Photo: Slava Gerovitch

Thirteenth Annual Fall-Term PRIMES Conference

October 14-15, 2023

Thirteenth Annual Fall Term PRIMES Conference, October 14-15, 2023

Saturday, October 14

Mathematics

Room 4-370 , MIT

9:00 am: Welcoming Remarks

- Prof. Michel Goemans, Head of the MIT Mathematics Department
- Prof. Pavel Etingof, PRIMES Chief Research Advisor
- Dr. Slava Gerovitch, PRIMES Program Director

9:15-10:15 am: Session 1. Mathematics and Computer Science

- Catherine Li, Machine Learning and Gradient Descent for Infectious Disease Risk Prediction (mentor Daniel Lazarev)
- Eric Wang, Speeding up and Reducing Memory for Scientific Machine Learning via Mixed Precision (mentor Prof. Lu Lu, University of Pennsylvania)
- Michelle Wei, Solving Second-order Cone Programs in Matrix Multiplication Time (mentor Guanghao Ye)
- Coleman DuPlessie and Aidan Gao, Stability Techniques in Differentially Private Machine Learning (mentor Hanshen Xiao)

10:30-11:45 am: Session 2. Algebra

- Michael Yang, The Rank and Rigidity of Group-Circulant Matrices (mentor Dr. Minh-Tâm Trinh)
- Rohan Das, Christopher Qiu, and Shiqiao Zhang, The Distribution of the Cokernel of a Random Integral Symmetric Matrix Modulo a Prime Power (mentors Prof. Gilyoung Cheong and Prof. Nathan Kaplan, University of California Irvine)
- Henry Jiang, Shihan Kanungo, and Harry Kim, On a Weaker Notion of the Finite Factorization Property (mentors Prof. Jim Coykendall, Clemson University, and Dr. Felix Gotti, MIT)
- Joshua Jang, Jason Mao, and Skyler Mao, Betti Graphs and Atomization of Puiseux Monoids (mentors Prof. Scott Chapman, SHSU, and Dr. Felix Gotti, MIT)

12:00-1:00 pm: Session 3. Applied Mathematics

- Raina Wu, Link Prediction and Influencer Identification on Weighted Graphs (mentor Prof. Laura Schaposnik, University of Illinois at Chicago)
- Steve Zhang, Absolute Tectonic Plate Motion Optimization (mentor Prof. James Unwin, University of Illinois at Chicago)
- Alex Zhao, Oscillating Near Circles at Intermediate Reynolds Numbers (mentor Dr. Nicholas J. Derr)
- Roger Fan, Multidisperse Random Sequential Adsorption and Generalizations (mentor Nitya Mani)

2:00-3:10 pm: Session 4. Combinatorics

- David Dong, On Generalized Eulerian Numbers (mentor Dr. Tanya Khovanova)
- Srinivas Arun, Improved Bounds on Helly Numbers of Exponential Lattices (mentor Travis Dillon)
- Aryan Bora and Lucas Tang, On the Spum and Sum-Diameter of Paths (mentor Yunseo Choi, Harvard University)
- Evan Chang and Neel Kolhe, Upper Bounds on the 2-colorability Threshold of Random d-regular k-uniform Hypergraphs for $k \geq 3$ (mentor Dr. Youngtak Sohn)

3:20-4:05 pm: Session 5. Combinatorics

- Linus Tang, Extremal Bounds on Peripherality Measures (mentor Dr. Jesse Geneson, SJSU)
- Eric Zhan, On the Winning and Losing Parameters of Schmidt Games (mentor Vasiliy Nekrasov)
- Raymond Luo, Cyclic Base Orderings and Equitability of Matroids (mentor Yuchong Pan)

4:15-5:00 pm: Session 6. Category Theory

- William Gvozdjak, On the Classification of Low-Rank Odd-Dimensional Modular Categories (mentors Prof. Julia Plavnik, Indiana University Bloomington, and Agustina Czenky, University of Oregon)
- Akshaya Chakravarthy, On Modular Categories with Frobenius-Perron Dimension Congruent to 2 Modulo 4 (mentors Prof. Julia Plavnik, Indiana University Bloomington, and Agustina Czenky, University of Oregon)

- Joseph Vulakh, Simple Racks over the Alternating Groups (mentors Prof. Julia Plavnik and Dr. Héctor Peña Pollastri, Indiana University Bloomington)

Sunday, October 15

Mathematics (in parallel with Computer Science sessions in the morning)

Room 4-370 , MIT

9:00-10:15 am: Session 7. Algebra

- Ethan Liu, On the Structure and Generators of the *n*-th Order Chromatic Algebra (mentor Merrick Cai)
- Justin Zhang, An Extension of Benson's Conjecture to Finite 3-Groups for Monomial Modules with Null Inner Partition (mentor Dr. Kent Vashaw)
- Andrew Lin, Henrick Rabinowitz, and Qiao Zhang, The Furstenberg Property in Puiseux Monoids (mentor Dr. Felix Gotti)
- Hannah Fox, Agastya Goel, and Sophia Liao, Arithmetic of semisubtractive semidomains (mentor Prof. Harold Polo, University of Florida)

10:25-11:30 am: Session 8. Geometry and Other Topics

- Anton Levonian, Existence of Circle Packings on Certain Translation Surfaces (mentor Prof. Sergiy Merenkov, CCNY – CUNY)
- Nicholas Hagedorn, Algorithmically Generated Pants Decompositions of Combinatorial Surfaces (mentor Elia Portnoy)
- Alexander Li, Canonical Forms for Toric and Surface Codes in ZX Calculus (mentor Andrey Boris Khesin)
- Iz Chen and Krishna Pothapragada, Classification of Non-degenerate Symmetric Bilinear Forms in the Verlinde Category Ver4+ (mentor Arun Kannan)

11:40 am - 12:40 pm: Session 9. Group Theory and Representations

- Matvey Borodin, The Action of the Cactus Group on Arc Diagrams (mentor Prof. Leonid Rybnikov)
- Alan Bu, The Local-Global Principle and a Projective Twist on the Hasse Norm Theorem (mentor Dr. Thomas Rüd)

- Brian Li, Tensor Product Decompositions for Modules over Subregular W-Algebras (mentor Dr. Artem Kalmykov)
- Razzi Masroor, Hyperoctahedral Schur Algebra and the Hyperoctahedral Web Algebra (mentor Dr. Elijah Bodish)

Computer Science (in parallel with Math sessions in the morning)

Room 4-270 , MIT

9:00 am: Welcoming Remarks

- Dr. Slava Gerovitch, PRIMES Program Director
- Prof. Srini Devadas, PRIMES Computer Science Section Coordinator

9:10-10:05 am: Session 10. Computer Science a

- Dongchen Zou, Intersection Attack in Non-Uniform Setting (mentor Simon Langowski)
- Alan Song and Evan Ning, Exploring Data-driven Resource Management for Serverless Systems (mentors Varun Gohil, Nikita Lazarev, and Yueying (Lisa) Li)
- Rohith Raghavan and Eric Chen, Comparing Methods of Opportunistic Risk-Limiting Audits (mentor Mayuri Sridhar)

10:20-11:25 am: Session 11. Computer Science

- Boyan Litchev, Parallelizable and Updatable Private Information Retrieval (mentor Simon Langowski)
- Andrew Carratu and Albert Lu, Public-key Signature Scheme with Reduced Hardware Trust (mentors Sacha Servan-Schreiber and Jules Drean)
- Yifan Kang, NUMA-Aware Data Structure Design & Benchmarking (mentors Shangdi Yu and Prof. Julian Shun)
- Omar El Nesr, Fast GPU Accelerated Ising Models for Practical Combinatorial Optimization (mentor Axel Feldmann)

