When Data Ordering Helps or Hurts: A Matched-Support Study of Interatomic Potentials

Anonymous authors
Paper under double-blind review

Abstract

Curriculum learning reorganizes training data to guide the optimization of machine-learned interatomic potentials. Its effect on joint energy and force accuracy remains unclear because reported gains often combine ordering with changes in sample selection and exposure. We isolate ordering on OC20 using matched training pools and one-pass budgets, holding initialization, within-model loss weights, and evaluation conditions fixed. Five orderings are compared across four models and four in- and out-of-distribution validation splits. These comparisons reveal that ordering changes the Energy–Force balance even under the same training objective. Difficulty-based and coverage-to-tail curricula increase mean Force MAE by 10.1–16.6% versus uniform on complete validation splits, with penalties also present under constant learning rates. Increasing the pool from 200k to 500k improves the uniform Force baseline but preserves substantial ordering costs. At 500k, coverage-first reduces PaiNN Energy MAE by 24.3% while increasing Force MAE by 0.14%. No tested curriculum jointly lowers both mean errors relative to uniform. Data ordering therefore changes the Energy–Force balance under fixed exposure and loss weights, making curriculum value depend on the prediction objective and training regime.

1 Introduction

Machine-learned interatomic potentials replace repeated electronic-structure calculations with learned energy and force models, enabling large-scale atomistic simulation with a quantum-mechanical training target. Models differ in geometric inductive bias and scale. On the Open Catalyst 2020 (OC20) benchmark, progress has combined larger data with expressive geometric networks. These range from continuous-filter and equivariant message passing to directional and higher-order equivariant architectures (Schütt et al., 2017; 2021; Gasteiger et al., 2022; Liao et al., 2024; Chanussot et al., 2021). Yet model accuracy alone does not determine reliability. Performance varies across adsorbate, catalyst, and joint out-of-distribution splits, while recent materials studies show that generalization and scaling depend on how distribution shift is defined (Li et al., 2025; Kreiman & Krishnapriyan, 2026).

Curriculum learning changes the sequence in which existing examples are presented (Bengio et al., 2009; Wang et al., 2022). Recent atomistic work has also used curriculum-based training to construct reactive potentials for zeolite screening, showing relevance beyond standard classification benchmarks (Guo et al., 2024). Prior studies report optimization or accuracy gains for selected difficulty measures and pacing rules (Weinshall et al., 2018; Hacohen & Weinshall, 2019), but models can forget previously learned samples. Example-learning dynamics also show that data importance depends on training state and architecture (Toneva et al., 2019; Paul et al., 2021). Recent large-model studies find that no ordering dominates across tasks and that curriculum benefits interact with optimization choices (Luo et al., 2026; Jia et al., 2026).

A central experimental ambiguity remains: curricula often change ordering and support simultaneously. Different subsets may be exposed at different frequencies, so apparent gains can reflect selection, repeated exposure, or a larger effective budget. Controlling exposure

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and within-model loss weights is therefore necessary to attribute Energy–Force trade-offs to ordering. The issue is acute for potentials because structural and chemical coverage are coupled to target force distributions.

Interatomic potentials couple structure-level energy with atom-level force targets, providing complementary observables within one run. Joint evaluation tests whether scheduling benefits extend across objectives. OC20's ID and OOD splits probe transfer across adsorbate and catalyst chemistry.

We therefore ask whether ordering alone changes the Energy–Force balance under fixed support, exposure, and within-model loss weights. Each curriculum permutes the same 200,000 OC20 records, consumes each record once, and divides the sequence into four stages. The design keeps the data multiset and one-pass exposure fixed. We compare uniform, easy-to-hard, hard-to-easy, coverage-first, and coverage-to-tail orders across EquiformerV2, GemNet-OC, PaiNN, and SchNet. Complete-split evaluation contrasts the staged schedules with uniform, while targeted full-split analysis examines the bounded coverage-first signal. A nested 500k extension separates budget dependence from ordering using same-budget comparisons. We provide cross-model evidence that data-ordering value depends on the prediction objective and training regime.

2 Related Work

Curriculum learning and adaptive presentation. Curriculum learning was introduced as a continuation strategy that presents examples from simpler to more complex cases (Bengio et al., 2009). Self-paced and automated variants instead adapt selection to the learner or to measured progress (Kumar et al., 2010; Graves et al., 2017), while teacher-derived curricula estimate difficulty through transfer, self-tutoring, or a learned mentor (Weinshall et al., 2018; Hacohen & Weinshall, 2019; Jiang et al., 2018). These approaches differ in scoring, pacing, and whether presentation order also changes support or frequency, a distinction emphasized by curriculum surveys (Wang et al., 2022). Recent controlled language-model studies likewise find that direction, model capability, and learning-rate schedule interact with curriculum effects (Luo et al., 2026; Jia et al., 2026). In atomistic modeling, curriculum-based training has recently enabled reactive potential construction for zeolite screening (Guo et al., 2024), while multi-domain training with selective regularization improves cross-domain transfer (Kim et al., 2026). These approaches alter more than a fixed permutation of one training multiset. Our matched-support design isolates the narrower intervention of ordering a fixed multiset under a fixed presentation budget.

Data value, pruning, and learning dynamics. Curricula are closely related to, but scientifically distinct from, data selection. Proxy selection, gradient matching, reducible holdout loss, and large-scale pruning all change which examples receive computation (Coleman et al., 2020; Killamsetty et al., 2021; Mindermann et al., 2022; Sorscher et al., 2022). Example-level learning trajectories provide another family of signals: forgetting events and early-training statistics can identify atypical or high-value examples (Toneva et al., 2019; Paul et al., 2021). These works motivate our diagnostics but do not answer whether a permutation alone helps when every sample appears exactly once.

Forgetting and distribution shift. Catastrophic forgetting has a long history in sequential learning (French, 1999); modern continual-learning methods and surveys formalize retention, backward transfer, and memory-based mitigation (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; De Lange et al., 2021). Our four stage-end validation events support signed stage-to-stage changes, but they do not separate forgetting from recovery. More broadly, WILDS, DomainBed, and uncertainty-under-shift benchmarks show that conclusions can change when domains, model selection, or evaluation conditions are controlled (Koh et al., 2021; Gulrajani & Lopez-Paz, 2021; Ovadia et al., 2019).

Representations and interatomic potentials. Machine-learned potentials developed from high-dimensional neural potentials, Gaussian approximation potentials, invariant local descriptors, and energy-conserving force models (Behler & Parrinello, 2007; Bartók et al., 2010;

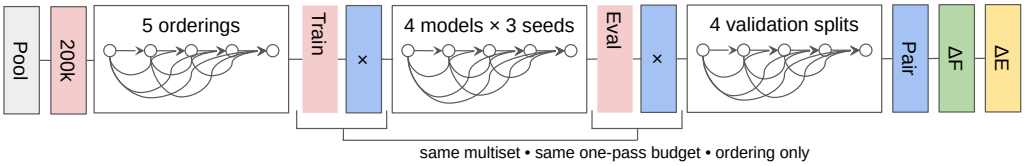

Figure 1: Core 200k matched-support design. A seed-local pool of 200,000 records is permuted into five orderings, crossed with four models and three seeds, and evaluated on four validation splits. Pairing against uniform isolates relative Force (Δ^F) and Energy (Δ^E) effects while the multiset and one-pass budget remain fixed.

2013; Chmiela et al., 2017; Smith et al., 2017). Atomic cluster expansions and Gaussian-process formulations provide systematic or kernel-based alternatives (Drautz, 2019; Deringer et al., 2021), while broad reviews describe the accuracy, conservation, data, and transferability requirements of modern force fields (Deringer et al., 2019; Unke et al., 2021). SchNet, PaiNN, GemNet-OC, and EquiformerV2 span invariant, Cartesian-equivariant, directional, and higher-order equivariant message passing (Schütt et al., 2017; 2021; Gasteiger et al., 2022; Liao et al., 2024). NequIP and Allegro further demonstrate the data efficiency and scaling of local equivariant representations (Batzner et al., 2022; Musaelian et al., 2023).

Dataset coverage and universal-potential generalization. OC20 supplies chemically and structurally distinct validation distributions for catalyst systems (Chanussot et al., 2021). Larger pretrained potentials such as M3GNet, CHGNet, and the GNoME potential initiated broad chemical and structural coverage (Chen & Ong, 2022; Deng et al., 2023; Merchant et al., 2023). More recent foundation models extend this direction across diverse materials and chemical domains (Batatia et al., 2025; Wood et al., 2025), but broad pretraining does not remove dataset dependence. A recent review identifies training-data generation and coverage as central reliability constraints (Kulichenko et al., 2024); direct comparisons further show that dataset transfer is architecture dependent and that target-domain fine-tuning remains useful (Niblett et al., 2025; Radova et al., 2025). Training-set selection, high-dimensional error analysis, entropy-driven coverage, and test-time adaptation likewise report sensitivity to configuration support (Montes de Oca Zapiain et al., 2022; Liu & Mo, 2024; Subramanyam & Perez, 2025; Cui et al., 2025). Recent studies expose systematic softening, domain-shift bias, and unresolved universality across elemental or latent-representation tests (Deng et al., 2025; Li et al., 2025; Kreiman & Krishnapriyan, 2026; Wong & Yang, 2026; Tahmasbi et al., 2026; Li & Walsh, 2026). We complement these architecture, adaptation, and dataset-construction studies by asking whether a fixed dataset becomes more useful when only its presentation order changes.

3 Matched-Support Curriculum Diagnosis

3.1 Controlled ordering intervention

For budget b and seed s, let $P_{s,b}$ denote a fixed pool of N_b training indices. A curriculum c is a permutation $\pi_{s,c,b}$ of $P_{s,b}$. We require

$$\mathrm{sort}(\pi_{s,c,b}) = \mathrm{sort}(P_{s,b}), \qquad |\pi_{s,c,b}| = N_b, \tag{1}$$

for every curriculum, so no method changes support, frequency, or sample budget within a budget. Each permutation is partitioned into four contiguous stages of $N_b/4$ presentations. Training uses one uninterrupted optimizer trajectory and sees every pool member exactly once.

3.2 Curricula

We compare uniform random ordering with two force-difficulty controls and two coverage-aware orders. Easy-to-hard and hard-to-easy sort records by per-structure force RMS, the

root-mean-square reference-force magnitude used here only as a label-amplitude proxy for difficulty; deterministic keys break ties. Coverage strata are formed from empirical-rank bins of log atom count, energy per atom, fixed-atom fraction, and the mean and standard deviation of atomic number, yielding at most $4\times4\times3\times4\times4 = 768$ nonempty cells. Coverage-first cycles through the strata, taking one record per stratum in a deterministic random order. Coverage-to-tail uses a first-stage coverage core, fills unused capacity with previously unseen exact compositions, and then orders the remainder through easy consolidation, intermediate difficulty, and a hard tail. This construction tests whether early representation breadth can survive later distribution changes.

3.3 Paired effects and descriptive uncertainty

For model m, curriculum c, seed s, and validation split d, the paired relative Force effect is

$$\Delta^F_{mcsd} = 100\frac{E^F_{mcsd} - E^F_{m,\mathrm{unif},s,d}}{E^F_{m,\mathrm{unif},s,d}}, \tag{2}$$

with an analogous Energy effect. Negative values favor the curriculum. For a budget b, the uniform run with the same model, seed, and validation split is the reference. We report these same-budget relative effects for ordering and matched-seed absolute MAE changes between budgets for budget dependence. For the model-level 200k summaries, we resample training-seed clusters while retaining all four splits belonging to each selected seed. The resulting 2.5th–97.5th percentile range is a descriptive interval. Appendix J specifies the model-level and overall resampling procedures.

3.4 Stage diagnostics

At the end of each stage, we record Energy and Force MAE at $t \in \{1, 2, 3, 4\}$, corresponding to cumulative presentations $N_b t/4$. Consecutive endpoints describe the net change over each complete stage. Appendix H gives the endpoint definition and its relationship to measurements of forgetting and recovery.

4 Experimental Setup

We study the OC20 structure-to-energy-and-forces task using the two-million sample training subset and the predefined ID, OOD-Ads, OOD-Cat, and OOD-Both validation distributions (Chanussot et al., 2021). The architecture matrix contains EquiformerV2, GemNet-OC, PaiNN, and SchNet, covering substantially different geometric inductive biases and computational scales. All models use AdamW, with effective batches of 4, 16, 32, and 24 structures, respectively. Optimization settings and loss weights are held fixed across curricula within each architecture. EquiformerV2, GemNet-OC, and PaiNN predict forces directly; SchNet obtains forces from energy gradients. We evaluate the final-stage checkpoint using structure-level Energy MAE and component-wise Force MAE over free atoms after reversing label normalization. Appendix B gives the complete configuration and metric definitions.

The 200k training matrix contains five curricula and seeds 17, 29, and 43, for $4 \times 5 \times 3 = 60$ runs. Each seed uses the same 200,000-record pool and four 50,000-record stages. The bounded diagnostic evaluates frozen 10,000-record subsets of all four validation splits, producing 240 cells; Force MAE is the primary metric and Energy MAE exposes trade-offs. The formal full evaluation uses four curricula on complete splits, for 192 absolute-metric cells. The separate post-hoc targeted evaluation reuses coverage-first checkpoints on those splits, adding 48 additional absolute-metric cells without new training identities.

The scale extension uses the same four architectures with uniform, coverage-first, and coverage-to-tail ordering on a 500,000-record pool divided into four 125,000-record stages. It uses one training seed (43). The two budgets are trained independently: 500k adds 300,000 records and approximately 2.5 times as many optimizer updates, while optimizer and scheduler configurations remain the same. Both budgets use the same four 10,000-record

Table 1: Matched-support training budgets and validation designs.

Factor	Values
Architectures	EquiformerV2, GemNet-OC, PaiNN, SchNet
Budgets	200k training; 500k nested training
Curricula	200k: five; 500k: uniform, coverage-first, coverage-to-tail
Training identities	200k: $4 \times 5 \times 3 = 60$; 500k: $4 \times 3 \times 1 = 12$
Training pools	200k: 200,000; 500k: 500,000 records
Stages	200k: $4 \times 50{,}000$; 500k: $4 \times 125{,}000$
Validation	ID, OOD-Ads, OOD-Cat, OOD-Both
Bounded cells	200k: 240; 500k: 48

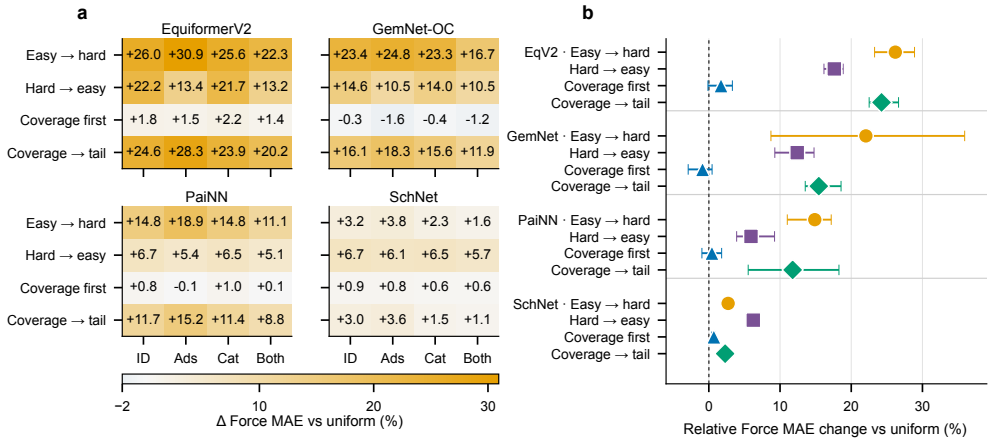

Figure 2: Bounded paired Force MAE effects. (a): seed-mean relative effects for each model, curriculum, and validation split. (b): model-level means with descriptive seed-cluster 95% percentile intervals. Negative values favor the curriculum; the intervals are not formal inferential confidence intervals.

validation subsets. Model configuration details and complete result tables are provided in the appendices.

The 500k pool contains the 200k seed-43 pool. Each budget uses its own permutation, giving a matched comparison between nested training pools.

5 Bounded Cross-Model Evidence

5.1 Uniform ordering is the robust bounded baseline

Across the 16 model–split comparisons, uniform ordering attains the lowest Force MAE in 11, while coverage-first attains the lowest value in the remaining five. All four GemNet-OC splits and PaiNN OOD-Ads favor coverage-first; no easy-to-hard, hard-to-easy, or coverage-to-tail condition wins a comparison. The model dependence is visible in Figure 2: the same coverage-first construction slightly improves GemNet-OC but is neutral or unfavorable for EquiformerV2, PaiNN, and SchNet. Appendix E reports seed-averaged effects for each model, curriculum, and split.

Difficulty ordering is consistently harmful in this fixed-support, one-pass regime. Easy-to-hard raises mean Force MAE by 26.19%, 22.05%, 14.88%, and 2.71% for EquiformerV2, GemNet-OC, PaiNN, and SchNet, respectively. Coverage-to-tail also raises Force MAE for every architecture, from 2.29% for SchNet to 24.25% for EquiformerV2. These effects show that a plausible ranking criterion and broad early coverage do not by themselves improve final force prediction when support and budget are controlled.

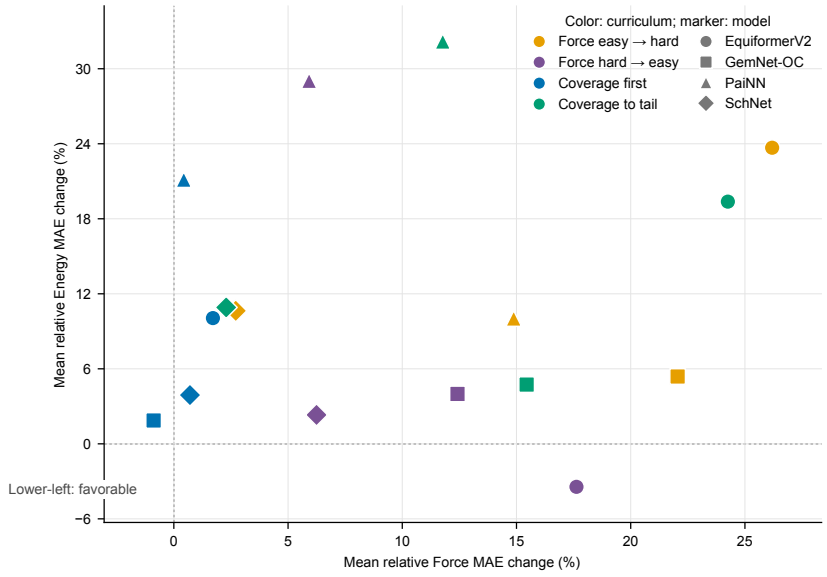

Figure 3: Bounded Energy–Force trade-off. Each point averages the twelve paired cells from three seeds and four splits for one model and curriculum. The lower-left quadrant represents simultaneous improvement; no tested aggregate occupies that quadrant.

Table 2: Mean bounded relative MAE change versus paired uniform ordering (%). Negative values are favorable.

