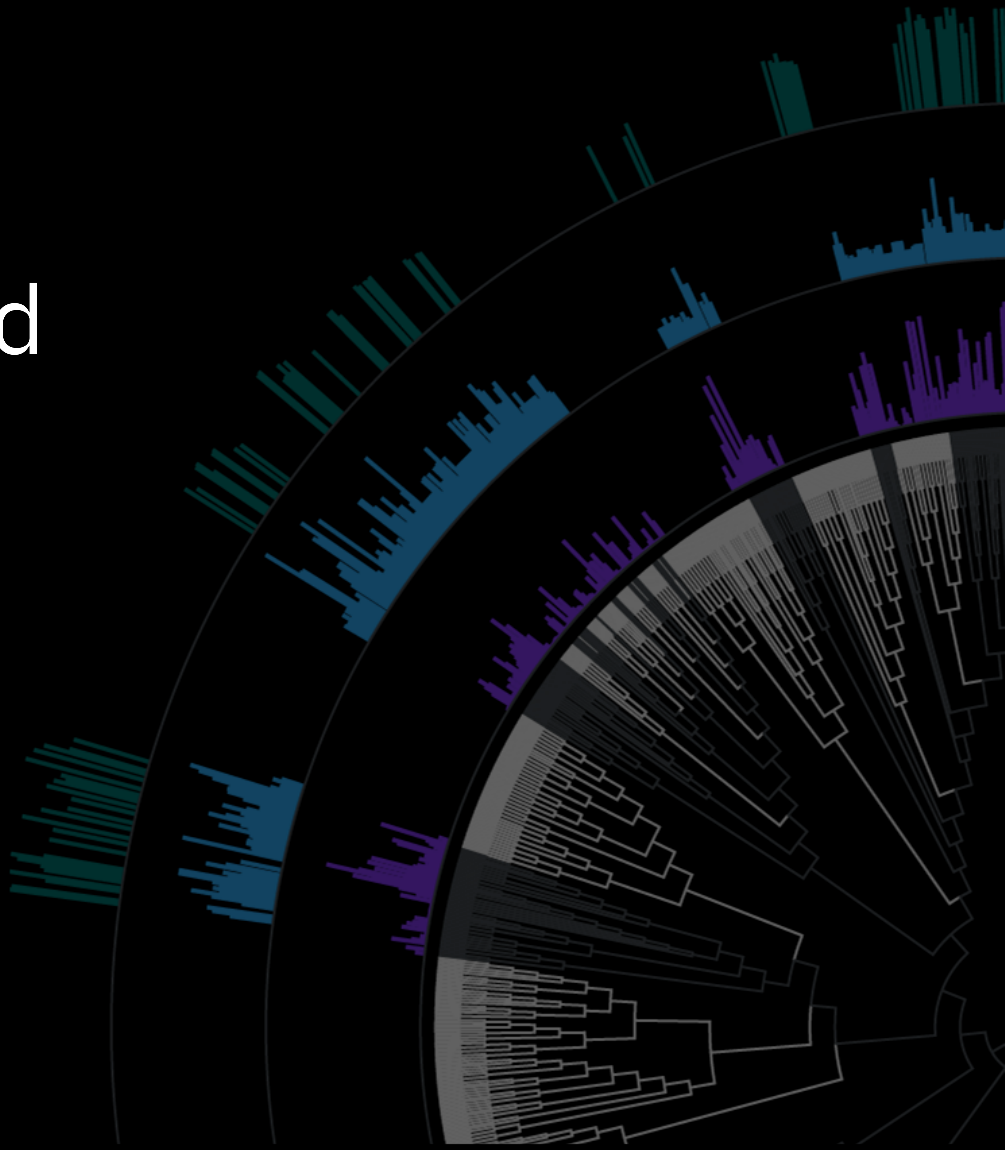
IBM Research

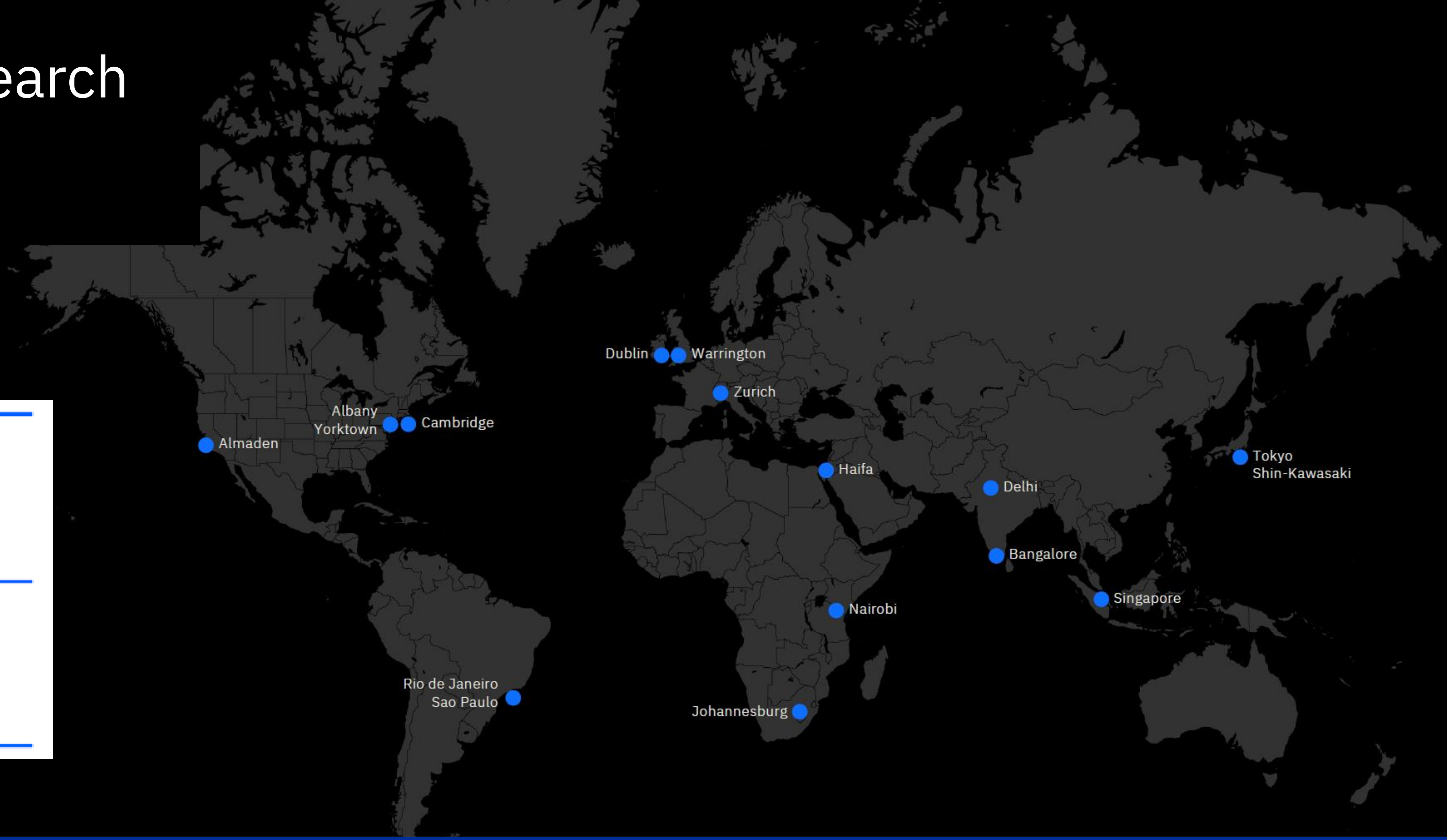
Foundation Models for Accelerated Discovery

John R. Smith, IBM Fellow, Accelerated Discovery
IBM T. J. Watson Research Center, Yorktown Heights, NY USA
jsmith@us.ibm.com

October 2022



IBM Research



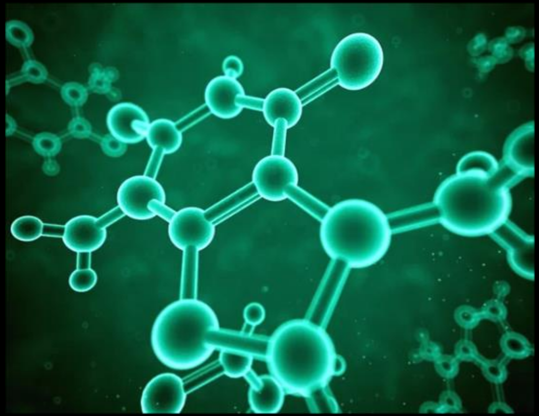
3,000
researchers

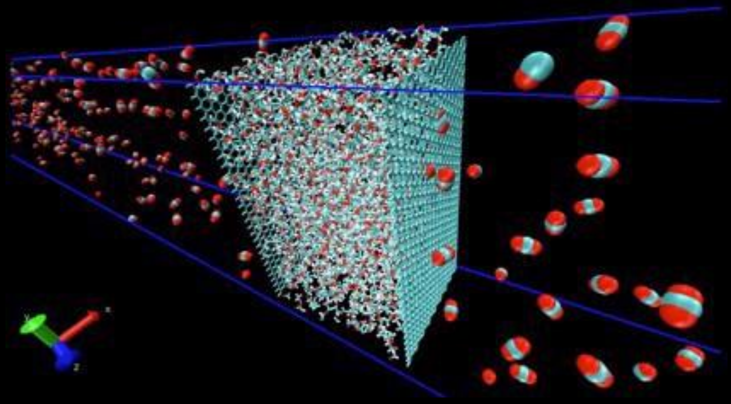
100s
of disciplines

Cloud Computing | Artificial Intelligence | Quantum Computing | Semiconductors and Systems | Security and Cryptography | Physical Sciences | Mathematical Sciences | Life Sciences | Computer Science

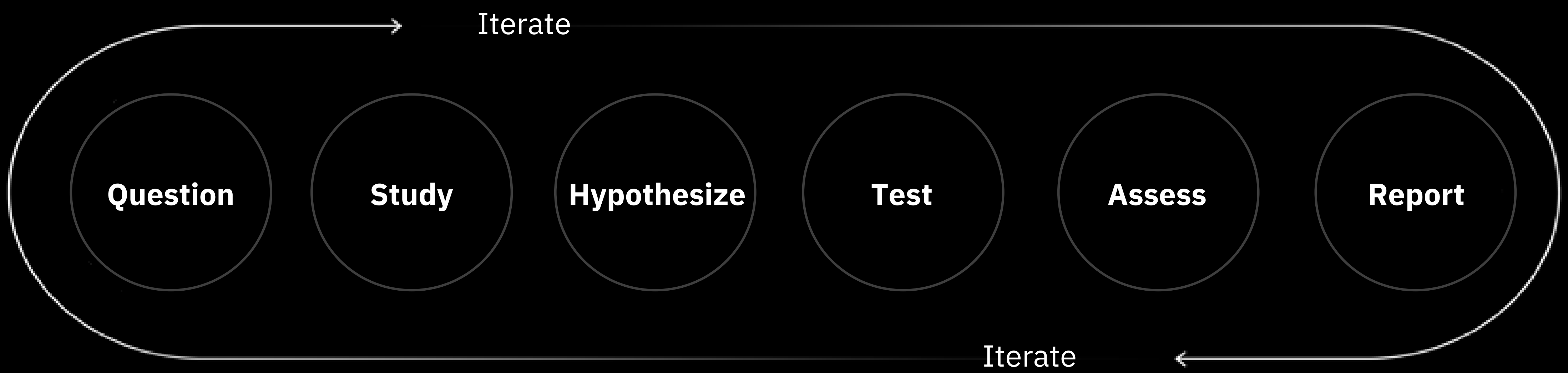
Accelerated Discovery

How do we discover solutions to complex problems?





The scientific method has been our best tool for discovery...



The scientific method has evolved over time

New Era

1st Paradigm	2nd Paradigm	3rd Paradigm	4th Paradigm	
Empirical Science	**Theoretical Science**	**Computational Science**	**Big data-driven Science**	**Accelerated Discovery**
• Observations • Experimentation	• Scientific laws • Physics • Biology • Chemistry	• Simulations • Molecular dynamics • Mechanistic models	• Big data • Patterns • Anomalies • Visualization	• Data + AI/ML • Hybrid Cloud • Quantum Computing • Automation & autonomy
Pre-Renaissance	~1600s	~1950	~2000	2020s

We are entering a new era of discovery

New Era

1st Paradigm

Empirical Science

- Observations
- Experimentation

Pre-Renaissance

2nd Paradigm

Theoretical Science

- Scientific laws
- Physics
- Biology
- Chemistry

~1600s

3rd Paradigm

Computational Science

- Simulations
- Molecular dynamics
- Mechanistic models

~1950

4th Paradigm

Big data-driven Science

- Big data
- Patterns
- Anomalies
- Visualization

~2000

Accelerated Discovery

- Data + AI/ML
- Hybrid Cloud
- Quantum Computing
- Automation & autonomy

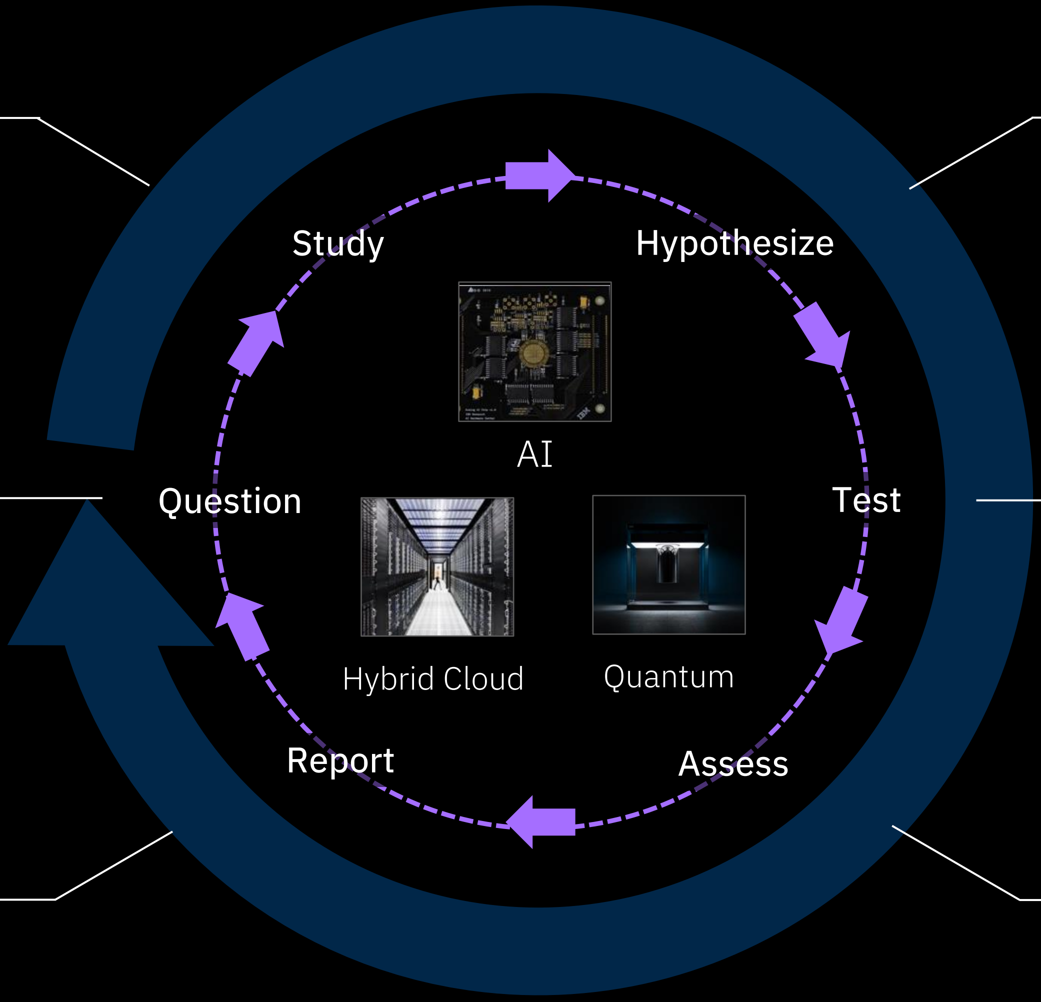
2020s

Automating and Accelerating Scientific Discovery

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 questions



Generative models propose new hypotheses and automatically explore vast discovery spaces

Robotic labs automate experimentation and bridge digital models and physical testing

AI surrogate models integrated with simulation fill in missing information

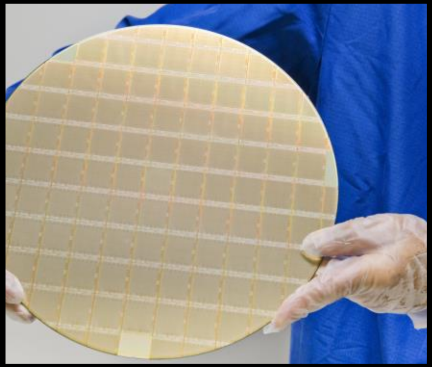
Examples in Materials Discovery

It takes roughly 10 years and upwards of $10–$100 million to discover one new material.

We aim to cut down both years and cost by 90%.

Example Accelerated Discovery challenges and results for materials

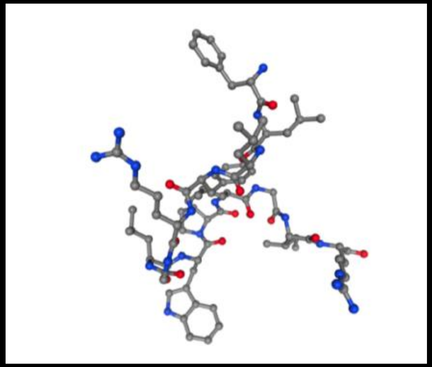
Sustainable semiconductors



Discovery and synthesis of a new photoacid generator molecule in less than 1 year

https://research.ibm.com/science/photoresist

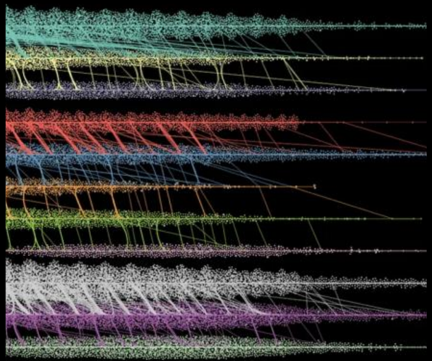
Therapeutics



Discovery and validation of two new antimicrobial compounds in 48 days instead of 2-4 years

https://research.ibm.com/publications/accelerated-antimicrobial-discovery-via-deep-generative-models-and-molecular-dynamics-simulations

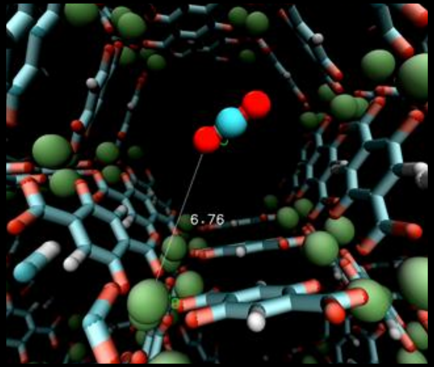
Biomarkers



Discovered biomarkers and new disease trajectories for Type 1 diabetes

https://research.ibm.com/blog/ai-predicting-onset-of-type-1-diabetes

Climate & Sustainability



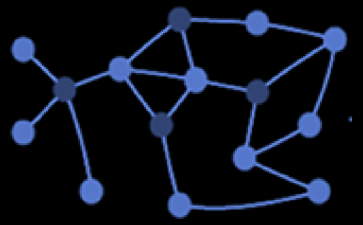
Discovery of 500 molecular candidates for membranes to better separate CO_2 from flue gas

https://research.ibm.com/blog/accelerating-materials-discovery

Discovery Technology Foundational Building Blocks

Accelerate scientific discovery through consumable "general purpose" discovery tools and platforms

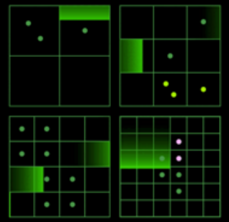
Deep Search



https://www.research.ibm.com/covid19/deep-search

Simulation (ST4SD)

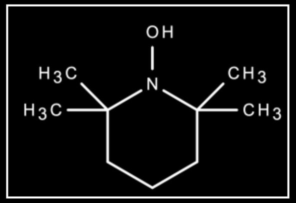
(Simulation Toolkit for Scientific Discovery)



https://pages.github.ibm.com/st4sd/overview/

Generative Models (GT4SD)

