ICCM Satellite workshop

Applied Geometry and Topology for Data Sciences

Agenda & Abstracts

https://www.simis.cn/workshop-on-applied-geometry-and-topology-for-data-sciences/

February 24-28, 2025

Shanghai Institute for Mathematics and Interdisciplinary Sciences (SIMIS)

Auditorium (18th floor) at SIMIS, Block A, No. 657 Songhu Road, Yangpu District, Shanghai

ICCM

The International Congress of Chinese Mathematicians (ICCM) was founded in 1998 by Shing-Tung Yau, president of the ICCM and chairman of the Board of Governors of the Shanghai Institute for Mathematics and Interdisciplinary Sciences. The ICCM's objectives are to support international mathematical communities for conferences and assist in identifying annual best paper in recent years. The ICCM holds an annual general meeting and a congress every three years. It has grown into international influential mathematical event over couple of years. With its headquarters in Shanghai, ICCM is going to serve as a platform for world-level Chinese mathematicians to discuss the latest research progress.

Workshop on Applied Geometry and Topology for Data Sciences

The emergence of data-driven sciences has transformed the landscape of research and has triggered an industrial revolution. Landmark accomplishments, such as AlphaFold's groundbreaking success in predicting protein structures, highlight this shift. Similarly, the impact of models like ChatGPT and Sora is opening unprecedented avenues for AI-generated content, ushering in a transformative era. Yet, significant hurdles remain, particularly in creating efficient methods for data representation and feature extraction. In this endeavor, computational and discrete geometry/topology provide essential tools for effectively capturing, representing, and modeling data. Geometric/topological deep learning, in particular, have become instrumental in equipping models to navigate the complex geometric and topological aspects found in challenging datasets. The combination of geometric and topological insights with machine learning can redefine the future of data sciences.

This workshop seeks to merge cutting-edge geometric and topological methods with machine learning models rooted in data-driven techniques. We will delve into recent progress in discrete geometry/topology, computational approaches, geometric/topological data analysis, and deep learning that leverages geometry and topology.

Organizers

- **Fei Han (NUS, Singapore)**

- **Theodore Papamarkou (ZJNU, China)**

- **Zhi Lv (Fudan, China)**

- **Jie Wu (BIMSA, China)**

- **Kelin Xia (NTU, Singapore)**

- **Hao Xu (Zhejiang University)**

Accommodation

We are pleased to provide accommodation for conference participants at **Yaduo Hotel**, located on the **20th Floor, Tower A, Tonghe International Building, No. 133 Guotong Road, Yangpu District, Shanghai**.

For additional options, we recommend the **Crowne Plaza Fudan Hotel**, conveniently located near the conference venue. Participants choosing this hotel may arrange their stay independently and claim expenses according to their institutional policies.

Meal

Lunch will be served in a buffet style at the canteen on the 4th floor of the SIMIS Building.

Transportation Guide

📍 **Hotel Address:** Yaduo Hotel, 20th Floor, Tower A, Tonghe International Building, No. 133 Guotong Road, Yangpu District, Shanghai.

From Shanghai Pudong International Airport (PVG) to Yaduo Hotel

Option 1: Taxi

- **For passengers arriving at Terminal 1 (T1):**
  - Proceed to **P1 Parking Lot, B1** or **B2 Level** to call and wait for a taxi.
- **For passengers arriving at Terminal 2 (T2):**
  - Proceed to **P2 Parking Lot** to call and wait for a taxi.
- **Estimated Fare:** ¥100 - ¥200
- **Travel Time:** Approximately **50 minutes**, depending on traffic conditions.

Option 2: Metro

- Follow airport signs to **Pudong International Airport Station (浦东国际机场站)**, located at **Terminal 1 & Terminal 2**.
- Take **Metro Line 2 (Green Line)** from **Pudong International Airport Station** to **East Nanjing Road Station (南京东路站)**.
- Transfer to **Metro Line 10 (Purple Line)** at **East Nanjing Road Station**.
- Take Line 10 towards **Xinjiangwancheng (新江湾城)** and get off at **Sanmen Road Station (三门路站)**.
- Exit the station and walk **approximately 11 minutes** to the hotel (**use navigation for precise directions**).
- **Estimated Fare:** Around **¥8-10**.
- **Travel Time:** About **1 hour 40 minutes**.

From Shanghai Hongqiao International Airport (SHA) to Yaduo Hotel

Option 1: Taxi

- **For passengers arriving at Terminal 1 (T1):**
  - Proceed to **T1-P1 Parking Garage, B2 Level**, near the escalator to call and wait for a taxi.
- **For passengers arriving at Terminal 2 (T2):**
  - Proceed to **P6, 2F (Giraffe Level), Pickup Points 4/5 or P7, 2F (Orange Level), Pickup Points 3/4** to call and wait for a taxi.
- **Estimated Fare**: ¥100 - ¥150
- **Travel Time:** Approximately **40 minutes**, depending on traffic conditions.

Option 2: Metro

- Follow airport signs to **Hongqiao Airport Terminal 2 Station (虹桥 2 号航站楼站)**.
- Take **Metro Line 10 (Purple Line)** and get off at **Sanmen Road Station (三门路站)**.
- Exit the station and walk **approximately 11 minutes** to the hotel (**use navigation for precise directions**).
- **Estimated Travel Time:** About **1 hour**.
- **Fare:** Around **¥6-8**.

From Yaduo Hotel to the Conference Venue

 Conference Venue: Auditorium (18th Floor), SIMIS, Block A, No. 657 Songhu Rd., Yangpu District, Shanghai

🚶 **By Walking**

- The conference venue is **approximately a 5-minute walk** from Yaduo Hotel.
- Please refer to the map below for the suggested walking route.

Schedule

Day 1	24/02/2025		
Time	Host	Speaker	Title
9:00 - 9:40	Dongqin Wei	Wojtek Chacholski	Data, Geometry and Homology
9:40 - 10:20		Anqi Qiu	Spectral Laplace-Beltrami Wavelets and Geometric Convolutional Neural Network for Signal Processing and Classification
10:20 - 11:00		Tea break	
11:00 - 11:40		Chang Liu	Geometric modeling for electronic structure in molecular systems
11:40 - 14:00	Lunch		
14:00 - 14:40	Wojtek Chacholski	Dongqing Wei	AIDD and drug candidates by super-computing: Anti-Aging, Cancer and Covid-19
14:40 - 15:20		Xinqi Gong	Sable: Bridging the Gap in Protein Structure Understanding with an Empowering and Versatile Pre-training Paradigm
15:20 - 16:00		Tea break	
16:00 - 16:40		Shihua Zhang	Deep Residual Networks Learn the Geodesic Curve in the Wasserstein Space
16:40 - 17:20		Yuguang Wang	How Out-of-Distribution Detection Learning Theory Enhances Transformer: Learnability and Reliability
17:20 - 20:00	Conference Banquet		

Day 2	25/02/2025		
Time	Host	Speaker	Title
9:00 - 9:40	Xinqi Gong	Woong Kook	Topological Lens functions for Clinical Data Analysis
9:40 - 10:20		Jae-Hun Jung	Topological data analysis of time series data- Graph-based and exact persistent homology methods
10:20 - 11:00		Tea break	
11:00 - 11:40		Zhongtao Wu	Alexander polynomial of spatial graph
11:40 - 14:00	Lunch		
14:00 - 14:40	Jae-Hun Jung	Zhipan Liu	Scientific Discovery Driven by LASP Simulations
14:40 - 15:20		Yu Rong	Scientific Data Modeling: From a Geometric Perspective to a Multimodal Perspective
15:20 - 16:00		Tea break	
16:00 - 16:40		Min Zhang	ProtPainter: Draw or Drag Protein via Topology-guided Diffusion
16:40 - 17:20		Xian Wei	Relaxed Group Equivariant Networks

Day 3	26/02/2025		
Time	Host	Speaker	Title
9:00 - 9:40	Woong Kook	Zhigang Yao	Manifold Fitting Reveals Metabolomic Heterogeneity and Disease Associations in UK Biobank Populations
9:40 - 10:20		Sunhyuk Lim	Classical MDS on Metric Measure Spaces
10:20 - 11:00		Tea break	
11:00 - 11:40		Yifei Zhu	Topological time series analysis and deep learning with applications
11:40 - 14:00	Lunch		
14:00 - 14:40	Jie Wu	Niu Huang	Integrating HPC and AI: A New Paradigm for Predicting Protein-ligand Binding Interactions
14:40 - 15:20		Tong Zhu	Automated Generation of Reaction Paths
15:20 - 16:00		Tea break	
16:00 - 16:40		Yi Feng	Acupoint Topology Based on Somatic Vascularity and Autonomic Nerves
16:40 - 17:20		Ke Wei	On the convergence of policy gradient methods

Day 4	27/02/2025		
Time	Host	Speaker	Title
9:00 - 9:40	Zhi Lv	Patrizio Frosini	Explainability of neural networks through the use of GENEOs
9:40 - 10:20		Subhroshekhar Ghosh	Filtering through a topological lens: point processes and persistent homology on the time-frequency plane
10:20 - 11:00		Tea break	
11:00 - 11:40		Yipeng Zhang	Multi-Cover: A Mathematical Framework for Topological Data Analysis
11:40 - 14:00	Lunch		
14:00 - 14:40	Patrizio Frosini	Wenbing Huang	Exploring the Atomic Foundation Model from a Geometric Perspective
14:40 - 15:20		Xiao He	ChemGPT: An AI-Driven Molecular Synthesis Platform
15:20 - 16:00		Tea break	
16:00 - 16:40		Mustafa Hajij	Topological Deep Learning : Frontier and opportunities in relational systems
16:40 - 17:20		Anton Ayzenberg	Posets, sheaves, cohomology, and learning

Day 5	28/02/2025		
Time	Host	Speaker	Title
9:00 - 9:40	Theodore Papamark ou	Wilderich Tuschmann	Quantifying molecular similarity via Gromov-Hausdorff distances
9:40 - 10:20		Yiyang Jia	Category Theory-Derived Machine Learning
10:20 - 11:00		Tea break	
11:00 - 11:40		Woojin Kim	Many Facets of the Generalized Rank Invariant
11:40 - 14:00	Lunch		

Speaker: Anqi Qiu

Affiliation: Hong Kong Polytechnic University

Title: Spectral Laplace-Beltrami Wavelets and Geometric Convolutional Neural Network for Signal Processing and Classification

Abstract: The Laplace-Beltrami operator is a generalization of the Euclidean representation of the Laplace operator to an arbitrary Riemannian manifold. It is a self-adjoint operator and its eigenfunctions form a complete set of real-valued orthonormal basis functions. In this talk, I will introduce spectral Laplace-Beltrami wavelets and its computational algorithm. I will then demonstrate its use for smoothing and classification of the data defined on smooth surfaces embedded in the 3-D Euclidean space. Furthermore, I will discuss that the spectral Laplace-Beltrami Wavelets and Hodge-Laplacian to incorporate graph network topology can be used for the construction of geometric convolutional neural network (CNN). I will show the use of this method on brain morphology and functional networks for the prediction of Alzheimer's Disease and Cognition in adolescents.

Speaker: Wojtek Chacholski

Affiliation: KTH Royal Institute of Technology

Title: Data, Geometry and Homology

Abstract: The development of informative vector representations for text, languages, protein interactions, and other complex data has revolutionized how we analyze, organize, and extract insights, exemplified by advancements in large language models and generative AI. Similarly, creating effective vectorizations for geometry holds the promise of equally transformative and far-reaching impacts. In my talk I will describe a geometrical vectorization framework based on homology called stable rank. I will provide several illustrative examples of how to use stable ranks to find meaningful results in biological data.

Speaker: Anton Ayzenberg

Affiliation: Noeon Research, Japan

Title: Posets, sheaves, cohomology, and learning

Abstract: In a recent work we analyze the toolbox of sheaf theory used in various machine learning and computer science applications. The goal was to estimate the potential of this theory to enhance our theoretical and applied work. It was found that the computational approach to sheaves over arbitrary posets is poorly developed, as opposed to cellular sheaves. Most papers on using sheaf diffusion in ML are restricted to cell posets, essentially simple graphs, and there is no available algorithm to compute sheaf cohomology of posets in an optimized manner.

This work is based on joint research with T. Gebhart, G. Magai, and G. Solomadin.

Speaker: Dong-Qing Wei

Affiliation: Shanghai Jiaotong University

Title: AIDD and drug candidates by super-computing: Anti-Aging, Cancer and Covid-19

Abstract: Identifying drug-target interactions (DTIs) is an important step for drug discovery and drug repositioning(BIB 2019,2020,2021,2022). To reduce the experimental cost, many computational approaches have been developed by our team, for example, deep forest, multi-label learning and graph deep neural networks. Moreover, we updated the gold standard data set by adding 15,000 more positive DTI samples in comparison to the data set, which has widely been used by most of previously published DTI prediction methods since 2008. The proposed methods/models are applied to both data sets, demonstrating its superiority over other machine learning methods and several existing methods.

Another parallel development in designing and screening bioactive peptides against various gene targets for different diseases（Nature MI 2022）. Human leukocyte antigen (HLA) can recognize and bind foreign peptides to present them to specialized immune cells, then initiate an immune response. Computational prediction of the peptide and HLA (pHLA) binding can speed up immunogenic peptide screening and facilitate vaccine design. However, there is a lack of an automatic program to optimize mutated peptides with higher affinity to the target HLA allele. Here, to fill this gap, we develop the TransMut framework—composed of TransPHLA for pHLA binding prediction and an automatically optimized mutated peptides (AOMP) program—which can be generalized to any binding and mutation task of biomolecules. First, TransPHLA is developed by constructing a transformer-based model to predict pHLA binding, which is superior to 14 previous methods on pHLA binding prediction and neoantigen and human papilloma virus vaccine identification. For vaccine design, the AOMP program is then developed by exploiting the attention scores generated by TransPHLA to automatically optimize mutated peptides with higher affinity to the target HLA allele and with high homology to the source peptide. The proposed framework may automatically generate potential peptide vaccines for experimentalists.

In searching for drug candidates against COVID-19, we propose a new nonlinear end-to-end model called LUNAR(BIB 2022). It uses graph convolutional neural networks to automatically learn the neighborhood information of complex heterogeneous relational networks and combines the attention mechanism to reflect the importance of the sum of different types of neighbourhood information to obtain the representation characteristics of each node. These AIDD tools are used to screening drug candidates of various kinds combing structural based atomic simulations. Some promising molecules have been found. An active compound kaempferol from the Traditional Chinese Medicine (TCM) was found to be an inhibitor of the main protease. It was found that kaempferol (62.50–125.00 μg/ml, IC50 31.25 μg/ml) significantly abridged the CPE instigated by an infection in Vero E6 cells, which indicates it could be a promising SARS-CoV-2 antiviral drug.

Intestinal epithelium encounters trillions of microorganisms, including various viruses. To cope with the microbial threat, intestinal cells have evolved to produce antimicrobial peptides, including HD5(Gastroenterology 2020), which is a lectin-like peptide able to bind lipids and glycosylated proteins. In this study, we found by large scale simulations a structure-dependent interaction between HD5 and ACE2.The binding of HD5 to ACE2 cloaked several sites in the ligand-binding domain, among which Asp30and Lys31 are crucial for SARS-CoV spike to bind ACE2. Accordingly, SARS-CoV-2 S1 binding and S pseudovirions entry to enterocytes were inhibited by HD5. To our knowledge, this is the first study demonstrating the innate defense function of human intestine against SARS-CoV-2. Our finding is a reasonable explanation to the clinical phenomenon that few intestinal symptoms are observed inpatients with COVID-19. HD5 also inhibited SARS-CoV-2S pseudovirions entry to human renal proximal tubularepithelial cells, demonstrating an extensive protection of HD5. For the shortage of effective drugs to prevent and treat COVID-19, we think that it may be a useful strategy to increase the content of HD5 in vivoby oral administration. This work was recommended by Prof. Jens-Michael Schröder in Faculty Opinions: "They identified the lectin-like antimicrobial and antiviral enteric alpha-defensin HD5 as the effector molecule of the gut's innate defense system towards SARS-CoV-2 infection. These findings may suggest a new option to prevent or treat COVID-19 either by using HD5 as a therapeutic or by stimulating endogenous HD5 synthesis"(Gastroenterology 2020).

The rare event dynamics simulations can be also be used to understand, for example antivirus drug by calculating the kinetic and thermodynamic free energy profiles on the drug binding sites in the M2 proton channel. Our results give a theoretical framework to interpret and reconcile existing and often conflicting results regarding these two binding sites, thus helping to expand our understanding of M2 drug binding, and may help guide the design and screening of novel drugs to combat the virus (JACS, 2011).

Although my research is related to AI and computation, I was able to get involved in translation medicine due to our discovery of wgx-50, a new agonist of a membrane protein, α7nAChR, which is extracted from the Sichuan pepper, a result from extensive so called network pharmacology based in AIDD. Extensive experimental studies show it could combine with α7nAChR on nerve cells, induce depolymerization of Aβ, inhibit Aβ-induced neurocyte apoptosis, and suppress the release of TNF-αand IL-1β from microglia. In vivo experiments showed that it could improve the cognition ability in APP-Transgenic Mice. These results suggest that wgx50 is a promising drug candidate for AD treatment. Anti-aging potential was explored with C. elegans and mouse model, the life span are significant increased, about 30% and 100% respectively. Mouse models were also used to evaluated the potential cardio-protective effect of wgx-50. It is found that it effectively reduces iron accumulation, preserves mitochondrial function, regulates lipid metabolism, and maintains cell redox status, thereby protecting cardiomyocytes from DOX-induced ferroptosis.

Speaker: Xinqi Gong

Affiliation: Renmin University of China

Title: Sable: Bridging the Gap in Protein Structure Understanding with an Empowering and Versatile Pre-training Paradigm

Abstract: Protein pre-training has emerged as a transformative approach for solving diverse biological tasks. While many contemporary methods focus on sequence based language models, recent findings highlight that protein sequences alone are insufficient to capture the extensive information inherent in protein structures. Recognizing the crucial role of protein structure in defining function and interactions, we introduce Sable, a versatile pre-training model designed to comprehensively understand protein structures. Sable incorporates a novel structural encoding mechanism that enhances inter-atomic information exchange and spatial awareness, combined with robust pre-training strategies and lightweight decoders optimized for specific downstream tasks. This approach enables Sable to consistently outperform existing methods in tasks such as regression, classification, and generation, demonstrating its superior capability in protein structure representation. The code and models will be released in the GitHub repository.

Speaker: Shihua Zhang

Affiliation: Chinese Academy of Sciences

Title: Deep Residual Networks Learn the Geodesic Curve in the Wasserstein Space

Abstract: The fundamental principle of deep neural network (DNN) has not been fully characterized in terms of optimization and generalization. Here, we model the forward propagation process of deep residual networks (ResNet) with continuity equation, where the measure is conserved and infinite curves in the measure space connects the input distribution to the output one of ResNet. We find ResNet with L2 regularization attempts to learn the geodesic curve in the Wasserstein space, which is induced by the optimal transport map. Compared with plain network, ResNet could achieve better approximation to the geodesic curve, which explains why ResNet can be optimized and generalize better. In a word, we conclude that ResNet learns the geodesic curve in the Wasserstein space and discretely engineer the data transformation in high-dimensional spaces. Inspired by this principle, we design a novel two-step Optimal Transport induced Adversarial Defense (OTAD) model that can fit the training data accurately while preserving the local Lipschitz continuity. OTAD is extensible to diverse architectures of ResNet and Transformer, making it suitable for complex data. OTAD opens a novel avenue for developing reliable and secure deep learning systems through the regularity of optimal transport map. Inspired by this principle, we propose a conjecture, progressive feedforward collapse (PFC), claiming the degree of collapse increases during the forward propagation of ResNet and derive a transparent model for the well-trained ResNet. This extends neural collapse to PFC to model the collapse phenomenon of intermediate layers and its dependence on the input data.

Speaker: Min Zhang

Affiliation: Zhejiang University

Title: ProtPainter: Draw or Drag Protein via Topology-guided Diffusion

Abstract: Recent advances in protein backbone generation have achieved promising results under structural, functional, or physical constraints. However, existing methods lack the flexibility for precise topology control, limiting navigation of the backbone space. We present ProtPainter, a diffusion-based approach for generating protein backbones conditioned on 3D curves. ProtPainter follows a two-stage process: curve-based sketching and sketch-guided backbone generation. For the first stage, we propose CurveEncoder, which predicts secondary structure annotations from a curve to parametrize sketch generation. For the second stage, the sketch guides the generative process in Denoising Diffusion Probabilistic Modeling (DDPM) to generate backbones. During the process, we further introduce a fusion scheduling scheme, Helix-Gating, to control the scaling factors.

Speaker: Woong Kook

Affiliation: Seoul National University

Title: Topological Lens functions for Clinical Data Analysis

Abstract: Topological data analysis is applied to clinical data as an innovative methodology for visualizing patient topology and stratifying disease. In this regard, mapper has been instrumental as a medium for understanding global landscape and discovering relevant subtypes of patients. We will present previous works demonstrating how these goals could be achieved through mappers. In many mapper-based studies, however, lens functions are often selected based on data analytic preferences but not necessarily on domain insight. To address this issue, we will present recent works in close collaboration with medical experts to construct lens functions based on localized hazard ratio, and propose an application of this function in responder analysis.

Speaker: Jae-Hun Jung

Affiliation: Department of Mathematics & Director, POSTECH MINDS

Title: Topological data analysis of time series data- Graph-based and exact persistent homology methods

Abstract: Persistent homology for time series analysis has actively been studied, with several advantages over traditional Fourier analysis. In this talk, we will present our recent work on applying persistent homology to time series data. We first explain that analyzing time series, such as music data, through a graph-based persistent homology approach provides useful insights into filtration (Tran et al., 2023, 2024; Heo et al., 2025), providing a hierarchical structure in persistence measures. We will specifically explore the injection properties when distances are defined path dependent and discuss their implications in music analysis. We also introduce a novel method connected to dynamical systems (Kim & Jung, 2024) that differs slightly from traditional embedding theory. This approach enables efficient filtration learning within a machine learning framework, minimizing computational costs while identifying the optimal filtration space. This work is a collaboration with Byeongchan Choi, Eunwoo Heo, Keunsu Kim, Changbom Park, and Mai Lan Tran.

Speaker: Zhongtao Wu

Affiliation: The Chinese University of Hong Kong

Title: Alexander polynomial of spatial graph

Abstract: Alexander polynomial has been one of the most important tools in the development of knot theory since its discovery 100 years ago. For spatial graphs, Bao and the speaker defined an analogous invariant. In many aspects, the Alexander polynomial of spatial graphs shares similar topological properties with the classical one for knots; but it also contains certain unique graph theoretical information, such as, its evaluation at t=1 gives the number of spanning trees of the graph. This talk aims to give a general introduction to this invariant and explore its potential applications in topological data analysis.

Speaker: Zhipan Liu

Affiliation: Fudan University

Title: Scientific Discovery Driven by LASP Simulations

Abstract: Atomic simulation aims to understand and predict complex physical phenomena. Its success largely depends on the accuracy of the potential energy surface (PES) description and the efficiency of capturing important rare events. In 2018, the research group led by Liu Zhipan released the LASP software (Large scale atomic simulation with a Neural Network Potential). By combining the advanced neural network potential function G-MBNN with the efficient SSW global optimization method, the software encompasses two key elements for achieving the ultimate goal of atomic simulation, namely automation and intelligence. This lecture mainly introduces the latest methods and software developments of LASP in the direction of high-level intelligence and automation for solving complex problems in materials and reactions. Thanks to LASP's powerful PES evaluation and exploration capabilities, it has now become a powerful intelligent simulation tool for scientific discovery. We present LASP's recent developments in combination with the diffusion generative models and their applications in materials and catalysis [1 - 6].

Speaker: Yu Rong

Affiliation: Alibaba DAMO Academy

Title: Scientific Data Modeling: From a Geometric Perspective to a Multimodal Perspective

Abstract: In current fundamental scientific research, the knowledge of systems is often not confined to a single modality. For example, modeling proteins simultaneously involves sequences, structures, and textual information. At the same time, a vast amount of fundamental scientific knowledge is stored and recorded in text. Therefore, how to integrate information from multiple modalities for modeling and understanding is crucial for leveraging AI to solve scientific problems. In this presentation, I will introduce the latest advancements and related applications in scientific data modeling from both geometric and multimodal perspectives. Additionally, I will provide an outlook on the role of multimodal modeling in the field of AI for Science.

Speaker: Yuguang Wang

Affiliation: Shanghai Jiaotong University

Title: How Out-of-Distribution Detection Learning Theory Enhances Transformer: Learnability and Reliability

Abstract: Transformer networks excel in natural language processing and computer vision tasks. However, they still face challenges in generalizing to Out-of-Distribution (OOD) datasets, i.e. data whose distribution differs from that seen during training. The OOD detection aims to distinguish outliers while preserving performance on in-distribution (ID) data. This paper introduces the OOD detection Probably Approximately Correct (PAC) Theory for transformers, which establishes the conditions for data distribution and model configurations for the learnability of transformers in terms of OOD detection. The theory demonstrates that outliers can be accurately represented and distinguished with sufficient data. The theoretical implications highlight the trade-off between theoretical principles and practical training paradigms. By examining this trade-off, we naturally derived the rationale for leveraging auxiliary outliers to enhance OOD detection. Our theory suggests that by penalizing the misclassification of outliers within the loss function and strategically generating soft synthetic outliers, one can robustly bolster the reliability of transformer networks. This approach yields a novel algorithm that ensures learnability and refines the decision boundaries between inliers and outliers. In practice, the algorithm consistently achieves state-of-the-art performance across various data formats.

Speaker: Xian Wei

Affiliation: East China Normal University

Title: Relaxed Group Equivariant Networks

Abstract: Group Equivariant Networks empower models to explore symmetries hidden in visual data and assume uniform and strict symmetry across all features as the transformations under the specific group. However, in real-world scenarios, objects or scenes often exhibit perturbations of a symmetric system, specifically a deviation from a symmetric structure, which can be characterized by a non-trivial action of a symmetry group, known as Symmetry-Breaking. Many Group Equivariant networks like Equivariant CNN are limited by the strict operation rules in the group space, only ensuring features remain strictly equivariant under limited group transformations, making it difficult to adapt to Symmetry-Breaking or non-rigid transformations. Motivated by this, we propose simple but highly effective methods, namely SBDet and RREConv, to address this problem.

Speaker: Zhigang Yao

Affiliation: NUS

Title: Manifold Fitting Reveals Metabolomic Heterogeneity and Disease Associations in UK Biobank Populations

Abstract: This study is the first to utilize a manifold-fitting framework within NMR-based metabolomics to explore metabolic heterogeneity in the UK Biobank population. Our method clusters 251 metabolic biomarkers into seven distinct categories that reflect the modular organization of human metabolism. Applying manifold fitting reveals low-dimensional structures in each category, capturing crucial metabolic variations associated with diverse disease risks. Notably, fitted manifolds in three categories distinctly stratify the population, each identifying two subgroups with unique metabolic profiles linked to a broad spectrum of diseases, from metabolic complications to cardiovascular and autoimmune disorders. This nuanced stratification enhances our understanding of the interactions between metabolism and disease, potentially guiding personalized health interventions and advancing preventive medicine strategies.

上海数学与交叉学科研究院

SHANGHAI INSTITUTE FOR MATHEMATICS AND INTERDISCIPLINARY SCIENCES

Speaker: Subhroshekhar Ghosh

Affiliation: NUS

Title: Filtering through a topological lens: point processes and persistent homology on the time-frequency plane

Abstract: We introduce a very general approach to the analysis of signals from their noisy measurements from the perspective of Topological Data Analysis (TDA). While TDA has emerged as a powerful analytical tool for data with pronounced topological structures, here we demonstrate its applicability for general problems of signal processing, without any a-priori geometric feature. Our methods are well- suited to a wide array of time-dependent signals in different scientific domains, with acoustic signals being a particularly important application. We invoke time-frequency representations of such signals, focusing on their zeros that are gaining salience as a signal processing tool in view of their stability properties. Leveraging state-of-the- art topological concepts, such as stable and minimal volumes, we develop a complete suite of TDA-based methods to explore the delicate stochastic geometry of these zeros, capturing signals based on the disruption they cause to this rigid, hyperuniform spatial structure. Unlike classical spatial data tools, TDA is able to capture the full spectrum of the stochastic geometry of the zeros, thereby leading to powerful inferential outcomes that are underpinned by a principled statistical foundation. This is reflected in the power and versatility of our applications, which include competitive performance in processing a wide variety of audio signals, esp. in low SNR regimes. Futuristic applications include effective detection and reconstruction of gravitational wave signals (a reputed signal processing challenge with non-Gaussian noise), and medical time series data from EEGs, indicating a wide horizon for the approach and methods introduced in this inaugural work. Based on joint work with J. Miramont, K.A. Tan, S. Mukherjee, R. Bardenet.

Speaker: Yifei Zhu

Affiliation: Southern University of Science and Technology

Title: Topological time series analysis and deep learning with applications

Abstract: We give an overview of topological approaches to analyzing time-dependent data, with an emphasis on detection of periodic phenomena. This methodology enjoys robustness afforded by continuous deformation and change of measures, captures interesting geometric features underlying the data, and requires a reasonable computational cost. We illustrate these by reporting progress on two specific applications: (i) to biomedical engineering, automated and real-time detection of mouse scratching behavior, joint with Fangyi Chen and Zhen Zhang, and (ii) to speech signal processing, classification of voiced and voiceless phonetic data, joint with Meng Yu. We will also indicate recent applications to design of convolutional layers for deep learning, as well as design of 3D human–computer interaction devices.

![上海数学与交叉学科研究院 SHANGHAI INSTITUTE FOR MATHEMATICS AND INTERDISCIPLINARY SCIENCES]

Speaker: Niu Huang

Affiliation: National Institute of Biological Sciences (NIBS), Beijing

Title: Integrating HPC and AI: A New Paradigm for Predicting Protein-ligand Binding Interactions

Abstract: In the process of small molecule drug discovery, the prediction of protein-ligand interactions urgently demands enhancements in computational accuracy and efficiency, given its crucial role in identifying novel lead compounds for new targets. However, current artificial intelligence (AI) models are constrained by the scarcity of large, high-quality protein-ligand complex structures and binding data, which consequently impairs their generalization ability, limiting their effectiveness in real-world applications. We have been actively exploring the potential of physics-based high performance computing (HPC). The remarkable computational power of HPC allows us to generate vast, top-tier datasets that are invaluable for both training and testing AI modes. When integrated with AI's proficiency in pattern recognition and predictive modeling, this combination allows for the rapid and in-depth analysis of molecular structures, more accurate prediction of drug-target interactions. Our ongoing research and practice will highlight the profound synergy between HPC and AI in facilitating more accurate and efficient calculations of molecular interactions, illuminating viable strategies to surmount existing data limitations and improve the predictive capabilities of AI models.

Speaker: Chang Liu

Affiliation: Microsoft Research

Title: Geometric modeling for electronic structure in molecular systems

Abstract: Accurately calculating properties of molecular systems is the foundation for many industrial innovations, including drug design and material discovery. Properties of a molecule are essentially determined by the electronic structure in it, which is governed by the Schrödinger equation. Density functional theory (DFT) solves it in a reduced form and achieves the current sweet point between accuracy and efficiency, but it is still costly to scale to large molecules. The talk covers two approaches to address it. 1) We used a higher-order-tensor equivariant model to predict the electron mean-field Hamiltonian, the fundamental solution from a DFT calculation. We introduce a self-consistency training method that allows continual improvement without data but using the basic DFT equation. 2) We used a geometric deep learning model to revive orbital-free DFT, a further reduced formulation with lower complexity. The model represents electron density as geometric quantities on a molecule, and accurately approximates the kinetic energy density functional, solving the longstanding bottleneck problem of the formulation, while leveraging the lower complexity to scale to larger systems.

Speaker: Yi Feng

Affiliation: Fudan University

Title: Acupoint Topology Based on Somatic Vascularity and Autonomic Nerves

Abstract: Acupoints have been recognized for over 2000 years and serve as the foundation for clinical treatments such as acupuncture, moxibustion, and massage in traditional Chinese medicine. Numerous studies have demonstrated that most acupoints are densely enriched with nerve endings. However, exploring the deeper significance of acupoints from a three-dimensional (3D) perspective has long been hindered by a lack of effective technical tools and analytical approaches. In this study, we employed topological data analysis to quantitatively map the 3D tubular networks of the autonomic nervous system and vasculature, complemented by in vivo near-infrared II (NIR-II) imaging, to explore the biological mechanisms underpinning these structural characteristics. By applying four novel topology-derived indices—bifurcation loops, angles, curvature, and entanglement—we accurately differentiated acupoints from non-acupoints. Strikingly, the entanglement of parasympathetic nerves with vessels was more pronounced at acupoints than that of sympathetic nerves, suggesting that the capacity to modulate blood vessels is a precursor to sympathetic regulation. These pioneering topology-based indices thus provide a more profound understanding of the neurovascular architecture associated with acupoints and its relevance to the efficacy of acupuncture.

Speaker: Ke Wei

Affiliation: School of Data Science, Fudan University

Title: On the convergence of policy gradient methods

Abstract: Reinforcement learning (RL) is a type of machine learning technique for solving sequential decision problems which has achieved great success in many applications. In this talk, we will report recent progress on the analysis of different policy gradient methods.

上海数学与交叉学科研究院
SHANGHAI INSTITUTE FOR MATHEMATICS AND INTERDISCIPLINARY SCIENCES

Speaker: Patrizio Frosini

Affiliation: University of Pisa

Title: Explainability of neural networks through the use of GENEOs

Abstract: Group equivariant non-expansive operators (GENEOs) are mathematical tools used to approximate data observers, especially when data are represented as real-valued or vector-valued functions (https://rdcu.be/bP6HV). These operators are founded on the idea that the geometric characteristics of observers significantly influence how data is interpreted. In this talk, we will explore the core properties of GENEOs, examine their role in Machine Learning, and discuss their promising applications in Explainable Artificial Intelligence.

Speaker: Sunhyuk Lim

Affiliation: Department of Mathematcis, Sungkyunkwan University (SKKU)

Title: Classical MDS on Metric Measure Spaces

Abstract: We study a generalization of the classical multidimensional scaling procedure (cMDS) which is applicable in the setting of metric measure spaces. Metric measure spaces can be seen as natural 'continuous limits' of finite data sets. Given a metric measure space X = (X, d_X, μ_X), the generalized cMDS procedure involves studying an operator which may have infinite rank, a possibility which leads to studying its traceability.

We establish that several continuous exemplar metric measure spaces such as spheres and tori (both with their respective geodesic metrics) induce traceable cMDS operators, a fact which allows us to obtain the complete characterization of the metrics induced by their resulting cMDS embeddings. To complement this, we also exhibit a metric measure space whose associated cMDS operator is not traceable. Finally, we establish the stability of the generalized cMDS method with respect to the Gromov–Wasserstein distance.

Speaker: Yipeng Zhang

Affiliation: NTU

Title: Multi-Cover: A Mathematical Framework for Topological Data Analysis

Abstract: Topological Data Analysis (TDA) has made significant contributions to molecular and materials science. Multi-Cover Persistence (MCP) and its associated Rhomboid Tiling (RT) structure, as a generalization of the alpha shape in TDA, provide a powerful framework for capturing the shape and higher-order geometric features of objects. Leveraging these concepts, we developed two innovative models: a featurization-based machine learning model and an end-to-end deep learning model. Specifically, the MCP framework is utilized to extract topological features from the persistent homology of molecular multi-covers, enabling a machine learning model to predict polymer properties with high accuracy. Meanwhile, the RT structure forms the basis for a hierarchical graph pooling model, designed for molecular graph classification tasks. Both models demonstrate excellent performance, highlighting the potential of MCP and RT in advancing polymer informatics and geometric graph learning.

Speaker: Wenbing Huang

Affiliation: Renmin University of China

Title: Exploring the Atomic Foundation Model from a Geometric Perspective

Abstract: Atomic systems, such as molecules, proteins, and crystals, represent fundamental material forms in the microscopic physical world. Accurately representing and efficiently generating atomic systems is crucial for supporting a wide range of scientific tasks. Given the unique geometric configurations and dynamical properties of atomic systems, this talk will explore and design atomic basis models, generative algorithms, and efficient dynamical prediction methods from geometric perspectives, including combinatorial structures, hierarchical organization, and symmetry. To achieve this, we will leverage recent advances in artificial intelligence, such as Transformer models, diffusion models, and DPO alignment algorithms, integrating these techniques to enhance the representation, generation, and prediction of atomic systems.

Speaker: Mustafa Hajij

Affiliation: University of San Francisco

Title: Topological Deep Learning : Frontier and opportunities in relational systems

Abstract: Topological Deep Learning (TDL) is an emerging frontier that generalizes graph-based models to higher-order structures such as simplicial complexes, cell complexes, and hypergraphs, enabling the extraction of global information from complex data domains. By building on topological data analysis (TDA), topological signal processing, network science, and geometric deep learning, TDL allows for the analysis of higher-order network data, moving beyond binary relationships to represent multi-entity interactions. This paradigm shift facilitates novel applications in diverse fields, including social science, transportation, physics, molecular design, epidemiology, art, and scientific discovery. Despite its potential, TDL faces significant challenges, including the computational complexity of higher-order structures, limited availability of large-scale labeled datasets, and the need for scalable architectures that effectively generalize across domains. Addressing these challenges presents opportunities to advance TDL by developing new algorithms, integrating domain knowledge, and exploring cross-disciplinary applications. As TDL evolves, it promises to expand the scope of deep learning, pushing the boundaries of how we model and understand complex systems.

Speaker: Tong Zhu

Affiliation: East China Normal University

Title: Automated Generation of Reaction Paths

Abstract: Chemical reactions, especially those involving complex mixtures, present a significant challenge for systematic understanding due to their intricate reaction pathways and mechanisms. Relying solely on current experimental methods and computational techniques to unravel the complexity of such reactions is often inadequate. Manual identification and analysis of all possible reaction pathways is a formidable task, while employing quantum chemistry methods to calculate each pathway is impractical due to computational demands. In light of the increasing demand for sustainable and environmentally friendly chemical processes, there is an urgent need for efficient and high-precision mechanisms for chemical reactions. In this context, we highlight some recent advancements made in the field. By leveraging the synergy between machine learning and physical models, we have managed to accelerate the search for chemical reaction pathways and quickly predict reaction rate constants. The incorporation of machine learning enhances the method's efficiency, whereas the physical model ensures its accuracy and ability to extrapolate. The development of this integrated approach is expected to offer more reliable and efficient tools for the rapid construction of chemical reaction mechanisms, paving the way for innovations in sustainable chemistry and material science.
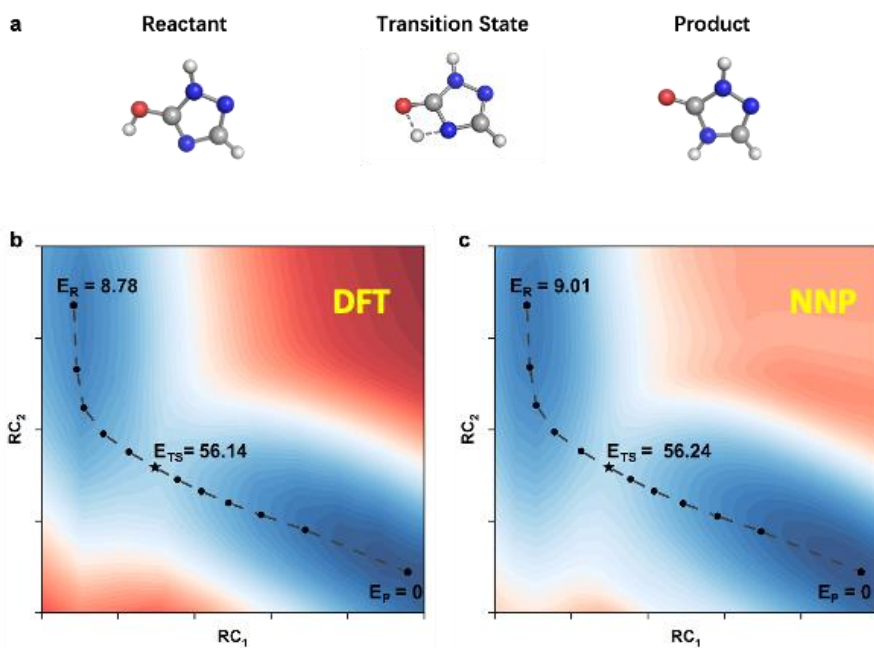
Speaker: Wilderich Tuschmann

Affiliation: KIT, Germany

Title: Quantifying molecular similarity via Gromov-Hausdorff distances

Abstract: I will discuss several metric geometry and, in particular, a cohomology-based Gromov-Hausdorff ultrametric approach based on the use of combinatorial Laplace operators to analyze and address typical clustering questions arising in molecular data analysis.

This is joint work with X. Gong, J. Wee and K. Xia.

Speaker: Yiyang Jia

Affiliation: Department of Mathematical, Physical, and Information Science Japan Women's University

Title: Category Theory-Derived Machine Learning

Abstract: In this presentation, we provide an overview of category theory-derived machine learning from four mainstream perspectives: gradient-based learning, probability-based learning, invariance and equivalence-based learning, and topos-based learning. In certain machine learning methods, the compositionality of functors plays a crucial role, inspiring the development of specific categorical frameworks. However, when examining how global network properties manifest in local structures, and how geometric properties and semantics are articulated through logic, the topos structure emerges as particularly significant and profound

Speaker: Woojin Kim

Affiliation: KAIST

Title: Many Facets of the Generalized Rank Invariant

Abstract: The Generalized Rank Invariant (GRI) is a natural invariant for multi-parameter persistence modules, extending the concept of the persistence diagram from one-parameter to the multi-parameter setting. In this talk, we will discuss various aspects of the GRI, including its Möbius invertibility, discriminating power, connections to other invariants, computational aspects, and estimation methods, as time permits.

Speaker: Rongling Wu

Affiliation: BIMSA

Title: A statistical framework to unify Yau-Yau filter and GLMY homology

Abstract: The Yau-Yau filter is a theory to find true states of internal workings within a complex stochastic system from noisy data, whereas GLMY homology aims to find topological features of digraphs. These two theories have increasingly emerged as powerful tools to find fundamental principles behind observed data, but they have been applied independently in their own fields. In this talk, I will present a unified theory to contextualize these two theories into a cohesive whole, demonstrating how this new theory can disentangle stochastic phenomena and find natural laws from them.

Speaker: Xiao He

Affiliation: School of Chemistry and Molecular Engineering, East China Normal University

Title: ChemGPT: An AI-Driven Molecular Synthesis Platform

Abstract: In this talk, I will introduce the latest development from East China Normal University—ChemGPT 2.0. This includes the construction of a high-quality chemical dialogue dataset, where ChemGPT 2.0 integrates over one million high-quality dialogue entries. Based on extensive collection and deep understanding of specialized knowledge in the field of chemistry, the dataset provides robust support for comprehensive and accurate chemical knowledge Q&A.

Next is the creation of the compound retrosynthesis database. To address the challenge of compound retrosynthesis, we employed techniques such as data splicing, overlaying, weighting, and synthesis to build a new retrosynthesis database. This large-scale database enhances the model's robustness and responsiveness, while its high-quality annotated data improves the model's accuracy and reliability.

Finally, we made innovative improvements to both the dialogue model and the retrosynthesis model. Through the implementation of multi-model and multi-module integration technology, ChemGPT 2.0 can support knowledge Q&A in the areas of professional chemistry, chemical retrosynthesis, biopharmaceuticals, and general knowledge. Based on this, we have completed the overall construction and framework design of an automated chemical synthesis reaction technology system. In conjunction with advanced "Beyond Limits Manufacturing" technology, microfluidic chip-based chemical synthesis reduced experimental time by 80%. The automated synthesis factory, driven by AI chemist "Xiaohua," has enabled automated compound synthesis, showcasing the vast potential of artificial intelligence in the biopharmaceutical field.