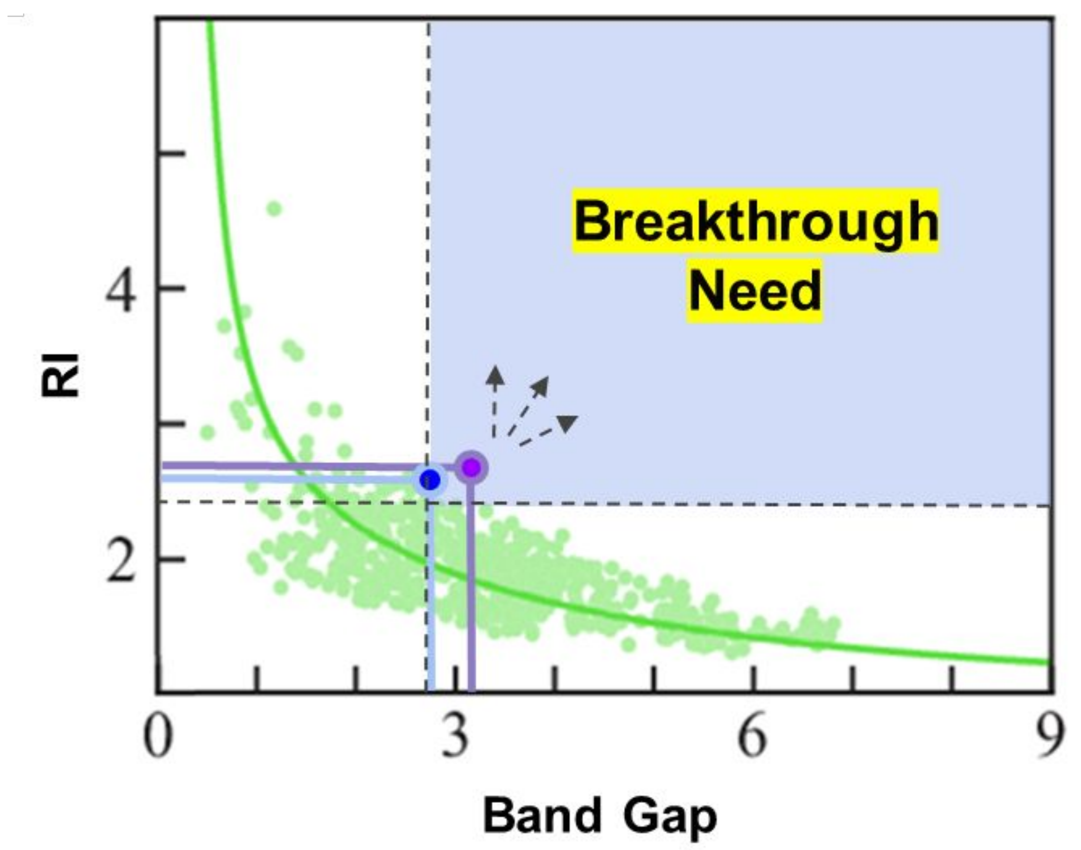
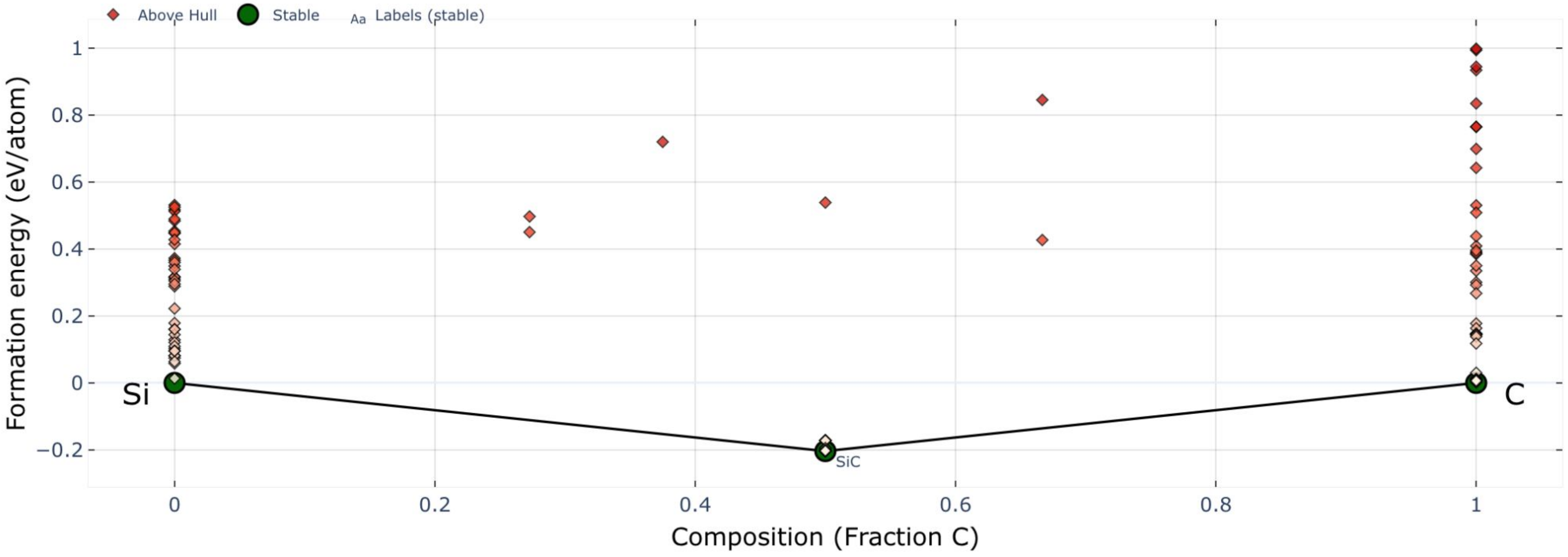
Exploring the Frontier of Universal Machine Learning Potentials

Part 1: Insights from OMat24 and eSEN

Luis Barroso-Luque
FAIR at Meta

Why does Meta care about materials?





Orion AR Headset (SiC)

Refractive Index ~ 2.6
70 deg Field of View

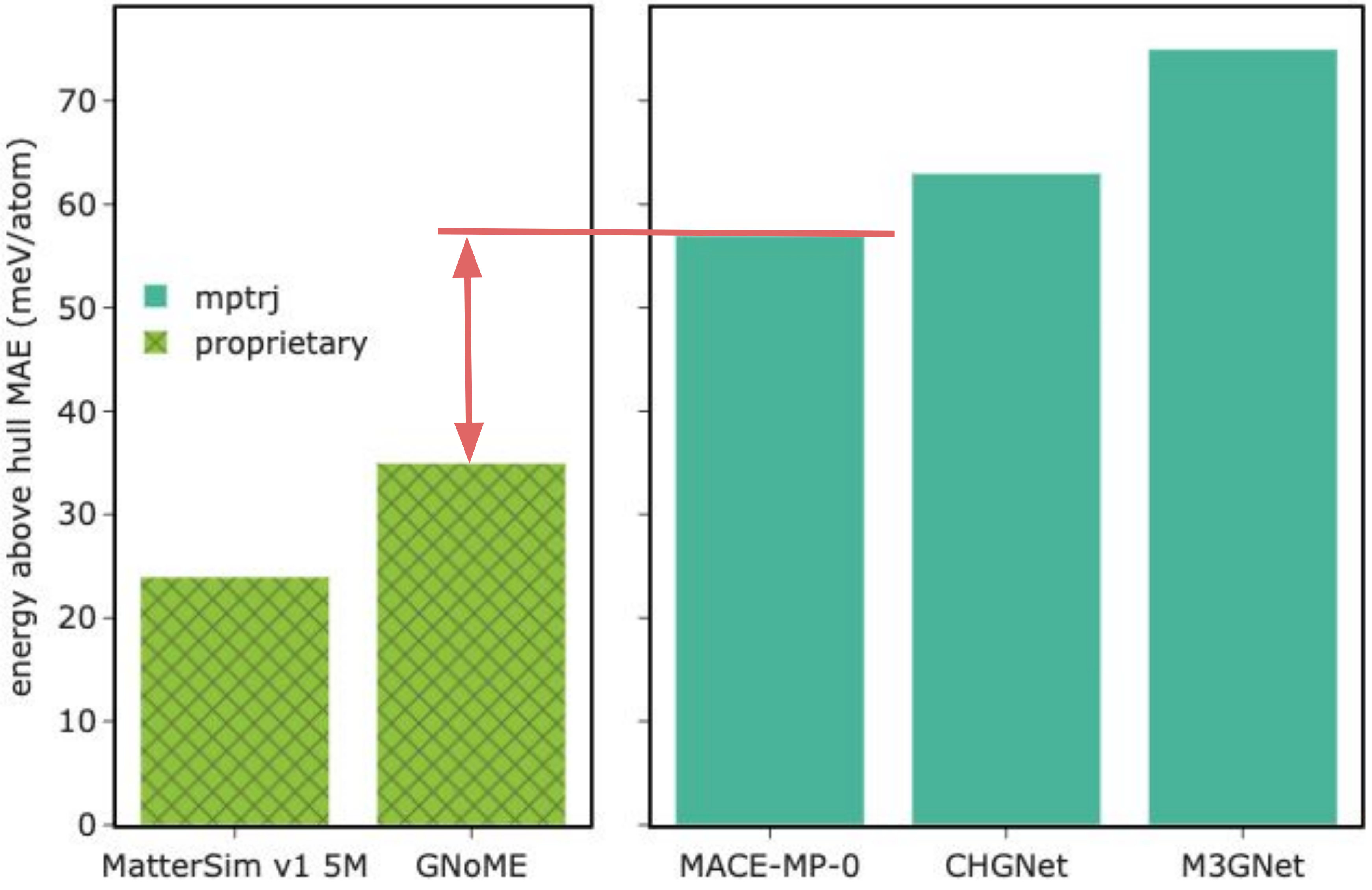
https://about.meta.com/realitylabs/orion/silicon-carbide/

A need for a large DFT training dataset of inorganic materials

Small open datasets for training inorganic materials MLIPs

Dataset	Size (No Structures)
MPTrj	1.6M
GNoME	89M
Mattersim	17M

Accuracy gap between models trained on closed data and open models



But we already had a lot of open data!

The composition space of inorganic bulk materials is well covered → Many open databases!

OQMD The Open Quantum Materials Database

MATERIALSCLOUD

The Materials Project

AFLOW

NOMAD NOVEL MATERIALS DISCOVERY

Alexandria

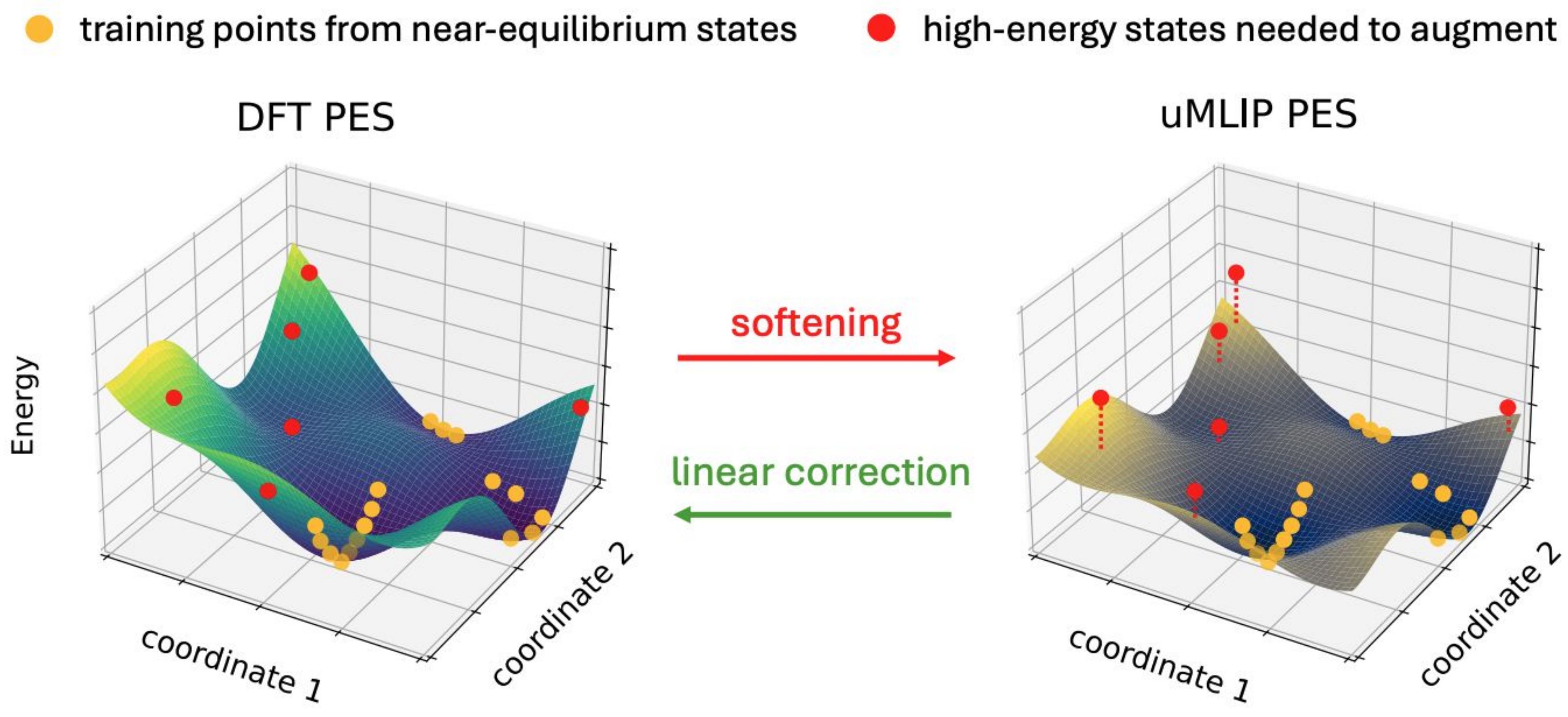
Alexandria PBE bulks

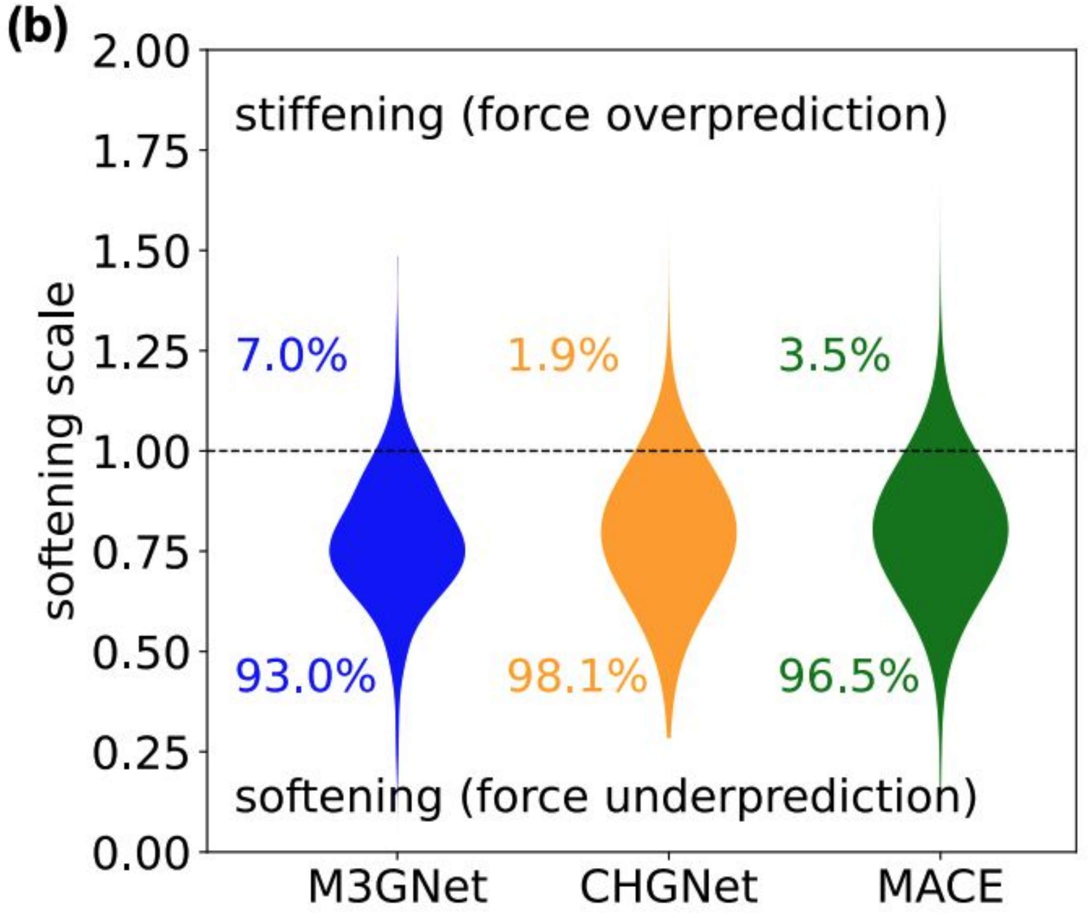
- 4.5M materials
- 2.6M chemical compositions
- 1.7K elementary compounds
- 241K binary compounds
- 3M ternary compounds
- 1.3M quaternary compounds
- 15K >4 multinary compounds

https://doi.org/10.1016/j.mtphys.2024.101560

Limitations of near equilibrium data

- Near equilibrium relaxation data is necessary to find stable materials
- Biases MLIP training and limits generalization
  - Underprediction of forces, phonons, defect energies, activation barriers





arXiv > cond-mat > arXiv:2405.07105

Condensed Matter > Materials Science

[Submitted on 11 May 2024]

Overcoming systematic softening in universal machine learning interatomic potentials by fine-tuning

Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A. Persson, Gerbrand Ceder

Machine learning interatomic potentials (MLIPs) have introduced a new paradigm for atomic simulations. Recent advancements have seen the emergence of universal MLIPs (uMLIPs) that are pre-trained on diverse materials datasets, providing opportunities for both ready-to-use universal force fields and robust foundations for downstream machine learning refinements. However, their performance in extrapolating to out-of-distribution complex atomic enviror

https://arxiv.org/abs/2405.07105

Open Materials 2024 (OMat24)

Open source models + data @

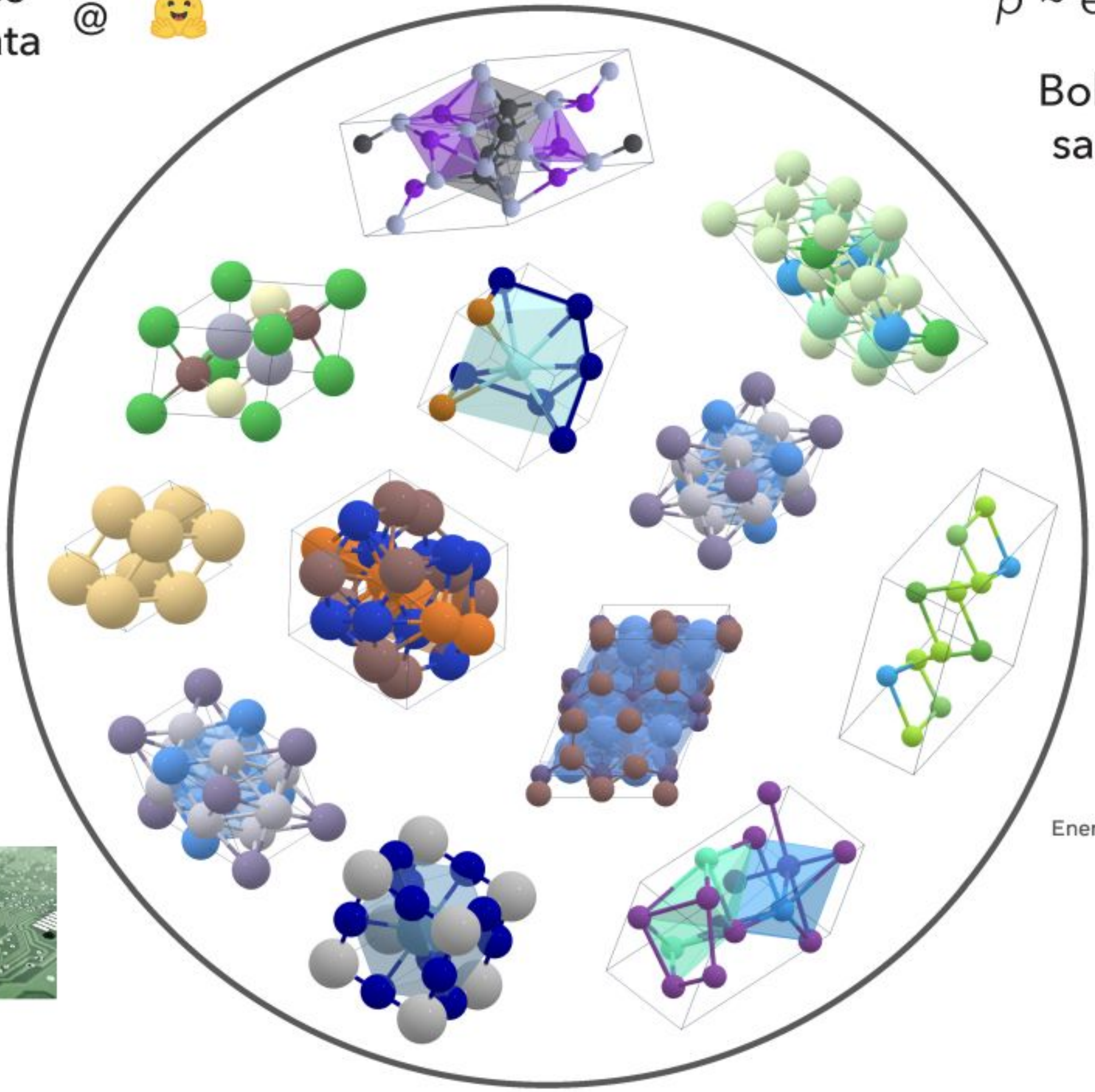
100M+ non-equilibrium structures

Applications:

Renewable energy storage & CO_2 reduction

Optics & Electronics

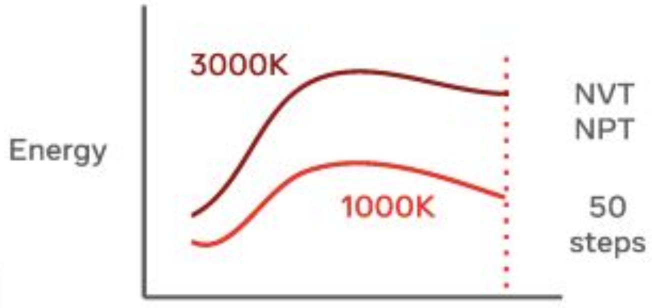
... and much more!



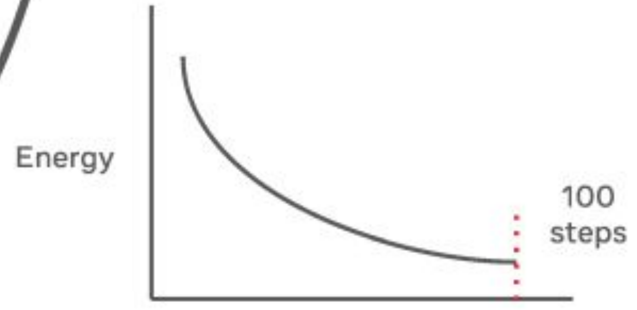
$p \sim \exp(-e/kT)$

Boltzmann sampling

300K
500K
1000K

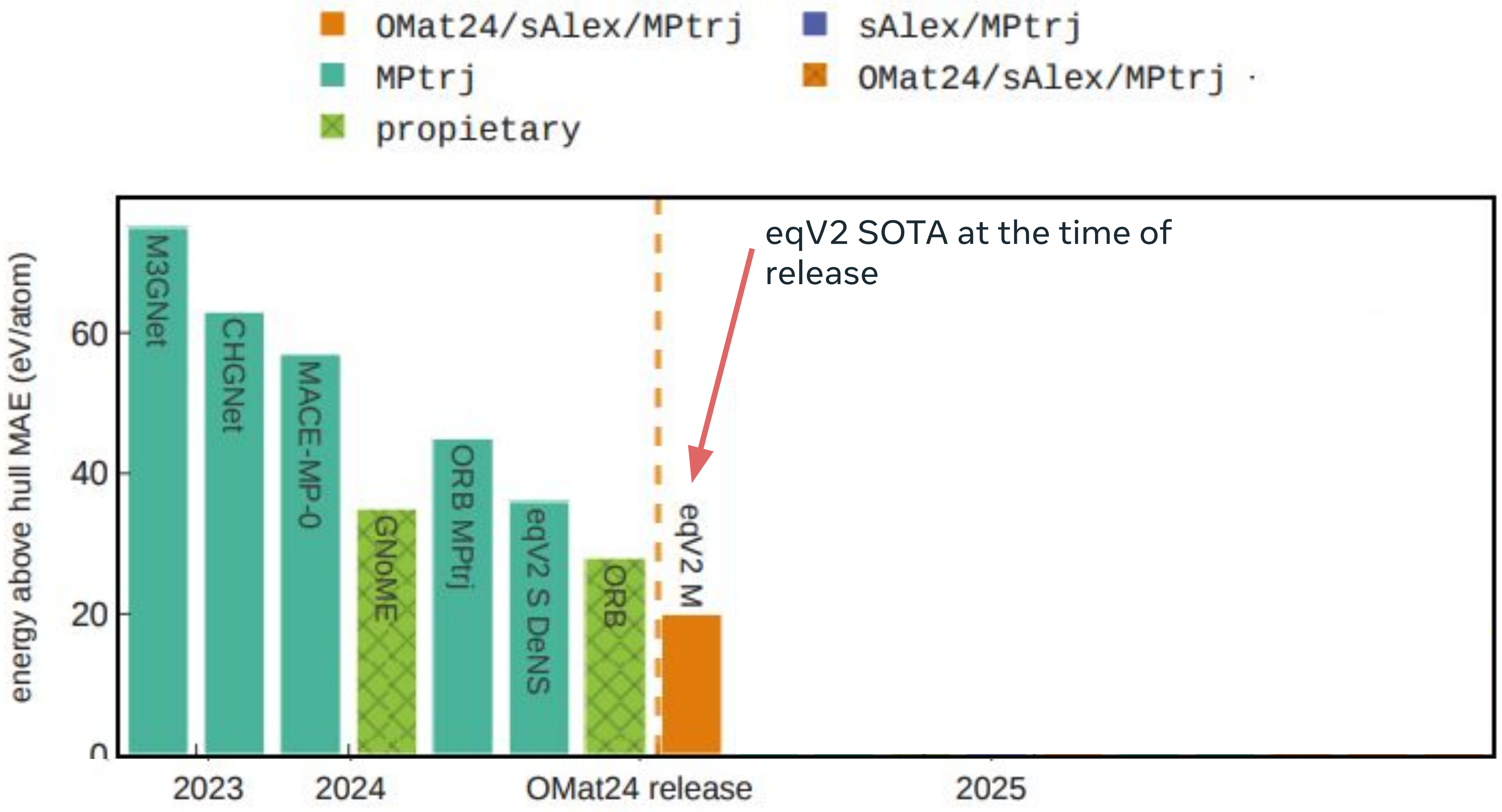


AIMD



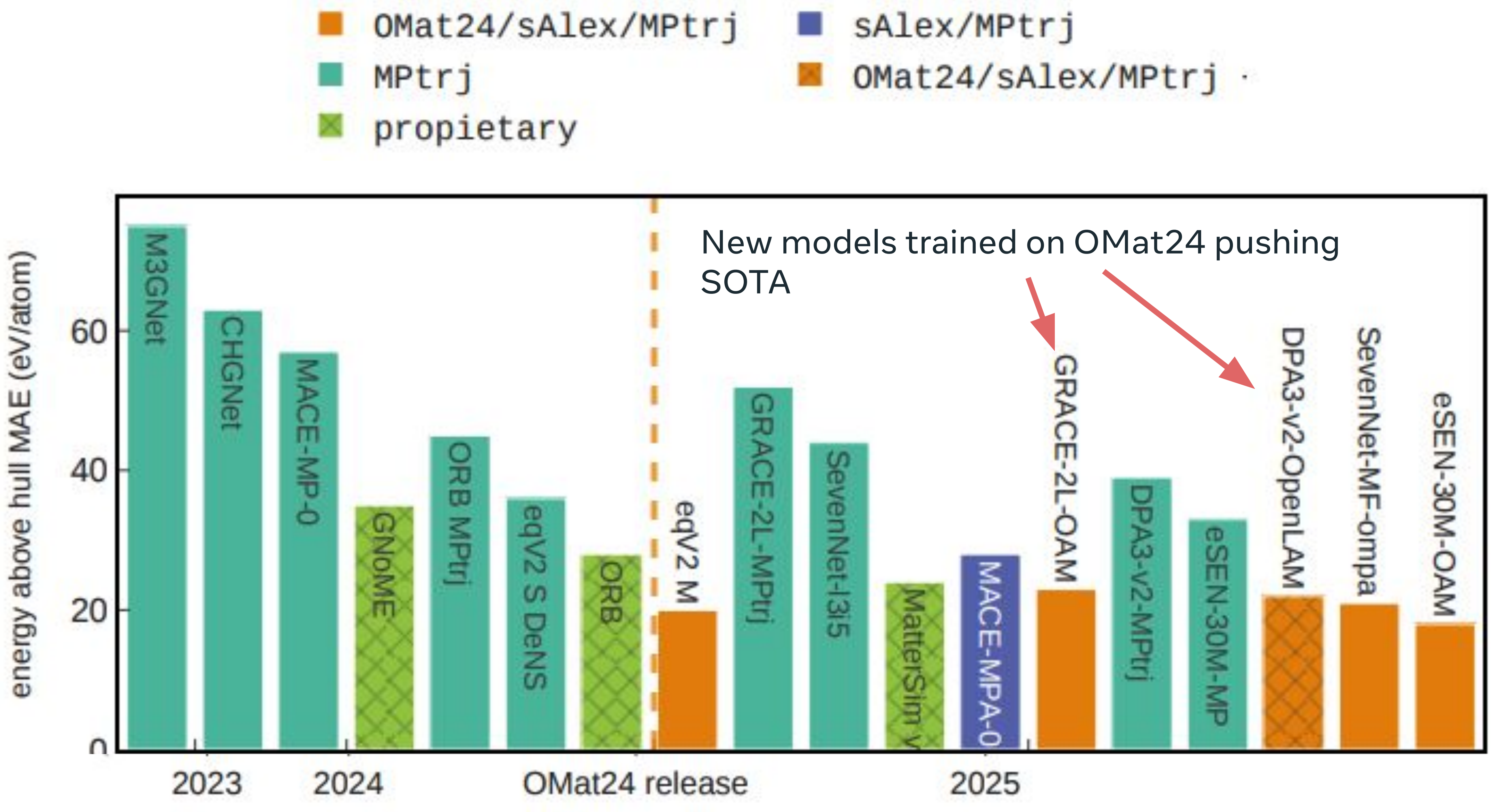
Rattled relaxation

OMat24 set a new SOTA for formation energy accuracy



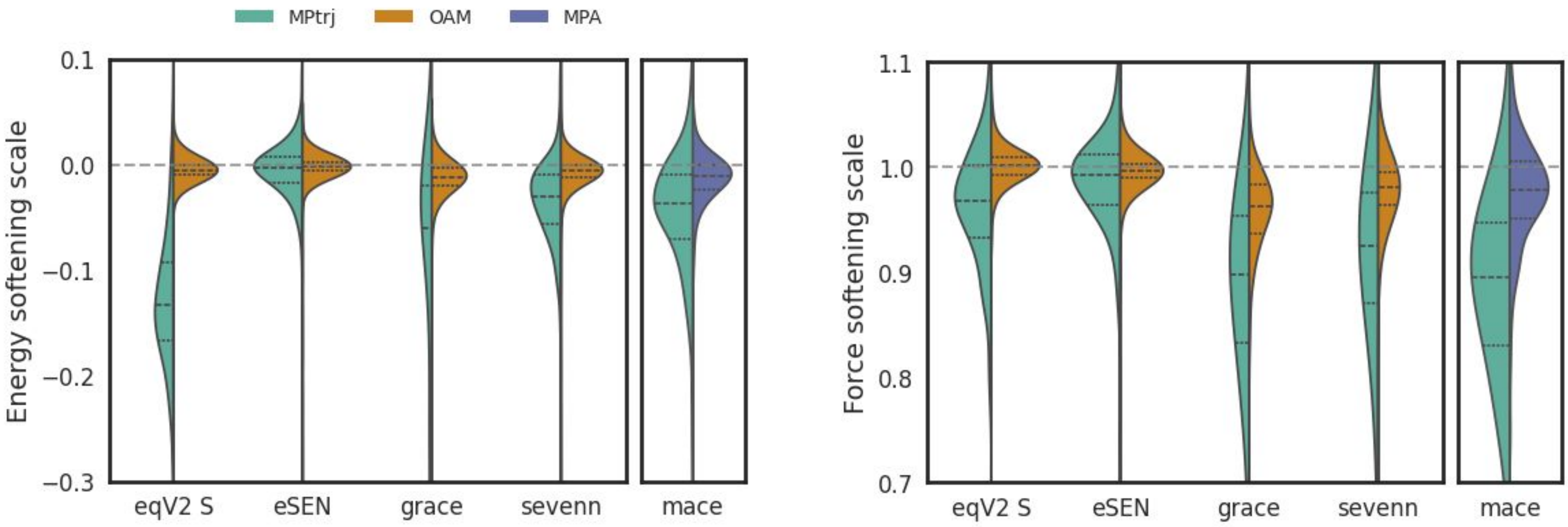
https://matbench-discovery.materialsproject.org/

OMat24 set a new SOTA for energy accuracy



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

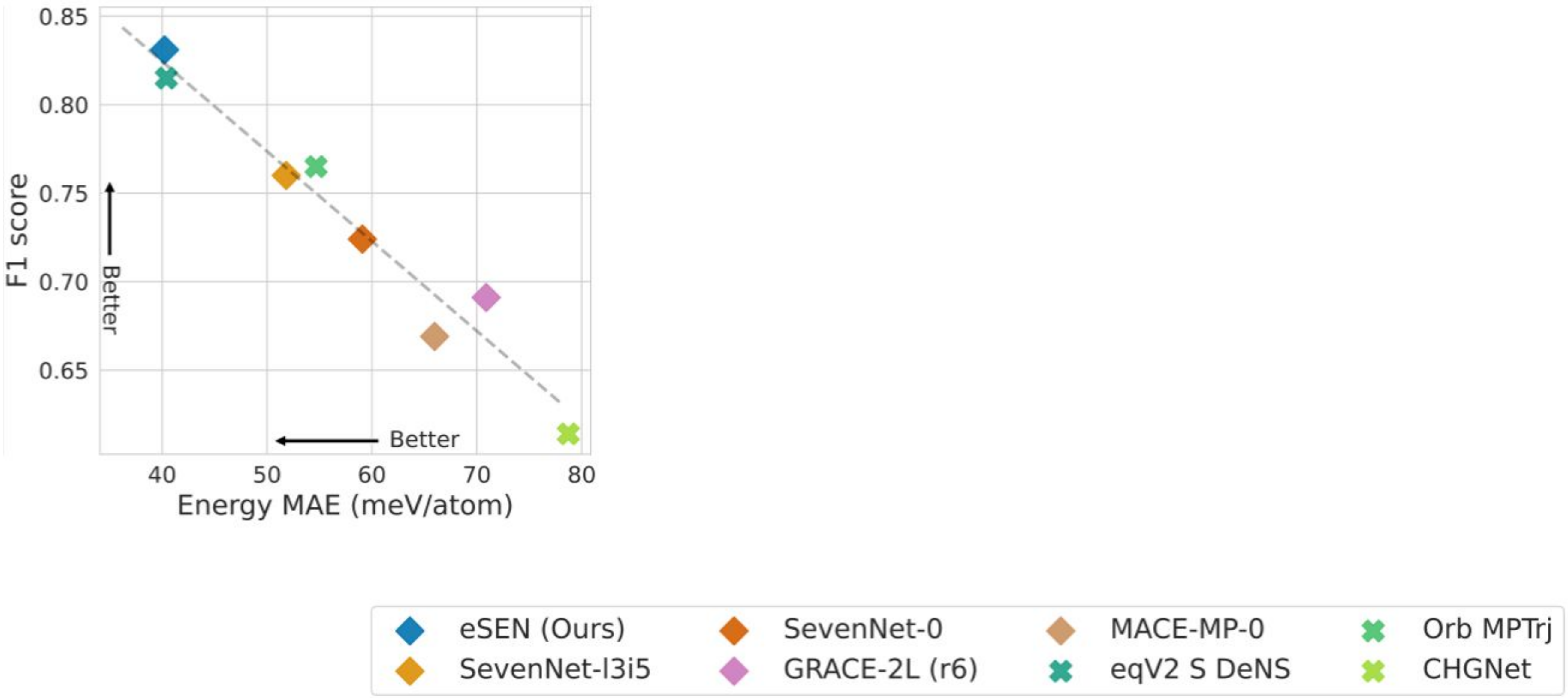
OMat24 also improves softening across models



Is a large dataset enough to train robust MLIPs for downstream applications?

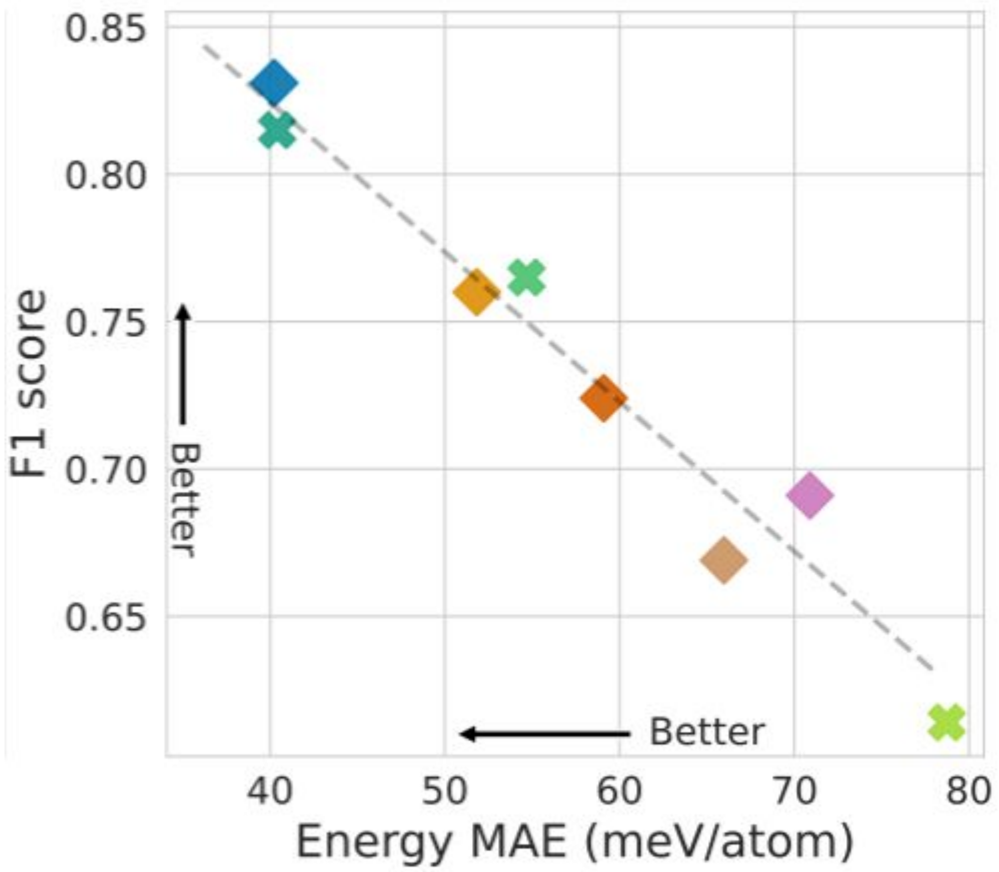
A gap between accuracy and reliability

Relaxation / Energy prediction

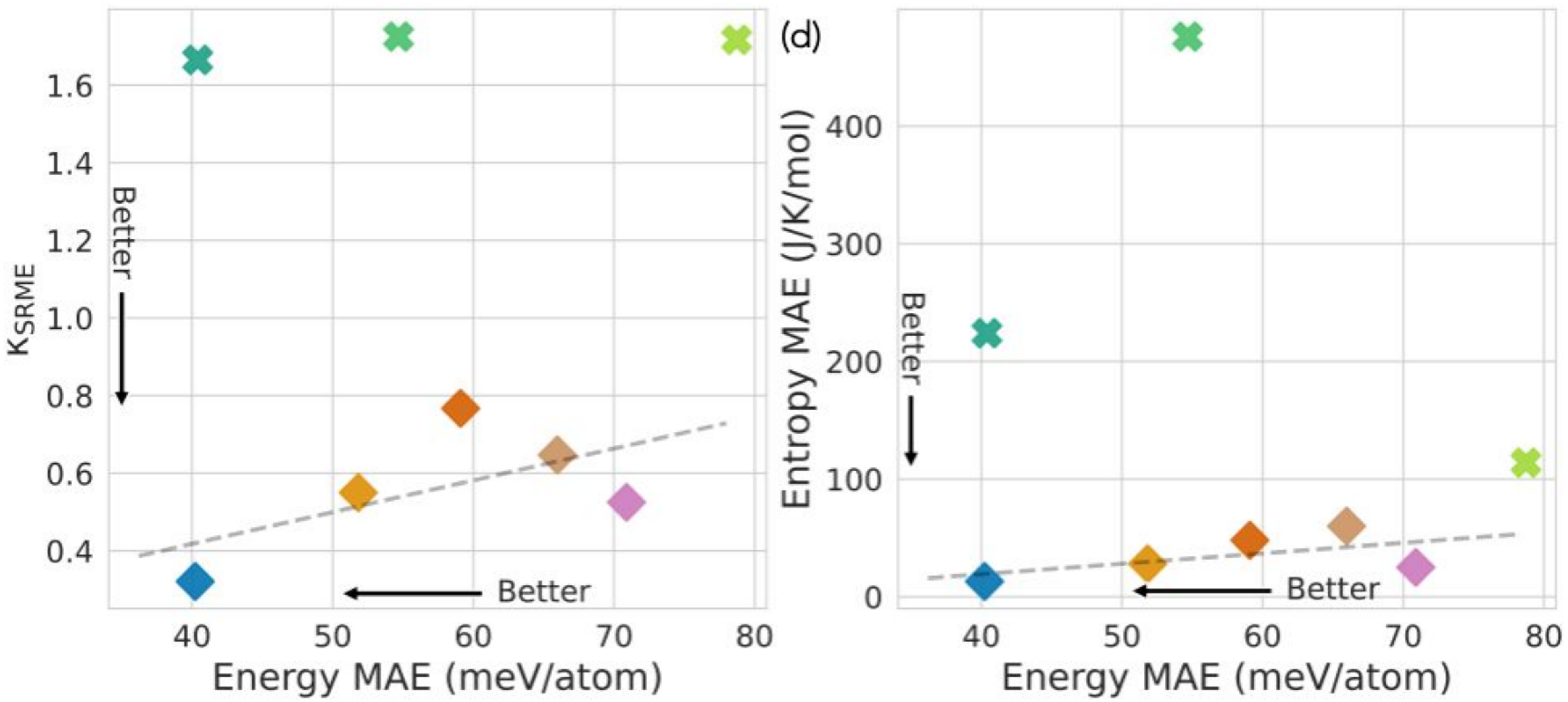


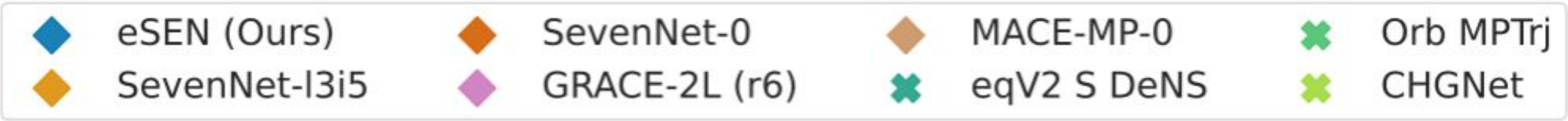
A gap between accuracy and reliability

Relaxation / Energy prediction



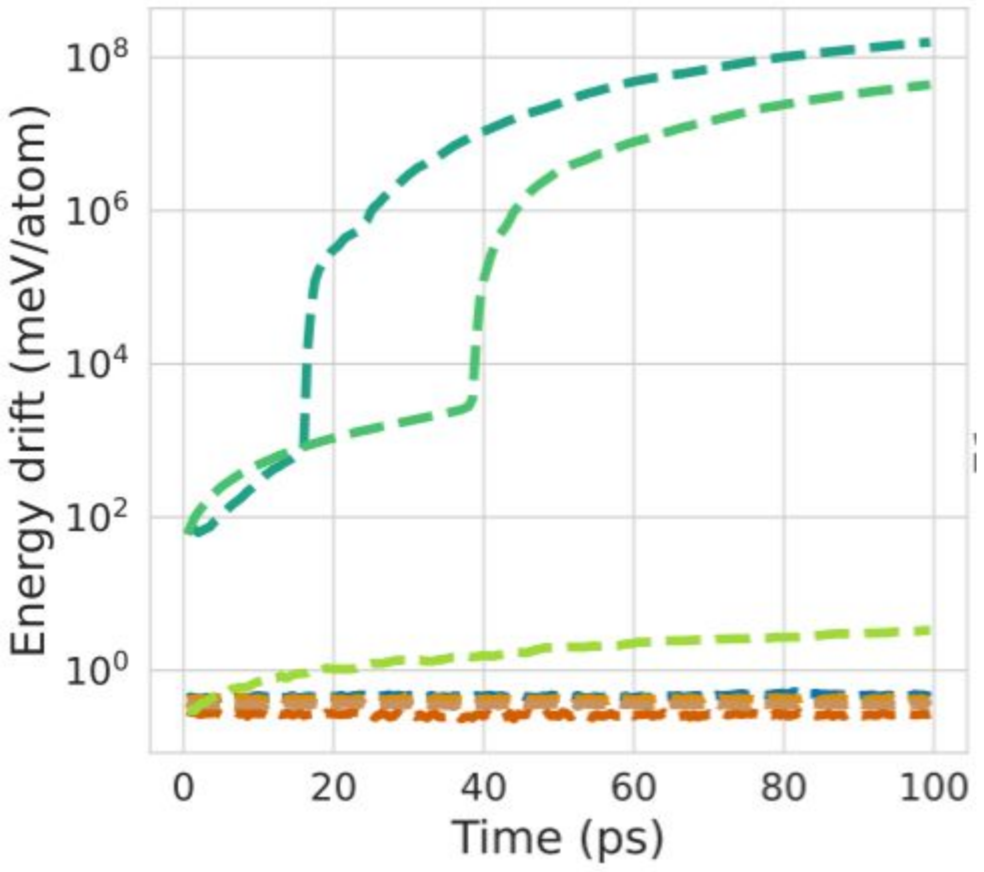
Phonon-related properties



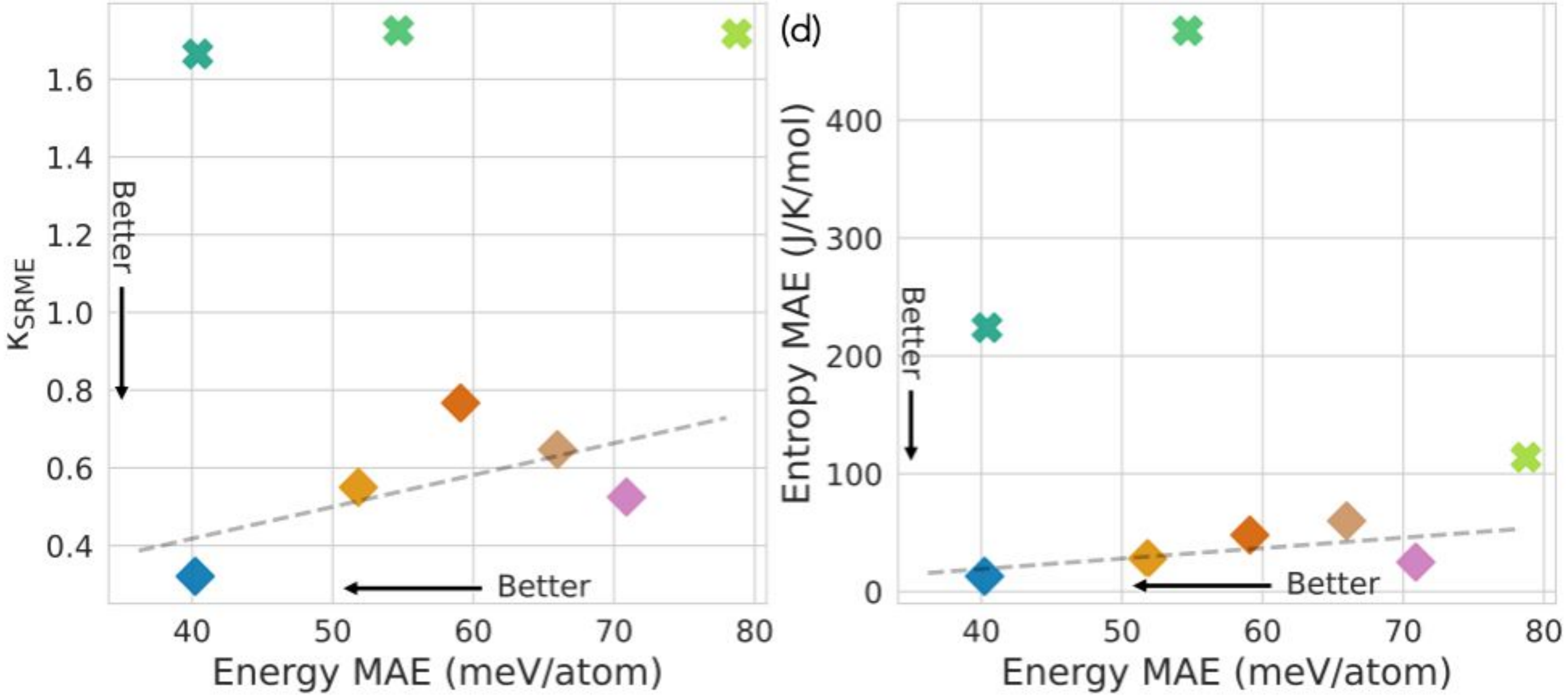


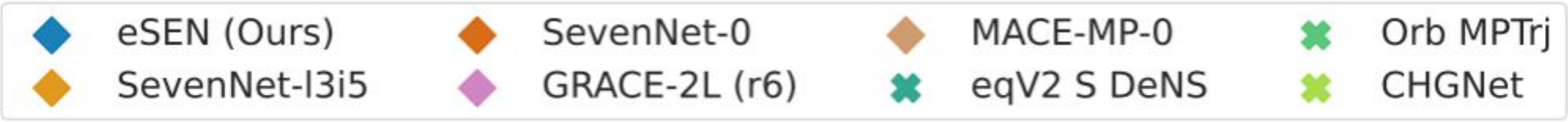
A gap between accuracy and reliability

Energy conservation in MD



Phonon-related properties

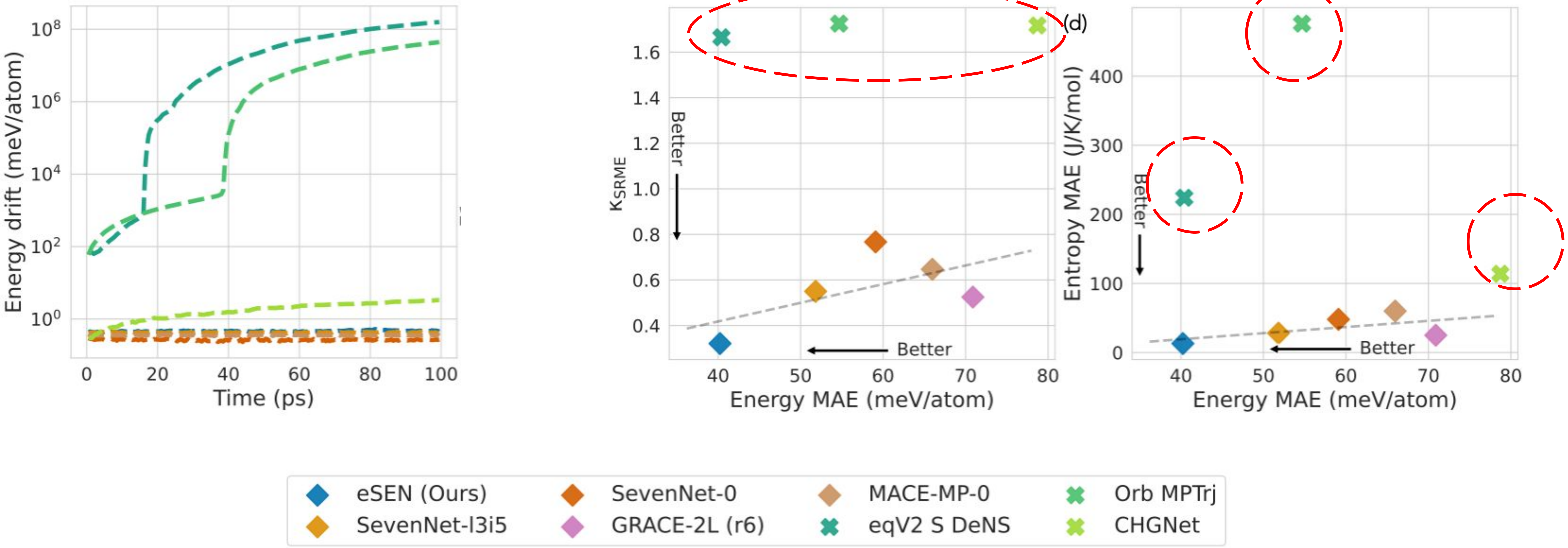




A gap between accuracy and reliability

Energy conservation in MD

Phonon-related properties



What does conservation in MD entail

$$|E(\boldsymbol{r}_T, \boldsymbol{a}) - E(\boldsymbol{r}_0, \boldsymbol{a})| \leq C\Delta t^2 + C_N \Delta t^N T$$

Conservation Error

Fluctuation

Long-term drift

What does conservation in MD entail

$$|E(\boldsymbol{r}_T, \boldsymbol{a}) - E(\boldsymbol{r}_0, \boldsymbol{a})| \leq C\Delta t^2 + C_N \Delta t^N T$$

Conservation Error

Fluctuation

Long-term drift

What does conservation in MD entail

$$|E(\boldsymbol{r}_T, \boldsymbol{a}) - E(\boldsymbol{r}_0, \boldsymbol{a})| \leq C\Delta t^2 + C_N \Delta t^N T$$

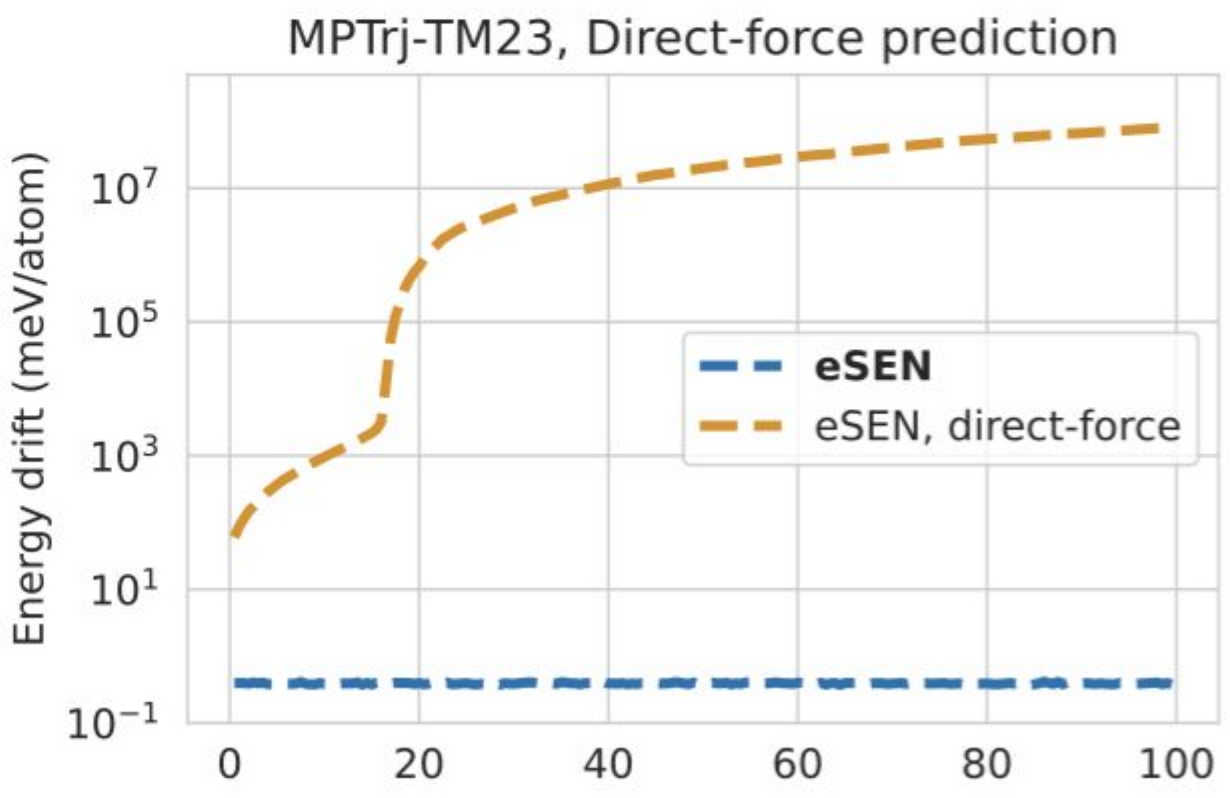
Long-term
drift

Bounded by the bounds on
higher-order PES derivatives

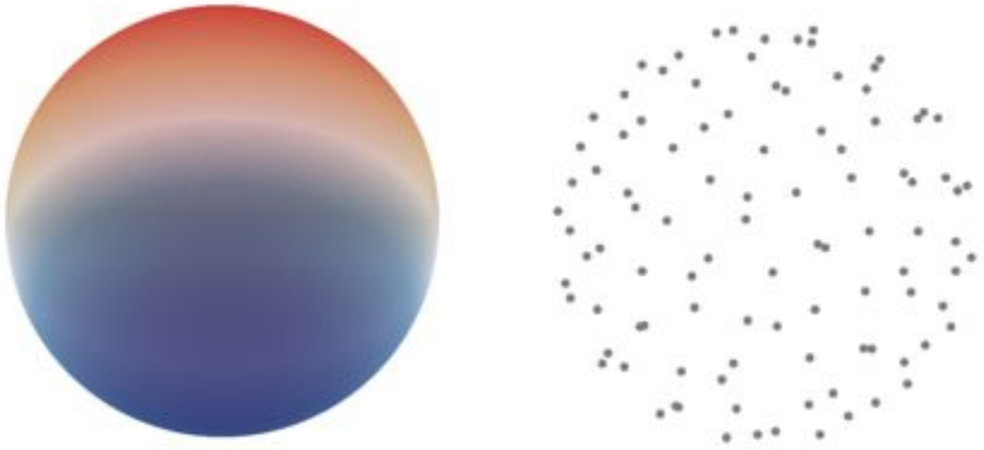
Design choices for smooth potentials

$$\mathbf{F} = \nabla_{\mathbf{r}} E$$

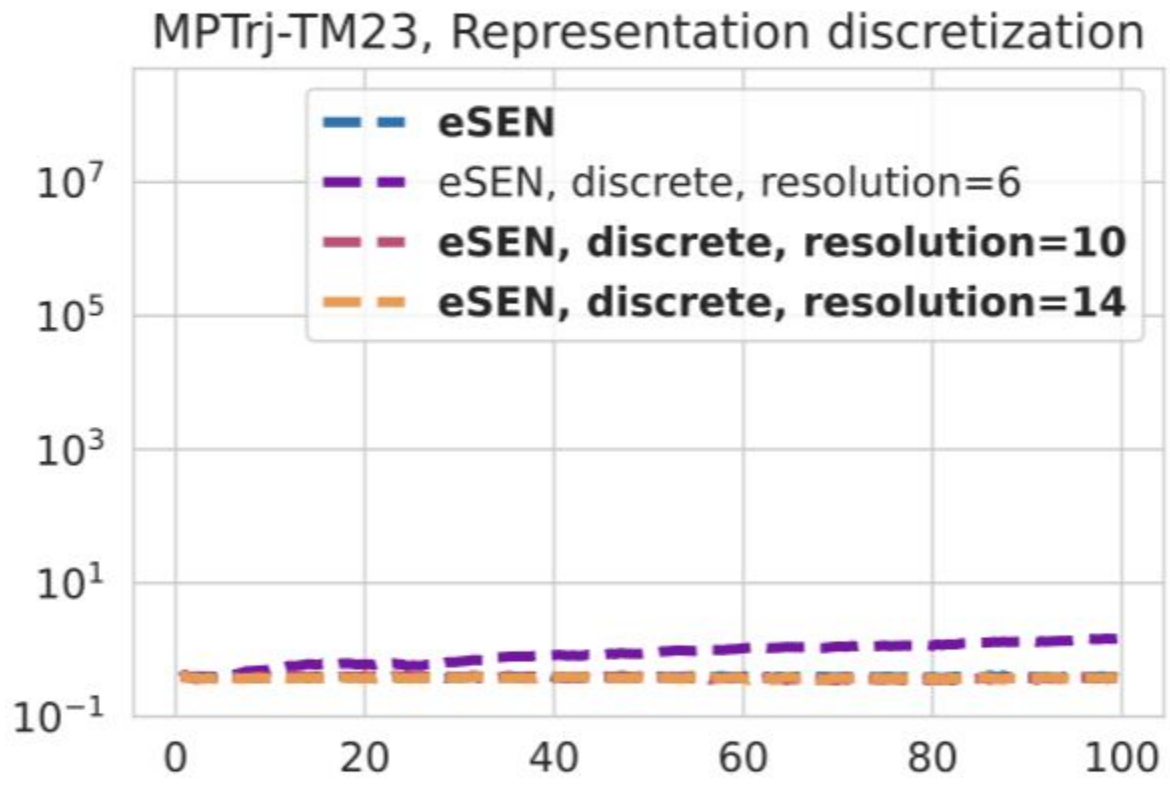
Energy conservation



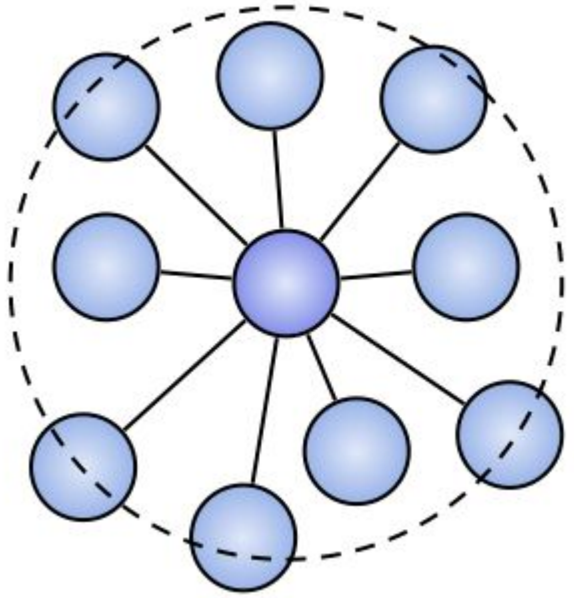
Design choices for smooth potentials



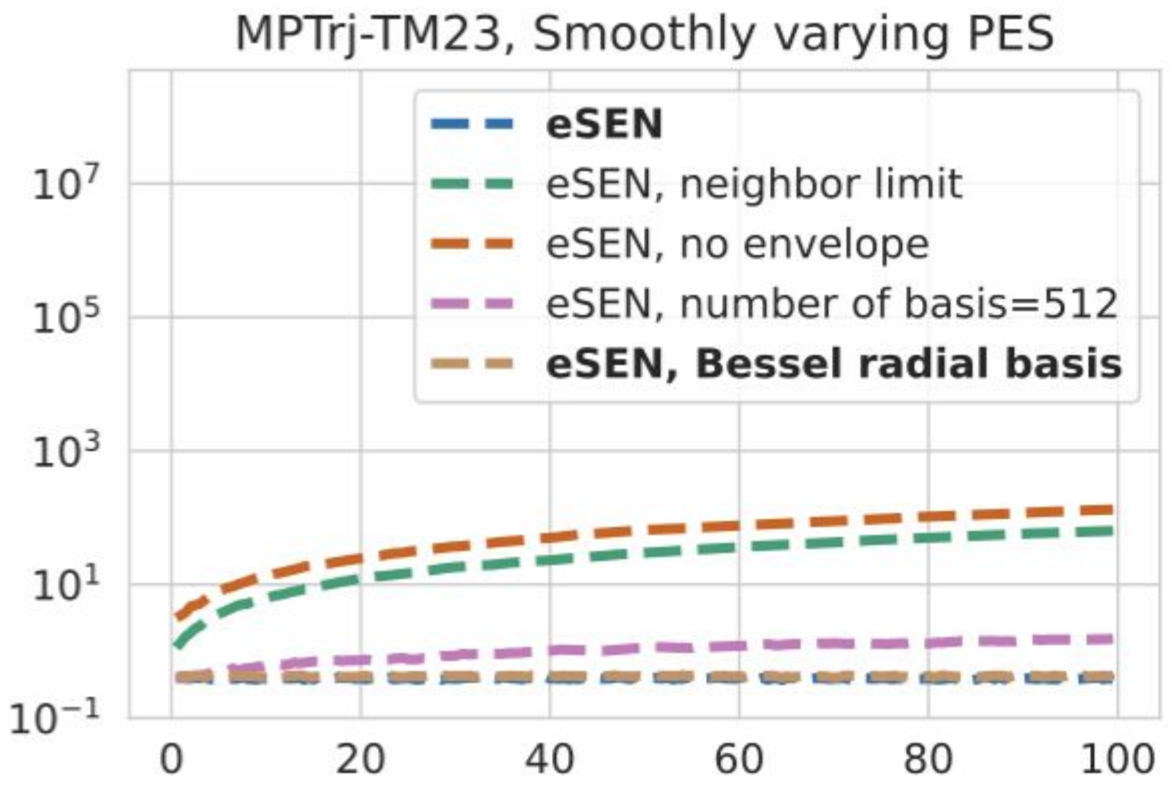
Representation Discretization



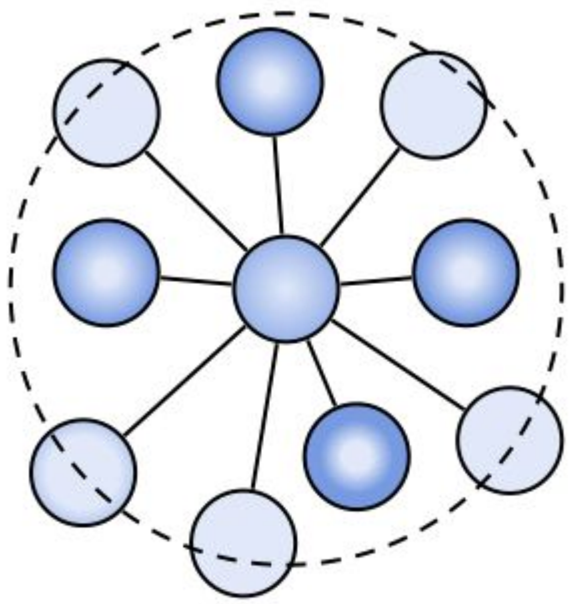
Design choices for smooth potentials



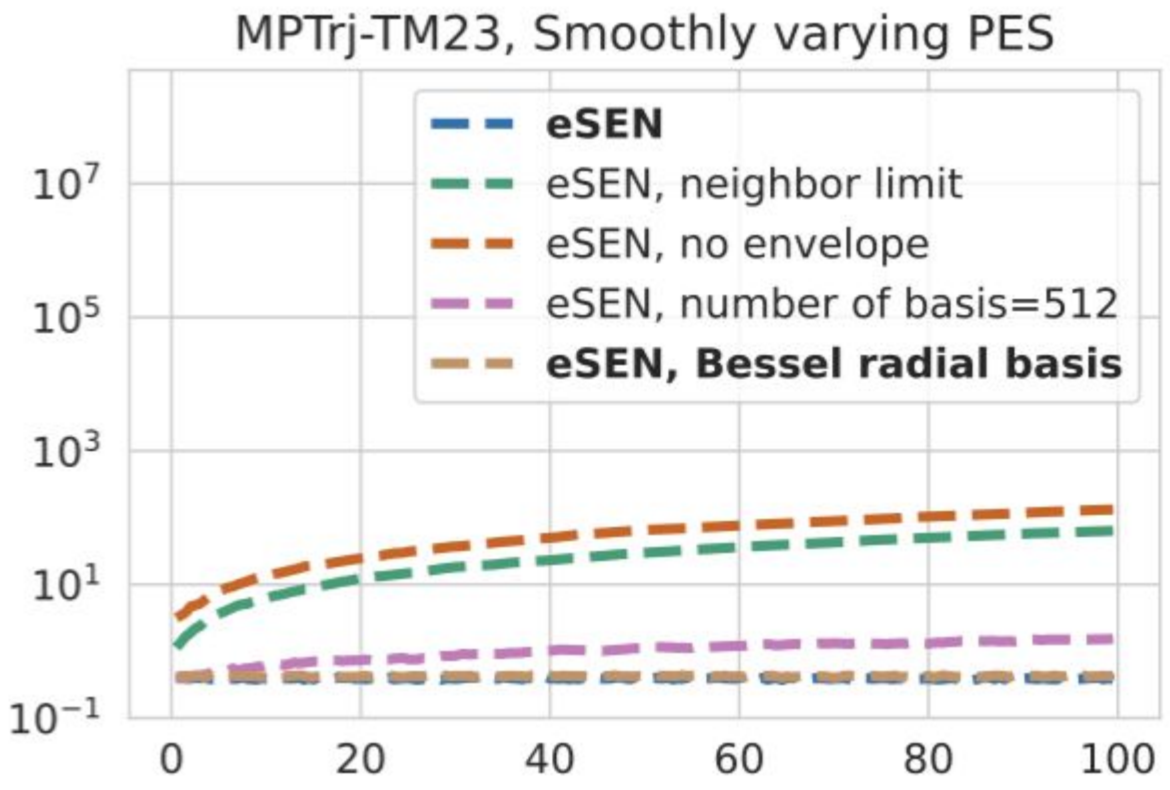
Cutoff-based graph



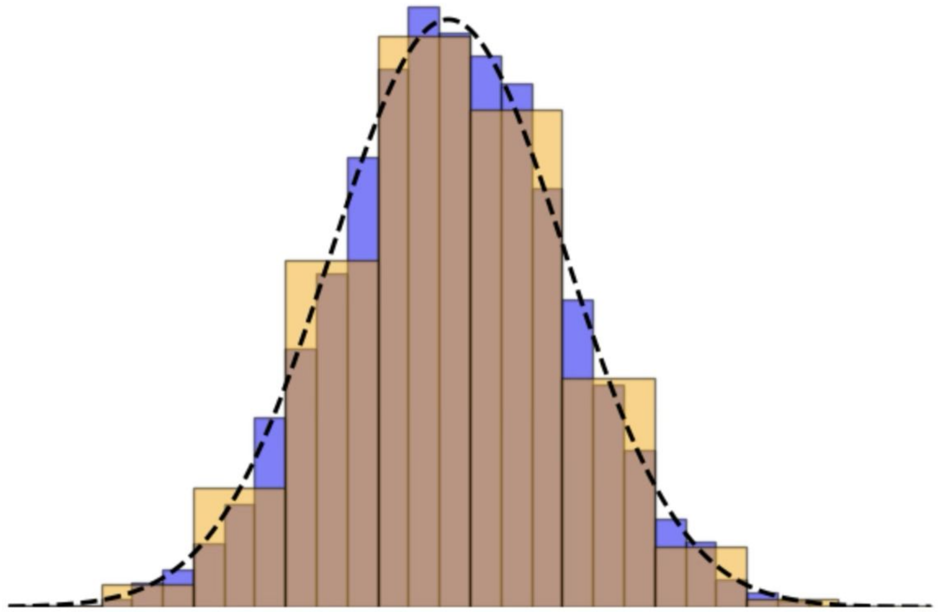
Design choices for smooth potentials



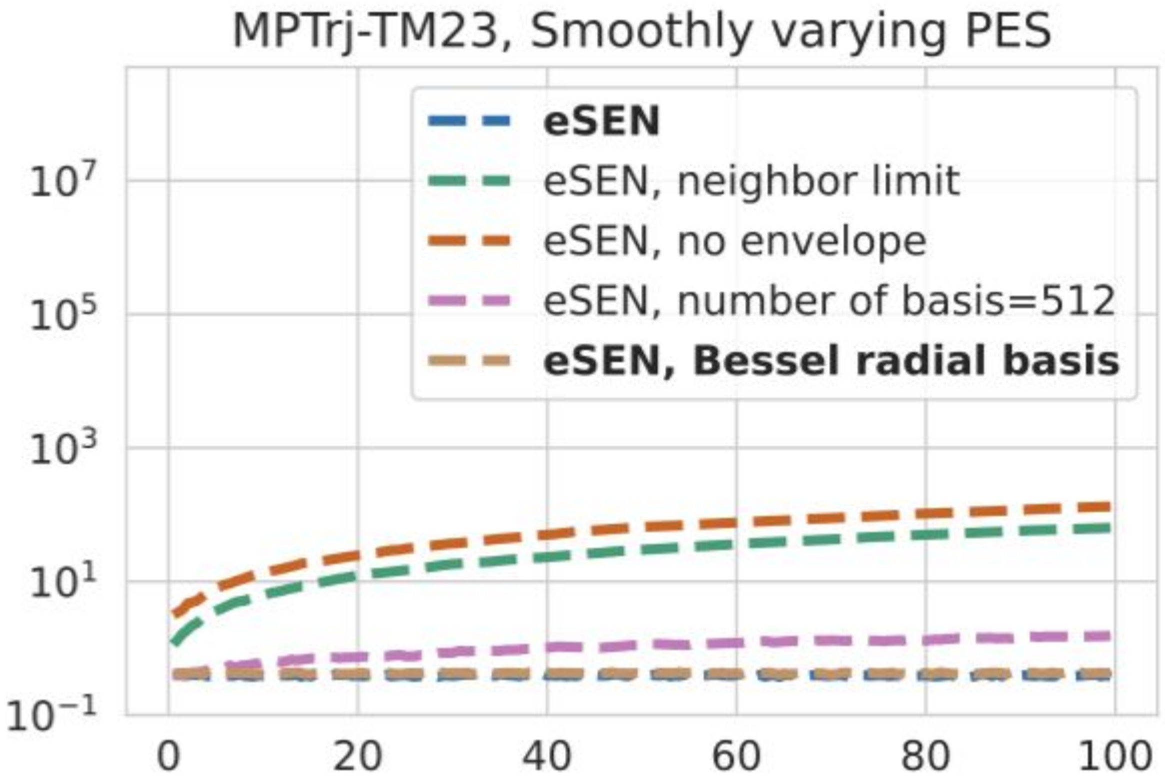
Envelope functions



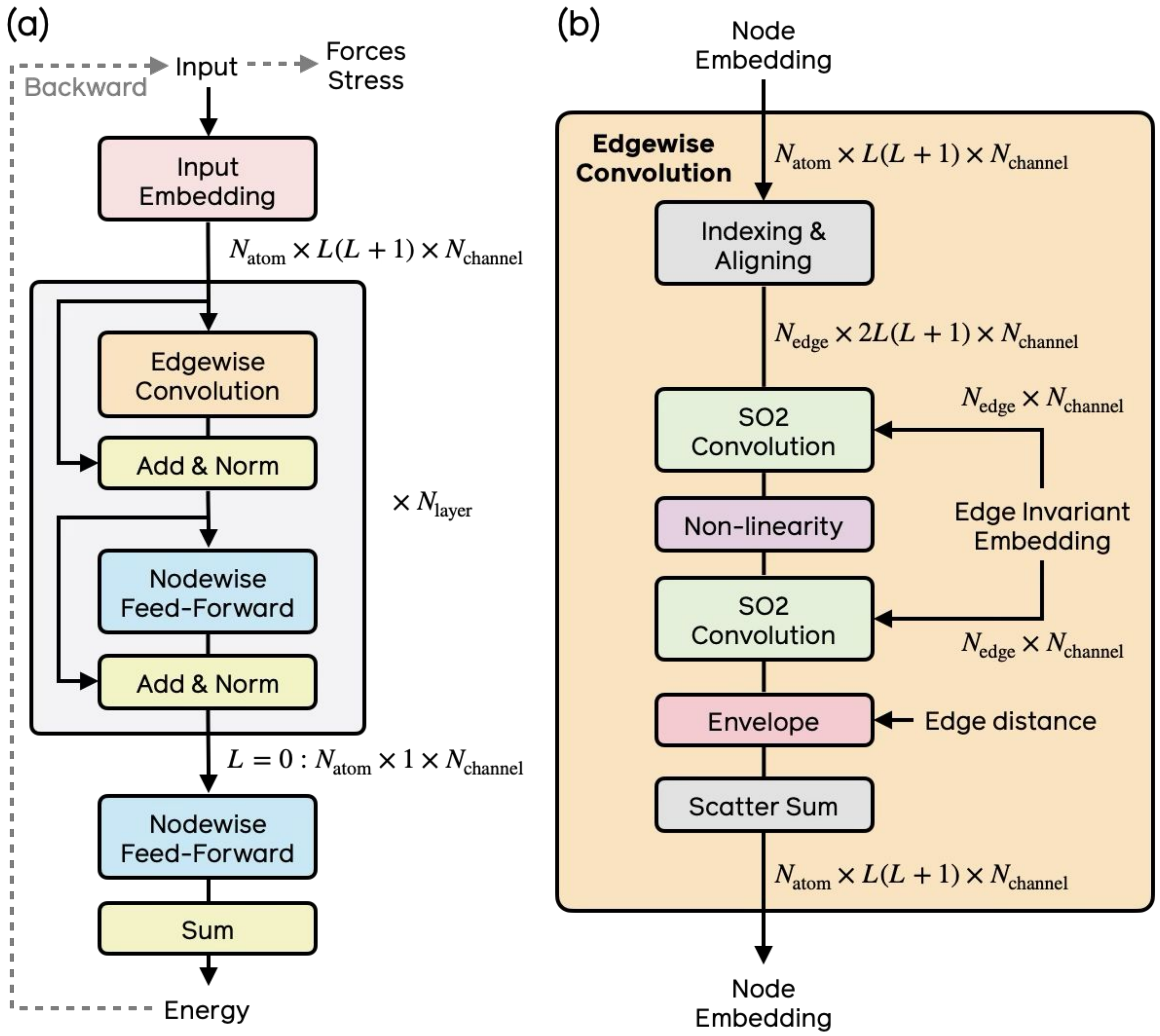
Design choices for smooth potentials



Radial-basis smoothing

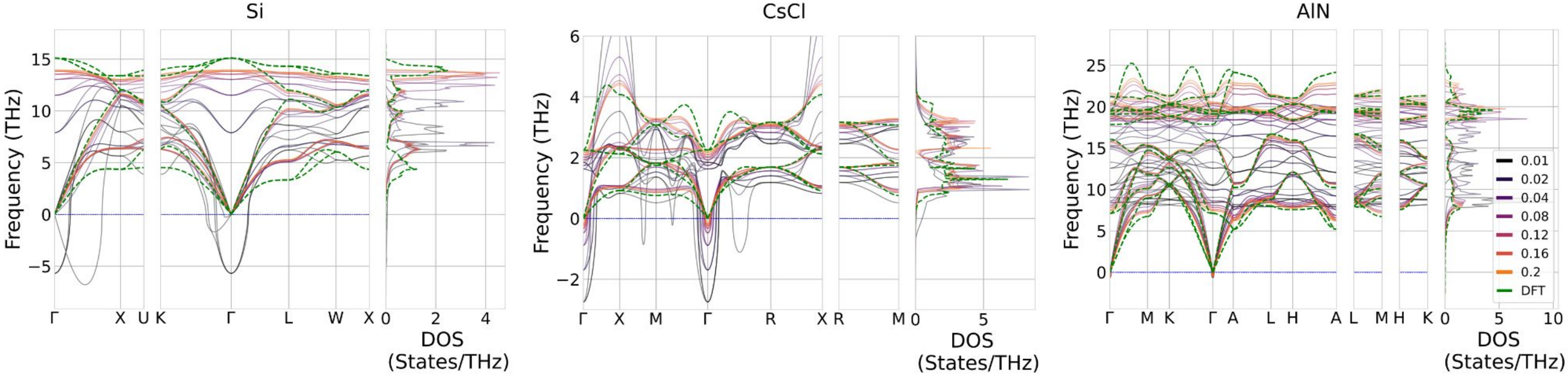


equivariant Smooth Energy Networks (eSEN)



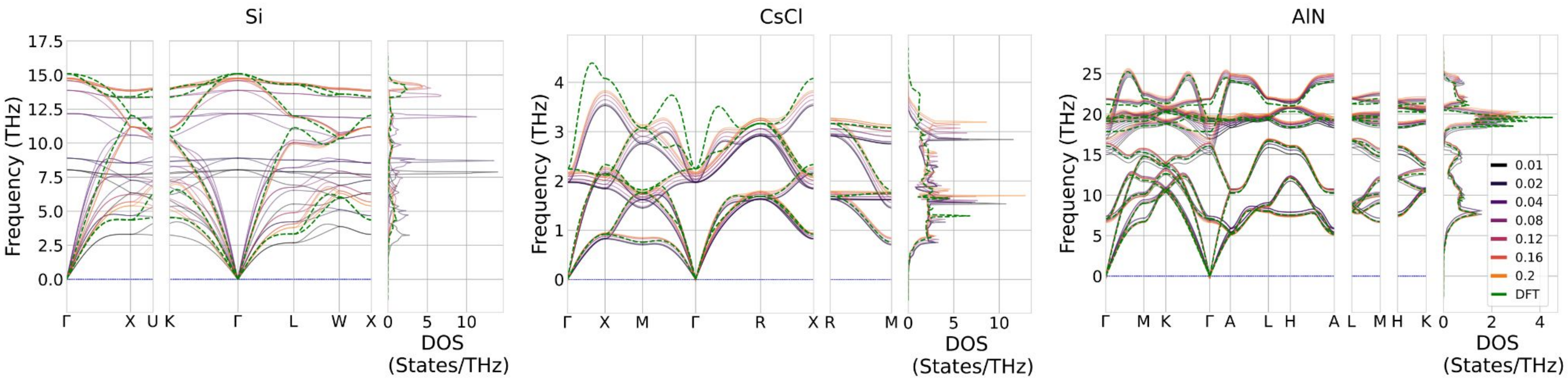
Does eSEN really learn a more faithful representation of the PES?

Remarkable improvements in phonon dispersion & DOS



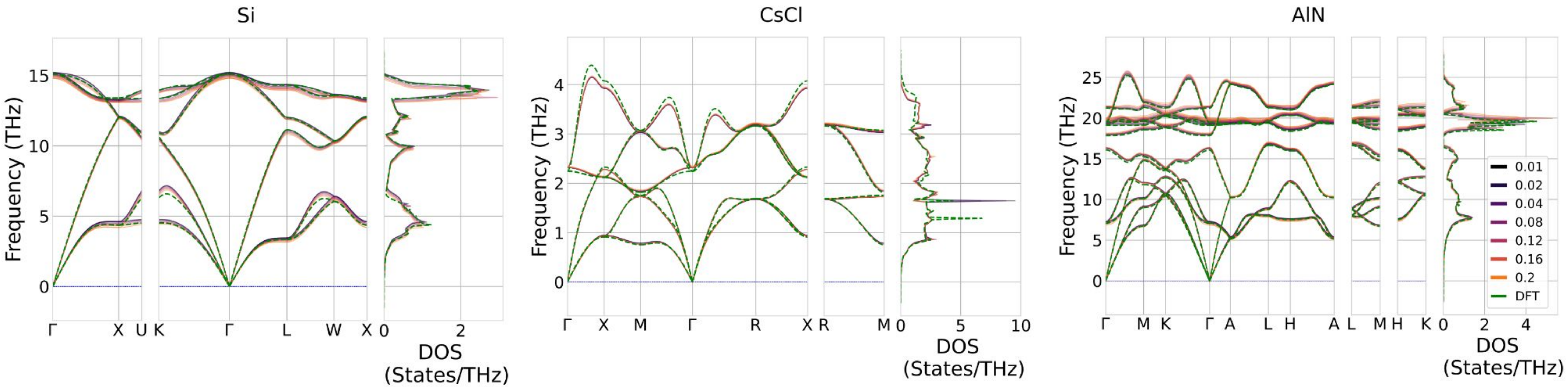
Phonon bands from EqV2 (direct forces)

Remarkable improvements in phonon dispersion & DOS



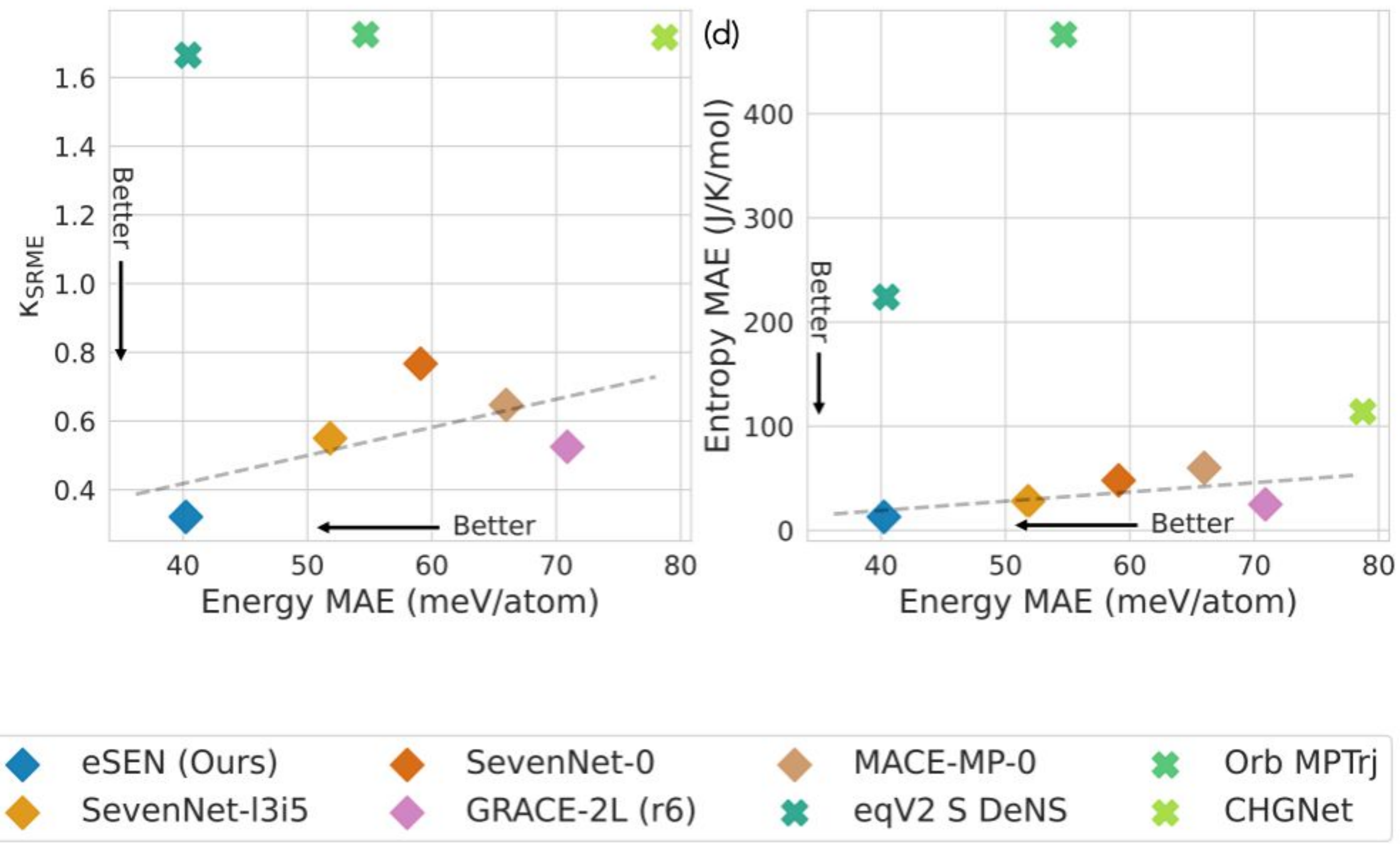
Phonon bands from eSEN (direct forces)

Remarkable improvements in phonon dispersion & DOS



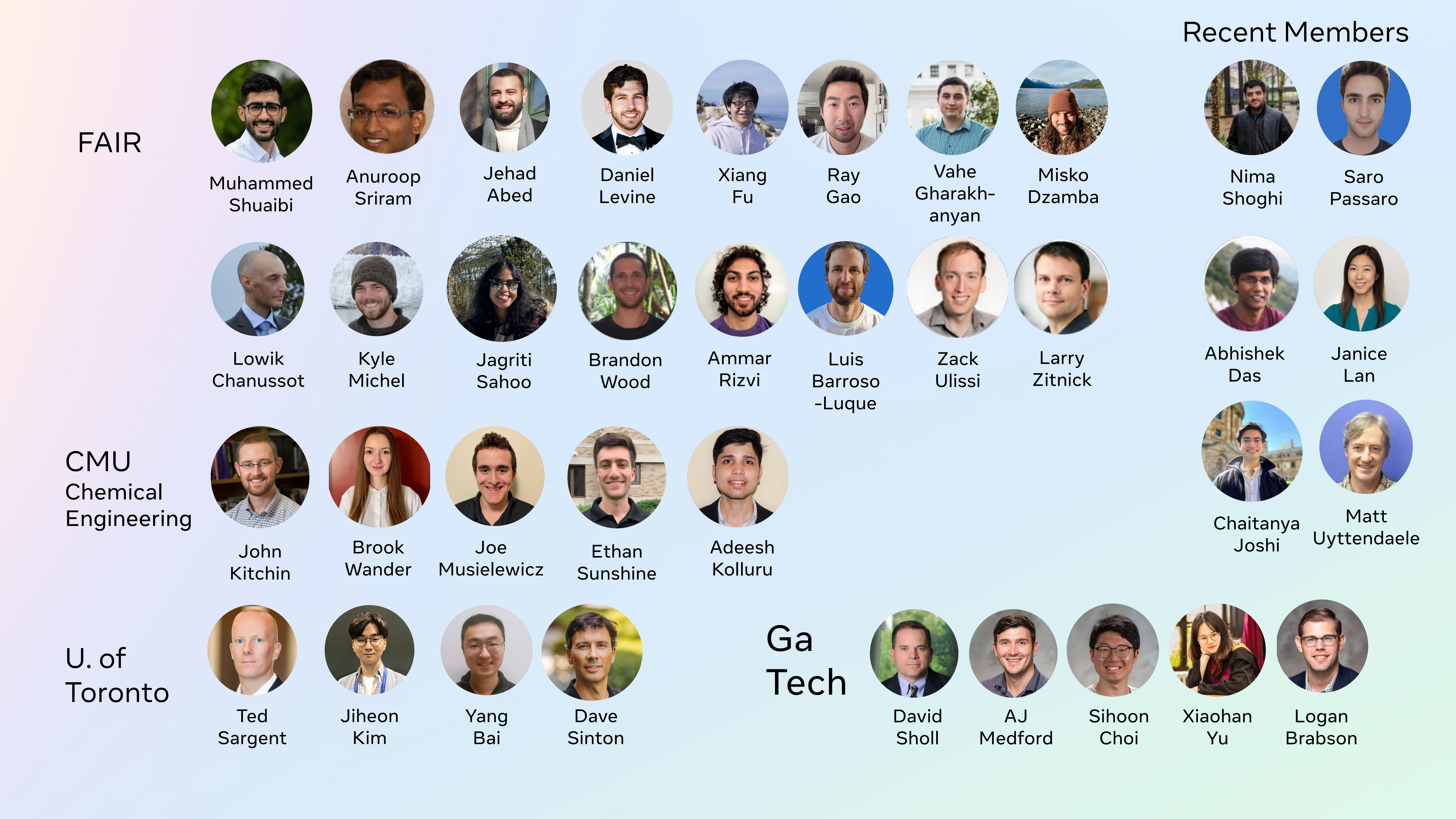
Phonon bands from eSEN (gradient forces)

Test error as a proxy metric works for *smooth* models



What we learned

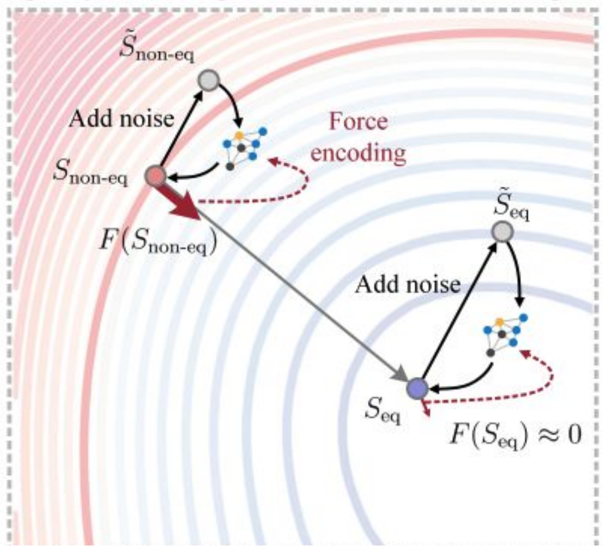
- Diverse and large datasets improve prediction accuracy and reduce PES softening.
- Model architecture design choices are critical for accurate downstream prediction and simulation.
- Energy conservation from numerical (NVE) integration can be used as an empirical proxy to test the smoothness of learned PES.
- Test set accuracy and prediction accuracy are strongly correlated for energy conserving models.

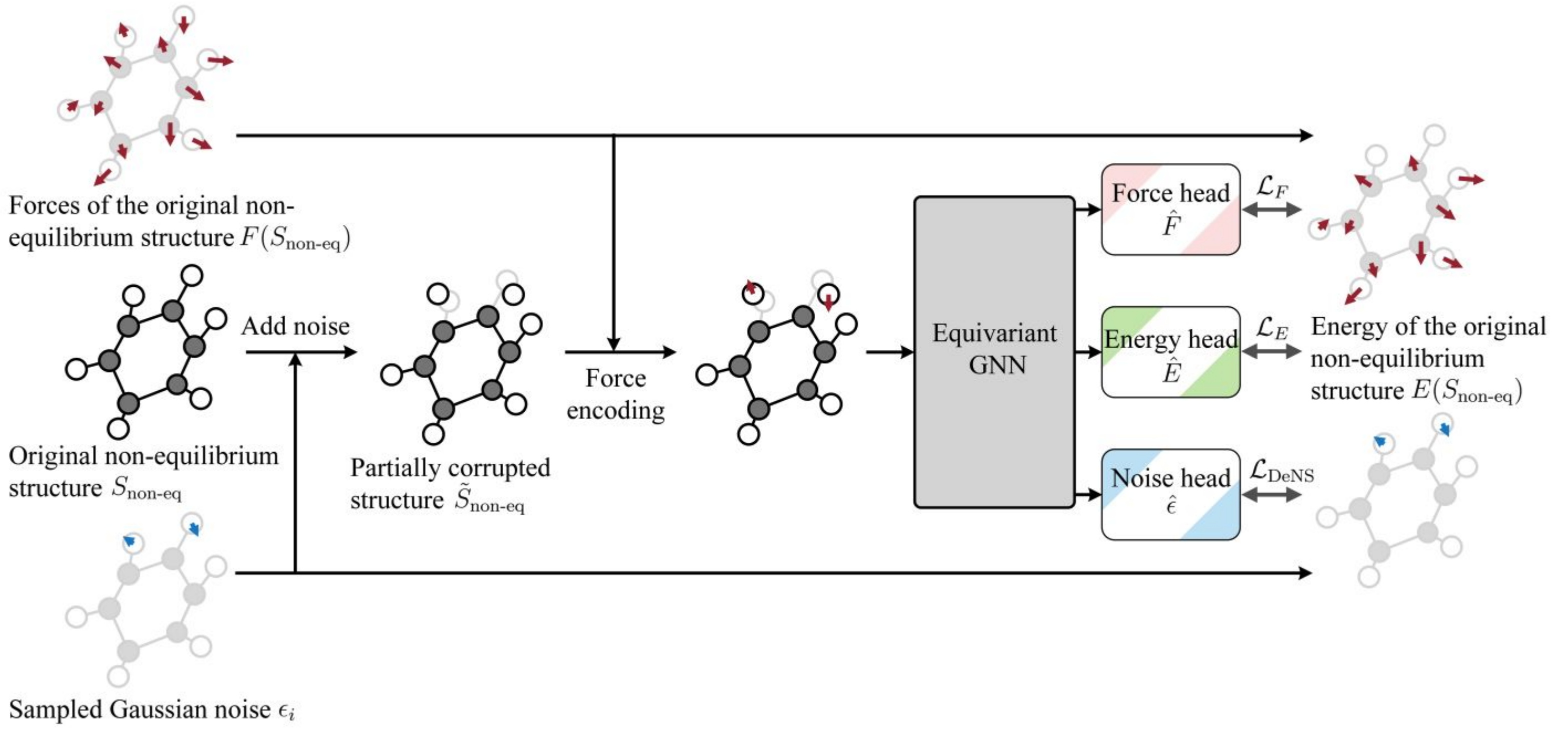


Recent Members
FAIR
Muhammed Shuaibi
Anuroop Sriram
Jehad Abed
Daniel Levine
Xiang Fu
Ray Gao
Vahe Gharakh-anyan
Misko Dzamba
Nima Shoghi
Saro Passaro
Lowik Chanussot
Kyle Michel
Jagriti Sahoo
Brandon Wood
Ammar Rizvi
Luis Barroso-Luque
Zack Ulissi
Larry Zitnick
Abhishek Das
Janice Lan
CMU Chemical Engineering
John Kitchin
Brook Wander
Joe Musielewicz
Ethan Sunshine
Adeesh Kolluru
Chaitanya Joshi
Matt Uyttendaele
U. of Toronto
Ted Sargent
Jiheon Kim
Yang Bai
Dave Sinton
Ga Tech
David Sholl
AJ Medford
Sihoon Choi
Xiaohan Yu
Logan Brabson

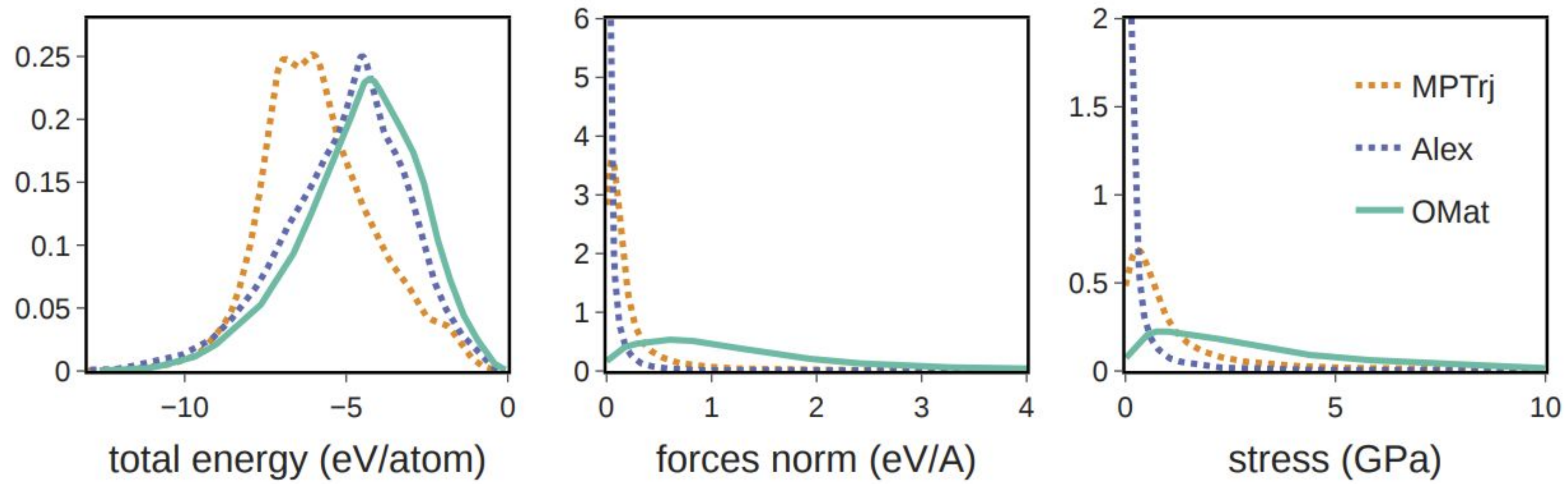
Two ways to overcome limitations of near-equilibrium data

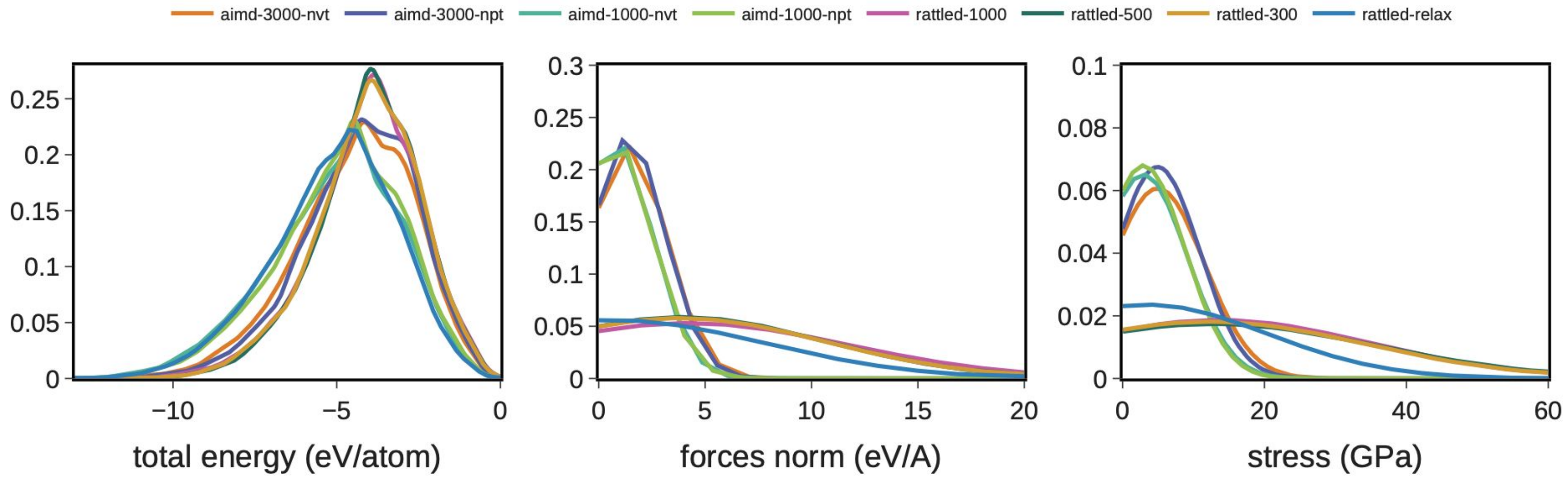
Denoising/dataset augmentation





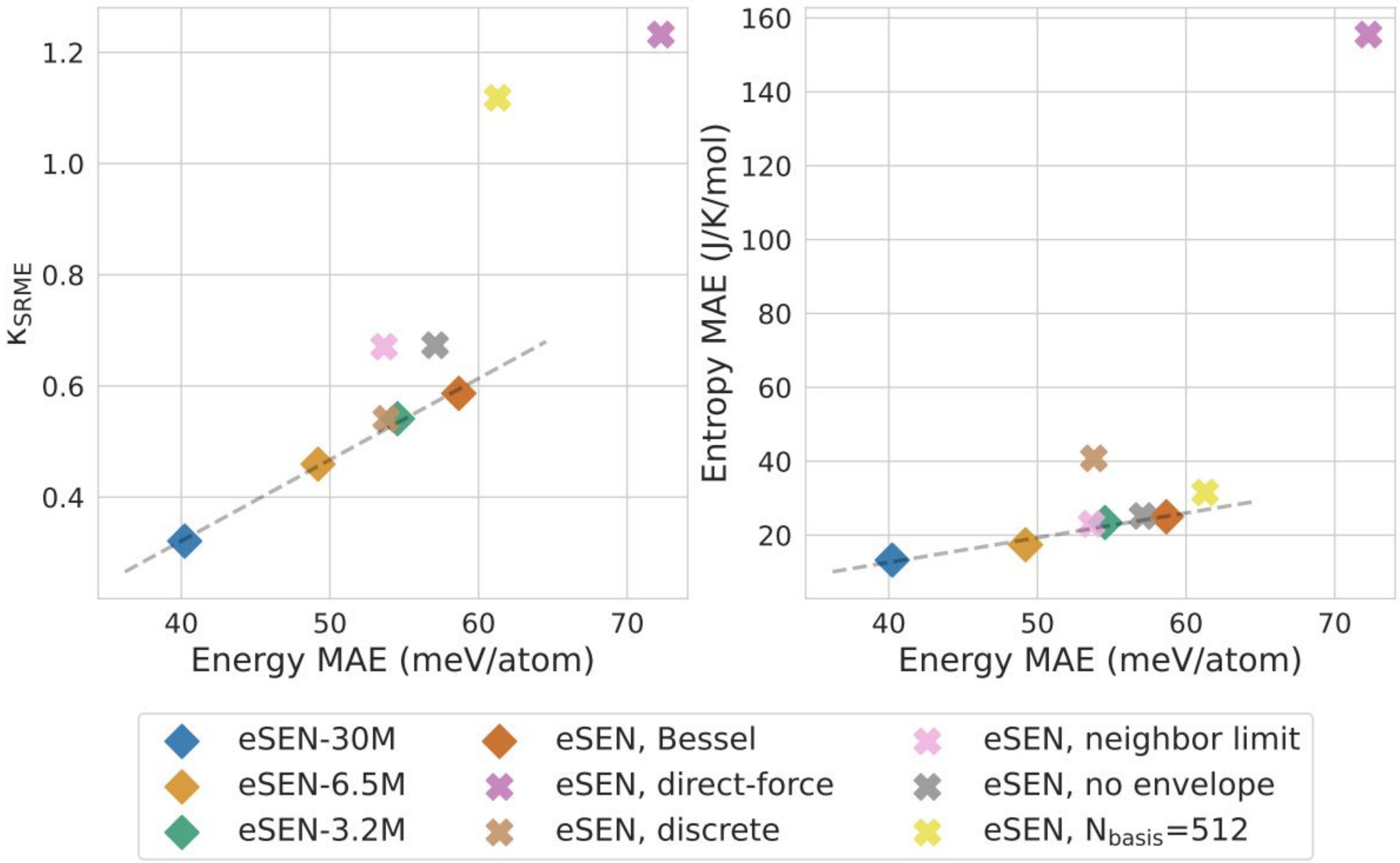
Diverse non-equilibrium data – OMat24 dataset!



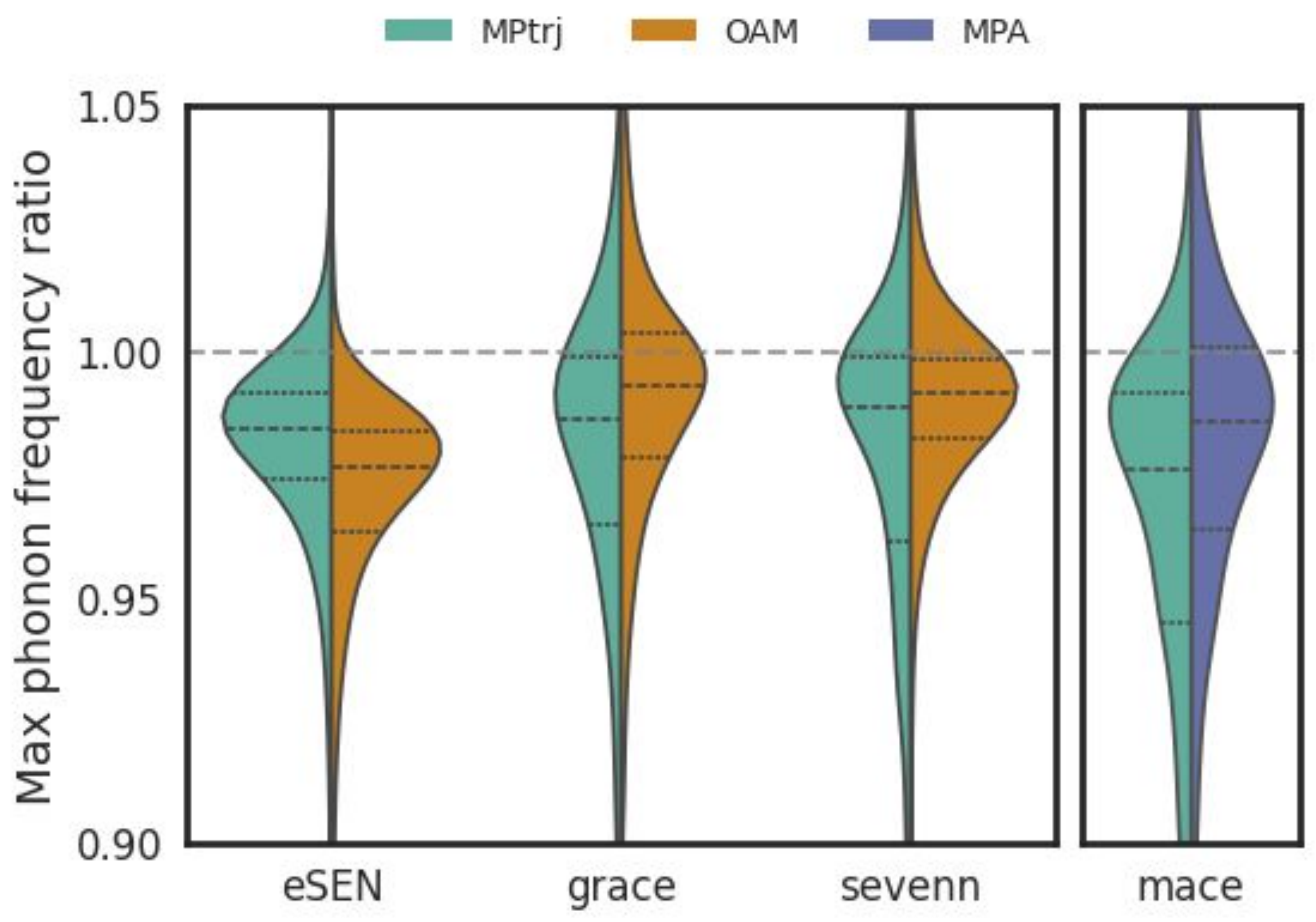


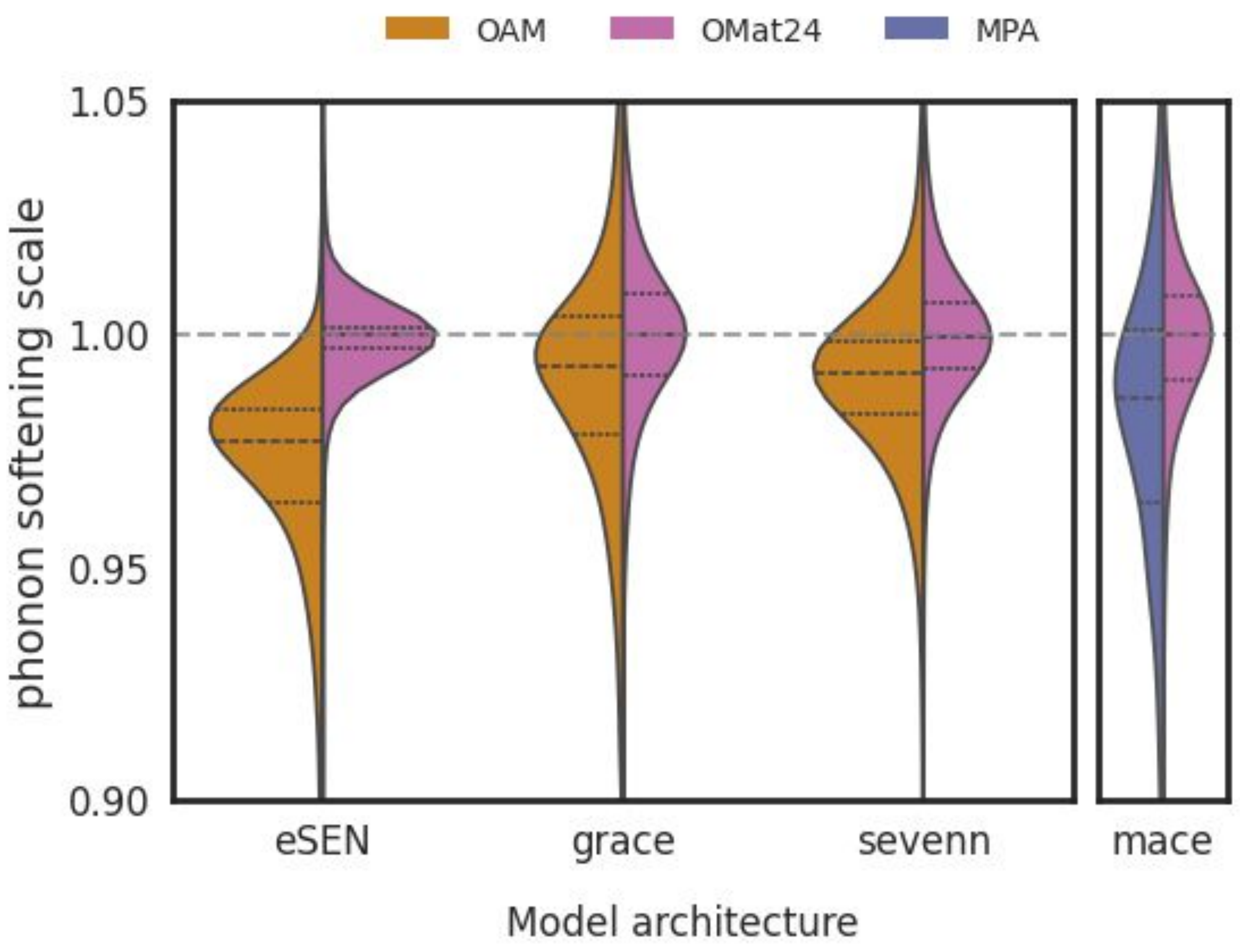
https://arxiv.org/pdf/2403.09549

Test error as a proxy metric

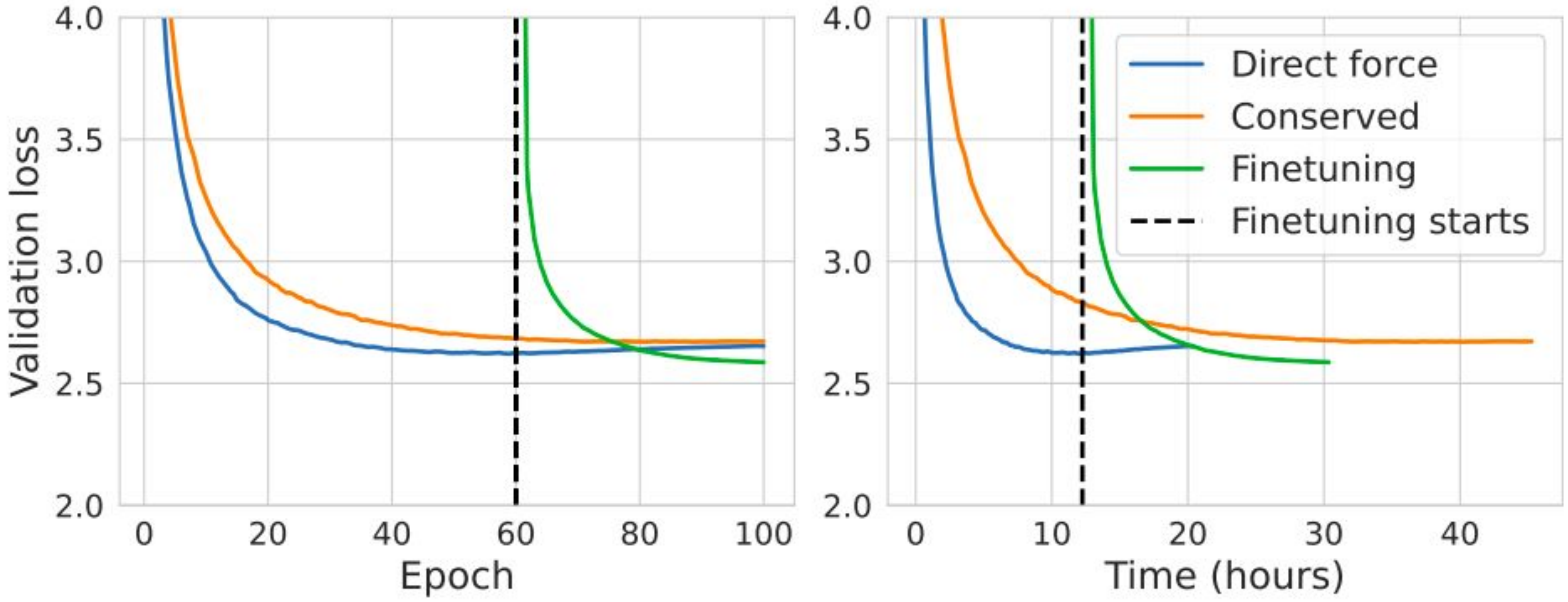


Improved phonons with eSEN + OMat





Direct force pretraining improves model accuracy in less training time



Smoothness is not necessarily captured in test errors

Model	MPTrj Energy	MPTrj Force	MPTrj Stress	SPICE Energy	SPICE Force
eSEN	17.02	43.96	0.14	0.23	6.36
eSEN, direct	18.66	43.62	0.16	0.56	10.98
eSEN, neighbor limit	17.30	44.11	0.14	0.24	6.52
eSEN, no envelope	17.60	44.69	0.14	0.23	6.33
eSEN, $N_{basis} = 512$	19.87	48.29	0.15	0.19	5.40
eSEN, Bessel	17.65	44.83	0.15	0.20	5.54
eSEN, discrete, res=6	17.05	43.10	0.14	0.26	6.34
eSEN, discrete, res=10	17.11	43.13	0.14	0.33	6.57
eSEN, discrete, res=14	17.12	43.09	0.14	0.33	6.51