**HUILIN TAI**

Machine Learning Engineer | Data Scientist | AI Researcher

(646) 866‑9171 | New York, NY | [ht2666@columbia.edu](mailto:ht2666@columbia.edu) | [www.linkedin.com/in/huilintai](https://www.linkedin.com/in/huilintai) | haleyyy2001.github.io

**education**

**Columbia University** New York, United States

*Master of Science, Computer Science (Thesis-Based)* (GPA 4.0) Sep 2024 - Dec 2025 (expected)

**McGill University** Montreal, QC, Canada

*Bachelor of Science, Statistics and Computer Science* (GPA 3.86) Sep 2020 - May 2024

**TECHNICAL SKILLS**

**Programming & Languages:** Python, R, SQL, Java, JavaScript, C, MATLAB, Ocaml, Bash

**Machine Learning & AI:** PyTorch, Scikit-learn, Hugging Face Transformers, CNNs, RNNs, Transformers, VAEs, Diffusion Models, Bayesian Inference, Survival Analysis, Time Series, MCMC

**Infrastructure & Data:** Docker, Kubernetes, Google Cloud Platform, MLflow, Git/GitHub, HPC clusters, MongoDB, Neo4j, MySQL, NumPy, Pandas, SciPy’

**PUBLICATIONS**

**[1]** **Huilin Tai**, Qian Li, Jingtao Wang, Jiahui Tan, Ryann Lang, Basil J. Petrof, Jun Ding. "CellSexID: Sex-Based Computational Tracking of Cellular Origins in Chimeric Models." Cell Reports Methods (Accepted, 2025).

**[2]** Mingxiao Huo, Jiayi Zhang, Hewei Wang, Jinfeng Xu, Zheyu Chen, **Huilin Ta**i, Ian Yijun Chen. "Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding." TTODLer Workshop at ICML 2025 (2025).

**[3]** Pengliang Ji, Chuyang Xiao, **Huilin Tai**, Mingxiao Huo. "T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation." ACM Multimedia Conference (2024).

**[4]** Adam M.R. Groh, Nina Caporicci-Dinucci, Brianna Lu, Maxime Bigotte, Elia Afanasiev, Joshua Gertsvolf, Dale J. Hatrock, Victoria Mamane, Sienna Drake, **Huilin Tai**, Jun Ding, Alyson Fournier, Catherine Larochelle, Jo Anne Stratton. "Ependymal cells undergo an astrocyte-like gliosis in response to chronic and acute neuroinflammation." Journal of Neurochemistry (2024).

**[5]** Xiaorong Guo, **Huilin Tai**, Xiaoqing Li, Peng Liu, Jin Liu, Shan Yu. "SPARC is a novelty prognostic biomarker for ovarian cancer and associated with immune signatures and drug response." Clinical and Experimental Obstetrics & Gynecology (2024).

**[6]** Dehai Wu, Congyi Zhang, Guanqun Liao, Kaiming Leng, Bowen Dong, Yang Yu, **Huilin Tai**, Lining Huang, Feng Luo, Bin Zhang, Tiexiang Zhan, Qiuhui Hu, Sheng Tai. "Targeting uridine-cytidine kinase 2 induced cell cycle arrest through dual mechanism and could improve the immune response of hepatocellular carcinoma." Cellular & Molecular Biology Letters, 105 (2022).

**professional experience**

**The Feinstein Institutes**  New York, United States

*Data Scientist Intern* May 2025 - Aug 2025

**Automated Clustering Performance Evaluation System | PyTorch, Docker, MLflow, Scikit-learn, Plotly**

* Designed experimental framework from hypothesis formulation to outcome validation, defining metrics and statistical testing protocols for multi-class response prediction across treatment conditions.
* Built CI/CD-ready PyTorch pipeline with memory-optimized preprocessing and automated QC; deployed on HPC via Docker and MLflow for reproducible ML workflows.
* Developed comparative evaluation of SpaGCN, Leiden, BANKSY, COVET; integrated MLflow + Plotly dashboards, achieving 0.85 Silhouette Score with interactive visualizations.

