ML/AI for Materials: Introduction and Overview

LANL Computational Condensed Matter Summer School 2026

Ying Wai Li (CAI-1)

June 25, 2026

Los Alamos NATIONAL LABORATORY EST. 1943

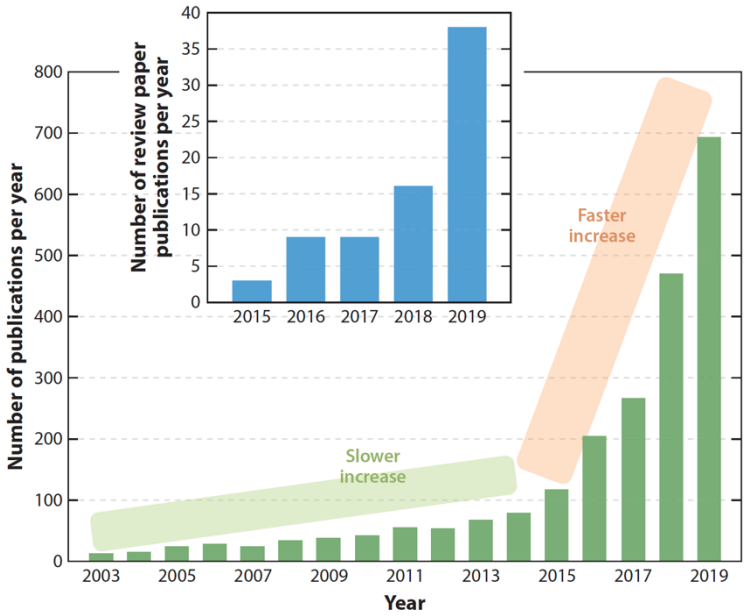
NNSA National Nuclear Security Administration

Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA

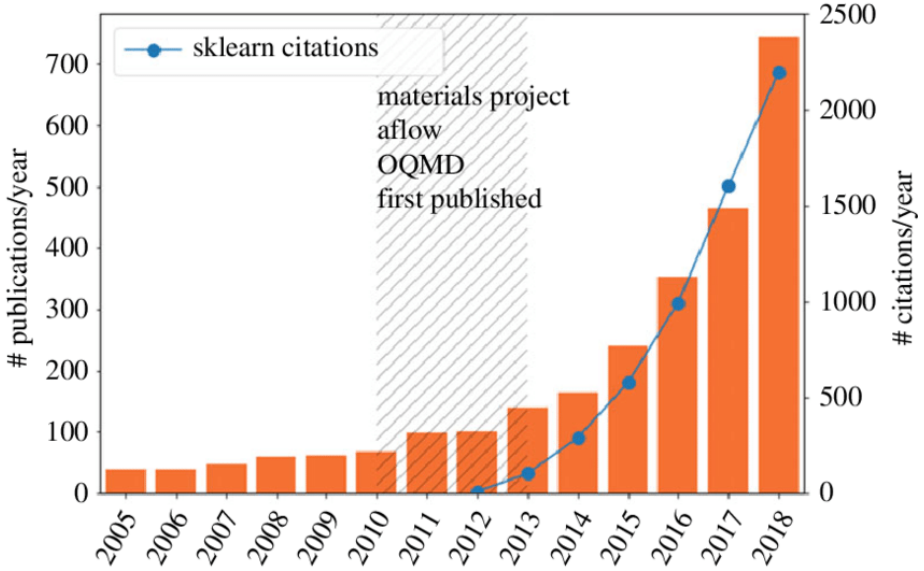
LA-UR-26-25488

Increase in machine learning adoption in materials science

Number of publications per year in materials science:

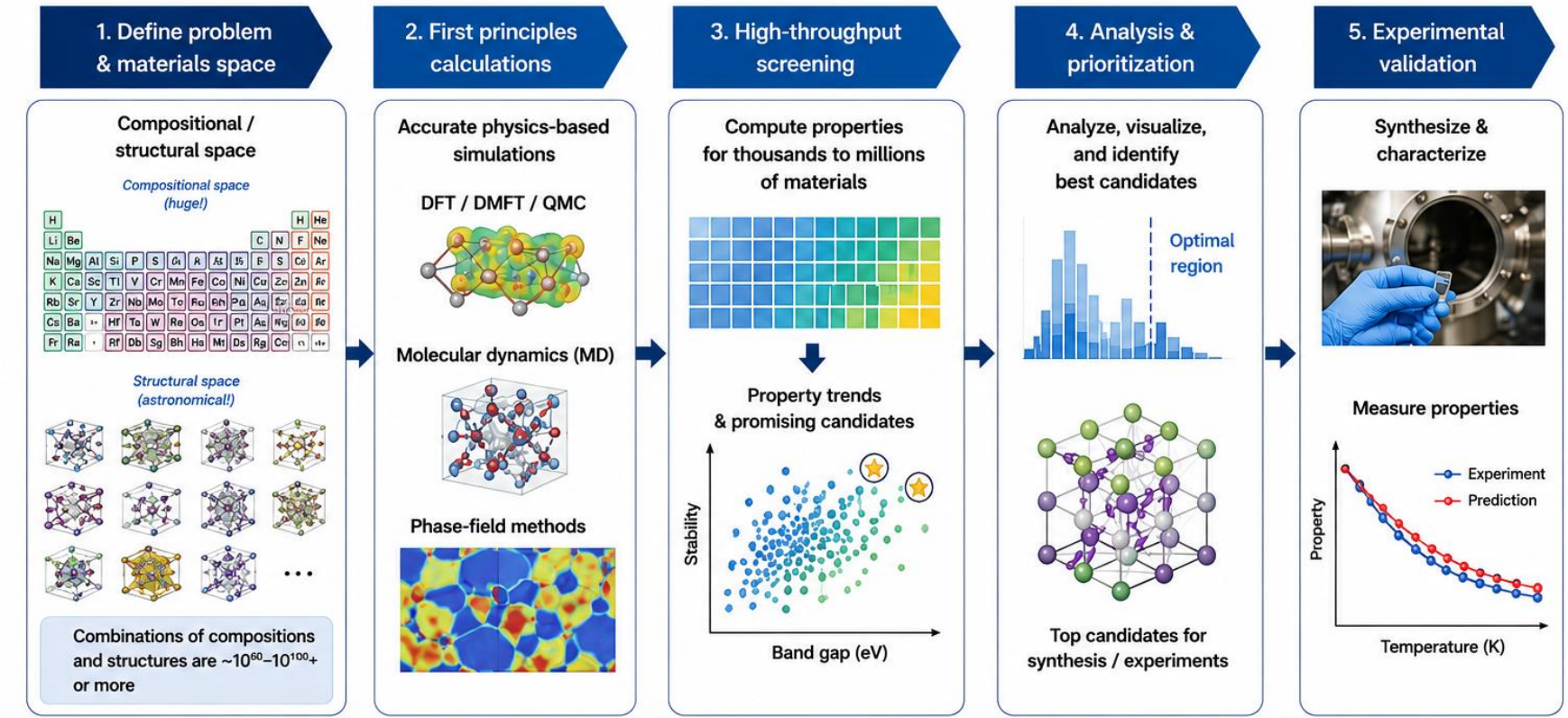


D. Morgan & R. Jacobs,
Annu. Rev. Mater. Res. **50**, 71-103 (2020)



T. Hey et al.,
Phil. Trans. R. Soc. A. **378,** 20190054 (2020)

Traditional computational materials discovery



Credit: graphics created by ChatGPT 5.5. Note the wrong periodic table!

Important materials databases

First-principles computational databases

The Materials Project
materialsproject.org

OQMD
The Open Quantum Materials Database

AFLOW
Automatic - FLOW for Materials Discovery

NOMAD
NOVEL - MATERIALS DISCOVERY LABORATORY
Fritz Haber Institute of the Max Planck Society

JARVIS

scientific data
(a journal)

Experimental/Crystallographic Databases

Inorganic Crystal Structure Database
ICSD

CSD
Cambridge Structural Database

COD
Crystallography Open Database

bilbao crystallographic server

Why ML/AI in condensed matter physics?

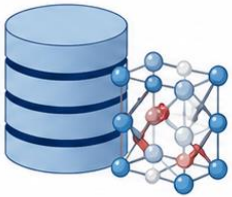
Fundamental bottlenecks in traditional approaches:

1. **Combinatorial explosion**
   - 10^{60}–10^{100} possible materials
   - Impossible to search exhaustively
2. **Computationally expensive simulations**
   - DFT: minutes to hours per structure
   - Beyond-DFT: days to weeks on supercomputers
3. **Data explosion in materials science**
   - Modern experiments generate GB–TB of data
   - Computational databases reaching millions of calculations

AI helps prioritize and accelerate physics-based modeling and experiments; analyze and extract patterns from data

Evolution of ML/AI for materials discovery

1. Materials Databases



2000s

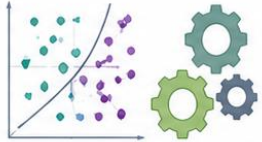
- Emergence of high-throughput computing
- Growth of material databases

2. Materials Informatics

2010s

- Extract insights using data analytics and statistical methods
- Features/descriptors

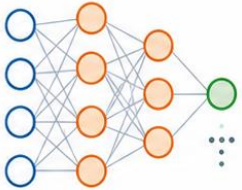
3. Machine Learning



Mid–2010s

- Predict properties
- Classify phases
- Accelerate simulations

4. Neural Networks



Late 2010s

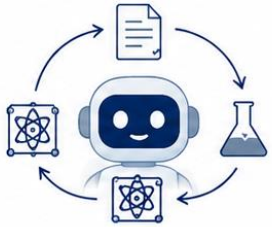
- Learn complex, non-linear relationships from data
- ML interatomic potentials
- Breakthrough in computer vision

5. Foundation Models

Early 2020s

- Big data
- Large pretrained models that can be finetuned for specific tasks

6. Scientific Agents



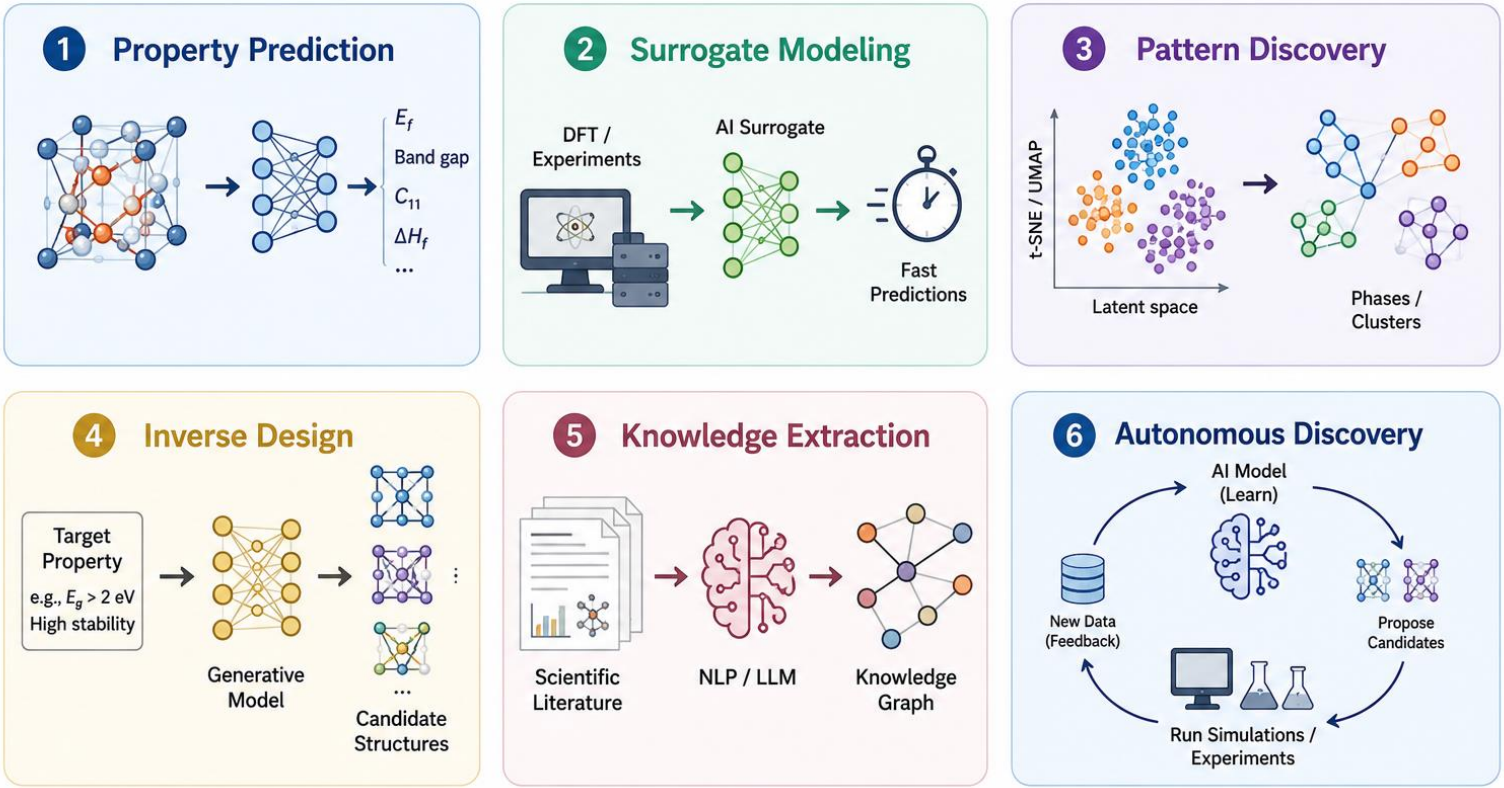
Mid–2020s and beyond

- Autonomous agents for planning and executing workflow
- ”Reasoning” and learning from results

