Enhancing the local expressivity of geometric graph neural networks

Sam Walton Norwood
Technical University of Denmark
swano@dtu.dk

Lars L. Schaaf
University of Cambridge
lls34@cam.ac.uk

Ilyes Batatia
University of Cambridge
ib467@cam.ac.uk

Gábor Csányi
University of Cambridge
gc121@cam.ac.uk

Arghya Bhowmik
Technical University of Denmark
arbh@dtu.dk

Abstract

A central operation in geometric graph neural networks (GNNs) is the equivariant *pairwise embedding function*, which encodes the local environment of each node as a learned representation. In this work, we examine the role of pairwise embedding and consider a series of generalizations of its functional form beyond previous work. The new embeddings that we design considerably advance the state of the art in challenging distributions: as a highlight, when applied as an interatomic potential, we achieve a 29% relative reduction of force errors on diverse allotropes of lithium-intercalated carbon with a 4-fold reduction in parameter count. Furthermore, we demonstrate improved transferability in molecular datasets by varying the locality of the network according to the depth of the representation.

1 Introduction

In physical sciences, machine learning often deals with graph-structured data that are embedded in vector spaces, known as *geometric graphs* [1]. Geometric graph neural networks are designed to handle these data efficiently by imposing relevant inductive biases, such as equivariance or invariance to permutation, as well as equivariance under the action of other groups, such as the Euclidean ($E(3)$) group of rotations, reflections, and translations [2]. The MACE architecture [3] has proven to be state of the art for a wide variety of tasks in fields as distinct as atomistic modeling, computer vision, and particle physics [4]. It has been shown that most of the existing geometric GNN architectures correspond to hyperparameter choices within the MACE design space [5]. Unlike other geometric GNNs, MACE allows control of the body order separately from the depth of the network [6].

In this paper, we explore architectural choices for geometric GNNs with the goal of obtaining improved performance on atomistic data. We report state-of-the-art results for a number of challenging datasets, often with a reduction in parameter count. Concretely, we make the following contributions:

- We show that equipping the embedding functions at different layers of the network with **different degrees of locality** enhances accuracy in a physically explainable manner.

- We demonstrate that applying **geometric self-attention** over the local receptive field results in substantial reductions in generalization error and parametric complexity.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

2 Method

2.1 The MACE architecture

In a forward pass of the MACE model, initial node features $h_i^{(0)}$ are generated as linear projections of one-hot vectors δ_{zz_i}, and edges r_{ij} are obtained from the neighbor graph constructed with cutoff radius r_{cut}. Edge lengths $||r_{ij}||$ are expanded in a set of radial basis functions j_0^n, often chosen as Bessel or Gaussian functions, and edge directions $\hat{r}_{ij}$ are expanded in a basis of spherical harmonics Y_l^m. To construct the edge embeddings, the expanded edge lengths are passed through an MLP $R(r_{ij}) = \mathrm{MLP}(j_0^n(||r_{ij}||))$ to generate a set of coefficients that parameterize an $E(3)$-equivariant tensor product between the node features h_j and the expanded edge directions. Collectively, this set of operations defines the pairwise embedding function $\phi_{ij}^{(s)}$ of the ij'th edge at layer s:

$$\phi_{ij,k\eta_1 l_3 m_3}^{(s)} = \sum_{l_1 l_2 m_1 m_2} C_{l_3 m_3 \eta_1, l_1 m_1 l_2 m_2} R_{k\eta_1 l_1 l_2 l_3}^{(s)}(r_{ij}) \times Y_{l_1}^{m_1}(\hat{r}_{ij}) h_{j,kl_2 m_2}^{(s)} \tag{1}$$