(Generative Toolkit for Scientific Discovery)

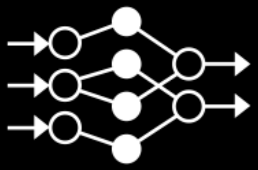


https://github.com/GT4SD

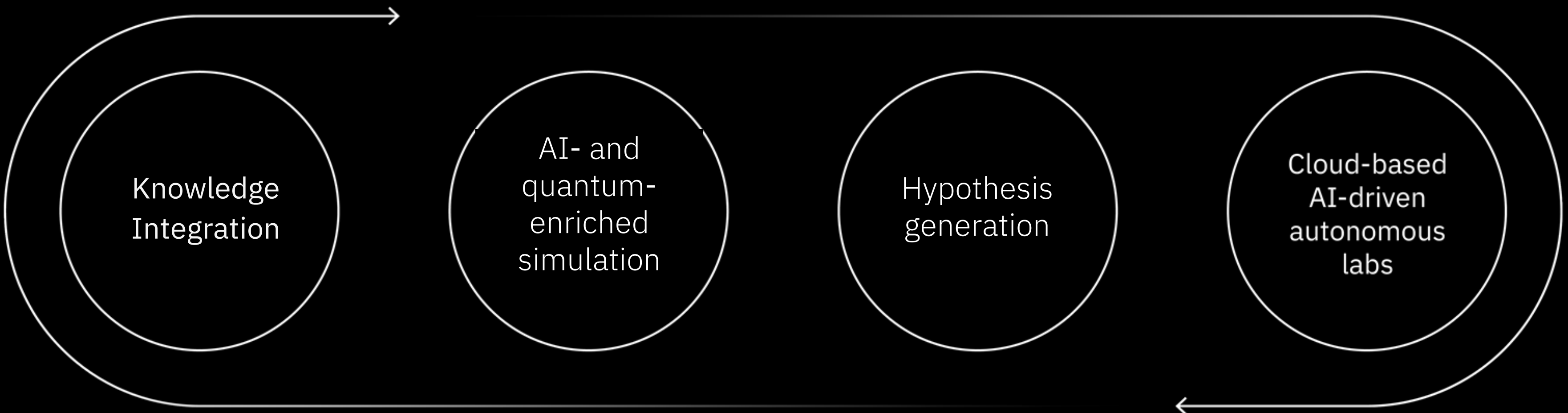
Synthesis & Testing (RXN)

https://rxn.res.ibm.com/rxn/robo-rxn

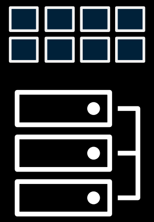
Foundation Models



"c1ccccc1 (C (=O) O"



Discovery Optimization

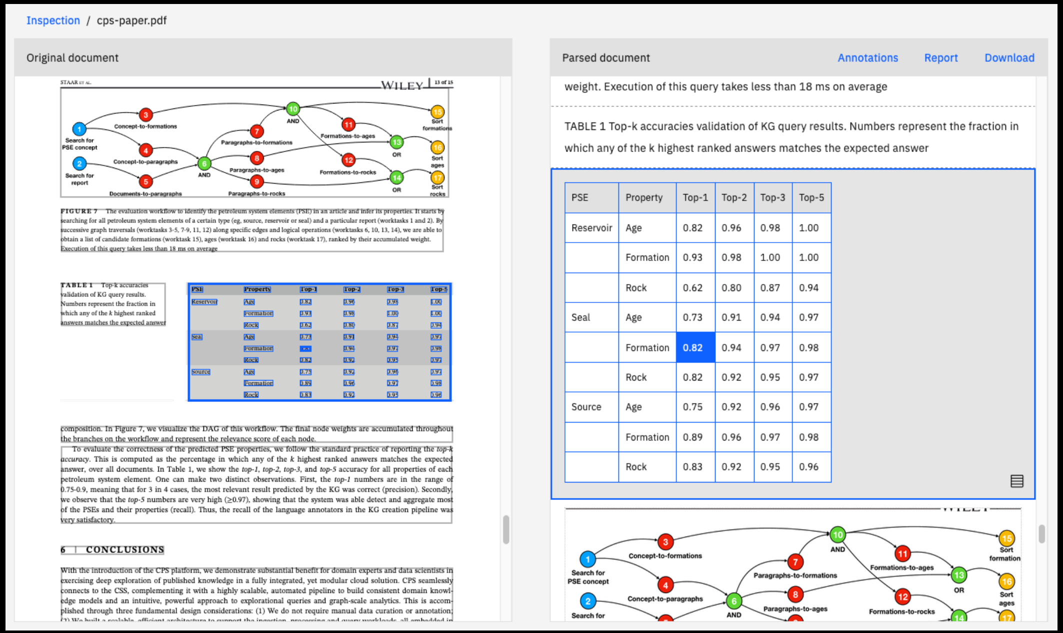


Accelerated Discovery Workflows

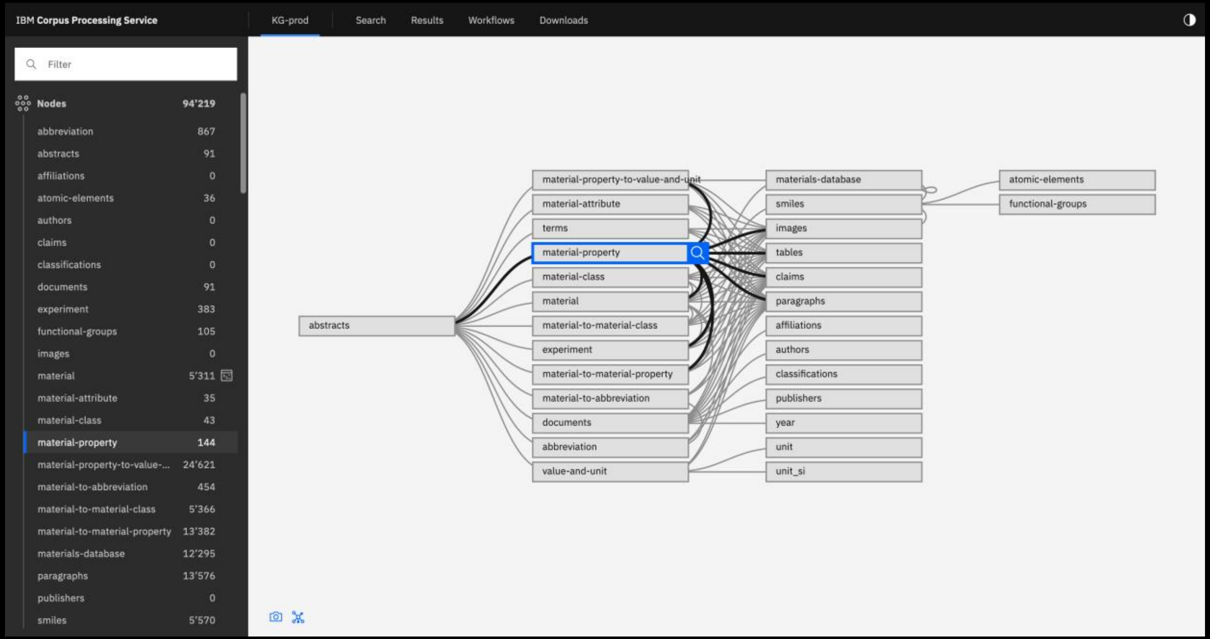
Deep Search: Structuring and Reasoning with Scientific Knowledge

Large-scale curation and insight-extraction from unstructured multi-modal documents

Deep Search can parse large collections of scientific articles



Discovers entities and allows search via knowledge graphs



Parse → Interpret → Index → Integrate

Deep Search Experience
https://deepsearch-experience.res.ibm.com/

Deep Search Toolkit
https://github.com/DS4SD/deepsearch-toolkit

Fork 5 | Star 47

Simulation Toolkit for Scientific Discovery (ST4SD)

Filling knowledge gaps with AI-enriched modeling and simulation

ST4SD runtime

A runtime for simulation workflows that includes memoization and surrogate support as well as pluggable HPC backends

AI surrogates

Includes and enables AI surrogate and hybrid workflows including configurable strategies for performance monitoring and risk management

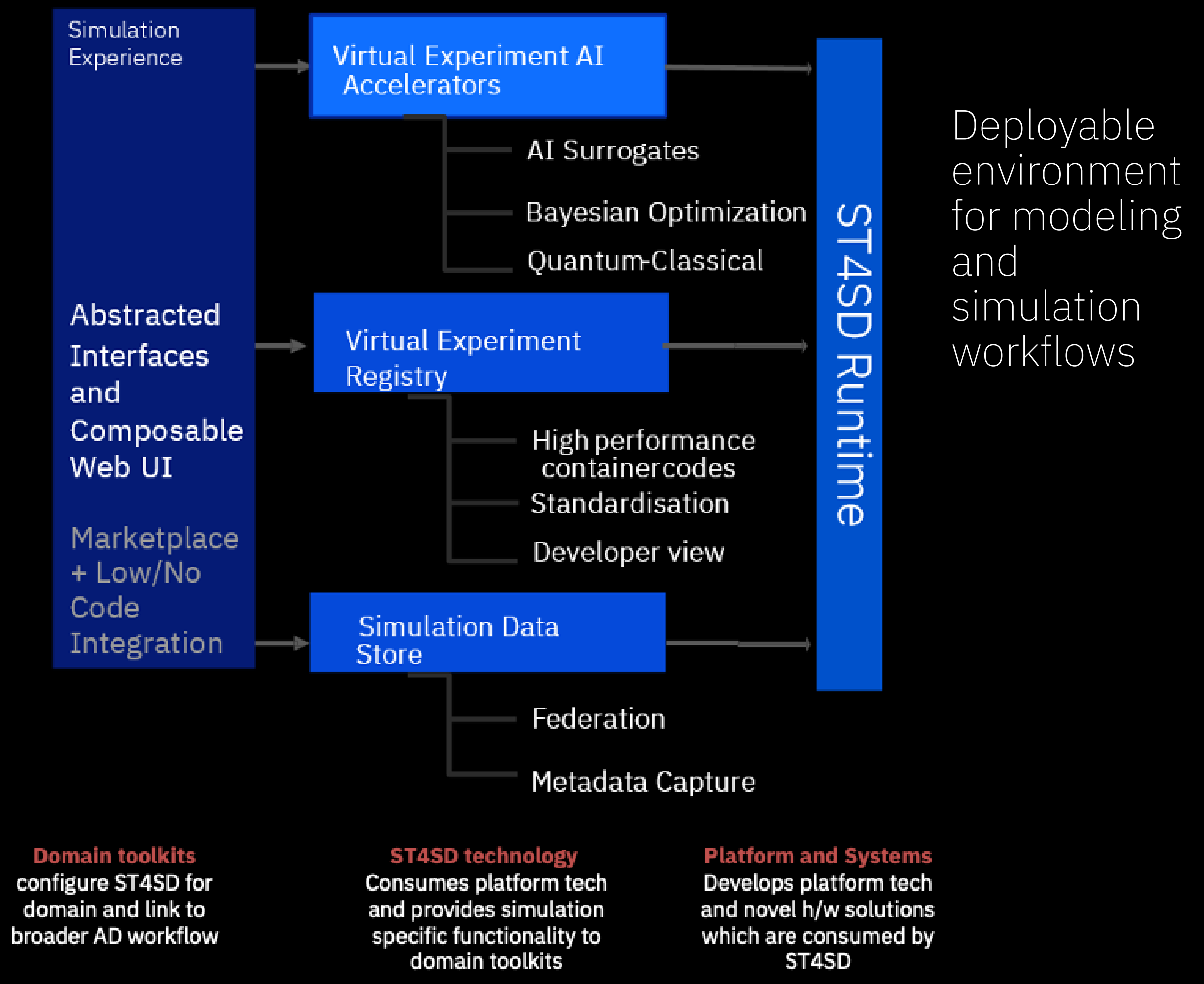
Virtual experiment registry

Includes pre-built workflows and building blocks for composing and hosting virtual experiments

ST4SD experience

User interactive tool with pre-packaged experiments

ST4SD Toolkit



https://github.com/GT4SD

113 stars 34 forks

Generative Toolkit for Scientific Discovery (GT4SD)

Open-source library to accelerate hypothesis generation in scientific discovery

pypi package 0.50.0 | Running tests: style, pytests and entry-points passing | License MIT
code style black | contributions welcome | website live | downloads 18k | downloads/month 2k
launch binder | DOI 10.5281/zenodo.7073764 | Award 2022 IEEE Open Software Services Award

Study → Hypothesize → Test

GT4SD/gt4sd-core

Applications include hypothesis generation for inverse design and discovery of materials

1. Train generative models

```
gt4sd-trainer  --training_pipeline_name paccmann-vae-trainer --epochs 25
```

2. Create inference pipelines

```
gt4sd-saving --training_pipeline_name paccmann-vae-trainer --model_path
```

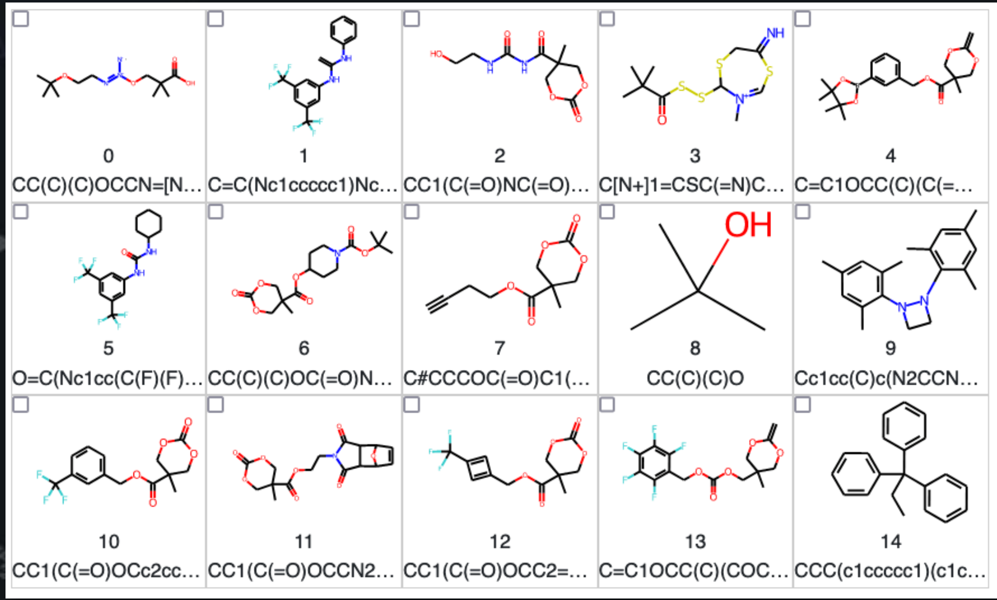
3. Run inference pipelines

```
gt4sd-inference --algorithm_name PaccMannGP --algorithm_application Pacc
```

4. Share your models with the community

```
gt4sd-upload --training_pipeline_name paccmann-vae-trainer --model_path
```

Example molecules generated using GT4SD



IBM RoboRXN – intelligent lab automation for the cloud

AI + Hybrid Cloud + Robotic Labs for automated synthesis prediction and execution

35,000
Users via cloud

9+ Million
Reaction predictions

7
Industrial partners

Foundation Models

The same AI breakthroughs happening in language are changing scientific discovery

Foundation models are powering new capabilities

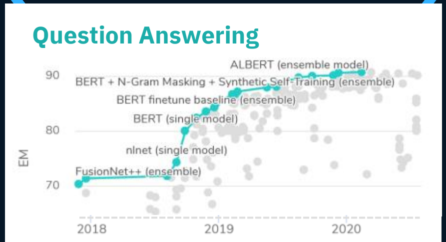
Natural Language Processing (NLP)

Transformers

AI

Data → AI Models

Letters → Words → Sentences → Language



Improved Language Tasks

Accelerated Chemistry

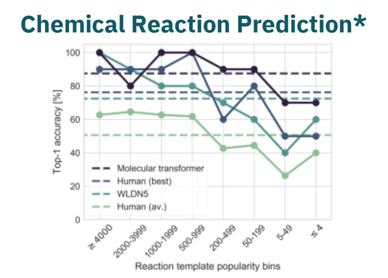
Transformers

AI

Data → AI Models

Atoms → Molecules → Properties/Reactions → Chemistry

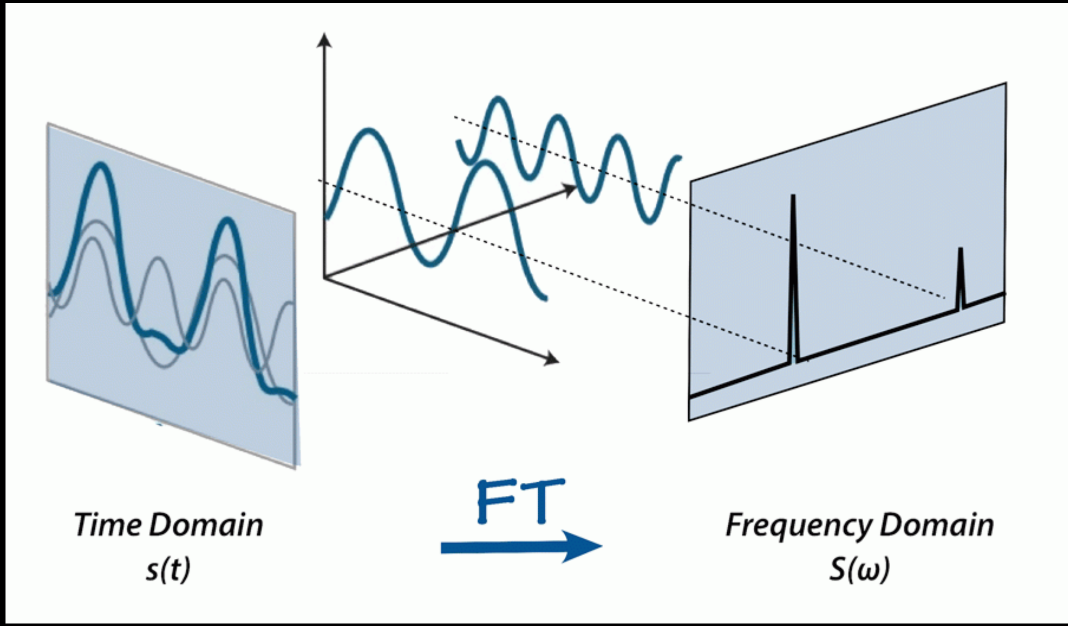
IBM RXN



Improved Chemistry Tasks

Well chosen representations can simplify complex problems

Example from the world of signal processing

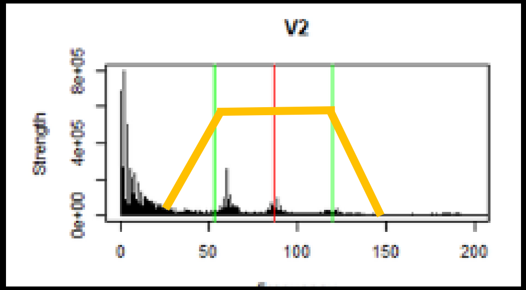


Data represented in the frequency domain can reduce complexity for filtering and denoising

