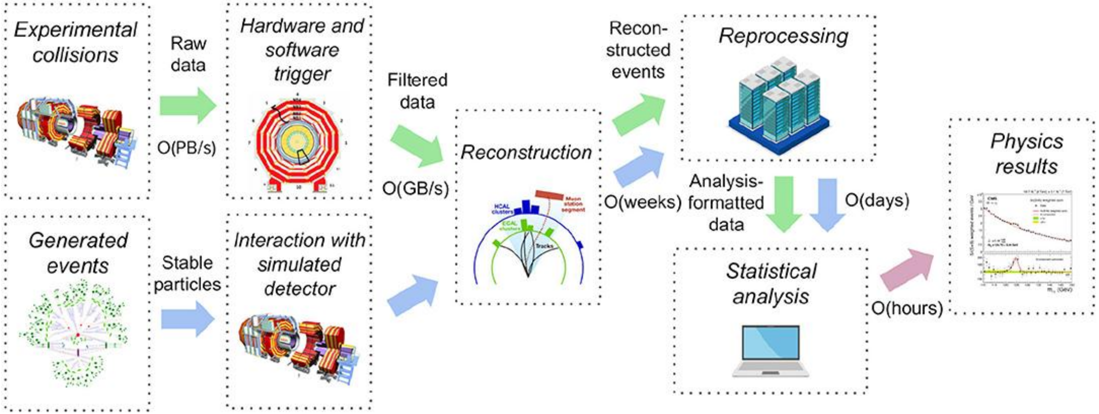
Developing foundational transformer models for unified top-quark physics analysis

Abasov E.E., Volkov P.V., Dudko L.V., Zaborenko A.D., Iudin E.S., Markina A.A., Perfilov M.A., Vorotnikov G. A.

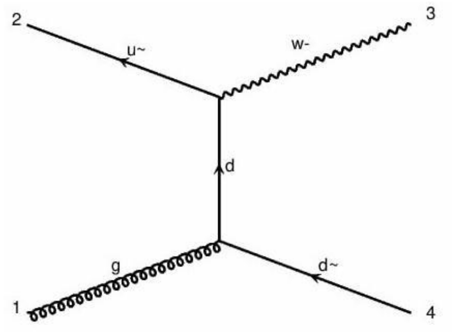
Motivation

- Modern experiments in High Energy Physics generate enormous amounts of data
- Classical ML models require complex manual customisation for each task

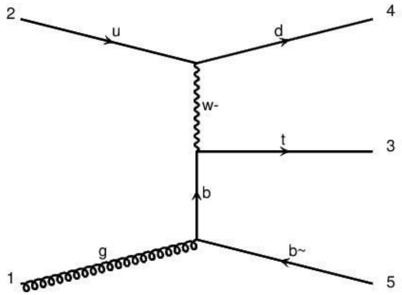
- CV and NLP have successfully applied foundation models that can generalise knowledge from large datasets
- We want to apply a similar approach to analysing collider events



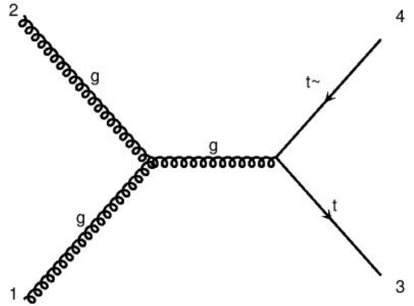
Datasets



0 top-quarks

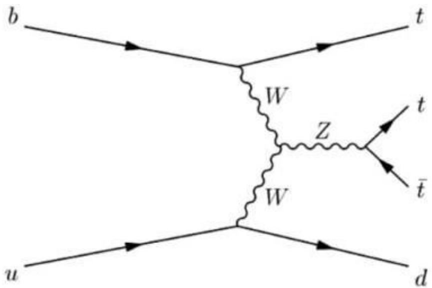


1 top-quarks

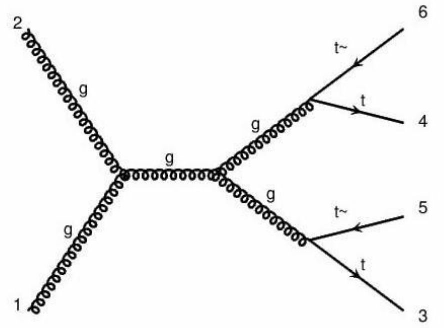


2 top-quarks

- Five groups of processes differing in the number of top quarks produced (0, 1, 2, 3, 4-t)
- Generators: MadGraph5 and CompHEP, detector response – DELPHES
- Total number of events ~8M



