Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs

Yi-Lun Liao   Tess Smidt

Atomic Architects
Massachusetts Institute of Technology

https://github.com/atomicarchitects/equiformer

2

Outline

- Introduction

- Equiformer

  – Equivariant Features and Operations

  – Equivariant Graph Attention

  – Overall Architecture

- Experiments

- Conclusion

3

Outline

- **Introduction**

- Equiformer

  – Equivariant Features and Operations

  – Equivariant Graph Attention

  – Overall Architecture

- Experiments

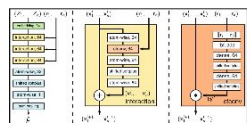
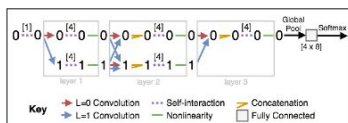
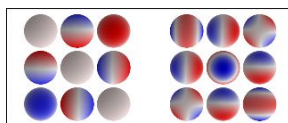
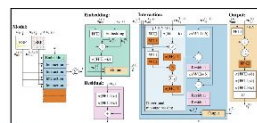
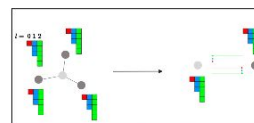
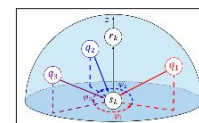
- Conclusion

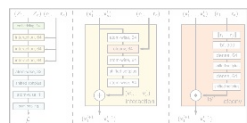
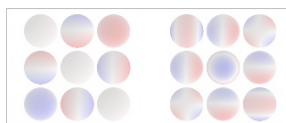
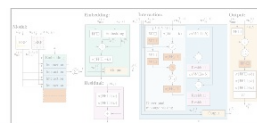
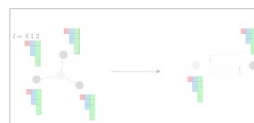
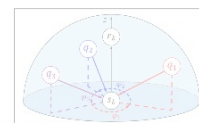
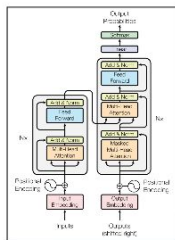
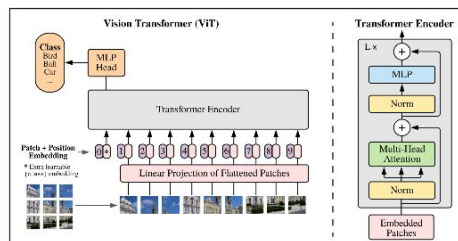
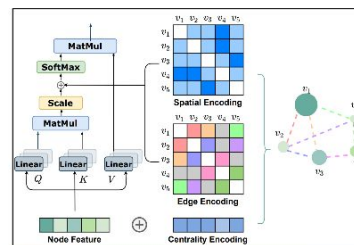
4

Introduction

- Incorporating equivariance/invariance to E(3) transformations is important to learning representations of 3D atomistic systems.

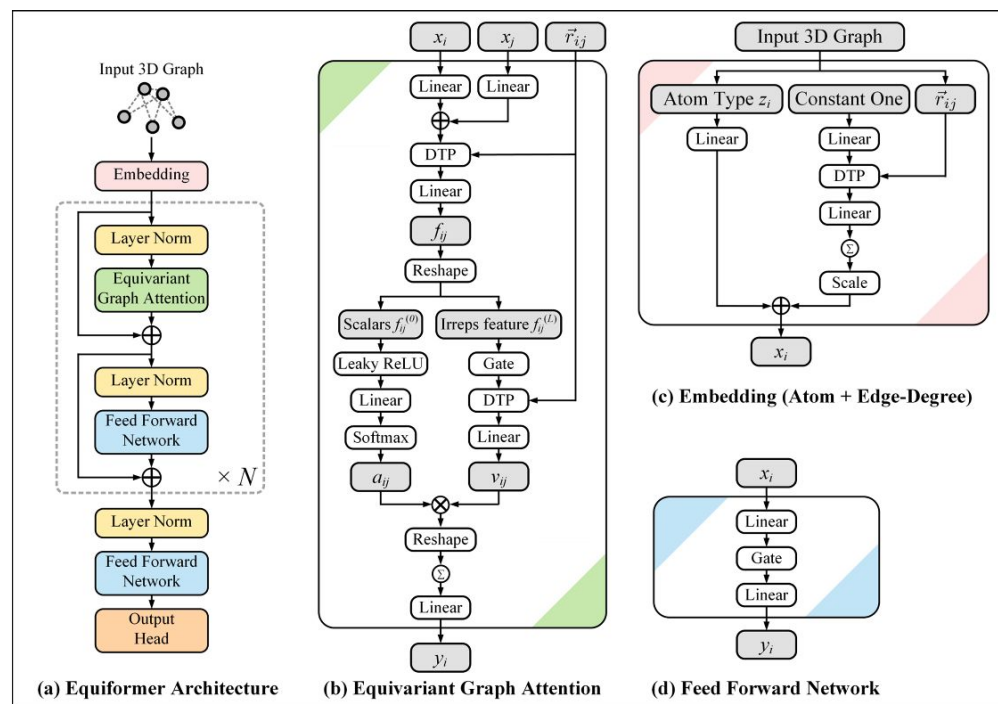
SchNet | Tensor Field Network | 3D Steerable CNN | DimeNet | NequIP | SphereNet

