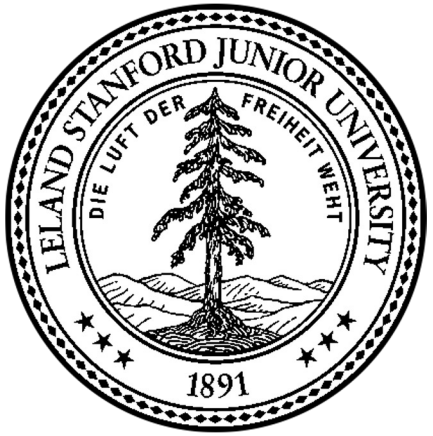
is AI what's NEXT for FLUIDS?



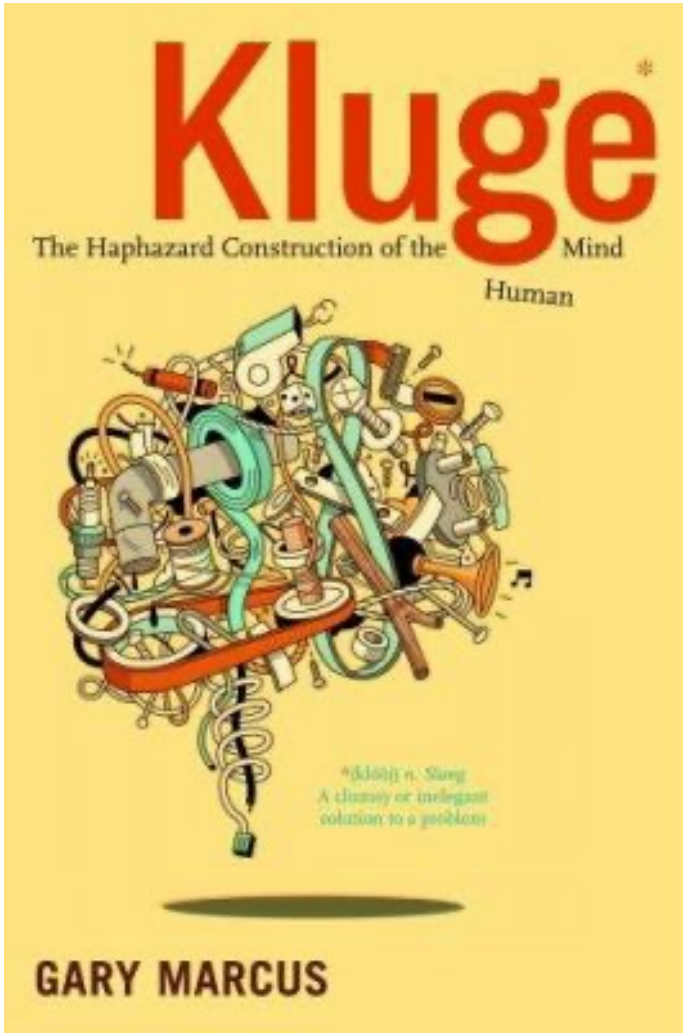
Gianluca Iaccarino
jops@stanford.edu
Mechanical Engineering
Stanford University

ML4FLUIDS: 3rd ERCOFTAC Workshop of Machine Learning for Fluid Dynamics

ML ≠ AI ≠ LLM

(A)Intelligence

intelligence is not well defined or measured



- intelligence as an ability: perceive, memorize, adapt, etc.
- choice (agency) is a critical human ability

This talk:

Focus on AI & Agency

stream of consciousness
rather than comprehensive assessment

Published as a conference paper at ICLR 2025

AGENTIC AI FOR SCIENTIFIC DISCOVERY: A SURVEY OF PROGRESS, CHALLENGES, AND FUTURE DIRECTIONS

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes & Christina Mack
IQVIA
{firstname.lastname}@iqvia.com

ABSTRACT

The integration of Agent
research automation. The
tonomous decision-makin
view, generate hypothese
vey provides a comprehe
categorizing existing syst
fields such as chemistry,
tion metrics, implementat
detailed understanding of
cal challenges, such as lite
concerns, while outlining
laboration and enhanced s

1 INTRODUCTION

The rapid advancements of Large L
Achiam et al., 2023) have opened
et al., 2024; Guo et al., 2023; Wa
tools for automating complex rese
designed to operate with a high de
such as hypothesis generation, lit
systems have the potential to signi
access to advanced tools across var

Recent efforts have demonstrated
with tasks such as literature review
including LitSearch (Ajith et al.,
et al., 2024c), CiteME (Press et a
tory (Schmidgall et al., 2025), have
citation management, document dis
often lack the domain-specific focu
domain, where the structured asses
Agent Laboratory demonstrated hi
writing. However, its performance
inherent challenges of automating
reliability, reproducibility, and ethi

This survey aims to provide a co
categorize existing systems into au

Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

---

turbulence.ai: an end-to-end AI Scientist for fluid mechanics

Jingsen Feng (冯晶森)[a], Yupeng Qi (亓宇鹏)[b], Ran Xu (徐冉)[c], Sandeep Pandey[d], Xu Chu (初旭)[a,e,*]

[a] Faculty of Environment, Science and Economy, University of Exeter, , Exeter, EX4 4QF, United Kingdom
[b] Cluster of Excellence SimTech, University of Stuttgart, , Stuttgart, , Germany
[c] Faculty for Aerospace Engineering and Geodesy, University of

---

CFDLLMBench: A Benchmark Suite for Evaluating Large Language Models in Computational Fluid Dynamics

Nithin Somasekharan[1] Ling Yue[1] Yadi Cao[2] Weichao Li[1]
Patrick Emami[5] Pochinapeddi Sai Bhargav[3] Anurag Acharya[4]
Xingyu Xie[1] Shaowu Pan[1]*
[1]Rensselaer Polytechnic Institute [2]University of California San Diego
[3]Indian Institute of Science [4]Pacific Northwest National Laboratory
[5]National Renewable Energy Laboratory

Abstract

Large Language Models (LLMs) have demonstrated strong performance across general NLP tasks, but their utility in automating numerical experiments of complex physical system—a critical and labor-intensive component—remains underexplored. As the major workhorse of computational science over the past decades, Computational Fluid Dynamics (CFD) offers a uniquely challenging testbed for evaluating the scientific capabilities of LLMs. We introduce *CFDLLMBench*, a benchmark suite comprising three complementary components—*CFDQuery*, *CFDCodeBench*, and *FoamBench*—designed to holistically evaluate LLM performance across three key competencies: graduate-level CFD knowledge, numerical and physical reasoning of CFD, and context-dependent implementation of CFD

Large language model-empowered next-generation computer-aided engineering

Jiachen Guo[a], Chanwook Park[b,c], Dong Qian[c,d], Thomas J.R. Hughes[e], Wing Kam Liu[b,c]

[a] Theoretical and Applied Mechanics Program, Northwestern University, 2145 Sheridan Road, Evanston, 60201, IL, USA
[b] Department of Mechanical Engineering, Northwestern University, 2145 Sheridan Road, Evanston, 60201, IL, USA
[c] HIDENN-AI, LLC, 1801 Maple Ave, Evanston, 60201, IL, USA
[d] Department of Mechanical Engineering, University of Texas, Dallas, 800 W. Campbell Road, Richardson, 75080, TX, USA
[e] Oden Institute for Computational Engineering and Sciences, University of Texas at Austin, 201 E 24th St, Austin, 78712, TX, USA

Abstract

Software development has entered a new era where large language models (LLMs) now serve as general-purpose reasoning engines, enabling natural language interaction and transformative applications across diverse domains. This paradigm is now extending into computer-aided engineering (CAE), offering a potential solution to the significant human effort and computational expense that constrain traditional finite element methods (FEM), particularly for large-scale parametric problems. Recent applications of LLMs in CAE have successfully automated routine tasks, including CAD model generation and FEM simulations. Nevertheless, these contributions, which primarily serve to reduce manual labor, are often insufficient for addressing the significant computational challenges posed by large-scale, high-dimensional systems. To this aim, we first introduce the concept of LLM-empowered CAE agent, where LLMs act as autonomous collaborators that plan, execute, and adapt CAE workflows. Then, we propose LLM-empowered CAE agent for data-free model order reduction (MOR), a powerful yet underused approach for ultra-fast large-scale parametric analysis due to the intrusive nature and labor-intensive redevelopment of solvers. LLMs can alleviate this barrier by automating derivations, code restructuring, and implementation, making intrusive MOR both practical and broadly accessible. To demonstrate feasibility, we present an LLM-empowered CAE agent for solving ultra-large-scale space-parameter-time (S-P-T) physical problems using Tensor-decomposition-based A Priori Surrogates (TAPS). Our results show that natural language prompts describing parametric partial differential equations (PDEs) can be translated into efficient solver implementations, substantially reducing human effort while producing

---

AI Agents in Engineering Design: A Multi-Agent Framework for Aesthetic and

Mohamed Elrefaie*
Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Janet Qian
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Qian Chen
Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Angela Dai
Department of Computer Science, Technical University of Munich, Munich, Germany

ABSTRACT

We introduce the concept of "Design Agents" for engineering applications, particularly focusing on the automotive design process, while emphasizing that our approach can be readily extended to other engineering and design domains. Our framework integrates AI-driven design agents into the traditional engineering workflow, demonstrating how these specialized computational agents interact seamlessly with engineers and designers to augment creativity, enhance efficiency, and significantly accelerate the overall design cycle. By automating and streamlining tasks traditionally performed manually, such as conceptual sketching, styling enhancements, 3D shape retrieval and generative modeling, computational fluid dynamics (CFD) meshing, and aero-

Keywords: AI Agents, Generativ

1. INTRODUCTION

The design of a car is a m
ances engineering performanc
aircraft design, where function
solute priority, automotive des
and brand identity, making it
process [1–9]. Consumers are
technical specifications but also
ics play a crucial role in marke
design involves a complex inter
ufacturability, and subjective u
workflows rely on iterative re

Agentic AI for Discovery

Published as a conference paper at ICLR 2025

AGENTIC AI FOR SCIENTIFIC DISCOVERY: A SURVEY OF PROGRESS, CHALLENGES, AND FUTURE DIRECTIONS

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes & Christina Mack
IQVIA
{firstname.lastname}@iqvia.com

ABSTRACT

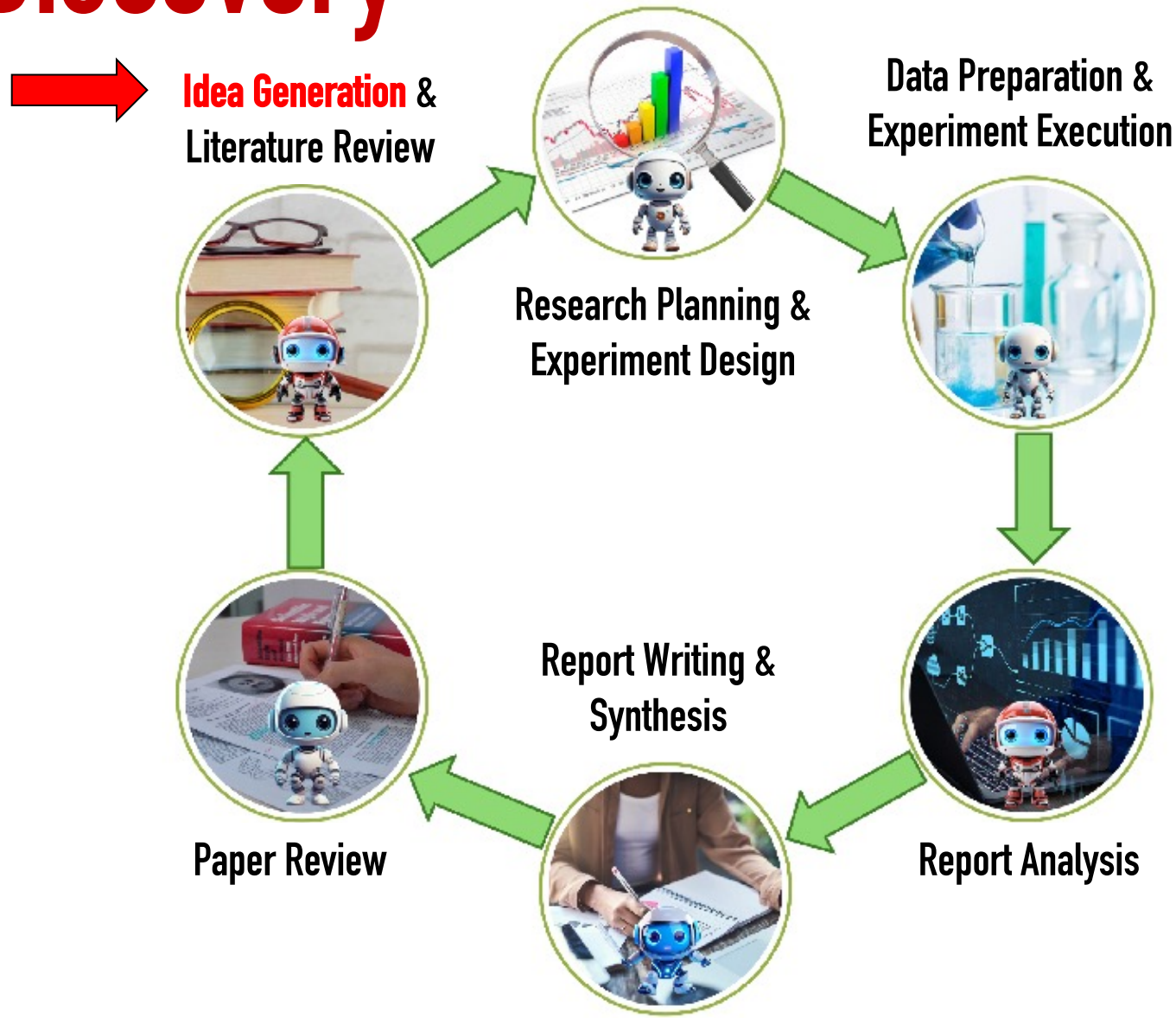
The integration of Agentic AI into scientific discovery marks a new frontier in research automation. These AI systems, capable of reasoning, planning, and autonomous decision-making, are transforming how scientists perform literature review, generate hypotheses, conduct experiments, and analyze results. This survey provides a comprehensive overview of Agentic AI for scientific discovery, categorizing existing systems and tools, and highlighting recent progress across fields such as chemistry, biology, and materials science. We discuss key evaluation metrics, implementation frameworks, and commonly used datasets to offer a detailed understanding of the current state of the field. Finally, we address critical challenges, such as literature review automation, system reliability, and ethical concerns, while outlining future research directions that emphasize human-AI collaboration and enhanced system calibration.

1 INTRODUCTION

The rapid advancements of Large Language Models (LLMs) (Touvron et al., 2023; Anil et al., 2023; Achiam et al., 2023) have opened a new era in scientific discovery, with Agentic AI systems (Kim et al., 2024; Guo et al., 2023; Wang et al., 2024; Abramovich et al., 2024) emerging as powerful tools for automating complex research workflows. Unlike traditional AI, Agentic AI systems are designed to operate with a high degree of autonomy, allowing them to independently perform tasks such as hypothesis generation, literature review, experimental design, and data analysis. These systems have the potential to significantly accelerate scientific research, reduce costs, and expand access to advanced tools across various fields, including chemistry, biology, and materials science.

Recent efforts have demonstrated the potential of LLM-driven agents in supporting researchers with tasks such as literature reviews, experimentation, and report writing. Prominent frameworks, including LitSearch (Ajith et al., 2024), ResearchArena (Kang & Xiong, 2024), SciLitLLM (Li et al., 2024c), CiteME (Press et al., 2024), ResearchAgent (Baek et al., 2024) and Agent Laboratory (Schmidgall et al., 2025), have made strides in automating general research workflows, such as citation management, document discovery, and academic survey generation. However, these systems often lack the domain-specific focus and compliance-driven rigor essential for fields like biomedical domain, where the structured assessment of literature is critical for evidence synthesis. For example, Agent Laboratory demonstrated high success rates in data preparation, experimentation, and report writing. However, its performance dropped significantly in the literature review phase, reflecting the inherent challenges of automating structured literature reviews. Moreover, questions about system reliability, reproducibility, and ethical governance continue to pose significant hurdles.

This survey aims to provide a comprehensive review of Agentic AI for scientific discovery. We categorize existing systems into autonomous and collaborative frameworks, detailing the datasets, implementation tools, and evaluation metrics that support these innovations. By highlighting the current state of the field and discussing open challenges, we hope to inspire further research and development in Agentic AI, ultimately encouraging more reliable and impactful scientific contributions.



Agentic AI for Discovery

AGENTIC AI FOR SCIENTIFIC DISCOVERY: A SURVEY OF PROGRESS, CHALLENGES, AND FUTURE DIRECTIONS

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes & Christina Mack
IQVIA
{firstname.lastname}@iqvia.com

ABSTRACT

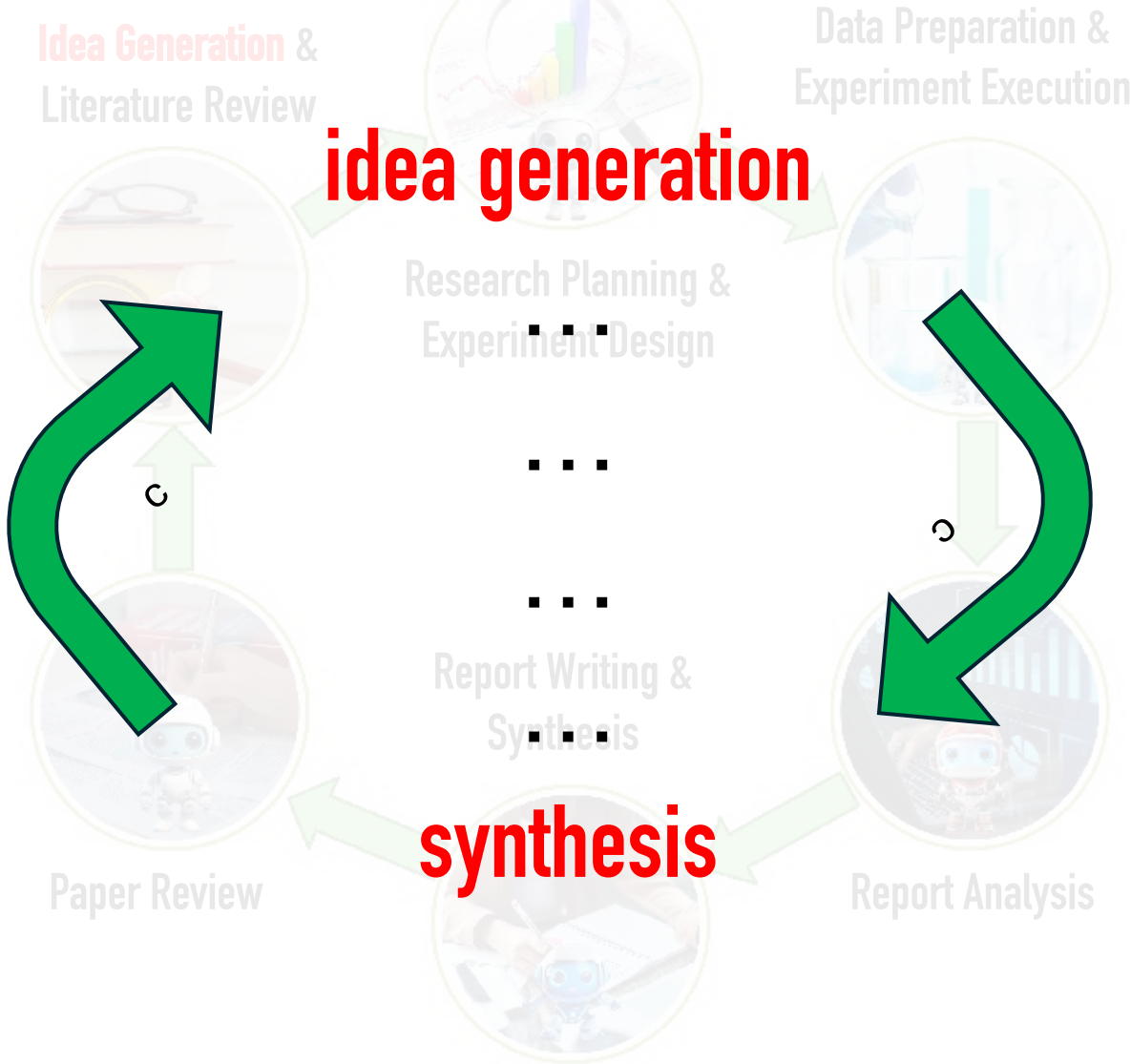
The integration of Agentic AI into scientific discovery marks a new frontier in research automation. These AI systems, capable of reasoning, planning, and autonomous decision-making, are transforming how scientists perform literature review, generate hypotheses, conduct experiments, and analyze results. This survey provides a comprehensive overview of Agentic AI for scientific discovery, categorizing existing systems and tools, and highlighting recent progress across fields such as chemistry, biology, and materials science. We discuss key evaluation metrics, implementation frameworks, and commonly used datasets to offer a detailed understanding of the current state of the field. Finally, we address critical challenges, such as literature review automation, system reliability, and ethical concerns, while outlining future research directions that emphasize human-AI collaboration and enhanced system calibration.

1 INTRODUCTION

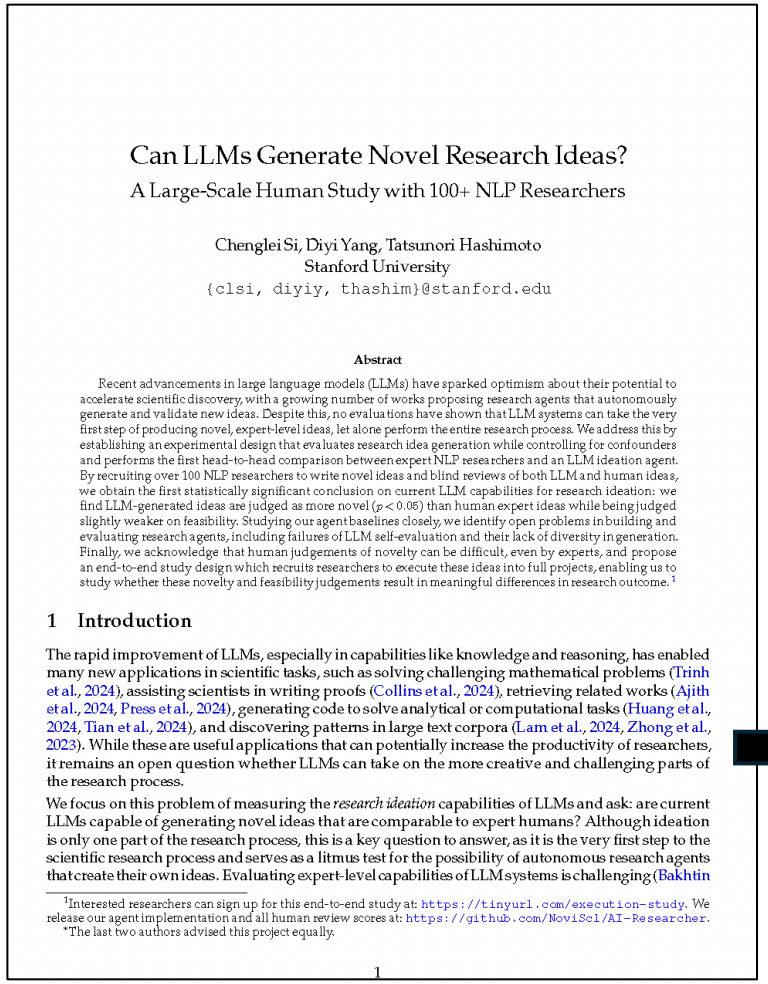
The rapid advancements of Large Language Models (LLMs) (Touvron et al., 2023; Anil et al., 2023; Achiam et al., 2023) have opened a new era in scientific discovery, with Agentic AI systems (Kim et al., 2024; Guo et al., 2023; Wang et al., 2024; Abramovich et al., 2024) emerging as powerful tools for automating complex research workflows. Unlike traditional AI, Agentic AI systems are designed to operate with a high degree of autonomy, allowing them to independently perform tasks such as hypothesis generation, literature review, experimental design, and data analysis. These systems have the potential to significantly accelerate scientific research, reduce costs, and expand access to advanced tools across various fields, including chemistry, biology, and materials science.

Recent efforts have demonstrated the potential of LLM-driven agents in supporting researchers with tasks such as literature reviews, experimentation, and report writing. Prominent frameworks, including LitSearch (Ajith et al., 2024), ResearchArena (Kang & Xiong, 2024), SciLitLLM (Li et al., 2024c), CiteME (Press et al., 2024), ResearchAgent (Baek et al., 2024) and Agent Laboratory (Schmidgall et al., 2025), have made strides in automating general research workflows, such as citation management, document discovery, and academic survey generation. However, these systems often lack the domain-specific focus and compliance-driven rigor essential for fields like biomedical domain, where the structured assessment of literature is critical for evidence synthesis. For example, Agent Laboratory demonstrated high success rates in data preparation, experimentation, and report writing. However, its performance dropped significantly in the literature review phase, reflecting the inherent challenges of automating structured literature reviews. Moreover, questions about system reliability, reproducibility, and ethical governance continue to pose significant hurdles.

This survey aims to provide a comprehensive review of Agentic AI for scientific discovery. We categorize existing systems into autonomous and collaborative frameworks, detailing the datasets, implementation tools, and evaluation metrics that support these innovations. By highlighting the current state of the field and discussing open challenges, we hope to inspire further research and development in Agentic AI, ultimately encouraging more reliable and impactful scientific contributions.



Idea Generation



Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

Abstract

Recent advancements in large language models (LLMs) have sparked optimism about their potential to accelerate scientific discovery, with a growing number of works proposing research agents that autonomously generate and validate new ideas. Despite this, no evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process. We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first head-to-head comparison between expert NLP researchers and an LLM ideation agent. By recruiting over 100 NLP researchers to write novel ideas and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on current LLM capabilities for research ideation: we find LLM-generated ideas are judged as more novel ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility. Studying our agent baselines closely, we identify open problems in building and evaluating research agents, including failures of LLM self-evaluation and their lack of diversity in generation. Finally, we acknowledge that human judgements of novelty can be difficult, even by experts, and propose an end-to-end study design which recruits researchers to execute these ideas into full projects, enabling us to study whether these novelty and feasibility judgements result in meaningful differences in research outcome. [1]

1 Introduction

The rapid improvement of LLMs, especially in capabilities like knowledge and reasoning, has enabled many new applications in scientific tasks, such as solving challenging mathematical problems (Trinh et al., 2024), assisting scientists in writing proofs (Collins et al., 2024), retrieving related works (Ajith et al., 2024, Press et al., 2024), generating code to solve analytical or computational tasks (Huang et al., 2024, Tian et al., 2024), and discovering patterns in large text corpora (Lam et al., 2024, Zhong et al., 2023). While these are useful applications that can potentially increase the productivity of researchers, it remains an open question whether LLMs can take on the more creative and challenging parts of the research process.

We focus on this problem of measuring the *research ideation* capabilities of LLMs and ask: are current LLMs capable of generating novel ideas that are comparable to expert humans? Although ideation is only one part of the research process, this is a key question to answer, as it is the very first step to the scientific research process and serves as a litmus test for the possibility of autonomous research agents that create their own ideas. Evaluating expert-level capabilities of LLM systems is challenging (Bakhtin

[1]Interested researchers can sign up for this end-to-end study at: https://tinyurl.com/execution-study. We release our agent implementation and all human review scores at: https://github.com/NoviSci/AI-Researcher.

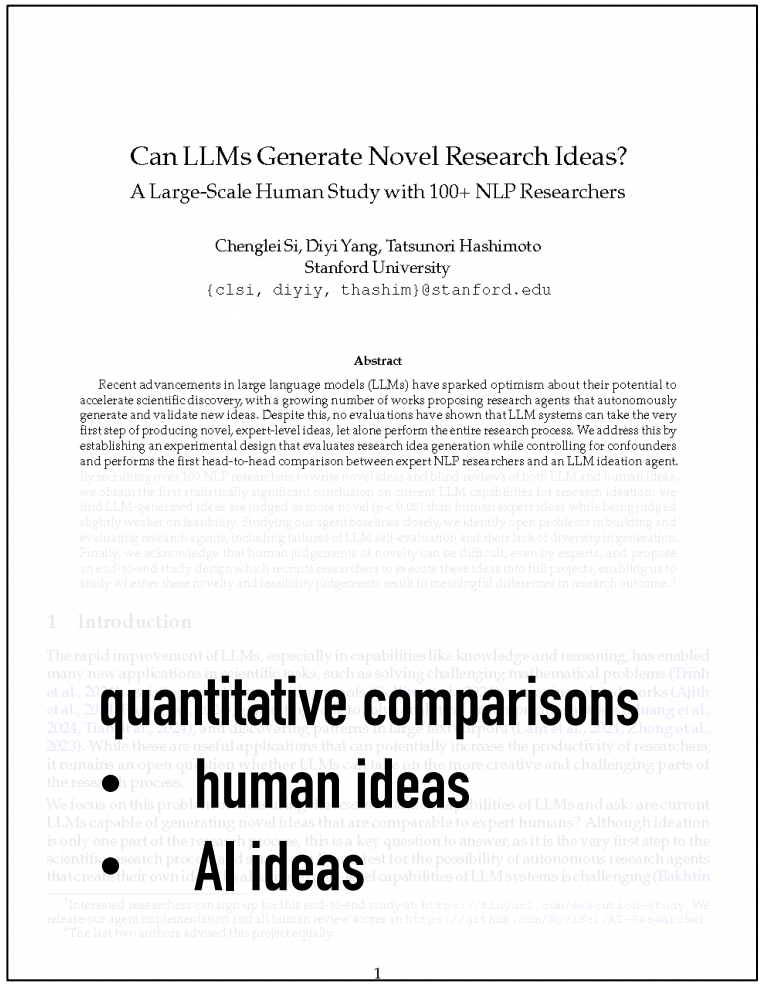
*The last two authors advised this project equally.

1

quantifying idea generation

- literature retrieval
- proposal formulation (idea)
- hypothesis generation
- ranking and downselect
- refine and combine

Idea Generation



Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

Abstract

Recent advancements in large language models (LLMs) have sparked optimism about their potential to accelerate scientific discovery, with a growing number of works proposing research agents that autonomously generate and validate new ideas. Despite this, no evaluations have shown that LLM systems can take the very first step of producing novel, expert-level ideas, let alone perform the entire research process. We address this by establishing an experimental design that evaluates research idea generation while controlling for confounders and performs the first head-to-head comparison between expert NLP researchers and an LLM ideation agent. By recruiting over 100 NLP researchers to write novel ideas and blind reviews of both LLM and human ideas, we obtain the first statistically significant conclusion on current LLM capabilities for research ideation: we find LLM-generated ideas are judged as more novel ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility. Studying our agent baselines closely, we identify open problems in building and evaluating research agents, including failures of LLM self-evaluation and their lack of diversity in generation. Finally, we acknowledge that human judgements of novelty can be difficult, even by experts, and propose an end-to-end study design which recruits researchers to execute these ideas into full projects, enabling us to study whether these novelty and feasibility judgements result in meaningful differences in research outcome.[1]

1 Introduction

The rapid improvement of LLMs, especially in capabilities like knowledge and reasoning, has enabled many new applications in scientific fields, such as solving challenging mathematical problems (Trinh et al., 2024; Yang et al., 2023). While these are useful applications that can potentially increase the productivity of researchers, it remains an open question whether LLMs can help with the more creative and challenging parts of the research process.

We focus on this problem of evaluating the research capabilities of LLMs and ask: are current LLMs capable of generating novel ideas that are comparable to expert humans? Although ideation is only one part of the research process, this is a key question to answer, as it is the very first step to the scientific research process. Moreover, it is a necessary test for the possibility of autonomous research agents that create their own ideas. Evaluating the research capabilities of LLM systems is challenging (Bakhtin

[1]Interested researchers can sign up for this end-to-end study at: https://tinyurl.com/execution-study. We release our agent implementation and all human review scores at: https://github.com/NoviScl/AI-Researcher.

[*]The last two authors advised this project equally.

1

- quantitative comparisons
- human ideas
- AI ideas

some conclusions (redacted)

- AI lacks diversity in idea generation
- AI ideas are not well-motivated and vague on feasibility details
- AI experts cannot evaluate/rank ideas reliably
  - missing or inappropriate baselines
  - making unrealistic assumptions
  - not adequately following existing best practices

- aren't these all traits of "early career" researchers?

Synthesis

Published as a conference paper at ICLR 2025

AGENTIC AI FOR SCIENTIFIC DISCOVERY: A SURVEY OF PROGRESS, CHALLENGES, AND FUTURE DIRECTIONS

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes & Christina Mack
IQVIA
{firstname.lastname}@iqvia.com

ABSTRACT

The integration of Agentic AI into scientific discovery marks a new frontier in research automation. These AI systems, capable of reasoning, planning, and autonomous decision-making, are transforming how scientists perform literature review, generate hypotheses, conduct experiments, and analyze results. This survey provides a comprehensive overview of Agentic AI for scientific discovery, categorizing existing systems and tools, and highlighting recent progress across fields such as chemistry, biology, and materials science. We discuss key evaluation metrics, implementation frameworks, and commonly used datasets to offer a detailed understanding of the current state of the field. Finally, we address critical challenges, such as literature review automation, system reliability, and ethical concerns, while outlining future research directions that emphasize human-AI collaboration and enhanced system calibration.

1 INTRODUCTION

The rapid advancements of Large Language Models (LLMs) (Touvron et al., 2023; Anil et al., 2023; Achiam et al., 2023) have opened a new era in scientific discovery, with Agentic AI systems (Kim et al., 2024; Guo et al., 2023; Wang et al., 2024; Abramovich et al., 2024) emerging as powerful tools for automating complex research workflows. Unlike traditional AI, Agentic AI systems are designed to operate with a high degree of autonomy, allowing them to independently perform tasks such as hypothesis generation, literature review, experimental design, and data analysis. These systems have the potential to significantly accelerate scientific research, reduce costs, and expand access to advanced tools across various fields, including chemistry, biology, and materials science.

Recent efforts have demonstrated the potential of LLM-driven agents in supporting researchers with tasks such as literature reviews, experimentation, and report writing. Prominent frameworks, including LitSearch (Ajith et al., 2024), ResearchArena (Kang & Xiong, 2024), SciLitLLM (Li et al., 2024c), CiteME (Press et al., 2024), ResearchAgent (Baek et al., 2024) and Agent Laboratory (Schmidgall et al., 2025), have made strides in automating general research workflows, such as citation management, document discovery, and academic survey generation. However, these systems often lack the domain-specific focus and compliance-driven rigor essential for fields like biomedical domain, where the structured assessment of literature is critical for evidence synthesis. For example, Agent Laboratory demonstrated high success rates in data preparation, experimentation, and report writing. However, its performance dropped significantly in the literature review phase, reflecting the inherent challenges of automating structured literature reviews. Moreover, questions about system reliability, reproducibility, and ethical governance continue to pose significant hurdles.

This survey aims to provide a comprehensive review of Agentic AI for scientific discovery. We categorize existing systems into autonomous and collaborative frameworks, detailing the datasets, implementation tools, and evaluation metrics that support these innovations. By highlighting the current state of the field and discussing open challenges, we hope to inspire further research and development in Agentic AI, ultimately encouraging more reliable and impactful scientific contributions.

1

quantify results synthesis

- comprehensive analysis of extensive amount of data
- identify and interpret patterns
- filter novel contributions
- extract meaningful insights
- draw conclusions

Synthesis

Published as a conference paper at ICLR 2025

AGENTIC AI FOR SCIENTIFIC DISCOVERY: A SURVEY OF PROGRESS, CHALLENGES, AND FUTURE DIRECTIONS

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes & Christina Mack
IQVIA
{firstname.lastname}@iqvia.com

ABSTRACT

The integration of Agentic AI into scientific discovery marks a new frontier in research automation. These AI systems, capable of reasoning, planning, and autonomous decision-making, are transforming how scientists perform literature review, generate hypotheses, conduct experiments, and analyze results. This survey provides a comprehensive overview of Agentic AI for scientific discovery, categorizing existing systems and tools, and highlighting recent progress across fields such as chemistry, biology, and materials science. We discuss key evaluation metrics, implementation frameworks, and commonly used datasets to offer a detailed understanding of the current state of the field. Finally, we address critical challenges, such as literature review automation, system reliability, and ethical concerns, while outlining future research directions that emphasize human-AI collaboration and enhanced system calibration.

1 INTRODUCTION

The rapid advancements of Large Language Models (LLMs) (Touvron et al., 2023; Anil et al., 2023; Achiam et al., 2023) have opened a new era in scientific discovery, with Agentic AI systems (Kim et al., 2024; Guo et al., 2023; Wang et al., 2024; Abramovich et al., 2024) emerging as powerful tools for automating complex research workflows. Unlike traditional AI, Agentic AI systems are designed to operate with a high degree of autonomy, allowing them to independently perform tasks such as hypothesis generation, literature review, experimental design, and data analysis. These systems have the potential to significantly accelerate scientific research, reduce costs, and expand access to advanced tools across various fields, including chemistry, biology, and materials science.

Recent efforts have demonstrated the potential of LLM-driven agents in supporting researchers with tasks such as literature reviews, experimentation, and report writing. Prominent frameworks, including LitSearch (Ajith et al., 2024), ResearchArena (Kang & Xiong, 2024), SciLitLLM (Li et al., 2024c), CiteME (Press et al., 2024), ResearchAgent (Baek et al., 2024) and Agent Laboratory (Schmidgall et al., 2025), have made strides in automating general research workflows, such as citation management, document discovery, and academic survey generation. However, these systems often lack the domain-specific focus and compliance-driven rigor essential for fields like biomedical domain, where the structured assessment of literature is critical for evidence synthesis. For example, Agent Laboratory demonstrated high success rates in data preparation, experimentation, and report writing. However, its performance dropped significantly in the literature review phase, reflecting the inherent challenges of automating structured literature reviews. Moreover, questions about system reliability, reproducibility, and ethical governance continue to pose significant hurdles.

This survey aims to provide a comprehensive review of Agentic AI for scientific discovery. We categorize existing systems into autonomous and collaborative frameworks, detailing the datasets, implementation tools, and evaluation metrics that support these innovations. By highlighting the current state of the field and discussing open challenges, we hope to inspire further research and development in Agentic AI, ultimately encouraging more reliable and impactful scientific contributions.

1

some conclusions (redacted)

AI agents excel at breadth but struggle with depth

- parameter sweeps
- literature harvesting
- principled data reductions
- figure generation

- framing sharper hypotheses
- flagging result that are too context-specific
- admitting when the evidence is insufficient
- quantifying uncertainty

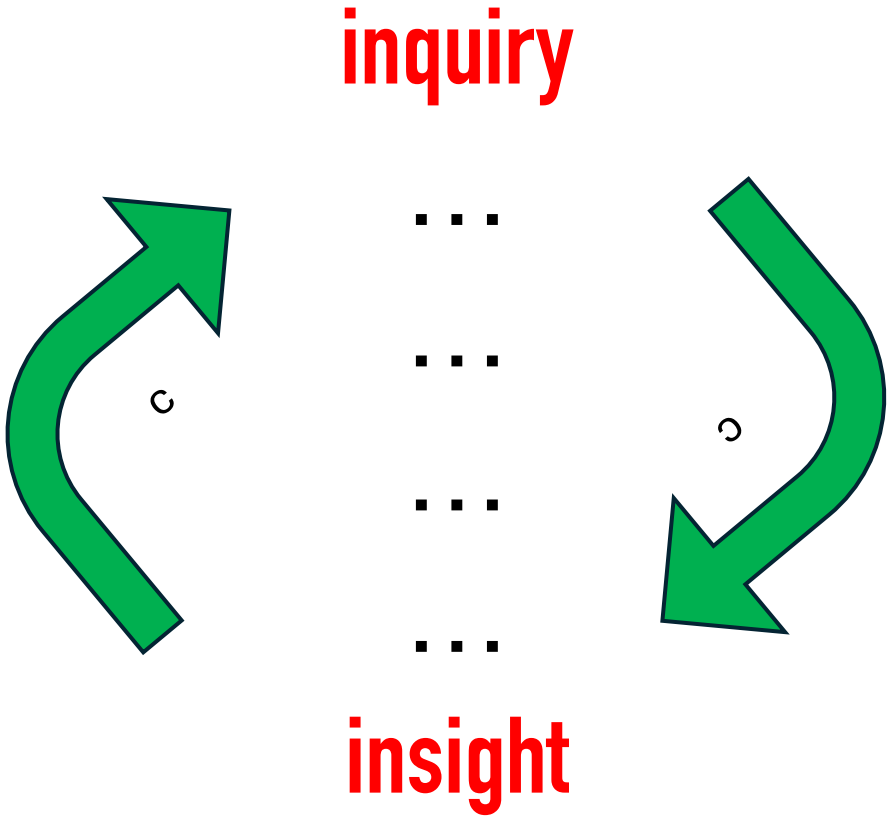
muscle vs. brain: isn't this typical of "early career" researchers?

Agentic AI for Discovery

- idea-generation ⟹ inquiry
- synthesis ⟹ insight

inquiry + insight = intelligence

The Scientific Process



by deduction:

from general to particular

from theory to data

by induction:

from examples to general

from data to theory

Outline

from well defined to open-ended inquiries

- forecast
- design
- discovery

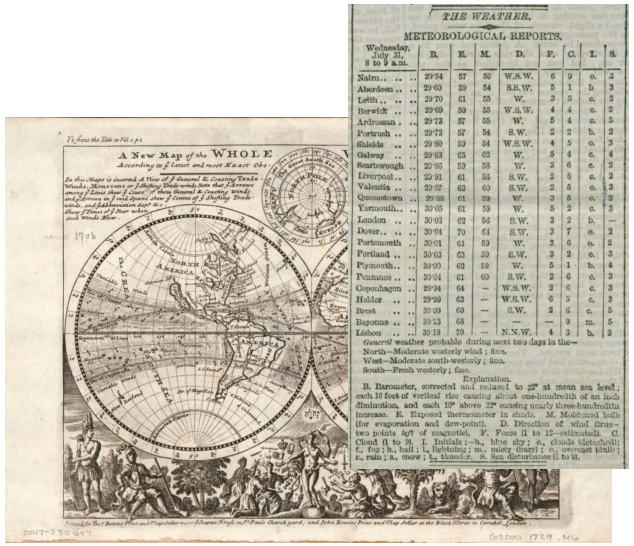
Outline

from well defined to open-ended inquiries

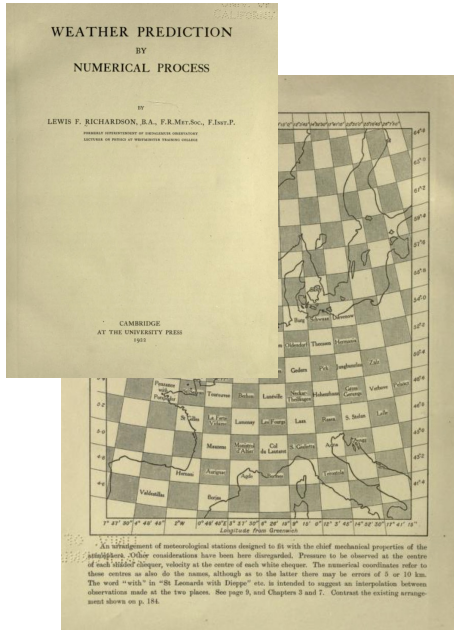
- forecast
- design
- discovery

Weather Forecast

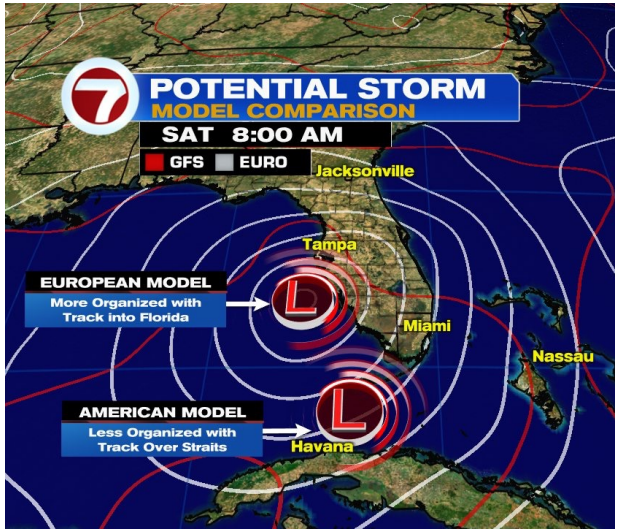
inquiry: can we reduce sailing times with accurate wind charts?



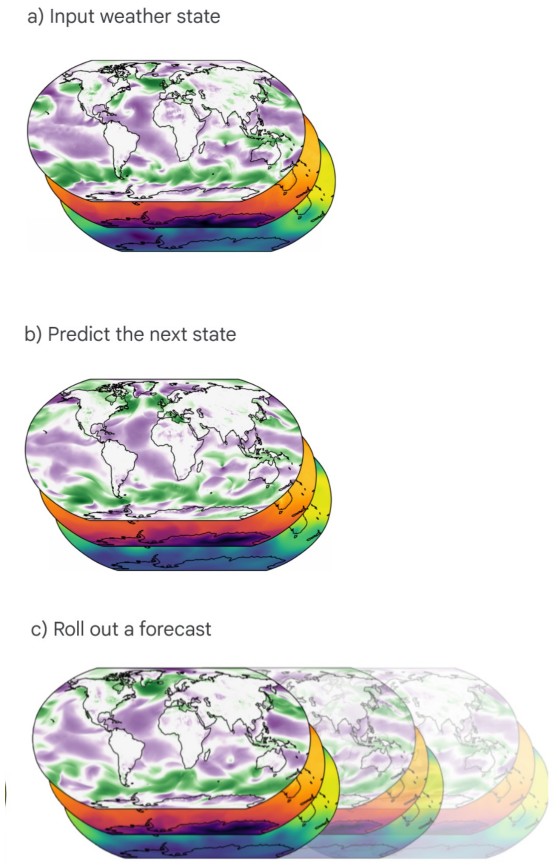
induction: observations and map matching (1854)



deduction: modeling based on physics of the atmosphere (1922)



combined: numerical weather models, HPC and satellite data assimilation (~2000)



induction: graphcast (2022)

from inquiry to insight: Hadley Circulation, Coriolis Effect, Rossby Waves, [. . .]

Forecast & Computers + AI

Google GraphCast (& now WeatherNext)

RESEARCH

WEATHER FORECASTING

Learning skillful medium-range global weather forecasting

Remi Lam[1*†], Alvaro Sanchez-Gonzalez[1*†], Matthew Willson[1*†], Peter Wirnsberger[2†], Meire Fortunato[3†], Ferran Alet[3†], Suman Ravuri[3†], Timo Ewalds[3], Zach Eaton-Rosen[3], Weihua Hu[3], Alexander Merose[2], Stephan Hoyer[2], George Holland[4], Oriol Vinyals[1], Jacklynn Stott[1], Alexander Pritzel[3], Shakir Mohamed[3*], Peter Battaglia[1*]

Global medium-range weather forecasting is critical to decision-making across many social and economic domains. Traditional numerical weather prediction uses increased compute resources to improve forecast accuracy but does not directly use historical weather data to improve the underlying model. Here, we introduce GraphCast, a machine learning–based method trained directly from reanalysis data. It predicts hundreds of weather variables for the next 10 days at 0.25° resolution globally in under 1 minute. GraphCast significantly outperforms the most accurate operational deterministic systems on 90% of 1380 verification targets, and its forecasts support better severe event prediction, including tropical cyclone tracking, atmospheric rivers, and extreme temperatures. GraphCast is a key advance in accurate and efficient weather forecasting and helps realize the promise of machine learning for modeling complex dynamical systems.

It is 05:45 UTC (coordinated universal time) in mid-October 2022 in Bologna, Italy, at the recently opened high-performance computing facility of the European Centre for Medium-Range Weather Forecasts (ECMWF). For the past several hours, the Integrated Forecasting System (IFS) has been running sophisticated calculations to forecast Earth's weather over the next days and weeks, and its first predictions have just begun to be disseminated to users. This process repeats every 6 hours, every day, to supply the world with the most accurate weather forecasts available.

The IFS, and modern weather forecasting more generally, are triumphs of science and engineering. The dynamics of weather systems are among the most complex physical phenomena on Earth, and each day, countless decisions made by individuals, industries, and policy-makers depend on accurate weather forecasts, from deciding whether to wear a jacket to deciding whether to flee a dangerous storm. The dominant approach for weather forecasting today is numerical weather prediction (NWP), which involves solving the governing equations of weather using supercomputers. The success of NWP lies in the rigorous and ongoing research practices that provide increasingly detailed descriptions of weather phenomena and in how well NWP scales to greater accuracy with greater computational resources (*1*, *2*). As a result, the accuracy of weather forecasts has increased year after year, to the point where the path of a hurricane can be predicted many days ahead—a possibility that was unthinkable even a few decades ago.

But while traditional NWP scales well with compute, capitalizing on the vast amount of historical weather data to improve accuracy is not straightforward. Rather, NWP methods are improved by highly trained experts innovating better models, algorithms, and approximations, which can be a time-consuming and costly process.

Machine learning–based weather prediction (MLWP)—wherein forecast models are trained from historical data, including observations and analysis data—offers an alternative to traditional NWP. MLWP has the potential to improve forecast accuracy by capturing patterns in the data that are not easily represented in explicit equations. MLWP also offers opportunities for greater efficiency by exploiting modern deep learning hardware, rather than supercomputers, and striking more favorable speed-accuracy trade-offs. Recently, MLWP has helped improve on NWP-based forecasting in regimes where traditional NWP is relatively weak, for example, in subseasonal heat wave prediction (*3*) and precipitation nowcasting from radar images (*4*–*7*), where accurate equations and robust numerical methods are not as available.

In medium-range weather forecasting—the prediction of atmospheric variables up to 10 days ahead—NWP-based systems such as the IFS are still most accurate. The top deterministic operational system in the world is ECMWF's high-resolution forecast (HRES), a configuration of IFS that produces global 10-day forecasts at 0.1° latitude and longitude resolution, in around an hour (*8*). However, over the past several years, MLWP methods for medium-range forecasting trained on reanalysis data have been steadily advancing, facilitated by benchmarks such as WeatherBench (*9*). Deep learning architectures based on convolutional neural networks (*9*–*11*) and Transformers (*12*) have shown promising results at latitude and longitude resolutions coarser than 1.0°, and recent works—which use graph neural networks (GNNs), Fourier neural operators, and Transformers (*13*–*16*)—have reported performance that begins to rival IFS's at 1.0° and 0.25° for a handful of variables and lead times up to 7 days.

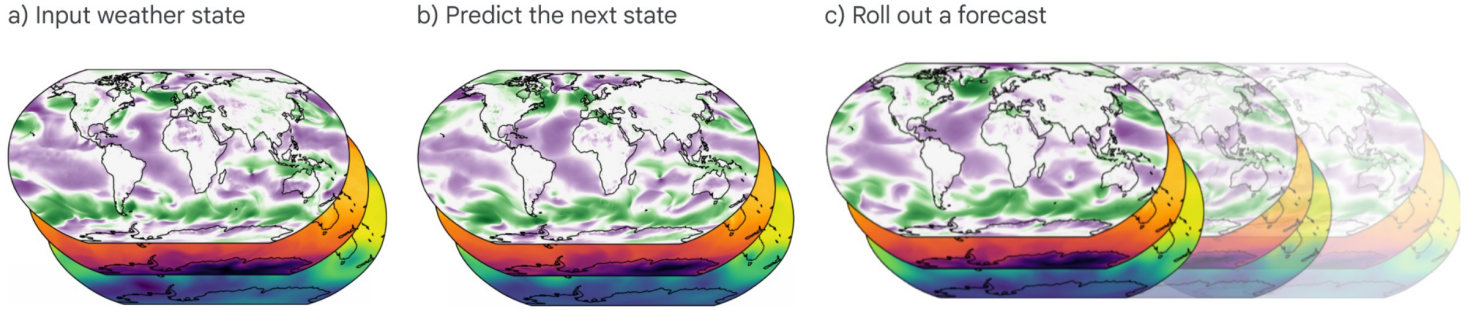
GraphCast

Here, we introduce an MLWP approach for global medium-range weather forecasting called GraphCast, which produces an accurate 10-day forecast in under a minute on a single Google Cloud TPU (Tensor Processing Unit) v4 device and supports applications including predicting tropical cyclone tracks, atmospheric rivers, and extreme temperatures.

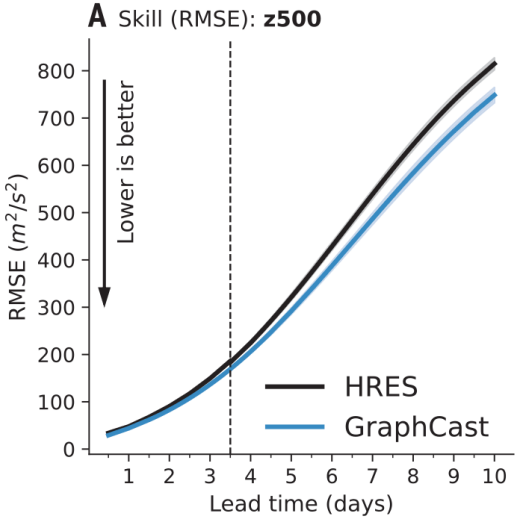
GraphCast takes as input the two most recent states of Earth's weather—the current time and 6 hours earlier—and predicts the next state of the weather 6 hours ahead. A single weather state is represented by a 0.25° latitude-longitude grid (721 by 1440), which corresponds to roughly 28 km by 28 km resolution at the equator (Fig. 1A), where each grid point represents a set of surface and atmospheric variables (listed in Table 1). Like traditional NWP systems, GraphCast

Table 1. Weather variables and levels modeled by GraphCast. The numbers in parentheses in the column headings are the number of entries in the column. Boldfaced variables and levels indicate those that were included in the scorecard evaluation. All atmospheric variables are represented at each of the pressure levels.

Surface variables (5)	Atmospheric variables (6)	Pressure levels (37)



- purely inductive, data driven
- not based on physics principles or constrained by numerical considerations
- preserves invariance & symmetries
- use simulations as training data
- more accurate than high-resolution models



Forecast

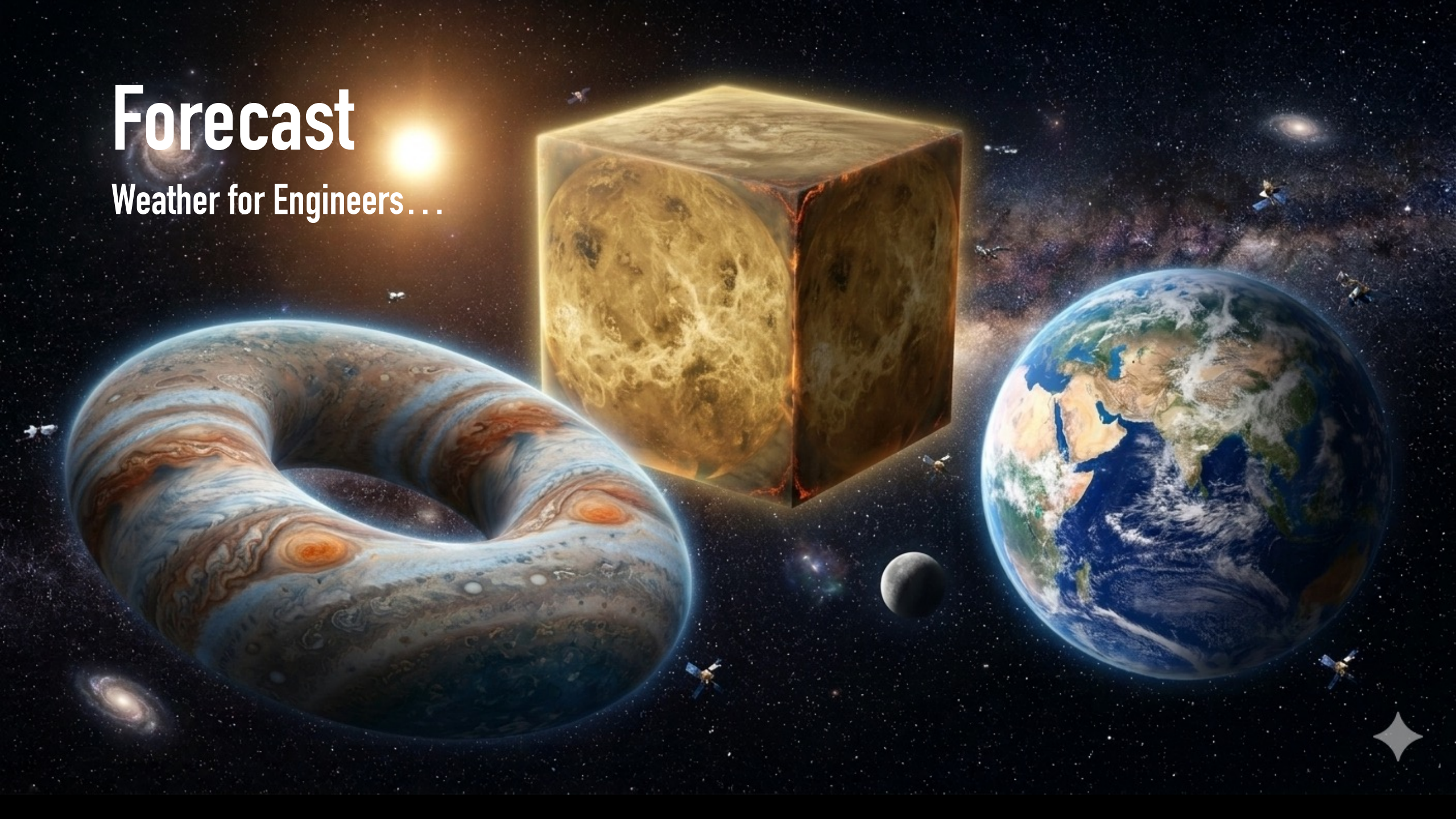
Weather for Engineers?

Forecast

Weather for Engineers . . .

Forecast

Weather for Engineers . . .

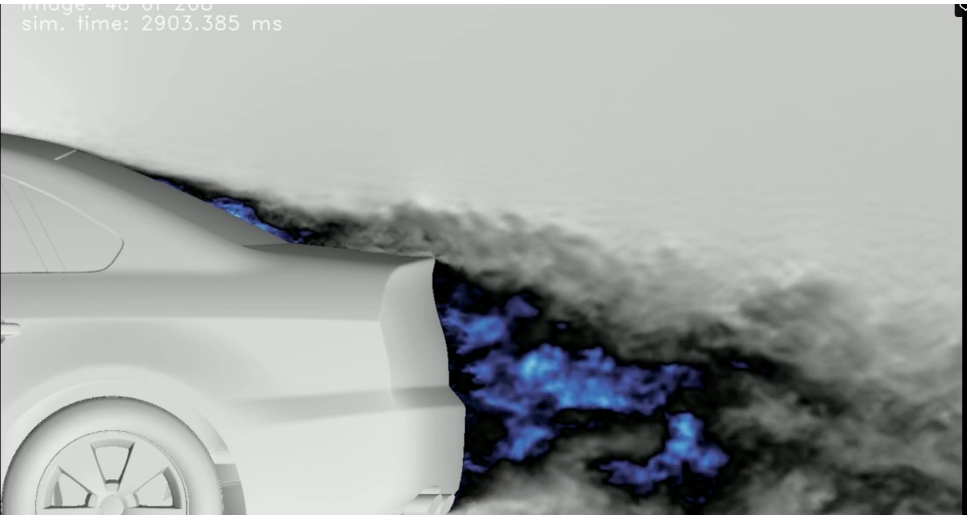


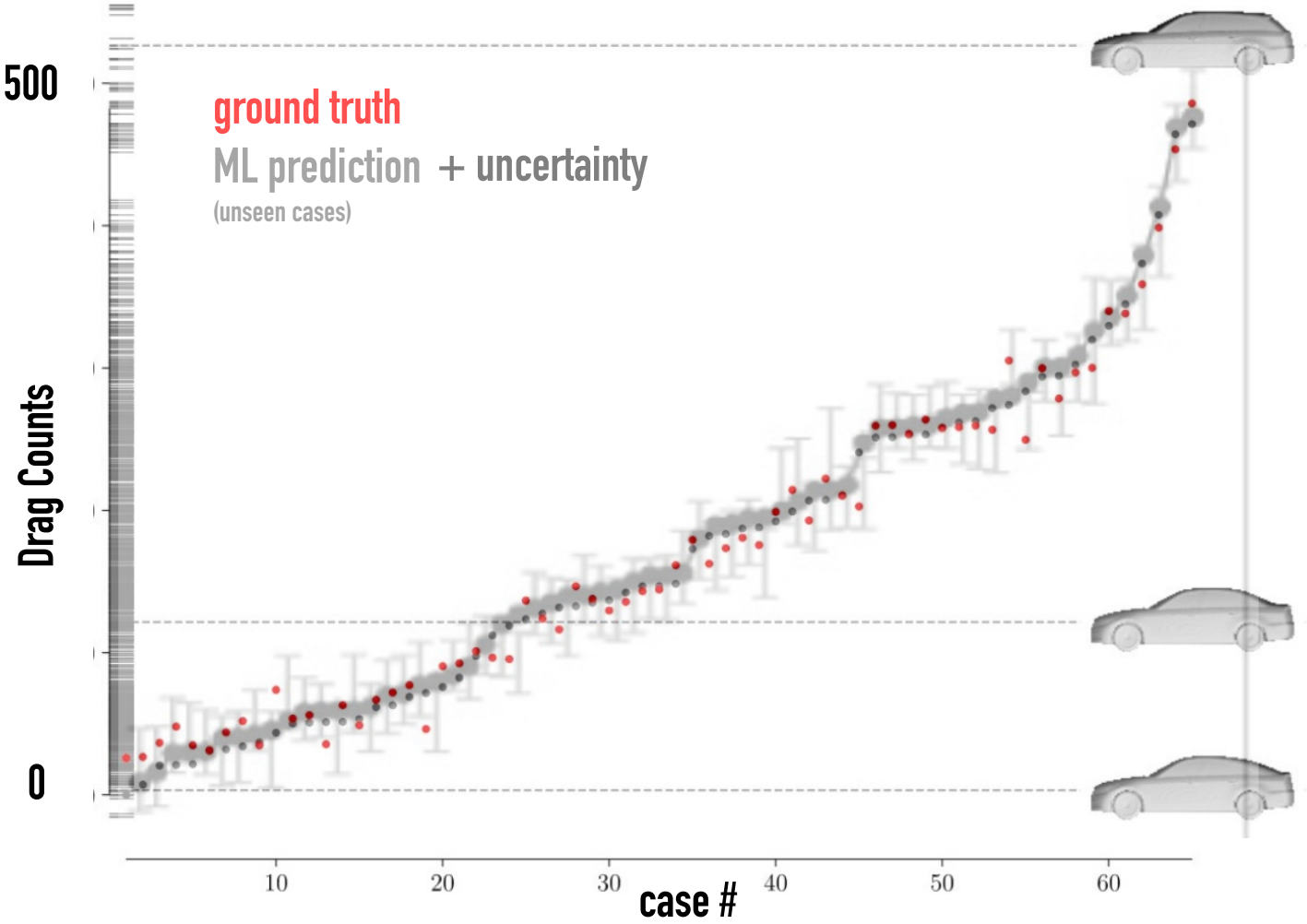
Forecast

inquiry: prediction of aerodynamic drag for road vehicle
insight: drag crisis, c-pillar vortices

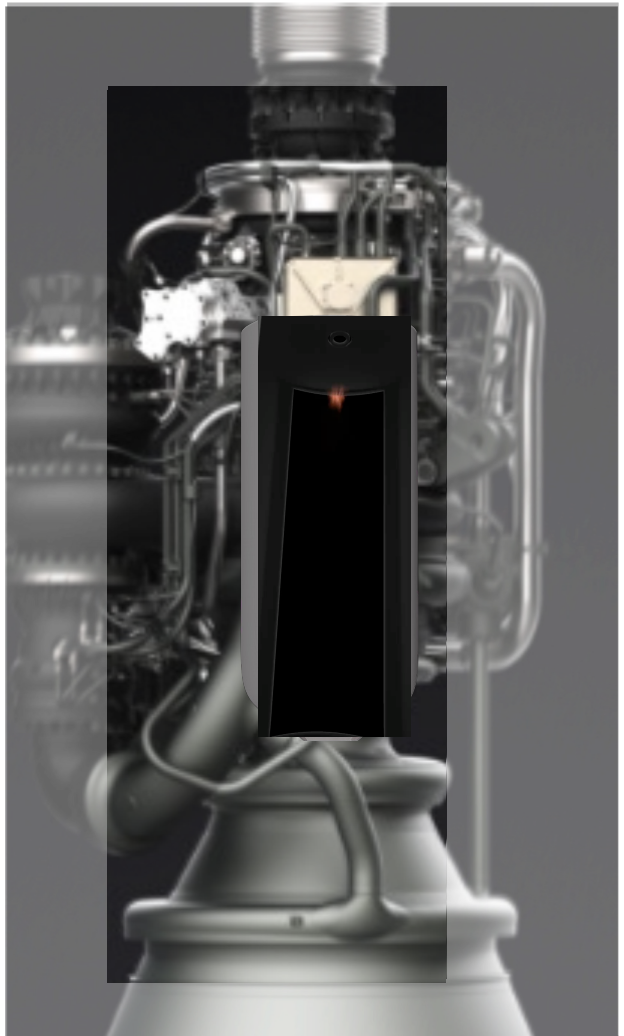
ML Predictions with quantified uncertainties and high-quality datasets

M Benjamin, G Iaccarino, A systematic dataset generation technique applied to data-driven automotive aerodynamics, APL Machine Learning, Vol 3 (1), 2025



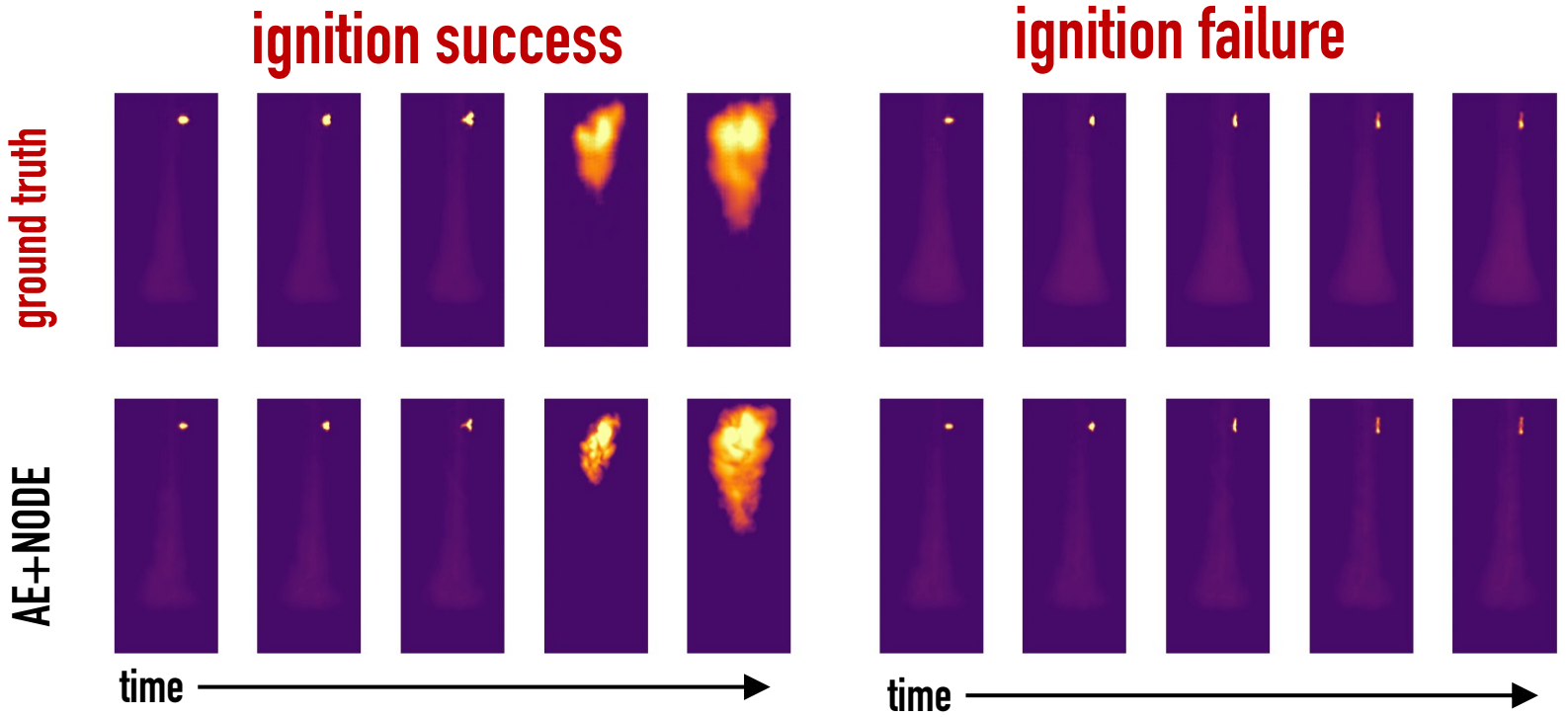


Forecast



inquiry: reliability map of laser-induced ignition in high-speed jet

insight: spark “propulsion”, flame stabilization



ML predictions based on AutoEncoder+NODE trained on high-resolution simulations

Zahtila et al, Generative Prediction of Laser-Induced Rocket Ignition with Dynamic Latent Space Representations, 2025

Outline

from well defined to open-ended inquiries

- forecast → physics + simulations + ML (AI)
  - well defined inquiry, extensive data, supercomputers, and time to fail and develop physics understanding (insights)
- design
- discovery

Outline

from well defined to open-ended inquiries

- forecast
- **design**
- discovery

Aerodynamic Design

inquiry: shape with low-drag, high-lift, etc.

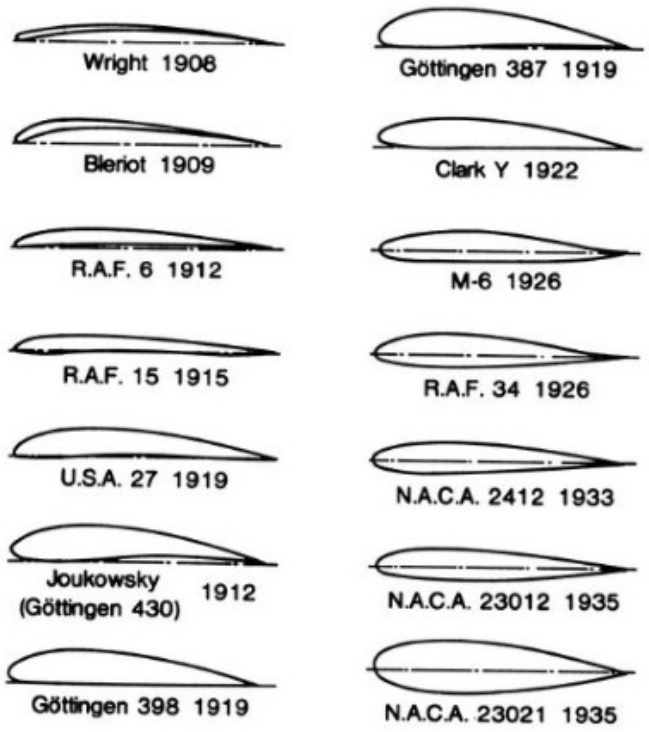


Image Credit: NACA

- experiments
- theory
- computational Tools (CFD)

VOL. 17, NO. 12, DECEMBER 1979 AIAA JOURNAL 1293
ARTICLE NO. 79-0129R

Dryden Lectureship in Research

Computational Aerodynamics Development and Outlook

Dean R. Chapman
Ames Research Center, NASA, Moffett Field, Calif.

Introduction

IT is an honor and challenge to present the Dryden Lecture in Research for 1979. Since my topic concerns a new trend in fluid mechanics, it should not be surprising that some aspects of this paper involve basic mechanics of turbulence, a field enriched by numerous contributions of Dr. Hugh L. Dryden. Having worked in related fields of fluid mechanics during past years, and long respected both his professional contributions and personal integrity, it is a special pleasure to present this Dryden lecture.

The field of computational fluid dynamics during recent years has developed sufficiently to initiate some changes in traditional methods of aerodynamic design. Both computer power and numerical algorithm efficiency are simultaneously improving with time, while the energy resource for driving large wind tunnels is becoming progressively more valuable. Partly for these reasons it has been advocated that the impact of computational aerodynamics on future methods of aircraft design will be profound.[1,2] Qualitatively, the changes taking place are not foreign to past experience in other fields of engineering. For example, trajectory mechanics and neutron transport mechanics already have been largely revolutionized [illegible] been much reduced over former years; experiments now are performed mainly on clear, physically describable arrays of elements aimed at further confirmation of computational techniques; and better designs are achieved than with former experimental methods alone. Similar changes in the relative roles of experimental and computational aerodynamics are anticipated in the future.

There are three compelling motivations for vigorously developing computational aerodynamics. One is to provide important new technological capabilities that cannot be provided by experimental facilities. Because of their fundamental limitations, wind tunnels have rarely been able to simulate, for example, Reynolds numbers of aircraft flight, flowfield temperatures around atmosphere entry vehicles, aerodynamics of probes entering planetary atmospheres, aeroelastic distortions present in flight, or the propulsive-external flow interaction in flight. In addition, transonic wind tunnels are notoriously limited by wall and support interference; and stream nonuniformities of wind tunnels severely affect laminar-turbulent transition. Moreover, the dynamic-aerodynamic interaction between vehicle motion in flight and transition-dependent separated flow also is inaccessible to wind-tunnel simulation.[3] In still different ways ground facilities for turbomachinery experiments are limited in their ability, for example, to simulate flight inlet-flow nonuniformities feeding into a compressor stage, or to determine detailed flowfields between rotating blades. Numerical flow simulations, on the other hand, have none of these fundamental limitations, but have their own: computer speed and memory. These latter limitations are fewer, but previously have been much more restrictive overall because the full Navier-Stokes equations are of such great complexity [illegible] with the largest current computers. Since the fundamental limitations of computational speed and memory are rapidly decreasing with time, whereas the fundamental limitations of experimental facilities are not, numerical simulations offer the potential of mending many ills of wind-tunnel and turbomachinery experiments, and of providing thereby important new technical capabilities for the aerospace industry.

A second compelling motivation concerns energy conservation. The large developmental wind tunnels require large amounts of energy, whereas computers require comparatively

Dr. Dean R. Chapman is Director of Astronautics at the Ames Research Center. In this post, Dr. Chapman administers organizations conducting space flight projects. Dr. Chapman joined Ames as an aeronautical engineer in 1944. He received his B.S. and M.S. degrees from California Institute of Technology in 1944 and his Ph.D. from Cal Tech in 1948 under a National Research Council Fellowship. He has made fundamental contributions in aerodynamic flow separation at supersonic speeds, the effects of trailing edge bluntness on drag and lift, atmosphere entry physics, and the origin of tektites. In 1952, Dr. Chapman received the Lawrence Sperry Award of the Institute of Aeronautical Sciences for contributions to aeronautics. In 1959 he received a Rockefeller Public Service Award to conduct research for a year at the University of Manchester, England and at Jodrell Bank. In 1963 he was awarded the NASA medal for Exceptional Scientific Achievement for his work on tektites, atmosphere entry physics, and space mechanics. In 1971 he received the H. Julian Allen Award of the Ames Research Center for his tektite work. He was elected in 1975 to the U.S. National Academy of Engineering. He was appointed as 1978/79 Hunsaker Professor, an honorary professorship, at the Massachusetts Institute of Technology and presented the Dryden Research Lecture of the American Institute of Aeronautics and Astronautics. Dr. Chapman is author of numerous technical papers and an AIAA Fellow.

Presented as Paper 79-0129 at the AIAA 17th Aerospace Sciences Meeting, Jan. 15-17, 1979; submitted March 17, 1979; revision received Sept. 5, 1979.
Index categories: Computational Methods; Aerodynamics; Computer Technology.

ITEA Journal 2010; 31: 329–346

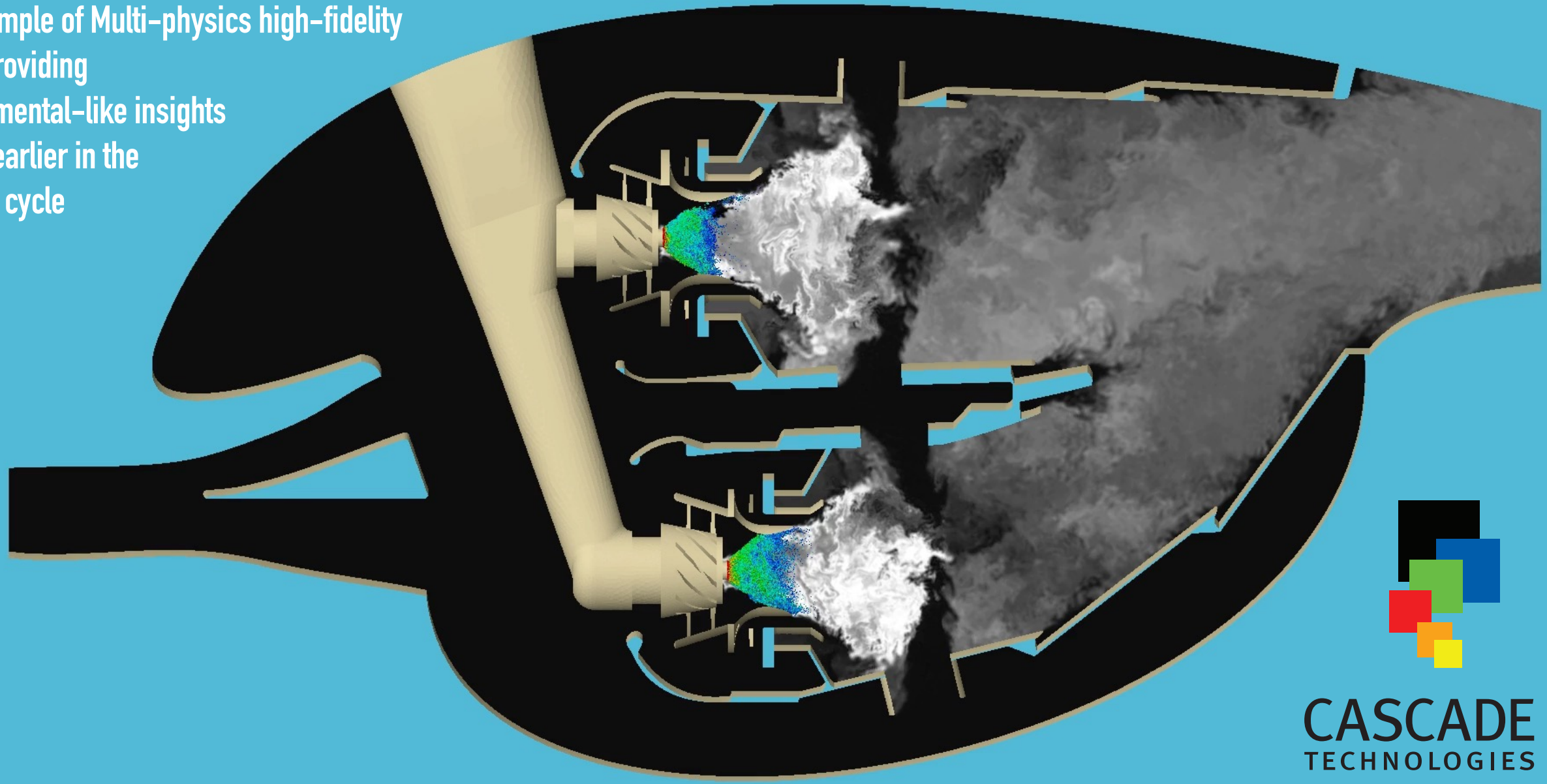
After 40 Years Why Hasn't the Computer Replaced the Wind Tunnel?

Edward M. Kraft, Ph.D.
USAF Arnold Engineering Development Center, Arnold Air Force Base, Tennessee

The debate between wind tunnels and computers to develop aeronautical systems has persisted for over 40 years. On the one hand, the majority of wind tunnels used today in aeronautical research, development, test, and evaluation were designed and commissioned in the 1950s and '60s. These facilities remain the backbone of the aeronautical development process, although they are becoming more challenging to maintain. On the other hand, rapid advances in computer hardware and software offer the potential to dramatically alter the design and development process for flight systems through the application of computational science and engineering. However, after 40 years of promises to eliminate the need for test facilities, advanced computational science and engineering have still not diminished significantly the need for test facilities or reduced the overall cycle time for development of flight systems. As many wind tunnel test hours are used today to develop a flight system as were used 20 years ago.

Key words: Aeronautical development; computational science and engineering;

An example of Multi-physics high-fidelity CFD, providing experimental-like insights much earlier in the design cycle



CASCADE TECHNOLOGIES

cadence

AI & Computational Fluid Dynamics

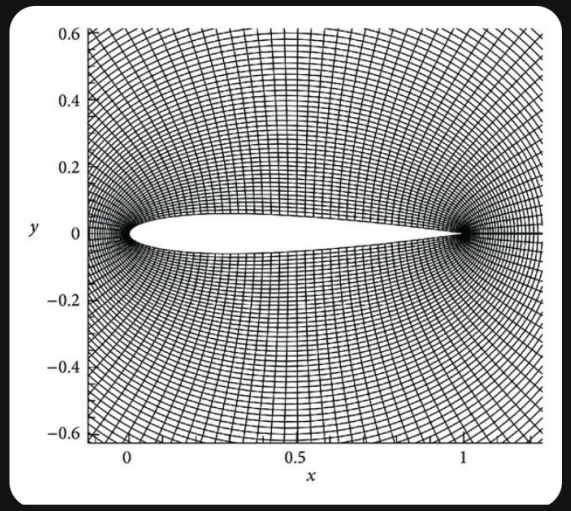
Ongoing research (and some objective successes)

- **rewrite code/kernels:** optimize performance, enable portability, etc.
- **identify best input:** (best practices): mesh resolution, convergence parameters, etc.

Gemini

Me

Stanford



I generated this mesh around an airfoil. Is it appropriate for simulating the flow at Reynolds number one million and angle of attack 10 degree using the RANS equations?

Analysis

✓ Query succe

To e candid

topology, it is

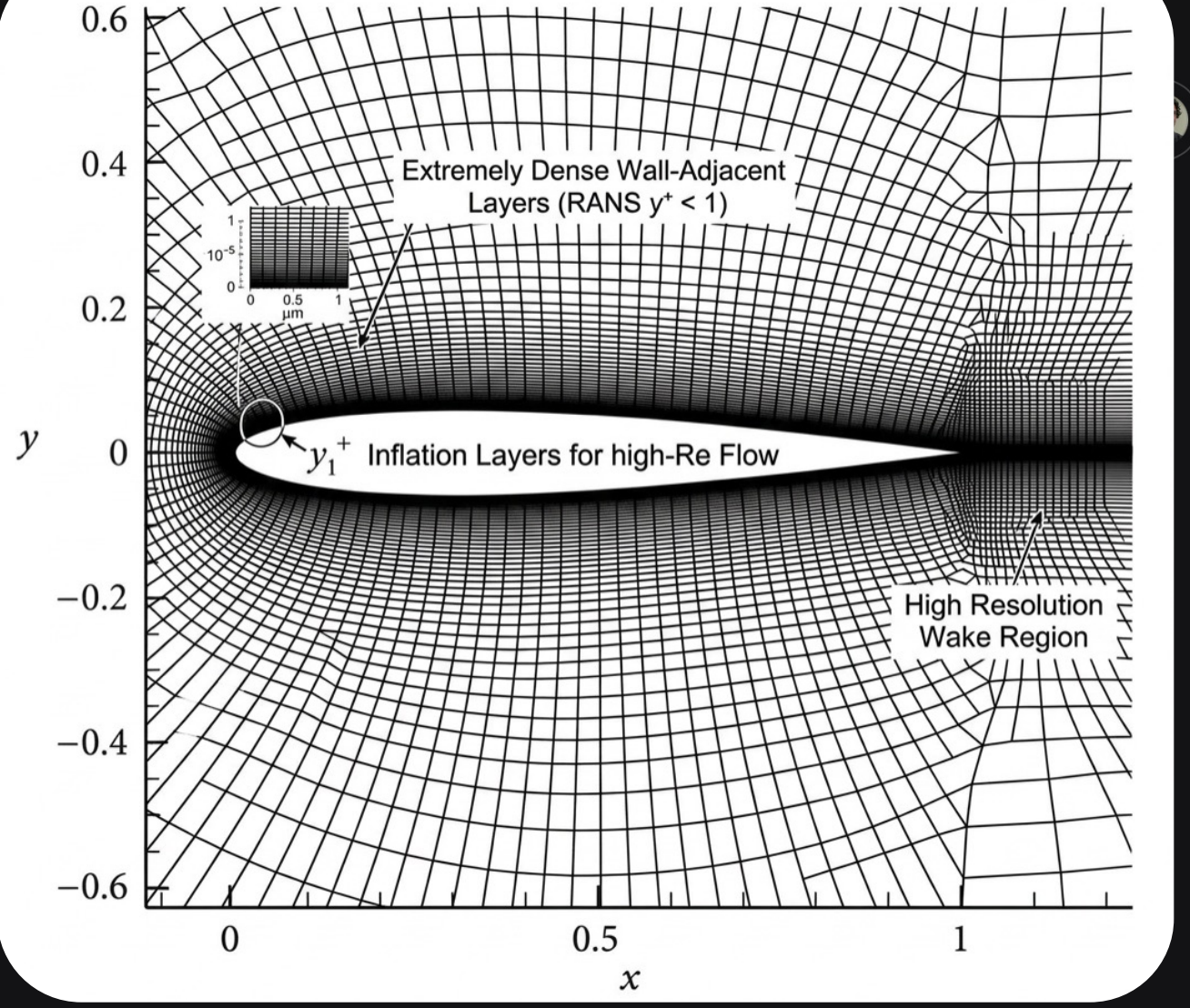
Reynolds nur

At $Re = 10^6$

the **boundar**

Why this me

- **Insuffic**
  is critica
  wall fun
  height s
  level.
- **Lack of Normal Stretching:** Your grid cells appear relatively uniform in height as they



Remark: gemini did not generate a new mesh, just a good target visual!

AI & Computational Fluid Dynamics

Ongoing research (and some objective successes)

- rewrite code/kernels: optimize performance, enable portability, etc.
- identify best input: (best practices): mesh resolution, convergence parameters, etc.
- aid with model parameters: inference of material properties, turbulence models constants, etc.
- embed a physics replacement: data-driven turbulence model, multiphysics effects, etc.

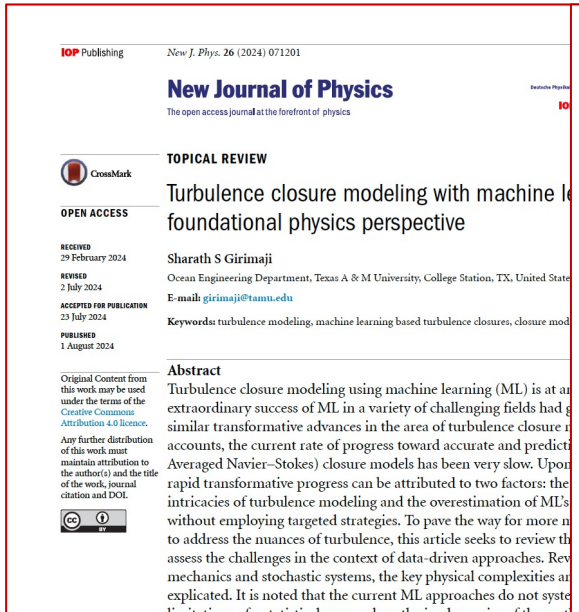
Machine Learning for Turbulence

Early Days

- Exciting opportunity to bring enthusiasm to the field of physics modeling
- Early work perhaps too "optimistic"
- Challenges include physical constraints, generalization, data scarcity
- "Embedded" models introduce numerical difficulties

A developing field

- ML and data-driven techniques can be useful tools
- Symbolic regression strategies help with interpretability
- Datasets (size and characteristics) are critical to success
- Equivariant graph models provide mathematical guarantees useful for modeling purposes


IOP Publishing New J. Phys. 26 (2024) 071201

New Journal of Physics
The open access journal at the forefront of physics

TOPICAL REVIEW

Turbulence closure modeling with machine l
foundational physics perspective

Sharath S Girimaji

Ocean Engineering Department, Texas A & M University, College Station, TX, United State

E-mail: girimaji@tamu.edu

OPEN ACCESS

RECEIVED
29 February 2024

REVISED
2 July 2024

ACCEPTED FOR PUBLICATION
23 July 2024

PUBLISHED
1 August 2024

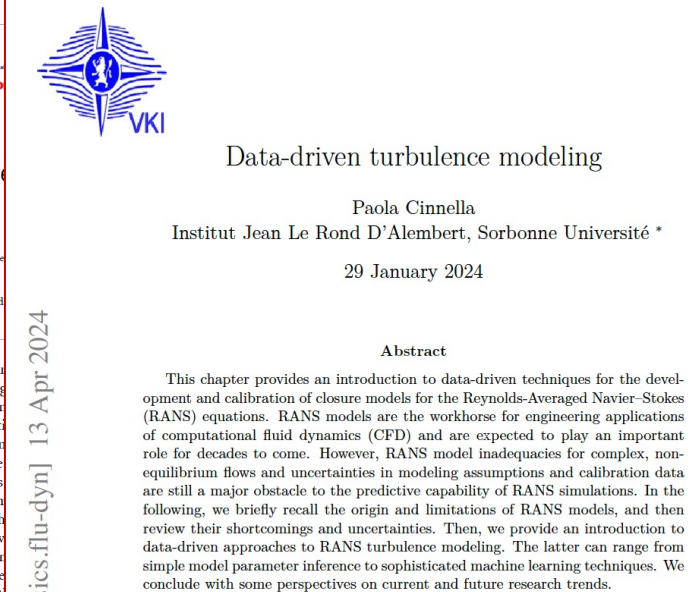
Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Keywords: turbulence modeling, machine learning based turbulence closures, closure mod

Abstract

Turbulence closure modeling using machine learning (ML) is at a
extraordinary success of ML in a variety of challenging fields had g
similar transformative advances in the area of turbulence closure
accounts, the current rate of progress toward accurate and predicti
Averaged Navier–Stokes) closure models has been very slow. Upon
rapid transformative progress can be attributed to two factors: the
intricacies of turbulence modeling and the overestimation of ML's
without employing targeted strategies. To pave the way for more
to address the nuances of turbulence, this article seeks to review th
assess the challenges in the context of data-driven approaches. Rev
mechanics and stochastic systems, the key physical complexities a
explicated. It is noted that the current ML approaches do not syste


VKI

Data-driven turbulence modeling

Paola Cinnella
Institut Jean Le Rond D'Alembert, Sorbonne Université *

29 January 2024

Abstract

This chapter provides an introduction to data-driven techniques for the development and calibration of closure models for the Reynolds-Averaged Navier–Stokes (RANS) equations. RANS models are the workhorse for engineering applications of computational fluid dynamics (CFD) and are expected to play an important role for decades to come. However, RANS model inadequacies for complex, non-equilibrium flows and uncertainties in modeling assumptions and calibration data are still a major obstacle to the predictive capability of RANS simulations. In the following, we briefly recall the origin and limitations of RANS models, and then review their shortcomings and uncertainties. Then, we provide an introduction to data-driven approaches to RANS turbulence modeling. The latter can range from simple model parameter inference to sophisticated machine learning techniques. We conclude with some perspectives on current and future research trends.


Acta Mech 236, 3295–3320 (2025)
https://doi.org/10.1007/s00707-025-04325-6

ORIGINAL PAPER

Boqian Zhang · Juanmian Lei

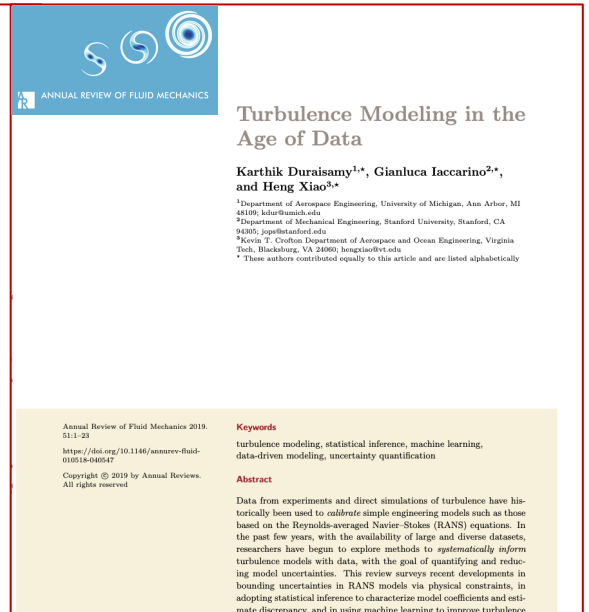
Interpretable data-driven turbulence modeling for separated flows using symbolic regression with unit constraints

Received: 21 November 2024 / Revised: 17 February 2025 / Accepted: 14 March 2025 / Published online: 21 April 2025
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2025

Abstract Machine learning techniques have been applied to enhance turbulence modeling in recent years. However, the "black box" nature of most machine learning techniques poses significant interpretability challenges in improving turbulence models. This paper introduces a novel unit-constrained turbulence modeling framework using symbolic regression to overcome these challenges. The framework amends the constitutive equation of linear eddy viscosity models (LEVMs) by establishing explicit equations between the Reynolds stress deviation and mean flow quantities, thereby improving the LEVM model's predictive capability for large separated turbulence. Unit consistency constraints are applied to the symbolic expressions to ensure physical realizability. The effectiveness of the framework and the generalization capability of the learned model are demonstrated through its application to the separated flow over 2D periodic hills and a backward-facing step. Compared to the standard k-ε model, the learned model shows significantly improved predictive accuracy for anisotropic Reynolds stresses, velocity and skin friction, while exhibiting promising generalization capabilities across various scenarios.

1 Introduction

Turbulence is prevalent across natural phenomena and engineering applications, and computational fluid dynamics (CFD) is an essential tool for studying turbulence. Achieving accurate and efficient numerical simulations of turbulent flows is a significant issue that persists in both academic and industrial spheres. Over the past few decades, a variety of numerical simulation techniques for turbulence have been developed, including direct numerical simulation (DNS), large eddy simulation (LES), and Reynolds-averaged Navier–Stokes (RANS) approaches, with RANS being the most extensively researched and applied. Despite earlier predictions suggesting that LES might supplant RANS in industrial applications within the forthcoming decades [1],


ANNUAL REVIEW OF FLUID MECHANICS

Turbulence Modeling in the Age of Data

Karthik Duraisamy[1,*], Gianluca Iaccarino[2,*], and Heng Xiao[3,*]

[1]Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI 48109; kdur@umich.edu
[2]Department of Mechanical Engineering, Stanford University, Stanford, CA 94305; jops@stanford.edu
[3]Kevin T. Crofton Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, VA 24060; hengxiao@vt.edu
* These authors contributed equally to this article and are listed alphabetically

Annual Review of Fluid Mechanics 2019. 51:1–23

https://doi.org/10.1146/annurev-fluid-010518-040547

Keywords
turbulence modeling, statistical inference, machine learning, data-driven modeling, uncertainty quantification

Abstract

Data from experiments and direct simulations of turbulence have historically been used to *calibrate* simple engineering models such as those based on the Reynolds-averaged Navier–Stokes (RANS) equations. In the past few years, with the availability of large and diverse datasets, researchers have begun to explore methods to *systematically inform* turbulence models with data, with the goal of quantifying and reducing model uncertainties. This review surveys recent developments in bounding uncertainties in RANS models via physical constraints, in adopting statistical inference to characterize model coefficients and estimate discrepancy, and in using machine learning to improve turbulence

AI & Computational Fluid Dynamics

Ongoing research (and some objective successes)

- rewrite code/kernels: optimize performance, enable portability, etc.
- identify best input: (best practices): mesh resolution, convergence parameters, etc.
- aid with model parameters: inference of material properties, turbulence models constants, etc.
- embed a physics replacement: data-driven turbulence model, multiphysics effects, etc.
- replace CFD simulations!
- autonomously orchestrate simulation campaign

Agentic Orchestration

- complex software workflows
- multi-agent interactions
- information retrieval
- agency

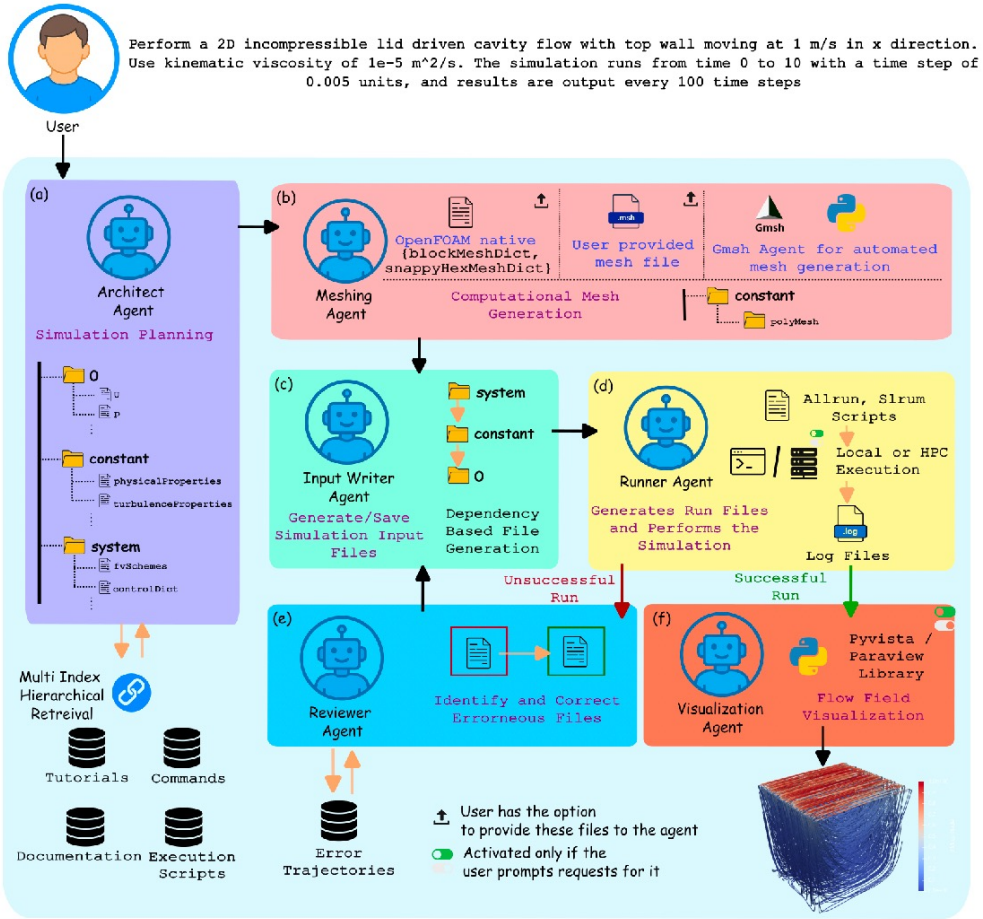
design process

AI agents

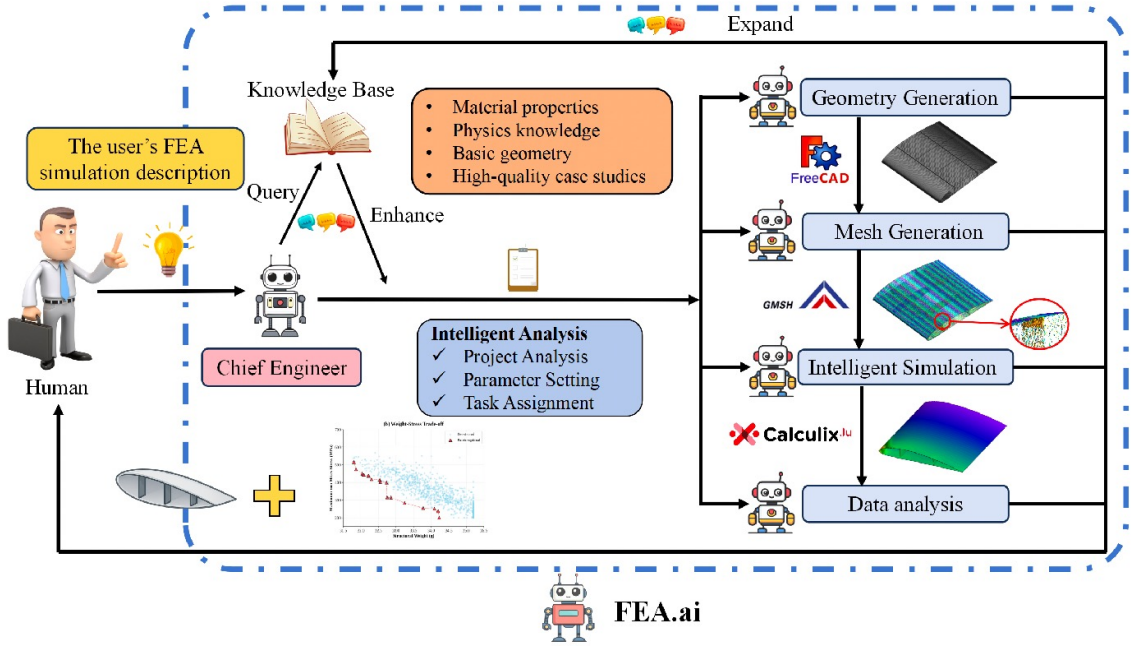
- understand design goals
- retrieve relevant examples in the literature
- define geometric parametrization
- build computational meshes
- select solver models and define input parameters
- launch simulations on HPC/cloud nodes
- analyze optimal designs
- compare and learn from past experiences
- decide on further iterations

Agentic Orchestration

- complex software workflows
- multi-agent interactions
- information retrieval
- agency

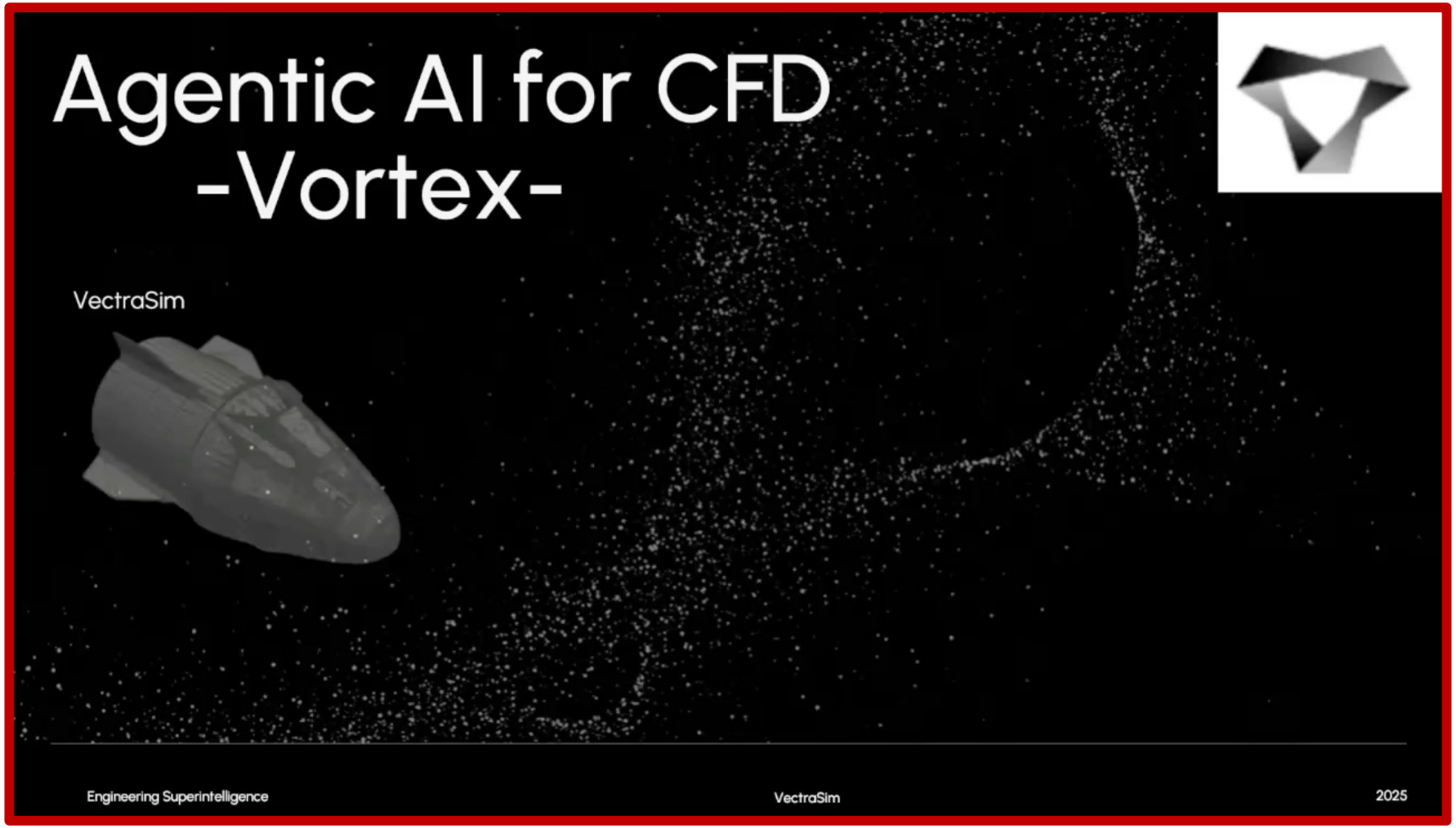


Yue et al. Foam-Agent 2.0: An End-to-End Composable Multi-Agent Framework for Automating CFD Simulation in OpenFOAM, 2025



Qi, Xu, Chu, FeaGPT: an End-to-End agentic-AI for Finite Element Analysis, 2025

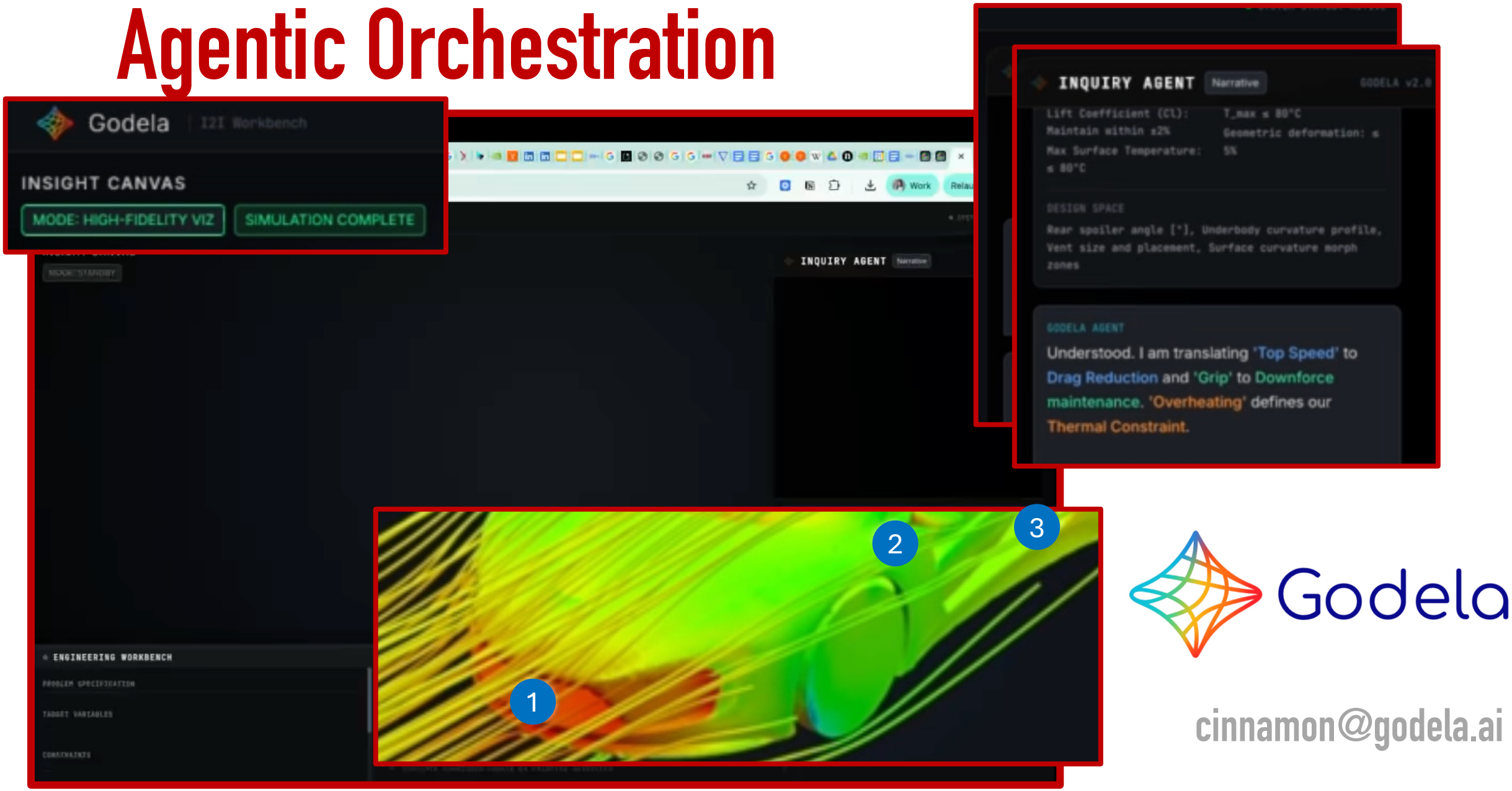
Agentic Orchestration



VECTRASIM

krissh@vectrasim.com

Agentic Orchestration



Godela

cinnamon@godela.ai

Model Context Protocol (MCP)

Foam-Agent 2.0: An End-to-End Composable Multi-Agent Framework for Automating CFD Simulation in OpenFOAM

Ling Yue[1†], Nithin Somasekharan[1†], Tingwen Zhang[1], Yadi Cao[2], Shaowu Pan[1*]

[1*]Rensselaer Polytechnic Institute.
[2]University of California San Diego.

*Corresponding author(s). E-mail(s): pans2@rpi.edu;
†These authors contributed equally to this work.

Abstract

Computational Fluid Dynamics (CFD) is an essential simulation tool in engineering, yet its steep learning curve and complex manual setup create significant barriers. To address these challenges, we introduce *Foam-Agent*, a multi-agent framework that automates the entire end-to-end `OpenFOAM` workflow from a single natural language prompt. Our key innovations address critical gaps in existing systems: 1. An Comprehensive End-to-End Simulation Automation: *Foam-Agent* is the first system to manage the full simulation pipeline, including advanced pre-processing with a versatile Meshing Agent capable of handling external mesh files and generating new geometries via `Gmsh`, automatic generation of HPC submission scripts, and post-simulation visualization via `ParaView`. 2. Composable Service Architecture: Going beyond a monolithic agent, the framework uses Model Context Protocol (MCP) to expose its core functions as discrete, callable tools. This allows for flexible integration and use by other agentic systems, such as Claude-code, for more exploratory workflows. 3. High-Fidelity Configuration Generation: We achieve superior accuracy through a Hierarchical Multi-Index RAG for precise context retrieval and a dependency-aware generation process that ensures configuration consistency. Evaluated on a benchmark of 110 simulation tasks, *Foam-Agent* achieves an 88.2% success rate with Claude 3.5 Sonnet, significantly outperforming existing frameworks (55.5% for Meta`OpenFOAM`). *Foam-Agent* dramatically lowers the expertise barrier for CFD, demonstrating how specialized multi-agent systems can democratize complex scientific computing. The code is public at https://github.com/csml-rpi/Foam-Agent.

Keywords: Large Language Model Agents, Simulation Automation, AI4Science, Computational Fluid Dynamics

https://github.com/webworn/openfoam-mcp-server

MCP is an open-source library for general interaction & communication between agents

MCP-based Agentic AI

- atomicity: each agent perform one task
- state awareness: tracking multi-stage simulations
- workflow decoupling: separating meshing, solving, and post-processing
- error checking
- interaction with LLMs and a vast ecosystem of applications

Claude LangGraph n8n

Outline

from well defined to open-ended inquiries

- forecast
- **design**
- discovery

agentic AI + simulations

not quite graphcast but exciting progress

open questions

Solver Failures & Benchmarking

CFDLLMBench: A Benchmark Suite for Evaluating Large Language Models in Computational Fluid Dynamics

Nithin Somasekharan[1] Ling Yue[1] Yadi Cao[2] Weichao Li[1]
Patrick Emami[5] Pochinapeddi Sai Bhargav[3] Anurag Acharya[4]
Xingyu Xie[1] Shaowu Pan[1]*
[1]Rensselaer Polytechnic Institute [2]University of California San Diego
[3]Indian Institute of Science [4]Pacific Northwest National Laboratory
[5]National Renewable Energy Laboratory

Abstract

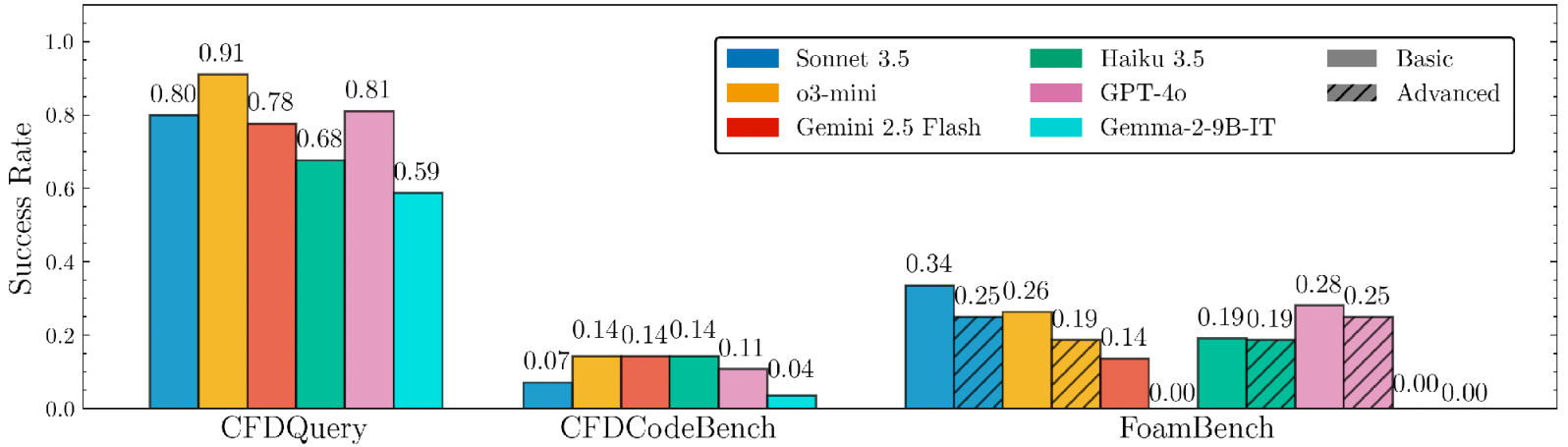
Large Language Models (LLMs) have demonstrated strong performance across general NLP tasks, but their utility in automating numerical experiments of complex physical system—a critical and labor-intensive component—remains underexplored. As the major workhorse of computational science over the past decades, Computational Fluid Dynamics (CFD) offers a uniquely challenging testbed for evaluating the scientific capabilities of LLMs. We introduce *CFDLLMBench*, a benchmark suite comprising three complementary components—*CFDQuery*, *CFDCodeBench*, and *FoamBench*—designed to holistically evaluate LLM performance across three key competencies: graduate-level CFD knowledge, numerical and physical reasoning of CFD, and context-dependent implementation of CFD workflows. Grounded in real-world CFD practices, our benchmark combines a detailed task taxonomy with a rigorous evaluation framework to deliver reproducible results and quantify LLM performance across code executability, solution accuracy, and numerical convergence behavior. *CFDLLMBench* establishes a solid foundation for the development and evaluation of LLM-driven automation of numerical experiments for complex physical systems. Code and data are available at https://github.com/NREL-Theseus/cfdllmbench/

1 Introduction

Recent advances in large language models (LLMs) have shown remarkable performance across general natural language processing tasks [19, 1]. However, their potential as scientific assistants—specifically, their ability to automate numerical simulation workflows—remains largely underexplored [10, 25]. Computational Fluid Dynamics (CFD) is critical in domains such as urban physics [7, 6], aerospace [46], climate [42], and aerial [43] and underwater robotics [28], and has labor-intensive workflows for computationally expensive numerical simulations of fluid dynamics. CFD workflows involve multiple steps, such as mesh generation, setup of boundary and initial conditions, and solver configuration. Such scientific workflows require an understanding of highly specialized knowledge [51], numerical and physical reasoning [50], and have context-dependent implementations involving domain-specific tool calling [20].

In this paper, we introduce *CFDLLMBench* (Figure 1), the first LLM benchmark for CFD composed of curated datasets designed to holistically evaluate LLMs' performance across three key competencies:

*Corresponding author: pans2@rpi.edu



~100 questions

- core concepts in fluids
- linear algebra, numerical methods,

24 CFD coding tests

- require Python coding
- selection of numerical method/parameters
- linear/non-linear PDEs
- 1D and 2D domains

~120 OpenFoam cases

- require full workflow
- includes grid generation
- selection of models and inputs parameters
- basic/advanced problems

Weak Baselines

Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations

Nick McGreivy[1,2*] and Ammar Hakim[2]

[1*]Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey, USA.
[2]Princeton Plasma Physics Laboratory, 100 Stellarator Rd, Princeton, New Jersey, USA.

*Corresponding author(s). E-mail(s): mcgreivy@princeton.edu;
Contributing authors: ahakim@pppl.gov;

Abstract

One of the most promising applications of machine learning (ML) in computational physics is to accelerate the solution of partial differential equations (PDEs). The key objective of ML-based PDE solvers is to output a sufficiently accurate solution faster than standard numerical methods, which are used as a baseline comparison. We first perform a systematic review of the ML-for-PDE solving literature. Of articles that use ML to solve a fluid-related PDE and claim to outperform a standard numerical method, we determine that 79% (60/76) compare to a weak baseline. Second, we find evidence that reporting biases, especially outcome reporting bias and publication bias, are widespread. We conclude that ML-for-PDE solving research is overoptimistic: weak baselines lead to overly positive results, while reporting biases lead to underreporting of negative results. To a large extent, these issues appear to be caused by factors similar to those of past reproducibility crises: researcher degrees of freedom and a bias towards positive results. We call for bottom-up cultural changes to minimize biased reporting as well as top-down structural reforms intended to reduce perverse incentives for doing so.

Keywords: machine learning, partial differential equations, metascience, reproducibility crisis

1 Introduction

Many fields of science have experienced reproducibility issues [1–3]. In some fields, reproducibility issues are thought to impact the validity of a significant percentage of published research [4–7]. These issues are often caused by pitfalls with data analysis and statistical techniques, as well as by a systemic bias towards publishing positive results [1–3, 8]. Because these issues can undermine the credibility and authority of an entire field, they are often referred to as a 'reproducibility crisis' [9].

As interest in machine learning (ML) has grown, more and more scientific fields are exploring whether ML can be used to advance science [10–15]. For some problems, ML has shown the potential to do so [16]. However, there are increasing concerns about reproducibility issues in ML [17–20] and in ML-based science [21, 22]. Compiling evidence from 22 articles across 17 fields analyzing reproducibility issues in 294 articles, Kapoor and Narayanan [21] argue that there is a 'reproducibility crisis' in ML-based science. Other large-scale analyses have found frequent reproducibility issues across hundreds of articles in medical ML [23–25]. Common pitfalls include data leakage [21, 26], poor data quality [21, 24, 27], weak baselines [23], and insufficient external validation [24, 25]. In each case, pitfalls result in overoptimistic assessments about the performance of ML.

1

focus on ML-for-fluids (PDEs)

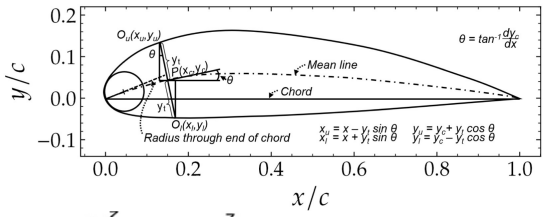
- **reviewed ~100 papers**
- **reproduced results for 1D/2D linear/non-linear equations (including NS)**

"Of articles that use ML to solve a fluid-related PDE and claim to outperform a standard numerical method, we determine that **79% (60/76) compare to a weak baseline**"

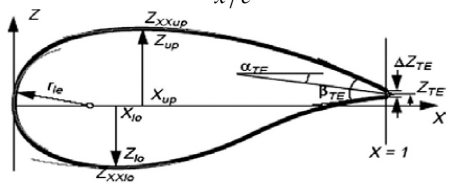
Article	Cited	PDE	Weaker baseline	Stronger baseline	Old outcome	New outcome
	941	b	PS 64×64	DG2 7×7	$10^3\times$ faster	$7\times$ faster
	911	c	FD n_x=100	DG2 n_x=13	$24\times$ faster	$10\times$ slower
	429	b	FV	PS	$80\times$ faster	slightly slower[†]
	382	a	WENO	DG2/DG3	4-8$\times$ fewer DOF	2-4$\times$ fewer DOF
	230	a	SP n_x=100	FV n_x=100	$10^3\times$ faster	$10\times$ slower
	124	b	PS 64×64	DG2 3×3	$10^3\times$ faster	$7\times$ faster
	101	e	MG	LU	faster	$10^3\times$ slower
	87	a,l	WENO, PS	WENO, FV	much faster	$10^3\times$ slower
	12	a	DG28 n_x=1	DG9 $n_x = 1$	22-75$\times$ faster	4-10$\times$ slower
	2	e	CG & MG	LU	$12\times$ faster	35-500$\times$ slower

Design Exploration

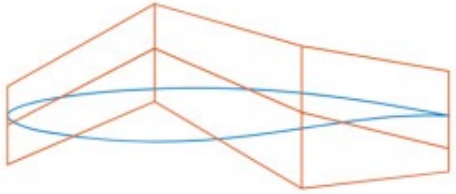
NACA



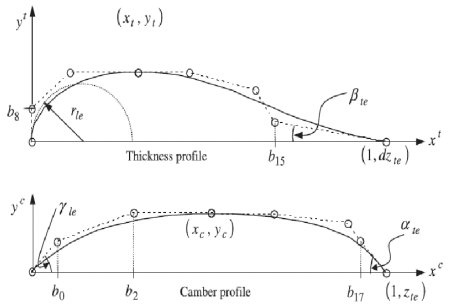
PARSED



FFD



BEZIER



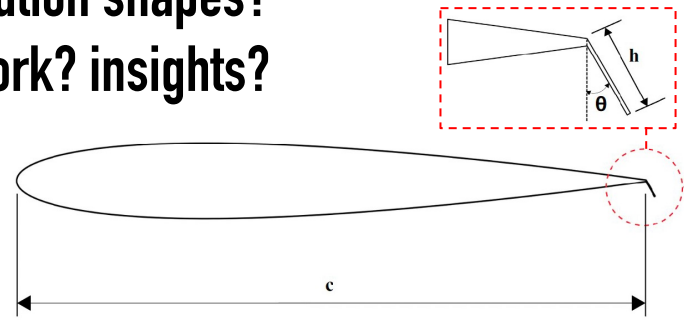
[...]

shape optimization requires flexible geometry parametrization

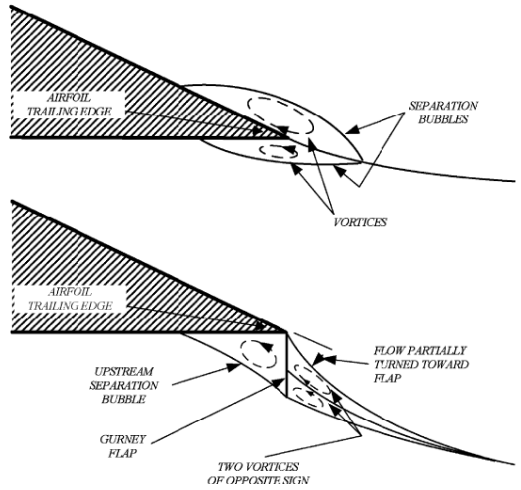
- different formulations (from human-centric to abstract)
- include O(10) parameters, i.e. large sampling space
- how do Agentic AI systems "learn" parametrizations?

aha-moment: Gurney flaps

- out-of-distribution shapes?
- why does it work? insights?



pressure increase on the lower surface leads t



Iyagi, A Novel Framework for Optimizing Gurney Flaps using RBF Neural Network, 2023

we need “Special Agents”

focused on critical and significant investigation

verification agent

- autonomously perform grid resolution studies, boundary location and conditions, order-of-accuracy assessments, etc.

validation agent

- find and retrieve relevant experimental data, establish the quality of the simulations, study limitations of models, perform inference/calibration

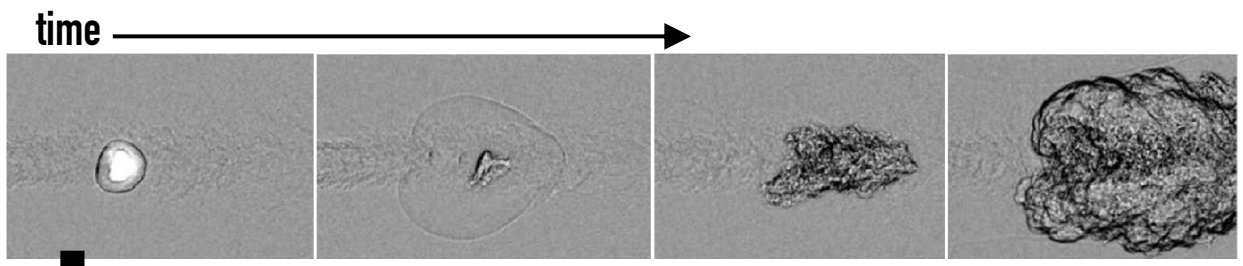
uncertainty quantification agent

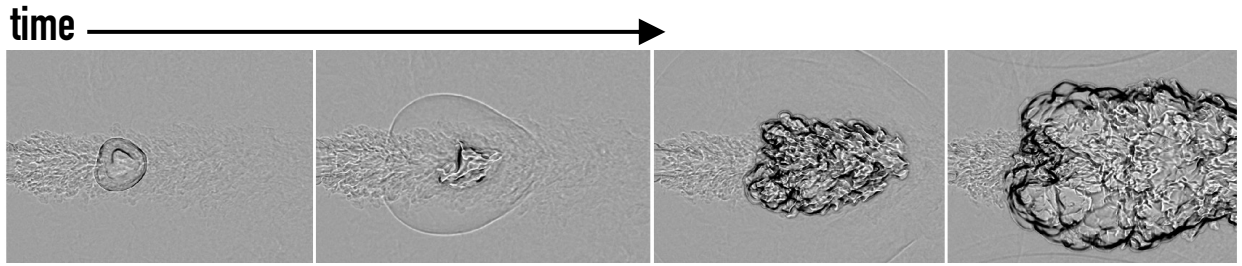
- retrieve information on parameter variability, plan multi-fidelity campaigns, build ensemble-based predictions

A Validation Agent

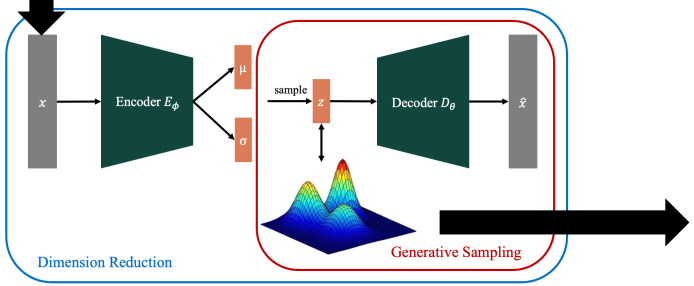
laser-induced ignition in high-speed co-axial jet

construct latent space for comparisons for 100s of 4D cases (AE, Fan et. al 2024)

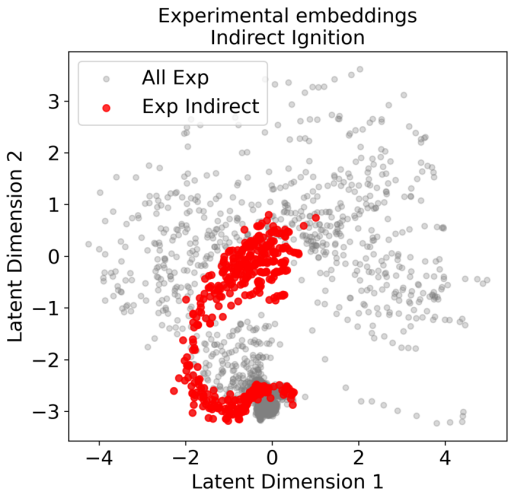




Experimental Schlieren

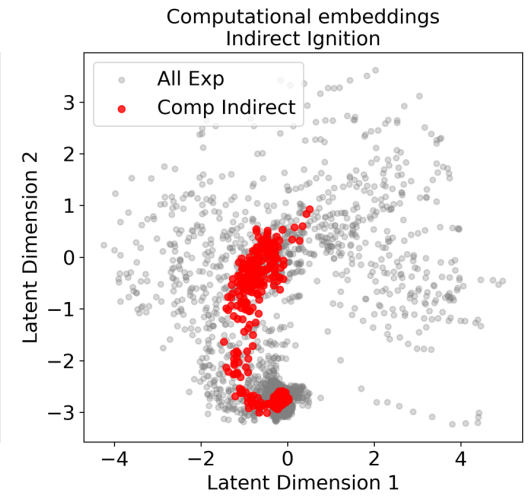


training



Computational Schlieren

pre-trained latent representation



Outline

from well defined to open-ended inquiries

- forecast
- **design**
- discovery

→

agentic AI + simulations

not quite graphcast but exciting progress

open questions

- accuracy, repeatability, trust
- “special agents” for V&V/UQ/certification
- explain/synthetize insights
- identifying the design space
- weak baselines

Outline

from well defined to open-ended inquiries

- forecast
- design
- discovery

AI Discovery

is there substantive evidence of AI-based discovery in fluids?

- equations are not the target — generalization (theory) is!
- insight is a discovery; (re)discovery is not discovery

is it simply too early? data-driven discovery is possible…

Discover | Discover Oncology | SPRINGER NATURE

LETTER · Discov Oncol. 2025 Mar 13;16:313. doi: 10.1007/s12672-025-02064-7

AI-driven biomarker discovery: enhancing precision in cancer diagnosis and prognosis

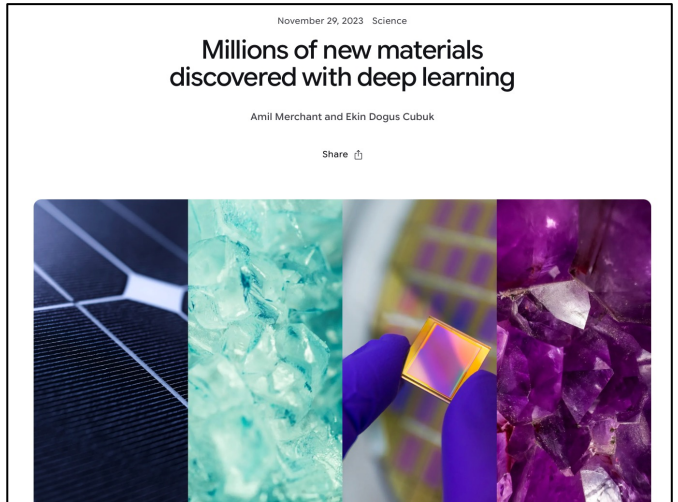
Esther Ugo Alum [1,✉]

Author information · Article notes · Copyright and License information

PMCID: PMC11906928 PMID: 40082367

Abstract

Cancer remains a significant health issue, resulting in around 10 million deaths per year, particularly in developing nations. Demographic changes, socio-economic variables, and lifestyle choices are responsible for the rise in cancer cases. Despite the potential to mitigate the adverse effects of cancer by early detection and the implementation of cancer prevention methods, several nations have limited screening facilities. In oncology, the use of artificial intelligence (AI) represents a transformative advancement in cancer diagnosis, prognosis, and treatment. The use of AI in biomarker discovery improves precision medicine by uncovering biomarker signatures that are essential for early detection and treatment of

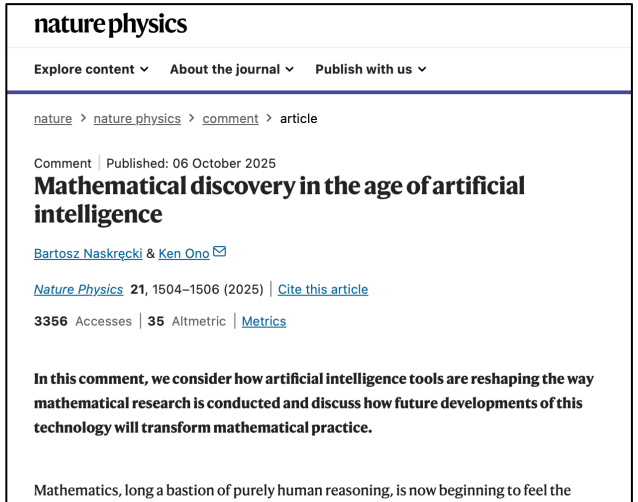


November 29, 2023 Science

Millions of new materials discovered with deep learning

Amil Merchant and Ekin Dogus Cubuk

Share



nature physics

Explore content · About the journal · Publish with us

nature › nature physics › comment › article

Comment | Published: 06 October 2025

Mathematical discovery in the age of artificial intelligence

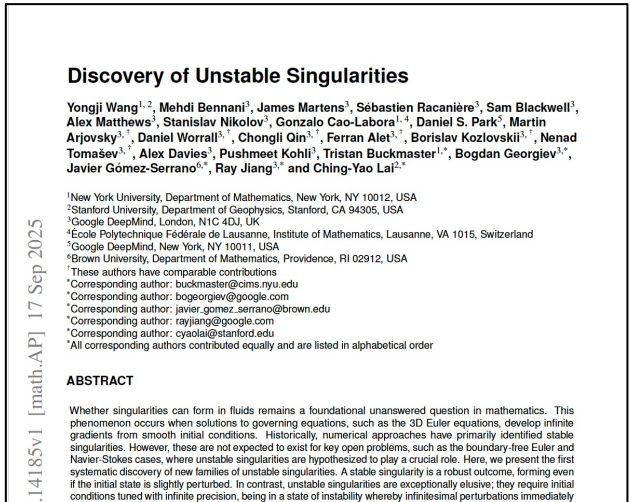
Bartosz Naskręcki & Ken Ono

Nature Physics **21**, 1504–1506 (2025) | Cite this article

3356 Accesses | **35** Altmetric | Metrics

In this comment, we consider how artificial intelligence tools are reshaping the way mathematical research is conducted and discuss how future developments of this technology will transform mathematical practice.

Mathematics, long a bastion of purely human reasoning, is now beginning to feel the



Discovery of Unstable Singularities

Yongji Wang[1, 2], Mehdi Bennani[3], James Martens[3], Sébastien Racanière[3], Sam Blackwell[3], Alex Matthews[3], Stanislav Nikolov[3], Gonzalo Cao-Labora[1, 4], Daniel S. Park[5], Martin Arjovsky[3, †], Daniel Worrall[3, †], Chongli Qin[3, †], Ferran Alet[3, †], Borislav Kozlovskii[3, †], Nenad Tomašev[3, †], Alex Davies[3], Pushmeet Kohli[3], Tristan Buckmaster[1, *], Bogdan Georgiev[3, *], Javier Gómez-Serrano[6, *], Ray Jiang[3, *] and Ching-Yao Lai[2, *]

[1]New York University, Department of Mathematics, New York, NY 10012, USA
[2]Stanford University, Department of Geophysics, Stanford, CA 94305, USA
[3]Google DeepMind, London, N1C 4DJ, UK
[4]École Polytechnique Fédérale de Lausanne, Institute of Mathematics, Lausanne, VA 1015, Switzerland
[5]Google DeepMind, New York, NY 10011, USA
[6]Brown University, Department of Mathematics, Providence, RI 02912, USA
[†]These authors have comparable contributions
[*]Corresponding author: buckmaster@cims.nyu.edu
[*]Corresponding author: bogeorgiev@google.com
[*]Corresponding author: javier_gomez_serrano@brown.edu
[*]Corresponding author: rayjiang@google.com
[*]Corresponding author: cyaolai@stanford.edu
[*]All corresponding authors contributed equally and are listed in alphabetical order

ABSTRACT

Whether singularities can form in fluids remains a foundational unanswered question in mathematics. This phenomenon occurs when solutions to governing equations, such as the 3D Euler equations, develop infinite gradients from smooth initial conditions. Historically, numerical approaches have primarily identified stable singularities. However, these are not expected to exist for key open problems, such as the boundary-free Euler and Navier-Stokes cases, where unstable singularities are hypothesized to play a crucial role. Here, we present the first systematic discovery of new families of unstable singularities. A stable singularity is a robust outcome, forming even if the initial state is slightly perturbed. In contrast, unstable singularities are exceptionally elusive; they require initial conditions tuned with infinite precision, being in a state of instability whereby infinitesimal perturbations immediately

Human Discovery

O. Reynolds

the Reynolds number was introduced to synthetize experimental data on laminar and turbulent flows

induction driven discovery

the energy cascade & 2/3 laws were derived from statistical theory and then verified experimentally

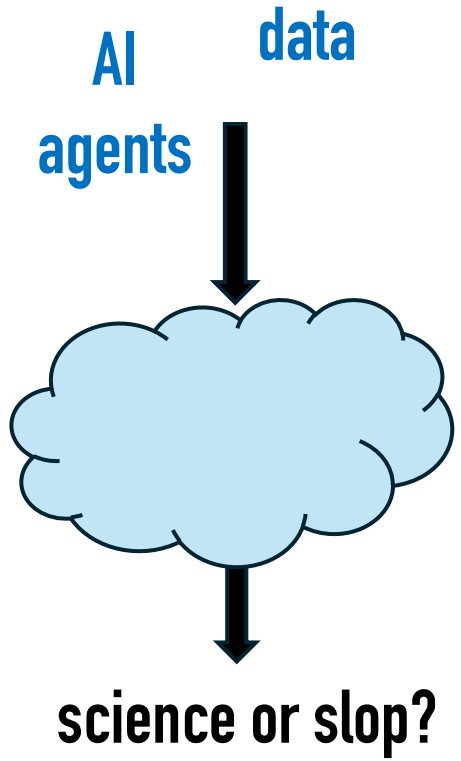
deduction driven discovery

A. N. Kolmogorov

Outline

from well defined to open-ended inquiries

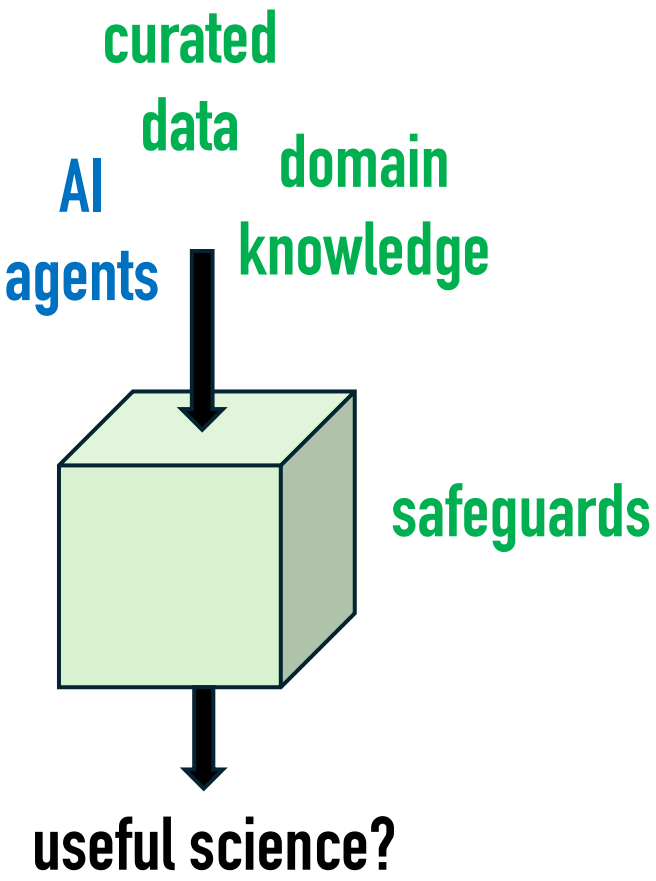
- forecast
- design
- discovery ⟶ AI



Outline

from well defined to open-ended inquiries

- forecast
- design
- discovery ⟶ Humans + AI



Closing Thoughts

AI & Agentic AI

- forecast
- design
- discovery

Thank You

Gianluca Iaccarino
jops@stanford.edu
Stanford University