TOPOLOGICAL REPRESENTATIONS ENHANCE MOLECULAR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular representation learning underpins computational chemistry and drug discovery, yet standard graph-based approaches struggle with oversmoothing and limited long-range interaction modeling. We explore topological deep learning (TDL) as an alternative, leveraging hypergraphs and cell complexes to incorporate higher-order molecular structures. By systematically comparing these representations against graph-based models, we evaluate their capacity to mitigate oversmoothing and capture richer molecular features. Our empirical analysis across QM9 and ZINC benchmarks demonstrates that topological representations enhance predictive performance, particularly in complex molecular graphs. These findings highlight the potential of TDL for more expressive and structurally aware molecular learning frameworks.

1 INTRODUCTION

Understanding molecular representations is fundamental to computational chemistry and drug discovery, where predicting chemical properties, drug efficacy, and biomolecular interactions relies on the quality of learned representations. Early descriptor-based approaches, such as Extended Connectivity Fingerprints (ECFP) (Rogers & Hahn, 2010) and MACCS keys, are built on predefined structural motifs but struggle to capture molecular flexibility and interactions beyond local patterns. The rise of deep learning techniques has transformed this landscape, shifting from static features to data-driven representations that model the underlying structure-function relationships of molecules.

Language models designed for molecular representations, such as MolBERTa (Balaji et al., 2023) and MoLFormer (Wu et al., 2023), leverage self-supervised learning over SMILES strings, mimicking homologous natural language processing models. These models have demonstrated remarkable improvements in virtual screening and property prediction by capturing chemical syntax and implicit molecular rules. However, the sequential nature of SMILES inherently discards 3D structural information and introduces artifacts from redundant molecular notations, limiting its generalizability (Andronico et al., 2011). To address this, hybrid architectures like MolTrans (Huang et al., 2020) integrate self-attention and convolutional layers, seeking to balance sequence-based expressivity with structural awareness.

Graph neural networks (GNNs) offer an alternative, encoding molecular structures as graphs where atoms serve as nodes and bonds as edges. Architectures like GROVER (Rong et al., 2020), Graphormer (Ying et al., 2021), and GeomGCL (Li et al., 2022) incorporate self-supervised and geometric learning paradigms to enrich molecular graph representations. Yet, despite their success, GNNs face inherent challenges, particularly oversmoothing and oversquashing. oversmoothing causes node embeddings to converge, diminishing their discriminative power and leading to feature homogenization (Qureshi et al., 2023; Rusch et al., 2023; Keriven, 2022). Oversquashing, on the other hand, restricts the capacity to model long-range dependencies by compressing distant molecular interactions into low-dimensional representations. This bottleneck prevents the network from capturing rich molecular features, ultimately hindering downstream performance (Jiang et al., 2021). While transformer-based models and geometric deep learning have been explored as alternatives, these challenges persist, highlighting the need for further innovation in molecular representation learning.

In response to these limitations, topological deep learning (TDL) has emerged as a promising direction, introducing higher-order representations that extend beyond traditional graph structures. These higher-order structures provide a more expressive framework to model molecular interactions, capturing multi-atom relationships that cannot be fully described by pairwise graphs alone. By leveraging hypergraphs and cell complexes, TDL allows interactions between more than two nodes through hyperedges or hierarchical relationships via 2-cells (faces) (Hajij et al., 2023; Papillon et al., 2024; Papamarkou et al., 2024). These structures naturally align with molecular architecture, as molecular rings can be represented as interactions involving three or more atoms, with hyperedges capturing group-level relationships and 2-cell elements modeling closed molecular cycles in hypergraphs or cell complexes.

Building on this idea, (Battiloro et al., 2025) propose representing molecules as combinatorial complexes by integrating rings with 2-cells and functional groups as hyperedges, demonstrating competitive results on the QM9 molecular property prediction benchmark using equivariant topological neural networks.

Given TDL's recent emergence, it still faces several open challenges, characteristic of its early development stage. One of the most pressing issues is the lack of standardized topological benchmarks. Additionally, the molecular representation field lacks a unified framework for defining and representing interactions between higher-order cells. Unlike graph-based models, which benefit from well-established datasets like QM9 (Ruddigkeit et al., 2012) and Open Graph Benchmark (OGB) (Hu et al., 2020), TDL still lacks diverse, large-scale datasets to evaluate model performance across different applications consistently.

Furthermore, there remains a gap in comparative analyses between various TDL architectures and their corresponding cell-lifting strategies (Papamarkou et al., 2024; Bernárdez et al., 2024). This absence of systematic benchmarking impedes the development of universally applicable design principles, a foundation that has already been well-established in classical graph theory.

Addressing these limitations requires not only theoretical advancements but also empirical approaches. In this work, we introduce a novel, systematic approach to validate the relevance of topological information in molecular representations. Our approach follows a structured pipeline. We begin by loading molecular datasets, representing molecules using different topological structures such as graphs, hypergraphs, and cell complexes. Next, we apply specialized neural network architectures tailored to each representation. Additionally, we assess the impact of oversmoothing in these representations by analyzing the behavior of embeddings across model depths, comparing the resilience of graphs, hypergraphs, and cell complexes to this phenomenon. Finally, we perform an extensive benchmarking analysis to compare these representations against traditional graph-based models. Through this process, we aim to assess the impact of the added topological information and determine whether these higher-order features yield measurable improvements in molecular property prediction.

2 Methods

2.1 Datasets

Our study utilizes two well-established molecular datasets, QM9 and ZINC, to evaluate the impact of higher-order molecular representations. QM9 comprises 134,000 small organic molecules containing up to nine heavy atoms, serving as a benchmark for quantum chemistry due to its 19 quantum-mechanical property annotations derived from density functional theory (DFT) calculations (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014). To extend our analysis to larger and more structurally diverse molecules, we incorporate a curated subset of the ZINC database, consisting of 250,000 drug-like molecules with up to 38 heavy atoms (Irwin & Shoichet, 2005). ZINC is widely used in drug discovery and virtual screening due to its pharmacologically relevant chemical diversity.

For the ZINC database benchmark, we define the target variable as a composite molecular property, calculated as:

$$y = \log P - \text{SAS} - \text{cycles} \tag{1}$$

where $\log P$ represents the octanol-water partition coefficient, quantifying hydrophobicity; SAS denotes the synthetic accessibility score, reflecting the ease of synthesis; and cycles count the number of molecular rings containing more than six atoms, which penalizes overly complex structures. This formulation provides a balanced measure of molecular drug-likeness by discouraging excessive lipophilicity, synthetic intractability, and high molecular rigidity.

2.2 Molecular Representations

Molecular graphs. Classical molecular representations leverage molecular graphs $G = (V, E)$, where nodes $v \in V$ correspond to atoms, and edges $e \in E$ represent chemical bonds. This formulation allows encoding of molecular topology using adjacency matrices $A \in \mathbb{R}^{|V|\times|V|}$ and feature matrices $X \in \mathbb{R}^{|V|\times d}$, where each node is associated with a feature vector capturing atomic properties such as atomic number, hybridization, and partial charge. The molecular graph formulation supports graph-based neural networks but remains limited in capturing higher-order molecular interactions beyond local relations.

To extend molecular graphs into higher-order topological domains, we apply a topological lifting that maps the initial graph structure into cell complexes and hypergraphs. This transformation introduces new topological objects, such as faces (2-cells) and hyperedges, while preserving the original graph's atomic and bonding information.

The lifting process constructs these higher-dimensional structures by identifying molecular rings as closed 2-cells in a cell complex or as multi-node hyperedges in a hypergraph. During this mapping, node-level features are propagated to these newly created topological entities, while additional domain-specific descriptors (e.g., ring size, aromaticity, or heteroatom composition) are computed to capture higher-order molecular properties.

Cell complexes. To enhance molecular representation, we extend graphs to cell complexes, a topological generalization that incorporates higher-dimensional structures. A cell complex C consists of nodes (0-cells), edges (1-cells), and faces (2-cells), enabling hierarchical part-whole relationships. The boundary operator $\partial_k : C_k \to C_{k-1}$ formalizes these relationships, where edges are defined by their boundary nodes, and faces by their boundary edges:

$$\partial_2(f) = \sum_{e\in f} w_e e, \quad \partial_1(e) = \sum_{v\in e} w_v v, \tag{2}$$

where w_e and w_v are orientation coefficients. Molecular rings, represented as 2-cells, preserve geometric and chemical constraints, providing an enriched structural descriptor beyond edge-based connectivity.

Hypergraphs. Hypergraphs $H = (V, E_H)$ generalize graphs by introducing hyperedges $e_h \in E_H$ that connect multiple nodes simultaneously, including edges which are limited to 2 nodes connections. This structure enables a more expressive modeling of multi-atom interactions, particularly in conjugated and delocalized electron systems. The incidence matrix $H \in \mathbb{R}^{|V|\times|E_H|}$ encodes node-hyperedge relationships, where each entry H_{ve} indicates node membership in a hyperedge.

Each topological domain is enriched with domain-specific features, capturing various molecular properties at different structural levels. At the node level, atomic descriptors such as atomic number, degree, formal charge, hybridization state, aromaticity, atomic mass, and chirality provide a detailed characterization of individual atoms within the molecular graph. At the edge level, bond-specific properties including bond type, conjugation status, and stereochemistry encode the nature of atomic connectivity, ensuring accurate representation of molecular bonding interactions. Beyond these conventional features, higher-order structures such as hyperedges and 2-cells, which represent molecular rings, contribute additional descriptors such as ring size, aromaticity, presence of heteroatoms, saturation status, hydrophobicity, electrophilicity, nucleophilicity, and polarity. These higher-order attributes enable a richer characterization of molecular complexity, allowing for a more detailed understanding of chemical and structural properties in molecular representations.

2.3 MODELS

To effectively model higher-order molecular representations, we employ architectures that extend beyond conventional GNNs to incorporate hypergraphs and cell complexes, capturing richer structural and relational information. It is important to clarify that we use simple GNN architectures as a raw baseline, focusing on their fundamental performance without introducing additional complexities.

At the foundation, the Graph Convolutional Network (GCN) (Kipf & Welling, 2016) serves as a baseline, leveraging spectral graph convolutions to aggregate information from local atomic neighborhoods. Formally, given a molecular graph $G = (V, E)$ with node features X, the propagation rule for GCN is:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{3}$$

where $\tilde{A} = A + I$ is the adjacency matrix with self-loops, $\tilde{D}$ is its diagonal degree matrix, $H^{(l)}$ represents node embeddings at layer l, and $W^{(l)}$ is the learnable weight matrix.

To capture multi-atom interactions beyond pairwise relations, the Hypergraph Message Passing Neural Network (HMPNN) (Heydari & Livi, 2022) extends the graph structure by introducing hyperedges. Given a hypergraph $H = (V, E_H)$, where hyperedges e_h connect multiple nodes simultaneously, message passing is defined by:

$$H_v^{(l+1)} = \sigma\left(\sum_{e_h \in E_H} \frac{1}{|e_h|} W_{e_h}^{(l)} H_{e_h}^{(l)} + W_v^{(l)} H_v^{(l)}\right) \tag{4}$$

where messages are weighted across hyperedges and nodes by their respective learnable matrix $W_e^{(l)}$, $W_v^{(l)}$, allowing efficient encoding of higher-order molecular dependencies.

For even more expressive modeling, CW Networks (CWN) (Bodnar et al., 2021) extend GNNs to operate over cell complexes. These networks generalize message passing beyond nodes and edges to include faces (2-cells), enabling hierarchical information flow. Given a cell complex C , the update rule for cell embeddings is defined as:

$$H_c^{(l+1)} = \sigma\left(\sum_{c' \in \mathcal{N}(c)} W_c^{(l)} H_{c'}^{(l)}\right) \tag{5}$$

where $\mathcal{N}(c)$ denotes the neighborhood of cell c, including lower-dimensional and higher-dimensional adjacent structures across different ranks. This enables molecular rings, represented as 2-cells, to propagate features in a way that captures topological constraints and part-whole relationships.

2.4 EXPERIMENTAL SETUP

To investigate oversmoothing and oversquashing effects, we conducted multiple experiments, exploring all possible combinations for a vast number of layers and hidden channels. Beyond these two key hyperparameters, we also explored additional configurations, adjusting various parameters. All these combinations are detailed detailed in A.

To systematically evaluate these models, we benchmark performance using standard regression metrics, including the coefficient of determination (R^2), Spearman Rank Correlation, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Relative Error. These metrics ensure robust comparison across datasets, quantifying the effectiveness of higher-order representations in molecular property prediction.

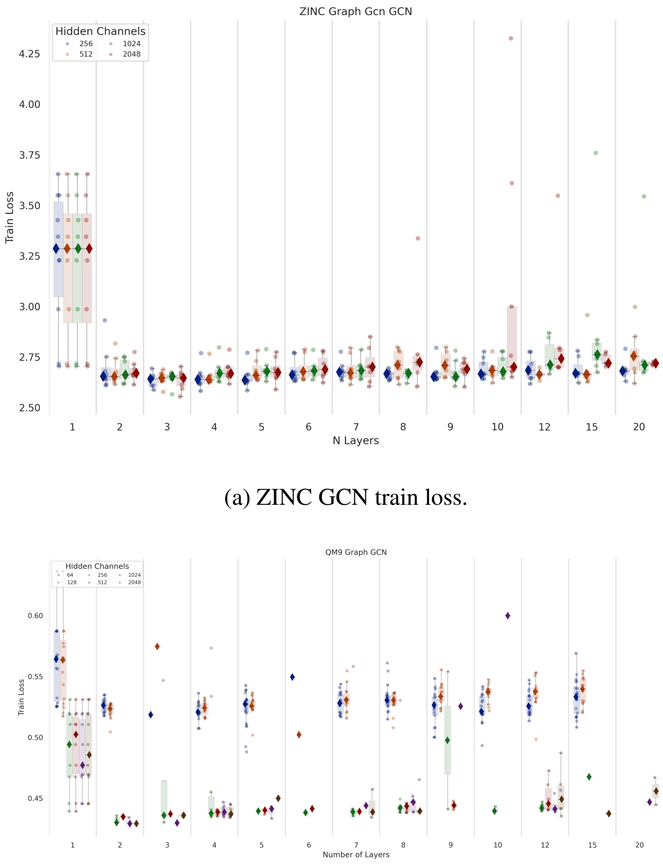
3 Results and discussion

3.1 Oversmoothing in molecular data

GCNs suffer from oversmoothing when the depth of the network increases, leading node embeddings to become indistinguishable. While this phenomenon is well-documented in traditional graph tasks, molecular graphs exhibit distinct oversmoothing characteristics due to their inherent structural constraints and localized connectivity.

For molecular datasets such as ZINC and QM9, our results show that oversmoothing manifests at different depths depending on the dataset. On ZINC, a regression task with complex molecular structures, oversmoothing becomes apparent beyond $N = 10$, where train loss begins to increase. This suggests that deeper layers propagate information excessively, reducing the network's ability to distinguish molecular features. Conversely, in QM9, which involves quantum chemical property prediction, oversmoothing occurs more gradually, with a noticeable but less severe degradation at $N \geq 9$ (Figure 1.

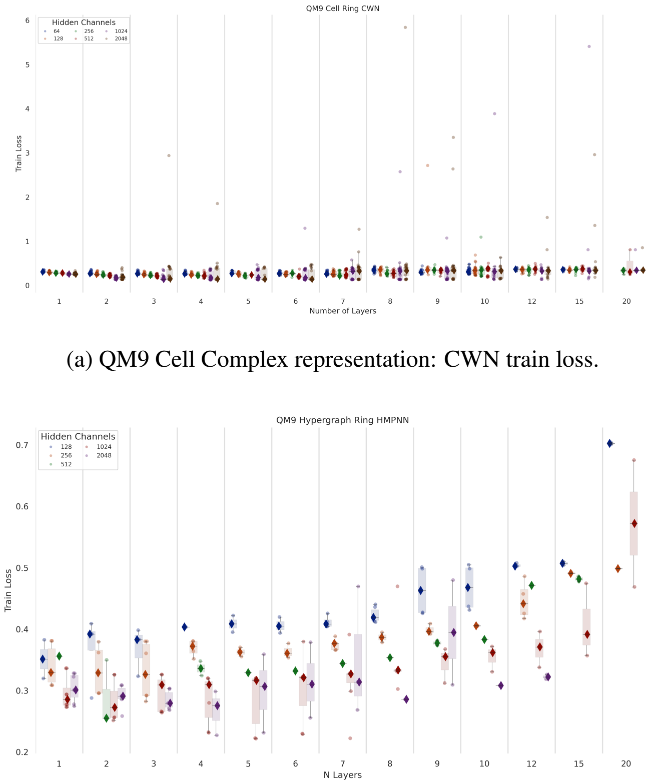
Unlike traditional social or citation networks, it is thought that molecular graphs are inherently sparse and chemically constrained, which should impact the effectiveness of message passing. However, in traditional networks, oversmoothing typically occurs at 3-5 layers, whereas we observed that molecular graphs can sustain effective message passing up to 3-8 layers before feature collapse.



(a) ZINC GCN train loss.

(b) QM9 GCN train loss.

Figure 1: Comparison of GCN train loss on ZINC and QM9 datasets.



(a) QM9 Cell Complex representation: CWN train loss.

(b) QM9 Hypergraph representation: HMPNN train loss.

Figure 2: Comparison of GCN train loss on ZINC and QM9 datasets.

3.2 Molecular graphs show fundamentally different connectivity potentially affecting information propagation

While oversmoothing has been extensively studied in social and citation networks, molecular graphs introduce a distinct structural paradigm that reshapes how this phenomenon unfolds. In order to gain insight in the information propagation of these graphs, we analyze the Laplacian spectrum across datasets with differing topologies—specifically, citation graphs (Cora and PubMed) and molecular datasets (QM9 and ZINC).

The Laplacian gap, a measure of eigenvalue separation, serves as a spectral indicator of a graph's connectivity and the ease with which signals propagate. A larger gap suggests better-separated eigenvalues, often correlating with reduced oversmoothing and more stable learning dynamics. In contrast, a near-zero or negative gap may indicate spectral degeneracy, where nodes converge to indistinguishable embeddings within a few layers.

In our analysis, Cora exhibits a strikingly small Laplacian gap (-5.16×10^{-14}), suggesting that at least two eigenvalues are nearly identical. This is characteristic of dense community structures and scale-free connectivity in citation networks, where a handful of highly connected nodes dominate information flow. Consequently, oversmoothing occurs rapidly, as signals diffuse too uniformly across the network, leading to an early collapse of node representations. PubMed, while still susceptible to oversmoothing, presents a slightly larger gap (-0.0275), implying marginally improved spectral separation, possibly due to its more dispersed connectivity compared to Cora.

Molecular graphs, by contrast, exhibit fundamentally different spectral properties. QM9 presents a significantly larger Laplacian gap (-0.3559, with a standard deviation of 0.1304), suggesting that its graph structures maintain a broader eigenvalue distribution. This aligns with the rigid connectivity imposed by atomic valency constraints and the presence of cyclic motifs, particularly aromatic rings, which introduce additional structural regularity. These topological features reshape the Laplacian spectrum, often mitigating the rapid spectral collapse observed in citation networks. The ZINC dataset, which contains more complex drug-like molecules, exhibits an intermediate gap magnitude (-0.0468), reflecting a balance between larger molecular structures and diverse connectivity patterns. While still susceptible to oversmoothing in deep architectures, the presence of rings and functional groups offers alternative pathways for information flow, delaying the onset of feature collapse compared to social graphs.

Table 1: Laplacian Gap Results for Different Datasets

Dataset	**Gap**	**Std Dev**
PubMed	-0.02752	—
Cora	-5.16e-14	—
QM9	-0.3559	0.1304
ZINC	-0.04677	0.02675

3.3 Topological representations are able to capture better representations than their graph counterparts

Topological networks display better results in molecular endpoint prediction. Our results show that topological representations, such as cell complexes and hypergraphs, provide enhanced molecular embeddings compared to traditional graph-based approaches. The ability to model higher-order molecular interactions, including aromaticity, conjugation, and multi-atom dependencies, enables more effective feature propagation and mitigates the limitations of standard GNNs.

