Utility of transfer learning in computational materials science

Reshma Devi, Keith Butler, Sai Gautam Gopalakrishnan

Materials Engineering, Indian Institute of Science

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

Materials in an explosively growing informatics world (CIMTEC) 2024, Montecatini Terme, Italy
Jun 21, 2024

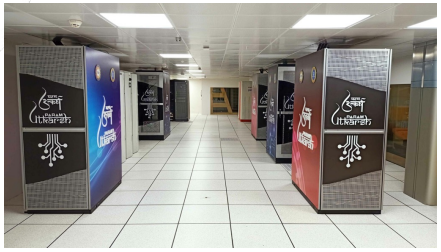
Acknowledgments

Group picture in Jun 2024

Dr. Keith Butler

Reshma

Archer (UK)



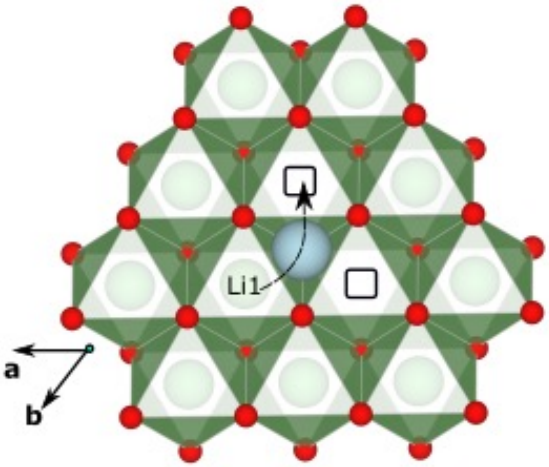
Param Utkarsh (CDAC)

SERC (IISc)

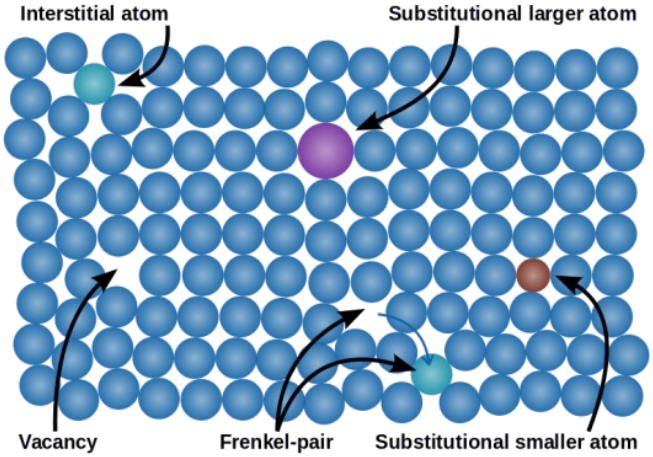
Materials science is data limited

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

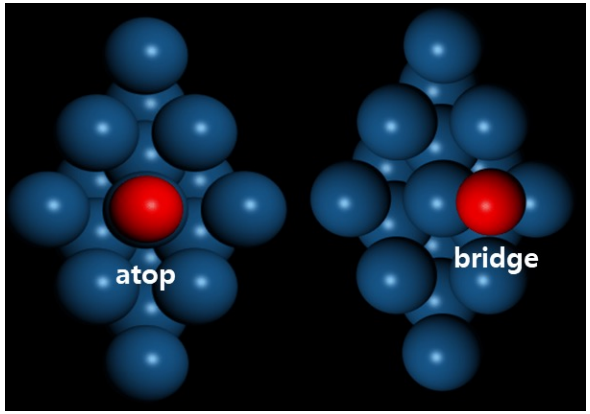


Devi et al., *npj Comput. Mater.* 2022



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

https://sites.psu.edu/



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

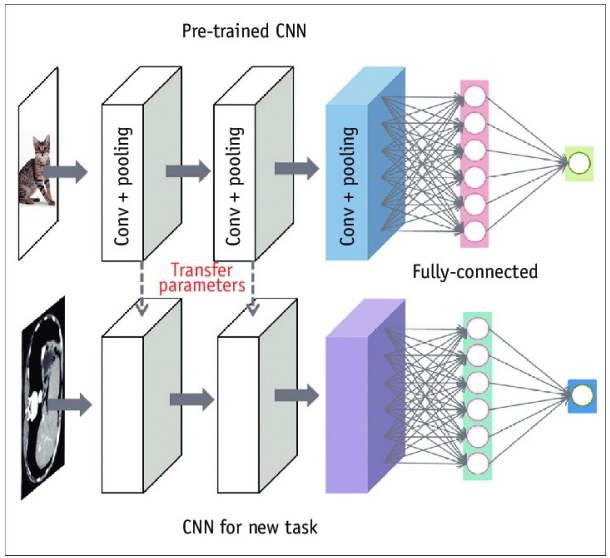
Materials science is data limited

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



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

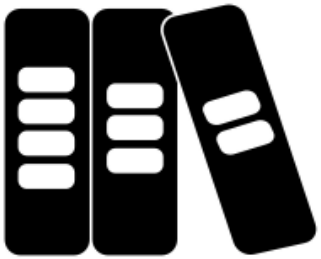
Transfer learning: efficiently use DL frameworks on small datasets

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

How useful is transfer learning in materials science?

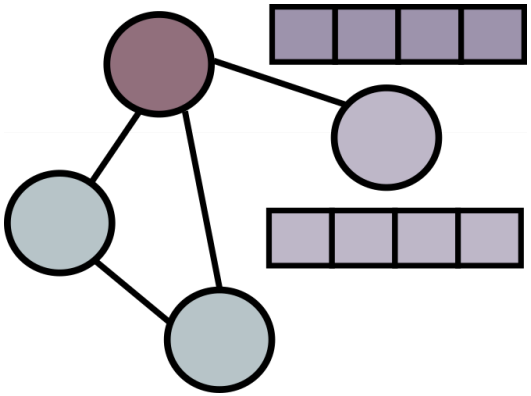
- Optimal ways to use?
- Ways to generate ‘generalized’ models?

Handles to consider



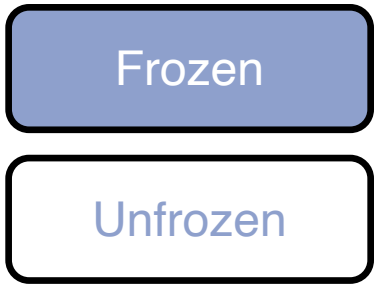
Dataset(s)

- What, how, how many?



Architecture

- Graph neural network

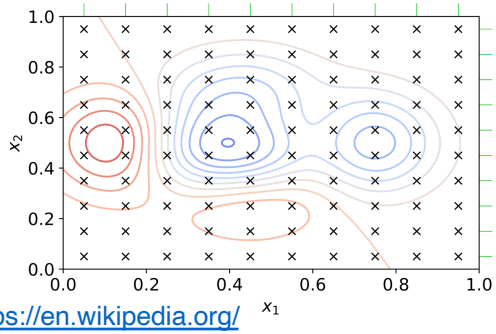


Strategy

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

(Learning) Hyperparameters

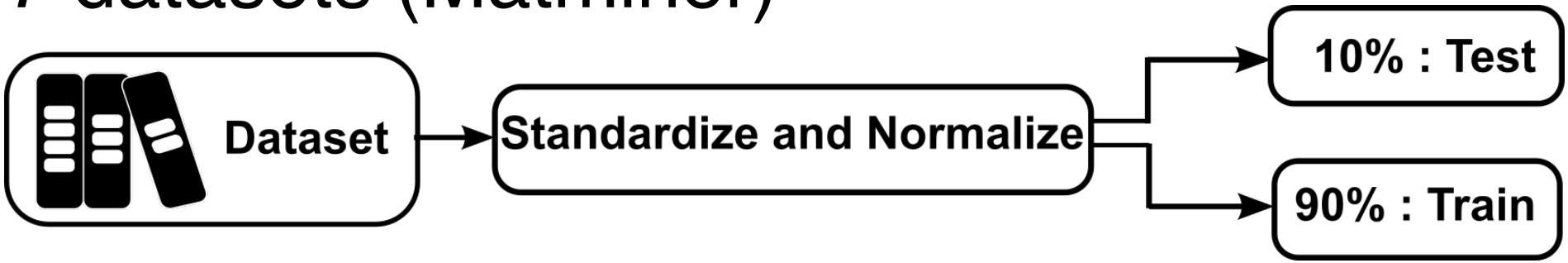
- Data sampling
- Learning rate
- Number of datapoints during PT, FT



https://en.wikipedia.org/

Handles

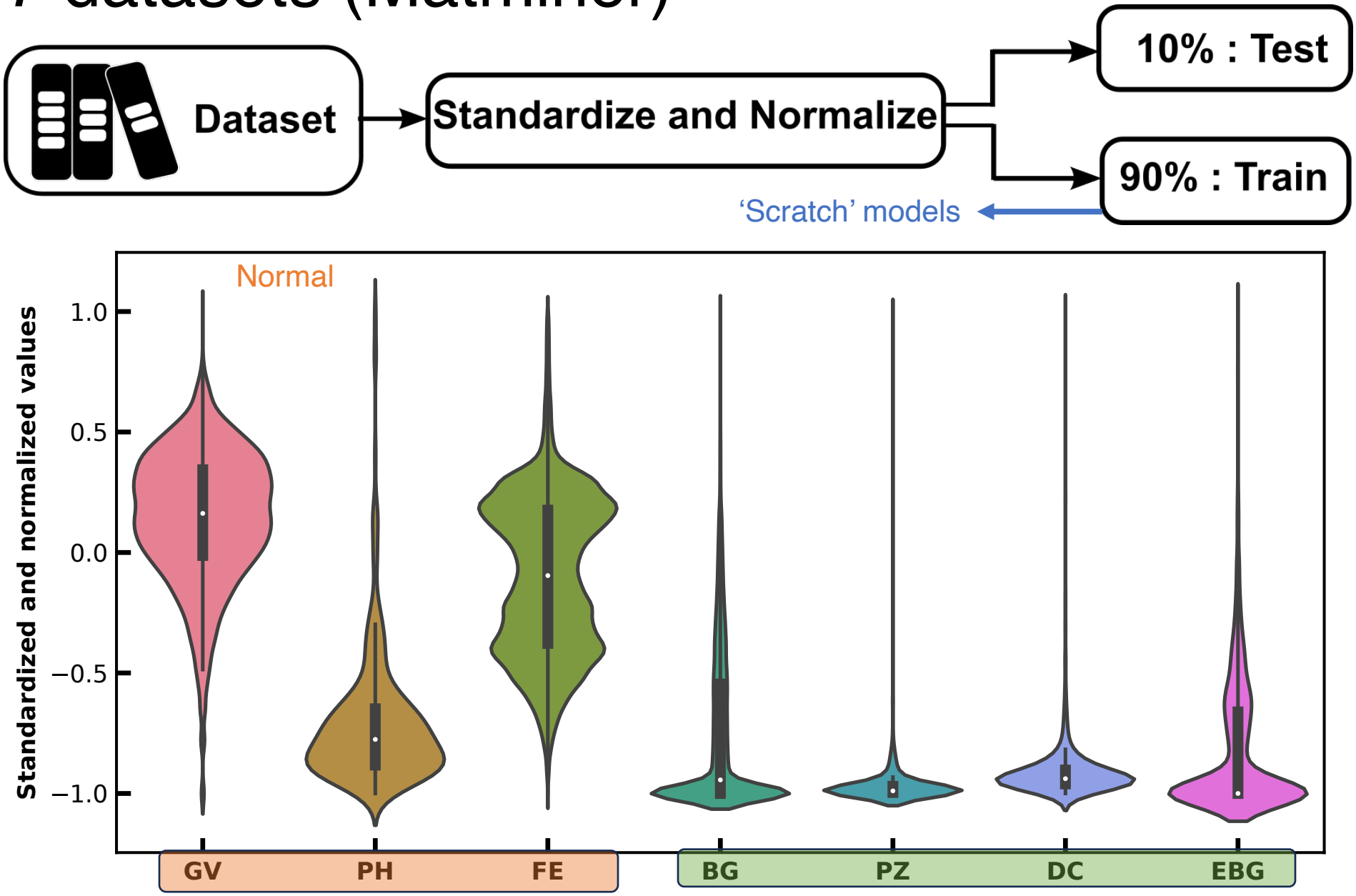
7 datasets (Matminer)



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

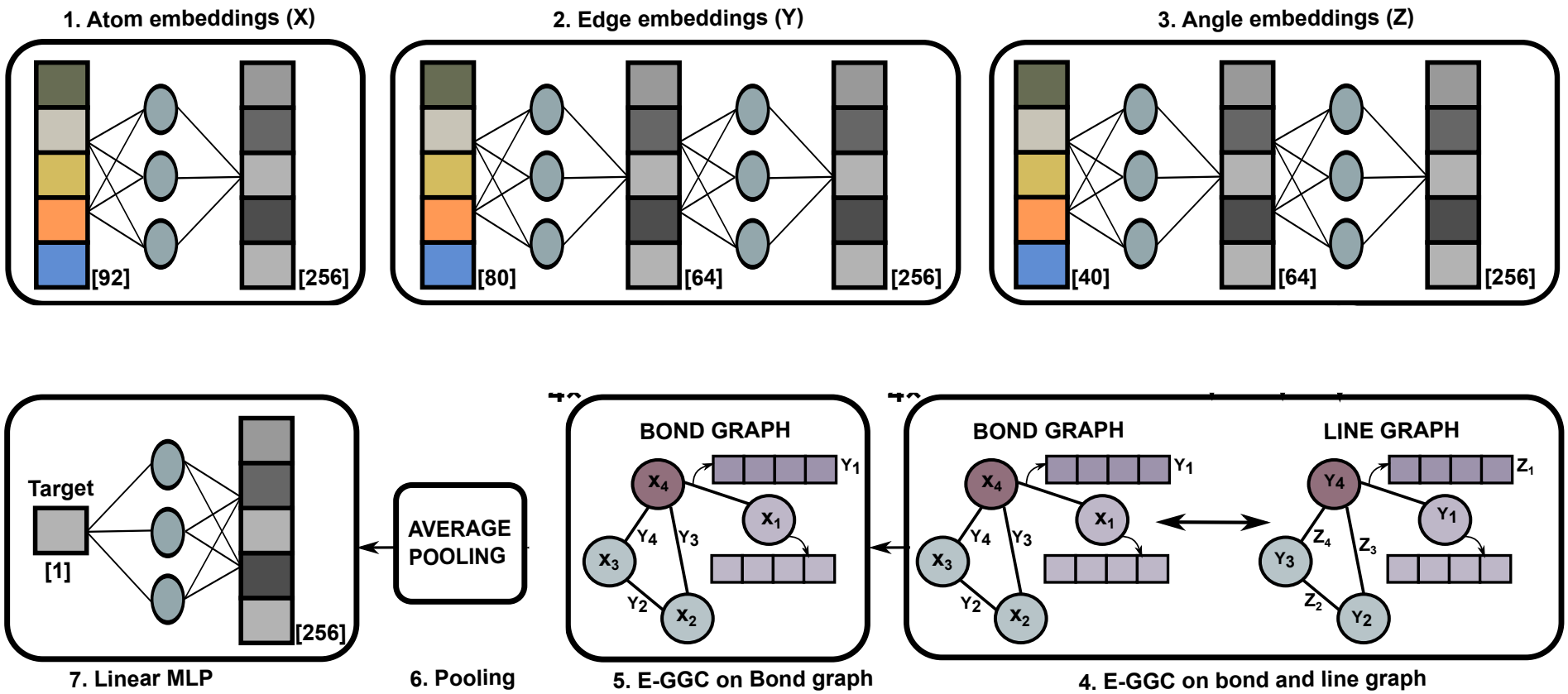
1. Ward *et al., Comput. Mater. Sci.* 152, 60–69 (2018).

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]

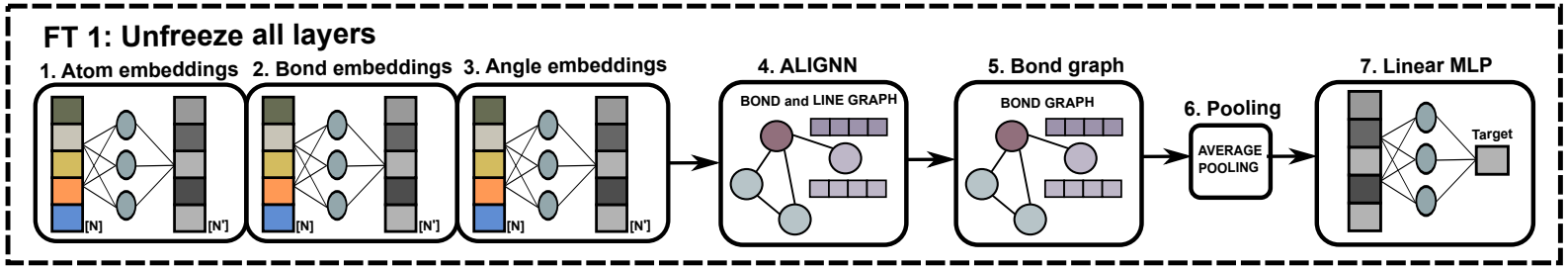
1. Choudhary and DeCost, *npj Comput. Mater.* 7, 185 (2021).

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

FT strategies

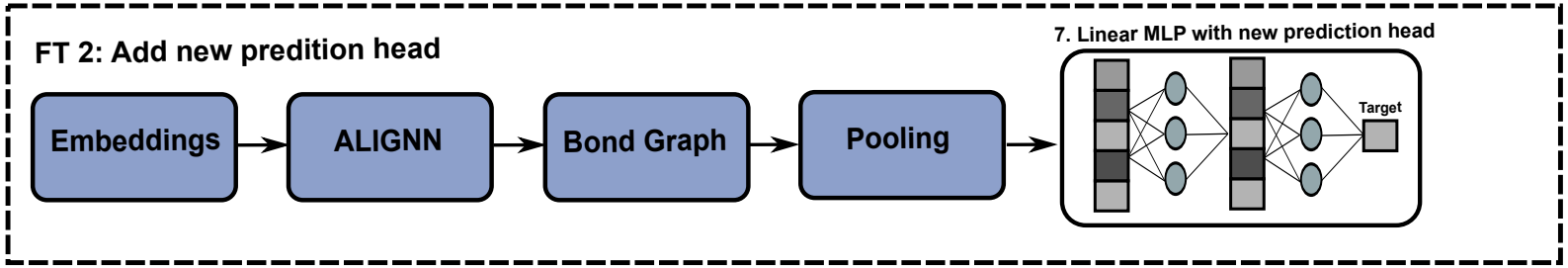
Frozen

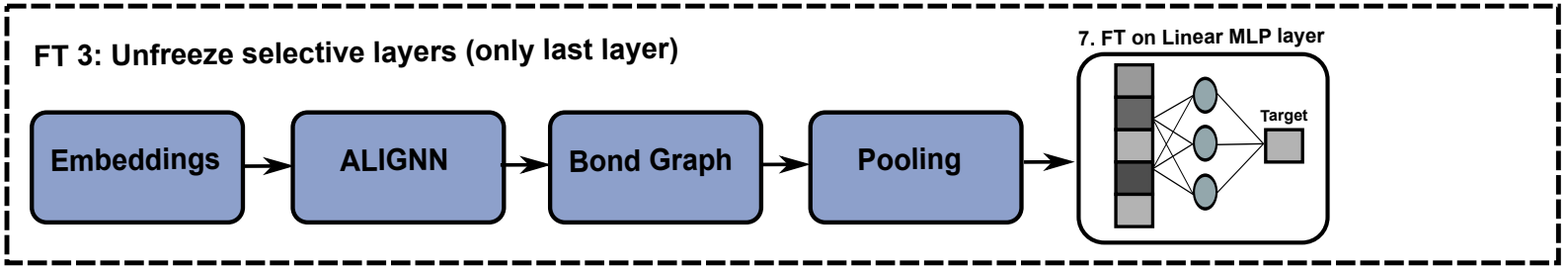
Unfrozen



Highest degree of freedom

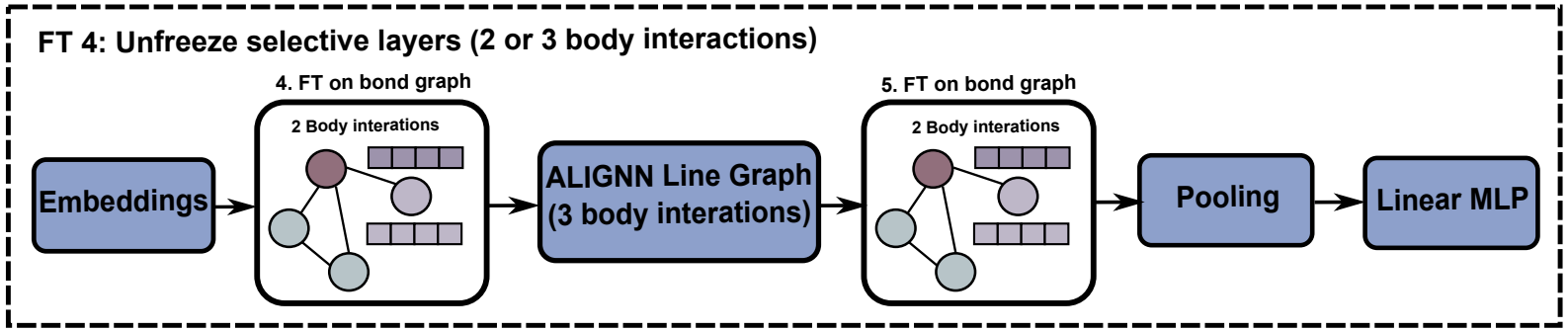
Additional flexibility at head





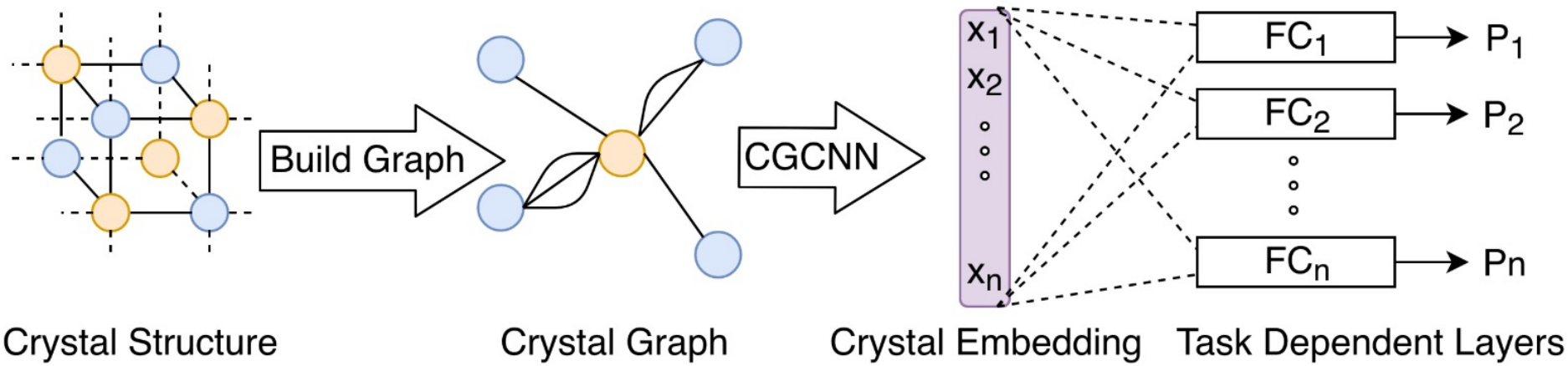
Classic; lowest degree of freedom

2-body and 3-body layers: central to ALIGNN



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

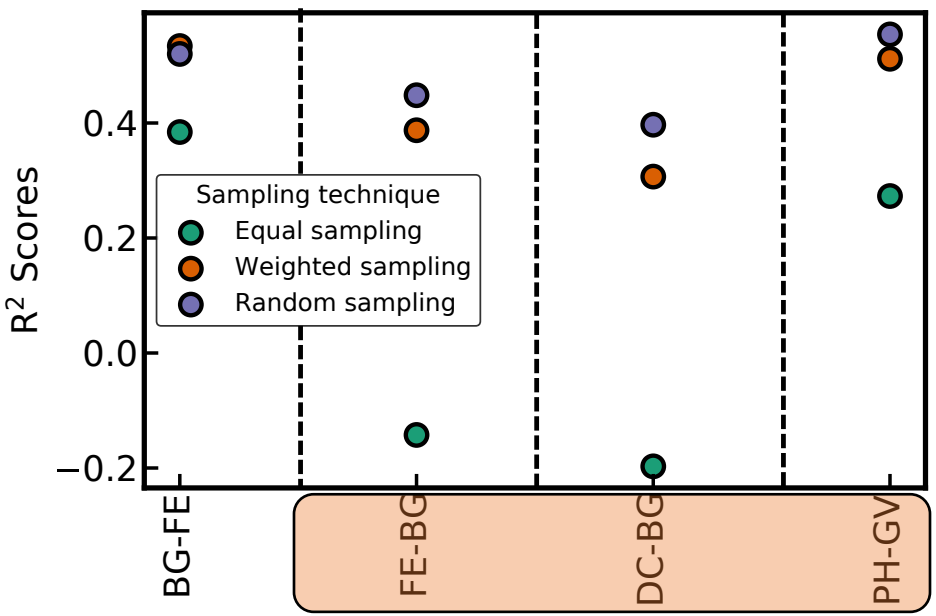
$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} |y_p^i - y_t^i| \delta^i$$

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

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

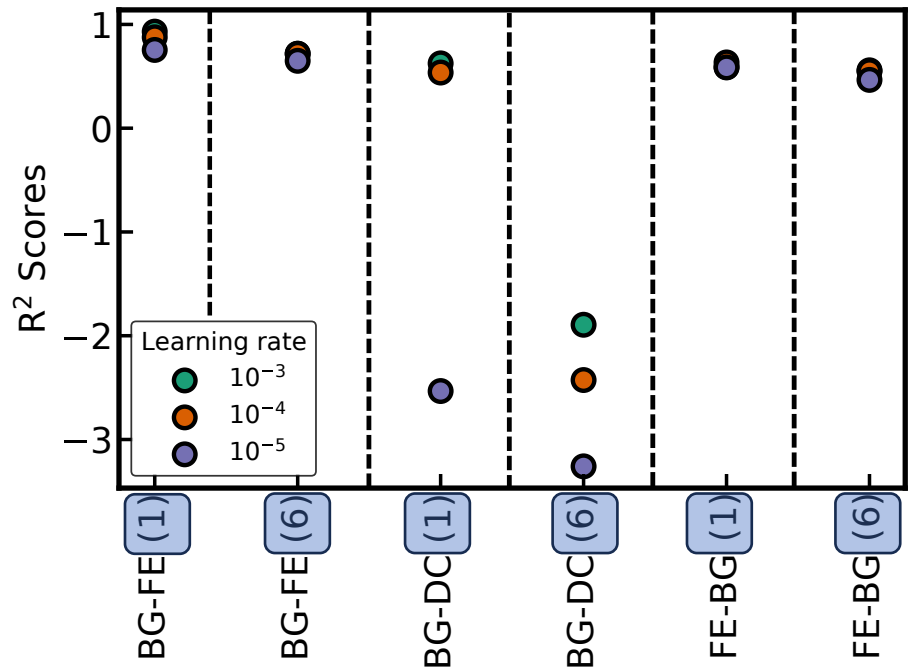
Hyperparameters

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

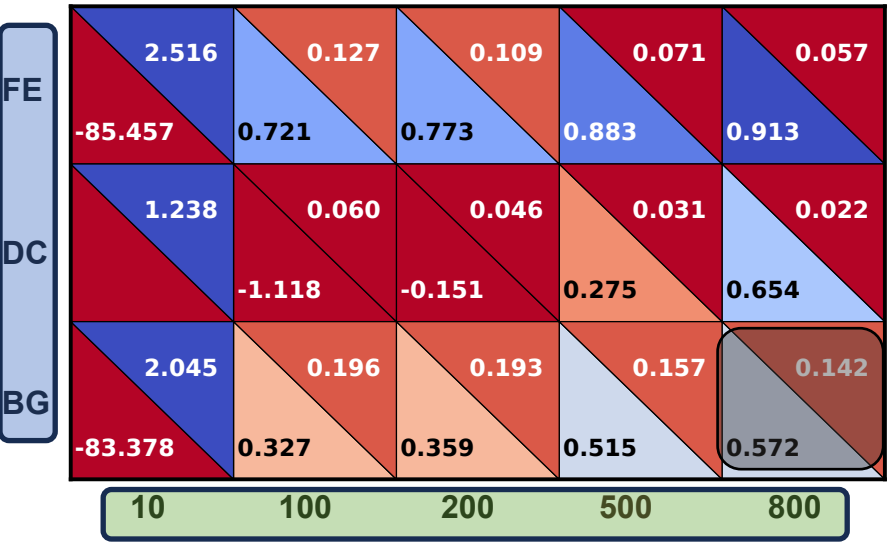
10^{-3} optimal; validation losses high at 10^{-2}

Let's look at model performances in more detail

Results

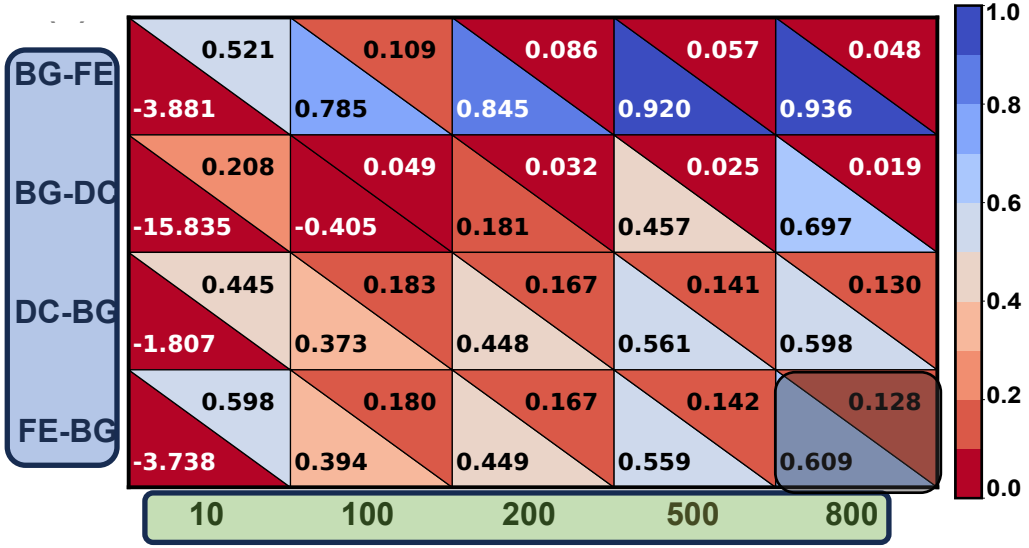
- Dataset size influence
- FT Strategy
- 7×6 Pair-wise models
- MPT framework

More FT data: better



Scratch models

Dataset name; Dataset size



Pair-wise PT-FT models

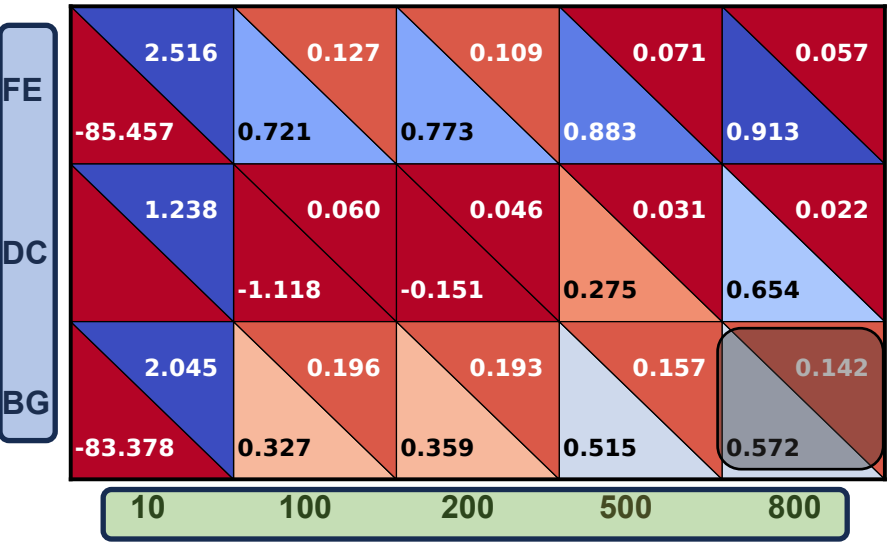
PT-FT dataset name; FT Dataset size

Test scores (5 trials)

- PT-FT models are consistently better than scratch models
- More FT data: better models (both scratch and PT-FT)

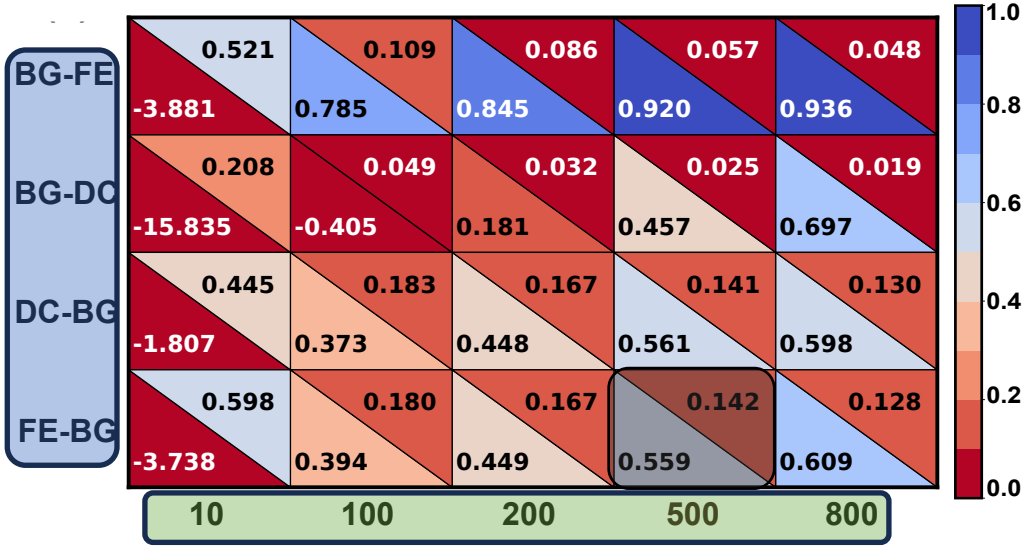
BG: Band gap; FE: Formation energy; DC: Dielectric constant; MAE: Mean absolute error

More FT data: better



Scratch models

Dataset name; Dataset size



Pair-wise PT-FT models

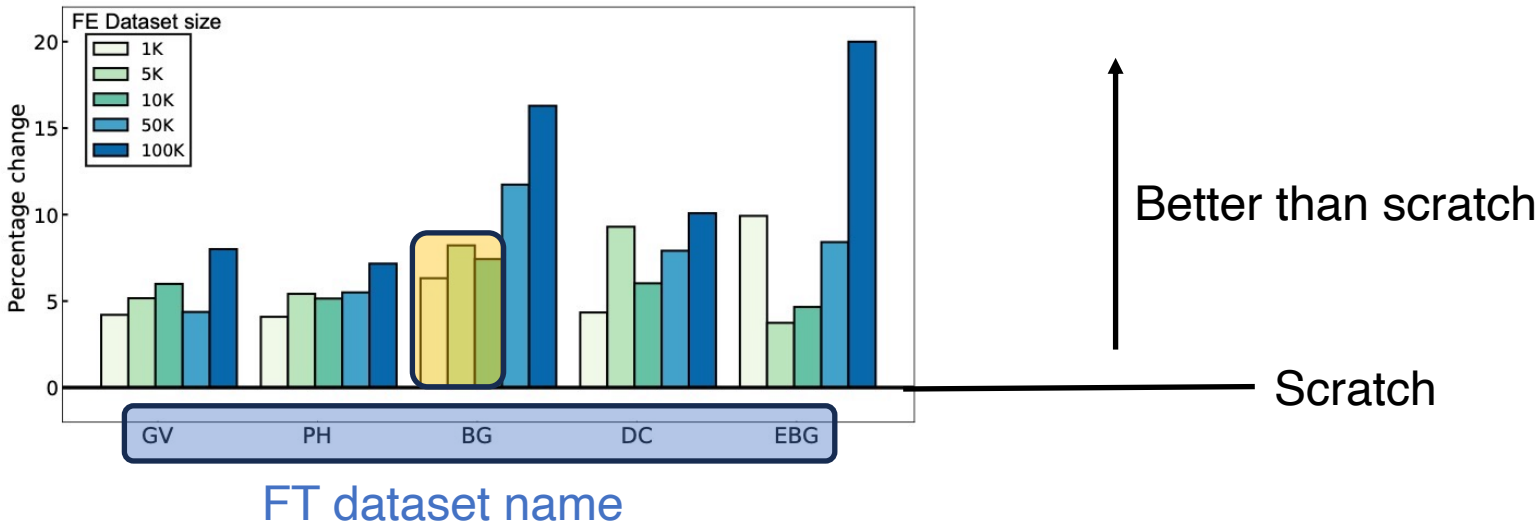
PT-FT dataset name; FT Dataset size

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

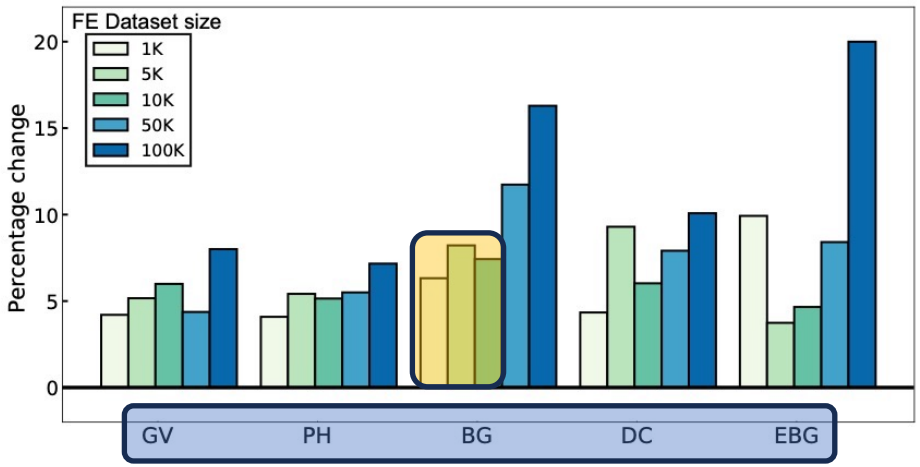


Formation energy as PT

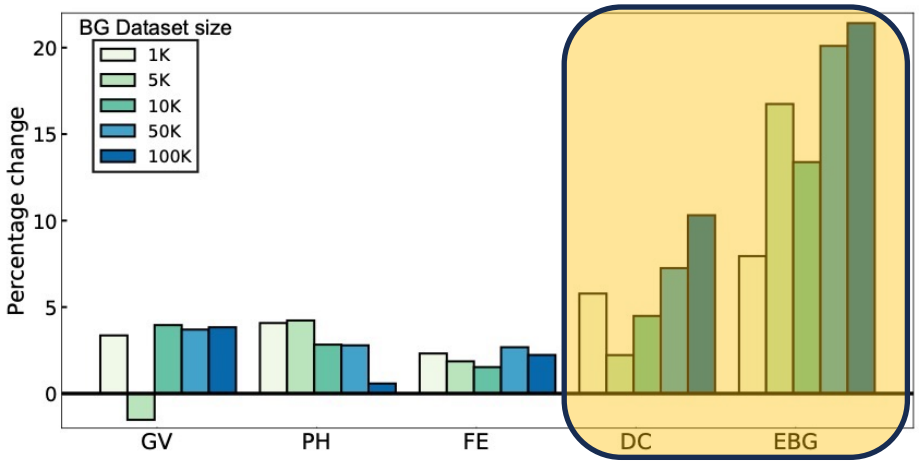
- Increasing dataset size: non-monotonicity
- Best models at 100K
- Always better than scratch

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

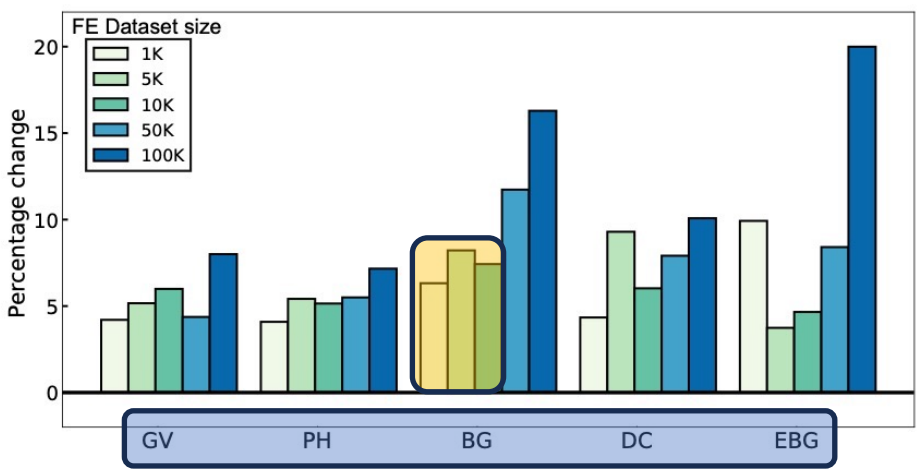
- Increasing dataset size: non-monotonicity
- Best models at 100K
- Always better than scratch

Band gap as PT

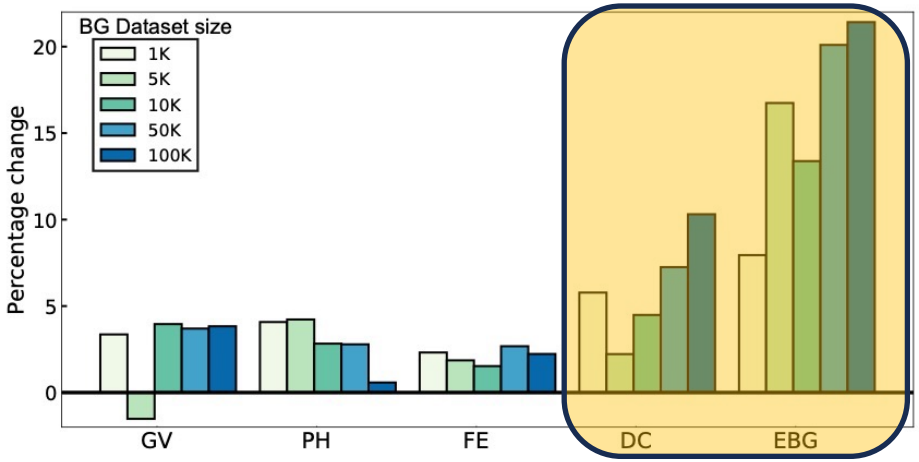
- Non-monotonicity
- Best models at 50K for non-correlated; 100K for correlated
- (Almost) always better than scratch

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

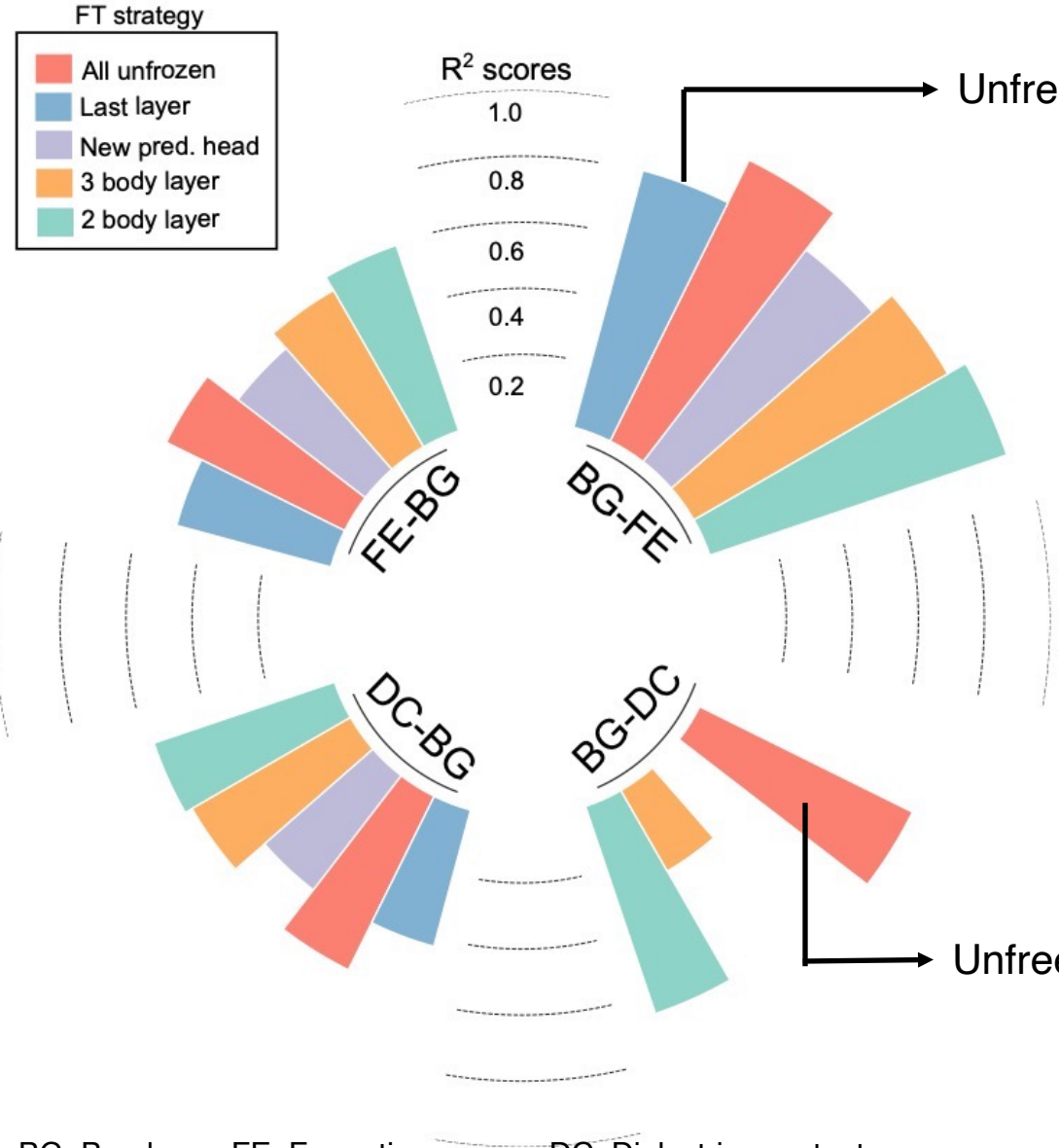


Larger PT data: generally better despite non-monotonic improvement

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

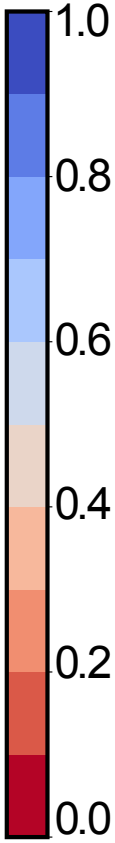
Creating foundational models: difficult in pair-wise

Unfreezing all layers: maximal degree of freedom

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

7×6 combinations of pair-wise models

	GV	PH	FE	BG	PZ	DC	EBG
800	0.096 / 0.742	0.043 / 0.904	0.057 / 0.913	0.142 / 0.572	0.040 / -0.086	0.022 / 0.654	0.104 / 0.415
EBG	0.090 / 0.766	0.032 / 0.940	0.049 / 0.934	0.126 / 0.619	0.032 / 0.030	0.020 / 0.681	
DC	0.090 / 0.768	0.032 / 0.940	0.048 / 0.936	0.130 / 0.598	0.032 / 0.007		0.092 / 0.449
PZ	0.089 / 0.772	0.030 / 0.947	0.048 / 0.938	0.128 / 0.607		0.019 / 0.681	0.093 / 0.436
BG	0.090 / 0.768	0.030 / 0.946	0.048 / 0.936		0.031 / 0.023	0.019 / 0.697	0.093 / 0.452
FE	0.089 / 0.774	0.031 / 0.946		0.128 / 0.609	0.031 / 0.033	0.019 / 0.677	0.092 / 0.459
PH	0.089 / 0.774		0.047 / 0.938	0.128 / 0.611	0.032 / 0.028	0.020 / 0.662	0.093 / 0.442
GV		0.031 / 0.947	0.048 / 0.936	0.129 / 0.609	0.031 / 0.025	0.019 / 0.687	0.091 / 0.461
	GV800	PH800	FE800	BG800	PZ800	DC800	EBG800



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

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

	GV	PH	FE	BG	PZ	DC	EBG
800	0.096 / 0.742	0.043 / 0.904	0.057 / 0.913	0.142 / 0.572	0.040 / -0.086	0.022 / 0.654	0.104 / 0.415
EBG	0.090 / 0.766	0.032 / 0.940	0.049 / 0.934	0.126 / 0.619	0.032 / 0.030	0.020 / 0.681	
DC	0.090 / 0.768	0.032 / 0.940	0.048 / 0.936	0.130 / 0.598	0.032 / 0.007		0.092 / 0.449
PZ	0.089 / 0.772	0.030 / 0.947	0.048 / 0.938	0.128 / 0.607		0.019 / 0.681	0.093 / 0.436
BG	0.090 / 0.768	0.030 / 0.946	0.048 / 0.936		0.031 / 0.023	0.019 / 0.697	0.093 / 0.452
FE	0.089 / 0.774	0.031 / 0.946		0.128 / 0.609	0.031 / 0.033	0.019 / 0.677	0.092 / 0.459
PH	0.089 / 0.774		0.047 / 0.938	0.128 / 0.611	0.032 / 0.028	0.020 / 0.662	0.093 / 0.442
GV		0.031 / 0.947	0.048 / 0.936	0.129 / 0.609	0.031 / 0.025	0.019 / 0.687	0.091 / 0.461

GV800 PH800 FE800 BG800 PZ800 DC800 EBG800 — FT dataset+size

1.0 0.8 0.6 0.4 0.2 0.0

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

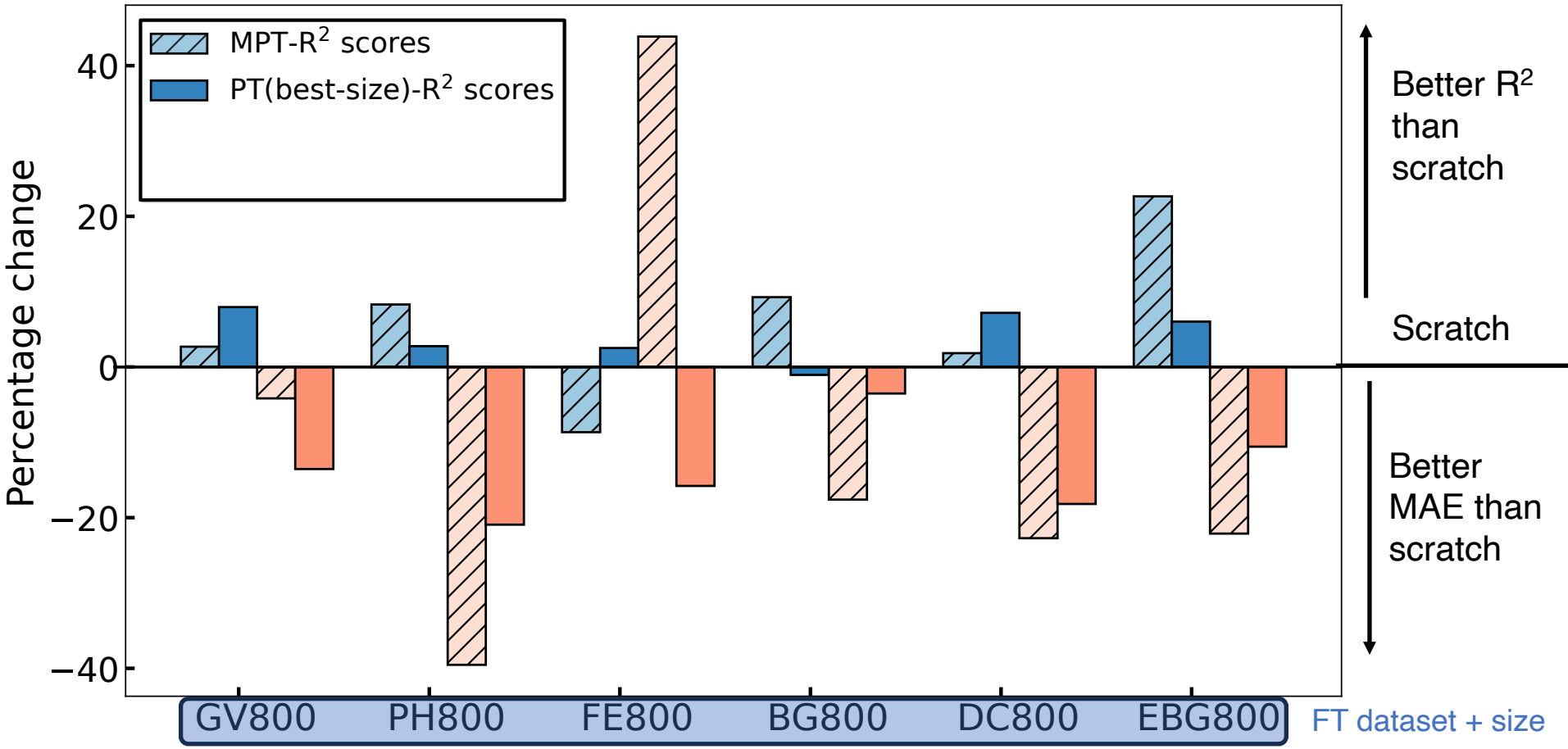
No symmetry

Test scores

MAE

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

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-distribution 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

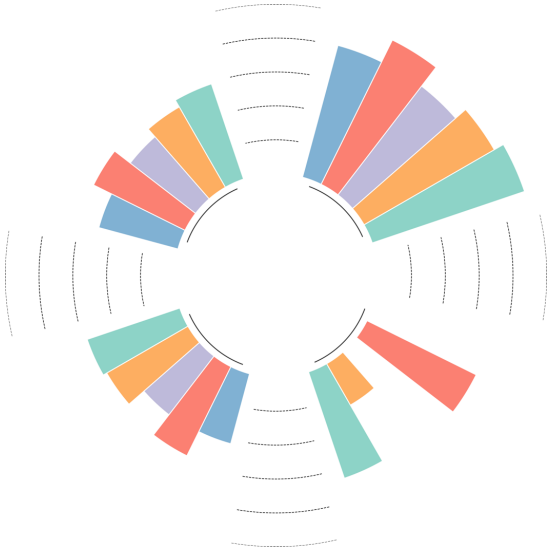
On average, MPT is 6% and 10% better on R^2 and MAE than PT-FT
Closest performer to MPT is FE: largest dataset within MPT
MPT models: may generalize quite well with more properties

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

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

Conclusions

- 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



"Optimal pre-train/fine-tune strategies for accurate material property predictions ", R. Devi, K.T. Butler, and G. Sai Gautam, **arXiv 2406.13142** (2024). *Under review*

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