Titles and Abstracts

December 16, 2024 Monday

Chair: Stephen S.-T. Yau(丘成栋), Beijing Institute of Mathematical Sciences and Applications (BIMSA)

Modeling of bio-related molecules with the help of AI

Yiqin Gao(高毅勤)
Peking University

In this talk, we will introduce some recent progress in protein and in particular protein complex structure prediction models. Protein complex structures are essential for understanding of biological activities and drug development. Improving complex structure prediction accuracy of AI models for cases such as antigen-antibody complexes is expected to further enhance their applicability. Meanwhile, a large variety of experimental methods are used to provide structural insights for protein complexes, but with only sparse or approximate knowledge obtained. To efficiently and flexibly incorporate the different forms of experimental information, we introduce here GRASP, which can integratively and effectively incorporates experimental information including those obtained from XL, CL, CSP, and DMS. Besides structure prediction, designing protein structures towards specific functions is of great values for science, industry and therapeutics. Aiming to sketch a compact space for designable protein structures, we develop probabilistic tokenization theory for metastable protein structures. We present an unsupervised learning strategy, which conjugates inverse folding with structure prediction, to encode protein structures into amino-acid-like tokens and decode them back to atom coordinates. We show that tokenizing protein structures variationally can lead to compact and informative representations (ProTokens). By unifying 1-dimensional and 3-dimensional representations of protein structures, ProTokens also enable all-atom protein structure design via various generative models without the concern of symmetry or modality mismatch. We demonstrate that generative pretraining over ProToken vocabulary allows scalable foundation models to perceive, process and explore the microscopic structures of biomolecules effectively. At the end of talk, we will also discuss on recent progress made on combining AI methods and fast docking methods to accelerate the prediction on protein-small molecule binding.

Protein-ligand binding energy and molecular solvation

John Z.H. Zhang(张增辉)
Shenzhen University of Advanced Technology

Accurate computation of protein-ligand binding free energy remains an elusive goal due to inherent difficulties involved in the calculation of gas-phase protein-ligand interaction energy, the entropy, and the solvation energy. In this study, we explore the use of fragment quantum chemical calculations for improved accuracy in protein-ligand binding free energy calculations. The present work demonstrated that the gas-phase protein-ligand interaction energies are accurately calculated by the MFCC (Molecular Fractionation with Conjugate Caps) method as verified by comparison with the full quantum calculations for several protein-ligand systems. The m06-2x/6-31+G* level of DFT calculation with basis set superposition error (BSSE) correction is found to give excellent protein-ligand interaction energies. The quantum calculated protein-ligand interaction energies are then combined with implicit solvation methods to obtain absolute binding free energies and the results are shown to be sensitive to the specific solvation models used. In particular, the accuracy of the quantum calculated binding free energies is significantly improved over that of the force field calculations using the same solvation models in terms of mean absolute errors (MAE). However, the correlation coefficients of the binding free energies computed with the quantum gas phase interaction energy with respect to the experimental data do not show any improvement over the corresponding result computed from the force field. Such result and the related analysis underscore the critical importance of solvation energies to the binding free energies and the need for developing more accurate methods to calculate solvation energies. Some efforts in .explicit solvation calculation of molecular solvation energy are presented.

Single-cell analysis via manifold fitting: A framework for RNA clustering and beyond

Zhigang Yao(姚志刚)
National University of Singapore

Single-cell RNA sequencing (scRNA-seq) data, susceptible to noise arising from biological variability and technical errors, can distort gene expression analysis and impact cell similarity assessments, particularly in heterogeneous populations. Current methods, including deep learning approaches, often struggle to accurately characterize cell relationships due to this inherent noise. To address these challenges, we introduce scAMF (Single-cell Analysis via Manifold Fitting), a framework designed to enhance clustering accuracy and data visualization in scRNA-seq studies. At the heart of scAMF lies the manifold fitting module, which effectively denoises scRNA-seq data by unfolding their distribution in the ambient space. This unfolding aligns the gene expression vector of each cell more closely with its underlying structure, bringing it

spatially closer to other cells of the same cell type. To comprehensively assess the impact of scAMF, we compile a collection of 25 publicly available scRNA-seq datasets spanning various sequencing platforms, species, and organ types, forming an extensive RNA data bank. In our comparative studies, benchmarking scAMF against existing scRNA-seq analysis algorithms in this data bank, we consistently observe that scAMF outperforms in terms of clustering efficiency and data visualization clarity. Further experimental analysis reveals that this enhanced performance stems from scAMF's ability to improve the spatial distribution of the data and capture class-consistent neighborhoods. These findings underscore the promising application potential of manifold fitting as a tool in scRNA-seq analysis, signaling a significant enhancement in the precision and reliability of data interpretation in this critical field of study.

