# Temporal Protein Evolution Prediction for Proactive Pathogen Surveillance

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Current pathogen biosurveillance systems reactively monitor evolving sequences from environmental samples and assess mutation risks to guide public health responses. To enable proactive surveillance, we developed PEVO, a deep learning framework that explores direct prediction of future protein evolution by learning mappings between temporal and sequence representations. Our approach combines TOTEM (Time-Ordered Evolutionary Modeling) embeddings to capture quarterly temporal patterns with ESM (Evolutionary Scale Modeling) protein sequence embeddings, training a neural network to map from TOTEM time space to ESM sequence space. We demonstrate our method on Ebola virus L protein sequences collected from 1976-2018 ($n = 2343$) with quarterly temporal binning, generating predictions for 2019-2030 and validating on known sequences after 2019 ($n = 596$). While our TOTEM-to-ESM mapping achieved an MSE of 0.002 (RMSE = 0.045), phylodynamics baseline outperformed our approach with an RMSE of 0.007. Additionally, our model's prediction variations (RMSE = 0.045) exceed the natural variability observed in training data (STD = 0.0292), indicating room for improvement in capturing evolutionary constraints. Despite current limitations, this work establishes a foundational framework for temporal-sequence learning that could potentially complement traditional phylodynamic approaches. With further refinement, such neural approaches may offer computational advantages for automated biosurveillance systems, contributing to the transformation of pathogen surveillance from reactive monitoring toward proactive preparedness for future pandemic threats.

## 1 Introduction

Microbial pathogens exist under continuous evolutionary pressure from environmental factors, host immune responses, and anthropogenic interventions such as vaccination and antimicrobial treatment [Smith et al., 2023]. These multifaceted selection pressures drive pathogens through perpetual cycles of mutation, selection, and adaptation, fundamentally shaping their genomic and proteomic landscapes over time. From this perspective, pathogen evolution can be conceptualized as a temporal mapping function that transforms chronological progression into specific genomic and proteomic mutations.

The traditional approach to modeling pathogen evolution relies on phylodynamic methods, which combine phylogenetic analysis with epidemiological dynamics to infer evolutionary parameters and reconstruct transmission histories [Volz et al., 2013]. Established phylodynamic frameworks such as BEAST2 [Bouckaert et al., 2019] and TreeTime [Sagulenko et al., 2018] construct phylogenetic trees from sequence data and estimate evolutionary rates, while platforms like Nextstrain [Hadfield et al., 2018] provide real-time phylogenetic tracking for pathogen surveillance. These methods have

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

proven invaluable for understanding historical evolutionary trajectories and are widely adopted in epidemiological practice due to their interpretability and theoretical grounding. However, phylodynamic approaches typically assume constant or slowly varying evolutionary rates and may struggle to capture complex, non-linear temporal patterns that characterize rapidly evolving pathogens under dynamic selection pressures.

Recent advances in protein language models (PLMs) have demonstrated remarkable success in learning intricate patterns within amino acid sequences, protein structures, and functional relationships. Models such as ESM model series [Rives et al., 2021, Lin et al., 2023], alongside structure prediction breakthroughs like AlphaFold [Jumper et al., 2021], RosettaFold [Baek et al., 2021], and generative approaches like RFdiffusion [Watson et al., 2023], have shown that deep learning can capture the underlying statistical regularities that govern protein sequence space. Critically, PLMs learn representations of the biologically viable protein sequence space—a constrained subset of the theoretically possible sequence combinations. While a protein of $n$ amino acid residues theoretically permits $20^n$ possible sequences, only a minute fraction of these combinations occur in nature, constrained by requirements for proper protein folding, stability, and biological function. This constraint implies that during evolutionary processes, pathogens navigate within this restricted, functionally viable region of sequence space rather than exploring the full combinatorial possibility space.

Parallel developments in temporal representation learning have introduced sophisticated methods for encoding time series data. Notable among these is TOTEM (Tokenized Time Series Embeddings) [Talukder et al., 2024], which provides a framework for learning rich temporal embeddings that capture complex time-dependent patterns. The convergence of these advances in protein sequence modeling and temporal representation learning presents an opportunity to test whether direct mappings between time and sequence representations can outperform traditional phylodynamic approaches for pathogen evolution prediction.

In this work, we introduce PEVO (Protein Evolution Predictor), a novel deep learning framework that combines TOTEM temporal embeddings [Talukder et al., 2024] with ESM-2 protein sequence representations [Rives et al., 2021] to explore direct temporal-to-sequence prediction for pathogen evolution. PEVO represents an exploratory effort to establish learnable mappings between time space and protein sequence space, bypassing explicit phylogenetic reconstruction while aiming to capture evolutionary dynamics through neural architectures.

Using a comprehensive dataset of historical Ebola virus L protein sequences spanning over four decades, we evaluate PEVO's ability to predict future protein sequence embeddings and compare its performance against established phylodynamics approaches. While our initial results show that phylodynamics currently outperforms PEVO (RMSE = 0.007 vs. 0.045), and the neural model introduces variations larger than the natural standard deviation observed in training sequences, this work establishes a foundational framework for temporal-sequence learning that could potentially surpass traditional methods with further development.

## 2 Dataset

We queried GenBank [Sayers et al., 2019] for all Ebola RNA-dependent RNA polymerase (L protein) sequences, which together with VP35 are essential for viral RNA replication and transcription. After removing synthetic sequences and those shorter than 500 base pairs, we processed sequences containing unknown amino acids (Xs) by aligning them with the GenBank reference sequence using pairwise alignment and replacing unknown residues with corresponding reference amino acids. The dataset was temporally split with sequences from 1976–2018 as training data (n=2343) and sequences from 2019 onward as testing data (n=596), creating an approximate 75/25 train-test split.

## 3 Phylodynamics Analysis

For phylodynamic baseline comparison, we created consensus sequences for each quarter using Clustal Omega [Sievers et al., 2011] and EMBOSS [Rice et al., 2000]. We constructed maximum-likelihood phylogenetic trees with IQ-TREE [Nguyen et al., 2015] and estimated evolutionary parameters using TreeTime [Sagulenko et al., 2018]. Future sequences were predicted using continuous-time

Markov chain models with extracted substitution rates, and RMSE was calculated by embedding predicted and actual sequences with ESM2. See details in Appendix A.

# 4 Model Architecture

Our PEVO-TOTEM (Protein Evolution with TOTEM Time Embeddings) model learns the mapping from temporal tokens to protein sequence representations. Unlike conventional approaches that embed sequences into temporal space, our architecture directly maps time periods to ESM2 embedding space, enabling the model to learn $f : \text{time} \rightarrow \text{sequence features}$.

The temporal tokenizer converts quarterly time periods into discrete tokens:

$$\text{Temporal Tokenizer} : \{1976\text{Q}1, 1976\text{Q}2, \ldots, 2030\text{Q}4\} \rightarrow \{0, 1, \ldots, 219\} \tag{1}$$

This creates a vocabulary of 220 temporal tokens spanning 55 years of quarterly periods between [1976, 2030]. Each time period is mapped to a unique integer identifier that serves as input to the TOTEM embedding layer to generate rich temporal representations:

$$\text{TOTEM Embedding} : \mathbb{R}^{B\times L} \rightarrow \mathbb{R}^{B\times L\times d_{\text{model}}} \tag{2}$$

where $B$ is batch size, $L$ is the temporal sequence length (number of quarterly tokens, default $L = 8$, and $d_{\text{model}}$ is the model dimension. In our example study, $d_{model} = 512$. The TOTEM embedding incorporates: 1) Base temporal embeddings: Learnable representations for each time token; 2) Positional encoding: Enhanced with multiple temporal cycles (quarterly=4, monthly=12, weekly=52); 3) Dropout regularization: Applied to prevent overfitting.

We used a six layer transformer encoder with 8 attention heads to processes temporal embeddings to capture temporal features in a 512 dimension space. The transformer uses causal masking to ensure that predictions at time $t$ only depend on information from times $t' \leq t$, maintaining temporal causality. The sequence projection layer uses linear model with layer normalization to project temporal features in 512 dimensions to ESM2-t6-8M embedding space in 320 dimensions.