3 top-quarks



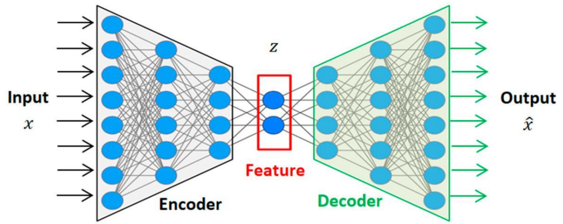
4 top-quarks

Universal input variable set

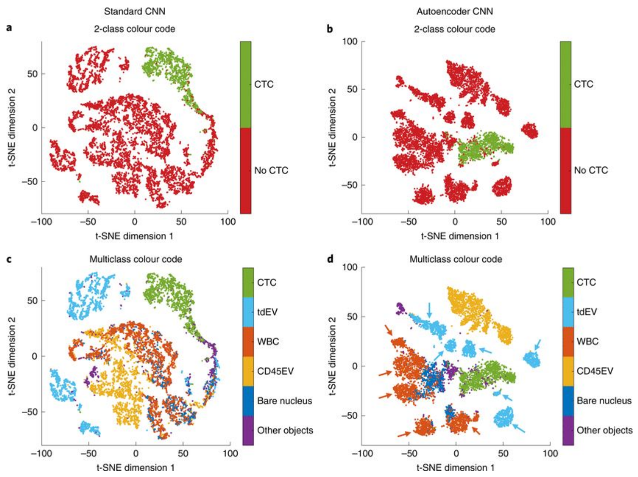
- **Analysis context:**
  - Study of collider events (jets and leptons) in high energy physics
- **Basic set of variables:**
  - For each object: p_T, η, φ, P_x, P_y, P_z, E and particle type
  - For the whole event: MET
- **Structure of objects in the event:**
  - Up to 12 jets (of which 4 b-jets) and 4 leptons
  - The objects within each group are ordered by p_T
- **Handling of missing objects:**
  - When an object is missing, the corresponding variables are filled with zeros
  - The null values are analysed and handled separately (based on particle type prediction)
- **Purpose of the set of variables:**
  - Universal basis for event analysis and identification algorithms

Shaping the space of representations

- **Objective:** create a common representation space in which patterns are expressed more clearly
- **Typical method:** using an intermediate layer of a neural network trained on a federated dataset
- **Possible realisations:**
  - Autoencoder (dimensionality reduction)
  - Masked modelling
- **The advantages of such a space:**
  - Advantages of such a space: explicit patterns in the data
  - Clearer segmentation into relevant clusters



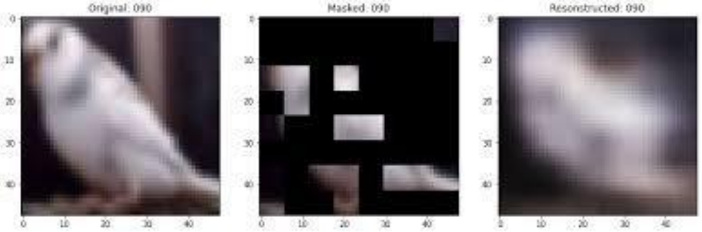
Schematic diagram of the autoencoder model



