# 2022 PRIMES FALL-TERM CONFERENCE ABSTRACTS

## Saturday, October 15

### Session 1: Applied Mathematics

**Grace Wang**

*Correlations between COVID-19 and dengue*

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

A dramatic increase in the number of outbreaks of Dengue has recently been reported, and climate change is likely to extend the geographical spread of the disease. In this context, this talk shows how a neural network approach can incorporate Dengue and COVID-19 data, as well as external factors (such as social behavior or climate variables), to develop predictive models that could improve our knowledge and provide useful tools for health policymakers. Through the use of neural networks with different social and natural parameters, in this talk, we define a Correlation Model through which we show that the number of cases of COVID-19 and Dengue have very similar trends. We then illustrate the relevance of our model by extending it to a Long short-term memory model (LSTM) that incorporates both diseases, and using this to estimate Dengue infections via COVID-19 data in countries that lack sufficient Dengue data.

**Linda He**

*Search of primordial black holes in extrasolar systems*

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

Primordial black holes (PBHs) are known to be dark matter candidates, and their existence is constrained by a number of observations, such as gravitational waves and dynamical constraints. In this talk, we consider how different dynamical parameters, such as velocity, affect the perturbation strength between a single flyby PBH with a star-planetary system by using Monte Carlo simulations and semi-analytical estimates. Moreover, given a population of PBHs in the galaxy, we study the frequency of close encounters between PBHs with star systems.

**Ethan Zhou**

*Online learning of smooth functions*

**Mentor: Dr. Jesse Geneson, SJSU**

In this talk, we study the online learning of real-valued functions where the hidden function is known to have certain smoothness properties. Specifically, for $q \geq 1$, let $\mathcal{F}_q$ be the class of absolutely continuous functions $f : [0,1] \to \mathbb{R}$ such that $\|f'\|_q \leq 1$, and for $q \geq 1$ and $d \in \mathbb{Z}^+$, let $\mathcal{F}_{q,d}$ be the class of functions $f : [0,1]^d \to \mathbb{R}$ such that any function $g : [0,1] \to \mathbb{R}$ formed by fixing all but one parameter of $f$ is in $\mathcal{F}_q$. For any class of real-valued functions $\mathcal{F}$ and $p > 1$, let $\text{opt}_p(\mathcal{F})$ be the best upper bound on the sum of $p$th powers of absolute prediction errors a learner can guarantee. In the single-variable setup, we show for $\varepsilon \in (0,1)$ that $\text{opt}_2(\mathcal{F}_{1+\varepsilon}) = \Theta(\varepsilon^{-1})$. In addition, we obtain new exact results by proving for $\varepsilon \in (0,1)$ that $\text{opt}_p(\mathcal{F}_{1+\varepsilon}) = 1$ whenever $p \geq 2 + \varepsilon^{-1}$. In the multi-variable setup, we establish inequalities relating $\text{opt}_p(\mathcal{F}_{q,d})$ to $\text{opt}_p(\mathcal{F}_q)$ and show that $\text{opt}_p(\mathcal{F}_{\infty,d})$ is infinite when $p < d$ and finite when $p > d$.

### Andrew Tung and Karthik Vedula

*New properties of the intrinsic information and their relation to bound secrecy*

**Mentor: Andrey Khesin**

Two parties, Alice and Bob, seek to generate a mutually agreed upon string of bits, unknown to an eavesdropper Eve, using realizations of a joint probability distribution. The secret-key rate has been defined as the rate at which Alice and Bob can extract secret bits over many realizations of the probability distribution. An information-theoretic quantity known as the reduced intrinsic information informally measures the amount of information one needs to give Eve to erase secrecy between Alice and Bob. It has previously been conjectured that the reduced intrinsic information is equal to the secret-key rate. In this talk, we introduce our result which implies that the reduced intrinsic information cannot be equal to the secret-key rate assuming bound secret distributions exist (which is widely believed to be the case).

### Benjamin Fan and Edward Qiao

*Deep learning for partial differential equations in economics*

**Mentor: Prof. Lu Lu, University of Pennsylvania**

Deep learning has been shown to be an effective method for solving partial differential equations (PDEs) by embedding the PDE residual into the neural network loss function. In this talk, we design a methodology that utilizes deep learning to simultaneously solve and estimate canonical continuous-time general equilibrium models in financial economics, including (1) industrial dynamics of firms and (2) macroeconomic models with financial frictions. Through these applications, we illustrate the advantages of our method: generality, simultaneous solution and estimation, leveraging the state-of-art machine-learning techniques, and handling large state space.

