

Patrick Yang

py235@cornell.edu • github.com/PatrickYang2007 • linkedin.com/in/patrick-yang-26849a287

Education

Cornell University, College of Engineering

Bachelor of Science in Computer Science

Ithaca, NY

Expected December 2028

Relevant Coursework: Reinforcement Learning, Machine Learning, Numerical Analysis, Probability & Statistics, Object-Oriented Programming & Data Structures

Experience

Undergraduate Researcher — Cornell University

August 2026 – Present

Bindel Group, Department of Computer Science

Ithaca, NY

- Leading research project on numerical stability of deep reinforcement learning with Prof. David Bindel: analyzing divergence in target-network training via fixed-point iteration and spectral analysis, toward computing rather than hand-tuning the target-network update rate

Research Software Engineer — Yale University

January 2026 – Present

Garg Lab

Hybrid

- Designed a matched-capacity controlled study across **9 activation-function arms** testing whether learned rational activations beat standard nonlinearities (GELU) on sequence-to-expression regression
- Uncovered **two silent data-corruption bugs** in a 5.6M-region genomic dataset through systematic data-quality validation: **38% duplicate rows** and a silently failing peak caller discarding **48% of data**
- Orchestrated a fault-tolerant **315-run experiment grid** on preemptible A100 GPUs (SLURM), implementing signal-driven checkpointing that automatically requeues preempted jobs
- Trained a 14M-parameter deep learning model (1D CNN) to **0.680 test Pearson**, benchmarked it against a 6-parameter GC/CpG baseline, and applied per-base attribution (DeepLIFT, in-silico mutagenesis) to interpret predictions

Projects

regsearch — Hybrid Search over Genomics Literature | *Python, PostgreSQL, PyTorch, FastAPI*

August 2026

- Engineered a RAG-style hybrid retrieval stack over 19,791 genomics abstracts (99,567 passages), fusing SQL full-text search and pgvector embedding search via reciprocal rank fusion, served through a FastAPI REST API
- Created a graded-relevance evaluation harness — 704 pooled candidates from 40 queries, labeled 0–3 by two independent LLM-as-judge passes with hand adjudication of disagreements — proving **hybrid beats either mode alone (0.82 Recall@50, 0.85 nDCG@10)**
- Fine-tuned a cross-encoder reranker on citation-derived labels for **+41% MRR** over the off-the-shelf model, mining hard negatives from the system's own retrieval while excluding source papers that would mislabel 38% of negatives

remapd — Congressional Redistricting Engine | *Python, LangGraph, Claude API, FastAPI, Next.js*

April 2026

- **Finalist, Claude Hackathon (~50 teams):** built a full-stack redistricting engine (FastAPI REST API + Next.js) scoring congressional maps for all 50 states on population balance, compactness, and voting-rights proxies
- Implemented a simulated-annealing optimizer over county-to-district assignments in NumPy: **+27% median** objective vs. random baselines (18–42% across 5 states), **under 1s** for 700 iterations on 254 counties
- Architected an agentic workflow — a 5-node LangGraph pipeline with Claude tool use — converting metric vectors into streamed analytical memos, grounding every generated number in a verifiable US Census API call

Thermodynamic Sequence Simulator | *Python, NumPy, PyTorch*

- Authored a Python research package for thermodynamic sequence simulation, verifying analytic labels against brute-force state enumeration to 2e-16 numerical precision

Heterogeneity Score — Chromatin Variability from DNA Sequence | *PyTorch, SLURM, Genomics*

- Developed an end-to-end deep learning pipeline predicting chromatin heterogeneity (0–1) from DNA alone: 5.6M regions as 2048 bp one-hot windows, whole-chromosome train/val/test splits to prevent leakage, and an attention-pooled CNN
- Diagnosed why the task resists learning: 41% of labels saturate at the score's ceiling (intrinsic to the label definition, not a data bug), with context-crop sweeps showing predictive signal plateaus at 256 bp — 8× less sequence than the window provides

Technical Skills

Languages: Python, C++, Java, JavaScript, SQL

ML & LLMs: PyTorch, NumPy, Pandas, scikit-learn, deep learning, fine-tuning, RAG, embeddings & vector search, LLM-as-judge evals, LangGraph

Systems & Tools: PostgreSQL, pgvector, FastAPI (REST APIs), Next.js, SLURM/HPC/Shell Scripting, Git