11:40 am - 12:50 pm: Session 12. Computer Science

- Sarah Pan, Let's Reinforce Step by Step (mentors Prof. Anna Rumshisky and Vlad Lialin, UMass Lowell)
- Eddie Wei, The Algebraic Value-Editing Conjecture in Deep Reinforcement Learning (mentor Andrew Gritsevskiy, Cavendish Labs)

- Henry Han and Adrita Samanta, Visualizing Distributed Traces in Aggregate (mentors Darby Huye, Max Liu, Roy Zhang, and Prof. Raja Sambasivan, Tufts University)
- Garima Rastogi and Sophia Lichterfeld, How Do I Pay Thee? Let Me Count the Ways: Leveraging Ethereum Smart Contracts to Facilitate Web Monetization Adoption (mentor Kyle Hogan)

Computational and Physical Biology and Bioinformatics

Room 4-270 , MIT

2:00-3:15 pm: Session 13: Computational and Physical Biology

- Valentina Zhang, Identifying Microglial Heterogeneity in Alzheimer's Disease (mentor Dr. Ayshwarya Subramanian, Broad Institute)
- Raj Saha, Surveying the Presence and Diversity of Viruses in Mammalian Transcriptomes (mentor Dr. Ayshwarya Subramanian, Broad Institute)
- Amith Saligrama, A Novel Statistical Framework for Identification of Mutated Cells (mentors Dr. Giulio Genovese, Broad Institute, and Prof. Steve McCarroll, Harvard Medical School)
- Anna Du, Utilizing Machine Learning to Identify Time Asymmetry of DNA Loop Extrusion (mentors Dr. Aleksandra Galitsyna and Henrik Pinholt)
- Elizaveta Rybnikova, Exploration of Hi-C Patterns Through Computer Simulations (mentors Dr. Aleksandra Galitsyna and Henrik Pinholt)

3:25-4:40 pm: Session 14: Bioinformatics

- Irene Jiang, Machine Learning Inference of Causal Genes for Tuberculosis using Mendelian Randomization and Single-cell Sequencing (mentor Prof. Gil Alterovitz, Harvard Medical School)
- Gavin Ye, Drug Design as Causal Language Modeling: Transferring Large Chemistry Models for De-Novo Drug Design with Supervised and Reinforcement Learning (mentor Prof. Gil Alterovitz, Harvard Medical School)
- Aaron Li, Identification of Biomarkers for Insulin Resistance using Meta-Analysis and Machine Learning (mentor Prof. Gil Alterovitz, Harvard Medical School)
- Stephanie Wan, Biomarker Identification of Oral Squamous Cell Carcinoma through Transcriptomic Expression Analysis (mentor Prof. Gil Alterovitz, Harvard Medical School)
- Rianna Santra, Transcriptomic Analysis of the Dengue Virus using Feature Selection and Random Forest (mentor Prof. Gil Alterovitz, Harvard Medical School)

2023 PRIMES FALL-TERM CONFERENCE ABSTRACTS

SATURDAY, OCTOBER 14

SESSION 1: MATHEMATICS AND COMPUTER SCIENCE

Catherine Li

Machine Learning and Gradient Descent for Infectious Disease Risk Prediction

Mentor: Daniel Lazarev

With the recent outbreaks of infectious diseases, methods of predicting the trajectory and risk factors of these diseases through machine learning and probabilistic models have become more pertinent. In this talk, we introduce epidemiology and disease models, as well as several factors of transmission for infectious diseases. Building on previous work that uses exponential risk scores to project upcoming virus variants, we design a new Geo Score that incorporates geographic, socioeconomic, and demographic factors to estimate infection and mortality risk by region and time. We assess the predictiveness of several such factors, then employ gradient descent to find the optimal weights of these factors in determining risk. Finally, we demonstrate the accuracy and improved interpretability of our results, as compared to current state-of-the-art techniques.

Eric Wang

Speeding up and Reducing Memory for Scientific Machine Learning via Mixed Precision

Mentor: Prof. Lu Lu, University of Pennsylvania

Scientific machine learning has emerged as a versatile approach to addressing complex scientific problems. Within this field, physics-informed neural networks (PINNs) have emerged as an effective method in solving partial differential equations (PDEs) by embedding a PDE residual into the loss function of a deep neural network. One drawback to PINNs is that they are computationally expensive to train for large problems. Modern GPUs are able to work with half precision data types that can be leveraged to reduce training time and memory usage. In this talk, we propose a mixed precision method of training PINNs that reduces computational cost while retaining accuracy.

Michelle Wei

Solving Second-order Cone Programs in Matrix Multiplication Time

Mentor: Guanghao Ye

We propose a deterministic algorithm for solving second-order cone programs of the form

$$\min_{Ax=b, x\in\mathcal{L}_1\times\cdots\times\mathcal{L}_r} c^\top x,$$

which optimize a linear objective function over the set of $x \in \mathbb{R}^n$ contained in the intersection of an affine set and the product of r second-order cones. Our algorithm achieves a runtime of $\widetilde{O}((n^\omega + n^{2+o(1)}r^{1/6} + n^{2.5-\alpha/2+o(1)})\log(1/\epsilon))$, where ω and α are the exponents of matrix multiplication, and ϵ is the relative accuracy. For the current values of $\omega \sim 2.37$ and $\alpha \sim 0.32$, our algorithm takes $\widetilde{O}(n^\omega \log(1/\epsilon))$ time. This nearly matches the runtime for solving the sub-problem $Ax = b$. To the best

of our knowledge, this is the first improvement on the computational complexity of solving second-order cone programs after the seminal work of Nesterov and Nemirovski on general convex programs. For $\omega = 2$, our algorithm takes $\widetilde{O}(n^{2+o(1)}r^{1/6}\log(1/\epsilon))$ time.

To obtain this result, we utilize several new concepts that we believe may be of independent interest:

- We introduce a novel reduction for splitting ℓ_p-cones.
- We propose a deterministic data structure to efficiently maintain the central path of interior point methods for general convex programs.

Coleman DuPlessie and Aidan Gao

Stability Techniques in Differentially Private Machine

Mentor: Hanshen Xiao

The field of machine learning has exploded in recent years, and with it, concerns about data privacy. Differential privacy is a guideline for machine learning models trained on private data that guarantees that the data remains private. However, large accuracy losses result unless models are stable, meaning that the addition or removal of a small amount of training data leads to a bounded, and ideally a small, change in the final model. In this presentation, we use the CIFAR-10 image recognition benchmark to investigate a variety of techniques to make differentially private machine learning preserve accuracy through the optimization of stability, including model optimization and pruning, binary trees composed of models, and data preprocessing. We focus on the effective stabilization of linear regressions, which creates gateways for minimal accuracy loss when optimizing more complex algorithms.

Session 2: Algebra

Michael Yang

The Rank and Rigidity of Group-Circulant Matrices

Mentor: Dr. Minh-Tâm Trinh

Circulant matrices are matrices whose applications range from signal processing to image compression. In this talk, we study properties of circulant matrices and of their generalizations, called group-circulant matrices. We first state a formula for the ranks of group-circulant matrices, for which we have given a new proof. We then discuss matrix rigidity, a concept motivated by theoretical computer science, and new results about the non-rigidity of nonabelian group-circulants.

Rohan Das, Christopher Qiu, and Shiqiao Zhang

The Distribution of the Cokernel of a Random Integral Symmetric Matrix Modulo a Prime Power

Mentors: Prof. Gilyoung Cheong and Prof. Nathan Kaplan, University of California Irvine

Given a prime p and positive integers n and k, consider the ring $\mathrm{M}_n(\mathbb{Z}/p^k\mathbb{Z})$ of $n \times n$ matrices over $\mathbb{Z}/p^k\mathbb{Z}$. We investigate the distribution of the cokernel of a random symmetric matrix uniformly selected from $\mathrm{M}_n(\mathbb{Z}/p^k\mathbb{Z})$. We prove a symmetric analogue of the result of Cheong, Liang, and Strand by adapting their methods. Our result leads to a refined version of a result of Clancy, Kaplan, Leake, Payne, and Wood. We are working on an analogous result for alternate (i.e., skew-symmetric) matrices.

