Qualcomm

ELLIS Symposium on Large Language and Foundation Models

October 18, 2023

In search of **structure** in the age of **scale**

Johann Brehmer
Qualcomm AI Research

Two stories of deep learning

A story of structure

ICML 2016

Group Equivariant Convolutional Networks

Taco S. Cohen T.S.COHEN@UVA.NL
University of Amsterdam

Max Welling M.WELLING@UVA.NL
University of Amsterdam
University of California Irvine
Canadian Institute for Advanced Research

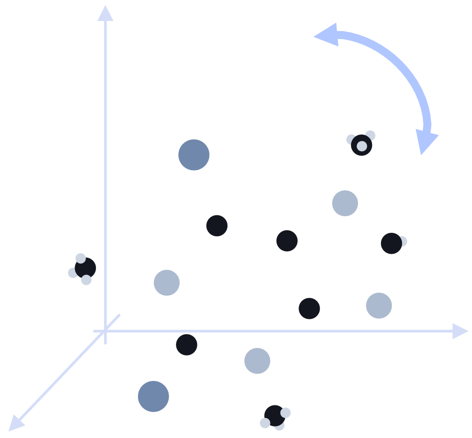
Abstract

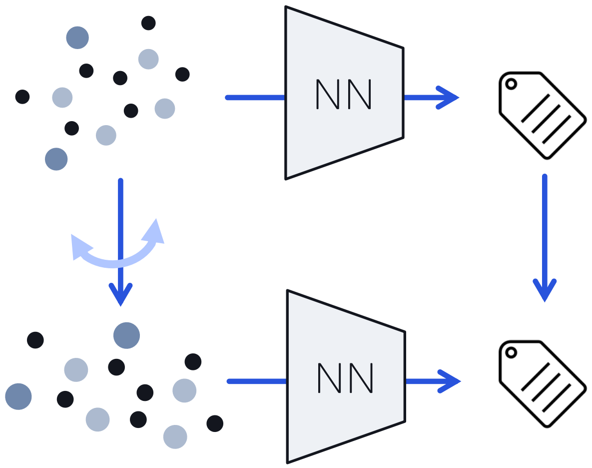
We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symmetries. G-CNNs use G-convolutions, a new type of layer that enjoys a substantially higher degree of weight sharing than regular convolution layers. G-convolutions increase the expressive capacity of the network without increasing the number of parameters. Group convolution layers are easy to use and can be implemented with negligible computational overhead for discrete groups generated by translations, reflections and rotations. G-CNNs achieve state of the art results on CIFAR10 and rotated MNIST.

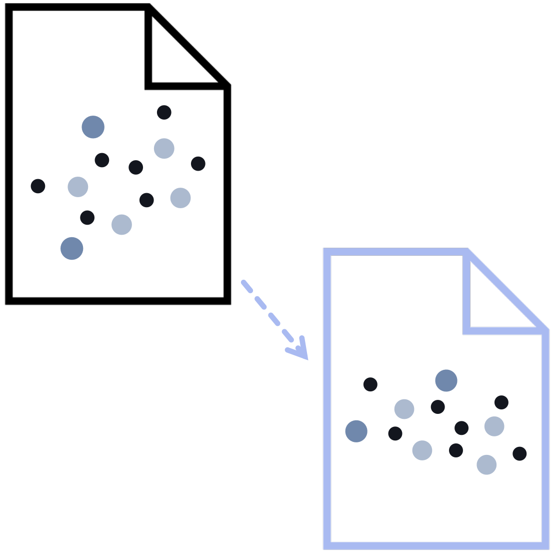
Convolution layers can be used effectively in a *deep* network because all the layers in such a network are *translation equivariant*: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature maps (at least up to edge-effects). In other words, the symmetry (translation) is preserved by each layer, which makes it possible to exploit it not just in the first, but also in higher layers of the network.

In this paper we show how convolutional networks can be generalized to exploit larger groups of symmetries, including rotations and reflections. The notion of equivariance is key to this generalization, so in section 2 we will discuss this concept and its role in deep representation learning. After discussing related work in section 3, we recall a number of mathematical concepts in section 4 that allow us to

T. Cohen & M. Welling, "Group Equivariant Convolutional Networks", ICML 2016





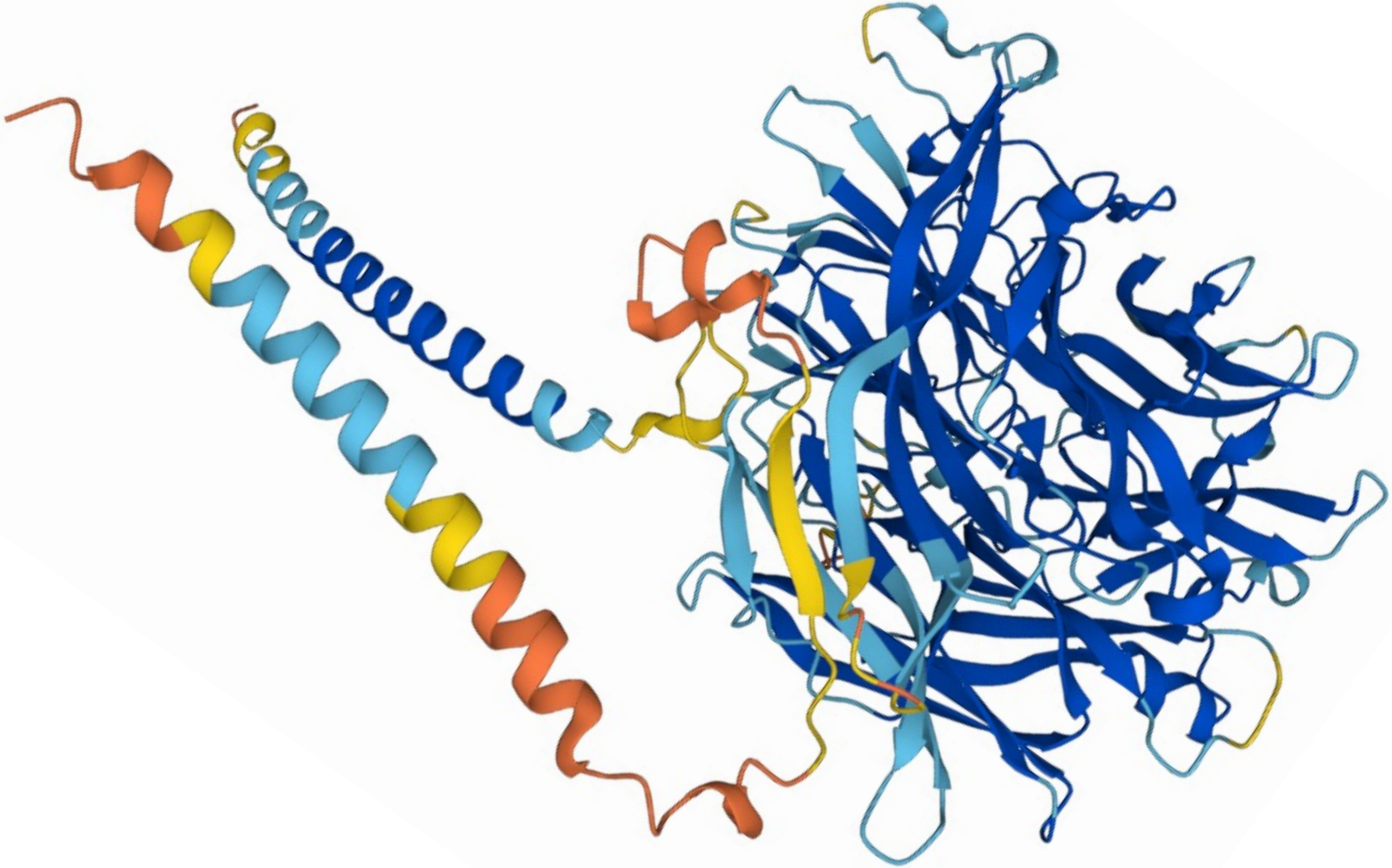


1. **Identify structure**:
   spaces and symmetries

2. **Design architectures that take structure into account**
   through equivariance / weight sharing

3. **Fill in the gaps from data**

M. Bronstein et al: "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges", arXiv:2104.13478



J. Jumper et al, "Highly accurate protein structure prediction with AlphaFold", Nature 2021;
M. Varadi et al, "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models", Nucleic Acids Research (2021).

A story of scale

NeurIPS 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

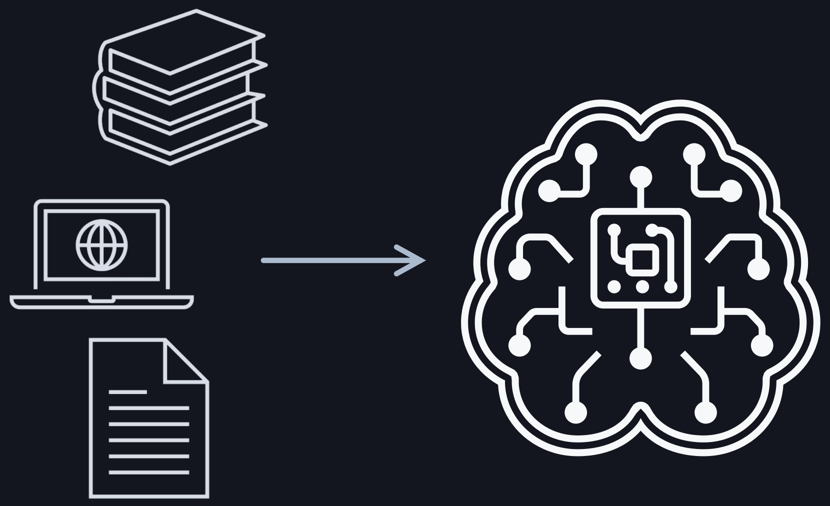
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

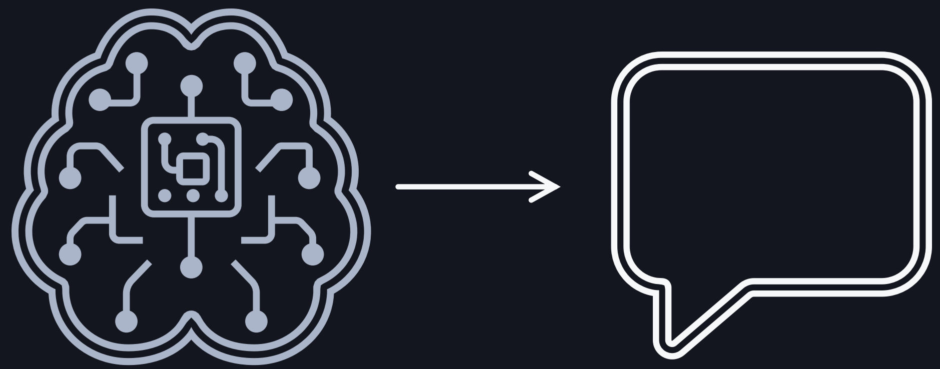
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task,

A. Vaswani et al: "Attention Is All You Need", NeurIPS 2017



1. Train big transformer on lots of data



2. Finetune for specific problems

Do strong inductive biases have a future?

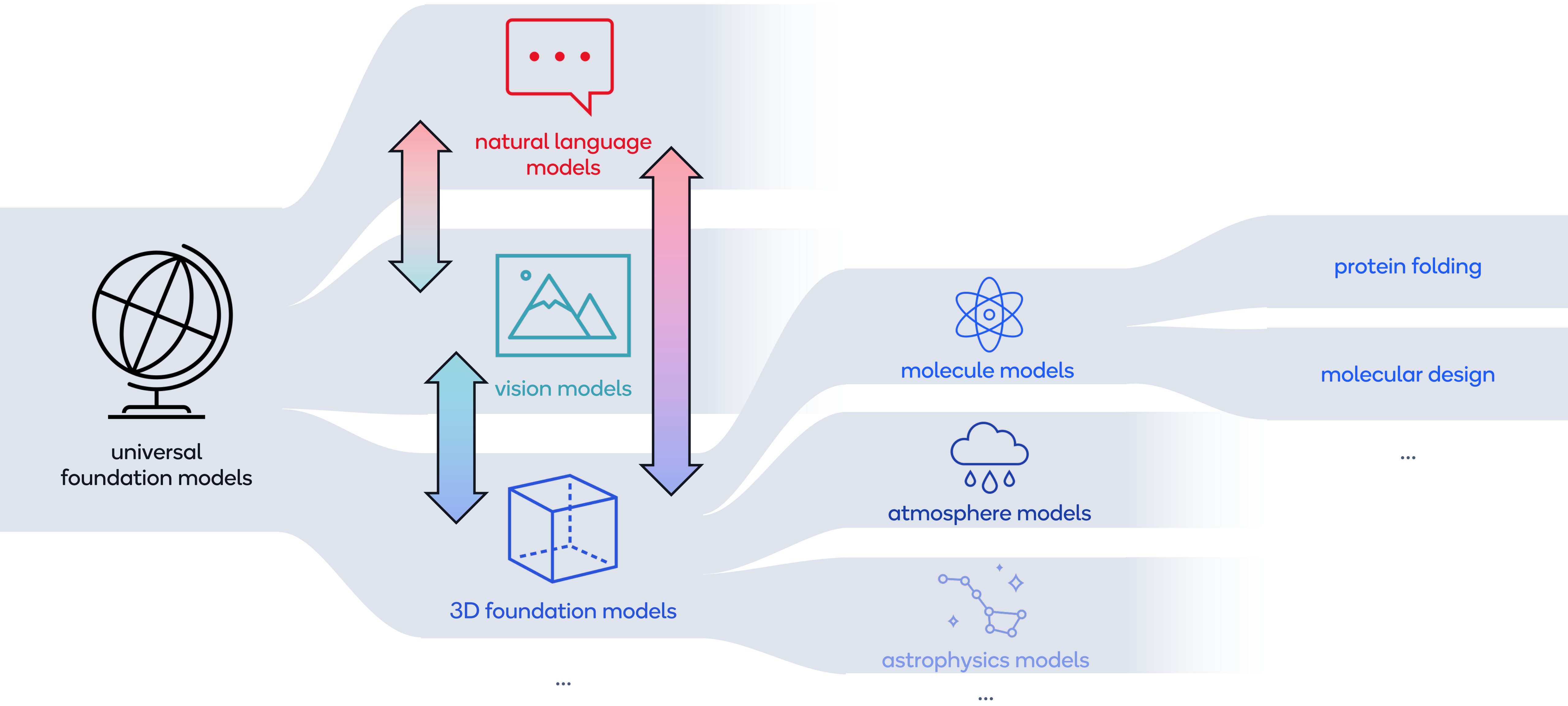
- There is **no clear evidence that LLMs can solve geometric problems** (yet?)

- **Strong inductive biases*** can help even when training on **100ks of samples** (AlphaFold, OpenCatalyst)
  *Transformers have inductive biases, too – but here we mean domain-specific structure like equivariance

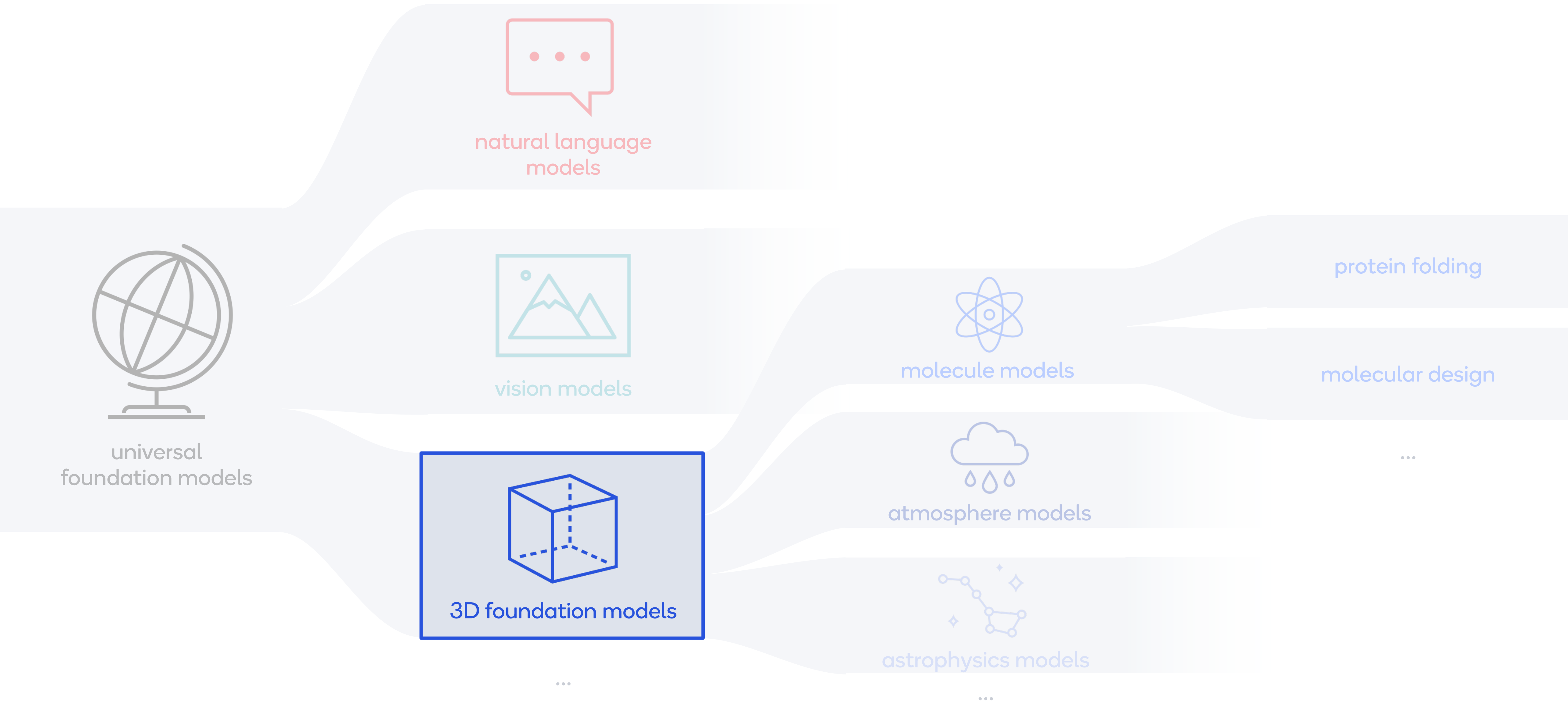
- Foundation models with strong inductive biases have not been studied in depth