## Session 2: Combinatorics I

### Matvey Borodin, Ethan Liu, and Justin Zhang

*Vanishing polynomials and polynomial functions*

**Mentor: Prof. Jim Coykendall, Clemson University**

We study the set of objects known as vanishing polynomials (the set of polynomials which annihilate all elements of the ring) over general commutative rings with identity. We show that this set is an ideal of the ring of polynomials whose natural projection maps the ring of polynomials to the ring of polynomial functions. We then present a new approach to finding the generating set of this ideal over the ring $\mathbb{Z}_n$. Generalizing this approach, we partially classify the vanishing polynomials over any general commutative ring with identity. We also establish a bijection between vanishing polynomials and polynomial functions over product rings to those of their constituent rings.

### Jeffrey Chen

*Positivity properties for $q$-hit numbers*

**Mentors: Prof. Alejandro Morales and Jesse Selover, UMass Amherst**

We consider the problem of counting matrices over a finite field with fixed rank and support contained in a fixed set. The count of such matrices gives a $q$-analogue of the classical rook number, but it is known not to be polynomial in general. We use inclusion-exclusion on the support of the matrices and the orbit counting method of Lewis, Liu, Morales, Panova, Sam, and Zhang to show that the residues of these functions in low degrees are polynomial. We define a generalization of the rook and hit numbers over certain classes of graphs. This provides us a formula for residues of the $q$-rook and $q$-hit numbers in low degrees. We analyze the residues of the $q$-hit number and show that the coefficient of $q-1$ in the $q$-hit number is always non-negative.

### Rich Wang

*Ending states of a special variant of the chip-firing algorithm*

**Mentor: Dr. Tanya Khovanova**

We help a group of violinists who are struggling to find a peaceful room to practice in an infinite hotel. The manager of the hotel continually has to separate two violinists in adjacent rooms that cannot stand each other's loud noises by sending one of them to the closest unoccupied room to the left, and the other to the closest unoccupied room to the right, but finds that the pairs of feuding violinists never seem to stop. Will he eventually help each violinist find a room in which they can play without someone else's music blaring through the walls from an adjacent room? And if he can, what set of rooms might the violinists occupy after every violinist has found a room that they are happy in?

## Session 3: Combinatorics II

### Advay Goel

*The geometry and limits of Young partition flow polytopes*

**Mentor: Zoe Wellner, Carnegie Mellon University**

In 2017 Mészáros, Simpson, and Wellner demonstrated that certain flow polytopes resulting from Young tableaux have an easy decomposition into simplices and others have a natural relation to the well-known Tesler and CRY polytopes. They introduced the notion of limiting polytopes within the family of polytopes resulting from a specific Young tableau as the polytopes where the decomposition into simplices is easiest. In this talk, we further examine the limiting process within each family of polytopes by geometric decomposition at each step toward the limiting polytope. Our main results analyze the family of hooks, and we demonstrate an algorithm by which to get geometric decompositions.

### Derek Liu

*Arrangements of simplices in fine mixed subdivisions*

**Mentor: Yuan Yao**

A regular simplex of side length $n$ can be subdivided into multiple polytopes, each of which is a Minkowski sum of some faces of a unit simplex. Ardila and Billey have shown that exactly $n$ of these cells must be simplices, and their positions must be in a "spread-out" arrangement. In this talk, we consider their question of whether every spread-out arrangement of simplices can be extended into such a subdivision, especially in the 3-dimension case. We prove that a specific class of these arrangements, namely those that project down to a 2-dimensional spread-out arrangement, all extend to a subdivision.

### Anthony Wang

*Consecutive patterns in Coxeter groups*

**Mentor: Yibo Gao**

For an arbitrary Coxeter group element $w$ and a connected subset $J$ of the Coxeter diagram, the parabolic decomposition $w = w^J w_J$ defines $w_J$ as a consecutive pattern of $w$, generalizing the notion of consecutive patterns in permutations. We then define the cc-Wilf-equivalence classes as an extension of the c-Wilf-equivalence classes for permutations, and identify a nontrivial family of cc-Wilf-equivalent classes. Furthermore, we study the structure of the consecutive pattern poset in Coxeter groups and prove that its M "obius function is bounded when the arguments lie in finite Coxeter groups, but can be arbitrarily large otherwise.

