

Valentin Badea

“A Multi-Scale Computational Framework for the Longitudinal Analysis of Enlarged Perivascular Space Distributions Under Sleep Deprivation”

Thesis Advisor: Charles Guttmann, MD

Abstract:

The glymphatic system, responsible for metabolic waste clearance in the central nervous system, has emerged as a critical pathway in the study of neurodegeneration. Enlarged perivascular spaces (ePVS), visible on standard MRI, represent one of the few accessible imaging biomarkers of glymphatic activity in humans. However, existing approaches to longitudinal glymphatic assessment have relied on coarse-grained summary metrics or highly localized diffusion-based indices, leaving the fine-grained spatiotemporal dynamics of ePVS distribution largely unexplored. In this work, we analyze a 7T MRI dataset of 30 healthy subjects scanned at three time points: before sleep deprivation (Baseline), immediately following 24 hours of total sleep deprivation, and 72 hours into recovery. Longitudinal ePVS changes are investigated at four biological scales: whole-brain, regional, sub-regional, and single-voxel.

In our first specific aim, we show that coarse-grained ePVS burden metrics do not reveal significant cohort-level responses to sleep deprivation. However, projecting subject-level trajectories into a novel transition space uncovers a robust elastic behavior of the glymphatic system: perturbations in ePVS burden induced by sleep deprivation tend to resolve following recovery, across both segmentation modalities and most brain regions. In our second specific aim, we introduce two novel computational frameworks. A mathematically principled SE(3) mid-space registration pipeline is developed and validated, ensuring unbiased alignment of longitudinal scan pairs. Building on this, a Deformation-Based Morphometry pipeline and an Optimal Transport framework produce single-voxel maps of volumetric changes in the brain parenchyma and ePVS mass redistribution respectively, summarized into fingerprint representations. Cross-transition correlation analysis reveals patterns consistent with the elastic glymphatic response and suggests a directional association between parenchymal adaptations during sleep deprivation and ePVS dynamics during recovery. These results demonstrate the potential of spatially resolved longitudinal ePVS analysis for advancing our understanding of glymphatic dynamics.

Mohammed Baharoon

“AI Systems for Understanding and Grounding Radiology Reports”

Thesis Advisor: Pranav Rajpurkar, PhD

Abstract:

Radiology reports are the primary medium through which radiologists communicate, conveying critical clinical information in natural language. AI holds the potential to both generate and analyze these reports, yet two key challenges persist. First, as AI systems increasingly generate radiology reports, evaluating their accuracy remains an open problem. Second, a fundamental disconnect exists between the findings described in radiology reports and their corresponding locations in the imaging studies, limiting referring physicians, patients, and trainees who must interpret findings without explicit visual guidance. This thesis addresses these challenges through three interconnected contributions spanning report evaluation, report visualization, and clinical application.

First, we introduce CRIMSON, a clinically grounded evaluation metric for radiology report generation, along with two new benchmarks: RadJudge, a 30-case clinical judgment test suite, and RadPref, a 100-case radiologist preference benchmark. CRIMSON incorporates patient context and weights errors by clinical significance when comparing generated against reference reports. Validated against radiologist error counts from the ReXVal dataset, RadJudge, and RadPref, CRIMSON achieves stronger alignment with expert judgment than prior metrics.

Second, we introduce ReXGroundingCT, the first publicly available dataset linking freetext radiology findings to manually annotated 3D segmentation masks in chest CT scans. Designed through a multi-stage annotation pipeline, the dataset comprises 3,142 scans and 8,028 segmented findings. We benchmark state-of-the-art text-prompted segmentation models, demonstrating that current approaches fall substantially short of clinical utility, even after fine-tuning. Subsequently, we release the dataset, along with a public leaderboard to drive continued progress on this task.

Third, we present RadGame, an AI-powered platform for radiology education that brings together both report evaluation and grounding as a concrete downstream application. The platform teaches two core skills: localizing findings through interactive bounding-box annotation, and writing radiology reports with automated structured feedback. In a prospective multi-institutional study with 18 medical students, participants using RadGame achieved a 68% improvement in localization accuracy and a 31% improvement in report-writing scores, outperforming traditional passive learning methods.

Together, these contributions address the radiology AI pipeline from multiple angles: evaluating generated reports against clinical standards, spatial grounding of reports in imaging, and applying both to advance radiology education.

Yourong Bao

“Longitudinal and Multi-Omics Analysis of Metastatic Melanoma Reveals Dynamic Tumor Evolution and Multi-Level Heterogeneity”

Thesis Advisor: David Liu, MD, MPH, MS

Abstract:

Melanoma is the most lethal cutaneous cancer. Advanced-stage (Stage III-IV) melanoma has a 10-year survival rate of less than 30%. Despite recent advances in targeted and immunotherapies, clinical outcomes of melanoma patients remain highly variable, largely due to its highly heterogeneous nature. This heterogeneity is observed in 3 levels: intratumoral, intertumoral, and inter-patient, complicating clinicians' efforts to identify consistent biomarkers and therapeutic targets. Therefore, traditional approaches that focus on individual genes or pathways are insufficient to capture the complexity of metastatic melanoma progression and treatment response. To address this limitation, integrative multi-omics strategies are required to systematically characterize tumor evolution and tumor-immune interactions across different biological scales.

In this study, we analyzed two complementary melanoma patient cohorts. A longitudinal in-transit metastasis (ITM) cohort with whole-exome sequencing (WES) and bulk RNA-seq data was used to reconstruct phylogenetic trees and investigate tumor evolutionary dynamics over time. A second cohort from the Human Tumor Atlas Network (HTAN), comprising WES, single-nucleus RNA sequencing (snRNA-seq), spatial transcriptomics (MERFISH), and spatial proteomics data, enables the characterization of tumor cell states and immune interactions across defined clinical response groups. We integrated genomic, transcriptomic, and spatial-omics analyses to identify recurrent patterns of tumor progression, including clinically relevant clonal expansion, biological or functional pathways, mutational signatures, and melanoma state transitions.

Our results reveal that melanoma progression is characterized by dynamic clonal evolution that is associated with therapeutic interventions. During melanoma progression, we observed a decreased genomic heterogeneity but increased phenotypic and mutational process diversities over time. Non-progressor tumors exhibited enhanced antigen presentation and interferon signaling, highlighting the importance of tumor-immune interactions in treatment outcomes. We also identified biologically interpretable tumor programs that capture coherent tumor-intrinsic states beyond single-gene signals, which can be validated across modalities and applied clinically. Together, this study demonstrates that melanoma progression is driven by coordinated genomic evolution, phenotypic plasticity, and tumor-immune interactions, which underscores the value of multi-modal integration for understanding complex cancers.

Grant Bell

“Channel-Level Seizure Onset and Offset Detection in Stereo-EEG with Sub-Second Temporal Precision Using Machine Learning”

Thesis Advisors: Mark Richardson, MD, PhD, and Boris Katz, PhD

Abstract:

The analysis of stereo-EEG (SEEG) recordings is integral to the surgical treatment of patients with drug-resistant epilepsy. These data, recorded from intracerebral depth electrodes, provide the spatiotemporal resolution necessary to identify the seizure onset zone and characterize the epileptogenic network. While prior machine learning approaches have demonstrated strong performance on seizure detection benchmarks, they lack the resolution and computational scalability required to precisely localize seizure onset and offset for each individual recording channel. We propose a two-stage framework for automated channel-level seizure detection in SEEG recordings with sub-second precision. Our model combines a permutation-based coarse-grain detector that flags potential seizure events within the recording and a fine-grain detector that performs window-level classification within each detected interval to identify seizure onset and offset for each channel. Our dataset contained 5723 hours of SEEG recordings and 677 annotated seizure events from 26 patients who underwent SEEG at Massachusetts General Hospital. When evaluated on a held-out test cohort of 12 seizure events from three patients, the coarse-grain detector achieved an F1 score of 0.608 and the fine-grain detector achieved an F1 score of 0.16 on the window level classification task. The fine-grain detector’s low F1 score was driven by low precision, reflecting a tendency to over-predict seizure windows within intervals flagged by the coarse-detector. We also train and evaluate the fine-grain detector on the Bonn EEG dataset, where it achieves an F1 score of 0.981. Although the Bonn dataset is not directly comparable to our task of sub-second channel-level seizure onset and offset detection, it demonstrates strong performance on standard seizure classification tasks. In summary, this work combines a computationally efficient coarse-grain detector with a transformer-based fine-grain classifier for automated channel-level seizure detection in clinical SEEG recordings. This framework enables the generation channel-level seizure onset and offset predictions at the spatial and temporal resolution required for clinical analyses of the epileptic network while also addressing the computational constraints of analyzing large-scale multi-channel SEEG data.

Shuo Cai

“Compositional Phenotype Representations and Clinical Prior Distillation for Missense Variant-Disease Association Retrieval”

Thesis Advisor: Tianxi Cai, ScD

Abstract:

Missense variants are by far the most prevalent type of pathogenic coding variants, but they are predominantly annotated as variants of uncertain significance (VUS) in ClinVar. This categorization prevents any medical intervention despite the variant being causal. Existing computational tools predict pathogenicity well but cannot answer the question that drives genetic diagnosis: harmful for which disease? We present PheMART2, a contrastive retrieval framework that ranks candidate diseases for each query missense variant. To solve these problems, we implement two major architectural designs to improve generalization. Diseases are encoded using weighted attention-based combinations of HPO terms instead of relying on concept IDs, allowing the model to generalize to disease labels outside the training distribution. Population-scale variant–phenotype relationships from the MVP phenome-wide association study are mapped to the HPO disease space through a curated CUI mapping and SapBERT-based semantic translation pipeline, and then distilled into our model using bidirectional Kullback-Leibler divergence in both the variant→disease and disease→variant directions. Model evaluation is performed by randomly splitting data on the basis of genes instead of per-variant random splits with negative sampling, across all 1,563 diseases in the dataset. In comparison with this more stringent evaluation approach, PheMART2 demonstrates an MRR of 0.272 for unseen genes, an 18.1% improvement over the no-KL, no-auxiliary version ($p = 0.001$), with the clinical prior distillation being the key contribution. Disease label out-of-distribution inference shows an MRR score of 0.39, suggesting that our compositional encoding successfully generalizes to diseases constructed from novel phenotypic primitives. Amongst 676 independently curated Phenopackets cases with no overlap between variants used in the training data, the top 10 candidate list includes the true disease for 51.5% and the top 1 for 18.5%. Performance concentrates on variants in genes with existing biological annotations (MRR = 0.418) and falls sharply on genes without much connection in the knowledge graph (MRR = 0.089), which reflects the inherent constraint of the transductive graph setting and points to inductive extension as the natural next step.

Shyam Chandra

“Speech as a Digital Biomarker: A Computational Framework for Severity and Progression in Parkinson's Disease and Cerebellar Ataxia”

Thesis Advisor: Anoopum Gupta, MD, PhD

Abstract:

Parkinson's disease and cerebellar ataxia are progressive neurodegenerative conditions monitored through clinician-administered rating scales that cannot support the continuous assessment needed to detect subtle change or serve as sensitive clinical trial endpoints. Speech, non-invasively collectible at scale and disrupted early and prominently in both conditions, offers a path toward objective monitoring. This thesis develops and evaluates a computational framework for speech-based severity assessment using longitudinal data from the Neurobooth platform at Massachusetts General Hospital, where standardized diadochokinetic speech tasks and same-day clinical assessments (MDS-UPDRS Part III; BARS) are collected within a single clinic visit for 163 PD patients, 217 ataxia patients, and 87 healthy controls across 921 sessions.

Three modeling approaches are evaluated across seven cross-sectional prediction targets: a handcrafted acoustic feature pipeline with classical machine learning (LASSO, XGBoost, SVM); a deep learning framework based on mel spectrogram gradient representations and ResNet-18; and frozen transformer-based transfer learning using wav2vec2 and WavLM. From diadochokinetic speech tasks alone, both disease cohorts are separable from healthy controls (AUC 0.824 for PD, 0.857 for ataxia) and from each other (AUC 0.863). For ataxia severity, BARS total R^2 reaches 0.621 (ResNet-18 with strong L2 regularization) and 0.620 (WavLM Transfer), competitive with the published benchmark on a separate cohort, and WavLM achieves the best BARS speech result ($R^2 = 0.601$). A cross-disease biomarker analysis identifies 62 shared acoustic features, 66 PD-specific, and 85 ataxia-specific, characterizing the acoustic divergence between hypokinetic and ataxic dysarthria.

Post-hoc sensitivity analyses show that deep learning models trained on cross-sectional severity produce statistically significant upward drift in predictions across patient sessions (Cohen's $d = 0.249$ - 0.418 , all four targets) despite no longitudinal supervision. WavLM Transfer additionally separates pharmacological state at the group level in Mild PD without medication supervision: OFF-state sessions receive significantly higher predicted severity than ON-state sessions ($p = 0.003$, rank-biserial $r = -0.235$). Taken together, these findings support the further development of speech-based measurement as a quantitative tool for remote assessment, longitudinal monitoring, and clinical trial outcome measurement in neurodegenerative disease.

Jiayi Chen

“Sc-Multiome Reveals a Time-dependent Regulatory Role of MECOM in Vascular Development”

Thesis Advisor: Kaifu Chen, PhD

Abstract:

Vascular development requires the precise coordination of endothelial and mural cell differentiation through tightly regulated intercellular signaling. MECOM is an established endothelial identity gene associated with endothelial lineage specification, yet its broader regulatory role during vascular development remains unclear. Using integrative single-cell transcriptomic and chromatin accessibility analyses across human vascular organoid and mouse embryonic heart, this study demonstrates that MECOM loss disrupts both endothelial and mural cell fate specification, resulting in dysregulation of vascular lineage composition during vascular development. Moreover, MECOM ablation promotes the activity of key intercellular signaling pathways for vascular development at early development stage, whereas such pathways exhibit reduced signaling activity upon MECOM knockout (KO). Specifically, TGF- β signaling is identified as a key pathway subject to this MECOM-mediated temporal regulation in both human organoid and mouse heart. Single-cell chromatin accessibility analysis further revealed that Mecom-KO increases accessibility at TGF- β pathway gene loci, suggesting its role in epigenetic regulation over key vascular signaling pathways. These findings establish MECOM as a temporal gatekeeper of vascular intercellular signaling, coordinating the differentiation of both endothelial and mural lineages through precise regulation of developmental pathway activity and chromatin accessibility during vascular development.

Stephanie Chen

“Predicting Progression from Mild Cognitive Impairment to Alzheimer’s Disease Using Blood-Based Biomarkers and Amyloid-Beta PET: A Survival Analysis Approach”

Thesis Advisor: Chirag Patel, PhD

Abstract:

Early identification of mild cognitive impairment (MCI) patients at high risk of progressing to Alzheimer’s disease (AD) is critical for therapeutic planning, clinical trial enrollment, and prognostic assessment. Although amyloid-beta positron emission tomography (PET) is a gold-standard for detecting amyloid plaques, a hallmark of AD neuropathology, its cost, invasiveness, and limited accessibility constrain widespread use. Blood-based biomarkers offer a more practical and scalable alternative, with phosphorylated tau 217 (p-Tau217) emerging as a promising marker of amyloid-related pathology. However, its utility for predicting the timing of progression from MCI to AD dementia remains incompletely characterized.

In this study, we compared the prognostic performance of p-Tau217 with amyloid-beta PET and the p-Tau217/amyloid- β 42 blood ratio test, which is currently the only FDA-approved diagnostic test for AD. We also assessed whether adding complementary AD-related biomarkers, including amyloid- β 42/40, neurofilament light chain (NfL), and glial fibrillary acidic protein (GFAP), improves prediction.

We analyzed adults with MCI from the Alzheimer’s Disease Neuroimaging Initiative, classified as stable MCI (sMCI) or progressive MCI (pMCI) based on conversion to AD dementia during follow-up. Random survival forest and Cox proportional hazards models were used to evaluate the prognostic value of amyloid-beta PET and blood-based biomarkers for predicting progression from MCI to AD. Models were benchmarked against a baseline model including age, sex, education, and APOE ϵ 4 status.

The cohort included 157 sMCI participants (mean age 74.4 years, 61.4% female) and 44 pMCI participants (mean age 75.3 years, 61.4% female), with a mean follow-up of 4.34 years. At four years, all biomarker-based models outperformed the demographics + APOE ϵ 4 model, including the combined blood biomarker panel (AUROC = 0.818, 95% CI [0.748-0.888]; C-index = 0.819 ± 0.093), p-Tau217 alone ($0.787 [0.705-0.868]$; 0.808 ± 0.097), and amyloid-beta PET imaging ($0.758 [0.669-0.848]$; 0.798 ± 0.114). p-Tau217 and amyloid-beta PET performed comparably for predicting MCI-to-AD progression, and there was no evidence that the multi-marker panel or p-Tau217/amyloid- β 42 ratio improved upon p-Tau217 alone. Partial dependence analysis showed that as p-Tau217 increased from 0.08 to 0.56 pg/mL, the predicted 4-year probability of conversion rose from about 10% to 45%, supporting p-Tau217 as a meaningful and minimally invasive prognostic marker in prodromal AD.

Yanshi Chen

“Evaluating the Statistical Fidelity and Utility of Aggregated Patient Count as a Surrogate for Line-Level EHR in Kidney-Transplant Analyses”

Thesis Advisor: Kenneth Mandl, MD, MPH

Abstract:

Electronic Health Records (EHR) contain valuable information for clinical and population-level studies, but the sharing of raw data is strictly constrained by privacy regulations. Cumulus addresses this challenge by generating privacy-preserving aggregated patient-count data, termed “CUBE”, which is a power set matrix of variable combinations with k-anonymity constraints. However, the statistical fidelity and utility of CUBE data compared to traditional line-level EHR remains unquantified. We conducted evaluations comparing CUBE and line-level EHR using data from a cohort of patients with kidney transplant at the Boston Children’s Hospital. Our approach includes: (1) implement Bayesian inference to model hidden information in the CUBE, (2) evaluate the statistical fidelity between the original data and the CUBE, and (3) evaluate the utility of CUBE in kidney-transplant-related analyses relative to using the original data. The count inference pipeline outperformed baseline imputation strategies in recovering suppressed counts, achieving mean absolute error (MAE) of 0.12 in a high-dimensional, low-cardinality CUBE (10 variables, 57.1% of suppressed cells deterministically recovered) and MAE of 1.53 in a lower-dimensional, high-cardinality CUBE (3 variables). Our results from the fidelity evaluation demonstrated that the CUBE preserved marginal and joint distributions with consistently low Jensen-Shannon divergence and retains pairwise associations with low mean absolute difference of Cramér’s V against the original data. The results from the utility evaluation demonstrated that the CUBE reproduced conclusions consistent with the original data in descriptive statistics, odds ratios, and classification performance. These findings show that CUBE data retains sufficient statistical fidelity and utility for clinical research applications.

Yunyi Cheng

“Get **AHEAD** of Acquired Epilepsy: An **A**cute **H**ospitalization **E**pilepsy **A**ssessment and **D**etection Tool”

Thesis Advisor: Brandon M. Westover, MD, PhD

Abstract:

Preventing acquired epilepsy will require identifying high-risk patients during the latent window between acute brain injury and the first unprovoked seizure, yet clinicians still lack reliable tools to do so. Current prediction approaches rely on narrow clinical information and ignore both the heterogeneity of underlying diagnoses and the high competing risk of death in acutely ill populations. I developed and evaluated a multimodal prognostic framework, AHEAD (Acute Hospitalization Epilepsy risk Assessment and Detection), which integrates demographic and clinical variables, electroencephalogram (EEG) features, and structured neuroimaging features extracted from radiology reports. The analytic cohort comprised 6,315 adults admitted to the Mass General Brigham health system between 2014 and 2024 who underwent continuous EEG within 7 days of admission for an acute neurologic emergency. The primary outcome was acquired epilepsy, defined as ≥ 1 unprovoked seizure occurring more than 7 days after the index acute neurologic event. I ascertained outcomes using a staged longitudinal phenotyping framework that combined automated electronic health record (EHR) screening with manual chart adjudication, and estimated time-to-event risk with Fine-Gray subdistribution-hazard models that treated death as a competing event rather than as ordinary censoring.

Within 2 years, 549 patients (8.7%) developed acquired epilepsy, 1,882 (29.8%) died, 2,259 (35.8%) were censored, and 1,625 (25.7%) remained alive and event-free. In the full cohort, discrimination improved from a baseline concordance index (C-index) of 0.614 to 0.681 with EEG features and 0.691 with combined imaging and EEG features. EEG augmentation produced the clearest gains in encephalopathy-, hemorrhage-, and tumor-related presentations; imaging features alone gave smaller and less consistent improvement. Cumulative incidence analyses revealed substantial heterogeneity across diagnosis groups: tumor-associated presentations had the highest 2-year epilepsy incidence, while septic and metabolic presentations had the highest competing mortality.

Together, these results show that risk for acquired epilepsy after acute neurologic hospitalization is shaped by strong clinical heterogeneity and heavy competing mortality, and that EEG-derived features add clinically meaningful prognostic information beyond baseline covariates. Subgroup-aware, competing-risk multimodal modeling is a practical foundation for external validation, risk stratification, and prevention-oriented acquired-epilepsy research that can enrich future antiepileptogenesis trials.

Jiabo Cui

“Foundation Model Guided Flow Matching on Tree Manifold for Differentiable Phylogenetics”

Thesis Advisors: Marinka Zitnik, PhD, and Maha R. Farhat, MD, MSc

Abstract:

Phylogenetic inference aims to reconstruct evolutionary relationships among entities with phylogenetic trees. However, this task remains challenging due to the large combinatorial space of tree topologies and continuous branch lengths. Traditional statistical methods usually face trade-offs between accuracy and computational efficiency. Recent deep learning approaches improve efficiency but often rely on prealigning sequences and constrained topological spaces.

In this thesis, we propose a generative framework for phylogenetic inference that combines foundation-model sequence embeddings with structured modeling of tree space. We first introduce PhylaDiff, a diffusion-based model that represents trees as tokenized graphs and jointly generates topology and branch lengths. While PhylaDiff achieves strong reconstruction, we show that non-differentiable decoding limits its ability to incorporate likelihood-based objectives.

To address this, we develop PhylaFlow, a flow-matching model that formulates phylogenetic inference as a continuous transport problem in Billera–Holmes–Vogtmann (BHV) space. The model learns a vector field over tree space, enabling differentiable generation of both topology and branch lengths. Results show stable training, reduced Robinson–Foulds distance, and lower KL divergence to the reference posterior.

Overall, this work reframes phylogenetic inference as continuous generative modeling in a geometric space. The proposed framework improves efficiency and provides a new approach for modeling distributions over evolutionary trees.

Neha Dantuluri

“Development and Validation of an Interpretable and Longitudinal Risk Prediction Model for Stillbirth Using Machine Learning”

Thesis Advisors: Vesela P. Kovacheva, MD, PhD, and Arjun K. Manrai, PhD

Abstract:

Background: Stillbirth, defined as fetal death at ≥ 20 weeks gestation, affects approximately 1 in 175 deliveries in the United States. Disparities linked to social determinants of health (SDOH) contribute an estimated 60% excess risk in historically marginalized populations. Despite growing interest in applying machine learning (ML) to obstetric risk stratification, existing models are limited by retrospective single-site designs, absence of time-series features derived from longitudinal clinical measurements, and lack of external validation. This study developed and externally validated an explainable ML model for stillbirth prediction using features available by 20 weeks of gestation.

Methods: We analyzed 162,374 pregnancies (635 stillbirths; 0.4%) from the Mass General Brigham electronic health record (EHR) system from 2015–2025. Four models were evaluated: Logistic Regression, ElasticNet, XGBoost with random under-sampling (RUS), and Random Forest with RUS. Features included summary statistics derived from repeated prenatal vital sign measurements (blood pressure, BMI, weight trajectories), clinical comorbidities, parity, and SDOH variables including the Area Deprivation Index (ADI) and insurance type. Missing data were addressed using Multiple Imputation by Chained Equations (MICE). External validation was conducted on the nuMoM2b cohort (8,588 nulliparous pregnancies; 45 stillbirths). Model discrimination was compared via DeLong's test with Bonferroni correction; clinical utility was assessed by decision curve analysis (DCA).

Results: Random Forest + RUS demonstrated the best overall performance and was selected as the final model. Internal validation yielded AUROC 0.797 (95% CI: 0.764–0.832), sensitivity 68.9% (63.8–74.4%), and specificity 71.8% (70.2–73.8%). External validation on nuMoM2b confirmed aggregate-level generalizability with AUROC 0.789 (95% CI: 0.683–0.879), though subgroup performance was substantially attenuated due to structural data differences between cohorts. DCA demonstrated net clinical benefit across probability thresholds of 0.1–2%. SHAP analysis identified weight variability during pregnancy, prenatal visit frequency, and ADI rank as the top predictors.

Conclusions: We present an externally validated, EHR-based ML model for early stillbirth risk prediction that integrates SDOH, clinical, and time-series features to achieve clinically meaningful discrimination. The model demonstrated stable internal performance across racial, ethnic, and insurance-defined subgroups, though external subgroup performance was limited by cohort-specific data constraints. Post-hoc recalibration and prospective validation in diverse populations are necessary before clinical deployment.

Gaïa-Marie Gerbaka

“Cost-Effectiveness and Methodological Analysis of fMRI-Guided Transcranial Magnetic Simulation Biomarker Studies”

Thesis Advisor: Shan Siddiqi, MD

Abstract:

The field of precision psychiatry relies on biomarkers to diagnose disease and predict treatment success (Cappon and Pascual-Leone 2024). Transcranial magnetic stimulation (TMS) can be used to treat psychiatric symptoms by inducing neural activity in a targeted region, and modulating activity in downstream circuits. TMS targeting using functional magnetic resonance imaging (fMRI) enables precise spatial targeting and mapping of each patient’s brain connectivity. Associations between baseline functional connectivity and symptom improvement have been proposed as candidate predictive biomarkers, as they may allow physicians to anticipate treatment response. However, progress in this field remains limited by challenges in replication and validation, driven in part by heterogeneity in how brain connectivity is processed and analyzed. In this work, we investigate methodological heterogeneity in image processing parameters, including differences in regression, frequency filtering, and spatial smoothing, as well as TMS modeling choices. Using data from the Symptom-Specific targets for TMS (SSTMS) clinical trial (n = 35) (Taylor et al. 2026), connectivity between stimulation sites and anxiousomatic and dysphoric circuits was computed and related to clinical outcomes (BAI and BDI). In parallel, using the Three-D dataset (n = 300) (Blumberger et al. 2018), we use a downsampling with replacement approach across increasing sample sizes to model variance, precision, and proportion of statistically significant results. A Value of Information cost-effectiveness interface was developed to inform sample size selection based on cost and expected gains in precision and proportion of statistically significant results: <https://cost-effectiveness-dashboard.onrender.com>. Our results show that preprocessing choices, particularly whole-brain regression (Li et al. 2019) and filtering order, significantly influence connectivity estimates and their influence on biomarker assessment. In contrast, spatial smoothing and white matter regression had minimal impact on overall conclusions. Although we did not propose a standardized methodological protocol, these findings highlight the need for further systematic evaluation of preprocessing pipelines. Overall, this work emphasizes the importance of methodological choices in fMRI-guided TMS studies and provide practical tools for improving reproducibility and optimizing experimental design.

Surabhi Ghatti

“The Impact of Illicit Drug Use on Tuberculosis Transmission in South Africa”

Thesis Advisor: Maha R. Farhat, MD, MSc

Abstract:

Tuberculosis (TB) remains the leading cause of death by infectious disease globally. South Africa carries a disproportionately high burden of TB incidence, compounded by high rates of co-infection with HIV, poverty, crowding, unemployment, and undernutrition. The role of illicit drug use, specifically concurrent methamphetamine and mandrax smoking, in shaping TB transmission dynamics has not previously been studied using genomic approaches. Using whole- genome sequencing data from 462 culture-confirmed TB patients across two observational studies in Worcester, Western Cape, South Africa (TRUST and TOTAL), transmission clusters were constructed at a 25 SNP distance threshold and analyzed using the Bayesian transmission inference framework BREATH to reconstruct pairwise transmission probabilities. These posterior probabilities were used as outcomes in a multivariate regression analysis, including infector-level negative binomial and logistic models and a dyadic generalized estimating equation (GEE) model, incorporating behavioral concordance between infector-infectee pairs. Concurrent methamphetamine and mandrax use was independently associated with increased secondary infection counts at the infector levels (IRR = 1.55, 95% CI: 1.02-2.36), while the dyadic analysis revealed that concordance in polysubstance use was more likely than other non- concordance or no polysubstance use between infector and infectee pairs ($\beta=0.37$, 95% CI: -0.02 to 0.76). These findings identify illicit polysubstance users as a high-risk, hard-to-reach population contributing disproportionately to onward TB transmission and highlight the potential value of integrating TB screening and harm reduction strategies within drug-use networks to interrupt transmission in high-burden settings.

Rishabh Goel

“Generating Actionable Clinical Summaries from EHR Data of Pediatric Patients Using Large Language Models”

Thesis Advisor: Isaac S. Kohane, MD, PhD

Abstract:

Electronic health record (EHR) information overload is a well-documented driver of physician burnout, and pediatric settings face unique challenges including longitudinal developmental tracking, weight-based dosing, and age-specific reference ranges. This thesis presents a multi-agent large language model (LLM) framework for generating longitudinal clinical summaries from structured pediatric EHR data and evaluates these summaries against physician-authored summaries using the Provider Documentation Summarization Quality Instrument (PDSQI-9). The framework uses GPT-4.1 for summary generation and iterative improvement and Claude Opus 4.6 as an independent LLM-as-a-judge evaluator.

Using the Pediatric Physicians Organization at Children's (PPOC) dataset, we developed a three-component pipeline consisting of a generation agent, an LLM-as-a-judge evaluation agent, and an improvement agent that operate in an iterative refinement loop. We selected a cohort of 10 pediatric patients stratified by clinical complexity (five with growth disorders, three with cancer-related histories, and two clinically healthy) and collected 30 physician-authored summaries from three practicing pediatricians through a custom web-based labeling tool.

AI summaries refined through the feedback loop (AI+Feedback) achieved the highest mean PDSQI-9 score of 3.7 on a 5-point scale, significantly outperforming physician summaries. AI+Feedback scored higher than base AI summaries without feedback, though this difference was not statistically significant. AI+Feedback summaries prevailed on 8 of 10 patients and showed the largest quality gains for patients with cancer-related histories. However, AI+Feedback summaries were approximately 1.9 times longer than physician summaries and scored lower on the comprehensibility dimension. Exploratory analyses of temperature parameters revealed that the iterative refinement process is robust across a wide range of generation and improvement temperatures, with the feedback loop providing the greatest benefit when initial summary quality is low. These findings suggest that LLMs can produce pediatric clinical summaries that meet or exceed physician quality standards as measured by the PDSQI-9. However, the non-significant difference between AI+Feedback and base AI summaries, combined with the comprehensiveness-conciseness tradeoff, indicates that the iterative feedback loop's primary value may lie in regularizing quality for complex cases rather than uniformly improving all summaries.

Andrew B. Gumbert

“Investigating Molecularly Defined Clusters of Parkinson’s Disease Based on Multi-Omics Data With Clinical and Biological Implications”

Thesis Advisor: Bruno A. Benitez, MD

Abstract:

Introduction: Parkinson’s disease (PD) is the most prevalent neurodegenerative movement disorder, involving motor and cognitive symptoms that are heterogeneous across individuals. Current clinical tests for assessing PD often miss important information about molecular and biological heterogeneity in the disease. To address this problem, we sought to answer the question of whether molecular data could provide generalizable clusters of PD with clinical and biological implications. We hypothesized that the molecular data would provide more informative classifications of PD relative to clinical tests alone.

Methods: We performed a clustering analysis with bulk transcriptomic data from whole blood and proteomic data from urine and cerebrospinal fluid from the Parkinson’s Progression Markers Initiative. In this analysis, we used relevant features that correlated with longitudinal PD severity. We analyzed the resulting clusters for clinical and biological implications, considering the Unified PD Rating Scale (UPDRS), the Montreal Cognitive Assessment, dopamine transporter scans, seed amplification assays, gene set enrichment, and cell-type deconvolution. To assess generalizability, we created logistic regression and k-nearest neighbor classification models, using transcriptomic data across the original cohort and from the Parkinson’s Disease Biomarkers Program to predict cluster groupings.

Results: The clusters from the analysis displayed robust differences in terms of UPDRS scores, especially regarding motor symptoms. Gene set enrichment and deconvolution analyses revealed differences in immune system activity by cluster. The classification models across cohorts suggested generalizable implications in terms of the UPDRS, supporting the applicability of the results beyond the initial cohort.

Conclusions: The correlations between clusters, clinical outcomes, and immune system activity support our hypothesis that molecular data can provide more informative classifications of PD relative to clinical tests alone. To build on these findings, future research can continue to investigate clinical, biological, and practical implications of clusters, potentially with other selections of omics data or in additional cohorts.

Rahul Gundamraj

“Discovery of Novel Rare Disease Genes Through Scalable Prioritization and Geometric Scoring of AlphaFold3 Structures”

Thesis Advisors: Piotr Sliz, PhD, and Shira Rockowitz, PhD

Abstract:

Rare diseases collectively represent a significant global health burden, with approximately 80% possessing a suspected genomic basis. Despite advances in sequencing, many patients remain without a molecular diagnosis, often due to the difficulty of functionalizing variants of uncertain significance (VUS) and thus implicating new undiscovered disease genes. This thesis presents an end-to-end framework designed to bridge the gap between clinical genomic variation and biophysical mechanisms by using protein-protein interactions as a way to implicate candidate disease genes. Utilizing data from the Children’s Rare Disease Collaborative at Boston Children’s Hospital, we used an existing interactome as well as agentic architectures as independent methods for the prioritization of candidate interactions for downstream structural prediction. Agentic platform ToolUniverse maintained 98.8% coverage across datasets of up to 25,000 queried interactions, successfully recovering known biological complexes within the top 200 ranked results, while effectively excluding all non-interacting pairs. To refine these prioritized lists, we developed the Interface Confidence Score (ICS), an XGBoost-based regressor trained on AlphaFold 3 (AF3) features. On a held-out test set, the ICS achieved a Spearman rank correlation of 0.74 and an accuracy of 95.9% when tasked with separating binding/non-binding proteins. Comparative analysis revealed that the ICS significantly outperformed ipTM in separation power, correctly rejecting 88.9% of “hallucinated” high confidence AF3 interfaces. When applied to 989 Predictomes structures, the ICS identified critical unique hits that were discarded by the previously established Structured Prediction Omics Classifier threshold. A high resolution case study of the MCM2-7 complex further validated the ICS’s ability to provide granular, subcomplex stability rankings that aligned with known hexameric assembly mechanics. Structural and transcriptomic validation of our lead candidates provided a mechanistic rationale for their associated phenotypes. Collectively, this work demonstrates that integrating agentic reasoning with geometrically grounded structural scoring provides a transformative roadmap for resolving the molecular basis of rare pediatric disorders and closing the diagnostic gap.

Ningxi Huang

“Characterizing Cancer Phenotypes and Their Tumor Microenvironment Across Malignant Progression Using Single-Cell RNA Sequencing Data”

Thesis Advisor: Felix Dietlein, MD, PhD

Abstract:

Sarcoma is a rare group of cancers originating from mesenchymal lineages, with high metastatic potential and poor clinical outcomes. Previous studies have linked collagen and extracellular matrix (ECM) remodeling to sarcoma invasion, metastasis, and immune evasion. However, with bulk RNA-seq datasets, it remains unclear whether these signals come from tumor cells themselves or from the tumor microenvironment (TME). In this study, I designed and implemented a single-cell RNA sequencing (scRNA-seq) analysis pipeline to address this question, utilizing publicly available datasets from three sarcoma subtypes: undifferentiated pleomorphic sarcoma (UPS), leiomyosarcoma (LMS), and dedifferentiated liposarcoma (DDLPS).

