

Anya Korsakova

Research scientist with experience in machine learning, computational biology and data engineering.

1743 Golden Gate Avenue, San Francisco CA 94115

anya@codehelix.ai | codehelix.ai | +1 4154180014 | github.com/anyakors

Skills

Programming: python, C++, bash, R, Cypher

ML/AI Frameworks: pyTorch, TensorFlow, scikit-learn, openCV

Cloud & Infrastructure: Google Cloud Platform, Neo4J, Docker, Slurm, LaTeX

Domain: scVI-tools, BioPython, scanpy, pyMOL, bedtools, AMBER MD, MEME suite

Experience

Postdoctoral researcher, Calico Life Sciences LLC (Alphabet) – San Francisco, CA May 2023 – Present
ML architecture design, training, evaluation, and data preparation for biological ML models.

- Built an ML model, TF motif discovery framework, and visualization server for mechanistic interpretability of the Borzoi DNA foundation model using sparse autoencoders (manuscript in preparation).
- Enabled genetic variant effect prediction for structural variants and indels with sequence augmentation; created new benchmarks and improved AUROC performance on existing benchmarks by 9% [1]. Collaborated on an ML-informed statistical fine-mapping framework [2].
- Improved doublet detection in the existing single-cell doublet detection library *solo* via ensembling; extended the approach to snATAC doublet detection.

Postdoctoral Researcher, Cancer Science Institute – Singapore Oct 2022 – May 2023
Mutational and structural variant signature method development for somatic mutations in cancer.

- Devised and implemented a probabilistic mutational signature assignment method ALPS
gitlab.com/PittGenomics/alps.
- Collaborated on building ensemble approaches for mutational signature assignment [3].

Quantitative Finance Developer, Juniper Investment Pte Ltd – Singapore Feb 2022 – Sept 2022
• Programmed high yield multi-timeframe, multi-instrument trading bots for foreign exchange and stock markets in C++ (MQL5) using technical and fundamental analysis.

PhD Scholar, Nanyang Technological University – Singapore Aug 2017 – July 2022
Development of machine learning frameworks for structural and cellular biology.

- Built a G-quadruplex DNA structure prediction model using deep learning and auxiliary epigenetic inputs, achieving state-of-the-art performance in terms of accuracy and enabling cell type specificity [4].
- Developed an RNA splicing prediction framework with an energy-based model augmented with RNA-binding protein levels [5].

Junior Researcher, Lebedev Physical Institute RAS – Moscow, Russia Jan 2014 – Aug 2017
• Modeled combustion (chemical and physical components) by numerically solving systems of partial differential equations in Mathcad and MATLAB; detected diffusive-thermal instabilities in hydrogen-air flames [6].

Junior C++ Developer, NRNU MEPhI – Moscow, Russia Dec 2012 – Dec 2013
• Implemented an algorithm for eye iris recognition and tracking in live video stream with C++ and openCV.

Projects

Hybrid RAG pipeline for a recommendation system hayvin.ai (pre-seed)
• Built a data ingestion, knowledge graph construction, and hybrid RAG pipeline for an agentic networking app recommendation system based on Github and OpenReview profiles.

Education

Nanyang Technological University, Singapore – PhD in Physics	July 2022
NRNU MEPhI, Moscow, Russia – MS in Applied Mathematics and Physics (First Class Honors)	June 2016
NRNU MEPhI, Moscow, Russia – BS in Applied Mathematics and Physics	June 2014

Awards

SINGA Scholarship Award – Nanyang Technological University, Singapore	2017 – 2021
Best Student Award – NRNU MEPhI, Moscow, Russia	2015

Peer Review

Nature Machine Intelligence ISSN: 2522-5839, reviewer, 2024.

Selected Publications

- [1] *Korsakova A et al.*, “**Shift augmentation for improved indel scoring in DNA sequence-based ML models**” bioRxiv, 2025, DOI:10.1101/2025.04.07.647656.
- [2] *Srivastava D., Korsakova A et al.*, “**Borzoï-informed fine mapping improves causal variant prioritization in complex trait GWAS**” bioRxiv, 2025, DOI:10.1101/2025.07.09.663936v1.
- [3] *Wu AJ, Perera A, Kularatnarajah L, Korsakova A, Pitt JJ*, “**Mutational signature assignment heterogeneity is widespread and can be addressed by ensemble approaches**” in Briefings in Bioinformatics, 2023, DOI:10.1093/bib/bbad331.
- [4] *Korsakova A, Phan AT*, “**Prediction of G4 formation in live cells with epigenetic data: a deep learning approach**” in NAR Genomics and Bioinformatics, 2023, DOI:10.1093/nargab/lqad071.
- [5] *Chan A, Korsakova A et al.*, “**RNA alternative splicing prediction with discrete compositional energy network**” at ACM CHIL ’21, DOI:10.1145/3450439.3451857.
- [6] **Google Scholar:** scholar.google.com/citations?user=5A3VUGMAAAAJ.

Datasets

- **Chan A., *Korsakova A. et al.*, “**Context Augmented Psi Dataset (CAPD) to benchmark RNA alternative splicing prediction models**” 2021, <https://doi.org/10.21979/N9/FFN0XH> **contributed equally*

Poster / Talk contributions

- *Korsakova A et al.*, “**Shift augmentation for improved indel scoring in DNA sequence-based ML models**”, ASHG (American Society of Human Genetics) Annual Meeting, Denver, CO 2024.
- *Srivastava D, Yuan H, Korsakova A et al.*, “**Borzoï-guided fine mapping improves variant and gene prioritization in GWAS**”, ASHG (American Society of Human Genetics) Annual Meeting, Denver, CO 2024.