5

Introduction

- A parallel line of research lies in applying Transformers to different domains and has demonstrated widespread success.

SchNet | Tensor Field Network | 3D Steerable CNN | DimeNet | NequIP | SphereNet

Transformer

Vision Transformer

Graphormer

6
Introduction

• We present Equiformer, an equivariant graph neural network combining the inductive bias of equivariance with the strength of Transformer.

Equiformer

7

Outline

- Introduction

- Equiformer

  – Equivariant Features and Operations

  – Equivariant Graph Attention

  – Overall Architecture

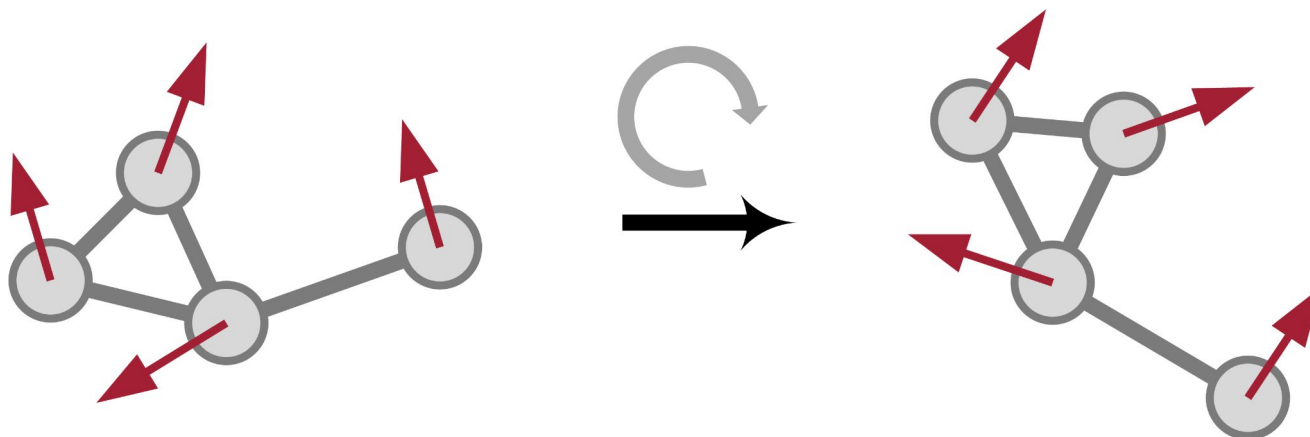
- Experiments

- Conclusion

8

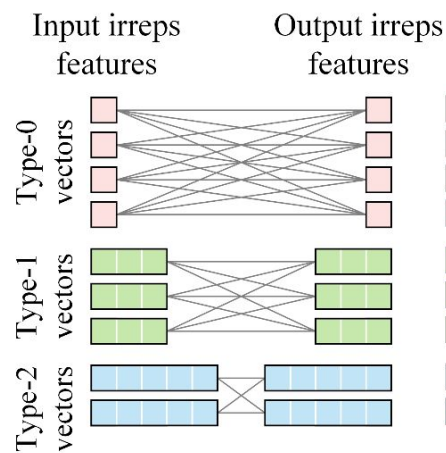
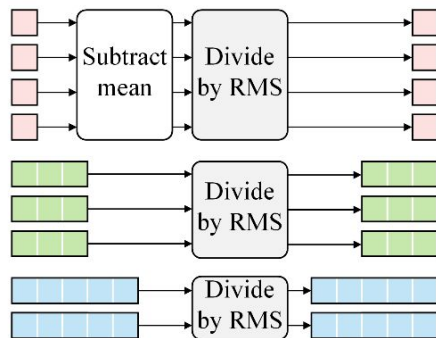
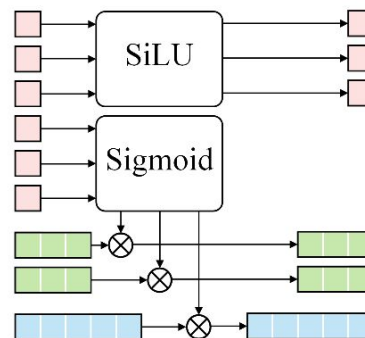
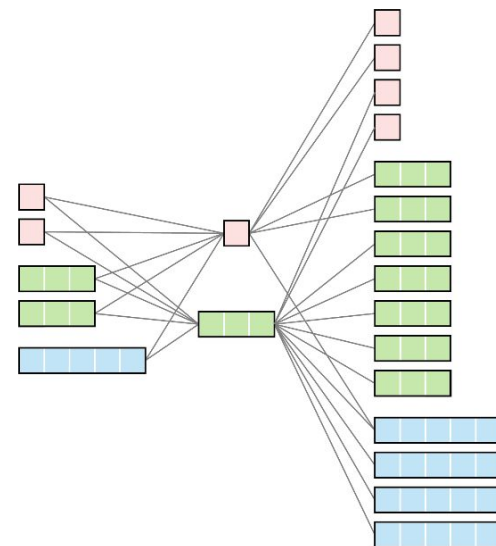
Equiformer – Equivariant Irreps Features

