DISTANCE-DEPENDENT ANGULAR RESOLUTION FOR EQUIVARIANT INTERATOMIC POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariant machine-learned interatomic potentials (MLIPs) typically use the same angular resolution on every edge, regardless of interatomic distance. We introduce radial gates with learnable cutoff radii that adapt angular resolution to interatomic distance within each message-passing block. We further introduce a band mixture-of-experts (Band-MoE) that uses these learned radial cutoffs to route edges deterministically to parameter-sharing experts with different maximum retained degrees. Separately, we also show that a reflection-symmetric simplification of the edge-frame operator can lead to 25.9% parameter reduction. We reach $2.10\times$ batched inference speedup with 50% training memory reduction, while maintaining accuracy close to the noise thresholds for forces and energy at maximum angular degree of 4. We show that at higher angular degree ($L_{\max} = 8$), our method outperforms the baseline model by 27% with an inference speedup of $2.31\times$. We evaluate the method on non-periodic molecules from OMol25 and periodic inorganic crystals from OMat24. The Band-MoE and the learnable radial cutoff methodologies are applicable to other MLIPs.

1 INTRODUCTION

Machine-learned interatomic potentials (MLIPs) predict energies and forces from atomic configurations and replace density functional theory (DFT) calculations in molecular dynamics and geometry optimization, when a DFT calculation at every step is too expensive. Most current MLIPs are equivariant: their features transform under rotations as irreducible representations indexed by the spherical-harmonic degree ℓ up to a global maximum degree $L_{\max}$ (Thomas et al., 2018; Batzner et al., 2022; Batatia et al., 2022; Passaro & Zitnick, 2023). Throughout this work, the maximum retained degree of an edge, k_e, is the highest degree ℓ evaluated on that edge; every degree above k_e is exactly zero there. The angular resolution of a feature is the finest angular variation its retained degrees can represent, and is therefore set by the highest degree the feature retains.

In popular MLIP methods including NequIP, Allegro, MACE, eSCN, EquiformerV2, and UMA (Batzner et al., 2022; Musaelian et al., 2023; Batatia et al., 2022; Passaro & Zitnick, 2023; Liao et al., 2024; Wood et al., 2025), $L_{\max}$ is a global setting. So, every edge uses the same maximum retained degree regardless of its length (Figure 1a). Raising $L_{\max}$ increases both the number of features and the size of the linear maps between degrees. At $L_{\max} = 4$, the edgewise computation accounts for 89.5% of a forward pass of eSEN backbone at 6500 atoms (Appendix B). Lowering the maximum retained degree of an edge according to its length could therefore lower both the computational and memory cost, provided that the reduction preserves equivariance and introduces no discontinuity in the potential-energy surface (Figure 1a).

Our work is motivated based on the distance dependent electrostatic multipole expansion of the electrostatic potential Jackson (1999). In the multipole expansion of the electrostatic potential of a localized charge distribution, the degree-ℓ term decays as $r^{-(\ell+1)}$:

$$\Phi(\mathbf{x}) = \frac{1}{4\pi\epsilon_0} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \frac{4\pi}{2\ell+1} q_{\ell m} \frac{Y_{\ell m}(\theta, \phi)}{r^{\ell+1}}, \tag{1}$$

where $Y_{\ell m}$ denotes the spherical harmonics of degree ℓ and order m with polar angle θ and azimuthal angle ϕ, $q_{\ell m}$ denotes the multipole moments of the source charge distribution, r denotes the distance

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.

from the source, and ϵ_0 denotes the vacuum permittivity. Spherical harmonics' based equivariant MLIPs (Passaro & Zitnick, 2023; Fu et al., 2025) build their feature set similarly to the electrostatic multipole expansion, but they are usually not enforced to represent the physical charge moments or decay. The shared angular structure between equivariant MLIPs and the multipole expansion motivates our hypothesis that less angular resolution would suffice between distant atoms. Currently, there is no equivariant MLIP that lowers the maximum retained degree of an edge by its length. Concretely, we hypothesize that: (i) training assigns shorter useful ranges to higher degrees and (ii) the force mean absolute error (MAE) of a model with such ranges stays within the noise of the dense model.

In this work, we introduce range-adaptive $L_{\max}$ (RAL). In RAL, each message-passing block learns one gate radius per nonzero degree (Figure 1a). A radial gate multiplies the degree-ℓ features and becomes exactly zero beyond this radius. A mean-gate penalty encourages smaller gate radii. We also introduce a band mixture-of-experts (Band-MoE, Figure 1c), which groups edges by their maximum retained degree and evaluates each group with the degree-k operator, the edge operator restricted to degrees 0 to k. We combine these methods through staged training (Figure 1d). Our contributions are as follows:

- **RAL.** Each block learns one gate radius per degree; a twice continuously differentiable radial gate sets the degree-ℓ features to zero beyond it, so the maximum retained degree of an edge depends on its length.
- **Band-MoE.** Learned radial ranges define a smooth routing function over reduced-degree operators that share one parameter set. After conversion from the gated model, training resumes with reduced edgewise cost.
- **Reflection-symmetric restriction.** Our contribution is inspired by that the energy difference in parity violating mirror symmetric objects (enantiomers) is negligible compared to MLIP and molecular simulations errors. This chemically inspired symmetry simplification enables us to remove the imaginary part of the edge-frame operator. The resulting equivariant $\mathrm{O}(2)$ message operations use 25.9% fewer parameters and increase the energy and force accuracy on OMol25 and force accuracy on the OMat24 datasets, while simultaneously increasing throughput through faster edge computations (Figure 1b).

2 Related work

Efficiency of MLIPs. Equivariant MLIPs reach lower energy and force errors than non-equivariant MLIPs on common molecular benchmarks but usually have slower training and inference speed (Elhag et al., 2026). In catalyst (Chandrasekhar et al., 2026), pharmaceutical or material discovery (Lee et al., 2026), MLIPs can replace expensive DFT calculations with a cost that is orders of magnitude lower than DFT (Kang et al., 2026). Therefore, a faster MLIP directly expedites these discovery applications. For example, replacing the e3nn operations of MACE by cuEquivariance resulted in significant training time reduction (Benoit, 2025). Faster Clebsch-Gordan tensor products (Bochkarev et al., 2026; Xie et al., 2026), relaxed equivariance (Qu et al., 2026), and architectures without Clebsch-Gordan products (Aykent & Xia, 2025) lower the cost per tensor product, while other approaches reduce the overall cost of the model (Li et al., 2025; Kavanagh et al., 2026). In the same spirit, our proposed RAL lowers the maximum retained degree for each edge.

Angular resolution and operator cost. NequIP, MACE, and EquiformerV2 use features of every degree up to $L_{\max}$ (Batzner et al., 2022; Batatia et al., 2022; Liao et al., 2024). eSCN reduces the cost of angular mixing from $O((L_{\max})^6)$ to $O((L_{\max})^3)$ by rotating each edge into its own frame, where the $\mathrm{SO}(3)$ convolution becomes an $\mathrm{SO}(2)$ convolution (Passaro & Zitnick, 2023), and eSEN builds its blocks on this convolution (Fu et al., 2025). Gaunt tensor products offer another operator construction (Luo et al., 2024). In contrast, RAL retains the $\mathrm{SO}(2)$ convolution and global $L_{\max}$, but uses learned gate radii to reduce the maximum retained degree for each edge (Figure 1a).

Conditional computation. Sparse mixtures of experts allocate computation according to the input (Jacobs et al., 1991a;b; Eigen et al., 2013; Masoudnia & Ebrahimpour, 2014; Shazeer et al., 2017). UMA uses a mixture of linear experts, where system-level routing combines same-shaped expert

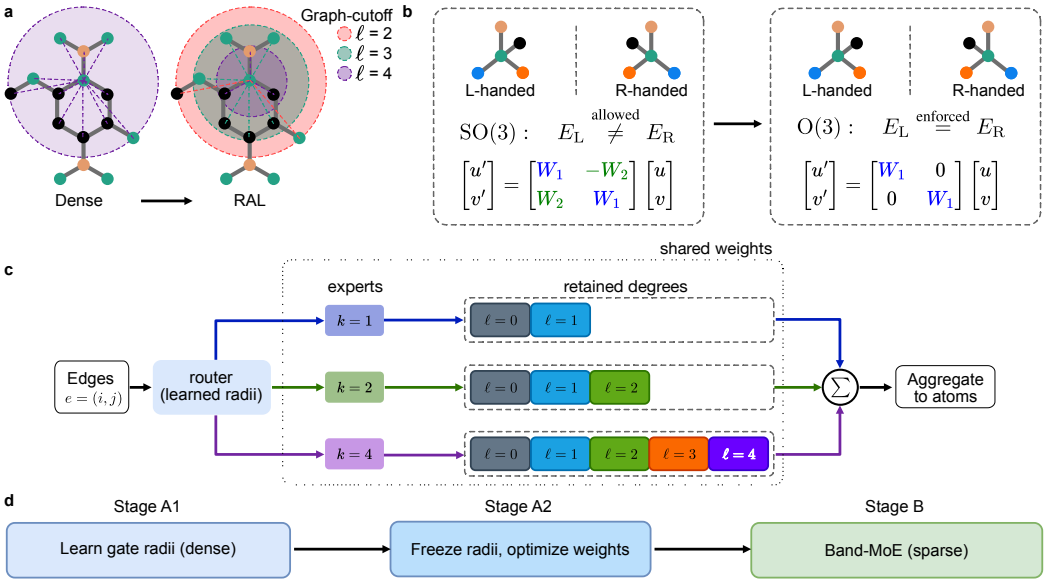


Figure 1: Range-adaptive $L_{\max}$ (RAL). **(a)** Dense $L_{\max}$ against RAL: the dense model evaluates degrees $\ell = 0, \ldots, 4$ on every edge, while RAL assigns each degree a gate radius. **(b)** Reflection-symmetric restriction. **(c)** Band-MoE: the band boundaries assign each edge to one degree-k operator, and the operators use nested slices of one weight set. **(d)** Staged training: Stage A1 learns weights and gate radii with the masked computation, Stage A2 freezes the gate radii, and Stage B trains through Band-MoE.

weights before the forward pass (Wood et al., 2025). Expert routing has been applied to UMA (Liu et al., 2026), to Allegro (de Miranda Nascimento et al., 2026), and to other attention-based MLIPs (Zhou et al., 2026). For a comparison on different mixture of experts in MLIPs applications see Liu et al. (2026). In our proposed Band-MoE, the experts are the degree-k operators of one parameter set, and the router compares the edge length with the learned gate radii, so Band-MoE lowers the computational cost on the edges without adding capacity.

Reflection symmetry in edge frames. A concurrent work derives the real-valued weight constraint for natural-parity edge convolutions and a general local-$\mathrm{O}(2)$ formulation (Xu et al., 2026b;a). Our chemically inspired reflection-symmetric restriction makes the $\mathrm{SO}(2)$ message operations to $\mathrm{O}(2)$-equivariant message passing (reflection symmetry) while still retaining the fast eSCN rotation edge-frame computation (Figure 1b).

3 Method

3.1 Backbone and staged training

We build our RAL (Figure 1c) MLIP on the eSEN-base architecture (Fu et al., 2025), using an adapted version of FAIRChem's UMA backbone (Wood et al., 2025) with the eSCN message passing block (Passaro & Zitnick, 2023). We use a three-stage training pipeline (Figure 1d). Stage A1 (Figure 1d) learns the gate radii and the weights with the masked computation. Stage A2 (Figure 1d) freezes the gate radii and continues training the weights with the masked computation, so that Stage B starts from weights adapted to the frozen radii; Stages A1 and A2 together take 25% of the steps and Stage B the rest. Refer to Appendix E.3 for the fraction of training allocated to Stage A.

