CS181 Artificial Intelligence

AI for Science

Jie Zheng (郑杰)

PhD, Associate Professor

School of Information Science and Technology (SIST), ShanghaiTech University

Dec. 29, 2023

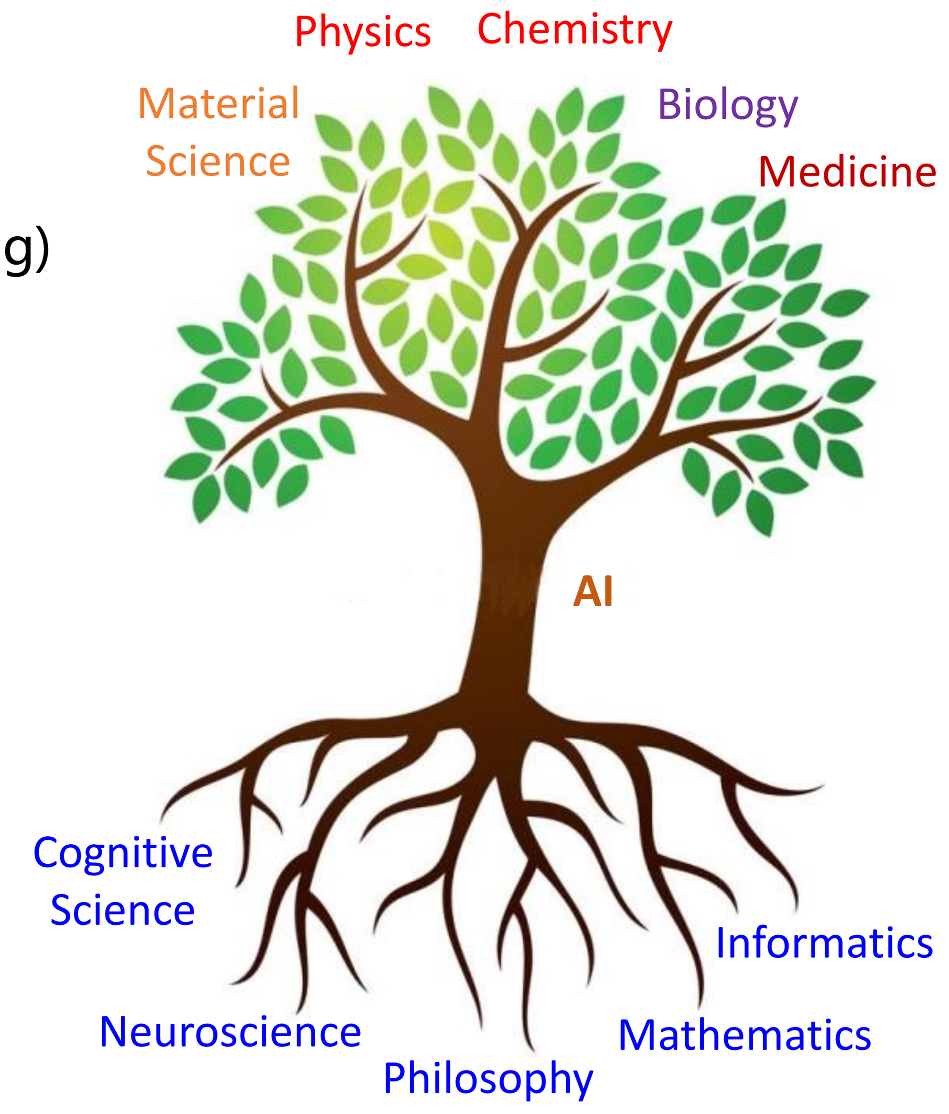
Outline

- **Background**
- AI for scientific discovery pipelines
  - Data collection and curation
  - Learning meaningful representations
  - AI-based generation of scientific hypotheses
  - AI-driven experimentation and simulation
- Large language models (LLMs) for Science
- Grant challenges

Why does Science need AI?

上海科技大学 ShanghaiTech University

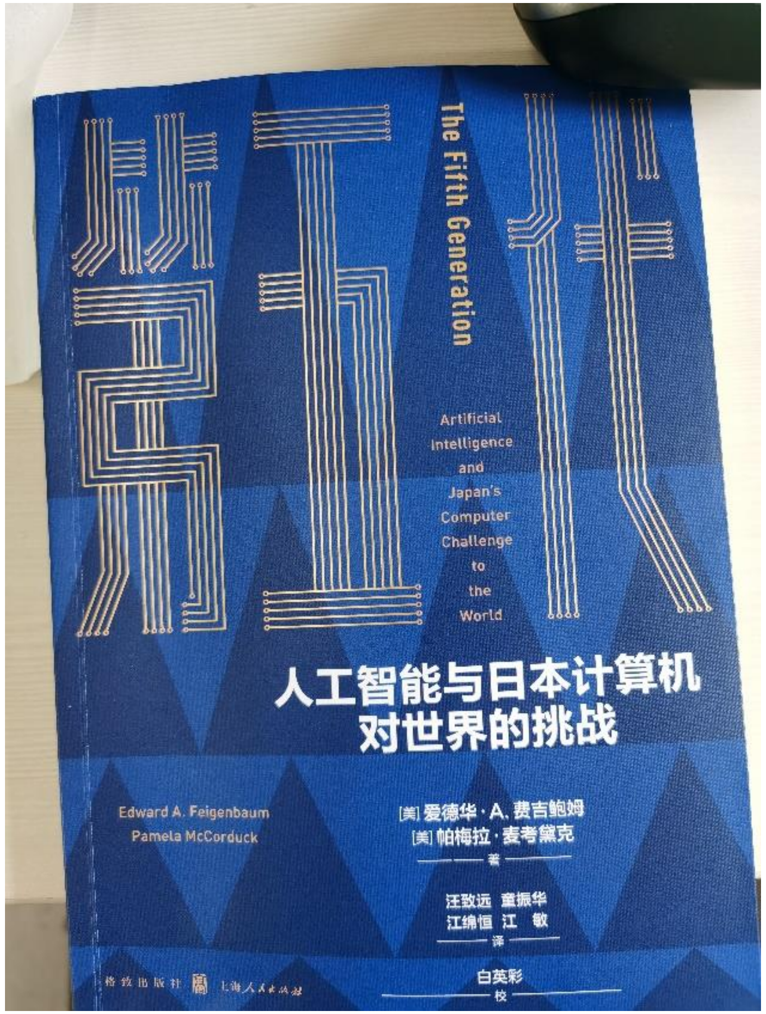
- Massive amounts of **data** have been generated in all different fields of natural science (e.g. astrophysics, genomics, material science)
  - The mainstream techniques of AI (e.g. deep learning) are **data-driven**
  - The overwhelming volume and complexity of scientific data can' t be handled by traditional computer software that lacks intelligence
- Lots of daily works in scientific research are repetitive routines, which should be done by machines instead of humans
- AI can help humans better understand the natural world:
  - Things too small: atoms
  - Processes too fast: protein folding
  - Phenomena too complex: cancer



立志 成才 报国 裕民

Expert systems

- An **expert system** is a computer program that uses AI methods to solve problems within a specialized domain that ordinarily requires human expertise (from Britannica)
  - Designed to solve complex problems by reasoning through bodies of knowledge, represented mainly as if-then rules (from Wikipedia)
  - Consisting of knowledge base (representing facts and rules) and inference engine (applying rules to known facts to deduce new facts)



DENDRAL

- **DENDRAL**, an early expert system, developed from 1965 by AI researcher **Edward Feigenbaum** and geneticist (Nobel Laureate) **Joshua Lederberg**, both at Stanford University
  - To identify the structures of chemical compounds
  - Starting from spectrographic data obtained from substance (e.g. a compound of carbon, hydrogen and nitrogen), it would hypothesize the substance' s molecular structure
  - Its performance rivaled that of human chemists at this task
- Many expert systems were derived from DENDRAL, such as:
  - **MYCIN**: Medical system for diagnosing blood disorders, first used in 1979
  - **MOLGEN**: A system that can plan or design laboratory experiments in molecular genetics (e.g. gene cloning)

Edward A. Feigenbaum (1936-), American Computer Scientist, Turing Award (1994)

Joshua Lederberg (1925-2008), American Geneticist, Nobel Prize in Physiology or Medicine (1958)

Cases of AI for Science (AI4S)

- Many success stories of using AI techniques (often deep learning) to make scientific discovery