Chair: Qi Wang(王奇), University of South Carolina

Finite Difference Approximation with ADI Scheme for Two-dimensional Keller-Segel Equations

Xiaofan Li(李晓璠)
Illinois Institute of Technology

Keller-Segel systems are a set of nonlinear partial differential equations used to model chemotaxis in biology. In this talk, we present two alternating direction implicit (ADI) schemes to solve the 2D Keller-Segel systems directly with minimal computational cost, while preserving positivity, energy dissipation law and mass conservation. One scheme unconditionally preserves positivity, while the other does so conditionally. Both schemes achieve second-order accuracy in space, with the former being first-order accuracy in time and the latter second-order accuracy in time. Besides, the former scheme preserves the energy dissipation law asymptotically. We validate these results through numerical experiments, and also compare the efficiency of our schemes with the standard five-point scheme, demonstrating that our approaches effectively reduce computational costs.

An Iteratively Derived Knowledge-based Scoring Function at Atomic Level for Protein-DNA Complexes Evaluations

Xiaoqin Zou(邹晓勤)
University of Missouri

Protein-DNA interactions play a significant role in biological processes and drug design owing to their prevalence. Computational methods for predicting protein-DNA complex structures serve as a valuable alternative to experimental methods, which, although more accurate, are also time-consuming and resource-intensive. The established framework for predicting protein-protein complex structures can be adapted for protein-DNA complexes, typically involving a Fast Fourier Transform (FFT)-based rigid docking, followed by a scoring function to re-rank the modeled structures. Despite the efficiency and success of this framework, its success rate is influenced by conformational changes induced during the binding process—a common phenomenon in protein-DNA interactions. To address this challenge, we have developed an iterative method, ITScorePD, for training a knowledge-based scoring function on an augmented set. This set includes experimentally resolved crystal structures, and reasonable decoy structures and enriched near-native structures generated through the rotation-translation blocked (RTB) method to account for conformational changes in both proteins and DNAs. Our results indicate that including near-native structures in the training set significantly improves the performance of ITScorePD compared to the scoring function derived from a training set without near-native structures. The detailed test results will be presented.

Chair: John Z.H. Zhang(张增辉), Shenzhen University of Advanced Technology

Kinetic pathway of HIV-1 TAR cotranscriptional folding

Shi-Jie Chen(陈世杰)
University of Missouri

The Trans-Activator Receptor (TAR) RNA, located at the 5'-end untranslated region (5' UTR) of the human immunodeficiency virus type 1 (HIV-1), is pivotal in the virus's life cycle. As the initial functional domain, it folds during the transcription of viral mRNA. Although TAR's role in recruiting the Tat protein for trans-activation is established, the detailed kinetic mechanisms at play during early transcription, especially at points of temporary transcriptional pausing, remain elusive. Moreover, the precise physical processes of transcriptional pause and subsequent escape are not fully elucidated.
This study focuses on the folding kinetics of TAR and the biological implications by integrating computer simulations of RNA folding during transcription with Nuclear Magnetic Resonance (NMR) spectroscopy data. The findings reveal insights into the folding mechanism of a non-native intermediate that triggers transcriptional pause, along with different folding pathways leading to transcriptional pause and readthrough. The profiling of the cotranscriptional folding pathway and identification of kinetic structural intermediates reveal a novel mechanism for viral transcriptional regulation, which could pave the way for new antiviral drug designs targeting kinetic cotranscriptional folding pathways in viral RNAs.

A Novel Representation for Proteins Related to Phosphorylation Process

Mengcen Guan(关梦岑)
Tsinghua University

Phosphorylation is one of the most important post-translational modifications in cells, and proteins like protein kinases, protein phosphatases, and phosphoprotein-binding domains play significant roles in this process. Here a new vector representation method is developed for these phosphorylation-related proteins, consisting of 648 elements that encompass information of the distributions of 20 amino acids and their physicochemical properties. With finite dimensions, this novel vector outperforms the former alignment free methods in the classification tasks of three protein datasets. Besides, on specific human kinase dataset and bacterial tyrosine kinase dataset, the novel vector can give precise phylogenetic analysis results which could give insights to the inner relationships of human kinases related to diseases and bacterial tyrosine kinases

Efficient clustering on a super large collection of molecular structures on a multiple GPU platform

Changjun Chen(陈长军)
Huazhong University of Science and Technology

Structure clustering is very time consuming for a large dataset in molecular dynamics simulation. In this work, we specially develop a powerful tool to do the work on GPU. To show the performance, we apply the tool to a 33-residue fragment in protein Pin1 WW domain mutant. The dataset contains 1,400,000 snapshots, which are extracted from an enhanced sampling simulation and distribute widely in the conformational space. Various testing results present that our program is quite efficient. Particularly, with two NVIDIA RTX4090 GPUs and single precision data type, the clustering calculation on one million snapshots is completed in a few seconds (including the uploading time of data from memory to GPU and neglecting the reading time from hard disk). This is hundreds of times faster than CPU. Our program could be a useful tool for fast extraction of representative states of a molecule from large mount of trajectories.

