Ziyang Song

School of Computer Science, McGill University, Montreal, Canada

ziyang.song@mail.mcgill.ca Website LinkedIn Google scholar

I am a PhD candidate specializing **healthcare AI**, with a focus on NLP and time-series analysis. My research mainly focuses on two areas: **Generative AI and Probabilistic AI**. In Generative AI, I develop foundation models on biosignals and medical record data. In Probabilistic AI, I develop probabilistic graph models and deep generative models on phenotype and genotype data.

My expertise includes:

- **Deep Learning:** Transformer and foundation models, SSM, RNN, Neural ODE
- **Deep Generative Models:** VAE, Normalizing Flows, deep latent variable models
- **Probabilistic models:** latent variable models, topic models, mixture models, statistical inference

Education:

- **Ph.D. Computer Science** 09/2020 – Expected 04/2025
  School of Computer Science, McGill University, Montreal, Canada
  Advisor: Yue Li
  Thesis: Probabilistic and generative models for medical health records
- **M.Sc. Quality Systems Engineering** 09/2018 – 10/2019
  Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada
  Advisor: Nizar Bouguila
  Thesis: Nonparametric Bayesian models based on asymmetric Gaussian distributions
- **B.Eng. Computer Science** 09/2014 – 06/2018
  University of Shanghai for Science and Technology, Shanghai, China

Research Experience:

- **PhD Research Assistant -** McGill University 03/2019 – Present
  - **Generative AI:**
    - Developed a foundation model (BiTimelyGPT) using a novel bidirectional generative pre-training on biosignals and longitudinal EHR data. Published at MLHC 2024 **(JMLR)**.
    - Developed a foundation model (TimelyGPT) for forecasting patient healthcare trajectories using biosignals and longitudinal EHR data. Published at ACM BCB 2024.
  - **Probabilistic AI:**
    - Developed a multi-modal, expert-guided topic model (MixEHR-Seed) for medical records. Published at KDD and **won KDD Healthday Best Paper Award**.
    - Developed a multi-modal, self-supervised topic model (MixEHR-S) for medical records. Published at ACM BCB 2021.
- **Master Research Assistant -** Concordia University 03/2019 – 12/2019
  - Conducted research on probabilistic graphical models and statistical inference approaches for anomaly detection and quality analysis in the CV domain.

Selected Publications:

Conference Papers:

1. ***Song, Z.***, Lu, Q., Xu, Q., Buckeridge, DL, Li, Y. (2024). TimelyGPT: Extrapolatable Transformer Pre-training for Long-term Time-Series Forecasting in Healthcare. ACM-BCB, oral presentation.
2. Wang, Z., Wang, R., ***Song, Z.***, Buckeridge, DL, Li, Y. (2024). MixEHR-Nest: Identifying subphenotypes

within electronic health records through hierarchical guided-topic modeling. ACM-BCB.

1. ***Song, Z.***, Lu, Q., Zhu, M., Buckeridge, DL, Li, Y. (2024). Bidirectional generative pre-training for improving healthcare time-series representation learning. MLHC, Proceedings of Machine Learning Research (**JMLR Proceedings track**).
2. ***Song, Z.***, Hu, Y., Verma, A., Buckeridge, DL., Li, Y. (2022). Automatic phenotyping by a seed-guided topic model. **KDD (HealthDay Best Paper award).** DOI: 10.1145/3534678.3542675.
3. ***Song, Z.***, Toral, XS., Xu, Y., Liu, A., Guo, L., Powell, G., Verma, A., Buckeridge, DL., Marelli, A., Li, Y. (2021). Supervised multi-specialist topic model with applications on large-scale electronic health record data. ACM BCB. DOI: 10.1145/3459930.3469543.
4. ***Song, Z.***, Bregu, O., Ali, S., Bouguila, N. (2019). Variational inference of finite asymmetric gaussian mixture models. SSCI. DOI: 10.1109/SSCI44817.2019.9002954.
5. ***Song, Z.***, Ali, S., Bouguila, N. (2019). Bayesian learning of infinite asymmetric gaussian mixture models for background subtraction. ICIAR. DOI: 10.1007/978-3-030-27202-9_24.

Journal Papers:

1. Zou, Y., Pesaranghader, A., ***Song, Z.***, Verma, A., Buckeridge, DL., Li, Y. (2022). Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. Scientific Reports. DOI: 10.1038/s41598-022-22956-w.
2. ***Song, Z.***, Ali, S., Bouguila, N. (2021). Bayesian inference for infinite asymmetric gaussian mixture with feature selection. Soft Computing. DOI: 10.1007/s00500-021-05598-4.
3. ***Song, Z.***, Ali, S., Bouguila, N. (2020). Background subtraction using infinite asymmetric Gaussian mixture models with simultaneous feature selection. IET Image Processing. DOI: 10.1049/iet-ipr.2019.1029.
4. ***Song, Z.***, Ali, S., Bouguila, N. Fan, W. (2020). Nonparametric hierarchical mixture models based on asymmetric gaussian distribution. Digital Signal Processing. DOI: 10.1016/j.dsp.2020.102829.

Preprints:

3. ***Song, Z.***, Lu, Q., Zhu, M., Buckeridge, DL, Li, Y. (2024). TrajGPT: Healthcare Time-Series Representation Learning for Trajectory Prediction. ICLR 2025, under review.
4. ***Song, Z.***, Xu, M., Latour, F., Gravel, S., Ho, V., Lettre, G., Li, Y. (2024). AI-driven approach for computational phenotyping with CARTaGENE cohort. Scientific Meeting of the Canadian Translational Geroscience Network. Abstract for oral presentation.
5. ***Song, Z.***, Yang, Z., Wang, R., Zabad, S., MA Legault, MA., Li, Y. (2023). MixEHR-SAGE: A seed-guided topic model to improve phenotyping for PheWAS analysis in UK Biobank data. ISMB and GLBIO. Abstract for oral presentation.

Selected Presentations:

Oral Presentation:

1. ***Song, Z.*** et al. (2024, Sep 5-6). AI-driven approach for computational phenotyping with CARTaGENE cohort. Scientific Meeting of the Canadian Translational Geroscience Network. Montreal, Canada.
2. ***Song, Z.*** et al. (2024, Apr 8-13). Practical - phenotyping and PheWAS using MixEHR-seed. Tokyo Symposium on Genomic Medicine, Therapeutics and Health. Tokyo, Japan.
3. ***Song, Z.*** et al. (2023, May 15-18). MixEHR-SAGE: A seed-guided topic model to improve phenotyping for PheWAS analysis in UK Biobank data. ISMB GLBIO. Montreal, Canada.
4. ***Song, Z.*** et al. (2022, Aug 14-18). Automatic phenotyping by a seed-guided topic model. KDD (HealthDay Best Paper award). Washington DC, U.S.
5. ***Song, Z.*** et al. (2021, Aug 1-4). Supervised multi-specialist topic model with applications on large-scale electronic health record data. ACM BCB. Virtual.

Poster Presentation:

1. ***Song, Z.*** et al. (2024, Aug 16-17). Bidirectional generative pre-training for improving healthcare time-

series representation learning. MLHC. Toronto, Canada
2. ***Song, Z.*** et al. (2022, Oct 21). MixEHR-Seed: automatic phenotyping by a seed-guided topic model. 50th Anniversary of SOCS at McGill University. Montreal, Canada.
3. ***Song, Z.*** et al. (2019, Aug 27-29). Bayesian learning of infinite asymmetric gaussian mixture models for background subtraction. ICIAR. Waterloo, Canada

Honors and Awards:

Academic Honros:

- ***FRQNT Provincial Doctoral Scholarship ($58,334)*** 05/2023 – Present
- ***McGill Stipend from Prof. Li ($40,000)*** 09/2020 – Present
- ***Grad Excellence Award in Computer Science ($38,700)*** 09/2020 – Present
- Faculty of Science Grad Supplement Award ($3,250) 09/2022 – 08/2024
- SOCS Grad Stimulus Initiative Award ($22,000) 09/2020 – 08/2022
- Jackie Cheung Graduate Award ($18,000) 09/2020 – 08/2021
- Concordia Master Research Stipend ($15,000) 09/2018 – 10/2019

Awards:

- KDD healthy day best paper award 09/2020 – 08/2021

Professional Experience:

- **Research Intern -** Centre de recherche du CHU Sainte–Justine 06/2023 – Expected 12/2024
  - Developed a clinical decision-support system using AI-driven models to track patient diagnoses, monitor disease progression, and analyze cancer registration.
  - Applied a probabilistic AI model (MixEHR-SAGE) to infer expert-guided phenotype distributions for 50K individuals in the CARTaGENE cohort. Presented in oral presentation.
  - Leveraged a foundation model (TimelyGPT) to analyze patient healthcare trajectories and predict disease diagnoses within the CARTaGENE cohort.
- **International Research Intern -** Nanyang Technology University 01/2020 – 06/2020
  - Developed probabilistic models using stochastic optimization techniques to analyze retail data, optimizing product recommendations and promotions tailored to patient profiles.
  - This project led to a significant improvement in the accuracy of the item recommendations and patient-targeted promotions.
- **Data Analyst Intern -** Shanghai MetaLab 07/2017 – 12/2017
  - Developed a click-model for a recommender system and implemented machine learning algorithms to analyze patent data. Supported data analysis and visualization to aid in marketing decisions.
  - Developed a web crawler, database system, and back-end data interfaces to process around 10 million public patents. Maintained the database for efficient and reliable data analysis.
- **Data Analyst Intern -** Shanghai Qingyue Environment Protection Center (NGO) 01/2017 – 03/2017
  - Developed an environmental data platform that provides public access to weather data, aimed at supporting efforts to reduce environmental pollution in China.
  - Developed a web crawler, database system, and back-end data interfaces to efficiently process and manage publicly available weather datasets.

Teaching Experience:

- **Guest Lecturer -** McGill University Fall 2023
  - Course: COMP565 Machine Learning in Genomics and Healthcare
  - Title: Time-series Transformer in EHR
- **Teaching Assistant** – McGill University Winter 2023

- Course: COMP551 Applied Machine Learning
- Duties: Office hours, tutorials, grading, design and evaluate final project.

- **Teaching Assistant** – McGill University Fall 2022
  - Course: COMP551 Applied Machine Learning
  - Duties: Office hours, tutorials, grading, research project instructor.

Mentor Experience:

- Undergraduate and master student Ziqi Yang 09/2022 – Present
  - Mentored in applying MixEHR-Seed on UKB dataset, presented to GLBIO 2023.
  - Placement: master student, McGill University
- Undergraduate and master student Ruohan Wang 09/2022 – Present
  - Mentored in applying MixEHR-Seed on MIMIC dataset, presented to ISMB GLBIO 2023.
  - Mentored in designing MixEHR-Nest model on MIMIC dataset, Published at ACM BCB 2024.
  - Placement: master student, McGill University
- Undergraduate student Hao Xu 05/2023 – 09/2023
  - Mentored in applying TimelyGPT model on biosignals, Published ACM BCB 2024.
  - Placement: master student, University of Pennsylvania
- Undergraduate student Ziyu Zhao 05/2023 – 09/2023
  - Mentored in applying McGill models on Quebec PopHR database.
  - Placement: master student, McGill University
- Undergraduate student Yuanyi Hu 05/2021 – 05/2022
  - Mentored in designing MixEHR-Seed model, published at KDD 2022.
  - Placement: master student, Columbia University
- Undergraduate student Yixin Xu 09/2020 – 09/2021
  - Mentored in applying MixEHR-S model, published at ACM-BCB 2021.
  - Placement: master student, Duke University

Reviewer Experience:

Reviewer:

- ICLR (2024), NeurIPS 2024 Workshop TSALM (2024), Machine Learning for Health (ML4H) (2024)

Sub-reviewer:

- KDD (2022, 2024), ISMB (2022)

Skills:

- **Programming Language:** Python, R, C++, Java
- **Computing Libraries:** NumPy, SciPy, Pandas**,** Scikit-learn, PyTorch, TensorFlow, Gensim, SpaCy
- **Databased Management:** MySQL, PostgreSQL, MongoDB, Dask

Referees:

- Dr. Yue Li. Assistant Professor in School of Computer Science, McGill University. yue.yl.li@mcgill.ca
- Dr. David Buckeridge. Professor in School of Population and Global Health, McGill University. david.buckeridge@mcgill.ca