

MLIP Arena

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

Yuan (Cyrus) Chiang

(cyrusyc@berkeley.edu)

University of California, Berkeley

Lawrence Berkeley National Laboratory

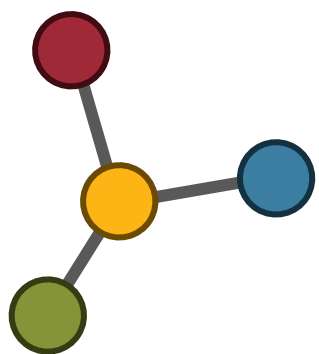
Tobias Kreiman, Christine Zhang, Matthew Kuner, Elizabeth Weaver, Ishan Amin, Hyunsoo Park, Yunsung Lim, Jiham Kim, Daryl Chrzan, Aron Walsh, Samuel M. Blau, Mark Asta, Aditi S. Krishnapriyan

2025 MRS Fall Meeting & Exhibit | NeurIPS 2025 Spotlight

Online leaderboard: <https://huggingface.co/spaces/atomind/mlip-arena>

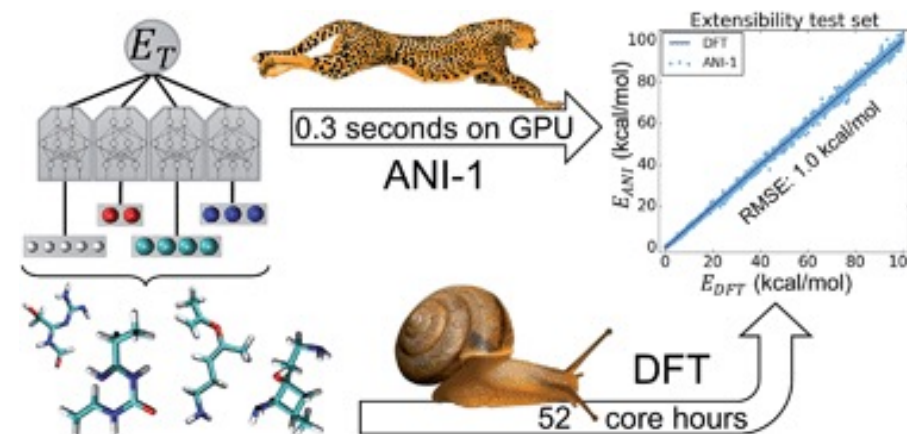
GitHub: <https://github.com/atomind-ai/mlip-arena>

Machine Learning Interatomic Potentials (MLIPs)



Structure $\rightarrow$ Properties

$$\{\mathbf{R}_i\}, \{Z_i\} \rightarrow E, \{\mathbf{F}_i\}, \boldsymbol{\sigma}, \dots$$



Smith, J. S., et al. (2017). *Chemical science*

Kohn-Sham Density Functional Theory (KS-DFT)

$$\left(-\frac{\hbar^2}{2m_e} \nabla^2 + v_{\text{eff}}(\mathbf{r}) \right) \psi_i(\mathbf{r}) = \varepsilon_i \psi_i(\mathbf{r}),$$

$$\rho(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2$$

$$E = \min_{\rho} \{ E^{\text{KS}}[\rho] \}$$

slow, cubic scaling with N

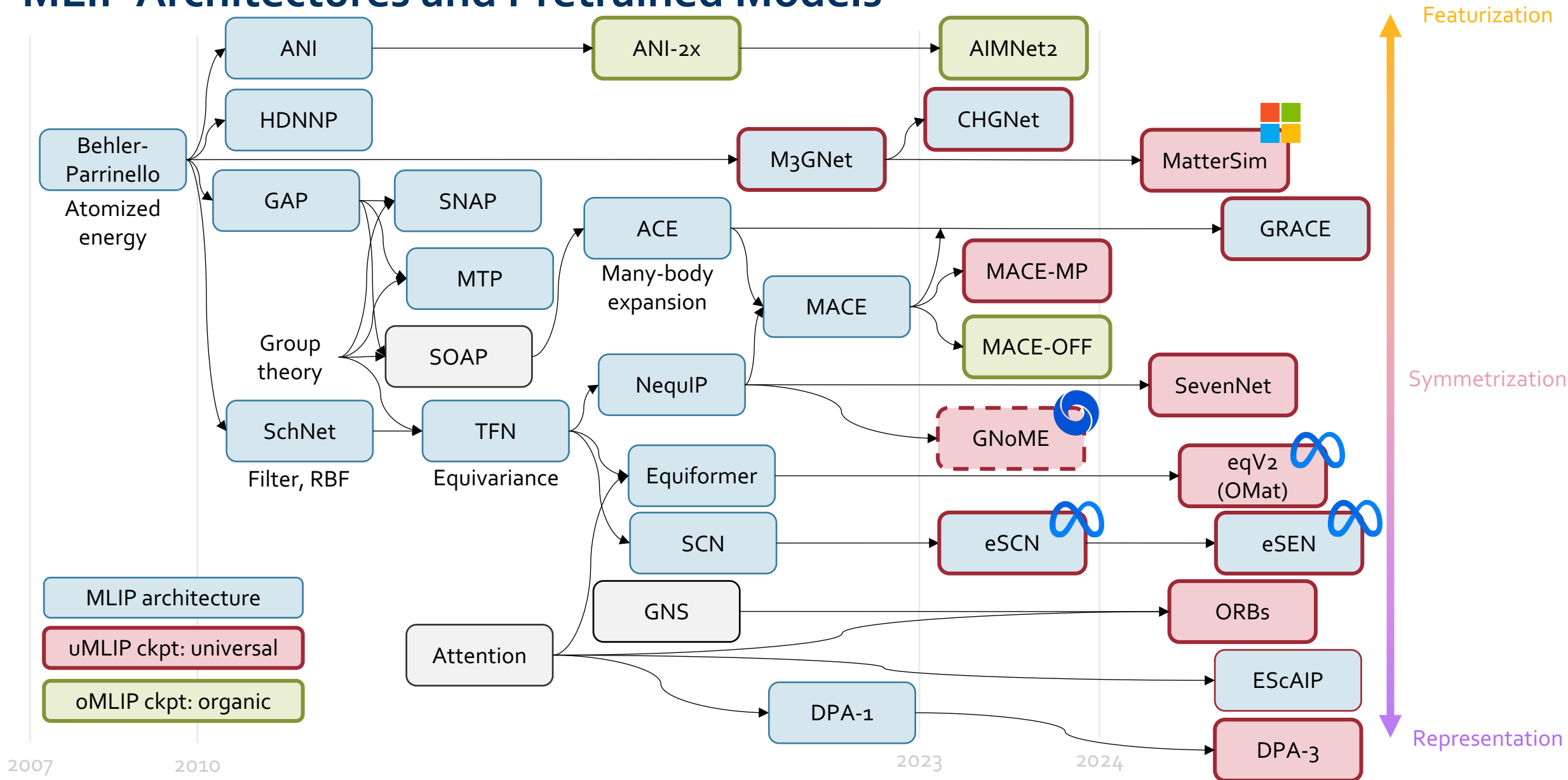
Machine Learning Interatomic Potentials (MLIPs)

$$E = E^{\text{NN}}(\{\mathbf{R}_i\}, \{Z_i\})$$

- Graph Neural Networks
- Message Passing Neural Networks
- Transformers
- ... others

fast, (mostly) linear scaling with N

MLIP Architectures and Pretrained Models



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

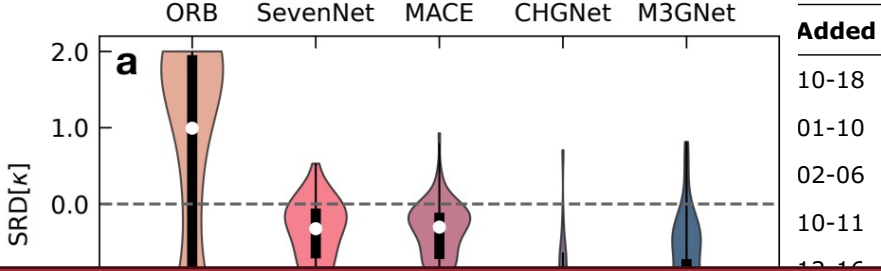
Janosh Riebesell



Rhys Goodall

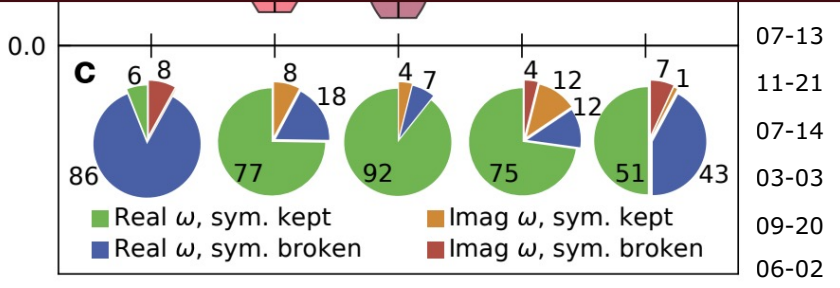


Model	F1 ↑	DAF ↑	Prec ↑	Acc ↑	TPR ↑	TNR ↑	MAE ↓	RMSE ↓	R ² ↑
eqV2 M	0.917	6.047	0.924	0.975	0.910	0.986	0.020	0.072	0.848
DPA3-v1-OpenLAM	0.883	5.754	0.880	0.963	0.885	0.978	0.023	0.067	0.869
GRACE-2L-OAM	0.880	5.774	0.883	0.963	0.878	0.979	0.023	0.068	0.862
ORB	0.880	6.041	0.924	0.965	0.841	0.987	0.028	0.077	0.824
Matbench Simul-1 EM	0.863	5.953	0.885	0.959	0.831	0.983	0.034	0.069	0.863



- 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

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



Riebesell, Janosh, et al. *Nature Machine Intelligence* 7.6 (2025): 836-847.

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)

Community Challenges around MLIP Benchmarks

- Error-based regression metrics (e.g. Energy, Force MAEs)
 - ⚠ Not correlate well with *physical soundness* (symmetry, conservation law, PES smoothness ...)
 - ⚠ Fail to be the good indicators of generalizability and robustness for *real-world applications*
 - ⚠ Could widen the gap between predicted properties and *experimental observables*
- Static DFT reference benchmarks
 - ⚠ Are subject to data leakage
 - ⚠ Become outdated quickly with newer, larger datasets released (OMat, OMol, etc)
 - ⚠ Lack cross-comparability for models trained on different datasets and levels of theory (PBE vs. r2SCAN vs. others)

MLIP Arena

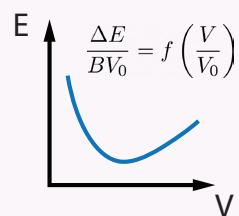
Hugging Face Space NeurIPS Spotlight arXiv 2509.20630 ICLR AI4Mat

build passing pypi v0.1.3 downloads 314/month DOI 10.5281/zenodo.17331903

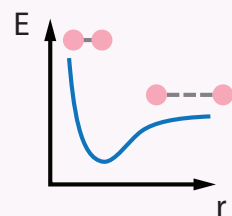
Online leaderboard: <https://huggingface.co/spaces/atomind/mlip-arena>

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

Asymptotic Behaviors



Equation of state



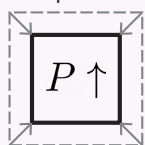
Diatomic curves

Distribution Shifts

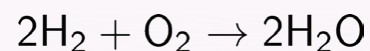
Energy conservation
Differential entropy
 $\delta \mathcal{H}(\mathbf{Y}|\{\mathbf{X}\}) \stackrel{?}{\sim} \delta E(t) \neq 0$
 $\mathcal{R}\mathbf{F}(\mathbf{r}) \neq \mathbf{F}(\mathcal{R}\mathbf{r})$
Force rotational equivariance

Stability and Reactivity

Compression



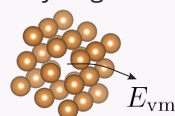
Heating



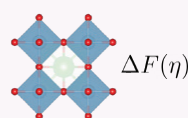
Combustion

Thermodynamic Properties

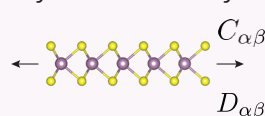
Vacancy migration



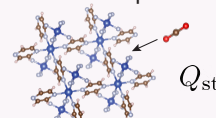
Phase transition



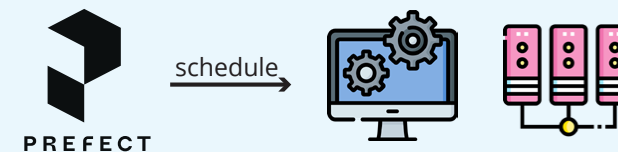
Dynamical stability



Gas adsorption



Workflow Orchestration



Database Storage



Parquet



JSON



ASE DB

Codebase

sync

Leaderboard



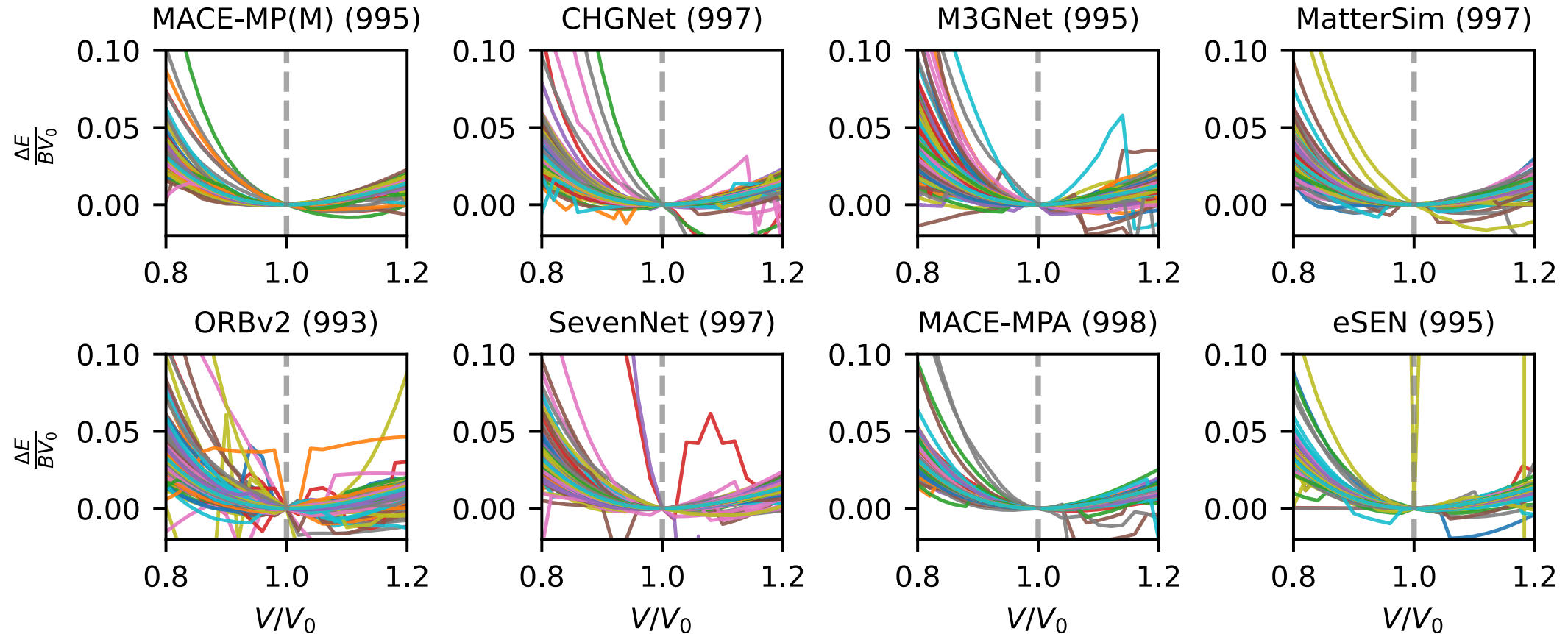
Hugging Face

Model	Training Set	Arena Rank	Meta Rank Agg	Homonuclear diatomics	Equation of state	Energy-volume scans	Stability	Combustion	Prediction	NVT	NPT
MACE-MPA	MPTraj Alexandria	1	9	1	1	2	3	2	EFS	✓	✓
MatterSim	MPTraj Alexandria Proprietary	2	16	3	4	4	2	3	EFS	✓	✓
MACE-MP(M)	MPTraj	3	20	2	3	9	5	1	EFS	✓	✓
CHGNet	MPTraj	4	25	7	5	3	5	5	EFSM	✓	✓

1. Asymptotic analyses on off-equilibrium conditions

Equation of state (EOS)

Birch–Murnaghan EOS
$$\frac{\Delta E}{BV_0} = \frac{E - E_0}{BV_0} = \frac{9}{16} \left[(\eta^2 - 1)^2 (6 + B' (\eta^2 - 1) - 4\eta^2) \right]$$



1,000 structures from WBM dataset

1. Asymptotic analyses on off-equilibrium conditions

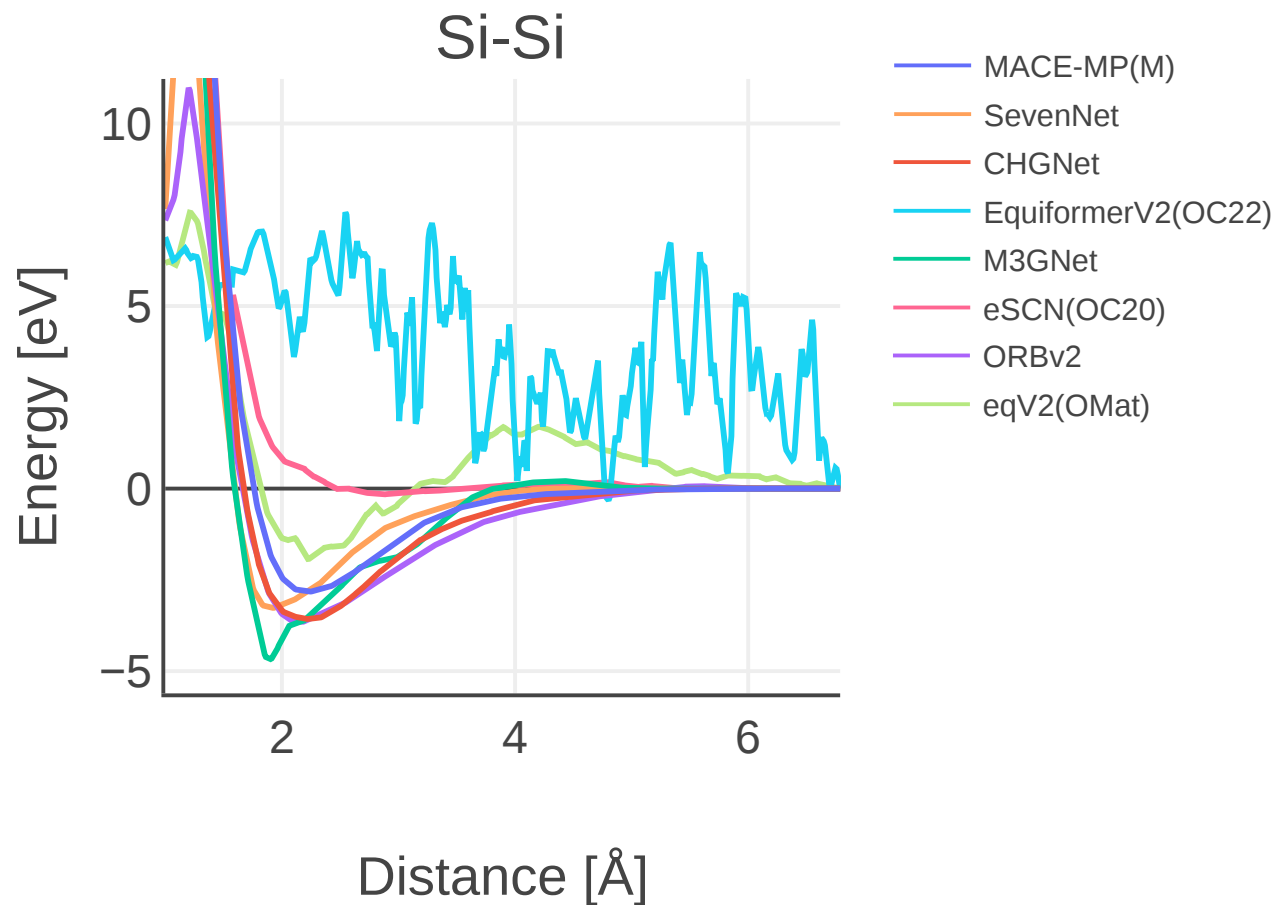
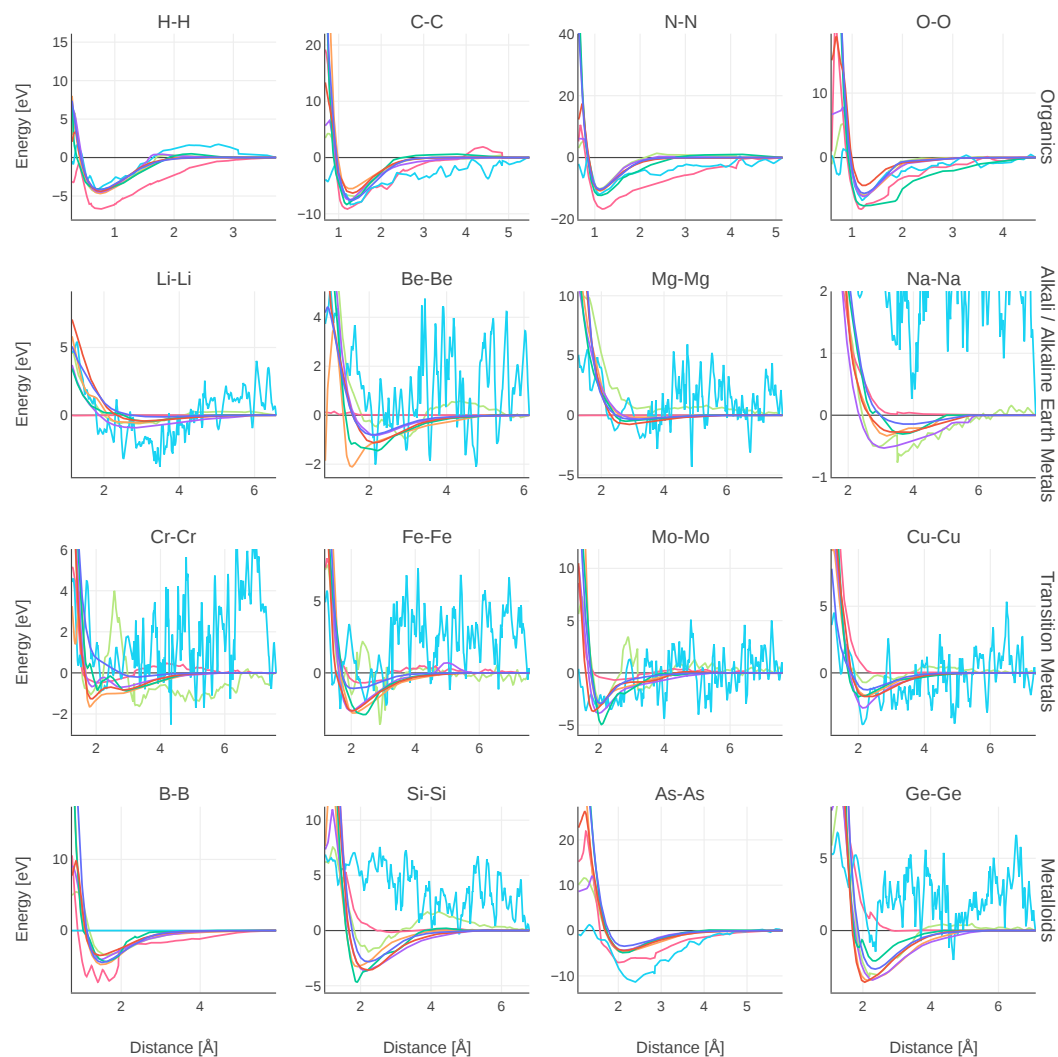
Physical and geometric measures as dataset/model agnostic metrics

Table 1: Equation of state (EOS) benchmark on 1,000 WBM structures [14]. **Boldface** and underline indicate the **best** and worst metrics across all MLIPs, respectively. Standard deviations are given in parentheses. Derivative flips are ranked by their absolute deviation from 1. For up-to-date models and aggregated rankings, see the [online leaderboard](#).

Model	Derivative flips ↓	Tortuosity ↓	Spearman's coefficient			Missing ↓
			E: compression ↓	$\frac{dE}{dV}$: compression ↑	E: tension ↑	
MACE-MPA	1.037 (0.283)	1.005 (0.054)	-0.999368 (0.012)	0.996332 (0.039)	0.993186 (0.077)	2
eSEN	1.042 (0.314)	1.008 (0.090)	-0.999330 (0.012)	0.996857 (0.037)	0.992097 (0.073)	5
MACE-MP(M)	1.042 (0.345)	1.009 (0.129)	-0.999330 (0.011)	0.994116 (0.059)	0.991586 (0.088)	5
MatterSim	1.045 (0.376)	1.006 (0.055)	-0.997350 (0.039)	0.992790 (0.078)	0.988098 (0.115)	3
CHGNet	1.105 (0.540)	1.015 (0.123)	-0.996499 (0.051)	0.992997 (0.052)	0.986642 (0.117)	3
SevenNet	1.109 (0.555)	1.019 (0.275)	-0.998128 (0.026)	0.988912 (0.077)	0.985958 (0.117)	3
M3GNet	1.175 (0.676)	1.018 (0.149)	-0.996321 (0.052)	0.989743 (0.065)	0.980169 (0.133)	5
ORBv2	<u>1.316 (0.870)</u>	<u>1.037 (0.215)</u>	<u>-0.991846 (0.082)</u>	<u>0.970143 (0.132)</u>	<u>0.963746 (0.198)</u>	<u>7</u>

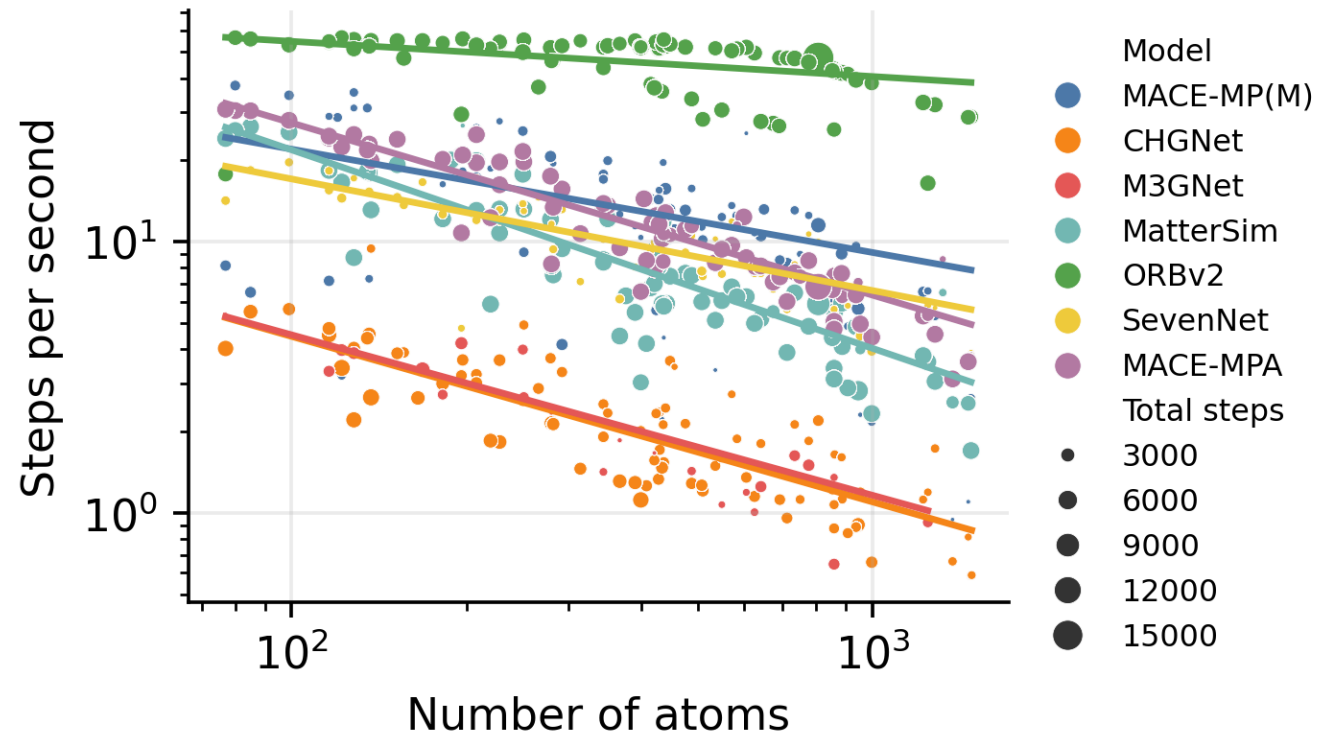
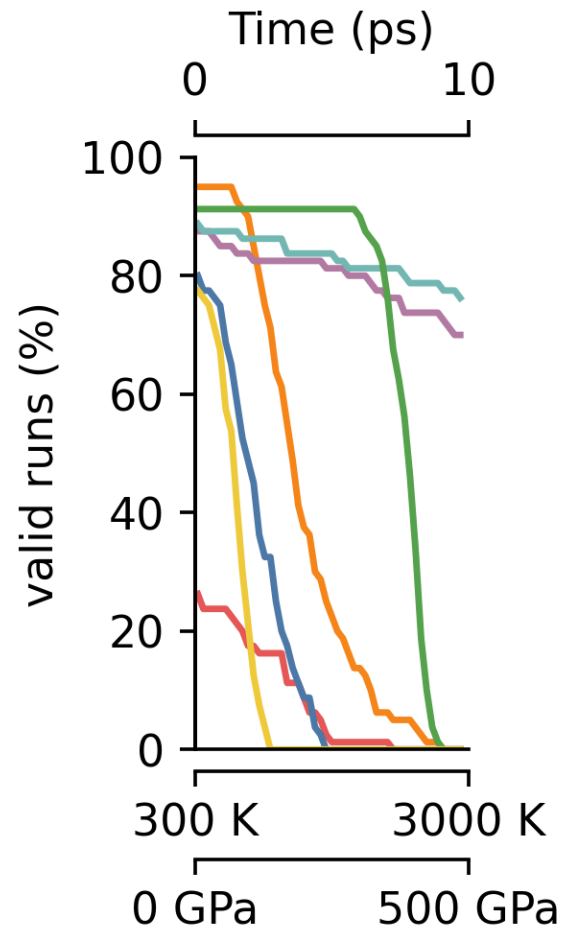
1. Asymptotic analyses on off-equilibrium conditions

Homonuclear Diatomic Potential Energy Curves (PECs)



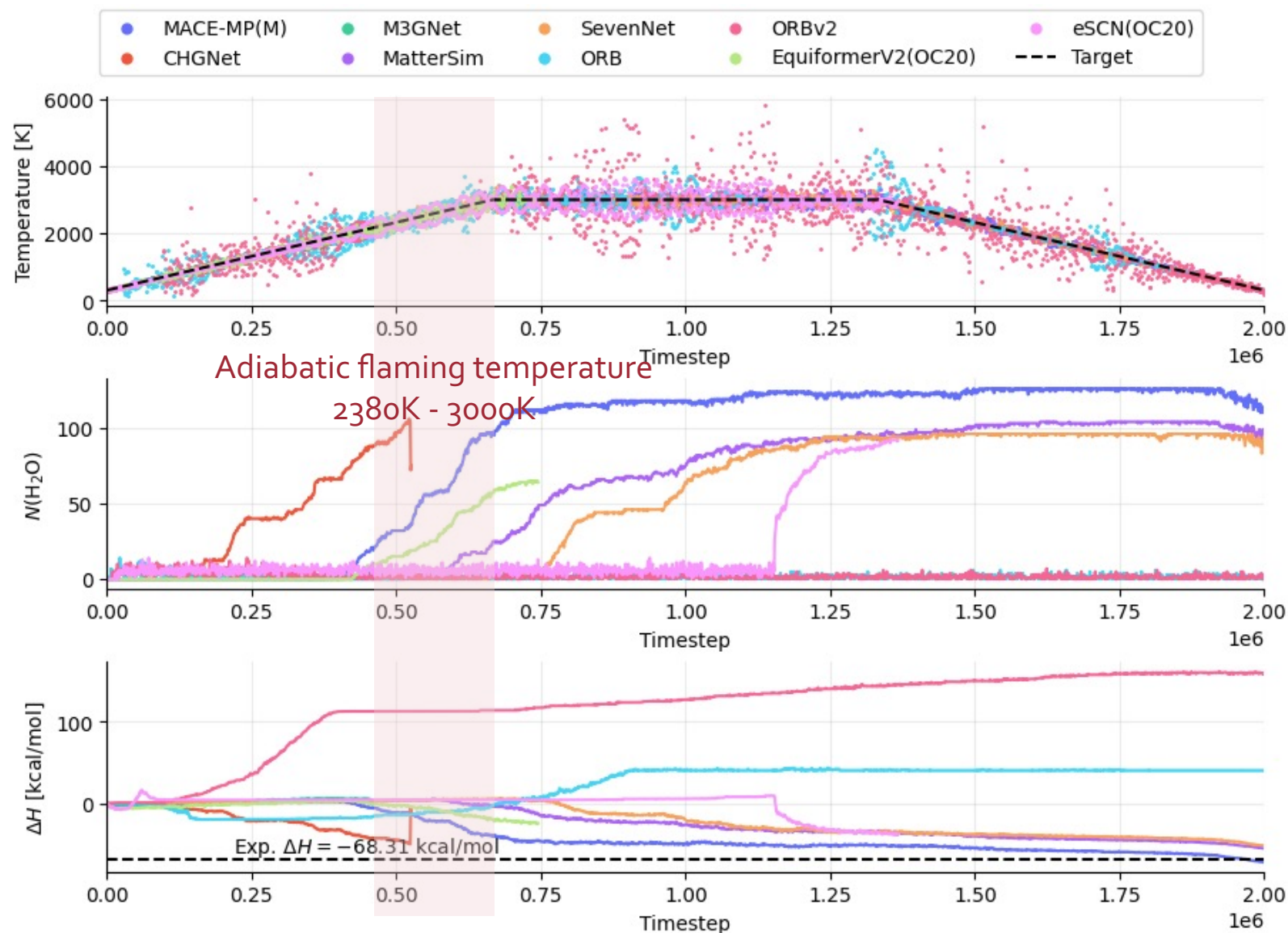
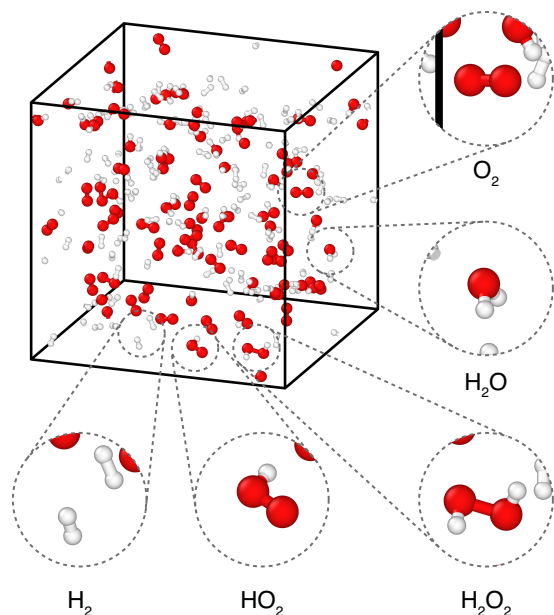
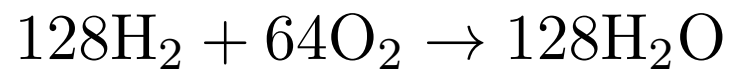
2. Stability and reactivity from molecular dynamics simulation

Hot compression test



2. Stability and reactivity from molecular dynamics simulation

Hydrogen combustion

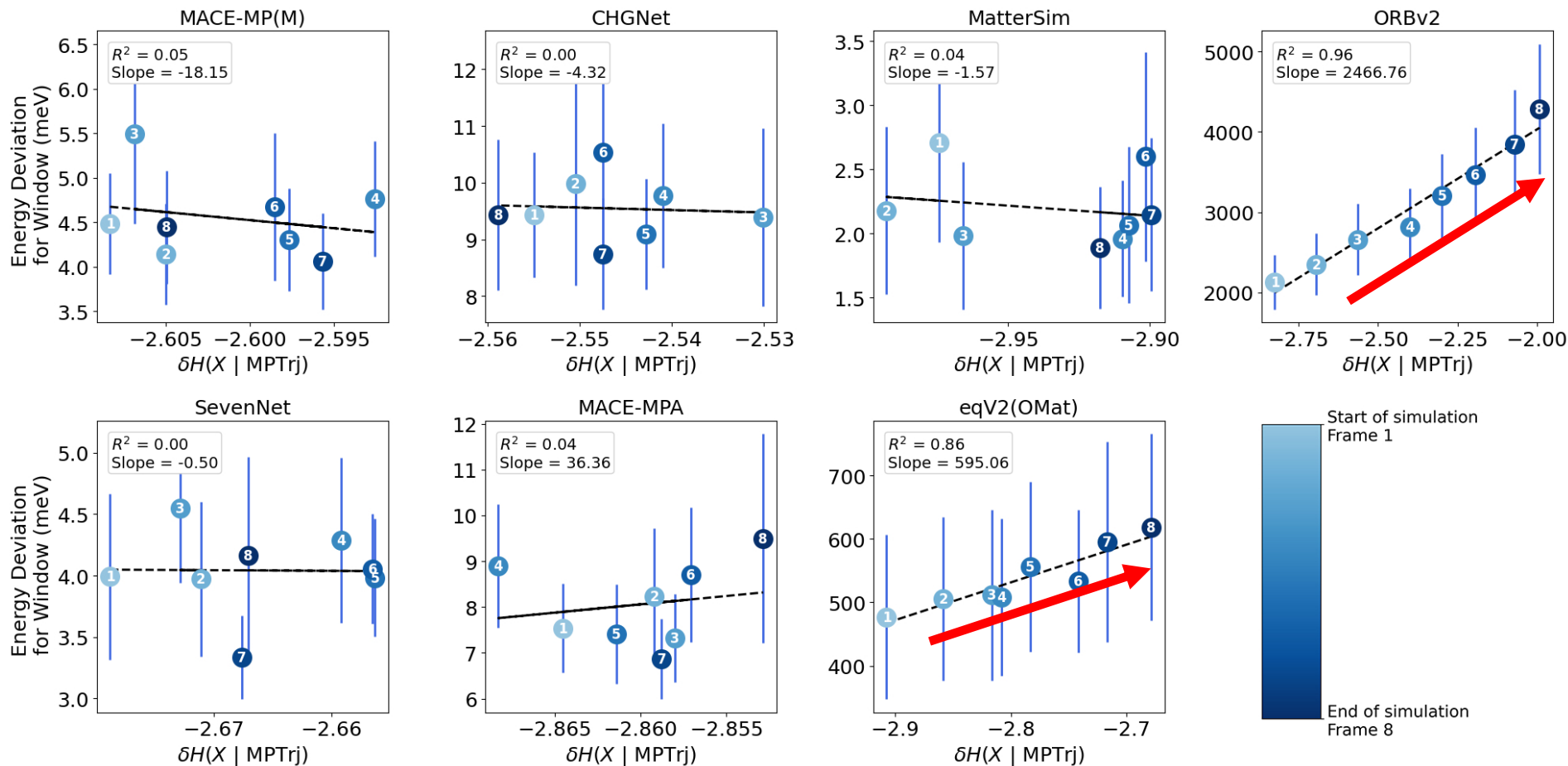


3. Distribution Shift

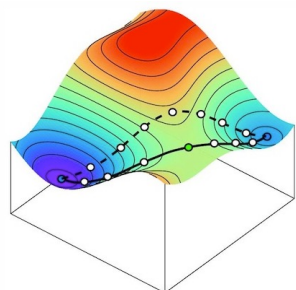
Energy conservation

Differential entropy of
local atomic environment

$$\delta\mathcal{H}(\mathbf{X}|\text{MPTrj}) \stackrel{?}{\rightarrow} \delta E(t) \neq 0$$

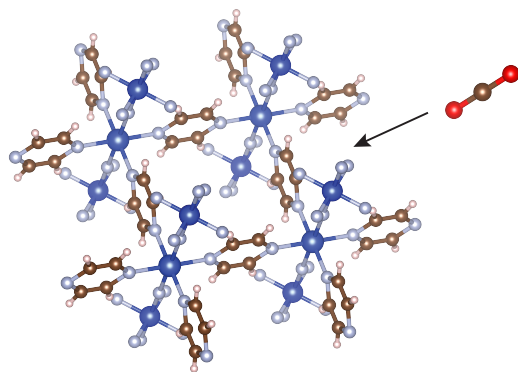


4. Thermodynamic Properties

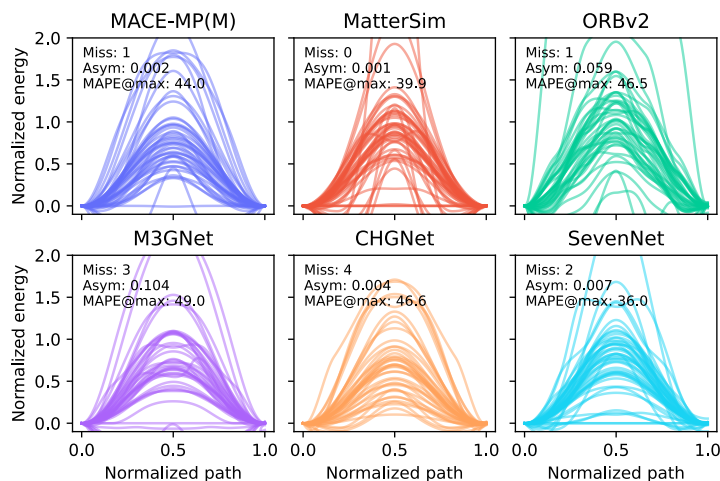


Vacancy Migration in Elemental FCC solids

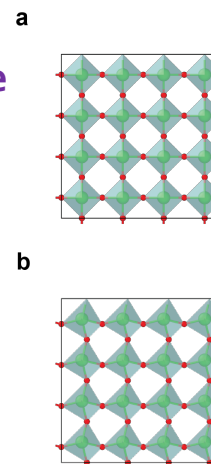
CO₂ Adsorption in Metal-Organic Frameworks (MOFs)



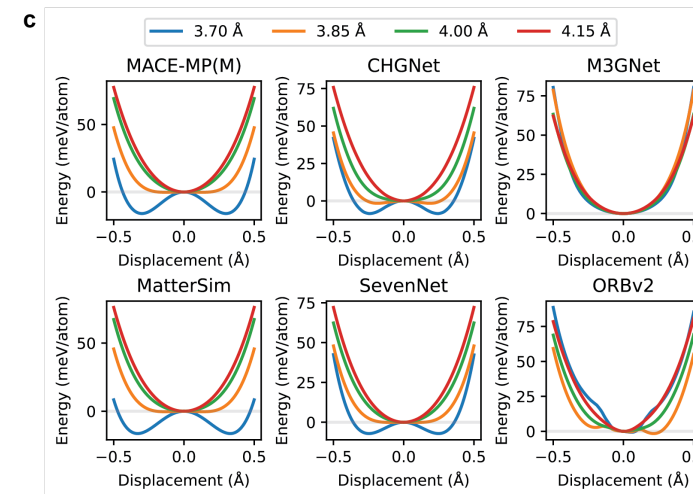
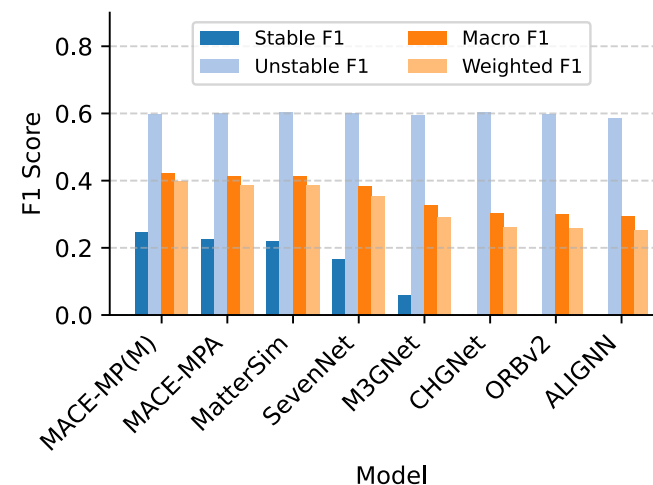
Climbing Image Nudged Elastic Band (CI-NEB)



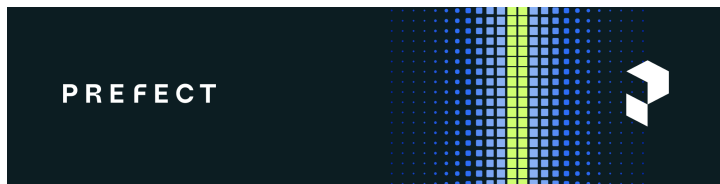
Landau-like Structural Phase Transition



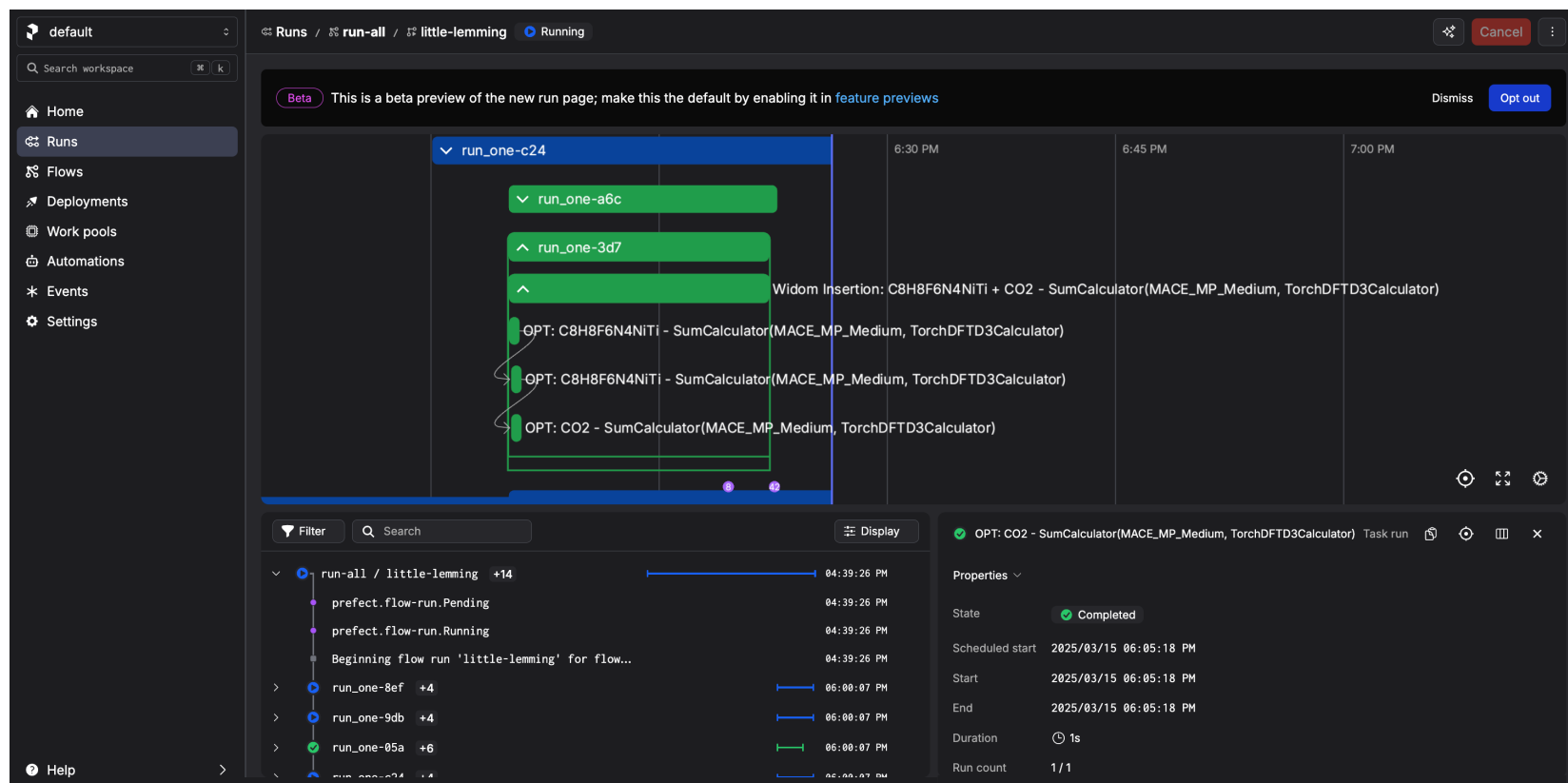
Dynamical Stability of 2D Materials



MLIP Arena Performs Benchmarks at Scale on HPC



- HPC workflow orchestration
- Task chaining/caching



Reusable Modular Prefect Tasks

- Structure optimization
- Equation of state
- Molecular dynamics (advanced T/P scheduling)
- Phonon
- Nudged elastic band
- Elasticity

Acknowledgements

Contributors

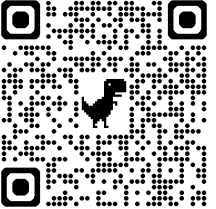
Tobias Kreiman, Christine Zhang, Matthew Kuner, Elizabeth Weaver, Ishan Amin, Hyunsoo Park, Yunsung Lim, Jiham Kim, Daryl Chrzan, Aron Walsh, Samuel M. Blau, Mark Asta, Aditi S. Krishnapriyan

Early feedback and discussion

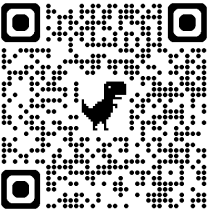
Janosh Riebesell, Philipp Benner, Patrick Huck, Rouxi Yang, Evan Walter Clark Spotte-Smith, Bowen Deng



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



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



@cyrusyc_tw



linkedin.com/in/yuanchiang/

Funding and compute