Chair: Shi-Jie Chen(陈世杰), University of Missouri

TBA

Dmytro Kozakov
Stony Brook

TBA

December 17, 2024 Tuesday

Chair: Yiqin Gao(高毅勤), Peking University

Motion of active elastic particles driven by self-generatedforce couples

Qi Wang(王奇)
University of South Carolina

In this talk, we introduce a novel hydrodynamically coupled volume-conserving Allen-Cahn phase-field model to investigate the dynamics of active particles propelled by force couples. Our model integrates variable-viscosity Navier-Stokes equations, rigid or elastic properties of the particle, active behavior, and nonlinear Allen-Cahn equations to capture multiphase interactions comprehensively. Central to our approach is the incorporation of a nonlocal Lagrange multiplier for phase volume conservation, coupled with force couples to drive particle self-propulsion. The model demonstrates remarkable scalability, accommodating particles with diverse elasticity, rigidity, and activity levels, thereby offering a versatile framework for analyzing active particle dynamics. Derived from the Onsager principle, our model ensures energy dissipation in the absence of active force couples. To efficiently solve this complex system, we employ a linearly decoupled, first-order time-marching numerical scheme that combines a projection approach with a stabilization technique. The rigorously validated decoupled scheme ensures asymptotic preservation and thermodynamic consistency. Numerical simulations underscore the efficacy of both the developed model and the numerical scheme, showcasing their effectiveness in capturing and elucidating the dynamics of active particles propelled by force couples.

RABC: Rheumatoid Arthritis Bioinformatics Center

Yongshuai Jiang(姜永帅)
Harbin Medical University

Advances in sequencing technologies have led to the rapid growth of multi-omics data on rheumatoid arthritis (RA). However, a comprehensive database that systematically collects and classifies the scattered data is still lacking. Here, we developed the Rheumatoid Arthritis Bioinformatics Center (RABC, http://www.onethird-lab.com/RABC/), the first multi-omics data resource platform (data hub) for RA. There are four categories of data in RABC: (i) 175 multi-omics sample sets covering transcriptome, epigenome, genome, and proteome; (ii) 175 209 differentially expressed genes (DEGs), 105 differentially expressed microRNAs (DEMs), 18 464 differentially DNA methylated (DNAm) genes, 1 764 KEGG pathways, 30 488 GO terms, 74 334 SNPs, 242 779 eQTLs, 105 m6A-SNPs and 18 491 669 meta-mQTLs; (iii) prior knowledge on seven types of RA molecular markers from nine public and credible databases; (iv) 127 073 literature information from PubMed (from 1972 to March 2022). RABC provides a user-friendly interface for browsing, searching and downloading these data. In addition, a visualization module also supports users to generate graphs of analysis results by inputting personalized parameters. We believe that RABC will become a valuable resource and make a significant contribution to the study of RA.

Chromosomal fusion recognition based on the alignment-free natural vector method

Hongyu Yu(余泓谕)
Tsinghua University

Chromosomal fusion is a significant type of structural variation, yet existing algorithms for its identification often depend on manual annotations and inefficient sequence alignments through synteny analysis. This report presents a novel alignment-free algorithm for recognizing chromosomal fusions by transforming the identification problem into a series of assignment tasks, which are solved efficiently using the Kuhn-Munkres algorithm. The proposed method significantly enhances processing speeds by eliminating time-consuming alignments and the need for manual annotations. By considering entire chromosomes instead of fragments, this approach offers substantial potential for advancing research in the field of chromosomal structural variations.

Chair: Shi Huang(黄石), Xiangya Medical School, Central South University

A Novel Framework for Predicting Genome Sequences and Gene Mutations Based on Natural Vector and Convex Hull Theory

Yishuai Niu(牛一帅)
Beijing Institute of Mathematical Sciences and Applications (BIMSA)

This report presents a novel computational framework grounded in natural vector and convex hull theory for the prediction of genome sequences and gene mutations. This methodology offers a sophisticated and efficient algorithmic structure that provides critical insights and capabilities for genome sequence and mutation prediction.
The framework is composed of four primary modules:
1. An enhanced natural vector method for the analysis of genomic fasta sequences;
2. A fast and parallel convex hull algorithm for convex hull generation and vertices identification derived from natural vectors;
3. A probabilistic model utilizing convex hull vertices combination weights for the prediction of genomic sequences;
4. Representation and validation of predicted genome sequences and gene mutations using convex hull combination constructs.
The proposed framework was empirically evaluated on the COVID-19 dataset,

demonstrating promising predictive performance and highlighting its potential applicability in genomic analysis.

