

JOHNATHAN JIA, M.S.

Houston, TX · jjia.vercel.app · jdjia93@gmail.com · github.com/jjia1

SUMMARY

Computational biologist and ML scientist with 5+ years applying ML to biological problems, a peer-reviewed publication record, and open-source contributions to comp-bio tooling. Combines LLM post-training (SFT, RLVR, eval-driven development) and interpretability research with single-cell and clinical-genomics depth and end-to-end production ownership.

TECHNICAL SKILLS

ML & LLM Post-Training: SFT (LLaMA-2), RLVR, GRPO-style rollouts, reward/advantage analysis, LLM-as-judge & eval-driven development, cross-encoder reranking, RAG/hybrid search; PyTorch, TensorFlow/Keras, Hugging Face, LangChain, CUDA; hidden-state probes, SHAP/LIME

Computational Biology: single-cell & bulk RNA-seq, scATAC-seq, pseudotime/trajectory analysis, differential expression, PPI networks (STRING), Bioconductor, NextFlow; NCBI, Ensembl, UniProt, PDB

Languages & Data: Python, R, SQL, Bash; Pandas, NumPy, Polars, PySpark/Spark, Dask, SciPy; large-scale tokenization (150M+ clinical events)

Backend & Infrastructure: FastAPI, PostgreSQL, FAISS, Docker, Singularity, AWS (EC2), Azure, Slurm, Palantir Foundry, Git/GitHub

RESEARCH

Detectability of Doomed Reasoning Chains from Hidden States · Sole author; initiated and led

- Initiated and led an independent study of whether linear probes on a reasoning model's hidden states can flag doomed rollouts mid-generation to shape RLVR training: generated 16K GRPO-style, verifier-graded rollouts on MATH (DeepSeek-R1-Distill-Qwen-1.5B) and trained residual-stream probes predicting eventual correctness (pooled AUC 0.708; 0.63 → 0.80 along the chain).
- Simulated probe-guided truncation across a threshold sweep, quantifying the compute-vs-gradient-fidelity tradeoff (80% kill precision at 7.4% token savings, 0.93 GRPO-advantage correlation; 23% savings degrades correlation to 0.75), and isolated a difficulty-vs-trajectory confound via problem-level splits.

EXPERIENCE

AI Engineer · Baylor College of Medicine

2024 – Present

- Led TEDDY's early model-development and distributed-training phase, building a versioned Spark/Palantir pipeline that transformed 150M raw clinical events into a 73M-event, 1.6M-patient corpus; trained PyTorch Lightning models across 8 A100 GPUs with AWS SageMaker and W&B. **Designed leakage-controlled evaluation across 797 ICD-10 tasks; the final model reached 0.72 median AUC and 0.793/0.847 AUROC for asthma/ADHD.**
- Optimized the open-source **decontX** Python package by eliminating a dense-to-sparse conversion and adding concurrency, **cutting runtime ~99%** (12+ hours to reach L1 convergence → 4–5 minutes for all 500 iterations) on a 60k-cell × 40k-gene dataset.
- Prototyped an LLM-as-judge for MentorMatcher's ranking step, evaluated it as unstable, and replaced it with a fine-tuned cross-encoder reranker (RoBERTa) over hybrid retrieval; shipped the end-to-end service on FastAPI, LangChain, PostgreSQL, Docker, and AWS EC2.
- Built a semi-supervised lesion-detection pipeline (pseudo-labeling, active learning, temporal propagation) over 3k+ endoscopy images and dozens of videos, training a classifier to 0.88 AUROC toward real-time video inference.

Data Analyst Associate · Baylor College of Medicine

2023 – 2024

- Built a RAG-powered search and chat service (FAISS + LangChain) matching students and faculty; deployed live at the TCH Research Symposium.
- Analyzed 30+ single-cell and bulk omics datasets across 10+ collaborations, including the AAV-serotype transduction study of olfactory sensory neurons published in *Frontiers in Neuroscience* (co-first author).

Graduate Research Assistant · UTHSC Graduate School of Biomedical Sciences

2020 – 2022

- Developed a self-attention CNN (TensorFlow/Keras) predicting viral transcription-factor targets from sequence, improving accuracy 23% (published in *Cell Patterns*); addressed 10× class imbalance with downsampling and 10-fold cross-validation, with the model extrapolating to held-out viral integration sites at 0.77 AUPR/AUROC.

EDUCATION

M.Sc. Quantitative Sciences · UTHSC Graduate School of Biomedical Sciences (*Best Seminar & Best Poster, 2022*)

2020 – 2022

B.S. Biology & B.A. · Emory University

2011 – 2015

SELECTED PUBLICATIONS

Neeley MB*, Botas J*, Jia J, et al. (2026). TEDDY: A Pediatric Foundation Model for Risk Forewarning from ICD-Coded Diagnostic Histories. *arXiv:2607.14191*.

Belfort BDW, Jia JD*, et al. (2025). Comparative Analysis of AAV Serotypes for Transduction of Olfactory Sensory Neurons. *Frontiers in Neuroscience*.

Xu H, Jia J*, Jeong H, Zhao Z. (2023). Deep learning for detecting and elucidating human T-cell leukemia. *Cell Patterns*.

* co-first author / equal contribution

CERTIFICATIONS & LANGUAGES

Certification: Enhancing Data Science Outcomes with Efficient Workflow — NVIDIA

Languages: English (Native), Mandarin Chinese (Professional Working Proficiency)