ON THE STEEPER SCALING LAWS OF EQUIVARIANT NEURAL FORCE FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariance to Euclidean symmetries is a critical inductive bias for neural networks that model 3D atomistic systems. Whether to build equivariance into a network or learn it through data augmentation remains an active debate. Prior work has empirically shown that equivariant networks achieve steeper scaling laws on certain tasks, raising the open question of what causes this gap. This paper takes a step toward an answer in the setting of machine learning interatomic potentials (MLIPs), where the targets are equivariant forces. In particular, we show that a theory of scaling laws based on source-capacity conditions explains this gap. The capacity and source exponents of each architecture that reflect the decay rate of the network's kernel spectrum and its alignment with the target forces, respectively, account for the observed exponents in the empirical scaling laws. Our results show that for equivariant targets, equivariant architectures scale better than their unconstrained counterparts, because their kernels align better with the targets. Since these exponents can be estimated from an early training checkpoint, they provide a practical way to forecast the scaling laws of MLIP architectures without full-scale training.

1 INTRODUCTION

Machine learning interatomic potentials (MLIPs) have emerged as a promising surrogate for quantum-mechanical calculations such as density functional theory (DFT), predicting the energies and atomic forces of atomistic systems at a fraction of the cost (Gilmer et al., 2017; Batzner et al., 2022). Such models are now used throughout computational pipelines, for example in molecular dynamics simulations (Batzner et al., 2022) and enzyme catalysis (Gao et al., 2026), and have substantially accelerated materials discovery. Motivated by this potential, several groups have released increasingly large 3D atomistic datasets (Levine et al., 2025) to train and evaluate MLIPs at scale, leading to the rapid development of foundation MLIPs (Barroso-Luque et al., 2025; Wood et al., 2025; Rhodes et al., 2025).

A defining characteristic of MLIPs is that they must respect the Euclidean symmetries of the underlying physics, including translations, rotations, reflections, and any composition of these (Kondor, 2025). Concretely, the network must be equivariant, i.e., transforming its input must transform its output in a correspondingly predictable way. Two distinct approaches to equivariance have emerged in recent years. The first, rooted in representation theory, builds equivariance into the architecture by operating on group representations with specialized operations such as tensor products, giving rise to a rapidly growing family of $E(3)$-equivariant graph neural networks (Thomas et al., 2018; Anderson et al., 2019; Fuchs et al., 2020; Batatia et al., 2022; Batzner et al., 2022; Passaro & Zitnick, 2023; Fu et al., 2025; Liao et al., 2024; 2026). The second argues that equivariance can instead be learned; unconstrained models built from hardware-friendly dense matrix multiplications perform competitively on 3D atom systems when trained with rotational data augmentation on sufficiently large datasets (Qu & Krishnapriyan, 2024; Rhodes et al., 2025; Elhag et al., 2026; Qu et al., 2026). Together, these two approaches frame a central debate in atomistic modeling over how equivariance should be treated as models are scaled to larger datasets and compute budgets.

Despite scaling progress, how MLIPs should scale remains far less understood than in language and vision models. In those domains, empirical neural scaling laws relating test loss to data, parameters, and compute (Kaplan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022) have matured into con-

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.

crete recipes for allocating a fixed compute budget. Furthermore, predictive theories of scaling laws have been actively developed to explain and predict the scaling exponents of these networks(Bahri et al., 2024; Paquette et al., 2024; Bordelon et al., 2024; 2025). For MLIPs, the picture is much less settled. Existing empirical results (Batzner et al., 2022; Frey et al., 2023; Ngo & Ravanbakhsh, 2026) demonstrate power-law improvements with data and model size for equivariant networks. Interestingly, higher-order equivariant networks tend to exhibit steeper scaling laws (i.e., larger power-law exponents) than their invariant or unconstrained counterparts on small datasets or under single-epoch training. Whether this gap widens or closes when unconstrained networks are trained longer with data augmentation (a.k.a. infinite compute regime) remains actively debated (Qu & Krishnapriyan, 2024; Brehmer et al., 2025; Pengmei et al., 2025; Shen et al., 2025; Qu et al., 2026). More importantly, no theoretical account currently explains *where these architecture-dependent exponents come from*.

1.1 Contributions

In this paper, we take a step toward answering this question. We study two representative MLIP architectures, the higher-order equivariant eSEN (Fu et al., 2025) and the unconstrained Orb (Rhodes et al., 2025), and explain the gap between their scaling exponents using the kernel–task alignment framework (Caponnetto & Vito, 2005; Pillaud-Vivien et al., 2018; Cui et al., 2021; 2023; Canatar et al., 2021). The framework views a network near a given checkpoint as a kernel regressor and characterizes it by two exponents. The *capacity* exponent α describes how quickly the kernel's eigenvalues decay, i.e., how much of the model's expressive power is concentrated in a few directions. A larger spectral exponent means that the network learns its favored patterns much faster than the rest, whereas a smaller exponent implies more balanced learning rates across a broader range of patterns. The *source* exponent β describes how well the regression target is aligned with those leading directions. A large β means the target is mostly explained by the directions the kernel already favours, so the task is easy for that kernel. Together they predict how the loss scales with model size (as $P^{-\alpha\beta}$) and set a lower bound on how fast it scales with training time. Under this view, two architectures with similar capacity can scale differently if their kernels are aligned differently with the target. Our contributions are as follows:

- **A kernel that captures the full equivariant network.** Last-layer features are not a faithful probe of higher-order equivariant networks, so we instead use the empirical neural tangent kernel (NTK) of the predicted forces at an early checkpoint. Measured on OMol25 forces, the two architectures have similar capacity exponents, but eSEN's kernel aligns much better with the force labels.
- **Scaling laws in a realistic training regime.** To test whether this alignment gap explains the scaling gap, we fit neural scaling laws for both architectures trained for multiple epochs on the neutral set of OMol25 (Levine et al., 2025), with rotational data augmentation for Orb. This goes beyond prior work, which considered single-epoch training (Ngo & Ravanbakhsh, 2026), small datasets (Frey et al., 2023; Batzner et al., 2022; Großmann et al., 2026), or equivariant networks alone without a direct comparison to unconstrained models (Wood et al., 2025).
- **Exponents can track scaling laws.** Our empirical fits indicate that the fitted model-size scaling exponents agree with the kernel prediction $\alpha\beta$ within uncertainty for all architectures under study, and the ordering of the training-time exponents matches the ordering of β. Furthermore, estimations of α and β from early checkpoints also hold for four architectures trained on the largest data size, where it separates design choices that only lower the loss (Orb's edge stream) from those that change the exponent (higher-order features). Taken together, these results suggest that the equivariant network's steeper scaling comes not from higher capacity but from better alignment of its kernel with the force targets.

2 Background

2.1 Source and Capacity Conditions for Scaling Laws

Our work builds on the framework of Bordelon et al. (2024; 2025); Paquette et al. (2024). This line of theory explains empirical scaling laws through the lens of kernel–task alignment (Canatar et al., 2021). The core idea is that the scaling exponents are set by how fast the network kernel's eigenvalues decay and how the target is spread across the corresponding eigendirections in feature

space. Mathematically, consider a linear model on a fixed set of features $\psi(x) \in \mathbb{R}^M$ in the limit $M \to \infty$, such as the features defining an infinite-width NTK. The target is a linear combination of these features, $y(x) = \bar{\mathbf{w}}^\top \psi(x)$. The features have covariance $\Sigma = \mathbb{E}[\psi\psi^\top]$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots$. We define $\bar{w}_k = \mathbf{v}_k^\top \bar{\mathbf{w}}$ as the target's component along the k-th eigendirection. Following the source and capacity conditions of kernel regression (Caponnetto & Vito, 2005; Cui et al., 2021; 2023), Bordelon et al. (2024; 2025); Paquette et al. (2024) assume power-law decays in the spectrum of Σ and in the target variance:

$$\lambda_k \sim k^{-\alpha}, \qquad \sum_{j>k} \lambda_j \bar{w}_j^2 \sim k^{-\alpha\beta}. \tag{1}$$

The capacity exponent α measures the decay rate of the spectrum, and the source exponent β quantifies how fast the target variance $\lambda_k \bar{w}_k^2$ decays along the eigendirections relative to the spectrum. A large β means the target can be captured by the leading eigendirections, while a small β indicates that its variance extends towards the tail. Therefore, β indicates the difficulty of the task under the network's NTK; tasks with $\beta < 1$ are defined as hard. Furthermore, measuring β on real networks and datasets typically yields $\beta < 1$ (Bordelon et al., 2025). Under squared loss, Bordelon et al. (2024; 2025) show that, up to constant prefactors, the loss of a network with P parameters trained on D samples for t steps scales as

$$\mathcal{L} \approx t^{-g(\beta)} + P^{-\alpha\beta} + D^{-\alpha\beta} + \mathcal{L}_\infty. \tag{2}$$

The data and model-size exponents are both $\alpha\beta$, whereas the training-time exponent $g(\beta)$ depends on the learning regime. In lazy learning regimes, $g(\beta) = \beta$ (Maloney et al., 2022; Bordelon et al., 2024). Bordelon et al. (2025) show that strongly feature-learning networks exhibit faster decay in training time for $\beta < 1$, reaching $g(\beta) = 2\beta/(1+\beta)$. This derivation can capture the dynamics of certain toy models and small networks on MNIST, but it still underestimates the exponents of practical networks on larger datasets (Bordelon et al., 2025). When the dataset is large enough that the data term $D^{-\alpha\beta}$ remains negligible throughout training, equation 2 reduces to

$$\mathcal{L}(t, P) \approx \mathcal{L}_\infty + A_t\, t^{-g(\beta)} + A_P\, P^{-\alpha\beta}. \tag{3}$$

This is the regime of compute-optimal scaling analyses (Hoffmann et al., 2022; Bordelon et al., 2024; Paquette et al., 2024) and the relevant one for pretraining on a fixed large dataset, where the question is how to allocate compute between P and t. We therefore focus on the model-size exponent $\alpha\beta$ and the training-time exponent $g(\beta)$.

2.2 Higher-Order Equivariant Networks

A 3D atomistic system is a point cloud $\{(z_1, \mathbf{x}_1), \ldots, (z_n, \mathbf{x}_n)\}$ of atomic numbers and positions in $\mathbb{R}^3$. Translation invariance is obtained by operating on relative positions $\mathbf{r}_{ij} = \mathbf{x}_j - \mathbf{x}_i$, so the remaining nontrivial symmetry is the rotation group $\mathrm{SO}(3)$; reflections can additionally be handled by tracking the parity of each feature, which we omit here for simplicity. Higher-order equivariant networks (Thomas et al., 2018; Kondor et al., 2018; Anderson et al., 2019) represent the features of each atom as a direct sum of irreducible representations (irreps) of $\mathrm{SO}(3)$. The irrep of degree $\ell \in \{0, 1, 2, \ldots\}$ acts on $V_\ell = \mathbb{R}^{2\ell+1}$ through the real Wigner D-matrix $D^\ell(g)$, an orthogonal $(2\ell+1) \times (2\ell+1)$ matrix. Features of degree $\ell = 0$ are scalars, $\ell = 1$ features are vectors, and $\ell \geq 2$ features are higher-order tensors. A node feature with c_ℓ channels of each degree $\ell \leq \ell_{\max}$ has the form, and transforms under a rotation $g \in \mathrm{SO}(3)$ as,

$$\mathbf{h} = \bigoplus_{\ell=0}^{\ell_{\max}} \mathbf{h}_\ell, \qquad \mathbf{h}_\ell \in \mathbb{R}^{c_\ell(2\ell+1)}, \qquad \mathbf{h}_\ell \mapsto \left(I_{c_\ell} \otimes D^\ell(g)\right) \mathbf{h}_\ell. \tag{4}$$

Networks with $\ell_{\max} = 0$ are called invariant, those with $\ell_{\max} = 1$ vector networks, and those with $\ell_{\max} \geq 2$ higher-order networks. Schur's lemma states that a linear map $A : V_\ell \to V_{\ell'}$ that commutes with the group action, $AD^\ell(g) = D^{\ell'}(g)A$ for all g, is zero if $\ell \neq \ell'$ and a scalar multiple of the identity if $\ell = \ell'$ (Fulton & Harris, 2013). This statement has implications for the design of higher-order equivariant networks. An equivariant linear layer cannot mix degrees. In particular, on features of the form in Eq. equation 4, it is necessarily block-diagonal, $\mathbf{h}_\ell \mapsto (W_\ell \otimes I_{2\ell+1})\, \mathbf{h}_\ell$ with $W_\ell \in \mathbb{R}^{c'_\ell \times c_\ell}$, so it mixes channels within a degree but never across degrees.

Tensor product. Since equivariant linear layers cannot mix degrees, features of different degrees interact through the tensor product. The product of degree-ℓ_1 and degree-ℓ_2 features splits into one component of each degree $|\ell_1 - \ell_2| \le \ell \le \ell_1 + \ell_2$. By Schur's lemma, each component is unique up to scale and is given by the Clebsch–Gordan coefficients:

$$\big(\mathbf{h}_{\ell_1} \otimes \mathbf{h}'_{\ell_2}\big)^{\ell}_{m} = \sum_{m_1,m_2} C^{\ell m}_{\ell_1 m_1,\ell_2 m_2}\, h_{\ell_1,m_1}\, h'_{\ell_2,m_2}. \tag{5}$$

In message passing, atom j sends atom i the message $\sum_{\ell_1,\ell_2} w_{\ell_1\ell_2\ell}(\|\mathbf{r}_{ij}\|)\big(\mathbf{h}_{j,\ell_1} \otimes Y^{\ell_2}(\hat{\mathbf{r}}_{ij})\big)^{\ell}$, where Y^{ℓ_2} are spherical harmonics of the edge direction and w are learnable radial functions (Thomas et al., 2018; Batzner et al., 2022). Tensor products are thus the only route by which directional information reaches the scalar channels and by which different degrees interact. Moreover, distinguishing certain geometric configurations provably requires higher-order tensor features (Joshi et al., 2023). The cost of a full tensor product grows as $\mathcal{O}(\ell^6_{\max})$, motivating the SO(2) reduction of Passaro & Zitnick (2023), which lowers it to $\mathcal{O}(\ell^3_{\max})$ and underlies certain foundation models (Fu et al., 2025; Wood et al., 2025).

3 Approach

For practical networks, one can empirically estimate α and β from a large dataset by constructing the kernel Gram matrix $K(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})\psi(\mathbf{x}')^\top$ (Bordelon et al., 2024; 2025). To construct such a Gram matrix, we need a feature map ψ for each architecture. A simple choice is the hidden representations at the last layer. The network's output is a linear readout of it, so its Gram matrix is the kernel of the final regression. For unconstrained models, this is unproblematic, as the force from every atom is read out from the same d-dimensional feature vector. For the higher-order equivariant network, last-layer features are poor probe.

