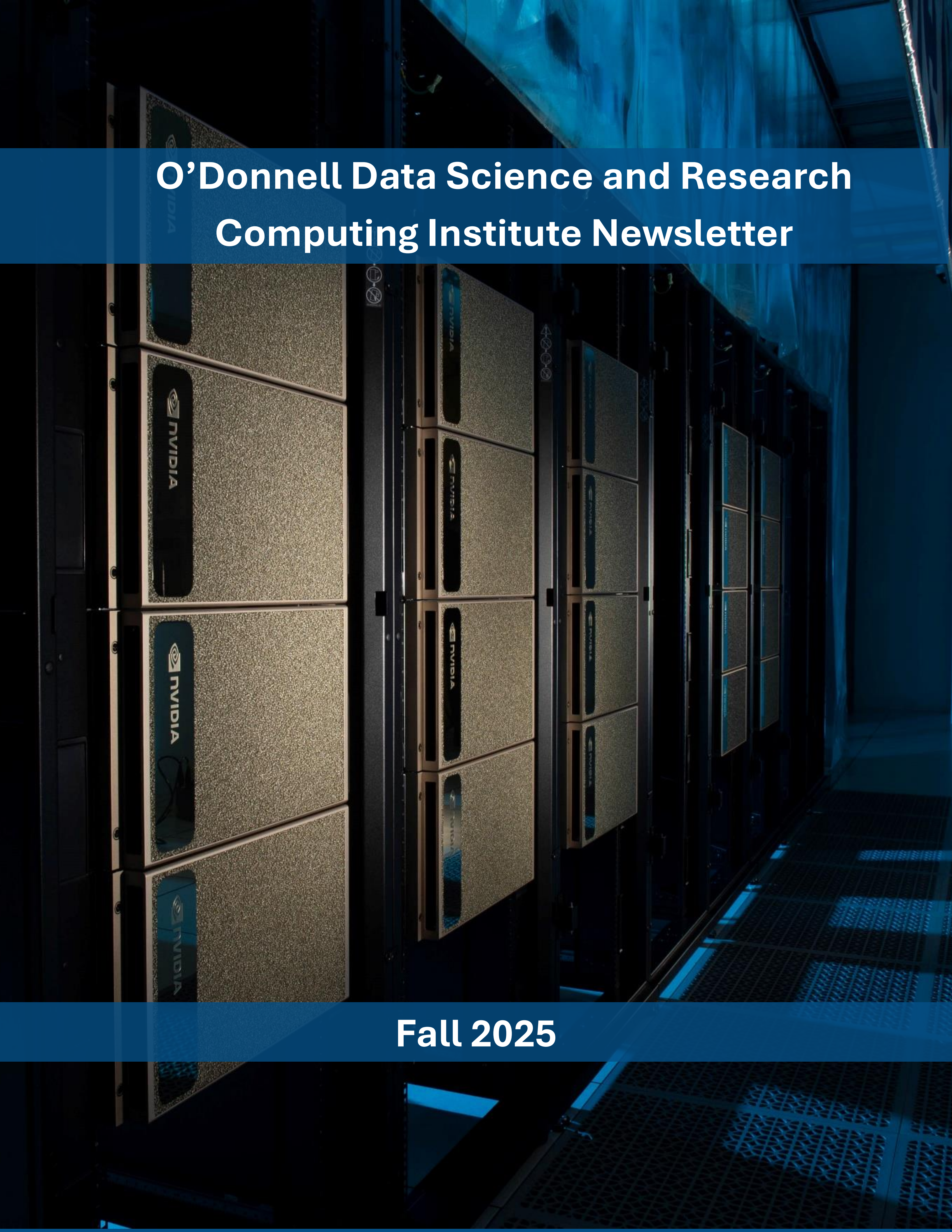
O’Donnell Data Science and Research Computing Institute Newsletter



Fall 2025

A WORD FROM THE DIRECTOR

Dear Colleagues,

I am pleased to share the Fall 2025 O'Donnell Data Science and Research Computing Institute (ODSRCI) newsletter. This semester was both dynamic and productive, reflecting the Institute's continued commitment to advancing AI-enabled research and high-performance computing to enable impactful, real-world applications.