The designed pipeline identifies malignant cells without relying on marker genes by using an iterative inferCNV reference refinement strategy. The pipeline achieved good annotation concordance with original study labels in UPS/LMS samples, with reduced concordance in liposarcoma samples, potentially attributable to the focal copy number architecture characteristic of DDLPS and substantial inter-patient transcriptional heterogeneity.

Downstream differential gene expression analysis showed that major fibrillar collagen genes, including COL1A1, COL1A2, and COL3A1, were upregulated in malignant cells compared with non-malignant cells across all analyzed datasets, while other ECM-related genes and collagen subtypes followed subtype- or patient-specific patterns. Collagen biosynthesis and ECM assembly pathways were upregulated in malignant cells, while Hallmark oncogenic pathways showed patient-specific patterns.

In conclusion, these results support tumor-intrinsic collagen expression as a consistent feature of sarcoma malignant cell states. The study may have important implications for malignant progression, immune evasion, and the development of matrix-targeted therapeutic strategies.

Charis Li

“Learning and Benchmarking Temporal Representations in Claims Data: Telehealth Trajectories in Buprenorphine Treatment for Opioid Use Disorder”

Thesis Advisor: Mohammad Jalali, PhD

Abstract:

The growing availability of longitudinal administrative claims data has created unprecedented opportunities to study healthcare utilization, treatment engagement, and patient trajectories in real-world settings. Yet despite this abundance of real-world data, extracting meaningful temporal insights remains challenging because claims records are sparse, irregularly timed, and highly sensitive to operational definitions. These challenges are particularly relevant when studying clinical patterns during buprenorphine treatment for opioid use disorder (OUD), where healthcare interactions are recorded as sparse longitudinal events, yet it remains unclear how different temporal representation strategies characterize these patterns and whether they identify meaningful utilization phenotypes.

To address this gap, this work uses telehealth utilization during buprenorphine treatment for OUD as a case study to evaluate how different temporal representation strategies characterize sparse longitudinal claims data. The primary objective was to compare whether rule-based definitions, deterministic timing features, linear embeddings, latent trajectory models, and recurrent neural network embeddings recover distinct telehealth-use phenotypes beyond conventional early-versus-late exposure definitions. A secondary objective was to assess whether these phenotypes were robust to alternative buprenorphine episode definitions.

Buprenorphine treatment episodes were constructed from prescription claims using allowable refill gaps of 7, 14, 30, and 60 days. Telehealth encounters were linked to treatment episodes and analyzed primarily within the COVID-19-era study window. Five approaches were compared: binary and early-versus-late telehealth definitions, normalized timing embeddings, sparse matrix embeddings using truncated singular value decomposition, latent class growth analysis (LCGA), and gated recurrent unit (GRU) autoencoder embeddings. Methods were benchmarked using internal clustering metrics, bootstrap stability, trajectory diagnostics, early-threshold sensitivity analyses, and cross-gap reassignment patterns.

Across methods, telehealth trajectories were sparse and heterogeneous. Simple early-versus-late definitions mainly captured time to first telehealth use and could not distinguish transient early use from sustained longitudinal engagement. SVD and normalized embeddings provided limited additional resolution. LCGA identified four interpretable timing–intensity phenotypes and showed the strongest cross-gap preservation, while GRU embeddings achieved stronger quantitative clustering stability but largely collapsed trajectories into broader low- versus high-utilization patterns. Overall, this case study shows that when studying sparse, irregular, and definition-sensitive claims data, temporal representation choice materially shapes the resulting phenotype space. More complex models do not automatically produce more

clinically interpretable subgroups; instead, benchmarking should explicitly evaluate stability, interpretability, robustness to episode construction, and the specific behavioral dimensions each method captures.

Jiatong Li

“Toward a Framework for Evaluating the Clinical Relevance of Cancer Cell Lines at Single-Cell Resolution”

Thesis Advisors: Srivastan Raghavan, MD, PhD, and Peter Winter, PhD

Abstract:

Human cancer cell lines (CCLs) play a critical role in propagating our understanding of cancer biology and enabling preclinical testing of novel anti-cancer compounds. Previous studies have shown how myriad factors (e.g., media formulation, genetic and transcriptomic evolution) can compromise how faithfully CCLs represent in vivo phenotypes. Prior computational strategies have compared CCLs and TCGA bulk RNA-seq expression profiles, but the relationship between CCLs and patient tumors at single-cell resolution remains poorly characterized. To this end, we established a robust benchmarking framework that compares the effectiveness of diverse embedding approaches at relating CCLs to patient single-cell RNA-seq (scRNA-seq) profiles. To facilitate robust benchmarking, we designed a suite of control experiments, standardized preprocessing pipelines, and adapted metrics to evaluate mapping performance for each benchmarked method. We benchmarked seven methods: principal component analysis (PCA), canonical correlation analysis (CCA), Gene program, Celligner, Symphony, expiMap, and Geneformer. We observed that most methods excelled at tasks in simpler experiments, but showed reduced accuracy for certain lineages when mapping CCLs to their corresponding lineage in patient samples. We identified CCA as a promising candidate for integrating large-scale cross-system mapping of CCLs and tumor datasets. In sum, we have taken initial steps toward a framework to compare preclinical cancer models and patient tumor samples at single-cell resolution. Ultimately, with further development, this framework will guide the selection of preclinical models for specific clinical contexts.

John Lin

“Non-Invasive Identification of Clinically Actionable Gene Fusions in Gliomas Using Deep Learning-Enabled Analysis of Magnetic Resonance Imaging Scans”

Thesis Advisor: Albert Kim, MD, and Christopher Bridge, DPhil

Abstract:

Gliomas are the most common primary malignant brain tumors in adults, with prognosis varying substantially across molecular subtypes. Oncogenic gene fusions have recently emerged as a promising class of actionable therapeutic targets, yet their detection relies on invasive tissue sampling via biopsy or surgical resection. Here we present a deep learning framework that leverages magnetic resonance imaging (MRI) to non-invasively predict gene fusion status in gliomas. Using contrastive self-supervised learning, we pretrained a foundation model on a diverse cohort of 30,270 brain MRI images from 10,336 patients across multiple institutions and tumor types. Validation on brain tumor type classification showed that the model learned meaningful representations (AUC: 0.84), outperforming existing foundation models. Application to gene fusion status classification, however, yielded more limited discrimination (AUC: 0.54), suggesting that gene fusions may manifest as comparatively subtle or heterogeneous MRI patterns. Our findings demonstrate that foundation models can learn transferable representations across diverse neuroimaging tasks, although non-invasive identification of gene fusions in gliomas remains an open challenge.

Sunni Chinjo Lin

“Building a Generalized Decoder for Immune Profiling in Pathology”

Thesis Advisor: Kun-Hsing Yu, MD, PhD

Abstract:

Accurate immune profiling of the tumor microenvironment (TME) requires integrating nuclei, tissue, and spatial-level information across sequential reasoning steps. Recent advances in vision language models (VLMs) have enabled natural language interaction with histopathology images, yet existing pathology VLMs are predominantly limited to single-turn question-answering and lack the structured, multi-step reasoning required for clinical immune profiling. We present a dataset construction, fine-tuning, and evaluation framework for multi-turn visual question answering (VQA), applied to computational immune profiling of the TME. Our multi-turn conversational pipeline decomposes immune profiling into sequential, clinically-grounded subtasks: nucleus detection and classification, cell-type quantification, tissue region analysis, and TME summarization. We introduce a structured approach to generate multi-turn VQA training data from annotated histopathology images. Our dataset comprises 238,747 patches spanning 22 organ types from 7 datasets, including hematoxylin and eosin (H&E) and immunohistochemistry (IHC) slides, from which we construct 515,382 multi-turn conversation samples across 13 turn types that span pixel-level and patch-level spatial scales. We further introduce a modular, turn-aware evaluation framework that assesses turn-level accuracy, cross-turn consistency, and conversational coherence. To our knowledge, this is the first such framework for multi-turn histopathology VQA. On an external test set, our fine-tuned model achieves a median nuclei detection F1 of 0.667, outperforming HoverNet on nuclei detection and segmentation. On a held-out test set, it surpasses all four VLM baselines across multi-turn evaluation metrics. This work establishes the feasibility of fine-tuning a general-purpose VLM for structured immune profiling. We envision this framework as a generalizable basis for developing and benchmarking conversational VLMs across immune profiling tasks and broader digital pathology applications.

Seshagiri Rao Mallina

“Early Detection of Pediatric Growth Disorders: A Deep Learning Approach to Longitudinal Electronic Health Records”

Thesis Advisor: Isaac Kohane, MD, PhD

Abstract:

Pediatric growth abnormalities serve as clinical indicators of a broad spectrum of endocrine, genetic, and gastrointestinal conditions, many of which are treatable when identified early, but carry significant long-term consequences when diagnosis is delayed. Despite advances in screening protocols, diagnostic delays remain prevalent, driven by the longitudinal nature of growth-trajectory deviations and cognitive biases. This study developed a transformer classification model, evaluated it across performance metrics, and analyzed how the model reached its predictions for the early detection of growth-related conditions across 24 diagnoses derived from the International Classification of Pediatric Endocrine Diagnosis (ICPED), using electronic health record data from 9,747 pediatric patients drawn from a pediatric primary care dataset. A central aim was to determine whether strong discriminative ability between patients with growth abnormalities and without could be achieved, and, if so, whether the model's reasoning patterns aligned with clinically valid reasoning.

First, a time-to-event analysis was conducted to establish patient eligibility criteria. Next, transformer and logistic regression models were trained and evaluated across multiple feature configurations. Following optimization, the best-performing model achieved a test AUROC of 0.9823. The interpretability analyses revealed a reasoning mechanism inconsistent with clinical norms. The model operated under a guilty-until-proven-innocent heuristic, assigning elevated predicted risk to all patients by default and reducing that risk after accumulating evidence of normal growth and routine healthcare engagement. Routine encounter codes dominated attribution rankings. The results indicated that performance reflected healthcare utilization patterns more than underlying pathology. Lead time calculation quantified how far in advance of formal diagnosis the model identified at-risk patients, yielding a bootstrapped median estimate of 64.10 months. Given the model's inverted clinical reasoning, these estimates are better interpreted as upper bounds on true pre-diagnostic capabilities rather than as reliable estimates of clinical utility.

These findings demonstrate that transformer architectures can perform classification on longitudinal EHR data, while highlighting that high predictive performance is insufficient evidence of clinical validity. The interpretability analysis developed here offers a replicable methodology for auditing EHR-based diagnostic models before clinical deployment.

Ananya Prasad

“Divergent Effects of KLF1 Variants: Gain and Loss of DNA Binding Determine the Severity of Erythropoietic Disruption”

Thesis Advisor: Martha L. Bulyk, PhD

Abstract:

Mutations in transcription factor (TF) DNA-binding domains (DBDs) have been implicated in a wide range of human diseases. One such TF, KLF1, plays a central role in erythropoiesis, the process by which hematopoietic stem cells differentiate into erythrocytes (red blood cells). Mutations in the KLF1 DBD are associated with a spectrum of blood disorders, ranging from relatively benign conditions, such as the Lutheran inhibitor blood type and hereditary persistence of fetal hemoglobin, to severe hemolytic anemias, including congenital dyserythropoietic anemia type IV. Although specific KLF1 variants have been linked to distinct clinical phenotypes, the extent to which these mutations differentially reshape KLF1 genomic occupancy and downstream transcriptional programs remains unclear. Addressing this gap requires an integrative analysis of how variant-specific alterations in DNA binding influence gene regulatory outcomes. This study investigates five clinically diverse KLF1 DBD variants (p.A298P, p.H299Y, p.R328H, p.E325D, and p.E325K) to elucidate the molecular mechanisms underlying disrupted erythropoiesis. Combined analyses of genomic occupancy, DNA sequence affinity, gene expression, and chromatin state reveal distinct mechanistic modes of variant KLF1 dysfunction: gain- and loss-of-function. Variants associated with milder clinical phenotypes (p.A298P, p.H299Y, and p.R328H) primarily reduce DNA-binding affinity, resulting in decreased occupancy at regulatory elements and underexpression of genes essential for erythrocyte development and differentiation. In contrast, variants linked to more severe phenotypes (p.E325D and p.E325K) modify DNA-binding affinity and specificity, enabling disruption of canonical KLF1 regulatory networks and ectopic occupancy at novel genomic sites that drive transcriptional programs opposing normal erythropoiesis. Collectively, these results establish a framework for understanding the mechanisms of KLF1 DBD variant perturbations, and clarify how changes in DNA binding, chromatin state, and gene expression drive diverse blood disorders.

Siavash Raissi

“From Hidden Activations to Human-Interpretable Concepts: Explaining and Interpreting Computational Pathology Models Using Concept Learning”

Thesis Advisors: Kun-Hsing Yu, MD, PhD, and Shuji Ogino, MD, PhD

Abstract:

Modern computational pathology models, particularly those based on foundation models and multiple-instance learning (MIL), achieve strong performance across diagnostic tasks but remain difficult to interpret due to their black-box nature. This lack of transparency limits their adoption in clinical workflows, where understanding the basis of a model's prediction is essential for trust and validation. This thesis investigates whether human-interpretable nuclear morphology concepts can be recovered from the hidden-layer activations of MIL-based pathology classifiers and used to improve explainability without constraining model performance. In Aim 1, a proof-of-concept auxiliary concept model (CM) was developed to predict discrete nuclear morphology concepts from the hidden activations of a CHIEF-based LUAD/LUSC classifier trained on TCGA whole-slide images. Results demonstrate that concept signals are recoverable from intermediate representations, with higher recoverability in earlier attention layers and selective preservation of interpretable features such as nuclear size and shape. However, discretization introduced calibration limitations, restricting the CM's ability to capture the continuous structure of nuclear morphologies. Aim 2 extends this framework by modeling concepts as continuous variables and benchmarking concept recoverability across MIL classifiers trained using four pathology foundation models: CHIEF, UNI, Virchow2, and GigaPath. A series of experiments evaluated the impact of CM architecture, loss functions, and concept-space designs using mean absolute error (MAE) and its variability across concepts. Nonlinear CMs improved performance over linear baselines, while concept clustering reduced bias toward highly correlated features. Across all settings, UNI consistently achieved the lowest MAE and variability, indicating stronger preservation of morphology-related information. Qualitative analyses further showed that predicted concepts are biologically grounded and spatially aligned with model attention, providing insight into both where and what features drive predictions. Overall, this work demonstrates the promise of learning concepts from hidden activations to enhance the interpretability and explainability of computational pathology models.