Time-domain

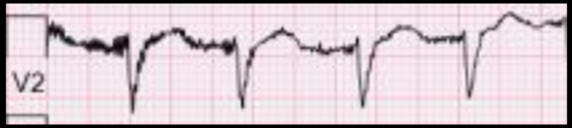
FT →

Frequency-domain



FT^{-1} →

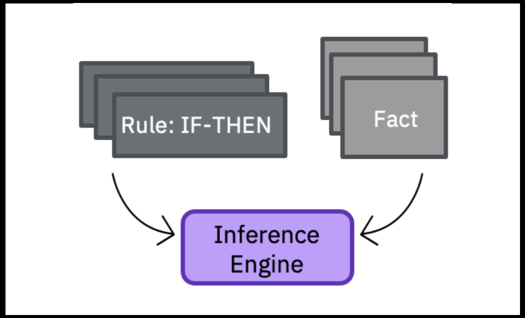
Time-domain



Story of AI has also been a story of data representations

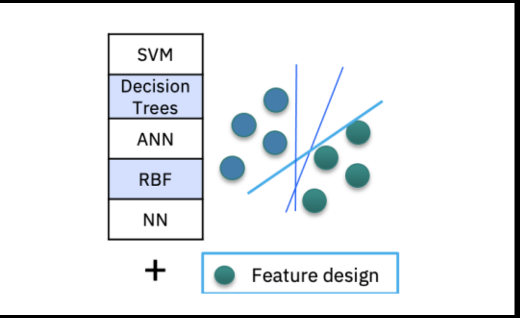
1980s | Big data

1980s to ~2010 | Massive labeled data + Compute



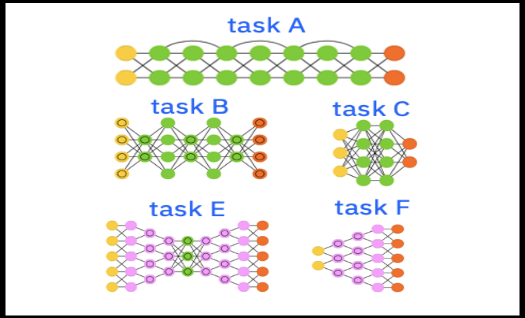
Expert Systems

Hand-crafted symbolic representations



Machine Learning

Task-specific hand-crafted feature representations



Deep Learning

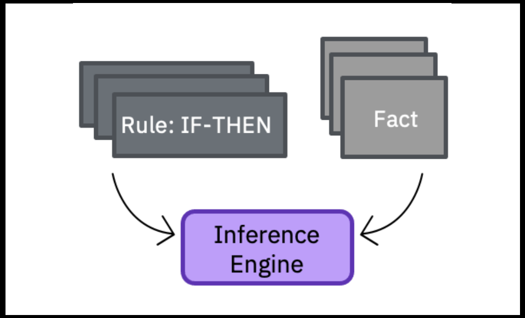
Task-specific learnt feature representations

Foundations models are an emerging class of representation that is changing the AI landscape

1980s | Big data

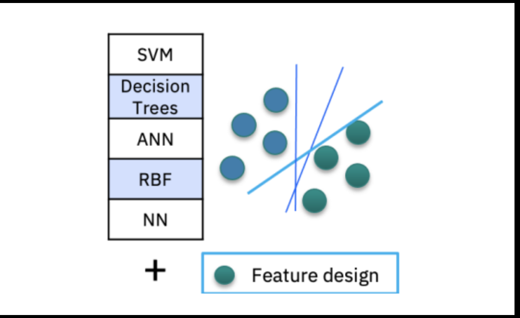
1980s to ~2010 | Massive labeled data + Compute

2017+ | Self-supervision at scale + Massive unlabeled data + Compute



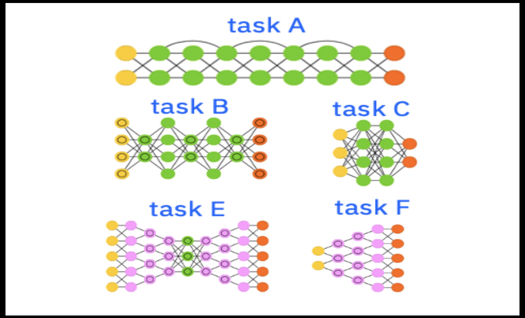
Expert Systems

Hand-crafted symbolic representations



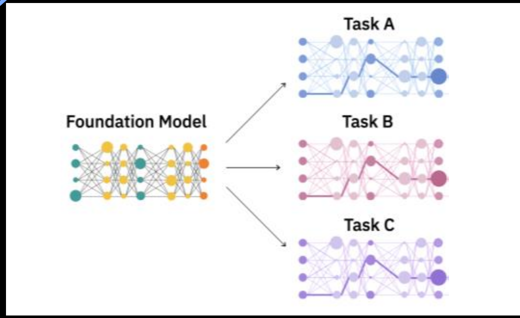
Machine Learning

Task-specific hand-crafted feature representations



Deep Learning

Task-specific learntedfeature representations

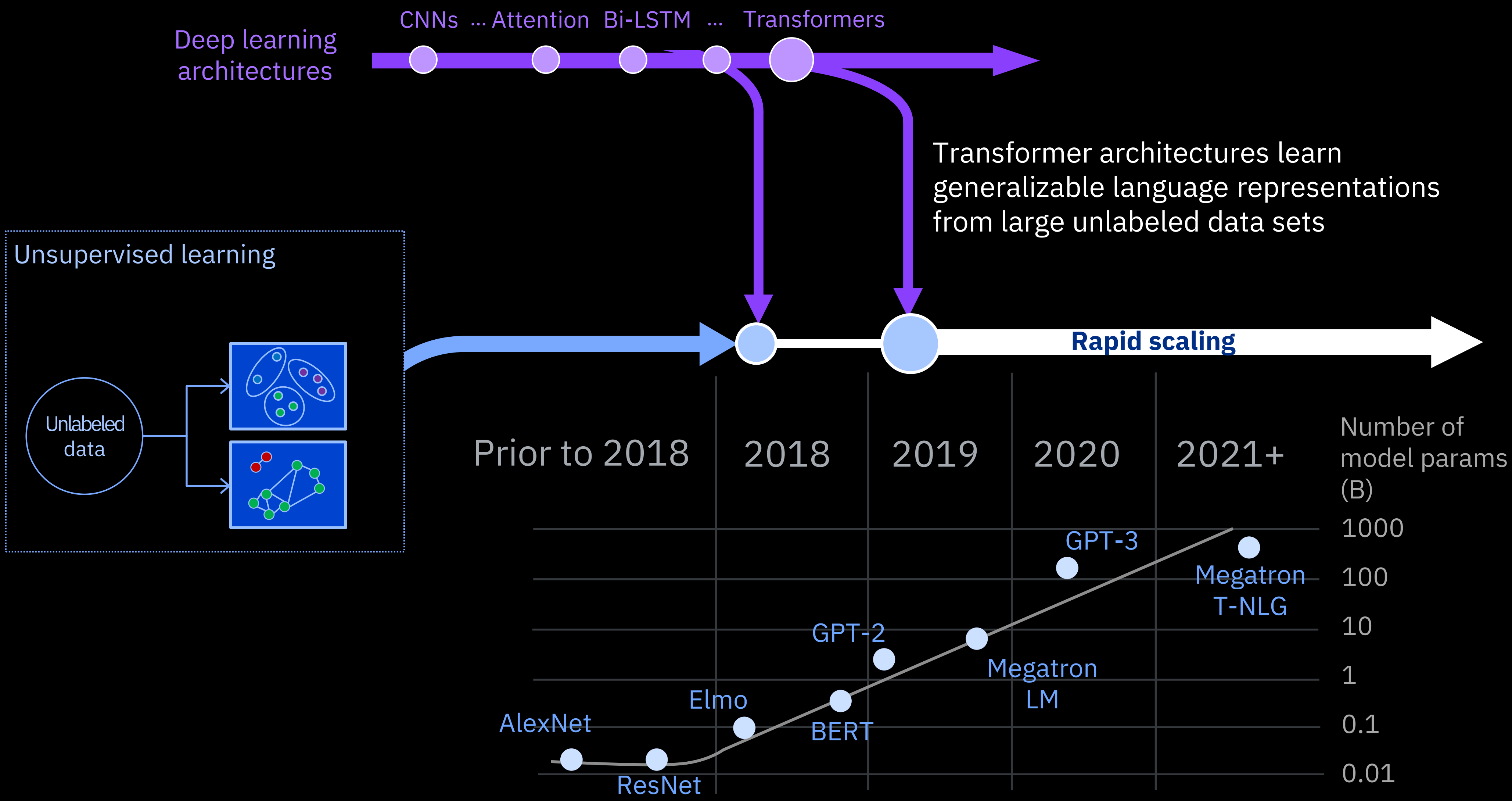


Foundation Models

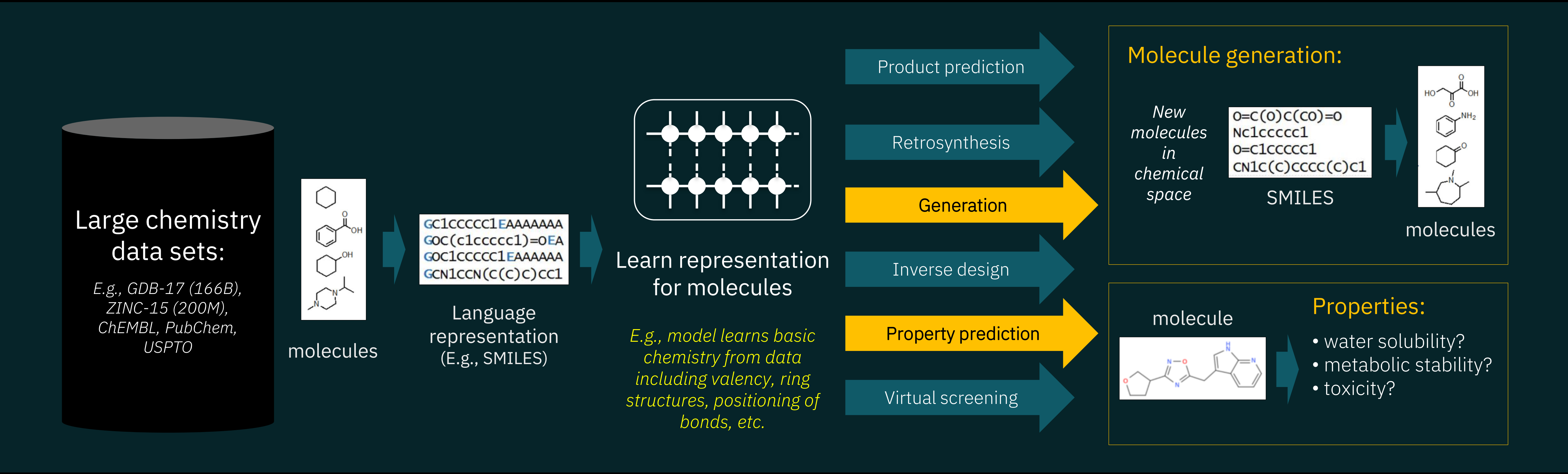
Generalizable & adaptable learned representations

Success of Foundation Models for Language from self-supervision *at scale*

Enabled by a novel learning architectures + data + compute



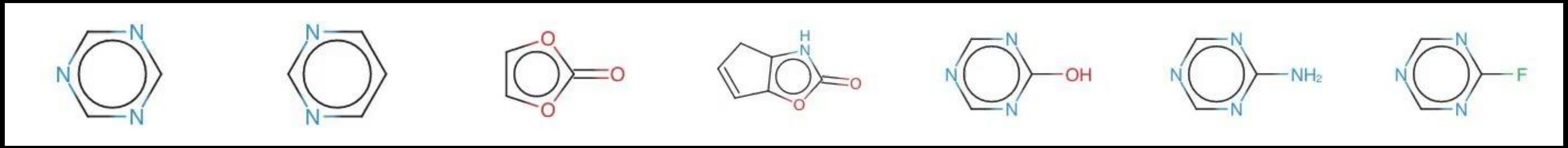
Foundation models for molecules enable multiple downstream tasks for molecule design, property prediction, reaction prediction, and more.



Generative models are a powerful tool for molecular inverse design and discovery

Learn from data to generate hypothetical and novel candidates for targeted properties

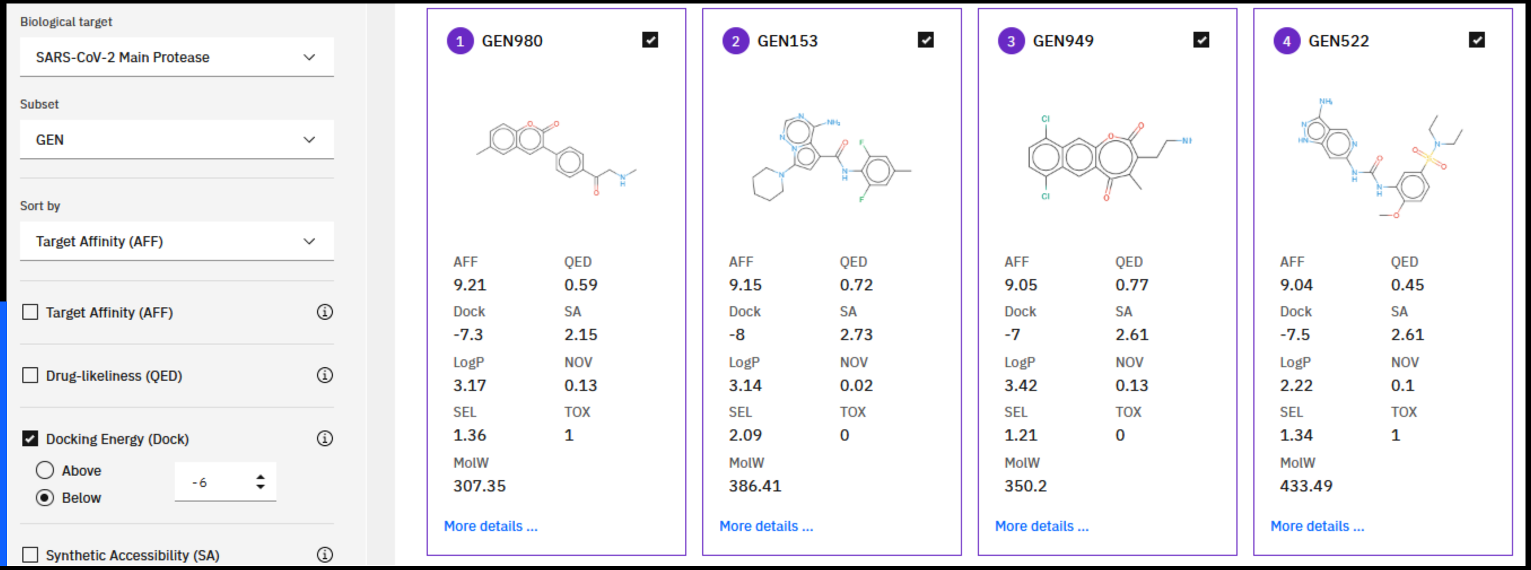
IBM Research, "Molecular Inverse-Design Platform for Material Industries", *KDD*, Aug 2020



IBM Research, "Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations," *Nature Biomedical Engineering* 2021

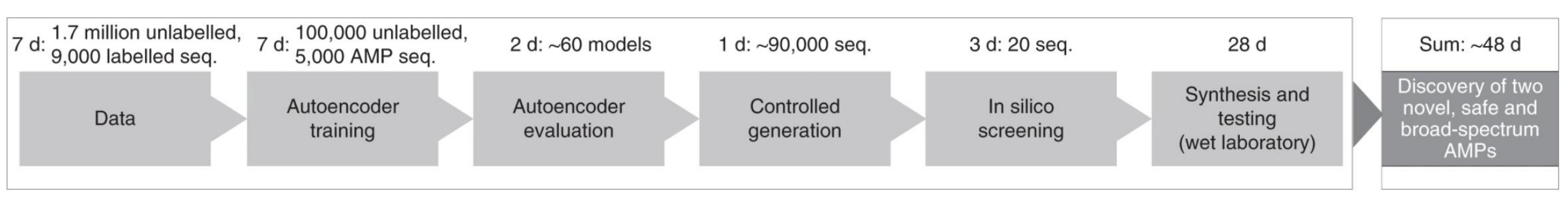
IBM Research, "Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models," May 2020

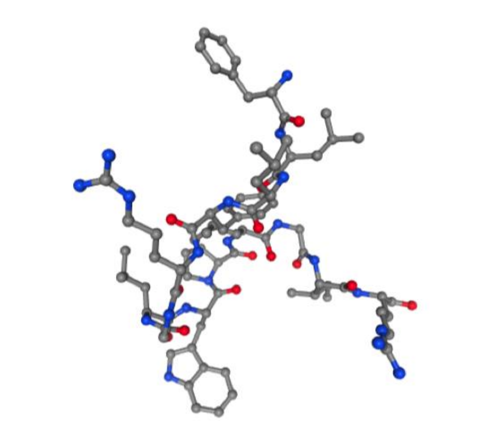
Explore novel drug candidates for COVID-19



Example: Generative Models for New Antibiotic Discovery

Accelerating end-to-end anti-microbial discovery with generative models and molecular dynamics.
Results demonstrate a 48 day time to discover and validate two new compounds, versus 2-4 years.





- Deep-neural-net-based generative models screen for antimicrobial function, broad-spectrum efficacy, presence of secondary structure and toxicity.
- Simulation confirms mode of action
- Synthesis and test reveal two new potent compounds

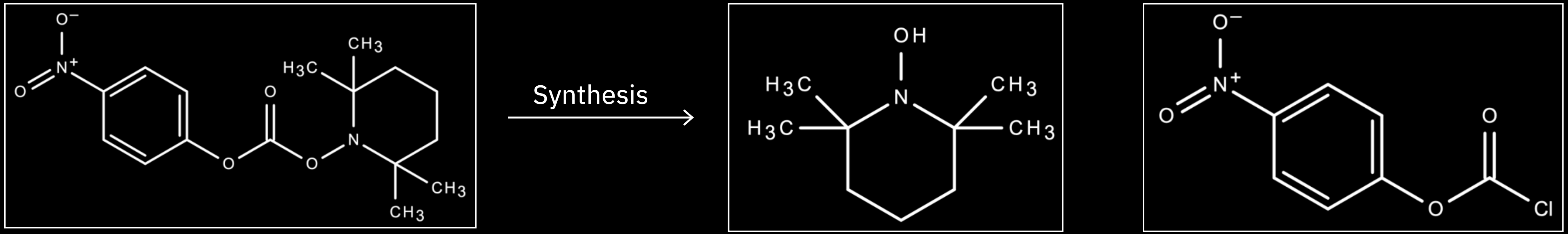
Two compounds displayed high potency against diverse Gram-positive and Gram-negative pathogens and a low propensity to induce drug resistance in Escherichia coli.

IBM Research & Oxford, "Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations," Nature Biomedical Engineering 2021

IBM RXN for Chemistry uses AI foundation models to predict chemical reactions and chemical procedures

Chemical Reaction predictions

1



Chemical Procedure (recipe) predictions

2

Add 2,2,6,6-Tetramethylpiperidin-1-ol (2.0 g)

Mix 4-nitrophenyl chloroformate (1.1 g) and dichloromethane (5 ml) in a separate vial

Add mixture dropwise at -10°C

Stir for 1 hour at -10°C

Quench with aqueous Na_2CO_3 (15 ml)

Extract with dichloromethane (20 ml)

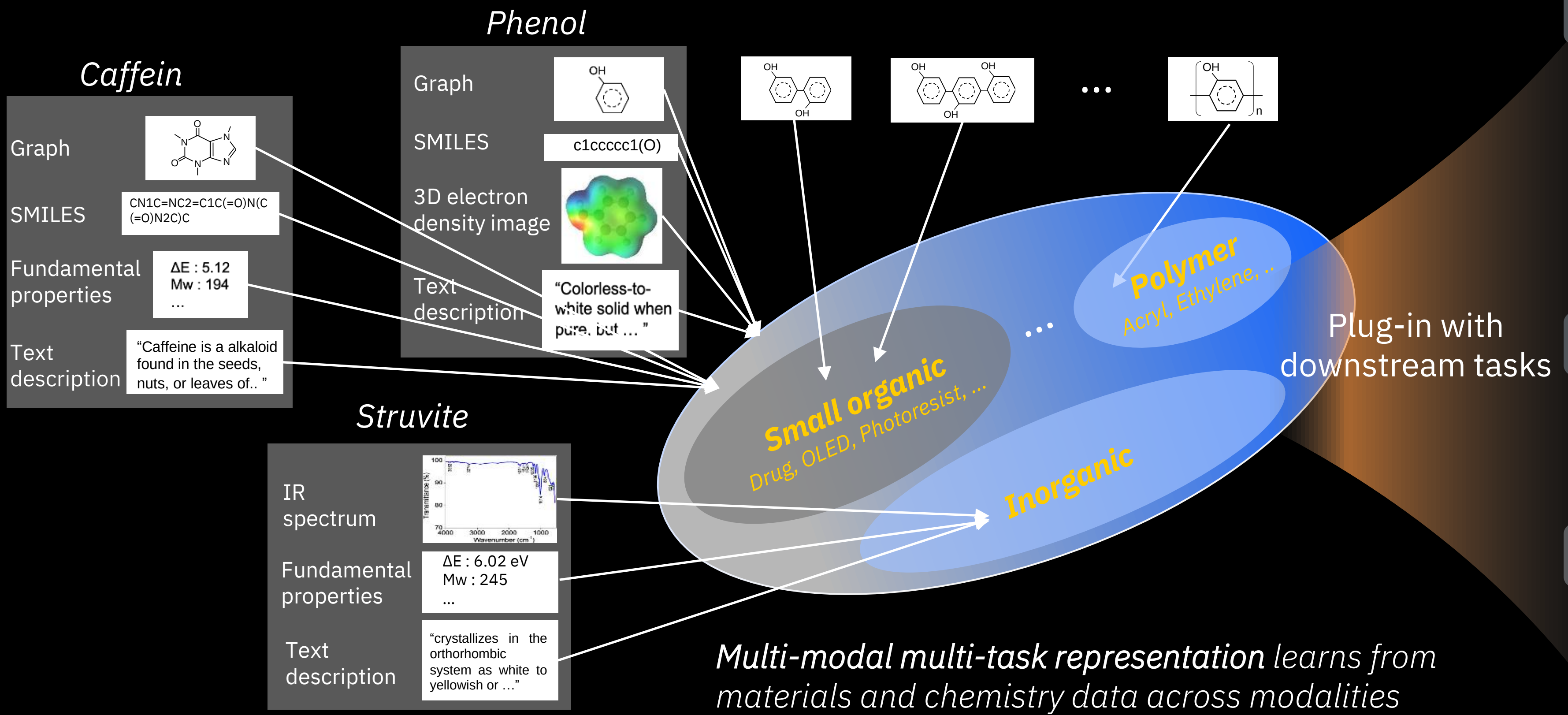
Concentrate

Next Steps

(1) Learning universal representations for materials, chemistry, biology, etc

Multi-modal foundation models that can power a broad field of downstream tasks

- Comprehensive reusable representation across domains – breaks down silos of independent modeling
- Integration of multiple modalities fuses diverse sources of knowledge and data



Multi-modal multi-task representation learns from materials and chemistry data across modalities

Generative model

- Generate new materials' information with multiple modalities (image, text description, etc.)
- Target properties with limited data
- Interpretable generation

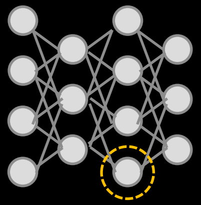
Predictive model

- Predict properties in small/zero data domains
- Interpretable prediction

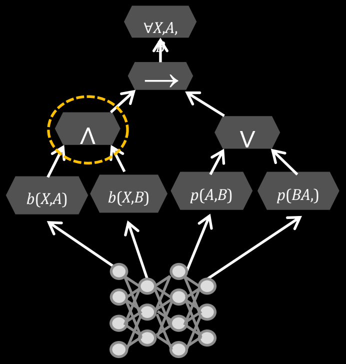
Simulation accelerator

- Intellectual speculative skip of simulation steps by orbital information (surrogate model)

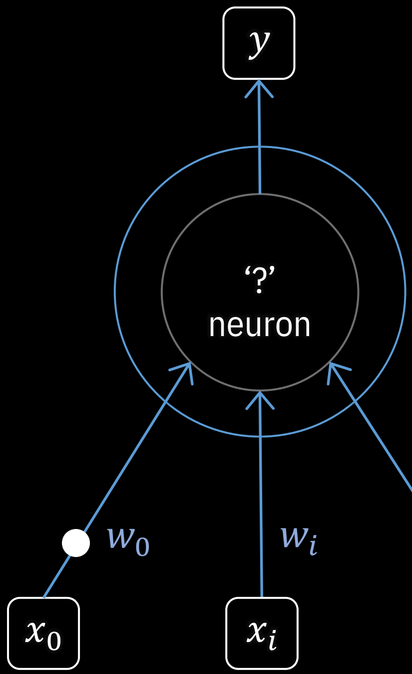
(2) Integrating Real-valued Logic into Foundation Model Representations using Logical Neural Networks (LNNs) for Improved Explainability



Standard Neuron



Neuron architecture in Logical Neural Networks



Activation Function introduces Non-linearities

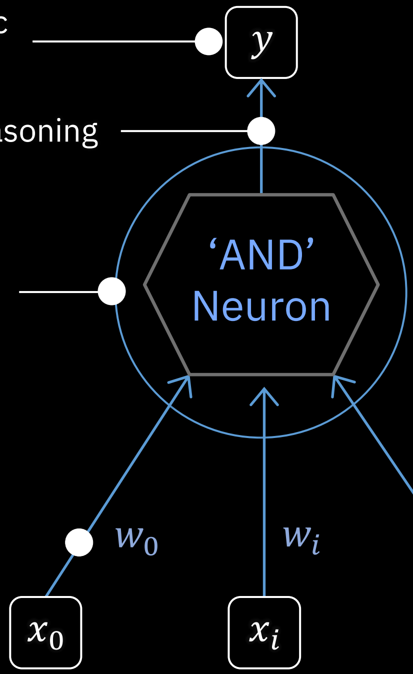
Input truth values

Werbos, 1974
Rumelhart, Hinton and Williams, 1986

Real-valued Logic
0<=y<=1

Learning and Reasoning

Constraints reflecting expert knowledge



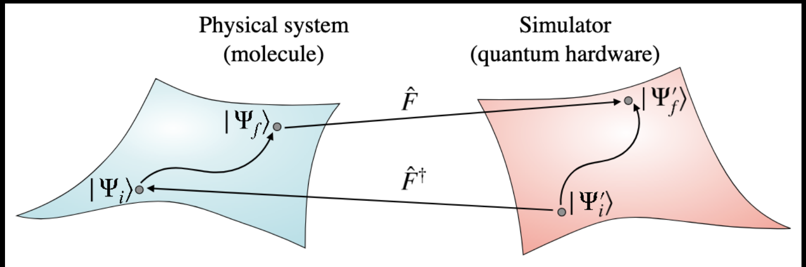
"AND" neuron

Activation Function introduces Logical operations

Input truth values

Riegel, et al., 2020, Logical Neural Networks

(3) **Quantum circuit-based representations** can potentially compute properties in physics and chemistry more naturally, accurately and efficiently



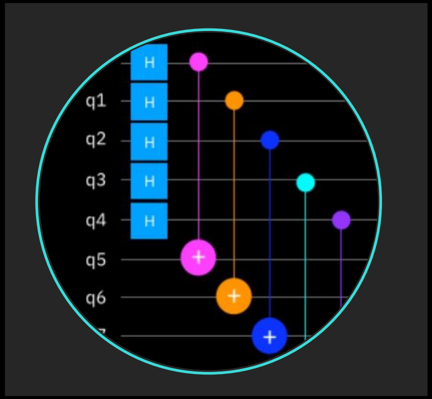
Quantum computers are universal simulators of physics and chemistry – can provide efficient representations of dynamics by quantum circuits

Molecule

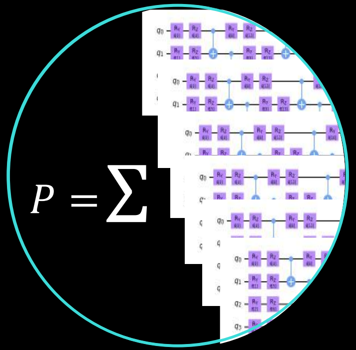
$$H = \sum_{p,q} h_{pq}\, c_p^+ c_q + \frac{1}{2} \sum_{p,q,r,s} h_{pqrs} c_p^+ c_q^+ c_q c_s$$

$$H = \sum_{\mu} c_\mu P_\mu$$

Mapped to qubit operators



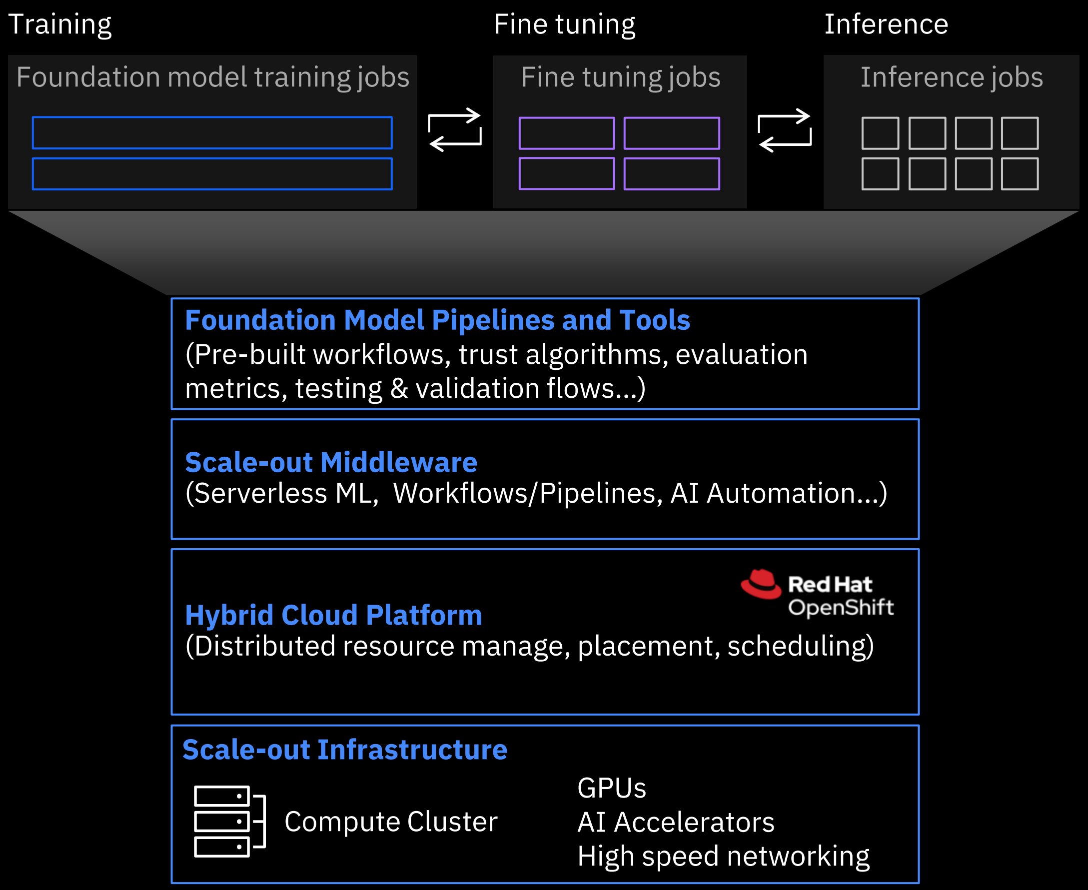
Mapped to quantum circuits representation



