O’Donnell Data Science and Research Computing Institute Newsletter

AI/ML and High Performance Computing | Fall 2024

Dear Colleagues,

I am happy to share the Fall 2024 O'Donnell Data Science and Research Computing Institute (ODSRCI) newsletter with you. This newsletter highlights the AI/ML research ongoing at SMU utilizing our HPC platforms. Our NVIDIA DGX SuperPOD and the ManeFrame III cluster are great resources for SMU researchers developing and training large-scale AI/ML models.

Fall 2024 was a very active semester for the ODSRCI. We provided training sessions covering the fundamentals of HPC, how to write parallel code using C++, and the concept of federated learning. In September, we organized a special workshop on Digital Twins sponsored by NVIDIA Corporation and Mark III Systems. Additionally, the ODSRCI seminar series featured presentations by our faculty discussing how AI/ML is incorporated in their research projects. Slides for these presentations can be found here. The ODSRCI recently awarded Faculty Research Acceleration Grants to three AI-focused HPC projects. Drs. Eric Larson, Elfie Kraka, and Heather DeShon are winners of these grants. I also want to congratulate the latest cohort of ODSRCI's graduate fellows: Yinglu Tan, Eli Laird, Kang Liang, Mark De Vernon, and Ding Lin. These five SMU graduate students won the ODSRCI fellowship by submitting competitive proposals in advanced AI/ML technologies, digital twins, and quantum information sciences.

I expect to announce additional research grant opportunities in the near future. Please stay tuned for updates.

Dr. Neena Imam
Peter O'Donnell Jr. Director
O'Donnell Data Science & Research Computing Institute

CONTENTS

LEVERAGING AI AGENTS FOR PERSONALIZED EDUCATION FOR STUDENTS WITH READING DISABILITIES

Dr. Corey Clark, SMU Guildhall

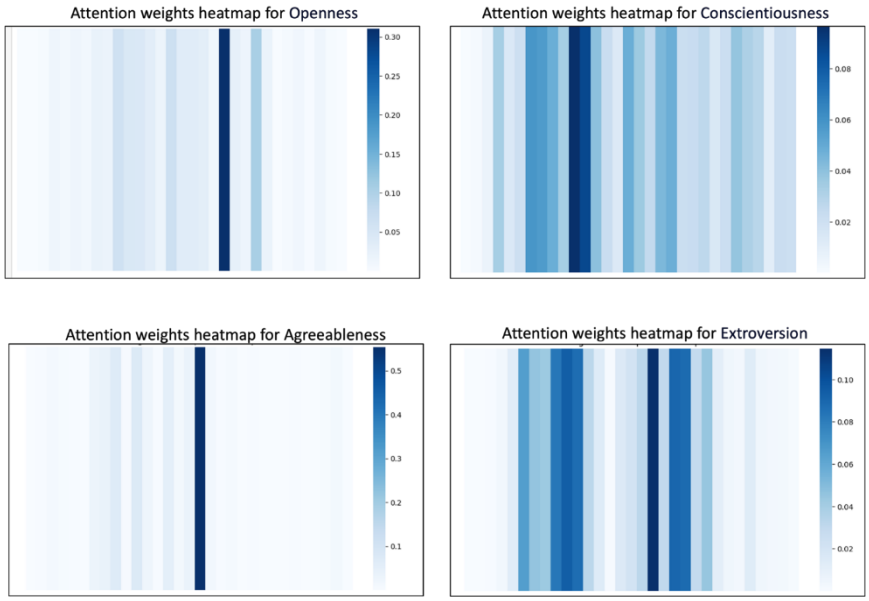
When appropriate learning materials are accessible, students with reading disabilities can effectively demonstrate their understanding in ways comparable to their peers. AI agents, powered by Large Language Models (LLMs), present a unique solution for providing universal access to students with reading disabilities by offering personalized learning experiences tailored to individual needs. This personalization enhances learning without compromising the generalizability of the AI agents' design, allowing a single agent design to adapt to diverse scenarios. Dr. Clark's research aims to advance AI agent design and deploy these agents in middle school classrooms to promote equity and better prepare students for technology-driven careers. These AI agents will also support teachers by offering data-driven insights into student progress, making it easier to track and address diverse learning needs. This research integrates game-based AI agents–driven by LLMs–into an educational version of Minecraft to personalize learning experiences, discover optimal metrics for supporting students with reading disabilities, and provide teachers with summaries and analytics on student progress. The AI agents are divided into two types: Personalized Education Agents and Teacher Support Agents. Integrated into Minecraft, these agents assist students through personalized feedback and game scenarios tailored to their reading needs, while helping teachers understand student progress through accessible summaries and visualizations.

