Dongshen (Dan) Peng

Chapel Hill, NC | (919)-360-5451 | dan.peng.1202@gmail.com | LinkedIn

Research Interests

NLP / LLM / Multimodal Learning / AI4Science / Precision Diagnostics & Medicine / Single-cell Analysis

Skills

Languages: Python, Java, Bash, C++, TypeScript, JavaScript, HTML/CSS, R, GO
NLP: Transformer, Sparse MoE, BERT, GPT2, GPT3, CoT, Llama3
Multi-Modal: CLIP, ViLT, ViT, LLaVA
Generation Models: GAN, DCGAN, DDPM, Guided Diffusion, DALLE2
Post-training: LoRA, RLHF, DPO, SimPO
Medical VLMs: LLaVA-Med, MedFlamingo, PLIP, AlphaFold3, scGPT, LLEMR, RULE
Video: TimeSformer, VideoMAE, LLaVA-NeXT, VideoMamba, CogVideoX
Tools: Git, PyTorch, Tensorflow, Jenkins, Spring Boot, Redis, Kafka, AWS, PostgreSQL, JPA, Hibernate, MongoDB, Express, React, Node.js, Terraform, Gin, gRPC, Docker, Kubernetes, Angular, Pydantic, SQLAlchemy, FastAPI, RESTful API, PowerPoints

Education

University of North Carolina at Chapel Hill Aug 2022 – May 2026
Honors B.S. in Computer Science & B.S. in Statistics and Analytics (GPA:3.8)

- **UNC Summer Undergraduate Research Fellowship** Highly selective granted academic fellowship (20/200+)
- **UNC Phi Beta Kappa:** Less than 1% of all college students qualify for acceptance

Workshop

1. LIFTED: Multimodal Mixture-of-Experts for Clinical Trial Outcome Prediction. Zheng W, **Peng D**, Xu H, Li Y, Zhu H, Fu T, Yao H. *ICML Foundation Models in the Wild workshop*, 2024
2. Epigenetic Profiling of Crohn's Disease: Analyzing Differential Chromatin Accessibility in Perianal Fistulizing and Non-Fistulizing Phenotypes, **Peng D**, Hamilton N, Furey T. *UNC Summer Undergraduate Research Fellowship*, 2024
3. Manifold Learning Benchmarks of the ScRNA-seq Analysis, **Peng D**, Klissouras A, Parker W, Li D. *UNC Biostatistics Undergraduate Summer Internship*, 2023
4. CRISPR/Cas9 Gene Editing in Drosophila via Visual Selection, Kockel L, ..., **Peng D**, Kim SK. Harvard Summer Session, 2023.

Experience

UNC Chapel Hill — NLP & LLM RA (Advisor: Huaxiu Yao) Jul 2023 – Present
LLM-Powered Chat Agents For Health Care, Department of Computer Science

- Developed weakly-supervised deep learning projects by **PyTorch** using visual-language-based features for better performance while balancing inference of **Transformer-based** models (**GPT**, **Llama**, **Llava**)
- Fine-tuned backbone VLMs with **200K+** multimodal data (tabular & Image-caption) for **EHR analysis and Image QA** by common post-training methods (**LoRA**, **RLHF** & **DPO**)
- Worked with multiple clinicians and scientists on **data collection**, **project management**, **report writing** and **data analysis** to non-domain collaborators, enabling faster iteration and code review on model development

UNC Chapel Hill — Data Science RA (Advisor: Didong Li) May 2023 – Present
Developing scRNA analysis with Machine Learning, Departement of Biostatistics

- Streamlined the data cleaning, normalization, processing, and visualization to accommodate **1,000+** high-dimensional ($\geq$ **10,000** expressed genes) single-cell RNA sequence data efficiently by **Python** and **R**
- Integrated vision-language models (**CLIP**) with spatial transcriptomics datasets of **1M+** image-gene expression pairs, implementing **contrastive learning** to maximize cosine similarity between aligned pairs
- Fine-tuned models (**CLIP, PLIP, UNI**) with higher mean **F1 scores** compared to zero-shot baselines and significantly improved tissue subtype classification across spatial transcriptomic datasets

- Designed and performed rigorous benchmarks on the **manifold learning** methods (PCA, UMAP, t-SNE, PHATE, Laplacian Eigenmaps, Hessian LLE...) in **12** human and mouse single-cell RNA datasets

UNC Chapel Hill — Bioinformatics RA (Advisor: Terry Furey) Dec 2023 – Aug 2024

Identifying New Disease Phenotypes by Epigenetics Analysis, Department of Genetics

