The INSTITUTE for DATA-INTENSIVE ENGINEERING and SCIENCE

OCT 1, 2023 - SEPT 30,

2024 ANNUAL REVIEW

idies The Institute for Data-Intensive Engineering and Science

JOHNS HOPKINS
INSTITUTE *for*
DATA-INTENSIVE
ENGINEERING & SCIENCE

TABLE OF CONTENTS

ON THE COVER

PAGE 38. SUPPORTING USE OF COMPUTATIONAL METHODS ON THE OPIOID INDUSTRY

This photomosaic contains hundreds of images extracted from documents in the Opioid Industry Documents Archive, a collaborative undertaking between the University of California, San Francisco and Johns Hopkins University to collect, organize, and provide free online access to millions of previously internal documents made public through legal settlements. The documents, and thousands of images extracted from them, are available through SciServer.org. For more information on IDIES' role with OIDA, see page 38.

2024 IDIES ANNUAL REVIEW

Layout and Design KATE STEVENS
Layout and Editing JORDAN RADDICK
Additional Editing Help from RAFAEL SANTOS

Johns Hopkins University
3400 N. Charles St. Baltimore, MD21218
https://www.idies.jhu.edu/

2024 IDIES ANNUAL REVIEW

LETTER FROM THE DIRECTOR

PHOTO CAPTION (right): IDIES director Alex Szalay listens intentively to a keynote speech at the the 2024 IDIES Annual Symposium; PHOTO CREDIT: Alec Zabrecky

IDIES DIRECTOR,
Alex Szalay

idies
The Institute for Data-Intensive Engineering and Science

JOHNS HOPKINS
INSTITUTE *for*
DATA-INTENSIVE
ENGINEERING & SCIENCE

2024 IDIES

FROM SENSORS TO TENSORS: IDIES AT THE FOREFRONT OF DATA SCIENCE

This summer IDIES has moved to the Whiting School of Engineering to become an affiliated center within the Johns Hopkins Data Science and AI Institute (DSAI), which is becoming the focal point for data science and AI research at JHU. Our collaboration with DSAI will blend our tools and expertise with theirs to keep IDIES at the forefront of these fields.

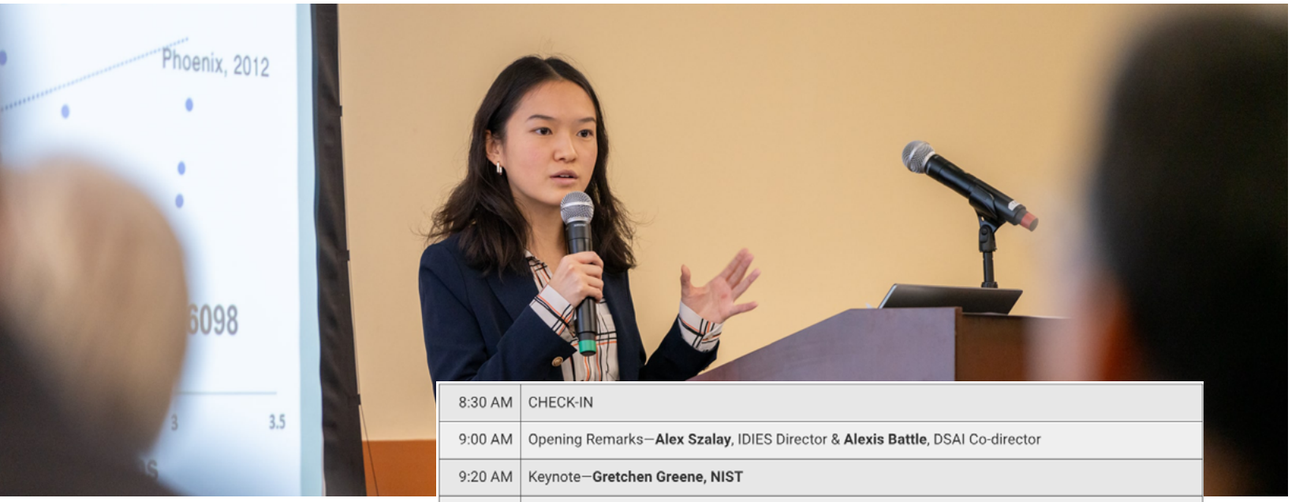
The Scientific Software Engineering Center (SSEC), funded by Schmidt Sciences to engage senior software developers with industry experience to develop software tools for scientific research, has already been part of DSAI from the beginning. DSAI has also just started to hire its own group of Research Software Engineers (RSEs). Having all these software development experts physically co-located with scientists and engineers from IDIES and DSAI will result in faster development of more reliable software tools, leading to faster and better discoveries. For now, these three groups will collaborate while maintaining their separate identities; in the long run, they will blend together to create a large critical mass of software expertise where the whole will become far greater than the sum of its parts.

At the same time, developments in AI by the industry have accelerated, resulting in tools and platforms we could have never dreamed of just a year ago. With industrial giants like Google, Facebook/Meta, and OpenAI investing billions of dollars in AI research, it is natural to wonder how universities can stay competitive – but this year's Nobel Prize in Physics and Chemistry show the role we can play in the future of AI research. AlphaFold made the final discovery, but its success required huge, well-curated open datasets generated and explored over the course of decades by researchers and citizen scientists all over the world. Industry will never have the patience and endurance to conduct major scientific experiments – it is our task to continue to do so!

At IDIES, our greatest potential contributions lie in our ability to manage Petabyte-scale datasets in all areas of science, providing the community with the tools and documentation to maximize the potential for new discoveries. The practice of researchers writing their own analysis and visualization tools can still result in useful prototypes, but is no longer sufficient to serve the needs of the broader research community. Creating robust, scalable tools that work in all local and cloud environments takes significant software engineering expertise, and making those tools run efficiently even more so. Efficiency is even more important as larger GPU architectures come online. If we plan to spend millions of dollars on hardware, we need to make sure we get our money's worth by writing code that works quickly and reliably anywhere.

Writing good software for data management and will become even more important in the coming years thanks to the proliferation of reliable, inexpensive sensors that can connect into networks. Soon every phone and tablet can be a scientific instrument. These scalable experiments using large inexpensive sensor arrays will generate lots of data, but they also create a massive data management challenge to make the data amenable to AI. In the near future scientists will not only use AI to analyze their data, but use it to drive our microscopes and telescopes, and optimize our experiments. Our task as data scientists must be to develop a smooth and natural transition "from sensors to tensors" – turning the raw data into an AI-ready form easily integrated into AI systems for new discoveries. These continuing challenges will offer us novel opportunities for decades to come.

Alex Szalay
ALEX SZALAY



2024 ANNUAL SYMPOSIUM

2024 IDIES

8:30 AM	CHECK-IN
9:00 AM	Opening Remarks—**Alex Szalay**, IDIES Director & **Alexis Battle**, DSAI Co-director
9:20 AM	Keynote—**Gretchen Greene, NIST**
9:50 AM	Seed Update—**Josiah Lim** *(Presenting on behalf of Tamás Budavári & Dr. Rohan Mathur)*
10:05 AM	Summer Student Fellowship Update—**Akshaya Ajith**
10:15 AM	BREAK
10:45 AM	Seed Update—**Mick Bonner,** JHU Department of Cognitive Science
11:00 AM	Summer Student Fellowship Update **—Millie Mi**
11:10 AM	SciServer Update—**Gerard Lemson**, IDIES Director of Science
11:20 AM	Keynote Speech—**Charles Meneveau**, Whiting School of Engineering
11:50 AM	Summer Student Fellowship Update—**Krishna Mukunda**
12:00 PM	**Poster Madness**
12:30 PM	LUNCH & **Poster Gallery**
2:00 PM	Keynote Speech—**Denis Wirtz**, JHU Vice Provost for Research
2:30 PM	Sheridan Libraries Update—**Megan Forbes**, JHU Open Source Programs Office (*presenting virtually*)
2:40 PM	Seed Update—**Dr. Kemar Green**, Assistant Professor of Neurology
2:55 PM	Summer Student Fellowship Update—Chris Le Wang
3:05 PM	Seed Update—**Paulette Clancy**, Edward J. Schaefer Professor in Engineering
3:20 PM	Summer Student Fellowship Update—**David Zhou**
3:30 PM	IDIES Infrastructure Update
3:40 PM	BREAK
4:05 PM	Seed Update—**Adam Sheingate**, Professor of Political Science
4:20 PM	Keynote Speech—**Elana Fertig**, University of Maryland School of Medicine
4:50 PM	Closing Remarks—**Alex Szalay**, IDIES Director
5:00 PM	RECEPTION

PHOTO CAPTION (above):SSF recipient Millie Mi speaks at the 2024 IDIES Annual Symposium; PHOTO CREDIT: Alec Zabrecky; VIEW MORE SYMPOSIUM PHOTOS; GALLERY CREDIT: Alec Zabrecky

2024 KEYNOTE SPEAKERS

Elana FERTIG

Director of the Institute for Genome Sciences
THE UNIVERSITY OF MARYLAND (UMD)

Forecasting pancreatic carcinogenesis through spatial multi-omics

Gretchen GREENE

Director for Research Data and Computing, Associate Director of Laboratory Programs
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

NIST Data Driven Science

Charles MENEVEAU

Louis M. Sardella Professor in Mechanical Engineering
JOHNS HOPKINS UNIVERSITY

Making massive simulation datasets in fluid turbulence accessible to human and artificial intelligence

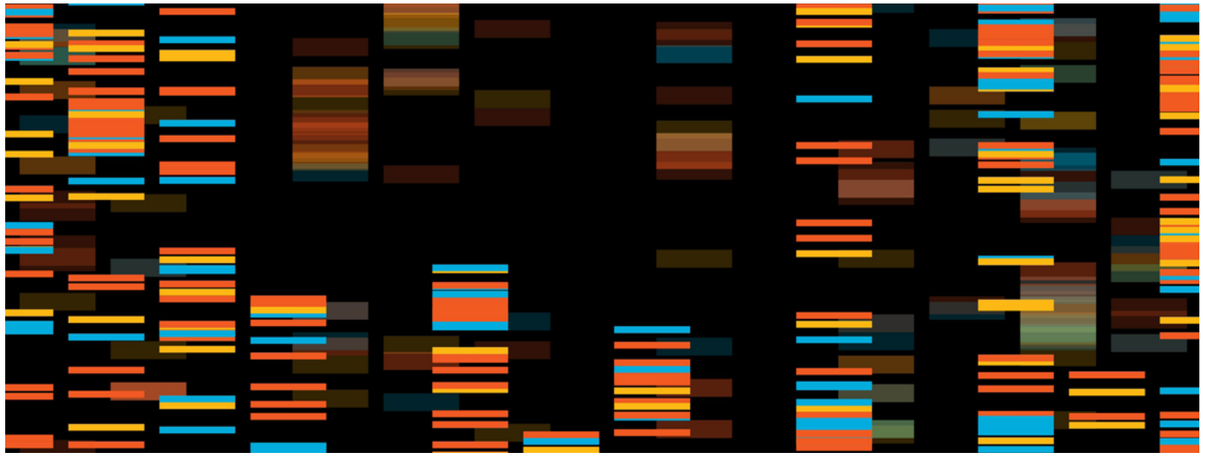
Denis WIRTZ

Vice Provost For Research And Theophilus Halley Smoot Professor of Engineering Science
JOHNS HOPKINS UNIVERSITY

CODA: 3D reconstruction of tumors and whole organs and organisms at single-cell resolution using AI

KEYNOTE SPEAKER

2024 IDIES ANNUAL SYMPOSIUM



ELANA FERTIG

Director of the Institute for Genome Sciences at the University of Maryland School of Medicine (UMD)

Elana Fertig, PhD, FAIMBE, is internationally-recognized for her work in integrating spatial multi-omics technologies with mathematical models to develop a new predictive medicine paradigm in cancer. She began her role as the new Director of the Institute for Genome Sciences (IGS) at the University of Maryland (UMD) School of Medicine in 2024.

Prior to her appointment at UMD, Dr. Fertig also served as a Professor of Oncology at the Johns Hopkins University with dual appointments in the Department of Applied Mathematics and Statistics and the Department of Biomedical Engineering. She was the Co-Director of the Single Cell Consortium at the Johns Hopkins University School of Medicine, as well as the Co-Director of the Convergence Institute and the Associate Director of Quantitative Sciences at the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins.

Fertig's landmark research involves using computational methods to identify biomarkers and molecular mechanisms of therapeutic resistance from a vast trove of multi-platform genomics data.

KEYNOTE SPEAKER

GRETCHEN GREENE

Director of Research Data and Computing; Associate Director of Laboratory Programs at the National Institute of Standards and Technology (NIST)

2024 IDIES ANNUAL SYMPOSIUM

Gretchen Greene is the Director for National Institute of Standards and Technology (NIST) Research Data and Computing, a new office supporting NIST Laboratory Programs. She coordinates and oversees multiple data science initiatives to facilitate laboratory information management systems architecture, design, and implementation and is the project manager for NIST open access infrastructure. She is actively engaged in organizational strategic planning for mission priority areas and consults regularly with domain experts primarily in the physical sciences and engineering fields to enable innovative approaches with data-oriented technology. She recently served on the Research Data Alliance technical advisory board and is an advocate for standards and community driven digital architecture to promote innovation and sustainability. In support of CHIPS R&D, she serves as lead for the implementation of NIST CHIPS Metrology Exchange to Innovate in Semiconductors (METIS), envisioned as a digital ecosystem supporting complex knowledge sharing across industry, government and academia.

KEYNOTE SPEAKER

CHARLES MENEVEAU

Louis M. Sardella Professor
in Mechanical Engineering at
Johns Hopkins University

2024 IDIES ANNUAL SYMPOSIUM

Charles Meneveau is an expert in the multiscale aspects of turbulence, large-eddy simulations, and wind farm fluid dynamics.

He employs computational and theoretical tools for his research and pursues subgrid-scale modeling, downscaling methods, fractal geometry, and their applications to large eddy simulation (LES). Research advances were made possible by elucidating the properties of the small-scale motions in turbulent flows and applying the new insights to the development of advanced subgrid-scale models, such as the Lagrangian dynamic model. This model has been implemented in various research and open-source CFD codes (e.g. OpenFoam) and expanded the applicability of LES to complex-geometry flows of engineering and environmental interest, where prior models could not be used. Current LES research is focused on improving wall models and subgrid-scale models for velocity gradients which is of interest to some turbulent multiphase flows.

Among the application areas of LES being pursued in Meneveau's group is the study of complex flows in large wind farms. Using the improved simulation tools as well as wind tunnel tests, Meneveau and his colleagues identified the important process of vertical entrainment of mean flow kinetic energy into an array of wind turbines. This research has clarified the mechanisms limiting wind plant performance at a time when there is enormous growth in wind farms. The research has led to new engineering models that will allow for better-designed wind farms thus increasing their economic benefit and helping to reduce greenhouse gas emissions from fossil fuels.

Meneveau has participated in efforts to democratize access to valuable "big data" in turbulence. As deputy director of JHU's Institute for Data-Intensive Engineering and Science, he led the team of computer scientists, applied mathematicians, astrophysicists, and fluid dynamicists that built the Johns Hopkins Turbulence Databases (JHTDB). This open numerical laboratory provides researchers from around the world with user-friendly access to large data sets arising from direct numerical simulations of various types of turbulent flows. To date, hundreds of researchers worldwide have used the data, and more than 250 peer-reviewed papers have been based on data sampled from JHTDB. The system has demonstrated how data resulting from large, world-class numerical simulations can be shared with many researchers who lack the massive supercomputing resources needed to generate such data.

Meneveau has also performed groundbreaking research on understanding several multiscale aspects of turbulence. As part of his doctoral work at Yale, Meneveau and his advisor established the fractal and multifractal theory for turbulent flows and confirmed the theory using experiments. Interfaces in turbulence were shown to have a fractal dimension of nearly 7/3, where the 1/3 exponent above the value of two valid for smooth surfaces could be related to the classic Kolmogorov theory. And a universal multi-fractal spectrum was established, leading to a simple cascade model, which has since been applied to many other physical, biological, and socio-economic systems. Later, as a postdoc at Stanford University's Center for Turbulence Research, Meneveau pioneered the application of orthogonal wavelet analysis to turbulence, introducing the concept of wavelet spectrum and other scale-dependent statistical measures of variability.