		Curriculum			
Model	Metric	Easy→hard	Hard→easy	Coverage-first	Coverage-tail
EquiformerV2	Force	26.19	17.63	1.71	24.25
	Energy	23.68	-3.44	10.06	19.37
GemNet-OC	Force	22.05	12.42	-0.88	15.44
	Energy	5.38	3.99	1.87	4.74
PaiNN	Force	14.88	5.92	0.44	11.77
	Energy	10.02	29.04	21.13	32.18
SchNet	Force	2.71	6.25	0.71	2.29
	Energy	10.64	2.31	3.90	10.91

5.2 Coverage gains do not remove the Energy–Force trade-off

Coverage-first is the only ordering that approaches uniform in aggregate Force MAE: its model-level changes are +1.71%, −0.88%, +0.44%, and +0.71% in the same architecture order. The GemNet-OC Force improvement is accompanied by a +1.87% Energy MAE increase. More generally, none of the 16 model–curriculum aggregates lies in the jointly favorable quadrant of the Energy–Force plane (Figure 3). The only aggregate Energy improvement, hard-to-easy for EquiformerV2, coincides with a 17.63% Force MAE increase. Appendix F lists the corresponding aggregates.

6 Full-Split Evaluation and Budget Dependence

The full evaluation contains 192 absolute-metric cells: four architectures, four curricula (uniform, easy-to-hard, hard-to-easy, and coverage-to-tail), three seeds, and four validation splits. Unlike the bounded diagnostic, each split is evaluated at its complete cardinality: 999,866 records for ID, 999,838 for OOD-Ads, 999,809 for OOD-Cat, and 999,944 for OOD-

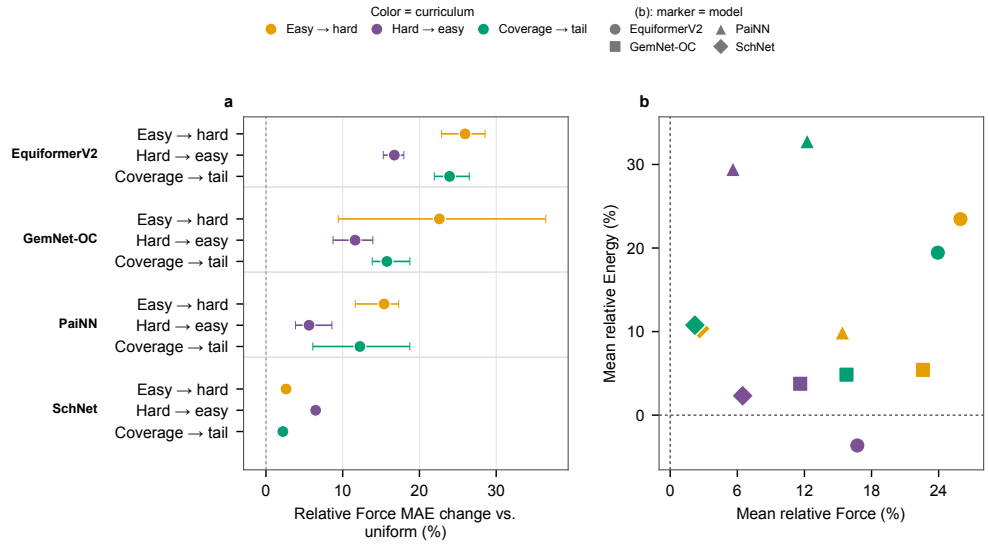

Figure 4: Full-split effects of three staged curricula relative to paired uniform ordering. (a) contains 12 Force forest intervals (four models by three curricula), and (b) contains 12 Energy–Force Pareto aggregates. Every interval or point summarizes $n = 3$ seeds $\times$ 4 validation splits; (a) shows descriptive seed-cluster 95% intervals. Positive values indicate higher error than uniform ordering; the lower-left quadrant in (b) denotes simultaneous Energy–Force improvement.

Both. Pairing the three staged curricula against uniform yields 144 full paired effects, 48 per curriculum.

6.1 Uniform ordering is lower for all three staged curricula

Across the three staged curricula, mean Force MAE relative to paired uniform ordering increased by +16.6407% for easy-to-hard, +10.1092% for hard-to-easy, and +13.5334% for coverage-to-tail. Uniform is lower in all 16 model–split units for each curriculum. The 48 paired coverage-to-tail cells have a Force relative mean of +13.5334% with a hierarchical-bootstrap 95% interval of [+5.3337%, +21.7820%]. The model-level effects are +23.9264%, +15.7585%, +12.2426%, and +2.2060% for EquiformerV2, GemNet-OC, PaiNN, and SchNet, respectively. Energy moves in the same unfavorable direction, with a +17.6605% aggregate relative regression. Broad early coverage therefore did not translate into a robust full-split gain under the matched one-pass budget.

The penalties also occur in GemNet-OC and PaiNN, whose learning rates remain at 5×10^{-4} and 10^{-4} throughout the 200k runs. For each staged curriculum, all twelve paired Force effects are positive in both models. The ordering penalty therefore persists under constant learning rates.

6.2 Targeted full-split confirmation of coverage-first

The bounded analysis suggested a small, model-specific Force benefit for coverage-first. We evaluated the existing coverage-first checkpoints on the complete validation splits against the paired uniform checkpoints, adding 48 absolute-metric cells.

Across the 48 cellwise pairs, the mean relative changes were +0.47% for Force MAE and +9.43% for Energy MAE. At the reported four-decimal precision, Force improved in 21 cells, worsened in 26, and tied in one; Energy improved in 8 cells and worsened in 40. GemNet-OC's benefit spans all four full splits, with mean Force improvements of 0.29–1.75% and Energy increases of 0.44–3.29%. The largest Force gain occurs on OOD-Ads, while the

Table 3: Coverage-first comparison across bounded and targeted full-split evaluation. Entries are mean cellwise relative changes versus paired uniform ordering; positive values indicate higher error. The bounded column uses the 10,000-record validation subsets, whereas the targeted full columns use the complete released splits. Each model's full-split summary contains 12 cells (3 training seeds × 4 validation splits), and "F improved" counts lower-Force cells among those 12.

Model	Bounded ΔF (%)	Full ΔF (%)	Full ΔE (%)	F improved
EquiformerV2	+1.71	+1.73	+10.13	3/12
GemNet-OC	-0.88	-0.92	+2.00	10/12
PaiNN	+0.44	+0.34	+21.66	6/12
SchNet	+0.71	+0.74	+3.95	2/12

Energy cost is present even on ID. This distribution-level pattern explains the small Force gain and Energy regression in Table 3. None of the 16 full model–curriculum aggregates, including coverage-first, jointly improves Energy and Force.

6.3 500k nested-pool budget comparison

The 500k extension compares the four models and three curricula at seed 43 across the four validation splits, using the same 10,000-record validation indices as the 200k comparison. Within the 500k budget, coverage-first changes mean cellwise Force and Energy MAE by +1.89% and −1.86%, respectively, relative to same-budget uniform. Coverage-to-tail changes them by +25.10% and +13.28%. Thus, the larger pool does not remove ordering heterogeneity: coverage-first trades a small Force increase for lower Energy error on average, whereas coverage-to-tail incurs a large Force penalty. In absolute terms, the matched-seed uniform Force MAE decreases from 200k to 500k by 14.33%, 12.17%, 7.98%, and 27.78% for EquiformerV2, GemNet-OC, PaiNN, and SchNet, respectively; the corresponding Energy changes are −18.68%, −19.23%, +6.36%, and −6.50%. Coverage-to-tail nevertheless increases absolute Force MAE by 0.85 meV Å^{-1} for EquiformerV2 and 12.00 meV Å^{-1} for PaiNN, even though it decreases for GemNet-OC and SchNet. More budget therefore improves the uniform baseline without automatically removing an ordering penalty.

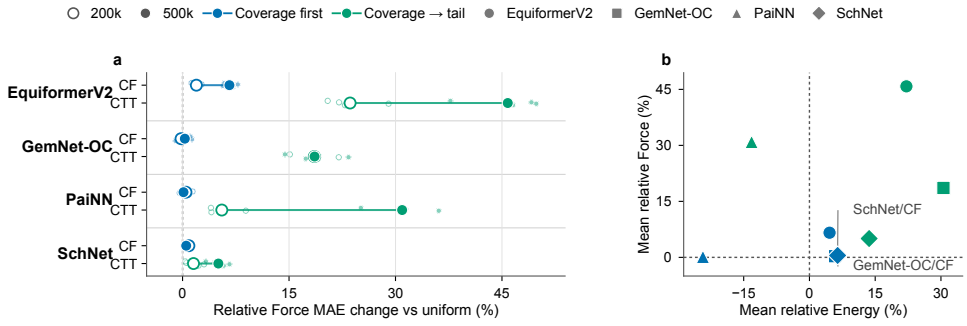

Figure 5: 500k nested-pool budget comparison. (a) shows Force MAE changes relative to same-budget uniform; open and filled markers denote 200k and 500k for the same fixed training seed, respectively. Large markers are means over four validation splits and faint points are the individual splits. (b) shows 500k model–curriculum aggregates in the Energy–Force plane. Positive values indicate higher error.

The model-level summaries make the trade-off explicit. Coverage-first lowers Energy only for PaiNN (−24.30%), while its Force change there is nearly neutral (+0.14%); the other three models show increases in both errors. Coverage-to-tail raises Force for every model,

with the largest penalties for EquiformerV2 and PaiNN. The remaining model–curriculum values and split signs are shown in Figure 5 and Appendix I.

Relative to the matched seed-43 200k anchor, the mean coverage-first Force effect changes from +0.78% to +1.89% (+1.12 percentage points), while its Energy effect changes from +14.98% to −1.86% (−16.83 percentage points). The coverage-to-tail Force effect increases from +12.31% to +25.10% (+12.79 percentage points), whereas its Energy effect is nearly unchanged (+13.00% to +13.28%). These matched-seed contrasts compare separate nested-pool runs.

7 Discussion

Our matched-support comparisons identify data ordering as a factor in the terminal Energy–Force balance under a fixed training objective. Within each model, support, exposure, and loss weights remain the same across curricula. The two errors therefore reflect different optimization endpoints. Across the tested models and budgets, no curriculum jointly lowers both mean errors relative to uniform mixing. An ordering's value therefore depends on the target observable. Energy values and forces probe complementary aspects of the same physical system. Energy MAE measures errors in configuration-level values, while Force MAE measures errors in local spatial responses. Small errors in energy values alone do not ensure accurate gradients. Joint evaluation is therefore informative even for models that derive forces from an energy function.