Molecular Epidemiological Investigation of Adverse Late Effects among Survivors of Childhood Cancer

Zhaoming Wang(王朝明)
St. Jude Children's Research Hospital

The number of childhood cancer survivors in the US is expected to reach 580,000 by 2040 thanks to continued therapeutic improvement and innovation. It is imperative for the survivorship research community to develop personalized care and surveillance for this medically vulnerable population due to the multitude of long-term health issues. Over the past decade, genome-scale molecular profiling of large childhood cancer survivorship cohorts has led to unprecedented advances in our understanding of the genetic and epigenetic bases of therapy-related adverse health outcomes in this vulnerable population. To facilitate the integration of knowledge generated from these studies into formulating next-generation precision care for survivors of childhood cancer, I will review key findings of genetic and epigenetic studies of long-term therapy-related genotoxicities and adverse effects including many original investigations from my own laboratory. I will highlight aging biomarkers and accelerated aging among survivors of childhood cancer, leading to ongoing and future non-pharmacologic and pharmacologic interventions to slow down/reverse the aging trajectory. These insights will hopefully inspire future studies that harness both expanding omics resources and evolving data science methodology to accelerate translation of precision medicine for survivors of childhood cancer.

Chair: Zhaoming Wang(王朝明), St. Jude Children's Research Hospital

Testing the molecular models of modern human origins

Shi Huang(黄石)
Xiangya Medical School, Central South University

Testing the models of modern human origins Shi Huang, Center for Medical Genetics, School of Life Sciences, Central South University, 110 Xiangya Road, Changsha, Hunan 410078, P.R. China
Over the past four decades, molecular studies have yielded two primary models for understanding the uniparental DNA phylogenetic trees of modern humans: the Recent Out of Africa (ROA) and the Recent Out of East Asia (ROE) models. These models differ in their underlying assumptions, neutral theory versus the maximum genetic diversity theory, particularly in relation to early stem haplotypes, even though they share many haplotype relationships. We made use of ancient DNA data to test these two

models. In addition, leveraging the wealth of new genetic variants unveiled through the comprehensive sequencing of 43 diverse human Y chromosomes, we investigated the presence of shared variants among different haplotypes to determine which model better aligns with the genetic data. We validated our approach by confirming numerous well-established haplotype relationships that are consistent with both the ROA and ROE models. Remarkably, our analysis revealed a compelling pattern: we were able to corroborate the existence of stem haplotypes specific to the ROE model, but not those exclusive to the ROA model. For instance, we found that A0b and A1a shared the most variants with each other, aligning with the notion that both fall under the A00A1a stem haplotype of the ROE model. So, it becomes evident that these tests lend robust support to the ROE model as the more accurate representation of modern human origins.

A mechanic program for phenotypic evolution

Qi Wu(吴琦)
Beijing Institute of Mathematical Sciences and Applications (BIMSA)

In evolutionary biology, it is widely acknowledged that certain forces drive evolution and shape genetic diversity. However, a quantitative theoretical framework analogous to the four fundamental mechanics remains elusive.

We have observed three intriguing and pivotal aspects, where analogous similarities exist between phenotypic evolution and physical mechanical phenomena.

Building upon population and quantitative genetics theories—particularly the concepts of episodes of selection (fitness components of fertility and viability) and Fisher's geometric model—here we have devised a mechanical framework for the evolution of complex traits, drawing analogies with classical and quantum mechanics. This framework encompasses six fundamental assumptions and associated qualitative and quantitative subjects.

Within this framework, we have explored several significant topics and identified directions for future endeavors to enhance this program. While it is currently neither comprehensive nor rigorous, we hope that our efforts will provide valuable impetus and inspiration for the formulation of a unified theory bridging the life and physical sciences.

ChemGPT: An AI-Driven Molecular Synthesis Platform

Xiao He(何晓)
East China Normal University

In this talk, I will introduce the latest development from East China Normal University—ChemGPT 1.0. This includes the construction of a high-quality chemical dialogue dataset, where ChemGPT 1.0 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 1.0 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.

Chair: Yishuai Niu(牛一帅), Beijing Institute of Mathematical Sciences and Applications (BIMSA)

Early Warning Indicators for Critical Transitions in Stochastic Biological Systems

Jinqiao Duan(段金桥)
Great Bay University

The interactions of uncertainty and nonlinearity in biological dynamical systems lead to complex phenomena, such as critical transitions between qualitatively different dynamical regimes.