Throughout the fall, ODSRCI strengthened its role at the intersection of research, education, and practice. We delivered a custom Generative AI course for the Federal Reserve Bank, providing hands-on training that led participants from foundational concepts to enterprise-scale AI systems. Our seminar series also featured leaders from industry, including NVIDIA, who highlighted how accelerated computing is reshaping data science and machine learning workflows. We also launched several new interdisciplinary research initiatives, including *Federated Learning for Drug Discovery* and *AI-enabled Data Compression*. These projects address critical challenges in privacy, security, and scalability and demonstrate how advanced algorithms, combined with high-performance computing, can unlock new capabilities across diverse scientific domains.

Equally important is our investment in people. I am proud to congratulate the newest cohort of ODSRCI graduate fellows, whose innovative projects span physics, engineering, computer science, and advanced manufacturing.

I look forward to expanding collaborations across SMU and beyond and to announcing additional research opportunities in the coming months.

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

CONTENTS

RECENT RESEARCH AT THE O'DONNELL INSTITUTE

Federated Learning for Drug Discovery

Dr. Neena Imam, O'Donnell Institute | Nahed Abdelgaber, Computer Science

The O'Donnell Institute is spearheading GEN-FLARE (**G**enerative **E**ngineering of **N**ovel proteins via **F**ederated **L**earning with **A**dversarial **R**esilience and **E**nhanced Security), a new initiative designed to accelerate drug discovery through secure, privacy-preserving artificial intelligence. Traditional drug development is slow and costly, and much of the most valuable biomedical data remains locked within institutional silos due to privacy regulations, ethical constraints, and competitive concerns. GEN-FLARE addresses these barriers by creating a shared AI network that allows institutions to collaborate without exposing their sensitive data.

Central to the project is Federated Learning (FL), an approach that enables multiple partners to train a common model while keeping their data local. However, FL can be vulnerable to adversarial attacks that manipulate model updates. SMU researchers have developed algorithms that filter out malicious updates and adapt dynamically to evolving threats. This FL framework will incorporate a generative AI model that designs novel CRISPR-associated proteins for potential gene therapy applications.

By demonstrating that institutions can work together without compromising privacy or security, GEN-FLARE aims to reshape how the biomedical community approaches innovation. The University of North Texas Health Science Center is partnering with SMU as a collaborator on this project.

