Haonan(Eric) Gao

Email: eric.gao@yale.edu
Phone: (+1) 203-675-5865
Github | Google Scholar | LinkedIn

EDUCATION

Yale University **2024.09 - 2026.05**
Master of Science in Biostatistics (Data Science track)
University of Toronto **2019.09 - 2024.05**
Honor Bachelor of Science (with high distinction)

- Major in **Computer Science** and **Statistics**; Minor in **Mathematics**; Cumulative GPA: 3.9/4.0

RESEARCH EXPERIENCES

Research Assistant | Zhao Lab | Yale School of Public Health | Yale University **2025.05 - 2026.05**

Orthogonal Sparse Subgraph Alignment for SC–FC Coupling (NeurIPS 2026 under review)

- Proposed Orthogonal Sparse Subgraph Alignment (OSSA), the first principled statistical framework for localizing cross-modal subnetworks in structure–function (SC–FC) coupling. Within each functional community, OSSA jointly learns an orthogonal transformation that absorbs SC–FC coordinate mismatch and a sparse node-selection vector that identifies compact structural cliques maximally concordant with functional connectivity across subjects.
- Generalized the classical orthogonal Procrustes problem to an unknown, induced-subgraph target. Reformulated the NP-hard combinatorial objective as a continuous relaxation over the capped simplex and solved it via an efficient projected block-coordinate ascent algorithm, with an overlap-regularized K-support extension for recovering multiple cliques.
- Established the statistical consistency of the OSSA estimator, proving that the relaxed solution recovers the true population-level SC clique and consistently estimates the aligned FC subspace under mild regularity conditions.
- Validated OSSA on extensive simulations and on multi-modal connectomes from the Adolescent Brain Cognitive Development (ABCD) study, outperforming nine size-matched and deep-learning baselines (e.g., BrainGNN, BrainNetCNN) on out-of-sample SC–FC alignment, and recovering compact single-core structural modules that remained highly reproducible across the Year-2 longitudinal follow-up.

Disentangling Latent Risk Pathways via Bayesian Hypergraph Inference (ICML 2026 Spotlight and Oral paper)

- Developed Bayesian Hypergraph Pathway Inference (BHPI), a probabilistic framework that reframes large-scale multi-disease EHR modeling as the discovery of latent, risk-factor-modulated disease pathways, representing pathways as hyperedges in a latent hypergraph where disease can participate in multiple overlapping, higher-order risk groups beyond pairwise associations.
- Designed a repulsion prior over hyperedges to enforce parsimonious and identifiable pathway structure, mitigating latent-structure collapse in the long-tail regime where rare diseases provide weak signal.
- Derived a scalable structured variational inference algorithm that preserves the logical hierarchy among hyperedge existence, disease membership, and effect, overcoming the miscalibration of standard mean-field approximations while scaling linearly in cohort size and latent dimension.
- Validated BHPI on simulations and the UK Biobank, where it delivered well-calibrated uncertainty, competitive overall prediction, and significantly improved risk estimation for rare diseases, while recovering interpretable risk-factor–specific pathways.

Research Assistant | Yale School of Management, Yale University **2024.09 - Present**

- Architected an end-to-end LLM–RAG pipeline that turns 2TB+ of heterogeneous financial text (SEC 10-K filings, USPTO patents, earnings-call transcripts, WSJ articles) into structured economic estimates, normalizing sources into a unified schema and embedding passages into 1024-dim vectors in a multi-namespace vector index (Pinecone).
- Designed a multi-stage semantic retrieval system with progressive query expansion (pseudo-relevance feedback), cosine-similarity nearest-neighbor search, metadata filtering, and strict no-look-ahead temporal constraints, plus reranking and fuzzy deduplication for high-precision evidence bundles.
- Built the downstream inference stack, faithful context compression, constrained JSON generation, and high-throughput batched serving (vLLM) with optional QLoRA calibration, improving project-identification accuracy by 23% over a keyword-retrieval baseline.

Research Assistant | Lee Lab | Department of Computer Science | University of Toronto **2022.09 - 2024.03**

OasisSimp: Asian–English Sentence Simplification (LREC 2026, Oral)

- Built OasisSimp, an open-source sentence-simplification dataset for 5 languages (English, Sinhala, Tamil, Thai, Pashto), with expert human simplifications.
- Curated complex-sentence pools and produced parallel simplifications under detailed annotation guidelines preserving meaning, fluency, and grammaticality.
- Benchmarked eight open-weight multilingual LLMs in zero-/few-shot settings using SARI and BERTScore, revealing low-resource simplification gaps.

MultiMWP: Multilingual Math Word Problem Generation (IEEE TASLP 2025)

- Built MultiMWP, the largest multi-way parallel math-word-problem corpus to date, spanning 9 languages (6 low-resource) across arithmetic and algebraic domains.
- Curated and quality-controlled multi-way parallel translations with automatic (embedding-similarity) and manual quality estimation. Benchmarked pretrained multilingual seq2seq models (mBART50, mT5, M2M-100, IndicBART) for autoregressive generation, integrating a math-constraint module.

