Neural Representations Of Atoms

Making atomistic models work for us

Kelvin Lee

AI Research Scientist

Intel Labs—
AI4Science

About Me

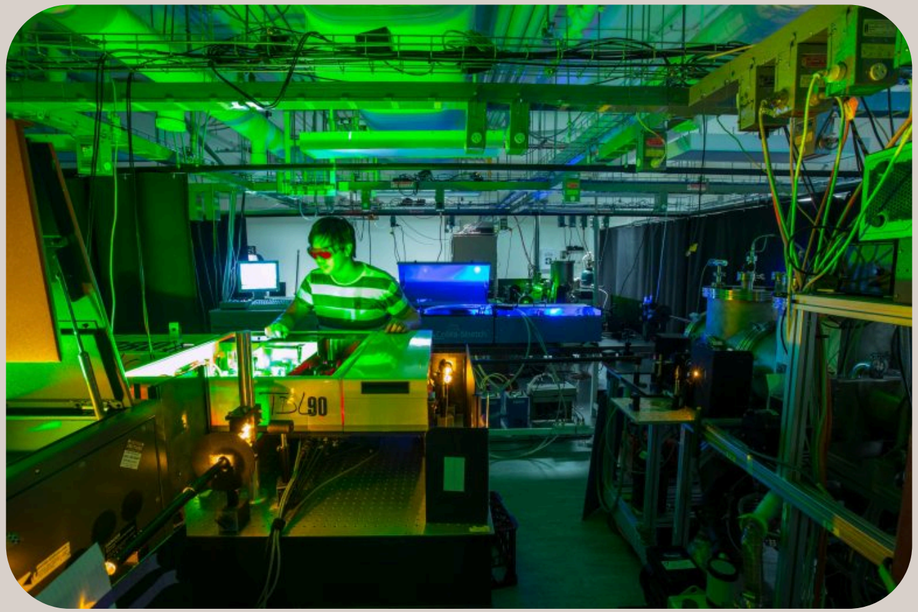
Born in Hong Kong, grew up in New Zealand & Australia
Parents are **not** scientists, despite name

About Me

Born in Hong Kong, grew up in New Zealand & Australia
Parents are **not** scientists, despite name
Currently lives in Portland, OR

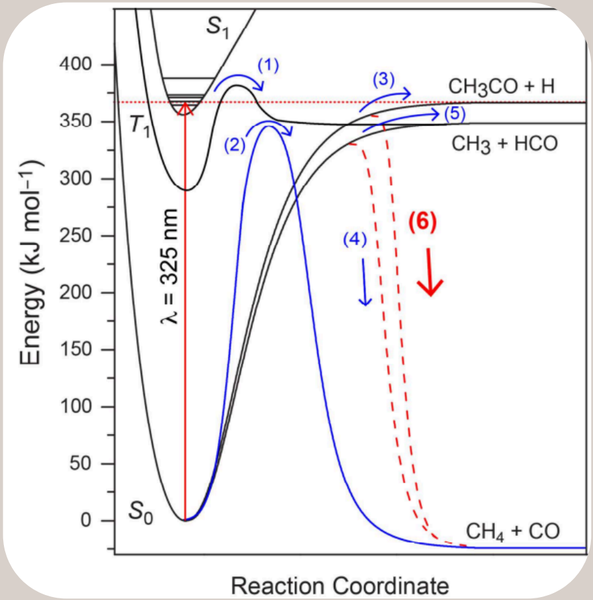
About Me

Born in Hong Kong, grew up in New Zealand & Australia
Parents are **not** scientists, despite name
Currently lives in Portland, OR
Enjoys nature in the Pacific Northwest
Enjoys ☕ & food



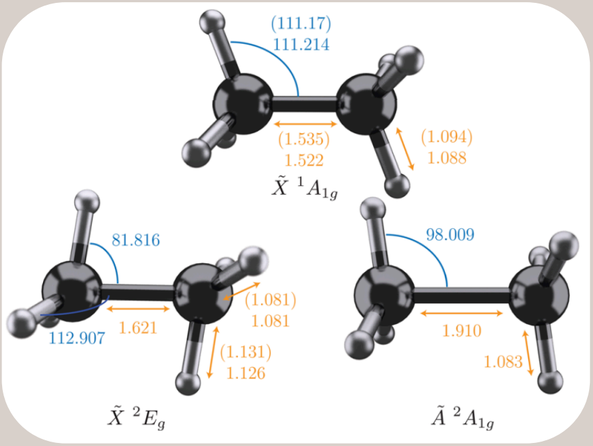
*Professional photoshoot—do not attempt at home

PhD in physical chemistry
University of New South Wales, Sydney



State-to-state reaction dynamics

Ion imaging



High accuracy ab initio thermochemistry

Lee, et al.; Two roaming pathways in the photolysis of CH_3CHO, Chemical Science (2014)

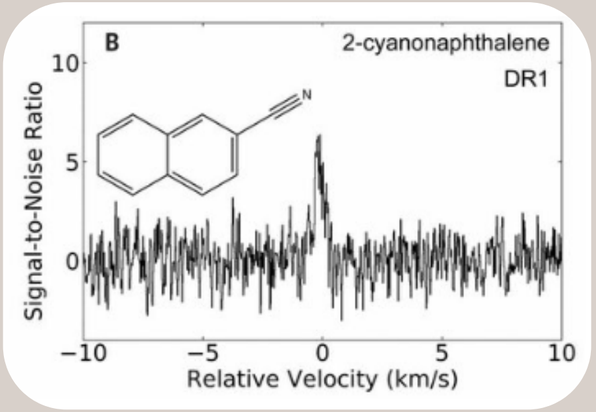
Lee, Rabidoux, Stanton; Cation Ion States of Ethane, JPCA (2016)

Submillimeter array @ Mauna Kea, HI

Postdoctoral research at:

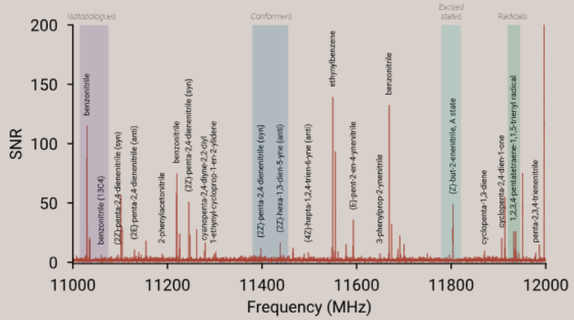
Center for Astrophysics | Harvard & Smithsonian (2017–2020)

MIT Chemistry (2020–2021)



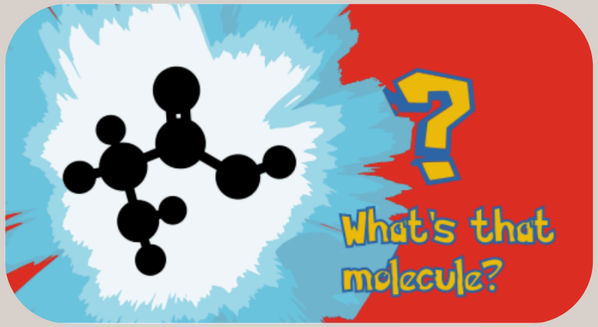
Growth and destruction of aromatic molecules in space

McGuire et al.; Detection of two interstellar polycyclic aromatic hydrocarbons via spectral matched filtering, Science (2021)



Automated spectroscopic mixture separation

Lee, McCarthy; A Study of Benzene Fragmentation, Isomerization, and Growth, JPCL (2019)

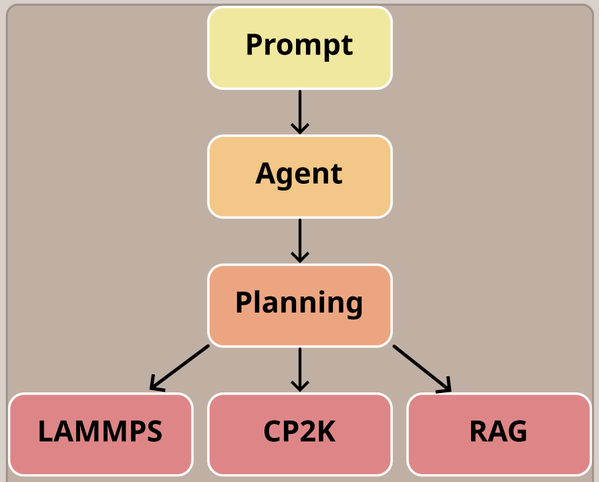


Structure determination with deep learning

McCarthy, **Lee**; Molecule Identification with Rotational Spectroscopy and Deep Learning, JPCA (2020)

Present day—AI4Science @ Intel Labs

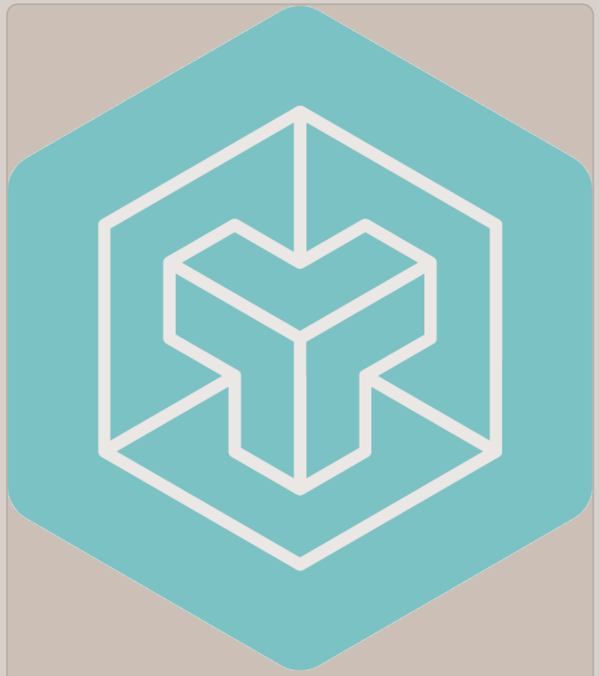
High-performance computing



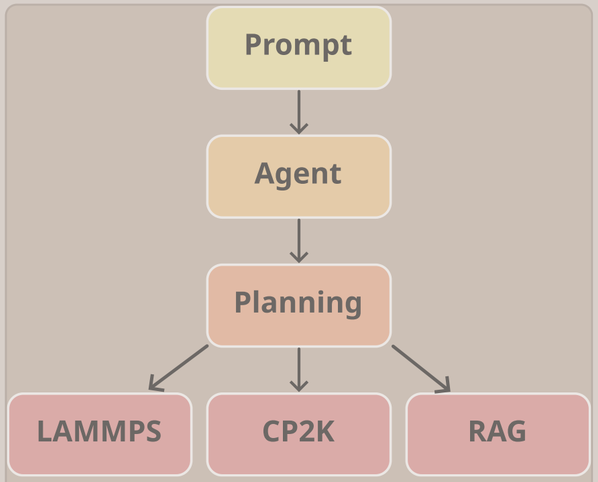
Agentic workflows for science

Deep learning with molecules & materials

Present day—AI4Science @ Intel Labs



High-performance computing



Agentic workflows for science

Deep learning with molecules & materials

Neural representations of atomistic systems

What, why, and how?

Problem statement

Problem statement

We want a molecule/material with property X for application Y

Problem statement

We want a molecule/material with property X for application Y

Could be an optimization problem if we had a model function —how do we obtain a good model?

Problem statement

We want a molecule/material with property X for application Y

Could be an optimization problem if we had a model function —how do we obtain a good model?

A map is not the territory it represents,
but, if correct,
it has a similar structure to the territory, which accounts for its usefulness.

—

Alfred Korzybski

Problem statement

We want a molecule/material with property X for application Y

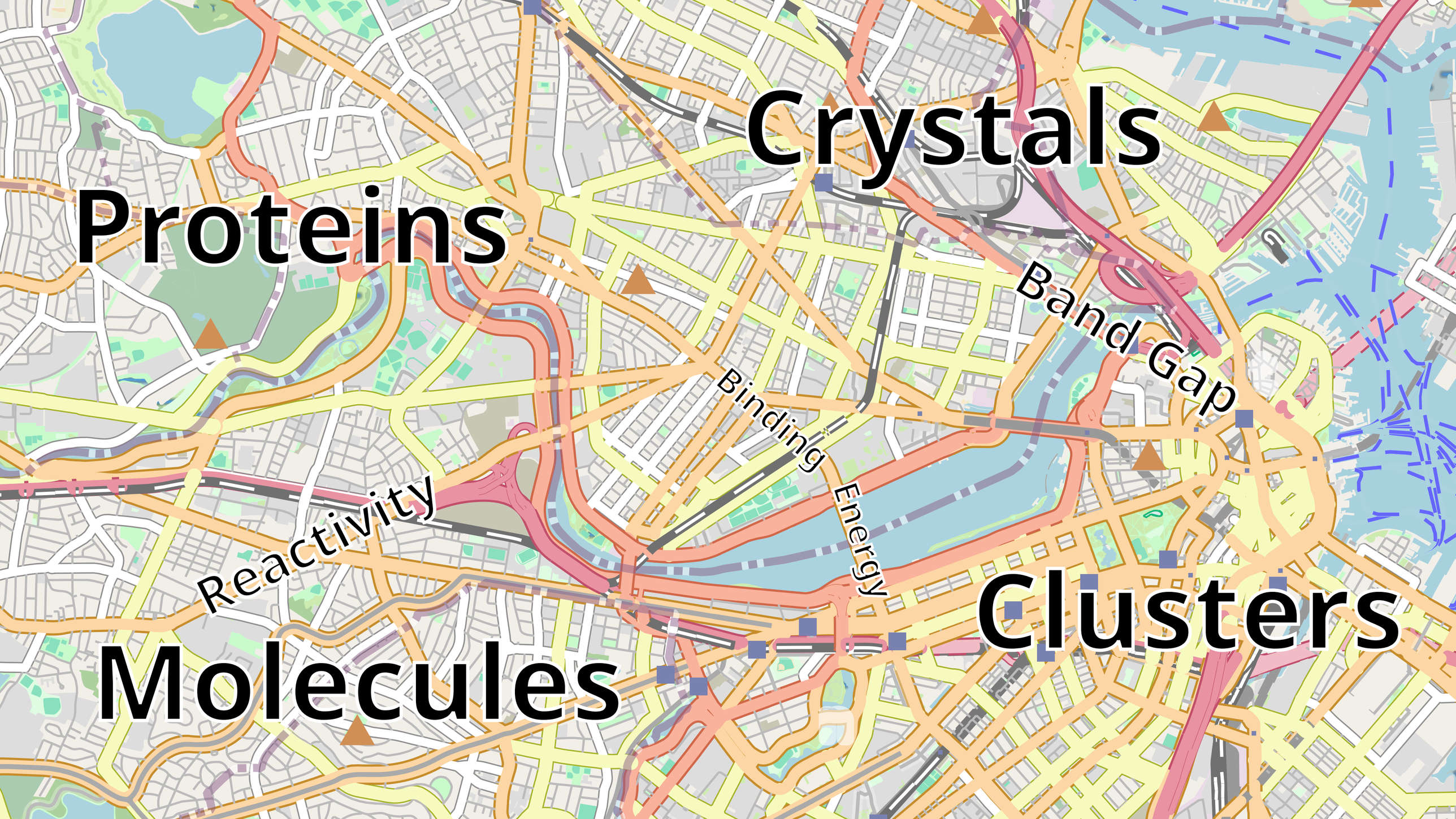
Could be an optimization problem if we had a model function
—how do we obtain a good model?

A map is not the territory it represents,
but, if correct,
it has a similar structure to the territory, which accounts for its usefulness.

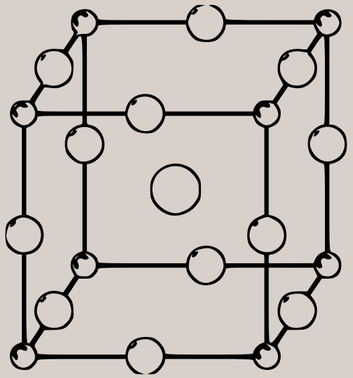
—

Alfred Korzybski

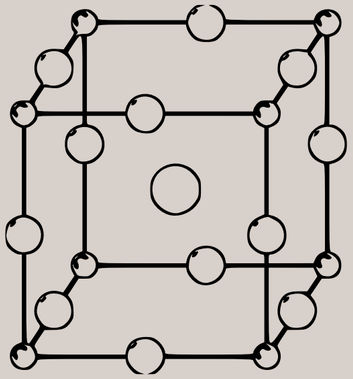
We need a useful map of chemical space
—or a sufficiently holistic representation


Crystals
Proteins
Band Gap
Binding Energy
Reactivity
Clusters
Molecules

How do we build a useful representation?



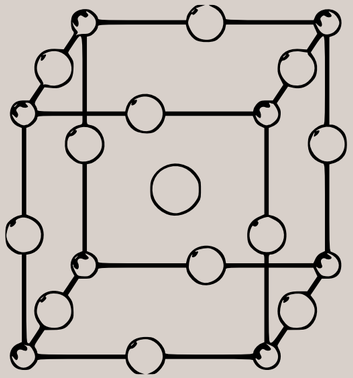
Structure and composition correlate with properties



Structure and composition correlate with properties

Band gap
Formation energy
Energy/force
Stress
Polarization
Multipole moments
Atomic charges

Sample chemical space to curate datasets



Structure and composition correlate with properties

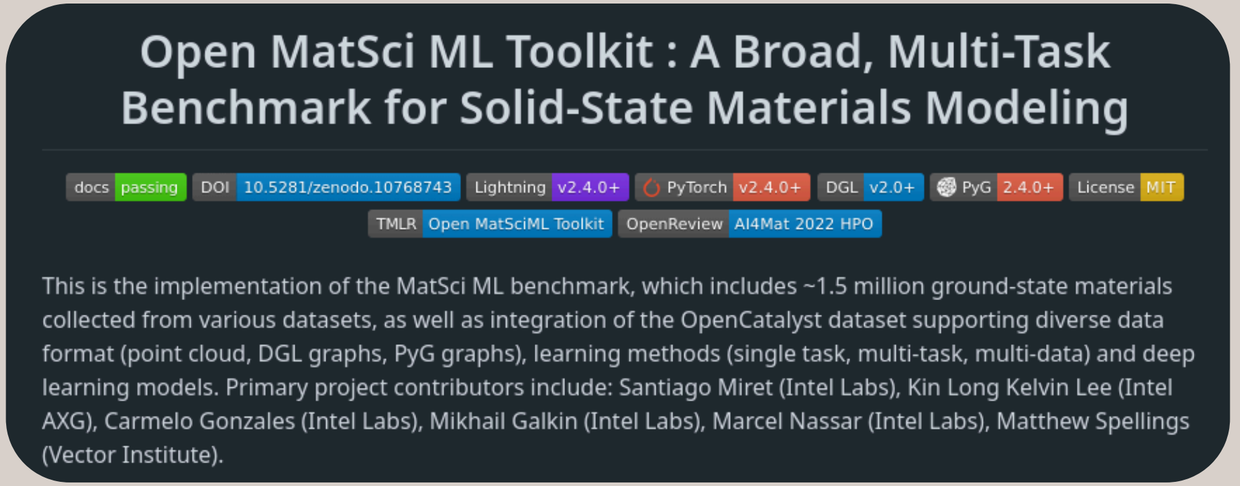
Band gap
Formation energy
Energy/force
Stress
Polarization
Multipole moments
Atomic charges

Sample chemical space to curate datasets

Trained neural networks learn to represent chemical space

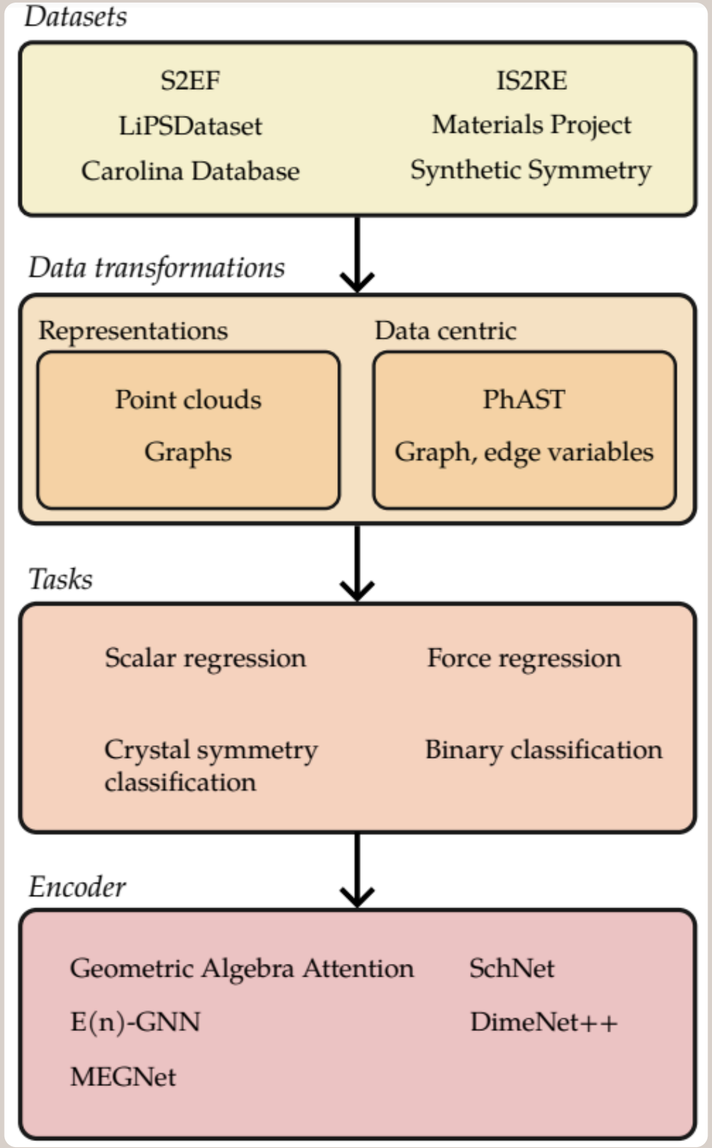
Learning to map from data

IntelLabs/matsciml



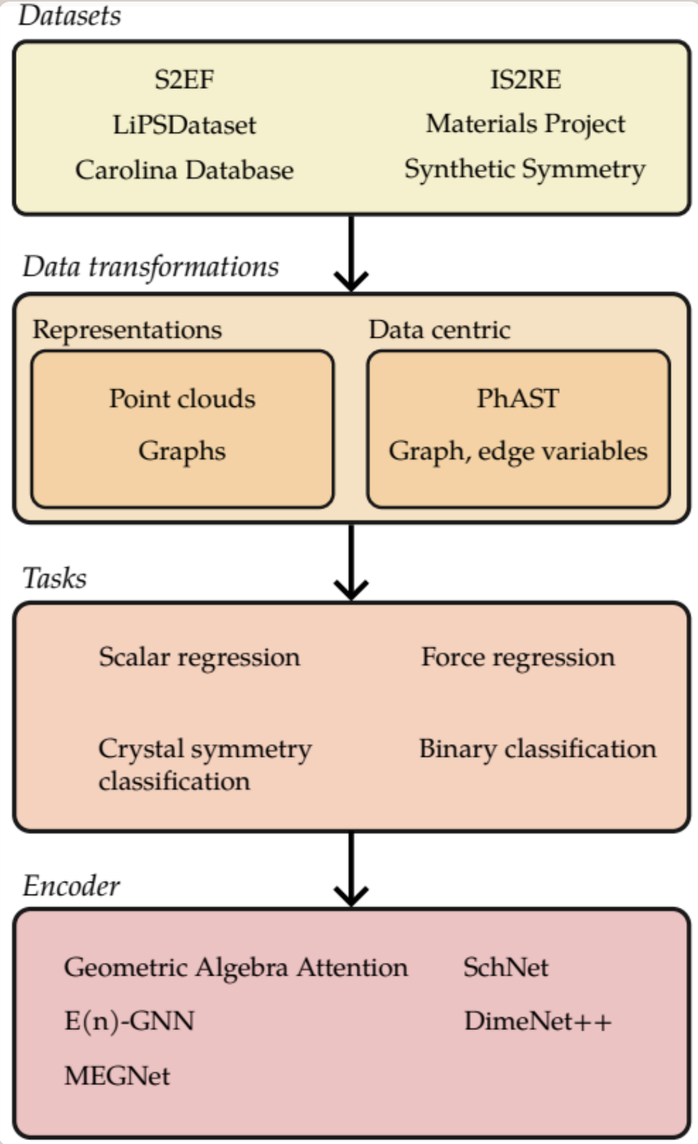
Framework for glueing data with models

Miret, **Lee**, Gonzales, Nassar, Spellings;
The Open MatSciML Toolkit, TMLR (2023)

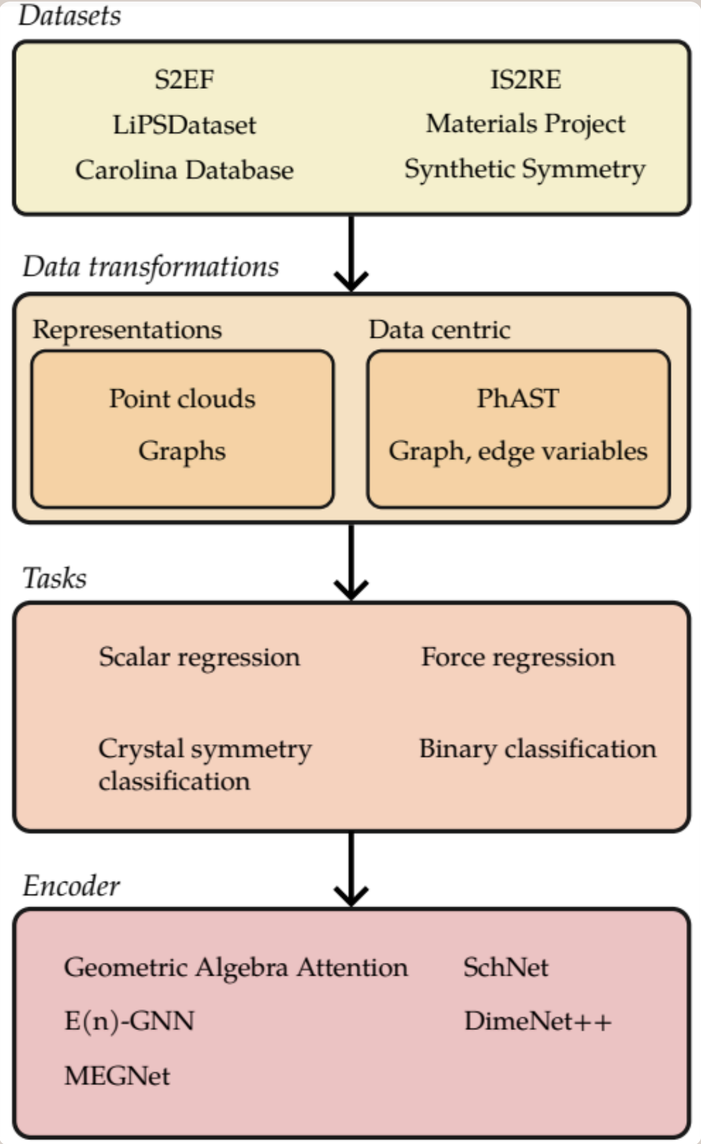


Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)

Modular pipeline for foundation model training



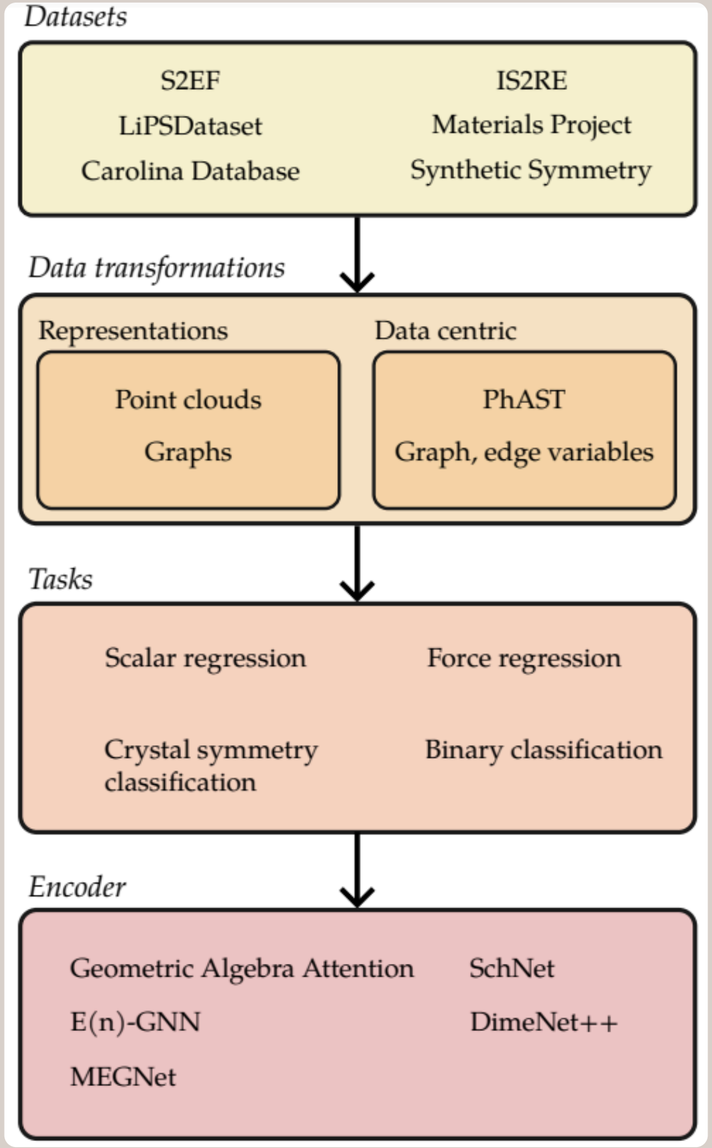
Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)



Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)

Modular pipeline for foundation model training

Interfaces to major datasets: Open Catalyst, Materials Project, Alexandria, NOMAD

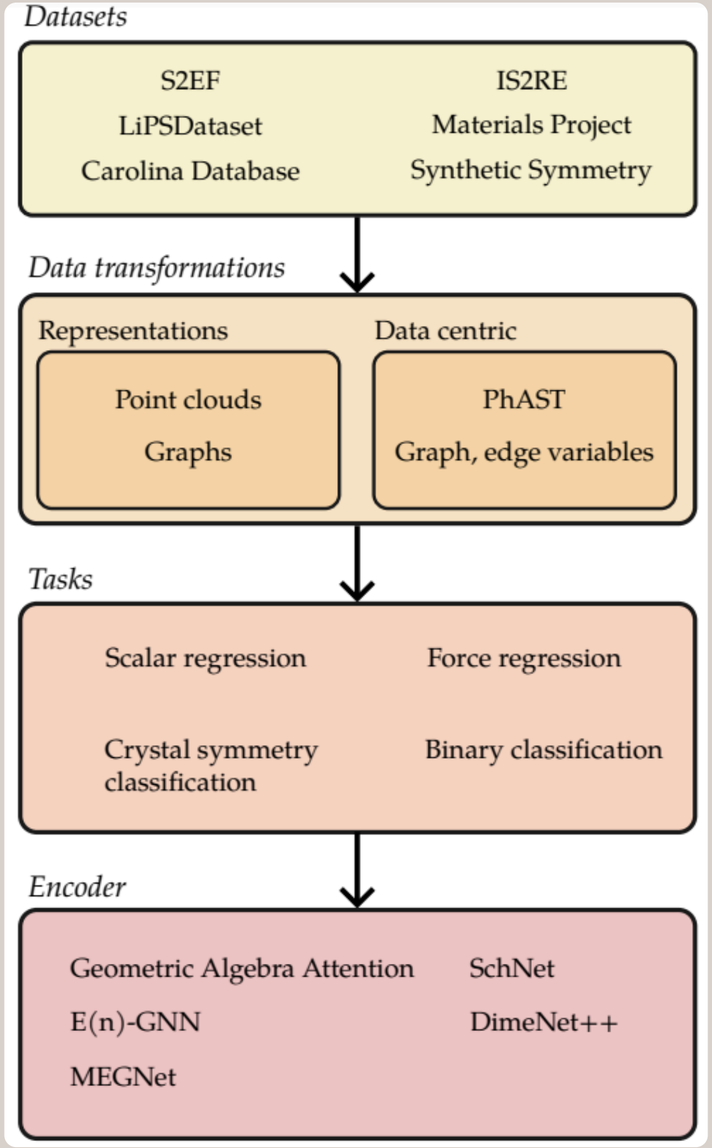


Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)

Modular pipeline for foundation model training

Interfaces to major datasets: Open Catalyst, Materials Project, Alexandria, NOMAD

Reference model interfaces and implementations



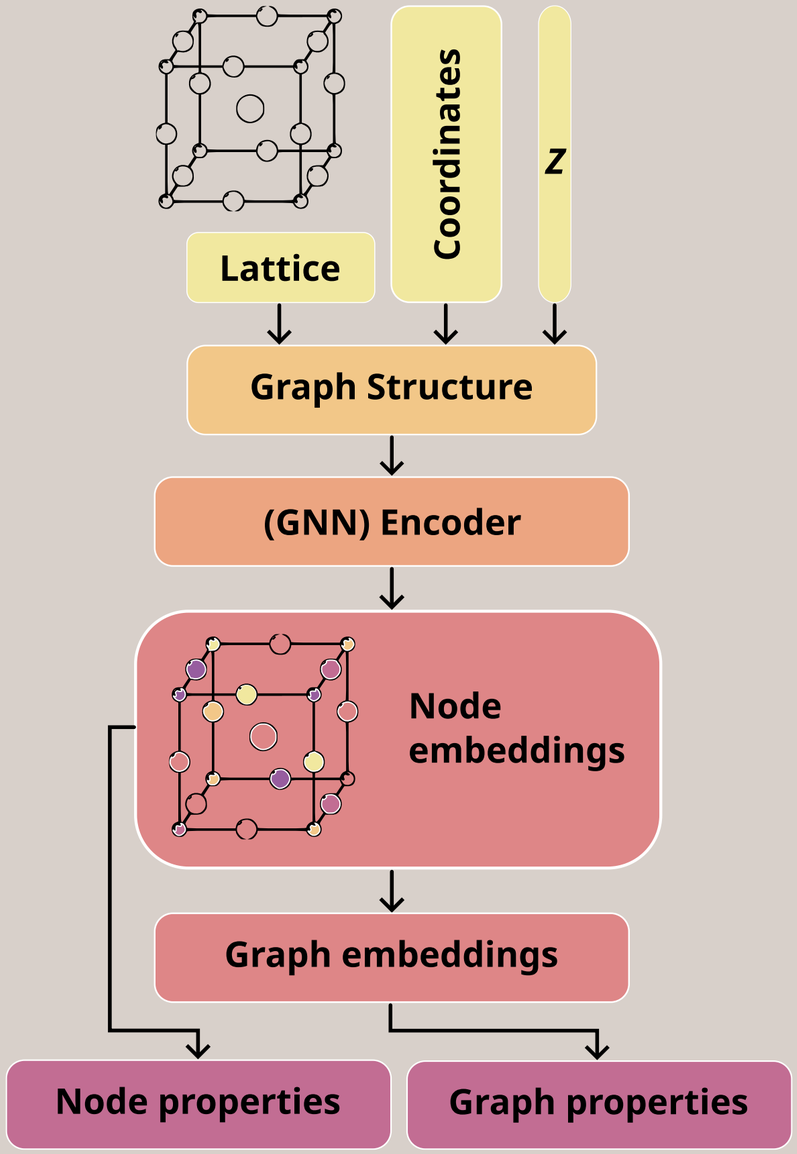
Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)

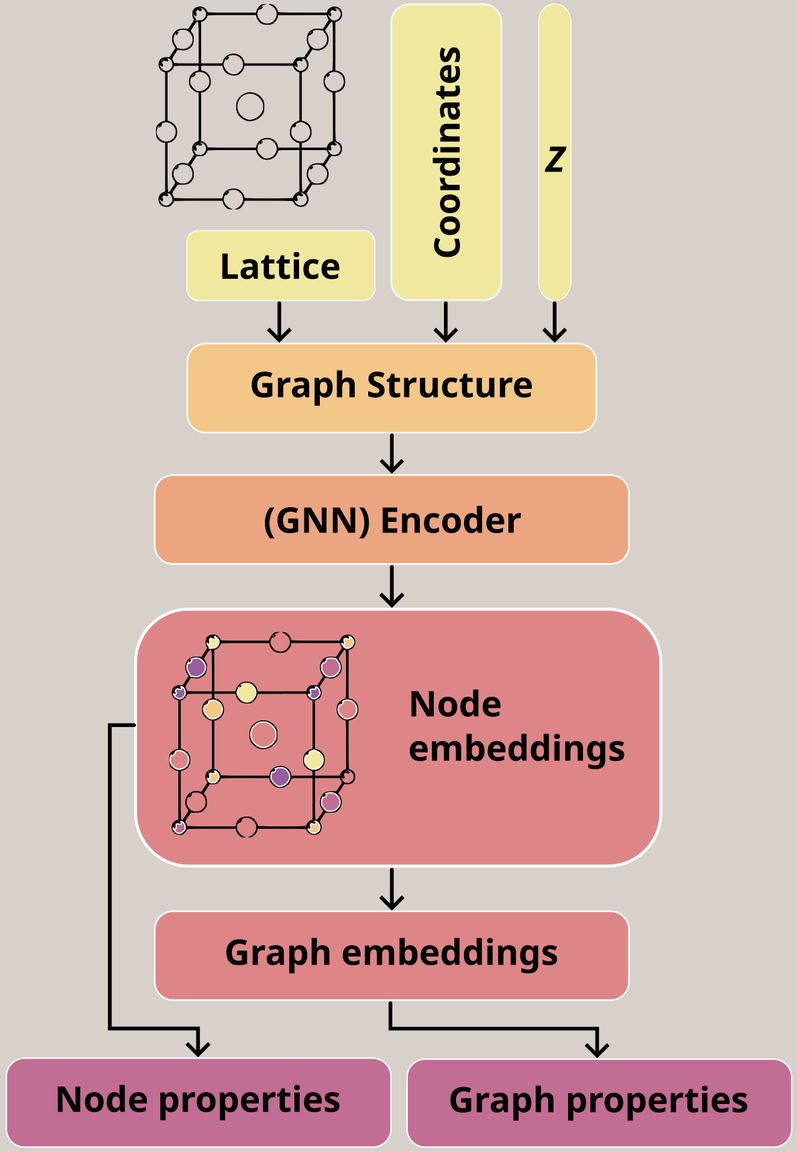
Modular pipeline for foundation model training

Interfaces to major datasets: Open Catalyst, Materials Project, Alexandria, NOMAD

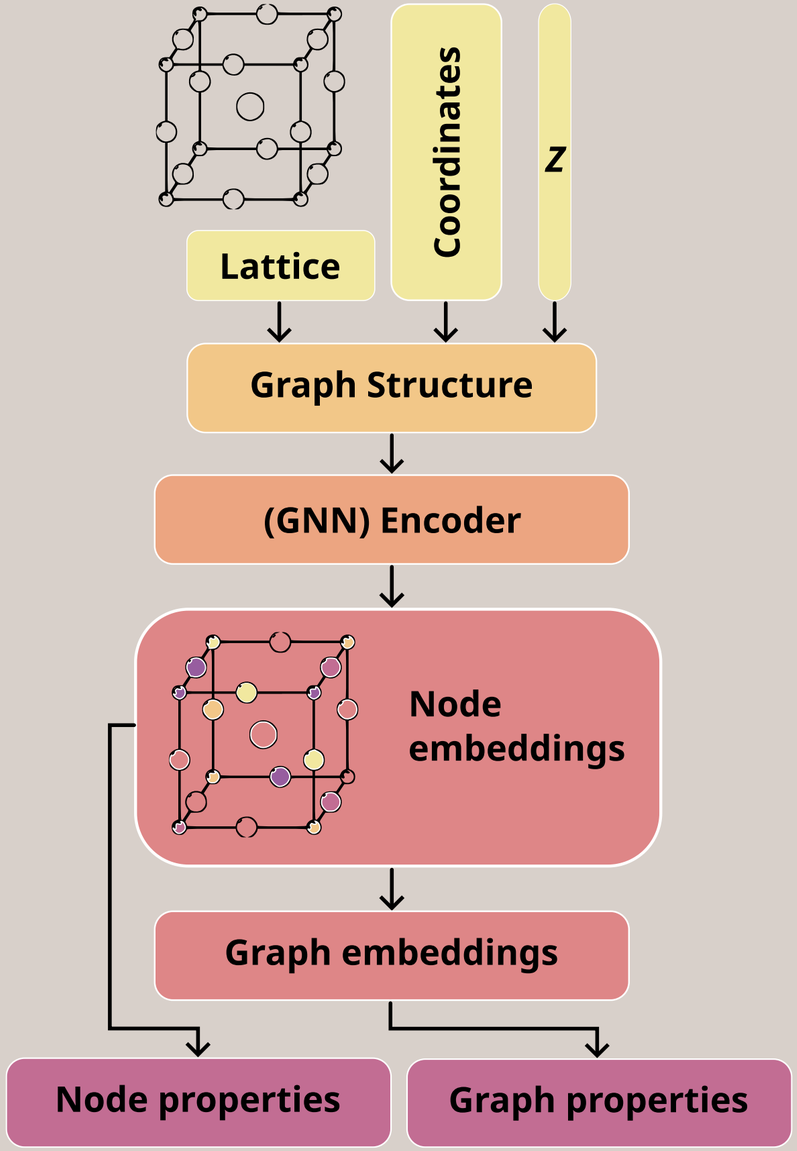
Reference model interfaces and implementations

Free composition of multiple tasks and datasets


Lattice
Coordinates
Z
Graph Structure
(GNN) Encoder
Node embeddings
Graph embeddings
Node properties
Graph properties

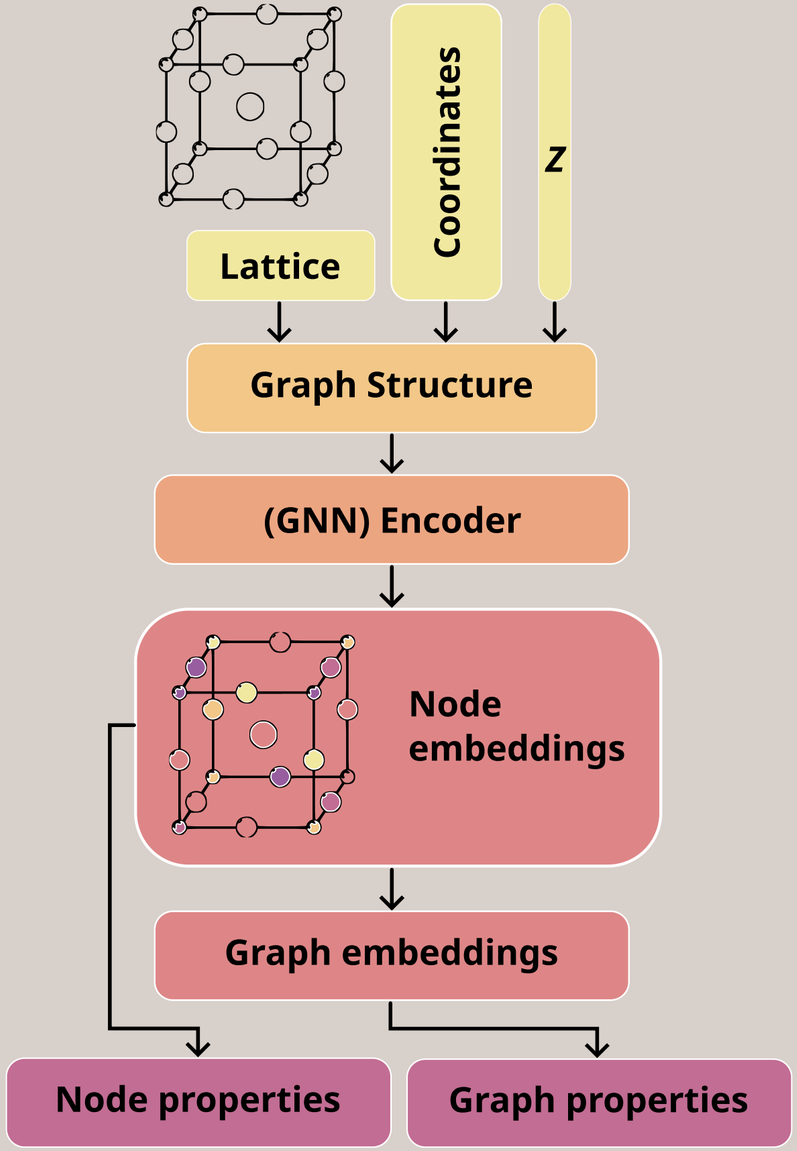


Graph wiring according to periodic boundary conditions



Graph wiring according to periodic boundary conditions

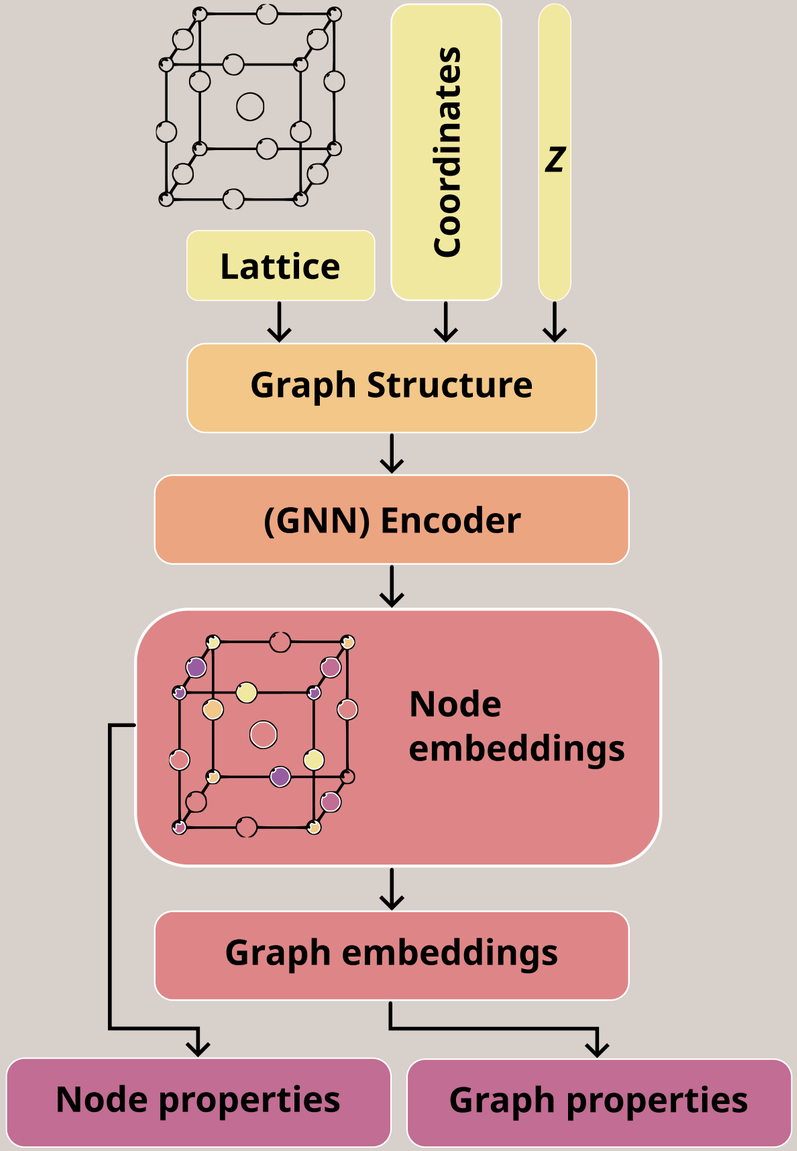
Abstract encoder processes node features



Graph wiring according to periodic boundary conditions

Abstract encoder processes node features

Output heads share embedding space to predict task labels



Graph wiring according to periodic boundary conditions

Abstract encoder processes node features

Output heads share embedding space to predict task labels

e.g. Energy/Forces as a task produces an MLIP

How do we know we have a useful representation?

Aligning data

Aligning data

S2EF and **IS2RE** are energy/force data from Open Catalyst

Aligning data

S2EF and **IS2RE** are energy/force data from Open Catalyst

S2EF is conventional MLIP; IS2RE is initial structure-to-relaxed energy (task)

Aligning data

S2EF and **IS2RE** are energy/force data from Open Catalyst

S2EF is conventional MLIP; IS2RE is initial structure-to-relaxed energy (task)

MP20 is Materials Project, with formation energy labels (data/task)

Aligning data

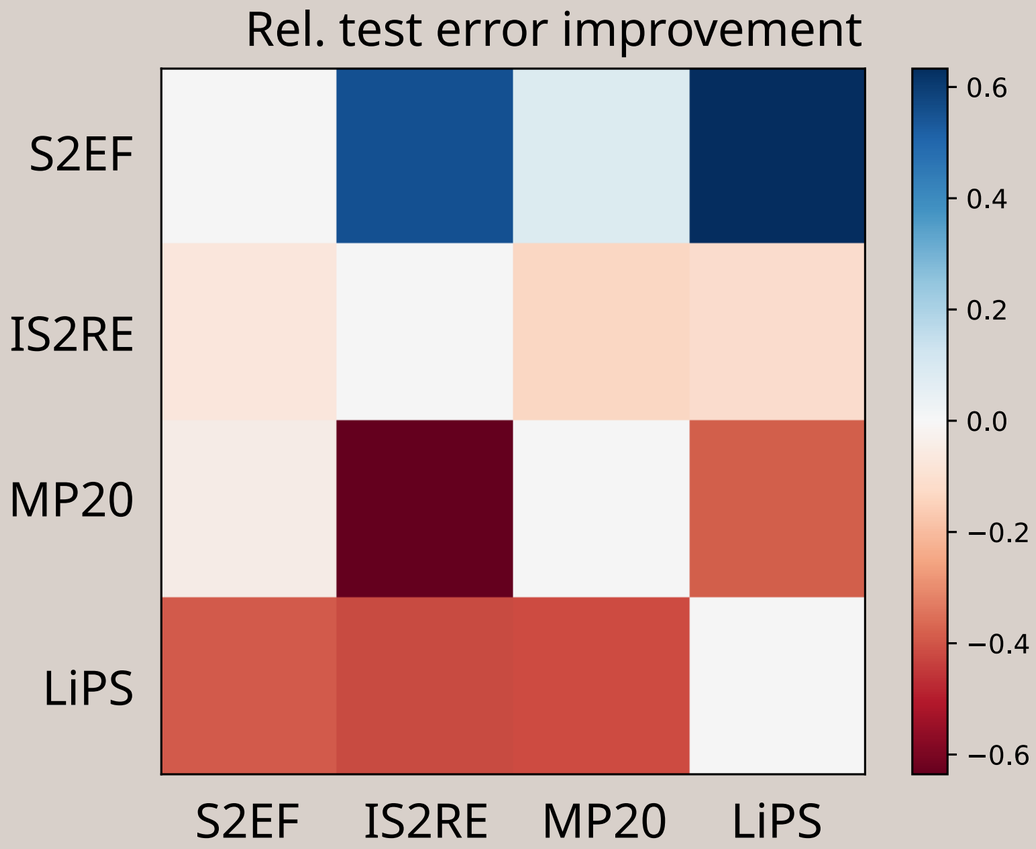
S2EF and **IS2RE** are energy/force data from Open Catalyst

S2EF is conventional MLIP; IS2RE is initial structure-to-relaxed energy (task)

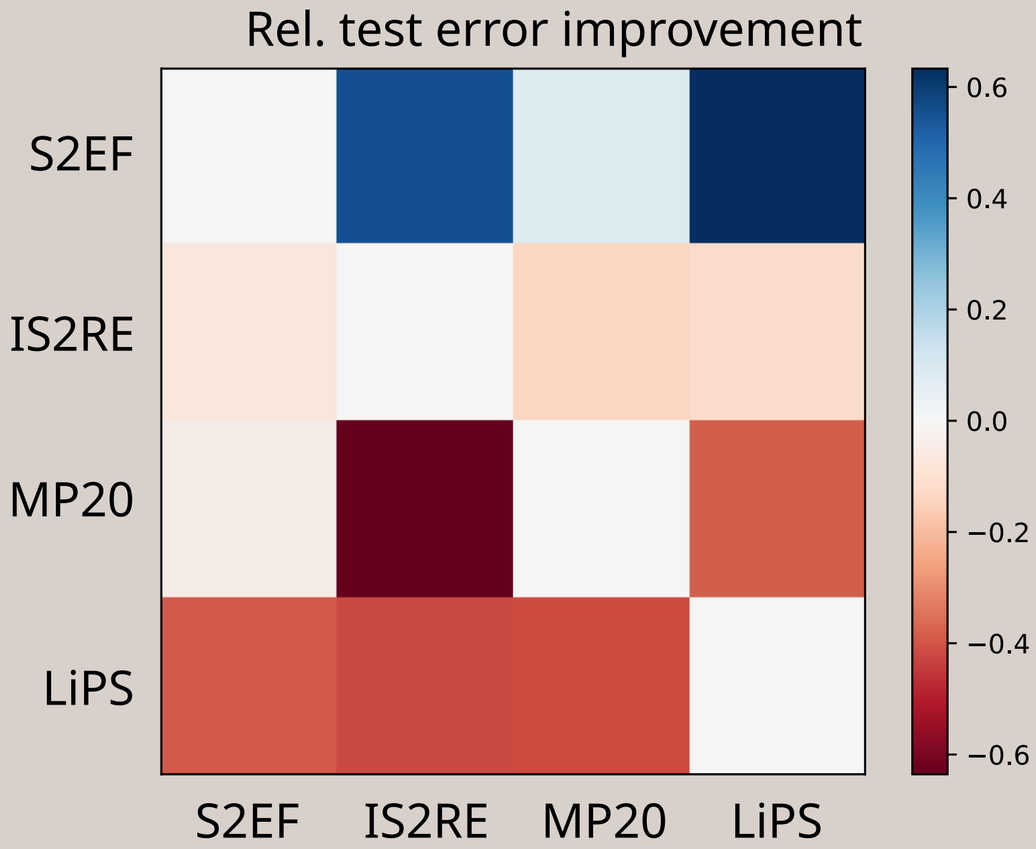
MP20 is Materials Project, with formation energy labels (data/task)

LiPS is a single composition energy/force trajectory (data)

Base encoder is
$E(n)$-GNN



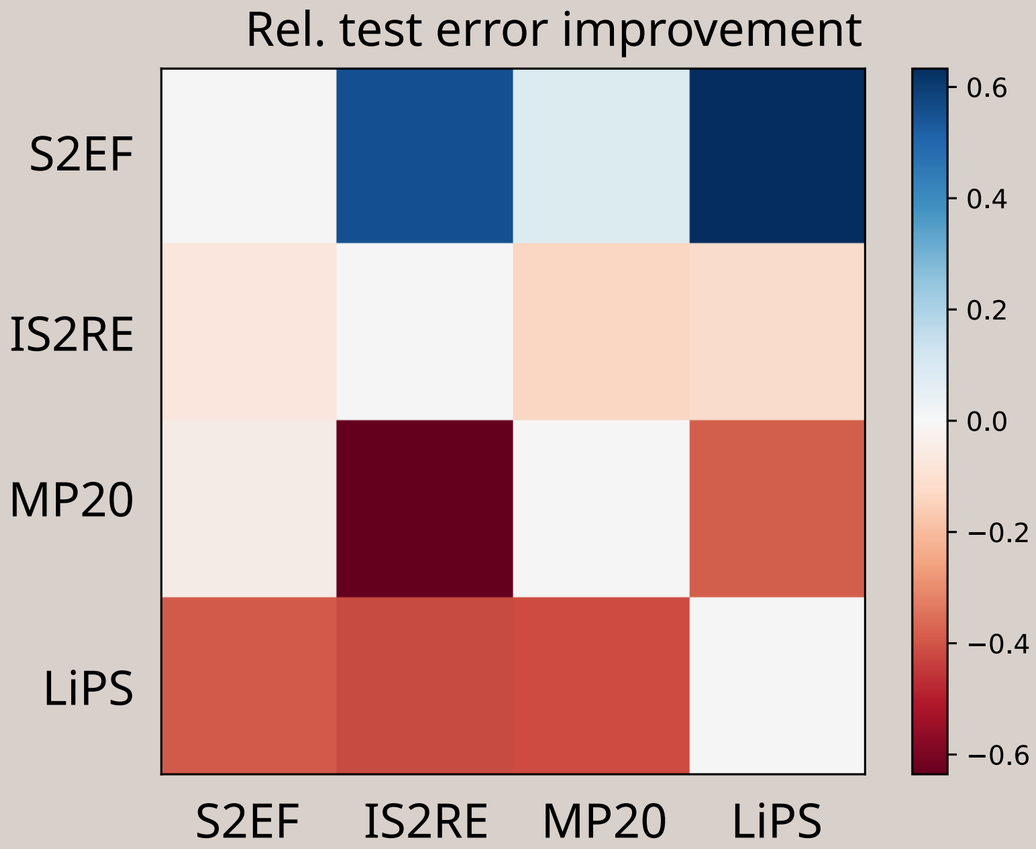
Lee, Gonzales, Nassar, Spellings, Galkin, Miret
MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling
arXiv preprint arXiv:2309.05934 (2023)



Base encoder is $E(n)$-GNN

Dataset composition and **task** alignment generally improves model performance

Lee, Gonzales, Nassar, Spellings, Galkin, Miret
MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling
arXiv preprint arXiv:2309.05934 (2023)

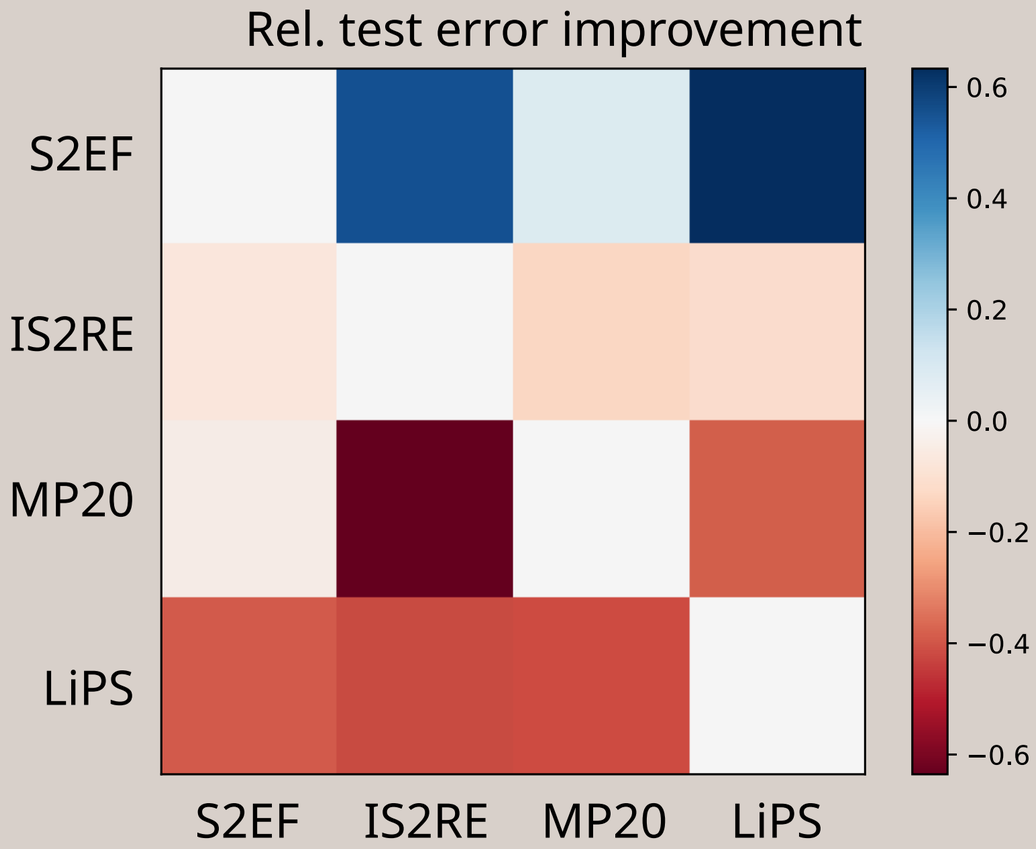


Base encoder is $E(n)$-GNN

Dataset composition and **task** alignment generally improves model performance

Poor or no alignment **degrades** performance

Lee, Gonzales, Nassar, Spellings, Galkin, Miret
MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling
arXiv preprint arXiv:2309.05934 (2023)



Base encoder is $E(n)$-GNN

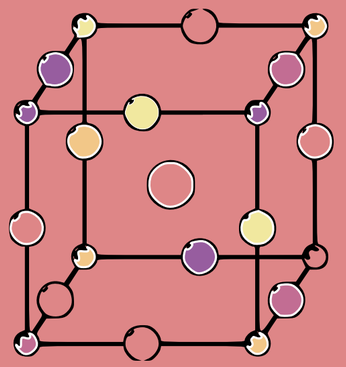
Dataset composition and **task** alignment generally improves model performance

Poor or no alignment **degrades** performance

Embedding visualizations offer a partial explanation

Lee, Gonzales, Nassar, Spellings, Galkin, Miret
MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling
arXiv preprint arXiv:2309.05934 (2023)

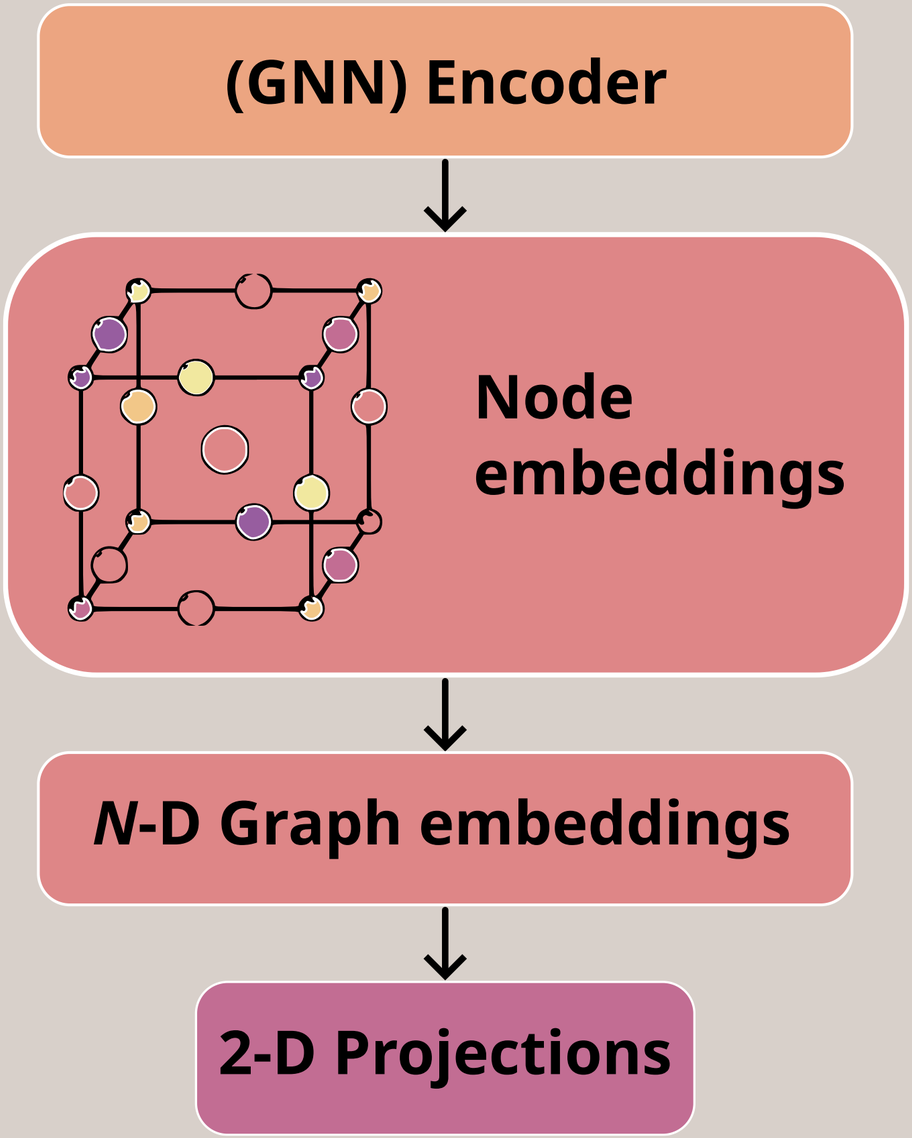
(GNN) Encoder



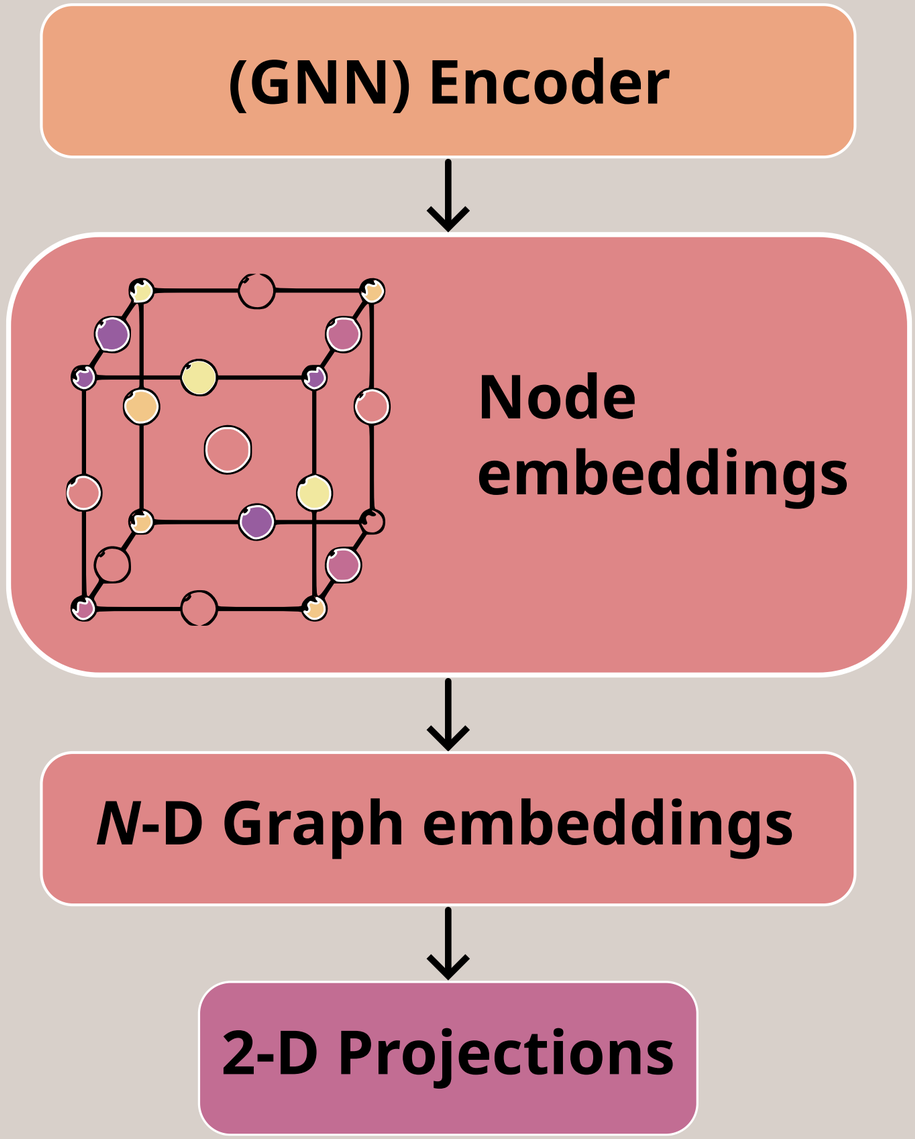
Node embeddings

N-D Graph embeddings

2-D Projections



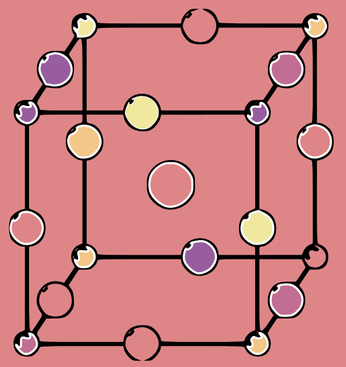
Qualitative explanation of observations using low-dimensional visualizations



Qualitative explanation of observations using low-dimensional visualizations

Techniques such as PCA, UMAP, and PHATE

(GNN) Encoder



Node embeddings

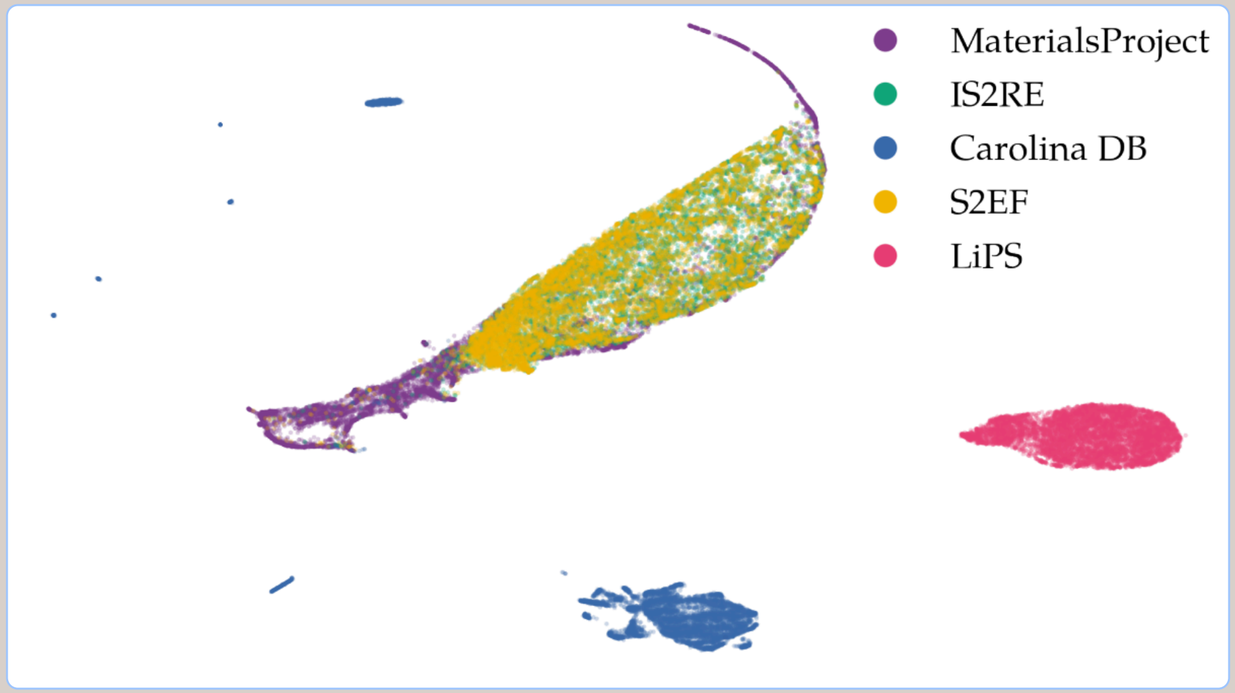
***N*-D Graph embeddings**

2-D Projections

Qualitative explanation of observations using low-dimensional visualizations

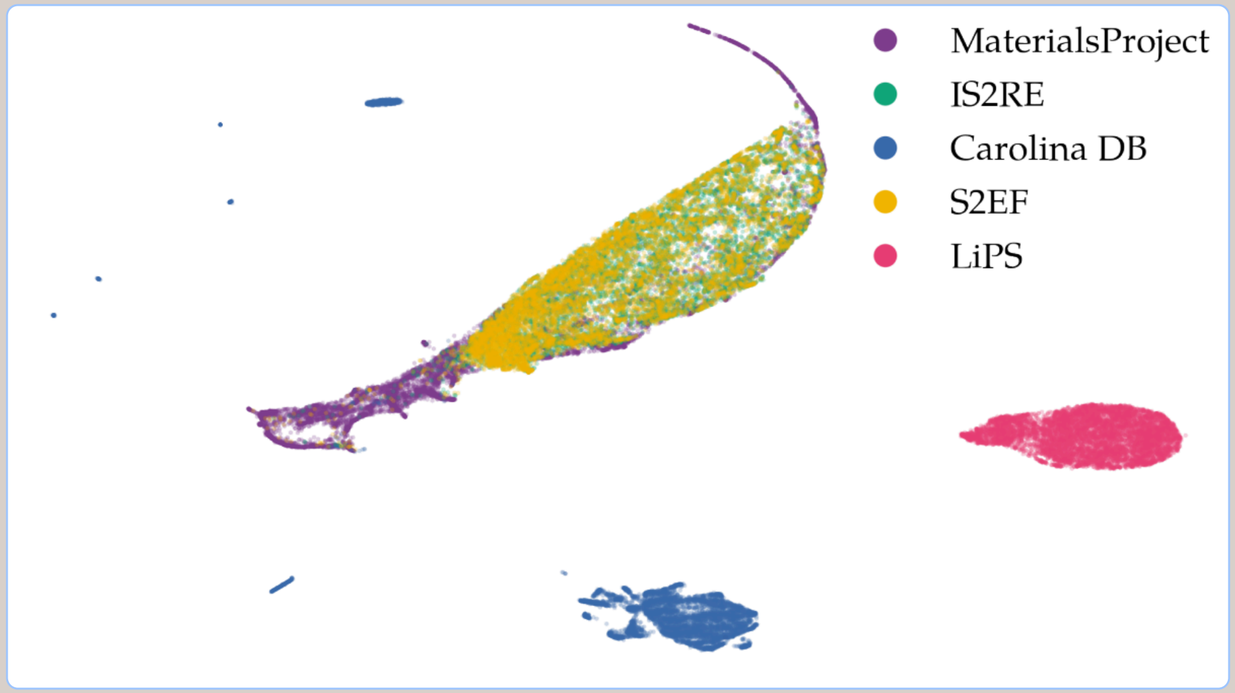
Techniques such as PCA, UMAP, and PHATE

Maps of composition and structural space—as seen by models trained on multiple tasks and datasets

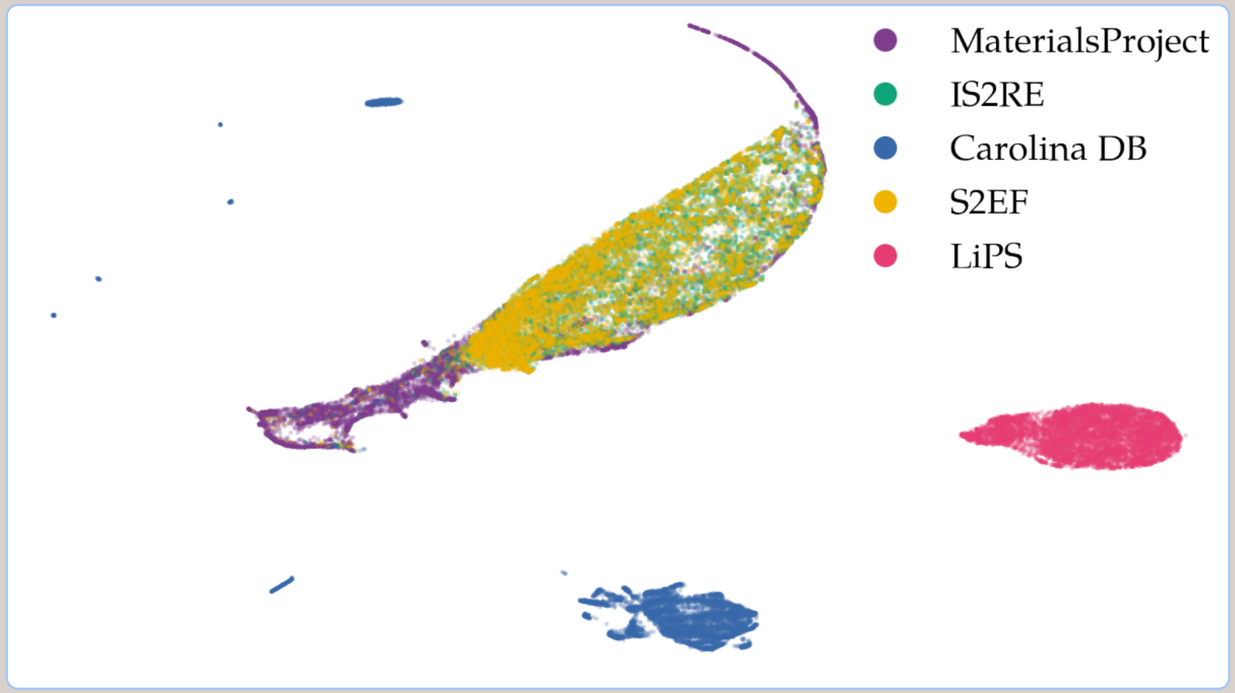


Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)

Projections of latent structures show dataset distribution modes



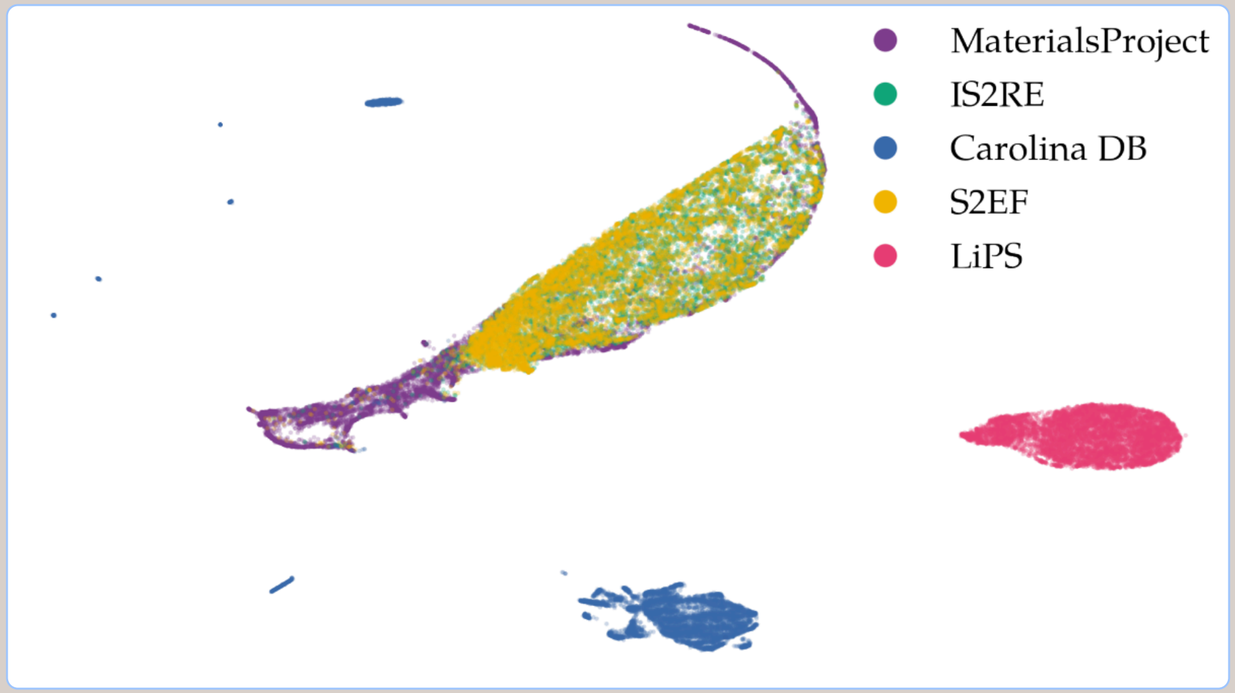
Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)



Projections of latent structures show dataset distribution modes

Model test performance improves with data overlap

Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)



Projections of latent structures show dataset distribution modes

Model test performance improves with data overlap

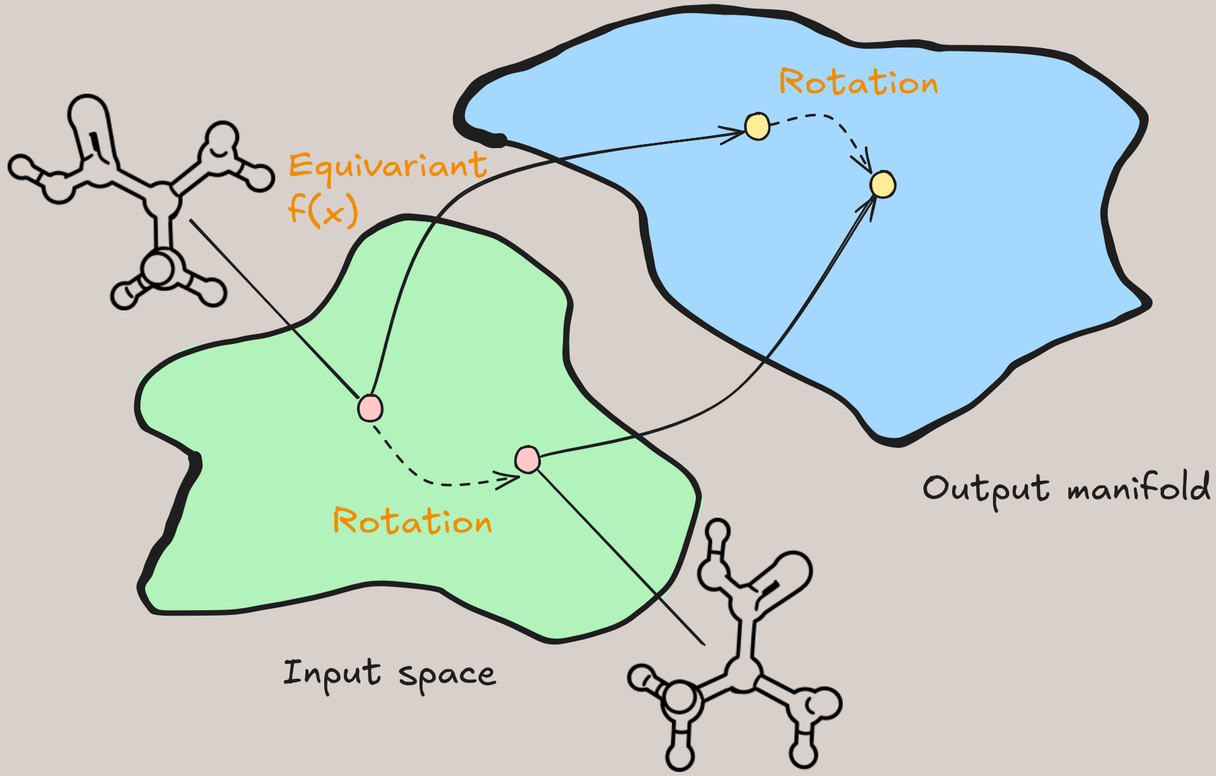
Representation engineering can potentially bridge these gaps

Lee, Gonzales, Spellings, Galkin, Miret, Kumar;
Towards Foundation Models for Materials Science,
Proceedings of the SC'23 Workshops (2023)

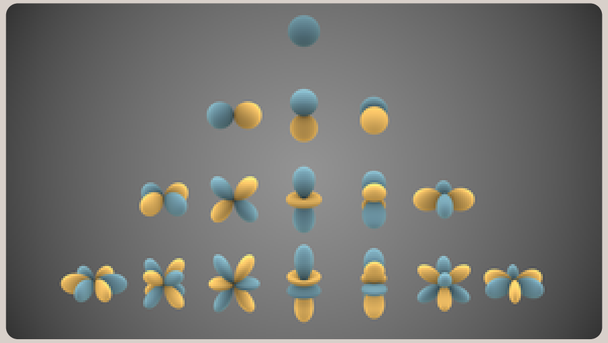
Architectural Explorations

Decomposing (equivariant) representations

Equivariant graph neural networks

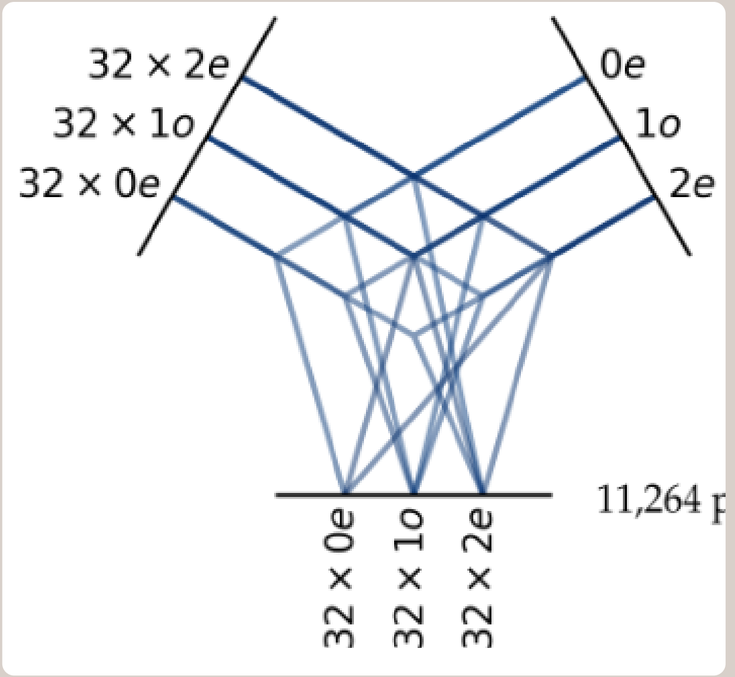


Preserved correspondence in object translations and rotations



Structures embedded with spherical harmonics become $SO(3)$ equivariant

Basis/feature size dependent on angular momentum l



Tensor products between angular-momentum conserving feature sets preserve equivariance

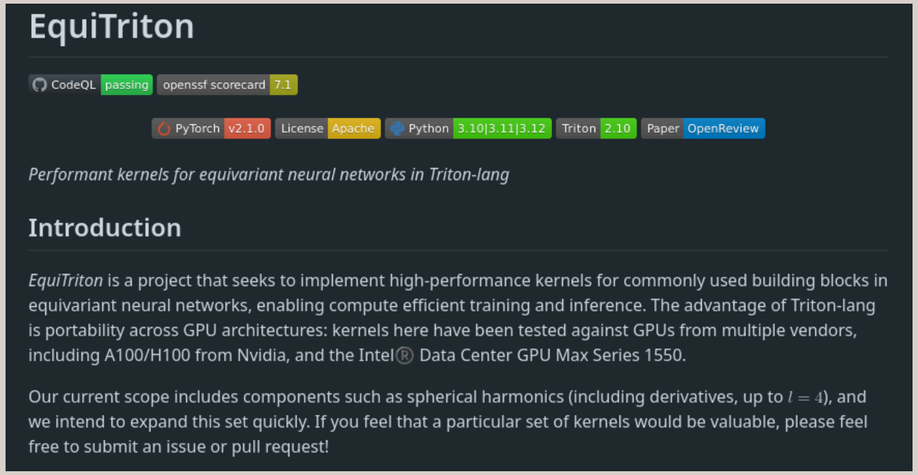
e3nn library

What do equivariant models learn?

$l = 0, 1, 2$ have physical interpretations—what about $l > 2$?

What do their latent variables represent?

IntelLabs/EquiTriton



EquiTriton

CodeQL passing | openssf scorecard 7.1

PyTorch v2.1.0 | License Apache | Python 3.10|3.11|3.12 | Triton 2.10 | Paper OpenReview

Performant kernels for equivariant neural networks in Triton-lang

Introduction

EquiTriton is a project that seeks to implement high-performance kernels for commonly used building blocks in equivariant neural networks, enabling compute efficient training and inference. The advantage of Triton-lang is portability across GPU architectures: kernels here have been tested against GPUs from multiple vendors, including A100/H100 from Nvidia, and the Intel® Data Center GPU Max Series 1550.

Our current scope includes components such as spherical harmonics (including derivatives, up to $l = 4$), and we intend to expand this set quickly. If you feel that a particular set of kernels would be valuable, please feel free to submit an issue or pull request!

Lee, Galkin, Miret; Deconstructing equivariant representations in molecular systems, NeurIPS AI4Mat (2024)

Triton-lang kernels for equivariant neural networks

Efficient spherical harmonics for up to $l = 10$

Apply projection methods to decomposed equivariant embeddings

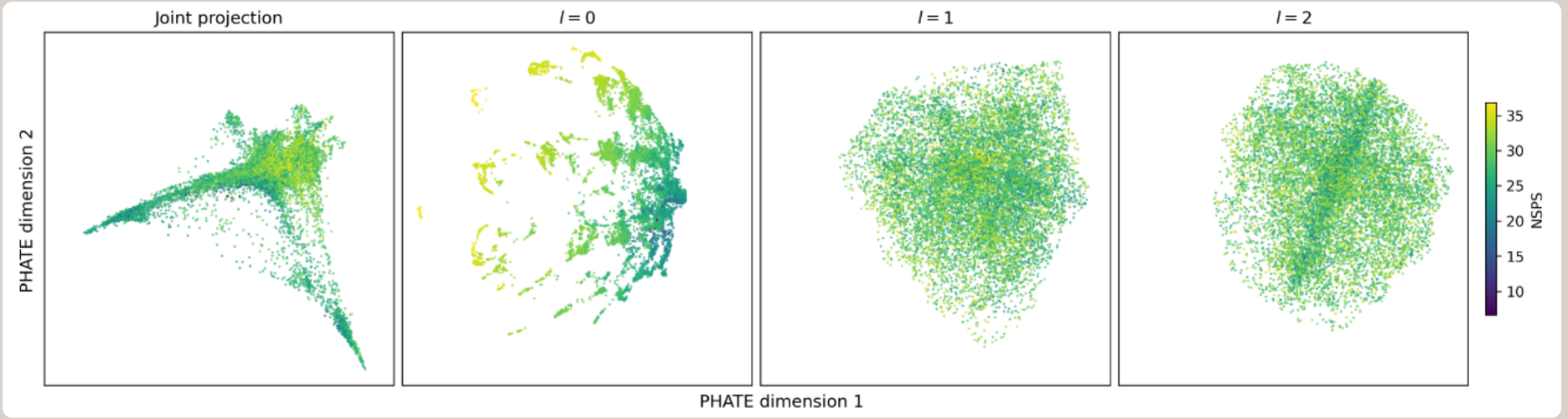
```python
def forward(
    self,
    atomic_features: torch.Tensor,
    coords: torch.Tensor,
    edge_index: torch.LongTensor,
) -> torch.Tensor:
    """
    High-level description:

    1. Project cartesian coordinates onto spherical harmonic basis
    2. Project interatomic distances onto radial (bessel) basis
    3. Transform radial basis functions with learnable weights
    4. Compute tensor product between scalar atom features and spherical harmonic basis
    5. Update node features
    """
    edge_dist = coords[edge_index[0]] - coords[edge_index[1]]
    sph_harm = self.spherical_harmonics(edge_dist)
    # calculate atomic distances, embed, and transform them
    edge_basis = self.edge_basis(edge_dist.norm(dim=-1))
    edge_z = self.fc(edge_basis)
    # compute tensor product
    messages = self.tensor_product(atomic_features[edge_index[0]], sph_harm, edge_z)
    # update node features
    hidden_feats = (
        scatter(messages, edge_index[1], dim=0, dim_size=atomic_features.size(0))
        / self.degree_norm
    )
    return hidden_feats
```

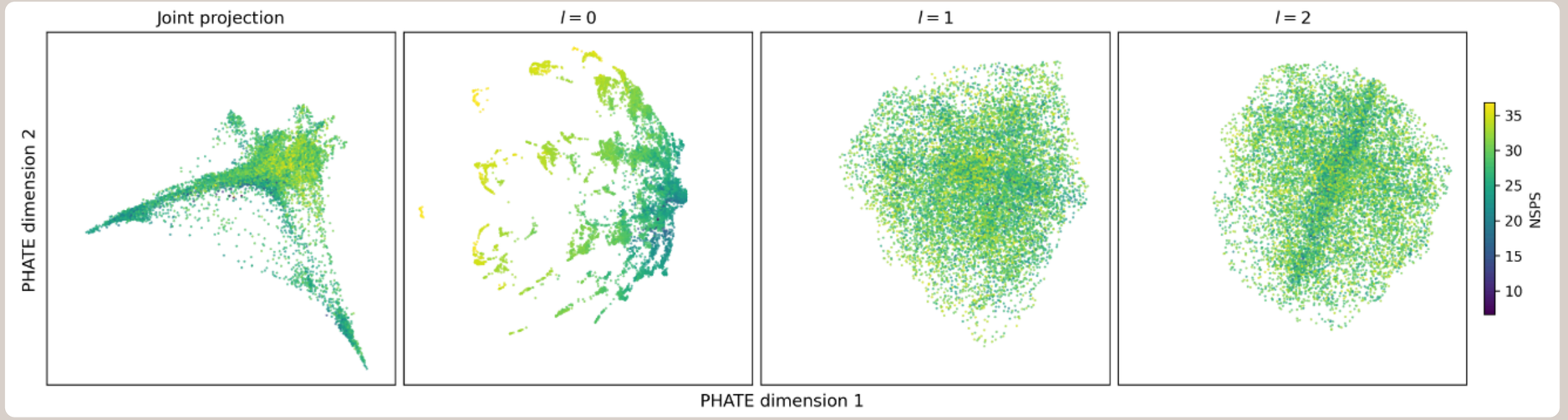
Equivariant interaction block

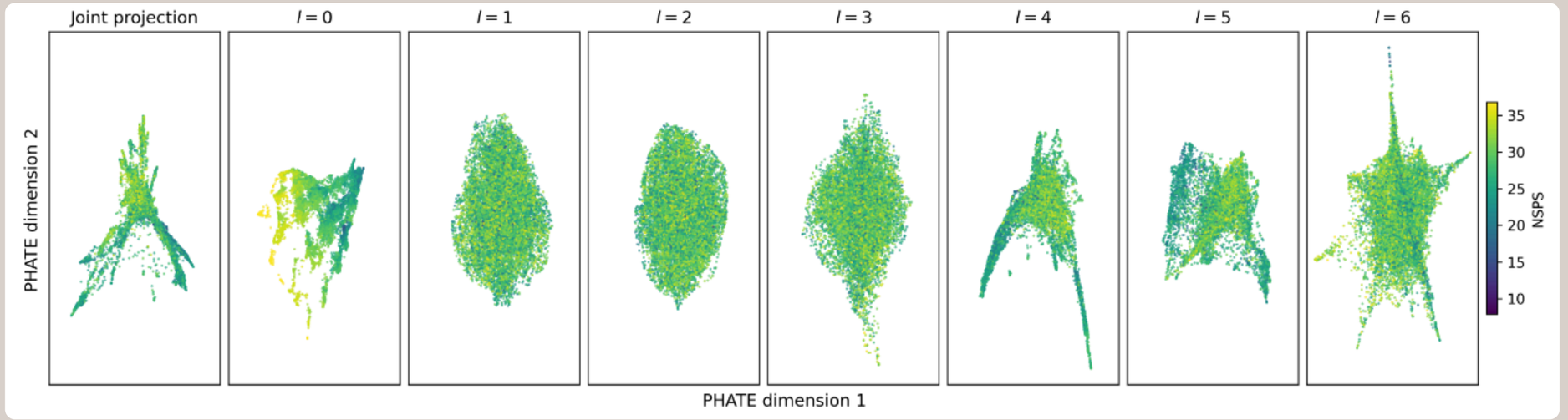
Model learns representation via QM9 energy prediction

Dimensionality of node/graph embeddings given by $h[2l+1]$

Projection analysis performed on features within given l


Joint projection
$l = 0$
$l = 1$
$l = 2$
PHATE dimension 2
PHATE dimension 1
NSPS
35
30
25
20
15
10


Joint projection
l = 0
l = 1
l = 2
PHATE dimension 2
PHATE dimension 1
NSPS
35
30
25
20
15
10


Joint projection
l = 0
l = 1
l = 2
l = 3
l = 4
l = 5
l = 6
PHATE dimension 2
PHATE dimension 1
NSPS
35
30
25
20
15
10

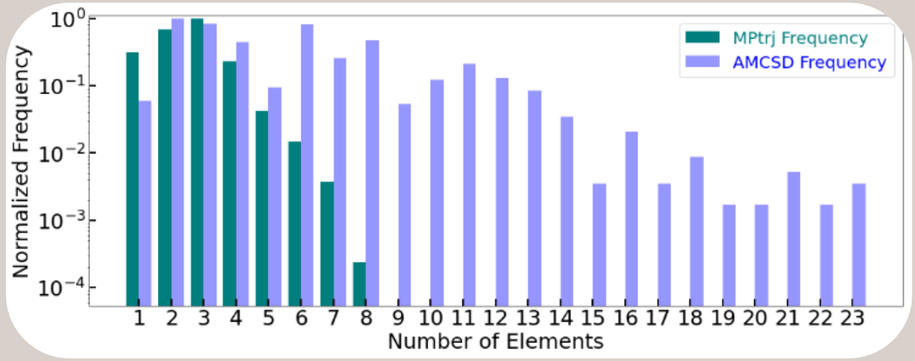
Future & ongoing work

Improving generalization through classifier pretraining and regularization

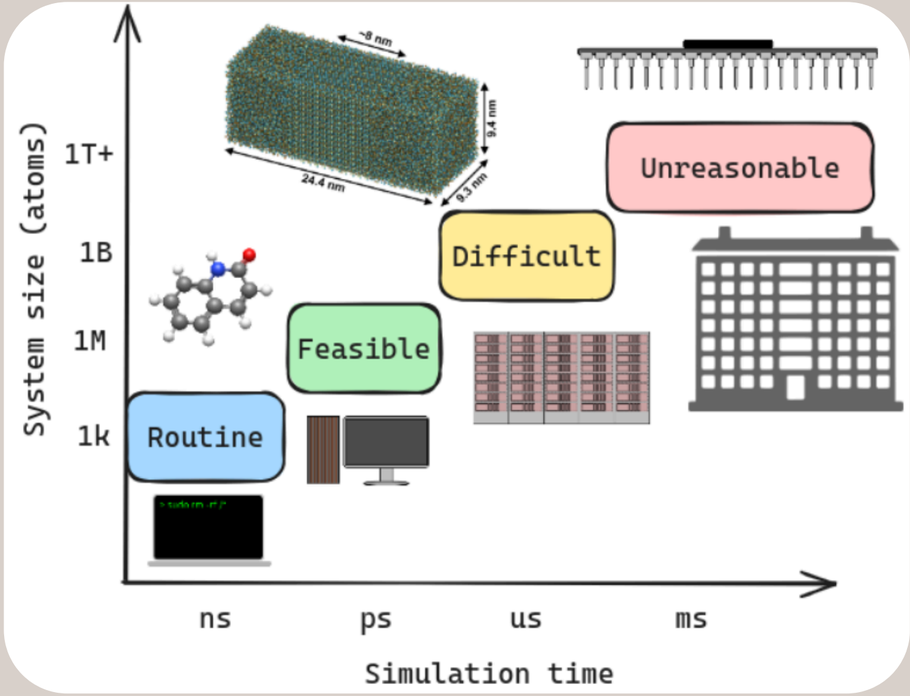
Quantiative metrics on embeddings—lessons from semi-self-supervised learning

Path to a useful map

...or production



Aligning data domains



Coupled cluster at device scale

Miret, **Lee**, Gonzales, Mannan, Krishnan
Energy & force regression on DFT trajectories is not enough for universal machine learning interatomic potentials,
Submitted to ICML (2025)
arXiv:2502.03660

Summary

Developed Open MatSciML Toolkit as a framework for experimentation

Summary

Summary

Developed Open MatSciML Toolkit as a framework for experimentation

Composable framework enabled search for data synergies in foundation model training

Summary

Developed Open MatSciML Toolkit as a framework for experimentation

Composable framework enabled search for data synergies in foundation model training

Modeling performance stems not just from data, but also from learned latents

Summary

Developed Open MatSciML Toolkit as a framework for experimentation

Composable framework enabled search for data synergies in foundation model training

Modeling performance stems not just from data, but also from learned latents

Experiments show physical inductive biases may still need regularization

Summary

Developed Open MatSciML Toolkit as a framework for experimentation

Composable framework enabled search for data synergies in foundation model training

Modeling performance stems not just from data, but also from learned latents

Experiments show physical inductive biases may still need regularization

Need to evaluate latent structures qualitatively!

Acknowledgements

Intel Labs—

Santiago Miret · Carmelo Gonzales
Mikhail Galkin · Lory Wang

MIT—

Pete Miedaner · Shiang Fang
Keith Nelson · Tess Smidt

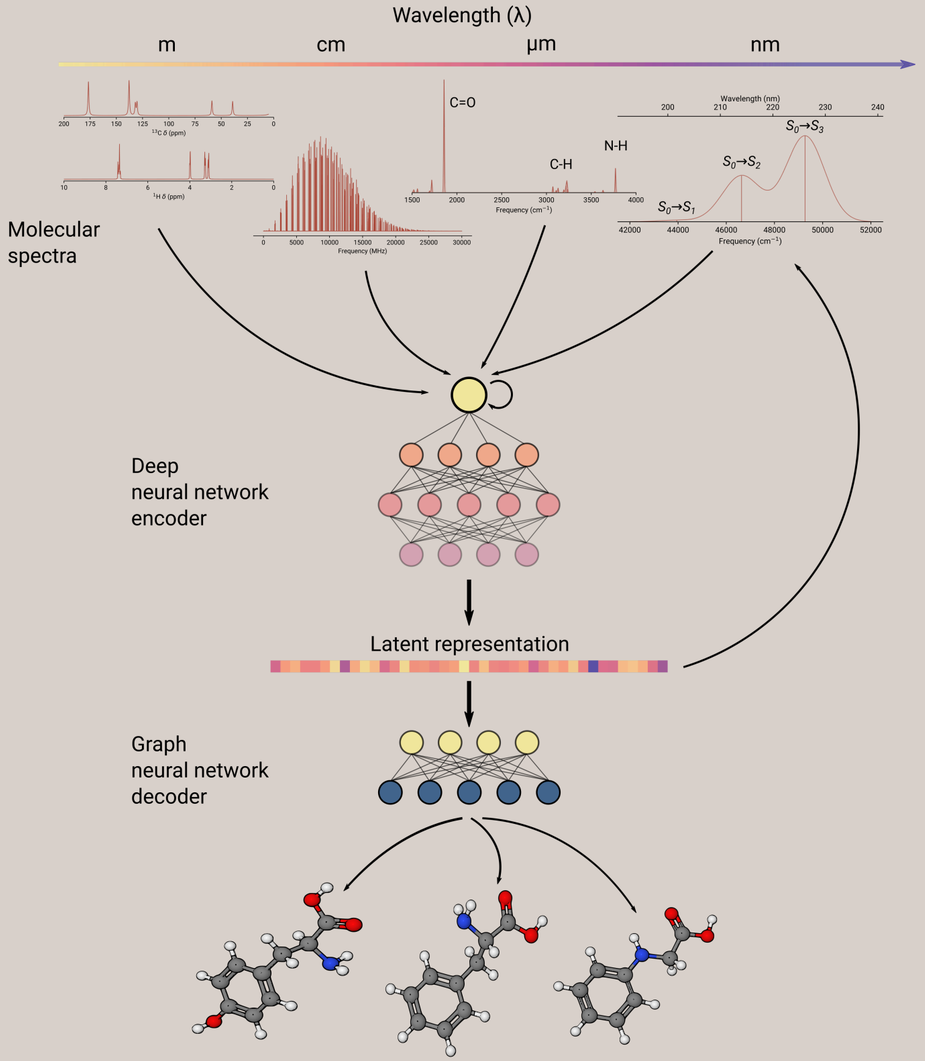
IIT Delhi—

Sajid Mannan · N. M. Anoop Krishnan

Pitch

What's needed to bridge the gap between model research and application

Holistic, "full stack" chemical autonomy needs **generation**, **simulation**, and **validation**



Unified neural representations beyond MLIPs

Google Scholar

LinkedIn

ICML Positions