Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Federated Fine-Tuning and Zero-Shot Application of Nicheformer for Spatial Transcriptomics

Course: Biostatistik und Künstliche Intelligenz in der Medizin
Instructor: Jan Baumbach
Project Advisor: Mohammad Bakhtiari (mohammad.bakhtiari@uni-hamburg.de)
Semester: WinterSemester 2025/2026

Overview

In this project, students will explore the intersection of artificial intelligence, spatial transcriptomics, and privacy-preserving computation by implementing a federated version of Nicheformer, a recently published transformer-based foundation model for spatial single-cell transcriptomics.

Nicheformer, introduced by Tejada-Lapuerta et al., Nature Methods (2025) is designed to model cell–cell interactions and microenvironmental niches using large-scale spatial omics data. The model is pretrained on more than 110 million spatial transcriptomic observations and supports downstream tasks such as cell type classification, reference mapping, and niche feature prediction.

Your task is to extend Nicheformer into a federated learning setting, where data from multiple sites (clients) collaboratively contribute to model training and fine-tuning without sharing raw data. This simulates real-world medical research collaborations where patient-level data cannot be centralized due to privacy regulations.

All members of each group are expected to contribute actively to both the implementation and the presentations. Therefore, attendance at all presentation sessions is mandatory for every group member. Each student must also submit an individual written report, even if the implementation work is completed collaboratively.

The project is divided into four milestones, and students must achieve at least 50% of the points in each milestone to be eligible for the final presentation and report submission.

Objectives

Students will:

1. Understand and explore the Nicheformer architecture
   - Review the publication Tejada-Lapuerta et al., Nature Methods (2025) and the accompanying codebase.
   - Learn how Nicheformer models spatial context by integrating molecular and positional features.
   - Familiarize themselves with the model's pretraining setup, architecture, and downstream tasks.
2. Access and prepare data
   - Download public datasets from the SpatialCorpus-110M resource available on Hugging Face:
     - URL: https://huggingface.co/datasets/theislab/SpatialCorpus-110M
   - Alternatively, use the provided data preprocessing scripts in the `data/` directory of the GitHub repository:

        - GitHub: https://github.com/theislab/nicheformer/
    - Explore one or more datasets (e.g., *Visium*, *MERFISH*, or *Slide-seq*) for downstream fine-tuning tasks.
3. Design and implement a federated fine-tuning framework

    - Simulate multiple institutions (clients), each holding a subset of the spatial transcriptomics data.
    - Implement federated training using FedAvg or similar aggregation algorithms.
        - Bonus: Consider privacy-preserving mechanisms such as Secure Multiparty Computation (SMPC) or Differential Privacy (DP) if possible.
    - Use frameworks such as Flower or FeatureCloud for implementation.
4. Conduct two main experiments:

    - Federated Fine-Tuning: Fine-tune pretrained Nicheformer weights on cell-type-labeled datasets distributed across multiple clients.
    - Zero-Shot Evaluation: Test the pretrained model on unseen datasets (without further training) to assess its generalization and reference mapping capabilities.
5. Evaluate and compare results

    - Compare federated vs. centralized performance.
        - Fine-tuning
        - Zero-shot
    - Report metrics such as accuracy, F1-score, and confusion matrices.
    - Analyze communication cost, convergence, and model performance.

Data and Code Access

All resources are publicly available:
Code Repository: https://github.com/theislab/nicheformer/
Datasets: Public datasets used in the paper are hosted on Hugging Face:
https://huggingface.co/datasets/theislab/SpatialCorpus-110M

Project Milestones

To ensure steady progress throughout the semester, the project is divided into four milestones, each aligned with one of the key objectives.
Students are expected to submit deliverables and give a short presentation at the end of each milestone to demonstrate their progress and understanding.