Henry Jiang, Shihan Kanungo, and Harry Kim

On a Weaker Notion of the Finite Factorization Property

Mentors: Prof. Jim Coykendall, Clemson University, and Dr. Felix Gotti, MIT

An (additive) commutative monoid is called atomic if every given non-invertible element can be written as a finite sum of atoms (i.e., irreducible elements), in which case, such a sum is called a factorization of the given element. The number of atoms (counting repetitions) in the corresponding sum is called the length of the factorization. Following Geroldinger and Zhong, we say that an atomic monoid M is a length-finite factorization monoid if each $b \in M$ has only finitely many factorizations of any prescribed length. An additive submonoid consisting of nonnegative real numbers is called a positive monoid. Factorizations in positive monoids have been actively studied in recent years. The main purpose of this talk is to give a better understanding of the non-unique factorization phenomenon in positive monoids through the lens of the length-finite factorization property. To do so, we identify a large class of positive monoids which satisfy the length-finite factorization property. Then we compare the length-finite factorization property to the bounded and the finite factorization properties, which are two properties that have been systematically investigated for more than thirty years.

Joshua Jang, Jason Mao, and Skyler Mao

Betti Graphs and Atomization of Puiseux Monoids

Mentors: Prof. Scott Chapman, SHSU, and Dr. Felix Gotti, MIT

Let M be a Puiseux monoid, that is, a monoid consisting of nonnegative rationals (under addition). A nonzero element of M is called an atom if its only decomposition as a sum of two elements in M is the trivial decomposition (i.e., one of the summands is 0), while a nonzero element $b \in M$ is called atomic if it can be expressed as a sum of finitely many atoms allowing repetitions: this sum of atoms is called an (additive) factorization of b. The monoid M is called atomic if every nonzero element of M is atomic. In this talk, we study factorizations in atomic Puiseux monoids through the lens of their associated Betti graphs. The Betti graph of $b \in M$ is the graph whose vertices are the factorizations of b with edges between factorizations that share at least one atom. Betti graphs have been useful in the literature to understand several factorization invariants in the more general class of atomic monoids.

Session 3: Applied Mathematics

Raina Wu

Link Prediction and Influencer Identification on Weighted Graphs

Mentor: Prof. Laura Schaposnik, University of Illinois at Chicago

A major question in the study of social networks is the influence maximization problem, which seeks the set of k nodes that maximizes information spread. However, the same set of nodes can produce different results when using different information diffusion models (e.g. simple contagion, complex contagion). Local, heuristic metrics in link prediction and influencer identification allow for predictions of future edges and gauge the influence of nodes respectively with relatively low time complexity. This talk explores how link prediction can be used in tandem with common centrality metrics to identify potential future influencers. We then consider the performances of different combinations of metrics using two contagion models.

Steve Zhang

Absolute Tectonic Plate Motion Optimization

Mentor: Prof. James Unwin, University of Illinois at Chicago

The theory of continental drift states that Earth's continents were once united in a supercontinent called Pangea, which began breaking apart about 200 million years ago. Plate motion models enhance the precision of tectonic research, spanning both local and global scales. In this talk, we review a code which optimizes an absolute plate motion model via the minimization of an objective function incorporating three different constraints. We then focus on one of the constraints, hotspot trail misfit, demonstrating how modifications to existing optimization procedures can drastically improve modeling results.

Alex Zhao

Oscillating Near Circles at Intermediate Reynolds Numbers

Mentor: Dr. Nicholas J. Derr

At low Reynolds number (Re), viscous forces dominate inertial effects, and Purcell's scallop theorem requires broken symmetries for a system to evolve via oscillating drivers. At high Re, inertial forces dominate. At intermediate Reynolds numbers, not one of viscous or inertial forces can dominate the other. A number of recent works have investigated structure assembly via oscillation-driven steady flows at intermediate Reynolds numbers. Here, we analyze the case of confined fluid in an oscillating rectangular domain around two fixed ellipses. A perturbation analysis decomposes the problem into a series of linear problems, which are solved using the finite element method. Force and torque on each ellipse are computed for varying geometric positions and Reynolds numbers.

Roger Fan

Multidisperse Random Sequential Adsorption and Generalizations

Mentor: Nitya Mani

We present a unified study of the limiting density in one-dimensional random sequential adsorption (RSA) processes where segment lengths are drawn from a given distribution. In addition to generic bounds, we are also able to characterize specific cases, including multidisperse RSA, in which we draw from a finite set of lengths, and power-law RSA, in which we draw lengths from a power-law distribution.

SESSION 4: COMBINATORICS

David Dong

On Generalized Eulerian Numbers

Mentor: Dr. Tanya Khovanova

Let $A(n,m)$ denote the Eulerian numbers, which count the number of permutations on $[n]$ with exactly m ascents. It is well known that $A(n,m)$ also counts the number of permutations on $[n]$ with exactly m excedances, due to the Foata transform.

We generalize these numbers to the form $A_r(n,m,k)$, which count the number of permutations on $[n]$ with exactly m ascents and the last element k. A Foata-like transform proves that this is equal to the number of permutations on $[n]$ with exactly m and last element $n+1-k$. We then mention several other results regarding the numbers $A_r(n,m,k)$, including a generalization of Worpitzky's identity.

Srinivas Arun

Improved Bounds on Helly Numbers of Exponential Lattices

Mentor: Travis Dillon

The Helly number $h(S)$ of a set $S \subseteq \mathbb{R}^d$ is defined as the smallest positive integer h, if it exists, such that the following statement is true: for any finite family of convex sets in S, if every subfamily of h sets intersect in S, then all sets in the family intersect in S. We focus on Helly numbers of product sets, which are sets of the form A^d for some one-dimensional set A.

Inspired by Dillon's research on the Helly numbers of product sets, Ambrus, Balko, Frankl, Jung, and Naszodi recently obtained the first upper and lower bounds for Helly numbers of exponential lattices in two dimensions, which are sets of the form $S = \{\alpha^n : n \in \mathbb{N}\}^2$ for some $\alpha > 1$. We use a different, simpler method to obtain better upper bounds for exponential lattices. In addition, we generalize the construction that lower-bounds the Helly number of exponential lattices of Ambrus et al. to higher dimensions.

Aryan Bora and Lucas Tang

On the Spum and Sum-Diameter of Paths

Mentor: Yunseo Choi, Harvard University

In a sum graph, the vertices are labeled with distinct positive integers, and two vertices are adjacent if and only if the sum of their labels is equal to the label of another vertex. The spum of a graph G is defined as the minimum difference between the largest and smallest labels of a sum graph that consists of G and a minimum number of isolated vertices. More recently, Li introduced the sum-diameter of a graph G, which modifies the definition of spum by removing the requirement that the number of isolated vertices must be minimal. In this talk, we settle conjectures by Singla, Tiwari, and Tripathi and Li by evaluating the spum and the sum-diameter of paths.

Evan Chang and Neel Kolhe

Upper bounds on the 2-colorability threshold of random d-regular k-uniform hypergraphs for $k \geq 3$

Mentor: Dr. Youngtak Sohn

For a large class of random constraint satisfaction problems (CSP), deep but non-rigorous theory from statistical physics predict the location of the sharp satisfiability transition. The works of Ding, Sly, Sun (2014, 2016) and Coja-Oghlan, Panagiotou (2014) established the satisfiability threshold for random regular k-NAE-SAT, random k-SAT, and random regular k-SAT for large enough $k \geq k_0$ where k_0 is a large non-explicit constant. Establishing the same for small values of $k \geq 3$ remains an important open problem in the study of random CSPs.