Where k indexes feature channels, l, m index angular momenta, $C_{l_3 m_3 \eta_1, l_1 m_1 l_2 m_2}$ are Clebsch-Gordan coefficients enforcing $E(3)$ equivariance, and η indexes combinations of l, m which preserve equivariance. The environment embedding $A_i^{(s)}$ of node i at layer s is constructed by aggregating edge embeddings over the local environment and linearly projecting:

$$A_{i,kl_3 m_3}^{(s)} = \sum_{\tilde{k},\eta_1} W_{k\tilde{k}\eta_1 l_3}^{(s)} \sum_{j \in \mathcal{N}(i)} \phi_{ij,\tilde{k}\eta_1 l_3 m_3}^{(s)} \tag{2}$$

Next, a tensor product is applied to the environment embedding ν times to increase the correlation order, and the resulting features are symmetrized via tensor contraction with generalized Clebsch-Gordan coefficients $\mathcal{C}_{\eta_\nu lm}^{LM}$.

$$\boldsymbol{B}_{i,\eta_\nu kLM}^{(s),\nu} = \sum_{lm} \mathcal{C}_{\eta_\nu lm}^{LM} \prod_{\xi=1}^{\nu} A_{i,kl_\xi m_\xi}^{(s)} \tag{3}$$

Finally, the symmetrized high-order features $\boldsymbol{B}_i^{(s),\nu}$ are linearly projected and combined residually with the node features from the previous layer to produce the node features of the next layer. Readout functions are then defined to regress target quantities from node features.

2.2 An extended design space of pairwise embeddings

The design choices relevant to body-ordered equivariant networks include 1) the choice of symmetry group ($E(3)$ or $SE(3)$ for MACE, or any reductive Lie group for G-MACE [4]); 2) the order ν of the symmetrized tensor product; 3) the architecture of the readout functions; and 4) the architecture of the pairwise embedding function ϕ_{ij}. Given that the pairwise embedding function directly controls the representation learned by the network, in this work, we focus on its design.

In 1, the MLP $R(r_{ij})$ is the only trainable component of ϕ_{ij}, and it depends only on the edge distance, such that ϕ_{ij} depends only on r_{ij} and h_j. Recent work [6] found benefit from augmenting ϕ_{ij} with nonlinear dependence on the ordered pair features (h_i, h_j). We consider pairwise embedding functions of the extended form:

$$\phi_{ij}^{(s)} \propto (r_{ij}, h_i^{(s)}, h_j^{(s)}, \{h_p^{(s)}, \phi_{ip}^{(s)} \forall p \in \mathcal{N}(i)^{(s)}\}) \tag{4}$$

Compared to 1, this form introduces dependence of the equivariant tensor product on arbitrary functions of all nodes and edges in the environment, and further defines the extent of the local environment (i.e. the set of neighbors $\mathcal{N}$) separately for each layer. We describe an implementation of an embedding of this form in the following section.

2.3 Geometric attention

A number of recent works have proposed the incorporation of attention mechanisms into geometric graph neural networks [7, 8, 9, 10]. Attention can be applied to neighboring nodes in the local environment to amplify or suppress particular edges when aggregating the environment embedding [7].

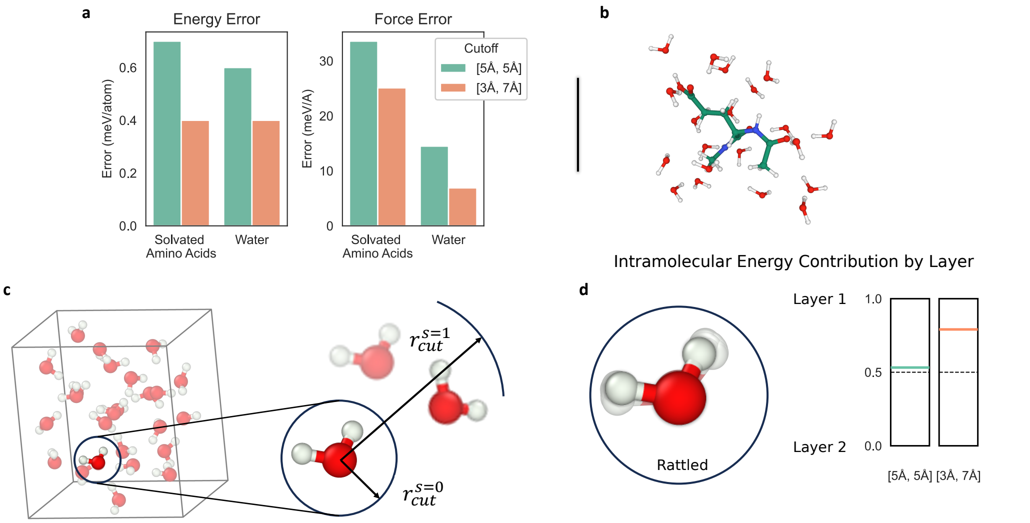

Figure 1: **Physically motivated partitioning of receptive field.** Comparison between two models with the same total receptive field of 10 Å show that a receptive field partitioning with a cutoff of 3 Å for the first layer and 7 Å cutoff for the second layer significantly outperforms using the same 5 Å cutoffs for both layers (**a**). The data-set contains solvated aminio-acid (b). The improved performance may be attributed to better separation of inter- and intra- molecular contributions (c). We show that intra-molecular energy contributions due to changes of the internal coordinates of a water molecule originate equally from the first and second layer, for the 5 Å, 5 Å model (d). In the case of the 3 Å, 7 Å partitioning, the intra-molecular contributions originate predominantly from the first layer.

To capture information about the geometry of the local environment, we compute attention weights between the central node's features and the pairwise embeddings of neighboring nodes:

$$\alpha_{ij,\mu}^{(s)} = \frac{\exp(\langle W_{Q,\mu}^{(s)} h_i^{(s)}, W_{K,\mu}^{(s)} \phi_{ij}^{(s)} \rangle)}{\sum_{p \in \mathcal{N}(i)^{(s)}} \exp(\langle W_{Q,\mu}^{(s)} h_i^{(s)}, W_{K,\mu}^{(s)} \phi_{ip}^{(s)} \rangle)} \tag{5}$$

$$A_{i,kl_3m_3}^{(s)} = \sum_{\tilde{k}_\mu, \eta_1} W_{k\tilde{k}_\mu \eta_1 l_3}^{(s)} \sum_{j \in \mathcal{N}(i)} \bigoplus_{\mu} \alpha_{ij,\mu}^{(s)} \phi_{ij,\tilde{k}\eta_1 l_3 m_3}^{(s)} \tag{6}$$

Where μ indexes separate parallel heads of attention, query and key operators W_Q and W_K project equivariant features to a common reduced dimension, $\bigoplus$ indicates concatenation, and $\langle \cdot, \cdot \rangle$ is an equivariant inner product. Evaluating similarity between query and key features using an equivariant inner product ensures the attention operation as a whole remains equivariant.

3 Layer-dependent locality

We illustrate how incorporating added flexibility in the pairwise embedding and graph structure enhances transferability and facilitates learning in low-data regimes. The first kind of flexibility we investigate is allowing for layer-dependent cutoff distances. Typically, for the task of learning molecular energies and forces, message passing occurs on an identical graph across layers: given an atomic point cloud, all atoms within a specified cut-off distance are connected with an edge. We find that adopting a smaller cut-off for the initial layer and a larger one for the subsequent layers yields a significant reduction in test error when training on solvated amino acids (Figure 1a). This modification helps integrate the inductive bias of inter- and intra-molecular interactions.

Our training set consists of a subset of the molecular SPICE dataset [11]. We train two layer MACE [3] models on four distinct solvated amino acids and water clusters. Comparing two models with the same total receptive field of 10Å, we find that force and energy errors can be significantly

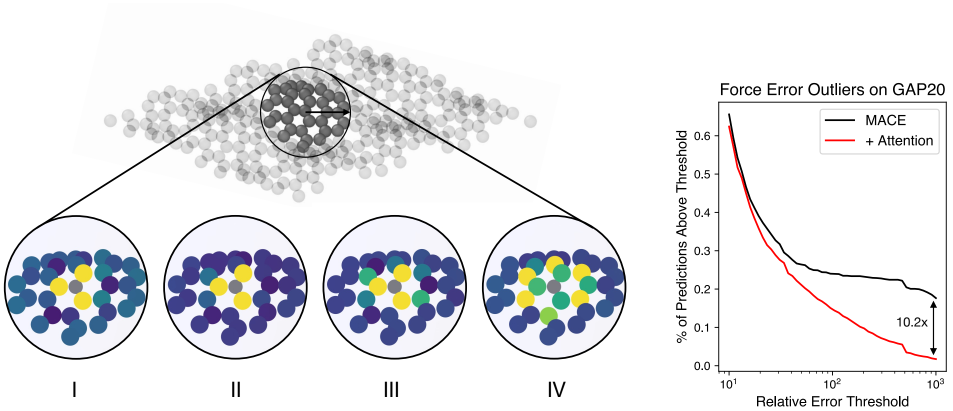

Figure 2: **Local attention improves generalization on the GAP20 carbon dataset**. Left: a rendering of all 4 learned attention patterns of the trained model for a representative atomic environment in a defected graphite sheet. For this environment, heads I-III weighs down all edges except the central atom's nearest neighbors, emulating the effect of a tighter cutoff radius. In contrast, head IV increases the weight of next-nearest neighbors. Right: the distribution of force error outliers on the GAP20 test set. Outliers are environments with a relative force error norm greater than the threshold on the x axis, shown from $10^1 - 10^3$. Local attention reduces the occurance of the largest generalization errors by an order of magnitude.

reduced by choosing a 3Å and 7Å cutoff for the first and second layer respectively, when compared to a standard equal partitioning of 5Å and 5Å.

We suggest that the observed improvement is a consequence of a more physically informed partitioning of the receptive field, which facilitates the differentiation between inter- and intra-molecular interactions. To substantiate this hypothesis, we evaluated a set of rattled single water molecules using both models. As depicted in Figure 1d, in the 5Å, 5Å model, the first layer accounts for 51% of the energy variation in the perturbed configurations, while in the 3Å, 7Å model, the contribution from the first layer rises to 71%.

The training set comprises 500 configurations, placing it in the low-data regime for systems of such complexity. This regime is particularly intriguing, especially in the context of water, where state-of-the-art force fields are developed based on computationally demanding coupled cluster calculations [12]. As such, incorporating a physical prior of the different length scales and interactions into the model may prove pivotal in learning on clusters and predicting bulk water behavior.

4 Local attention

Table 1: **Numerical errors and parameter counts** for MACE models trained on GAP20 carbon and lithium-carbon datasets. Energy and force validation RMSE are reported in meV/atom and meV/Å.

Model	**GAP20 C**				**Li-C**			
	Channels	**Params.**	**E**	**F**	**Channels**	**Params.**	**E**	**F**
Baseline MACE, 3 Layers	64L3	616K	10.7	328.3	64L3	710K	13.9	462.4
+ Attention (4 Heads)	16L3	**163K**	**8.6**	**257.9**	16L3	**176K**	**12.8**	**358.1**

Having shown the benefits of carefully controlled locality, we investigate the learnable control of the receptive field by constructing MACE networks with attention applied to the local environment. As a first test, we turn to the GAP20 dataset of carbon allotropes, which contains highly diverse carbon structures at a wide range of densities [13]. We find that a MACE network augmented with four attention heads outperforms a strong baseline MACE network of the same feature dimension[1]on GAP20 by 22% in validation force RMSE while using 73% fewer parameters (Table 1). Parametric complexity is of acute concern for atomistic applications, as it directly affects the number of particles that can be modelled simultaneously for a given amount of processor memory. Notably, our proposed model's force RMSE is roughly 1/3 that of the GAP model released with the dataset. We also investigate performance on a dataset of lithium-intercalated carbon structures derived from GAP20, where an attention model with matched feature dimension and 1/4 as many parameters yields a 29% improvement in force RMSE. Furthermore, for GAP20, we find that attention reduces the number of large relative force errors ("holes" in the learned potential) observed over the test set (Figure 2, right).

In Figure 2, we visualize the attention pattern learned for a particular environment in the test set on which the attention model has a force error residual roughly 48 times smaller than the baseline MACE model (3.3 meV/Å vs. 160 meV/Å). We note that the pattern learned is both topologically interpretable and nontrivial: for the environment in question, three of the four attention heads assign high weight to nearest neighbors, whereas the last head assigns high weight to *next-nearest* neighbors. This is suggestive of a role for attention in dynamically reducing the model's effective cutoff radius, incorporating a locality bias as appropriate to the chemical environment being considered. We leave systematic investigation of the patterns learned by MACE augmented with geometric attention to future work.

5 Conclusion

Our results indicate that the extensions to the MACE framework presented here allow interatomic interactions to be better captured, either via the direct incorporation of prior physical intuition, or via learnable sparsification of the atomic environment as a function of the local geometry. We expect these new capabilities to be especially relevant for many-element datasets, such as the Open Catalyst Project's OC20 [14], where the combinatorial chemical space is inevitably sparsely sampled, and cross-species interactions must be learned from few direct examples. We foresee that representations constructed with variable locality, conditional on local chemistry, will be a crucial addition in this setting.

Acknowledgments and Disclosure of Funding

LLS acknowledges support from the EPSRC Syntech CDT with grant reference EP/S024220/1. Computational resources were provided by the Cambridge Service for Data Driven Discovery (CSD3), which was accessed through the University of Cambridge EPSRC Core Equipment Award EP/X034712/1. SWN acknowledges support by funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions (Grant Agreement 945357) as part of the DESTINY PhD program. AB and SWN acknowledge support from the European Union's Horizon 2020 research and innovation program under Grant Agreement 957189 (BIG-MAP). We acknowledge DeIC (Denmark) for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium.

References

[1] Chaitanya K. Joshi, Cristian Bodnar, Simon V. Mathis, Taco Cohen, and Pietro Liò. On the expressive power of geometric graph neural networks, 2023.

[2] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

[3] Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[4] Ilyes Batatia, Mario Geiger, Jose Munoz, Tess Smidt, Lior Silberman, and Christoph Ortner. A general framework for equivariant neural networks on reductive lie groups, 2023.

[1] For our models using multi-head attention, the feature dimension is increased by a factor equal to the number of heads, thus the attention models use fewer feature channels. In terms of Equation 6, $\tilde{k}_\mu$ is equal to $\tilde{k}$ times μ.

[5] Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor N. C. Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e(3)-equivariant atom-centered interatomic potentials, 2022.

[6] 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):044118, 07 2023.

[7] Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks, 2020.

[8] Philipp Thölke and Gianni De Fabritiis. Torchmd-net: Equivariant transformers for neural network based molecular potentials, 2022.

[9] J. Thorben Frank, Oliver T. Unke, and Klaus-Robert Müller. So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems, 2023.

[10] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs, 2022.

[11] SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials - Scientific Data — nature.com. `https://www.nature.com/articles/s41597-022-01882-6#citeas`. [Accessed 27-09-2023].

[12] Coupled Cluster Molecular Dynamics of Condensed Phase Systems Enabled by Machine Learning Potentials: Liquid Water Benchmark — journals.aps.org. `https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.129.226001`. [Accessed 27-09-2023].

[13] Patrick Rowe, Volker L. Deringer, Piero Gasparotto, Gábor Csányi, and Angelos Michaelides. An accurate and transferable machine learning potential for carbon. *The Journal of Chemical Physics*, 153(3):034702, 07 2020.

[14] 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*, 2021.