The speaker will overview recent advances in in detecting or predicting the most probable transitions between metastable regimes, and for applications in neural and other biophysical systems.

Exploring the convex hull principle in scRNA-seq scenarios

Tao Zhou(周涛)
Tsinghua University

The Convex hull is a fundamental and important concept in computational geometry that denotes the smallest convex polygon (polyhedron) that contains all the points in a point set. Taking the two-dimensional case as an example, figuratively speaking, if the points in a set are regarded as some nails on a board, then the convex hull is the figure formed by enclosing all the nails with a ring of rubber bands. In bioinformatics, through the natural vector representation of genome sequences proposed by Prof. Yau, we can convert the genome sequences of organisms into a series of points in a high-dimensional Euclidean space. Prof. Yau's previous work shows that the convex hulls formed by the natural vectors of the genome sequences of most organisms satisfy the convex hull principle, i.e., the convex hulls of the genomes of different families of organisms do not intersect with each other. We try to further extend this principle to scRNA-seq scenarios. By combining the natural vector representation of RNA sequences and scRNA transcriptional profiling data, we can validate the convex hull principle on single-cell data in different scenarios, which brings different perspectives from the traditional downscaling visualization methods such as tSNE and UMAP.

December 18, 2024 Wednesday

Chair: Shan Zhao(赵山), University of Alabama

Protein language models and their prompt-based learning

Dong Xu(许东)
University of Missouri

Protein language models (PLMs) offer powerful representations of protein sequences and their evolutionary patterns through pre-training on large-scale protein sequence datasets. We introduced S-PLM, a structure-aware PLM that enhances protein prediction by integrating both sequence and structural information. This model employs a multi-view contrastive learning strategy to align protein sequences with their structures at both the protein and residue levels within a shared embedding space. S-PLM uses a Swin-Transformer for contact map images of AlphaFold-predicted structures, combined with sequence-based embeddings from ESM2. Equipped with an extensive set of fine-tuning tools, S-PLM achieves superior prediction performance compared to other PLMs. In addition, we developed Prot2Seq to extend PLMs' capabilities for multitasking in protein prediction and design. Prot2Seq utilizes an autoregressive language modeling approach, where task-specific tokens are added to the decoder to enhance simultaneous multitasking training within a single model. Prot2Seq demonstrated improved performance in these scenarios. Furthermore, we implemented a Parameter-Efficient Fine-Tuning framework on the ESM2 model, employing various prompting methods such as Prompt Tuning, LoRA, and Adapter Tuning for tasks like predicting protein signal peptides and localization signals. These methods delivered significant improvements over state-of-the-art techniques, particularly when training data is limited. Our studies show great promise for PLMs and prompt-based learning in protein research.

Integrating differential operators and deep learning in biology application

Jiahui Chen(陈嘉会)
University of Arkansas

This talk will discuss differential operators and deep learing in biology applications and focus on the recent developments. The multiscale analysis of graph neural networks and the de Rham-Hodge theory provide a unified paradigm for the evolving manifolds constructed from filtration, which induces a family of evolutionary complexes. While the present evolutionary de Rham-Hodge method can be easily applied to close manifolds, the emphasis is given to more challenging compact manifolds with 2- manifold boundaries, which require appropriate analysis and treatment of boundary conditions on differential forms to maintain proper topological properties. Meanwhile, we will discuss the multiscale graph neural network in the modeling of biomolecules.

Investigation of biochemical reaction through quantum computing

Minghui Yang(杨明晖)
Innovation Academy of Precision Measurement Science and Technology, Chinese Academy of Sciences

One of the most promising applications of quantum computing lies in the field of quantum chemistry, particularly in solving the electronic Schrödinger equation. In this work, I will introduce: (1) the fundamentals of quantum chemistry and quantum computing; (2) an illustration of the quantum computing process by calculating the energy of the hydrogen molecule; and (3) our current research on biochemical reactions using quantum computing

Chair: Buyong Ma(马步勇), Shanghai Jiao Tong University

Modeling Single-cell and Spatial Transcriptomics Data Using Optimal Transport

Zixuan Cang(仓子晅)
North Carolina State University

Single-cell and spatial transcriptomics data examines high-throughput gene expression profiles at fine resolutions providing an unprecedented opportunity to elucidate the underlying complex biological processes. Optimal transport has proven to be an effective tool for various applications with such data, such as multi-omics integration. In this talk, we will discuss several optimal transport variants motivated by the biological applications, where there are detailed application-specific constraints, multiple distribution species, and multiple embedding spaces of the same system. We will illustrate the applications of these tools for addressing multi-compatible molecular species in cell-cell communication analysis and devising coherent trajectories of the same biological system from multi-omics datasets.