### Nilay Mishra

*On the uniqueness of certain types of circle packings on translation surfaces*

**Mentor: Prof. Sergiy Merenkov, CCNY – CUNY**

Consider a collection of finitely many polygons in $\mathbb{C}$, such that for each side of each polygon, there exists another side of some polygon in the collection (possibly the same) that is parallel and of equal length. A translation surface is the surface formed by identifying these opposite sides with one another. The $\mathcal{H}(1,1)$ stratum consists of genus two translation surfaces with two singularities of order one. A circle packing corresponding to a graph $G$ is a configuration of disjoint circles such that each vertex of $G$ corresponds to a circle, two circles are externally tangent if and only if their vertices are connected by an edge in $G$, and $G$ is a triangulation of the surface. We prove that for certain circle packings on $\mathcal{H}(1,1)$ translation surfaces, there are only a finite number of ways the packing can vary without changing the contacts graph, if two circles along the slit are fixed in place. These variations can be explicitly characterized using a new concept known as *splitting bigons*. Finally, we generalize our uniqueness theorem to a specific type of translation surfaces with arbitrary genus $g \geq 2$.

## Session 4: Miscellaneous

### Max Misterka

*A generalization of q-calculus using formal group laws*

**Mentor: Sanath Devalapurkar, Harvard University**

In this talk, we will give a short introduction to $q$-calculus, and then discuss a generalization of the $q$-derivative. Recall that the derivative of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ is

$$Df(x) = \frac{df}{dx}(x) = \lim_{x' \to x} \frac{f(x') - f(x)}{x' - x}.$$

Suppose that we do not let $x'$ approach $x$, and instead we set $x' = qx$, where $q \neq 1$ is a fixed real number. Then we get the $q$-*derivative* of $f$:

$$\nabla_q f(x) = \frac{f(qx) - f(x)}{qx - x}.$$

The $q$-derivative has a power rule and a product rule. There are also $q$-analogs of binomial coefficients which satisfy analogs of many combinatorial identities. We will define a new type of derivative called the $s$-derivative on polynomials by using a modified version of the $q$-power rule. We will also state generalizations of several theorems in $q$-calculus to $s$-calculus. We found analogs of Pascal's identity, Vandermonde's identity, and Lucas's theorem, and also the Poincaré lemma and the Cartier isomorphism for the algebraic de Rham complex.

### Eric Chen and Alexander Zitzewitz

*Unitarity conditions of Heun and Lamé differential operators*

**Mentor: David Darrow**

In this talk, we explore the connections between the so-called "accessory parameter" of the Heun Equation and the properties of its monodromy groups. In particular, we investigate which numerical values of the accessory parameter yield unitary monodromy groups (i.e., those that preserve a Hermitian inner product). To this end, we employ both analytical and computational methods, extending the work of Beukers on the Lamé Equation. In particular, for a large class of Heun Equations (generalizing the Lamé Equation), we prove a connection between unitarity and the traces of certain monodromy matrices. We exploit this theorem to create an algorithm that finds accessory parameters that yield unitary monodromy groups, inspired again by Beukers' work on the Lamé Equation. Using this algorithm, we calculate and report the values of the accessory parameter that give rise to unitary monodromy groups. We also draw convergence maps, demonstrating the convergence and overall robustness of our algorithm.

### Eric Shen and Kevin Wu

*Congruences between logarithms of Heegner points*

**Mentor: Dr. Daniel Kriz, Institut de Mathématiques de Jussieu – Paris Rive Gauche**

Elliptic curves are an important class of Diophantine equations. We study certain special solutions of elliptic curves called Heegner points, which are the traces of images under modular parametrizations of complex multiplication points in the complex upper half-plane. We prove, for pairs of elliptic curves with isomorphic Galois representations, a general congruence of stabilized formal logarithms. This is done by first showing the isomorphism of Galois representations implies a congruence of stabilized modular forms and then translating these to the congruence of formal logarithms using Honda's theorem relating formal groups of elliptic curves to $L$-series and the modular parametrization. We use this congruence to show that examples of elliptic curves with analytic and algebraic rank 1 propagate in quadratic twist families.

## Session 5: Graph Theory

### Paul Gutkovich

*Computing truncated metric dimension on trees*

**Mentor: Zi Song Yeoh**

Let $G = (V, E)$ be a simple, finite, connected graph with vertex set $V$. Let $d(u, v)$ denote the distance between vertices $u, v$. A resolving set of $G$ is a subset $S$ of $V$ such that knowing the distance from a vertex $v$ to every vertex in $S$ uniquely identifies $v$. The metric dimension of $G$ is defined as the size of the smallest resolving set of $G$. We define the $k$-truncated resolving set and $k$-truncated metric dimension of a graph similarly, but with the notion of distance replaced with $d_k(u, v) := \min(d(u, v), k + 1)$.

In this talk, we demonstrate that computing the $k$-truncated metric dimension of trees is NP-Hard for general $k$. We then present a polynomial-time algorithm to compute the $k$-truncated metric dimension of trees when $k$ is a fixed constant.

### Max Xu

*Gonality sequences of multipartite graphs*

### Mentors: Amanda Burcroff, Harvard University, and Dr. Felix Gotti

In this presentation, we go over a particular sequence associated with a graph, the gonality sequence. The gonality sequence is part of a larger structure around the chip-firing game on a graph. The gonality sequence of a graph measures how much the degree of a divisor on that graph needs to change in order to increase its rank. In this presentation, we go over the methodologies of determining and proving the gonality sequences of certain classes of graphs. In particular, we go over Dhar's Burning Algorithm to find conjectures, in addition to bounding methods to prove gonality sequences. Finally, we take a look at some results of this year's efforts.

### Edward Yu

*A Turán-type problem in mixed graphs*

### Mentor: Nitya Mani

We investigate a natural Turán-type problem on mixed graphs, generalizations of graphs where edges can be either directed or undirected. We study a natural *Turán density coefficient* that measures how large a fraction of directed edges an $F$-free mixed graph can have; we establish an analog of the Erdős-Stone-Simonovits theorem and give a variational characterization of the Turán density coefficient of any mixed graph (along with an associated extremal $F$-free family).

This characterization enables us to highlight an important divergence between classical extremal numbers and the Turán density coefficient. We show that Turán density coefficients can be irrational, but are always algebraic; for every $k \in \mathbb{N}$, we construct a family of mixed graphs whose Turán density coefficient has algebraic degree $k$.

### David Dong, Alan Lee, and Michelle Wei

*Connectedness and cycle spaces of friends-and-strangers graphs*

### Mentor: Dr. Colin Defant

If $X = (V(X), E(X))$ and $Y = (V(Y), E(Y))$ are $n$-vertex graphs, then their *friends-and-strangers graph* $\mathsf{FS}(X, Y)$ is the graph whose vertices are the bijections from $V(X)$ to $V(Y)$ in which two bijections $\sigma$ and $\sigma'$ are adjacent if and only if there is an edge $\{a, b\} \in E(X)$ such that $\{\sigma(a), \sigma(b)\} \in E(Y)$ and $\sigma' = \sigma \circ (a\ b)$, where $(a\ b)$ is the permutation of $V(X)$ that swaps $a$ and $b$. We prove general theorems that provide necessary and/or sufficient conditions for $\mathsf{FS}(X, Y)$ to be connected. As a corollary, we obtain a complete characterization of the graphs $Y$ such that $\mathsf{FS}(\mathsf{Dand}_{k,n}, Y)$ is connected, where $\mathsf{Dand}_{k,n}$ is a dandelion graph; this substantially generalizes a theorem of Defant and Kravitz in the case $k = 3$. For specific choices of $Y$, we characterize the spider graphs $X$ such that $\mathsf{FS}(X, Y)$ is connected. In a different vein, we study the cycle spaces of friends-and-strangers graphs. Naatz proved that if $X$ is a path graph, then the cycle space of $\mathsf{FS}(X, Y)$ is spanned by 4-cycles and 6-cycles; we show that the same statement holds when $X$ is a cycle and $Y$ has domination number at least 3. When $X$ is a cycle and $Y$ has domination number at least 2, our proof sheds light on how walks in $\mathsf{FS}(X, Y)$ behave under certain *Coxeter moves.*

## Session 6: Algebra I

### Sophie Zhu

*Pointed fusion categories over non-algebraically closed fields*

**Mentors: Prof. Julia Plavnik and Sean Sanford, Indiana University Bloomington**

Pointed fusion categories over $\mathbb{C}$ are completely classified. We consider these categories over non-algebraically closed fields. We classify pointed braided fusion categories over arbitrary fields $F$ by imposing $K \otimes_F K$-module structures on the endomorphism algebras.

