Universiteit Leiden
The Netherlands

Topological explanations of deep neural networks

Rabiza, M.

Citation
Rabiza, M. (2026). Topological explanations of deep neural networks. *Synthese*, *208*. doi:10.1007/s11229-026-05646-2

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/4308272

Note: To cite this publication please use the final published version (if applicable).

Synthese (2026) 208:54
https://doi.org/10.1007/s11229-026-05646-2

ORIGINAL RESEARCH

Topological explanations of deep neural networks

Marcin Rabiza[1,2]

Received: 17 March 2025 / Accepted: 9 May 2026
© The Author(s) 2026

Abstract
Recent philosophical work on explainable AI has explored mechanistic methods targeting the causal structure of deep neural networks. However, the ever-growing complexity of deep learning models challenges purely decompositional strategies. This paper advances an alternative perspective based on topological explanations—an approach that both complements and, in some respects, surpasses the mechanistic perspective. By examining how topological properties of deep neural networks relate to model behavior in a non-causal, abstract, and counterfactual manner, I argue that this framework is particularly well suited to addressing why-questions in AI contexts. Drawing on cases of model-driven and data-driven explanations, I show that the topological explanatory strategy offers distinctive epistemic advantages for explainability research.

Keywords Topological explanation · Explainable artificial intelligence · Deep neural networks · Topological data analysis · Topological deep learning

1 Introduction

In recent years, *deep learning* has become the dominant approach in *artificial intelligence* (AI). A persistent challenge, however, is the *black box problem*, which stems from the opacity of *deep neural networks* (DNNs) containing millions or even billions of parameters learned through automated training. Despite their predictive success, the complexity of these systems renders their internal workings inscrutable.

Researchers have focused on *explainable AI* (XAI), seeking ways to validate opaque AI systems and make their behavior understandable to stakeholders (Adadi & Berrada, 2018). Despite significant technical advancements (e.g., Gilpin et al., 2018;

✉ Marcin Rabiza
marcin.rabiza@gssr.edu.pl

1 Institute for Philosophy, Leiden University, Leiden, The Netherlands

2 Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, Poland

Published online: 10 July 2026

Linardatos et al., 2020; Montavon et al., 2018), the field still lacks solid conceptual foundations and integration with established accounts of scientific explanation (Erasmus et al., 2021). Many current XAI methodologies excel at generating approximate local explanations but fail to provide a holistic understanding of complex AI systems (Mittelstadt et al., 2019). Thus far, AI explanations have predominantly followed a technology-focused approach, often disregarding established theories and insights from philosophy (Miller, 2019).

Recent philosophical work has proposed interpreting XAI through a *mechanistic* lens (Erasmus et al., 2021; Kästner & Crook, 2024; Millière & Buckner, 2025). However, given the epistemic limitations and intrinsic complexity of DNNs, purely mechanistic frameworks are unlikely to address all relevant forms of opacity (cf. Rabiza, 2026). This paper therefore advances an alternative perspective grounded in topological considerations, which can complement and, in some respects, surpass the explanatory reach of mechanistic approaches.

Specifically, I examine recent advances in explainable deep learning that employ graph-theoretic and network-analytic methods, arguing that the resulting accounts can qualify as *topological explanations* in the sense articulated by Kostić (2018a, 2022). To demonstrate the value of this framework, I analyze candidate *model-driven* and *data-driven* topological explanations of DNNs. The latter are especially important, as *data explainability*—understood as clarifying the structure, distribution, and salient features of datasets—has received limited attention despite its substantial impact on model opacity and performance.

Although topological techniques are widely used in deep learning, their explanatory value has been largely overlooked by XAI theorists. Topological explanations appeal to high-level structural features that, while more abstract than many alternative explanatory strategies, are linked to performance outcomes through counterfactual dependence. This abstraction confers notable epistemic advantages. It reduces the number of variables that must be tracked, supports generalization across diverse cases, and avoids reliance on system-specific parameters. By foregrounding structural relations, topological explanations facilitate the formulation of general principles governing model behavior across contexts. More broadly, they offer a distinctive, non-causal explanatory framework with relevance beyond XAI, thus holding significance for philosophers of science.

The structure of the paper is as follows: Sect. 2 provides a minimal account of topological explanation, clarifying how mathematical properties describe the operations of complex, interconnected systems. Section 3 conceptualizes the explainability of DNNs through a topological lens, discussing candidate data-driven and model-driven explanations. Section 4 addresses potential objections. Section 5 examines the epistemic significance of topological explanations for the XAI research agenda. Section 6 concludes the paper.

2 A minimal account of topological explanation

Topological explanations, also known as network explanations, constitute a form of mathematical and structural explanation that focuses on topological properties of complex systems. By quantifying networks using graph theory and other formal methods, they aim to reveal how patterns of connectivity in highly interconnected systems shape global behavior and system dynamics. Their emergence is closely tied to the rise of network science across multiple disciplines, in which the analysis of structural features has become central to understanding complex phenomena.

Topology, in this context, refers both to structural properties of networked systems related to their connectivity and to features that remain invariant under bicontinuous transformations. In the sense commonly used in network science, topology concerns connectivity patterns in a graph independently of spatial embedding. Graph properties are described as "topological" insofar as they capture invariant aspects of relational organization. In the strict mathematical sense, topology studies *topological spaces*—sets equipped with collections of open subsets satisfying specific axioms—which ground the notions of continuity and homeomorphism. Two spaces related by a homeomorphism are topologically equivalent: although they may differ geometrically, they share the same qualitative structure. Accordingly, this paper treats topological properties as those that remain unchanged under transformations that preserve the relevant notion of structure, whether in continuous topological spaces or discrete network structures.

Topology provides a rigorous framework for characterizing both discrete graph structures and high-dimensional data spaces without reducing them to geometry or physical layout. Drawing on graph-theoretical, network-scientific, and algebraic methods, topological analyses can uncover deeper insights into the underlying architecture and internal dynamics of highly interconnected systems (Carlsson, 2009). For example, they can identify hub nodes, which possess disproportionately many connections, as well as motifs, understood as recurring connectivity patterns. Among graph-theoretical measures commonly used in network analysis are average path length, defined as the mean number of edges required to connect pairs of nodes, and clustering coefficients, which quantify the tendency of nodes to form tightly connected groups. Beyond connectivity patterns, topological properties may also concern shape invariants such as the number of holes, compactness, orientability, separability, or dimensionality.

One of the most elaborate explications of the theory of topological explanation has been provided by Kostić (2022). According to a minimal version of his framework, an explanation of a physical fact is a function of the system topology insofar as it reveals counterfactual dependencies between topological properties and empirical characteristics.[1] Kostić proposes the following counterfactual explanatory schema (the *T-schema*) for analyzing topological explanations:

[1] Kostić (2018b) characterizes this relation in terms of *topological realization*: a system S realizes a topology T when the elements of S are interconnected in ways that instantiate the characteristic pattern of connectivity of T. On this view, topological explanations ultimately rest upon properties of the realized topological structure itself.

a's being *F* topologically explains why *b* is *G* (*a* and *b* are often identical) if and only if:

(T1) *a* is *F* (where *F* is a topological property);
(T2) *b* is *G* (where *G* is an empirical property);
(T3) Had *a* been *F'* (rather than *F*), then *b* would have been *G'* (rather than *G*);
(T4) That *a* is *F* is an answer to the question why is *b G* (Kostić, 2022, p. 97).

Each condition of the T-schema serves a specific purpose. Condition T1 distinguishes the properties considered explanatory, thereby determining whether an explanation is topological or of some other kind. T2 ensures that *G* is a valid scientific explanandum, representing a description of an empirical phenomenon. T3 secures explanatoriness by capturing counterfactual dependency relations. Finally, T4 provides contextual criteria for using the counterfactual via one of two explanatory modes: horizontal and vertical. The *horizontal mode* (T4H) relates topological and empirical properties at the same level, while the *vertical mode* (T4V) connects them across a range of levels.

The distinction between the vertical and horizontal explanatory modes is exemplified by local and global properties of a network. Local properties refer to specific parts of a network, while global properties pertain to the characteristics of the whole network. For instance, the *node degree*, which measures the number of edges connected to a single node, represents a local topological property. Conversely, the *average path length* serves as an example of a global property. Understanding these distinctions is important, as it allows us to comprehend how various levels of a system are functionally interrelated and the extent to which external changes affect these levels. Vertical and horizontal explanatory modes can yield complementary multilevel explanations, offering insights into the behavior of complex systems.

Explanations conforming to the T-schema are, in general, non-causal. Although one might say that topological properties "cause" outcomes in a counterfactual sense, equating such modal dependence with productive causation would be misleading. As Huneman (2010) emphasizes, claiming that the topology of food webs causes particular resilience properties differs fundamentally from tracing the propagation of concrete perturbations in species interactions through a system to restore community structure. While system topology can sometimes shape causal capacities, many proponents of topological explanation maintain that there are cases in which topological properties provide explanatory leverage independent of causal considerations (Darrason, 2018; Huneman, 2010, 2018a, b; Kostić, 2018a, b, 2022; Kostić & Khalifa, 2023; Ross, 2021).

If an explanation is non-causal, the required dependency must be secured by other means. Identifying a topological feature *F* and its correlation with an outcome *G* may satisfy conditions T1 and T2 of Kostić's schema, but it is insufficient for T3, which requires showing that *G* counterfactually depends on *F*. Even strong, stable correlations can be nonexplanatory if *F* does not make a difference to *G*, as illustrated by Salmon's well-known barometer–storm example. Satisfying T3 therefore requires showing that a variation in *F* would change *G*, ideally through intervention (but cf. Mikhalevich, 2025). Where this is not possible, explanatoriness must be supported

non-interventionally through convergent empirical evidence, simulation, plausible functional or causal reasoning, or arguments from mathematical necessity that non-arbitrarily constrain the outcome while supporting relevant counterfactuals.

When discussing novel forms of scientific explanation, philosophers of science often position them within established scientific disciplines such as physics or molecular biology. Although graph theory is not a distinct scientific discipline with unique empirical content, it gives rise to a distinctive form of explanation that transcends traditional disciplinary boundaries. Many methodologists maintain that certain mathematical representations can yield epistemically distinctive explanations not confined to any single empirical domain (Batterman & Rice, 2014; Humphreys, 2014; Ladyman et al., 2007; Lange, 2009; Rathkopf, 2018; Rice, 2021; but cf.; Kuorikoski, 2021). Topological explanations are abstract enough to enable explanatory inferences that are not confined to a particular discipline, offering a domain-general, representation-centric, non-causal approach to explanation.