**Columbia University** New York, United States

*ML Research Assistant (Supervised by Professor Mohammed AlQuraishi)* Aug 2024 - Present

**Deeplearning Benchmarking Development for AMR Prediction**

* Engineered production-scale genomic classification pipeline using autoregressive Evo model with custom transformer block extraction; implemented MLflow experiment tracking and model versioning across distributed GPU infrastructure for reproducible large-scale genomic inference.
* Optimized memory-efficient distributed system featuring block-wise compression, float16 precision, and HDF5-based active caching; achieved 10x memory reduction enabling full-genome processing on multi-GPU clusters through manual transformer iteration and parallel batch processing.
* Developed robust cross-validation benchmarking framework with leave-n-species-out evaluation and model comparison pipeline; integrated multiple pre-trained model families into unified evaluation system with automated hyperparameter tuning and performance monitoring to compare classic AMR prediction model Kover and Evo.

**Columbia University** New York, United States

*Independent Research*

**T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation (ACM MM 2024)**

* Contributed to the computational backbone of the evaluation pipeline, integrating VLMs (LLaVA-1.5, InstructBLIP) with automated metrics (CLIPScore, BLIPScore, VQAScore) to benchmark temporal coherence.
* Supported large-scale experiments by processing 1,600+ prompts and 5,000+ videos, enabling correlation analysis between model outputs and human preference scores.
* Assisted with statistical validation and annotator consensus checks, strengthening the reliability of temporal evaluation results.

**Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding (ICML 2025 Workshop)**

* Ran large-scale computational experiments validating speculative decoding speedups (3.28×) for multimodal LLaVA-1.5 models across diverse benchmarks.
* Contributed to pipeline optimization and evaluation scripts for draft vs. target models, ensuring reproducible performance testing.
* Assisted with deployment-focused experiments to test scalability of the draft model architecture for on-device settings.

**Large-Scale NLP & Pattern Mining Systems | Python, SpanBERT, Google Gemini API, A-priori Algorithm**

* Built iterative information extraction pipeline on Google Cloud Platform with SpanBERT + Gemini API, applying dual-model validation and confidence-based filtering to extract structured relations from unstructured web text with 90%+ precision across multi-page corpora.
* Architected end-to-end NLP workflow combining Google Custom Search, Beautiful Soup, and spaCy NER with relation detection, automating classification of 4+ entity-relation types from web-scale data.
* Implemented scalable A-priori pattern mining on NYC Open Data (1K+ records), generating interpretable association rules and actionable insights for urban analytics with configurable thresholds.

**Mila - Quebec AI Institute** Montreal, Canada

*Research Assistant (Supervised by Professor Jun Ding)*  Dec 2022 - May 2024

**Data Mining and Unsupervised Cell Profiling**

* Collaborated with the Lady Davis Institute, Segal Cancer Centre and Dartmouth Cancer Center to perform large-scale data mining in Python, employing advanced data structures and parallel processing to handle multi-terabyte single-cell datasets.
* Fine-tuned unsupervised models for clustering PBMCs using PCA and UMAP, discovering pathogenic cell subsets and transcriptional signatures.
* Developed a custom pipeline in Python to automate the analysis workflow,Visualized complex cellular hierarchies and gene expression patterns using Seaborn and Matplotlib.

**Large-Scale Data Mining & Dimensionality Reduction for Cellular Profiling**

* Extended the Cellar single-cell analysis platform in collaboration with Stratton Lab, implementing robust file-format support and sharable visualization endpoints to streamline cross-functional research.Developed standardized R and Python pipelines for single-cell RNA-seq data preprocessing and visualisation, adopted lab-wide.
* Created UMAP visualizations and gene expression heatmaps to identify distinct epidermal cell populations; facilitated knowledge transfer through documentation and training.

**Machine Learning Application for Sex Prediction**

* Implemented ML models for sex prediction using XGBoost, Random Forest, SVM, and Logistic Regression. Processed raw 10X Genomics single-cell RNA-seq data, performing quality control, normalization, and feature selection across 2k+ cells.
* Conducted comparative analysis of sex-specific gene signatures across autosomal, X-chromosome, and Y-chromosome gene sets. Applied dimensionality reduction techniques to visualize cell clusters and identify sex-specific expression patterns.
* Achieved 96% prediction accuracy via ensemble modeling and Bayesian hyperparameter optimization, validating sex-determining gene candidates through differential expression analysis against published data

**Deep Learning Model for Spatial Gene Expression Prediction**

* Designed a Variational Autoencoder with integrated Stochastic Variational Inference in PyTorch, enabling scalable learning for spatial transcriptomics data across diverse tissue types.
* Engineered custom Evidence Lower Bound and shift sigmoid transformation functions, enhancing model convergence and interpretability of latent space representations.
* Crafted interactive visualizations using Matplotlib to map clustering results onto scaled tissue images, revealing spatial patterns of genomic feature interactions in physical tissue contexts.