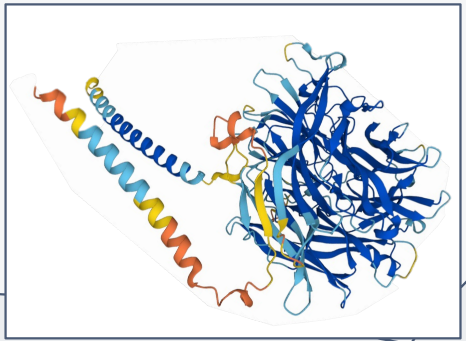
- It's not a dichotomy! Let's find out what happens when we **combine scale and structure**?

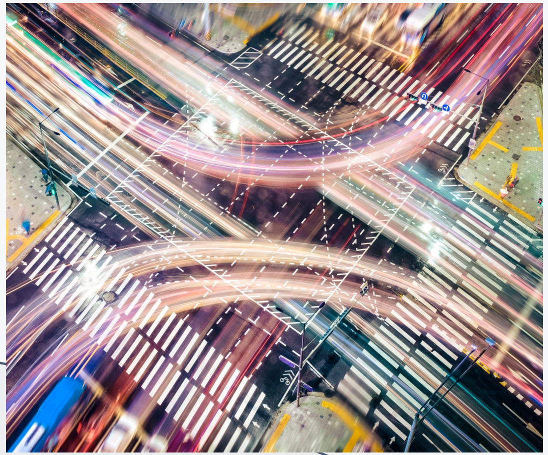
Ali Borji et al, "A Categorical Archive of ChatGPT Failures", arXiv:2302.03494
Nouha Dziri et al, "Faith and Fate: Limits of Transformers on Compositionality", arXiv: 2305.18654;
Karthik Valmeekam et al., "Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)", arXiv:2206.10498
Lawrence Zitnick et al, "An Introduction to Electrocatalyst Design using Machine Learning for Renewable Energy Storage", arXiv:2010.09435



natural language models
vision models
3D foundation models
...
universal foundation models
molecule models
atmosphere models
astrophysics models
...
protein folding
molecular design
...


natural language models
vision models
universal foundation models
3D foundation models
...
molecule models
atmosphere models
astrophysics models
...
protein folding
molecular design
...