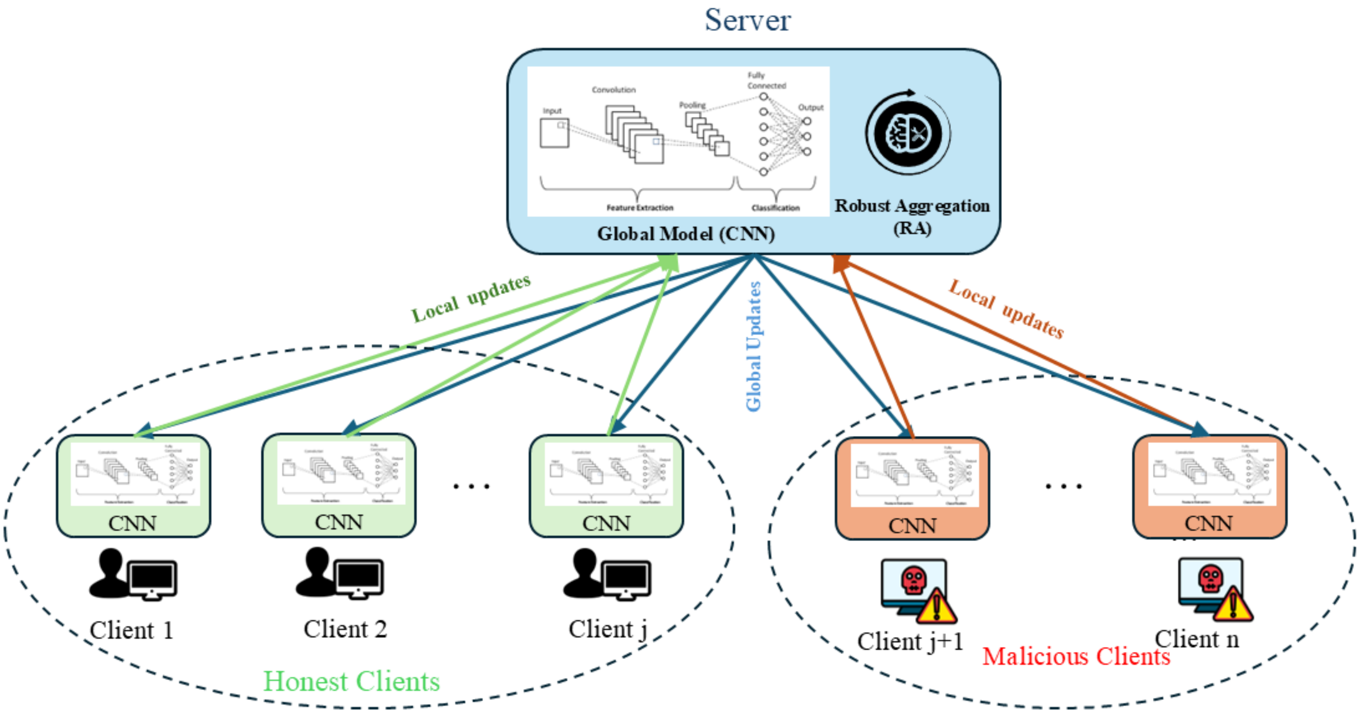


Figure 1: Federated Learning framework for secure and privacy preserving drug discovery.

AI-Enabled Data Compression

Dr. Neena Imam, O'Donnell Institute | Dr. Abrar Alam, O'Donnell Institute

In an era where digital data grows and diversifies every year, efficient data compression is a necessity. From scientific simulations and web archives to digital libraries, storage and transmission costs can dominate operational budgets if data is left uncompressed or inadequately handled. For decades, dominant approaches to data compression revolved around statistical and rule-based techniques, which leverage symbol frequencies and local correlations to reduce redundancies. While these methods have been proven optimal under certain constraints, they struggle to learn rich semantic patterns and long-range dependencies, particularly in natural language. The last decade saw the emergence of machine learning–based compression. Early neural compressors hinted at the potential of learned representations to outperform traditional strategies on structured datasets. However, these models were often limited by vanishing-gradient issues, short context windows, and prohibitive training costs. More recently, large, attention-based generative models have transformed natural language processing by capturing long-range dependencies through self-attention mechanisms. These models can produce highly accurate next-token probability distributions, making them well-suited to serve as the predictive backbone of entropy-coding methods such as Arithmetic Coding, which converts such probabilities into near-optimal and compact bitstreams.

Recent research at the O'Donnell Institute adopts this hybrid paradigm by integrating advanced language models with Arithmetic Coding to achieve lossless text compression with state-of-the-art quality. We evaluated diverse model architectures within a unified, end-to-end compression pipeline in which causal, next-token predictions feed directly into Arithmetic Coding to produce compact bitstreams that can be perfectly reconstructed during decompression. This hybrid approach leverages the probability estimation strengths of modern transformer-based systems while preserving the theoretical compression performance of Arithmetic Coding. One of the main challenges of model-driven compression is the computational intensity of autoregressive inference. To address this challenge, we conducted comprehensive numerical experiments on the NVIDIA DGX SuperPOD. Results from our experiments demonstrate that hybrid model–Arithmetic Coding systems scale effectively across multi-GPU and multi-node environments and can achieve compression gains beyond traditional techniques. As text data continues to expand and diversify across domains—from scientific archives to multimodal corpora—this hybrid AI-driven compression paradigm points toward scalable and efficient compressors capable of adapting to the demands of modern data infrastructure.

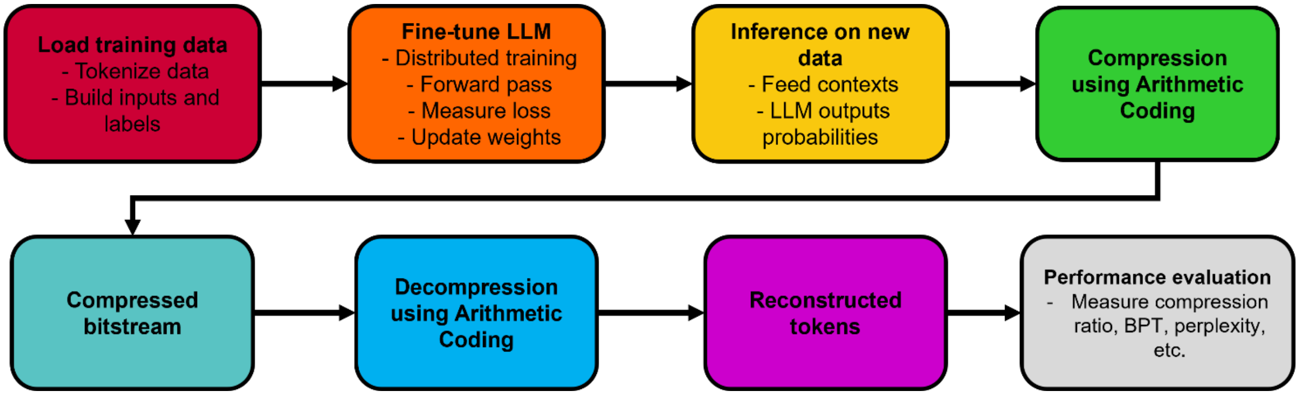


Figure 2: LLM-enabled data compression pipeline.

O'DONNELL INSTITUTE GRADUATE FELLOWS PROJECT HIGHLIGHTS

Nonlinear QCD Evolution with Machine Learning for the Electron–Ion Collider

Junaid Saif Khan, Physics | Advisor: Dr. Fredrick Olness

The small-x regime of Quantum Chromodynamics (QCD) is one of the most dynamic frontiers in nuclear physics, where rapidly growing gluon densities and nonlinear saturation effects reshape the internal structure of hadrons. Understanding these dynamics is essential for interpreting measurements at the upcoming Electron–Ion Collider. Traditional parton-model methods based on Parton Distribution Function (PDF) fits and Dokshitzer–Gribov–Lipatov–Altarelli–Parisi (DGLAP) evolution have provided decades of precision at moderate x, while the dipole model combined with Balitsky-Kovchegov (BK) equation evolution captures the nonlinear physics that emerges at small-x. A central goal of this project's research is to unify these two perspectives and establish a consistent theoretical picture that transitions smoothly across their overlapping regions, validated against experimental Hadron-Electron Ring Accelerator (HERA) measurements.

A major challenge in this effort is computational. Nonlinear QCD evolution is slow and expensive to evaluate at the scale required for global analyses, limiting detailed parameter studies and restricting the incorporation of nonlinear dynamics into modern fitting workflows. To overcome this obstacle, this project developed high-fidelity machine learning surrogates that emulate nonlinear QCD evolution with exceptional speed and accuracy. Using extensive high-precision datasets generated from Running Coupling Balitsky-Kovchegov (rcBK) and BK solvers, both deep neural networks and Deep Operator Networks were trained to learn the mapping from initial conditions to fully evolved dipole amplitudes and cross sections across the relevant kinematic range. These models reproduce the evolution with percent-level accuracy while achieving speedups of two to three orders of magnitude, effectively transforming a slow differential-equation solver into a near-instantaneous predictive tool.

The impact of this computational gain is substantial. The ML surrogates are integrated into χ^2-based global fitting pipelines, allowing for rapid scans, real-time sensitivity studies, and broadened coverage of the small-x region. This makes it possible to perform comprehensive fits that naturally combine linear and nonlinear QCD, guided by both data and computational efficiency. By unifying physics frameworks and removing longstanding computational bottlenecks, this work advances the Institute's mission in AI-driven research computing and provides a scalable foundation for EIC-era precision studies.

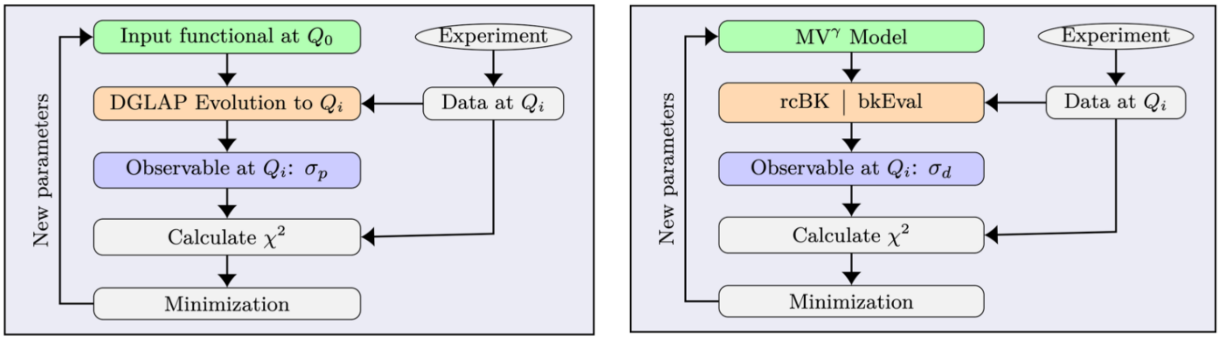


Figure 3: Fitting workflows for linear and nonlinear QCD. The parton model with DGLAP evolution (left) captures perturbative dynamics at moderate x, whereas the dipole model with BK evolution (right) accounts for gluon saturation at small-x. Both frameworks connect theory to data through χ^2 minimization.