KEYNOTE SPEAKER

DENIS WIRTZ

Vice Provost For Research and
Theophilus Halley Smoot Professor
of Engineering Science
at Johns Hopkins University

2024 IDIES ANNUAL SYMPOSIUM

Dr. Wirtz directs the Johns Hopkins Physical Sciences-Oncology Center, and co-directs the Cancer Nanotechnology Training Center and the Cellular Cancer Biology Imaging Research, all National Cancer Institute-funded entities. He is a co-founder and former associate director of the Johns Hopkins Institute for NanoBioTechnology.

Wirtz studies the biophysical properties of healthy and diseased cells, including interactions between adjacent cells and the role of cellular architecture on nuclear shape and gene expression. Cell biophysics, single molecule manipulation, intracellular particle trafficking, instrument development, tissue engineering, and nanotechnology in biology and medicine are among his research interests. He is developing artificial intelligence solutions to map the complex architecture of tissues and tumors in 3D and at single-cell resolution, and recently cofounded AbMeta Therapeutics, a startup that develops antibodies to block metastatic spread in patients with cancer.

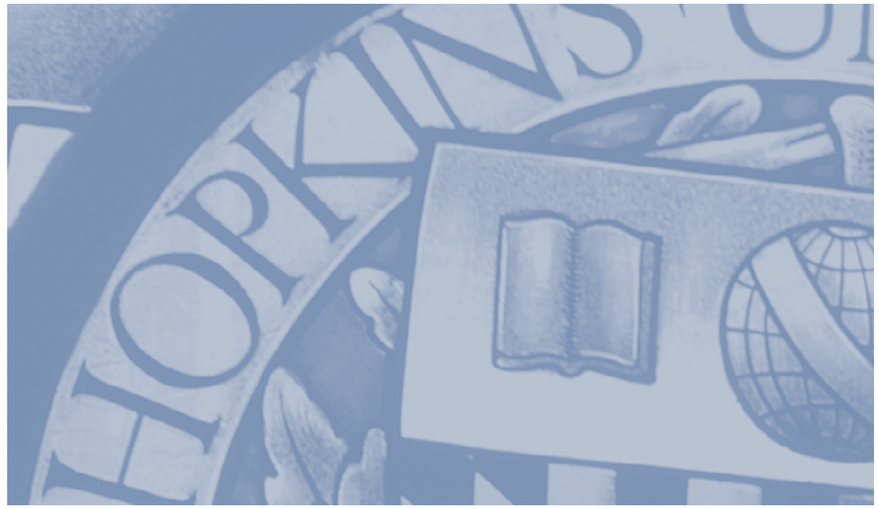
Wirtz was elected Fellow of the American Association for the Advancement of Science for his contributions to cell micromechanics and cell adhesion. AAAS also recognized Wirtz for his development and applications for particle tracking methods to probe the micromechanical properties of living cells in normal conditions and disease state. In addition, Wirtz is a fellow of the American Physical Society and the American Institute for Medical and Biological Engineering. He is a past recipient of the National Science Foundation's CAREER Award and the Whitaker Foundation Biomedical Engineering Foundation Award. In 2019, he was elected to the Royal Academy of Medicine of Belgium.

Wirtz has authored more than 250 peer-reviewed articles published in top journals including Science, Nature, Nature Cell Biology, Nature Methods, Nature Reviews Cancer, Nature Materials, Nature Protocols, PNAS, Nature Communications, and Journal of Cell Biology. Ten of his PhD students and postdoctoral fellows are in faculty positions at research universities and more than a dozen have successful careers in industry. As of 2024, his scholarly work has been cited more than 35,000 times and has an h-index of over 100. He has also given numerous presentations at universities and research institutions around the world.

Wirtz earned his bachelor's degree in Engineering Physics from the Université Libre de Bruxelles in Belgium and master's and doctoral degrees in Chemical Engineering from Stanford University.

SEED FUNDING PROGRAM

2024 IDIES



Every fall, The Institute for Data-Intensive Engineering and Science (IDIES) offers its members the chance to receive up to $25,000 in grant funding for eligible data-intensive research projects.

The goal of the IDIES Seed Funding initiative is to provide pilot funding for data-intensive computing projects that:

(a) will involve areas relevant to IDIES and JHU institutional research priorities;

(b) are multidisciplinary; and

(c) build ideas and teams with good prospects for successful proposals to attract external research support by leveraging IDIES intellectual and physical infrastructure.

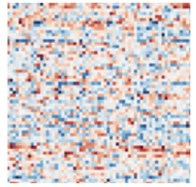
This year, IDIES was prepared to award up to four seed grants, however generous sponsorship from the School of Medicine allowed for the funding of an additional project, resulting in a total of five seed-funded teams, consisting of seven IDIES PIs

For details on the 2025 IDIES funding awards cycle, please visit: https://www.idies.jhu.edu/what-we-offer/funding/

2024 SEED PROJECTS

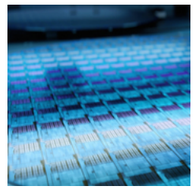
Hybrid Statistical AI for Identifying Neurosurgery Patients in Need of Critical Care

TAMAS BUDAVARI & ROHAN MATHUR



Exploring Uncharted Dimensions of Human Brain Representation

MICHAEL F. BONNER & BRICE MÉNARD



Playing with Chemical Lego Blocks: Rational Design of Semiconducting Polymers for Thermoelectric Materials

PAULETTE CLANCY, CO-PI HOWARD KATZ

The Global Elections Dashboard: Harnessing Data Science to Enhance Electoral Oversight

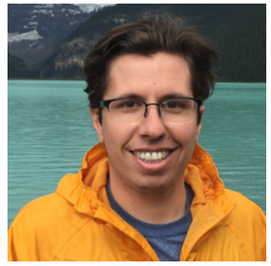
ADAM SHEINGATE, CO-I TOM LIPPINCOTT

Generating Digital Ocular Motor Biomarkers for Deep Learning-Based Neurologic Phenotyping

KEMAR GREEN, CO-I VISHAL PATEL

2024 SEED FUNDING RECIPIENTS

IDIES IS EXCITED TO SUPPORT THE FOLLOWING PIS AND THEIR WORK IN 2024



MICHAEL F. BONNER
Cognitive Science
KRIEGER SCHOOL OF ARTS AND SCIENCES

Mick Bonner is Assistant Professor of Cognitive Science and the director of the Cognitive Neuroscience and Machine Learning Lab. His work uses computational methods, including deep neural networks and advanced statistical techniques, in combination with neuroimaging and behavioral studies to understand the visual system of the human brain. The goal of this work is to identify the statistical principles that govern the representations of visual cortex and to build theoretically grounded models of how these representations are computed from sensory inputs.

BRICE MÉNARD
Physics and Astronomy
KRIEGER SCHOOL OF ARTS AND SCIENCES

Brice Ménard joined the Physics and Astronomy Dept. in 2010. He received his PhD from both the Institut d'Astrophysique de Paris and the Max Planck Institute for Astrophysics in Germany. His research combines physics and statistics, with a primary focus on the physics of learning and the properties of neural networks in artificial & biological systems. He holds secondary appointments in the Cognitive Science and Computer Science departments. Awards: Johns Hopkins President Frontier Award (2019), Packard Fellowship (2014), Sloan Research Fellowship (2012), 2012 Outstanding Young Scientist of Maryland, American Astronomical Society Henri Chrétien Grant Award

PI Bonner, and his lab members use machine learning, nuroimaging, and behavioral studies to gain insights into the representations of human vision and memory. PI Ménard, is a physicist who uses machine learning to understand complex phenomena. He has extensive experience with statistics in high dimensions. He now works primarily on neural networks and neural representations. With shared scientific interests in the statistics of visual computation, Bonner and Ménard have been collaborating for the past two years.

TAMÁS BUDAVÁRI
Applied Mathematics and Statistics
WHITING SCHOOL OF ENGINEERING

Tamás Budavári, an associate professor in the Dept. of Applied Mathematics and Statistics, is known for his studies in the intersection of observational astronomy and statistics with a focus on a large-scale structure, cosmology, and galaxy evolution. He holds a joint appointment in the Dept. of Physics and Astronomy and a secondary appointment in the Dept. of Computer Science. His interests include Bayesian inference, low-dimensional embeddings, streaming algorithms, parallel processing on GPUs, scientific databases, and survey astronomy. His research has been used to understand the dynamics of vacant housing in Baltimore City, to extract high-resolution images of stars in the night sky using repeated exposures, and crossmatch astronomy catalogs.

ROHAN MATHUR
Neurology-Neurocritical Care
SCHOOL OF MEDICINE

Rohan Mathur specializes in Neurological Critical Care and Neurology and currently serves as the Associate Director for the Precision Medicine Center of Excellence in Neurocritical Care, the Medical Director of the Neurosciences Intermediate Care Unit at Hopkins and as the Neurocritical Care Liason for the Hopkins' Biocontainment Unit. In addition to taking care of critically ill patients, Dr. Mathur does translational and clinical research to develop and validate new tools to study and treat high intracranial pressure, as well as coma and other disorders of consciousness. He is the Associate Director and a Scientific Lead for the Precision Medicine Center of Excellence in Neurocritical Care, leading a team of physician scientists, data scientists, and engineers to develop decision support tools for patients with acute brain injuries using Machine Learning and other AI methods.

ADAM SHEINGATE
Political Science
KRIEGER SCHOOL OF ARTS AND SCIENCES

Adam Sheingate is a professor of Political Science at Johns Hopkins University. He is the author of Building a Business of Politics: The Rise of Political Consulting and the Transformation of American Democracy. He has also published widely on various aspects of U.S. politics, public policy, and U.S. political history. His current research includes the application of machine-learning and other tools to the analysis of data on campaign spending in the United States and other large democracies around the world. Sheingate joined the Johns Hopkins University faculty in 2000 and served as Department Chair from 2015-2020.

PAULETTE CLANCY
Chemical and Biomolecular Engineering
WHITING SCHOOL OF ENGINEERING

Paulette Clancy, the Edward J. Schaefer Professor in Engineering, is known for her work in computational materials processing. Clancy is the director of research for the JHU Data Science and AI Initiative, associate director of the Johns Hopkins Center for Integrated Structure-Mechanical Modeling and Simulation (CISMMS), and a fellow of the Hopkins Extreme Materials Institute (HEMI). She is a Fellow of the Royal Society of Chemistry and the American Institute of Chemical Engineers. Clancy leads one of the top groups in the country studying atomic- and molecular-scale modeling of semiconductor materials, ranging from traditional silicon-based compounds to all-organic materials.

KEMAR GREEN
Neurology
SCHOOL OF MEDICINE

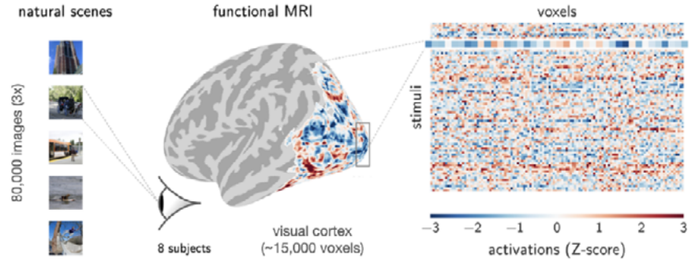
Kemar Green, D.O., is a neurologist with expertise in vestibular diseases, eye movement disorders and undifferentiated/rare neurologic disorders. He specializes in treating patients with disorders causing vestibular symptoms (dizziness, imbalance, vertigo, oscillopsia, imbalance, ataxia), abnormal eye movements, double vision and vision loss. Dr. Green earned his osteopathic medicine degree from the New York Institute of Technology College of Osteopathic Medicine. He completed his neurology residency at Michigan State University, followed by a fellowship in neurotology and efferent neuro-ophthalmology at the Hopkins. He also completed training in Artificial Intelligence and Precision Care Medicine at the Whiting School of Engineering.

SEED FUND PROJECT

2024 IDIES

EXPLORING UNCHARTED DIMENSIONS OF HUMAN BRAIN REPRESENTATION

MICK BONNER
&
BRICE MÉNARD



Big data in human neuroscience. Initiatives to collect massive neuroimaging data on the human brain make it possible to characterize the nature of neural representations at an unprecedented scale. Our work leverages big data and new statistical methods to explore dimensions of human brain representation that were previously inaccessible to researchers.

The human brain is one of the most complex objects in the universe, and the fundamental principles of how it supports intelligent behavior remain unknown. One challenging aspect is that its functions are carried out in high dimensions—they rely on the coordinated interplay of immense high-dimensional populations of neurons. All high-dimensional systems, including the brain, present several challenges to researchers: 1) they are difficult to observe and measure at scale, 2) analyzing data from high-dimensional systems requires major computational resources and advanced statistical methods, and 3) the intuitions we obtain from lower-dimensional systems often fail to generalize to higher dimensions. We believe that these challenges have hampered progress toward a neuroscientific theory of natural intelligence and that, with a new statistical framework and large-scale neuroscience data, we can transform how neuroscientists approach the study of the human brain and its high-dimensional functions.

Neuroscientists have struggled to understand the brain in part because the standard methodological tools of neuroscience are ill-equipped for characterizing statistical phenomena in high dimensions. In fact, the methods of neuroscience primarily focus on low-dimensional problems in small-scale datasets (e.g., identifying whether a specific brain region or neuron responds more to stimuli from category A or category B). As a result, there is a critical gap in our understanding of how the human brain operates in high dimensions. The goal of this proposal is to bridge this knowledge gap by applying a novel statistical technique to massive recordings of human brain activity and behavioral assessments of cognitive abilities. Specifically, we will analyze a publicly available large-scale dataset to test the hypothesis that our statistical approach can reveal new aspects of high-dimensional human brain representations that are inaccessible with conventional methods but may be critical for understanding individual differences in cognition.

The findings from this project have broad implications for understanding the fundamental properties of brain representation that vary across people and underlie differences in sensory and cognitive abilities. This project also has the potential to transform how neuroscientists study individual differences in brain function—opening new opportunities to explore high-dimensional representations and the precise ways in which they are shared or unique across people. Importantly, our findings will provide critical preliminary evidence for a proposal that we are currently developing to study the neural underpinnings of visual intelligence.

We propose to conduct a proof-of-principle analysis showing that individual differences in general intelligence are linked to high-dimensional aspects of human brain representations that would otherwise be undetectable with standard analysis methods. We will do so using a publicly available NIH-funded big-data initiative known as the Human Connectome Project (Van Essen et al., 2013). The Human Connectome Project contains one of the largest datasets in the world for functional neuroimaging of the human brain, and it includes rich behavioral and demographic data, allowing us to precisely characterize the relationship between the functional properties of the human brain and cognitive performance. The PIs, Bonner and Ménard, have recently developed a new statistical approach for characterizing a key functional signal in brain activity that spans thousands of orthogonal dimensions (Gauthaman et al., in prep). In contrast, standard analysis methods only allow researchers to see the tip of the iceberg of meaningful dimensions in human brain activity. We are now seeking to leverage these methodological innovations to make major advances in understanding the nature of diversity in human brains and behavior.

Our approach, which we refer to as cross-decomposition, involves two key procedures that allow us to uncover a high-dimensional spectrum of meaningful signals in the brain:

Hyperalignment: The fundamental dimensions of brain representation are not neurons or cortical patches but rather latent dimensions that need to be aligned across individuals. Our procedure optimally rotates the high-dimensional brain representations of one individual to be functionally aligned with another. By doing so, we can uncover fine-scale information in human brain representations that is undetectable with conventional methods relying on coarse anatomical alignment.

Cross-validated dimensionality spectrum: When using standard methods to characterize the latent dimensions of brain representations, it is impossible to distinguish reliable signal from sources of nuisance variance, such as physiological noise and equipment noise. Our procedure uses cross-validation across repeated presentations of experimental stimuli to characterize dimensionality without corruption from nuisance variance, giving us the first unbiased view of the full spectrum of latent dimensions in human brain representations.

Using this approach, we will characterize the high-dimensional spectrum of brain representations measured with functional neuroimaging, and we will test the prediction that newly detectable fine-scale information in brain activity is tightly connected to individual differences in fluid intelligence, which is the general ability to think abstractly and reason through problems.

SEED FUND PROJECT

2024 IDIES

HYBRID STATISTICAL AI FOR IDENTIFYING NEUROSURGERY PATIENTS IN NEED OF CRITICAL CARE

TAMÁS BUDAVÁRI
&
ROHAN MATHUR

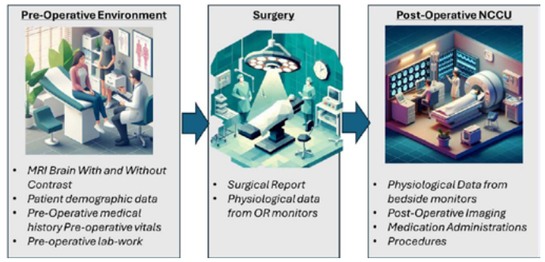


Figure 1. Illustration of the breadth of data collected into the PMCoE-NCC Data Repository at each stage of a patient's care. We will retrospectively analyze intra-operative and post-operative data to determine which patients truly ended up requiring NCCU level of care. These labels will serve as endpoints for our model development. Data from the preoperative environment will be used to build the models discussed in Specific Aims 2 and 3.

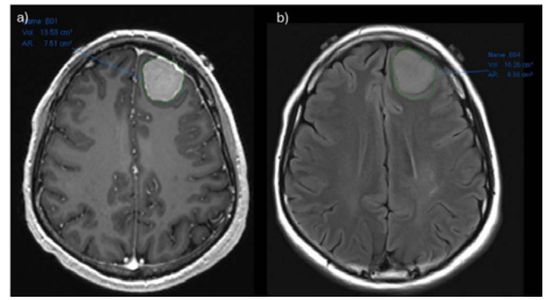


Figure 2. Sample representative MRI Brain sequences for a patient with brain tumor in the pre-operative environment: (a) MRI T1- post contrast sequence demonstrates the segmentation and boundaries of the brain tumor itself; (b) MRI T2 FLAIR sequence that demonstrates the overall extent of the tumor in addition to surrounding brain swelling.

Patients diagnosed with intracranial tumors, after consultation with a neurosurgeon, present to the hospital at a scheduled date for surgical resection of their intracranial tumors. Post-operatively, they are routinely admitted to the Neurosciences Critical Care Unit (NCCU) for continuous monitoring. The problem that motivates our project is that only a small unknown percentage of those patients truly develop medical conditions that require care that can only be provided in the NCCU, whereas most patients would be able to receive necessary care in less resource-intensive facilities in the hospital. Identifying those patients would allow for a more effective utilization of limited resources (i.e., total of 24 beds in NCCU) and better care for those who truly need it, and hence save more lives. Financially, today this is a multimillion-dollar issue even just at the Johns Hopkins Hospital!

We aim to identify the need for critical care of recovering patients by using recent advancement in AI/ML by utilizing the Precision Medicine Center of Excellence in Neurocritical Care (PMCoE-NCC) Data Repository that contains clinical electronic medical records or EMR data, physiological and imaging data from all patients who have been admitted to the NCCU at the Johns Hopkins Hospital. Statistical classification methods are mature approaches that consider uncertainty and provide well calibrated probabilistic results. That said, the recent success of deep learning cannot be overlooked, but their blind application is simply too dangerous! We will combine the power of statistics and deep learning approaches without their drawbacks. The usual statistical methods suffer from naive feature extraction processes, which limit their applicability. For example, image classification is traditionally performed on simple summaries of segmented regions, such as their average colors or brightness, etc. Deep learning, however, is arguably so successful because it starts with the original raw data, e.g., the image itself, and learns to extract optimal features along the way. The problems begin when a fully connected neural network makes blackbox predictions at the end of the feed-forward network architecture. We will design and train neural networks to achieve our objective in a way that the extracted features can also be used in proper probabilistic classification approaches. Using appropriate network architectures, we will identify the layers where important image properties are encoded and use them for statistical analysis. Given that the functional form of a neural network is analytically known, we can even propagate uncertainties in images to the optimal features, which can then be used in probabilistic classification, potentially even combining with other traditional features.

Specific Aim 1. Using the PMCoE-NCC database, we will first identify those patients, admitted after elective brain tumor surgery, that truly needed treatment that can only be administered at the NCCU to separate them from those who could have received care in less resource-intensive facilities. This labeling will be performed in a projection database (SQL Server) already created for our approved IRB.

Specific Aim 2. To establish a baseline, we will first use the usual traditional features available for all patients in EPIC and apply statistical ML methods, such as Bayes classifiers, boosted decision trees, logistic regression, and Generalized Additive Models (GAMs) – all of which provide interpretable and explainable results.

Specific Aim 3. Using available MRI images, we will train specially designed neural networks to classify the patients based on the historically available labeling; see Specific Aim 1. This is a compute- and data intensive task, which will also include optimizing the architecture of the deep learning network. The performance of the neural network classifier will serve as a baseline for comparison to the following statistically explainable AI methods.

Specific Aim 4. Combining the statistical methods with the optimal features extracted by the deep learning network from the MRI images, we are aiming to develop a novel tool for the classification by utilizing a probabilistic outcome of the model.

Patient Selection & Data Sources Inclusion Criteria: (1) adult patients 18 years or older with (2) diagnosis of a brain tumor based on the ICD10 code screening, and (3) elective resection of the tumor, who (4) were directly admitted to the NCCU after their surgery with (5) all relevant information stored in the PMCOE-NCC Data Repository.

A preliminary screening of the PMCoE-NCC Data Repository has identified 2582 patients between July 1, 2018, and August 31,2023 that meet these inclusion criteria.

Screening Criteria for NCCU Needs: For the labeling in Specific Aim 1, consensus-driven criteria are being developed by a team of four neurointensivists, which will include: 1. Administration of specific medications that can only be administered in the NCCU, 2. NCCU specific procedures including intubation; high-flow nasal canula; non-invasive positive pressure ventilation; central venous catheter placement; ventriculostomy management, 3. NCCU specific diagnoses, - 3 - In addition, the time window to screen for post-operative NCCU complications is restricted to the first 72- hours post-op, as complications beyond this time frame may be due to factors other than the surgery.

MRI Data: All pre-operative MRI Brain images are available to our Team via the PMCoE-NCC Data Repository as DICOM images. These are expected to be invaluable inputs, which can be accessed and analyzed via the Databricks Cloud computing Platform, the Johns Hopkins SAFE Desktop and the PMAP Crunchr powered by SciServer.

SEED FUNDING PROJECT

2024 IDIES

PLAYING WITH CHEMICAL LEGO BLOCKS: RATIONAL DESIGN OF SEMICONDUCTING POLYMERS FOR THERMOELECTRIC

PAULETTE CLANCY
WITH CO-PI
HOWARD KATZ

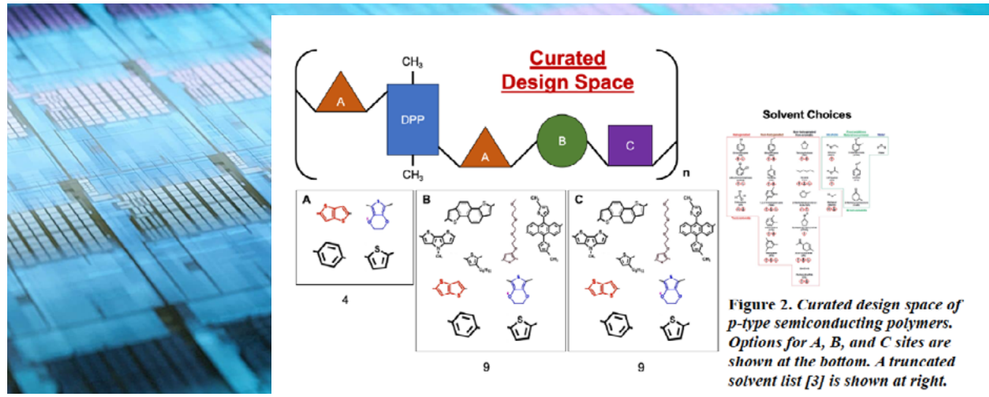


Figure 2. *Curated design space of p-type semiconducting polymers. Options for A, B, and C sites are shown at the bottom. A truncated solvent list [3] is shown at right.*

Alan Heeger and colleagues won the Nobel Prize in 2000 for their 1977 discovery of a conducting polymer.

Since that time, semiconducting polymers have become popular in electronic applications ranging from thermoelectrics and organic solar cells to light-emitting diodes, field-effect transistors and sensors. Given that the palette of organic moieties available to modify conducting polymers is essentially limitless, finding which combination of organic groups will result in better performance is quite an art form. Our co-PI, Howard Katz, has a track record of excelling at that art, producing some of the highest performing, energy efficient thermoelectrics in the world from his experience and chemical intuition.

With the advent of "materials discovery," under which umbrella this proposal falls, we can start to shift from relying on expert human knowledge into a data-centric, AI-guided one. In that vein, this work seeks to identify novel, high-performing polymer candidates using Bayesian optimization over a judiciously curated design space. Bayesian optimization is a well-known machine learning (ML) technique for optimizing global black-box functions. It is especially suitable in data-scarce situations. Much can be achieved with tens of data points rather than the thousands of data points invariably needed for a deep learning neural net approach. Bayesian optimization requires inputs from a training set. Here, we will use first-principles density functional theory (DFT) calculations, which describe the distribution of electron density in a polymer 'repeat unit;' see Fig. 1 for some sample molecules. The target of the optimization is the minimum Gibbs free energy of solvation (ΔGsolv), a metric that describes the strength of a polymer's interaction with surroundingsolvent molecules. This is known to be an indicator of good film quality, but hard to determine directly fromexperiments. If such an ML model can be trained and validated, it opens the opportunity for a data-driven search for disruptively successful polymeric materials, which are notoriously hard to model at the atomic level. Our goal is to create a ML-generated computational model that can inform "human-in-the-loop" experimental organic chemists and materials scientists in a feedback cycle that is significantly more efficient than relying on trial-and-error and expert chemical intuition alone.

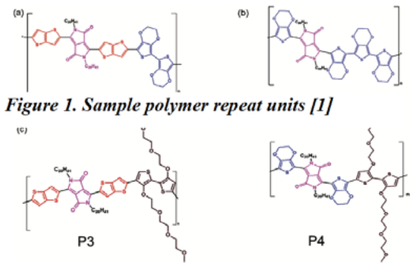


Figure 1. Sample polymer repeat units [1]

Polymers based on the diketopyrrolopyrrole (DPP) motif exhibit successful p-type semiconducting polymers. In those papers, expensive but accurate DFT calculations were used to characterize the electronic properties of repeat units of various DPP-based polymers, which were simultaneously synthesized experimentally and subjected to experimental characterization techniques for thin-film device performance. This is a great starting point for our work; there is a clear similarity between the polymers studied in these works (e.g., Fig. 1) and our interests; the polymers vary only in their subunits. However, there are "infinitely" more functional groups, repeat units, monomer lengths, and side-chains that could have been investigated, even if the design space was restricted to include a DPP unit in every polymer. Further, there are numerous solvent choices (eco-friendly or otherwise) for solution processing these polymers, vastly increasing the complexity of the design space. The problem of investigating such a large space of polymer compositions and available solvents very quickly becomes intractable, whether experimentally or computationally.

The goal of this work, therefore, will be to create hypothetical DPP-based polymers, made by combining 'building blocks' (Lego pieces) of functional groups similar to those studied in Refs. 1-3, and then training a Bayesian optimization model using DFT calculations as a means to identify high-performing polymer/solvent compositions. We are, in essence, putting Lego pieces together with different characteristics and determining if they lead to an improved design or not. The role of Bayesian optimization here is to sort through the limitless Lego pieces in our box.

A major assumption of this work is that a higher magnitude ΔGsolv will yield a higher-performing polymer because of superior thin film morphology brought about by favorable solute-solvent interactions. This assumption has its foundations in previous work done both by us and others, demonstrating the impact of solution-phase chemistry on end-device thin film uniformity. Our end-goal is to create a multiscale model that correlates DFT results to end-device performance, an elusive task across materials research.

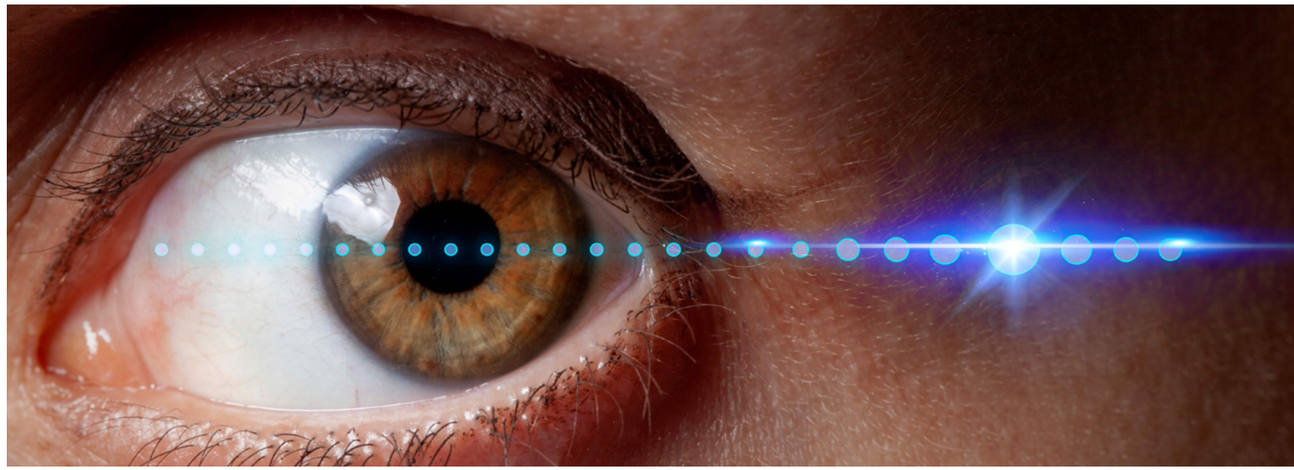
Our design strategy is summarized in Fig. 2, where we represent the otherwise dauntingly complex polymer chains as segments of different types of moieties (labeled A, B, C in Fig. 2) that typically appear in sequence in real conducting polymers. This pares the taxonomy down to a manageable set of categories, with a list of potential candidate moieties in A, B and C, as shown in Fig. 2.

We have considered 8,424 possible combinations (4x9x9x26) for A x B x C x solvent choices. A polymer repeat unit is created wherein one of each of the A, B, and [CONTINUED ON PAGE 40]

A DIFFUSION PROBABILISTIC FRAMEWORK FOR PATIENT-INDEPENDENT EYE MOVEMENT VIDEO GENERATION

SEED FUNDING PROJECT
2024 IDIES

KEMAR GREEN, MD
WITH CO-I, VISHAL PATEL



Eye movement pathways are affected in various conditions owing to the widespread brain areas represented by these neural connections. Changes in the function of various eye movement types can provide accurate information about the location of the brain abnormality, and in the right context, the neurologic diagnosis. Eye movements are therefore beneficial in providing physiologic data for various neurologic illnesses.

While the pattern of eye movements in various neurologic and psychiatric disorders is well described, some of these diseases are very rare, and the existence of robust eye movement disease datasets are equally rare or non-existent – making it difficult to develop reliable deep learning classification systems. In addition to data scarcity, data security also poses a challenge in the use of eye movement video data, as ocular features as well as eye movements bear biometric information. A possible solution to the data scarcity issues is the use of synthetic data. Synthetic eye tracking videos has been simulated for non-medical purposes, and have been shown to mimic human ocular motor dynamics.

We have shown in previous studies that image recognition models can successfully detect torsional eye movements from artificially generated retinal photos. It has also been shown that subtle abnormal eye movements can identified via clinician-guided telemedicine evaluation and with deep learning video classifiers from video-oculography recordings of varying qualities. Even though the technology exists for robust video generation, and synthetic medical data usage in artificial intelligence (AI) have been shown to increase model performance metrics.

We hypothesize that novel generative AI techniques can be adopted for generating synthetic neurologic disease specific eye movement datasets that can then be used to develop deep learning-based brain disease phenotypic system.

During the funding cycle, we will focus on eye movement signatures of myasthenia gravis (MG), a rare neurologic disease, that causes muscle fatigue with repetitive muscle use. Several eye movement types (smooth pursuit, saccades, OKN, etc.) have been implicated in MG4. During the funding cycle, we will focus on developing of optokinetic nystagmus (OKN) eye movement signatures based on a recently published method. First, however, due to the rarity the disease and available eye movement datasets, coupled with privacy concerns surrounding the use data containing biometric information, it is necessary to augment the available clinical data (being collected simultaneously from neurologic patients) by generating realistic synthetic eye movement datasets of myasthenia gravis using a pose-guided video generation framework. In this framework, the pose will be obtained from 1-5% of the collected patient waveform data, and the videos will be generated from validated open-source eye movement databanks. Our methodology leverages the efficacy of video diffusion models, traditionally employed for generating high-quality, short videos (Fig. 1). Our initial phase employs a model proficient in generating videos from segmented eye movement masks. At the heart of this process is a latent video diffusion mechanism that translates segmented mask inputs into visual sequences.

We will then develop multimodal myasthenia gravis deep learning classifiers using imaging and waveform data (Fig.1). Multimodal data (raw videos, filtered images and extracted OKN time series data) will be used to build a fusion model by adopting existing methods. The approach is based on long short-term memory (LSTM) network that is capable of processing video clips and time-series data representing OKN. The multi-modal representation will be generated as a sequence of uni-modal representations (or tokens), such that the fusion module aggregates these representations through the recurrence mechanism of LSTM. The use of multimodal data in our case will allow for better understanding of the ocular motor biomarkers of OKN from videos and waveforms. The data being extracted from the OKN fatigue tasks (video, filtered image, waveform, etc.,) have different signatures for MG that individually might not provide the most accurate diagnosis; therefore, the combination of different data types using the multimodal deep learning approach will allow for the development of the most robust model. Furthermore, the multimodal data approach will address the potential noise that can occur in the eye movement data, as well as the variability of the eye movement signs in MG. The models will be validated on the remaining 95-99% of the real MG data as well as a commensurate value of normal patients to assess the generalizability of the synthetic model. Various explainable AI methods will be applied to assess the model's prediction and unveil any novel "non"-physiologic. If successful, we will apply similar methods to curate similar datasets for various brain diseases such as Parkinson's disease, Alzheimer's disease, Huntington's disease, progressive supranuclear palsy (PSP), multiple system atrophy (MSA), etc. [CONTINUED ON PAGE 40]

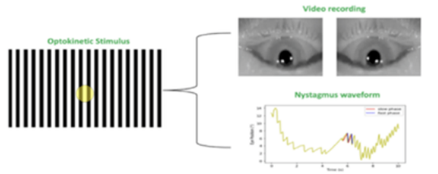


Figure 1 Example of an optokinetic nystagmus (OKN) stimulus and binocular eye movement (screenshot) and corresponding nystagmus waveform.

THE GLOBAL ELECTIONS DASHBOARD: HARNESSING DATA SCIENCE TO ENHANCE ELECTORAL OVERSIGHT

SEED FUNDING PROJECT

2024 IDIES

ADAM SHEINGATE WITH CO-I, TOM LIPPINCOTT

Efforts to weaken or undermine the electoral system are one of several existential threats to democracies around the world. Effective electoral oversight is therefore essential if citizens are to perceive elections as both free and fair and is key to the survival of the democratic project. The application of AI and data science tools can help deliver effective and trusted elections, advance social science research, and inform the public about campaigns and elections. For example, natural language processing can automate the classification of millions of campaign records in ways that facilitate election monitoring and scholarly analysis. The development of multi-lingual NLP models makes it possible to extend these tools to the analysis of global election spending and text-based descriptions of expenditures in different languages.

To that end, we are using our IDIES seed grant to explore how data science tools can assist in the collection and dissemination of campaign finance and political spending data from around the world. In June, we organized a workshop on "The Global Elections Dashboard: Harnessing Data Science to Enhance Electoral Oversight" at the JHU Bloomberg Center in DC. The meeting brought together faculty from both the Krieger and Whiting schools along with data journalists and transparency advocates from the US, UK, and Brazil. We discussed a variety of approaches and tools that could assist with the analysis of campaign data such as large language models, machine-learning classifiers, optical character recognition, and entity disambiguation among others. We also discussed how a global elections dashboard that enables users to visualize and analyze campaign data could serve various stakeholders including election regulators, transparency advocates, and academic researchers.

Currently, a team of research assistants is assessing the availability of campaign finance data in consolidated democracies around the world. Researchers will determine whether the election management body in each country has bulk data available for download or through an API, whether the data is in a language other than English, the type of data available (contributions or spending), and for which years. From this survey, we hope to identify a subset of candidate countries with sufficient data available to serve as the basis for a second meeting planned for early 2025. At this next meeting, we will focus on identifying specific tools and tasks needed to produce a unified dataset to populate the dashboard such as dataset linkages, data dictionaries, and multi-language models.

Working with colleagues in the UK and the International Institute of Democracy and Electoral Assistance (International IDEA), we are also exploring ways to develop a global standard for the collection and reporting of campaign finance data. We hope to identify a set of best practices for data on political contributions, categories of political spending, and vendors who provide products and services to parties and candidates. Standardized reporting will improve the transparency and accountability of electoral processes while also making it easier for social scientists, journalists, transparency advocates, and citizens to track the flow of money in politics.

Finally, the 2024 U.S. Election affords us an opportunity to test some of our ideas about the analysis and visualization of campaign spending data. Using a supervised machine-learning model, we are able to classify millions of individual expenditure records into categories such as media, fundraising, or digital advertising. We can then compare these categories of spending over time and across election cycles as well examine differences in spending by political party or political office. Figure 1 shows a screenshot comparing media spending by party in campaigns for the U.S. House of Representatives in 2020 and 2022. We will continue to update the dashboard as more data becomes available for the 2024 election cycle. We will also add additional features to the dashboard that will enable users to examine spending by individual candidates, specific campaign committees, or particular states or congressional districts. You can explore the U.S. dashboard for yourself at https://election-spending-data.shinyapps.io/dashboard/.

SUMMER STUDENT FELLOWSHIP PROGRAM

2024 IDIES

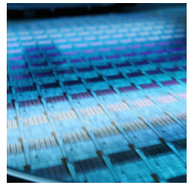
The IDIES **Summer Student Fellowship (SSF) program** provides $6,000 to support undergraduate students from all divisions of JHU for summer research projects in data science. Students work full-time for ten weeks with guidance from an IDIES faculty mentor. Projects can address any research question – from any field of science, social science, or humanities – that can benefit from big datasets and the techniques used to query and analyze those datasets.

Prospective fellows complete a short proposal by the end of January, explaining the goals and objectives of the project and an outline of possible methods, along with a CV and a recommendation from the proposed faculty mentor. Research is not restricted the student's major area of study, and working with faculty mentors in different departments and schools is encouraged. All proposals are evaluated by a committee of IDIES researchers for intellectual merit and broader impact.

Five students were chosen as IDIES Summer Student Fellowship recipients for summer 2024. This year's recipients are shown here, first by their project title, and then with additional information on the student and their project.

Two SSF final reports were chosen to include in full in this year's review—that of Akshaya Ajith (page 34) and David Zhou (page 36).

2024 SSF PROJECTS



Accelerated Materials Discovery for High-Entropy Hydrogen Fuel Cell Catalysts (HE-FCC)

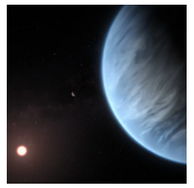
AKSHAYA AJITH

Advancing Otolith Function Assessment: Integrating Machine Learning with Video Oculography for Enhanced Vestibular Diagnosis

KRISHNA MUKUNDA

Quantifying and Visualizing Urban Emissions of US Cities

MILLIE MI



Computational Study on Accretion Modes of Planet Formation and Their Impact on Exoplanet Population

CHRIS LE WANG

Simulating Murine and Human Cerebrocortical Development through an Agent-based Computational Model

DAVID ZHOU

2024 SUMMER STUDENT FELLOWSHIP (SSF) RECIPIENTS

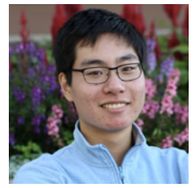
IDIES IS EXCITED TO SUPPORT THE FOLLOWING STUDENTS* AND THEIR WORK IN 2024

AKSHAYA AJITH is a first-year student pursuing a double major in Materials Science & Engineering and Computer Science. She is enthusiastic about the applications of machine learning in accelerating the discovery of materials geared towards renewable energy applications. Originally from Washington state, she enjoys reading and hiking in her free time.

KRISHNA MUKUNDA is a third-year Biomedical Engineering student from Cleveland, OH. His interests lie in integrating deep learning and neuroscience to enhance medical diagnostics. Post-graduation, he aspires to become a doctor, leveraging his engineering background to advance patient care.

CHRIS LE WANG is a third-year computer science and physics major.

MILLIE MI is a first-year student majoring in environmental engineering with an interest in urban sustainability. In her free time, she is in an a capella group on campus, and enjoys volunteering at conservation centers.



DAVID ZHOU is a second-year undergraduate majoring in Neuroscience and Applied Mathematics & Statistics. He is interested in investigating systems and cellular neuroscience, particularly through a hybrid computational-experimental approach.

Accelerated Materials Discovery for High-Entropy Hydrogen Fuel Cell Catalysts (HE-FCC)

AKSHAYA AJITH
WITH PROF. COREY OSES

Hydrogen fuel cells offer a path to sustainable, low-carbon electricity generation. However, their benefits are undermined by the expense of acquiring platinum-based catalysts required to facilitate the reaction. Ongoing development of cheaper substitute catalysts faces challenges regarding inferior catalytic activity – but recent studies have suggested that materials with higher disorder maintain higher catalytic activity.

This project seeks to identify cost-effective High-Entropy Hydrogen Fuel Cell Catalysts (HE-FCCs) that can substitute for platinum-based catalysts. We constructed a machine learning algorithm to generate a list of ideal high-entropy catalysts, which will be experimentally verified. This algorithm will accelerate the discovery of potential platinum catalyst substitutes.

Advancing Otolith Function Assessment: Integrating Machine Learning with Video Oculography for Enhanced Vestibular Diagnosis

KRISHNA MUKUNDA
WITH PROF. KEMAR GREEN

Vertigo has a lifetime prevalence of around 7.4%, and is significantly difficult to diagnose. The presence or absence of torsional nystagmus can allow for quick diagnosis and prevent unnecessary medical testing. Our proposal aims to develop and explain a deep learning model for torsional identification from video oculography (VOG) recordings.

This approach could potentially reveal new clinical markers of torsion, improving the accessibility and accuracy of vestibular diagnostics. The model will differentiate between torsional nystagmus, static torsion, and non-torsion. Additionally, we will create a synthetic dataset using generative AI methods to develop an open dataset for broader research use.

Computational Study on Accretion Modes of Planet Formation and Their Impact on Exoplanet Population

CHRIS LE WANG
WITH PROF. KEVIN SCHLAUFMAN

The relative importance of pebble versus planetesimal accretion presents the most important uncertainty in current planet formation models. Since pebble accretion has much lower efficiency than planetesimal accretion, we expect that planets formed in circumstellar disks with little planet-making materials available cannot be formed via pebble accretion.

We conducted a large suite of computational planet formation experiments to test this hypothesis by examining the impact of varying pebble accretion efficiency on the structure of fully-grown planetary systems. We compared our simulated systems to observed exoplanet systems to identify a subset of systems that could not have been formed by pebble accretion alone. Our results will be a crucial step forward towards demystifying the formation history of the diverse population of exoplanets observed.

Quantifying and Visualizing Urban Emissions of US Cities

MILLIE MI
WITH DYLAN GAETA

Many US cities have set greenhouse gas (GHG) emissions goals, making it necessary for them to understand their emissions. However, the existing GHG inventories are incomprehensive or lack data disaggregated by city. While some cities have created their own inventories using one or more of the existing protocols, the inventories are decentralized, which makes it challenging to compare GHG emissions between different cities.

We aim to create a standardized database of GHG emissions estimates from cities by scraping the web for publicly-available GHG inventory data. Through visualizations, the data will be used to compare cities' GHG inventories and their progress on their GHG emissions goals. Additionally, we plan to identify best practices in emissions accounting based on the range of practices we encounter while creating the database.

Simulating Murine and Human Cerebrocortical Development through an Agent-based Computational Model

DAVID ZHOU
WITH PROF. GENEVIEVE STEIN-O'BRIEN

Development of the mammalian cerebral cortex is a complex and highly-regulated process requiring precise coordination of biomolecular pathways. Dysregulation of those pathways yields a mature cortex abnormal in both structure and function, correlating with an array of neurodevelopmental disorders. Research on cerebrocortical development is valuable to underlying such disorders. However, current in vivo methods are often hindered by concerns relating to time, cost, and ethics.

To address this, we developed a computational model of the process of corticogenesis at a cellular and molecular level, with parameters obtained via spatial transcriptomic data from developing mice. Our model allows researchers to conduct corticogenic studies in silico, offering an alternative to in vivo experimentation and its associated challenges.

**Recipients' class year is as of the IDIES spring funding announcement in 2024*

SSF FUNDING PROJECT

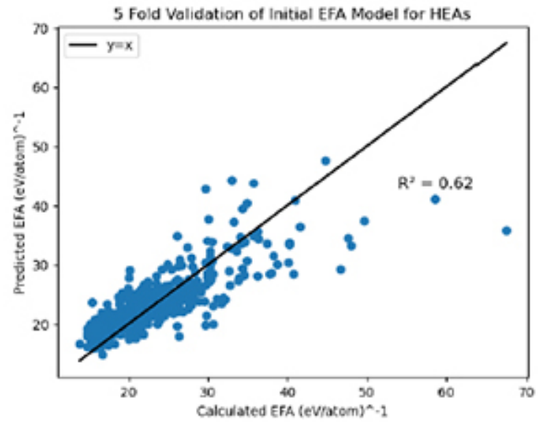
2024 IDIES

ACCELERATED MATERIALS DISCOVERY FOR HIGH-ENTROPY HYDROGEN FUEL CELL CATALYSTS (HE-FCC)

AKSHAYA AJITH
WITH MENTOR
PROF. COREY OSES

Hydrogen fuel cells are a promising clean energy alternative to traditional electricity generation methods. They are more efficient than their traditional counterparts and produce no carbon emissions, outputting only water. However, achieving large-scale adoption across industries requires a cost reduction. Most hydrogen fuel cells utilize a platinum catalyst to facilitate the electrochemical reactions, and platinum's scarcity contributes to its high cost – a notable 42% of the cost of producing a hydrogen fuel cell can be attributed to the platinum catalyst.

Developing a non-precious metal catalyst would cut costs and promote the implementation of hydrogen fuel cells in more industries. Recent studies have suggested that materials with higher disorder from additional components see an increase in catalytic performance and durability. This project seeks to develop a predictive machine learning algorithm that identifies cost-effective High-Entropy Alloys (HEAs) that can substitute for platinum-based catalysts.



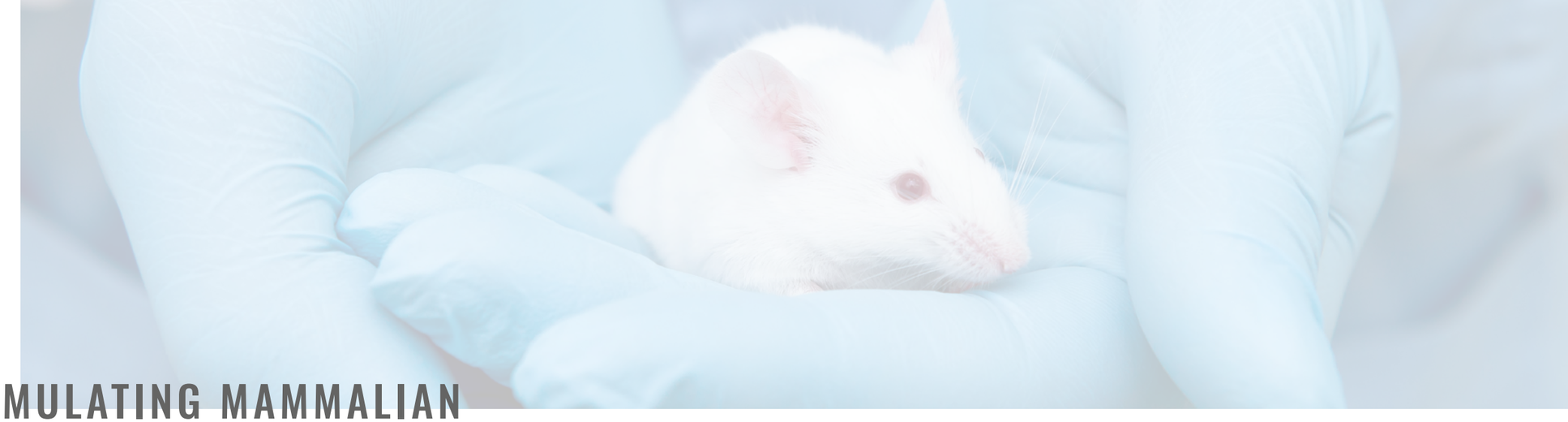
Left: Regression analysis of the calculated EFA for 436 High Entropy Alloys against the original values predicted by the Random Forest model. The initial model achieved R2 = 0.62.

Our initial efforts focused on creating an optimization algorithm to accelerate the discovery of stable HEAs using the AFLOW database. The aflow-POCC framework allows us to estimate the stability of HEAs by analyzing their Entropy-Forming Ability (EFA), with higher EFA values indicating greater stability based on Density Functional Theory (DFT) calculations.

To develop an EFA prediction model, we employed a random forest algorithm trained on 436 HEAs derived from a database of five-component metal alloys. The model utilized 20 parameters describing material properties, such as ionic character, melting temperature, and electronegativity. A five-fold cross-validation yielded an initial R^2 value of 0.57 on the testing set for the model's predictions. By incorporating 12 additional parameters, we improved the R^2 to 0.62.

To improve the R^2 further, the project turned to identifying parameters with higher feature importance for EFA prediction of HEAs. We developed a feature selection program to generate feature vectors based on parameters calculated from JSON files containing the molecular information of each HEA in our training set. The goal was to identify parameters with high predictive power and eliminate those with low variance. This process identified 415 additional parameters of interest, which were subsequently incorporated into the model. Training the random forest model with the expanded set of 415 features and on all 436 HEAs resulted in a R2 value of 0.58.

The project's immediate objective is to enhance the predictive accuracy (R^2) of the model by refining feature selection and dividing the heterogeneous HEA dataset into specialized subgroups. This specialization will reduce overfitting and increase the model's accuracy in predicting the stability and performance of different compound types.



SSF FUNDING PROJECT

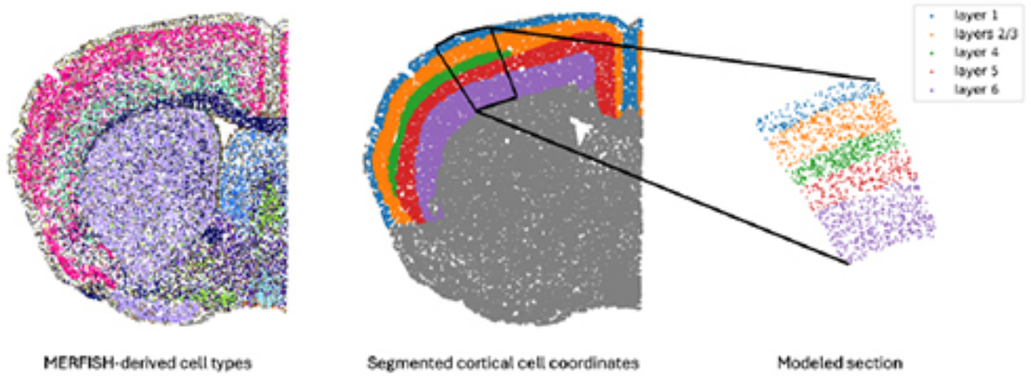
2024 IDIES

SIMULATING MAMMALIAN CEREBROCORTICAL DEVELOPMENT THROUGH AN AGENT-BASED COMPUTATIONAL MODEL

DAVID ZHOU
WITH MENTOR
PROF. GENEVIEVE STEIN-O'BRIEN

This project sought to further develop and refine a computational model of murine cerebral cortex development using PhysiCell, an agent-based modeling software. We had three main goals: to refine computational mechanisms for simulating key cellular behaviors; to encode additional behaviors to enhance biological accuracy; and to tune model parameters using in situ spatial cellular data.

First, we developed a new asymmetric division model where progenitors directly produce daughters of distinct cell types, more accurately representing what actually occurs in a developing cortex. Our model directly computes neuronal fate based on local environmental signals, generating an instance of that neuron type along with a regenerated progenitor when division occurs. Second, we selected several new behaviors that exhibit strong influence on cerebrocortical development, and created parameterizable models for each characteristic. Third, we developed a parameter tuning pipeline using metrics extracted from spatial murine cerebrocortical data. Specifically, we sought to reproduce layer-specific cell count and density metrics through variation of a selected subset of model parameters. Using a Nelder-Mead algorithm, we successfully tuned the model using data from distinct regions of the murine cortex to obtain different sets of parameter values. By comparing our models, we examined how different cortical regions control the timing of different developmental cues in neurogenesis.

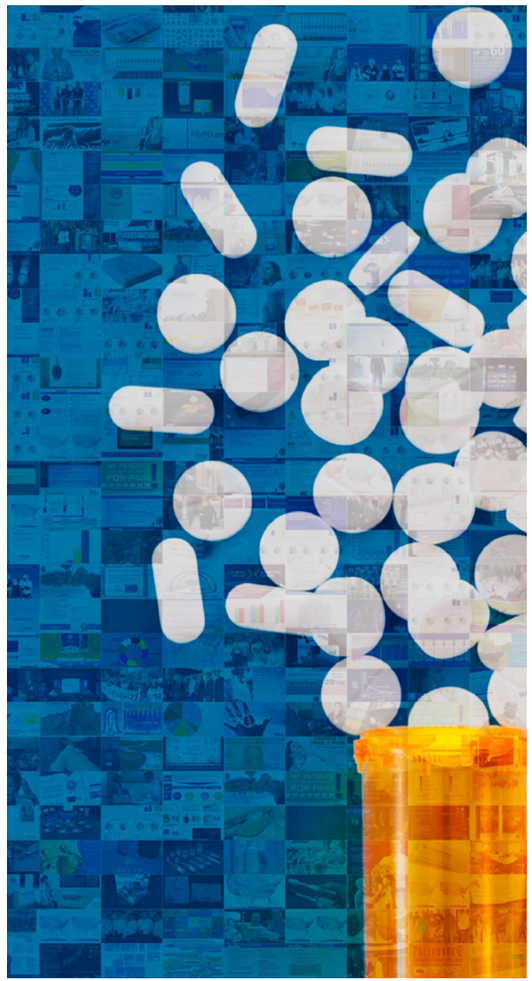


Above: Computational modeling of cerebrocortical development. Data used for parameter optimization originates from MERFISH spatial data from the murine brain. Here, a cortical column from the somatosensory cortex was selected and layer-specific cell counts were extracted, then used to tune our model.

Our current ongoing efforts focus on bringing the model closer to biological reality by implementing additional smaller-scale behaviors such as Notch signaling and spatial gradients of competence factors. We hope to eventually obtain human developmental data and use it to extrapolate our murine model to the real human cerebral cortex.

SUPPORTING USE OF COMPUTATIONAL METHODS ON THE OPIOID INDUSTRY DOCUMENTS

KEVIN S. HAWKINS, JHU PROGRAM DIRECTOR
UCSF-JHU OPIOID INDUSTRY DOCUMENTS ARCHIVE



The Opioid Industry Documents Archive collects, organizes, preserves, and provides free online access to millions of previously internal documents made public through legal settlements to enable multiple audiences to explore and investigate information which shines a light on the opioid crisis. This groundbreaking digital archive of opioid industry documents advances understanding of the root causes of the U.S. opioid epidemic, promotes transparency and accountability, and informs and enables evidence-based research and investigation to protect and improve public health. For example, materials related to pharmaceutical distributors speak to the methods that they used to monitor the opioid supply chain, and the degree to which indicators of potential high-risk opioid distribution were acted upon. Policy analyses might examine how manufacturers engaged with advocacy organizations to achieve their policy objectives and strategies that manufacturers may have used to respond to regulatory concerns regarding opioid safety. The varied nature of the documents, which include corporate e-mail chains and internal company documents in connection with brochures and pamphlets, allow researchers to compare internal marketing strategies against the claims of safety and due diligence presented to practitioners and regulatory bodies. Because the litigation has focused on abatement of the opioid crisis, the documents also contain extensive information regarding how to best prevent further harms, and at what cost.

It currently includes over 3.4 million documents, with millions more expected to be added in the coming years. When dealing with such a massive corpus of documents, searching and browsing a digital library interface isn't always sufficient to find what you need. Therefore, the OIDA team at JHU has created the OIDA Toolbox, which provides a set of options for accessing the raw data behind the archive in order to support the needs of those who wish to use computational methods to study this invaluable corpus of data.

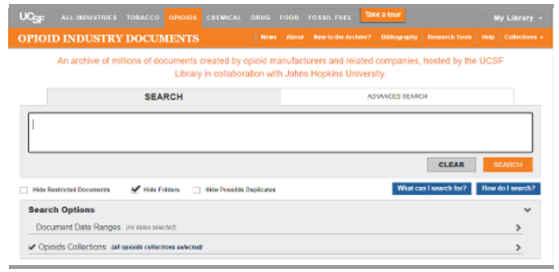
A key tool in the OIDA Toolbox is using SciServer to work with the OIDA corpus. Thanks to the generous support of the SciServer team,

OIDA's raw data is available in IDIES storage, and we have developed a few Jupyter notebooks to allow users to quickly get started on accessing and analyzing OIDA documents in ways that you couldn't do through OIDA's main digital library interface.

We look forward to providing more through SciServer as we learn more about the ways in which researchers and the general public are interested in exploring and studying the documentary evidence of the role that companies played in the U.S. opioid epidemic!

https://oida-resources.jhu.edu/

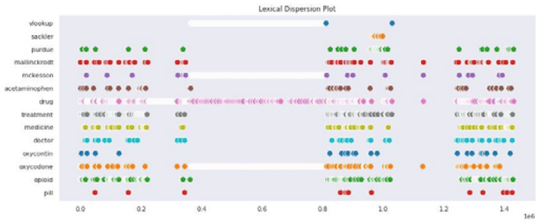
https://www.industrydocuments.ucsf.edu/opioids



Above: OIDA's main digital library interface allows for searching and browsing documents based on the full text and metadata.

Right: The OIDA Toolbox documentation includes this matrix overview of the features of the tools including SciServer.

Below: A lexical dispersion plot for some common terms in OIDA documents based on a sample of 1,000 documents. Note that "sackler" was not used over time similarly to the other terms.



FEATURE	SCISERVER	PREPARED DATASETS FROM IDL	AWS	INTERNET ARCHIVE	IDL SOLR API
Query metadata and full text	Yes	–	Yes (using Parquet)	–	Yes
Download metadata	Yes	Whole collections	Yes (in Parquet format)	Yes	Yes
Download full text	Yes	Whole collections	Yes	–	–
Download native formats and page images	Yes except audio and video	No	Yes except audio and video	Audio and video	–
Sample notebooks/code	Available: showing flexible query via Solr API and bulk downloading, plus basic text analysis and visualization	–	–	–	Access and sample code provided
Environment	Cloud-based (everything included)	–	Bring your own cloud instance	Command line or Python	API-based
File system access	Yes	–	Yes	–	–
Metadata formats	Parquet (plus other formats through Solr API)	–	Parquet only	XML and SQLite	XML, JSON, Python, Ruby, PHP, and CSV
"Batteries included"	High (cloud-based, ready to use)	Medium (prepared datasets)	Low (requires own setup)	Medium (basic technologies)	Medium (API access)

A DIFFUSION PROBABILISTIC FRAMEWORK FOR PATIENT-INDEPENDENT EYE MOVEMENT VIDEO GENERATION CONTINUED

[CONTINUED FROM PAGE 27] The IDIES-supported project has made significant strides in overcoming the challenges of data scarcity and privacy concerns in eye movement research, particularly around spontaneous eye movement classification and Myasthenia Gravis diagnosis from optokinetic nystagmus (OKN) eye movement (Figure 1). Developing deep learning models for detecting and characterizing these specific eye movements has long been hindered by the limited availability of publicly accessible data, as eye movement patterns can uniquely identify individuals. To address this, we leverage a diffusion probabilistic framework for generating synthetic, patient-independent, pose-guided eye movement videos.

This framework simulates clinically relevant eye movement patterns by generating long, realistic videos that mimic various waveforms (Figure 2). These synthetic videos are produced without the use of real patient data, utilizing publicly available datasets as a foundation. This approach allows us to create extensive datasets that maintain privacy while providing a rich source of data for training and validating AI models.

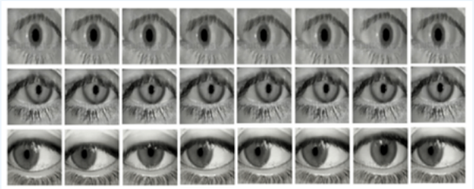


Figure 2 Samples of extracted frames from nystagmus synthetic eye movement dataset

The effectiveness of these generated videos has been validated through their application in downstream tasks, using real patient datasets for comparison. Results demonstrate that models trained on these synthetic videos perform comparably to those trained on real data, proving the viability of synthetic datasets for eye movement research. This breakthrough enables us to scale our research efforts while ensuring patient privacy is uncompromised.

In addition to analyzing existing eye movement patterns, we are expanding the range of eye movement types covered in our research. Our aim is to develop disease-specific biomarkers for each type, leveraging synthetic data to create targeted classifiers for a range of neurologic conditions (Figure 3).

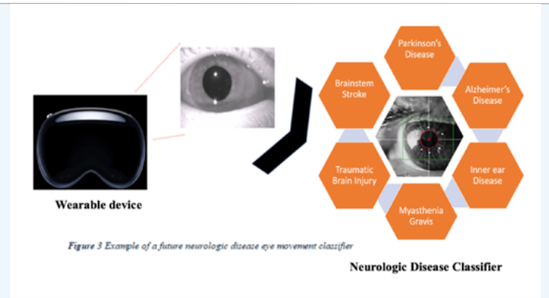


Figure 3 Example of a future neurologic disease eye movement classifier

Looking ahead, our goal is to make these synthetic eye movement videos, disease-specific biomarkers, and classifiers available to the public research community, encouraging further exploration and analysis in the field. Additionally, our next steps involve integrating these synthetic datasets into a wearable device capable of real-time monitoring and therapy adjustments, bridging the gap between research and clinical application.

PLAYING WITH CHEMICAL LEGO BLOCKS CONTINUED

[CONTINUED FROM PAGE 25] C, as shown in Fig. 2., and the sequence is immersed in an implicit solvation model of one of the solvent choices before the polymer/solvent system undergoes DFT relaxation. From the DFT results, we can characterize aspects of the polymer repeat unit's electron density that relate to solvation and are theoretically consistent with solubility metrics like, the Hansen and Hildebrand solubility parameters, which are influenced by the system's electron density. These are the model's inputs. We can then directly calculate ΔGsolv using a Solvation Model Density framework. The Bayesian -1 optimization model's inputs and target value are readily available experimental properties: isotropic quadrupole moment, isotropic polarizability, electrostatic potential minimum and maximum, polarity index, percentage of the surface area that is polar, and the dielectric constant.

We are using our in-house Bayesian code, called PAL 2.0, for this task. It allows a user, whose expertise with these systems may range from amateur to expert, to provide a posited (uncurated) list of possible property data (like the list above) that PAL 2.0 can use to create a chemistry-informed surrogate model. We have used PAL 2.0 with success for chemical systems in solution, for space actuators and orthopedic screw alloy choices, which gives us the confidence to attack the high-dimensional parameter space of this project.

So far, we have conducted DFT sampling of 7000 of the 8,424-candidate design space, to provide the "right answer" for the most favorable (most negative) free energy of solvation. The distribution of ΔGsolv values shown for each polymer/solvent choice is shown in Fig. 3. Based on the assumption that this is the right 'metric of success,' we are at the point that we can, if funded, perform exploratory experiments to confirm that the top choices are high-performers. If not, then our assumption is incorrect, and we will look for a different target (objective function).

Looking at the results in Fig. 3, the range of ΔGsolv values we calculated covers differences amongst candidate polymer/solvent pairings of up to almost 3.0 eV. This a large range, and hence highly discriminatory. This is encouraging from the point of view of candidate triage. Preliminary analysis suggests that the top performers (1%) in the training set generally exhibit the highest magnitude electrostatic potential extrema, polarizabilities, and quadrupoles, which is consistent with the solvation theories discussed earlier—the more the polymer's electron density can distort to accommodate solvent near neighbors, the greater the free energy of solvation. Further analysis will reveal, in greater detail, trends which are not readily apparent; indeed, this is the strength of AI and ML. Symbolic regression will then look to ground these trends in physical realizability.

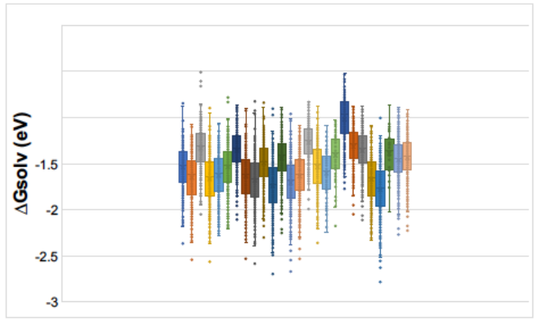


Figure 3. Box plot of exhaustive sampling of ΔG_{solv} vs. solvent choice. Each column represents a different solvent, and each point the solvation energy of a given polymer in that solvent.

Our planned research will take a small portion of this very extensive data set and use Bayesian optimization to predict the polymer/solvent combination with the lowest value of ΔGsolv and compare it to the known best result. Further, we can use the fact that we have the best result for other investigations:

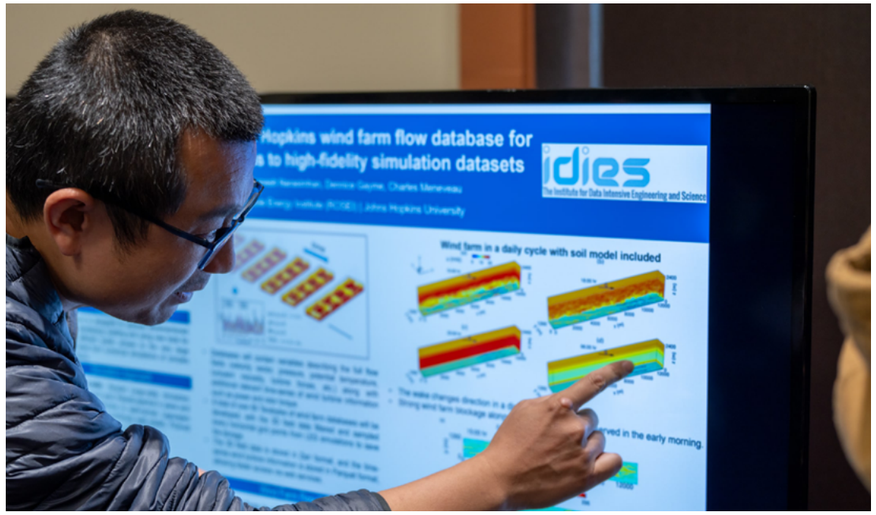
How few data can we use and still find the best result?

What can we learn from "feature engineering" to find the properties that correlate best with the data?

Is there a symbolically-regressed, closed-form function we can identify that provides physical interpretation and realizability to our Bayesian "gray box" model?

This project represents the first time that Bayesian optimization is used to rationally design semiconducting polymers with atomic-scale granularity using chemical functional group building blocks. If our guess at the target being the lowest ΔGsolv is not representative of a high performer, we will harness our expertise in "feature engineering" to find a better metric. Once we make headway with this, it will open the door to considering a more complicated model for polymer design. Possible funding sources will include several NSF divisions CMMI (Civil Mechanical and Manufacturing), Future Manufacturing, DMR (Materials Research) and Chemistry. Our last IDIES seed proposal, a joint WSE/APL collaboration, produced two papers in high-quality journals and an NSF EAGER award. A full proposal to NSF CMMI is planned once we have preliminary data from this IDIES proposal.

POSTER GALLERY COMPETITION



Above: 2024 Poster Gallery contest winner, Shuolin Xiao, expanding on a figure in his digital poster; PHOTO CREDIT: Alec Zabrecky

Every year, IDIES members of all levels—faculty, associate, student, and alumni—have the opportunity to present data-intensive research in a competitive poster gallery format.

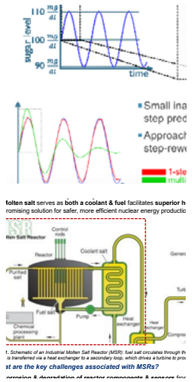
Posters reflect the IDIES mission (https://www.idies.jhu.edu/who-we-are) of furthering data-intensive and computationally intensive research and education. The area of research may be in any discipline that would be of interest to IDIES members and other attendees as long as the methodology or analysis pertains to the use of big data.

IDIES judges score posters based on scientific merit, scientific content, and poster display and organization, and then select a winner of the year's poster gallery.

This year's poster gallery contest winner was Shuolin Xiao for his poster *JHTDB-wind: The Johns Hopkins Wind Farm Flow Database for Open Community Access to High-fidelity Simulation Datasets* (page 52).

2024 IDIES ANNUAL SYMPOSIUM

2024 POSTERS



Lookahead-Driven Inference of Networked Operators for Continuous Stability (LINOCS) Across Various Statistical Models

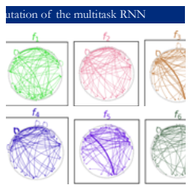
NOGA MUDRIK, EVA YEZERETS, YENHO CHEN, CHRISTOPHER ROZELL, ADAM CHARLES

Development of a Machine Learning-Based Predictive Modeling Framework for Molten Salt Reactor Sensor Data Analysis

OMOBOLADE A. ODEDOYIN, RAMON K. YOSHIURA, BENJAMIN D. GRIMMER

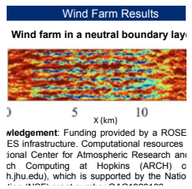
Building an AI for the Study of World-Historical Patterns of Social Protest

BEVERLY J. SILVER AND HAOHANG GUO



Unveiling Latent RNN Dynamics by Leveraging Decomposed Linear Dynamical Systems

DANIEL WANG, NOGA MUDRIK, ADAM CHARLES



JHTDB-wind: The Johns Hopkins Wind Farm Flow Database for Open Community Access to High-fidelity Simulation Datasets

SHUOLIN XIAO, XIAOWEI ZHU, GHANESH NARASIMHAN, DENNICE GAYME, CHARLES MENEVEAU

POSTER GALLERY ENTRIES

Lookahead-driven Inference of Networked Operators for Continuous Stability (LINOCS) across Various Dynamical Models

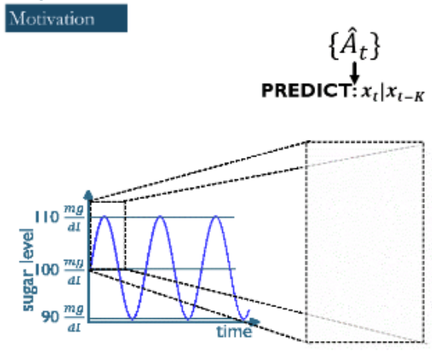
Noga Mudrik[1*], Eva Yezerets[1], Yenho Chen[2], Christopher Rozell[2], Adam Charles[1]
Presenter: nmudrik1@jhu.edu; Corresponding: adamsc@jhu.edu
[1] Biomedical Engineering, Kavli NDI, Johns Hopkins [2] Biomedical Engineering, Georgia Institute of Technology

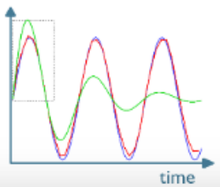
OUR BACKGROUND PAPER | PAPER (NOW AT TMLR!) | CODE | TUTORIAL | VIDEO

Introduction

- Identifying hidden interactions in complex systems is key to understanding dynamic (e.g., neuron interactions).
- These systems display noisy, high-dimensional, and non-stationary behavior, making interaction identification difficult.
- Existing methods capture short-term behavior but fail with long-term trends, indicating incomplete models, while models that integrate long-term trends present limited interpretability.
- We propose LINOCS, leveraging multi-step predictions with adaptive weights into the system identification procedure.
- LINOCS capture non-stationary and non-linear behavior across systems, including decomposed Dynamical Systems (dLDS).
- It 1) learns interpretable operators, 2) enhance dynamic robustness and 3) accurately recovers interactions in both synthetic and human neuronal data.

Motivation

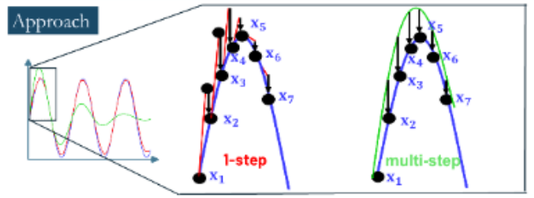




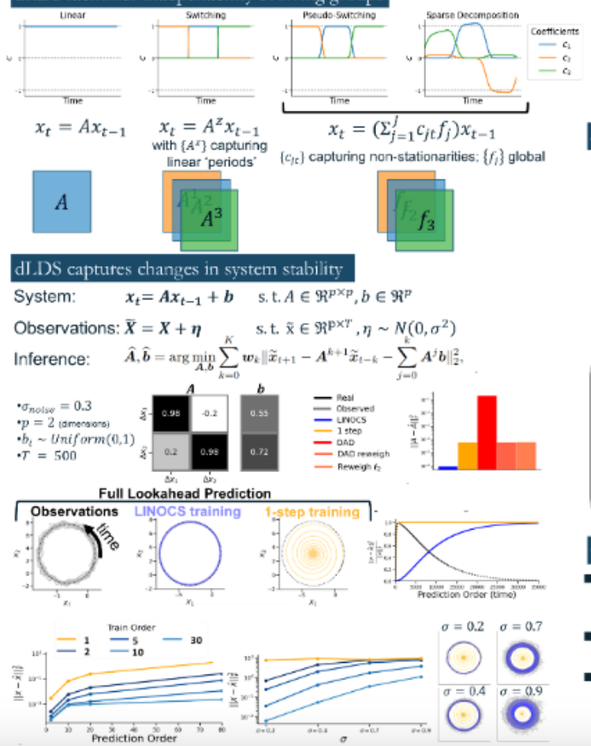
- Small inaccuracies accumulatie in multi-step prediction
- Approach: multi-step training with dynamic step-reweighting

1-step prediction
multi-step prediction
ground truth

Approach



dLDS identifies independently evolving groups



dLDS captures changes in system stability

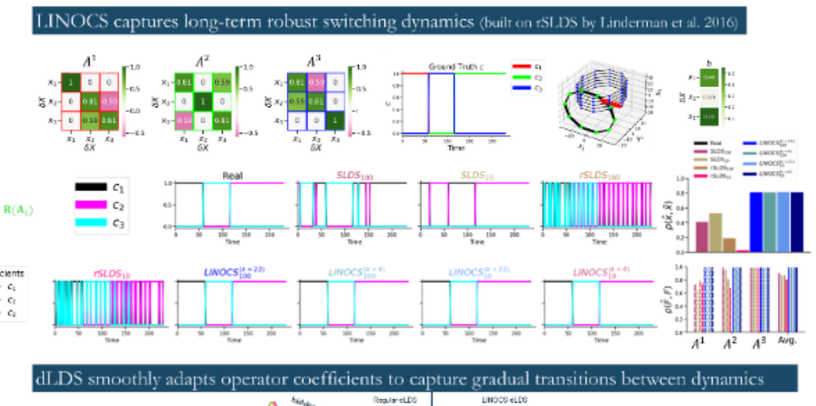
System: $x_t = Ax_{t-1} + b$ s.t. $A \in \Re^{p \times p}, b \in \Re^p$

Observations: $\tilde{X} = X + \eta$ s.t. $\tilde{x} \in \Re^{p \times T}, \eta \sim N(0, \sigma^2)$

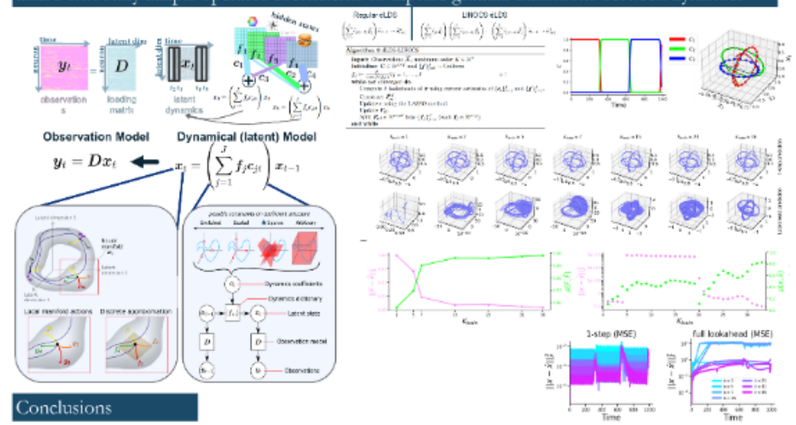
Inference: $\hat{A}, \hat{b} = \arg\min_{A,b} \sum_{k=0}^{K} w_k \|\tilde{x}_{t+1} - A^{k+1}\tilde{x}_{t-k} - \sum_{j=0}^{k} A^j b\|_2^2$,

- $\sigma_{noise} = 0.3$
- $p = 2$ (dimensions)
- $b_i \sim Uniform(0,1)$
- $T = 500$

LINOCS captures long-term robust switching dynamics (built on rSLDS by Linderman et al. 2016)



dLDS smoothly adapts operator coefficients to capture gradual transitions between dynamics



Conclusions

- We introduce LINOCS, a learning procedure to improve dynamical systems inference with lookahead integration, and demonstrated its performance across various dynamical models (linear, switching, decomposed).
- Additional examples in the paper include linear experiments with structured noise and additional dimensions; more diverse dLDS experiment; locally-linear time varying systems; application to real-world neural Data.
- Future directions include applying LINOCS to more datasets for discovering component interactions (e.g., in ecological interactions); extending LINOCS to improve RNN training and tackle vanishing/exploding gradients; handling non-linear local transformations ($x_t = f(x_{t-1})$ for non-linear f); expanding LINOCS to consider non-Gaussian data statistics.

POSTER GALLERY ENTRIES

Development of a Machine Learning-Based Predictive Modeling Framework for Molten Salt Reactor Sensor Data Analysis

Omobolade A. Odedoyin[1,2], Ramon K. Yoshiura[2], Benjamin D. Grimmer[1]

[1]Johns Hopkins University | Whiting School of Engineering Dept. of Applied Mathematics & Statistics | Baltimore, MD
[2]Idaho National Laboratory | Integrated Energy & Market Analysis (C170) | Idaho Falls, ID

ABSTRACT

This research focuses on developing a machine learning (ML)-based predictive framework for analyzing sensor data from Molten Salt Reactors (MSRs). Specifically, the project uses data from Cyclic Voltammetry (CV) & Open Circuit Potentiometry (OCP) experiments and to predict current in molten salt environments. Accurate peak current prediction provides critical insights into the electrochemical behavior of molten salts, improving the operational efficiency & safety of MSRs. This study provides critical insights that help inform the development of systems for real-time monitoring and early detection of anomalies in impurity levels in MSRs. By enhancing the stability and reliability of MSR operations, these efforts support advancements in nuclear technology, contributing to more sustainable and efficient energy production.

INTRODUCTION & OBJECTIVES

MSRs are advanced nuclear reactors that offer significant advantages over traditional pressurized water reactors…

- MSRs operate at **near-atmospheric pressures** & at **higher temperatures**, which **improves their thermal efficiency & safety** [1].
- **Molten salt** serves as **both a coolant & fuel** facilitates **superior heat transfer**, making them a promising solution for safer, more efficient nuclear energy production [2].

