

PROFESSIONAL EXPERIENCE

Data Engineer — Google (Contractor) — San Jose, CA— 2025

- **Agentic STEM Evaluation Systems:** Architected multi-agent pipeline (analyzer→planner→verifier) with reusable prompt engineering and verification strategies for scalable AI evaluation
 - **Auto-Hinting System:** Designed 3-stage domain-agnostic hint generation system using prompt engineering, improving STEM problem-solving accuracy by 20–40%
 - **Auto-Verification System:** Built multi-agent consensus framework with Google Search contradiction detection, reducing false positive rates by 80%
 - **Automated Evaluation Framework:** Formulated reasoning-error metrics and engineered MCP automation system, accelerating PhD researcher review cycles from days to under 1 hour
- **Progressive Visual Reasoning STEM Solver:** Developed educational multimodal agent integrating visual-textual chain-of-thought reasoning through Analyzer (multimodal perception + SVG parsing) → Planner (image augmentation via tool calling) → Verifier (iterative refinement)
- **Technical Leadership for RLHF:** Led 2–5 member teams to evaluate Gemini Pro reasoning patterns, structured evaluation workflows, and delivered actionable insights through technical reports that informed RLHF/RLAIF training for next-generation models

Sr ML Data Scientist — Protein Evolution, San Jose, CA (Remote) — 2024

- **NLP for Sequence Design:** Developed domain-specific BERT model using variable masking strategies to generate high-fitness protein sequences, accelerating design workflows and reducing iteration time for protein engineering applications tailored to specific queries
- **Data-Centric Classifiers:** Shifted from model-centric deep learning to data-centric workflows using ESM2 mutational embeddings with feature engineering and class balance optimization, achieving F1-scores above 0.75 (baseline 0.65) with explainable AI to guide protein library optimization in engineering workflows

Computational Protein / ML Engineer — Aether Biomachines — Menlo Park, CA — 2021–2023

- **Recommender Systems Development:** Designed and deployed chemistry-aware sequence recommender using substrate-to-enzyme mapping across public databases, reducing sequence retrieval time by 95% and accelerating project initiation for multiple customers
- **Generative ML for Protein Design:** Developed GPT-2 sequence generator with motif-based BPE tokenization, pre-trained on curated enzyme-reaction datasets (Rhea, BRENDA, RetroRules) to generate enzyme candidates from target chemical reactions for automated design workflows
- **ML-Driven Deployment Engineering:** Built production-ready end-to-end pipeline from sequence to 3D structure modeling with catalytic site prediction (AlphaFold), enabling reproducible integration across downstream protein engineering platforms

SKILLS

- **AI/ML Frameworks:** PyTorch, HuggingFace Transformers, TRL, Scikit-learn, OpenAI API, Gemini API, FastMCP
- **Programming & Data:** Python, R, SQL, Bash, Pandas, NumPy, Matplotlib/Seaborn, Neo4j (Graphs)
- **NLP & LLM:** GPT-2/BERT fine-tuning & pre-training, prompt engineering, BPE tokenization, multi-agent systems, preference data preparation for RLHF/RLAIF
- **Infrastructure & DevOps:** Docker, AWS (EC2, S3), Git, Linux, FastAPI, Conda
- **Specialized Domains:** Frontier AI model evaluation (STEM), deep reinforcement learning, recommender systems, MD simulations, protein engineering

EDUCATION

- **MS, Computer Science** — San Jose State University, CA 2018–2020
Thesis: “Multi-Agent Deep Reinforcement Learning for Walker Systems”
Focus: Deep reinforcement learning, multi-agent systems, AI
- **PhD, Chemical Physics** — The Ohio State University, OH 2006–2010
Dissertation: “Computational Simulations of Protein-Ligand Molecular Recognition via Enhanced Samplings, Free Energy Calculations and Applications to Structure-Based Drug Design”
Focus: Protein modeling, drug design, molecular dynamics simulation, free energy calculation
- **BS, Chemistry & Physics** — SookMyung Women’s University, Seoul, Korea