Extending the N-body Benchmark: A Cross-Model Study of Geometric Deep Learning Architectures

Patrik Šimurka, Martin Karas, Juraj Mečír, Andrej Kozák, Ondrej Hovorka, Pavol Skubák, Ján Kmeťko
Simona Biosystems, Bratislava, Slovakia
patrik.simurka@simonabio.com

Code available at *https://github.com/Simona-Biosystems/Extending-the-N-body-Benchmark-A-Cross-Model-Study-of-Geometric-Deep-Learning-Architectures*

Abstract

In our previous work, we argued that the N-body problem has not been formalized properly as a benchmark for geometric deep learning, and we proposed a benchmark centered around long-horizon rollouts and macro-property tests to evaluate the fidelity of predicted dynamics. In this paper, we extend that benchmark to a cross-model study. We evaluate multiple modern geometric architectures—EquiformerV2, SEGNN, Ponita, and CGENN—alongside a non-equivariant Transformer baseline under matched training time and parameter budgets. We standardize inputs/targets, datasets, and evaluation protocols; and we report macro-property similarity via the Kolmogorov–Smirnov two-sample test (with Fisher-combined p-values). We open-source all code and configurations.

1 Introduction

The practicality of simulating many-body systems is constrained by small time-step requirements, which make long-time horizons computationally expensive. Neural surrogates that step coarsely while preserving realistic dynamics are attractive in practice and naturally motivate a benchmark for geometric deep learning.

In our previous paper[1], we proposed assessing learned simulators through long-horizon rollouts and statistical macro-property tests, rather than only static one-step errors. Here, we extend that framework to a fair, controlled comparison of multiple geometric architectures and a non-equivariant baseline. Our goals are to (a) provide a robust, reproducible leaderboard for

N-body dynamics respecting physical symmetries and rollout stability, (b) stress-test how design choices (equivariance vs. no equivariance, attention vs. message-passing, spherical harmonics vs. multivectors) influence macro-property fidelity at larger time steps than the integrator, and (c) shed some light on efficiency trade-offs that are relevant in practice.

We conduct controlled experiments under matched parameter and training time budgets. We reuse the macro-property tests and evaluation flow from our previous study—collisions, group collisions, escapes, sharp turns (30°/45°), and stickings—quantified via the Kolmogorov–Smirnov (K–S) test and Fisher-combined p-values.

2 Related Work

Equivariant message passing and attention models—such as SEGNN[2], SE(3)-Transformer[3], tensor field networks[4], Clifford group equivariant neural networks[5], and Ponita[6]—have demonstrated strong data efficiency and improved one-step accuracy on geometric tasks. Recent developments in equivariant attention (e.g., EquiformerV2[7]) further improve expressivity and throughput. Despite these advances, systematic benchmarking on long-horizon dynamical fidelity remains sparse and fragmented, often emphasizing limited subsets of physical metrics (e.g., only energy conservation) or single-step prediction in isolation. Building on our earlier benchmark formalization, we aim to provide a consistent, fair evaluation across architectures with a focus on rollout stability, macro-properties, and efficiency.

3 Benchmarks and Extensions

We briefly recapitulate the benchmark we introduced previously and describe the extensions required for a cross-model study.

Rollouts over long horizons
We evaluate models in free-rollout mode by iteratively feeding predictions back into the model to generate long trajectories. This emphasizes stability, error accumulation, and realistic dynamical behavior.

Macro-properties over micro-alignment
Given the sensitivity to initial conditions, we prioritize macro-properties—collisions, group collisions, escapes, sharp turns (30°/45°), and stickings—over direct trajectory alignment. We compute these properties identically for ground truth and model rollouts and quantify their distributional similarity with the K–S two-sample test and Fisher-combined p-values.

Standardization across models
All models predict position deltas (*Δpos*) and velocities from the same underlying data: particle positions, velocities, and masses in gravitational N-body systems. Each architecture processes

these inputs using its reference graph construction and feature encoding from the original papers/codebases. Equivariant models (PONITA, SEGNN, EquiformerV2, CGENN) construct fully-connected graphs with architecture-specific geometric encodings (typed scalars/vectors, spherical harmonics, or multivectors) and varying treatments of mass and relative coordinates. The Graph Transformer baseline operates directly on concatenated position and velocity coordinates without mass, graph edges, or geometric typing. Despite these architectural differences in data preprocessing, we maintain consistent physical simulation parameters (interaction strength 2.0, softening 0.2), Verlet integration at dt=0.01 with 10× coarse stepping, double precision, and five-body systems throughout.

4 Models

We consider four equivariant architectures and one non-equivariant baseline. All models are trained with the same loss family and data pipeline and are evaluated with identical macro-property and diagnostic suites.

Ponita[6]: a state-of-the-art E(n)-equivariant attention/message-passing architecture operating on position–orientation features. We follow the reference configuration from the original paper, adapting feature typing (masses as scalars, velocities/relative positions as vectors).

SEGNN[2]: steerable E(3)-equivariant message passing with irreducible representations. We use l=0 (scalars) and l=1 (vectors) for hidden features, with *l_max_h* and *l_max_attr* tuned to fit capacity budgets.

EquiformerV2[7]: an equivariant transformer using tensor product features with efficient attention. We adopt the recommended radial basis and cutoff set to realize a complete graph.

CGENN[5]: a Clifford-group equivariant network.

Graph Transformer (non-equivariant baseline): a standard multi-head attention transformer operating on node features with no geometric equivariance constraints. Serves as a capacity-matched baseline to isolate the benefits of equivariant architectures.

5 Methodology

Training objective

We train to predict next-step position deltas (*Δpos*) and velocities simultaneously. This dual-target approach is necessary for autoregressive rollouts: position deltas advance particle positions, while predicted velocities serve as input features for the next step. Without velocity prediction, the model would need to compute finite differences from positions, which amplifies

numerical errors in long rollouts. Models operate on complete graphs with edge features encoding relative positions and distances.

Summary of data generation

Ground truth trajectories are generated using a Verlet integrator with a step size of 0.01. As demonstrated in our previous paper[1], this step size is small enough that macroscopic properties remain stable compared to smaller steps on our distributions of initial conditions; we verify this through energy and macroproperty diagnostics. For coarse stepping, we subsample by N=10 to form training pairs and evaluate rollouts at the same coarse step (N=10). We train on five-body systems for comparability with our earlier work. Initial conditions consist of random positions and velocities sampled to create a mixture of bound and unbound regimes and a range of mass distributions; the gravitational constant is fixed to 1.0, the interaction strength to 2.0, and the softening parameter to 0.2. Trajectories are generated on-the-fly during training to avoid overfitting, with each simulation providing 10,000 timesteps at *dt*=0.01.

Fairness controls

For fairness controls, we impose two matched capacity regimes per model (≈2M and ≈10M parameters). Each model is trained for 8 hours under a fixed training budget. We use the AdamW optimizer with inverse square-root decay and warmup; *warmup_steps* is set to 2048. We keep this configuration consistent with our previous setup[1]. To ensure determinism and reproducibility, we enforce fixed seeds, identical preprocessing, and comprehensive metadata logging. All models use double precision (*float64*) for numerical stability in long rollouts.

Hyperparameter Optimization

To ensure fair comparison across architectures, we conduct systematic hyperparameter optimization (HPO) for each model within the capacity budgets. We use Optuna with TPE (Tree-structured Parzen Estimator) sampling, allocating 12 trials per model per budget regime. For each trial, we:

- Sample learning rate log-uniformly in [0.05, 2.0]
- Sample architecture-specific hyperparameters (depth, width, attention heads, *l_max*) from discrete grids matched to the reference implementations
- Adjust width parameters iteratively to meet the target parameter count within ±7% tolerance
- Train for 45 minutes per trial
- Evaluate using the Fisher-combined p-value from macro-property K-S tests on rollouts

We select the configuration maximizing the Fisher-combined p-value (equivalently, minimizing the negative log p-value) and train the final model for the full 8-hour budget. All HPO trials share the optimizer schedule (AdamW with inverse square-root decay, *warmup_steps*=2048).

Losses

Mean squared error on both *Δpos* and velocities. We do not use auxiliary physics losses (energy, momentum, center-of-mass) as experimentation showed these were counterproductive for macro-property alignment in this setup.

Evaluation Protocol

Each model is evaluated at convergence (end of 8-hour training) using rollouts of 1000 coarse steps (10,000 integrator steps at *dt*=0.01) across 64 independent trajectories per evaluation. We compute six macro-properties—collisions, group collisions, escapes, sharp turns at 30° and 45°, and stickings—identically for both ground truth and model rollouts. Experiments have shown that these macroproperties correlate very well with other macroproperties like energy conservation, so we do not include more. For each macro-property, we apply the Kolmogorov–Smirnov two-sample test, then aggregate results via Fisher's method to obtain a combined p-value. We report this combined p-value as our primary metric, where higher values indicate better alignment between model and ground truth macro-property distributions. As in our previous work, we treat p-values as continuous measures of distributional similarity rather than binary thresholds.

