PROCEEDINGS OF SCIENCE

Tokenization Strategies for Masked-Token Prediction for Anomaly Detection at the Large Hadron Collider

Ambre Visive,[a,b,*] **Roberto Ruiz de Austri,**[c] **Polina Moskvitina,**[d,e] **Clara Nellist**[a,b] **and Sascha Caron**[d,e]

[a] *Institute of Physics, University of Amsterdam, Science Park 904, 1090 GL Amsterdam, The Netherlands*
[b] *Nikhef, Dutch National Institute for Subatomic Physics, Science Park 105, 1098 XG Amsterdam, The Netherlands*
[c] *Instituto de Física Corpuscular, IFIC-UV/CSIC, Catedrático José Beltrán 2, E-46980 Paterna, Spain*
[d] *High Energy Physics, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands*
E-mail: ambre.visive@cern.ch, rruiz@ific.uv.es, polina.moskvitina@cern.ch, clara.nellist@cern.nl, scaron@nikhef.nl

Advances in Machine Learning, particularly Large Language Models (LLMs), enable more efficient interaction with complex datasets through tokenization and next- or masked-token prediction, providing a novel framework for analysing high-energy physics datasets. We explore strategies for representing particle physics data as token sequences, enabling LLM-inspired models to learn event distributions and detect anomalies in proton-proton collisions at the Large Hadron Collider (LHC). By training solely on background events, the model reconstructs expected physics processes, learning properties of the given Standard Model (SM) processes. Deviations in reconstruction scores during inference flag anomalous events, providing a data-driven approach to identify rare signatures or physics beyond the Standard Model (BSM). The method is tested using simulated LHC Run 2 ($\sqrt{s}$ = 13 TeV) proton-proton collision data from the Dark Machines Collaboration, replicating ATLAS conditions, focusing on SM and BSM four-top-quark final states. These tokenization strategies enable anomaly detection and suggest a path toward foundation models for the LHC and beyond, integrating state-of-the-art ML with physics principles to advance adaptive, data-driven searches for new physics.

EPS-HEP 2025, 6–11 July 2025, Palais du Pharo, Marseille, France

*Speaker

ISSN 1824-8039. Published by SISSA Medialab.

PoS(EPS-HEP2025)657

1. Introduction

The increasing data volume measured in experiments on the Large Hadron Collider (LHC) [1] presents new opportunities for applying advanced machine learning models originally developed for other domains. Large Language Models (LLMs), designed for natural language tasks [2], have demonstrated exceptional capabilities in modelling complex, high-dimensional data distributions [3]. Their success stems from transformer architectures [4] and large-scale training on diverse datasets [5]. In this work, we investigate LLM-inspired networks for unsupervised anomaly detection in high-energy physics, following the masked-token prediction strategy as developed for BERT [6]. The model learns the underlying physics of the background processes through training on predictions of masked particles from the context provided by the remaining particles in each event. Deviations in reconstruction scores during inference serve as indicators of potential signal events, enabling detection without prior signal knowledge. This work builds upon our previous study [7], which introduced this method. Here, we extend that approach by focusing on the impact of different strategies for representing collision events through tokenization and, inspired by the Omnijet-α approach [8], we also explore a VQ-VAE-based representation. This framework aims to enhance sensitivity to rare Standard Model (SM) processes, such as the four-top-quark simultaneous production process and to uncover new potential Beyond the Standard Model (BSM) signatures.

2. Dataset

2.1 Physics Motivation

To evaluate unsupervised anomaly detection methods on SM processes, we select benchmarks characterized by rare and complex event topology. The simultaneous production of four top quarks, first observed by the ATLAS experiment in 2023 [9] provides an ideal test due to these requirements. Given the dominant decay channel of the top quark ($t \to W + b$, as $|V_{tb}|^2 \sim 1$), these events decay into 0 to 4 leptons and 4 to 12 jets, including four hadronized bottom quarks. Differing in the number of b-jets in the final state, $t\bar{t}WW$ and $t\bar{t}W$ can mimic the $t\bar{t}t\bar{t}$ signature in the detector. Equivalently, $t\bar{t}Z$, and $t\bar{t}H$ exhibit comparable topologies to $t\bar{t}t\bar{t}$, as Z and Higgs bosons can decay leptonically and hadronically, and the Higgs boson can decay into Z- and W-pairs. Due to these overlaps, $t\bar{t}WW$, $t\bar{t}W$, $t\bar{t}Z$ and $t\bar{t}H$ are treated as **background events** in this work, while $t\bar{t}t\bar{t}$ constitutes the **signal process** to be isolated.