Ambiguous inner product. A last-layer kernel requires an inner product between feature vectors. The Euclidean inner product on the concatenated features is rotation invariant, since each $\boldsymbol{D}^{\ell}(g)$ is orthogonal, but it is not the only one. Assuming one channel per degree ($\mathbf{h}_\ell \in \mathbb{R}^{2\ell+1}$), Schur's lemma (Section 2.2) implies that the rotation-invariant inner products on features of the form in equation 4 are exactly the weighted sums $\langle \mathbf{h}, \mathbf{h}' \rangle = \sum_{\ell=0}^{\ell_{\max}} w_\ell \langle \mathbf{h}_\ell, \mathbf{h}'_\ell \rangle$ with $w_\ell > 0$, the Euclidean case being $w_\ell = 1$.[1] Nothing in the network fixes these weights, and the Euclidean default is not neutral because degrees differ in dimension ($2\ell+1$) and typical magnitude, $w_\ell = 1$ implicitly favors some degrees over others. The Gram matrix, and hence α and β, would therefore depend on an arbitrary weighting rather than on what the network has learned.

Missing features from higher degrees. The usual choice in representation analyses of equivariant MLIPs is to use the invariant $\ell = 0$ block alone (Li & Walsh, 2026; Edamadaka et al., 2025). This is well suited to comparing embeddings across models, since the resulting features are rotation-invariant regardless of architecture. For our purpose it is too restrictive, as kernel built from the $\ell = 0$ block only sees the scalar channels and cannot capture the higher-degree features that give equivariant networks their additional expressivity. Since the goal here is to characterize the full network rather than a rotation-invariant summary of it, we need a kernel that includes every degree.

3.1 Empirical NTK as Feature Map

We resolve both issues by using the Jacobian of the model's predicted forces with respect to its parameters. Let $\theta_t \in \mathbb{R}^P$ denote vectors of P parameters of a checkpoint at an early training step t, and let $\hat{\mathbf{f}}_i(\theta) \in \mathbb{R}^3$ denote the predicted force on atom i as a function of the parameters. For n atoms, we define the Jacobian $J \in \mathbb{R}^{3n\times P}$, with one row per atom and Cartesian component,

$$J_{(i,c)} = \left.\frac{\partial \hat{f}_{i,c}}{\partial \theta}\right|_{\theta=\theta_t}, \qquad c \in \{x, y, z\}, \tag{6}$$

and the corresponding Gram matrix

$$K = JJ^\top \in \mathbb{R}^{3n\times 3n}, \qquad K_{(i,c),(j,d)} = \big\langle \partial_\theta \hat{f}_{i,c},\, \partial_\theta \hat{f}_{j,d} \big\rangle, \tag{7}$$

[1] For $c_\ell > 1$ channels, w_ℓ becomes a symmetric positive-definite matrix $W_\ell \in \mathbb{R}^{c_\ell \times c_\ell}$, which only enlarges this freedom.

which is the empirical neural tangent kernel (NTK) of the force output evaluated at θ_t (Jacot et al., 2018). The Euclidean inner product on the parameter gradients is not an arbitrary choice. It is the inner product that the gradient uses, so the kernel $K = JJ^\top$ is the kernel that governs the training dynamics. Furthermore, by the chain rule, the gradient aggregates contributions from every layer, including the tensor-product interactions across degrees, and therefore the kernel K characterizes the full model rather than only its $\ell = 1$ readout.

Equivariance of NTK feature map. The remaining question is whether the empirical NTK feature map can preserve the symmetry of an equivariant network. The answer is yes when the network f_θ is G-equivariant for every parameter θ, which holds by construction for equivariant networks. Concretely, a G-equivariant network f_θ is defined as $f_\theta(\rho_{\text{in}}(g)\mathbf{x}) = \rho_{\text{out}}(g)f_\theta(\mathbf{x})$, where ρ_{in} and ρ_{out} are the linear actions on the network's inputs and outputs, respectively. One can write the NTK feature of f_θ as $\psi_\theta(\mathbf{x}) = \nabla_\theta f(\mathbf{x})$. Because $\rho_{\text{in}}(g)$ and $\rho_{\text{out}}(g)$ are independent of θ, taking gradients of f_θ w.r.t θ leads to:

$$\psi_\theta(\rho_{\text{in}}(g)\mathbf{x}) = \nabla_\theta f_\theta(\rho_{\text{in}}(g)\mathbf{x}) = \nabla_\theta \big[\rho_{\text{out}}(g)\, f_\theta(\mathbf{x})\big] = \rho_{\text{out}}(g)\, \nabla_\theta f_\theta(\mathbf{x}) = \rho_{\text{out}}(g)\, \psi_\theta(\mathbf{x}). \quad (8)$$

Spectrum of the kernel and alignment in force-fields. We diagonalize $K = U\Lambda U^\top$, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{3n} \geq 0$ and orthonormal eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_{3n} \in \mathbb{R}^{3n}$. The capacity condition posits a power-law decay $\lambda_k \propto k^{-\alpha}$, so a larger α corresponds to a kernel whose variance is concentrated in fewer directions. Let $\mathbf{f} \in \mathbb{R}^{3n}$ stack the Cartesian components of the reference forces on the same atoms. Following (Canatar et al., 2021), the fraction of label variance left after projecting out the leading k eigenvectors is

$$R(k) = \frac{\big\|(I - U_k U_k^\top)\,\mathbf{f}\big\|^2}{\|\mathbf{f}\|^2} = \frac{\sum_{j>k}(\mathbf{u}_j^\top \mathbf{f})^2}{\sum_j (\mathbf{u}_j^\top \mathbf{f})^2}, \qquad U_k = [\mathbf{u}_1, \ldots, \mathbf{u}_k], \quad (9)$$

and the source condition posits $R(k) \propto k^{-\alpha\beta}$. At fixed α, a larger β means the target is concentrated in the directions to which the kernel assigns the largest eigenvalues. Additionally, from equation 8, the kernel identify follows:

$$K_\theta(\rho_{\text{in}}(g)\mathbf{x}, \rho_{\text{in}}(g)\mathbf{x}') = \rho_{\text{out}}(g) K_\theta(\mathbf{x}, \mathbf{x}') \rho_{\text{out}}(g)^\top. \quad (10)$$

For force prediction, $\rho_{\text{in}}(g)$ and $\rho_{\text{out}}(g)$ act as 3×3 orthogonal matrices on each atom 3D coordinates and force outputs, respectively. This means that for equivariant networks, rotations transform the kernel K by an orthogonal similarity, and thus its spectrum and α remains unchanged. Since the reference forces transform by the same $\rho_{\text{out}}(g)$, the residual $R(k)$ in equation 9 and β are invariant under rotations as well. In other words, the estimated exponents do not depend on the choice of reference frame.

4 Experiments

4.1 Experimental Setup

We extend the setup of Ngo & Ravanbakhsh (2026) to multi-epoch training on the OMol25 neutral set (Levine et al., 2025). To our knowledge, it is the largest available dataset of neutral molecules, and it includes recomputed structures from earlier datasets such as SPICE and ANI-2x. Restricting to neutral molecules lets us compare architectures without the added complexity of charge and spin. We use two subsets, 10% and 100% of the set, with approximately 3.4M and 34M training structures. Because the targets are per-atom forces, we measure dataset size in atoms: approximately 93M and 930M, respectively. We train for 15–20 epochs on the 93M-atom subset and, as in Ngo & Ravanbakhsh (2026), for a single epoch on the full set. We track the validation loss on a held-out set of 27K structures.

We study two representative architectures, namely eSEN (Fu et al., 2025) and Orb (Rhodes et al., 2025). The former is a higher-order equivariant architecture whose node representations follow the definition in section 2.2, whereas the latter is an unconstrained message-passing network that achieves competitive performance across benchmarks. The architectures are trained on the direct-force prediction task. The reason for this is twofold. First, while learning the energy is sufficient for predicting conservative forces, direct-force prediction is substantially cheaper, as it avoids a

second backward pass through the energy. Although directly predicted forces are not conservative, this is the setting used in practice for pretraining. The direct-force head can later be removed and the model fine-tuned to predict conservative forces as the negative gradient of the predicted energy (Duval et al., 2023; Bigi et al., 2025; Fu et al., 2025). Second, since our goal is to study how each network's kernel aligns with the target forces, we use a simple linear layer on top of each architecture to output the force predictions. This keeps the force a linear readout of the last-layer features, so that the kernel we measure is not altered by complex readout layers or by differentiation through the energy.

