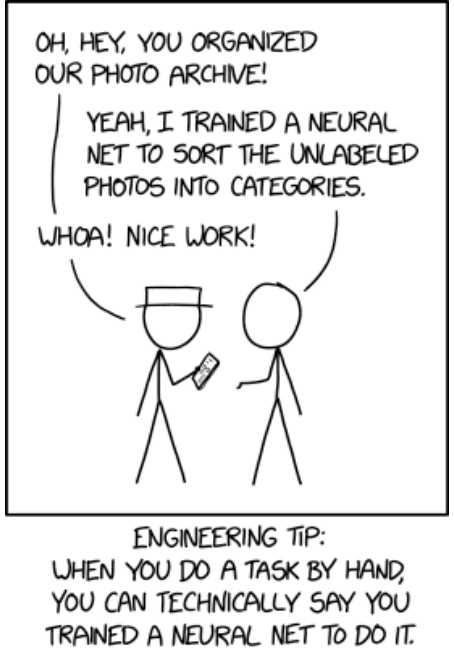


Using machine learning to advance materials design

Sai Gautam Gopalakrishnan

Materials Engineering, Indian Institute of Science

saigautamg@iisc.ac.in; https://sai-mat-group.github.io

IIT Alumni Centre Bengaluru
Jan 18, 2025

Acknowledgments

Group picture in Jan 2025

SERC (IISc)

Param Utkarsh+
Param Siddhi
(CDAC)

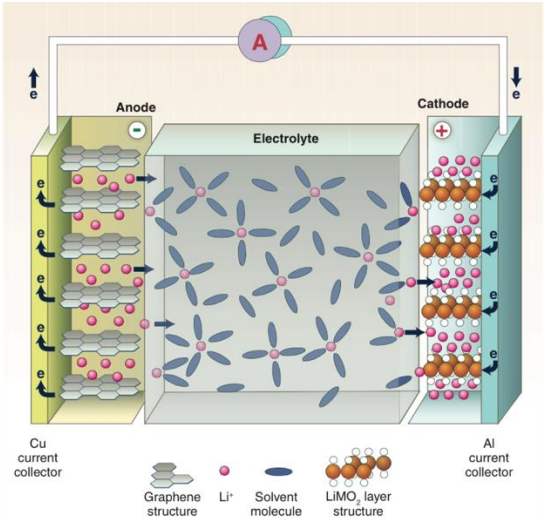
Fugaku (Japan)

Jureca (Germany)

Archer (UK)

Why bother about materials science?

Key performance bottlenecks in key applications: governed by materials used

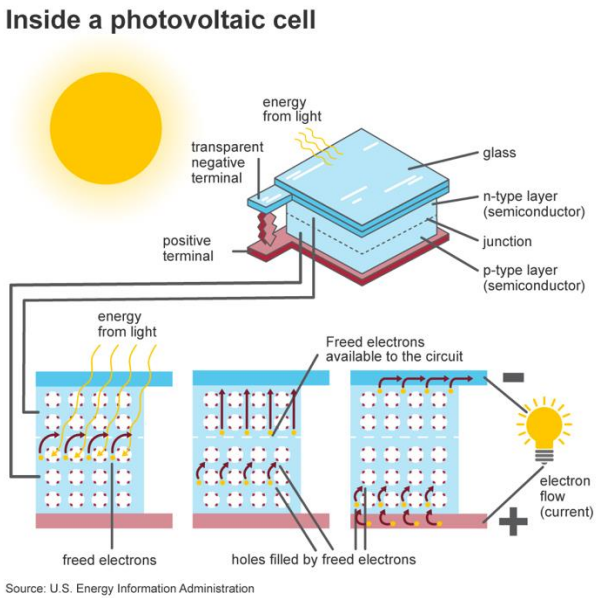


B. Dunn et al., Science 2011

Energy and power density of a battery: limited by materials used as electrodes (and at times, electrolytes)

Key material properties: stability, ionic mobility, reaction energies

Usage of better materials → better performance

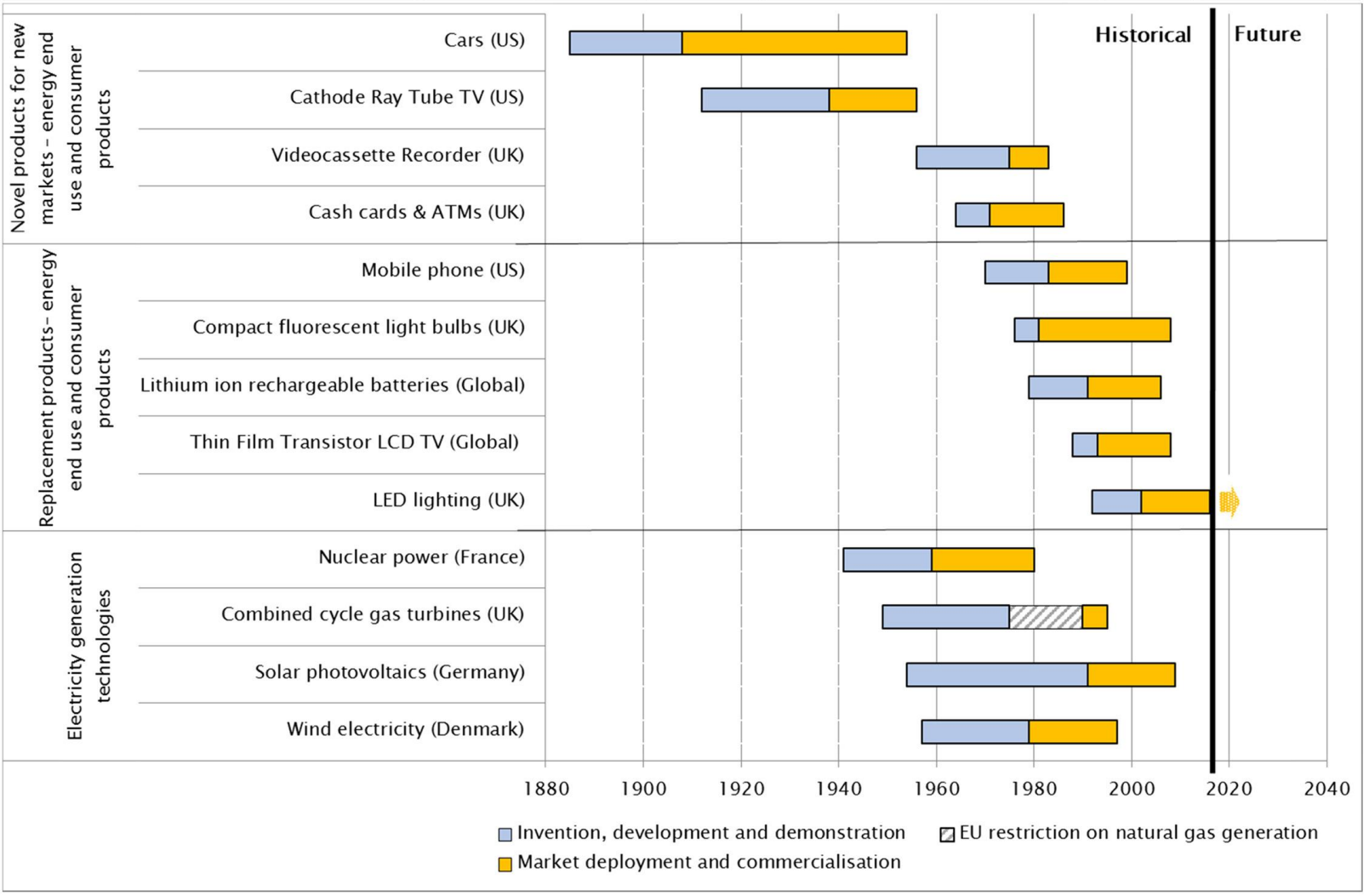


Efficiency of a photovoltaic: choice of semiconductor used as the light absorber

Key material properties: band gap, stability, resistance to point defects

Why use machine learning (ML) in materials science?

Technological innovation and deployment is a ‘slow’ process: often limited by materials

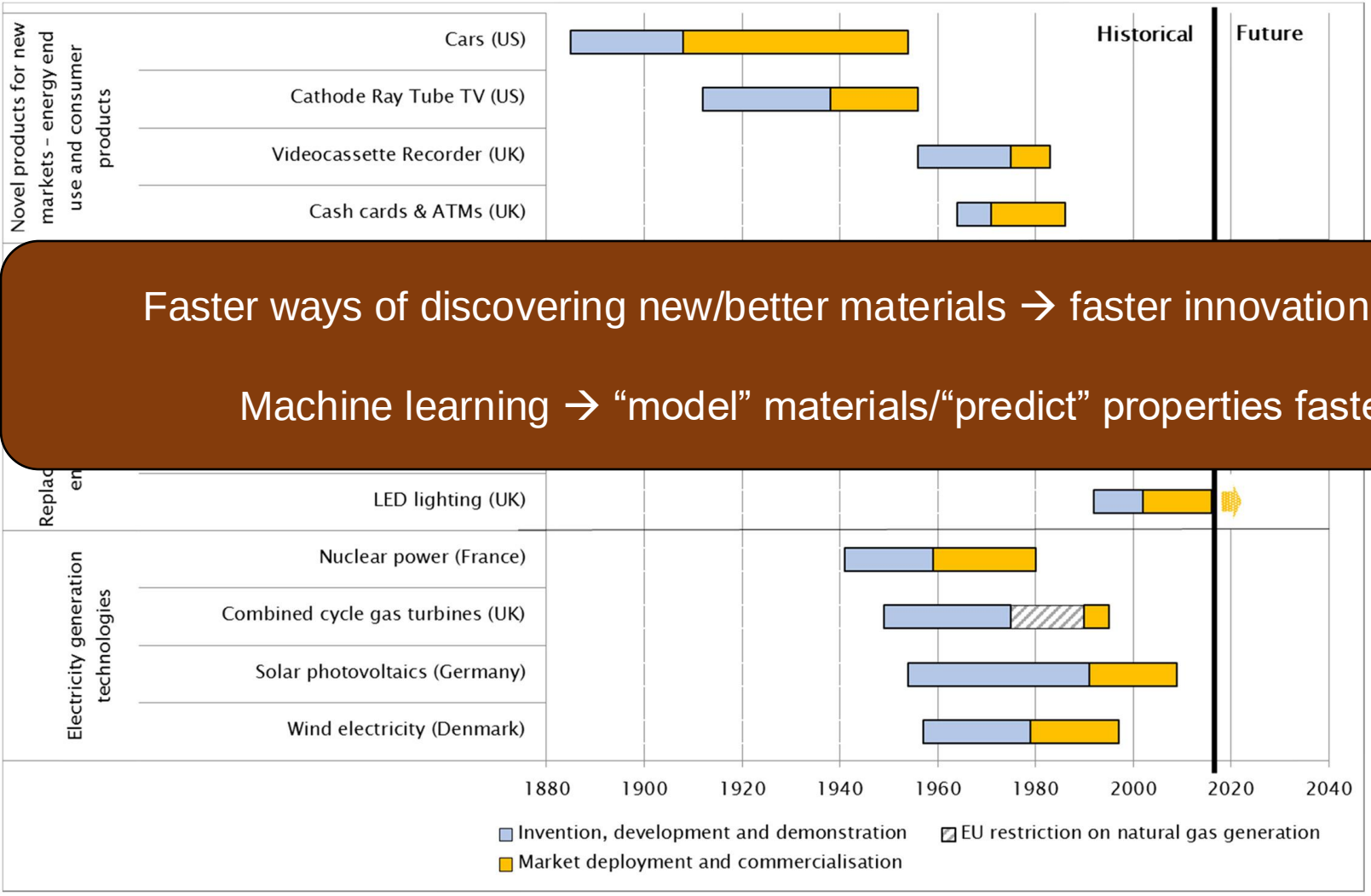


Innovation is particularly slow in energy generation sector!

Gross et al., Energy Policy 123, 682-699 (2018)

Why use machine learning (ML) in materials science?

Technological innovation and deployment is a ‘slow’ process: often limited by materials



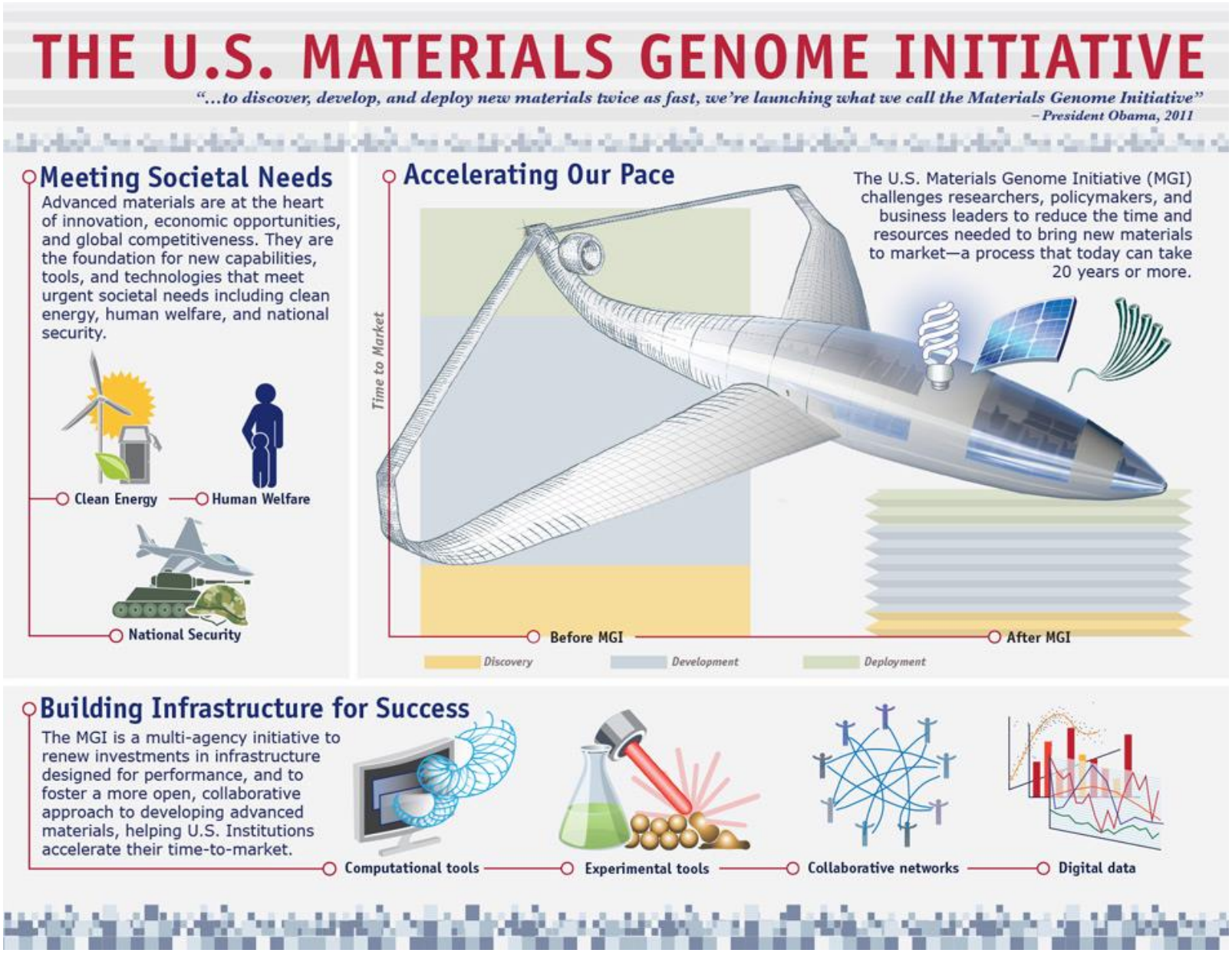
Faster ways of discovering new/better materials → faster innovation cycles

Machine learning → “model” materials/“predict” properties faster

Innovation is particularly slow in energy generation sector!

Gross et al., Energy Policy 123, 682-699 (2018)

Materials Genome (2011-present)



Evolution of ‘modelling’ in materials science

On the determination of molecular fields. —II. From the equation of state of a gas

J. E. Jones

Published: **01 October 1924** | https://doi.org/10.1098/rspa.1924.0082

1924

1957

RESEARCH ARTICLE | AUGUST 13 2004

Phase Transition for a Hard Sphere System

Special Collection: JCP 90 for 90 Anniversary Collection

B. J. Alder; T. E. Wainwright

Check for updates

J. Chem. Phys. 27, 1208–1209 (1957)

https://doi.org/10.1063/1.1743957 **Article history**

Inhomogeneous Electron Gas

P. Hohenberg and W. Kohn

Phys. Rev. **136**, B864 – Published 9 November 1964

1964

1975

Clustering and ordering in solid solutions

D. de Fontaine

Computer simulation of local order in condensed phases of silicon

Frank H. Stillinger and Thomas A. Weber

Phys. Rev. B **31**, 5262 – Published 15 April 1985; Erratum Phys. Rev. B **33**, 1451 (1986)

1986

1996

Generalized Gradient Approximation Made Simple

John P. Perdew, Kieron Burke, and Matthias Ernzerhof

Phys. Rev. Lett. **77**, 3865 – Published 28 October 1996; Erratum

From ultrasoft pseudopotentials to the projector augmented-wave method

G. Kresse and D. Joubert

Phys. Rev. B **59**, 1758 – Published 15 January 1999

1999

2003

Shock Waves in High-Energy Materials: The Initial Chemical Events in Nitramine RDX

Alejandro Strachan, Adri C. T. van Duin, Debashis Chakraborty, Siddharth Dasgupta, and William A. Goddard, III

Phys. Rev. Lett. **91**, 098301 – Published 28 August 2003

Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces

Jörg Behler and Michele Parrinello

Phys. Rev. Lett. **98**, 146401 – Published 2 April 2007

2007

Predicting Crystal Structures with Data Mining of Quantum Calculations

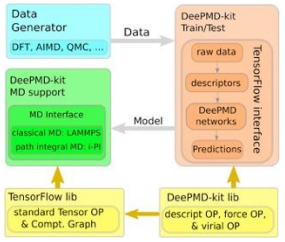
Stefano Curtarolo, Dane Morgan, Kristin Persson, John Rodgers, and Gerbrand Ceder

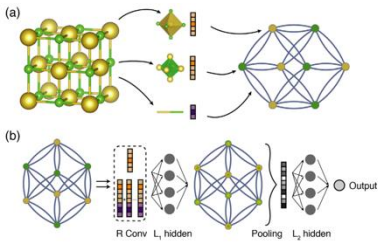
Phys. Rev. Lett. **91**, 135503 – Published 24 September 2003

THE U.S. MATERIALS GENOME INITIATIVE

2011-2018

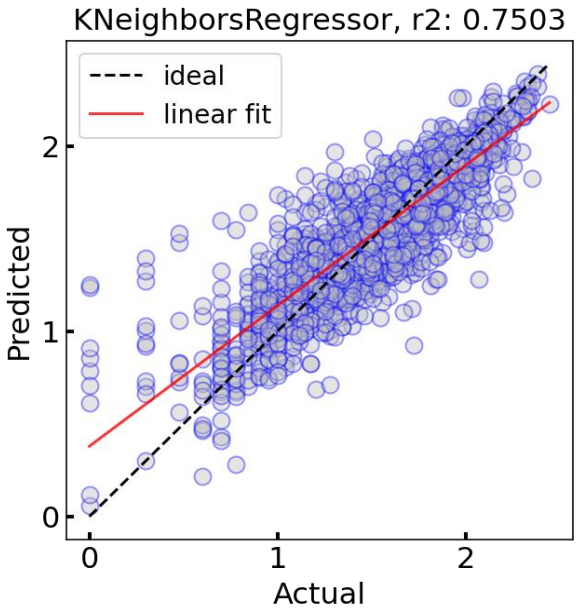
2018-present



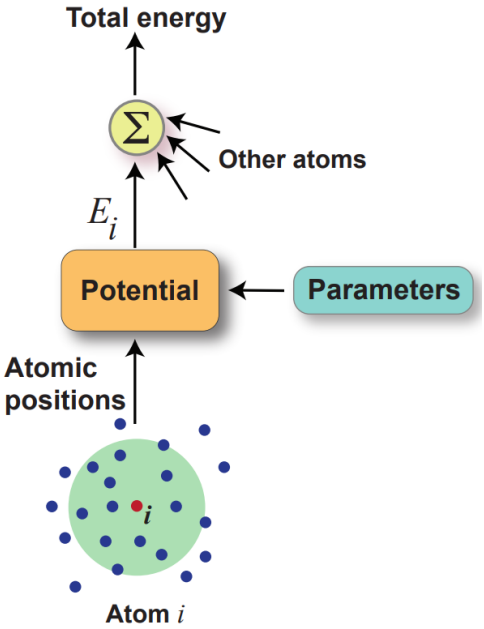


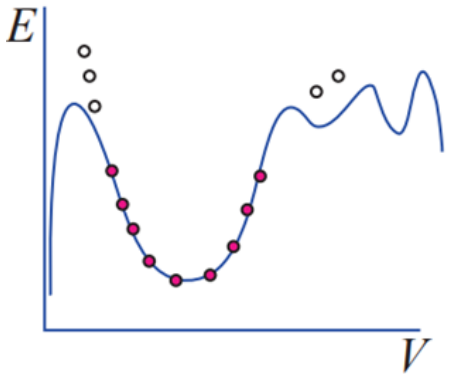
Types of ML in materials science

Regressions: make property predictions better with ‘simple’ inputs (also classifications)

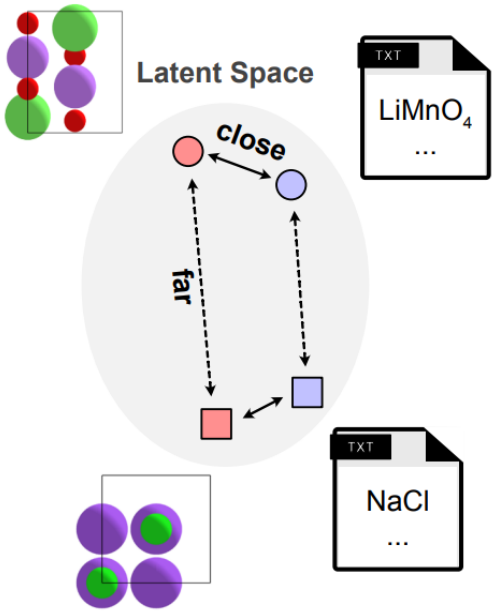


Interatomic potentials: describe potential energy surface accurately

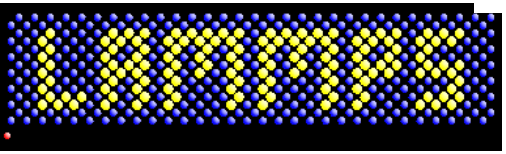


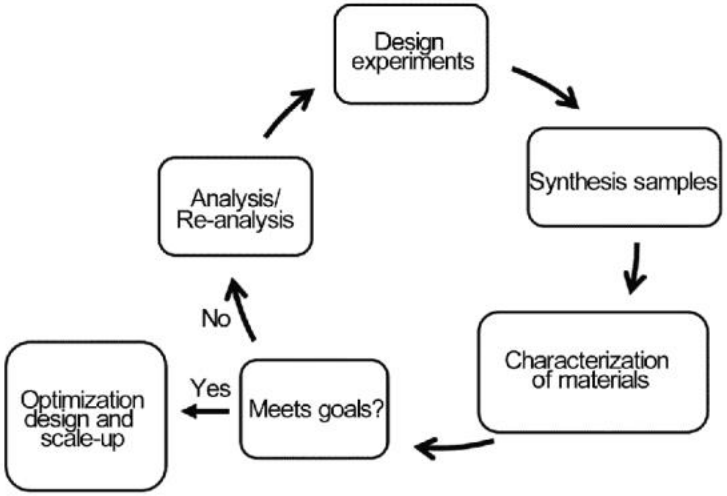


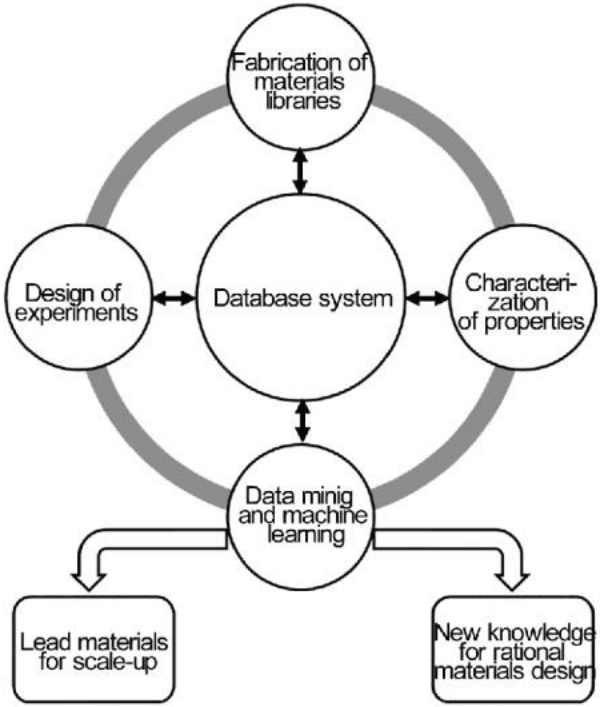
Advanced topics: Diffusion (generative) models, language models, transfer learning



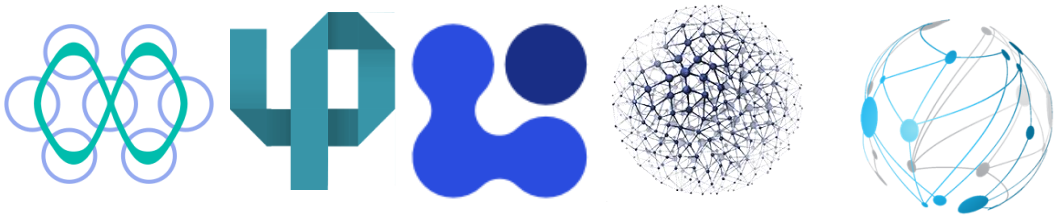
Where does the data come from?



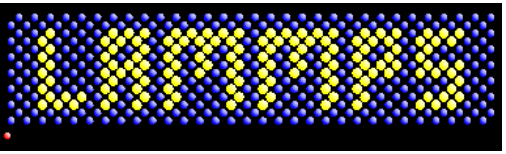




Liu et al., Sci. China Tech. Sci. 62, 4 (2019)



Where does the data come from?



Home

Search

Home | Benchmark Info | Full Benchmark Data | How To Use | Leaderboards Per Task | Reference

Home

Leaderboard-Property: General Purpose Algorithms on `matbench_v0.1`

Find more information about this benchmark on the benchmark info page

Task name	Samples	Algorithm	Verified MAE (unit) or ROCAUC	Notes
matbench_steels	312	MODNet (v0.1.12)	87.7627 (MPa)	
matbench_jdft2d	636	MODNet (v0.1.12)	33.1918 (meV/atom)	
matbench_phonons	1,265	MegNet (kgcnn v2.1.0)	28.7606 (cm^-1)	structure required
matbench_expt_gap	4,604	MODNet (v0.1.12)	0.3327 (eV)	
matbench_dielectric	4,764	MODNet (v0.1.12)	0.2711 (unitless)	
matbench_expt_is_metal	4,921	AMMExpress v2020	0.9209	
matbench_glass	5,680	MODNet (v0.1.12)	0.9603	
matbench_log_gvrh	10,987	coNGN	0.0670 (log10(GPa))	structure required
matbench_log_kvrh	10,987	coNGN	0.0491 (log10(GPa))	structure required
matbench_perovskites	18,928	coGN	0.0269 (eV/unit cell)	structure required
matbench_mp_gap	106,113	coGN	0.1559 (eV)	structure required
matbench_mp_is_metal	106,113	CGCNN v2019	0.9520	structure required
matbench_mp_e_form	132,752	coGN	0.0170 (eV/atom)	structure required

https://matbench.materialsproject.org/

Data organization: python/API

ML: python

ICSD

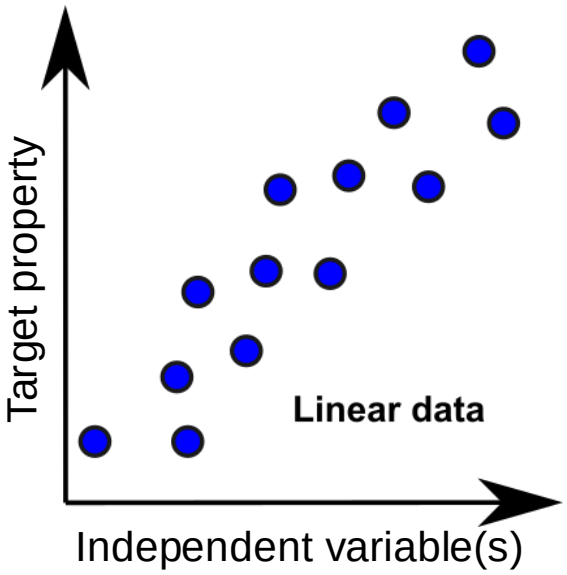
NIST National Institute of Standards and Technology

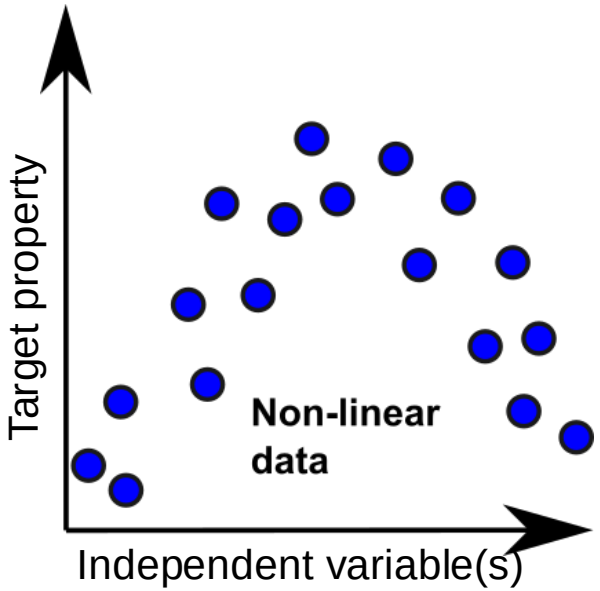
HTEM

Classic machine learning models and use cases in materials

Linear and non-linear models

Relationship of target data can be linear/non-linear with underlying independent variables (descriptors)





Linear regression/linear model works best

$$y = b + \sum_i a_i x_i$$

Popular models:

- Linear regression (RMSE reduction)
- LASSO regression (L_1 norm)
- Ridge regression (L_2 norm)

Non-linear regression/non-linear model works best

$$y = b + \sum_i f(a_i, x_i)$$

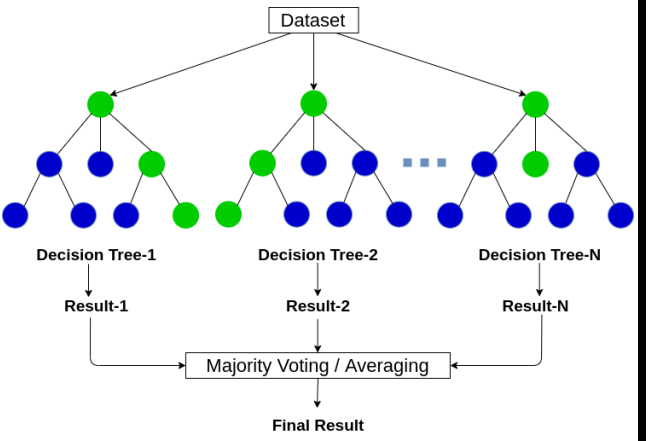
Popular models:

- Random forest
- Support vector machine (SVM)
- K-nearest neighbors (KNN)
- Neural networks*

Overview of non-linear (simple) models

Most non-linear models can be used both for regression and classification

Random forest



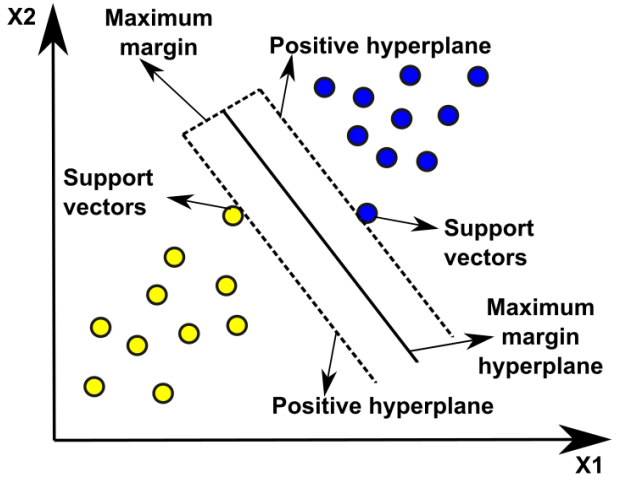
Ensemble model: final decision is an average of several trees

Each tree: if-else decisions

- Handle noisy and 'large' data
- Resistant to overfitting
- Less sensitive to training data