Digital Twin-Enabled Accelerated Thermal Surrogate Modeling of 3D Integrated Circuits Using Physics-Informed Neural Networks

Ahmet Ata Ersoy, Mechanical Engineering | Advisor: Dr. Ali Beskok

As semiconductor technology approaches the limits of Moore's Law, industry has increasingly adopted 3D integrated circuits (3D ICs) to improve logic density and reduce data-movement distances. However, this stacking approach increases heat density, making fast and accurate thermal analysis essential for ensuring reliability in next-generation chips. Traditional thermal solvers based on the Finite Element Method (FEM) offer high accuracy but are computationally intensive. This is a major bottleneck for design optimization and real-time thermal management.

This project develops a Physics-Informed Neural Network (PINN) surrogate model to bridge the gap between numerical solvers and data-driven approaches. PINNs embed the governing 3D transient heat conduction equations directly into their loss function, enabling them to maintain physical accuracy while requiring far less training data than conventional machine-learning models. Once trained on high-fidelity FEM data generated using Cadence Celsius, the PINN is expected to deliver full-field temperature predictions in milliseconds, providing near real-time performance that conventional solvers cannot approach.

The resulting model will be integrated into a unidirectional Digital Twin (DT) platform for real-time thermal monitoring. In Phase I, the DT operates in a predictive, one-way configuration. It ingests power-trace data and uses the PINN to perform near–real-time inference of temperature fields, visualizing quantities such as hotspot locations, thermal gradients, and spatial temperature distributions across the 3D IC stack. Building on this foundation, Phase II will extend the DT into a fully bidirectional, closed-loop system. In this stage, the DT will also drive actuation, informing or controlling hardware elements such as fans, pumps, and DVFS settings through feedback channels. While Phase I establishes the physics-informed surrogate model and real-time predictive DT, Phase II will introduce hardware-in-the-loop mechanisms to realize a closed-loop architecture.

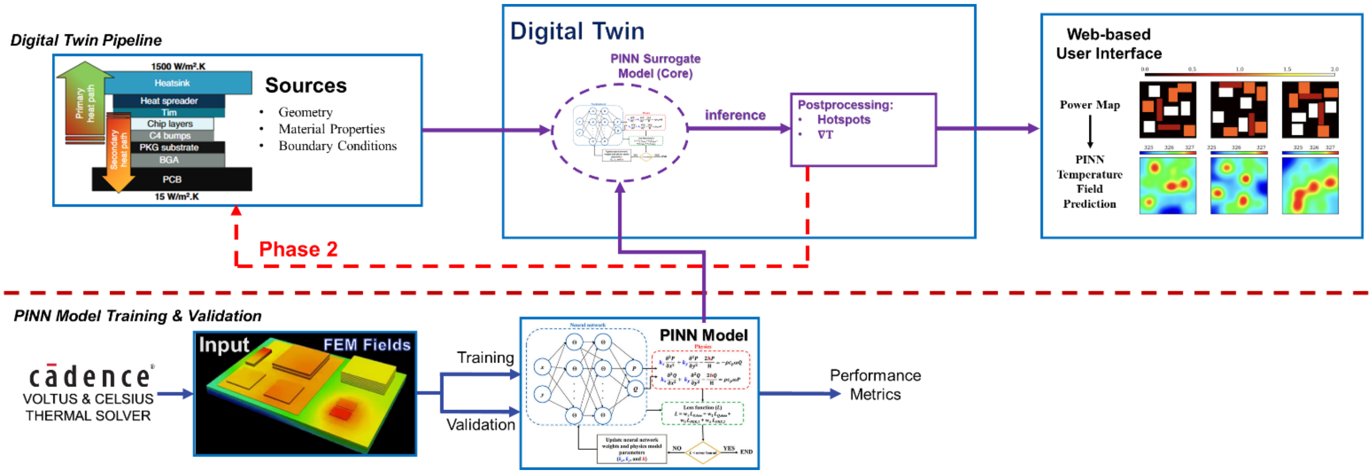


Figure 4: Schematic overview of the Digital Twin pipeline for accelerated thermal analysis of 3D integrated circuits. High-fidelity FEM simulations from Cadence Celsius are used to train a Physics-Informed Neural Network (PINN), which incorporates the governing PDEs and boundary conditions directly into its loss function to ensure physical consistency. The trained PINN serves as the core inference engine of the Digital Twin, delivering near-real-time predictions of temperature fields, hotspot locations, and thermal gradients. Phase I establishes this predictive, one-way DT pipeline, while Phase II will expand it to include bidirectional feedback to the physical system.

Integrating AI Driven Modeling with Additive Manufacturing of Fiber Reinforced Composites

Behzad Parvaresh, Mechanical Engineering | Advisor: Dr. Adel Alaeddini

Advances in continuous fiber reinforced composite additive manufacturing (CFRC AM) have made it possible to fabricate lightweight structures with exceptional mechanical performance. However, the strength, stiffness, and overall quality of these printed composites depend on a complex interaction of material and process parameters. Conducting thousands of physical experiments across this design space is impractical. This creates a need for intelligent modeling tools that can accurately predict performance outcomes while reducing experimental workload.

This project presents an AI-based framework that combines Latin Hypercube Sampling (LHS) with a Squeeze and Excitation Wide and Deep Neural Network (SE WDNN) to jointly predict multiple mechanical and manufacturing properties of CFRC AM parts. From a total of 4,320 potential printing conditions, 155 representative specimens were fabricated and tested. This dataset includes several key mechanical properties such as Young's modulus, yield strength, tensile strength, toughness, and resilience, along with manufacturing metrics including part weight, build time, and material cost.

The SE WDNN architecture integrates a wide component that captures high level interactions with a deep component that learns nonlinear relationships. The squeeze and excitation blocks dynamically recalibrate feature importance, allowing the model to focus on the most influential characteristics. This hybrid approach achieved the highest predictive accuracy among all models evaluated. It outperformed XGBoost, CatBoost, Kolmogorov Arnold Networks, and random forests, producing the lowest test error. SHapley Additive exPlanations (SHAP) based interpretability analysis showed that reinforcement strategy is the most influential factor affecting mechanical performance, providing practical insight for manufacturing optimization.

By combining efficient experimental design with advanced machine learning, this research demonstrates a powerful method for exploring complex manufacturing design spaces. It improves predictive reliability and accelerates the development of high-performance composite structures.

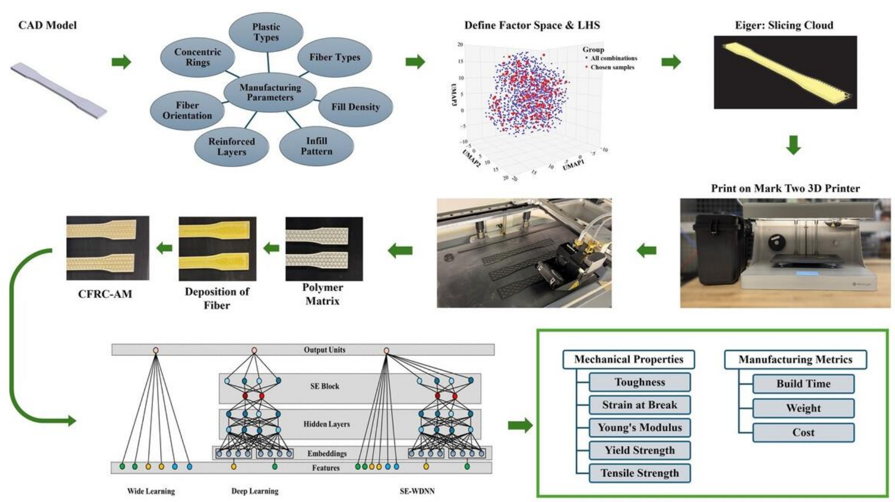


Figure 5: The full workflow of the project, from defining composite manufacturing parameters and fabricating CFRC-AM specimens on a Markforged Mark Two printer to training a Wide and Deep Neural Network augmented with squeeze and excitation blocks that predicts key mechanical properties and production metrics from only 155 strategically sampled builds.

Solving the AI Forged Meeting Participants in Video Calls Using Two Methods: Acoustic Echo and Wi-Fi Signals CSI

Jingwei Zhang, Computer Science | Advisor: Dr. Chen Wang

With the rise of deep learning, AI tools for making fake videos and voices have become widely available. Deepfake systems can now generate people who look and sound real, even when they do not exist. This brings serious risks: a recent McAfee report shows that deepfake-related scams grew by nearly 3,000% in 2024, as criminals copy voices from short clips or turn public photos into convincing videos.

Yet even if attackers can fake audio and video, they still cannot fake the physical world around them. In many scams, the "person" on the call is not in any real room. Their media is routed into the meeting through software tools instead of real microphones or cameras.

We propose a method that continuously probes the remote side of a call to confirm the presence of a real human and a real physical space. Our system uses two natural by-products of VoIP communication: acoustic echoes, created when we send audio and record the returned sound, and Wi-Fi Channel State Information (CSI), which changes with human motion and the surrounding environment.

Our earlier work showed that simple notification tones played through a bone-conduction headset can serve as reliable probing signals, even under low sampling rates and built-in noise reduction. Other studies show that CSI contains rich physical cues related to presence and movement.

In this project, we use deep learning to analyze both echoes and CSI. We collect data across varied indoor scenes, multiple users, and different devices. Audio is processed with a CNN–LSTM structure, while CSI is handled with depth-wise convolution networks for high-dimensional wireless data. Multi-head attention then fuses the two modalities into a single multimodal recognizer.

This study aims to make multimodal sensing a practical tool for defending against AI-driven impersonation attacks and improving trust in everyday video communication.

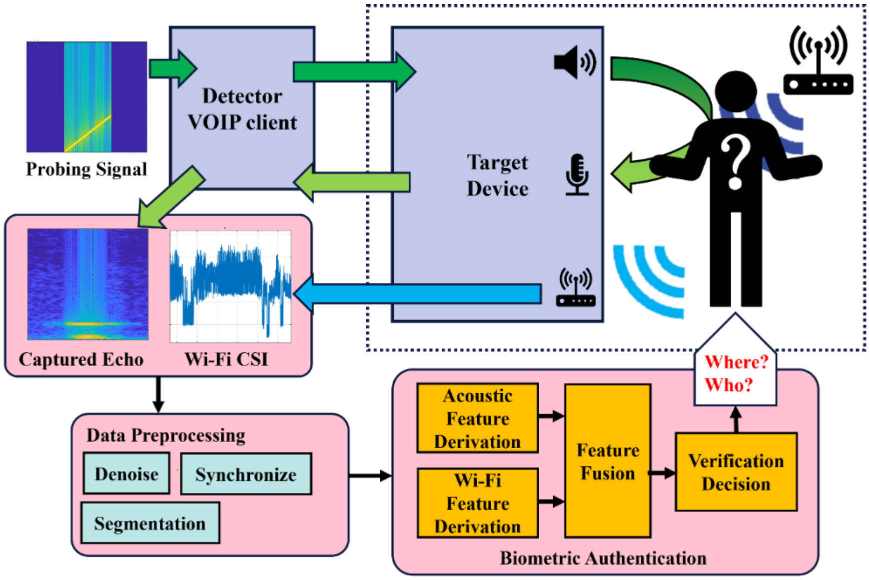


Figure 6: System framework of the multimodal Deepfake participant detector using acoustic echo and Wi-Fi CSI.

FALL 2025 EVENTS

SEMINAR SERIES

October 6: Accelerated Data Science: A Paradigm Shift for Data Wrangling, Analysis, and Machine Learning | William Hill, NVIDIA

Mr. Hill discussed how GPU-accelerated ML and Data Science libraries reduced the data-wrangling bottleneck using the code participants already had in place. Through benchmarks and case studies, he demonstrated double- to triple-digit speedups in data processing, graph analytics, and model training with the tools attendees already used. Slides from Mr. Hill's presentation can be found here.

William Hill started studying computer science at NC State University when he was activated into military duty in 2003 during the Iraq and Afghanistan Wars. He was a computer programmer for the US Air Force but went into industry after his four years. He worked as a software engineer for various startups and large companies like IBM and SAS Institute. He transitioned to machine learning and data science while working on SAS's Enterprise Miner application and went to be a data scientist in the early days of the profession at places like IBM, Red Hat, and Google. William is currently a Senior Developer Advocate at NVIDIA with a focus on accelerating open-source ML, data science, and AI libraries with existing APIs. He finished his Computer Science BS while on active duty at Park University and earned a Master's degree from Boston University.

