A Path Already Walked

On Inheriting Network-Neuroscience Tools for Mechanistic Interpretability

Phongsakon Mark Konrad[1], Toygar Tanyel[2], Serkan Ayvaz[1]
[1]Centre for Industrial Software, University of Southern Denmark [2]Promake, Newark, DE, USA

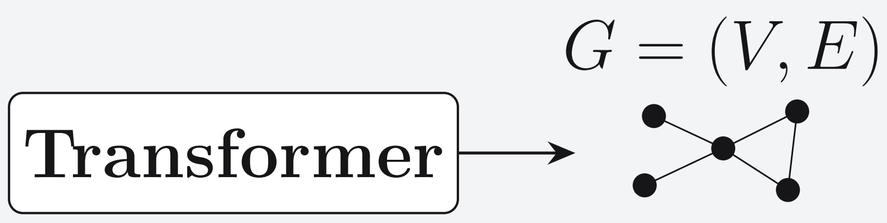
Scan for paper

TL;DR

Mech-interp is moving from neurons and heads toward circuits and features; many phenomena are **relational**, not component-local. We argue for a **disciplined import** of network-neuroscience graph tools, *not* a loose "transformers are brains" analogy. The contribution is a **methodological standard**: a graph contract and **eight falsifiable translations**; **no transformer experiments are reported**.

The question: one model, many graphs

A transformer induces *many* graphs, not one: "small-worldness" or "rich club" are meaningful only once the graph object is declared.



$V \in \{$neurons, heads, SAE features, blocks$\}$ and $E \in \{$weights, correlation, patching, Jacobian$\}$. Each choice is a *different* object, so declare it before any claim.

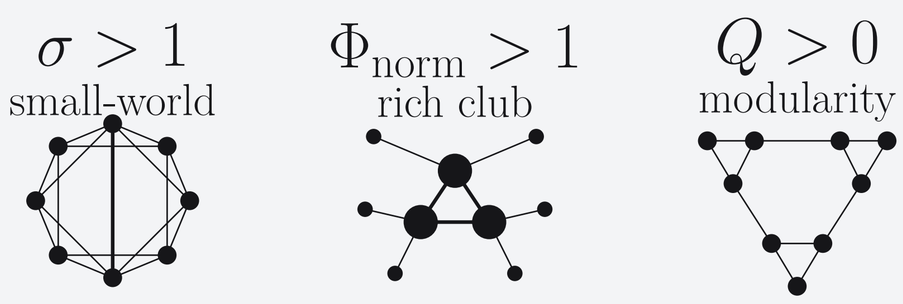
Methodology

The Transformer Graph Contract. A graph $G = (V, E)$ with adjacency $A \in \mathbb{R}^{n \times n}_{>0}$ must declare:

- **Node set** V: neurons, heads, SAE features, residual directions, blocks, attribution-graph nodes.
- **Edge construction**: weights, activation correlation, patching effect, or Jacobian proxy (A_{ij}: $i \to j$).
- **Prompt distribution** defining the graph.
- **Projection / directionality** and **thresholding**: split signed effects, declare symmetrisation, match a density target.
- **Null model** preserving architecture, density, degree, and layer order.
- **Behavioural target** the graph should predict.

Statistics gain meaning only against a null:

$$\sigma = \frac{C/C_{\text{rand}}}{L/L_{\text{rand}}}, \quad \Phi_{\text{norm}}(k) = \frac{\Phi(k)}{\Phi_{\text{rand}}(k)}, \quad Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \gamma\frac{k_i k_j}{2m}\right)\delta(c_i, c_j).$$



Each statistic reads off a *different* structure, meaningful only when measured against a matched null.

Worked example. A local effective-connectivity proxy: $\mathrm{EC}_{ab}(\mathcal{X}) = \mathbb{E}_{x \sim \mathcal{X}} \|J_{ba}(x)\|_F^2$ (sensitivity, **not** causal).

The null preserves **degree** and **density** but scrambles structure. A statistic means something only *relative to its null*.

Limitations

No transformer experiments are reported; neuroscience findings do not transfer automatically. The Jacobian edge is a sensitivity proxy, not a causal intervention; the many degrees of freedom (threshold, symmetrisation, null model) demand that *discovery* be separated from *validation*.

Results: primitive → measurement

Each network-neuroscience primitive becomes a concrete measurement *only* after the graph object, projection, and null are declared.

Primitive	Transformer question
Small-world	Distributed routing as width grows?
Rich club	Redundancy among important heads / features
Modularity	Functional groups of heads, features, blocks
Controllability	Candidate steering sets (driver nodes)
Motifs	Recurring circuit motifs vs. enriched subgraphs
Edge-centric FC	Prompt-edge variability, overlapping communities
Higher-order / dynamic	Polyadic feature interactions, token states

Results: testable translations

The catalogue turns the analogy into **falsifiable** transformer questions, each with an explicit failure condition, meaningful once the contract is fixed.

Question	Claim to test	Would fail if
Small-world scaling *(Direct)*	EC graphs exceed architecture-preserving nulls; hub summaries improve held-out prediction.	σ is null-level and hubs fail beyond unit baselines.
Rich-club ordering *(Direct)*	High-degree heads/features interconnect above degree nulls and overlap importance sets.	$\Phi_{\text{norm}}(k) \leq 1$ or overlap random-level.
Modularity *(Moderate)*	Communities align with labelled functions (induction, copying, refusal).	Alignment random-level under labelled probes.
Motif enrichment *(Moderate)*	Motif z-score vectors cluster fine-tuned models by behaviour.	Profiles collapse under calibrated nulls.
Min. steering set *(Moderate)*	High-controllability nodes overlap patching-derived steering targets.	Controllability and steering sets are disjoint.
Cost-efficiency *(Exploratory)*	Trained models sit nearer a declared compute-depth cost frontier than matched alternatives.	Reasonable cost definitions remove the effect.
Communication mix *(Moderate)*	Mixtures of shortest-path, diffusion, navigation, and communicability policies predict behaviour better than any single policy.	A single policy matches mixture performance.
Dynamic topology *(Exploratory)*	Token or prompt segments induce recurring connectivity states with behavioural dwell-time structure.	State labels fail to predict held-out behaviour.

Future work

Pick one primitive, declare the contract, and run the null-model test. A first benchmark should be narrow: start with GPT-2 small, attention heads as nodes, held-out patching effects as directed edges; fix density before topology; measure σ, Φ_{norm}, and Q against architecture-preserving nulls with a held-out behavioural target.

Conclusion

The useful import is **methodological, not metaphorical**: a declared graph object scored against an architecture-preserving null with a behavioural target. **Discipline is the contribution**, not any single result.

References

[1] Friston (2011). Functional and effective connectivity.
[2] Bullmore & Sporns (2009). Complex brain networks.
[3] Elhage et al. (2022). Toy models of superposition.
[4] Ameisen et al. (2025). Circuit tracing / attribution graphs.
[5] Conmy et al. (2023). Automated circuit discovery.