

Siddharth Sahay

github.com/sid2364 – sahaysid@gmail.com

www.linkedin.com/in/siddharth-sahay

Brussels, Belgium

SUMMARY

Bioinformatics PhD candidate at ULB, exploring **drug repurposing for rare diseases** using **multi-agent LLM systems**. Software engineering background with **8+ YoE** building distributed systems, data-intensive applications, and scalable backend platforms across **fintech, e-commerce, and HPC** environments.

EDUCATION

Université Libre de Bruxelles (ULB)

Brussels, Belgium

PhD Candidate – Machine Learning Group (MLG)

Sep 2026 – Present

- **Research area:** drug repurposing for heritable thoracic aortic diseases (HTADs) using multi-agent LLM systems
- Supervised by Prof. Dr. Tom Lenaerts

Université Libre de Bruxelles (ULB)

Brussels, Belgium

Master of Computer Science, Focus: AI & Data Science

Sep 2024 – Jun 2026

- Graduated with **High Distinction** (*Grande Distinction*)
- Received the **Babbage Award** (2026) from the Master in Computer Science jury for a remarkable study path

Manipal Institute of Technology

Manipal, India

Bachelor of Technology in Information Technology

Jul 2012 – May 2016

RESEARCH

Master's Thesis: Drug repurposing using knowledge graph embeddings

2026

- Designed an **end-to-end ML pipeline** combining **knowledge graph embeddings** (TransE, DistMult, ComplEx) as feature extractors with downstream classifiers (SVM, MLP) to predict drug-disease associations on a **90K-node biomedical knowledge graph**
- **Corrected a data leakage flaw present in prior literature** by removing all treatment edges before embedding training, preventing the model from indirectly memorising the prediction target
- Proposed a **novel degree-matched negative sampling strategy** with metapath-based filtering to eliminate node-degree as a trivial classification shortcut, and built an **embedding-space stratified cross-validation** scheme using K-Means to produce leakage-free generalization estimates
- Applied **split conformal prediction** to convert classifier scores into **distribution-free confidence sets**, replacing ad hoc threshold tuning with statistically grounded uncertainty quantification
- Screened **~23,000 drugs** against three target diseases, producing **literature-supported drug candidates** via cross-model consensus voting to eliminate blind-spots

Hybrid data mining algorithm in cloud computing using MapReduce framework

2016

- Proposed a **hybrid variant of the Apriori algorithm** optimized for distributed execution on a **Hadoop MapReduce** cluster, improving frequent itemset mining performance over standard implementations on large-scale datasets, (DOI: 10.1109/ICACCCT.2016.7831691)

WORK EXPERIENCE

Nosto Solutions

Helsinki, Finland

Software Engineer Team Lead/Technical Product Manager – Client Technology

Mar 2022 – Sep 2024

- **Led team of 4 engineers** developing headless commerce solution using React, TypeScript, and Remix framework
- Architected and launched **open-source React component library** integrated with Shopify Hydrogen, serving **50M+ monthly users** across Shopify Markets
- Designed and implemented **scalable backend services and APIs** supporting personalisation and recommendation workflows serving global e-commerce platforms

Goldman Sachs – Global Markets Division

Bangalore, India

Vice President – FICC Systematic Market Making Strategist

Jun 2018 – Feb 2022

- Developed **low-latency trading and post-trade processing systems** handling high-volume fixed income market data and transaction workflows

- Built **real-time high-volume trade reconciliation system** using Apache Kafka, processing **100K+ trades daily**
- Enhanced Java-based trade message routing infrastructure, **reducing message processing latency by more than 40%**

Altair Engineering

Software Development Engineer

Bangalore, India

Jan 2016 – May 2018

- Maintained and enhanced **PBS Pro**, an **open-source HPC workload manager** serving enterprise computing clusters
- Developed **DRMAA2 API prototype library** for distributed resource management across grid computing systems

SELECTED PROJECTS

Distributed RAG Exam Generator

2025

- Built a **distributed Retrieval-Augmented Generation (RAG) pipeline** using **Apache Spark** to process, chunk, embed, and retrieve information from large collections of course documents
- Compared retrieval quality across **Word2Vec** and **SBERT** embedding spaces, multiple similarity metrics, and **Locality Sensitive Hashing (LSH)** for scalable approximate nearest-neighbor search
- Integrated a **locally hosted LLM** to generate context-aware practice exam questions from retrieved course content, creating an end-to-end educational AI application

Hand Pose Estimation from EMG signal

2025

- Used deep learning techniques to reproduce hand and finger positions/angles from **live forearm EMG signals**, enabling **real-time decoding** of human hand movements for prosthetic control and HCI
- Achieved an **NRMSE of ~0.09**, and landed among the **top-5 on the class leaderboard**
- Engineered **EMG signal processing and feature extraction pipelines** using statistical, energy-based, and covariance-matrix representations of multichannel time-series data
- Designed and evaluated **PyTorch CNN/CNN-LSTM** architectures alongside **ensemble methods** (XGBoost, Random Forests, bagging, averaging, and stacking) for robust pose estimation

Chess Training Bot

2026

- Built a **full-stack chess training platform** using **Flask** and the **Lichess API**, enabling users to practice personalized opening repertoires against an **adaptive chess bot**
- Implemented **real-time game synchronization** and local state management for live Lichess games, adjusting opening selection and playing strength based on user preferences and rating level
- Developed an opening knowledge base by crawling and processing Lichess Opening Explorer data, integrating **Stockfish** for position evaluation and move generation

TECHNICAL SKILLS

Programming: Python, Java, JavaScript/TypeScript, SQL

Machine Learning: PyTorch, TensorFlow, Scikit-Learn, Pandas, HuggingFace, MLflow, PyKEEN, Knowledge Graph Embeddings, Conformal Prediction

LLMs & Agents: Multi-agent systems, ReAct agents, RAG, LangChain, llama.cpp, vLLM, MCP

Data & Databases: Apache Spark, Kafka, Airflow, PostgreSQL, Neo4j, MongoDB, Elasticsearch, Apache Solr, LanceDB

Infrastructure & HPC: SLURM, Docker, Kubernetes, AWS (EC2, S3, Lambda), RabbitMQ, ELK stack, Git

Web & Backend: React, Node.js, Remix, Flask, Spring Boot, GraphQL, REST APIs, Microservices, Event-Driven Architecture

ADDITIONAL INFORMATION

Certifications: Engineering Management (TheFrontiers), Product Management (Indian School of Business, Hyderabad)

Languages: English (Native), French (Limited Working), Dutch (Limited Working)