This result complements recent evidence on curriculum–optimization interactions. Luo et al. (2026) show that learning-rate decay can weaken curriculum benefits in LLM pretraining, while Jia et al. (2026) find that ordering preferences vary across models and tasks. Here, substantial Force penalties also occur in the constant-learning-rate GemNet-OC and PaiNN runs. The observed cost thus extends to training without learning-rate decay. Both ascending and descending force-magnitude orders incur penalties, so reversing the ranking direction does not restore the uniform baseline. This agreement across directions makes temporal organization relevant alongside the difficulty score itself. Sustained mixing and concentration within successive stages expose the learner to different local distributions. At each budget, these orderings consume the same structures with the same number of updates within each model.

The budget comparison separates improvements in the baseline from changes in curriculum value. Enlarging the pool improves uniform Force prediction across all four models, while coverage-first and coverage-to-tail produce different Energy–Force responses. These findings make the target observable central to curriculum selection. Evaluating both errors at the intended model and budget reveals whether an ordering improves the required prediction task and what accuracy cost accompanies that improvement.

The cross-model patterns concern models together with their training regimes, including learning-rate schedules, effective batches, and force-prediction heads. A direct next test crosses ordering with learning-rate trajectories while holding support, exposure, loss weights, and within-model batch settings fixed. This comparison would distinguish effects that persist across optimization schedules from ordering–schedule interactions. For learning physically related observables, the present results support joint evaluation of curriculum outcomes against a matched uniform reference.

8 Conclusion

Matched-support evaluation shows that ordering changes the Energy–Force balance when each model sees the same data once and its within-model loss weights remain fixed. In the 200k core, staged schedules incur 10.1–16.6% higher mean Force MAE, while GemNet-OC's coverage-first Force gain carries an Energy cost. At 500k, absolute uniform Force MAE improves for all models, yet coverage-to-tail penalties grow; PaiNN coverage-first trades +0.14% Force for −24.3% Energy. Curriculum value therefore depends on target and training regime.

AI use statement

Generative AI tools were used to assist with initial manuscript drafting and editing. All AI-assisted text, as well as the scientific claims, citations, analyses, and results, were reviewed and verified by the authors, who take responsibility for the final submission.

Reproducibility statement

The accompanying code provides curriculum construction, training and evaluation interfaces, and analysis utilities. Training parameters and evaluation definitions are given in Appendix B, and statistical procedures in Appendix J. Appendix L provides code access.

References

Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical Review Letters, 104(13), 2010. doi: 10.1103/PhysRevLett.104.136403.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. Physical Review B, 87(18), 2013. doi: 10.1103/PhysRevB.87.184115.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, et al. A foundation model for atomistic materials chemistry. The Journal of Chemical Physics, 163(18), 2025. doi: 10.1063/5.0297006.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. $E(3)$-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature Communications, 13(1), 2022. doi: 10.1038/s41467-022-29939-5.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical Review Letters, 98(14), 2007. doi: 10.1103/PhysRevLett.98.146401.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th International Conference on Machine Learning, pp. 41–48, 2009. doi: 10.1145/1553374.1553380.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (OC20) dataset and community challenges. ACS Catalysis, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. Nature Computational Science, 2(11):718–728, 2022. doi: 10.1038/s43588-022-00349-3.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. Science Advances, 3(5), 2017. doi: 10.1126/sciadv.1603015.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=HJg2b0VYDr.

Taoyong Cui, Chenyu Tang, Dongzhan Zhou, Yuqiang Li, Xingao Gong, Wanli Ouyang, Mao Su, and Shufei Zhang. Online test-time adaptation for better generalization of interatomic potentials to out-of-distribution data. Nature Communications, 16(1), 2025. doi: 10.1038/s41467-025-57101-4.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. doi: 10.1109/TPAMI.2021.3057446.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. Nature Machine Intelligence, 5(9):1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.

Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A. Persson, and Gerbrand Ceder. Systematic softening in universal machine learning interatomic potentials. npj Computational Materials, 11(1), 2025. doi: 10.1038/s41524-024-01500-6.

Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. Advanced Materials, 31(46), 2019. doi: 10.1002/adma.201902765.

Volker L. Deringer, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. Chemical Reviews, 121(16):10073–10141, 2021. doi: 10.1021/acs.chemrev.1c00022.

Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. Physical Review B, 99(1), 2019. doi: 10.1103/PhysRevB.99.014104.

Robert M. French. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 3(4):128–135, 1999. doi: 10.1016/S1364-6613(99)01294-2.

Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C. Lawrence Zitnick, and Abhishek Das. GemNet-OC: Developing graph neural networks for large and diverse molecular simulation datasets. Transactions on Machine Learning Research, 2022. URL https://openreview.net/forum?id=u8tvSxm4Bs.

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 1311–1320, 2017. URL https://proceedings.mlr.press/v70/graves17a.html.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Jiawei Guo, Tyler Sours, Sam Holton, Chenghan Sun, and Ambarish R. Kulkarni. Screening Cu-zeolites for methane activation using curriculum-based training. ACS Catalysis, 14 (3):1232–1242, 2024. doi: 10.1021/acscatal.3c05275.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 2535–2544, 2019. URL https://proceedings.mlr.press/v97/hacohen19a.html.

Yaning Jia, Chunhui Zhang, Xingjian Diao, Xiangchi Yuan, Zhongyu Ouyang, Chiyu Ma, and Soroush Vosoughi. What makes a good curriculum? disentangling the effects of data ordering on llm mathematical reasoning. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, pp. 34472–34488, 2026. doi: 10.18653/v1/2026.acl-long.1591.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 2304–2313, 2018. URL https://proceedings.mlr.press/v80/jiang18c.html.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 5464–5474, 2021. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Jaesun Kim, Jinmu You, Yutack Park, Yunsung Lim, Yujin Kang, Jisu Kim, Haekwan Jeon, Suyeon Ju, Deokgi Hong, Seung Yul Lee, Saerom Choi, Yongdeok Kim, Jae W. Lee, and Seungwu Han. Optimizing cross-domain transfer for universal machine learning interatomic potentials. Nature Communications, 17(1), 2026. doi: 10.1038/s41467-026-70195-8.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Tobias Kreiman and Aditi S. Krishnapriyan. Understanding and mitigating distribution shifts for universal machine learning interatomic potentials. Digital Discovery, 5(1):415–439, 2026. doi: 10.1039/D5DD00260E.

Maksim Kulichenko, Benjamin Nebgen, Nicholas Lubbers, Justin S. Smith, Kipton Barros, Alice E. A. Allen, Adela Habib, Emily Shinkle, Nikita Fedik, Ying Wai Li, Richard A. Messerly, and Sergei Tretiak. Data generation for machine learning interatomic potentials and beyond. Chemical Reviews, 124(24):13681–13714, 2024. doi: 10.1021/acs.chemrev.4c00572.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In Advances in Neural Information Processing Systems, volume 23, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/e57c6b956a6521b28495f2886ca0977a-Abstract.html.

Kangming Li, Andre Niyongabo Rubungo, Xiangyun Lei, Daniel Persaud, Kamal Choudhary, Brian DeCost, Adji Bousso Dieng, and Jason Hattrick-Simpers. Probing out-of-distribution generalization in machine learning for materials. Communications Materials, 6(1):9, 2025. doi: 10.1038/s43246-024-00731-w.

Zhenzhu Li and Aron Walsh. Platonic representation of foundation machine learning interatomic potentials. Nature Machine Intelligence, 8(5):830–840, 2026. doi: 10.1038/s42256-026-01235-7.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=mCOBKZmrzD.

Yunsheng Liu and Yifei Mo. Learning from models: High-dimensional analyses on the performance of machine learning interatomic potentials. npj Computational Materials, 10(1), 2024. doi: 10.1038/s41524-024-01333-3.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In Advances in Neural Information Processing Systems, volume 30, 2017. URL

https://proceedings.neurips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

Kairong Luo, Zhenbo Sun, Haodong Wen, Xinyu Shi, Jiarui Cui, Chenyi Dang, Kaifeng Lyu, and Wenguang Chen. How learning rate decay wastes your best data in curriculum-based llm pretraining. In International Conference on Learning Representations, 2026. URL https://mlanthology.org/iclr/2026/luo2026iclr-learning/.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. Nature, 624(7990): 80–85, 2023. doi: 10.1038/s41586-023-06735-9.

Sören Mindermann, Jan M. Brauner, Muhammed T. Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 15630–15649, 2022. URL https://proceedings.mlr.press/v162/mindermann22a.html.

David Montes de Oca Zapiain, Mitchell A. Wood, Nicholas Lubbers, Carlos Z. Pereyra, Aidan P. Thompson, and Danny Perez. Training data selection for accuracy and transferability of interatomic potentials. npj Computational Materials, 8(1), 2022. doi: 10.1038/s41524-022-00872-x.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. Nature Communications, 14(1), 2023. doi: 10.1038/s41467-023-36329-y.

Samuel P. Niblett, Panagiotis Kourtis, Ioan-Bogdan Magdău, Clare P. Grey, and Gábor Csányi. Transferability of data sets between machine-learned interatomic potential algorithms. Journal of Chemical Theory and Computation, 21(12):6096–6112, 2025. doi: 10.1021/acs.jctc.5c00272.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In Advances in Neural Information Processing Systems, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/ac56f8fe9eea3e4a365f29f0f1957c55-Abstract.html.

Mariia Radova, Wojciech G. Stark, Connor S. Allen, Reinhard J. Maurer, and Albert P. Bartók. Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning. npj Computational Materials, 11(1), 2025. doi: 10.1038/s41524-025-01727-x.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In Advances in Neural Information Processing Systems, volume 30, 2017. URL https://papers.nips.cc/paper/6700-schnet-a-continuous-filter-convolutional-neural-network-for-modeling-quantum-interactions.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 9377–9388, 2021. URL https://proceedings.mlr.press/v139/schutt21a.html.

Justin S. Smith, Olexandr Isayev, and Adrian E. Roitberg. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. Chemical Science, 8(4):3192–3203, 2017. doi: 10.1039/C6SC05720A.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In Advances in Neural Information Processing Systems, volume 35, 2022. doi: 10.52202/068431-1419.

Aparna P. A. Subramanyam and Danny Perez. Information-entropy-driven generation of material-agnostic datasets for machine-learning interatomic potentials. npj Computational Materials, 11(1), 2025. doi: 10.1038/s41524-025-01602-9.

Hossein Tahmasbi, Andreas Knüpfer, Thomas D. Kühne, and Hossein Mirhosseini. Benchmarking universal machine learning interatomic potentials on elemental systems. npj Computational Materials, 12(1), 2026. doi: 10.1038/s41524-026-02251-2.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. Chemical Reviews, 121(16):10142–10186, 2021. doi: 10.1021/acs.chemrev.0c01111.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9):4555–4576, 2022. doi: 10.1109/TPAMI.2021.3069908.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 5238–5246, 2018. URL https://proceedings.mlr.press/v80/weinshall18a.html.

Nicolas H. Wong and Julia H. Yang. Bias in universal machine-learned interatomic potentials and its effects on fine-tuning. Journal of Chemical Theory and Computation, 22(13):6820–6834, 2026. doi: 10.1021/acs.jctc.6c00425.

Brandon Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John Kitchin, et al. UMA: A family of universal models for atoms. In Advances in Neural Information Processing Systems, volume 38, pp. 143528–143564, 2025. doi: 10.52202/085713-4310.

A Data Construction and Curriculum Protocol

The core experiment uses the two-million-record OC20 S2EF training source. For each seed s, the builder maps the source shards to contiguous global indices, ranks those indices with a seed-keyed domain-separated SHA-256 rank, and retains the first $N = 200{,}000$ distinct indices (ties are resolved by the global index). Feature extraction and all five schedules are then applied to this one seed-local pool. Each schedule is a signed-int64 permutation of the pool and is reshaped into four contiguous stages of $L = N/4 = 50{,}000$ records. The 500k extension uses the same constructor with $N = 500{,}000$, $L = 125{,}000$, and the 200k pool retained as a set-level parent; its schedule is recomputed over the larger pool.

1. Row features. For atom a in structure i, let $\mathcal{A}_i$ be all atoms and let $\mathcal{F}_i = \{a : \text{fixed}_{ia} = 0\}$. The Force RMS coordinate is

$$r_i = \sqrt{\frac{1}{3|S_i|} \sum_{a \in S_i} \sum_{\alpha=1}^{3} F_{ia\alpha}^2}, \qquad S_i = \begin{cases} \mathcal{F}_i, & |\mathcal{F}_i| > 0, \\ \mathcal{A}_i, & |\mathcal{F}_i| = 0. \end{cases}$$

   An all-fixed structure uses all atoms. The remaining row features are $\log(1 + n_i)$, E_i/n_i, the fraction of atoms with fixed $\neq 0$, the count-weighted mean atomic number, and its population standard deviation. The exact composition signature is $Z_1 : c_1 | Z_2 : c_2 | \cdots$, with strictly increasing Z and positive counts.

2. Strata. The five numeric coverage coordinates are

$$\big(\log(1 + n_i), E_i/n_i, \text{fixed fraction}_i, \bar{Z}_i, \sigma_{Z,i}\big).$$

   They are binned with $B = (4, 4, 3, 4, 4)$. For a feature value whose equal-value block occupies zero-based sorted positions $[a, b)$ among M rows, its bin is

$$q_B = \min\{B - 1, t_B\}, \quad t_B = \left\lfloor \frac{(a + b - 1)B}{2M} \right\rfloor.$$

   Values are ordered by stable mergesort, equal values share one block, and a constant feature maps to bin zero. The five bins are combined as a mixed-radix label, yielding at most $4{\cdot}4{\cdot}3{\cdot}4{\cdot}4 = 768$ nonempty strata. Observed mixed-radix codes are relabeled in increasing order as $0, \ldots, G-1$; these contiguous IDs define the stratum ordering keys. Composition signatures are kept separate from this label; they record the full element counts of each structure.

3. Deterministic keys and force controls. A seed-keyed SHA-256 rank of the global index defines the ordering key, with distinct domains for uniform, force, coverage-member, coverage-stratum, tail-member, and tail-stratum rankings. Global index resolves hash ties. Uniform sorts by its keyed rank; easy-to-hard and hard-to-easy sort by increasing and decreasing r_i, respectively, with the force-domain key resolving equal values.

4. Coverage-first. Let K_{cm} and K_{cs} denote the coverage-member and coverage-stratum keyed ranks. Within each stratum, rows are ordered by (K_{cm}, j), not by Force; strata are ordered by $(K_{\text{cs}}, \text{stratum ID})$. The flattened schedule consumes these queues round-robin: take one row from every active stratum in stratum order, then one from every still-active stratum again, until all rows are consumed; the four stage boundaries are applied afterward.

5. Coverage-to-tail. Let $d_i = q_4(r_i)$, m_g be the size of stratum g, and $u_i = 1/m_{g(i)}$, where $g(i)$ is the stratum of row i. Let K_{tm} and K_{ts} denote the tail-member and tail-stratum keyed ranks. Within each stratum, rows are ordered by $(\mathbb{1}[d_i > 1], r_i, K_{\text{tm}}, j)$. Strata with at least one $d_i \leq 1$ are placed first, followed by (K_{ts}, g). The first $\min(L, \#\text{strata})$ strata are selected, and the first row of each selected stratum is emitted once in that stratum order as Stage 1 (coverage_core). Remaining Stage-1 capacity is filled from the unselected rows sorted by $(d_i, -u_i, r_i, K_{\text{tm}}, j)$: the first occurrence of each previously unseen exact composition is taken first, then repeated compositions. After Stage 1, the unselected rows retain this same order and are sliced into easy_consolidation, mid, and hard_tail stages.

Table 4: Public feature schema used to construct curricula.

Feature	Role
Force RMS	Difficulty coordinate
Log atom count	Structural-size coverage
Energy per atom	Energetic coverage
Fixed-atom fraction	Constraint coverage
Atomic-number mean/std.	Chemical coverage
Composition signature	Exact-composition novelty

B Architecture and Training Hyperparameters

Table 5 gives the parameters used for the 200k and 500k experiments. Within each architecture, all curricula use the same model, optimizer, loss, and batch settings. Each run starts from a random initialization and carries the optimizer and scheduler state across all four stages. Evaluation uses the checkpoint at the end of stage four.

Table 5: Model and optimization parameters. Batch sizes count structures. The effective batch is the microbatch size multiplied by gradient accumulation on one GPU.

Parameter	EquiformerV2	GemNet-OC	PaiNN	SchNet
Layers / interaction blocks	12	4	6	5
Node / spherical channels	128	256	512	1024
Radial basis functions	512	128	128	200
Cutoff (Å)	12	12	12	6
Main-graph neighbor cap	20	30	50	50
Training microbatch	1	16	32	24
Gradient accumulation	4	1	1	1
Effective batch	4	16	32	24
Evaluation batch	1	16	32	24
Initial learning rate	4×10^{-4}	5×10^{-4}	10^{-4}	10^{-4}
AdamW weight decay	0.001	0	0	0.01
Energy loss weight	2	1	1	1
Force loss weight	100	100	100	100
Gradient clipping norm	100	10	10	none
EMA decay	0.999	0.999	0.999	0
Data-loader workers	8	2	2	16

Architectures. EquiformerV2 uses $l_{\max} = 6$, $m_{\max} = 2$, eight attention heads, 64 attention hidden channels, 16 value channels, and a 128-channel feedforward network. Its attention dropout and stochastic-depth rates are 0.1 and 0.05. GemNet-OC uses 512 edge channels, seven spherical basis functions, and neighbor caps of 8, 20, and 1000 for quadruplet, atom–edge, and atom–atom interactions; all interaction cutoffs are 12 Å. SchNet uses 256 filters. EquiformerV2, GemNet-OC, and PaiNN predict forces directly; SchNet differentiates its energy output with respect to atomic positions. All models use periodic boundaries.

Optimization. All models use AdamW; GemNet-OC and PaiNN enable AMSGrad. EquiformerV2 uses cosine learning-rate decay with a warmup factor of 0.2, warmup length of 0.1 stage epochs, and final learning-rate factor of 0.01. GemNet-OC and PaiNN use ReduceLROnPlateau on validation Force MAE with factor 0.8 and patience 3; their learning rates remain at the initial values throughout these runs. SchNet uses a warmup factor of 0.2 for 0.3999872 stage epochs, followed by multiplicative reductions of 0.1 at 0.6666411, 1.0666283, and 1.3332821 stage epochs, ending at 10^{-7}. One stage epoch is one pass through the current stage. Final partial minibatches are retained. Update counts at both budgets are given in Appendix K.

Loss and normalization. The training objective is $w_E L_E + w_F L_F$, with the weights in Table 5. L_E is the mean absolute error of the normalized structure-level energy. For

EquiformerV2, GemNet-OC, and PaiNN, L_F is the mean Euclidean norm of the normalized force-error vector over free atoms; SchNet uses the mean absolute component error. Energy labels are normalized as $(E - \mu_E)/\sigma$, with $\mu_E = -0.755445\,\text{eV}$ and $\sigma = 2.887317\,\text{eV}$. Forces use zero mean and scale $2.887317\,\text{eV}\,\text{Å}^{-1}$. The same normalization constants are used across models, curricula, and budgets.

Evaluation metrics. Predictions are transformed back to physical units before evaluation. For S structures and their set $\mathcal{A}$ of free atoms,

$$\text{MAE}_E = \frac{1}{S}\sum_{s=1}^{S}|\widehat{E}_s - E_s|, \qquad \text{MAE}_F = \frac{1}{3|\mathcal{A}|}\sum_{i\in\mathcal{A}}\sum_{\alpha\in\{x,y,z\}}|\widehat{F}_{i\alpha} - F_{i\alpha}|. \tag{3}$$

Energy MAE uses the OC20 structure-level energy label in eV. Force MAE is in $\text{eV}\,\text{Å}^{-1}$ and includes atoms with fixed=0. The same metric definitions and atom mask apply to all four architectures.

C Bounded Factorial Completeness

Table 6 gives the dimensions of the bounded comparison. Each trained model is evaluated on four 10,000-record subsets. Pairing each non-uniform ordering with uniform gives 192 effects and 16 model–curriculum aggregates.

Table 6: Dimensions of the bounded comparison.

Quantity	Count	Factors
Training runs	60	4 models × 5 curricula × 3 repetitions
Bounded evaluations	240	60 runs × 4 splits
Paired effects	192	4 models × 4 non-uniform curricula × 3 repetitions × 4 splits
Pareto aggregates	16	4 models × 4 non-uniform curricula

D Complete Bounded Results

Table 7 summarizes all 240 terminal bounded cells into 80 model–curriculum–split rows. Each cell used 10,000 validation records. Values are mean ± sample standard deviation across three repetitions; the accompanying TSV retains all row-level values for paired recomputation.

Table 7: Bounded Energy and Force MAE across three repetitions. Energy is in eV and Force in $\text{eV}\,\text{Å}^{-1}$.

Model	Curriculum	Split	Energy MAE	Force MAE
EquiformerV2	Uniform	ID	0.6630 ± 0.0224	0.0637 ± 0.0006
EquiformerV2	Uniform	OOD-Ads	0.8051 ± 0.0436	0.0665 ± 0.0009
EquiformerV2	Uniform	OOD-Cat	0.6549 ± 0.0227	0.0633 ± 0.0007
EquiformerV2	Uniform	OOD-Both	0.9302 ± 0.0452	0.0791 ± 0.0009
EquiformerV2	Easy→hard	ID	0.8551 ± 0.0046	0.0802 ± 0.0019
EquiformerV2	Easy→hard	OOD-Ads	0.9812 ± 0.0317	0.0871 ± 0.0021
EquiformerV2	Easy→hard	OOD-Cat	0.8328 ± 0.0009	0.0795 ± 0.0018
EquiformerV2	Easy→hard	OOD-Both	1.0806 ± 0.0347	0.0967 ± 0.0018
EquiformerV2	Hard→easy	ID	0.6348 ± 0.0172	0.0778 ± 0.0003
EquiformerV2	Hard→easy	OOD-Ads	0.7466 ± 0.0333	0.0754 ± 0.0002
EquiformerV2	Hard→easy	OOD-Cat	0.6396 ± 0.0183	0.0770 ± 0.0003
EquiformerV2	Hard→easy	OOD-Both	0.9270 ± 0.0384	0.0895 ± 0.0002
EquiformerV2	Coverage-first	ID	0.7366 ± 0.0138	0.0648 ± 0.0008
EquiformerV2	Coverage-first	OOD-Ads	0.8771 ± 0.0425	0.0675 ± 0.0002
EquiformerV2	Coverage-first	OOD-Cat	0.7263 ± 0.0198	0.0647 ± 0.0007
EquiformerV2	Coverage-first	OOD-Both	1.0090 ± 0.0305	0.0802 ± 0.0002
EquiformerV2	Coverage-tail	ID	0.8357 ± 0.0286	0.0793 ± 0.0008
EquiformerV2	Coverage-tail	OOD-Ads	0.9418 ± 0.0545	0.0854 ± 0.0013
EquiformerV2	Coverage-tail	OOD-Cat	0.8061 ± 0.0284	0.0784 ± 0.0008
EquiformerV2	Coverage-tail	OOD-Both	1.0355 ± 0.0542	0.0951 ± 0.0010

Model	Curriculum	Split	Energy MAE	Force MAE
GemNet-OC	Uniform	ID	0.7763 ± 0.0193	0.0685 ± 0.0012
GemNet-OC	Uniform	OOD-Ads	0.8627 ± 0.0323	0.0689 ± 0.0018
GemNet-OC	Uniform	OOD-Cat	0.7592 ± 0.0111	0.0683 ± 0.0014
GemNet-OC	Uniform	OOD-Both	0.9978 ± 0.0559	0.0816 ± 0.0017
GemNet-OC	Easy→hard	ID	0.8674 ± 0.0733	0.0844 ± 0.0078
GemNet-OC	Easy→hard	OOD-Ads	0.8716 ± 0.0603	0.0859 ± 0.0090
GemNet-OC	Easy→hard	OOD-Cat	0.8474 ± 0.0785	0.0841 ± 0.0075
GemNet-OC	Easy→hard	OOD-Both	0.9713 ± 0.0529	0.0952 ± 0.0081
GemNet-OC	Hard→easy	ID	0.7899 ± 0.0294	0.0785 ± 0.0005
GemNet-OC	Hard→easy	OOD-Ads	0.8939 ± 0.0228	0.0761 ± 0.0004
GemNet-OC	Hard→easy	OOD-Cat	0.7759 ± 0.0266	0.0778 ± 0.0005
GemNet-OC	Hard→easy	OOD-Both	1.0801 ± 0.0298	0.0902 ± 0.0005
GemNet-OC	Coverage-first	ID	0.7725 ± 0.0193	0.0682 ± 0.0004
GemNet-OC	Coverage-first	OOD-Ads	0.8940 ± 0.0378	0.0678 ± 0.0001
GemNet-OC	Coverage-first	OOD-Cat	0.7680 ± 0.0198	0.0680 ± 0.0002
GemNet-OC	Coverage-first	OOD-Both	1.0275 ± 0.0636	0.0806 ± 0.0002
GemNet-OC	Coverage-tail	ID	0.8523 ± 0.0367	0.0795 ± 0.0026
GemNet-OC	Coverage-tail	OOD-Ads	0.8750 ± 0.0208	0.0815 ± 0.0021
GemNet-OC	Coverage-tail	OOD-Cat	0.8272 ± 0.0375	0.0789 ± 0.0026
GemNet-OC	Coverage-tail	OOD-Both	0.9839 ± 0.0314	0.0913 ± 0.0024
PaiNN	Uniform	ID	1.4853 ± 0.1590	0.0771 ± 0.0020
PaiNN	Uniform	OOD-Ads	1.5933 ± 0.2114	0.0755 ± 0.0011
PaiNN	Uniform	OOD-Cat	1.4503 ± 0.1432	0.0767 ± 0.0020
PaiNN	Uniform	OOD-Both	1.8239 ± 0.2618	0.0891 ± 0.0014
PaiNN	Easy→hard	ID	1.6657 ± 0.1872	0.0885 ± 0.0029
PaiNN	Easy→hard	OOD-Ads	1.6982 ± 0.1225	0.0897 ± 0.0022
PaiNN	Easy→hard	OOD-Cat	1.6502 ± 0.1247	0.0880 ± 0.0030
PaiNN	Easy→hard	OOD-Both	1.9058 ± 0.0589	0.0990 ± 0.0016
PaiNN	Hard→easy	ID	1.8652 ± 0.1951	0.0822 ± 0.0006
PaiNN	Hard→easy	OOD-Ads	1.9921 ± 0.1297	0.0795 ± 0.0004
PaiNN	Hard→easy	OOD-Cat	1.8826 ± 0.1435	0.0816 ± 0.0006
PaiNN	Hard→easy	OOD-Both	2.4123 ± 0.1288	0.0937 ± 0.0004
PaiNN	Coverage-first	ID	1.7210 ± 0.4094	0.0777 ± 0.0015
PaiNN	Coverage-first	OOD-Ads	1.8564 ± 0.3944	0.0754 ± 0.0006
PaiNN	Coverage-first	OOD-Cat	1.7350 ± 0.4578	0.0774 ± 0.0014
PaiNN	Coverage-first	OOD-Both	2.2879 ± 0.6241	0.0892 ± 0.0008
PaiNN	Coverage-tail	ID	1.9690 ± 0.7727	0.0861 ± 0.0057
PaiNN	Coverage-tail	OOD-Ads	2.0834 ± 0.8447	0.0869 ± 0.0055
PaiNN	Coverage-tail	OOD-Cat	1.9733 ± 0.8920	0.0854 ± 0.0051
PaiNN	Coverage-tail	OOD-Both	2.4315 ± 1.2227	0.0970 ± 0.0048
SchNet	Uniform	ID	1.8950 ± 0.0128	0.1124 ± 0.0006
SchNet	Uniform	OOD-Ads	2.0354 ± 0.0056	0.1095 ± 0.0006
SchNet	Uniform	OOD-Cat	1.9053 ± 0.0184	0.1290 ± 0.0012
SchNet	Uniform	OOD-Both	2.3017 ± 0.0321	0.1381 ± 0.0005
SchNet	Easy→hard	ID	2.2363 ± 0.0146	0.1160 ± 0.0008
SchNet	Easy→hard	OOD-Ads	2.1824 ± 0.0110	0.1136 ± 0.0006
SchNet	Easy→hard	OOD-Cat	2.1971 ± 0.0257	0.1320 ± 0.0018
SchNet	Easy→hard	OOD-Both	2.3473 ± 0.0289	0.1403 ± 0.0009
SchNet	Hard→easy	ID	1.9165 ± 0.0103	0.1199 ± 0.0008
SchNet	Hard→easy	OOD-Ads	2.0643 ± 0.0111	0.1161 ± 0.0008
SchNet	Hard→easy	OOD-Cat	1.9483 ± 0.0104	0.1374 ± 0.0022
SchNet	Hard→easy	OOD-Both	2.4036 ± 0.0134	0.1460 ± 0.0012
SchNet	Coverage-first	ID	1.9297 ± 0.0078	0.1134 ± 0.0004
SchNet	Coverage-first	OOD-Ads	2.1139 ± 0.0161	0.1103 ± 0.0004
SchNet	Coverage-first	OOD-Cat	1.9651 ± 0.0334	0.1298 ± 0.0020
SchNet	Coverage-first	OOD-Both	2.4572 ± 0.0803	0.1389 ± 0.0012
SchNet	Coverage-tail	ID	2.2329 ± 0.0162	0.1158 ± 0.0012
SchNet	Coverage-tail	OOD-Ads	2.1787 ± 0.0171	0.1134 ± 0.0010
SchNet	Coverage-tail	OOD-Cat	2.2158 ± 0.0219	0.1310 ± 0.0021
SchNet	Coverage-tail	OOD-Both	2.3583 ± 0.0297	0.1396 ± 0.0012

E Complete Bounded Paired Effects

Table 8 aggregates the 192 paired comparisons into 64 model–curriculum–split rows over three repetitions. F_0 and E_0 denote the uniform reference; F_1 and E_1 denote the curriculum target. Relative effects are averaged after cell-wise pairing; negative values favor the target curriculum.

Table 8: Bounded paired effects across three repetitions. F_0, F_1 are in eV Å^{-1} and E_0, E_1 in eV.

Model	Curriculum	Split	F_0	F_1	ΔF (%)	E_0	E_1	ΔE (%)
EquiformerV2	Easy→hard	ID	0.0637	0.0802	+25.97	0.6630	0.8551	+29.06
EquiformerV2	Easy→hard	OOD-Ads	0.0665	0.0871	+30.94	0.8051	0.9812	+22.09
EquiformerV2	Easy→hard	OOD-Cat	0.0633	0.0795	+25.59	0.6549	0.8328	+27.27
EquiformerV2	Easy→hard	OOD-Both	0.0791	0.0967	+22.27	0.9302	1.0806	+16.31
EquiformerV2	Hard→easy	ID	0.0637	0.0778	+22.21	0.6630	0.6348	-4.22
EquiformerV2	Hard→easy	OOD-Ads	0.0665	0.0754	+13.39	0.8051	0.7466	-7.04
EquiformerV2	Hard→easy	OOD-Cat	0.0633	0.0770	+21.71	0.6549	0.6396	-2.30
EquiformerV2	Hard→easy	OOD-Both	0.0791	0.0895	+13.20	0.9302	0.9270	-0.20
EquiformerV2	Coverage-first	ID	0.0637	0.0648	+1.79	0.6630	0.7366	+11.23
EquiformerV2	Coverage-first	OOD-Ads	0.0665	0.0675	+1.47	0.8051	0.8771	+9.22
EquiformerV2	Coverage-first	OOD-Cat	0.0633	0.0647	+2.23	0.6549	0.7263	+11.07
EquiformerV2	Coverage-first	OOD-Both	0.0791	0.0802	+1.36	0.9302	1.0090	+8.71
EquiformerV2	Coverage-tail	ID	0.0637	0.0793	+24.56	0.6630	0.8357	+26.08
EquiformerV2	Coverage-tail	OOD-Ads	0.0665	0.0854	+28.34	0.8051	0.9418	+16.97
EquiformerV2	Coverage-tail	OOD-Cat	0.0633	0.0784	+23.87	0.6549	0.8061	+23.12
EquiformerV2	Coverage-tail	OOD-Both	0.0791	0.0951	+20.25	0.9302	1.0355	+11.32
GemNet-OC	Easy→hard	ID	0.0685	0.0844	+23.38	0.7763	0.8674	+11.65
GemNet-OC	Easy→hard	OOD-Ads	0.0689	0.0859	+24.79	0.8627	0.8716	+0.97
GemNet-OC	Easy→hard	OOD-Cat	0.0683	0.0841	+23.32	0.7592	0.8474	+11.53
GemNet-OC	Easy→hard	OOD-Both	0.0816	0.0952	+16.72	0.9978	0.9713	-2.61
GemNet-OC	Hard→easy	ID	0.0685	0.0785	+14.64	0.7763	0.7899	+1.75
GemNet-OC	Hard→easy	OOD-Ads	0.0689	0.0761	+10.50	0.8627	0.8939	+3.66
GemNet-OC	Hard→easy	OOD-Cat	0.0683	0.0778	+14.00	0.7592	0.7759	+2.19
GemNet-OC	Hard→easy	OOD-Both	0.0816	0.0902	+10.53	0.9978	1.0801	+8.38
GemNet-OC	Coverage-first	ID	0.0685	0.0682	-0.32	0.7763	0.7725	-0.43
GemNet-OC	Coverage-first	OOD-Ads	0.0689	0.0678	-1.65	0.8627	0.8940	+3.65
GemNet-OC	Coverage-first	OOD-Cat	0.0683	0.0680	-0.37	0.7592	0.7680	+1.16
GemNet-OC	Coverage-first	OOD-Both	0.0816	0.0806	-1.20	0.9978	1.0275	+3.10
GemNet-OC	Coverage-tail	ID	0.0685	0.0795	+16.10	0.7763	0.8523	+9.78
GemNet-OC	Coverage-tail	OOD-Ads	0.0689	0.0815	+18.26	0.8627	0.8750	+1.49
GemNet-OC	Coverage-tail	OOD-Cat	0.0683	0.0789	+15.56	0.7592	0.8272	+8.93
GemNet-OC	Coverage-tail	OOD-Both	0.0816	0.0913	+11.86	0.9978	0.9839	-1.24
PaiNN	Easy→hard	ID	0.0771	0.0885	+14.77	1.4853	1.6657	+12.51
PaiNN	Easy→hard	OOD-Ads	0.0755	0.0897	+18.92	1.5933	1.6982	+7.49
PaiNN	Easy→hard	OOD-Cat	0.0767	0.0880	+14.77	1.4503	1.6502	+14.20
PaiNN	Easy→hard	OOD-Both	0.0891	0.0990	+11.05	1.8239	1.9058	+5.90
PaiNN	Hard→easy	ID	0.0771	0.0822	+6.72	1.4853	1.8652	+25.80
PaiNN	Hard→easy	OOD-Ads	0.0755	0.0795	+5.37	1.5933	1.9921	+26.02
PaiNN	Hard→easy	OOD-Cat	0.0767	0.0816	+6.45	1.4503	1.8826	+30.25
PaiNN	Hard→easy	OOD-Both	0.0891	0.0937	+5.15	1.8239	2.4123	+34.07
PaiNN	Coverage-first	ID	0.0771	0.0777	+0.80	1.4853	1.7210	+17.28
PaiNN	Coverage-first	OOD-Ads	0.0755	0.0754	-0.13	1.5933	1.8564	+18.66
PaiNN	Coverage-first	OOD-Cat	0.0767	0.0774	+0.98	1.4503	1.7350	+21.09
PaiNN	Coverage-first	OOD-Both	0.0891	0.0892	+0.08	1.8239	2.2879	+27.49
PaiNN	Coverage-tail	ID	0.0771	0.0861	+11.71	1.4853	1.9690	+32.06
PaiNN	Coverage-tail	OOD-Ads	0.0755	0.0869	+15.20	1.5933	2.0834	+30.15
PaiNN	Coverage-tail	OOD-Cat	0.0767	0.0854	+11.40	1.4503	1.9733	+35.53
PaiNN	Coverage-tail	OOD-Both	0.0891	0.0970	+8.79	1.8239	2.4315	+30.96
SchNet	Easy→hard	ID	0.1124	0.1160	+3.20	1.8950	2.2363	+18.01
SchNet	Easy→hard	OOD-Ads	0.1095	0.1136	+3.78	2.0354	2.1824	+7.22
SchNet	Easy→hard	OOD-Cat	0.1290	0.1320	+2.30	1.9053	2.1971	+15.33
SchNet	Easy→hard	OOD-Both	0.1381	0.1403	+1.57	2.3017	2.3473	+2.00
SchNet	Hard→easy	ID	0.1124	0.1199	+6.67	1.8950	1.9165	+1.14
SchNet	Hard→easy	OOD-Ads	0.1095	0.1161	+6.09	2.0354	2.0643	+1.42
SchNet	Hard→easy	OOD-Cat	0.1290	0.1374	+6.51	1.9053	1.9483	+2.26
SchNet	Hard→easy	OOD-Both	0.1381	0.1460	+5.72	2.3017	2.4036	+4.44
SchNet	Coverage-first	ID	0.1124	0.1134	+0.86	1.8950	1.9297	+1.83
SchNet	Coverage-first	OOD-Ads	0.1095	0.1103	+0.79	2.0354	2.1139	+3.86
SchNet	Coverage-first	OOD-Cat	0.1290	0.1298	+0.59	1.9053	1.9651	+3.14
SchNet	Coverage-first	OOD-Both	0.1381	0.1389	+0.58	2.3017	2.4572	+6.77
SchNet	Coverage-tail	ID	0.1124	0.1158	+3.02	1.8950	2.2329	+17.84
SchNet	Coverage-tail	OOD-Ads	0.1095	0.1134	+3.56	2.0354	2.1787	+7.04
SchNet	Coverage-tail	OOD-Cat	0.1290	0.1310	+1.50	1.9053	2.2158	+16.30
SchNet	Coverage-tail	OOD-Both	0.1381	0.1396	+1.09	2.3017	2.3583	+2.46

F Complete Bounded Energy–Force Aggregates

Table 9 reports all 16 model–curriculum aggregates used by the Energy–Force Pareto analysis. Each row averages three repetitions and four validation splits (12 paired cells); it is an aggregate diagnostic, not a replacement for the long table above.

Table 9: Complete bounded Energy–Force aggregates.

Model	Curriculum	Force Δ (%)	Energy Δ (%)	Reps.	Splits	Cells
EquiformerV2	Easy→hard	+26.19	+23.68	3	4	12
EquiformerV2	Hard→easy	+17.63	-3.44	3	4	12
EquiformerV2	Coverage-first	+1.71	+10.06	3	4	12
EquiformerV2	Coverage-tail	+24.25	+19.37	3	4	12
GemNet-OC	Easy→hard	+22.05	+5.38	3	4	12
GemNet-OC	Hard→easy	+12.42	+3.99	3	4	12
GemNet-OC	Coverage-first	-0.88	+1.87	3	4	12
GemNet-OC	Coverage-tail	+15.44	+4.74	3	4	12
PaiNN	Easy→hard	+14.88	+10.02	3	4	12
PaiNN	Hard→easy	+5.92	+29.04	3	4	12
PaiNN	Coverage-first	+0.44	+21.13	3	4	12
PaiNN	Coverage-tail	+11.77	+32.18	3	4	12
SchNet	Easy→hard	+2.71	+10.64	3	4	12
SchNet	Hard→easy	+6.25	+2.31	3	4	12
SchNet	Coverage-first	+0.71	+3.90	3	4	12
SchNet	Coverage-tail	+2.29	+10.91	3	4	12

G Complete Full-Evaluation Results

The full comparison contains 4 models × 4 curricula ×3 repetitions ×4 validation splits = 192 cells. Comparing each of the three staged curricula with uniform gives $4 \times 3 \times 3 \times 4 = 144$ paired effects. The subsequent coverage-first evaluation adds 48 cells from the existing checkpoints. Tables 10 and 11 summarize the two comparisons; Table 12 gives the validation split sizes.

Table 10: Full-split mean Force effects for the three staged curricula versus paired uniform ordering. Positive values indicate higher error; each curriculum mean averages four model-level summaries, each based on $n = 3$ seeds × 4 validation splits, and the win count spans 16 model–split units.

Curriculum	Mean relative Force change (%)	Uniform wins
Easy→hard	+16.6407	16/16
Hard→easy	+10.1092	16/16
Coverage→tail	+13.5334	16/16

Table 11: Post-hoc targeted full-split confirmation of coverage-first. Entries are mean cell-wise relative changes versus paired uniform ordering; positive values indicate higher error. The full-split summaries use 12 cells per model (3 training seeds × 4 validation splits).

Model	Force change (%)	Energy change (%)	Force improved	Energy improved
EquiformerV2	+1.73	+10.13	3/12	0/12
GemNet-OC	-0.92	+2.00	10/12	4/12
PaiNN	+0.34	+21.66	6/12	4/12
SchNet	+0.74	+3.95	2/12	0/12
All 48 cells	+0.47	+9.43	21/48	8/48

Effects are averaged after pairing each curriculum with the uniform run for the same model, repetition, and split. The accompanying tables data/full_eval_summary/full_metrics.tsv and data/full_eval_summary/full_paired_effects.tsv contain the 192 absolute and 144 paired results. Coverage-first values are in data/targeted_full_eval/targeted_metrics.tsv and data/targeted_full_eval/targeted_paired_effects.tsv.

Table 12: Complete validation split sizes.

Validation split	Full sample count
ID	999,866
OOD-Ads	999,838
OOD-Cat	999,809
OOD-Both	999,944

Table 13: Full-split coverage-to-tail Force effects versus paired uniform ordering. Positive values indicate higher error for coverage-to-tail; values are relative changes in percent.

Model or summary	Force relative change (%)
EquiformerV2	+23.9264
GemNet-OC	+15.7585
PaiNN	+12.2426
SchNet	+2.2060
Overall (48 paired cells)	+13.5334
Descriptive bootstrap 95% interval	[+5.3337, +21.7820]
Uniform wins (model–split units)	16/16
Energy aggregate relative regression (%)	+17.6605

H Stage-End Evaluation

Energy and Force MAE are recorded at the four stage endpoints, corresponding to 50k, 100k, 150k, and 200k cumulative presentations in the core experiment. For metric M, the endpoint change is $M_{t+1} - M_t$ for $t = 1, 2, 3$. These values describe the net change over a complete stage. Isolating forgetting and recovery would additionally require measurements bracketing the same distribution switch and following the subsequent updates.

I 500k Scale Extension

The scale extension tests a larger nested training pool while retaining the matched-support ordering intervention. It uses the seed-43 200k pool as a set-level parent of the 500k pool. Each budget uses its own permutation of the corresponding pool.

Factorial and support. The 500k design crosses uniform, coverage-first, and coverage-to-tail with EquiformerV2, GemNet-OC, PaiNN, and SchNet at seed 43. The pool contains 500,000 unique records divided into four 125,000-record stages. All 48 bounded evaluation cells are complete, with four splits and 10,000 validation records per split. Both budgets use the same validation indices and the evaluation procedure described in Appendix B. The 200k anchor is the three-curriculum, seed-43 subset of the core study.

Table 14: Training and evaluation dimensions at each budget.

Budget	Models	Curricula	Repetitions	Training identities	Bounded cells
200k core study	4	5	3	60	240
500k scale	4	3	1	12	48

Cross-budget results. Within each budget, curriculum effects use the same-budget uniform run as the reference. The matched seed-43 200k-to-500k contrast is reported separately from those ordering effects. Table 16 reports the corresponding absolute errors. The 48 split-wise pairs are provided in data/scale_500k/budget_comparison.tsv; same-budget relative effects are in data/scale_500k/scale_paired_effects.tsv.

Table 15: 500k same-budget relative MAE changes versus uniform. Values are means of four split-wise paired effects; positive values indicate higher error.

Model	Curriculum	Δ^F (%)	Δ^E (%)
EquiformerV2	Coverage-first	+6.6022	+4.5869
EquiformerV2	Coverage-to-tail	+45.8156	+22.1194
GemNet-OC	Coverage-first	+0.3161	+5.8592
GemNet-OC	Coverage-to-tail	+18.5969	+30.5697
PaiNN	Coverage-first	+0.1358	-24.3021
PaiNN	Coverage-to-tail	+30.9353	-13.1978
SchNet	Coverage-first	+0.5118	+6.4261
SchNet	Coverage-to-tail	+5.0437	+13.6161

Table 16: Absolute errors at the two training budgets, averaged over the same four validation subsets at seed 43. Energy is in eV and Force in meV Å^{-1}.

		Force MAE		Energy MAE	
Model	Curriculum	200k	500k	200k	500k
EquiformerV2	Uniform	68.250	58.575	0.7646	0.6228
EquiformerV2	Coverage-first	69.550	62.400	0.8174	0.6506
EquiformerV2	Coverage-to-tail	84.275	85.125	0.8834	0.7536
GemNet-OC	Uniform	71.425	62.775	0.8688	0.7022
GemNet-OC	Coverage-first	71.275	63.000	0.9010	0.7440
GemNet-OC	Coverage-to-tail	84.575	74.325	0.9207	0.9099
PaiNN	Uniform	80.725	74.325	1.3671	1.4590
PaiNN	Coverage-first	81.100	74.425	2.0162	1.1041
PaiNN	Coverage-to-tail	85.125	97.125	1.6271	1.2578
SchNet	Uniform	121.800	87.675	2.0299	1.8979
SchNet	Coverage-first	122.875	88.125	2.0833	2.0242
SchNet	Coverage-to-tail	123.550	92.025	2.2232	2.1465

J Statistical Analysis

Paired effects and aggregation. Energy and Force effects are computed as $100(M_{\text{curriculum}} - M_{\text{uniform}})/M_{\text{uniform}}$ within each model, repetition, validation split, and training budget. At 200k, model-level effects average twelve paired values from three repetitions and four splits; at 500k, they average four split-wise values. Overall effects give each model equal weight. The model–split winner is the ordering with the lowest mean Force MAE across repetitions. An Energy–Force aggregate is jointly favorable when both mean relative effects are negative.

Model-level intervals. The intervals in Figures 2 and 4 use 2,000 and 10,000 bootstrap draws, respectively, with random seed 20260821. Each draw samples three training-seed clusters with replacement and retains all four validation splits within each selected cluster. The 2.5th and 97.5th percentiles are reported as descriptive intervals.

Overall interval. The coverage-to-tail interval in Table 13 uses 10,000 crossed bootstrap draws with random seed 20260821. Each draw samples four model indices, three seed indices, and four split indices with replacement, then averages the 48 relative effects at their Cartesian product. The sampled seed and split indices are shared across models, preserving the crossed pairing. The 2.5th and 97.5th percentiles give the descriptive range $[+5.3337\%, +21.7820\%]$.

Success criteria. The prespecified coverage-to-tail comparison requires lower mean absolute Force MAE for at least three of four models and at least three of four validation splits, a negative upper endpoint of the overall Force interval, and an aggregate Energy increase of at most 2%. For the Energy criterion, the aggregate is the ratio of mean Energy MAEs across all 48 pairs, minus one. This differs from the mean cellwise relative effect used in the comparison tables. Coverage-to-tail meets none of these four criteria.

Budget comparisons. The 200k and 500k runs are paired by model, curriculum, and validation split at seed 43. Same-budget effects use uniform at that budget as the reference; cross-budget changes compare the absolute MAEs of the same curriculum. The relative budget effects in the tables are means of split-wise percentages.

K Computational Resources

Each training and evaluation run uses one NVIDIA H100 GPU with 80 GB memory. The environment uses Python 3.9.25, PyTorch 2.0.1, and CUDA 11.8. The core comparison comprises 60 training runs and 240 bounded evaluations; full-split evaluation comprises 192 initial evaluations and 48 subsequent coverage-first evaluations. The 500k extension adds 12 training runs and 48 bounded evaluations. Each bounded evaluation processes 10,000 structures.

Table 17: Optimizer updates per training run. Each budget is divided into four equal presentation stages; the final minibatch of each stage is retained.

Model	200k presentations	500k presentations
EquiformerV2	50,000	125,000
GemNet-OC	12,500	31,252
PaiNN	6,252	15,628
SchNet	8,336	20,836

L Reproducibility

The anonymous repository provides model configurations, curriculum code, analysis utilities, and aggregate result summaries.

Code availability. Code and reproduction instructions are available at https://anonymous.4open.science/r/mlp-curriculum-learning/.