The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu
UC Berkeley
ericqu@berkeley.edu

Aditi S. Krishnapriyan
UC Berkeley, LBNL
aditik1@berkeley.edu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as surrogate models for *ab initio* quantum mechanical calculations, predicting the energy and forces between atoms in molecules and materials based on atomic configurations. The dominant paradigm in this field is to incorporate numerous physical domain constraints into the model, such as symmetry constraints like rotational equivariance. We contend that these increasingly complex domain constraints inhibit the scaling ability of NNIPs, and such strategies are likely to cause model performance to plateau in the long run. In this work, we take an alternative approach and start by systematically studying NNIP scaling properties and strategies. Our findings indicate that scaling the model through attention mechanisms is both efficient and improves model expressivity. These insights motivate us to develop an NNIP architecture designed for scalability: the Efficiently Scaled Attention Interatomic Potential (EScAIP). EScAIP leverages a novel multi-head self-attention formulation within graph neural networks, applying attention at the neighbor-level representations. Implemented with highly-optimized attention GPU kernels, EScAIP achieves substantial gains in efficiency—at least 10x speed up in inference time, 5x less in memory usage—compared to existing NNIP models. EScAIP also achieves state-of-the-art performance on a wide range of datasets including catalysts (OC20 and OC22), molecules (SPICE), and materials (MPTrj). We emphasize that our approach should be thought of as a *philosophy* rather than a specific model, representing a proof-of-concept towards developing general-purpose NNIPs that achieve better expressivity through scaling, and continue to scale efficiently with increased computational resources and training data.

1 Introduction

In recent years, the principle of scaling model size, data, and compute has become a key factor for improving performance and generalization in machine learning (ML), across fields from natural language processing (NLP) [Kaplan et al., 2020] to computer vision (CV) [Dosovitskiy et al., 2021, Zhai et al., 2022]. Scaling in ML is, in a large part, defined by the ability to best exploit GPU computing capabilities. This typically involves efficiently increasing model sizes to large parameter counts, as well as optimizing model training and inference to be optimally compute-efficient.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Parallel to these developments, ML models have also been rapidly developing for atomistic simulation, addressing problems in drug design, catalysis, materials, and more [Deringer et al., 2019, Unke et al., 2021]. Among these, machine learning interatomic potentials, and particularly neural network interatomic potentials (NNIPs), have gained popularity as surrogate models for computationally intensive *ab initio* quantum mechanical calculations like density functional theory. NNIPs are designed to predict the energies and forces of molecular systems with high efficiency and accuracy, allowing downstream tasks such as geometry relaxations or molecular dynamics to be carried out on systems that would be intractable to simulate directly with density functional theory.

Current NNIPs are predominantly based on graph neural networks (GNNs). The atomistic system is represented as a graph, where nodes correspond to atoms and edges representing interactions between atoms. Many effective models in this field have increasingly tried to embed physically-inspired constraints into the model, often justified by the belief that these constraints improve accuracy and data efficiency. Common constraints include incorporating predefined symmetries into the NN architecture, such as rotational equivariance, as well as using complex input feature sets.

NNIP models that integrate symmetry constraints [Batzner et al., 2022, Batatia et al., 2022, Liao et al., 2024] often rely on computationally intensive tensor products of rotation order L [Geiger and Smidt, 2022] to maintain rotational equivariance. Although recent advancements have reduced the computational complexity of these operations [Passaro and Zitnick, 2023a, Luo et al., 2024], the remaining computational overhead still significantly limits their scalability. Other approaches [Gasteiger et al., 2020a, 2021, 2022] use basis expansions of the edge directions, angles, and dihedrals as features. Generally, incorporating these constraints tends to be compute-inefficient. As a result, many of the models in the field remain highly-constrained and small, despite the availability of larger datasets [Chanussot et al., 2021, Jain et al., 2013] and more computational resources.

We contend that these increasingly complex domain constraints inhibit the scaling ability of NNIPs, and such strategies are likely to plateau over time in terms of model performance. As the scale of the models increase, we hypothesize that imposing these constraints hinders the learning of effective representations, restricts the model's ability to generalize, and impedes efficient optimization. Many of these feature-engineered approaches are not optimized for efficient parallelization on GPUs, further limiting their scalability and efficiency, especially when applied to larger systems.

In many other fields of ML, general-purpose architectures that best exploit computing capabilities outperform models with handcrafted, domain-specific constraints [Dosovitskiy et al., 2021, Zhai et al., 2022]. These observations motivate us to ask: **How can we develop principled methods and design choices that enable the creation of general-purpose neural network interatomic potentials that scale effectively with increased computational resources and training data?**

To answer this question, we conduct an initial ablation study to identify which components in NNIPs are most conducive to scaling. In NNIPs with built-in rotational equivariance, it is commonly believed that increasing the rotation order (L) improves model performance, even though it incurs additional computational cost. However, our investigations show that increasing the rotation order also adds more parameters to the model, and NNIPs are not always adjusted to account for this difference in parameter count. Our investigations also show that how parameters are added to the model is critical, as different types of parameter increases can differently impact the model's expressivity. We find that increasing the parameters of other components of the model besides the rotation order——particularly those involved in attention mechanisms——greatly improves model performance.

Based on these insights, we develop the Efficiently Scaled Attention Interatomic Potential (EScAIP), an NNIP architecture explicitly designed for scaling by incorporating highly optimized attention mechanisms. To the best of our knowledge, our model is the first to leverage attention mechanisms on the neighbor representations of atoms rather than only the nodes, resulting in more expressivity. We also leverage advancements in attention mechanisms [Lefaudeux et al., 2022], which have computational and memory efficiencies for scaling on large datasets.

Our model achieves the best performance on a wide range of chemical applications, including the top performance on the Open Catalyst 2020 (OC20), Open Catalyst 2022 (OC22), SPICE molecules, and Materials Project (MPTrj) datasets. It also demonstrates a 10x speed up in inference time and 5x less in memory usage compared to existing NNIP models. To analyze rotational equivariance, on the validation set, we 1) predict forces on a set of atomistic systems (A), 2) rotate the atomistic systems and predict forces (B), and then 3) compute the cosine similarity between

