Published as
Chung, M.K., Maccotta, L., Struck, A. 2026, IEEE International Symposium on Biomedical Imaging (ISBI), pages 1-5

COUNTERFACTUAL ANALYSIS OF BRAIN NETWORK DYNAMICS

Moo K. Chung[1] *Luigi Maccotta*[2], *Aaron Struck*[2]

[1] University of Wisconsin, Madison, USA
[2]Washington University, St. Louis, USA

ABSTRACT

Causal inference in brain networks has traditionally relied on regression-based models such as Granger causality, structural equation modeling, and dynamic causal modeling. While effective for identifying directed associations, these methods remain descriptive and acyclic, leaving open the fundamental question of intervention: *what would the causal organization become if a pathway were disrupted or externally modulated?* We introduce a unified framework for *counterfactual causal analysis* that models both pathological disruptions and therapeutic interventions as an *energy-perturbation problem* on network flows. Grounded in Hodge theory, directed communication is decomposed into dissipative and persistent (harmonic) components, enabling systematic analysis of how causal organization reconfigures under hypothetical perturbations. This formulation provides a principled foundation for quantifying network resilience, compensation, and control in complex brain systems.

1. INTRODUCTION

Understanding causality in brain networks is one of the central challenges in computational neuroimaging. Traditional approaches—including Granger causality, structural equation modeling (SEM), and dynamic causal modeling (DCM)—have provided powerful tools for testing directed influences between brain regions [1]. However, these methods remain fundamentally associational: they infer directed dependencies from observational data but do not provide a principled way to reason about *interventions*. Consequently, the mechanistic question—*how would the brain's causal architecture change if a specific pathway or feedback loop were disrupted?*—remains unresolved.

Counterfactual analysis provides a principled framework for answering "what if" questions that go beyond assocations to model the consequences of hypothetical interventions. It has emerged as a central theme across modern data science, from epidemiology [2] to machine learning [3], where interventions are often infeasible or unethical to observe directly. In neuroscience, counterfactual reasoning is increasingly recognized as critical for bridging observational neuroimaging data with mechanistic insight [4, 3]. By explicitly simulating interventions—even when no experimental manipulation is possible— counterfactual analysis moves beyond descriptive connectivity toward predictive models of network stability, compensation, and control.

In this study, we develop a formal framework for counterfactual causal analysis grounded in *Hodge theory* [5] and the *minimum-energy principle*. We represent directed functional interactions as energy-carrying edge flows and model perturbations as controlled changes in the network's energy landscape. This allows us to test how causal organization would reconfigure following structural damage, neuromodulation, or adaptive reorganization—even when such interventions cannot be performed in vivo. We demonstrate this framework through applications to temporal lobe epilepsy (TLE), comparing pathological recurrence with therapeutic disconnection.

2. METHODS

2.1. Data and preprocessing

We analyzed resting-state fMRI (rs-fMRI) data from 400 participants of the Human Connectome Project (HCP) [6]. Each rs-fMRI dataset was acquired with 2 mm isotropic spatial resolution and consisted of 1200 time points. Preprocessing followed the HCP minimal preprocessing pipelines [7], including motion correction, spatial normalization, and artifact removal. Time frames with framewise displacement exceeding 0.5 mm, along with their adjacent volumes, were excluded from further analysis [6]. Subjects exhibiting excessive head motion were also removed. The preprocessed images were then parcellated using the Automated Anatomical Labeling (AAL) atlas into 116 non-overlapping anatomical regions, and voxel-wise signals within each region were averaged to obtain regional time series. Additional details of the imaging and preprocessing procedures can be found in our previous study [8].

2.2. Spatial Scaffold

We constructed simplicial complexes from the AAL parcellation to capture not only pairwise but also higher-order motifs such as loops. AAL parcels served as *nodes*, with mean fMRI signals providing one representative time series per region.

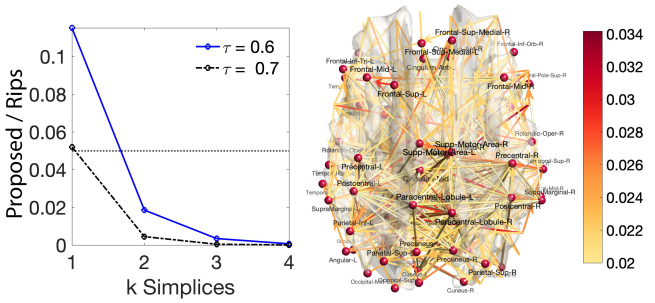

Fig. 1. **Left:** Average fraction of simplices in the Spatial Scaffold relative to the Rips complex (theoretical upper bound) across subjects in rs-fMRI connectivity. **Right:** Average time-lagged correlation across all time points and subjects. The mean correlation is extremely low (maximum $r = 0.034$), indicating that none of the connections reach statistical significance.

Dynamic edge flows were estimated from resting-state fMRI using time-lagged Pearson correlations within 20-second sliding windows, corresponding to the minimum hemodynamic response scale [9]. This yielded a directed, time-varying matrix $X(t) = (X_{ij}(t))$, where $X_{ij}(t)$ denotes the influence of region i on j. Because of intersubject and temporal asynchrony, averaging $X(t)$ cancels directional effects, producing extremely weak mean connectivity (maximum $r = 0.034$; Fig. 1, right). To obtain a statistically robust directed flow, we used 2-simplices (triangles) to encode higher-order interactions. Edges with weights below $\tau = 0.6$ were removed, and directed triangles were included only when all three constituent edges exceeded τ, ensuring hierarchical nesting and directional consistency across scales.

To manipulate simplicial complexes algebraically, we used *boundary (incidence) matrices*, which encode how simplices assemble across dimensions: $\mathbf{B}_1 \in \mathbb{R}^{|V|\times|E|}$ for node–edge and $\mathbf{B}_2 \in \mathbb{R}^{|E|\times|F|}$ for edge–face incidence [10, 11]. These matrices provide a compact algebraic representation of higher-order interactions. Complexes and their corresponding boundary matrices were constructed hierarchically using the Intersecting Neighbor Sets algorithm [12], enabling scalable computation.

Unlike the commonly used Rips complex [10] in topological data analysis (TDA), which enumerates all possible k-simplices and rapidly becomes intractable due to combinatorial explosion [13], our approach constructs a sparse, data-driven complex that retains only statistically significant higher-order interactions. The resulting Spatial Scaffold contains fewer than 0.02% of the simplices generated by a full Rips complex (Fig. 1-left), providing a scalable and biologically meaningful topological representation.

2.3. Dirichlet potential energy of network

We construct a causal framework based on the Dirichlet energy of time-varying edge flows $X(t)$. The 1-Hodge Laplacian [5].

$$\mathcal{L}_1 = \mathbf{B}_1^\top \mathbf{B}_1 + \mathbf{B}_2 \mathbf{B}_2^\top$$

acts on edges, with $\mathbf{B}_1^\top \mathbf{B}_1$ capturing divergence (source–sink activity) and $\mathbf{B}_2 \mathbf{B}_2^\top$ capturing rotationality (cyclic circulation around triangles). For comparison, the traditional graph Laplacian, $\mathcal{L}_0 = \mathbf{B}_1 \mathbf{B}_1^\top$, actes on vertices and captured only acyclic pairwise interactions. The associated Dirichlet potential energy is defined as

$$E(X) = \tfrac{1}{2}\langle X, \mathcal{L}_1 X\rangle = \tfrac{1}{2}\|\mathbf{B}_1 X\|_2^2 + \tfrac{1}{2}\|\mathbf{B}_2^\top X\|_2^2,$$

which jointly measured divergence and cyclic circulation. This defines the energy landscape on which causal flows are modeled.

An edge flow X can be interpreted as a fluid-like transport of information, with causality corresponding to the system's intrinsic tendency to evolve toward minimum-energy configurations (Fig. 2). Minimization of $E(X)$ therefore provided a variational principle for causal propagation. The gradient flow of the potential energy yields Dirichlet diffusion through Onsager's principle [14]:

$$\frac{dX(t)}{dt} = -\nabla E(X(t)) = -\mathcal{L}_1 X(t). \tag{1}$$

Along this trajectory, the energy decreases monotonically in steepest descent:

$$\frac{d}{dt}E(X(t)) = \left\langle \frac{dX}{dt}, \nabla E(X(t)) \right\rangle = -\|\mathcal{L}_1 X(t)\|^2 \leq 0,$$

ensuring that flows converges toward steady state X_H.

Using the spectral decomposition of the 1-Hodge Laplacian [15] $\mathcal{L}_1 \phi_k = \lambda_k \phi_k$ with $\lambda_k \geq 0$ [5], edge flow $X(t)$ admits solution

$$X(t) = e^{-\mathcal{L}_1 t} X(0) = \sum_k e^{-\lambda_k t} \alpha_k \phi_k, X(0) = \sum_k \alpha_k \phi_k.$$

As $t \to \infty$, all modes with $\lambda_k > 0$ vanish while those with $\lambda_k = 0$ remain:

$$X_H = \mathcal{P}_H X(0), \quad \mathcal{P}_H = \sum_{\lambda_k = 0} \phi_k \phi_k^\top.$$

X_H thus captures the persistent, non-dissipative backbone of information flow—the portion of network dynamic that remains stable after transient dynamics decay (Fig. 2).

2.4. Counterfactual analysis on harmonic flow

Counterfactual analysis [2, 3, 4] provides a principled framework to probe how causal organization reconfigures under

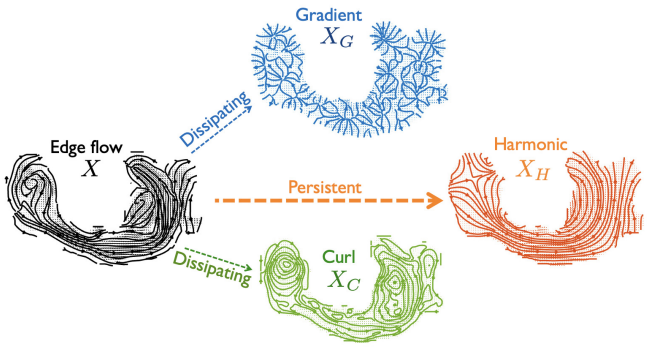

Fig. 2. Minimization of the Dirichlet potential energy of edge flow X converges to the persistent harmonic flow X_H, which captures long-range, stable, and recurrent communication patterns. The residual $X - X_H$, composed of the gradient X_G and curl X_C, represents transient dynamics that dissipate over time.

hypothetical interventions or disruptions that cannot be performed in vivo. Such interventions can be simulated at multiple scales: at the edge level by perturbing individual connections; at the motif level by damping or removing recurrent cycles to assess their stability; and at the system level by evaluating changes in the total Dirichlet energy.

We formulate a counterfactual operator $\mathcal{C}$ to act linearly on the edge flow X, producing a modified flow $X^{(\mathcal{C})} = \mathcal{C}X$ that scales, deletes, or alters selected connections—for example, by setting an edge weight to zero, damping a cycle by a factor $0 \leq \alpha \leq 1$, or attenuating all edges incident to a hub node. Comparing $X^{(\mathcal{C})}$ with the baseline X reveals how information flow reorganizes across the network and quantifies its resilience or restructing.

The corresponding change in Dirichlet energy provides a global measure of network stability:

$$\Delta E = E(X^{(\mathcal{C})}) - E(X) = \tfrac{1}{2}\big(X^{(\mathcal{C})\top}\mathcal{L}_1 X^{(\mathcal{C})} - X^\top \mathcal{L}_1 X\big).$$

Large positive ΔE indicates increased dissipation and fragility, whereas small or near-zero ΔE implies that causal flow is rerouted along energetically equivalent pathways, reflecting robust topological compensation.

For harmonic flows X_H, which satisfy $\mathcal{L}_1 X_H = 0$, the Dirichlet energy is zero ($E(X_H) = 0$), representing non-dissipative, self-sustaining cycles of information flow. Counterfactual analysis of X_H isolates the brain's most stable communication backbone— the part of causal organization that persists even when transient or dissipative components (gradient and curl flows) are removed. By perturbing X_H, we can test how these long-range recurrent pathways reorganize under hypothetical disruptions, providing a direct measure of the system's intrinsic resilience and compensatory capacity.

When acted on by a counterfactual operator $\mathcal{C}$, the resulting flow

$$X_H{}^{(\mathcal{C})} = \mathcal{C}(\mathcal{P}_H X) = \mathcal{P}_H(\mathcal{C}X)$$

remains within the harmonic subspace and preserves its zero-energy property ($\Delta E = 0$). Hence, harmonic counterfactuals reorganize how information circulates without changing the global energy balance—yielding energetically conserved but topologically distinct configurations. These invariant harmonic flows form the persistent, self-correcting backbone of brain communication and provide a principled foundation for quantifying redundancy, compensation, and resilience in functional networks.

3. APPLICATIONS: TEMPORAL LOBE EPILEPSY

We applied the proposed framework to model counterfactual disruptions and interventions in temporal lobe epilepsy (TLE) —a disorder characterized by recurrent synchronization and impaired large-scale stability [16]. Our analysis examined how disease-like perturbations reshape persistent causal pathways (harmonic flow, X_H) while preserving overall energy.

3.1. Pathological Disruption (Disease Model)

TLE involves excessive synchronization within mesial temporal and limbic structures, including the hippocampus, amygdala, and orbitofrontal cortex [17]. We simulated this pathology using a harmonic counterfactual that amplified recurrent flows within these regions by 30%, enhancing temporal→limbic drive consistent with patient evidence of hyper connectivity [18].

Although the total Dirichlet energy remained unchanged ($\Delta E = 0$), the topology of persistent flows reorganized substantially (Table 1, Fig. 3). In the healthy baseline, dominant harmonic flows were primarily interhemispheric sensory loops (Calcarine, Cuneus, Postcentral) and vermian–rectal connections. After perturbation, these were supplanted by stronger limbic and subcortical pathways, notably Olfactory→Amygdala, Rectus→Amygdala, and Amygdala→Olfactory, reflecting hyperactive feedback within temporal–limbic hubs.

This shift from symmetric sensory exchange to asymmetric limbic–subcortical recurrence mirrors electrophysiological and neuroimaging findings in TLE, where seizures propagate through hippocampal–amygdalar circuits and engage subcortical regions including the cerebellum and pallidum [18, 19].

3.2. Invasive Therapeutic Intervention

We examined the effects of left *anterior temporal lobectomy (ATL)*—a surgical treatment for drug-resistant temporal lobe epilepsy (TLE) [20, 21]. ATL often removes the anterior temporal neocortex and mesial temporal structures, including the hippocampus and amygdala, thereby disrupting the recurrent temporal–limbic feedback loops that sustain seizure propagation. Since such resection cannot be performed in healthy par-

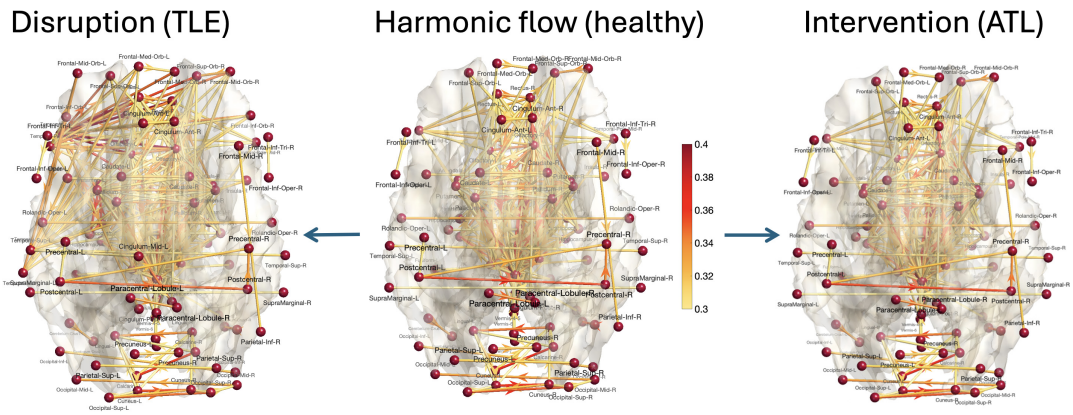

Fig. 3. **Middle:** Ten dominant average harmonic flows (X_H) across time and 400 healthy subjects, showing strong recurrent interhemispheric and predominantly homotopic organization—most evident in visual (Calcarine, Cuneus, Lingual) and sensorimotor (Postcentral, Paracentral) regions. **Left:** After counterfactual perturbation of the temporal–limbic feedback loop, pathological hyperrecurrence shifts dominance from interhemispheric sensory pathways toward olfactory–limbic–vermal circuits. **Right:** After counterfactual simulation of left anterior temporal lobectomy (ATL) on healthy controls, the global topology remains largely preserved, indicating compensatory stabilization through contralateral and midline pathways.

Table 1. Top 10 dominant harmonic flows before and after counterfactual amplification of the temporal–limbic feedback loop. Pathological recurrence strengthens limbic and subcortical interactions (e.g., Olf→Amyg, Rect→Amyg) while reducing the dominance of interhemispheric sensory loops. In contrast, simulated anterior temporal lobectomy (ATL) in healthy controls yields minimal reorganization, reflecting preserved global topology through contralateral and midline compensation.

Baseline (Healthy)				**Disease Model (TLE)**			
From	→	To	Mag.	From	→	To	Mag.
Vrm1–2	→	Vrm10	0.38	Olf-R	→	Amyg-R	0.38
Olf-R	→	Vrm10	0.38	Vrm1–2	→	Vrm10	0.38
Olf-R	→	Vrm1–2	0.36	Olf-R	→	Vrm10	0.38
Calc-L	→	Calc-R	0.36	Rect-R	→	Amyg-R	0.37
Vrm10	→	Rect-R	0.36	Amyg-L	→	Olf-R	0.36
Cune-L	→	Cune-R	0.36	Olf-R	→	Vrm1–2	0.36
Olf-R	→	Rect-R	0.35	Vrm10	→	Rect-R	0.36
Post-L	→	Post-R	0.35	Calc-L	→	Calc-R	0.36
Rect-R	→	Vrm1–2	0.35	Cune-L	→	Cune-R	0.36
Vrm10	→	Pall-L	0.35	Olf-R	→	Rect-R	0.35

Abbreviations: Olf = Olfactory, Amyg = Amygdala, Calc = Calcarine, Cune = Cuneus, Post = Postcentral, Rect = Rectus, Pall = Pallidum, Vrm = Vermis

ticipants, we simulated its functional impact by attenuating left temporal–limbic and hippocampal–amygdalar flows by 70%, approximating the disconnection produced by surgery.

In healthy controls, applying the left-ATL counterfactual left this top 10 harmonic flows largely unchanged (Table 1, Fig. 3-right), indicating that persistent harmonic organization in the normative brain is supported by contralateral (right) limbic and midline vermis circuits that maintain energetic balance despite unilateral disruption. This pattern reflects hemispheric redundancy and compensatory topology typical of healthy networks, where removal of one temporal pole minimally perturbs global harmonic equilibrium.

By contrast, in the TLE disease model, the same perturbation induces a major redistribution toward limbic and subcortical dominance, demonstrating reduced compensatory capacity in pathological networks.

4. CONCLUSION & DISCUSSION

We proposed a unified framework for *counterfactual causal analysis* that models pathological disruptions and therapeutic interventions as energy perturbations on network flows. Within the Hodge framework, the harmonic flow X_H emerged as the non-dissipative backbone of brain communication, allowing perturbations to reveal how stable recurrent pathways reorganize under disruption.

Applied to temporal lobe epilepsy (TLE), pathological recurrence amplified temporal–limbic and subcortical feedback loops, whereas simulated anterior temporal lobectomy (ATL) in healthy controls produced minimal change—reflecting contralateral and midline compensation. Harmonic counterfactuals thus provide a principled tool to quantify network resilience and redundancy, offering a general framework for modeling virtual lesions, stimulation, and adaptive reorganization in brain networks.

Acknowledgment. This study is funded by NIH MH133614 and NSF DMS-2010778.

5. REFERENCES

[1] K.J. Friston, A.M. Bastos, A. Oswal, B. Van Wijk, C. Richter, and V. Litvak, "Granger causality revisited," *NeuroImage*, vol. 101, pp. 796–808, 2014.

[2] T.J. VanderWeele, "Invited commentary: counterfactuals in social epidemiology—thinking outside of "the box"," *American Journal of Epidemiology*, vol. 189, pp. 175–178, 2020.

[3] N.J. Dhinagar, S.I. Thomopoulos, E. Laltoo, and P.M. Thompson, "Counterfactual MRI generation with denoising diffusion models for interpretable alzheimer's disease effect detection," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–6.

[4] T. Matsui, M. Taki, T.Q. Pham, J. Chikazoe, and K. Jimura, "Counterfactual explanation of brain activity classifiers using image-to-image transfer by generative adversarial network," *Frontiers in Neuroinformatics*, vol. 15, pp. 802938, 2022.