- May not be interpretable
- Computationally slow for 'large' datasets

Support vector machine

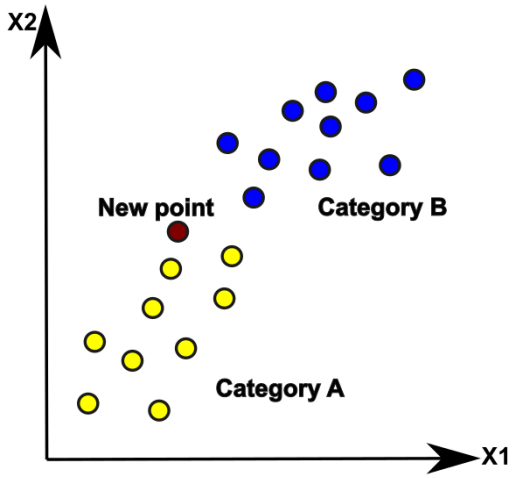


Identify hyperplanes that separate data into clusters

- Efficient at identifying key descriptors in high-dimensional space
- Memory efficient

- Sensitive to noise in data

K-nearest neighbors



Uses feature similarity (i.e., 'distance' from other points) to make predictions about unseen data

- Easy to implement
- Resistant to noisy data

- Memory inefficient (needs to store entire training data)

(Classic) machine learning in action: predicting vacancy formation

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS Article

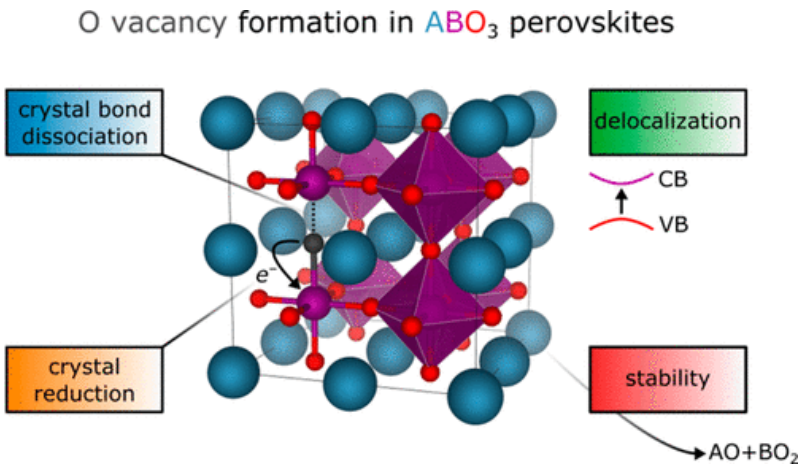
Factors Governing Oxygen Vacancy Formation in Oxide Perovskites

Robert B. Wexler, Gopalakrishnan Sai Gautam, Ellen B. Stechel, and Emily A. Carter*

Cite This: *J. Am. Chem. Soc.* 2021, 143, 13212–13227 Read Online

- ABO_3 perovskites
  - A= Ca, Sr, Ba, La, or Ce
  - B= Ti, V, Cr, Mn, Fe, Co, or Ni
- **Database**: 341 Datapoints obtained from density functional theory (DFT) calculations

- **Model**: A simple linear model with physically intuitive descriptors
  - Crystal bond dissociation energy
  - Crystal reduction potential
  - Band gaps
  - Energy above hull
- **Performance**:
  - Mean absolute error (MAE) - 0.45 eV
  - $BiFeO_3$ and $BiCoO_3$ identified as viable candidates for solar thermochemical water splitting



(Classic) machine learning in action: predicting elastic moduli

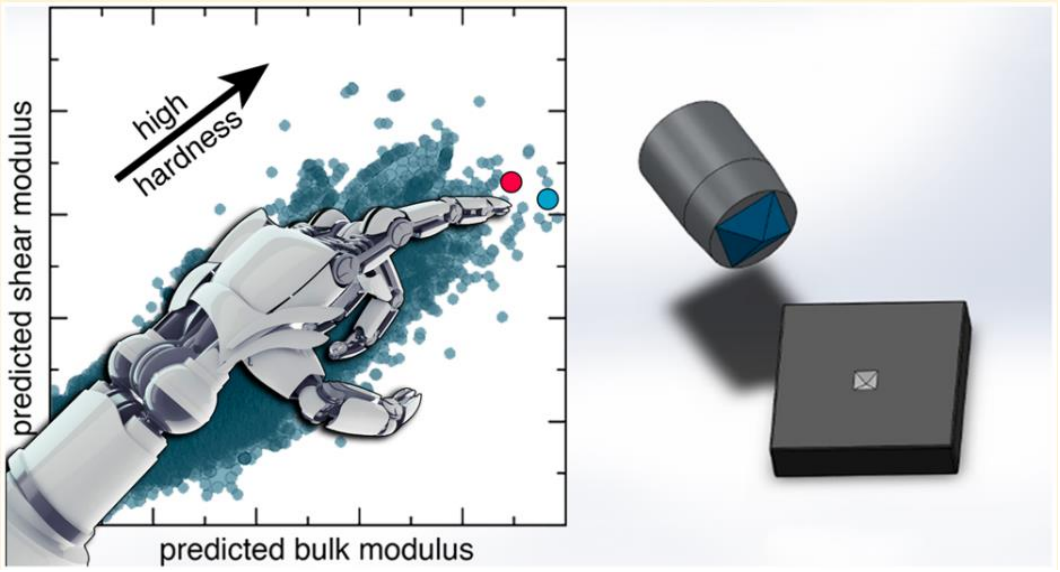
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
Cite This: J. Am. Chem. Soc. 2018, 140, 9844–9853
pubs.acs.org/JACS

Machine Learning Directed Search for Ultraincompressible, Superhard Materials

Aria Mansouri Tehrani,[†,⊥] Anton O. Oliynyk,[†,⊥] Marcus Parry,[‡] Zeshan Rizvi,[†] Samantha Couper,[§] Feng Lin,[§] Lowell Miyagi,[§] Taylor D. Sparks,[‡] and Jakoah Brgoch*[,†]

Database: 3248 Bulk (*B*) and shear modulus (*G*) data obtained from the Materials Project (MP) database

- **Model**: Support vector machine regression using 150 composition and structural descriptors
- **Performance**:
  - r^2 score = 0.94
  - Identified incompressible – high hardness metal $ReWC_{0.8}$ and $Mo_{0.9}W_{1.1}BC$ with *B* = 380 and 370 GPa, respectively
  - Experimentally verified



(Classic) machine learning in action: predicting elastic moduli

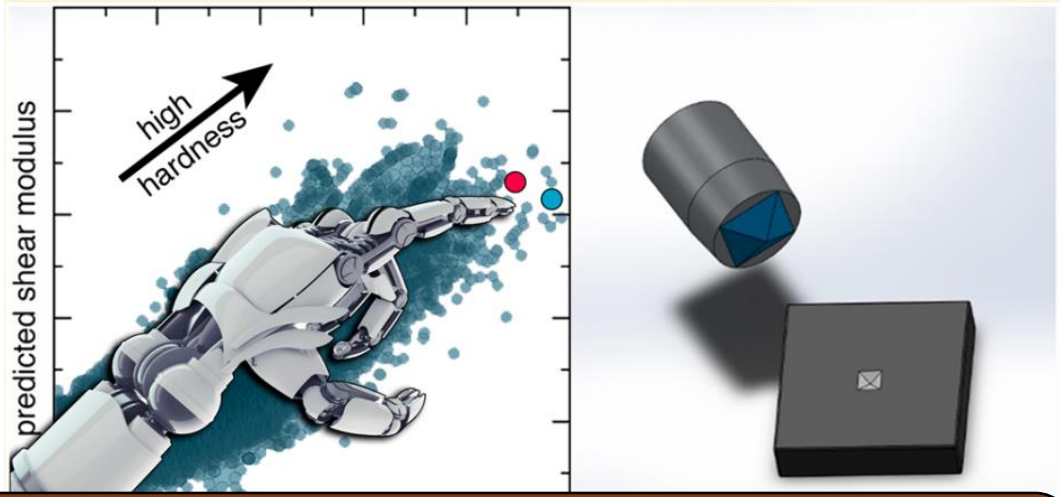
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
Article
Cite This: J. Am. Chem. Soc. 2018, 140, 9844−9853
pubs.acs.org/JACS

Machine Learning Directed Search for Ultraincompressible, Superhard Materials

Aria Mansouri Tehrani,[†,⊥] Anton O. Oliynyk,[†,⊥] Marcus Parry,[‡] Zeshan Rizvi,[†] Samantha Couper,[§] Feng Lin,[§] Lowell Miyagi,[§] Taylor D. Sparks,[‡] and Jakoah Brgoch*[,†]

Database: 3248 Bulk (*B*) and shear modulus (*G*) data obtained from the Materials Project (MP) database

- **Model**: Support vector machine regression using 150 composition and structural descriptors
- **Performance**:
  - r^2 score = 0.94
  - Identified incompressible – high hardness metal $ReWC_{0.8}$ and $Mo_{0.9}W_{1.1}BC$ with *B* = 380 and

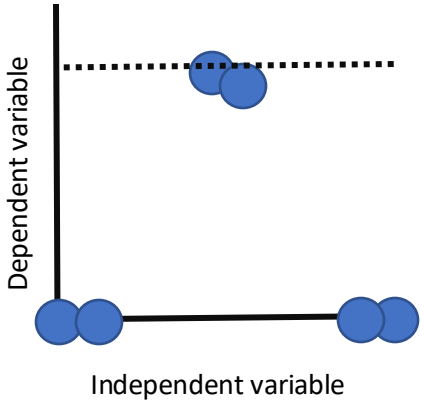


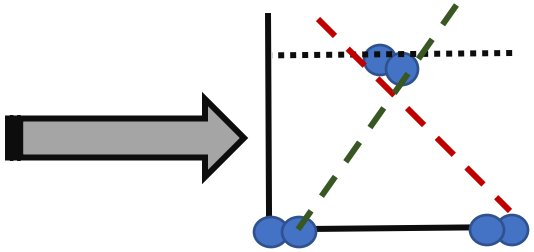
Summary:
- Classical ML models have been used in specific property predictions with varied accuracy

Graph models

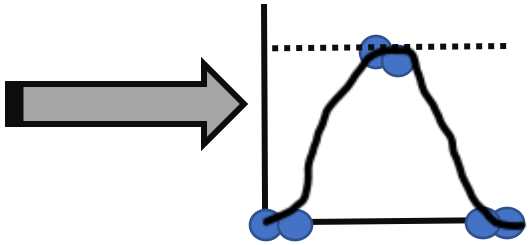
Neural networks

Suppose we want to fit the following data





Fitting a simple regression model

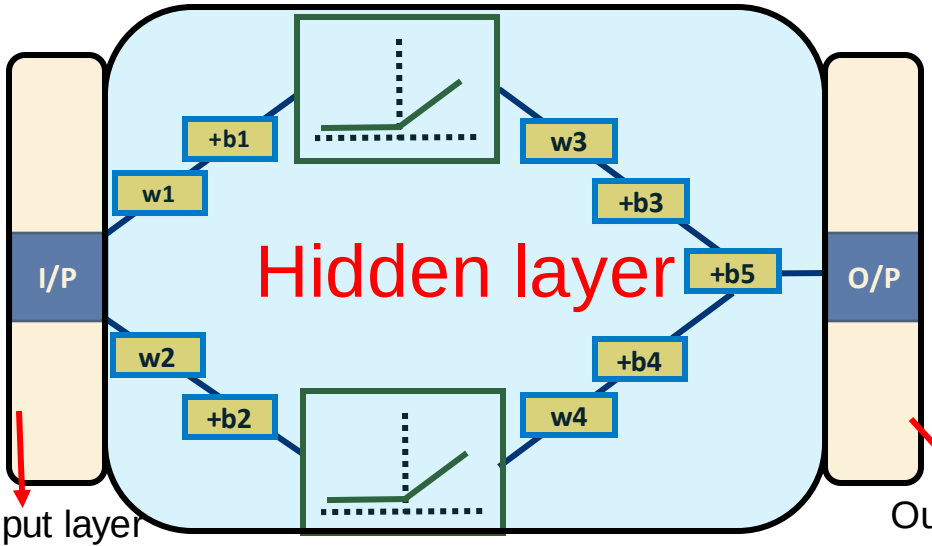


Neural networks can fit a squiggle

Pros
- Robust and accurate
- Parallel processing

Cons
- Minimal interpretability
- Tendency to overfit

Single layer neural network



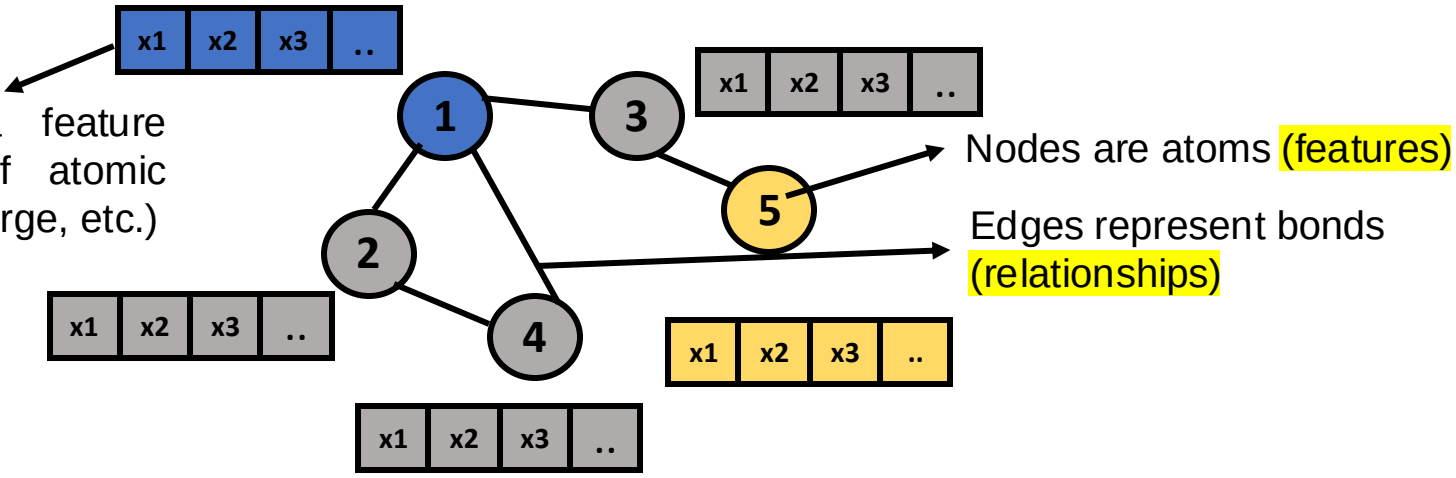
Optimized biases and weights are obtained via back propagation

Weights and biases determine the part of the activation function that will contribute to the squiggle

Several types of NNs exist
Graph NNs particularly relevant for materials

Graphs are an intuitive way to model atoms and bonds

Each node has a feature vector: collection of atomic properties (mass, charge, etc.)



Graphical representation of the data (structure) → Suitable convolutional scheme → Embeddings (pass information across bonds) → Feed forward to a NN (fit weights and biases) → Regression or classification

Graph neural networks can make predictions at three levels

- Graph level (overall structure)
- Edge level (for a given bond)
- Node level (for a given atom)

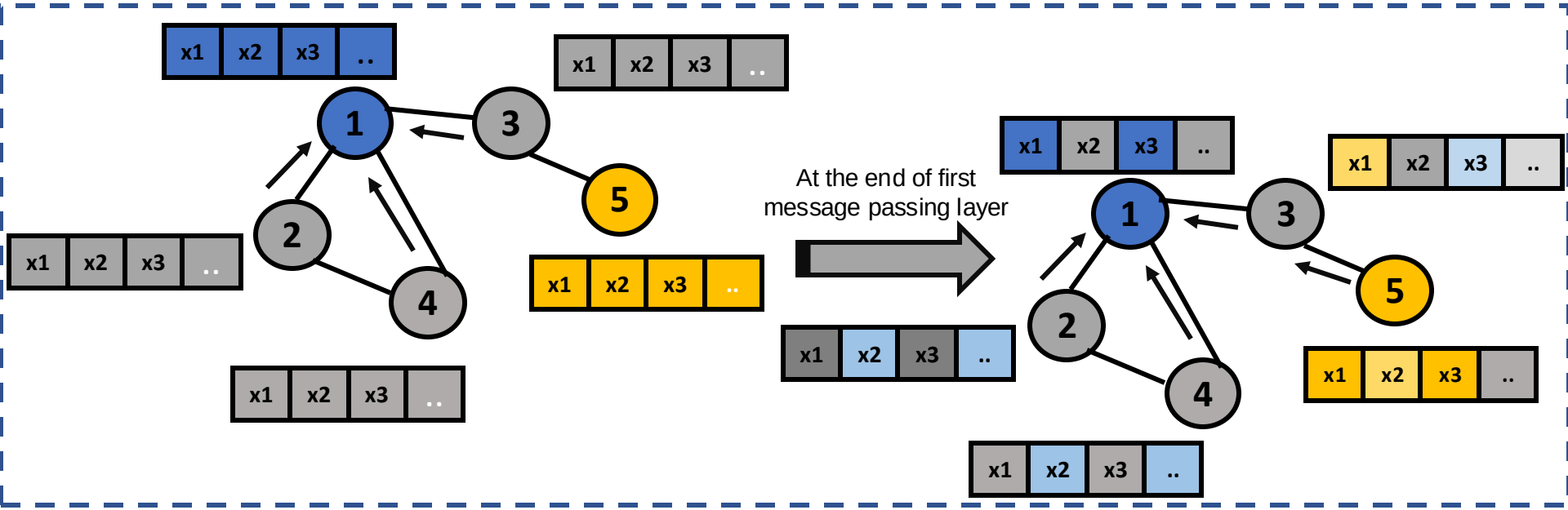
Pros

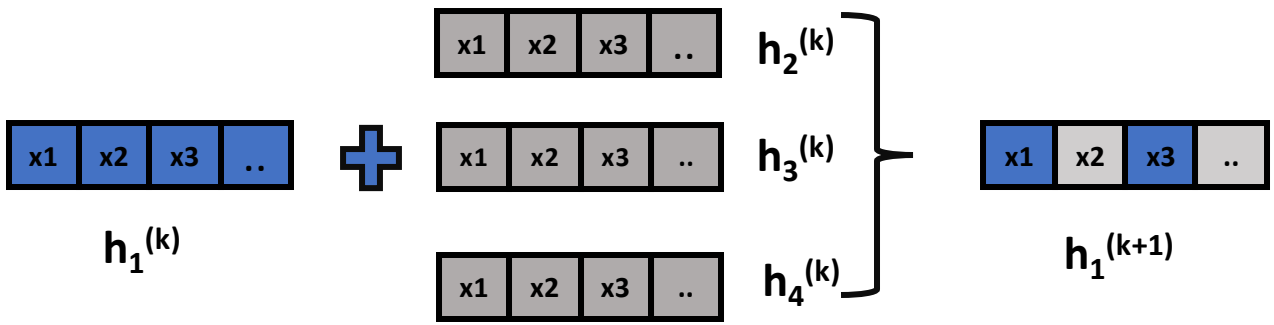
- Highly accurate
- Message passing: use information from neighbors
- Can take into account underlying symmetry

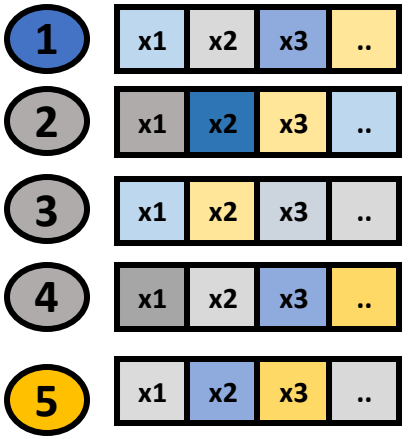
Cons

- Storage/input graph size
- Inability to distinguish multiple types of bonds
- Need to ensure permutational invariance and equivariance

Message passing: learn from neighbors





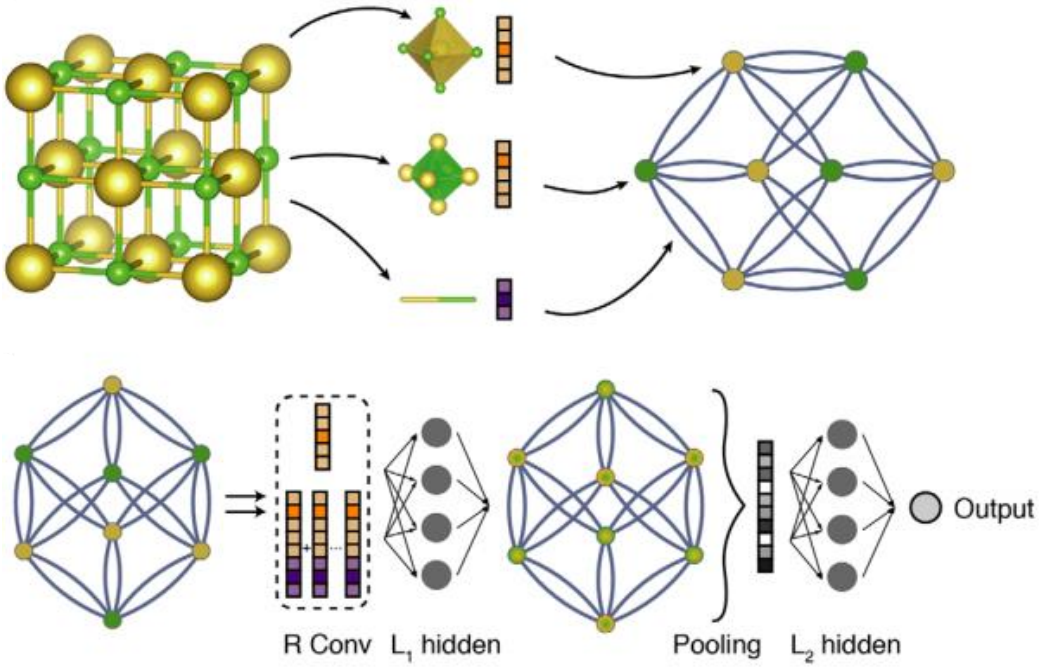


Predicting material properties:
Diverse material properties with graph neural network

PHYSICAL REVIEW LETTERS **120,** 145301 (2018)

Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties

Tian Xie and Jeffrey C. Grossman
Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA



Properties : Formation energy, band gap, Fermi energy, bulk and shear moduli, and Poisson's ratio

Database: 10^4 DFT-calculated datapoints from MP

Model: Crystal Graph convolutional neural network (CGCNN)

Performance:

- Formation energy: 0.039 eV/atom
- Band gap: 0.388 eV
- Fermi energy: 0.363 eV
- Elastic moduli: ~1-2 GPa
- Poisson's ratio: 0.03
- Identified 228 'synthesizable' perovskites out of 18928 in the training database

Predicting material properties:
Mechanical properties for energy storage

ACS central science

ACS AuthorChoice

Research Article

Cite This: *ACS Cent. Sci.* 2018, 4, 996−1006

http://pubs.acs.org/journal/acscii

Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes

Zeeshan Ahmad,[†] Tian Xie,[‡] Chinmay Maheshwari,[†] Jeffrey C. Grossman,[‡] and Venkatasubramanian Viswanathan*[,†,§]

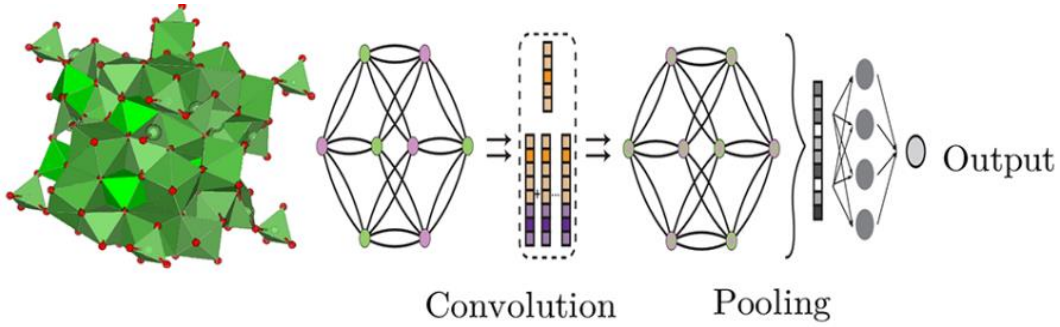
Mechanically anisotropic interfaces suppress dendrite growth

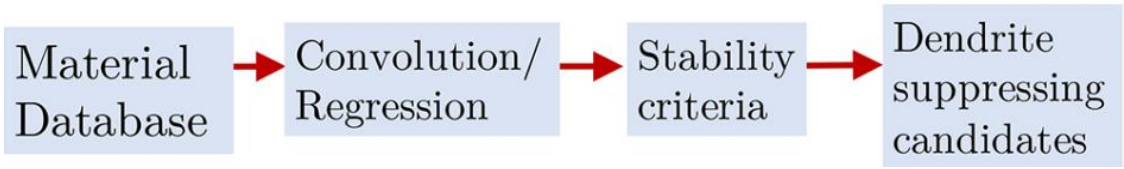
- Dependent on *G*, *B*, and elastic constants.

Database: Subset of MP containing 12,000 compounds with Li

Model:

- Graph neural network for *G* and *B* prediction
- Gradient boost and Kernel-ridge regression for elastic constant predictions





Performance:

- RMSE in log(GPa): 0.1268 (*G*) and 0.1013 (*B*)
- 20 interfaces with six solid electrolytes predicted to be stable against dendrite initiation

Predicting material properties:
Mechanical properties for energy storage

ACS central science

ACS AuthorChoice

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 996−1006

http://pubs.acs.org/journal/acscii

Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes

Zeeshan Ahmad,[†] Tian Xie,[‡] Chinmay Maheshwari,[†] Jeffrey C. Grossman,[‡] and Venkatasubramanian Viswanathan*[,†,§]

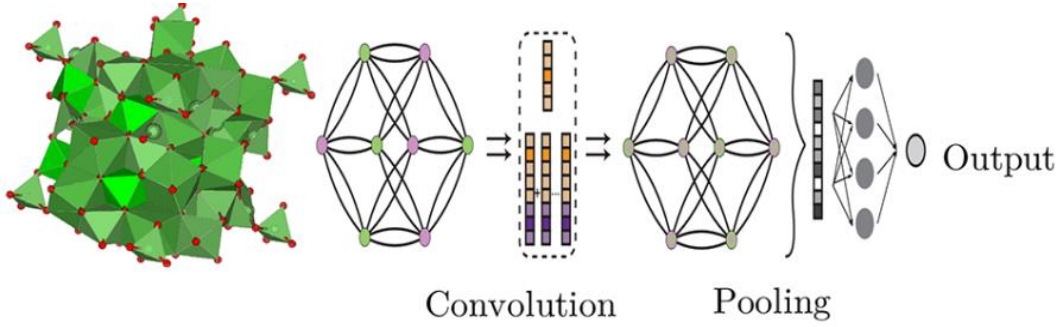
Mechanically anisotropic interfaces suppress dendrite growth

- Dependent on *G*, *B*, and elastic constants.

Database: Subset of MP containing 12,000 compounds with Li

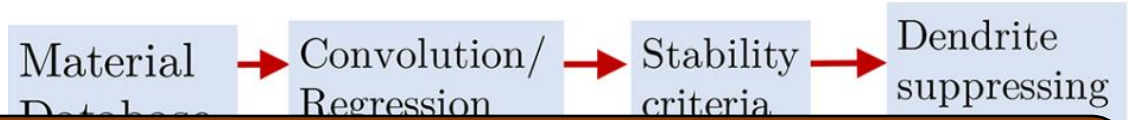
Model:

- Graph neural network for *G* and *B* prediction
- Gradient boost and Kernel-ridge regression for elastic constant predictions



Performance:

- RMSE in log(GPa): 0.1268 (*G*) and 0.1013 (*B*)
- 20 interfaces with six solid



Summary:

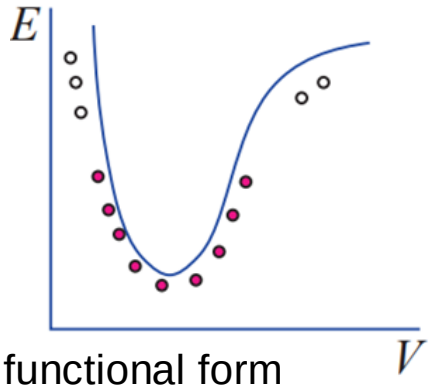
- Graph networks are an intuitive way to represent materials
- Have advanced accuracy of models and enabled predicting multiple properties with similar architecture

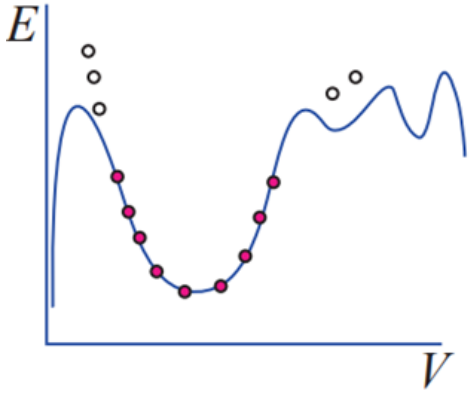
Graph models and interatomic potentials

Why machine learned interatomic potentials (MLIPs)?

Classical force-fields have difficulties in modelling ‘complex’ potential energy surfaces

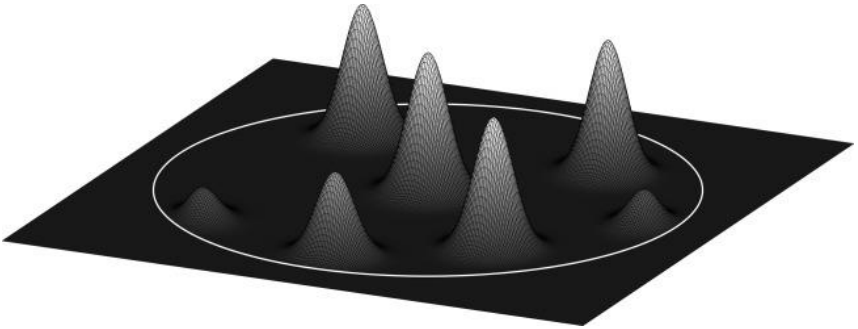
- Diversity of species and bonding environments
- Limited accuracy vs. DFT





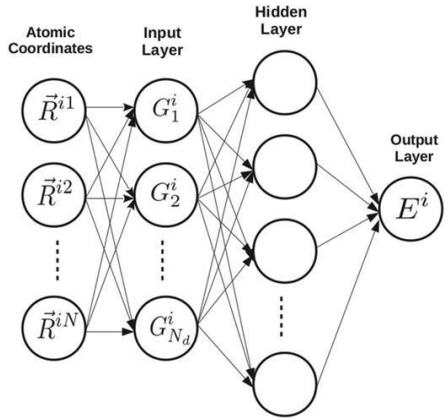
MLIPs: Flexible functional form

- Can handle diversity of species and bonding environments
- Introduce permutation, rotation invariance
- Improved accuracy vs. DFT compared to classical force-fields



Bartók and Csányi, Int. J. Quantum Chem. 116, 1049 (2016)

Mishin, Acta Mater. 214, 116980 (2014)

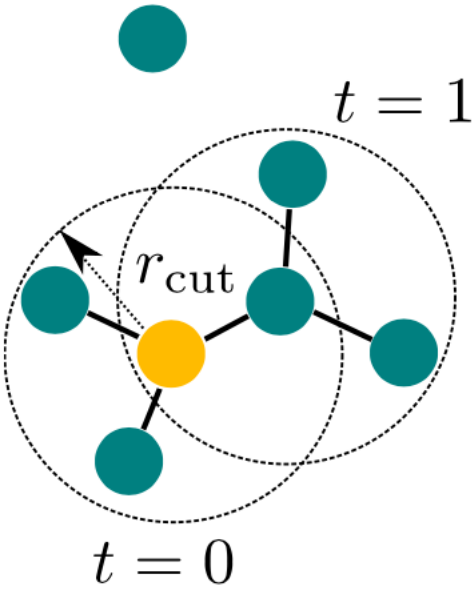


Kocer et al., J. Chem. Phys. 150, 154102 (2019)

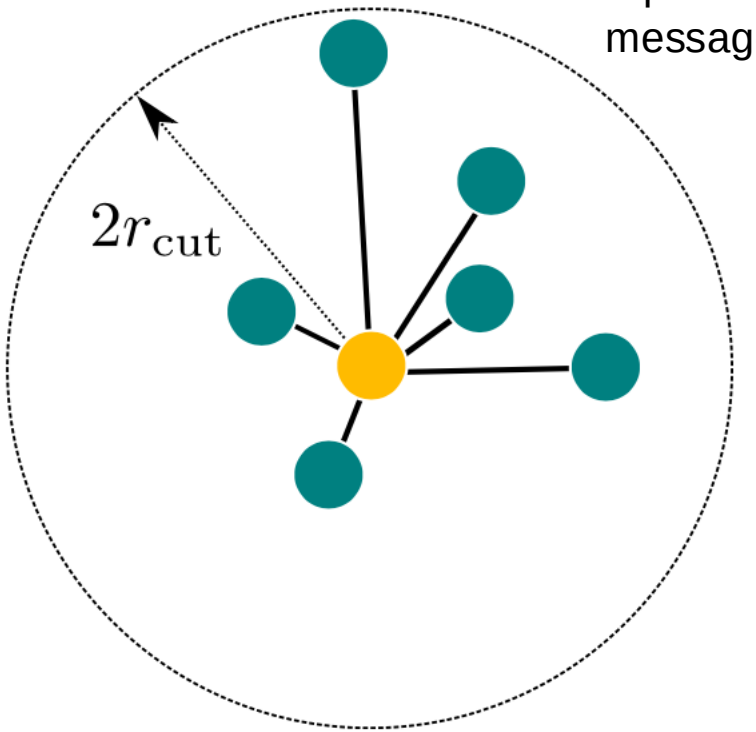
Fingerprint a local environment around a reference atom
+ machine-learning model
= (classic) MLIP

Message passing is quite useful

With message passing
(t: iteration)



Atom-centered representation (without message passing)



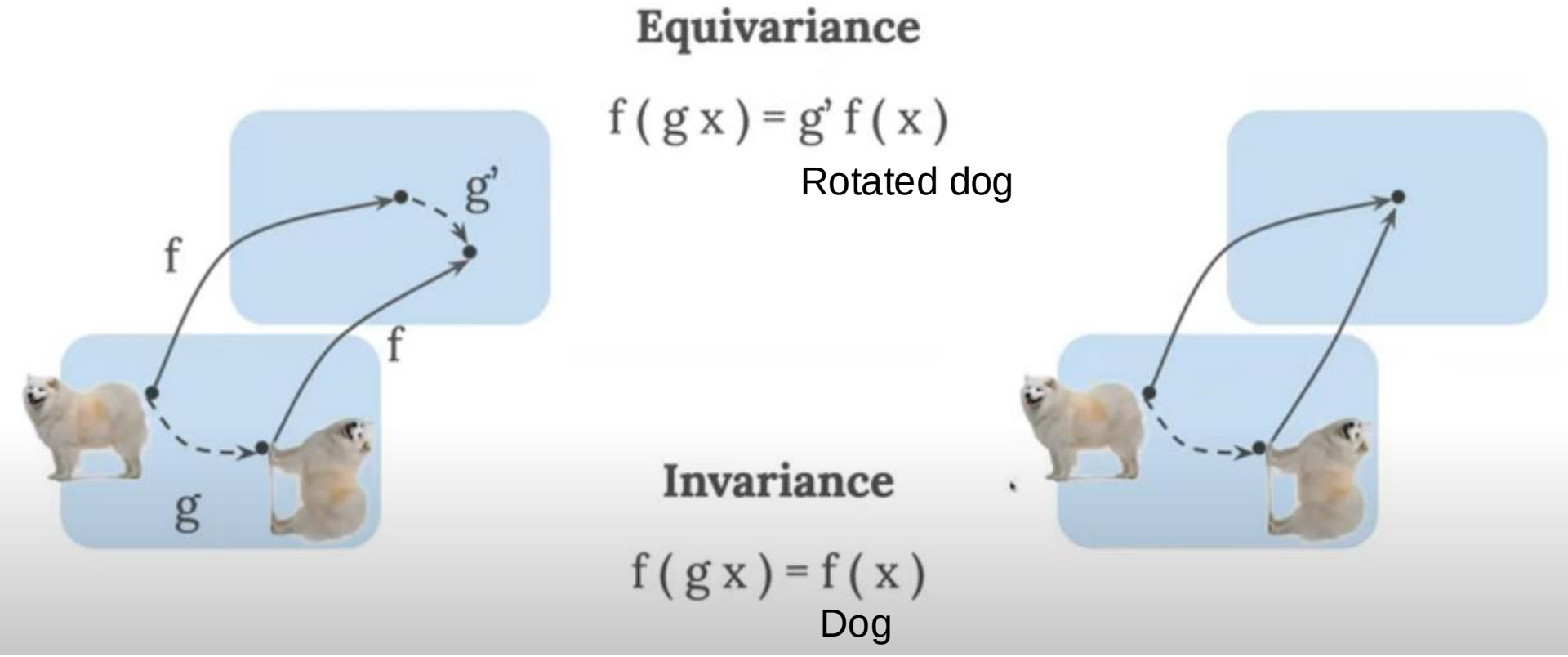
Message passing helps learn long-range interactions

- Effective interaction from $t \times r_{cut}$
- Computationally efficient
- Eliminates unnecessary neighbors

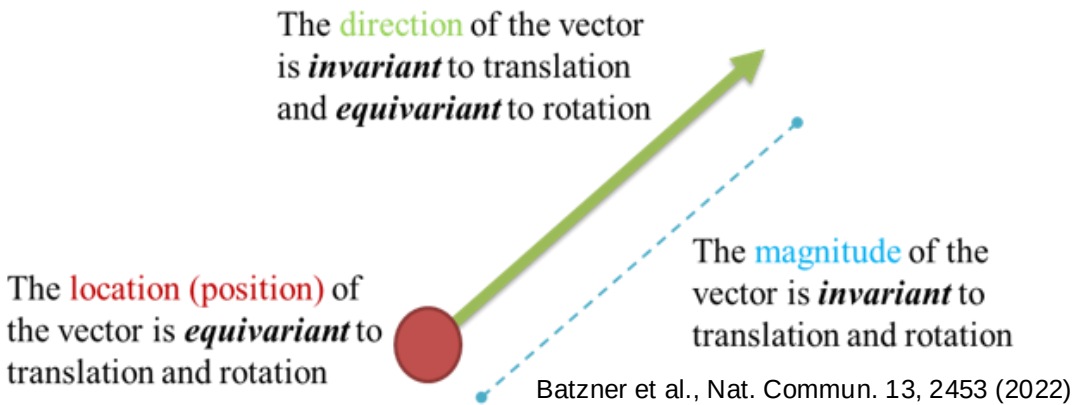
MLIPs incorporating message passing should have higher learning rates and describe longer range interactions better

Figure: Batatia et al.,, arXiv, 2205.06643v2 (2022)

Invariance vs. equivariance



https://datascience.stackexchange.com/questions/16060/what-is-the-difference-between-equivariant-to-translation-and-invariant-to-tr



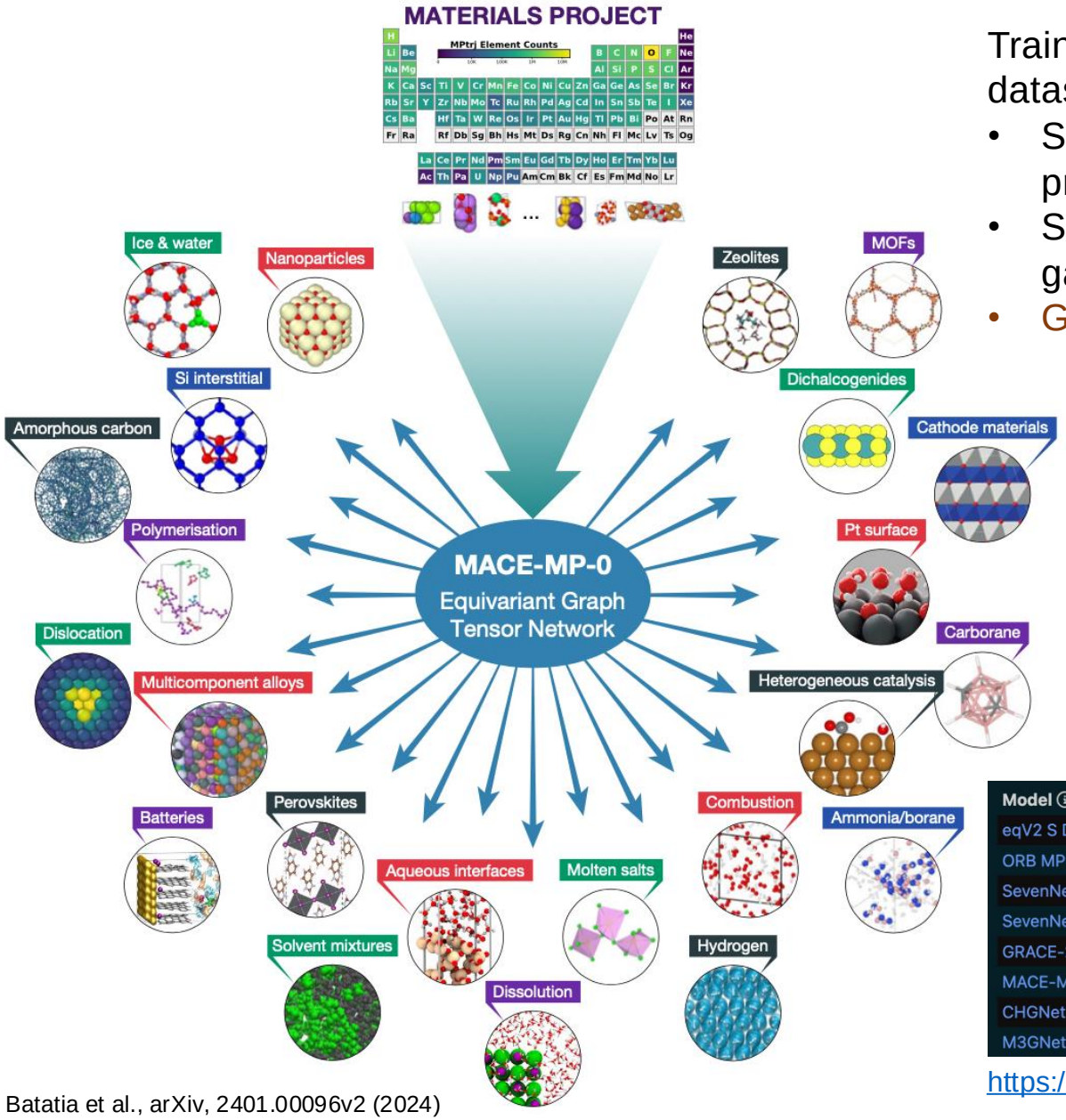
Batzner et al., Nat. Commun. 13, 2453 (2022)

In materials parlance:

- Scalars (energies) are invariant
- Vectors (forces) and tensors (stresses) are equivariant
- Several useful material properties are equivariant

E(3): Translation, Rotation, Reflection

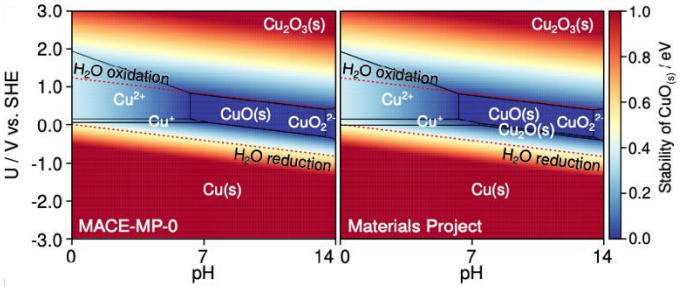
Foundational models: MACE-MP-0



Batatia et al., arXiv, 2401.00096v2 (2024)

Trained on Materials Project trajectory dataset (~1.5M structures)

- Stable performance on 30 different property predictions/application areas
- Stable dynamics in solids, liquids, and gases
- GPU; limited system size



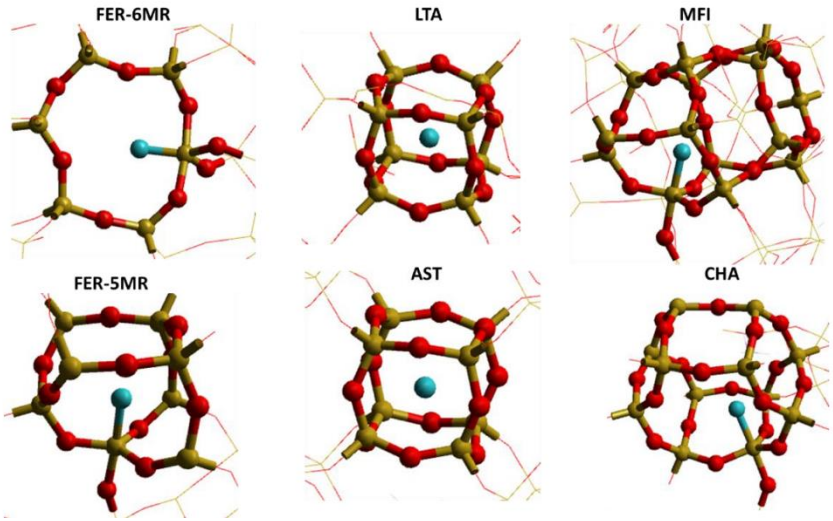
Many more to come…

Model	F1 ↑	DAF ↑	Prec ↑	Acc ↑	MAE ↓	R^2 ↑	κ_{SRME} ↓	Training Set
eqV2 S DeNS	0.815	5.042	0.771	0.941	0.036	0.788	1.665	146k (1.58M) (MPtrj)
ORB MPtrj	0.765	4.702	0.719	0.922	0.045	0.756	1.725	146k (1.58M) (MPtrj)
SevenNet-l3i5	0.76	4.629	0.708	0.92	0.044	0.776	0.55	146k (1.58M) (MPtrj)
SevenNet-0	0.724	4.252	0.65	0.904	0.048	0.75	0.767	146k (1.58M) (MPtrj)
GRACE-2L (r6)	0.691	4.163	0.636	0.896	0.052	0.741	0.525	146k (1.58M) (MPtrj)
MACE-MP-0	0.669	3.777	0.577	0.878	0.057	0.697	0.647	146k (1.58M) (MPtrj)
CHGNet	0.613	3.361	0.514	0.851	0.063	0.689	1.717	146k (1.58M) (MPtrj)
M3GNet	0.569	2.882	0.441	0.813	0.075	0.585	1.412	62.8k (188k) (MPF)

https://matbench-discovery.materialsproject.org/

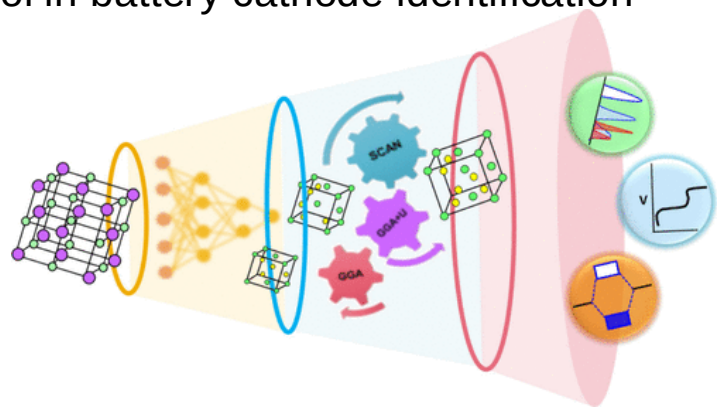
MACE in action

Modelling zeolites (using MACE-ML-IP model)



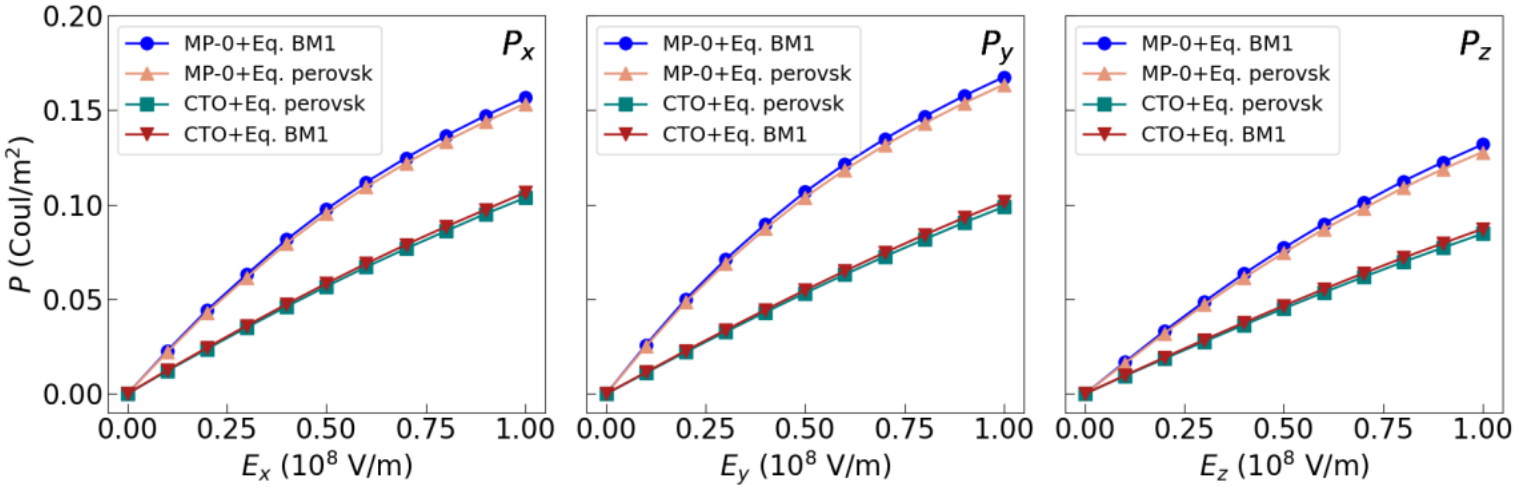
Nasir et al., arXiv, 2411.00436 (2024)

Using MACE-MP-0 as a pre-screening tool in battery cathode identification



Singh et al., ACS Appl. Electron. Mater. 6, 7065-7074 (2024)

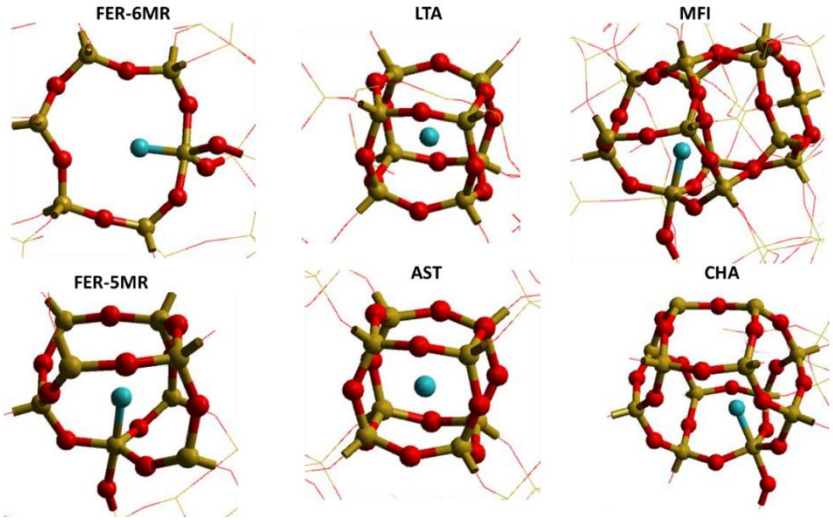
Polarization of $CaTiO_3$ with applied electric field (MACE-MP-0 and custom models)



Kutana et al., arXiv, 2412.03541 (2024)

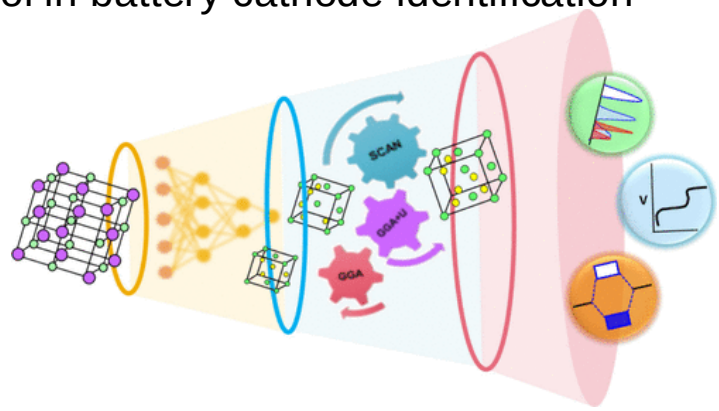
MACE in action

Modelling zeolites (using MACE-ML-IP model)

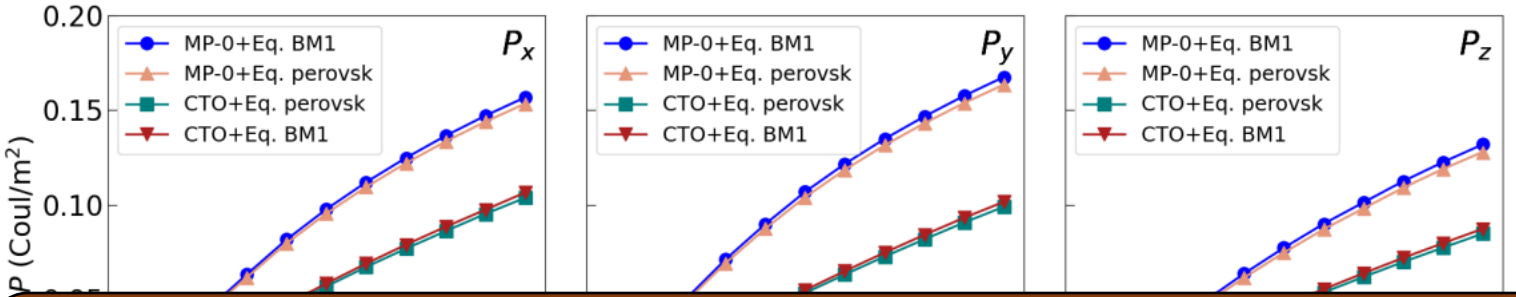


Nasir et al., arXiv, 2411.00436 (2024)

Using MACE-MP-0 as a pre-screening tool in battery cathode identification



Singh et al., ACS Appl. Electron. Mater. 6, 7065-7074 (2024)



Polarization of $CaTiO_3$ with applied electric field (MACE-MP-0 and custom models)

Summary:

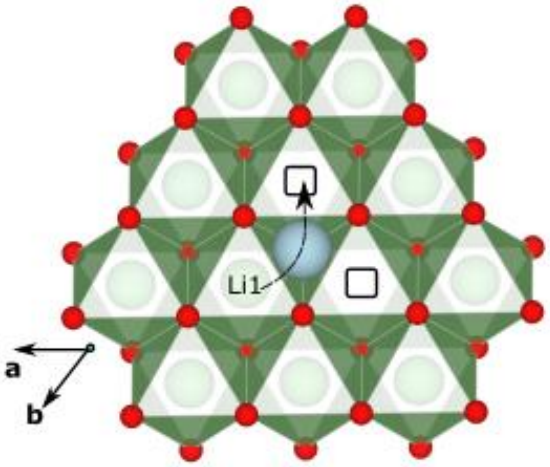
- Graph networks enabling message passing, equivariance, and high body-order interactions have advanced the accuracy of MLIPs and aided in creation of foundational models

Graph models and transfer learning

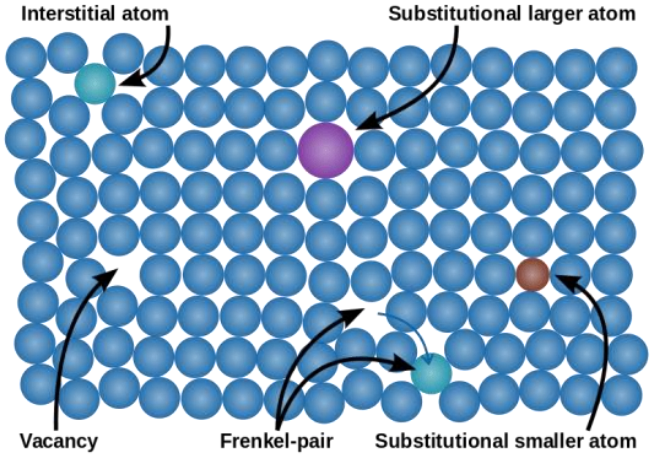
Materials science is data limited

Several key material properties that govern performance in applications have limited data

- 'Small' datasets (< 10^4 datapoints)
  - Ionic mobilities, defect formation energies, adsorption energies,…
- Limits application of deep learning (DL) frameworks

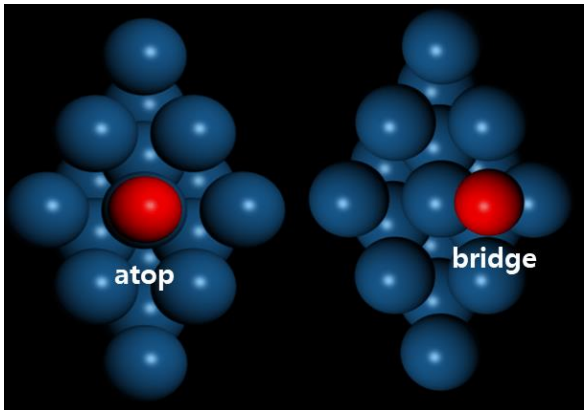


Devi et al., **npj Comput. Mater.** 2022



https://www.differencebetween.com/difference-between-point-defect-and-line-defect/

https://sites.psu.edu/



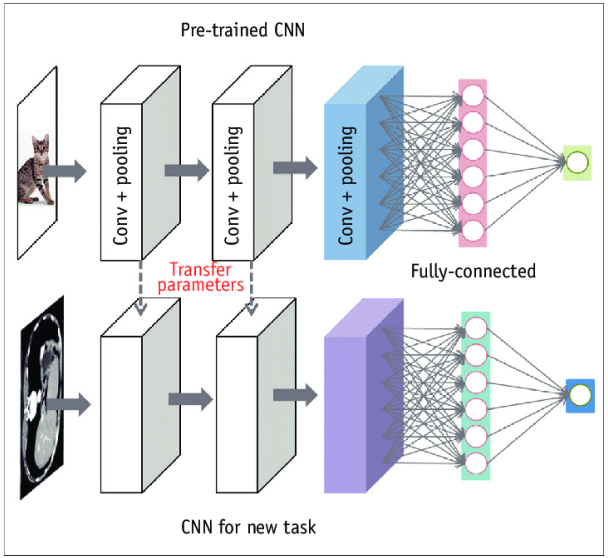
Materials science is data limited

Several key material properties that govern performance in applications have limited data

- ‘Small’ datasets (< 10^4 datapoints)
  - Ionic mobilities, defect formation energies, adsorption energies,…
- Limits application of deep learning (DL) frameworks

Transfer learning: efficiently use DL frameworks on small datasets

- Pre-train (**PT**) on ’large’ dataset, fine-tune (**FT**) on ‘small’ dataset



Do et al., **Korean J. Radiol.** 2020

Materials science is data limited

Several key material properties that govern performance in applications have limited data

- 'Small' datasets (< 10^4 datapoints)
  - Ionic mobilities, defect formation energies, adsorption energies,…
- Limits application of deep learning (DL) frameworks

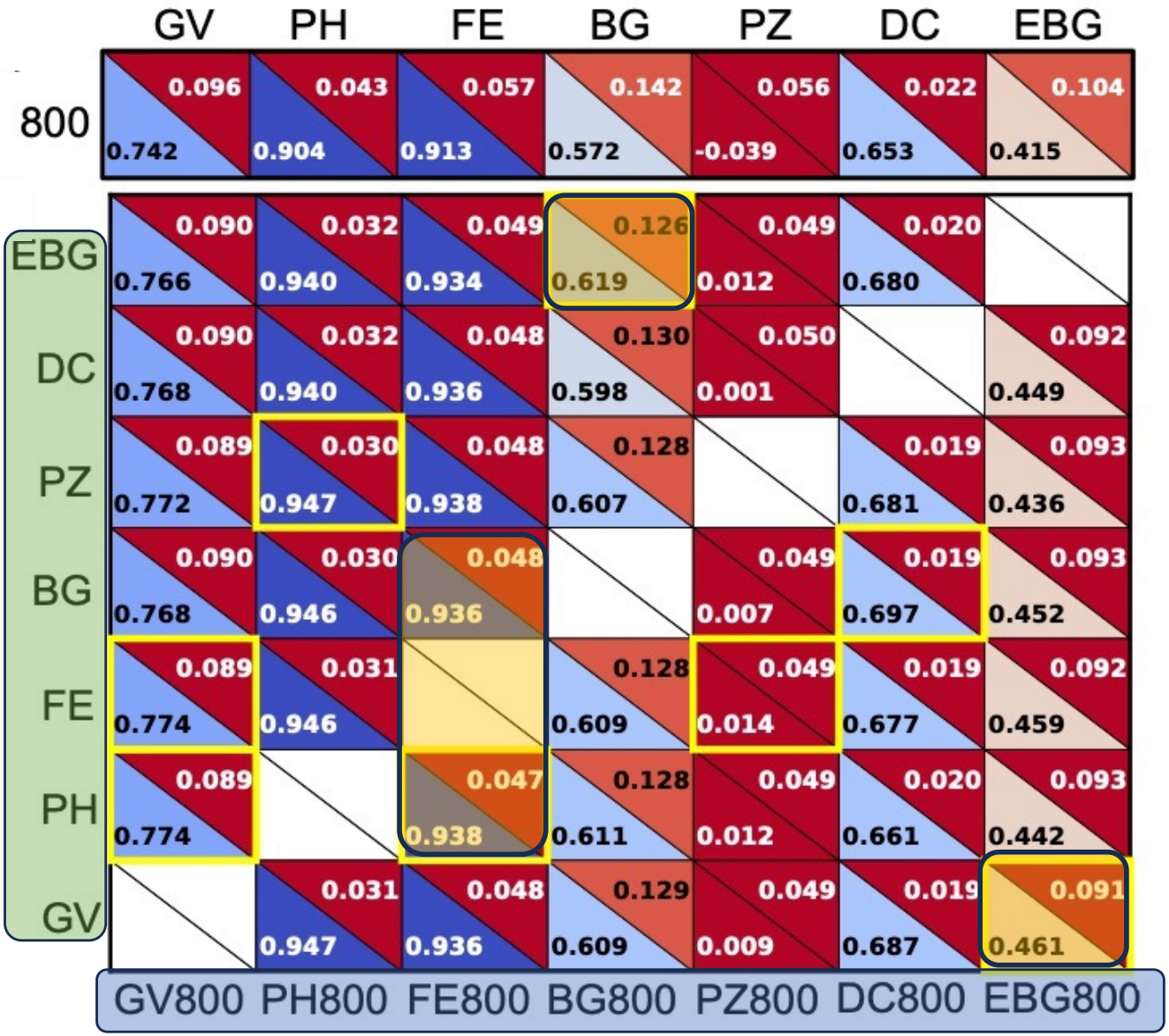
Transfer learning: efficiently use DL frameworks on small datasets

- Pre-train (**PT**) on 'large' dataset, fine-tune (**FT**) on 'small' dataset

How useful is transfer learning in materials science?

- Optimal ways to use?
- Ways to generate 'generalized' models?

7×6 combinations of pair-wise models



Pair-wise models: better than scratch

- Average increase in R^2: 25%
- Average decrease in MAE: 16%

Best models: GV, PH, FE ($R^2 > 0.75$)

Average models: BG, DC, EBG

Specific PT property: little influence on FT

No symmetry

FT dataset+size

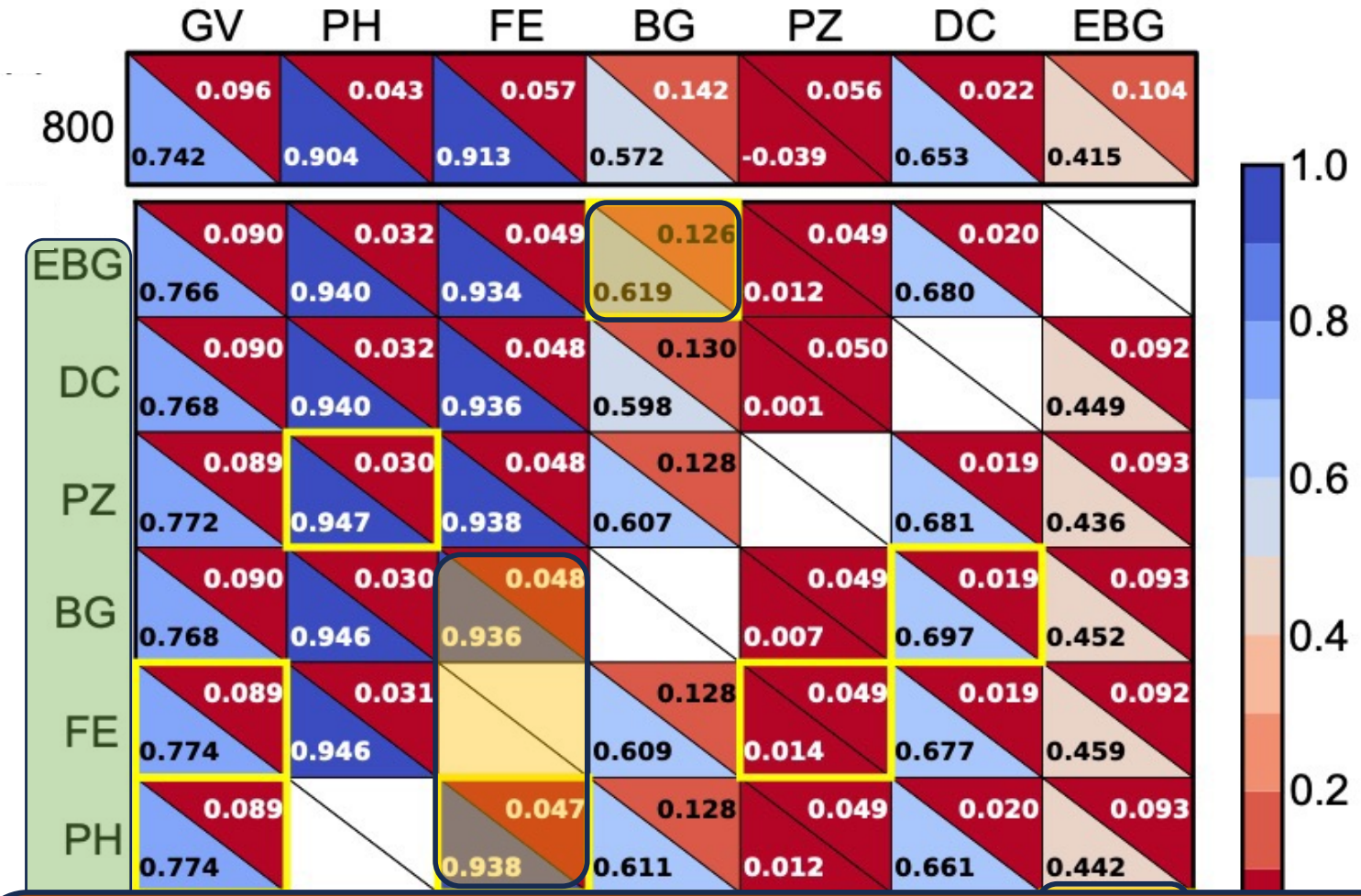
PT dataset (941)

Best model

Test scores

GV: Shear modulus; PH: Phonons: FE: Formation energy; BG: Band gap
PZ: Piezoelectric modulus; DC: Dielectric constant; EBG: Experimental band gap

7×6 combinations of pair-wise models



Pair-wise models: better than scratch

- Average increase in R^2: 25%
- Average decrease in MAE: 16%

Best models: GV, PH, FE ($R^2 > 0.75$)

Average models: BG, DC, EBG

Specific PT property: little influence on FT

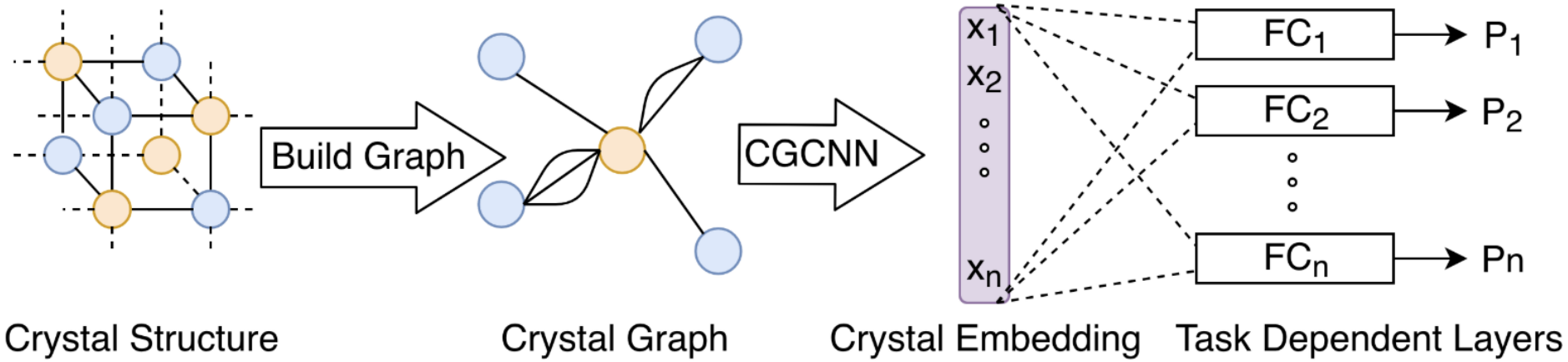
No symmetry

At capped dataset size, specific PT property is a weak handle; Normal distribution is better

Pair-wise transfer learning has significant utility

MPT: (Beta) Generalized models

Inspiration from literature: multi-task crystal graph convolutional neural network[1]



MPT models can generalize dependence of several properties on the structure

- Build cumulative dataset: 132,270 points
  - Remove overlaps
- Add task-dependent prediction heads with a one-hot encoded vector
  - Presence/absence of property
- Modify loss function

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left|y_p^i - y_t^i\right|\delta^i$$

- PT on all (but one) property, FT on one property

1. Sanyal et al., **arXiv** 1811.05660 (2018)

MPT: better on out-of-domain than PT-FT

Band gap of 2D materials (1,103 datapoints) from JARVIS-DFT[1]

Model	Test R^2	Test MAE
Scratch	0.635	0.148
MPT (all seven datasets)	**0.671**	**0.125**
FE(100K)	0.670	0.127
BG(50K)	0.617	0.138
PH(1256)	0.628	0.145
GV(10,987)	0.626	0.143
EBG(2,481)	0.619	0.143

On average, MPT is 6% and 10% better on R^2 and MAE than PT-FT
Closest performer to MPT is FE: largest dataset within MPT
MPT models: may generalize quite well with more properties

FE: Formation energy; BG: Band gap; GV: Shear modulus;
EBG: Experimental band gap

1. Choudhary et al. **npj Comput. Mater.** *6*, 173 (2020).

MPT: better on out-of-domain than PT-FT

Band gap of 2D materials (1,103 datapoints) from JARVIS-DFT[1]

Model	Test R^2	Test MAE
Scratch	0.635	0.148
MPT (all seven datasets)	**0.671**	**0.125**
FE(100K)	0.670	0.127
BG(50K)	0.617	0.138
PH(1256)	0.628	0.145
GV(10,987)	0.626	0.143
EBG(2,481)	0.619	0.143

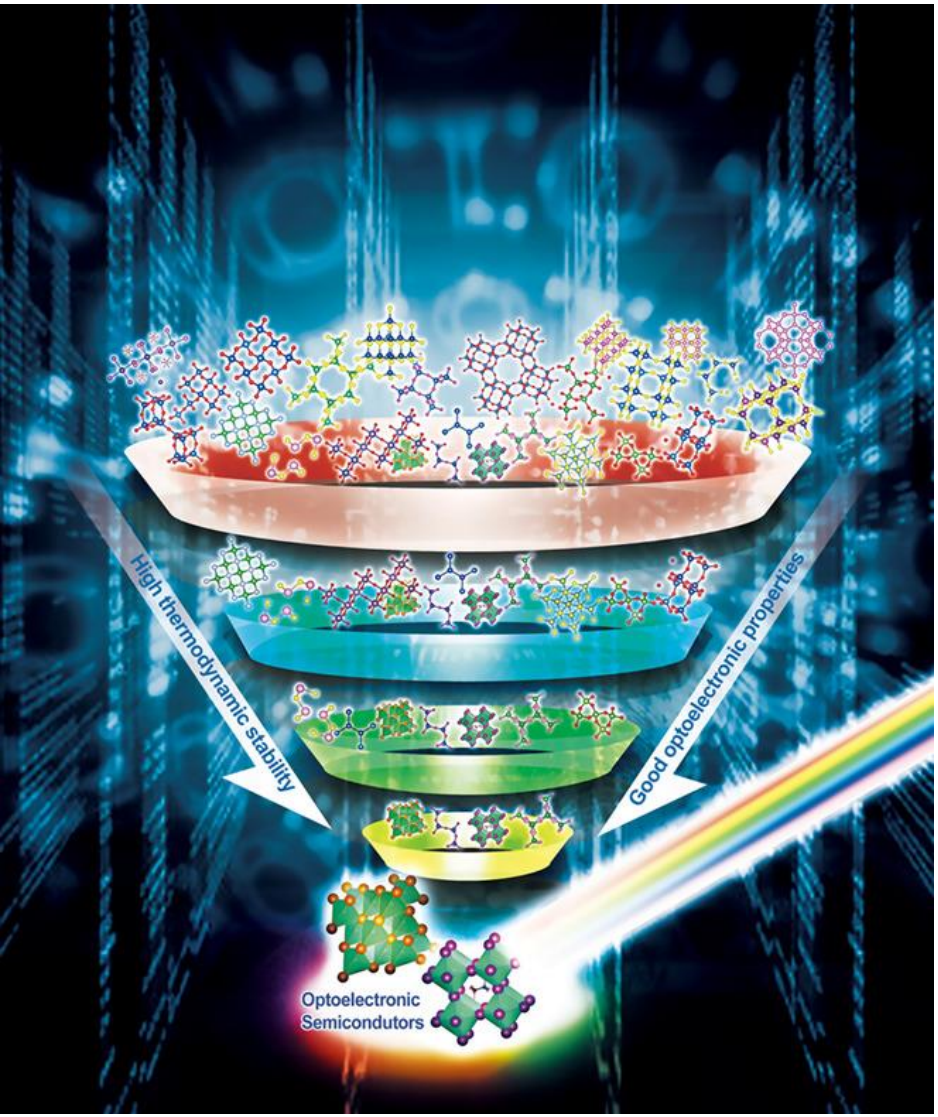
On average, MPT is 6% and 10% better on R^2 and MAE than PT-FT

Summary:

- Transfer learning is useful in mitigating data-availability constraint in materials
- MPT provides a systematic way to create generalizable models

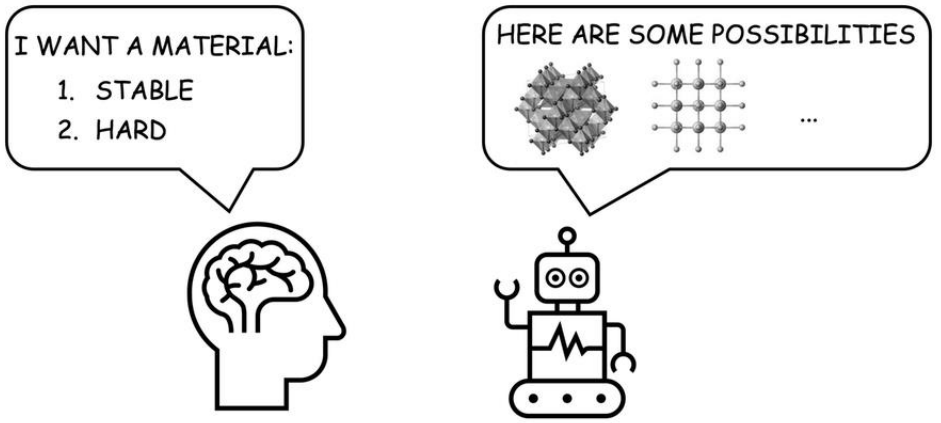
Generative models

Inverse materials design



https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1489

Property → Structure



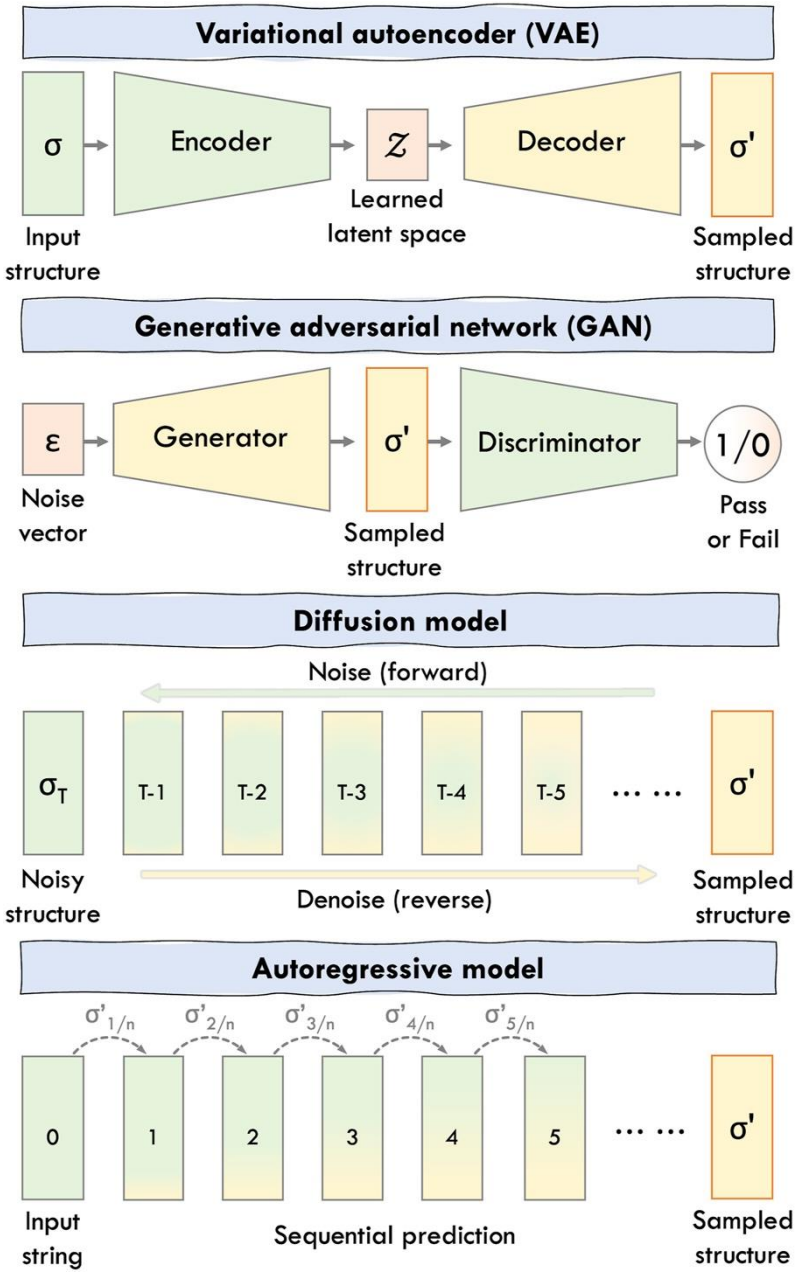
https://news.mit.edu/2022/new-way-perform-general-inverse-design-high-accuracy-0118

Generative models (classic)

- Step 1: encode a configuration (σ) into a latent/feature space (Z)
  - $Z = f(\sigma)$
- Critical info of any structure
  - Composition
  - Lattice parameters
  - Atomic positions
  - Use graph neural networks to obtain Z
  - Z can be mapped to labelled properties
- Step 2: decode configuration from latent space using a learnable function
  - $\sigma' = f'(Z)$
  - Introduces noise
  - Provides a probability distribution (compositions, lattice parameters, and positions)
- Step 3: generate configuration by sampling probabilities
  - $\sigma_{sampled} = p(Z)$
  - Given constraints on target properties, composition, and/or lattice geometry

Park et al., Matter 7, 2355-2367 (2024)

Advancements in generative models



Classic

Generator confuses discriminator with synthetic data

- Beaten by diffusion models ☹

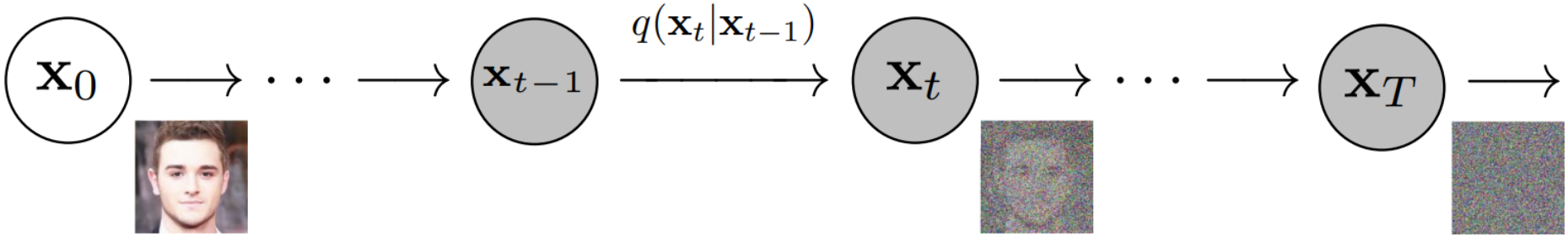
Progressive noise addition/removal

Sequential probability (language models)

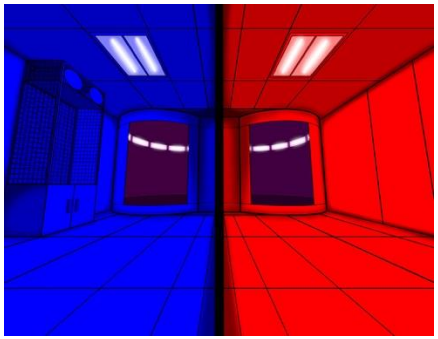
Park et al., Matter 7, 2355-2367 (2024)

What is diffusion?

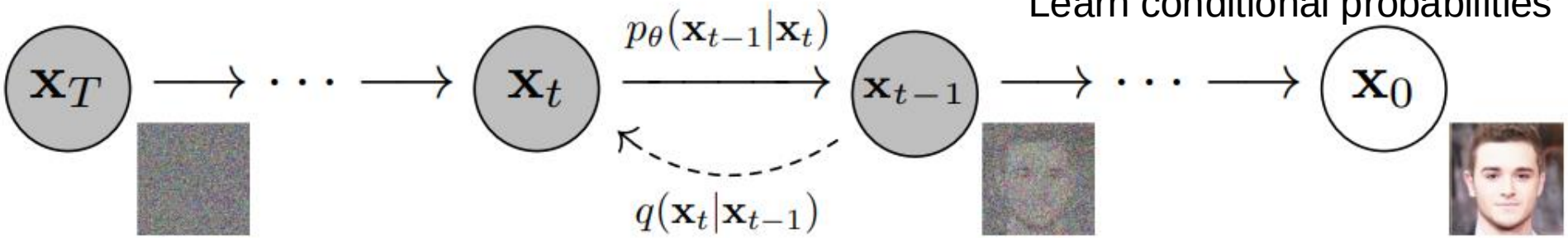
Forward diffusion



Gaussian noise added in a Markov chain



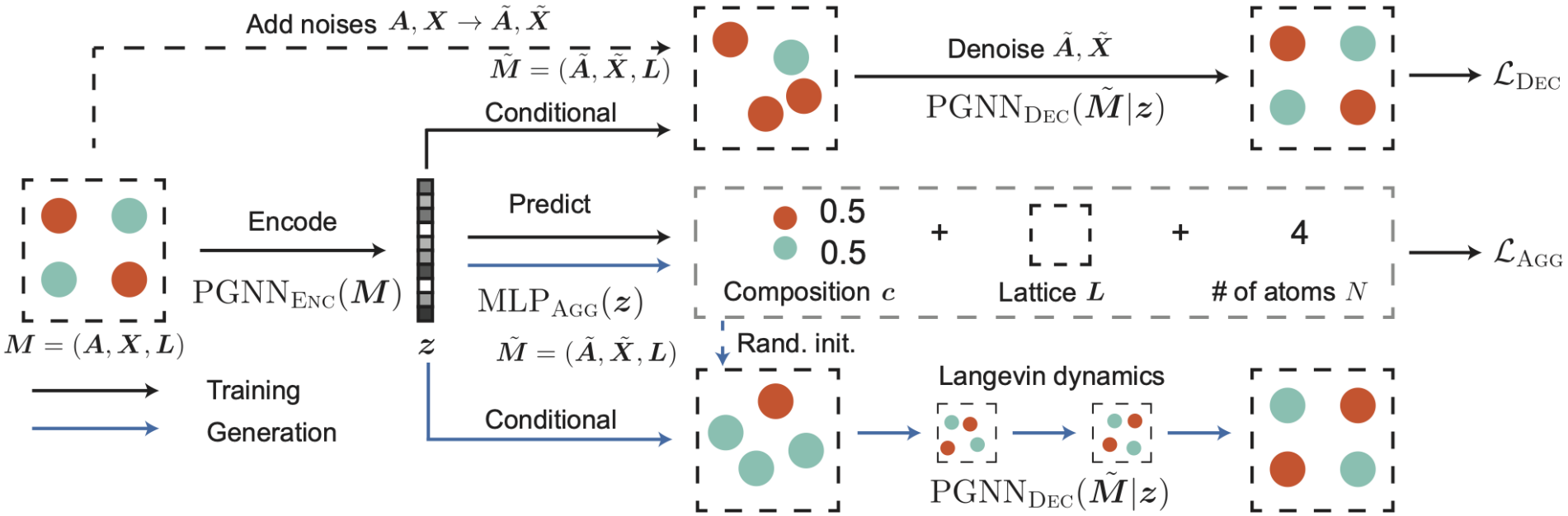
Learn conditional probabilities



Reverse diffusion

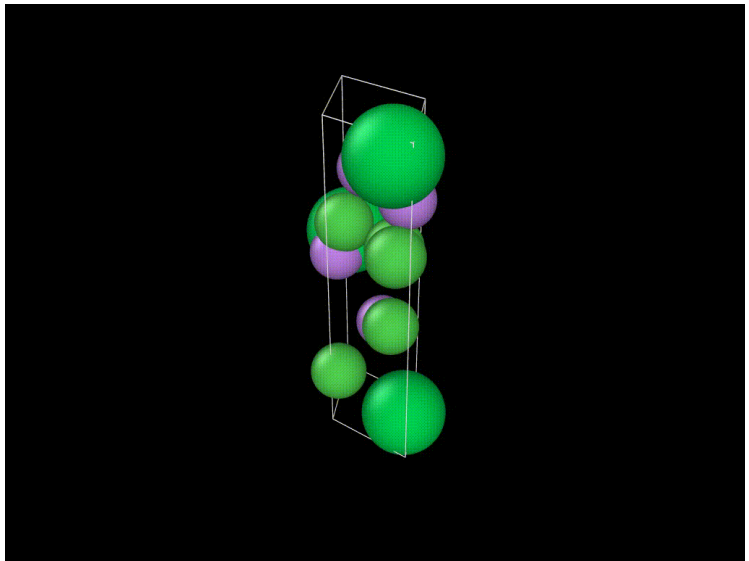
https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction https://medium.com/@luisfelipechary/my-experience-with-diffusion-super-resolution-3386b6574696

In materials, diffusion models can be used for structure generation



Crystal diffusion variational autoencoder (CDVAE)

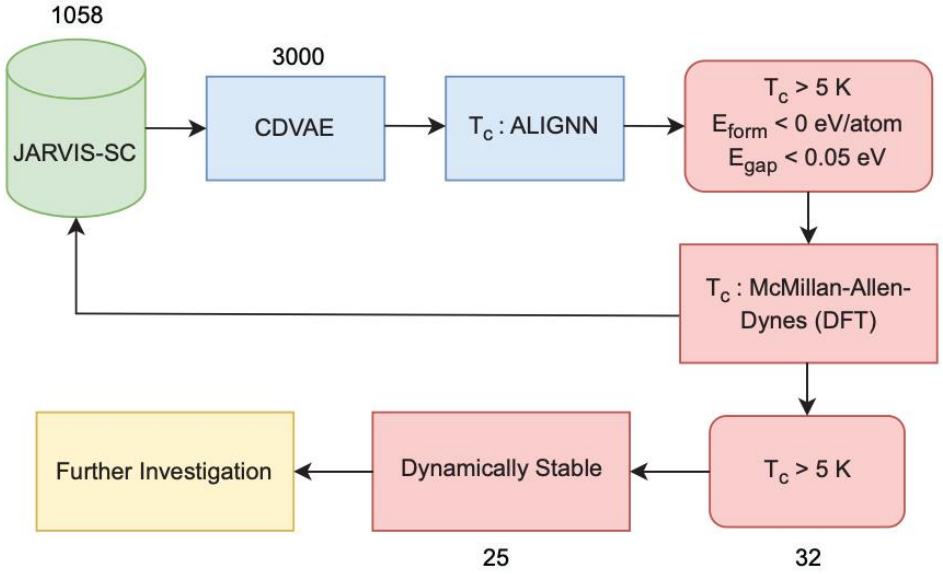
- One of the first diffusion models to be developed for structure prediction
- Periodic graph networks for encoding a latent space and denoising
- Property predictor: for composition, lattice, and number of atoms from latent space
- Langevin dynamics: final structure

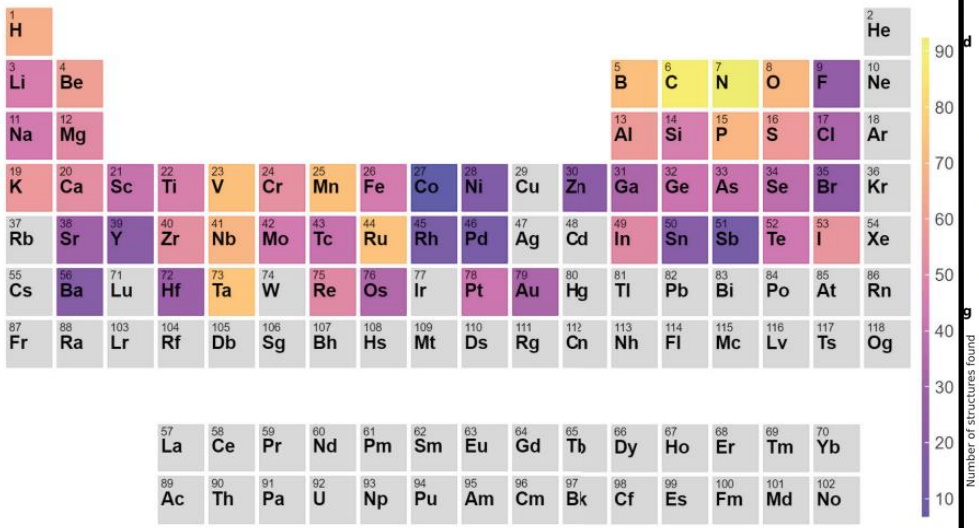


Xie et al., ICLR 2022; https://github.com/txie-93/cdvae

Diffusion models in action

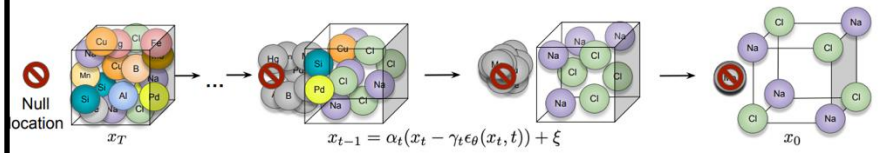
Inverse design of new superconductors





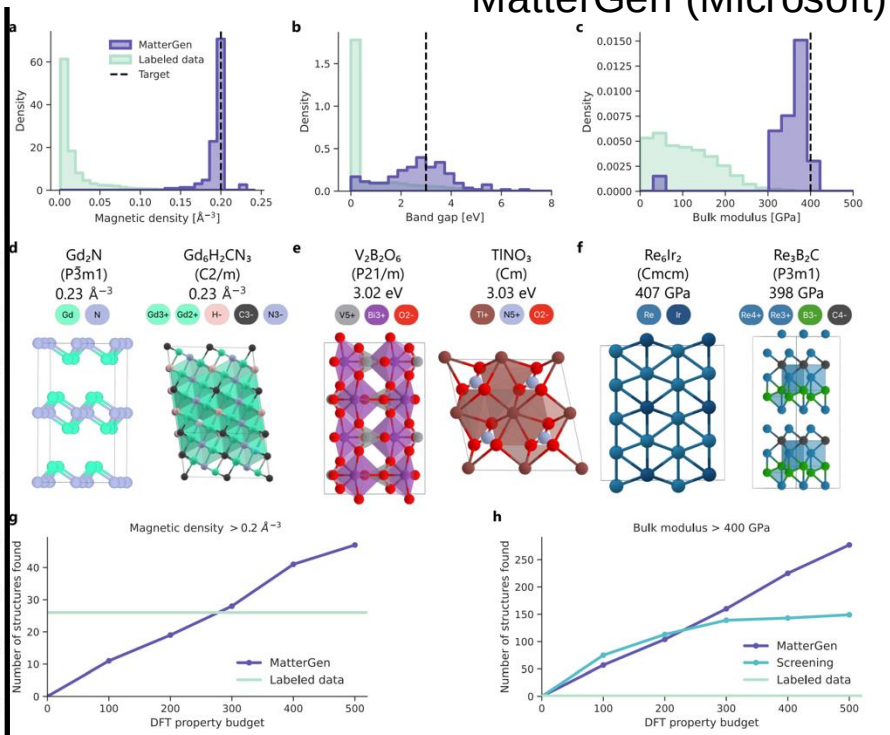
Wines et al., arXiv, 2304.08446v4 (2023)

UniMat + Diffusion (Google Deepmind)



Yang et al., arXiv, 2311.09235v2 (2023)

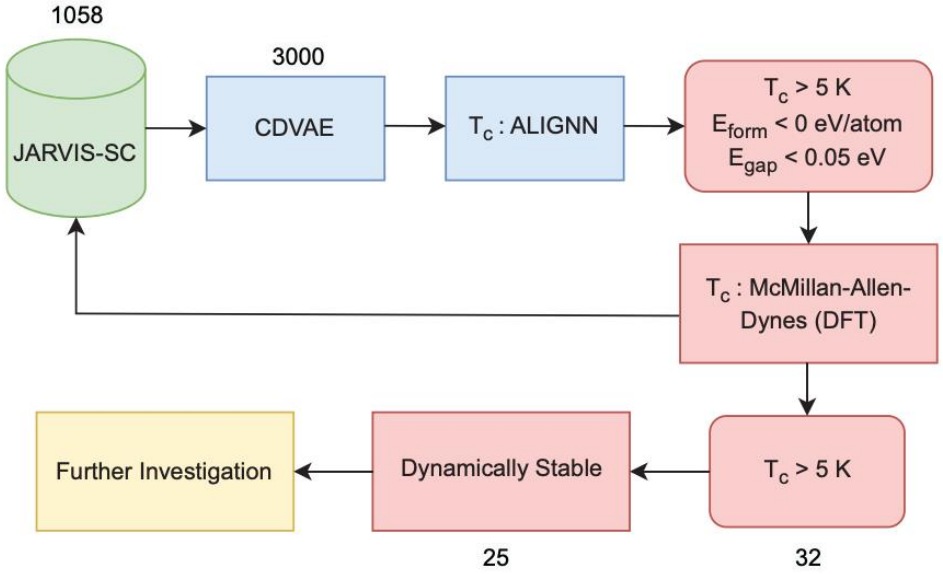
MatterGen (Microsoft)

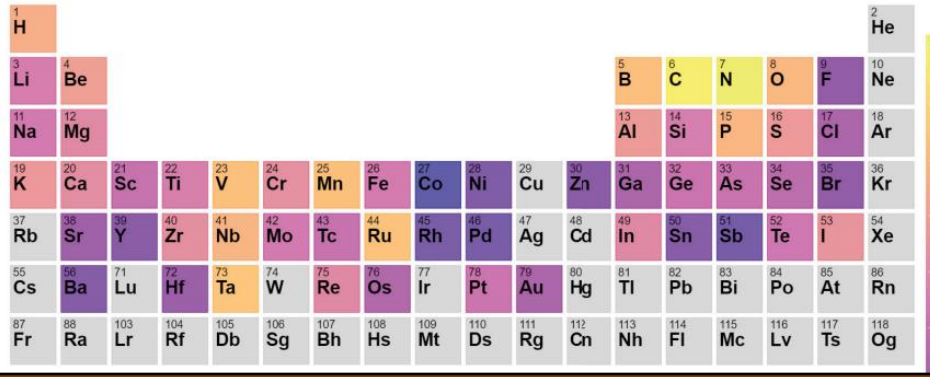


Zeni et al., arXiv, 2312.03687v2 (2024)

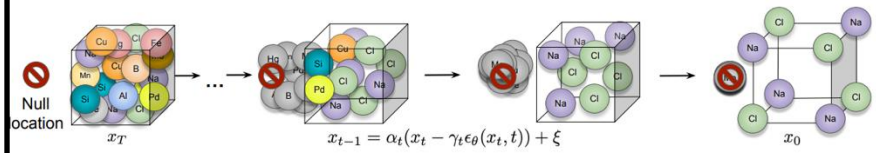
Diffusion models in action

Inverse design of new superconductors



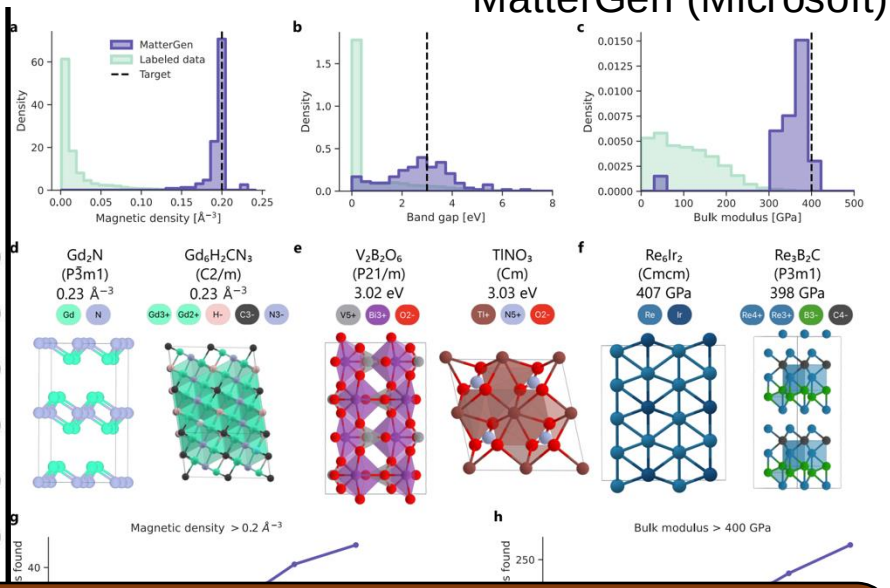


UniMat + Diffusion (Google Deepmind)



Yang et al., arXiv, 2311.09235v2 (2023)

MatterGen (Microsoft)



Summary:

- Diffusion models generating structures can accelerate materials discovery
- Still in nascent stages, can generate ‘bad’ structures

Conclusions and some thoughts to chew

- Designing better materials critical for performance improvement in several applications
  - Computations + ML can significantly accelerate materials design

- Different ways to use ML (or precursors to ML)
  - Regressions (or classifications): predict properties using experimental/calculated properties
  - Interatomic potentials: model larger/longer phenomena on a dynamic lattice
  - Diffusion and language models, transfer learning

- Materials science is a data-limited domain
  - Garbage in = Garbage out; data normalization
  - Real vs. synthetic data
  - What model to choose? Simple models are usually better
  - 'Real' success stories: still few, possibly in development
  - Don't do ML just because you can (hammer doesn't beget a nail)
  - Construct models with care: overfitting, lack of transferability
  - Test and validate, validate and test, and …

saigautamg@iisc.ac.in; https://sai-mat-group.github.io; https://github.com/sai-mat-group