Reactive Force Error and Reaction-Path Completion in Pretrained Potentials

Anonymous authors
Paper under double-blind review

Abstract

We test local force response, reactive force accuracy and reaction-path completion in pretrained interatomic potentials. Across five broad-domain checkpoints, local force-response directions closely match their references, while forward barriers are compressed on 200 Transition1x reactions (slopes 0.660–0.857). On near-path configurations, median absolute-force cosines range from 0.311 to 0.605. Under the primary fixed-endpoint CI-NEB protocol, 3–7 of 20 paths converge per checkpoint. Local response, reactive forces and path completion therefore need separate evaluation.

Keywords: machine learning interatomic potentials, reaction pathways, force evaluation, pretrained models, nudged elastic band

TL;DR: Local force-response agreement coexists with compressed reaction barriers and limited fixed-endpoint CI-NEB completion.

1 Introduction

Pretrained interatomic potentials make energy and force evaluation available across broad chemical spaces at a small fraction of the cost of electronic-structure calculations. MACE, SevenNet, MatterSim, M3GNet, CHGNet, Orb and UMA use distinct model and data pipelines (Batatia et al., 2022; 2025a; Park et al., 2024; Yang et al., 2024; Chen & Ong, 2022; Deng et al., 2023; Rhodes et al., 2025; Wood et al., 2025). Using these potentials for reactions requires reliable forces away from local molecular geometries and an optimization procedure that can recover a path between endpoints.

Landscape17 shows that accurate local tests can miss errors in complete reaction networks (Cărare et al., 2026). We ask which local-to-reactive differences survive when the comparison uses the same checkpoint head, quantum-chemical reference and force definition. These controls matter because local displacement responses and absolute reactive forces measure different quantities, while a stalled path optimizer introduces another source of failure.

An exploratory matched comparison finds higher reactive force-component error. With SevenNet-Omni's same molecular head and reference level, restricting both domains to common force-magnitude support leaves a 0.0566 $[0.0235, 0.0849]$ eV Å^{-1} MAE gap. The force-direction difference has an interval spanning zero.

Our contributions connect this controlled force comparison to reaction-path evaluation:

1. **A matched estimate of reactive force error.** Same-geometry reference bridges, checkpoint-head comparisons and common-support standardization separate reference, model and force-definition effects. The positive component-MAE contrast survives every single-cluster deletion in the evaluated sample.
2. **Separate tests of prediction and path completion.** Five broad-domain checkpoints compress barriers across 200 reactions despite close local directional agreement. A paired CI-NEB experiment compares optimizers within each checkpoint's runtime, with path completion reported by cohort.

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.

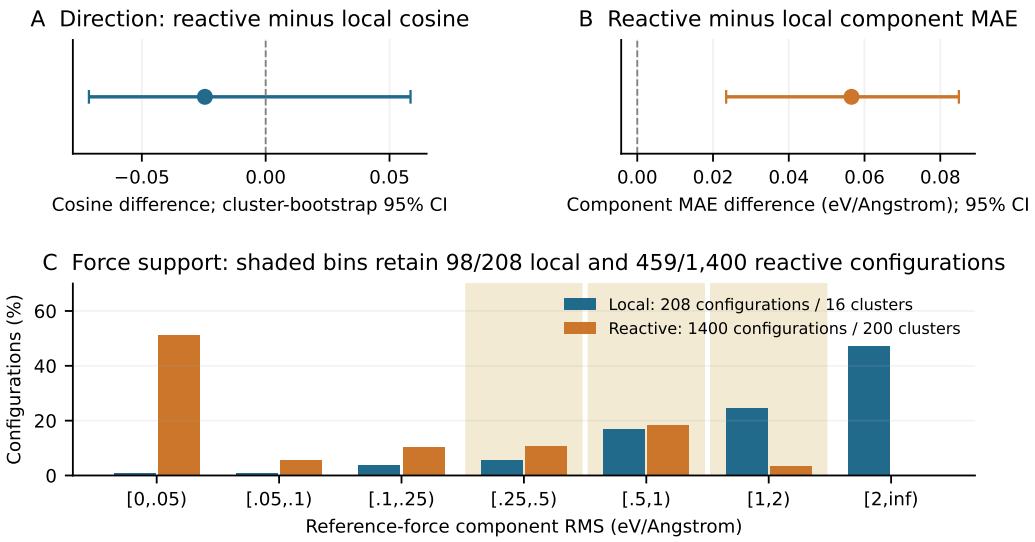

Figure 1: Matched-head comparison of force direction and component error. SevenNet-Omni uses the same `omol25_low` head and ωB97X reference in both domains. (A,B) Common-support reactive-minus-local absolute-force contrasts, with independent molecule/reaction cluster-bootstrap 95% intervals. (C) The complete force-magnitude distributions; shaded bins define the reported support. Configurations outside those bins remain in the unstandardized results. Counts and sample-influence analyses are in Appendix G.

An unfitted pair-repulsion control further shows why scramble sensitivity alone cannot validate physical fidelity (Appendix D). The main evaluation uses explicit electronic-structure references throughout.

2 Reference-Grounded Verification Benchmarks

2.1 Molecules and internal perturbation paths

We use 16 neutral H/C/N/O molecules from structures distributed with ASE's G2 collection (Larsen et al., 2017): four diatomics and twelve polyatomics, listed in Appendix A. The base geometries are used as distributed; we do not optimize them separately for any method. This avoids method-dependent coordinates and makes every comparison a response about exactly the same q_0.

For molecule m and direction index $d \in \{1, 2\}$, a SHA-256-derived seed generates a standard-normal Cartesian vector $z_{m,d}$. We form an orthonormal basis Q_m for the identifiable rigid translations and rotations about the geometric center, then project

$$u_{m,d} = \frac{(I - Q_m Q_m^\top) z_{m,d}}{\sqrt{N_m^{-1} \|(I - Q_m Q_m^\top) z_{m,d}\|_2^2}}. \tag{1}$$

Rank-deficient rotations of linear molecules are removed by singular-value decomposition. The sampled configurations are $q_{m,d,a,s} = q_{m,0} + s a u_{m,d}$ for $a \in \{0.02, 0.05, 0.10\}$ Å RMS per atom and $s \in \{-1, +1\}$. The ± 0.10 Å path endpoints must satisfy $r_{ij} \geq \max[0.50\,\text{Å}, 0.55(r_{\text{cov},i} + r_{\text{cov},j})]$; otherwise the direction is deterministically redrawn. Thus every molecule contributes one base and 12 displaced configurations, for 208 reference calculations. Three additional random rigid rotations per molecule test energy invariance and force equivariance in each MLIP but require no new reference.

Table 1: Evaluated broad-domain families. Each covers every registered static molecular and reaction configuration. Coverage counts finite predictions.

Family	Molecule	Reaction	dtype
MACE	complete	complete	float64
SevenNet	complete	complete	float32
MatterSim	complete	complete	float32
M3GNet	complete	complete	float32
Orb	complete	complete	float32

2.2 Molecular references and evaluated checkpoints

Reference energies and analytic gradients are computed in PySCF 2.14 (Sun et al., 2020) with PBE (Perdew et al., 1996), def2-SVP (Weigend & Ahlrichs, 2005), density fitting, grid level 3, and SCF tolerance 10^{-9} Hartree. O_2 uses UKS with spin 2; all other molecules use RKS with spin 0. We also evaluate two reference levels using samples and settings specified before computation. PBE0/def2-TZVP (Adamo & Barone, 1999) repeats all 208 energy-and-gradient calculations under the same grid and SCF rules. RHF-CCSD(T)/cc-pVTZ (Raghavachari et al., 1989) evaluates energies for 72 registered configurations from eight closed-shell molecules, excluding C/N/O $1s$ orbitals from electron correlation. We retain all configurations and report the conventional T_1 diagnostic as a sensitivity analysis; the CC subset provides energy comparisons. Each level defines a computational reference.

Table 1 identifies the five broad-domain model families evaluated on molecular and reaction inputs. Exact checkpoint revisions, weight hashes, packages, requested precision, and observed parameter dtypes are recorded in manifests. Checkpoint and head comparisons are exploratory (Appendix F).

2.3 Endpoints and uncertainty

For each displaced configuration, let $x = \Delta E_{\text{ref}}/N$ and $y = \Delta E_{\text{MLIP}}/N$. The primary energy scale is the through-origin slope $\beta_E = (\sum xy)/(\sum x^2)$; companion measures are correlation, MAE, NRMSE, and the central curvature $\kappa(a) = [E(+a) + E(-a) - 2E(0)]/(Na^2)$ with its slope and sign agreement.

For forces, we compare responses rather than absolute base forces: $\Delta\mathbf{F}(q) = \mathbf{F}(q) - \mathbf{F}(q_0)$, summarized by full-vector cosine, norm ratio, and component MAE. The prespecified 10^{-8} eV Å^{-1} cosine exclusion removes no row. Uncertainty uses 10,000 fixed-seed draws over whole molecule or reaction clusters; all compatible checkpoints and registered rows are retained. The local panels contain 16 molecule clusters, or eight for CCSD(T); percentile intervals from these small panels can under-cover. The fixed-head comparison additionally has cluster-deletion checks.

Samples, reference settings and primary metrics were specified before their respective evaluations. The matched-force, basis, amplitude, role, checkpoint and optimizer comparisons are exploratory. Appendix F describes the analysis design.

2.4 The reactive-domain benchmark

We use the official Transition1x test split (Schreiner et al., 2022), whose reference energies and forces were computed at ωB97x/6-31G(d). Eligibility was fixed to reactions containing only H/C/N/O with 3–7 C/N/O heavy atoms, finite reactant (R), transition-state (TS), product (P), and intermediate arrays, consistent atom order, and at least four non-colliding intermediate configurations. A barrier/heavy-atom stratification followed by a fixed SHA-256 ordering selected 200 of 287 eligible reactions. For each reaction we retain R/TS/P and four deterministic near-path configurations, giving 1,400 formal structures per model. The intermediate arrays pool saved NEB iterations and are therefore treated as an unordered near-path cloud, not an intrinsic reaction coordinate.

Energies are zeroed at each model's reactant endpoint. Primary measures are forward- and reverse-barrier and reaction-energy error, through-origin slope, and TS ordering; near-path measures are energy NRMSE plus absolute-force component error, cosine, and norm ratio. Intervals use 10,000

reaction-cluster bootstrap draws (seed 2027). A separate 20-reaction hash-selected set uses identical ten-image, fixed-endpoint, IDPP-initialized CI-NEB calculations; only paths reaching the fixed final force threshold contribute optimized barriers. Path completion throughout denotes convergence of the MLIP CI-NEB optimization. The Transition1x R/P geometries remain fixed across checkpoints. A follow-up validation cohort uses ranks 21–100 of the same model-independent hash order: 80 additional reactions per model under the unchanged BFGS protocol. Optimizer sensitivity uses ASE FIRE on all 20 primary reactions with the same initialization, tangent, spring, thresholds and step caps. A paired experiment also runs both optimizers within each checkpoint's runtime (Appendix I). Results are reported separately by cohort and runtime.

2.5 Crossed reference bridge

A descriptive 2×2 reference comparison uses samples selected independently of model predictions: R1a evaluates all 208 local configurations at ωB97X/6-31G(d), and R1b evaluates 224 configurations from a hash-stratified 32-reaction subset at PBE0/def2-TZVP. The five checkpoints are rescored under both references on identical geometries. R1c adds RHF-CCSD(T)/cc-pVTZ R/TS/P energies for 12 reactions, excluding C/N/O $1s$ correlation. All reference calculations use CPU PySCF 2.14.0. Intervals resample molecule or reaction clusters. R1b pools all seven geometry roles. An exploratory stratification separates four near-path roles from R/TS/P stationary-point geometries using their geometry labels. Both the pooled and stratified estimates are reported.

