Joseph Lee

jxlee.me · jxlee@cs.washington.edu · Google Scholar · GitHub

Research Interests

Generative models for biology · Agents for scientific discovery

Current Position

2025 – **Ph.D. Student, Computer Science**
University of Washington, Allen School
Advisors: Pang Wei Koh, Hanna Hajishirzi

- Developing native multi-modal architectures for text and molecular structure, with the Baker Lab.

Education

2023 – 2024 **M.Sc. Statistical Science**
University of Oxford
Thesis: *Learning to learn in-context: insights and algorithms*
Advisor: Yee Whye Teh · Grade: Distinction

2018 – 2022 **B.Sc. (Hons.) Computational Biology**
Minor in Mathematics, Minor in Philosophy
National University of Singapore
Thesis: *Modeling degradation bias for accurate transcript quantification from long-read RNA-seq*
Advisor: Jonathan Göke · Grade: Honours (Highest Distinction)

Experience

2026 **Research Intern**
Allen Institute for AI (Ai2)
Advisors: Rory Donovan-Maiye, Allyson Ettinger, Noah A. Smith
In collaboration with Yanjing Li, Bo Qiang, David Baker

- Built `Prompteina`, the first all-atom protein structure model conditioned on natural language.
- Constructed `Prompteina-Data`: 1.8M structures paired to 50M+ captions.
- Investigated RL post-training of flow-based structure generation (Flow-GRPO, AWM).
- Outperformed baselines on 15 of 18 structural properties and 13 of 20 binder targets.
- Validated designs experimentally: nine de novo GM2A binders, best $K_D \approx 107$ nM.

2024 – 2025 **Research Assistant**
University of Oxford, Computational Statistics and Machine Learning Group
Advisor: Yee Whye Teh

- Steered prompt robustness at inference time through activation edits alone, with no weight updates.
- Built model diffing methods for mechanistic comparison of related LLMs.

2024 **M.Sc. Dissertation Student**
University of Oxford, Computational Statistics and Machine Learning Group
Advisor: Yee Whye Teh

- Derived a task-agnostic steering vector for in-context learning, with no gradient updates.
- Matched meta-training on 50× less data; outperformed it under distribution shift.
- Showed via semantic interpolation that meta-training smooths decision boundaries.

2022 – 2023 **Bioinformatics Specialist**
Genome Institute of Singapore, Laboratory of Systems Biology and Data Analytics
Advisors: Jagadish Sankaran, Shyam Prabhakar
Invariant representations of single-cell morphology via signal processing and graph-based methods.

2021 **Research Intern**
Genome Institute of Singapore, Laboratory of Systems Biology and Data Analytics
Advisors: Vipul Singhal, Shyam Prabhakar
Method development for BANKSY, a spatially-aware cell-typing method.

2020 – 2022 **Research Intern and B.Sc. Thesis Student**
Genome Institute of Singapore, Laboratory of Computational Transcriptomics
Advisors: Chen Ying, Andre Sim, Jonathan Göke
Degradation-aware isoform quantification; analysis and software for LRGASP, `bambu`, `proActiv`.

2019 – 2021 **Research Assistant**
National University of Singapore, Computational and Statistical Systems Biology Lab
Advisor: Hyungwon Choi
Gene-essentiality classification for CRISPR screens; `eGFRsmooth` for clinical trajectory inference.

Selected Publications & Software

2026 Designing protein structures from natural language
Joseph Lee, Yanjing Li, Bo Qiang, Gregory Lau, Allyson Ettinger, Noah A. Smith, Rory Donovan-Maiye, David Baker, Hanna Hajishirzi, and Pang Wei Koh. *arXiv*:2610.13917

2026 Gacha decoding: eliciting diverse generations through instruction following
Scott Geng, Yufei Zhang, **Joseph Lee**, Jerry Li, Marjan Ghazvininejad, and Pang Wei Koh. *arXiv*:2610.01382

2026 StochasTok: Improving fine-grained subword understanding in LLMs
Anya Sims, Thom Foster, Klara Kaleb, Tuan-Duy H. Nguyen, **Joseph Lee**, Jakob N. Foerster, Yee Whye Teh, and Cong Lu. *ICLR* 2026

2024 BANKSY unifies cell-typing and tissue domain segmentation for spatial genomics analysis
Vipul Singhal, Nigel Chou, **Joseph Lee**, et al. *Nature Genetics* 56:431–441

2024 Systematic assessment of long-read RNA-seq methods for transcript identification and quantification
LRGASP consortium. *Nature Methods* 21:1349–1363

2023 Context-aware transcript quantification from long-read RNA-seq data with Bambu
Ying Chen, Andre Sim, Yuk Kei Wan, Keith Yeo, **Joseph Lee**, et al. *Nature Methods* 20:1187–1195

Software

2026 `prompteina` 73⋆ — text-conditioned protein structure generation.
2024 `Banksy` 171⋆ — spatially-aware cell typing. *Bioconductor*.
2020 `proActiv` 61⋆ — promoter activity estimation. *Bioconductor*.

Awards & Scholarships

2026 Google.org Impact Challenge: AI for Science. USD 3,000,000, 1 of 17 internationally. Computational lead; co-wrote the funded proposal on protein design from natural language.
2026 – 2027 Young NUS Fellow. SGD 10,000, NUS Development Grant.
2025 Open Philanthropy Technical AI Safety Grant. USD 20,000, language model behaviour.
2024 Graduate Research Award. Brasenose College, University of Oxford.
2022 Lijen Industrial Development Medal. Best Honours thesis, NUS.
2021 A*STAR Research Internship Award. Spatial transcriptomics method development.
2020 Lim Soo Peng Book Prize. Best student, Computer Science, NUS Faculty of Science.
2018 NUS Global Merit Scholarship. Full undergraduate scholarship.