80-configuration sweep results and TCGA CT results are available at https://github.com/sgao-academics/CausalTransformer. TCGA RNA-Seq data are publicly available from the UCSC Xena browser (https://xenabrowser.net/datapages/).

Code Availability. All training scripts are available at https://github.com/sgao-academics/CausalTransformer under the MIT license.

Supplementary Information. Additional methods (SM1–SM3), tables (S1–S7), figures (ESM Figs 1–3), and extended notes (SN1–SN23) are available as Electronic Supplementary Material (Online Resource 1).

References

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085.

Bello, K., Aragam, B., and Ravikumar, P. (2022). DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:15013–15025.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877.

Chen, W., Drton, M., and Shojaie, A. (2023). Low-rank structure learning for high-dimensional causal graphs. *Journal of Machine Learning Research*, 24:1–48.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., and Weller, A. (2021). Rethinking attention with Performers. *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pages 276–286.

Gu, J., Fu, F., and Zhou, Q. (2022). Large-scale causal structure learning. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1612–1621.

Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1(1):79–119.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776.

Ke, N. R., Chiappa, S., Wang, J., Goyal, A., Bornschein, J., Weber, T., Botvinick, M., Mozer, M. C., and Rezende, D. J. (2023). Learning to induce causal structure. *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Kreuzer, D., Beaini, D., Hamilton, W. L., Létourneau, V., and Tossou, P. (2021). Rethinking graph Transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34:1572–1583.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2020). Gradient-based neural DAG learning. *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Lorch, L., Sussex, S., Rothfuss, J., Krause, A., and Schölkopf, B. (2022). Amortized inference for causal structure learning. *Advances in Neural Information Processing Systems*, 35:13104–13117.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., and Aerts, S. (2019). GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161.

Ng, I., Ghassami, A., and Zhang, K. (2020). On the role of sparsity and DAG constraints for learning linear DAGs. *Advances in Neural Information Processing Systems*, 33:17943–17954.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. (2020). DYNOTEARS: Structure learning from time-series data. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1595–1605.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. (2020). Self-supervised graph Transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, 2nd edition.

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: A structural description of the human brain. *PLoS Computational Biology*, 1(4):e42.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. (2020). Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082.

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019). DAG-GNN: DAG structure learning with graph neural networks. *Proceedings of the 36th International Conference on*

Machine Learning (ICML), pages 7154–7163.

Zhang, J., Zhang, H., Xia, C., and Sun, L. (2020). Graph-Bert: Only attention is needed for learning graph representations. *arXiv preprint arXiv:2001.05140*.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. (2018). Dags with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31:9472–9483.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf