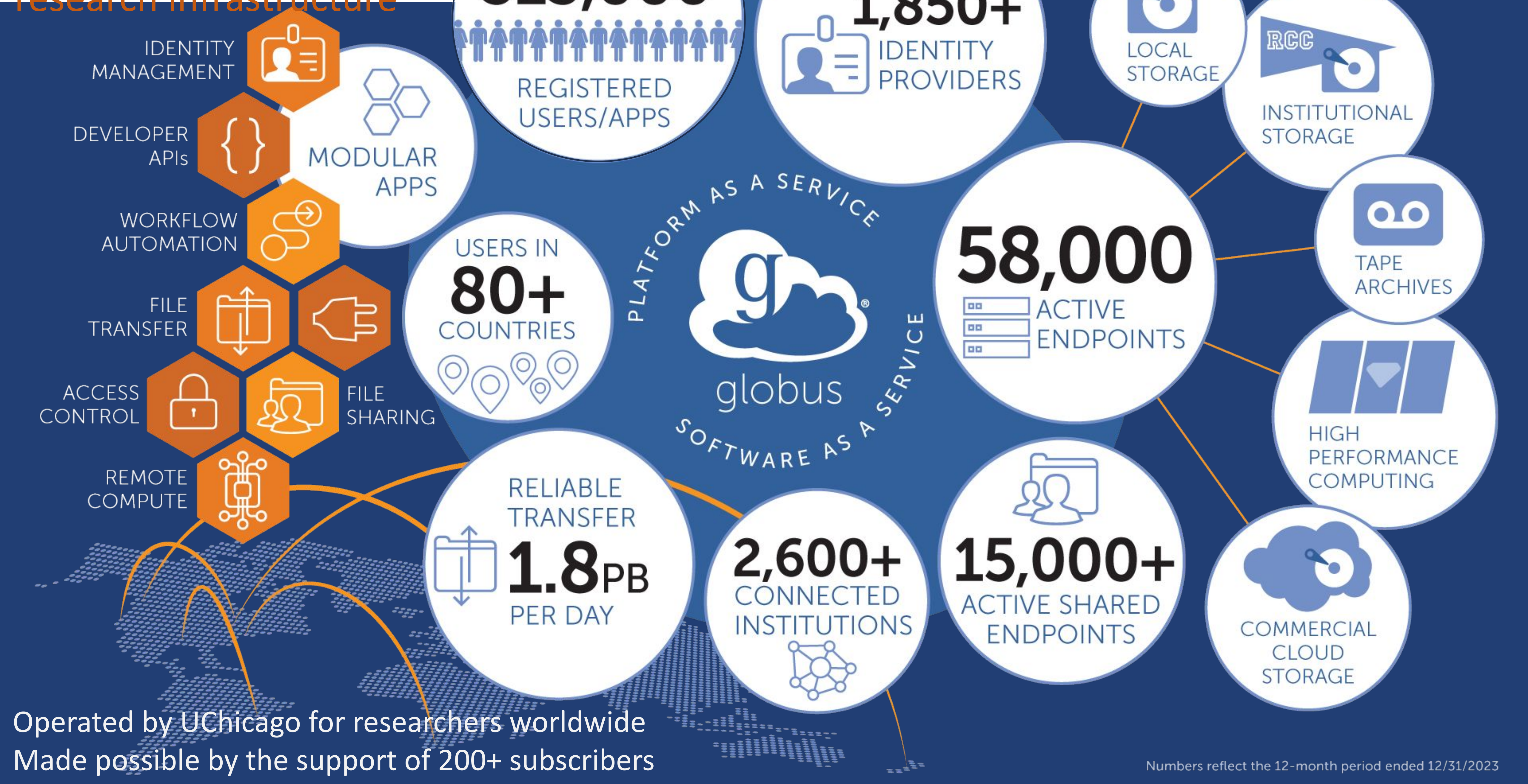
AuroraGPT
A foundation model for science

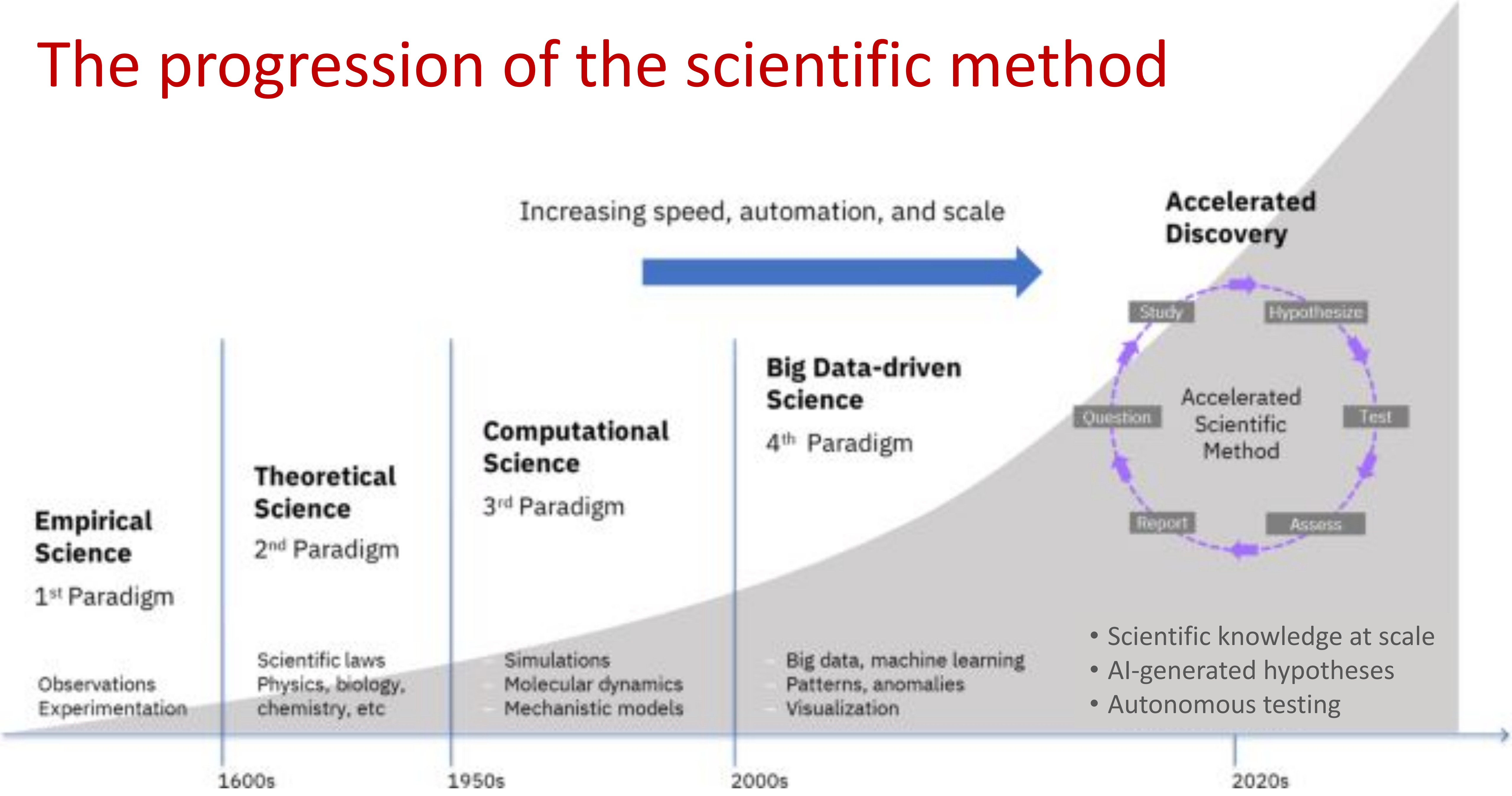
Ian Foster
Argonne National Laboratory
The University of Chicago

Globus: A global integrated research infrastructure



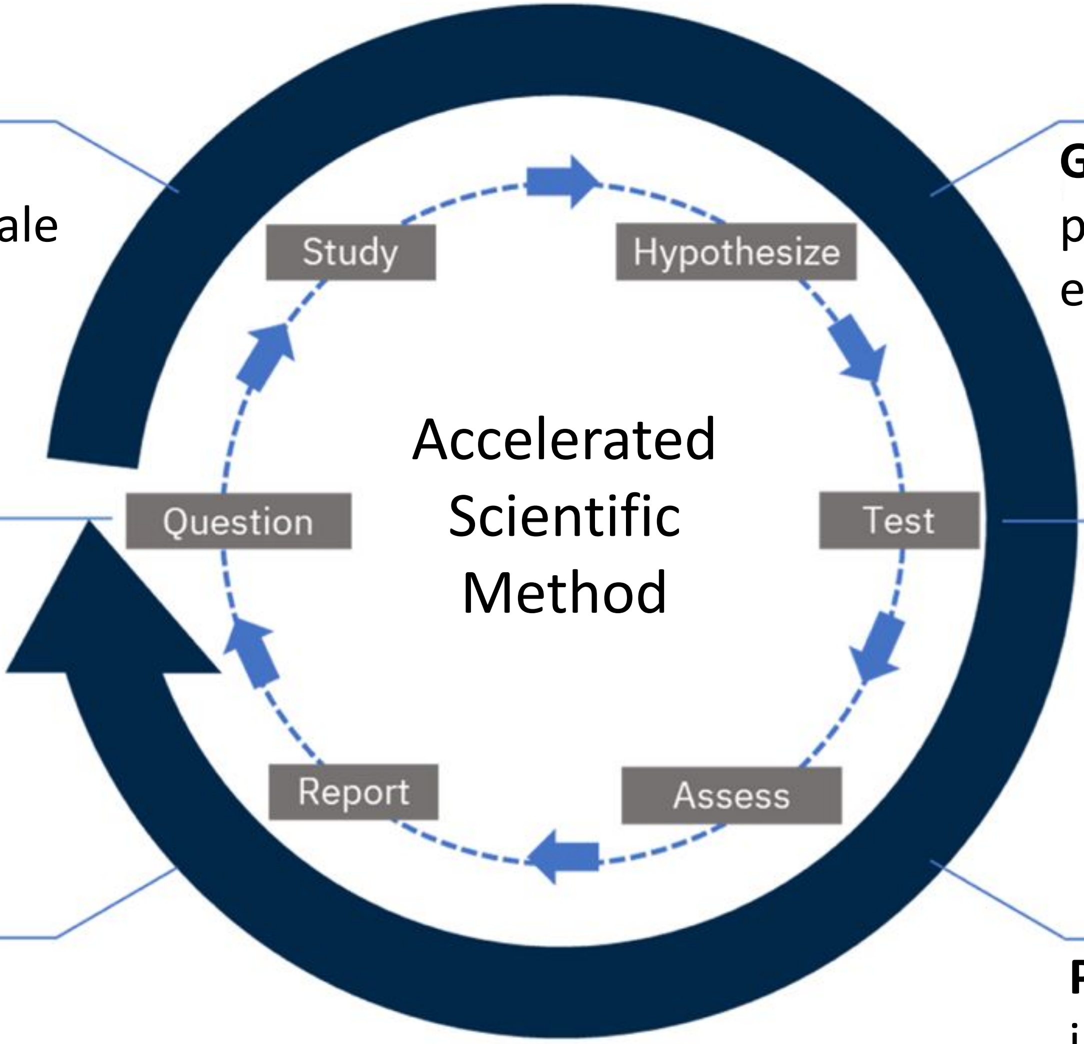
Operated by UChicago for researchers worldwide
Made possible by the support of 200+ subscribers

The progression of the scientific method



https://doi.org/10.1038/s41524-022-00765

Accelerating discovery using AI, HPC, and robotics



Extraction, integration and reasoning with knowledge at scale

Tools help **identify new questions** based on needs and gaps in knowledge

Machine representation of knowledge leads to **new hypotheses and questions**

Generative models automatically propose new hypotheses that expand the discovery space

Robotic labs automate experimentation and bridge digital models and physical testing

Pattern and anomaly detection is integrated with simulation and experimentation to extract new insights

https://doi.org/10.1038/s41524-022-00765

Embodied agents as first-class participants in discovery

Persistent, stateful

Sense, act

A **computational system** that can **interact** with its **environment** and **learn** from those interactions

Physical or virtual

Adapt responses

For example: A **peptide expert**

(Prototyped with PubMed and ChatGPT)

PubMed ChatGPT

Query PubMed for ChatGPT feedstock

Retrieve abstracts **A** from PubMed that reference specified **peptide**

Use ChatGPT to build hypotheses by using retrieval-augmented generation: e.g.:

"Given **A**, on which organism is {**peptide**} acting?"

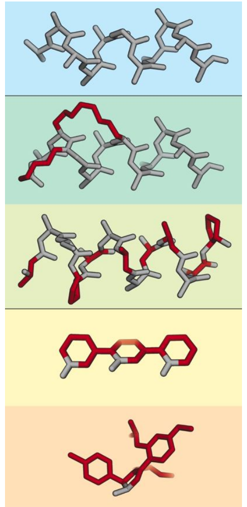
We want a model with deep expertise regarding peptides and related topics

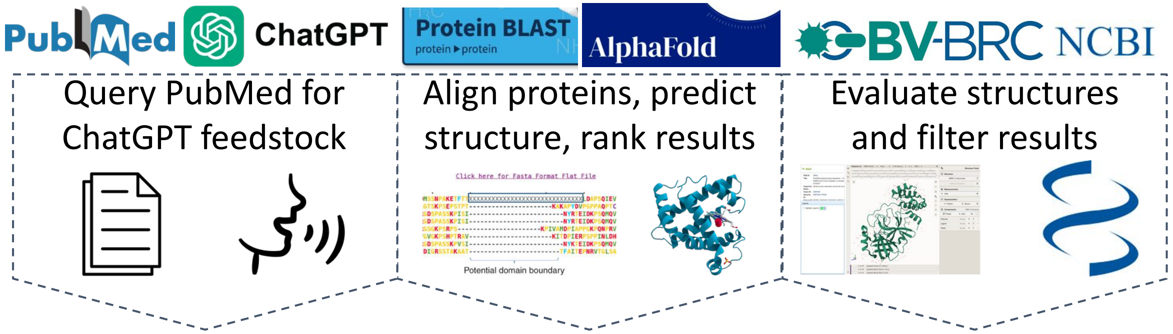
We want to be able to make millions of such requests

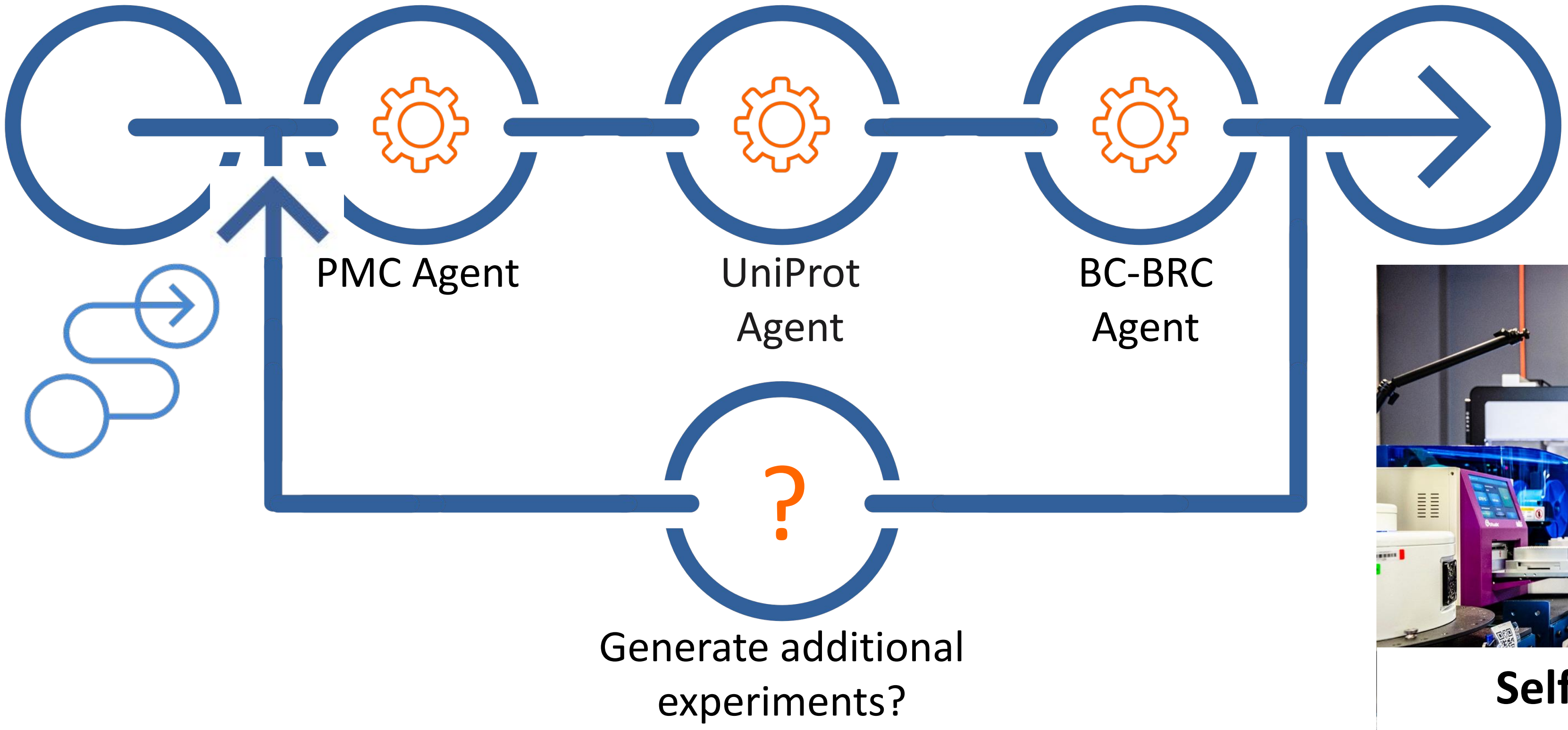
Arvind Ramanathan, Priyanka Setty, et al.

Peptide agent may be used with other agents to identify antimicrobial peptides

Set of peptides as input

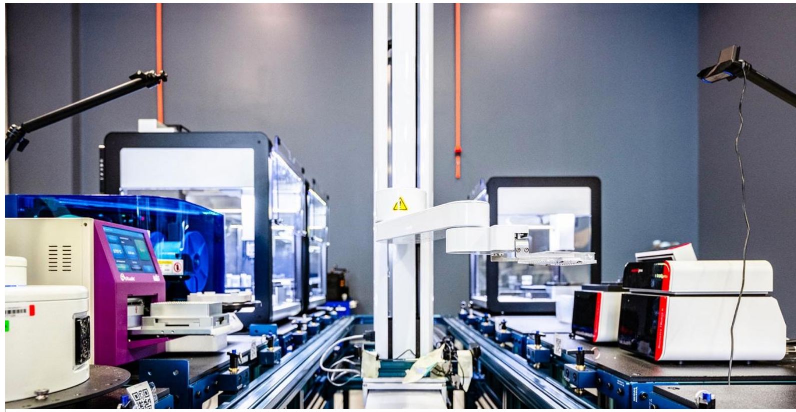






Agents run on HPC/AI resources

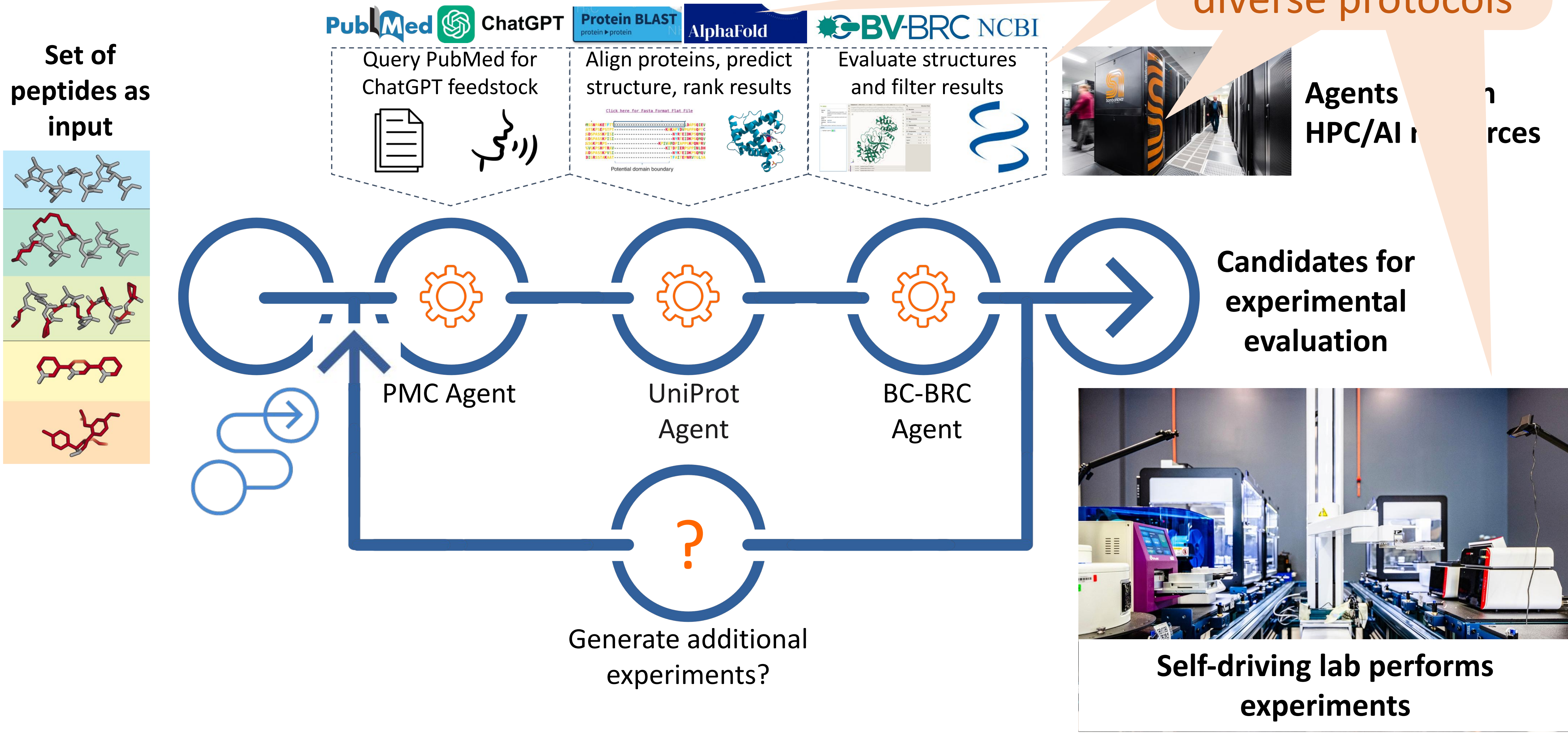
Candidates for experimental evaluation



Self-driving lab performs experiments

Peptide agent may be used with other ager
identify antimicrobial peptides

We want models that know about diverse protocols



AI for science: One or many foundation models?

Discovery
- Biology
- Materials
- Chemistry
- Batteries
- Drugs
- Devices

Control
- Accelerators
- Experiments
- Reactors
- Simulation
- Mobility
- Energy

Augmented Simulations
- Landscape Search
- Surrogates
- Optimize

Science and Math Comprehension
- Physics
- Mathematics
- Biology Chemistry Materials

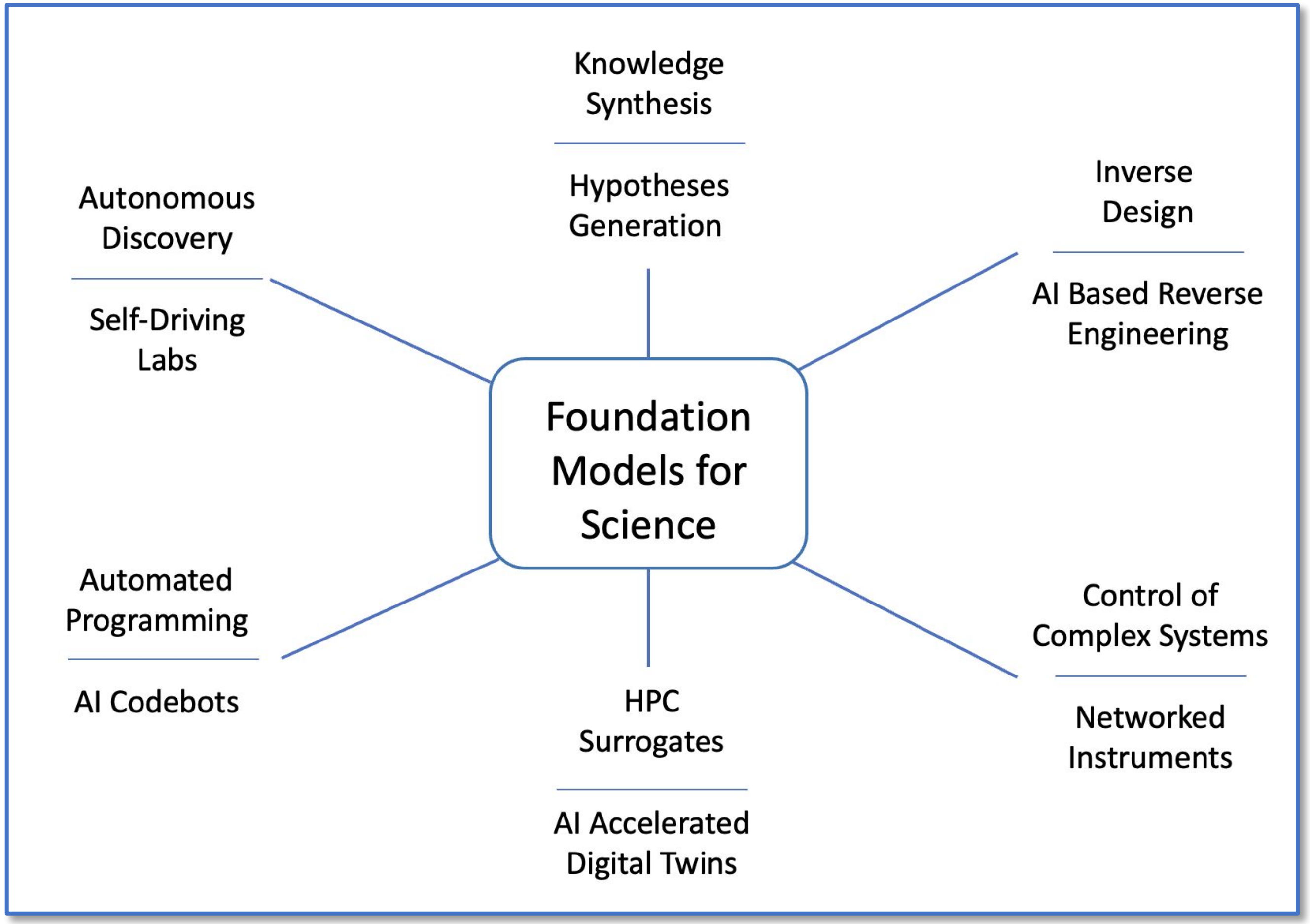
Generative Models
- Cosmology
- Detector Simulations
- Biodesign Materials

Inverse Problems
- Spectra 2 Structures
- Image2Phase
- Waveform 2 Source

Multimodal Learning
- Text
- Waveforms
- Images
- Graphs
- Structured
- Time-series

Decision-Making
- Research Priorities
- Risk Assessment
- The Next Big Problem

We hypothesize that many science use cases can be driven directly or indirectly from sufficiently powerful Foundation Models



Open science foundation model(s)

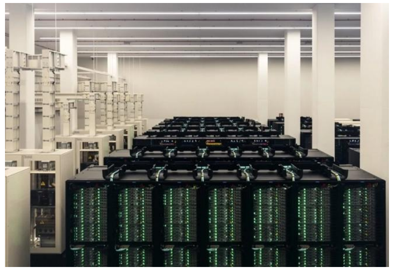
Scientific & Engineering Datasets

- Mathematics
- Biology
- Materials
- Chemistry
- Particle Physics
- Nuclear Physics
- Computer Science
- Climate
- Medicine
- Cosmology
- Fusion Energy
- Accelerators
- Reactors
- Energy Systems
- Manufacturing

Text and Code Corpora

- General Text
- Media
- News
- Humanities
- History
- Law
- Digital Libraries
- OSTI Archive
- Scientific Journals
- arXiv
- Code repositories
- Data.gov
- PubMed
- Agency Archives

Training



Open Science Foundation Model

Tuned and Adapted Downstream Models

Downstream Scientific Tasks

- Scientific Discovery
- Digital Twins
- Inverse Design
- Code Optimization
- Accelerated Simulations
- Autonomous Experiments
- Co-Design

AuroraGPT:
A foundation model for open science

- **General purpose scientific LLM:** Broadly trained, on general corpora; scientific papers and texts; structured science data
- **Explore pathways** towards a "Scientific Assistant"
- **Built with international partners**
- **Multilingual**: English, 日本語, French, German, Spanish, Italian, …
- **Multimodal**: Images, tables, equations, proofs, time-series, graphs, fields, sequences, …

A founding member of:

https:// tpc.dev

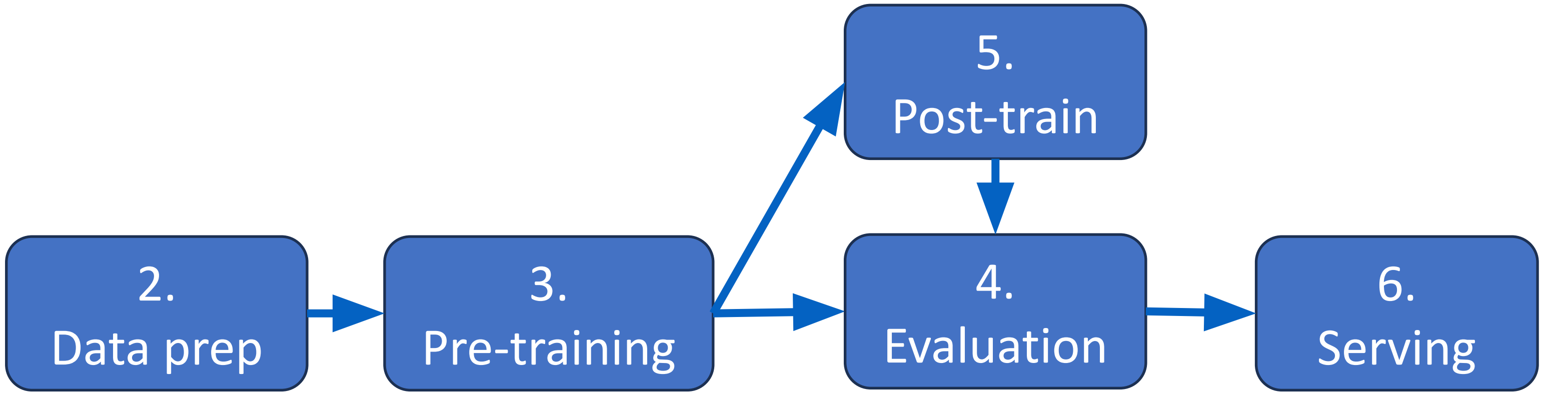
Trillion Parameter Consortium

AuroraGPT: A foundation model for open science

- A series of LLMs (7B, 70B, 200B, 1000B, etc. params)
- Trained on a mixture of general text, code, and scientific domain knowledge (Biology, Physics, Materials/Chemistry, Climate, Computer Science, Nanoscience, Cancer, Biomedicine, Energy Technologies)
- Domain knowledge beyond information in Common Crawl (RP2, Dolma, Pile), ArXiv, PMC, etc., to include text-encoded forms of structured scientific data from variety of domain data resources
- Multiple phases of development:
  - Phase 1 – Text oriented models – raw and instruct models (2023/2024)
  - Phase 2 – Basic multimodal models (2024/2025)
  - Phase 3 – Advanced scientific multi-model models (2025/2026)

AuroraGPT working groups

- 01 Planning
- 02 Data prep
- 03 Model Training
- 04 Evaluation
- 05 Post-Pretraining
- 06 Inference
- 07 Distribution
- 08 Communication



Aurora is:

166 Racks

10,624 Nodes

21,248 CPUs

63,744 GPUs

84,992 NICs

8 PB HBM

10 PB DDR5

Feasibility of Training Models on Aurora/Polaris

AuroraGPT set of models (1.5B, **7B**, 13B, **70B**, **200B**, **1T**, …)

