

RUSTAMBEK UROKOV

Tashkent, Uzbekistan

+998-77-089-0841 | r.urokov@newuu.uz | github.com/rustambekurokov | huggingface.co/rustambekurokov | linkedin.com/in/rustambek-urokov

Summary

Undergraduate researcher working on protein language models and computational biology. Most of my work is about **benchmarking**: building held-out evaluation protocols for protein and sequence models, and testing whether a computational score actually measures what it is assumed to measure. Both of my manuscripts report negative or corrective results. Comfortable in Python/PyTorch and with genomic and protein data formats.

Experience

Openbank

2026 – Present

Machine Learning Intern

Tashkent, Uzbekistan

- Surveyed diffusion and flow-matching TTS literature, multilingual speech corpora including FLEURS, and TTS scaling studies to set the model and data strategy for Uzbek OmniVoice training.
- Designed and ran a quality-controlled single-speaker corpus pipeline combining diarization, ECAPA speaker verification, speech-boundary segmentation, selective source separation, and 2-of-3 consensus ASR filtering, with an immutable provenance ledger and audit gates before any data is promoted to a training manifest.
- Completed the design of the Uzbek OmniVoice text-to-speech system, covering corpus construction, curation thresholds, the evaluation protocol, and the training-manifest gate.

Manuscripts

- **Sequence-Derived Representations versus Pfam-Domain Content for Biosynthetic Gene Cluster Retrieval.** *First author. Submitted to the ICBINB: Failure Modes of AI in Biology workshop, NeurIPS 2026.*
- **Second manuscript, co-authored, under review at a NeurIPS 2026 workshop.** *Title and subject withheld until the review is complete.*

Research Projects

BGC-SetNet: ESM-2 vs. Pfam Content for BGC Retrieval | *PyTorch, ESM-2, Set Transformers*

2026

- Built the full genomics pipeline end to end: 333 *Streptomyces griseus* genomes from NCBI, Prodigal ORF calling, DeepBGC cluster detection, BLASTp against MIBiG for silver compound-family labels, and genome deduplication, yielding an atlas of 6,953 BGCs and 316,257 proteins with frozen 1,280-dimensional ESM-2 embeddings stored in HDF5.
- Designed BGC-SetNet, a Set Transformer (two induced self-attention blocks, 32 inducing points, 8 heads, multihead attention pooling) over frozen ESM-2 gene embeddings and relative gene positions, trained by masked-gene reconstruction and then a family-level contrastive objective; also implemented a differentiable weighted-Pfam Jaccard as a learned baseline.
- Evaluated under a group-disjoint train/validation/test split fixed before training, validation-only checkpoint and ensemble selection, five seeds, and family-level paired inference on 16 eligible held-out groups with exact Wilcoxon sign-flip tests and 95% bootstrap intervals.
- **Reported a negative result:** plain Pfam-domain Jaccard reached Recall@50 0.879 versus 0.847 for the learned representation, and the validation-selected ensemble was 0.877 ($\Delta = -0.002$, exact $p = 0.625$). Documented that low domain overlap is candidate prioritisation, not evidence of an alternative biosynthetic pathway, and named the silver-label and 16-family power limitations instead of reporting a more favourable subgroup.

ESM-2 Fine-Tuning for Protein Subcellular Localization | *PyTorch, HuggingFace, LoRA, Gradio*

2026

- Fine-tuned ESM-2 from 8M to 650M parameters for 10-class subcellular localization on 24,667 DeepLoc 2.0 proteins, comparing linear probing, full fine-tuning, and parameter-efficient adaptation under one training and evaluation harness.
- Rank-16 LoRA on attention query/key/value projections of the 650M model reached 76.5% accuracy (MCC 0.704) with 2.4% trainable parameters, matching full fine-tuning of the 150M model (76.6%, MCC 0.706), so the extra scale bought nothing at this data size.
- Released both checkpoints on the HuggingFace Hub and a public Gradio Space for interactive evaluation; training on an NVIDIA DGX Spark, evaluation on AWS EC2.

TabPFN vs. NARX for Mooring-Line Tension Prediction | *Python, TabPFN, Weights & Biases*

2026

- Benchmarked zero-shot TabPFN against a domain-specific NARX recurrent network on an ~100,000-row offshore tension regression task, measuring ~1.4s to fit against ~166s to predict and exposing the mismatch between TabPFN’s small-tabular prior and a large physical time-series workload.
- Wrote up the validity threats explicitly, including the missing NARX baseline metrics, an ambiguous NRMSE normalization that inverts the interpretation of the headline number, and sea-state distribution shift, rather than presenting the comparison as settled.

Uzbek Speech Recognition | Python, PyTorch, HuggingFace Transformers2026

- Fine-tuned Whisper-medium on the 52-hour FeruzaSpeech corpus (22,432 augmented samples) to 9.64% test WER and 1.60% CER, with a reproducible pipeline for text normalization, Cyrillic-to-Latin handling, deduplication, duration filtering, and speed/pitch/MUSAN-noise augmentation.

Technical Skills

Languages: Python, SQL, Bash

ML: PyTorch, HuggingFace Transformers and PEFT/LoRA, scikit-learn, FAISS

Protein & Genomic Data: ESM-2 embeddings, FASTA, HDF5, Pfam domain annotations, MIBiG, DeepBGC, Prodigal, BLASTp

Evaluation & Statistics: group-disjoint and held-out splits, retrieval metrics (Recall@k, MRR, MAP, nDCG), bootstrap confidence intervals, exact paired non-parametric tests, multi-seed protocols, preregistration

Tools & Infrastructure: Git, Linux, CUDA, Weights & Biases, AWS EC2, HuggingFace Hub and Spaces, Gradio

Education

New Uzbekistan UniversityOct. 2023 – June 2027

Bachelor of Science in Economics and Data ScienceTashkent, Uzbekistan

- Dean’s List (top 10% of class), Spring 2026. IELTS 7.5 (C1) | SAT 1420.

Leadership & Achievements

IT Community of UzbekistanJune 2023 – Present

Project Management VolunteerTashkent, Uzbekistan

- Coordinated logistics for technical events including Google I/O Extended for 100+ attendees, and interviewed Uzbek engineers at major tech companies for a community page reaching 5,000+ followers.

CAU Business Case Competition – Finalist, top 5 teams of 50+ participants2025

Mathematics for Machine Learning Specialization – DeepLearning.AI (Coursera)2025