Asymmetric Natural Vector Method for Predicting Ambiguous Non-standard Base Codes

Guoqing Hu(胡国庆)
Beijing Institute of Mathematical Sciences and Applications (BIMSA)

In this report, we introduce a novel approach based on the Asymmetric Natural Vector (ANV) method to address the problem of ambiguity in DNA sequences.
We propose using ANV to predict the bases represented by non-standard codes in DNA sequences. Our approach involves developing a deep learning framework to establish a correspondence between DNA sequences (in FASTA format) and natural vectors, which encode relevant sequence properties. By training on a large dataset, we learn the distribution of these ambiguous base codes within the dataset.
This method allows us to accurately predict masked or ambiguous nucleotide bases in genomic fragments. It is particularly applicable to datasets, such as the COVID-19 genome data, which contain numerous non-standard base codes like R, Y, S, W, K, M, B, D, H, and V. By employing our algorithm, we can effectively estimate the corresponding standard bases and assign confidence scores to each prediction, aiding in the resolution of sequencing uncertainties.

December 19, 2024 Thursday

Chair: Jie Wu(吴杰), Beijing Institute of Mathematical Sciences and Applications (BIMSA)

Learning governing equations of cellular dynamics from single cell data

Jianhua Xing(邢建华)
University of Pittsburgh

High-throughput techniques, especially at the single cell level, have greatly expanded our knowledge of cellular processes. With the increasing availability of data, a fundamental question arises: how can we leverage this data to gain mechanistic insights? Unlike static data typically targeted by statistics-based machine learning approaches, single cell data are snapshots from the dynamical state space of a cell having interacting components that dictate the temporal evolution of the system. Consequently, we witness a growing convergence of two disciplines: data science and systems biology. The latter seeks to unravel qualitative and quantitative causal relationships among cellular components, as well as their functions within the broader context of cell regulatory

networks, all through the lens of dynamical systems theory. Consequently, an emerging new direction is to integrate the vast amount available single cell data into systems biology modeling to study complex cellular processes.
During this presentation, I will delve into our endeavors on deducing the comprehensive governing dynamical equations of cells using both snapshot and time series single-cell data. I will first briefly summarize our published work, then focus on a set of new-generation approaches. Distinguishing our approach from simple data manifold learning, reconstructing a dynamical system imposes additional constraints on the manifold and its associated tangent space. Furthermore, it necessitates dynamical and topological invariance under representation transformation, a fundamental principle in theoretical physics.

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

Xinqi Gong(龚新奇)
Renmin University

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.

A new boundary condition for the nonlinear Poisson-Boltzmann equation in electrostatic analysis of proteins

Shan Zhao(赵山)
University of Alabama

As a well-established implicit solvent model, the Poisson-Boltzmann equation (PBE) models the electrostatic interactions between a solute biomolecule and its surrounding solvent environment over an unbounded domain. One numerical challenge in solving the nonlinear PBE lies in the boundary treatment. Physically, the boundary condition of this solute solvent system is defined at infinity where the electrostatic potential decays to zero. Computationally, a finite domain has to be employed in grid-based numerical algorithms. However, the Dirichlet boundary conditions commonly used in protein simulations are known to produce unphysical solutions in some special cases. This motivates the development of a few asymptotic conditions in the PBE literature, which are global boundary conditions and have to resort to iterative algorithms for calculating volume integrals from the previous step. To overcome these limitations, a simple Robin condition is proposed in this work as a local boundary condition for the nonlinear PBE, which can be implemented in any finite difference or finite element method. The derivation is based on the facts that away from the biomolecule, the asymptotic decaying pattern of the nonlinear PBE is essentially the same as that of the linearized PBE, and the monopole term will dominate other terms in the multipole expansion. Asymptotic analysis has been carried out to validate the application range and robustness of the proposed Robin condition. Moreover, a second order boundary implementation by means of a matched interface and boundary (MIB) scheme has been constructed for three-dimensional biomolecular simulations. Extensive numerical experiments have been conducted to examine the robustness, accuracy, and efficiency of the new boundary treatment for calculating electrostatic free energies of Kirkwood spheres and various protein systems.

Chair: Dong Xu(许东), University of Missouri

Integrated approach to DNA 3D structure prediction

Yi Xiao(肖奕)
Huazhong University Science and Technology

The accuracy of current methods of DNA 3D structure prediction cannot be compared with that of protein yet. Here we propose an integrated method to accurately and rapidly predict DNA 3D structures by combining the traditional template-based and molecular-dynamics-based methods with deep-learning method to overcome the limitation of available DNA experimental structures and aligned homologous sequences. The benchmarks demonstrate that the accuracy of our method is much higher than AlphaFold3.