- To incorporate 3D-related inductive biases (e.g., rotational equivariance), we use vector spaces of irreducible representations (irreps) as internal feature representations.

- Irreps features contain vectors of different types (degrees) L.

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

9

Equiformer – Equivariant Operations

- To operate on equivariant irreps features, we:
  - replace original operations in Transformers with their equivariant counterparts and
  - use additional tensor product operations.

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

10

Outline

- Introduction

- **Equiformer**

  – Equivariant Features and Operations

  – **Equivariant Graph Attention**

  – Overall Architecture

- Experiments

- Conclusion

11
Equiformer – Equivariant Graph Attention

- We propose equivariant graph attention, which:
  - improves the expressivity of attention used by original Transformers and
  - consists of tensor products, non-linear attention and non-linear messages.

- The feature sent from node s to node t is:
$$m_{ts} = a_{ts} \times v_{ts}$$

  - a_{ts} : attention weights (scalars)
  - v_{ts} : value vectors (irreps features or geometric tensors)
  - Both are obtained with tensor products and non-linear functions.

12

Outline

- Introduction

- **Equiformer**

  - Equivariant Features and Operations

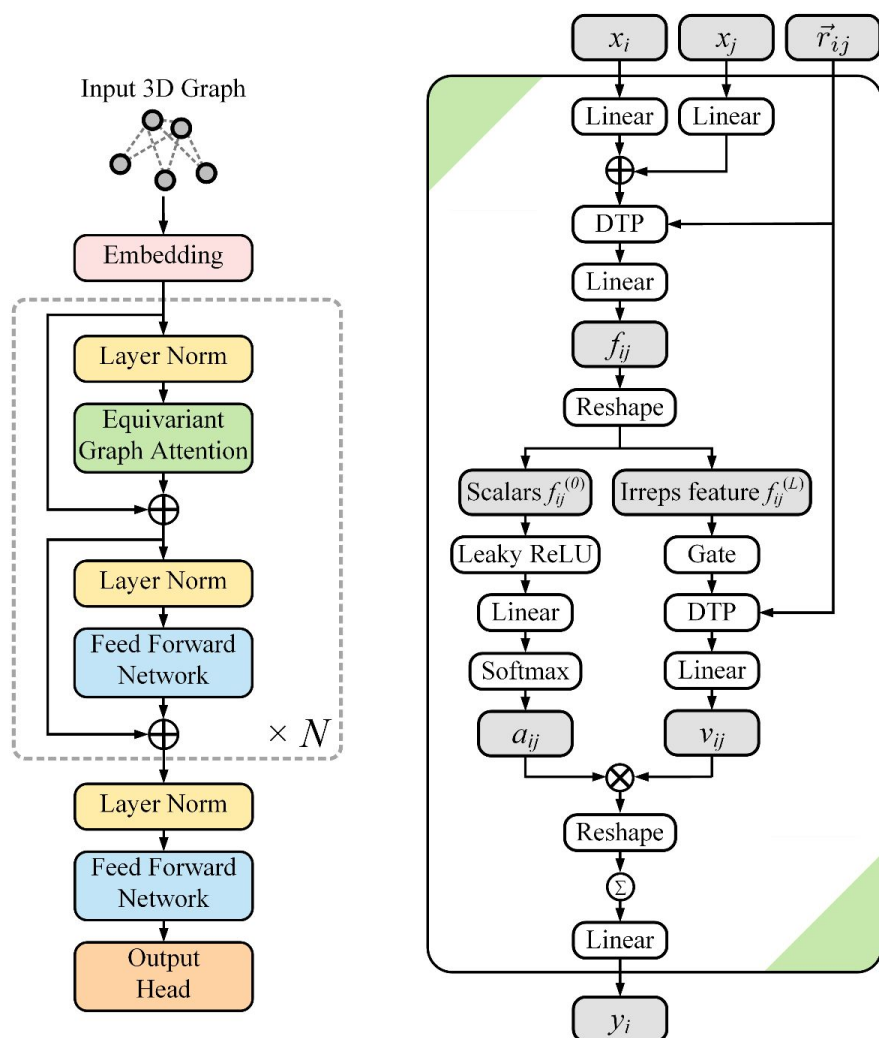
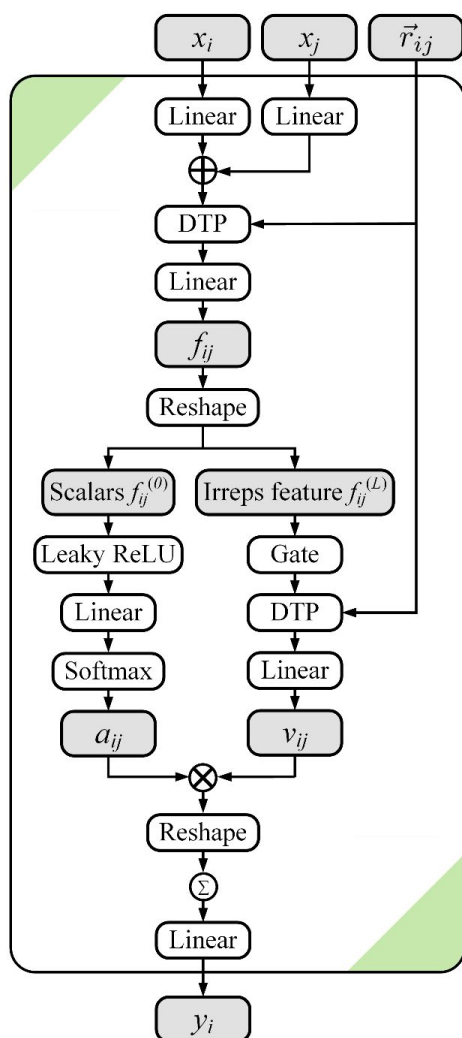
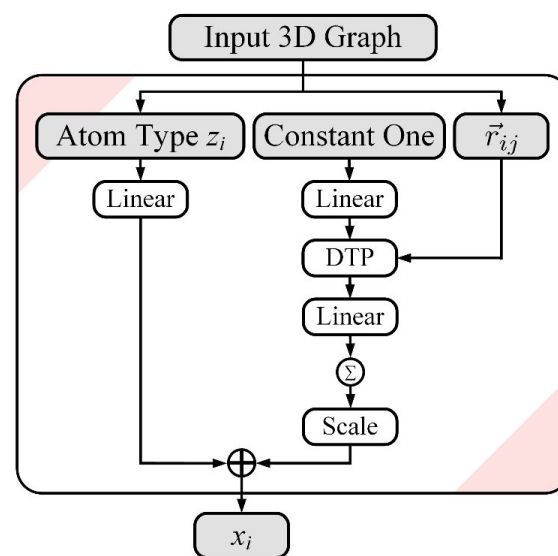
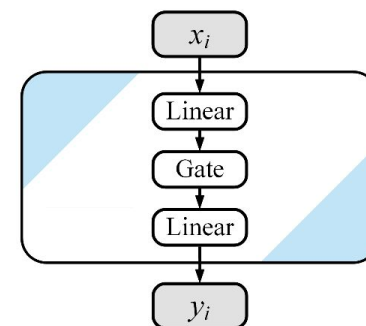
  - Equivariant Graph Attention

  - **Overall Architecture**

- Experiments

- Conclusion

13

Equiformer – Overall Architecture

(a) Equiformer Architecture

(b) Equivariant Graph Attention

(c) Embedding (Atom + Edge-Degree)

(d) Feed Forward Network

14

Outline

- Introduction

- Equiformer

  – Equivariant Features and Operations

  – Equivariant Graph Attention

  – Overall Architecture

- **Experiments**

- Conclusion

15

Experiments – QM9

- The task is to predict quantum properties of small molecules given their atom types and 3D positions.

- Equiformer achieves overall better results compared to previous equivariant Transformers and other equivariant message passing networks and invariant message passing networks.

Methods	Task Units	α a_0^3	$\Delta\varepsilon$ meV	ε_{HOMO} meV	ε_{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.

16

Experiments – MD17

- The dataset consists of molecular dynamics simulations of small organic molecules.

- The task is to predict their energy and forces given atom types and 3D positions.

- Equiformer achieves overall better results compared to previous equivariant Transformers and other equivariant message passing networks and invariant message passing networks.

Methods	Aspirin		Benzene		Ethanol		Malonaldehyde		Naphthalene		Salicylic acid		Toluene		Uracil	
	energy	forces	energy	forces	energy	forces	energy	forces	energy	forces	energy	forces	energy	forces	energy	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 ($L_{max} = 3$)	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 ($L_{max} = 3$)	**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.

17

Experiments – OC20

- The task of IS2RE (initial structure to relaxed energy) is to predict the energy of a relaxed structure given an initial structure.

- With the same setting, Equiformer improves upon previous works.

Methods	Energy MAE (eV) ↓				
	ID	OOD Ads	OOD Cat	OOD Both	Average
DimeNet++	0.5621	0.7252	0.5756	0.6613	0.6311
PaiNN	0.575	0.783	0.604	0.743	0.6763
SpinConv	0.5583	0.7230	0.5687	0.6738	0.6310
SphereNet	0.5625	0.7033	0.5708	0.6378	0.6186
SEGNN	0.5327	0.6921	0.5369	0.6790	0.6101
Equiformer	**0.5037**	**0.6881**	**0.5213**	**0.6301**	**0.5858**

Results on OC20 IS2RE testing set.

18

Experiments – OC20

- When trained with additional relaxed structure data, Equiformer improves upon GNS + Noisy Nodes and Graphormer.

- Equiformer has 2.3× to 15.5× less training time.

Methods	Energy MAE (eV) ↓					Training time (GPU-days)
	ID	OOD Ads	OOD Cat	OOD Both	Average	
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.

