# E-ProTran: Efficient Probabilistic Transformers for Forecasting

Batuhan Koyuncu*, Tim Nico Bauerschmidt*, Isabel Valera

UNIVERSITÄT DES SAARLANDES · DAAD · ZUSE SCHOOL ELIZA

Can we design probabilistic transformers with higher efficiency?
Yes, **E-ProTran** has improved inference speed and scalability with comparable performance!

## Motivation and Contributions

**Transformers** excel in handling long-range dependencies in sequential data and show promise in time series analysis. However, **their complexity often results in overparameterization, extended training times, and scalability challenges**, especially with complicated generative model assumptions. In this work, we propose Efficient Probabilistic Transformers (E-ProTran), a re-design of Probabilistic Transformers (ProTran) to mitigate these problems.

## Formulation and Problem Setting

We work with multivariate time series $x_{1:T} \in \mathbb{R}^{T\times d}$, where $x_t \in \mathbb{R}^d$ is a $d$-dimensional data vector at the discrete time step $t \in \mathbb{N}^+$. We parameterize a generative model with joint of $p_{\psi,\theta}(x_{1:T}, z_{1:T}|x_{1:t_0}, c_{1:T})$ can be used for forecasting in temporal settings. The baseline work ProTran uses

$$p_\psi(z_{1:T}|x_{1:t_0}, c_{1:T}) = \prod_{l=1}^{L}\prod_{t=1}^{T} p_\psi(z_t^{(l)}|z_{1:t-1}^{(l)}, z_{1:T}^{(l-1)}, x_{1:t_0}, c_{1:T}) \quad (1)$$

$$p_\theta(x_{1:T}|z_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t|z_t^{(L)}) \quad (2)$$

where $L$ is number of layers in the encoder, $T$ is length of the temporal data.

## E-ProTran

For our model, we have the generative model with components

$$p_\psi(z_{1:T}|x_{1:t_0}, c_{1:T}) = \prod_{t=1}^{T} p_\psi(z_t|x_{1:t_0}, c_{1:T}) \quad (3)$$

$$p_\theta(x_{1:T}|z_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t|z_t). \quad (4)$$

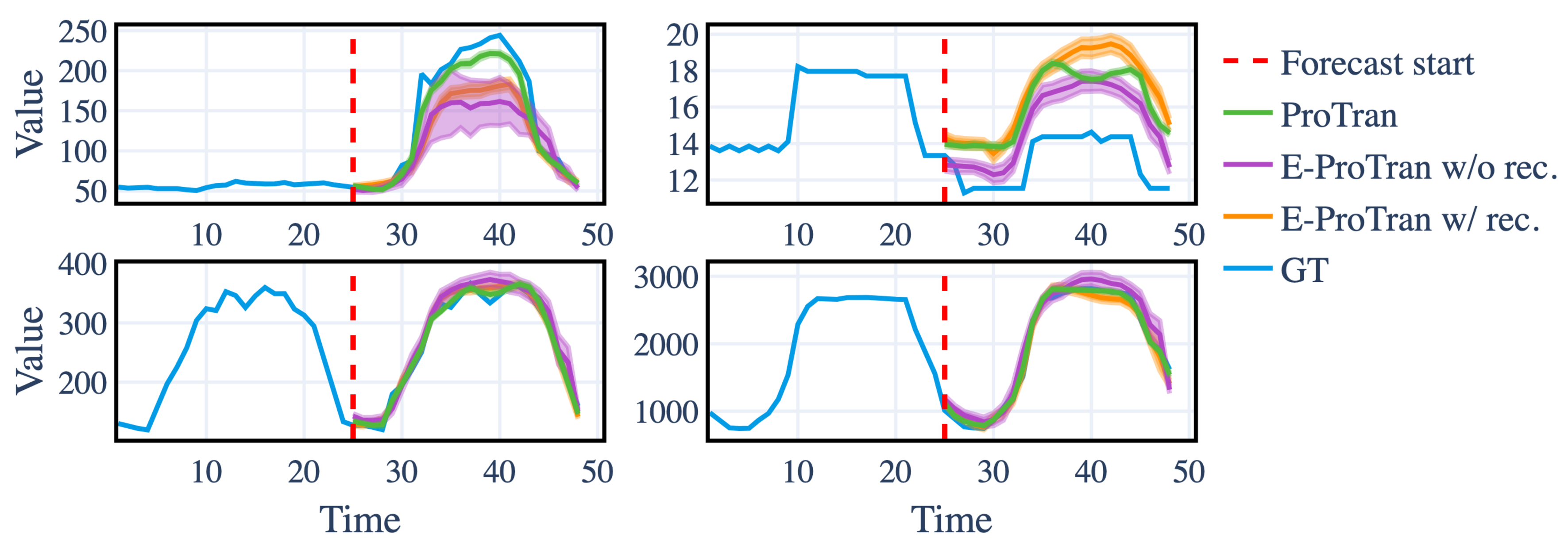
We opt to use

- Non-autoregressive attention to reduce computational overhead while allowing information flow over time through layer attention.
- Causal attention to lower layers to minimize propagation of error from later time steps, and make sure our predictions are independent of the forecasting length.
- Stochastic $z_t$ only at the latent bottleneck; otherwise, it introduces noise in the forward pass and thus affects the gradients during training.

## Results

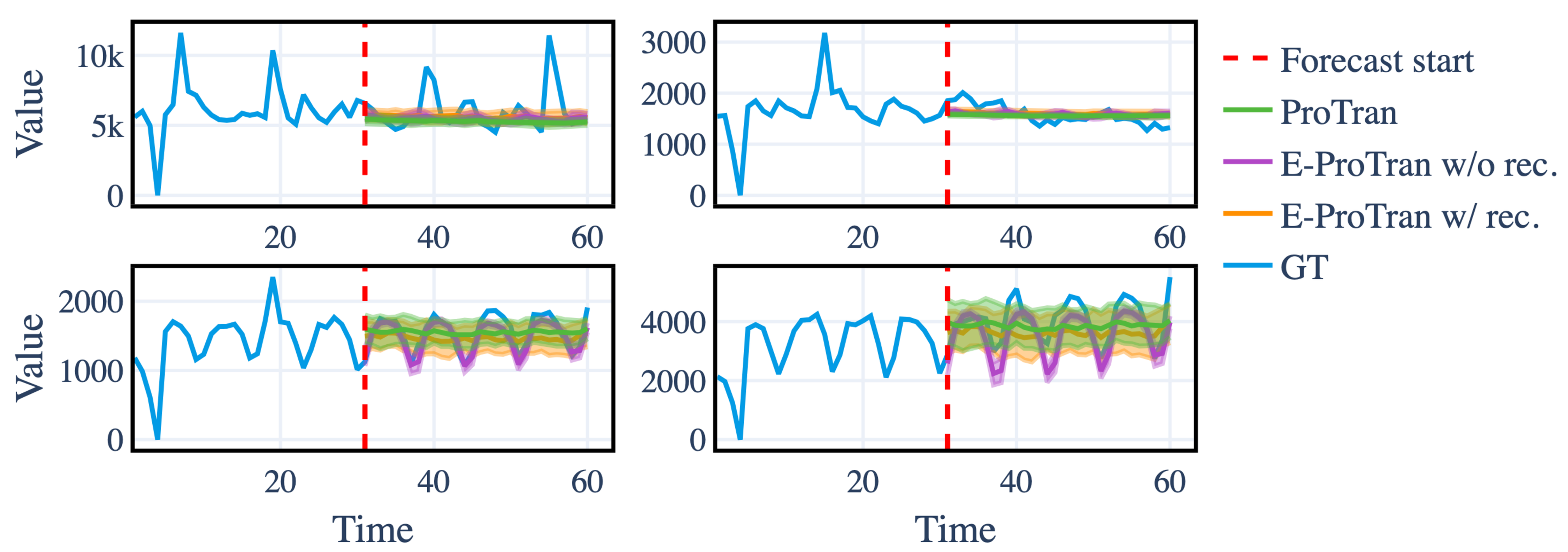
We present the performance metrics for Electricity and Wikipedia. The average predictions are from 100 forward passes; shaded areas show 1 standard deviation.

| Model | | | Electricity | | | | |
|---|---|---|---|---|---|---|---|
| Type | Att. | Rec. | CRPS$_{\text{sum}}$ | CRPS | RMSE | #Params | Forw. Pass (sec) |
| ProTran | L I A | ✓ | 0.030 | **0.069** | **439.2** | 501,924 | 0.064 ± 0.001 |
| E-ProTran | L I A | ✓ | **0.024** | 0.075 | 501.7 | 382,084 | 0.044 ± 0.001 |
| E-ProTran | L I | ✓ | – | – | – | – | – |
| E-ProTran | L | ✓ | **0.024** | 0.072 | 530.2 | **315,652** | **0.002 ± 0.000** |
| E-ProTran | L I A | ✗ | 0.030 | 0.077 | 582.8 | 382,084 | 0.044 ± 0.001 |
| E-ProTran | L I | ✗ | – | – | – | – | – |
| E-ProTran | L | ✗ | 0.029 | 0.079 | 598.7 | **315,652** | **0.003 ± 0.001** |




**Forecasting on Electricity** for the first test sequence. We show 4 of 370 dims.

| Model | | | Wikipedia | | | | |
|---|---|---|---|---|---|---|---|
| Type | Att. | Rec. | CRPS$_{\text{sum}}$ | CRPS | RMSE | #Params | Forw. Pass (sec) |
| ProTran | L I A | ✓ | 0.066 | 0.320 | 5909.2 | 1,522,080 | 0.187 ± 0.003 |
| E-ProTran | L I A | ✓ | 0.075 | 0.353 | 6020.0 | 1,259,584 | 0.141 ± 0.005 |
| E-ProTran | L I | ✓ | 0.063 | 0.328 | 5932.2 | 1,126,720 | **0.007 ± 0.001** |
| E-ProTran | L | ✓ | 0.063 | **0.311** | 5936.8 | **1,060,416** | **0.007 ± 0.002** |
| E-ProTran | L I A | ✗ | 0.081 | 0.354 | 5983.3 | 1,259,584 | 0.139 ± 0.005 |
| E-ProTran | L I | ✗ | 0.054 | 0.327 | 5945.1 | 1,126,720 | **0.007 ± 0.002** |
| E-ProTran | L | ✗ | **0.053** | 0.316 | **5906.7** | **1,060,416** | **0.007 ± 0.002** |




**Forecasting on Wikipedia** for the first test sequence. We show 4 of 2000 dims.

## Looking for More?

* Equal contribution.

Batuhan Koyuncu is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.