Aurora BFP16 HGEMM ~ 180 TF per tile x (127,488 tiles) ⟹ 22.9 TF/s

Model Size (# of Parameters in Billions)	Training Tokens (Trillions)	Training F/P/T	Total Training Compute (Flops in BF16)	Total Training Compute (EF-days)	Aurora Time (Days)	Aurora Time (Hours)	Polaris Time (Days)	Polaris Time (Hours)	Cloud Cost ($3 GPU/hr)
1.5	1	6	9E+21	0.10	0.01	0.25	1	36	$46,871
1.5	2	6	1.8E+22	0.21	0.02	0.49	3	71	$93,741
1.5	3	6	2.7E+22	0.31	0.03	0.74	4	107	$140,612
7	1	6	4.2E+22	0.49	0.05	1.14	7	167	$218,729
7	2	6	8.4E+22	0.97	0.10	2.29	14	333	$437,459
7	3	6	1.26E+23	1.46	0.14	3.43	21	500	$656,188
70	2	6	8.4E+23	9.72	0.95	22.88	139	3,333	$4,374,588
70	3	6	1.26E+24	14.58	1.43	34.31	208	5,000	$6,561,882
70	4	6	1.68E+24	19.44	1.91	45.75	278	6,667	$8,749,176
200	6	6	7.2E+24	83.33	8.17	196.08	1,190	28,571	$37,496,471
200	10	6	1.2E+25	138.89	13.62	326.80	1,984	47,619	$62,494,118
200	15	6	1.8E+25	208.33	20.42	490.20	2,976	71,429	$93,741,176
1000	10	6	6E+25	694.44	68.08	1633.99	9,921	238,095	$312,470,588
1000	20	6	1.2E+26	1388.89	136.17	3267.97	19,841	476,190	$624,941,176
1000	30	6	1.8E+26	2083.33	204.25	4901.96	29,762	714,286	$937,411,765

We are assuming about 40% efficiency for LLM BFP16 flops utilization relative to HGEMM measurements

Slide: Rick Stevens

Trillion Parameter Consortium

Generative AI for Science

November 10, 2023

Rick Stevens, Charlie Catlett
Argonne National Laboratory

Founding partners come from many organizations

AI Singapore: Leslie Teo
Allen Institute For AI: Noah Smith
AMD: Michael Schulte
Argonne National Laboratory: Ian Foster
Barcelona Supercomputing Center: Mateo Valero Cortes
Brookhaven National Laboratory: Shantenu Jha
CalTech: Anima Anandkumar
CEA: Christoph Calvin
Cerebras Systems: Andy Hock
CINECA: Laura Morselli
CSC - IT Center for Science: Per Öster
CSIRO: Aaron Quigley
ETH Zürich: Torsten Hoefler
Fermilab : Jim Amundson
Flinders University: Rob Edwards
Fujitsu Limited: Koichi Shirahata
HPE: Nic Dube
Intel: Koichi Yamada
Juelich Supercomputing Center: Thomas Lippert
Kotoba Technologies, Inc.: Jungo Kasai

LAION: Jenia Jitsev
Lawrence Berkeley National Laboratory: Stefan Wild
Lawrence Livermore National Laboratory: Brian Van Essen
Leibniz Supercomputing Centre: Dieter Kranzlmüller
Los Alamos National Laboratory: Jason Pruet
Microsoft: Shuaiwen Leon Song
National Center for Supercomputing Applications: Bill Gropp
AIST - Japan: Yoshio Tanaka
National Renewable Energy Laboratory: Juliane Mueller
National Supercomputing Centre, Singapore: Tin Wee Tan
NCI Australia: Jingbo Wang
New Zealand eScience Infrastructure: Nick Jones
Northwestern University: Pete Beckman
NVIDIA: Giri Chukkapalli

Oak Ridge National Laboratory: Prasanna Balaprakash
Pacific Northwest National Laboratory: Neeraj Kumar
Pawsey Institute: Mark Stickells
Princeton Plasma Physics Laboratory: William Tang
RIKEN: Makoto Taiji
Rutgers University: Shantenu Jha
SambaNova: Marshall Choy
Sandia National Laboratories: John Feddema
Seoul National University: Jiook Cha
SLAC National Accelerator Laboratory: Daniel Ratner

Stanford University: Sanmi Koyejo
STFC Rutherford Appleton Laboratory, UKRI: Jeyan Thiyagalingam
Texas Advanced Computing Center: Dan Stanzione
Thomas Jefferson National Accelerator Facility: Malachi Schram
Together AI: Ce Zhang
Tokyo Institute of Technology: Rio Yokota
Université de Montréal: Irina Rish

University of Chicago: Rick Stevens
University of Delaware: Ilya Safro
University of Illinois Chicago: Michael Papka
University of Illinois Urbana-Champaign: Lav Varshney
University of New South Wales: Tong Xie
University of Tokyo: Kengo Nakajima
University of Toronto: Alan Aspuru-Guzik
University of Utah: Manish Parashar
University of Virginia: Geoffrey Fox

TPC contact: Charlie Catlett
Learn more at tpc.dev

https:// tpc.dev