For the BSM benchmark, we consider the simultaneous production of gluino pairs in a Supersymmetry (SUSY) scenario, where each gluino decays into a pair of top quark. This decay channel introduces a unique challenge: the four-top process itself becomes part of the background for the SUSY signal. Consequently, the dominant **SM background events** for gluino pair production include: $t\bar{t}WW$, $t\bar{t}W$, $t\bar{t}Z$, $t\bar{t}H$ and $t\bar{t}t\bar{t}$ events; while $\tilde{g}\tilde{g}$ is treated as the **BSM signal events** in this analysis.

2.2 The Dark Machines Datasets and Data Format

This study employs datasets from the Dark Machines challenge [10], which include more than 10^9 simulated pp collisions at $\sqrt{s} = 13$ TeV. Event generation is performed using `MG5_aMC@NLO` 2.7, followed by hadronization with `Pythia` 8.239, and fast detector simulation via `Delphes` 3.4.2 with

modified ATLAS-Run 2 settings. The resulting Monte Carlo samples, originally stored in ROOT format, are converted to CSV files after applying the event selection criteria described in [10] and [11]. In a subsequent pre-processing step, the event-types are extracted and mapped to integer labels, as summarized in Table 1, producing labelled datasets. Each event is represented as an ordered sequence of particle-object tags (see Table 2) together with their four-momentum components (energy E, transverse momentum p_T, pseudo-rapidity η, azimuthal angle ϕ), sorted by object type and p_T. Additionally, the missing transverse energy ($\|E_T^{miss}\|$) and its azimuthal angle ($\phi_{E_T^{miss}}$) are included for each event. Bottom-quark jets are tagged without uncertainty. All sequences are padded to a fixed length of 18 particles. The dataset is partitioned into training (80%), validation (10%), and test (10%) subsets. Further details can be found in [10–12].

SM process	$t\bar{t}t\bar{t}$	$t\bar{t}H$	$t\bar{t}W$	$t\bar{t}WW$	$t\bar{t}Z$
`process ID`	1	2	3	4	5

Table 1: Labels of SM processes.

Object	jet	b-tagged jet	positron	electron	muon	anti-muon	photon
symbol ID	`j`	`b`	`e+`	`e-`	`mu+`	`mu-`	`g`
`tag`	1	2	3	4	5	6	7

Table 2: Tags of particle-objects.

The input to the LLM-inspired model consists of batches of token sequences, each representing a particle physics event. Tokens represent the missing transverse energy and its azimuthal angle, or individual particles, characterized by their type, charge, transverse momentum, pseudo-rapidity, and azimuthal angle. Constructing these sequences from the Dark Machines dataset requires a dedicated tokenization procedure, which will be detailed in Sections 3.2 and 3.3.

3. Model Architecture and Tokenization Techniques

3.1 Large-Language-Model for Anomaly Detection

As a unsupervised anomaly detector, we adopt a lightweight architecture inspired by LLMs, leveraging the encoder component of a transformer network and the masked-token prediction strategy introduced in BERT [6]. Each input (a particle physics event) is represented as a tokenized sequence and embedded into a high dimensional space. The core network of the model comprises two transformer layers, each equipped with four self-attention heads, allowing the model to capture contextual relationships across tokens. A positional encoding layer can be added to incorporate information about the token ordering within each sequence. A subsequent linear projection layer and a softmax activation produce a probability distribution over the token classes, which represent the likelihood of each token being the correct reconstruction.

To enable the model to learn the distribution of background events, training is performed exclusively on background events, using a masked-token prediction strategy. For each event, a single token is randomly masked, and the network is optimized to reconstruct it based on the remaining context. The loss function employed is Sparse Categorical Cross-Entropy, and optimization is

carried out using the `Adam` algorithm [13], with early stopping applied to prevent overfitting. During inference, all tokens in an event are sequentially masked and reconstructed, generating individual reconstruction scores. These scores are averaged to obtain an event-level reconstruction score. Events with low reconstruction are flagged as anomalous, signalling deviation from the learned background distribution. The resulting score distributions (see Figure 1) enable the definition of thresholds for identifying candidate signal events.