- Adapted **version control** and **code testing** using GitHub, ensuring a robust peak calling software, ROCCO, for identifying open chromatin regions mainly from Inflammatory Bowel Disease (IBD) patients
- Processed **ATAC-seq** data of **70+** IBD patients by **PEPATAC** and **bedtools**; Removed batch effects by **PCA**
- Proposed and identified a marked, quantifiable difference in the accessible chromatin data between IBD patients with and w/o perianal phenotypes by **RUVseq** and **DEseq2** in **R** for differential expression analysis

BGI Group — Bioinformatics Fellow (Advisor: Jiguang Peng) May 2022 – Aug 2022

Optimizing Prenatal Chromosomal Disorder Testing, Department of Maternal and Child Health

- Operated robust **quality control** and **validation** on Non-Invasive Prenatal Testing (NIPT) sequenced genome samples (75-bp single end reads) by **Linux CLI** and **FastQC**; Implemented GC correction by **deeptools**
- Partnered with engineers and client team to power the **multiple sequence alignment algorithm by Hidden Markov Model** and reduce average alignment error on nucleotide biological datasets by **4.5%**

Research Work Reproduction & Improvement

SigmaFold: Simplified Implementation of Alphafold Github | *Protein Structure Prediction*

- Developed a PyTorch-based framework to reproduce core architectural concepts of **AlphaFold2 (Nature 2021)** and **AlphaFold3 (Nature 2024)**, enabling unified modeling of proteins, nucleic acids, and small molecules
- Implemented **Evoformer** to facilitate joint attention and coordinate prediction for ligand docking
- Created dataset loaders compatible with PDB, Rfam, and ZINC databases to support model training and eval

Multimodal Large Language Models (MLLM) Eval Github | *Multimodal LLM Inference*

- Implemented **Qwen2-1.5B** on the **GSM8K** dataset to evaluate LLM performance in math problem-solving
- Integrated **MiniGPT4** with the **MME** dataset to advance vision-language model capabilities in multimodal tasks
- Conducted extensive data preprocessing and model fine-tuning, achieving a BERT F1 score (0.82) for Qwen2-1.5B and a total perception score (468.73) for MiniGPT4

Long VideoQA by Divide-Conquer Github | *Video Processing*

- Developed a **Divide-and-Conquer** method to extend video recognition capabilities from 2-minute clips to 10-minute videos by segmenting long videos and applying LLM-based summarization techniques
- Reproduced existing video understanding models, such as **VideoChat** and **Video-ChatGPT (ACL 2024)**
- Implemented a pipeline that combines segment summaries to achieve comprehensive understanding of extended video content without significantly increasing computational costs

Tree Crown Detection & Segmentation by Deep Learning Github | *Object Detection, Image Segmentation*

- Led a team of 4 students on tree crown detection using multiple methods including **CNN**, **UNet**, **pre-trained SAM2**, and pseudo-masking by **YOLOv11** to overcome **unlabeled geographical dataset challenges**
- Implemented over 400 human annotations for ground-truth mask data by Roboflow
- Designed a comprehensive benchmark to evaluate models by Test Accuracy, Average F1 Score, and Average IoU, where we found that UNet trained with 70 masks showcased the best performance

Identifying Breast Cancer in Pathology Images by CNN Github | *Image Classification*

- Conducted rigorous data preprocessing, including image normalization and augmentation techniques by **Keras**
- Compiled and trained the **CNN** model (ROC-AUC: 0.93) using the Adam optimizer and binary cross-entropy loss

Software Development Projects

Home Picks (E-Commerce Start-up) | *Java, Spring Boot, Redis, Kafka, AWS Lambda, PostgreSQL, Hibernate*

Facebook Clone (Social Media Full-stack Platform) Github | *Java, MERN, AWS, Terraform*

OnlyBank (Digital Bank Web Service) Github | *Go, PostgreSQL, Redis, Gin, gRPC, Docker, Kubernetes, AWS*

Pro Hub (Organization Roster Platform) Github | *Angular, TypeScript, Pydantic, SQLAlchemy, FastAPI*

AutoDPD (Automatic PyPI Dependency Detector) Github | *Python*

Services

Instructional Assistant at UNC Chapel Hill Jan 2024 – Present

- Instructional Assistant for MATH 235 Mathematics for Data Science, STOR 435 Intro to Probability
- Graded homework and Held Q&As Sessions for 200+ students