3 Results

3.1 Completeness, symmetry, and the negative control

All 208/208 PBE/def2-SVP calculations converge, and each MLIP returns finite energies and forces on all 256 configurations. On 48 rotated structures per model, maximum per-atom energy residuals are 5.0×10^{-13}, 9.54×10^{-7}, and 2.38×10^{-6} eV for MACE, SevenNet, and MatterSim. Corresponding maximum force- equivariance errors are 9.56×10^{-11}, 1.22×10^{-5}, and 1.67×10^{-4} eV Å^{-1}. These residuals quantify numerical invariance and equivariance at each model's precision.

As a negative control, we applied Equation 3 to the same local paths. It is sensitive to scrambling but has energy slope 20.54 [5.52, 24.38], NRMSE 26.61 [8.07, 30.37], force norm ratio 4.27, and curvature slope 13.80 against PBE/def2-SVP. Its median force cosine is nevertheless 0.961. High directional agreement therefore coexists with large scale errors.

3.2 Local responses agree directionally but scale is reference-dependent

Across 192 displaced configurations, the three primary models have energy correlation above 0.990, slopes 0.964–1.036, and NRMSE 0.100–0.125; their clustered intervals overlap (Appendix Table 2 and Figure 4). All 96 curvature signs match, with correlations 0.996–0.998 and slopes 0.935–1.034. Median force-response cosine is 0.99939–0.99975, while norm ratios are 0.929–0.945. The models closely agree with local reference directions. Amplitude slices are retained in Appendix B because the largest displacements dominate the pooled slope.

A def2-TZVP repeat changes the reference response scale: relative to def2-SVP its energy slope is 0.9165 [0.8999, 0.9505], while force cosine remains 0.99987. Model slopes then become 1.045–1.123. The basis change shifts response magnitudes while preserving direction and curvature signs (Appendix B).

3.3 Independent reference levels preserve the local directional result

All 208 PBE0 energy- and-gradient calculations and all 72 CCSD(T) energy calculations (64 displaced configurations and 8 base configurations) converge, and each of the five compatible models covers every registered structure. Relative to PBE0, energy- response slopes are 1.010–1.122 and NRMSE is 0.304–0.385 (Appendix Table 3). Median force-response cosines remain 0.9984–0.9996 and every model matches all 96 PBE0 curvature signs. Thus the strong local directional result is not specific to semilocal PBE, although the larger PBE0 energy NRMSE shows greater scale error.

On the smaller CCSD(T) subset, energy slopes are 1.042–1.169, NRMSE is 0.147–0.253, and all 32 curvature signs agree. The T_1 diagnostic has median 0.0091 and maximum 0.0354; 12/72 configurations exceed 0.02 and remain in the primary result. Restricting to $T_1 \leq 0.02$ does not uniformly improve NRMSE, and we report both the full sample and this sensitivity subset. The CC result concerns energies.

3.4 Barrier and force errors on the reactive benchmark

Each checkpoint returns finite energy and force predictions for all 1,400 structures, giving 7,000 predictions across five models. The reference TS is above both endpoints for 200/200 reactions. Nevertheless, every model compresses the reference forward-barrier scale (Appendix Figure 5): through-origin slopes range from 0.660 to 0.857, and the upper bound of every reaction-clustered 95% interval remains below one. Forward- barrier MAE ranges from 0.552 to 1.295 eV and model TS-ordering rates from 87.0% to 99.5%. Appendix Table 29 also reports the reverse-barrier and reaction-energy errors on all 200 reactions per checkpoint.

Directional disagreement is larger away from the fixed endpoints. Near-path energy NRMSE is 0.159–0.396, while median absolute-force cosine falls to 0.311–0.605 and median force- norm ratio rises to 1.431–2.617. In contrast, the local G2 and PBE0 force-response cosines are near one. These domains use different force measures; the reference comparison below evaluates reference dependence, and the matched-force analysis compares absolute forces in both domains.

In the checkpoint comparison (C1), MACE-MH-1 (Batatia et al., 2025b) and SevenNet-Omni (Kim et al., 2026), including molecular-data heads associated with OMol25 (Levine et al., 2025), narrow the static contrast: forward-barrier MAE is 0.268 and 0.275 eV, and near-path force cosine is 0.806 and 0.827. MatterSim-5M gives 0.606 eV and 0.506. Under the same 20-reaction BFGS protocol, MACE-MH-1 and SevenNet-Omni complete 11/20 and 12/20 paths; paired improvements relative to their original family checkpoints have intervals that include or touch zero. Appendix Table 14 reports all three comparisons.

Matched forces retain higher reactive component error. We compute absolute forces in both domains and standardize on common reference-force magnitude bins. With SevenNet-Omni's identical `omol25_low` head and ωB97X/6-31G(d) reference, the common bins retain 98/208 local and 459/1,400 reactive configurations. On this support, reactive component MAE exceeds local MAE by 0.0566 [0.0235, 0.0849] eV Å^{-1}. Before standardization, reactive and local MAEs are 0.1619 and 0.2145 eV Å^{-1}, respectively. Matching force magnitudes reverses this ordering: the positive contrast measures excess reactive error at comparable force scale. The standardized cosine difference is -0.0245 [-0.0714, 0.0585]. The interval includes differences in either direction. Molecules and reactions are resampled as independent clusters; the domains contain different geometries. Appendix G gives the support definition, paired checkpoint changes, and both unstandardized and standardized estimates.

Figure 1 shows the common force-magnitude range. Deleting one molecule at a time gives MAE point contrasts 0.0506–0.0674 eV Å^{-1}; deleting one reaction gives 0.0542–0.0575 eV Å^{-1}. All 216 deletions retain the original support requirements. These ranges describe sensitivity to individual clusters.

3.5 A crossed reference bridge preserves the domain contrast

R1a includes 208/208 local ωB97X calculations, R1b has 224/224 reactive PBE0 calculations over 32 fixed reactions, and R1c has 36/36 CCSD(T) stationary-point calculations over 12 fixed reactions. On identical local geometries, ωB97X relative to PBE0 has energy-response slope 1.099, NRMSE 0.165, median force-response cosine 0.99994, and median force-norm ratio 1.054. On identical reactive geometries, PBE0 relative to the Transition1x ωB97X reference has energy-response slope 0.932 and NRMSE 0.085. Its pre-specified force aggregate over all seven roles has median cosine 0.423 and norm ratio 4.353. The geometry-role stratification gives 0.864 and 1.035 on 128 near-path configurations, versus 0.088 and 14.060 on 96 R/TS/P stationary-point geometries (intervals in Appendix Table 5). Because Transition1x stationary points are ωB97X-optimized but need not

be stationary under PBE0, we interpret these strata as near-path agreement and reference-specific stationarity, respectively, with the pooled estimate reported alongside both strata.

Under the original five checkpoints and force measures, reference changes retain the contrast. Local force-response cosines across five checkpoints are 0.9984–0.9996 under either reference. On the matched 128 near-path configurations, reactive absolute-force cosines are 0.379–0.583 against Transition1x and 0.125–0.278 against PBE0, both well below the 0.864 direct reference-to-reference value (Figure 2). On 12 reactions, Transition1x-minus-CCSD(T) forward-, reverse-, and reaction-energy MAEs are 0.206, 0.091, and 0.178 eV, respectively, with no exclusions. The same-geometry reference comparisons use local displacement responses and reactive absolute forces. The matched-head analysis uses absolute forces in both domains and finds a smaller directional contrast for SevenNet-Omni.

Reaction-level force errors covary with fixed-geometry barrier errors. We average force-component absolute error over the stored near-path configurations of each reaction and compare it with forward-barrier absolute error on all 200 reactions per checkpoint. Spearman correlations range from 0.144 to 0.380; four of five model-adjusted 99% intervals exclude zero. The corresponding force-error difference between nonconverged and converged BFGS paths remains unresolved in all five models on the complete 100-reaction primary-plus-validation inventory. These within-checkpoint associations retain every reaction and distinguish a link to static barrier error from the separate question of path completion. Appendix M reports both endpoint families.

3.6 Climbing-image optimization limits fixed-endpoint completion

All 100 ordinary-NEB stages reach 0.5 eV Å^{-1}. Under the subsequent climbing-image criterion of 0.05 eV Å^{-1} and a further 500-step cap, 3–7 of 20 paths per model converge (15–35%; Figure 3). Converged-only barrier MAE is 0.716–1.533 eV with negative bias in every model. The converged subsets differ across models. Static endpoint error supplies no reliable convergence screen: model-specific AUCs are 0.484–0.720, with every interval spanning chance.

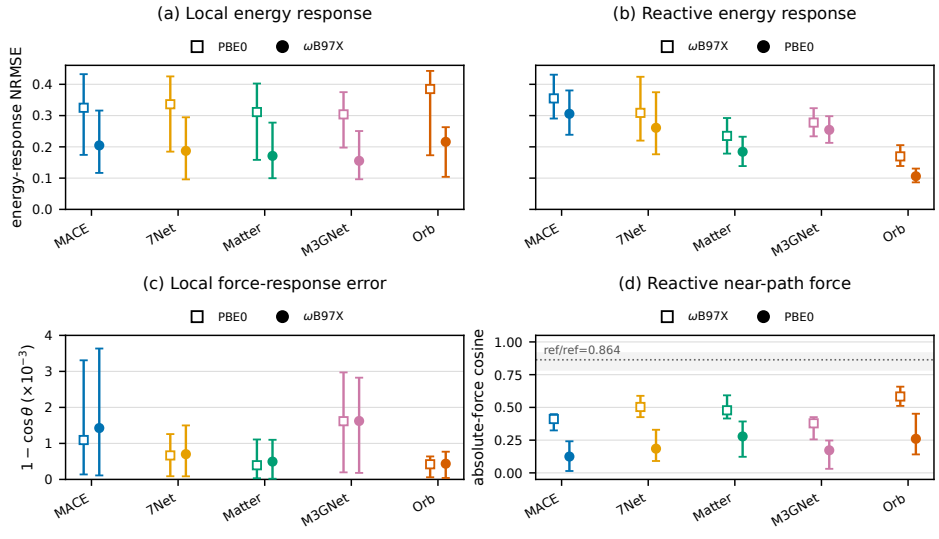

Figure 2: Reference changes do not erase the local–reactive contrast. Each checkpoint is evaluated on identical geometries under two references; bars are 95% molecule- or reaction-clustered intervals. Local panels use energy and force responses from each molecule base. Panel (c) shows the directional error $1-\cos\theta$ in units of 10^{-3}, where θ is the angle between model and reference force changes; lower is better. Median errors span $(0.39\text{–}1.62)\times10^{-3}$. Reactive energy retains the six fixed nonreactant roles; reactive force cosines use the role-stratified 128 near-path configurations. The dotted line and band in (d) give direct PBE0–ωB97X force agreement. The domains use different samples and force definitions.

The separate 80-reaction validation yields 15–37 converged paths per model (18.8–46.2%) and conditional barrier MAE 0.697–1.751 eV. The combined cohorts give 19–44 successes per 100 in the combined sample. On the original 20 reactions, FIRE gives 49/100 model–reaction successes versus 28/100 for BFGS, with paired gains of 5–40 percentage points. MatterSim changes from 5/20 to 13/20. Common-convergence subsets contain only 0–7 reactions per model, precluding a general conditional-barrier comparison. The appendix reports denominators and intervals for both optimizers.