Traditionally, scientific fields in which topological explanations might be applied have been dominated by other approaches, such as mechanistic explanations in neuroscience (Kostić & Halffman, 2023). However, the growing significance of network modeling in fields like social science, cell biology, molecular genomics, and neuroscience highlights the potential of topological explanatory practices. As interest in network methods rises across scientific domains, philosophers of science are also recognizing the topological approach as a distinctive explanatory strategy. Noteworthy contributors include Bechtel (2020), Craver (2016), Darrason (2018), Green et al. (2018), Huneman (2010, 2018a, b), Jones (2014), Kostić (2018a, 2019, 2022), Kostić and Khalifa (2021, 2023), Kostić et al. (2020), Lacková and Zámečník (2020), and Ross (2021).

Since the proliferation of topological explanation, the philosophical literature has extensively explored numerous instances of topological reasoning across scientific domains. The relevance of this inquiry extends to deep learning and XAI. Given the mathematical and structural foundations of DNNs, treating their topology as explanatorily salient is not only appropriate but arguably more fitting than approaches modeled on traditional neuroscience, which have long been dominated by mechanistic and causal paradigms. The complexity of DNNs, together with the widespread use of topological methods in AI research and data engineering, further strengthens the case for adopting the topological explanatory strategy for XAI explicated in the subsequent sections.

3 Topological interpretation of deep learning

3.1 Topological constraints on class separability

Deep learning systems are based on *artificial neural networks* composed of nodes and links that mimic the behavior of neurons and synapses at various levels of abstraction. Input signals are processed through layers of nodes, each computing its activation function. Signals propagate via weighted links, culminating in an output layer that produces the network's decision. In graph representations of such models,

vertices represent neuron-like nodes, while edges represent the connections between them.

Given the complex structure of DNNs and their inspiration from biological brains with scale-free and small-world properties, understanding their behavior may require attention to abstract, system-level properties of complexity. Causal explainability paradigms, such as *mechanistic interpretability* (e.g., Bricken et al., 2023; Cammarata et al., 2021; Casper, 2023; Nanda, 2023), aim to decompose networks into causally or functionally relevant components, but may prove inadequate in certain opacity contexts. In particular, they may fail to scale with the increasing complexity of contemporary models, as large DNNs frequently behave as *non-decomposable systems* in which each component's behavior depends on numerous interactions, leading to a combinatorial explosion (cf. Rathkopf, 2018). Evidence from toy models does not guarantee that large-scale DNNs can always be effectively decomposed in a way that yields holistic mechanistic understanding.

When part–whole decomposition becomes infeasible, alternative strategies are needed to address the complexity of highly interconnected DNNs operating on large, high-dimensional datasets. Topological explanations are well suited to this role: rather than being undermined by complexity, they exploit it by focusing on abstract, global properties of networks and data. This approach is timely given the growing interest among AI researchers in the relation of topology to performance. Additionally, topological methods provide a natural lens on DNN behavior, as neural networks effectively transform data topology by reshaping complex input spaces into more manageable forms across layers (Naitzat et al., 2020). In classification tasks, this process can be understood as the progressive separation of low-dimensional manifolds embedded in high-dimensional data spaces, as suggested by the *manifold hypothesis*.[2]

As a toy example of topological explanation in deep learning, consider Olah's (2014) work, which applies visualization techniques to low-dimensional neural networks in order to explore how network behavior reflects underlying structural properties. He examines a simple neural network with only one hidden layer, showing how each layer transforms data to create new representations. In architectures using smooth, monotonic activation functions such as sigmoid or hyperbolic tangent, these transformations continuously "stretch" and "compress" the input space without introducing topological discontinuities. That is, the network preserves the input space's topological structure: it does not tear, break, or fold it.

Consider a one-dimensional dataset with two classes, as visualized in Fig. 1:

[2] While both graph and manifold representations may offer topological explanations, due to the presence of topological properties in the explanans, they may provide a range of insights into the system's operation. For instance, graph-based representations emphasize the relationships and connections between discrete elements, offering a clearer view of data's movement through the network. In contrast, manifold representations focus on the continuous, smooth structures of data, providing a more geometric understanding of networks' classification and organization of complex inputs. This distinction does not change the fundamental status of topological explanations but introduces pragmatic differences in the interpretation.

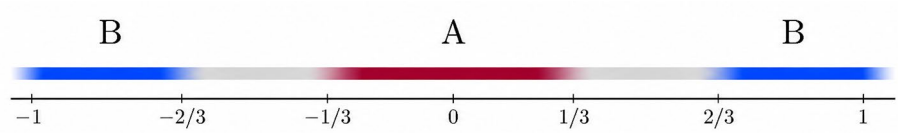

Fig. 1 Visualization of a one-dimensional dataset with two classes (based on Olah, 2014)

$$A = \left[-\frac{1}{3}, \frac{1}{3}\right], B = \left[-1, -\frac{2}{3}\right] \cup \left[\frac{2}{3}, 1\right]$$

In this scenario, a neural network cannot classify a dataset with fewer than two hidden units, regardless of its depth. Classification using a sigmoid unit or a softmax layer requires a decision boundary between classes A and B in the output space to be drawn. With only one hidden unit, the network is topologically incapable of creating this separation, leading to classification failure. However, with two hidden units, the network can adequately represent the data, allowing for the separation of the two classes with a curved boundary. Specifically, one hidden unit activates when $x > -\frac{1}{2}$, and the other activates when $x > \frac{1}{2}$. If the first unit activates but not the second, the data point is classified as belonging to class A.

Applying Kostić's (2022) T-schema, we can obtain a perfectly reasonable topological explanation for the failed classification. Let a denote the network (a DNN classifier), F the topological property of having one hidden unit, b equal to a, and G the empirical event of classification failure:

(T1) The network has one hidden unit.
(T2) The model fails to classify the dataset accurately.
(T3) If the network had two or more hidden units, the model would classify the dataset accurately.
(T4) The number of units in a hidden layer that the network has is an answer to why the classifier fails.

Not only is this counterfactual valid in this specific case, but it also allows us to infer a plausible generalization about this kind of data representation and DNN architecture: n-dimensional datasets cannot be classified without using $n+1$ units in the hidden layer. Following this, for a two-dimensional dataset with classes $A, B \subset \mathbb{R}^2$, visualized in Fig. 2, another counterfactual holds: if the network had three or more hidden units, the model would classify the new dataset accurately.

$$A = \left\{x | d\left(x, 0\right) < \frac{1}{3}\right\}$$

$$B = \left\{x | \frac{2}{3} < d\left(x, 0\right) < 1\right\}$$

Olah shows that the network achieves linear separability only after it learns a higher-dimensional hidden representation with a sufficient number of units. Drawing on an analytical argument, he demonstrates that with too few units the network necessarily

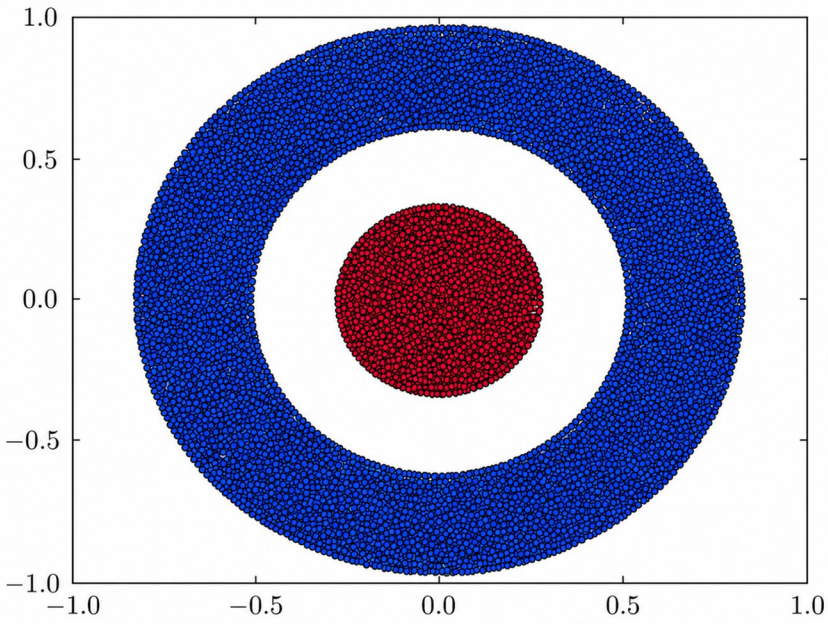

Fig. 2 Visualization of a two-dimensional dataset with two classes (based on Olah, 2014)

must fail: the continuity and low dimensionality imposed by smooth, monotonic activation functions make it topologically impossible to "unwrap" the inner class from the outer one.[3] In other words, the dimensionality of the hidden layer imposes a topological constraint on separability. Although no interventionist test is performed, the case supports a structural and modal counterfactual grounded in *mathematical necessity* (cf. Lange, 2009, 2013, 2017; Woodward, 2003, 2018; Povich, 2021, 2025).[4] Because success or failure is, *ceteris paribus*, modally fixed by the topology of the network, the example provides a simple instance of topological explanation.

Interestingly, this explanation integrates both empirical (or "mechanistic") and topological elements. In this specific case, the network's causal–mechanistic structure is captured directly by its topology. More precisely, the simplicity of the architecture ensures a one-to-one alignment between topological and causal structure: adding or removing hidden units straightforwardly alters the classifier's ability to

[3] For example, in Fig. 2, the smooth and invertible activation function prevents the network from "tearing" the inner circle out of the surrounding ring.

[4] As Lange (2009, 2017) argues, mathematical facts can impose a stronger form of necessity than ordinary causal factors, yielding explanations whose force is derived from mathematical impossibility (*explanation by constraint*; see also *dimensional explanation* in Lange, 2009). This applies to Olah's case, in which the classification outcome is explained by constraint and follows inevitably from the network's architecture and design of activation functions. Woodward (2003, 2018) likewise makes an allowance that non-causal explanations may appeal to generalizations that hold for mathematical or conceptual reasons, and Povich (2021, 2025) account of *distinctively mathematical explanation* similarly maintains that an outcome "had to happen" unless certain mathematical principles were violated.

separate classes in a multidimensional dataset. In this minimal setting, no alternative pathways or architectural configurations could yield the same functional outcome, so the causal organization is transparently mirrored by the network's topology.

Although this direct correspondence is clearest in small, easily inspectable classifiers, the point remains relevant for large, billion-parameter black box models.[5] In such systems, the same topological description may be compatible with multiple causal–mechanistic instantiations, and inspecting individual neurons rarely illuminates overall behavior (and is often infeasible altogether). As a result, the fine-grained causal structure becomes obscured. Even so, a topological link persists: specific causal factors become subsumed under the network's structural organization, which can still be analyzed using graph-theoretical methods.

Topological methods are already widely employed in deep learning, offering readily available tools for analyzing both data structure and model architecture. To assess how these methods align with philosophical accounts of topological explanation, this paper examines selected case studies from the XAI literature while introducing a distinction between data explainability and model explainability. The analysis proceeds across three levels: deep learning models themselves, XAI techniques that exploit topological properties, and the philosophical T-schema used to evaluate explanatoriness and articulate possible topological explanations of DNNs. For clarity and readability, I avoid extensive mathematical formalism and instead present a proof of concept for how a topological explanatory strategy can be articulated in practice.

3.2 Data-driven topological explanations

Data explainability, often neglected in philosophical debates on XAI, involves understanding the nature, content, distribution, and salient properties of datasets used in AI model training and inference. Because training data strongly shapes model behavior, achieving data explainability may be crucial for effective XAI design. This is where *topological data analysis* (TDA) plays a key role, as it enables detailed examination of datasets to assess their structure and suitability for model training. Rooted in the idea that "data has shape, and shape has meaning" (Carlsson, 2009), TDA uncovers intrinsic relationships within datasets by studying their mathematical shapes. It visualizes complex topologies—such as connected components, loops, and voids—in high-dimensional data as human-interpretable graphs, facilitating machine learning algorithm engineering and model optimization (Gabrielsson & Carlsson, 2019; Watanabe & Yamana, 2022).

A central concept in TDA is *persistent homology*, which measures the longevity of topological features as parameters like scale or resolution change, analyzing the evolution of a dataset's topology across a range of scales. By representing high-dimensional data as point clouds, persistent homology identifies significant topological properties, tracking the emergence and dissolution of features like clusters,

[5] For instance, GPT-3 (a large language model and the immediate predecessor of models used in ChatGPT) features 175 billion parameters and operates within 96 layers.

loops, and voids. This multiscale analysis offers a comprehensive understanding of data relationships and reveals critical structural insights that might otherwise remain obscured.[6]

AI data explainability addresses the epistemic opacity of the data used to train deep learning models. Key aspects include understanding encoded information, identifying salient features, and analyzing information distribution. This often involves visualizing relationships within training or output data to enhance explainability by identifying influential topological features. Here, the focus shifts from the neural network's internal architecture to *data-driven explanations*, emphasizing the structure and distribution of data to uncover hidden patterns that inform the model's decision-making process.

One example of leveraging graph-theoretic concepts for explanatory purposes involves modeling input data as graphs using *graph neural networks* (GNNs), which are well suited to capturing both graph structure and node-level features through recursive aggregation of information from neighboring nodes. Interpreting GNN predictions typically involves identifying subgraphs that are most responsible for particular outputs. In this vein, Ying et al. (2019) introduce GNNExplainer, a model-agnostic tool that generates local explanations for individual GNN predictions by identifying influential edges and node features, and presenting them as small explanatory subgraphs (see Fig. 3). For example, an explanation produced by GNNEx-

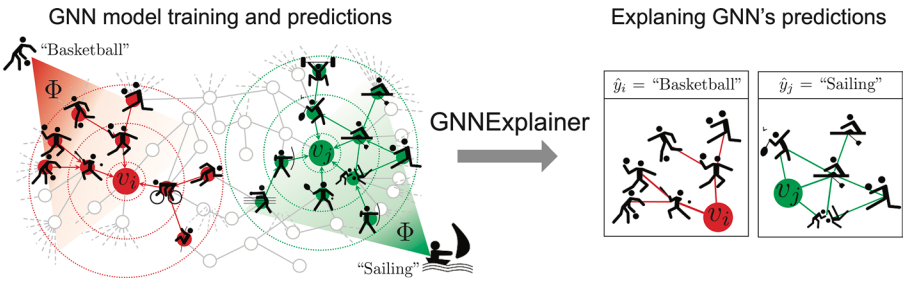

Fig. 3 Example of a GNNExplainer-generated explanation (reproduced from Ying et al., 2019, with authors' permission). A GNN model Φ trained on a social interaction graph predicts $\hat{y}_i$ that v_i will engage in "Basketball." GNNExplainer identifies a subgraph and key node features indicating that many friends in one part of v_i's social circle enjoy ball games, leading to the basketball prediction

[6] Persistent homology is especially valuable in graph classification tasks, in which understanding global structure is essential. It highlights the arrangement of nodes and edges rather than individual properties, uncovering hidden patterns and structures that are not immediately apparent. This enables the extraction of simple topological measures from complex data, offering holistic insights into internal relationships even in the presence of noise or high complexity. In graph-based machine learning, persistent homology facilitates an understanding of the overall structure of data graphs and supports informed predictions based on that structure. Its advantages include a strong theoretical foundation, practical computability, robustness in the face of minor perturbations in learning parameters, deterministic feature extraction that efficiently compresses meaningful data structures, and naturally interpretable outputs such as persistence diagrams. A persistence diagram visually represents the shape and structure of data by illustrating how features persist and change across scales. Each point in the diagram corresponds to a feature, or "hole," at a specific scale, with the horizontal axis indicating the feature's birth and the vertical axis its death. Features with short lifespans appear near the diagonal, while those with longer persistence appear farther away, often forming clusters that correspond to recurring patterns in the data.

plainer might state: "node v_i is similar to node v_j because this part of the structure, as well as attributes of v_i's subgraph matches those of v_j."[7]

For a GNN trained on a social interaction graph to predict future sports activities for a new player, a possible topological explanation proceeds as follows:

(T1) In the graph, there exists a particular subgraph—a tightly connected friend clique containing many "ball game" nodes—and a corresponding subset of node features (e.g., "plays basketball") that are most influential for the classification.
(T2) The system classifies the player as participating in "Basketball."
(T3) Had this subgraph been absent or altered (e.g., if its edges were removed or its features masked), the model would not have predicted "Basketball" for the player or would have assigned it a substantially lower probability.
(T4) The fact that the player is embedded in this specific ball game friend-clique subgraph provides the explanation for why the model classifies the player as a "Basketball" participant.

GNNExplainer offers an intuitive interpretation of the counterfactual dependence required by T3—an interpretation endorsed by Ying et al., who note that "if removing an edge between v_j and v_k strongly decreases the probability of prediction $\hat{y}_i$ then the absence of that edge is a good counterfactual explanation for the prediction at v_i" (2019, p. 4). However, the original method infers difference making only indirectly, relying on gradient-based influence estimates rather than performing actual graph interventions.

Subsequent extensions fill this gap by constructing and testing counterfactual graph modifications. CF-GNNExplainer (Lucic et al., 2022) identifies minimal edge deletions that flip model predictions and verifies their effects by reevaluating the model. XPlore (De Sanctis et al., 2026) generates counterfactual edits to edges or node features and validates them through observed changes in predictions. COMBINEX (Giorgi et al., 2025) formulates counterfactual generation as a constrained optimization problem that jointly modifies graph structure and attributes. In all three cases, counterfactuals are explicitly instantiated and empirically evaluated, providing direct evidence that predictions depend on specific graph configurations.

Such intervention-based evidence, however, remains rare in the TDA literature. Despite the frequent use of explanatory language, many topology-based approaches infer explanatory relationships on the basis of correlation alone. A representative example is Elhamdadi et al. (2022), who apply TDA and persistent homology to emotion recognition from facial poses. Using a visual analytics tool, they extract topological features of facial landmarks over time in the form of persistence diagrams and compute pairwise dissimilarities using tailored distance metrics, yielding a dissimilarity matrix that captures topological variation across facial expressions.

In Fig. 4, the authors present facial poses classified as "surprise," illustrating how persistence diagrams and representative cycles relate topological features to emotion

[7] A similar strategy is employed in *self-explainable graph neural networks* (SE-GNNs) by Dai and Wang (2021).

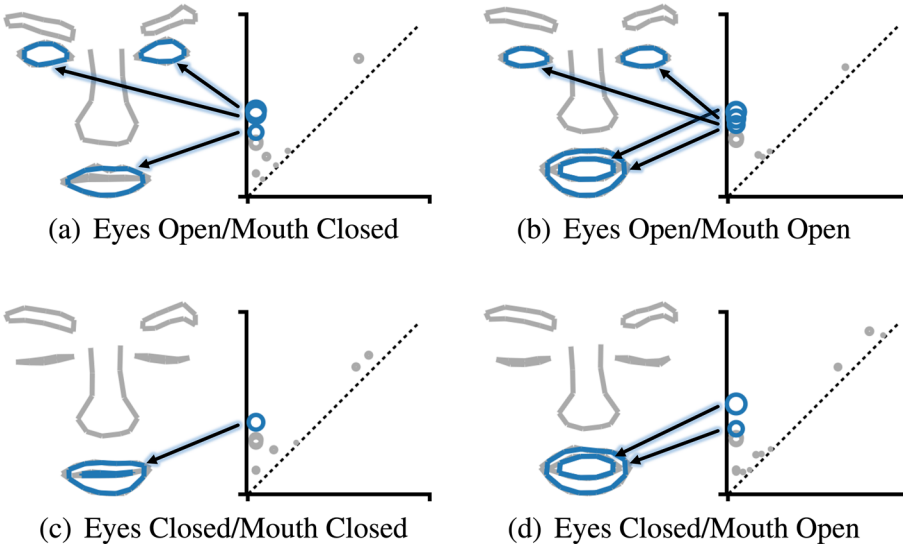

Fig. 4 Four female facial poses correctly classified as "surprise." Each pose includes persistence diagrams indicating the number and persistence of topological features, alongside representative cycles that highlight landmarks like the eyes and mouth movements commonly associated with a surprised expression (reproduced from Elhamdadi et al., 2022, arXiv version licensed under CC BY 4.0)

classification. These visualizations allow researchers to associate the geometry of facial landmarks with high-persistence features, thereby supporting both empirical and topological interpretations.

Although Elhamdadi et al. present their topology-based method as enhancing "explainability,"[8] the evidence they provide remains correlational. They identify stable associations between topological features and emotion labels and suggest that, had persistence patterns differed, the classifier might have produced a different emotion label. However, they neither intervene on these features nor supply a mathematical argument showing that modifying them would necessarily change the model's predictions.

Still, two considerations lend partial support for a counterfactual reading. First, persistent homology is explicitly designed to track geometric change, implying that alterations in facial geometry must induce corresponding changes in persistence diagrams. Second, the reported regularities are systematic and consistent: specific topological signatures align with known facial movements, and emotion-specific clusters recur across subjects and embeddings. Taken together, these features render the proposed dependence plausible.

Üçer et al. (2022) introduce GSNAc, a classifier that transforms tabular data into graph representations by modeling samples as nodes and similarities as edges. At

[8] Specifically, Elhamdadi et al. (2022, p. 9, Fig. 10) state that "the number and persistence of features explains the difference between poses" and that these topological features "strongly align" with emotion-relevant facial configurations. They further suggest that TDA "provides a means … to discern what parts of the face may be causing the machine learning algorithm to recognize the data as a particular emotion or explain the shortcomings or misdiagnoses" of such recognition.

inference time, test nodes are embedded into the graph and labeled according to their similarity to training instances, using vectorial and topological metrics such as cosine similarity between neighborhood structures. Applied to neuron classification in the *C. elegans* nervous system, GSNAc focuses on the "AY Ganglion" class, representing a group of peripheral neuron cell bodies. Here, a neuron's local topological similarity functions as the explanans, while assignment to the AY label is the explanandum, with the implicit counterfactual that a neuron lacking such similarity would receive a different classification.

Because GSNAc deterministically maps similarity scores to labels, the explanandum depends on the explanans by design: sufficiently altering the similarity would necessarily alter the output. This satisfies only a minimal and trivial notion of counterfactual dependence, one engineered by the classifier's decision rule rather than grounded in mathematical necessity intrinsic to the graph. No topological invariant requires a node with a given local configuration to belong to a specific class; GSNAc simply imposes this mapping. As a result, the method identifies a predictively useful feature but does not establish the robust, non-accidental counterfactual dependence required for a topological explanation. Meeting this stronger standard would require showing that the same classification follows across reasonable similarity-based methods, or that certain local topological configurations necessitate specific labels independently of GSNAc's decision rule.

Finally, Spannaus et al. (2024) employ the *Mapper algorithm*[9] to create topological representations of a model's prediction space. In these Mapper graphs, clusters of inputs form vertices, and overlapping inputs create edges, effectively capturing the "closeness" of data points in the prediction space. The method is applied to deep *convolutional neural networks* (CNNs) for tasks such as cancer phenotype prediction and 20 Newsgroups topic classification. The resulting graphs reveal systematic correlations between the proximity of data points and classification outcomes, providing insights into model decision rules by identifying regions with high predictive accuracy (see Fig. 5).

In the 20 Newsgroups case, documents from the "alt.atheism" and "talk.religion.misc" classes form dispersed and overlapping clusters, reflecting shared vocabulary and correlating with elevated misclassification rates. On this basis, the authors suggest that such disconnected regions are "likely to be misclassified" (p. 12), implying that tighter clustering around ground truth classes would improve accuracy. However, neither the data nor the embeddings are manipulated to test whether changes in cluster structure would in fact alter the model's predictions.

The counterfactual claim derives some plausibility from the stability of these regularities across samples and from the fact that Mapper faithfully represents the geometry of the prediction space, such that changes in embeddings would be mirrored in the visualization. Nonetheless, Mapper is a post hoc analytical tool that summarizes

[9] The Mapper algorithm, introduced by Singh et al. (2007), is a TDA method that constructs a simplified graph capturing the "shape" of high-dimensional data. It operates by applying a filter (or "lens") function to the data, dividing its range into overlapping intervals, clustering the data points within each interval, and connecting clusters that share data points. The resulting graph offers a compressed summary of global structure, in which loops, branches, and clusters often reveal meaningful patterns or subpopulations. Such representations can guide feature selection by highlighting distinctions between data points.

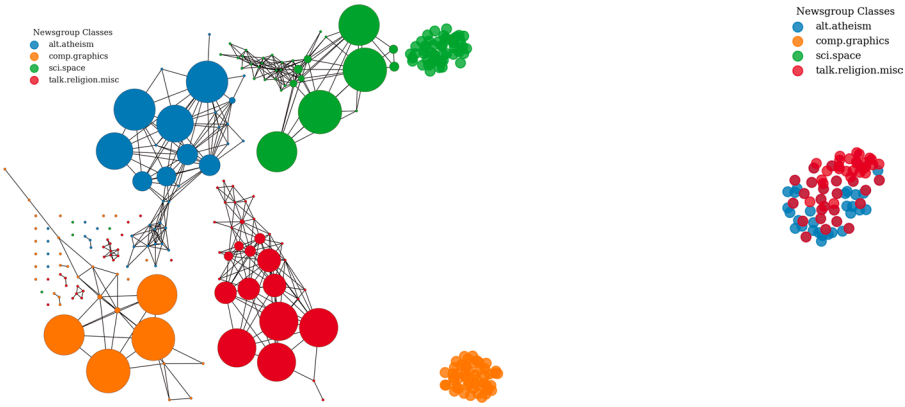

Fig. 5 The Mapper graph for the 20 Newsgroups dataset, showing color-coded ground truth labels and node sizes based on document counts (on the left). The accompanying two-dimensional word embeddings (on the right) highlight regions of high predictive accuracy, although some overlap exists between "alt.Atheism" and "talk.Religion.misc" due to their similar vocabularies (adapted by merging two figures from Spannaus et al., 2024, arXiv version licensed under CC BY 4.0)

model behavior rather than influencing or constraining it. The apparent dependency therefore does not reflect genuine difference making but arises from the construction of the visualization itself. The analysis improves the intelligibility of classification failures and supports a pattern-based understanding of model behavior, but it does not satisfy the requirements of the T-schema.

3.3 Model-driven topological explanations

Analyzing input data and output predictions in isolation often fails to reveal the factors driving decision making. Many contemporary XAI techniques focus instead on the internal structures and functional organization of DNNs, targeting *model explainability*. These approaches include inherently interpretable models, proxy models that approximate complex black boxes, and post hoc methods that probe internal model states (e.g., gradients, activations, or weights).

In this context, graph-theoretical methods and TDA techniques extend beyond dataset characterization to provide *model-driven explanations* by probing the trained network itself. This involves understanding both the model's structure (computational graph architecture and training dynamics) and its function (inner representations encoded in activations and weight spaces).[10] Topological summaries of these

[10] While the architecture constrains its inner representations, it is the learned parameters that define them.

elements can reveal how training organizes internal representations and how this organization, in turn, shapes model behavior on novel inputs.[11]

In *topological deep learning* (TDL), researchers employ topological methods to track information flow within DNNs despite their inherent complexity. They focus on higher-order relations, including multiway interactions among model entities, capturing long-distance or seemingly disparate connections. Such relations align with message-passing principles in complex networks and have contributed substantially to recent advances in AI. Specifically, architectures informed by specific topological principles can constrain or enhance computation by enforcing particular message-passing schemes, preserving topological invariants, or promoting desirable structural properties in learned representations.

More generally, DNNs function as "knowledge-distilling pipelines," progressively abstracting features with increasing depth (Watanabe & Yamana, 2022, p. 76). Each layer transforms raw input data into increasingly invariant and semantically meaningful representations (Bengio et al., 2013; Buckner, 2019; LeCun et al., 2015). In image recognition, for example, early layers typically respond to simple features such as edges or color gradients, while deeper layers encode more complex structures such as contours or facial components. By combining these features, DNNs encode representational "knowledge" in their architecture.

Topological methods offer a distinctive perspective on this abstraction process by revealing how feature combinations emerge, persist, and disappear across layers. By computing the persistent homology of layer-wise activations, one can track how class-specific data manifolds evolve as they propagate through the network. On this view, representational content is encoded in the "shape" of activation spaces, and topological invariants such as connected components or holes correspond to the presence, integration, or separation of features. Their emergence, merging, or collapse reflects the network's way of organizing information and may systematically correlate with task performance.

In this light, central concerns in AI model explainability include identifying the features a system attends to or disregards as data are transformed across layers, assessing the effects of design choices or parameter changes, and examining the relationships among model parameters, learning objectives, and outputs. Because direct interpretation of individual DNN weights quickly becomes intractable due to combinatorial explosion, alternative strategies are required. Topological methods address this challenge by providing insight into a network's internal "cognitive" processes through analysis of the shapes and transformations of its internal representations as

[11] To avoid confusion, it is important to distinguish the explanatory topological property from the representational artifact used to depict it. In both data-driven and model-driven explanations, topological summaries serve as representations of deeper structural or algebraic features of DNNs (which themselves aim to model aspects of the world). For example, a persistence diagram or Mapper graph may reveal clusters or loops in a dataset, while a network's computational graph or activation space may be summarized using measures such as node degree, collapsed Betti numbers, or branching patterns in weight space. Yet it is the underlying topological property that does the explanatory work, not the visual or numerical artifact that makes it visible. The artifact simply provides an additional representational layer that provides epistemic access to the property.

information propagates through successive layers. The following paragraphs examine selected examples that illustrate this.

One promising direction in TDL involves integrating insights from network science into the architectural design of DNNs. Stier and Granitzer (2019), for example, embed *directed acyclic graphs* (DAGs) into feed-forward neural networks to study the influence of structural properties on image classification performance. They begin with randomly generated graphs exhibiting characteristic topological features, such as power-law degree distributions (indicative of scale-free structure) and short characteristic path lengths (suggestive of small-world connectivity). These graphs are transformed into DAGs and mapped onto neural network layers, yielding sparse architectures that are subsequently trained on MNIST (see Fig. 6).

Across 10,000 generated architectures, ranging from 50 to 500 vertices and 97 to 4,365 edges, the authors identify strong correlations between performance and several topological measures, most notably variance in degree, eccentricity, and neighborhood structure. These properties capture heterogeneity in path lengths and local connectivity, which affect the propagation of information through the network. In particular, high variance in *eccentricity*—measuring the spread of longest shortest paths from each node (Takes & Kosters, 2013)—is consistently associated with higher classification accuracy. This pattern suggests that the eccentricity distribution, as a global property, may topologically explain why certain network architectures achieve superior performance. If so, the topological explanation proceeds as follows:

(T1) The network has a high variance in eccentricity distribution.
(T2) The model achieves high classification accuracy.
(T3) If the network had lower variance in eccentricity distribution, it would classify the dataset with lower accuracy.
(T4) Thus, the specific eccentricity distribution provides an explanation as to why the model achieved high accuracy levels.

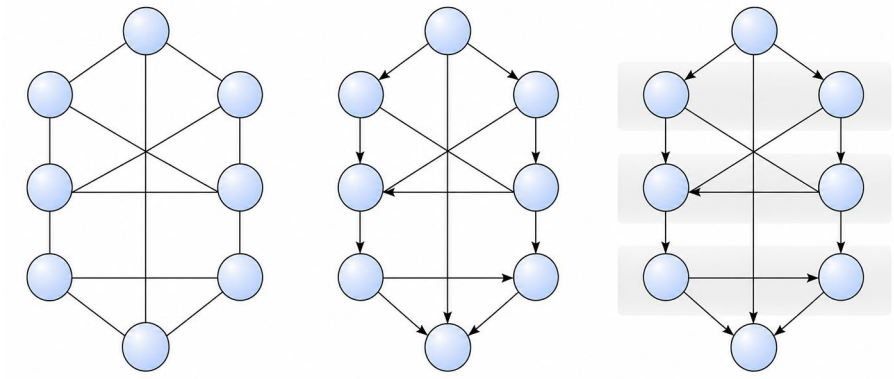

Fig. 6 Schematic illustration of a random graph transformed into a directed acyclic graph and embedded into a neural network classifier (based on Stier and Granitzer, 2019)

A strong functional, network-theoretic rationale supports a counterfactual interpretation in this case. High eccentricity variance implies a mixture of short and long information pathways—a configuration known to benefit signal processing in complex networks.[12] In deep learning, short paths have been shown to facilitate signal and gradient propagation—analogous to skip connections and small-world architectures—while longer paths enable deeper hierarchical processing, richer transformations, and greater expressive capacity.[13] Networks with diverse path lengths thus behave like ensembles of shallow and deep subcircuits, better accommodating tasks that require multiple levels of abstraction. Because reducing this variance would impair either optimization or expressivity and thereby likely degrade performance, the eccentricity distribution serves as an explanans in a topological explanation of model performance.[14]

Another approach involves using topological information to assess and improve model performance. Gabrielsson and Carlsson (2019) apply persistent homology and the Mapper algorithm to the weight spaces of trained CNNs. Focusing on first-layer convolutional filters, they find that across architectures, depths, datasets, and training stages, these filters reliably organize into a simple geometric structure: a single loop corresponding to rotating edge detectors, captured by a dominant one-dimensional homology class (the "primary circle"). The persistence of this feature correlates strongly with test accuracy and cross-dataset generalization, meaning that networks exhibiting a clearer and more persistent loop tend to perform better on held-out data and generalize more effectively to new data.

The relationship between topological structure and generalization is further supported by controlled experiments. In a first set of tests, identical CNNs trained on MNIST are evaluated under three conditions: the first convolutional layer is manually fixed to an idealized primary-circle filter bank; it is fixed to random Gaussian filters; or it is left unconstrained—with all other architectural and training parameters held

[12] Consider simulations of dynamical processes in computation, communication, and transportation networks. Network science studies show that architectures combining short and long paths (of which small-world networks are a specific case) enhance communication efficiency, signal propagation, synchronization, and robustness. Canonical examples include Watts and Strogatz's (1998) small-world model and subsequent work formalizing notions of global and local efficiency (Latora & Marchiori, 2001), as well as simulations of diffusion, synchronization, and routing on heterogeneous networks. In neuroscience in particular, path-length diversity is routinely interpreted as underwriting efficient information processing at the network level (Bassett & Bullmore, 2006; Sporns & Honey, 2006).

[13] This aligns with the broader success of architectures that explicitly increase path-length diversity. *Residual neural networks* (ResNets) and *highway networks* introduce skip connections that create short routes through otherwise very deep models, stabilizing training and improving optimization (He et al., 2016; Veit et al., 2016). Similarly, *small-world neural networks* rewire layers to combine high clustering with short characteristic path lengths, achieving faster convergence and comparable accuracy with fewer parameters (Javaheripi et al., 2019).

[14] This reasoning parallels Kostić's (2018a) canonical example of infectious disease spread, in which a population's small-world topology explains observed transmission rates via the counterfactual that individual network structures would yield individual dynamics. No direct intervention is required because disease propagation is well known to be sensitive to network topology—a claim that aligns with common intuitions about differences between disease transmission in densely populated urban environments and sparsely populated rural ones. As Rathkopf (2018) emphasizes, such reasoning qualifies as explanatory by satisfying both probabilistic and counterfactual relevance criteria.

constant. When evaluated on SVHN, networks initialized with the primary-circle filters achieve substantially higher out-of-distribution accuracy than either baseline. A complementary asymmetry reinforces this result: filters learned from SVHN exhibit a clearer and more persistent primary circle than those learned from MNIST, and SVHN-trained networks generalize better to MNIST than vice versa.

Gabrielsson and Carlsson interpret these results as evidence for a systematic "connection between the simplicity of the topological structure of a network's learned weights and its ability to generalize to unseen data" (2019). Reconstructed explanatorily, the proposal rests on the counterfactual claim that if first-layer filters lacked this loop structure, generalization performance would be worse (conversely, poor generalization would be accompanied by a less simple or less persistent topological organization of the filters).

The evidential support for this dependence derives primarily from simulated manipulations of contrast classes defined by the explanans—namely, a range of topological configurations of first-layer filters—and the observation of corresponding contrasts in the explanandum—that is, generalization performance. This establishes sufficiency: introducing the topological motif reliably improves performance. However, the effect of removing an already learned motif is not tested. Even so, the study provides one of the clearest cases of manipulation-based evidence in the literature, suggesting that topological features account for performance differences in deep learning models.

In classical machine learning, generalization is typically assessed on a held-out test set. Corneanu et al. (2020; see also Ballester et al., 2024) challenge this paradigm by proposing a topology-based method for estimating the generalization gap without test data. They model a network's internal operation as a functional topology, represented by a graph whose nodes correspond to neurons and whose edges encode functional relations such as activation correlations.[15] To assess how well the model captures the underlying data distribution, they compute persistence diagrams of this functional graph and interpret persistent cavities, quantified via *Betti numbers*,[16] as indicators of structural complexity in learned representations (see Fig. 7).

The authors hypothesize that networks which genuinely learn the structure of the data develop well-organized functional topologies with persistent features, whereas overfitted networks exhibit weak or degenerate topology. Across several computer vision benchmarks, they find consistent support for this claim, as higher average persistence robustly correlates with smaller generalization gaps and better test performance. Topological summaries predict test accuracy with errors between 4.6% and 8.45%, comparable to estimates obtained from labeled test sets. On this basis, they argue that information about topological properties can help "make deep learning more explainable" (2020, p. 2684).

[15] More precisely, Corneanu et al. define a metric space over the network by computing pairwise correlations between the activations of network nodes—typically filters or mean outputs in convolutional layers—which determine functional relationships. Using this metric, they apply *Vietoris–Rips filtration* to construct a sequence of simplicial complexes on which persistent homology is computed.

[16] Higher Betti numbers indicate more holes or voids, suggesting potential overfitting where the network memorizes training data specifics rather than learning general patterns. Conversely, lower Betti numbers suggest better generalization, indicating the network can apply learned patterns to new, unseen data.

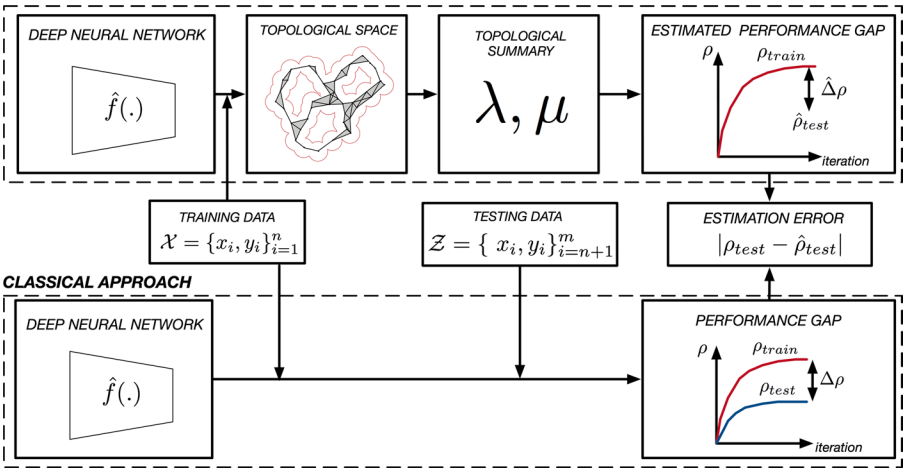

Fig. 7 Computing the test performance of a DNN on a computer vision problem without test samples (top), in contrast to the classical computer vision approach (bottom), where model performance is calculated using a curated test dataset (reproduced from Corneanu et al., 2020, arXiv preprint version, with authors' permission)

The implicit explanatory claim relies on the counterfactual that, had the persistence of topological cavities been lower (indicating weaker topology), the model would have been more susceptible to overfitting. The evidence, however, remains correlational, leaving it unclear whether the observed association reflects a genuine dependence. At the same time, it aligns with a growing body of empirical and analytical work linking topological complexity to generalization capacity (e.g., Bianchini & Scarselli, 2014; Birdal et al., 2021; Guss & Salakhutdinov, 2018; Gutiérrez-Fandiño et al., 2021), which suggests that certain topological properties constrain learning in principled ways. The proposed dependency is also predictive and operationally exploitable, as Corneanu et al. suggest monitoring topological measures during training to guide early stopping.[17]

Finally, Purvine et al. (2023) extend Gabrielsson and Carlsson's analysis of topological structure in CNN representations using persistent homology and Mapper graphs. Analyzing a ResNet-18 trained on CIFAR-10, they identify branching structures in representation spaces that track class separation and reveal the influence of background features on classification. Mapper graphs show that early layers can separate classes based on background pixels, while final convolutional layers consistently exhibit clear, class-separated branching patterns. These effects generalize across architectures, including InceptionV1, InceptionV3, AlexNet, and ImageNet models.

[17] Since topological features correlate with generalization, one could monitor these features during training and stop training at the point where the topology stops becoming more "rich," signaling the potential onset of overfitting. This idea aligns with prior work, such as Rieck et al. (2019), who propose an early stopping criterion based on the persistent homology of network weights.

Across 100 training runs, persistent-homology-based distance metrics between layer representations remain invariant under irrelevant variations such as random initialization or sampling, while varying systematically with functionally relevant factors, including layer depth, architectural differences, and input perturbations. These topological structures are stable across sampling strategies but degrade under increasing input noise, coinciding with declines in classification accuracy. This pattern of invariance and sensitivity supports a plausible counterfactual reading, according to which less structured internal topology would impair performance.

In sum, the reviewed examples illustrate how graph-theoretic and topological methods can enhance our understanding of deep learning models and datasets by foregrounding their structural and algebraic properties. A growing body of research shows that topological approaches can reveal relational organization within data, clarify the relation of input features to model outputs, and provide robust tools for handling noise and high dimensionality. Recognizing that DNNs often learn structures more naturally characterized in topological terms enables deeper analysis of internal representations, training dynamics, and systematic correlations between model performance and data characteristics.

TDA techniques allow such structures to be extracted in a principled and reproducible manner, improving traceability and comparability across models and experiments, while benefiting from the relative stability of many topological features under parameter variation and noise. Moreover, topologically informed models and analyses often exhibit practical advantages. They can achieve comparable or improved performance with fewer parameters or less training data, while supporting stronger generalization and greater interpretability than standard approaches. As Papamarkou et al. (2024) note, emphasizing topological methods may be key to advancing XAI research:

> Training TDL models on the computational graph, controlling topological properties of weight trajectories traced out during stochastic optimization of neural networks, or appropriately enforcing desirable topological properties of a network's internal representation of the data may hold the keys to understanding the puzzling generalization behavior in deep learning and to moving a step closer to explainable AI. This direction emphasizes the potential to uncover explanations that describe how network dynamics depend on specific topological properties, even within the high complexity of large DNNs utilized for image and video recognition tasks, ultimately leading to the development of more interpretable and efficient neural network architectures. (Papamarkou et al., 2024, p. 6)

At the same time, there are principled reasons to question whether these practices qualify as genuine topological explanations in the philosophical sense.

4 Objections and reconciliation

4.1 Correlation vs. counterfactual dependence

The first, already indicated, objection concerns the explanatory status of topology-based accounts when evaluated against philosophical standards. Although the technical literature frequently characterizes topological properties as explanatory, this claim often weakens under closer scrutiny. When assessed using criteria such as Kostić's (2022) T-schema, many of the cited studies establish robust correlations between topological features and aspects of model behavior but fall short of demonstrating genuine difference making. In the absence of evidence that systematically varying a topological property would alter the explanandum, these accounts do not fully satisfy the requirements for explanatoriness in the strict philosophical sense.

That said, not all counterfactual dependencies can be established through direct intervention. In systems with deep interdependencies, isolating a single relevant variable may be practically or conceptually impossible, a difficulty that is especially acute in TDA and TDL contexts. Although software systems may appear readily intervenable, topological explanantia are typically emergent features of data representations or model structures, so intervening on them would simultaneously alter multiple correlated properties and impede attempts to isolate difference-making variables. As a result, support for the T3 condition varies across cases. Some, such as the class-separation toy model discussed in Sect. 3.1, establish structural and modal counterfactuals grounded in mathematical necessity, while others, including subgraph-based GNN explanation methods discussed in Sect. 3.2, provide empirical evidence by manipulating graph structure and observing corresponding changes in model outputs.

Many analyses, however, provide only partial or indirect support for the dependence relation. Some offer limited manipulation-based evidence establishing sufficiency, such as showing that introducing a topological motif improves performance. Others rely on robust correlations, where large-scale experiments and stability across datasets or architectures render a counterfactual interpretation plausible. In still other cases, dependence is present by design, reflecting imposed mappings of a classifier rather than genuine topological necessity. Finally, some explanatory claims gain plausibility when supplemented by broader causal or functional considerations, such as established principles of information processing in complex networks.

Although presented as "explanations," these accounts thus fail to demonstrate the modally robust dependence relations required for genuine topological explanation and are therefore best understood as *candidate* explanations: empirically suggestive and interpretively valuable, but not yet explanatory in the strict sense. Reconstructing each case through the lens of the T-schema distinguishes more clearly between stronger and weaker explanatory claims; it clarifies the underlying reasoning, makes evidential limitations explicit, and, where possible, identifies any additional support that would be required to secure explanatoriness.

If topological properties are indeed explanatory, the topological strategy could substantially advance XAI by leveraging tractable algebraic and structural relations to account for performance where other methods offer limited insight. Given the ubiquity of topological structure in data, representations, and architectures, rigorous

demonstrations of counterfactual dependence could establish topological explanation as a broadly applicable framework for both data and model explainability.

4.2 Topology vs. mechanisms

A second objection is that some of the proposed explanations include empirical, directional, or seemingly causal elements, suggesting that they may be better interpreted as mechanistic rather than topological. This concern reflects a longstanding debate in the philosophy of science, in which topological and mechanistic explanations are often treated as competing (Chirimuuta, 2018; Kostić, 2022; Kostić & Halffman, 2023; Weiskopf, 2011). Mechanistic "imperialists" typically subsume topological explanations under mechanistic accounts or question their explanatory legitimacy altogether (Bechtel, 2020; Craver, 2016; Krickel et al., 2023; Povich & Craver, 2017; Povich, 2018, 2021; Zednik, 2019), whereas topological "autonomists" argue that topological explanation constitutes a distinct, irreducible explanatory strategy (Darrason, 2018; Huneman, 2010; Jones, 2014; Kostić, 2018a, b, 2019, 2022; Kostić & Khalifa, 2023; Rathkopf, 2018).[18]

Relevant for this paper is the fact that, although some deep learning cases combine empirical, causal, and directional elements associated with mechanistic frameworks, it is topology that "does the explanatory work" (cf. Ross, 2021).[19] Even in "borderline" cases that include mechanistic features, the explanations consistently privilege robust relational patterns grounded in topological properties. A pluralist perspective therefore accommodates purely mechanistic explanations, purely topological explanations, and mixed forms such as the "causal–topological" explanations (Ross, 2021), which are best understood as complementary modes suited to different *why*-questions in XAI, depending on the system under study, the context of inquiry, and the particular source of opacity.

With this in mind, the next section examines the distinctive epistemic advantages of topological explanations, situating them within ongoing philosophical debates on XAI.

5 Epistemic significance for XAI

Topological approaches offer significant epistemic advantages for XAI, defining a new and potentially fruitful area in the philosophy of AI. First, they meet the *comprehensibility* criteria outlined by Doran et al. (2017), who distinguish between interpretable, comprehensible, and "truly explainable" systems. In their framework, a comprehensible system produces symbols alongside its outputs, enabling users to connect input properties to outputs using their implicit knowledge and reasoning. The assessment of comprehensibility is based on how easily users can interpret

[18] A full assessment of this debate lies beyond the scope of this paper; the reader is therefore referred to the relevant philosophical literature.

[19] Provided that the topological properties cited in the explanantia make a difference to the corresponding explananda.

these symbols, often relying on cognitive "intuitions" about input–symbol–output relationships.

Topological explanations enhance the comprehensibility of otherwise opaque DNNs by appealing to topological properties that capture input–output relations or general dynamics through patterns of counterfactual dependence. Such properties may include scale-freeness, small-world connectivity, node-degree distributions, network eccentricity, the simplicity of learned filters, computational graph structures, influential subgraphs, similarity to training-data shapes, and proximity of data points within prediction space.

Unlike certain other forms of explanation, they rely on abstract yet human-understandable, well-defined structural concepts and mathematically formalized measures. Explainees need only grasp the relevant mathematical notions associated with topology, rather than the extensive domain-specific knowledge required for, for example, causal or mechanistic understanding (e.g., knowledge of specific entities and activities). This conceptual economy allows topological explanations to retain explanatory depth while improving accessibility and applicability for interpreting complex AI systems (see also Kostić, 2019, for a discussion on complexity and explanatory depth).[20]

Second, the topological explanatory strategy offers significant advantages in enhancing AI systems' *algorithmic transparency*. According to Creel (2020), algorithmic transparency involves understanding the overall functioning of a system by grasping the high-level logical rules governing its transformations—essentially knowing the algorithm that the program instantiates. Because algorithms can be implemented in various ways, understanding the algorithm does not require knowledge of the specific components or their interrelations. Programs that implement the same algorithm can differ substantially in structure, especially when written in a range of programming languages.

For example, Creel cites *local interpretable model-agnostic explanations* (LIME) by Ribeiro et al. (2016) as a method that secures algorithmic transparency without relying on mechanistic decomposition. Similarly, topological methods provide local explanations for individual predictions in machine learning tasks. By abstracting away from the specific organizational details of subcomponents and interactions, topological explanations offer high-level insight into deep learning algorithms, thereby providing a form of algorithmic transparency. They address system-level properties such as performance, robustness, and functional redundancy without requiring detailed knowledge of internal mechanics.

[20] As a case in point for comprehensible models, Doran et al. reference visual explanations derived from techniques like *t-distributed stochastic neighborhood embedding* (t-SNE), a statistical method that visualizes high-dimensional data by positioning each data point on a two- or three-dimensional map. This approach is akin to the explanation of data topologies and their graph representations, as seen in the work of Dai and Wang (2021), who apply t-SNE to node features in order to visualize node positions for similarity comparisons; Elhamdadi et al. (2022), who utilize t-SNE among other dimensionality reduction methods to depict the topological structure of facial landmark data; or Spannaus et al. (2024), who also employ t-SNE to represent word embeddings from regions of high-predictive accuracy in a graph of a newsgroup dataset. These visualizations effectively make the data's underlying topological structures and relationships visible and inspectable.

This aligns with Chirimuuta's (2018, p. 415) observation in neuroscience that "low-level descriptions don't go very far in explaining how the systems work … In order to explain object recognition, neuroscientists … refer to the more abstract structural features of the system, such as its being hierarchical and multilayered." This parallel supports the claim that explaining the behavioral capacities of networked systems requires high-level structural accounts, such as the topological explanations discussed here. In contexts where opacity arises from complexity and dense interconnection, a high-level understanding of DNN dynamics renders low-level details of causal organization less relevant.

Relatedly, a third advantage of the topological approach lies in its capacity to manage the *complexity* of deep learning systems. Dense interdependencies among millions of parameters render such models opaque, since learning dynamics emerge from automated parameter adjustment that enables effective problem solving while undermining interpretability. As models scale to multi-billion-parameter systems (such as *large language models*), they increasingly exhibit *non-decomposability* (cf. Rathkopf, 2018). In non-decomposable systems, component behavior depends on interactions with many others, making part–whole decomposition largely infeasible. Topological explanations address this limitation by capturing global structural properties of the model as a whole. By abstracting away from individual components while remaining informative about system-level behavior, they restore a degree of epistemic accessibility to otherwise opaque models.

Finally, topological methods offer *counterfactual* and highly *generalizable* insights. Buijsman (2022) argues that many XAI techniques fail to explain why AI systems succeed and, drawing on Woodward's (2003) interventionist account, maintains that robust explanations should rely on abstract generalizations that support counterfactual reasoning. Topological explanations fit this ideal by linking topological properties to empirical outcomes in a way that supports counterfactual inferences. These dependencies, however, do not constitute productive causation, since they do not involve interventions on causal variables. Instead, topological properties operate at the same level as their explananda and do not temporally precede them. Consequently, they address system-level phenomena that emerge from interactions among multiple variables rather than isolated causal relations.

Topological explanations appeal to high-level structural features that, although more abstract than those invoked by many other explanatory strategies, remain connected to AI decision making through counterfactual dependence. This level of abstraction yields two epistemic advantages. First, it can reduce the number of variables that must be tracked, thereby improving cognitive feasibility (although the extent of this benefit remains an empirical question). Second, abstraction supports generalization: by focusing on structural relations rather than system-specific parameters, topological explanations enable general claims about model behavior across cases, aligning with Buijsman's notion of "reasonable generalization" in XAI (Buijsman, 2022, p. 569).

Despite these epistemic advantages, philosophers of science have largely overlooked the potential of topological explanations in XAI. This oversight is a mistake, as the examples discussed indicate that certain empirical characteristics of AI systems may counterfactually depend on readily measurable topological properties. Rather

than decomposing the internal mechanisms of AI systems, topological explanations capture structural relationships that account for DNN dynamics without requiring detailed causal or mechanistic analysis. Future work in the philosophy of AI should therefore pay greater attention to topological approaches, particularly as tools for advancing XAI in contexts where traditional methods struggle, such as highly complex models or the analysis of data distributions.

6 Conclusions

Historically, enthusiasm for DNNs was driven by their perceived similarity to the human brain, suggesting that traditional causal and mechanistic frameworks drawn from neuroscience could explain their computational phenomena. However, DNNs operate at levels of abstraction far removed from biological brain processes, casting doubt on the prospects for accurate low-level explanations. This skepticism has contributed to a backlash against decompositional approaches, fostering greater acceptance of functional, fictional, non-causal, and mathematical explanations—a trend particularly visible in computational neuroscience. Consequently, the imposition of causal–mechanistic frameworks on XAI research may be misguided, thus motivating a shift toward more abstract structural or mathematical explanatory strategies that apply across both biological and artificial systems.

Topological explanations of DNNs exemplify this shift by providing a robust framework for explaining opaque AI systems through counterfactual dependencies between topological and empirical properties. Given the complexity and high interconnectivity of DNNs, explaining their behavior requires an understanding of the abstract aspects of their connectivity and relationships with empirical characteristics like accuracy, robustness, and generalizability. Case studies in topological data analysis and topological deep learning suggest that topological reasoning may enhance explainability regarding internal structures, network weights, algorithms, and dataset characteristics.

Topological explanations are valuable for XAI not only because of their specific insights, but also because they outline a more general explanatory strategy. When appropriately adapted, the T-schema provides a flexible framework for capturing counterfactual dependencies involving not only topological properties but also other abstract, non-causal properties. Moreover, topological methods are already widely used in deep learning practice and provide readily available tools for analyzing both data and model structures, making them well suited to addressing multiple dimensions of the black box problem in AI. For these reasons, the topological approach should be regarded as a distinct explanatory strategy capable of answering a broad range of why-questions about AI systems where traditional approaches falter.

Acknowledgements I am grateful to Daniel Kostić for invaluable discussions and comments on earlier versions of this paper. I also thank Marcin Miłkowski, Jim Woodward, Kareem Khalifa, Wiktor Lachowski, and Daniel Piecka for helpful comments. I thank the anonymous reviewers for constructive feedback that improved the manuscript.

Author contributions Single author, 100% contribution.

Funding This work was supported by the National Science Centre, Poland, under PRELUDIUM grant no. 2023/49/N/HS1/02461.

Data availability Not applicable.

Declarations

Competing interests The author declares no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ballester, R., Clemente, X. A., Casacuberta, C., Madadi, M., Corneanu, C. A., & Escalera, S. (2024). Predicting the generalization gap in neural networks using topological data analysis. *Neurocomputing*, *596*, 127787. https://doi.org/10.1016/j.neucom.2024.127787

Bassett, D. S., & Bullmore, E. (2006). Small-world brain networks. *The Neuroscientist*, *12*(6), 512–523. https://doi.org/10.1177/1073858406293182

Batterman, R. W., & Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, *81*(3), 349–376. https://doi.org/10.1086/676677

Bechtel, W. (2020). Hierarchy and levels: Analysing networks to study mechanisms in molecular biology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1796), 20190320. https://doi.org/10.1098/rstb.2019.0320

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bianchini, M., & Scarselli, F. (2014). On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(8), 1553–1565. https://doi.org/10.1109/TNNLS.2013.2293637

Birdal, T., Lou, A., Guibas, L. J., & Şimşekli, U. (2021). Intrinsic dimension, persistent homology and generalization in neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. WORTMAN. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 6776–6789). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/35a12c43227f217207d4e06ffefe39d3-Paper.pdf

Bricken, T., Templeton, A., Batson, J., Olah, C., Henighan, T., Carter, S., Hume, T., Burke, J. E., McLean, B., Nguyen, K., Tamkin, A., Joseph, N., Maxwell, T., Schiefer, N., Kravec, S., Wu, Y., Lasenby, R., Askell, A., & Denison, C., … Chen, B. (2023, October 5). Decomposing language models into understandable components. *Transformer circuits Thread, Anthropic*. Retrieved from https://www.anthropic.com/index/decomposing-language-models-into-understandable-components

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), e12625. https://doi.org/10.1111/phc3.12625

Buijsman, S. (2022). Defining explanation and explanatory depth in XAI. *Minds and Machines*, *32*(3), 563–584. https://doi.org/10.1007/s11023-022-09607-9

Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., & Olah, C. (2021). Curve circuits. *Distill*. Retrieved from https://distill.pub/2020/circuits/curve-circuits/

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, *46*(2), 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X

Casper, J. L. (2023). Takeaways from the mechanistic interpretability challenges. *AI alignment forum*. Retrieved from https://www.alignmentforum.org/posts/EjsA2M8p8ERyFHLLY/takeaways-from-the-mechanistic-interpretability-challenges

Chirimuuta, M. (2018). Explanation in computational neuroscience: Causal and non-causal. *The British Journal for the Philosophy of Science*, *69*(3), 849–880. https://doi.org/10.1093/bjps/axw034

Corneanu, C. A., Escalera, S., & Martinez, A. M. (2020). Computing the testing error without a testing set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2677–2685). https://doi.org/10.1109/CVPR42600.2020.00275

Craver, C. F. (2016). The explanatory power of network models. *Philosophy of Science*, *83*(5), 698–709. https://doi.org/10.1086/687856

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, *87*(4), 568–589. https://doi.org/10.1086/709729

Dai, E., & Wang, S. (2021). Towards self-explainable graph neural network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 302–311). Association for Computing Machinery. https://doi.org/10.1145/3459637.3482306

Darrason, M. (2018). Mechanistic and topological explanations in medicine: The case of medical genetics and network medicine. *Synthese*, *195*(1), 147–173. https://doi.org/10.1007/s11229-015-0983-y

De Sanctis, M., De Sanctis, R., Faralli, S., & Prenkaj, B. (2026). *Beyond edge deletion: A comprehensive approach to counterfactual explanation in graph neural networks*. OpenReview. https://openreview.net/forum?id=L4KJT9QpqE

Doran, D., Schulz, S. C., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. In *CEUR Workshop Proceedings, 2071*. https://doi.org/10.48550/arXiv.1710.00794

Elhamdadi, H., Canavan, S., & Rosen, P. (2022). AffectiveTDA: Using topological data analysis to improve analysis and explainability in affective computing. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 769–779. https://doi.org/10.1109/TVCG.2021.3114784

Erasmus, A., Brunet, T. D., & Fisher, E. (2021). What is interpretability? *Philosophy & Technology*, *34*(4), 833–862. https://doi.org/10.1007/s13347-020-00435-2

Gabrielsson, R. B., & Carlsson, G. (2019). Exposition and interpretation of the topology of neural networks. In *2019 18th IEEE International Conference on Machine Learning and Applications* (pp. 1069–1076). IEEE. https://doi.org/10.1109/ICMLA.2019.00180

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics* (pp. 80–89). IEEE. https://doi.org/10.1109/DSAA.2018.00018

Giorgi, F., Silvestri, F., & Tolomei, G. (2025). COMBINEX: A unified counterfactual explainer for graph neural networks via node feature and structural perturbations. arXiv, abs/2502.10111. https://arXiv.org/abs/2502.10111

Green, S., Şerban, M., Scholl, R., Jones, N., Brigandt, I., & Bechtel, W. (2018). Network analyses in systems biology: New strategies for dealing with biological complexity. *Synthese*, *195*(4), 1751–1777. https://doi.org/10.1007/s11229-016-1307-6

Guss, W. H., & Salakhutdinov, R. (2018). On characterizing the capacity of neural networks using algebraic topology. *arXiv*, abs/1802.04443. https://arXiv.org/abs/1802.04443

Gutiérrez-Fandiño, A., Pérez Fernández, D., Armengol-Estapé, J., & Villegas, M. (2021). Persistent homology captures the generalization of neural networks without a validation set. *arXiv, Abs/2106.00012*. https://arXiv.org/abs/2106.00012

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90

Humphreys, P. (2014). Explanation as condition satisfaction. *Philosophy of Science*, *81*(5), 1103–1116. https://doi.org/10.1086/677698

Huneman, P. (2010). Topological explanations and robustness in biological sciences. *Synthese*, *177*(2), 213–245. https://doi.org/10.1007/s11229-010-9842-z

