



On the Value of Tokeniser Pretraining in Physics Foundation Models

Hadi Sotoudeh, Payel Mukhopadhyay, Ruben Ohana, Michael McCabe
Neil D. Lawrence, Shirley Ho, Miles Cranmer

Polymathic

Read Paper

Motivation

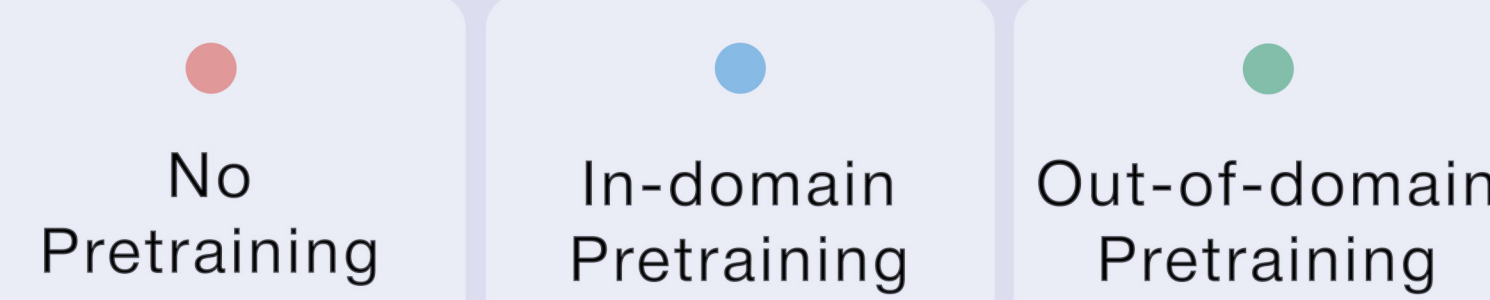
Physics foundation models learn to compress high-resolution data while modeling dynamics.

Unlike vision, tokenisers in physics are typically trained end-to-end rather than pretrained.

💡 Can pretraining the tokeniser make physics models faster or more accurate?

Setup

Compare emulation performance across three training settings:



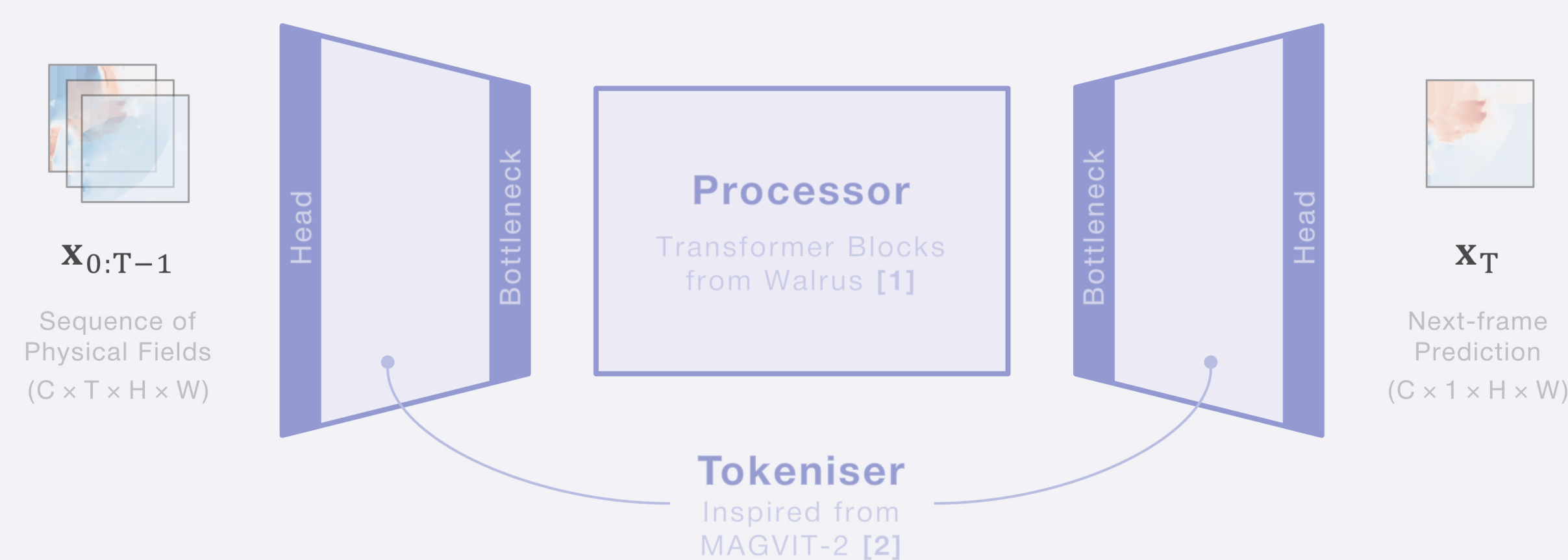
Key Findings

Pretraining reduces VRMSE, but gains depend strongly on domain alignment:

- In-domain: **64%** lower error
- Out-of-domain: **19%** lower error



Experimental Setup



Training Settings:

- No Pretraining:** Train emulator with randomly initialised weights
- In-domain:** Pretrain tokeniser on the same dataset as emulator
 - Keep tokeniser fully trainable
 - Only heads & bottlenecks trainable
- Out-of-domain:** Pretrain tokeniser on a different dataset than emulator
 - Keep tokeniser fully trainable
 - Only heads & bottlenecks trainable

Model Size:

105 M Total	100 M Processor	4.9 M Tokeniser Core	0.1 M Tokeniser Heads & Bottlenecks
-----------------------	---------------------------	-----------------------------------	--

Loss Functions & Metrics

Tokeniser Pretraining
(Autoencode x)

Training

Emulator Training
(Predict x_T given $x_{0:T-1}$)

$$\ell_{Tok} = \|x - \hat{x}\|_2^2$$

(Mean Squared Error)

$$\ell_{Emu} = \|x_T - \hat{x}_T\|_1$$

(Mean Absolute Error)

Evaluation

VRMSE
(Variance-normalised RMSE)

$$VRMSE = \frac{\|x - \hat{x}\|_2}{\sigma_x}$$

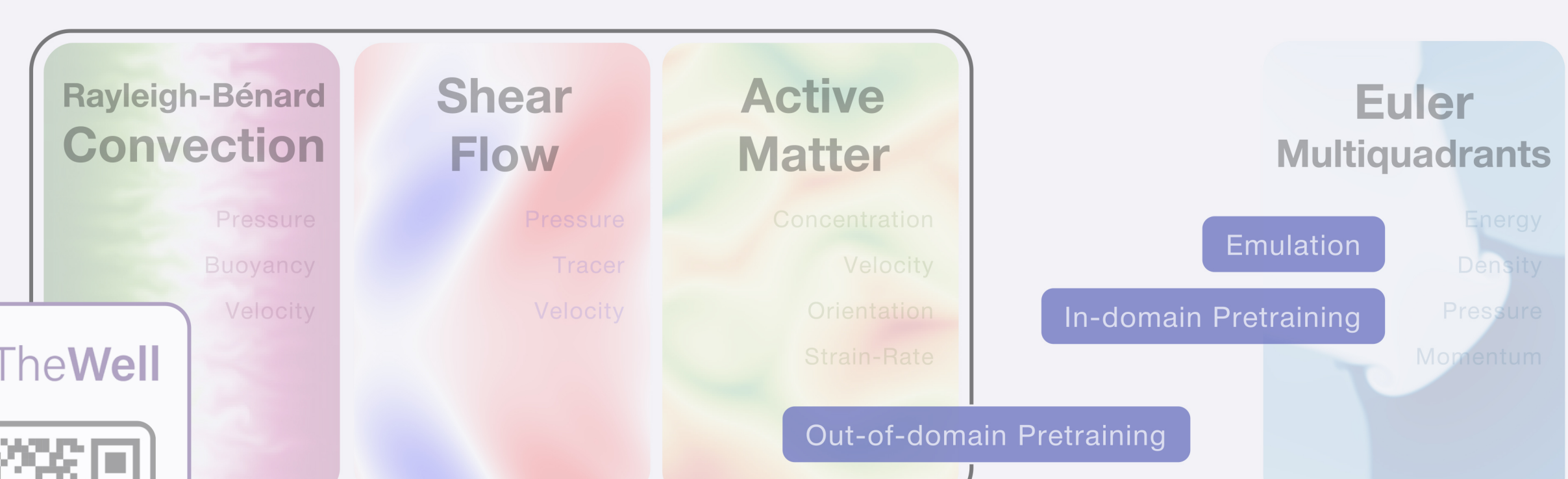
(Pixel space error normalised by variability in the target)