19

Outline

- Introduction

- Equiformer

  – Equivariant Features and Operations

  – Equivariant Graph Attention

  – Overall Architecture

- Experiments

- **Conclusion**

20

Conclusion

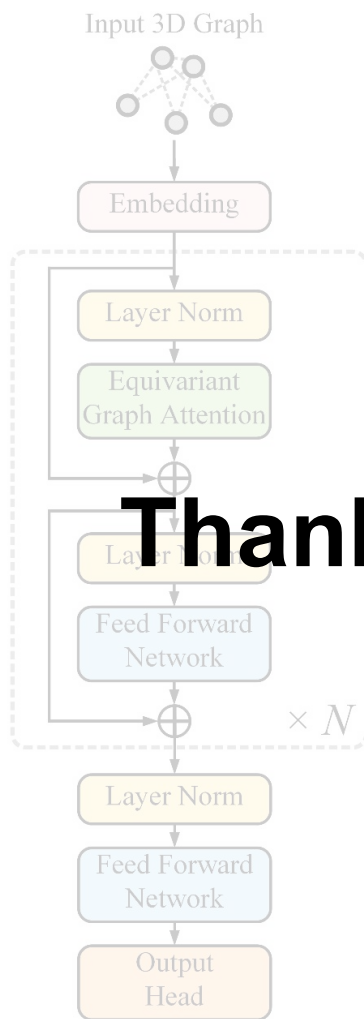
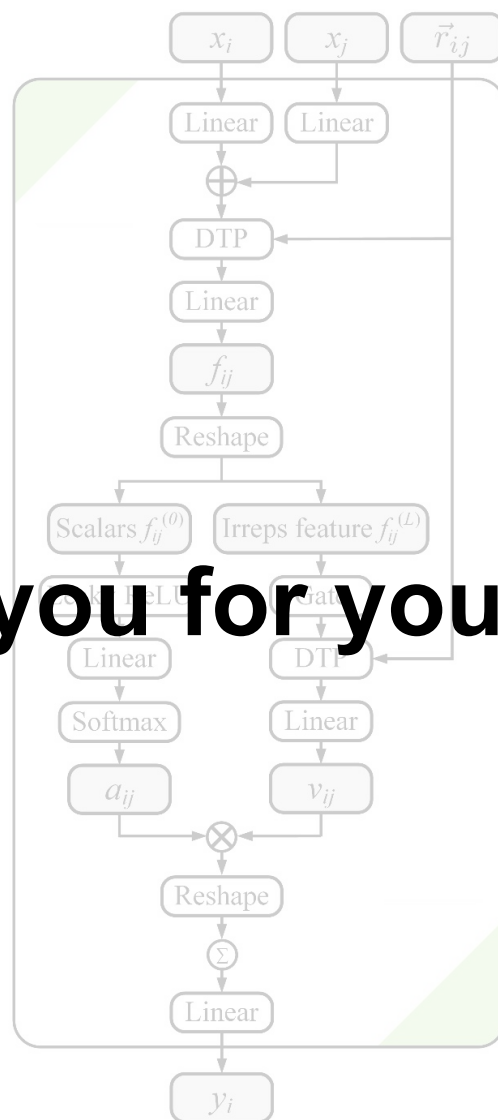
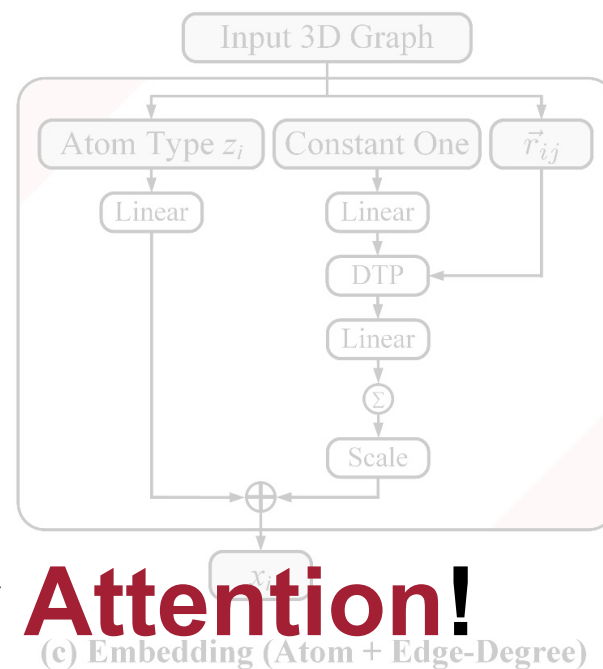
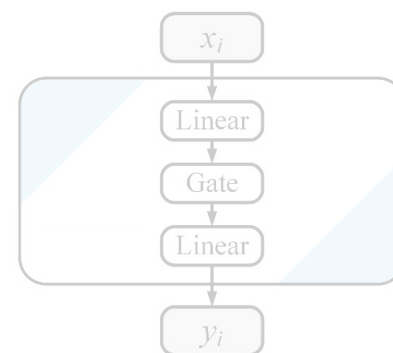
- Equiformer demonstrates the possibility of generalizing Transformers to the domain of 3D atomistic systems.

- Equivariant graph attention improves upon typical self-attention in original Transformers.

- The encouraging result leads to the question of whether other approaches developed in other domains can generalize to problems related to 3D atomistic systems.

- Official PyTorch implementation:

    https://github.com/atomicarchitects/equiformer

Thank you for your Attention!

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

Backup Slides

23

Experiments – Ablation Study

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

	Methods											
Index	Non-linear message passing	MLP attention	Dot product attention	Task Unit	α a_0^3	$\Delta\varepsilon$ meV	ε_{HOMO} meV	ε_{LUMO} meV	μ D	C_ν cal/mol K	Training time (minutes/epoch)	Number of parameters
1	✓	✓			.046	30	15	14	.011	.023	12.1	3.53M
2		✓			.051	32	16	16	.013	.025	7.2	3.01M
3			✓		.053	32	17	16	.013	.025	7.8	3.35M

Ablation study results on QM9.

	Methods			Energy MAE (eV) ↓						
Index	Non-linear message passing	MLP attention	Dot product attention	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	0.6382	0.5297	0.5692	0.5689	99.3	8.72M

Ablation study results on OC20 IS2RE validation set.

24

Experiments – Ablation Study

- Non-linear messages improves upon linear messages.

	Methods											
Index	Non-linear message passing	MLP attention	Dot product attention	Task Unit	α a_0^3	$\Delta\varepsilon$ meV	ε_{HOMO} meV	ε_{LUMO} meV	μ D	C_ν cal/mol K	Training time (minutes/epoch)	Number of parameters
1	✓	✓			.046	30	15	14	.011	.023	12.1	3.53M
2		✓			.051	32	16	16	.013	.025	7.2	3.01M
3			✓		.053	32	17	16	.013	.025	7.8	3.35M

Ablation study results on QM9.

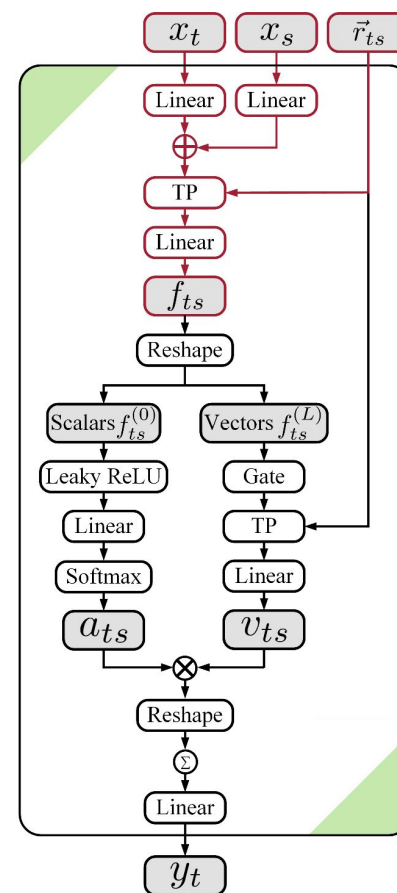
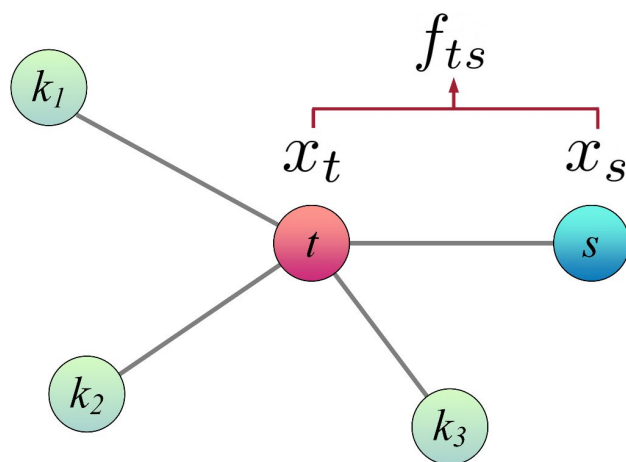
	Methods			Energy MAE (eV) ↓						
Index	Non-linear message passing	MLP attention	Dot product attention	ID	OOD Ads	OOD Cat	OOD Both	Average	Training time (minutes/epoch)	Number of parameters
1	✓	✓		0.5088	0.6271	0.5051	0.5545	0.5489	130.8	9.12M
2		✓		0.5168	0.6308	0.5088	0.5657	0.5555	91.2	7.84M
3			✓	0.5386	0.6382	0.5297	0.5692	0.5689	99.3	8.72M

Ablation study results on OC20 IS2RE validation set.

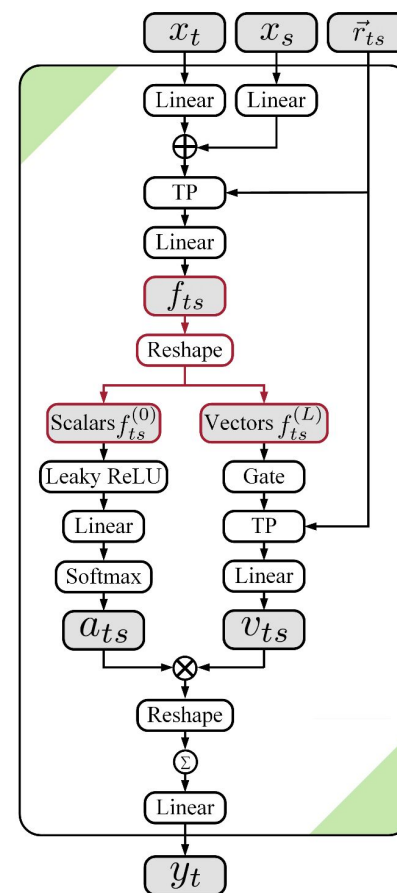
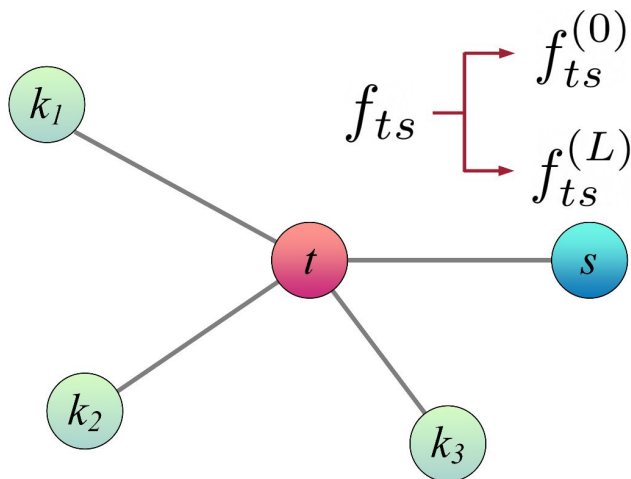
25