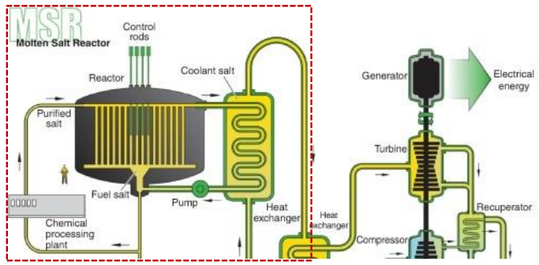


Fig 1. *Schematic of an Industrial Molten Salt Reactor (MSR): fuel salt circulates through the reactor core, generating heat. The heat is transferred via a heat exchanger to a secondary loop, which drives a turbine to produce [1]*

What are the key challenges associated with MSRs?

- **Corrosion & degradation of reactor components & sensors** from dynamic chemical environments within the molten salt [3,4].
- This complicates long-term reactor operation & requires advanced monitoring systems to ensure safe & efficient performance.

How can continuous monitoring & machine learning (ML) framework improve MSR operation?

- Continuous monitoring of molten salt composition is critical for detecting impurities early
- Our research employs an ML-based predictive framework with supervised learning models on data from CV & OCP experiments with molten salt reactions to predict the resulting measured current.
- This project supports a larger initiative using Halide Salt Sensors (HSS) to ensure real-time monitoring & early detection of anomalies, improving the stability, safety, & efficiency of MSRs, contributing to sustainable nuclear energy.

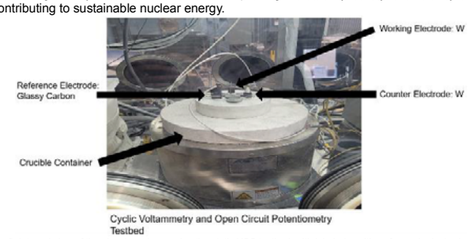
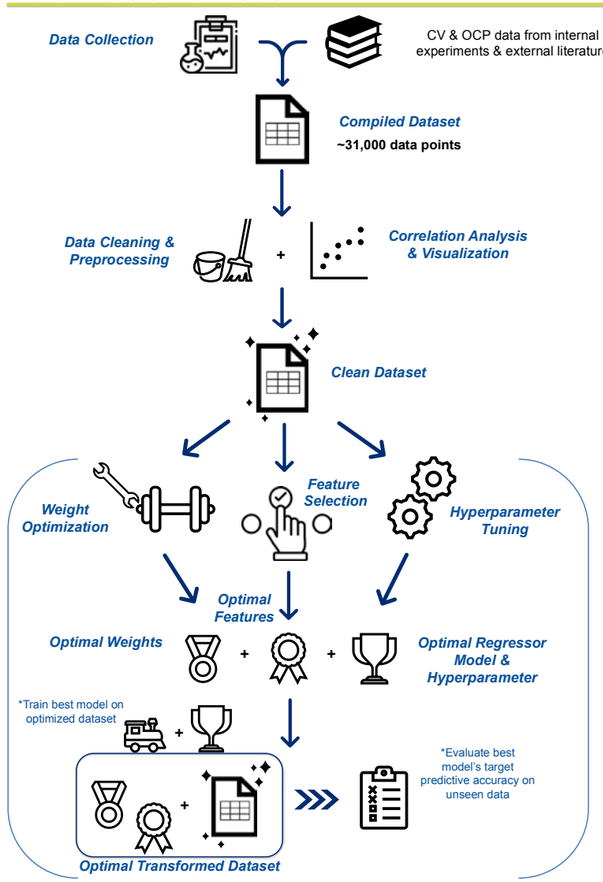


Fig 2. Internal view of the glovebox showcasing the static HSS testbed, with labeled positions for each electrode.

MATERIALS & METHODS



PRELIMINARY RESULTS & PROGRESS

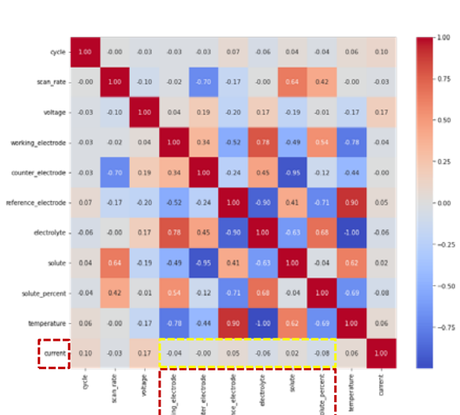


Fig 3. Heatmap of feature correlations within the dataset, including both numeric & non-numeric variables. This visual representation highlights the relationship strength between different features & the target variable (current), allowing for insight into predictor-predictor relationships & potential multicollinearity issues that could affect model accuracy.

- Baseline correlations obtained using one-hot encoding showed low correlations between the qualitative compound features & the target
- Motivated further feature engineering to improve the correlation & thus model accuracy.
- Weighted average approach was applied, using molar mass & ionization energy for the compound features.
- Optimizing these weights can significantly increase the correlation between certain qualitative features & the target.
- For ex, the working & counter electrodes showed a marked improvement in correlation with the target as the weights were adjusted.

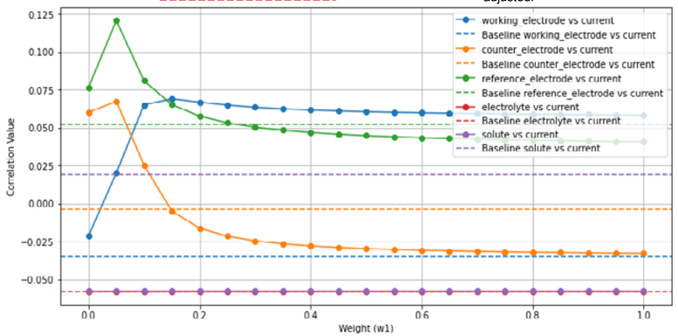


Fig 4. Correlation trends of qualitative predictors (working electrode, counter electrode, reference electrode, electrolyte, & solute) with the target variable (current) across varying weight combinations. The solid lines represent the transformed correlations at different weights, while the dashed lines show the baseline correlation using one-hot encoding for comparison. This analysis helps identify the weight combinations that maximize correlation with the target variable.

FUTURE WORK

- Establish the most optimal weights that maximize the mean absolute correlation with the target & minimize Mean Absolute Error (MAE) for predicting the target.
- Use Recursive Feature Elimination with Cross-Validation (RFECV) to identify the most critical features that minimize MAE for target prediction.
- Grid search to determine the best regressor & hyperparameters that minimize MAE for predicting the target.
- Train the best model using the dataset with qualitative features transformed by the optimal weight combinations & reduced to the most optimal feature set.
- Evaluate the model's accuracy on unseen data to assess its generalization capabilities.

References: [1] Zohuri, B. (2021). Chapter 1 - Molten Salt Reactor History, From Past to Present. In B. Zohuri (Ed.), "Molten Salt Reactors and Integrated Molten Salt Reactors" (pp. 1-58). Academic Press. https://doi.org/10.1016/B978-0-323-85638-6.00001-0. [3] Sridharan, K., & Allen, T. R. (2013). Corrosion in molten salts. In F. Lantelme & H. Groult (Eds.), Molten Salts Chemistry (pp. 241-267). Elsevier. https://doi.org/10.1016/B978-0-12-398538-5.00012-3
[2] Sabharwall, P., Clark, D., Glazoff, M., Zhang, G., Sridharan, K., & Anderson, M. (2014). Advanced heat exchanger development for molten salts. Nuclear Engineering and Design. United States. https://doi.org/10.1016/j.nucengdes.2014.09.026. [4] Holcomb, David Eugene, Kisner, Roger A., & Cetiner, Sacit M. (2018). Instrumentation Framework for Molten Salt Reactors. United States. https://doi.org/10.2172/1462984

POSTER GALLERY ENTRIES

Building an AI for the Study of World-Historical Patterns of Social Protest

Dr. Beverly J. Silver and Haohang Guo
Arrighi Center for Global Studies, Johns Hopkins University

Introduction

The Global Social Protest (GSP) working group at the Arrighi Center is developing a new database on social unrest, covering events *worldwide* from 1792 to the present, reported in international press outlets such as *The New York Times* (US) and *The Guardian* (UK). Our goal is to map the spatiotemporal distribution of protest events, action types, and underlying grievances. Motivated by an effort to understand the current (post-2008) major wave of global protest, the GSP research team is deploying the database to systematically compare the current period with analogous past major global protest waves.

The GSP database currently contains over 570,000 articles extracted from the newspapers using relevant social protest keywords. After establishing annotation guidelines, the research team reviewed and hand-labeled 246,845 of these articles, identifying 53,554 unique "true positives". Leveraging this data, we cleaned and engineered features, compared and evaluated multiple machine learning models, and developed an AI system to make predictions, with the goal of improving the speed and accuracy of data annotation.

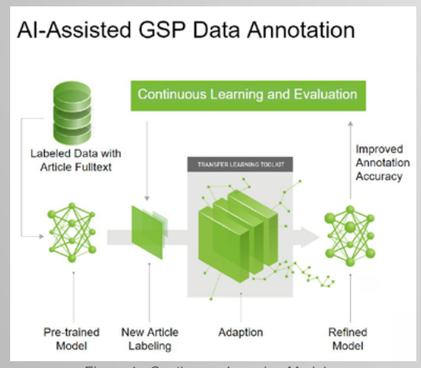


Figure 1: Continuous Learning Model
Modified from: NVIDIA Continuous Learning
(Source: https://docs.nvidia.com/clara/clara-train-sdk/index.html)

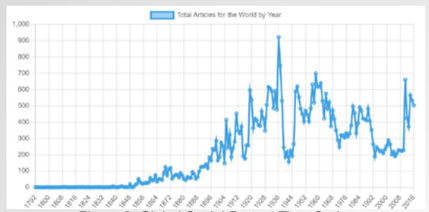


Figure 2. Global Social Protest Time Series
(Manually Labeled by Human Researchers)

Figure 3. Social Protest Time Series for China
(Manually Labeled by Human Researchers)

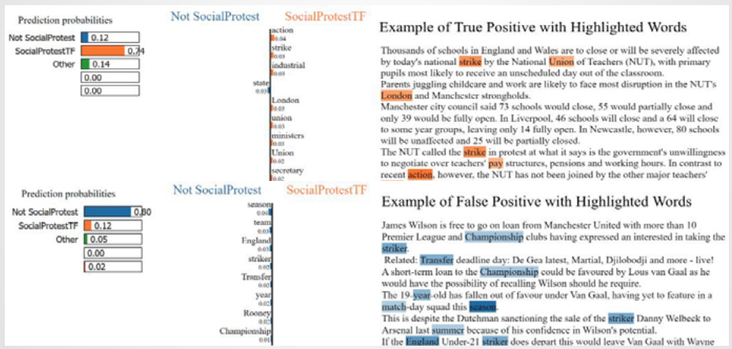


Figure 4: Local Interpretable Model-Agnostic Explanations (LIME)

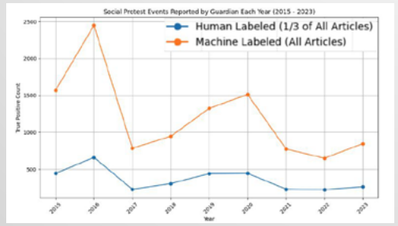


Figure 5: Social Protest Events
Reported by The Guardian Each
Year (2015 - 2023)

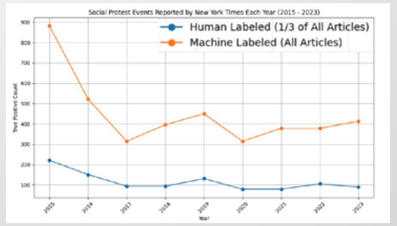


Figure 6: Social Protest Events
Reported by The New York Times
Each Year (2015 - 2023)

	Precision	Recall	F1-score	Support
0	0.89	0.85	0.87	710
1	0.83	0.87	0.85	580
Accuracy			0.86	1290
Macro Avg	0.86	0.86	0.86	1290
Weighted Avg	0.86	0.86	0.86	1290

Table 1: Evaluation report from a random forest classifier with TF-IDF vectorizer, predicting 0 and 1 in social protest using 20% testing data (pilot study to test the efficacy of the machine learning model and data)

Discussion

1. We found that the labels that the AI predicts conform with the general trend of labels produced by human annotators when it comes to identifying true positive events of social protest. Our initial assessments suggest that the AI can be deployed to annotate the remaining unlabeled articles in the database (approx. 380,000 articles) efficiently and (relatively) reliably.

2. We remain concerned about the black box, i.e., the low explainability for how AI labels are produced. By highlighting relevant words and associated likelihoods, Figure 4 attempts to peek inside the black box.

3. Our next research steps include: (a) training the AI to label other variables such as 'action type' and 'grievances'; (b) expanding the predictive capacity of the AI to languages other than English; (c) developing a topic modeler.

Acknowledgements

This research has been funded by grants from The National Science Foundation (#1460434) and seed money grants from IDIES and from the JHU Data Science and AI Institute. We would also like to acknowledge other core members of the research team including Dr. Sahan Savas Karatasli, Dr. Corey Payne, Dr Juan Grigera, Ms. Wei Zhou and multiple cohorts of JHU graduate and undergraduate student team members.

POSTER GALLERY ENTRIES

Unveiling latent RNN dynamics by leveraging decomposed linear dynamical systems

Daniel Wang[1*], Noga Mudrik[1], Adam Charles[1]
Presenter: dwang118@jhu.edu; Corresponding: adamsc@jhu.edu
[1] Biomedical Engineering, Kavli NDI, Johns Hopkins University

DLDS PAPER

MULTITASK RNN PAPER

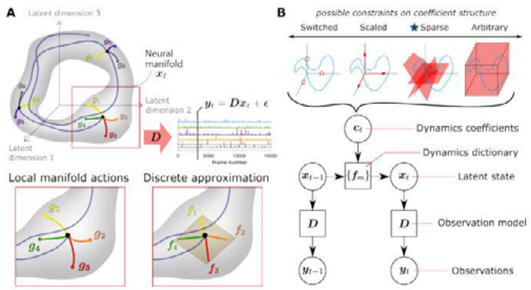
Introduction

- Recurrent neural networks (RNN) have emerged as a powerful tool to analyze time-series data, but existing work with RNNs have focused primarily on performance rather than interpretability.
- We focus on two computing properties: computational modularity and non-stationarity.
  - **Computational modularity** refers to the organization of a system into distinct, independent components or modules, each responsible for a specific task of function.
  - **Non-stationarity** refers to a system whose statistical properties change over time.
- Recent work by Driscoll et al. found that the computations performed by RNNs trained to perform multiple tasks exhibit underlying dynamical motifs that evolve over time.[1]
  - These dynamical motifs have a modular and compositional structure.
- Decomposed linear dynamical systems (dLDS) is a method by which one can uncover the fundamental building blocks of such dynamical motifs.[2]
  - This is done by decomposing the complex dynamics of the RNN's state activations into the simpler, interpretable latent variables that drive the observed dynamics.

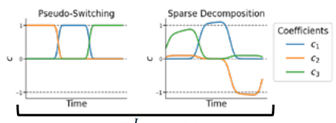
Motivation

- Existing evidence that RNNs operate through internal modularity and these modules may capture or reflect non-stationarities. dLDS serves as a tool to uncover these internal modules.
- Identification of the latent modular components gives a better understanding of the inner workings of the RNN that is otherwise a "black box".
- Understanding the modular motifs that drive RNN computation and behavior may give us better insight on how other computational systems (such biological neural systems) perform computation.

dLDS



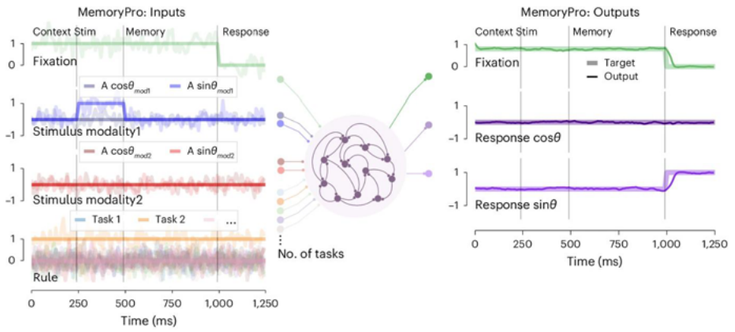
dLDS identifies independently evolving groups



$$x_t = (\Sigma_{j=1}^{J} c_{jt} f_j) x_{t-1}$$

$\{c_{jt}\}$ capturing non-stationarities; $\{f_j\}$ global

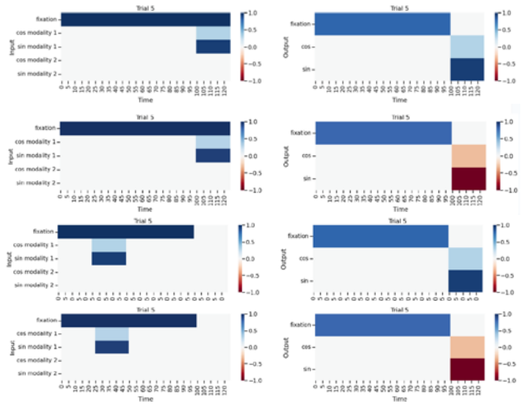
Multitask computation in recurrent neural networks (Driscoll et al. 2024)



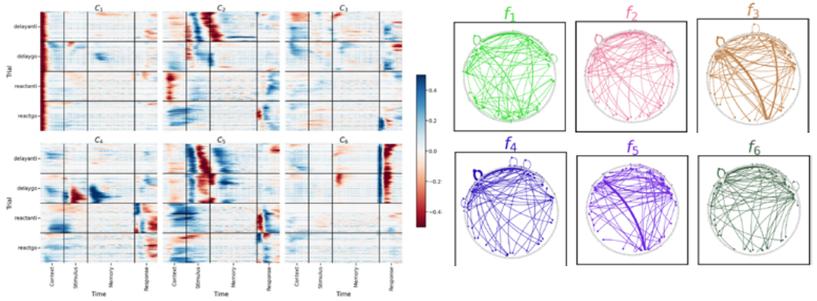
Input/Output of RNN for 4 tasks (Driscoll et al. 2024)

- Tasks differ in the timing of the stimulus and the expected output of the RNN
- Tasks
  - ReactGo
  - ReactAnti
  - DelayGo
  - DelayAnti

Plots on the righthand side show the inputs into the RNN and the RNN's output for each task respectively



dLDS identifies modular dynamical motifs in the computation of the multitask RNN



- Visualization of the time-varying weights of decomposed latent components for each task.
- Visualization of the latent components and their connectivity.

References

1.Driscoll, L. N., Shenoy, K. & Sussillo, D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. Nature News (2024). Available at: https://www.nature.com/articles/s41593-024-01668-6. (Accessed: 16th October 2024)

2. Mudrik, N., Chen, Y., Yezerets, E., Rozell, C. J., and Charles, A. S. Decomposed linear dynamical systems (dlds) for learning the latent components of neural dynamics. Journal of Machine Learning Research (2024).

POSTER GALLERY ENTRIES

JHTDB-wind: The Johns Hopkins wind farm flow database for open community access to high-fidelity simulation datasets

Shuolin Xiao, Xiaowei Zhu, Ghanesh Narasimhan, Dennice Gayme, Charles Meneveau

The Ralph O'Connor Sustainable Energy Institute (ROSEI) | Johns Hopkins University

Introduction

Large Eddy Simulations (LES) have been used to provide the insights required to improve wind farm design and operation. However, these high-fidelity simulations require domain expertise and are computationally expensive. Each simulation also produces very large datasets (tens to hundreds of Terabytes) that are difficult to share with the community using prevailing approaches.

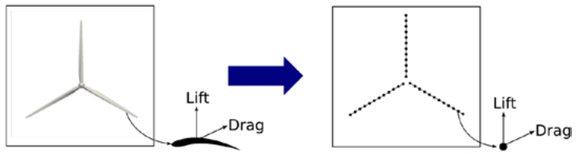
Objective

The JHTDB-wind is designed to address aforementioned challenges by developing, building and using new tools for analysis and efficient public access to the very large datasets that arise from numerical simulations of complex flows in wind farms.

Simulation Approach

JHU-LESGO code: Parallel Large-eddy simulation code on Cartesian mesh grid; Pseudo-spectral in plane and Second-order finite difference in wall-normal; Lagrangian scale-dependent model; Equilibrium wall model; Filtered actuator line model.

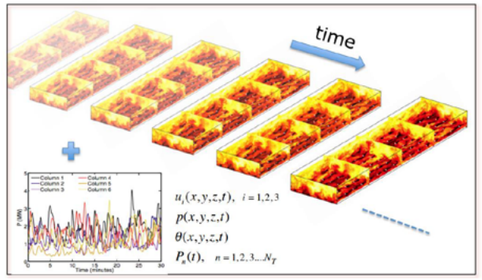
Actuator line model:



$$F_L = \frac{1}{2}\rho c C_L U_\infty^2;\ F_D = \frac{1}{2}\rho c C_D U_\infty^2;$$

Technical Goals

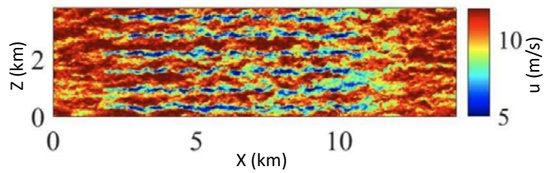
- Publicly accessible, Web-Services equipped databases will be constructed.
- Databases will contain various turbine spacings, arrangements, and a number of representative atmospheric conditions (neutral, stable, and convectively stratified, and an entire representative daily cycle).



- Databases will contain variables describing the full flow fields (velocity vector, pressure, potential temperature, turbulent viscosity, turbine forces, etc.) along with additional relevant time-series of wind turbine information such as power and rotor torque.
- A total of over 60 Terabytes of wind farm databases will be developed, with the 3D field data filtered and sampled every horizontal grid points from LES simulations to save the storage.
- The 3D field data is stored in Zarr format, and the time-series wind turbine information is stored in Parquet format, allowing faster access via web services.

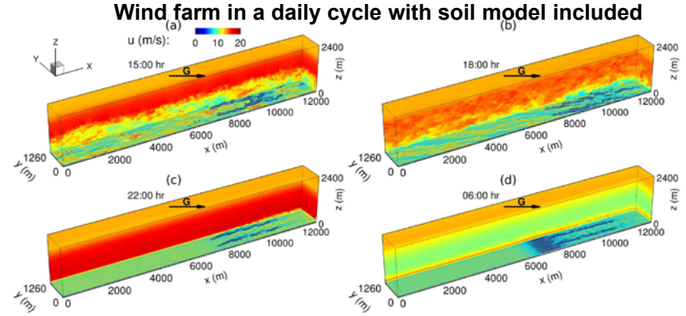
Wind Farm Results

Wind farm in a neutral boundary layer

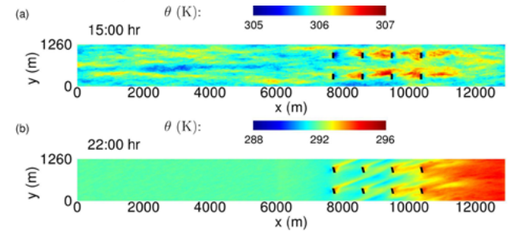


Acknowledgement: Funding provided by a ROSEI seed fund and IDIES infrastructure. Computational resources provided by the National Center for Atmospheric Research and Advanced Research Computing at Hopkins (ARCH) core facility (rockfish.jhu.edu), which is supported by the National Science Foundation (NSF) grant number OAC1920103.

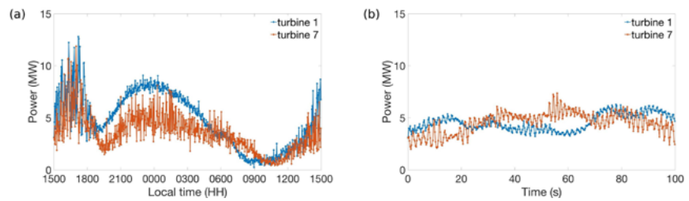
Wind farm in a daily cycle with soil model included



- The wake changes direction in a diurnal cycle.
- Strong wind farm blockage along with a low-level jet is observed in the early morning.



- A strong warming effect at ground surface is observed behind the wind farm at night.



- Uncommonly higher power time series in the last row of wind turbines during the early morning, compared to those in the first row.

Conclusion: Two wind farm configurations have been set up for database storage: one features a large array of wind farms in a neutral boundary layer, and the other has a smaller array spanning an entire representative daily cycle.

2024 MEMBER NEWS & ACHIEVEMENTS

NOTABLE MEMBER ACTIVITY FROM OCTOBER 1, 2023–SEPTEMBER 30, 2024 PLUS THE HIGHLY ANTICIPATED RETURN OF IDIES BY THE NUMBERS

Paulette Clancy, 2024 IDIES seed funding recipient, secured a $2.7 million National Science Foundation (NSF) National Research Traineeship (NRT) grant as a PI for a cross-institutional team of JHU and Morgan State researchers dedicated to furthering graduate-level training in the areas of low-dimensional and quantum materials; materials for advanced semiconductor manufacturing; advanced computing hardware; and next-generation electronics for new environments and applications.

Charles Meneveau, Louis M. Sardella Professor of Mechanical Engineering, was awarded the *Journal of Fluid Mechanics* Batchelor Prize for 2024.

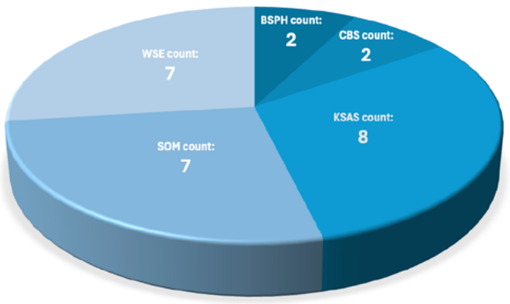
Also, Charles Meneveau's article, "Forward and inverse energy cascade in fluid turbulence adhere to Kolmogorov's refined similarity hypothesis" was published online in the 19 April 2024 issue of *Physical Review Letters* (Vol. 132, No. 16): https://physics.aps.org/articles/v17/s45 by the American Physical Society.

Somdatta Goswami and her team, were awarded the Early-concept Grant for Exploratory Research (EAGER) through NSF's National Artificial Intelligence Research Resource (NAIRR) Pilot program, which serves to democratize access to the compute infrastructure necessary for AI-guided research. Goswami's EAGER-funded work uses ML to expedite exascale multiphysics simulations allowing past solutions, derived from traditional small-scale simulation approaches, to be applied to large, multifaceted systems.

Corey Oses, 2023 IDIES seed funding recipient, received the Early-Career Investigator award for materials modeling from the International Society of Materials Modeling (ISMM).

Also, Professor Oses held the first-ever Data-Driven Materials Workshop this past May. With the support of IDIES, The Ralph O'Connor Sustainable Energy Institute (ROSEI), and the Advanced Research Computing at Hopkins (ARCH) facilities, attendees gathered at the Homewood campus for three full days of speakers, and hands-on training.

2024 IDIES BY THE NUMBERS



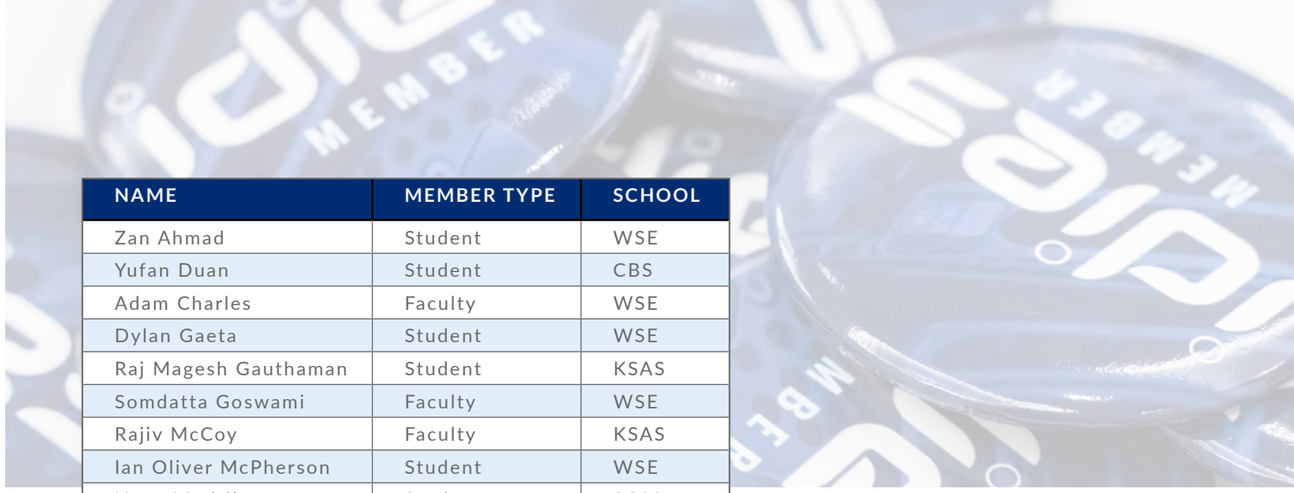
SEED FUNDING PROPOSALS RECEIVED BROKEN DOWN BY PI APPOINTMENT

26
PROPOSALS RECEIVED FOR SEED FUNDING

$155K
GIVEN IN FUNDING AWARDS

16
NEW MEMBERS

2PB
TOTAL OF PUBLIC DATASETS ON SCISERVER



NAME	MEMBER TYPE	SCHOOL
Zan Ahmad	Student	WSE
Yufan Duan	Student	CBS
Adam Charles	Faculty	WSE
Dylan Gaeta	Student	WSE
Raj Magesh Gauthaman	Student	KSAS
Somdatta Goswami	Faculty	WSE
Rajiv McCoy	Faculty	KSAS
Ian Oliver McPherson	Student	WSE
Noga Mudrik	Student	SOM
Shreesh Mysore	Faculty	KSAS
Omobolade Odedoyin	Student	WSE
Milton Roy	Student	SOM
Rafael Santos	Alumni*	
James C. Spall	Faculty	WSE, APL
Webster Stayman	Faculty	SOM
Yashil Sukurdeep	Affiliate	WSE
Guanjin Wang	Affiliate	WSE
Ziqing Ye	Student	WSE

WELCOME TO THE 16 NEW IDIES MEMBERS (*LEFT*) WHO JOINED BETWEEN OCTOBER 1, 2023 – SEPTEMBER 30, 2024

To learn more about membership, visit https://www.idies.jhu.edu/who-we-are/join/

*The Alumni category was added this year and is extended to former IDIES members and collaborators.

Adam Sheingate held the first workshop for his IDIES seed-funded project "The Global Elections Dashboard: Harnessing Data Science to Enhance Electoral Oversight" at the JHU Bloomberg Center in DC.

The meeting brought together social scientists, engineers, data journalists, and transparency advocates from the US, UK, and Brazil to discuss how AI and data science can be utilized to collect, analyze, and disseminate campaign finance and spending data from around the world. Professor Sheingate also discussed this project at a KSAS Dean's Faculty Forum.

See the election spending dashboard, now live, here: https://election-spending-data.shinyapps.io/dashboard/

THANK YOU FOR YET ANOTHER DATA-INTENSIVE YEAR AT IDIES AND JOHNS HOPKINS!

A collaborative team of researchers at IDIES and NIST received the 2024 NIST Material Measurement Laboratory (MML) Technology Transfer accolade, for developing an innovative public science domain for NIST's Additive Manufacturing Benchmark data system, hosted on www.sciserver.org.

Team members recognized were Benjamin Long, Brandon Lane, Gerard Lemson, Gretchen Greene, Lyle Levine, Andrew Reid, Chandler Becker, Jai Won Kim, Samia Benjida, Carlyn Campbell, Arik Mitschang, and Jordan Raddick.

PLEASE KEEP AN EYE OUT FOR UPCOMING ANNOUNCEMENTS ON OUR 2025 FUNDING CYCLE AND REQUESTS FOR PROPOSALS FOR IDIES SEED FUNDING AND SUMMER STUDENT FELLOWSHIP PROGRAMS.

idies

" ai · dee · uhz "

NEXT ISSUE

NOVEMBER 2025