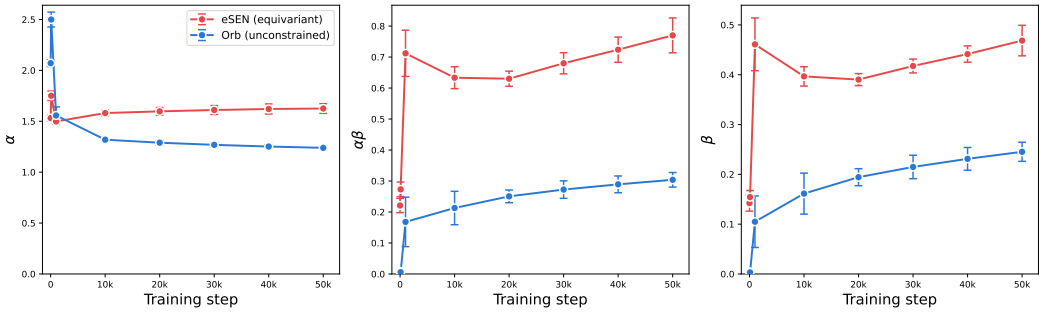

Figure 1: **Empirical NTK exponents early in training illustrates that gap between architectures persists.** Capacity exponent α (left), target variance decay rate $\alpha\beta$ (middle), and the source exponent β (right), estimated from the empirical force NTK (equation 7) at checkpoints over the first 50K training steps of eSEN (red, equivariant) and Orb (blue, unconstrained), using 3M-5M parameter models. After a fast initial phase ($<$ 10K steps), α is nearly constant (≈ 1.6 for eSEN, ≈ 1.3 for Orb), while β drifts slowly upward for both architectures, indicating feature learning happens in both architectures. eSEN has the larger $\alpha\beta$ and β at every checkpoint after initialization. The exponents fitted at these early steps match the exponents of neural scaling laws when the models are trained upto *1M training steps*, as discussed in section 4.3.

4.2 Estimating α and β from Empirical NTKs

Both exponents α and β are estimated from a single early checkpoint and a fixed set of held-out validation atoms, without relying on the training curve. This is possible because the empirical NTK changes rapidly during a short initial phase of learning and only slowly afterwards Fort et al. (2020); Atanasov et al. (2022), so an early checkpoint can capture the kernel that governs most of training. Figure 1 shows the exponents across early checkpoints. After the fast phase (before 10K steps), the kernel spectrum decay rate α is stable for eSEN(≈ 1.6) and Orb (≈ 1.3). Meanwhile, the β exponents drifts upward from initialization (step = 0) for both architectures, indicating that feature learning happens in both architectures on atomistic data. Crucially for our purpose, the gap between architectures and their orderings stay fixed across steps, and thus the comparison does not depend on which checkpoint is used. We therefore use the exponents at step $t \approx 10^4$, which is the earliest one past the fast phase, corresponding to rougly $1 - 2\%$ of the full training run of $\approx 1M$ steps.

Figure 2 shows, for each architecture at $t \approx 10^4$, the empirical NTK eigenvalue spectrum and the target variance across its eigenmodes. While the eigenvalue spectra of the two architectures are similar, the target variance of the equivariant network (eSEN) decays substantially faster beyond the leading modes than that of the unconstrained network (Orb). This indicates that the equivariant network's kernel aligns better with the force labels on this force-field task. Power-law fits to the eigenvalue spectrum and to the target-variance tail give, respectively, $\alpha = 1.58\ [1.56, 1.60]$ and $\alpha\beta = 0.63\ [0.61, 0.66]$ for eSEN (i.e., $\beta \approx 0.40$), compared with $\alpha = 1.32\ [1.31, 1.33]$ and $\alpha\beta = 0.21\ [0.17, 0.25]$ for Orb ($\beta \approx 0.16$) (brackets denote 95% confidence intervals over nine random seeds). In Section 4.3, we show that these exponents account for the model-size exponents and the ordering of the training-time exponents in the empirical scaling laws of both networks.

4.3 Empirical Scaling Laws

Matching compute. Because the two architectures perform different amounts of computation per parameter, we measure a FLOP coefficient κ such that training on D atoms costs $C \approx \kappa PD$ (FLOPs) for each, in analogy to the $6PD$ rule for transformers (Kaplan et al., 2020). Given a compute budget C, this fixes the number of parameters $P \approx C/(\kappa D)$

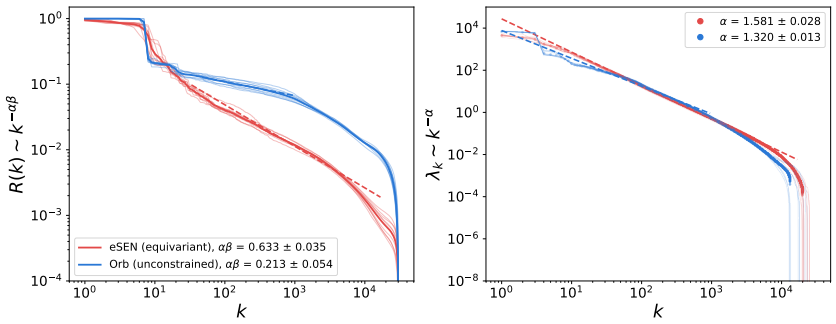

Figure 2: **The two kernels have similar spectra but differ sharply in their alignment with the force labels.** Empirical force NTK at $t \approx 10^4$ for eSEN (red) and Orb (blue), computed on $n \approx 10K$ held-out validation atoms (a $3n \times 3n$ Gram matrix). *Left:* residual label variance $R(k)$ (equation 9), i.e., the fraction of reference-force variance not captured by the top-k kernel eigenvectors. eSEN's residual decays much faster, so its leading eigendirections capture more of the force targets. *Right:* eigenvalue spectrum λ_k, which decays at similar rates for both architectures. Dashed lines are power-law fits $R(k) \propto k^{-\alpha\beta}$ and $\lambda_k \propto k^{-\alpha}$ over $k \in [k_{\min}, k_{\max}]$, giving $\beta \approx 0.4$ for eSEN and $\beta \approx 0.16$ for Orb. The measurements are conducted over nine random seeds.

that can be trained on D atoms, ensuring a compute-matched comparison. We determine κ by varying P and D and estimating the training FLOPs via `torch.profiler`. Figure 3 shows that fit of κ for both architectures. eSEN's κ is roughly three times larger than that of Orb. Both are 39 and 12 times larger than a usual transformer trained on this task, meaning that training a $3M$-parameter eSEN is as much as training a 117M-parameter transformer.

Scaling laws. At $D = 93M$, we trained eSEN models ranging from 800K to 3M parameters and Orb models ranging from 1M to 86M parameters. We trained each model for 15-20 epochs. These model sizes are close to those reported in public benchmarks where eSEN models have roughly 6M-50M parameters and Orb reportedly has about 25M. Because Orb is not rotationally equivariant, we trained it with rotational data augmentation. This setup keeps us close to the regime described in equation 3.

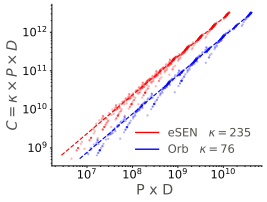

Figure 3: Estimating κ per architecture.

