Research Statements

We are living in an age in which data are pervasive. Also, big and complex data of all kinds are being collected at a brisk pace to enable advances in scientific, medical, industrial, financial and many other domains. Extracting information from data and turning data into actionable knowledge are fundamental challenges that we must address. Apart from Big Data, another trend that is reshaping our life is Deep Learning, for which a model with massive parameters could mimic human intelligence and resolve many problems that was thought impossible before. However, to figure out how deep neural networks actually work and make it more reliable is another big challenge. Mathematics has much to contribute to these challenges in data science and machine learning that opens up many new research perspectives.

My research interests are in problems that involve mathematics, big data and deep learning. The tools I use including but not limited to Algebraic topology, Riemannian geometry, Optimal transport and Statistics. I have worked on problems related to:

1. low-dimensional topology
2. topological and geometric data analysis (TGDA)
3. applications of TGDA for problems in ecology and materials science;
4. computer vision/deep learning

Below, I introduce some aspects of my research which demonstrate my contributions to each of the above areas.

Low-dimensional topology

Topology is one of the fundamental branches in pure mathematics. Compact surfaces are fundamental objects in the study of low-dimensional topology. In a paper of Nielsen [12] in year 1937, he classified all orientation preserving periodic self-maps on compact surfaces. However, for the orientation reversing periodic maps, few people have been able to come up with a simple way to classify them. In [8], I extended Nilsen's work to orientation reversing case and give a complete classification of all orientation reversing periodic self-maps on compact surfaces. This work was done during my master's degree and paper [8] was published in journal ***Topology and its Applications***, which is a well-known journal in Topology area.

Topological and Geometric Data Analysis

Persistent homology is one of the main tools in topological data analysis (TDA) that allows us to summarize and analyze complex features of data through computationally accessible barcodes. Thus far, I have worked on:

1. **introducing ways of effectively using TDA to analyze functional and structural data with barcodes that are rich in information about their shapes:** We investigated ways of using persistent homology to analyze functional data on compact topological spaces and structural data represented as compact metric measure spaces. The usual way of applying persistent homology may lead to highly inaccurate barcode expressions caused by the topology of regions where the signals are weak, often unrelated to the shape of the signal. In our paper, we introduced a novel cone construction, which has a "barcode trimming" effect that enhances the "actual" shape of functional data. We defined a metric on the space of compact functional spaces and proved a stability theorem for persistent

homology, stronger than the stability theorem for standard barcodes. We also mapped metric-measure spaces to functional space via their Fréchet functions, proving consistency results and obtaining rates of convergence for barcodes derived from empirical distributions. This is joint work with Pro. Washington Mio and Pro. Facundo Mémoli, published in the ***Journal of Applied and Computational Topology***, which is a well-known top journal in TDA area.

2. **extending persistence homology to a more general setting that unifies basic structures such as persistence modules and zigzag modules:** This is the core material of my doctoral dissertation in which I introduce the notions of correspondence modules (c-modules) and persistence sheaves (p-sheaves) that not only provide a unifying framework for the study of various types of 1-dim algebraic persistence structures such as persistence modules and zigzag modules, but are also broader, including many additional structures and providing a more natural framework for several fundamental problems in TDA. For example, they formulate level-set persistent homology of a function (as opposed to using discrete zigzags), they enable us to define restrictions of a 2-dim persistence module to lines of negative slope, etc. A paper [5] has been published in ***Japan Journal of Industrial and Applied Mathematics***.
3. **introducing decorated merge trees (DMTs) as a novel invariant for TDA:** DMTs combine both 0-dimensional and high-dimensional information into a single data structure that distinguishes structures that merge trees and persistent homology cannot distinguish alone. We introduce three variants on DMTs, which emphasize category theory, representation theory and persistence barcodes, respectively, offer different advantages in terms of theory and computation. We also introduced two notions of distance—interleaving distance and bottleneck distance—for DMTs are defined and a hierarchy of stability results that both refine and generalize existing stability. We introduce computational frameworks for generating, visualizing and comparing DMTs derived from synthetic and real data. Paper has been published in the ***Journal of Applied and Computational Topology***, which is a well-known top journal in TDA area.
4. **developing new algorithms for computations of TDA invariants:** Many problems in computational TDA can be solved via linear algebra. However, matrices arise from practical applications are typically large, with rows and columns numbered in billions. Low-rank approximation of such arrays typically destroys essential information; thus, new mathematical and computational paradigms are needed for very large, sparse matrices. We present the U-match matrix factorization scheme to address this challenge. U-match has two desirable features. First, it admits a compressed storage format that reduces the number of nonzero entries held in computer memory by one or more orders of magnitude over other common factorizations. Second, it permits direct solution of diverse problems in linear and homological algebra, without decompressing matrices stored in memory. As an application, we show that individual cycle representatives in persistent homology can be retrieved at time and memory costs orders of magnitude below current state of the art. A software development project based on the corresponding algorithms can be found here https://www.nsf.gov/awardsearch/showAward?AWD_ID=1854748&HistoricalAwards=false. A paper [9] introducing the details is under review in ***Journal of Applied and Computational Topology***.

5. **Principal formulation of covariance fields on Riemannian manifold:** The mean and covariance matrix are the most basic statistics of Euclidean data and crucial in principal component analysis (PCA). Nonetheless, the standard formulation of covariance uses the linear structure of the ambient space in an essential way, making it unclear that how to extend the concept to data on non-linear spaces modeled as Riemannian manifolds in a principled manner. In this work, we proposed a formulation for general covariance tensor fields on Riemannian manifolds. For a random variable distributed according to a Borel probability measure this definition that only requires a locally linear structure coincides with the usual covariance matrix. Our paper [7] about the behavior of these covariance tensor fields have been published in ***Oberwolfach Report***.

Practical application of TGDA in Ecology and Material Science

1. **Detecting Carbon Nanotube Orientation with Topological Data Analysis of SEM Images:** High-performance carbon nanotube (CNT) materials are in high demand as a result of their extraordinary mechanical, electrical and thermal properties. CNT alignment is an important property in the fabrication of ultra-strong CNT composites. Hence, it is fundamentally important to evaluate and quantify the degree of alignment using various characterization methods. In collaboration with researchers from Florida State University High-Performance Materials Institute, we developed a novel method to detect CNT orientation combining topological data analysis with scanning electron microscopy (SEM). We use barcodes derived from persistent homology of SEM images to quantify CNT alignment. The results we have obtained are highly consistent with those from polarized Raman spectroscopy and X-ray scattering. Our approach offers a simpler and more effective way of understanding the role that alignment plays in CNT properties. Paper [10] reporting the results is published in journal ***Nanomaterials***, which is a Q1 ranking journal in material science (see https://www.scimagojr.com/journalsearch.php?q=21100253674&tip=sid for ranking details).
2. **Characterizing Phenotypic Plasticity of Ginkgo Biloba Leaves with Topological Data Analysis：** Phenotypic plasticity of living organisms can be seen in many guises and represents some of the raw material for the evolutionary process. For example, different morphologies have different functional properties and so can be favored, or not, under certain environmental conditions. In this project, joint with an ecologist from Open University, UK, we carried out a pilot study with Ginkgo biloba leaves that are known to produce leaves that are enormously diverse in their shape. We employed TGDA to develop models of morphological variation for leaves. This has allowed us to quantify morphological variation in leaves of the same tree and compare the morphology of leaves of extant plants and fossils. Such evolutionary mappings enable us to investigate evolution over geological time scales and potentially relate contrasts in phenotypes to ecological events. Paper has been published in journal ***Royal Society Open Science,*** which is a Q1 ranking journal (see https://www.scimagojr.com/journalsearch.php?tip=sid&q=21100446014 for ranking details).