To prepare the data for model training, we begin with a CSV file containing protein sequences and their associated quarterly collection dates. Multiple sequences from the same quarter are averaged using their ESM2 embeddings, while missing quarters use the previous quarter's embedding. Each sequence is embedded using ESM2 and organized chronologically into quarterly time series. We then construct a sliding window dataset where each input contains 8 consecutive historical quarters and targets the subsequent 4 future quarters, with the window advancing by 1 quarter per step. The `SlidingWindowDataset` creates training pairs:

$$\text{Input} : \quad \text{Historical quarters } [t - w + 1, \ldots, t] \tag{3}$$

$$\text{Target} : \quad \text{Future quarters } [t + 1, \ldots, t + h] \tag{4}$$

where $w$ is the window size (8 quarters) and $h$ is the prediction horizon (4 quarters).

For prediction, the model uses the 8 historical quarters as context to predict the next 4 future quarters. We obtain temporal embeddings for future time tokens using the TOTEM embedding layer, then compute a context vector by mean-pooling the transformer-encoded representations from the historical sequence. This context vector is broadcast across all future time steps and added to the future temporal embeddings. The enhanced embeddings are then projected into the ESM2 sequence space to yield predicted future sequence features.

The model optimizes a multi-component loss function: $\mathcal{L} = \boldsymbol{\lambda}^T \mathbf{L}$, where $\boldsymbol{\lambda} = [\lambda_r, \lambda_p, \lambda_s, \lambda_c, \lambda_d]^T$ are loss weights and $\mathbf{L} = [\mathcal{L}_{\text{recon}}, \mathcal{L}_{\text{pred}}, \mathcal{L}_{\text{smooth}}, \mathcal{L}_{\text{consist}}, \mathcal{L}_{\text{div}}]^T$ are individual loss components.

The reconstruction loss $\mathcal{L}_{\text{recon}}$ uses combined MSE and L1 loss between predicted and ground truth ESM2 embeddings for the 8 historical quarters, evaluating the model's ability to reconstruct known sequences. The prediction loss $\mathcal{L}_{\text{pred}}$ applies the same combined loss to the 4 future quarters, assessing forecasting accuracy for unseen sequences.

Three regularization losses ensure prediction quality: $\mathcal{L}_{\text{smooth}}$ encourages smooth temporal changes in predicted embeddings to match ground truth dynamics; $\mathcal{L}_{\text{consist}}$ enforces temporal consistency by matching first- and second-order temporal dynamics; and $\mathcal{L}_{\text{div}}$ promotes prediction diversity across batch samples using cosine similarity to prevent mode collapse.

Default weights are $\lambda_r = 1.0$, $\lambda_p = 1.0$, $\lambda_s = 0.1$, $\lambda_c = 0.2$, $\lambda_d = 0.05$. The model uses AdamW optimization with cosine annealing learning rate schedule, weight decay of $10^{-5}$, and gradient clipping at 1.0.

## 5 Results and Discussion

We trained the model for 145 epochs on sequences spanning 1976-2019 and evaluated performance on sequences after 2019Q1-2021Q2 (Fig. 1). The model demonstrated near perfect convergence across both the composite total loss (Fig. 1A) and individual loss components (Fig. 1B, D-G) and achieved an MSE of 0.002 for ESM embedding prediction, corresponding to an RMSE of 0.045 (Fig. 1E). In comparison, the phylodynamics baseline produced an RMSE of 0.007 (Fig. 2 in Appendix).

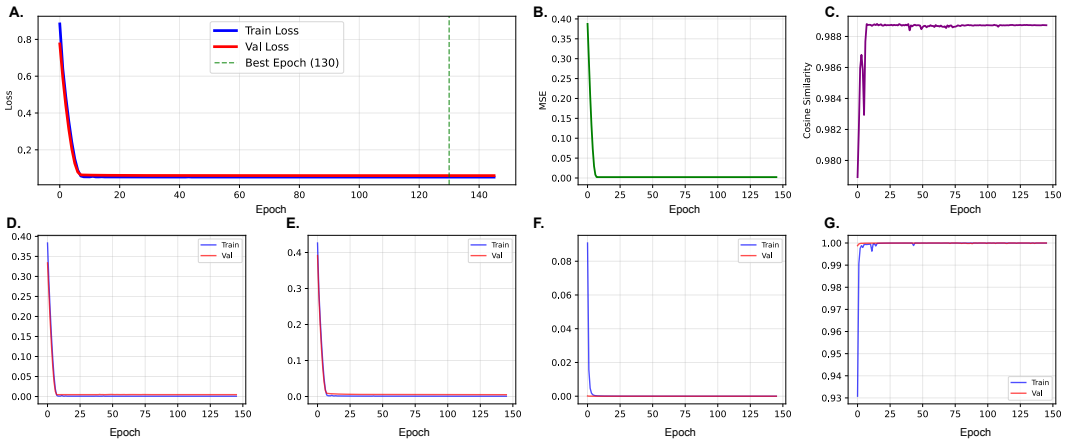


Figure 1: Final training summary from 146 epochs of training. A: Overall loss curve. B: Time-> ESM2 MSE, C: Time-> ESM2 consine similarity, D: Reconstruction ESM MSE, E: Future Prediction ESM MSE, F: Temporal consistency; G: Diversity.

To properly interpret these results, we note that the ground truth ESM embeddings exhibited a mean standard deviation of 0.0292 across all dimensions (Fig. 3 in Appendix). Our model's RMSE (0.045) exceeds this natural variability by 1.5-fold, indicating prediction errors surpass inherent variation in the ESM embedding space. In contrast, phylodynamics achieved superior performance with an RMSE of 0.007—4 times smaller than natural data variability and 6.4 times better than our neural approach.

This performance gap demonstrates that traditional evolutionary modeling remains highly competitive for protein sequence prediction. The gap likely reflects our relatively small dataset, high-dimensional ESM embeddings, sparse data between 1977Q1 and 1994Q3, or specific temporal patterns in Ebola L protein evolution that mechanistic phylodynamic models capture more effectively.

Given the phylodynamics baseline's superior performance, we plan two key refinements: **Dataset Refinement** using only sequences from 1994 onward to address data sparsity and temporal gaps, and **Consensus Sequence Approach** replacing averaged embeddings with consensus sequences per quarter, matching phylodynamics methodology.

While PEVO's direct time-to-sequence mapping approach currently underperforms, it remains valuable for biosurveillance by predicting functional and epidemiological characteristics of future mutations. Comparing predicted embeddings with historical sequences enables quantitative assessment of evolutionary similarity and identification of significant deviations for public health monitoring without requiring explicit amino acid reconstruction.

This work establishes a foundational framework for neural temporal-sequence learning in pathogen evolution prediction. While phylodynamics currently outperforms our approach, PEVO demonstrates the feasibility of direct time-to-sequence mapping and offers potential advantages in computational efficiency and automated integration for real-time biosurveillance systems. The planned refinements targeting data quality and methodological alignment provide clear pathways for improving predictive accuracy.

# References


Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004. doi: 10.1093/nar/gkh340.

James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Subha Kalyaanamoorthy, Bui Quang Minh, Thomas KF Wong, Arndt von Haeseler, and Lars S Jermiin. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589, 2017. doi: 10.1038/nmeth.4285.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yilun Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 2015. doi: 10.1093/molbev/msu300.

Peter Rice, Ian Longden, and Alan Bleasby. Emboss: the european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000. doi: 10.1016/S0168-9525(00)02024-2.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.

Pavel Sagulenko, Vadim Puller, and Richard A Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1):vex042, 2018. doi: 10.1093/ve/vex042.

Eric W Sayers, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, and Ilene Karsch-Mizrachi. Genbank. *Nucleic Acids Research*, 48(D1):D84–D86, 10 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz956. URL `https://doi.org/10.1093/nar/gkz956`.

Fabian Sievers, Andreas Wilm, David G. Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Harry McWilliam, Michael Remmert, Johannes Söding, Julie D. Thompson, and Desmond G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7:539, 2011. doi: 10.1038/msb.2011.75.

William P J Smith, Benjamin R Wucher, Carey D Nadell, and Kevin R Foster. Bacterial defences: mechanisms, evolution and antimicrobial resistance. *Nature Reviews Microbiology*, 21(8):519–534, 2023.

Abera Talukder, Yisong Yue, and Georgia Gkioxari. Totem: Tokenized time series embeddings for general time series analysis. *arXiv preprint arXiv:2402.16412*, 2024.

Erik M Volz, Katia Koelle, and Trevor Bedford. Viral phylodynamics. *PLoS computational biology*, 9(3):e1002947, 2013.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

# A Phylodynamic Baseline Details

Multiple sequences per quarter were aligned using Clustal Omega [Sievers et al., 2011], and consensus sequences were generated with EMBOSS [Rice et al., 2000]. Unknown amino acids were replaced using the previously described method.