A tokenization step is needed to represent particle physics events as sequences of `tokens` suitable for LLM or LLM-like models.

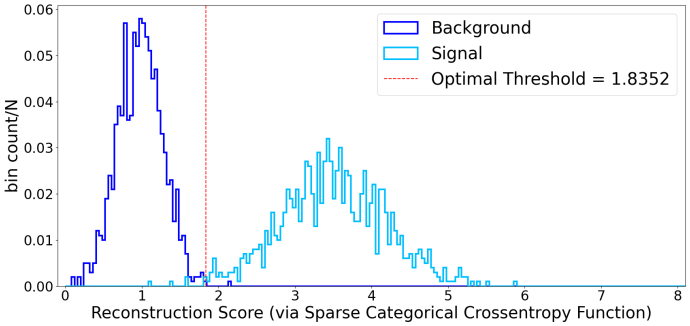


Figure 1: Illustrative distribution of the aggregated reconstruction scores in a perfectly trained model, evaluated with sparse categorical cross-entropy, for background (blue) and signal (cyan) events. The red dashed line indicates the optimal threshold used to separate the two classes.

3.2 Fixed-Bin Tokenization Method

To provide a structured representation of particle features, we explored a discretization-based tokenization strategy referred to as binning in this work. Particles are mapped to seven predefined object categories (see Table 2) and their kinematic features (p_T, η, and ϕ) are discretized into same-size bins. Careful selection of bin edges is essential to preserve a physically meaningful separation of particle properties. The performance of each binning configuration, as well as the effect of including missing transverse energy ($\|E_T^{miss}\|$) and its azimuthal angle ($\phi_{E_T^{miss}}$) as additional tokens in the sequence representation, was initially assessed using a compact classifier network for rapid evaluation. Configurations that demonstrated satisfactory results were subsequently tested with the downstream LLM-like model. In the current most effective binning configuration:

- p_T, η and $\|E_T^{miss}\|$ were each divided into 4 bins, with edges defined such as each bin contains 25% of the background data; consequently, the bins are not of equal size.
- ϕ and $\phi_{E_T^{miss}}$ were divided into 4 bins of width $\frac{1}{4}\pi$.
- With bins indexed from 1 onward, each particle token is defined as followed:
  $token_{part} = 64 \times (bin_{obj} - 1) + 16 \times (bin_{p_T} - 1) + 4 \times (bin_{\eta} - 1) + bin_{\phi}$.
- It yields: $\cdot token_{part} \in [1, 448]$; $\cdot token_{\|E_T^{miss}\|} \in [449, 452]$; $\cdot token_{\phi_{E_T^{miss}}} \in [453, 456]$.
- An event is represented as a sequence:
  $[token_{part,1}, token_{part,2}, token_{part,3}, ..., token_{part,18}, token_{\|E_T^{miss}\|}, token_{\phi_{E_T^{miss}}}]$

To maintain uniform sequence lengths required by transformer architectures and because the original dataset is zero-padded, padded entries corresponding to absent particles were assigned a special token value ($token_{part,i} = 0$).

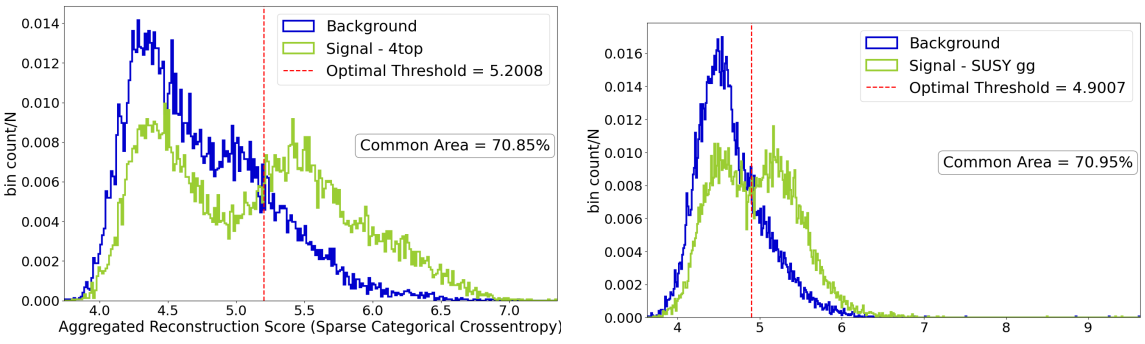
Results obtained with this binning-based tokenization will be systematically compared to those achieved using a deep-learned representation approach (obtained using a Vector-Quantized Variational-Auto Encoder (VQ-VAE), discussed in Section 3.3). This comparison aims to evaluate whether deep-learned tokenization can outperform discretization in capturing the structure of particle physics events.

3.3 VQ-VAE for Learned Tokenization

Building on the approach employed by Birk, Hallin, and Kasieczka [8] for jet generation and tagging, we employ a Variational Autoencoder with a Vector Quantization layer (VQ-VAE) [14] to tokenize particle physics event data. The VQ-VAE architecture consists of three main components: an encoder, a vector quantization layer, and a decoder. The encoder (a NormFormer model [15]) maps the input event sequence into a continuous latent representation, which is then discretized by the quantization layer through assignment to the nearest entry in a learned codebook. The resulting codebook indices form a discrete representation of the event and serve as tokens for downstream processing by the LLM-like model. Training is performed on a subset of background events, with the objective of reconstructing the original event from its quantized latent representation. After training, the VQ-VAE weights are frozen, and the full dataset is passed through the encoder network and quantization layer to generate tokenized representations. The size of the codebook is a critical hyperparameter, as it determines the size of the vocabulary available to the LLM-like model in later stages. Larger codebooks offer finer granularity but increase complexity, while smaller codebooks promote compact representations. Based on downstream performance, the current most effective configuration in this study uses a codebook of size 850.

4. Results and Evaluation

4.1 Search for Four-Top Production



(a) Results using binned tokenization

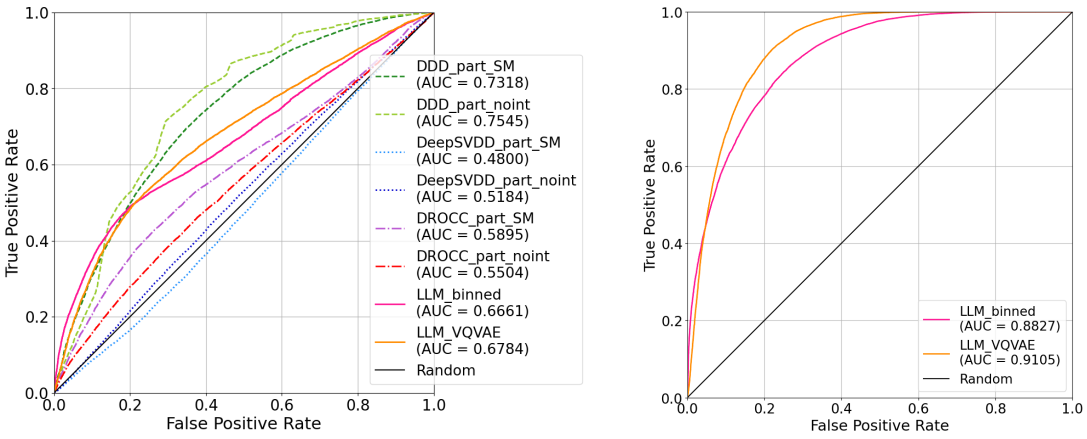
(b) Results using VQ-VAE-based tokenization

Figure 2: Distribution of the aggregated reconstruction scores, evaluated with sparse categorical cross-entropy, for background (blue) and four-top-signal (green) events. The red dashed line indicates the optimal threshold used to separate the two classes.

To assess the method's performance on the simultaneous four-top-quark production process, both signal and background events are tokenized using the strategies outlined in Sections 3.2 and 3.3. After training the LLM-like model exclusively on background data, it is evaluated on mixed samples containing both signal and background. For VQ-VAE-based tokens, the model achieves better results with sinusoidal positional encoding, whereas for binned tokens, performance improves without positional encoding. Following the approach described in Section 3.1, the distribution of the average reconstruction score of the signal and background events is obtained, as shown in Figure 2a for binned tokens and Figure 2b for VQ-VAE-based token. The overlap region between these distributions (70.85% with a binning tokenization and 70.50% with a VQ-VAE-base tokenization) quantifies the model's discriminative power. An optimal threshold for class separation can be derived from these distributions and used as reference for future analysis to identify signal events. From these score distributions, the Receiving Operator Characteristic (ROC) curve of each model can be derived (see Figure 3a) and the area under the curve (ROC-AUC) of the model can be calculated. The model yields ROC-AUC = 0.6661 when a binning tokenization is used and ROC-AUC = 0.6781 with VQ-VAE-based tokens. Preliminary results indicate superior performance when using the VQ-VAE tokenization strategy.

4.2 Comparison with Established Unsupervised Methods

The results for the four-top-quark search are compared against established unsupervised anomaly detection methods implemented in [11], namely DDD, DeepSVDD, and DROCC, as all approaches are evaluated on the same dataset. As shown in Figure 3a, while the proposed approach does not outperform DDD-based techniques, for any tested tokenization strategy, it yields better performance than DeepSVDD and DROCC, demonstrating competitive behaviour among unsupervised anomaly detection methods.



(a) Evaluated by isolating $t\bar{t}t\bar{t}$

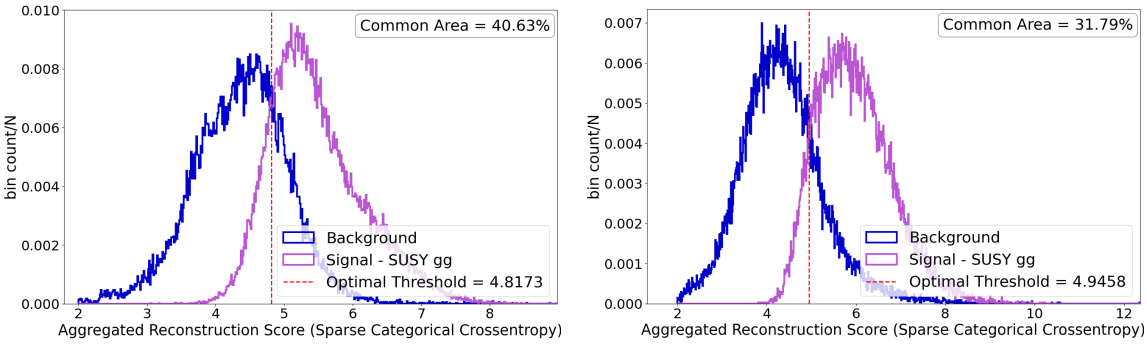
(b) Evaluated by isolating $\tilde{g}\tilde{g}$

Figure 3: Comparison of the ROC curves of different anomaly detection and/or tokenization methods

4.3 Search for BSM signals

To evaluate the method's performance on the pair-production of SUSY gluinos, signal and background events are tokenized using the strategies described in Sections 3.2 and 3.3. Separate

models are trained for this benchmark since the background differs from the four-top case. In a similar way to the four-top benchmark, the model performs best with sinusoidal positional encoding when using VQ-VAE-based tokens, whereas for binned tokens, performance improves without positional encoding. Likewise, the distribution of the average reconstruction score for BSM signal and SM background events is obtained, as shown in Figure 4a for binned tokens and Figure 4b for VQ-VAE-based token, and the overlap region between these distributions (40.63% with a binning tokenization and 31.79% with a VQ-VAE-base tokenization) quantifies the model's discriminative power. Similarly, ROC curves are constructed from the score distributions (see Figure 3b), the model achieves ROC-AUC = 0.8827 when a binning tokenization is used and ROC-AUC = 0.9105 with VQ-VAE-based tokens. Preliminary results indicate superior performance when using the VQ-VAE tokenization strategy in this BSM benchmark.



(a) Results using binned tokenization

(b) Results using VQ-VAE-based tokenization

Figure 4: Distribution of the aggregated reconstruction scores, evaluated with sparse categorical cross-entropy, for background (blue) and SUSY gluinos signal events (purple). The red dashed line indicates the optimal threshold used to separate the two classes.

5. Conclusion

A novel application of LLM-inspired models for unsupervised anomaly detection in high-energy physics has been introduced. Although the findings are preliminary, promising performance has been observed in the identification of rare Standard Model (SM) processes such as four-top-quark production and unobserved Beyond the Standard Model (BSM) processes like the pair-production of SUSY gluino. While superiority over existing approaches has not been achieved yet, competitive results have been demonstrated through the use of a flexible token-based representation of the data. In addition, tokenization strategies have been investigated, with initial evidence suggesting that a deep-learned tokenization scheme with the use of a VQ-VAE model to tokenize the data provides better performance. Further optimization of tokenization schemes and model architectures is expected to enhance sensitivity and robustness, establishing this approach as a strong candidate for model-independent searches in future high-energy physics analyses. These results highlight the potential of token-based representations of collider data, combined with predictive modelling, as a powerful framework for discovering new physical phenomena. Once trained on SM background data, the model can be reused across different BSM searches, limiting computational resources and enabling scalable, model-independent anomaly detection.

References

[1] The ATLAS and CMS Collaborations, *Highlights of the HL-LHC physics projections by ATLAS and CMS*, *ATL-PHYS-PUB-2025-018* (2025)

[2] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, *Improving language understanding by generative pre-training*, *OpenAi Blog* (2018)

[3] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain and J. Gao, *Large Language Models: A Survey*, *arXiv: 2402.06196* (2025)

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, *Attention Is All You Need*, *CoRR* **abs/1706.03762** (2023)

[5] E. M. Bender, A. McMillan-Major, T. Gebru and S. Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021) 610-623

[6] J. Devlin, M-W. Chang and K. Lee and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, *arXiv: 1810.04805* (2019)

[7] A. Visive, P. Moskvitina, C. Nellist, R. Ruiz de Austri and S. Caron, *Event Tokenization and Next-Token Prediction for Anomaly Detection at the Large Hadron Collider*, *arXiv: 2509.26218* (2025)

[8] J. Birk, A. Hallin and G. Kasieczka, *OmniJet-α: the first cross-task foundation model for particle physics*, *Machine Learning: Science and Technology* 5 (2024) 035031

[9] The ATLAS Collaborations, *Observation of four-top-quark production in the multilepton final state with the ATLAS detector*, *Eur. Phys. J. C* **83** (2023) 496

[10] T. Aarrestad, M. van Beekveld, M. Bona, A. Boveia, S. Caron, Sascha, J. Davies, A. de Simone, C. Doglioni, J. Duarte, A. Farbin, H. Gupta, L. Hendriks, L. A. Heinrich, J. Howarth, P. Jawahar, A. Jueid, J. Lastow, A. Leinweber, J. Mamuzic, E. Merényi, A. Morandini, P. Moskvitina, C. Nellist, J. Ngadiuba, B. Ostdiek, M. Pierini, B. Ravina, R. Ruiz de Austri, S. Sekmen, M. Touranakou, M. Vaškeviciute, R. Vilalta, J-R. Vlimant, R. Verheyen, M. White, E. Wulff, E. Wallin, K. A. Wozniak and Z. Zhang, *The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider*, *SciPost Physics* **12** (2022) 043

[11] S. Caron, J-E. García Navarro, M. Moreno Llácer, P. Moskvitina, M. Rovers, A. Rubio Jímenez, R. Ruiz de Austri and Z. Zhang, *Universal anomaly detection at the LHC: transforming optimal classifiers and the DDD method*, *Eur. Phys. J. C* **85** (2025) 415

[12] L. Builtjes, S. Caron, P. Moskvitina, C. Nellist, R. Ruiz de Austri, R. Verheyen and Z. Zhang, *Attention to the strengths of physical interactions: Transformer and graph-based event classification for particle physics experiments*, *SciPost Physics* **19** (2025) 028

[13] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, *arXiv: 1412.6980* (2017)

[14] A. van den Oord, Oriol Vinyals and K. Kavukcuoglu, *Neural Discrete Representation Learning*, *arXiv: 1711.00937* (2018)

[15] S. Shleifer and J. Weston and M. Ott *NormFormer: Improved Transformer Pretraining with Extra Normalization*, *arXiv: 2110.09456* (2021)