Deep learning/Computer vision

1. **Geometric interpolation of Generative Adversarial Networks:** Deep generative models including Variational Autoencoder (VAE), Generative Adversarial Networks (GAN) and their variants have achieved great success in generative tasks such as image and video synthesis, super-resolution (SR), image-to-image translation, text generation, neural rendering, etc. The above approaches try to generate samples which mimic real data by minimizing various discrepancies between their corresponding statistical distributions. We propose an optimal transport based generative model called MvM — Manifold-matching via Metric-learning: Similar to GAN, the MvM contains two networks: distribution generator and metric generator. During the training process, these networks work interchangeably and obtains a win-win situation at the end. Specifically, we treat the real data set as some manifold embedded in high-dimensional Euclidean space, and generate a fake distribution measure condensed around the real data manifold by optimizing a Manifold Matching (MM) objective. The MM objective is built on shape descriptors, such as centroid and p-diameter with respect to some proper metric learnt by a metric generator using Metric Learning (ML) approaches. The paper [3] has been published in ***International Conference on Computer Vision (ICCV2021)***, which is ranked top 2 conference in computer vision/deep learning (see https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computervisionpatternrecognition for ranking details).
2. **Implicit data augmentation of GAN in low-shot setting:** Training of generative models especially Generative Adversarial Networks can easily diverge when training data is very limited. To mitigate this issue, we propose a novel implicit data augmentation approach which facilitates stable training and synthesize high-quality samples without need of label information. Specifically, we view the discriminator as a metric embedding of the real data manifold, which offers proper distances between real data points. We then utilize information in the feature space to develop a fully unsupervised and data-driven augmentation method. Experiments on low-shot generation tasks show the proposed method significantly improve results from strong baselines with hundreds of training samples. Paper [1] has appeared at ***European Conference on Computer Vision (ECCV2022)***, which is ranked top 3 conference in computer vision/deep learning (see https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computervisionpatternrecognition for ranking details).
3. **Theoretical study of GAN with high-dimensional critic output:** The study of multidimensional discriminator (critic) output for Generative Adversarial Networks has been under explored in the literature. In this work, we generalize the Wasserstein GAN framework to take the advantage of multidimensional critic output and explore its properties. We also introduce a square-root velocity transformation (SRVT) block which favors training in the multi-dimensional setting. Proofs of properties are based on our proposed maximal p-centrality discrepancy, which is bounded above by p-Wasserstein distance and fits the Wasserstein GAN framework with multi-dimensional critic output n. Especially when n = 1 and p = 1, the proposed discrepancy equals 1-Wasserstein distance. Theoretical analysis and empirical evidence show that high-dimensional critic output has its advantage on distinguishing real and fake distributions, and benefits faster convergence and diversity of results. Paper [3] has appeared at AdvML workshop in ***International Conference on Machine Learning (ICML2022)***, which is ranked top 3 conference in artificial intelligence/deep learning (see

https://scholar.google.es/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence for ranking details).

1. M. Dai, H. Hang, X. Guo, Adaptive Feature Interpolation for Low-Shot Image Generation, European Conference on Computer Vision (ECCV) (2022) [ECCV2022]
2. M. Dai, H. Hang, A. Srivastava, Rethinking Multidimensional Discriminator Output for Generative Adversarial Networks, Proceedings of the 39th International Conference on Machine Learning Workshops (AdvML) (2022) [AdvML]
3. M. Dai, H. Hang, Manifold Matching via Deep Metric Learning for Generative Modeling. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) [ICCV2021] [code]
4. J. Curry, H. Hang, W. Mio, T. Needham, O. Okutan, Decorated merge trees for persistent topology, Journal of Applied and Computational Topology (2022), 1-58 [doi] [code]
5. H. Hang, W. Mio, Correspondence modules and persistence sheaves: a unifying perspective on one-parameter persistent homology, Japan Journal of Industrial and Applied Mathematics (2022). [doi]
6. H. Hang, F. Mémoli, W. Mio, A topological study of functional data and Fréchet functions of metric measure spaces. J Appl. and Comput. Topology (2019). [doi] [arXiv]
7. H. Hang, F. Mémoli, W. Mio, Covariance tensors on Riemannian manifolds, Oberwolfach Reports (2018), 153-156. [doi] [slides]
8. H. Hang, Homology and orientation reversing periodic maps on surfaces, Topology and its Applications, 229 (2017), 1-19. [doi] [arXiv]
9. Hang, C. Giusti, L. Ziegelmeier, G. Henselman, U-match factorization: sparse homological algebra, lazy cycle representatives, and dualities in persistent(co)homology. ArXiv:2108.08831 (2021). [arXiv]
10. L. Dong, H. Hang, J. G. Park, W. Mio, R. Liang, Detecting Carbon Nanotube Orientation with Topological Analysis of Scanning Electron Micrographs, Nanomaterials (2022) [slides] [doi] [code]
11. H. Hang, M. Bauer, W. Mio, L. Mander, Geometric and topological approaches to shape variation in Ginkgo leaves, Royal Society open science (2021) [doi] [code]
12. J. Nielsen, Die Struktur periodischer Transformationen von Flächen
Math.Fys. Medd. Danske Vidensk. Selsk., 15 (1937), pp. 65-102 English transl. in: Jakob Nielsen Collected Works, Vol. 2