Case studies	Time	Domain	AI Technique	Research team
AlphaFold 2 won CASP14 protein-folding competition	Dec. 2020	Structural biology	NLP modeling (Transformer)	Google DeepMind
Mankind has the first photo of a black hole	Apr. 2019	Astrophysics	CHIRP algorithm based on Bayesian statistical model	MIT etc.
The AI program called **Atom2Vec** recreated the periodic table of chemical elements	Jun. 2018	Chemistry, material science	Natural language processing (NLP), knowledge representation	Stanford University
An AI system speeded up the discovery of metallic glass by 200 times	May 2018	Material science & engineering	Supervised machine learning	SLAC, NIST, Northwest University

Cover image (Aug. 2021): DeepMind's AlphaFold2

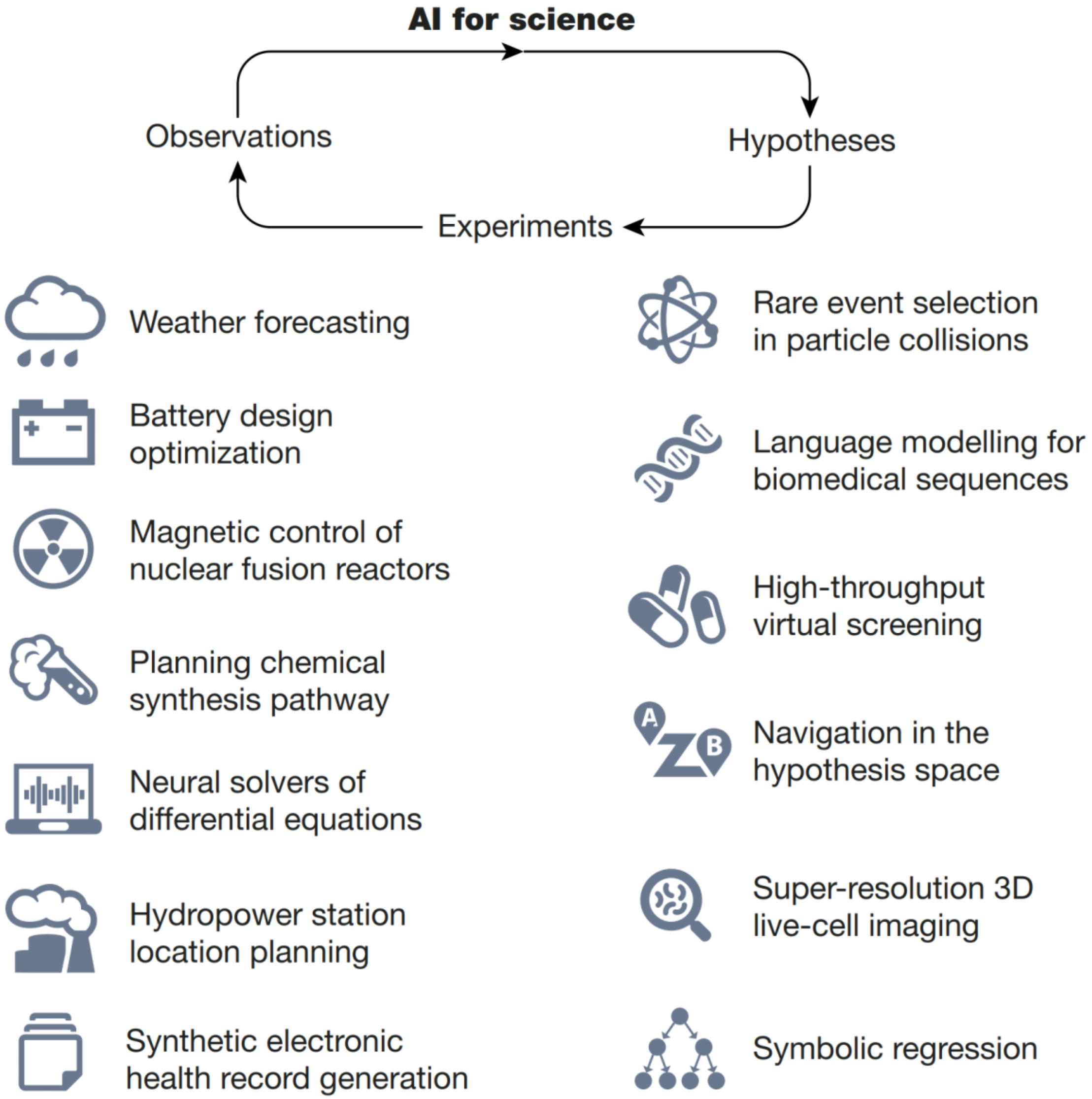
Outline

- Background
- **AI for scientific discovery pipelines**
    - Data collection and curation
    - Learning meaningful representations
    - AI-based generation of scientific hypotheses
    - AI-driven experimentation and simulation
    - Grant challenges
- Large language models (LLMs) for Science
- Grant challenges

AI is reshaping scientific discovery

上海科技大学
ShanghaiTech University

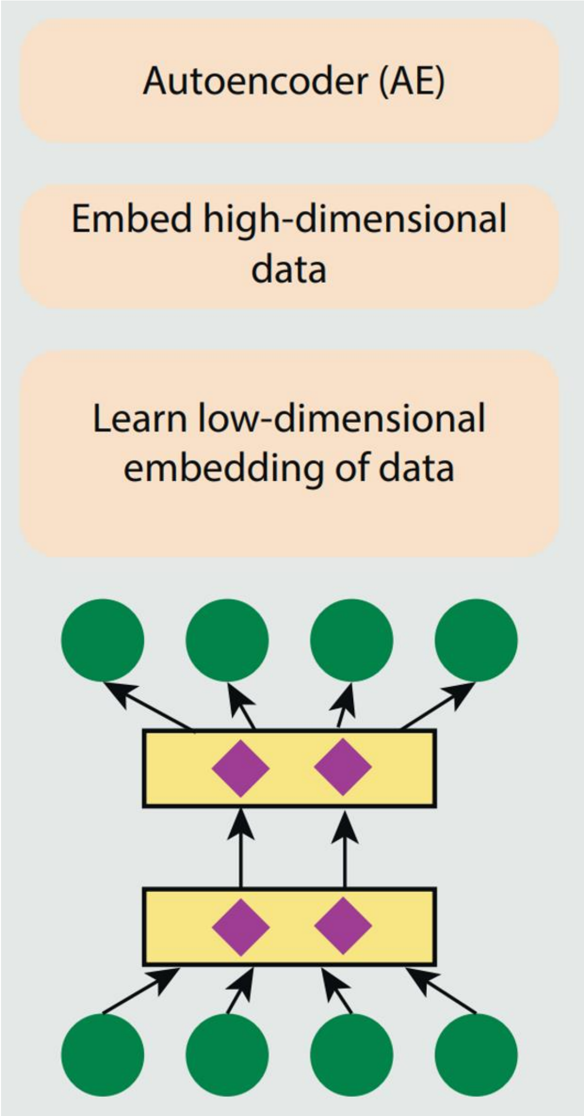
- Data collection & curation
  - Automation in data preparation
- Representation learning
  - Data-driven and knowledge-driven methods
- Hypothesis generation & search
  - Efficient strategies
- Experimentation & simulation
  - Automatic testing + efficient solver



立志 成才 报国 裕民

AI-aided data collection & curation

- **Data selection**
  - A particle collision experiment generates over 100 terabytes of data per second
  - Over 99.99% of raw data are background events, to be detected and discarded in real time
  - **Autoencoder** is used to model the background processes and detect rare events (unseen signals)
- **Data annotation**
  - Training supervised models requires data with annotated labels
  - Labeling scientific data is expensive
  - AI techniques to automate data labeling:
    - Pseudo-labeling
    - Label propagation
    - Active learning



AI-aided data collection & curation

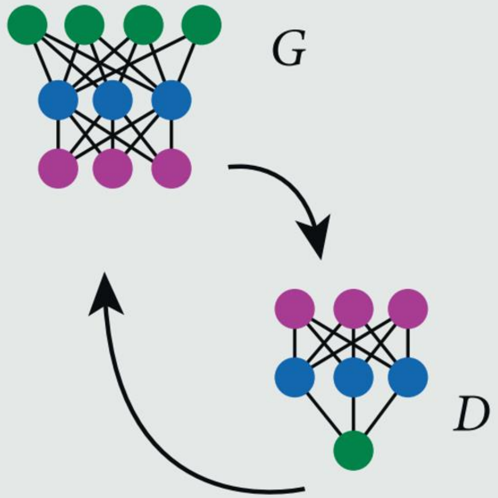
- **Data generation**
  - Variational Autoencoders (VAEs)
  - Generative Adversarial Networks (GANs)
  - Normalizing flows
  - Diffusion models
  - Probabilistic programming
- **Data refinements**
  - Transform poor spatiotemporally resolved measurements into high-quality, super-resolved and structure images
  - **Denoising**: differentiating relevant signals from noise, using denoising autoencoders (DAEs) or VAEs, etc.

Denoising autoencoders (DAE) are autoencoder models that learn low dimensional embeddings of noisy high dimensional data, i.e. inputs that differ by a small amount of noise give rise to a similar embedding vector.

Generative Adversarial Network (GAN)

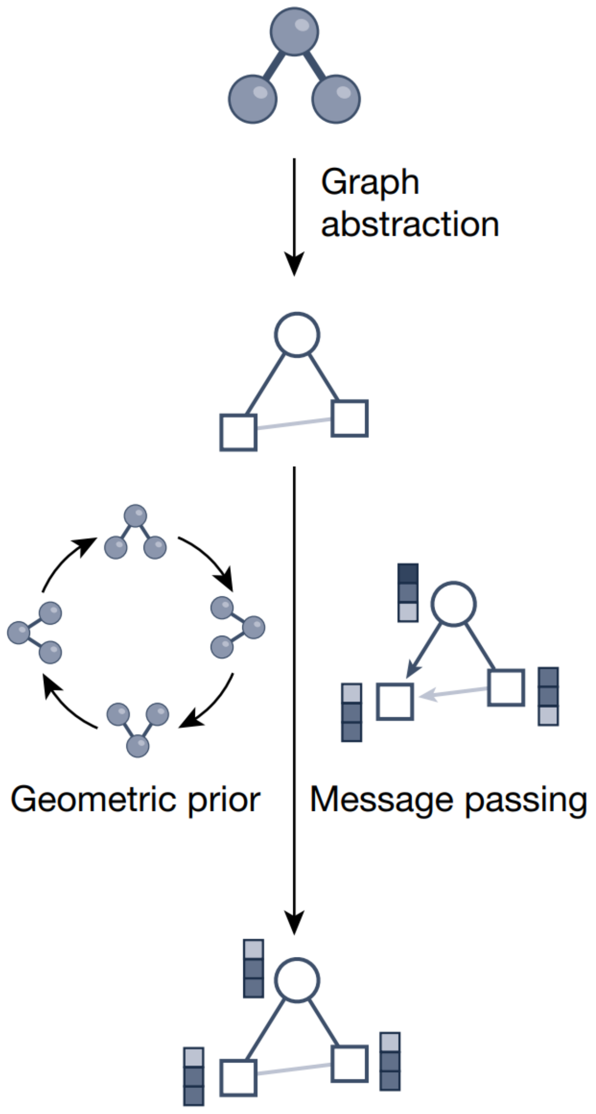
Generate samples from data distribution

Simultaneously train generator and discriminator



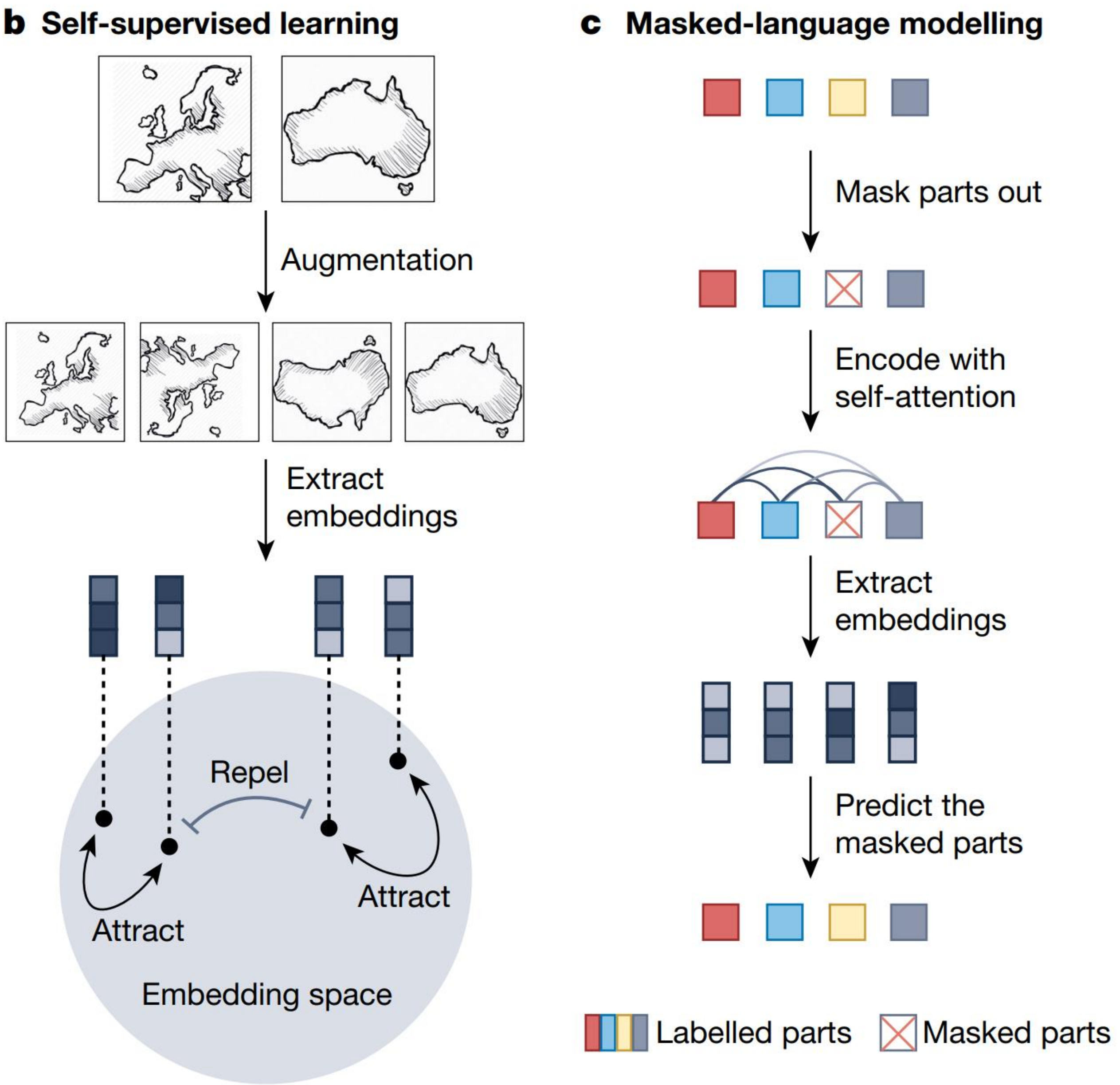
Learning representations of scientific data

- **Geometric deep learning**:
  - A field of machine learning that deals with geometric data, such as graphs or manifolds.
  - It typically preserves the invariance of geometric data under transformations and can be applied to 3D structures.
- **Geometric priors**, e.g. **symmetries**:
  - **Equivariance** characterizes the symmetry of functions. An equivariant function transforms the input equivalently under an operation from a group.
  - **Invariance**: A function is invariant to a group of transformations if the output remains unchanged when the inputs are transformed.



Wang, H. et al. Nature 2023.

Learning representations of scientific data

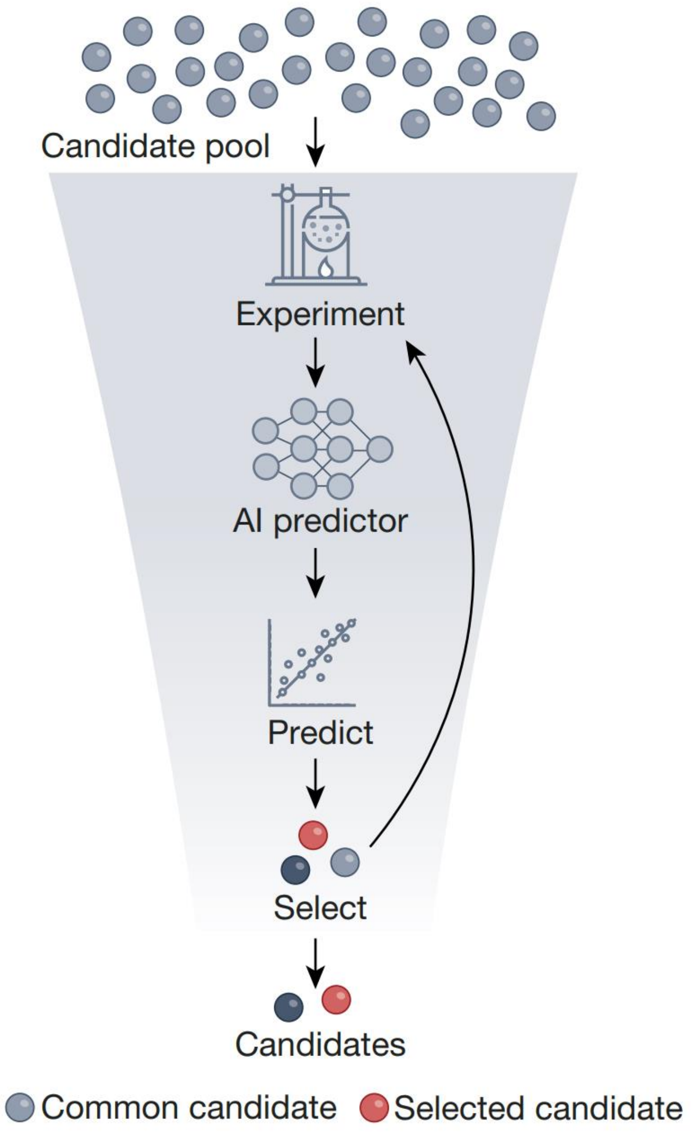


- **Self-supervised learning**
  - Aims to learn meaningful features without needing labelled data:
  - **Generative learning**: predicts a part of the raw data based on the rest
  - **Contrastive learning**: defines positive and negative views of the input, and then aligns the positives and separates the negatives
- **Masked-language modeling**
  - Captures the semantics of sequential data, e.g. natural language and biological sequences
  - Based on Transformers

Wang, H. et al. Nature 2023.

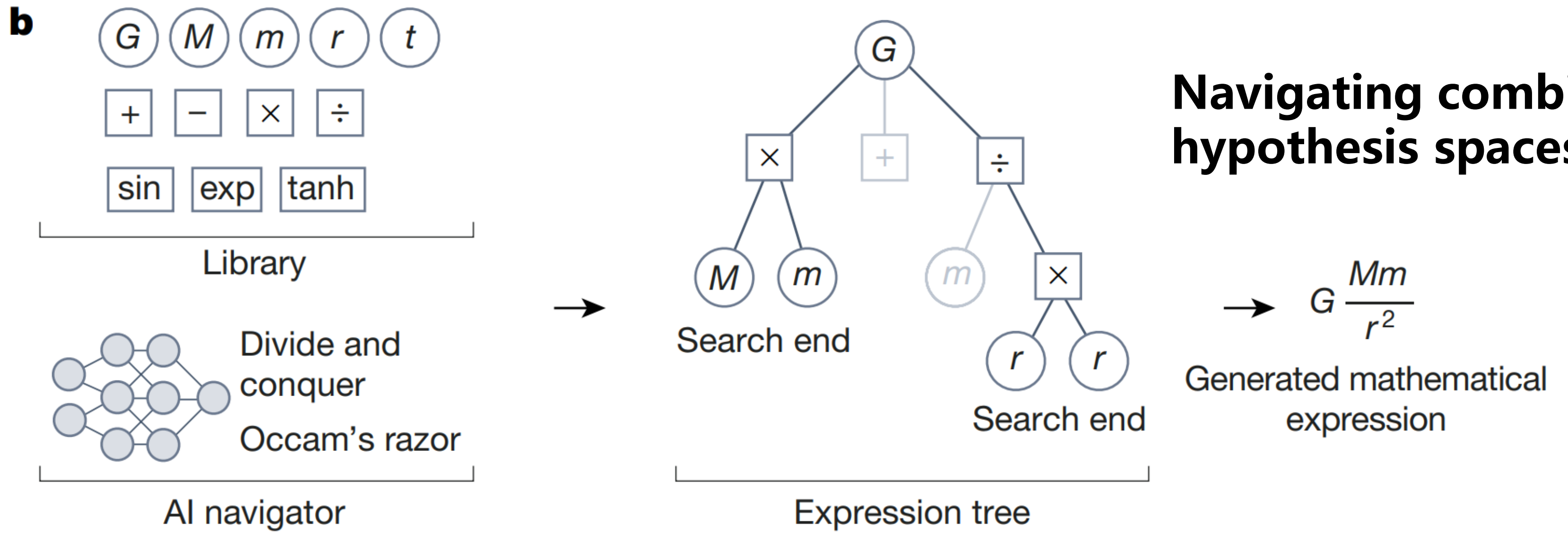
AI-based generation of scientific hypotheses

- Black-box predictors of scientific hypotheses
  - High-throughput screening uses AI predictors to select a small number of screened objects with desirable properties
  - The predictors can be pre-trained by self-supervised learning on many unscreened objects
  - Then, fine-tune the predictors on labelled data of screened objects
  - Benefits: Reduce the search space, making it cheaper and faster to identify chemical compounds, materials and biomolecules

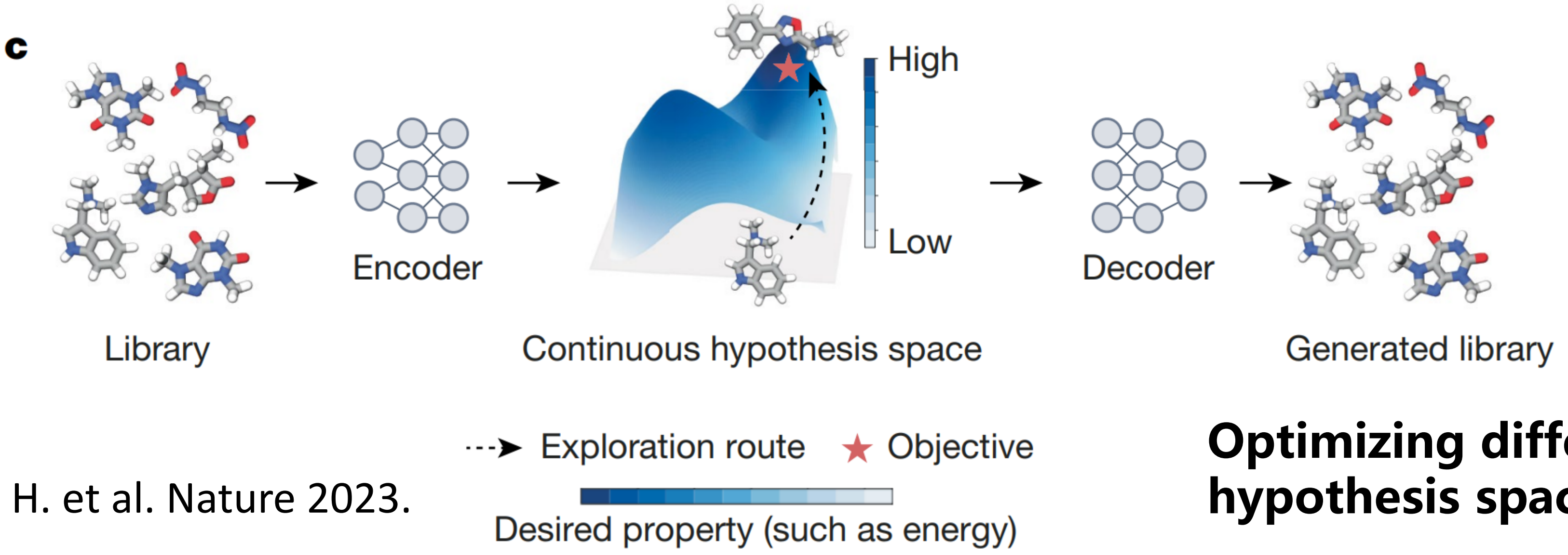


Wang, H. et al. Nature 2023.

AI-based generation of scientific hypotheses



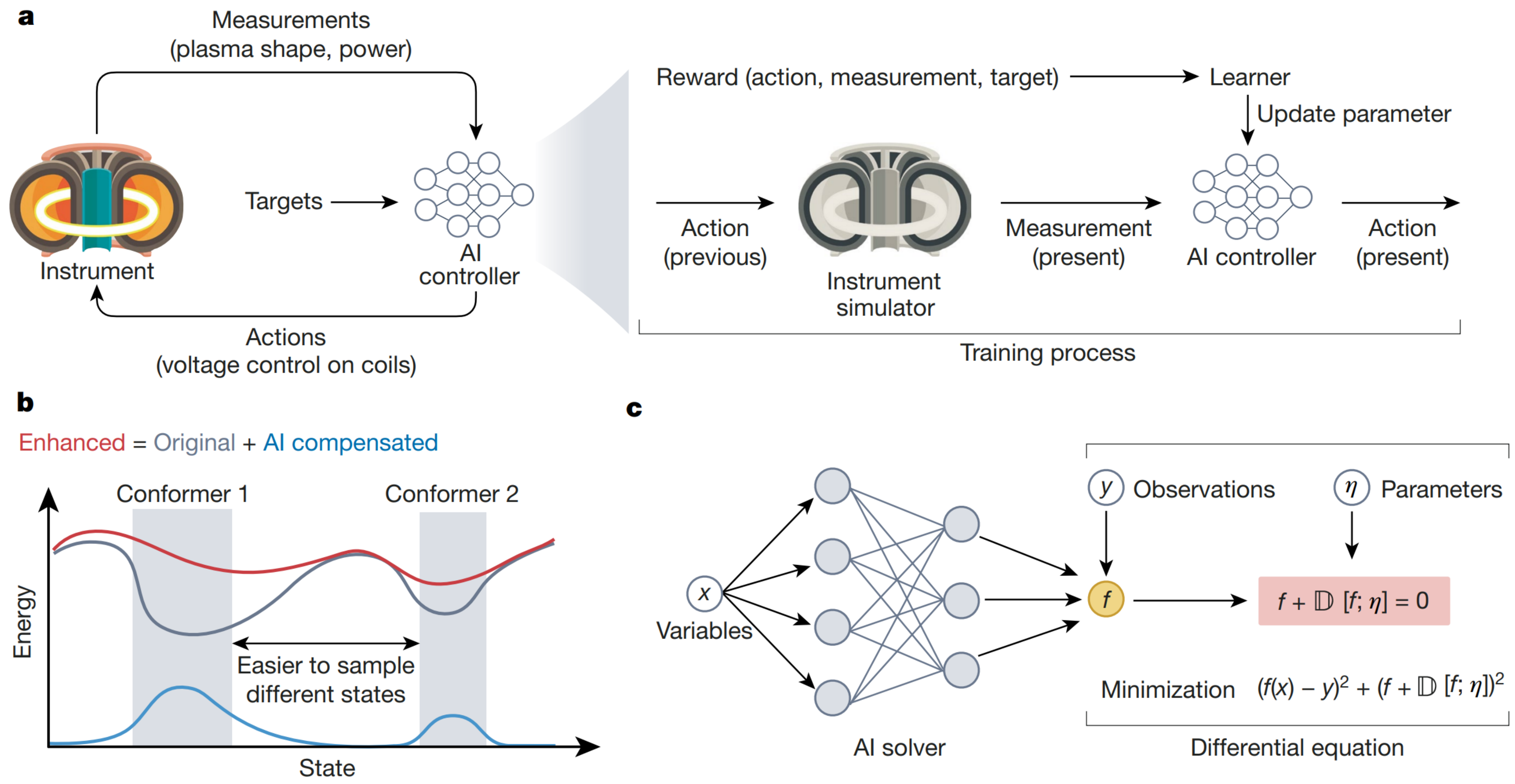
Navigating combinatorial hypothesis spaces



Optimizing differentiable hypothesis spaces

Wang, H. et al. Nature 2023.

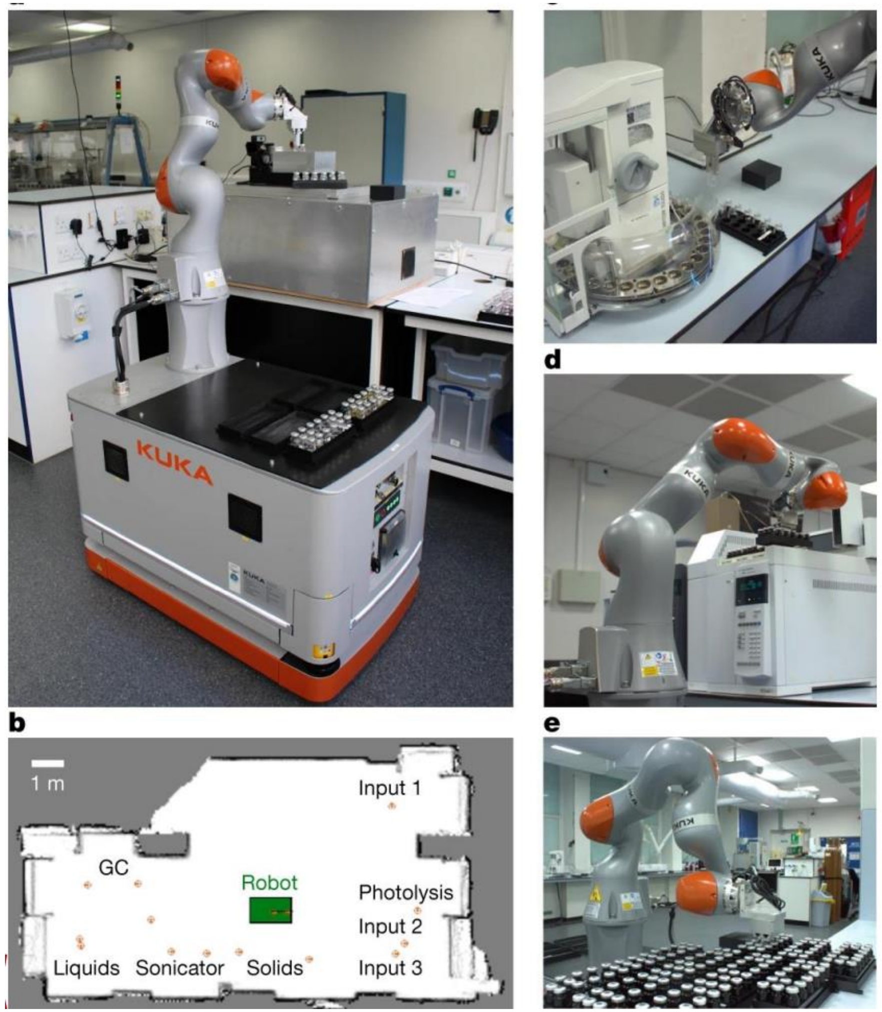
AI-driven experimentation and simulation



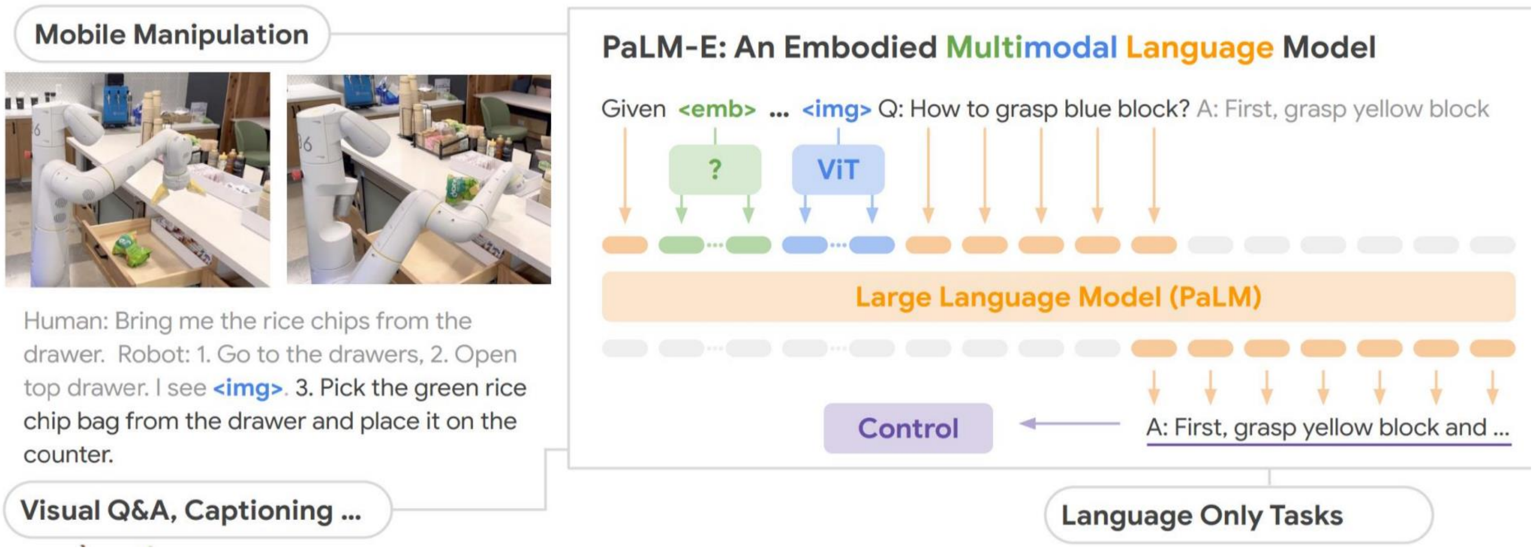
Wang, H. et al. Nature 2023.

AI-driven experimentation and simulation

- Self-driving laboratories
- Automation of scientific workflows



Google PaLM-E



Burger, B., Maffettone, P.M., Gusev, V.V. et al. A mobile robotic chemist. Nature 583, 237–241 (2020)

Outline

- Background
- AI for scientific discovery pipelines
  - Data collection and curation
  - Learning meaningful representations
  - AI-based generation of scientific hypotheses
  - AI-driven experimentation and simulation
  - Grant challenges
- **Large language models (LLMs) for Science**
- Grant challenges

LLM for Science

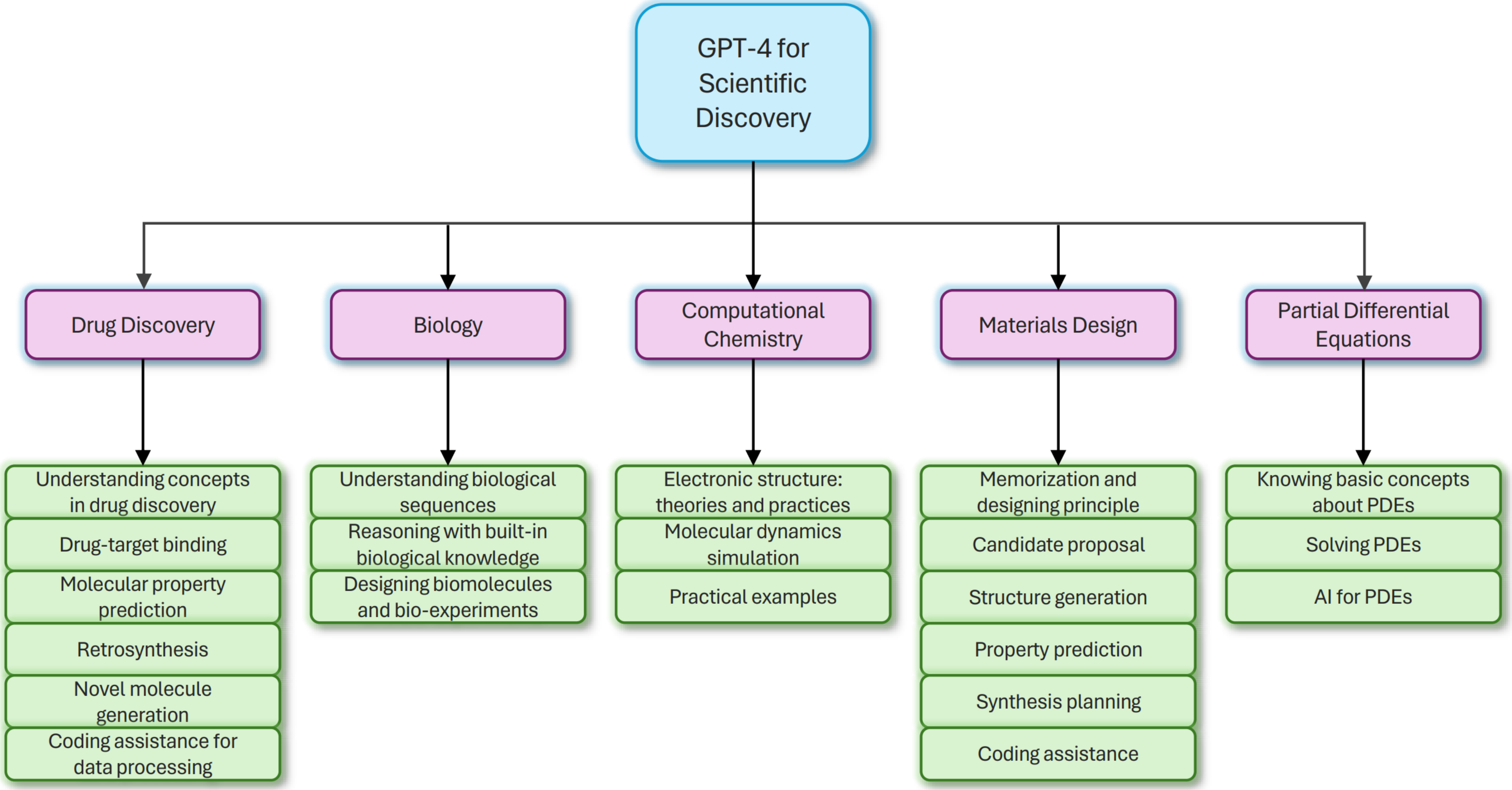
ChatGPT入选了《Nature》2023年度十大人物，是有史以来首个非人类实体入选。

包括ChatGPT在内的AI工具被《Nature》评为2024年最值得关注的科学事件。

大语言模型(LLM)有非常强大的能力，可以帮助我们在生化环材领域更高效的科研

LLM for Science

微软研究人员深入研究了 LLM 在科学发现/研究背景下的表现，重点关注最先进的语言模型 GPT-4。研究涵盖多个科学领域，包括药物发现、生物学、计算化学、材料设计和偏微分方程。

