DINAMO: Dynamic and INterpretable Anomaly MOnitoring for Large-Scale Particle Physics Experiments

Arsenii Gavrikov[1,*], Julián García Pardiñas[2], Alberto Garfagnini[1]

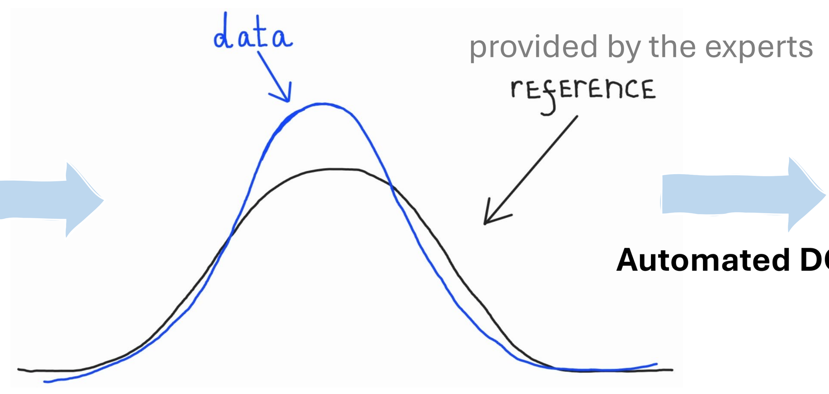
[1]University & INFN Padova, Italy; [2]Massachusetts Institute of Technology (MIT), USA

[*]agavriko@cern.ch

2nd EuCAIFCon, Cagliari, Italy

I. Data quality monitoring

- In particle physics experiments, **DQM** is a gatekeeper **against corrupted, anomalous data**
- Traditional DQM is manual: **shifters compare data with references** provided by experts:
  - High cost of person power
  - Limited accuracy
  - Any detector or software updates → **challenges to adapt to changes in operational conditions**



Shifters are trained to judge if it is a good data or a bad one

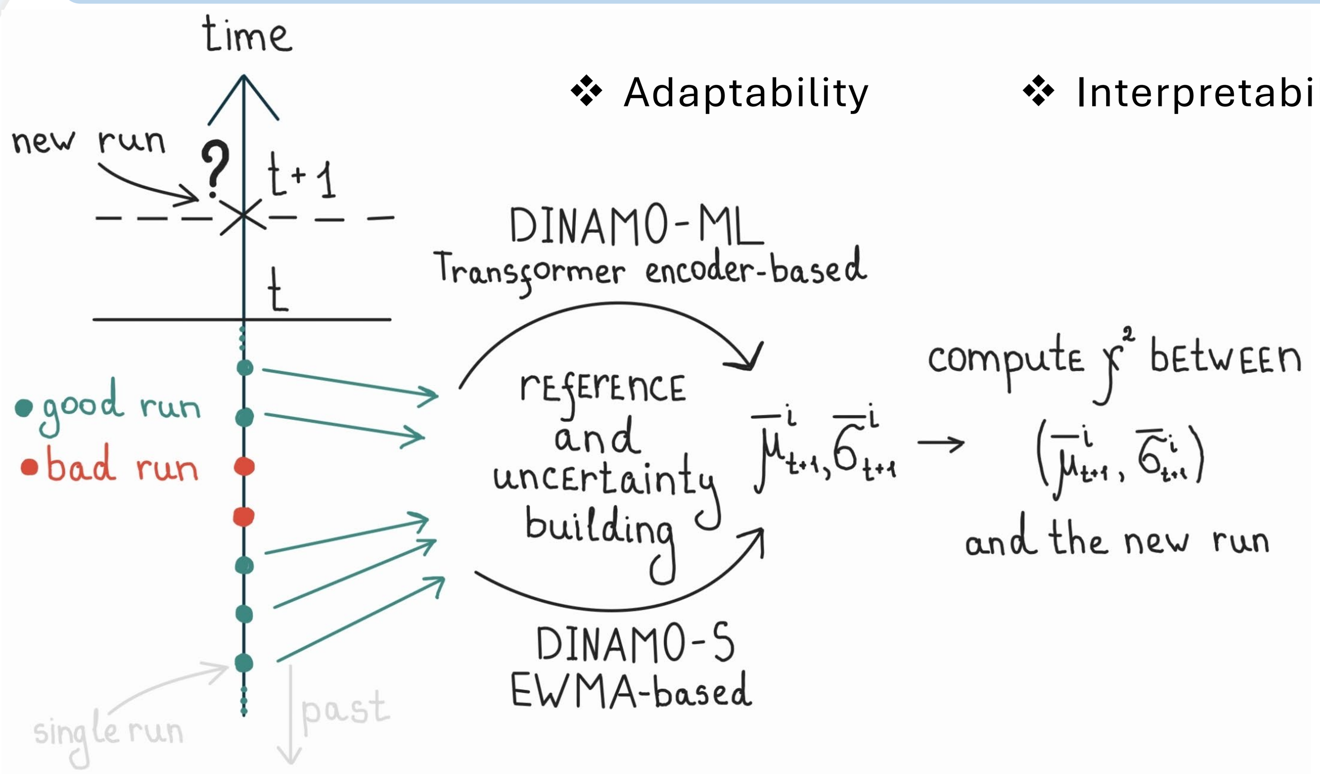
Automated DQM?

- **An automated DQM** to be:
  - **Accurate** → as high as possible
  - **Specific** → where is the problem
  - **Interpretable** → "why" the algorithm decided so
  - **Dynamic** → adaptability to changing conditions
  - **Fast** → although analyzing vast amount of data

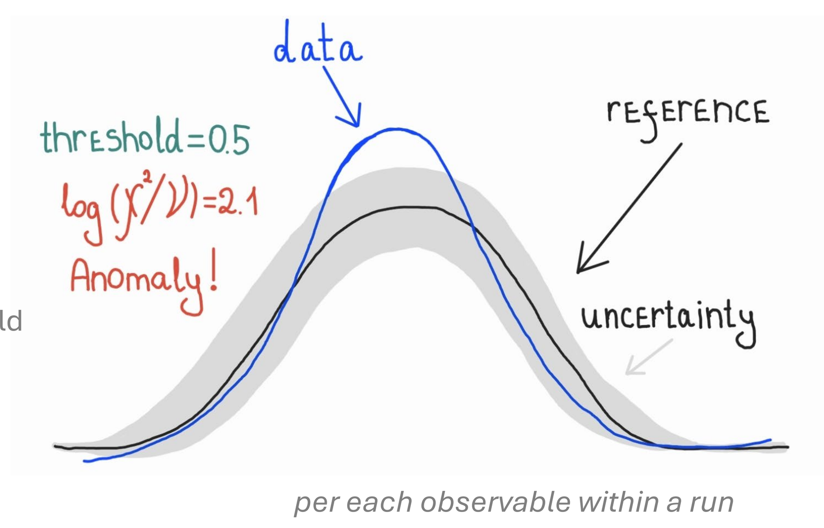
Automated DQM will help to reduce shifters burden and the amount of inconsistencies

II. DINAMO: Dynamic and Interpretable Anomaly Monitoring [1]

- Adaptability
- Interpretability
- Relative simplicity



+ set a threshold



per each observable within a run

DINAMO-S

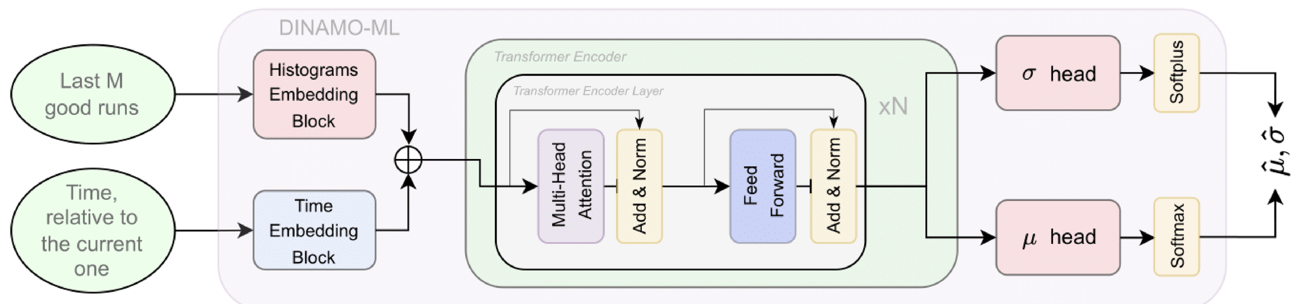
Build the reference **dynamically to account to the changing conditions** based on **Exponentially Weighted Moving Average (EWMA)** method:

1. Include also **the weights of each run** according to their statistical noise: $\omega_t = \frac{1}{\sigma^2_{X_t}}$
2. Iterative update of **sum of weights**: $W_{t+1} = \alpha W + (1-\alpha)\omega_t$
3. Iterative update of **weighted sum**: $S_{t+1} = \alpha S_t + (1-\alpha)\omega_t X_t$
4. Iterative update of **sum of weighted squared residuals** to model **the uncertainty**: $S_{\sigma,t+1} = \alpha S_{\sigma,t} + (1-\alpha)\omega_t (X_t - \mu_t)^2$
5. Compute **the reference mean and uncertainty:**
$$\mu_{t+1} = \frac{S_{t+1}}{W_{t+1}} \qquad \sigma_{\mu_{t+1}} = \sqrt{\frac{S_{\sigma,t+1}}{W_{t+1}}}$$
6. Compute **test statistics** to compare the two:
$$\chi^2_\nu = \frac{1}{N_b}\sum_{j=1}^{N_b}\frac{\left(\tilde{x}_{i,j} - \hat{\mu}_{i,j}\right)^2}{\sigma^2_{\tilde{x}_{i,j},p} + \hat{\sigma}^2_{\mu_{i,j}}}$$

Parameter α is a **hyperparameter to control EWMA**

DINAMO-ML

Substitutes the EMWA-based part with the transformer encoder:

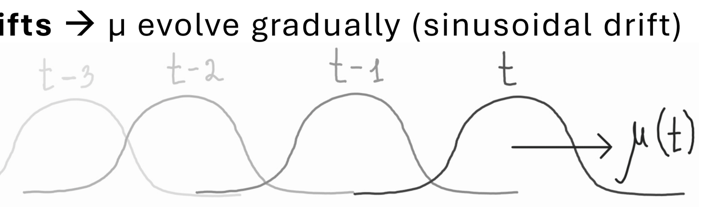


- Training via *Online learning*:
  - Takes **mini-batch of *K*** good last runs
  - **Per each run** in the mini-batch we build **a context of *M*** preceding good runs
  - Learning to predict the bin-by-bin means and widths for future runs
- Anomaly detection as follows:
  - **Context creation:** for a new run, we identify up to *M* most recent good runs as input to the transformer
  - **Predict reference:** output **μ** and **σ** by the model
  - **Anomaly score:** compute the test statistic the same as with DINAMO-S
- Gaussian negative log-likelihood as loss: $NLL = \frac{1}{2N_bK}\sum_{k=1}^{K}\sum_{j=1}^{N_b}\left[\left(\frac{\tilde{x}_{k,j} - \hat{\mu}_{k,j}}{\hat{\sigma}_{k,j}}\right)^2 + \log\left(\hat{\sigma}^2_{k,j}\right)\right]$
- More complex → slower performance but **more accurate** and quicker in **adapting**

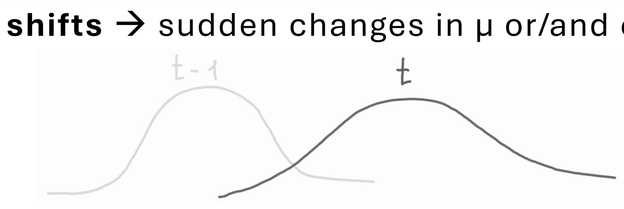
III. Synthetic data generator

Realistic synthetic data:
- Based on **1d Gaussian distributions** as histograms
- Main focus is the modeling of changes in conditions

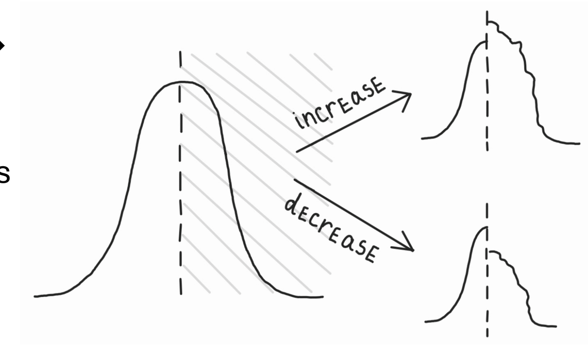
Implemented features:
- **Slow drifts** → μ evolve gradually (sinusoidal drift)



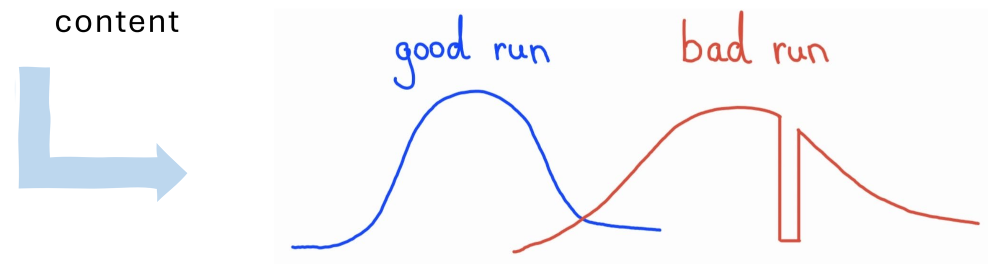
- **Abrupt shifts** → sudden changes in μ or/and σ



- **Varying events statistics** → initial number of events is sampled from a uniform distribution
- **Systematic uncertainty** → accidental increase or decrease of events in the right half of the histogram's window using binomial distribution



Anomalies:
- **extra distortion** in μ or/and σ
- **dead bins**: random number of bins (up to 20) get missed content



IV. Main metrics

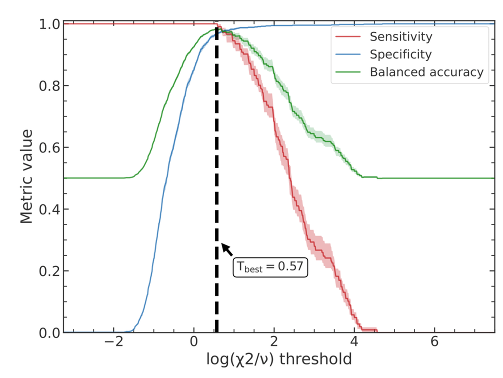
- **Balanced accuracy** → to balance the uneven class ratio
- **Adaptation time** → an average amount of good runs to be misclassified before the algorithm adapts

- **Uncertainty coverage** → how well the actual variability of the good runs is described using Jaccard distance between the true good runs and the predicted references in the z-score space

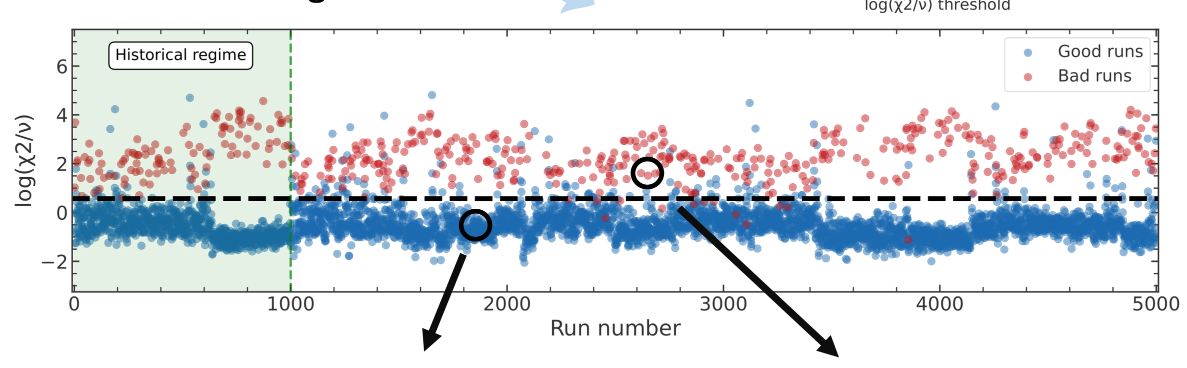
V. Results with synthetic data

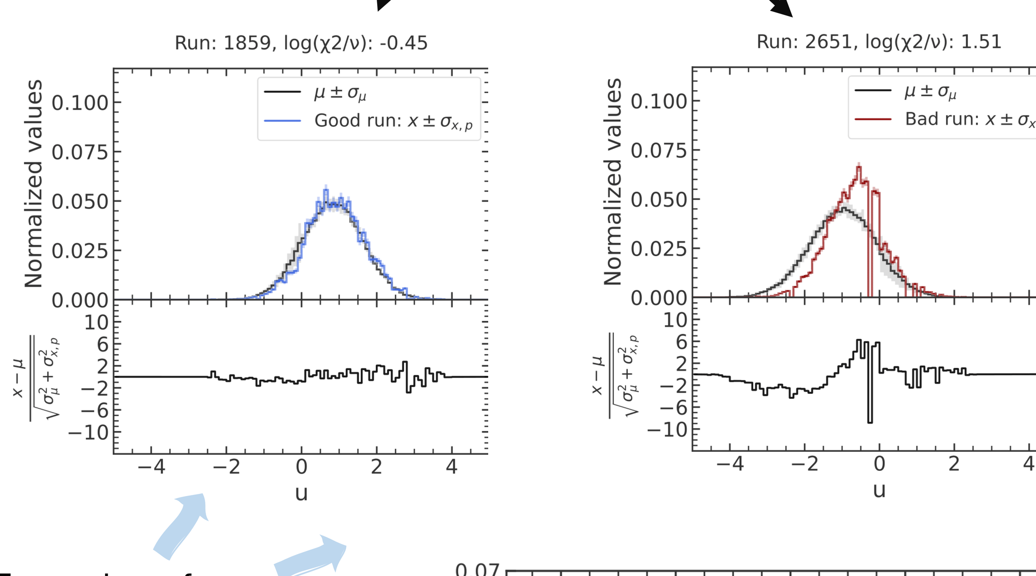
Results on a single dataset: *DINAMO-ML*

- **"Historical regime"** (first 20% of runs) → **to tune hyperparameters** and the threshold based on balanced accuracy

