# Research Statement

Cheng Xin
Computer Science Department, Rutgers University
cx122@cs.rutgers.edu

## Research Overview

My research develops geometric and topological foundations for machine learning and artificial intelligence. Working at the intersection of topological data analysis and AI, I create mathematically rigorous frameworks that address fundamental challenges in reliability, interpretability, and structural understanding of machine learning systems. My approach systematically integrates classical mathematical tools—algebraic topology, computational geometry, and geometric analysis—with contemporary machine learning methodologies. While modern machine learning achieves substantial empirical success, many systems lack theoretical guarantees and fail to capture intrinsic geometric and topological structures present in data. My research addresses these limitations by developing principled mathematical frameworks that provide both theoretical insights and practical improvements in AI system performance and trustworthiness.

## Research Philosophy and Approach

Geometric and topological approaches capture essential structural properties that resist conventional statistical analysis. Geometry provides principled frameworks for understanding intrinsic data relationships, while topology reveals stable, qualitative features persisting across scales. These mathematical tools prove particularly valuable for scientific applications where data exhibits complex non-Euclidean structure. My research trajectory demonstrates commitment to developing rigorous mathematical theory while maintaining focus on practical applications—evidenced by publications spanning theoretical computational topology venues (SoCG, JACT) and applied machine learning conferences (ICML, NeurIPS, CVPR).

## Major Research Contributions

### 1. Foundational Theory: Multiparameter Persistent Homology

**Research Challenge:** Classical persistent homology analyzes data varying along single parameters, but many applications require tracking topological features across multiple parameters simultaneously. However, multiparameter persistence lacks the algebraic structure of single-parameter cases—computing interleaving distances is NP-hard, and there is no complete discrete invariants.

**Contributions:** During my doctoral research, I developed algorithmic solutions to core computational challenges in multiparameter persistent homology. These contributions formed the foundation of my PhD dissertation [7]:

1. **Efficient Decomposition Algorithm:** I designed algorithms for computing decompositions of multiparameter persistence modules [4], extending classical persistence algorithms through matrix reduction techniques handling partial order structures. This work provides computational foundations for practical multiparameter analysis.
2. **Polynomial-Time Distance Computation:** For 2-D interval decomposable modules—a theoretically and practically important class—I created polynomial-time algorithms for computing bottleneck distance [3], introducing concepts of "effective intersections" and "trivializable intersections" enabling exact distance computation.

**Impact:** These works, published in **SoCG 2018** and **JACT 2022**, established computational methods for the multiparameter TDA and remain among the most efficient known solutions. Subsequent NP-hardness results by other researchers have confirmed the computational challenges.

## 2. Applied Theory: Topologically-Enhanced Graph Neural Networks

**Research Challenge:** Graph Neural Networks achieve strong performance but face fundamental limitations in capturing complex structural information due to reliance on local message-passing. Many graphs exhibit rich topological structures—cycles, voids, higher-dimensional features—encoding essential semantic information but remaining invisible to standard architectures.

**Contributions:** I developed general frameworks integrating topology with deep learning:

1. **GRIL (PMLR 2023):** I introduced Generalized Rank Invariant Landscape [8], a vectorization framework achieving improved expressivity over traditional rank invariant-based representations while maintaining stability and differentiability required for gradient-based learning. The framework enables end-to-end training of topology-aware neural networks.
2. **Applications and Validation:** Experiments demonstrate consistent improvements over state-of-the-art methods, particularly in scenarios requiring understanding of complex structural patterns, with promising results in scientific applications where global structural properties prove crucial for predictions.

**Impact:** GRIL was selected for oral presentation at the ICML 2024 TAG-ML workshop, establishing it as a notable contribution to topologically-enhanced deep learning.

## 3. Interpretable AI: Topological Explainability Framework

**Research Challenge:** As AI systems grow increasingly complex, interpretability becomes critical for trust and adoption in high-stakes applications. Current interpretability methods (attention mechanisms, saliency maps) are heuristic and post-hoc, lacking theoretical guarantees. They indicate "where" models attend but not "why" in terms of structural reasoning.

**Contributions:** My work on **TopInG [6] (ICML 2025)** develops principled mathematical frameworks for topological interpretability:

1. **Theoretical Foundation:** I established rigorous frameworks for topological interpretability, introducing **topological discrepancy** quantifying statistical differences through 1-Wasserstein distance between graph distributions with respect to topological structure, with efficient approximating algorithm and theoretical guarantees showing optimization recovers ground truth under certain conditions.
2. **Architecture Innovation:** The framework models GNN decision-making as a "persistent rationale generation process," using differentiable topological representations to track statistically significant structural differences between decision-relevant and irrelevant subgraphs.
3. **Performance:** TopInG achieves up to **20% improvement** over state-of-the-art methods on both prediction accuracy and interpretation quality. The framework successfully handles variable and complex rationale subgraphs where previous methods encounter difficulties, demonstrating improved robustness to spurious correlations.

**Impact:** TopInG provides principled, theoretically-grounded frameworks for topological explainability in graph neural networks, with applications to drug discovery, materials science, and domains where understanding model decisions proves as critical as accuracy.

### 4. Geometric Innovation: Non-Euclidean Representations

**Research Challenge:** Classical dimensionality reduction techniques like MDS assume Euclidean structure, but real-world data often exhibits inherently non-Euclidean or non-metric properties. Forcing such data into Euclidean space causes information loss and can lead to paradoxical behaviors like increased error with higher embedding dimensions. Similarly, fundamental results in high-dimensional geometry like the Johnson-Lindenstrauss lemma have been limited to Euclidean spaces.

**Contributions:** I have developed theoretical frameworks extending classical geometric methods to non-Euclidean settings:

1. **Neuc-MDS (NeurIPS 2024):** I extended classical MDS theory from Euclidean to pseudo-Euclidean spaces [1]. I introduced bilinear forms to unify representation of positive and negative distance information, enabling effective use of negative eigenvalues carrying crucial "non-Euclidean" information typically discarded by classical methods. I designed efficient algorithms for jointly optimizing eigenvalue selection and bilinear form choice, with comprehensive theoretical analysis of error bounds and asymptotic behavior.
2. **Johnson-Lindenstrauss Beyond Euclidean Geometry (NeurIPS 2025):** I extended the celebrated Johnson-Lindenstrauss lemma—a cornerstone of dimensionality reduction—to general metric spaces beyond Euclidean geometry [2]. This work establishes fundamental theoretical guarantees for dimensionality reduction in non-Euclidean settings, with applications to hyperbolic embeddings, graph metrics, and other non-Euclidean structures prevalent in modern machine learning.

**Impact:** These works provide theoretical foundations for studying non-Euclidean embeddings in modern ML models, with applications to graph representation learning, hyperbolic neural networks, recommendation systems, and biological networks. The extension of the JL lemma opens new possibilities for efficient processing of non-Euclidean data in high-dimensional settings.

## Future Research Directions

Building upon these foundations, I plan to expand topological and geometric methods as frameworks for developing stable, interpretable, and trustworthy AI systems:

### 1. Topological Representations for Foundation Models and Sequential Data

Large language models and foundation models achieve remarkable capabilities through complex internal representations whose structure remains poorly understood. A fundamental challenge arises when these models process data that lies on or near an underlying manifold in high-dimensional space: how do we optimally order and process observations to reveal the manifold's intrinsic structure?

I am developing a unified mathematical framework for **learnable topological representations of sequential data** through the lens of filtrations on data manifolds. The key insight is that sequential processing—whether in autoregressive language models, visual reasoning systems, or diffusion models—fundamentally involves studying an underlying manifold incrementally, from local neighborhoods to global structure. By formalizing this through soft ordering functions $\alpha : X \to [0, 1]$ that induce filtrations $M_\alpha$, we can track topological evolution via persistent homology and optimize these orderings for specific tasks.

This framework addresses three core theoretical challenges: **1. Learnable Filtration Construction:** Designing parameterized families $M_\theta(X, t)$ that are differentiable, topologically expressive, and respect meaningful structure (temporal coherence, causal precedence). **2. Differentiable Topological Vectorization:** Creating stable, informative vector representations from persistence diagrams that enable gradient-based optimization. **3. Efficient Approximation:** Developing near-linear time algorithms ($O(n^{1+\delta})$) for computing topological features and gradients, enabling practical deployment at scale.

Beyond representation analysis, topological methods provide new perspectives on causal inference. Current approaches focus on graph structures, but topological stability could identify features persisting across interventions. My multiparameter persistence theory proves particularly valuable for analyzing data varying along multiple causal or semantic dimensions simultaneously, connecting topological stability with causal invariance for robust causal discovery in complex AI systems.

### 2. Geometric and Topological Foundations for Scientific Discovery

Scientific data exhibits rich geometric and topological structure beyond Euclidean assumptions, requiring principled mathematical frameworks for analysis and prediction. I plan to develop integrated geometric-topological methods tailored to specific scientific domains where data manifests complex non-Euclidean properties.

**Non-Euclidean Geometric Representations:** Building on my work extending classical dimensionality reduction and the Johnson-Lindenstrauss lemma to non-Euclidean spaces, I will develop geometric frameworks for molecular conformations in pseudo-Euclidean spaces, hyperbolic embeddings for hierarchical biological networks, and metric geometry approaches for physical simulations where data naturally resides in curved spaces.

**Topological Analysis of Complex Systems:** Persistent homology provides powerful tools for understanding multi-scale structure in scientific data. I plan to apply topological methods to protein interaction networks (identifying functionally important cycles and voids), materials science (characterizing porous structures and phase transitions), and dynamical systems (detecting bifurcations through topological signatures). The combination of my multiparameter persistence theory with domain-specific filtrations enables richer characterizations than single-parameter approaches.

**Integrated Geometric-Topological Frameworks:** The true power emerges from combining geometric and topological perspectives. For instance, studying molecular dynamics requires both geometric understanding of conformational spaces and topological tracking of persistent structural motifs across trajectories. My experience with the DL3DV-10K [5] dataset for 3D vision demonstrates the value of such integration, which I will extend to broader scientific applications including protein folding, materials discovery, and climate modeling.

This research direction aims to establish geometric and topological data analysis as fundamental tools for scientific machine learning, providing both theoretical guarantees and practical computational methods for domains where traditional Euclidean assumptions fail.

## Collaboration and Broader Impact

My interdisciplinary background—training under computational geometers (Tamal K. Dey) and current collaboration with experts in network computing (Jie Gao) and non-Euclidean geometry (Feng Luo)—positions me to build bridges between abstract mathematics and practical AI applications. I maintain productive collaborations with researchers at Purdue, Rutgers, and other

institutions, contributing to NSF-funded projects including the AI Institute for Agent-based Cyber Threat Intelligence.

By systematically developing rigorous geometric and topological foundations while maintaining focus on practical applications, my research aims to advance the reliability, interpretability, and theoretical understanding of modern machine learning systems.

## References


[1] Chengyuan Deng, Jie Gao, Kevin Lu, Feng Luo, Hongbin Sun, and **Cheng Xin**†. Neucmds: Non-euclidean multidimensional scaling through bilinear forms. In *Advances in Neural Information Processing Systems **(NeurIPS)***, volume 37, pages 121539–121569, 2024.

[2] Chengyuan Deng, Jie Gao, Kevin Lu, Feng Luo, and **Cheng Xin**†. Johnson-lindenstrauss lemma beyond euclidean geometry. 2025. **(NeurIPS)**.

[3] Tamal K. Dey and **Cheng Xin**†. Computing Bottleneck Distance for 2-D Interval Decomposable Modules. In *34th International Symposium on Computational Geometry **(SoCG)***, volume 99 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 32:1–32:15, 2018.

[4] Tamal K. Dey and **Cheng Xin**†. Generalized persistence algorithm for decomposing multiparameter persistence modules. *Journal of Applied and Computational Topology*, 6(3):271–322, Sep 2022.

[5] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, **Cheng Xin**, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition **(CVPR)***, pages 22160–22169, 2024.

[6] **Cheng Xin**, Fan Xu, Xin Ding, Jie Gao, and Jiaxin Ding. Toping: Topologically interpretable graph learning via persistent rationale filtration. In *the 42nd International Conference on Machine Learning **(ICML)***, 2025.

[7] **Cheng Xin**. *Decomposition and Stability of Multiparameter Persistence Modules*. Thesis, Purdue University Graduate School, 2023.

[8] **Cheng Xin**, Soham Mukherjee, Shreyas N. Samaga, and Tamal K. Dey. Gril: A 2-parameter persistence based vectorization for machine learning. volume 221 of *Proceedings of Machine Learning Research*, pages 313–333. PMLR, 28 Jul 2023.


---

† Authors listed in alphabetical order.