Jiarong Nicole Ye

“Domain Adaptation for Breast Cancer Computational Pathology: Evaluating Feature-Level and Pixel-Level Approaches Under Leave-One-Domain-Out Framework”

Thesis Advisors: Christopher Bridge, DPhil, and Albert Kim, MD

Abstract:

Deep learning models for computational pathology suffer substantial performance degradation when deployed across institutions with different scanners, staining protocols, and patient populations. This thesis investigates whether domain adaptation can improve cross-institutional performance of a multiple instance learning classifier for estrogen receptor (ER) status prediction in breast cancer, evaluated across four geographically diverse whole-slide image datasets (TCGA-BRCA, Zanmi Lasante Haiti, BCNB, and PostNAT-BRCA) under a Leave-One-Domain-Out framework.

Three aims were pursued. Aim 1 established a CLAM-CONCH baseline using a frozen foundation model encoder. The baseline achieves held-out AUC exceeding 0.76 across all four folds, with cross-domain gaps of 13–17 percentage points for the three large-cohort folds. Feature space analysis confirms strong domain-driven clustering despite histopathology-specific pretraining, motivating explicit adaptation. Aim 2 evaluated feature-level adaptation via Deep CORAL, integrated as a covariance alignment loss within CLAM training. Slide-level CORAL produced the largest improvement for TCGA (+4.54 pp), the domain with the greatest distributional shift, while providing marginal gains for Haiti and BCNB. Tile-level CORAL failed across all domains, attributable to the architectural bottleneck of a single linear layer between the frozen encoder and the attention pooling mechanism.

Aim 3 evaluated pixel-level adaptation via CyCADA: a CycleGAN augmented with a semantic consistency loss using frozen CONCH features to constrain content preservation during translation. CycleGAN-only adaptation recovered approximately 51% of the TCGA domain gap (+8.59 pp). Adding semantic consistency achieved the best result for BCNB (+6.68 pp) and was the only pixel-level variant to improve on baseline for Haiti (+0.62 pp), preventing the degradation produced by unconstrained translation.

The central finding is that adaptation effectiveness depends on the interaction between method and domain shift type. Pixel-level style transfer suits high-resource domains with large appearance-driven shift; feature-level alignment suits distributional shift; and neither approach adequately addresses acquisition-process degradation of diagnostic signal quality, as observed in Haiti. Haiti's resistance to all tested methods identifies a failure mode that standard benchmarks conceal and motivates adaptation strategies designed for resource-constrained clinical environments.

Wenzhu Ye

“Bridging the Resolution Gap: Super-Resolution Enhancement of Panoramic Radiographs to Periapical-Like Clarity Using Deep Learning”

Thesis Advisors: Juan Eugenio Iglesias Gonzalez, PhD, and Roya Zandparsa, DDS, MSc, DMD, FICD

Abstract:

Panoramic radiographs (PANs) are widely used in dental practice for their broad coverage of the maxillofacial region, low cost, and clinical convenience. However, compared with periapical radiographs (PAs), PANs offer lower spatial resolution and reduced visibility of fine anatomical structures, including the lamina dura, periodontal ligament space, root morphology, and subtle periapical changes, which can limit their diagnostic utility for precision-dependent evaluations. Developing a super-resolution (SR) and enhancement method to enhance PAN images toward PA-like anatomical clarity could ultimately aid in early detection and prevention of potential diseases and reduce the need for additional intraoral imaging.

A major challenge in developing an SR solution to align the level of detail in PAN to the PA standard is the lack of paired PAN and PA images that depict the same anatomy at their own levels of detail. A two-step method was designed to solve this problem. First, the image style translation model CycleGAN was used to transform HR PA images into synthetic PAN-style counterparts that simulate panoramic-like degradation in contrast, noise, and detail level. These generated images were paired with their original PA to create supervised training data for SR training. Next, two SR architectures, SwinIR and NAFNet, were then trained and evaluated using quantitative image quality metrics, including peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), feature similarity index (FSIM), and mean squared error (MSE).

Both models demonstrated strong reconstruction performance and successfully recovered finer anatomical structures from degraded PAN-styled inputs. NAFNet achieved better overall performance and produced outputs that have both a quantitatively higher score and a closer visual appearance to the ground truth PA images. These results suggest that the proposed framework is a viable approach for enhancing dental radiographic detail and improving anatomical visibility in PANs.

Although further validation on real PAN images with clinically verified reference labels is still needed, this work represents an important first step toward enhancing the anatomical detail visible in panoramic radiographs. With future clinical validation and expert assessment, this framework may help improve PAN-based screening and expand the diagnostic value of routinely acquired dental imaging.

Xingli Yu

“Investigating Antigen Recognition Patterns and T Cell Dynamics in Neoantigen Vaccine-Treated Chronic Lymphocytic Leukemia”

Thesis Advisors: Catherine Wu, MD, and Gad Getz, PhD

Abstract:

Cancer vaccination in chronic lymphocytic leukemia (CLL) remains challenging because CLL has an ultra-low tumor mutational burden and a profoundly immunosuppressive microenvironment, limiting both targetable neoantigens and anti-tumor T cell function. To investigate how vaccination shapes tumor evolution and immune remodeling in this setting, we performed longitudinal genomic and single-cell immune profiling across two complementary CLL vaccine cohorts: NeoVax, a personalized neoantigen vaccine trial in treatment-naïve high-risk CLL, and GVAX, an autologous whole-tumor cell vaccine administered after allogeneic hematopoietic stem cell transplantation.

Using whole-exome sequencing with PhylogicNDT-based clonal reconstruction, we tracked tumor clonal dynamics over time. In the NeoVax cohort, vaccine-targeted mutations spanned both clonal and subclonal tumor populations, meaning immune pressure against these targets would be expected to act broadly rather than selectively. Across most patients, clonal architecture remained stable during and after vaccination, and immunogenic vaccine-targeted mutations did not show consistent selective contraction. However, one progressing patient showed expansion of a non-immunogenic driver-harboring lineage, suggesting that intrinsic driver-mediated fitness can overcome immune pressure against immunogenic neoantigens. In the GVAX cohort, clonal analysis identified relapse-associated expansion of a TP53-mutant subclone with accompanying allele-specific copy-number changes, consistent with a strong fitness advantage at relapse.

Single-cell RNA sequencing of longitudinal bone marrow samples from six GVAX-treated patients showed that long-term remission was associated with sustained expansion of cytotoxic GZMB+ CD8+ T cell states after vaccination, whereas relapse patients lacked consistent effector enrichment.

Results show that clonal dynamics alone may not fully explain vaccine response in CLL and support integrating tumor evolution with longitudinal immune-state changes to identify mechanisms of response and resistance in low-mutational-burden malignancies.