Properties measured on quantum computer

Potential applications include drug discovery, next-gen battery design, corrosion analysis, structural analysis, new materials design, solar conversion, catalysts ad enzyme design

(4) IBM tools and infrastructure for foundation models



Scale-out infrastructure

- Optimized container networking implementations
- AI accelerators

Scalable middleware stack for distributed training

- Store and compute disaggregation
- Elastic and fault-tolerant distributed training
- Data caching
- Efficient resource management (job placement & scheduling)

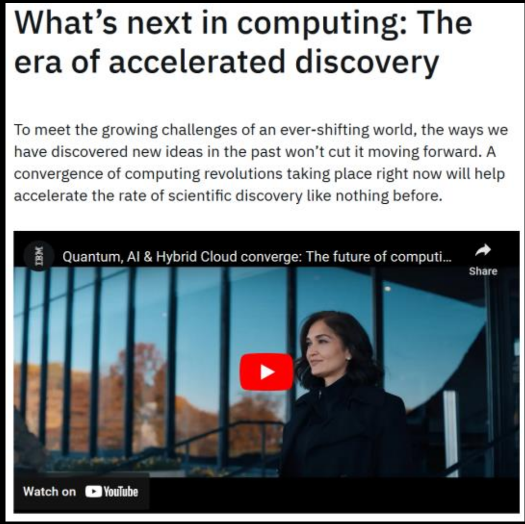
Tools and workflows for train, test, and consume

- Tools to support data scientist productivity
- Standardized reusable pipelines for train, test, validate, fine tune
- Tools for automatic metric collections for trust and quality control
- Scalable tools for model exploration & evaluation

https://research.ibm.com/topics/foundation-models

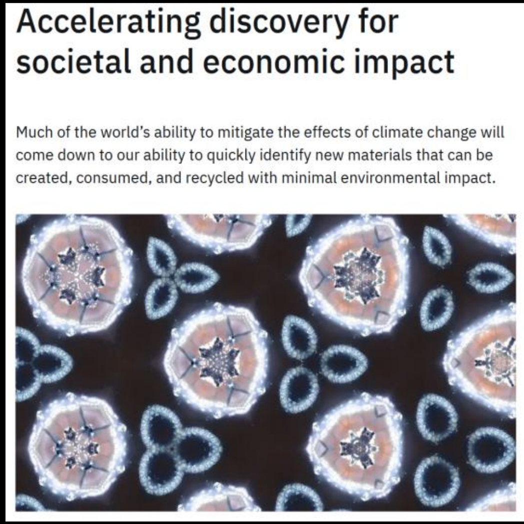
For more information ...

Accelerated Discovery


What's next in computing: The era of accelerated discovery

To meet the growing challenges of an ever-shifting world, the ways we have discovered new ideas in the past won't cut it moving forward. A convergence of computing revolutions taking place right now will help accelerate the rate of scientific discovery like nothing before.

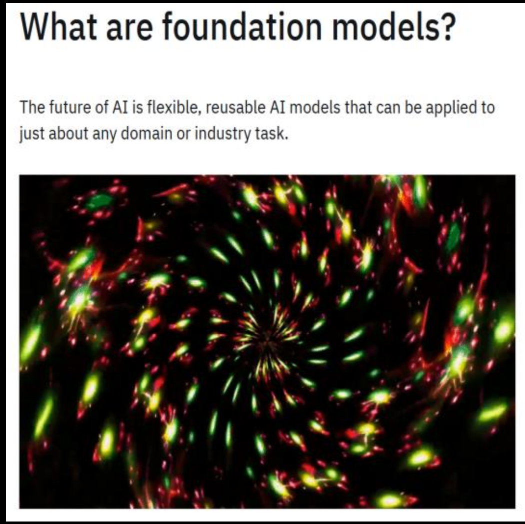
https://research.ibm.com/blog/what-is-accelerated-discovery


Accelerating discovery for societal and economic impact

Much of the world's ability to mitigate the effects of climate change will come down to our ability to quickly identify new materials that can be created, consumed, and recycled with minimal environmental impact.

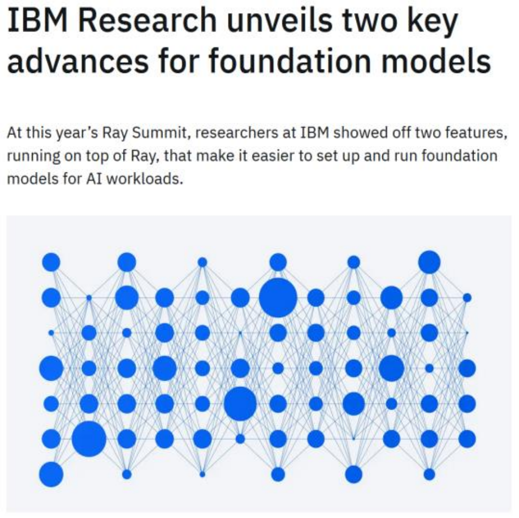
https://research.ibm.com/blog/new-sustainable-materials

Foundation Models


What are foundation models?

The future of AI is flexible, reusable AI models that can be applied to just about any domain or industry task.

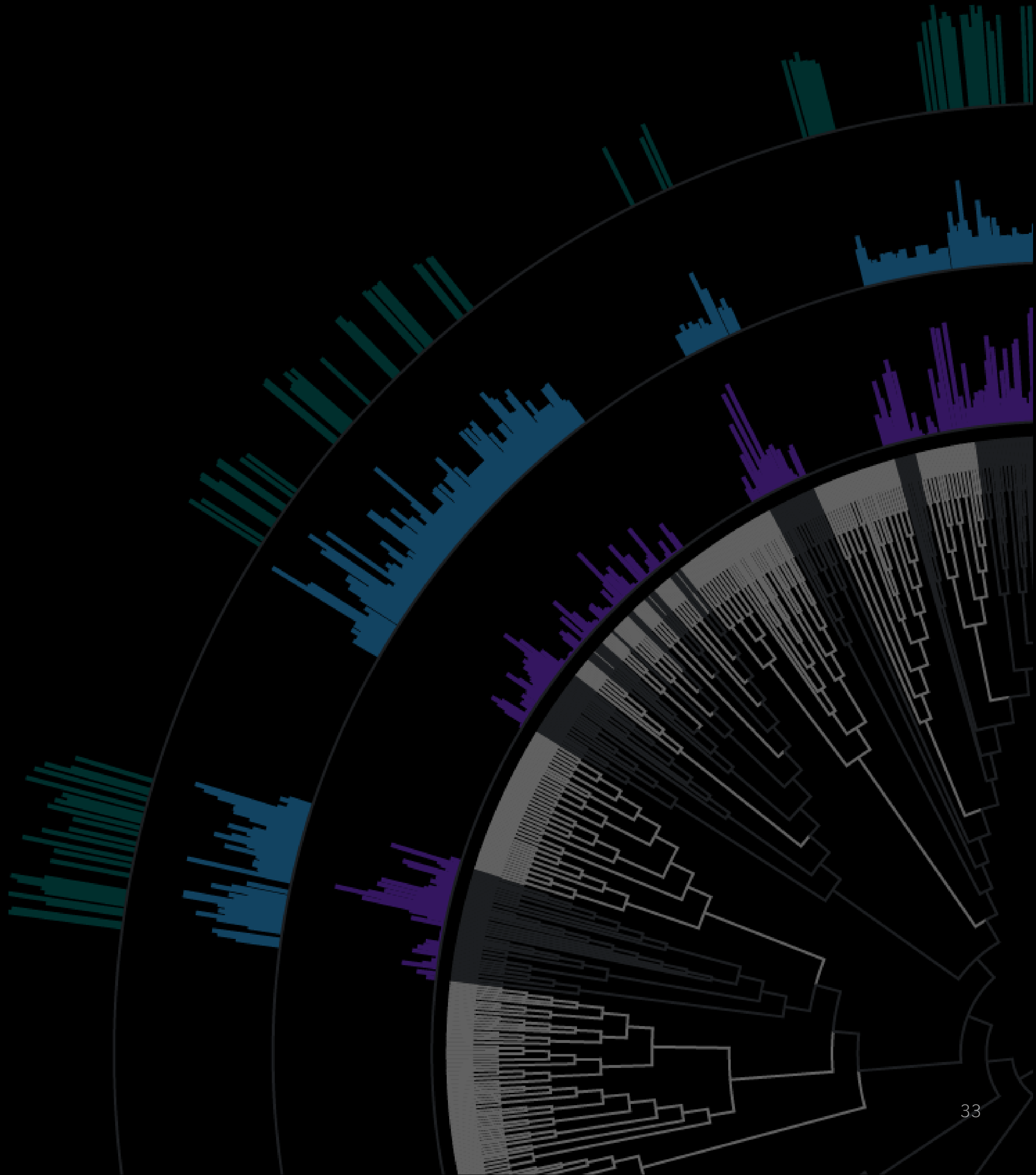
https://research.ibm.com/blog/what-are-foundation-models


IBM Research unveils two key advances for foundation models

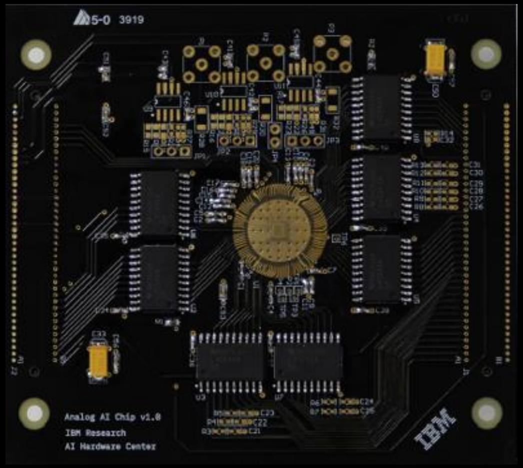
At this year's Ray Summit, researchers at IBM showed off two features, running on top of Ray, that make it easier to set up and run foundation models for AI workloads.

https://research.ibm.com/blog/ray-summit-codeflare-foundation-models

Thank You!



What’s Next in Computing?



Bits + Neurons + Qubits

Hybrid Cloud: cloud + high performance computing + robotics, instruments and labs

Artificial Intelligence (AI): advanced data-driven modeling, analytics and automation

Quantum Computing: making intractable problems tractable through a new computing modality