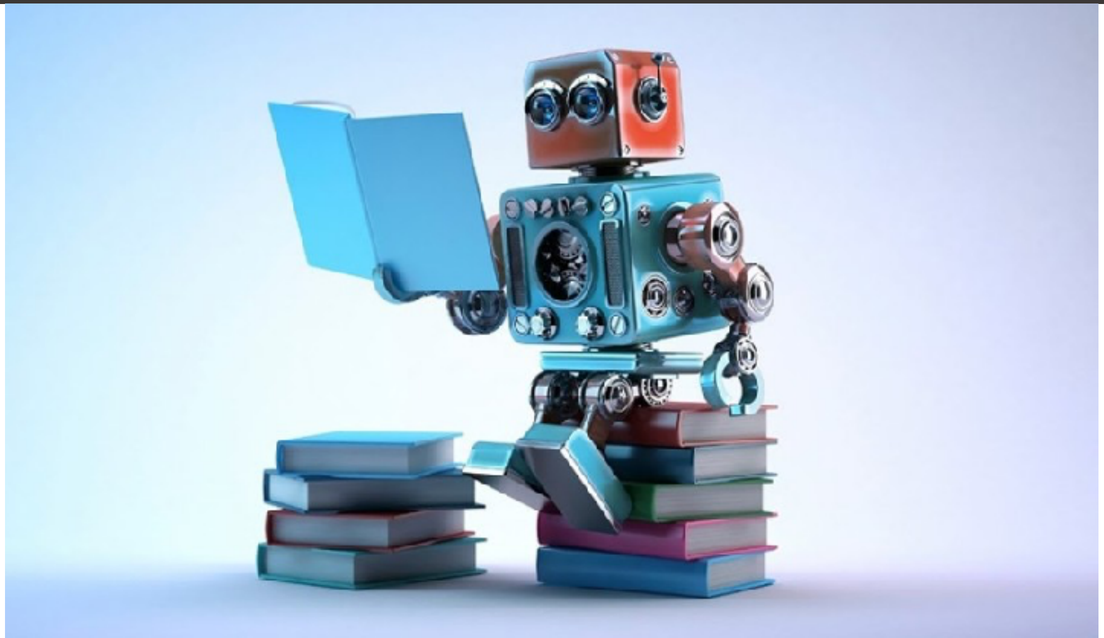
Machine Learning at the Belle II Experiment

ML at the LHC workshop

Simon Wehle
University of Nagoya
05.02.2020



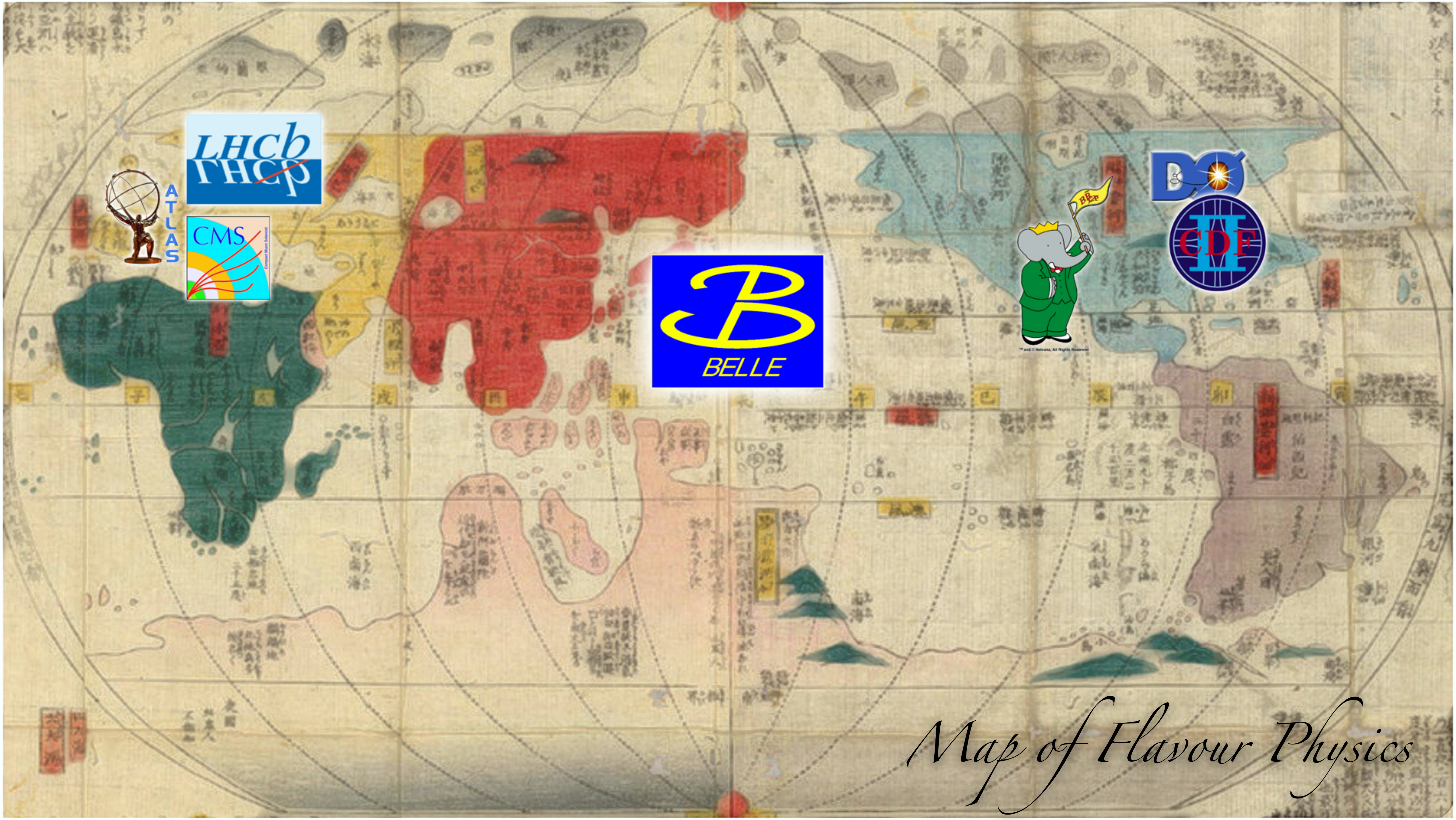
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Machine Learning Applications at Belle II

Contents

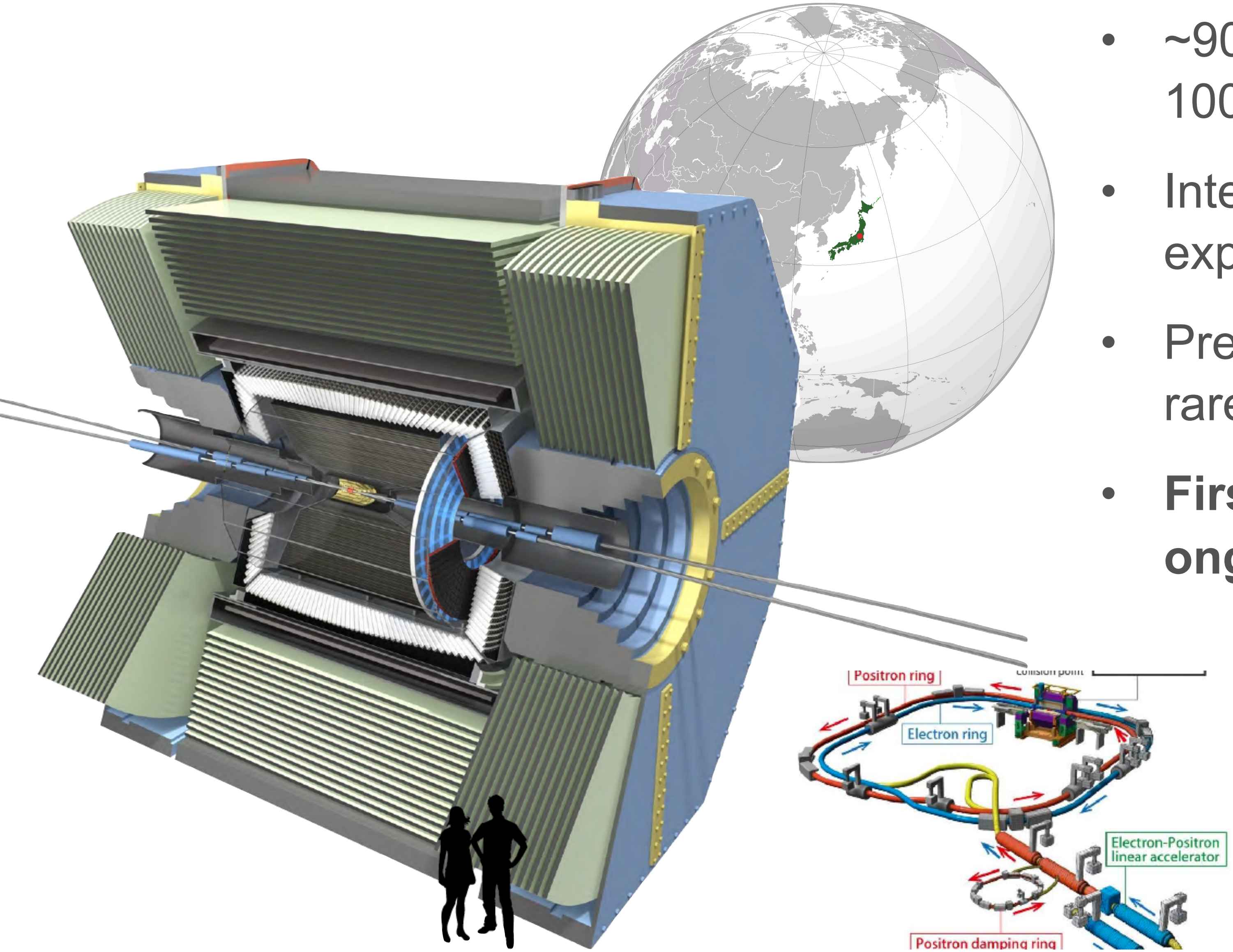
- Charge Particle Identification
- Clustering, Cluster position, Cluster direction
- Neutral hadron/photon separation
- Image calibration
- Full Event Interpretation
- Flavour Tagging
- *Disclaimer: Work from many collaborators is also presented*

Introduction To Belle II


LHCb
ATLAS
CMS
BELLE
CDF
Map of Flavour Physics

The Belle II Experiment

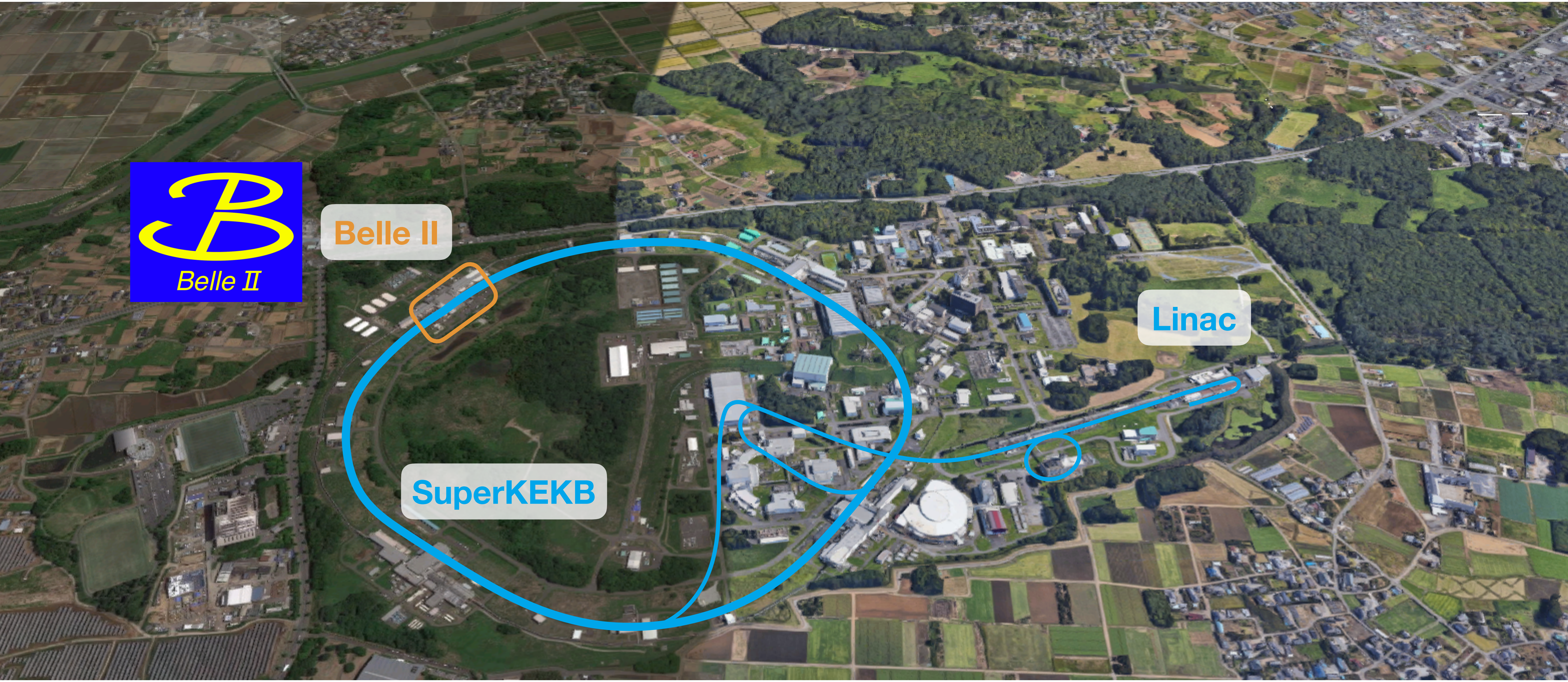
Pushing the intensity frontier to the next level



- ~900 researchers from 30 countries, with 100+ from Germany, ~50 from DESY
- Intensity frontier flagship “B-factory” experiment: **30kHz event rate**
- Precision physics and searches for (very) rare decays including Dark Matter
- **First data taken 2018, data taking ongoing**

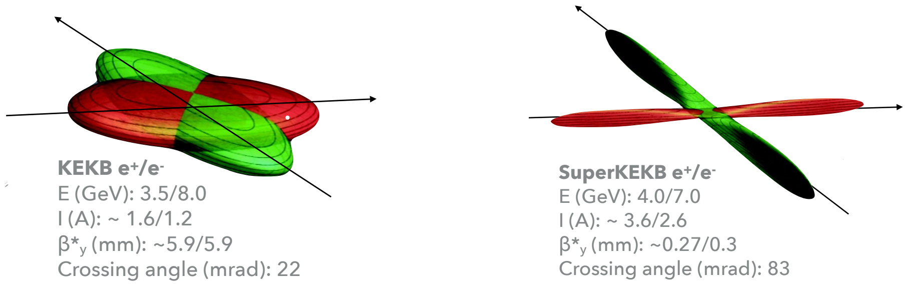
Belle II Experiment

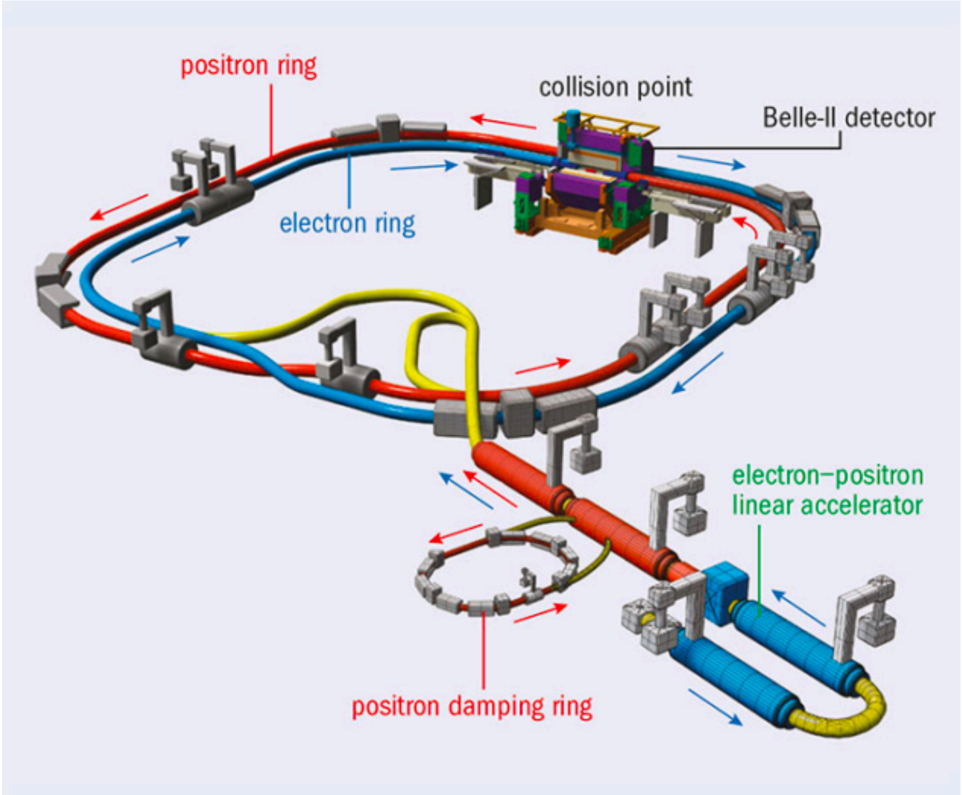
KEK, Tsukuba



SuperKEKB

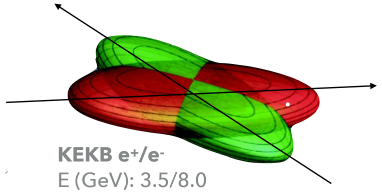
- Asymmetric (4.0 GeV/7.0 GeV) e^+e^- collider, sqrt(s) = 10.58 GeV
- Large crossing angle of 83mrad
- Major upgrade to the accelerator with 40× the KEKB design luminosity ($8x10^{35}$ $cm^{-2}s^{-1}$)
  - 2× higher beam currents
  - 20× smaller beam spot (σ_y=50 nm): "Nano-beam scheme"
- Ultimate goal: $50ab^{-1}$ (50× Belle)

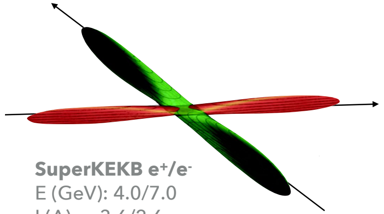


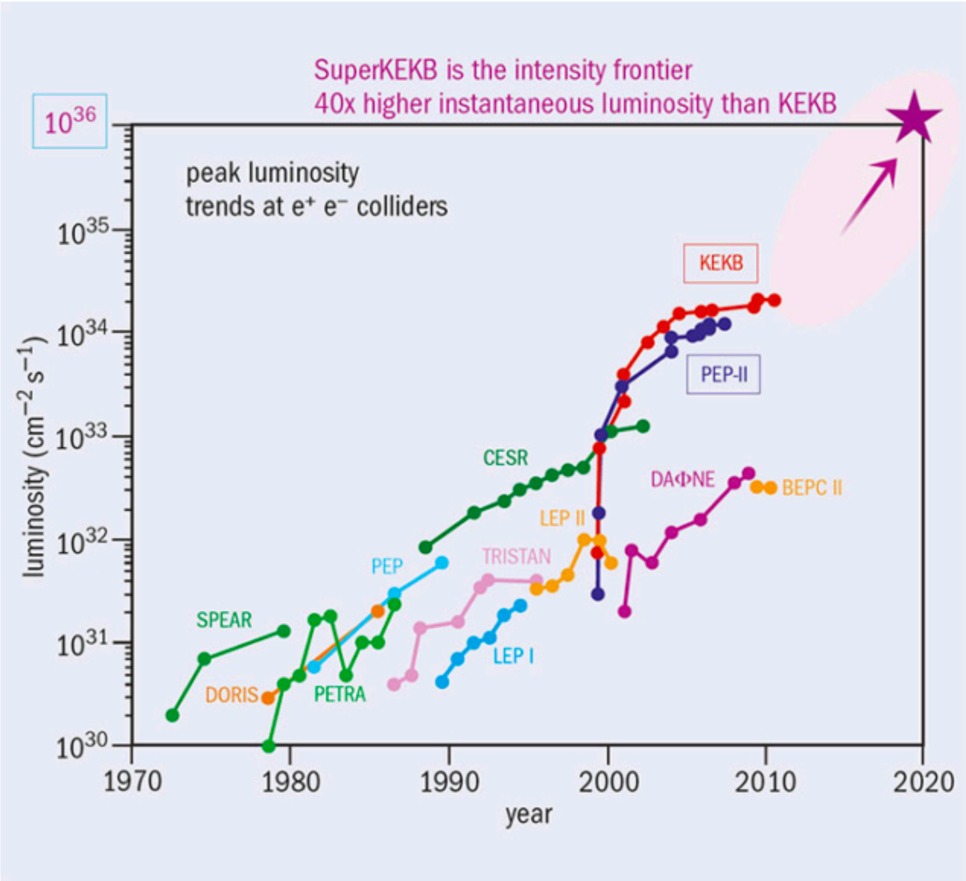


SuperKEKB

- Asymmetric (4.0 GeV/7.0 GeV) e^+e^- collider, sqrt(s) = 10.58 GeV
- Large crossing angle of 83mrad
- Major upgrade to the accelerator with 40× the KEKB design luminosity (8×10^{35} $cm^{-2}s^{-1}$)
  - 2× higher beam currents
  - 20× smaller beam spot (σ_y=50 nm): “Nano-beam scheme”
- Ultimate goal: 50ab^{-1} (50× Belle)

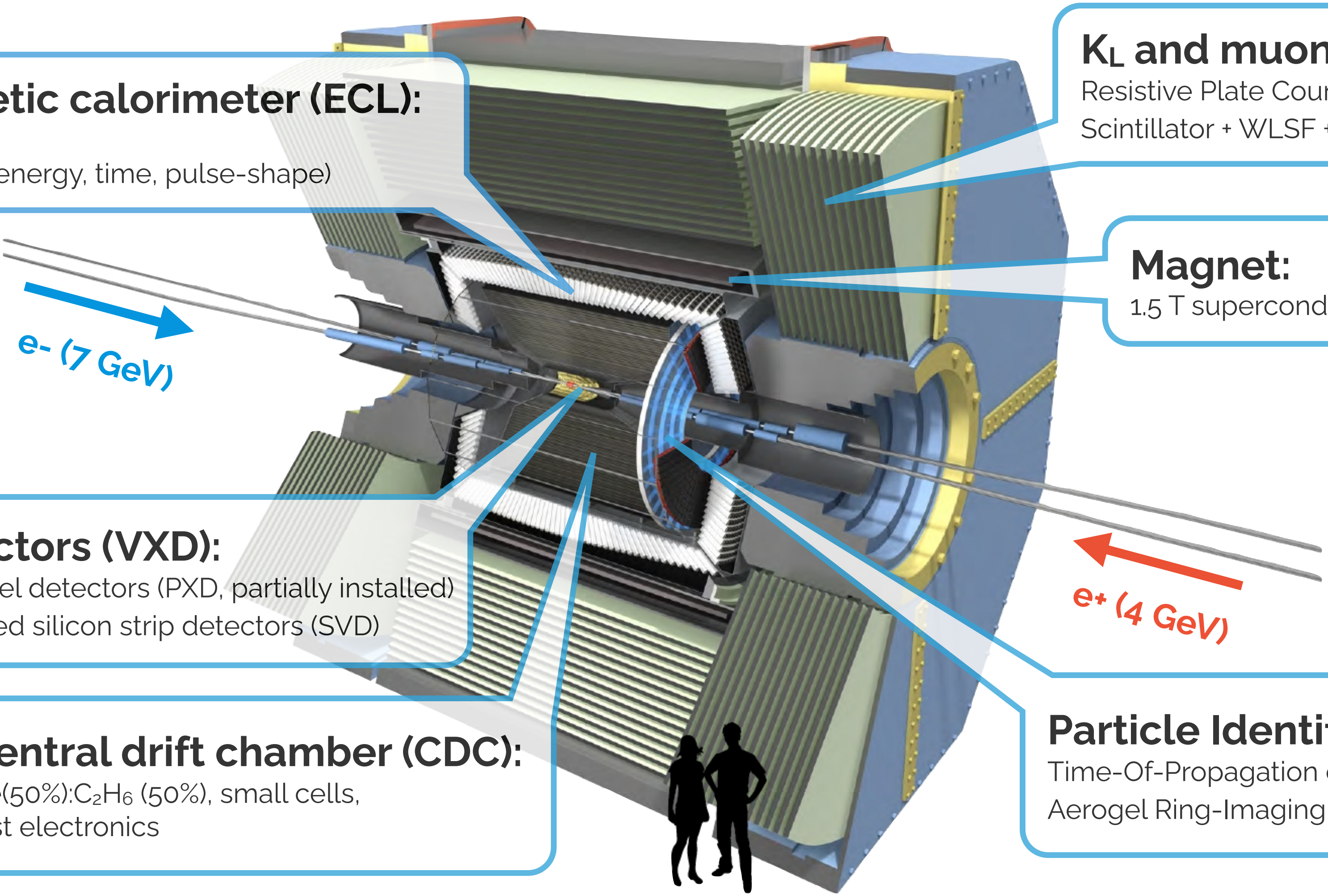






Belle II

Detector



Electromagnetic calorimeter (ECL):
CsI(Tl) crystals
waveform sampling (energy, time, pulse-shape)

K_L and muon detector (KLM):
Resistive Plate Counters (RPC) (outer barrel)
Scintillator + WLSF + MPPC (endcaps, inner barrel)

Magnet:
1.5 T superconducting

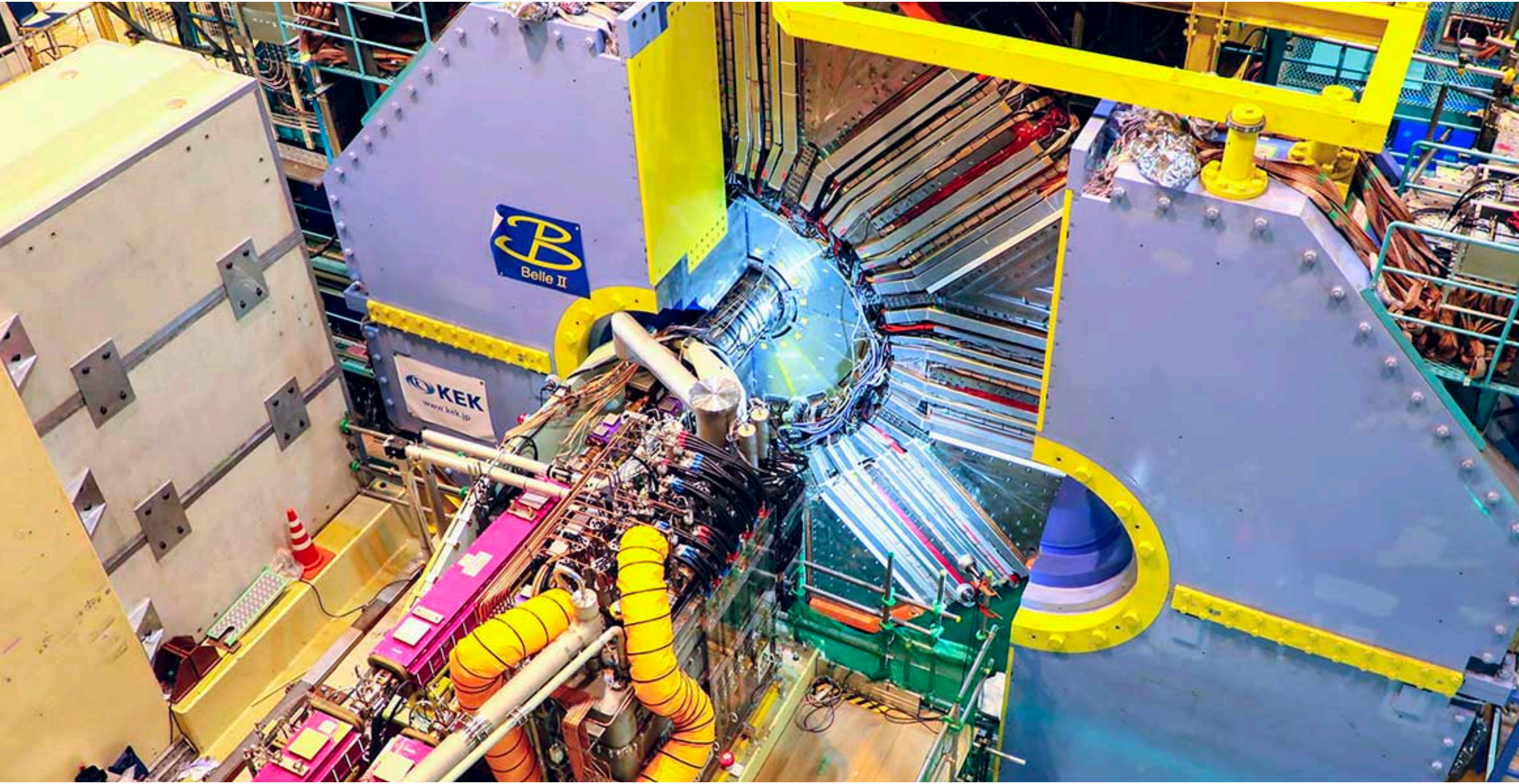
Vertex detectors (VXD):
2 layer DEPFET pixel detectors (PXD, partially installed)
4 layer double-sided silicon strip detectors (SVD)

Trigger:
Hardware: < 30 kHz
Software: < 10 kHz

Central drift chamber (CDC):
He(50%):C_2H_6 (50%), small cells,
fast electronics

Particle Identification (PID):
Time-Of-Propagation counter (TOP) (barrel)
Aerogel Ring-Imaging Cherenkov Counter (ARICH) (FWD)

DEPFET: depleted p-channel field-effect transistor
WLSF: wavelength-shifting fiber
MPPC: multi-pixel photon counter


Belle II
KEK
www.kek.jp

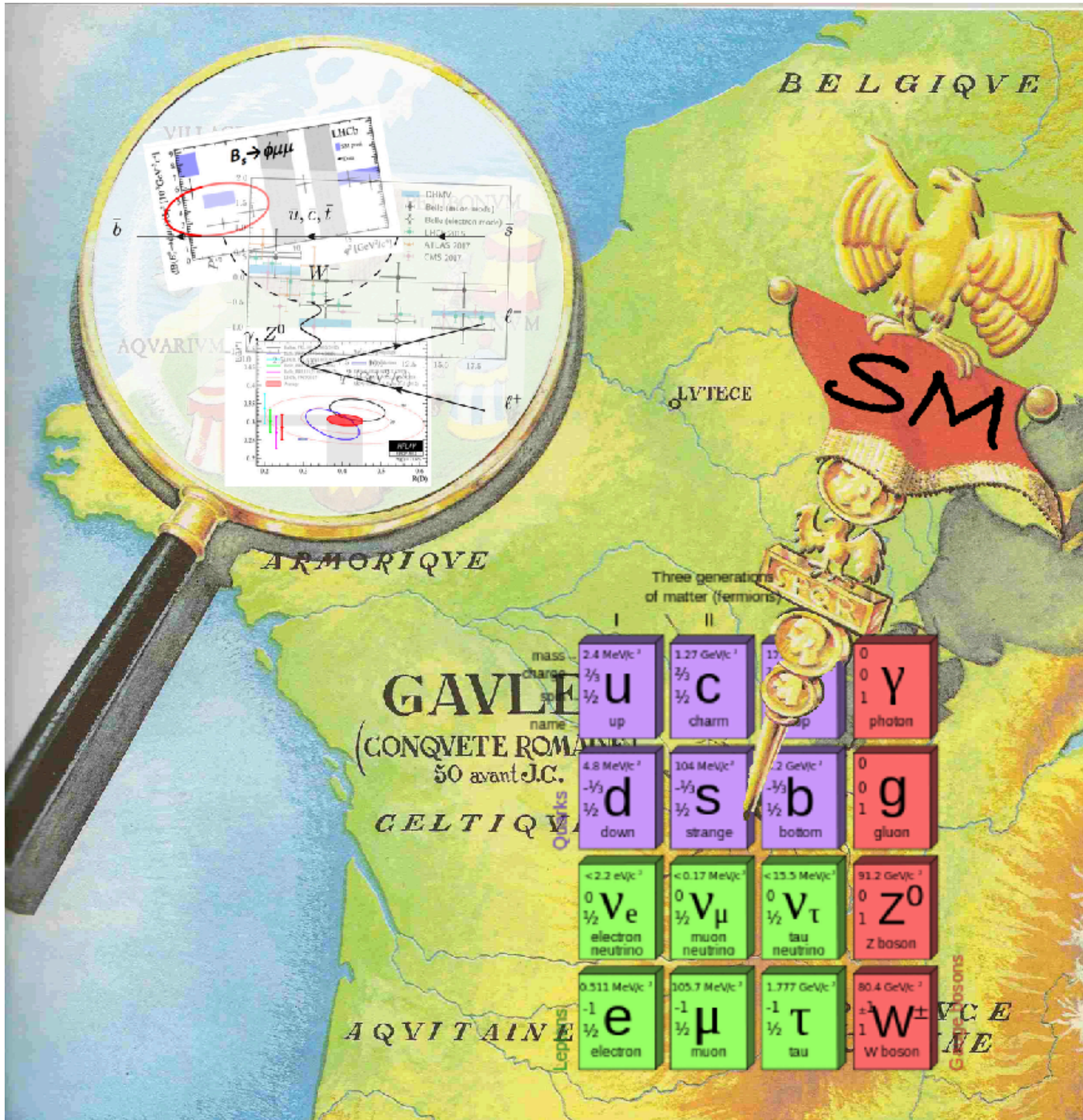
Physics Case

Motivation

- The Standard Model (SM) is very successful in describing the world at particle level

- Although many questions remain unanswered
    - Why do we have three generations of leptons and quarks? Hierarchy, masses, 22 free parameters …

- Almost all SM predictions seem to fit experimental data precisely… Almost?

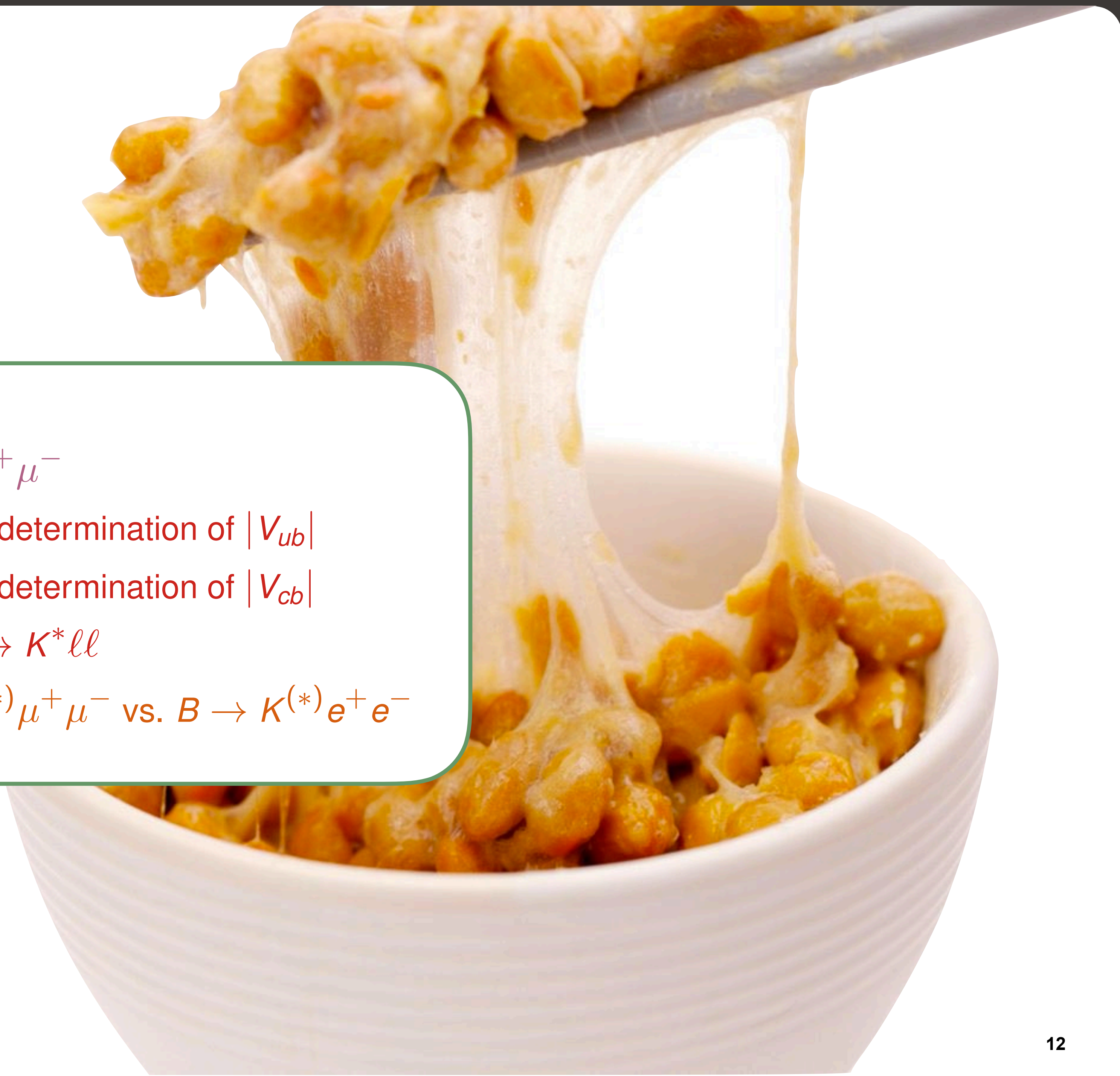


The Flavour Anomalies

(maybe only “local” anomalies…)

The Flavour Anomalies

(maybe only “local” anomalies…)



$> 3.5\sigma$ enhanced $B \to D^{(*)}\tau\nu$ rates

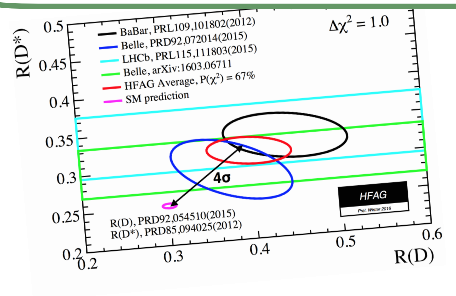
3.3σ suppressed branching ratio of $B_s \to \phi\mu^+\mu^-$

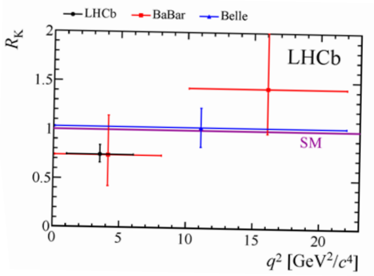
$\sim 3\sigma$ tension between inclusive and exclusive determination of $|V_{ub}|$

$\sim 3\sigma$ tension between inclusive and exclusive determination of $|V_{cb}|$

$> 3\sigma$ anomalies in angular distributions of $B \to K^*\ell\ell$

2.6σ lepton flavor non-universality in $B \to K^{(*)}\mu^+\mu^-$ vs. $B \to K^{(*)}e^+e^-$

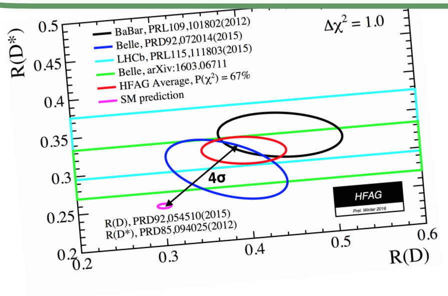


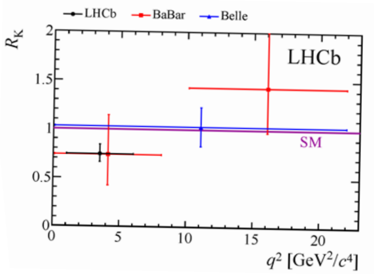


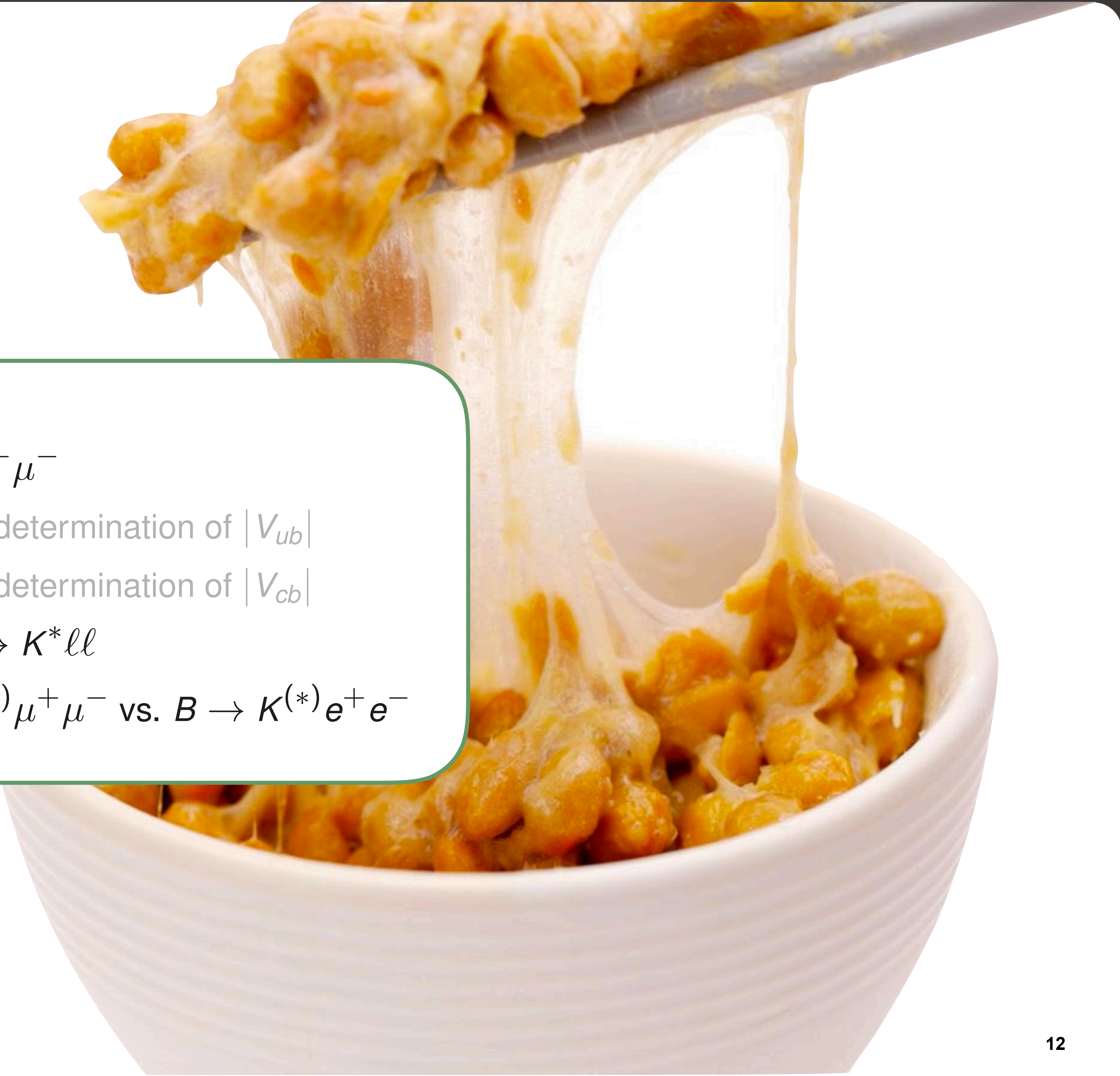
The Flavour Anomalies

(maybe only “local” anomalies…)

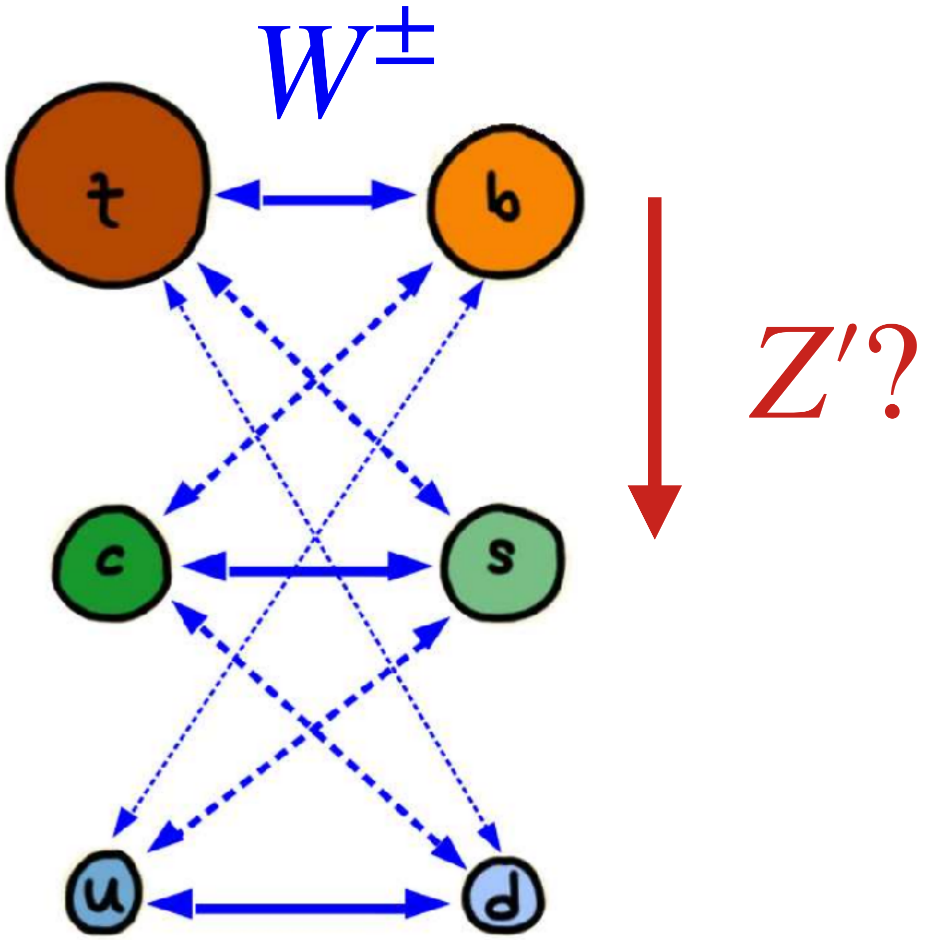
- $> 3.5\sigma$ enhanced $B \to D^{(*)}\tau\nu$ rates
- 3.3σ suppressed branching ratio of $B_s \to \phi\mu^+\mu^-$
- $\sim 3\sigma$ tension between inclusive and exclusive determination of $|V_{ub}|$
- $\sim 3\sigma$ tension between inclusive and exclusive determination of $|V_{cb}|$
- $> 3\sigma$ anomalies in angular distributions of $B \to K^*\ell\ell$
- 2.6σ lepton flavor non-universality in $B \to K^{(*)}\mu^+\mu^-$ vs. $B \to K^{(*)}e^+e^-$



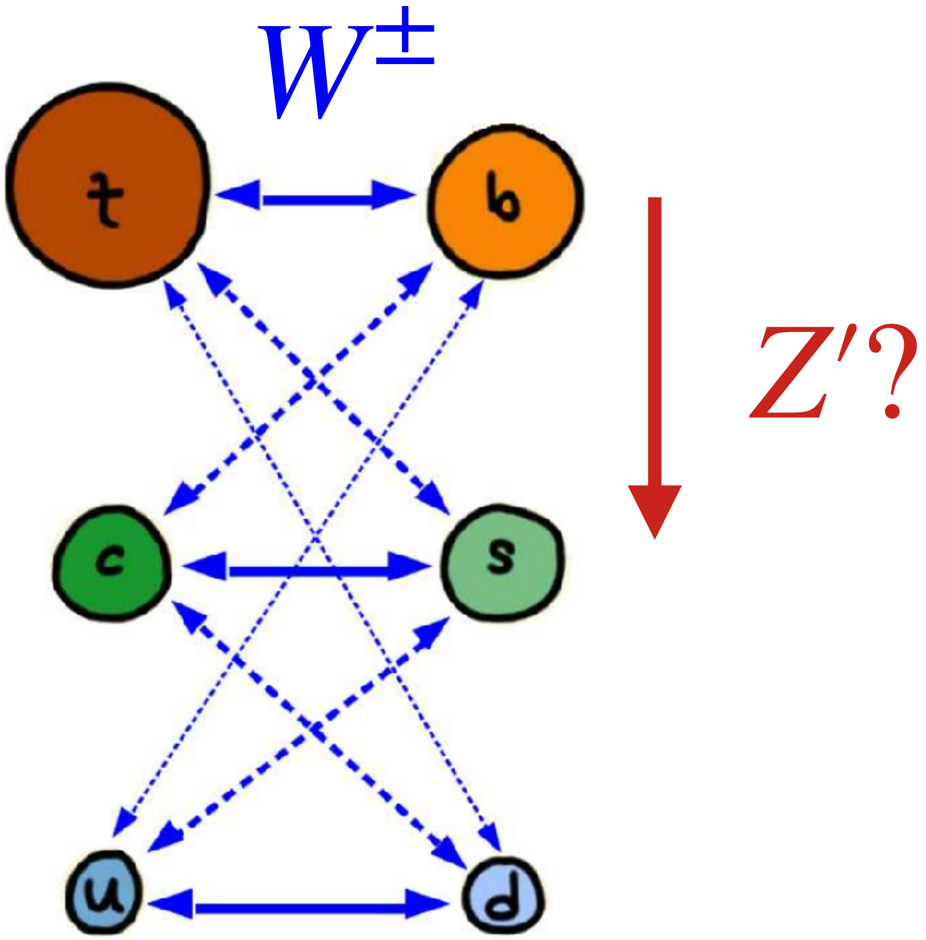


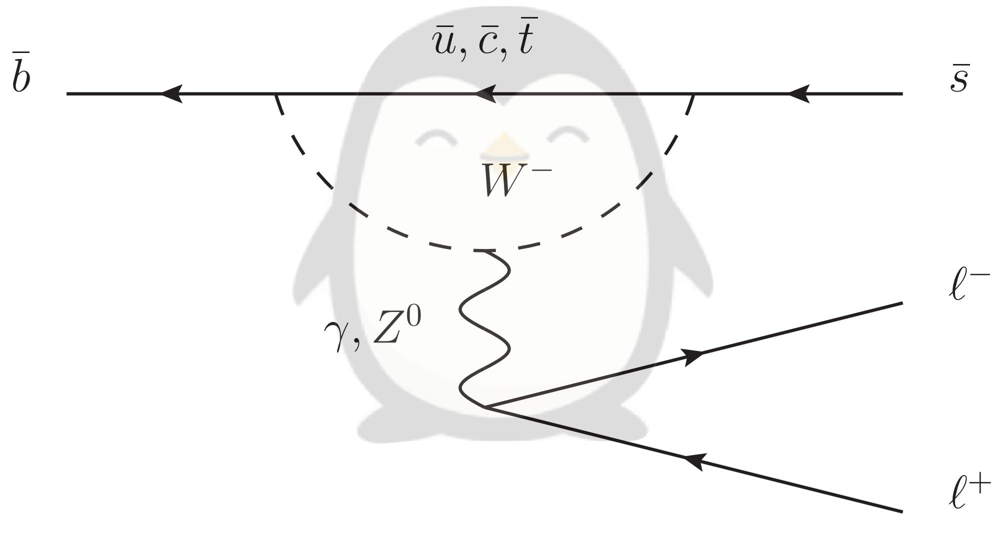


The b→s transition

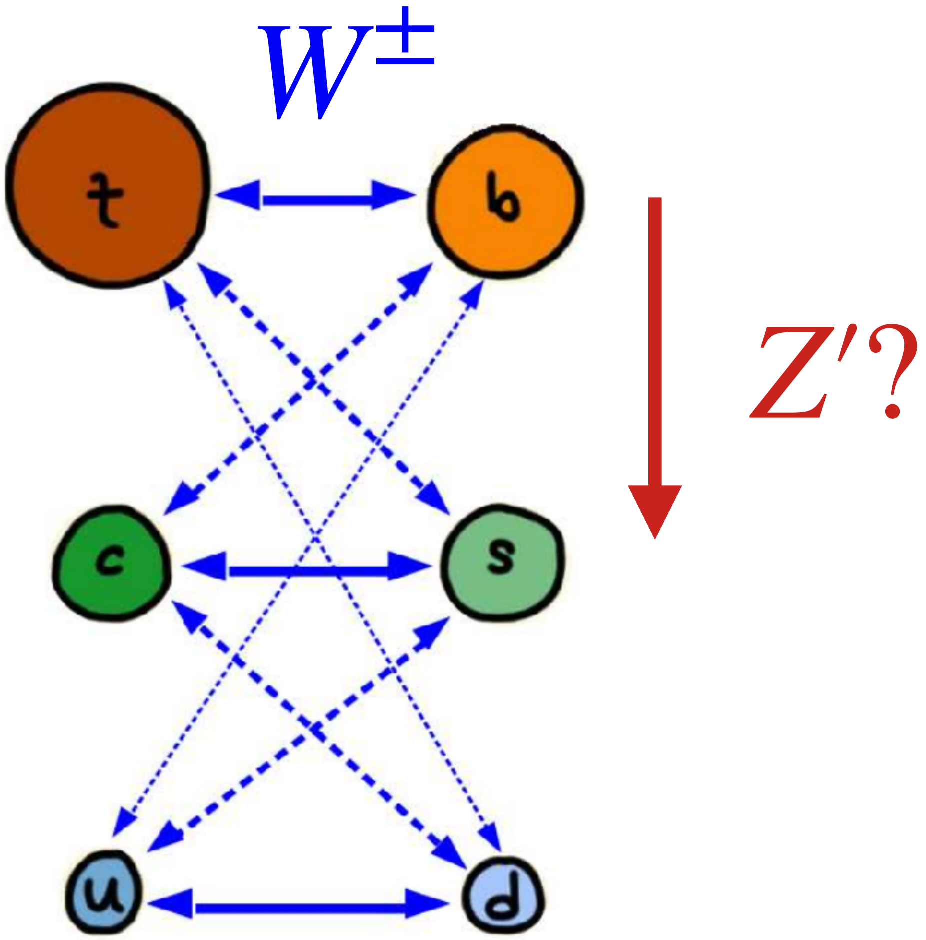


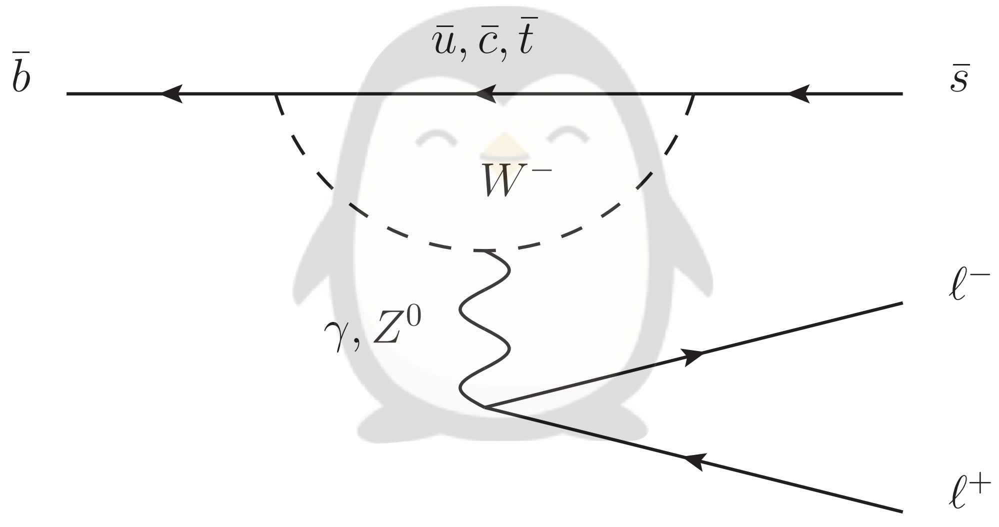
The b→s transition



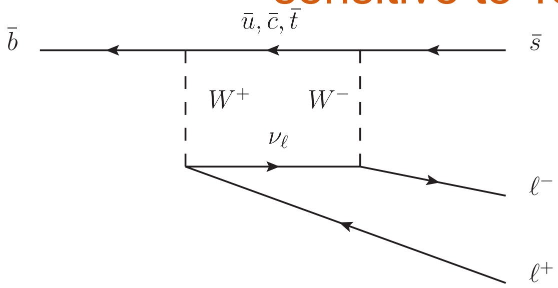


The b→s transition

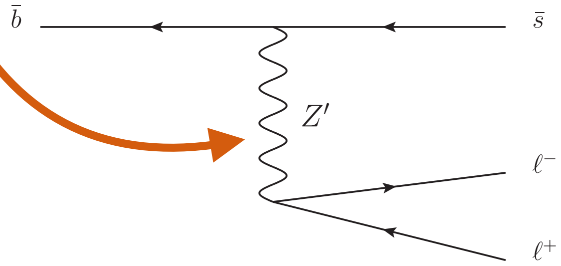




sensitive to TeV scale in loops



(c) SM example



(d) NP example

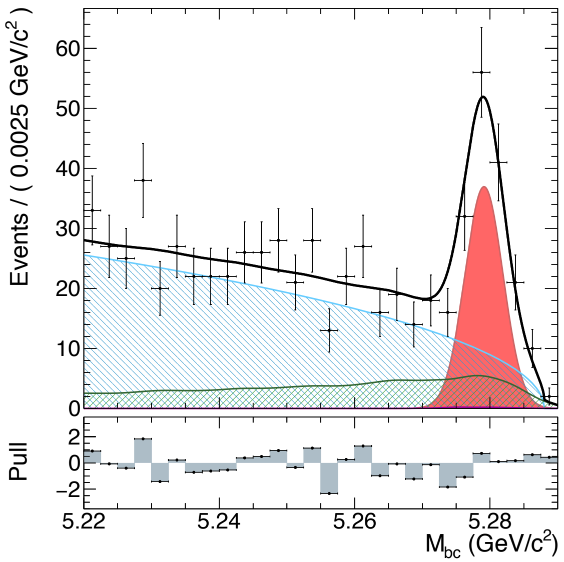
$$\mathscr{B}_{SM}(b \to s\ell\ell) = \mathcal{O}(10^{-6})$$

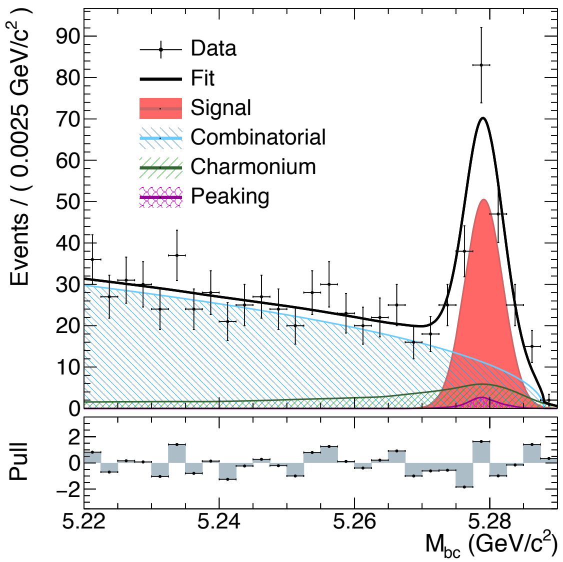
B→K*ll Analysis at Belle

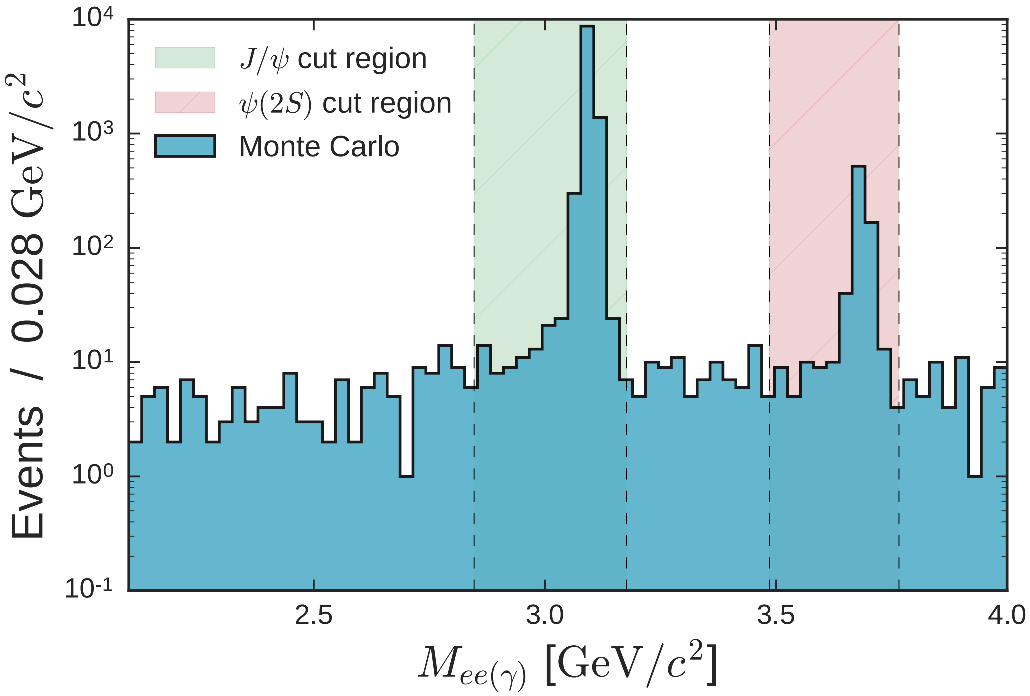
Belle Analysis

- Similar electron and muon performance
- limited statistics
- Neural network based reconstruction in order to maximise efficiency

$$B^+ \to K^{*+}(K^+\pi^0)\ell^+\ell^- \qquad B^0 \to K^{*0}(K_S\pi^0)\ell^+\ell^-$$
$$B^+ \to K^{*+}(K_S\pi^+)\ell^+\ell^- \qquad B^0 \to K^{*0}(K^+\pi^-)\ell^+\ell^-$$







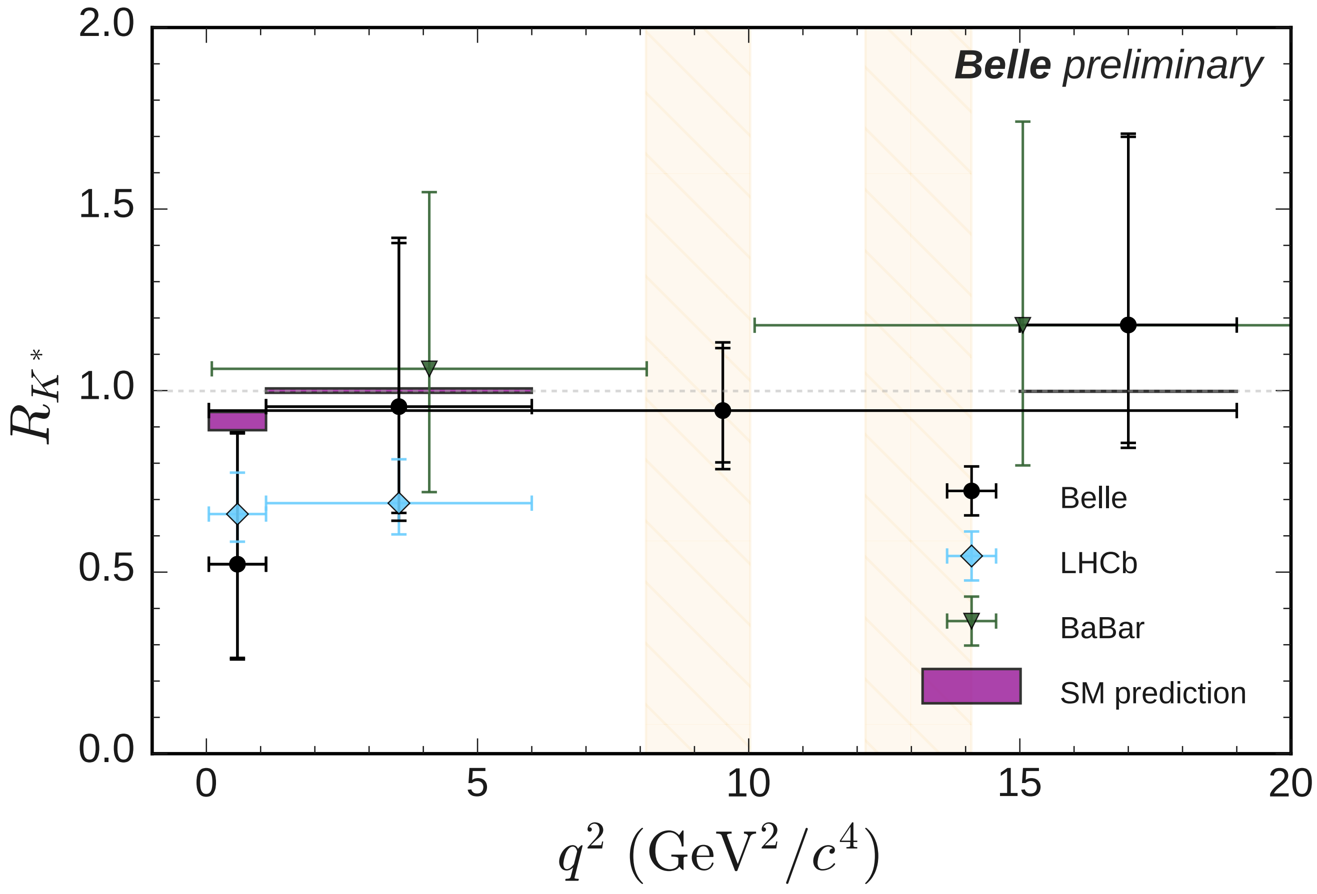
Most simple approach: Ratio of Branching Ratios

After Moriond

- Measurements in accordance with the SM
- First measurement of $R_{K^{*+}}$

TABLE II. Result for R_{K^*}, $R_{K^{*0}}$ and $R_{K^{*+}}$. The first uncertainty is statistical and the second is systematic.

q^2 in GeV$^2/c^4$	All modes	B^0 modes	B^+ modes
$[0.045, 1.1]$	$0.52^{+0.36}_{-0.26} \pm 0.05$	$0.46^{+0.55}_{-0.27} \pm 0.07$	$0.62^{+0.60}_{-0.36} \pm 0.10$
$[1.1, 6]$	$0.96^{+0.45}_{-0.29} \pm 0.11$	$1.06^{+0.63}_{-0.38} \pm 0.13$	$0.72^{+0.99}_{-0.44} \pm 0.18$
$[0.1, 8]$	$0.90^{+0.27}_{-0.21} \pm 0.10$	$0.86^{+0.33}_{-0.24} \pm 0.08$	$0.96^{+0.56}_{-0.35} \pm 0.14$
$[15, 19]$	$1.18^{+0.52}_{-0.32} \pm 0.10$	$1.12^{+0.61}_{-0.36} \pm 0.10$	$1.40^{+1.99}_{-0.68} \pm 0.11$
$[0.045,]$	$0.94^{+0.17}_{-0.14} \pm 0.08$	$1.12^{+0.27}_{-0.21} \pm 0.09$	$0.70^{+0.24}_{-0.19} \pm 0.07$

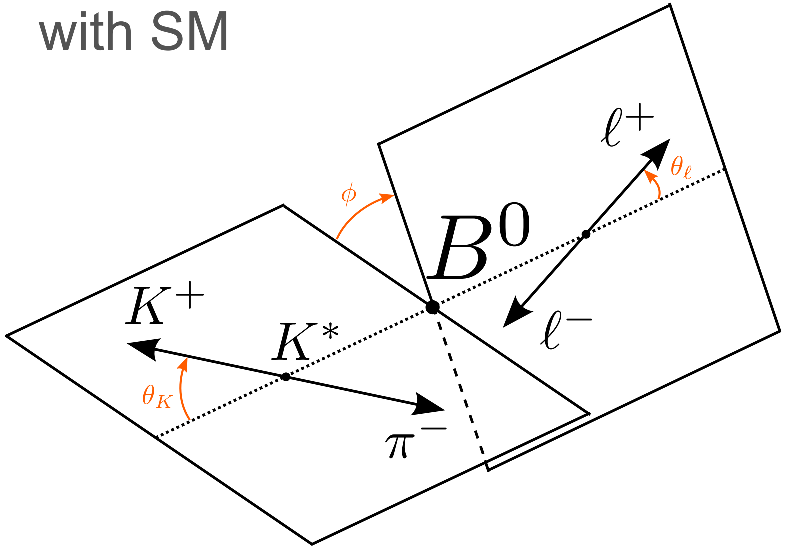


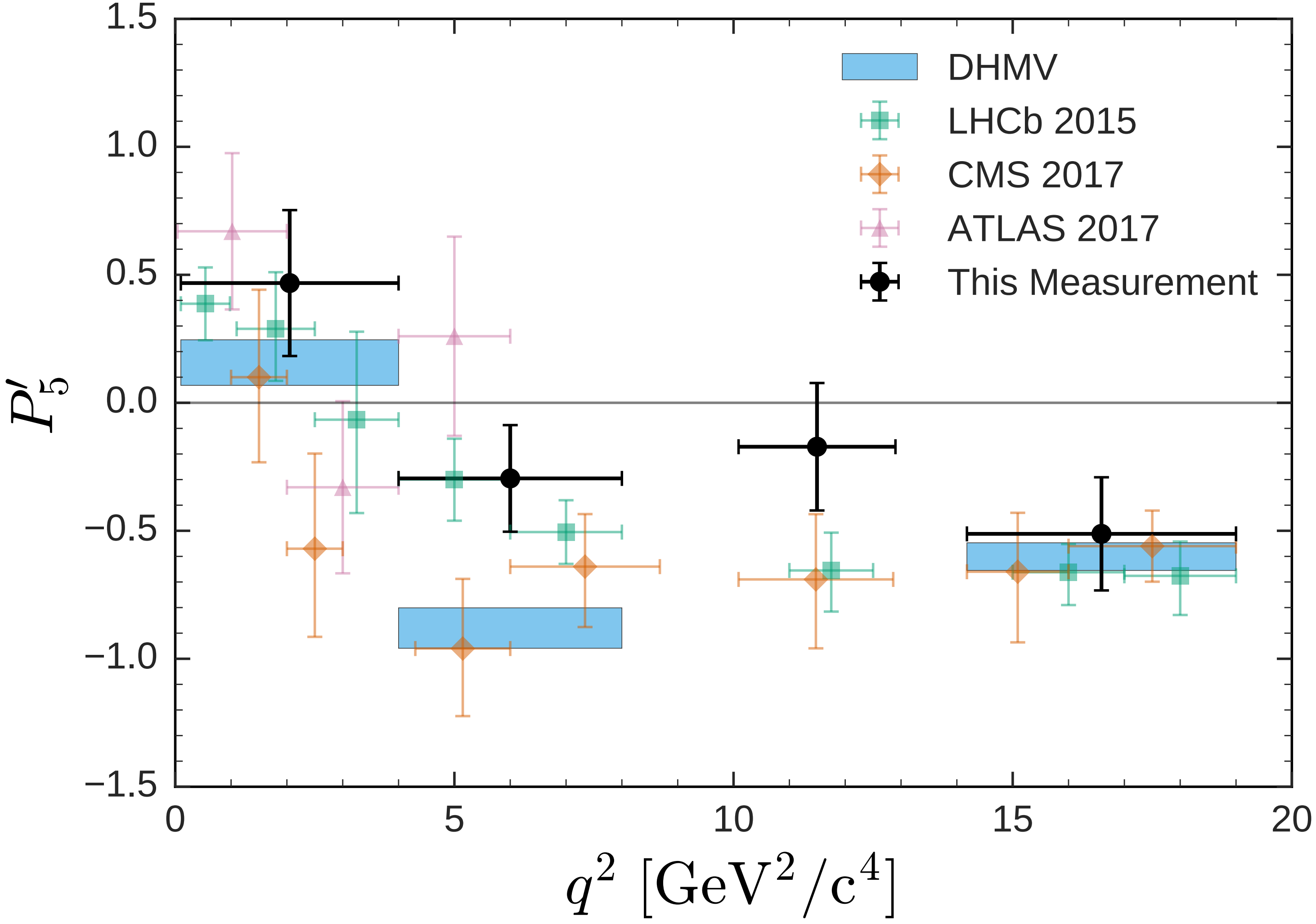
arXiv:1904.02440

Belle 1 Angular Analysis

Results

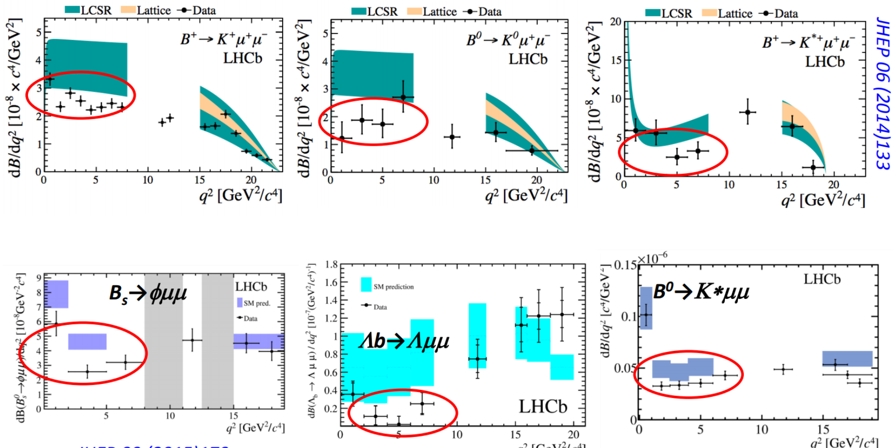
- LHCb sees the largest deviation in the low q^2 region
- Atlas and Belle can confirm the anomaly with less significance
- CMS is in good agreement with SM





Flavour Anomalies in b->sll

New Physics or systematic problem?

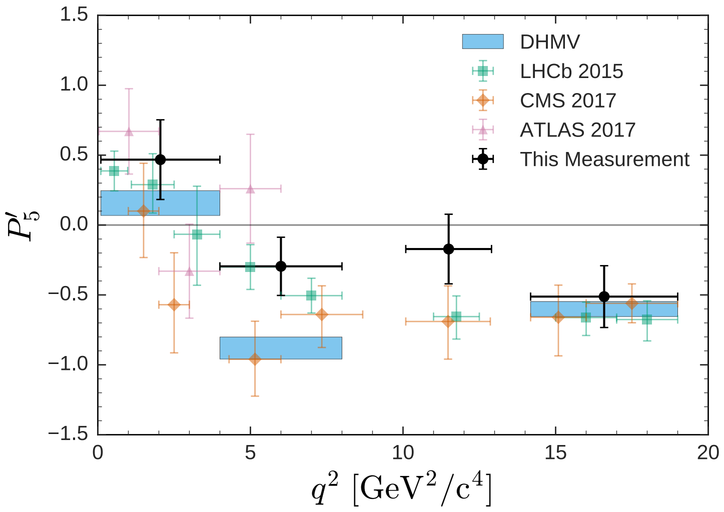


JHEP 06 (2014)133

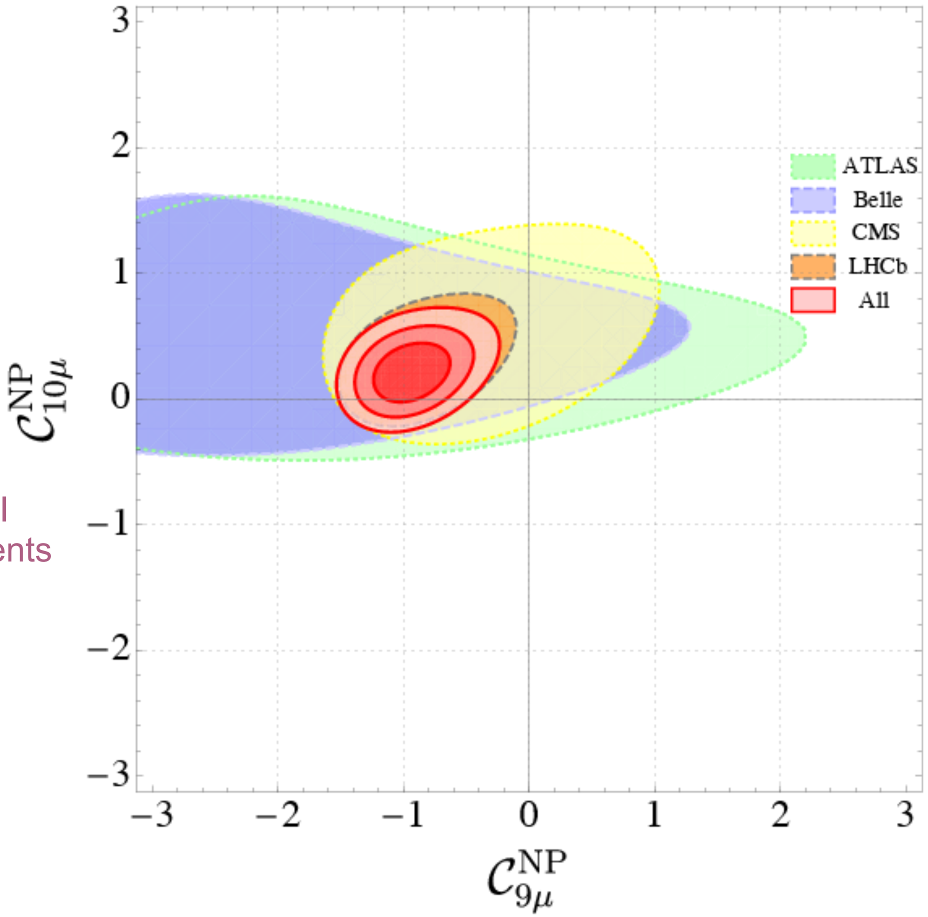
JHEP 09 (2015)179

JHEP 06 (2015) 115

JHEP 11(2016)047
JHEP 04(2017)142



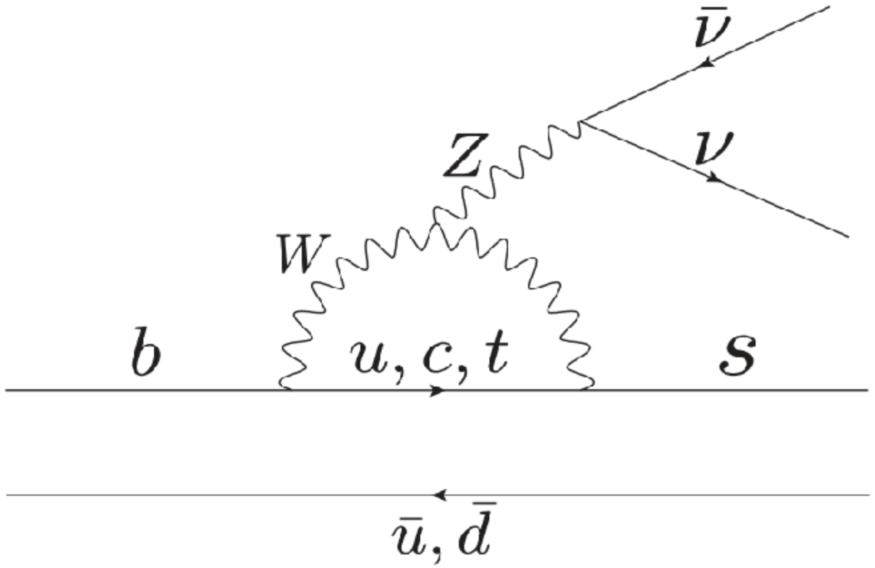
Combined Fit to all available measurements

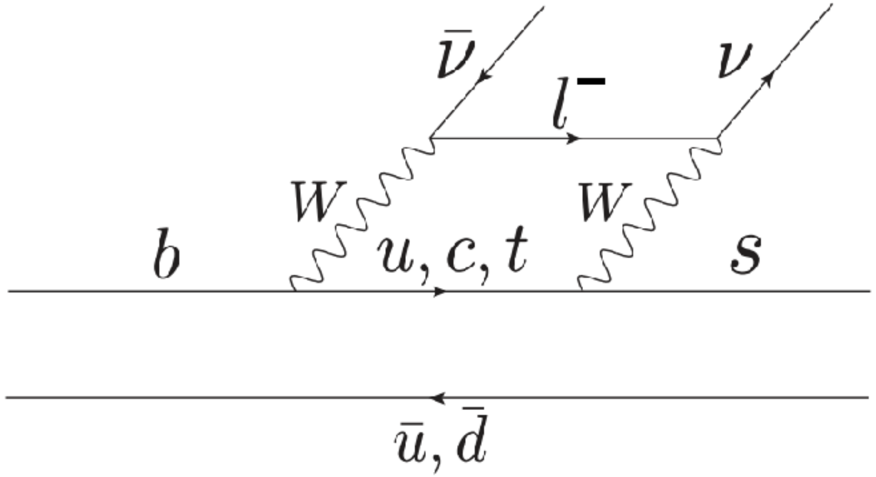


$$\mathcal{C}_9 = \mathcal{C}_9^{SM} + \mathcal{C}_9^{NP}$$

The neutrino case

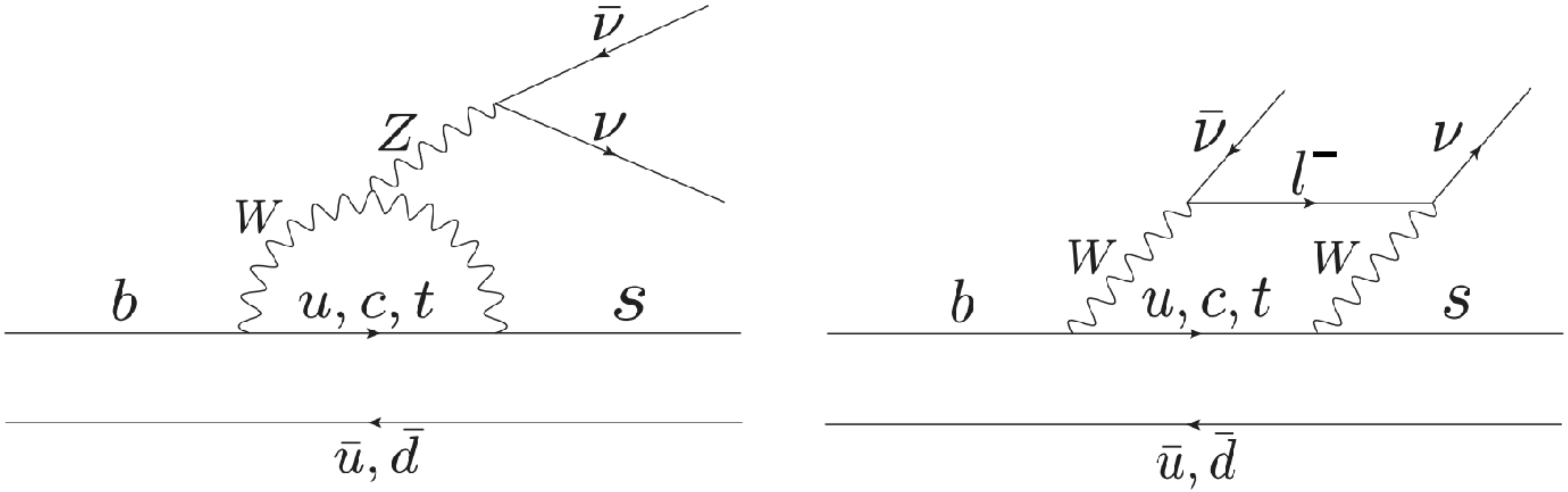
Golden mode for Belle II





The neutrino case

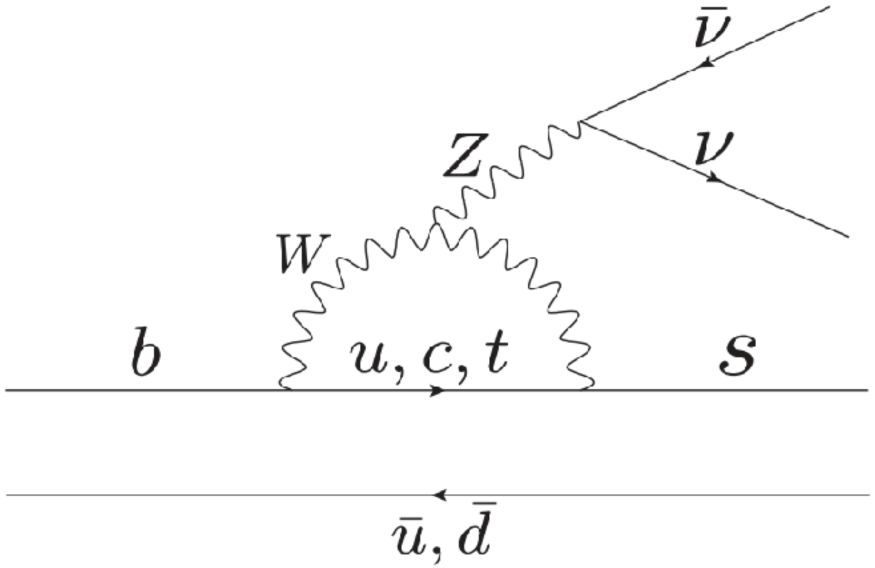
Golden mode for Belle II

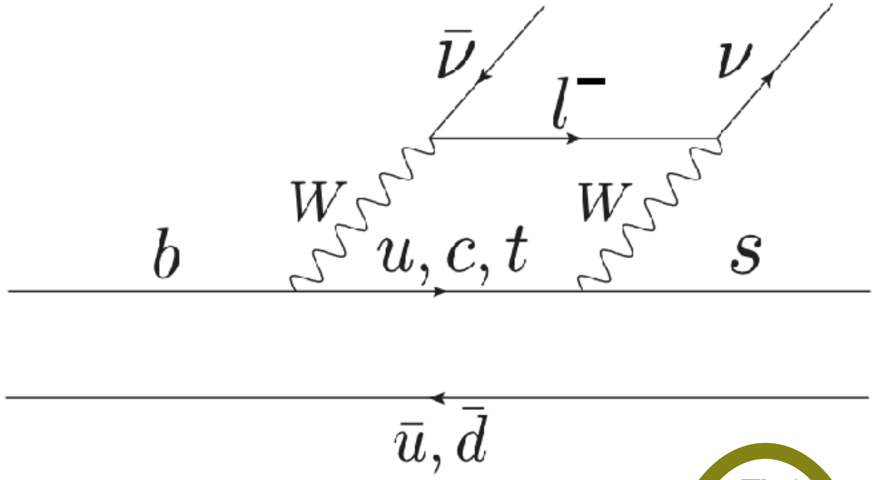


- Sensitive to similar NP as tension in C9:
  - b→s transition shows signs of NP
- Theoretically very clean (no charm loops)

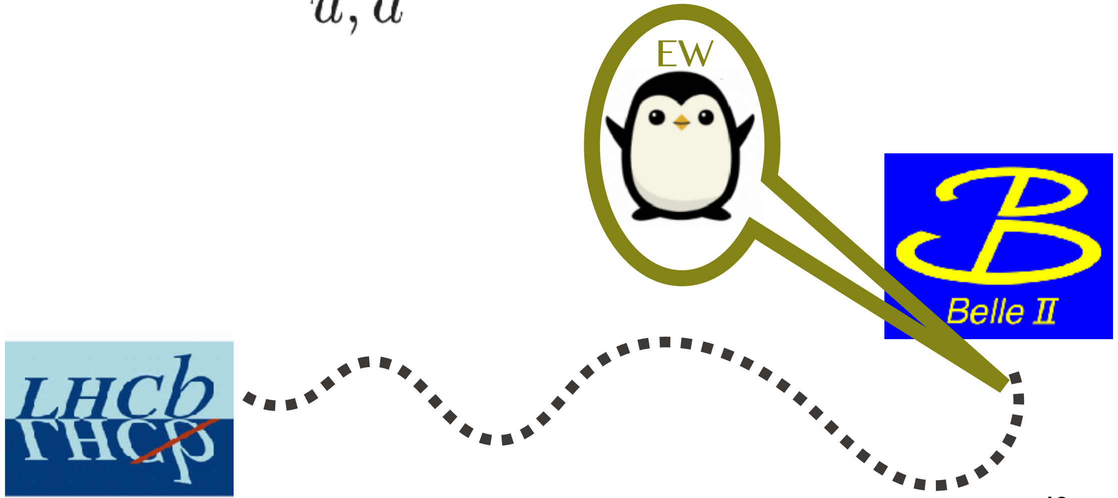
The neutrino case

Golden mode for Belle II

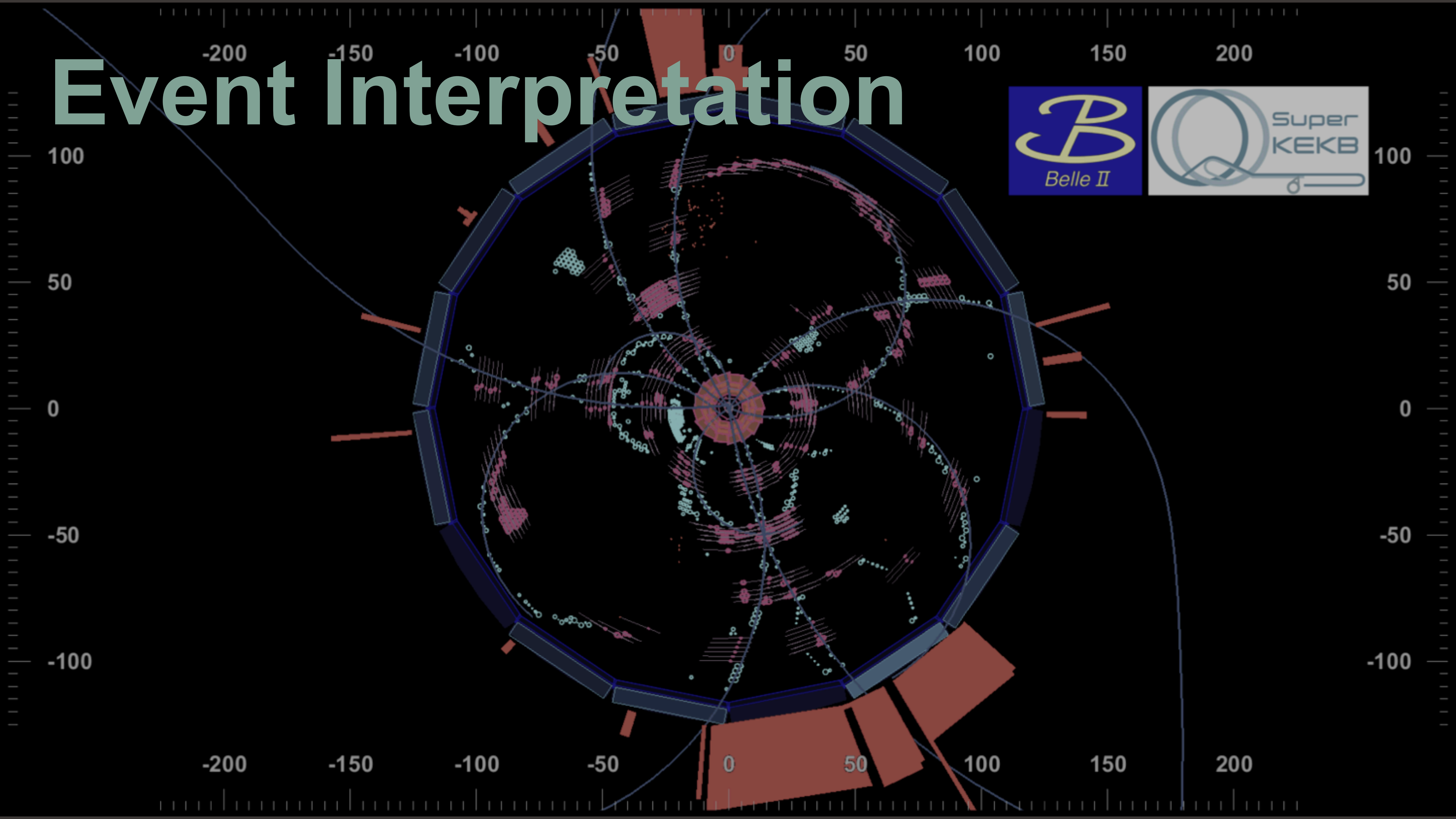




- Sensitive to similar NP as tension in C9:
  - b→s transition shows signs of NP
- Theoretically very clean (no charm loops)

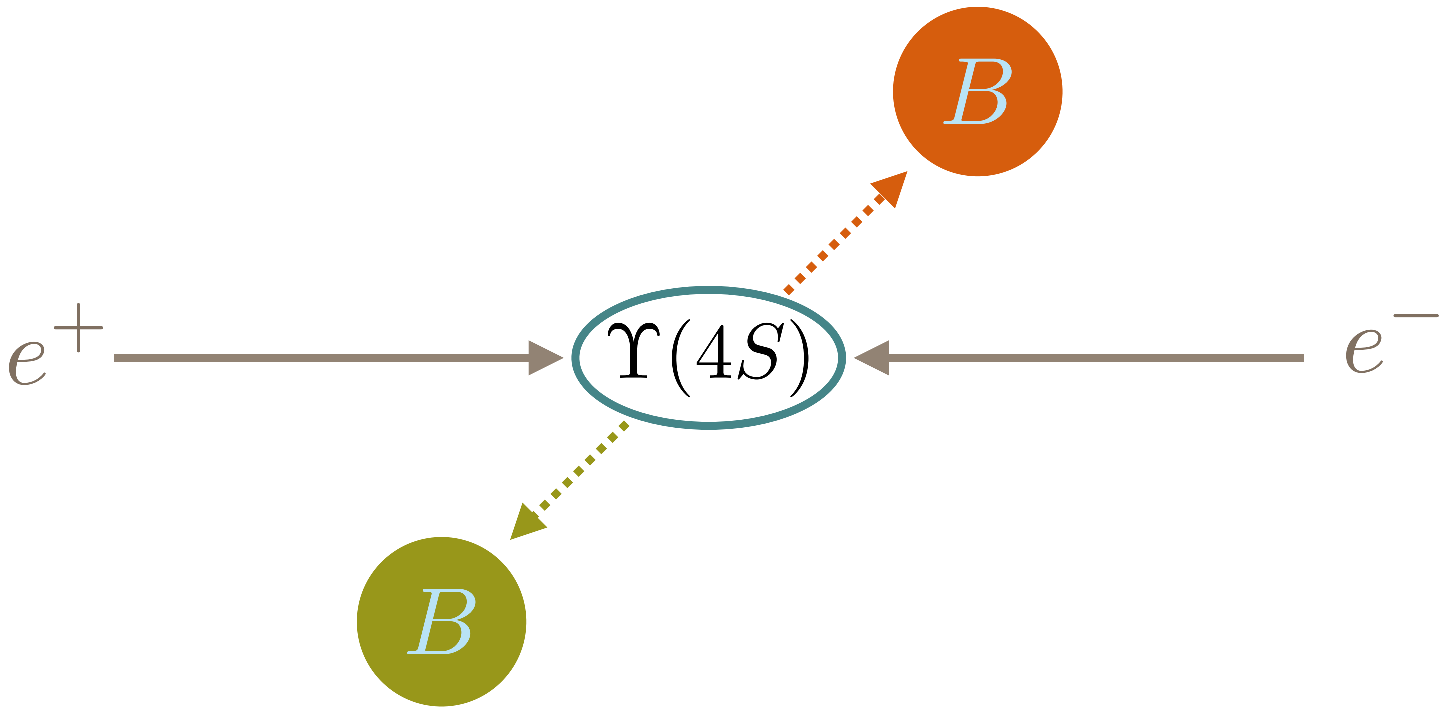


Event Interpretation



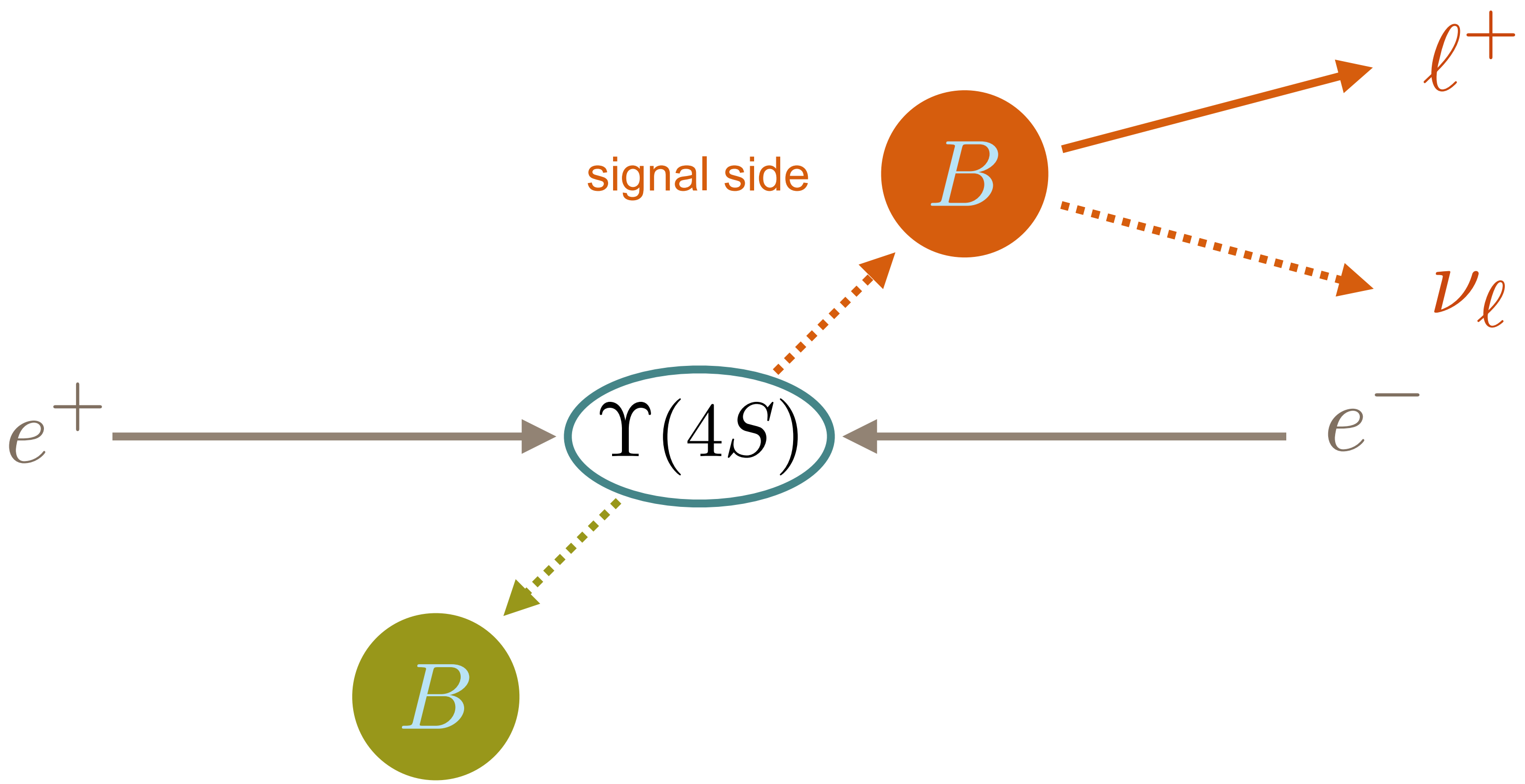
Experimental Setup for Belle II

Advantages at e^+e^- colliders



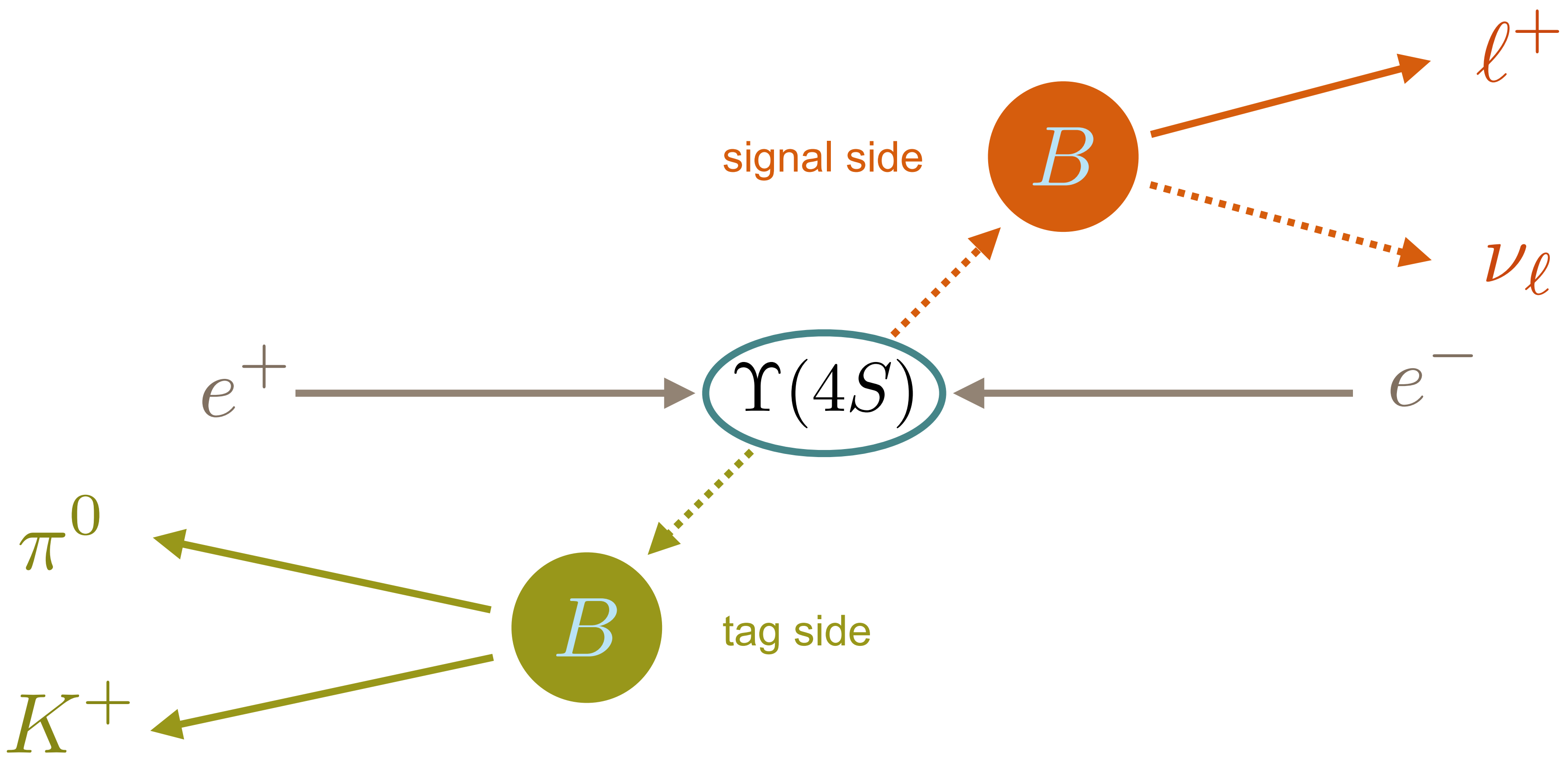
Experimental Setup for Belle II

Advantages at e^+e^- colliders



Experimental Setup for Belle II

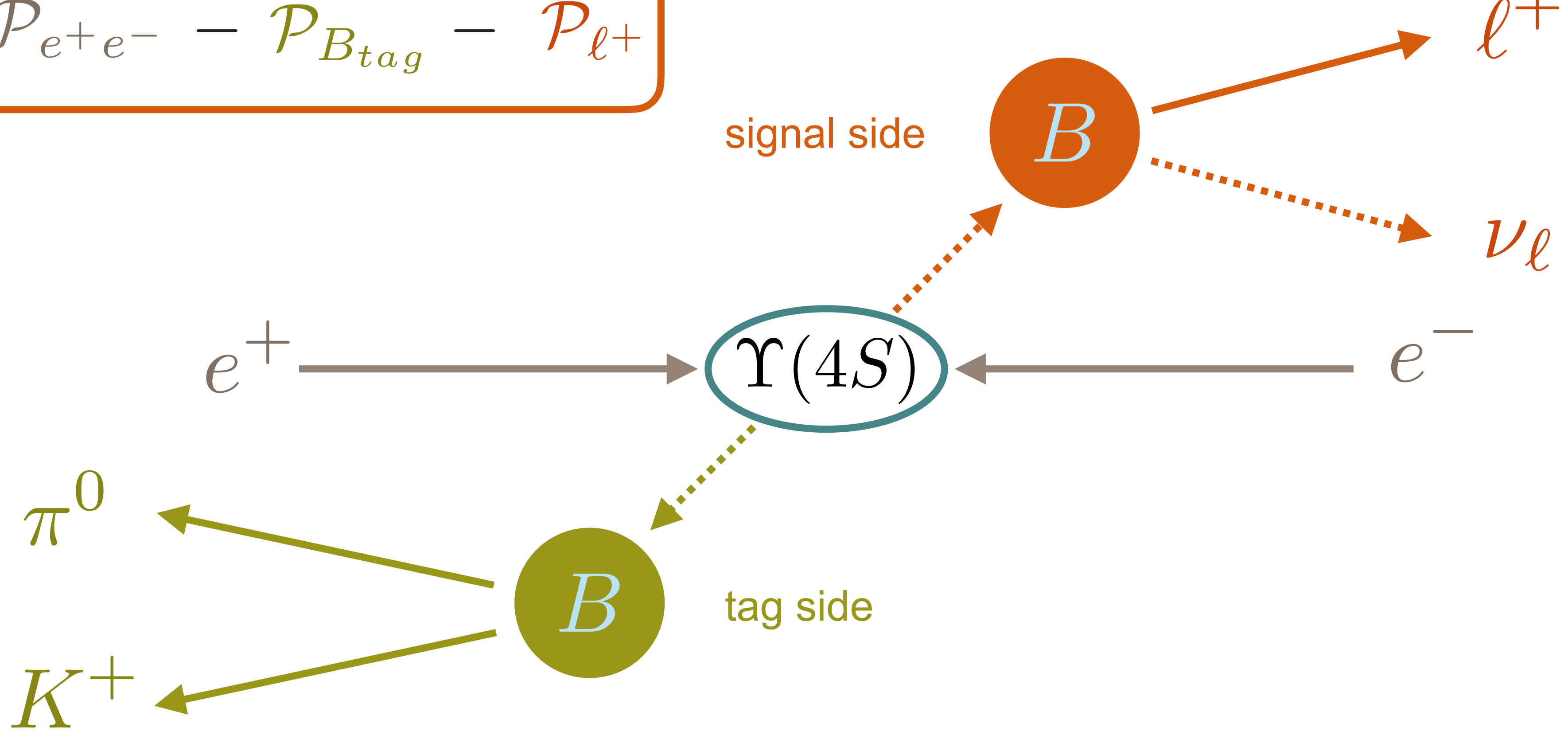
Advantages at e⁺e⁻ colliders



Experimental Setup for Belle II

Advantages at e⁺e⁻ colliders

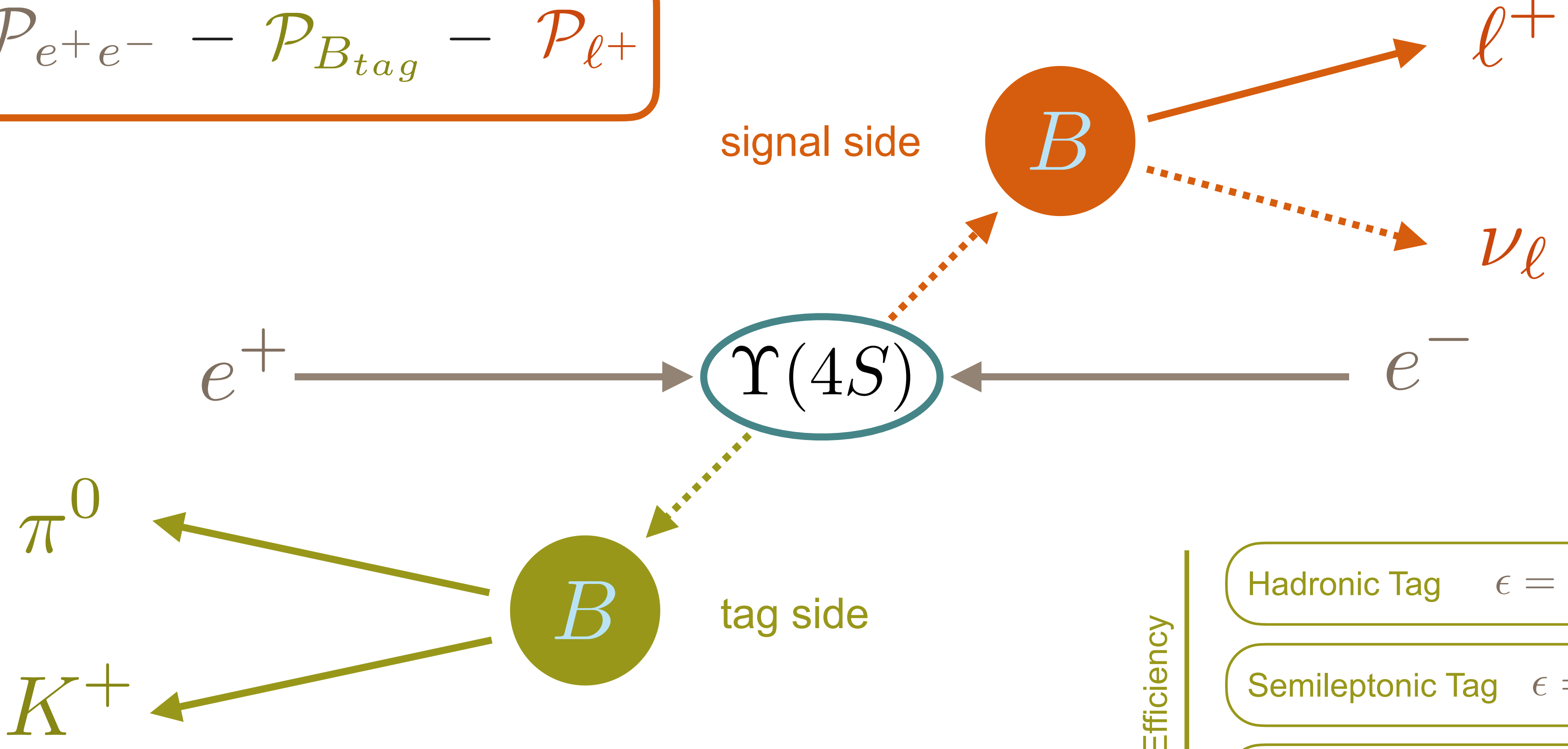
$$\boxed{\mathcal{P}_{\nu_\ell} = \mathcal{P}_{e^+e^-} - \mathcal{P}_{B_{tag}} - \mathcal{P}_{\ell^+}}$$

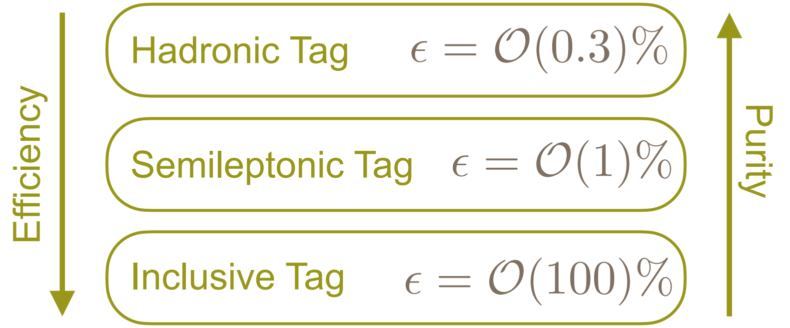


Experimental Setup for Belle II

Advantages at e⁺e⁻ colliders

$$\boxed{\mathcal{P}_{\nu_\ell} = \mathcal{P}_{e^+e^-} - \mathcal{P}_{B_{tag}} - \mathcal{P}_{\ell^+}}$$





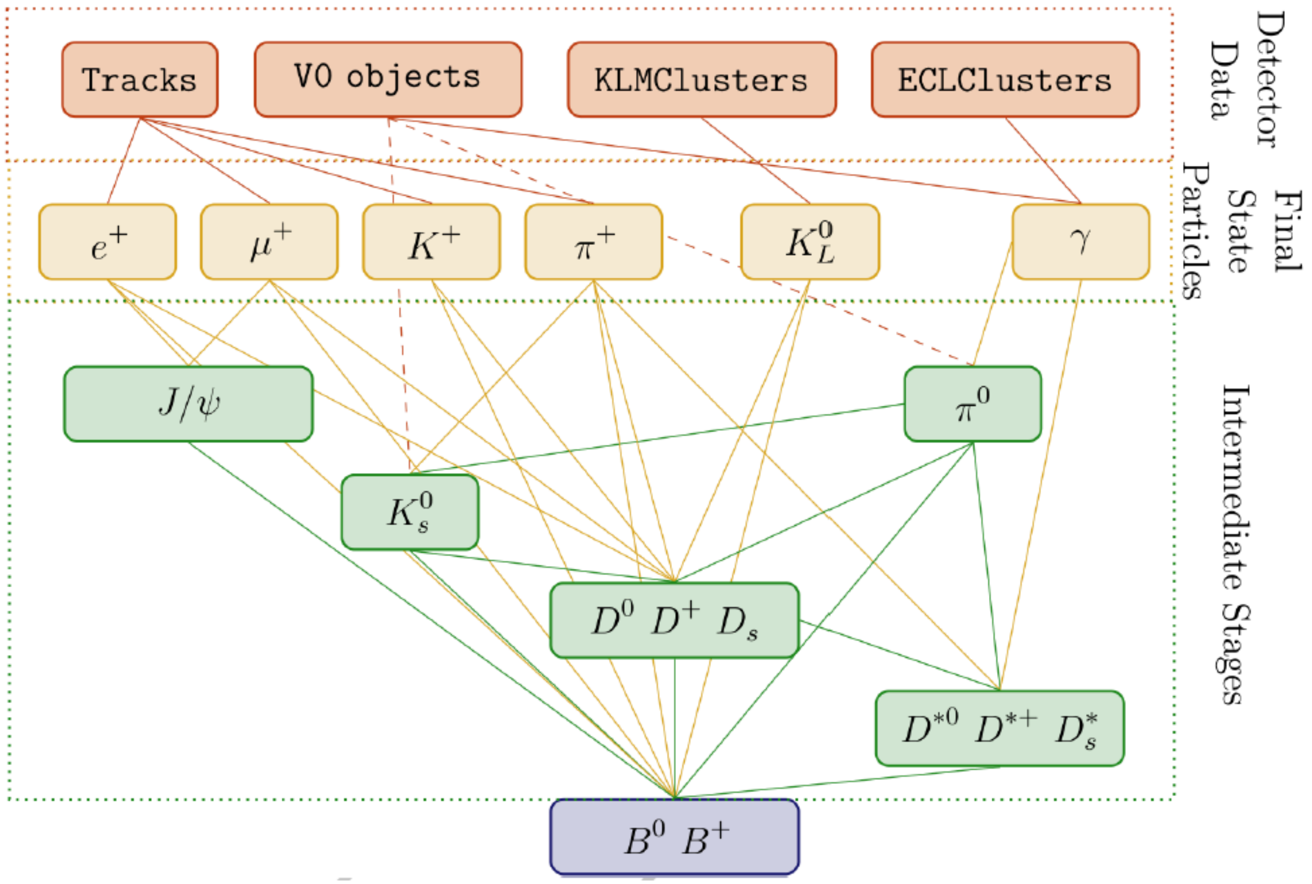
Missing Energy Channels

Full Event Interpretation (FEI)

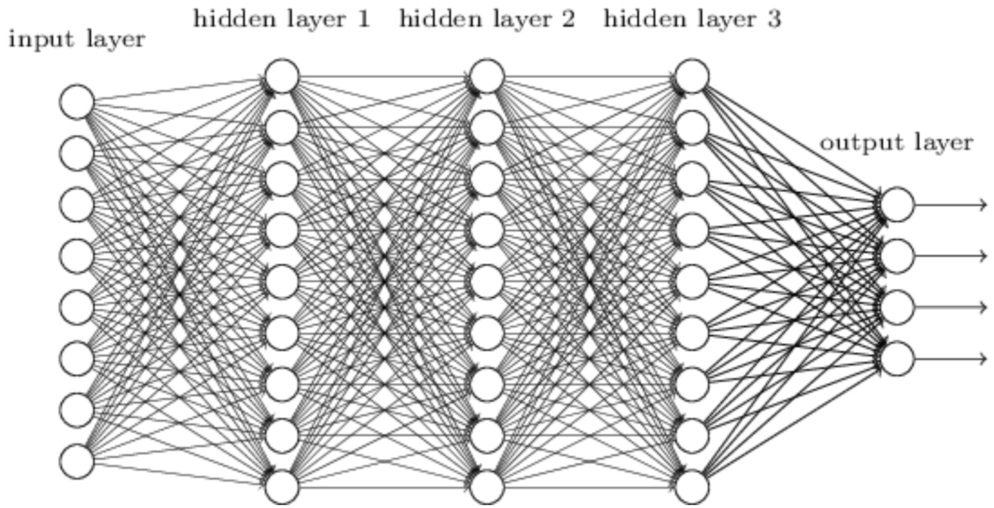
- Hierarchical approach
  - Multivariate classifier for each state
  - Gather all information in the signal probability
- FEI can provide hadronic and semileptonic final states

Maximum reconstruction efficiency

Tag	FR @ Belle	FEI @ Belle	FEI @ Belle II
Hadronic B^+	0.28 %	0.49 %	0.61 %
Semileptonic B^+	0.67 %	1.42 %	1.45 %
Hadronic B^0	0.18 %	0.33%	0.34 %
Semileptonic B^0	0.63 %	1.33%	1.25 %

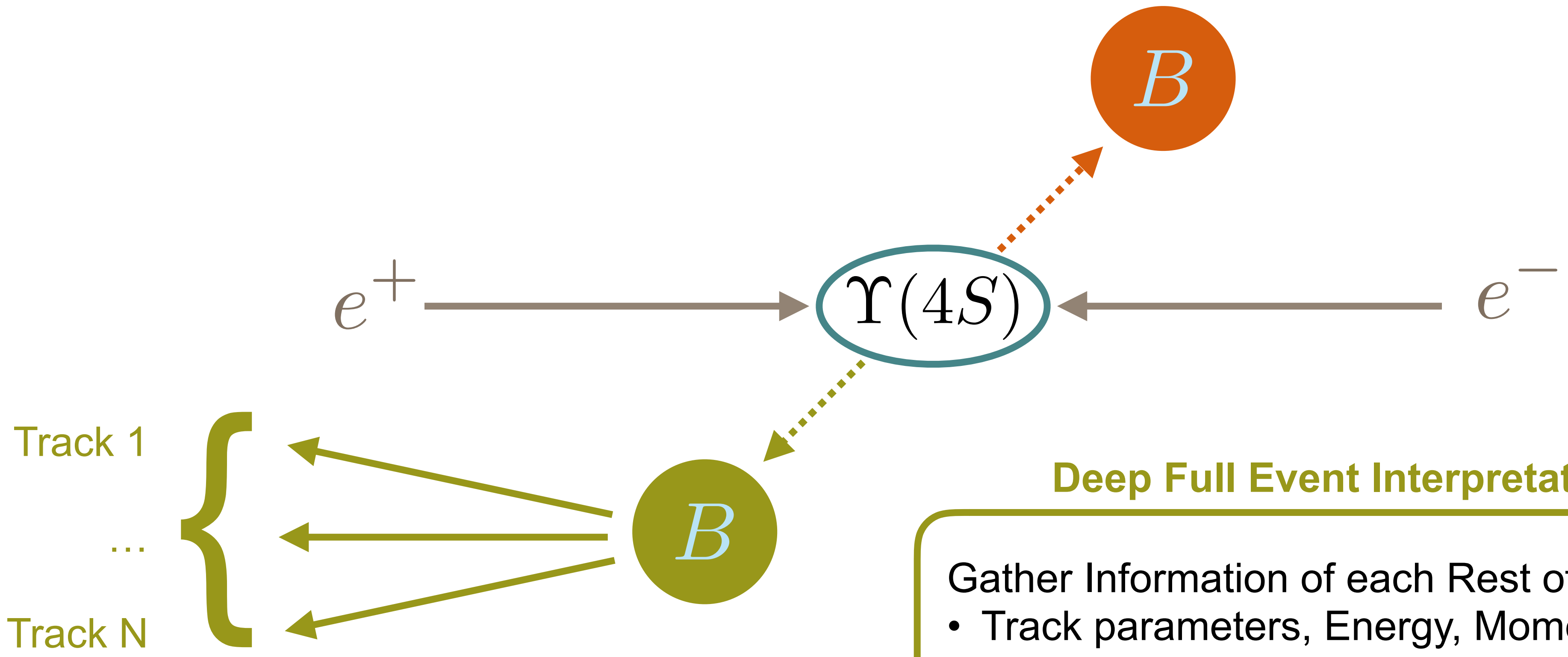


Improving the sensitivity: Deep Full Event Interpretation



Deep Full Event Interpretation

Similar to Inclusive Tagging Approach

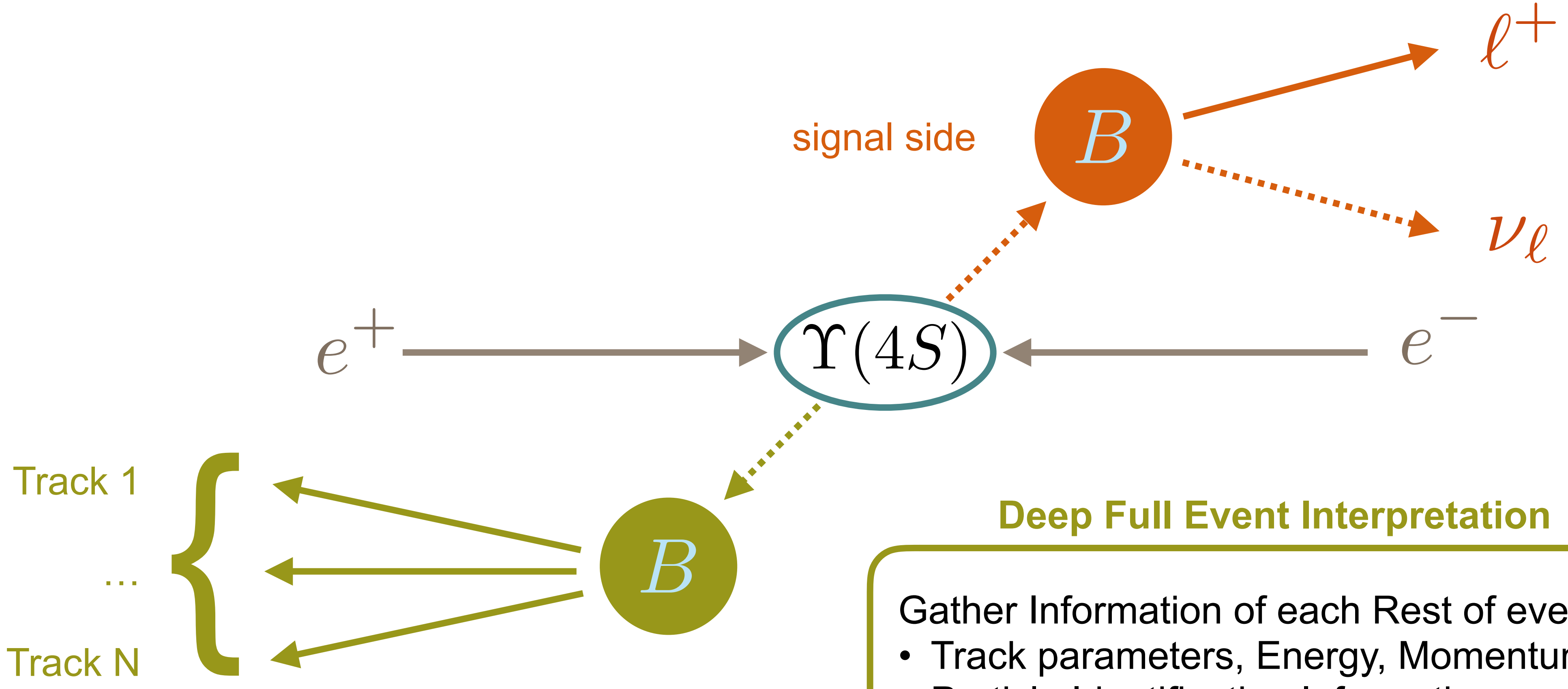


Deep Full Event Interpretation

Gather Information of each Rest of event Track
- Track parameters, Energy, Momentum
- Particle Identification Information

Deep Full Event Interpretation

Similar to Inclusive Tagging Approach

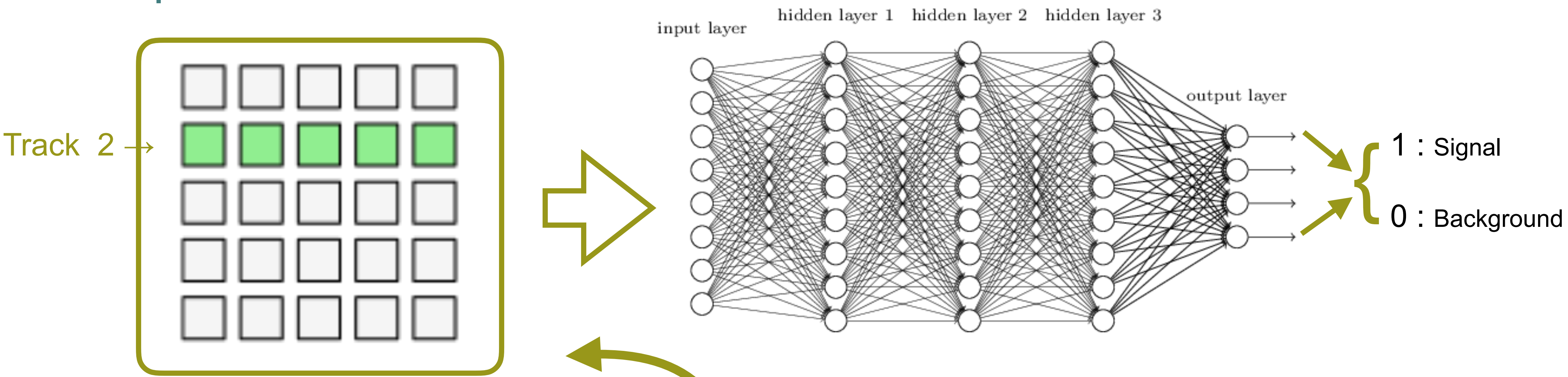


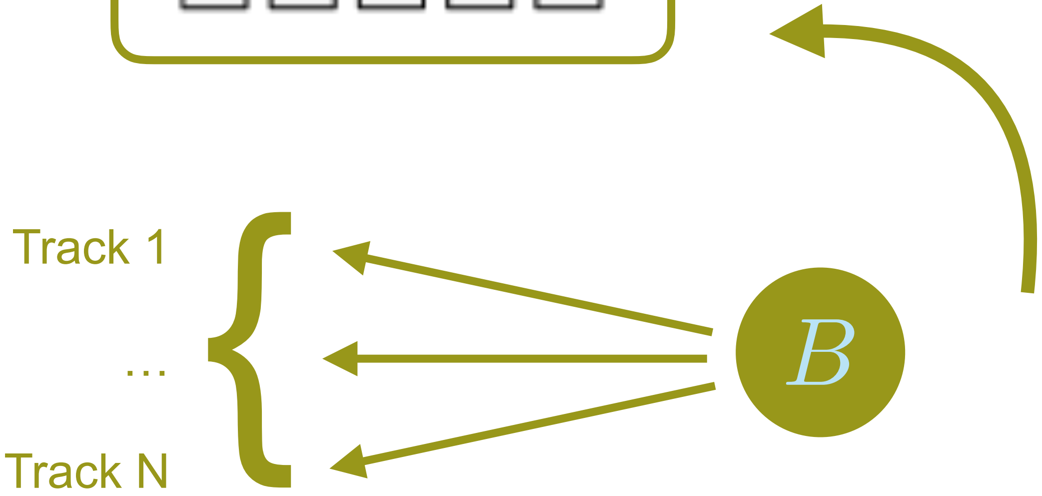
Deep Full Event Interpretation

Gather Information of each Rest of event Track
- Track parameters, Energy, Momentum
- Particle Identification Information

Deep Full Event Interpretation

Dense Deep Neural Network



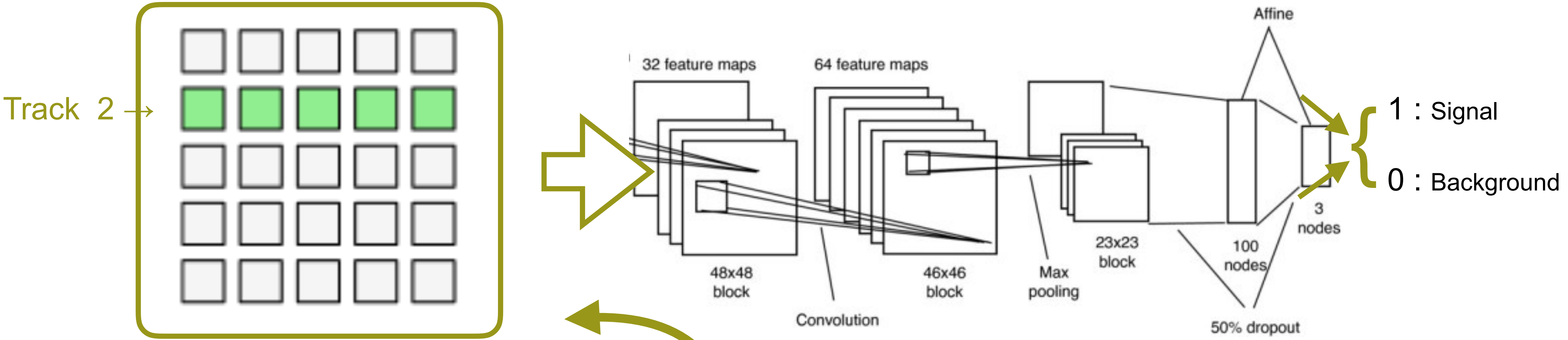


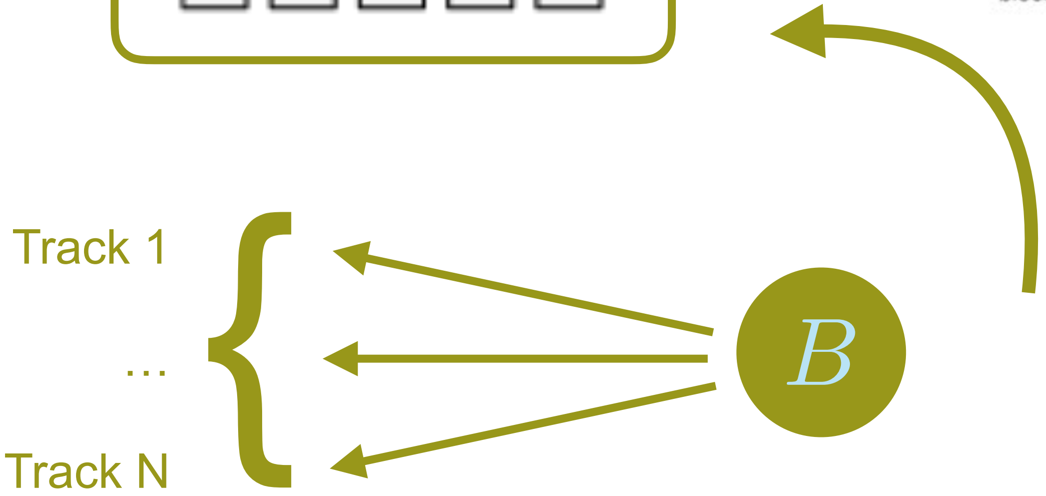
Deep Full Event Interpretation

Gather Information of each Rest of event Track
- Track parameters, Energy, Momentum
- Particle Identification Information

Deep Full Event Interpretation

Convolutional Neural Network





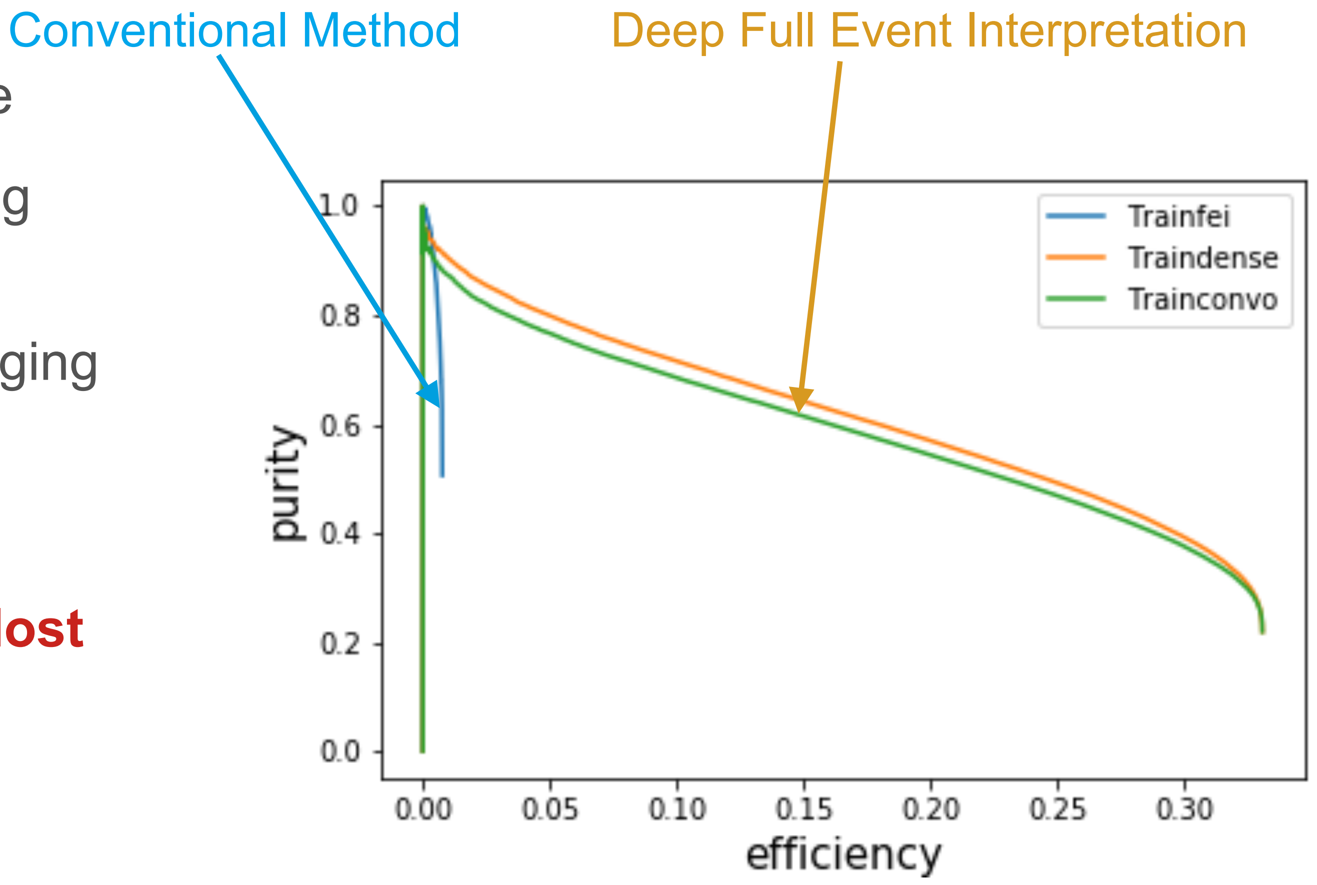
Deep Full Event Interpretation

Gather Information of each Rest of event Track
- Track parameters, Energy, Momentum
- Particle Identification Information

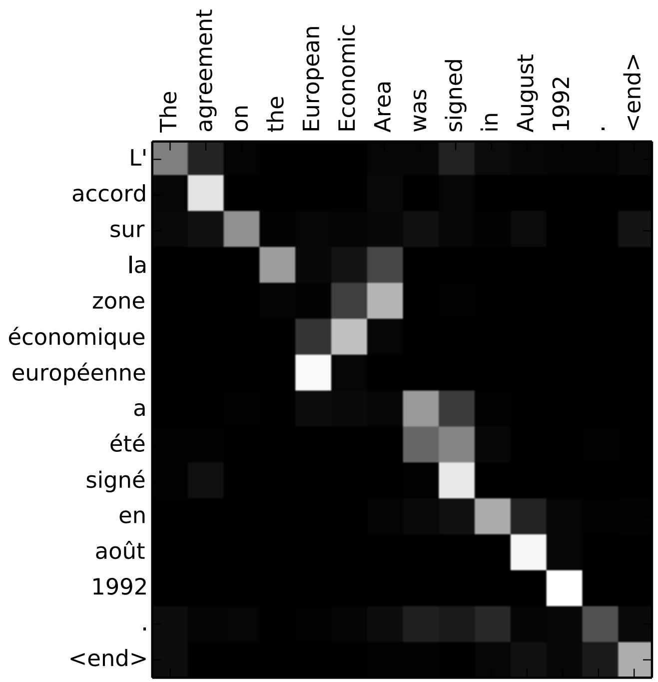
First Results

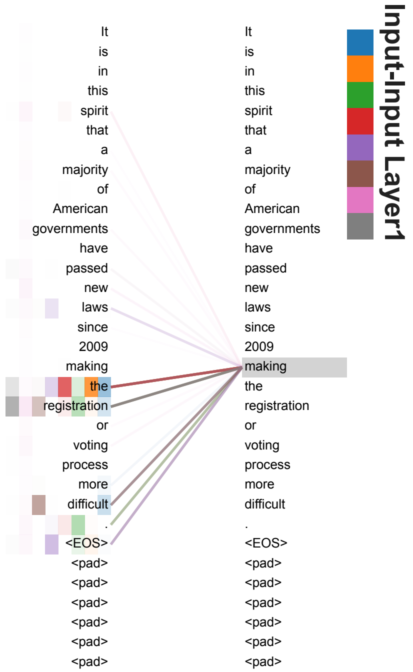
Study performed for an example rare decay of B mesons

- Result on simulated data:
  - Order of magnitude better Performance
  - Only slight loss of information on the tag candidate
- Many “golden modes” for Belle II need tagging
- The baseline of the traditional tagging methods is ~1%
- **Detailed constrain of tag-side 4-vector lost**
- Improvements of only a few percent to the method can increase the statistics corresponding to years of data taking!



Excursion: Attention and Transformer Networks

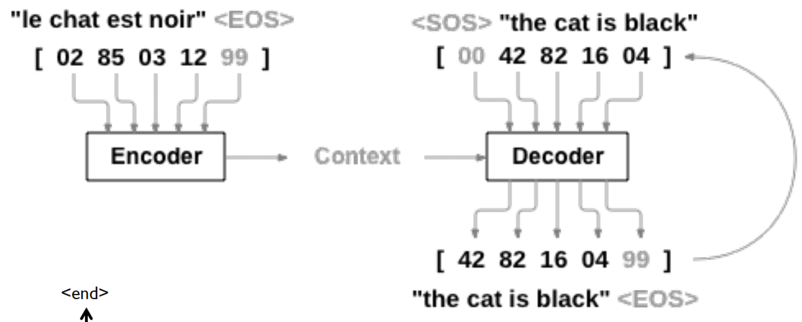




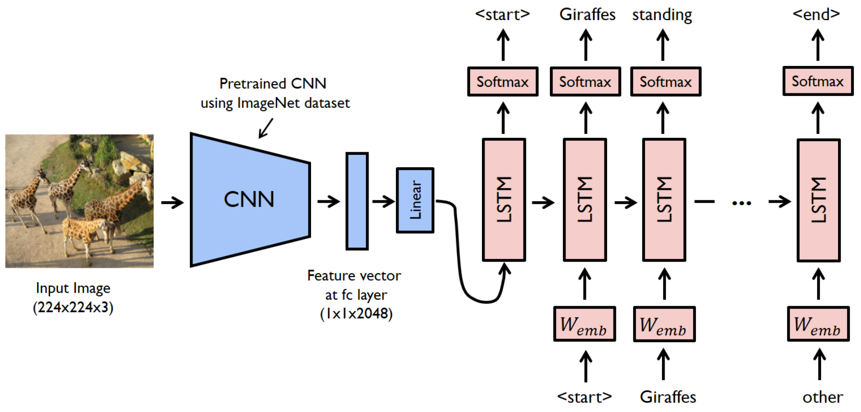
Applications of RNNs

Excursion: Attention and Transformer Networks

- Translation



- Caption Generation



- Sentiment analysis

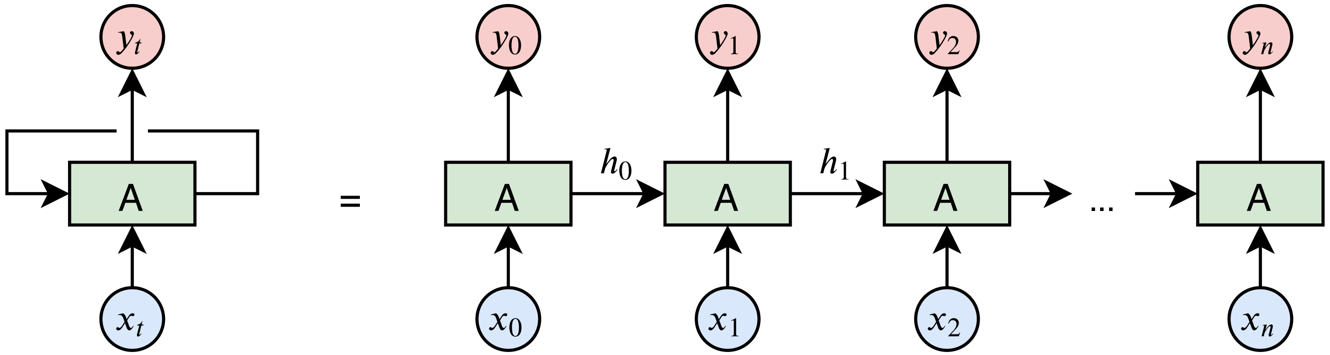
- But: information of whole sentence stored in fixed-length context vector

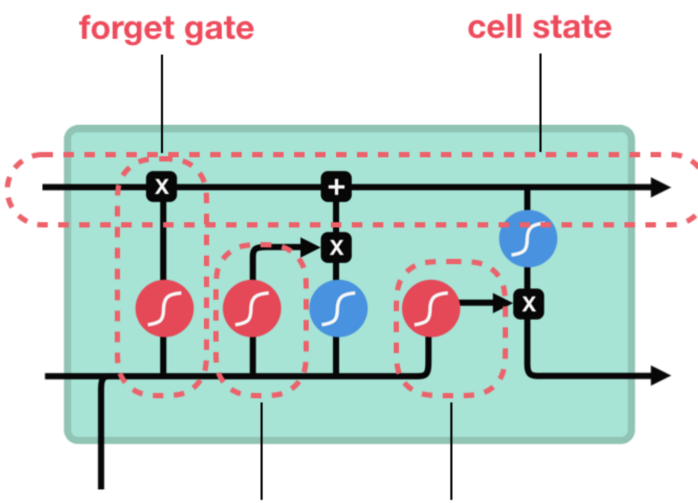
Recurrent Neural Networks (RNN)

Excursion: Attention and Transformer Networks

internal state $\boldsymbol{h_t} = f(W_{hx}\boldsymbol{x_t} + W_{hh}\boldsymbol{h_{t-1}} + \boldsymbol{b_h})$

output $\boldsymbol{y_t} = g(W_{yh}\boldsymbol{h_t} + \boldsymbol{b_y})$





(a) LSTM

- internal state / ”memory” $\boldsymbol{h_n}$ $\rightarrow$ learn context

BUT require sequence

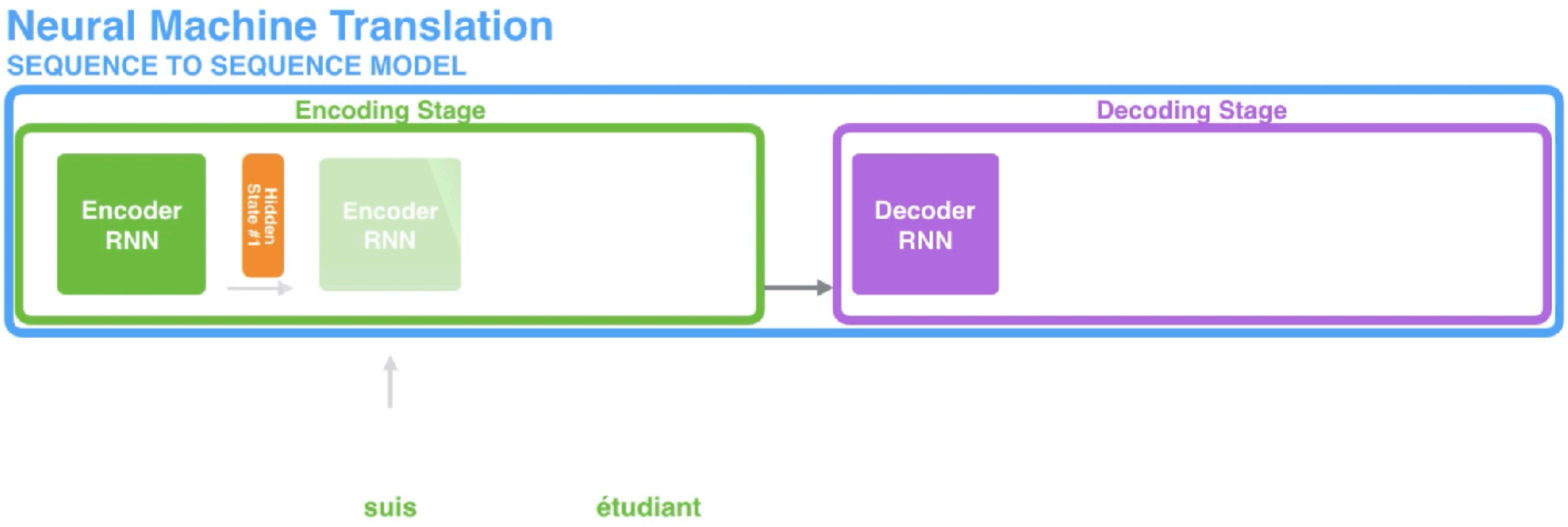
slow computation

long-range dependencies are tricky

Credit: Tobias Böckh

Recurrent Neural Networks (RNN)

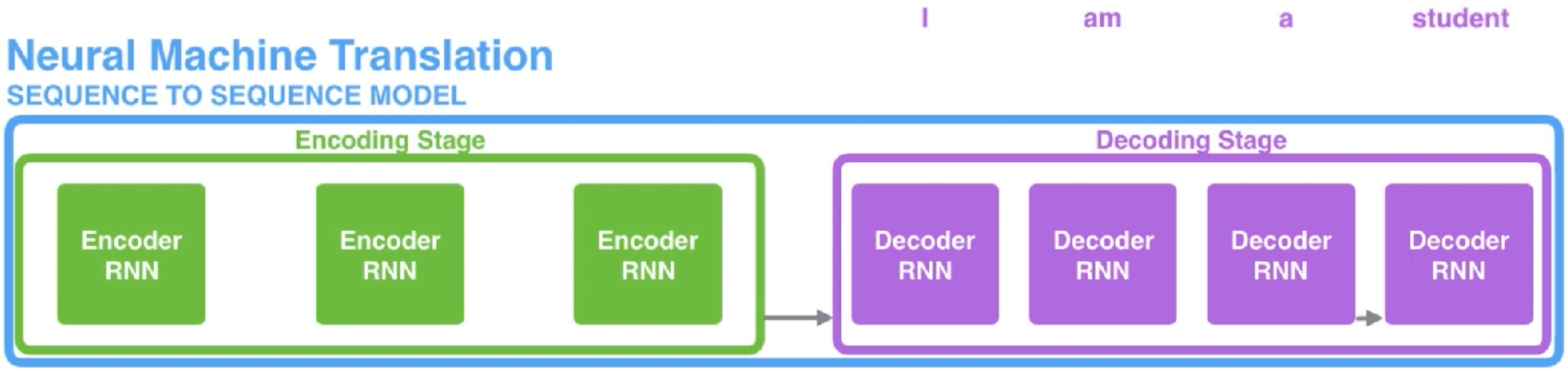
Sequence to Sequence



Credit: Jay Alammar

Recurrent Neural Networks (RNN)

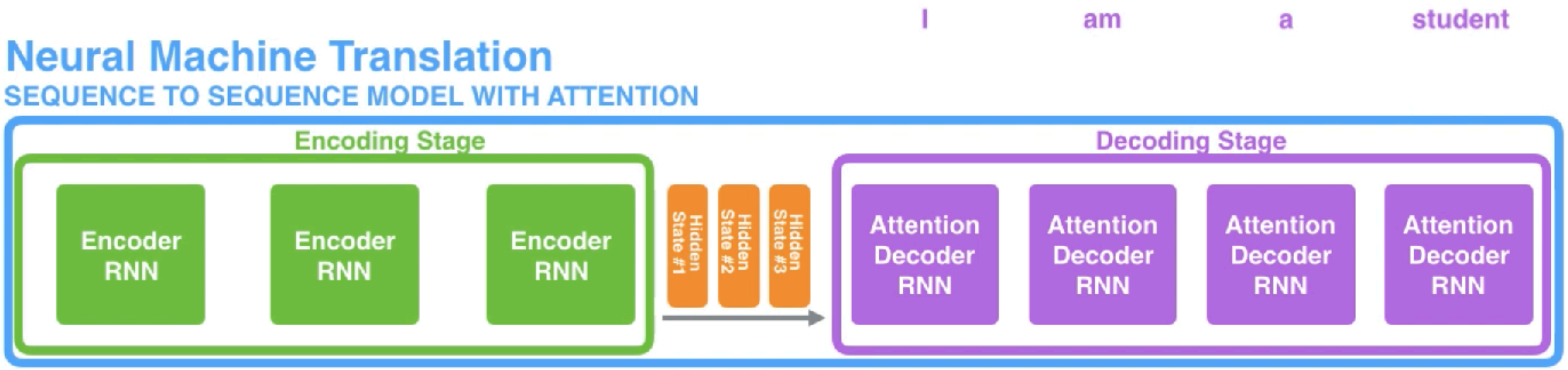
Sequence to Sequence



Credit: Jay Alammar

Adding Attention

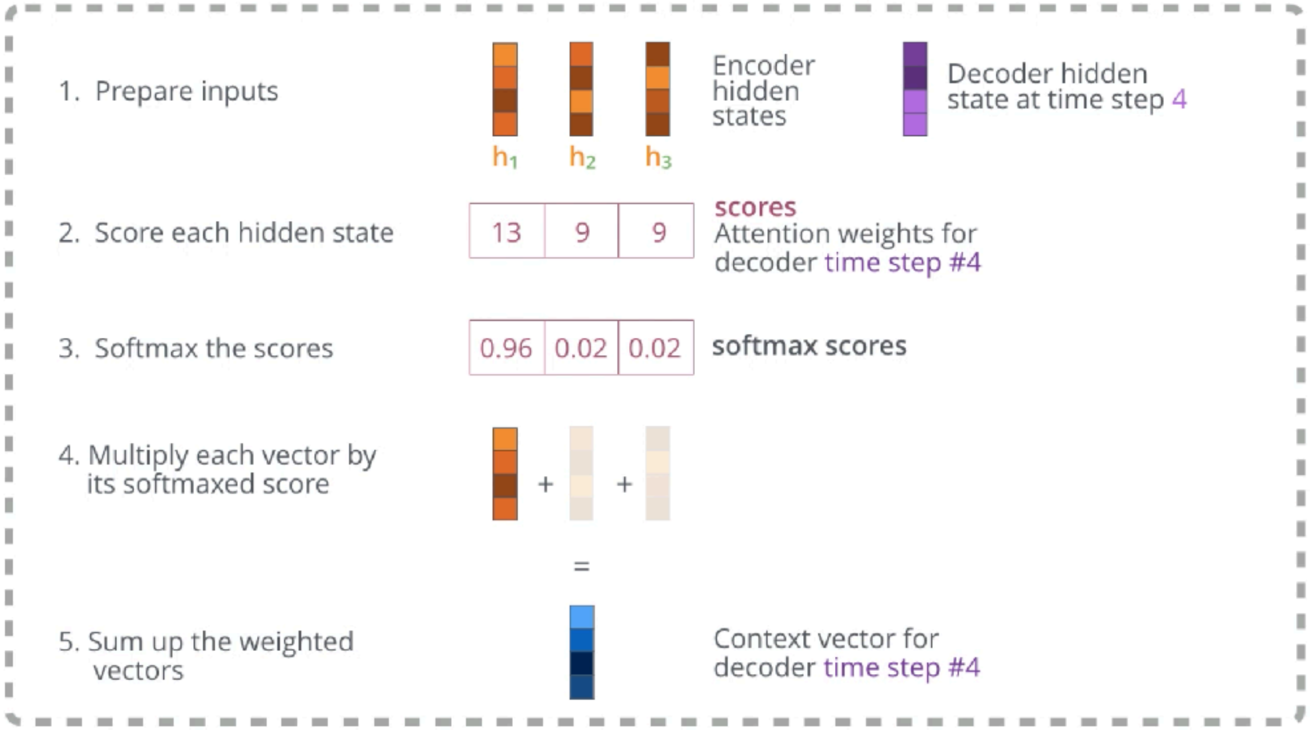
Attention is all you need



Credit: Jay Alammar

Adding Attention

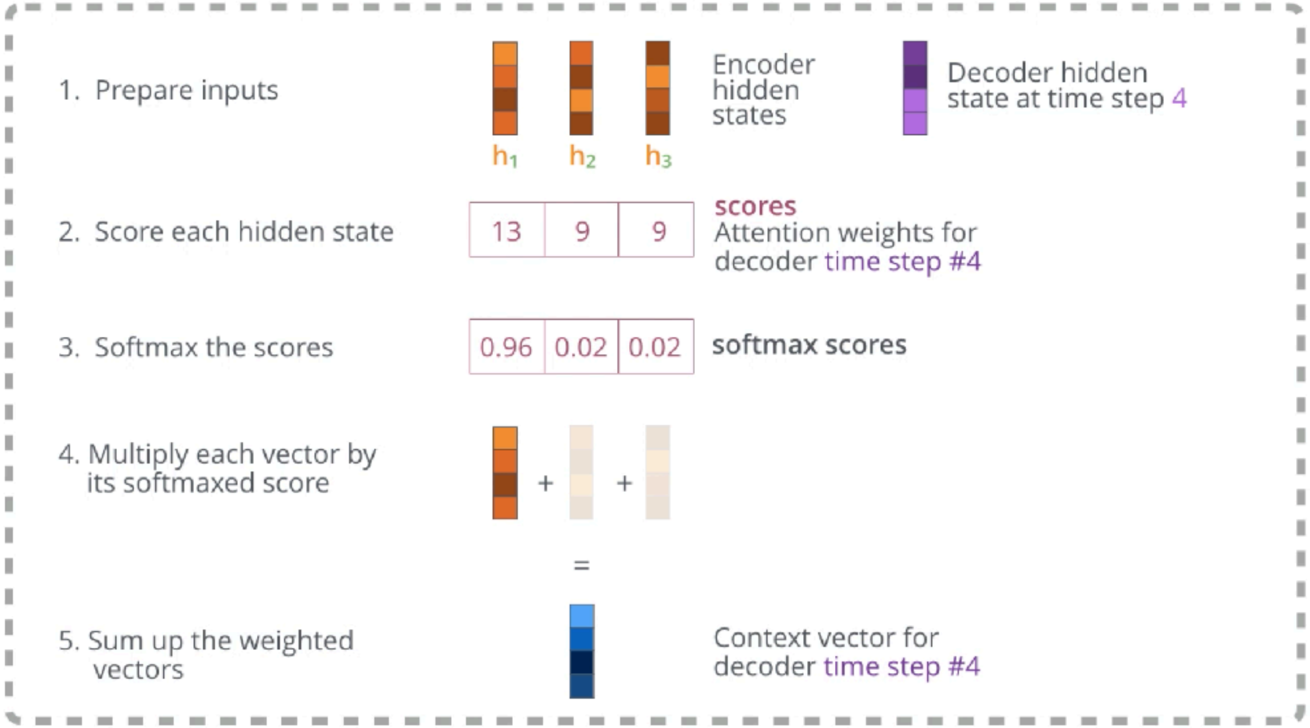
Attention is all you need

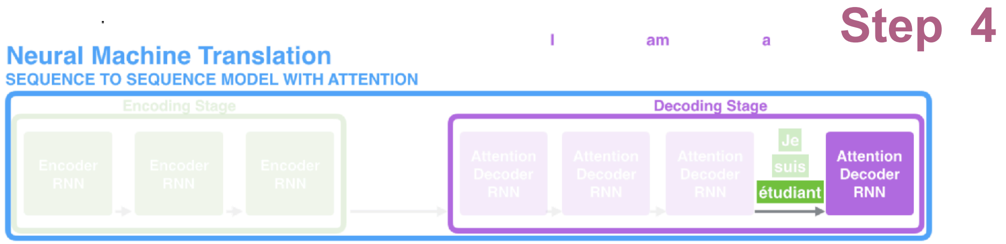


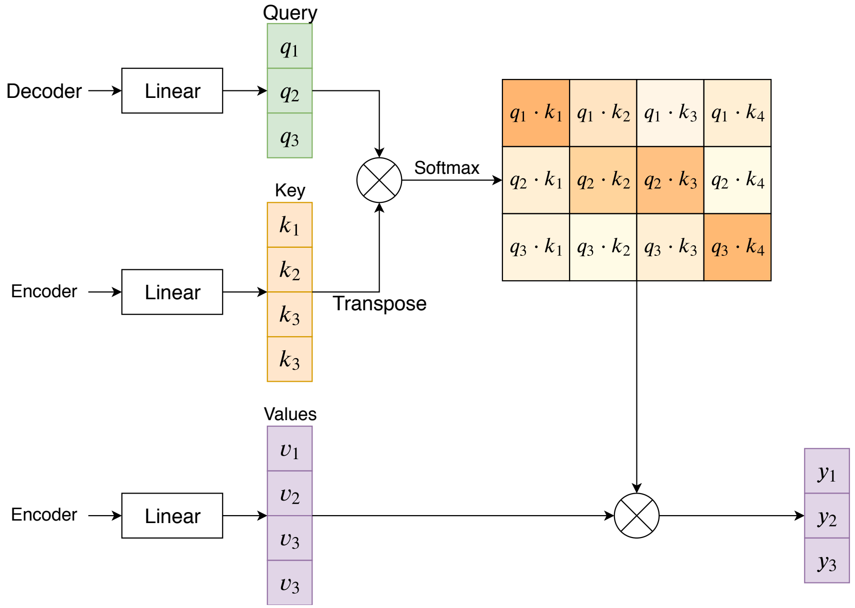
[Attention is All You Need]

Adding Attention

Attention is all you need





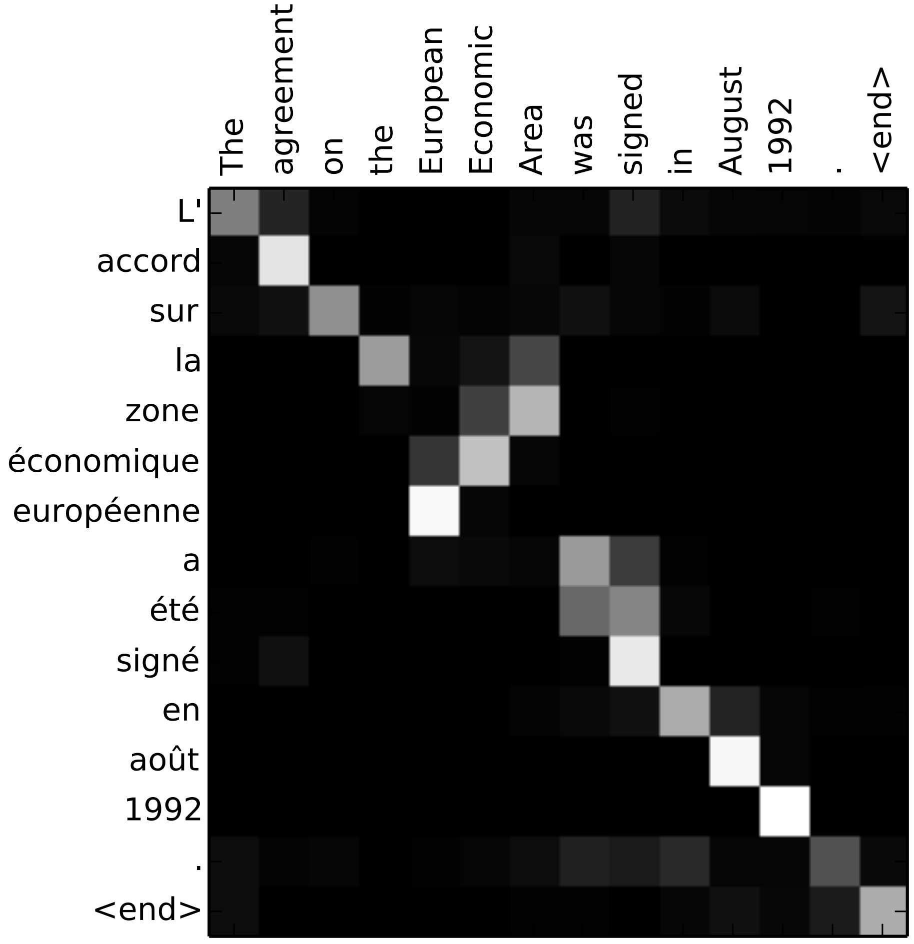


$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V$$

Dot product Attention:
Luong, Pham, and Manning 2015

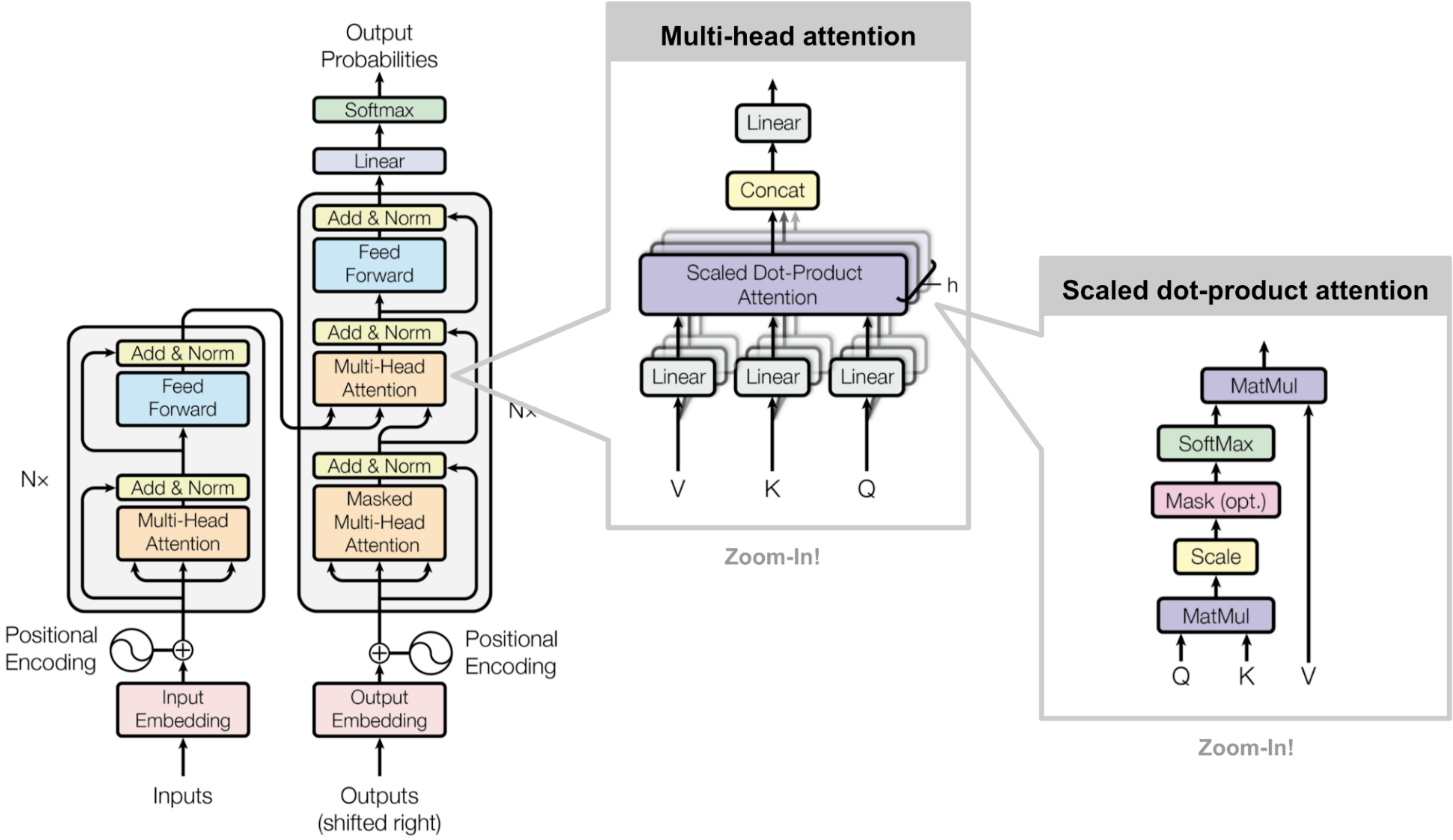
Attention is All You Need

Visualisation



Transformer

Vaswani et al. 2017

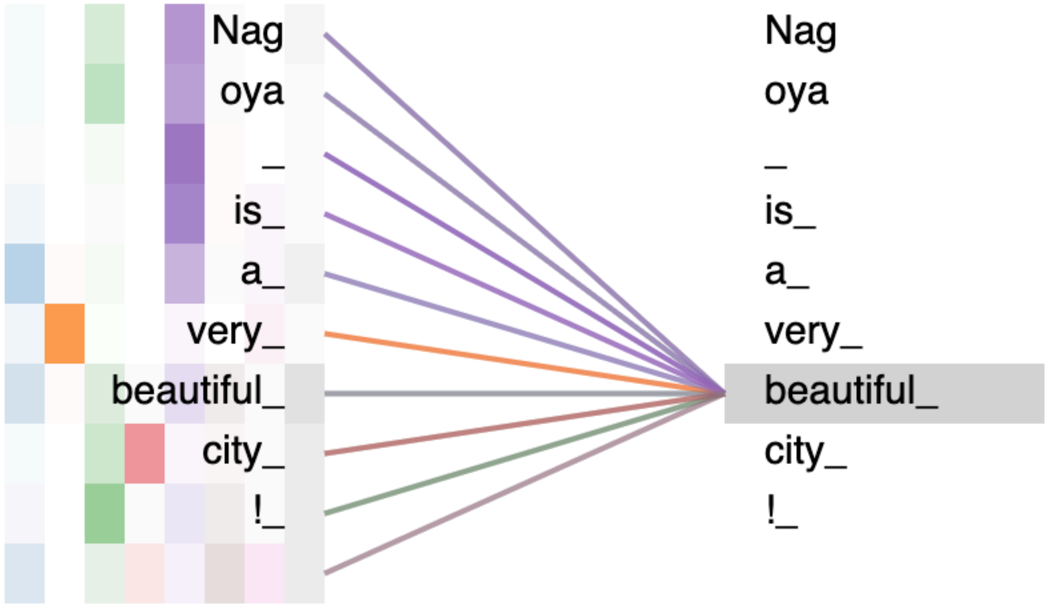


Credit: Jay Alammar

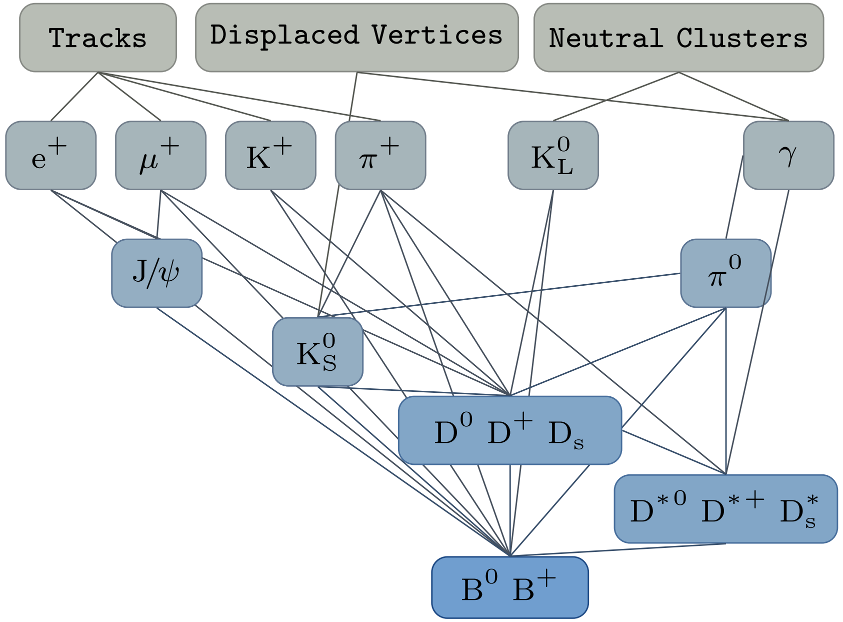
Transformer

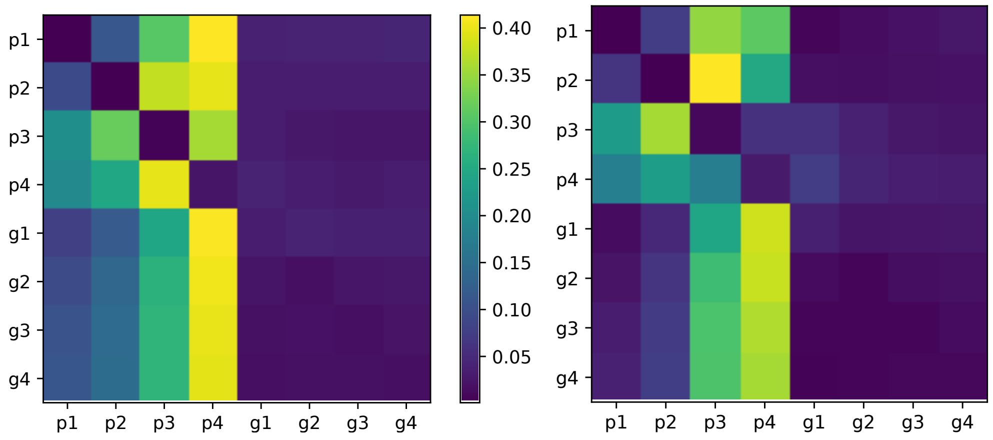
Vaswani et al. 2017

- Multiple stacks of attention
- Implements “Self-Attention”
- Handel long range relations
- Computationally efficient
- State of the art performance



Transformer Applications in Belle II

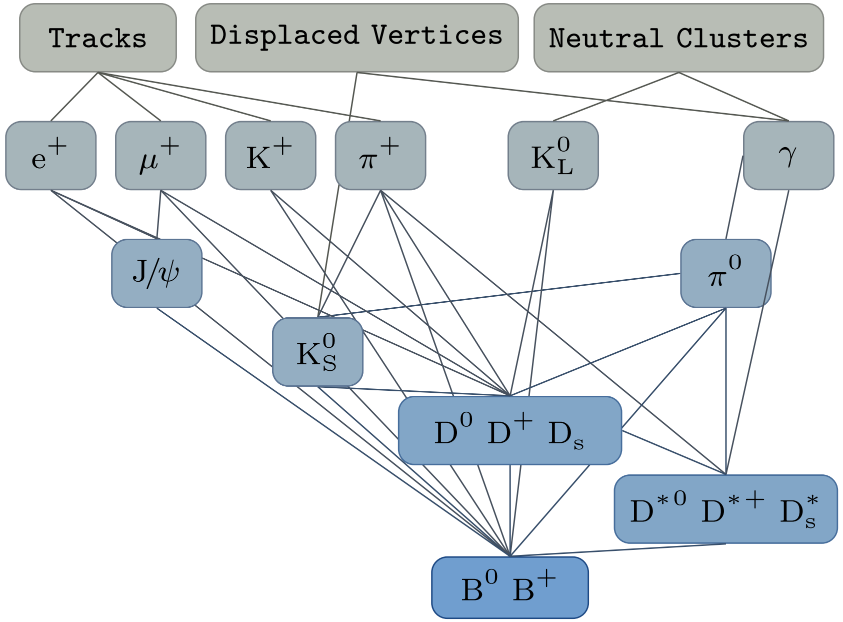




Deep FEI Developments

Future work

- Learn generic decay reconstruction by example
- Currently FEI contains hard-coded sub-decays
- Utilising self-attention maps to cluster particles
- Implemented B → D(→ K ππ0)π reconstruction with transformer network
- Utilising permutation invariant loss function: Kuhn-Munkres/Hungarian algorithm (shipped with SciPy)

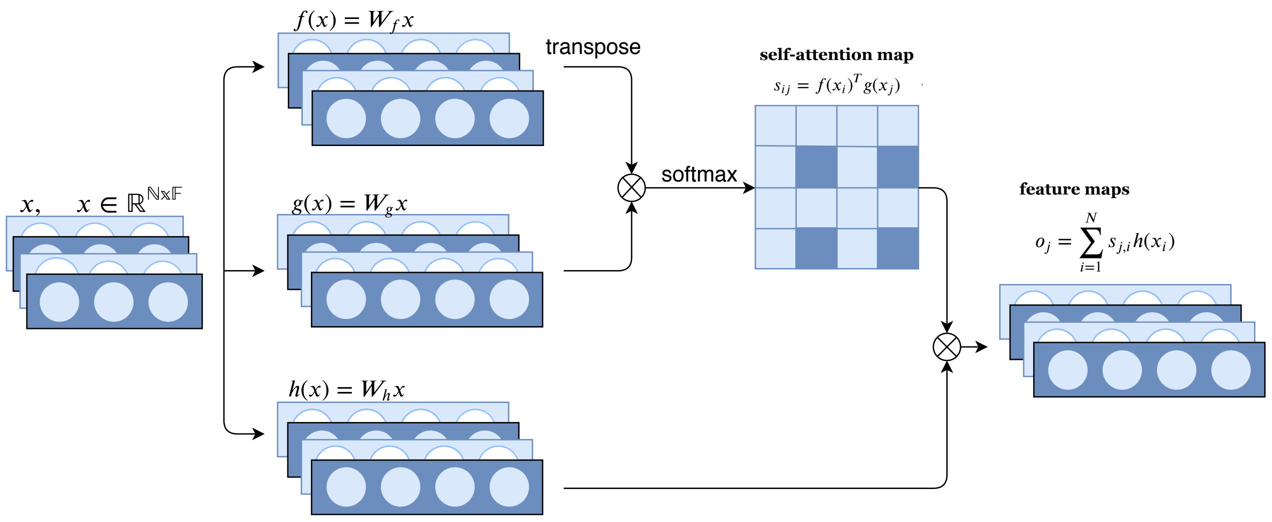


Attention Is All You Need

Deep FEI Developments

Future work

- Learn generic decay reconstruction by example
- Currently FEI contains hard-coded sub-decays
- Utilising self-attention maps to cluster particles
- Implemented B → D(→ K ππ0)π reconstruction with transformer network
- Utilising permutation invariant loss function: Kuhn-Munkres/Hungarian algorithm (shipped with SciPy)

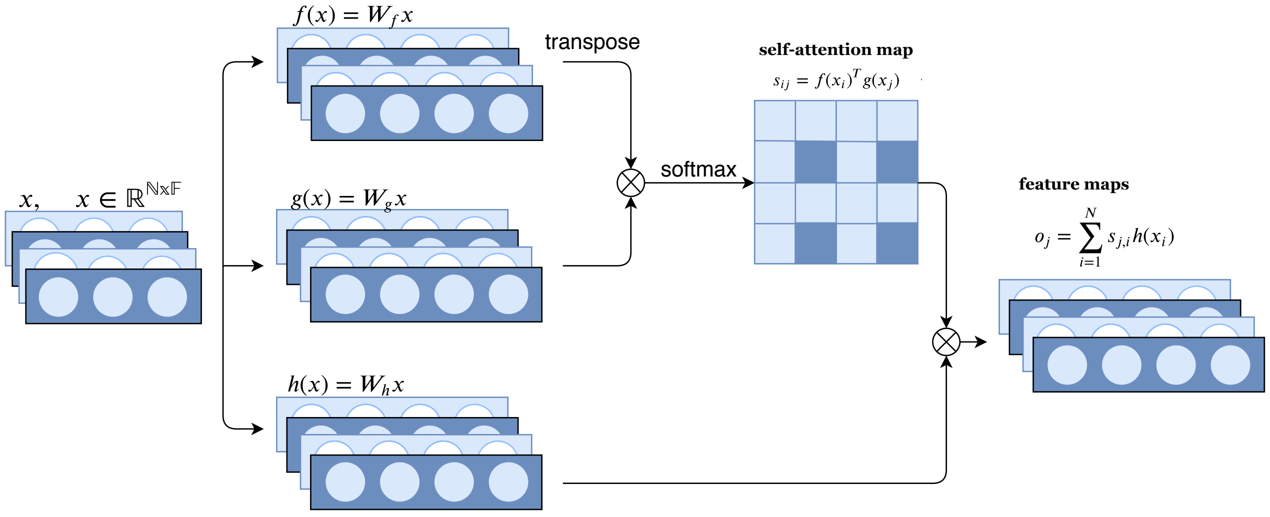


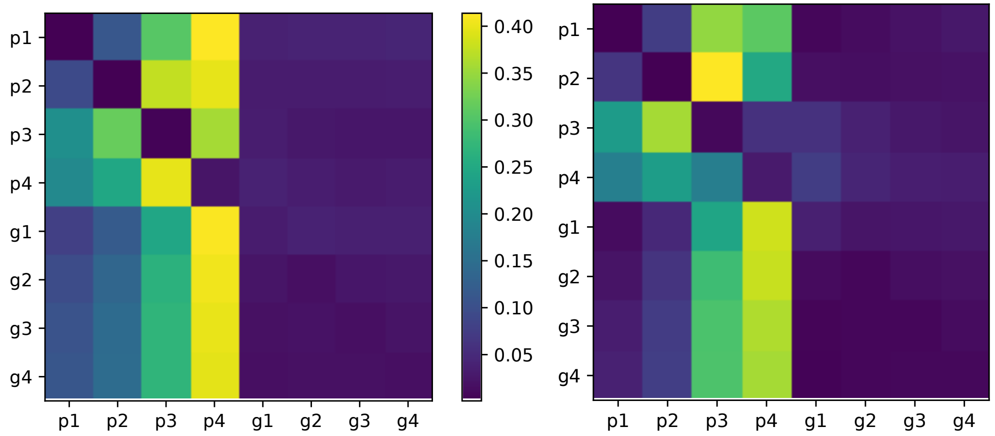
Attention Is All You Need

Deep FEI Developments

Future work

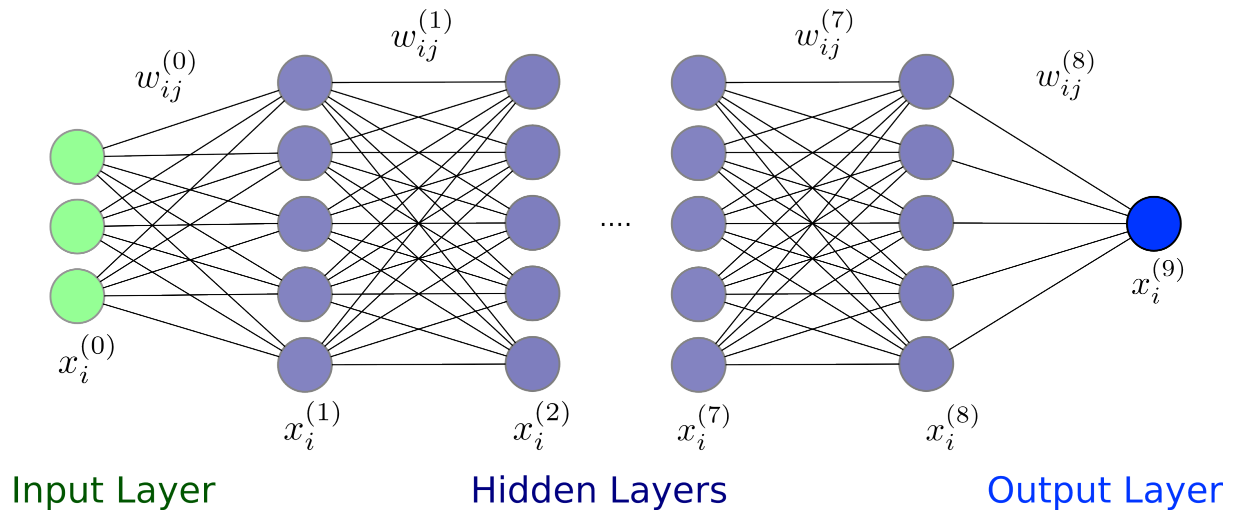
- Learn generic decay reconstruction by example
- Currently FEI contains hard-coded sub-decays
- Utilising self-attention maps to cluster particles
- Implemented B → D(→ K ππ0)π reconstruction with transformer network
- Utilising permutation invariant loss function: Kuhn-Munkres/Hungarian algorithm (shipped with SciPy)





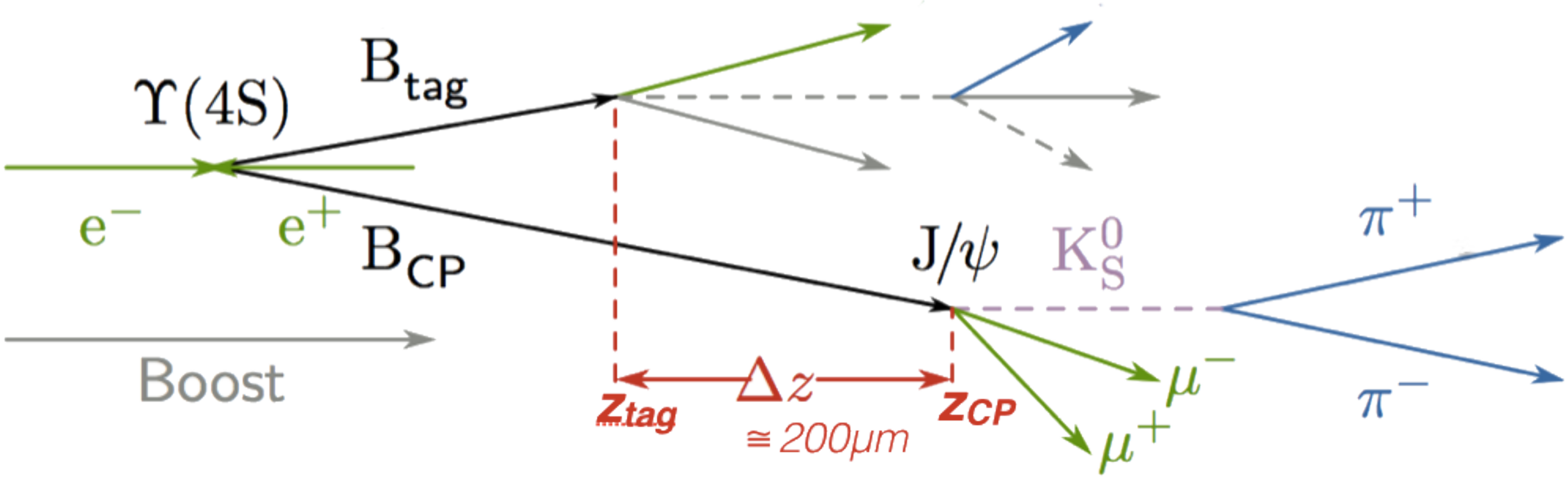
Attention Is All You Need

Other Deep NN Applications at Belle II



DEEP Flavour tagger

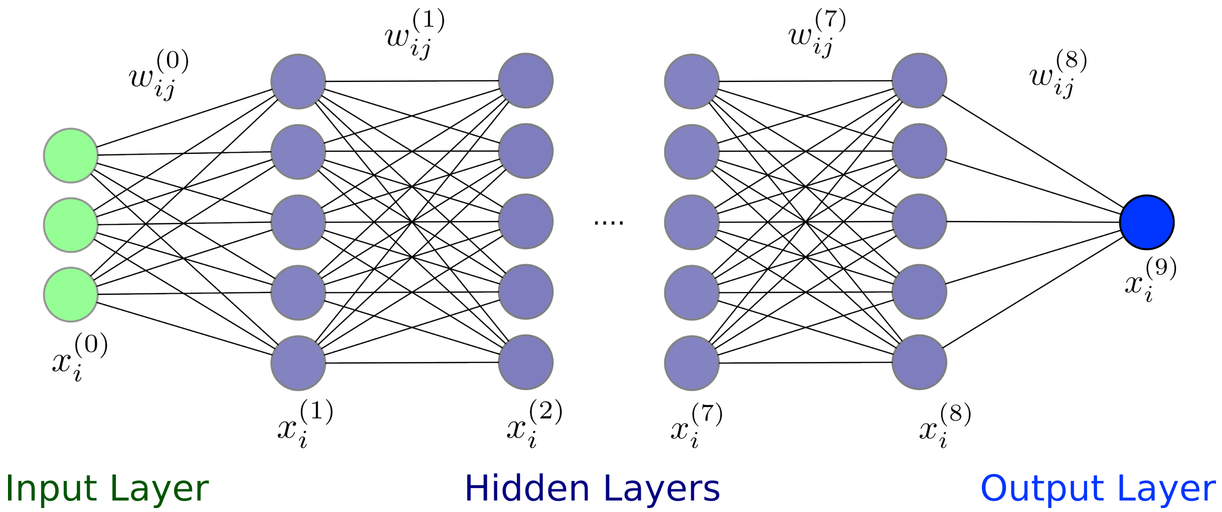
Flavour Tagging

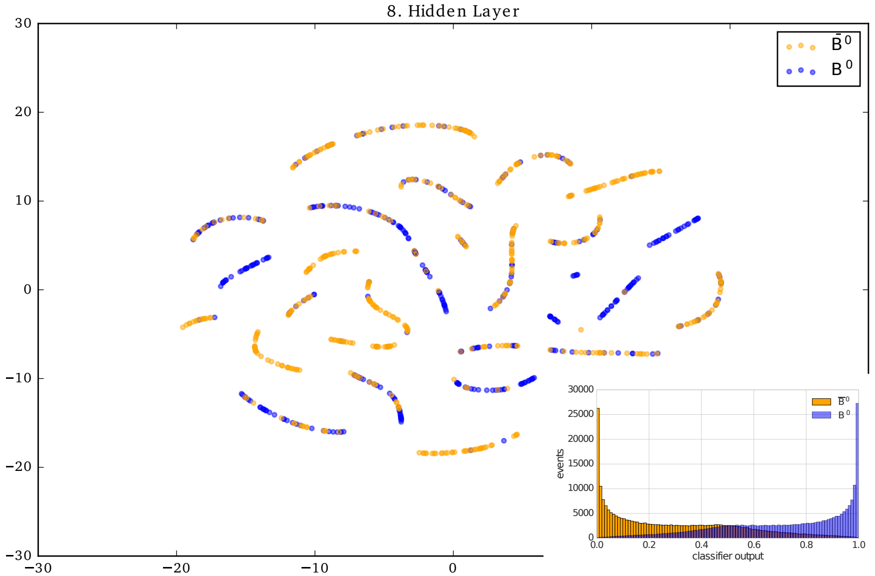


- Quantum Entanglement:
  - Neutral B mesons are entangled in flavour with their production
  - With mixing, the possible outcomes are $B\bar{B}, BB, \bar{B}\bar{B}$

DEEP Flavour tagger

Flavour Tagging



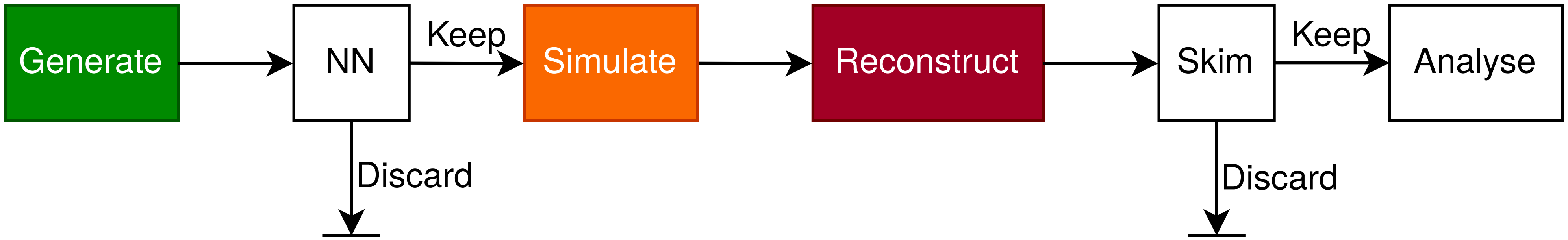


- Deep neural network approach using track information in Input
- Simple approach already outperforms “classical” method

		Category Based	Deep Neural Network
Belle (MC)	$J/\Psi K_S^0$	0.293 ± 0.01 [1]	0.3442 ± 0.0009

Credit: Jochen Gemler

Selective background Monte Carlo simulation at Belle II

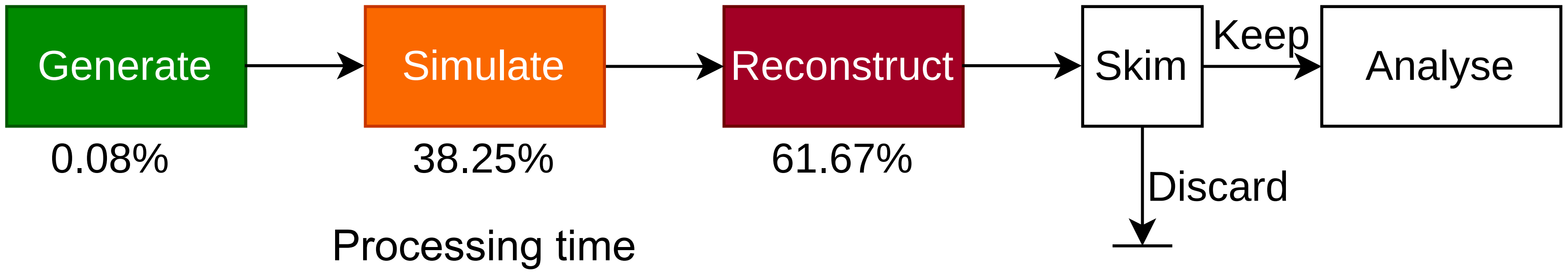


Problem

Selective background Monte Carlo simulation at Belle II

- Approach at Belle:
  - Background MC ≈ 10 × data
- Infeasible at Belle II → still require high statistics

- Skims
  - Physics working-group specific datasets (26)
  - General selections applied to discard trivial backgrounds
  - Retain O(0.1–10%) of full dataset

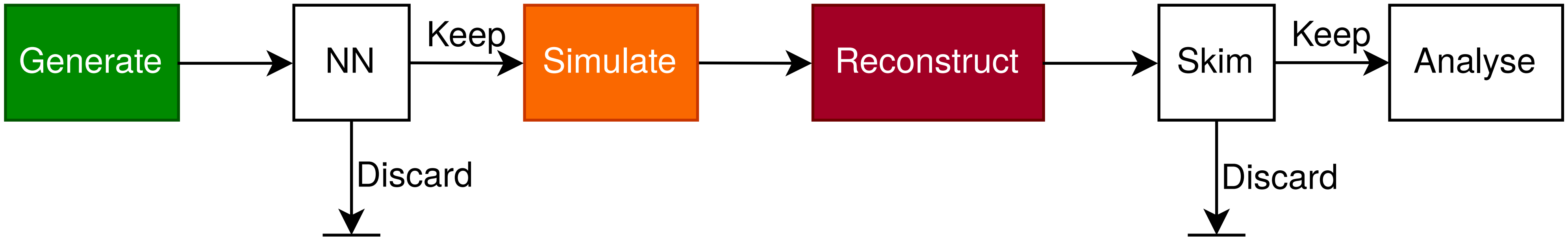


Problem

Selective background Monte Carlo simulation at Belle II

- Approach at Belle:
  - Background MC ≈ 10 × data
- Infeasible at Belle II → still require high statistics

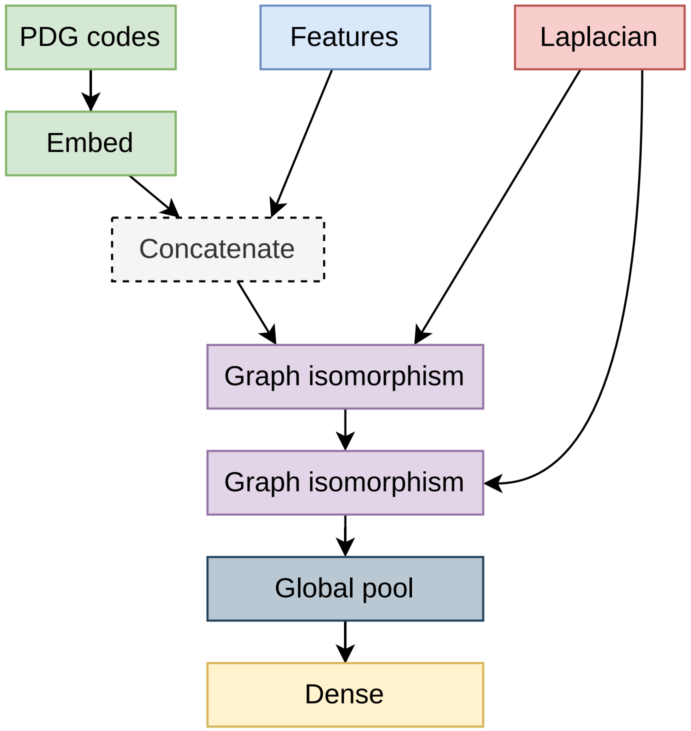
- Skims
  - Physics working-group specific datasets (26)
  - General selections applied to discard trivial backgrounds
  - Retain O(0.1–10%) of full dataset

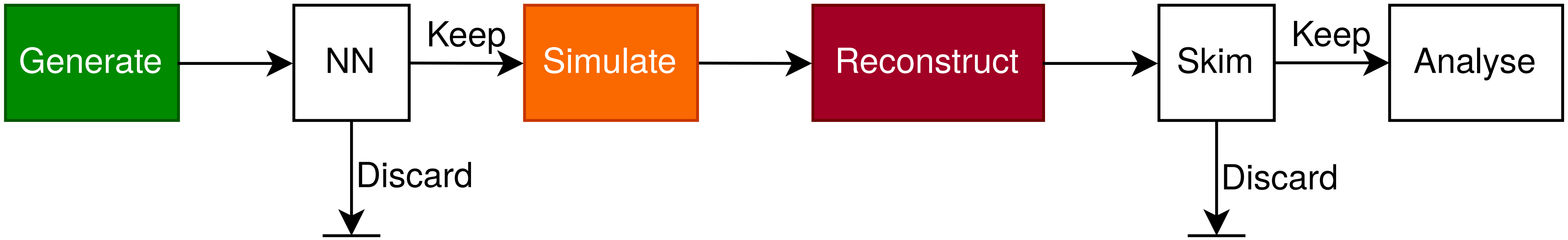


Selective Event Reconstruction

Selective background Monte Carlo simulation at Belle II

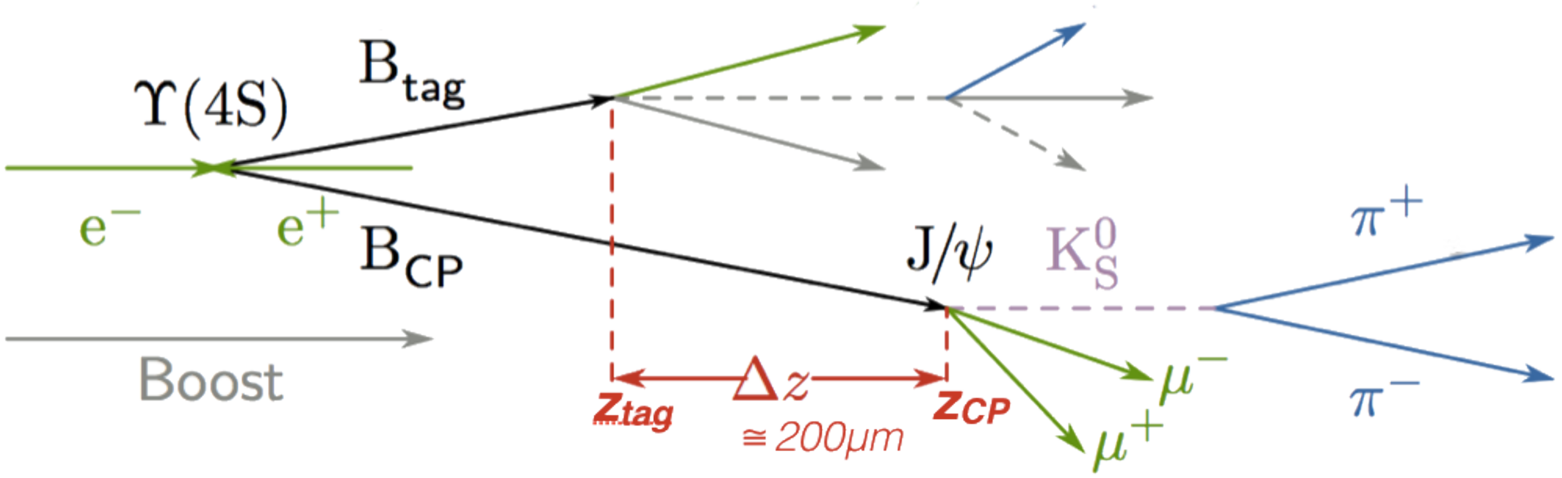
- Proposed Solution:
  - Let ML algorithm decide before time intensive steps
- Use Graph NN for classification





Exploring Graph Neural Networks

Selective background Monte Carlo simulation at Belle II

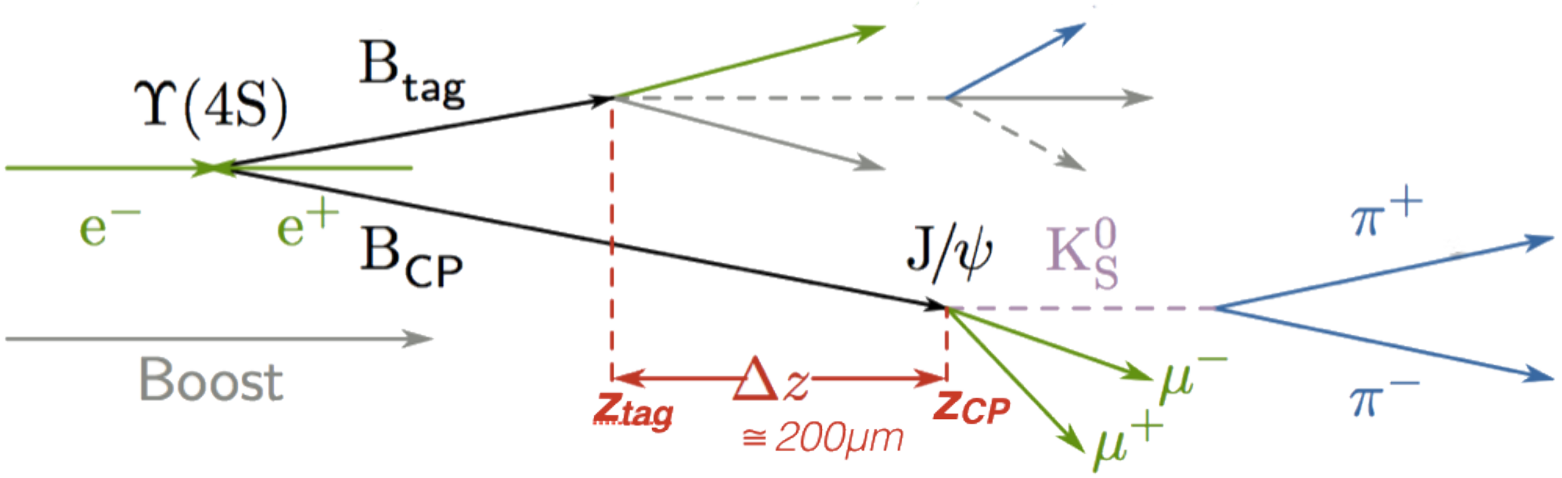


Graph terminology

- Nodes = Particles
- Node attributes = Particle properties
- Edges = Parent-daughter relations (decays)
- Graph type = Tree

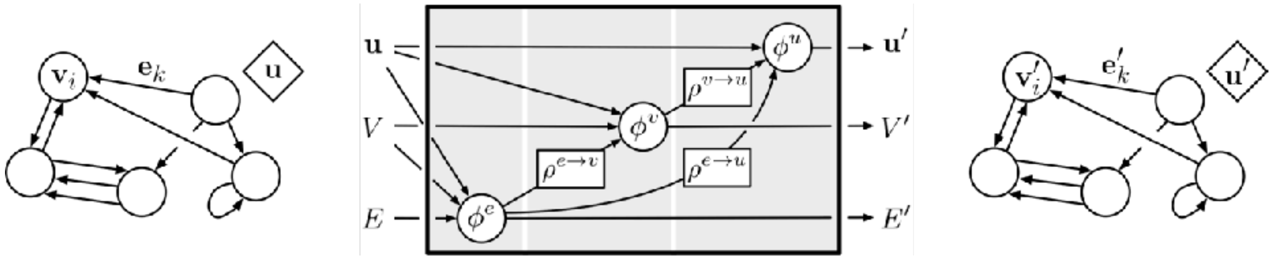
Exploring Graph Neural Networks

Selective background Monte Carlo simulation at Belle II



Graph terminology

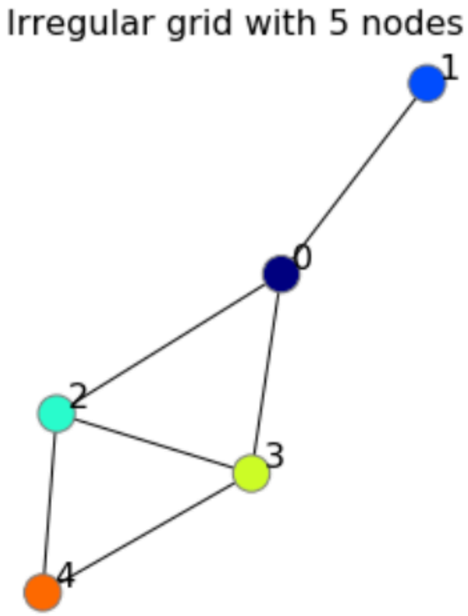
- Nodes = Particles
- Node attributes = Particle properties
- Edges = Parent-daughter relations (decays)
- Graph type = Tree

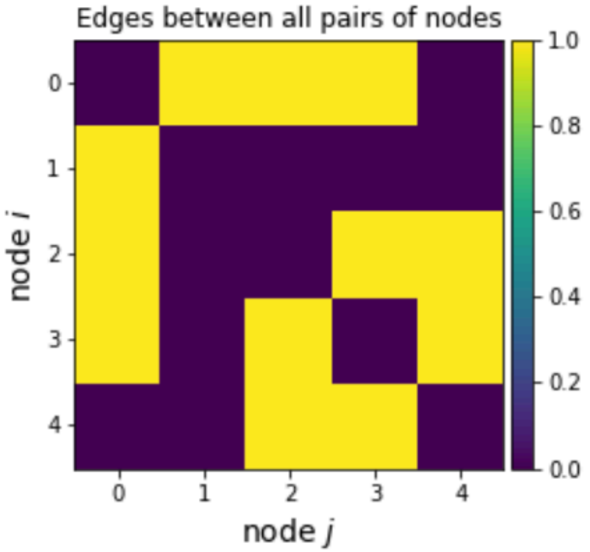


https://github.com/deepmind/graph_nets

Exploring Graph Neural Networks

Selective background Monte Carlo simulation at Belle II





A = adjacency matrix

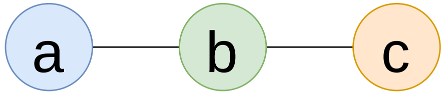
$$X^{(l+1)} = AX^{(l)}W^{(l)}$$

X = Feature matrix
W = Weight matrix

Original Graph Convolutional Networks (GCN)

Propagation rule of layer activations $H^{(l)}$

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}\right)$$



$$\begin{aligned} H^{(0)} &= X \\ \tilde{A} &= A + I_N \\ \tilde{D}_{ii} &= \textstyle\sum_j \tilde{A}_{ij} \end{aligned}$$

$$\tilde{A}^{N\times N} = A + I = \begin{array}{c c} & \begin{array}{ccc} a & b & c \end{array} \\ \begin{array}{c} a \\ b \\ c \end{array} & \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \end{array}$$

Thomas N. Kipf, Max Welling, Semi-Supervised Classification with Graph Convolutional Networks (ICLR 2017)

Dataset

$\sim 300,000$ particle collision events with binary classification labels:

- Hadronic $B+$ meson reconstruction ($\sim 5\%$)
- Time-dependent *CP* violation ($\sim 0.2\%$)

Graph terminology

- Nodes = Particles
- Node attributes = Particle properties
- Edges = Parent-daughter relations (decays)
- Graph type = Tree

$\Upsilon(4S)$ (300553)
  $\bar{B}^0$ (-511)
    J/ψ (443)
      μ^+ (-13)
      μ^- (13)
    K^0_S (310)
      pi^- (-211)
      pi^+ (211)
  B^0 (511)
    $\bar{D}^0$ (-421)
      pi^- (-211)
      K^+ (321)
    pi^- (-211)
    μ^+ (-13)
    ν_μ (14)

Feature	Definition
PDG code	Identifier of particle type and charge.
Mother PDG code	Particle parent PDG code.
Mass	Particle mass in GeV/c^2.
Charge	Electric charge of the particle.
Energy	Particle energy in GeV.
Momentum	Three momentum of the particle in Gev/c.
Production time	Production time in ns relative to $\Upsilon(4S)$ production.
Production vertex	Coordinates of particle production vertex.
Status bit	Bitmask representing MC production conditions.

Graph Isomorphism Network

Node N update rule of layer ℓ (Red = trainable):

$$N^{(\ell+1)} = \mathrm{MLP}^{(\ell)}\left(W_p^{(\ell)} N_p^{(\ell)} + W^{(\ell)} N^{(\ell)} + W_d^{(\ell)} \sum_{\text{daughters}} N_d^{(\ell)} \right)$$

Intuition: **Create representation of node considering its neighbours**

- Custom weights for parent (W_p), node (W), daughters (W_d)
- Independent of daughter ordering
- Normalise adjacency matrix
  - Prevent over-representation in high multiplicity decays

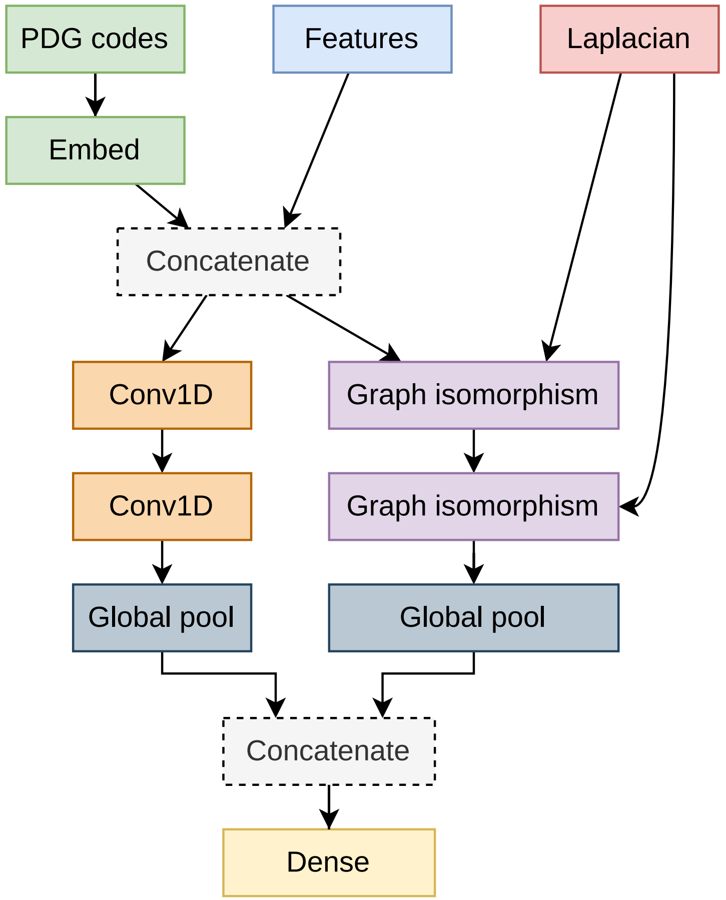
Normalised Laplacian

$\tilde{A} = A + I_N$
$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$
$\tilde{L} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$

Special case of:
K. Xu, W. Hu, J. Leskovec, S. Jegelka, How Powerful are Graph Neural Networks? (CoRR 2018)

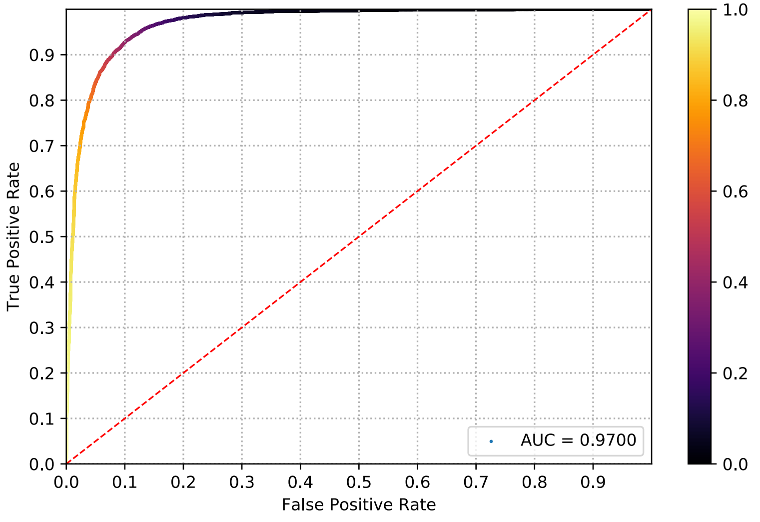
Training

- Train on 250k events (validate on 10%)
- Test on 50k independent events
- Batch normalisation, dropout, class weights, early stopping, reduce LR on plateau, model checkpoint (save only best), . . .
- Additional convolutional 1D for full reconstruction dataset

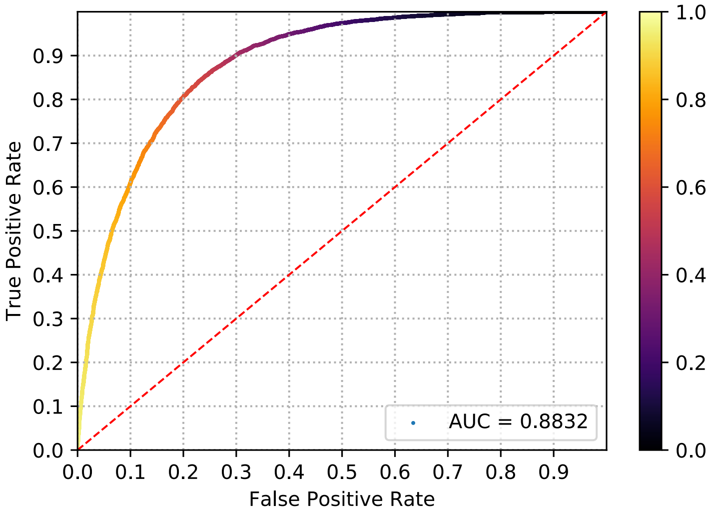


Training

- Train on 250k events (validate on 10%)
- Test on 50k independent events
- Batch normalisation, dropout, class weights, early stopping, reduce LR on plateau, model checkpoint (save only best), . . .
- Additional convolutional 1D for full reconstruction dataset



(a) TDCPV



Credit: James Kahn

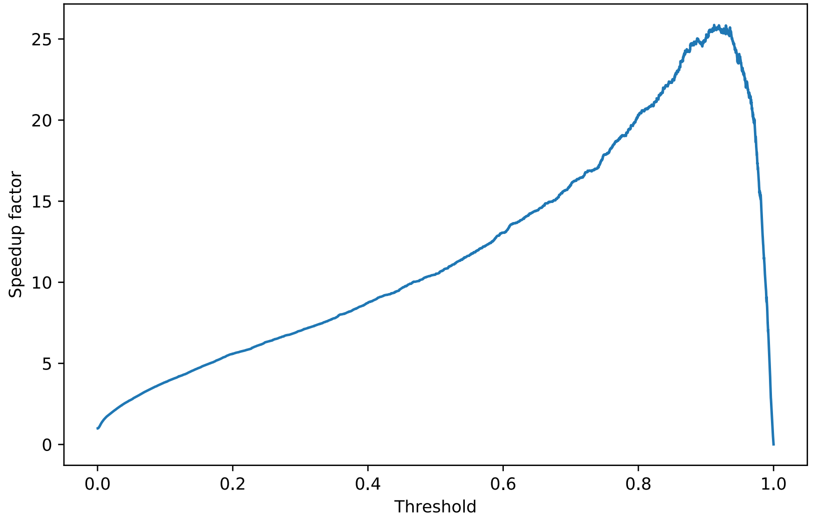
Bias check

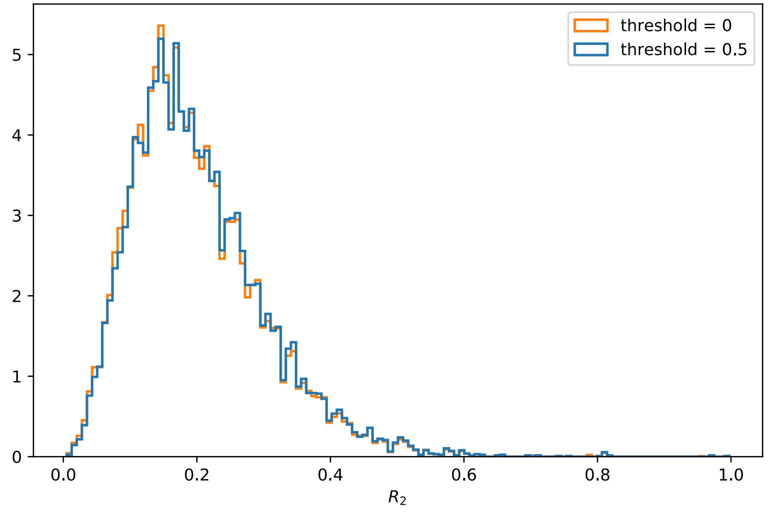
Compare event-level kinematics:

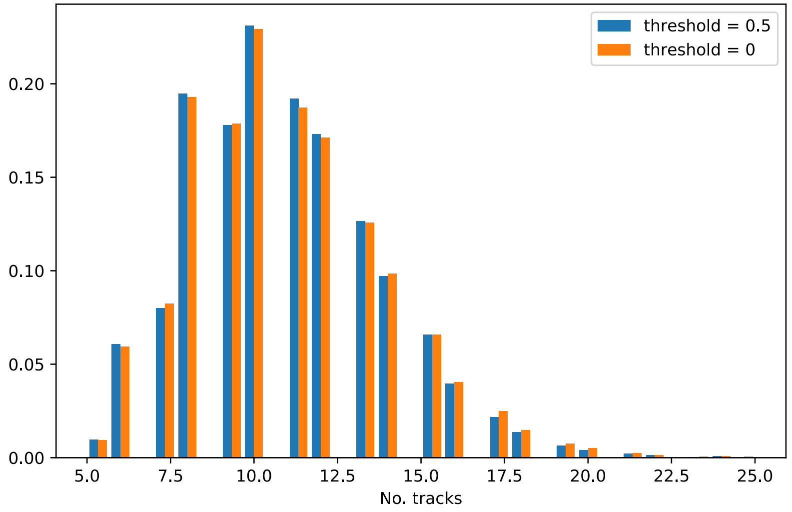
- Pass skim = True
- Pass skim and NN = True positive

Kullback-Leibler divergence of Q from P:

$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log \left(\frac{Q(x)}{P(x)} \right)$$





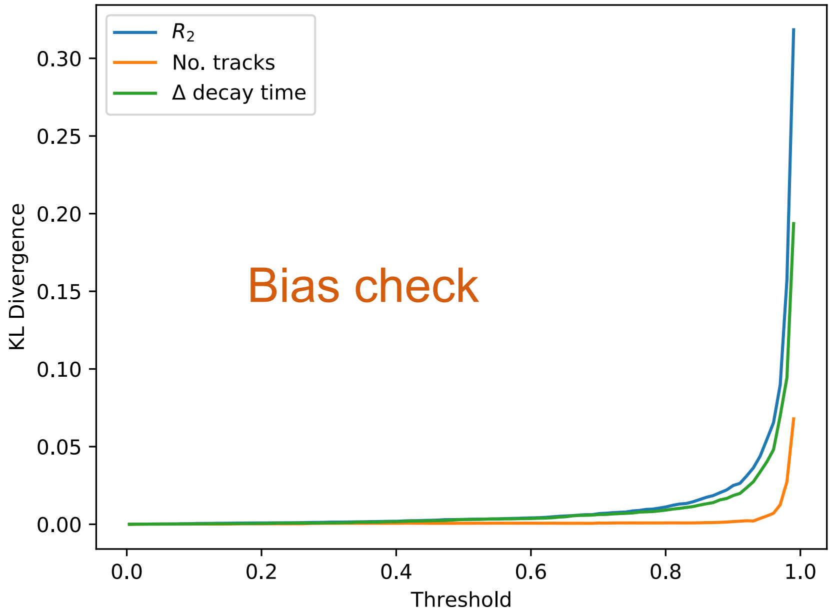


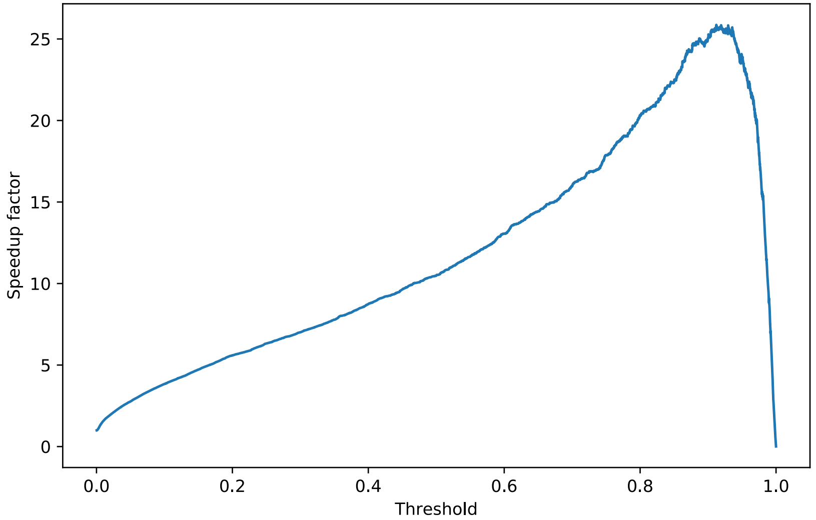
Credit: James Kahn

Selective Skim Performance

Selective background Monte Carlo simulation at Belle II

- Graph NN classification
  - > 90% accuracy on test skim
  - Orders of magnitude speed-up possible
  - Inspect event-level kinematics for bias (Kullback-Leibler divergence)
  - Presented at [CHEP]() (James Kahn)

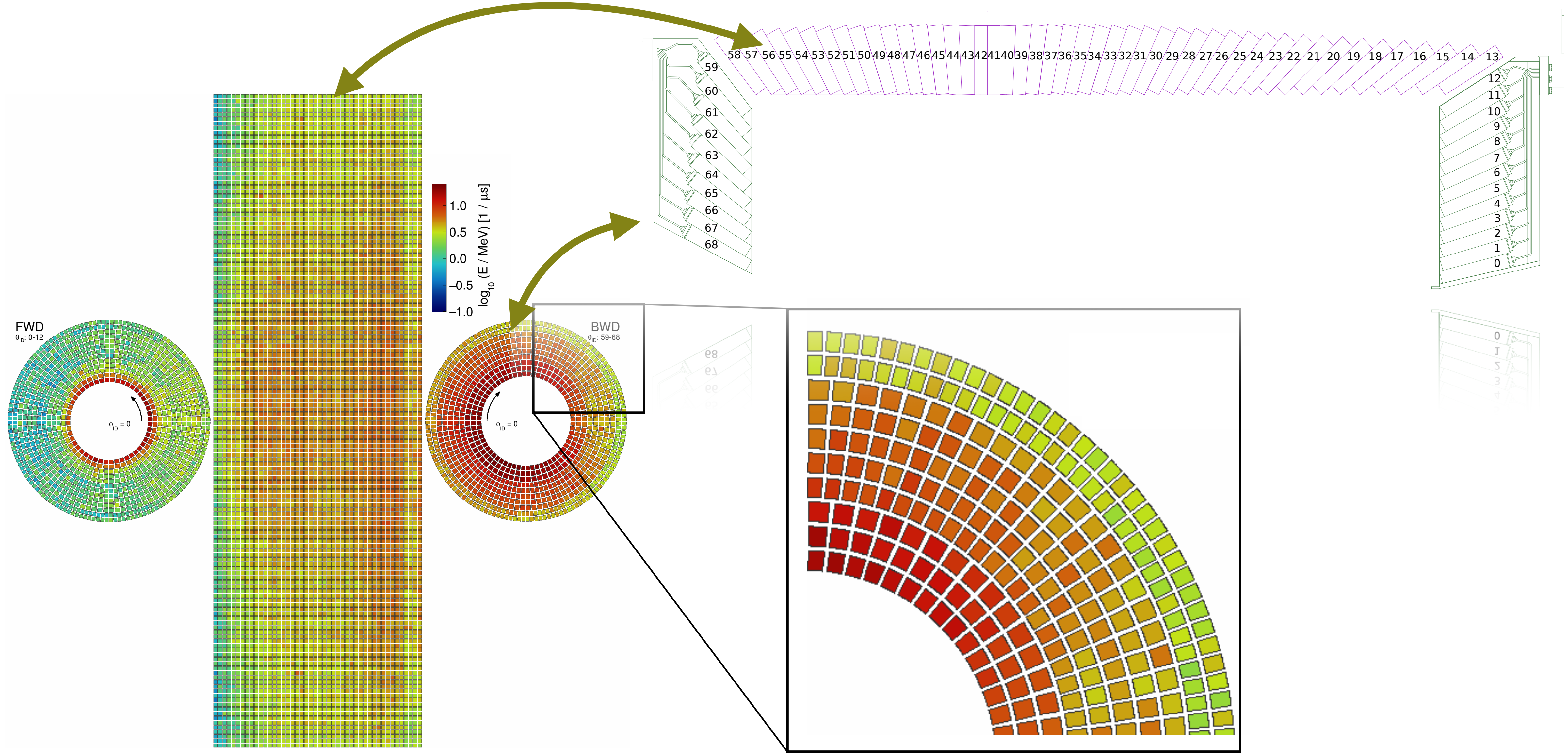




Credit: James Kahn

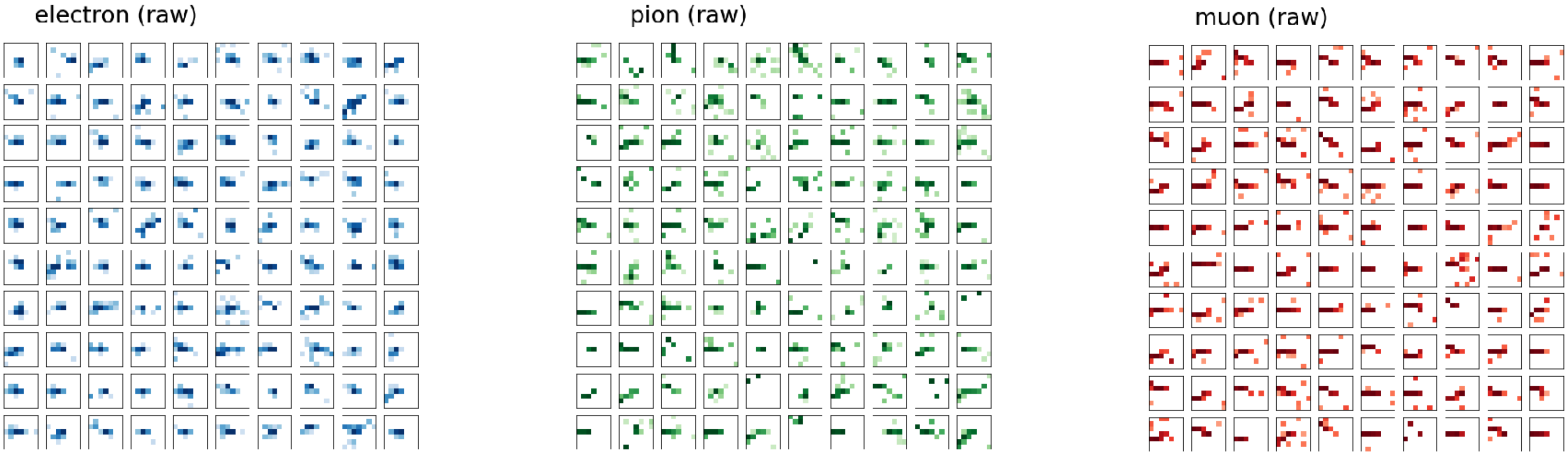
Detector based ML

ML with the Belle II Calorimeter



Charged PID using ECL images

Clustering

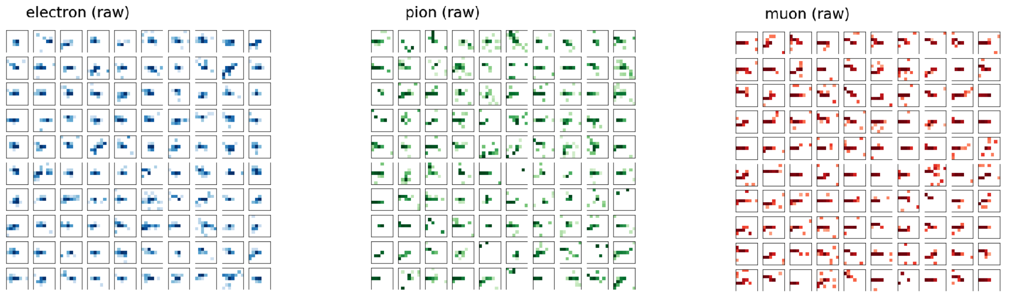


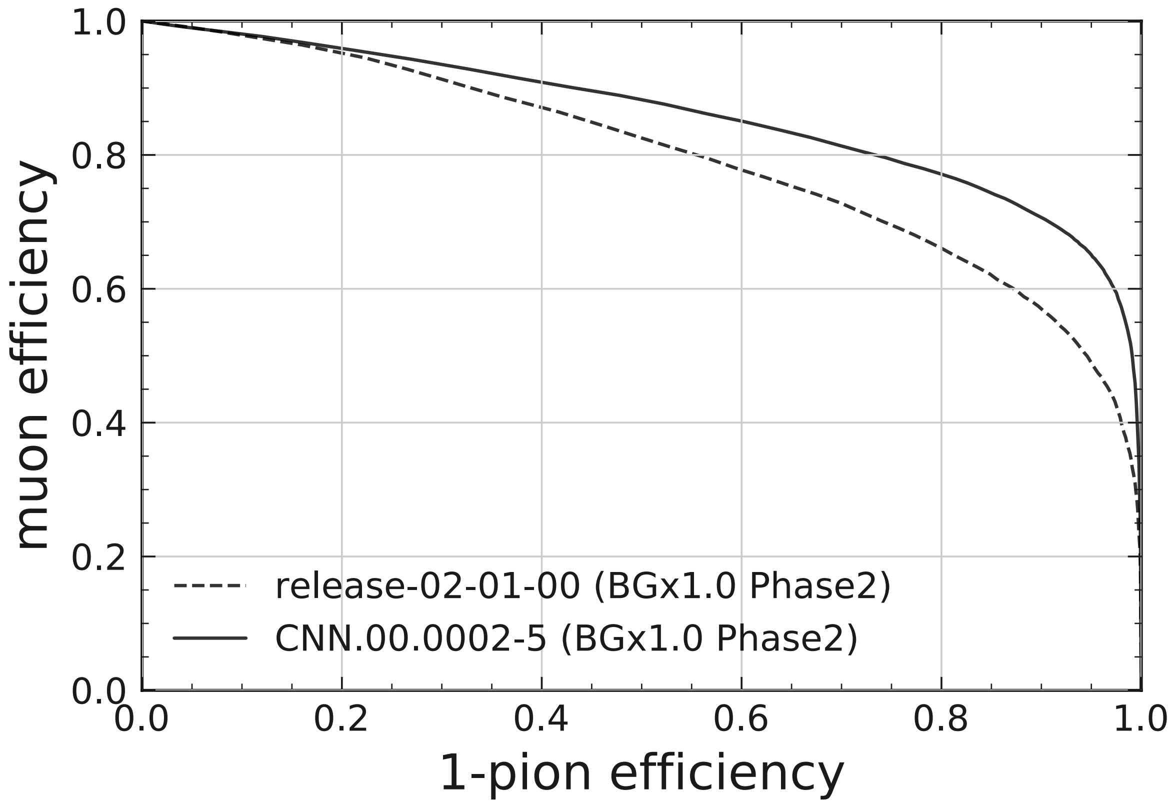
Preprocessing: Image rotation, normalization, thresholds

Charged PID using cluster images

First proof-of-concept

- **Seedless clustering** around extrapolated track
- **Preprocessing** to correct charge asymmetries and background fluctuations
- **Image recognition** using convolutional networks
- **Future:** Add non-image information to the fully connected layers, use asymmetric images, use high dimensional image information (7×7×3..9) from digitized waveforms.



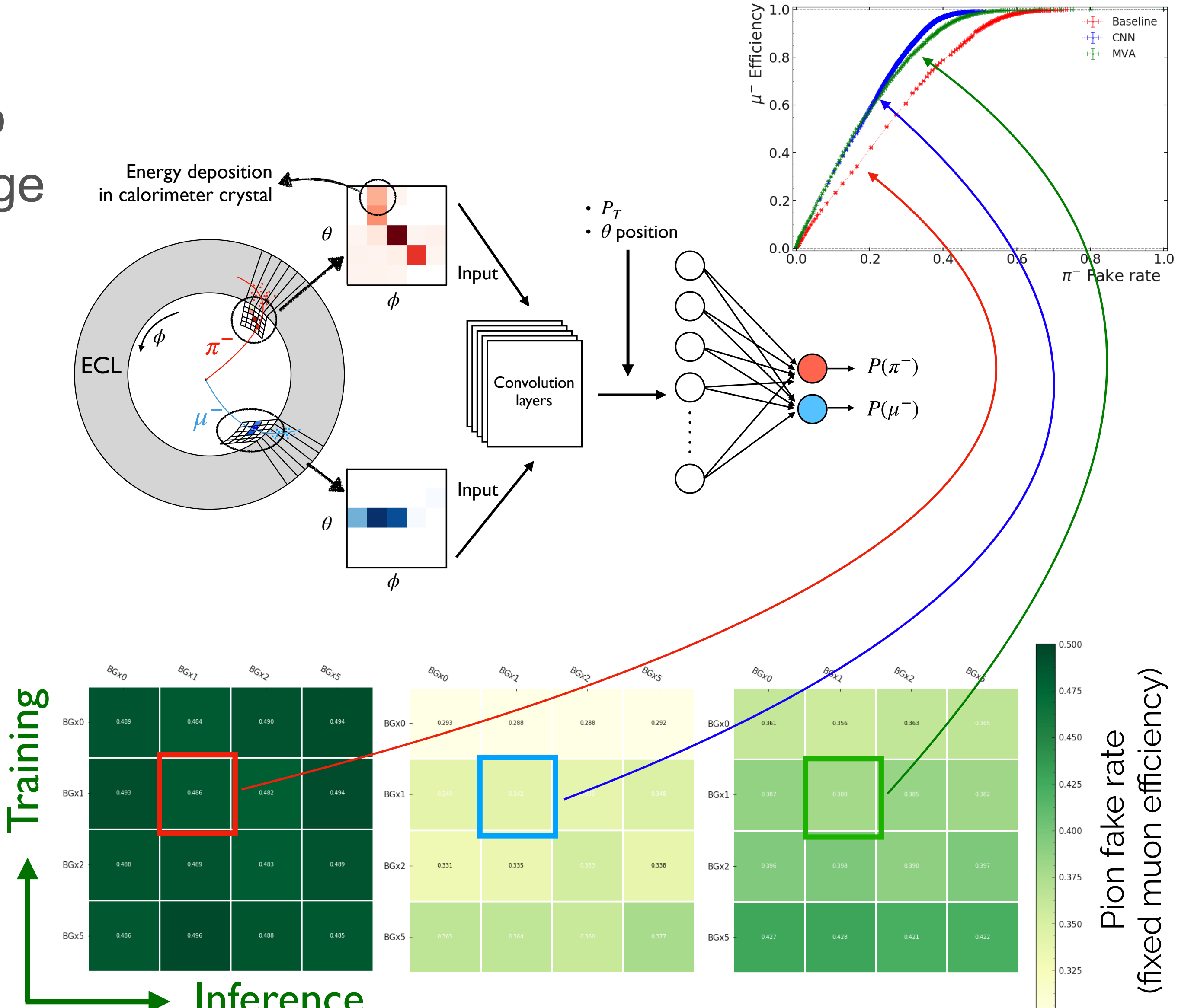


Credit: Torben Ferber

Muon/pion separation for low pt tracks

ML with the Belle II calorimeter

- Low transverse momentum muons do not reach outer muon detector → Large fake rate from pions
- Particle ID based on calorimeter: E/p, BDT, or CNN?
- Design goals:
  - Separation power
  - Robustness against varying beam background
  - Flat efficiency as function of polar angle

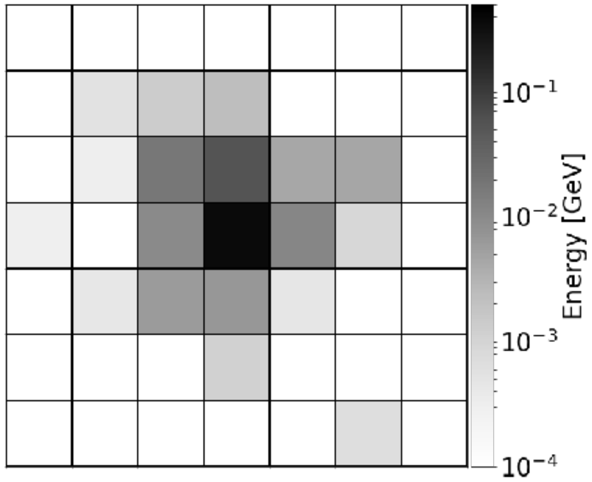


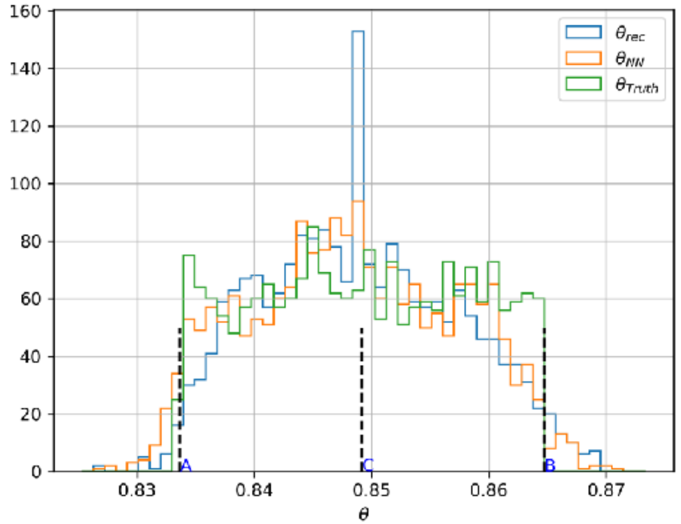
Credit: Torben Ferber

ECL Photon position reconstruction

ML with the Belle II Calorimeter

- Crystal calorimeter: most information contained in central crystal.
- Problem: Very sparse information leads to strong bias towards towards central crystal in non-ML approaches.
- Current ML approach uses “brute force” input 5×5×3 (energy, θ, Φ) and two targets θTruth and ΦTruth. Barrel only (fully connected approach only)
- Move to generalised local position + bias reconstruction next





Credit: Torben Ferber

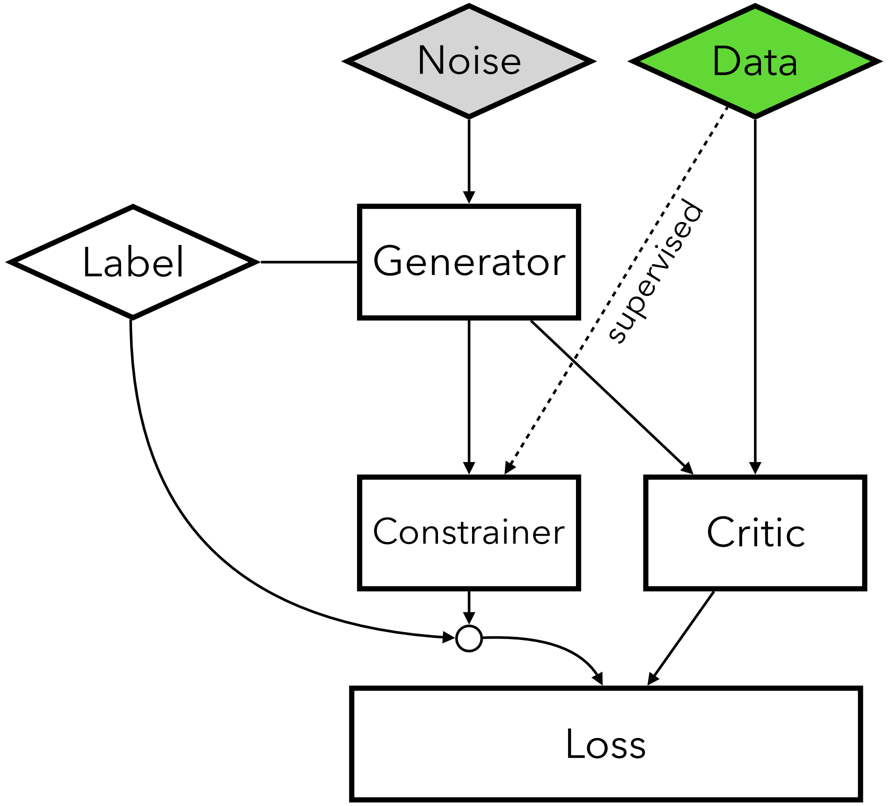
ECL cluster shape calibration

Wasserstein Generative Adversarial Network: WGAN

- High level user analysis is performed on reduced datasets with several expert-engineered shower shape variables per shower
  - Used to separate photons and neutral hadrons
- Differences in data and simulation of shower shapes reduces experimental precision by introducing multiple ad-hoc corrections (one per shower shape)
- Under study: Use Wasserstein refiner networks to calibrate shower images instead, before further analysis steps

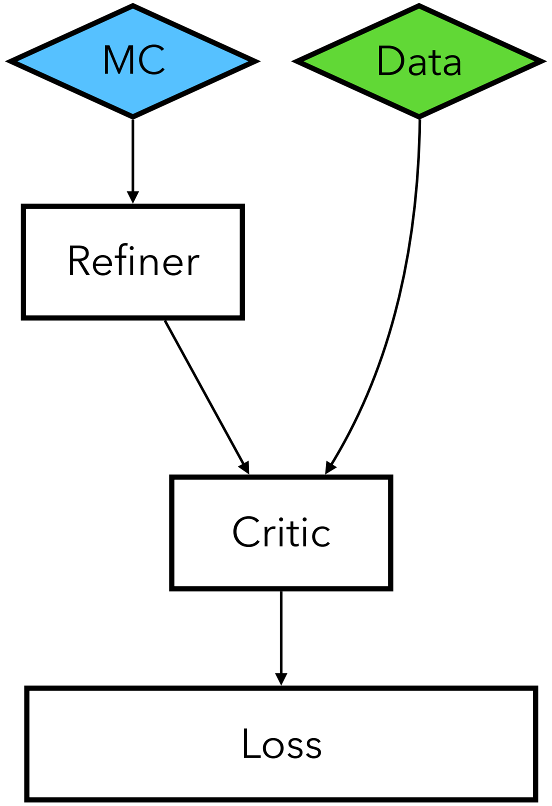
ECL cluster shape calibration

Wasserstein Generative Adversarial Network: WGAN



Wasserstein Generative Adversarial Network: **WGAN**
(with supervised auxiliary constrainers: AC-WGAN)

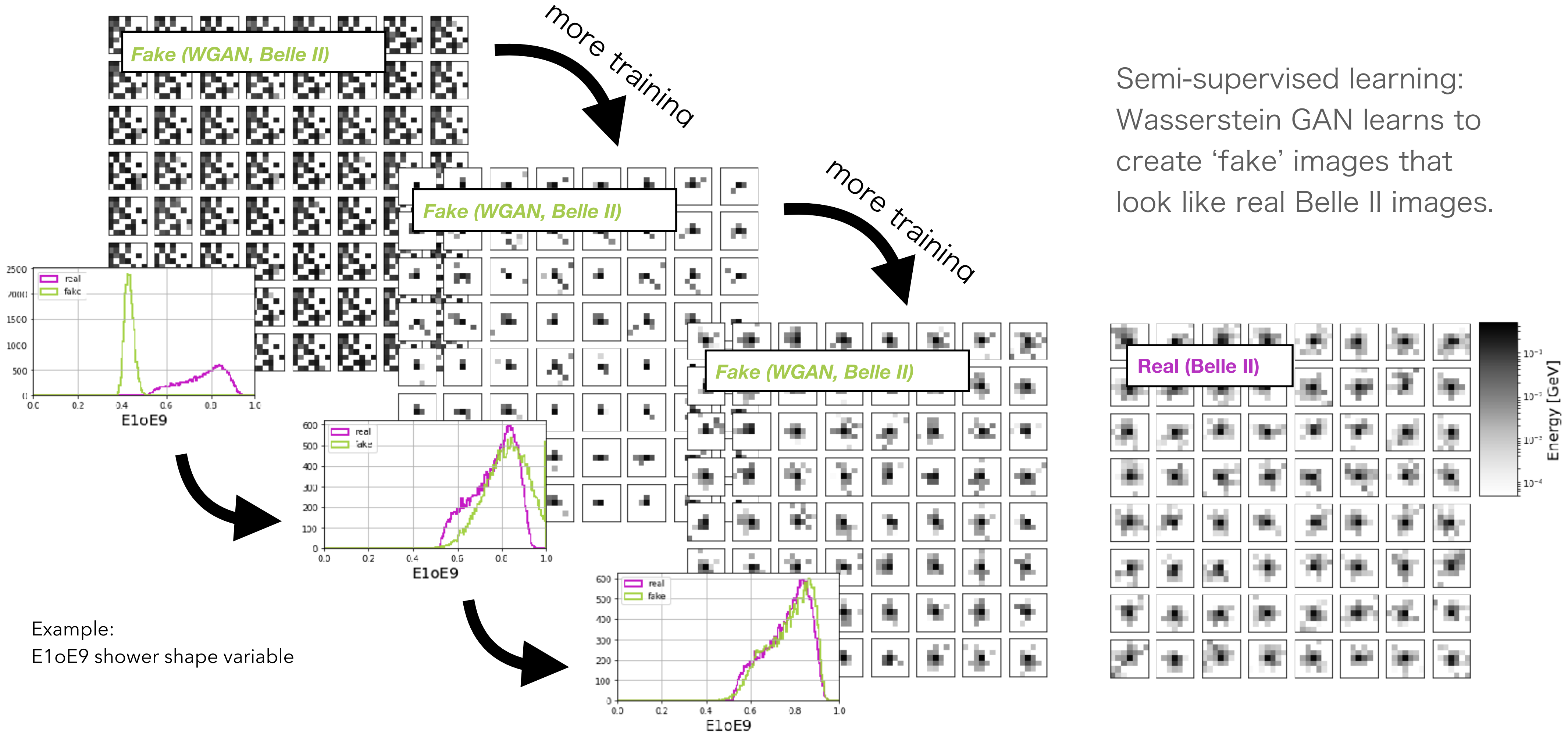
Improve existing MC simulations using data before further analysis steps.



Wasserstein Refiner Network

ECL cluster shape calibration

Wasserstein Generative Adversarial Network: WGAN



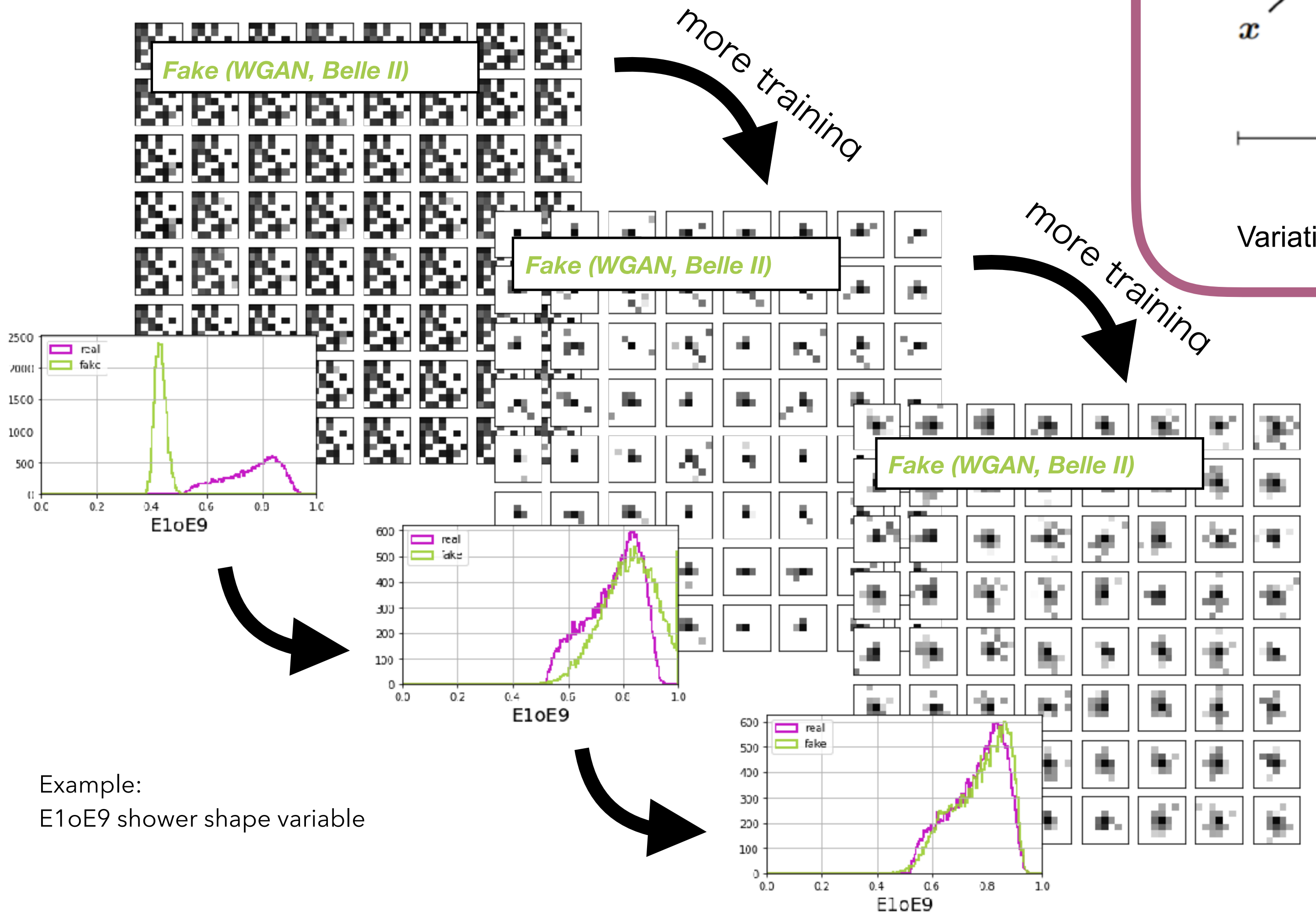
Semi-supervised learning: Wasserstein GAN learns to create 'fake' images that look like real Belle II images.

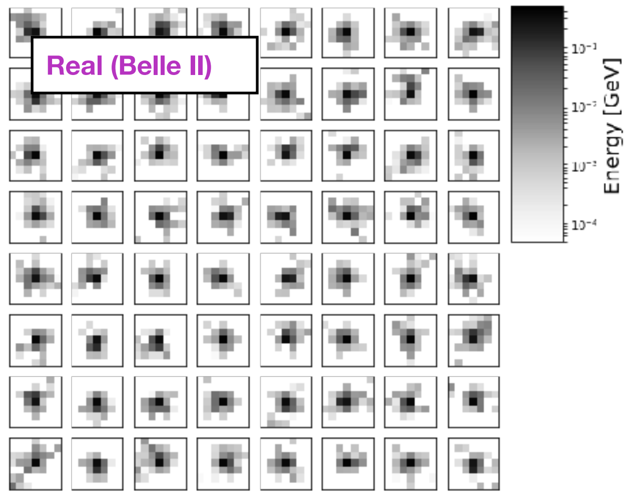
Example:
E1oE9 shower shape variable

Credit: Torben Ferber

ECL cluster shape calibration

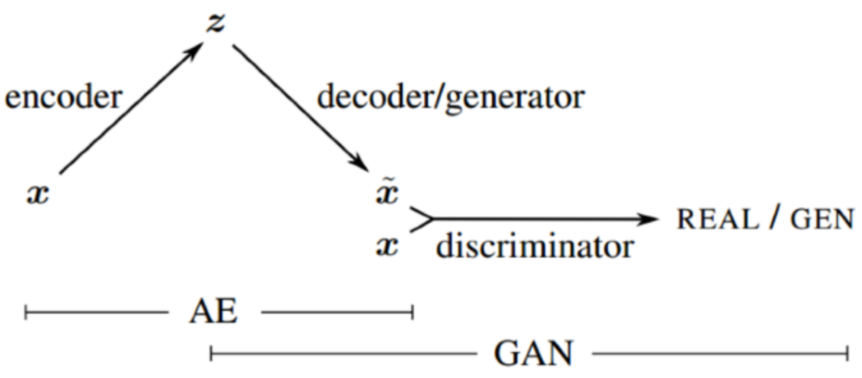
Wasserstein Generative Adversarial Network: WGAN





Example:
E1oE9 shower shape variable

Future



Variational Autoencoder Generative Adversarial Network

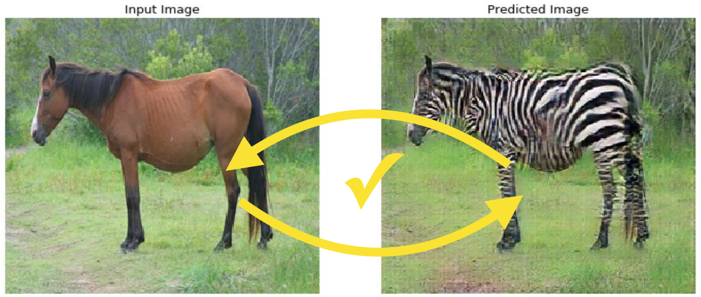
Credit: Torben Ferber

Photon calibration samples using CycleGANs

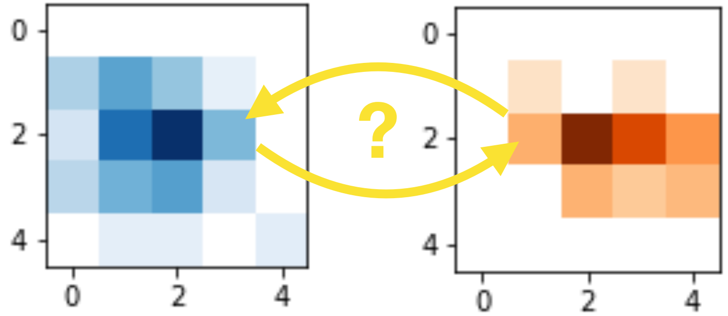
ML with the Belle II calorimeter

- Difficult to get very pure photon calibration sample of low-medium energy at Belle II
- Train CycleGAN to convert electrons into photons: Same physics, different curvature due to magnetic field
- Visually appealing, but is the physics right? (Same question for GANs)
- Design goals:
  - Proof of concept for a use case of CycleGANs in HEP

Horse ↔ Zebra

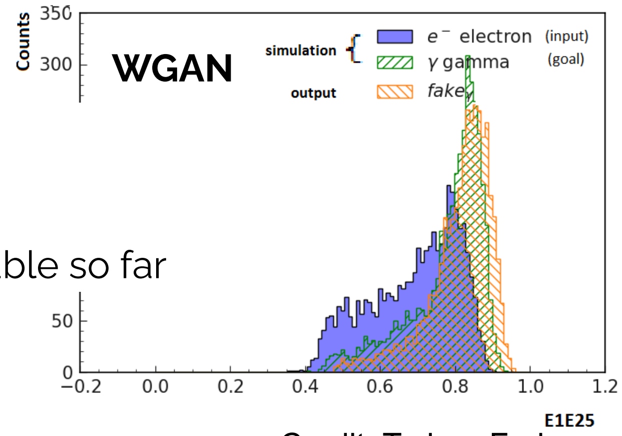


Electron ↔ Photon



Work in progress:
Electron input
WGAN photon output

Cycle-constraints unstable so far



Credit: Torben Ferber

Fast inference for L1 trigger

ML with the Belle II calorimeter

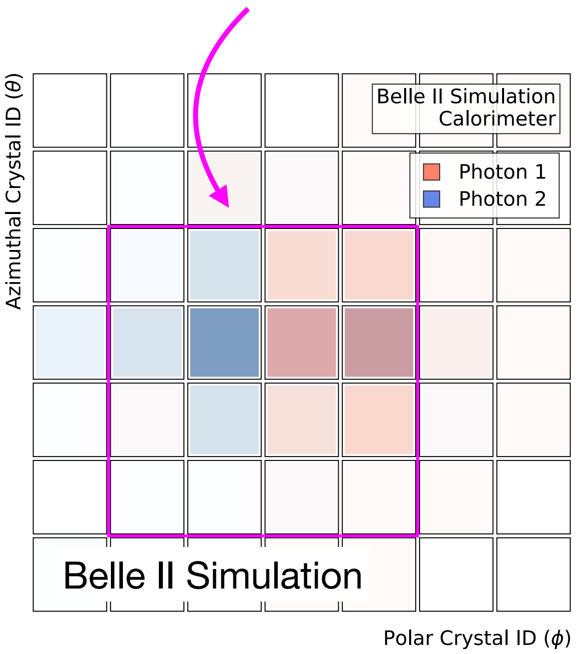
CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

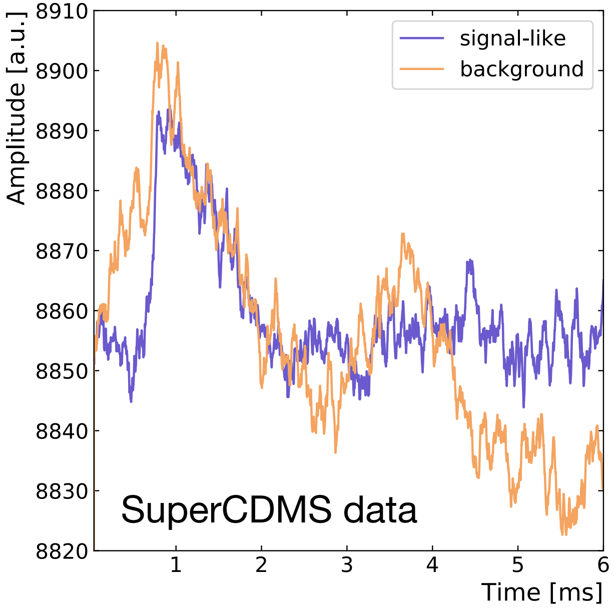
- New “Cross disciplinary” project within the Quantum Universe cluster: Belle II (intensity frontier) and SuperCDMS (low background frontier)

- Belle II: Real time photon identification for merged clusters and low threshold Dark Photon searches

- SuperCDMS: Ultra-low threshold triggers for light ALP searches

- Please get in touch if you are interested.





Credit: Torben Ferber

Final Excursion: Fallacies and Prospects

Dangers of Deep Learning Methods

Final Excursion: Fallacies and Prospects

- Deep learning methods sometimes have extraordinary amount of free parameters (VC dimensions >> data)
- Learned “features" of the data is not always obvious

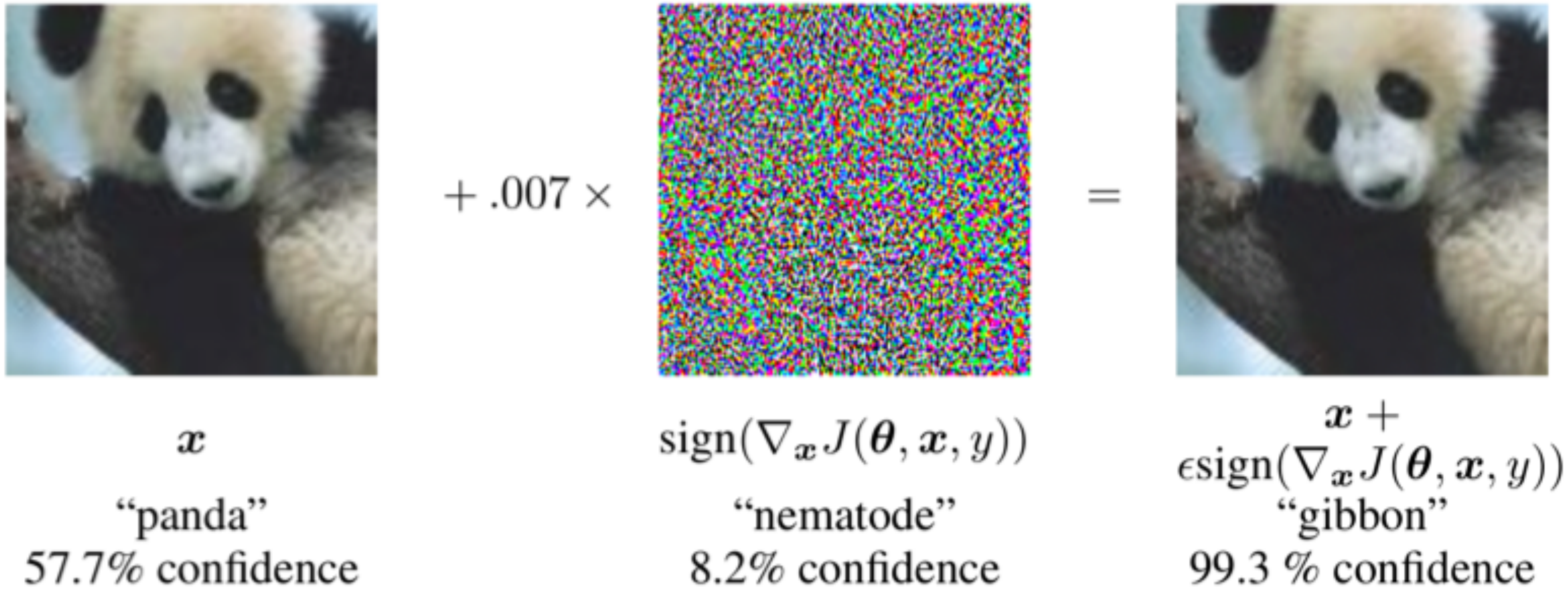
Dangers of Deep Learning Methods

Final Excursion: Fallacies and Prospects

- Deep learning methods sometimes have extraordinary amount of free parameters (VC dimensions >> data)
- Learned “features" of the data is not always obvious

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

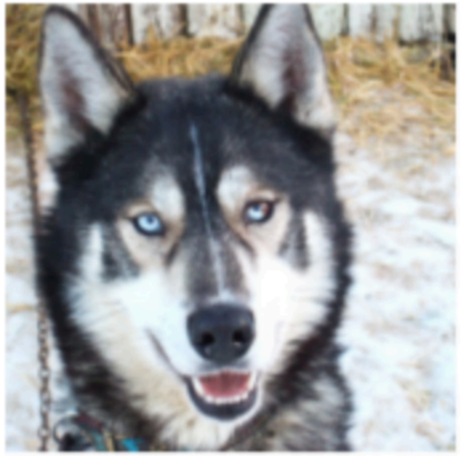
Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com



Dangers of Deep Learning Methods

Final Excursion: Fallacies and Prospects

- Deep learning methods sometimes have extraordinary amount of free parameters (VC dimensions >> data)
- Learned “features" of the data is not always obvious



(a) Husky classified as wolf

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

Re-Interfere model response on the input may help understand the expertise

"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Dangers of Deep Learning Methods

Final Excursion: Fallacies and Prospects

- Deep learning methods sometimes have extraordinary amount of free parameters (VC dimensions >> data)
- Learned “features" of the data is not always obvious

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the “Husky vs Wolf” task.

Re-Interfere model response on the input may help understand the expertise

"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Dangers of Deep Learning Methods

Final Excursion: Fallacies and Prospects

- Deep learning methods sometimes have extraordinary amount of free parameters (VC dimensions >> data)
- Learned “features" of the data is not always obvious
- These problems often can’t be spotted with classical test and training data sets

Heatmaps showing the class evidence extracted from of each part of the image.

Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet
Wieland Brendel, Matthias Bethge

Dangers of Deep Learning Methods

Final Excursion: Fallacies and Prospects

- Deep learning methods sometimes have extraordinary amount of free parameters (VC dimensions >> data)
- Learned “features" of the data is not always obvious
- These problems often can’t be spotted with classical test and training data sets

Heatmaps showing the class evidence extracted from of each part of the image.

Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet
Wieland Brendel, Matthias Bethge

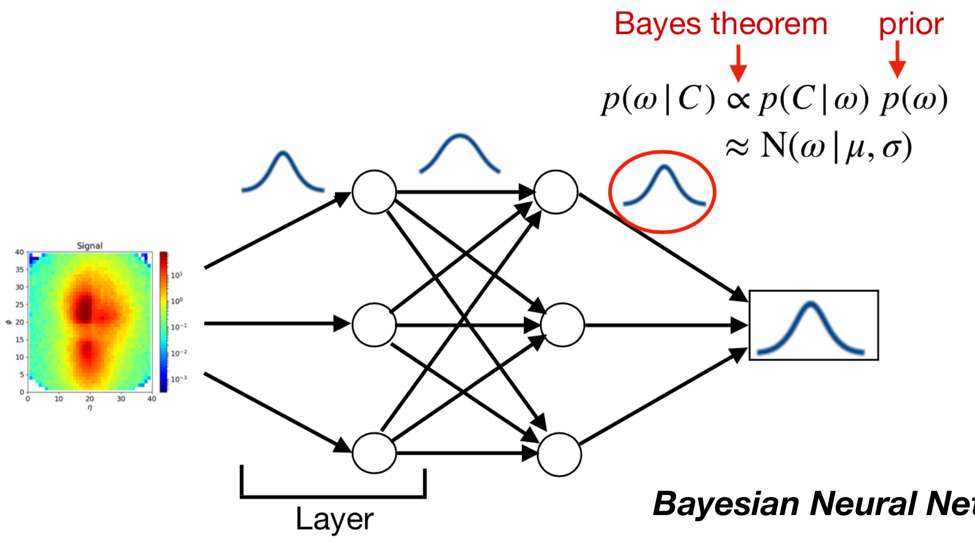
Dangers of Deep Learning Methods

Final Excursion: Fallacies and Prospects

- Deep learning methods sometimes have extraordinary amount of free parameters (VC dimensions >> data)
- Learned “features" of the data is not always obvious
- These problems often can’t be spotted with classical test and training data sets

Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet
Wieland Brendel, Matthias Bethge

Plans for the Future of ML at Belle II



$$p(\omega\,|\,C) \propto p(C\,|\,\omega)\; p(\omega) \approx \mathrm{N}(\omega\,|\,\mu,\sigma)$$

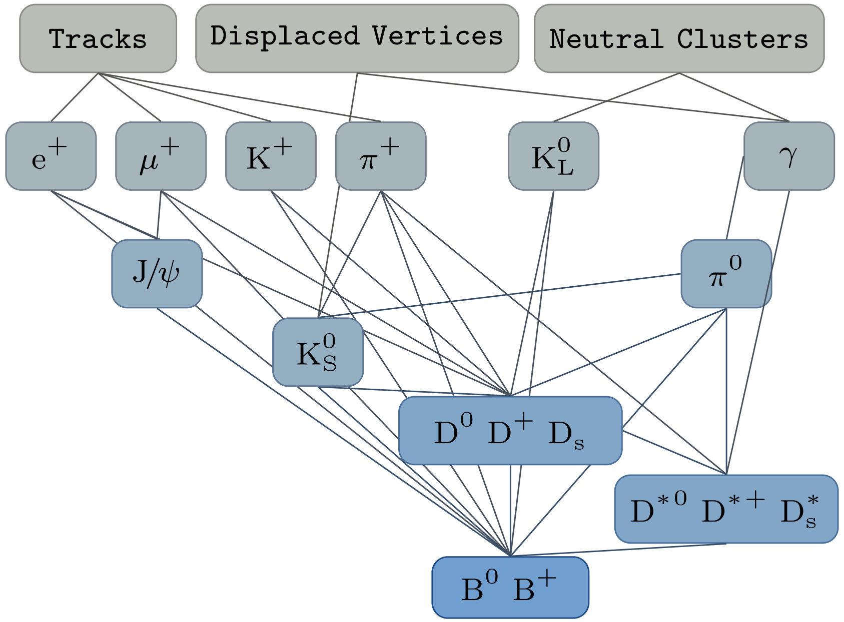
Bayesian Neural Net

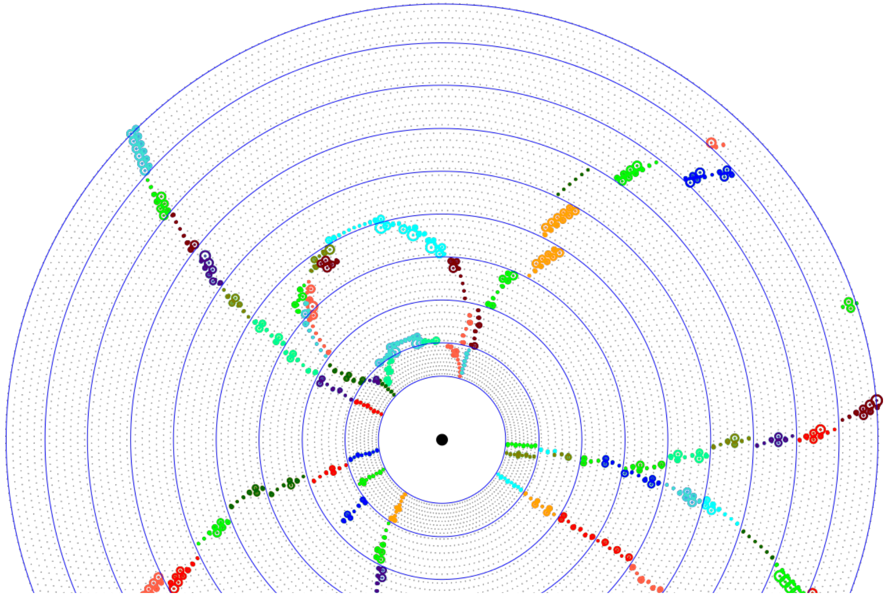
Graph Neural Network

Plans for the Future of ML at Belle II

- GNNs might be applicable in many parts of our analyses
  - Deep Full Event Interpretation
  - Skimming of Data
  - Tracking

HEP.TrkX project

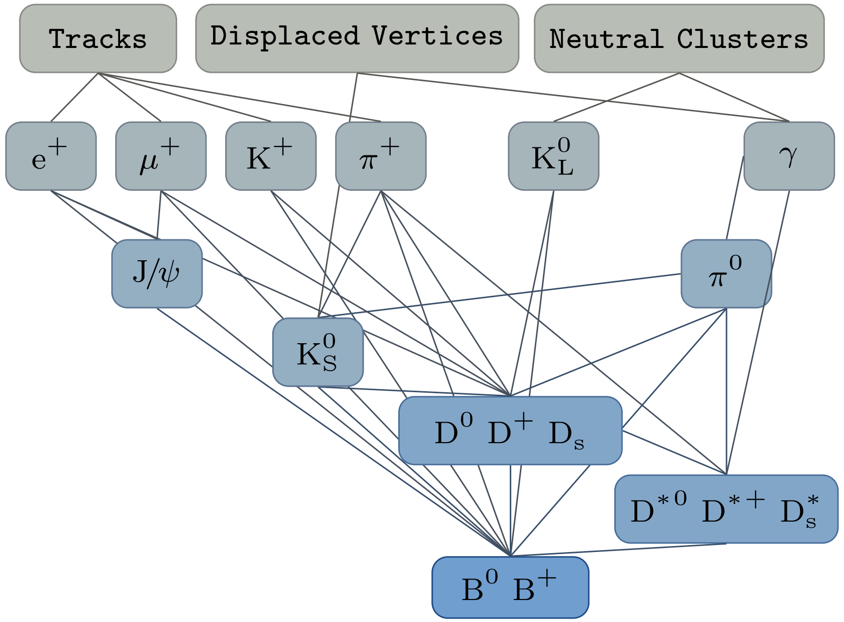




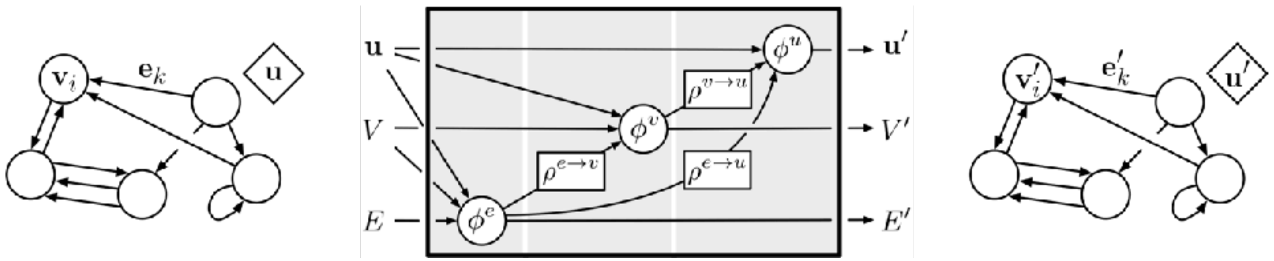
Graph Neural Network

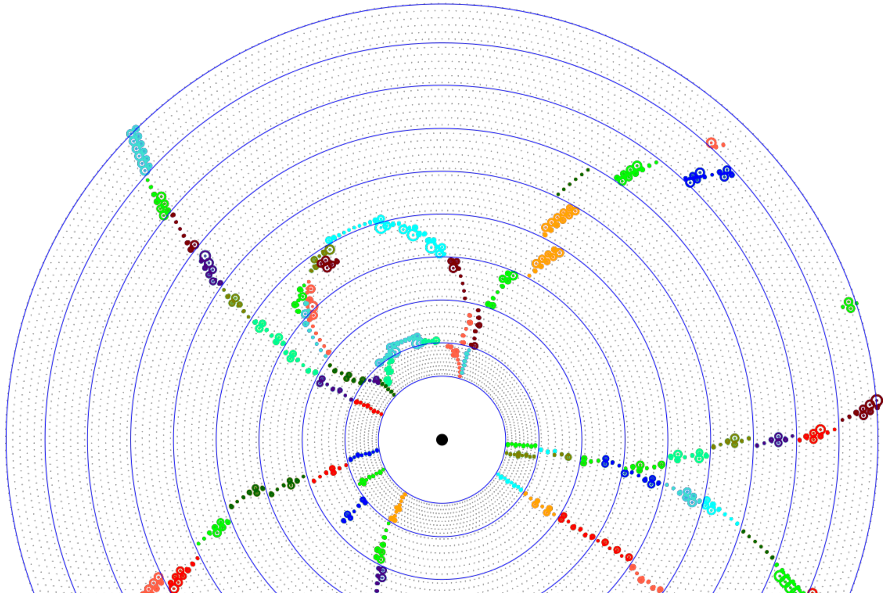
Plans for the Future of ML at Belle II

- GNNs might be applicable in many parts of our analyses
  - Deep Full Event Interpretation
  - Skimming of Data
  - Tracking



HEP.TrkX project



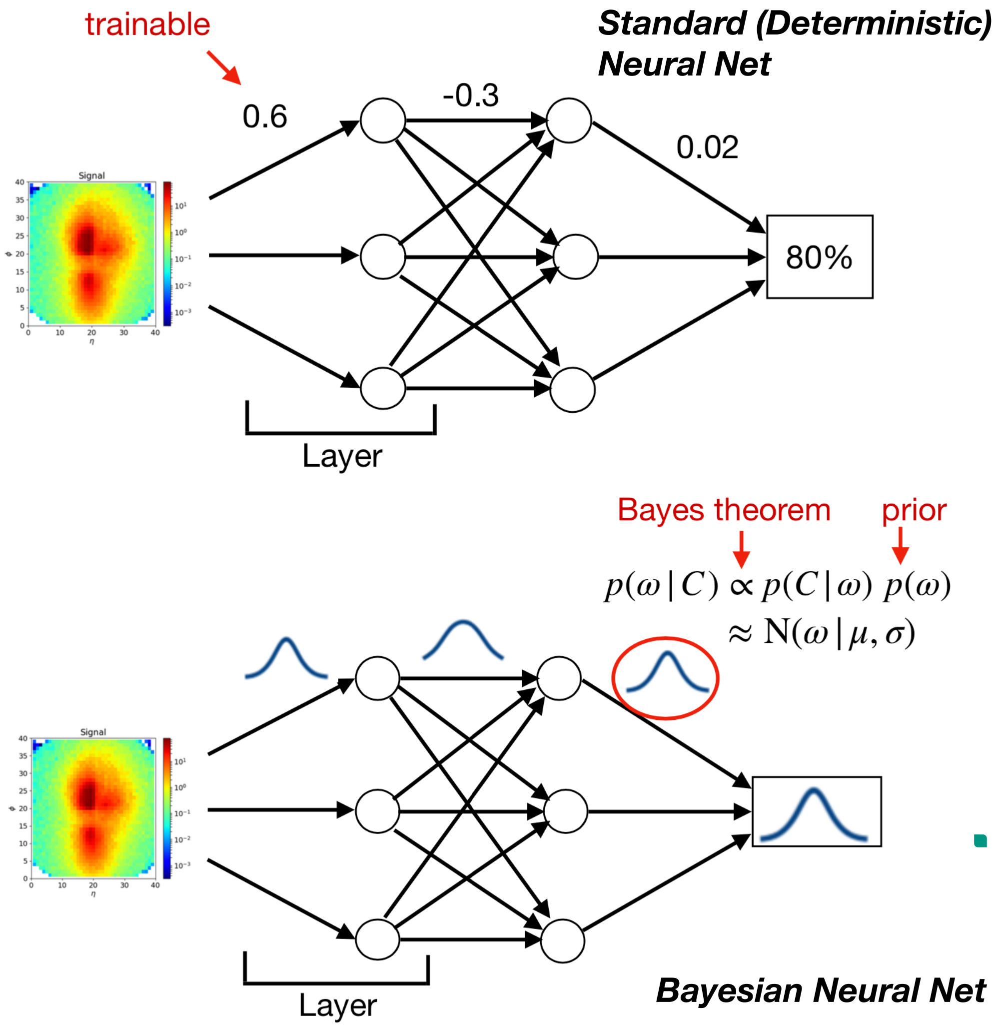


Systematic Uncertainties and Deep Learning

Plans for the Future of ML at Belle II

- Systematic uncertainty on multivariate methods are a serious challenge
  - How to propagate uncertainties

- Bayesian Neural Networks are more and more used in HEP

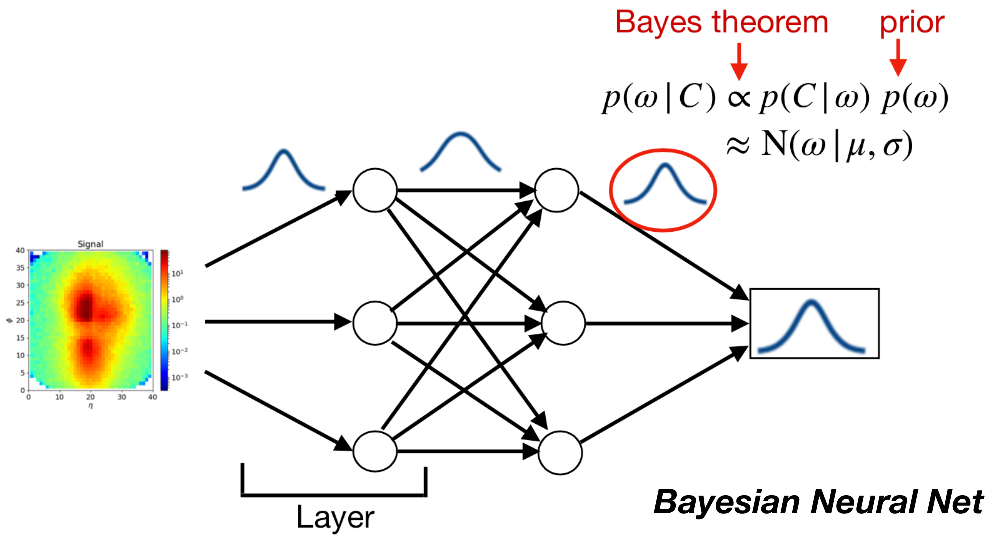


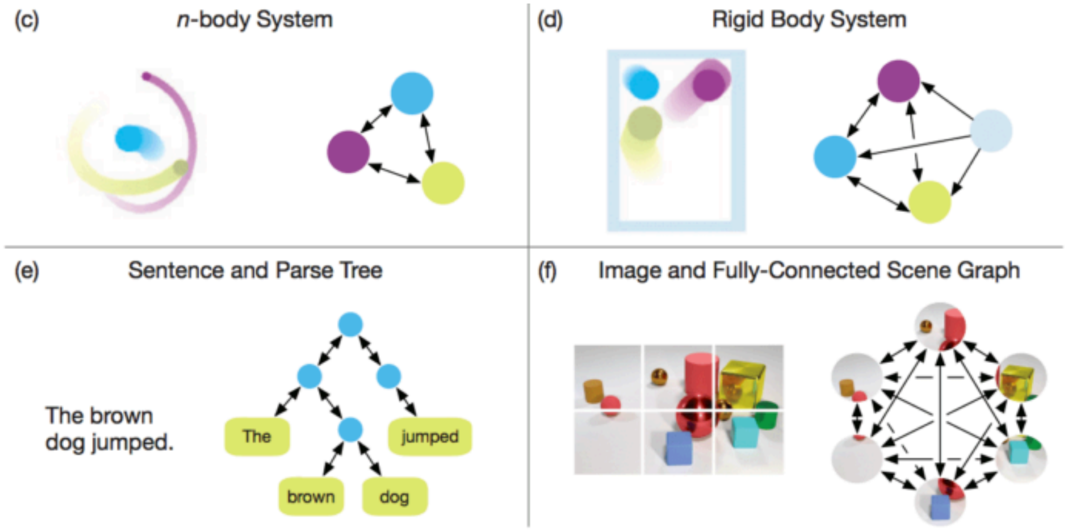
- Talk from Gregor Kasieczka (U. Hamburg, CMS)

Conclusion

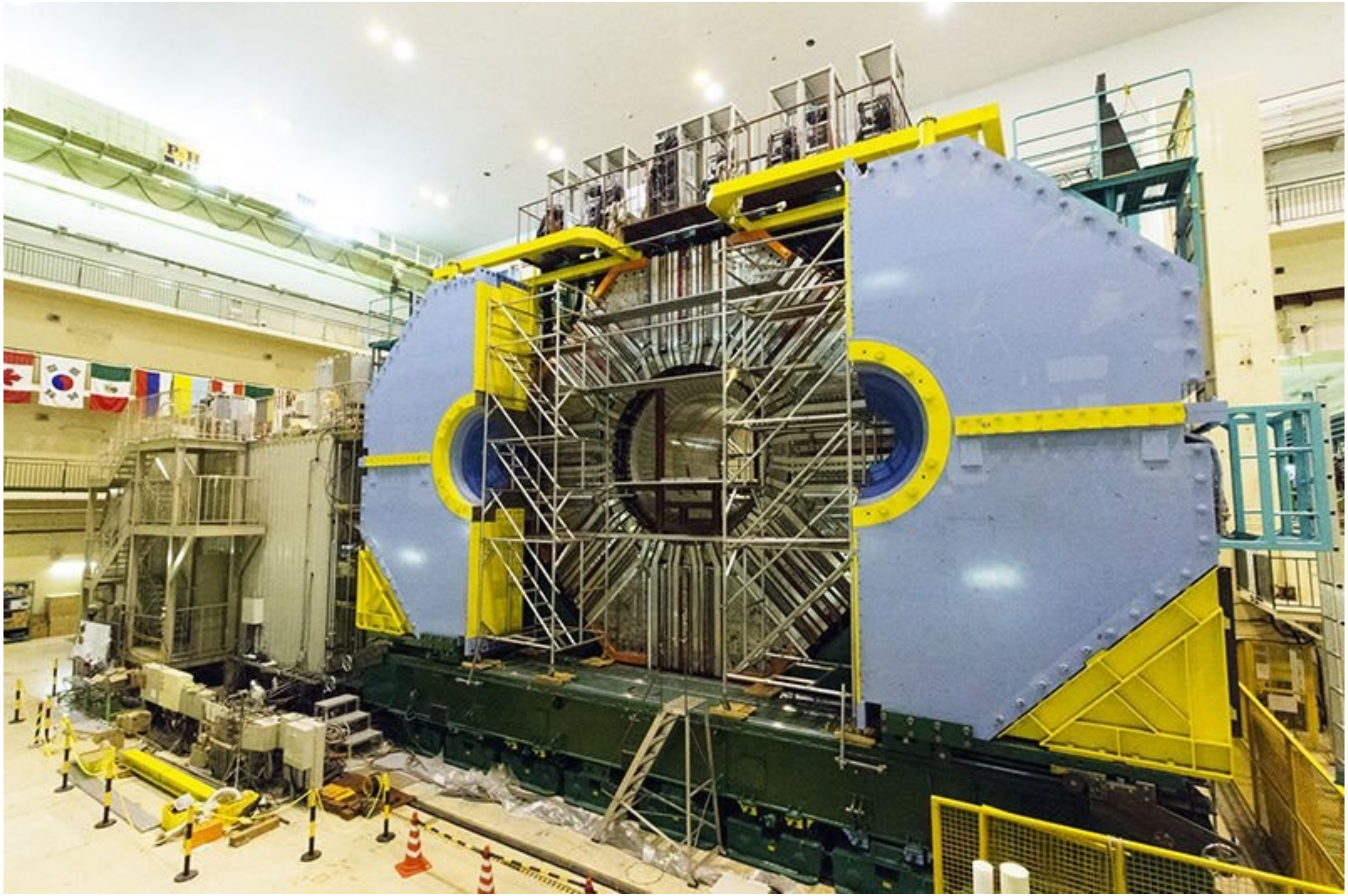
Prospects for machine learning

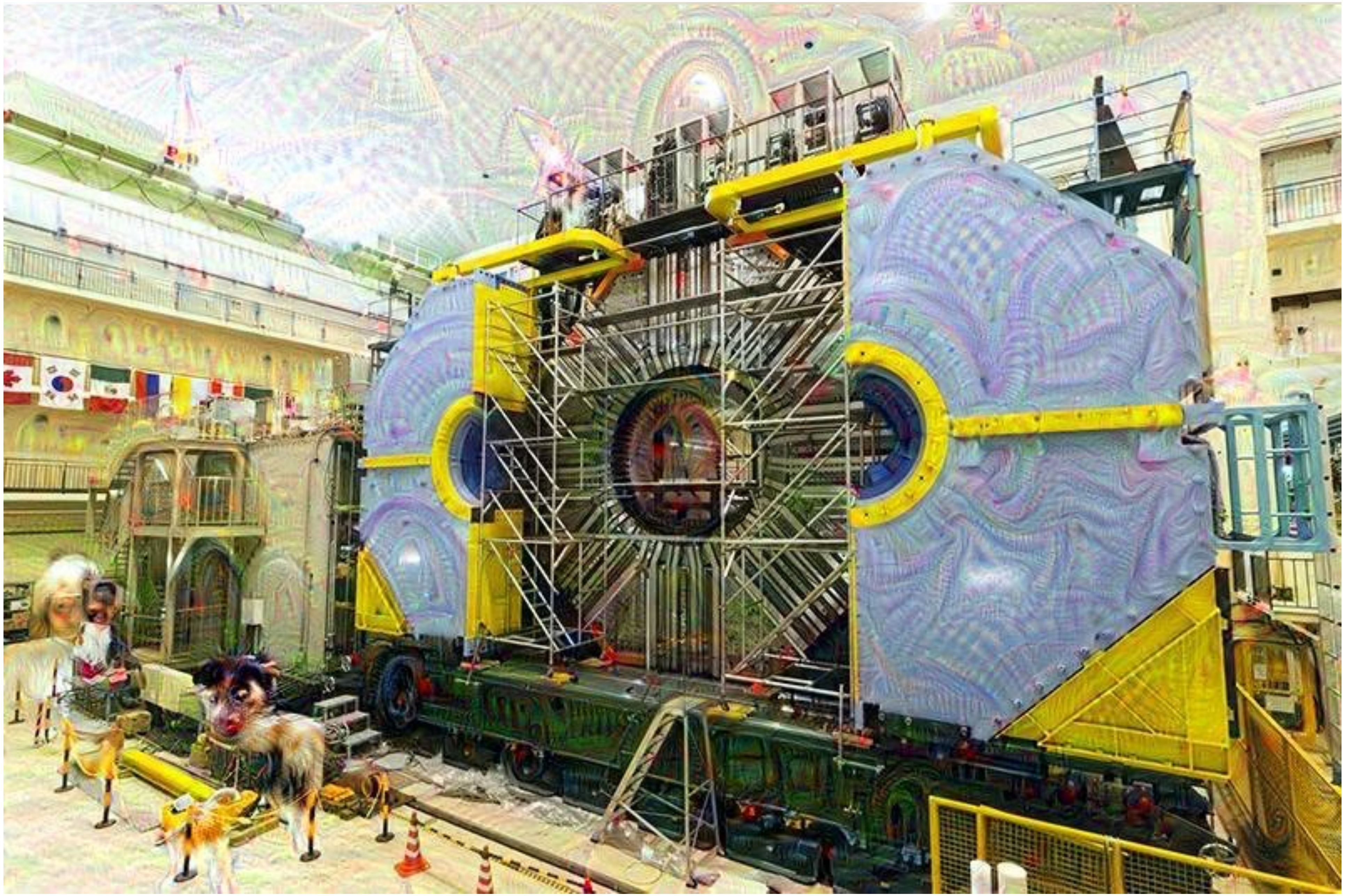
- “Classical" ML is very successful in Belle II analyses already
- Potential of Deep Learning is explored in many studies like simulation, reconstruction and analyses
  - Is the problem well described?
  - Does it get the physics right?
  - There is for sure a lot room for improvement
- GraphNN approaches offer interesting new opportunities
  - Work in progress
  - Promising approaches exist already from industry and other experiments

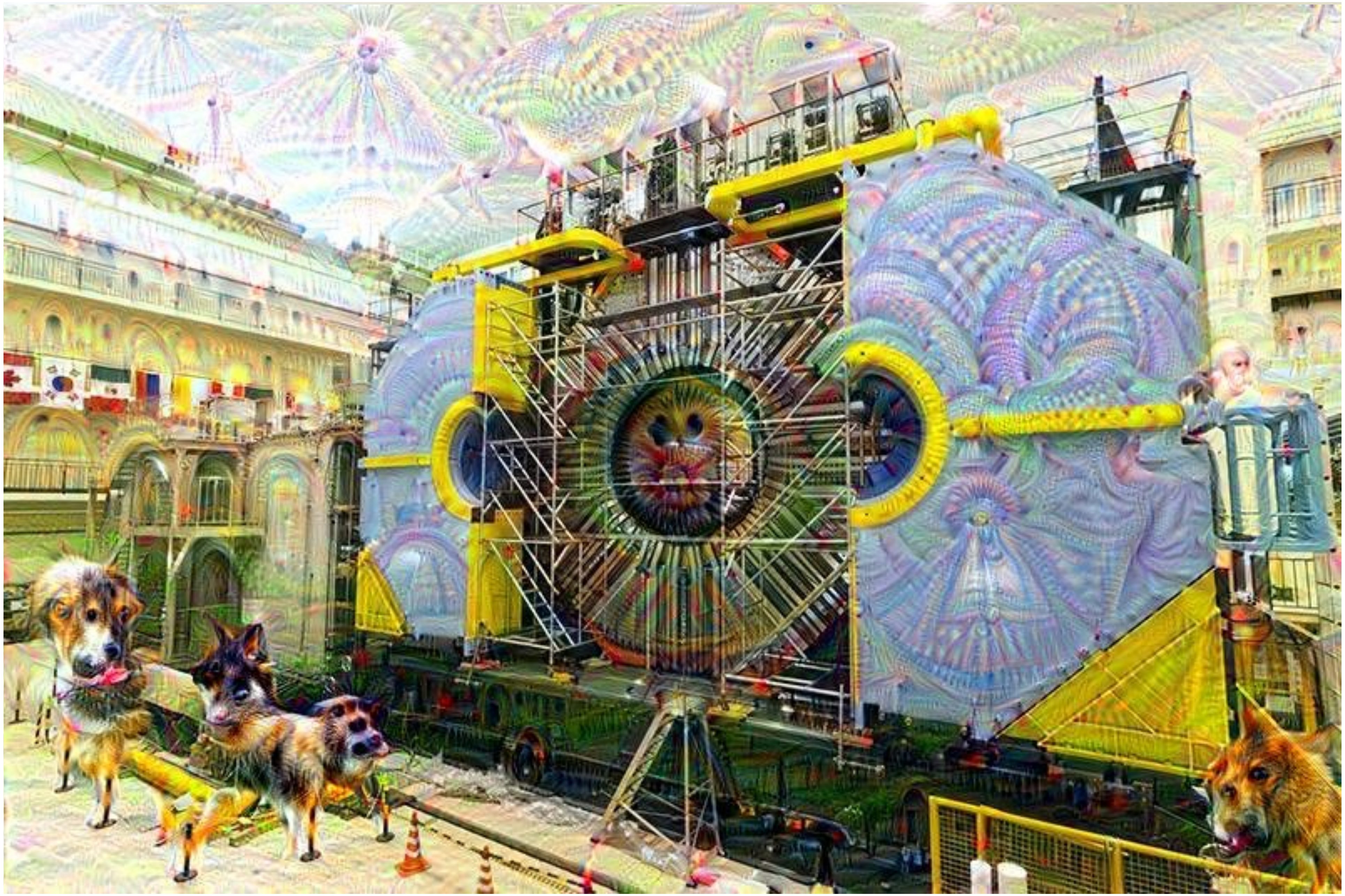


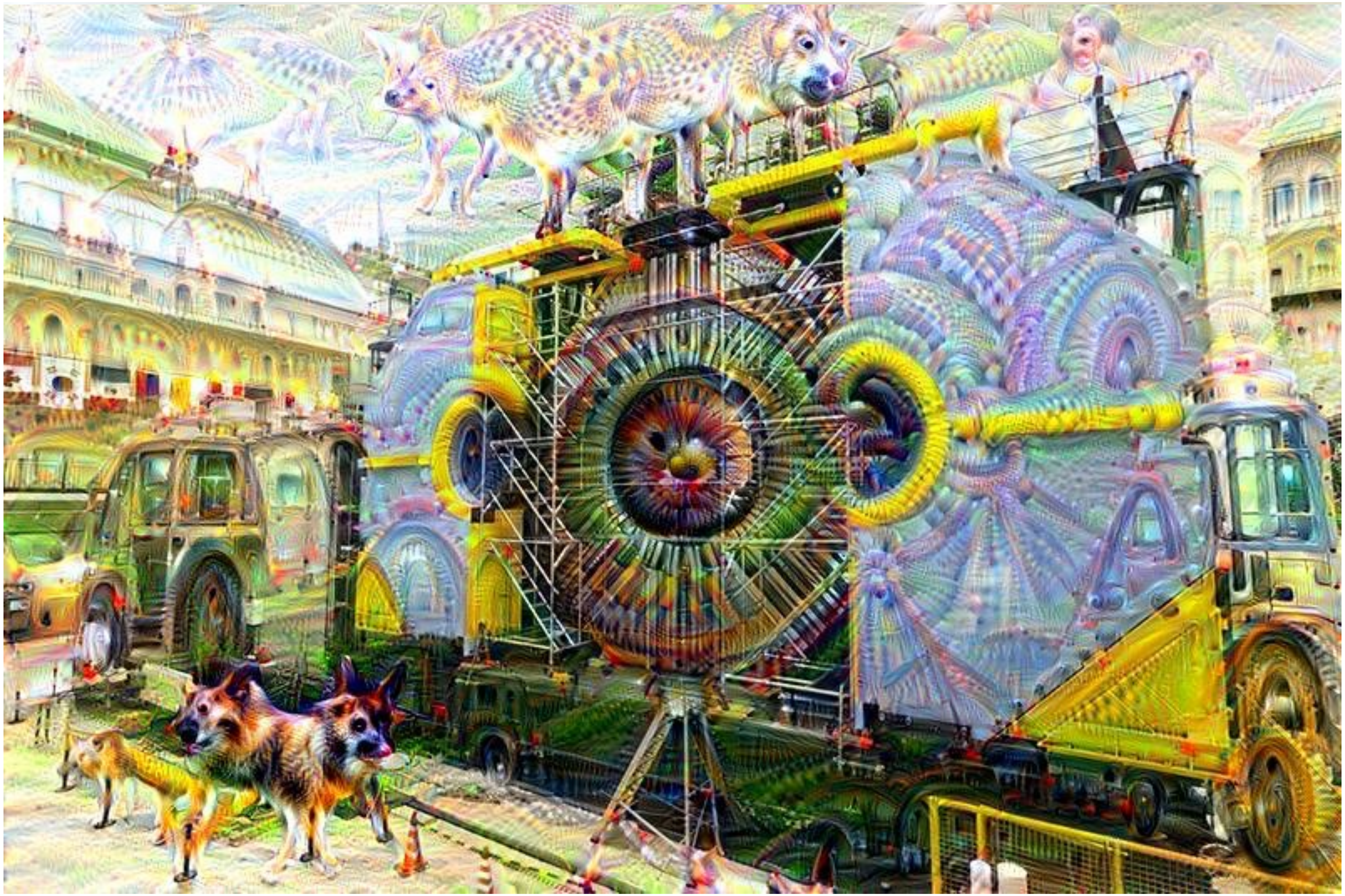


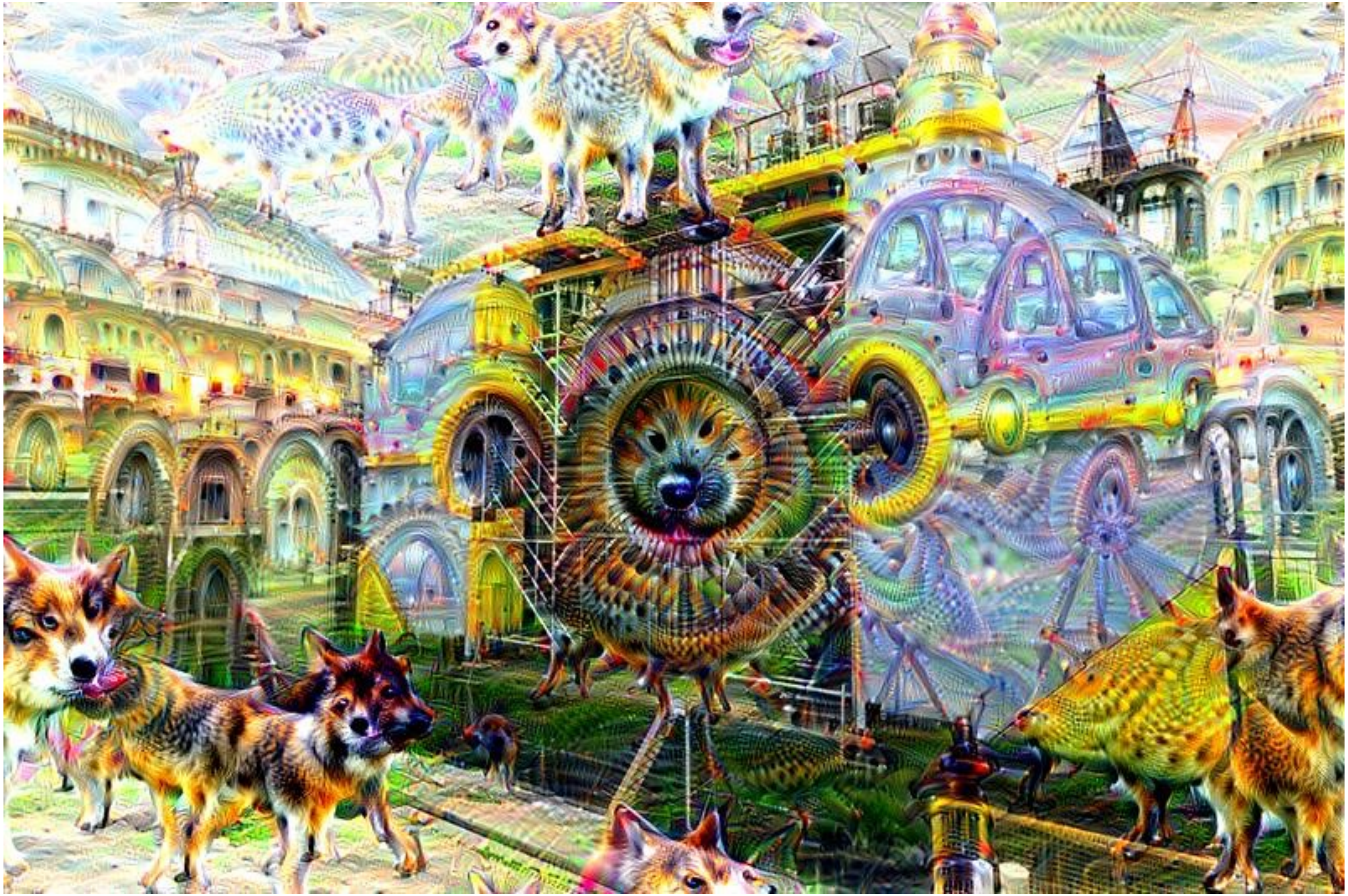
Credit: Google DeepMind

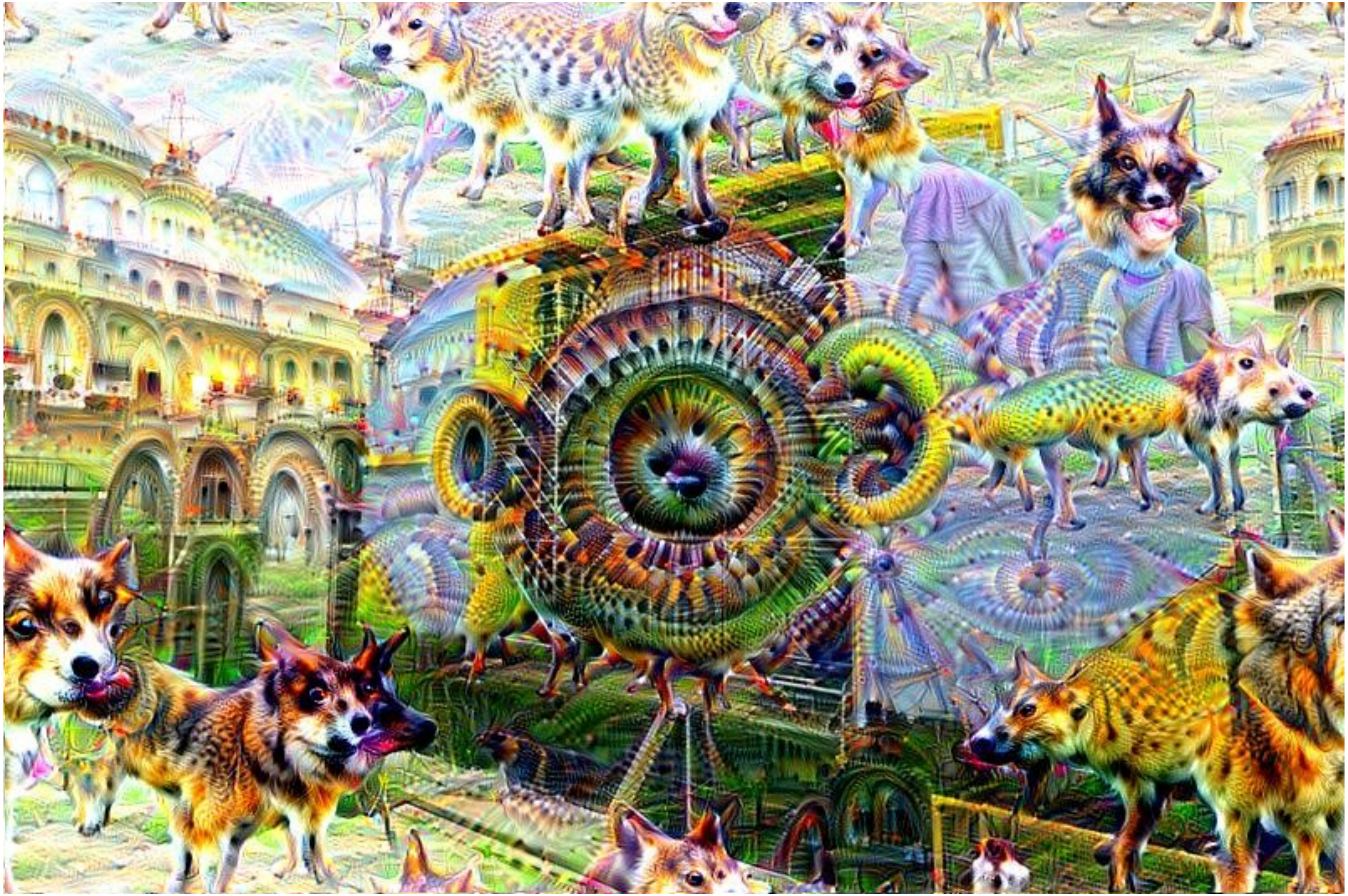












Thank you very much for your attention!

Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de

Simon Wehle
Belle & Belle II
simon.wehle@desy.de
+49 40 4994 3789