[5] D.V. Anand and M.K. Chung, "Hodge-Laplacian of brain networks," *IEEE Transactions on Medical Imaging*, vol. 42, pp. 1563–1473, 2023.

[6] D.C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T.E.J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, and S.W. Curtiss, "The human connectome project: a data acquisition perspective," *NeuroImage*, vol. 62, pp. 2222–2231, 2012.

[7] M.F. Glasser, S.M. Smith, D.S. Marcus, J.L.R. Andersson, E.J. Auerbach, T.E.J. Behrens, T.S. Coalson, M.P. Harms, M. Jenkinson, and S. Moeller, "The human connectome project's neuroimaging approach," *Nature Neuroscience*, vol. 19, pp. 1175, 2016.

[8] S.-G. Huang, S.-T. Samdin, C.M. Ting, H. Ombao, and M.K. Chung, "Statistical model for dynamically-changing correlation matrices with application to brain connectivity," *Journal of Neuroscience Methods*, vol. 331, pp. 108480, 2020.

[9] S. Keilholz, C. Caballero-Gaudes, P. Bandettini, G. Deco, and V. Calhoun, "Time-resolved resting-state functional magnetic resonance imaging analysis: current status, challenges, and new directions," *Brain Connectivity*, vol. 7, pp. 465–481, 2017.

[10] H. Edelsbrunner and J. Harer, *Computational topology: An introduction*, American Mathematical Society, 2010.

[11] J. Huang, M.K. Chung, and A. Qiu, "Heterogeneous graph convolutional neural network via hodge-laplacian for brain functional data," in *International Conference on Information Processing in Medical Imaging*. Springer, 2023, pp. 278–290.

[12] T. Schank and D. Wagner, "Finding, counting and listing all triangles in large graphs, an experimental study," in *International workshop on experimental and efficient algorithms, LNCS*. Springer, 2005, vol. 3503, pp. 606–609.

[13] A. Choudhary, M. Kerber, and S. Raghvendra, "Improved approximate Rips filtrations with shifted integer lattices and cubical complexes," *Journal of Applied and Computational Topology*, vol. 5, pp. 425–458, 2021.

[14] M. Noirhomme, E. Opsomer, and N. Vandewalle, "Onsager variational principle for granular fluids," *Physical Review E*, vol. 110, pp. 054901, 2024.

[15] Z. Su, Y. Tong, and G.-W. Wei, "Hodge decomposition of single-cell RNA velocity," *Journal of Chemical Information and Modeling*, vol. 64, pp. 3558–3568, 2024.

[16] M.K. Chung, C.G. Ramos, F.B. De Paiva, J. Mathis, V. Prabhakaran, V.A. Nair, M.E. Meyerand, B.P. Hermann, J.R. Binder, and A.F. Struck, "Unified topological inference for brain networks in temporal lobe epilepsy using the Wasserstein distance," *NeuroImage*, vol. 284, pp. 120436, 2023.

[17] F. Bartolomei, S. Lagarde, F. Wendling, A. McGonigal, V. Jirsa, M. Guye, and C. Bénar, "Defining epileptogenic networks: contribution of SEEG and signal analysis," *Epilepsia*, vol. 58, pp. 1131–1147, 2017.

[18] H.F.J. González, S. Chakravorti, S.E. Goodale, K. Gupta, D.O. Claassen, B. Dawant, V.L. Morgan, and D.J. Englot, "Thalamic arousal network disturbances in temporal lobe epilepsy and improvement after surgery," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 90, pp. 1109–1116, 2019.

[19] H.J. Jo, D.L. Kenny-Jung, I. Balzekas, E.E. Benarroch, D.T. Jones, B.H. Brinkmann, S.M. Stead, J.J. Van Gompel, K.M. Welker, and G.A. Worrell, "Nuclei-specific thalamic connectivity predicts seizure frequency in drug-resistant medial temporal lobe epilepsy," *NeuroImage: Clinical*, vol. 21, pp. 101671, 2019.

[20] D.W. Kim, S.K. Lee, K.-Y. Jung, K. Chu, and C.K. Chung, "Surgical treatment of nonlesional temporal lobe epilepsy," *Seizure*, vol. 86, pp. 129–134, 2021.

[21] L.E. Sainburg and V.L. Morgan, "Investigation of network reorganization after epilepsy surgery is worth the effort," *Brain*, vol. 147, pp. 2261–2263, 2024.