**McGill University** Montreal, Canada

*Research Assistant (Supervised by Professor Hamed Hatami)*  Sep 2023 - Apr 2024

**Randomized Group-Testing Algorithm Design for Hamming Distance Communication Protocol**

* Devised a protocol based on a group testing algorithm to estimate Hamming distance between two n-bit strings and reduced upper bound of communication complexity from O(log n) to O(log log n).
* Adopted adaptive testing algorithm to improve algorithm efficiency, focusing on randomized algorithms for Hamming distance estimation.

**Theoretical Bounds on Excess-Error Replicability in Agnostic Learning**

* Investigated excess-error dependent replicability in agnostic learning to identify conditions that exhibit excess-error dependency.
* Designed an algorithm for covering hypothesis classes with finite VC dimensions, achieving accurate approximation of real error rates using empirical data.

**McGill University** Montreal, Canada

*Research Assistant (Supervised by Professor Anmar Khadra)*  Nov 2022 - Jan 2024

**Markov Chain Monte Carlo Modeling of Nanoparticle Binding Dynamics**

* Implemented advanced probabilistic frameworks in MATLAB to model nanoparticle binding dynamics, correlating IFN-γ dosage and pMHC valence with strong experimental alignment.
* Optimized a serial engagement model through Markov Chain Monte Carlo (MCMC) simulations, analyzing T-cell activation under various geometric constraints.

**Quantifying Multivalent Nanoparticle T-Cell Therapies**

* Utilized mathematical models employing Poisson and Rayleigh distributions to investigate more than 20 selective binding phenomena towards clustered and uniformly distributed TCR surface topologies.
* Developed a MATLAB-driven computational framework leveraging randomization algorithms to calculate binding capacities and visualize TCR surface distribution probabilities.

**Harbin Medical University** Hybrid

*Data Scientist Intern* May 2021 - May 2022

* Constructed survival analysis to estimate if targeting uridine-cytidine kinase 2 could improve immune response of hepatocellular carcinoma. Implemented an interactive monogram with R language for 1-year, 3-year, and 5-year survival analysis.
* Launched a GSEA analysis using R language, processed data with different tumor risk levels to elucidate enrich-ment patterns within KEGG and Hallmark pathways, and identified pathways in order of Normal-TumorLR- Tumor-HR according to enrichment output.
* Led a pivotal study on SPARC gene's impact on ovarian cancer prognosis, employing Kaplan-Meier survival analysis and validating findings via log-rank tests.

**Yooden Technology** Shanghai, China

*Data Analyst* Oct 2020 - Feb 2021

**Real-Time Energy Optimization Pipeline for Data Center Robotics | Python, SQL, ETL, Anomaly Detection, Streaming Data**

* Engineered a streaming data pipeline for energy-saving robots in data centers, ingesting multi-sensor streams with SQL-based storage, ETL transformations, and anomaly detection logic, supporting real-time monitoring of power consumption.
* Contributed to interactive dashboards and automated load-balancing alerts using Python (Pandas, NumPy, Matplotlib) and containerized deployment (Docker); integrated scheduling workflows to enable proactive energy optimization and improved system reliability.

**EXTRACURRICULAR**

**King Abdullah University of Science and Technology** Remote

*Research Associate* May 2024 - Sep 2024

* Designed generative model architecture combining state-of-the-art video editing frameworks with Beta-VAE for disentangled representation learning, implementing custom loss functions and training protocols in PyTorch.
* Conducted comprehensive literature analysis of 10+ cutting-edge papers in video editing and generative modeling, delivering weekly technical presentations to research team with critical analysis of model architectures, experimental methodologies, and performance benchmarks.

**McGill University** Montreal, Canada

*Course Assistant* Sep 2022 - May 2024

* Colaborated with 16 colleagues and graded over 20 assignments for Math235 Algebra, Math240 Discrete Math, Math308 Fundamentals of Statistical Learning, and Math356 Honor Probability.
* Discussed Bayesian and Generalized Linear Models for survival analysis, with a focus on managing missing data and detecting risk factors.

**honors**

Mackey-Glass Research Bursary, Issued by McGill Faculty of Medicine Apr. 2023

Hugh Brock Scholarship, Issued by McGill University Sep. 2020