LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Reconstruction of Full Decays using Transformers and Hyperbolic Embedding at Belle II

Boyang Yu

Ludwig-Maximilians-Universität München

EPD Seminar, IHEP Beijing, April 18th, 2023

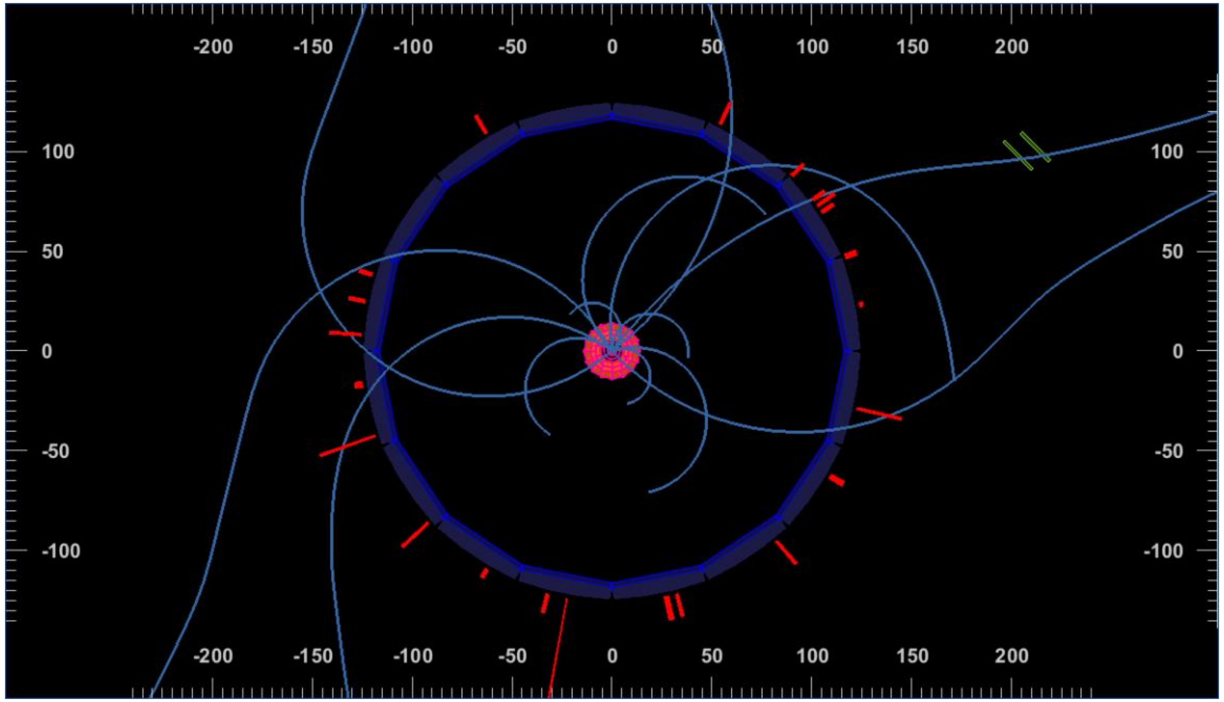
Belle II

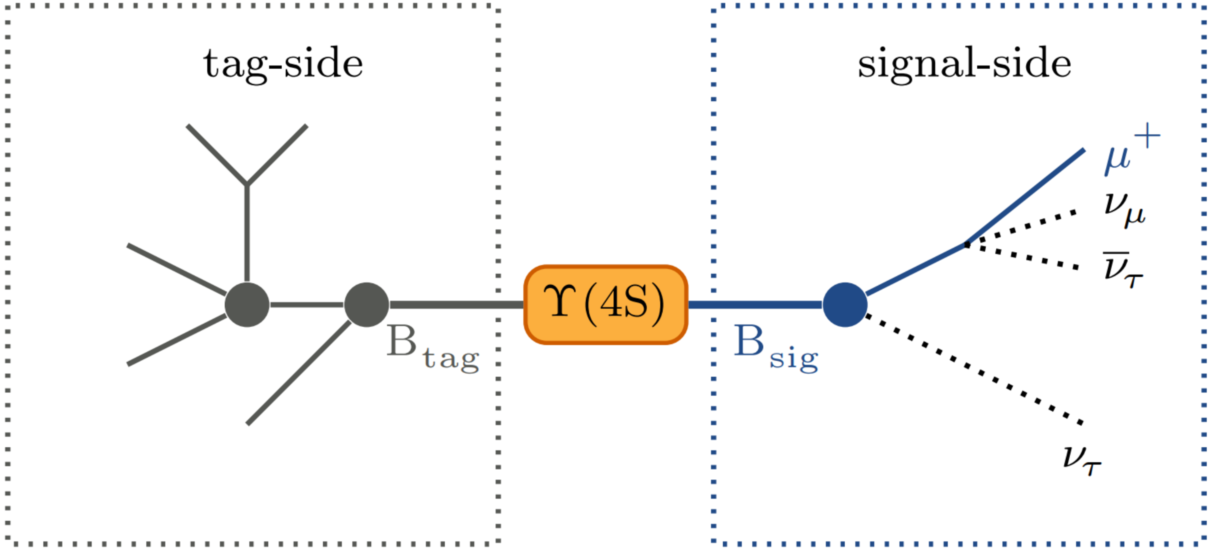
Bundesministerium für Bildung und Forschung

PUNCH 4NFDI

Boyang.Yu@physik.uni-muenchen.de

Reconstruction of full decays





Detector information → Decay information

Reconstruction of full decays

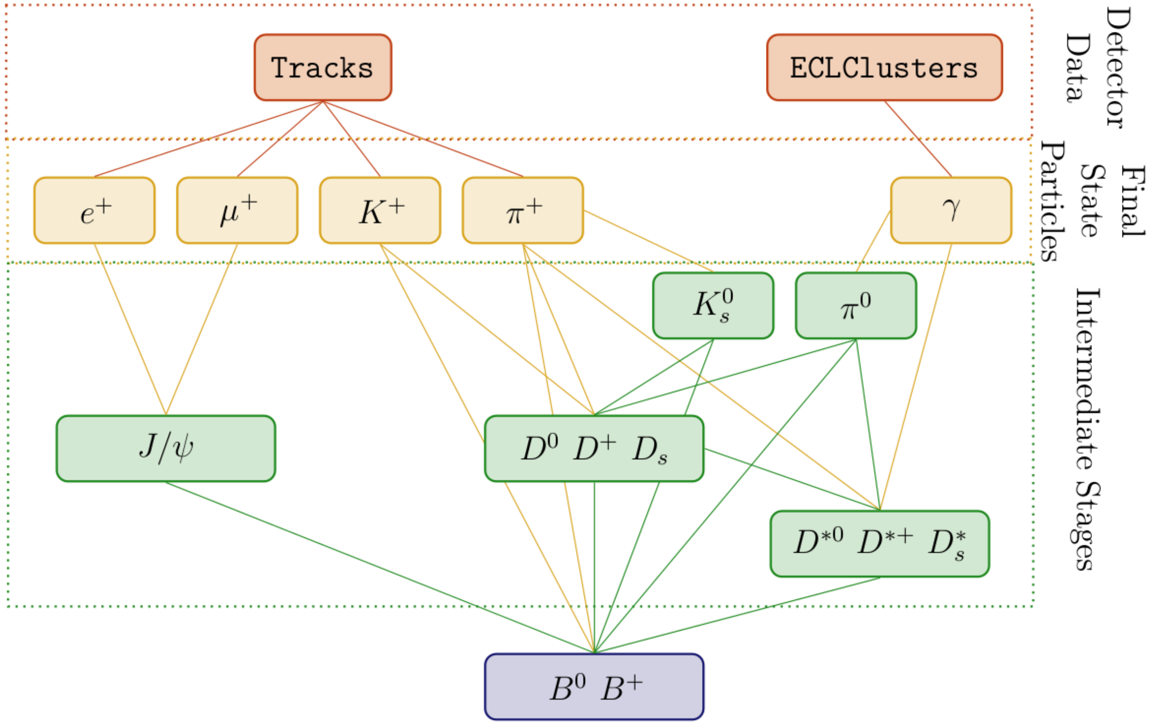
Full Event Interpretation:

- Estimate probabilities of individual decays using boosted decision trees (BDTs)
- Hierarchical reconstruction of the whole decay tree in 7 stages
- In total $\mathcal{O}(10^3)$ BDTs

Limitations:

- Hard-coded decay channels for each particle
- Hard-coded particle types at each stage

-> Low reconstruction efficiency: $\mathcal{O}(1\%)$



Improvement

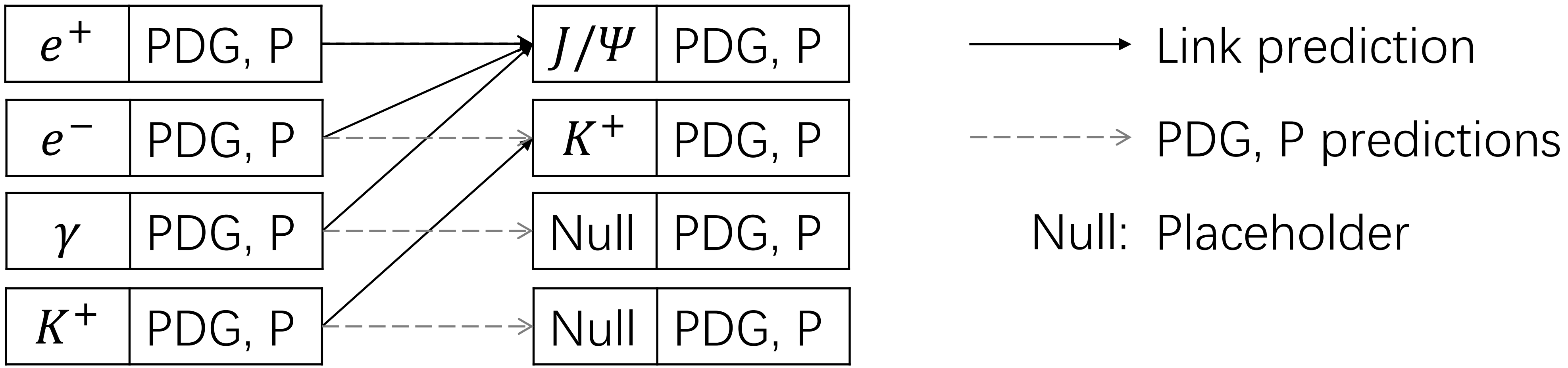
Goal:

- No restrictions on decay channels
  -> Predictions of particles instead of estimations of decay probabilities
  -> PDG (Particle type) + P (Four momentum) + Link predictions
- Only train a single model for all decay channels

Example:

Given final state particles (including particle information): $e^+e^-\gamma K^+$

- FEI: $p(J/\Psi K^+ \rightarrow e^+e^-\gamma K^+) > p(\pi^0 K^+ \rightarrow e^+e^-\gamma K^+) > \cdots$
- New:



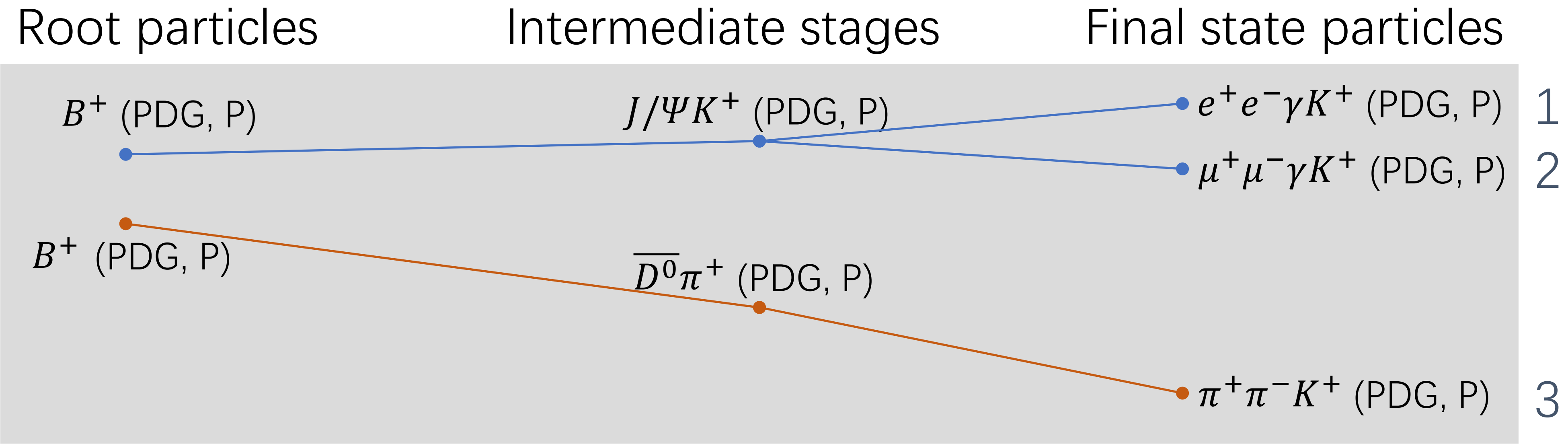
Improvement

Goal:

- No restrictions on available particle types at each stage
  -> Looser definition of stages, but still hierarchical reconstruction
  -> Continuous representation of the decay information in an embedding space

Example:

Embedding space:

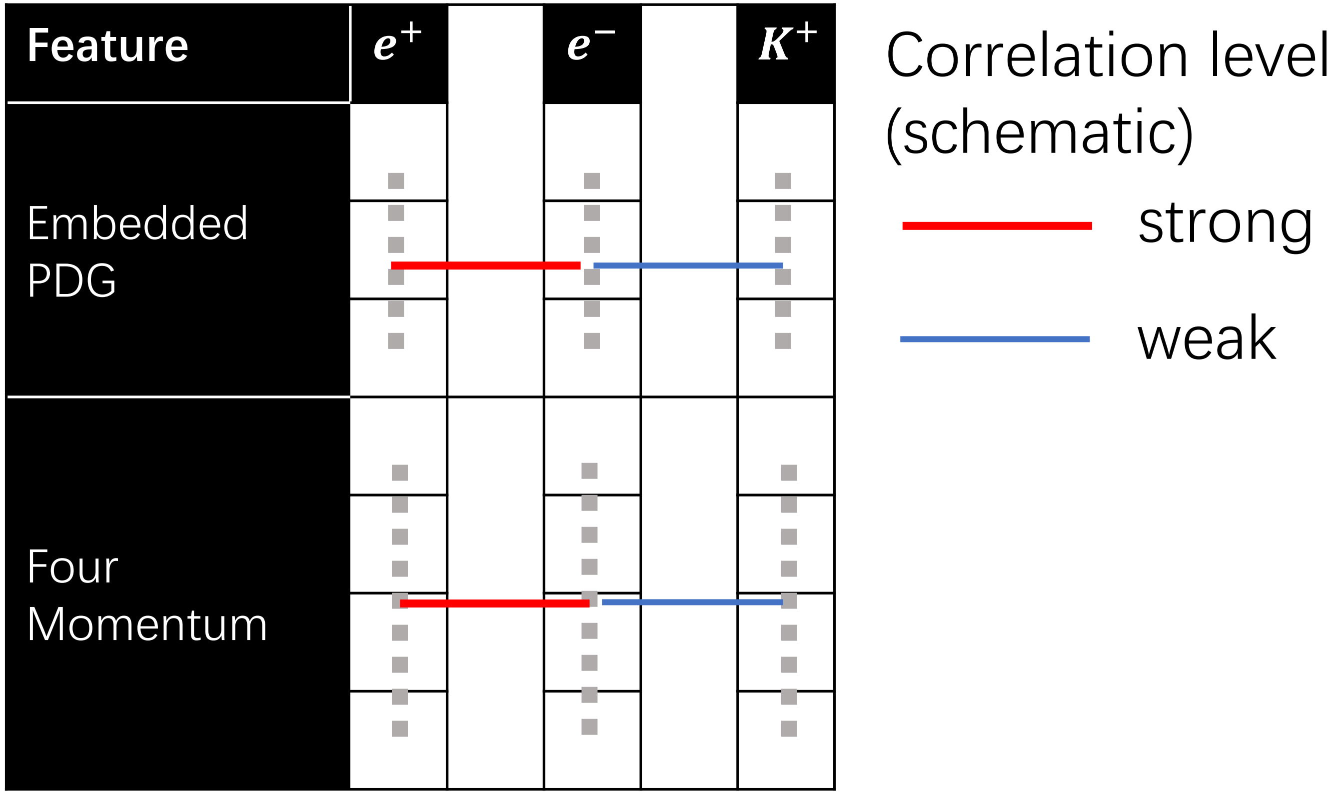


Building blocks

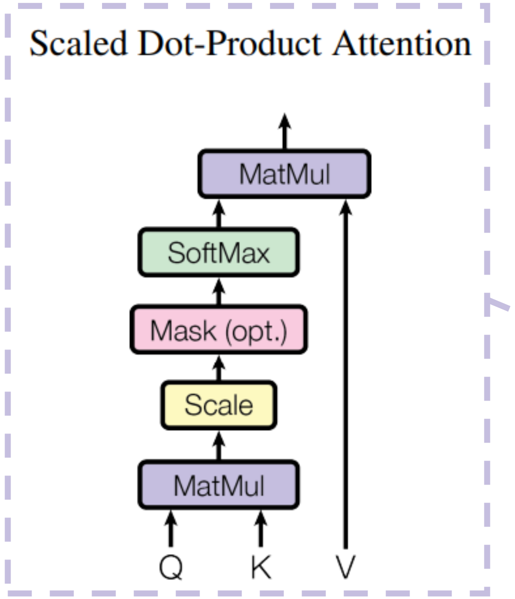
Transformer-based models:

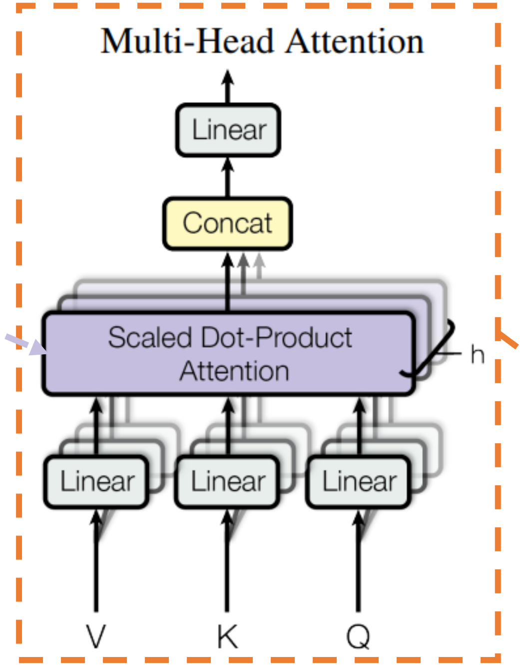
- High representative power (e.g., ChatGPT)
- Suitable for a variable number of particles as input
- Extracting high order correlations among particle features with attention mechanism
- Suit for various kinds of tasks (classification, regression, clustering) with the same basic network block

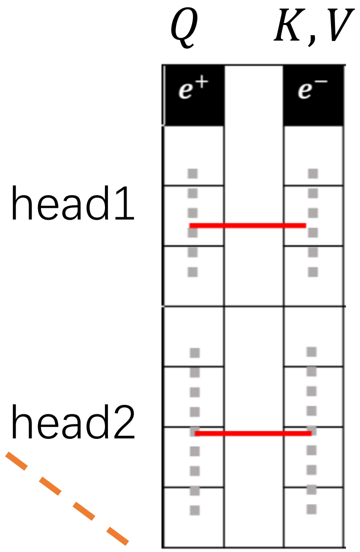
$$J/\Psi K^+ \rightarrow e^+ e^- \gamma K^+$$

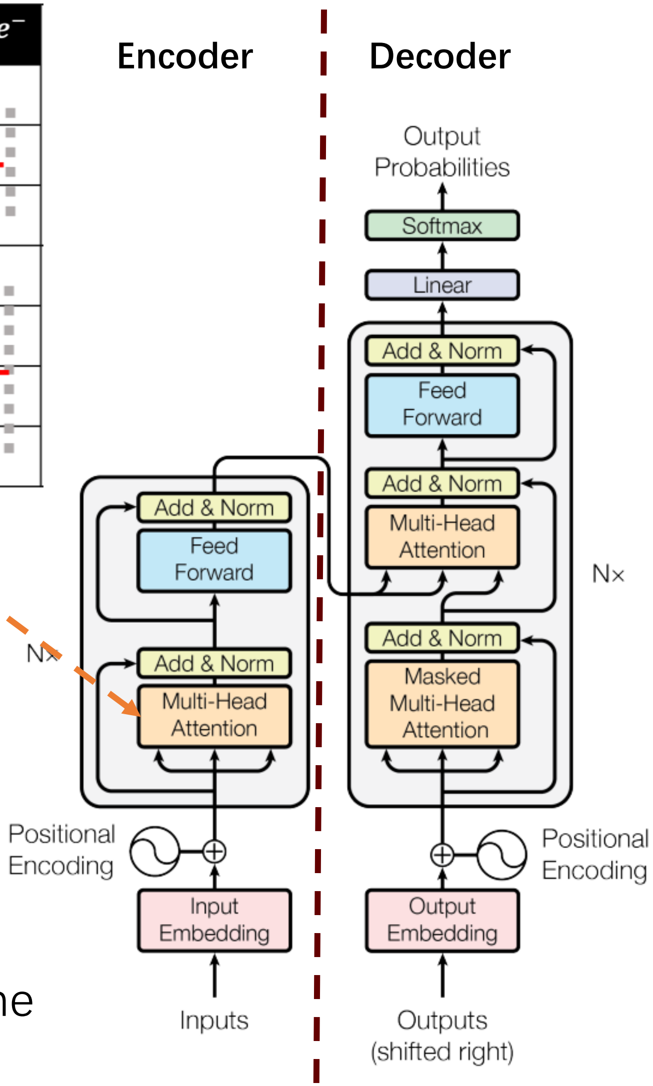


Transformer









- Encoder for embedding
- Decoder for reconstruction
- No positional encoding

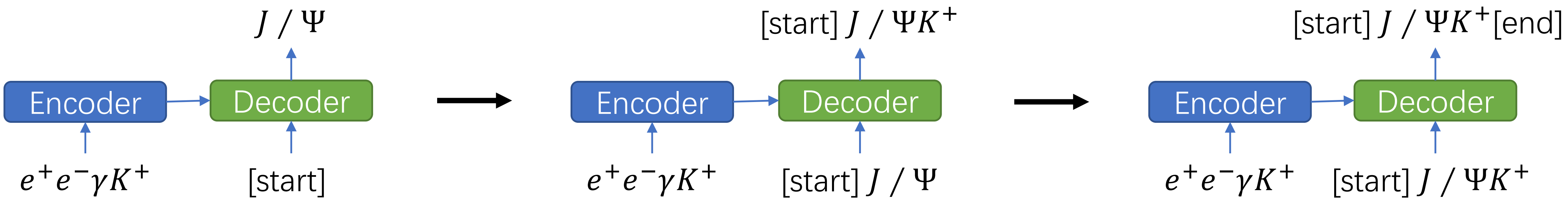
- Vectors:
  - Q: Query
  - K: Key
  - V: Value

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

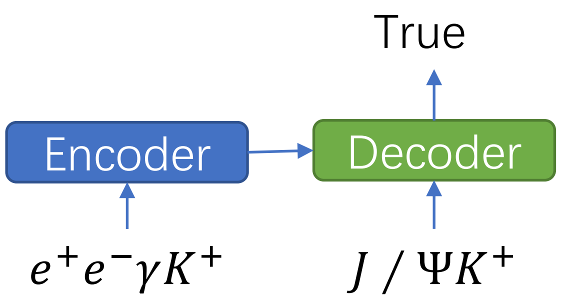
- Softmax represent the similarity of Q and K
- Multi-Head enables different combinations of the subspaces of the inputs through linear projections

Transformer structures

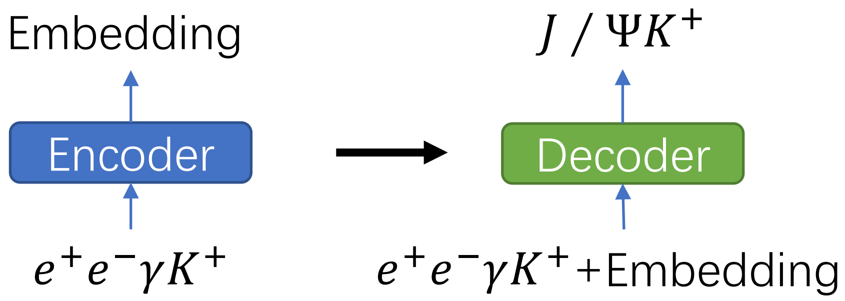
- GPT-like (original design, tested by Nikolai)



- BERT-like



- HyperTagging (ideal)

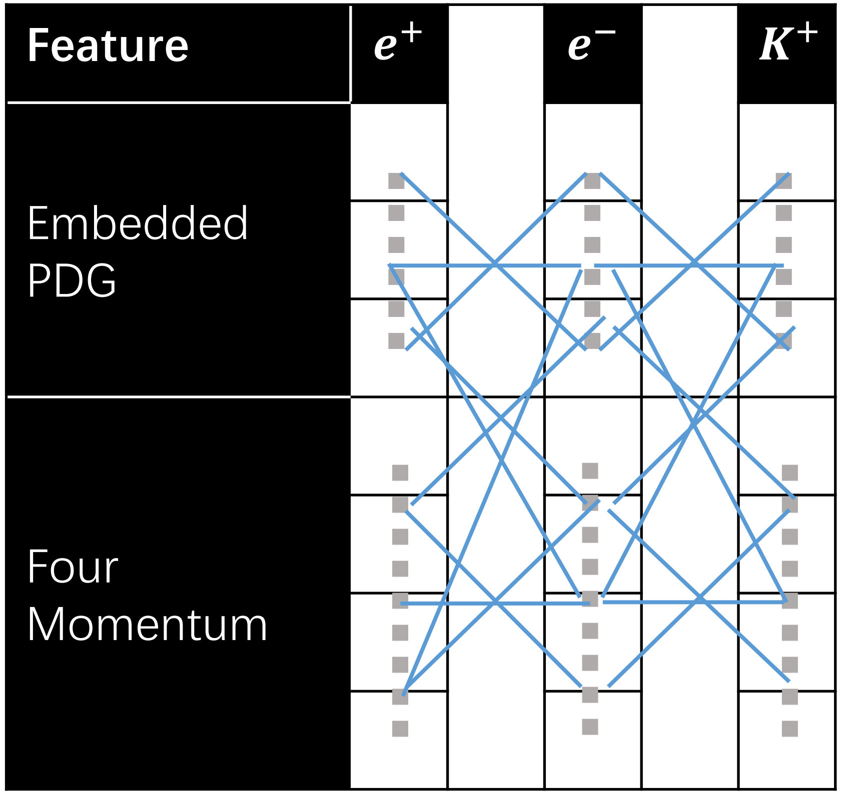


Building blocks

Interactor:

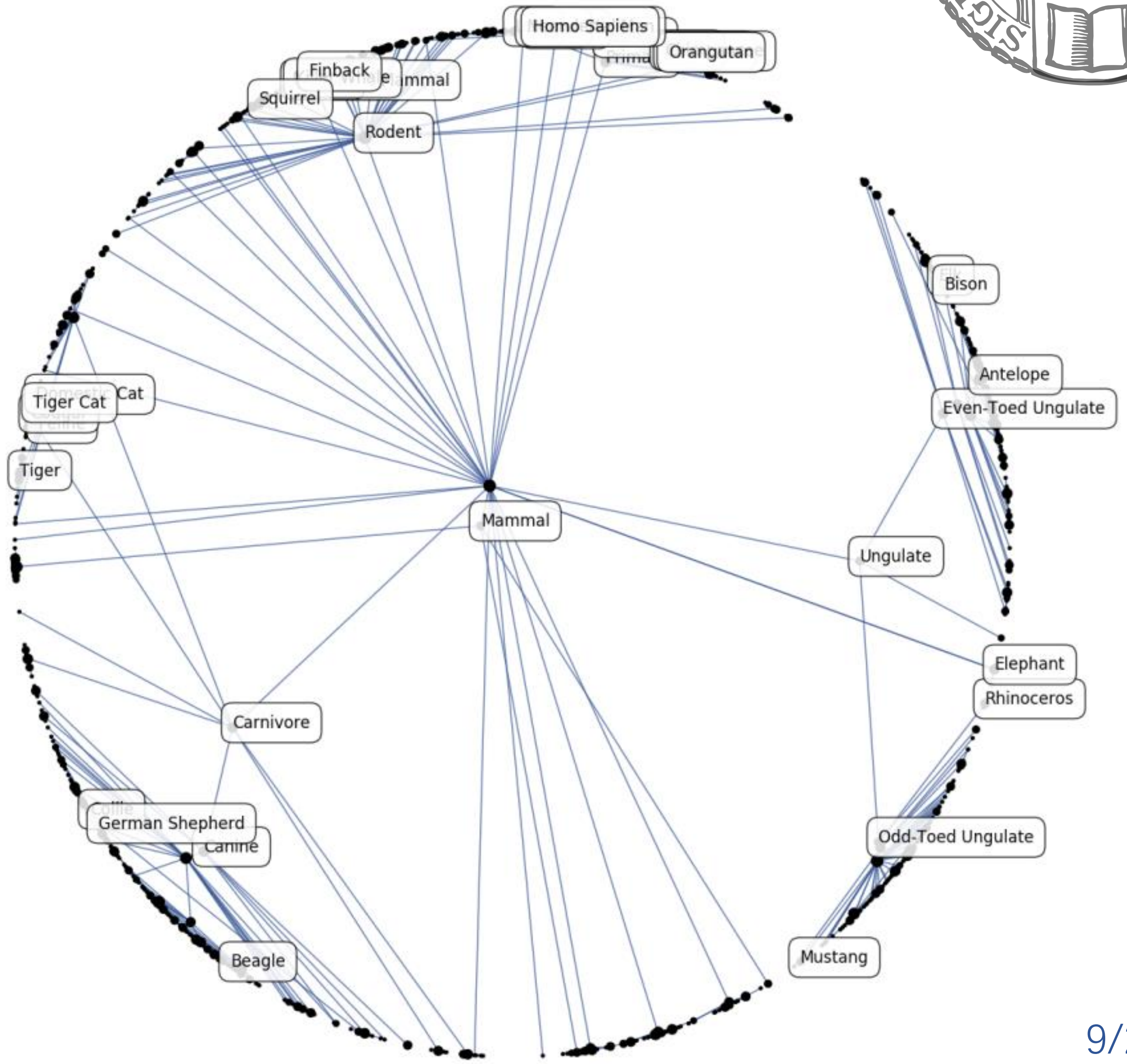
- Similiar to Transformers
- Powerful for sparse features (PDG, Charge, #Daughters···)
- Extracting high order correlations among different features from different particles with attention mechanism
- Better extracting particle level information

$$J/\Psi K^+ \rightarrow e^+ e^- \gamma K^+$$



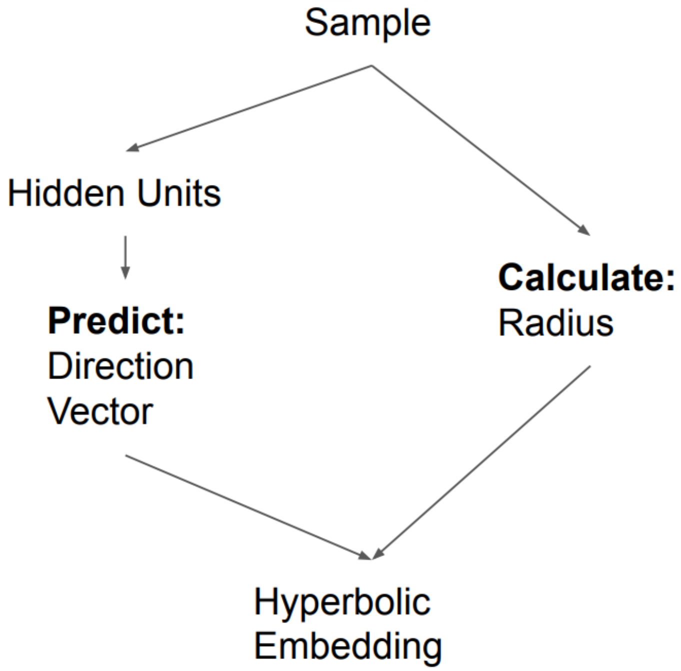
Hyperbolic embedding:

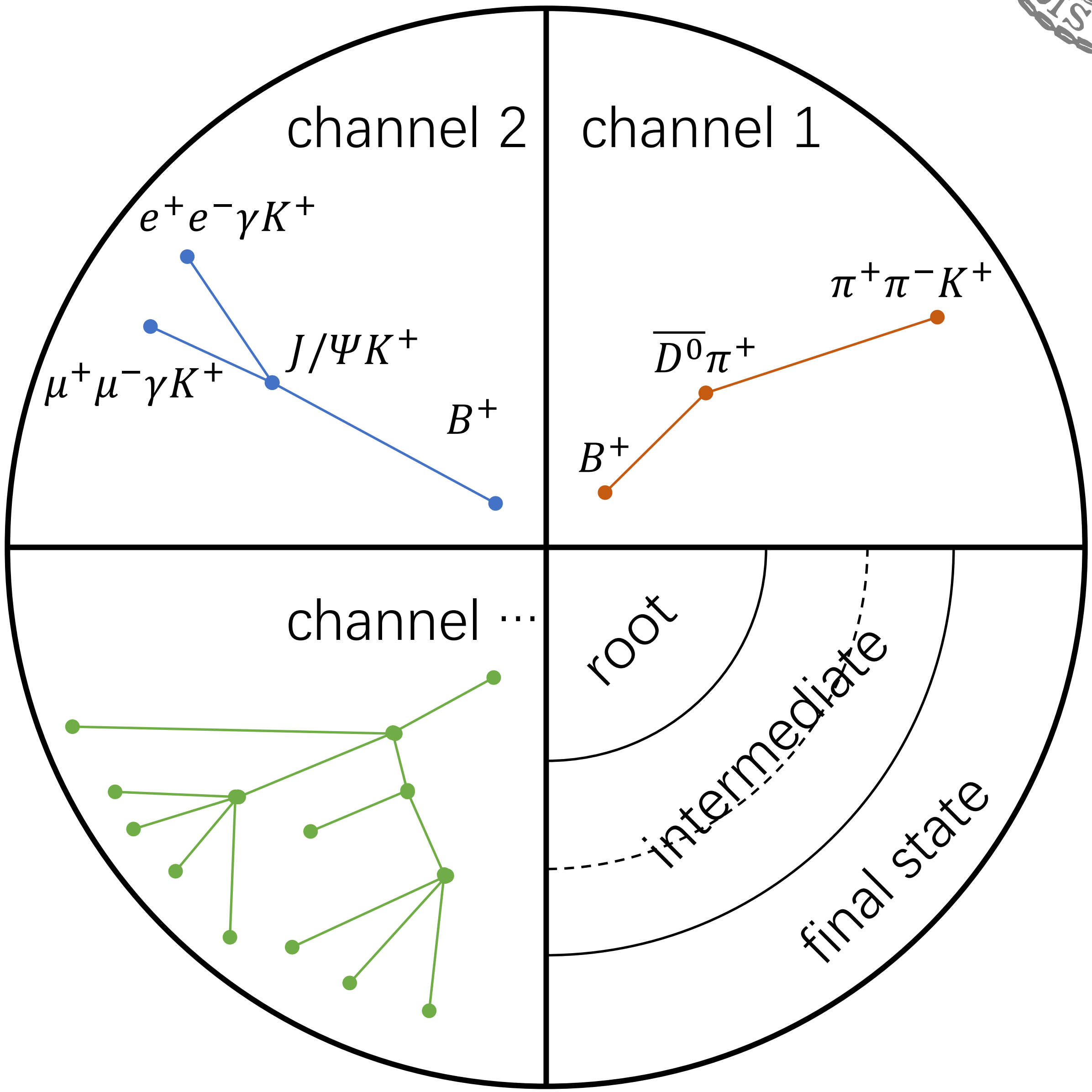
- High representative power for hierarchical clustering tasks



Hyperbolic embedding:

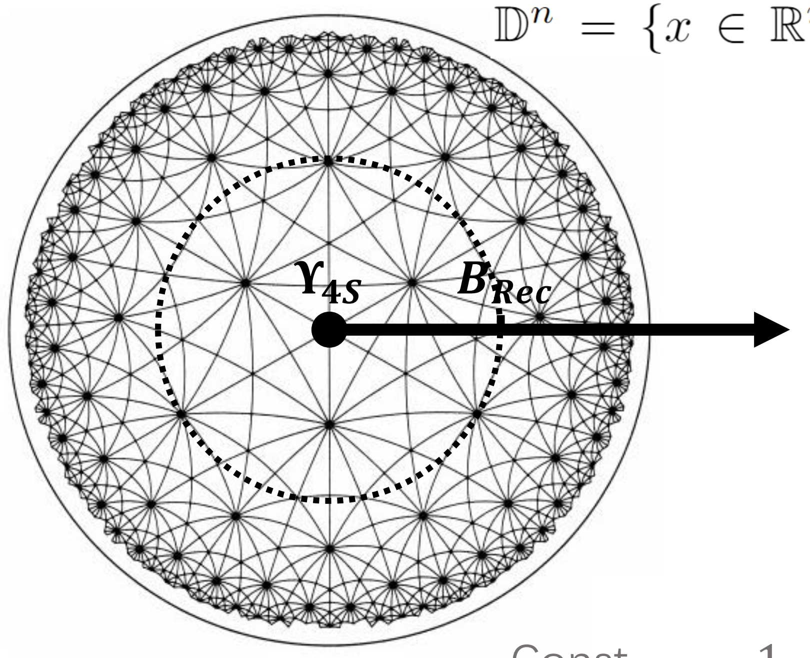
- High representative power for hierarchical clustering tasks
- Forcing the network to self-study physics information by clustering task





Hyperbolic Space (2D example – Poincare disc)

$$\mathbb{D}^n = \{x \in \mathbb{R}^n : c\|x\|^2 < 1, c \geq 0\}$$



Const. $\sim 1 - Radius^2$

$$E_{\text{ROE}} = E_{\Upsilon(4S)} - E_{\text{Reconstructed}}$$

$$\rightarrow r \sim \sqrt{E_{\text{ROE}}}$$

Properties:

- The size of an object with distance *d* to the center ~ $1 - d^2$
  -> Embedded events will never reach the boundary
  -> Effective space near the boundary is infinite
- Volume of the space scales exponentially with radius
  -> Comparable to tree-structured data (decay relations)

Metrics:

$$\mathbf{x} \oplus_c \mathbf{y} = \frac{(1 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c\|\mathbf{y}\|^2)\mathbf{x} + (1 - c\|\mathbf{x}\|^2)\mathbf{y}}{1 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c^2\|\mathbf{x}\|^2\|\mathbf{y}\|^2}$$

- Hyperbolic distance

$$D_{hyp}(\mathbf{x}, \mathbf{y}) = \frac{2}{\sqrt{c}} \text{arctanh}(\sqrt{c}\| - \mathbf{x} \oplus_c \mathbf{y}\|)$$

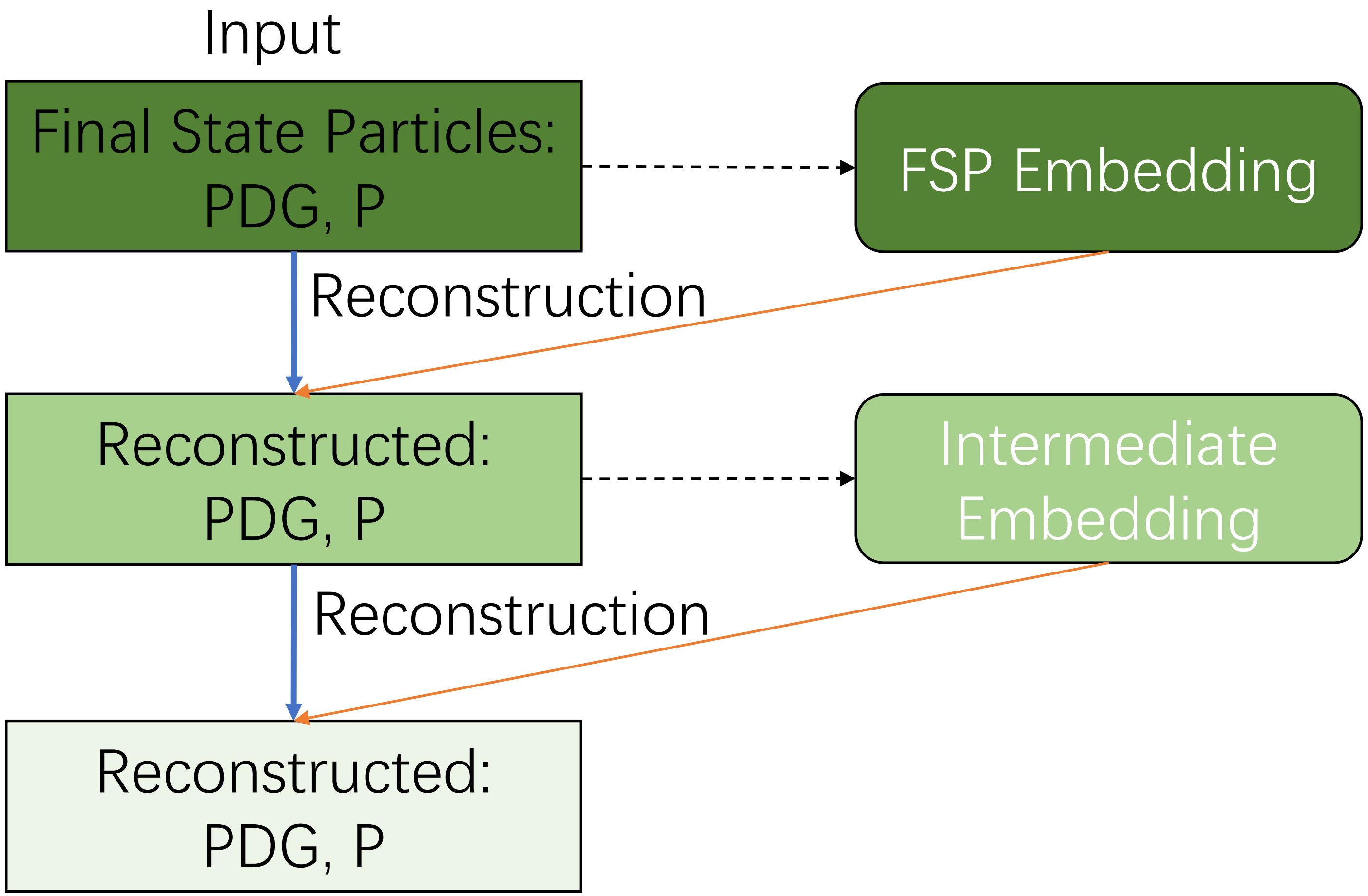
- Hyperbolic angle/cosine similarity (the same as euclidical)

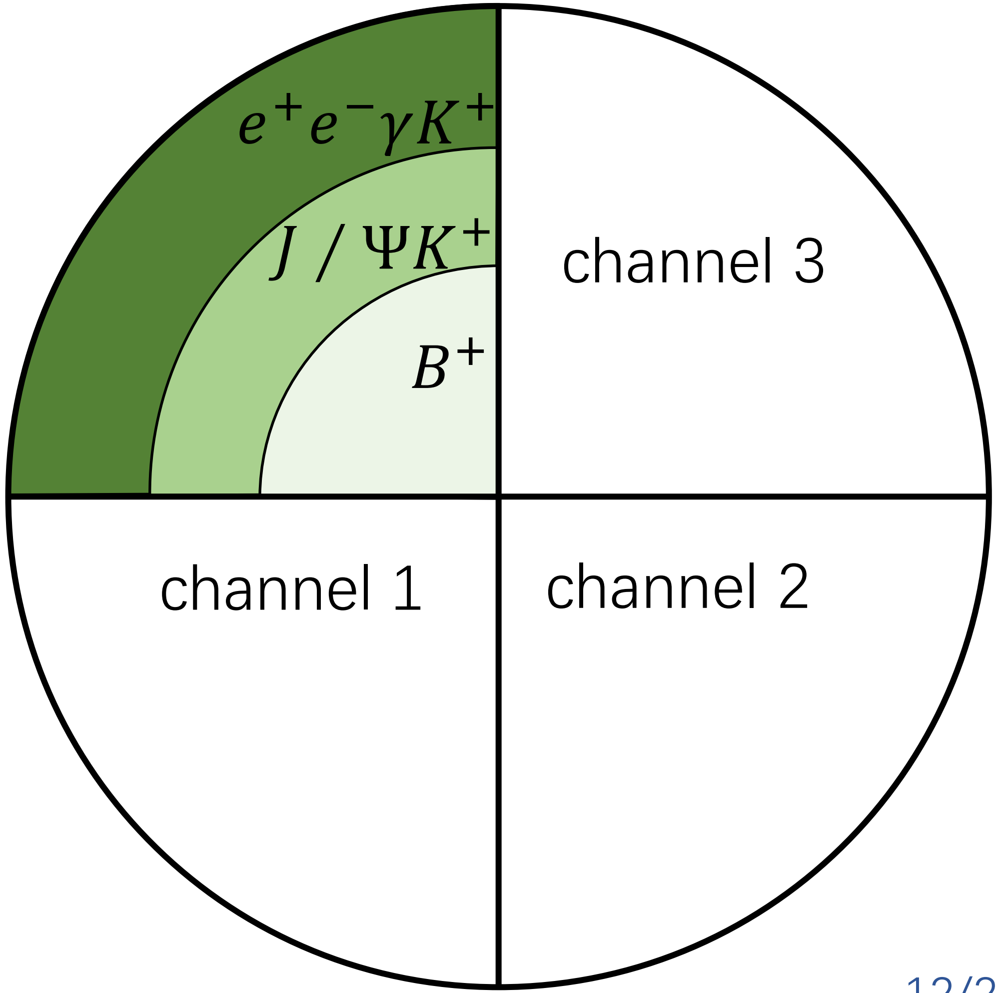
$$D_{cos}(\mathbf{z}_i, \mathbf{z}_j) = \left\| \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \right\|_2^2 = 2 - 2\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\|\mathbf{z}_i\|_2 \cdot \|\mathbf{z}_j\|_2}$$

Workflow

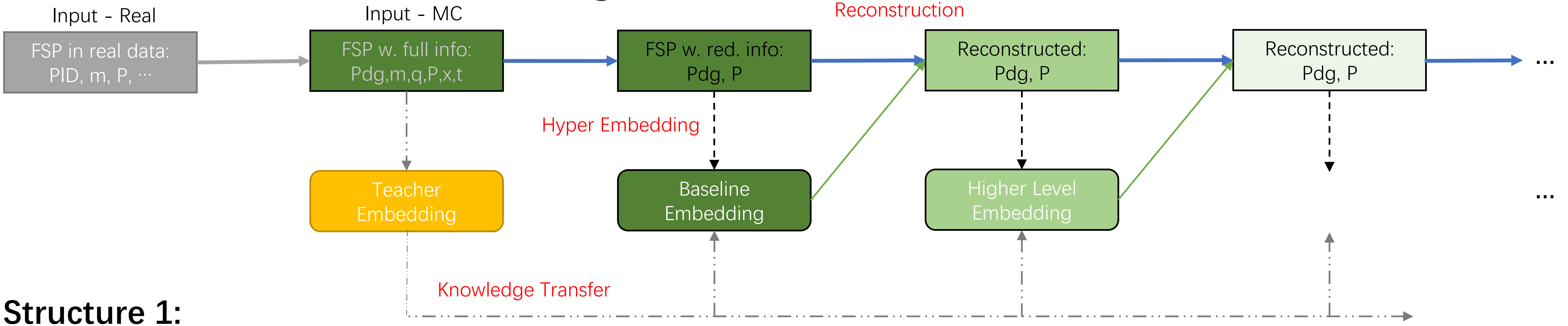
Hierarchical reconstruction with well trained embedding

$e^+e^-\gamma K^+$ → $J/\Psi K^+$ → B^+

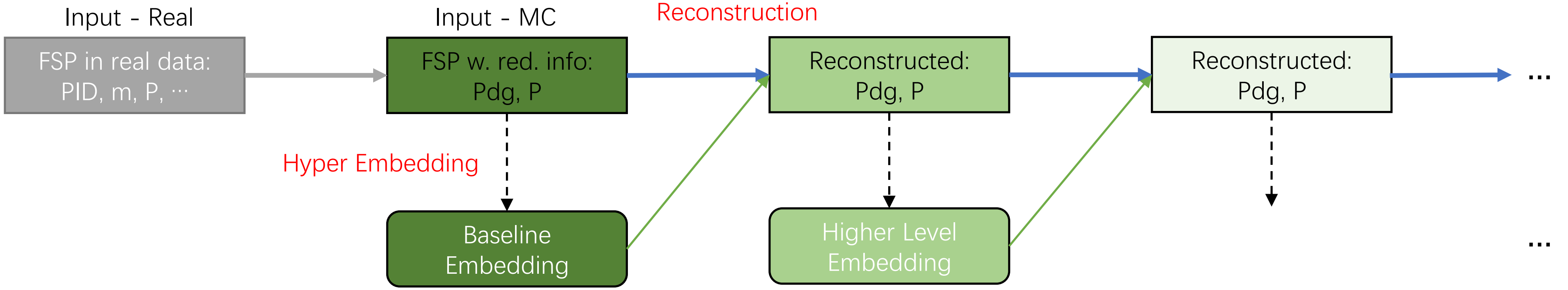




Hierarchical reconstruction with well trained embedding

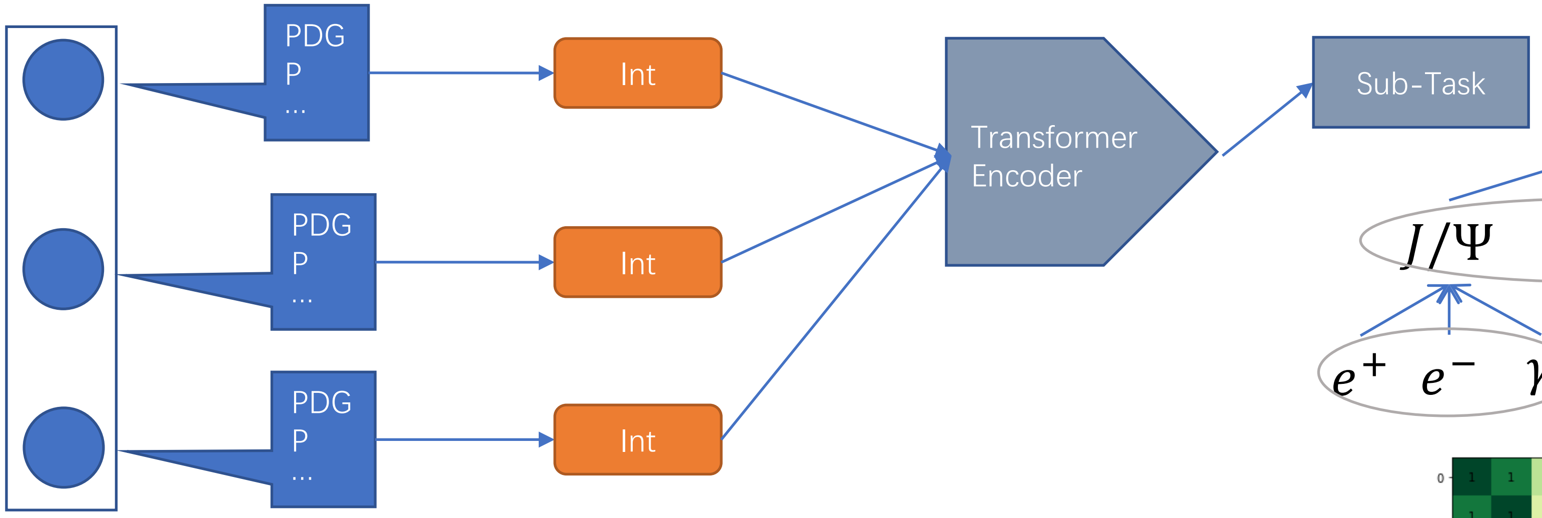


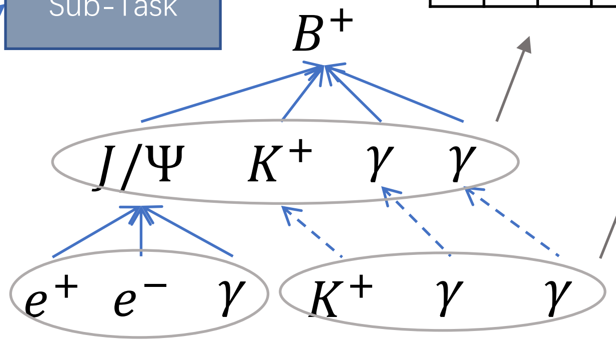
Structure 1:



Structure 2:

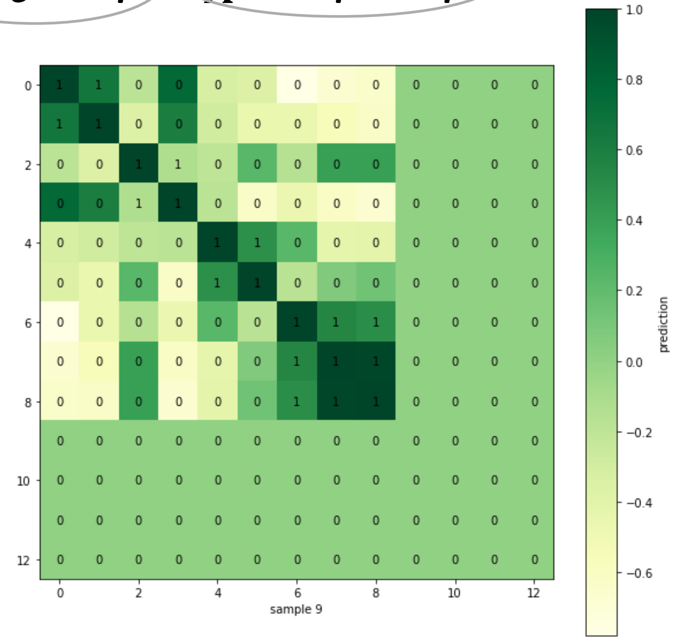
Particle level embedding
(pre-training of Sample level embedding)



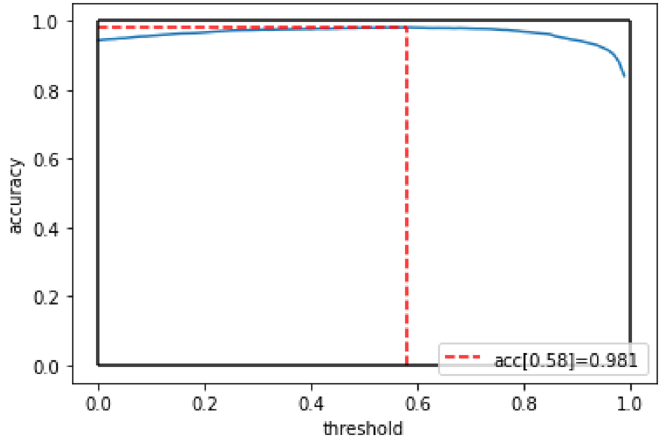


1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

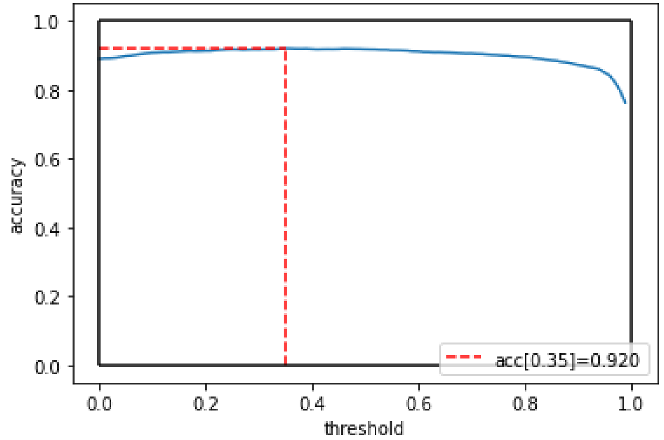
1	1	1	0	0	0
1	1	1	0	0	0
1	1	1	0	0	0
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1



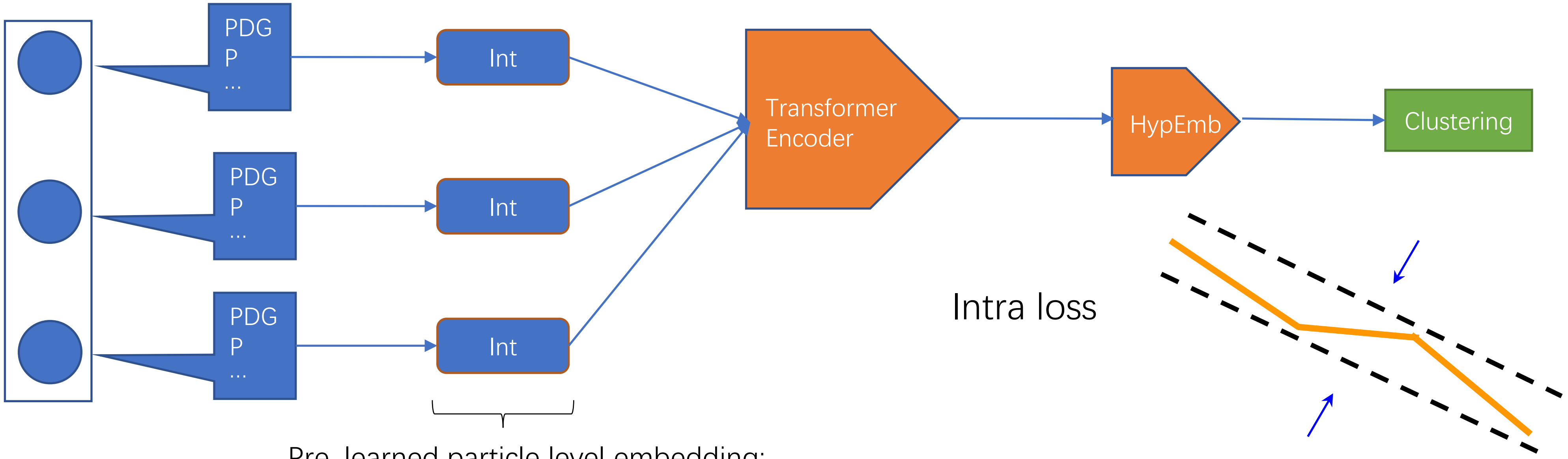
Structure 1



Structure 2

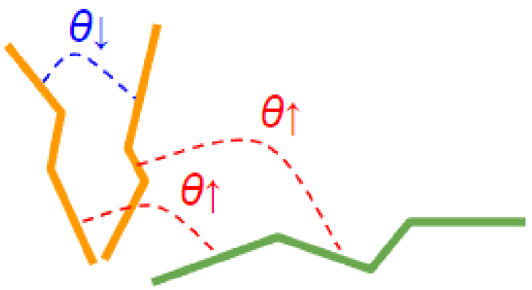


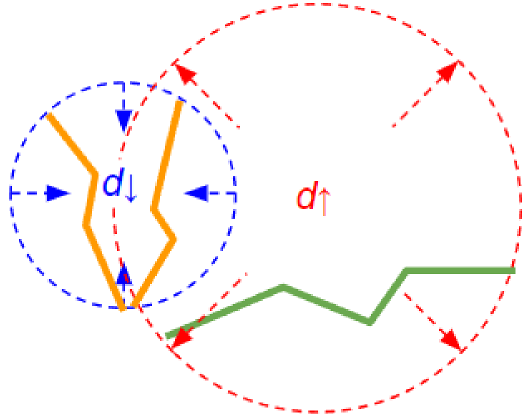
Sample level embedding



Pre-learned particle level embedding:
Frozen at the beginning of the trainings

Intra loss





Inter loss - Angle

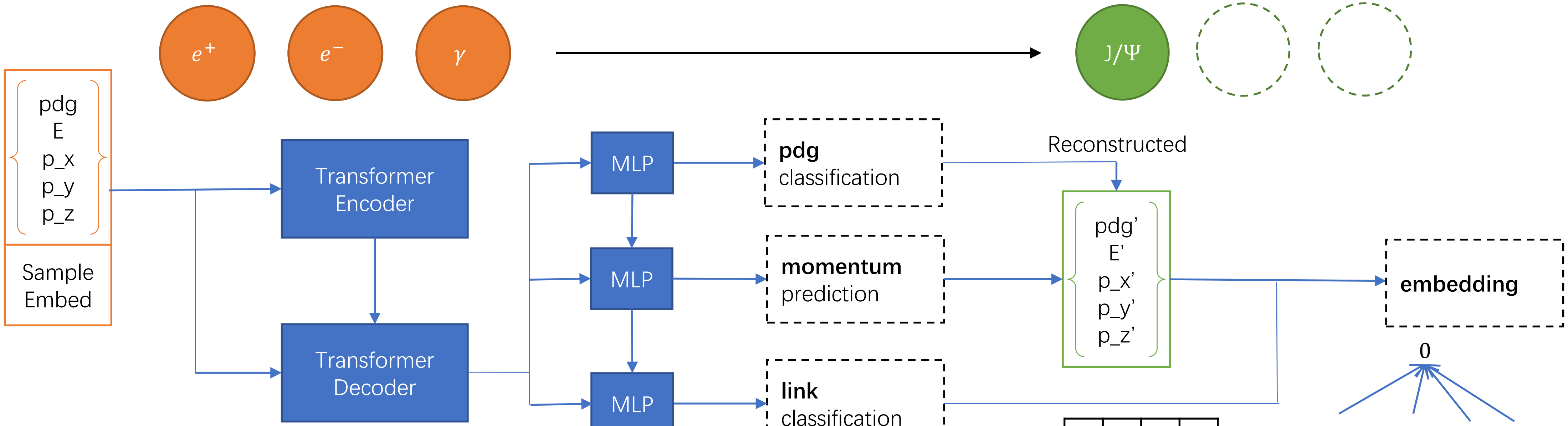
Inter loss - Distance

Clustering according to combined losses:

- Classes Intra vs Inter:
  Labeled with event number vs channel
  Representing same decay event vs same decay channel
- Metric Angle and Distance

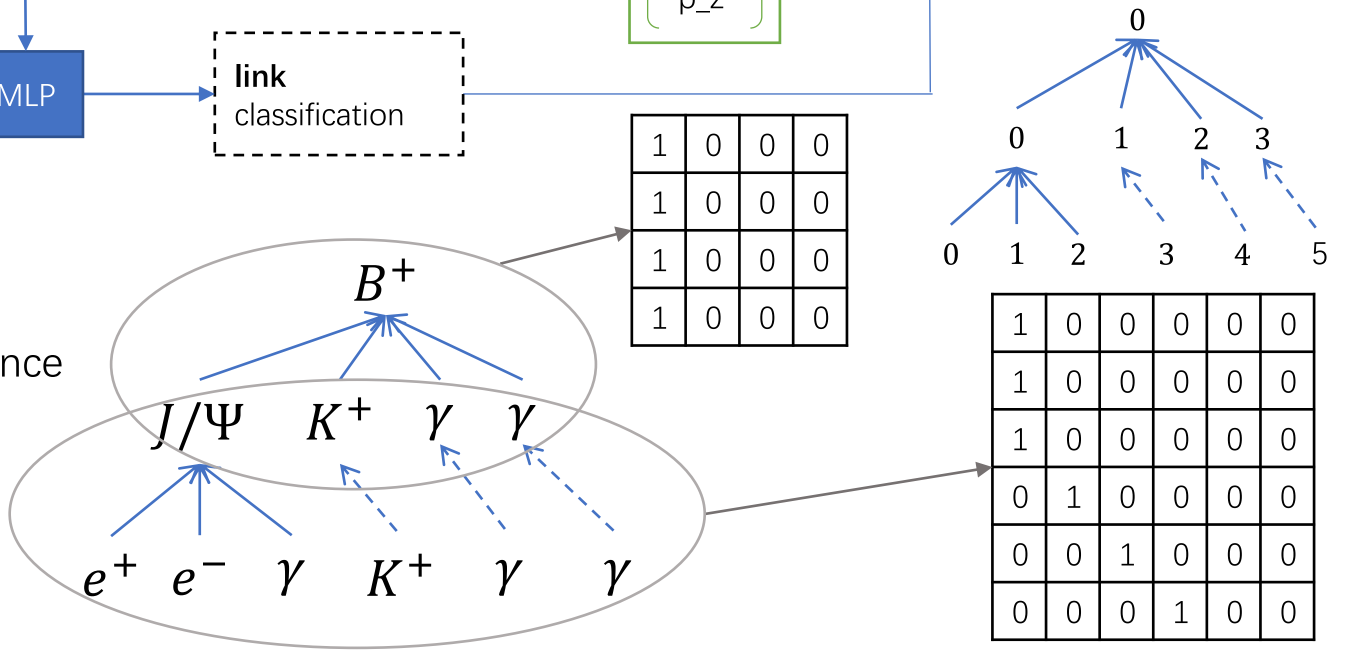
Reconstruction

Input



Multi-task-losses:

- PDG classification: Cross entropy
- Momentum prediction: Mean square error
- Link prediction: Cross entropy of link matrix
- Embedding loss: (Discriminator): Hyperbolic distance



Structures overview

Stage	Neural Networks	Model Size	Task	Technics
Particle Level Embedding	Automatic Feature Interaction (AutoInt) + Transformer Encoder	11K 3.5K	Prediction of combinations of daughter particles	Supervised pre-training
Sample Level Embedding	Transformer Encoder + Hyperbolic Embedding (HypTr)	900K 460K	Learning the representation of decays in hyperbolic space	Unsupervised training
Variable Reduction (Structure 1 only)	Transformer Encoder + Hyperbolic Embedding (HypTr)	300K -	Imitating the sample level embedding with less information	Knowledge transfer
Reconstruction (Structure 2 only)	Transformer Encoder + Decoder + MLP	- 200K	Prediction of reconstructed PDG, four momentum (and link)	Multi-task supervised training
Link Prediction (Structure 2.1 only)	Transformer Encoder	- 30K	Prediction of link	Supervised training

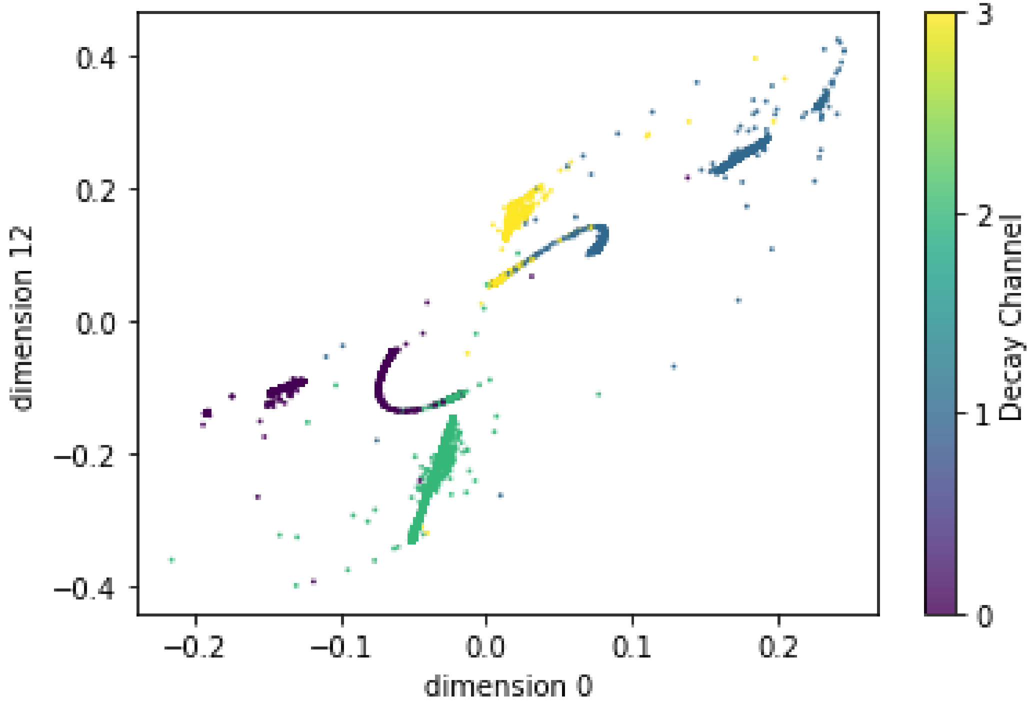
Dataset (proof of concept)

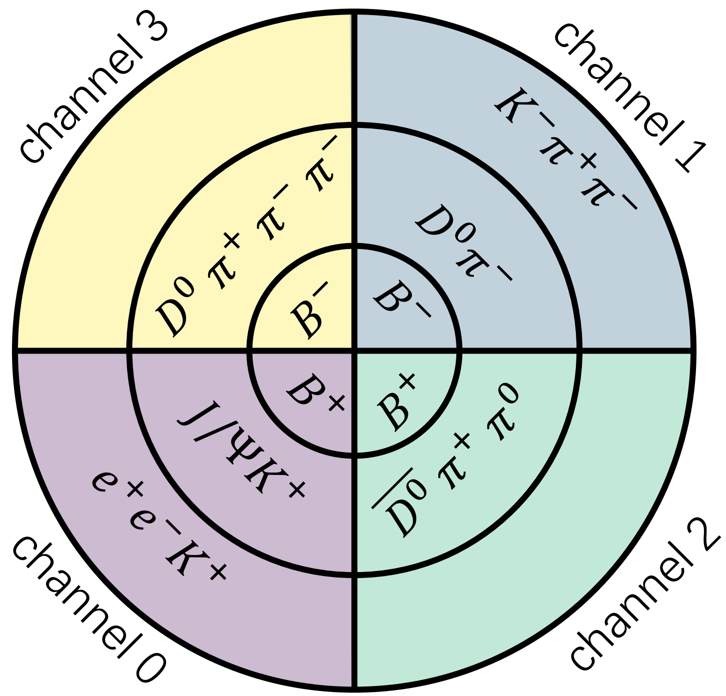
- Monte Carlo truth information from four B decay channels (25% * 4)

Performances

- **Clustering performance:**

  2D slice from 16D embedding space, colored with channels





Dataset (proof of concept)

- Monte Carlo truth information from four B decay channels (25% * 4)

Performances

- **Reconstruction performance**

Task	Evaluation Metric	3-Task Training	2+1-Task Training
PDG prediction	Accuracy: $\frac{\#\text{correctly predicted PDGs incl. Nulls}}{\#\text{all PDGs incl. Nulls}}$	**92%**	84%
Four momentum prediction	Mean absolute error: $mean(abs(P_{\text{pred}} - P_{\text{truth}}))$	0.087GeV	0.083GeV
Link prediction	Accuracy: $\frac{\#\text{correctly predicted links}}{\#\text{all links}}$	80%	**95%**

Experiments

Dataset (generic)

- Monte Carlo truth information from $\Upsilon(4S)$ decays
  - Four $B \rightarrow K\nu\nu$ signal channels (2.5% * 4)
  - Two generic $B^{0,\pm}$ datasets (45% * 2)
- No definition of “same channel“ -> new metric $M_{i,j}$ to replace channels in the loss function:

$$l_{i,j} = -\delta_{c_i,c_j} \log \frac{\exp(-D(z_i, z_j))}{\sum_k \exp(-D(z_i, z_k))} \rightarrow -M_{i,j} \log \frac{\exp(-D(z_i, z_j))}{\sum_k \exp(-D(z_i, z_k))}$$

Summary

Achieved:

- Successful encoding of decay information into hyperbolic space
- Accurate predictions of PDG, four momentum and links after reconstructions

On going:

- Balancing the performances of the 3 predictions
  (trying to replace particle level embedding with link prediction for pre-training)
- Coding for the evaluation of the whole reconstruction
- Studying the performance on generic dataset

Plans:

- Adding more available information for each particle/event
- Changing from simulation truth to reconstructed information
  -> Enable ansatz on real data

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

Thank You for your Attention

Boyang Yu
Ludwig-Maximilians-Universität München
EPD Seminar, IHEP Beijing, April 18th, 2023

Boyang.Yu@physik.uni-muenchen.de

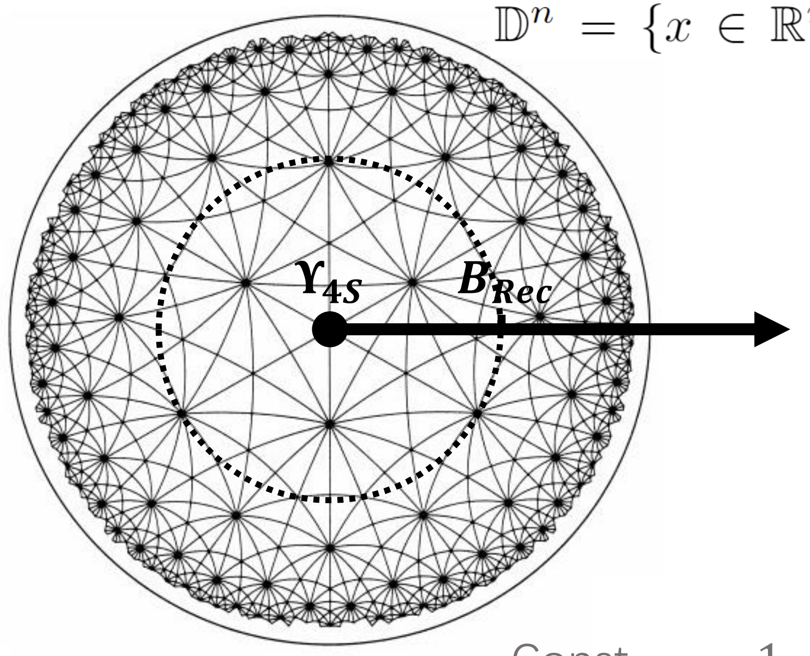
Reference:

- **FEI:** T. Keck et al. "The Full Event Interpretation --An exclusive tagging algorithm for the Belle II experiment", *arXiv:1807.08680*
- **Transformers:** A. Vaswani et al. "Attention Is All You Need", *arXiv:1706.03762*
- **Interactor:** W. Song et al. "AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks", *arXiv:1810.11921*
- **GPT3:** Tom B. Brown et al. "Language Models are Few-Shot Learners", *arXiv:2005.14165*
- **BERT:** Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv:1810.04805*
- **HyperVIT:** A. Ermolov et al. "Hyperbolic Vision Transformers: Combining Improvements in Metric Learning", *arXiv:2203.10833*
- **Hyperbolic metrics:** W. Peng et al. "Hyperbolic Deep Neural Networks: A Survey", *arXiv:2101.04562*

Backup

Hyperbolic Space (2D example – Poincare disc)

$$\mathbb{D}^n = \{x \in \mathbb{R}^n : c\|x\|^2 < 1, c \geq 0\}$$



Const. $\sim 1 - Radius^2$

$$E_{\text{ROE}} = E_{\Upsilon(4S)} - E_{\text{Reconstructed}}$$

$$\rightarrow r \sim \sqrt{E_{\text{ROE}}}$$

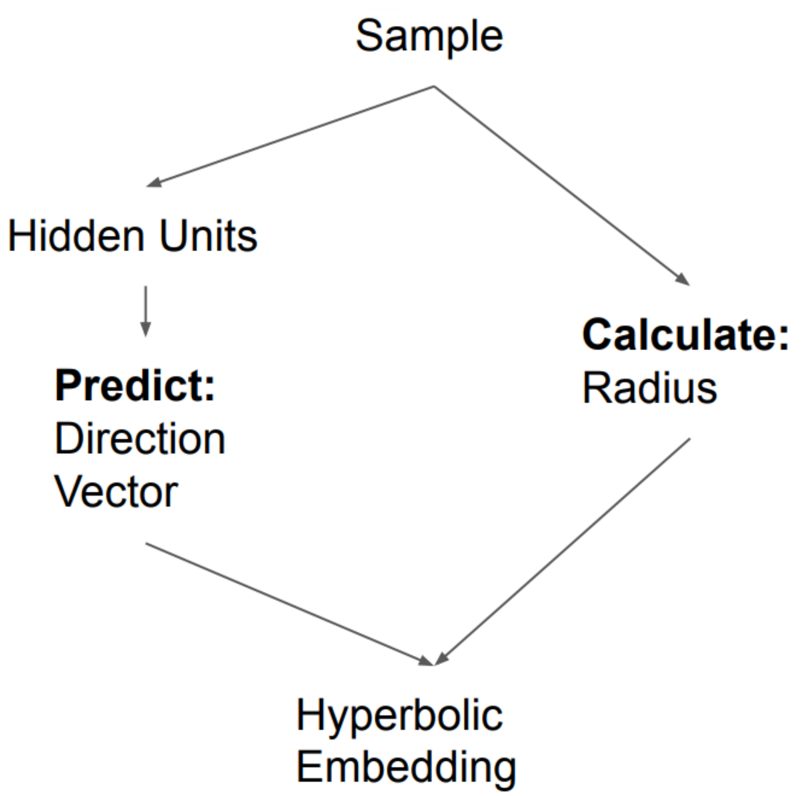
Properties:

- The size of an object with distance d to the center ~ $1 - d^2$
  -> Embedded events will never reach the boundary
  -> Effective space near the boundary is infinite
- Volume of the space scales exponentially with radius
  -> Comparable to tree-structured data (decay relations)

Ideal Embedding:

- Center: Singularity containing all full reconstructions of $\Upsilon(4S)$
  -> Empty rest of event (ROE)
- Bulk points: Partially reconstructed decays
- Points near boundary: Starting points of reconstructions
  -> The less reconstructed, the smaller branching ratio (taking less place in embedded space)
  -> Enable all possible decays

Hyperbolic Embedding



Hyperbolic metrics

Addition:

$$\mathbf{x} \oplus_c \mathbf{y} = \frac{(1 + 2c\langle \mathbf{x}, \mathbf{y}\rangle + c\|\mathbf{y}\|^2)\mathbf{x} + (1 - c\|\mathbf{x}\|^2)\mathbf{y}}{1 + 2c\langle \mathbf{x}, \mathbf{y}\rangle + c^2\|\mathbf{x}\|^2\|\mathbf{y}\|^2}$$

Distance:

$$D_{hyp}(\mathbf{x}, \mathbf{y}) = \frac{2}{\sqrt{c}}\operatorname{arctanh}(\sqrt{c}\| - \mathbf{x} \oplus_c \mathbf{y}\|)$$

Exponential:

$$\exp_{\mathbf{x}}^c(\mathbf{v}) = \mathbf{x} \oplus_c \left(\tanh\left(\sqrt{c}\frac{\lambda_{\mathbf{x}}^c\|\mathbf{v}\|}{2}\right)\frac{\mathbf{v}}{\sqrt{c}\|\mathbf{v}\|}\right)$$

with x the base point, usually set to 0

Hyperbolic metrics

- Hyperbolic distance $\mathbf{x} \oplus_c \mathbf{y} = \frac{(1 + 2c\langle \mathbf{x}, \mathbf{y}\rangle + c\|\mathbf{y}\|^2)\mathbf{x} + (1 - c\|\mathbf{x}\|^2)\mathbf{y}}{1 + 2c\langle \mathbf{x}, \mathbf{y}\rangle + c^2\|\mathbf{x}\|^2\|\mathbf{y}\|^2}$

$$D_{hyp}(\mathbf{x}, \mathbf{y}) = \frac{2}{\sqrt{c}} \operatorname{arctanh}(\sqrt{c}\| - \mathbf{x} \oplus_c \mathbf{y}\|)$$

- Hyperbolic angle/cosine similarity (the same as euclidical)

$$D_{cos}(\mathbf{z}_i, \mathbf{z}_j) = \left\| \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \right\|_2^2 = 2 - 2\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\|\mathbf{z}_i\|_2 \cdot \|\mathbf{z}_j\|_2}$$

- Cross entropy losses w.r.t the two metrics for positive pairs (i, j)

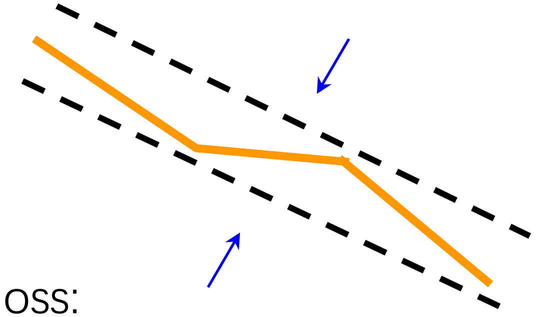
$$l_{i,j} = -\log \frac{\exp\left(-D(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1, k\neq i}^{K} \exp\left(-D(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)}$$

- Cross entropy losses for general cases
  (no well defined decay channels to specify “positive pairs”)

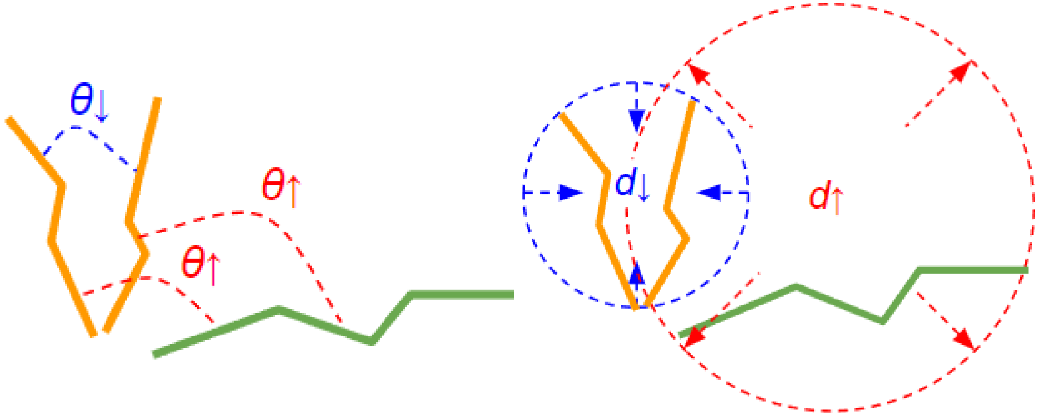
$$l_{i,j} = -\delta_{c_i, c_j} \log \frac{\exp\left(-D(\mathbf{z}_i, \mathbf{z}_j)\right)}{\sum_k \exp\left(-D(\mathbf{z}_i, \mathbf{z}_k)\right)} \rightarrow -\log \frac{M_{i,j} \exp\left(-D(\mathbf{z}_i, \mathbf{z}_j)\right)}{\sum_k M'_{i,k} \exp\left(-D(\mathbf{z}_i, \mathbf{z}_k)\right)}$$

Embedding losses

- Intra loss:
  Align the samples from the same decay event



- Inter loss:
  Cluster/Separate the samples according to their decay channels



Proof of concept: Toy Monte Carlo

Dataset:

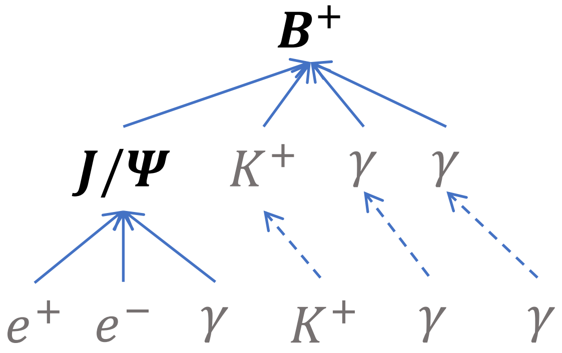
Four channels:

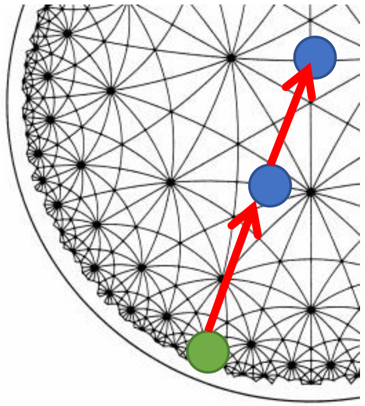
- $B^+ \to (J/\Psi \to e^+e^-)K^+$
- $B^- \to (D^0 \to K^-\pi^+)\,\pi^-$
- $B^+ \to \overline{D^0}\,\pi^+\,\pi^0$
- $B^- \to D^0\,\pi^+\,\pi^-\,\pi^-$

```
particle_list = ['N.A.', 'Upsilon(4S)', 'gamma', 'K_L0', 'pi0', 'J/psi', 'K_S0',
        'e+', 'K+', 'pi+', 'mu+', 'p+', 'Lambda0', 'Sigma+', 'D+', 'D0',
        'D_s+', 'Lambda_c+', 'D*+', 'D*0', 'D_s*+', 'B0', 'B+', 'B_s0',
        'e-', 'K-', 'pi-', 'mu-', 'anti-p-', 'anti-Lambda0', 'anti-Sigma-', 'D-', 'anti-D0',
        'D_s-', 'anti-Lambda_c-', 'D*-', 'anti-D*0', 'D_s*-', 'anti-B0', 'B-', 'anti-B_s0'
      ] # len(particle_list) = 40 + 1 (empty)
```

Each event (Y4S Decay) produces several samples according to the depth of particles to its root B meson, e.g.

- Depth 1 (Sample 1)
  -> Embedding 3
- Depth 2 (Sample 2)
  -> Embedding 2
- Depth 3 (Sample 3)
  -> Embedding 1





$$r_i = 0.6\sqrt{1 - \frac{E_{\text{Reconstructed},i}}{E_{B^{+/-/0}}}} + 0.3$$

Each particle carries 12 features (**Bold** as reduced)

PDG, mass, charge, **energy**, production time, x, y, z, **px**, **py**, **pz**, nDaughters

Particles in each sample are sorted according to energy

Notice:

For each sample in the toy MC, all the FSPs come from the same mother B, i.e. the same channel

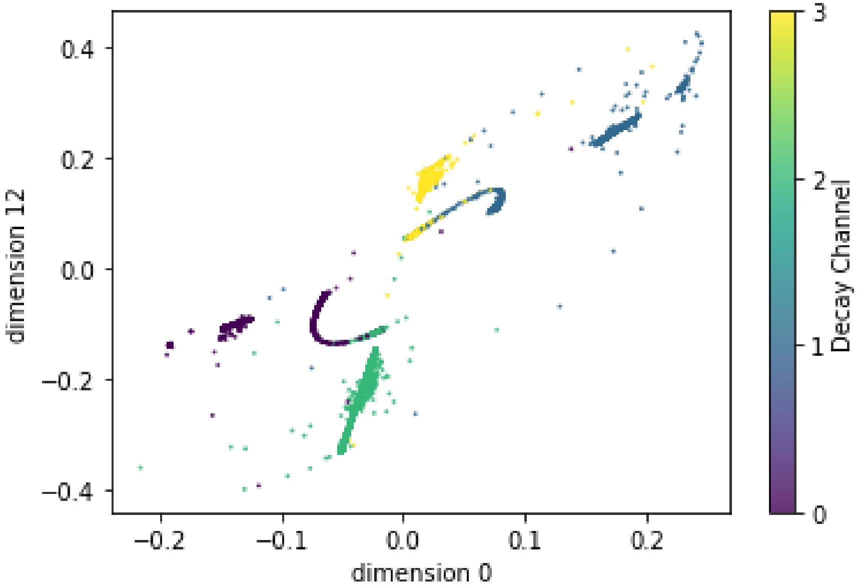
In the real case they are from two Bs or even background

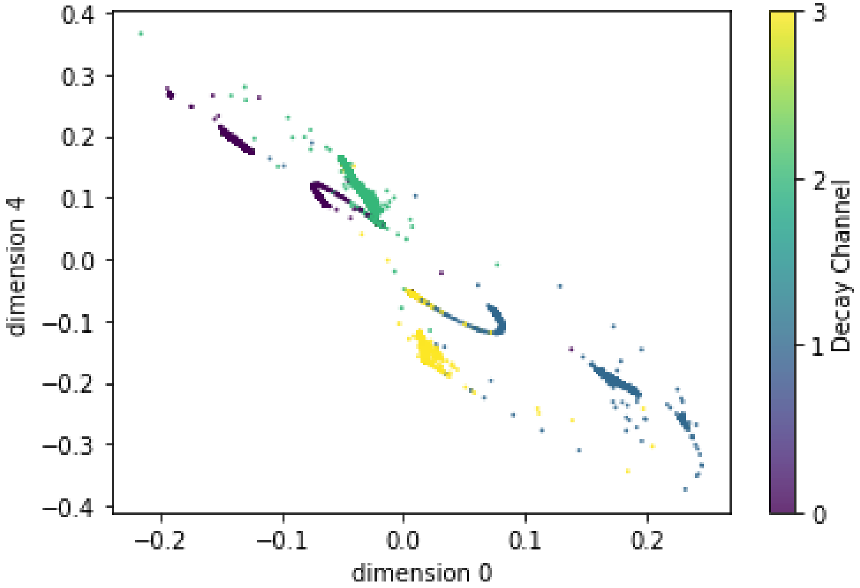
-> Need a better metric to do clustering instead of the channel of one mother B for real data

Sample Level Embedding:

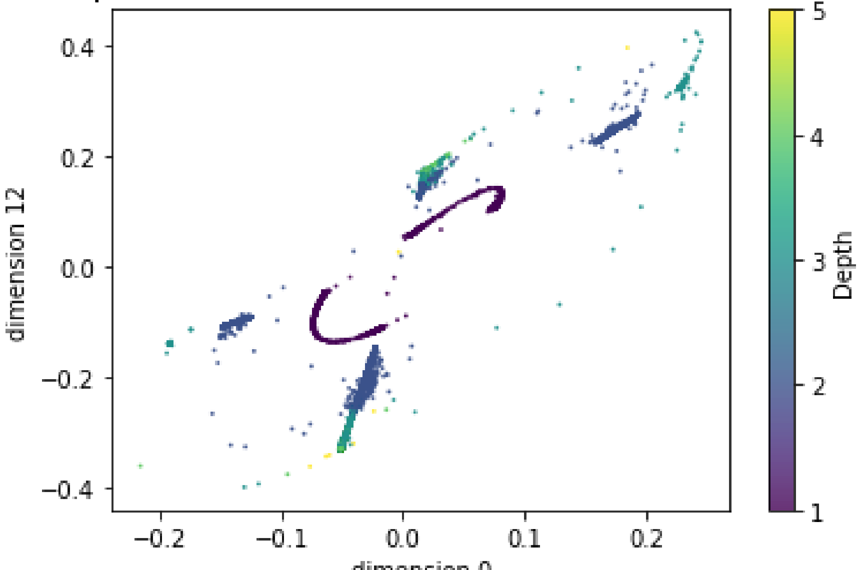
Visualisation with 16 dimensional hyperbolic embedding

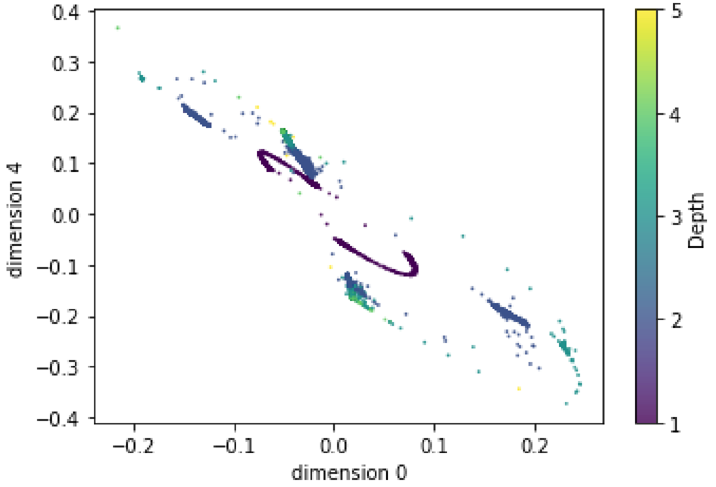
- Clustering vs. Decay channels





- Clustering vs. Depth



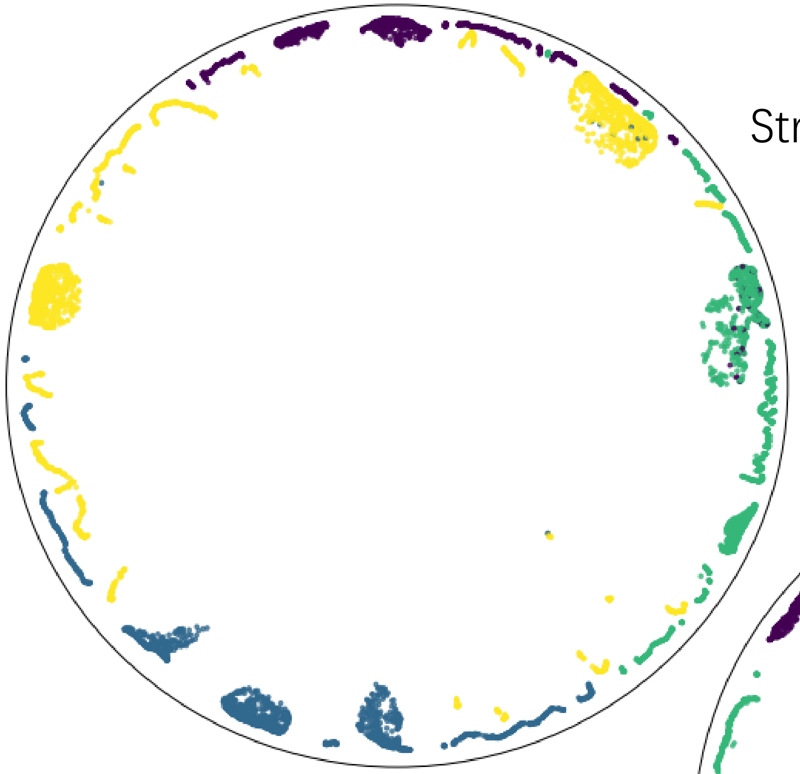


Sample Level Embedding:

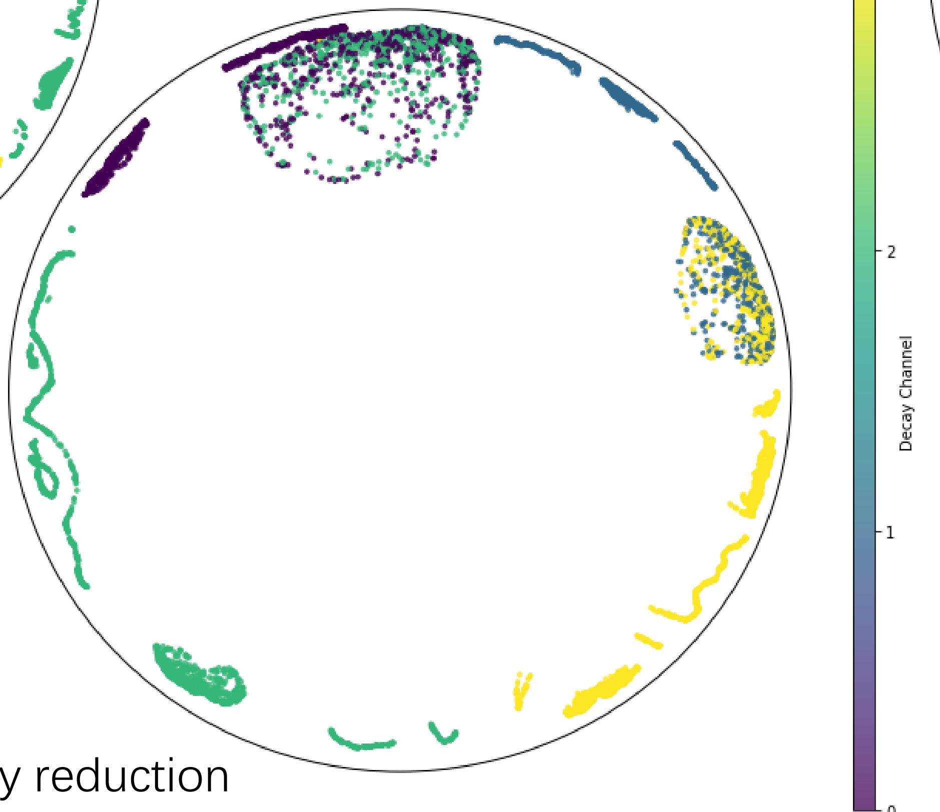
Visualisation with UMAP* for 16 dimensional hyperbolic embedding

Clustering vs. Decay channels

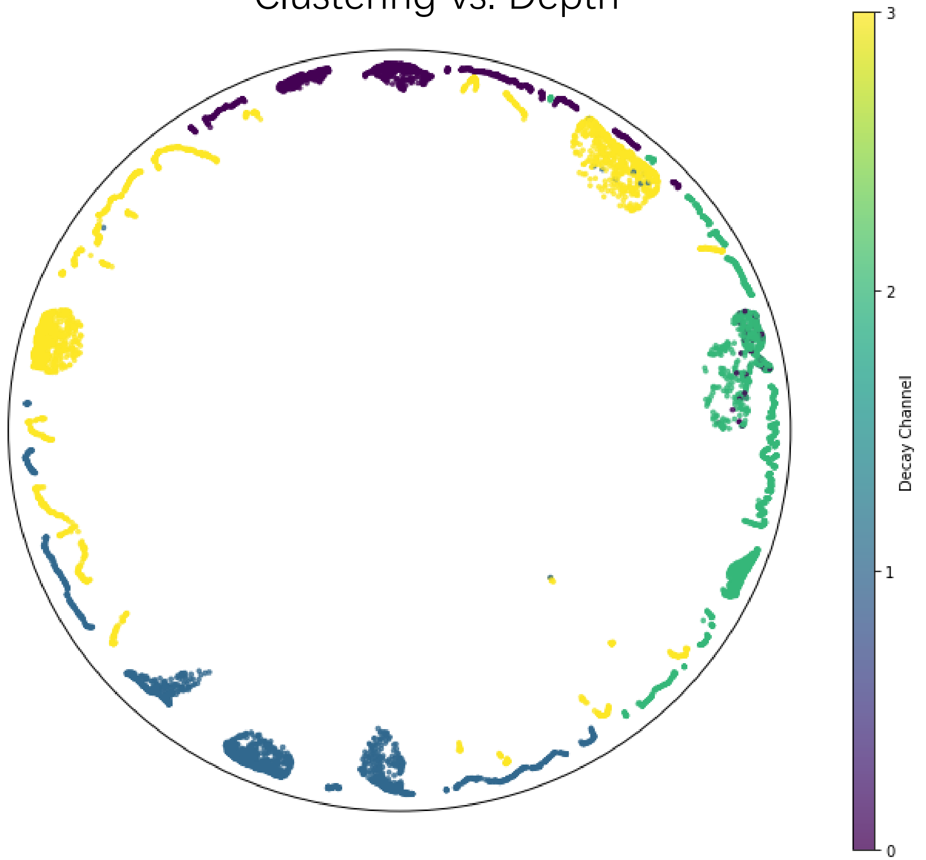
Structure 1 (full info)



Structure 2



Clustering vs. Depth



UMAP*: A tool for dimensionality reduction

Knowledge Transfer:

Teacher with well trained NN
on full information
available for MC

Student with smaller NN to be trained
on reduced information
available for reconstructed data