ML/AI is a collection of tools for learning structure-property relationships and accelerating materials discovery in simulation, experiment, and theory

Credit: graphics created by ChatGPT 5.5

Major ML/AI applications in materials science



Credit: graphics created by ChatGPT 5.5

From linear regression to supervised learning

- Linear regression: $y = f(x) = mx + c$

$$x \rightarrow \boxed{f(x)} \rightarrow y$$

- Fitting $f(x)$ amounts to finding m and c with least-square fit
- Supervised learning:

$$x \rightarrow \boxed{ML} \rightarrow y$$

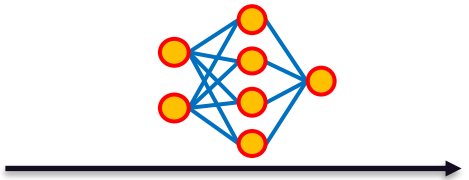
$$x \rightarrow \boxed{\text{  }} \rightarrow y$$

- Training a model:
  fitting of model parameters such that the loss function is minimized

Supervised learning

- Regression: $y = f(x)$
- Classifications
- Given a labeled dataset $D = \{(x_1, y_1), (x_2, y_2), ...\}$, learn the mapping function f

MNIST data: handwritten digits (28x28 pixels)



Labels

0
1
2
3
4
5
6
7
8
9

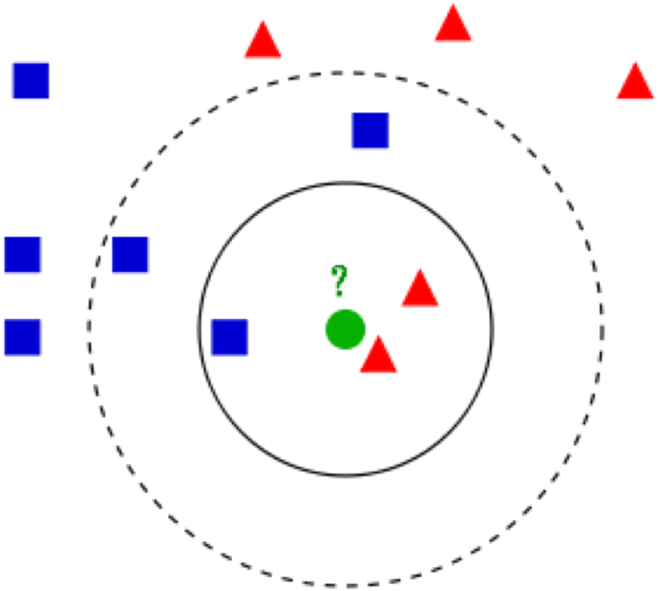
http://yann.lecun.com/exdb/mnist/

Kernel methods

	Polynomial regression	Kernel method
Data	$\boldsymbol{x} = (x_0, x_1, \ldots, x_i, \ldots, x_D)$ $\boldsymbol{x}' = (x'_0, x'_1, \ldots, x'_i, \ldots, x'_D)$	
Feature space	$\boldsymbol{f}(\boldsymbol{x}) = (1, x, x^2, \ldots, x^i, \ldots, x^N)$ $\boldsymbol{f}(\boldsymbol{x}') = (1, x', x'^2, \ldots, x'^i, \ldots, x'^N)$	bypasses features to directly define similarities
Similarities	$\boldsymbol{f}(\boldsymbol{x}) \cdot \boldsymbol{f}(\boldsymbol{x}') = (x^{\mathrm{T}}x')^2$	$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = (x^{\mathrm{T}}x')^2$
Model	$\hat{y}(x) = \sum_{i=0}^{N} w_i f_i(x)$	$\hat{y}(x) = \sum_{j=0}^{M} w_j K(x, x_j)$
	N: number of features	M: number of training data points

Clustering with Kernel methods

- A kernel $K(x_1,x_2)$ measures *similarity* between x_1 and x_2
- An instance-based learners
- Effectively *interpolates* nearby data
- Model ***grows automatically*** with new data



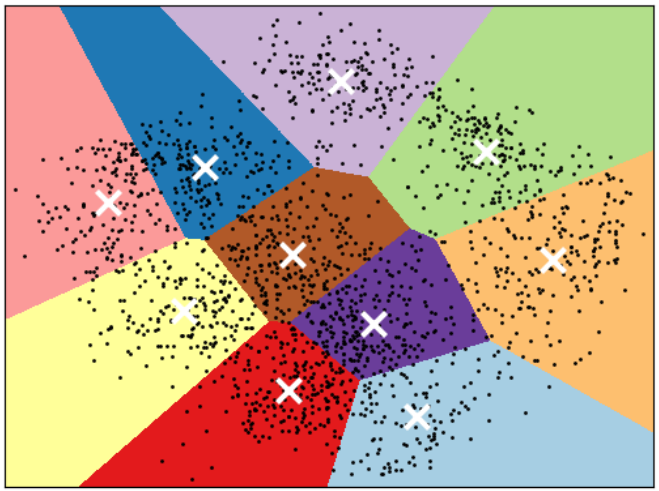
***k*-NN algorithm**:
Majority vote of *k*-nearest neighbors

k = 3: y = red
k = 5: y = blue

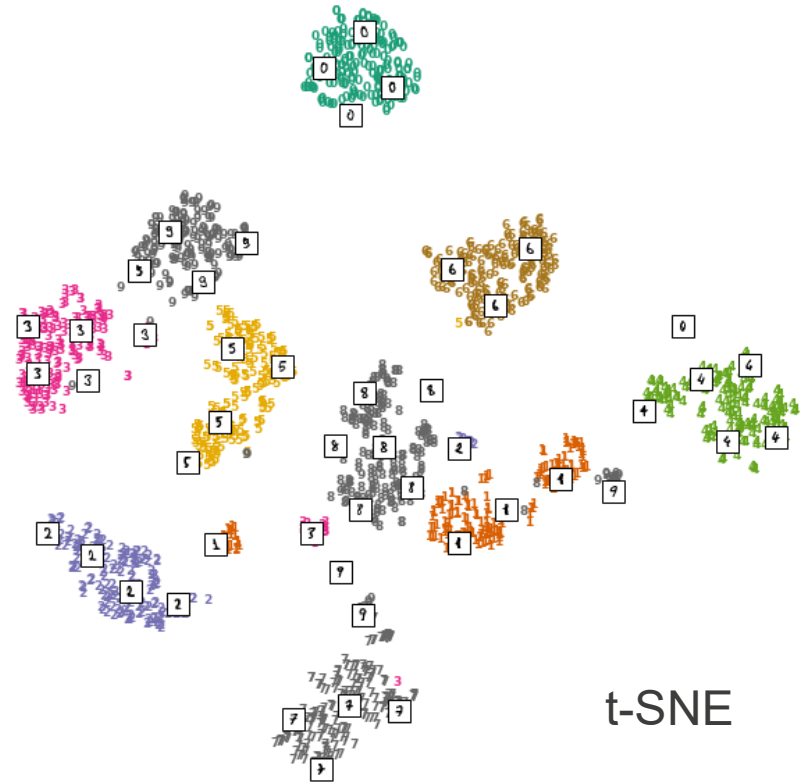
Methods belonging to this class
- Support Vector Machines
- Kernel Ridge Regression
- Gaussian Process Regression

Unsupervised learning

- Dimensionality reduction algorithms; manifold learning (e.g. t-SNE)
- Principal Component Analysis
- Clustering (e.g. k-means)



k-means

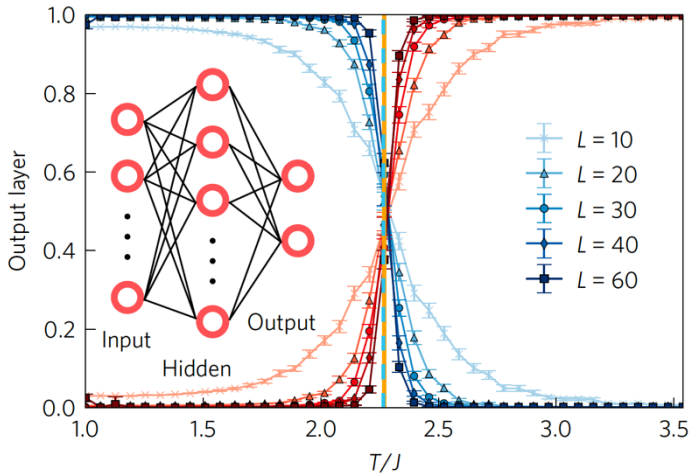


t-SNE

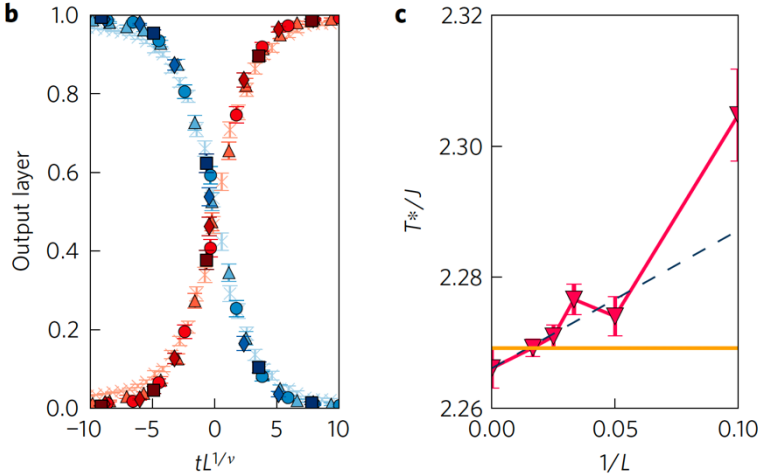
https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html

Applications of supervised learning: Identification of phases and phase transitions of matter

- Classifying “phases” in classical Ising spin models
- Supervised learning with fully-connected or convolutional neural network
- Possible when configurations from all different phases are available
- Could be useful when order parameters are not known (e.g. topological phase transitions)



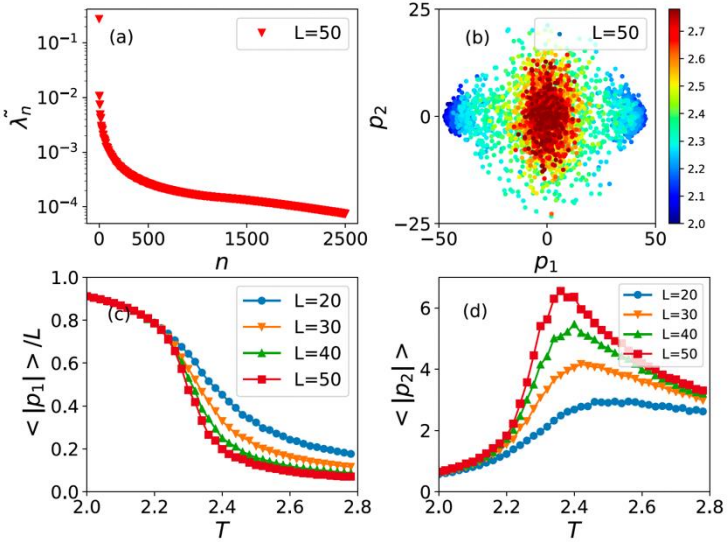
J. Carrasquilla and R.G. Melko, Nature Physics **13**, 431-434 (2017)



E.P.L. van Nieuwenburg et al., Nature Physics **13**, 435-440 (2017)

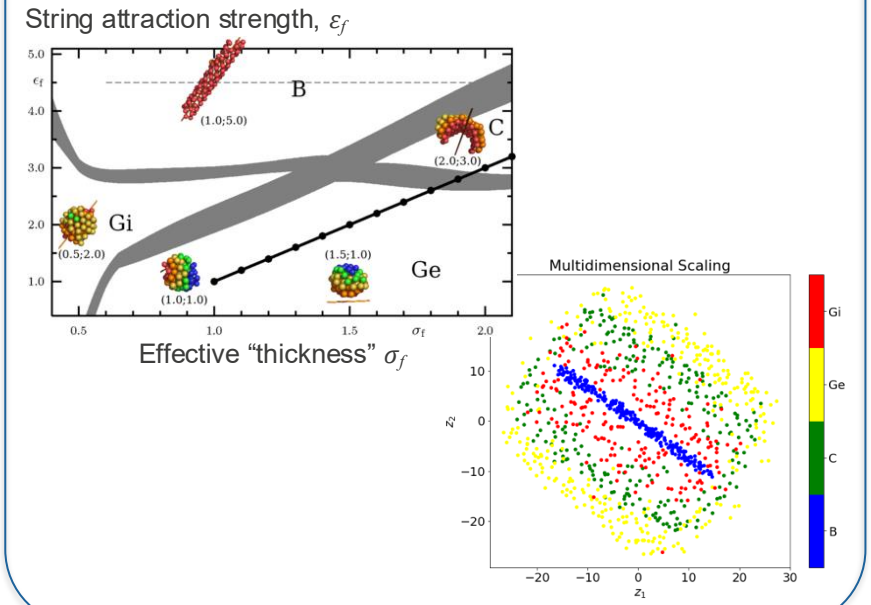
Applications of unsupervised learning: identifying phase transitions

- Classifying phases in classical spin models
- Principal Component Analysis (PCA) identifies order parameters for fully frustrated systems; distinguishes between crossovers, first- and second-order transitions



W. Hu, R. R. P. Singh, and R. T. Scalettar. Phys. Rev. E **95,** 062122 (2017).

- PCA, multidimensional scaling, *t*-SNE, isomap, etc. identify structural "phases" in polymers in similarly as human-defined "order parameters"



Q. Parker, D. Perera, **YWL**, and T. Vogel. Phys. Rev. E **105**, 035304 (2022).

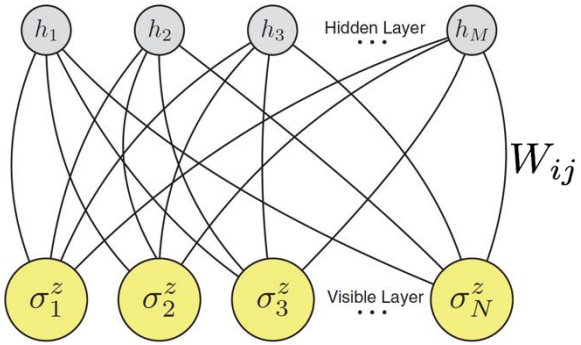
Applications of supervised learning: quantum many-body wavefunction

- Variational representation of a quantum wavefunction using an artificial neural network (restricted Boltzmann machine)
- Able to find ground state and time-dependent states of a quantum Hamiltonian
- Neural-network quantum states (NQS):

$$\Psi_M(\mathcal{S};\mathcal{W}) = \sum_{\{h_i\}} \exp\left(\sum_j a_j \sigma_j^z + \sum_i b_i h_i + \sum_{ij} W_{ij} h_i \sigma_j^z\right)$$

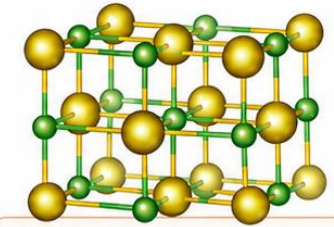
$\mathcal{S} = \sigma_i^z, ..., \sigma_N^z$ ← spin configurations (input)

$\mathcal{W} = \{a, b, W\}$ ← to be trained by reinforcement learning



Arbitrary precision can be obtained by increasing the number of hidden units h

Materials representations (“descriptors”, “fingerprints”)



Na⁺

Cl⁻

1 Composition features

Na Cl

1 : 1

x = [1, 1, 11, 17, …]

fractions, atomic number, electronegativity, ionic radius, etc.

2 Hand-crafted descriptors

SOAP ACSF Coulomb Matrix

d = [0.21, 0.34, …, 0.05]

encoding local / atomic arrangements

3 Graph representation

Na Atoms → nodes

Cl Bonds → edges

captures local geometry (tetrahedra), compositions and connectivity explicitly

4 Local environment representation

r_1 r_2 r_3 … r_M

radial / basis expansion around each atom

5 Learned embeddings

z = [0.12, −0.34, 0.51, …]

learned automatically from data

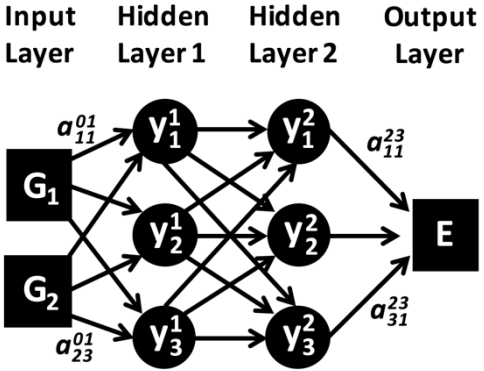
More human-designed ⟵ ⟶ More data-driven

Credit: graphics created by ChatGPT 5.5

Interatomic potentials for atomistic simulations

1. Neural networks (NN)
   - First ML method applied to fit DFT potential energy surfaces [1]
   - "Behler-Parrinello" NN, deep NN [2]

$$E = f\left(b_1^3 + \sum_{k=1}^{3} a_{k1}^{23} \cdot f\left(b_k^2 + \sum_{j=1}^{3} a_{jk}^{12} \cdot f\left(b_j^1 + \sum_{i=1}^{2} G_i \cdot a_{ij}^{01}\right)\right)\right)$$



2. Kernel-based ML potentials
   - Gaussian Approximation Potentials (GAPs) [3]

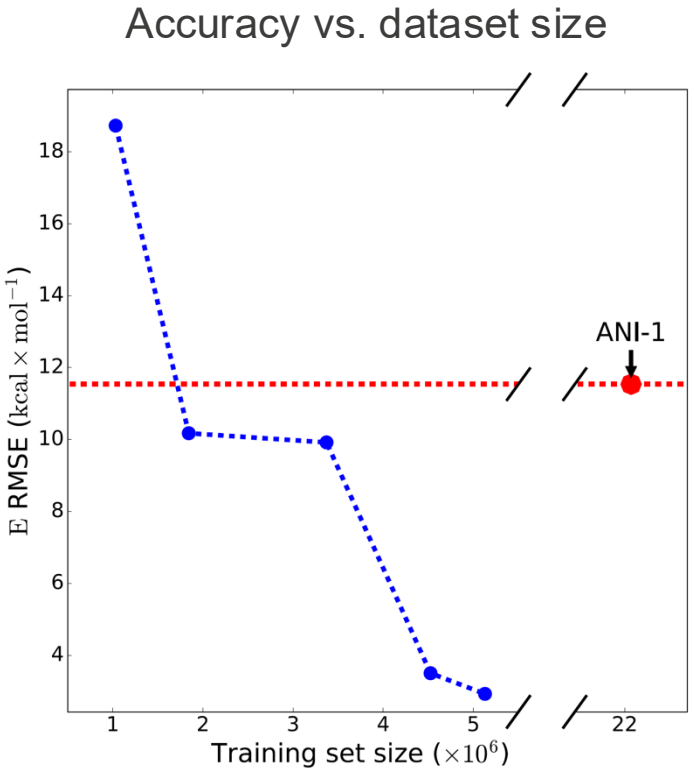
$$E = \sum_{i=1}^{N_{\text{atoms}}} E_i \quad , \qquad E_i = \sum_{j=1}^{N_{\text{train}}} \alpha_j K(\mathbf{d}, \mathbf{d}^j) \quad , \qquad K_{\text{SE}}(\mathbf{d}, \mathbf{d}') = \exp\left(-\sum_j \frac{(d - d')^2}{\theta_j}\right)$$

3. Descriptor-based ML potentials
   - SNAP, SOAP, ACE, MACE…

[1] T.B. Blank et al., J. Chem. Phys. **103**, 4129 (1995).
[2] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 1464-1 (2007); J. Behler, J. Chem. Phys. **145**, 170901 (2016).
[3] A.P. Bartok et al., Phys. Rev. Lett. **104**, 136403 (2010).

Active learning for data acquisition

- Quality of data determines model accuracy and training efficiency of ML potentials
- Developing **diverse, representative** datasets is a challenge for any ML problem
- ANI-1x, ANI-1ccx in organic chemistry[1]:
  - Efficient sampling using active learning reduces dataset size by 4x yet achieving higher accuracy
  - "Query-by-committee"[2] is a widely-used method for acquiring new training data
- Reminiscent to reinforcement learning



J. S. Smith et al., Nat. Comm. 12, 1257 (2021).

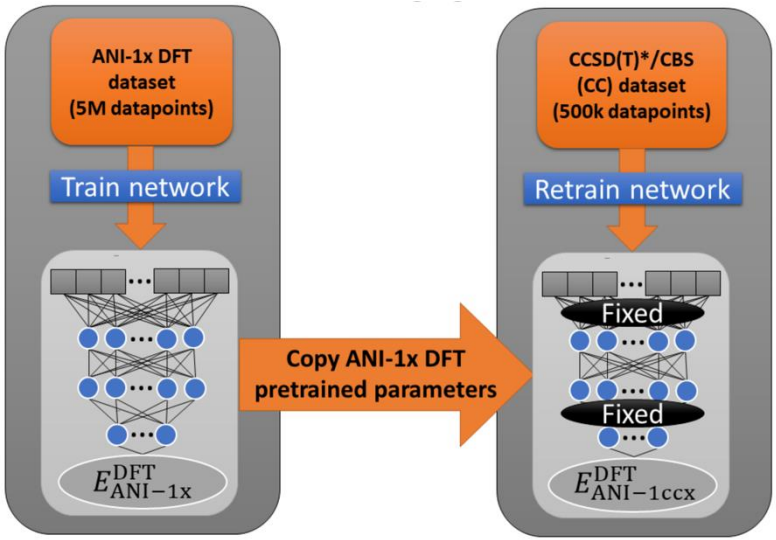
[1] J. S. Smith et al., Sci. Data 7, 134 (2020)

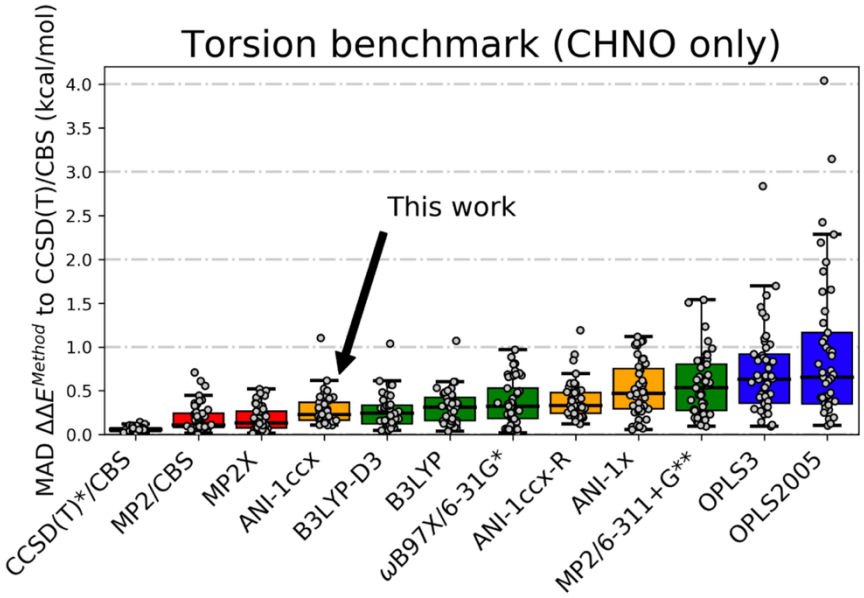
[2] H. S. Seung et al., In Proceedings of COLT'92, pp. 287–294 (1992).

Transfer learning for model finetuning

Training neural network potentials with transfer learning

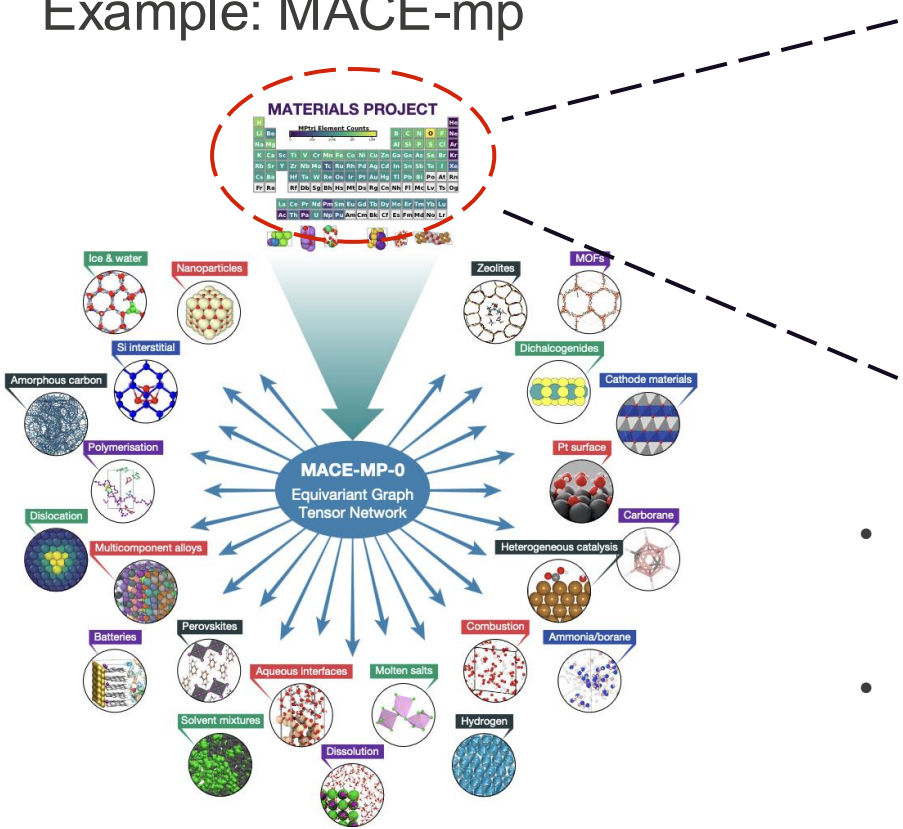
- NN first trained with lots of “cheap” data from density functional theory
- Retrained with a “small” number of high accuracy CCSD(T) data
- Obtained quantum chemistry accuracy

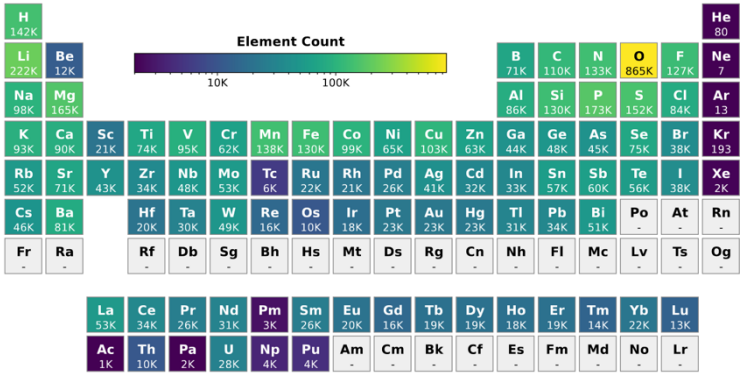




Foundation models for materials

Example: MACE-mp



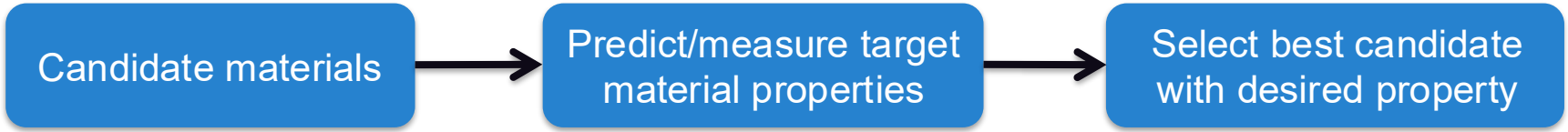


- Instead of training a new model for every property or every material, train a large, general model on massive datasets
- Adapted to downstream tasks via transfer learning

I. Batatia et al., arXiv: 2401.00096 (2024).

Materials inverse design

Traditional forward design of materials:



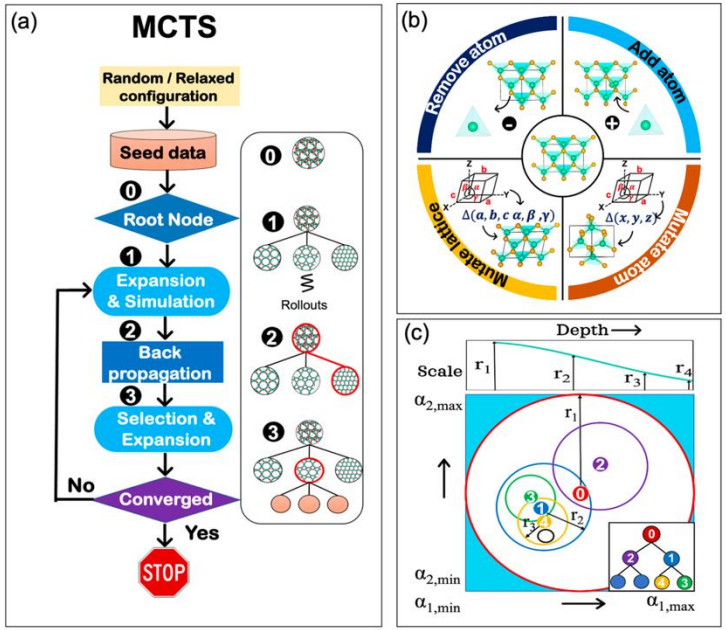
Drawback: High-throughput *ab initio* calculations is an exhaustive search in a massive design space → high computational cost

Can we design a material by starting from a desired property?

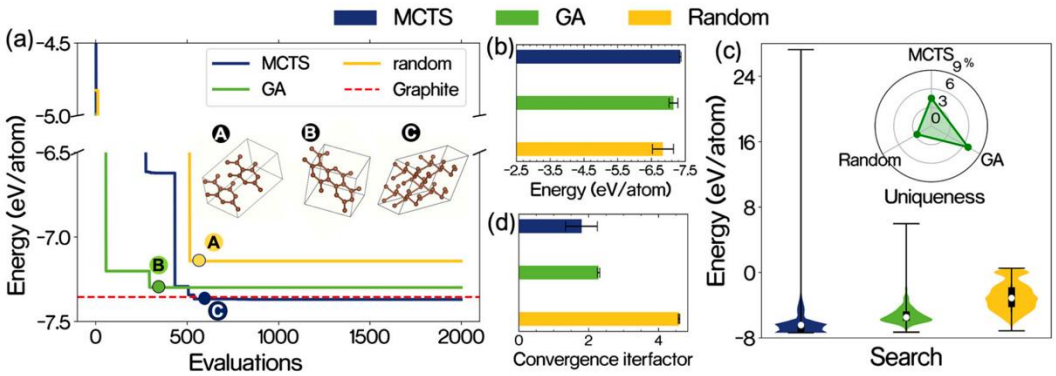
Intelligent search algorithms for inverse materials design

Crystal structure optimization using Monte Carlo tree search algorithm

MCTS algorithm



Structure prediction for carbon polymorphs compared with genetic algorithm (GA) and random search

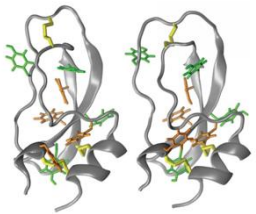


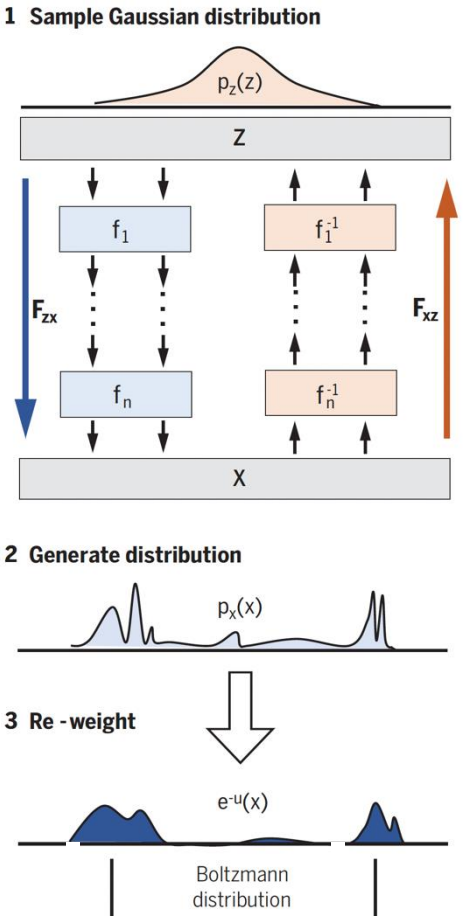
S. Banik et al., NPJ Computational Materials **9**, 177 (2023).

Learning distributions and generating molecules

Example: Boltzmann generators

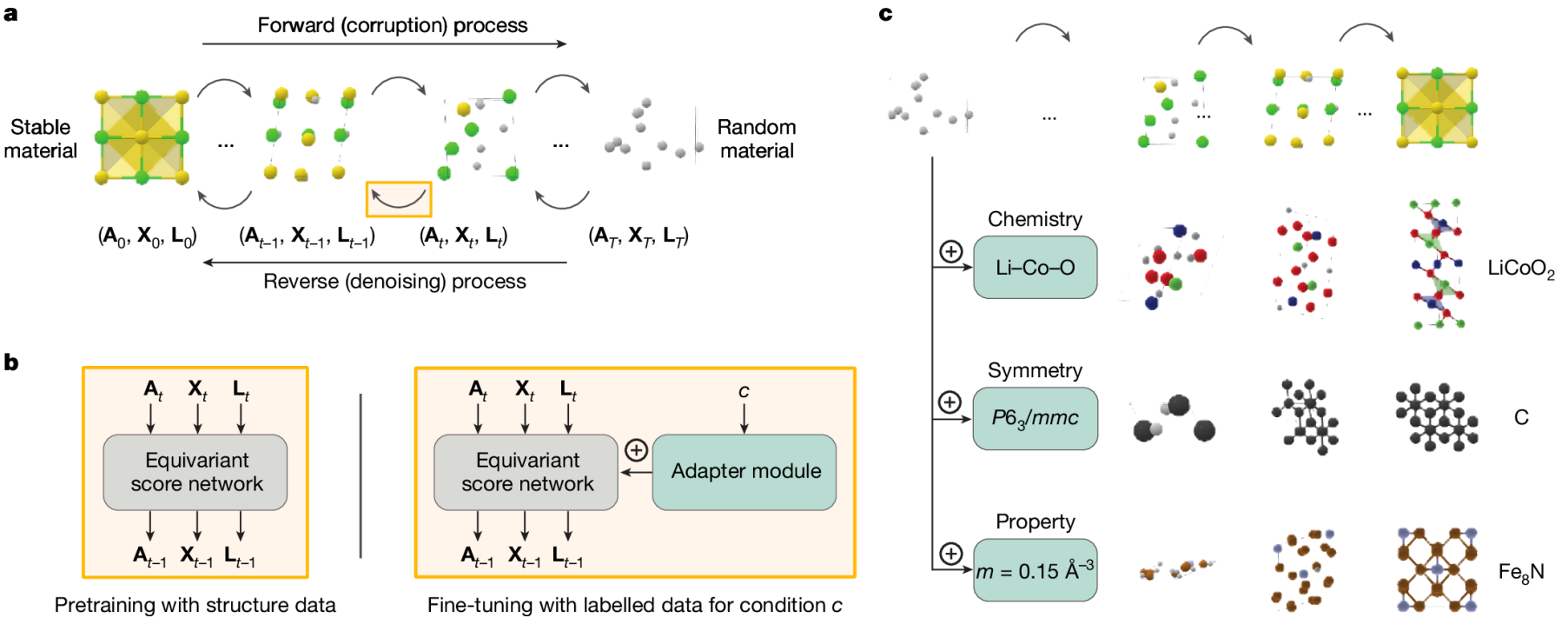
- Goal: sampling (long-lived) equilibrium states
- Learn series of transformations to transform original, rugged distribution into smooth one (e.g. Gaussian)
- Sample the smooth distribution
- Transform back to the original distribution to yield equilibrium states





Diffusion models for inverse materials design

Example: MatterGen



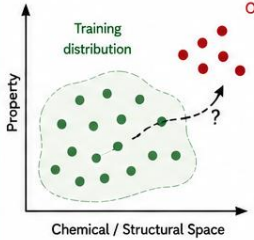
Limitations and open challenges

1 Data quantity and quality

- Limited data
- Experimental noise
- DFT approximation errors
- Data imbalance
- Missing labels and missing physics

Models are only as good as the data they learn from.

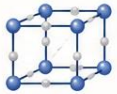
2 Out-of-distribution generalization



- Interpolates well
- Extrapolates poorly
- New chemistries
- Extreme conditions
- Defects, interfaces, surfaces

Real discovery requires reliable predictions beyond the training domain.

3 Physical consistency



Symmetry
$E(R) = E(gR)$

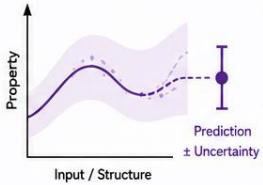
Conservation laws
$\sum_i F_i = 0$

Thermodynamics
$\Delta G < 0$

- Rotational and translational symmetry
- Conservation of energy, momentum, charge
- Thermodynamic stability
- Charge neutrality and boundary conditions

Predictions must obey known physical laws and symmetries.

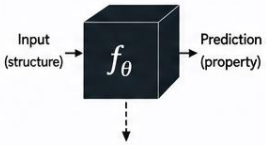
4 Uncertainty quantification



- Confidence estimates are essential
- Active learning
- Experimental planning
- Risk assessment

AI predictions should come with quantified uncertainty.

5 Interpretability and causality



- Why did the model predict this?
- Which features or patterns matter?
- Correlation ≠ causation
- Identify causal mechanisms

Understanding and causal insight are crucial for scientific discovery.

6 Reliable AI for science

- ✕ Fabricated references
- ✕ Incorrect equations
- ✕ Misinterpreted physical concepts
- ✕ Overconfident explanations

LLMs can generate plausible but incorrect scientific information. Human verification remains essential.

Credit: graphics created by ChatGPT 5.5

Take home messages

- ML/AI are now core research tools in data analysis, constructing surrogate models, data-driven information discovery, etc.
- Foundation models and scientific agents are emerging rapidly
- Try to understand the underlying principles of ML algorithms instead of using it as a black box
- Physics remains essential!

- Always ask yourself:
  - **Is ML necessary to perform the task?**
  - **Can ML be beneficial over traditional methods? (Computational cost, accuracy, generality, etc.)**
  - **Can I build in more physics insights into the model/ML workflow?**
  - **Do the results align with physical insights?**