We fitted the validation loss to the scaling law $\mathcal{L}(P,t) = \mathcal{L}_\infty + A_P P^{-a_P} + A_t t^{-a_t}$, where P is the parameter count and t is the number of gradient steps. Figure 4 and Table 1 show the fits for both architectures; brackets denote 95% confidence intervals. The results highlight three points:

- **Irreducible floor.** $\mathcal{L}_\infty \approx 0$ for both architectures, consistent with ab initio force labels being a deterministic function of the atomic configuration within the Born-Oppenheimer approximation, and hence practically noise-free (Unke et al., 2021; Chmiela et al., 2023).
- **Model-size scaling.** The fitted exponents are $a_P = 0.21$ $[0.20, 0.22]$ for Orb and $a_P = 0.71$ $[0.62, 0.84]$ for eSEN. Both closely agree with the theoretical prediction $a_P = \alpha\beta$, which gives 0.21 $[0.17, 0.25]$ for Orb and 0.63 $[0.61, 0.66]$ for eSEN (Section 4.2).
- **Training-time scaling and feature learning.** The fitted exponents are $a_t = 0.61$ $[0.56, 0.67]$ for Orb and $a_t = 0.75$ $[0.62, 0.92]$ for eSEN. Their ordering matches that of β, since $\beta_{\text{eSEN}} \approx 0.43 > \beta_{\text{Orb}} \approx 0.16$, reflecting the better alignment of the eSEN kernel with the force labels early in training. For hard tasks ($\beta < 1$), the theoretical model in Section 3 predicts that feature learning yields a training-time exponent of at least $2\beta/(1+\beta)$, which is about 0.27 for Orb and 0.60 for eSEN. Both measured exponents exceed these values, and hence also exceed their initial β, consistent with feature learning in both architectures.

Predictability at the largest data size. To test whether the early-checkpoint exponents predict scaling beyond the setting above, we trained four architectures at $D = 930$M atoms, including Orb, eSEN, a vanilla MPNN, and a multi-channel EGNN (Satorras et al., 2021; Levy et al., 2023). The vanilla MPNN is unconstrained like Orb but lacks its separate edge-feature stream (Neumann et al., 2024), making it much cheaper ($\kappa \approx 6$ vs. 76 for Orb). EGNN is a lower-order equivariant network

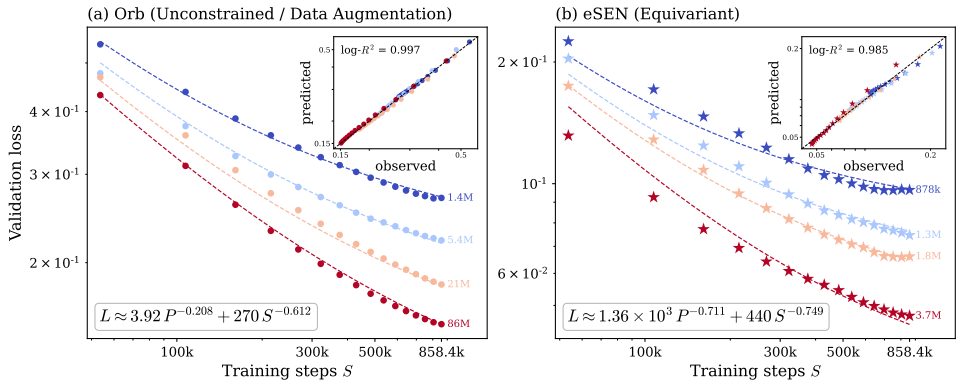

Figure 4: **eSEN scales more steeply than Orb with model size and has a larger fitted training-time exponent.** Validation loss on the 93M-atom subset versus training step for (a) Orb and (b) eSEN, fitted jointly to $\mathcal{L}(P,t) = \mathcal{L}_\infty + A_P P^{-a_P} + A_t t^{-a_t}$, where P is the parameter count and t the number of gradient steps. Markers are evaluations, dashed lines are the fitted law, colors denote model size, and each curve is labeled with its parameter count. The fitted law is written in each panel ($\mathcal{L}_\infty \approx 0$ for both). **Insets:** predicted versus observed loss for the same points; log-R^2 is the coefficient of determination on log-losses. The fitted exponents are $a_P = 0.71$ and $a_t = 0.75$ for eSEN versus $a_P = 0.21$ and $a_t = 0.61$ for Orb.

with separate scalar ($\ell = 0$) and vector ($\ell = 1$) streams. Following Ngo & Ravanbakhsh (2026), we trained each model for a single epoch under this infinite data regime, and fitted the final validation loss to $\mathcal{L} = A_P P^{-a_P} + \mathcal{L}_{/P}$, where $\mathcal{L}_{/P}$ collects the terms that do not depend on model size, such as the training-time and irreducible losses. For every architecture, the model-size component $\mathcal{L} - \mathcal{L}_{/P}$ follows a clean power law, as shown in figure 5, and its fitted exponent a_P at the final training steps $\approx 550K$ with a batch size of 64 agrees with the prediction $\alpha\beta$ from the 10K-step checkpoint. The comparison also separates two architectural effects. Orb and the vanilla MPNN follow parallel lines; Orb's edge-feature stream lowers the loss but leaves the exponent unchanged. In contrast, going from EGNN ($\ell_{\max} = 1$) to eSEN ($\ell_{\max} \geq 2$) increases the exponent itself. The early-checkpoint kernel therefore distinguishes design choices that only shift the loss from those that change how it scales.

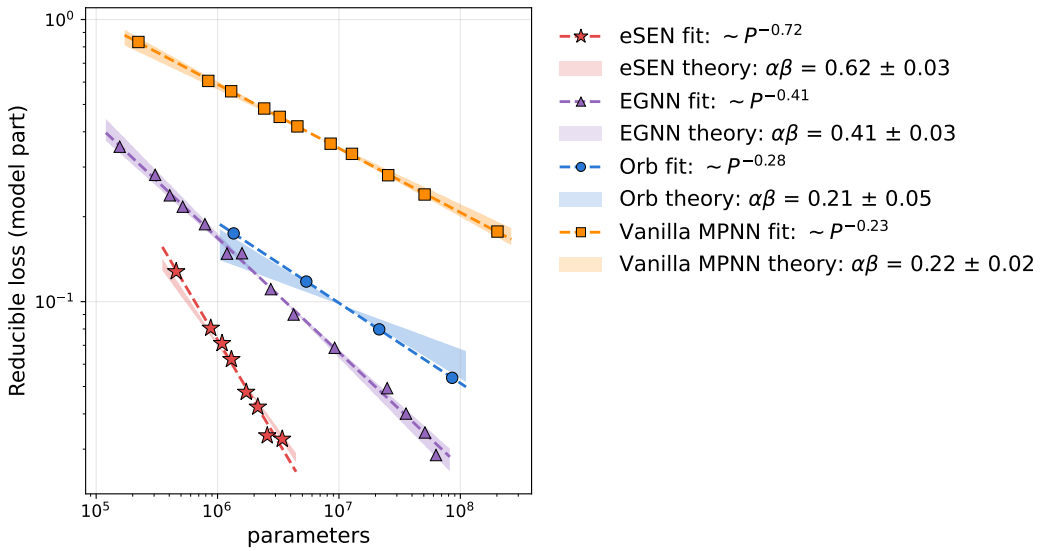

Figure 5: **Early-checkpoint NTK exponents track model-size scaling across four architectures.** Model-size component of the final validation loss, $\mathcal{L} - \mathcal{L}_{/P}$, versus parameter count P for models trained for a single epoch on the full 930M-atom set. The four architectures are eSEN (higher-order equivariant), EGNN (equivariant), and Orb and a vanilla MPNN (both unconstrained). Markers are individually trained models, and dashed lines are power-law fits $A_P P^{-a_P}$. Shaded bands show the slope predicted by $\alpha\beta$ from the empirical NTK at step 10^4 (1% of training budget). Orb and the vanilla MPNN have nearly parallel fits, so Orb's edge-feature stream lowers the loss without changing the exponent. By contrast, going from EGNN to eSEN increases the exponent itself.

5 Related Work

5.1 Neural Scaling Laws

Previous studies empirically showed that the performance of neural networks improves predictably as power laws in model size, dataset size, and compute (Ahmad & Tesauro, 1988; Henighan et al., 2020; Hoffmann et al., 2022; Kaplan et al., 2020; Sharma & Kaplan, 2022). Based on these observations, subsequent theoretical works aimed to explain the origins of these power laws. One line of work draws on approximation theory and the manifold-fitting hypothesis, suggesting that when networks fit data manifolds, model-size power laws are determined by the data's intrinsic dimension (Sharma & Kaplan, 2022; Bahri et al., 2024). Measuring the intrinsic dimension of 3D atomistic data remains elusive. This is because while the rotation group is only three-dimensional, the degrees of freedom of n-atom molecules are approximately $3n$. A second line of work, on which we build, models the network as kernel or random-feature regression and traces the exponents to spectral properties of the kernel and of the target (Maloney et al., 2022; Bordelon et al., 2024; Paquette et al., 2024; Lin et al., 2024). Under the source and capacity conditions of kernel regression (Caponnetto & Vito, 2005; Pillaud-Vivien et al., 2018), the decay of the kernel spectrum and the alignment of the target with its leading eigendirections jointly determine the exponents in training time, data, and parameters, and Bordelon et al. (2025) show that feature learning can further improve the time exponent when the target is hard, which also was empirically observed in (Vyas et al., 2023). Measuring source and capacity exponents for practical architectures on large datasets, and using them to explain observed exponents, remains largely unexplored.

6 Discussions, Limitations and Future Directions

This paper studies why MLIP architectures scale differently through the lens of kernel–task alignment. We find that the higher-order equivariant network has steeper scaling exponents than its unconstrained counterpart, and our results indicate that this gap stems from better alignment of its empirical NTK with the equivariant force labels rather than from differences in kernel spectra. Moreover, exponents estimated from an early checkpoint predict the fitted model-size exponents for all architectures we study. This suggests a principled way to compare MLIP architectures before committing to expensive full-scale training, or even to avoid such experiments altogether if the goal is to identify a promising new architecture in academic settings, where the compute cost may not be viable.

Overall, we believe our use of empirical NTK for analysis and probing in equivariant networks may be useful in other context beyond the scaling law studies. For example, recent work probes the MLIPs representations, using atom-level embeddings to predict molecular properties (Kim et al., 2025), condition generative models (Edamadaka et al., 2026), and analyze what the network encodes (Lee et al., 2024). Li & Walsh (2026) and Edamadaka et al. (2025) further find a shared representation space across equivariant and non-equivariant MLIPs. To our knowledge, these analyses probe only the *invariant* features of equivariant networks. For equivariant targets such as forces, this characterizes a truncated model, and probes of the full network, such as the empirical NTK, are more meaningful.

Our analysis assumes power-law decay in the kernel spectrum and the target variance, and relies on empirical fits to each network's NTK at an early checkpoint. Understanding which factors shape these power laws across architectures is an important direction. Furthermore, our results are established for supervised force-field pretraining in the regime targeted by compute-optimal scaling analyses. We do not consider other ways to improve unconstrained networks. First, our fits indicate strong feature-learning behaviours in the networks. This suggests that the near-term improvement is generating more DFT data. The second relevant direction is infinite-compute pretraining, where well-regularized, overparameterized models achieve better scaling on language tasks (Kim et al., 2026). Finally, the relation between our findings and self-supervised or generative pretraining (Joshi et al., 2025; Morehead et al., 2026; Perez & Gomez-Bombarelli, 2026) is an interesting direction for future work, since Neumann et al. (2024) report that a denoising pretraining stage substantially improves Orb and it partly explains its strong performance on public benchmarks.

AI USAGE STATEMENT

Generative AI tools were used to assist with writing experiment and analysis code, interpreting results, refining hypotheses, retrieving related literature, and preparing figures and layout. For the manuscript, AI tools were used to proofread the text, improve clarity, and correct grammar. All suggested references were checked by the authors for relevance and accuracy. The research questions and core ideas were conceived by the authors, who take full responsibility for the content.

REFERENCES

Subutai Ahmad and Gerald Tesauro. Scaling and generalization in neural networks: A case study. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988. URL https://proceedings.neurips.cc/paper_files/paper/1988/file/d1f491a404d6854880943e5c3cd9ca25-Paper.pdf.

Brandon Anderson, Truong-Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In *Conference on Neural Information Processing (NeurIPS)*, 2019.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1NvflqAdoom.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C Lawrence Zitnick, and Zachary W Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models, arxiv, 2024. *arXiv preprint arXiv:2410.12771*, 10, 2025.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4a36c3c51af11ed9f34615b81edb5bbc-Paper-Conference.pdf.

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), May 2022. doi: 10.1038/s41467-022-29939-5. URL https://doi.org/10.1038/s41467-022-29939-5.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, Apr 2007. doi: 10.1103/PhysRevLett.98.146401. URL https://link.aps.org/doi/10.1103/PhysRevLett.98.146401.

Vivek Bharadwaj, Austin Glover, Aydın Buluç, and James Demmel. An efficient sparse kernel generator for o (3)-equivariant deep networks. In *2025 Proceedings of the Conference on Applied and Computational Discrete Algorithms (ACDA)*, pp. 32–46. SIAM, 2025.

Filippo Bigi, Marcel F. Langer, and Michele Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=OEl3L8osas.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/bordelon24a.html.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dEypApI1MZ.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e(3) equivariant message passing. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=_xwr8gOBeV1.

Johann Brehmer, Sönke Behrends, Pim De Haan, and Taco Cohen. Does equivariance matter at scale?, 2025. URL https://openreview.net/forum?id=iIWeyfGTof.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1. URL https://doi.org/10.1038/s41467-021-23103-1.

Andrea Caponnetto and Ernesto De Vito. Fast rates for regularized least-squares algorithm, April 2005. URL http://hdl.handle.net/1721.1/30539.

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):eadf0873, 2023. doi: 10.1126/sciadv.adf0873. URL https://www.science.org/doi/abs/10.1126/sciadv.adf0873.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10131–10143. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/543bec10c8325987595fcdc492a525f4-Paper.pdf.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Error scaling laws for kernel classification under source and capacity conditions. *Machine Learning: Science and Technology*, 4(3):035033, aug 2023. doi: 10.1088/2632-2153/acf041. URL https://doi.org/10.1088/2632-2153/acf041.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 9974–10007. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/136b9a13861308c8948cd308ccd02658-Paper-Conference.pdf.

Alexandre Agm Duval, Victor Schmidt, Alex Hernández-García, Santiago Miret, Fragkiskos D. Malliaros, Yoshua Bengio, and David Rolnick. FAENet: Frame averaging equivariant GNN for materials modeling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9013–9033. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/duval23a.html.

Sathya Edamadaka, Soojung Yang, Ju Li, and Rafael Gómez-Bombarelli. Universally converging representations of matter across scientific foundation models. *arXiv preprint arXiv:2512.03750*, 2025.

Sathya Edamadaka, Krithik Ramesh, Ju Li, and Rafael Gómez-Bombarelli. Atomistic language models understand and generate materials. *arXiv preprint arXiv:2606.21395*, 2026.

Ahmed A. A. Elhag, Arun Raja, Alex Morehead, Samuel M Blau, Garrett M Morris, and Michael M. Bronstein. Learning inter-atomic potentials without explicit equivariance. In *EurIPS 2025 Workshop on SIMBIOCHEM*, 2026. URL https://openreview.net/forum?id=sqXXsPReH5.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5850–5861. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/405075699f065e43581f27d67bb68478-Paper.pdf.

Nathan C Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gomez-Bombarelli, Connor W Coley, and Vijay Gadepally. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, 2023.

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. URL https://openreview.net/forum?id=R0PBjxIbgm.

Fabian Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *Conference on Neural Information Processing (NeurIPS)*, 2020.

William Fulton and Joe Harris. *Representation theory: a first course*. Springer Science & Business Media, 2013.

Meng Gao, Armin Shayesteh Zadeh, Aniruddha Seal, Siva Dasetty, Siddarth K Achar, Misko Dzamba, Benjamin K Miller, Leif D Jacobson, C Lawrence Zitnick, Brandon M Wood, et al. Quantum-accurate atomistic modeling of enzyme catalysis using a machine learned potential. *arXiv preprint arXiv:2609.09293*, 2026.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Max Großmann, Malte Grunert, and Erich Runge. Broken neural scaling laws in learning the optical properties of solids. *PRX Intelligence*, 1:013011, Aug 2026. doi: 10.1103/wv8r-7zr1. URL https://link.aps.org/doi/10.1103/wv8r-7zr1.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Chaitanya K. Joshi, Cristian Bodnar, Simon V Mathis, Taco Cohen, and Pietro Lio. On the expressive power of geometric graph neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun

Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15330–15355. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/joshi23a.html.

Chaitanya K. Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary Ward Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28393–28417. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/joshi25a.html.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ck0aZTAnwK.

SY Kim, YJ Park, and J Li. Leveraging neural network interatomic potentials for a foundation model of chemistry (2025). *arXiv preprint arXiv:2506.18497*, 2025.

Risi Kondor. The principles behind equivariant neural networks for physics and chemistry. *Proceedings of the National Academy of Sciences*, 122(41):e2415656122, 2025. doi: 10.1073/pnas.2415656122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2415656122.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. In *Advances in Neural Information Processing Systems 32*, pp. 10117–10126, 2018.

Kin Long Kelvin Lee, Mikhail Galkin, and Santiago Miret. Deconstructing equivariant representations in molecular systems. *arXiv preprint arXiv:2410.08131*, 2024.

Seung Yul Lee, Hojoon Kim, Yutack Park, Dawoon Jeong, Seungwu Han, Yeonhong Park, and Jae W. Lee. FlashTP: Fused, sparsity-aware tensor product for machine learning interatomic potentials. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33143–33156. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lee25l.html.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The open molecules 2025 (omol25) dataset, evaluations, and models, 2025. URL https://arxiv.org/abs/2505.08762.

Daniel Levy, Sékou-Oumar Kaba, Carmelo Gonzales, Santiago Miret, and Siamak Ravanbakhsh. Using multiple vector channels improves e (n)-equivariant graph neural networks. *arXiv preprint arXiv:2309.03139*, 2023.

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

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KwmPfARgOTD.

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. URL https://openreview.net/forum?id=mCOBKZmrzD.

Yi-Lun Liao, Alexander J Hoffman, Sabrina C Shen, Alexandre Duval, Sam Walton Norwood, and Tess Smidt. Equiformerv3: Scaling efficient, expressive, and general se (3)-equivariant graph attention transformers. *arXiv preprint arXiv:2604.09130*, 2026.

Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason Lee. Scaling laws in linear regression: Compute, parameters, and data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 60556–60606. Curran Associates, Inc., 2024. doi: 10.52202/079017-1937. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6fcb1afcc1e9c2c82c8ddddf03bcf0f6-Paper-Conference.pdf.

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. URL https://openreview.net/forum?id=mhyQXJ6JsK.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Alex Morehead, Miruna Cretu, Antonia Panescu, Rishabh Anand, Maurice Weiler, Tynan Perez, Samuel Blau, Steven Farrell, Wahid Bhimji, Anubhav Jain, et al. Zatom-1: Towards a multimodal foundation model for 3d molecules and materials. *arXiv preprint arXiv:2602.22251*, 2026.

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. *arxiv preprint arxiv:2204.05249*, 2022.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential, 2024. URL https://arxiv.org/abs/2410.22570.

Khang Ngo and Siamak Ravanbakhsh. Scaling laws and symmetry, evidence from neural force fields. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qyjaVda7t2.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 16459–16537. Curran Associates, Inc., 2024. doi: 10.52202/079017-0526. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1dccfc3ee01871d05e33457c61037d59-Paper-Conference.pdf.

Saro Passaro and C. Lawrence Zitnick. Reducing SO(3) convolutions to SO(2) for efficient equivariant GNNs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27420–27438. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/passaro23a.html.

Zihan Pengmei, Zhengyuan Shen, Zichen Wang, Marcus D. Collins, and Huzefa Rangwala. Pushing the limits of all-atom geometric graph neural networks: Pre-training, scaling, and zero-shot transfer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4S2L519nIX.

Tynan Perez and Rafael Gomez-Bombarelli. Self-conditioned denoising for atomistic representation learning. *arXiv preprint arXiv:2603.17196*, 2026.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/10ff0b5e85e5b85cc3095d431d8c08b4-Paper.pdf.

Sergey Pozdnyakov and Michele Ceriotti. Smooth, exact rotational symmetrization for deep learning on point clouds. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79469–79501. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fb4a7e3522363907b26a86cc5be627ac-Paper-Conference.pdf.

Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zIUyj55nXR.

Eric Qu and Aditi S. Krishnapriyan. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 139030–139053. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fad8e1915f66161581bb127ccf01092e-Paper-Conference.pdf.

Eric Qu, Brandon M Wood, Aditi S Krishnapriyan, and Zachary W Ulissi. A recipe for scalable attention-based mlips: unlocking long-range accuracy with all-to-all node attention. *arXiv preprint arXiv:2603.06567*, 2026.

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

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Conference on Neural Information Processing (NeurIPS)*, 2017.

Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022. URL http://jmlr.org/papers/v23/20-1111.html.

Max W Shen, Ewa Nowara, Michael Maser, and Kyunghyun Cho. Training dynamics of learning 3d-rotational equivariance. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=DLOIAW18W3. J2C Certification.

J. S. Smith, O. Isayev, and A. E. Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical Science*, 8(4):3192–3203, 04 2017. ISSN 2041-6520. doi: 10.1039/c6sc05720a. URL https://doi.org/10.1039/c6sc05720a.

Behrooz Tahmasebi and Stefanie Jegelka. The exact sample complexity gain from invariances for kernel regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55616–55646. Curran Associates, Inc., 2023. doi: 10.52202/075280-2427. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/adf5a38a2e2e7606fbfc3eff72998afa-Paper-Conference.pdf.

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. *arXiv preprint arXiv:1802.08219*, 2018.

Oliver T Unke and Hartmut Maennel. E3x: E(3)-equivariant deep learning made easy. *arXiv preprint arXiv:2401.07595*, 2024.

Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 08 2021. ISSN 0009-2665. doi: 10.1021/acs.chemrev.0c01111. URL https://doi.org/10.1021/acs.chemrev.0c01111.

Nikhil Vyas, Yamini Bansal, and Preetum Nakkiran. Empirical limitations of the NTK for understanding scaling laws in deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=Y3saBb7mCE.

Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, et al. Uma: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

Yuqing Xie, Ameya Daigavane, Mit Kotak, and Tess Smidt. The price of freedom: Exploring expressivity and runtime tradeoffs in equivariant tensor products. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68599–68625. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/xie25c.html.

Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17084–17097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8df7c2e3c3c3be098ef7b382bd2c37ba-Paper.pdf.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

A More on Related Work

A.1 Machine Learning Interatomic Potentials.