ENHANCED PERSONALITY ASSESSMENT WITH NLP IN PSYCHOLOGICAL RESEARCH

Dr. Mehak Gupta, Computer Science

Dr. Mehak Gupta's research team has made a significant breakthrough in harnessing the power of Artificial Intelligence (AI) for psychological assessments, leveraging cutting-edge Natural Language Processing (NLP) techniques to enhance the evaluation of personality traits. Traditional methods rely on closed-ended questionnaires, but NLP offers a more nuanced approach by analyzing language-based evaluations. However, LLMs are limited by their token processing capacity, restricting the length of text they can effectively contextualize. To overcome this challenge, Dr. Gupta's team developed a hierarchical transformer model that combines the strengths of RoBERTa and Recurrent Neural Networks (RNNs) with attention layers. This innovative approach enables the efficient processing of lengthy life narrative texts, over 2000 tokens, to assess the Big Five personality traits in older adults. By fine-tuning RoBERTa with a sliding-window approach and processing contextual embeddings through RNNs, the team achieved impressive predictive accuracy, with R^2 values **(cont.)**

ranging from 0.30 to 0.56 for different personality traits. Furthermore, the attention layers in the RNN provided valuable insights into the decision-making process of the model, highlighting specific parts of a participant's life narrative that inform personality trait predictions. Domain experts verified the model's interpretations, supporting its validity and paving the way for significant advancements in psychological research and practice.



HARNESSING HPC FOR MACHINE LEARNING APPLICATIONS IN HEALTHCARE AND BEYOND

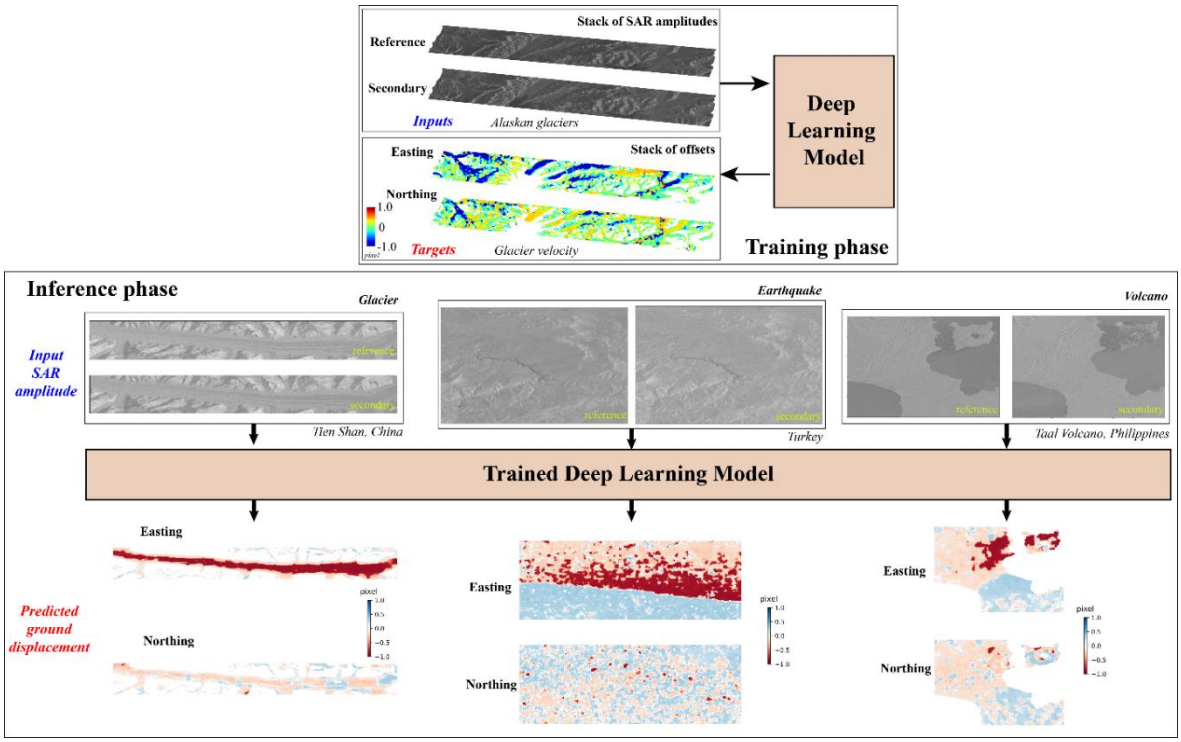
Dr. Eric Larson, Computer Science; Dr. Mehak Gupta, Computer Science