The comparison of endpoint performance across different architectures, GCNs, CWNs, and HMPNNs reveals significant differences. While GCNs exhibit gradual degradation in performance as depth increases, CWNs and HMPNNs show varied improvements. CWNs, which encode rings as 2-cells, initially outperform GCNs but exhibit instability at high depths. HMPNNs maintain consistent performance across depths, suggesting that hypergraph message passing effectively preserves long-range molecular interactions (Figure 1).

TNNs and their liftings scale better with width. The number of hidden channels influences smoothing behavior significantly. GCNs show a rapid decline in feature diversity as hidden channels increase, leading to uniform embeddings that reduce predictive power. In contrast, CWNs and HMPNNs scale more effectively with increased feature dimensionality, maintaining distinct representations across layers. This suggests that incorporating topological structures provides a more stable inductive bias that allows models to leverage higher-dimensional feature spaces efficiently.

Oversmoothing seems mitigated by topological learning in molecules. Network depth plays a critical role in determining model stability. GCNs exhibit progressive oversmoothing beyond 8 layers, where node representations become nearly indistinguishable. CWNs experience increasing variance in performance beyond 10 layers, likely due to overcomplicated message pathways within high-dimensional complexes. HMPNNs, however, exhibit robust behavior across depths, leveraging hyperedges to prevent feature collapse and maintain meaningful differentiation between molecular structures. The resilience of hypergraph-based methods suggests that they offer a more scalable solution for deep molecular learning tasks.

Topological representations mitigate oversmoothing but still struggle with molecular complexity. The fitting behavior observed across QM9 and ZINC highlights key differences in how graph-based and topological representations generalize across molecular structures. In QM9, where molecular graphs are smaller and structurally constrained, GCNs exhibit moderate oversmoothing at deeper layers, while CWNs and HMPNNs maintain stable performance, effectively capturing ring structures and long-range dependencies. However, in ZINC, which consists of larger and more chemically diverse molecules, GCNs struggle with overfitting, and CWNs display high variance across depths, indicating that simple graph convolutions may be insufficient to model complex molecular interactions. HMPNNs, while effective in QM9, show signs of instability in ZINC, suggesting that hypergraph message passing requires additional constraints when applied to highly heterogeneous molecular graphs. These findings emphasize the data specific nature of topological deep learning methods. In QM9, structured topologies like CWNs and HMPNNs enable better feature propagation without excessive smoothing, aligning well with the dataset's limited molecular complexity. Conversely, in ZINC, the increased connectivity and variability of molecular structures introduce challenges for both graph-based and topological models, with deeper networks prone to over-parameterization (see figures Appendix C).

Embedding similarity shows richer representations learned trough topological networks. For QM9, we observe that the Euclidean distance between GCN and CWN embeddings varies significantly with network depth and hidden dimensionality. Notably, distances are highest at intermediate depths (4–8 layers) and larger hidden channels (1024–2048), suggesting that CWNs produce distinct representations that diverge from traditional graph embeddings, particularly when richer feature capacity is available. The comparison between HMPNN and CWN further reinforces this, showing substantial shifts in embedding space at larger depths, indicating that hypergraph-based models introduce non-trivial transformations to molecular representations. Conversely, distances between HMPNN and GCN remain relatively lower, implying that while hypergraphs capture additional topological features, their representations still retain some similarity to graph-based encodings.

For ZINC, the embedding distances display a different pattern. Here, the divergence between GCN and CWN embeddings is smaller, particularly at greater depths, implying that deeper CWNs may be converging towards GCN-like representations in highly complex molecular graphs. However, distances remain relatively high in the low-depth, high-dimensional regime, indicating that CWNs initially leverage additional topological structure before eventual feature collapse at depth. The hypergraph-based HMPNN, in contrast, demonstrates significant shifts from both GCN and CWN embeddings across all configurations, particularly in the mid-depth, high-channel range. This suggests that hypergraph-based methods offer unique structural representations in larger molecular graphs, diverging more markedly from traditional graph-based methods than they do in QM9. Complete results for similarity comparisons between representations can be found at Appendix B.

4 CONCLUSIONS

Our findings highlight the advantages of topological deep learning in molecular representation learning, demonstrating that higher-order structures such as hypergraphs and cell complexes provide richer molecular embeddings compared to conventional graph-based models. By capturing multi-atom interactions beyond pairwise connectivity, TDL-based architectures mitigate key limitations of graph neural networks, including oversmoothing and restricted information propagation.

Empirical evaluations across QM9 and ZINC reveal that topological representations enhance predictive performance in molecular property tasks, particularly in capturing long-range dependencies and preserving structural information. Hypergraph message passing networks maintain more stable performance across network depths, while combinatorial complex-based architectures, such as CWNs, demonstrate competitive results but exhibit increased variance in complex molecular graphs. These observations suggest that the choice of topological domain significantly influences model expressivity and generalization.

Despite their advantages, topological representations introduce additional modeling complexity and mild computational overheads, requiring further optimization for large-scale applications. Standardized benchmarks and a unified framework for topological molecular learning remain open challenges, necessitating systematic evaluations across diverse datasets.

MEANINGFULNESS STATEMENT

A meaningful representation of life should capture all possible degrees of complexity of the object it is modelling. Thus, we see topological representations as a leap forward in the modelling capacity that graph representations offer in molecular representation. This work aims to establish a first comparison from an empirical lens of such representations and architectures over more traditional counterparts with a will of pushing forward research in this field.

REFERENCES

Alessio Andronico, Arlo Randall, Ryan W Benz, and Pierre Baldi. Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress. *J Chem Inf Model*, 51 (4):760–776, March 2011.

Suryanarayanan Balaji, Rishikesh Magar, Yayati Jadhav, , and Amir Barati Farimani. GPT-MolBERTa: GPT molecular features language model for molecular property prediction. *ArXiv*, abs/2310.03030, 2023.

Claudio Battiloro, Ege Karaismailoglu, Mauricio Tec, George Dasoulas, Michelle Audirac, and Francesca Dominici. E(n) equivariant topological neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Guillermo Bernárdez, Lev Telyatnikov, Marco Montagna, Federica Baccini, Mathilde Papillon, Miquel Ferriol-Galmés, Mustafa Hajij, Theodore Papamarkou, Maria Sofia Bucarelli, Olga Zaghen, Johan Mathe, Audun Myers, Scott Mahan, Hansen Lillemark, Sharvaree Vadgama, Erik Bekkers, Tim Doster, Tegan Emerson, Henry Kvinge, Katrina Agate, Nesreen K Ahmed, Pengfei Bai, Michael Banf, Claudio Battiloro, Maxim Beketov, Paul Bogdan, Martin Carrasco, Andrea Cavallo, Yun Young Choi, George Dasoulas, Matouš Elphick, Giordan Escalona, Dominik Filipiak, Halley Fritze, Thomas Gebhart, Manel Gil-Sorribes, Salvish Goomanee, Victor Guallar, Liliya Imasheva, Andrei Irimia, Hongwei Jin, Graham Johnson, Nikos Kanakaris, Boshko Koloski, Veljko Kovač, Manuel Lecha, Minho Lee, Pierrick Leroy, Theodore Long, German Magai, Alvaro Martinez, Marissa Masden, Sebastian Mežnar, Bertran Miquel-Oliver, Alexis Molina, Alexander Nikitin, Marco Nurisso, Matt Piekenbrock, Yu Qin, Patryk Rygiel, Alessandro Salatiello, Max Schattauer, Pavel Snopov, Julian Suk, Valentina Sánchez, Mauricio Tec, Francesco Vaccarino, Jonas Verhellen, Frederic Wantiez, Alexander Weers, Patrik Zajec, Blaž Škrlj, and Nina Miolane. ICML topological deep learning challenge 2024: Beyond the graph domain, 2024.

Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yuguang Wang, Pietro Lio, Guido F Montufar, and Michael Bronstein. Weisfeiler and lehman go cellular: CW networks. *Advances in neural information processing systems*, 34:2625–2640, 2021.

Mustafa Hajij, Ghada Zamzmi, Theodore Papamarkou, Nina Miolane, Aldo Guzmán-Sáenz, Karthikeyan Natesan Ramamurthy, Tolga Birdal, Tamal K. Dey, Soham Mukherjee, Shreyas N. Samaga, Neal Livesay, Robin Walters, Paul Rosen, and Michael T. Schaub. Topological deep learning: Going beyond graph data, 2023.

Sajjad Heydari and Lorenzo Livi. Message passing neural networks for hypergraphs. In *International Conference on Artificial Neural Networks*, pp. 583–592. Springer, 2022.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for machine learning on graphs. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133. Curran Associates, Inc., 2020.

Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. MolTrans: Molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 10 2020.

John J Irwin and Brian K Shoichet. ZINC- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, 13:1–23, 2021.

Nicolas Keriven. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *Advances in Neural Information Processing Systems*, 35:2268–2281, 2022.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong. GeomGCL: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 4541–4549, 2022.

Theodore Papamarkou, Tolga Birdal, Michael M Bronstein, Gunnar E Carlsson, Justin Curry, Yue Gao, Mustafa Hajij, Roland Kwitt, Pietro Lio, Paolo Di Lorenzo, et al. Position: Topological deep learning is the new frontier for relational learning. In *Forty-first International Conference on Machine Learning*, 2024.

Mathilde Papillon, Guillermo Bernárdez, Claudio Battiloro, and Nina Miolane. TopoTune: A framework for generalized combinatorial complex neural networks. *arXiv preprint arXiv:2410.06530*, 2024.

Shaima Qureshi et al. Limits of depth: Over-smoothing and over-squashing in gnns. *Big Data Mining and Analytics*, 7(1):205–216, 2023.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

Fang Wu, Dragomir Radev, and Stan Z Li. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5312–5320, 2023.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

A Hyperparameter Tuning

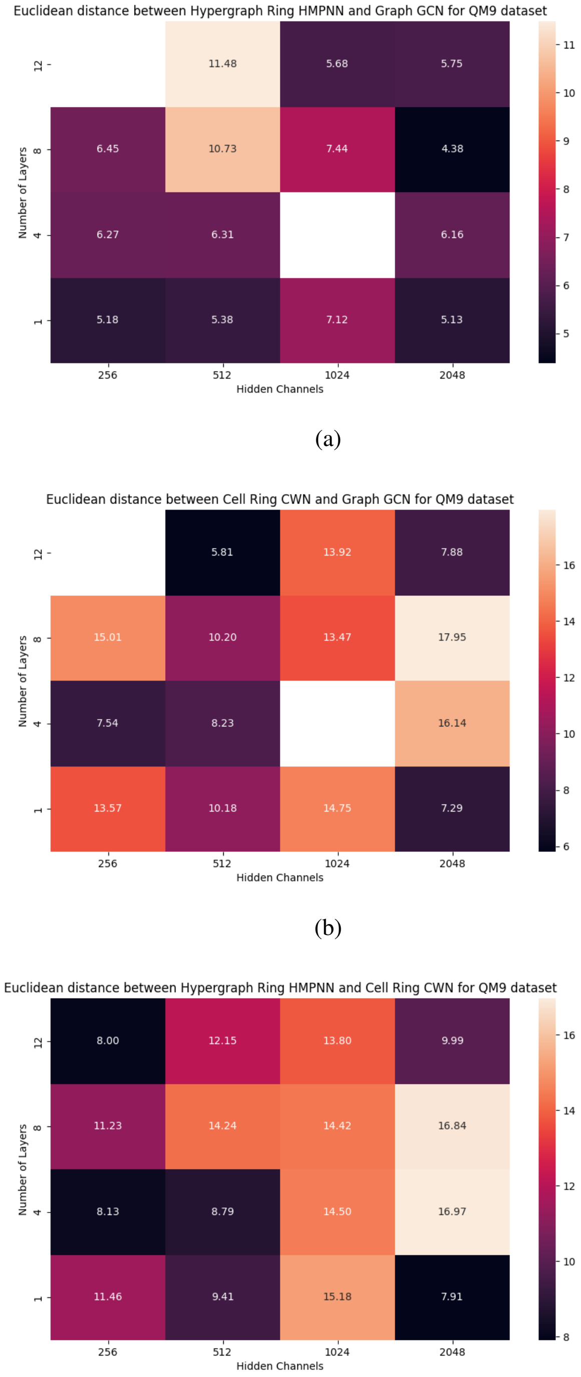
In order to do a fair comparison between the different dataset, we have applied the same train, validation, and test split in both QM9 and ZINC datasets: 70%, 5% and 15%, respectively. This has been chosen after benchmarking different splits. All runs were executed on NVIDIA H100 GPUs.

Table 2: Hyperparameter configurations used in models.

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	[0.001, 0.0001]
Weight Decay	1e-5
Batch size	[4, 8, 16, 32, 64, 128, 256]
Epochs	1000
Early Stopping Patience	50
Number of Layers	[1 - 10, 12, 15, 20]
Hidden Channels	[256, 512, 1024, 2048]

This systematic exploration ensured that each model was effectively trained for its respective representation.

B Model Embedding Comparison


Euclidean distance between Hypergraph Ring HMPNN and Graph GCN for QM9 dataset

(a)

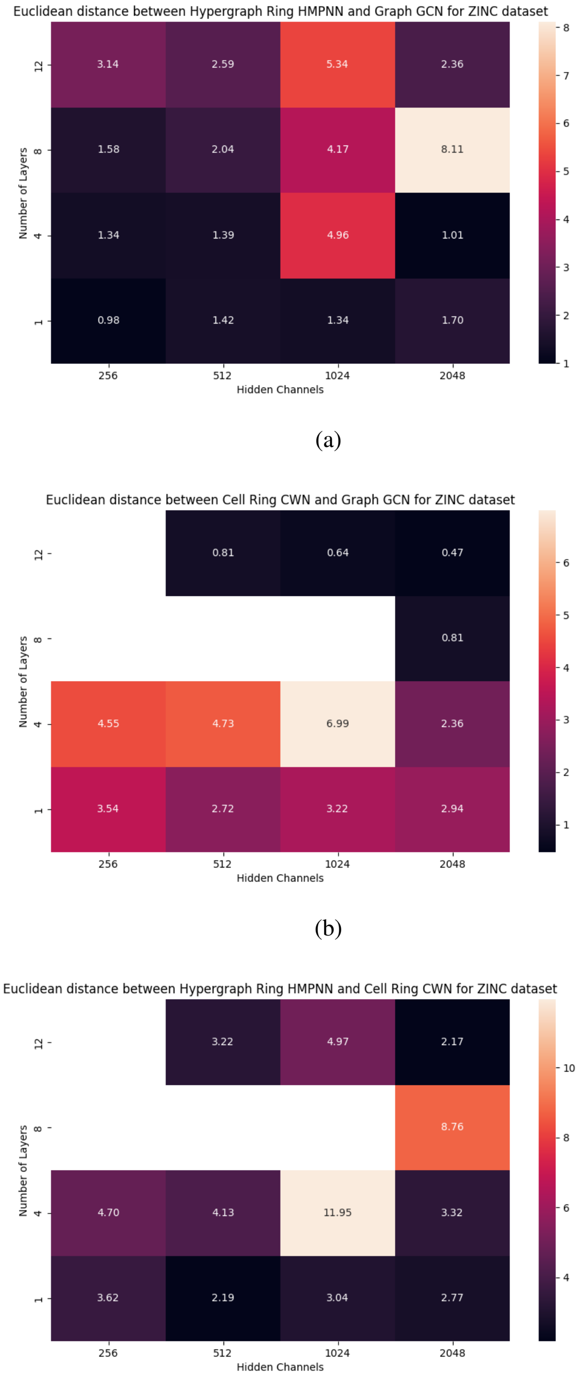
Euclidean distance between Cell Ring CWN and Graph GCN for QM9 dataset

(b)

Euclidean distance between Hypergraph Ring HMPNN and Cell Ring CWN for QM9 dataset

(c)

Figure 3: Comparison of embeddings for five training points from the QM9 dataset across three different topological liftings and their corresponding models.