3D space

Positions

Movement

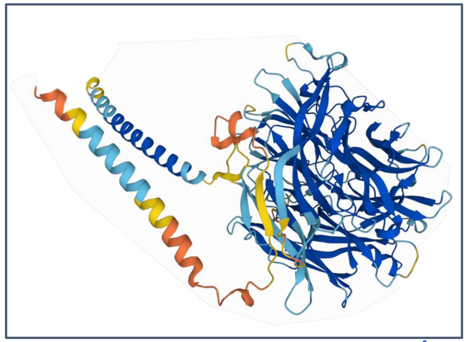
Distances

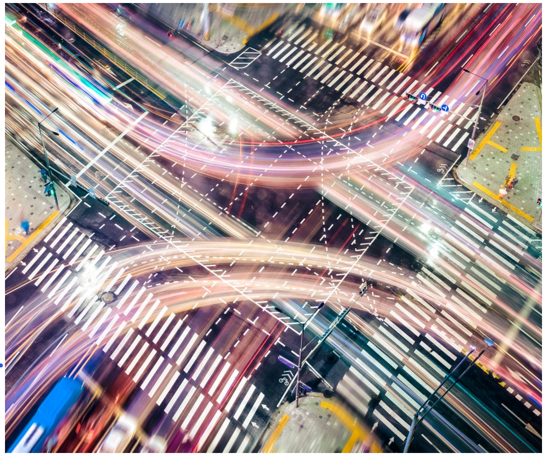
Distance-dependent forces

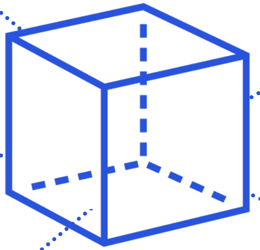
Orientations

Collisions

Spatial symmetries







One foundation model
for any 3D system?

Foundation model checklist	In **language / vision foundation models**
☐ Lots of **data**	the internet
☐ Universal **representations**	tokenizers and embeddings
☐ A scalable, espressive **architecture**	transformers
☐ Self-supervised **training** protocol	max likelihood, classification
☐ Multiple **downstream tasks** with shared structure	chatbots, ...
☐ Bonus points for **predictable scaling**	neural scaling laws

Foundation model checklist	Geometric foundation model
☐ Lots of **data**	the data is out there (?)
☐ Universal **representations**	?
☐ A scalable, espressive **architecture**	?
☐ Self-supervised **training** protocol	?
☐ Multiple **downstream tasks** with shared structure	?
☐ Bonus points for **predictable scaling**	?

GATr:
a versatile architecture for geometric data

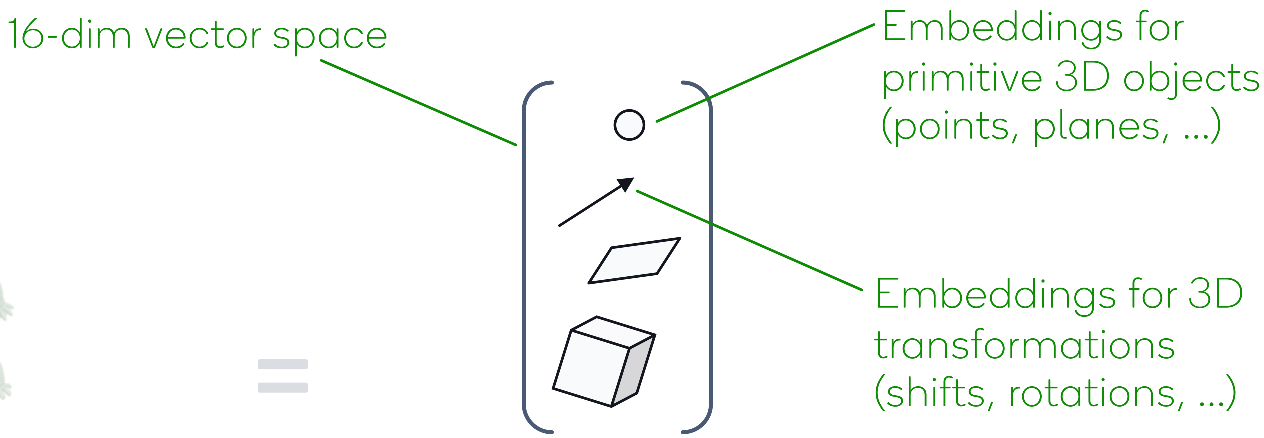
J. Brehmer et al, "Geometric Algebra Transformer", NeurIPS 2023