The eSCN message-passing uses $\mathrm{SO}(2)$ edge-frame convolution (Passaro & Zitnick, 2023), where atom i carries features $x_i \in \mathbb{R}^{D \times C}$ and $D = (L_{\max} + 1)^2$ is the number of spherical-harmonic coefficients and C is the channel count. For edge (i, j), a block-diagonal Wigner-D matrix rotates the features into the edge frame, where the $\mathrm{SO}(3)$ convolution reduces to an $\mathrm{SO}(2)$ convolution. Two $\mathrm{SO}(2)$ mixing operations, separated by a gated nonlinearity, form the message. Our modifi-

cations operate within this edgewise computation (Figure 6a) and the main difference between the small eSEN and our RAL model is that we use $L_{\max} = 4$ and 8 rather than 2. First, the reflection-symmetric restriction halves the number of $m > 0$ mixing weights (Section 3.2). Second, the radial gates multiply the degree-resolved features after the forward rotation, after each mixing operation, and before the inverse rotation (Section 3.3). Third, Band-MoE (Fig 1c-d) replaces the masked computation in Stage B, after the gate radii are frozen (Section 3.4). RAL, Band-MoE and the reflection symmetry restriction are compatible with both direct force and conservative force prediction training. In direct force prediction, the model has separate output heads for energy and force predictions. In conservative force prediction, the model has a single energy head and force is calculated by differentiating the energy with respect to atomic position by $\mathbf{F}_i = -\partial E/\partial \mathbf{r}_i$; where $\mathbf{r}_i$ are the atomic positions.

An edge with maximum retained degree k has $(k+1)^2$ spherical-harmonic coefficients per channel, and its $\mathrm{SO}(2)$ mixing cost scales as $\sum_{m=0}^{k}(k-m+1)^2C^2$, which leads to the following cost model

$$\mathrm{cost}(k) \;\propto\; (k+1)^3. \tag{2}$$

Refer to Appendix B for the derivation, the runtime attribution, and the masking overhead.

3.2 Reflection symmetry restriction

For each order $m > 0$, the edge-frame operator applies the complex map $W_1 + iW_2$ to the paired real spherical-harmonic components (x_{+m}, x_{-m}); this map is the general linear map that commutes with rotations about the edge axis. The stabilizer of the edge axis in $\mathrm{O}(3)$ is $\mathrm{O}(2)$, which adds the mirrors through planes containing the axis. Such a mirror maps $(x_{+m}, x_{-m}) \mapsto (x_{+m}, -x_{-m})$, and commuting with it forces $W_2 = 0$. We therefore delete W_2 and apply the same real matrix W_1 to both components, which halves the number of $m > 0$ edge weights at an unchanged channel count (Figure 1b). The resulting backbone is $\mathrm{O}(3)$-equivariant with $\mathrm{O}(2)$ edge messages. TECE uses the same restriction, but its architecture is not fully $\mathrm{O}(3)$-equivariant (Xu et al., 2026b) as it retains complex-valued weights in certain $\mathrm{SO}(2)$ layers. While the $\mathrm{SO}(3)$ flexibility might improve in-domain accuracy, it relies on data-driven learning rather than structural constraints to handle reflection. However, mirror-image (chiral pairs) configurations differ in energy only through parity-violating weak interaction, at a scale of approximately 10^{-11} J mol^{-1} for organic molecules (Quack et al., 2022), which is 12 to 15 orders of magnitude below DFT and MLIP energy errors. This chemical domain knowledge inspires us to enforce reflection symmetry for chiral configurations. An $\mathrm{SO}(3)$-equivariant model can nevertheless assign different energies to mirror images. The restriction does not prevent distinguishing diastereomers or a chiral molecule in a chiral environment, since these configurations are not mirror images of each other.

3.3 Radial gates and training objective

We keep $L_{\max}$ as a fixed upper bound and make the contribution of each degree depend on the edge length. In message-passing block b, a radial gate multiplies the degree-ℓ features. We use the quintic smoothstep function (Perlin, 2002), which reaches exactly zero at a trainable gate radius $R_\ell^{(b)}$ over a fixed transition width $w_\ell > 0$. With the block index suppressed, the radial gate $g_\ell(r)$ on an edge of length r is:

$$g_\ell(r) \;=\; \sigma_5(t)\,, \qquad t = \mathrm{clamp}\left(\frac{R_\ell - r}{w_\ell},\, 0,\, 1\right), \qquad \sigma_5(t) = 6t^5 - 15t^4 + 10t^3, \tag{3}$$

where σ_5 denotes the quintic smoothstep polynomial and t the normalized distance. We keep $g_0(r) \equiv 1$. The radial gate equals one for $r \le R_\ell - w_\ell$ and zero for $r \ge R_\ell$, and its first two derivatives vanish at both transition boundaries, so the gate introduces no discontinuity into the forces or their first derivatives. Because the edge length is rotation-invariant and the gate multiplies all components of the degree-ℓ features, radial gating commutes with rotations (Corollary 1, Appendix K). A twice continuously differentiable potential-energy surface also requires the same smoothness of the remaining network operations and of the graph construction, which gate smoothness alone does not establish.

The gate radii are learned independently, without an ordering constraint across degrees (Appendix N). In the final configuration, the effective gate radius of degree ℓ is $\max(R_\ell, r_\ell^{\text{floor}})$ with

the radius floor $r^{\text{floor}} = [6.5, 0, 0, 0]$ Å, and the transition widths are $w = [0.5, 2, 2, 2]$ Å, so the dipole gate equals one throughout the 6 Å graph cutoff. Refer to Appendix E for the ablations of the dipole penalty and the transition widths.

We add a degree-weighted mean-gate penalty to the weighted sum of the energy and force losses $\mathcal{L}_{\text{EF}}$:

$$\mathcal{L} = \mathcal{L}_{\text{EF}} + \gamma(s) \sum_{\ell=1}^{L_{\max}} \lambda_\ell \cdot \frac{1}{|B|} \sum_{b \in B} \mathbb{E}_e \big[g_\ell^{(b)}(r_e) \big], \tag{4}$$

where B denotes the set of message-passing blocks, $\mathbb{E}_e$ the expectation value over the edges of a batch, λ_ℓ the penalty weight of degree ℓ, and $\gamma(s)$ a schedule of the training step s that rises linearly from zero to one over the first 3000 steps of the 400k runs, the first 2000 steps of the 102M run, and the first 1000 steps of the 4M runs (Table 13).

3.4 BAND-MOE

Radial gating saves computation only if the model skips the operations on inactive degrees. Throughout this work, a degree is active on an edge if its gate value is nonzero, and the maximum retained degree k_e of edge e refers to the largest k such that degrees 1 to k are all active. Edges with the same k_e can be evaluated with the degree-k_e operator, which uses the corresponding slices of the full weights. When the active degrees form the set $\{0, \ldots, k_e\}$, masking the inactive degrees after every operation and evaluating the degree-k_e operator give the same outputs and gradients (Proposition 1, Appendix K); this masking-reduction equivalence assumes the degree-diagonal gated nonlinearity of eSEN. If a gate radius increases with degree within a block, the active set is not contiguous, and the degree-k_e operator discards active degrees (Appendix K.1).

Band-MoE deploys the degree-k operators as experts (Figure 1c). For $L_{\max} = 4$, the experts have maximum degrees 1, 2, and 4, and the band boundaries are the gate radii averaged over blocks R_2 and R_3 learned in Stage A1 (Fig 1d); for $L_{\max} = 2$ and $L_{\max} = 8$, the experts have maximum degrees $\{1, 2\}$ and $\{1, 2, 4, 8\}$. A twice continuously differentiable partition of unity with a transition width of 0.15 Å assigns each edge to one expert, or to two experts inside a transition window, so the energy stays twice continuously differentiable in the positions. At the start of Stage B, Band-MoE replaces the radial gates with a smooth partition of unity over the experts. This replacement changes the model output, so Stage B trains through the deployed computation.

4 EXPERIMENTAL SETUP

For OMol25, we train on the first 400k structures of its 4M training split, the 4M split, and the 102M split, and evaluate on its 2.76M validation structures (Levine et al., 2025). For OMat24, we use its 1.01M training structures and validation structures (Barros-Luque et al., 2024). OMol25 contains non-periodic molecules with charge and spin; OMat24 contains periodic inorganic crystals.

As a baseline, we use a custom eSEN-like model with $L_{\max} = m_{max} = 4$ with four layers, using the original eSCN $\mathrm{SO}(2)$ edge convolution Fu et al. (2025). We use the eSEN model as a baseline because it is currently the state-of-the art model on the OMol25 benchmark. We compare this dense $\mathrm{SO}(2)$ eSEN model to our dense $\mathrm{O}(2)$ variant to study the effect of the reflection symmetry. We further analyze the effect of RAL and BandMoE compared directly to the dense $\mathrm{SO}(2)$ eSEN model and our eSEN $\mathrm{O}(2)$ variant as well to study the speed–accuracy tradeoff. We also evaluate pretrained MLIPs from other architectures, trained either solely on OMol25 or on datasets that include OMol25. These models include the original eSEN-sm-d, eSEN-sm-cons, eSEN-md-d Fu et al. (2025); UMA-m-1.1, UMA-s-1.1, UMA-s-1.2.1 Wood et al. (2025); MACE-omol-0-XL, MACE-POLAR-1-M Batatia et al. (2026); and AllScAIP-md-d, AllScAIP-md-cons Qu et al. (2026) pretrained models. We use each models force prediction according to their original setting (d: direct, cons: conservative force head). Within each comparison, the data, seed, step count, hardware, maximum atoms per GPU batch, learning rate, and activation checkpointing are matched. Training uses distributed data parallelism and BF16. Architecture settings, training hardware, and hyperparameters are listed in Appendix L.

We report energy MAE in meV/atom and force MAE in $\mathrm{meV}/\text{Å}$. Throughout this work, the noise thresholds refer to single-run differences exceeding the variation measured across five seeds of the

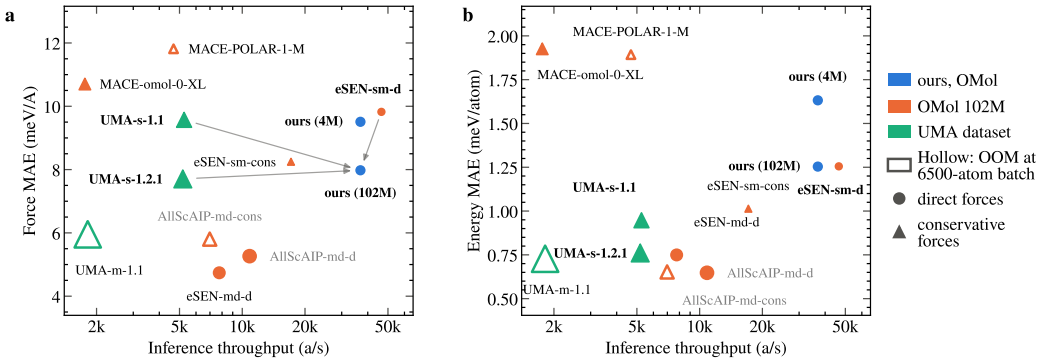


Figure 2: **(a)** Force MAE against inference throughput. **(b)** Energy MAE against inference throughput. Both panels show results of the 12 models on the OMol25 validation split. Arrows shows the RAL improvement compared to similarly sized models. Marker area is proportional to the parameter count. UMA-m-1.1 was evaluated on 1000 atoms per GPU, AllScAIP-md-cons and MACE-POLAR-1-M with a maximum batch size of 4000 atoms (Appendix J).

dense $\mathrm{O}(3)$ model trained on 400k structures: 2.6% for force MAE and 7.9% for energy MAE per atom (Appendix I). Forward inference is timed in BF16 on one NVIDIA H200 using batches of 6500 atoms of validation geometry, unless stated otherwise. Forward speedup is the throughput of RAL divided by that of the dense model with the same GPU and training parameters.

5 Results

Figure 2 shows the force MAE against inference throughput across 12 models. RAL trained on the 102M OMol25 split sits on the speed–accuracy Pareto front. The three equivariant MLIP models that are closest to the scale of our models are UMA-s-1.1, UMA-s-1.2.1, and eSEN-sm-d. RAL beats UMA-s-1.1 on both accuracy and speed and is only slightly less accurate than UMA-s-1.2.1 but nearly 6× faster. eSEN-sm-d is slightly faster than our model but less accurate on forces. Applying RAL on the scale of larger models such as UMA-m-1.1, eSEN-md-d or AllScAIP, could further reduce their inference cost, although the resulting speed-accuracy tradeoff is yet to be evaluated. These large models are deeper and wider than our RAL-trained model; therefore, they have better energy and force MAEs.

Table 1: RAL against dense models on the $4M$ OMol25 subset trained for $200k$ steps and $1M$ OMat24 training split for $60k$ steps. Reflection-symmetric restrictions are denoted with $(\mathrm{O}(3))$ or without $(\mathrm{SO}(3))$. Force MAE in meV/Å, energy MAE in meV/atom, throughput in atoms/s. Δ is relative to the dense model in the same pair. Forward speedup is relative to the dense model.

model	$L_{\max}$	symmetry	force head	throughput	fw. speedup	F MAE	ΔF (%)	E MAE	ΔE(%)
OMol25									
dense baseline	4	SO(3)	direct	22.9k	–	10.276	–	1.749	–
RAL	4	SO(3)	direct	32.9k	1.44×	10.656	+3.7	1.787	+2.2
dense baseline	4	SO(3)	direct	22.9k	–	10.276	–	1.749	–
dense baseline	4	O(3)	direct	29.4k	1.28×	10.250	−0.25	1.727	−1.26
RAL	4	O(3)	direct	48.2k	1.64× to O(3) 2.10× to SO(3)	10.503	+2.47 to O(3) +2.21 to SO(3)	1.763	+2.08 to O(3) +0.8 to SO(3)
dense baseline	2	O(3)	direct	70.0k	–	13.078	–	2.015	–
RAL	2	O(3)	direct	73.5k	1.05×	13.274	+1.50	2.056	+2.05
dense baseline	8	O(3)	direct	6.7k	–	13.825	–	2.241	–
RAL	8	O(3)	direct	15.5k	2.31×	10.084	−27.06	1.832	−18.24
dense baseline	4	O(3)	conservative	11.4k	–	9.792	–	1.671	–
RAL	4	O(3)	conservative	18.7k	1.64×	10.406	+6.27	1.860	+11.31
OMat24									
dense baseline	4	SO(3)	direct	30.1k	–	81.371	–	15.91	–
RAL	4	O(3)	direct	48.9k	1.62×	79.267	−2.59	16.05	+0.9

Table 1 compares the components against the dense baseline. RAL without the reflection-symmetric restriction ran $1.44\times$ faster than the dense baseline at 3.70% higher force MAE. With the reflection symmetry restriction, the throughput rose to $2.10\times$ that of the dense baseline, and the force MAE increase fell to 2.21% (energy MAE $+0.81\%$). On OMat24, RAL with direct force prediction achieved a $1.62\times$ forward speedup, with force MAE decreasing by 2.59% and energy MAE increasing by 0.9%. The speedup over the dense model at the same $L_{\max}$ increased from $1.05\times$ at $L_{\max} = 2$ to $2.31\times$ at $L_{\max} = 8$. Surprisingly, at $L_{\max} = 8$, RAL reached a better accuracy at 27.06% lower force MAE and 18.24% lower energy MAE than the dense $L_{\max} = 8$ model. We hypothesize that restricting higher-degree contributions on longer edges might act as a regularizer, while simultaneously retaining higher $L_{\max}$ on shorter edges. Using the conservative force head, RAL reached a similar forward speedup of $1.64\times$ compared to the direct force head version. However, the conservative RAL scored significantly worse on both energy and force MAE on the OMol25 dataset. Reflection symmetry restriction alone had a minor but steady improvement on the force MAE on both datasets and on energy on OMol25, OMat24 energy is within noise (Table 6). Due to the added symmetry bias, the restriction had a more pronounced improvement in low-data regimes (Table 6).

Table 2: Penalty weights and the resulting band boundaries, force MAE, and forward speedup (fw). OMol25 4M, 40k steps with 5k Stage A1 steps, 8×H100; forward speedup relative to the dense $\mathrm{O}(3)$ model. Accuracy and speedup of RAL with $\lambda_2 = 0.5, \lambda_3 = 0.7, \lambda_4 = 0.9$ differ from Table 1 due to different number of steps and, therefore, different learned radii. Band bound. denotes the band boundaries R_2 and R_3.

model	$\lambda_{2,3,4}$	band bound. (Å)	F MAE (meV/Å)	ΔF (%)	fw. speedup
dense $\mathrm{O}(3)$	–	–	12.880	–	$1.00\times$
RAL	0.25, 0.35, 0.45	4.093, 3.820	12.881	$+0.01$	$1.39\times$
RAL	0.50, 0.70, 0.90	3.754, 3.582	12.983	$+0.80$	$1.45\times$

We further explored the effect of the λ-penalty weights that control how aggressively $L_{\max}$ is reduced in stage A1. Table 2 shows the results of the effect of different λ values. Increasing the penalty weights reduced the band boundaries and raised the forward speedup from $1.39\times$ to $1.45\times$, while the force MAE increase relative to the dense baseline rose from 0.01% to 0.80%. We also ablated the choice of radial boundaries by comparing fixed radii taken from a reference run and an alternative fixed allocation against RAL with learned radii and a uniformly reduced angular bandwidth. Both fixed-radius variants achieve lower energy and force MAEs than the dense but more importantly, above RAL with learned radii (Appendix Table 9).

Because Stage A1 jointly optimizes the model weights and gate radii, we examined how the gate radii evolve across runs with different cutoffs, optimizers and dataset size (Figure 3a and b). Initially, all gate radii were initialized at the 6.0 Å cutoff. Overall, all six experiments resulted in a monotone shell radius reduction with respect to the angular degree (Figure 3a). Interestingly, the active angular fraction decreased rapidly at the start of Stage A1 and had completed most of its decline by step 5000 in every run, while still contracting slowly at the freeze (Figure 3b). We also observed no significant improvement on varying the number of steps during the different stages even with different learning rate scheduling (Appendix E.3). After Stage A1, the gate radii were $[6.2, 3.47, 3.28, 3.03]$ Å on OMol25 and $[6.15, 4.00, 3.76, 3.64]$ Å on OMat24, both on $400k$ with no floor. The gate radii of degrees 2 to 4 decreased with degree in both domains, while the dipole gate radius rose above the cutoff. This ordering agrees with the $r^{-(\ell+1)}$ decay of Equation (1). In the final configuration, the effective dipole gate radius is $\max(R_1, 6.5\,\text{Å})$.

On a sample of 500 molecules from the OMol25 dataset comprising 1.81M edges, the gate radii kept the degree-1 active on every edge and degrees 2, 3, and 4 active on 38%, 34%, and 31% of the edges, respectively. Refer to Appendix N for the per-degree and per-block values and the gate plots (Figure 6).

We also trained RAL on the 102M OMol25 dataset and the results showed a similar trend compared to the $4M$ split (Table 3). The energy and force MAEs were 1.95% and 3.44% higher than those of the dense baseline model. Our results suggest that less angular resolution is needed at longer interatomic distances, consistent with the faster decay of higher-degree terms in the electrostatic

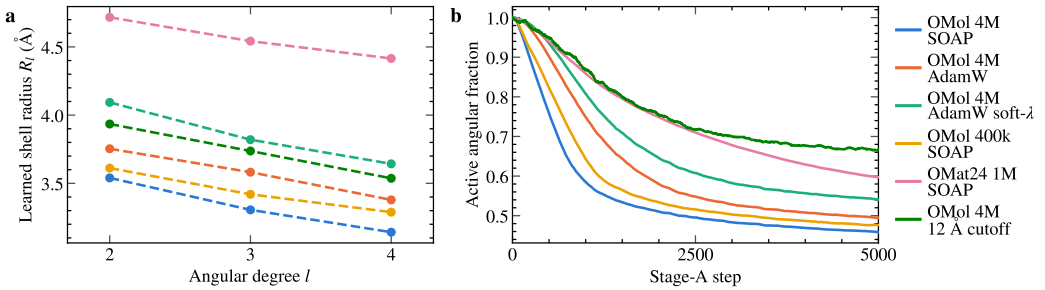


Figure 3: Angular truncation learned during Stage A, for six runs differing in dataset, scale, optimizer and cutoff. **(a)** Average shell radius R_ℓ in Å at the end of Stage A against angular degree ℓ. **(b)** Active angular fraction against Stage-A step. The plotted quantity is the averaged gate value $g_\ell(r_e)$, normalized by its own value at step 0. The AdamW $4M$ OMol25 run was trained for $40k$ steps, $400k$ subset for $30k$ steps, SOAP for $85k$. The run on the 1M OMat24 training split used 60k steps. The radii are frozen at step 5000 in every run.

multipole expansion. While RAL introduces a measurable MAE increase, it offers a systematic approach to improve inference speed significantly while still being fully equivariant.

Table 3: RAL against the dense baseline, both with the reflection-symmetric restriction, on the 102M OMol25 dataset. Δ is relative to the dense baseline. Fw denotes forward.

model	E MAE (meV/atom)	ΔE (%)	F MAE (meV/Å)	ΔF (%)
dense baseline	1.246	–	8.267	–
RAL	1.271	+2.01	8.551	+3.44

We performed zero-shot molecular dynamics simulation on the MD22 dataset's seven molecules (Figures 4 and 5). Figure 4a–d shows the interatomic distance distributions from molecular dynamics simulations of four MD22 molecules using six models and the reference distributions (with the only exception of buckyball-catcher (Figure 5b)). We must highlight that our model was only trained on $4M$ OMol25 datapoints, while the other models were trained on all 102M OMol25 datapoints. RAL produces distributions similar to those of the matched dense baseline, suggesting that angular truncation preserves structural sampling under the tested conditions. Its distributions are also comparable to those obtained with larger models, such as AllScAIP-md-cons and UMA-m-1.1.

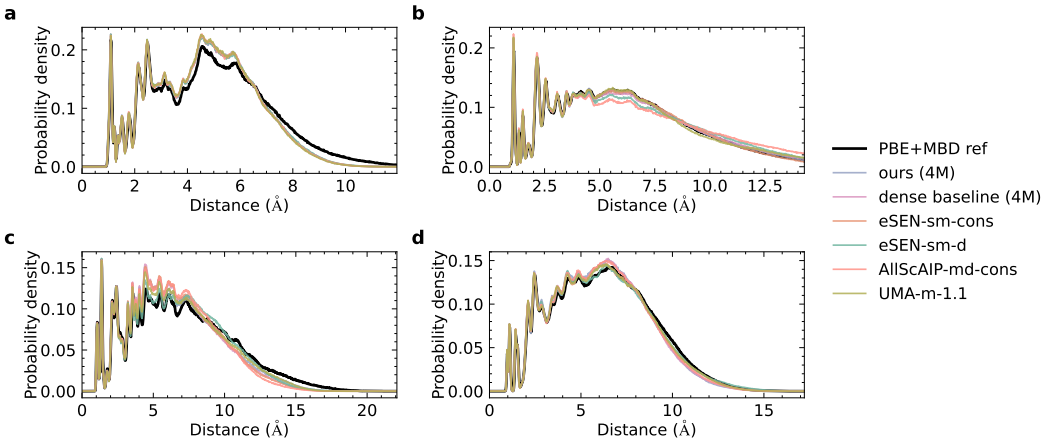


Figure 4: Interatomic distance distributions from zero-shot MD with six MLIPs against the MD22 reference trajectories. **a** Ac-Ala3-NHMe. **b** DHA. **c** AT-AT. **d** Stachyose. The remaining MD22 systems are shown in Appendix H.

6 DISCUSSION

Our results suggest that allocating different angular resolutions to edges can be considered alongside other model parameters, such as depth or channel count. RAL retains higher-degree features on shorter edges and reduces computational cost. The learned radii and λ penalties provide a compromise between energy and force prediction error and computational cost. These learned radii therefore describe the retained angular resolution under the training objective, rather than being physical interaction ranges. Our trained RAL models also have significantly higher inference throughput compared to similarly sized UMA models while having comparable energy and force accuracy. With a higher computational budget, training RAL on deeper and wider models would allow us to further push the speed–accuracy Pareto front of MLIPs.

Physicochemical knowledge also motivates which symmetry operations the model should consider. Enforcing reflection symmetry simplifies learning energy predictions for mirror-image configurations from data. The larger accuracy improvements in the low-data regime support this interpretation. The reflection-symmetry restriction can be used independently of RAL. The MD22 comparisons support the use of RAL in similar settings, but the buckyball-catcher exception requires further investigation. For example, evaluating models trained on the full OMol25 dataset would help determine whether the buckyball-catcher behavior persists.

Therefore, the practical use of RAL depends on the intended application. For example, in large-scale molecular simulations, where current MLIPs are still too expensive or memory-limited, RAL can offer up to $2\times$ speedup with only a minor accuracy loss. Furthermore, RAL is architecture-agnostic and could be applied to other MLIPs as a general approach.

7 CONCLUSION

In this work, we introduced RAL and Band-MoE, which adapt angular resolution to edge length and reduce edgewise computation in equivariant MLIPs. RAL scores higher in terms of speed and accuracy than models of similar size. A separate, chemically motivated reflection-symmetric restriction reduces the model's parameter count by 25.9% while slightly improving speed and energy and force prediction accuracy. We found that the gate radii averaged over blocks decreased with degree on OMol25 and OMat24 under degree-increasing penalty weights. We showed that the penalty weights set the speed–accuracy trade-off. On the 4M OMol25 training split, RAL with the restriction reached $2.10\times$ the batched inference throughput of the dense eSEN baseline at 2.21% higher force MAE, and on the 102M split it reached $1.64\times$ at 3.44% compared to the dense eSEN with the reflection-symmetric restriction. These findings support allocating angular resolution according to distance as an approach to reducing MLIP evaluation cost.

AI USE STATEMENT

Generative AI tools assisted with editing, restructuring, and formatting of this manuscript. Generative AI tools were used to implement code snippets, scripts; and to help with general coding and debugging during development. Generative AI tools were not used to generate ideas, create figures, plan the research, design experiments, evaluate the results. The authors are responsible for the final text, claims, and all supporting artifacts.

ETHICS STATEMENT

This work uses published molecular and materials datasets to study interatomic potentials. It involves no human subjects, no personal data, and no dual-use concern beyond computational chemistry in general.

REPRODUCIBILITY STATEMENT

Section 4 describes the experimental setup; Appendices L, K, and B provide training settings, proofs and gate-support conditions, and the cost model and speed predictor, respectively.

REFERENCES

Gene M. Amdahl. Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, AFIPS '67 (Spring), pp. 483–485, New York, NY, USA, 1967. ACM. doi: 10.1145/1465482.1465560.

Sarp Aykent and Tian Xia. GotenNet: Rethinking Efficient 3D Equivariant Graph Neural Networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Luis Barros-Luque, Muhammed Shuaibi, Xiang Fu, Brandon Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, Matt Uyttendaele, C. Lawrence Zitnick, and Zachary Ulissi. The Open Materials 2024 (OMat24) inorganic materials dataset and models, 10 2024.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gabor Csanyi. 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. Curran Associates, Inc., 2022. doi: 10.52202/068431-0830.

Ilyes Batatia, William J. Baldwin, Domantas Kuryla, Joseph Hart, Elliott Kasoar, Alin M. Elena, Harry Moore, Mikołaj J. Gawkowski, Benjamin X. Shi, Venkat Kapil, Panagiotis Kourtis, Ioan-Bogdan Magdău, and Gábor Csányi. MACE-POLAR-1: A Polarisable Electrostatic Foundation Model for Molecular Chemistry, 2026.

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

Alexandre Benoit. Speeding Up MACE: Low-Precision Tricks for Equivarient Force Fields, 2025.

Anton Bochkarev, Yury Lysogorskiy, and Ralf Drautz. Fast contracted Clebsch–Gordan tensor products for equivariant graph neural networks, 2026.

Achuth Chandrasekhar, Janghoon Ock, and Amir Barati Farimani. Catalyst-Agent: Autonomous heterogeneous catalyst screening with an LLM Agent, 2026.

Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T. Unke, Adil Kabylda, Huziel E. Sauceda, Alexandre Tkatchenko, and Klaus-Robert Müller. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2), January 2023. ISSN 2375-2548. doi: 10.1126/sciadv.adf0873. URL http://dx.doi.org/10.1126/sciadv.adf0873.

Gabriel de Miranda Nascimento, Marc L. Descoteaux, Laura Zichi, Chuin Wei Tan, William C. Witt, Nicola Molinari, Sriteja Mantha, Daniil Kitchaev, Mordechai Kornbluth, Karim Gadelrab, Charles Tuffile, and Boris Kozinsky. Mixture of Experts Framework in Machine Learning Interatomic Potentials for Atomistic Simulations, 2026.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning Factored Representations in a Deep Mixture of Experts, 12 2013.

Ahmed A. Elhag, Arun Raja, Alex Morehead, Samuel M. Blau, Hongtao Zhao, Christian Tyrchan, Eva Nittinger, Garrett M. Morris, and Michael M. Bronstein. Learning Inter-Atomic Potentials without Explicit Equivariance, 2026.

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction. In *Forty-second International Conference on Machine Learning*, 2025.

Mario Geiger and Tess Smidt. e3nn: Euclidean Neural Networks, 07 2022.

Gil Harari, Yoel Zimmermann, Ola Tangen Kulseng, Laura Zichi, Chuin Wei Tan, Marc L. Descoteaux, and Boris Kozinsky. Beyond Adam: SOAP and Muon for Faster, Label-Efficient Training of Machine Learning Interatomic Potentials. *ArXiv*, abs/2607.02499, 2026. URL https://api.semanticscholar.org/CorpusID:289748880.

John David Jackson. *Classical Electrodynamics*. John Wiley & Sons, 3rd edition, 1999. ISBN 978-0471309321.

Robert A. Jacobs, Michael I. Jordan, and Andrew G. Barto. Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, 15 (2):219–250, 1991a. ISSN 0364-0213. doi: 10.1016/0364-0213(91)80006-Q.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87, 1991b. doi: 10.1162/neco.1991.3.1.79.

Hanwen Kang, Tenglong Lu, Sheng Meng, and Miao Liu. Are Machine Learning Interatomic Potentials Truly Practical? A Benchmark of 23 Mainstream Models, 2026.

Seán R. Kavanagh, Chuin Wei Tan, Menghang Wang, Marc L. Descoteaux, Gabriel de Miranda Nascimento, Ulrik Unneberg, Laura Zichi, Francesco Libbi, Norma Rivano, Austin Glover, Vivek Bharadwaj, Anders Johansson, William C. Witt, Albert Musaelian, and Boris Kozinsky. Fast and Accurate Equivariant Foundation Models for Atomistic Simulation, 2026.

Jaewoong Lee, Taeun Bae, and Jihan Kim. SimMOF: AI agent for Automated MOF Simulations, 2026.

Daniel Levine, Muhammed Shuaibi, Evan Spotte-Smith, Michael Taylor, Muhammad Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan Frey, Xiang Fu, Vahe Gharakhanyan, Aditi Krishnapriyan, Joshua Rackers, Sanjeev Raja, Ammar Rizvi, Andrew Rosen, Zachary Ulissi, and Brandon Wood. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models, 05 2025.

Wenwen Li, Nontawat Charoenphakdee, Yong-Bin Zhuang, Ryuhei Okuno, Yuta Tsuboi, So Takamoto, Junichi Ishida, and Ju Li. LightPFP: A Lightweight Route to Ab Initio Accuracy at Scale, 2025.

Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Yuzhi Liu, Duo Zhang, Anyang Peng, Weinan E, Linfeng Zhang, and Han Wang. Mixture of experts architectures for machine learning interatomic potentials. *npj Artificial Intelligence*, August 2026. ISSN 3005-1460. doi: 10.1038/s44387-026-00144-9.

Shengjie Luo, Tianlang Chen, and Aditi S. Krishnapriyan. Enabling Efficient Equivariant Operations in the Fourier Basis via Gaunt Tensor Products. In *The Twelfth International Conference on Learning Representations*, 2024.

Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: A literature survey. *Artificial Intelligence Review*, 42, 08 2014. doi: 10.1007/s10462-012-9338-y.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023. doi: 10.1038/s41467-023-36329-y.

Saro Passaro and C. Lawrence Zitnick. Reducing SO(3) convolutions to SO(2) for efficient equivariant GNNs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202. PMLR, 2023.

Ken Perlin. Improving noise. *ACM Trans. Graph.*, 21(3):681–682, July 2002. ISSN 0730-0301. doi: 10.1145/566654.566636.

Eric Qu, Brandon Wood, Aditi Krishnapriyan, and Zachary Ulissi. A recipe for scalable attention-based MLIPs: unlocking long-range accuracy with all-to-all node attention, 03 2026.

Martin Quack, Georg Seyfang, and Gunther Wichmann. Perspectives on parity violation in chiral molecules: theory, spectroscopic experiment and biomolecular homochirality. *Chemical Science*, 13(36):10598–10643, 2022. ISSN 2041–6539. doi: 10.1039/d2sc01323a.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*, 2017.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds, 02 2018.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and Stabilizing Shampoo using Adam for Language Modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

Brandon Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John Kitchin, Daniel Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Sahoo, Zachary Ulissi, and C. Lawrence Zitnick. Uma: A family of universal models for atoms. In *Advances in Neural Information Processing Systems*, 06 2025. doi: 10.48550/arXiv.2506.23971.

YuQing Xie, Ameya Daigavane, Mit Kotak, and Tess Smidt. Asymptotically Fast Clebsch-Gordan Tensor Products with Vector Spherical Harmonics, 2026.

Zemin Xu, Peijun Hu, and Wenbo Xie. Complete O(3) interactions from Wigner-6j recoupling to local O(2) frames, 2026a. arXiv:2608.16592.

Zemin Xu, Wenbo Xie, and P. Hu. Edge cluster expansion with radial rotary attention for interatomic potentials, 2026b. arXiv:2607.10664.

Yuanchang Zhou, Hongyu Wang, Yiming Du, Yan Wang, Mingzhen Li, Siyu Hu, Xiangyu Zhang, Weijian Liu, Chen Wang, Zhuoqiang Guo, Long Wang, Jingde Bu, Yutong Lu, Guangming Tan, and Weile Jia. Breaking the Training Barrier of Billion-Parameter Universal Machine Learning Interatomic Potentials, 2026.

Table 4: Six 400k runs, 30k steps each, 4×H200, no activation checkpointing, one run per row. All six use the $\mathrm{SO}(3)$ backbone, and the RAL rows execute the degree-k operators through per-edge bucketing rather than Band-MoE; $\lambda_1 = 0$ with the radius floor $r_1^{\text{floor}} = 6.5\,\text{Å}$. ΔE and ΔF are relative to the dense baseline of the same domain. Inference measurement: forward is one forward pass; gradient-serving is one forward and one backward pass. F MAE is in (meV/Å) and E MAE is in (meV/atom). Accuracy was measured on $10k$ slice of official validation sets for both domains. Fw and bckw denote forward and backward respectively.

domain	model	$\lambda_{2,3,4}$	E MAE	ΔE (%)	F MAE	ΔF (%)	forward	fw + bckw
OMat24	dense baseline	–	16.667	–	93.062	–	1.00×	1.00×
	RAL	0.5, 0.7, 0.9	16.652	−0.09	94.695	+1.75	1.45×	1.36×
	RAL	1.0, 1.4, 1.8	16.871	+1.22	95.550	+2.67	1.58×	1.49×
OMol25	dense baseline	–	3.026	–	16.494	–	1.00×	1.00×
	RAL	0.5, 0.7, 0.9	2.870	−5.16	17.030	+3.25	1.43×	1.42×
	RAL	1.0, 1.4, 1.8	3.035	+0.30	17.147	+3.96	1.52×	1.50×

A Results for 400k OMol25 subset

Table 4 lists the penalty weights, energy and force MAE, and speedups of the six 400k runs that Section 4 describes.

B Computational cost and speedup

At maximum retained degree k, the $\mathrm{SO}(2)$ mixing work of an edge is $\sum_{m=0}^{k}(k-m+1)^2C^2 \approx \frac{1}{3}(k+1)^3C^2$, because the order-$m$ block acts on $k-m+1$ degrees, while the two Wigner rotations cost $(k+1)^4C$. The ratio of the two terms is $C/(3(k+1))$, which is 8.5 at $C = 128$ and $k = 4$, so the mixing term gives Equation (2). Under this cube model, reducing an edge from $k = 4$ to $k = 2$ lowers its cost $4.63\times$, while the coefficient count ($25 \to 9$) would suggest $2.78\times$.

B.1 Runtime breakdown and gating overhead

We estimated the share of runtime spent on edgewise computation by setting the edgewise output of every block to zero. The estimation attributed 88% of the training-step time at 4000 atoms and 89.5% of the forward-pass time at 6500 atoms to edgewise computation. Both measurements used the $\mathrm{SO}(3)$ backbone in BF16 with activation checkpointing disabled. Each batch contained one non-periodic system of uniformly random atoms in a cube at a number density of $0.033\,\text{Å}^{-3}$, with a 6 Å cutoff. Profiling the same script and batch attributed 87.4–87.8% of the forward-pass time to edgewise computation, providing a cross-check of the ablation.

The synthetic geometry had 26.4 edges per atom, compared with 35.5–36.0 in OMol25 validation batches. With an edgewise share of 0.895, retaining a fraction s of the edgewise cost bounds the total speedup by $1/(0.895\,s + 0.105)$, assuming the remaining cost is unchanged. At $s = 0.4$, this bound is approximately $2.2\times$.

The masked computation evaluates the full expansion before applying the gates. Gating increased the forward-pass time relative to the dense baseline by 9–11% for batches of 1000–6500 atoms and by 6% for smaller batches. The degree-k operators incorporate two of the four gate multiplications into the Wigner matrices using $\mathrm{diag}(g)(Dx) = (\mathrm{diag}(g)D)x$. This change is bit-exact and recovers 3.7% of the forward-pass time, leaving approximately 7% overhead. The degree-k operators must therefore run more than $1.07\times$ faster than the dense computation to offset this overhead.

Table 5: Fraction of angular computation retained and realized fraction for the four RAL runs of Table 4. S_{hard} is computed from the checkpoint radii and validation edge lengths without a forward pass. The end-to-end ceiling is Equation (6) at $\rho = 1$; ρ is calculated from the measured forward speedup of Table 4. Edgewise ceiling and end-to-end ceiling refer to the upper bound on edgewise speedup and upper bound on total speedup, respectively.

domain	$\lambda_{2,3,4}$	S_{hard}	edgewise ceiling $1/S_{\mathrm{hard}}$	end-to-end ceiling	ρ
OMat24	0.5, 0.7, 0.9	0.395	2.53×	2.18×	0.381
OMat24	1.0, 1.4, 1.8	0.319	3.13×	2.56×	0.372
OMol25	0.5, 0.7, 0.9	0.410	2.44×	2.12×	0.384
OMol25	1.0, 1.4, 1.8	0.359	2.79×	2.35×	0.386

B.2 Retained computation and speedup prediction

We estimate the fraction of angular computation retained by a checkpoint from its gate radii and a sample of edge lengths. Relative to evaluating every edge at $L_{\max}$, this fraction is:

$$S_{\mathrm{hard}} \;=\; \mathbb{E}_e\left[\frac{(k_e+1)^3}{(L_{\max}+1)^3}\right], \tag{5}$$

where k_e denotes the maximum retained degree of edge e, the largest k with $r_e < R_\ell$ for every $\ell \le k$ (equivalently $g_\ell(r_e) > 0$ for every $\ell \le k$; $g_0 \equiv 1$). S_{hard} needs no forward pass and no additional training, and at fixed radii it does not depend on the transition widths. The mean-gate penalty in Equation (4) does not directly measure the computation skipped. Narrowing the degree-1 transition increased the summed mean gate value $S = \sum_\ell \mathbb{E}_e[g_\ell]$ from 1.21 to 1.37 without changing the measured speed. Moreover, the same value of $S = 1.375$ in the two domains corresponded to speeds differing by $1.5\times$.

The reduction applies only to edgewise computation, and the degree-k operators introduce gather, scatter, and concatenation costs. We account for the unchanged computation using an Amdahl model (Amdahl, 1967) and fit one factor to the fraction of the predicted gain realized:

$$\text{speedup} \;=\; 1 \;+\; \rho\left(\frac{1}{f\,S_{\mathrm{hard}} + (1-f)} - 1\right), \qquad f = 0.895, \tag{6}$$

where ρ denotes the realized fraction of the predicted speedup above one and f the edgewise share of a forward pass at 6500 atoms. The model estimates speed after the radii have been learned; it does not estimate accuracy. Table 5 lists S_{hard}, the two ceilings, and ρ for the four RAL runs of Table 4.

B.3 Speedup model calibration

Across the two domains and a $1.3\times$ range of S_{hard}, ρ varied by 4% (Table 5). The pooled value $\rho = 0.379$ reproduced all four measured forward speedups within 1.0%, and ρ calibrated on one run predicted the other three within 0.8% (calibration on OMat24, $\lambda_{2,3,4} = 0.5, 0.7, 0.9$) to 1.7% (calibration on OMat24, $\lambda_{2,3,4} = 1.0, 1.4, 1.8$). The functional form matters: calibrated on one run and evaluated on another, speedup $\propto 1/S_{\mathrm{hard}}$ erred by -23%, the model without the Amdahl term by -1.1%, and Equation (6) by $+0.1\%$. Given a checkpoint, the estimate takes seconds on a CPU. These four runs are the $\mathrm{SO}(3)$ runs of Table 4, which execute the degree-k operators without Band-MoE; for the Band-MoE path, ρ back-solved on three Band-MoE runs on measured OMol25 geometry was 0.40 to 0.44 (mean 0.425), so ρ must be refit for each execution path.

C Speed–accuracy tradeoff

Per 1% of force MAE increase, the forward speedup above one in Table 4 was 0.26 and 0.22 on OMat24 and 0.13 at both penalty settings on OMol25, so OMat24 gave 1.7 to 2.0 times more speedup per unit of force MAE than OMol25. Neither run of either domain reached the force MAE of its dense baseline. A linear fit of the force MAE increase against S_{hard} through the two points of each domain, propagated through Equation (6), extrapolates to a speedup of 1.2 to $1.3\times$ (OMat24)

Table 6: Effect of the reflection-symmetric restriction: **(a)** compares the dense $\mathrm{O}(3)$ model with the corresponding dense baseline; each pair shares all training settings except the restriction, and each entry is a single run. ΔE and ΔF are relative to the dense baseline. **(b)** tests whether the gain can be explained by reduced parameter count alone. The phase-locked control uses $W_1 = \cos(\pi/4)\, W$ and $W_2 = \sin(\pi/4)\, W$.

(a) Dense $\mathrm{O}(3)$ restriction across datasets, training scales, and force heads.

			E MAE (meV/atom)			F MAE (meV/Å)		
domain	force head	steps	dense baseline	dense $\mathrm{O}(3)$	ΔE (%)	dense baseline	dense $\mathrm{O}(3)$	ΔF (%)
OMol25 40k	direct	30k	8.043	7.607	-5.4	32.74	29.40	-10.2
OMol25 40k	cons.	40k	6.440	6.084	-5.5	26.58	24.65	-7.3
OMol25 4M	direct	240k	1.393	1.340	-3.8	8.72	8.63	-1.0
OMat24 1M	direct	60k	13.18	13.28	$+0.7$	69.00	66.58	-3.5
OMat24 1M	cons.	60k	12.50	12.39	-0.9	62.98	62.06	-1.5

(b) Capacity controls on OMol25 40k with a direct force head and 30k training steps.

capacity control	E MAE (meV/atom)	F MAE (meV/Å)	parameters
$\mathrm{SO}(3)$, $C = 128$ (dense baseline)	8.043	32.74	22.78M
$\mathrm{SO}(3)$, $C = 110$ (width-matched)	8.088	33.23	16.89M
$\mathrm{SO}(3)$, phase-locked	8.548	34.99	16.88M
$\mathrm{O}(3)$, reflection-symmetric restriction	7.607	29.40	16.88M

and 1.15 to 1.2× (OMol25) at zero force MAE increase. This extrapolation lies outside the measured range and rests on two points per domain.

D Reflection symmetry and combined results

Table 6 compares the dense baseline with the dense $\mathrm{O}(3)$ model, which differs from it only in the reflection-symmetric restriction (Section 3.2), in five pairs of runs, and lists three capacity controls at OMol25 40k.

The restriction alone lowered the force MAE by 1.0 to 10.2% in the five comparisons and the energy MAE in four of the five (Table 6); the width-matched and phase-locked $\mathrm{SO}(3)$ controls did not reproduce the decrease, so the parameter count does not explain it. Combining RAL, Band-MoE, the restriction, and the SOAP optimizer gave 25.9% fewer parameters, 50% lower training memory, 38% higher Stage B training throughput, 25.4% lower force MAE, and 38.4% lower energy MAE than the dense baseline trained with AdamW (direct force head, 6500 atoms per GPU batch). Most of this accuracy change comes from the optimizer, because SOAP alone lowered the force MAE of the dense 400k model by 25.6% (Appendix F).

E Ablations

E.1 Dipole penalty

An initial penalty proportional to the coefficient count, $\lambda_\ell = \lambda\,(2\ell + 1)$, moved R_1 below the graph cutoff, although degree 1 costs 8 of the 125 units of the full expansion under Equation (2). Table 7 pairs runs with identical $\lambda_{2,3,4}$ that differ only in λ_1. Lowering λ_1 to 0.001 left $R_{2,3,4}$ unchanged and reduced the force MAE increase from +5.6% to +1.9% and from +10.0% to +3.0%, so λ_1 controls R_1 and $\lambda_{2,3,4}$ control $R_{2,3,4}$. The runs with the larger λ_1 were faster because an edge beyond R_1 keeps only degree 0, that is, 1 spherical-harmonic coefficient instead of the 4 of degrees 0 and 1; this speedup cost 3.7 and 7.0 percentage points of force MAE. The final configuration therefore sets the radius floor $r_1^{\text{floor}} = 6.5$ Å, which keeps the dipole gate at one on every edge within the cutoff.

Table 7: Dipole penalty ablation on OMol25 400k, one run per row. Each pair shares $\lambda_{2,3,4}$ and differs only in λ_1; $R_{2,3,4}$ differ by at most 0.53% within a pair.

λ_1	$\lambda_{2,3,4}$	R_1 (Å)	$R_{2,3,4}$ (Å)	ΔF vs. dense baseline (%)	relative speed
0.15	0.25, 0.35, 0.45	5.60	4.23, 4.01, 3.79	+5.6	1.40×
0.001	0.25, 0.35, 0.45	6.43	4.24, 4.00, 3.80	+1.9	1.36×
0.45	0.75, 1.05, 1.35	4.97	3.74, 3.55, 3.36	+10.0	1.60×
0.001	0.75, 1.05, 1.35	6.46	3.76, 3.56, 3.37	+3.0	1.49×

Table 8: Stage A share and learning-rate (LR) schedule ablation on OMol25 4M, 200k steps, one run per row. Stage B start LR vs. single cosine compares the learning rate at the start of Stage B with that of single uninterrupted cosine schedule at the same step; LR sum is the learning rate summed over the run, relative to the same single-cosine schedule. F MAE is reported in (meV/Å), E MAE in (meV/atom). ΔF is relative to the best row.

Stage A share (%)	Stage B peak LR (× base)	Stage B start LR vs. cosine (%)	LR sum vs. single cosine (%)	E MAE	F MAE	ΔF vs. best (%)
20	0.6	−33.7	−13.3	1.7842	10.467	+1.21
20	0.8	−11.6	+2.5	1.7783	10.369	+0.26
30	0.6	−24.6	−2.5	1.7798	10.416	+0.72
30	0.8	+0.5	+11.4	1.7609	10.342	0.00
50	0.6	+18.8	+11.4	1.7471	10.388	+0.44
50	0.8	+58.4	+21.3	1.7451	10.352	+0.10

E.2 Transition width

The transition width w_ℓ sets which edges give R_ℓ a gradient, because $\partial g_\ell / \partial R_\ell \neq 0$ only for $R_\ell - w_\ell < r < R_\ell$, and it sets how strongly the active degrees are attenuated. At fixed radii, it does not change which degrees are active, because that condition is $r < R_\ell$. With $w_1 = 0.5$ and 1.0 Å, the force MAE differed by 0.02 percentage points (two seeds), and both settings had a lower force MAE than $w_1 = 2.0$ Å at equal inference speed. For $\ell \geq 2$, $R_\ell = 3$ to 3.5 Å and $w_\ell = 2$ Å place $R_\ell - w_\ell$ at 1 to 1.5 Å, so the gate equals one only for the shortest edges and most active edges at these degrees are attenuated. We fixed $w_{1,2,3,4} = 0.5, 2, 2, 2$ Å throughout.

E.3 Training-stage duration and learning-rate scheduling

Across Stage A shares of 20 to 50% and Stage B peak learning-rate factors of 0.6 and 0.8, the force MAE varied by at most 1.21%, below the 2.6% noise threshold (Table 8; OMol25 4M, 200k steps, 8×GH200, full validation split).

E.4 Angular-allocation

F Optimizers

Table 1 uses AdamW, as the UMA configuration does (Wood et al., 2025). SOAP (Vyas et al., 2025) lowered the force error of NequIP and Allegro (Harari et al., 2026), and we tested whether this transfers to the eSEN backbone.

On OMol25 400k, replacing AdamW with SOAP in a matched grid lowered the force MAE by 25.6% for the dense model and by 23.4% for RAL ($\lambda_{2,3,4} = 0.5, 0.7, 0.9$). The force MAE increase from the same truncation therefore grew from +3.56% under AdamW to +6.73% under SOAP (1.9×); the +3.56% is the AdamW reference run repeated in this comparison, while the separate anchor of Table 4 gives +3.25%. The forward speedup above one per 1% of force MAE increase fell from 0.12 to 0.073. SOAP also lowered S_{hard} at fixed λ from 0.405 to 0.365; the 0.405 is the same re-run anchor, and the separate checkpoint of Table 5 gives 0.410. Table 10 repeats the comparison on OMol25 4M.

Table 9: Angular-allocation ablations on the full OMol25 validation split after training on the 4M split. All models use direct force prediction and the reflection-symmetric $\mathrm{O}(2)$ edge operator. Fixed-radius rows specify the band boundaries (R_2, R_3) in Å. All three RAL variants use Band-MoE with maximum expert degrees $\{1, 2, 4\}$. Their staged training schedule differs from that of the main RAL run.

Model	Band boundaries (Å)	$L_{\max}$	E MAE (meV/atom)	F MAE (meV/Å)
Dense baseline	–	4	1.749	10.276
RAL	learned	4	1.763	10.503
Dense baseline	–	3	1.7981	11.102
Fixed	$(3.4096, 3.1598)$	4	1.7551	10.564
Fixed	$(5.00, 2.64)$	4	1.7420	10.667

Table 10: Optimizer comparison on OMol25 4M. Force MAE in meV/Å on the full validation split (2.76M structures), EMA weights, evaluated in bf16 autocast. Column headings give the learning rate (LR) and the step budget; the two columns use different budgets and are not comparable with each other. ΔF is relative to the dense $\mathrm{O}(3)$ model trained with the same optimizer and budget.

model	AdamW, LR 1.13×10^{-3}, 40k steps	SOAP, LR 3.39×10^{-3}, 85k steps
dense $\mathrm{O}(3)$ model	12.880	9.891
RAL, $\lambda_{2,3,4}$=0.5, 0.7, 0.9	12.983	10.192
ΔF vs. dense $\mathrm{O}(3)$ model (%)	+0.80	+3.04

SOAP's optimizer state was 824 MB at 16.9M parameters, 12.8 times the model weights and 6.4 times the 129 MB state of AdamW. The four 3840×128 radial-MLP output layers held 56% of it, 113 MB each for the 3840×3840 left preconditioner and its eigenbasis; limiting the preconditioned dimension to below 3840 reduces the state to 373 MB. At 6500 atoms per GPU batch, the state was 1.7% of the 21 GB peak allocation of the dense $\mathrm{O}(3)$ model trained with activation checkpointing and 0.9% of the 42 GB peak of the RAL (phase B) runs trained without it.

G Conservative force head

Section 5 uses the direct force head, except for the conservative row of Table 1. With the conservative force head, the forces are the derivatives of the energy through the radial gates. In FP64, the double backward pass runs in Stage A and in Stage B, finite-difference forces agreed with the analytic forces to 8×10^{-10} in Stage A and to 10^{-9} in Stage B, and the two stages gave identical autograd forces.

In one OMat24 comparison without activation checkpointing in either model, the conservative RAL model had 1.1% higher force MAE and 3.0% lower energy MAE than the dense baseline. This comparison is one run on the OMat24 40k subset over 40k steps at 2500 atoms per batch on a single GPU, a smaller scale than the runs of Table 1. We speculate that a direct force head might compensate for a truncated energy surface, while a conservative force head differentiates that surface. On OMol25 4M, however, the conservative force head cost 6.27% force MAE and 11.31% energy MAE per atom (Table 1), which contradicts a shift of the cost from energy to forces.

H Molecular Dynamics

Following Qu et al. (2026); Chmiela et al. (2023), we assess molecular-dynamics predictions under distribution shift using zero-shot NVT simulations of seven MD22 molecules. Figure 5 shows the results for the three molecules not included in Figure 4. The one exception is the buckyball catcher (Figure 5) RAL releases the C_{60} guest within the 50 ps trajectory, with a host–guest centre-of-mass distance of 9.7 ± 0.6 Å against 5.3 ± 0.2 Å in the reference and 5.3 ± 0.3 Å for the matched dense model, which keeps the guest bound. This failure follows from the learned band boundaries: 96% of the cross-fragment edges of this system are longer than the outer boundary of 3.54 Å, so the

dispersion-bound host–guest contact is evaluated at dipole resolution only. Because the boundaries were learned on OMol25, where such contacts are rare, the penalty weights or the boundaries should be re-validated before RAL is applied to non-covalent complexes.

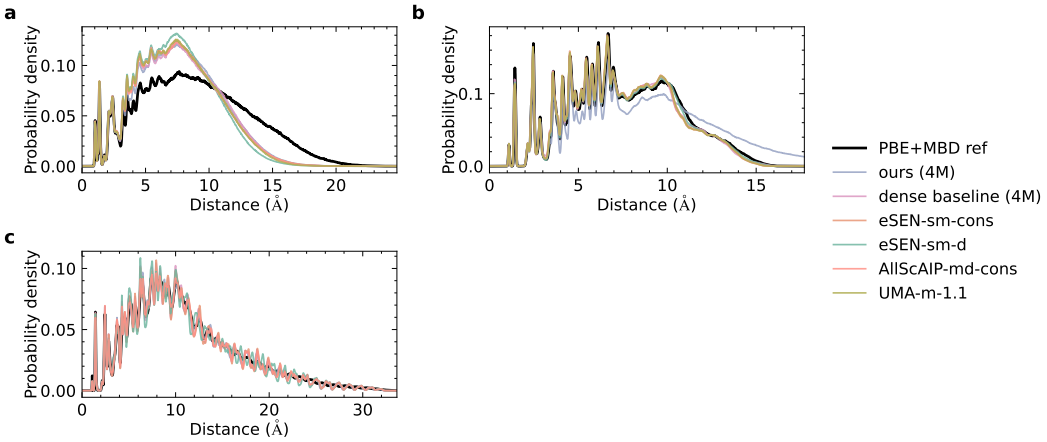


Figure 5: **Zero-shot MD on MD22: radial distribution h(r)**. **(a)** AT-AT-CG-CG; **(b)** buckyball-catcher, **(c)** double-walled nanotube.

I VARIATION ACROSS RANDOM SEEDS

To measure the noise, we trained the dense $\mathrm{O}(3)$ model five times on the first 400k structures of the 4M OMol25 training split, changing only the random seed for parameter initialization and data order. All other settings, including 30k steps on $4\times$H100, 6500 atoms per GPU batch, the SOAP optimizer (Vyas et al., 2025) with learning rate 2.4×10^{-3} and weight decay 10^{-3}, a cosine learning-rate schedule, BF16, and equal loss weights on the energy per atom and on the forces, were kept constant. We evaluated the EMA weights (decay 0.999) in FP32 on the same $500{,}000$ structures, drawn uniformly without replacement from the OMol25 validation split (2.76M structures). We repeated the same five-seed analysis with AdamW to test whether the observed run-to-run variation was specific to SOAP.

Table 11: Run-to-run variation across five random seeds for SOAP and AdamW. The 95% threshold is $t_{0.975,4}\sqrt{2}\,\mathrm{CV}$, the approximate single-run relative difference required to exceed the observed seed-to-seed variation.

optimizer	metric	mean ± sd	CV	range / mean	95% threshold
SOAP	F MAE (meV/Å)	13.20 ± 0.09	0.66%	1.55%	2.6%
	E/atom MAE (meV)	2.94 ± 0.06	2.01%	4.26%	7.9%
	E/structure MAE (meV)	178.3 ± 4.9	2.73%	6.51%	10.7%
AdamW	F MAE (meV/Å)	17.556 ± 0.111	0.63%	1.68%	2.5%
	E/atom MAE (meV)	4.231 ± 0.068	1.60%	4.31%	6.3%
	E/structure MAE (meV)	264.32 ± 4.15	1.57%	3.57%	6.2%
	median E/atom (meV)	0.876 ± 0.014	1.64%	3.71%	6.4%

With SOAP, force MAE was $13.20 \pm 0.09\,\mathrm{meV/Å}$ (CV 0.66%), while energy MAE was $2.94 \pm 0.06\,\mathrm{meV/atom}$ (CV 2.01%). The corresponding 95% single-run detection thresholds were 2.6% for force and 7.9% for energy. AdamW showed the same pattern: force MAE was $17.556 \pm 0.111\,\mathrm{meV/Å}$ (CV 0.63%) and energy MAE was $4.231 \pm 0.068\,\mathrm{meV/atom}$ (CV 1.60%), giving thresholds of 2.5% and 6.3%, respectively. Validation sampling explains only a small fraction of this variance, so the spread is dominated by training noise.

Table 12: Inference throughput (atoms/s) versus atoms per batch. OOM: out of memory at that batch size; "–": not run due to OOM. Measured on $1\times$ NVIDIA H100 (80GB) with a fp32 precision on a OMol25 validation systems packed into one batch. $4k$ and $8k$ values differ from Table 1 as measured on different hardware.

		atoms per batch							
model	parameters	1k	2k	4k	8k	16k	32k	64k	128k
RAL + Band-MoE (OMol25 4M)	17M	24,683	27,557	32,074	34,869	35,874	32,066	OOM	–
eSEN-sm-direct (Fu et al., 2025)	6M	30,519	36,340	41,966	43,962	45,672	43,603	46,130	OOM
eSEN-sm-cons	6M	12,742	13,557	15,732	OOM	–	–	–	–
eSEN-md-direct	51M	6,555	6,996	7,387	7,401	7,366	OOM	–	–
AllScAIP-md-direct (Qu et al., 2026)	87M	14,134	13,903	11,598	OOM	–	–	–	–
AllScAIP-md-cons	86M	5,423	5,537	OOM	–	–	–	–	–
UMA-s-1.1 (Wood et al., 2025)	147M	8,269	6,214	5,480	OOM	–	–	–	–
UMA-s-1.2.1	291M	7,959	6,841	5,388	OOM	–	–	–	–
UMA-m-1.1	1397M	1,604	OOM	–	–	–	–	–	–
MACE-OMol-0-XL (Batatia et al., 2022)	52M	911	1,740	2,620	2,834	OOM	–	–	–
MACE-POLAR-1-S (Batatia et al., 2026)	8M	1,773	10,785	11,324	10,149	OOM	–	–	–
MACE-POLAR-1-M	15M	1,600	5,721	6,035	OOM	–	–	–	–

Force MAE is therefore the primary comparison metric in this work. Across both optimizers, single-run force differences below roughly 2.5% and energy differences below roughly 6–8% should be interpreted cautiously.

J Inference throughput

Table 12 lists the inference throughput of RAL and of published models for batches of 1000 to $128{,}000$ atoms, and the batch size at which each model ran out of memory.

K Proofs

Let $L_{\max}$ denote the global maximum degree and C the number of channels per degree. The feature space is $V \coloneqq \bigoplus_{\ell=0}^{L_{\max}} \mathbb{R}^{(2\ell+1)\times C}$, and a feature $x = \bigoplus_\ell x_\ell \in V$ is decomposed into its degree-ℓ blocks x_ℓ. The atomic positions are $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_N) \in (\mathbb{R}^3)^N$, the joint space is $X \coloneqq V \times (\mathbb{R}^3)^N$, and $r_{ij} \coloneqq \|\mathbf{r}_i - \mathbf{r}_j\|$ denotes the length of the edge (i, j). A rotation $Q \in \mathrm{SO}(3)$ acts on V through the degree-ℓ Wigner matrices $D^{(\ell)}(Q) \in \mathbb{R}^{(2\ell+1)\times(2\ell+1)}$ in the real spherical-harmonic basis, $Q \cdot x \coloneqq \bigoplus_\ell D^{(\ell)}(Q)\, x_\ell$, which are orthogonal, $D^{(\ell)}(Q)^{-1} = D^{(\ell)}(Q)^{\mathrm{T}}$ (Passaro & Zitnick, 2023); it acts on positions as $Q \cdot \mathbf{r}_n \coloneqq Q\mathbf{r}_n$ and on X as $Q \cdot (x, \mathbf{r}) \coloneqq (Q \cdot x, (Q\mathbf{r}_1, \dots, Q\mathbf{r}_N))$. For each degree ℓ, a gate is a scalar function $c_\ell : X \to \mathbb{R}$, and the gated map $G : X \to V$ is

$$G(x, \mathbf{r}) \coloneqq \bigoplus_\ell c_\ell(x, \mathbf{r})\, x_\ell. \tag{7}$$

The statements below also hold for channel-wise gates $c_\ell : X \to \mathbb{R}^C$, with the scalar product replaced by right multiplication by $\mathrm{Diag}(c_\ell(x, \mathbf{r}))$, because left multiplication by $D^{(\ell)}(Q)$ acts on the degree axis of x_ℓ and right multiplication by a diagonal matrix on its channel axis.

Lemma 1 (Equivariance)**.** *The gated map G of Equation* (7) *is* $\mathrm{SO}(3)$*-equivariant, $G(Q \cdot (x, \mathbf{r})) = Q \cdot G(x, \mathbf{r})$ for all $Q \in \mathrm{SO}(3)$ and all $(x, \mathbf{r}) \in X$, if and only if $c_\ell(Q \cdot (x, \mathbf{r})) = c_\ell(x, \mathbf{r})$ for every ℓ, every $Q \in \mathrm{SO}(3)$, and every $(x, \mathbf{r}) \in X$ with $x_\ell \neq 0$.*

Proof of Lemma 1. Fix Q and $(x, \mathbf{r})$. Because a scalar commutes with the linear map $D^{(\ell)}(Q)$, the degree-ℓ block of $G(Q \cdot (x, \mathbf{r})) - Q \cdot G(x, \mathbf{r})$ equals $\big(c_\ell(Q \cdot (x, \mathbf{r})) - c_\ell(x, \mathbf{r})\big)\, D^{(\ell)}(Q)\, x_\ell$. If the gates are invariant, this block vanishes for every ℓ, also when $x_\ell = 0$, so G is equivariant. Conversely, if G is equivariant, the block vanishes for every $(x, \mathbf{r})$; for $x_\ell \neq 0$ the vector $D^{(\ell)}(Q)\, x_\ell$ is nonzero because $D^{(\ell)}(Q)$ is invertible, so the scalar factor must vanish, which is the invariance condition. □

Corollary 1 (Radial gating is equivariant)**.** *If $c_\ell(x, \mathbf{r}) \coloneqq g_\ell(r_{ij})$ for an edge (i, j), where $g_\ell : \mathbb{R}_{>0} \to \mathbb{R}$ is any function, in particular the radial gate of Equation* (3)*, then G is $\mathrm{SO}(3)$-equivariant.*

Proof of Corollary 1. Since $Q^{\mathrm{T}}Q = I$, $c_\ell(Q\cdot(x, \mathbf{r})) = g_\ell(\|Q\mathbf{r}_i - Q\mathbf{r}_j\|) = g_\ell(\|\mathbf{r}_i - \mathbf{r}_j\|) = c_\ell(x, \mathbf{r})$ for every ℓ, so the condition of Lemma 1 holds. □

Proposition 1 (Masking-reduction equivalence)**.** *Fix an edge with active degrees $\{0, \dots, k\}$ and gate values $g_0, \dots, g_k$. The degree-$\ell \le k$ outputs of the masked computation, in which the edgewise computation at $L_{\max}$ is evaluated and every degree above k is set to zero after every operation, equal the outputs of the degree-k operator, which is the same computation under the following parameter restriction: orders $|m| > k$ are dropped; each remaining per-$|m|$ linear map is restricted to degrees $\ell = |m|, \dots, k$ on its input and on its output; the $m = 0$ map is additionally restricted to the activation-gate scalar rows of degrees $1, \dots, k$; the radial MLP keeps the output rows of the remaining (m, ℓ) pairs; and the same gate values g_ℓ multiply the remaining degrees. The identity holds for the forward values and for the gradients with respect to the remaining inputs and parameters, and in the masked computation the gradients of all discarded entries vanish.*

Proof of Proposition 1. Each per-$|m|$ linear map acts on the degrees of that order, so setting the input degrees $\ell > k$ to zero removes the corresponding columns and masking the output degrees removes the corresponding rows, which leaves exactly the restricted map; under the reflection-symmetric restriction this is the leading sub-block of W_1, and otherwise the same restriction applies to W_1 and W_2 separately. The $m = 0$ map has a second output section of activation-gate scalars indexed by degree, which is restricted to $\ell = 1, \dots, k$ in the same way, and the radial MLP produces coefficients indexed by (m, ℓ, c) that act diagonally on the corresponding features, so keeping $\ell \le k$ selects the matching output rows. The Wigner rotations are block-diagonal in ℓ and the gate values act degree-wise, so both commute with the restriction, and because the mask is reapplied after every operation, the bias and activation-gate entries of the discarded degrees, which produce output independently of the input, are set to zero downstream, which gives the vanishing gradients. Every restriction is a linear projection and composition preserves the identity, so the vector-Jacobian products of all remaining quantities are identical. □

K.1 Gate support and nested radii

Stage B evaluates degrees $0, \dots, k_e$ on each edge and skips the rest. This computation equals the masked computation only if the radial gate is exactly zero beyond its gate radius, $g_\ell(r) = 0$ for $r \ge R_\ell$, and if the gate radii are nested within each block, $R_1 \ge R_2 \ge \cdots \ge R_{L_{\max}}$. Because $g_0 \equiv 1$ and the radius floor sets $R_1 \ge 6.5\,\text{Å}$, degrees 0 and 1 are never skipped, but nothing orders $R_2, \dots, R_{L_{\max}}$. A sigmoid gate, $g_\ell(r) = \mathrm{sigmoid}((R_\ell - r)/\tau)$, is positive at every distance, so a threshold of zero skips nothing and a threshold of one half drops half of the degree-ℓ contribution at $r = R_\ell$, a jump at which the force $-\partial E/\partial \mathbf{x}$ is undefined; a direct force head does not differentiate through this jump, but the predicted energy keeps it. The radial gate of Equation (3) is zero for $r \ge R_\ell$ and positive for $r < R_\ell$, so skipping at threshold zero removes only exact zeros. Replacing the sigmoid gate by Equation (3) reduced the difference between the masked and the degree-k features from 4.5% to 1.4×10^{-6} and made the energy C^2 instead of discontinuous at every gate radius.

K.2 Non-nested radii

The degree-k operator requires the active degrees to be $\{0, \dots, k\}$, so each edge is assigned the largest such run, $k_e = \max\{k : g_\ell(r_e) > 0 \;\forall\, \ell \le k\}$. If the gate radii are nested, this set is exactly the set of active degrees and Proposition 1 applies. If they are not, the degree-k_e operator drops a degree ℓ' with a positive gate only on edges with $r_e \ge R_\ell$ for some $\ell < \ell'$, so the dropped gate value is at most $\sigma_5\big((R_{\ell'} - R_\ell)/w_{\ell'}\big)$. The bucket assignment is made independently in each block, so nesting has to hold per block: of the 77 runs of Appendix N, the 36 that use the deployed gate shape (33 molecular, 3 periodic; floored dipole, $r_1^{\text{floor}} = 6.5\,\text{Å}$) are nested in every block in 25 cases, and in the remaining 11 the non-nesting is confined to the last block in every case, and to the last block alone in 9. The bound never exceeded 6.2×10^{-3}: at most 3.0% of the OMol25 validation edges lose one degree in one block, at a mean dropped gate value of 0.0006. Stage B trains with the deployed assignment, so only the transition from Stage A2 to Stage B can change the output.

Table 13: Training settings. All runs use four blocks, 128 channels, $L_{\max} = m_{\max} = 4$, a 6 Å cutoff, and 256 maximum neighbors. Shared settings are AdamW, weight decay 10^{-3}, gradient clipping 100, EMA 0.999, energy and force loss weights 30 each, BF16, and a cosine learning-rate schedule. Forces are predicted directly unless stated otherwise. Paired entries give dense/RAL values. RAL uses the reflection-symmetric restriction and staged training: A uses radial gates, and B uses Band-MoE with frozen boundaries derived from Stage A1. "–" denotes not applicable or unspecified.

	400k runs	Dense/RAL, 102M	Dense/RAL, 4M
Training data	400k per domain	102M	4M
Total steps	30k	200k	200k
Hardware	4×H200	128×GH200	8×GH200
Atoms per GPU	6500/4500[a]	2300	2300/6500
Base LR	8×10^{-4}	2×10^{-3}	10^{-3} / 1.13×10^{-3}
Warmup	–	0.1; 0.2 (RAL: Stage A)	0.01; 0.2 (RAL: Stage A)
Final LR	–	1% of peak	1% of peak / 1% of base
Activation checkpointing	off	off / A on, B off	on / A on, B off
Reflection restriction	no	no/yes	optional/yes
Parameters (M)	22.8	22.8/16.9	22.8 or 16.9 / 16.9
Radial-gate settings; not used by dense models			
Stage A1 steps	5k	10k	5k
Penalty ramp (steps)	3000	2000	1000
$\lambda_{1,2,3,4}$		$[0, 0.5, 0.7, 0.9]$	
Widths $w_{1,2,3,4}$		$[0.5, 2, 2, 2]$ Å	
Radius floor		$R_1 \geq 6.5$ Å	
Band-MoE settings; RAL only			
Stage B starts at	–	40k	$\{40, 60, 100\}$k
Stage B peak LR	–	1.81×10^{-3}	$\{0.8, 0.6\}\times$ base
Expert degrees	–	$\{1, 2, 4\}$	$\{1, 2, 4\}$
Transition width	–	0.15 Å	0.15 Å

[a]OMol25/OMat24. The 400k runs are reported in Table 4.

L HYPERPARAMETERS

Table 13 lists the training settings of the reported runs.

M SYNTHETIC VALIDATION

The training datasets do not specify how the contribution of each spherical-harmonic degree depends on distance. We therefore construct a synthetic target with a prescribed distance dependence for each degree and test whether the radial gates recover the ordering of these ranges.

The degree-ℓ content of the target has the range $\xi_\ell = \xi_0 \kappa^\ell$ with $\xi_0 = 4.0$ Å and $\kappa = 0.7$:

$$M_i^{(\ell)} = \sum_j e^{-r_{ij}/\xi_\ell} \, \mathrm{env}(r_{ij}) \, Y_\ell(\hat{\mathbf{r}}_{ij}), \qquad E = \sum_{i,\ell} a_\ell \big\| M_i^{(\ell)} \big\|^2, \tag{8}$$

where Y_ℓ denotes the real spherical harmonics of degree ℓ, $\mathrm{env}(r) = \frac{1}{2}(1 + \cos(\pi r / r_c))$ with $r_c = 5.0$ Å, and $a_\ell = 1$; the forces are the analytic gradient of E, and the exponential is the only term that couples ℓ to r. Each system has 28 atoms in a 5.5 Å box, rejection-sampled with a minimum separation of 0.9 Å, with 64 training and 48 validation systems per seed. The model

Table 14: Synthetic validation on the validation systems, three seeds (mean ± standard deviation). $R_{1,2,3,4}$: gate radii averaged over blocks per degree; ordered seeds: seeds in which $R_1 > R_2 > R_3 > R_4$; RAL is initialized at $R_\ell = r_c = 5.0$ Å. The last column gives the change in force MAE and in GFLOP relative to the dense variant. Ord. denotes ordered; unif denotes uniform

model	$R_{1,2,3,4}$ (Å)	ord. seeds	F MAE	GFLOP	ΔF / Δgflop v dense
dense	–	–	0.803 ± 0.068	1.926	–
unif. $L_{\max}$=1	–	–	1.195 ± 0.072	0.521	$+48.9\%$ / -72.9%
unif. $L_{\max}$=2	–	–	0.962 ± 0.078	0.967	$+19.9\%$ / -49.8%
unif. $L_{\max}$=3	–	–	1.204 ± 0.065	1.443	$+50.0\%$ / -25.1%
RAL, λ=0	5.20, 5.30, 5.26, 5.32	0/3	0.834 ± 0.172	1.815	$+3.9\%$ / -5.7%
RAL, λ=0.02	5.28, 5.25, 5.17, 4.89	1/3	0.851 ± 0.198	1.751	$+6.0\%$ / -9.1%
RAL, λ=0.05	5.29, 5.23, 4.98, 4.32	3/3	0.939 ± 0.173	1.635	$+17.0\%$ / -15.1%
RAL, λ=0.15	5.17, 4.92, 3.97, 3.15	3/3	0.813 ± 0.051	1.265	$+1.25\%$ / -34.3%

is a two-layer tensor-field network implemented in e3nn (Geiger & Smidt, 2022), trained in three variants that differ only in the gate radii: dense ($g_\ell \equiv 1$ for every degree), uniform $L_{\max}$=k ($g_\ell \equiv 1$ for $\ell \leq k$), and RAL with every gate radius initialized at $R_\ell = r_c$, so that any ordering must be learned.

In FP64 the model is exactly equivariant ($\Delta E = 0$, force residual 1.5×10^{-16}), and finite-difference forces match the autograd forces to 1.1×10^{-8} across a gate radius. The gate radii were unordered in every seed at $\lambda = 0$ and ordered in every seed at $\lambda \geq 0.05$, moving by up to 1.85 Å from the initialization (Table 14); the penalty weight sets the absolute radii, while the ordering is the property of the target that training recovers. At $\lambda = 0.15$, RAL used 34.3% fewer GFLOP than the dense variant at 1.25% higher force MAE (0.813 vs. 0.803; $t = 0.21$, $n = 3$), and its force MAE was 15.5% lower than that of uniform $L_{\max}$=2 ($t = -2.78$) and 32.5% lower than that of uniform $L_{\max}$=3 ($t = -8.27$). We report forces only, because the loss weights the forces 30:1 over the energies, and the energy is not learned at this weighting.

N LEARNED RADII

The gate radii are learned independently, yet the block-mean ordering $R_1 > R_2 > R_3 > R_4$ held in 70 of the 77 runs at $L_{\max} = 4$, and in all 29 runs at 400k and above; the seven exceptions are all in the 40k screen. In runs without the radius floor ($\lambda_1 = 0.001$) and with $\lambda_{2,3,4} = [1.5, 2.1, 2.7]$, the gate radii averaged over blocks were $[6.23, 3.47, 3.28, 3.03]$ Å on OMol25 400k and $[6.15, 4.00, 3.76, 3.64]$ Å on OMat24 400k (Figure 6b). Because R_1 lay close to the 6.0 Å cutoff, the final configuration sets $r_1^{\text{floor}} = 6.5$ Å; the floor is applied in every forward pass, so $g_1 \equiv 1$ on every edge within the cutoff from the start of Stage A1.

The gate radii also depend on the block and on the domain. The following comparisons use the final Stage A runs on OMol25 4M and OMat24 1M, both with the radius floor applied. On OMat24, the block-mean R_2 increased from 4.51 to 4.95 Å over the four blocks, and the number of active spherical-harmonic coefficients increased with it, while OMol25 showed no monotonic trend across blocks OMat24 edges are $1.13\times$ longer on average than OMol25 edges, which with the OMol25 gate radii would leave 16.9% of the OMat24 edges at the full expansion; the OMat24 gate radii are $1.31\times$ larger, which raises this fraction to 38.7%, against 30.8% on OMol25. The larger gate radii on OMat24 therefore offset its longer edges, and the two full-expansion fractions differ by $1.3\times$ instead of the $1.8\times$ implied by the edge lengths alone.

Over 1.81M OMol25 validation edges from 500 systems, the reduction in active spherical-harmonic coefficients was 55.7%, 52.7%, 56.1%, and 58.7% in blocks 0 to 3. A uniform edge density ($\propto r^2$) predicts that degrees 2, 3, and 4 are active on 24%, 21%, and 18% of the edges, against the measured 38%, 34%, and 31% (Figure 6c), because the lengths of the validation edges are concentrated at intermediate values (mean 4.07 Å).

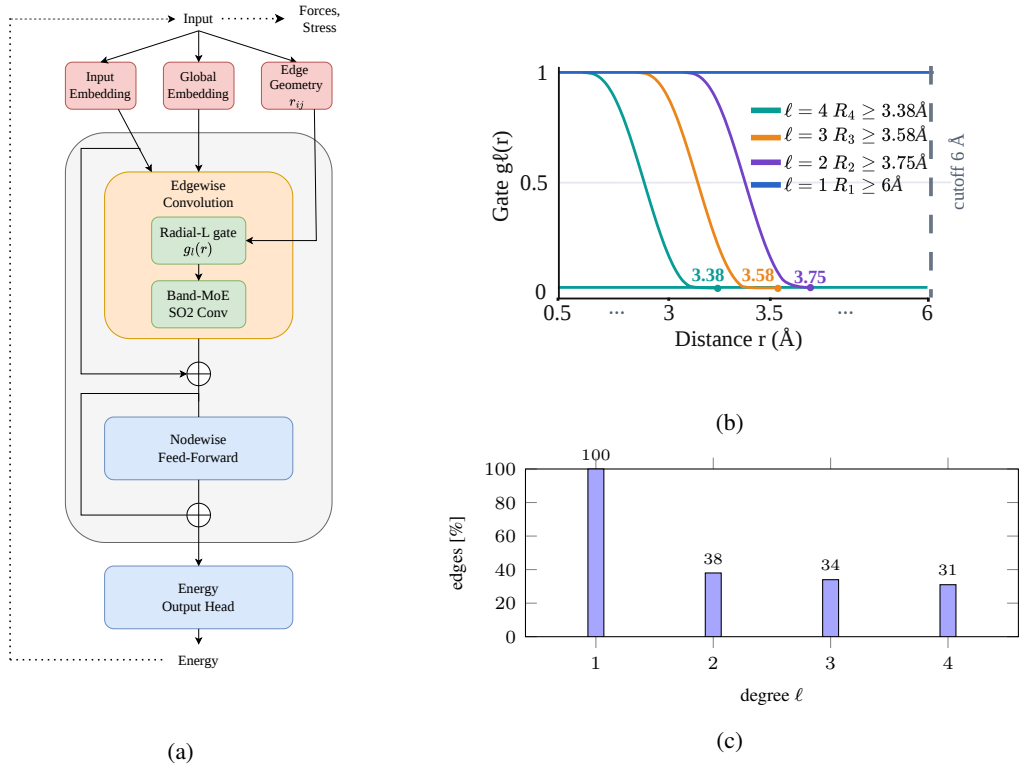


Figure 6: Learned radial gates and their effect on the active spherical-harmonic coefficients. **(a)** Edgewise computation of eSEN with the radial gate $g_\ell(r_{ij})$ and the Band-MoE router, which take the edge length r_{ij} as an additional input. **(b)** Learned radial gates on OMol25 400k: $g_1 = 1$ within the cutoff, and g_ℓ for $\ell \geq 2$ reaches zero at 3–3.5 Å. **(c)** Share of OMol25 validation edges (500 systems) with a nonzero gate per degree, computed from the 4M AdamW RAL gate radii; on average 11.1 of 25 coefficients per edge are active (55.8% fewer).