CUSTOM GENERATIVE AI TRAINING FOR THE FEDERAL RESERVE BANK

October 28-29, November 4-5

The O'Donnell Institute provided a custom-built course*, From Foundations to Applications in Generative AI*, designed exclusively for the Federal Reserve Bank of Dallas. Developed and led by Dr. Neena Imam and her team, this two-week program provided a rigorous, hands-on pathway from core GenAI principles to advanced Retrieval-Augmented Generation (RAG) systems, empowering analysts, researchers, and technical staff to build secure, efficient, and domain-aware GenAI tools. Participants progressed through foundational model families, data ingestion and indexing practices, vector databases, prompt engineering, guardrails, evaluation metrics, and full-stack RAG implementations using OpenWebUI and Langchain. They also leveraged SMU's NVIDIA DGX SuperPOD and its powerful A100 GPUs to train, test, and deploy large language models at scale, gaining first-hand experience with enterprise-grade AI infrastructure. This high-performance environment allowed participants to experiment with sophisticated model architectures and accelerate workflows that would be impractical on standard hardware. Collaborations like this highlight what is possible when leading-edge research and motivated professionals work together.

SC25 (THE INTERNATIONAL CONFERENCE FOR HIGH PERFORMANCE COMPUTING, NETWORKING, STORAGE, AND ANALYSIS) BIRDS OF A FEATHER SESSION

November 20: Continuum Computing: The Resilience Challenge

The new paradigm of continuum computing (also known as the digital continuum) is a distributed, multi-layered ecosystem that spans sensors at the edge, interconnected instruments, cloud platforms, datacenters, exascale supercomputers, and recently quantum computers. Continuum computing has evolved to keep pace with the growth and expansion of geographically distributed science and hybrid computing infrastructures. In the continuum paradigm, computation and data are orchestrated in various stages from the edge to the core to optimize data movement and response times. Novel solutions are needed for system design, software frameworks, workflows that can react to dynamic data sizes, monitoring tools, multisite governance policies, actionable experimental metrics, etc. We organized the inaugural BoF on continuum computing at SC22, which discussed state-of-the-art in the digital continuum. The SC23 BoF discussed the aggregation and synthesis of previously distinct techniques and tools (such as HPC, AI/ML, and digital twins) to advance continuum computing. At SC24, we focused on the role of quantum information science in advancing continuum computing. For SC25, our theme was resilience in the digital continuum.

Session Leader

Dr. Neena Imam

Peter O'Donnell Jr. Director
O'Donnell Data Science and Research Computing Institute
Southern Methodist University

Additional Speakers

Dr. Nagi Rao - Oak Ridge National Laboratory (ORNL)

Rao joined ORNL in 1993. He is currently a Corporate Fellow. He received his B. Tech (1982) from NIT Warangal, M.E. (1984) from the Indian Institute of Science, and his Ph.D. (1988) in Computer Science from Louisiana State University. He is PI of DOE PiQSci: Performance Integrated Quantum Scalable Internet project. He is a Fellow of IEEE and received 2005 IEEE Computer Society Technical Achievement Award and 2014 R&D 100 Award.

Dr. Ian Foster - Argonne National Laboratory (ANL)

Ian Foster is Senior Scientist, Distinguished Fellow, and Director of the Data Science and Learning Division at Argonne National Laboratory, and the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago. He has a B.Sc. degree from the University of Canterbury and a Ph.D. from Imperial College, both in computer science. He is a Fellow of the AAAS, ACM, BCS, and IEEE, and has received the BCS Lovelace Medal; the IEEE CS Babbage, Goode, and Kanai Awards; and the ACM/IEEE CS Ken Kennedy Award.

Dr. Benjamin Brown – Department of Energy (DOE), Advanced Scientific Computing Research

Benjamin Brown is the Director of the Facilities Division in the Office of Advanced Scientific Computing Research (ASCR). Ben leads and executes the DOE's world-leading research supercomputing, data, and networking infrastructure to enable the DOE mission and the national research enterprise. Ben has been a member of the federal Senior Executive Service since 2020. He also served as the founding program manager of DOE's Project Leadership Institute (2014-21). He holds a Ph.D. in optics from the University of Rochester and a bachelor's in physics from Harvard University.