6 Results

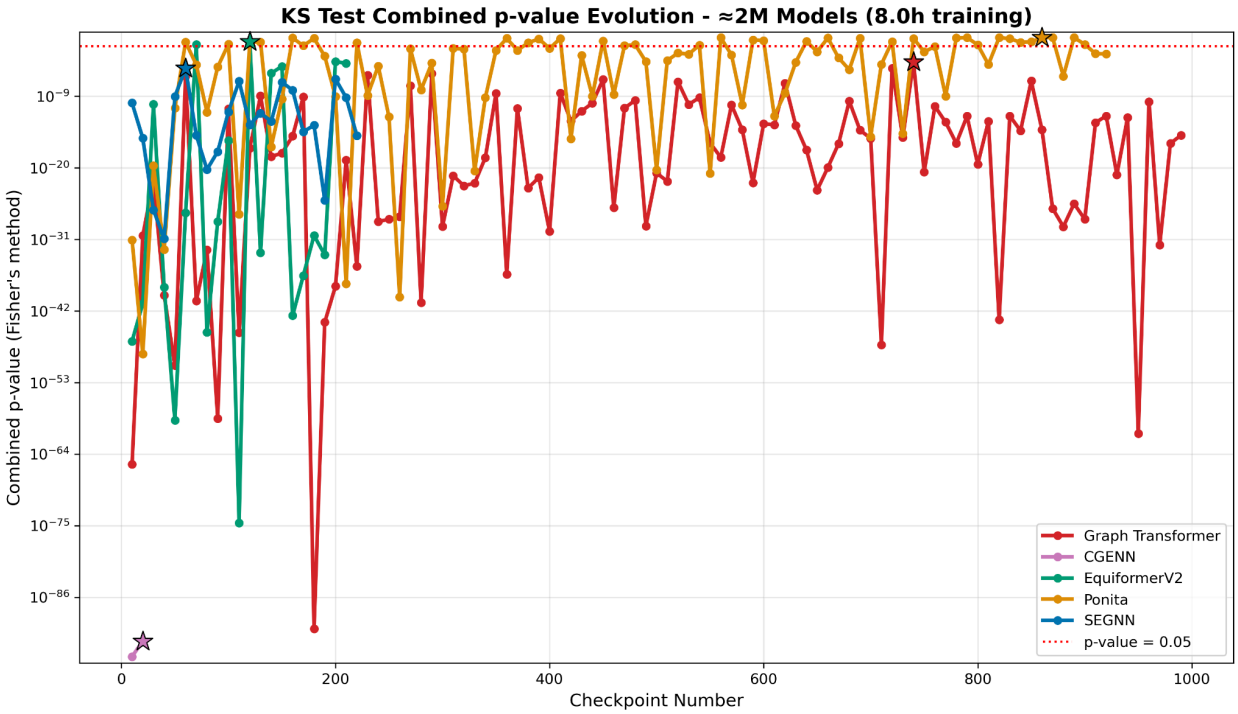
Table 1 presents the Fisher-combined p-values for all models at both capacity budgets. For reference, ground truth vs. ground truth comparisons consistently yield p-values >0.95, confirming the stability of our evaluation procedure.

Model	**≈2M Parameters best p-value**	**≈10M Parameters best p-value**
EquiformerV2	0.236	0.082

SEGNN	1.8818e-05	0.17
Ponita	0.999	0.997
CGENN	1.81e-93	5e-100
Graph Transformer	1.82e-04	6.91e-03

Table 1: Fisher-Combined P-Values at Convergence

Values represent the best Fisher-combined p-values from K-S tests across six macro-properties. Higher is better (max=1.0). Capacity scaling reflections compare best p-values within the 8-hour budget and do not disentangle depth-versus-width trade-offs.



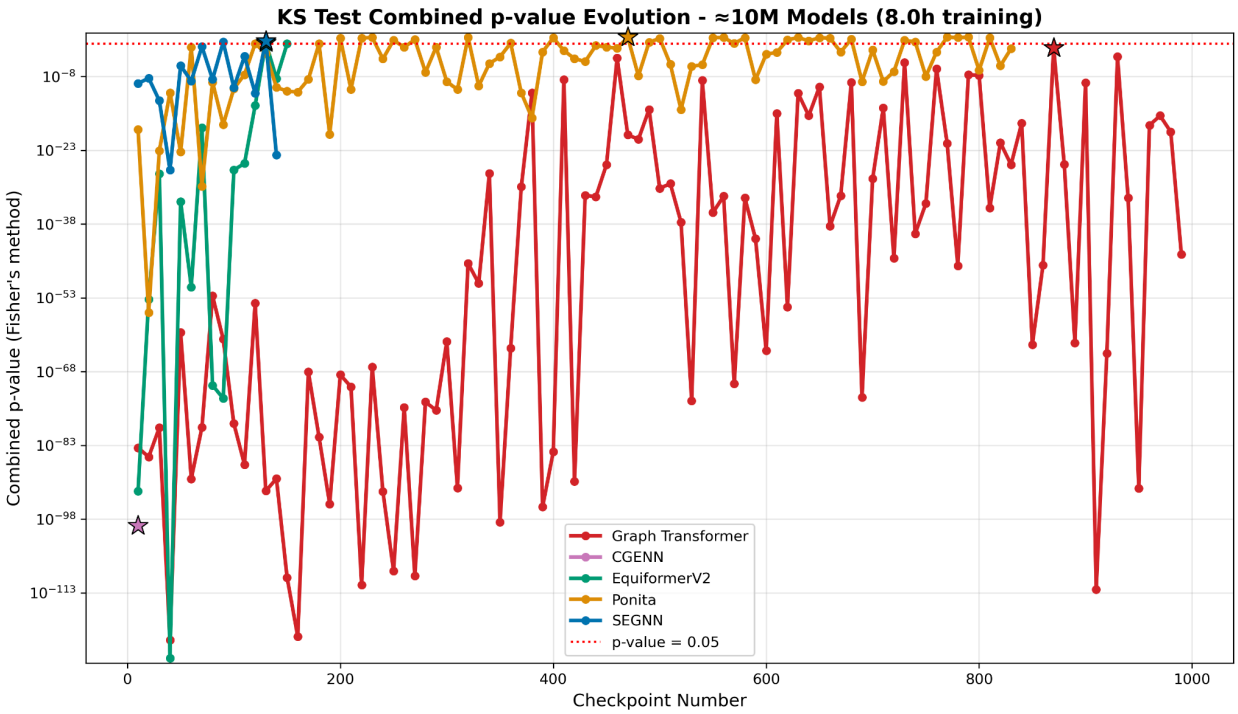


Figure 1 and 2: Evolution of Fisher-combined Kolmogorov–Smirnov p-values over the first 8 hours of training for ≈2M (top) and ≈10M (bottom) parameter runs. Each line traces checkpoints saved under the 8-hour budget; stars mark the best-within-budget p-value, matching Table 1 entries. The horizontal dotted line denotes p = 0.05. The vertical axes are logarithmic and auto-ranged per plot to maintain readability.

Observations

Equivariance advantage: At ≈2M parameters, equivariant models span $p \in [1.88\times10^{-5}, 0.999]$, while the Graph Transformer achieves 1.82×10^{-4}. Interestingly, the baseline outperforms SEGNN at this scale, suggesting equivariance alone is insufficient without adequate capacity or training time. At ≈10M parameters, the gap widens: the best equivariant models (Ponita, SEGNN) exceed the baseline by 1–2 orders of magnitude ($p \approx 0.17$–0.997 versus 6.91×10^{-3}).

Capacity scaling: Models respond differently to increased parameters. SEGNN improves dramatically (~9,000-fold), confirming it was capacity-starved at 2M. Ponita maintains near-perfect performance (~0.999) at both scales, having already saturated at the smaller size. EquiformerV2 and CGENN show slight degradation with more parameters, likely because larger models need more than 8 hours to converge. The Graph Transformer improves 38-fold, demonstrating that non-equivariant models can partially compensate via capacity.

Architectural diversity: Within the equivariant family, performance varies by orders of magnitude. At ≈2M, Ponita (0.999) vastly exceeds EquiformerV2 (0.236), which beats SEGNN (1.88×10^{-5}) by ~10^4×. At ≈10M, Ponita remains strongest (0.997), but SEGNN (0.17) overtakes

EquiformerV2 (0.082). This spread indicates that architectural details within the equivariant family matter as much as equivariance itself.