For phylogenetic projections, consensus training sequences were aligned with MUSCLE [Edgar, 2004], then maximum-likelihood trees were constructed using IQ-TREE with 1000 bootstrap replicates and ModelFinder for automatic model selection [Nguyen et al., 2015, Kalyaanamoorthy et al., 2017]. Quarter dates were converted to the first day of each quarter and input with phylogenetic trees into TreeTime (clockfilter=5, gtr=infer) [Sagulenko et al., 2018].

The substitution rate ($\mu$) extracted from TreeTime was used with uniform distributions for the rate matrix of amino acid substitutions (normalized $\mathbf{Q}$) and state frequencies. Using the final training consensus sequence (2022Q4) as the starting point, each site was modeled as a continuous-time Markov chain over 20 amino acids with $\Delta t = 0.25$ years. For each site with rate multiplier $r$, the transition probability matrix was calculated as $\mathbf{P}(\Delta t, r) = \exp(\mathbf{Q} \cdot \mu \Delta t r)$.

Predicted sequences were combined with original training data, realigned with MUSCLE, and phylogenetically analyzed using IQ-TREE with ModelFinder and TreeTime (Appendix Fig. 2).

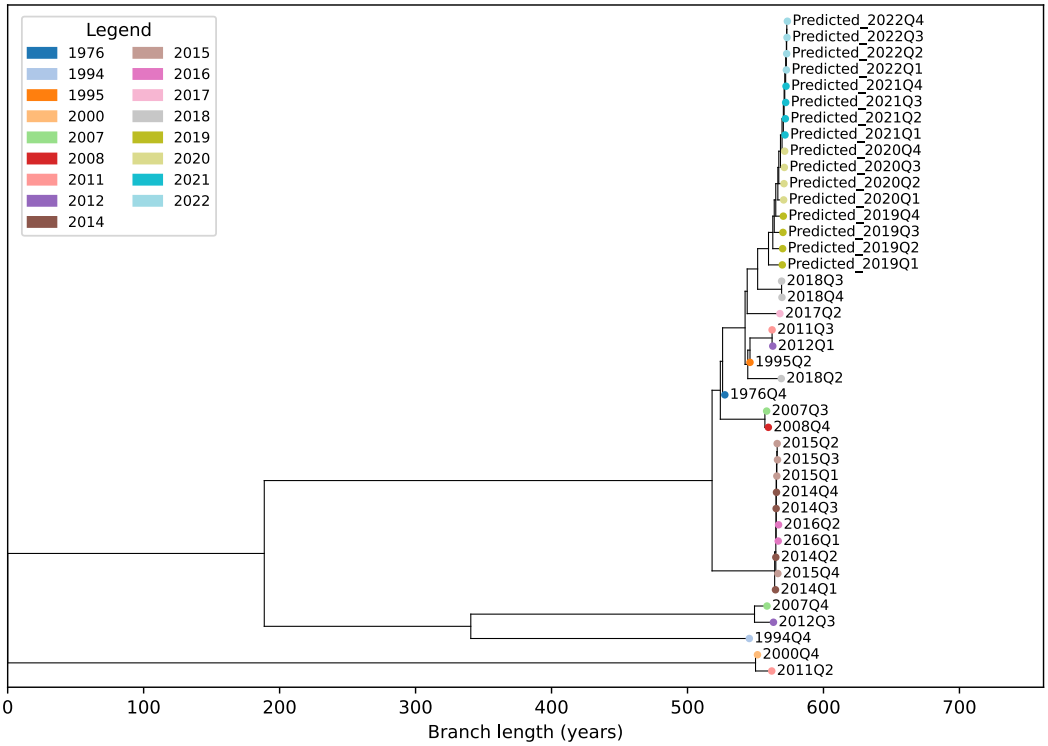


Figure 2: Phylodynamics predictions

# B Loss Terms

Table 1 in Appendix lists the details of the error terms used in PEVO.