In this work, we study two closely related models of random CSPs, namely the 2-coloring on random d-regular k-uniform hypergraphs and the random d-regular k-NAE-SAT model. For every $k \geq 3$, we prove that there is an explicit $d_\star(k)$ which gives a satisfiability upper bound for both of the models. Our upper bound $d_\star(k)$ for $k \geq 3$ matches the prediction from statistical physics for the hypergraph 2-coloring by Dall'Asta, Ramezanpour, Zecchina (2008), thus conjectured to be sharp. Moreover, $d_\star(k)$ coincides with the satisfiability threshold of random regular k-NAE-SAT for large enough $k \geq k_0$ by Ding, Sly, Sun (2014).

Linus Tang

Extremal Bounds on Peripherality Measures

Mentor: Dr. Jesse Geneson, SJSU

We investigate several measures of peripherality for vertices and edges in networks. We improve asymptotic bounds on the maximum value achieved by edge peripherality over connected n-vertex-graphs, edge sum peripherality over connected n-vertex graphs, edge sum peripherality over n-vertex graphs with diameter at most 2, edge sum peripherality over bipartite n-vertex graphs with diameter at most 3, and the Trinajstić index over n-vertex graphs. We compute the maximum value achieved by the peripherality index over connected n-vertex graphs and n-vertex trees for all $1 \leq n \leq 8$. We refute two conjectures of Furtula, the first on necessary conditions for minimizing the Trinajstić index and the second about maximizing the Trinajstić index. Finally, we find an asymptotic expression for the expected value of the irregularity of the random graph $G_{n,p}$ for arbitrary $0 < p < 1$.

Eric Zhan

On the Winning and Losing Parameters of Schmidt Games

Mentor: Vasiliy Nekrasov

In this talk, we discuss the Schmidt Game, a game played in metric spaces developed in the 1960s with applications in number theory, analysis, and dynamical systems. With a few trivial exceptions, the conditions on parameters under which one player wins or loses in general are largely unknown. We describe some new nontrivial winning and losing conditions for two parameterized families of sets: one constructed similarly to the set of badly approximable numbers and one constructed by bounding the frequencies of zeroes in base-2 expansions of real numbers.

Raymond Luo

Cyclic Base Orderings and Equitability of Matroids

Mentor: Yuchong Pan

Matroids are combinatorial structures that generalize both the notion of spanning trees from graph theory and linear independence from linear algebra. Let $M = (E, \mathcal{I})$ be a matroid, and let $r(E)$ be the rank of M. Kajitani, Ueno, and Miyano provided a characterization of all matroids M that exhibited a linear ordering of their elements such that every consecutive $r(E)$ elements in the ordering is a base. However, if we replace "linear ordering" with "cyclic ordering", then the problem becomes significantly harder and remains unsolved to this day. We refer to such a cyclic ordering as a cyclic base ordering. Kajitani et al. showed that if the ground set of a graphic matroid can be partitioned into two bases, then there exists a cyclic base ordering. We give an extension of this theorem and show that there must exist a cyclic base ordering with a certain structure. Moreover, we also introduce the notion of "equitable" matroids, and show how they are related to cyclic base orderings.

William Gvozdjak

On the Classification of Low-Rank Odd-Dimensional Modular Categories

Mentors: Prof. Julia Plavnik, Indiana University Bloomington, and Agustina Czenky, University of Oregon

Modular categories are important algebraic structures that appear in a diverse quantity of applications, including topological quantum computing, which makes it interesting to classify them. We prove that any odd-dimensional modular category of rank at most 23 is pointed. We also show that an odd-dimensional modular category of rank 25 is either pointed, perfect, or equivalent to $\mathrm{Rep}(D^{\omega}(\mathbb{Z}_7 \rtimes \mathbb{Z}_3))$. Finally, we give partial classification results for modular categories of rank up to 73.

Akshaya Chakravarthy

On Modular Categories with Frobenius-Perron Dimension Congruent to 2 Modulo 4

Mentors: Prof. Julia Plavnik, Indiana University Bloomington, and Agustina Czenky, University of Oregon

In this talk, we consider the classification of modular categories $\mathcal{C}$ with Frobenius-Perron dimension, FPdim, congruent to 2 modulo 4. We show that such categories have their group of invertibles of even order and that they factorize as $\mathcal{C} \cong \widetilde{\mathcal{C}} \boxtimes \text{semion}$, where $\widetilde{\mathcal{C}}$ is an odd-dimensional modular category and semion is the rank 2 pointed modular category. This reduces the classification of these categories to those of odd-dimensional modular categories. It follows that modular categories $\mathcal{C}$ with $\mathrm{FPdim}(\mathcal{C}) \equiv 2 \pmod 4$ and rank up to 46 are pointed. More generally, we prove that if $\mathcal{C}$ is a weakly-integral modular category and p is an odd prime dividing the order of the group of invertibles that has multiplicity one in $\mathrm{FPdim}(\mathcal{C})$, then we have the factorization $\mathcal{C} \cong \tilde{\mathcal{C}} \boxtimes \mathrm{Vec}_{\mathbb{Z}_p}^{\chi}$, for $\tilde{\mathcal{C}}$ a modular category with Frobenius-Perron dimension not divisible by p and χ a non-degenerate quadratic form on $\mathbb{Z}_p$. Lastly, we show the existence of pointed modular subcategories in pseudo-unitary modular categories.

Joseph Vulakh

Simple Racks over the Alternating Groups

Mentors: Prof. Julia Plavnik and Dr. Héctor Peña Pollastri, Indiana University Bloomington

Racks are fundamental algebraic structures with significance in many areas of mathematics. In this talk, we discuss conjugation in groups and define racks as a natural generalization of the group conjugation operation. We present an application of racks to a long-standing algebraic problem—the classification of finite-dimensional pointed Hopf algebras. In particular, we study the type D property, which allows certain racks to be ruled out from being sources of finite-dimensional pointed Hopf algebras, for simple racks over the alternating groups.

SESSION 7: ALGEBRA

Ethan Liu

On the Structure and Generators of the n-th Order Chromatic Algebra

Mentor: Merrick Cai

The chromatic algebra is an important object in the study of graph theory and relates graphs to topics such as the Potts model and topological quantum field theory. We first provide proofs for many fundamental facts about chromatic algebras that appear to be missing from the literature. These include establishing a bijection between the n-th order chromatic algebra basis diagrams and the noncrossing planar partitions without singletons of $2n$ points, determining a canonical representation of the chromatic algebra as strings of length $2n$, and showing that the dimension of the n-th order chromatic algebra is the $2n$-th Riordan number, which exhibits exponential growth. Beyond studying the basis, we investigate the question of the size of generating sets of the chromatic algebra. We establish a generating set of size $\binom{n}{2} + 1$, and we provide a procedure to construct the basis from the generating set. Additionally, we provide a new method for computing the trace of elements in the chromatic algebra.

