Feature Selection Using Explainable AI to Refine Associations Between Prominent Genes and Alzheimer’s Disease Neuropathology

Aditi Chandrashekar
Mentor: Su-In Lee

A critical focus of Alzheimer’s Disease (AD) research is the development of a generalized understanding of the genetic and molecular mechanisms involved in AD. The Multi-task Deep learning for Alzheimer’s Disease neuropathology (MD-AD) model, proposed by Beebe-Wang et al. (2021), learns interrelationships between brain gene expression data and neuropathological phenotypes related to AD in a multi-cohort setting. Leveraging Explainable AI (XAI) gives insight into how the model makes these associations. There is reason to believe that reducing the number of genes inputted into the model as features could result in clearer explanations of the model. Here we compare several methods to reduce the number of input genes to the model and discuss their tradeoffs. We show that using post hoc XAI-based gene selection methods to reduce the size of the feature space results in comparable performance to the original, more comprehensive MD-AD model, suggesting that, within the context of this task, the genes most strongly associated with the presentation of neuropathological phenotypes are representative of the full set of features.

Data-Adaptive Model Predictive Control of Linear Time Varying Systems With Imperfect Predictions

Emile Anand
Mentors: Adam Wierman and Yiheng Lin

Model Predictive Control (MPC) is a novel method used to control processes while satisfying a set of constraints, with the slated advantage that at each discrete time-step, the algorithm receives exact predictions of the costs, dynamics, and disturbances of the system for k future time-steps. These predictions are typically imperfect in real-world applications; moreover, the longer the predictive window, the noisier the predictions get. In this work, we propose an online pipeline and analysis framework to find the optimal prediction window even when the predictive errors accumulate with time. We generalize our work to apply for all controllers for which the trajectories of an initial state and its perturbed state exponentially converge to induce input-to-state stability. We then leverage existing algorithms for online convex optimization through multi-armed bandits for tackling the exploitation vs. exploration problem. Next, we propose a broader algorithm to dynamically tune the model’s confidence of the predictions it receives. Finally, we prove that these online data-adaptive algorithms produce a bounded sublinear dynamic regret.

Improve Online Optimization With Uncertainty-Quantified Advice

Jerry Huang
Mentors: Adam Wierman and Nicolas Christianson

We study the problem of online optimization, in which an agent must make sequential decisions such that the sum of per-round costs is minimized. The agent is aided by advice from an oracle paired with quality hints, such as the predictions from an uncertainty-quantified machine learning model, which provide potentially untrusted information on the optimal decision for the given round along with the oracle’s confidence in its proposed decision. We explore how hints can be used to improve performance while maintaining robustness guarantees in the classical problem of prediction with expert advice and convex function chasing.

Multi-agent Online Convex Chasing With Untrusted Machine-Learned Advice

Junxuan Shen
Mentors: Adam Wierman and Nicolas Christianson

We study the problem of k-chasing of convex functions in a leaning-augmented setting, where multiple decision makers must make decisions each round they receive a request, to minimize the sum of the convex hitting cost and the cost of switching decisions between the rounds. Unlike the traditional online algorithms, we assume that the k decision makers have access to some machine-learned predictions with untrusted quality. We design a novel memoryless algorithm for the 2-chasing problem on R, and prove that, it exploits the ML predictions when they are accurate, while also guaranteeing performance compare to the offline optimal decisions even when the ML predictions are poor. Our competitive ratios are error-dependent, which is a function of some parameters that is adjustable by the users, depending on their confidence level of the ML predictions.
In July, 2022, JPL and NASA launched an instrument to the International Space Station. The EMIT mission’s purpose is to determine the effect mineral atmospheric dust clouds have on a region’s climate. EMIT tracks ground mineralogy and uses Aerosol Optical Depth (AOD - distribution of aerosols) models to determine dust cloud composition. The first task was to determine these models’ accuracy by comparing predicted and actual AOD. If accuracy is known, we can better predict dust composition. Time-efficient code was written to process AERONET sites’ data (actual AOD measurements), interpolate AOD for the desired wavelength, and return AOD information and sites in the same location, time, and date as an EMIT image. A program to compare actual and predicted AOD information is being developed. Moreover, because clouds in EMIT images distort data, the second task was to utilize machine learning to develop a program that identifies and removes clouds. Pixels in 10-12 images were annotated and classified. Code was written to process each pixel in annotated images, separate annotated pixels from the rest, calculate reflectance values at each pixel, train a model, test it, classify/map test data, and determine accuracy. The accuracy is currently 96%, which will be further improved!

Learning to Optimize with Preference Feedback
Chu Xin Cheng
Mentors: Yanan Sui and Yisong Yue

The Dueling Bandits problem is a stochastic Multi-Armed Bandits framework for learning from preference feedback, where the learner hopes to choose the optimal action with data given in the form of comparisons from the user. SELFSPARRING is an algorithm for the Dueling Bandits problem which views the optimization process as a multi-player cooperative game and can be easily generalized to a setting in which the arms are dependent with each other. Coactive Learning is another online learning framework that, instead of receiving preference feedback, obtains suggestions from the user in each round. Combining these two frameworks, the CoSPAR algorithm can be implemented in situations with preference or coactive feedback and has seen its application in exoskeleton gait optimization. We provide theoretical analysis for the algorithms KERNELSELFSPARRING and CoSPAR and attempt to generalize these algorithms to different online learning settings.

Compact and Efficient Chemical Boltzmann Machines With an Autonomous Learning Rule
Inhoo Lee
Mentors: Erik Winfree and Salvador Buse

Stochastic chemical reaction networks (CRNs) – a model of computation describing molecular systems with limited counts of species - can perform complex computations. CRNs can capture how biological cells are able to navigate, understand, and manipulate their environments, using species with counts often as low as one or several. In particular, stochastic CRNs are well-suited to behaving as stochastic neural networks, which are of interest due to their high computational power relative to the size of their chemical implementations. The Boltzmann Machine is a stochastic neural network capable of representing distributions, performing inference, and learning from data. Chemical Boltzmann Machines are stochastic CRNs that either exactly replicate or approximate the behavior of an abstract Boltzmann Machine. Our work proposes a new stochastic CRN implementation of a Chemical Boltzmann Machine, with an autonomous learning rule and a lower molecularity (fewer species per reaction) than prior work. The lower molecularity increases the plausibility of physical realizations of the CRN. An autonomous learning rule, with learning from data happening end-to-end within a chemical system, can further our understanding of how intelligent chemical agents such as cells can come to understand and make decisions.

Stochastic Local Search in Surface Chemical Reaction Networks to Solve 3-SAT Problems
Mohini Misra
Mentors: Erik Winfree and Salvador Buse

Chemical reactions drive biological phenomena at a molecular level. To study the sophisticated functionality that is thus enabled in molecular environments, the scientific community has developed models of computation to understand and engineer such systems. Chemical Reaction Networks (CRN) are one such model of computation, which have been studied extensively in well-mixed environments. Chemical reactions are inherently stochastic, so CRNs are modeled to be as well, allowing us to exploit randomness in algorithms we create. This is particularly useful for solving Constraint Satisfaction Problems (CSP) with Stochastic Local Search (SLS). In well-mixed environments, a CRN for the 3-SAT CSP has been previously developed. However, this CRN must scale up the number of species and reactions with the size of the problem. Here we show that such scaling isn’t necessary if we solve 3-SAT using a surface CRN, where reactions can be localized and thus reused in different contexts. First we present a simple surface CRN (31 reactions, 11 species) which only resolves constraint conflicts, and second we present a more complex surface CRN which resolves constraint conflicts more efficiently by harnessing directionality and bias. This work has implications for how CSPs are approached in biological phenomena that depend on geometry and location, such as self-assembly and self-healing.
Using Stochastic Surface Chemistry to Solve the k-Coloring Constraint Satisfaction Problem
Jean-Sebastien Paul
Mentors: Erik Winfree and Salvador Buse

Well-mixed stochastic Chemical Reaction Networks (CRNs) have been shown able to exploit their inherent stochasticity to solve hard NP-Complete constraint satisfaction problems. Seeing how the inherent randomness of low-level chemistry can be used to form efficient stochastic local search algorithms, one intuitively asks how this might apply to surface CRNs— an equally stochastic, yet spatially restricted counterpart to stochastic CRNs.

In this vein we explore the graph k-Coloring problem. By treating chemical reactions as a means to identify violations of defined constraints, we are able to randomly resolve vertex coloring conflicts until a valid solution is found— a defined equilibrium. Harnessing the surface CRN’s spatial constraints, we achieve a surface CRN (DirectColoringCRN) with 7 species and 24 reactions that can solve any solvable planar 3-coloring problem. We are also able to construct a surface CRN with 15 species and 72 reactions that implements biased local search to improve performance in the 3-coloring case (RejectWaveColoringCRN). Finally we are able to create a larger CRN that uses directional propagation of information (in addition to local search bias) to effectively solve k-coloring in all arbitrary solvable graphs (RejectSignalColoringCRN); that architecture provides a promising springboard for the implementation of a surface CRN neural network.

Enhancing Gamification Aspects for Mobile App ZTF Augmented Reality Transient Hunter (ZARTH)
Anika Arora
Mentors: Ashish Mahabal and Ivona Kostadinova

The Zwicky Transient Facility (ZTF) detects transients in the Northern sky using a 47-deg² CCD camera. Through the app ZTF Augmented Reality Transient Hunter (ZARTH), users can locate planets, stars, and transients found by ZTF. Upon locating transients, users earn points and view their properties, a concept similar to Pokémon Go. To improve ZARTH’s UI, there must be a representative, sparse distribution of transients across the sky. Using Alert-Classifying Artificial Intelligence, a deep learning framework to classify ZTF alerts, we have created an algorithm which removes bogus transients and includes variable, orphan, hosted, and nuclear transients. This algorithm filters ZTF’s 100,000 transients per night into a uniformly distributed subset of 1,000. A separate algorithm assigns points to transients in inverse proportionality to their rarity, which is calculated from their magnitude, classification, and spectra. We can improve these two algorithms by considering other transient properties, like its filter.

Bringing Machine Learning to Planetary Protection Analysis
Subin Kim
Mentors: Ashish Mahabal and Nitin Singh

As humans continue to explore outer space, it is crucial that we are conscious of the impacts we make on the outer space environments such as the interplanetary contaminations. JPL generates a multitude of biological data from spacecraft-associated environments and organisms, however, currently there is no database that can be used to derive Planetary Protection relevant information from all these different databases. In our project, we aim to standardize these data, which can be partly achieved through the development of datasheets, which describe a species and its biological properties that are relevant to the measurement of the contamination risk it poses. Even with data sheets that compile multiple biological databases, a huge portion of our data sheets are incomplete due to the difficult nature of measuring data on the vast number of bacterial species. Thus, in our project, we develop a machine learning model to predict these missing data using data present for other species.

Analytic Continued Fractions for Multivariate Regression Models
Mabel Lu
Mentors: Pablo Moscato and Michelle Effros

Mathematical models for describing two-body interacting potentials have been notoriously difficult to obtain from experimental data and there is a current lack of an algorithmic solution that is simple and effective as well as problem-independent. For instance, in the Thompson Problem the potential between a pair of electrons is known, but the minimum configurations of a set of electrons spread out across the surface of a sphere is not. In the case of noble gases, like Argon, the problem is to identify the two-body potential as a function of the distance between the molecules given some experimental data. In using different combinations of truncated analytic continued fractions (CF) to model these problems, one is of particular interest due to its adaptability, which follows the form of the first CF multiplied by e to the power of the other CF. To develop an algorithm to fit a function of that model, we explored common trends that reduce both complexity and error, including the use of integer and prime coefficients. We also explore simple manipulation of data variables that serve to simplify the problem. These traits in addition to the model are especially powerful in that they are accurate and computationally advantageous.
Computing Tri-Partitions and Bases of an Ordered Complex
Erick Jimenez Berumen
Mentor: Peter Schröder

Given an ordered complex, $K$, for any dimension, $p$, the set of $p$-cells of the complex can be partitioned into three disjoint sets: (1) a maximal $p$-tree set, (2) a maximal $p$-cotree set, and (3) a set of remaining $p$-cells whose cardinality is the $p$-th reduced Betti number of $K$. This partition directly leads to a construction for bases of $p$-th cycle, boundary, and homology groups. Interestingly, the structure of this partition closely mimics the Hemholtz-Hodge decomposition of differential forms. We implement an algorithm introduced by Edelsbrunner and Ölsböck to compute tri-partitions for surface and volume meshes, and for the case of surface meshes, we use build on this algorithm to construct harmonic differential forms on the complex $K$. 