

Yanjie Chen

New York, NY 10025
github.com/YanjieChen2002

(646) 533-7307
yc4594@columbia.edu

RESEARCH INTERESTS

Computational Genomics, Bayesian Inference, Causal Inference, Modern ML Solutions

SKILLS

Programming:	R	Python (PyTorch)	Bash/Linux	LaTeX	MATLAB
ML/Stats:	Bayesian Inference	Random Forests	Time Series Analysis	NLP	Transformers

EDUCATION

Columbia University
Master of Arts in Statistics
Coursework: Advanced Machine Learning, Deep Learning, Probability Theory, Computational Statistics

New York, US
GPA: 3.8/4.0
Sep 2024– Dec 2025

University of Liverpool
Bachelor of Science in Mathematics
Coursework: Stochastic Modelling, Game Theory, Statistics and Probability, Operational Research, Modeling in R

Liverpool, UK
GPA: 3.9/4.0
Sep 2022 – May 2024

Xi'an Jiaotong-Liverpool University
Bachelor of Science in Applied Mathematics
Coursework: Dynamic Modeling, Linear Algebra, Programming in Java, Analysis

Suzhou, China
GPA: 3.7/4.0
Sep 2020 - Jun 2022

- Received Undergraduate Academic Excellence Performance Scholarship

RESEARCH EXPERIENCES

Irving Institute for Cancer Dynamics, Columbia University
Research Intern & Staff Associate
Advisor: Khanh Ngoc Dinh

New York, US
Jan 2025 – Jun 2027

- Developing DECODE, a machine learning algorithm for clustering mutations from bulk DNA-seq data by implementing mathematical results from population genetics
- Integrating random forest methods to enhance Bayesian parameter inference and Generalized Information Criterion to select parsimonious DECODE parametrization, resulting in ~10× improvement in runtime
- Applying synthetic tests to compare DECODE against >50 existing methods, from which DECODE was shown to outperform all in predicted clonality, clonal composition and mutational assignment
- Implementing DECODE to analyze >15,000 cancer samples from multi-sample AML dataset, The Cancer Genome Atlas (TCGA), and International Cancer Genome Consortium (ICGC); Performed multivariable Cox proportional hazards modeling to quantify the impact of inferred clonality and evolutionary parameters on patient survival
- Preparing parallel computing framework and performing large-scale analysis on Columbia’s High-Performance Computing system
- First-author manuscript completed, preprint (<https://www.biorxiv.org/content/10.64898/2026.06.15.732415v1>)

Memorial Sloan Kettering Cancer Center
Research Assistant
Advisor: Li-Xuan Qin

New York, US
Aug 2025 – Jul 2026

- Debugging and optimizing SyNG-BTS, a data-augmentation framework for tumor RNA/miRNA-seq, stabilizing both offline augmenters (autoencoder, Gaussian) and online generative models (VAE, flow-based models, GAN)
- Performing large-scale repeated experiments to evaluate augmentation performance across >5 TCGA subtypes, >20 parameter configurations, 3 data-normalization strategies, and >20 augmentation batches; developing Slurm-based HPC execution options
- Refining RNA-seq augmentation by tuning CVAE architectures, including customized network depth and latent structure, to better model high-dimensional expression spaces
- Co-first-author manuscript in preparation; target submission: Q4 2026

DitecT Lab, Columbia Engineering

New York, US

Data Researcher

Advisor: Xuan Sharon Di

Sep 2025 – Dec 2025

- Compiling NYC TLC yellow/green taxi trips with holiday and weather data to evaluate CBD congestion pricing
- Building lean cleaning/feature pipeline and performing time-series & causal analyses (DiD, event-study, survival) to quantify demand shifts; exploring substitution and behavioral mechanisms

TEACHING EXPERIENCES

Azure Partners

New York, US

Teaching Assistant

Jun 2025 – Jul 2025

- Taught a two-week AI module covering image classification, object detection, and introductory RNN/attention/transformers
- Designed hands-on PyTorch assignments for beginner-to-intermediate Python learners; hosted 1:1 and group office hours
- Recorded lecture videos and tutorial walkthroughs to support asynchronous learners

PUBLICATIONS & PREPRINTS

- Chen Y., Dinh K. N. Automatic perturbation kernels for biological applications of Approximate Bayesian Computation sequential Monte Carlo with random forests. 21st Conference on Computational Science and Intelligence Systems FedCSIS 2026 (IEEE#69809). Accepted for publication.
- Chen, Y., & Dinh, K. N. Refining breast cancer stratification through improved mutation clustering of genomic samples. Under review.
- Chen, Y., Jaksik, R., Terranova, P., El Baghdadi, S., Koval, A., Kurpas, M. K., Tavaré, S., Kimmel, M., & Dinh, K. N. (2026). Accurate detection of tumor clonality and ongoing expansion mode from genomic data. Under review. PubMed: <https://pubmed.ncbi.nlm.nih.gov/42367953/>
- Generative data augmentation methods for miRNA/RNA-seq.
Co-first-author manuscript in preparation with Qin Lab collaborators.

PRESENTATIONS

Joint Statistical Meetings (JSM)

Boston, US

Speaker

Aug 2026

- Chen, Y. “Reliable Inference of Tumor Expansion Rate and Clonality from DNA-Sequencing Data.”
Topic-Contributed Paper Session, “Stochastic Processes Describing Cancer-Related Biological Systems: Modeling and Inference”

INTERESTS

Piano, Go, Soccer