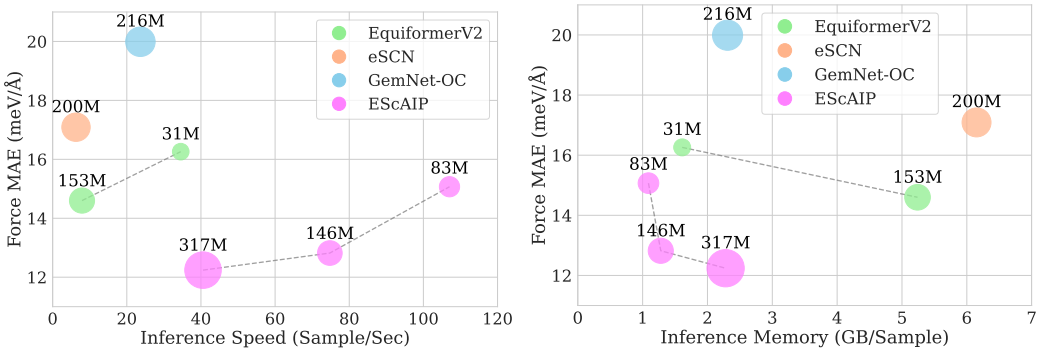

Figure 1: Efficiency, performance, and scaling comparisons between EScAIP and baseline models on the Open Catalyst dataset (OC20). Force MAE Error (meV/Å ↓) vs. Inference Speed (Sample/Sec ↑) and Force MAE vs. Memory (GB/Sample ↓) is reported. Results with Energy MAE can be found in the Appendix Fig. 7. EScAIP achieves better performance with smaller time and memory cost.

the force predictions (B) and the rotated version of force predictions (A). After training EScAIP on different datasets, we find that the cosine similarity is consistently ≥ 0.99, meaning EScAIP is essentially always predicting the rotations correctly. We also provide evidence that EScAIP scales well with compute, and is designed in such a way that it will further improve in efficiency as advances in GPU computing continue to increase. Our code and model checkpoints are publicly available at `https://github.com/ASK-Berkeley/EScAIP`.

2 Related Works

Neural Network Interatomic Potentials. There have been significant advancements in the development of neural network interatomic potentials (NNIPs), and we give a very general overview of the field. These models are usually trained to predict the system energy and per-atom force based on system properties, including atomic numbers and positions. We classify these models into two categories: (1) models that are based on Group Representation node features, and (2) models that are based on node features represented by Cartesian Coordinates. In the former, the node features are equivariant to different groups acting on the atomic positions, such as rotations and translations. In the latter, most architectures obey basic group symmetries, such as rotation and translation invariance.

- **Group Representation Architectures.** The first model that used group representation node features was the Tensor Field Network [Thomas et al., 2018], followed by an improved version, NequIP [Batzner et al., 2022]. Then, MACE [Batatia et al., 2022] incorporated the Atomic Cluster Expansion [Drautz, 2019] into the architecture. SCN [Zitnick et al., 2022] used spherical functions to represent equivariant node features, followed by an efficiency improvement in the tensor products, eSCN [Passaro and Zitnick, 2023b]. Equiformer [Liao and Smidt, 2022, Liao et al., 2024] incorporated graph attention into the architecture.

- **Cartesian Coordinates Architectures.** SchNet [Schütt et al., 2017] was the initial work that only used edge distances as input to maintain invariant node features. DimeNet [Gasteiger et al., 2020a,b] and GemNet [Gasteiger et al., 2021, 2022] added invariant bond direction feature sets as input. They designed an output head that maintains rotational equivariance with invariant node features. Another line of work tries to maintain equivariant features in Cartesian space by explicitly modeling spherical functions [Frank et al., 2022, Bekkers et al., 2024, Chen and Ong, 2022, Cheng, 2024, Haghighatlari et al., 2022].

Datasets for NNIP training. There has also been a growing focus in the NNIP domain on generating larger datasets with quantum mechanical simulations, and using this to train models. These datasets span domains such as molecules [Eastman et al., 2023, Smith et al., 2020], catalysts [Chanussot et al., 2021, Tran et al., 2023], and materials [Barroso-Luque et al., 2024, Yang et al., 2024, Merchant et al., 2023, Jain et al., 2013, Choudhary et al., 2020].

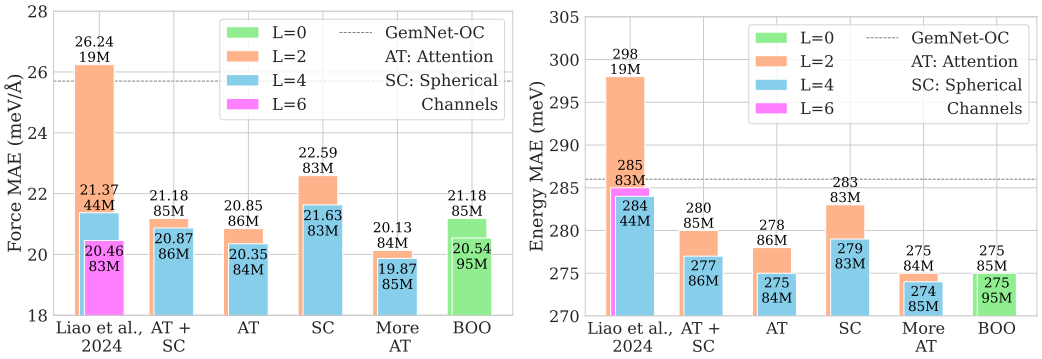

Figure 2: Results of ablation study of EquiformerV2 [Liao et al., 2024] on the OC20 2M dataset. Energy (eV) and force (eV/Å) mean absolute error (MAE) are reported, along with the model's parameter counts. The leftmost column shows the original results from [Liao et al., 2024], where different L had a different number of trainable parameters. We look at scaling parameters through the attention mechanisms (*AT*) and spherical channels (*SC*) for the original $L = 2$ and $L = 4$ models, such that the number of parameters is approximately equal to the original $L = 6$ model. Scaling parameters in different ways affects the overall energy and forces error, and increasing attention parameters is particularly effective in improving model performance (*More AT*). We also modify the architecture to be invariant ($L = 0$), allowing us to examine the effects of excluding rotational equivariance while controlling for the number of parameters (*BOO*). After controlling for parameter counts, many of the models have comparable error to the original $L = 6$ model.

Constrained vs. Unconstrained Architectures. There has been a trend of incorporating physically-inspired constraints into NNIP model architectures, such as all Group Representation Architectures that incorporate symmetry constraints into the model. However, there have been other lines of work that do not try to build in symmetry directly into the NN, and instead either try to "approximate" the symmetry [Pozdnyakov and Ceriotti, 2023, Wang et al., 2022, Finzi et al., 2021] or learn the symmetry via data augmentation techniques [Puny et al., 2022, Duval et al., 2023].

3 Investigation on How to Scale Neural Network Interatomic Potentials

We systematically investigate strategies for scaling neural network interatomic potential (NNIP) models through an ablation study. We examine how higher-order symmetries (rotation order L) impact scaling efficiency and identify the most effective methods for increasing model parameters (§3.1). We also assess the importance of incorporating directional bond features (§3.2). We conduct experiments using a leading NNIP architecture, the EquiformerV2 model [Liao et al., 2024], on the Open Catalyst 2020 (OC20) Dataset [Chanussot et al., 2021] 2M split to evaluate the performance of different scaling strategies.

3.1 Optimal Components for Scaling Neural Network Interatomic Potentials

A prevalent approach to improve the capability of NNIP models with group representation features is to increase the order of representations (L). Liao et al. [2024] did a study on the EquiformerV2 model, varying L to examine its impact on model performance. However, they did not control for the total number of trainable parameters in the model. This variation introduces discrepancies that can confound the true effect of L on the model's performance.

Ablation Study Settings. To clarify the impact of increasing L on model performance and determine the most effective strategy for increasing parameters in NNIP models, we conduct a parameter-controlled experiment using the EquiformerV2 model on the OC20 S2EF 2M dataset. We standardize the number of trainable parameters across different values of L to isolate the effects of increasing L, and systematically add parameters to different components of the original $L = 2$ and $L = 4$ EquiformerV2 models from Liao et al. [2024]. Our approach targets four distinct configurations: increasing parameters solely in the attention mechanisms (*AT*), solely in the spherical channels that act on all group representations in the NN (*SC*), evenly across both attention mechanisms and spherical

channels (*AT + SC*), and a configuration where spherical channels are reduced while significantly boosting attention parameters (*More AT*).

Results of Ablation Study. The comparative analysis reveals a clear hierarchy in performance gains with different parameter scaling strategies. The *More AT* configuration yields the highest performance improvement, followed by *AT*, *AT + SC*, and *SC*. The results, summarized in Fig. 2, show that once the number of parameters across models are controlled, many of the models have comparable error to the original $L = 6$ model. Increasing the parameters of the attention mechanisms is most beneficial and provides more substantial improvements than simply adding more parameters across all components.

3.2 Bond Directional Features

We explore what the most minimal representations are of the atomistic system to enable the model to learn a scalable, data-driven feature set, and find that incorporating bond directional information is useful for NNIP models. As opposed to other domains, such as social networks, the edges (or bonds) in molecular graphs possess distinct geometric attributes, i.e., pairwise directions. However, the raw value of the bond direction changes with the rotation and translation of the molecule, making it challenging to directly utilize these features in NNIP models.

We propose a straightforward and data-driven approach to embed the bond directional information. To avoid the computational inefficiency of taking a tensor product, we aim to use the simplest possible representation of bond direction that is rotationally invariant. Inspired by Steinhardt et al. [1983], we use an embedding of the Bond-Orientational Order (BOO) to represent the directional features. Formally, for a node v, the BOO of order l is,

$$\begin{aligned} \mathrm{BOO}^{(l)}(v) &= \sum_{m=-l}^{l} \sqrt{\frac{4\pi}{2l+1}\left|\frac{1}{n_v}\sum_{u\in\mathrm{Nei}(v)} Y_m^{(l)}(\hat{\boldsymbol{d}}_{uv})\right|^2}, \\ \mathrm{BOO}(v) &= \mathrm{Concat}\left(\{\mathrm{BOO}^{(l)}(v)\}_{l=0}^{L}\right), \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{d}}_{uv}$ is the normalized bond direction vector between node v and u, n_v is the number of neighbors of v, $\mathrm{Nei}(v)$ is the neighbors of v, $Y_m^{(l)}$ is the spherical harmonics of order l and degree m. This can be interpreted as the minimum-order rotation-invariant representation of the l-th moment in a multipole expansion for the distribution of bond vectors $\rho_{\mathrm{bond}}(\mathrm{n})$ across a unit sphere. In other words, BOO is the *simplest* way to encode the neighborhood directional information in a rotationally invariant manner. The BOO features $\mathrm{BOO}(v) \in \mathbb{R}^{L+1}$ for a node v is the concatenation of $\mathrm{BOO}(v)^{(l)}$. In theory, the BOO feature contains all the directional information of the neighborhood, and the embedding network can learn to extract such information.

Testing the bond-orientational order (BOO) features. We conduct a study to test the BOO features. We modify the EquiformerV2 model to be $L = 0$ and replace the spherical harmonics directional features with embeddings of the BOO features. The results are in Fig. 2. The $L = 0$ model achieves comparable results with the $L = 6$ model. This finding suggests that the BOO features are a straightforward and effective way to incorporate bond directional information in NNIP models, and that it is also possible to learn additional information solely through scaling.

4 Efficiently Scaled Attention Interatomic Potential (EScAIP)

We introduce a new NNIP architecture, Efficiently Scaled Attention Interatomic Potential (EScAIP), which leverages highly optimized self-attention mechanisms for expressivity, with design choices centered around scalability and efficiency. To avoid costly tensor products, we operate on scalar features that are invariant to rotations and translations. This enables us to take advantage the optimized self-attention mechanisms from natural language processing, making the model substantially more time and memory efficient than equivariant group representation models such as EquiformerV2 [Liao et al., 2024]. An illustration of our model is shown in Fig. 3. We describe the key components of the model and the motivation behind their design:

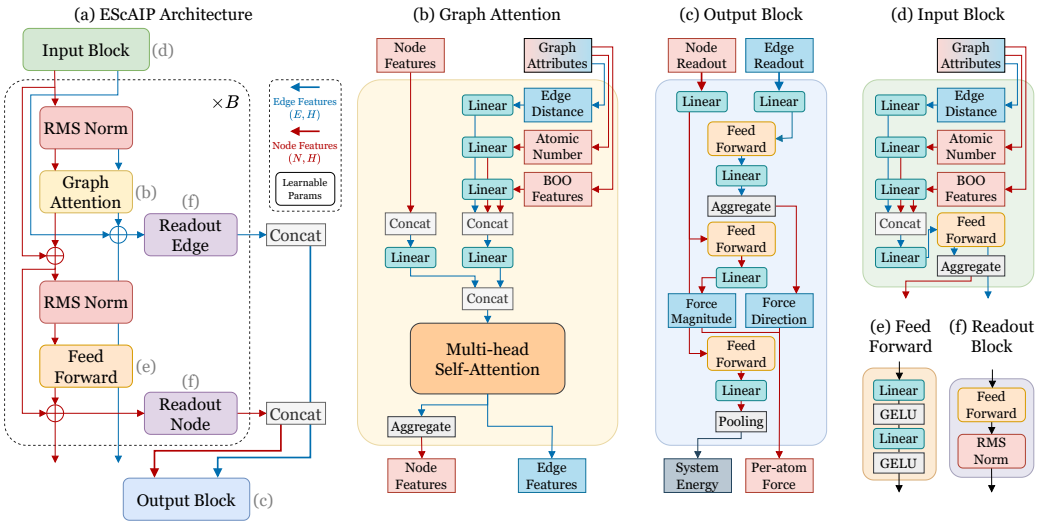

Figure 3: Illustration of the Efficiently Scaled Attention Interatomic Potential (EScAIP) model architecture. The model consists of B graph attention blocks (dashed box), each of which contains a graph attention layer, a feed forward layer, and two readout layers for node and edge features. The concatenated readouts from each block are used to predict per-atom forces and system energy.

