Qualcomm

ORIGINS Data Science Lab seminar

May 24, 2024

Scalable equivariance
with Geometric Algebra Transformers

Johann Brehmer
Qualcomm AI Research

Structure:
Problem-specific inductive biases in algorithms and architectures

Scale:
Flexible architectures, lots of data, lots of compute

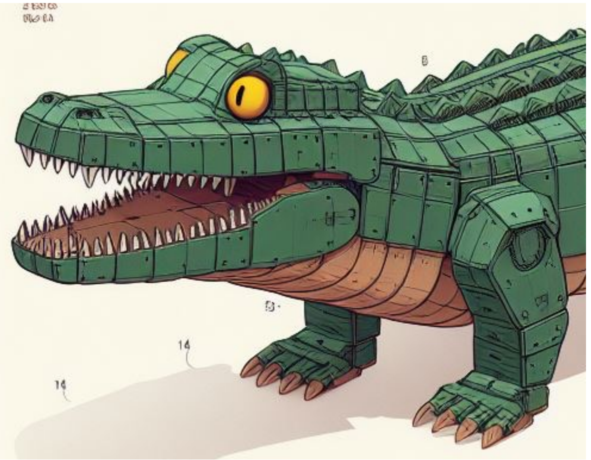
Structure:
Problem-specific inductive biases in algorithms and architectures

Scale:
Flexible architectures, lots of data, lots of compute

Structure:
Problem-specific inductive biases in algorithms and architectures

Geometric Algebra Transformer:
Our version of a versatile architecture for geometric problems

Scale:
Flexible architectures, lots of data, lots of compute



GATr 101
How to build an equivariant Transformer

Euclidean GATr
Applications far below the speed of light

Lorentz-GATr
for particle physics

GATr 101

Awareness of data types

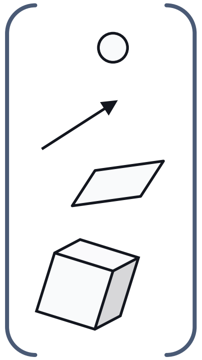
Equivariance to symmetry transformations

Geometric Algebra Transformer

Transformer architecture (scaled dot-product attention)

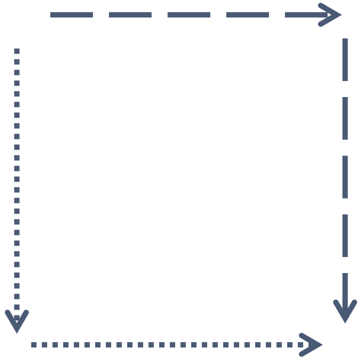
Geometric Algebra
Transformer

=



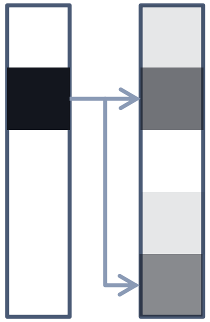
Geometric algebra
representations

+



Equivariant
layers

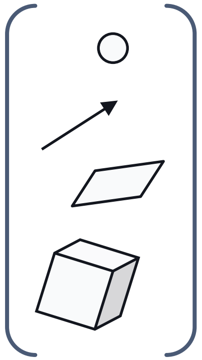
+



Transformer
architecture

Geometric Algebra Transformer

=



Geometric algebra
representations

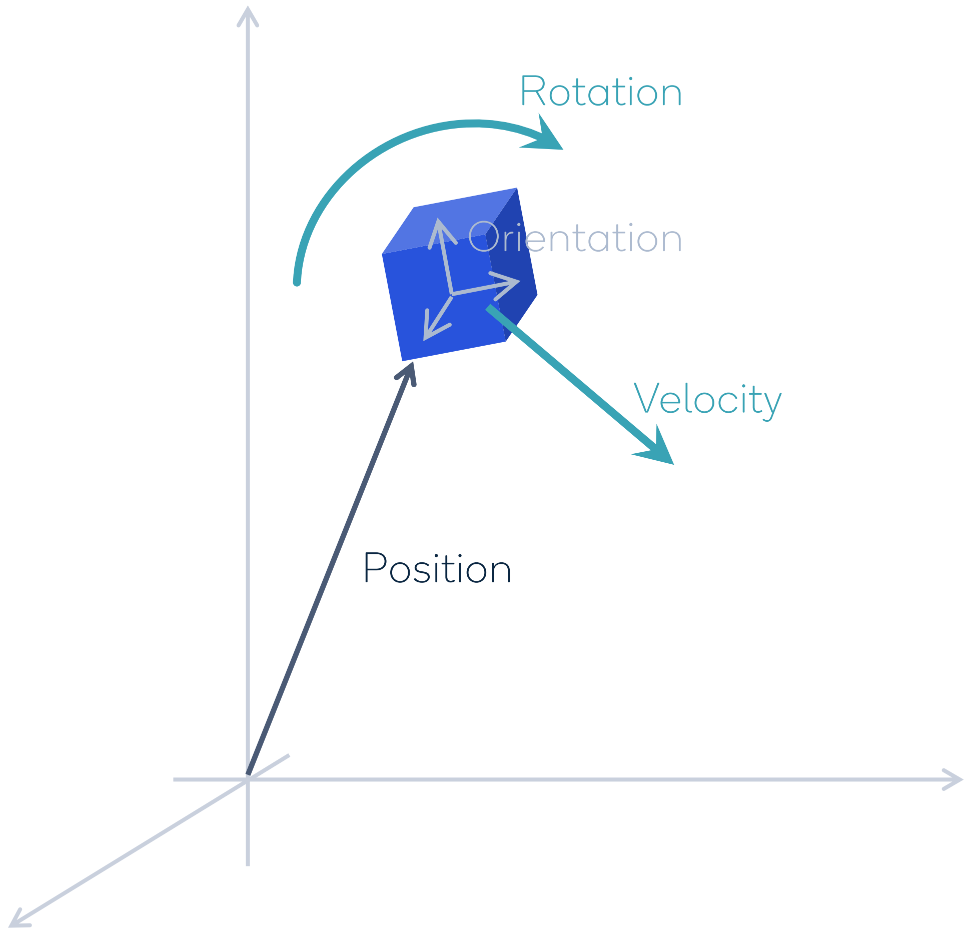
+

Equivariant
layers

+

Transformer
architecture

Representing geometric data



How do you parameterize such a 3D object?

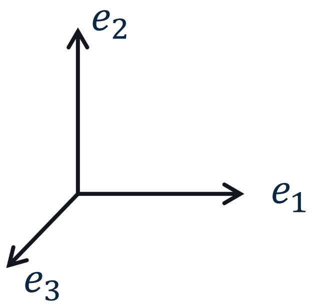
- **Most deep learning:** 14 numbers
- Previous **geometric deep learning:** e.g. 2 vectors, 2 rotation matrices
- **GATr:** 1 position, 3 directions, 1 translation, 1 rotation

Why use different types?

- Types have different **common patterns** (compute distances between positions, but not between direction vectors)
- Types differ in behaviour under **transformations**
- Types provide an inductive bias, potentially improving **sample efficiency** and **generalization**

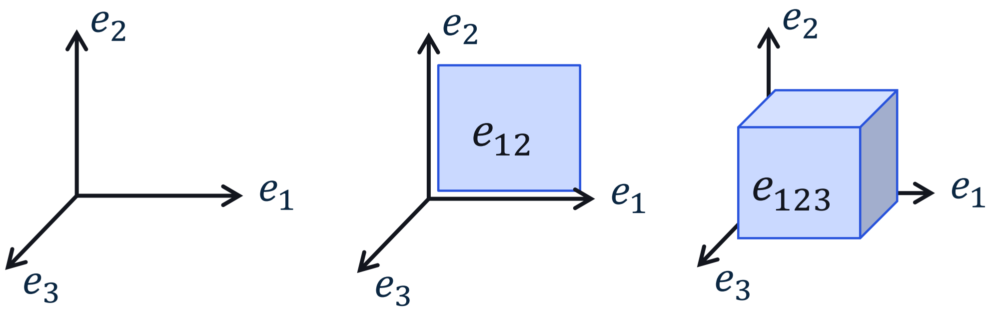
Geometric algebra

- Vector space V with d dimensions and inner products
  - Basis $e_1, e_2, \dots, e_d$



- Geometric algebra $\mathcal{G}(V)$ has 2^d dimensions, basis

$$\underset{\text{scalar}}{1,} \quad \underbrace{e_1, e_2, \dots, e_d,}_{\text{vector}} \quad \underbrace{e_{12}, e_{13}, \dots,}_{\text{bivector}} \quad \underbrace{e_{123}, \dots, e_{12\dots d}}_{\text{trivector...}}$$



- Geometric product $\mathcal{G}(V) \times \mathcal{G}(V) \to \mathcal{G}(V)$
  - Generalizes dot product and cross product

Graßmann

Clifford

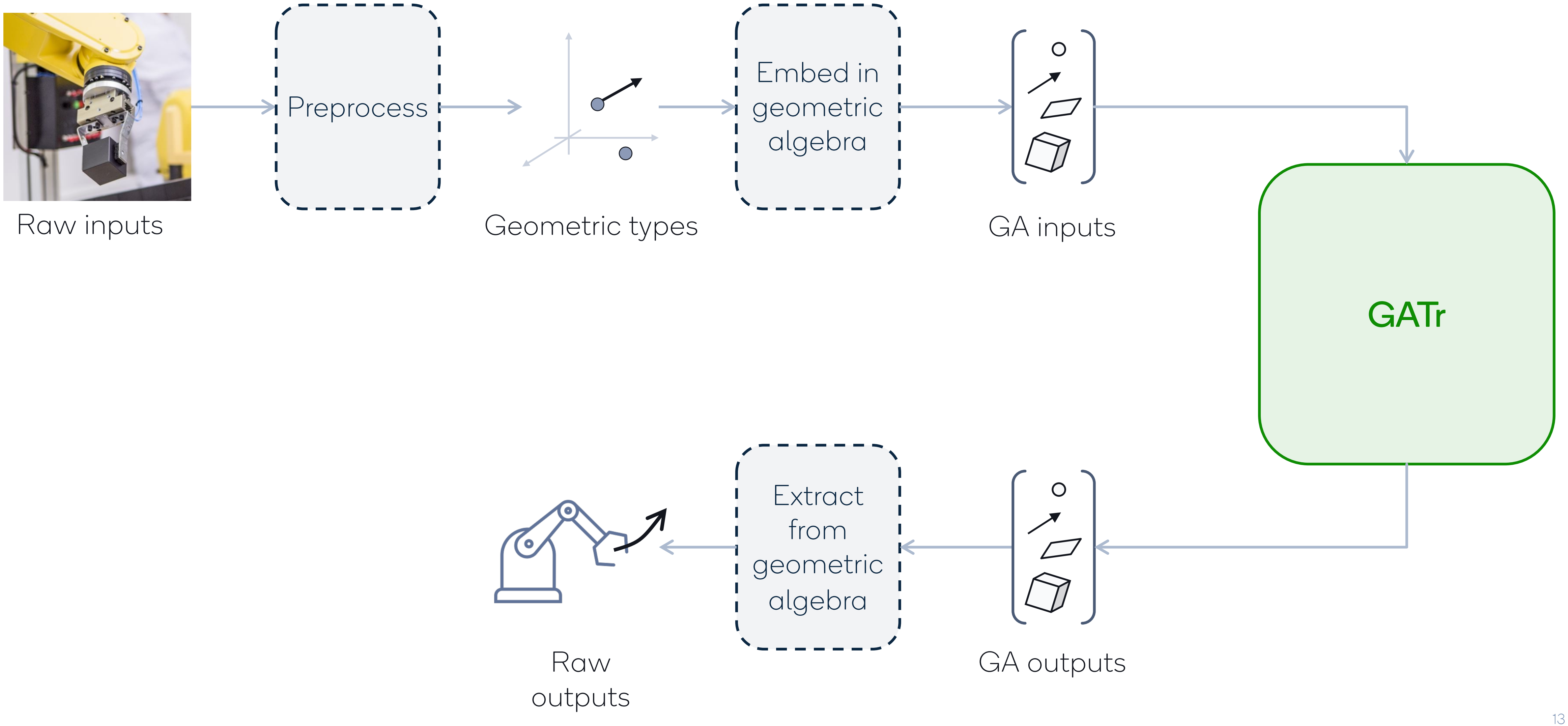
Projective GA as **representation for ML**

- We use GA representations in addition to the usual unstructured vector space
- Offers **$16n$-dimensional representation** of 3D geometric data
  - "**Typing**": a point is not a direction of movement is not the orientation of a plane
  - Established embeddings for **3D primitives and transformations**

Object / operator	Scalar	Vector		Bivector		Trivector		PS
	1	e_0	e_i	e_{0i}	e_{ij}	e_{0ij}	e_{123}	e_{0123}
Scalar $\lambda \in \mathbb{R}$	λ	0	0	0	0	0	0	0
Plane w/ normal $n \in \mathbb{R}^3$, origin shift $d \in \mathbb{R}$	0	d	n	0	0	0	0	0
Line w/ direction $n \in \mathbb{R}^3$, orthogonal shift $s \in \mathbb{R}^3$	0	0	0	s	n	0	0	0
Point $p \in \mathbb{R}^3$	0	0	0	0	0	p	1	0
Pseudoscalar $\mu \in \mathbb{R}$	0	0	0	0	0	0	0	μ
Reflection through plane w/ normal $n \in \mathbb{R}^3$, origin shift $d \in \mathbb{R}$	0	d	n	0	0	0	0	0
Translation $t \in \mathbb{R}^3$	1	0	0	$\frac{1}{2}t$	0	0	0	0
Rotation expressed as quaternion $q \in \mathbb{R}^4$	q_0	0	0	0	q_i	0	0	0
Point reflection through $p \in \mathbb{R}^3$	0	0	0	0	0	p	1	0

  - Geometric product: **canonical operation** on these representations

GA representations **in practice**

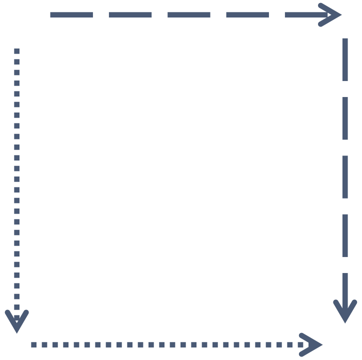


Geometric Algebra
Transformer

=

Geometric algebra
representations

+

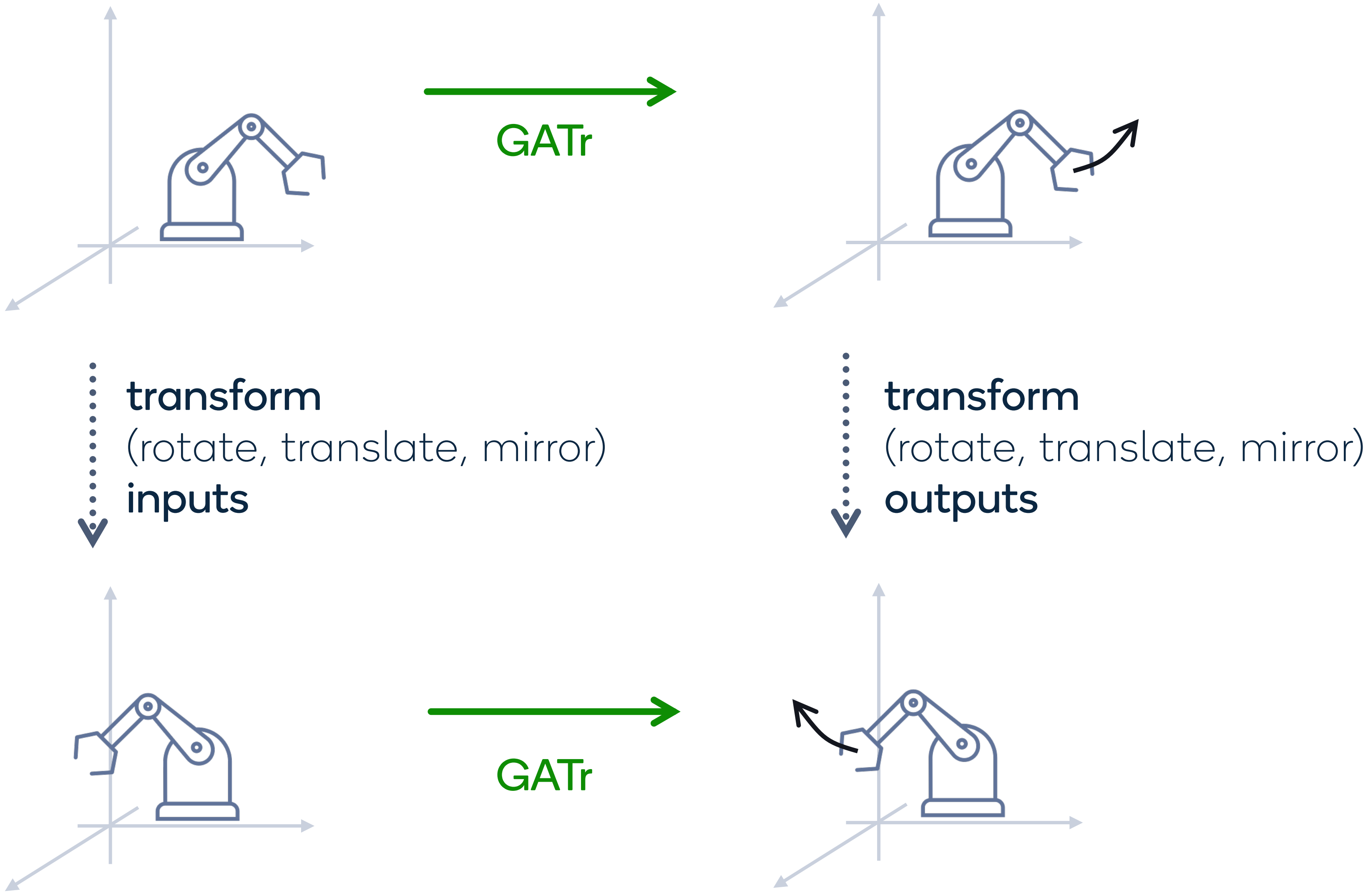


Equivariant
layers

+

Transformer
architecture

E(3) equivariance



GATr: built out of new **E(3)-equivariant layers** between GAs

- We theoretically characterize how equivariance constrains **linear layers**:

essentially E(d)

Proposition 1. *Any linear map* $\phi : \mathbb{G}_{d,0,1} \to \mathbb{G}_{d,0,1}$ *that is equivariant to* $\mathrm{Pin}(d, 0, 1)$ *is of the form*

$$\phi(x) = \sum_{k=0}^{d+1} w_k \langle x \rangle_k + \sum_{k=0}^{d} v_k e_0 \langle x \rangle_k \tag{4}$$

for parameters $w \in \mathbb{R}^{d+2}, v \in \mathbb{R}^{d+1}$*. Here* $\langle x \rangle_k$ *is the blade projection of a multivector, which sets all non-grade-k elements to zero.*

- Plus: equivariant attention, nonlinearities, normalization...

Geometric Algebra
Transformer

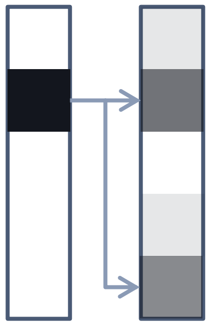
=

Geometric algebra
representations

+

Equivariant
layers

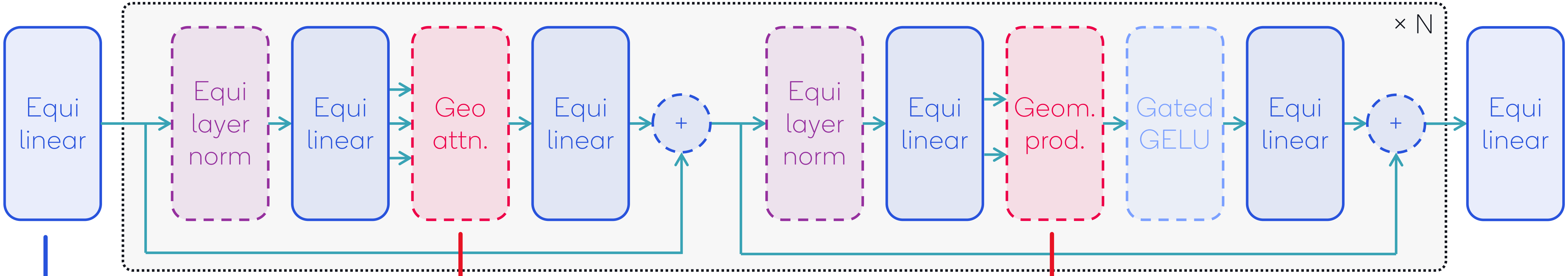
+



Transformer
architecture

Input and output data
can have one or multiple token dimensions

Attention blocks
can be stacked to large depth, gradients are propagated efficiently



Linear layers
between GA representations with equivariance constraint

Geometric attention
generalizes scaled dot-product attention

Geometric product
allow for construction of new geometric types

GATr is powered by **dot-product attention**

Message-passing neural networks

$$m_{i\to j} = \Phi(x_i, x_j, e_{ij})$$

$$x_j' = \Psi\left(x_j, \sum_i m_{i\to j}\right)$$

- Node and edge representations
- (Equivariant) network on each edge

GATr (and other transformers)

$$V_i, K_i, Q_i = \Phi(x_i)$$

$$x' = \sigma\left({QK^T}/{\sqrt{d}}\right)V$$

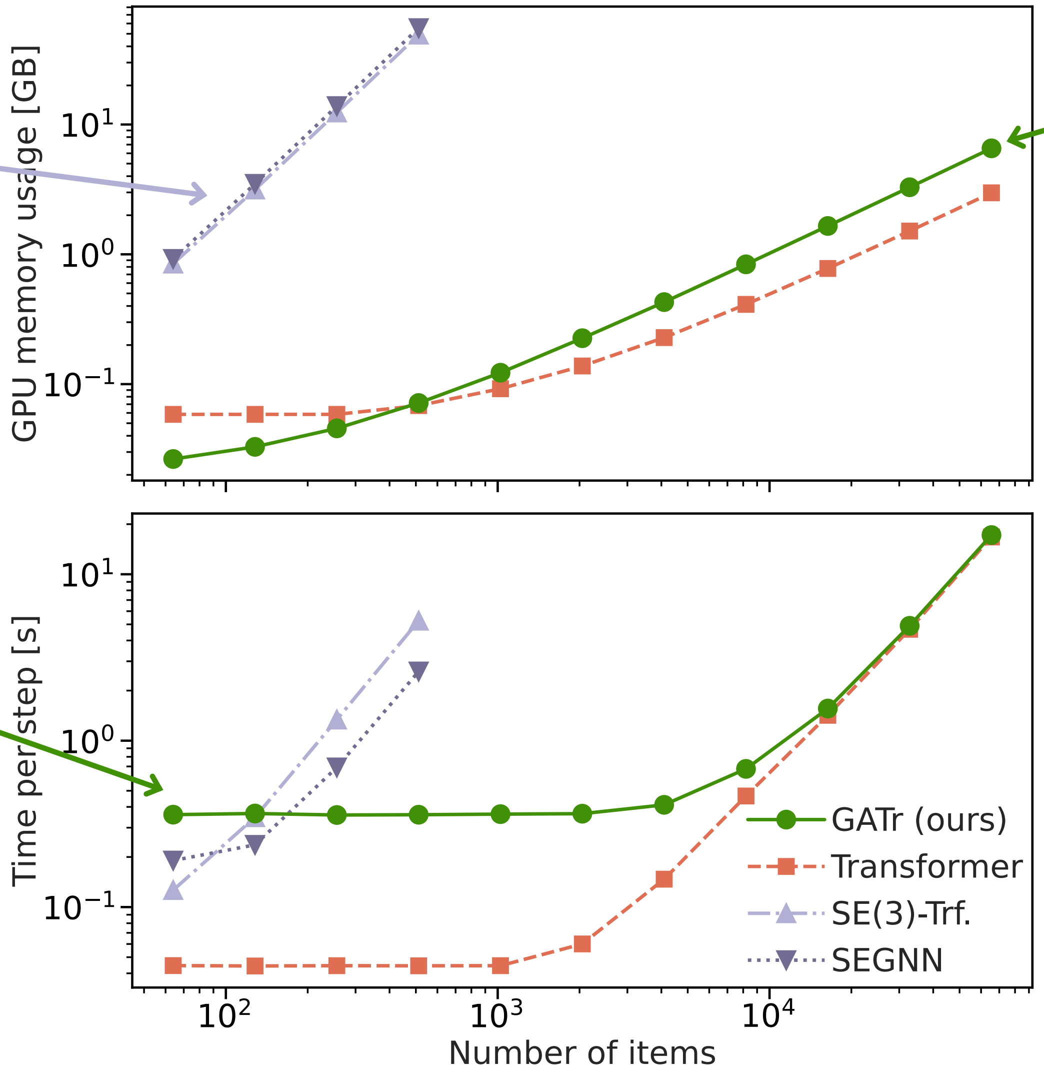
- Only node representations
- Dot-product per edge,
  with highly optimized implementations
  (e.g. flash attention)

Transformers have same theoretical complexity, but dramatically more efficient in practice

GATr is more scalable than GDL baselines

Heavily optimized Nvidia implementation of a classic GDL baseline

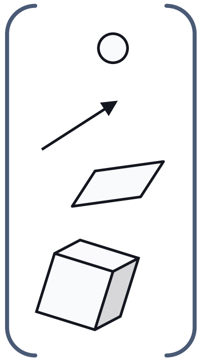
Overhead in small problems, but still room for optimization