=

Geometric Algebra Transformer

Geometric Algebra Transformer

=



Geometric algebra representations

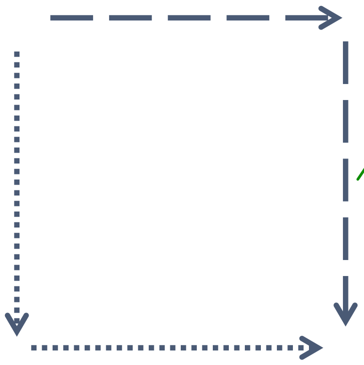
L. Dorst and S. De Deninck, "A Guided Tour to the Plane-Based Geometric Algebra PGA", bivector.net/PGA4CS.pdf

Geometric Algebra Transformer

=

Geometric algebra
representations

+

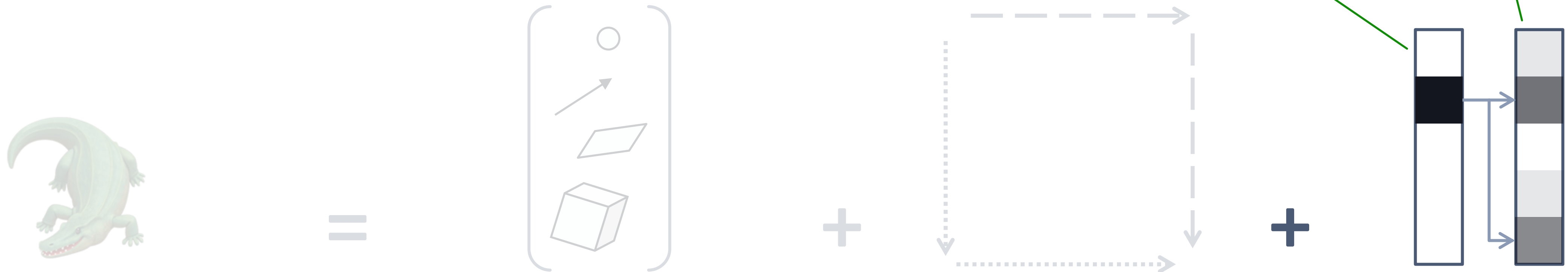


Equivariant
layers

Respect E(3)
symmetries of 3D space

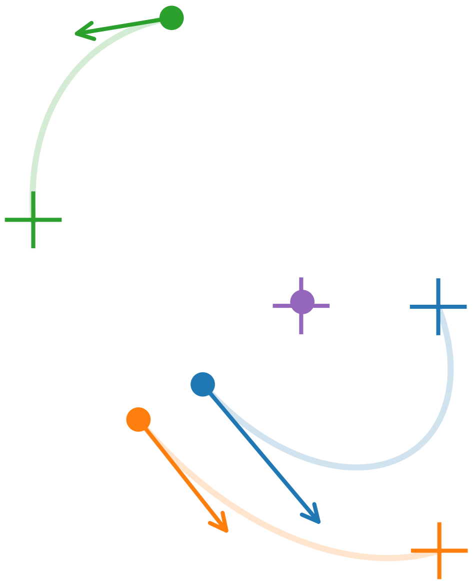
Expressive and versatile

Scalable to thousands of tokens (unlike many previous geometric architectures)

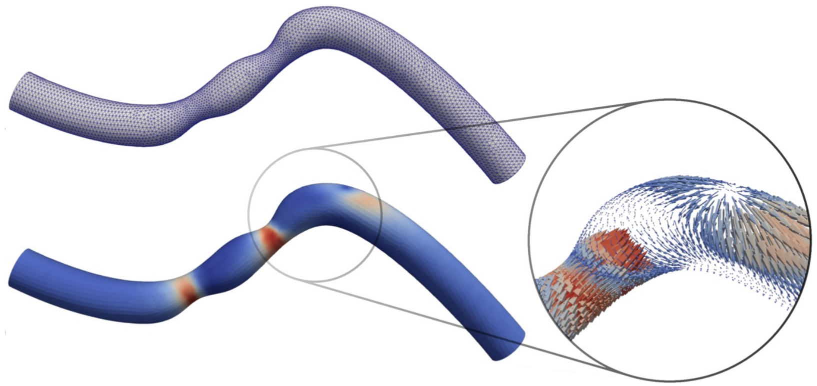


Geometric Algebra Transformer = Geometric algebra representations + Equivariant layers + **Transformer** architecture

