Caltech
Student-Faculty Programs

Session I Abstracts

Hardware-efficient generative quantum Eigensolver for ground state energy estimation
Stuart C. Florescu
Mentors: Alan Aspuru-Guzik, R. Michael Alvarez, and Austin Cheng

Accurately determining ground state energies of molecular Hamiltonians lies at the heart of quantum chemistry, condensed matter physics, and optimization, but doing so on today's noisy quantum devices remains a major challenge. Reliable energy estimation not only advances fundamental science but also underpins applications such as material design and drug discovery. Variational approaches such as the VQE offer a path forward but are hindered by barren plateaus and circuit depths that exceed hardware capabilities. Generative methods avoid some of these issues by directly learning to produce low-energy circuits, yet existing models rely on large, composite gate sets that are not suited to near-term devices.

That's why we develop the Hardware-Efficient Generative Quantum Eigensolver (HE-GQE), which generates quantum circuits composed solely of device-native gates. Our approach produces shallow, hardware-executable circuits that maintain competitive accuracy in estimating Hamiltonian ground states. Benchmarks on H_2 demonstrate reliable dissociation curves, and preliminary extensions to BeH_2, LiH, and N_2 suggest improved scalability over standard VQE baselines. By combining generative modeling with hardware-efficient design, this work establishes a practical framework for ground state estimation on noisy intermediate-scale quantum devices and points toward broader applications in quantum chemistry and combinatorial optimization.

Supporting teaching assistants in theoretical computer science courses
Joseph W. Giambrone III
Mentors: Adam Blank, Claire Ralph, and Matthew M. Gherman

Theoretical computer science classes at Caltech, like CS 38 (Algorithms), are notoriously difficult. In particular, there are typically not enough resources to support the teaching assistants ("TAs") for these classes. Additionally, CS 38 does not currently have recitations, leading to a potential disconnect between lecture materials and homework. Bridging this gap would help students take away more concepts from the class. Our objective is to make improvements to CS 38 informed by a literature review of the current research in computer science education. Specifically, we want to focus on what resources are effective for TAs in classes on algorithms and other theoretical computer science classes. Through this project, we have created new rubric items for archetypes of problems, hints for TAs to provide during office hours, new problems and solutions to replace the existing homework sets, and a full set of weekly recitation materials that include example problems and walkthroughs. We hope that as a result of our work, CS 38 will be a more enjoyable and less frustrating experience for students next spring.

Extending BayesRays for reliable spatial uncertainty in ill-posed inverse problems
Kyle T. Berkson
Mentors: Katherine L. Bouman and Brandon Zhao

Spatial uncertainty quantification is indispensable when implicit neural representations are used to solve ill-posed inverse-imaging problems. We investigate—a post-hoc Laplace-approximation method originally designed for 3-D NeRFs—as a drop-in estimator for 2-D and 3-D imaging pipelines whose reconstructions are parameterised by coordinate MLPs. On synthetic Fourier- and Radon-based data, BayesRays exhibits two systematic failure modes: (i) an inverse correlation with ensemble variance that over-confidently scores high-frequency structure while inflating uncertainty in flat regions, and (ii) "marbling'' artefacts that arise from the diagonal Fisher approximation. We trace both pathologies to a

mismatch between BayesRays' deformation grid and the geometry of general inverse problems, where the forward operator breaks the one-to-one mapping between image and measurement domains. Using this diagnosis we propose a banded-Hessian variant that suppresses marbling without sacrificing computational efficiency, and outline a roadmap for extending these ideas to VLBI and dynamic black-hole tomography. These results show that naïvely transplanting BayesRays is insufficient, but that a geometry-aware adaptation can yield trustworthy spatial error bars for scientifically critical reconstructions.

Understanding the computational limits of bipartite graph alignment
Eric M. He
Mentors: Daniel Cullina and Adam C. Wierman

The growing use of anonymized datasets raises concerns about privacy, particularly their susceptibility to de-anonymization attacks via alignment of correlated data. While graph alignment has been studied extensively in the undirected setting, many real-world settings involve bipartite structures with two distinct node sets, such as users and features, that are both independently anonymized and shuffled. We investigate the information-theoretic limits of aligning correlated bipartite graphs. Specifically, we aim to understand the computational boundary of this problem, where statistically correct recovery is possible but no known polynomial-time algorithms succeed, and explore how different alignment algorithms behave in this regime. Using a correlated Erdős–Rényi model as a base, we analyze the relationship between various graph parameters and alignment accuracy, and we propose novel algorithmic approaches informed by such probabilistic analysis. Our work provides a deeper understanding of the computational feasibility of different regimes in bipartite graph alignment for de-anonymization.

Multivariate continued fraction regression with target derivatives
Kieran A. Hale
Mentors: Pablo A. Moscato, Adam C. Wierman, and Andreas Heinecke

Continued Fraction Regression (CFR) is a recent method in symbolic multivariate regression that is based on representing the model as a continued fraction. CFR can be implemented with flexible depth naturally limiting the need to impose a structure on the model before training, which symbolic regression methods often do. The appeal of symbolic regression itself over methods like ANNs or SVR lies in the potential simplicity of the resulting model and its amenability to downstream analysis. For univariate problems, it has been shown that supplying both approximate and exact derivative information as part of a memetic algorithm for CFR can improve the performance of the resulting models – this method has been called CFR with target derivatives. This work extends CFR with target derivatives to multivariate regressions.

Towards a digital twin: Applying ML methodologies to complete solar event data
Chigozirim N. Ifebi
Mentors: Hillary Mushkin and Ashish Mahabal

This project investigates methods for analyzing and predicting high-latitude particle precipitation using Defense Meteorological Satellite Program (DMSP) data and physics-based and machine learning models. I began by exploring the Ovation Prime model and precipNet framework, developing visualizations of auroral boundaries and electron energy flux in both polar and time-series contexts. To evaluate predictive modeling, I trained machine learning baselines, including random forests, to forecast electron energy flux month-ahead; while initial results highlighted challenges due to the highly variable nature of the flux, they provided a benchmark for comparison. I then implemented the SAITS (Self-Attention-based Imputation for Time Series) architecture to address data gaps and explore sequence-to-sequence forecasting in a multivariate setting that included geomagnetic indices (AE, Bz, SymH). This required restructuring the dataset into sequential formats, masking values for imputation, and adapting GPU-based training in Google Colab. Alongside, I developed interactive

visualizations to contextualize model outputs with DMSP observations. The results demonstrate both the promise and challenges of applying advanced sequence models to space physics: while computational resource limitations constrained large-scale experiments, the pipeline establishes a foundation for using attention-based architectures to better capture precipitation variability.

Data visualization for an exploratory sun emissions dashboard
Anya B. Mischel
Mentors: Hillary Mushkin, Allan Labrador, Ashish Mahabal, and Santiago V. Lombeyda

Understanding solar flare activity is a critical area of astrophysics due to its potential to disrupt satellites, communication systems, and power grids on Earth. With over 25 years of solar probe data totaling approximately 1 TB, efficient data summarization and visualization are essential for identifying trends in solar activity. We developed an interactive web-based visual analysis interface and dashboard utilizing data from the STEREO and ACE missions. This tool enables researchers to filter solar events by various determinants (including emission strength, classification, etc), examine specific time windows, and sort by various parameters. Researchers can explore relationships between key variables such as event duration, peak emission magnitudes of detected elements, element ratios (e.g., iron-to-oxygen), and overall flare intensity. This tool facilitates deeper insights into the dynamics of solar flare activity and lays the groundwork to integrate additional datasets from other solar probes and add additional features.

Using an expected volume heuristic to guide adaptive sampling of molecular dynamics simulations of proteins
Gavin N. John
Mentors: Gregory Bowman and Mitchell Guttman

The vast majority of drugs target proteins. As such, identification of protein binding sites and an understanding of protein conformational heterogeneity are important first steps in virtual drug discovery. The Bowman lab specializes in the identification and characterization of cryptic pockets: binding sites normally internal to the protein, but that become exposed as the protein undergoes conformational changes. Due to their nature, these are much harder to identify, but represent a massive number of potential drug targets. The lab has developed several tools to assist in the discovery of cryptic pockets, of which two are particularly important. The first, PocketMiner, is a neural network that can quickly and accurately predict residues that might form part of cryptic pockets. The second, FAST (Fluctuation Amplification of Specific Traits), is a conformational analysis tool that uses adaptive sampling and a customizable geometric component to direct molecular dynamics simulations to conformational spaces of interest. Here, we developed a new geometric component that uses the predictions from PocketMiner to generate an estimate of the pocket volume that was used in a FAST simulation of a candidate protein target, TEM-1 beta-lactamase, which successfully demonstrated pocket opening behavior.

Functional flow matching for self-supervised representation learning on time series data
Duy H. Nguyen
Mentors: Anima Anandkumar, Jiachen Yao, and Jiayun Wang

Self-supervised approaches like masked autoencoding have proven effective for learning time series representations. However, their architectures typically yield a single static feature vector, limiting their ability to support downstream tasks that require multi-level classification from the same input signal. We propose a self-supervised learning framework based on Functional Flow Matching (FFM), a continuous-time generative model that operates directly in function spaces. By pretraining an FFM model, we learn rich and label-free representations of time series trajectories. We address the challenge of distribution shift between the noisy inputs seen during pretraining and the clean data samples used for downstream tasks. We hypothesize that because FFM allows for extracting features at any intermediate time step of the generative process, we can access a feature hierarchy corresponding to different levels of abstraction. A key novelty of our approach is exploiting this to

perform multi-level feature extraction—capturing everything from low-level patterns to high-level semantic concepts—all from a single pretrained model. We will demonstrate that this approach enables more powerful and versatile representations, and we evaluate our method on downstream tasks, including multi-label classification.

Towards formal verification of neural networks in Lean
Jennifer A. Cruden
Mentors: Anima Anandkumar and Robert Joseph George

Modern neural networks are increasingly deployed in safety-critical domains, yet empirical testing alone leaves potential failures undetected. We present a machine learning framework in Lean, an interactive theorem prover, that supports both rigorous proofs of network behavior and concrete execution. The framework provides a flexible tensor library and modular layer definitions for common architectures such as multilayer perceptrons, convolutional networks, and transformers. Networks are defined over a generic scalar type. This allows the same model to be analyzed with real numbers for formal proofs or executed with rationals or floats for computation and interoperability with tools like PyTorch. To illustrate feasibility, we specify and run a two-layer perceptron in both Lean and PyTorch, while reasoning about its properties over the reals. Although current results focus on idealized reals, future work aims to address floating-point semantics. This unification of proof and execution within a single system lays the foundation for end-to-end formal verification of neural networks in Lean, offering a basis for trustworthy AI.