# AI for Science and Engineering

**Henry Kautz**
https://henrykautz.com
henry.kautz@gmail.com

Revised 29 January 2024

## Introduction

AI is revolutionizing science and engineering. No longer limited to data analytics where it provides a toolkit of methods for complex cases of regression and classification, AI is now being used in every stage of scientific inquiry, including literature review, hypothesis generation, experimental design, experiment execution, analysis, and communication of results. AI systems can now collaborate with engineers for ideation, exploration of design spaces, validation for manufacturability, and generation of detailed schematics and programs to drive machine tools. This report provides a survey of exciting, leading-edge work in AI for science, with a focus on the role of data sets in powering the revolution.

## Contents

# Creation of the report

We began by attending the NSF Workshop on the AI-Enabled Scientific Revolution (Alexandria, Virginia, March 2023), performing a literature review, and studying the work of a set of AI for science researchers who had been in earlier conversations with Science and Technologies Futures.

Through this process, we identified an initial group of scientists and engineers working at the leading edge of AI for science. Interviews with this group identified additional important AI for science projects and researchers, and in total, we interviewed 40 scientists and engineers from 33 different research groups. Each interview lasted one hour, and interviewees were provided with the following questions for initiating the conversations.

- How are you using AI in your and your colleagues' work; for example, for prediction, optimization, design, or hypothesis generation?
- What scientific, technological, and societal problems can the application of this work help solve?
- What datasets would have an impact on the role that AI could play in your field as large as that of the Protein Data Bank on AlphaFold and generative AI for the design of artificial proteins?
- Who should create or curate such data? Are there existing community efforts that could be built on?
- What benchmarks need to be created to evaluate progress – similar to CASP or Image Net?
- What other kinds of AI-related resources are needed; for example, hardware innovation such as self-driving labs, open software, large language models tailored for science, or others?

This report provides a summary of the interviews; full notes from each are on file with Science and Technologies Futures. The summaries are grouped as follows:

- Biology and Biochemistry
- Material Science, Inorganic Chemistry, and Physics
- Neuroscience and Cognitive Science
- Agriculture, Environmental, and Geological Science
- Mechanical Engineering and Design

- General AI for Science

This report was written by the author under contract to Science and Technologies Futures, Cambridge, MA, which has given permission for redistribution by the author. It is not an official publication of that organization.

# Common Themes

Several common themes emerged in the interviews, including

**Data curation**: Data curation is as important as generating data. Data is often siloed among hundreds of research groups using different formats and with different standards of quality. Unlabeled noisy data is often plentiful and cheap to obtain, while clean labeled data is rare and expensive to obtain.

**Theory-guided machine learning**: Methods for incorporating scientific knowledge in the form of simulators or mathematical equations directly into machine learning algorithms are vital for advancing AI for science. Incorporating knowledge can result in more accurate and usable models, and can drastically reduce the amount of needed labeled training data. In the limit, theory-guided machine learning supports few-shot learning, where only a handful of high-quality training examples are needed instead of hundreds of thousands or millions of examples.

**Accelerating scientific simulation with machine learning**: The complement of using theory to guide machine learning is using machine learning to accelerate scientific simulations. Doing so leverages the fact that a deep neural network (DNN) can learn to approximate any computable function. The DNN can be trained on data generated by a traditional first-principles simulator - for example, on the results of a physics simulator computing interactions between a small number of particles. The DNN can then predict the results of a new simulation in a fraction of the time of the first-principles simulator and can scale to problems that are too large for exact simulation.

**Proxy data**: When gathering primary data at scale is impractical, it may be possible to substitute data that is cheap and plentiful by using machine learning to predict primary measurements from the so-called proxy data. An example of this would be to use satellite data as a proxy for on-the-ground measurements. Accurate prediction from proxy data is often a significant scientific challenge in its own right.

**Large language models**: The use and potential of large language models (LLMs) are ubiquitous. In addition to using LLMs to extract structured information from scientific literature, LLMs can be used for a wide variety of tasks in the scientific method, including experiment design and hypothesis generation. LLMs can also be used for generating programs for data analysis chains and for analyzing artifacts such as CAD (computer-aided design) files that are essentially specialized programs.

**Self-driving laboratories**: The importance of self-driving labs is not just their speed or automation, but that they can intelligently explore a huge hypothesis space where exhaustive or random exploration is too costly or entirely infeasible. A self-driving lab can run a set of experiments - for example, where each is a hypothesized way to produce a chemical with desirable properties - and after analyzing the results, design and execute a new set of experiments. Self-driving labs can thus be thought of as just-in-time data-gathering systems. These systems are not limited to the laboratory but include autonomous information-gathering robotic systems such as are being deployed for deep ocean exploration.

# Filling Funding Gaps

The vast majority of industry investment in AI for science is focused on the development of pharmaceuticals. General language language models produced by OpenAI and other groups such Meta FAIR have begun to find use in scientific applications. Google Deepmind has performed groundbreaking pre-competitive research on AI for Science, most notably its work on protein folding and materials discovery.

The major funder of academic research in AI is the National Research Foundation. Although its funding for AI research, approximately $200 million a year, has not increased in decades, beginning in 2019 it began funding interdisciplinary National AI Research Institutes at $20 million each over 5 years. Of the 25 awards made by 2023, 8 are focused on AI for various scientific domains. In January 2024, NSF announced the National Artificial Intelligence Research Resource (NAIRR) Pilot, a joint effort with other federal agencies, corporations, and non-profit laboratories, which will have a large AI for science component.

The largest philanthropic investments in AI for science have come from Eric Schmidt and several other philanthropists with their support of two new focused research organizations. The first is Future House, which describes itself as “a moonshot focused on building an AI scientist”, with an initial focus on biology, and the second is Kyutai, located in Paris and dedicated to energizing AI for science and engineering in Europe. In addition to these efforts, the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship provides support for graduate students in the sciences who are working on AI for science projects. The Chan Zuckerberg Initiative, the Allen Institute, and the University of Washington have recently announced a $75 million partnership named the Seattle Hub for Synthetic Biology for AI-centric work on understanding gene expression.

There remain enormous funding gaps in AI for science and many opportunities for philanthropies to make a large impact with relatively modest investments - in the range of millions rather than billions of dollars.

The following list of gaps and opportunities links to corresponding projects described in this report. There are, of course, other worthy projects in each of the four categories.

**Support of the creation and curation of open, ethically-sourced, linked, and multi-scale data sets, including data gathered through benchmarking competitions.**

- Basecamp
- PathBank
- Metabolome Project (Cornell)
- Bacteria
- Institute for Digital Innovation
- International Brain Laboratory
- Harvest Africa Initiative
- AI Institute for Resilient Agriculture
- Circularity Informatics Lab
- Environmental Data Initiative
- Earthscope Consortium
- Design Computation and Digital Engineering Lab
- Digitally Integrated Manufacturing Environment Lab
- OpenCatalyst

**Support of development and maintenance of open AI software for science, including analysis tools, large language models, and simulators.**

- OpenFold
- OpenBioML
- PyThesus
- Neurodata Without Borders
- Earthranger
- Open Language Model

**Support of robotic laboratories and autonomous explorers that can be accessed by many different research groups.**

- Molecular Maker Institute
- Model-Based Embedded and Robotic Systems Group

**Support for foundational research on general methods for combining science knowledge in the form of equations and simulations with machine learning.**

- Theory Guided Machine Learning
- Deep Reasoning Networks
- Deep Geometric Reasoning

# Existing Open Datasets Supporting AI for Science

Following is a partial list of open scientific datasets currently being used by AI for science projects.

- Biology and Biochemistry
    - RCSB Protein Data Bank

    - The Human Metabolome Database (HMDB)
    - The Small Molecule Pathway Database (SMPDB)
    - ChemFOnt (the Chemical Functional Ontology)
- Material Science, Inorganic Chemistry, and Physics
    - NIST JARVIS (Joint Automated Repository for Various Integrated Simulations) Data Sets
    - Open Quantum Materials Database (OQMD)
    - The Materials Project
    - Center for Hierarchical Materials Design Databases
    - Inorganic Crystal Structure Database (ICSD)
    - NOMAD Materials Science Data
- Neuroscience and Cognitive Science
    - OpenNeuro Datasets
    - Collaborative Research in Computational Neuroscience (CRCNS) Data Sets
    - Human Connectome Project ConnectomeDB
- Agriculture, Environmental Science, and Geological Science
    - FooDB
    - USDA Ag Data Commons
    - NOAA OneStop Data Search
    - EarthScope Data Sets
- Mechanical Engineering and Design
    - DeCoDE Lab Mechanical Engineering Data Sets

# Other Recent Reports

Several other interesting reports about AI for science have come out in the past year, including the following:

Matthew Hudson, Hypotheses devised by AI could find 'blind spots' in research, *Nature Index*, 17 Nov 2023.

> Describes the history of research in AI on automated hypothesis generation including recent work by James Evans (University of Chicago) on network analysis of scientific papers in order to generate hypotheses that are both likely to be true and unlikely to be readily discovered by humans.

Tom Hope, Doug Downey, Oren Etzioni, Daniel S. Weld, and Eric Horvitz, A Computational Inflection for Scientific Discovery, *Communications of the ACM*, Vol. 86, No. 8, August 2023.

> Provides an overview of what the authors call "task-guided scientific knowledge retrieval, in which systems counter humans' bounded capacity by ingesting corpora of scientific knowledge and retrieving inspirations, explanations, solutions, and evidence synthesized to directly serve task-specific utility."

Microsoft Research AI4Science, The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4, Technical Report, November 2023.

Detailed overview and evaluation of the use of large language models for scientific research in drug discovery, biology, computational chemistry (density functional theory and molecular dynamics), materials design, and partial differential equations.

Merchant, A., Batzner, S., Schoenholz, S.S. et al. Scaling deep learning for materials discovery. *Nature* (2023). https://doi.org/10.1038/s41586-023-06735-9.

Paper from DeepMind announcing ability to identify 2.2 million novel crystals using deep learning.

# Biology and Biochemistry

## Glen Gowers and Oliver Vince (Basecamp): Novel proteins from nature, ethically sourced

https://www.basecamp-research.com/

Basecamp is a venture-backed startup with the mission to bridge biodiversity and biotechnology. It performs data collection and curation for biology. Our knowledge of how nature works is almost non-existent; most genetic knowledge is based on 20 organisms, but there are trillions. Further, function cannot be determined just by looking at individual organisms; to understand DNA's role one needs contextual information. Historically data has been "pirated" from around the world. However, the Nagoya Protocol 2014 on ethical access to biodiversity changed the landscape for collecting biological specimens. It now takes years to get approvals and is a serious bottleneck for science. Basecamp solves this problem by partnering with biodiversity hotspots, enabling researchers and companies to gain access to biological data in a way that fully complies with Nagoya Protocol. It builds bio literacy broadly and returns money to biodiversity areas.

Currently, Basecamp makes its data available to paid users, academics (with limitations), and to scientists in the countries where samples are gathered. However, it is in the process of creating the Basecamp Foundation to make data more widely available. A foundation could be a founding partner in launching the Basecamp Foundation, thus opening up the world's largest dataset of in-context biological data. Basecamp's data is organized as a knowledge graph that connects protein sequences to the contexts in which they act.

In addition to its data infrastructure, Basecamp has built solar-powered bio-collection labs from off-the-shelf components. These units are currently deployed on a research ship, in Wales, and Cameroon. These labs cost $15,000 each. Support from a foundation would dramatically expand the ability of scientists in developing nations to do DNA sequencing (today there are only two laboratories in all of Africa with DNA sequencing capability) by funding a set of these kits.

## Mohammed AlQuraishi (Columbia): OpenFold: Open-source software for protein prediction and drug discovery

https://openfold.io/

OpenFold is a non-profit AI research and development consortium developing free and open-source software tools for biology and drug discovery.  It has 12 employees and 10 corporate partners as well as numerous academic partners. Its areas of focus are protein design, protein-molecule interactions, predictions of ensembles of forms, and prediction of how mutations affect shapes.  OpenFold has many ideas for expanding and curating biochemistry databases, including:

- A dynamics and interaction protein database that would complement the current Protein Data Bank (which contains only static information).
- A protein-ligament DB created using federated learning across private corporate repositories.  OpenFold is beginning to organize this effort.
- A molecular dynamics data repository, to curate data that is now distributed and hard to access.
- A cryogenic electron microscopy (Cyro-EM) database.  Cyro-EM is a cheaper alternative to crystallography that is taking off, but for which there is not yet a large public data repository.

## Frank Schroeder and Carla Gomes (Cornell University): Revealing the metabolome with deep reasoning networks

https://btiscience.org/schroeder/

https://www.cs.cornell.edu/gomes/

DNA, RNA, and proteins are all constructed from a small set of amino acids.  There are, however, hundreds of thousands of other kinds of organic molecules, the so-called metabolome, that control cell signaling, immune response, fetal development, and countless other functions. So little is known about the chemistry of the metabolome that biochemist Frank Schroeder calls it “the dark matter of life”.  Schroeder and AI expert Carla Gomes are using new ultra-high precision spectroscopy and new algorithms that combine expert knowledge of chemistry with deep learning to vastly expand our knowledge of the structure and function of these small organic molecules.  The results will lead to cures for diseases for which traditional pharmaceutics offer modest benefits, including diabetes, obesity, asthma, cancer, autoimmune diseases, and developmental diseases.  Gomes leads several projects that combine machine learning with scientific knowledge and constraint-based reasoning under the general name of “deep reasoning networks”.

## David Wishart (University Alberta): Pathbank: A Comprehensive Pathway Database

https://www.wishartlab.com/
https://pathbank.org/

PathBank is a comprehensive, visually rich pathway database containing more than 110 000 machine-readable pathways found in 10 model organisms. PathBank aims to provide a pathway for every protein and a map for every metabolite. This resource is designed specifically to support pathway elucidation and pathway discovery in transcriptomics, proteomics, metabolomics and systems biology. It provides detailed, fully searchable, hyperlinked diagrams of metabolic, metabolite signaling, protein signaling, disease, drug and physiological pathways.

## Russ Greiner (University of Alberta): Machine Learning for Disease Prognosis and Treatment

https://apps.ualberta.ca/directory/person/rgreiner
https://rgreiner6.wixsite.com/greiner

Greiner is one of the world's top machine learning researchers working in healthcare. A few of his most significant contributions to date include:

- A novel way to predict the survival curve for an individual – basically a Precision Health version of Kaplan-Meier plot, that takes advantage of all the information available about a patient, including their stage of cancer, age, gender, histology, and many other factors. This tool can be used to estimate the time-to-death associated with different treatments, to identify the best one; or to estimate time-to-onset of cancer, to encourage a person to modify their lifestyle.
- Prognostic models that can accurately predict cancer relapse, mental health conditions and response to treatment, damage from ischemic stroke, and others.
- Models for optimal treatment planning of diseases including diabetes and cancer.

## Paul Jensen (University of Michigan): Bacteria: using reinforcement learning and robotics for microbiology

http://jensenlab.net/

Dr. Jenson is a microbiologist who has been using reinforcement learning to determine the optimal growth media for microbes. As did the scientists at Basecamp, he stressed how little knowledge we have of nearly all microbes: for example, for 3/4 of microbe species, there is not even a single paper, and 50% of the literature is about just 13 of the 45,000 known species. New species are rapidly being discovered by large sequencing campaigns, but we don't know how each responds to antibiotics and environmental conditions.

## Amarda Shehu (George Mason University): Institute for Digital Innovation: Bridging biology and AI research

https://idia.gmu.edu/

Dr. Shehu noted that a key AI approach moving forward is the incorporation of scientific theories represented as differential equations into learned models, thus bridging simulation and machine learning.  Two bottlenecks in AI for science are biologists not using AI correctly and AI researchers not working on real biology problems. A central mission of the Institute for Digital Innovation that she helps lead is to support collaboration and cross-training of graduate students in the computational and biological sciences.

# Material Science, Inorganic Chemistry, and Physics

## Martin Burke and Huimin Zhao (University of Illinois):  Molecular Maker

https://moleculemaker.org/

There is a limit to what can be learned from a retrospective analysis of a chemical database. However large, a static database must capture only a tiny fraction of the exponential number of combinations of molecules and conditions that could exist.  Martin Burke stresses the importance of closed-loop discovery through robotic laboratories, AI-assisted hypothesis generation, and AI-assisted experiment design.  Such an approach can make discoveries by intelligently exploring a tiny fraction of the combinatorial space and has been proven to work in his own work on small-molecule design.  The key technologies are

- “Function first” molecule discovery.  For the last 200 years, chemistry has been structure-first, but there are many different structures that perform the same function. The Molecular Maker Lab Institute (MMLI) is unique in making functional approach primary: representation based on functions such as reactivity, ability to penetrate blood-brain barrier, lifetime until degradation, etc.
- Biofoundaries.  Central facilities with robotic laboratories and full-time staff that make their resources available remotely to scientists across the nation, including faculty based at community colleges and under-resourced colleges and universities. MMLI built the first biofoundry in 2014 and initiated a global biofoundry alliance in 2019 (https://www.biofoundries.org).

## Krishna Rajan (University of Buffalo): Materials Design and Innovation: Materials Connectomics

https://engineering.buffalo.edu/materials-design-innovation.html

Krishna Rajan is the founding chair of a new kind of material science department and graduate program based on integrating science and information technology. To date, material science datasets are highly varied, unconnected, and based on different benchmarks. The biggest need is for linked data sets at different length scales, including the quantum, atom, and molecular scales.

## James Warren (NIST, National Institute of Standards & Technology): The Material Genome Initiative

https://www.mgi.gov/

The MGI started in 2011 and has helped support the creation of a number of important data resources for density functional theory, such as Calphad and NanoMine. He noted that MGI has a messaging problem: many people think that quantum databases are all that MGI is about, but while quantum is a useful starting point, we need a hierarchy of databases at all scales.

## Jonathan Godwin (Orbital Materials): Molecular materials design

https://orbitalmaterials.com/

Jonathan Godwin founded Orbital Materials to use AI for material discovery. He described the field as divided between the “machine learning data absolutists” and the “simulation-based” researchers. He spoke of the need to incorporate true physics into the models in order to handle hard real-world problems. He stressed the need to have widespread community buy-in for the design of a challenge dataset - the “build it and they will come” approach doesn’t work.

## Larry Zitnick (Meta): The Open Catalyst Project

https://opencatalystproject.org/

Larry Zitnick organized the COCO visual Q/A challenge, the FAST MRI challenge, and now the Open Catalyst challenge. The goal of the challenge is to encourage the creation of machine learning models that can match the performance of density function theory computations (DFT) but many orders of magnitude more quickly; ultimately, the hope is that the ML models will handle molecules that are too large for DFT calculations. Meta provides the extraordinary computing power necessary for the calculations - only Meta or Google have such resources - and the calculations still take weeks. The data for the challenge is thus not laboratory data but the result of DFT calculations, which are accepted as ground truth with the caveat that even with Meta’s computing resources the algorithms are approximations based on relaxations of DFT. Larry said the ingredients of a successful ML challenge are:

- Having domain experts involved in defining the task;
- Ensuring the data is ready for analysis by AI researchers who do not understand the domain;

- Defining an evaluation metric that is accepted by the domain scientific community and that also allows progress by AI researchers to be measured;
- Including both a continuous leaderboard and annual competition at a prominent venue, where the annual competition may include evaluation of results by human experts as well as automated methods;

He noted that although Meta worked with only one research group at CMU to create the challenge, they ran their ideas past a group of senior chemistry experts and took their feedback into account to help with community buy-in. Over three years the best performance by ML methods has risen from 0% to 16%.

## Andrew White (University of Rochester): Open Bio/ML: Large language models for chemistry:

https://thewhitelab.org/

Andrew White is the leading proponent of using large language models in chemistry. He is a member of the Open Bio/ML group (https://openbioml.org/) that is funded by Stability AI and is working with AI2 (see below) on knowledge extraction from the Semantic Scholar Open Data Platform. He mentioned ChemEngine from Oloren as a great open-source library for molecular property prediction, and that a funder might work with them to create and fund a chemical prediction competition. White is currently helping launch a new biotech startup, Future House.

## Mario Krenn (Max Planck Institute): PyThesus AI for designing quantum experiments

https://mpl.mpg.de/research-at-mpl/independent-research-groups/krenn-research-group/

Mario Krenn created the PyThesus system that designs new quantum optics experiments to answer fundamental questions in physics. PyTheus has in hours designed experiments that would take people months or years to design. Automatic experiment design requires a well-defined metric, a very large search space that humans cannot explore, and a reliable simulator. Current physics simulators are a weak link; he suggests creating a consortium of physicists and material scientists to create a general and robust simulator.

# Neuroscience and Cognitive Science

## Bing Bruton, Elizabeth Buffalo, and Oliver Ruebel (Neurodata without Borders): Making brain data more usable and accessible

https://www.nwb.org/

NWB provides users with a mature software ecosystem that enables users to 1) efficiently create and use NWB data files via Python and Matlab APIs, 2) extend the data standard via Neuro Data Extensions (NDX), 3) share and deploy these extensions to the community via the extensions catalog, and 4) explore, convert, and document NWB data via high-level tools and utilities. Neuroscience labs are increasingly using NWB to process and analyze data in their research and to store and share their data with collaborators and the public.

## Michael Hausser and Hannah Bayer (International Brain Laboratory): Reusable and replicable brain data through coordinated experiments

https://www.internationalbrainlab.com/

The standard model for animal neuroscience research is broken: researchers place sensors in one brain area - typically recording a few dozen neurons - and gather data for one behavior. The relatively small amount of data collected cannot be combined with data gathered by other groups because of differences in methodology, and so is lost forever. The IBL, by contrast, organizes different research groups to use a common methodology, record across the entire brain, enforce data quality, and make all resulting integrated data sets open. It has put together a standard hardware suite and employs a team of 10 staff members to build core software for data collection and analysis. IBL works closely with NWB.

## Aude Oliver (Massachusetts Institute of Technology): Cognitive science open data

http://olivalab.mit.edu/audeoliva.html

In contrast to neuroscience data sets, there are relatively few cognitive scientists who make data available about experiments where fMRI data is collected while human subjects perform tasks. An exception is Aude Oliver, who has published datasets for activities such as watching short video clips, hearing meaningful sounds, and using short-term memory. The data sets are still relatively small (100 to 1000 subjects). She theorizes that there are only a few thousand categories of basic human brain experience; to date, her largest experiments considered a few hundred categories.

# Agriculture, Environmental, and Geological Science

## Catherine Nakalembe (University of Maryland): Harvest Africa Initiative

https://geog.umd.edu/facultyprofile/nakalembe/catherine

Catherine Nakalembe works on data and analytics to measure agriculture in Africa and other parts of the developing world that lack the data collection infrastructure (such as instrumented tractors and USDA reporting systems) that are used in the US and Europe.  Gathering data on crop types, crop boundaries, yields, and crop disease is a challenge.  She has begun to crack the problem by integrating on-the-ground data from cell phone apps her group created (including what she calls a “Streetview for crops”) with satellite data in order to create models that can predict desired features (e.g. crop type) from the satellite images.  Her group has a “kit” program to send people what they need to participate in the Street to Sat project. It has sent out 60 kits so far and could do a lot more with additional support.  Her CropHarvest project is an open-source remote sensing dataset for agriculture with benchmarks. It collects data from a variety of previously fragmented agricultural land use datasets and remote sensing products.  She received the 2020 Africa Food Prize Laureate award.

## Baskar Ganapathysubramanian (Iowa State University): Global co-registered data to support AI for agriculture

https://aiira.iastate.edu/

Baskar Ganapathysubramanian leads the USDA-NIFA AI Institute for Resilient Agriculture (AIIRA). He envisions several transformational “moonshot” data sets; the most compelling was to create a global co-registered dataset from satellite, drone lidar, ground images, and soil sensors.  Example questions this data could help answer include: How can we extract hyperspectral signatures from satellite data that predict what the fine-grained ground-based sensors reveal? How can we predict yield from the first 6 weeks of growth? This work will dramatically drive down the cost of gathering agricultural data and support more powerful ways for AI to improve agricultural production while helping the environment.

## Elsa A. Olivetti (Massachusetts Institute of Technology): Environmental and economic sustainability of materials

https://dmse.mit.edu/people/elsa-a-olivetti

Elsa Olivetti works at the intersection of materials science and environmental science. A unique focus of her research is the creation of materials that lend themselves to more effective recovery - in other words, producing materials that make recycling cost-effective. She envisioned the creation of a general data set for improving recovery processes by linking materials composition, quality, location, and byproducts with collection rates and recovery data to inform where materials recovery and shared symbiotic use of inputs and outputs would be more effective.

## Jenna Jambeck (University of Georgia): Circularity Informatics Lab

https://www.circularityinformatics.org

Jenna Jambeck was named a 2022 MacArthur Fellow for her work investigating the scale of plastic pollution and galvanizing efforts to address plastic waste. Her lab has created software that is being used to collect data in collaboration with local partners worldwide.

## Jordan Read (CUAHSI, Consortium of Universities for the Advancement of Hydrologic Science, Inc): AI for water science

https://www.cuahsi.org/

CUAHSI provides cross-institution data and education services in water science. It has 100+ members (graduate degree-granting institutes) and 15 employees. Water data is published by the US Geological Survey, NASA, and NOAH. The CAMELS Large-Sample Hydrometeorological Dataset has long been used for benchmarking water prediction models, but it is limited to large catchment areas and North America. He argued that current water data sets are targeted at the lowest common denominator and miss vital details such as human impact on the water cycle. With better data, he envisions the ability of AI models to handle dream challenges such as predicting rare extreme events (droughts and floods), toxic algae bloom, and optimizing when and where to sample water (a costly manual process) to maximize information gain.

## Paul Hanson (University of Wisconsin): Environmental Data Initiative: AI for aquatic ecology

https://hanson.limnology.wisc.edu/

There are 7 million lakes and reservoirs in the US, but there are only 100,000 for which we have more than one data point, and only 15,000 for which we have good bi-weekly data. Paul Hanson is PI for the NSF-funded Environmental Data Initiative, which has the mission to preserve environmental data for open and reproducible science, to promote synthesis across space and time, and to aid in the assessment of environmental change and its consequences. It is a virtual organization with 10 FTE scientists. He wished that there were more collaborations between water, ecology, and AI researchers using the data and informing EDI on how to improve their data sets.

## Amy McGovern (University of Oklahoma): Trustworthy AI for Weather, Climate, and Coastal Oceanography

https://www.ai2es.org

Amy McGovern leads the NSF AI Institute on Trustworthy AI for Weather, Climate, and Coastal Oceanography (AI2ES). Although AI2ES's focus is broader than water science, there was much overlap in Amy McGovern's comments with those of Jordan Read and Paul Hanson. She noted the scarcity of good on-the-ground data in the developing world and the importance (as yet a dream) of being able to predict extreme weather water-related events such as droughts and floods.

## Becks Bendick (EarthScope Consortium): Open global geoscience data

https://www.earthscope.org/

The Earthscope consortium holds the world's more detailed geophysical data, and is in the process this year of making all 10 petabytes of its data available in a cloud data analytics platform. By dramatically increasing access to this data, they hope that research will be able to dramatically increase our understanding of how earth systems interact with each other by cross-correlating meteorological, atmospheric, seismological, and geological data.

## Jes Lefourt and Ted Schmidt (Allen Institute for AI): Earthranger: Protecting wildlife with real-time data

https://www.earthranger.com/

Earthranger is a data infrastructure system for wildlife protection that integrates information from camera traps, acoustic sensors, tracking devices on wildlife, vehicle trackers, and wildlife rangers using mobile devices. The original use case was anti-poaching, but now it is being used for other operational needs in wildlife management, including herd tracking and monitoring invasive species. Earthranger was created at Paul Allen's Vulcan company and now has its home at AI2. Skylight is a more recent system developed for monitoring illegal fishing using satellite data.

## Brian Williams (Massachusetts Institute of Technology) and Rich Camilli (Woods Hole Oceanographic Institute): Model-based autonomous systems for oceanography

https://groups.csail.mit.edu/mers
https://www.whoi.edu/

Brian William's research group at MIT builds autonomous robotic systems for long-term deployment in remote and hostile environments, and Rich Camilli's team from Woods Hole deploys them for oceanographic research. What makes their approach to oceanography unique is that not all mission parameters are fixed in advance: the A-UWV (autonomous underwater vehicle) can adjust its course in reaction to what it has discovered to maximize the value of the overall mission. This is important because the ocean is simply too big to survey in a mindless grid pattern - current data sets, gathered by traditional fixed survey patterns, are very sparse once one goes beyond coastlines. Past and current campaigns include studying the diversity of coral reefs, locating underwater volcanoes, measuring the impact of underwater noise from drilling and wind turbines on sea mammals, and exploring under arctic ice.

# Mechanical Engineering and Design

## Faez Ahmed (Massachusetts Institute of Technology): Design Computation and Digital Engineering: Automating mechanical design

https://decode.mit.edu/

Although computer-aided design is a huge area of commercial activity, there is relatively little research to date on automated design. Faez Ahmed is trying to change this through his work on the automated design of healthcare devices, bicycles, and homes. The biggest challenge for AI design is the lack of datasets because industry does not release them. He wrote a wishlist of datasets:

- Kinds of movement and force transfer
- Manufacturability
- Annotated CAD models

He stressed that CAD models without annotation are of little use - for example, researchers at NYU collected a million-entry database of unannotated CAD models, but it has been rarely cited. Annotations could be many things and often depend on the end applications, and could include fluid/structural performance, material properties, fluid-related performance, manufacturability, systems systems-related performance. Creating such a resource would require lots of computing resources, mechanical domain expertise in collecting a meaningful and large dataset, and expertise in training very large multimodal LLMs.

## Daniele Grandi (Autodesk): AI and knowledge representation for mechanical design

https://www.autodesk.com/products/fusion-360

Daniele Grandi is a researcher at the largest CAD software company, where his work focuses on generative design assistants. He believes that AI for design must aim for interactive rather than fully automated systems because the user doesn't know in advance the full specification of what he wants; the requirements change as the user explores the design space. His holy grail would be a data-driven tool that looks at millions of designs and learns an underlying knowledge base - for example, it would learn that boats need to be corrosion-resistant, cars need to use fasteners that are vibration-resistant, and so on. He also noted that while there are a few existing data sets of classes of objects (e.g. ship hulls), the field needs to shift to more generalizable data sets.

## Binil Starly (North Carolina State University): AI for Manufacturing Design

https://www.dimelab.org/

Binil Starly heads the Digitally Integrated Manufacturing Environment Laboratory, which is advancing data-driven product design and the use of knowledge graphs in manufacturing. The biggest need for advancing AI in mechanical engineering, design, and manufacturing, he argues, is for broad, high-quality open datasets. Most existing datasets are either very narrow - many versions of a few objects - or based on designs of questionable quality created by amateurs. His FabWave database of high-quality annotated designs is an exception but still needs to grow much larger. The main bottleneck is that the vast majority of good designs are proprietary. He envisions the creation of a design marketplace that would incentivize companies to share design data.

## Adriana Schulz (University of Washington): Understanding CAD with self-supervised learning

https://homes.cs.washington.edu/~adriana/

Datasets for AI for design need to be of high quality, have high precision, and be extensively annotated with the information needed for the object to be manufactured. Adriana Schulz is pursuing an approach to AI for design that drastically reduces the amount of data needed by incorporating deep geometric reasoning in the learning algorithm. Standard computer-aided design (CAD) files are essentially programs for generating 3D objects by combining geometric primitives. She uses deep learning methods similar to those used in large language models to convert these geometric programs directly into AI-ready data with annotations needed for assembly and manufacturing.

## David Lattanzi (George Mason University): AI for civil infrastructure monitoring and design

https://volgenau.gmu.edu/profiles/dlattanz

David Lattanzi works on monitoring the health of civil infrastructure such as bridges, buildings, and pipes. Today the most sophisticated way AI is used in practice is the use of robots and drones to gather inspection data. He has begun early work on using AI to innovate in the engineering design space, exploring questions such as: can we teach AI systems human design standards? Can they create innovative designs? Can they help designers see design choices they overlooked? NIST is trying to help build a community of practice for AI for civil engineering, but as a whole the civil engineering community is very conservative and slow to embrace AI or even data science. He argued that we need a convening to try to define what should be in an AI for civil engineering database, beginning with an ontology of what is a building, bridge, or city.

## Pascal Van Hentenryck (Georgia Institute of Technology): AI for building robust supply chains

https://www.ai4opt.org/

Pascal Van Hentenryck leads the NSF AI Institute for Advances in Optimization. The Covid-19 pandemic has shown that many of the supply chains in the US did not have the necessary resilience to sustain a shock. They were optimized for efficiency and had many single points of failure. There is much interest in using AI-driven optimization to design robust supply chains. However, this area of engineering severely lacks test cases and data because companies are very protective of their data. However, one of the interesting observations in supply chains is that many companies share the same models (within reasonable variations), although they consider these aspects of their operations as highly confidential. Companies have expressed interest in developing some generic test cases that capture the complexities of actual supply chains without revealing confidential information. At this point, it should be possible to establish the following methodology:

- Develop an abstract model of a supply chain in a specific sector;
- Populate the model with synthetic data that will reproduce some of the computational and scalability challenges encountered in actual operations;
- Develop distributions of inputs that are realistic to explore resilience and efficiency issues and their tradeoffs;
- Develop stress tests and scenarios for recovery from significant disruptions.

# General AI for Science

## Dan Weld (Allen Institute for AI): Open scientific literature and language models

https://allenai.org/

Semantic Scholar is a scholarly paper index optimized for scientific use that has been building in size and functionality at AI2 for a decade. OLMo is a new large language model designed for AI for science. Both have received recent funding from NSF for building and supporting core infrastructure - an STTR Phase II award ($1M) for Semantic Scholar and a CISE Community Research Infrastructure award ($2M) for OLMo. OLMo will be both a tool for AI for science and a resource for the science of language models - it will make both models, parameters, and training data completely open, provide data from a large number of ablation experiments to help understand what goes on under the hood, and provide many tools for data cleaning and management.

## Vipin Kumar (University of Minnesota): Theory-guided machine learning for science

http://www.cs.umn.edu/~kumar/

Vipin Kumar is a leader of the AI and data science for science movement, having made many contributions to ML theory, applications in climate, energy, and biodiversity, and organizing events such as the AI-Enabled Scientific Revolution workshops at NSF headquarters in February and at KDD in August of this year. The key computational approach is theory-guided machine learning, which combines ML methods such as neural networks or probabilistic models with the kind of differential-equation-based models used in traditional scientific simulation. His paper "Theory Guided Data Science: A New Paradigm for Scientific Discovery From Data" is an excellent technical introduction that illustrates four different ways of using theory to strengthen ML models (using theory for initialization, regularization, optimization, and refinement) and two ways to use ML to strengthen simulations (data assimilation and parameter calibration).