Dr. Eric Larson, a leading researcher in the interdisciplinary application of machine learning and signal/image processing, recently highlighted the pivotal role of SMU's High-Performance Computing (HPC) systems in his research endeavors at a seminar hosted by the O'Donnell Data Science & Research Computing Institute. His work, which spans fields such as security, health, education, and ubiquitous computing, often involves complex computational tasks that demand significant processing power - a need that is met by SMU's advanced HPC resources, including the NVIDIA DGX SuperPOD and M3 HPC systems. These systems have been instrumental in enabling Dr. Larson to perform critical tasks such as ablation studies and hyperparameter searches, which would be impossible without the support of SMU's data science specialists and HPC capabilities. Currently, Dr. Larson is collaborating with Dr. Mehak Gupta on a project that aims to harness machine learning to analyze medical data responsibly, ensuring patient privacy and ethical considerations while extracting valuable insights to improve healthcare outcomes. With the support of the O'Donnell Data Science & Research Computing Institute, Dr. Larson is able to process large datasets efficiently and explore innovative solutions that would otherwise be unattainable, underscoring the institute's crucial role in driving groundbreaking research across various scientific fields.

DEEP LEARNING FOR GROUND SURFACE DISPLACEMENT MEASUREMENT USING SAR IMAGERY

Dr. Jin Woo Kim, Earth Sciences; Dr. Zhong Lu, Geophysics

A team of researchers led by Drs. Jin Woo Kim and Zhong Lu, in collaboration with Dr. Hyung-Sup Jung from the University of Seoul, has developed a groundbreaking deep learning model that accurately measures ground surface displacement from Synthetic Aperture Radar (SAR) imagery. This innovative approach combines multi-kernel offset tracking for preprocessing with a sophisticated architecture featuring U-Net, cross-connections, residual blocks, and attention mechanisms to predict displacement between pairs of SAR amplitude images. Trained on 200,000 pairs of reference and secondary SAR images from Alaska's glaciers using SMU's powerful SuperPOD, equipped with NVIDIA A100 GPUs, the model demonstrated exceptional performance in estimating ground displacement. The team successfully applied the model to various real-world scenarios, including glaciers from different regions, earthquakes in Turkey and California, and volcanic activity in the Philippines and Hawaii. With continued support from SMU O'Donnell Data Science and Research Computing Institute on the SuperPOD, the researchers plan to scale up their model training with larger, more diverse datasets from around the world, further enhancing its prediction accuracy and paving the way for significant advancements in fields such as geology, seismology, and volcanology.