Huneman, P. (2018a). Outlines of a theory of structural explanations. *Philosophical Studies*, *175*(3), 665–702. https://doi.org/10.1007/s11098-017-0887-4

Huneman, P. (2018b). Realizability and the varieties of explanation. *Studies in History & Philosophy of Science Part A*, *68*, 37–50. https://doi.org/10.1016/j.shpsa.2018.01.004

Javaheripi, M., Darvish Rouhani, B., & Koushanfar, F. (2019). Swnet: Small-world neural networks and rapid convergence. *arXiv, Abs/1904.04862*. https://arXiv.org/abs/1904.04862

Jones, N. (2014). Bowtie structures, pathway diagrams, and topological explanation. *Erkenntnis*, *79*(5), 1135–1155. https://doi.org/10.1007/s10670-014-9598-9

Kästner, L., & Crook, B. (2024). Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science*, *14*(4), 52. https://doi.org/10.1007/s13194-024-00614-4

Kostić, D. (2018a). Mechanistic and topological explanations: An introduction. *Synthese*, *195*(1), 1–10. https://doi.org/10.1007/s11229-016-1257-z

Kostić, D. (2018b). The topological realization. *Synthese*, *195*(1), 79–98. https://doi.org/10.1007/s11229-016-1248-0

Kostić, D. (2019). Minimal structure explanations, scientific understanding and explanatory depth. *Perspectives on Science*, *27*(1), 48–67. https://doi.org/10.1162/posc_a_00299

Kostić, D. (2022). Topological explanations: An opinionated appraisal. In I. Lawler, E. Shech, & K. Khalifa (Eds.), *Scientific understanding and representation: Mathematical modeling in the life and physical sciences* (pp. 96–115). Routledge. https://doi.org/10.4324/9781003202905-9

Kostić, D., & Halffman, W. (2023). Mapping explanatory language in neuroscience. *Synthese*, *202*(4), 112. https://doi.org/10.1007/s11229-023-04329-6

Kostić, D., Hilgetag, C., & Tittgemeyer, M. (2020). Unifying the essential concepts of biological networks: Biological insights and philosophical foundations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1796), 20190314. https://doi.org/10.1098/rstb.2019.0314

Kostić, D., & Khalifa, K. (2021). The directionality of topological explanations. *Synthese*, *199*(5–6), 14143–14165. https://doi.org/10.1007/s11229-021-03414-y

Kostić, D., & Khalifa, K. (2023). Decoupling topological explanations from mechanisms. *Philosophy of Science*, *90*(2), 245–268. https://doi.org/10.1017/psa.2022.29

Krickel, B., de Bruin, L., & Douw, L. (2023). How and when are topological explanations complete mechanistic explanations? The case of multilayer network models. *Synthese*, *202*(1), 14. https://doi.org/10.1007/s11229-023-04241-z

Kuorikoski, J. (2021). There are No mathematical explanations. *Philosophy of Science*, *88*(2), 189–212. https://doi.org/10.1086/711479

Lacková, Ľ., & Zámečník, L. (2020). Logical principles of a topological explanation: Peirce's iconic logic. *Chinese Semiotic Studies*, *16*(3), 493–514. https://doi.org/10.1515/css-2020-0027

Ladyman, J., Ross, D., Spurrett, D., & Collier, J. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press.

Lange, M. (2009). Dimensional explanations. *Noûs*, *43*(4), 742–775. https://doi.org/10.1111/j.1468-0068.2009.00726.x

Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science*, *64*(3), 485–511. https://doi.org/10.1093/bjps/axs012

Lange, M. (2017). *Because without cause: Non-causal explanations in science*. Oxford University Press.

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

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. B. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lucic, A., Ter Hoeve, M. A., Tolomei, G., De Rijke, M., & Silvestri, F. (2022). CF-GNNExplainer: Counterfactual explanations for graph neural networks. In G. Camps-Valls, F. J. R. Ruiz, & I. Valera (Eds.). *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics* (Vol. *151*, pp. 4499–4511). PMLR. https://proceedings.mlr.press/v151/lucic22a.html

Mikhalevich, I. (2025). Intervention and experiment. *European Journal for Philosophy of Science*, *15*(1), 1–25. https://doi.org/10.1007/s13194-025-00647-3

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Millière, R., & Buckner, C. (2025). Interventionist methods for interpreting deep neural networks. In G. Piccinini (Ed.), *Neurocognitive foundations of mind* (1st ed., pp. 190–221). Routledge. https://doi.org/10.4324/9781003458531

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 279–288). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287574

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Naitzat, G., Zhitnikov, A., & Lim, L. H. (2020). Topology of deep neural networks. *The Journal of Machine Learning Research*, *21*(184), 1–40. https://jmlr.csail.mit.edu/papers/volume21/20-345/20-345.pdf

Nanda, N. (2023, July 6). Concrete open problems in mechanistic interpretability: A technical overview. *Effective altruism forum*. Retrieved from https://forum.effectivealtruism.org/posts/EMfLZXvwiEioPWPga/concrete-open-problems-in-mechanistic-interpretability-a

Olah, C. (2014, March). Neural networks, manifolds, and topology. Retrieved from https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Papamarkou, T., Birdal, T., Bronstein, M., Carlsson, G., Curry, J., Gao, Y., Hajij, M., Kwitt, R., Liò, P., Di Lorenzo, P., Maroulas, V., Miolane, N., Nasrin, F., Ramamurthy, K. N., Rieck, B., Scardapane, S., Schaub, M. T., Veličković, P.Wang, B., & Zamzmi, G. (2024). Position: Topological deep learning is the new frontier for relational learning. In *Proceedings of Machine Learning Research* (Vol. *235*, pp. 39529–39555). https://pmc.ncbi.nlm.nih.gov/articles/PMC11973457/

Povich, M. (2018). Minimal models and the generalized ontic conception of scientific explanation. *The British Journal for the Philosophy of Science*, *69*(1), 117–137. https://doi.org/10.1093/bjps/axw019

Povich, M. (2021). The narrow ontic counterfactual account of distinctively mathematical explanation. *British Journal for the Philosophy of Science*, *72*(2), 511–543. https://doi.org/10.1093/bjps/axz008

Povich, M. (2025). Distinctively mathematical explanation. In *Rules to infinity: The normative role of mathematics in scientific explanation*. Oxford University Press. https://doi.org/10.1093/oso/9780197679005.003.0002

Povich, M., & Craver, C. F. (2017). Mechanistic levels, reduction, and emergence. In S. Glennan, & P. Illari (Eds.), *The Routledge handbook of mechanisms and Mechanical philosophy* (pp. 185–197). Routledge. Retrieved from https://philarchive.org/archive/POVMLR

Purvine, E., Brown, D., Jefferson, B., Joslyn, C., Praggastis, B., Rathore, A., Shapiro, M., Wang, B., & Zhou, Y. (2023). Experimental observations of the topology of convolutional neural network activations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(8), 9470–9479. https://doi.org/10.1609/aaai.v37i8.26134

Rabiza, M. (2026). A mechanistic explanatory strategy for XAI. In V. C. Müller, L. Dung, G. Löhr, & A. Rumana (Eds.), *Philosophy of artificial intelligence: The state of the art* (pp. 389–412, Synthese Library, Vol. 533). Springer. https://doi.org/10.1007/978-3-032-10073-3_23

Rathkopf, C. (2018). Network representation and complex systems. *Synthese*, *195*(1), 55–78. https://doi.org/10.1007/s11229-015-0726-0

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rice, C. (2021). Understanding realism. *Synthese*, *198*(5), 4097–4121. https://doi.org/10.1007/s11229-019-02331-5

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., & Borgwardt, K. (2019). Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *Proceedings of the 7th International Conference on Learning Representations*, OpenReview. https://openreview.net/forum?id=ByxkijC5FQ

Ross, L. N. (2021). Distinguishing topological and causal explanation. *Synthese*, *198*(10), 9803–9820. https://doi.org/10.1007/s11229-020-02685-1

Singh, G., Mémoli, F., & Carlsson, G. (2007). Topological methods for the analysis of high dimensional data sets and 3D object recognition. In M. Botsch, R. Pajarola, B. Chen, & M. Zwicker (Eds.), *Eurographics symposium on Point-based Graphics* (pp. 91–100). The Eurographics Association. https://doi.org/10.2312/SPBG/SPBG07/091-100

Spannaus, A., Hanson, H. A., Tourassi, G., & Penberthy, L. (2024). Topological interpretability for deep learning [Article 21]. *Proceedings of the Platform for Advanced Scientific Computing Conference*, Association for Computing Machinery. https://doi.org/10.1145/3659914.3659935

Sporns, O., & Honey, C. J. (2006). Small worlds inside big brains. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(51), 19219–19220. https://doi.org/10.1073/pnas.0609523103

Stier, J., & Granitzer, M. (2019). Structural analysis of sparse neural networks. *Procedia Computer Science*, *159*, 107–116. https://doi.org/10.1016/j.procs.2019.09.165

Takes, F. W., & Kosters, W. A. (2013). Computing the eccentricity distribution of large graphs. *Algorithms*, *6*(1), 100–118. https://doi.org/10.3390/a6010100

Üçer, S., Ozyer, T., & Alhajj, R. (2022). Explainable artificial intelligence through graph theory by generalized social network analysis-based classifier. *Scientific Reports*, *12*(1), 15210. https://doi.org/10.1038/s41598-022-19419-7

Veit, A., Wilber, M. J., & Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/37bc2f75bf1bcfe8450a1a41c200364c-Paper.pdf

Watanabe, S., & Yamana, H. (2022). Topological measurement of deep neural networks using persistent homology. *Annals of Mathematics and Artificial Intelligence*, *90*(1), 75–92. https://doi.org/10.1007/s10472-021-09761-3

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

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, *183*(3), 313–338. https://doi.org/10.1007/s11229-011-9958-9

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2018). Some varieties of non-causal explanation. In A. Reutlinger, & J. Saatsi (Eds.), *Explanation beyond causation: Philosophical perspectives on non-causal explanations*. Oxford University Press. https://doi.org/10.1093/oso/9780198777946.003.0007

Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/d80b7040b773199015de6d3b4293c8ff-Paper.pdf

Zednik, C. (2019). Models and mechanisms in network neuroscience. *Philosophical Psychology*, *32*(1), 23–51. https://doi.org/10.1080/09515089.2018.1512090

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.