Milestone 1: Understanding the Nicheformer Architecture

- Goal:
  - Gain a deep understanding of the Nicheformer model, its components, and how it processes spatial transcriptomics data.
- Tasks:
  - Review the paper
    - (Bonus) Explore the official GitHub repository and familiarize yourself with the Python package structure and tutorial notebooks.
  - Present an overview of the model's architecture, pretraining objectives, and downstream tasks.
- Deliverables:
  - Short presentation summarizing Nicheformer's architecture, training setup, and key innovations.
  - Date: December 19, 2025

Milestone 2: Data Preparation and Federated Data Partitioning

- Goal:
  - Access, preprocess, and partition spatial transcriptomics datasets to simulate multiple independent data sites (clients).
- Tasks:
  - Download datasets
  - Select one or more datasets (e.g., Visium, MERFISH, Slide-seq) for downstream experiments.
  - Create federated data splits representing different clients (e.g., by batch, assay, tissue type, organ, or spatial region).
- Deliverables:
  - Cleaned and preprocessed dataset ready for federated training.
  - Code or notebook documenting data partitioning strategy.
  - Short presentation describing dataset characteristics and rationale for data splitting.
- Date: January 2, 2026

Milestone 3: Implementation of the Federated Fine-Tuning Framework

- Goal:
  - Develop a working federated training pipeline for Nicheformer.
- Tasks:
  - Run the centralized model to reproduce the reported results in the paper on target dataset and analysis
  - Implement federated fine-tuning
    - Suggested: using frameworks such as Flower or FeatureCloud.

    - Integrate a federated averaging algorithm (FedAvg) for model aggregation.
        - (Bonus) Add privacy-preserving techniques such as Secure Multiparty Computation (SMPC) or Differential Privacy (DP).
    - Implement Federated zero-shot
    - Run the centralized model to reproduce the reported results in the paper on target dataset and analysis
- Deliverables:
    - Code for centralized training and federated fine-tuning implementation.
    - Short Presentation explaining:
        - Technical design and implementation details.
        - Centralized results
            - Bonus: Federated results
- Date: January 16, 2026

Milestone 4: Experimental Evaluation – Fine-Tuning and Zero-Shot Application

- Goal:
    - Evaluate the federated model's performance through fine-tuning and zero-shot generalization experiments.
- Tasks:
    - Fine-tune pretrained Nicheformer weights on distributed datasets.
        - Limited fine tuning in terms of epochs and rounds of training is acceptable!
    - Compare Federated Fine-Tuning results with Centralized Training and client's local training(centralized training limited to a client's local data) .
    - Perform Zero-Shot Evaluation on unseen datasets to assess generalization and reference mapping capability.
    - Compute evaluation metrics such as accuracy, F1-score, and confusion matrices.
- Deliverables:
    - Presentation (20 min) explaining:
        - Federated setup and training
        - Experimental results comparing federated and centralized models.
        - Performance analysis, insights, and challenges encountered.
- Date: January 23, 2026

Final Report and Presentation

Students should consolidate their work into a comprehensive written report and a final oral presentation, summarizing the project's objectives, methods, results, and conclusions. All developed code must be made publicly available in a GitHub repository with clear documentation and instructions for reproducibility.

Report

Students are required to prepare a **5–10 page report (PDF format)** that includes the following sections:

- Project objectives and background
- Federated system architecture and implementation details
- Dataset description and experimental setup
- Results and evaluation metrics
- Discussion of challenges, limitations, and future directions

Submission deadlines:

- Draft submission: *20 February 2026, 23:59* (optional, for feedback)
- Final submission: *7 March 2026, 23:59*

Presentation

Prepare a final presentation (25 minutes) summarizing the outcomes and insights from all project milestones.

Presentations will take place on:

📅 **Date:** *6 February 2026*
🕘 **Time:** *09:00–18:00*
📍 **Location:** *AER 8–10*

The presentation should clearly communicate:

- The motivation and problem formulation
- Implementation strategy and key technical decisions
- Experimental findings and comparisons
- Lessons learned and perspectives for future work

References

- Tejada-Lapuerta, A., Schaar, A. C., Gutgesell, R., Palla, G., Halle, L., Minaeva, M., ... & Theis, F. J. (2025). Nicheformer: a foundation model for single-cell and spatial omics. *Nature Methods*, 1-14.
- https://github.com/theislab/nicheformer/
- https://huggingface.co/datasets/theislab/SpatialCorpus-110M