Justin Zhang

An Extension of Benson's Conjecture to Finite 3-Groups for Monomial Modules with Null Inner Partition

Mentor: Dr. Kent Vashaw

For finite 2-groups, Dave Benson conjectured that over an algebraically closed field $\Bbbk$ of characteristic 2, for any indecomposable module V of odd dimension, $V \otimes V^*$ has a unique odd dimensional summand $\Bbbk$, and thus any tensor power of V has a unique summand of odd dimension. It is known that an extension of Benson's conjecture to finite 3-groups fails: namely, there exist indecomposable modules W with dimension coprime to 3 for which $W \otimes W^*$ does not have a unique summand with dimension coprime to 3, and thus, it does not hold that any tensor power of W has a unique summand with dimension coprime to 3. However, it is difficult to characterize all such modules. In this talk, we propose and investigate a characterization for all such modules W when W is a monomial module, a type of graded representation over the group $\mathbb{Z}_{3^a} \times \mathbb{Z}_{3^b}$ for positive integers a and b which corresponds to a skew Young diagram, with null inner partition. We show that the family of monomial modules corresponding to partitions $(a_1, a_2, \ldots, a_n)$ with $a_i \equiv 0, 5 \pmod 9$ for $1 \leq i \leq n$ does not satisfy the extension of Benson's conjecture to finite 3-groups. We prove that for all such modules, there exists an isomorphic summand of dimension 5 in their decompositions. We then use the syzygy functor to show that the extension also fails for monomial modules corresponding to $(a_1, a_2, \ldots, a_n)$ with $a_i \equiv 0, 4 \pmod 9$ for $1 \leq i \leq n$. We propose that these two cases are the only monomial representations with null inner partition which fail the proposed extension of Benson's conjecture, which we support with abundant computational evidence.

Andrew Lin, Henrick Rabinowitz, and Qiao Zhang

The Furstenberg Property in Puiseux Monoids

Mentor: Dr. Felix Gotti

Let M be a commutative monoid. The monoid M is called atomic if every non-invertible element of M factors into atoms (i.e., irreducible elements), while M is called a Furstenberg monoid if every non-invertible element of M is divisible by an atom. Additive submonoids consisting of nonnegative rationals are called Puiseux monoids, and their atomic structure has been actively studied during the past few years. The primary purpose of this talk is to investigate the property of being Furstenberg

in the context of Puiseux monoids. In this direction, we consider some properties weaker than being Furstenberg, and then we connect these properties with some atomic results which have been already established for Puiseux monoids.

Hannah Fox, Agastya Goel, and Sophia Liao

Arithmetic of Semisubtractive Semidomains

Mentor: Prof. Harold Polo, University of Florida

A subset S of an integral domain is called a semidomain if the pairs $(S, +)$ and $(S \setminus \{0\}, \cdot)$ are commutative and cancellative semigroups with identities. The multiplication of S extends to the group of differences $\mathcal{G}(S)$, turning $\mathcal{G}(S)$ into an integral domain. In this talk, we study the arithmetic of semisubtractive semidomains (i.e., semidomains S for which either $s \in S$ or $-s \in S$ for every $s \in \mathcal{G}(S)$). Specifically, we provide necessary and sufficient conditions for a semisubtractive semidomain to satisfy atomicity, the ascending chain condition on principals ideals, to be a bounded factorization semidomain, and to be a finite factorization semidomain, which are subsequent relaxations of the property of having unique factorizations. In addition, we present a characterization of half-factorial and factorial semisubtractive semidomains. Throughout this talk, we present examples to provide insight into the arithmetic aspects of semisubtractive semidomains.

Session 8: Geometry and Other Topics

Anton Levonian

Existence of Circle Packings on Certain Translation Surfaces

Mentor: Prof. Sergiy Merenkov, CCNY – CUNY

A translation surface is a surface formed by identifying edges of a collection of polygons in the complex plane that are parallel and of equal length using only translations. Every translation surface is in a stratum determined by the order of its singular points. A circle packing is a collection of interiorwise disjoint discs on a translation surface which can be represented by a contacts graph. In this talk, we will discuss the realizability of the same circle packing on varying translation surfaces in the $\mathcal{H}(2)$ stratum. We will also discuss the possible complexity of contacts graphs in $\mathcal{H}(2)$ and $\mathcal{H}(1,1)$, providing a bound on this complexity in the $\mathcal{H}(2)$ stratum. Finally, we establish the possibility of certain contacts graphs' complexities in the generalized genus g strata $\mathcal{H}(g-1, g-1)$ and $\mathcal{H}(2g-2)$.

Nicholas Hagedorn

Algorithmically Generated Pants Decompositions of Combinatorial Surfaces

Mentor: Elia Portnoy

Buser proved that the Bers' constant of all Riemannian surfaces S with genus $g \geq 2$ is less than $C(g\text{Area}(S))^{1/2}$ for some constant C. The proof uses an algorithm that constructs pants decompositions of S. The algorithm is understood theoretically, but its behavior is unknown. We introduce combinatorial surfaces, a method of storing Riemannian 2-manifolds in discrete data structures. We then provide a discrete algorithm that finds the length of a pants decomposition of combinatorial surfaces. We find the algorithm's time complexity and prove it gives pants decompositions with length less than $C(g\text{Area}(S))^{1/2}$ for some constant C. We then show that our algorithm rarely does better than Buser's upper bound.

Alexander Li

Canonical Forms for Toric and Surface Codes in ZX Calculus

Mentor: Andrey Boris Khesin

Quantum computation brings with it the possibility of efficiently solving problems in various fields of study, in ways that modern computers and supercomputers cannot. Quantum computers have inherently noisy processes, and information transferred through quantum processes must go through error-correcting codes to avoid being overly distorted by the noise. These quantum error-correcting codes can be represented in graphical languages, such as ZX calculus, which displays quantum systems and processes using nodes and connections. Certain codes, such as the Toric and surface codes, have wide applicability to the creation of large-scale quantum computers due to the nature of their error detection methods. We aim to derive the canonical forms of these Toric and surface codes in the ZX calculus. To do this, we manipulate the codes using ZX rewrite rules, which simplify and condense the diagrams, and the diagrammatic software Quantomatic. The resulting canonical forms of the Toric and surface codes minimize the number of nodes used and show clear symmetries. This work builds on previous works in converting stabilizer tableaus into a form in ZX calculus that is intuitive and explainable, and this talk provides critical work in deriving improved canonical forms for error-correcting codes.

Iz Chen and Krishna Pothapragada

Classification of Non-degenerate Symmetric Bilinear Forms in the Verlinde Category Ver4+

Mentor: Arun Kannan

Symmetric tensor categories (STCs) provide a framework for new types of linear algebra and Lie theory. In characteristic 0, all STCs with moderate growth can be understood as representations of an affine supergroup scheme. However, this is not always the case in characteristic p. The most fundamental counterexamples are the Verlinde categories, which play a key role in understanding STCs in prime characteristics. Of these, the simplest in characteristic $p = 2$ is the Verlinde category Ver_4^+, which can be realized as $\mathrm{Rep}\,\mathbb{K}[t]/(t^2)$ with a modified braiding. In this talk, we classify the isomorphism classes of non-degenerate symmetric bilinear forms on objects in Ver_4^+.

Session 9: Group Theory and Representations

Matvey Borodin

The Action of the Cactus Group on Arc Diagrams

Mentor: Prof. Leonid Rybnikov

We study the problem of classifying orbits for some particularly interesting combinatorial group action, namely, the action of the cactus group on the set of arc diagrams. We will start with definitions of all of the objects involved: groups, group actions, the cactus group, and arc diagrams. Then, we will briefly discuss some results regarding the classification of the orbits of this group action and additional relations imposed on the group by this action. The motivation for this study comes from Combinatorial Representation Theory. Namely, the action of the cactus group on arc diagrams is a pictorial representation of the action of the cactus group on tensor products of crystals of irreducible $U_q(\mathfrak{sl}_2)$-modules (though this connection will not be discussed in the talk).

Alan Bu

The Local-Global Principle and a Projective Twist on the Hasse Norm Theorem

Mentor: Dr. Thomas Rüd

The Local-Global principle has been the focus of numerous theorems and study since its conception by Hasse with his Hasse-Minkowski theorem. It has wide connections to elliptic curves and the analysis of diophantine equations. We present a historical introduction to the major breakthroughs leading up to the Hasse Norm Theorem, and investigate whether the local-global principle holds on projective ratios of multiple norms over multiple number fields.

Brian Li

Tensor Product Decompositions for Modules over Subregular W-Algebras

Mentor: Dr. Artem Kalmykov

Decomposition of the tensor product between two finite-dimensional representations into simple ones is a classical problem in representation theory. In this talk, we recall the well-studied case of $\mathfrak{sl}_2$ representations and investigate their tensor product decomposition through highest-weight vectors. We generalize it to the case of the tensor product of a Whittaker module and a finite-dimensional representation of $\mathfrak{gl}_N$ by explicitly computing its Whittaker vectors. In particular, we show how it provides a non-standard quantization of the group $GL(N)$.

Razzi Masroor

Hyperoctahedral Schur Algebra and the Hyperoctahedral Web Algebra

Mentor: Dr. Elijah Bodish

In this talk, we will discuss how diagrammatic depictions of algebraic objects can motivate their generators and relations presentations. We start by discussing string diagrams for the symmetric group and how one could guess the Coxeter presentation from this perspective. We continue this story with the Schur Algebra in place of the symmetric group and web diagrams in place of string diagrams. Finally, we discuss our work on applying these concepts to the hyperoctahedral group.

Session 10: Computer Science

Dongchen Zou

Intersection Attack in Non-Uniform Setting

Mentor: Simon Langowski

Recently consumer demand for privacy has spurred growth in private messaging systems. However, formally, privacy degrades in such systems when users log on and off: this change of status exposes the ongoing conversations. Intersection attacks (also known as statistical disclosure attacks) use messaging patterns or liveness information to reconstruct relationships, deanonymize users, and track user behaviors. Prior attacks assume users have an underlying uniform communication pattern for simplicity, leaving the question open of how effective such attacks would be in a non-uniform real world. We observe that effects like clustering in real social graphs, and correlation between repeated conversations change the behavior and potential of such attacks. This talk provides a new approach that can take into account some of these additional factors, by constructing a polynomial to determine the social graph. We provide an analysis of the performance, accuracy, and convergence rate of our attack, and compare our attack with prior work. Our attack applies to many existing anonymous communication systems, and our technique can be extended to incorporate additional factors.

Alan Song and Evan Ning

Exploring Data-driven Resource Management for Serverless Systems

Mentor: Varun Gohil, Nikita Lazarev, and Yueying (Lisa) Li

Serverless computing is a paradigm of cloud computing that allows users to avoid challenging server management and overprovisioning of resources. In the serverless model, users submit functions to the cloud provider, who deploy and execute instances of them in short-lived containers before returning the output to the user. The cloud provider is thus responsible for managing computing resources such that (1) user-provider agreements on quality of service (QoS) objectives are met and (2) resources (i.e. containers) are neither over- nor under-provisioned. Current serverless systems in production address resource management with naive autoscalers that provide heuristic solutions at best. We propose a Deep Q-Learning model that learns to make optimal resource allocation decisions through iteratively interacting with the dynamic serverless environment. By collecting experience across many iterations, the model is able to make informed scaling decisions given QoS objectives and the environment state that satisfy QoS constraints. We implement the system using the Kubernetes cloud platform and evaluate its performance with vSwarm serverless benchmarks.

Rohith Raghavan and Eric Chen

Comparing Methods of Opportunistic Risk-Limiting Audits

Mentor: Mayuri Sridhar

Auditing elections is an important part of preserving faith in the electoral system and verifying the accuracy of the reported results of an election. Conventional election audits involve taking a set number or percentage of ballots and checking if the samples match the reported winner. However, these methods are unreliable for close races and excessive for races with a wide margin. Risk-limiting audits use statistical tests in order to assign a certain risk limit, the maximum probability that the results are incorrect, by sampling ballots one at a time until the risk limit is achieved. Our research focuses on opportunistic auditing, the ability to audit multiple races simultaneously, and attempts to determine what strategies are most effective for opportunistic auditing.

Session 11: Computer Science

Boyan Litchev

Parallelizable and Updatable Private Information Retrieval

Mentor: Simon Langowski

Private Information Retrieval (PIR) — a cryptographic protocol that lets users retrieve items from a database without revealing which item was retrieved — is an important building block of recent systems in anonymous messaging and private streaming. Currently, PIR utilizes fully homomorphic encryption (FHE) and homomorphic multiplications in order to achieve this. To speed up the multiplications, number-theoretic transforms (NTTs) are utilized. After multiplication, the ciphertext noise increases multiplicatively, meaning that few multiplications can be applied successively. To reduce this noise, schemes apply modulus and key-switching after multiplication. However, these optimizations cannot be applied to the NTT forms of ciphertexts, so ciphertexts have to be converted out of NTT form, using a significant amount of processing time and preventing parallelization. In the setting of PIR, small ciphertext values, low multiplicative depth, and the usage of fresh ciphertexts in multiplications mitigate noise even without key and modulus-switching. We explore the efficiency of removing key and modulus-switching from the computation process for PIR, eliminating the need for intermediate number-theoretic transforms. This also aids in updating the result of a query when the database is modified.

Andrew Carratu and Albert Lu

Public-key Signature Scheme with Reduced Hardware Trust

Mentors: Sacha Servan-Schreiber and Jules Drean

This project aims to define a new "memory-less" paradigm to create cryptographic primitives that are resistant to transient execution attacks and microarchitectural side channels. The main idea is to constrain secrets and intermediary results used during sensitive computations within the perimeter of a single core and specifically in its registers. This makes it impossible for an attacker to access these secrets through transient execution attacks as they never reside in memory, are hence not addressable and not accessible beyond the window of execution of the cryptographic function. It would also protect our secure function against side channels exploiting the memory hierarchy including any type of cache or DRAM side channels. For this project, we specifically aim at developing digital signature scheme that can be implemented following this new paradigm. Digital signatures are a fundamental primitive used in many applications where transient attacks and microarchitectural side channels can be extremely detrimental such as remote attestation or trusted execution environments.

Yifan Kang

NUMA-Aware Data Structure Design and Benchmarking

Mentors: Shangdi Yu and Prof. Julian Shun

High-performance servers are non-uniform memory access (NUMA) machines. To fully leverage these machines, programmers need efficient concurrent data structures that are aware of the NUMA performance artifacts. This talk explores the design and implementation of NUMA (Non-Uniform Memory Access)-aware data structures, such as kd-trees, in computer systems with NUMA architectures.

Omar El Nesr

Fast GPU Accelerated Ising Models for Practical Combinatorial Optimization

Mentor: Axel Feldmann

Combinatorial optimization (CO) problems exist in every scientific field, spanning disciplines such as biology, chemistry, finance, and mathematics. These problems are NP-hard, and difficulty explodes with size. Methods for quickly finding high-quality solutions to many problems enable breakthroughs in research. The Ising model is a flexible and general system from statistical mechanics that optimization problems can be mapped onto. Solving the Ising model then becomes equivalent to solving the original problem. However, current solvers do not scale well to large problem sizes. In this talk, we take advantage of the sparsity of large graphs and present an optimized GPU Ising solver for both MAXCUT and the Traveling Salesman Problem (TSP). Benchmarking shows the solver outperforms state-of-the-art tools in both solution quality and runtime, achieving a 3x geometric mean speedup and a maximum speed up of over 2,000x over the prior leading implementation. Our method scales well to problems of over 40,000 binary variables and runs thousands of simulations simultaneously, giving near-optimal solutions in seconds. To the best of our knowledge, this is the fastest open-source tool for Ising optimization and will enable researchers to solve these computational problems with ease.

Sarah Pan

Let's Reinforce Step by Step

Mentors: Prof. Anna Rumshisky and Vlad Lialin, UMass Lowell

While recent advances have boosted language model (LM) proficiency in linguistic benchmarks, LMs consistently struggle to reason correctly on complex tasks like mathematics. We turn to Reinforcement Learning from Human Feedback (RLHF) as a method with which to shape model reasoning processes. In particular, we explore two reward schemes, outcome-supervised reward models (ORMs) and process-supervised reward models (PRMs), to optimize for logical reasoning. Our results show that the fine-grained reward provided by PRM-based methods enhances accuracy on simple mathematical reasoning (GSM8K) while, unexpectedly, reducing performance in complex tasks (MATH). ORM-based approaches, on the other hand, increase accuracy in MATH but do not affect GSM8K performance. Furthermore, we show the critical role reward aggregation functions play in model performance. Providing promising avenues for future research, our study underscores the need for further exploration into fine-grained reward modeling for more reliable language models.

Eddie Wei

The Algebraic Value-Editing Conjecture in Deep Reinforcement Learning

Mentor: Andrew Gritsevskiy, Cavendish Labs

Reinforcement learning has been around for over half a century and much research has been done to make agents learn through trial and error. In the past decade, people have combined deep learning with reinforcement learning to create deep reinforcement learning, allowing agents to learn much more complex tasks. However, despite our understanding of the overall structure of these models we train on, due to the complexity of the internal calculations, we still lack comprehension of the internal mechanisms through which agents truly learn. We conjecture that there is a way to edit the values of the activations of our models to alter what our agent does. We do this by using Stable Baseline's benchmark models for the Minigrid environment from Farama Foundation. By analyzing the loss values after testing, we can see if our results support this theory.

Henry Han and Adrita Samanta

Visualizing Distributed Traces in Aggregate

Mentors: Darby Huye, Max Liu, Roy Zhang, and Prof. Raja Sambasivan, Tufts University

Understanding system behavior can be difficult to do when using distributed tracing. Even with sampling and performance profilers, developers still encounter thousands of various traces. Even when there are patterns in the system, it is often difficult to detect similarities since current tools only allow developers to view individual traces. Debugging and optimization becomes harder for developers without an understanding of the whole trace dataset. In order to help present these similarities, this talk proposes a method to aggregate traces in a way that groups together and visualizes similar traces. We do so by assigning a few traces that are representative of each set. We suggest that traces can be grouped based on how many services they share, how many levels the graph has, how structurally similar they are, or how close their latencies are. We also develop an aggregate trace data structure as a way to comprehensively visualize these groups and a tool to predict missing elements of traces, enabling developers to look through different groups of completed traces. The unique traces of each group are especially useful to developers for troubleshooting. Overall, our approach allows for a more efficient method of analyzing system behavior.

Garima Rastogi and Sophia Lichterfeld

How Do I Pay Thee? Let Me Count the Ways: Leveraging Ethereum Smart Contracts to Facilitate Web Monetization Adoption

Mentor: Kyle Hogan

Web Monetization (WM) aims to provide users with an alternative method—besides ads and subscription models—of compensating creators. Traditional WM schemes have faced significant challenges in acquiring widespread adoption because they require full website participation to be implemented. However, website owners are often unfamiliar with cryptocurrencies and online wallets, making it difficult for them to set up WM quickly. Our scheme addresses this barrier by providing users with the option to initiate WM on a web page even before the owner has had the chance to establish their end of the system. Users employ WM stream micropayments into a Solidity smart contract on the Ethereum blockchain where it will be stored securely in escrow. Owners wanting to retrieve this revenue must adopt W3C's WM API payment pointer standard for future use; thus, our approach ultimately aims to encourage the propagation of WM as a viable option for supporting content creators online.

Session 13: Computational and Physical Biology

Valentina Zhang

Identifying Microglial Heterogeneity in Alzheimer's Disease

Mentor: Dr. Ayshwarya Subramanian, Broad Institute

Microglia are predominant resident immune cells of the central nervous system (CNS) and play a critical role in the pathological process of Alzheimer's Disease (AD). Changes in microglia have been observed in pathologically relevant brain regions of both AD mouse models and patients. In recent years, several studies that characterized the features of microglia have indicated that microglia are heterogeneous. However, the functions of microglia and the molecular changes underlying the responses of microglia in the AD brain are not very clear. We hypothesize that unraveling the heterogeneity of microglial cells in AD could help elucidate AD's pathogenic mechanisms and neuropathological phenotypes. Using single-cell RNA-sequencing (scRNA-seq), we integrated and analyzed brain samples of 34 AD patients and 29 healthy controls sourced from four public datasets. We identified 13 distinct microglial clusters. We use two existing studies as reference for comparative analysis, and determine new microglial subsets not represented in current literature. In the next step, we will associate those clusters with different aspects of AD progression. Upon the completion of the analysis, our findings may suggest the potential roles of microglial cell states in analyzing AD's pathology.

Raj Saha

Surveying the Presence and Diversity of Viruses in Mammalian Transcriptomes

Mentor: Dr. Ayshwarya Subramanian, Broad Institute

Zoonotic spillover of pathogens from animals to humans is a serious threat to public health and society, as evidenced by the COVID-19 pandemic. Uncovering the intermediate hosts that transmit zoonoses between different species is a vital step in mitigating the spread of viral pathogens. For instance, studies suggest that pangolins have passed SARS-CoV-2 from bats to humans, acting as an intermediate host. We utilize public RNA sequencing datasets to identify the presence of viral RNA in a diverse array of species. To classify the viruses present in the RNA-seq samples, we use a metagenomic classifier that provides an abundance profile for all detected species in the sample. The amount of identified viral RNA is quantified by the number of sequencing reads in a matrix, with the sample species as rows and viruses as columns. Upon normalizing these raw read counts, we use visualizations

to discover viruses most prevalent throughout observations; identify differing viral read abundance patterns; and analyze viral presence across various mammalian taxonomic levels. Our approach can help provide insight into conjectures regarding species susceptibility to zoonotic viruses.

Amith Saligrama

A Novel Statistical Framework for Identification of Mutated Cells

Mentors: Dr. Giulio Genovese, Broad Institute, and Prof. Steve McCarroll, Harvard Medical School

We propose a novel statistical model for single-cell sequencing data to identify mutated cells in samples with known autosomal loss and alterations. Identifying cells with autosomal loss or alterations is fundamentally challenging because: (a) autosomal chromosomes are pairs, and loss or alteration in one of the pairs does not nullify its expression; (b) expression data is invariably noisy, and the noise arises from both technical and biological variations. Consequently, while we expect a reduction in counts, a cell's condition is not directly inferable. Our key insight for detecting chromosomal loss in a cell is based on the idea of normalizing against another chromosome, whose expression is known to be statistically independent of target chromosomal loss/mutation. This leads us to a precise characterization in terms of binomial distributions, and we can perform a hypothesis test for each cell and detect ploidy. We extend this framework for detection of allelic alterations. We then develop a classification algorithm that detects chromosomal loss under control on false positivity rate (FPR). We validate our model by utilizing counts of single RNA molecules from haplotypes affected in a fraction of the cells analyses, and then use the algorithm to identify cells that have lost chromosome 18 in brain cells or carry a 9q CN-LOH alteration in chromosome 9q in induced pluripotent stem cells derived from peripheral blood mononuclear cells. Cell-by-cell identification of chromosomal loss is a critical step for inferring gene expressivity, and surprisingly we identify a consistent pattern of abnormal trans-chromosomal expression in cells for known autosomal loss/alterations. Our study also leads to a surprising finding: prior studies associate 9q CN-LOH with diverse detrimental effects, and in contrast our study reveals that the mutated cells behave no differently from non-mutated cells.

Anna Du

Utilizing Machine Learning to Identify Time Asymmetry of DNA Loop Extrusion

Mentors: Dr. Aleksandra Galitsyna and Henrik Pinholt

DNA loop extrusion, facilitated by the cohesin protein complex, plays a pivotal role in organizing the genome, influencing processes like gene regulation, somatic recombination, and DNA repair. The loop extruder cohesin and its interaction with the boundary protein CTCF has been found to lead to the establishment and regulation of compacted regions of DNA known as topologically associated domains (TADs). The fundamental goal behind this research is to study a novel method for detection of loop extrusion in living systems. This uses the fact that "movies" of DNA under action of loop extrusion would look different when played forward in time compared to backwards. Microscopy data of DNA motion could therefore be used to detect the presence of loop extrusion based on our ability to discern reversed from regular trajectories. To explore the feasibility of this approach, it is necessary to perform simulations both with and without loop extrusion to generate a synthetic dataset. By conducting these simulations, we assess whether detectability is possible, and if possible, which experimental conditions would be needed to see loop extrusion in living systems. This study bridges the gap in our understanding of loop extrusion and offers promising avenues for future exploration in genome organization.

Elizaveta Rybnikova

Exploration of Hi-C Patterns Through Computer Simulations

Mentors: Dr. Aleksandra Galitsyna and Henrik Pinholt

DNA in the nucleus of eukaryotic cells is a compact structure that forms a variety of local structures — which appear as domains, fountains, and stripes in Hi-C maps — responsible for the regulation of genes and other crucial biological processes. In mammals, these structures are known to be formed by the mechanism of loop extrusion. In this process, a motor protein binds to DNA, connecting two regions of DNA together in a loop. This loop is processively extruded by the motor protein until it falls off. In this work, we build a quantitative model to characterize the features of 3D DNA that emerge through this mechanism. We compare our output against the experimental contact map of the mammalian genome (the readout of the Hi-C sequencing technique). Our polymer simulation model predicts the shape of fountains, domains, and stripes, which we further validate with a convolutional neural network. The resulting database provides a map of the variety of local 3D genome structures, providing an understanding of the parameters of loop extrusion, helping to reveal its biological significance.

Session 14: Bioinformatics

Irene Jiang

Machine Learning Inference of Causal Genes for Tuberculosis using Mendelian Randomization and Single-cell Sequencing

Mentor: Prof. Gil Alterovitz, Harvard Medical School

The treatment of Mycobacterium Tuberculosis is becoming increasingly challenging due to limited access to essential medication and facilities, the rise of antibiotic resistance, and co-occurrence with other diseases such as HIV and cancer. As the need for personalized and cost-effective TB intervention strategies intensifies, advancements in biotechnology and *in silico* techniques are proving pivotal in understanding disease mechanisms and drug discovery. Here, we combine Mendelian Randomization (MR) to screen for causal genes, with single-cell RNA-sequencing (RNA-seq) to understand how these gene modules change during infection in specific cell types. Our MR analysis provided evidence of causal links between the *ZNF575* and *ZNF577* gene module, the *KLK11-KLK10* pathway, and the iron-regulation pathways by *USP24* or *GLRX5*. We then used single-cell RNA sequencing data of a murine model of TB infection to investigate the response of these genes to TB infection. Taken together, our work prioritizes these potentially causal pathways for TB infection and personalized susceptibility. Future *in vivo* studies are required to investigate the functional impact of these genes.

Gavin Ye

Drug Design as Causal Language Modeling: Transferring Large Chemistry Models for De-Novo Drug Design with Supervised and Reinforcement Learning

Mentor: Prof. Gil Alterovitz, Harvard Medical School

Recent years have shown various successes in applying generative machine learning models to the design of novel drug molecules. Specifically, most successful drug design models model drug molecules using a sequence-like language called SMILES. Due to the sequential nature of the data, models such as recurrent neural networks and transformers are candidates to generate drug molecules with optimized properties such as drug efficacy. However, with recent advances in Large Language Models, their implications for drug design have largely been unexplored. Specifically, one study has successfully pretrained a Large Chemistry Model, but its application to specific tasks in drug discovery is unexplored.

In this study, the drug design task is modeled as an NLP causal language modeling problem. Thus, a procedure of supervised finetuning, reward modeling, and proximal policy optimization is used to transfer the Large Chemistry Model to drug design, similar to Open AI's ChatGPT and InstructGPT

procedure. By combining the SMILES sequence feature with chemical descriptors, the reward model exceeded performance compared to previous studies. After proximal policy optimization, the drug design model generated molecules with 99.2% having efficacy pIC50 > 7 and took less than 14 iterations of data to learn, with 100% of the generated molecules being valid. This demonstrated the applicability of the Large Chemistry Model in drug discovery, with benefits including less data consumption when fine-tuning. This opens the door for larger studies involving reinforcement-learning with human feedback, where chemists provide feedback to chemistry models, making them learn the chemists' intuitions and generate higher-quality molecules.

Aaron Li

Identification of Biomarkers for Insulin Resistance using Meta-Analysis and Machine Learning

Mentor: Prof. Gil Alterovitz, Harvard Medical School

Insulin resistance (IR) is a metabolic disorder associated with reduced cellular responsiveness to insulin, leading to impaired glucose uptake and metabolism. Identifying biomarkers for IR is crucial for improving treatment strategies and understanding the underlying mechanisms. This study aimed to identify biomarkers for IR through a combination of two methods. Firstly, the ImaGEO webtool was used to identify genes that were significantly differentially expressed in microarray samples across 3 GEO datasets from IR and insulin sensitive patients through statistical analysis. ImaGEO identified 576 DEGs at p=0.005, with top DEGs including DAP3, SYS1, ZNF451, PDE6D, and AH1. Enrichment analysis revealed that these genes were members of organ development, cell fate commitment and system development pathways. To augment the data produced by DEG analysis, three machine learning (ML) models of different architectures were trained to classify microarray samples from IR and insulin sensitive patients. The highest performing model, ANN, achieved an accuracy of 84%. Features were extracted from machine learning models to determine genes that the trained models had found to be significant, and were compared to the genes discovered through statistical analysis. A comparison revealed AH1, TGFB1, and RYK as top features in ML models that were also discovered to be differentially expressed by meta- analysis. These genes were found to be connected to the late-stage neurodegeneration that is often characteristic of diabetic patients. Analyses of gene expression levels through these two methods have identified potential biomarkers and inform further research on insulin resistance.

Stephanie Wan

Biomarker Identification of Oral Squamous Cell Carcinoma through Transcriptomic Expression Analysis

Mentor: Prof. Gil Alterovitz, Harvard Medical School

Due to the high mortality rate of oral squamous cell carcinoma (OSCC), early detection of the disease is critical. Despite previous research on potential diagnostic biomarkers, there is still no consensus regarding the role and validity of specific biomarkers for OSCC. The purpose of this study was to explore and verify potential diagnostic biomarkers for OSCC. mRNA expression data of 57 oral tissues from OSCC patients and 22 from individuals without OSCC was analyzed using a moderated t-test to determine potential biomarkers. This study suggests that MMPs are especially promising diagnostic biomarkers and therapeutic targets for OSCC. Additionally, 163 differentially expressed genes between OSCC and normal tissues found through statistical analysis using R, 68 of which were upregulated in OSCC tissue, were identified for further research.

Rianna Santra

Transcriptomic Analysis of the Dengue Virus using Feature Selection and Random Forest

Mentor: Prof. Gil Alterovitz, Harvard Medical School

Advancements in sequencing technologies have enabled the generation of large transcriptomic datasets, such as those publicly available on NCBI GEO. In this study, we use publicly accessible gene expression data from NCBI GEO to perform a comprehensive analysis of differential gene expression for the dengue virus. The primary objective is to identify genes exhibiting significant expression patterns associated with the virus, and, subsequently, to assess the importance of these genes by using a random forest classifier. These differentially expressed genes are subjected to further analysis to uncover their biological relevance, including enrichment analysis for potential molecular pathways and processes affected by these genes. These results contribute to a better understanding of the molecular mechanisms underlying diverse biological conditions and also highlight genes that could serve as potential biomarkers.