SIB-200: A Topic-Classification Benchmark for 200+ Languages (EACL 2024)

- Built SIB-200, a topic-classification benchmark covering 205 languages and dialects (first NLU evaluation set for many of them), built on the Flores-200 parallel corpus.
- Annotated and adjudicated English sentence-level labels, then propagated annotations across 204 languages with inter-annotator agreement checks. Benchmarked multilingual PLMs under supervised, cross-lingual transfer, and LLM-prompting settings, quantifying the high- vs. low-resource performance gap.

Research Assistant | Department of Computer Science | Carnegie Mellon University **2021.09 - 2022.03**

Diagnostic Question Analysis Model

- Implemented a modified Item Response Theory (IRT) model to analyze extensive data derived from student interactions with diagnostic questions, thereby facilitating a precise assessment of individual learning statuses and automating curriculum recommendations.
- Compared to the baseline model, the fine-tuned IRT model shows a better performance after 40 iterations, and with approximately 5% increase in accuracy after 100 iterations.

WORK EXPERIENCE

Machine Learning Engineer | DeepManifold Inc., New York, NY | Part-time Intern **2025.09 - 2025.12**

- Architected and deployed a customer embedding platform leveraging sequence models (Word2Vec, Transformer) and contrastive self-supervised learning (SimCLR, MoCo), producing dense representations for 50M+ users with weekly updates.
- Designed and operationalized segmentation workflows (KMeans, GMM, DBSCAN, extended RFM), uncovering high-value and churn-risk cohorts; enabled personalized campaigns that increased CTR by 14% and reduced churn by 9% in controlled experiments.
- Integrated embeddings into recommendation and marketing pipelines, driving +11% CTR uplift for cold-start users and a 17% ROI improvement across multi-channel campaigns.

Cloud Data Engineer | Huawei Technologies Canada, Toronto | Full-time **2022.04 - 2023.05**

- Participated in developing a distributed, low-latency, reliable data engine and developed a connector for the time-series Database InfluxDB based on the query engine, to analyze and migrate across data centers and hybrid clouds with over 40 different Databases.
- The engine is now deployed and used in the top 5 banks in China with over 3 million users, the InfluxDB connector provided efficient usage for operations monitoring, IoT sensor data and real-time analytics, such as stocks.
- Carried out a solution to push down SQL query's predicates under cloud cluster environment, the overall improvements that ultimately increased the processing speed of TPC-H 1000GB benchmark queries by an average of 20%.

PUBLICATIONS

[1] **Haonan Gao**, Daeyoung Ham, Yifei Zhang, Xinyuan Tian, Shengxian Ding, Yize Zhao, Tianxi Li. "Orthogonal Sparse Subgraph Alignment for Structure-Function Coupling in Brain Networks." *Under Review*, the 40th Conference on Neural Information Processing Systems (NeurIPS 2026).
[2] Shengxian Ding, **Haonan Gao**, Pangpang Liu, Xinyuan Tian, Yize Zhao. "Disentangling Latent Risk Pathways via Bayesian Hypergraph Inference." Proceedings of the 43rd International Conference on Machine Learning (ICML 2026). Spotlight & Oral presentation.
[3] Hannah Liu, Muxin Tian, Iqra Ali, **Haonan Gao**, Qiaoyiwen Wu, Blair Yang, Uthayasanker Thayasivam, Annie En-Shiun Lee, Pakawat Nakwijit, Surangika Ranathunga, Ravi Shekhar. "OasisSimp: An Open-source Asian-English Sentence Simplification Dataset." Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026). Oral presentation.
[4] David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, **Haonan Gao**, Annie En-Shiun Lee. "SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects." Proceedings of the 18th Conference of the European Chapter of the ACL (EACL 2024), pp. 226–245.
[5] Gamage O. Ishendra, Surangika Ranathunga, Annie En-Shiun Lee, Mehreen Alam, **Haonan Gao**, et al. "A Multilingual Dataset (MultiMWP) and Benchmark for Math Word Problem Generation." Proceedings of IEEE Transactions on Audio, Speech and Language Processing, VOL. 33, 2025, pp. 1838–1848.
[6] **Gao**, **Haonan**. "A Diagnostic Question Analysis Model based on a Modified Item Response Theory." Proceedings of the 6th International Seminar on Education, Management and Social Sciences (ISEMSS 2022). Atlantis Press, 2022.

PROFESSIONAL ACTIVITY

Peer Review

- The 40th Conference on Neural Information Processing Systems (NeurIPS 2026)
- The 43rd International Conference on Machine Learning (ICML 2026)

Presentation/Talk

- Thesis presentation, Yale University, April 2026

HONORS / AWARDS

- **Dean's List Scholar Awards** **2019 - 2024**
  University College, University of Toronto
- **Dr. James A. & Connie P. Dickson Scholarship In Science & Mathematics** **2020.10**
  University College, University of Toronto, Award amount: 500 CAD
- **Covid-19 Student Bursary Scholar** **2020.05**
  The University of Toronto Scholars Program, University of Toronto, Award amount: 2,000 CAD
- **J. S. Mclean Scholarship** **2019.10**
  University College, University of Toronto, Award amount: 5,000 CAD
- **University Of Toronto Scholar** **2019.09**
  The University of Toronto Scholars Program, University of Toronto, Award amount: 7,500 CAD