Mathematical AI for virtual screening of drug discovery

Jian Jiang(江健)
Wuhan Textile University

Pain is a significant global health issue, and the current treatment options for pain management have limitations in terms of effectiveness, side effects, and potential for addiction. There is a pressing need for improved pain treatments and the development of new drugs. Voltage-gated sodium channels, particularly Nav1.3, Nav1.7, Nav1.8, and Nav1.9, play a crucial role in neuronal excitability and are predominantly expressed in the peripheral nervous system. Targeting these channels may provide a means to treat pain while minimizing central and cardiac adverse effects. In this study, we construct protein–protein interaction (PPI) networks based on pain-related sodium channels and develop a corresponding drug–target interaction network to identify potential lead compounds for pain management. To ensure reliable machine learning predictions, we carefully select 111 inhibitor data sets from a pool of more than 1000 targets in the PPI network. We employ 3 distinct machine learning algorithms combined with advanced natural language processing (NLP)–based embeddings, specifically pretrained transformer and autoencoder representations. Through a systematic screening process, we evaluate the side effects and repurposing potential of more than 150,000 drug candidates targeting Nav1.7 and Nav1.8 sodium channels. In addition, we assess the ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of these candidates to identify leads with near-optimal characteristics. Our strategy provides an innovative platform for the pharmacological development of pain treatments, offering the potential for improved efficacy and reduced side effects.

Chair: Yi Xiao(肖奕), Huazhong University Science and Technology

IntComplex for high-order interactions

Jie Wu(吴杰)
Beijing Institute of Mathematical Sciences and Applications (BIMSA)

Graphs serve as powerful tools for modeling pairwise interactions in diverse fields such as biology, material science, and social networks. However, they inherently overlook many-body interactions involving more than two entities. Simplicial complexes and hypergraphs have emerged as prominent frameworks for modeling manybody interactions; nevertheless, they exhibit limitations in capturing specific high-order interactions, particularly those involving transitions from n-interactions to m-interactions. Addressing this gap, we propose IntComplex as an innovative framework to characterize such high-order interactions comprehensively. Our framework leverages homology theory to provide a quantitative representation of the topological structure inherent in such interactions. Furthermore, we introduce persistent homology through a filtration process and establish its stability to ensure robust quantitative analysis of these complex interactions. The proposed IntComplex framework introduces a foundational paradigm for analyzing topological properties in high-order interactions, showcasing significant potential to advance the field of complex network analysis.

Ensemble average solvation energy functional and its computational model

Yuanzhen Shao(邵元桢)
University of Alabama

In 2023, Zhan Chen, Shan Zhao, and I introduced the first variational implicit solvent model for computing ensemble average solvation energy. This model, formulated as a total variation functional with an obstacle constraint, effectively captures the stochastic behavior of the solute-solvent interface due to thermodynamic fluctuations. However, the non-differentiability of the functional and the presence of the obstacle pose significant computational challenges.
In this talk, I will present a p p-energy computational model designed to address these difficulties. I will discuss key results, including the existence and uniqueness of solutions, as well as the gradient flows of the model. If time permits, I will also cover recent advancements in the modeling of ensemble average solvation energy functionals.

Deep learning methods for elliptic problems and the applications in predicting electrostatics

Jinyong Ying(应金勇)
Central South University

In this talk, we mainly introduce the deep learning methods for solving elliptic problems, including elliptic interface problems and advection-dominated problems. Here for different types of problems, slightly different deep learning methods are presented, including the method based on the randomized neural networks as well as the corresponding convergence analysis. Meanwhile, different types of adaptive methods are also introduced for the purpose of improving the solution accuracy. Using the deep learning methods to solve PBE and SMPBE are briefly presented as one application.

Chair: Xinqi Gong(龚新奇), Renmin University

Physics-principle-based prediction for RNA 3D structures

Zhijie Tan(谭志杰)
Wuhan University

RNAs are important biomolecules with crucial biological functions such as gene regulation and catalysis, and the functions of RNAs are closely dependent on their structures and structural stability. Therefore, to predict 3D structures and stability of RNAs is helpful for a deep understanding RNA and related applications of RNA functions.

Recently, we have developed a residue-separation-based scoring function (cgRNASP) for RNA 3D structure evaluation, a coarse-grained force-field-based model for predicting RNA 3D structures (TiRNA), and a fragment-ensemble assembly model for building RNA 3D structures (FebRNA+). First, the performance of cgRNASP is superior to other existing statistical potentials and scoring functions, and very importantly, cgRNASP is strikingly efficient than other top scoring functions. Second, TiRNA involves a coarse-grained force field with the effects of temperature and ions, and the model can predict 3D structures and thermal stability of RNAs in monovalent/divalent ion solutions. Third, FebRNA1 can make consistently reliable 3D structure predictions for RNAs with lengths ranging from tens of nucleotides to thousands of nucleotides, as compared with the existing top traditional prediction methods and the newly developed AI-based methods such as AlphaFold3. The packages of cgRNASP, TiRNA, and FebRNA+ are available at https://github.com/Tan-group. The further developments of our models are still ongoing.

IGHV3-53 Sequence Characteristics and Neutralizing Activity: Predictive Modeling and Broad-Spectrum Antibody Design Against SARS-CoV-2

Xin Zhao(赵鑫)
Beijing Institute of Mathematical Sciences and Applications (BIMSA)

The IGHV3-53 gene plays a pivotal role in the development of neutralizing antibodies, particularly against SARS-CoV-2. This study investigates the sequence characteristics of IGHV3-53-derived antibodies with varying neutralizing activities to identify key structural and functional determinants that contribute to their potency and breadth. Using a combination of computational bioinformatics, predictive modeling, and experimental approaches, the research develops a functional prediction framework to assess sequence variation impacts on antibody performance. Key amino acid residues and structural motifs critical for neutralization are identified, forming the basis for the rational design of novel broad-spectrum neutralizing antibodies targeting SARS-CoV-2 and its variants. This work provides insights into IGHV3-53's functional mechanisms and establishes a foundation for next-generation antibody therapeutics aimed at combating rapidly evolving viral pathogens.

December 20, 2024 Friday

Chair: Jianhua Xing(邢建华), University of Pittsburgh

Deep learning of protein energy landscape and conformational dynamics from experimental structures in PDB

Buyong Ma(马步勇)
Shanghai Jiao Tong University

Protein structure prediction has reached revolutionary levels of accuracy, implying biophysical energy function can be learned from known protein structures. However apart from single static structure, conformational distributions and dynamics often control protein biological functions. Towards this goal, we develop DeepConformer, a diffusion generative model for sampling protein conformation distributions from a given amino acid sequence. Despite the lack of molecular dynamics (MD) simulation data in training process, DeepConformer captured conformational flexibility similar to MD simulation and reproduced experimentally observed conformational variations. Our study demonstrated that DeepConformer learned energy landscape can be used to efficiently explore protein conformational distribution and dynamics.

Topology-enhanced machine learning model for anticancer peptide prediction

Xue Gong(公雪)
Nanyang Technological University

Recently, therapeutic peptides have demonstrated great promise for cancer treatment. To explore powerful anticancer peptides, artificial intelligence (AI)-based approaches have been developed to systematically screen potential candidates. However, the lack of efficient featurization of peptides has become a bottleneck for these machine-learning models. In this paper, we propose a topology-enhanced machine learning model (Top-ML) for anticancer peptide prediction. Our Top-ML employs peptide topological features derived from its sequence "connection" information characterized by vector and spectral descriptors. Our Top-ML model has been validated on two widely used AntiCP 2.0 benchmark datasets and has achieved state-of-the-art performance. Our results highlight the potential of leveraging novel topology-based featurization to accelerate the identification of anticancer peptides. The talk is based on a joint work with Joshua Zhi En Tan, JunJie Wee, and Kelin Xia.

Quantitative cancer-immunity cycle modeling for predicting disease progression in advanced metastatic colorectal cancer

Jinzhi Lei(雷锦志)
Tiangong University

Patients diagnosed with advanced metastatic colorectal cancer often exhibit heterogeneous disease progression and face poor survival prospects. In order to comprehensively analyze the varied treatment responses among individuals and the challenge of tumor recurrence resistant to drugs in advanced mRCR, we devel- oped a novel quantitative cancer-immunity cycle model. The proposed model was meticulously crafted utilizing a blend of differential equations and randomized modeling techniques to quantitatively elucidate the intricate mechanisms governing the cancer-immunity cycle and forecast tumor dynamics under different treatment modalities. Furthermore, by integrating diverse clinical datasets and rigorous model analyses, we introduced two pivotal concepts: the treatment response index and the death probability function. These concepts are crucial tools for translating model predictions into clinically relevant evaluation indexes. Using virtual patient technology, we extrapolated tumor predictive biomarkers from the model to predict survival outcomes for mCRC patients. Our findings underscore the significance of tumor-infiltrating CD8+ CTL cell density as a key predictive biomarker for short-term treatment responses in advanced mCRC while empha- sizing the potential predict value of the tumor-infiltrating CD4+ Th1/Treg ratio in determining patient survival. This study presents a pioneering methodology bridging the divide between diverse clinical data sources and the generation of virtual patients, offering invaluable insights into understanding inter-individual treatment variations and forecasting survival outcomes in mCRC patients.