A047 · LINGO: A Knowledge Graph-Grounded Foundation Model for Single-Cell Lineage Inference

Presenter: Kaifu Chen — Boston Children's Hospital

Authors:

Session: Day 1 · Thu Oct 1, 2026 · 2:15–4:15 PM

Single-cell lineage tracing (scLT) enables reconstruction of developmental histories by coupling lineage barcodes with transcriptomic profiling. The rapid accumulation of scLT datasets across diverse species, tissues, and experimental systems has created unprecedented opportunities to uncover general principles of cell fate determination. However, lineage inference remains challenging because lineage information is often incomplete owing to barcode dropout, barcode silencing, and destructive sampling, while existing computational methods are typically tailored to individual datasets and transfer poorly across biological contexts. Meanwhile, current single-cell foundation models are largely transcriptome-driven and do not explicitly incorporate lineage-relevant biological knowledge. Here we present LINGO, a lineage foundation model that integrates a lineage-informed knowledge graph with large-scale scLT pretraining for single-cell lineage inference. LINGO combines developmental hierarchies, gene regulatory interactions, pathway information, and lineage-relevant cellular relationships within a unified graph framework and uses graph-based contrastive pretraining to learn lineage-aware representations. We pretrained LINGO on more than 2 million cells from lineage-tracing datasets spanning 18 tissue types across mouse, human, and zebrafish, and evaluated it on held-out benchmarks encompassing cell linkage prediction, lineage origin prediction, and clonal plasticity prediction across diverse lineage-tracing technologies, tissues, and species. LINGO consistently outperformed state-of-the-art lineage-analysis methods and transcriptome-based foundation models, achieving gains of up to 26 percentage points on held-out benchmarks. Furthermore, pretrained representations remained highly informative even when model weights were frozen during down- stream fine-tuning, demonstrating the robustness and transferability of the learned representations. Together, these results demonstrate the value of inte- grating structured biological knowledge into large-scale representation learning and establish a foundation-model framework for single-cell lineage inference.

LINGO: A Knowledge Graph-Grounded Foundation Model for Single-Cell Lineage Inference

Ziyang Liu[1,2], Jiatong Liang[1,2,3], Rongbin Zheng[1,2], Yunxia Wang[1,2], De-an Li[1,2], Kulandaisamy Arulsamy[1,2], Krishan Gupta[1,2], Gaopeng Xian[1,2], Lili Zhang[1,2], William T. Pu[1,2*], Kaifu Chen[1,2*]

[1]Department of Cardiology, Boston Children's Hospital, 300 Longwood Avenue, Boston, 02115, MA, United States.
[2]Department of Pediatrics, Harvard Medical School, 300 Longwood Avenue, Boston, 02115, MA, United States.
[3]Department of Statistics, University of Michigan, 1085 S University Ave, Ann Arbor, 48109, MI, United States.

*Corresponding author(s). E-mail(s):
William.Pu@cardio.chboston.org;
Kaifu.Chen@childrens.harvard.edu;

Abstract

Single-cell lineage tracing (scLT) enables reconstruction of developmental his- tories by coupling lineage barcodes with transcriptomic profiling. The rapid accumulation of scLT datasets across diverse species, tissues, and experimental systems has created unprecedented opportunities to uncover general princi- ples of cell fate determination. However, lineage inference remains challenging because lineage information is often incomplete owing to barcode dropout, bar- code silencing, and destructive sampling, while existing computational methods are typically tailored to individual datasets and transfer poorly across bio- logical contexts. Meanwhile, current single-cell foundation models are largely transcriptome-driven and do not explicitly incorporate lineage-relevant biological knowledge. Here we present LINGO, a lineage foundation model that inte- grates a lineage-informed knowledge graph with large-scale scLT pretraining for single-cell lineage inference. LINGO combines developmental hierarchies, gene regulatory interactions, pathway information, and lineage-relevant cellular rela- tionships within a unified graph framework and uses graph-based contrastive pretraining to learn lineage-aware representations. We pretrained LINGO on more than 2 million cells from lineage-tracing datasets spanning 18 tissue types across mouse, human, and zebrafish, and evaluated it on held-out benchmarks encompassing cell linkage prediction, lineage origin prediction, and clonal plas- ticity prediction across diverse lineage-tracing technologies, tissues, and species. LINGO consistently outperformed state-of-the-art lineage-analysis methods and transcriptome-based foundation models, achieving gains of up to 26 percent- age points on held-out benchmarks. Furthermore, pretrained representations remained highly informative even when model weights were frozen during down- stream fine-tuning, demonstrating the robustness and transferability of the learned representations. Together, these results demonstrate the value of inte- grating structured biological knowledge into large-scale representation learning and establish a foundation-model framework for single-cell lineage inference.