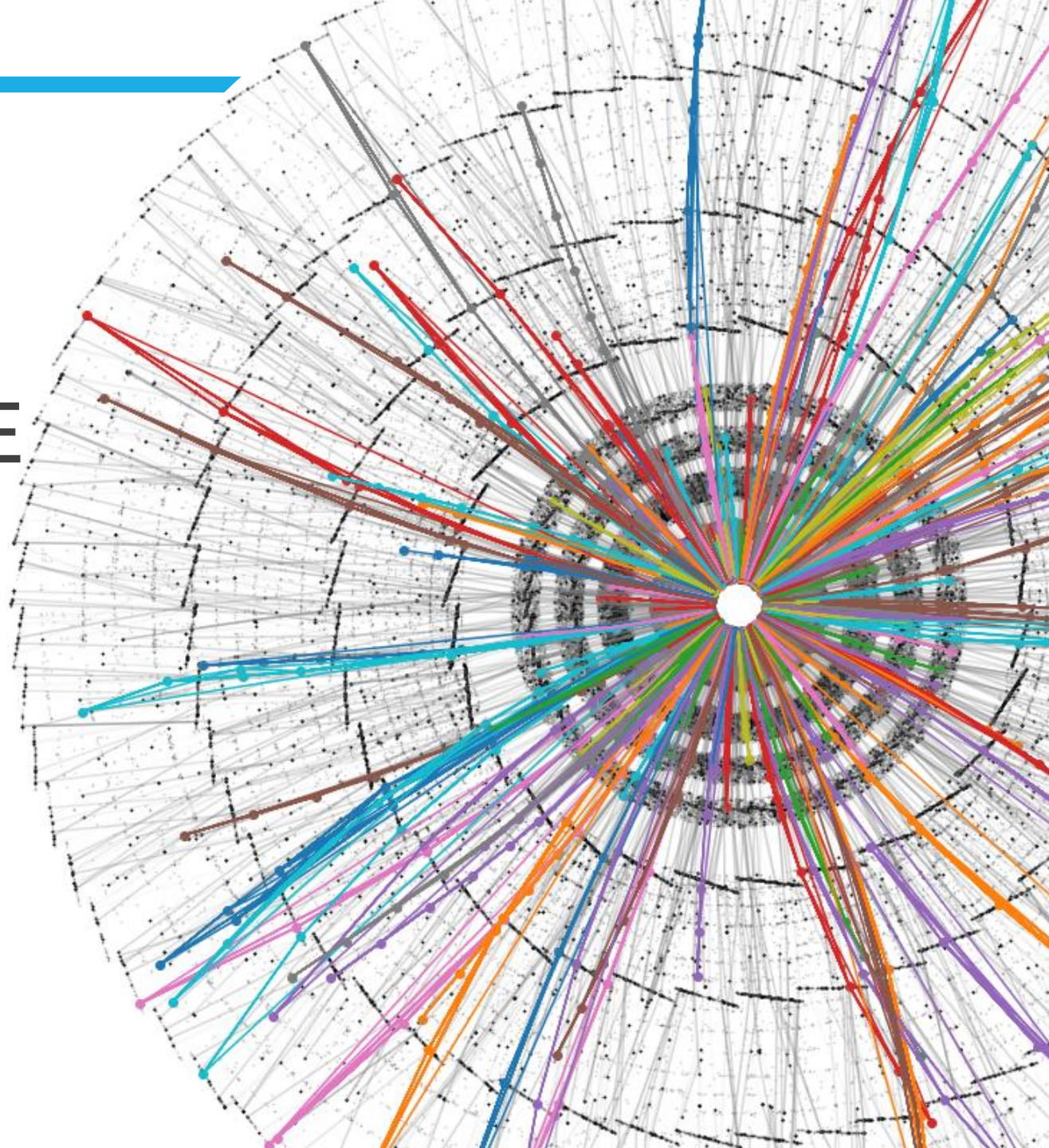
NEXT TOKEN PREDICTION FOR ARBITRARY PARTICLE PHYSICS TASKS

Daniel Murnane
Niels Bohr Institute, University of Copenhagen



UNIVERSITY OF COPENHAGEN

Danish Data Science Academy

TrkX

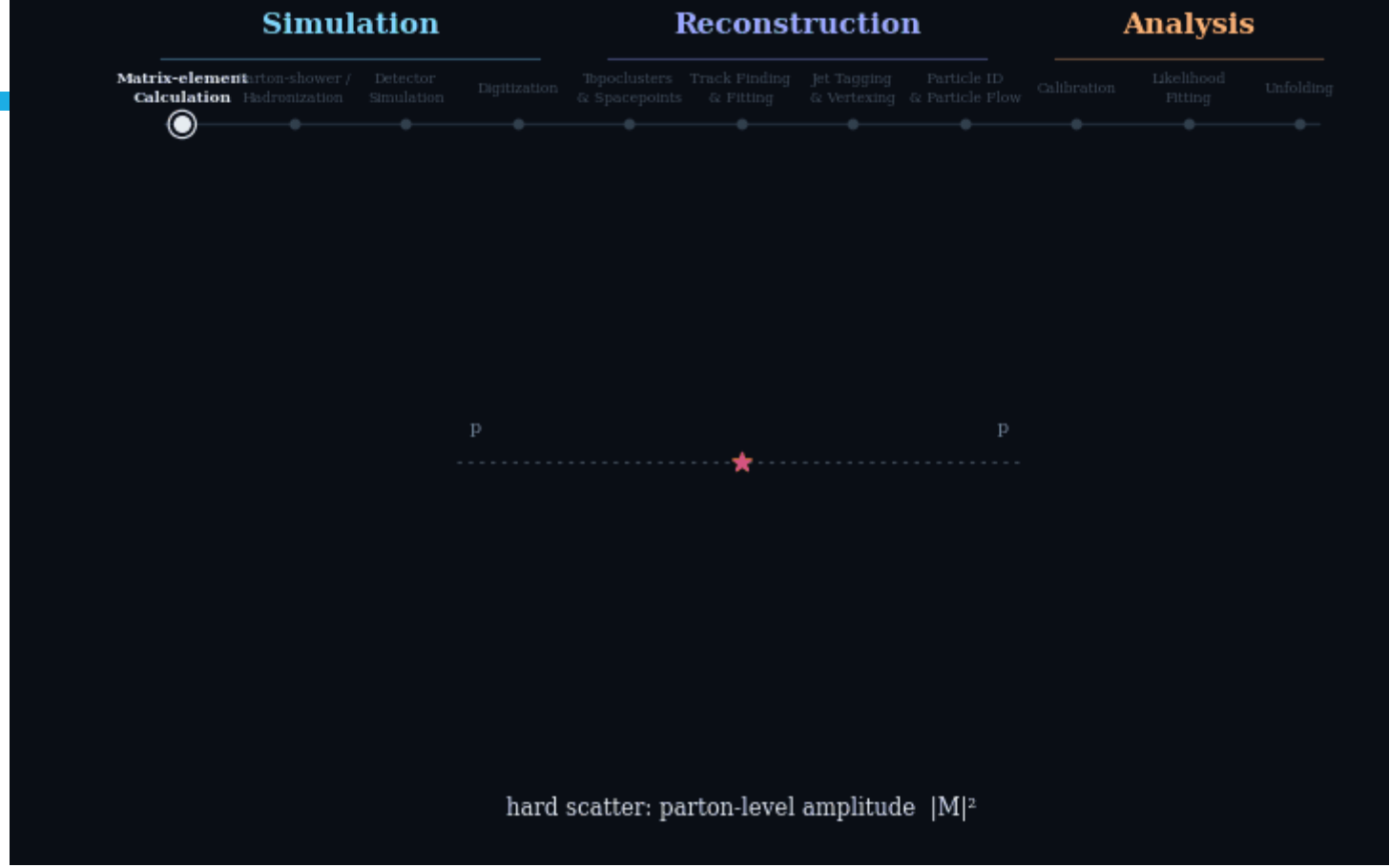
BERKELEY LAB

ATLAS EXPERIMENT

CAN LANGUAGE HELP US WITH PHYSICS?