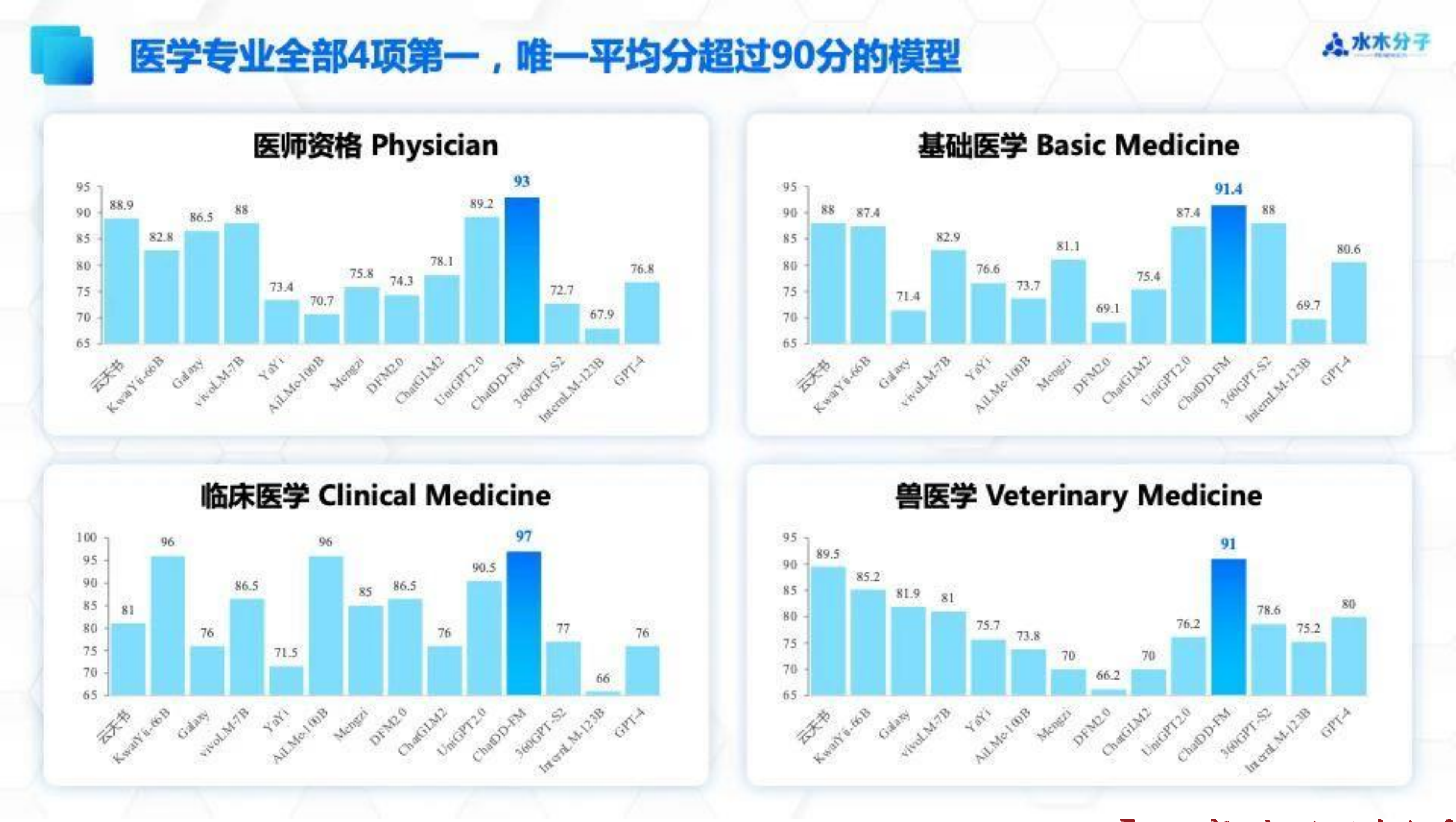


AI4Science M R, Quantum M A. The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4[J]. arXiv preprint arXiv:2311.07361, 2023.

LLM for Science

清华系初创团队水木分子开发了新一代对话式药物研发助手 ChatDD (Drug Design)，覆盖药物立项、临床前研究、临床试验的各阶段，作为制药专家的得力 AI 助手，提升药物研发效率。

ChatDD-FM 100B在C-Eval评测中达到医学专业全部 4 项考试第一



LLM for Science

BioMap 百图生科开发了生命科学大模型驱动的 AIGP (AI Generated Protein) 平台

AIGP平台的能力

BioMap 百图生科

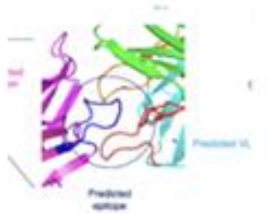
F2P - Function to Protein

Ideal SHAPE, Physiochemical PROPERTY and FUNCTION

- 新功能蛋白设计
- 新结构蛋白设计
- 新酶蛋白设计
- 蛋白质参数优化能力

P2P – Protein to Protein

Ideal BINDING SITE, AFFINITY and SPECIFICITY



- 高亲和力设计
- 高精准表位设计
- 高特异性设计
- 高序列差异性设计

C2P – Cell to Protein

Read CELL STATUS and predict modulating targets and design proteins to reprogram the cell and the DISEASE

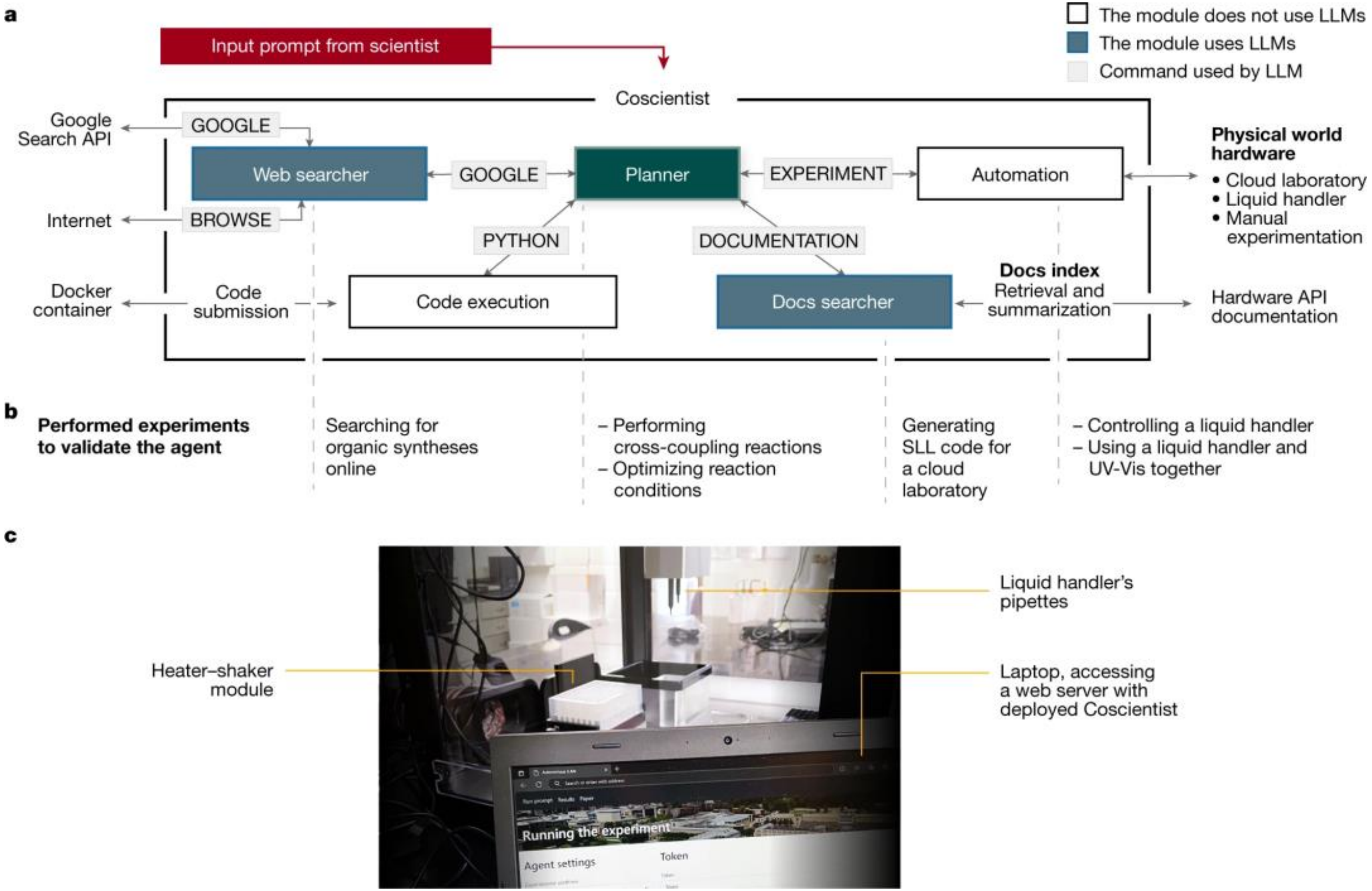
- 细胞调控靶点预测
- 药物组合功能预测
- 组织特异性靶点预测
- 细胞分类预测

https://mp.weixin.qq.com/s?__biz=MzI3MjM3ODk0NQ==&mid=2247494693&idx=1&sn=7fcc0bc203882a3dd43efd8c961b605d&chksm=eb31d10bdc46581d085d7add881cc4929abd0d1382012d65b068eac4a292ddbba65319b7f1a7&scene=21#wechat_redirect

LLM for Science

卡内基梅隆大学的研究团队提出了一种基于 GPT-4 的智能代理（Coscientist），用一个简单的语言提示就可以执行整个实验过程。能够自主设计、规划和执行复杂的科学实验。

Coscientist 可以设计、编码和执行多种反应，在湿实验中使用其机器人设备制造包括扑热息痛和阿司匹林在内的化合物。

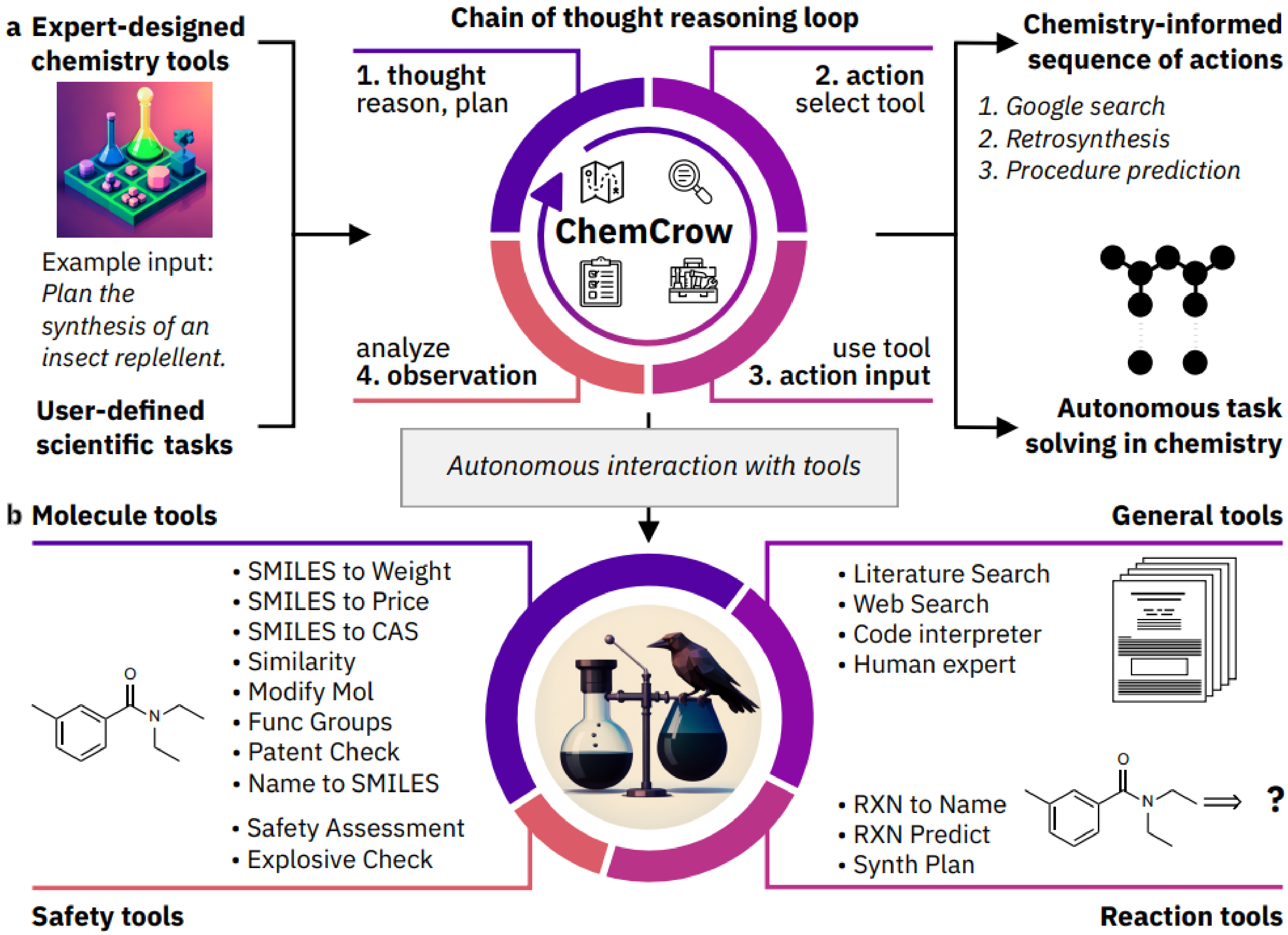


Boiko D A, MacKnight R, Kline B, et al. Autonomous chemical research with large language models. Nature, 2023, 624(7992): 570-578

立志成才 报国裕民

LLM for Science

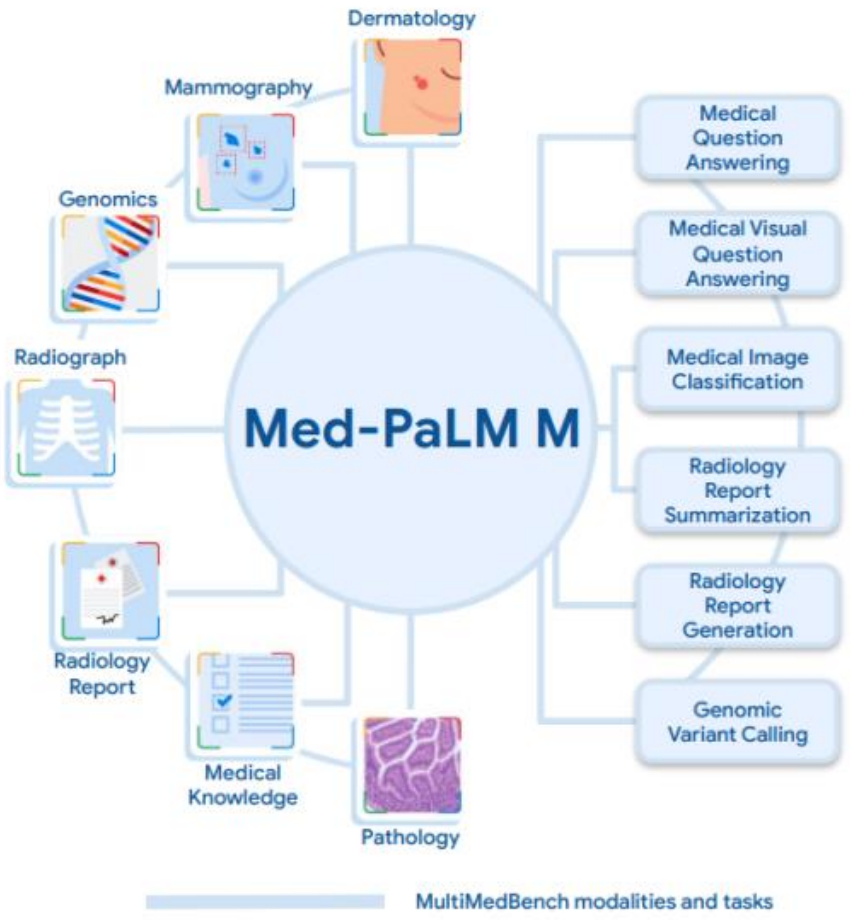
ChemCrow是一个由大语言模型驱动的化学引擎，它集成了许多专家设计的化学工具，可以利用自然语言交互的方式实现许多药物设计和材料领域的化学任务, 如反应预测、逆合成规划、分子特性预测、从头分子生成、材料设计等



Bran, Andres M., et al. "ChemCrow: Augmenting large-language models with chemistry tools." arXiv preprint arXiv:2304.05376 (2023).

LLM for Science

在医疗领域，谷歌的研究人员提出了用于评估医学大模型的基准MultiMedQA，并开发了准确率能够与人类医生相当的大语言模型Med-PaLM M



Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI[J]. arXiv preprint arXiv:2307.14334, 2023.

Grand challenges

- Practical considerations
  - Incomplete and biased data, privacy and safety concerns
  - Challenge for human-in-the-loop automated scientific workflows
  - Complex software and hardware engineering
  - Issue of reproducibility
- Algorithmic innovations
  - Out-of-distribution generalization: causality in AI models
  - How to integrate multimodal observations
  - How to incorporate scientific knowledge into data-driven models
  - Deep learning models are black-boxes: need interpretability
- Conduct of science and scientific enterprise
  - Big data, big models: a large energy footprint and high computational cost
  - AI could rival, surpass and replace humans for routine laboratory work
  - Misuse and security risks of AI: need ethics review and responsible implementation

Summary and discussion

- Today, we have learned:
  - Why need AI (esp. deep learning) for Science?
  - AI can contribution in the whole pipeline of Scientific Discovery
    - Data collection and curation
    - Learning meaningful representations of scientific data
    - Generation of scientific hypotheses
    - AI-driven experimentation and simulation
  - However, there are still many challenges in AI for Science
- Open-end discussions
  - How should the Education in Science be changed to adapt to the age of AI?
  - What are key differences between Science and other applications of AI?
  - How can AI4Science contribute to the economy and social well-being?

References

- Hanchen Wang et al. "Scientific discovery in the age of artificial intelligence", Nature, Vol. 620, pp. 47-60, 2023.
- Microsoft Research AI4Science, Microsoft Azure Quantum. "The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4". arXiv preprint arXiv:2311.07361, 2023.

Related courses:

- CS177 “Bioinformatics: Software Development and Applications” (Spring Semester), 4 units: 3 for lectures, 1 for projects
- CS286 “AI for Science and Engineering” (Fall Semester), 4 units: 2 for lectures, 2 for projects

Thanks!