Performance of an Attention-based Model on Atomic Systems

Praharsh Suryadevara
ps4586@nyu.edu

Abstract

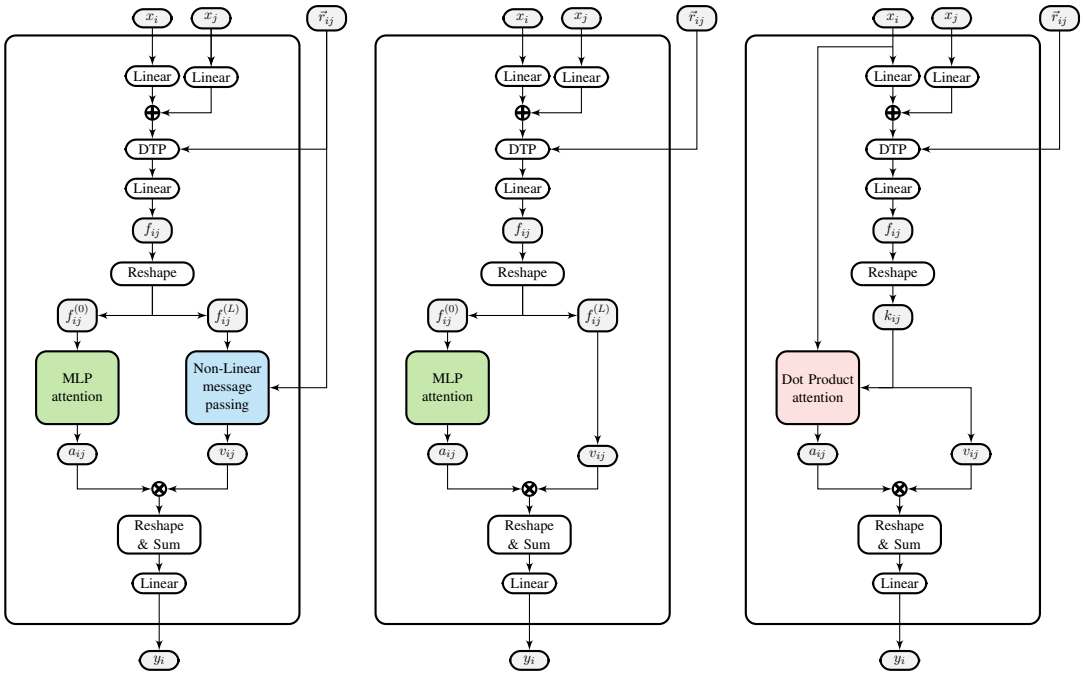
Transformer models have excelled in natural language processing, leading to applications in various domains. Liao et al. introduced the Equiformer, a model incorporating translational and rotational symmetries with a novel attention mechanism, outperforming previous GNN models on QM9, MD17, and OC20 datasets. Ablation studies by Liao et al. revealed that their equivariant graph attention mechanism surpassed dot product attention on the OC22 dataset but showed less improvement on QM9, potentially due to OC20's larger and more diverse atom types. In this report, we reproduce the Equiformer's results on aspirin in the MD17 dataset, with ablation studies demonstrating performance gains from the equivariant attention and nonlinear message passing introduced in the original paper. We also show results on other molecules in the MD17 dataset

1 Introduction

Understanding the properties of atomic systems is essential in a wide range of fields such as drug discovery, materials design, and computational chemistry. Machine Learning techniques, Machine learning (ML) techniques, particularly deep neural networks, have emerged as a powerful tool to model the complex interatomic interactions governing these systems. Specifically ML-based interatomic potentials provide accurate representations of the underlying landscapes at a fraction of the cost of traditional Density Functional Theory (DFT) methods. This has the potential to allow researchers to rapidly analyze large-scale atomic systems and discover novel molecules or materials, which would have been infeasible previously using traditional methods.

To efficiently model systems, Deep learning techniques often integrate inductive biases that leverage symmetries inherent to data to reduce training cost. For image data, this is done by convolutional neural networks that exploit the translational invariance of features to reduce the number of connections needed to model data. Atomic systems reside in a 3D Euclidean space which suggests that incorporating inductive biases to the 3D euclidean group would help us efficiently model these systems.