NEPS
(Normalised Error Power Spectrum)

$$NEPS(k) = \frac{P_{x-\hat{x}}(k)}{P_x(k)}$$

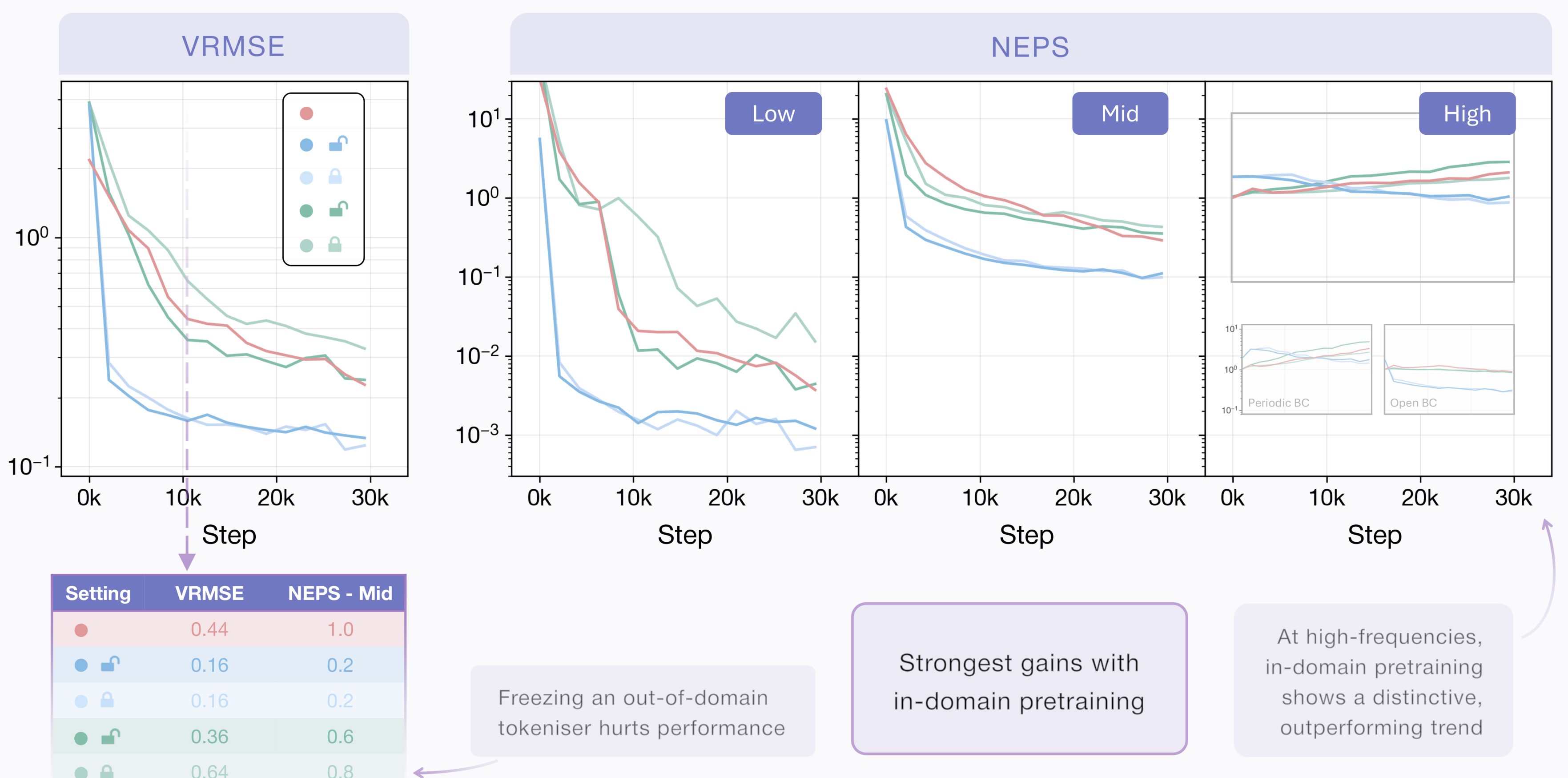
(Ratio of error power to signal power)

Data

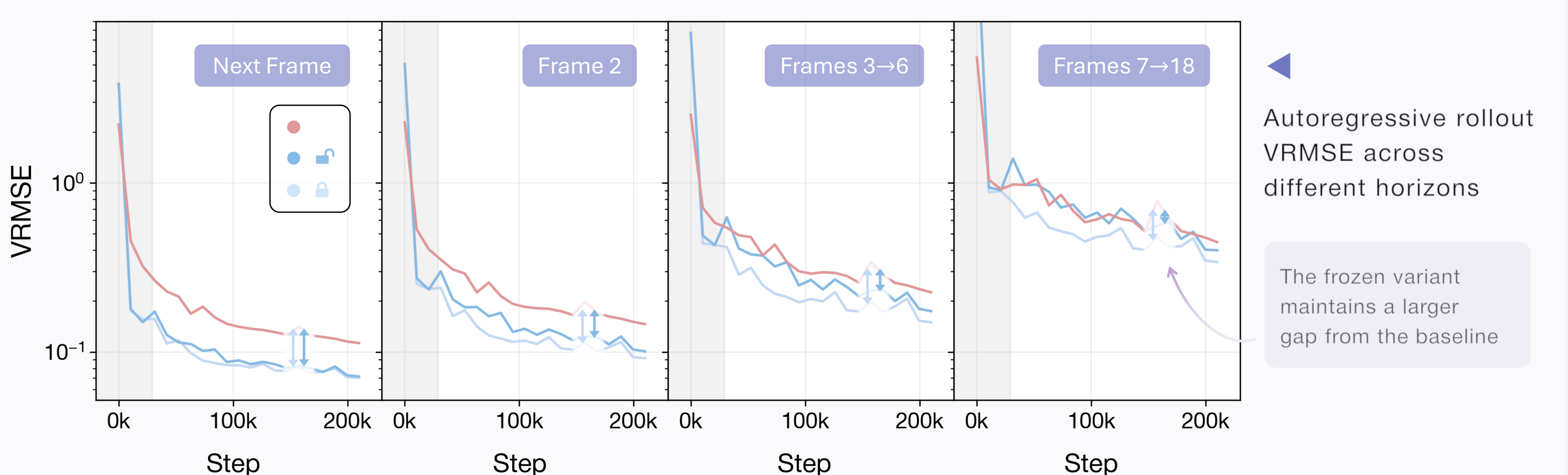


Results

Emulation Performance



Freezing Strategy as Regularisation



Takeaways

- Tokeniser pretraining improves early downstream learning, especially when pretraining data match the target system. Out-of-domain pretraining is only partially transferable.
- For in-domain pretraining, freezing the tokeniser acts as a regulariser, reducing VRMSE on longer rollout horizons. This comes with a 98% reduction in trainable parameters (5M $\rightarrow$ 85k).
- Understanding how pretraining data composition affects transfer will be key to building generalisable tokenisers for physics foundation models.

References

[1] M. McCabe et al. (2025). "Walrus: A Cross-Domain Foundation Model for Continuum Dynamics"

[2] L. Yu et al. (2023). "Language Model Beats Diffusion - Tokenizer is Key to Visual Generation"