GATr (with flash attention) **scales like a transformer, to 10ks tokens!**

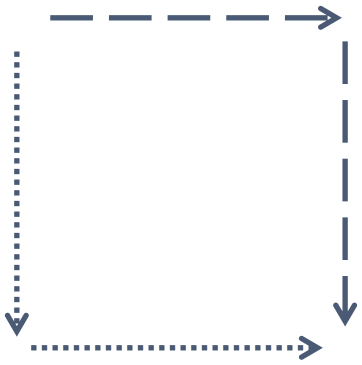
Geometric Algebra
Transformer

=



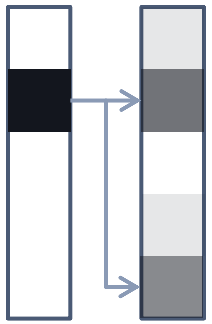
Geometric algebra
representations

+



Equivariant
layers

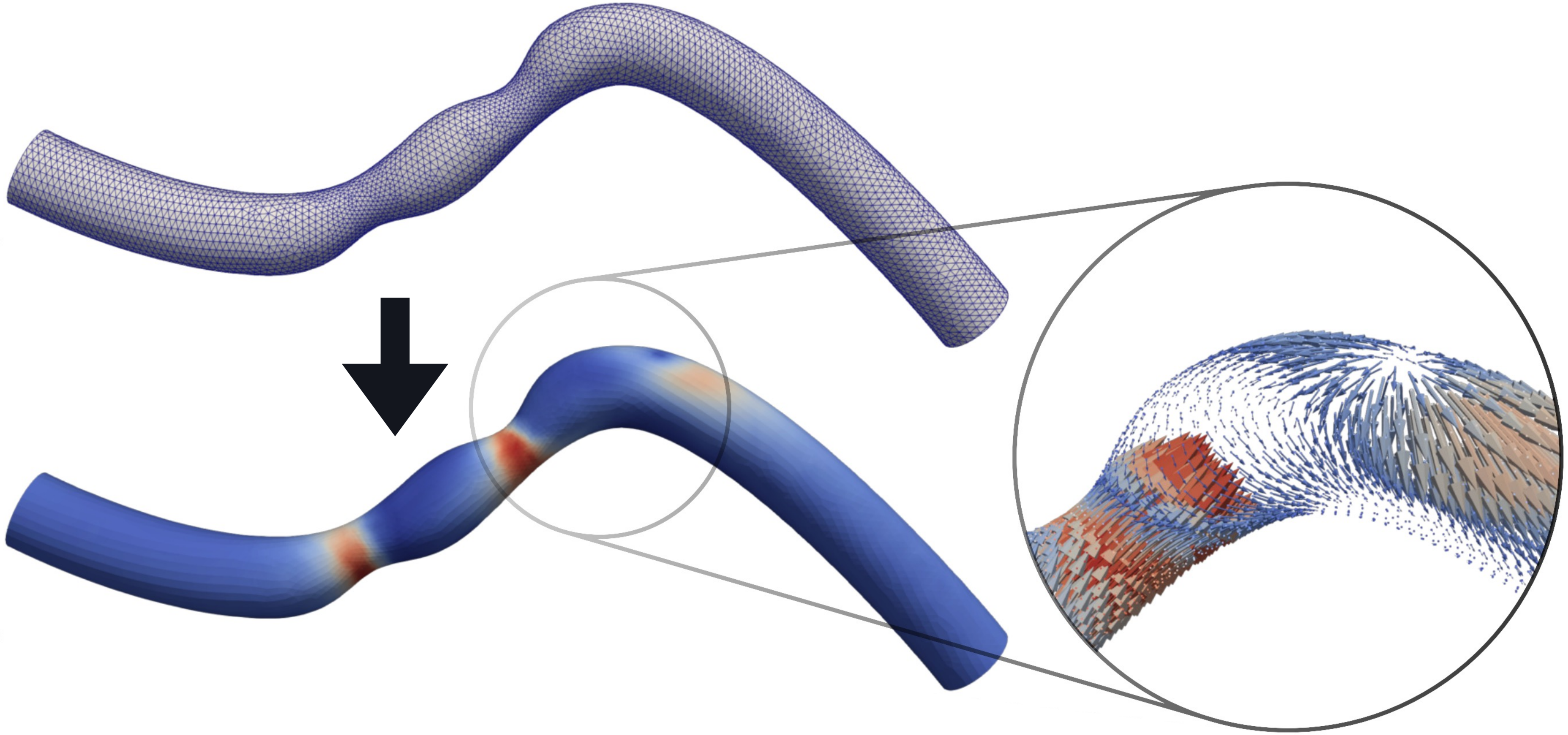
+



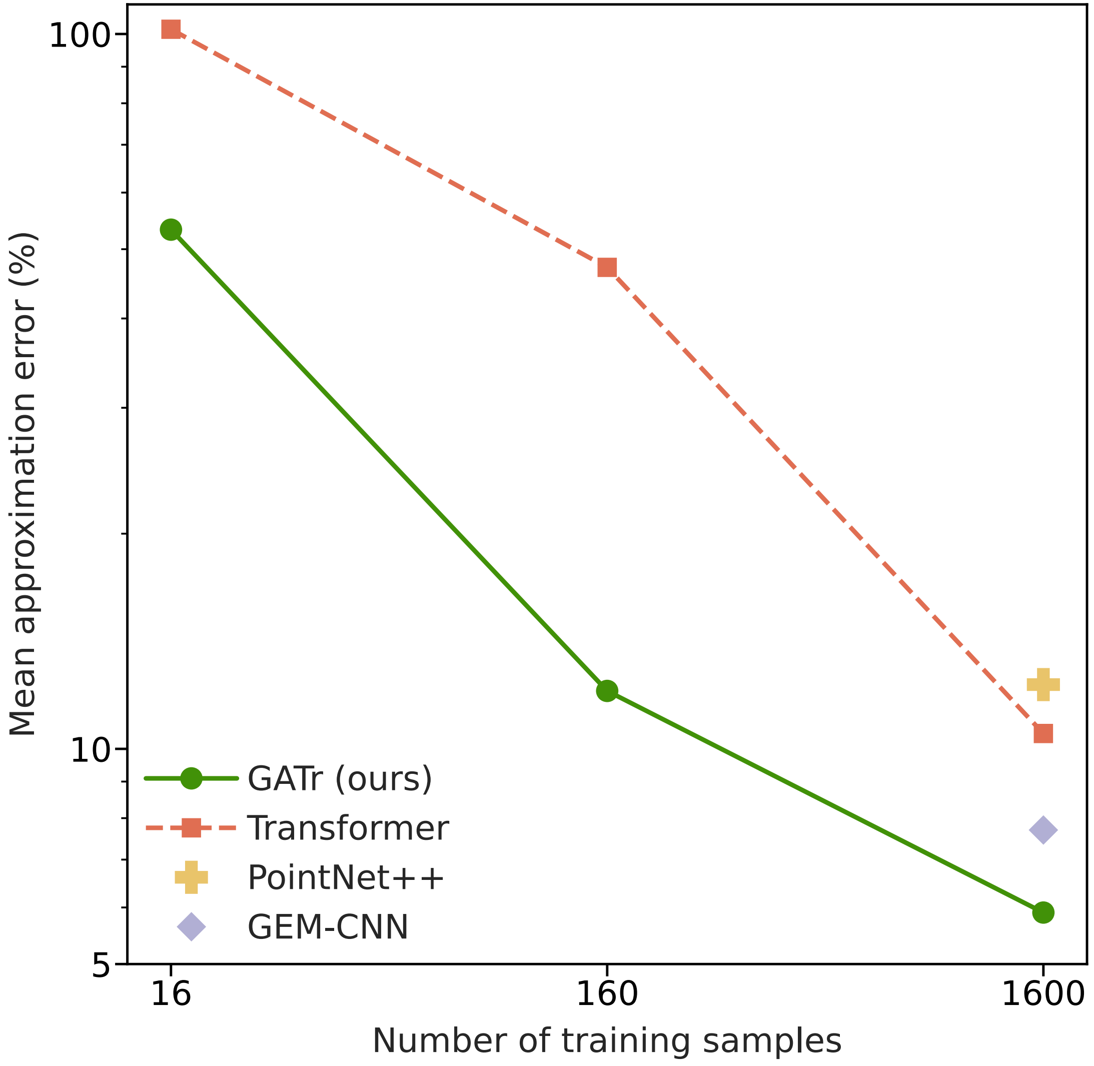
Transformer
architecture

Euclidean GATr

Arterial wall-shear stress estimation with 7k tokens

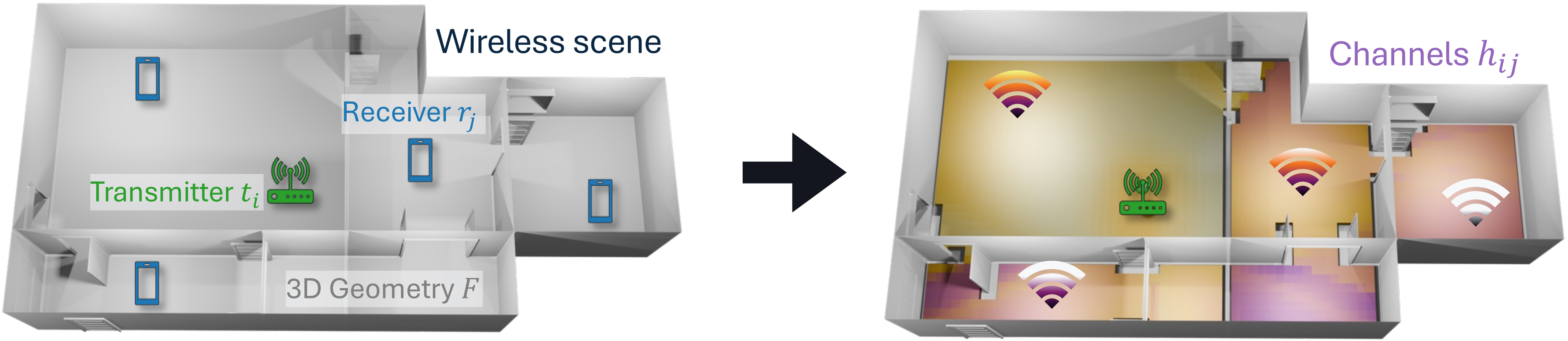


Arterial wall-shear stress estimation with 7k tokens

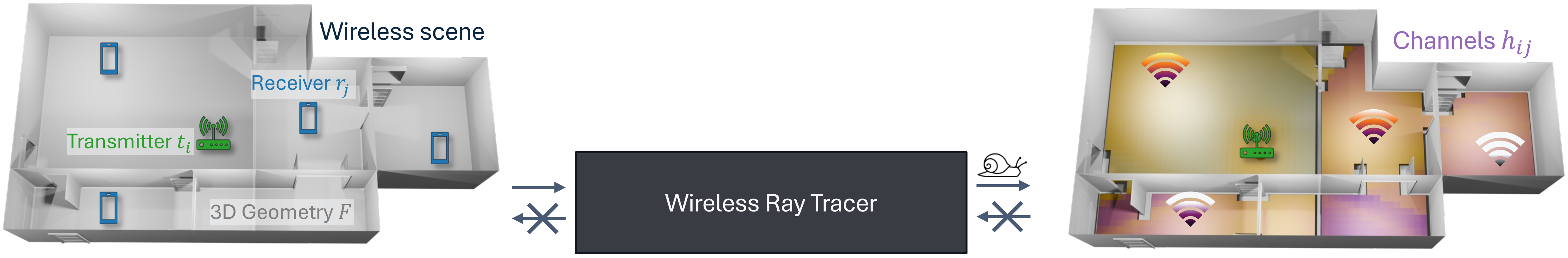


GATr works well on low-data, high-complexity problems

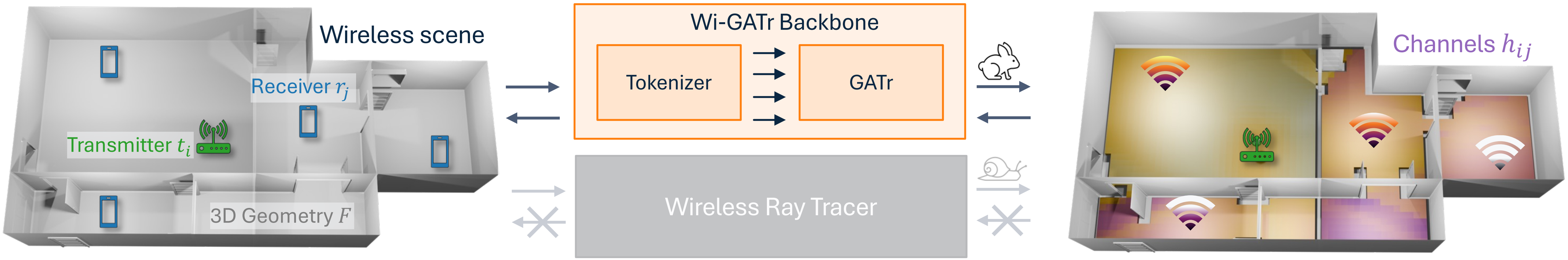
Differentiable **Wireless signal modelling**



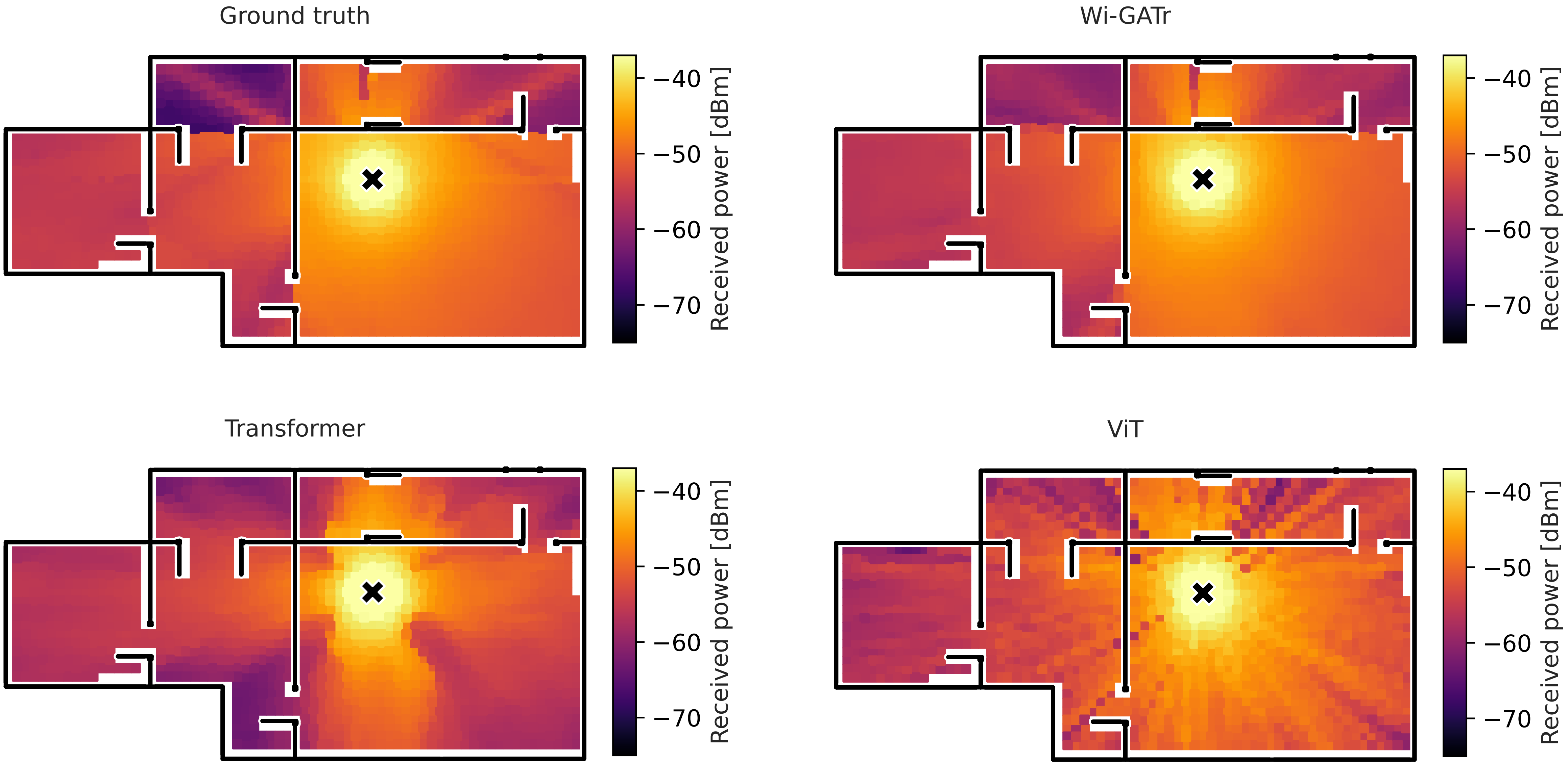
Differentiable **Wireless signal modelling**



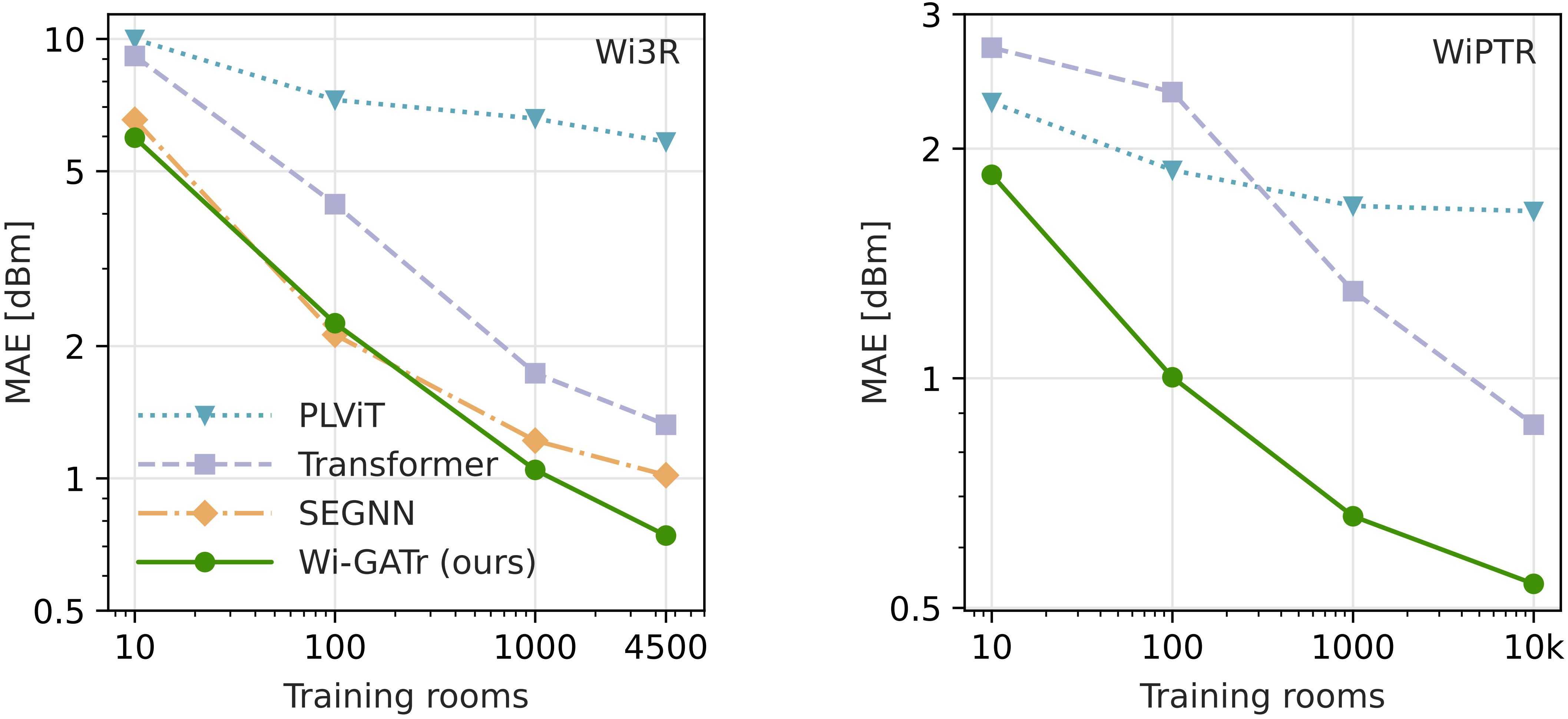
Differentiable **Wireless signal modelling**



Differentiable **Wireless signal modelling** from only 100 training samples

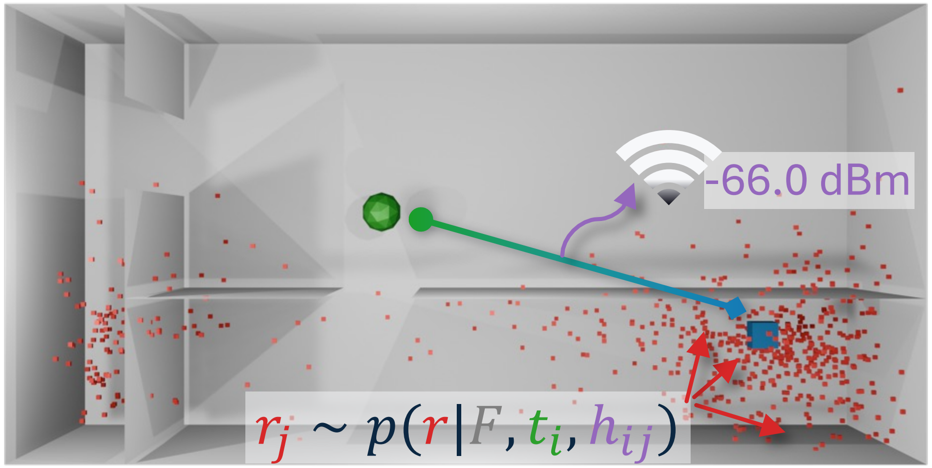


Differentiable **Wireless signal modelling**

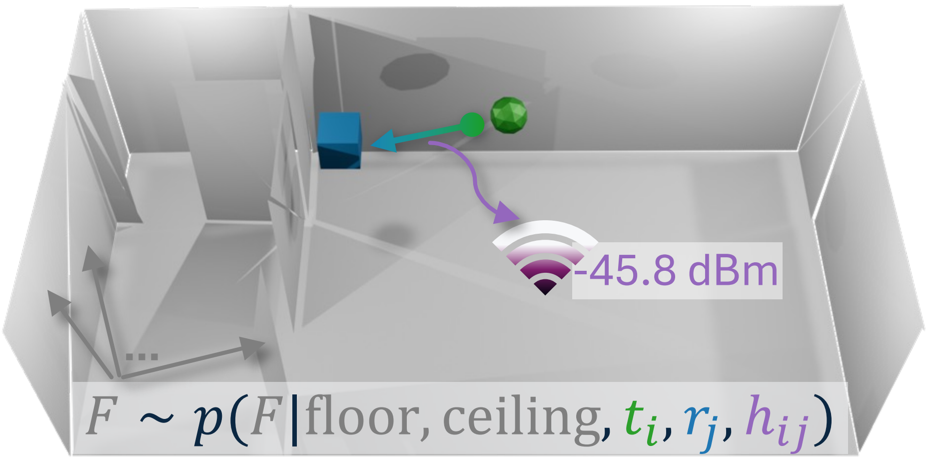


Next level: **Probabilistic wireless modelling** with diffusion models

GATr-based diffusion model lets us solve various inference tasks through conditional sampling from a single joint density (a la "inpainting")

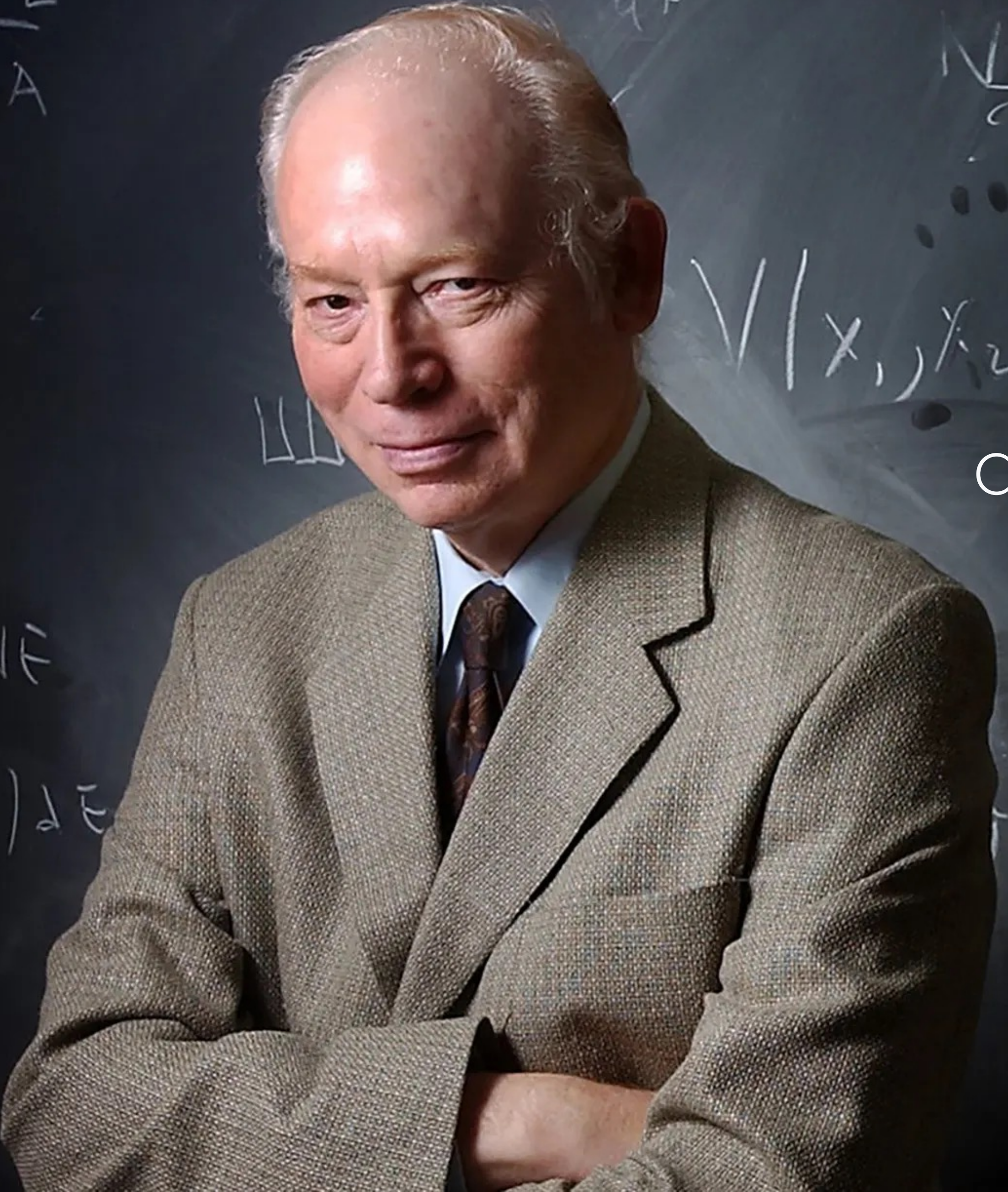


Receiver localization from environment and wireless signal



Geometry reconstruction from wireless signal

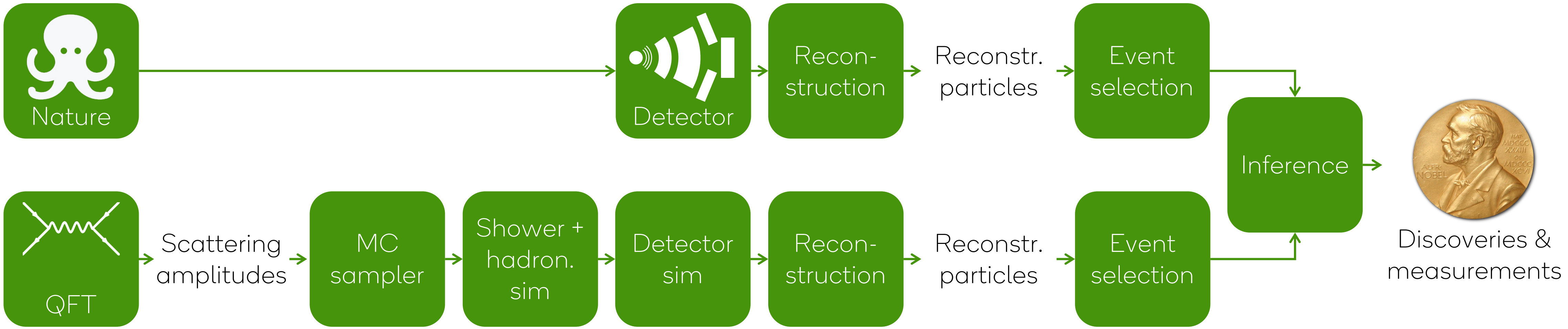
Lorentz-GATr



"The universe is
an enormous direct product
of representations
of symmetry groups"

Steven Weinberg

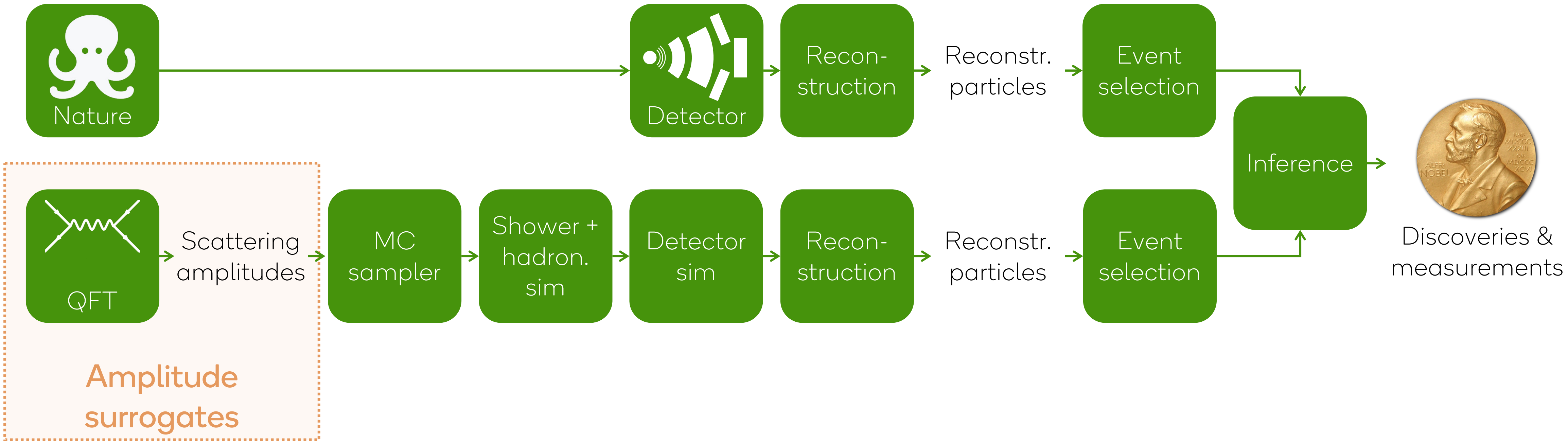
hep data analysis pipeline



- Most steps have some degree of **Lorentz symmetry**
  - Detector / measurement process can partially break it or make it approximate
- Yet, equivariant architectures for high-energy physics seem understudied
  - Notable exceptions: LorentzNet, PELICAN, …

LorentzNet: S. Gong et al, "An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging", arXiv:2201.08187
PELICAN: A. Bogatskiy et al, "Explainable Equivariant Neural Networks for Particle Physics: PELICAN", arXiv:2307.16506

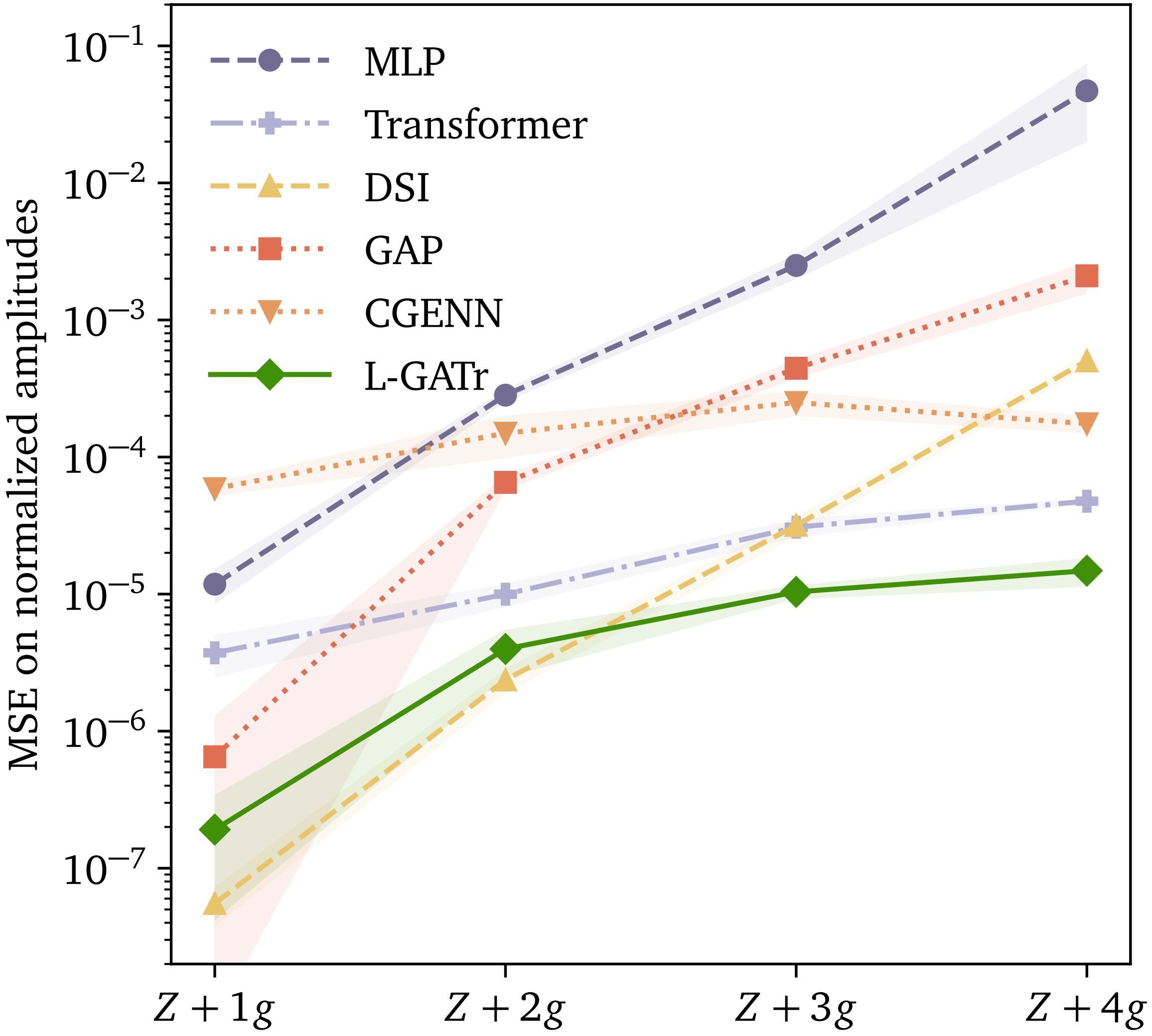
	life at low speed	high-energy physics
symmetry group	**E(3)**	Poincaré but we only care about **Lorentz, $O^+(1,3)$**
data	scalars points directions …	scalars (e.g. particle ID) four-vectors (e.g. four-momenta) …
GATr rep	**projective GA** $\mathcal{G}_{3,0,1}$ (16D)	**space-time GA** $\mathcal{G}_{1,3,0}$ (also 16D, but different metric)
GATr layers	equi linear attention normalization …	equi linear (derived for different symmetry) attention (based on new metric) normalization (robust to neg. norms) …
GATr architecture	transformer	transformer


Nature
Detector
Recon-
struction
Reconstr.
particles
Event
selection
Inference
Discoveries &
measurements
QFT
Scattering
amplitudes
MC
sampler
Shower +
hadron.
sim
Detector
sim
Recon-
struction
Reconstr.
particles
Event
selection
Amplitude
surrogates

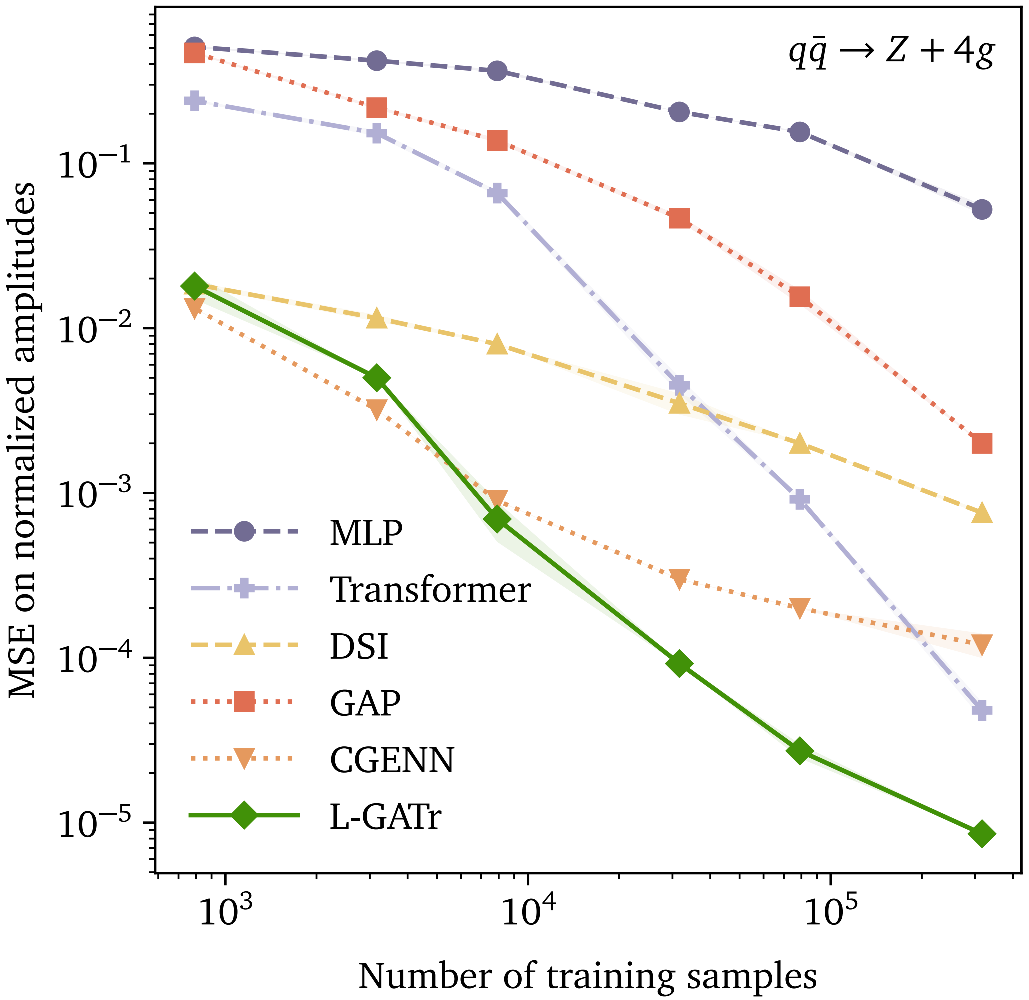
Amplitude regression

Surrogate models for QFT amplitudes are most useful (and most challenging) at high multiplicity

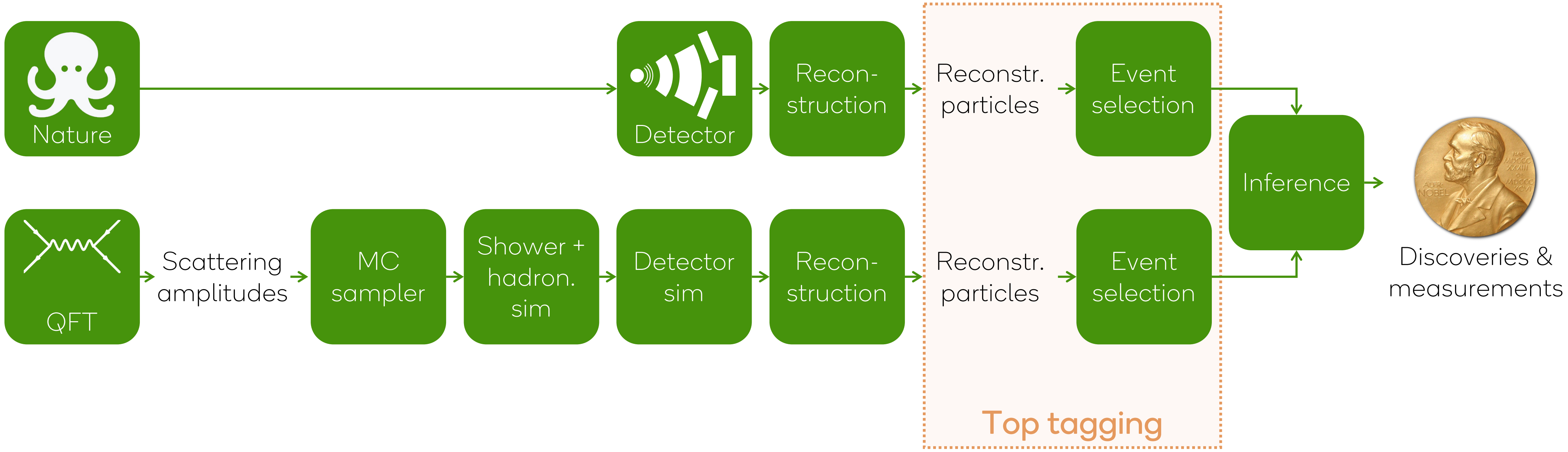
That's where **L-GATr** shines!



Amplitude regression



L-GATr is both a **small-data** and a **big-data** architecture



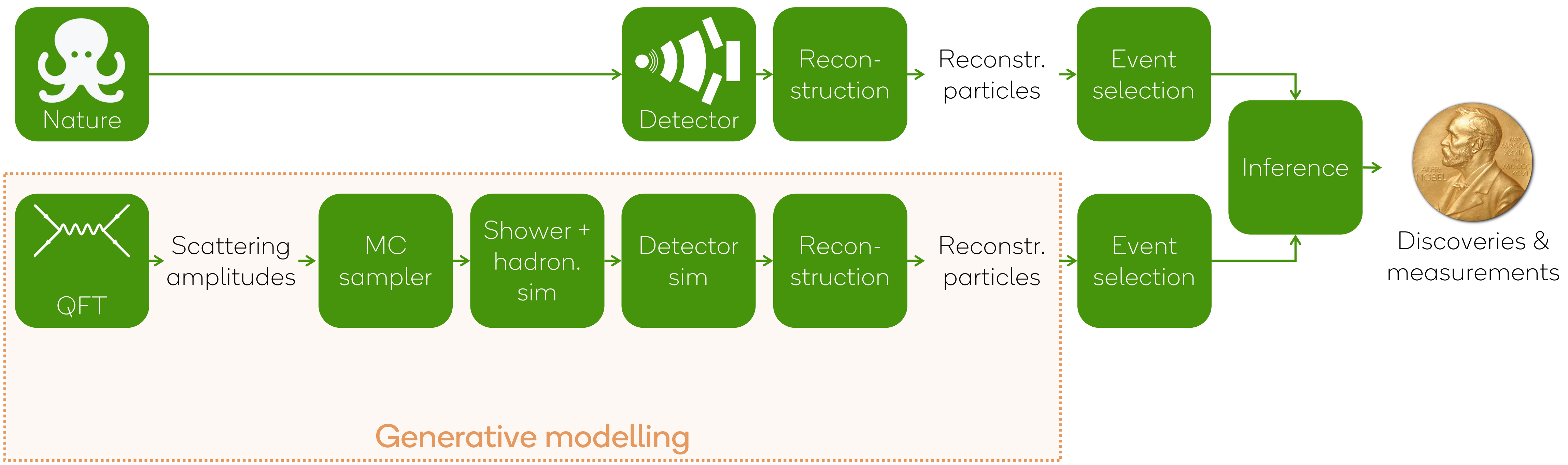
Nature
Detector
Recon-
struction
Reconstr.
particles
Event
selection
Inference
Discoveries &
measurements
QFT
Scattering
amplitudes
MC
sampler
Shower +
hadron.
sim
Detector
sim
Recon-
struction
Reconstr.
particles
Event
selection
Top tagging

Top tagging on everyone's favorite dataset

Model	Accuracy	AUC	$1/\epsilon_B$ ($\epsilon_S = 0.5$)	$1/\epsilon_B$ ($\epsilon_S = 0.3$)
TopoDNN [48]	0.916	0.972	–	295 ± 5
LoLa [15]	0.929	0.980	–	722 ± 17
P-CNN [1]	0.930	0.9803	201 ± 4	759 ± 24
N-subjettiness [60]	0.929	0.981	–	867 ± 15
PFN [50]	0.932	0.9819	247 ± 3	888 ± 17
TreeNiN [56]	0.933	0.982	–	1025 ± 11
ParticleNet [62]	0.940	0.9858	397 ± 7	1615 ± 93
ParT [63]	0.940	0.9858	413 ± 16	1602 ± 81

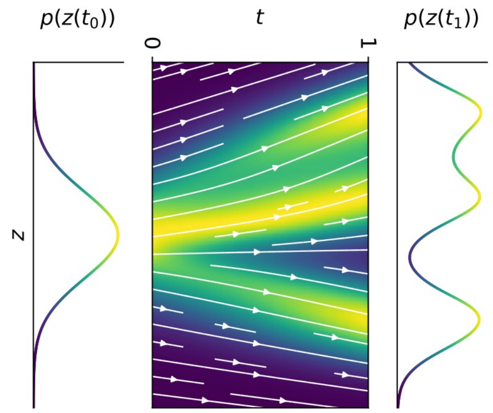
L-GATr is on par (but not better than) the best equivariant (*) baselines

G. Kasieczka et al, "The machine learning landscape of top taggers", SciPost Physics 2019


Nature
Detector
Recon-
struction
Reconstr.
particles
Event
selection
Inference
Discoveries &
measurements
QFT
Scattering
amplitudes
MC
sampler
Shower +
hadron.
sim
Detector
sim
Recon-
struction
Reconstr.
particles
Event
selection
Generative modelling

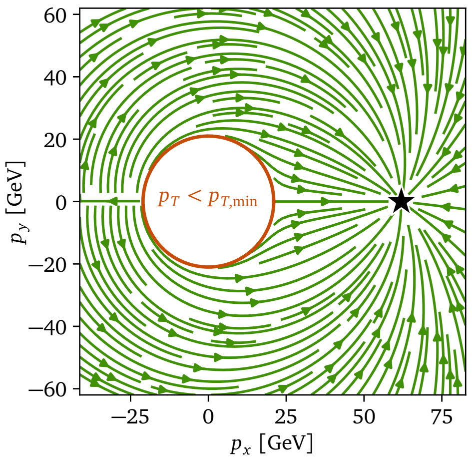
Generative modelling with conditional flow matching

- **Continuous normalizing flows:**
  simple base density
  + continuous dynamics described by differential equation
  = generative model



- **Conditional flow matching:**
  a brilliantly simple way to train them

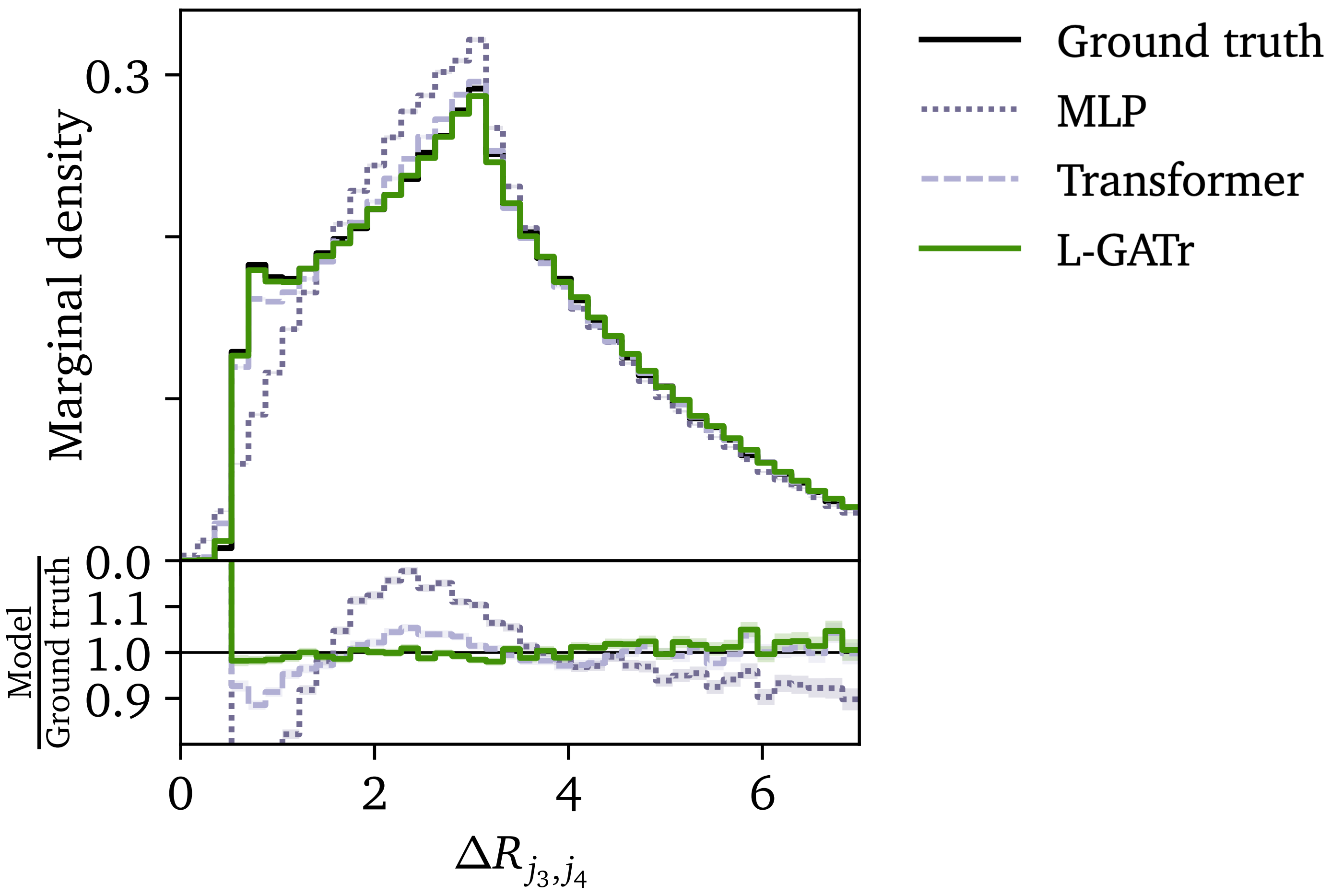
- **Riemannian flow matching:**
  manifold-based approach that lets us incorporate boundaries



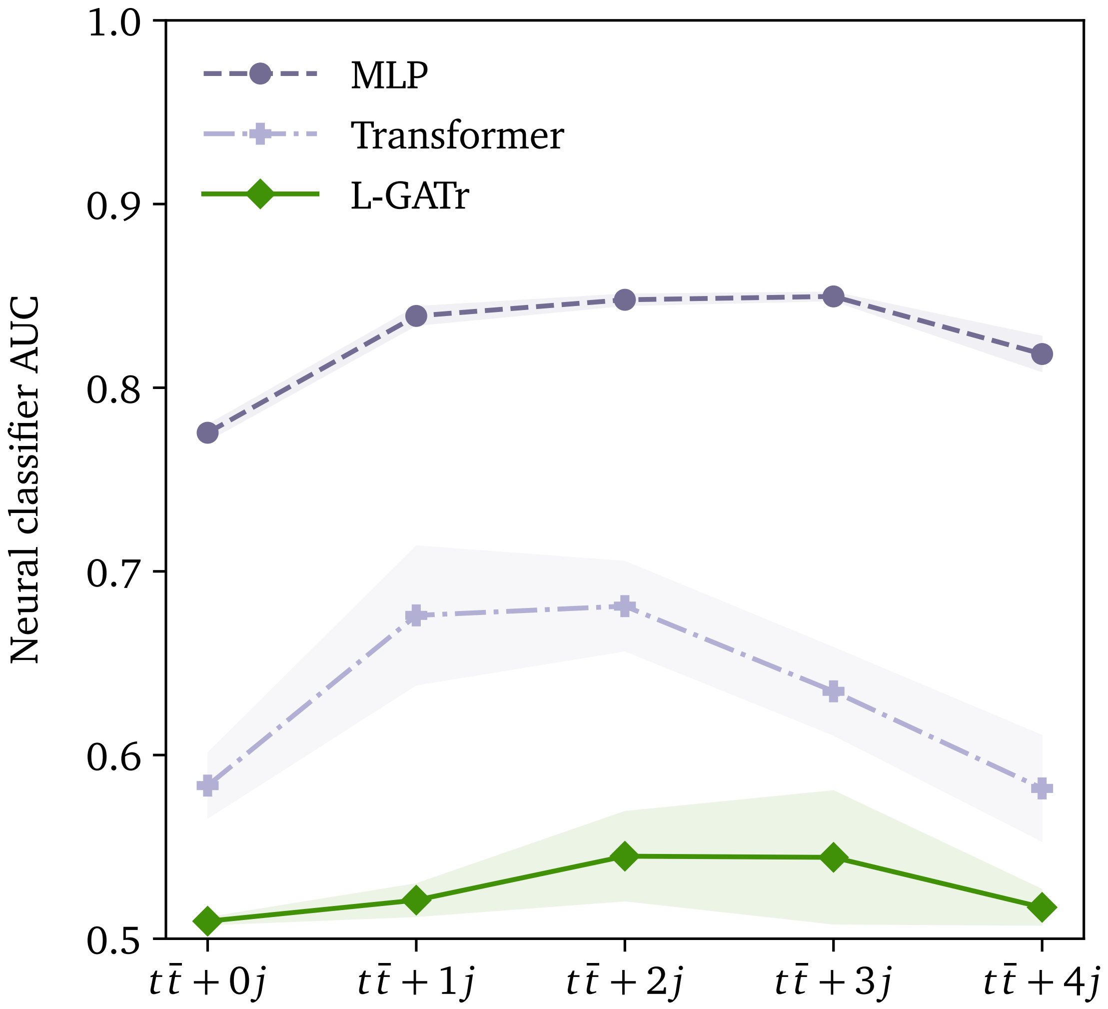
R. Chen et al, "Neural Ordinary Differential Equations", NeurIPS 2018
W. Grathwohl et al, "FFJORD: Free-form continuous dynamics for scalable reversible generative models", ICLR 2019
Y. Lipman et al, "Flow matching for generative modelling", ICLR 2023
R. Chen et al, "Flow matching on general geometries", ICLR 2024

Generative modelling with conditional flow matching

L-GATr leads to improvements
on tricky kinematic features...



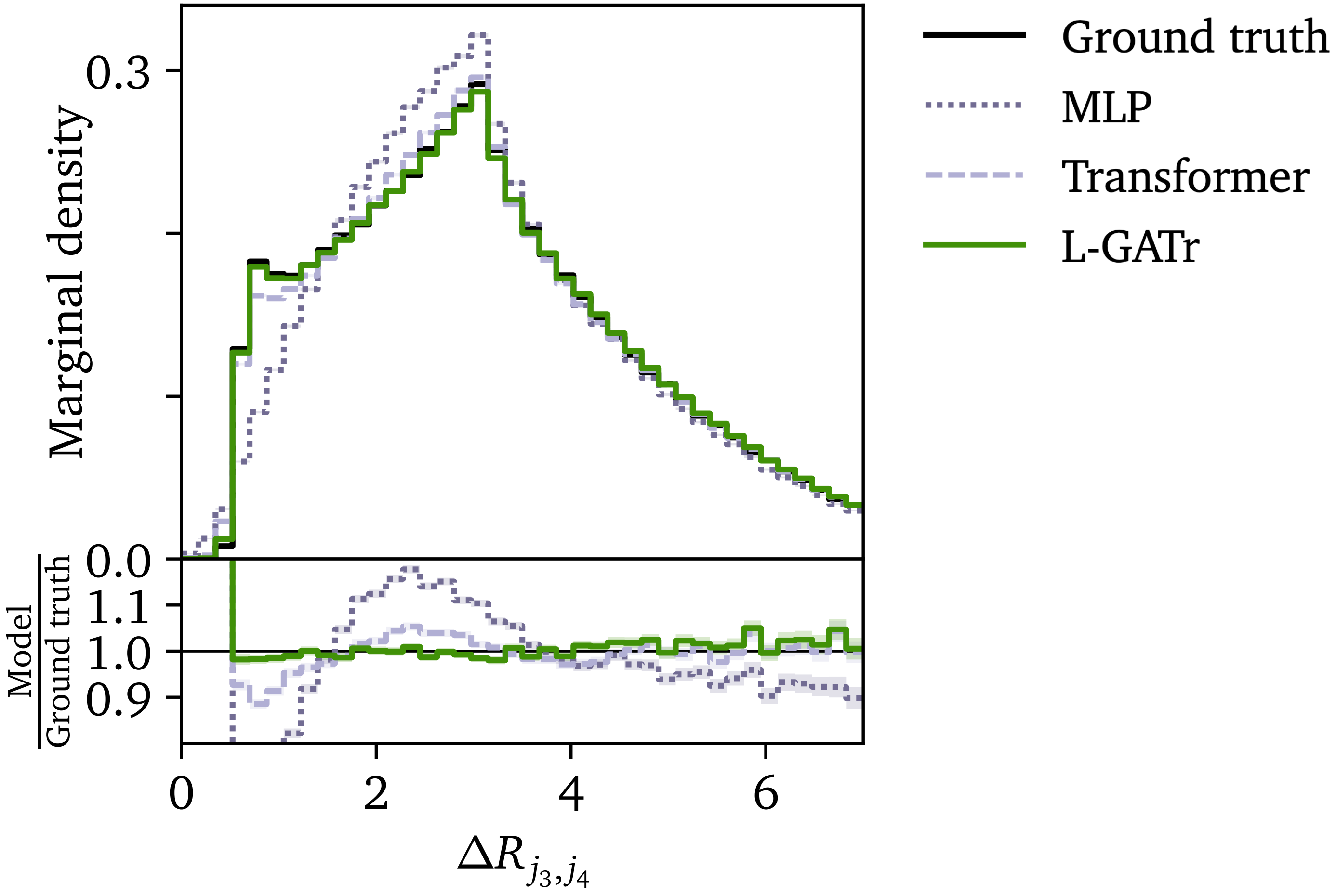
Generative modelling with conditional flow matching

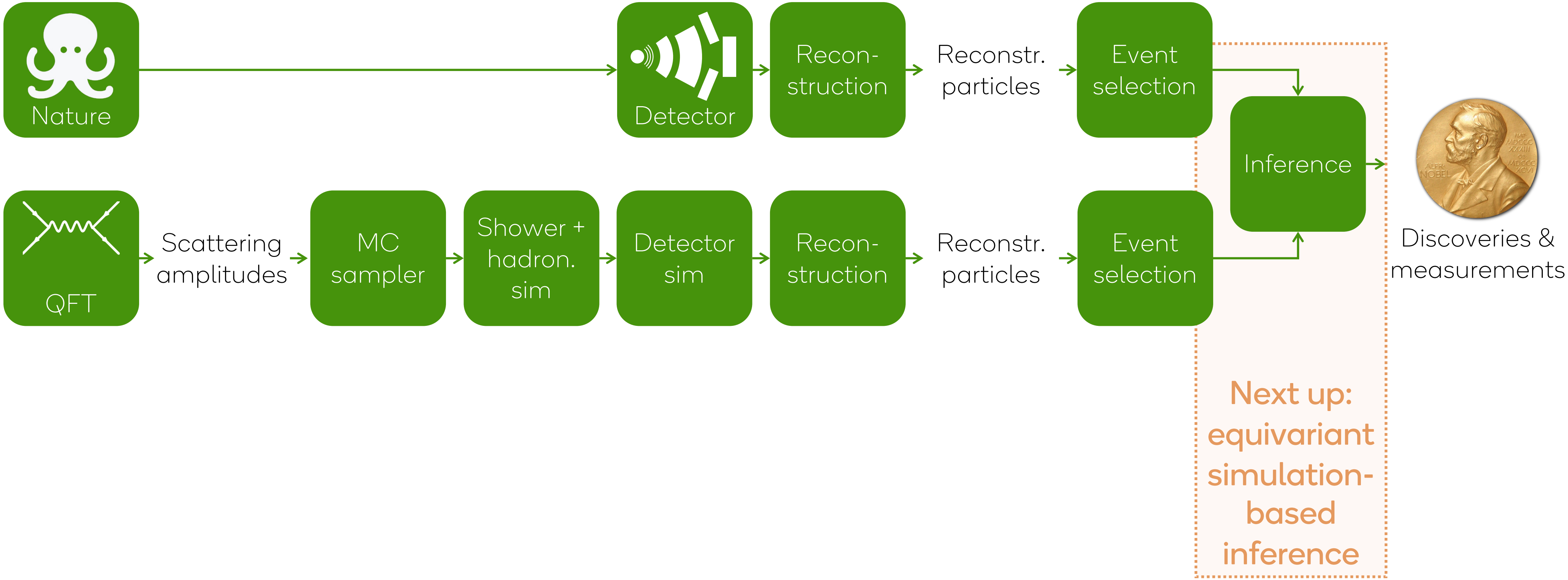


L-GATr generates samples that a classifier almost cannot distinguish from the ground truth…

Generative modelling with conditional flow matching

...and gets tricky kinematic features right




Nature
Detector
Recon-
struction
Reconstr.
particles
Event
selection
Inference
Discoveries &
measurements
QFT
Scattering
amplitudes
MC
sampler
Shower +
hadron.
sim
Detector
sim
Recon-
struction
Reconstr.
particles
Event
selection
Next up:
equivariant
simulation-
based
inference

Structure:
Problem-specific inductive biases in algorithms and architectures

Geometric Algebra Transformer:
Our version of a versatile architecture for geometric problems

Scale:
Flexible architectures, lots of data, lots of compute

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

Euclidean, Projective, Conformal:
Choosing a Geometric Algebra for your Equivariant Transformer
Pim de Haan, Taco Cohen, Johann Brehmer
AISTATS 2024, arXiv:2311.04744

Probabilistic and Differentiable Wireless Simulation
with Geometric Transformers
Thomas Hehn, Markus Peschl, Tribhuvanesh Orekondy,
Arash Behboodi, Johann Brehmer
Under review

Lorentz-Equivariant Geometric Algebra Transformers
for High-Energy Physics
Jonas Spinner*, Victor Bresó*, Pim de Haan, Tilman Plehn,
Jesse Thaler, Johann Brehmer
arXiv:2405.14806

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

LaB-GATr: geometric algebra transformers for large biomedical surface and volume meshes
Julian Suk, Baris Imre, Jelmer M. Wolterink
arXiv:2403.07536


Pim de Haan


Sönke Behrends


Taco Cohen


Thomas Hehn


Markus Peschl

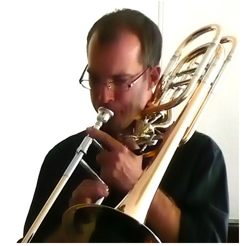

Tribhuvanesh Orekondy


Arash Behboodi


Jonas Spinner

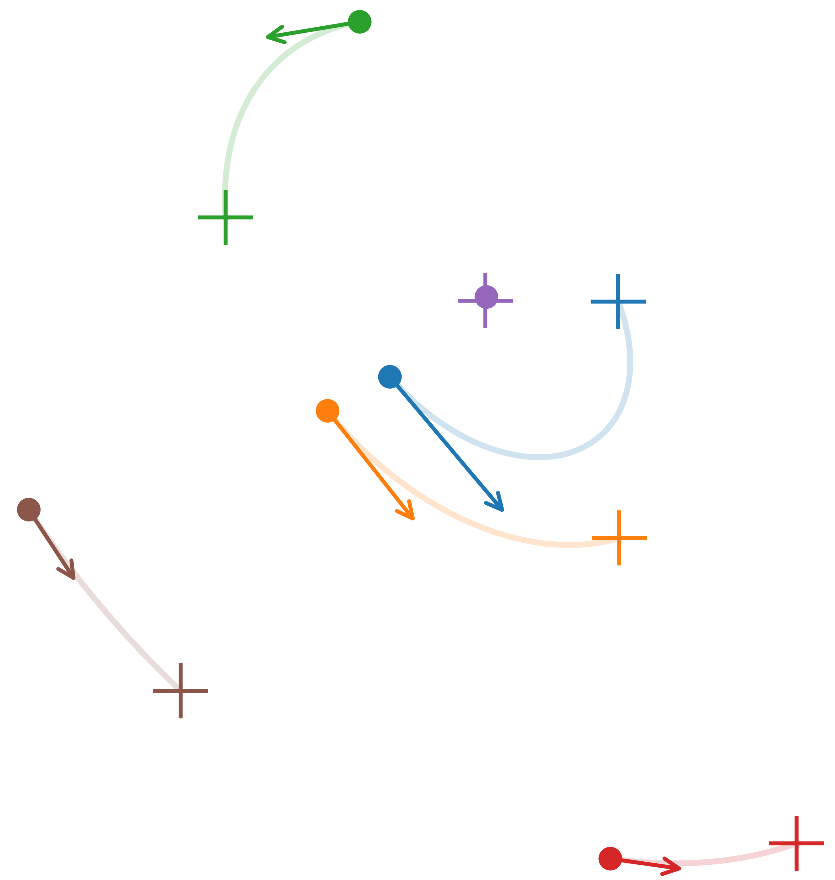

Victor Bresó

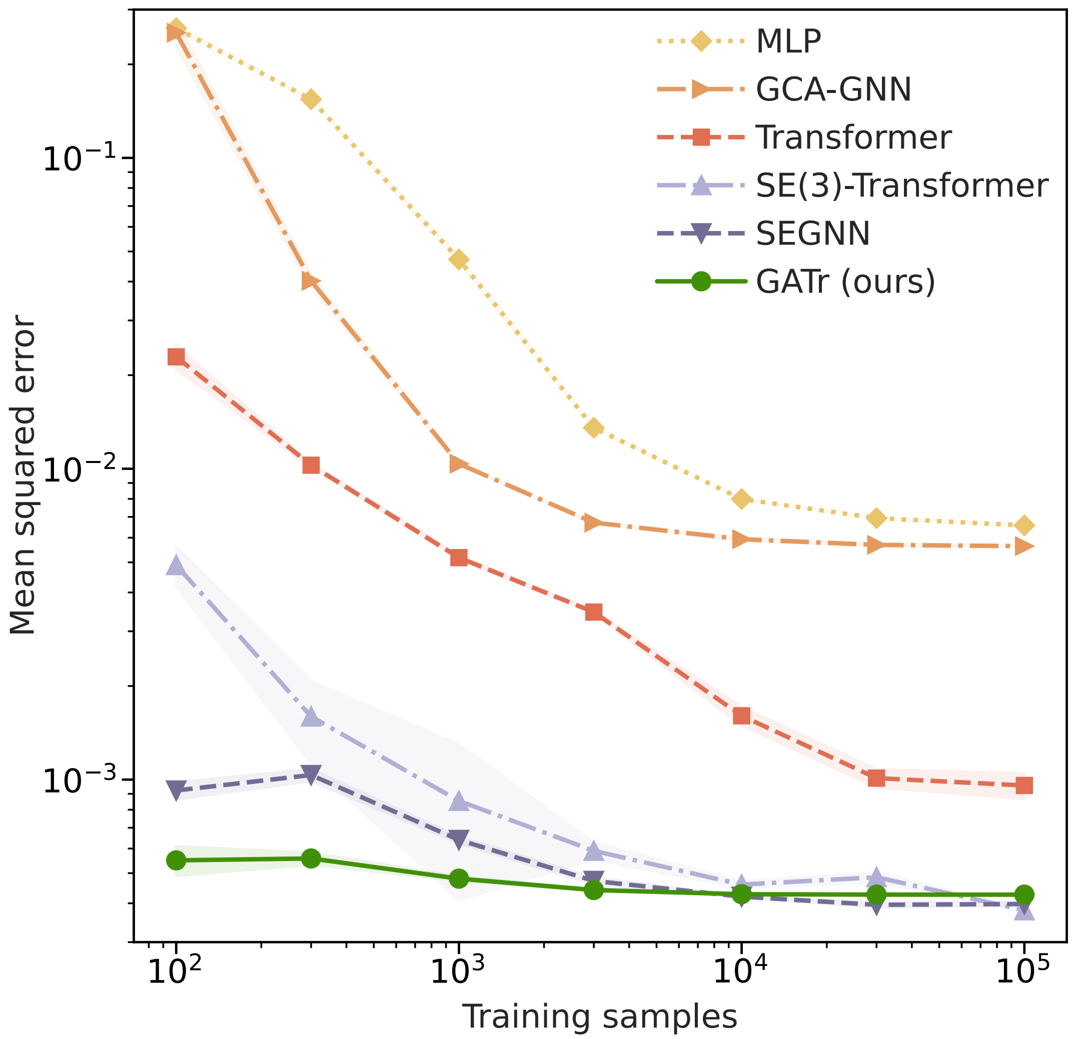

Jesse Thaler


Tilman Plehn

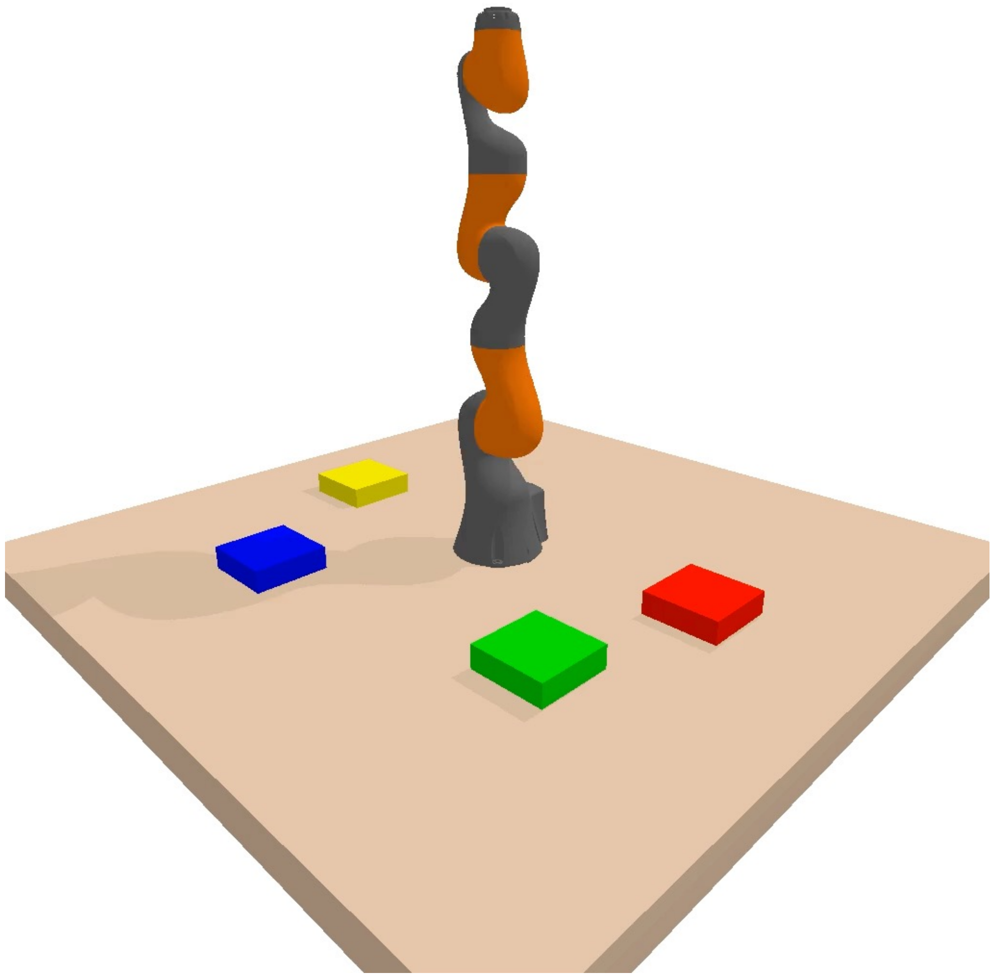
Bonus material

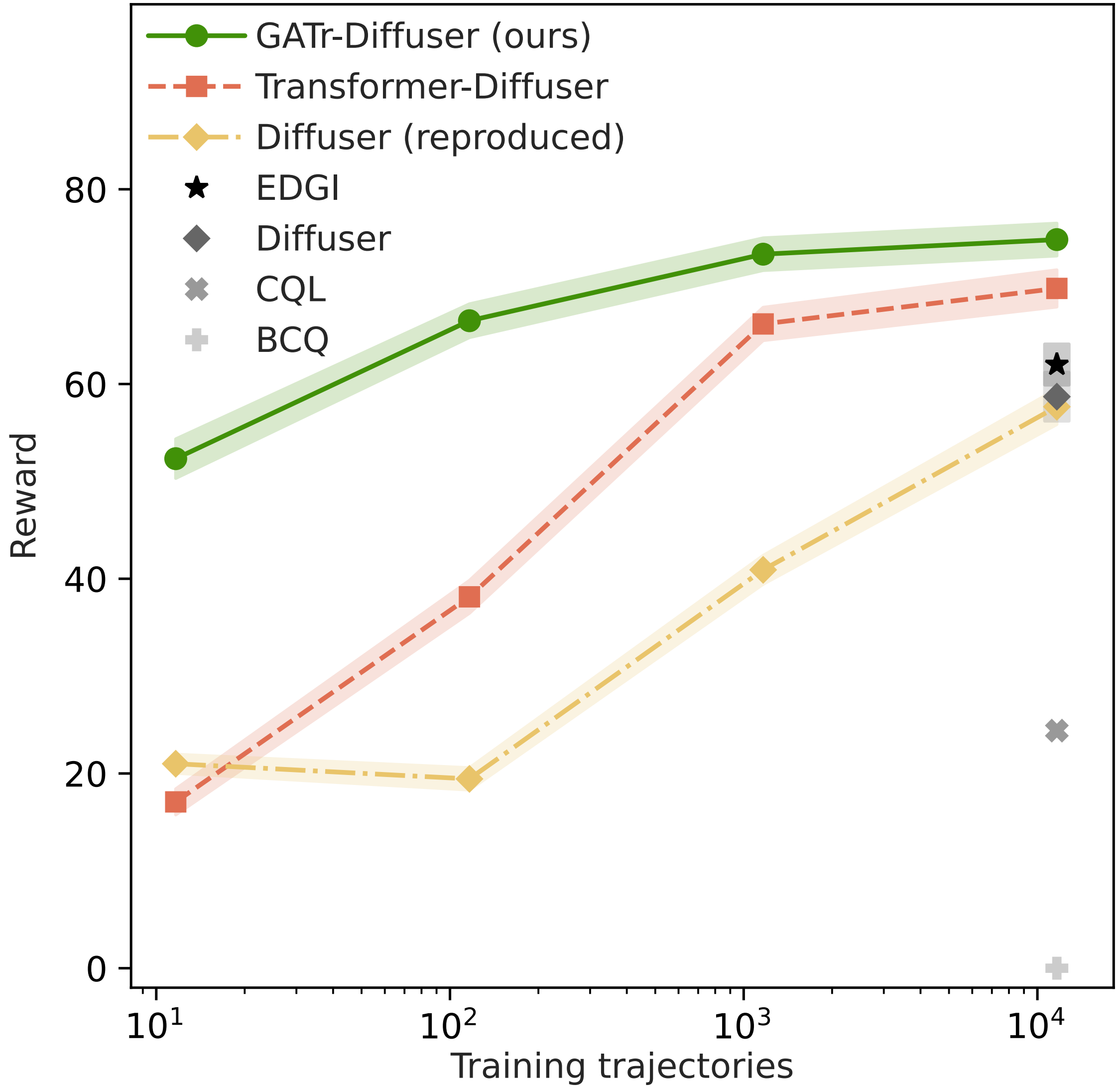
Benchmarking: **n-body modelling**



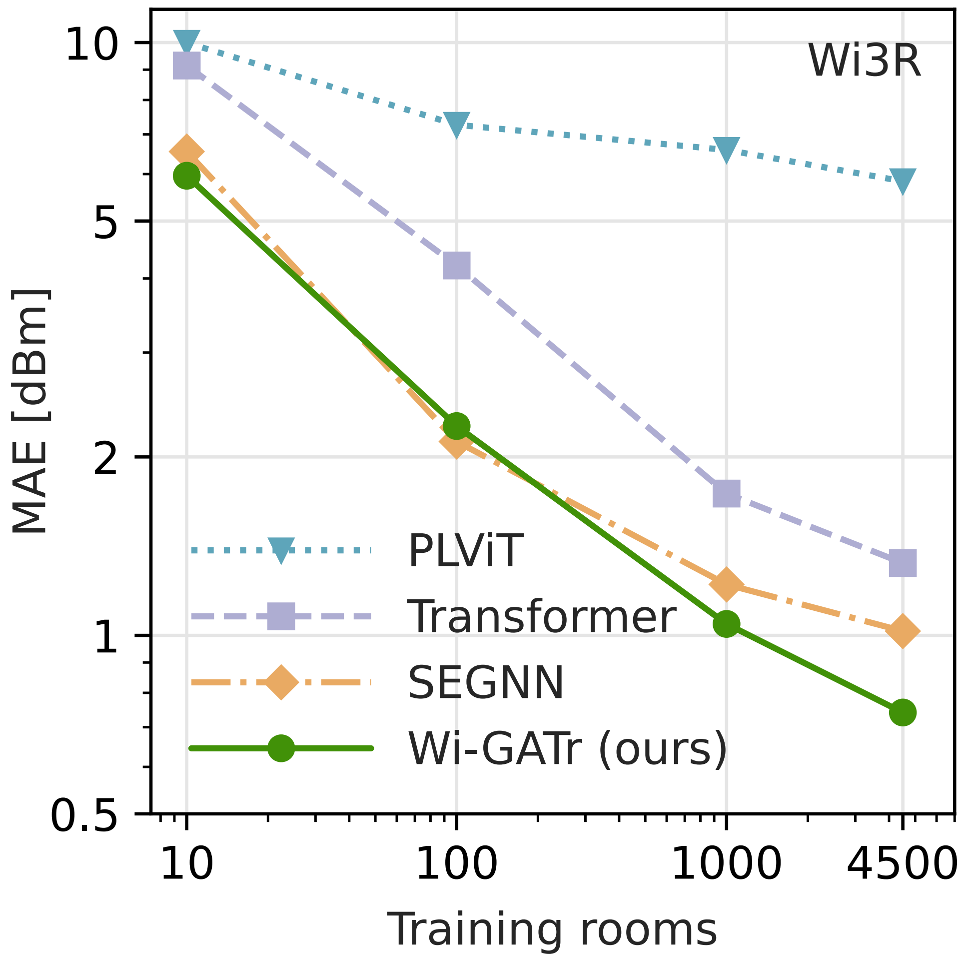


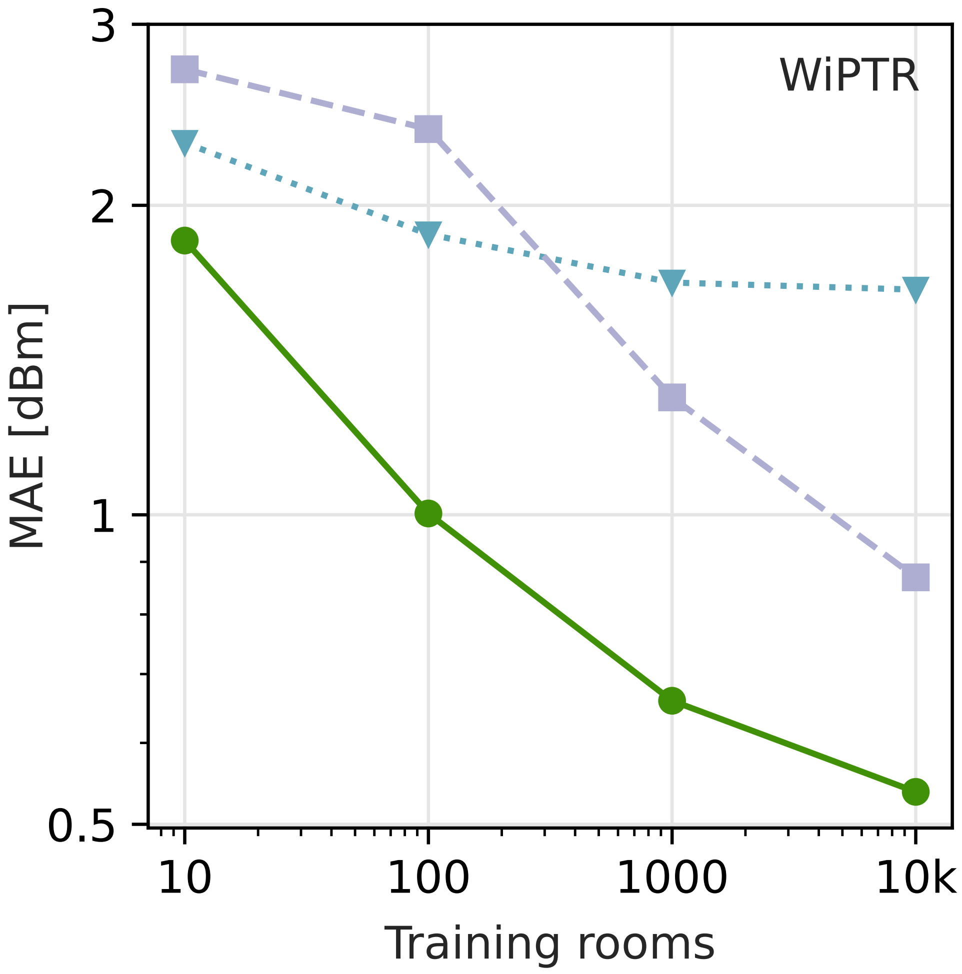
Diffusion-based planning: **Robotic control**





Differentiable **Wireless signal modelling**

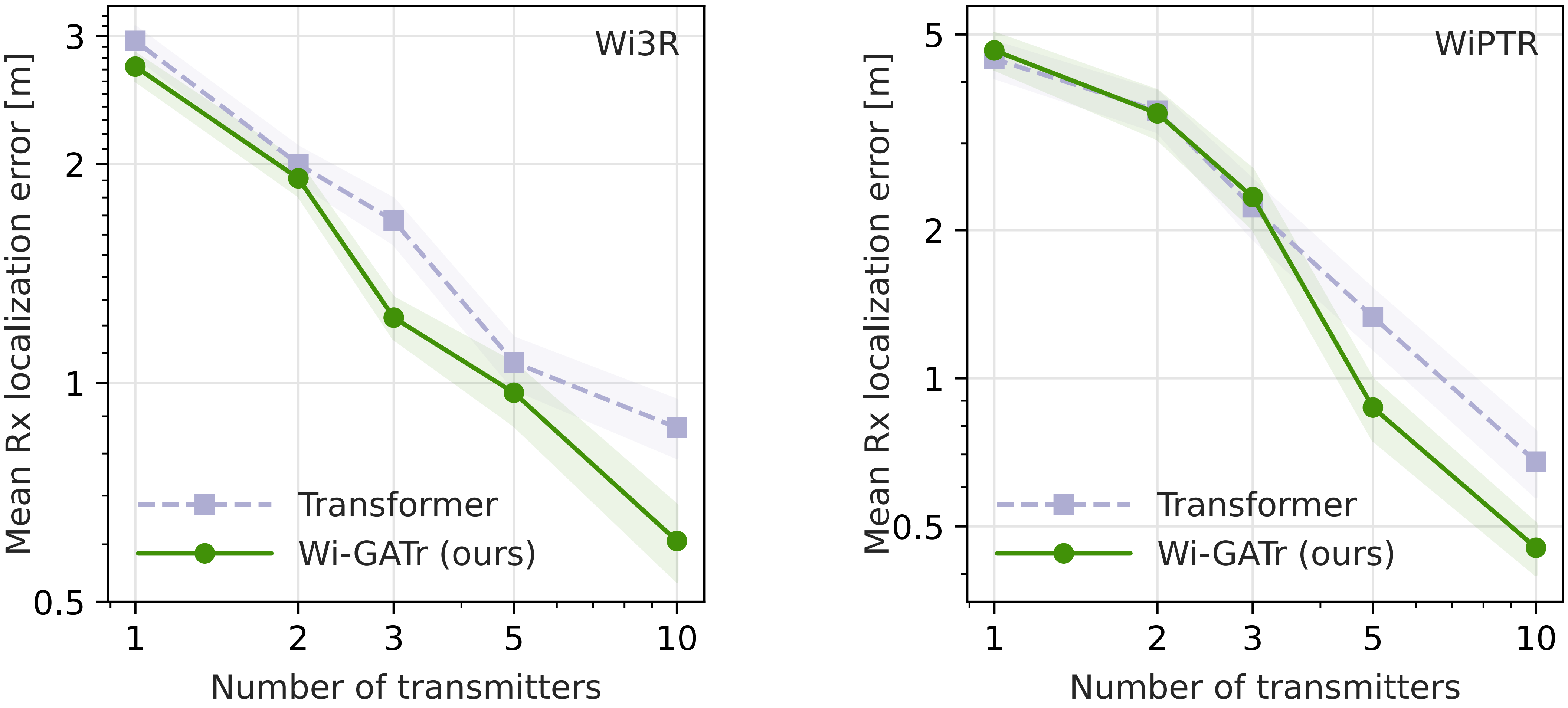




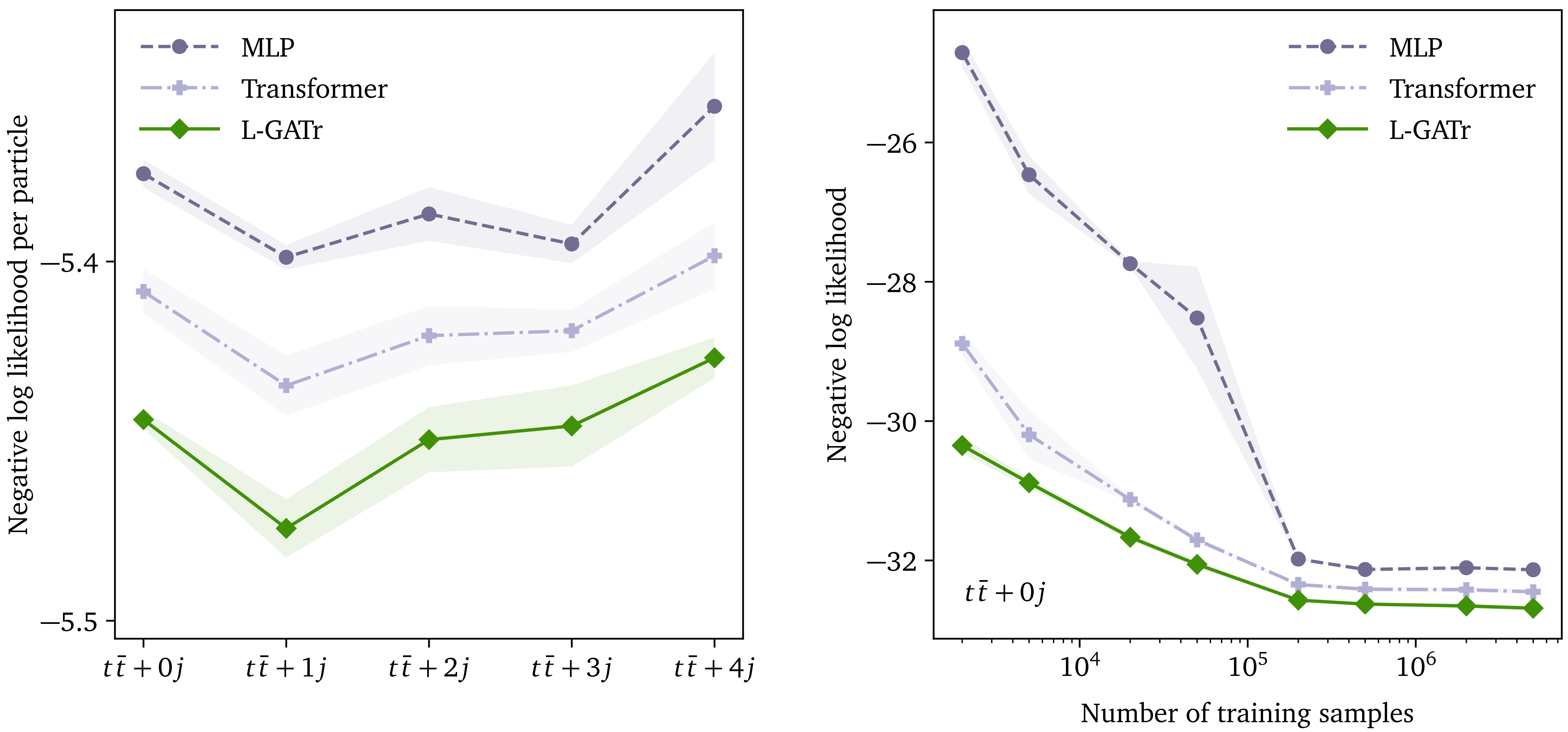
Differentiable **Wireless signal modelling**

	Wi3R dataset				WiPTR dataset		
	Wi-GATr (ours)	Transf.	SEGNN	PLViT	Wi-GATr (ours)	Transf.	PLViT
In distribution							
Rx interpolation	**0.63**	1.14	0.92	5.61	**0.53**	0.84	1.67
Unseen floor plans	**0.74**	1.32	1.02	5.84	**0.54**	0.87	1.66
Symmetry transformations							
Rotation	**0.74**	78.68	1.02	5.84	**0.54**	28.17	1.66
Translation	**0.74**	64.05	1.02	5.84	**0.54**	4.04	1.66
Permutation	**0.74**	1.32	1.02	5.84	**0.54**	0.87	1.66
Reciprocity	**0.74**	1.32	1.01	8.64	**0.54**	0.87	1.65
Out of distribution							
OOD layout	9.24	14.06	**2.34**	7.00	**0.54**	1.01	1.58

Differentiable **Wireless signal modelling**

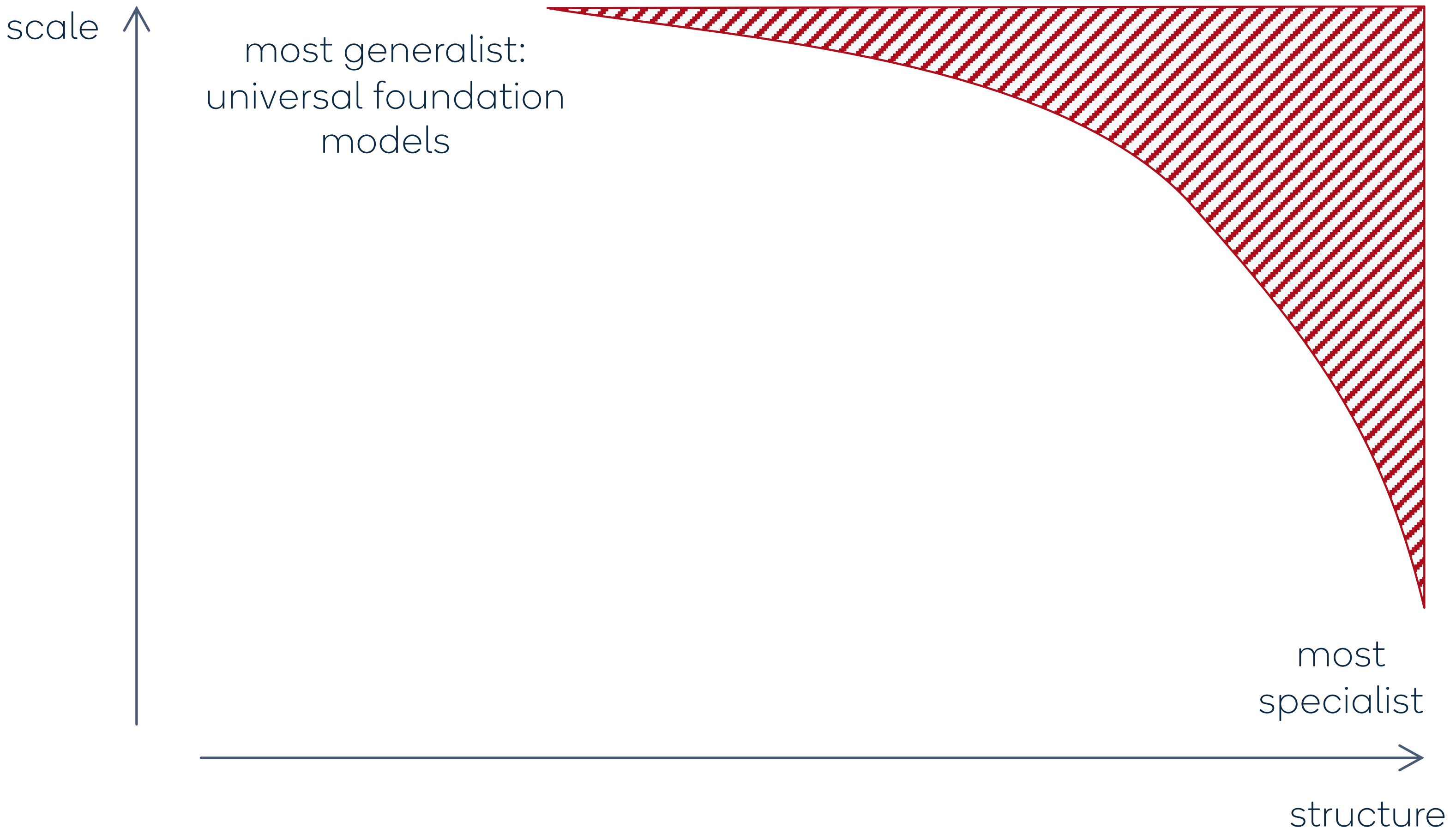


Generative modelling for hep with conditional flow matching



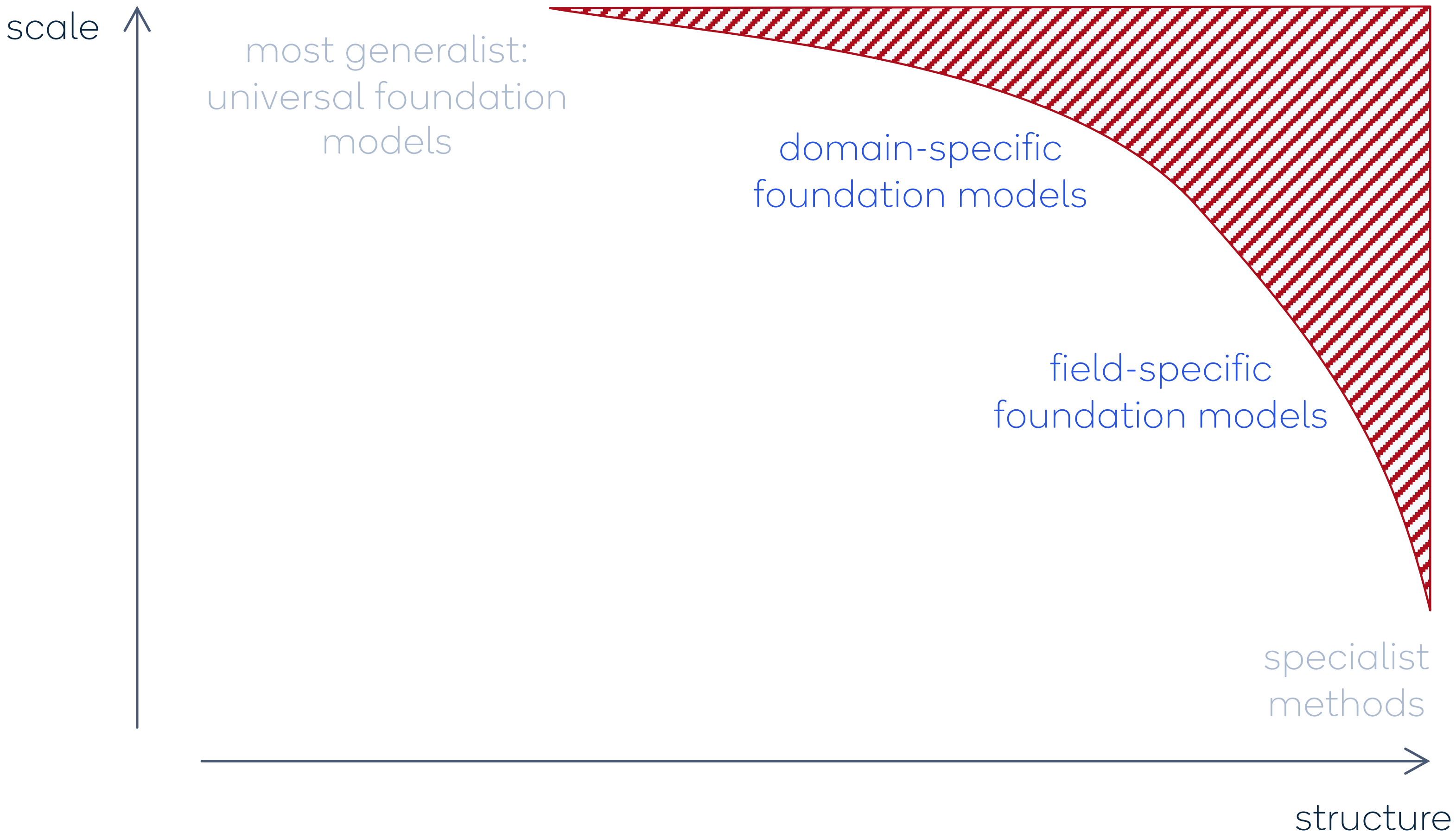
Structure vs scale

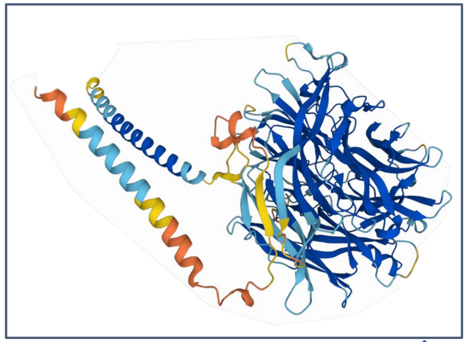
It is tempting to think of scale and structure as opposites

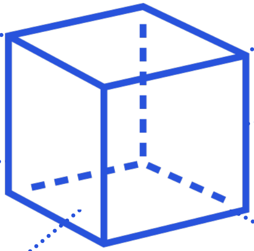


Structure vs scale

That's a false dichotomy:
there are interesting, underexplored settings where **structure and scale can be complementary**

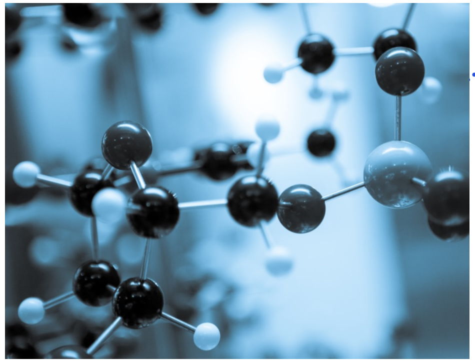


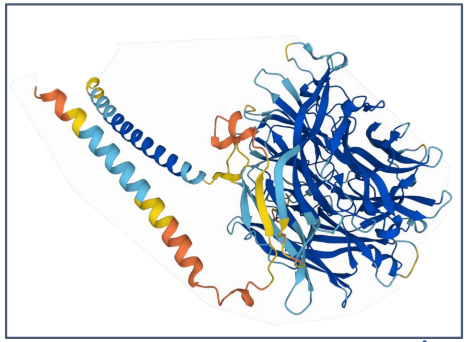




Geometric foundation model

- Trained on (and finetuned to) various scientific and engineering problems
- Representations tailored to spatio-temporal data
- Architecture reflects symmetries of 3D space





Orientations

Spatial symmetries

Temporal symmetries

Time

3D space

Positions

Movement

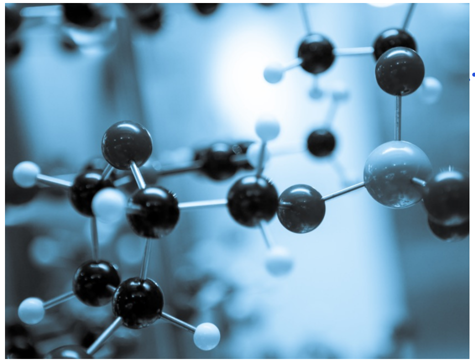
Distances

Distance-dependent forces

3D shapes

Permutation symmetries

Collisions



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

Foundation model checklist	Geometric foundation model
☐ Lots of **data**	
☐ Universal **representations**	
☐ A scalable, expressive **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**	?
☐	Universal **representations**	**could be geometric algebra–based**
☐	A scalable, expressive **architecture**	**equivariant Transformers**
☐	Self-supervised **training** protocol	?
☐	Multiple **downstream tasks** with shared structure	**GATr works on very different domains**
☐	Bonus points for **predictable scaling**	?

Foundation model checklist	Geometric foundation model
☐ Lots of **data**	data and simulators are out there?
☐ Universal **representations**	**could be geometric algebra–based**
☐ A scalable, expressive **architecture**	**equivariant transformers**
☐ Self-supervised **training** protocol	predicting masked-out tokens?
☐ Multiple **downstream tasks** with shared structure	**GATr works on very different domains**
☐ Bonus points for **predictable scaling**	**?**

Thank you

Qualcomm

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog