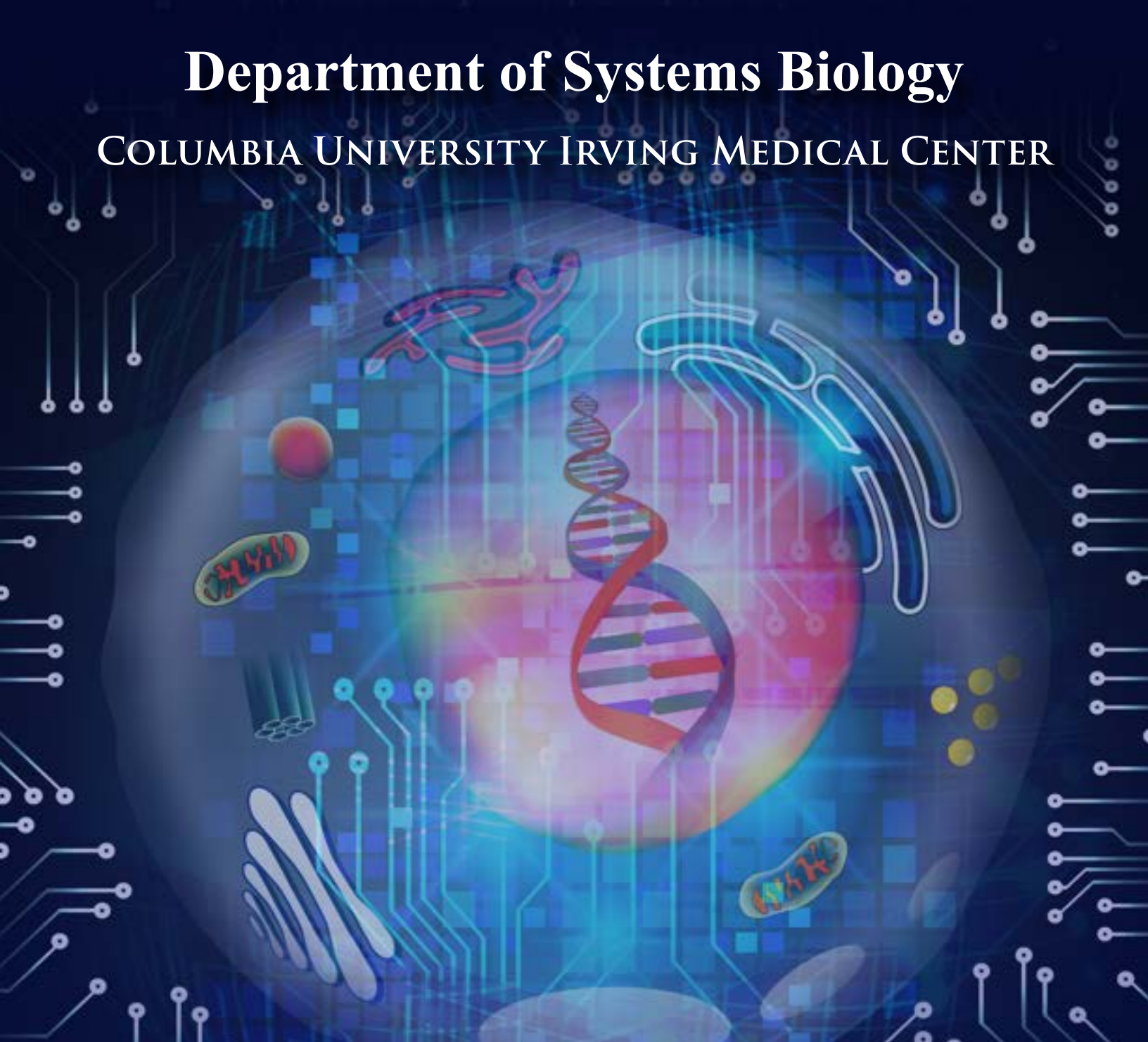
Department of Systems Biology

Columbia University Irving Medical Center



Newsletter 2023–2024

COLUMBIA UNIVERSITY
IRVING MEDICAL CENTER
Department of Systems Biology

Table of Contents

Highlights 2023–2024

Inside this issue, learn about the latest research, noteworthy grants and milestones from the Department of Systems Biology.

Afternoon of Science

The Afternoon of Science series at the Vagelos College of Physicians and Surgeons took place on September 25 with presentations from the Department of Systems Biology.

The event was hosted by the interim chair, Harris Wang, Dr, associate professor of systems biology, pathology & cell biology, and biomedical engineering, who shared his vision for the department. The Department of Systems Biology was founded in 2013 and now has more than a dozen primary faculty and 19 additional cross-appointed faculty members, who work to advance the integration of computational and experimental research methods in the biological and biomedical sciences. Research in the department is highly collaborative and integrates specialists in molecular biology, genetics, computational biology and bioinformatics, mathematics, chemistry and chemical biology, physics, and other fields.

Looking to the future, Wang shared enthusiasm for areas of growth in the field, including foundational AI models for DNA, RNA, and proteins; ultra-high-throughput experimental biology and automation; cell therapies and living medicines; and genome medicine technologies.

"We're excited for the launch of the Vagelos Institute for Basic Biomedical Research Education and its potential to catapult the next generation of basic science and graduate education here on campus for decades to come," Wang said.

Faculty presentations were made by:

Andrea Califano, President of the Chan Zuckerberg Biohub New York; Clyde '56 and Helen Wu Professor of Chemical Biology (in Systems Biology); Professor of Biomedical Informatics, Biochemistry & Molecular Biophysics, and Medicine (in the Institute for Cancer Genetics); and founding Chair of the Department of Systems Biology.

Presentation: "A Brief History of Time (and of the DSB)"

Peter Sims, Associate Professor of Systems Biology and Biochemistry & Molecular Biophysics and Scientific Director, Sulzberger Columbia Genome Center

Presentation: "Systems Immunology in Human Tissues"

Sara Zaccara, Assistant Professor of Systems Biology

Presentation: "Unraveling the Epitranscriptome and its Impact on Disease"

Dennis Vitkup, Associate Professor of Systems Biology and Biomedical Informatics

Presentation: "Systems Biology of Psychiatric Disorders"

Andrea Califano

Mohammed AlQuraishi, Assistant Professor of Systems Biology (in Computer Science) and Member of the Program in Mathematical Genomics

Presentation: "Current Frontiers in Biomolecular Machine Learning"

Raul Rabadan, Professor of Systems Biology and Biomedical Informatics, Gerald and Janet Carrus Professor of Surgical Sciences (in Surgery and in the Institute for Cancer Genetics), and Director of the Program in Mathematical Genomics

Presentation: "Illuminating the Dark Cancer Genome using Foundation Models"

Mohammed AlQuraishi is Taking His Deep Learning OpenFold Model to the Next Level

The process by which proteins fold into various shapes had tantalized scientists for decades with so many questions to answer. What physical forces govern this complex process? Why does it happen so fast—within a thousandth of a second? And is there a way to predict how proteins will fold, based on their amino acid composition?

"At first, the paradigm was very physics-based," says Mohammed AlQuraishi, PhD, an assistant professor of systems biology at Columbia University. "You start with a protein sequence, you understand the types of energies involved in protein folding, and you try to emulate that on a computer."

The arrival of self-learning AI models enabled AlQuraishi to change that paradigm. He used AI to comb through amino acid sequences, learn the patterns and predict how proteins might fold. In 2018, AlQuraishi released an AI model that made good estimates of protein structures, six times faster than other methods.

Two years later, Google DeepMind released AlphaFold2, an AI model that predicted protein structures with exactitude no model could do before. However, AlphaFold2 had limits: it couldn't predict mutations or explain how proteins might interact with each other or with drug molecules. Plus, although DeepMind released some source code, it didn't release enough code for researchers to retrain the model with different data.

AlQuraishi, and his Harvard Medical School collaborator Nazim Bouatta, decided to go a step further and create a fully open-source model. Working with a group of students, led by Gustaf Ahdritz—at the time a Columbia master's student—they built their own model, OpenFold, releasing it six months later. OpenFold was faster than AlphaFold2 and just as accurate. Moreover, it could be retrained on other data, a feature that paved the way to more advances. For example, by combining OpenFold with a language model akin to ChatGPT, Meta AI was able to reveal the structures of more than 600 million little-known proteins—including those of viruses that dwell deep in the ocean.


Mohammed AlQuraishi

OpenFold helped researchers understand the protein-folding process and predict protein structures from amino acid sequences, but it still left some unanswered questions. For example, the other two key pieces of the puzzle—understanding the underlying physics and why proteins fold so fast—remain to be deciphered. Scientists also want to know what causes some proteins to fold incorrectly, which can lead to neurodegenerative diseases.

AlQuraishi and collaborators are working on the next version of OpenFold. They are also examining the shapes of the structures OpenFold builds as it trains—to understand how their model learns. Interestingly, they found that OpenFold doesn't necessarily need to know physics to assemble protein structures—it creates one-dimensional views of folded proteins and proceeds to build two- and three-dimensional views. The scientists, however, believe that teaching deep learning systems some physics would make them better. And, they plan to keep their model open source, so other teams can build upon their work. "The open-source component is critical and is only going to become more so moving forward," AlQuraishi says

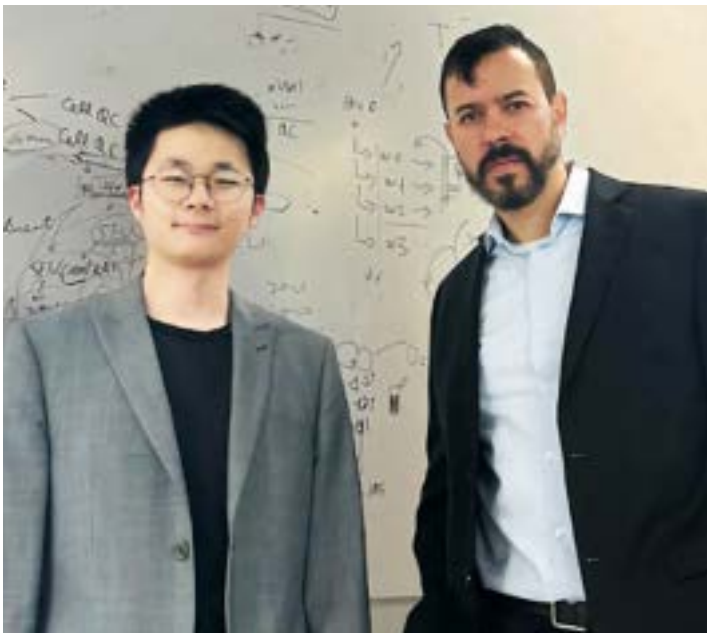
Raul Rabadan's New AI Model Predicts How Genes Inside a Cell Would Govern its Behavior

In a tour de force of computational genomics, scientists at Columbia University trained a deep learning AI model to predict how the genes inside a cell would govern its behavior—a feat that aims to explain how certain diseases develop and find ways to treat them.

Called General Expression Transformer or GET, the model was built similarly to ChatGPT, only instead of English grammar rules, it focused on learning the language of genes that govern cells. Using 213 different cell types found in the human body, with over 1.3 million cells in total, Raul Rabadan, PhD, director of the Program for Mathematical Genomics at Columbia, and Xi Fu, a graduate student in Rabadan's lab, trained GET to predict how a cell's genes affect its behavior. Their paper, "A Foundation Model of Transcription Across Human Cell Types," was published in the journal *Nature* on January 8, 2025.

The *Nature* paper tackles one of biology's greatest challenges: understanding how the same genome can drive diverse behaviors in different cell types. Human body contains about 30 trillion cells. In a given individual, all of these cells have the same DNA, yet they function differently depending on the type of cells they are. A skin cell performs very different functions than a retinal or a liver cell. These behaviors are governed by about 20,000 genes, some of which are turned on in certain cells and switched off in others or dialed up or down, increasing or decreasing their activity—in a process called gene expression.

When gene expression goes awry, diseases develop, so understanding how the genes inside a cell drive its behavior is important for treatment. However, this regulatory "grammar" is poorly understood, so GET makes an important step toward elucidating this language. Understanding the language of gene regulation holds the potential to greatly benefit human health. "Biology is being transformed into a predictive science," says Rabadan. His team found that even if they omitted one cell type from the data—for example astrocytes, crucial to the central nervous system—GET still made accurate predictions about astrocytes based on what it had learned from all other cells.



Xi Fu, a graduate student and lead author of the *Nature* paper, and author Rual Rabadan, both from the Program for Mathematical Genomics at Columbia.

Scientists hope that GET will help in several ways. It can be a powerful tool for understanding the development of cancer and genetic diseases. Being able to predict which genes are turned on, off, up or down in different cells could help determine the cell where a disease originated. GET also can aid in designing cell-specific gene therapies to correct a mutation that caused a particular health problem. Such therapies would be designed with great precision, so they only fix the cells harmed by the mutation, bypassing the unaffected cells. "We can design gene therapies that deliver a gene that is only expressed in one cell type, and not in another," Rabadan says.

Finally, GET holds the promise of reducing the complexity of the too common needle-in-a-haystack problem. A cancer, for example, may contain more than 1,000 mutations, some of which may be harmful and others not. That makes it difficult for scientists to determine which mutations should be fixed to treat the disease. "The number of potential genetic combinations is more than the number of atoms in the universe," Rabadan says. "What are the ones that are relevant?" As deep learning systems keep evolving, they might help answer that too.

From Lab to Clinic: New Genomic Tools Transform Disease Research

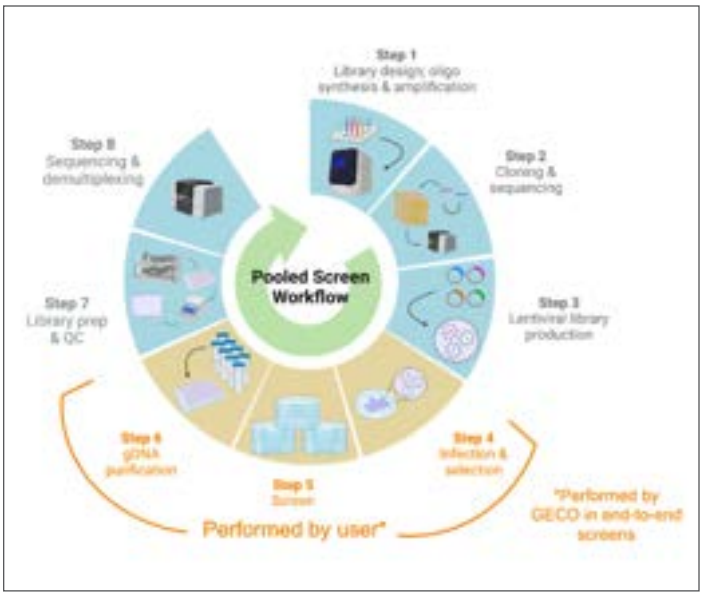
Scientists have long struggled to understand the vast complexity of biology at the systems level. Now two transformative technologies are changing that. "Genome engineering and artificial intelligence may be the most important technologies in generations, and we have only begun to see what they can do together," says Jeremy Worley, PhD, an Assistant Professor at CUIMC in the Department of Systems Biology and Director of the Genome Engineering CoLab (GECO). GECO is a new platform designed to collaborate with researchers on projects that require cutting-edge genomics technologies. Worley's work aims to develop advanced genomics tools that unravel the complex genetic underpinnings of diseases.

It started with Worley's postdoctoral research in Andrea Califano's lab, in which he and colleagues created the Transcriptional REgulator Knockdown (TREK) project to improve computationally inferred gene-regulatory networks (GRNs) with data from large-scale CRISPRi time course experiments. Computational biologists are now using that data to test and fine-tune AI models to predict cellular processes at the systems level.

Since the TREK Project began in 2017, there have been tremendous advances in CRISPR technologies. Although in popular culture, we primarily hear about the use of CRISPR in therapeutics, like CASGEVY®, a recently approved treatment for beta-thalassemia and sickle-cell disease, behind the scenes, hundreds of new CRISPR tools are allowing researchers to better understand how cells work at virtually all levels—including the vital question of how our immune system battles tumor cells.

To answer that question, Worley and GECO are working closely with experimental biologists and AI researchers at Columbia and the Chan Zuckerberg Biohub New York. Combining large pooled CRISPRi screens with new machine learning models to uncover how macrophages are reprogramed and inactivated by tumor cells is "the most complex, interdisciplinary project we've ever worked on," says Worley.

And that is just one of many projects. GECO is currently collaborating with researchers studying prostate,



colorectal, and pancreatic cancer, as well as neuroscientists investigating neurodegenerative diseases. Last year, GECO's projects at the nexus of massive-scale CRISPR-based genome engineering and AI have soared to the new heights.

"Five years ago, we were doing experiments that involved hundreds of thousands of cells at a cost of close to one dollar per cell," Worley says. "Now we can profile millions of cells in a couple of hours at less than ten cents per cell, and the costs are still plummeting. And our ability to use CRISPR to perturb cells has dramatically improved with the newest methods." These technological advances—the ability to profile tens or hundreds of millions of cells combined with enhanced CRISPR techniques—will be critical to building and fine-tuning the next generation of machine learning models.

REFERENCE:

Xiangtian Tan, Jeremy Worley, Mikko Turunen, Kelly Wong, Ester Calvo Fernández, Evan Paull, Sunny Jones, Junqiang Wang, Heeju Noh, Beatrice Salvatori, Alejandro Chavez, Andrea Califano. Interrogation of genome-wide, experimentally dissected gene regulatory networks reveals mechanisms underlying dynamic cellular state control. bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.449297

FACULTY SPOTLIGHT

Q&A with Raul Rabadan

Raul Rabadan is a professor in the Departments of Systems Biology, Biomedical Informatics and Surgery and director of the Program for Mathematical Genomics (PMG) and, previously, the Center for Topology of Cancer Evolution and Heterogeneity. He established PMG in the fall of 2017 with the goal of bringing together scientists, mathematicians, and researchers from multiple disciplines to work toward a quantitative understanding of complex biological systems.

Q: *You have a background in physics. How did you end up in medicine?*

A: I began my academic career in theoretical physics, which I realize is not the usual background for working in the field of cancer genomics and epigenomics. I was a researcher at CERN, the European Organization for Nuclear Research, in Geneva, Switzerland and in 2003 I joined the Physics Group of the School of Natural Sciences at the Institute for Advanced Study in Princeton, New Jersey, whose faculty included Albert Einstein, John von Neumann and Robert Oppenheimer, among others. While at the Institute, I became interested in biology and the potential of developing quantitative models of biology.

Q: *How did you end up in medicine?*

A: When I came to Columbia in 2008, I was highly active in trying to understand the origins of pandemic viruses, mostly influenza pandemics. The main question I was trying to address is how viruses that are infecting other animals (birds, pigs, rodents, etc.) can acquire the capacity to infect and transmit it in humans. One year after I came to Columbia there was a pandemic caused by the H1N1 influenza virus. I was fascinated with understanding the genomic changes of a virus and how influenza viruses that infect animals, such as pigs and birds, could generate new genetic combinations that enable them to infect and propagate in humans. My colleagues and I developed computational methods to identify genetic changes in these rapidly evolving systems and to trace back these changes in viruses propagating in different species. This was the seed for the of the genomic work I have carried out with many colleagues at Columbia in cancer evolution since 2010.

Raul Rabadan

Q: *Why is prediction so important in medicine?*

A: I have always thought that biology does a wonderful job in describing processes and diseases. But the ability to predict is even more important. Wouldn't it have been great if we could have predicted the SARS-CoV-2 pandemic and taken steps to mitigate its devastating effects?

The future of medicine is in prediction and in the design of new biological systems. We are at an extremely exciting historical moment in biology where high throughput data and large models are making biology a predictive science. The 2024 Nobel Prize in Chemistry was awarded to Demis Hassabis and John Jumper (Google DeepMind) along with David Baker (University of Washington) for their groundbreaking work in predicting and designing protein structures using artificial intelligence. Although protein prediction and design are only one aspect of biology, the transformation will be pervasive across many layers of biology.

Q: *What is systems biology and why is it important?*

A: The NIH defines systems biology as "field of study that uses computational and mathematical models to analyze and understand the complex interactions within biological systems." Systems biology is a comprehensive approach that considers how different organisms and their parts interact at many scales."

Connecting and predicting differences is the goal and dream of systems biology. Biology is not a set of random processes but working systems with an underlying

Continued on page 8

Continued from page 8

structure. This involves data generation and constructing quantitative models that go beyond the traditional realm of descriptive science. I want to know what are the functions that connect different layers of biological information? What are the right mathematical models that capture these connections?

Q: *What are foundation models and why are they important?*

A: Foundation models are large-scale machine learning models trained on massive datasets and designed to be versatile across various tasks. Examples are large language models like chat-GPT that take all documents from the internet, learn the grammar, and can generate new documents.
What we have shown is that we can create foundations models that are able to capture and predict gene regulation using genetic and epigenetic data in a cell-type specific manner. Working with more than 200 cell types in adult and fetal tissue we can understand and model how cell types differentiate.

Q: *Can you tell me about experiments in silico?*

A: "In silico" experiments refer to computer simulations or experiments conducted on a computer or via computer software rather than in a physical laboratory setting or in a test tube, cell lines or animal models. The questions we ask are: what is the effect of the mutations that we find in patients? Can we design new cell therapies that are cell type specific?
We have a large programmatic effort at Columbia to expand this work into other conditions, mostly in cancer: brain tumors, lymphomas of B- and T-cells, DNA repair genes, etc. But it can be applied to many other domains of biomedical research.

Q: *You have a paper coming out about a cell type-specific foundation model of expression. Can you tell us about it?*

A: This effort is led by Xi Fu, an extraordinary graduate student in my laboratory. It is about GET, which stands for General Expression Transformer. GET is an interpretable foundation model designed to uncover regulatory grammars across 213 human fetal and adult cell types. Through GET we can learn the grammar and apply it to new cell types, new conditions, and diseases. For example, in the paper we study how cell specific TF-TF interactions can get dysregulated with mutations in cancers and perform in silico mutagenic experiments to understand the role of non-coding mutations in cancer, among many other application

Service Core Updates

Columbia Genome Center: In 2024, the NextGen and Single Cell Service cores merged to form the Columbia Genome Center, creating a unified entity dedicated to advancing genomic research. The newly established center continues to enhance its capabilities in advanced genomics and transcriptomics, leveraging the Element Aviti system and NovaSeq X technology. It also provides comprehensive single-cell profiling services using the 10x Genomics Chromium and Chromium-X Controllers, as well as spatial transcriptomics through the 10x Genomics Xenium platform. The center supports GECO by delivering single-cell genomics services for CRISPR screens and offering computational resources for spatial profiling.

Information Technology Service Core: For over 15 years, C2B2/DSBIT has provided essential IT services including co-location, big data storage, high-performance computing (HPC), and desktop administration. These services have been widely utilized by the Department of Systems Biology, various departments and centers, as well as individual research labs. Notably, co-location, enterprise storage, and HPC services are uniquely offered by DSBIT on the CUIMC campus.

Recently, through the leadership of Dr. Andrea Califano and contributions from multiple researchers, a $2 million NIH grant (Award #1S10OD032433-01A1) was secured, enabling the acquisition of a cutting-edge HPC cluster. This state-of-the-art system boasts 12,000 CPU cores, 80 NVIDIA L40S GPUs with a combined 1.5 million CUDA cores, 50 TB of memory, a 1 PB NVMe-based Weka storage system offering high IOPS, and a 200 Gbps low-latency InfiniBand fabric interconnecting all subsystems.

For additional information about the new HPC cluster, please visit: Columbia University Wiki – RCAC Overview.

ANNUAL RETREAT

The Department of Systems Biology's 2024 annual retreat was held in September at Woodloch Pines Resort, PA. The retreat gave DSB faculty, post-docs, and students a chance to get away from the bustle of New York City, learn about their peers' research, and network.

DSB researchers and graduate students participated in a poster competition held the first evening and reviewed by Systems Biology faculty judges. The three poster winners were: Daniel Caron for "Maintenance of human macrophage tissue identity and plasticity during acute polarization"; Yiwei Sun for "Fecal exfoliated RNA (eRNA) profiling captures immune dynamics of healthy and inflamed gut"; and Ruchika Mishra for "Deciphering RNA recognition code for Pentatricopeptide repeat (PPR) editing factors."

From left: to come

Top row, from left: Daniel Caron & Harris Wang; Harris Wan & Yiwei Sun; and Harris Wan & Ruchika Mishra.

Around the Department, 2023-2024

Grants, Awards, and More

Mohammed AlQuraishi, PhD, received 200,000 node-hours on Polaris to develop OpenFold-powered machine learning of protein-protein interactions and complexes.

Mohammed AlQuraishi, PhD, **Virginia Cornish,** PhD and **Harris Wang,** PhD, received a 3-year grant for DARPA for *MEBES: Modular engineered biosensors for environmental sensing.*

Mohammed AlQuraishi, PhD, received a 1-year grant from AbbVie for *Federated learning of protein-small molecule and protein-protein interactions.*

Cory Abate-Shen, PhD, received a 2-year grant from NCI for *Investigating mitochondrial dysfunction in high-risk prostate cancer.*

Andrea Califano, Dr, has been honored with the 26th Alfred G. Knudson Award in Cancer Genetics by the National Cancer Institute for his exceptional contributions to the field of cancer research.

Andrea Califano, Dr, has been elected fellow of the AACR Academy.

Barry Honig, PhD, received a 4-year grant from NSF for *Collaborative Research: NSF-BSF: How cell adhesion molecules control neuronal circuit wiring: Binding affinities, binding availability and sub-cellular localization.*

Barry Honig, PhD, received a 2-year grant from the Gates Foundation for *FEP for Biologics and Biosensors.*

Tal Korem, PhD, received a 5-year grant from NIAID (M-PI w/ Uhlemann) for *Microbial biomarkers of intestinal MDR colonization after solid organ transplantation.*

Tal Korem, PhD, received a 5-year grant from NICHD for *A large scale investigation of the vaginal ecosystem in preeclampsia.*

Tal Korem, PhD, received a 3-year grant from NIDDK (PI. Brennan & Kalan) for *Microbiome based biomarkers of wound healing.*

Raul Rabadan, PhD, received a 5-year grant (Co-I with Stavroula K), from NCI for Aging-related hematopoietic stem cell intrinsic and microenvironmental signals in AML transformation.

Peter Sims, PhD, received a 5-year grant (Co-I with Farber D.), from CZ Biohub for *Context-Specificity of Immune Cell Regulatory Networks.*

Michael Shen received a 5-year grant from NCI for *Investigating mechanisms of tumor plasticity in human bladder cancer.*

Michael Shen received a 2-year grant from Prostate Cancer Foundation (PCF) Challenge for *Modulation of anti-androgen response by NSD2 in castration-resistant prostate cancer.*

Harris Wang, PhD, received a 5-year grant from CZ Biohub for *Cell-corder: A transcriptional diary of immune cells using engineered RNA memory.*

Harris Wang, PhD, received a 2-year grant from NIAID for *Profiling diarrhea-predominant irritable bowel syndrome (IBS-D) using stool-based RNAs.*

Harris Wang, PhD, received a 2-year grant from LLNL for *From Sequence to Cell Population: Secure and Robust Biosystems.*

Harris Wang, PhD, received a 2-year grant from DOD for *An Integrated Solution for Automated Cold Cultivation using Robotics, AI, and Genomic.*

Xuebing Wu has been awarded the MIND Prize by Pershing Square Foundation and the Schaefer Research Scholar by Columbia University.

Chaolin Zhang, PhD, along with **Harris Wang,** PhD, were jointly awarded the Precision Medicine Pilot Grant by the Columbia Precision Medicine Initiative (CPMI), the Herbert Irving Comprehensive Cancer Center (HICCC), and the Irving Institute for Clinical and Translational Research.

Congratulations to **Tal Korem** and **Harris Wang** for being named Highly Cited Researchers by Clarivate (for the 4th consecutive year)!

Yocelyn Recinos was selected as a recipient of the 2024 Dean's Award for Excellence in Research in recognition of the high quality and significance of her PhD thesis research in the Coordinated Doctoral Programs in Biomedical Sciences.

Postdocs & ARS Receive Faculty Positions:

Alessandro Vasciaveo - Assistant Professor of Computational Biology and Artificial Intelligence at the Sanford Burnham Prebys Medical Discovery Institute in La Jolla, CA.

Haiqing Zhao - Assistant Professor, Department of Biochemistry & Molecular Biology, University of Texas Medical Branch

Guillaume Urtecho - Assistant Prof. of Medicine at UC San Diego in the Division of Gastroenterology & Hepatology

Congratulations to the new graduate students:

Tyler Perdue (Wang lab)

Guojie Zhong (Shen Y. lab)

Bulat Ziganshin (Shen Y. lab)

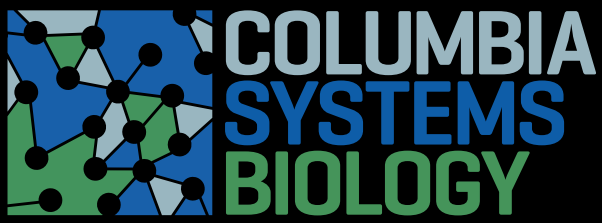
Sydney Blattman (Tavazoie lab)

Yige Zhao (Shen Y. lab)

Ester Calvo Fernandez (Califano lab)

Chrystal Mavros (Wang lab)

COLUMBIA UNIVERSITY
IRVING MEDICAL CENTER



Columbia University Department of Systems Biology
Irving Cancer Research Center
1130 St. Nicholas Avenue
New York, NY 10032

To learn more about our research and programs,
visit **systemsbiology.columbia.edu.**

Design by: Nicoletta Barolini
Edited by: Ann Rae Jonas