Input Block. The input to the model is a radius-r graph representation of the molecular system. We use three attributes from the molecular graph as input: atomic numbers [Zitnick et al., 2022], Radial Basis Expansion (RBF) of pairwise distances [Schütt et al., 2017], and Bond Order Orientation (BOO) features from §3.2. The atomic numbers embeddings are used to encode the atom type information, while the RBF and BOO embeddings are used to encode the spatial information of the molecular system. These input attributes are the minimal representations of the system, enabling the model to learn a scalable, data-driven feature set. We also note that the attributes can be pre-computed, requiring minimal computational cost. The input features are then passed through a feed forward neural network (FFN) to produce the initial edge and node features.

Graph Attention Block. The core component of the model is the graph attention block, illustrated in Fig. 4. It takes node features and molecular graph attributes as input. All the features are projected and concatenated into a large message tensor of shape (N, M, H), where N is the number of nodes, M is the max number of neighbor, and H is the message size. The message tensor is then processed by a multi-head self-attention mechanism. The attention is parallelized over each neighborhood, where M is the sequence length. By using customized Trition kernels [Tillet et al., 2019, Lefaudeux et al., 2022], the attention mechanism is highly optimized for GPU acceleration. The output of the attention mechanism is aggregated back to the atom level. The aggregated messages are then passed through the node-wise Feed Forward Network (FFN) to produce the output node features. To the best of our knowledge, this attention mechanism is unique because it acts on a neighborhood level, which is more expressive than the graph attention architectures that only act on the node level.

Readout Block. We use two readout layers for each graph attention block, which follows GemNet-OC [Gasteiger et al., 2022]. The first readout layer takes in the unaggregated messages from the graph attention block and produces edge readout features. The second readout layer takes in the output node features from the node-wise FFN and produces node readout features. The node and edge readout features from all graph attention blocks are concatenated and passed into the output block for output prediction.

Output Block. The output block takes the concatenated readout features and predicts the per-atom forces and system energy. The energy prediction is done by an FFN on the node readout features. The force prediction is divided into two parts: the force magnitude is predicted by an FFN on the node readout features, and the force direction is predicted by a transformation of the unit edge directions with an FFN on the edge readout features. As opposed to GemNet [Gasteiger et al., 2022], the

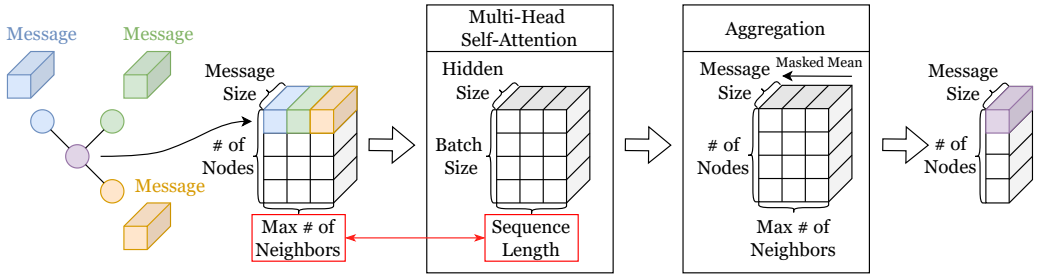

Figure 4: Detailed illustration of the graph attention block. The input attributes are projected and concatenated into a large message tensor. The tensor is fed into an optimized multi-head self-attention computation, where the max number of neighbors dimension is the sequence length dimension.

Table 1: EScAIP performance on the OC20 All+MD, OC20 2M, and OC22 datasets. The results are reported in Energy (eV) and Force (eV/Å) mean absolute error (MAE). EScAIP generally achieves the best Energy and Force MAE among all current models. Due to its efficiency, EScAIP requires less training time compared to the other models.

Dataset	Model	# of Parameters	Validation Energy MAE (meV) ↓	Validation Force MAE (meV/Å) ↓	Test Energy MAE (meV) ↓	Test Force MAE (meV/Å) ↓
OC20 All+MD	GemNet-OC-L-F	216M	252	19.99	241	19.01
	eSCN L=6 K=20	200M	243	17.09	228	15.60
	EquiformerV2 ($\lambda_E = 4$)	31M	232	16.26	228	15.50
	EquiformerV2 ($\lambda_E = 2$)	153M	230	14.60	227	13.80
	EquiformerV2 ($\lambda_E = 4$)	153M	227	15.04	219	14.20
	EScAIP-Small	83M	229	15.07	233	15.73
	EScAIP-Medium	146M	217	12.82	221	13.19
	EScAIP-Large	317M	**211**	**12.17**	**215**	**12.65**
OC20 2M	GemNet-dT	31M	358	29.50	-	-
	GemNet-OC	38M	286	25.70	-	-
	SCN	126M	279	21.90	-	-
	eSCN	51M	283	20.50	-	-
	EquiformerV2	85M	285	20.46	-	-
	EScAIP-Small	83M	263	20.15	-	-
	EScAIP-Medium	146M	**254**	**19.08**	-	-
OC22	GemNet-OC	39M	-	-	707	35.0
	EquiformerV2	122M	531	26.79	**462**	27.1
	EScAIP-Medium	146M	**514**	**24.32**	473	**25.73**

transformation is not scalar but vector-valued. Thus, the predicted force direction is not equivariant to rotations of the input data. In our experiments, we found this symmetry-breaking output block made the model perform better. The reason could be that this formulation has more degrees of freedom and so is easier to optimize. We note that though the force direction is initially not equivariant, the trained model is able to learn this symmetry from the data (See §5.4).

We also note that predicting the force magnitude from node readout features is very helpful for energy prediction. The reason could be that the energy prediction is a global property of the molecular system, while the force magnitude is a local property of the atom. By guiding the model towards a fine-grained force magnitude prediction, the model can learn a better representation of the system, which can in turn help it predict the system energy more accurately.

5 Experiments

We conduct experiments on a wide range of chemical systems, including catalysts (OC20 and OC22) §5.1, materials (MPTrj) §5.2, and molecules (SPICE and MD22) §5.3,§B.2.

5.1 Catalysts (OC20 and OC22)

Dataset. We evaluate the performance of our EScAIP model on the Open Catalyst dataset [Chanussot et al., 2021, Tran et al., 2023], which consists of 172 million systems with 73 atoms on average.

Table 2: EScAIP efficiency comparisons with baseline models on the OC20 dataset. All reported results are measured on NVIDIA V100 32G.

