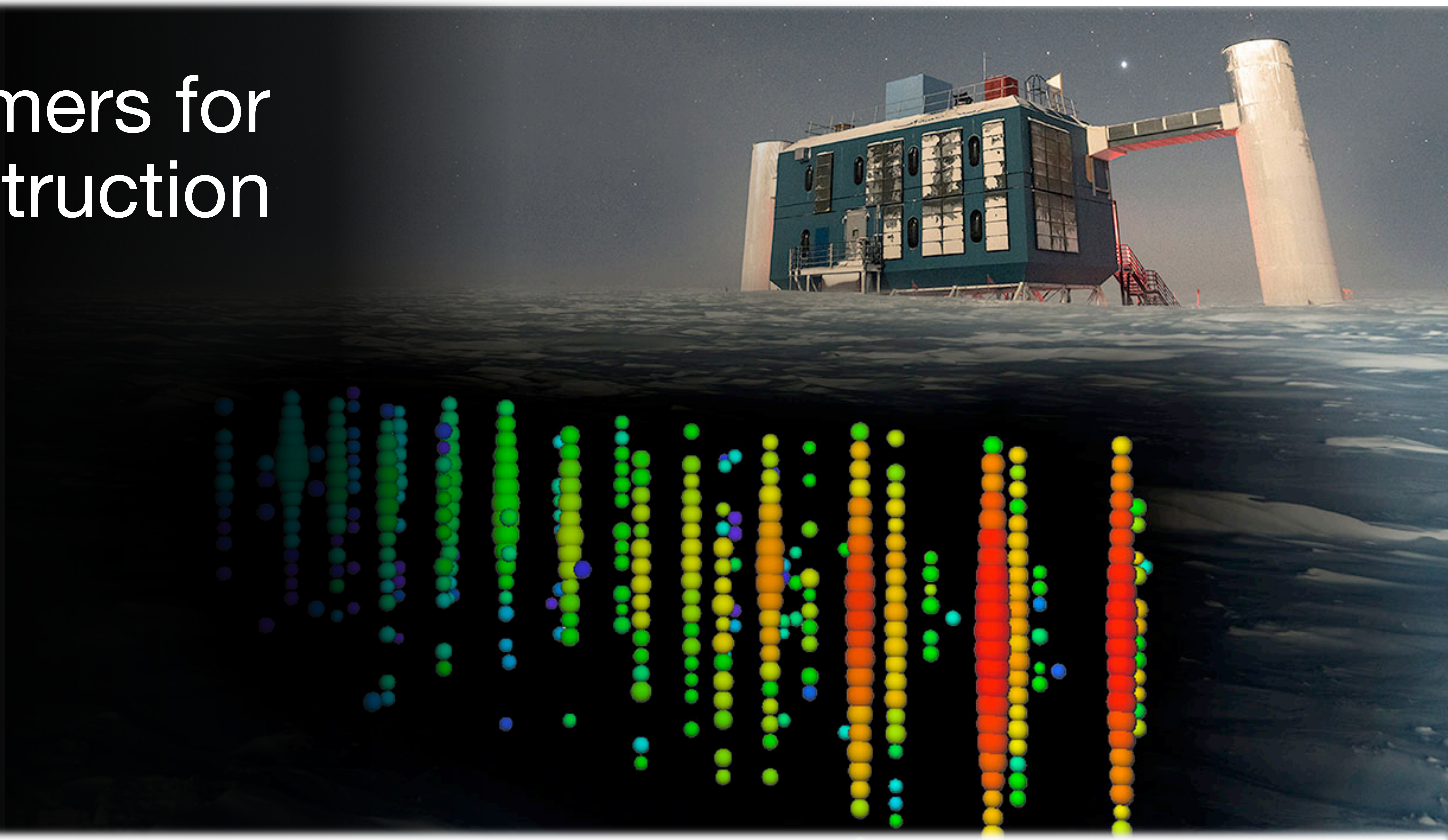
Using transformers for angular reconstruction with IceCube

Luc Voorend

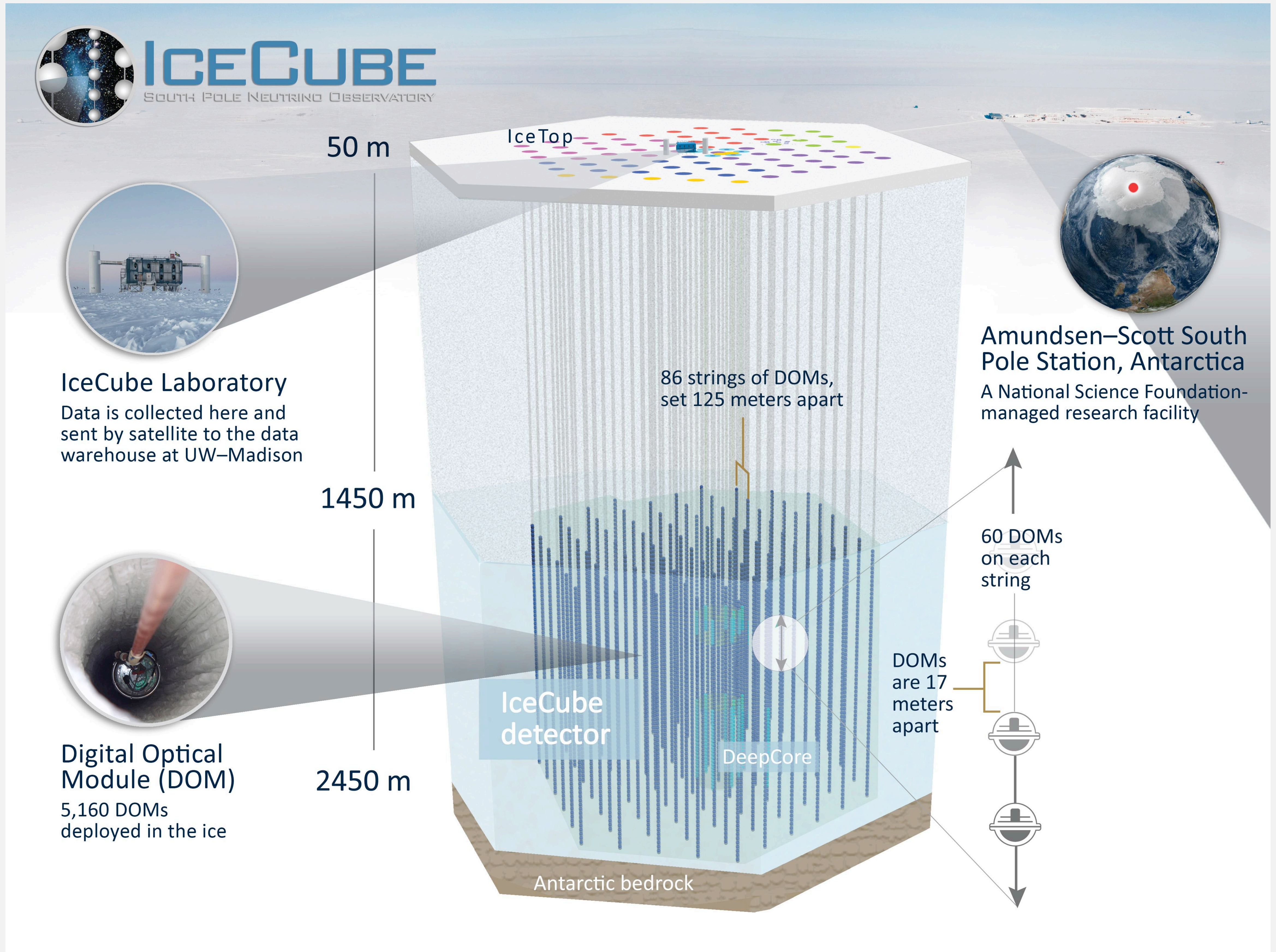


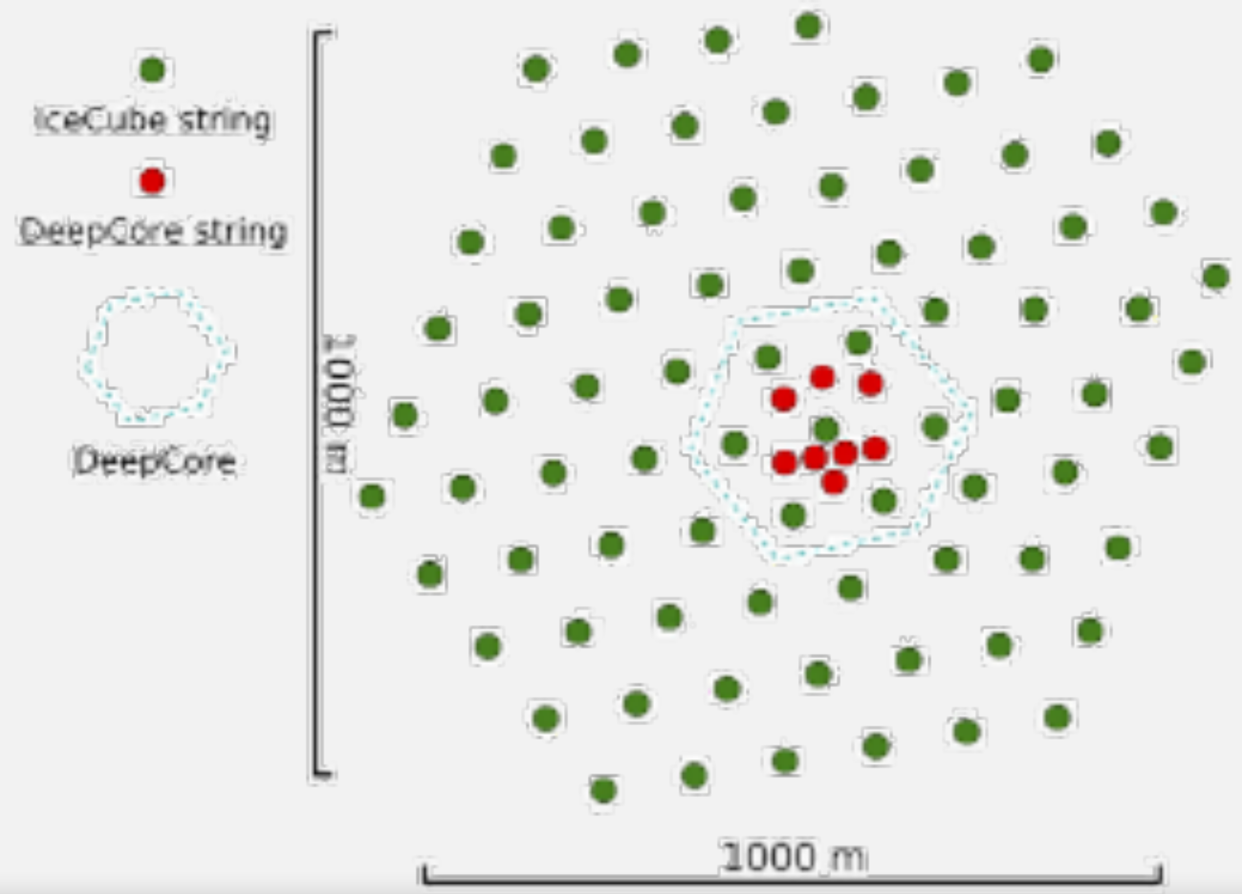
19-11-2025

Social outing talk

The IceCube observatory

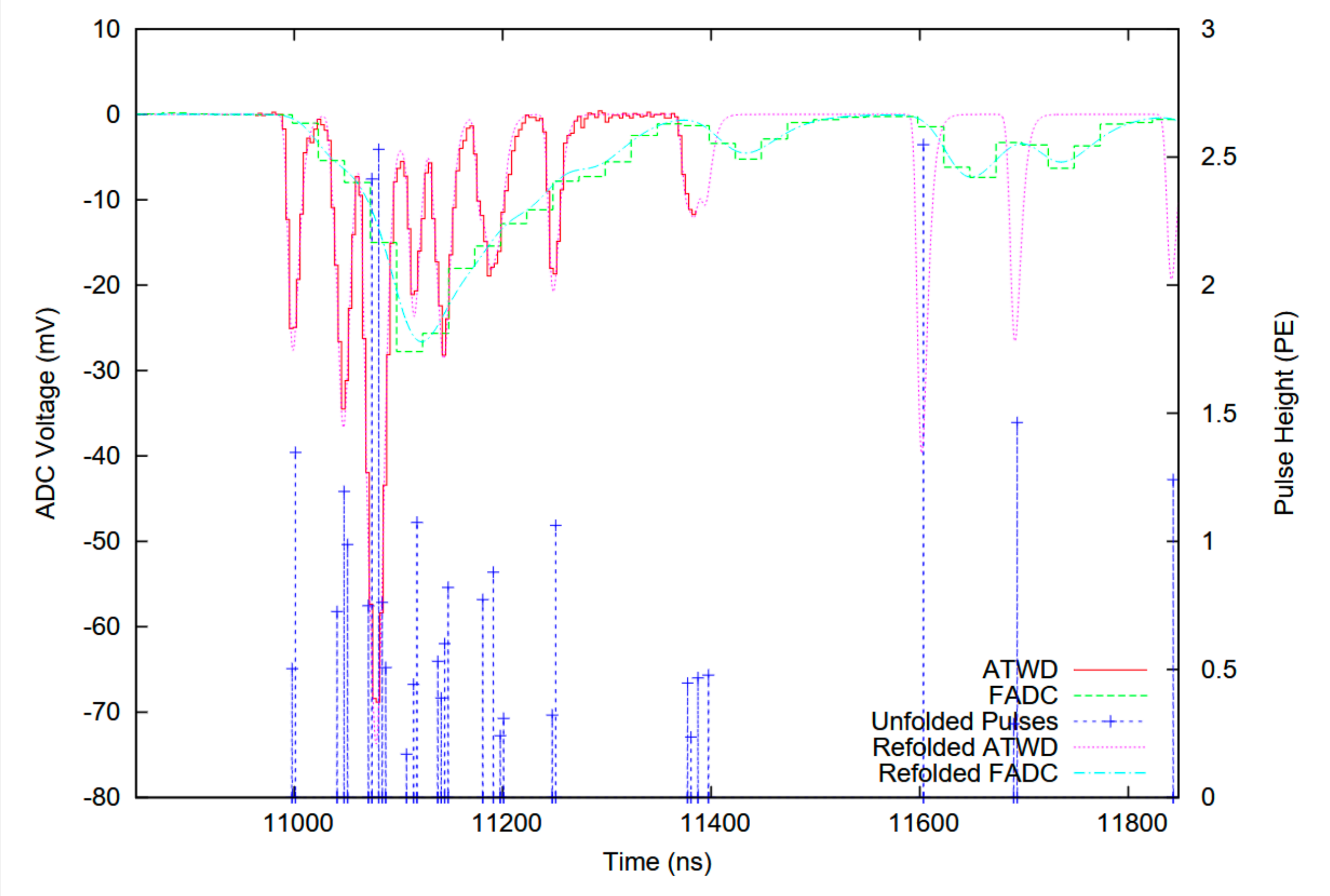
- Neutrino detector at the South Pole
- Completed in 2010
- 5160 DOMs on 86 vertical strings
  - Single PMT DOMs
- Dense low-energy region called DeepCore
- Surface array IceTop to veto cosmic ray air showers
- Installing IceCube Upgrade this winter with 7 additional strings inside DeepCore



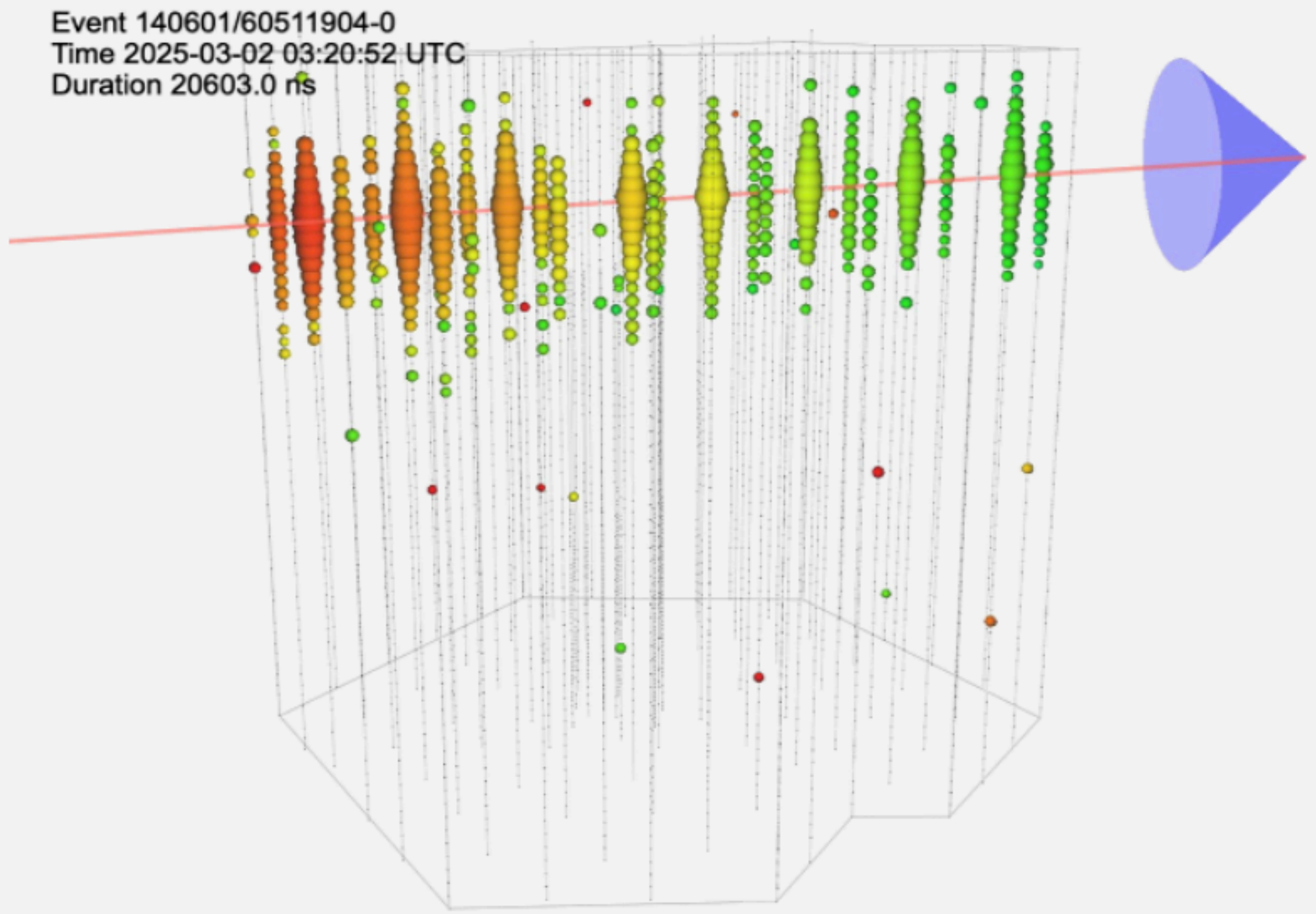


Neutrino events

- DOMs record waveforms and send them to the surface
- In-ice 'quality' check is done to check if a neighboring DOM also recorded some charge
  - HLC: full waveform send to surface
  - SLC: reduced waveform around peak send to surface
- If one of the event triggers is met, the surface lab combines **all** waveforms in a time window into a event
- First reduction: waveform unfolded into a pulse map



Reconstructing neutrino track direction



- Traditional reconstruction algorithms
  - Line-fit
  - Likelihood optimization of time residuals (SplineMPE)

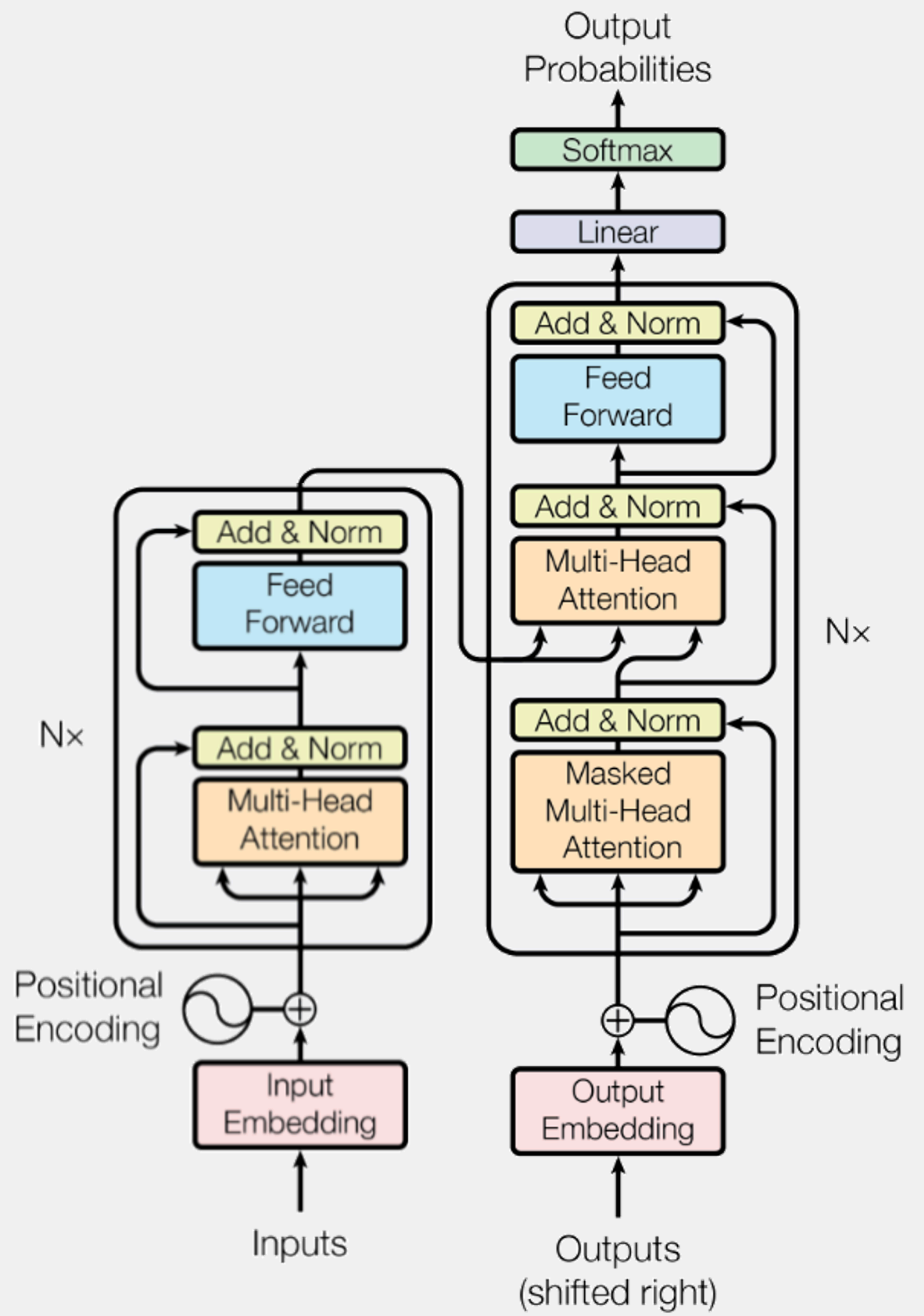
$$\mathcal{L}_{\mathrm{MPE}}(\vec{x}|\vec{\theta}) = \prod_{i}^{\text{1st hits}} n_i \cdot p(t_{\mathrm{res},i}|\vec{\theta}) \cdot (1 - \mathrm{P}(t_{\mathrm{res},i}|\vec{\theta}))^{n_i - 1}$$

$$P(t_{\mathrm{res}}|\vec{\theta}) = \int_{-\infty}^{t_{\mathrm{res}}} p(t|\vec{\theta})\mathrm{d}t$$

- Machine learning reconstruction algorithms
  - GraphNeT
  - Transformers

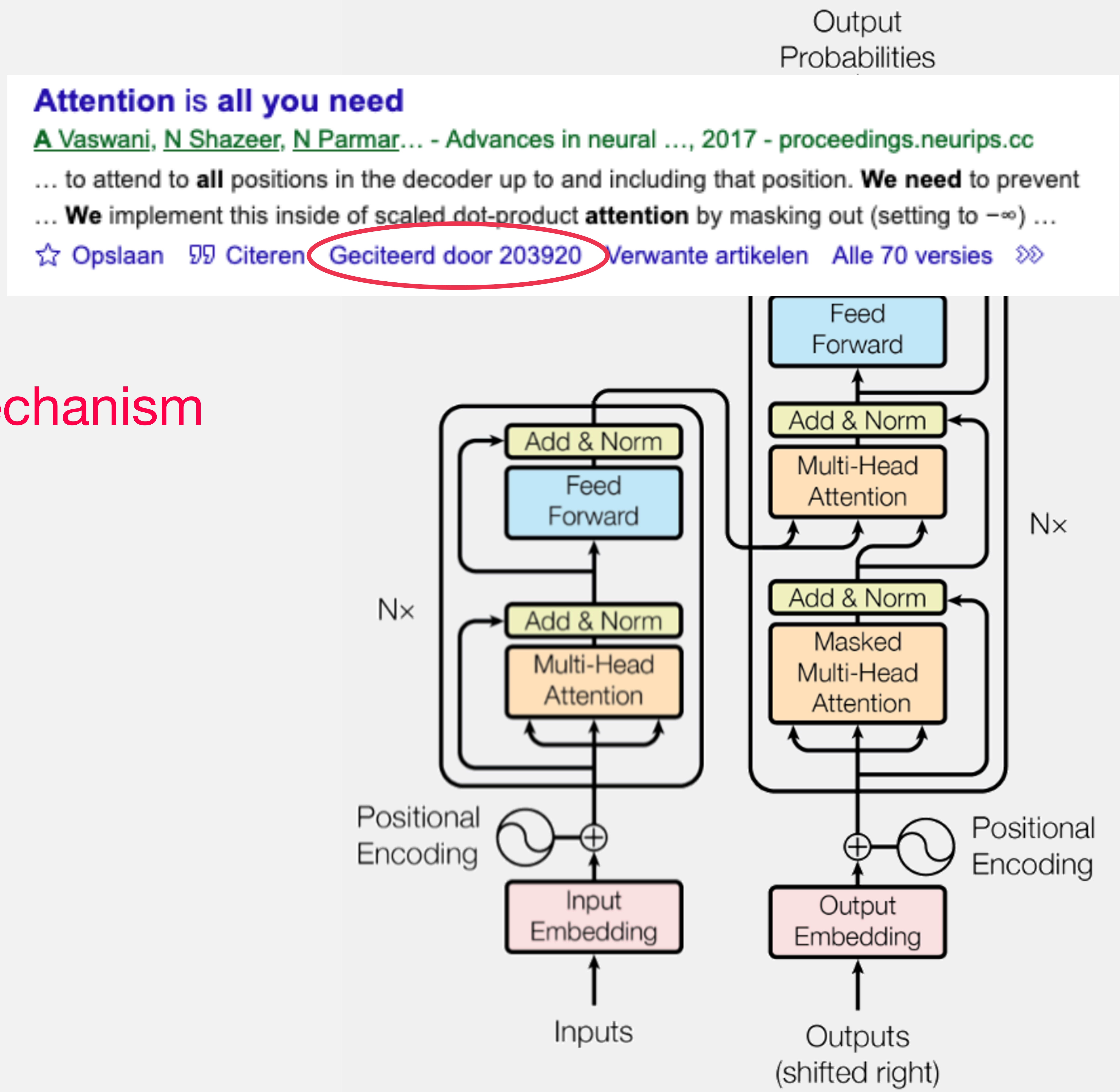
Transformer

- Designed for machine translation
- Takes a sequence as input
- Learning based on the attention mechanism
- Highly parallelizable



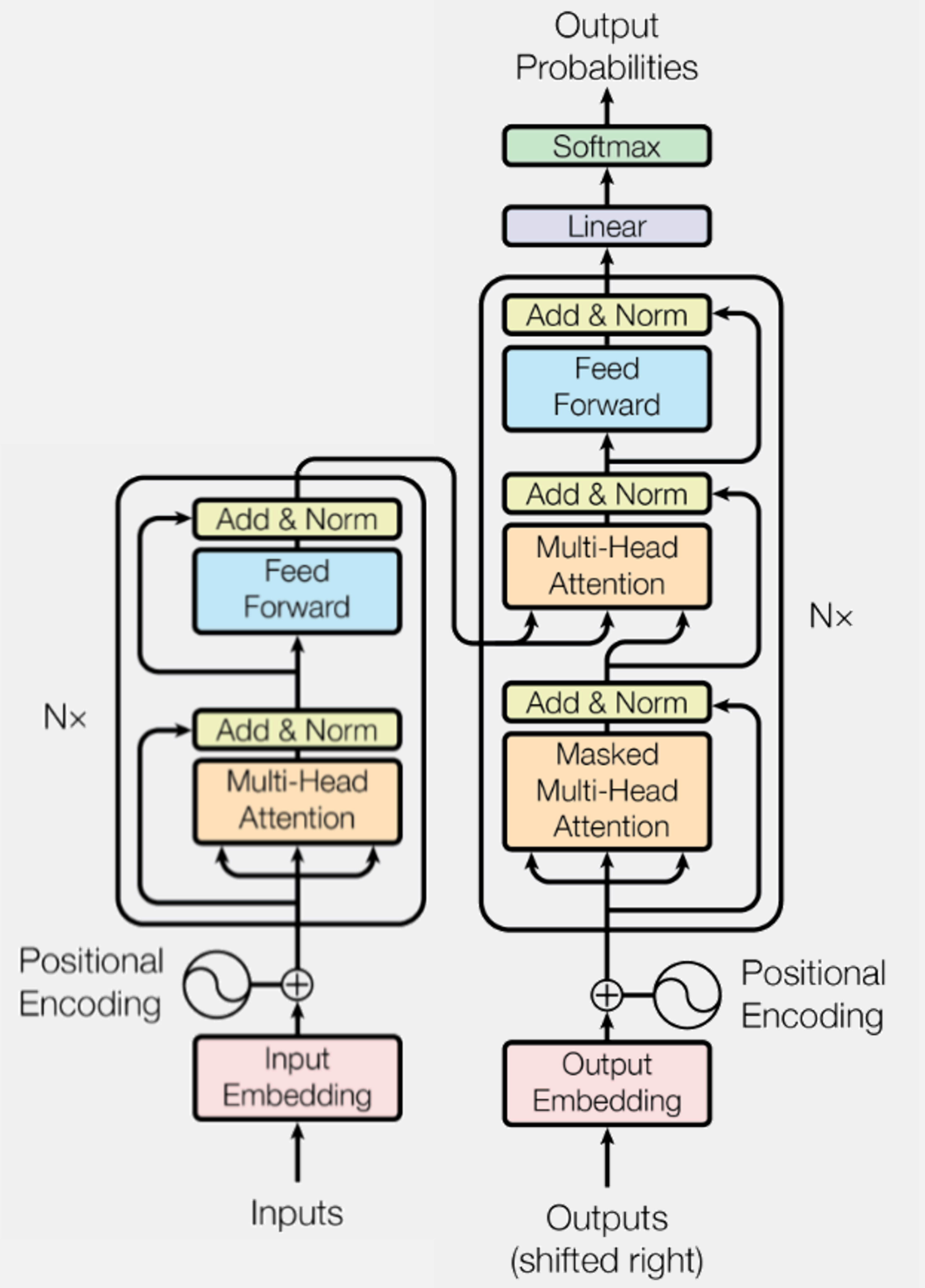
Transformer

- Designed for machine translation
- Takes a sequence as input
- Learning based on the attention mechanism
- Highly parallelizable



Transformer

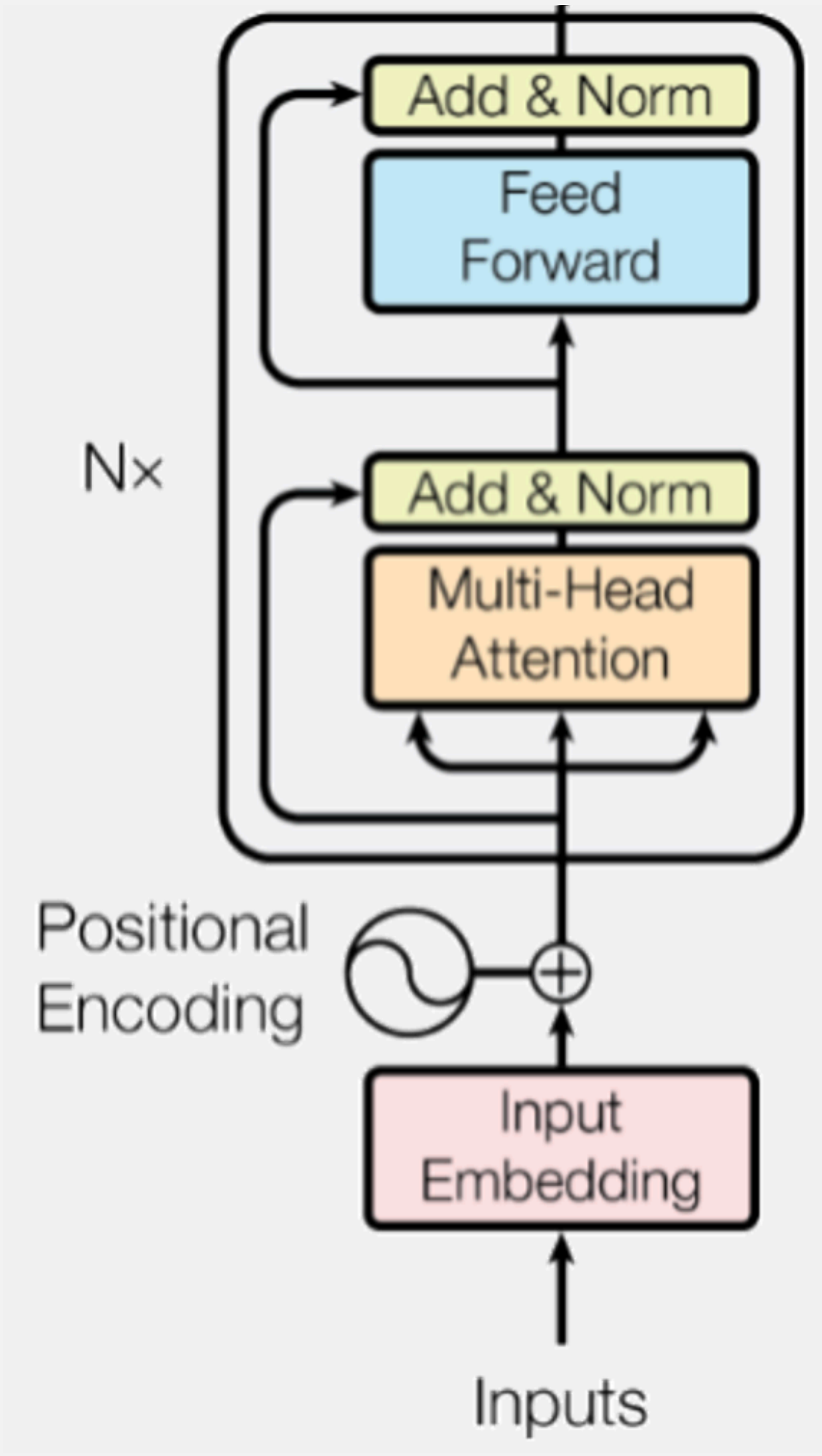
- Designed for machine translation
- Takes a sequence as input
- Learning based on the attention mechanism
- Highly parallelizable



Transformer

- Designed for machine translation
- Takes a sequence as input
- Learning based on the attention mechanism
- Highly parallelizable

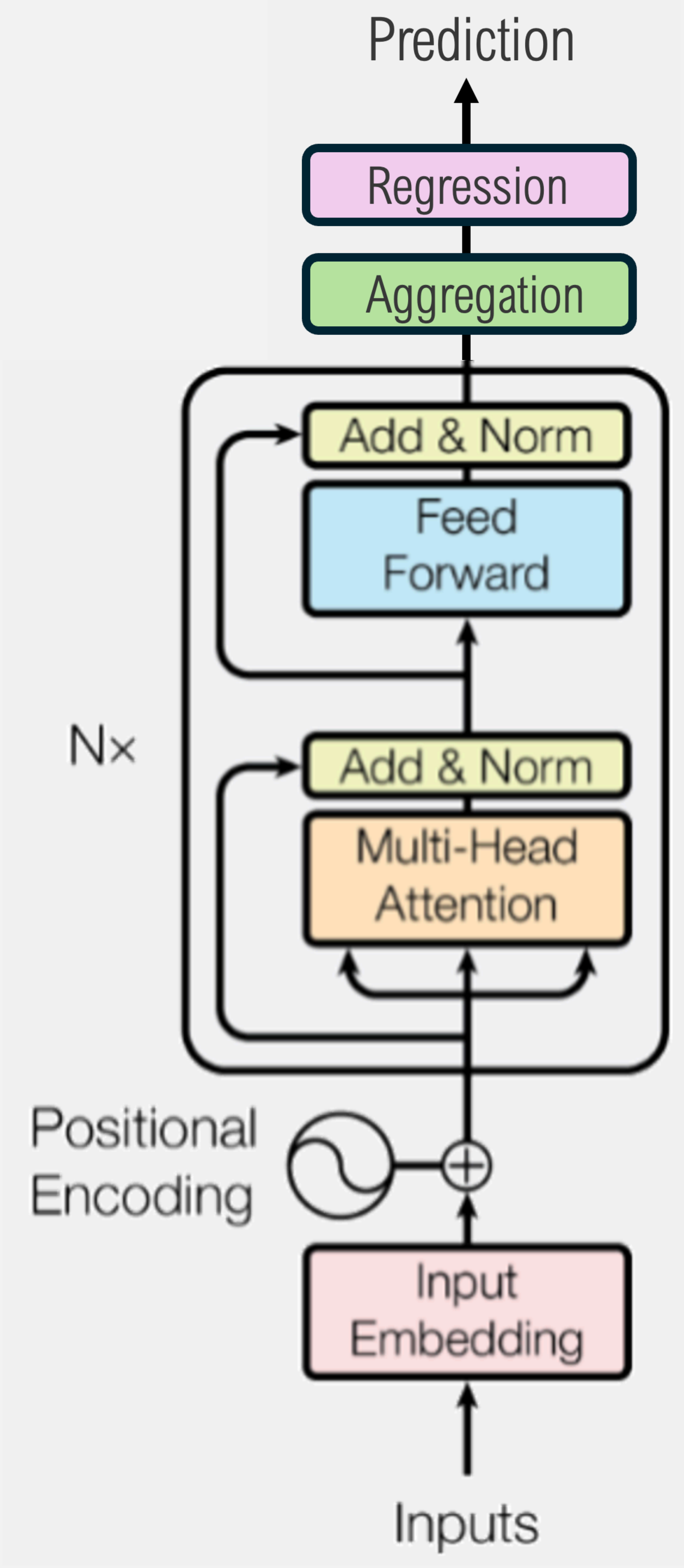
- Only interested in the encoder



Transformer

- Designed for machine translation
- Takes a sequence as input
- Learning based on the attention mechanism
- Highly parallelizable

- Only interested in the encoder
  - Add aggregation and regression layer



PMT-fication

- Training on ~4 million simulated neutrino events with energy between 100 GeV and 100 PeV
- Problem: highest energy events can have over 100k pulses
  - Way too long sequence to be computationally feasible for a transformer

PMT-fication

- Training on ~4 million simulated neutrino events with energy between 100 GeV and 100 PeV
- Problem: highest energy events can have over 100k pulses
  - Way too long sequence to be computationally feasible for a transformer
- Solution: PMT-fication
  - Summarize pulses per PMT (DOM) using summary features

PMT-fication

- Using 32 summary features per PMT
  - Reducing input to sequences of at most length 5160

$$\mathbf{PMT}_i = (\underbrace{x, y, z, x\_rel, y\_rel, z\_rel}_{\text{Dom position } (6)}, \underbrace{t1, q1, hlc1, \ldots, t5, q5, hlc5}_{\text{First 5 pulses } (15)}, \underbrace{Q25, Q75, Qtot, T10, T50, Tstd}_{\text{Quantiles } (6)}, \underbrace{rde, sat, \ldots}_{\text{Quality flags } (5)})$$

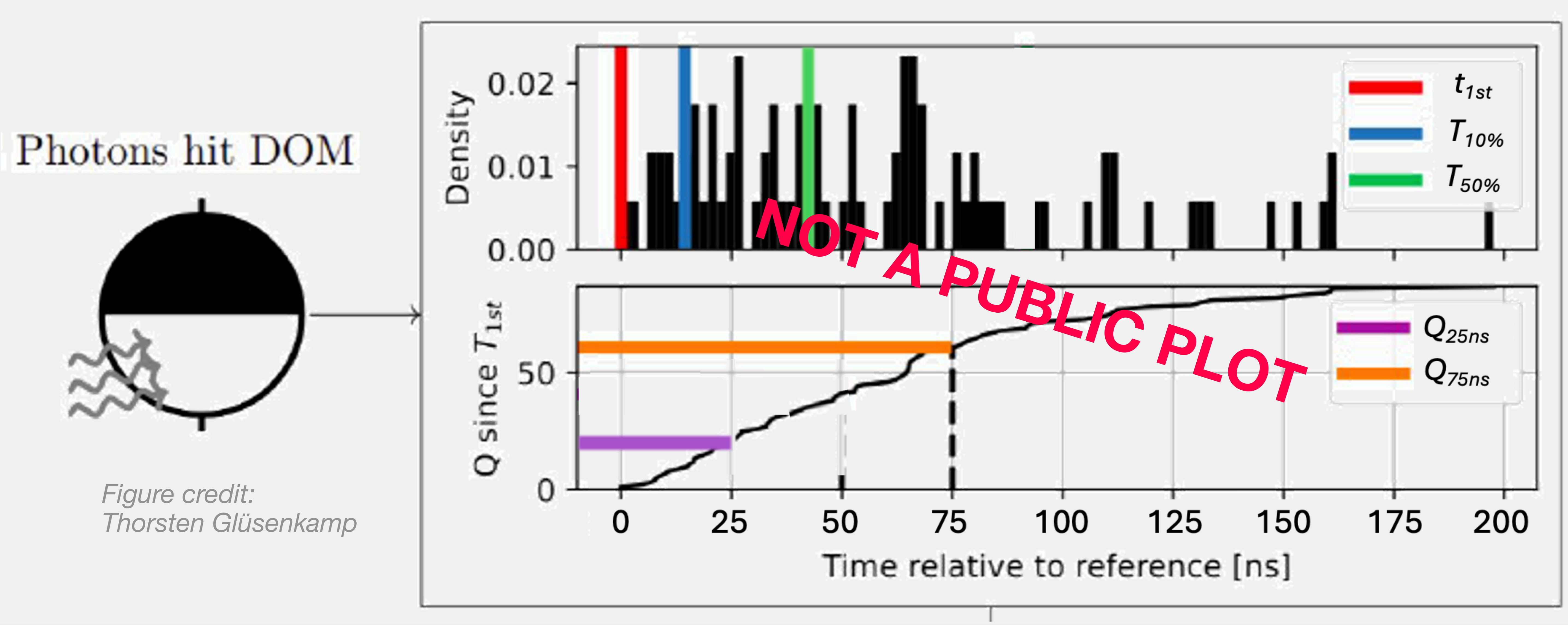


Figure credit: Thorsten Glüsenkamp

PMT-fication

IcePack

- Using 32 summary features per PMT
  - Reducing input to sequences of at most length 5160

PMT_i = (x, y, z, x_rel, y_rel, z_rel, t1, q1, hlc1, ..., t5, q5, hlc5, Q25, Q75, Qtot, T10, T50, Tstd, rde, sat, ...)

Dom position *(6)* | First 5 pulses *(15)* | Quantiles *(6)* | Quality flags *(5)*

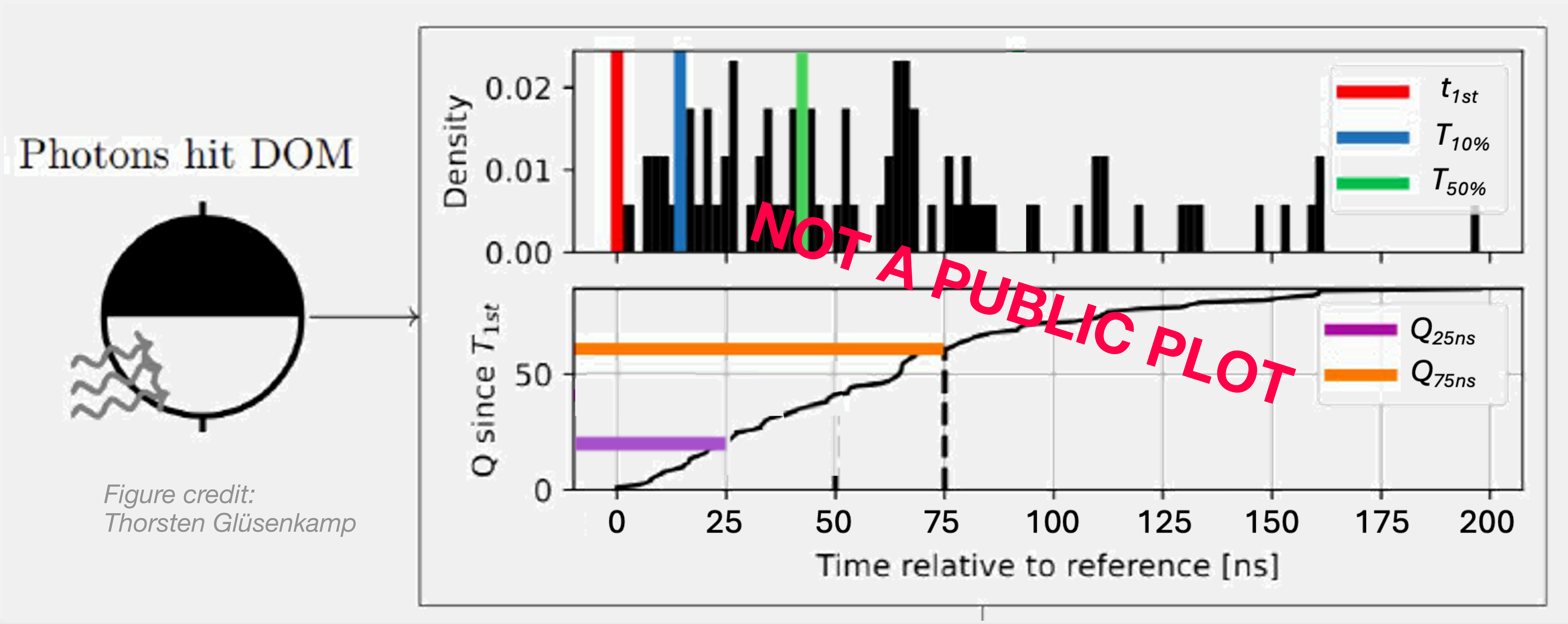


Figure credit: Thorsten Glüsenkamp

Training

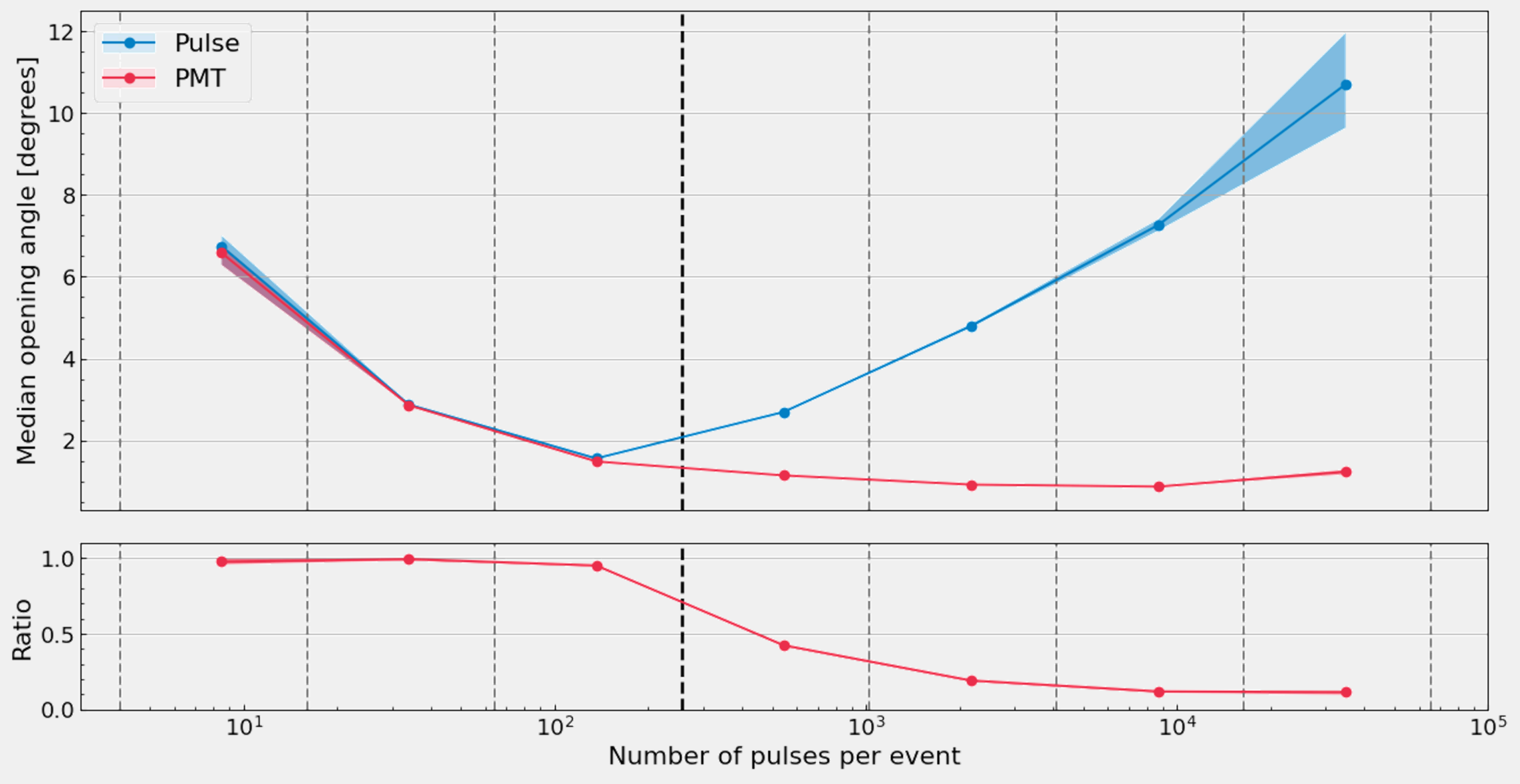
- Most training done on single NVIDIA GeForce RTX3090 GPU
- Trained small scale models for development
    - Limited model size and training set size
    - Up to 2 days of training time
- Large scale final model
    - 27.0 million trainable parameters
    - 3.9 million training events
    - ~3 weeks of training
    - Inference at ~1000 events/s

A selection of results

Effectiveness of PMT-fication as a reduction method:

- Opening angle of pulse models explode for events with more than 256 pulses
- Reconstruction of PMT-fied model keep improving for high-pulse events
- No “loss of information” for low-pulse events

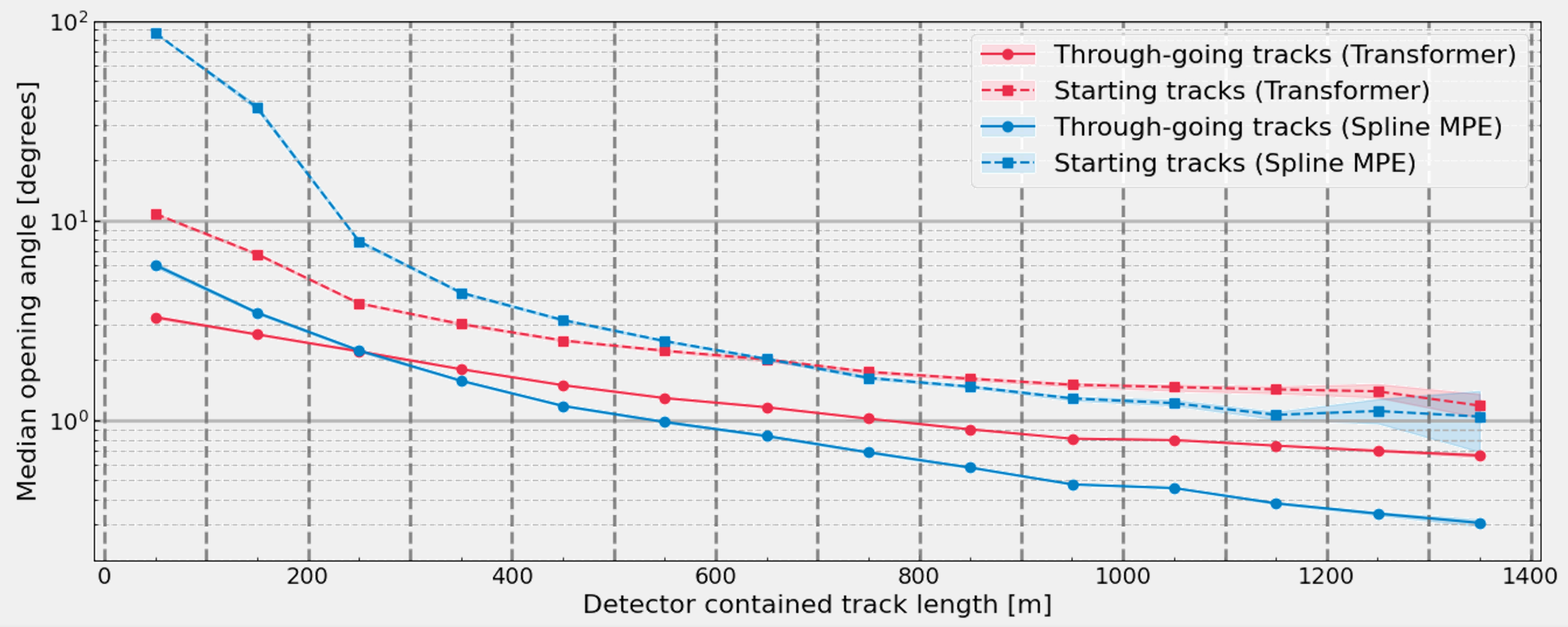
Plot showing results in the energy range 10 TeV - 1PeV



A selection of results

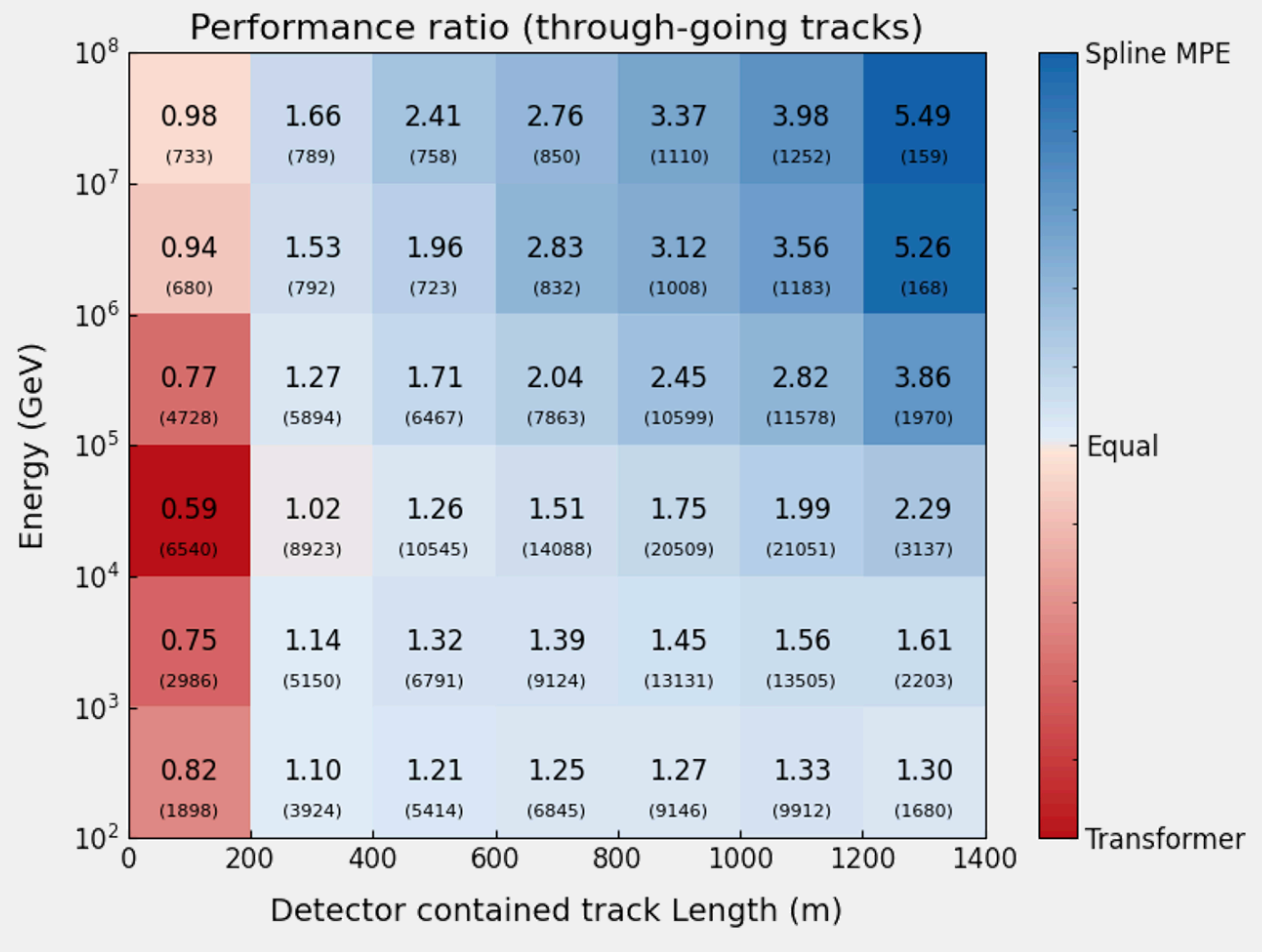
Comparing performance to SplineMPE:

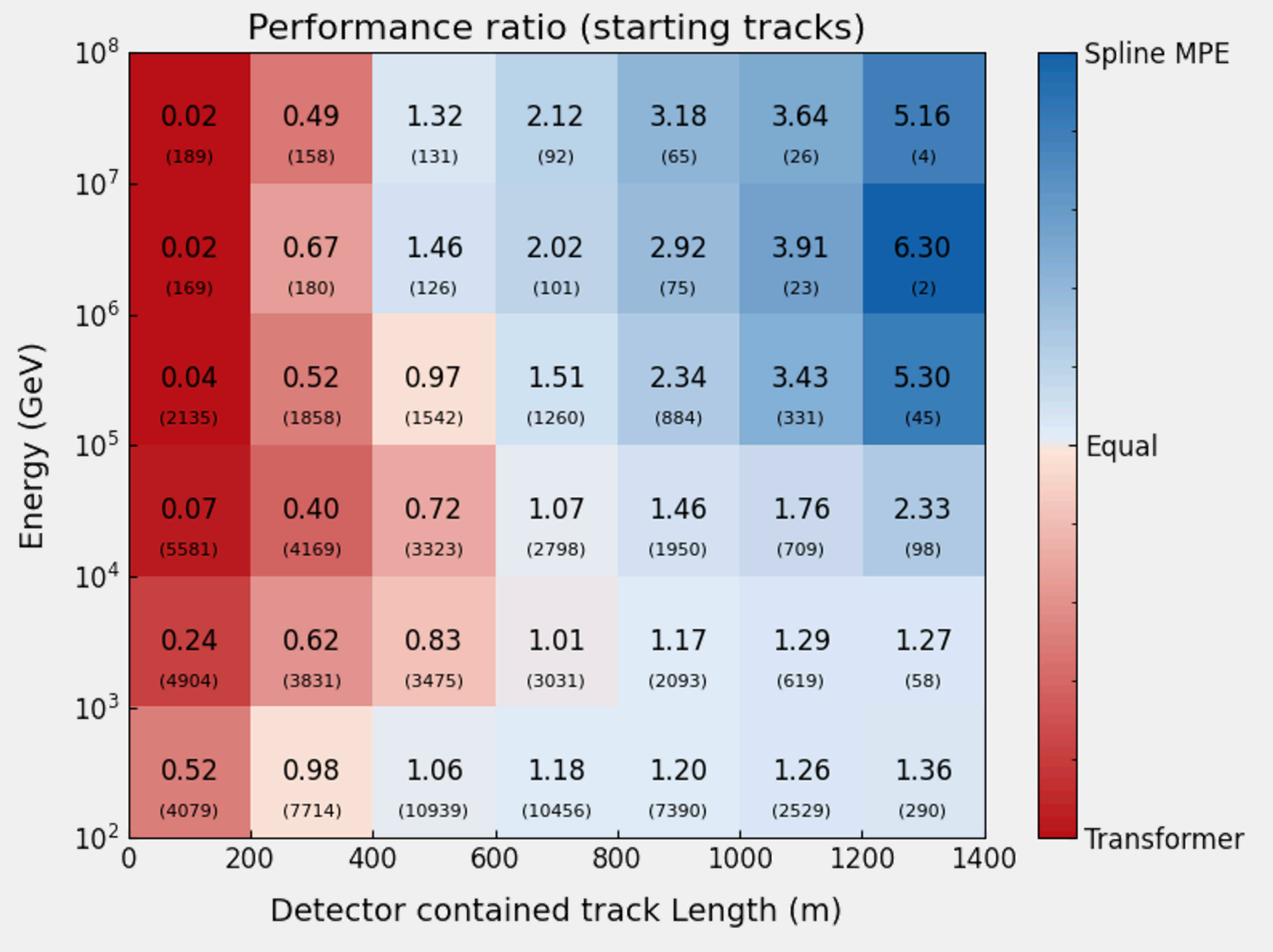
- Transformer not outperforming SplineMPE for through-going tracks
- Transformer does improve for starting tracks with track length < 700m
    - ~71% of all starting tracks



A selection of results

$$\text{Performance ratio} = \frac{\text{Median opening angle transformer}}{\text{Median opening angle SplineMPE}}$$





(N) = Number of events in the bin

Interesting findings

- High-energy neutrino events can be made suitable for transformers
- The transformer can reconstruct events fast
- The transformer can learn detector/event geometries without explicitly defining these
- The transformer generalizes better in ranges where assumptions of statistical models break
- More training = Better performance
  - Unpublished results show a transformer architecture matching/outperforming SplineMPE for all track types for all energies*

*using normalizing flows for likelihood free posterior prediction