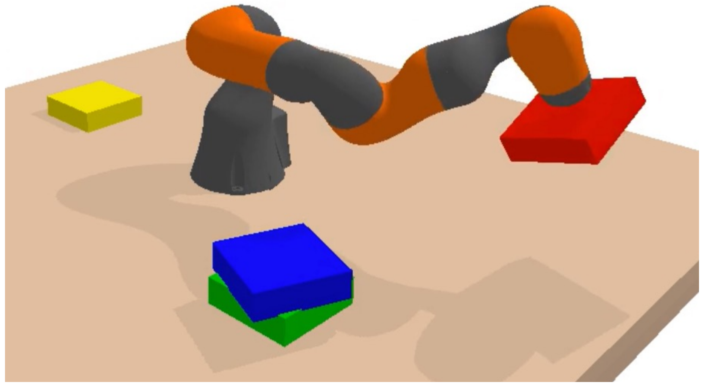
GATr outperforms baselines across domains



n-body modeling



predicting wall shear stress in arteries



robotic motion planning

Artery image from: J. Suk et al, "Mesh neural networks for SE(3)-equivariant hemodynamics estimation on the artery wall", arXiv:2212.05023

Towards geometric foundation models?

Foundation model checklist	Geometric foundation model
☐ Lots of **data**	the data is out there (?)
☐ Universal **representations**	?
☐ A scalable, espressive **architecture**	?
☐ Self-supervised **training** protocol	?
☐ Multiple **downstream tasks** with shared structure	?
☐ Bonus points for **predictable scaling**	?

Foundation model checklist	Geometric foundation model
☐ Lots of **data**	the data is out there (?)
☐ Universal **representations**	**could be geometric algebra–based**
☐ A scalable, espressive **architecture**	**equivariant transformers**
☐ Self-supervised **training** protocol	?
☐ Multiple **downstream tasks** with shared structure	?
☐ Bonus points for **predictable scaling**	?

Foundation model checklist	Geometric foundation model
☐ Lots of **data**	the data is out there (?)
☐ Universal **representations**	**could be geometric algebra–based**
☐ A scalable, espressive **architecture**	**equivariant transformers**
☐ Self-supervised **training** protocol	**?**
☐ Multiple **downstream tasks** with shared structure	**?**
☐ Bonus points for **predictable scaling**	**?**

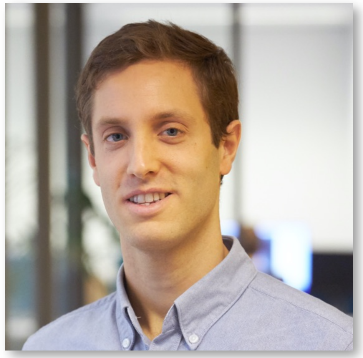
Scaling works.

Structure works.

Structure at scale is exciting, and there's lots we don't know yet.

Pim de Haan

Sönke Behrends



Taco Cohen

Ekdeep Singh Lubana

Geometric Algebra Transformer
Johann Brehmer*, Pim de Haan*, Sönke Behrends, Taco Cohen *equal contribution
NeurIPS 2023, arXiv:2305.18415

Surveying the Swamp: Choosing a Geometric Algebra for your Equivariant Transformer
Pim de Haan, Taco Cohen, Johann Brehmer
Under review

A Guided Tour to the Plane-Based Geometric Algebra PGA
Leo Dorst, Steven De Keninck
bivector.net/PGA4CS.pdf

Clifford group equivariant neural networks
David Ruhe, Johannes Brandstetter, Patrick Forré
NeurIPS 2023, arXiv:2305.11141

Language models in molecular discovery
Nikita Janakarajan, Tim Erdmann, Sarath Swaminathan, Teodoro Laino, Jannis Born
arXiv:2309.16235

Multiple Physics Pretraining for Physical Surrogate Models
Michael McCabe et al.
arXiv:2310.02994

xVal: A Continuous Number Encoding for Large Language Models
Siavash Golkar et al.
arXiv:2310.02989

Thanks to **Pim de Haan**, **Ekdeep Singh Lubana**, and **Taco Cohen** for their help preparing this presentation!

Thank you

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog