Research Statement

Dongsheng Luo
School of Computing and Information Systems, Singapore Management University
Tel: (65) 6828-0108; Email: dsluo@smu.edu.sg
1 July 2026

Background

As artificial intelligence systems become embedded in critical decisions, from predicting urban floods to diagnosing diseases, a fundamental tension has emerged. The most powerful AI models are often the least understandable. A water management operator cannot act on a flood forecast they cannot interrogate. A clinician will not trust a diagnostic recommendation without understanding the reasoning. My research addresses this challenge by developing Trustworthy AI for Scientific Discovery, building AI systems that are accurate, robust, interpretable, and accessible to the domain experts who need them. Rather than treating explainability as an afterthought, I build trustworthiness into AI systems from the ground up, spanning three interconnected areas. The first is foundational methods for interpretable deep learning. The second is rigorous evaluation frameworks that ensure explanation quality. The third is interdisciplinary applications that translate these advances into real-world impact in environmental science, healthcare, and molecular biology.

Research Area 1: Interpretable Deep Learning for Scientific Data

Scientific discovery demands understanding, not just prediction. When an AI system identifies a toxic molecule or forecasts a flood, scientists need to know why. My research has established foundational methods that provide these mechanistic insights, transforming deep learning from opaque predictors into interpretable tools for science.

My most influential contribution, PGExplainer (NeurIPS 2020), introduced the first generative approach to graph neural network explainability. Rather than producing case-by-case rationalizations, it learns a generalizable explanation function, mirroring how scientists seek universal principles. Building on this foundation, I have systematically addressed critical gaps: resolving locality bias in GNN explanations (AAAI 2024), tackling distribution shift to ensure explanations remain valid for rare but critical events like toxic molecular configurations (SIGKDD 2023, ICML 2024), and developing confidence-aware explanation frameworks (SIGKDD 2025). For temporal data, TimeX++ (ICML 2024) applies information bottleneck principles to generate counterfactual time series explanations that preserve causal relationships, revealing not just that flooding will occur but the critical sequence of conditions that cause it.

Going forward, I plan to develop explanation-enhanced learning paradigms where explanations actively improve the training process. My preliminary work demonstrates that incorporating explanation information during training reduces sample complexity,

meaning models learn faster with less data when guided by explanations. This is particularly valuable for scientific domains where labeled data is scarce but domain knowledge is rich. I will also investigate the theoretical connections between explainability and robustness, and develop explanation-aware architectures where interpretability constraints are embedded directly into model design.

Research Area 2: Evaluation Frameworks for Trustworthy AI

A critical and often overlooked challenge is: how do we know an explanation is actually faithful to the model's reasoning? Our group discovered that existing explainability metrics suffer from a fundamental flaw where distribution shift between original data and explanation substructures leads to unreliable fidelity measurements. We introduced Robust Fidelity (ICLR 2024), information-theoretic measures that rigorously account for this shift, and subsequently developed F-Fidelity (ICLR 2025), a unified evaluation methodology that generalizes across graphs, images, text, and time series. F-Fidelity theoretically enables inference of ground-truth explanation sizes, a breakthrough for domains lacking annotated explanations.

I plan to establish comprehensive evaluation infrastructure that simultaneously assesses prediction accuracy, explanation faithfulness, robustness under distribution shift, and uncertainty calibration, reflecting that trustworthiness is not a single property but a constellation of requirements. I will also create domain-specific evaluation protocols for scientific applications, where standard benchmarks fail to capture the unique requirements of fields like environmental science and healthcare.

Research Area 3: Trustworthy AI for Environmental Science and Urban Sustainability

My applied research demonstrates that trustworthy AI achieves its full potential when it enables scientists to uncover previously hidden patterns. I lead the explainable AI component of a collaborative project on compound flood forecasting, where multiple environmental drivers (rainfall, tides, storm surge, groundwater) interact across temporal scales to produce extreme events. Through collaboration with the South Florida Water Management District, we are developing world models for compound flooding simulation: generative AI systems that learn underlying hydrological dynamics and enable "what-if" scenario planning, such as optimizing pump station operations under novel compound conditions. Our team is also building interactive systems that allow water management operators to engage with AI-driven flood forecasts without machine learning expertise.

These capabilities are directly transferable to Singapore's context. As a low-lying island nation facing compound flooding risks from monsoon rainfall, storm surge, and rising sea levels, Singapore's challenges closely parallel those in South Florida. At SMU, I plan to develop foundation models for Southeast Asian environmental science, pre-trained on regional data including tropical climate patterns and high-density urban infrastructure interactions. I will also build human-AI collaborative systems for automated environmental monitoring and create accessible AI platforms with natural

language interfaces that enable domain experts to leverage advanced AI without programming expertise.

Selected Publications and Outputs

- D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, X. Zhang. "Parameterized Explainer for Graph Neural Network." NeurIPS, 2020.
- X. Zheng, F. Shirani, Z. Chen, C. Lin, W. Cheng, W. Guo, D. Luo. "F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI." ICLR, 2025.
- X. Zheng, F. Shirani, T. Wang, W. Cheng, Z. Chen, H. Chen, H. Wei, D. Luo. "Towards Robust Fidelity for Evaluating Explainability of Graph Neural Networks." ICLR, 2024.
- X. Zheng, T. Wang, W. Cheng, A. Ma, H. Chen, M. Sha, D. Luo. "Parametric Augmentation for Time Series Contrastive Learning." ICLR, 2024.
- Z. Chen, J. Zhang, J. Ni, X. Li, Y. Bian, et al. "Generating In-Distribution Proxy Graphs for Explaining Graph Neural Networks." ICML, 2024.
- Z. Liu, T. Wang, J. Shi, X. Zheng, Z. Chen, et al. "TimeX++: Learning Time-Series Explanations with Information Bottleneck." ICML, 2024.
- J. Ding, D. Luo, A. Zilverstand, F. Liu. "NeuroTree: Hierarchical Functional Brain Pathway Decoding for Mental Health Disorders." ICML, 2025.
- J. Zhang, Z. Chen, H. Mei, D. Luo, H. Wei. "RegExplainer: Generating Explanations for Graph Neural Networks in Regression Task." NeurIPS, 2024.
- Z. Chen, J. Ni, H. Salehi, X. Zheng, E. Schafir, F. Shirani, D. Luo. "Explanation-Preserving Augmentation for Semi-Supervised Graph Representation Learning." AAAI, 2026.
- D. Luo, W. Cheng, Y. Wang, D. Xu, J. Ni, et al. "Time Series Contrastive Learning with Information-Aware Augmentations." AAAI, 2023.
- J. Zhang, D. Luo, H. Wei. "MixupExplainer: Generalizing Explanations for Graph Neural Networks with Data Augmentation." SIGKDD, 2023.
- J. Zhang, X. Liu, D. Luo, H. Wei. "Is Your Explanation Reliable: Confidence-Aware Explanation on Graph Neural Networks." SIGKDD, 2025.
- X. Liu, D. Luo, W. Gao, Y. Liu. "3DGraphX: Explaining 3D Molecular Graph Models via Incorporating Chemical Priors." SIGKDD, 2025.
- D. Luo, T. Zhao, W. Cheng, D. Xu, et al. "Towards Inductive and Efficient Explanations for Graph Neural Networks." IEEE TPAMI, 2024.
- H. Hsu et al. "MedPlan: A Two-Stage RAG-Based System for Personalized Medical Plan Generation." ACL, 2025.
- D. Luo, W. Cheng, W. Yu, B. Zong, J. Ni, H. Chen, X. Zhang. "Learning to Drop: Robust Graph Neural Network via Topological Denoising." WSDM, 2021.
- R. Huang, F. Shirani, D. Luo. "Factorized Explainer for Graph Neural Networks." AAAI, 2024.
- G. Li, J. Yang, S. Liang, D. Luo. "Elevating Spectral GNNs through Enhanced Band-pass Filter Approximation." WWW, 2025.