Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs

Yi-Lun Liao, Tess Smidt
Massachusetts Institute of Technology

github.com/atomicarchitects/equiformer

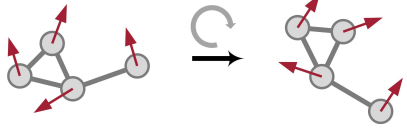
Motivation and Contribution

(1) Equivariant and invariant networks have demonstrated the importance of incorporating 3D-related inductive biases in learning representations of 3D atomistic systems.
(2) A parallel line of research lies in generalizing Transformers to many domains such as vision and graphs and achieves widespread success.
(3) This naturally leads to the question of how we can apply Transformer-like networks to 3D atomistic systems.
(4) We propose Equiformer, an equivariant graph neural network that combines the inductive bias of equivariance with the strength of Transformers.

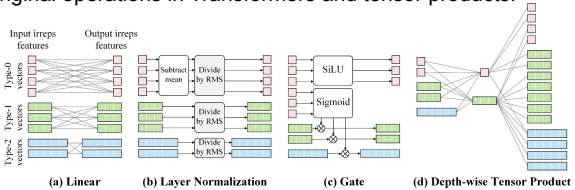
Proposed Method

Equivariant Features and Operations

Equivariant features (e.g., type-1 vectors) are rotated accordingly when input graphs are rotated.



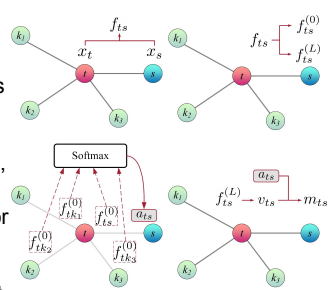
Equivariant operations include the equivariant version of original operations in Transformers and tensor products.



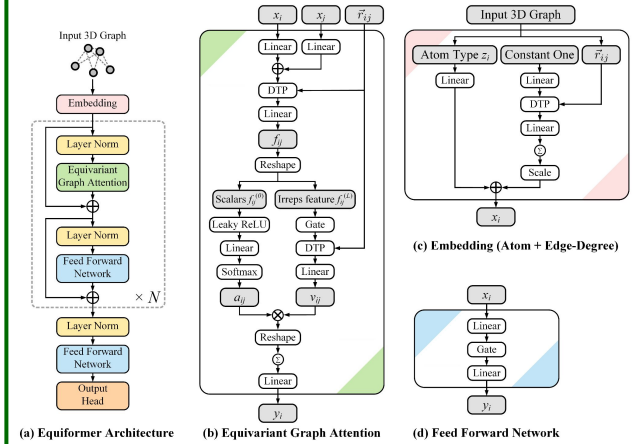
(a) Linear (b) Layer Normalization (c) Gate (d) Depth-wise Tensor Product

Equivariant Graph Attention

(1) Equivariant graph attention improves upon typical attention in Transformers.
(2) The feature sent from node s to node t is:
$m_{ts} = a_{ts} \times v_{ts}$, where
a_{ts}: attention weights (scalars),
v_{ts}: value (irreps features).
(3) Both are obtained with tensor products and non-linear functions.
(4) Steps are shown on the right.



Overall Architecture



(a) Equiformer Architecture (b) Equivariant Graph Attention (c) Embedding (Atom + Edge-Degree) (d) Feed Forward Network

Result

QM9 and MD17

Equiformer achieves better results than previous equivariant Transformers and other equivariant message passing networks and invariant message passing networks.

Methods	Task Units	α a_0^3	$\Delta\varepsilon$ meV	$\varepsilon_{\text{HOMO}}$ meV	$\varepsilon_{\text{LUMO}}$ meV	μ D	C_ν cal/mol K	G meV	H meV	R^2 a_0^2	U meV	U_0 meV	ZPVE meV
SE(3)-Transformer†		.142	53	35	33	.051	.054						
PaiNN		.045	46	28	20	.012	.024	**7.35**	**5.98**	.066	**5.83**	**5.85**	1.28
TorchMD-NET		.059	36	20	18	**.011**	.026	7.62	6.16	**.033**	6.38	6.15	1.84
SphereNet		.046	32	23	18	.026	**.021**	8	6	.292	7	6	**1.12**
SEGNN†		.060	42	24	21	.023	.031	15	16	.660	13	15	1.62
EQGAT		.053	32	20	16	**.011**	.024	23	24	.382	25	25	2.00
Equiformer		.046	**30**	**15**	**14**	**.011**	.023	7.63	6.63	.251	6.74	6.59	1.26

Mean absolute error results on QM9 testing set. † denotes using different data partitions.

Methods	Aspirin energy	Aspirin forces	Benzene energy	Benzene forces	Ethanol energy	Ethanol forces	Malonaldehyde energy	Malonaldehyde forces	Naphthalene energy	Naphthalene forces	Salicylic acid energy	Salicylic acid forces	Toluene energy	Toluene forces	Uracil energy	Uracil forces
DimeNet	8.8	21.6	3.4	8.1	2.8	10.0	4.5	16.6	5.3	9.3	5.8	16.2	4.4	9.4	5.0	13.1
PaiNN	6.9	14.7	-	-	2.7	9.7	3.9	13.8	5.0	3.3	4.9	8.5	4.1	4.1	4.5	6.0
TorchMD-NET	**5.3**	11.0	2.5	8.5	2.3	4.7	3.3	7.3	**3.7**	2.6	**4.0**	5.6	**3.2**	2.9	**4.1**	4.1
NequIP ($L_{\text{max}} = 3$)	5.7	8.0	-	-	**2.2**	3.1	3.3	5.6	4.9	**1.7**	4.6	**3.9**	4.0	**2.0**	4.5	**3.3**
Equiformer ($L_{\text{max}} = 3$)	**5.3**	**6.6**	2.5	8.1	**2.2**	**2.9**	**3.2**	**5.4**	4.4	2.0	4.3	**3.9**	3.7	2.1	4.3	3.4

Mean absolute error results on MD17 testing set.

OC20

(1) When trained with IS2RE data, Equiformer improves upon previous state-of-the-art models.
(2) When trained with IS2RE + IS2RS data, Equiformer improves upon GNS and Graphormer and has **2.3× to 15.5× less training time**.

Methods	Energy MAE (eV) ↓ ID	OOD Ads	OOD Cat	OOD Both	Average	Training time (GPU-days)
GNS + Noisy Nodes	0.4219	0.5678	0.4366	**0.4651**	0.4728	56 (TPU)
Graphormer†	**0.3976**	0.5719	**0.4166**	0.5029	0.4722	372 (A100)
Equiformer + Noisy Nodes	0.4171	**0.5479**	0.4248	0.4741	**0.4660**	**24** (A6000)

Results on OC20 IS2RE testing set when IS2RS is adopted during training. † denotes using ensemble of models trained on both IS2RE training and validation sets.

Ablation Study

(1) Non-linear attention (MLP attention) improves upon linear attention (dot product attention).
(2) Non-linear messages improves upon linear messages.

Index	Non-linear message passing	MLP attention	Dot product attention	Energy MAE (eV) ↓ ID	OOD Ads	OOD Cat	OOD Both	Average	Training time (minutes/epoch)	Number of parameters
1	✓	✓		0.5088	0.6271	0.5051	0.5545	0.5489	130.8	9.12M
2		✓		0.5168	0.6308	0.5088	0.5657	0.5555	91.2	7.84M
3			✓	0.5386	0.6328	0.5297	0.5692	0.5689	99.3	8.72M

Ablation study results on OC20 IS2RE validation set.