Dataset	Model	# of Parameters	Training Speed (Sample/Sec) ↑	Training Memory (GB/Sample) ↓	Inference Speed (Sample/Sec) ↑	Inference Memory (GB/Sample) ↓
OC20	GemNet-OC	216M	9.44	3.40	23.67	2.31
	eSCN L=6 K=20	200M	2.16	6.90	6.3	6.15
	EquiformerV2	31M	14.22	1.63	34.55	1.61
	EquiformerV2	153M	2.85	6.33	7.92	5.24
	EScAIP-Small	83M	35.84	1.23	107.04	1.09
	EScAIP-Medium	146M	25.36	1.54	74.77	1.28
	EScAIP-Large	312M	12.88	2.78	40.56	2.28

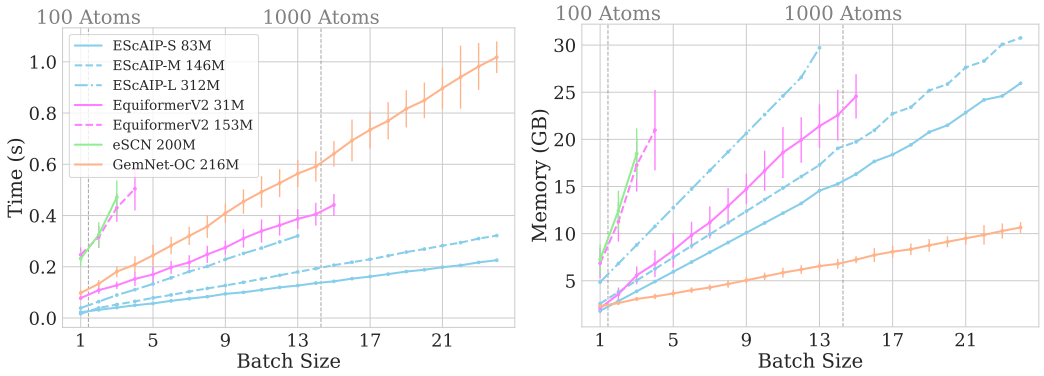

Figure 5: Inference runtime and memory usage comparison of EScAIP and baseline models on the OC20 dataset. Mean and standard deviation are reported across 16 randomly sampled batches per batch size. Grey lines indicate the cumulative number of atoms in the batch. EScAIP not only scales efficiently with batch size, but also exhibits minimal variation in performance across different batches. All reported results are tested on NVIDIA V100 32G.

We evaluate on the S2EF task, which is the prediction of system energy and per-atomic force from atomistic structure.

Settings. We use three variants of the EScAIP model: Small (83M), Medium (146M), and Large (317M). The models are trained to predict the energy and forces of each sample (S2EF task). We train the model on the OC20 All+MD, OC20 2M, and OC22 splits. We evaluate the performance on the four validation sets and test sets (both have 4M samples in total) and compare the results with EquiformerV2 [Liao et al., 2024], eSCN [Passaro and Zitnick, 2023b], SCN [Zitnick et al., 2022], and GemNet-OC [Gasteiger et al., 2022], the best performing models on this dataset.

Results. The results of EScAIP on the Open Catalyst dataset are summarized in Tab. 1, where EScAIP achieves state-of-the-art performance across all splits: OC20 2M, OC20 All+MD, and OC22. We note that we exclude models that train with a denoising objective, as we focus on the performance of the model architecture itself. There is a clear trend that increasing the model size improves the performance of EScAIP. Notably, even the Small model achieves competitive performance against other models while remaining significantly more efficient, making it suitable for downstream, practical applications. More results on the scalability of the EScAIP model can be found in the Appendix B.1.

Efficiency Comparisons. We provide the runtime and memory usage of EScAIP and other baseline models on the OC20 dataset in Tab. 2. EScAIP is approximately 10x faster and has 5x less memory usage than an EquiformerV2 [Liao et al., 2024] model of comparable size. Additionally, as shown in Fig. 5, EScAIP scales effectively with batch size while maintaining minimal performance variation across batches. This consistency is because EScAIP's input is padded to the maximum system size, enabling efficient use of PyTorch's compile feature. These qualities could make EScAIP well-suited for practical applications.

5.2 Materials (MPTrj)

Dataset. We evaluate EScAIP's performance on the Matbench-Discovery benchmark [Riebesell et al., 2023], a widely recognized benchmark for assessing models in new materials discovery. The

Table 3: EScAIP performance on the Matbench-Discovery benchmark. Mean absolute error (MAE) and Root Mean Squared Error (RMSE) are reported in eV/atom.

Model	F1 ↑	DAF ↑	Precision ↑	Recall ↑	Accuracy ↑	TPR ↑	FPR ↓	TNR ↑	FNR ↓	MAE ↓	RMSE ↓	R2 ↑
MACE	0.669	3.777	0.577	0.796	0.878	0.796	0.107	0.893	0.204	57	101	0.697
SevenNet	0.724	4.252	0.65	0.818	0.904	0.818	0.081	0.919	0.182	48	92	0.75
ORB MPtrj	0.765	4.702	**0.719**	0.817	0.922	0.817	0.059	0.941	0.183	45	91	0.756
EquiformerV2	0.77	4.64	0.709	0.841	0.926	0.841	0.063	0.937	0.159	42	87	0.778
EScAIP	**0.782**	**5.634**	0.712	**0.869**	**0.939**	**0.869**	**0.050**	**0.949**	**0.131**	**38**	**85**	**0.783**

Table 4: EScAIP performance on the SPICE dataset. The results are reported in Energy (meV/atom) and Force (meV/Å) mean absolute error (MAE).

Model	Metric	PubChem	DES370K Monomers	DES370K Dimers	Dipeptides	Solvated Amino Acids	Water	QMugs
MACE	Force MAE	14.75	6.58	6.62	10.19	19.43	13.57	16.93
	Energy MAE	0.88	0.59	0.54	0.42	0.98	0.83	0.45
EScAIP	Force MAE	**5.86**	**3.48**	**2.18**	**5.21**	**11.52**	**10.31**	**8.74**
	Energy MAE	**0.53**	**0.41**	**0.38**	**0.31**	**0.61**	**0.72**	**0.41**

model is trained on the MPTrj dataset [Deng et al., 2023], which consists of 1.6 million samples. This approach adheres to the "compliant" setting of the Matbench-Discovery benchmark.

Settings. Given the relatively small dataset size, we use a small version of EScAIP with 45M parameters. The model is trained to predict the energy, force, and stress of each sample. After training for 100 training epochs, we increase the energy coefficient in the loss function and fine-tune the model for another 50 epochs. We evaluate the performance on the Matbench-Discovery benchmark and compare the results with EquiformerV2 [Liao et al., 2024, Barroso-Luque et al., 2024], ORB MPTrj [Orbital-Materials, 2024], SevenNet [Park et al., 2024], and MACE [Batatia et al., 2023, 2022]—-the top compliant models on this benchmark. We note that we exclude models that train with a denoising objective, as we focus on the performance of the model architecture itself[1].