Early MLIPs modeled the potential energy surface by fitting multi-layer perceptrons (MLPs) on hand-crafted invariant descriptors (Behler & Parrinello, 2007; Smith et al., 2017). Gilmer et al. (2017); Schütt et al. (2017) extended this approach by building message-passing neural networks that exchange interatomic distances and scalar node features. To respect rotational symmetry exactly, Kondor et al. (2018); Anderson et al. (2019) built $E(3)$/$SE(3)$-equivariant neural networks from irreducible representations and tensor products. Batzner et al. (2022) showed that such higher-order features are more data-efficient for learning force fields than invariant ones. Subsequent work refined these networks with non-linear messages (Brandstetter et al., 2022), strict locality (Musaelian et al., 2022), many-body interactions through tensor contractions (Batatia et al., 2022), and equivariant attention (Fuchs et al., 2020; Liao & Smidt, 2023; Liao et al., 2024; 2026). Although highly accurate, higher-order networks built upon tensor products are expensive to train and run, because the cost of $SO(3)$ tensor products grows rapidly with the maximum degree $\ell_{\max}$ of the features. Recent work has attempted to make $SO(3)$ operations cheaper by reducing them to $SO(2)$ convolutions via an edge-alignment trick (Passaro & Zitnick, 2023), writing dedicated GPU kernels (Lee et al., 2025; Bharadwaj et al., 2025), or developing more efficient tensor products, such as Gaunt tensor products (Luo et al., 2024) or matrix tensor products (Unke & Maennel, 2024). Xie et al. (2025) showed a considerable trade-off between expressivity and efficiency among these tensor products. EquiformerV2 (Liao et al., 2024), eSEN (Fu et al., 2025), UMA (Wood et al., 2025), and most recently EquiformerV3 (Liao et al., 2026) build on these advances, with eSEN and its successors additionally enforcing smooth cutoffs and strict equivariance so that the learned potential energy surface is smooth and energy-conserving. A second family of MLIPs does not build equivariance into the network. Instead, equivariance is imposed or encouraged through symmetrization techniques, such as frame averaging (Puny et al., 2022; Duval et al., 2023), test-time symmetrization (Pozdnyakov & Ceriotti, 2023), or symmetry-regularizing losses (Elhag et al., 2026). With the rapid growth of data and compute in recent years, fully unconstrained models built upon standard matrix multiplications and attention have emerged as strong competitors, including Orb (Neumann et al., 2024; Rhodes et al., 2025), PET (Pozdnyakov & Ceriotti, 2023), and EScAIP (Qu & Krishnapriyan, 2024; Qu et al., 2026). These networks learn approximate equivariance from data augmentation and achieve competitive performance across benchmarks. Note that PET and EScAIP apply attention over edges rather than atoms, which increases their cost on large systems. The two architectures in our study, eSEN and Orb, represent these two families.

A.2 Scaling Laws for MLIPs and Equivariance

For MLIPs and equivariance specifically, Frey et al. (2023) first reported power-law improvements of deep chemical models with data and model size, and later studies documented scaling for multi-domain (Wood et al., 2025) settings, for unconstrained attention-based models (Qu & Krishnapriyan, 2024), and for property prediction tasks where the power laws break (Großmann et al., 2026). Closest to our question, Brehmer et al. (2025) compare equivariant and non-equivariant transformers under a compute budget on rigid-body dynamics, and (Ngo & Ravanbakhsh, 2026) showed that higher-order equivariant MLIPs exhibit steeper scaling exponents than their lower-order equivariant or unconstrained counterparts under single-epoch training over effectively large datasets. In addition, Shen et al. (2025) study how approximate equivariance is acquired under augmentation, and showed that equivariance is an easy learning task, where unconstrained networks can achieve low equivariance error at early steps of the training phase. On the theoretical side, invariance is known to reduce the sample complexity of kernel regression (Tahmasebi & Jegelka, 2023), but these results concern idealized invariant kernels rather than the kernels of trained networks, and they do not explain why two architectures trained on the same task exhibit different exponents. Our work connects these threads. Particularly, we measure source and capacity exponents directly from the empirical NTK of practical equivariant and unconstrained MLIPs and show that they account for the exponents observed in our fitted neural scaling laws.

B Confidence Intervals for Neural Scaling Laws

Table 1: Scaling-law fits of $\mathcal{L}(P,t) = \mathcal{L}_\infty + A_P\, P^{-a_P} + A_t\, t^{-a_t}$ to the validation loss on the 93M training atoms. P is parameter count, t is the number of training steps, and $\mathcal{L}_\infty$ is irreducible error. The fitted results are reported with 95% confidence interval.

Architecture	Coefficient	Estimate	95% CI
Unconstrained (Orb)	A_P	3.92	[3.42, 4.52]
	a_P	0.21	[0.20, 0.22]
	A_t	270	[164, 497]
	a_t	0.61	[0.56, 0.67]
	$\mathcal{L}_\infty$	≈ 0	-
Equivariant (eSEN)	A_P	1.4×10^3	$[4.1 \times 10^2, 7.3 \times 10^3]$
	a_P	0.71	[0.62, 0.84]
	A_t	440	$[106, 3.1 \times 10^3]$
	a_t	0.75	[0.62, 0.92]
	$\mathcal{L}_\infty$	≈ 0	-

C Experimental Details

Training. Molecular structures are converted to radius graphs with a cutoff of 6 Å and at most 32 neighbours per atom, using the same preprocessing for all architectures. Following Ngo & Ravanbakhsh (2026), we tune the learning rate on the base model of each architecture, obtaining $\eta^\star = 10^{-3}$, and transfer it to larger widths with μP (Yang et al., 2021). All models are trained with a batch size of 64 using Schedule-Free AdamW (Defazio et al., 2024), which replaces the learning-rate schedule with an average of the iterates, after a linear warmup of 100 steps. This avoids the dependence of fitted scaling laws on a predefined decay schedule, which Hoffmann et al. (2022) found to bias the estimated exponents when the schedule length does not match the training length. For each molecule X with n atoms, we minimize the sum of the per-atom energy error and the per-atom force error (Fu et al., 2025; Wood et al., 2025):

$$\mathcal{L}(\theta, X) = \frac{\lambda_e}{n} \left| e_\theta(X) - e(X) \right| + \frac{\lambda_f}{n} \sum_{i=1}^{n} \left\| f_{\theta,i}(X) - f_i(X) \right\|_2, \tag{11}$$

where $\lambda_e, \lambda_f > 0$ set the relative weight of the two terms; we use $\lambda_e = \lambda_f = 10$. Energies are first referenced by subtracting a per-element linear fit to the training set, then standardised as $e' = (e - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the referenced energies over the training set. Since forces $f_i = -\partial e/\partial x_i$ are unchanged by the reference subtraction, only the scaling applies to them, $f_i' = f_i/\sigma$ [2].

Scaling strategy. Under μP parameterization, we vary the width of each network and keep its depth fixed at the value found to work best in prior work, namely five layers for the unconstrained networks (Ngo & Ravanbakhsh, 2026; Rhodes et al., 2025) and 12 for the equivariant ones (Fu et al., 2025). The larger depth of equivariant networks is consistent with the view that equivariant layers propagate geometric information beyond local neighbourhoods, so that additional layers increase expressivity (Joshi et al., 2023), whereas unconstrained MPNNs gain little from depth beyond a few layers (Ngo & Ravanbakhsh, 2026).

Empirical NTK measurement. For each architecture we take a checkpoint at step $t \approx 10^4$ of around 3M-5M parameters and compute the Jacobian of the predicted forces with respect to all parameters on $n \approx 10$K held-out validation atoms, giving the $3n \times 3n$ Gram matrix K of equation equation 7. We diagonalise K and fit $\lambda_k \propto k^{-\alpha}$ and $R(k) \propto k^{-\alpha\beta}$ (equation equation 9) by least squares on log–log axes. The theory assumes a clean power law over all k, which real kernels do not satisfy: the leading modes and the finite-sample tail both deviate from it. We therefore empirically fit over the range $[k_{\min}, k_{\max}]$ in which the log–log slope is stable for each architecture.

[2] https://github.com/facebookresearch/fairchem