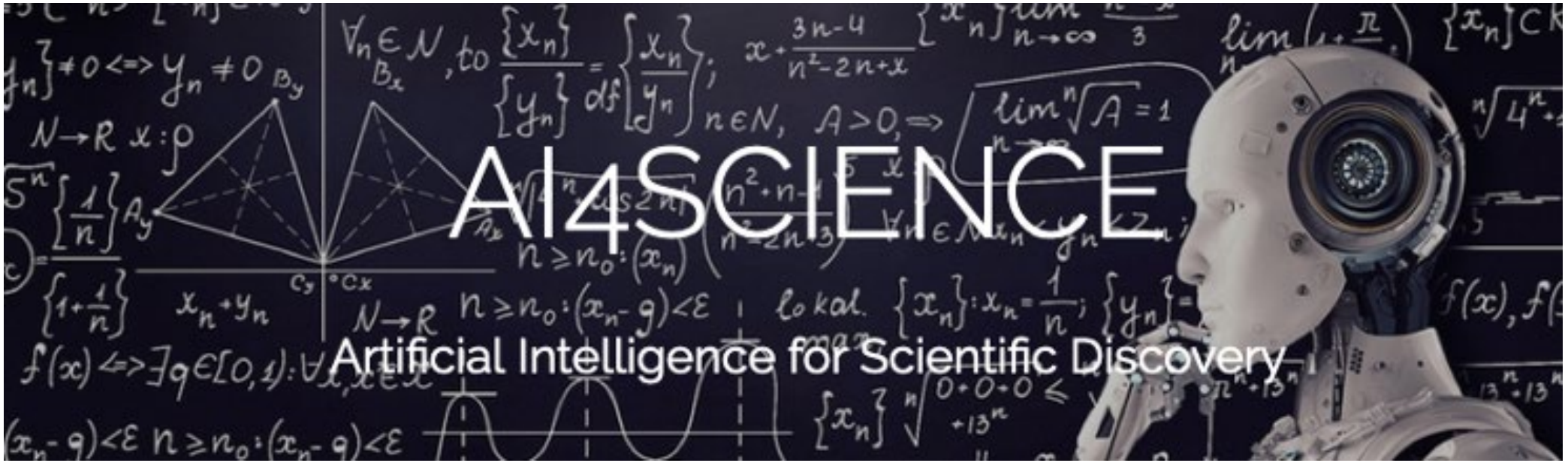
Artificial Intelligence for Scientific Discovery

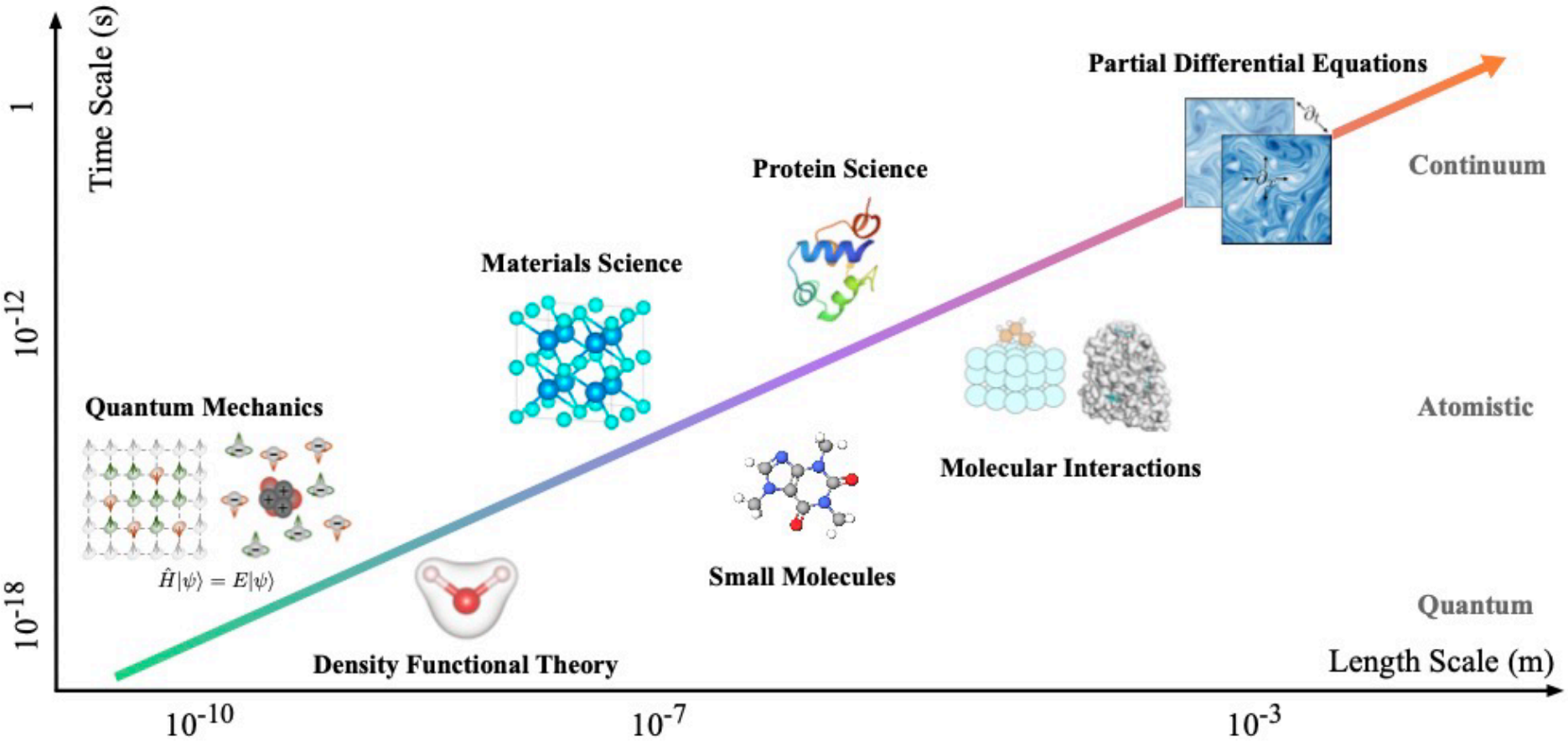
Yi Liu, Assistant Professor

AMS/Data Science, Stony Brook University



Use recent advances in artificial intelligence and deep learning to solve problems in natural sciences: quantum chemistry, applied math, quantum physics, material science,…

Science: Understand and Explain the Natural World



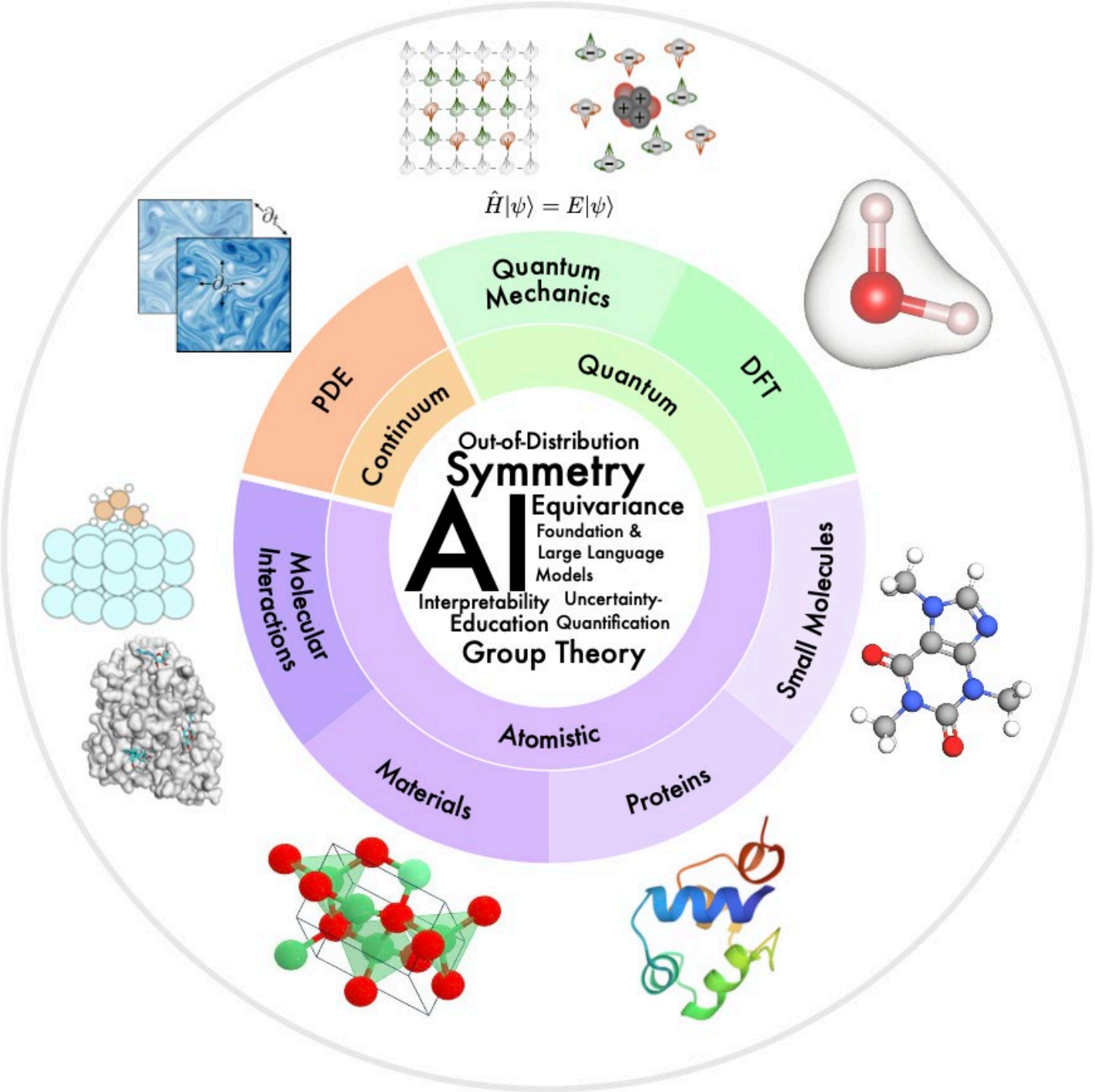
- What are the wavefunctions of electrons?
- How atoms move and interact?
- How small molecules bind to their targets?
- How fluids flow?

- Schrödinger equation, DFT
- Newtonian mechanics
- Finite-element, fluid mechanics, PDEs

Why AI is Needed in Science?

"The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble." -- Paul Dirac

- The Schrödinger equation governs quantum systems, but incurs exponential complexity
- Navier-Stokes equations describe fluid mechanics, but solving them is highly demanding
- AI4Science: Develop innovative AI methods to accelerate scientific discoveries



Key mission:

- How to properly (and efficiently) integrate domain knowledge in science (like symmetry) into AI models
- New AI models - innovations in both AI/science

Other perspectives:

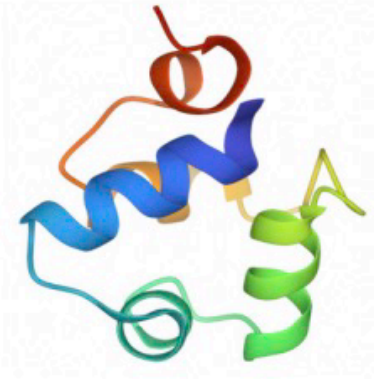
- Explainability
- Imperfect scientific data
- Large language models
- Uncertainty estimate

Symmetries in Science

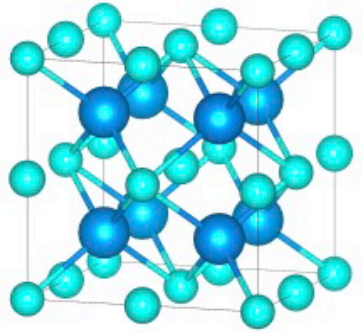
"It is only slightly overstating the case to say that physics is the study of symmetry."

—Philip W. Anderson (1972)

Molecule



Protein

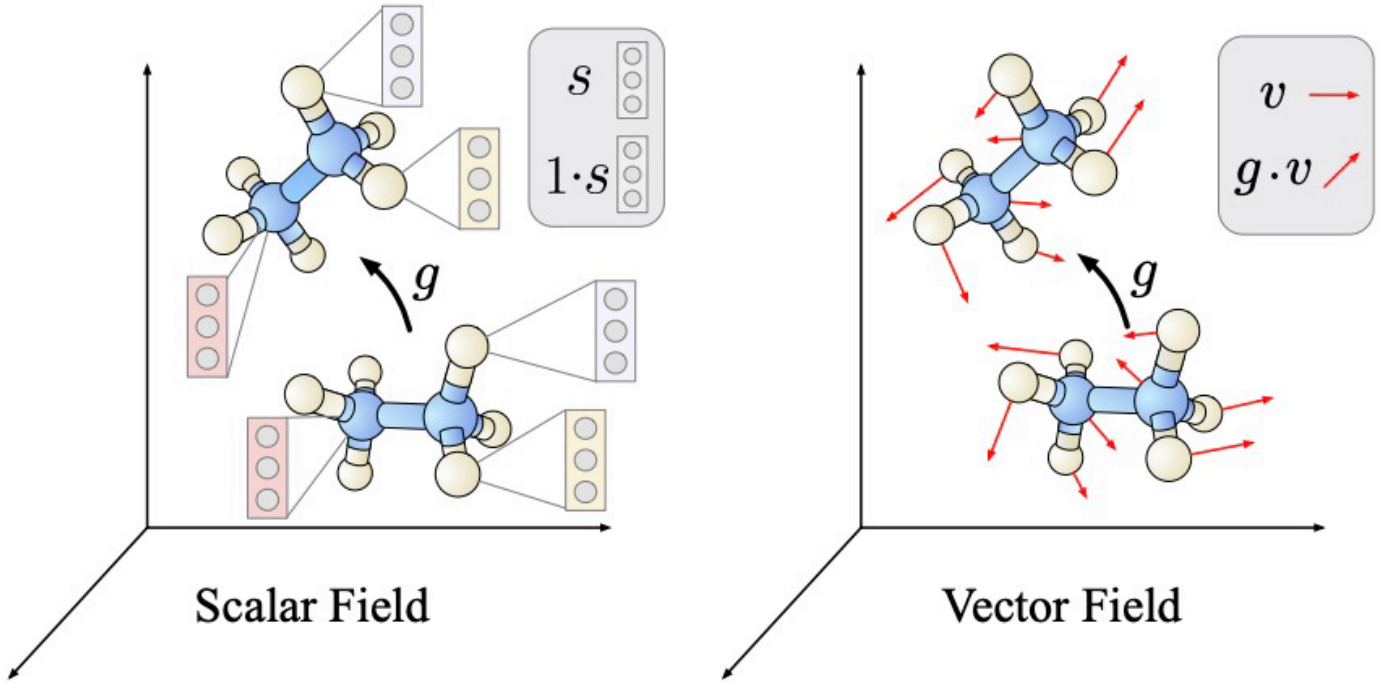


Material

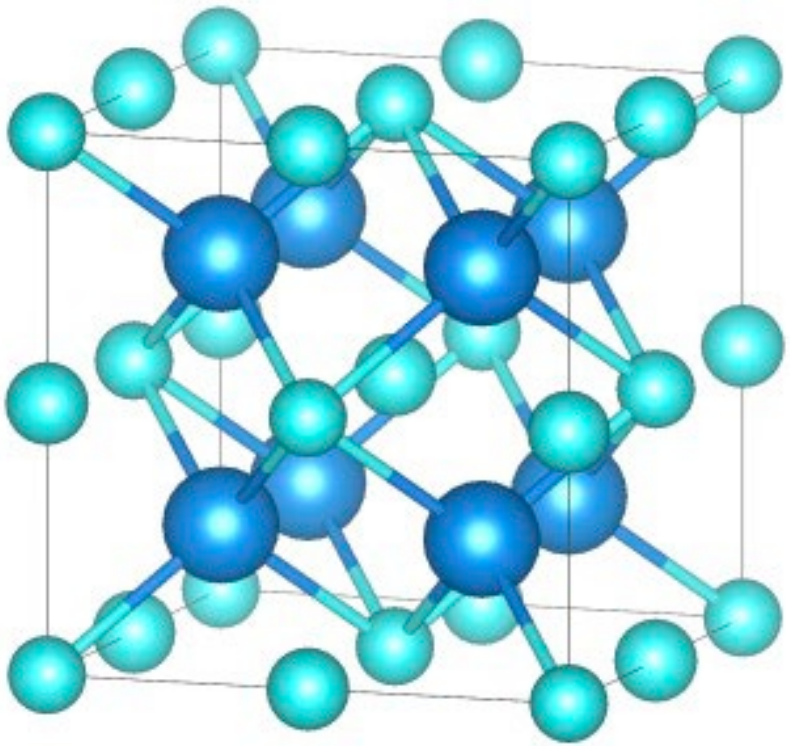
- **Symmetries** lie at the core of physics
- Nature does not use coordinates
- Define a set of transformations that **change** mathematical descriptions but **maintain** physical properties
- Result in **invariance** or **equivariance** of physical properties

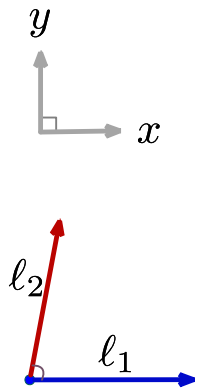
Invariance and Equivariance

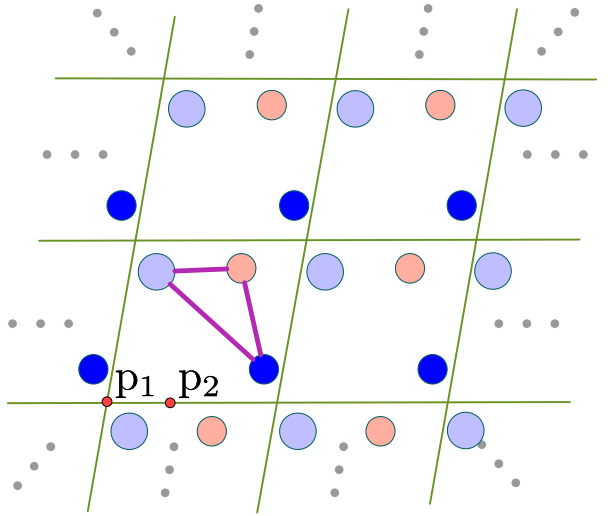
- When applying a symmetry transformation g to an object,
  - **Invariance:** physical properties remain the same
  - **Equivariance:** physical properties changed by symmetry transformations (same as or different from g)
- Example: Rotating a 3D molecule
  - retains its atom type (invariance)
  - rotates its atomic force fields (equivariance)



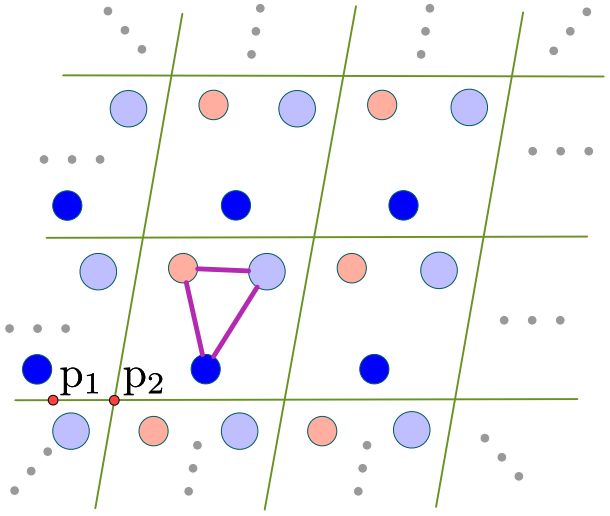
Symmetries in Crystal Material







(a)

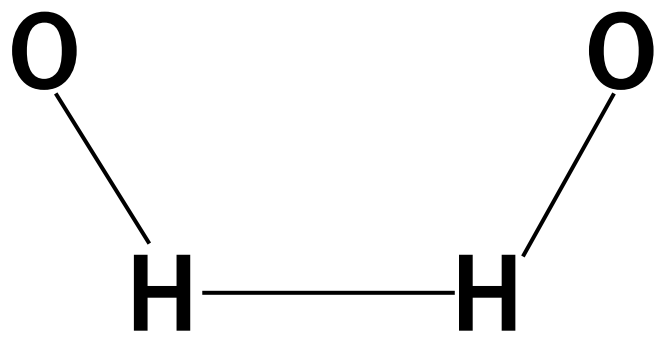


(b)

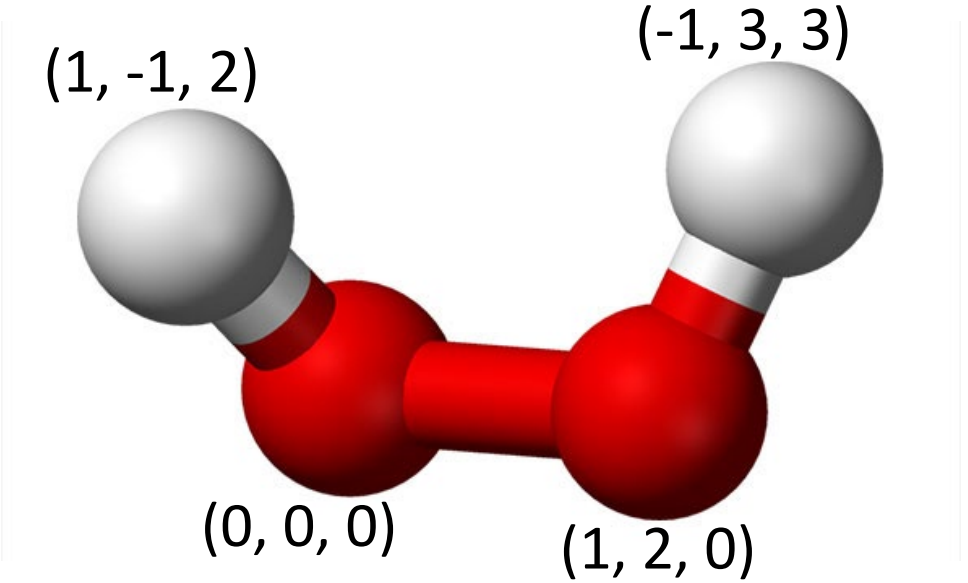
Periodic Invariance

A Bit of Techniques…

Molecules as 3D graphs



Molecules as 2D graphs

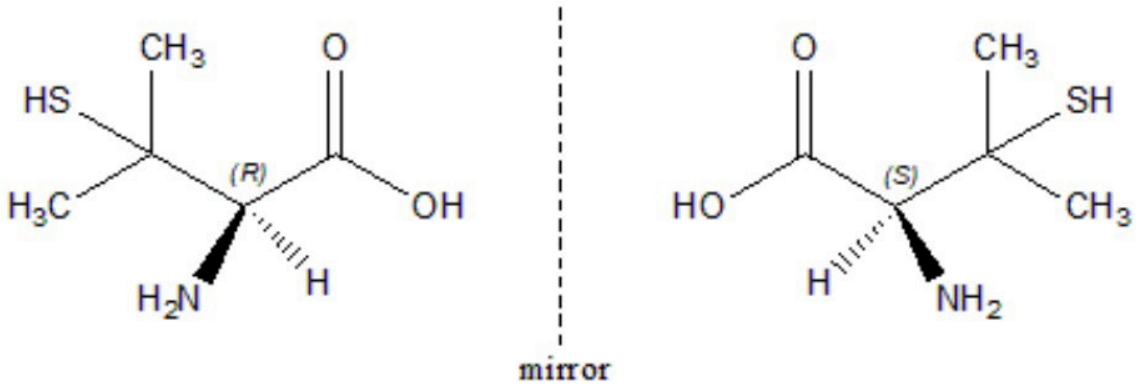


Position information:
Final conformation

https://www.msdsonline.com/2015/06/17/hydrogen-peroxide-health-safety-tips/

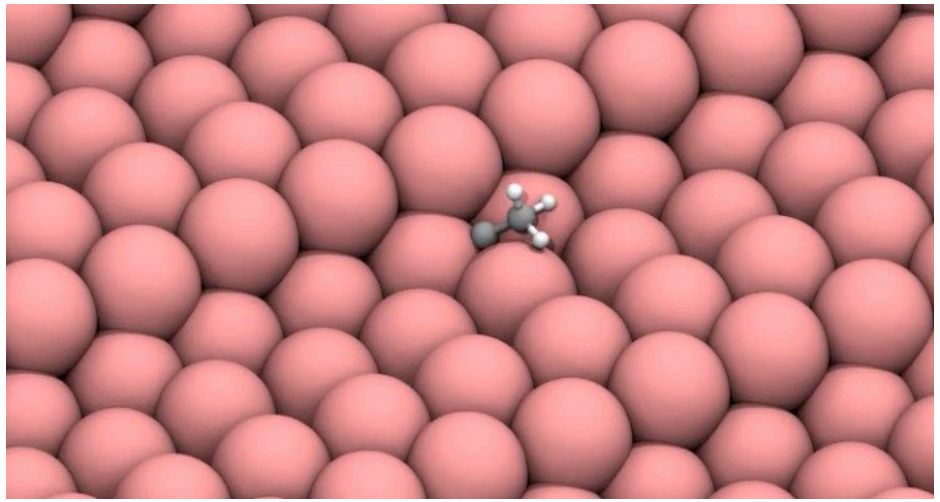
3D Graph Learning – Main Challenges

- **Symmetry:** invariant properties (e.g. energy) and equivariant properties (e.g. force)
- **Geometric Completeness:** ability to distinguish different 3D geometries of molecules
- **Efficiency (Complexity):** applicable to large-scale molecules and datasets

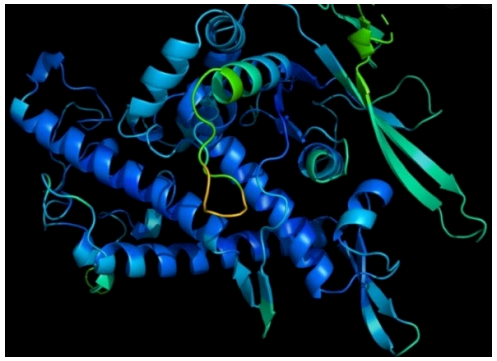


R-Penicillamine
Extremely toxic

S-Penicillamine
Antiarthritic - medicine

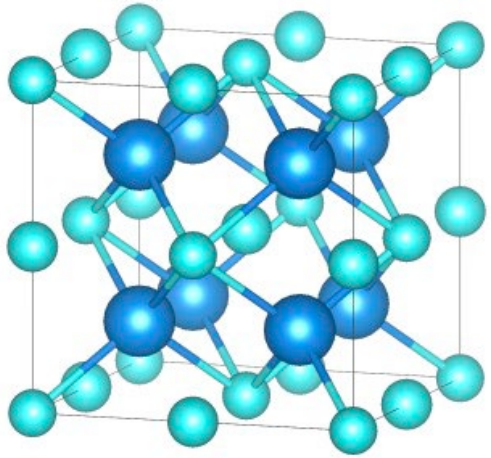


Macro Molecules and Interactions



- Psychological process understanding, drug discovery...
- Large graphs with hierarchies
- **Need collaboration for both data and domain knowledge**

Protein

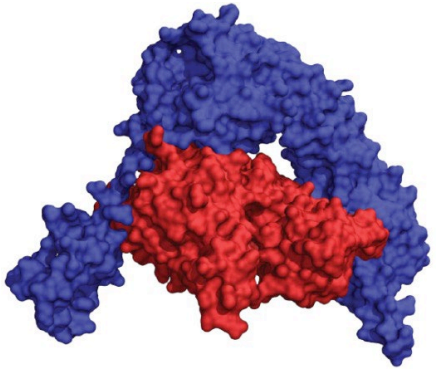


Material property prediction, new material design…

Infinite with periodic information - repeated lattices

Graph modeling is challenging

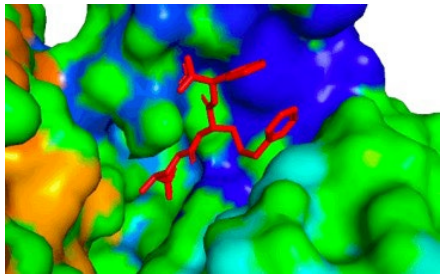
Material



Important in biological systems

Antibody-antigen interaction

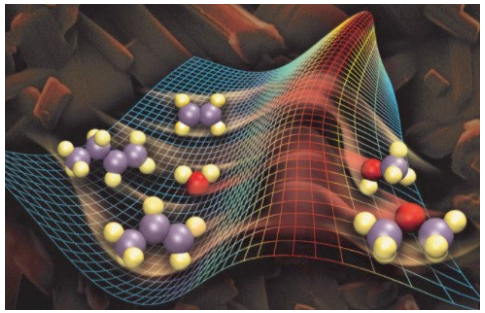
Protein-Protein



How molecules affect protein functions

Drug discovery

Protein-Molecule



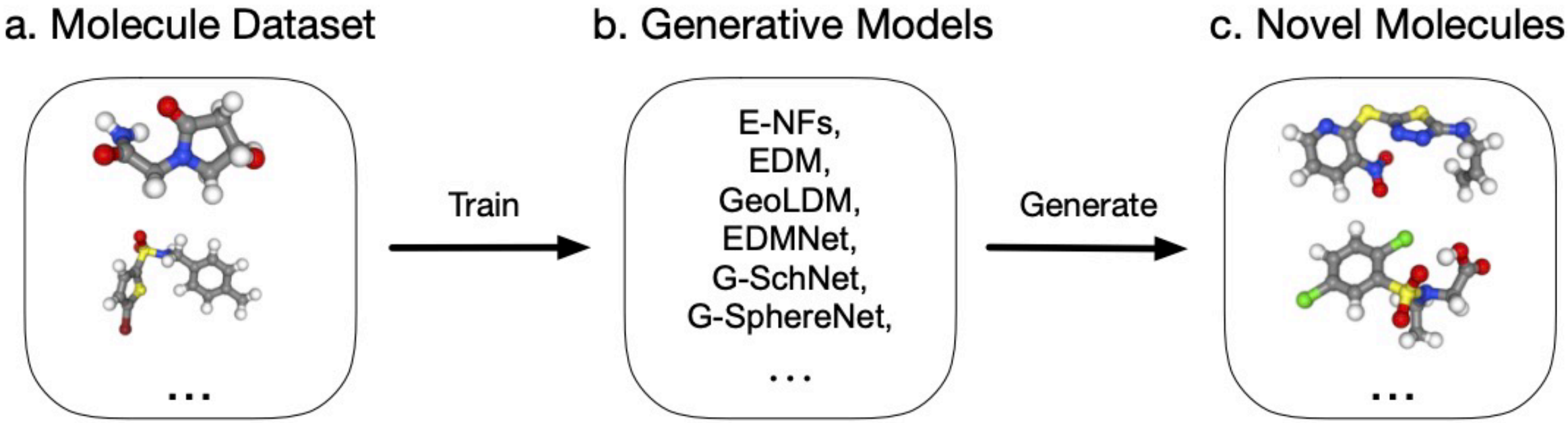
Molecules interact environment

Chemical reaction simulation, **material discovery** …

Material-Molecule

Current Projects

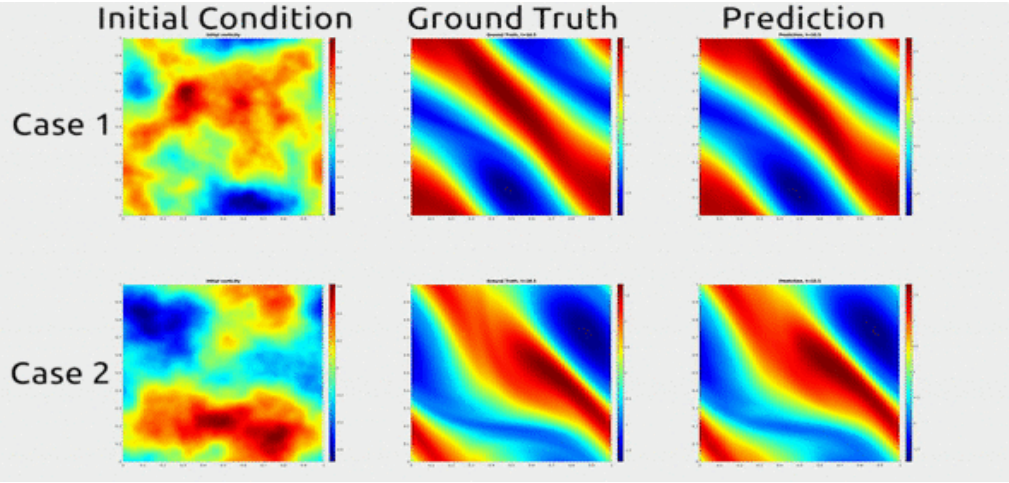
Molecule Generation



- Given a molecule dataset, we aim to learn the probability distribution of molecules by generative models and generate novel molecules with desirable properties

Partial Differential Equations for Continuum Systems

A PDE mathematically describes the behavior of a system by prescribing constraints relating partial derivatives.



Source: https://zongyi-li.github.io/blog/2020/fourier-pde/

PDE solvers

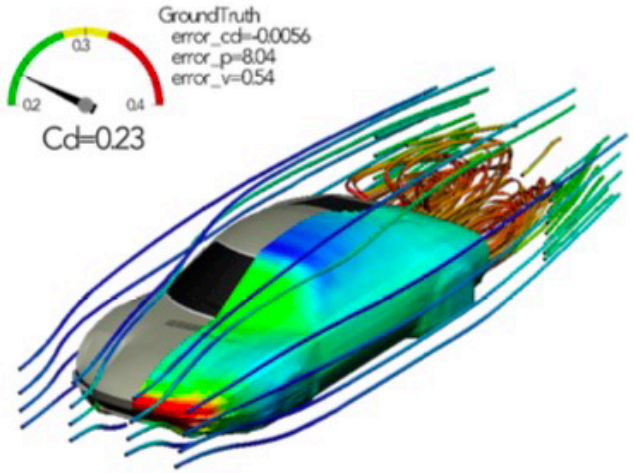
Traditional Solvers

- Discretized Methods:
  - Finite Element Method
  - Finite Volume Method
  - Finite Difference Method
- Spectral Methods
- Other Mesh-free Methods
- Pros:
  - Accuracy
- Cons:
  - Slow
  - Suffers from the Curse of Dimentionality(CoD)

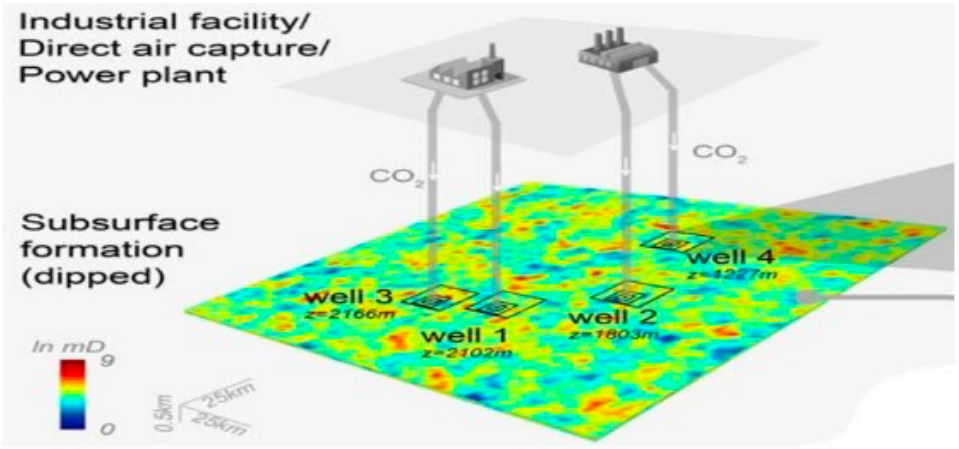
Neural PDE Solvers

- Physics-informed Neural Networks
  - Incorporate known physics
  - Mesh-free
  - Lessen the CoD issue
  - Can train without data
- Neural Operators
  - Solve a family of PDEs
  - Inference is several orders of magnitude faster than traditional solvers
  - Can be mesh-independent, resolution-invariant

Applications of Operator Learning



Computational Fluid Dynamics [a]



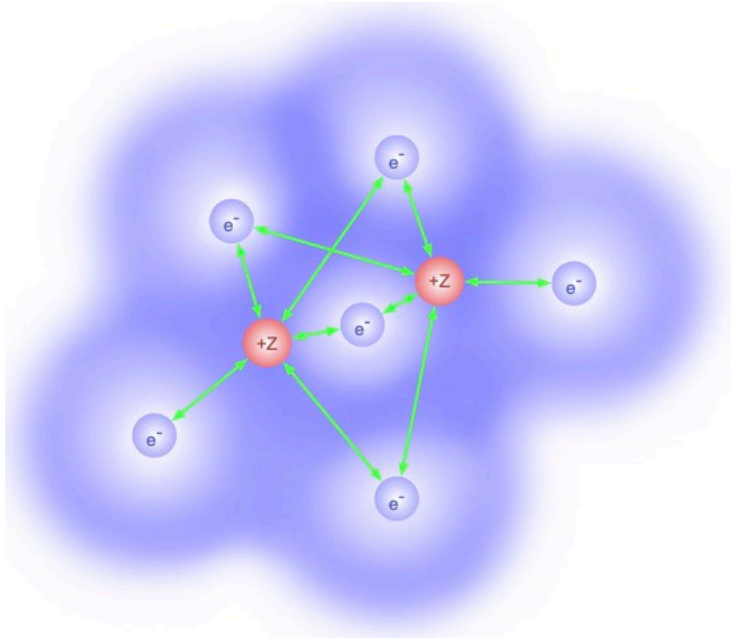
Carbon Storage Modeling [b]

Weather Modeling [c]

[a] Geometry-informed neural operator for large-scale 3D PDEs. arXiv, 2023. Zongyi Li, et al..
[b] Fourier-MIONet: Fourier-enhanced multiple-input neural operators for multiphase modeling of geological carbon sequestration. arXiv, 2023. Zhongyi Jiang, et al..
[c] DeepPhysiNet: Bridging deep learning and atmospheric physics for accurate and continuous weather modeling. arXiv, 2024. Wenyuan Li, et al..

Imperfect Scientific Data

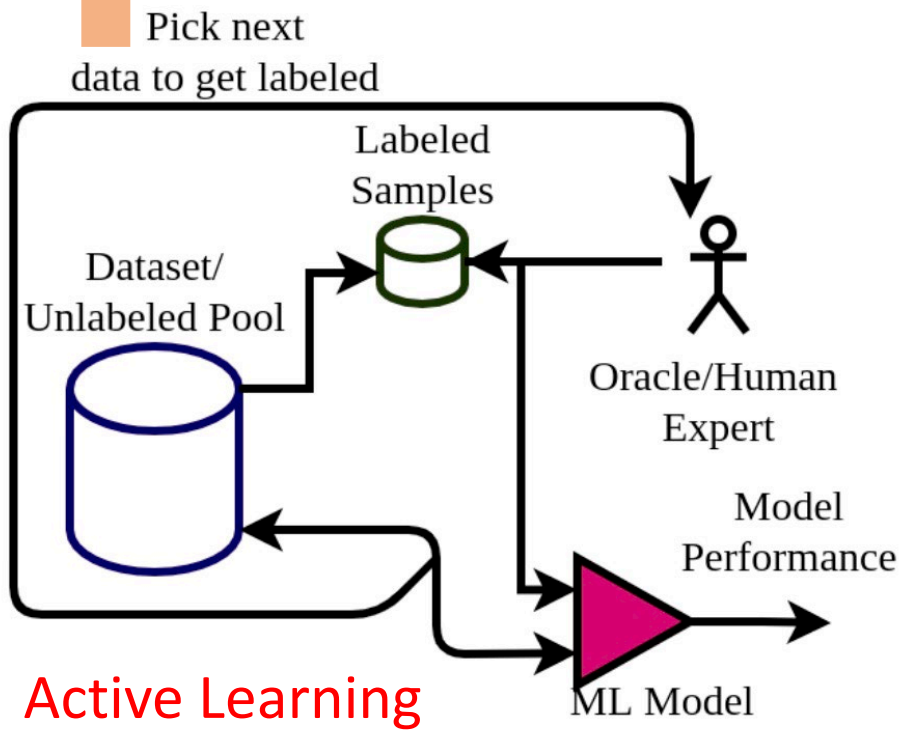
- Annotating/labeling scientific data is particularly hard
  - Requires heavy domain expertise
  - Annotation cost is extremely high



Density Functional Theory (DFT) can label molecule energy
But can be hundreds of thousands times slower than a NN

How?

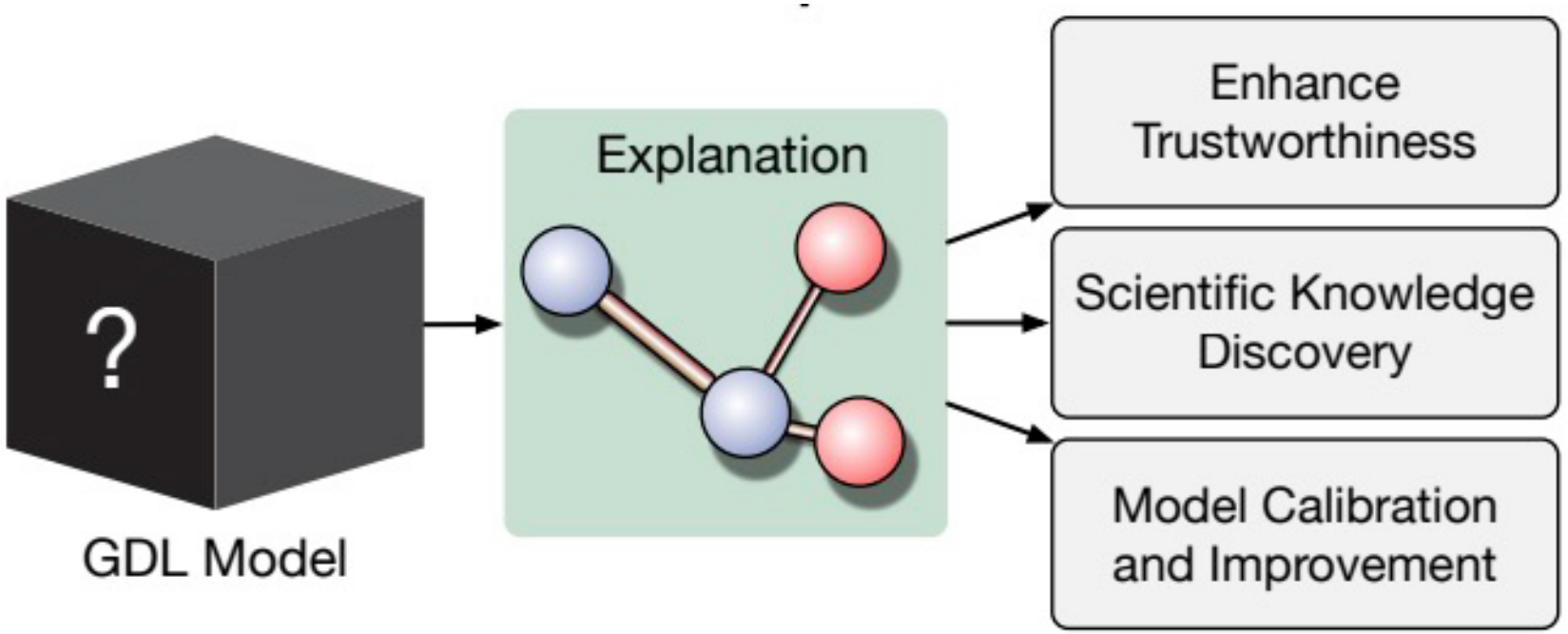
- **Uncertainty:** how the model is confident about a sample
  Molecules?
- **Diversity:** how a sample is different from others
  Molecules?



Active Learning

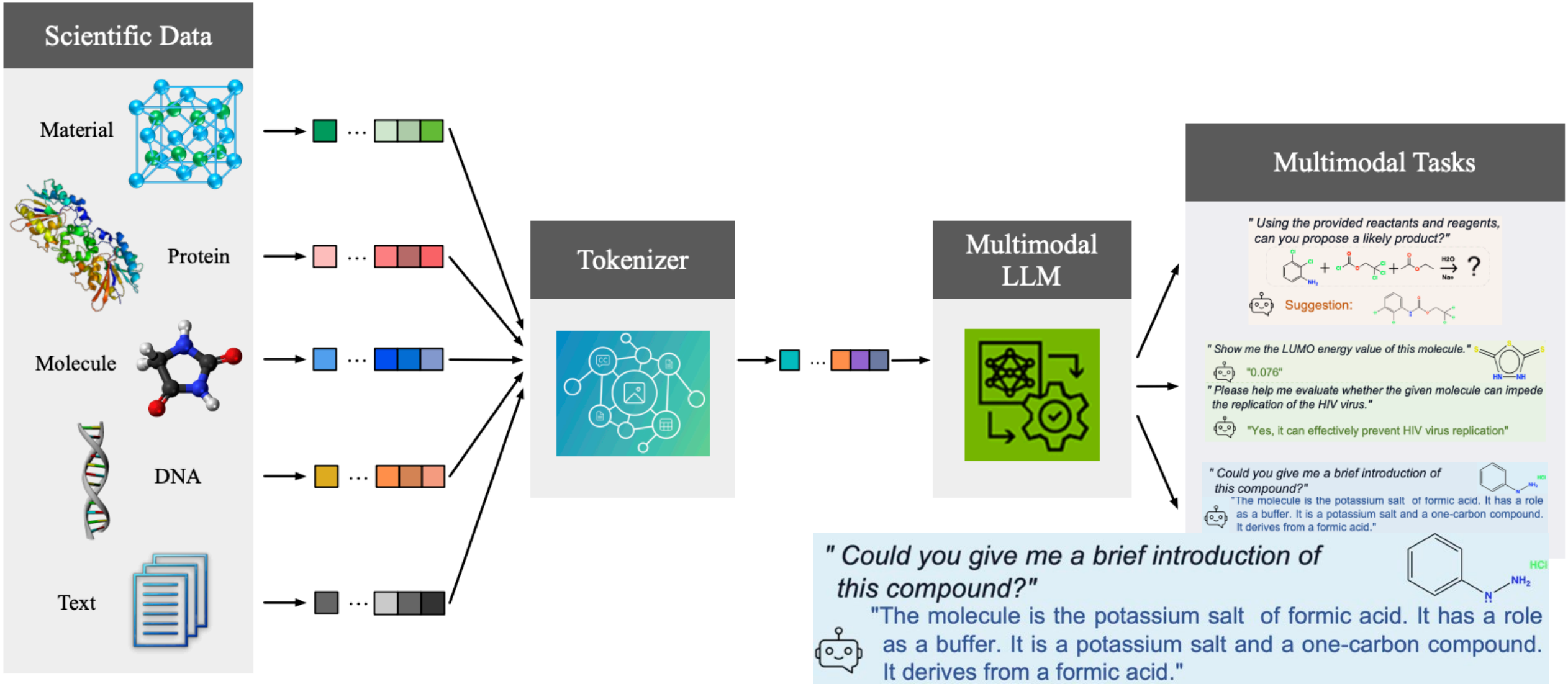
Explainability

Science seeks to understand and explain the natural world



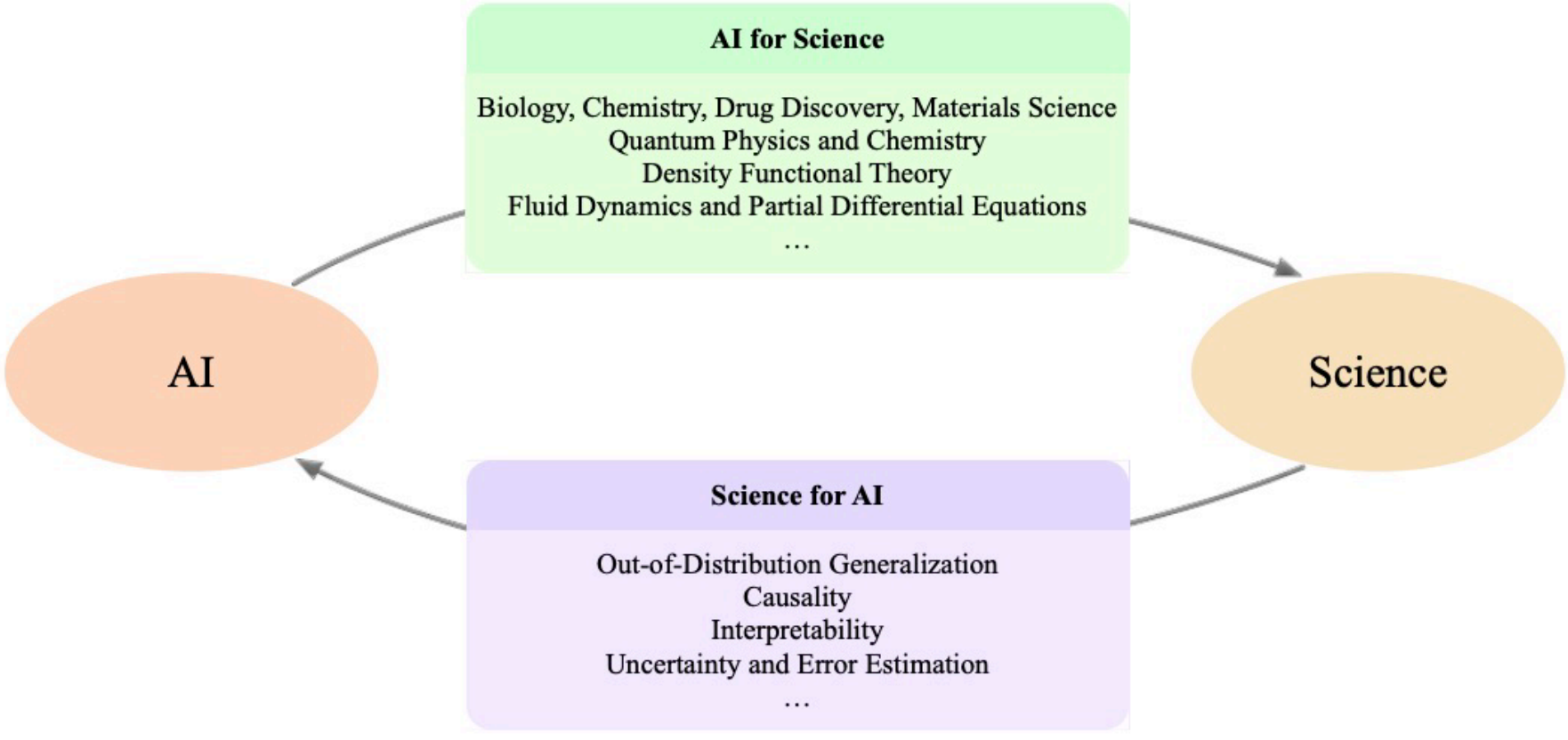
Customized interpretations considering geometries and equivariance would lead to better insights

LLMs for Science: Open Research



One purpose: Broader users and the general public can use scientific product!

Roadmap



- **AI for Science**: Develop new AI methods to accelerate discoveries in science and engineering
- **Science for AI**: Use new data generation and simulation schemes in science (ordinary/partial/controlled differential equations) to advance AI, which has largely relied on observational data