Results. The results of EScAIP on the Matbench-Discovery benchmark are summarized in Tab. 3. EScAIP achieves state-of-the-art performance on this benchmark, outperforming other models. In the code release, we also provide the EScAIP model before the energy fine-tuning step, as this may be more relevant for applications such as molecular dynamics simulations.

5.3 Molecules (SPICE)

Dataset. We evaluate the EScAIP model's performance on the SPICE dataset [Eastman et al., 2023], which consists of approximately one million molecules across seven different categories. To ensure comparability, we adopt the same training and evaluation settings as used for the MACE-OFF23 model [Kovács et al., 2023a].

Settings. We use a smaller EScAIP model with 45M parameters, trained to predict the energy and forces of each sample. The model's performance is then evaluated on the different SPICE test datasets and compared directly with MACE-OFF23 [Kovács et al., 2023a].

Results. A summary of EScAIP's results on the SPICE dataset is provided in Tab. 4, where it outperforms MACE-OFF23 in predicting the energy and forces on the different test sets.

5.4 Rotational Equivariance Analysis

Settings. To assess whether EScAIP learns rotational equivariance after training on various datasets, we design the following procedure: first, a batch is randomly sampled from the validation dataset and passed through the trained model to obtain a force prediction (A). Next, we rotate the batch by a random angle and obtain a second force prediction (B) from the model. To quantify rotational

[1]Based on the ORB technical report orb [2024], it is possible that the ORB MPTrj model result reported here was pre-trained with a denoising objective.

Table 5: To analyze rotational equivariance, on the validation set, we 1) predict forces on a set of atomistic systems (A), 2) rotate the atomistic systems and predict forces (B), and then 3) compute the cosine similarity between the force predictions (B) and the rotated version of force predictions (A). After training EScAIP on different datasets, we find that the cosine similarity is consistently ≥ 0.99, meaning EScAIP is essentially always predicting the rotations correctly.

Dataset	OC20 All+MD			MPTrj	SPICE
# of Params.	83M	146M	312M	45M	45M
Before Training	0.2109	0.2940	0.2132	0.2287	0.2364
After Training	0.9981	0.9987	0.9994	0.9999	0.9999

equivariance, we calculate the cosine similarity between prediction (B) and the rotated version of prediction (A). This process is repeated for 128 batches, and we report the average cosine similarity.

Results. The results of the rotational equivariance analysis are presented in Tab. 5, and the cosine similarity is consistently ≥ 0.99. These findings indicate that though EScAIP is not initially rotationally equivariant, after training it is able to correctly map input system rotations to output system predictions. This could suggest that these symmetries can be effectively learned from the data.

6 Conclusions

We have investigated scaling strategies for developing neural network interatomic potentials (NNIPs) on large-scale datasets. Based on our investigations, we introduced a new NNIP architecture, Efficiently Scaled Attention Interatomic Potential (EScAIP), that leverages highly optimized self-attention mechanisms for scalability and expressivity. We demonstrated the effectiveness of EScAIP on a wide range of chemical datasets (OC20, OC22, MPTrj, SPICE) and showed that EScAIP achieves top performance on these different prediction tasks, while being much more efficient in training and inference runtime, as well as memory usage. We highlight some important takeaways from our work and the future of machine learning interatomic potentials more broadly:

The "sweet lesson." We note that our line of investigation in this work follows some of the general principles of the bitter lesson [Sutton, 2019]. That is, strategies that focus on scaling and compute tend to outperform those that try to embed domain knowledge into models. However, in this field, we prefer to think of this as a "sweet lesson." Training large, constrained models requires significantly more computational resources, making this feasible for only a limited number of researchers. Efficient scaling strategies thus democratize large-scale training and make it accessible to a broader community.

We're still not giving enough credit to the data. Thus far, much of the effort in the NNIP field has concentrated on model development. However, atomistic systems are far more complex than the domain-specific information being embedded into models. Predefined symmetry constraints and handcrafted features offer only a simplistic representation of this complexity. A path forward to capture these complexities is to focus on generating comprehensive datasets, ideally accompanied by relevant evaluation metrics, allowing NNs to learn the rest of the information through gaining expressivity via scaling.

Future of NNIPs. As datasets continue to grow, training models from scratch on small datasets will likely become unnecessary. While constraints may be beneficial in the very small data regime (though data augmentation techniques can also help here), leveraging the representation of a pre-trained large model can serve as a starting point for fine-tuning on smaller datasets. This could make the very small dataset regime essentially a non-factor in the future, and it is likely that the need for NNIPs with built-in hard-constraints becomes even less necessary. Beyond focusing on data generation, other techniques are likely to gain importance in the NNIP domain. These include model distillation, general training and inference strategies that are model agnostic and can be applied to any NNIP, and approaches to better connect with experimental results. Finally, more comprehensive strategies will be important for evaluating NNIP accuracy and utility.

Acknowledgments and Disclosure of Funding

This work was initially supported by Laboratory Directed Research and Development (LDRD) funding under Contract Number DE-AC02-05CH11231. It was then supported in part by the Office of Naval Research (ONR) under grant N00014-23-1-2587. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231. We thank Ishan Amin, Sam Blau, Xiang Fu, Rasmus Hoegh, Mit Kotak, Toby Kreiman, Jennifer Listgarten, Ryan Liu, Sanjeev Raja, and Brandon Wood for helpful discussions and comments.

References

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

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*, pages 12104–12113, 2022.

Volker L Deringer, Miguel A Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.

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

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):2453, 2022.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436, 2022.

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`.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Saro Passaro and C Lawrence Zitnick. Reducing so (3) convolutions to so (2) for efficient equivariant gnns. In *International Conference on Machine Learning*, pages 27420–27438. PMLR, 2023a.

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`.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020a. URL `https://openreview.net/forum?id=B1eWbxStPH`.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ward Ulissi, C Lawrence Zitnick, and Abhishek Das. Gemnet-oc: Developing graph neural networks for large and diverse molecular simulation datasets. *Transactions on Machine Learning Research*, 2022.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. `https://github.com/facebookresearch/xformers`, 2022.

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.

Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019.

Larry Zitnick, Abhishek Das, Adeesh Kolluru, Janice Lan, Muhammed Shuaibi, Anuroop Sriram, Zachary Ulissi, and Brandon Wood. Spherical channels for modeling atomic interactions. *Advances in Neural Information Processing Systems*, 35:8054–8067, 2022.

Saro Passaro and C. Lawrence Zitnick. Reducing so(3) convolutions to so(2) for efficient equivariant gnns. In *International Conference on Machine Learning*, 2023b.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2022.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020b.

Thorben Frank, Oliver Unke, and Klaus-Robert Müller. So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems. *Advances in Neural Information Processing Systems*, 35:29400–29413, 2022.