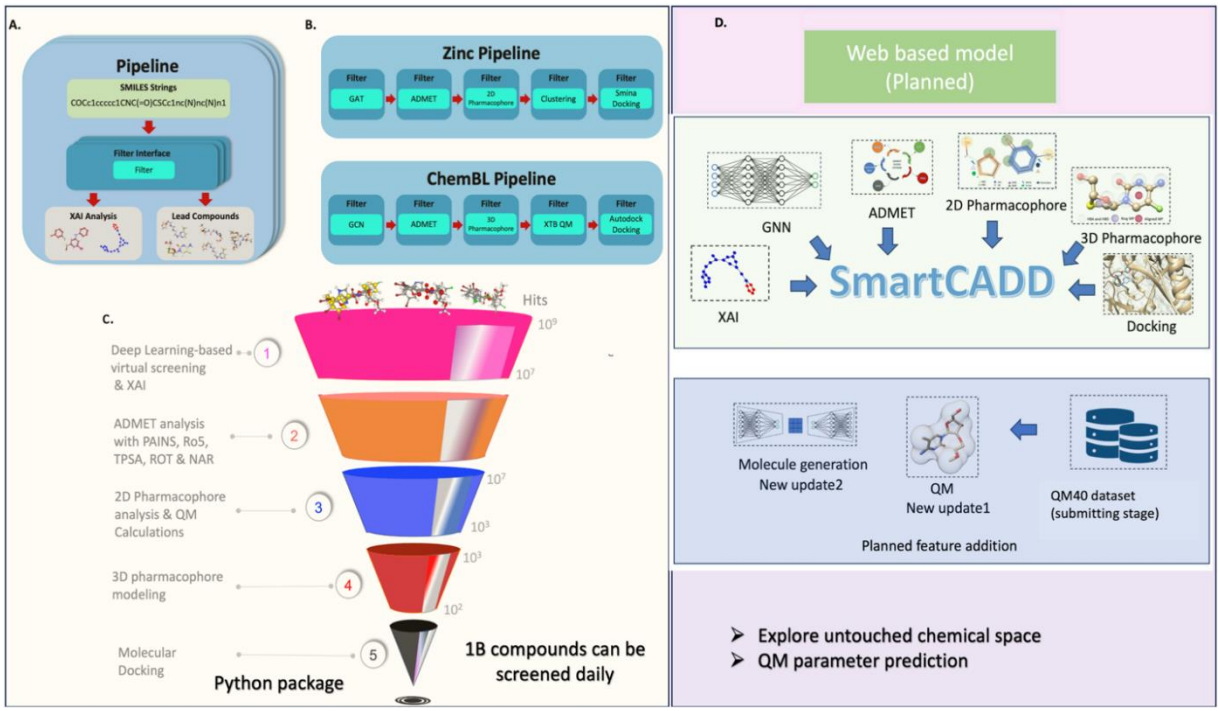


AI AND QUANTUM MECHANICS IN DRUG DISCOVERY AND CHEMICAL RESEARCH

Dr. Elfi Kraka, Chemistry; Dr. Corey Clark, SMU Guildhall

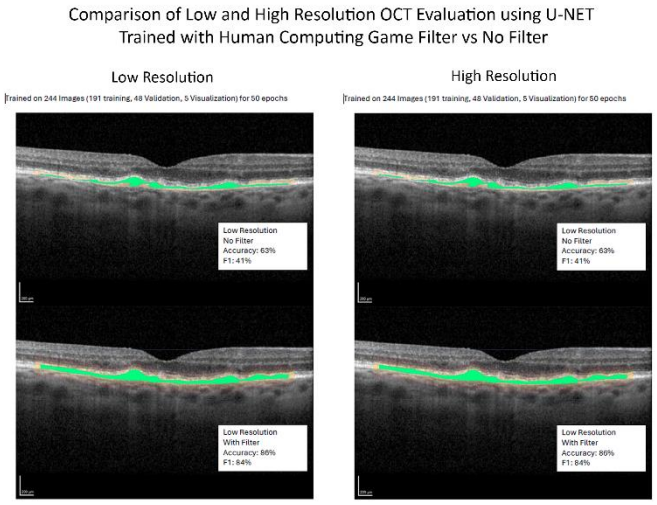
The Computational and Theoretical Chemistry Group (CATCO) at SMU, led by Dr. Elfi Kraka, is dedicated to developing innovative quantum chemical tools that can accelerate chemical discovery processes, reduce the need for costly and dangerous experiments, and provide insights not accessible through traditional methods. In addition to advancing scientific knowledge, CATCO is committed to educating and mentoring graduate students and postdoctoral researchers to become the next generation of leaders in their field. Recently, CATCO has joined forces with Dr. Corey Clark's Guildhall Human and Machine Intelligence (HuMIn) Game Lab to develop a novel drug design platform that combines the power of Artificial Intelligence (AI) and Quantum Mechanics (QM).