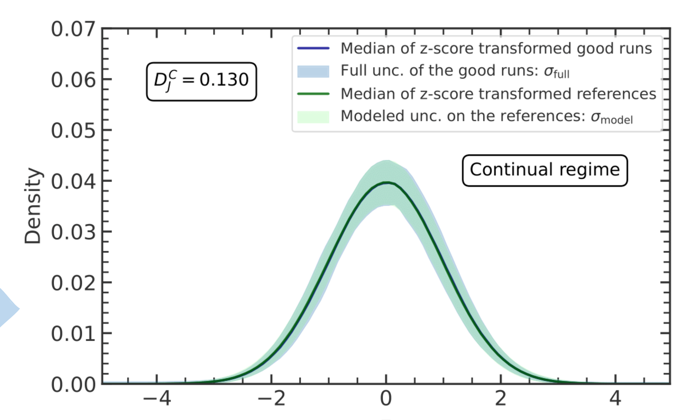


- Distribution of the anomaly scores **for a single dataset**



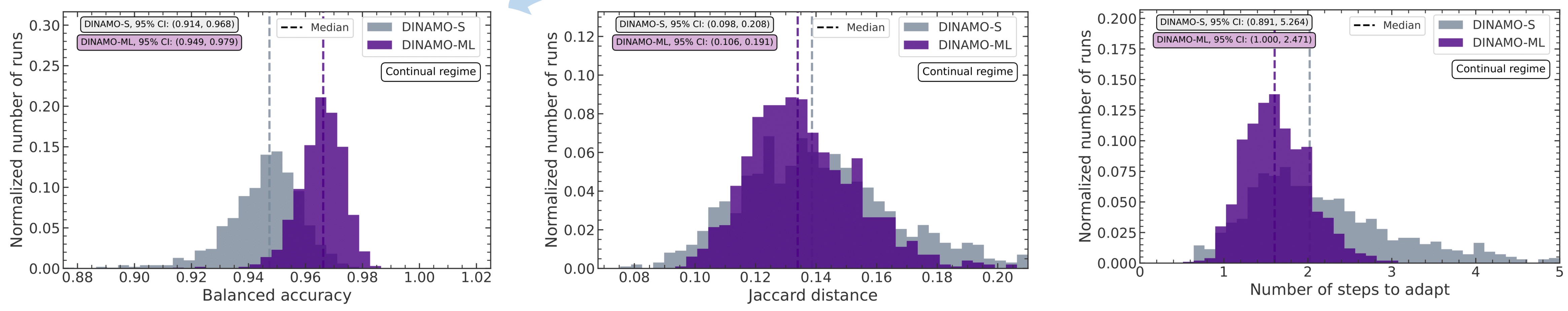


- Examples of **references and their uncertainties** for **correctly** classified runs
- Perfect description of the **variability** of good data



Aggregated results: *1000 synthetic datasets with different seeds*

- The comprehensive evaluation on synthetic data demonstrates that both **DINAMO algorithms** successfully **address the core challenges of automated DQM**
- **DINAMO-ML outperformed** DINAMO-S in all our metrics thanks to it more complex nature
- **DINAMO-ML** is particularly better **in adaptation speed** and **in balanced accuracy**



	Bal. acc. ↑	Specif. ↑	Sensit. ↑	Jaccard D. for σ ↓	Adapt. time ↓
DINAMO-S	$0.947^{+0.020}_{-0.033}$	$0.943^{+0.028}_{-0.058}$	$0.956^{+0.029}_{-0.075}$	$0.139^{+0.069}_{-0.041}$	$2.02^{+3.24}_{-1.13}$
DINAMO-ML	$\mathbf{0.966^{+0.012}_{-0.018}}$	$\mathbf{0.969^{+0.015}_{-0.037}}$	$\mathbf{0.966^{+0.024}_{-0.044}}$	$\mathbf{0.134^{+0.057}_{-0.028}}$	$\mathbf{1.61^{+0.87}_{-0.61}}$

VI. Conclusions

- We present **DINAMO** → a novel approach to **automate DQM** for large particle physics experiments:
  - EWMA-based, "standard" version: **DINAMO-S**
  - Transformer encoder-based, machine learning version: **DINAMO-ML**
- **DINAMO-ML outperforms the standard version** in all our metrics
- **Key advantages:**
  - **Adaptability** to changes in operational conditions
  - **Interpretability** through the references' dynamic creation (+ uncertainties)
  - **Relative simplicity** to enhance maintainability
- **DINAMO-S** is already **being commissioned** at the LHCb experiment for offline DQM

[1] A. Gavrikov, J. García Pardiñas, A. Garfagnini, "DINAMO: Dynamic and INterpretable Anomaly MOnitoring for Large-Scale Particle Physics Experiments", arXiv:2501.19237 (2025)

Arsenii Gavrikov is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No.101034319 and from the European Union – NextGenerationEU.