Sijie Chen

✉ chensj16@stanford.edu | LinkedIn | Google Scholar | GitHub | Website

Research Statement

I build machine-learning systems for biological discovery, focused on **representation learning**, **generative latent-variable modeling**, and **cross-domain alignment**. Drawing on discrete *optimal transport*, *variational inference*, and *geometric / equivariant deep learning*, I design models that turn high-dimensional, irregularly-structured biological signals into representations that modern architectures can exploit—reorganising gene expression as 2D feature images via Gromov–Wasserstein layouts, embedding cells with hierarchical conditional VAEs, and modelling morphogenetic dynamics with equivariant neural ODEs. I am excited to bring these methods to ML research and engineering in industry and to AI-for-medicine settings (diagnosis, treatment planning, image reconstruction, outcome prediction), where data is high-dimensional, irregular, and biologically structured.

Education

Tsinghua University Beijing, China
Ph.D., Department of Automation (Control Science & Engineering) — Advisors: Xuegong Zhang & Michael S. Waterman *09/2016 – 06/2023*

Chongqing University Chongqing, China
B.Eng., School of Big Data and Software Engineering (Outstanding Graduate) *09/2012 – 06/2016*

Work & Research Experience

I build models from high-dimensional, structurally heterogeneous biological signals (single-cell transcriptomes, spatial omics, ATAC, imaging phenotypes) to guide scientific discovery and clinical translation. My work spans three methodological threads:

- **Representation learning with spatial / structural inductive bias**: Gromov–Wasserstein gene-to-pixel layouts, image-native encoders for tabular biology, contrastive cross-species alignment.
- **Generative latent-variable models**: VAE families, hierarchical (MrVI-style) conditional models, foundation-model pre-training, multi-task generative pre-training for cells.
- **Spatio-temporal & equivariant dynamics**: equivariant transformer-based neural ODEs, optimal-transport-driven trajectory inference, in-silico perturbation.

Stanford University Stanford, CA
Postdoctoral Scholar *05/2024 – present, Full-time*

- **TransMap** *(NeurIPS 2026 submission, first author)*. Designed a discrete optimal-transport framework (Gromov–Wasserstein, Fused-GW with block-coordinate descent, SNN-geodesic distance refinement, injective rounding via Jonker–Volgenant assignment) that re-represents scRNA-seq cells as 2D feature images, letting CNN encoders exploit gene–gene co-expression as image locality. On five multi-organ benchmarks (up to 714k cells / 167 donors), a $\sim$15M-parameter CNN beats parameter-matched MLPs on scIB and scGraph metrics; a shared 5-species pixel grid aligns human–mouse pancreas cells (iLISI = 1.297; 76.4% label transfer) *without* ortholog conversion. Implemented as a complete training framework (PyTorch, distributed training, mixed precision, SLURM-scale runs).
- **Dynode** *(first author, in preparation)*. Built an SE(3)-equivariant transformer-based neural-dynamics model that learns 3D organogenesis trajectories from spatiotemporally resolved single-cell data using optimal-transport-based regularisation; supports *in silico* perturbation for congenital heart disease.
- **Chromatin regulatory dynamics**. Used the dynamics framework above plus single-cell RNA-seq to study how SMARCB1 (BAF complex) and JARID2 (PRC2) coordinate chromatin regulation to control human pluripotent stem-cell self-renewal and recovery of pluripotency.
- **Antibody design with immune epigenomics**. Used ATAC-seq as a functional readout to compare

antibody design variables on receptor-cell chromatin and regulatory networks; fed findings back into the molecular-engineering loop.

- Bridged modelling and wet-lab collaborators, translating high-dimensional signals into ML problem formulations, benchmarks, and analyses that drove assay iteration and imaging-adjacent phenotypes.

Tsinghua University Beijing, China
Research Assistant *07/2023 – 04/2024, Full-time*

Tsinghua University Beijing, China
Ph.D. Student *09/2016 – 06/2023, Full-time*

- Ph.D. thesis: *Constructing cell-state transition landscapes from single-cell transcriptomes* — recovering short- and long-term transitions through (1) large referential atlas assembly, (2) cross-batch transition integration, and (3) discovery of gene-regulatory modules along the trajectory.
- Co-led the development of **hECA** , an infrastructure for cell-centric assembly of cell atlases. Curated 1.09M hierarchically annotated healthy cells from 116 datasets across 38 human organs; supported cloud retrieval so that scattered datasets behave as one.
- Developed **VeloFusion**, a velocity-informed scRNA-seq integration algorithm grounded in multi-task graph generative models, linking cell transitions across datasets via RNA velocity (RECOMB 2024 abstract).
- Contributed to **scMulan** , a 368M-parameter multitask generative pre-trained model for single-cell analysis; built latent-variable models (joint-VAE universal coordinate systems supporting cross-platform and cross-modality alignment).
- Designed a novel repeat-detection statistic D_2^R under Prof. Michael S. Waterman; the linear-time algorithm detects repetitive sequences including candidate CRISPR regions in bacterial and metagenomic data.
- Co-led *Dynamic Regulation of Innate Lymphoid Cell Development During Ontogeny* (Prof. Li Wu); built the ILC differentiation vector field, developed the temporal-spatial-consensus (TSC) gene-module method, and identified the Notch–GATA3–RORγt axis orchestrating ILC subset determination.
- Worked across diverse organ systems (musculoskeletal, intestinal, respiratory, pancreatic) and pathologies (T2D, CTLA-4 irAE, PAH, asthma, chronic colitis).

SELECTED PUBLICATIONS AND MANUSCRIPTS

**equal contribution. Full list and citation counts at Google Scholar.*

[1] **Sijie Chen**. TransMap: image-native representations for single-cell genomics. *NeurIPS* (under review), 2026.

[2] **Sijie Chen***, Yanting Luo*, Haoxiang Gao*, Fanhong Li*, Yixin Chen*, Jiaqi Li*, Renke You, Minsheng Hao, Haiyang Bian, Xi Xi, et al. hECA: the cell-centric assembly of a cell atlas. ***iScience***, 25(5), 2022.

[3] Tao Wu*, **Sijie Chen***, Jie Ma, Xinyi Zhu, Maocai Luo, Yuanhao Wang, Yujie Tian, et al., Xuegong Zhang, Yunping Zhu, Li Wu. Dynamic regulation of innate lymphoid cell development during ontogeny. ***Mucosal Immunology***, 2023.

[4] Haiyang Bian, Yixin Chen, Xiaomin Dong, Chen Li, Minsheng Hao, **Sijie Chen**, Jinyi Hu, Maosong Sun, Lei Wei, Xuegong Zhang. scMulan: a multitask generative pre-trained language model for single-cell analysis. ***RECOMB / ISMB***, 2024.

[5] Xue Bai, **Sijie Chen**, Xinxin Chi, Bowen Xie, Xinyi Guo, Han Feng, Peng Wei, Di Zhang, Shan Xie, Mengting Gou, et al., Chen Dong. Reciprocal regulation of T follicular helper cells and dendritic cells drives chronic colitis development. ***Nature Immunology***, 2024.

[6] Weili Fu*, **Sijie Chen***, Runze Yang, Chen Li, Haoxiang Gao, Jian Li, Xuegong Zhang. Cellular features of localized microenvironments in human meniscal degeneration: a single-cell transcriptomic study. ***eLife***, 2022.

[7] Yifan Zhou, Yusra B Medik, Bhakti Patel, Daniel B Zamler, **Sijie Chen**, Thomas Chapman, et al. Intestinal toxicity to CTLA-4 blockade driven by IL-6 and myeloid infiltration. ***Journal of Experimental Medicine***, 220(2), 2022.

[8] **Sijie Chen**, Yixin Chen, Fengzhu Sun, Michael S. Waterman, Xuegong Zhang. A new statistic for efficient detection of repetitive sequences. ***Bioinformatics***, 35(22):4596–4606, 2019.

[9] Xiangyue Zhang, Christopher S. McGinnis, Guotao Yu, **Sijie Chen**, . . . , Edgar G. Engleman. Erythropoietin receptor on cDC1s dictates immune tolerance. ***Nature*** (accepted), 2025.

[10] Qing-Yu Wang, Wei Zhang, Yue Zhao, Hui-Lian Chen, Qian Liu, Zi-Hui Wang, Lv-Tao Zeng, Jin Li, **Sijie Chen**, Lei Wei, Tomoo Iwakuma, Jian-Ping Cai. Colonic L-cell impairment in aged subjects with type 2 diabetes leads to diminished GLP-1 production. ***Diabetes & Metabolic Syndrome***, 2023.

[11] X. Rosa Ma, Stephanie D. Conley, . . . , **Sijie Chen**, . . . , Jesse M. Engreitz. Molecular convergence of risk variants for congenital heart defects leveraging a regulatory map of the human fetal heart. ***Nature*** (2nd-round review), 2024.

[12] The MorPhiC Consortium. MorPhiC Consortium: towards functional characterization of all human genes. ***Nature***, 2025.

[13] Jiachen Liu, Jiaoyan Lv, **Sijie Chen**, Jie Ma, Jiaying Yao, Zan Yuan, Chunyuan Yang, Xiao Li, Xuegong Zhang, Jianhong Zhang, Li Wu. Single-cell profiling highlights pathogenic myeloid features in house-dust-mite-induced murine airway inflammation. ***Allergy***, 2024.

