

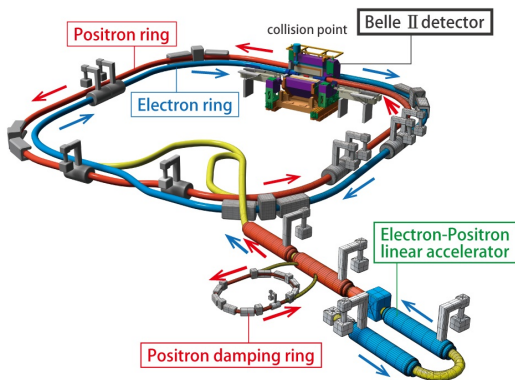
Transforming Flavor tagging at Belle II

Lukas Herberg

HEP Herbstschule 2026

The SuperKEKB accelerator

- $e^+ e^-$ collider running at centre-of-momentum energy close to the mass of the $\Upsilon(4S)$ ($b\bar{b}$) resonance
- Asymmetrical beam energies: high-energy ring running at 7 GeV with e^- , low-energy ring running at 4 GeV with e^+
- Goal of producing $\Upsilon(4S)$ at high luminosity which then decay to either $B^+ B^-$ or $B^0 \bar{B}^0$ pairs.

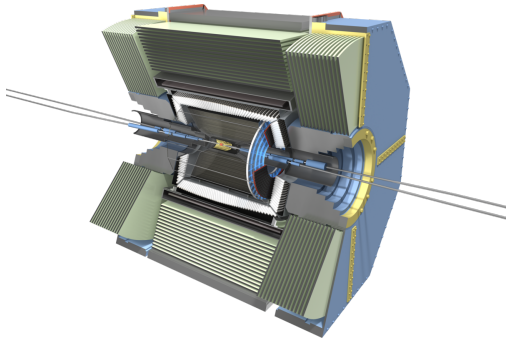


The Belle II Detector

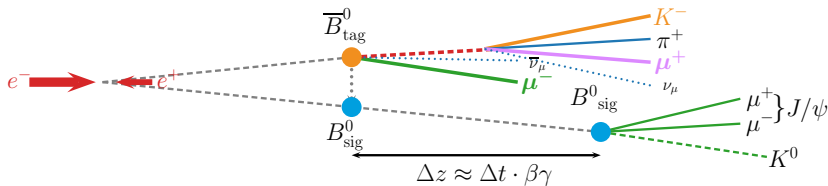
Belle II is an asymmetric detector located at the beam pipe crossing of SuperKEKB

Important characteristics of Belle II for this talk:

- Tracking: Reconstruction of the 4-momenta of charged particles in the detector
- Particle identification (ID): Determination of particle type
- Electromagnetic calorimeter: Measurement of Photons and Neutral particles
- Vertexing: Precise measurement of B meson decay position (which gives decay time difference)



Flavor tagging



B^0 are produced in entangled state at SuperKEKB.

$\Rightarrow$ The moment one of them decays, the other **must** have the opposite flavor.

Measurements of CP violating processes can take advantage of this:

- Infer the flavour of B_{tag} from its decay products
- From this, we know initial flavour of B_{sig}

$$\text{Asym} = \frac{N(B_{\text{tag}}^0) - N(\bar{B}_{\text{tag}}^0)}{N(B_{\text{tag}}^0) + N(\bar{B}_{\text{tag}}^0)} = (1 - 2w) \sin(\Delta m_d \Delta t) \sin(2\beta)$$

where Δt is the decay time difference and w is the fraction of wrongly tagged B_{tag} .

Traditional BDT approach to flavour tagging

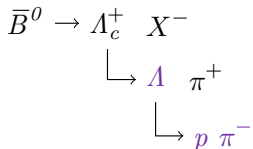
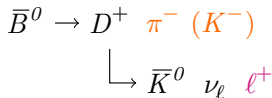
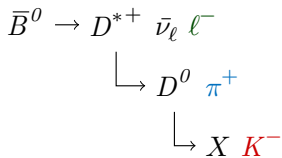
Flavour tagger ([paper](#)) inherited from Belle works in two steps:

① **13 category BDTs**: classify each track within 13 tagging signatures, mainly:

- Fast lepton from B
- intermediate kaon from D
- intermediate lepton from D
- fast hadron from B
- slow pion from D^{*+}

② **Combiner BDT**

- Inputs: The highest output value of each category
- Outputs: $q = \pm 1$ for B^0 and $\bar{B}^0$, and $r \in [0, 1]$ for the confidence

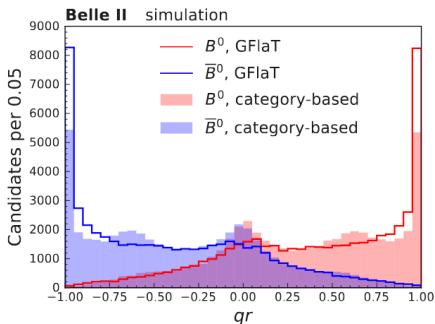


BDTs use low level (PID, momentum) and high level information (aggregate of cluster variables)

GFlat: Graph Neural Network Flavour Tagger

GNN ([paper](#)) still uses the 13 category BDTs, and builds a “particle cloud”, or graph with all charged tracks from the tag B decay:

- Nodes: 1 for each charged track
 - 13 category BDT output values
 - 3-momentum
- Edges: distance between the tracks
- Output: again q & r



Significant gain in efficiency from retaining the information of how the category output is shared among all tracks.

Metric: effective tagging efficiency $\varepsilon_{eff} = \sum_i \frac{n_i}{n_{tot}} (1 - 2w_i)^2$

$$\sigma_{stat}(\text{CPV}) \sim 1/\sqrt{\varepsilon_{eff}}$$

- Category-based: 31.1% (data)
- GFlat: 37.4% (data), 37.9% (MC)

TFlat: Transformer based flavour tagger

GFlat's two-step approach has two shortcomings:

- Technically difficult to re-train/maintain due to 13 BDT's + 1 GNN
- Require engineering of the tagging categories and input features

Additionally: Latest developments in machine learning may improve performance

TFlat: Transformer based flavour tagger

GFlat's two-step approach has two shortcomings:

- Technically difficult to re-train/maintain due to 13 BDT's + 1 GNN
- Require engineering of the tagging categories and input features

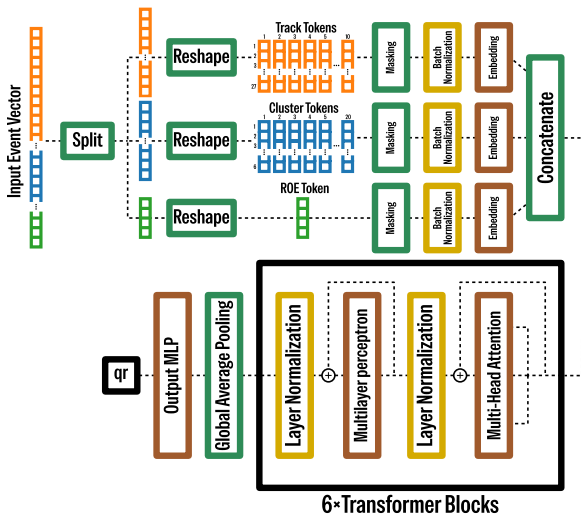
Additionally: Latest developments in machine learning may improve performance

⇒ New algorithm **TFlat**:

- Same direction as GFlat, but take it a step further and go for full deep learning
- 1-step approach using low-level information and attention mechanism (transformer) to let the algorithm focus on what is important
- Use information about neutral clusters (photons)

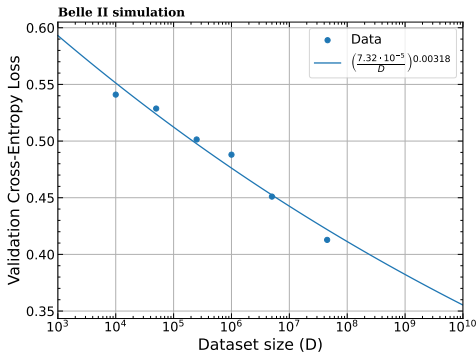
TFlat: Architecture

- Input data is split into tokens; 1 per track/cluster
- Tokens are embedded into 128-dim latent space
- Tokens then pass through a 'stock' transformer architecture, consisting of alternating Attention and MLP layers
- Finally all tokens are averaged and projected down into a single number $qr \in (-1, 1)$



Transformer scaling with more data

- For training, we use MC samples where one side decays $B^0 \rightarrow \nu \bar{\nu}$ (invisible) and let the other B^0 decay generically
- Transformers are known to scale well with training dataset size
- Performance scaling of TFlaT follows established transformer scaling laws
- Final training of TFlaT was performed on 50M events



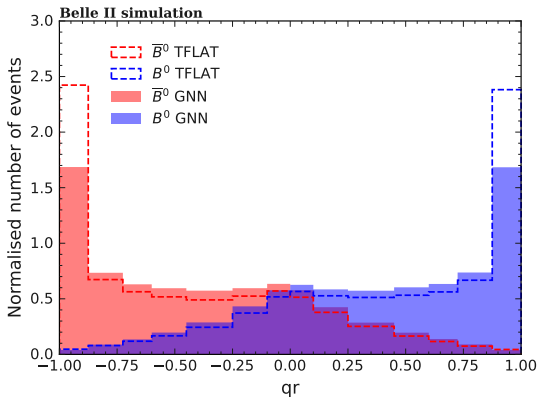
Reminder: $\varepsilon_{eff} = \sum_i \frac{n_i}{n_{tot}} (1 - 2w_i)^2$

- **GFlaT:** $\varepsilon_{eff} = 37.9\% \pm 0.2\%$

- **TFlaT:** $\varepsilon_{eff} = 45.7\% \pm 0.2\%$

TFlaT shows substantial improvement relative to previous methods

Much larger fraction of events is tagged with high certainty



- Reconstruct self-tagging

$B^0 \rightarrow D^{(*)-} h^+$ decays, where

$h^+ \in \{K^+, \pi^+\}$

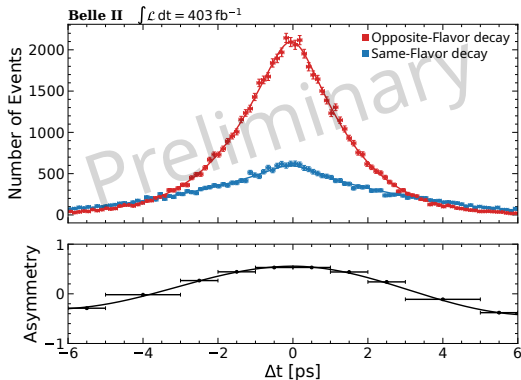
- Reconstruction gives us flavor of B_{sig}^0 which is then used to calibrate FT response

- Extract wrong tag fractions by fitting decay time distribution, which then gives ε_{eff}

- FBDT**: $\varepsilon_{eff} = 31.1\% \pm 0.5\%$

- GFlaT**: $\varepsilon_{eff} = 37.4\% \pm 0.4\%$

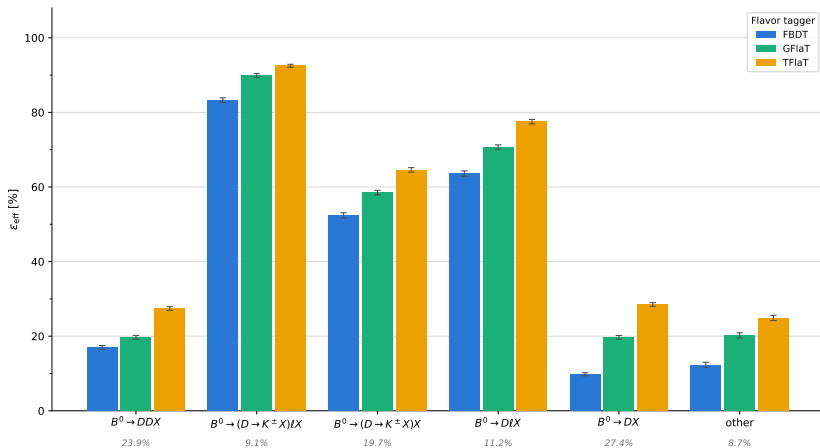
- TFlaT**: $\varepsilon_{eff} = 44.3\% \pm 0.4\%$



Comparison to previous methods

Relative improvement is most significant in decays without $e^\pm/\mu^\pm$ or $K^\pm$ that tag the flavor.

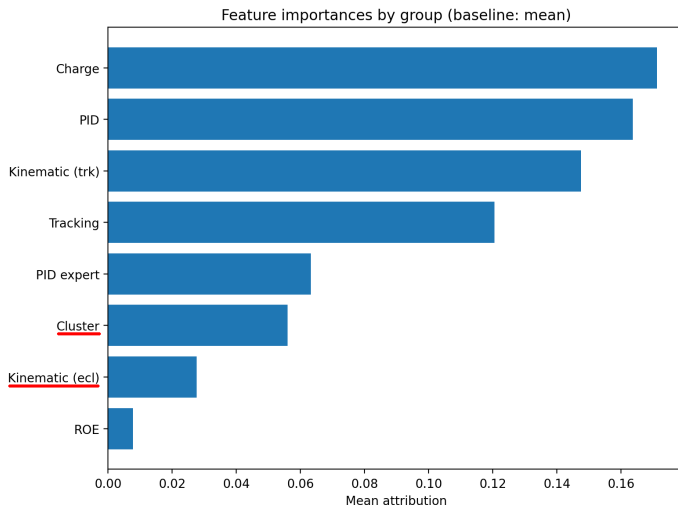
This was also the case for $\text{FBDT} \rightarrow \text{GFlaT}$



Feature attribution

Feature attribution using the integrated gradients method ([paper](#))

Top features are as expected; Attribution value of neutral features not insignificant



Conclusion

- Introduced new transformer based flavor tagging algorithm
- 1-step algorithm based only on low level information
- Gain relative 18% efficiency wrt GNN
~ 42% wrt Belle FBDT approach on real data
- TFlaT has also been trained for Belle with slightly worse improvements (due to worse simulation)

Tagger	ϵ_{eff}
FBDT	31.1%
GFlat	37.4%
TFlat tracks	44.3%

Charged track inputs:

- Charge, p^* , $\cos(\theta^*)$ and ϕ^*
- Particle ID
- Number of hits in the vertex detector
- Impact parameters (POCA of the B_{tag}^0 vertex)
- Cluster shapes of calorimeter cluster attached to the track

Neutral cluster inputs:

- p^* , $\cos(\theta^*)$ and ϕ^*
- Cluster shapes

Additional global event inputs:

- Number of photons/tracks
- Total transverse momentum