MLIP Arena

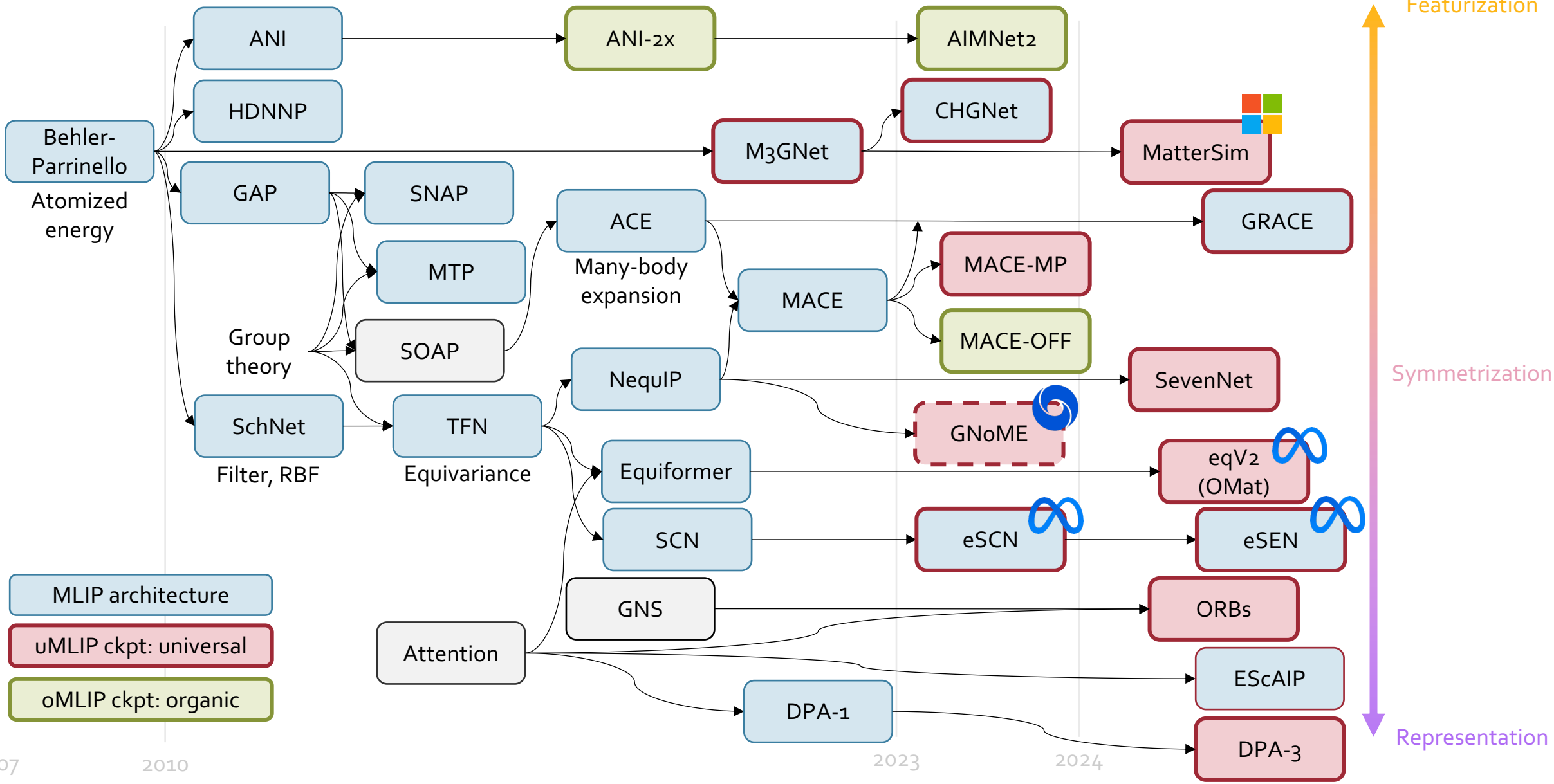
Advancing fairness and transparency in machine learning interatomic potentials via an open accessible benchmark platform

Yuan Chiang
(cyrusyc@berkeley.edu)
University of California, Berkeley
Lawrence Berkeley National Laboratory

Tobias Kreiman, Elizabeth Weaver, Ishan Amin, Matthew Kuner, Christine Zhang, Aaron Kaplan, Daryl Chrzan, Samuel M. Blau, Aditi S. Krishnapriyan, Mark Asta

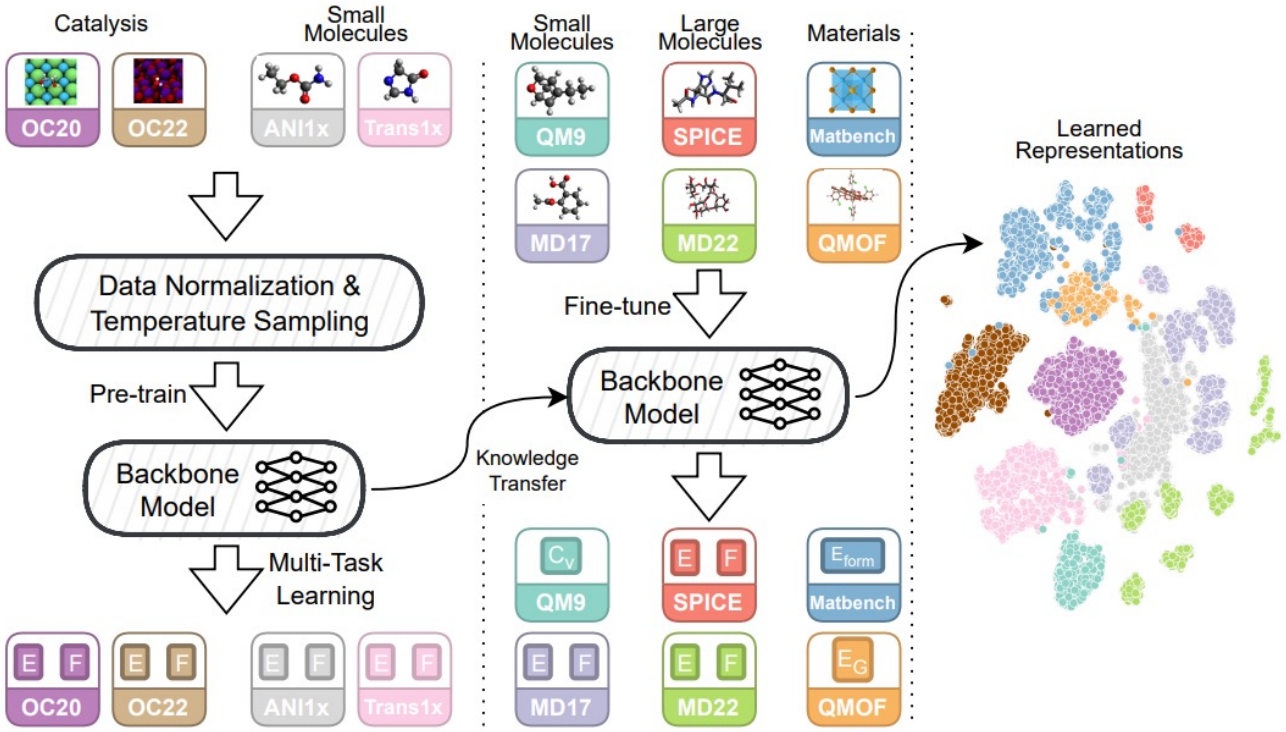
ICLR 2025 AI4Mat
Online leaderboard: https://huggingface.co/spaces/atomind/mlip-arena
OpenReview: https://openreview.net/forum?id=ysKfIavYQE

MLIP Architectures and Pretrained Models



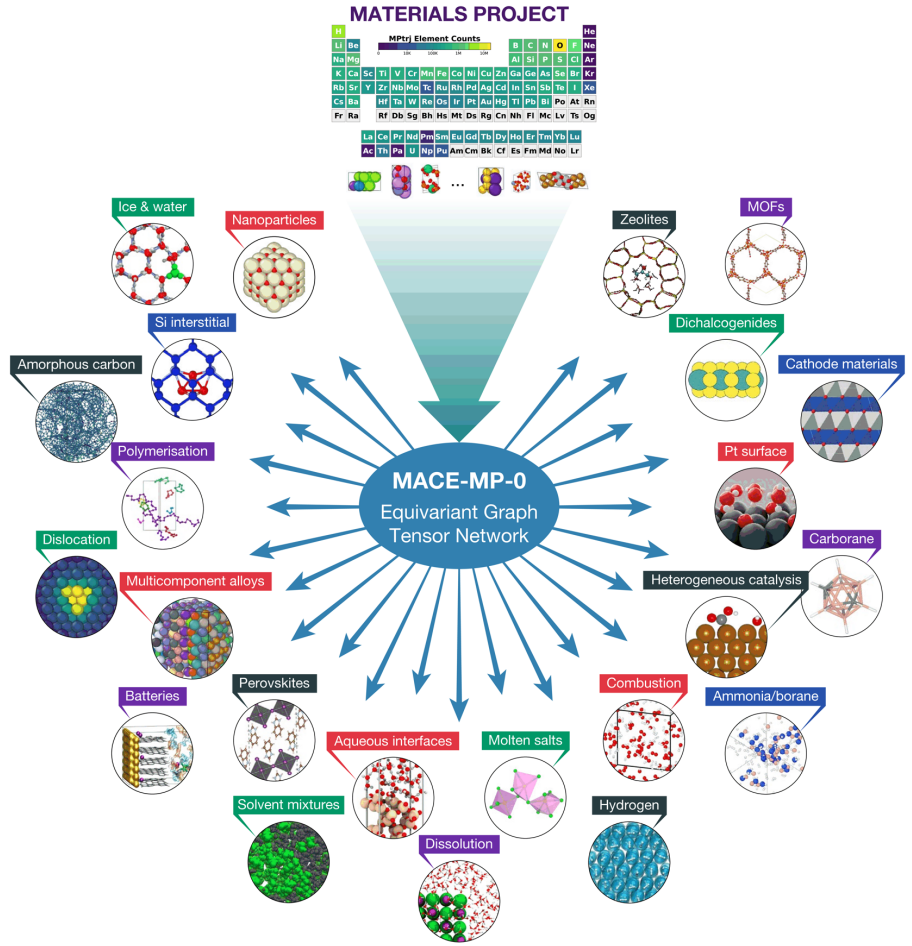
The Emergence of Foundation MLIPs

JMP-1



Shoghi, N., Kolluru, A., Kitchin, J. R., Ulissi, Z. W., Zitnick, C. L., & Wood, B. M. (2023). ICLR 2024

MACE-MP-0



Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kovács, D. P., Riebesell, J., ... & Csányi, G. (2023). arXiv:2401.00096.

Matbench Discovery – ML Crystal Stability Predictions

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, Kristin A. Persson

Janosh Riebesell

Rhys Goodall

Model	F1 ↑	DAF ↑	Prec ↑	Acc ↑	TPR ↑	TNR ↑	MAE ↓	RMSE ↓	R^2 ↑
eqV2 M	0.917	6.047	0.924	0.975	0.910	0.986	0.020	0.072	0.848
DPA3-v1-OpenLAM	0.883	5.754	0.880	0.963	0.885	0.978	0.023	0.067	0.869
GRACE-2L-OAM	0.880	5.774	0.883	0.963	0.878	0.979	0.023	0.068	0.862
ORB	0.880	6.041	0.924	0.965	0.841	0.987	0.028	0.077	0.824
MatterSim v1 5M	0.862	5.852	0.895	0.959	0.831	0.982	0.024	0.068	0.863
MACE-MPA-0	0.852	5.582	0.853	0.954	0.851	0.973	0.028	0.073	0.842
GNoME	0.829	5.523	0.844	0.955	0.814	0.972	0.035	0.085	0.785
GRACE-1L-OAM	0.824	5.255	0.803	0.944	0.846	0.962	0.031	0.073	0.842
eqV2 S DeNS	0.815	5.042	0.771	0.941	0.864	0.953	0.036	0.085	0.788
AlphaNet-MPTrj	0.799	4.863	0.743	0.933	0.864	0.945	0.041	0.093	0.745
ORB MPtrj	0.765	4.702	0.719	0.922	0.817	0.941	0.045	0.091	0.756
DPA3-v1-MPtrj	0.765	4.654	0.711	0.921	0.828	0.938	0.042	0.083	0.798
SevenNet-l3i5	0.760	4.629	0.708	0.920	0.821	0.938	0.044	0.087	0.776
SevenNet-0	0.724	4.252	0.650	0.904	0.818	0.919	0.048	0.092	0.750
GRACE-2L-MPtrj	0.691	4.163	0.636	0.896	0.757	0.921	0.052	0.094	0.741
MACE-MP-0	0.669	3.777	0.577	0.878	0.796	0.893	0.057	0.101	0.697
CHGNet	0.613	3.361	0.514	0.851	0.758	0.868	0.063	0.103	0.689
M3GNet	0.569	2.882	0.441	0.813	0.803	0.813	0.075	0.118	0.585
ALIGNN	0.567	3.206	0.490	0.841	0.672	0.872	0.093	0.154	0.297

Training Set	Model Params	Targets	Date Added
3M (102.4M) (OMat24+MPtrj)	86.6M	EFS_D	2024-10-18
7M (112.9M) (OMat24+MPtrj+sAlex)	8.2M	EFS_G	2025-01-10
7M (112.9M) (OMat24+sAlex+MPtrj)	12.6M	EFS_G	2025-02-06
3M (32.1M) (MPtrj+Alex)	25.2M	EFS_D	2024-10-11
17M (MatterSim)	4.5M	EFS_G	2024-12-16
3M (12.0M) (MPtrj+sAlex)	9.1M	EFS_G	2024-12-09
6M (89.0M) (GNoME)	16.2M	EF_G	2024-02-03
7M (112.9M) (OMat24+sAlex+MPtrj)	3.4M	EFS_G	2025-02-06
146k (1.6M) (MPtrj)	31.2M	EFS_D	2024-10-18
146k (1.6M) (MPtrj)	16.2M	EFS_G	2025-03-05
146k (1.6M) (MPtrj)	25.2M	EFS_D	2024-10-14
146k (1.6M) (MPtrj)	3.4M	EFS_G	2025-01-10
146k (1.6M) (MPtrj)	1.2M	EFS_G	2024-12-10
146k (1.6M) (MPtrj)	842.4k	EFS_G	2024-07-13
146k (1.6M) (MPtrj)	15.3M	EFS_G	2024-11-21
146k (1.6M) (MPtrj)	4.7M	EFS_G	2023-07-14
146k (1.6M) (MPtrj)	412.5k	EFS_GM	2023-03-03
63k (188.3k) (MPF)	227.5k	EFS_G	2022-09-20
155k (MP 2022)	4.0M	Energy	2023-06-02

Riebesell, Janosh, et al. *(2023) - arXiv:2308.14920* (Accepted at NMI)

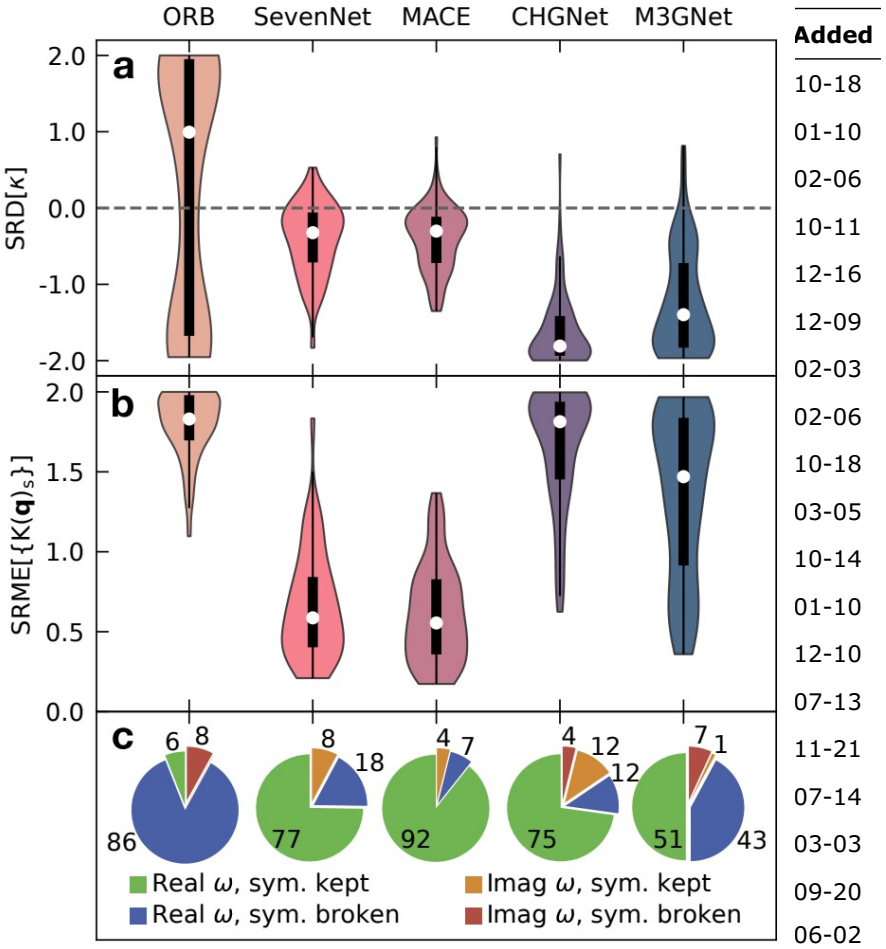
Póta, Balázs, et al. *(2024). - arXiv:2408.00755*.

Matbench Discovery – ML Crystal Stability Predictions

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, Kristin A. Persson

Stability prediction is not correlated with lattice thermal conductivity (phonon)

Model	F1 ↑	DAF ↑	Prec ↑	Acc ↑	TPR ↑	TNR ↑	MAE ↓	RMSE ↓	R^2 ↑	κ_{SRME} ↓	Added
eqV2 M	0.917	6.047	0.924	0.975	0.910	0.986	0.020	0.072	0.848	1.771	10-18
DPA3-v1-OpenLAM	0.883	5.754	0.880	0.963	0.885	0.978	0.023	0.067	0.869	0.741	01-10
GRACE-2L-OAM	0.880	5.774	0.883	0.963	0.878	0.979	0.023	0.068	0.862	0.294	02-06
ORB	0.880	6.041	0.924	0.965	0.841	0.987	0.028	0.077	0.824	1.732	10-11
MatterSim v1 5M	0.862	5.852	0.895	0.959	0.831	0.982	0.024	0.068	0.863	0.574	12-16
MACE-MPA-0	0.852	5.582	0.853	0.954	0.851	0.973	0.028	0.073	0.842	0.412	12-09
GNoME	0.829	5.523	0.844	0.955	0.814	0.972	0.035	0.085	0.785		02-03
GRACE-1L-OAM	0.824	5.255	0.803	0.944	0.846	0.962	0.031	0.073	0.842	0.516	02-06
eqV2 S DeNS	0.815	5.042	0.771	0.941	0.864	0.953	0.036	0.085	0.788	1.676	10-18
AlphaNet-MPTrj	0.799	4.863	0.743	0.933	0.864	0.945	0.041	0.093	0.745	1.310	03-05
ORB MPtrj	0.765	4.702	0.719	0.922	0.817	0.941	0.045	0.091	0.756	1.725	10-14
DPA3-v1-MPtrj	0.765	4.654	0.711	0.921	0.828	0.938	0.042	0.083	0.798	0.964	01-10
SevenNet-l3i5	0.760	4.629	0.708	0.920	0.821	0.938	0.044	0.087	0.776	0.550	12-10
SevenNet-0	0.724	4.252	0.650	0.904	0.818	0.919	0.048	0.092	0.750	0.767	07-13
GRACE-2L-MPtrj	0.691	4.163	0.636	0.896	0.757	0.921	0.052	0.094	0.741	0.525	11-21
MACE-MP-0	0.669	3.777	0.577	0.878	0.796	0.893	0.057	0.101	0.697	0.647	07-14
CHGNet	0.613	3.361	0.514	0.851	0.758	0.868	0.063	0.103	0.689	1.717	03-03
M3GNet	0.569	2.882	0.441	0.813	0.803	0.813	0.075	0.118	0.585	1.412	09-20
ALIGNN	0.567	3.206	0.490	0.841	0.672	0.872	0.093	0.154	0.297		06-02



Riebesell, Janosh, et al. *(2023) - arXiv:2308.14920* (Accepted at NMI)

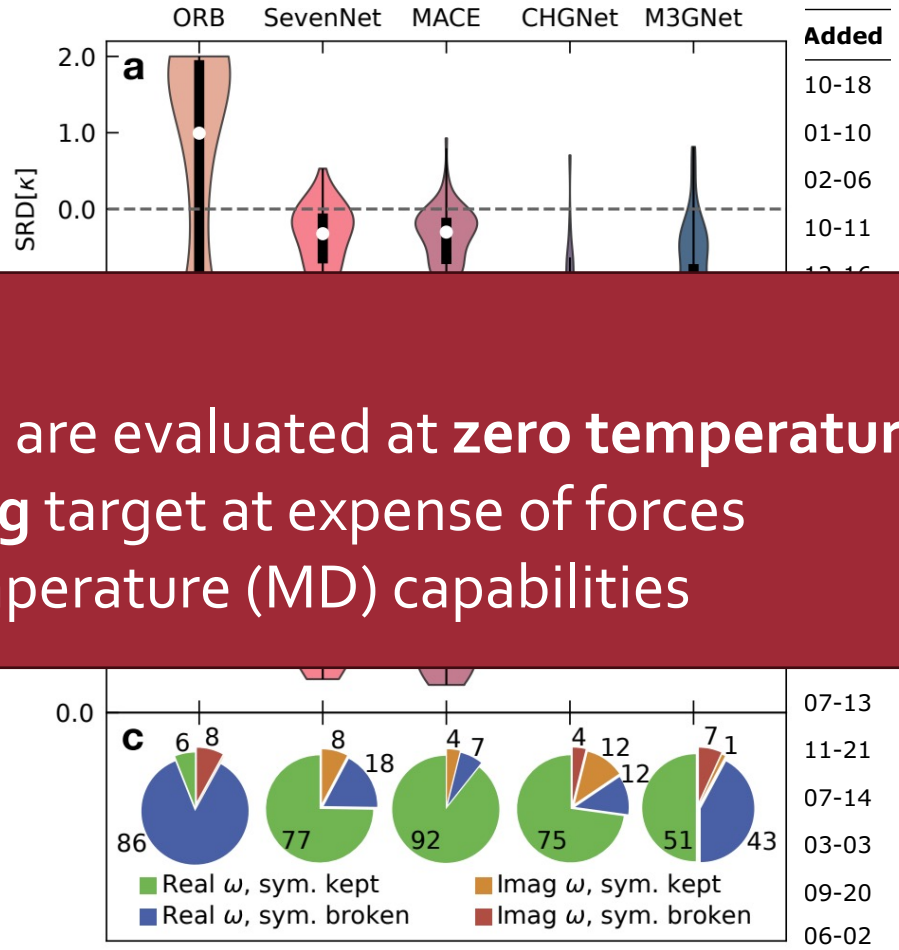
Póta, Balázs, et al. *(2024). - arXiv:2408.00755*.

Matbench Discovery – ML Crystal Stability Predictions

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta,
Alpha A. Lee, Anubhav Jain, Kristin A. Persson

Stability prediction is not correlated with lattice thermal conductivity (phonon)

Model	F1 ↑	DAF ↑	Prec ↑	Acc ↑	TPR ↑	TNR ↑	MAE ↓	RMSE ↓	R^2 ↑	κ_{SRME} ↓	Added
eqV2 M	0.917	6.047	0.924	0.975	0.910	0.986	0.020	0.072	0.848	1.771	10-18
DPA3-v1-OpenLAM	0.883	5.754	0.880	0.963	0.885	0.978	0.023	0.067	0.869	0.741	01-10
GRACE-2L-OAM	0.880	5.774	0.883	0.963	0.878	0.979	0.023	0.068	0.862	0.294	02-06
ORB	0.880	6.041	0.924	0.965	0.841	0.987	0.028	0.077	0.824	1.732	10-11
SevenNet-0	0.724	4.252	0.650	0.904	0.818	0.919	0.048	0.092	0.750	0.767	07-13
GRACE-2L-MPtrj	0.691	4.163	0.636	0.896	0.757	0.921	0.052	0.094	0.741	0.525	11-21
MACE-MP-0	0.669	3.777	0.577	0.878	0.796	0.893	0.057	0.101	0.697	0.647	07-14
CHGNet	0.613	3.361	0.514	0.851	0.758	0.868	0.063	0.103	0.689	1.717	03-03
M3GNet	0.569	2.882	0.441	0.813	0.803	0.813	0.075	0.118	0.585	1.412	09-20
ALIGNN	0.567	3.206	0.490	0.841	0.672	0.872	0.093	0.154	0.297		06-02



⚠

- Test set structures are **near equilibrium** and are evaluated at **zero temperature**
- Energy-only benchmark becomes **overfitting** target at expense of forces (phonons), symmetries, and other finite temperature (MD) capabilities

Riebesell, Janosh, et al. *(2023) - arXiv:2308.14920* (Accepted at NMI)

Póta, Balázs, et al. *(2024). - arXiv:2408.00755*.

Goodhart's law

"When a measure becomes a target, it ceases to be a good measure”

Charles Goodhart (1975)

"When a benchmark becomes the training objective function,
it ceases to be a good benchmark”

Community Challenges around MLIP Benchmarks

- Error-based regression metrics (e.g. Energy, Force MAEs)
  - ⚠ Not correlate well with *physical soundness* (symmetry, conservation law, smoothness …)
  - ⚠ Fail to be the good indicators of generalizability and robustness for *real-world applications*
  - ⚠ Could widen the gap with predicted properties and experimental *observables*
- Static DFT reference benchmarks
  - ⚠ Are subject to data leakage
  - ⚠ Become outdated quickly with newer, larger datasets released
  - ⚠ Lack cross-comparability for models trained on different datasets and levels of theory

MLIP Arena

- Evaluates the physical soundness in wider aspects
- Focuses on the readiness for downstream applications
- Promotes benchmarks comparable across architectures and training sets

MLIP Arena Benchmarks

What are the general behaviors and qualities the MLIPs should have?

Hugging Face online leaderboard:
huggingface.co/spaces/atomind/mlip-arena

⚔️ MLIP Arena ⚔️

1. Static Interactions

- Diatomic potential energy curves
- Force rotational equivariance

2. Molecular Dynamics

- Reactivity: Hydrogen combustion
- Stability: Hot compression on random mixture

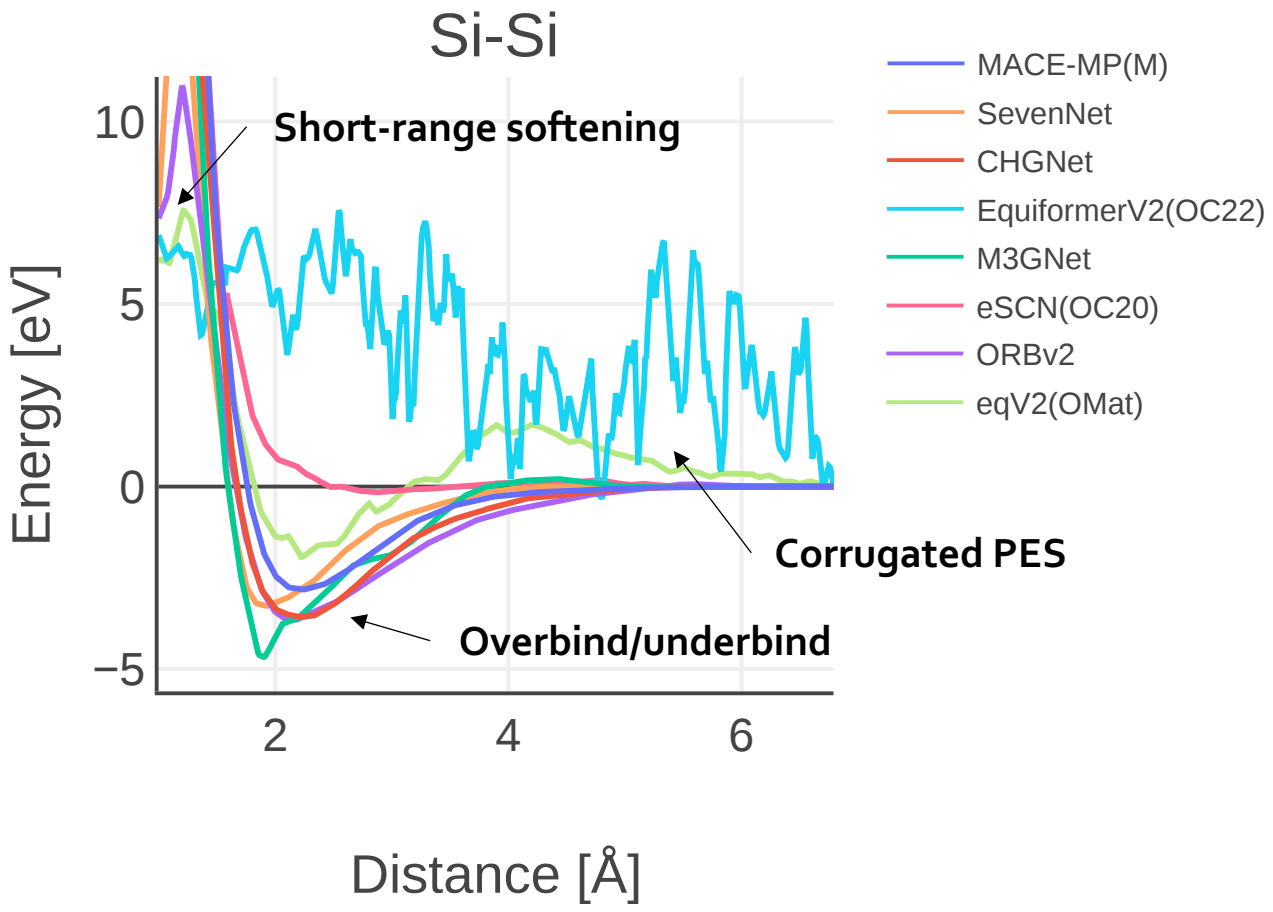
3. Thermodynamic properties and phenomena

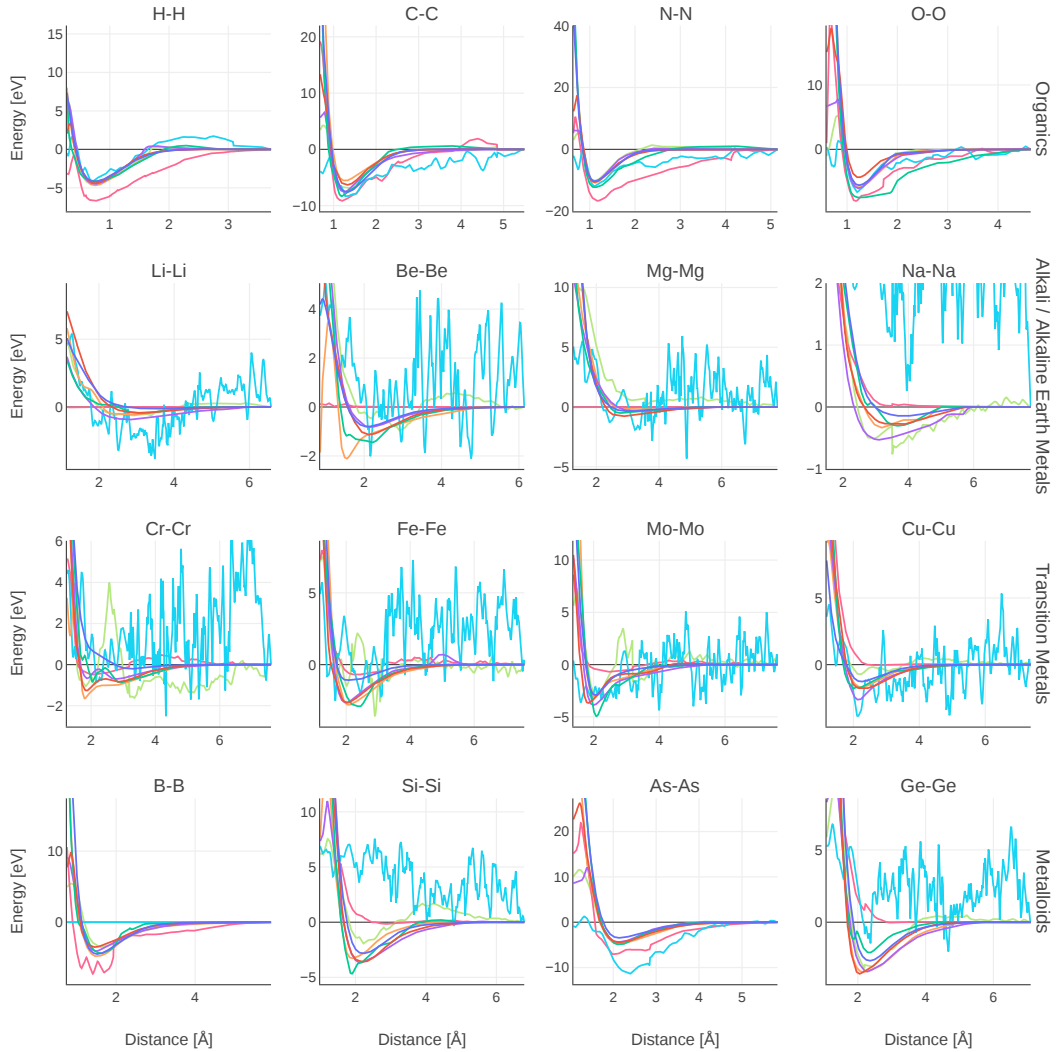
- Vacancy formation and migration
- Structural phase transition in Perovskite

🎯 Physical Soundness 🎯 Real-World Relevance
📊 Pairwise Rank Aggregation 🔄 Reproducible Workflow 🤗 Online Leaderboard

1. Static Interactions

Static Interactions – Homonuclear Diatomic PECs





Homonuclear Diatomic PECs – Rank Aggregation as Pairwise Win Rate

📊 Pairwise rank aggregation from multiple metrics → less chance for overfitting

Model	Rank	Rank aggregation	Conservation deviation [eV/Å]	Spearman's coefficient E: repulsion	Spearman's coefficient F: descending	Energy jump [eV]	Force flips	Tortuosity
MACE-MP(M)	1	12	0.070	**-0.997**	-0.980	0.038	1.449	1.161
MatterSim	2	16	**0.013**	-0.980	-0.972	**0.008**	2.766	1.021
M3GNet	3	19	0.026	-0.991	-0.947	0.029	3.528	**1.016**
eSCN(OC20)	4	27	2.045	-0.939	-0.984	0.806	0.640	5.335
ORBv2	4	27	9.751	-0.883	**-0.988**	0.991	**0.991**	1.287
CHGNet	6	29	1.066	-0.992	-0.925	0.291	2.255	2.279
SevenNet	7	35	b) 34.005	-0.986	-0.928	0.392	2.112	1.292
ORB	8	36	10.220	-0.881	-0.954	1.019	1.026	1.798
a) eqV2(OMat)	9	48	15.477	-0.880	-0.976	4.118	3.126	2.515
ALIGNN	10	53	5.164	-0.913	-0.310	9.876	30.669	1.818
EquiformerV2(OC20)	11	64	21.385	-0.680	-0.891	38.282	22.775	8.669
EquiformerV2(OC22)	12	69	27.687	-0.415	-0.855	64.837	21.674	15.880

a) Top Matbench-Discovery and Open Catalyst Project models don't necessarily capture correct diatomic interaction (corrugated PECs). Model design and training dataset matter as well.

b) Gradient force models can be non-conservative in diatomic interaction

Static Interactions – Rotational Equivariance

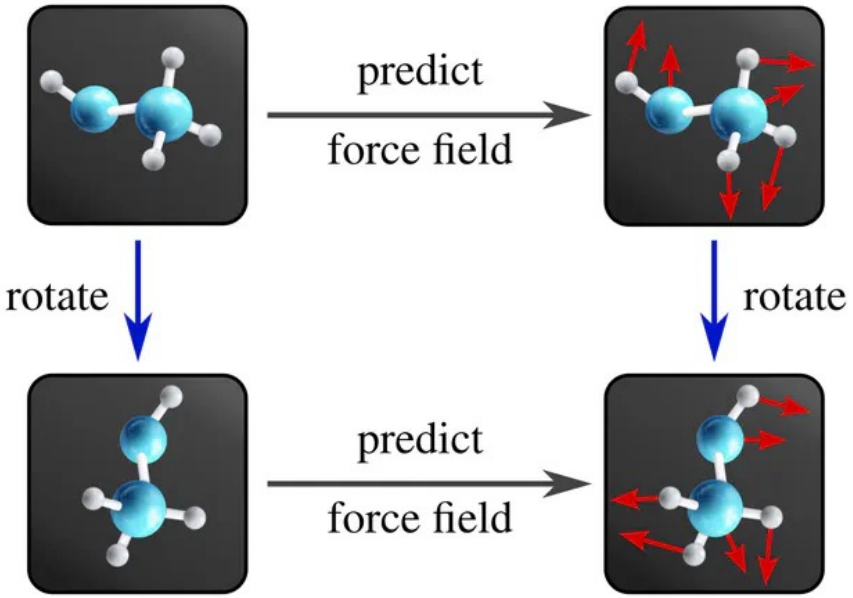
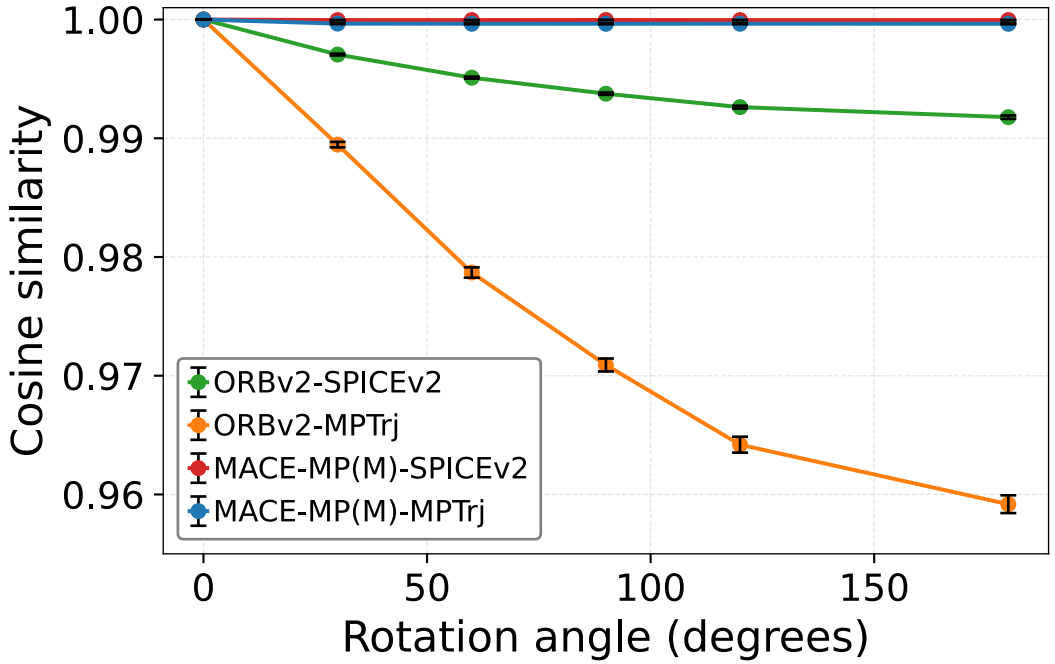


Figure credit: Maurice Weiler

$$\mathrm{sim}(\mathbf{F}) = \frac{\mathbf{RF}(\mathbf{r}) \cdot \mathbf{F}(\mathbf{Rr})}{\|\mathbf{RF}(\mathbf{r})\| \|\mathbf{F}(\mathbf{Rr})\|}$$



Rotational equivariance is relatively easier to learn w/o hard constraints

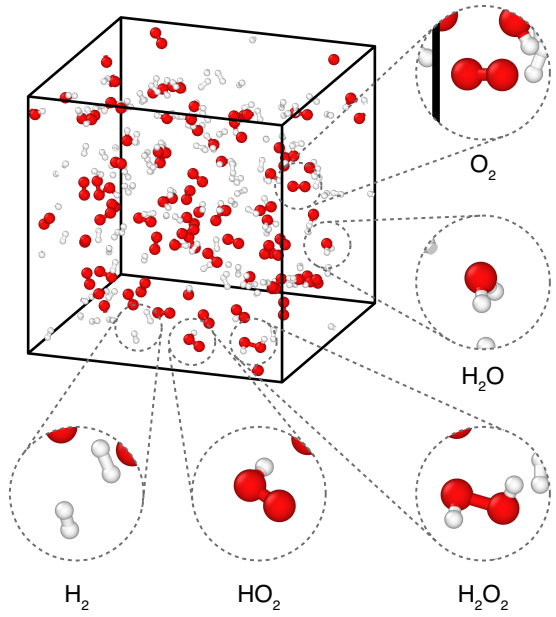
(Learning equivariance from data can reach high quality, e.g. EScAIP achieves >0.999 sim)

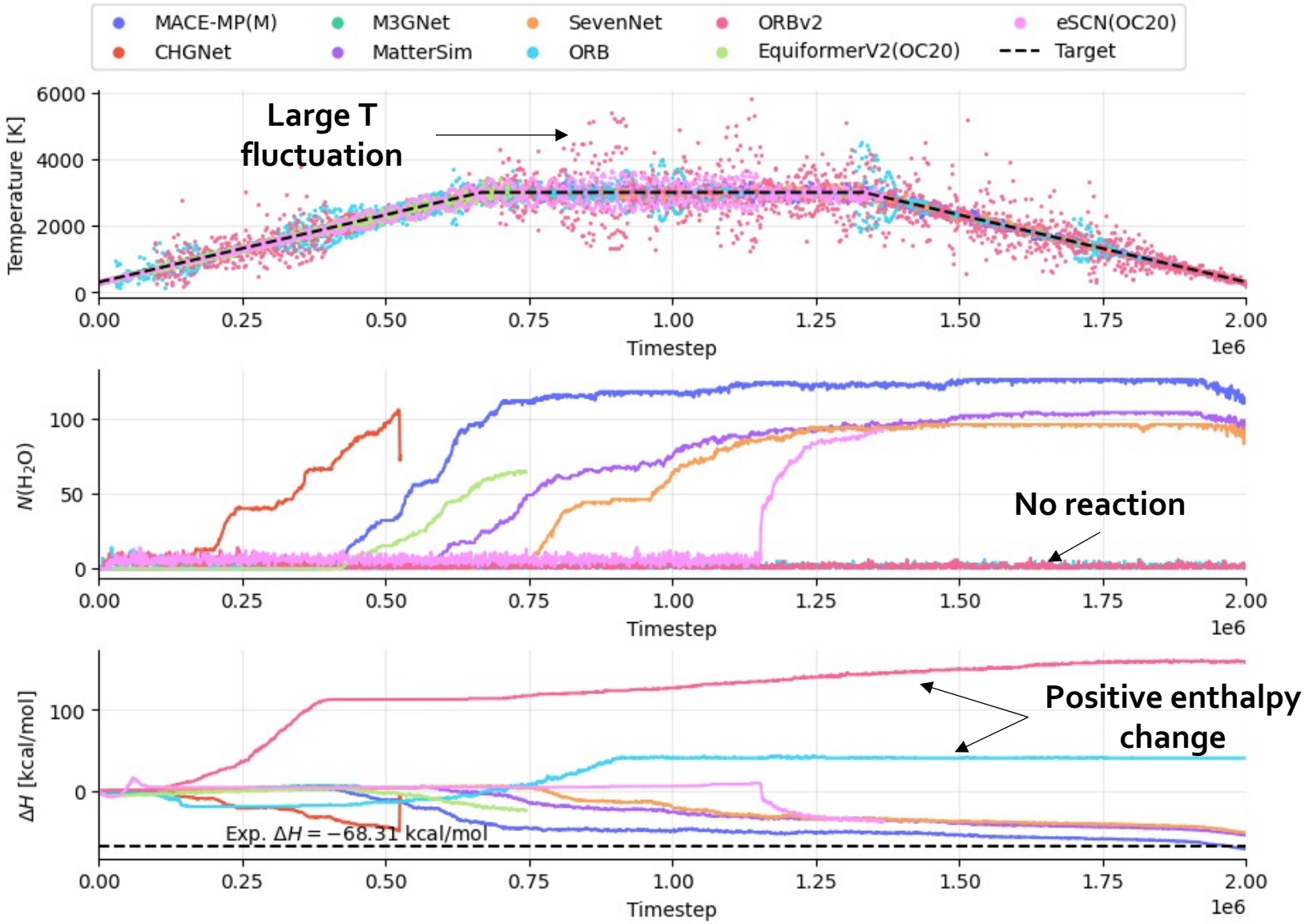
Qu, Eric, and Aditi Krishnapriyan. *NeurIPS* (2024)

2. Molecular Dynamics

MD Stability and Reactivity – Hydrogen Combustion

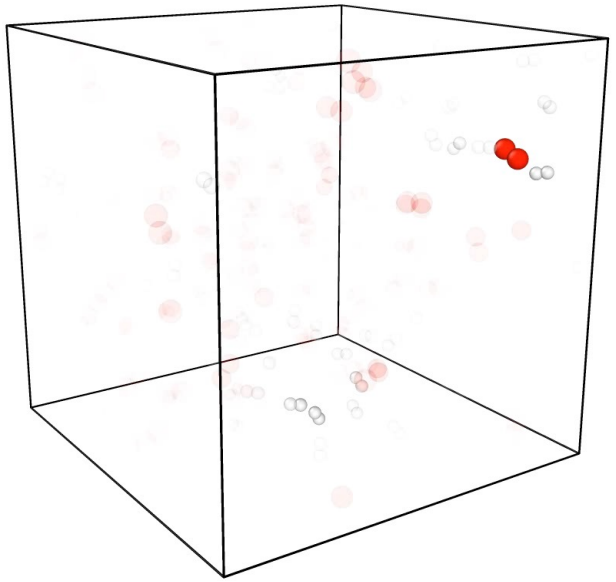
$$128H_2 + 64O_2 \rightarrow 128H_2O$$



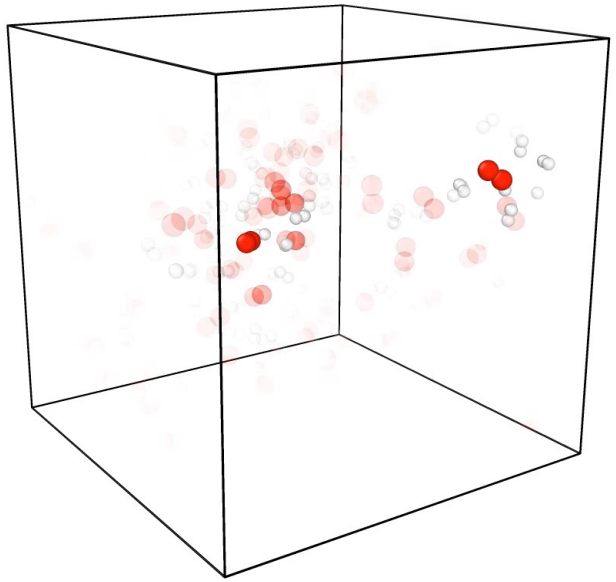


MD Stability and Reactivity – Hydrogen Combustion

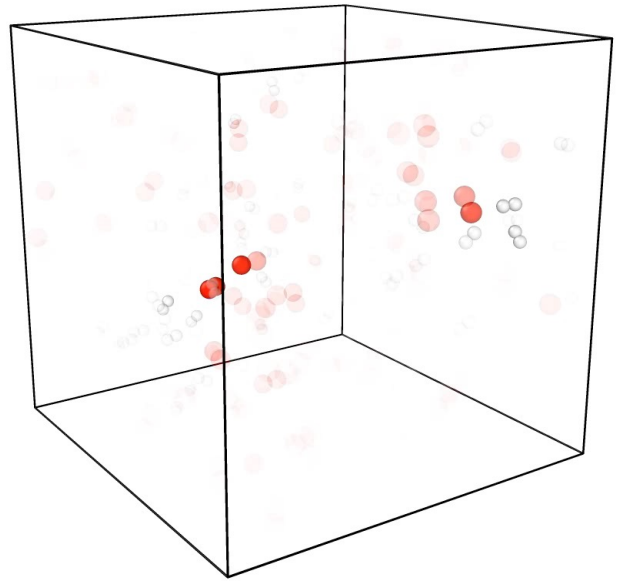
MACE-MP



CHGNet



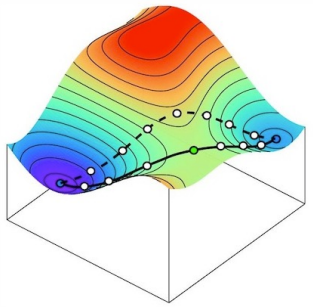
ORBv2



- Hydrogen Combustion is a challenging out-of-distribution test for MLIP reactivities
  - Annealing MD for 1 ns reveal subtle failure modes of MLIP

3. Thermodynamic Properties & Physical Phenomena

Vacancy Migration Energy in Elemental FCC/HCP Solids

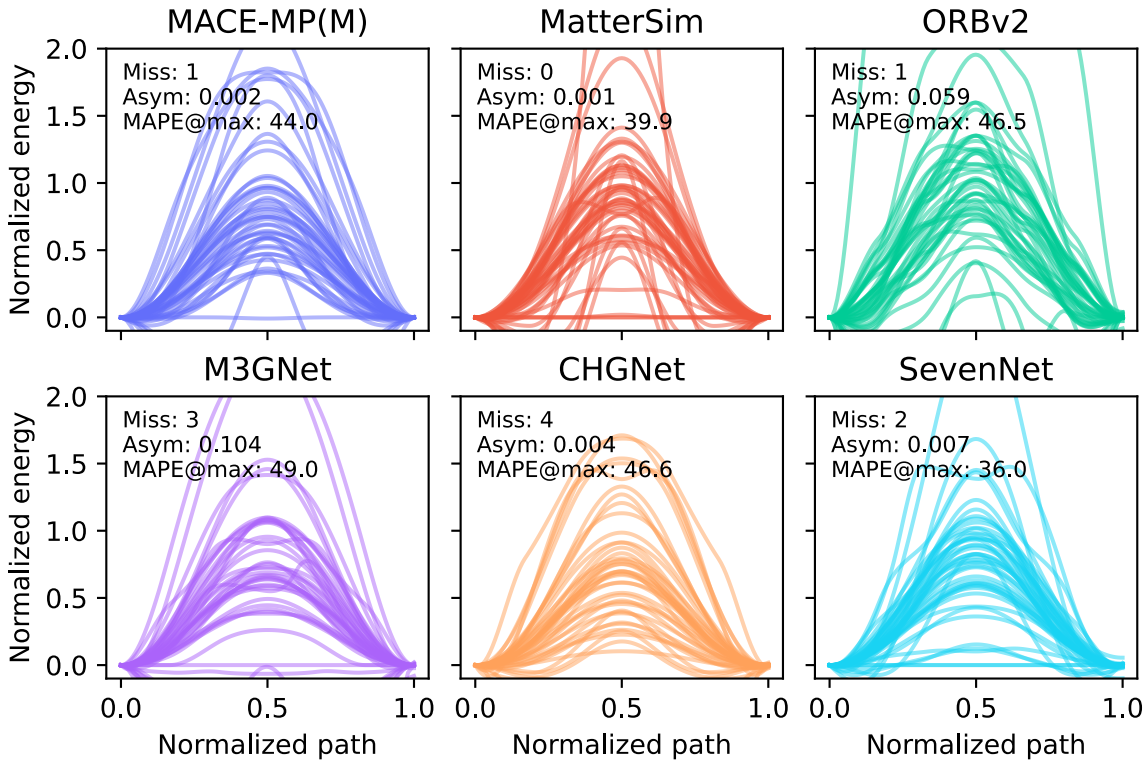


Climbing Image
Nudged Elastic Band
(CI-NEB)

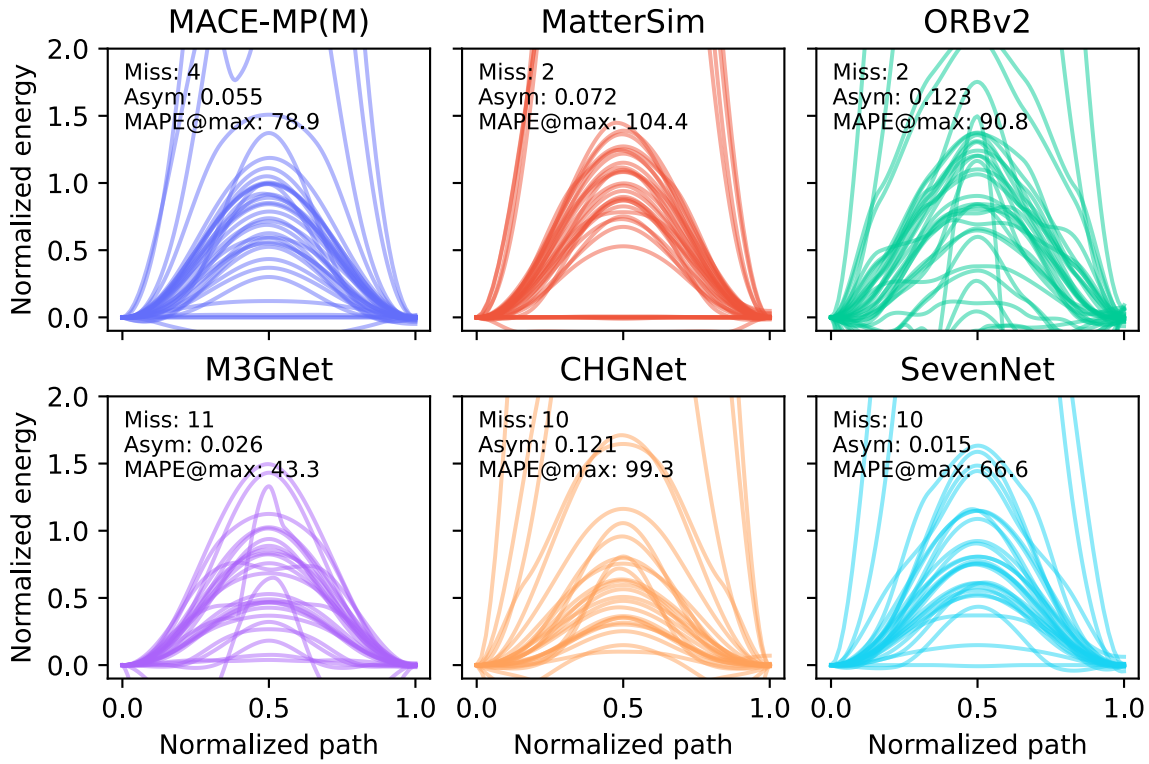
$$\text{path asymmetry} = 2\int_0^{0.5} |\epsilon(0.5-x) - \epsilon(0.5+x)|\, dx$$

$$\text{barrier asymmetry} = \frac{\Delta E}{E_{\text{forward}}} = \frac{E_f - E_i}{E_{\text{TS}} - E_i}$$

Elemental FCC solids

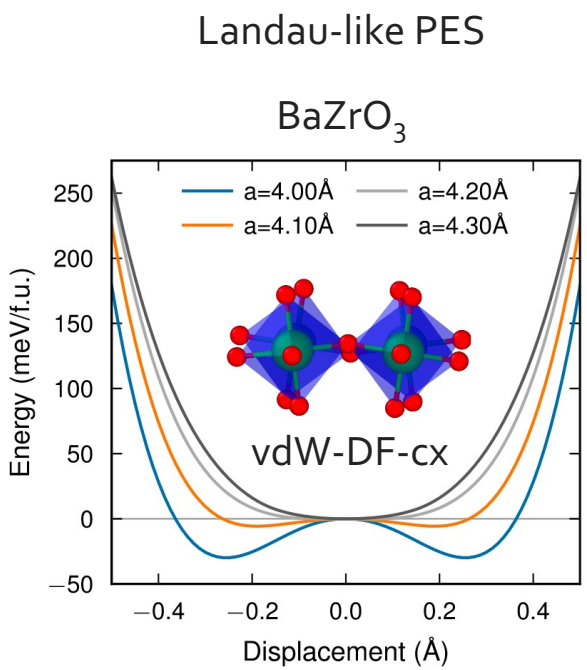


Elemental HCP solids (0001)

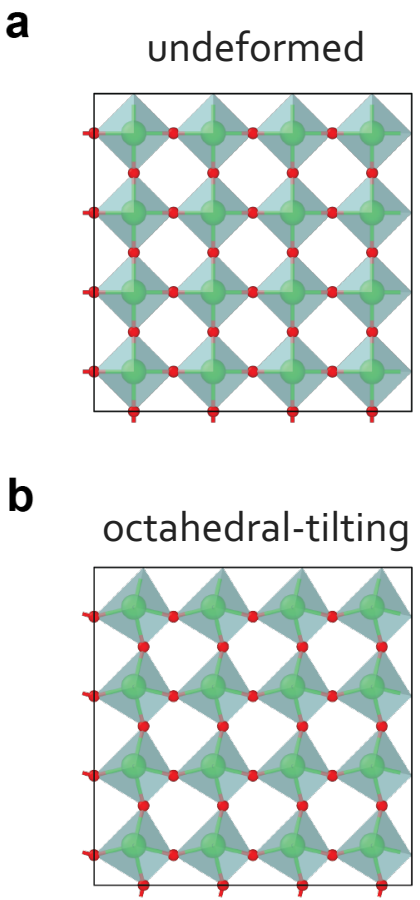


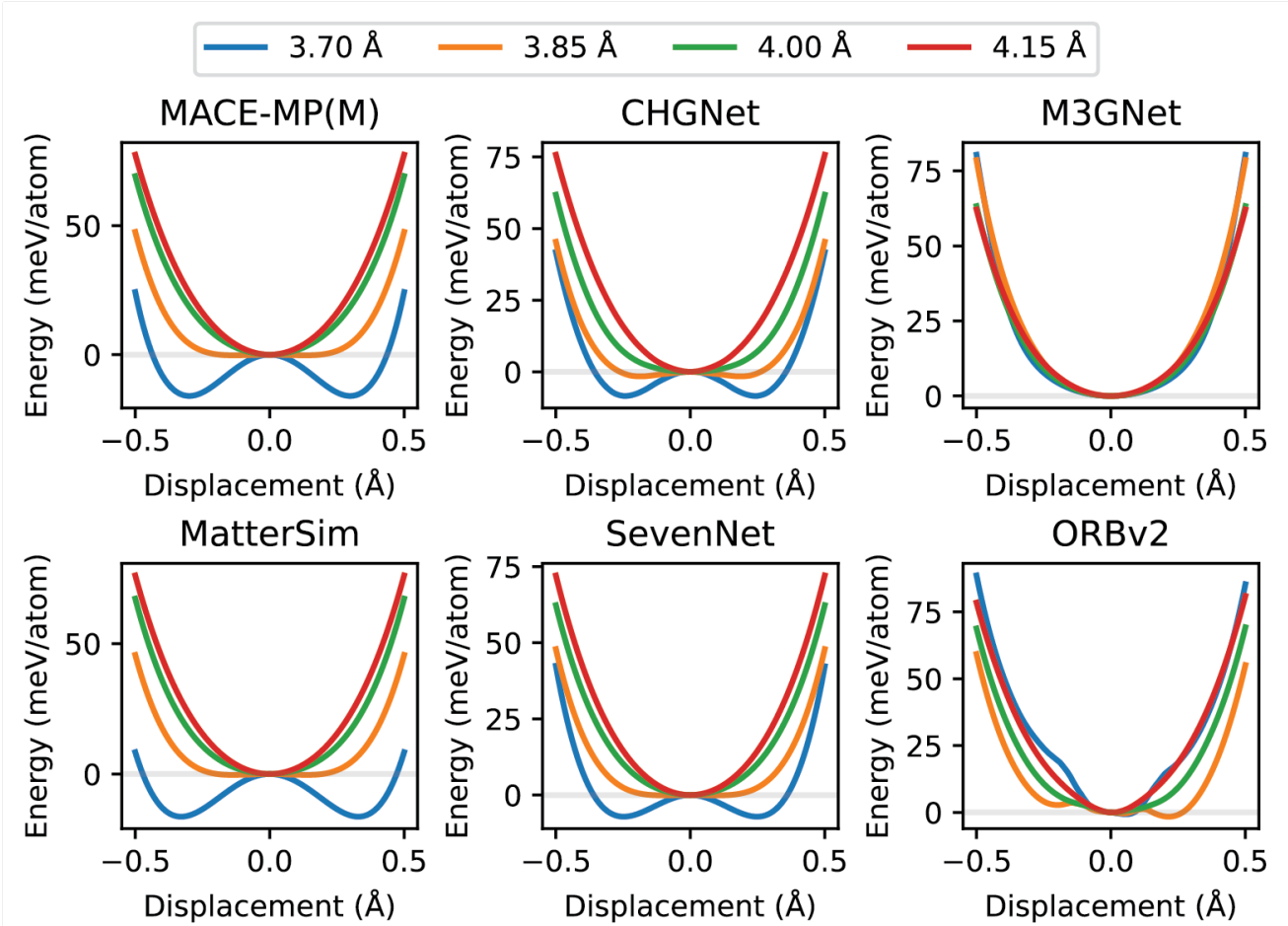
Angsten, Thomas, et al. *New Journal of Physics* 16.1 (2014): 015018.

Displacive Phase Transition in Perovskite



Fransson, Erik, et al. *Chemistry of Materials* 36.1 (2023): 514-523.





- Crystal symmetry and phonon reveal subtle MLIP failure modes

Acknowledgements

Collaborators

Tobias Kreiman, Elizabeth Weaver, Ishan Amin, Matthew Kuner, Christine Zhang, Aaron Kaplan, Daryl Chrzan, Samuel Blau, Aditi Krishnapriyan, Mark Asta

Early feedback and discussion

Janosh Riebesell, Philipp Benner, Patrick Huck, Rouxi Yang, Evan Walter Clark Spotte-Smith, Bowen Deng

MOF benchmark (coming)

Hyunsoo Park, Yunsung Lim

Special thanks

Prof. Mark Asta (PhD advisor)

Asta and Chrzan groups @ UC Berkeley

Prof. Aron Walsh @ ICL (generous comments)

Funding and compute

Hugging Face online leaderboard:
huggingface.co/spaces/atomind/mlip-arena

GitHub repo:
https://github.com/atomind-ai/mlip-arena

@cyrusyc_tw