A Path Already Walked: On Inheriting Network-Neuroscience Tools for Mechanistic Interpretability

Anonymous Authors[1]

Abstract

Mechanistic interpretability is moving from neurons and heads toward circuits, dictionary features, and attribution graphs. That transition is productive, but it also raises a familiar issue. Many important phenomena are relational rather than component-local. Network neuroscience has spent two decades building graph vocabulary, null models, and failure modes for related problems. We argue for a disciplined import rather than a loose brain analogy. We specify the transformer graph contract required before the import is meaningful, give a compact mapping from network-neuroscience primitives to transformer analyses, work through a local effective-connectivity proxy for gated MLPs, and state eight testable translations with failure criteria. We do not report transformer experiments, and we do not claim neuroscience results transfer automatically.

1. Position

Mechanistic interpretability has made progress, but its explanatory units keep moving. Early work analyses individual neurons (Olah et al., 2020). Later work studies attention heads, circuits, superposition, dictionary features, and attribution graphs (Elhage et al., 2021; Wang et al., 2023; Elhage et al., 2022; Bricken et al., 2023; Templeton et al., 2024; Ameisen et al., 2025). This shift shows that part-enumeration is useful but incomplete.

Recent work makes the issue more concrete. Sparse autoencoders can be inconsistent across runs (Song et al., 2025), may not identify canonical units (Leask et al., 2025), and require falsification-oriented explanation tests (Ma et al., 2025). Circuit work is also evolving into graph work through automated circuit discovery, sparse-feature circuits, and attribution graphs (Conmy et al., 2023; Marks et al., 2025; Lindsey et al., 2025). A useful next layer is not another named unit. It is a population-level graph vocabulary with explicit null models.

Network neuroscience has such a vocabulary. It distinguishes structural, functional, and effective connectivity (Friston, 2011), uses graph-theoretic summaries of brain organization (Bullmore & Sporns, 2009; Bassett & Sporns, 2017), and treats null-model choice as part of the method. The specific primitives include small-world structure, rich clubs, modularity, efficiency, controllability, edge communities, higher-order structure, and time-varying connectivity (Watts & Strogatz, 1998; Humphries & Gurney, 2008; Colizza et al., 2006; van den Heuvel & Sporns, 2013; Newman, 2006; Latora & Marchiori, 2001; Liu et al., 2011; Gu et al., 2015; Faskowitz et al., 2020; Battiston et al., 2020; Lurie et al., 2020). Recent LLM papers already import pieces of this network frame (Liu et al., 2025; Bhandari et al., 2025; Zheng et al., 2025), while adjacent work imports neuroscience-inspired residual-stream dynamics (Fernando & Guitchounts, 2025). Our contribution is the map. It states what has to be specified before these tools become transformer measurements, and which findings would weaken the analogy.

2. Why the Analogy Is Methodological

The relevant parallel is not that transformers are brains. It is that both fields begin with measurable parts, then encounter phenomena where relational structure matters. Neuroscience began with single cells, regions, and lesion effects. Mech-interp began with neurons and attention heads (Olah et al., 2020; Elhage et al., 2021). These primitives were tractable because they could be named, visualised, ablated, and counted.

Both fields then met mixed representations. Prefrontal-cortex cells can respond to mixtures of task variables (Rigotti et al., 2013), and transformer neurons can represent more features than there are dimensions through superposition (Elhage et al., 2022). Sparse coding was a natural response in both settings. Overcomplete sparse bases produce V1-like features from natural images (Olshausen &

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Table 1. Network-neuroscience primitives become useful transformer measurements only after the graph object, projection, and null model are declared. The import depends on a contract, not a metaphor.

Primitive	Network meaning	Transformer question	Must declare
Connectivity	structural, functional, effective (Friston, 2011)	weights, activation correlations, patching or Jacobian proxies	component set, prompt distribution, proxy/intervention distinction
Small-world	high clustering, short paths (Watts & Strogatz, 1998; Humphries & Gurney, 2008)	distributed routing as width grows	edge-length transform and density-matched null
Rich club	high-degree hubs interconnect (Colizza et al., 2006; van den Heuvel & Sporns, 2013)	redundancy among important heads or features	degree definition and degree-preserving null
Modularity	communities (Newman & Girvan, 2004; Newman, 2006)	functional groups of heads, features, or blocks	directed/weighted modularity null
Participation	connector vs. provincial nodes (Guimerà & Amaral, 2005)	distinguish local specialists from cross-module routers	stable partition and module labels
Efficiency	global/local path efficiency (Latora & Marchiori, 2001)	compare routing reachability across models or tasks	edge-to-distance transform
Controllability	driver nodes and Gramians (Liu et al., 2011; Gu et al., 2015)	candidate steering sets	linearisation regime and behavioural validation
Communication	paths, diffusion, navigation, communicability (Estrada & Hatano, 2008; Seguin et al., 2023)	learned mixtures of routing policies	policy family and held-out target
Motifs	enriched subgraphs (Milo et al., 2002; Sporns & Kötter, 2004)	recurring circuit motifs	motif size with density and layer-order null
Brain disorder connectomics	disorder as connectome disruption (Fornito et al., 2015)	fine-tuning as ΔEC fingerprint	matched base/fine-tuned graphs
Edge-centric FC	edge time series (Faskowitz et al., 2020)	prompt-edge variability and overlapping communities	prompt ensemble and edge feature space
Higher-order/dynamic	simplices and time-varying FC (Battiston et al., 2020; Sizemore et al., 2019; Lurie et al., 2020)	polyadic feature interactions and token-state transitions	filtration and token segmentation

Field, 1996), and sparse autoencoders attempt to recover monosemantic transformer features (Bricken et al., 2023; Templeton et al., 2024). The analogy is useful because the sparse-coding story has known limitations. Non-canonical decompositions, inconsistent feature atlases, underspecified causal roles, and fragile explanations are expected failure modes (Song et al., 2025; Leask et al., 2025; Ma et al., 2025; Sharkey et al., 2025).

Both fields also construct small computational subgraphs. Neuroscience calls enriched recurring subgraphs motifs and compares their z-scores against null ensembles (Milo et al., 2002; Sporns & Kötter, 2004). Mech-interp calls prompt-local subgraphs circuits, estimates them with patching or attribution, and now increasingly automates their discovery (Wang et al., 2023; Conmy et al., 2023; Marks et al., 2025; Ameisen et al., 2025). The overlap is not that every circuit is a motif. It is that circuit work already treats computation as a graph, while network neuroscience supplies population-level statistics and null-model discipline.

The most useful methodological import is therefore not metaphorical. It is a requirement for contracts. A graph claim about a transformer should say which components are nodes, how signed or directed effects become edges, what prompt distribution defines the graph, what null is being used, and which behavioural quantity the graph is supposed to predict. Without this contract, claims about small-worldness, rich clubs, modularity, or controllability mainly report preprocessing choices.

3. Transformer Graph Contracts

A transformer does not induce one graph. It induces many possible graphs. We use $G = (V, E)$ for a task-conditioned directed graph over declared components V, such as neurons, attention heads, SAE features, residual directions, blocks, or attribution-graph nodes. Its adjacency $A \in \mathbb{R}_{\geq 0}^{n \times n}$ records nonnegative edge strength with A_{ij} denoting influence from i to j, self-loops removed unless declared, and token/prompt effects aggregated by a stated statistic after a projection such as absolute patching effect, squared local derivative, or thresholded activation association.

Four choices are needed. First, signed quantities must be projected or separated into positive and negative graphs. Second, directed metrics must not be replaced silently by undirected ones. If symmetrisation is used, it must be declared. Third, binary metrics require a threshold or density target, ideally matched across models, plus a convention for disconnected pairs. Fourth, any null should preserve the relevant architecture, density, degree sequence, and layer order. With those choices fixed, standard graph primitives become transformer questions rather than informal descriptions.

Three examples show why the contract matters. In the simplest connected undirected case, small-worldness uses $\sigma = (C/C_{\text{rand}})/(L/L_{\text{rand}})$, where C_{rand} and L_{rand} are null-ensemble expectations. Thresholded disconnected graphs need a declared path convention or an efficiency statistic. Rich-club analysis uses $\Phi_{\text{norm}}(k) = \Phi(k)/\Phi_{\text{rand}}(k)$, so a high-degree "core" is only meaningful against a

degree-preserving null. Modularity uses $Q = \sum_{ij}(A_{ij} - \gamma k_i k_j/2m)\delta(c_i, c_j)/(2m)$ in the undirected case (Newman & Girvan, 2004), where $k_i = \sum_j A_{ij}$ is degree, $m = |E|$ is the edge count, c_i is the community label, δ is one for matching labels and zero otherwise, and γ is the resolution parameter, with $\gamma = 1$ recovering standard modularity. Weighted undirected graphs use strengths and total edge weight in the same form, while directed extensions require an explicit directed null.

4. A Worked Effective-Connectivity Proxy

Definition 1 (Gradient effective-connectivity proxy)**.** For upstream components $\mathcal{A}$ and downstream components $\mathcal{B}$ with declared intervention coordinates $a_a(x) \in \mathbb{R}^{d_a}$ and $a_b(x) \in \mathbb{R}^{d_b}$, let $J_{ba}(x)$ be the block Jacobian from $\mathcal{A}$ to $\mathcal{B}$. Define $\mathrm{EC}_{ab}(\mathcal{X}) = \mathbb{E}_{x\sim\mathcal{X}}\|J_{ba}(x)\|_F^2$.

This is a local proxy for effective connectivity, not an intervention. Patching estimates finite causal effects (Goldowsky-Dill et al., 2023; Heimersheim & Nanda, 2024). The Jacobian proxy estimates infinitesimal sensitivity. As a scalar special case, take x as the upstream coordinate vector for a gated-MLP neuron $a_n(x) = \phi(w_{g,n}^\top x)(w_{u,n}^\top x)$. Let $g = w_g^\top x$, $u = w_u^\top x$, and $c = w_g^\top w_u$. Then

$$\left\| \frac{\partial a_n}{\partial x} \right\|^2 = \phi'(g)^2 u^2 \|w_g\|^2 + 2\phi'(g)\phi(g)uc + \phi(g)^2\|w_u\|^2. \quad (1)$$

Eq. 1 is an input-sensitivity norm. An edge from upstream direction v_a to this neuron would use the squared directional derivative $|(\phi'(g)uw_g + \phi(g)w_u)^\top v_a|^2$. If measuring transmission into the residual stream, the MLP down-projection must also be included. Pairwise probing scales as $O(P|\mathcal{A}||\mathcal{B}|)$ evaluations. VJP/JVP-based Jacobian extraction can often reduce the number of derivative queries, depending on batching, block size, and memory costs.

5. Testable Translations

Table 2 is the shortest useful output of the map. These are not results. They are testable transformer questions that become meaningful once the graph contract in Section 3 is fixed. Priority reflects how direct the methodological analogy is, not confidence that the claim is true.

Worked protocol. A first instantiation uses attention heads as nodes and held-out patching effects as directed non-negative edges. This is the finite-intervention counterpart to the local Jacobian edge proxy in Section 4. Fix density before computing topology. Use a null that preserves layer order, layer-pair density, and prompt family. The target is held-out patching recovery. The claim fails if the statistic is null-level or adds no prediction beyond component counts.

Reading the catalogue. Priority marks transfer difficulty. Direct rows need only a declared graph, a matched null, and a behavioural target. Moderate rows need stronger validation because communities and routing policies are easier to create through preprocessing. Exploratory rows are stress tests drawn from wiring-cost and time-varying-connectivity debates (Bullmore & Sporns, 2012; Lurie et al., 2020). Negative results remain useful because they mark the boundary of the analogy.

6. Use Case, Related Work, and Limits

A concrete use case is deceptive or refusal fine-tuning. If local MLP interventions fail while attention-path interventions recover behaviour, the network-neuroscience vocabulary would describe the result as a change in effective connectivity rather than a single-component effect. If a fine-tune preserves most of the weight geometry while behaviour changes, the functional/effective connectivity distinction of Friston (2011) explains why activation- or intervention-defined routing shifts can be missed by weight-diff audits. If ablating the top predicted components redirects recruitment to nearby alternatives, the phenomenon can be tested as redundancy and, if recruited alternatives form a high-degree core, as rich-club-like redundancy. If refusal is concentrated in a low-rank direction (Arditi et al., 2024), the contrast with distributed routing becomes a topology question rather than a naming dispute.

A minimal report is simple. First, build matched base and fine-tuned graphs under the same component set, prompt distribution, and edge construction. Second, report the effect on local components, graph topology, and behaviour. Third, compare the observed shift with nulls that preserve layer order and density. A fine-tune that changes behaviour while leaving most structural weights similar is then, by analogy, a candidate connectopathy-like connectivity-shift phenotype (Fornito et al., 2015). A model that repairs after head ablation can be tested against redundancy statistics rather than described only as a Hydra effect (Michel et al., 2019; McGrath et al., 2023). The import turns qualitative routing stories into statistics with nulls.

The same frame suggests near-term empirical programs. Measure σ, Φ_{norm}, Q, and E_{glob} on matched effective-connectivity graphs across model scales. Cluster fine-tunes by motif profiles (Milo et al., 2002; Sporns & Kötter, 2004) or edge-centric prompt variability (Faskowitz et al., 2020). Adapt controllability diagnostics and compare them with patching-derived steering targets (Liu et al., 2011; Gu et al., 2015). Use persistent homology on threshold filtrations of EC as a multiscale descriptor of higher-order feature interactions (Sizemore et al., 2019; Battiston et al., 2020). Each starts from a graph already produced by circuit or attribution work and asks a population-level question.

Table 2. Eight translations turn the analogy into testable transformer questions. Source-backed primitive definitions are in Table 1. This table states proposed transformer tests and failure conditions.

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

The closest current papers cover one import at a time. Liu et al. (2025) test functional brain-inspired networks, Bhandari et al. (2025) analyse module communities, and Zheng et al. (2025) probes neural topology. Fernando & Guitchounts (2025) use residual-stream dynamics. These papers support the premise, but they do not make graph construction, null choice, and failure criteria the object of the contribution. Existing circuit work already speaks graph language through automated circuit discovery, sparse feature circuits, and attribution graphs (Conmy et al., 2023; Marks et al., 2025; Ameisen et al., 2025; Lindsey et al., 2025). We differ by making the cross-reference itself the contribution. The goal is not to beat these empirical papers, but to give them and their successors a shared vocabulary.

Other adjacent threads clarify the boundary. El et al. (2025) port network science to graph transformers, which is related but targets a different architecture. Connectome-inspired ML runs the reverse direction, using brain wiring to design models (Johnson et al., 2023). SAE critiques and audits identify why a single sparse-feature atlas is not enough (Sharkey et al., 2025; Song et al., 2025; Leask et al., 2025; O'Neill et al., 2025; Ma et al., 2025). Philosophical critiques argue that mechanistic interpretability needs sharper explanatory standards (Williams et al., 2025). Our proposal is one concrete response. Relational explanations should state their graph objects, nulls, and refutation conditions.

The limits are material. Transformers are fully observable, discrete, engineered systems. Brains are partly observable, continuous, and biologically constrained. A Jacobian proxy is not a causal intervention. Dense attention can make some efficiency statistics trivial. Brain controllability metrics have known caveats before they are ported to another nonlinear system (Suweis et al., 2019). These limits define the first checks on the import.

A first empirical benchmark should therefore be narrow. Use two or three model families, a fixed prompt family, one component space (e.g., heads or SAE features), and two graph constructions (e.g., patching-derived effective connectivity and activation-derived functional connectivity). Report the graph contract, then test small-world scaling and rich-club ordering with density- and architecture-preserving nulls. A positive result would show that a relational descriptor predicts behaviour or scale better than component counts alone. A negative result would rule out the simplest topology story before the field builds a more elaborate graph language on top of it.

The main methodological risk is the number of degrees of freedom. Graph construction, thresholding, symmetrisation, prompt selection, and null choice can each change the answer. A useful paper should report threshold sensitivity, distinguish between signed and unsigned effects, control for density, and separate discovery from validation. Otherwise, the import risks becoming a vocabulary upgrade without explanatory pressure.

7. Conclusion

The claim is narrow. Network neuroscience does not solve mechanistic interpretability. It gives the field a disciplined relational vocabulary. The useful next step is to choose one primitive in Table 1, define the graph contract, and run the null-model test in Table 2.

References

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024. URL https://arxiv.org/abs/2406.11717.

Bassett, D. S. and Sporns, O. Network neuroscience. *Nature Neuroscience*, 20(3):353–364, 2017. doi: 10.1038/nn.4502.

Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., and Petri, G. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1–92, 2020. doi: 10.1016/j.physrep.2020.05.004.

Bhandari, K. R., Chen, P.-Y., and Gao, J. Unraveling the cognitive patterns of large language models through module communities. *arXiv preprint arXiv:2508.18192*, 2025. URL https://arxiv.org/abs/2508.18192.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bullmore, E. and Sporns, O. Complex brain networks, graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009. doi: 10.1038/nrn2575.

Bullmore, E. and Sporns, O. The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349, 2012. doi: 10.1038/nrn3214.

Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A. Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115, 2006. doi: 10.1038/nphys209.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. URL https://arxiv.org/abs/2304.14997.

El, B., Choudhury, D., Liò, P., and Joshi, C. K. Towards mechanistic interpretability of graph transformers via attention graphs. *arXiv preprint arXiv:2502.12352*, 2025. doi: 10.48550/arXiv.2502.12352. URL https://arxiv.org/abs/2502.12352.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Estrada, E. and Hatano, N. Communicability in complex networks. *Physical Review E*, 77(3):036111, 2008. doi: 10.1103/PhysRevE.77.036111.

Faskowitz, J., Esfahlani, F. Z., Jo, Y., Sporns, O., and Betzel, R. F. Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture. *Nature Neuroscience*, 23(12):1644–1654, 2020. doi: 10.1038/s41593-020-00719-y.

Fernando, J. and Guitchounts, G. Transformer dynamics, a neuroscientific approach to interpretability of large language models. *arXiv preprint arXiv:2502.12131*, 2025. URL https://arxiv.org/abs/2502.12131.

Fornito, A., Zalesky, A., and Breakspear, M. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172, 2015. doi: 10.1038/nrn3901.

Friston, K. J. Functional and effective connectivity, a review. *Brain Connectivity*, 1(1):13–36, 2011. doi: 10.1089/brain.2011.0008.

Goldowsky-Dill, N., MacLeod, C., Sato, L., and Arora, A. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. URL https://arxiv.org/abs/2304.05969.

Gu, S., Pasqualetti, F., Cieslak, M., Telesford, Q. K., Yu, A. B., Kahn, A. E., Medaglia, J. D., Vettel, J. M., Miller, M. B., Grafton, S. T., et al. Controllability of structural brain networks. *Nature Communications*, 6:8414, 2015. doi: 10.1038/ncomms9414.

Guimerà, R. and Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005. doi: 10.1038/nature03288.

Heimersheim, S. and Nanda, N. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

Humphries, M. D. and Gurney, K. Network 'small-worldness', a quantitative method for determining canonical

network equivalence. *PLoS ONE*, 3(4):e2051, 2008. doi: 10.1371/journal.pone.0002051.

Johnson, E. C., Robinson, B. S., Vallabha, G. K., Joyce, J., Matelsky, J. K., Norman-Tenazas, R., Western, I., Villafañe-Delgado, M., Cervantes, M., Robinette, M. S., et al. Exploiting large neuroimaging datasets to create connectome-constrained approaches for more robust, efficient, and adaptable artificial intelligence. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V*, pp. 49. SPIE, 2023. doi: 10.1117/12.2663901. URL https://doi.org/10.1117/12.2663901.

Latora, V. and Marchiori, M. Efficient behavior of small-world networks. *Physical Review Letters*, 87(19):198701, 2001. doi: 10.1103/PhysRevLett.87.198701.

Leask, P., Bussmann, B., Pearce, M., Bloom, J., Tigges, C., Al Moubayed, N., Sharkey, L., and Nanda, N. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2502.04878. URL https://openreview.net/forum?id=9ca9eHNrdH.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., et al. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Liu, Y., Liu, Z., Wu, Z., Ning, J., Sun, H., Xia, S., Yang, Y., Gao, X., Qiang, N., Ge, B., Liu, T., Han, J., and Hu, X. Brain-inspired exploration of functional networks and key neurons in large language models. *arXiv preprint arXiv:2502.20408*, 2025. URL https://arxiv.org/abs/2502.20408.

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011. doi: 10.1038/nature10011.

Lurie, D. J., Kessler, D., Bassett, D. S., Betzel, R. F., Breakspear, M., Keilholz, S., Kucyi, A., Liégeois, R., Lindquist, M. A., McIntosh, A. R., et al. Questions and controversies in the study of time-varying functional connectivity in resting fmri. *Network Neuroscience*, 4(1):30–69, 2020. doi: 10.1162/netn_a_00116.

Ma, G., Pfrommer, S., and Sojoudi, S. Revising and falsifying sparse autoencoder feature explanations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=OJAW2mHVND.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

McGrath, T., Rahtz, M., Kramar, J., Mikulik, V., and Legg, S. The hydra effect, emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023. URL https://arxiv.org/abs/2307.15771.

Michel, P., Levy, O., and Neubig, G. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.10650.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. Network motifs, simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. doi: 10.1126/science.298.5594.824.

Newman, M. E. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi: 10.1073/pnas.0601602103.

Newman, M. E. and Girvan, M. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004. doi: 10.1103/PhysRevE.69.026113.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in, an introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. doi: 10.1038/381607a0.

O'Neill, C., Jayasekara, M., and Kirkby, M. Resurrecting the salmon: Rethinking mechanistic interpretability with domain-specific sparse autoencoders. *arXiv preprint arXiv:2508.09363*, 2025. doi: 10.48550/arXiv.2508.09363. URL https://arxiv.org/abs/2508.09363.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013. doi: 10.1038/nature12160.

Seguin, C., Sporns, O., and Zalesky, A. Brain network communication, concepts, models and applications. *Nature Reviews Neuroscience*, 24(9):557–574, 2023. doi: 10.1038/s41583-023-00718-5.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025. doi: 10.48550/arXiv.2501.16496. URL https://arxiv.org/abs/2501.16496.

Sizemore, A. E., Phillips-Cremins, J. E., Ghrist, R., and Bassett, D. S. The importance of the whole, topological data analysis for the network neuroscientist. *Network Neuroscience*, 3(3):656–673, 2019. doi: 10.1162/netn_a_00073.

Song, X., Muhamed, A., Zheng, Y., Kong, L., Tang, Z., Diab, M. T., Smith, V., and Zhang, K. Position: Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. *arXiv preprint arXiv:2505.20254*, 2025. doi: 10.48550/arXiv.2505.20254. URL https://arxiv.org/abs/2505.20254.

Sporns, O. and Kötter, R. Motifs in brain networks. *PLoS Biology*, 2(11):e369, 2004. doi: 10.1371/journal.pbio.0020369.

Suweis, S., Tu, C., Rocha, R. P., Zampieri, S., Zorzi, M., and Corbetta, M. Brain controllability: Not a slam dunk yet. *NeuroImage*, 200:552–555, 2019. doi: 10.1016/j.neuroimage.2019.07.012.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jermyn, A., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

van den Heuvel, M. P. and Sporns, O. Network hubs in the human brain. *Trends in Cognitive Sciences*, 17(12):683–696, 2013. doi: 10.1016/j.tics.2013.09.012.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. doi: 10.1038/30918.

Williams, I., Oldenburg, N., Dhar, R., Hatherley, J., Fierro, C., Rajcic, N., Schiller, S. R., Stamatiou, F., and Søgaard, A. Mechanistic interpretability needs philosophy. *arXiv preprint arXiv:2506.18852*, 2025. doi: 10.48550/arXiv.2506.18852. URL https://arxiv.org/abs/2506.18852.

Zheng, Y., Yuan, Y., Zhuo, Y., Li, Y., Kreiman, G., Poggio, T., and Santi, P. Probing neural topology of large language models. *arXiv preprint arXiv:2506.01042*, 2025. URL https://arxiv.org/abs/2506.01042.