Steps of Equivariant Graph Attention

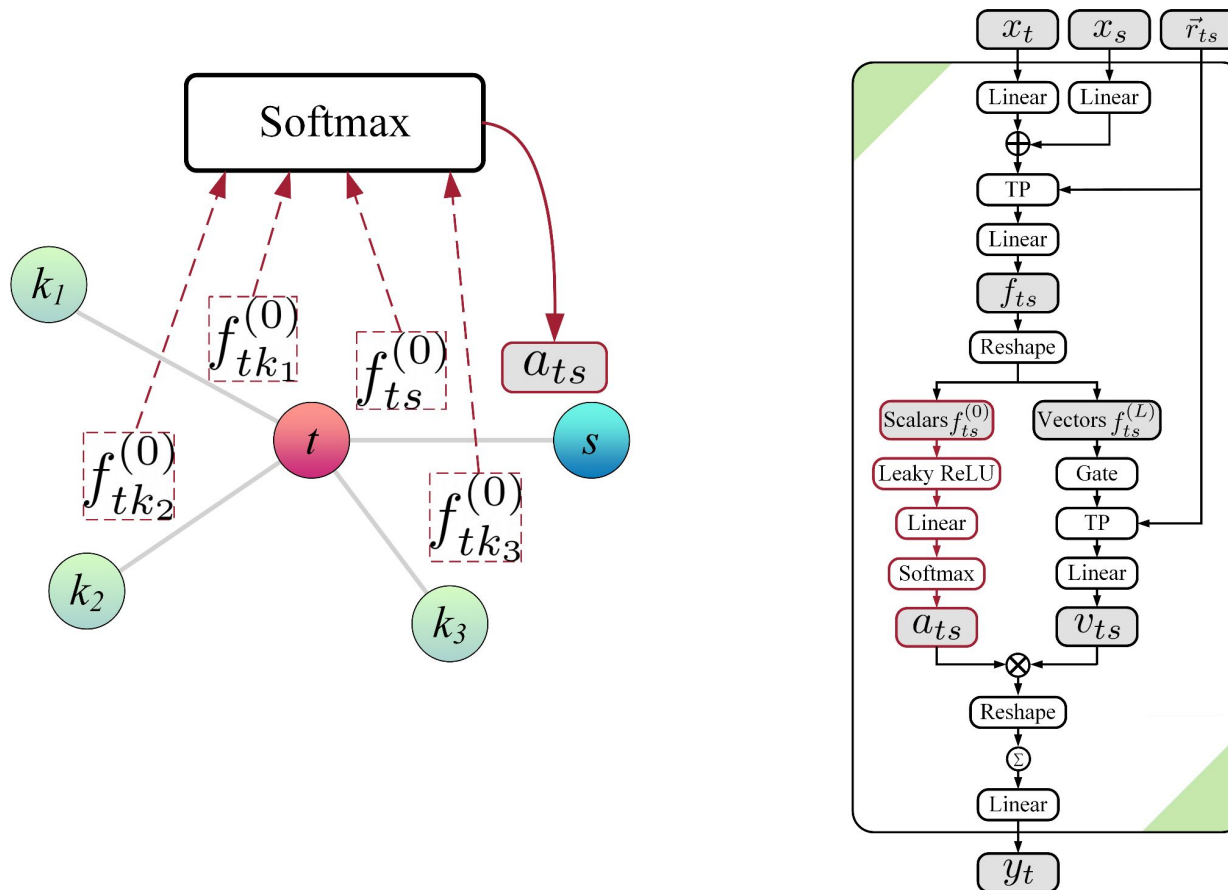
- Given nodes t and s with features x_t and x_s, we combine the features with linear layers and tensor products to obtain f_{ts}.

26

Steps of Equivariant Graph Attention

- We split feature f_{ts} into scalar features $f_{ts}^{(0)}$ and irreps features $f_{ts}^{(L)}$ (scalars + vectors of higher degrees).

27

Steps of Equivariant Graph Attention – Non-linear Attention

- Given scalar features $f_{tk}^{(0)}$ associated with neighbors of node t,
  - 1) project each feature into a scalar with **activation** and a linear layer.
  - 2) softmax over all scalars to obtain attention weights a_{ts}.

28

Steps of Equivariant Graph Attention – Non-linear Messages

- We transform irreps features $f_{ts}^{(L)}$ into value vectors v_{ts} with gate activation and tensor product.

- Finally, we multiply non-linear value vectors with non-linear attention weights.