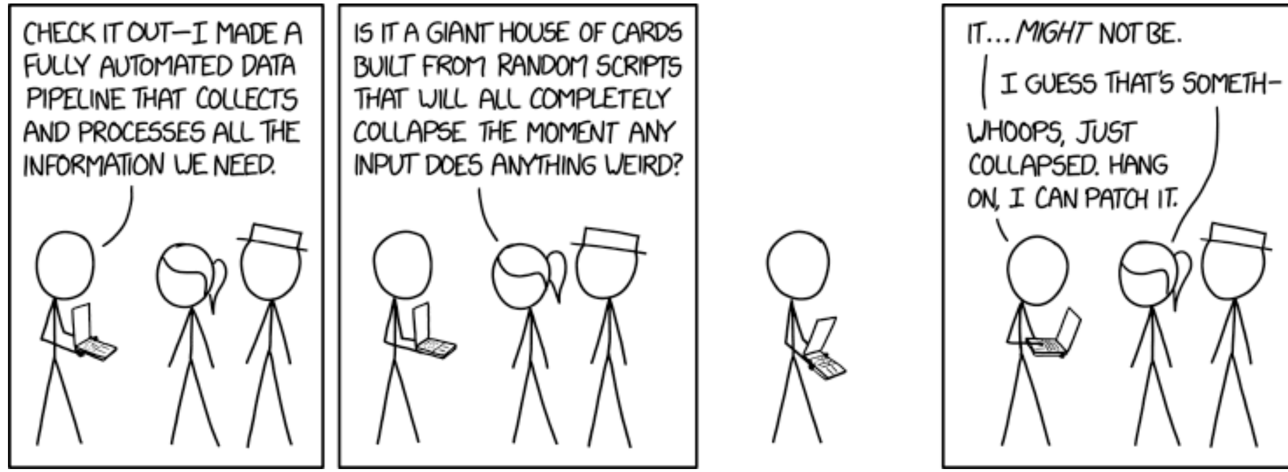


Transfer learning for materials science

Sai Gautam Gopalakrishnan[1], Keith Tobias Butler[2], Reshma Devi[1], and Pritam Ghosh[1]

[1]Materials Engineering, Indian Institute of Science

[2]Chemistry, University College London

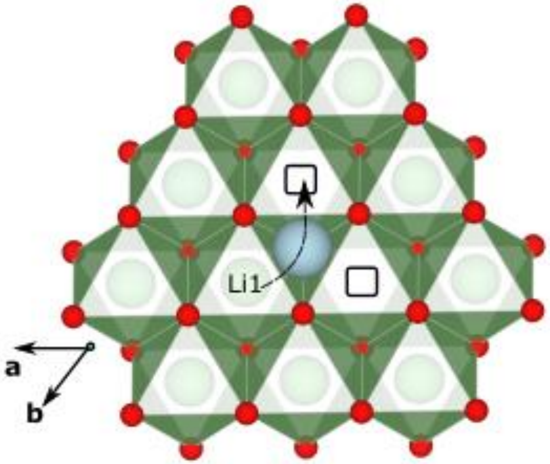
saigautamg@iisc.ac.in; https://sai-mat-group.github.io

Jan 8, 2025

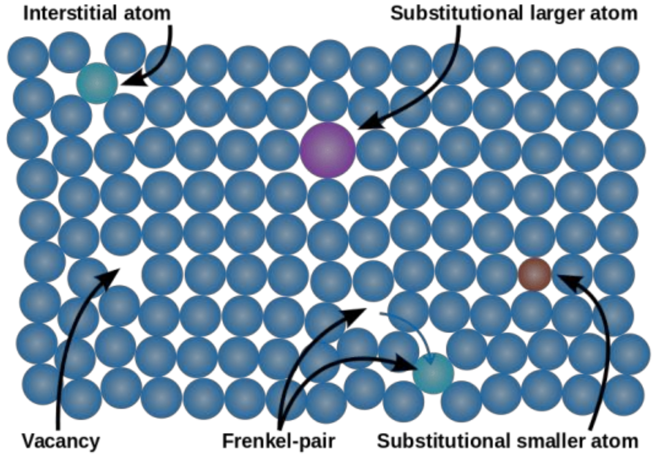
Materials science is data limited

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

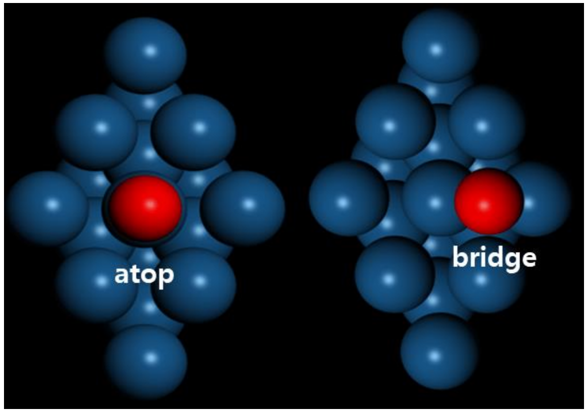


Devi et al., **npj Comput. Mater.** 2022



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

https://sites.psu.edu/



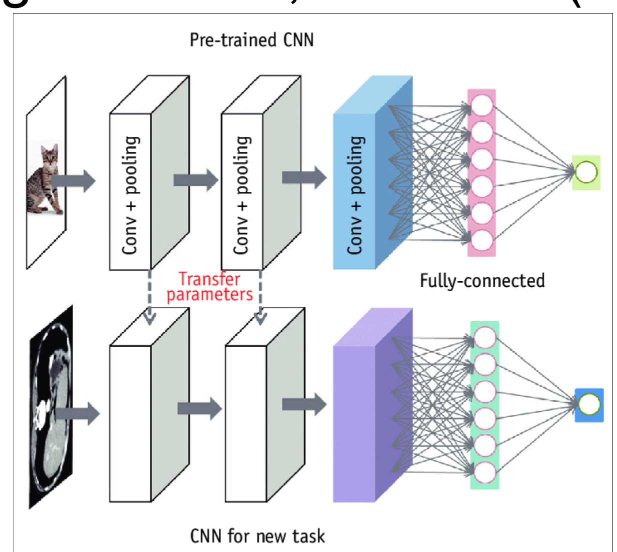
Materials science is data limited

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

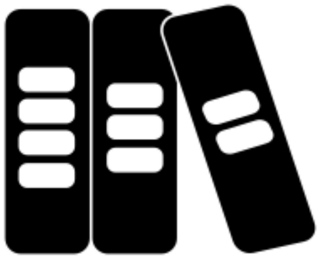
Transfer learning: efficiently use DL frameworks on small datasets

- 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?

Handles to consider

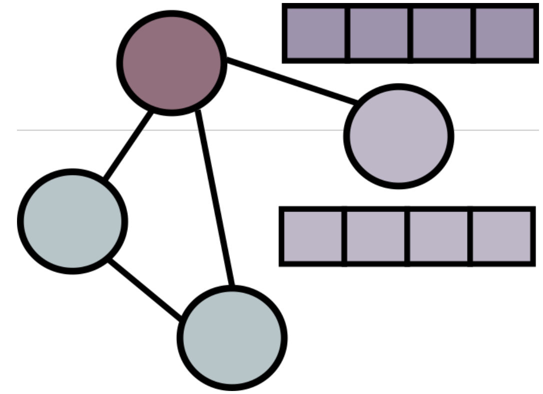


Dataset(s)

- What, how, how many?

Architecture

- Graph neural network



Frozen

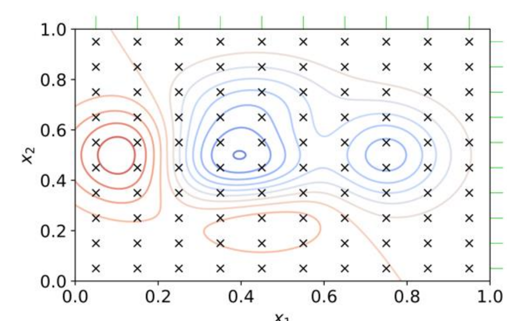
Unfrozen

Strategy

- FT techniques in pair-wise PT/FT models
- Multi-property PT (**MPT**) models

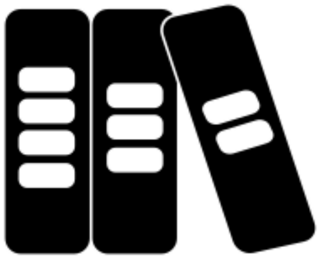
(Learning) Hyperparameters

- Data sampling
- Learning rate
- Number of datapoints during PT, FT



https://en.wikipedia.org/

Handles to consider

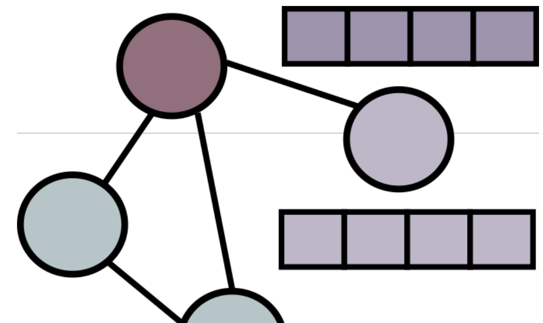


Dataset(s)

- What, how, how many?

Architecture

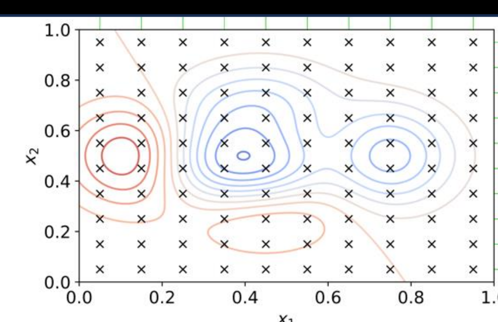
- Graph neural network



Let’s take a detailed look at the handles

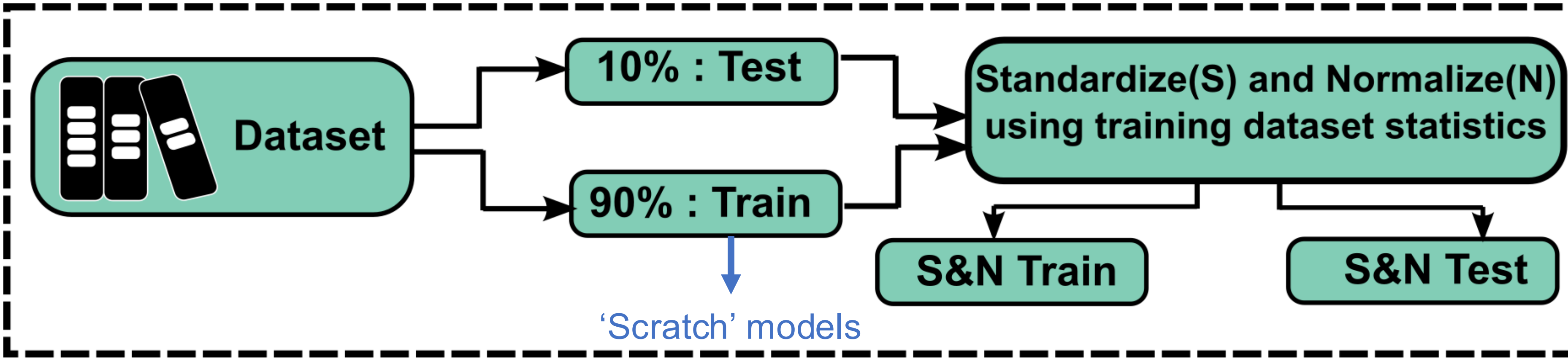
(Learning) Hyperparameters

- Data sampling
- Learning rate
- Number of datapoints during PT, FT



https://en.wikipedia.org/

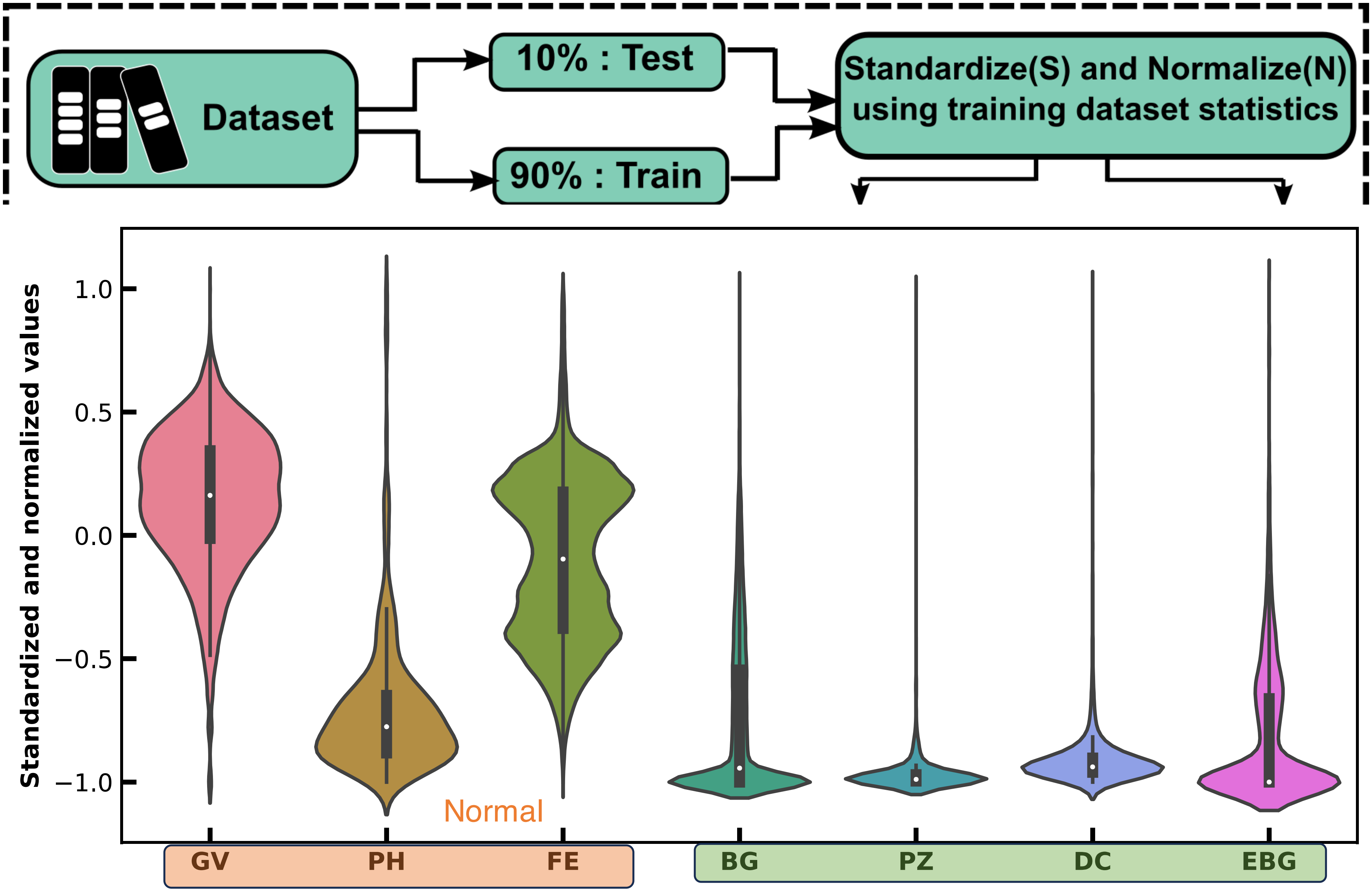
7 datasets (Matminer)



Dataset description	Abbreviation	Size	
Piezoelectric modulus	PZ	941	
Dielectric constant	DC	1,056	Computational
Highest frequency of optical phonon peak	PH	1,265	
Experimental band gap	EBG	4,604	Experimental
Average shear modulus	GV	10,987	
Band gap	BG	106,113	Computational
Formation energy	FE	132,752	

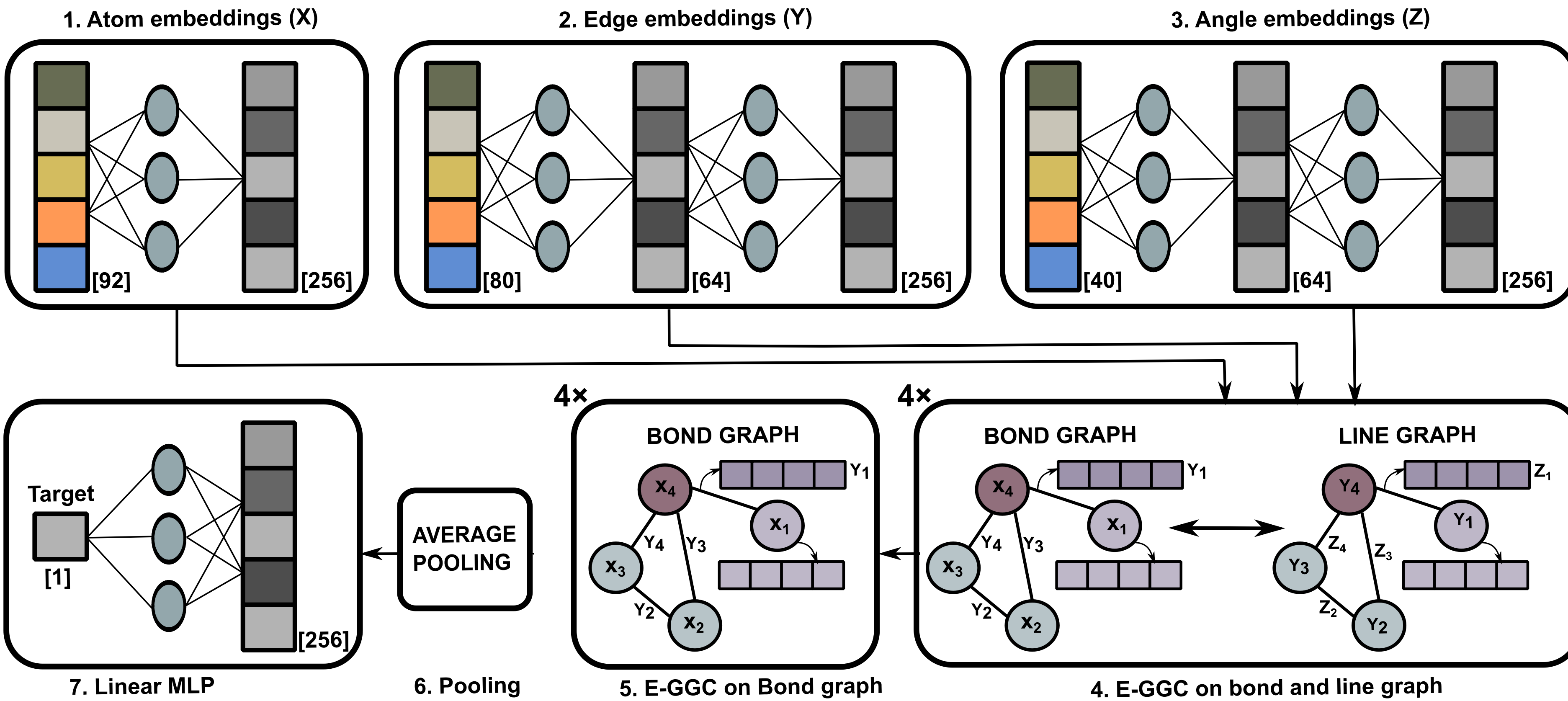
1. Ward *et al.*, **Comput. Mater. Sci.** *152*, 60–69 (2018).

7 datasets (Matminer)



1. Ward *et al.*, **Comput. Mater. Sci.** *152*, 60–69 (2018).

Atomistic line graph neural network (ALIGNN)



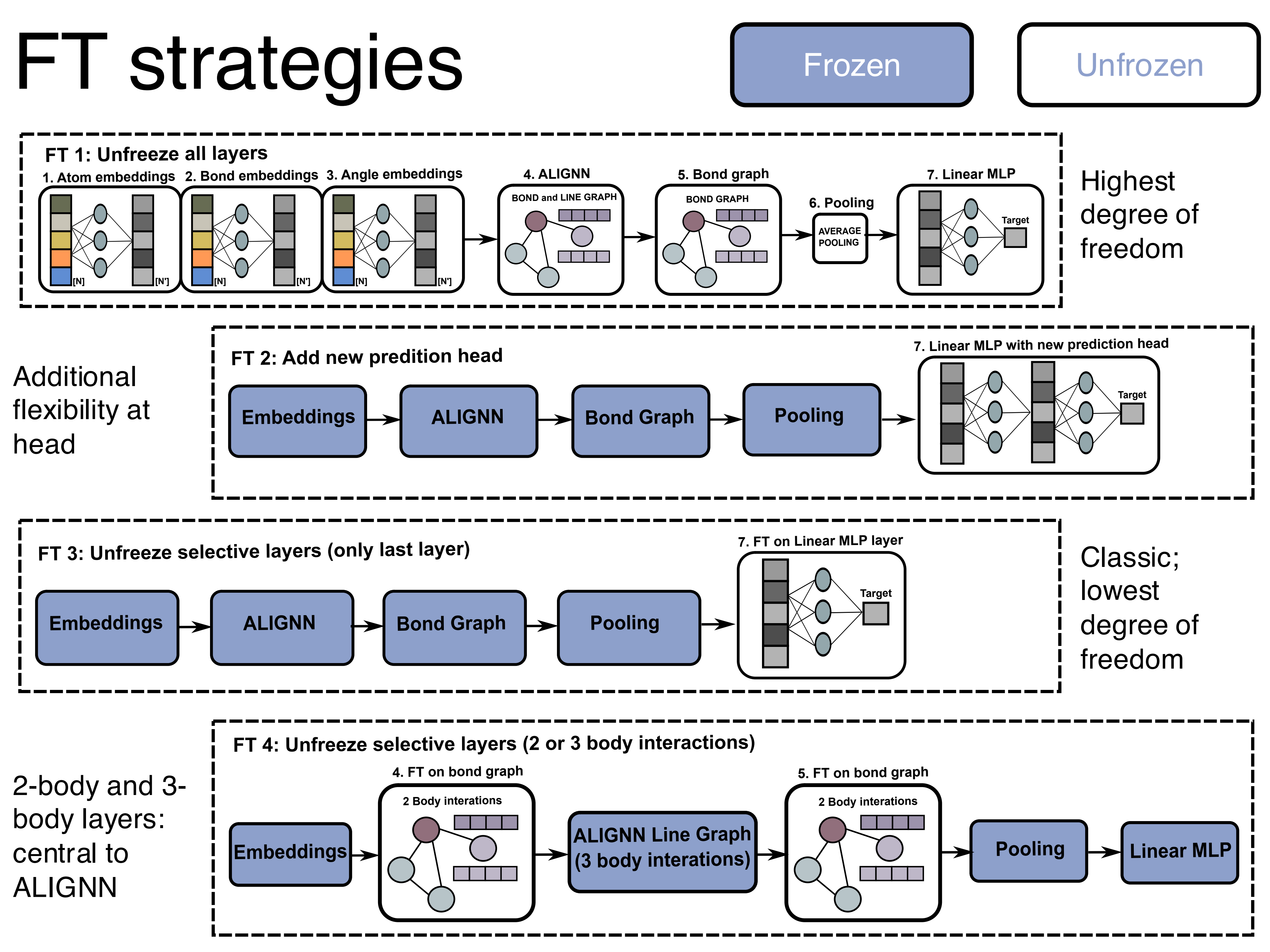
ALIGNN: Takes atoms, bonds, and bond angles into account

Bond graphs: atoms are nodes, bonds are edges; 2-body layers
Line graphs: bonds-nodes, bond angles-edges; 3-body layers
Communication: edge-gated graph convolution (E-GGC)

ALIGNN generalizes well 'out-of-distribution'[2]

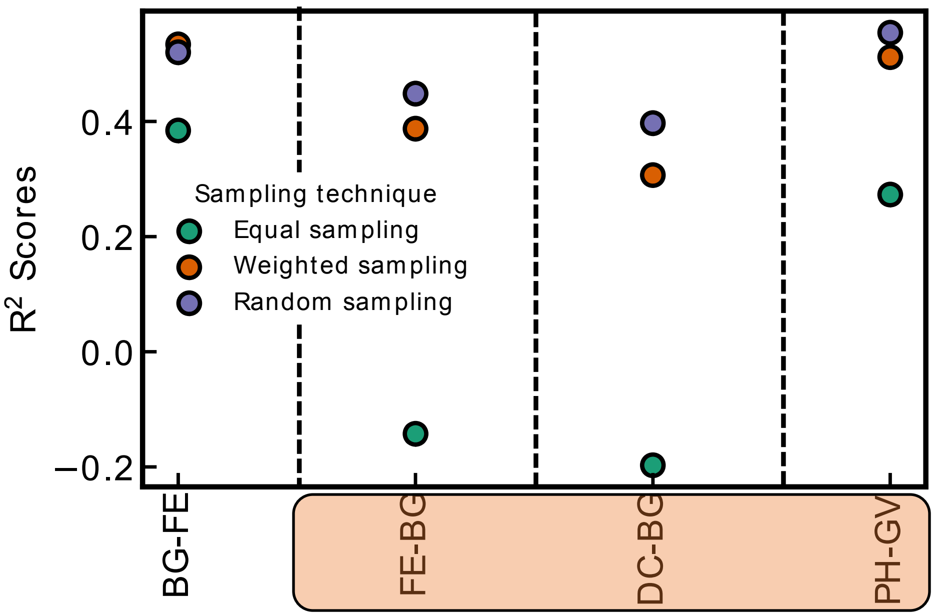
1. Choudhary and DeCost, **npj Comput. Mater.** *7*, 185 (2021).

2. Omee et al., **arXiv** 2401.08032 (2024).


FT strategies
Frozen
Unfrozen
FT 1: Unfreeze all layers
1. Atom embeddings
2. Bond embeddings
3. Angle embeddings
[N]
[N']
4. ALIGNN
BOND and LINE GRAPH
5. Bond graph
BOND GRAPH
6. Pooling
AVERAGE POOLING
7. Linear MLP
Target
Highest degree of freedom
Additional flexibility at head
FT 2: Add new predition head
Embeddings
ALIGNN
Bond Graph
Pooling
7. Linear MLP with new prediction head
Target
FT 3: Unfreeze selective layers (only last layer)
Embeddings
ALIGNN
Bond Graph
Pooling
7. FT on Linear MLP layer
Target
Classic; lowest degree of freedom
2-body and 3-body layers: central to ALIGNN
FT 4: Unfreeze selective layers (2 or 3 body interactions)
Embeddings
4. FT on bond graph
2 Body interations
ALIGNN Line Graph (3 body interations)
5. FT on bond graph
2 Body interations
Pooling
Linear MLP

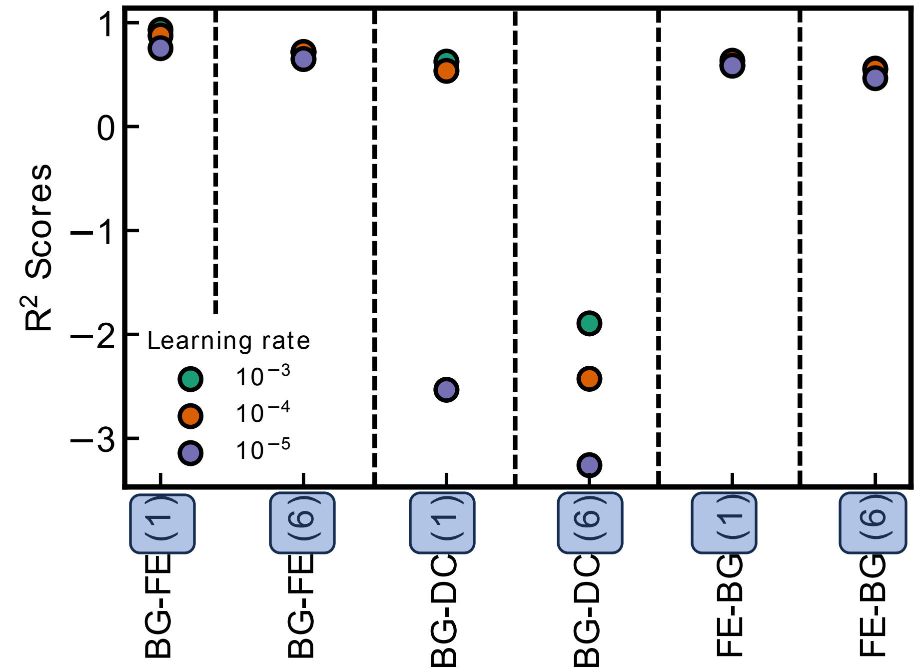
Hyperparameters

Sampling: random sampling is best



Random is better

Learning rate: higher is better



Different number of frozen layers

PT-FT: Pre-train dataset/Fine-tune dataset

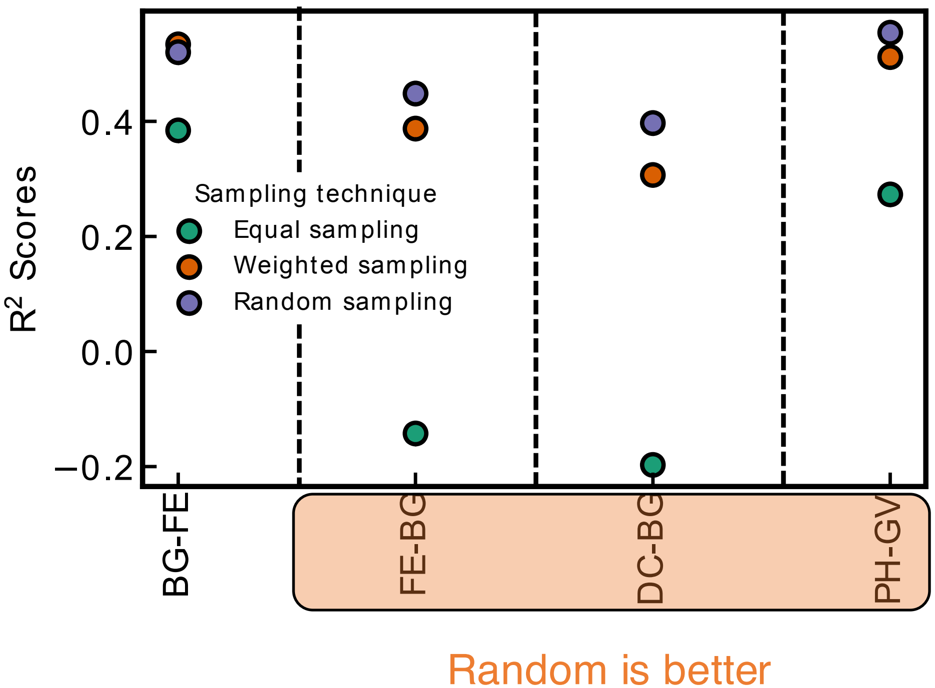
BG: Band gap
FE: Formation energy
DC: Dielectric constant
PH: Phonons

Higher learning rate: more re-training of parameters

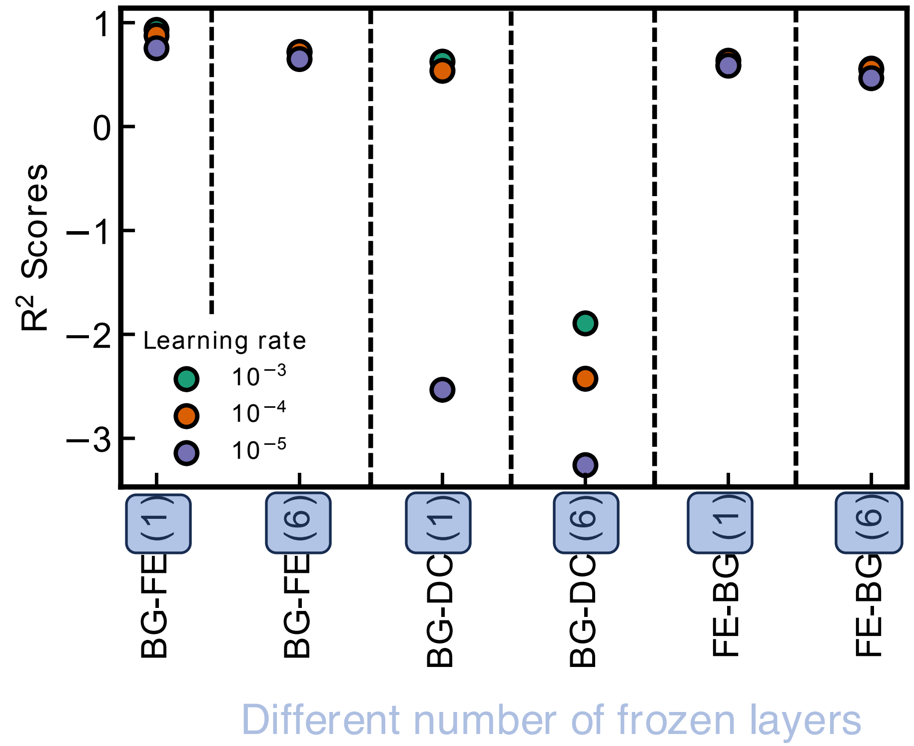
10^{-3} optimal; validation losses high at 10^{-2}

Hyperparameters

Sampling: random sampling is best



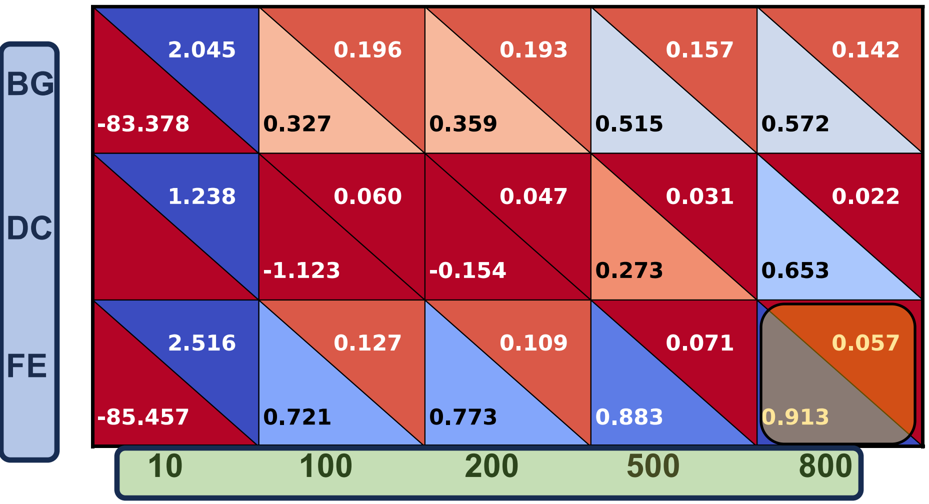
Learning rate: higher is better



Let’s look at pair-wise model performances in more detail

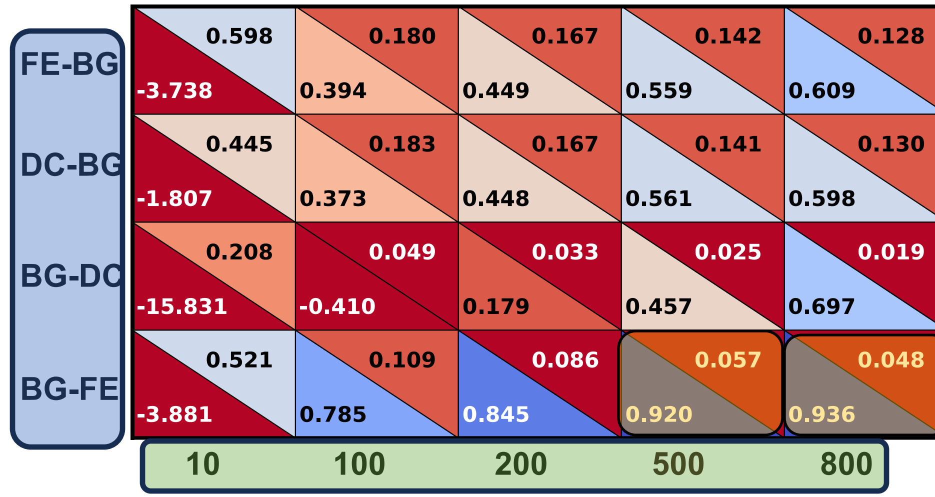
- Influence of PT/FT dataset size
- FT strategy
- 7×6 pair-wise models

More FT data: better



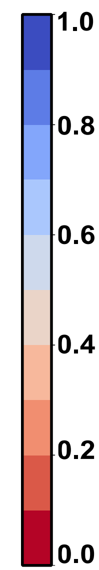
Scratch models

Dataset name; Dataset size



Pair-wise PT-FT models

PT-FT dataset name; FT Dataset size

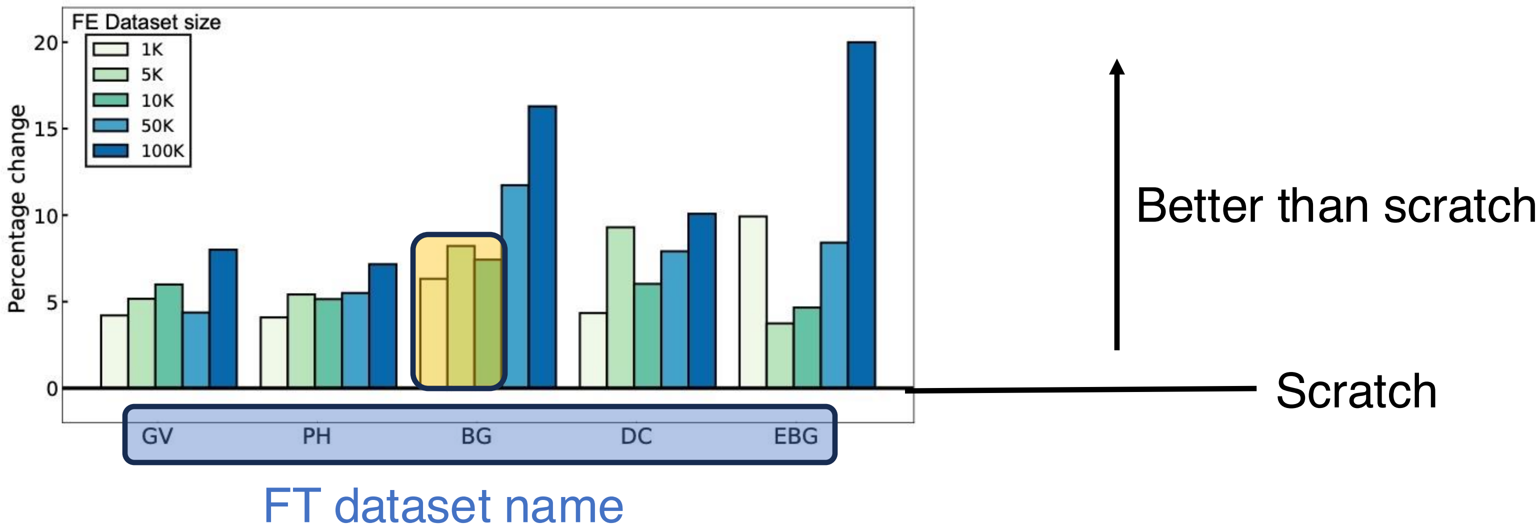


Test scores (5 trials)

- PT-FT models are consistently better than scratch models
- More FT data: better models (both scratch and PT-FT)
- Efficiency of PT-FT learning better than scratch (lower dataset sizes)

BG: Band gap; FE: Formation energy; DC: Dielectric constant; MAE: Mean absolute error

More PT data: non-monotonic improvement

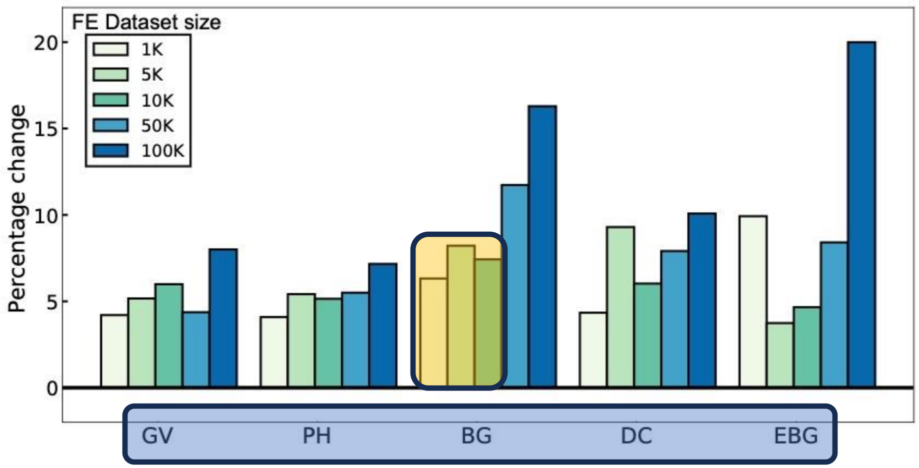


Formation energy as PT

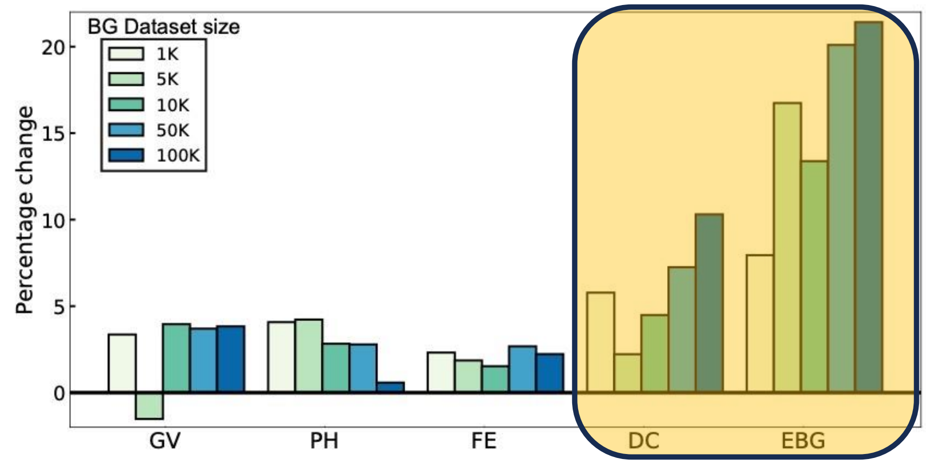
- Increasing dataset size: non-monotonicity
- Best models at 100K
- Always better than scratch

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

More PT data: non-monotonic improvement



FT dataset name



Formation energy as PT

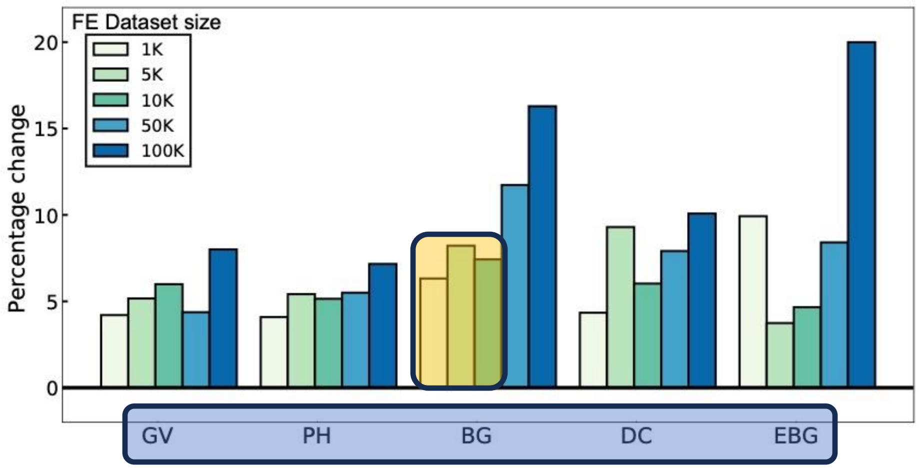
- Increasing dataset size: non-monotonicity
- Best models at 100K
- Always better than scratch

Band gap as PT

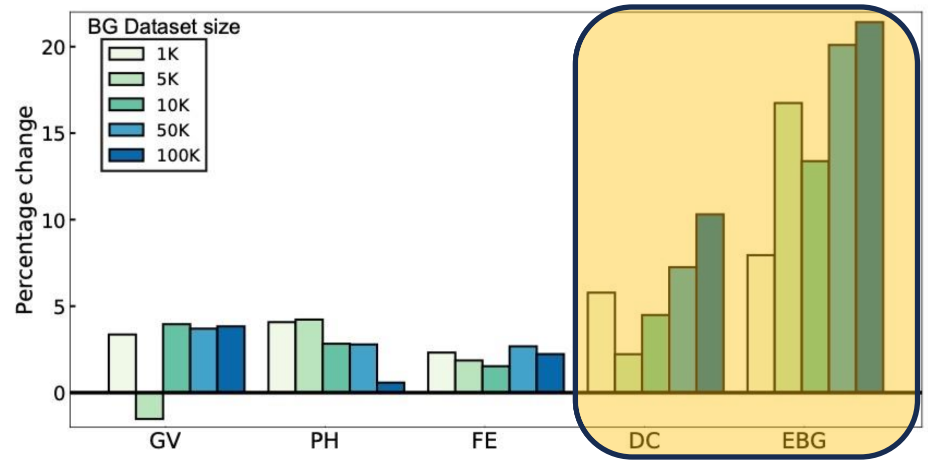
- Non-monotonicity
- Best models at 50K for non-correlated; 100K for correlated
- (Almost) always better than scratch

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

More PT data: non-monotonic improvement



FT dataset name

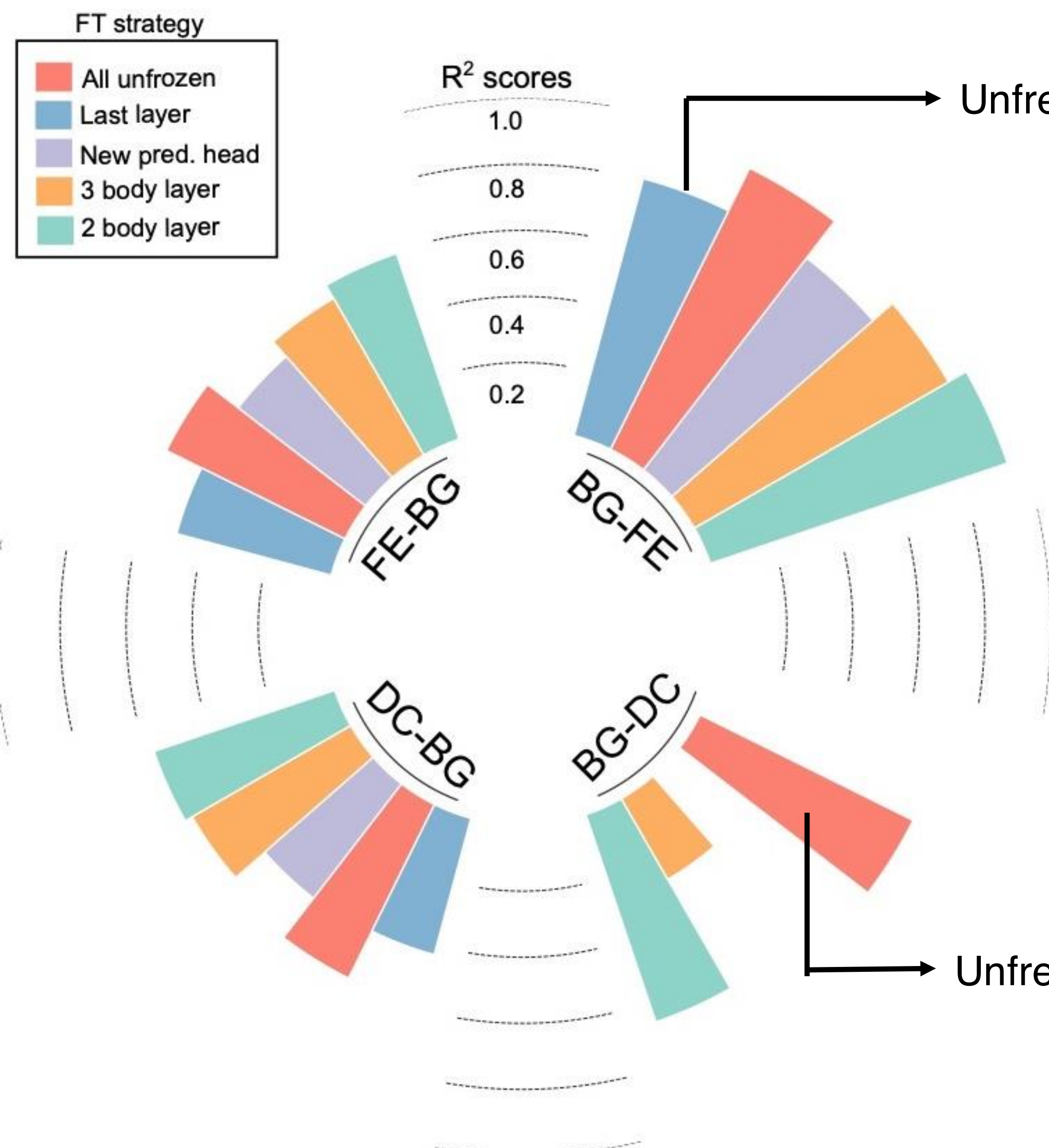


Larger PT data: generally better despite non-monotonic improvement

If FT property is correlated, more PT data helps

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

FT strategy: unfreezing all is best



Unfreezing last layer: minimal degree of freedom

Models need significant retraining

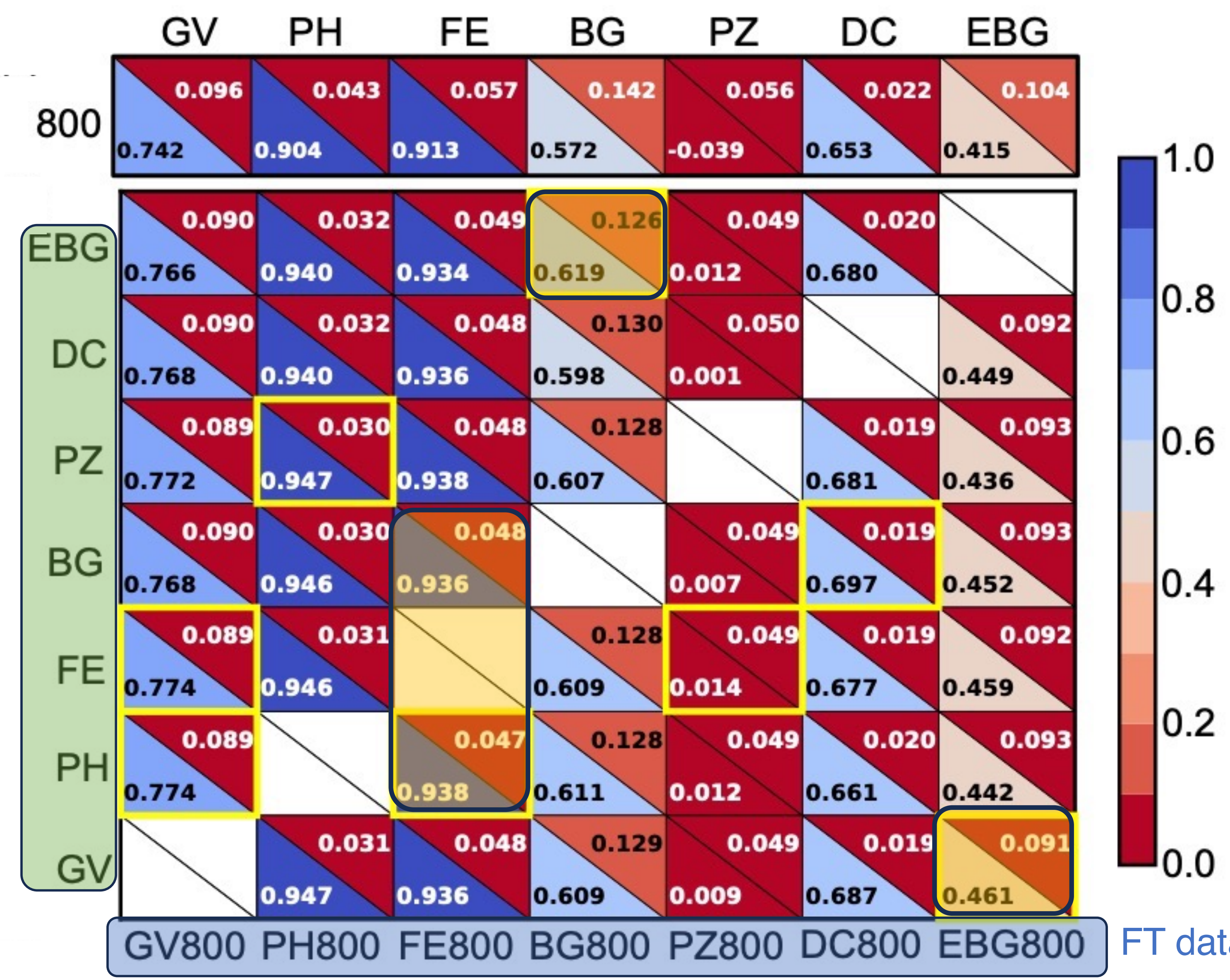
Creating foundational models: difficult in pair-wise

2-body and 3-body layers are large contributors to ALIGNN

Unfreezing all layers: maximal degree of freedom

BG: Band gap; FE: Formation energy; DC: Dielectric constant

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

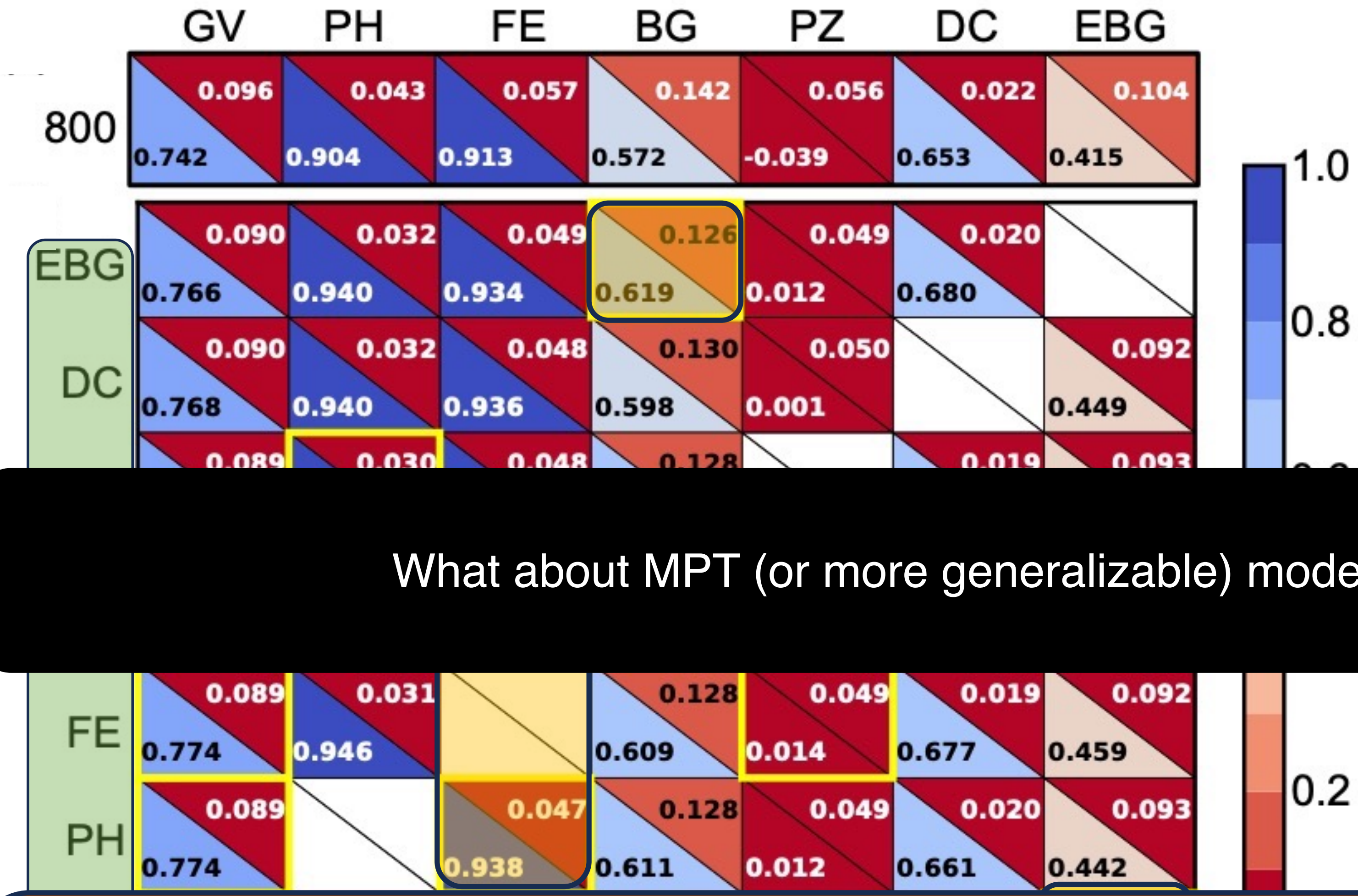
PT dataset (941)

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 ($R^2 > 0.75$)

What about MPT (or more generalizable) models?

property: little influence on FT

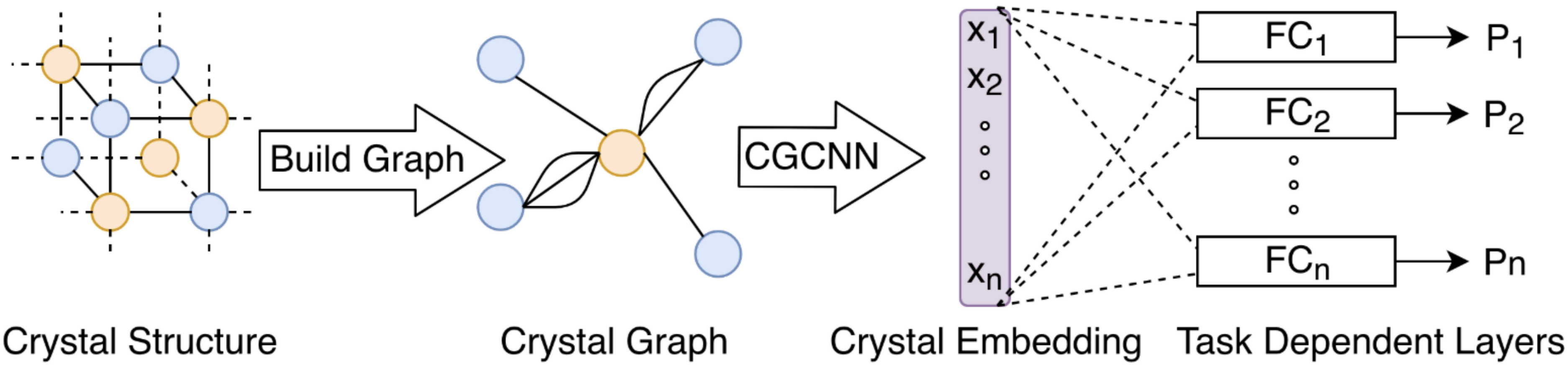
No symmetry

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

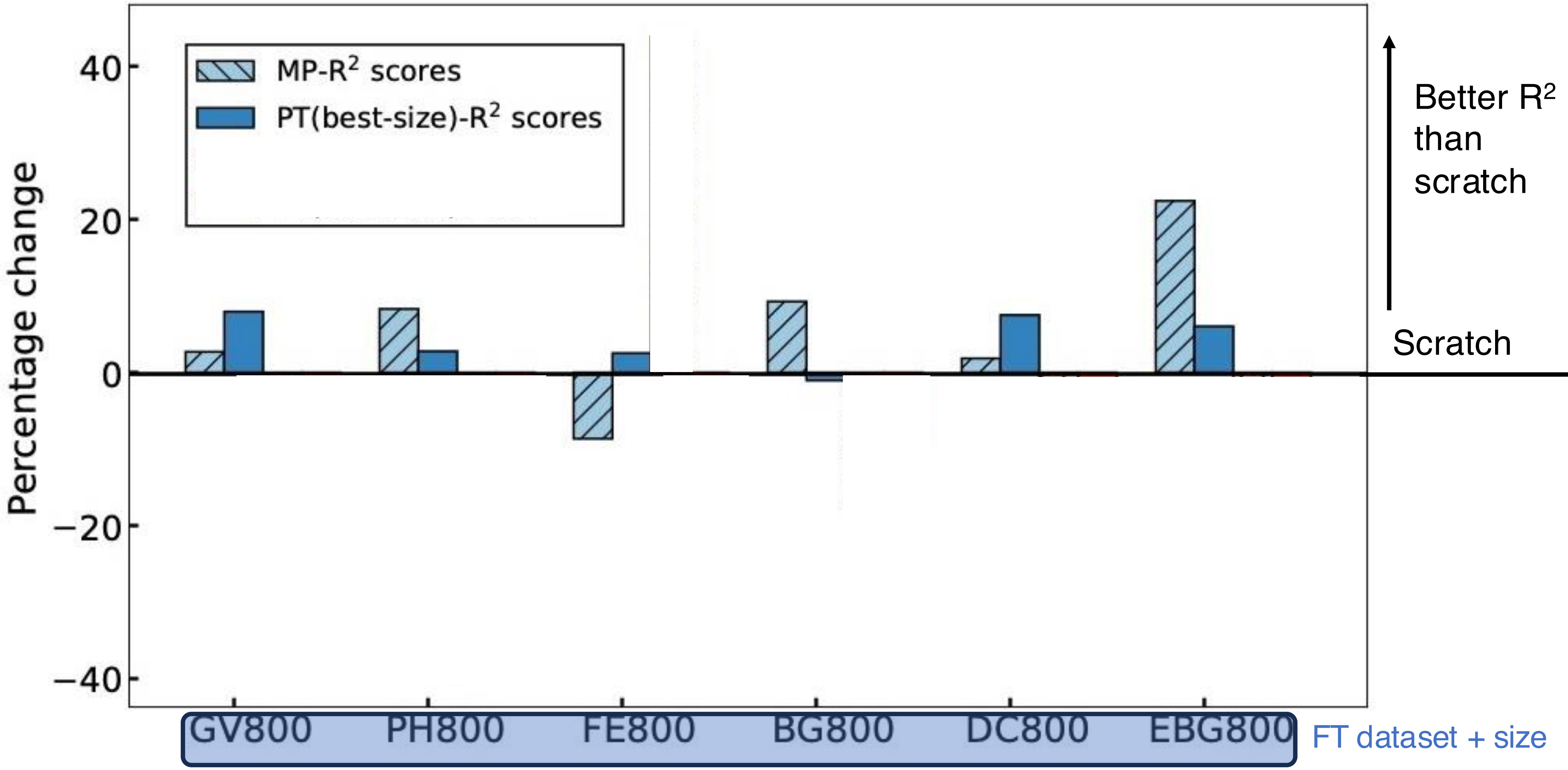
- Build cumulative dataset: 132,270 points
  - Remove overlaps
- Add task-dependent prediction heads with a one-hot encoded vector
  - Presence/absence of property
- Modify loss function

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

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

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

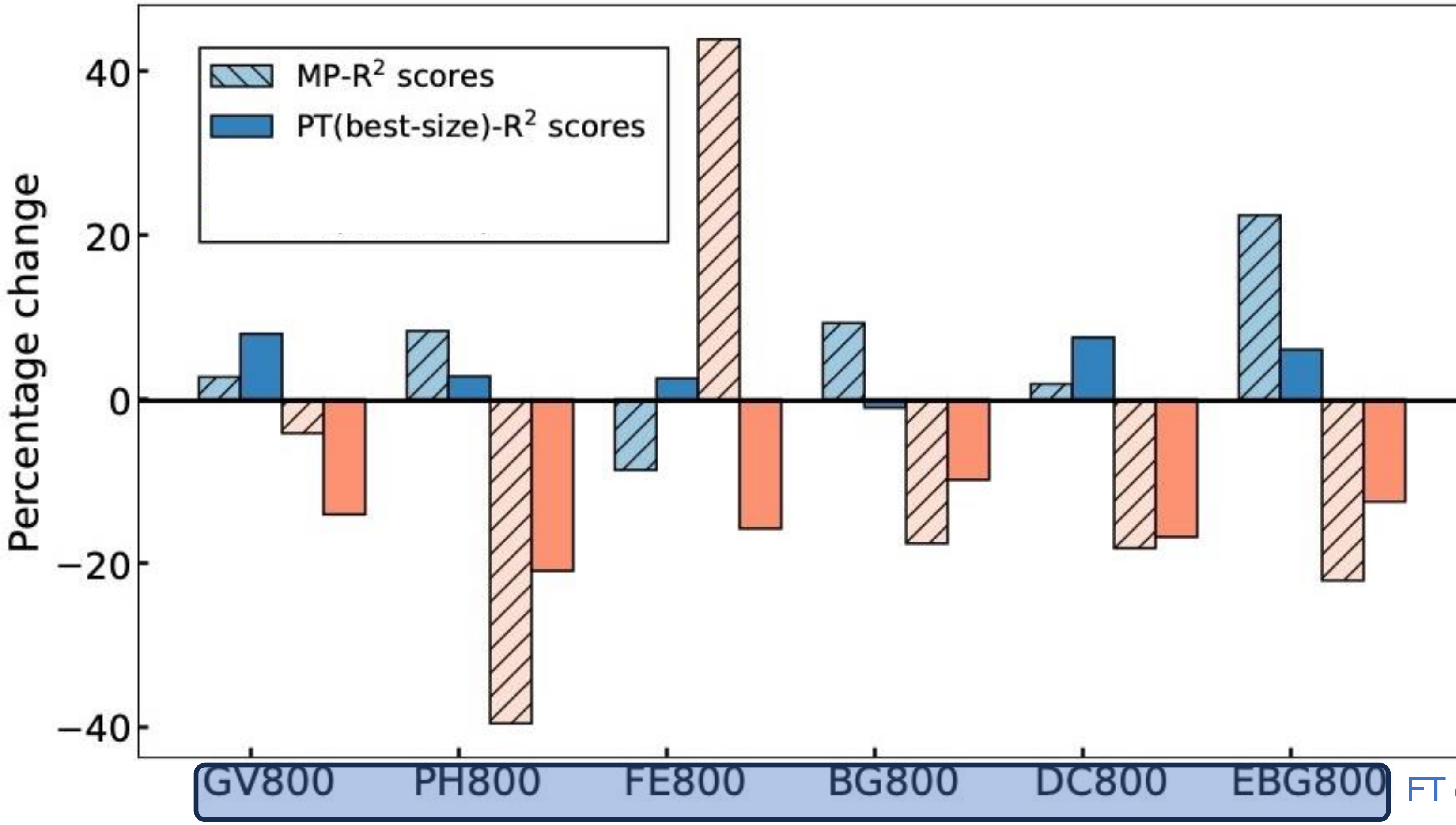
MPT: better on-average than PT-FT



MPT better than PT(best-size) in 3/6 on R^2

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

MPT: better on-average than PT-FT



Better R^2 than scratch

Scratch

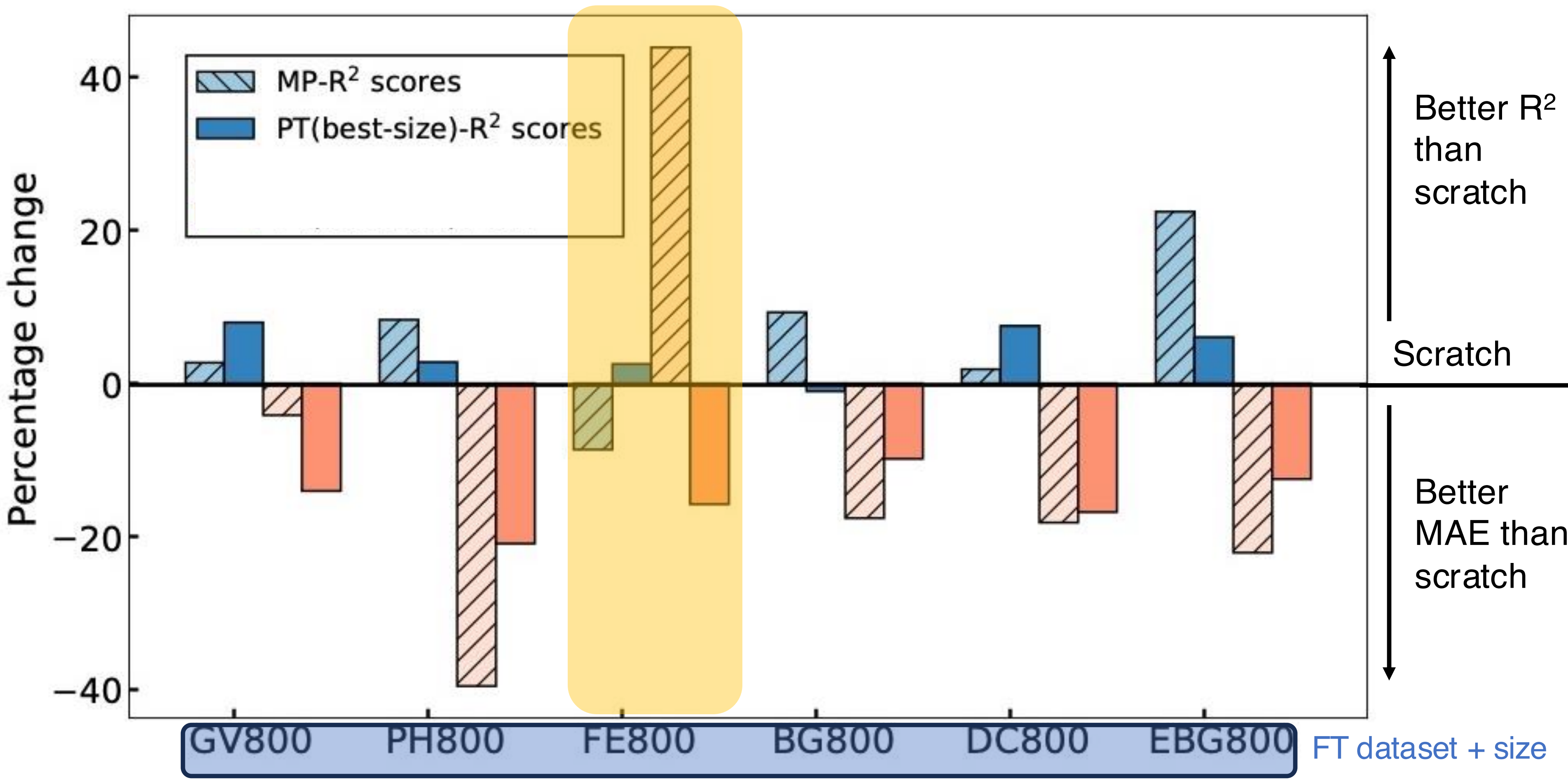
Better MAE than scratch

FT dataset + size

MPT better than PT(best-size) in 3/6 on R^2
in 4/6 on MAE

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

MPT: better on-average than PT-FT



MPT better than PT(best-size) in 3/6 on R^2
in 4/6 on MAE

Negative transfer in FE with MPT

- Due to exclusion of large number of datapoints

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

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

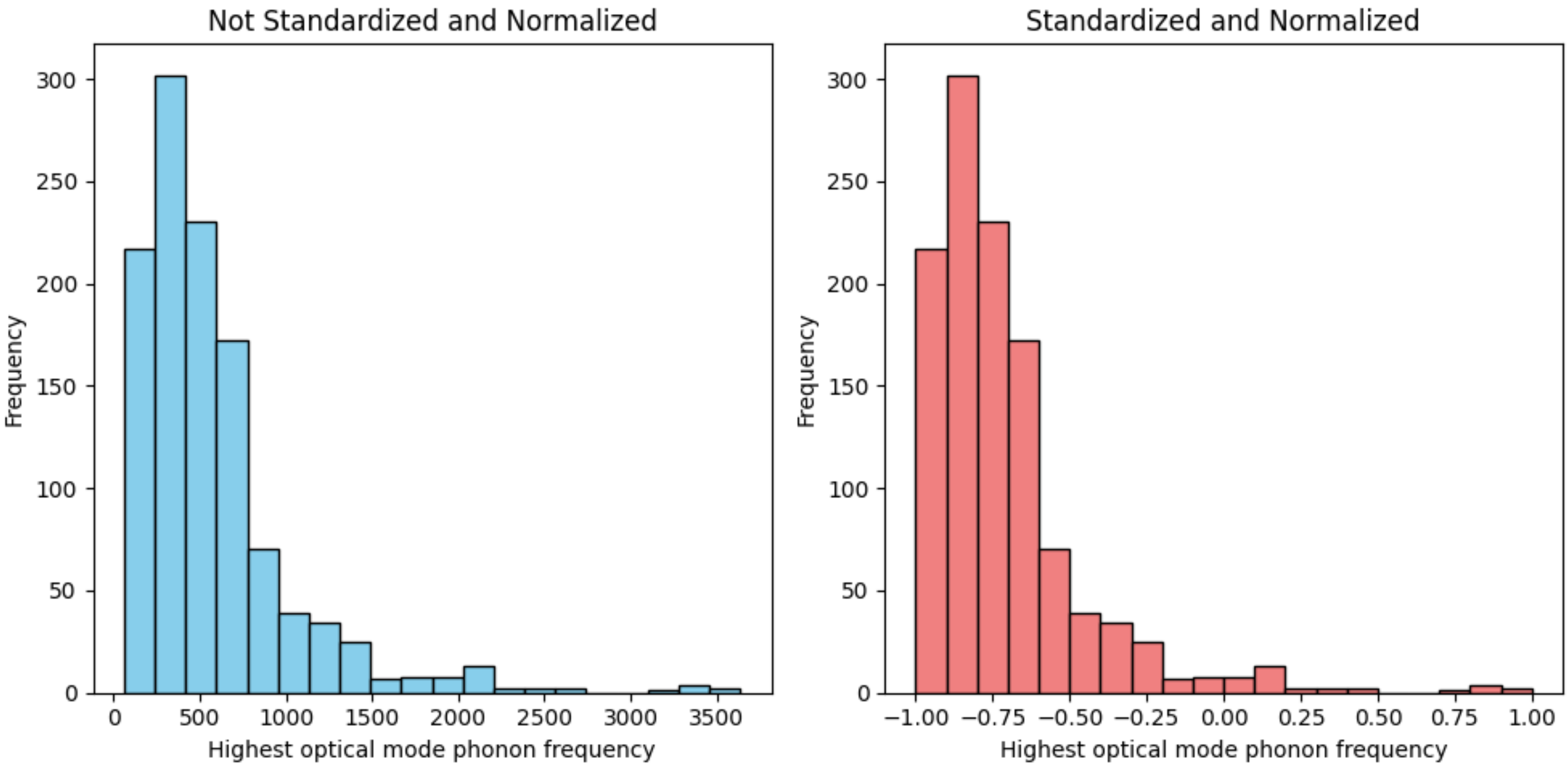
On average, MPT is 6% and 10% better on R^2 and MAE than PT-FT
Closest performer to MPT is FE: largest dataset within MPT
MPT models: may generalize quite well with more properties

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

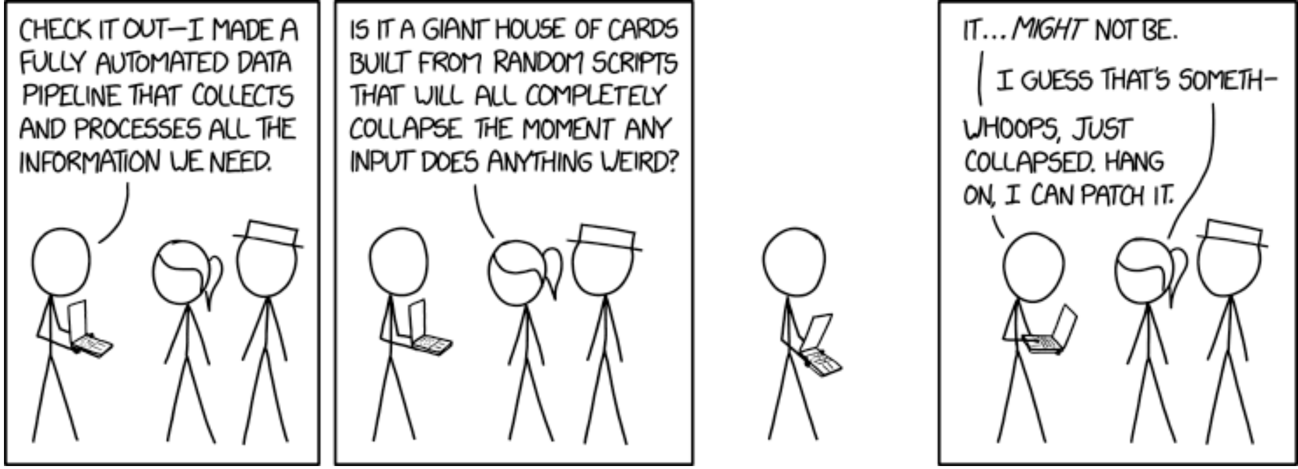
1. Choudhary et al. **npj Comput. Mater.** *6*, 173 (2020).

Hands-on session?

Predict phonon modes using scratch and fine-tuned models



Summary



- Materials science is limited by data availability on key properties
  - Transfer learning as a path to build robust models

- Optimal PT-FT strategies
  - Larger PT/FT dataset generally helps
  - Specific PT property: weak handle
  - More degrees of freedom in model: better

- MPT: a path to generalized models
  - On-average better than scratch and best PT-FT
  - Generalizes well out-of-distribution

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