This interdisciplinary collaboration has already yielded significant results, including the creation of SmartCADD, a user-friendly virtual screening platform for drug discovery that can efficiently screen billions of compounds and identify potential leads. In a pilot study, SmartCADD successfully identified derivatives of FDA-approved HIV inhibitors from the ZINC database. The team has also developed QM40, the largest QMx-type dataset to date, which provides a vast repository of structural, thermochemical, and electronic property data for nearly 90% of FDA-approved drugs. The next step is to integrate QM40 into SmartCADD and develop an AI-based algorithm that can predict QM properties without requiring expensive calculations, making the platform accessible to research groups without high-performance computing resources. This cutting-edge research has been made possible through the generous support of the O'Donnell Data Science & Research Computing Institute, which has provided critical computational resources and support for key team members. In addition to the publication, a press release was sent out, which can be found here.



ENHANCING AGE-RELATED MACULAR DEGENERATION OCT IMAGE GRADING BY INTEGRATING HUMAN COMPUTING GAME DATA INTO DEEP NEURAL NETWORKS

Dr. Myque Ouellette, SMU Guildhall



Insufficiently labeled OCT data limits the ability to create new labeled datasets. User errors and lack of proper medical understanding make robust AI/ML training difficult. Therefore, researchers seek alternate means to enhance DNN performance given limited data. Dr. Ouellette's project aggregated the Human Computing Game (HCG) data into a Probability Map Input (PMI) channel for a DNN framework to provide human guidance regarding the retinal boundary positions to the DNN, improving evaluation quality and robustness despite a small dataset. In the framework, several DNN models are implemented, commonly used in evaluating retinal OCT and other medical image segmentation applications. Initial testing using the popular U-Net DNN, shows the DNN using the HCG PMI demonstrates increased accuracy and robustness, and trains more efficiently (in fewer epochs) than the DNN alone. An HCG PMI DNN trained to evaluate standard OCT images also demonstrated improved performance in evaluating new, high resolution OCT scans, compared to standard DNNs.

INNOVATIVE APPROACHES TO PROTEIN DYNAMICS AND AI-ASSISTED DRUG DESIGN

Dr. Peng Tao, Chemistry

The Tao Research Group at SMU, a pioneering team in the field of protein dynamics and drug development, has made significant strides in advancing understanding of allosteric mechanisms and developing innovative tools for AI-assisted drug design. At the forefront of this research is PASSer (Protein Allosteric Sites Server), a user-friendly web application that utilizes machine learning to identify critical allosteric sites in protein ensembles. Since its release, PASSer has garnered widespread attention and global usage, providing valuable insights for researchers worldwide. Additionally, the group has created an AI-assisted drug design platform that features several computational models, including ADMETboost, a machine learning tool for predicting absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties.

The Tao Research Group's commitment to advancing research in protein dynamics and drug development is evident in their cutting-edge computer simulations and modeling. Over the past five years, the group has made notable strides in three major areas: protein allostery, evolution of **(cont.)**

enzyme catalysis, and fundamental chemical mechanism studies. Supported by funding from the National Institutes of Health (NIH), National Science Foundation (NSF), and American Chemical Society (ACS) Petroleum Research Fund, the group has developed novel theoretical frameworks that integrate machine learning with enzyme catalysis.

ATLAS EXPERIMENT: MEASUREMENT OF OFFSHELL HIGGS PRODUCTION WITH AI TECHNIQUES

Dr. Ryszard Stroynowski, Physics; Dr. Arthur Schaffer, External Collaborator, Univ. Paris-Sud

In a groundbreaking study, Dr. Ryszard Stroynowski and Dr. Arthur Schaffer have employed advanced AI techniques to measure offshell Higgs production as part of the ATLAS experiment. Building on the 2012 discovery of the Higgs boson, which was identified through its decay into a pair of Z bosons and subsequently into leptons, this research focuses on the 15% of Higgs bosons produced with a mass above its 125 GeV resonance mass. To isolate these offshell events from overwhelming backgrounds, including those that interfere quantum mechanically with Higgs boson production, the team leveraged simulated decays of the four-lepton decay channel. By harnessing AI-driven methods to capture these interference effects, they constructed probability distributions for each event, creating a test statistic that effectively discriminates between Higgs boson offshell production and background noise. This innovative approach has significantly improved the precision of previous measurements, enabling not only a more accurate assessment of offshell production but also the determination of the Higgs boson's width - a fundamental property that has long been a subject of interest in the physics community.