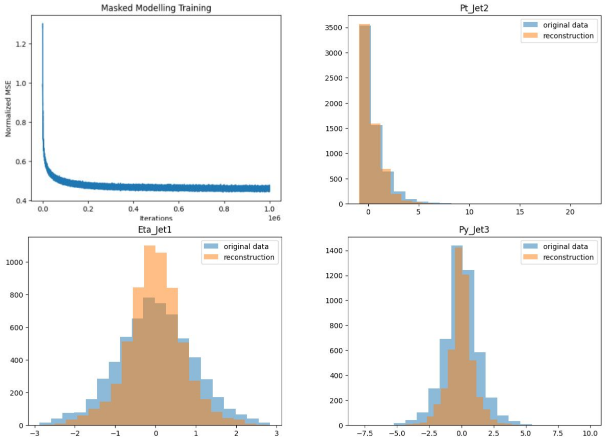
View space visualization for conventional network and autoencoder (image domain)

Shaping the space of representations

- **Reconstruction of masked variables:**
  - The mask is applied to all parameters at once
  - The model learns to recover missing (masked) values by considering the whole data structure
- **Comparison with autoencoder:**
  - The accuracy can be higher because the interrelationships are taken into account for all variables
  - Autoencoder is usually optimised for average reconstruction quality
- **Potential areas of pre-training:**
  - Utilisation of generator information
  - Neutrino component regression
  - Solution of combinatorial problems