[14] Chen Li, **Sijie Chen**, Yixin Chen, Haiyang Bian, Minsheng Hao, Lei Wei, Xuegong Zhang. TFcomb identifies transcription factor combinations for cellular reprogramming based on single-cell multi-omics data. ***Genome Research***, 2025.

[15] **Sijie Chen**, Lei Wei, Xuegong Zhang. Velocity-informed integration of single-cell RNA sequencing data (abstract). *RECOMB*, 2024.

[16] Minsheng Hao, Erpai Luo, Yixin Chen, Yanhong Wu, Chen Li, **Sijie Chen**, Haoxiang Gao, Haiyang Bian, Jin Gu, Lei Wei, Xuegong Zhang. STEM enables mapping of single-cell and spatial transcriptomics data with transfer learning. ***Communications Biology***, 2024.

[17] Jiaqi Li, Zhenbin Qiu, Chao Zhang, **Sijie Chen**, Mengmin Wang, Qiuchen Meng, Haiming Lu, Lei Wei, Hairong Lv, Wenzhao Zhong, et al. ITHscore: comprehensive quantification of intra-tumour heterogeneity in NSCLC by multi-scale radiomic features. ***European Radiology***, 33(2):893–903, 2023.

[18] Haoxiang Gao, Kui Hua, Lei Wei, Xinze Wu, **Sijie Chen**, Qijin Yin, Rui Jiang, Xuegong Zhang. Building a learnable universal coordinate system for single-cell atlas with a joint-VAE model. ***Communications Biology***, 2023.

[19] **Sijie Chen***, Yanting Luo*, Haoxiang Gao*, Fanhong Li*, Jiaqi Li*, Yixin Chen, Renke You, Hairong Lv, Kui Hua, Rui Jiang, et al. Toward a unified information framework for cell atlas assembly. ***National Science Review***, 9(3):nwab179, 2022.

[20] Yixin Chen*, **Sijie Chen***, Xuegong Zhang. Using DenseFly algorithm for cell searching on massive scRNA-seq datasets. ***BMC Genomics***, 21:1–9, 2020.

[21] Najeebullah Shah, Jiaqi Li, Fanhong Li, Wenchang Chen, Haoxiang Gao, **Sijie Chen**, Kui Hua, Xuegong Zhang. An experiment on ab initio discovery of biological knowledge from scRNA-seq data using machine learning. ***Patterns***, 1(5), 2020.

[22] Jiaqi Li, Haiming Lu, Xiang Fang, **Sijie Chen**, Xuegong Zhang. Pixel-level clustering reveals intra-tumour heterogeneity in non-small-cell lung cancer. *IEEE BIBM*, 2019.

[23] Shansong Liu, Kui Hua, **Sijie Chen**, Xuegong Zhang. Comprehensive simulation of metagenomic sequencing data with non-uniform sampling distribution. ***Quantitative Biology***, 6:175–185, 2018.

Teaching Experience

Machine Learning (graduate course) Tsinghua University
Teaching Assistant *Fall 2018, 2019, 2020*

Data Structure (undergraduate course) Tsinghua University
Teaching Assistant *Fall 2017*

Awards

- **Outstanding Graduate**, Chongqing University, 2016.
- **National Scholarship**, Ministry of Education of China, 2015.
- **ACM-ICPC Asia Regional Bronze Medal**, Xi'an site, 2015.

TECHNICAL SKILLS

Generative & Representation Learning

Variational autoencoders (*vanilla / hierarchical / conditional*, MrVI-style sample-level conditioning), contrastive learning (NT-Xent, SimCLR-style), foundation-model pre-training (single-cell, 368M-parameter scale), discrete & continuous latent spaces, joint-VAE.

Optimal Transport

Sinkhorn, Gromov–Wasserstein, Fused-GW, BCD solvers, discrete OT for cross-domain / cross-species alignment.

Equivariant & Geometric DL

SE(3)-equivariant transformers, neural ODEs / dynamical-system modelling, graph neural networks (GAT for gene-regulatory inference).

Multi-modal & Cross-domain Integration

scRNA + spatial transcriptomics + scATAC + imaging; species-, batch-, and sample-level transfer; cross-modal alignment.

Software & Systems

PyTorch (advanced, custom modules & training loops), TensorFlow / JAX, CUDA kernel development, distributed training (DDP, multi-node), mixed precision, SLURM clusters, Docker, Git, GitHub Actions.

Programming Languages

Python, C / C++, R, MATLAB, Java.

Bioinformatics

(Single-cell) RNA-seq, ATAC-seq, ChIP-seq, spatial transcriptomics; gene-regulatory-network inference; trajectory & velocity analysis; cross-species ortholog mapping.

Medical Imaging / Radiomics

CT/MRI feature extraction; multi-scale radiomic modelling (NSCLC heterogeneity quantification, ITHscore).