(a)

(b)

(c)

Figure 4: Comparison of embeddings for five training points from the ZINC dataset across three different topological liftings and their corresponding models.

C Model Results

C.1 Hypergraph Message Passing Neural Network Results

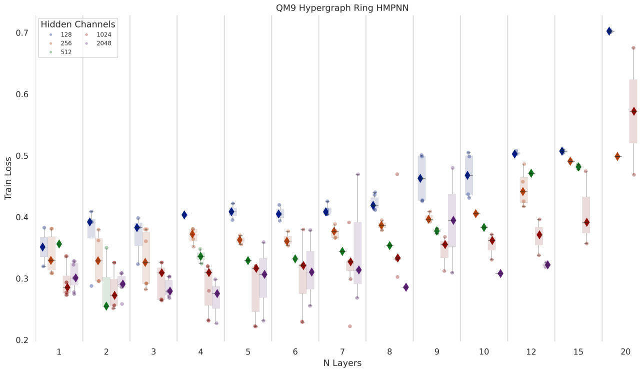


Figure 5: QM9 Hypergraph representation: Train Loss

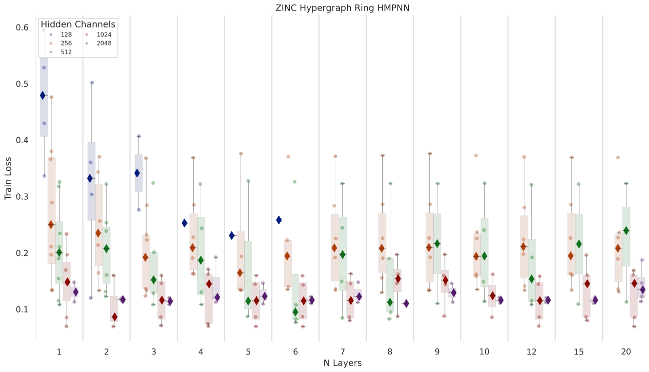


Figure 6: ZINC Hypergraph representation: Train Loss

C.2 CW Network Results

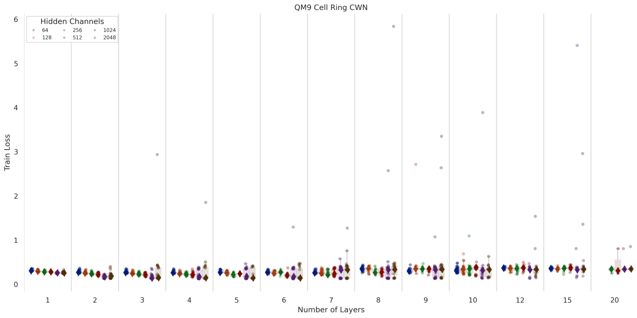


Figure 7: QM9 cell complex representation: Train Loss

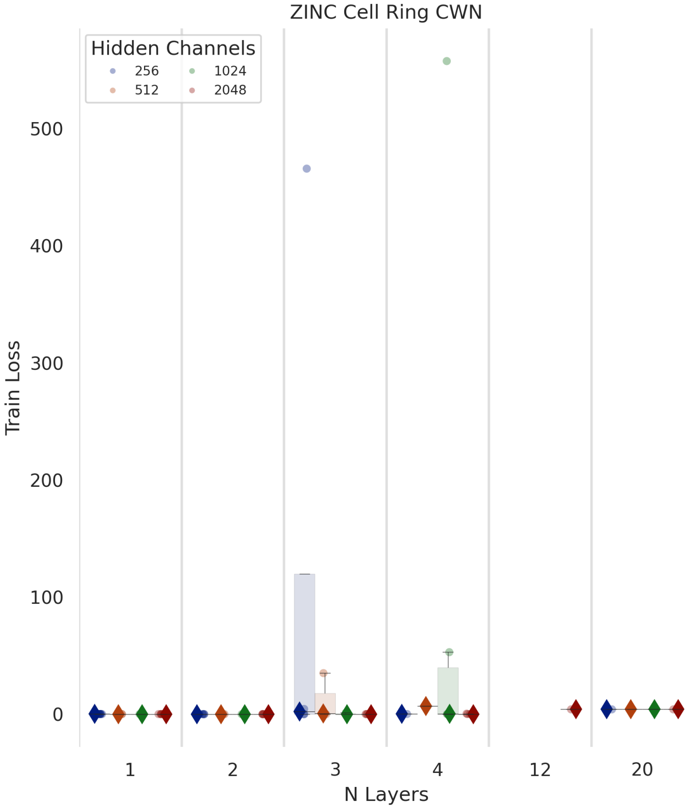


Figure 8: ZINC cell complex representation: Train Loss