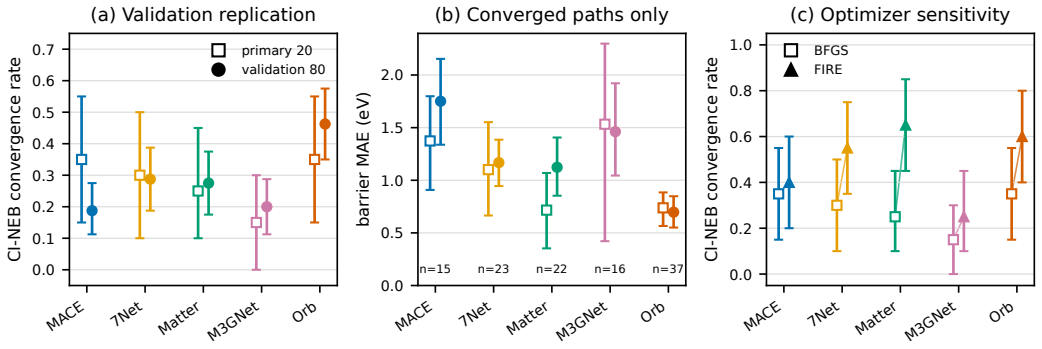

Figure 3: Path completion changes with the optimizer/runtime configuration. (a) The 80-reaction validation cohort is plotted separately from the primary 20; bars are 95% reaction-bootstrap intervals. (b) Barrier MAE is conditional on the annotated validation successes and is not a common-sample ranking. (c) FIRE and BFGS use the same original 20 reactions; connecting lines emphasize paired protocol changes.

Optimizer sensitivity on the validation cohort. FIRE evaluation of all 80 validation reactions per model gives 156/400 successes versus BFGS's 113/400. Gains range from 0.0625 to 0.1500; only MACE's paired test survives five-model Holm adjustment ($p = 0.0377$). Barrier MAE changes by -0.0273 to -0.0008 eV on common-convergence subsets of 9–30 reactions. Appendix J gives every interval. This comparison changes both optimizer and runtime configuration. The following experiment holds the runtime fixed to estimate the optimizer effect.

Optimizer effect within a fixed runtime. The paired optimizer experiment holds hardware, software, checkpoint and precision constant within every BFGS/FIRE pair on the same 20 reactions. FIRE completes 49/100 paths versus 30/100 for BFGS: a paired gain of 19 percentage points [10, 28], resampling reactions jointly across models. Per-model gains are 5–35 points; none of the five exact paired tests survives Holm adjustment. Changing the optimizer therefore changes aggregate completion even with the runtime held fixed. Appendix I reports the paired counts and intervals.

Adaptation and local transfer. Adaptation lowers reactive force error on the development data and local error when the 21 selected models are evaluated on 12 new chemical formulas. Full fine-tuning with replay remains competitive with the proposed residual designs. Appendix K reports all initial and joint training conditions, paired intervals for each dataset.

4 Related Work

Reference-based validation combines numerical errors and physical tasks (Morrow et al., 2023). Broad uMLIP studies examine crystals, relaxation, phonons and minerals (Yu et al., 2024; Deng et al., 2025; Loew et al., 2025; Mannan et al., 2025). Reactive benchmarks cover Transition1x paths (Schreiner et al., 2022), Landscape17 topology (Cărare et al., 2026) and Hessian-supervised search (Cui et al., 2026). Marks et al. (2026) benchmark six available potentials with freezing-string and CI-NEB searches across 58 reactions, including hybrid workflows with DFT refinement. Their high success with molecular-trained potentials establishes the value of matching model and workflow to the chemical task. Our fixed-endpoint tests separate force prediction from optimization behavior;

the matched reference and force-support comparisons identify which prediction differences persist before path search.

5 Discussion and Limitations

Accurate local force-response directions coexist with compressed barriers and limited CI-NEB completion. Reference swaps preserve the original measured contrast, while task-nearer heads improve reactive predictions. Under matched heads, references and common force support, reactive component MAE remains higher by 0.057 [0.023, 0.085] eV Å^{-1}; the directional difference of -0.025 $[-0.071, 0.058]$ is uncertain. The fixed-runtime experiment isolates a 19-point [10, 28] aggregate completion gain from changing BFGS to FIRE. Common-convergence barrier gains remain small. Adaptation lowers reactive development error and local error on new formulas; full fine-tuning with replay remains competitive with the residual designs.

Reaction-path use therefore requires separate reactive-force and workflow validation. Scramble sensitivity admits an unfitted repulsion control and cannot substitute for either. These estimates apply to the evaluated heads, common-support subset, sampled chemistry and optimizer/runtime combinations; Appendix C gives the full limits.

AI use statement

Generative AI tools were used for experiment planning, code generation and debugging, figure generation, literature discovery, and drafting and editing the manuscript. They were not used as experimental subjects or as substitutes for electronic-structure reference calculations. AI-assisted code and numerical claims were checked with deterministic unit tests, completeness checks, fixed-seed reruns, independent basis-set calculations, and SHA-256 provenance manifests; bibliographic metadata were checked against primary publication or official model records. The authors reviewed the AI-assisted artifacts and take responsibility for the final content.

Reproducibility statement

The anonymous supplement provides the manuscript, scientific summaries and row-level inputs for four CPU replays: fixed-head support and influence, reaction-level associations, static energy-error endpoints, and paired optimizer statistics. Appendix H maps the inputs and commands. Sampling, quantum-chemistry settings and analysis definitions appear in Appendices A and F.

References

Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics*, 110:6158–6170, 1999. doi: 10.1063/1.478522.

Ilyes Batatia, Dávid P. Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436, 2022.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, et al. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18): 184110, 2025a. doi: 10.1063/5.0297006.

Ilyes Batatia, Chen Lin, Joseph Hart, Elliott Kasoar, Alin M. Elena, Sam Walton Norwood, Thomas Wolf, and Gábor Csányi. Cross learning between electronic structure theories for unifying molecular, surface, and inorganic crystal foundation force fields, 2025b. URL `https://arxiv.org/abs/2510.25380`.

Vlad Cărare, Fabian L. Thiemann, Joe D. Morrow, David J. Wales, Edward O. Pyzer-Knapp, and Luke Dicks. Global properties of the energy landscape: A testing and training arena for machine learned potentials. *npj Computational Materials*, 12:9, 2026. doi: 10.1038/s41524-025-01878-x.

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

Taoyong Cui, Yunhong Han, Haojun Jia, Chenru Duan, and Qiyuan Zhao. A large scale molecular hessian database for optimizing reactive machine learning interatomic potentials. *Scientific Data*, 13:37, 2026. doi: 10.1038/s41597-025-06350-5.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, Gerbrand Ceder, et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5:1031–1041, 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*, 11:9, 2025. doi: 10.1038/s41524-024-01500-6.

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*, 17:3432, 2026. doi: 10.1038/s41467-026-70195-8. URL https://www.nature.com/articles/s41467-026-70195-8.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben L. Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schøtt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017. doi: 10.1088/1361-648X/aa680e.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The open molecules 2025 (OMol25) dataset, evaluations, and models, 2025. URL https://arxiv.org/abs/2505.08762.

Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials*, 11:178, 2025. doi: 10.1038/s41524-025-01650-1.

Sajid Mannan, Carmelo Gonzales, Kin Long Kelvin Lee, Vaibhav Bihani, Nitya Nand Gosvami, Santiago Miret, and N. M. Anoop Krishnan. Evaluating universal interatomic potentials for molecular dynamics of real-world minerals. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=me0flBb1hi.

Jonah Marks, Jonathon Vandezande, and Joseph Gomes. Reliable and efficient automated transition-state searches with machine-learned interatomic potentials. *arXiv preprint arXiv:2604.00405*, 2026. doi: 10.48550/arXiv.2604.00405. URL https://arxiv.org/abs/2604.00405.

Joe D. Morrow, John L. A. Gardner, and Volker L. Deringer. How to validate machine-learned interatomic potentials. *The Journal of Chemical Physics*, 158(12):121501, 2023. doi: 10.1063/5.0139611.

Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 20(11):4857–4868, 2024. doi: 10.1021/acs.jctc.4c00190.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865.

Krishnan Raghavachari, Gary W. Trucks, John A. Pople, and Martin Head-Gordon. A fifth-order perturbation comparison of electron correlation theories. *Chemical Physics Letters*, 157(6):479–483, 1989. doi: 10.1016/S0009-2614(89)87395-6.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: Atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025. doi: 10.48550/arXiv.2504.06231.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x—a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9:779, 2022. doi: 10.1038/s41597-022-01870-w.

Qiming Sun, Xing Zhang, Samragni Banerjee, Peng Bao, Marc Barbry, Nick S. Blunt, Nikolay A. Bogdanov, George H. Booth, Jia Chen, Zhi-Hao Cui, et al. Recent developments in the PySCF program package. *The Journal of Chemical Physics*, 153(2):024109, 2020. doi: 10.1063/5.0006074.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005. doi: 10.1039/B508541A.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. UMA: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025. doi: 10.48550/arXiv.2506.23971.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. MatterSim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024. doi: 10.48550/arXiv.2405.04967.

Haochen Yu, Matteo Giantomassi, Giuliana Materzanini, Junjie Wang, and Gian-Marco Rignanese. Systematic assessment of various universal machine-learning interatomic potentials. *Materials Genome Engineering Advances*, 2(3):1–12, 2024. doi: 10.1002/mgea.58.

A Full Protocol Details

A.1 Sample and configuration accounting

The 16 ASE molecule identifiers are `H2`, `N2`, `O2`, `CO`, `CO2`, `H2O`, `NH3`, `CH4`, `C2H2`, `C2H4`, `C2H6`, `C6H6`, `CH3OH`, `CH3CH2OH`, `H2CO`, and `N2H4`. Each has one base and 2 directions $\times$ 3 amplitudes $\times$ 2 signs, giving 13 reference configurations and 16$\times$13=208 total. Three rotations per base give 48 more configurations and 256 MLIP evaluations per model. The configuration file contains coordinates, atomic numbers, projection rank, seed, achieved RMS displacement, minimum distance, rotation matrix where applicable, and per-structure SHA-256.

The actual RMS differs from its target by less than 10^{-10} Å in every formal configuration. All path endpoints pass the collision guard. Rigid-mode projection is unweighted in Cartesian space; it is designed to remove representation-level nuisance motion, not to generate mass-weighted vibrational normal modes.

A.2 Electronic-structure execution

PySCF calculations used two CPU threads per process and four deterministic shards. Restricted Kohn–Sham PBE was used for all closed-shell molecules. O_2 used UKS with charge 0 and spin 2. The orbital basis was def2-SVP, density fitting used PySCF's corresponding automatically selected auxiliary basis, the DFT grid level was 3, direct SCF convergence tolerance was 10^{-9} Hartree, and the maximum cycle count was 100. No geometry optimization, dispersion correction, frequency calculation, or per-molecule method selection was performed. All calculations converged without a Newton retry.

A.3 Metric conventions

Energy differences are always method-internal and divided by atom count. The force response is not divided by atom count; its cosine and norm use all $3N$ components. The force-component MAE is first computed over one configuration's $3N$ components and then summarized by the median across configurations. Curvature uses energy per atom and therefore has units eV atom^{-1} Å^{-2}. Bootstrap resampling duplicates or omits entire molecule clusters, including both directions, all amplitudes, and both signs.

Table 2: Primary measures relative to PBE/def2-SVP, fixed in advance. Brackets are 95% molecule-clustered bootstrap intervals. Force cosine and norm ratio are medians over configuration-level values. r_E is Pearson correlation; β_κ is the curvature slope. Curvature-sign agreement is 100% (96/96) for every model.

Endpoint	MACE	SevenNet	MatterSim
β_E	.964 [.900,.992]	1.036 [.955,1.057]	1.004 [.936,1.031]
r_E	.990	.993	.994
Energy NRMSE	.125 [.074,.223]	.114 [.065,.200]	.100 [.048,.192]
Force cosine	.9994 [.9978,1.0000]	.9994 [.9989,1.0000]	.9998 [.9993,1.0000]
Force norm	.929 [.893,.949]	.945 [.909,.982]	.943 [.930,.966]
β_κ	.935 [.904,.967]	1.034 [.969,1.062]	.995 [.941,1.020]

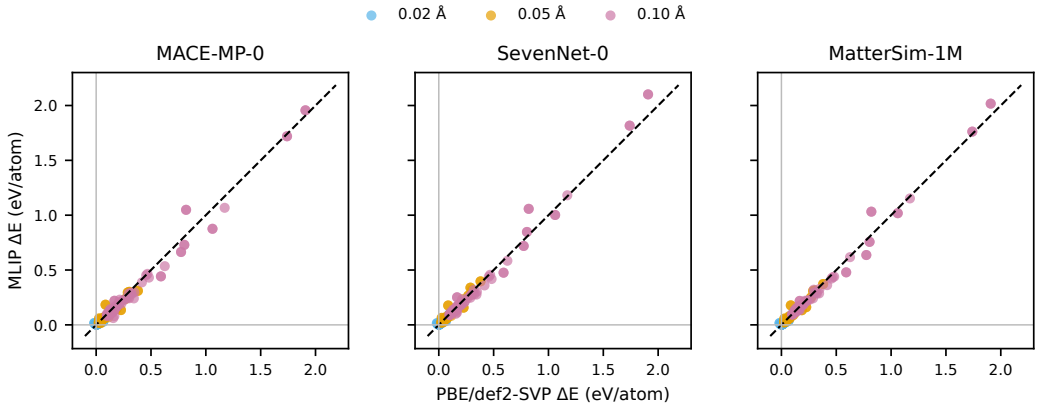

Figure 4: Local energy responses relative to PBE/def2-SVP. Each point is one signed internal displacement; colors denote RMS amplitude. Energies are differences from the same molecule's undisplaced structure within each method and are divided by atom count. The dashed line is equality. Pooled summaries use all amplitudes and cluster uncertainty by molecule.

A.4 Independent-reference extension

The PBE0 repeat uses def2-TZVP, density fitting, grid level 3, SCF tolerance 10^{-9} Hartree, at most 100 iterations, and one prespecified Newton retry. O_2 is UKS with spin 2 and recorded $\langle S^2 \rangle$; all other systems are RKS. The 72-row CC subset contains the base plus both signs at 0.05 and 0.10 Å for two directions of `H2`, `N2`, `CO`, `H2O`, `NH3`, `CH4`, `CO2`, and `H2CO`. RHF uses tolerance 10^{-10} Hartree; RCCSD uses energy tolerance 10^{-8} Hartree and amplitude-norm tolerance 10^{-6}; the perturbative triples correction follows the converged RCCSD calculation. The lowest spatial orbital of each C, N, or O atom is excluded from electron correlation. We report the full set and a sensitivity subset with $T_1 \leq 0.02$.

Table 3: Independent-reference molecular endpoints. Brackets are 95% molecule-clustered bootstrap intervals. PBE0 uses 192 displaced configurations from 16 molecules; CCSD(T) uses 64 from eight closed-shell molecules and provides no force endpoint.

Model	PBE0 β_E [95% CI]	PBE0 NRMSE	PBE0 force cosine	CC β_E [95% CI]	CC NRMSE
MACE MP-0	1.010 [0.944, 1.029]	0.325	0.999	1.042 [0.954, 1.067]	0.147
SevenNet-0	1.091 [0.991, 1.116]	0.336	0.999	1.129 [1.021, 1.166]	0.185
MatterSim-1M	1.058 [0.982, 1.082]	0.311	1.000	1.090 [1.005, 1.132]	0.149
M3GNet MatPES	1.091 [0.999, 1.119]	0.304	0.998	1.125 [1.073, 1.144]	0.159
Orb-v3 OMat	1.122 [1.024, 1.160]	0.385	1.000	1.169 [1.052, 1.230]	0.253

A.5 Reference-comparison details

R1a repeats every local configuration at ωB97X/6-31G(d); R1b applies PBE0/def2-TZVP to seven fixed roles from each of 32 reactions; and R1c applies RHF-CCSD(T)/cc-pVTZ to R/TS/P for 12 reactions with C/N/O $1s$ correlation excluded. All calculations use CPU PySCF 2.14.0. Local force metrics are base-relative responses, while reactive force metrics are absolute vectors. The pre-specified reactive summary pooled all seven roles. The exploratory geometry-role analysis reports the pooled estimate alongside the four near-path roles and three stationary-point roles; no force magnitude or model result enters the split. Table 4 uses near-path force values for the model com-

parison, and Table 5 retains all three reference-comparison summaries. The supplement provides configuration identities, reference outputs and checksums.

Table 4: Crossed reference bridge for the five compatible checkpoints. Local energy and force-response metrics use 192 displaced configurations. Reactive energy responses retain the six non-reactant roles; reactive force metrics use the role-stratified set of 128 near-path configurations and exclude R/TS/P stationary endpoints. Brackets are 95% molecule- or reaction-clustered intervals. The four cells use domain-specific force measures.

Geometry	Model	Reference	energy NRMSE [95% CI]	force cosine [95% CI]	force norm [95% CI]
Local	MACE MP-0	PBE0	0.325 [0.174, 0.432]	0.999 [0.997, 1.000]	0.964 [0.929, 0.989]
Local	MACE MP-0	ωB97X	0.205 [0.117, 0.316]	0.999 [0.996, 1.000]	0.918 [0.869, 0.939]
Local	SevenNet-0	PBE0	0.336 [0.185, 0.425]	0.999 [0.999, 1.000]	0.968 [0.936, 1.034]
Local	SevenNet-0	ωB97X	0.187 [0.096, 0.295]	0.999 [0.999, 1.000]	0.924 [0.890, 0.964]
Local	MatterSim-1M	PBE0	0.311 [0.158, 0.403]	1.000 [0.999, 1.000]	0.977 [0.970, 1.014]
Local	MatterSim-1M	ωB97X	0.171 [0.100, 0.277]	1.000 [0.999, 1.000]	0.929 [0.918, 0.952]
Local	M3GNet MatPES	PBE0	0.304 [0.198, 0.375]	0.998 [0.997, 1.000]	1.016 [0.993, 1.064]
Local	M3GNet MatPES	ωB97X	0.155 [0.096, 0.251]	0.998 [0.997, 1.000]	0.964 [0.947, 0.984]
Local	Orb-v3 OMat	PBE0	0.385 [0.173, 0.443]	1.000 [0.999, 1.000]	1.005 [0.985, 1.067]
Local	Orb-v3 OMat	ωB97X	0.216 [0.104, 0.263]	1.000 [0.999, 1.000]	0.955 [0.934, 1.009]
Reactive	MACE MP-0	ωB97X	0.355 [0.291, 0.431]	0.411 [0.324, 0.448]	1.892 [1.516, 2.365]
Reactive	MACE MP-0	PBE0	0.306 [0.239, 0.380]	0.125 [0.014, 0.241]	1.547 [1.347, 1.803]
Reactive	SevenNet-0	ωB97X	0.309 [0.220, 0.424]	0.503 [0.426, 0.588]	1.676 [1.517, 2.012]
Reactive	SevenNet-0	PBE0	0.261 [0.176, 0.375]	0.185 [0.091, 0.329]	1.561 [1.368, 1.657]
Reactive	MatterSim-1M	ωB97X	0.235 [0.178, 0.292]	0.478 [0.415, 0.592]	1.776 [1.485, 2.185]
Reactive	MatterSim-1M	PBE0	0.184 [0.139, 0.232]	0.278 [0.123, 0.393]	1.500 [1.371, 1.710]
Reactive	M3GNet MatPES	ωB97X	0.278 [0.234, 0.324]	0.379 [0.256, 0.426]	2.701 [2.250, 3.176]
Reactive	M3GNet MatPES	PBE0	0.254 [0.213, 0.298]	0.173 [0.031, 0.247]	2.347 [2.042, 2.712]
Reactive	Orb-v3 OMat	ωB97X	0.170 [0.139, 0.205]	0.583 [0.512, 0.658]	1.475 [1.275, 1.732]
Reactive	Orb-v3 OMat	PBE0	0.106 [0.086, 0.131]	0.260 [0.141, 0.452]	1.264 [1.138, 1.422]

Table 5: Reference comparisons on identical geometries. The upper block compares CPU PySCF references on identical geometries; N_E/N_F gives energy-response and force-vector counts. The local comparison is ωB97X versus PBE0; reactive comparisons reverse that order and report all geometries, near-path configurations and stationary endpoints separately. The lower block reports errors for Transition1x minus CCSD(T)/cc-pVTZ on the fixed reaction subset, with C/N/O $1s$ orbitals excluded from correlation; brackets are 95% cluster-bootstrap intervals.

Geometry	N_E/N_F	energy slope	energy NRMSE	force cosine	force norm
Local	192/192	1.099 [1.057, 1.109]	0.165 [0.095, 0.180]	1.000 [1.000, 1.000]	1.054 [1.048, 1.075]
Reactive: all roles	192/224	0.932 [0.914, 0.948]	0.085 [0.071, 0.101]	0.423 [0.337, 0.543]	4.353 [2.286, 5.437]
Reactive: near path	–/128	–	–	0.864 [0.787, 0.914]	1.035 [0.997, 1.131]
Reactive: R/TS/P	–/96	–	–	0.088 [0.012, 0.132]	14.060 [13.189, 16.910]

CCSD(T) comparison	N reactions	bias (eV) [95% CI]	MAE (eV) [95% CI]	RMSE (eV) [95% CI]
Forward barrier	12	0.197 [0.131, 0.259]	0.206 [0.151, 0.261]	0.228 [0.172, 0.278]
Reverse barrier	12	0.031 [-0.026, 0.086]	0.091 [0.062, 0.120]	0.105 [0.074, 0.130]
Reaction energy	12	0.166 [0.084, 0.253]	0.178 [0.105, 0.258]	0.224 [0.140, 0.294]

Maximum reaction-level T_1: 0.024; median: 0.017; reaction maxima above 0.02: 4.

A.6 Reaction registry and CI-NEB details

The exact Transition1x source file is checksum-locked, and the official loader revision is stored with the registry. Candidate reactions are stratified jointly by C/N/O heavy-atom count and reference forward-barrier quintile. Each stratum is seeded with eight reactions under the ordering `sha256(Transition1x-v4|2027|formula|reaction-id)`, then deterministically filled to 200. Near-path configurations target reference-energy ranks 0.20, 0.50, and 0.80 and reference-force-RMS rank 0.90 after removing endpoint duplicates. No model output enters selection.

The 20 CI-NEB reactions are the first fixed reaction hashes, again without model information. All models use ten images including endpoints, IDPP interpolation, ASE's improved-tangent method, spring 0.1 eV/Å^2, and BFGS with $\alpha = 70$ and maximum step 0.03 Å. Ordinary NEB has threshold 0.5 eV/Å and at most 500 steps; climbing- image NEB has threshold 0.05 eV/Å and another 500. We retain initialization diagnostics and all failure states. A final barrier or TS-geometry error is undefined unless the final threshold is reached.

The validation cohort takes the next 80 reactions in the already fixed hash order, with zero overlap with the primary 20 and no model-dependent selection. It uses the same BFGS settings. The optimizer-sensitivity experiment instead retains the primary 20 reactions and changes only BFGS to ASE FIRE 3.29.0 with prespecified FIRE parameters. It reruns every reaction, including BFGS successes, so convergence differences are paired over 20 reactions per model. Conditional barrier comparisons use only explicitly reported common-success intersections.

Table 6: Complete Transition1x static reaction endpoints. Brackets are 95% reaction-clustered bootstrap intervals. Barrier MAE is in eV/reaction; near-path force metrics use 800 configurations per model.