# C Per ESM Dimension Variation

We generated sequence-level ESM2 embeddings for all sequences (training and testing) and computed the standard deviation across each embedding dimension (Fig. 3). The mean standard deviation

| Loss Term | What it Measures | Input/Output Used | Formula / Code |
|---|---|---|---|
| $\mathcal{L}_{\text{recon}}$ | Reconstruction loss: how well the model reconstructs ESM2 embeddings for historical (input) periods | Predicted and target ESM2 for historical periods | $0.8\,\text{MSE}(\hat{\mathbf{y}}_{\text{hist}}, \mathbf{y}_{\text{hist}}) + 0.2\,\text{L1}(\hat{\mathbf{y}}_{\text{hist}}, \mathbf{y}_{\text{hist}})$ |
| $\mathcal{L}_{\text{pred}}$ | Prediction loss: how well the model predicts ESM2 embeddings for future periods | Predicted and target ESM2 for future periods | $0.8\,\text{MSE}(\hat{\mathbf{y}}_{\text{future}}, \mathbf{y}_{\text{future}}) + 0.2\,\text{L1}(\hat{\mathbf{y}}_{\text{future}}, \mathbf{y}_{\text{future}})$ |
| $\mathcal{L}_{\text{smooth}}$ | Smoothness loss: encourages smooth temporal changes in predictions | Differences between consecutive predicted and target ESM2 (historical & future) | $\text{MSE}(\Delta\hat{\mathbf{y}}_{\text{hist}}, \Delta\mathbf{y}_{\text{hist}}) + \text{MSE}(\Delta\hat{\mathbf{y}}_{\text{future}}, \Delta\mathbf{y}_{\text{future}})$ |
| $\mathcal{L}_{\text{temp-cons}}$ | Temporal consistency loss: matches the temporal evolution (velocity & acceleration) of predictions to targets | First and second differences (gradients, accelerations) of predicted and target future ESM2 | $\text{MSE}(\Delta\hat{\mathbf{y}}_{\text{future}}, \Delta\mathbf{y}_{\text{future}}) + 0.1\,\text{MSE}(\Delta^2\hat{\mathbf{y}}_{\text{future}}, \Delta^2\mathbf{y}_{\text{future}})$ |
| $\mathcal{L}_{\text{div}}$ | Diversity loss: encourages diversity among predictions in a batch | Pairwise cosine similarity of predicted future ESM2 | Mean of cosine similarity matrix between batch predictions |

Table 1: Summary of loss terms used in the PEVO-TOTEM model. Here, $\hat{\mathbf{y}}_{\text{hist}}$ and $\mathbf{y}_{\text{hist}}$ are predicted and ground truth ESM2 embeddings for historical periods, $\hat{\mathbf{y}}_{\text{future}}$ and $\mathbf{y}_{\text{future}}$ are for future periods, $\Delta$ denotes the first difference (temporal gradient), and $\Delta^2$ the second difference (acceleration).

across all dimensions is 0.0292, representing the natural variability present in the ESM2 embedding space for our Ebola L protein dataset.

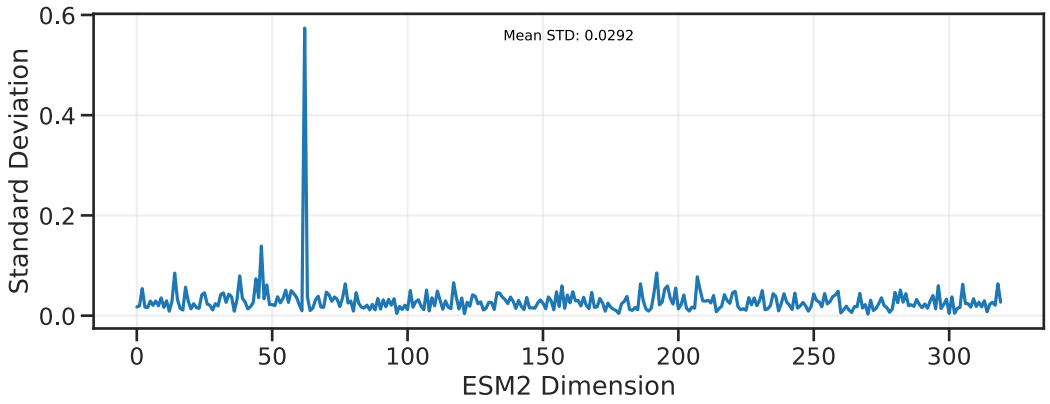


Figure 3: Per ESM embedding dimension standard deviations. X axis: 320 dimensions from ESM2 embedding space. Y axis: standard deviations