Erik J Bekkers, Sharvaree Vadgama, Rob D Hesselink, Putri A van der Linden, and David W Romero. Fast, expressive se(n) equivariant networks through weight-sharing in position-orientation space, 2024.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

Bingqing Cheng. Cartesian atomic cluster expansion for machine learning interatomic potentials, 2024.

Mojtaba Haghighatlari, Jie Li, Xingyi Guan, Oufan Zhang, Akshaya Das, Christopher J Stein, Farnaz Heidar-Zadeh, Meili Liu, Martin Head-Gordon, Luke Bertels, et al. Newtonnet: a newtonian message passing network for deep learning of interatomic potentials and forces. *Digital Discovery*, 1(3):333–343, 2022.

Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

Justin S Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E Roitberg, Olexandr Isayev, and Sergei Tretiak. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific data*, 7(1):134, 2020.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

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, 2024.

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

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials*, 6(1):173, 2020.

Sergey Pozdnyakov and Michele Ceriotti. Smooth, exact rotational symmetrization for deep learning on point clouds. *Advances in Neural Information Processing Systems*, 36, 2023.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, pages 23078–23091. PMLR, 2022.

Marc Finzi, Gregory Benton, and Andrew G Wilson. Residual pathway priors for soft equivariance constraints. *Advances in Neural Information Processing Systems*, 34:30037–30049, 2021.

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`.

Alexandre Duval, Victor Schmidt, Alex Hernandez Garcia, Santiago Miret, Fragkiskos D. Malliaros, Yoshua Bengio, and David Rolnick. Faenet: Frame averaging equivariant gnn for materials modeling, 2023.

Paul J Steinhardt, David R Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Physical Review B*, 28(2):784, 1983.

Philippe Tillet, Hsiang-Tsung Kung, and David Cox. Triton: an intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, pages 10–19, 2019.

Janosh Riebesell, Rhys EA Goodall, Anubhav Jain, Philipp Benner, Kristin A Persson, and Alpha A Lee. Matbench discovery–an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*, 2023.

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

Orbital-Materials. Orb forcefield models from orbital materialsorb forcefield models from orbital materials. https://github.com/orbital-materials/orb-models, 2024. Accessed: 2024-10-29.

Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 2024.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, William J. Baldwin, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. 2023.

Technical blog: Introducing the orb ai-based interatomic potential. https://www.orbitalmaterials.com/post/technical-blog-introducing-the-orb-ai-based-interatomic-potential, 2024. Accessed: 2024-10-29.

Dávid Péter Kovács, J Harry Moore, Nicholas J Browning, Ilyes Batatia, Joshua T Horton, Venkat Kapil, William C Witt, Ioan-Bogdan Magdău, Daniel J Cole, and Gábor Csányi. Mace-off23: Transferable machine learning force fields for organic molecules. *arXiv preprint arXiv:2312.15211*, 2023a.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

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.

Yunyang Li, Yusong Wang, Lin Huang, Han Yang, Xinran Wei, Jia Zhang, Tong Wang, Zun Wang, Bin Shao, and Tie-Yan Liu. Long-short-range message-passing: A physics-informed framework to capture non-local interaction for scalable molecular dynamics simulation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rvDQtdMnOl.

Dávid Péter Kovács, Ilyes Batatia, Eszter Sára Arany, and Gábor Csányi. Evaluation of the mace force field architecture: From medicinal chemistry to materials science. *The Journal of Chemical Physics*, 159(4), 2023b.

Sanjeev Raja, Ishan Amin, Fabian Pedregosa, and Aditi S Krishnapriyan. Stability-aware training of neural network interatomic potentials with differentiable boltzmann estimators. *arXiv preprint arXiv:2402.13984*, 2024.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi S Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023.

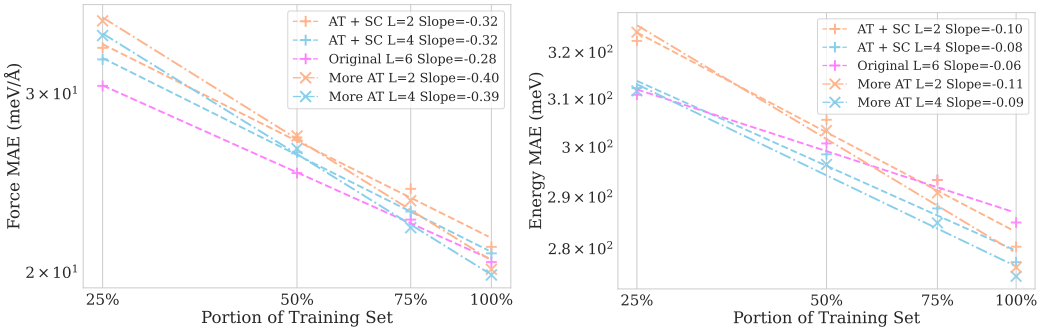

Figure 6: Force MAE vs.Training Dataset Size for EquiformerV2 ablation study on the OC20 2M dataset. Slope is fitted by linear regression. We scale the parameters of the original $L = 2$ and $L = 4$ models from Liao et al. [2024] through the attention mechanisms and/or spherical channels, such that the number of parameters is approximately equal to the original $L = 6$ model. As the training dataset size increases, the scaled $L = 2$ and $L = 4$ models have a steeper slope, indicating faster performance improvement with increasing data.

A Additional Details on Investigations

We provide additional details on our investigations from §3.

Results of Ablation Study comparing Force MAE vs. Training Dataset Size. To further investigate how scaling efficiency varies as a function of training dataset size, we train the parameter-controlled Equiformer V2 models with different amounts of training data. The results in Fig. 6 show that the scaled $L = 2$ and $L = 4$ models exhibit a steeper performance improvement (log-log slope) compared to the original $L = 6$ model. In particular, the *More AT* configuration (more attention) has a steeper log-log slope compared to the *AT + SC* configuration and the original $L = 6$ model. This suggests that increasing the complexity of the attention mechanisms is a more effective strategy for scaling with increasing training dataset size, rather than increasing L.

B Additional Details and Results on Experiments

B.1 Catalysts (OC20 and OC22)

To illustrate EScAIP's scalability, we train the model on varying sizes of training data and model configurations. The results, shown in Fig. 8, indicate a clear trend: as model and data sizes grow, EScAIP's performance continues to improve. We also include results to complement Fig. 1: the same efficiency, performance, and scaling comparisons between EScAIP and baseline models on the Open Catalyst dataset for Energy MAE (meV ↓) vs. Inference Speed (Sample/Sec ↑) and Energy MAE vs. Memory (GB/Sample ↓). The trend is similar to the forces MAE results, and EScAIP achieves better performance with smaller time and memory cost.

B.2 Large Molecules (MD22)

Dataset. We evaluate the performance of our EScAIP model on the MD22 dataset [Chmiela et al., 2023], which consists of seven molecular systems with varying sizes. It consists of energy and force labels calculated from DFT simulations.