### Brendan Halstead

*Moduli spaces of tropical and logarithmic morphisms*

**Mentor: Jeffery Yu**

We first study morphisms between *cone stacks*, objects defined by Cavelieri, Chan, Ulirsch, and Wise as a framework for moduli problems in tropical geometry. We construct a cone stack $[\Sigma, \Gamma]$ parameterizing morphisms between fixed cone stacks $\Sigma$ and $\Gamma$. Wise showed that there is a logarithmic algebraic space $[X, Y]$ parameterizing logarithmic morphisms between fixed logarithmic schemes $X$ and $Y$. Using the equivalence of categories between cone stacks and Artin fans, we define a logarithmic algebraic stack $a^*[\Sigma(X), \Sigma(Y)]$ modeling combinatorially the mapping stack $[X, Y]$. We then construct a natural, strict tropicalization morphism from $[X, Y]$ to $a^*[\Sigma(X), \Sigma(Y)]$. This yields a stratification of the moduli space of morphisms of logarithmic schemes by combinatorial type, in the same way that tropicalization yields a stratification of the moduli space of curves $\overline{\mathcal{M}}_{g,n}$.

### Alan Bu, Joseph Vulakh, and Alex Zhao

*Length-factoriality and pure irreducibility*

**Mentor: Dr. Felix Gotti**

An atomic monoid $M$ is called length-factorial if, for every non-invertible element $x \in M$, no two distinct factorizations of $x$ into irreducibles have the same length (i.e., number of irreducible factors, counting repetitions). The notion of length-factoriality was introduced by J. Coykendall and W. Smith in 2011 under the term 'other-half-factoriality': they used length-factoriality to provide a characterization of unique factorization domains. In this talk, we study length-factoriality in the more general context of commutative, cancellative monoids. We give several results about the structure of factorizations in length-factorial monoids.

## Session 7: Algebra II

### Annie Wang

*The Hilbert series of the irreducible quotient for the polynomial representation of the rational Cherednik algebra of type A*

**Mentor: Serina Hu**

We study the polynomial representation of the rational Cherednik algebra of type A in characteristic $p = 3$ for $p|n-2$, $t = 0$, and generic parameter $c$. We describe all the polynomials in the maximal proper graded submodule $\ker \mathcal{B}$, which is the kernel of the contravariant form $\mathcal{B}$, and we use this to find the Hilbert series of the irreducible quotient for the polynomial representation. We proceed degree by degree to explicitly determine the Hilbert series and work towards proving Etingof and Rains's conjecture in the case that $p = 3$, $t = 0$, and $n = kp + 2$.

**George Cao**

*The indecomposable summands of the tensor products of monomial modules over finite 2-groups*

**Mentor: Dr. Kent Vashaw**

In this talk, we investigate a conjecture of Benson and Symonds regarding the indecomposable summands of tensor products of representations of finite 2-groups over an algebraically closed field of characteristic 2. They conjectured that the function outputting the dimension of the unique odd-dimensional indecomposable summand of $V^{\otimes n}$ for a given integer $n$ is polynomial or quasi-polynomial. Monomial modules are a special representation of the group $\mathbb{Z}/2^r \times \mathbb{Z}/2^s$, and they combinatorially correspond to skew Young diagrams. We prove the Benson-Symonds conjecture for two monomial representations whose tensor powers were previously unknown. We present computational evidence of the conjecture for a broad range of monomial modules.

**Caroline Liu, Annabel Ma, and Andrew Zhang**

*Factorization invariants of arithmetical congruence monoids*

**Mentor: Prof. Scott Chapman, Sam Houston State University**

In this talk, we discuss various factorization invariants of arithmetical congruence monoids. The invariants we investigate are the catenary degree, a measure of the maximum distance between any two factorizations of the same element, the length density, which describes the distribution of the factorization lengths of an element, and the omega primality, which measures how far an element is from being prime.

## SESSION 8: COMPUTER SCIENCE I

### Matan Yablon and Alicia Li

*How optimal can you get: Stochastic and adversarial reinforcement learning*

**Mentor: Mayuri Sridhar**

Reinforcement Learning (RL) is a rapidly growing field that can solve a wide range of difficult tasks, such as playing Atari games or robotic arm manipulation. Moreover, making RL algorithms *robust* against perturbations is essential to its utility in the real world to ensure high performance. Adversarial RL, in which an attacker attempts to degrade a system's performance by perturbing the environment, can be used to understand how to robustify systems. However, it is important to design algorithms that perform well in both perturbed and unperturbed environments. Most previous works about this best of both worlds focus on bandit settings, however we focus on Markov Decision Processes with multiple layers. In this presentation, we will explain the basics of Adversarial RL and provide an adversarial algorithm along with its proof of optimality. Analyzing the optimal adversarial strategy is a key step to understanding how to design a robust RL algorithm.

### Eric Chen and Boyan Litchev

*Truly anonymous sealed sender in Signal*

**Mentors: Kyle Hogan and Simon Langowski**

Recently, demand for anonymous communication has grown tremendously, with one of the most popular trusted privacy-oriented apps being Signal. In 2018 Signal devised its Sealed Sender protocol, which provides some anonymity to its users. However, it has timing-related side channels that deprive users of long-term anonymity. Using Private Information Retrieval and Signal's existing Sealed Sender protocol, we create a new scheme which resolves the timing attacks on Signal's current protocol and provides strong and practical anonymity guarantees.

### Simon Beyzerov, Hyojae Park, and Eliyahu Yablon

*Private access control for function secret sharing*

**Mentor: Sacha Servan-Schreiber**

Function Secret Sharing (FSS; Eurocrypt 2015) allows a dealer to share a function $f$ with two or more evaluators. Given secret shares of a function $f$, the evaluators can locally compute secret shares of $f(x)$ on a common input $x$, without learning any information about $f$ in the process. In this talk, we initiate the study of access control for FSS. Given the shares of $f$, the evaluators can ensure that the dealer is authorized to share the provided function. For a function family $\mathcal{F}$ and an access control list defined over the family, the evaluators receiving the shares of $f \in \mathcal{F}$ from the dealer can efficiently check that the dealer has the appropriate access key for $f$. This model enables new applications of FSS, such as: (1) anonymous authentication in a multi-party setting, (2) access control in private databases, and (3) authentication and spam prevention in anonymous communication systems.

Our definitions and constructions abstract and improve the concrete efficiency of several recent systems that implement ad-hoc mechanisms for access control over FSS. The main building block behind our efficiency improvement is a discrete-logarithm zero-knowledge proof-of-knowledge over secret-shared elements, which may be of independent interest.

We implement and evaluate our constructions. We show a 70-80x reduction in computational overhead compared to existing access control techniques used in anonymous communication. In other applications, such as private databases, the processing cost of introducing access control is only 1.5-3x.

## Session9: Computational Biology I

### Anish Mudide

*Uncovering the genomic basis for extinction risk via practical machine learning*

**Mentor: Dr. Ayshwarya Subramanian, Broad Institute**

For decades, we have evaluated extinction risk based on ecological features (e.g. population size, geographic range and breeding rates) collected via fieldwork. Such features are expensive and time-consuming to gather, which has restricted our understanding to just 100,000 of the over 8,000,000 species inhabiting Earth. What if, instead, we could evaluate extinction risk by sequencing a single genome of a species? Genomic datasets pose a host of issues and peculiarities ranging from missing data to massive feature sets. In this talk, we overcome these issues by employing practical machine learning strategies on a recently published comparative genomics dataset from the Zoonomia Project. We detail common pitfalls we avoided to prevent data leakage and maximize generalizability. Our model can inform future conservation interventions by predicting the extinction risk of species not yet classified. Most importantly, our work provides support for a new, genomics-centered approach to conservation.

### Achyuta Rajaram

*Comparative analysis of mouse and human podocytes with scRNA-seq*

**Mentor: Dr. Ayshwarya Subramanian, Broad Institute**

The use of animal models, especially mouse models, has revolutionized human biology with massive advancements in drug testing and disease progression modeling. However, modern studies still suffer from a "valley of death" phenomenon, where transferring studies across species has high failure rates, resulting in resource-intensive clinical trials for products that end up with no useful products. The main way to remedy this is by way of greater understanding of model specimens, whether by drawing comparisons across species or comparing in-vitro organoids to their analogous organs in the native setting. The innovation surrounding scRNA-seq allows for significant progress in this field, as it unlocks higher resolution and specificity previously unavailable to previous studies. However, significant ambiguity exists surrounding the methods of answering such questions. We propose a framework for cross-species comparative analysis, as well as a concrete interpretable definition of cell type, which can be easily transferred to scRNA-seq studies across species. We test these definitions and frameworks on human-mouse kidney data, specifically by analyzing podocyte cells and comparing them across species.

### Tanmay Gupta and Raj Saha

*Surveying the presence and diversity of coronaviruses in mammalian transcriptomes*

**Mentor: Dr. Ayshwarya Subramanian, Broad Institute**

Zoonotic spillover 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 viruses. For instance, studies suggest that pangolins have passed SARS-CoV-2 from bats to humans, acting as intermediate hosts. We utilize public RNA sequencing datasets to identify the presence of viral RNA in a diverse array of species. To classify the coronaviruses present in the RNA-Seq samples, we use a metagenomics classifier that provides an abundance profile for all detected species in the sample. The amount of identified coronavirus RNA is quantified (by number of reads) in a matrix, with the sample species as rows and coronaviruses as columns. Additionally, we investigate the species' susceptibility to coronaviruses by analyzing the expression levels of receptors mediating the entry of coronaviruses. We quantify the abundance of receptor RNA by aligning the RNA-Seq samples to reference transcriptomes. Gaining a deeper understanding of the mammalian species susceptible to coronaviruses will help us better control the spread of virus. Furthermore, our approach can be extended to other zoonotic viruses, such as monkeypox and the West Nile virus.

## Session 10: Computational Biology II

### Steven Tan

*Models for somatic CAG repeat expansion in the onset and progression of Huntington's disease*

**Mentors: Bob Handsaker and Seva Kashin, Broad Institute**

Huntington's Disease (HD) is an inherited neurodegenerative disease caused by alleles with 36 or more repeats of the trinucleotide sequence CAG in the huntingtin (HTT) gene. A person with HD inherits an allele with a certain CAG length (>35) at birth, but somatic expansion within the brain is known to occur throughout their lifetime, resulting in a situation in which individual cells have longer and highly variable numbers of CAG repeats. Somatic expansion is increasingly thought to be a driver of disease onset, as age-at-onset associates with modifier alleles in DNA-repair genes that regulate somatic expansion. Thus, a better understanding of the mechanisms behind CAG repeat expansion could be crucial in revealing novel therapeutic targets. In our study, we adapted a stochastic birth-death model previously used for a different repeat-expansion disease (Myotonic Dystrophy Type 1, or DM1) to model CAG repeat expansion in HD. We made use of a new kind of biological data, in which CAG length has been measured precisely in many individual neurons of the vulnerable type from post-mortem brain samples. We found that single-process models consisting of only one length threshold and rate — models that succeeded in modeling DM1 — failed to model the repeat expansion observed in HD patients. Effectively fitting the data required models consisting of two separate processes, suggesting that there may be two distinct biological mechanisms underlying CAG repeat expansion in HD. These processes appear to have differing rates and CAG length thresholds: one at roughly 35 CAGs — a threshold for instability — and another at 70 CAGs, which we hypothesize is a threshold for accelerated expansion. The model deepens our understanding of disease progression and can inform the design of clinical trials for new therapies that target the somatic expansion process.

### Rianna Santra

*Diagnosing brain cancers with gene expression data using a novel neural network method*

**Mentor: Prof. Gil Alterovitz**

Brain cancer is a deadly type of cancer and can often go unnoticed until the tumors grow big enough to interrupt brain function, during which the patient no longer has enough time for treatment. In order to diagnose brain cancer early enough and for treatment to work successfully, as part of our research study, a neural network algorithm is designed and developed using gene expression data from the Curated Microarray Database (CuMiDa) and the Repository for Molecular Brain Neoplasia Data (REMBRANDT) Dataset. Although many classification algorithms have been developed previously, not all of these employ deep learning and only classify the broad type of cancer (breast, prostate, brain, lung, etc.) and not specific types of cancer (astrocytoma, glioblastoma, etc.). Since deep learning has succeeded in numerous fields with high dimensionality such as image, speech and text processing to predict cancer tumors, deep learning will also have a high accuracy in genomic data. In addition, the most relevant predictive genes are selected from the data so biologists can find a fundamental root cause for different types of brain cancer.

### Ho Tin (Alex) Fan and Rianna Santra

*Leveraging statistical distributions for RNA sequencing across time*

**Mentor: Prof. Gil Alterovitz**

In this talk, we analyze the RNA sequences of babies to see if there is any change in gene expression when the trainer is used. The trainer is an oral feeding machine to help the babies to feed after birth. Babies were randomized to either receive treatment from a trainer or a 'Sham' (a fake trainer, something similar to a placebo) to keep the control modality. They were assessed until they learned how to feed completely or were transferred to an outlying hospital. The primary goal of this study is to determine whether babies who learned how to feed successfully had a different gene expression than babies who did not. We explored different groups to see if gene expression changed by treatment status, sex, or both.

## Session 11: Computer Science II

### Michael Huang

*Theoretically efficient parallel density-peaks clustering*

**Mentors: Prof. Julian Shun and Shangdi Yu**

Clustering multidimensional points is a fundamental data mining task with applications in many fields, such as astronomy, neuroscience, bioinformatics, and computer vision. The goal of clustering algorithms is to group similar objects together. Density-based clustering is a clustering approach that defines clusters as dense regions of points. It has the advantage of being able to detect clusters of arbitrary shapes, rendering it useful in many applications.

In this talk, we propose fast and theoretically efficient parallel algorithms for Density-Peaks Clustering (DPC), a method for density-based clustering. DPC is effective in detecting clusters of arbitrary shapes and allows hyperparameter selection in a user-friendly fashion, unlike standard methods such as DBSCAN. However, existing exact DPC algorithms suffer from high computational cost both theoretically and in practice, which limits DPC's application to large-scale datasets. To remedy the performance issue, we propose three theoretically efficient exact DPC algorithms. Our most performant algorithm achieves lower work complexity (sequential runtime complexity) than the state-of-the-art DPC algorithm; it attains $O(\log(n))$ span complexity (parallel runtime complexity), a dramatic improvement from the $O(n^2)$ span complexity achieved by the previous best DPC algorithm. Our most performant DPC algorithm utilizes a novel data structure which we call a priority search $k$d-tree. We present the priority search $k$d-tree and provide complexity analysis for performing queries on this data structure.

We provide optimized implementations of our algorithms and evaluate their performances via extensive experiments. Running on a 30-core machine with two-way hyperthreading, we find that our best algorithm achieves a 8.3–4666.3x speedup over the previous best exact DPC algorithm. Compared to the state-of-the-art approximate DPC algorithm, our best algorithm achieves competitive results and is able to achieve a geometric mean speedup of 8.2x. Our DPC algorithms are scalable, attaining a 8.8–13.2x self-relative speedup.

### Joey Dong and Anshul Rastogi

*Locating regions of uncertainty in distributed systems using aggregate trace data*

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

Developers of distributed systems use distributed tracing to provide visibility into the complex interactions within the system. Oftentimes, they must manually instrument their codebases and have difficulty determining where to place instrumentation to maximize utility. We propose a tool that utilizes aggregate trace data from a distributed system to localize regions of high relative uncertainty. Our work will advise developers on where to investigate the system in case of unreliable structural or performance issues. We present an algorithm for this tool, limitations faced, and a prototype/proof-of-concept implementation in Python. We further discuss proposed techniques for evaluation with simulated distributed system platforms such as HotROD.

**Coleman DuPlessie and Eddie Wei**

*Deep learning transformers for non-cyclical kinematics*

**Mentor: Andrew Gritsevskiy, University of Toronto**

Transformers are powerful machine learning models that are especially good at capturing long-distance relationships in data. However, they have not been applied to human kinematics, a field which has seen significant success from the application of other machine-learning models. Unfortunately, common models such as LSTMs perform much worse on non-cyclical data than on cyclical data, which limits their use in the field of kinematics. We theorize that, because Transformers can better represent long-term dependencies, they will achieve superior performance on tasks in this field, where the time series data is significantly aperiodic. In this talk, we compare Transformers and similar models to LSTM models and a heuristic benchmark on non-cyclical, 3-dimensional positional data from CMU's Quality of Life Grand Challenge Kitchen dataset.