Transformer networks have shown significant potential in multiple domains such as Natural language processing and Computer Vision and have thus shown potential to be widely applicable. Because Transformers were originally designed for sequential data, architectural changes must be made to achieve performance gains in other fields. We expect physics to be invariant to 3D rotations, 3D translations, and inversions, transformations that generate the 3D Euclidean group. To capture these symmetries accurately in predictions, our models need to be equivariant to them. Liao et al. have introduced the following architectural changes in their paper to transformers to adapt them to atomic systems in a way that achieves state of the art results outperforming other GNN models (Batzner et al., 2022), (Gasteiger et al., 2022), (Schütt et al., 2021), that previous equivariant transformers have not achieved (Fuchs et al., 2020), (Thölke and Fabritiis, 2022). First, they replace operations in the original transformer with their equivariant counterparts. Second, they introduce a new equivariant graph attention that combines features considering group symmetries, replacing the dot product attention with. They also conduct ablation studies to show that each of the proposed architectural changes they introduce contributes to the model's performance. They have shown that the architectural changes they propose adds significant performance gain on tasks in the OC20 dataset (Chanussot et al., 2021) but not as much in the QM9 (Ruddigkeit et al., 2012) dataset with ablation studies. However they have not mea-



(a) Nonlinear message passing and MLP attention

(b) Linear message passing and MLP attention

(c) Dot Product attention

Figure 1: Overview of the Attention Mechanism Architectures evaluated in the ablation study, Dot Product attention, MLP attention and Non-linear message passing are illustrated in Figure 2

sured whether there exists a performance gap in the MD17 dataset(Chmiela et al., 2017). In this report we run ablation studies on the equiformer model on the Aspirin molecule in the MD17 dataset and show that there is a performance gain due to the architectural choices they make. We further report for training for other molecules in the dataset, which we have not been able to run for enough epochs to reproduce results due to GPU compute limits. We also comment on the architectural changes we expect are needed for running this model on larger molecules such as those in the MD22 dataset (Tran et al., 2023).

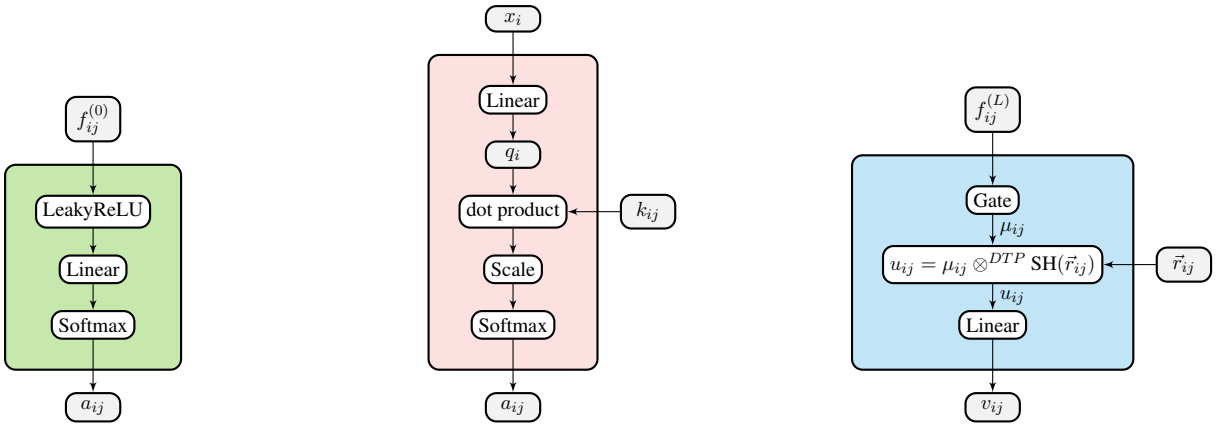
2 Methods

2.1 Dataset and Metrics

The MD17 dataset consists of separate tasks on eight organic molecules with a maximum of 21 atoms and 4 chemical elements. The data consists of positions, and corresponding energies and forces for each atom in the molecule for different configurations. These configurations were obtained using Path Integral Molecular dynamics simulations done energies and forces calculated with Density functional Theory. We use 950 configurations for training, 50 for validation and the rest for training as done in the original paper. The goal is to predict the energies and forces, for which the Mean Average Error (MAE) between the prediction and the ground truth was minimized. Considering that there is MAE in both Energy and Force, in practice, we take minimize a weighted sum. This has been done in the original paper as well as other papers evaluating their models on this Dataset. We focus on Aspirin because it has the largest number of atoms in the dataset and all models whose performance was compared with equiformer had data on this molecule in the original paper.

2.2 Preprocessing

Using a cutoff radius, the list of positions can be viewed as a 3D graph where each node corresponds to an atom, with information about atom type z_i, and we have displacement vectors between two atoms i, and j, $\vec{x_i} - \vec{x_j} = \vec{r_{ij}}$ corresponding to each edge. Using only z_i and $\vec{r_{ij}}$ addresses translational symmetry since they do not change with translations. The vector $\vec{r}_{ij}$ are projected into $O(3)$ equivariant type-L vectors by using spherical harmonics $f_{ij}^{(L)} = Y^{(L)}\left(\frac{\vec{r_{ij}}}{|\vec{r_{ij}}|}\right)$. We refer to (e3nn team, 2022) for discussion regarding Rotational Equivariance in the context of Neural networks. We also use a cutoff L_{max} to prevent



(a) MLP Attention Block, b_i are learnable weights the same dimension as $f_{ij}^{(0)}$. They correspond to a linear layer without bias

(b) Dot Product Attention Block, The value pair v_{ij} is multiplied later

(c) Non-Linear message passing Block, SH corresponds to a Spherical Harmonic Decomposition, and DTP corresponds to a depth-wise tensor product

Figure 2: Changed Blocks in Ablation studies

			Results		(Liao and Smidt, 2023)	
Attention Mechanism	Message passing	Parameters	Energy MAE	Force MAE	Energy MAE	Force MAE
MLP Attention	Non-Linear	3.5M	**5.4**	**7.2**	5.3	7.2
MLP Attention		2.9M	**5.4**	8.2	-	-
DP Attention		3.3M	5.8	9.2	-	-

Table 1: Ablation study MAE results on the test set on Aspirin in the MD17 dataset, for $L_{max} = 2$, Values are in meV

computations from becoming intractable.

2.3 Model and Parameters

We use the equiformer model from from Liao et, al, with attention mechanics as illustrated in figure 1. The novel contributions (e3nn team, 2022) of the paper were combining MLP attention (Fig. 2a) and Nonlinear message passing (Fig. 2c), both of which we ablate to evaluate performance contributions. We set $L_{max} = 2$, use 4 attention heads and 6 Transformer blocks. For the Optimizer we used AdamW, with a batch size of 4, and a maximum learning rate set to 5×10^{-4} except for ethanol and benzene where it was set to 1×10^{-4} as done in the original paper.

2.4 Experiments

Models trained and checkpoints can be found here

Ablation Studies on Aspirin in MD17. We trained three separate models from scratch on Aspirin in the MD17 dataset, the second with MLP attention and Nonlinear message passing, third with just MLP attention for 1500 epochs. Training time took roughly 36 hours for the full model, 31 hours for using just MLP attention and 29 hours for dot product attention using one NVIDIA Tesla V100 with 16GB memory. Since we had a wall time of 24 hours, we restarted from checkpoints while training. We verify results on the full model, and evaluate performance due to architectural changes in section.

Training on other Molecules in MD17 We also the full model on Benzene, Ethanol, Malonaldehyde, Naphthalene, Salicyclic Acid and Toluene in the dataset. The training ran for around 850-1000 epochs before hitting the wall time of 24 hours, we couldn't train for longer since we hit GPU quotas. We compare results for the model at 850 epochs to the results from the paper. We also attempted training on the DNA base pair AT-AT,

	Epochs	Benzene	Ethanol	Malona-ldehyde	Naphth-alene	Salicyclic Acid	Toluene
Results	850	2.5	6.5	4.8	7.1	40.4	13
(Liao and Smidt, 2023)	1500	2.2	2.2	3.3	3.7	4.5	3.8

Table 2: Energy MAE Results on the test for Benzene, Ethanol, Malonaldehyde, Naphthalene, Salicyclic Acid and Toluene in the MD17 datasets $L_{max} = 2$. Values are in meV

	Epochs	Benzene	Ethanol	Malona-ldehyde	Naphth-alene	Salicyclic Acid	Toluene
Results	850	7.3	4.7	7.0	3.0	6.0	2.89
(Liao and Smidt, 2023)	1500	6.6	3.1	5.8	2.1	4.1	2.1

Table 3: Force MAE Results on the test for Benzene, Ethanol, Malonaldehyde, Naphthalene, Salicyclic Acid and Toluene in the MD17 datasets $L_{max} = 2$. Values are in meV Å^{-1}

and Ac-Ala3-NHMe from the MD22 dataset before hitting memory limits on the GPU. Note that the only molecule in MD17 that we don't have results on is Uracil.

3 Results and Discussion

Ablation Studies on Aspirin in MD17 We reproduce the results of (Liao and Smidt, 2023) and further show that the MLP attention and Non-Linear message passing improve performance on aspirin in MD17 as shown in Table 1. We see a slight difference in the Energy MAE for the full model, which may be due to seeding differences due to loading from a checkpoint. We note that Non-Linear message passing does not improve on Force MAE as compared to linear message passing. We believe this is because MLP attention acts on scalars ($L = 0$) as compared to Non-Linear message passing which affects higher order vectors. Hence using just MLP attention can help with predicting $L = 0$ scalars like force that are invariant to rotations. (Liao and Smidt, 2023) see improvements due to MLP attention on OC20 but not so much on QM9. In their appendix they surmise that this is because QM9 contains less atoms and less diverse atom types, and hence dot product attention is enough. However we see performance improvements on Aspirin in MD17, A molecule with similar number of atoms as other molecules in QM9, suggesting that the task and training may also play a role. However, we need ablation studies on all of MD17, considering that the improvements we see may be specific to this molecule.

Results on other Molecules in MD17: We obtain the Energy (Table 2) and Force MAE (Table 3) at 850 Epochs, for the molecules mentioned. We the results on some molecules such as Benzene and Malonaldehyde have converged significantly rapidly whereas others such as Salicyclic acid are still far from the final results. On Salicyclic acid we have noticed large fluctuations in the energy around 850 epochs. We believe the results on Benzene have converged quickly because the molecule is symmetric and the true ground state is much more stable than any other configuration.

4 Conclusion

In this report we have reproduced the results of (Liao and Smidt, 2023) on Aspirin in the MD17 dataset, and conducted ablation studies to show that the architectural changes they introduce improve performance of their models. We have also shown and compared results on Benzene, Ethanol, Malonaldehyde, Naphthalene, Salicyclic Acid and Toluene in the MD17 dataset

The trained models, code, and logs can be found here. Note that the results in the training logs are in kcal mol $^{-1}$ kcal mol $^{-1}$ Å^{-1} for Energy MAE and force MAE respectively and need to be converted to meV and meV Å^{-1} for comparison

References

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. 2022. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. 2021. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. 2017. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015.

e3nn team. 2022. e3nn: a modular pytorch framework for euclidean neural networks.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. 2020. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981.

Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C Lawrence Zitnick, and Abhishek Das. 2022. Gemnet-oc: developing graph neural networks for large and diverse molecular simulation datasets. *arXiv preprint arXiv:2204.02782*.

Yi-Lun Liao and Tess Smidt. 2023. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. 2012. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875.

Kristof Schütt, Oliver Unke, and Michael Gastegger. 2021. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR.

Philipp Thölke and Gianni De Fabritiis. 2022. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Félix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence Zitnick. 2023. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084.