Settings. We use an EScAIP model 15M parameters on each system and evaluate the performance on the test set (train-test split 95:5). The model is compared with MACE [Batatia et al., 2022], VisNet-LSRM [Li et al., 2024], and sGDML [Chmiela et al., 2023]. We note that there were discrepancy of results of VisNet-LSRM in the MACE [Kovács et al., 2023b] paper and the VisNet-LSRM [Li et al., 2024], thus we reported both in the table. We use the same train/validation splits as the baselines [Chmiela et al., 2023], where the training set is hundreds to thousands of samples, and also apply data augmentation (randomly rotating each training sample 16 times).

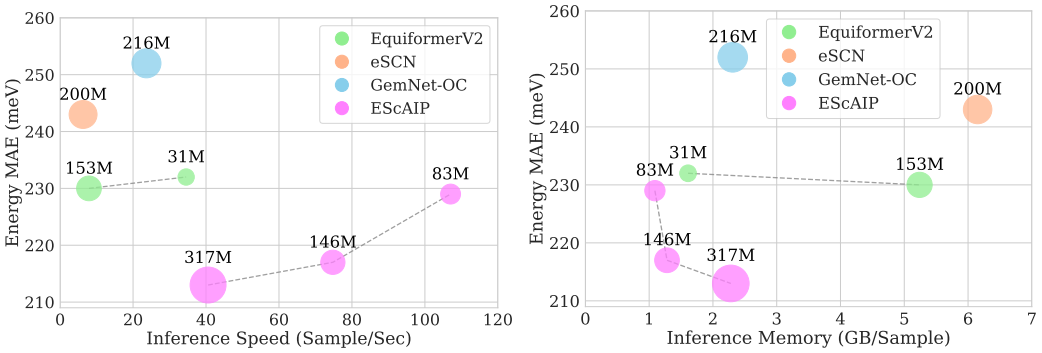

Figure 7: Efficiency, performance, and scaling comparisons between EScAIP and baseline models on the Open Catalyst dataset. Energy MAE (meV ↓) vs. Inference Speed (Sample/Sec ↑) and Energy MAE vs. Memory (GB/Sample ↓) is reported. EScAIP achieves better performance with smaller time and memory cost.

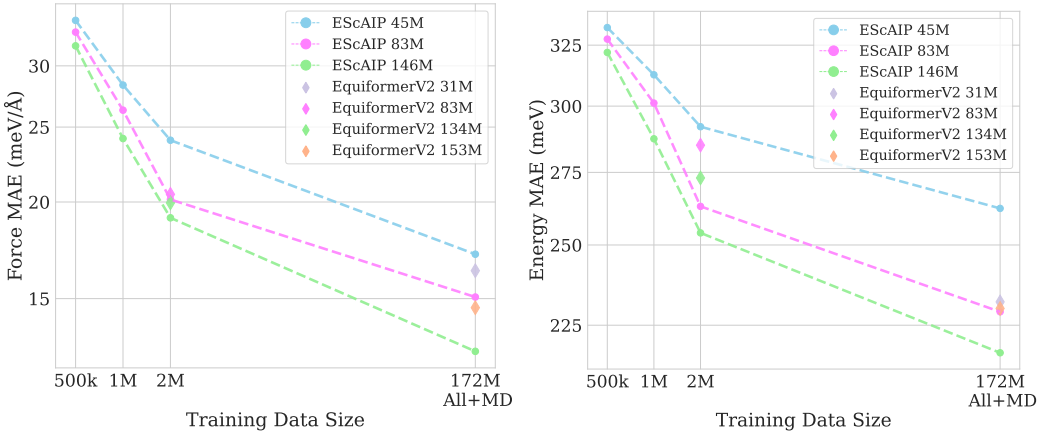

Figure 8: Scaling experiment of EScAIP on OC20. Forces MAE (meV/Å) and Energy (meV) across 4 validation splits are reported. For 500k, 1M, and 2M split, the EScAIP model is trained for 30 epochs; for All+MD, the EScAIP model is trained for 8 epochs. Force and Energy MAE consistently decreases as model size and training data size increases.

After we train the model, we also run molecular dynamics (MD) simulations to check the stability of the potential and evaluate how well the simulation recovers the distribution of interatomic distances, h(r), in the simulation [Raja et al., 2024, Fu et al., 2023]. We use the simulation setup from Fu et al. [2023] and run the simulation for 200000 steps (100 ps) using the Langevin integrator with a friction coefficient of 0.5. The temperature is set to 500 K.

Results. The results of EScAIP on the MD22 dataset are summarized in Tab. 6. EScAIP outperforms other models in both energy and force prediction, especially for large molecules. The low $h(r)$ error in the MD simulation also indicates that the model is able to capture this observable accurately. Interestingly, MD22 is not a particularly large dataset: the training dataset sizes are in the thousands. Despite this, a scalable architecture with high parameter counts is still able to achieve good performance.

Table 6: EScAIP performance on the MD22 dataset. The results are reported in Energy (meV/atom), Force (meV/Å) and $h(r)$ (unitless) mean absolute error (MAE).

	Metric	Tetra-peptide	Fatty acid	Tetra-saccharide	Nucleic acid (AT-AT)	Nucleic acid (AT-AT-CG-CG)	Buckyball catcher	Double-walled nanotube
# of Atoms		42	56	87	60	118	148	370
sGDML	Energy	0.40	1.0	2.0	0.52	0.52	0.34	0.47
	Force	34	33	29	30	31	29	23
MACE	Energy	0.064	0.102	0.062	0.079	0.058	0.141	0.194
	Force	3.8	**2.8**	3.8	4.3	5.0	3.7	12.0
VisNet-LSRM (MACE)	Energy	0.080	0.058	0.044	**0.055**	0.049	0.124	0.117
	Force	5.7	3.6	5.0	5.2	8.3	11.6	28.7
VisNet-LSRM (Paper)	Energy	0.068	0.068	0.053	0.056	**0.039**	-	-
	Force	3.9	2.5	3.3	**3.4**	4.6	-	-
EScAIP	Energy	**0.053**	**0.052**	**0.041**	0.062	0.042	**0.112**	**0.095**
	Force	**3.3**	3.2	**3.1**	3.8	**4.3**	**2.5**	**8.1**
	$h(r)$	0.07	0.09	0.11	0.13	0.15	0.05	0.21

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are detailed in the results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss some limitations in §6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not have a theoretical derivation.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we provide all details of our method and a schematic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets are available on Github. We also provide all code and model checkpoints.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we provide this in the experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to computational cost, we are not able to provide error bars for every experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we provide computational details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed this and the research conforms to this code.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, this is also discussed in the introduction.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This is not relevant to this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we cite the relevant datasets we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we provide a new model and include implementation details.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We don't have this.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve this.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.