ADVANCING NLP AND HPC INTEGRATION FOR DATA PROCESSING IN CHATSSRN

Dr. Jane Tan, Information Technology & Operations Management

Dr. Jane Tan, with the support of the O'Donnell Data Science and Research Computing Institute, is harnessing cutting-edge technology to revolutionize data processing for the ChatSSRN platform. By tapping into the power of High-Performance Computing (HPC), her team is able to efficiently index abstract data using Intel/bge-small-en-v1.5 LLM, ensuring seamless and regular system updates that incorporate newly collected data on a monthly basis. Furthermore, Dr. Tan's project is also leveraging the capabilities of Natural Language Processing (NLP) techniques to automate the extraction and processing of unstructured text - a critical task that has far-reaching implications for various industries, including healthcare, finance, and legal technology. By combining the strengths of HPC and NLP, Dr. Tan's work is empowering ChatSSRN to remain a dynamic and cutting-edge platform, providing users across sectors with the most up-to-date information and insights, while paving the way for significant advancements in fields such as law, where NLP's ability to process unstructured text is particularly vital.

ColdFront Arrived November 1

What is ColdFront?

ColdFront is a powerful application designed to simplify the management of High-Performance Computing (HPC) resource allocations. With ColdFront, authorized users can:

- Create new projects
- Request HPC compute and storage resources for specific research projects
- Assign users to collaborate on those projects

By automating the provisioning of new users onto HPC systems, ColdFront enables researchers to start working and collaborating more quickly.

Mandatory Use of ColdFront Allocations for Submitting Jobs Starting November 1

As of November 1, 2024, submitting and running jobs on SMU HPC systems require a SLURM account associated with an active compute allocation in a ColdFront project. This change also impacts jobs submitted before November 1 that are still queued or running, if they do not meet this requirement.

Getting Started with ColdFront

To ensure a smooth transition, we've prepared documentation for common tasks in ColdFront:

- Quick Start Guide
- Creating a Project
- Requesting an Allocation
- Adding or Removing Users
- Running Jobs
- Frequently Asked Questions

Need Help?

If you need assistance setting up projects and allocations in ColdFront or have any questions or concerns, please don't hesitate to reach out:

- Email: help@smu.edu (include [HPC] in the subject line)
- Attend an HPC or ColdFront drop-in support sessions for personalized assistance (schedule)

FACULTY RESEARCH ACCELERATION GRANTS

The O'Donnell Institute Faculty Research Acceleration Grants were awarded to enable/support scalable high performant AI/ML application development using HPC technology. The current projects are:

- SmartCADD: AI-QM Empowered Drug Discovery Platform with Explainability in a Secure, Web-Based SaaS Platform (Drs. Elfi Kraka and Corey Clark).
- Unveiling Seismotectonics of the Subduction Zones: Leveraging Machine Learning Analysis of Ocean Bottom Seismometers (Dr. Heather DeShon).
- Fairness-aware Multimodal AI for Healthcare Data (Drs. Eric Larson and Mehak Gupta).

CONGRATULATIONS TO OUR 2025 O'DONNELL INSTITUTE FELLOWS

The O'Donnell Institute Graduate Student Fellowships support doctoral students pursuing research in advanced AI/ML technologies, digital twins, and quantum information sciences. The current research projects are:

- A Self-Supervised Deep Learning Approach for Radar Interferometry Denoising. Student: **Kang Liang** (Earth Sciences), Advisor: Zhong Lu.
- Algorithms for Reduced-Order Modeling and Data Assimilation with Applications to Digital Twins of Complex Flow System. Student: **Mark De Vernon** (Mathematics), Advisor: Tom Hagstrom.
- Toward World Model-Based Learning: A New Paradigm for Generalized Artificial Intelligence. Student: **Eli Laird** (Computer Science and Guildhall), Advisor: Corey Clark.
- Quantum Reinforcement Learning in Power Distribution System Operations. Student: **Ding Lin** (Electrical and Computer Engineering) , Advisor: Jianhui Wang.
- Advanced Machine Learning Techniques for Discovering Aging Biomarkers in Human Skeletal Muscle. Student: **Yinglu Tang** (Biology) , Advisor: Zhihao Wu.