Model	N	barrier MAE [95% CI]	β_{barrier} [95% CI]	TS ordered	path NRMSE	force cosine [95% CI]	force norm
MACE MP-0	200	1.295 [1.204, 1.390]	0.660 [0.638, 0.681]	92.0%	0.347	0.358 [0.327, 0.401]	1.846
SevenNet-0	200	0.993 [0.909, 1.081]	0.742 [0.721, 0.763]	96.5%	0.299	0.442 [0.404, 0.472]	1.757
MatterSim-1M	200	0.764 [0.694, 0.838]	0.820 [0.799, 0.841]	99.5%	0.229	0.457 [0.433, 0.493]	1.722
M3GNet MatPES	200	0.927 [0.814, 1.045]	0.782 [0.751, 0.812]	87.0%	0.396	0.311 [0.269, 0.339]	2.617
Orb-v3 OMat	200	0.552 [0.505, 0.599]	0.857 [0.845, 0.869]	99.5%	0.159	0.605 [0.558, 0.641]	1.431

Table 7: Fixed-endpoint 20-reaction CI-NEB protocol outcomes. Common Transition1x R/P geometries are held fixed for every checkpoint. Convergence intervals use all selected reactions; barrier MAE and TS RMSD use converged paths only. The static-error convergence AUC is an exploratory endpoint (0.5 is chance orientation).

Model	converged	rate [95% CI]	barrier MAE [95% CI]	TS RMSD (Å)	static→convergence AUC [95% CI]
MACE MP-0	7/20	0.350 [0.150, 0.550]	1.374 [0.908, 1.798]	0.260	0.648 [0.357, 0.896]
SevenNet-0	6/20	0.300 [0.100, 0.500]	1.101 [0.665, 1.552]	0.301	0.690 [0.417, 0.920]
MatterSim-1M	5/20	0.250 [0.100, 0.450]	0.716 [0.353, 1.069]	0.191	0.720 [0.453, 0.941]
M3GNet MatPES	3/20	0.150 [0.000, 0.300]	1.533 [0.421, 2.298]	0.351	0.686 [0.111, 1.000]
Orb-v3 OMat	7/20	0.350 [0.150, 0.550]	0.738 [0.566, 0.885]	0.131	0.484 [0.208, 0.762]

Table 8: CI-NEB results for the primary 20-reaction cohort and the separate 80-reaction follow-up cohort. Brackets are 95% reaction-bootstrap intervals. Barrier MAE is in eV and is conditional on final convergence, with its effective n shown explicitly.

Model	Primary convergence	Validation convergence	Validation MAE	Combined convergence	Combined MAE
MACE MP-0	7/20: 35.0% [15.0, 55.0]	15/80: 18.8% [11.2, 27.5]	1.751 [1.338, 2.153] (n=15)	22/100: 22.0% [14.0, 30.0]	1.631 [1.308, 1.940] (n=22)
SevenNet-0	6/20: 30.0% [10.0, 50.0]	23/80: 28.7% [18.8, 38.8]	1.168 [0.945, 1.386] (n=23)	29/100: 29.0% [20.0, 38.0]	1.154 [0.964, 1.345] (n=29)
MatterSim-1M	5/20: 25.0% [10.0, 45.0]	22/80: 27.5% [17.5, 37.5]	1.124 [0.853, 1.406] (n=22)	27/100: 27.0% [19.0, 36.0]	1.048 [0.811, 1.300] (n=27)
M3GNet MatPES	3/20: 15.0% [0.0, 30.0]	16/80: 20.0% [11.2, 28.7]	1.462 [1.045, 1.921] (n=16)	19/100: 19.0% [12.0, 27.0]	1.473 [1.098, 1.885] (n=19)
Orb-v3 OMat	7/20: 35.0% [15.0, 55.0]	37/80: 46.2% [35.0, 57.5]	0.697 [0.551, 0.848] (n=37)	44/100: 44.0% [34.0, 54.0]	0.703 [0.578, 0.837] (n=44)

Table 9: Exploratory optimizer comparison on the same 20 reactions. The change is FIRE minus BFGS in percentage points with a paired reaction-bootstrap interval. Model-specific comparisons are exploratory and are not multiplicity-adjusted. FIRE barrier MAE uses its converged subset, with the count shown.

Model	BFGS	FIRE	Δ convergence (pp)	common n	FIRE MAE
MACE MP-0	7/20	8/20	+5.0 [-10.0, +20.0]	6	1.577 (n=8)
SevenNet-0	6/20	11/20	+25.0 [+10.0, +45.0]	6	1.089 (n=11)
MatterSim-1M	5/20	13/20	+40.0 [+20.0, +60.0]	5	1.080 (n=13)
M3GNet MatPES	3/20	5/20	+10.0 [-15.0, +35.0]	0	1.349 (n=5)
Orb-v3 OMat	7/20	12/20	+25.0 [+5.0, +45.0]	7	0.884 (n=12)

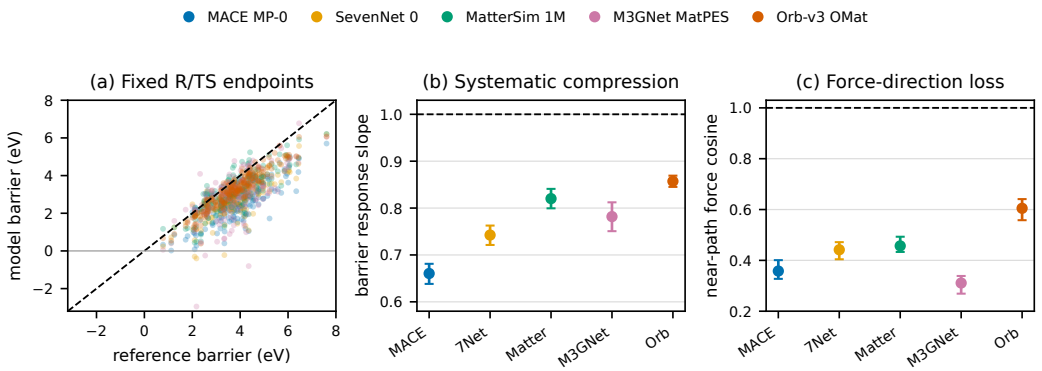

Figure 5: Complete static evaluation reveals reactive-domain degradation. (a) Model versus Transition1x forward barriers for 200 reactions; the dashed line is equality. (b) Through-origin barrier slopes and 95% reaction-clustered intervals are below one for all five models. (c) Median near-path force cosines and clustered intervals are far below the near-equilibrium molecular values.

B Sensitivity to displacement amplitude and basis set

This section reports exploratory amplitude and basis-set comparisons. Table 10 gives amplitude-specific measures. Energy NRMSE falls with amplitude as the signal grows, whereas force-direction and force-scale metrics remain stable.

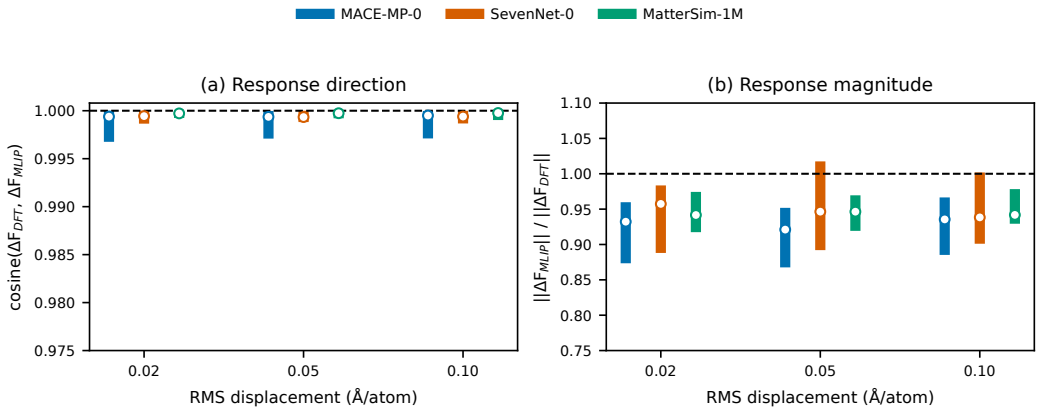

Figure 6: Force-response agreement by displacement amplitude. White circles are medians over 64 signed configurations and thick bars are interquartile ranges. Full-vector directions nearly coincide with PBE/def2-SVP, while norm ratios are mostly below one for this reference. Axes are magnified; dashed lines mark exact agreement.

Table 10: Amplitude slices against PBE/def2-SVP. Each row contains 64 signed configurations. Values are point estimates.

Model	RMS (Å)	β_E	NRMSE	force cosine	force norm
MACE	.02	.753	.376	.99938	.932
	.05	.893	.221	.99936	.921
	.10	.968	.117	.99950	.935
SevenNet	.02	.863	.323	.99939	.957
	.05	1.016	.181	.99934	.946
	.10	1.037	.109	.99941	.938
MatterSim	.02	.831	.335	.99972	.942
	.05	.962	.176	.99974	.946
	.10	1.006	.093	.99976	.942

Leaving out one molecule at a time gives pooled slope ranges 0.953–0.969 (MACE), 1.017–1.041 (SevenNet), and 0.985–1.013 (MatterSim). Excluding O_2 gives slopes 0.953, 1.030, and 0.996; restricting to non-diatomics gives 0.898, 0.945, and 0.953. Response slopes vary across molecule subsets.

B.1 Basis-set repeat

All 208 PBE/def2-TZVP calculations converge. Table 11 treats def2-SVP as x and def2-TZVP as y on the same paths. The nearly unit directional correlations coexist with a systematic response-scale change.

Recomputing the MLIP summaries with def2-TZVP as the reference yields the values in Table 12. The change is large enough that a simple "soft versus hard" interpretation reverses for all three models.

Table 11: PBE basis comparison on the same 192 displaced configurations and 96 curvatures. Brackets are molecule-clustered 95% bootstrap intervals.

Endpoint	TZVP relative to SVP	95% interval
Energy slope	0.9165	[0.8999, 0.9505]
Energy NRMSE	0.1070	[0.0649, 0.1276]
Force cosine (median)	0.99987	[0.99979, 0.99999]
Force norm ratio (median)	0.9422	[0.9373, 0.9507]
Curvature slope	0.9312	[0.9161, 0.9604]
Curvature sign agreement	100%	[100%, 100%]

Table 12: Model measures using PBE/def2-TZVP as reference.

Model	β_E	NRMSE	force cosine	force norm	β_κ
MACE	1.045	.168	.999	.978	1.004
SevenNet	1.123	.195	1.000	.988	1.108
MatterSim	1.089	.161	1.000	.999	1.068

C Scope and Limitations of the Extended Comparisons

Task-nearer MACE-MH-1 and SevenNet-Omni checkpoints improve static reactive force direction and barrier accuracy and finish more of the fixed-endpoint paths, although 40–45% of those paths remain unconverged. MatterSim's 1M-to-5M change is smaller. These descriptive comparisons identify checkpoint/head dependence. Holding the head fixed retains higher reactive component MAE on common support, while the cosine-difference interval spans zero. The fixed-endpoint optimizer comparison separately identifies workflow dependence.

PBE0 and ωB97X near-path forces align substantially better with each other than any original broad-domain checkpoint aligns with either reference. That comparison does not extend to every task-nearer head or standardized force measure. The geometry-role split separates reference-specific stationarity from near-path force agreement. The domains differ in samples and force measures, and pretraining membership is unknown.

CI-NEB conclusions condition on fixed Transition1x endpoints rather than model-specific endpoint relaxation. The fixed-runtime comparison gives 30/100 BFGS and 49/100 FIRE completions; the earlier cross-runtime comparison gives 28/100 and 49/100. The paired design estimates an optimizer effect within the evaluated runtimes. Completion means reaching the prescribed NEB force threshold; transition-state certification requires a reference Hessian and reaction-connectivity check. The evidence in this paper covers molecular and reaction systems.

External validity is bounded to 16 neutral H/C/N/O molecules and 200 H/C/N/O reactions with 3–7 heavy atoms; saved intermediates are a near-path cloud, not an IRC. We do not test charged or open-shell chemistry beyond O_2, long-range electrostatics, surfaces, liquids, or long-time dynamics. Converged-only CI-NEB results remain selection-vulnerable with 3–7 primary and 15–37 validation observations per model. Reference-force standardization does not match chemical identities or geometry roles, and near-stationary strata remain separately visible.

D Sensitivity and Fidelity Are Different Constructs

Let $q = (Z, R)$ denote atomic identities and Cartesian coordinates, and let $E_\theta(q)$ be a model energy. For a base configuration q_0, a perturbation distribution $\mathcal{P}$ can define a sensitivity statistic such as

$$R_{\mathcal{P}} = \frac{\text{median}_{q\sim\mathcal{P}_{\text{scramble}}} |E_\theta(q) - E_\theta(q_0)|/N}{\text{median}_{q\sim\mathcal{P}_{\text{rattle}}} |E_\theta(q) - E_\theta(q_0)|/N}. \tag{2}$$

Large $R_{\mathcal{P}}$ establishes dependence on something changed by the scramble. It cannot identify what feature caused the change or whether the magnitude and direction are correct. In contrast, a reference response compares $E_\theta(q) - E_\theta(q_0)$ with $E_{\text{ref}}(q) - E_{\text{ref}}(q_0)$ on the same path. The energy zero cancels within each method, so models with different atomic reference energies remain comparable.

D.1 Construct-validity test of the scramble statistic

The test contains 23 bulk crystals perturbed by shuffling species over fixed sites and 23 coordinate-scrambled structures (7 surfaces and 16 molecules). It compares five scrambles with three 0.05 Å Gaussian rattles and uses the rule $R_{\mathcal{P}} > 10$ plus an absolute scramble response above 10^{-3} eV atom^{-1}. We apply both thresholds structure by structure as a descriptive coverage diagnostic. The test measures sensitivity to structural changes.

Coordinate scrambling preserved coordinate pools but not local geometry. Across its 23 structures, the median Wasserstein-1 change in the pair-distance distribution within 6 Å was 0.200 Å, and the median absolute coordination change was 1.00 neighbor per atom. Species shuffling kept positions and composition fixed, but its median total-variation change in neighbor element-pair frequencies was 0.094. Short distances, coordination and local chemistry are direct inputs to short-ranged potentials.

We therefore evaluated an intentionally incomplete control,

$$E_{\text{pair}}(q) = (1\,\text{eV}) \sum_{\substack{i<j \\ r_{ij} \le 6\,\text{Å}}} \left(\frac{r_{\text{cov},i} + r_{\text{cov},j}}{r_{ij}} \right)^{12}, \tag{3}$$

with no fitting, angular term, electronic state, or many-body interaction. The 46-structure test uses the 6 Å neighbor cutoff and divides this energy by N; the local molecular repeat evaluates all pairs, whose maximum separation is 5.091 Å here, so the definitions coincide on the molecular sample. The control passes the same rule on 33/46 structures and both designated sanity structures. The three MLIPs pass on 41/46 (MACE), 42/46 (SevenNet), and 43/46 (MatterSim), not on every structure. The fixed 1 eV coefficient is arbitrary; this is itself a warning against an absolute energy threshold without a physical reference.

Figure 7 supplies a constructive counterexample: the same observable can be large for a model that is manifestly not a faithful quantum PES. The main evaluation scores small, symmetric paths against explicit reference responses.

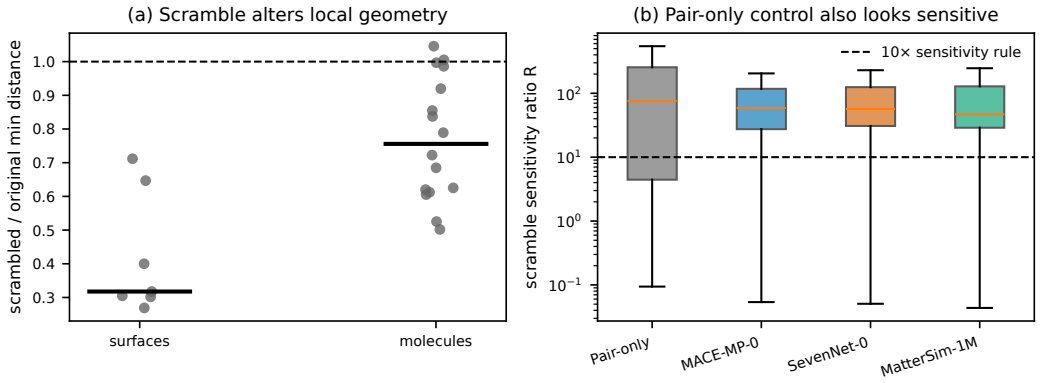

Figure 7: The scramble sensitivity statistic is non-identifying. (a) Coordinate scrambles frequently reduce the minimum pair distance; dots are structures and bars are family medians. (b) The unfitted pair-only control attains sensitivity ratios comparable in scale to the three MLIPs. Boxes show the interquartile range over all 46 structures, orange lines the median, and whiskers the range; the dashed line is the $R > 10$ component of the rule. High sensitivity alone therefore cannot establish reference fidelity or many-body physics.

E Symmetry and Exploratory Association Checks

Rigid rotations measure numerical equivariance. Table 13 reports the largest residual over the fixed rotations for each model.

Table 13: Rigid-rotation implementation checks over 48 rotations per model. Force error is the maximum absolute Cartesian-component residual after rotating the base force.

Model	max $\lvert\Delta E\rvert/N$ (eV)	max force error (eV/Å)
MACE	5.00×10^{-13}	9.56×10^{-11}
SevenNet	9.54×10^{-7}	1.22×10^{-5}
MatterSim	2.38×10^{-6}	1.67×10^{-4}

F Analysis design and sample coverage

The molecular and reference-extension studies specified their samples, calculations, metrics and exclusion rules before generating the respective outputs. The original CI-NEB analysis uses BFGS on 20 reactions; the 80-reaction follow-up uses the same settings and a disjoint sample. The pair control, basis and amplitude checks, geometry-role split, checkpoint and common-support comparisons, error associations, and FIRE comparisons are exploratory. Adaptation uses development data. The first 21 selected adaptations are evaluated on 12 new local formulas; joint training reuses that panel for exploratory evaluation.

The molecular analysis requires at least 15 valid molecule bases, 95% reference convergence and predictions for every valid configuration. It includes 16/16 bases and 208/208 references per model. Static reaction analysis requires 200 reactions and at least 99% prediction coverage; all five models cover 1,400/1,400 configurations. CI-NEB reports every assigned path, including nonconvergence: 20 per model in the primary cohort and 80 per model in validation. PBE0 and CCSD(T) require 95% reference convergence over at least 15/16 and 8/8 molecules, respectively; both achieve 100%. The reference comparisons contain 208/208 local, 224/224 reactive and 36/36 stationary-point calculations. The supplement records configurations, checkpoint identities, analysis settings and full checksums.

Table 14: Exploratory checkpoint comparison. Static metrics use all 200 reactions; local force cosines use the molecular sample and domain-specific heads (SevenNet: local qcml, reactive omol25_low). They are response versus absolute-force measures; Appendix G supplies the common-head, absolute-force comparison. CI-NEB MAE is conditional on the displayed converged denominator. MatterSim-5M was a capacity control and had no C1 CI-NEB block.

Checkpoint/head	Local cos.	Barrier MAE	Reactive cos.	CI-NEB	Cond. MAE
MACE-MH-1 (omol)	1.000	0.268	0.806	11/20	0.307
SevenNet-Omni (omol25-low)	1.000	0.275	0.827	12/20	0.332
MatterSim-5M	1.000	0.606	0.506	—	—

Table 15: Exploratory Spearman association between the molecule scramble-sensitivity ratio and molecule-level response fidelity ($n = 16$). Higher NRMSE is worse; higher force cosine is better. Values are correlation estimates.

Model	R vs energy NRMSE	R vs force cosine
MACE	0.244	−0.233
SevenNet	0.238	−0.059
MatterSim	0.321	−0.224

G Absolute-force and matched-head comparisons

The original local endpoint is a displaced-minus-base force response, whereas the reactive endpoint is an absolute force. This exploratory comparison uses absolute forces in both domains. We calculate one full-vector cosine and one component MAE per configuration, average within molecule or reaction, then equally weight clusters. Cosine is dimensionless; MAE is in eV Å^{-1}. Norms below

10^{-8} eV Å^{-1} make directional measures undefined; MAE remains defined. A separate reference-component-RMS stratum below 0.05 eV Å^{-1} identifies near-stationary forces. All aggregate and stratified counts remain in the accompanying scientific JSON.

Table 16: Paired task-nearer minus original-checkpoint differences on the same configurations and ωB97X reference. Local has 208 configurations/16 molecules; reactive has 1,400 configurations/200 reactions. SevenNet-Omni uses qcml locally and omol25_low reactively in this first comparison. Intervals are marginal 95% cluster bootstrap intervals.

Checkpoint	Domain	Δ cosine [95% CI]	Δ MAE [95% CI]
MACE-MH-1	Local	0.0460 [-0.0241, 0.1093]	-0.1422 [-0.2610, -0.0438]
MACE-MH-1	Reactive	0.1822 [0.1635, 0.2009]	-0.2447 [-0.2596, -0.2301]
SevenNet-Omni	Local	0.0741 [0.0225, 0.1261]	-0.1189 [-0.2198, -0.0170]
SevenNet-Omni	Reactive	0.1562 [0.1379, 0.1744]	-0.2198 [-0.2310, -0.2091]
MatterSim-5M	Local	-0.0264 [-0.0540, -0.0034]	0.0092 [-0.0104, 0.0264]
MatterSim-5M	Reactive	0.0120 [0.0051, 0.0188]	-0.0173 [-0.0238, -0.0112]

The force bins, fixed before this reanalysis, have reference component-RMS edges $[0, 0.05, 0.1, 0.25, 0.5, 1, 2, \infty)$ eV Å^{-1}. A bin must contain at least ten configurations and two clusters in each domain. Eligible bins are $[0.25, 0.5)$, $[0.5, 1)$ and $[1, 2)$, with local counts 12, 35, 51 and reactive counts 153, 258, 48. Thus the standardized contrast uses 98/208 local and 459/1,400 reactive configurations. We average the two domains' within-support configuration proportions to obtain fixed weights 0.227891, 0.459617, and 0.312492; clusters receive equal weight within each bin. The 10,000-draw bootstrap independently resamples molecule and reaction clusters, retaining common bin weights. It conditions on those weights and gives marginal intervals for each contrast.

Table 17: Standardized reactive-minus-local absolute-force contrasts. SevenNet-Omni in this table retains domain-specific heads; the next table fixes its head. All differences remain descriptive because chemical identities and geometry roles differ between domains.

Checkpoint	Cosine difference [95% CI]	MAE difference [95% CI]
M3GNet	-0.3504 [-0.4444, -0.2406]	0.4307 [0.3583, 0.5033]
MACE-MH-1	-0.0288 [-0.0767, 0.0552]	0.0592 [0.0217, 0.0913]
MACE-MP-0	-0.3590 [-0.4171, -0.2936]	0.3041 [0.2409, 0.3626]
MatterSim-1M	-0.2367 [-0.2739, -0.1980]	0.2110 [0.1556, 0.2669]
MatterSim-5M	-0.1699 [-0.2238, -0.1101]	0.1800 [0.1222, 0.2353]
Orb-v3	-0.1224 [-0.1795, -0.0324]	0.1306 [0.0717, 0.1854]
SevenNet-0	-0.2173 [-0.3323, -0.0530]	0.2232 [0.1622, 0.2778]
SevenNet-Omni	-0.0620 [-0.0718, -0.0524]	0.0477 [0.0164, 0.0746]

Seven of these eight cosine intervals are negative, with MACE-MH-1 unresolved. SevenNet-Omni uses different heads in this table; fixing its molecular head in the next comparison also leaves its directional contrast unresolved. The other seven checkpoints already use a single head across domains.

For the fixed-head analysis, the SevenNet-Omni checkpoint SHA-256 begins `ca81bd3aac9fc269`; the complete hash is in the scientific summary. The head is `omol25_low` with float32 inference in both domains. Its 208 local predictions are compared against the same local ωB97X references; the 1,400 reactive predictions use the identical checkpoint and head. Both domains reuse the corresponding electronic-structure references. The paired local head contrast (omol25_low minus qcml) is -0.0259 $[-0.0669, -0.0022]$ for cosine and -0.0084 $[-0.0240, 0.0066]$ eV Å^{-1} for component MAE.

Table 18: Fixed-head SevenNet-Omni. Top: unstandardized absolute-force summaries; 717 reactive configurations and two local configurations lie in the near-stationary reference stratum. Bottom: common-support differences, which distinguish direction from magnitude.

Domain	Configs/clusters	Cosine [95% CI]	MAE [95% CI]
Local	208/16	0.9481 [0.9037, 0.9833]	0.2145 [0.1275, 0.3171]
Reactive	1400/200	0.4077 [0.3922, 0.4231]	0.1619 [0.1550, 0.1690]

Metric	Reactive minus local [95% CI]
Cosine	-0.0245 [-0.0714, 0.0585]
Norm ratio	0.0239 [-0.0397, 0.0877]
Component MAE	0.0566 [0.0235, 0.0849]

Standardization reduces the directional contrast and reverses the MAE ordering. The estimates depend on force support and weighting; chemical identities and geometry roles still differ between domains, and pretraining membership is unknown.

G.1 Common-support coverage and cluster influence

Table 19: Complete force-RMS bins. Counts are configurations/independent clusters. The three common bins retain 98 local and 459 reactive configurations.

RMS bin (eV/Angstrom)	Local configs/clusters	Reactive configs/clusters	Common
$[0, .05)$	2/2	717/200	no
$[.05, .1)$	2/2	77/71	no
$[.1, .25)$	8/7	147/121	no
$[.25, .5)$	12/8	153/115	yes
$[.5, 1)$	35/13	258/161	yes
$[1, 2)$	51/16	48/39	yes
$[2, \infty)$	98/16	0/0	no

We remove each of the 16 molecule clusters and 200 reaction clusters in turn, keeping the original common-bin identities. The primary influence analysis fixes the original weights; a secondary analysis recomputes the equal-domain configuration-proportion weights. Every remaining bin must still contain ten configurations and two clusters in each domain. All deletions meet these requirements; no deletion is omitted. Cluster means remain equally weighted within each bin, so prolific configurations do not receive independent-cluster weight.

Table 20: Delete-one-cluster point-estimate ranges. These are sample-influence summaries, not sampling confidence intervals. MAE remains positive in every listed deletion; the original directional confidence interval still includes zero.

Deleted cluster	Weights	Estimable	Cosine range	MAE range (eV/Angstrom)
local	Fixed	16/16	[-0.0635, -0.0208]	[0.0506, 0.0674]
local	Reestimated	16/16	[-0.0638, -0.0202]	[0.0509, 0.0674]
reactive	Fixed	200/200	[-0.0250, -0.0232]	[0.0542, 0.0575]
reactive	Reestimated	200/200	[-0.0252, -0.0229]	[0.0543, 0.0575]

Reestimating weights yields similar ranges. The directional confidence interval includes zero despite negative point estimates in all deletion subsets. The supplement records configurations, bins, cluster identities and weights for each deletion.

H Artifact and Reproduction Map

The anonymous CPU analysis supplement contains the manuscript and its TeX dependencies, scientific summaries, and the following replay inputs:

- `docs/paper-atomistic-next/` contains configuration-level absolute-force records, the common-support protocol and the influence summary. `scripts/revision_at3_support_influence.py` reproduces support point estimates and all 216 cluster deletions. The original bootstrap intervals and additional fixed-head and checkpoint comparisons remain in the summaries under `docs/paper-atomistic-revision/`.
- Four CSVs in `docs/atomistic-extension/analysis/` contain static reaction endpoints, near-path errors, and primary and validation BFGS records. `scripts/replay_atomistic_endpoints.py` reproduces all three static energy-error endpoints and their intervals from 1,000 checkpoint/reaction rows. `scripts/analyze_revision_reaction_consequences.py` reproduces the force-error/barrier-error and path-completion associations. Expected results are in `docs/revision-reaction-consequences-20260923/`.
- `docs/atomistic-optimizer-control-20260924/` contains the protocol, all 100 paired optimizer rows and the scientific summary. `scripts/replay_atomistic_optimizer_control.py` reproduces completion counts, paired bootstrap intervals, exact tests and common-convergence barrier point estimates for the fixed-runtime experiment.
- `docs/submission-revision-20260913/atomistic-fire-validation.json` contains all five paired optimizer comparisons on 80 reactions. `docs/completed-results-20260916/` contains the adaptation, local-validation and joint-training summaries, evidence roles and hash inventory. `docs/atomistic-reactive-local-validation/` supplies the new local registry, reference labels, per-molecule results and evaluated model identities.

The package README gives the four CPU replay commands, and `SHA256SUMS.json` binds every packaged file. Other molecular and crossed-reference results retain their manuscript tables and figures. Model inference, training and electronic-structure calculations require the corresponding checkpoints and compute environments and are outside these CPU replays. Appendix F describes the analysis design.

The reference environment uses Python 3.11.16, PySCF 2.14.0, ASE 3.29.0, NumPy 2.4.6, SciPy 1.17.1, pandas 3.0.5, and Matplotlib 3.11.1. MLIP environments use PyTorch 2.12.1 and the package versions recorded with Table 1. The original references ran on an arm64 CPU; the extension used isolated Conda environments on arm64 and Linux x86-64 CPUs, and an A800-class GPU for checkpoint inference and CI-NEB. The paired optimizer experiment uses the per-checkpoint GPU, package and precision records in `docs/atomistic-optimizer-control-20260924/protocol.json`.

I Paired optimizer comparison within fixed runtimes

This exploratory experiment uses the unchanged primary 20-reaction selection and all five broad-domain checkpoints. Each of the 100 model–reaction pairs shares one GPU, software environment, checkpoint, precision, endpoints and IDPP initialization. Each optimizer starts from a fresh calculator and band. Execution order alternates with reaction rank, giving ten BFGS-first and ten FIRE-first pairs per checkpoint; both receive the same rank-based random seed. ASE is 3.29.0. MACE uses float64 and the other checkpoints use float32.

Each path has ten images, fixed endpoints, improved tangent and a 0.1 eV/Å^2 spring. Ordinary NEB receives at most 500 steps to reach 0.5 eV Å^{-1}; a converged first stage proceeds to climbing NEB with a further 500-step cap and 0.05 eV Å^{-1} threshold. BFGS uses $\alpha = 70$ and maximum step 0.03 Å. FIRE uses $dt = 0.1$, $dt_{\max} = 1.0$, maximum step 0.03 Å, $N_{\min} = 5$, $f_{\text{inc}} = 1.1$, $f_{\text{dec}} = 0.5$, $a_{\text{start}} = a = 0.1$ and $f_a = 0.99$, with downhill checking disabled.

All 200 terminal paths enter the completion analysis. The aggregate FIRE-minus-BFGS difference averages the five models, then uses 10,000 bootstrap draws of the 20 reaction identities jointly across models and optimizers (seed 2027). Its 95% interval is [10, 28] percentage points around a 19-point gain. Per-model differences use 99% paired reaction-bootstrap intervals and exact two-sided McNemar tests with Holm adjustment across five checkpoints. Table 21 gives every model, including the discordant pairs. Table 22 restricts barrier comparisons to the fixed intersection where both optimizers converge.

Table 21: Paired optimizer comparison within fixed runtimes. Counts use 20 reactions per model. B-only and F-only are discordant completions; Δ is FIRE minus BFGS in percentage points. Per-model intervals are 99% paired reaction-bootstrap intervals, and p_{Holm} adjusts exact McNemar tests across five models. The aggregate is 30/100 versus 49/100, with a 19-point gain and 95% interval [10, 28].

Model	BFGS	FIRE	B-only	F-only	Δ [99% CI]	p_{Holm}
MACE	7/20	8/20	1	2	5 [-15, 30]	1.000
SevenNet	7/20	11/20	0	4	20 [0, 45]	0.375
MatterSim	6/20	13/20	0	7	35 [10, 65]	0.078
M3GNet	3/20	5/20	2	4	10 [-20, 40]	1.000
Orb	7/20	12/20	0	5	25 [5, 50]	0.250

Table 22: Barrier MAE on the common-convergence intersection of the fixed-runtime experiment. All values are in eV. Δ averages the paired FIRE-minus-BFGS absolute-error differences on the displayed subset; these subsets differ across models. The M3GNet subset contains one reaction.

Model	Common n	BFGS MAE	FIRE MAE	Δ MAE
MACE	6	1.4440	1.4442	+0.0002
SevenNet	7	1.1089	1.1082	-0.0007
MatterSim	6	0.7868	0.7848	-0.0020
M3GNet	1	1.5674	1.5684	+0.0011
Orb	7	0.7379	0.7362	-0.0017

The supplement provides the 100 paired rows, scientific summary and protocol in `docs/atomistic-optimizer-control-20260924/`. `scripts/replay_atomistic_optimizer_control.py` reproduces the completion counts, bootstrap intervals, exact tests and common-convergence barrier point estimates.

J Optimizer sensitivity on 80 validation reactions

The same 80 reaction identities are paired across BFGS and FIRE for each of five models; the 400 model–reaction paths are not 400 independent reaction draws. The validation cohort is disjoint from the primary 20 reactions, and the optimizer comparison is exploratory. Endpoints, paths and model identities are fixed; no BFGS outcome selects which reaction receives FIRE. There are 156 converged and 244 nonconverged FIRE paths. Convergence-rate intervals use 10,000 paired reaction bootstrap draws. Two-sided exact McNemar tests concern each model's discordant reaction pairs; Holm adjustment covers the five model comparisons. The conditional barrier analysis uses only the fixed intersection where both methods converge and labels its small denominators explicitly.

Table 23: Paired convergence on the same 80 reactions for each model. Intervals resample reactions; exact McNemar tests are adjusted over five models. The aggregate is 113/400 versus 156/400 model–reaction paths, with 80 independent reaction units.

Model	BFGS/80	FIRE/80	Change	95% CI	Holm p
MACE	15	27	0.1500	[0.0500, 0.2500]	0.0377
SevenNet	23	28	0.0625	[-0.0250, 0.1625]	0.3785
MatterSim	22	31	0.1125	[0.0250, 0.2000]	0.1406
M3GNet	16	26	0.1250	[0.0125, 0.2375]	0.1917
Orb	37	44	0.0875	[-0.0250, 0.2000]	0.3785

Table 24: Barrier MAE in eV on the fixed common convergence intersection. Intervals concern FIRE-minus-BFGS absolute error within that intersection; they do not describe all attempted paths.

Model	Common n	BFGS MAE	FIRE MAE	Paired difference 95% CI
MACE	12	1.7922	1.7914	[-0.0014, -0.0002]
SevenNet	18	1.0764	1.0605	[-0.0482, 0.0009]
MatterSim	19	0.9767	0.9495	[-0.0647, -0.0030]
M3GNet	9	0.9883	0.9672	[-0.0719, 0.0094]
Orb	30	0.6959	0.6933	[-0.0053, -0.0004]

K Adaptation on reactive development data and new local formulas

Data and partitions. The adaptation study starts from SevenNet-Omni with the `omol25_low` head (checkpoint SHA-256 prefix `ca81bd3aac9f`). The reaction and local configurations form a development dataset. Reactions sharing an endpoint identifier form connected groups before splitting; all configurations of a reaction stay together. The training partition has 138 reactions in 44 groups (966 configurations) and 12 local molecules (156 configurations). Validation has 44 reactions in nine groups (308 configurations) and two local molecules (26 configurations). Held-out development evaluation has 18 reactions in nine groups (126 configurations) and two local molecules (26 configurations). The split seed is 20260915. Chemical formulas need not be disjoint across these partitions, and pretraining overlap is unknown.

Adaptation design. Seven conditions use seeds 0, 1 and 2 and 600 training steps, with validation every 100 steps. A batch samples 12 reactive configurations and their reactant energy anchors; local-constraint and replay conditions also use eight local configurations. The objective combines configuration-averaged Cartesian force MSE and within-reaction relative-energy error per atom with weight 0.1. The residual energy branch has two scalar/vector message-passing layers, 64 channels, 16 radial features and a smooth 5 Å cutoff. It reads elements and geometry and has 102,082 trainable parameters in both gated and ungated versions. The gated version multiplies an energy readout by an invariant sigmoid gate; the ungated control adds two readouts and divides by $\sqrt{2}$. Forces are full energy gradients, including the gate derivative. Full fine-tuning updates 25,759,631 backbone parameters. AdamW uses learning rates 10^{-3} for the residual and 3×10^{-5} for the backbone, weight decay 10^{-6} and gradient-norm clipping at 10.

The constrained conditions apply a Lagrangian penalty to the local MAE increase above $\max(0.05\,\mathrm{MAE}_{\mathrm{base}}, 0.002\ \mathrm{eV\,Å^{-1}})$. The multiplier starts at one, has update rate 10 and is clipped to $[0, 100]$. Replay instead adds local force MSE with weight one. Each condition selects the lowest reactive validation MAE among checkpoints satisfying the same local tolerance; the initial model is also a candidate. Static barriers use fixed reactant and transition-state geometries. Reactive metrics average within a reaction, then across reactions; local metrics average within a molecule, then across molecules. Reported adaptation means average the three seeds.

Development comparisons. The gated constrained residual reduces reactive force MAE from 0.1853 to 0.0743 eV/Å, while ordinary full fine-tuning reaches 0.0658 and fine-tuning with replay

reaches 0.0712. The constrained gated-minus-ungated reactive difference is -0.00091 with paired seed/connected-group 95% interval $[-0.00327, 0.00159]$. Against fine-tuning with replay, the gated constrained residual has higher reactive error by 0.00308 $[0.00088, 0.00550]$. Ungated residual replay has local MAE 0.3379 and meets the held-out development tolerance in zero of three seeds; all other initial conditions meet it in three of three. Adaptation lowers reactive error; the gated-minus-ungated difference has an interval spanning zero.

Prospective local evaluation. We select the 21 models before obtaining labels and predictions for 12 new chemical formulas, disjoint from all 1,608 preceding configurations. The fixed panel selects one isomer per formula by hash order from ASE 3.29.0 G2, requiring carbon, only H/C/N/O, even electrons, zero initial magnetic moment, two to four heavy atoms and at most ten atoms. Each base geometry has two fixed random internal directions at signed RMS displacement 0.02 Å, giving five configurations per molecule. Translation and infinitesimal rotation are removed. All 60 local energy-and-force references converge with ωB97X/6-31G(d), PySCF 2.14.0 density fitting, grid level three, SCF tolerance 10^{-9} Hartree and analytic forces. The primary comparison is gated-plus-constraint minus the original model, with 10,000 paired seed/molecule bootstrap draws and tolerance 0.004868 eV/Å. Its difference is -0.06161 $[-0.07678, -0.04741]$, meeting that criterion. Its difference from the ungated constrained residual is -0.00098 $[-0.00508, 0.00229]$; its error exceeds fine-tuning with replay by 0.01054 $[0.00683, 0.01423]$. The latter comparisons are descriptive. This panel evaluates local transfer.

Joint-training extension. A further 15 runs cross three seeds with five conditions: full fine-tuning, gated and ungated residuals with the backbone held constant, and jointly trained gated and ungated backbone/residual models. All use local replay, the original pretrained initialization, the same split and 600-step budget. Joint models have 25,861,713 trainable parameters. Checkpoint selection precedes evaluation on the reused 12-formula local panel; these comparisons are exploratory. Joint gated training has reactive MAE 0.07208, development-local MAE 0.13652 and reused-local MAE 0.02178. Against full fine-tuning with replay, the reactive difference is 0.00091 $[-0.00057, 0.00246]$, the development-local difference is 0.07475 $[0.06272, 0.08881]$, and the reused-local difference is -0.00343 $[-0.00710, 0.00010]$. The local-panel interval includes zero, while the two-molecule development panel has higher error under joint training. The tables report all conditions and twelve paired contrasts.

Table 25: Development evaluation of seven adaptation conditions, with three training seeds each. Force MAE is in eV/Angstrom and barrier MAE in eV. Reactions and local molecules receive equal weight within their respective panels. The 18 reactions and two local molecules belong to previously examined cohorts. Local passes count training seeds meeting the held-out development tolerance.

Method	Reactive force	Local force	Static barrier	Local passes
Original checkpoint	0.1853	0.2137	0.3203	–
Ungated residual	0.0748	0.1532	0.2698	3/3
Gated residual	0.0740	0.1538	0.2594	3/3
Ungated + constraint	0.0752	0.1442	0.2662	3/3
Gated + constraint	0.0743	0.1396	0.2666	3/3
Ungated + replay	0.0905	0.3379	0.2611	0/3
Full fine-tuning	0.0658	0.1061	0.2908	3/3
Full fine-tuning + replay	0.0712	0.0618	0.2113	3/3

Table 26: Local evaluation of the 21 selected adaptations on 12 new chemical formulas and 60 configurations. Models and evaluation settings were selected before obtaining these reference labels and predictions. Intervals pair training seeds and molecules. Only gated-plus-constraint versus the original checkpoint is the designated primary comparison; other contrasts are descriptive.

Method	Local force MAE	Difference from original [95% CI]
Original checkpoint	0.09736	–
Ungated residual	0.03688	-0.06048 [-0.07542, -0.04665]
Gated residual	0.03790	-0.05946 [-0.07397, -0.04602]
Ungated + constraint	0.03673	-0.06063 [-0.07567, -0.04631]
Gated + constraint	0.03575	-0.06161 [-0.07678, -0.04741]
Ungated + replay	0.04228	-0.05508 [-0.07499, -0.03638]
Full fine-tuning	0.03050	-0.06686 [-0.08272, -0.05257]
Full fine-tuning + replay	0.02520	-0.07216 [-0.08821, -0.05726]

Table 27: Joint-training extension: five conditions, three seeds each, all with local replay. All panels are development or exploratory for this extension, including the reused 12-formula local panel. Force MAE is in eV/Angstrom and static barrier MAE in eV. Fixed-backbone residual controls and full fine-tuning are retained.

Method	Reactive	Dev. local	Reused local	Static barrier
Full fine-tuning + replay	0.07117	0.06177	0.02520	0.21127
Fixed-backbone ungated + replay	0.09376	0.36449	0.03960	0.25378
Fixed-backbone gated + replay	0.09134	0.36451	0.04306	0.26918
Joint ungated + replay	0.07254	0.13229	0.02371	0.19267
Joint gated + replay	0.07208	0.13652	0.02178	0.20231

Table 28: All twelve joint-training paired contrasts, with 5,000 seed-and-connected-group bootstrap draws. Intervals are descriptive and unadjusted. Reactive configurations are clustered within nine endpoint-connected groups; the local panels contain two and twelve molecule groups, respectively.

Comparison	Panel	MAE difference [95% CI]
Joint gated minus Fine-tuning	Reactive	0.00091 [-0.00057, 0.00246]
Joint gated minus Fine-tuning	Dev. local	0.07475 [0.06272, 0.08881]
Joint gated minus Fine-tuning	Reused local	-0.00343 [-0.00710, 0.00010]
Joint gated minus Joint ungated	Reactive	-0.00045 [-0.00196, 0.00075]
Joint gated minus Joint ungated	Dev. local	0.00423 [-0.01600, 0.01494]
Joint gated minus Joint ungated	Reused local	-0.00193 [-0.00491, 0.00092]
Joint gated minus Fixed-backbone gated	Reactive	-0.01926 [-0.03020, -0.01484]
Joint gated minus Fixed-backbone gated	Dev. local	-0.22799 [-0.30030, -0.14437]
Joint gated minus Fixed-backbone gated	Reused local	-0.02129 [-0.03644, -0.01052]
Joint ungated minus Fixed-backbone ungated	Reactive	-0.02123 [-0.02926, -0.01527]
Joint ungated minus Fixed-backbone ungated	Dev. local	-0.23219 [-0.36244, -0.12109]
Joint ungated minus Fixed-backbone ungated	Reused local	-0.01589 [-0.02946, -0.00703]

L Complete static reaction-energy endpoints

Table 29: All three registered energy-error endpoints on the same 200 reactions per checkpoint. Entries are MAE [95% reaction-bootstrap interval], in eV, using 10,000 draws and seed 2027. The zero-prediction row is a descriptive reference-scale anchor computed on the same reactions; it is not a learned baseline.

Checkpoint	Forward barrier	Reverse barrier	Reaction energy
M3GNet	0.927 [0.814, 1.045]	0.810 [0.721, 0.905]	0.859 [0.757, 0.971]
MACE-MP-0	1.295 [1.204, 1.390]	0.999 [0.908, 1.091]	0.621 [0.555, 0.691]
MatterSim-1M	0.764 [0.694, 0.838]	0.572 [0.512, 0.635]	0.395 [0.339, 0.455]
Orb-v3	0.552 [0.505, 0.599]	0.390 [0.349, 0.432]	0.282 [0.249, 0.317]
SevenNet-0	0.993 [0.909, 1.081]	0.694 [0.616, 0.774]	0.546 [0.480, 0.614]
Zero prediction	3.775	2.079	1.808

The three endpoints retain all reactions, independently of path-search completion. Intervals are marginal for each endpoint and checkpoint. The accompanying CSV and CPU replay reproduce all 15 estimates and their interval bounds.

M Reaction-level error and completion associations

The analysis uses every one of the 200 fixed-geometry reactions per broad-domain checkpoint. For each reaction, component absolute errors are pooled across the stored near-path configurations and divided by their total number of force components. The paired barrier endpoint is the absolute forward-barrier error at reference geometries. We compute Spearman correlation separately within each checkpoint and resample reactions 10,000 times (seed 20260923).

The BFGS comparison uses all 20 primary and 80 disjoint validation reactions per checkpoint. It retains nonconverged paths and compares their reference-geometry force error with that of converged paths. Resampling is stratified by the primary/validation cohort. The 99% intervals adjust five models within each endpoint family. These are associations among sampled reactions; reaction difficulty and reference mismatch can contribute to both errors.

Table 30: Complete within-checkpoint associations. Force/barrier correlations use all 200 reactions; the completion comparison uses all 100 assigned CI-NEB reactions. Force MAE difference is non-converged minus converged, in eV/Å. All five completion differences have intervals spanning zero.

Model	Force/barrier ρ [99% CI]	Δ force MAE [99% CI]	Complete
M3GNet-MatPES	0.144 [-0.043, 0.326]	-0.074 [-0.233, +0.051]	19/100
MACE-MP0	0.243 [0.051, 0.423]	-0.012 [-0.133, +0.079]	22/100
MatterSim-1M	0.303 [0.130, 0.469]	-0.019 [-0.135, +0.068]	27/100
Orb-v3	0.252 [0.076, 0.411]	-0.037 [-0.101, +0.012]	44/100
SevenNet-0	0.380 [0.210, 0.538]	+0.034 [-0.020, +0.085]	29/100