A DAY IN THE LIFE OF AN EXPERIMENTAL PARTICLE PHYSICIST

- Particle physics has LOTS of tasks
- They need to be calibrated and interpreted, cannot simply be a black box
- We don't want to train O(NM) models (for N inputs, M outputs)
- We may not even want to train with truth labels (e.g. for calibration or simulation purposes)



Simulation

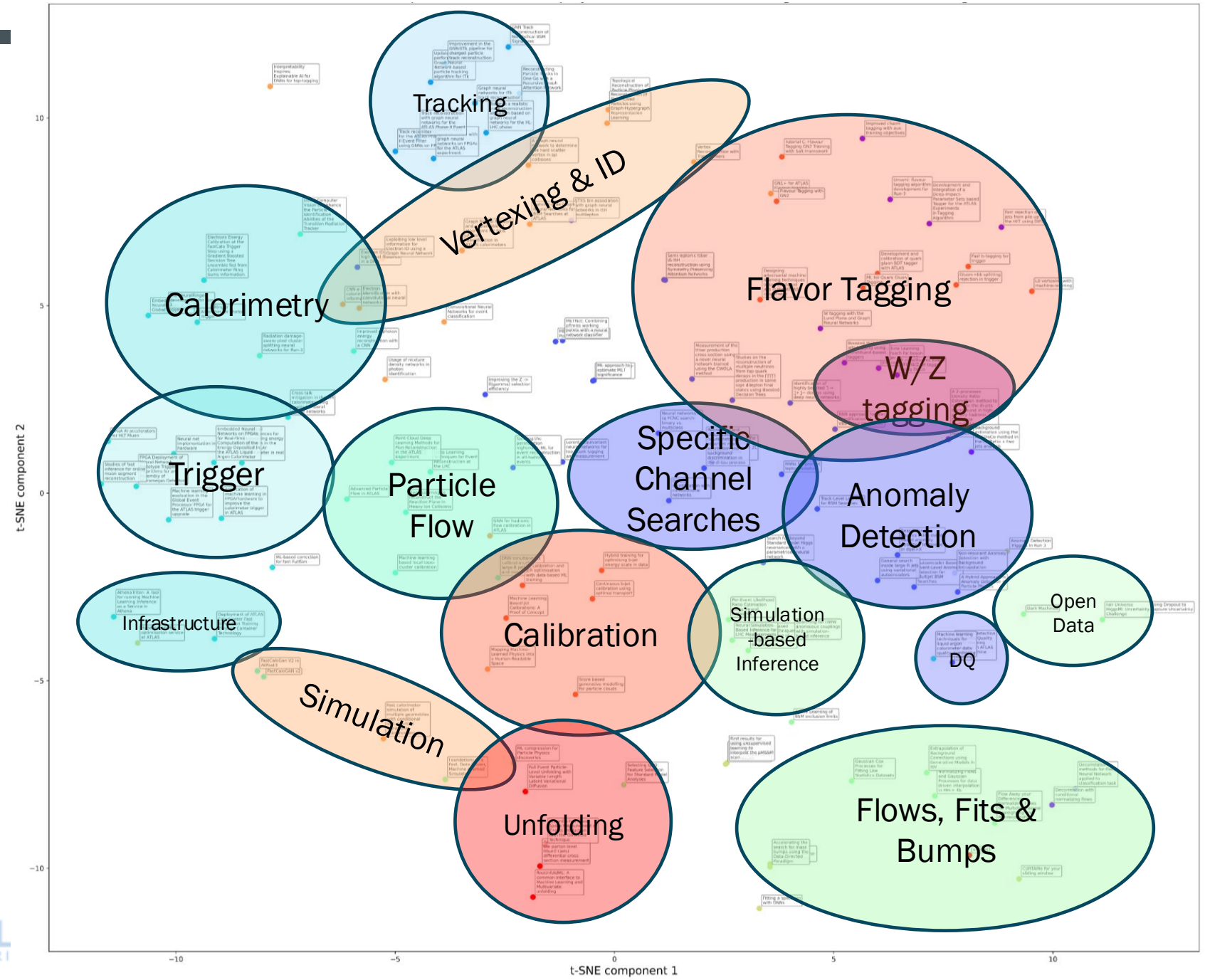
Reconstruction

Analysis

Matrix-element Calculation	Parton-shower / Hadronization	Detector Simulation	Digitization	Topoclusters & Spacepoints	Track Finding & Fitting	Jet Tagging & Vertexing	Particle ID & Particle Flow	Calibration	Likelihood Fitting	Unfolding
Numerical Integration	Markov Chain Monte Carlo		Topological clustering		Kalman Filtering & Fitting		Conformal Fits & Hough Transform		Statistical Techniques, Bayesian Inference	

WHAT DOES ML IN THE ATLAS EXPERIMENT “LOOK LIKE”?

Embed every talk in the ATLAS ML Workshop (135 talks) from the last 3 years

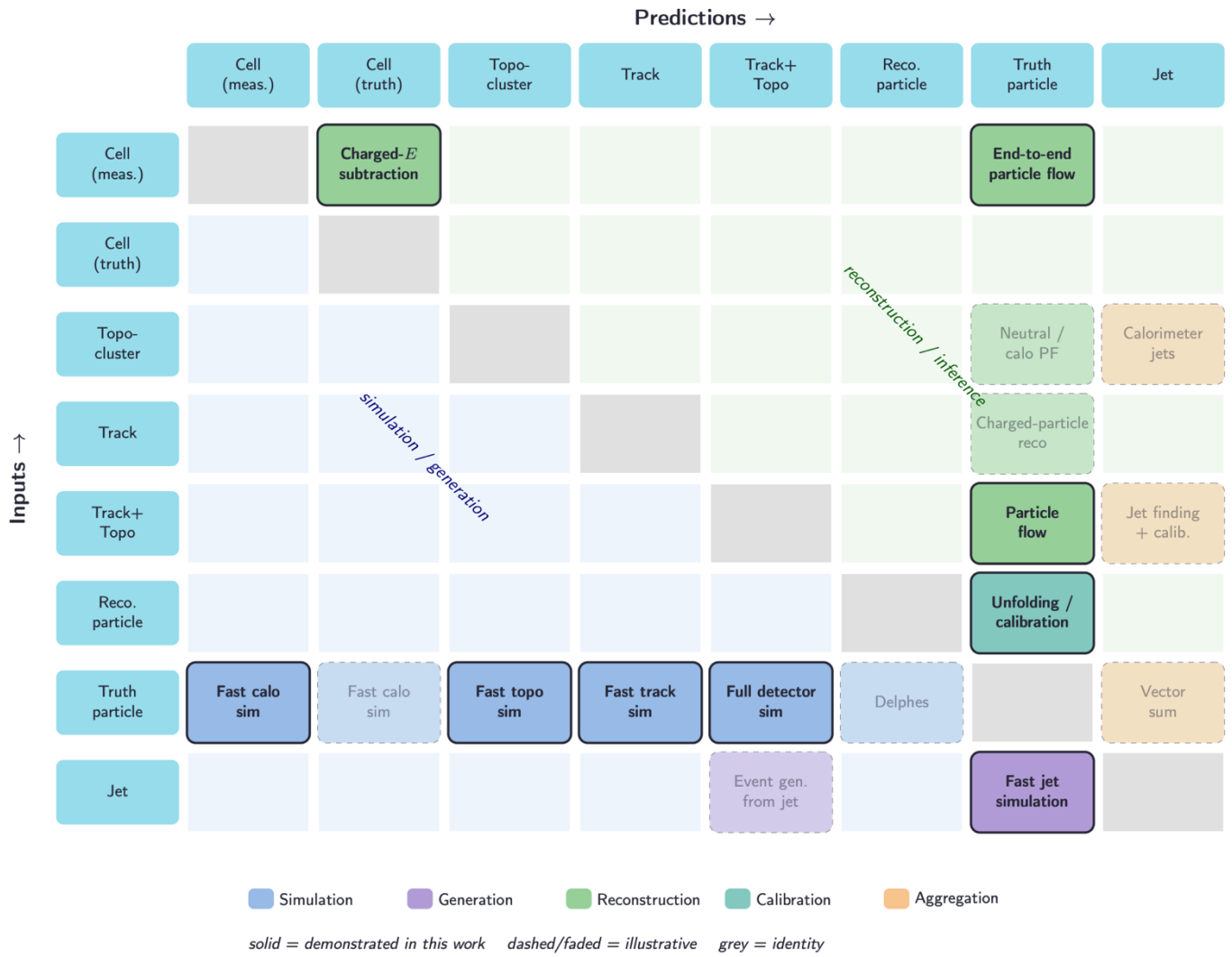


UNIVERSITY OF COPENHAGEN

BERKELEY LAB

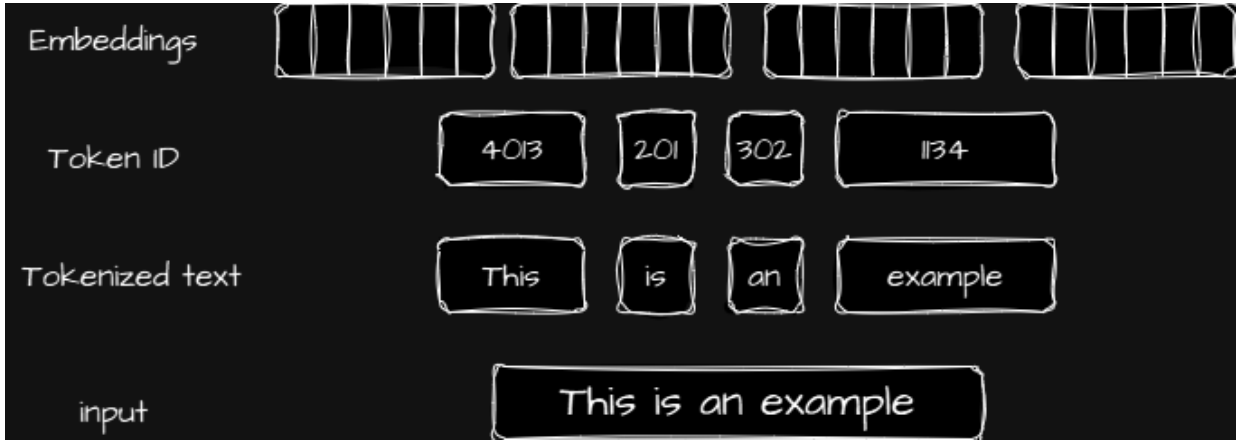
ATL EXPERI

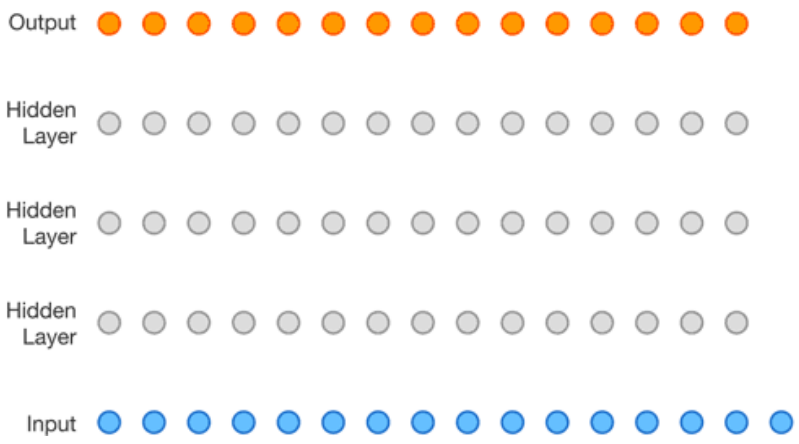
TASKS CAN BE CAPTURED BY INPUTS AND OUTPUTS



HOW DO LARGE LANGUAGE MODELS SOLVE THIS?
TOKENS & AUTOREGRESSION

- Every object that can be represented by language is converted to a common form = “token”
- English, Chinese, numbers, emojis → one big one-hot vector

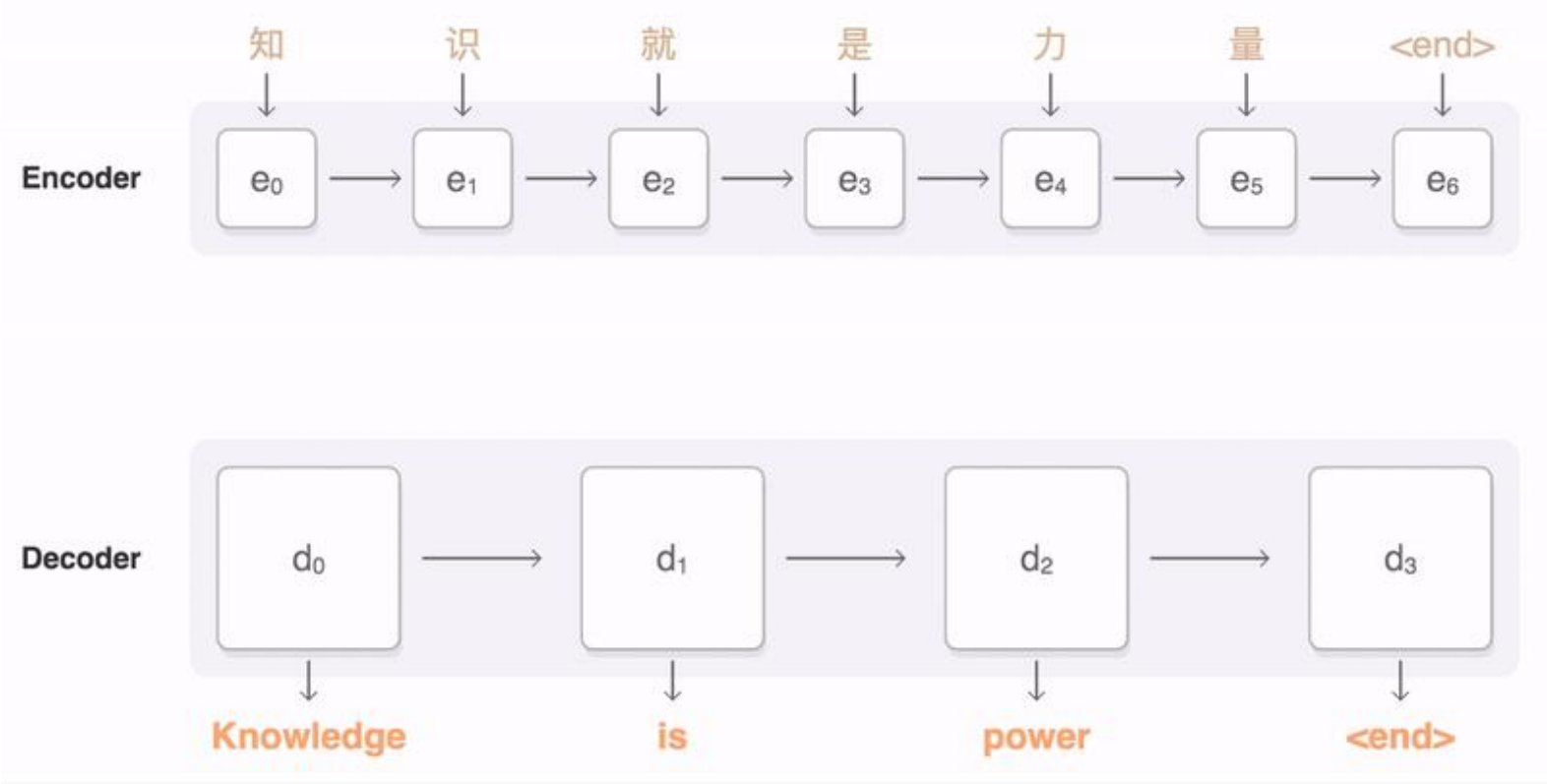




- The question of defining a task is then simply presenting a sequence of previous tokens = “prompt”
- Switching between input and output, different tasks or modalities, can be done with special tokens. E.g. “EOS” tells the model to stop generating

HOW DO LARGE LANGUAGE MODELS SOLVE THIS?
SOME INDUSTRY TRICKS...

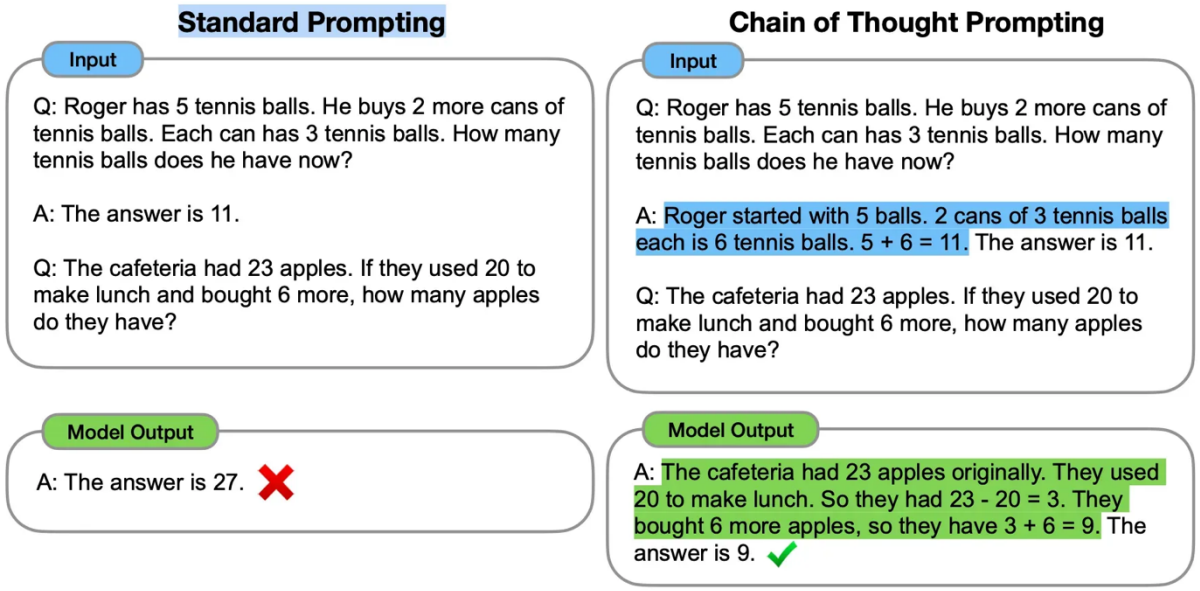
- Some small extra points:
  - Teacher forcing
  - KV caching
  - Causal attention



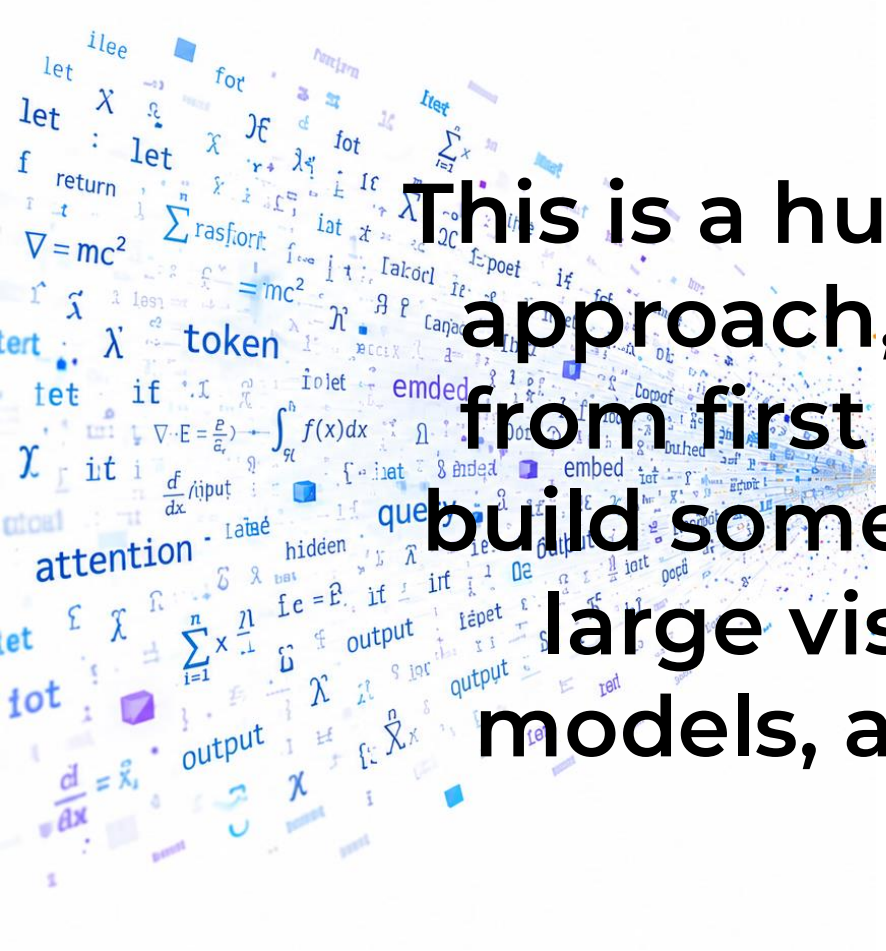
HOW DO LARGE LANGUAGE MODELS SOLVE THIS?
EMERGENT BEHAVIOUR

From only this training task, we seem to get “emergent” behavior:

- Some level of generalizability across tasks
- The ability to improve performance by “thinking” (version 1 = chain of thought, version 2 = reasoning)
- In-context learning (new tasks given as examples)



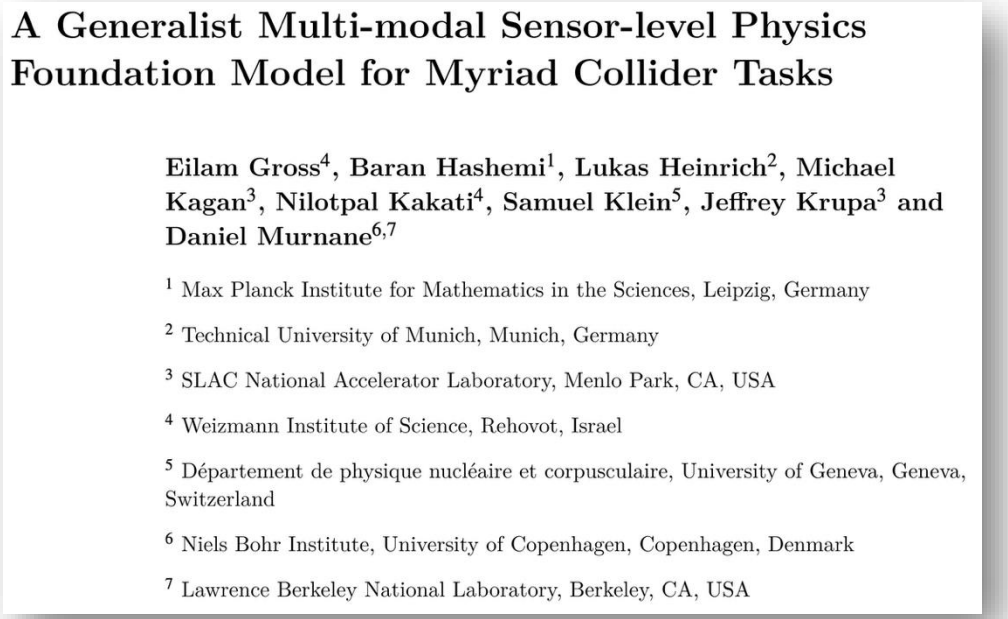
IS NATURAL LANGUAGE SIMILAR TO (PARTICLE) PHYSICS?



This is a huge and open question. Our approach, rather than approaching from first principles, was instead to build some models that borrow from large vision and large language models, and see how they behave!

BUILDING A GENERALIST PHYSICS MODEL

WORK BASED ON UPCOMING PRE-PRINT →



A Generalist Multi-modal Sensor-level Physics Foundation Model for Myriad Collider Tasks

Eilam Gross[4], Baran Hashemi[1], Lukas Heinrich[2], Michael Kagan[3], Nilotpal Kakati[4], Samuel Klein[5], Jeffrey Krupa[3] and Daniel Murnane[6,7]

[1] Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

[2] Technical University of Munich, Munich, Germany

[3] SLAC National Accelerator Laboratory, Menlo Park, CA, USA

[4] Weizmann Institute of Science, Rehovot, Israel

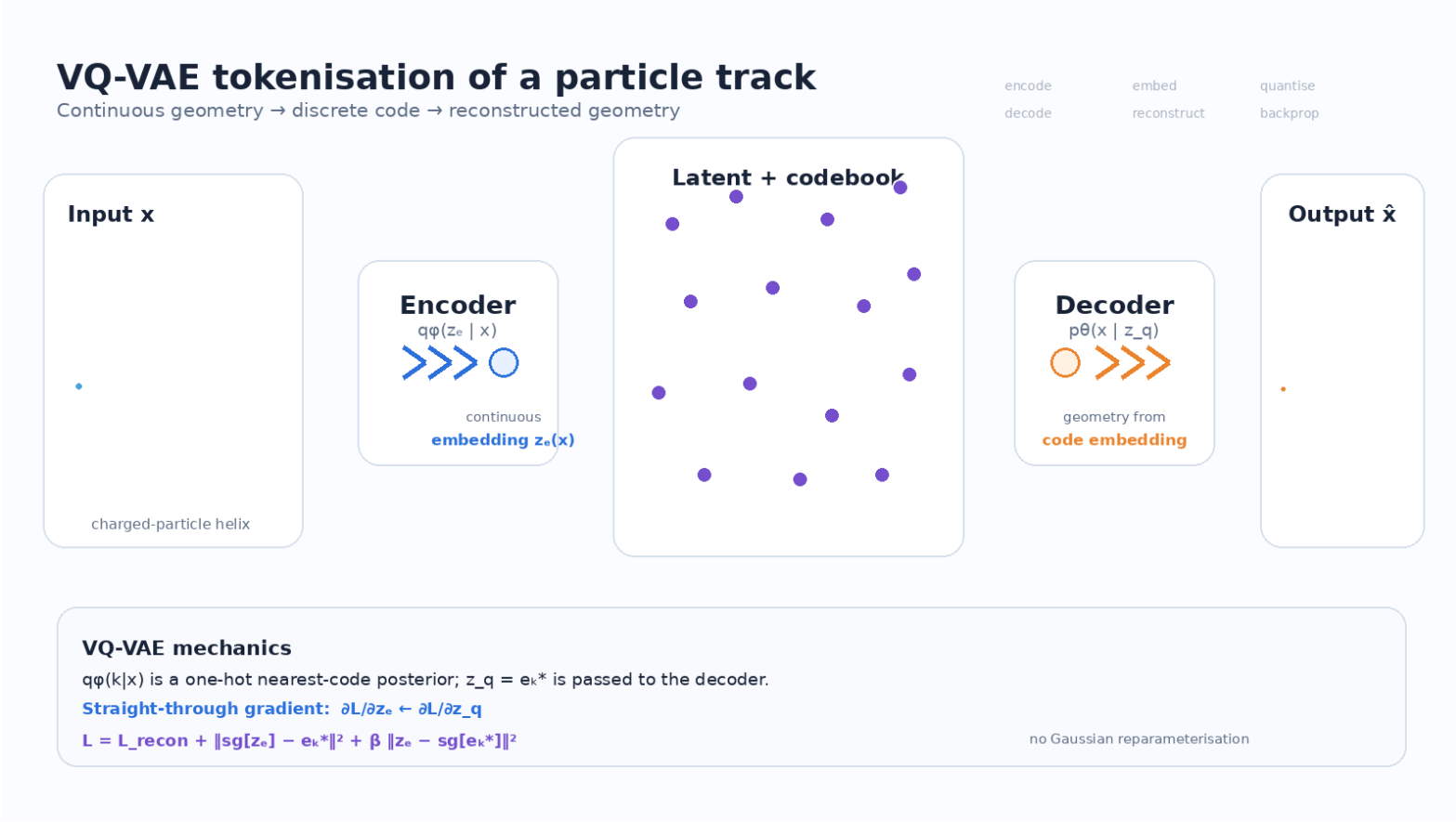
[5] Département de physique nucléaire et corpusculaire, University of Geneva, Geneva, Switzerland

[6] Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

[7] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

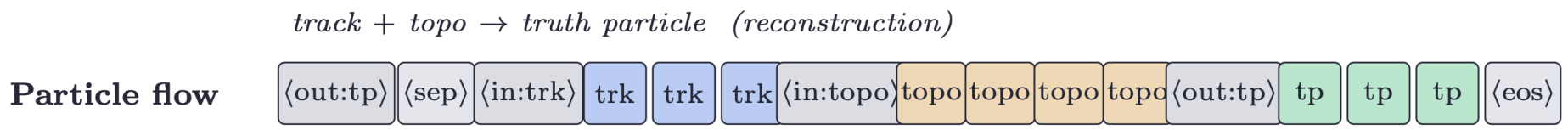
DATA & TOKENISATION

- Data is simulated in the simplified COCOA detector geometry (calorimeter full sim, tracker fast sim)
- 100M jets
- 7 modalities: tracks, topoclusters, truth particles, reco particles (from HGPFlow), cell energies, jet-level features, reconstructed cells (for calibration)
- Tokenise with RVQ-VAE



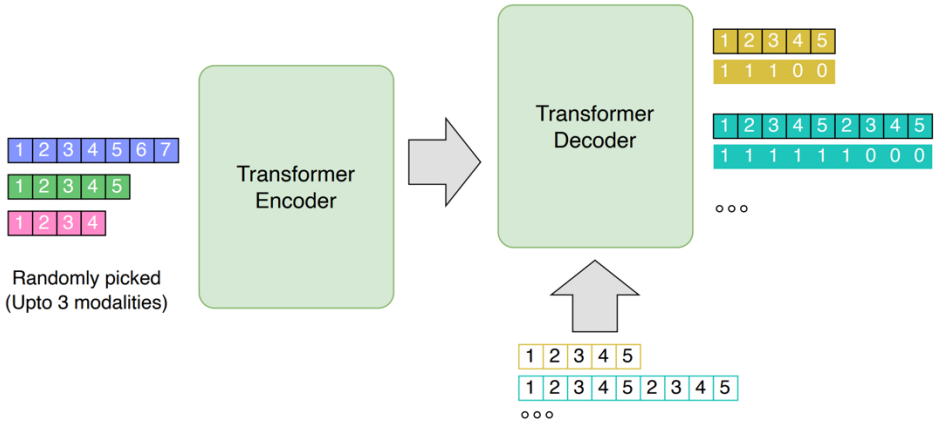
DEFINING TASKS

- We propose two ways to define tasks, which depend on how we decode and sample a sequence…
- We can decode and sample **in parallel**: All tokens produced at the same time. In that case, we have a ”cardinality” (length) head to predict how many of each modality should be produced first, then we decode that many of each
- We can decode and sample **autoregressively**: Produce one token at a time, conditioned on all previous input and output tokens. In that case, learn when to switch modalities, and finish the task, with dedicated modality and EOS tokens
- A header then “prompts” the model for the modalities we want to receive

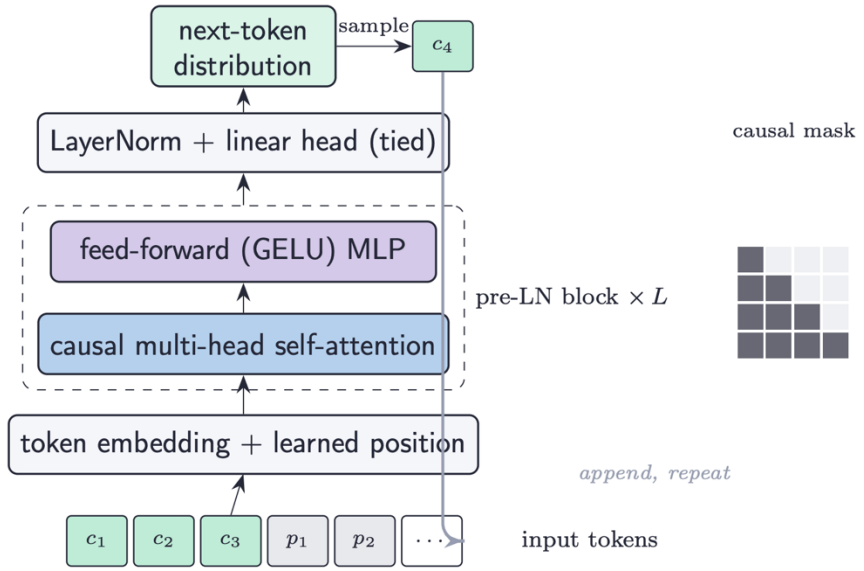


CHOOSING AN ARCHITECTURE

- Parallel architecture based on 4M model from Apple AI Research group (call it *HEP4M*)



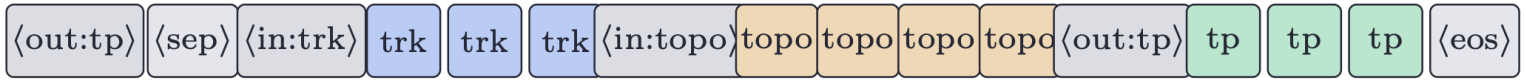
- Autoregressive architecture based on nanoGPT (call it *nanoHEP)*



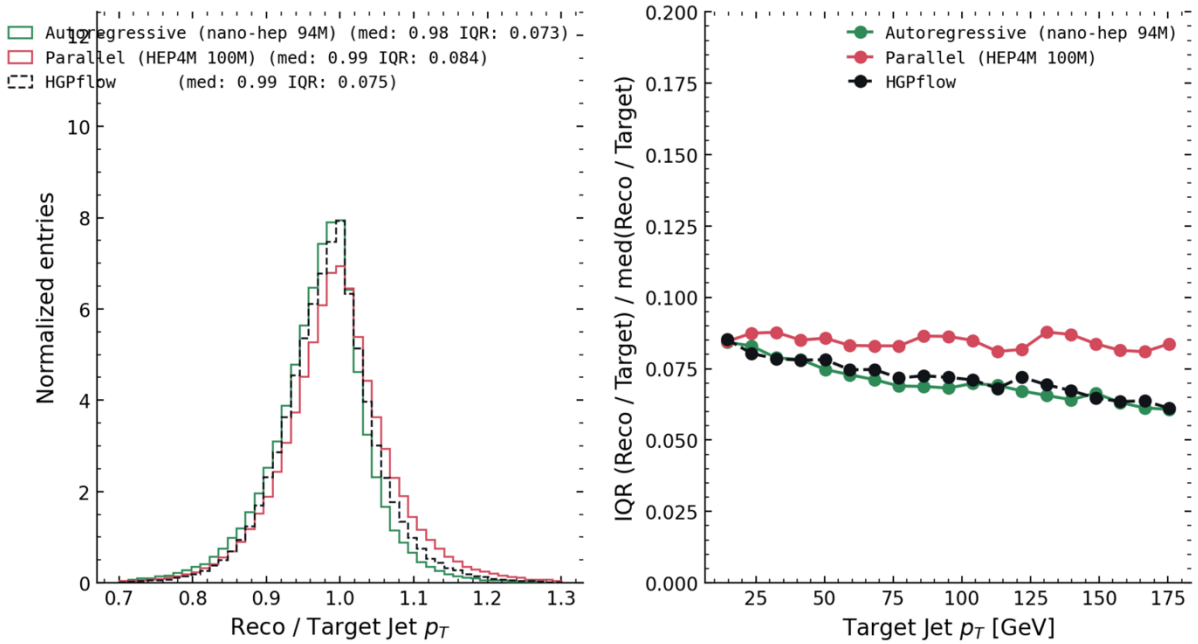
SAMPLE TASKS
PARTICLE FLOW RECONSTRUCTION

$$track + topo \rightarrow truth\ particle \quad (reconstruction)$$

Particle flow



- Particle flow is a nice task: High impact, well-defined, low-cardinality, with great baseline model
- HGPFlow is a current contender for pflow inside ATLAS – current state-of-the-art
- Contains several physics-based heuristics (e.g. conservation of energy)
- HEP4M is competitive
- NanoHEP outperforms
- Both are simply “filling in the blank”



SAMPLE TASKS
DETECTOR SIMULATION

- Can easily define the task of detector simulation:

$$truth\ particle \rightarrow track + cell + topo \quad (generation)$$

Detector simulation

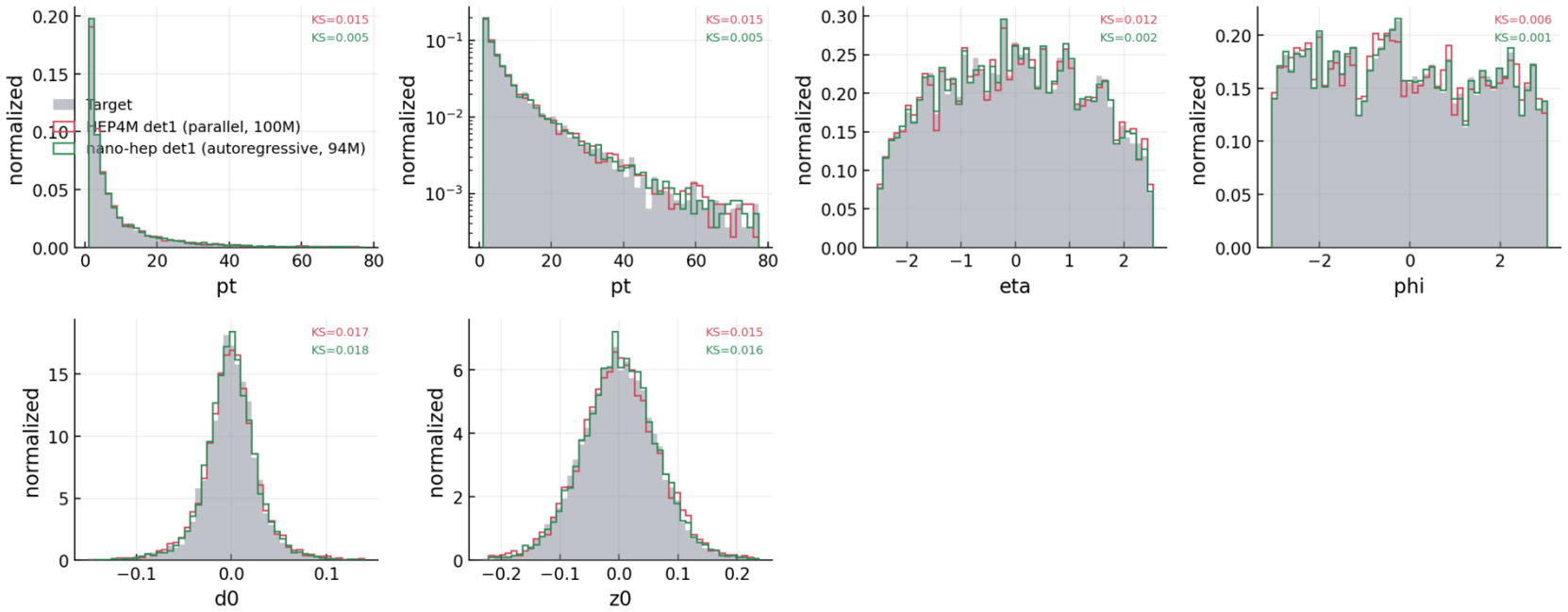
- But how should we evaluate output?

EVALUATING FOUNDATION MODEL BEHAVIOUR

EVALUATION
STANDARD APPROACHES

- We can inspect the marginal distributions of a generative model, and we would expect that the generator and training sample should be similar (can quantify with Kolmogorov-Smirnov = KS)

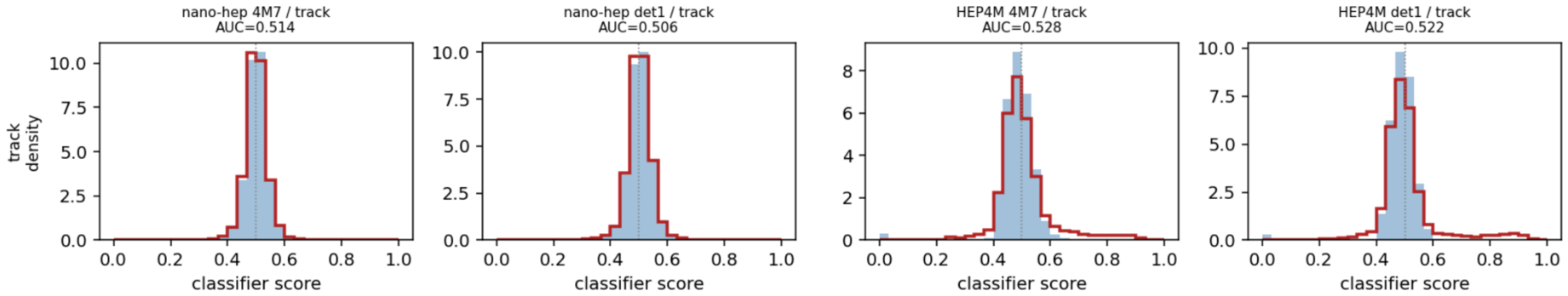
- Indeed, good match for both here!
- Hard to interpret this, and it doesn't tell us how each event looks



EVALUATION
MODEL-AS-A-JUDGE

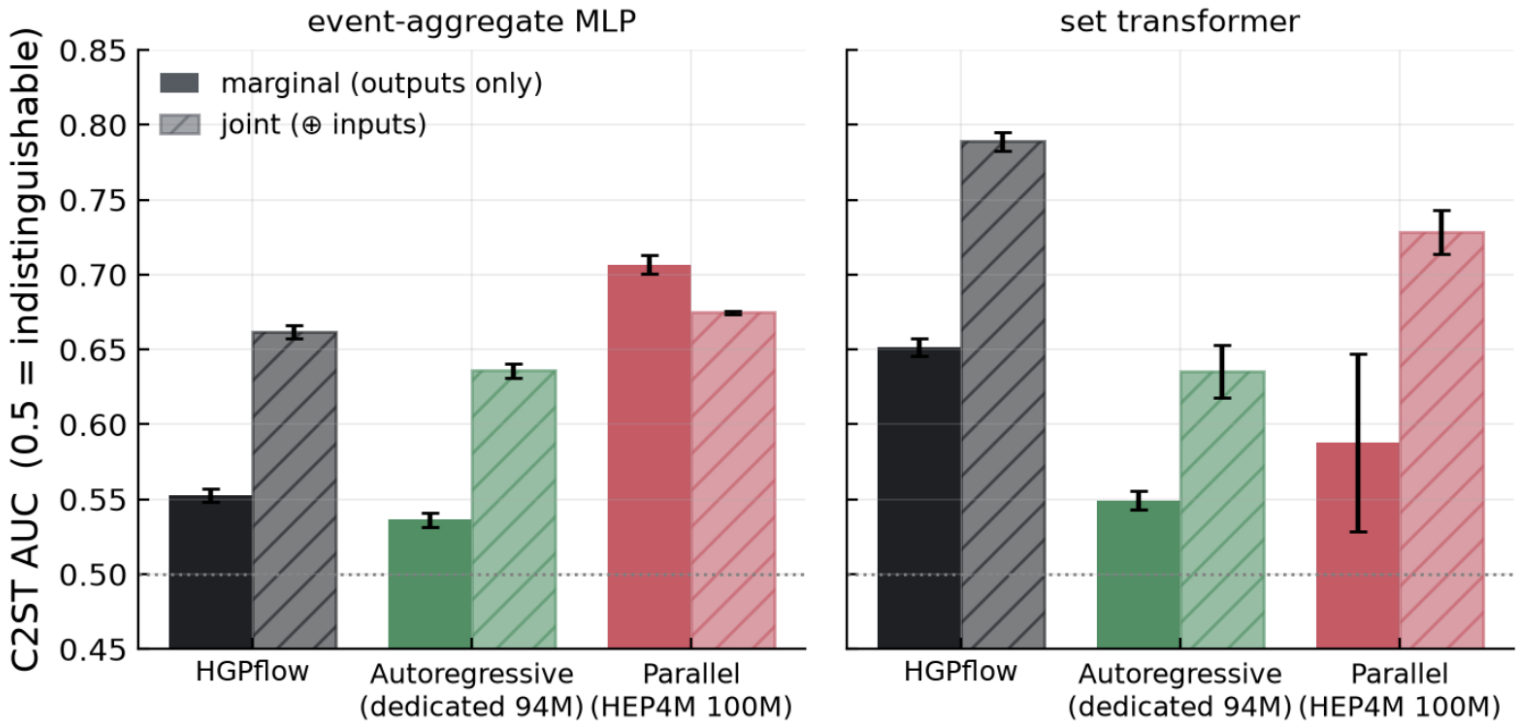
- We can try to define per-event quality by training a classifier to distinguish generated events from “real” (Geant4) events
- Even better: Let’s give the classifier the FULL sequence. That is, we want the generator to not only generate realistic events, but they should match the input conditions

- Nano-hep subtly better here than HEP4M



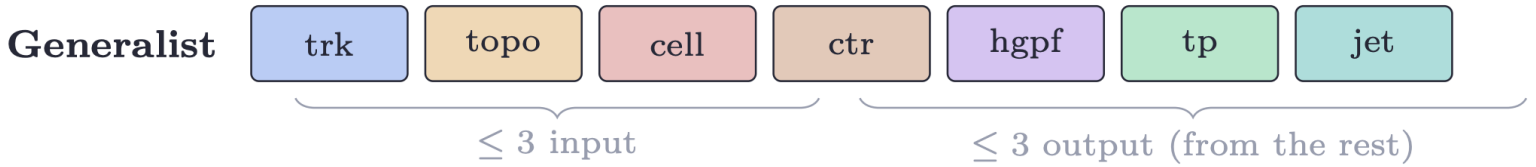
EVALUATION
MODEL-AS-A-JUDGE

- Nano-hep really shines when we look back at particle flow
- Interestingly, HGPflow looked so good with jet-level residuals
- But a classifier can tell its reconstruction apart from truth!
- NanoHEP reconstructs much closer to truth objects
- HEP4M in between

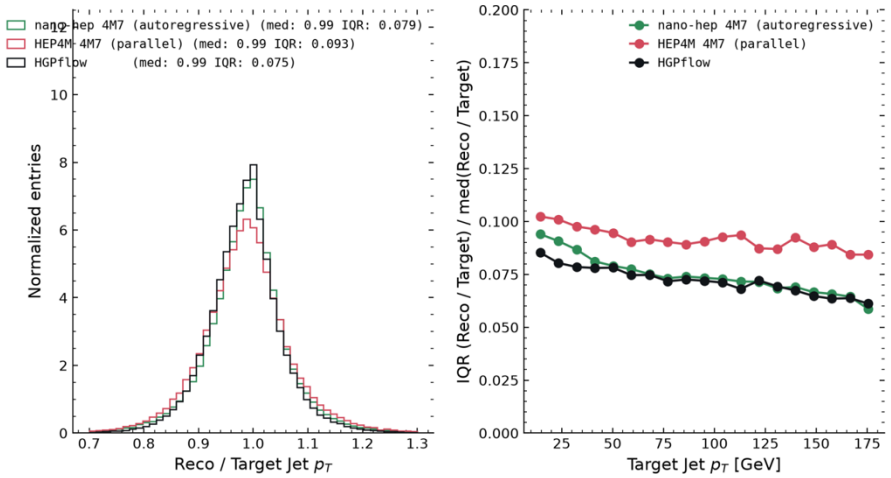


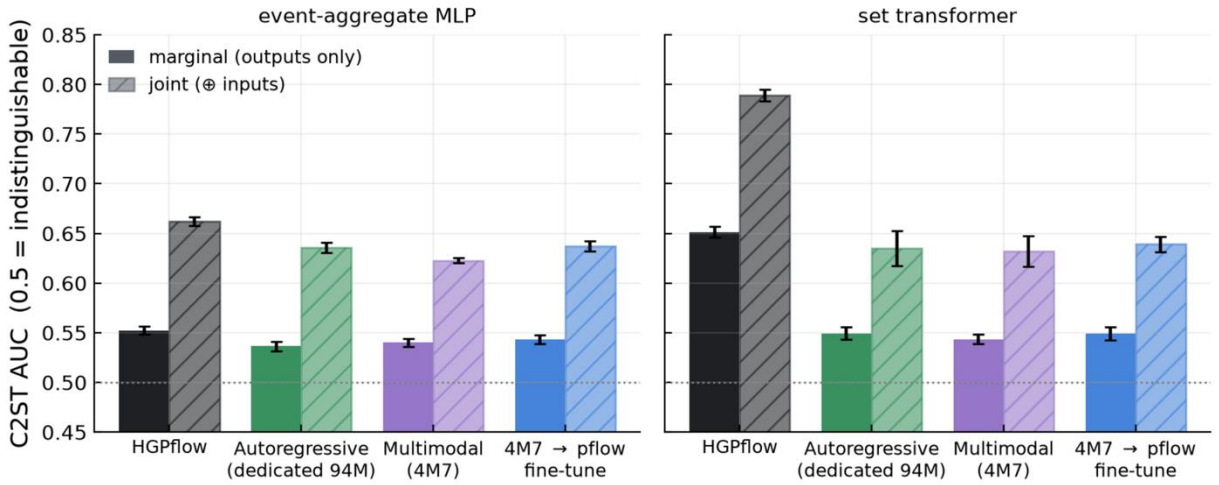
GENERALIST TRAINING & EXAMPLES

$7\ modalities \rightarrow sample \leq 3\ input\ and \leq 3\ output \rightarrow 1302\ distinct\ tasks$



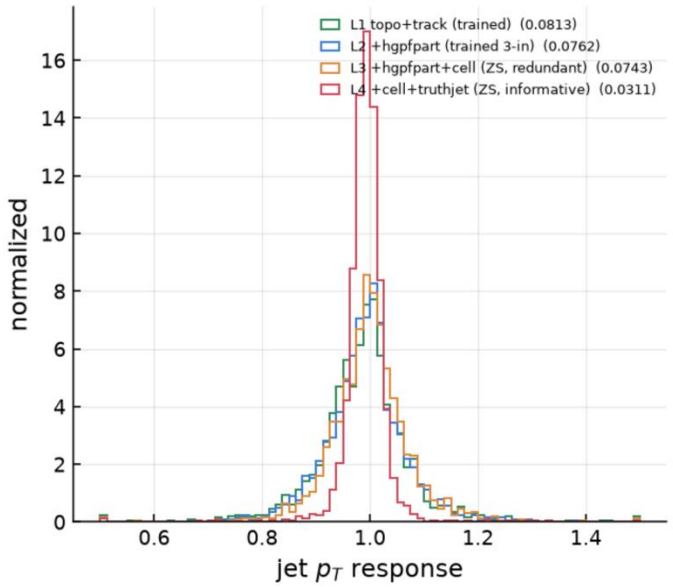
- Finally, show the main result: Train on all 1302 distinct tasks (7 choose up to 3 x remaining 4 choose up to 3)
- How does it do? Look again at the particle flow task (only one of the 1302 tasks) – competitive with HGPflow state-of-the-art

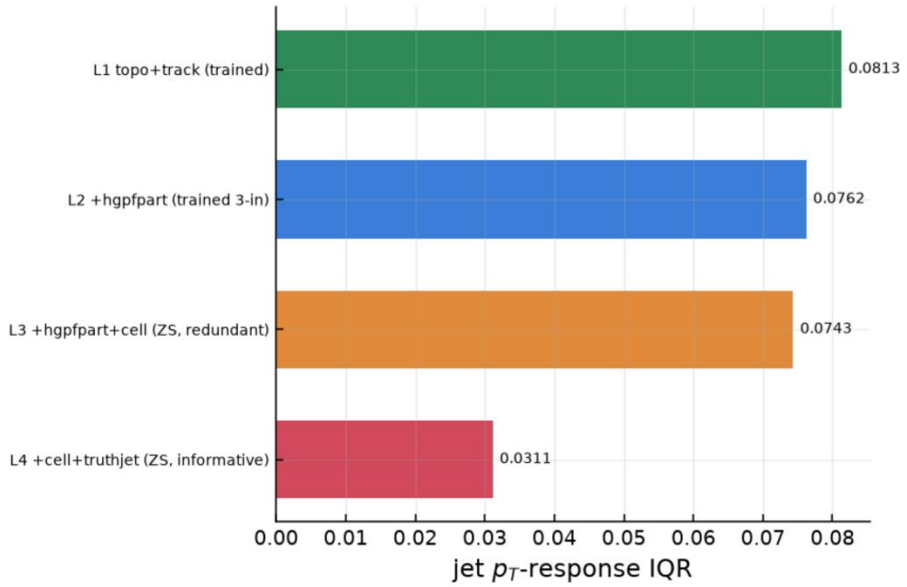




GENERALISABILITY
OUT-OF-DISTRIBUTION TASKS

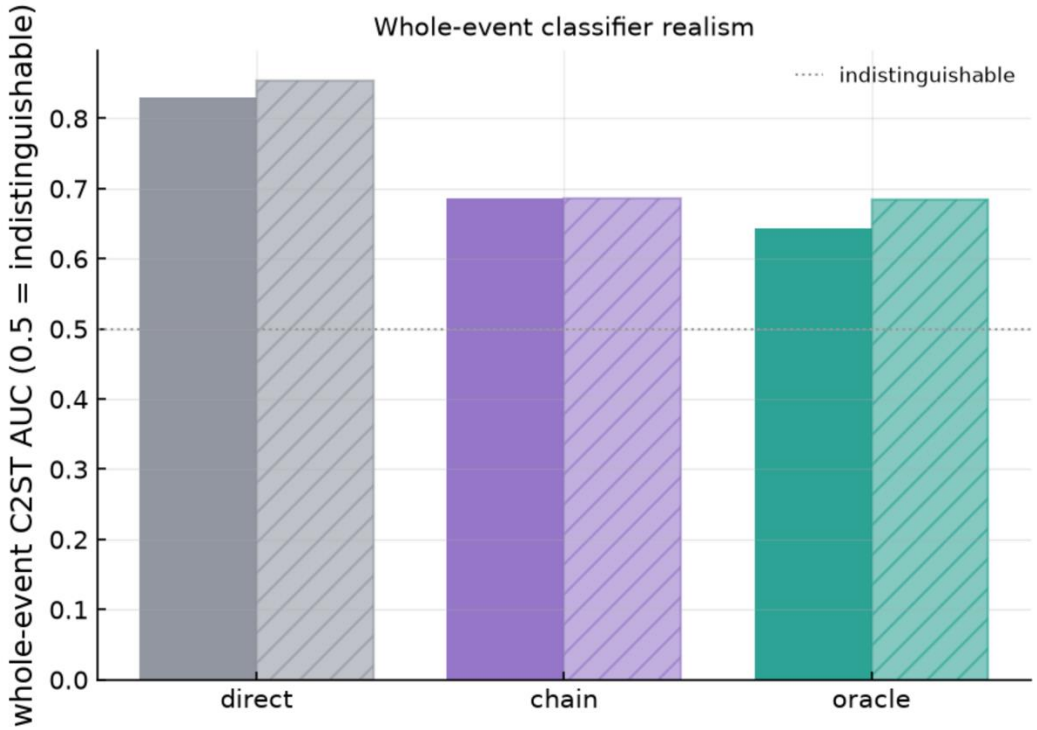
- The model had only ever seen up to three modalities as input. L1 is standard Pflow (2 input modalities) as before
- L2 gives the model HGPFlow particles as a hint – now it outperforms HGPFlow
- L3 gives a fourth modality, out-of-distribution, but it actually improves performance significantly





GENERALISABILITY
CHAIN-OF-PHYSICS

- Generating calorimeter showers is notoriously difficult
- The holy grail would be providing particles and receiving calorimeter cell energy deposits – nanoHEP is not very good at this: a classifier can tell its showers apart with an AUC of 0.8
- But ask nanoHEP to generate tracks first from particles, then tracks+particles → topoclusters, then tracks+particles+topoclusters → cells, and it significantly improves
- This is a kind of proto-chain of thought, motivated by physics heuristics



CONCLUSION

- Simple fill-in-the-blank allows us to define a generalist particle physics model
- HEP4M (parallel) is already competitive with state of the art pflow, and is very quick
- NanoHEP (autoregressive) scales very well with model and dataset size and outperforms state of the art
- However, it comes at a cost, which needs to be given some thought...