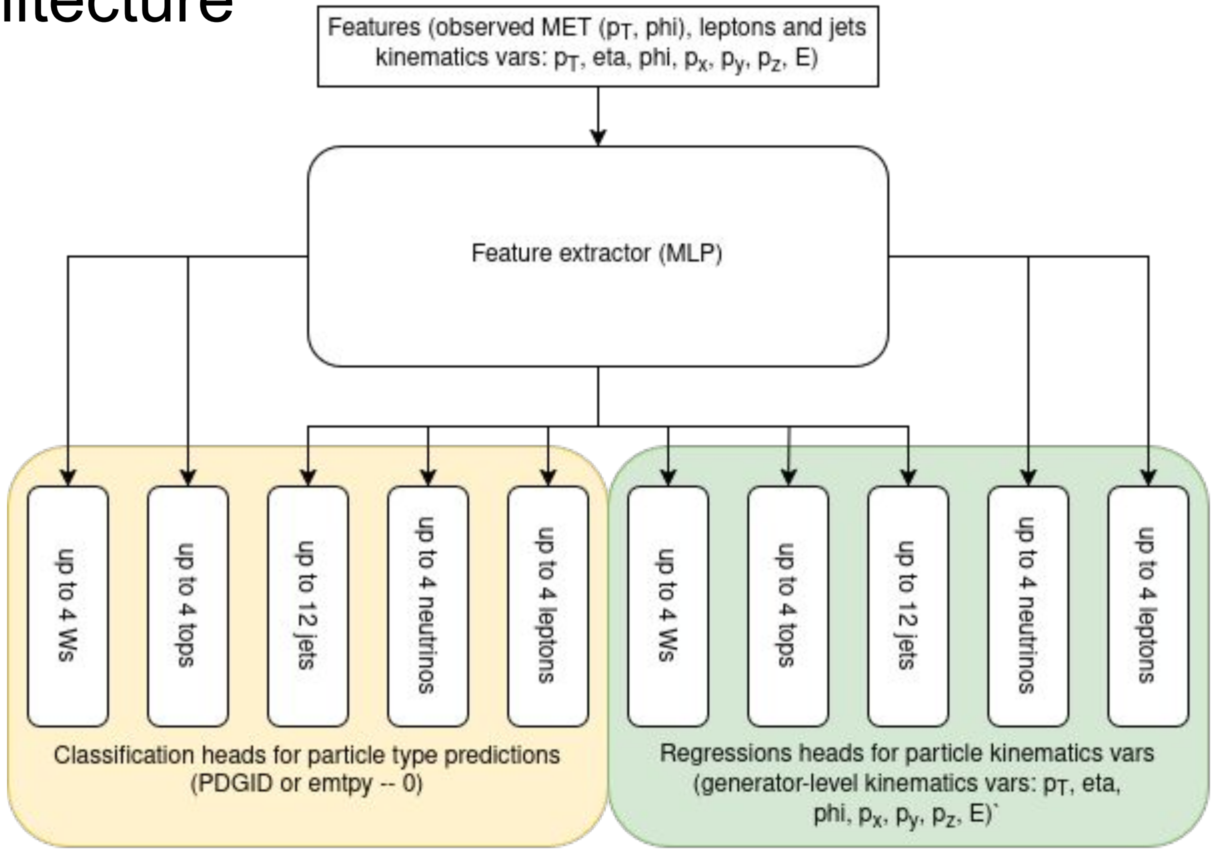
Example of reconstruction for images



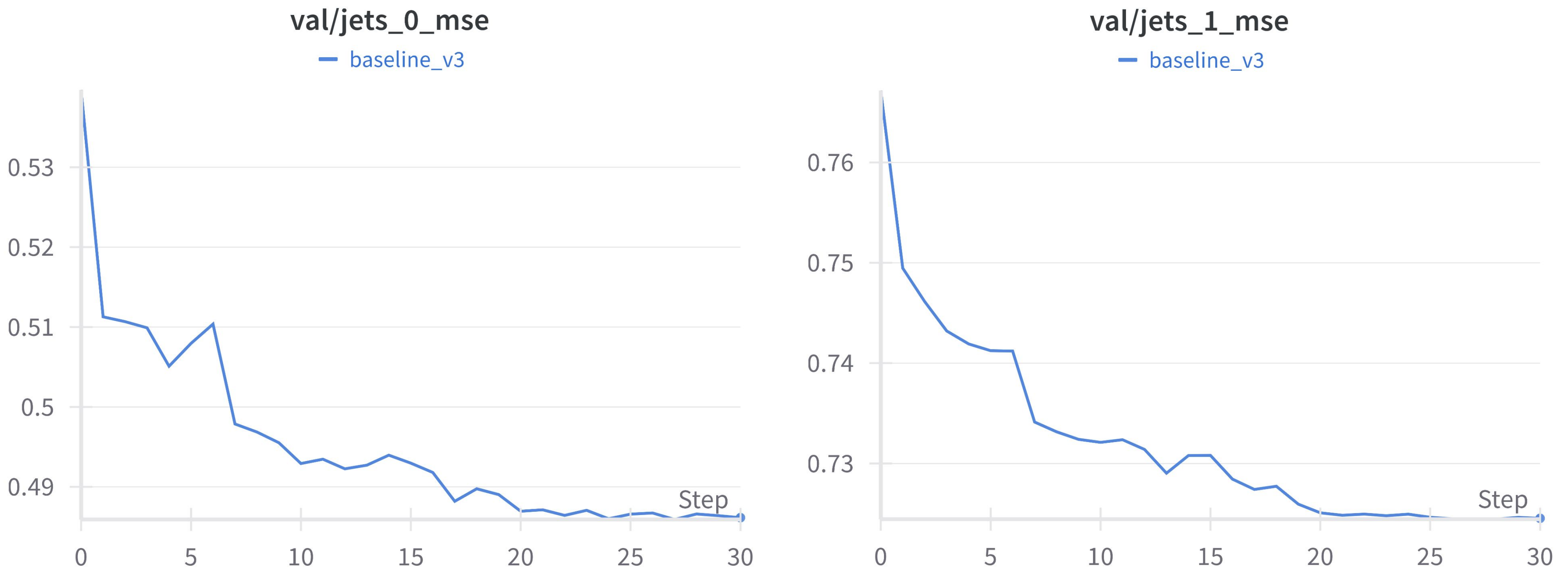
Mean reconstruction error and reconstructed variables. Probability of variable masking is 30%, variables are standardised

Level 1 Net: Architecture

- criterion: custom loss (mse + cross-entropy loss)
- optimizer: Adam with learning rate 1e-3 and weight decay 1e-5
- scheduler: ReduceLROnPlateau
- batch size: 1024

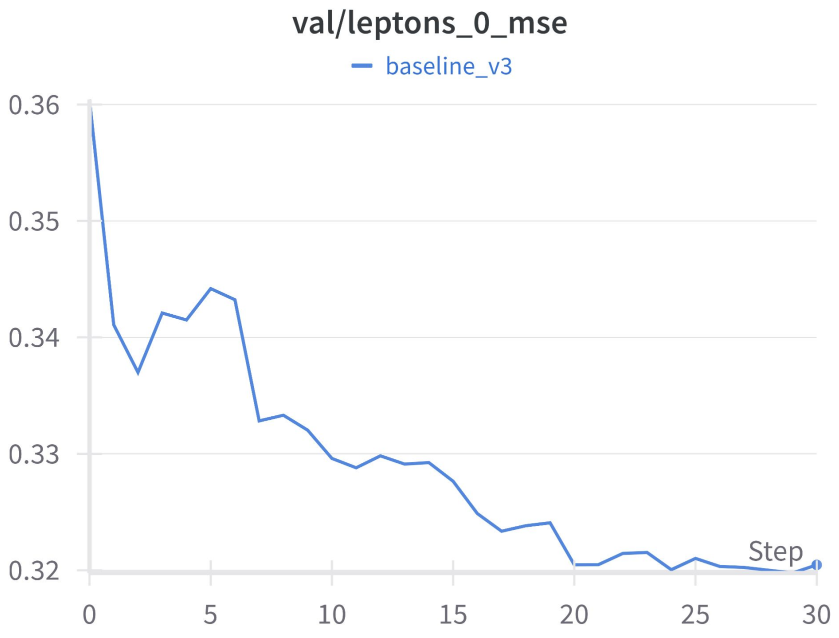


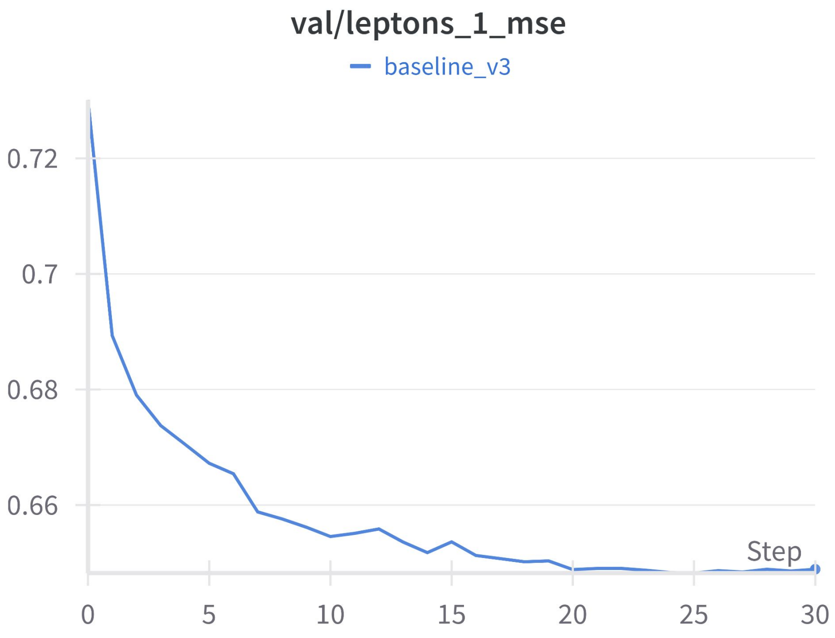
Level 1 Net: Metrics



Mean error of reconstruction of kinematic variables for the first two most energetic jets

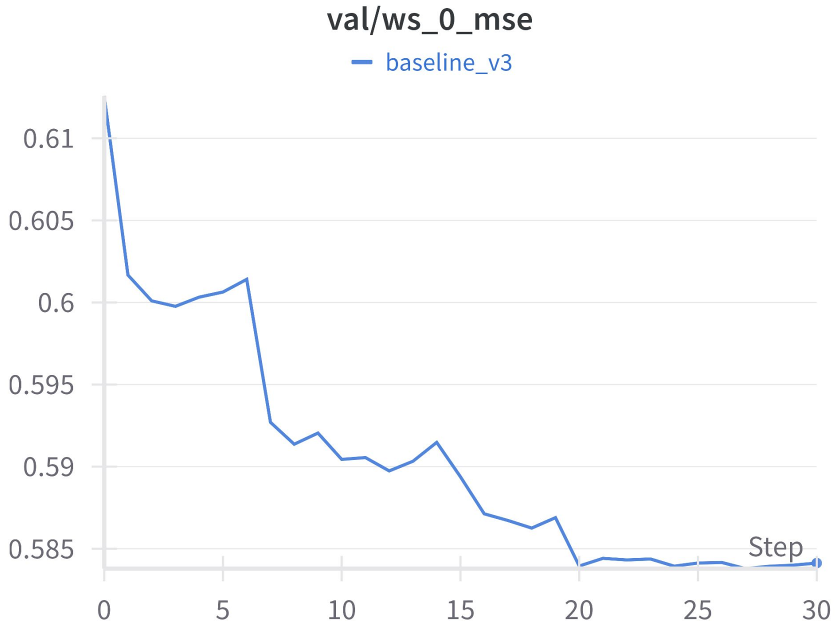
Level 1 Net: Metrics

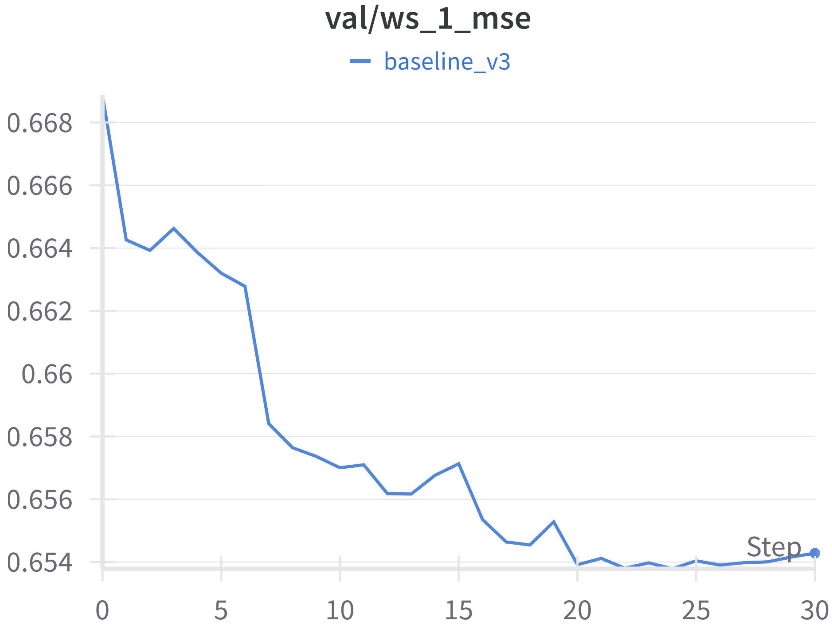




Mean error of reconstruction of kinematic variables for the first two most energetic leptons

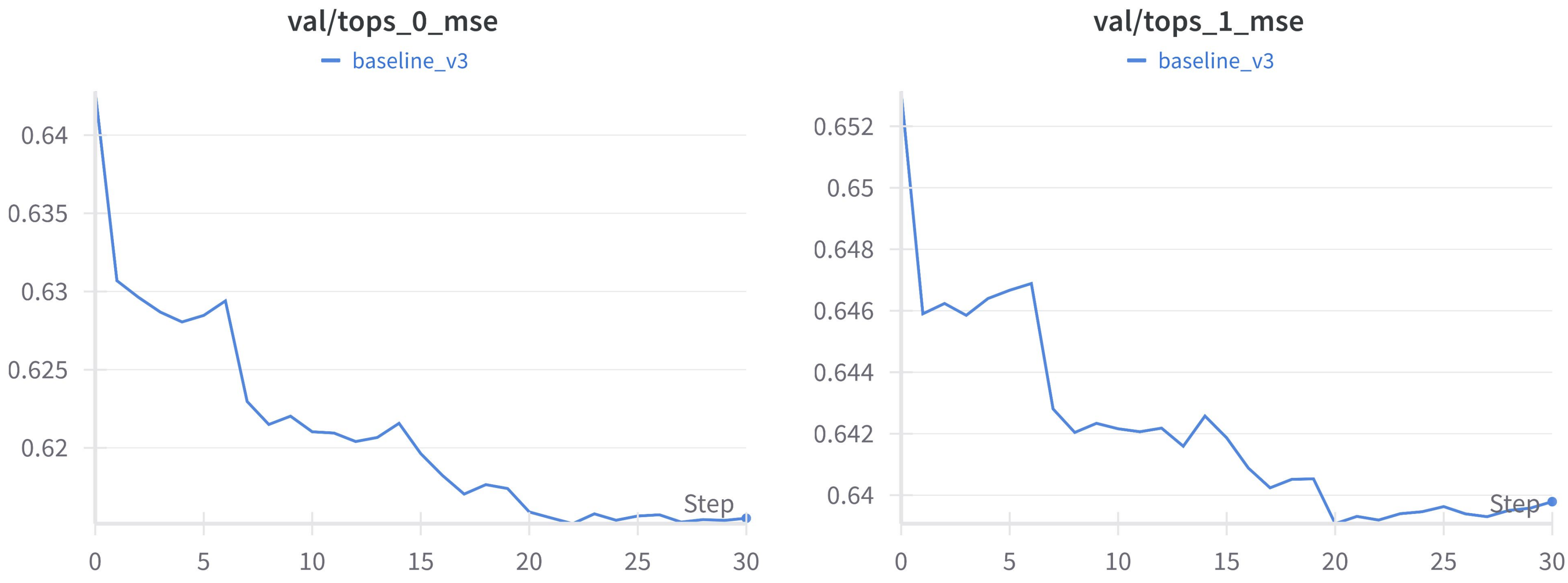
Level 1 Net: Metrics





Mean error of reconstruction of kinematic variables for the first two most energetic W-bosons

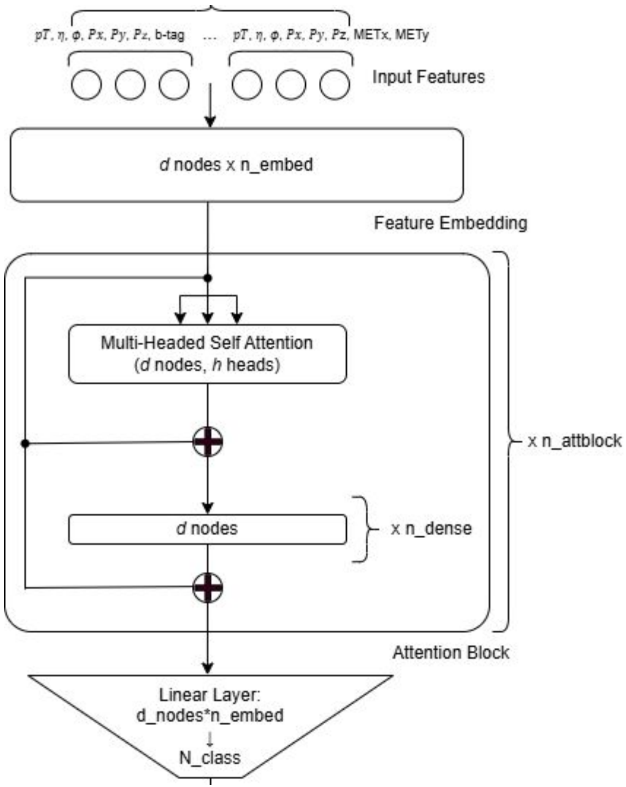
Level 1 Net: Metrics



Mean error of reconstruction of kinematic variables for the first two most energetic top-quarks

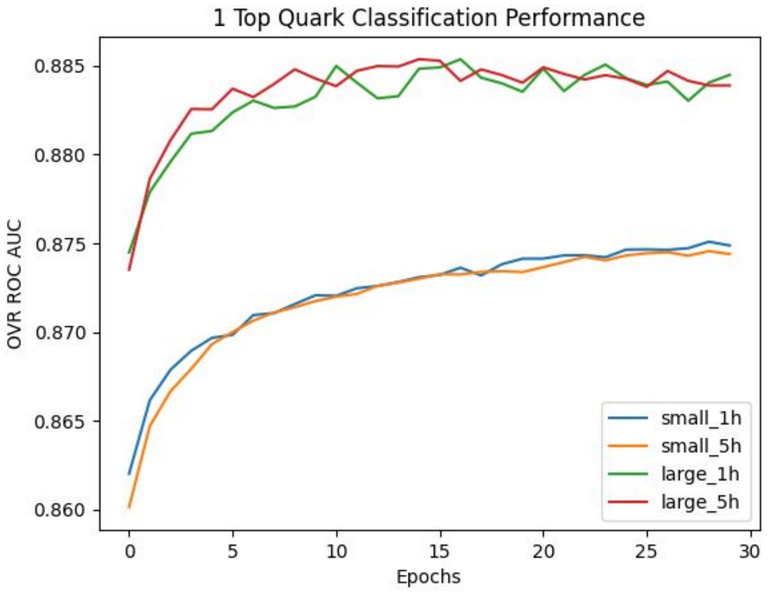
Level 2 Net: Architecture

- **Choice of architecture:**
  - *Transformer can be perfectly parallelized on large datasets*
  - *The attention module captures complex dependencies between input objects*
- **Application to a classification task:**
  - The model is trained to separate events with different numbers of top quarks
- **Advantages of the approach:**
  - High scalability for large data volumes
  - Easy handling of combinatorial relationships and non-linear dependencies



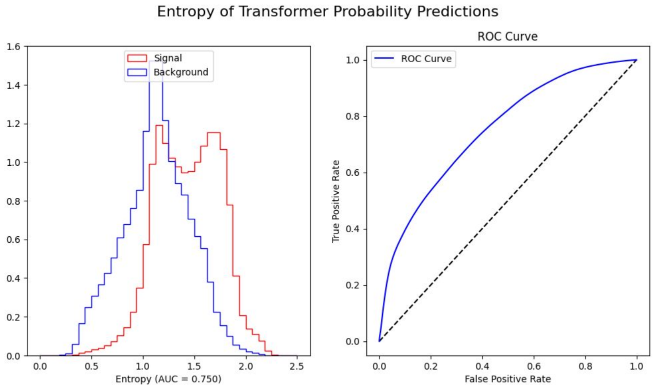
Level 2 Net: Metrics

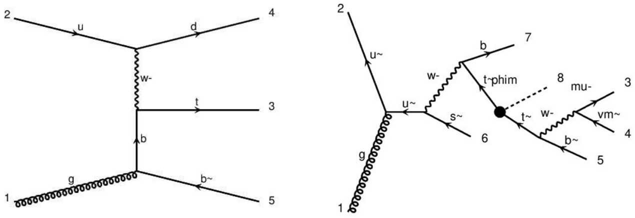
- Variants ‘small’ vs ‘large’ (embedding size, number of layers)
- Options 1h vs 5h (number of attention heads)
- **Results:**
    - Increasing model size improves metrics, but is harder to train and requires more resources
    - The number of attention heads has a mixed effect: in some cases overtraining



Application to the search for a ‘New Physics’

- **Objective:**
  - identify differences in kinematic and probability distributions
- **Analysing the predictions:**
  - Entropy: $H(X) = -\sum_{x \in X} p(x) \log_2(p(x))$
  - Used to measure the ‘uncertainty’ of a model prediction
- **Event Identification:**
  - Comparison of entropy distribution in events generated with and without the DM mediator
  - The model is not further trained; entropy is calculated directly from the current probability estimates
- **Data:**
  - t-channel top-quark birth in the Standard Model.
  - t-channel process in the Simplified DM Model with introduction of the scalar dark matter mediator.





Feynman diagrams for background and signalling processes

Conclusion

The talk demonstrated approaches to building versatile neural networks capable of unified analyses of high-energy physics (HEP) data at different signatures. Key steps included:

- Formation of a set of variables and a representation in an optimised abstract space
- Pre-training of the model to form this representation space
- Training of the Transformer model on the classification problem covering all the main processes of top-quark birth with different final signatures

In the future, it is planned to successively improve the intermediate steps and increase the complexity of the problems solved by the basic model in order to expand the capabilities of the method for top-quark research at colliders.

Thank You
For Your Attention

Backup

Additional classification metrics