Convergence speed: Training efficiency varies widely. Within 8 hours at ≈2M, the Graph Transformer completes 990 steps, Ponita 920, SEGNN 220, EquiformerV2 210, and CGENN only 20. At ≈10M, throughput drops further (Graph Transformer 990, Ponita 830, EquiformerV2 150, SEGNN 140, CGENN 10). Figures 1–2 show p-value evolution; note that y-axes auto-scale per plot for readability—compare magnitudes using the horizontal p = 0.05 reference line.

Caveats: We report best p-values within 8 hours (stars in Figures 1–2). We did not control depth-vs-width allocation; models reach similar parameter counts via different configurations, affecting convergence speed. SEGNN appears weak at 8 hours but exceeds p > 0.1 given ~2.5× more time in extended runs. CGENN trains so slowly (1–2 checkpoints per 8 hours) that its p-values lack reliability; we include it for completeness regarding practical training costs.

7 Discussion & Future Work

Our cross-model study supports macro-property tests as a practical measure of dynamical fidelity beyond one-step metrics. We outline several takeaways and open directions below.

Equivariance helps, but design matters. Explicit equivariance is expected to benefit long-horizon stability; however, design trade-offs across attention versus message passing and across representation choices (spherical harmonics, tensor products, Clifford algebra) remain to be quantified. Systematically mapping the Pareto frontier of stability versus throughput is a key goal of this benchmark.

Capacity and scaling. Beyond comparing fixed “small” (≈2M) and “medium” (≈10M) capacities, we plan controlled scaling studies at matched parameter counts: depth-vs-width sweeps (e.g., layers vs channels, heads, and *l_max*/channel-base for equivariant models), with p-values and throughput reported jointly. This will clarify which scaling direction yields better long-horizon stability per unit compute.

Horizon robustness. We will benchmark longer rollouts (beyond 1,000 coarse steps), reporting combined p-values and auxiliary diagnostics as a function of horizon. This will surface failure modes (e.g., energy/momentum drift, trajectory collapse) and architecture-specific robustness.

Generalization in N. While training on five-body systems simplifies controlled comparisons, scaling training N and systematically testing N-generalization are natural next steps.

Integrator and *dt*. The Verlet reference with step 0.01 is sufficient on our distributions; exploring stricter integrators or adaptive *dt* could refine ground truth and stress-test models further.

Toward molecular dynamics. The same evaluation principles—macro-properties, stability diagnostics, and fairness controls—should transfer beyond gravity. We plan to test short-range and mixed-interaction systems next.

8 Conclusion

We extend our N-body benchmark into a fair, reproducible cross-model evaluation of geometric deep learning architectures. By focusing on macro-property preservation in long-horizon rollouts, complemented by physical diagnostics and efficiency metrics, we provide a practical measure of dynamical fidelity that goes beyond one-step accuracy. Our results and open-source artifacts are intended to serve as a basis for future work on stable, efficient learned simulators and to inform architectural choices in geometric deep learning.

References

1. (PDF) Revising the N-body Problem as a Benchmark for Geometric Deep Learning. *ResearchGate* https://www.researchgate.net/publication/386552469_Revising_the_N-body_Problem_as_a_Benchmark_for_Geometric_Deep_Learning doi:10.13140/RG.2.2.24804.69760.
2. Brandstetter, J., Hesselink, R., Pol, E. van der, Bekkers, E. J. & Welling, M. Geometric and Physical Quantities Improve E(3) Equivariant Message Passing. Preprint at https://doi.org/10.48550/arXiv.2110.02905 (2022).
3. Fuchs, F. B., Worrall, D. E., Fischer, V. & Welling, M. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. Preprint at https://doi.org/10.48550/arXiv.2006.10503 (2020).
4. Thomas, N. *et al.* Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. Preprint at https://doi.org/10.48550/arXiv.1802.08219 (2018).
5. Ruhe, D., Brandstetter, J. & Forré, P. Clifford Group Equivariant Neural Networks. Preprint at https://doi.org/10.48550/arXiv.2305.11141 (2023).

6. Bekkers, E. J., Vadgama, S., Hesselink, R. D., Linden, P. A. van der & Romero, D. W. Fast, Expressive SE(n) Equivariant Networks through Weight-Sharing in Position-Orientation Space. Preprint at https://doi.org/10.48550/arXiv.2310.02970 (2024).
7. Liao, Y.-L., Wood, B., Das, A. & Smidt, T. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. Preprint at https://doi.org/10.48550/arXiv.2306.12059 (2024).