Stanford University
Human-Centered
Artificial Intelligence

NEWS

BLASTNet – The First Large Machine Learning Dataset for Fundamental Fluid Dynamics

DATE NOVEMBER 20, 2023

TOPICS MACHINE LEARNING

ISTOCK/ARTEM ZAKHAROV

By collecting data from the field of computational fluid dynamics into a single dataset, AI researchers at Stanford hope to do for rocket science, oceanography, and climate modeling what web-scale data did for language.

For decades, scientists in many fields have used complex mathematics to predict the turbulence of fire, water, air, and other fluids to forecast weather, improve rocket propulsion, and estimate the effects of climate change on air and ocean. Fluid flows are indeed ripe for AI, but the field of fundamental flow physics lacks a suitable large dataset such as CommonCrawl for text or ImageNet for photos.

All that may soon change, however. Researchers at Stanford University set out to fill the void with BLASTNet, the first large dataset tailored for training machine learning models for fundamental fluid dynamics. We talked to Wai Tong Chung, a doctoral candidate in mechanical engineering and first author of a new preprint paper introducing the dataset to the field.

Why does such a dataset not exist already?

The reason is scientific data is super high-dimensional. To compare scientific data with text data used for large language models (LLMs), GPT-3 was trained with six hundred gigabytes of text data — BLASTNet is five *terabytes*. It is a geometric problem. Text is one-dimensional (1D). A photo is 2D. Flowfields in fluid dynamics are typically 4D (three in space plus one in time). Thus, in a uniform grid, for every 100 data points in 1D space, there are 10,000 in 2D space, and 1,000,000 in 3D space. For 4D fluids, that means 100,000,000 data points. If you're simulating a fire, for instance, there are complex phenomena driving the flames — the chaotic nature of turbulence, variations in temperature, available fuel, pressure, chemical reactions, and so forth — that increase the complexity of the problem. So, the mathematics for generating and training with this data is computationally expensive — even for supercomputers. This is one reason AI for science is immature compared with language models — we're somewhere near the year 2009 in terms of AI maturity compared with the multimodal LLMs we see today, since we don't really have web-scale data for a lot of fundamental physical phenomena yet.

What does it take to put something like BLASTNet together?

We took a community-driven approach and crowdsourced data generation to experts in the field and encouraged them to share their computational fluid dynamics data with us. We wanted all this data in one place in an easy machine learning-ready format. You'll notice this is actually BLASTNet 2.0. It took a year for us to grow the database from BLASTNet 1.0 — which was a proof of concept that was not ready for machine learning purposes. With BLASTNet 2.0, the dataset has some 700 samples from over 30 different configurations. Now, BLASTNet is large and diverse enough for use by other scientists for all sorts of machine learning problems. There's some really nice sustainability stuff in there, because a lot of our data is focused on carbon-free hydrogen fuels. The U.S., the EU, and Japan are very interested in transitioning toward it as a fuel. One way BLASTNet could help is to accelerate this transition toward a carbon-free future.

What are the benefits of curating this data into a single webpage?

This is all tied to having a common consistent format and access from a single portal. As reviewed in our recent NeurIPS paper, there is some availability of fundamental flow physics data spread across different sources. Downloading and reformatting this data can involve a lot of labor, especially when dealing with terabyte-scale datasets. Now that fluid flows have a consistent dataset that is free, open-source, and large enough for deep learning purposes and benchmarking new models, there are lots of directions things might go. Published papers in machine learning usually include a comparison of the new approach against others that have gone before. When everyone works on the same dataset, that alone improves these comparisons. But today, they are really only done in natural language processing and computer vision, where the datasets are suitably large. We're beginning to see this being practiced in scientific machine learning — so this dataset can really help foster these practices into flow physics.

How might BLASTNet advance other scientific fields?

As you know from ChatGPT, DALL-E, and others, machine learning is a powerful tool, but there are still many unexplored applications in fluid flows. Maybe this data can be used to train AI models that help us better understand the behavior of hydrogen or discover new operating regimes that lead to a carbon-free jet engine. Maybe AI models can learn better turbulence models from this data that can optimize wind farms to improve renewable energy. Whether it's wind, water, hydrogen, or any other fluid, these phenomena are governed by the same conservation principles. My colleagues and I like using it to improve our understanding of jets and rockets, but it could improve

many other applications, such as climate modeling, ocean currents, weather forecasting, maybe even medicine — anywhere liquids and gases are found. The plan now is to expand to those domains. Part of that challenge was creating the dataset. Part of it is getting engineers and scientists to talk to one another, because those of us who work in propulsion don't really talk to people in ocean modeling (or other fluid domains), so there's a lot of opportunity for BLASTNet to grow. One way we've tried to foster these collaborations is by hosting a virtual workshop around this data — which recently attracted more than 700 participants!

How do you plan to use BLASTNet yourself?

My colleagues and I have already used BLASTNet data for different flow physics problems — since we've had "beta" access to this dataset as BLASTNet's curators. Some of our past work includes developing machine learning algorithms that can automatically discover new physics models. Another application involves training deep learning models to improve the quality of flow physics simulations — a problem that was also tackled in our recent NeurIPS paper via computer vision techniques. Outside of physics, this large dataset is also suitable for investigating quantization and compression techniques that are really in vogue for lowering the cost of LLMs these days. We also recently used this data to host an open Kaggle competition for modeling turbulence with machine learning. So there's definitely a bunch of open problems in fundamental physics and machine learning that we're pretty excited about.

BLASTNet is available on GitHub.

*Stanford HAI's mission is to advance AI research, education, policy and practice to improve the human condition. **Learn more**.*

SHARE

CONTRIBUTOR(S)

Andrew Myers

Related News

NEWS

Stanford AI Scholars Find Support for Innovation in a Time of Uncertainty

Nikki Goth Itoi

JUL 01, 2025

Stanford HAI offers critical resources for faculty and students to continue groundbreaking research across the vast AI landscape.

NEWS

Stanford AI Scholars Find Support for Innovation in a Time of Uncertainty

Nikki Goth Itoi

MACHINE LEARNING FOUNDATION MODELS EDUCATION, SKILLS JUL 01

Stanford HAI offers critical resources for faculty and students to continue groundbreaking research across the vast AI landscape.

NEWS

Digital Twins Offer Insights into Brains Struggling with Math — and Hope for Students

Andrew Myers

JUN 06, 2025

Researchers used artificial intelligence to analyze the brain scans of students solving math problems, offering the first-ever peek into the neuroscience of math disabilities.

NEWS

Digital Twins Offer Insights into Brains Struggling with Math — and Hope for Students

Andrew Myers

MACHINE LEARNING SCIENCES (SOCIAL, HEALTH, BIOLOGICAL, PHYSICAL) JUN 06

Researchers used artificial intelligence to analyze the brain scans of students solving math problems, offering the first-ever peek into the neuroscience of math disabilities.

NEWS

Better Benchmarks for Safety-Critical AI Applications

Nikki Goth Itoi

MAY 27, 2025

Stanford researchers investigate why models often fail in edge-case scenarios.

Better Benchmarks for Safety-Critical AI Applications

Nikki Goth Itoi

MACHINE LEARNING MAY 27

Stanford researchers investigate why models often fail in edge-case scenarios.

Stanford University
Human-Centered
Artificial Intelligence

NAVIGATE

About

Events

Careers

Search

PARTICIPATE

Get Involved

Support HAI

Contact Us

Stay Up To Date

Get the latest news, advances in research, policy work, and education program updates from HAI in your inbox weekly.

Sign Up For Latest News

Stanford University