From Atoms to Accelerated Discovery: Machine Learning Pathways for Materials Innovation

Discuss: Descriptors / Unsupervised Learning

Nong Artrith

Debye Institute for Nanomaterials Science
Faculty of Science, Utrecht University, The Netherlands
Email: n.artrith@uu.nl

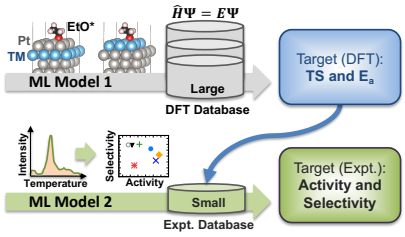
2026 CaMML - Chemistry and Materials Machine Learning School
Daresbury Laboratory, UK, April 14th , 2026

1

Computational Discovery of Energy Materials and Interpretation of Experimental Observations with Atomistic First-Principles Methods and Machine Learning

Research Activities

Nong Artrith and Team
Computational Materials Science and Machine Learning (ML)

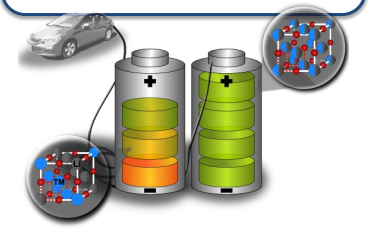


1. Machine Learning & Data Science

- Data from simulation and experiment (with collaborators)
- Accelerating DFT calculations
- Predicting materials properties
- Data mining for materials discovery

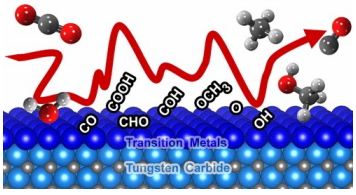
2. Energy Storage

- Storage for renewable energy
- Enable electric vehicles
- Improve portable electronics



3. Energy Conversion

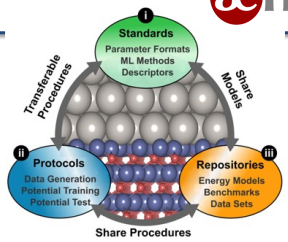
- Replace fossil fuels (oil/gas)
- Produce synthetic fuels using renewable energy



4. Method Development

- **Open-source** ML modeling tools

http://nartrith.atomistic.net



2

We know the laws of physics that govern (most) materials properties

- Quantum mechanics describes extremely accurately how chemical bonds form
- Already developed 100 years ago!
- If we can simulate QM, we can predict new materials *in silico*

But length and time scales are severely limited!

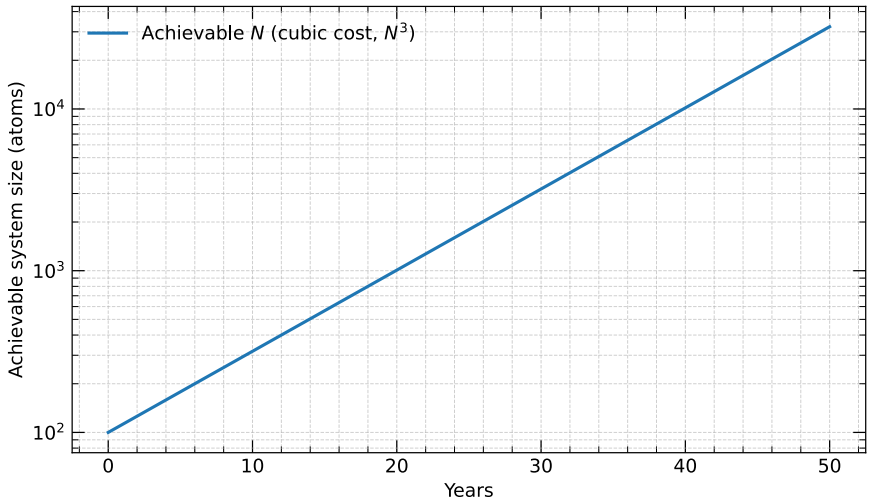
Cost of DFT

$$\propto (N_{\text{atoms}})^3$$

Moore's law

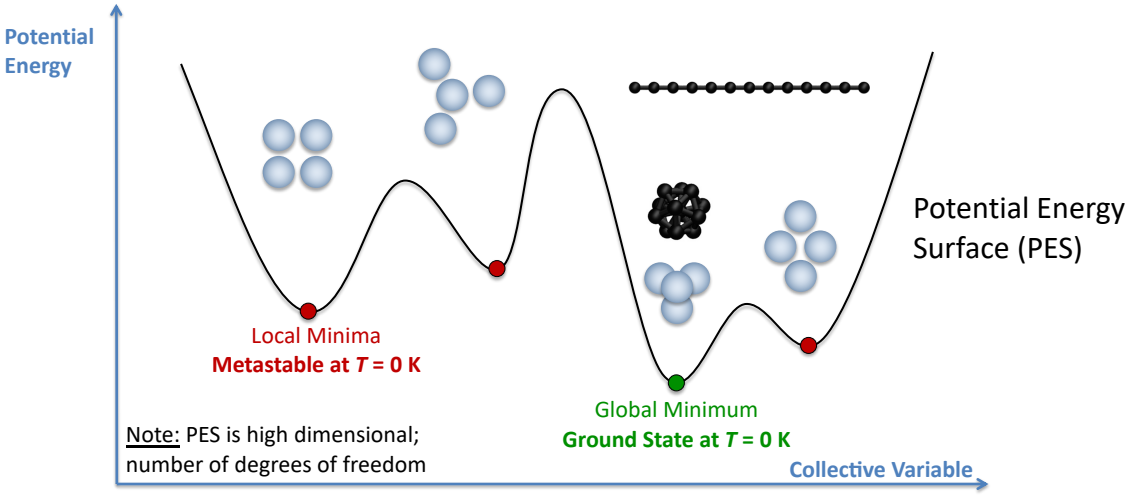
$$\propto 2^{N_{\text{years}}}$$

Practical limit:
Tens of thousands of atoms.



Atomic-Scale Simulations Approximate the Potential Energy Surface (PES)

- The atomic arrangement with the lowest energy is the ground state structure
- Metastable structures also occur in nature (e.g., polymorphs/crystal structures)



5

How can Machine Learning be used for Materials Chemistry?

- Machine Learning can be used as:
  A method for data analysis that automates analytical model building
- Learn from data, identify patterns, and make decisions with minimal human intervention
- Data can be from experimental measurement or from computational modeling

6

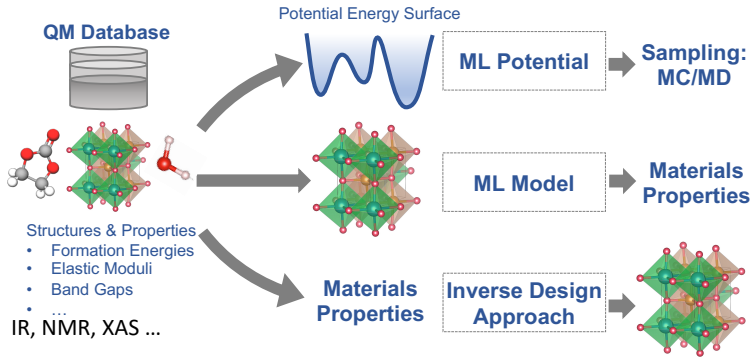
Textbooks

- Christopher Bishop,
  Pattern Recognition and Machine Learning,
  Springer **2006**
- Kevin P. Murphy,
  Machine Learning: A Probabilistic Perspective,
  MIT Press **2012**
- Grégoire Montavon, Geneviève Orr, Klaus-Robert Müller,
  Neural Networks: Tricks of the Trade,
  Springer **2012**

7

Machine Learning for Accelerated Modeling

- Machine learning (ML) can be used to accelerate atomic-scale simulations
- ML models can also be trained on measured data to predict materials
- ML is also useful for the interpretation of measurements (e.g., from spectroscopy)



H. Guo, Q. Wang, A. Stuke, A. Urban, and **N. Artrith**, *Frontiers in Energy Research* **9**, 2021, 695902

8

Outline: Principles of Machine Learning

1. Popular Learning Types and Their Applications
   - Supervised learning
   - Unsupervised learning
   - Semi-supervised learning
   - Reinforcement learning
2. Models and Model Training
3. Model Validation
4. Materials Science Examples

9

1.1 Supervised Learning

Supervised learning: Learn from labeled data

- Data pairs (x, y) are available to train the model
- x is the *feature*, y is the *label*
- ML model learns to predict y for unseen x
- Typical applications: Classification, regression

10

Classification: Predict Label of New Data Points

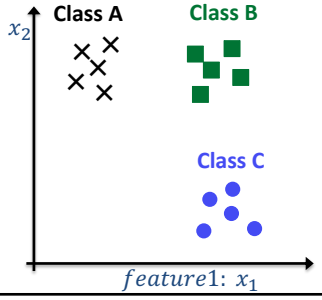
- **Labeled** data points used for *training*

(x, y) → Model

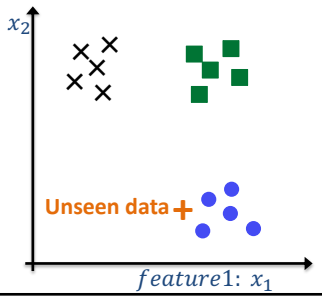
- *Prediction* of **discrete** labels for **new** data points

x → Model → y

Reference Data: *training*



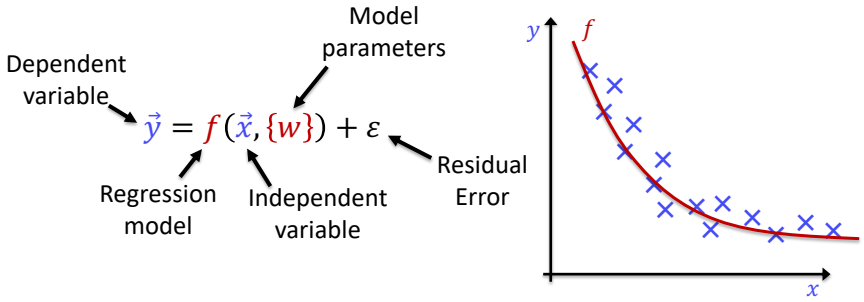
Predict Label of Unseen Data Point



11

Regression: Estimating Variable Relationships

- Given data points $(\vec{x}, \vec{y})$ determine a **continuous** function that relates the independent variables $\vec{x}$ to the dependent variables $\vec{y}$



- Machine learning provides *models* and methods for determining the model parameters by *training*

12

1.2 Unsupervised Learning

Unsupervised learning: Detect structure in data

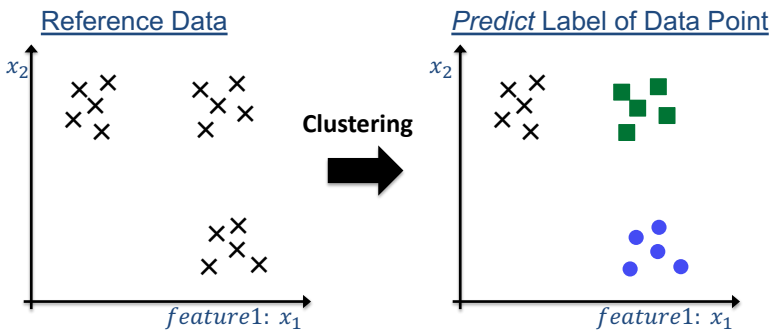
- Only features given, but *no labels*
- Given data points $\{x\}$; determine relationship between data points
- Typical applications: Clustering, dimension reduction

13

Clustering: Determine Which Data Belongs Together

- Only **features** used for *training*

- *Prediction* of the labels for **original** data points



14

2. Models and Model Training

15

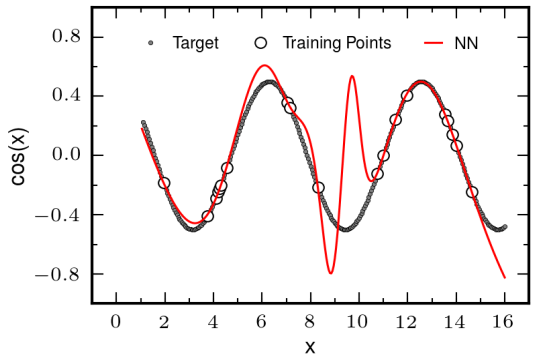
The Right Choice of ML Model Depends on the Data

- **Size of the reference data set**
  - Billions of Google search requests
    vs.
    Hundreds of materials compositions
- **Dimension of the learning problem**
  - Pixels in a photo (1600 x 800)
    vs.
    Free energy G(T,p)
- **…**

16

Complex Models can be *Overfitted*

- Model accurately reproduces data used for training
- But model interpolates incorrectly → poor **generalization**



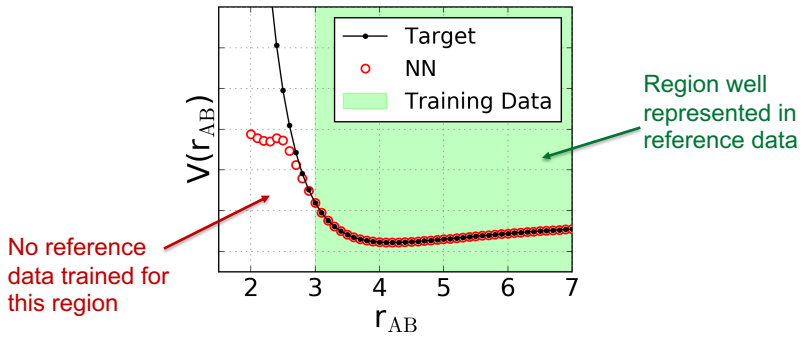
A.M. Miksch, T. Morawietz, J. Kästner, A. Urban, **N. Artrith***,
Machine Learning: Science and Technology, **2**, 031001**(2021),** DOI: https://doi.org/10.1088/2632-2153/abfd96.

Code: https://github.com/atomisticnet/MLP-beginners-guide

17

Machine Learning is Poor for *Extrapolation*

- ANNs are flexible functions
- Can well interpolate between reference data points
- But **no good for extrapolation** beyond trained region



A.M. Miksch, T. Morawietz, J. Kästner, A. Urban, **N. Artrith***,
Machine Learning: Science and Technology, **2**, 031001**(2021),** DOI: https://doi.org/10.1088/2632-2153/abfd96.

Code: https://github.com/atomisticnet/MLP-beginners-guide

18

3. Model Validation

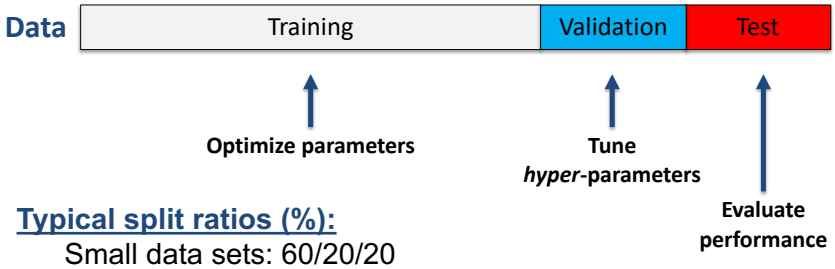
Machine Leaning models have to be validated.

19

ML Models Have to be Validated

Solution: Split reference data set into different parts

- **Training set:** data samples that are used for training
- **Validation set:** used to tune hyper parameters
- **Test set:** used to evaluate performance (error estimate)



Typical split ratios (%):

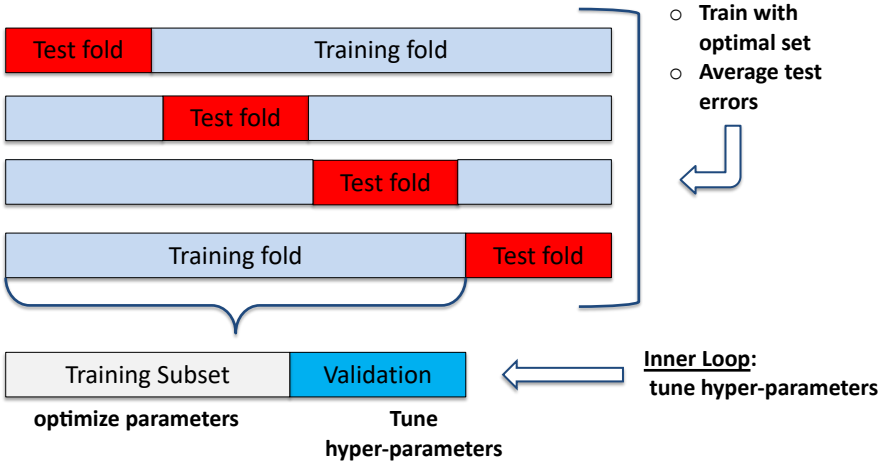
Small data sets: 60/20/20

Large data sets: 98/1/1

20

Cross-Validation

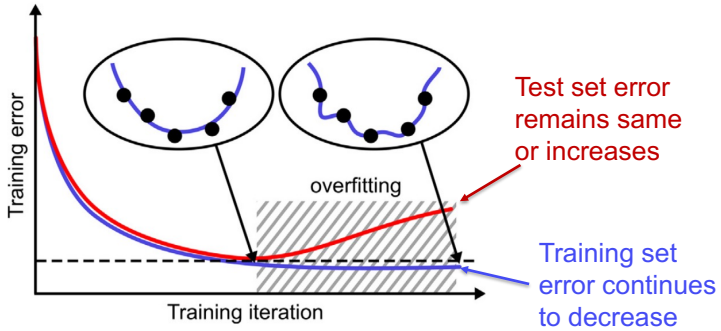
- **Problem**: hyper-parameters can be overfit to test data
- **Solution**: Use many splits
  - Example: Nested cross-validation



21

Early Stopping: Prevent Overfitting

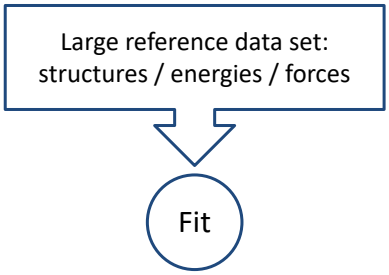
- Monitor mean error of samples in test set during training
- Test set error does no longer improve when the model is overfitted



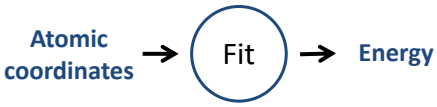
22

Machine-Learning Interatomic Potentials: Learn from QM, Inference with ML

Interpolation



Use fit as a black box



Requires specialized regression methods
- Large number of fitting parameters
- High-dimensional problem (many DoF)

→ **Machine learning**

Not a Novel Idea!

- Machine learning is used for classification (e.g., image recognition) and regression (function fitting)
- Well suited for regression:
  → *Gaussian Process Regression* (or *Kriging*)
  → *Artificial Neural Networks*
- Both methods have been used for potential fitting

Early ANN Potentials Were Published in the 1990s

ELSEVIER

6 June 1997

Chemical Physics Letters 271 (1997) 152–156

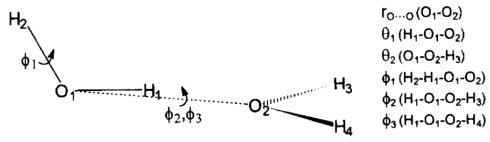
CHEMICAL PHYSICS LETTERS

Description of the potential energy surface of the water dimer with an artificial neural network

Kyoung Tai No [a,1], Byung Ha Chang [a], Su Yeon Kim [a], Mu Shik Jhon [b,1], Harold A. Scheraga [c]

[a] Department of Chemistry, Soong Sil University, Seoul 156-743, South Korea
[b] Department of Chemistry, Korea Advanced Institute of Science and Technology, Taejon 305-701, South Korea
[c] Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14853-1301, USA

ANN Input:
Most molecular degrees of freedom (distances, angles, dihedrals); fixed O-H bond length



Drawbacks:

- Not symmetric wrt. exchange of equivalent atoms.
- Potential can **only** be used for the water dimer
- Extension to other systems is not straightforward

The Breakthrough: Learn the Atomic Energy

Inspiration from bond-order potentials:
Decompose total energy into **atomic contributions**

$$E = \sum_{i}^{\text{atoms}} E_i(\sigma_i)$$

E_i is an ANN

σ_i describes the local environment of atom i

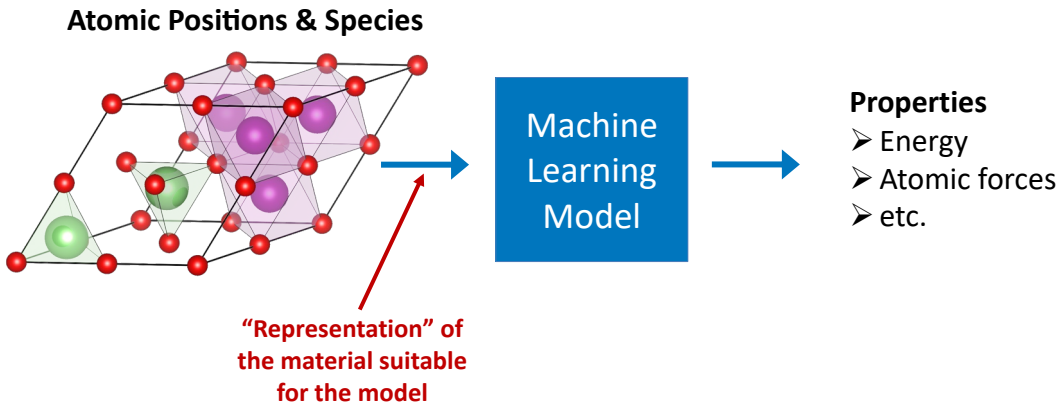
Note:

- The atomic energy is neither known nor uniquely defined.
- The target for the training is the total energy of the atomic structure.
- The model *learns* one possible atomic energy decomposition.

J. Behler, and M. Parrinello, *Phys. Rev. Lett.* **98** (2007) 146401.

How do you input a chemical compound into a machine-learning model?

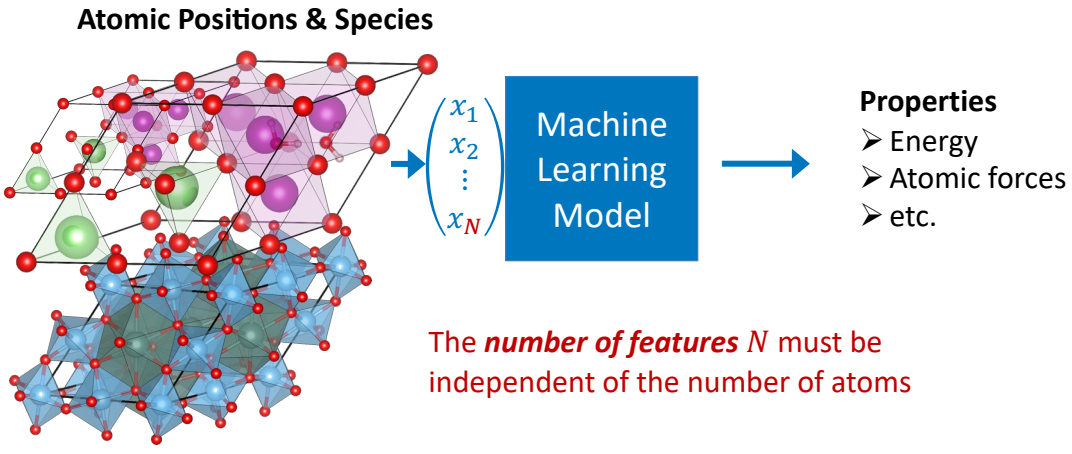
- Our model should work with any arbitrary atomic structure in a defined chemical space



27

Requirement 1: (Descriptor) *Feature vector* with constant dimension

- Many ML models expect an input vector with constant dimension

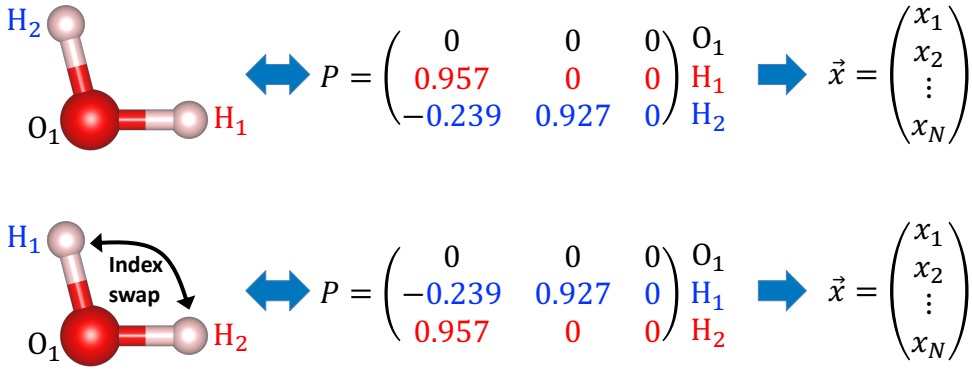


The ***number of features*** N must be independent of the number of atoms

28

Requirement 2: Exchanging equivalent atoms must not affect the result

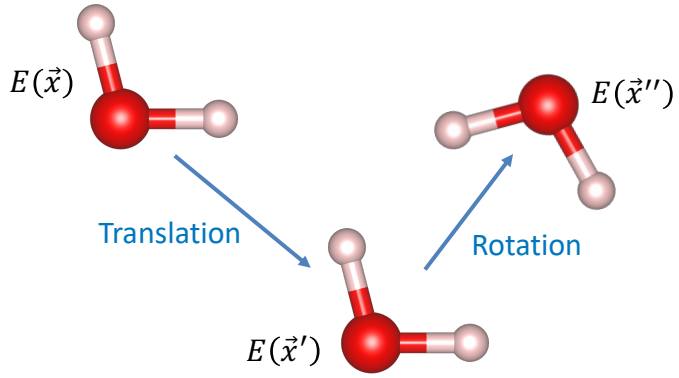
- Many ML models expect an input vector with constant dimension



29

Requirement 3: Scalar properties must be invariant wrt. translation & rotation

- The energy and other scalar properties, such as the charge, do not change upon translation, rotation, or reflection



$$E\left(P+\vec{t}\right)=E(P)$$

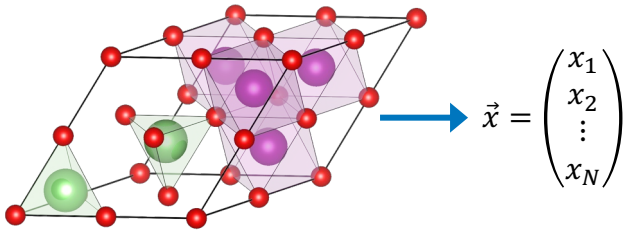
- $P=(\vec{r}_1,\vec{r}_2,\dots)^T$ is a matrix with the atomic positions
- $\vec{t}$ is an arbitrary translation
- $P+\vec{t}=\left(\vec{r}_1+\vec{t},\vec{r}_2+\vec{t},\dots\right)^T$

$$E(RP)=E(P)$$

- R is an arbitrary rotation matrix

30

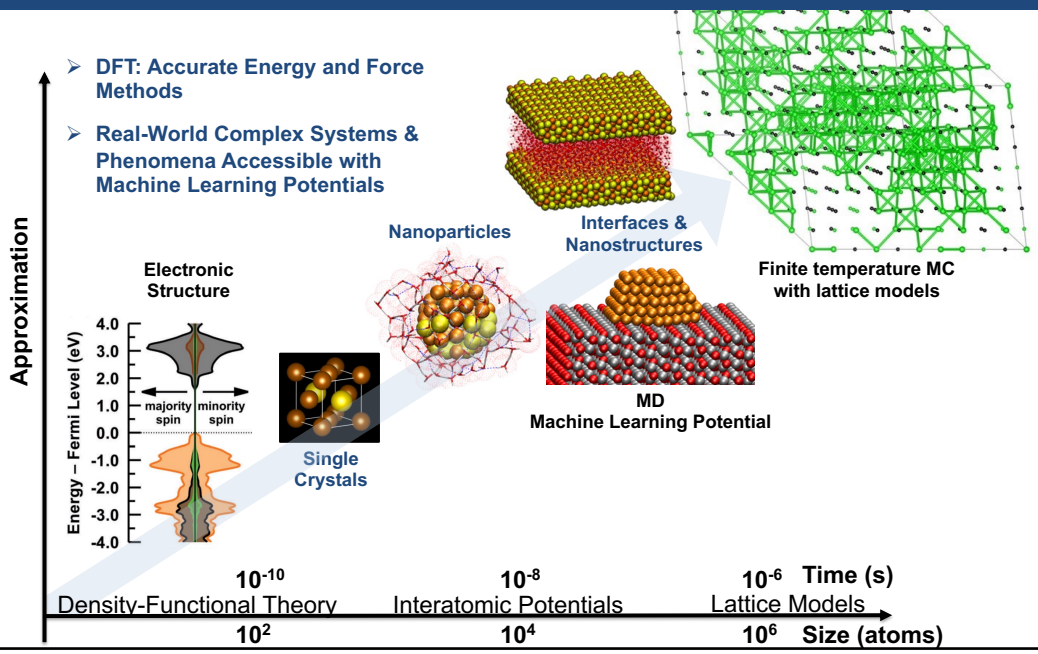
Summary of the requirements for materials features



$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

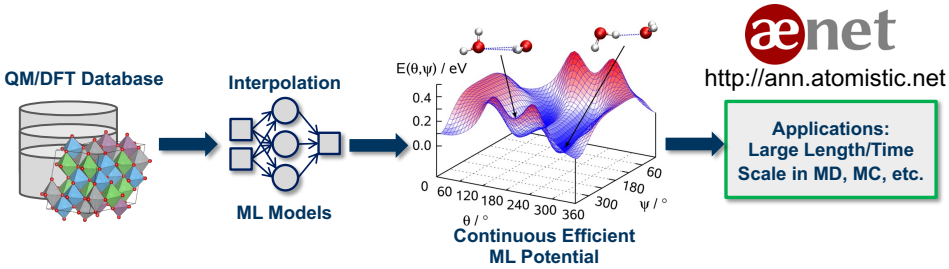
- Dimension independent of the number of atoms
- Invariant to the exchange of equivalent atoms
- Support for model E(3) equivariance; = energy invariant with respect to
  - Rotation,
  - Reflection, and
  - Translation

Accurate Simulation of Large-Scale Systems



Machine Learning Potentials for Complex Materials

• **Learn from electronic structure theory (DFT) how atoms interact**



• **Our Approach:**

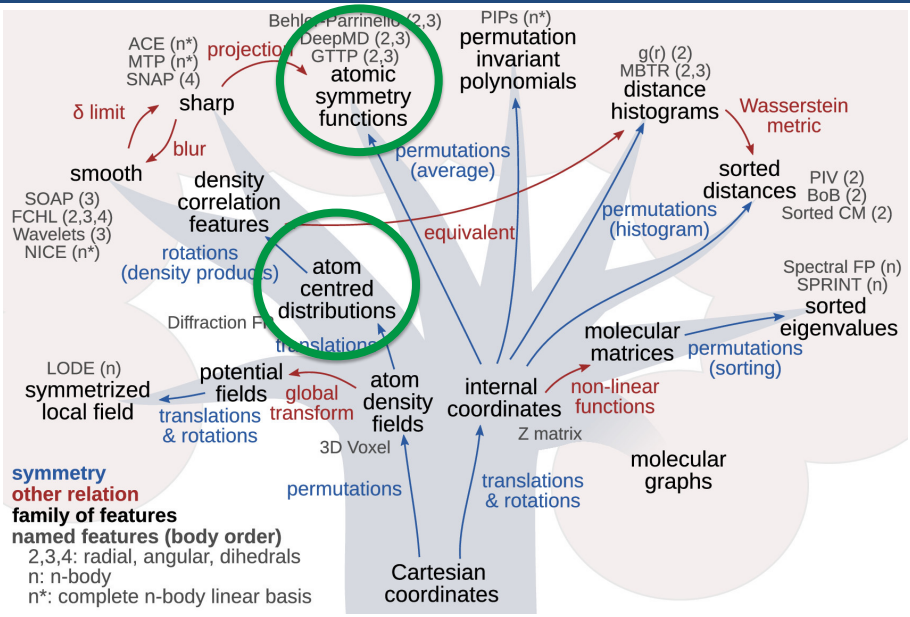
- Artificial neural networks for interpolation
- Only based on atomic positions

⇒ **As accurate as DFT**
⇒ **Reactive**
+ Full-dimensional

J. Behler, and M. Parrinello, *Phys. Rev. Lett.* 98, (**2007**) 146401.
N. Artrith, T. Morawietz, and J. Behler, *Phys. Rev. B* 83, (**2011**) 153101.
N. Artrith* and A. Urban, Comput. Mater. Sci. 114, (**2016**) 135-150.
AM. Cooper, J. Kästner, A. Urban, and **N. Artrith***, *npj Comput. Mater.* 6, (**2020**) 54.
M.S. Chen, T. Morawietz, H. Mori, T.M. Markland, and **N. Artrith***, J. Chem. Phys. 155, (**2021**) 074801
T. Morawietz and **N. Artrith***, *J. Comput. Aided. Mol. Des.* 35, (**2021**) 557.

33

Structural Representation Trees of Materials for MLIP Construction



F. Musil, M. Ceriotti et al, *Chem. Rev.* 2021, **121**, 9759–9815

34

Featurization: Representation Methods for Translating Materials into Vectors

Option 1 (relatively easy): **Ignore the atomic details of the crystal structure**

- Elements can be represented by their known properties, *i.e.*, from the periodic table
- Group & period, electronegativity, ionization energies, mass, quantum numbers, etc.
- Element-wise features need to be combined to represent compositions
  → weighted mean, standard deviation, other averages, etc.

Featurizing the structure (lattice & sites) as well is harder

Option 2 (harder): Also featurized the structure, *i.e.*, atomic positions/sites and lattice

- Need to (should) maintain invariants
  (rotation, translation, permutation of equivalent atoms)
- Ideally works with **different numbers of atoms**

Two principal directions:

- "Hand-crafted" representations = recipes that a human has come up with.
  For example, based on basis-set expansions.
- Learned representations = specialized ANNs such as E(3)-equivariant
  graph neural networks

Selection of hand-crafted (physics-based) material representations

Some hand-crafted representations (selection only)

Atomic sites (→ local representation)

- Atom-centered Symmetry functions, ACSF (Behler, Parrinello, *Phys. Rev. Lett.* **98**, 2007, 146401)
- Smooth Overlap of Atomic Positions, SOAP (Bartók, Kondor, Csányi, *Phys. Rev. B* **87**, 2013, 184115)
- Moment tensors (Shapeev, *Multiscale Model. Simul.* **14**, 2016, 1153– 1173)
- Atomic Cluster Expansion, ACE (Drautz, *Phys. Rev. B* **99**, 2019, 014104)

Entire structure (→ global representation)

- Diffraction patterns
- Coulomb matrices (Rupp … von Lilienfeld, *Phys. Rev. Lett.* **108**, 2012, 058301)
- Crystal-symmetry and sites (Gómez-Bombarelli and coworkers)

Recent reviews:

Himanen, …, Rinke, Foster, *Comput. Phys. Commun.* **247**, 2020, 106949.
Musli, …, Csányi, Ceriotti, *Chem. Rev.* ***121***, 2021, 9759–9815.
Damewood, …, Gómez-Bombarelli, *Annu. Rev. Mater. Res.* **53**, 2023, 12.1–12.28.

Local representation: Expansion of the RDF, ADF, etc. in Orthonormal Basis Set

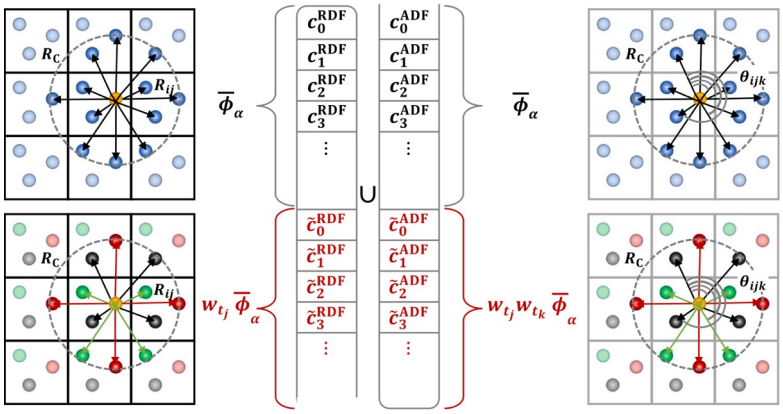
$$\mathrm{RDF}_i(r) = \sum_\alpha c_\alpha^{(2)} \phi_\alpha(r) \text{ for } 0 \le r \le R_c \qquad \mathrm{ADF}_i(\theta) = \sum_\alpha c_\alpha^{(3)} \phi_\alpha(\theta) \text{ for } 0 \le \theta \le \pi$$

$\{\phi_i\}$ is a complete orthonormal basis.

Features: Expansion coefficients **with and without** element-specific weights.

$$c_\alpha^{(2)} = \sum_{j \ne i} w_{t_j} \bar{\phi}_\alpha(R_{ij}) f_c(R_{ij})$$

$$c_\alpha^{(3)} = \sum_{j,k \ne i} w_{t_j} w_{t_k} \bar{\phi}_\alpha(\theta_{ijk}) f_c(R_{ij}) f_c(R_{ik})$$

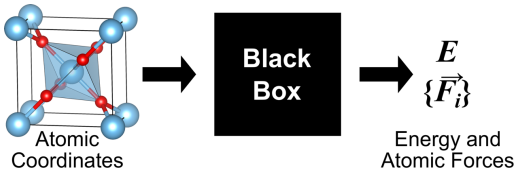


Method: N. Artrith, A. Urban, G. Ceder, *Phys. Rev. B* **96**, 2017, 014112.
Figure: A.M. Miksch, T. Morawietz, J. Kästner, A. Urban, N. Artrith, *Mach. Learn. Sci. Tech.* **2**, 2021, 031001.

Our Model ænet: Actively Developed Since ~2011. First Public Version in 2016.

Machine-learned interatomic potentials (MLIPs) as "drop-in replacement" for DFT

- Trained on first principles reference data
- Nearly as accurate as the reference method
- But very data hungry; can require thousands of reference calculations



ænet

https://github.com/atomisticnet

N. Artrith and A. Urban, *Comput. Mater. Sci.* **114** (2016) 135-150.
N. Artrith, A. Urban, G. Ceder, *Phys. Rev. B* **96** (2017) 014112.
A. Cooper, J. Kästner, A. Urban, N. Artrith, *npj Comput. Mater.* **6** (2020) 54.
I. W. Yeu et al., *npj Comput. Mater.* **11** (2025) 156.

The ænet Package

First publicly available implementation of the Machine Learning Potential (MLP) method!

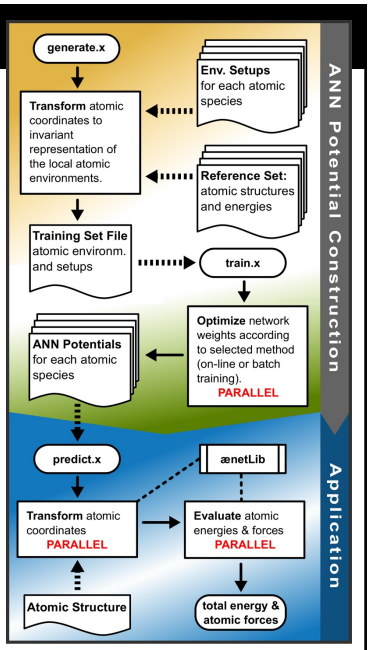
The Atomic Energy Network

ænet

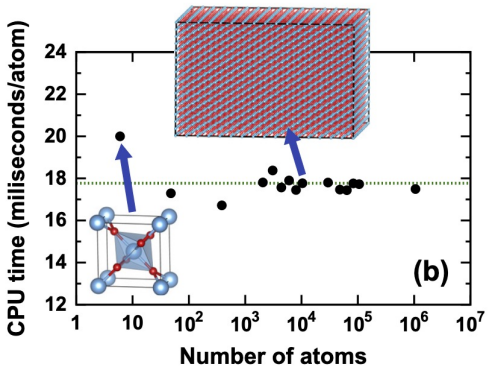
https://github.com/atomisticnet

Modules for potential fit and for application

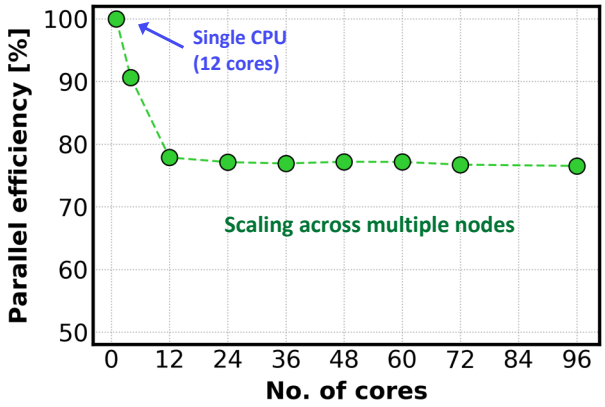
- **C compatible library & Python interface**
- **Interfaces with simulations software: ASE, LAMMPS, Tinker, DL_POLY**
- **Efficient (approximate) force pre-training**



Ænet Potentials Scale Linearly and are Highly Parallelizable

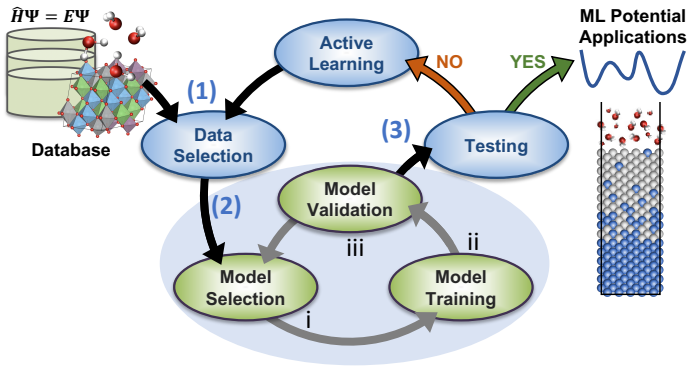


N. Artrith and A. Urban,
Comput. Mater. Sci. **114** (2016) 135-150.



M.S. Chen, T. Morawietz, H. Mori, T.E. Markland, N. Artrith,
J. Chem. Phys. **155**, 074801 (**2021**).

MLIPs are Typically Constructed in an Iterative (Active Learning) Fashion

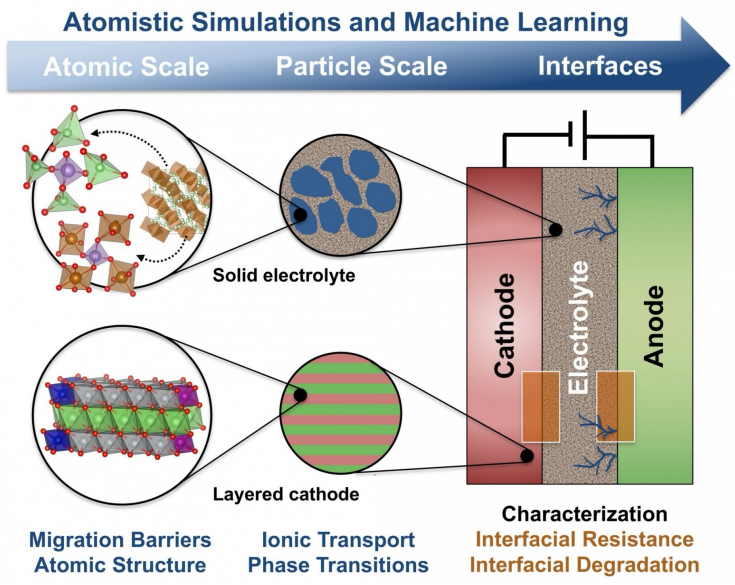


A.M. Miksch, T. Morawietz, J. Kästner, A. Urban, N. Artrith,
Mach. Learn.: Sci. Tech. **2** (2021) 031001.
Code: https://github.com/atomisticnet/MLP-beginners-guide

Some Applications

43

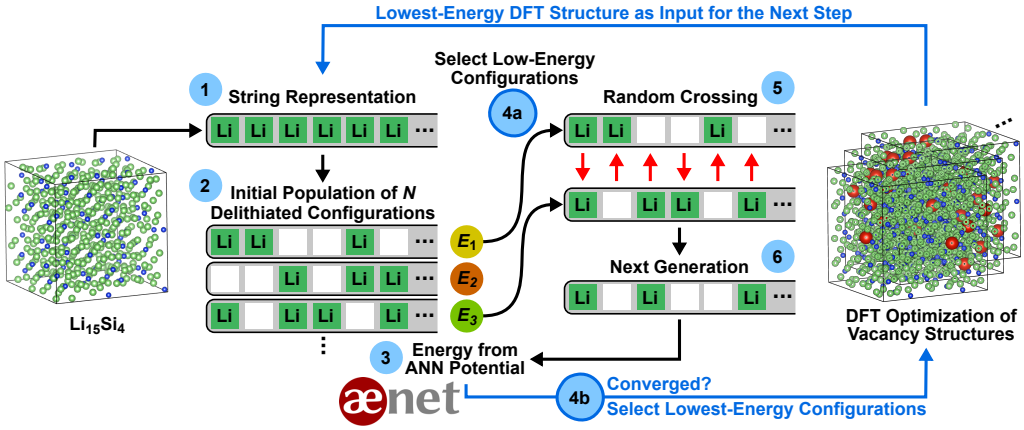
Simulating Batteries with MLIPs



44

Enable the Sampling of Amorphous Materials

Describing amorphous phases requires large numbers of atoms due to the lack of long-range order



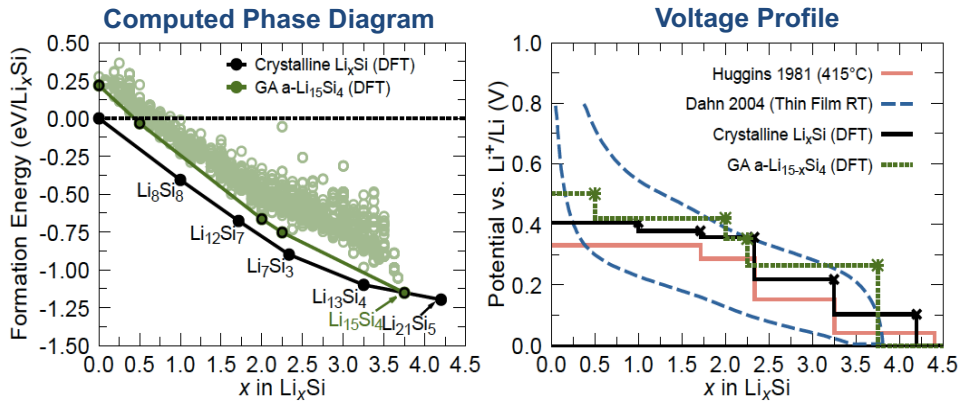
N. Artrith, A. Urban, G. Ceder, *J. Chem. Phys.* **148** (2017) 241711.

45

Quantitative Modeling of Amorphous Materials Feasible

ANN potential predicts phase diagram and voltage profiles in agreement with experiment

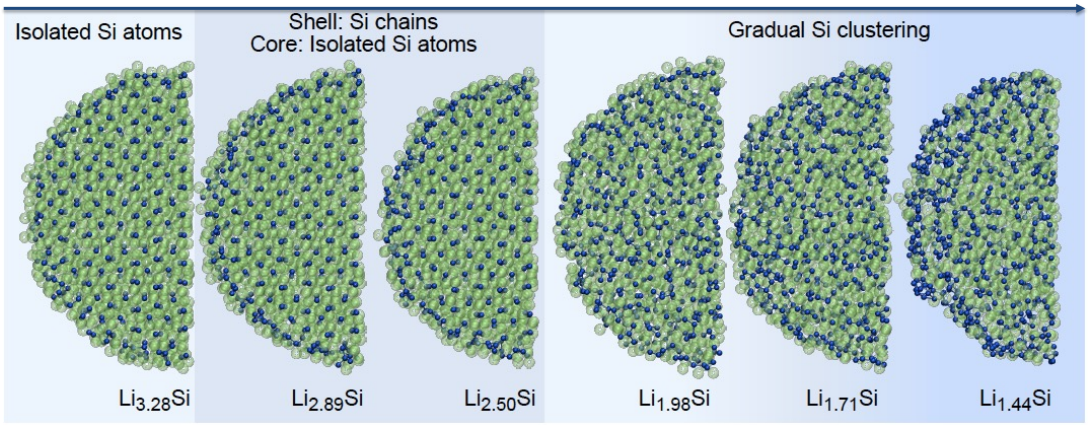
N. Artrith, A. Urban, and G. Ceder, *J. Chem. Phys.* **148,** (2018) 241711.



Please see this reference for how to calculate a DFT phase diagram:
Computational understanding of Li-ion batteries. *npj Comput Mater* **2**, 16002 (2016). https://doi.org/10.1038/npjcompumats.2016.2

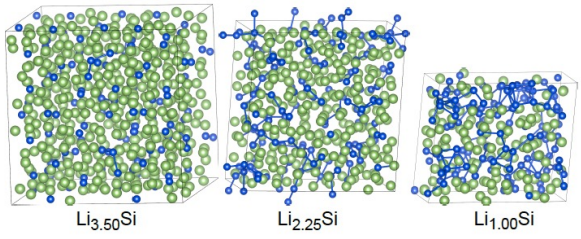
46

Makes it Possible to Model Entire Nanoparticles with Tens of Thousands of Atoms



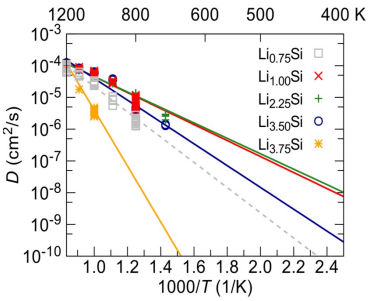
47

Predictions are in Quantitative Agreement with Experiment



Si
Li

Li Transport (Arrhenius Plot)



Experimental References

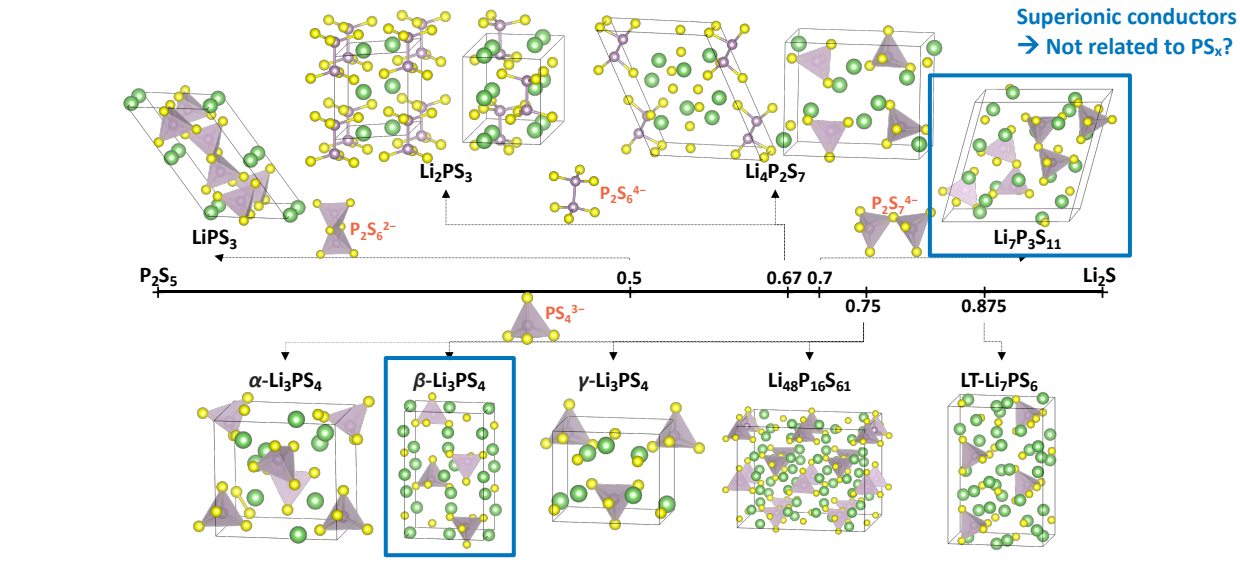
D (cm^2/s)	Method	Reference
10^{-10}	EIS	[1]
10^{-12}	CV, EIS, GITT	[2]
10^{-14}	EIS, PITT	[3]
10^{-14}–10^{-13}	PITT	[4]

ANN Results

x_{Li}	E_a (eV)	D (cm^{2}s^{-1})
0.75	0.789	1.154 x 10^{-14}
1.00	0.500	5.986 × 10^{-11}
2.25	0.483	9.607 × 10^{-11}
3.50	0.682	3.820 × 10^{-13}

48

Lithium Thiophosphates (LPS) are Promising Solid Electrolytes



H. Guo, Q. Wang, A. Urban, N. Artrith, *Chem. Mater.* **34** (2022) 6702-6712.

49

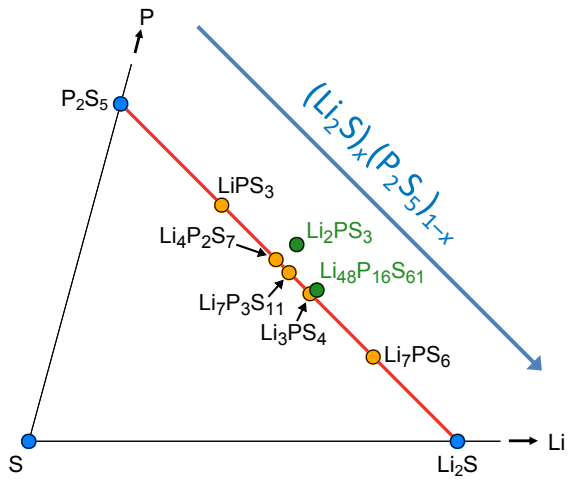
Lithium Thiophosphates (LPS) are Promising Solid Electrolytes

Challenge:

- Known superionic LPS are metastable and degrade rapidly
- Promising **amorphous** LPS phases also not stable enough
- Does an LPS phase exist that is superionic AND stable over time?

50

Approach: First Map the Phase Diagram of Crystalline and Amorphous LPS



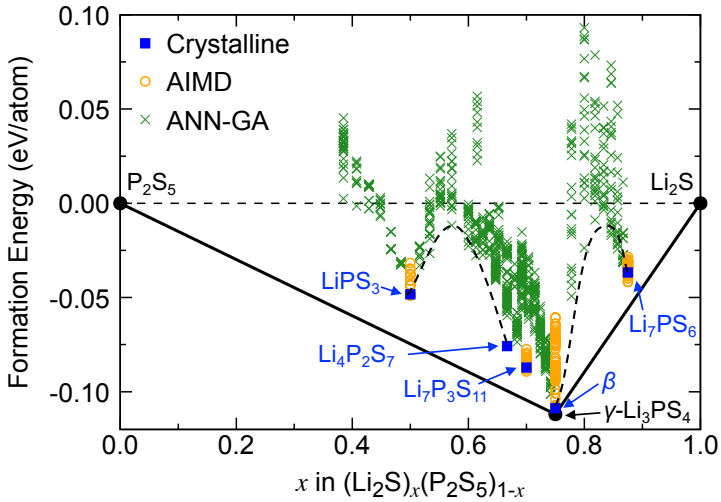
ML-accelerated sampling

Vary x in $(Li_2S)_x(P_2S_5)_{1-x}$

1. Start with supercell of crystalline LPS
2. 2 Li and 1 S (or 2 P and 5 S)
3. Search for most stable arrangement of the remaining atoms
4. Repeat steps 2 & 3

Only γ-Li_3PS_4 is Thermodynamically Stable at Zero Temperature

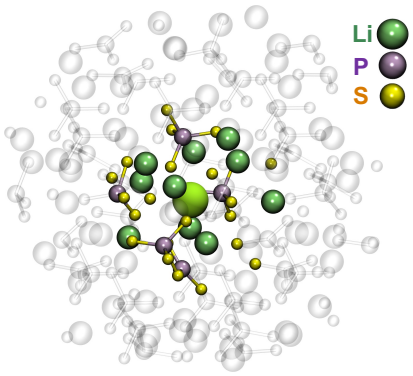
0K LPS phase diagram: which compositions can likely be made → "Energy above hull"

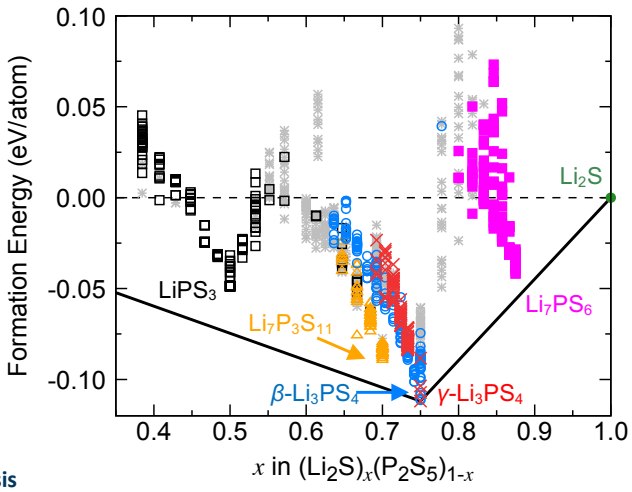


Which compositions are best for solid electrolytes?

H. Guo, Q. Wang, A. Urban, N. Artrith, *Chem. Mater.* **34** (2022) 6702-6712.

Classifying Amorphous Phases based on their Local Li Site Environment



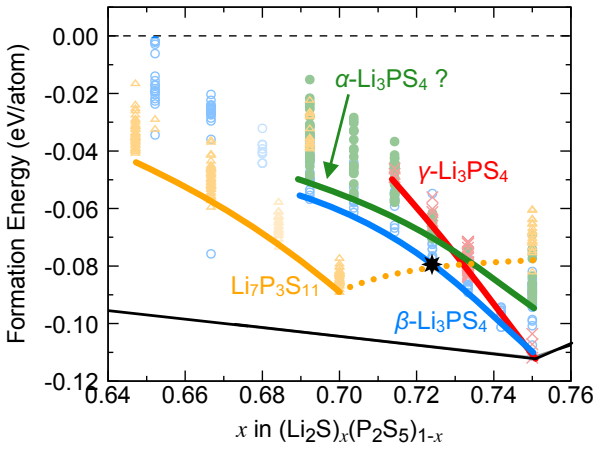


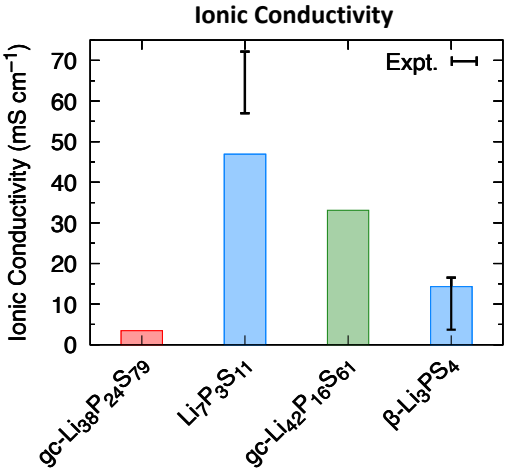
➢ **Unsupervised learning for similarity classification using our ænet descriptors and k-means clustering analysis of Li-site environments in our structural databases.**

53

LPS Compositions Between $Li_7P_3S_{11}$ and β-Li_3PS_4 are Likely Stable Conductors

Molecular dynamics simulations indicate that gc-$Li_{42}P_{16}S_{61}$ is a stable superionic conductor





H. Guo, Q. Wang, A. Urban, N. Artrith, *Chem. Mater.* **34** (2022) 6702-6712.

54

On the Horizon: Universal MLIPs

55

Universal MLIPs

Matbench Discovery

https://matbench-discovery.materialsproject.org

Full Test Set | Unique Prototypes | 10k Most Stable

Model	CPS ↑	Acc	F1	DAF	Prec	MAE	R^2	κ_{SRME}	RMSD	Training Set	Params	Targets	Date Added
EquiformerV3+DeNS-OAM	0.902	0.978	0.931	6.074	0.928	0.018	0.868	0.118	0.059	6.6M (113M) OMat24+MPtrj+sAlex	30.3M	EFS_G	2026-04-07
PET-OAM-XL	0.898	0.977	0.924	6.075	0.929	0.019	0.864	0.119	0.060	6.6M (113M) OMat24+sAlex+MPtrj	730M	EFS_G	2026-01-10
TACE-OAM-L	0.889	0.972	0.910	5.898	0.902	0.020	0.868	0.126	0.061	6.6M (113M) OMat24+sAlex+MPtrj	82.9M	EFS_G	2026-04-09
eSEN-30M-OAM	0.888	0.977	0.925	6.069	0.928	0.018	0.866	0.170	0.061	6.6M (113M) OMat24+MPtrj+sAlex	30.2M	EFS_G	2025-03-17
EquFlash	0.888	0.975	0.919	5.983	0.915	0.019	0.871	0.158	0.060	6.6M (113M) OMat24+MPtrj+sAlex	28.7M	EFS_G	2025-06-23
Nequip-OAM-XL	0.886	0.971	0.906	5.869	0.897	0.020	0.872	0.125	0.063	6.6M (113M) OMat24+sAlex+MPtrj	32.1M	EFS_G	2025-11-30
MatRIS-10M-OAM	0.877	0.976	0.921	6.039	0.923	0.019	0.871	0.218	0.060	6.6M (113M) OMat24+sAlex+MPtrj	10.4M	EFS_GM	2025-10-29
SevenNet-Omni-i12	0.873	0.971	0.906	5.954	0.910	0.021	0.868	0.192	0.062	243M COSMOSDataset	54.9M	EFS_G	2026-01-12
Nequip-OAM-L	0.870	0.967	0.893	5.823	0.890	0.022	0.865	0.166	0.065	6.6M (113M) OMat24+sAlex+MPtrj	9.6M	EFS_G	2025-09-08
GRACE-2L-OAM-L	0.865	0.964	0.883	5.840	0.893	0.022	0.862	0.169	0.064	6.6M (113M) OMat24+sAlex+MPtrj	26.4M	EFS_G	2025-09-09
ORB v3	0.860	0.971	0.905	5.912	0.904	0.024	0.821	0.210	0.075	6.47M (133M) MPtrj+Alex+OMat24	25.5M	EFS_G	2025-04-05
Allegro-OAM-L	0.840	0.966	0.895	5.674	0.867	0.022	0.868	0.319	0.065	6.6M (113M) OMat24+sAlex+MPtrj	9.7M	EFS_G	2025-09-08
GRACE-2L-OAM	0.837	0.963	0.880	5.774	0.883	0.023	0.862	0.294	0.067	6.6M (113M) OMat24+sAlex+MPtrj	12.6M	EFS_G	2025-02-06
EquiformerV3+DeNS-MP	0.830	0.956	0.863	5.479	0.838	0.029	0.840	0.275	0.070	146k (1.58M) MPtrj	30.3M	EFS_G	2026-04-07
DPA-3.1-3M-FT	0.802	0.963	0.884	5.667	0.866	0.023	0.869	0.469	0.069	163M OpenLAM	3.27M	EFS_G	2025-06-05
eSEN-30M-MP	0.797	0.946	0.831	5.260	0.804	0.033	0.822	0.340	0.075	146k (1.58M) MPtrj	30.1M	EFS_G	2025-03-17
MACE-MPA-0	0.795	0.954	0.852	5.582	0.853	0.028	0.842	0.412	0.073	3.37M (12M) MPtrj+sAlex	9.06M	EFS_G	2024-12-09
MatRIS-10M-MP	0.778	0.951	0.847	5.422	0.829	0.031	0.824	0.489	0.072	146k (1.58M) MPtrj	10.4M	EFS_GM	2025-10-29
AlphaNet-v1-OAM	0.769	0.968	0.901	5.747	0.879	0.024	0.831	0.643	0.079	6.6M (113M) OMat24+sAlex+MPtrj	4.65M	EFS_G	2025-05-12

56

Can we Train a Universal MLIP to Replace DFT in All Simulations?

nature computational science

Article

https://doi.org/10.1038/s43588-022-00349-3

A universal graph deep learning interatomic potential for the periodic table

Received: 18 March 2022

Accepted: 5 October 2022

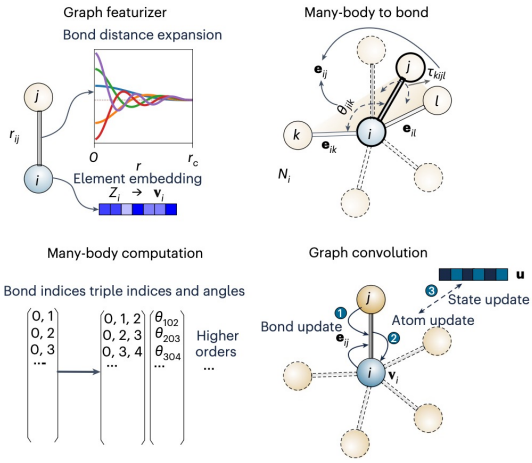
Published online: 28 November 2022

Check for updates

Chi Chen & Shyue Ping Ong

Interatomic potentials (IAPs), which describe the potential energy surface of atoms, are a fundamental input for atomistic simulations. However, existing IAPs are either fitted to narrow chemistries or too inaccurate for general applications. Here we report a universal IAP for materials based on graph neural networks with three-body interactions (M3GNet). The M3GNet IAP was trained on the massive database of structural relaxations performed by the Materials Project over the past ten years and has broad applications in structural relaxation, dynamic simulations and property prediction of materials across diverse chemical spaces. About 1.8 million materials from a screening of 31 million hypothetical crystal structures were identified to be potentially stable against existing Materials Project crystals based on M3GNet energies. Of the top 2,000 materials with the lowest energies above the convex hull, 1,578 were verified to be stable using density functional theory calculations. These results demonstrate a machine learning-accelerated pathway to the discovery of synthesizable materials with exceptional properties.

Trained on 1.8 million DFT calculations!



C. Chen and S. P. Ong, *Nat. Comput. Sci.* **2** (2022) 718–728.

Tech Industry Shows What is Possible When Money is Unlimited

UMA: A Family of Universal Models for Atoms

Brandon M. Wood[1,†,*], Misko Dzamba[1,*], Xiang Fu[1,*], Meng Gao[1,*], Muhammed Shuaibi[1,*], Luis Barroso-Luque[1], Kareem Abdelmaqsoud[2], Vahe Gharakhanyan[1], John R. Kitchin[2], Daniel S. Levine[1], Kyle Michel[1], Anuroop Sriram[1], Taco Cohen[1], Abhishek Das[1], Ammar Rizvi[1], Sushree Jagriti Sahoo[1], Zachary W. Ulissi[1], C. Lawrence Zitnick[1,†]

[1]FAIR at Meta, [2]Department of Chemical Engineering, Carnegie Mellon University
[*]Co-first Author, [†]Co-corresponding Author

The ability to quickly and accurately compute properties from atomic simulations is critical for advancing a large number of applications in chemistry and materials science including drug discovery, energy storage, and semiconductor manufacturing. To address this need, Meta FAIR presents a family of Universal Models for Atoms (UMA), designed to push the frontier of speed, accuracy, and generalization. UMA models are trained on half a billion unique 3D atomic structures (the largest training runs to date) by compiling data across multiple chemical domains, e.g. molecules, materials, and catalysts. We develop empirical scaling laws to help understand how to increase model capacity alongside dataset size to achieve the best accuracy. The UMA small and medium models utilize a novel architectural design we refer to as mixture of linear experts that enables increasing model capacity without sacrificing speed. For example, UMA-medium has 1.4B parameters but only ~50M active parameters per atomic structure. We evaluate UMA models on a diverse set of applications across multiple domains and find that, remarkably, a single model without any fine-tuning can perform similarly or better than specialized models. We are releasing the UMA code, weights, and associated data to accelerate computational workflows and enable the community to continue to build increasingly capable AI models.

Models: https://huggingface.co/facebook/UMA
Code: https://github.com/facebookresearch/fairchem
Correspondence: B.M.W. (bmwood@meta.com), C.L.Z. (zitnick@meta.com)

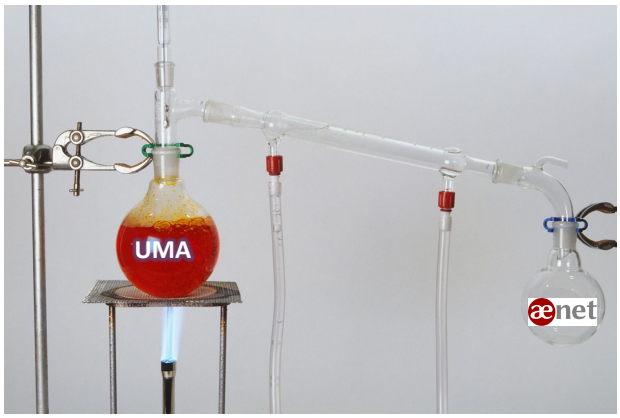
Meta

- Trained on ½ billion DFT calculations
- Our benchmarks show: Reproduces DFT for **molecular simulations** with great reliability
- Not yet reliable for inorganic materials

Other “players”

- Bytedance (Bytedance AI Molecular BOOster, BAMBOO 2024)
- Microsoft (e.g., SimPoly, 2025)
- Schrödinger

Is ænet a Relic of the Past? — Model Distillation



- Graph neural nets are complex and computationally demanding
- No strict linear scaling with the number of atoms
- Require GPU hardware

Renewed interest in ænet

- Get domain-specialized models for efficient simulations
- Distillation in Python, but inference in Fortran
- ~ 1000 times more efficient

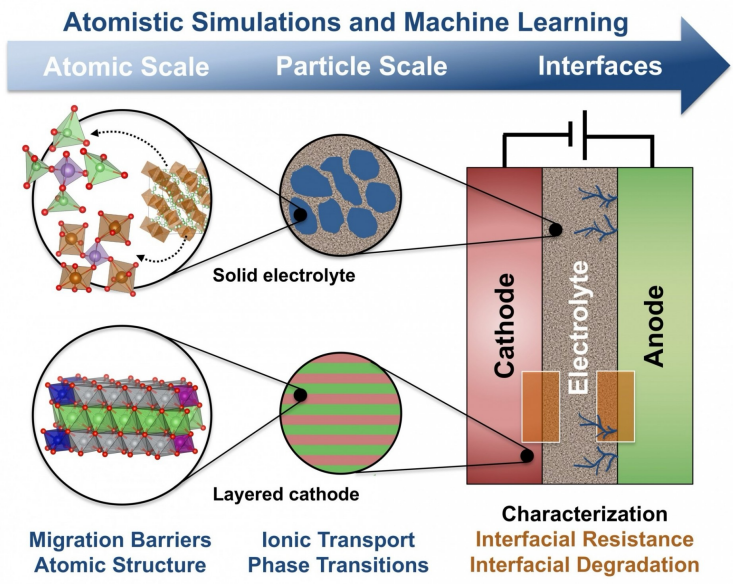
Figure source: https://www.thoughtco.com/what-is-distillation-601964 (2/26)

Summary

- The laws of quantum mechanics determine material properties
- Often, DFT is a good numerical approximation
- (Although sometimes, it is not yet good enough.)
- The computational scaling of DFT prevents routine materials discovery
- ML interatomic potentials can reduce the scaling to linear with atoms
- Recent enormous advances through model scaling (in industry)

61

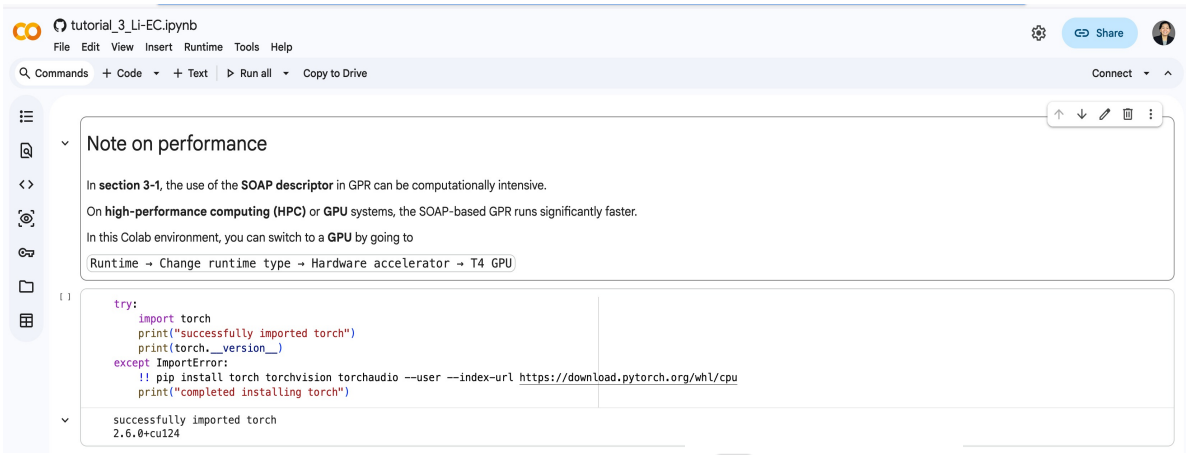
Simulating Batteries with MLIPs



62

Practical: SOAP descriptor

https://colab.research.google.com/github/atomisticnet/aenet-gpr/blob/main/tutorial/tutorial_3_Li-EC.ipynb



Homework: SOAP descriptor tutorials

GPR Tutorial: H_2

GPR Tutorial: EC–EC

GPR Tutorial: Li–EC

aenet-gpr

IW Yeu, A Urban, N Artrith et al., *npj Comput. Mater.* **11**, 156 (**2025**)

GPR-ANN and **GPR-aided NEB** Code: https://github.com/atomisticnet/aenet-gpr

63

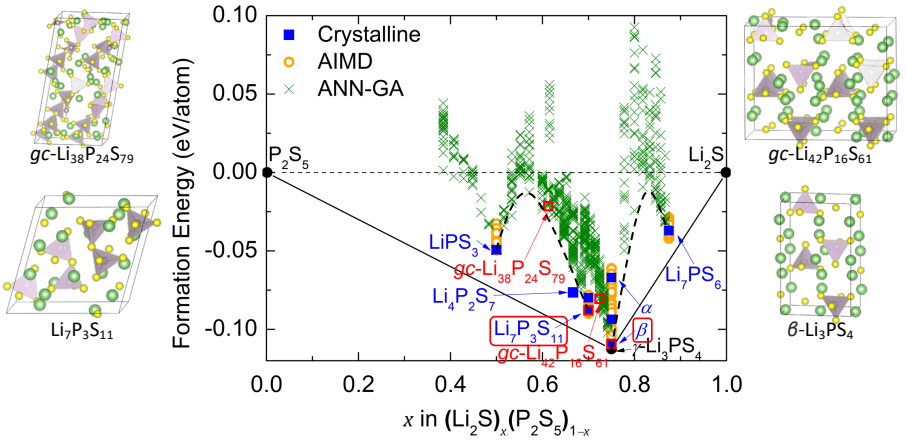
Local Geometry and/or Electronic Structure

K-means is an unsupervised learning algorithm

Nong Artrith (Email: n.artrith@uu.nl)

64

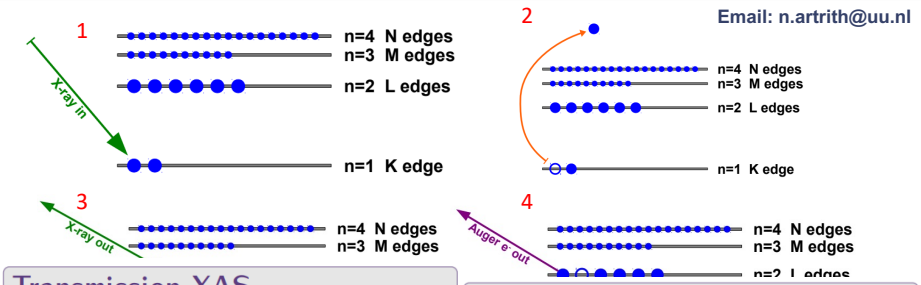
$(Li_2S)_x(P_2S_5)_{1-x}$: Phase Diagram



H. Guo, Q. Wang, A. Urban & **N. Artrith**, *Chem. Mater.* **34** (2022), 6702 - 6712.

65

X-ray Absorption Spectroscopy



Email: n.artrith@uu.nl

Transmission XAS

$\mu(E) = \ln(I_0/I_t)$, i.e. Beers' Law for X-rays

Fluorescence XAS

$\mu(E) \propto I_f/I_0$

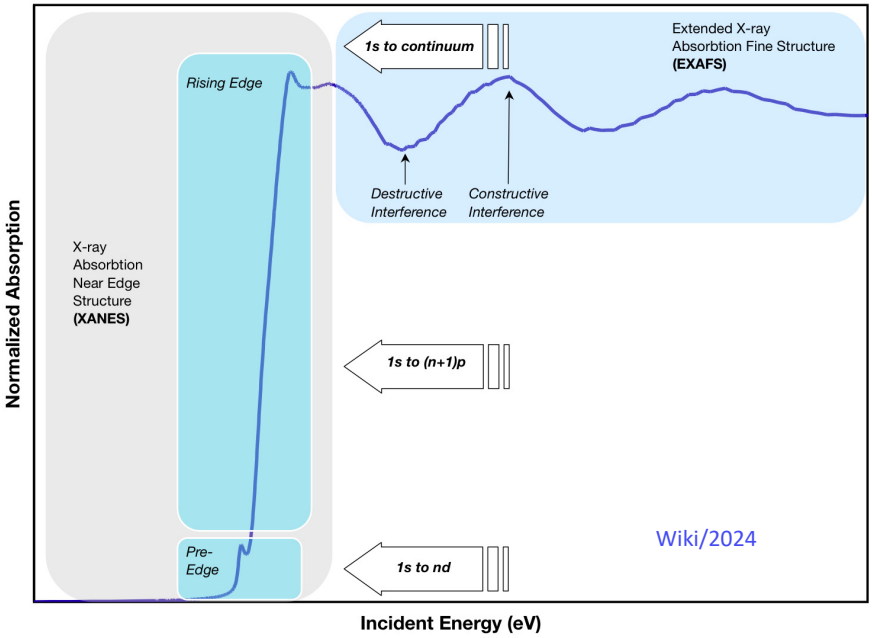
1. An incoming photon interacts with a deep-core electron. Here, a 1s electron is excited for a K-edge spectrum.
2. The deep-core electron is promoted to some unoccupied state above the Fermi energy, propagates away, and leaves behind a core-hole.
3. A short time later (1 or 2 femtoseconds), a higher-lying electron decays into the core-hole and emits a photon.
4. Alternately, the energy from the higher-lying electron can be used to emit an Auger electron.

- Relationship between formal valence of a metal and the position of the edge in XANES spectrum: The shift to higher energy is, to first order, a Coulomb effect. Less charge on the atom means less screening of the core.

[1] Bruce Ravel. Introduction to X-ray Absorption Spectroscopy (2015).

66

X-ray Absorption Spectroscopy: K-Edge



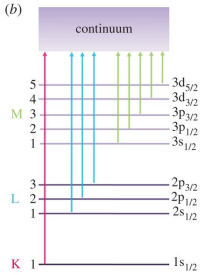
Wiki/2024

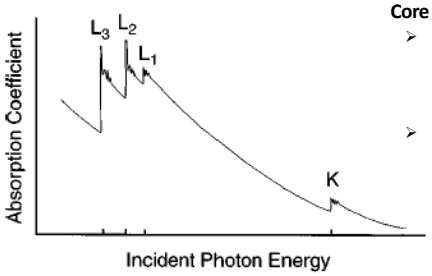
67

67

XAS Simulation

Email: n.artrith@uu.nl





Fermi's golden rule:

$$\sigma(\omega) = 4\pi^2 \frac{\omega}{c} \sum_F |\langle 0|d|F\rangle|^2 \delta(\omega + E_0 - E_F).$$

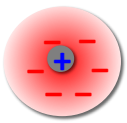
Dipole app. : $d = \hat{\varepsilon} \cdot \boldsymbol{r}$

Single particle app. :

$$M_{i \to f} = \langle \Psi_f | \hat{\epsilon} \cdot \mathbf{R} | \Psi_i \rangle \approx S \langle \psi_f | \hat{\epsilon} \cdot \mathbf{r} | \psi_i \rangle,$$

Core hole final state effect:

- Core hole potential approach: SC relaxation of valence electron (Xspectra, VASP 6.1.1)
- Perturbation approach: linear response of valence electrons (OCEAN, Exciting)



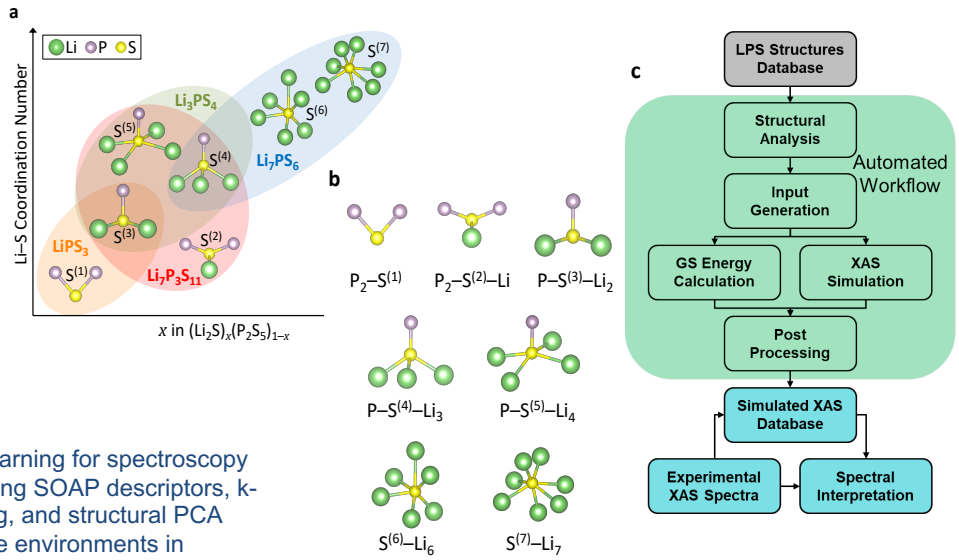
[1] Phil. Trans. R. Soc. A 371.1995 (2013).

68

68

Local Coordination: S atoms wt Li & P atoms in *gc*-LPS

Guo, Carbone, **N. Artrith**, Urban, Lu, *et al. Nature Scientific Data* **10**, 349 (2023)



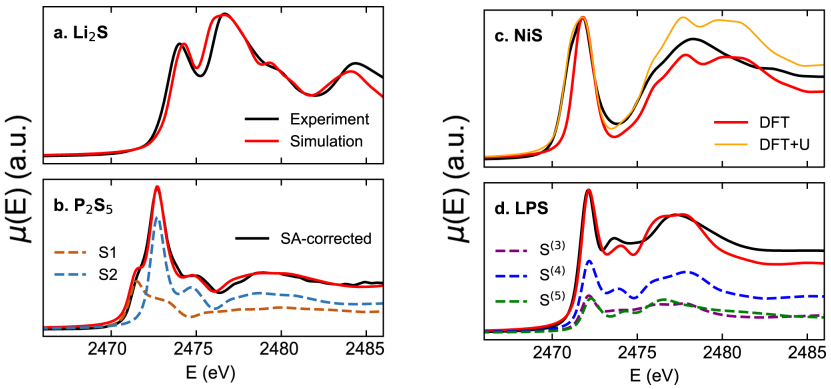
- Unsupervised learning for spectroscopy classification using SOAP descriptors, k-means clustering, and structural PCA analysis of S-site environments in structural databases.

https://github.com/atomisticnet/xas-tools

69

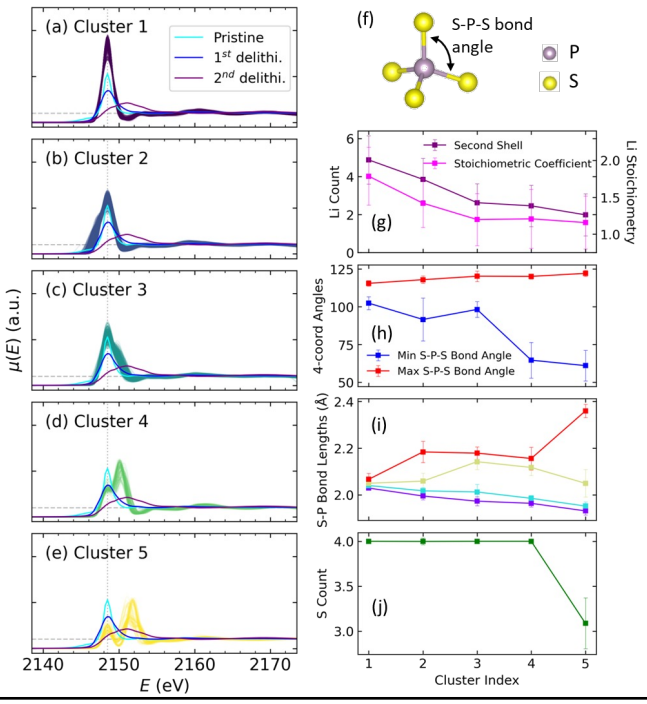
Validation: Benchmark of the XAS (DFT) simulations

- Simulated spectra reproduce the main features in reference systems



Guo, Carbone, **N. Artrith**, Urban, Lu, *et al. Nature Scientific Data* **10**, 349 (2023)

70



Clustering of *P* K-edge XANES

k-means: Unsupervised Learning Algorithm

➢ Structural principal-component analysis(PCA) using the SOAP descriptor for all the S sites in the structural database

Code:
https://archive.materialscloud.org/record/2023.130

➢ **a-e**: Clustered into **five groups** using K-Means

➢ **f-i**: Averaged structural descriptors in the **5 clusters**

➢ **h:** min/max **S-P-S** bond angles

➢ **i:** S-P bond length

➢ **j:** Number of neighboring S atoms.

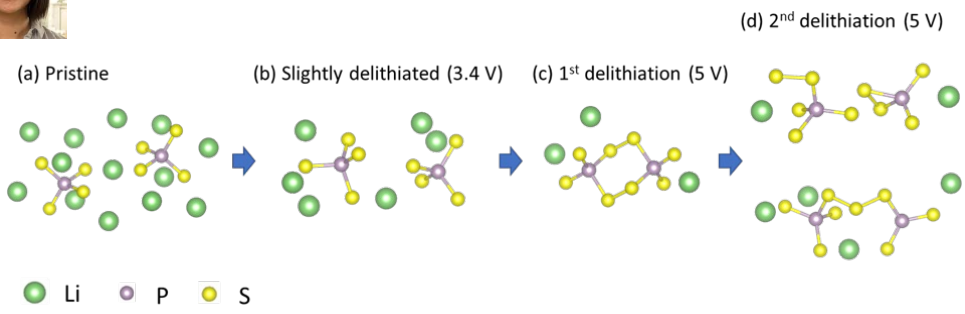
71

β-Li_3PS_4 Structures are Stable Conductors

Chuntian Cao

➢ **Atomic Insights into the Oxidative Degradation Mechanisms of Sulfide Solid Electrolytes**

https://github.com/AI-multimodal/Lightshow



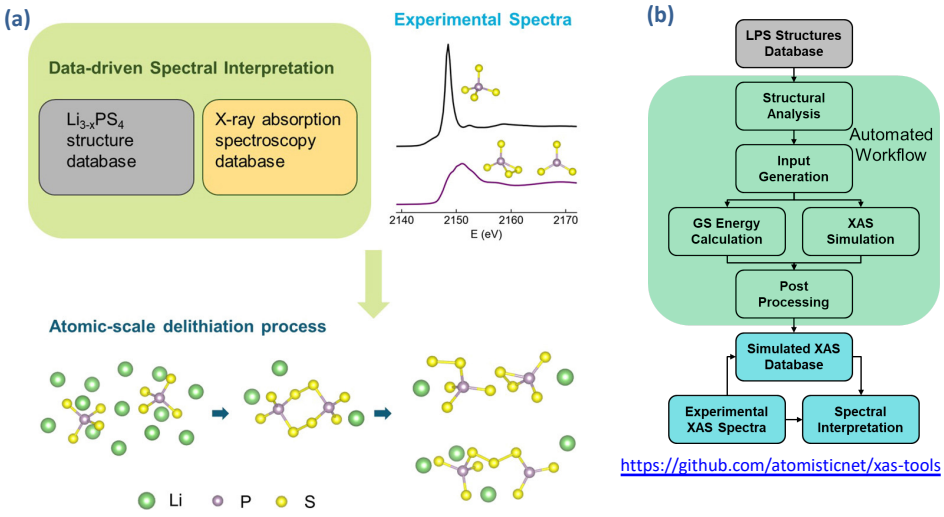
➢ **Unsupervised learning for spectroscopy classification using SOAP descriptors, k-means clustering, and structural PCA analysis of S-site environments in structural databases.**

Chuntian Cao *et al., Cell Rep. Phys. Sci,* **5**, 101909 **(2024)** https://doi.org/10.1016/j.xcrp.2024.101909

72

Summary

Guo, Carbone, N. Artrith, Urban, Lu, *et al. Nature Scientific Data* **10**, (**2023**) 349.
Cao, Carbone, N. Artrith, Urban, Lu, Wang, et al, *Cell Rep. Phys. Sci*. **5**, (**2024**) 101909.



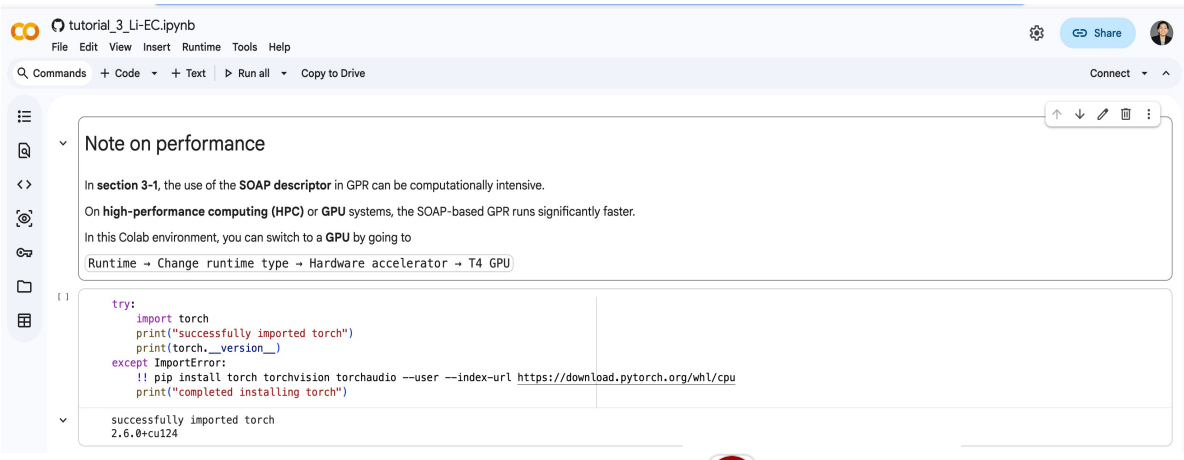
https://github.com/atomisticnet/xas-tools

➢ **ML-Assisted XAS Analysis Combines Structural & Spectral Databases**

73

Practical: SOAP descriptor

https://colab.research.google.com/github/atomisticnet/aenet-gpr/blob/main/tutorial/tutorial_3_Li-EC.ipynb



Homework: SOAP descriptor tutorials

GPR Tutorial: H_2
GPR Tutorial: EC–EC
GPR Tutorial: Li–EC

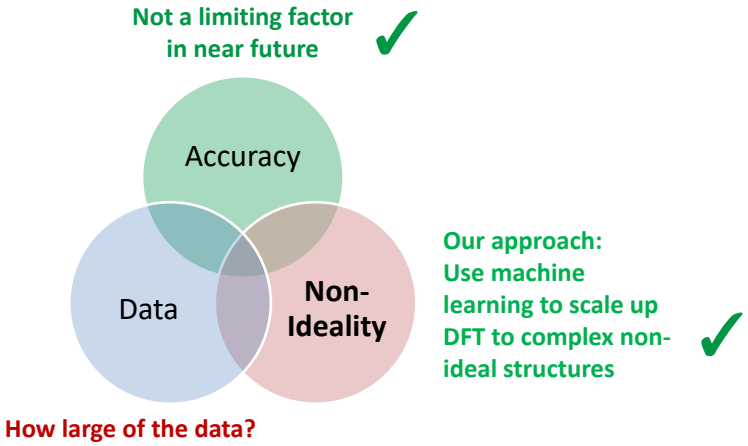
aenet-gpr

IW Yeu, A Urban, N Artrith et al., *npj Comput. Mater.* **11**, 156 (**2025**)

GPR-ANN and **GPR-aided NEB** Code: https://github.com/atomisticnet/aenet-gpr

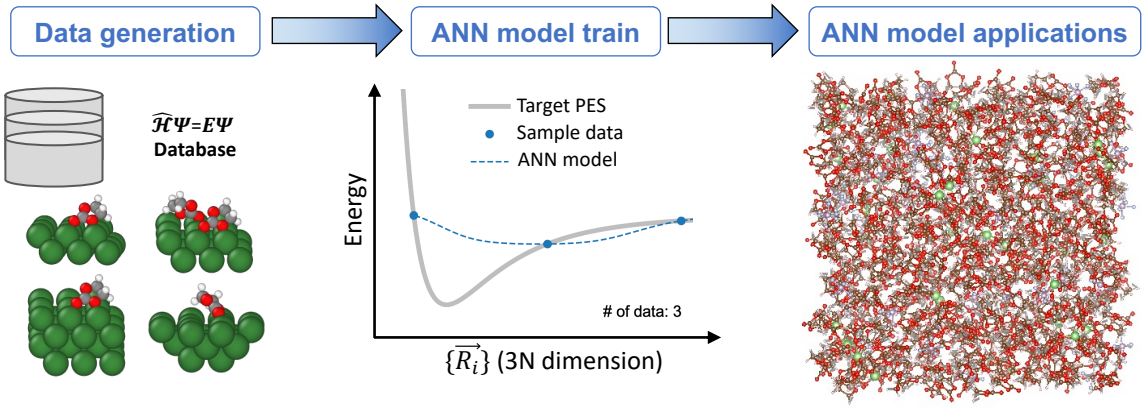
74

What's Holding Back Computational Materials Discovery?



75

Practical MLIP applications are hindered mainly by three obstacles



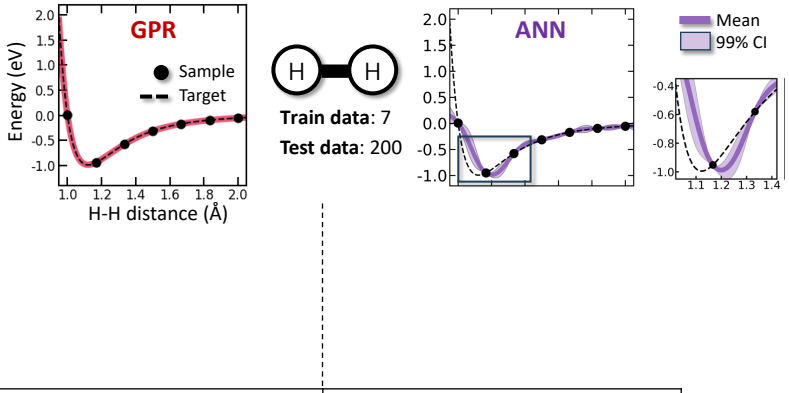
Problems

- **(Database)** High-fidelity data is lacking
- **(Sampling)** Data generation is arbitrary and inefficient
- **(Cost)** Training is not scalable

76

GPR vs ANN: Which is better MLIP?

- **GPR** execels for interpolation in low-data regimes
- **GPR** uncertainty can be used for active sampling
- But **GPR**'s scalability is limited



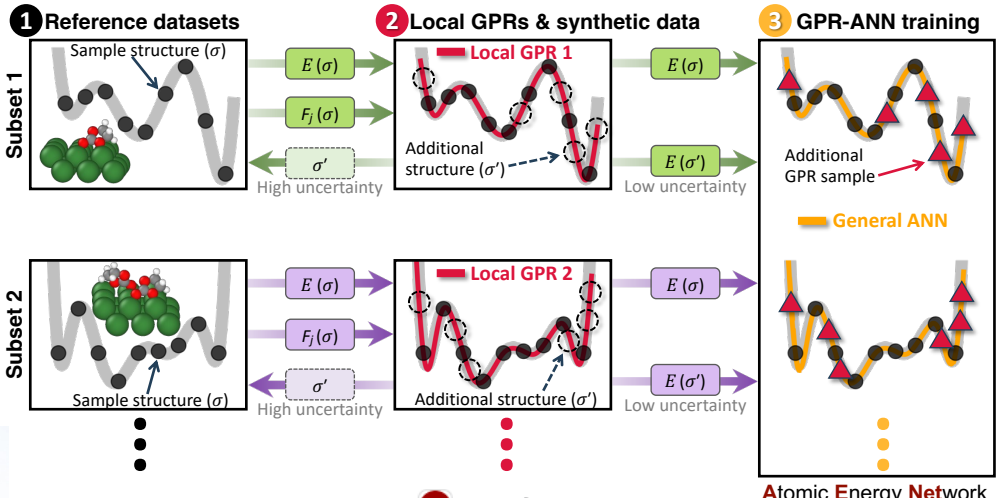
	GPR Non-parametric kernel-based	ANN Parametric model defined by {w}
Model uncertainty	O	X
Data requirements	Small	Large
Training cost	$\mathbb{O}(N^{2\sim3})$	$\mathbb{O}(N^{1\sim2})$
Transferability to large-scale	Not Great	Good

77

GPR-ANN: Combine the Best of Both Worlds

GPR for low-data interpolation, uncertainty-based data refinement

ANN obtained from efficient energy training, final model for large-scale evaluation



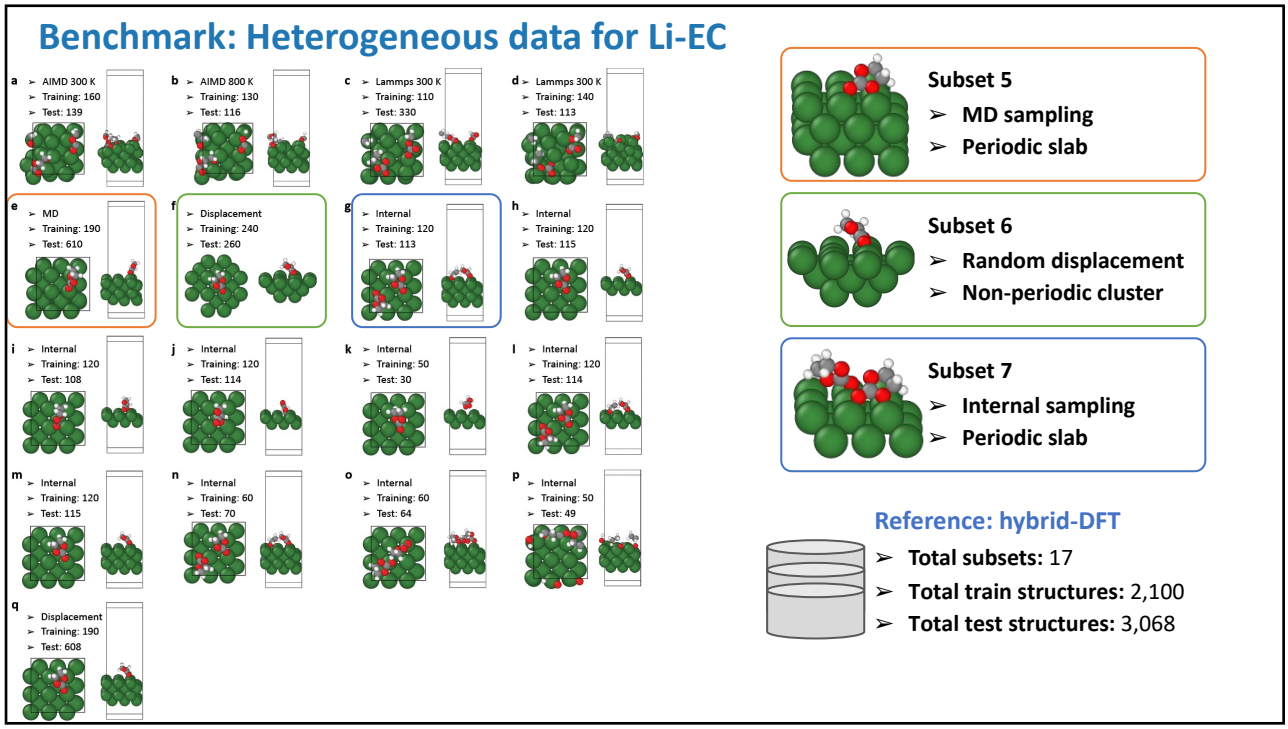
InWon Yue

aenet-gpr
https://github.com/atomisticnet/aenet-gpr

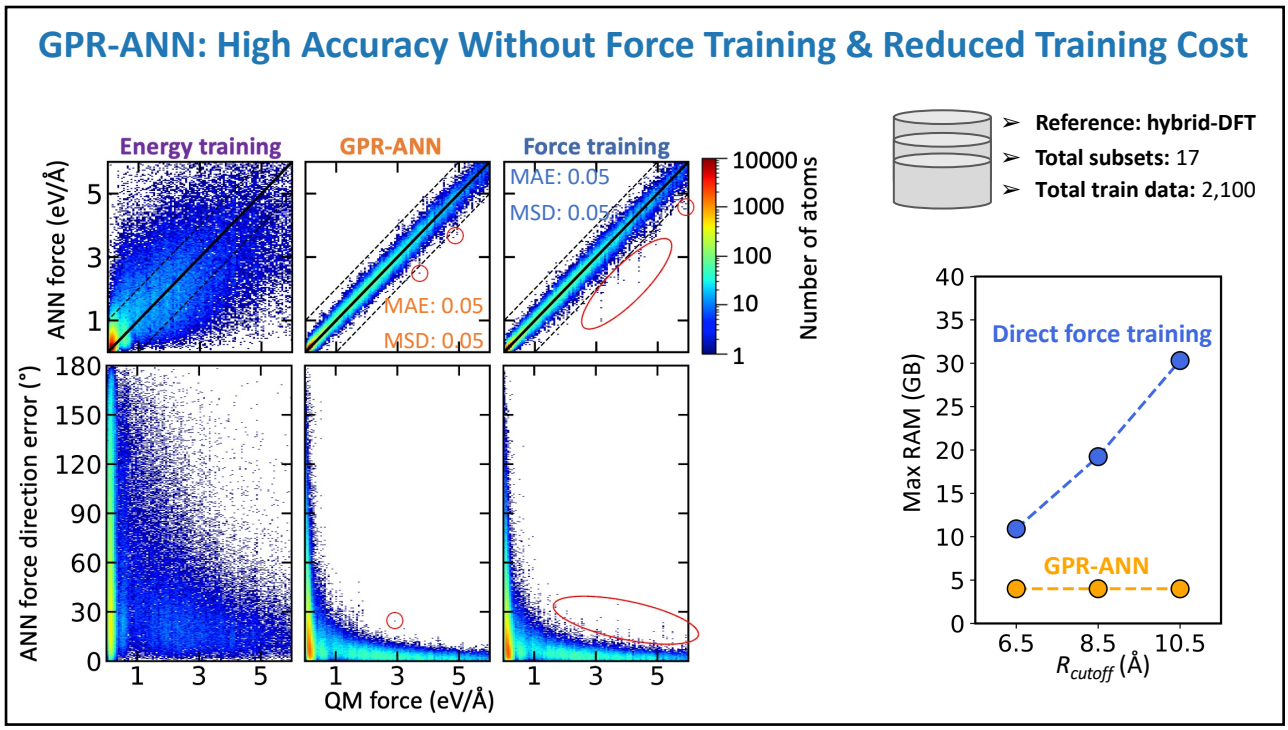
Yeu† et al., *npj Comput. Mater.* **11**, 156, **(2025)**

Atomic Energy Network
aenet
https://github.com/atomisticnet/aenet

78

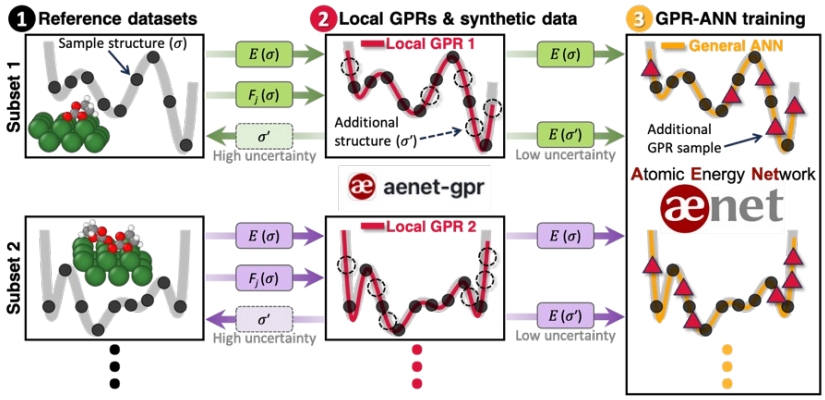

Benchmark: Heterogeneous data for Li-EC
a ➢ AIMD 300 K ➢ Training: 160 ➢ Test: 139
b ➢ AIMD 800 K ➢ Training: 130 ➢ Test: 116
c ➢ Lammps 300 K ➢ Training: 110 ➢ Test: 330
d ➢ Lammps 300 K ➢ Training: 140 ➢ Test: 113
e ➢ MD ➢ Training: 190 ➢ Test: 610
f ➢ Displacement ➢ Training: 240 ➢ Test: 260
g ➢ Internal ➢ Training: 120 ➢ Test: 113
h ➢ Internal ➢ Training: 120 ➢ Test: 115
i ➢ Internal ➢ Training: 120 ➢ Test: 108
j ➢ Internal ➢ Training: 120 ➢ Test: 114
k ➢ Internal ➢ Training: 50 ➢ Test: 30
l ➢ Internal ➢ Training: 120 ➢ Test: 114
m ➢ Internal ➢ Training: 120 ➢ Test: 115
n ➢ Internal ➢ Training: 60 ➢ Test: 70
o ➢ Internal ➢ Training: 60 ➢ Test: 64
p ➢ Internal ➢ Training: 50 ➢ Test: 49
q ➢ Displacement ➢ Training: 190 ➢ Test: 608
Subset 5
➢ MD sampling
➢ Periodic slab
Subset 6
➢ Random displacement
➢ Non-periodic cluster
Subset 7
➢ Internal sampling
➢ Periodic slab
Reference: hybrid-DFT
➢ Total subsets: 17
➢ Total train structures: 2,100
➢ Total test structures: 3,068

79


GPR-ANN: High Accuracy Without Force Training & Reduced Training Cost
Energy training
GPR-ANN
Force training
MAE: 0.05
MSD: 0.05
MAE: 0.05
MSD: 0.05
ANN force (eV/Å)
ANN force direction error (°)
QM force (eV/Å)
Number of atoms
10000
1000
100
10
1
➢ Reference: hybrid-DFT
➢ Total subsets: 17
➢ Total train data: 2,100
Direct force training
GPR-ANN
Max RAM (GB)
R_{cutoff} (Å)

80

Solved the common issues!



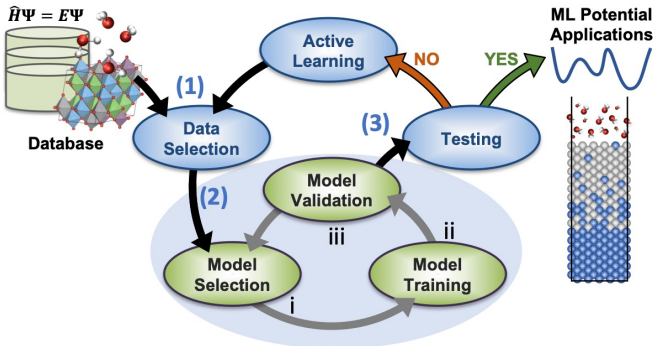
InWon Yue

Problems solved

- **(Database)** Reference data requirement is reduced
- **(Sampling)** Uncertainty-aware surrogate guides targeted sampling
- **(Cost)** Training is scalable

81

Can we systematically sample missing data that greatly improve MLIPs?



A.M. Miksch, T. Morawietz, J. Kästner, A. Urban, N. Artrith*,
Machine Learning: Science and Technology, **2**, 031001 **(2021),** DOI: https://doi.org/10.1088/2632-2153/abfd96.

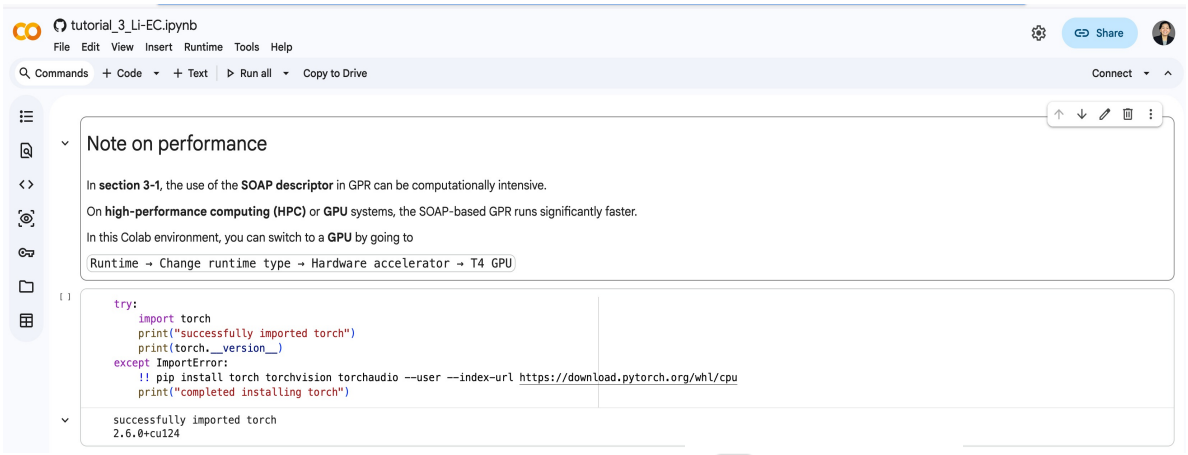
Code: https://github.com/atomisticnet/MLP-beginners-guide

- **(Model Selction)** kernel based GPR + neural networks
- **(Model Training)** direct force training → indirect force training (distill force into energy)
- **(Active learning)** Query of committee → GPR uncertainty

82

Practical: SOAP descriptor

https://colab.research.google.com/github/atomisticnet/aenet-gpr/blob/main/tutorial/tutorial_3_Li-EC.ipynb



Homework: SOAP descriptor tutorials

GPR Tutorial: H_2

GPR Tutorial: EC–EC

GPR Tutorial: Li–EC

aenet-gpr

IW Yeu, A Urban, N Artrith et al., *npj Comput. Mater.* **11**, 156 (**2025**)

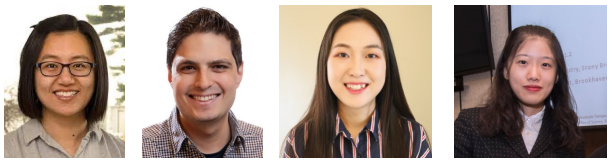
GPR-ANN and **GPR-aided NEB** Code: https://github.com/atomisticnet/aenet-gpr

83

Acknowledgements: Collaborators and Facilities

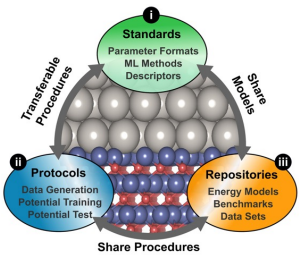
Thank You for Your Attention

Chuntian Cao **Matt Carbone** **Qian Wang** **Haoyue Guo**



DOE VTO: DE-SC0012704 (2020-2024)

aenet



The Materials Project
materialsproject.org

pymatgen

84

Acknowledgements: Collaborators and Facilities

OPEN ACCESS

Thank You for Your Attention

BAYER

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Advanced Research Center
Chemical Building Blocks Consortium

Dutch National Supercomputer Snellius

aenet

SURF
Sharing science, shaping tomorrow

Utrecht University

Infrastructure on High Performance Computing
Pan-European Research
HPC Europa 3

Dutch Sector Plan Project

85