XINSONG DU

Personal Website: https://bit.ly/3yVgkay
Email: xidu1@bwh.harvard.edu
255 Vale ST APT 6216
Chelsea, MA 02150

EDUCATION & TRAINING

Postdoctoral Research Fellow.

Harvard Medical School, Brigham and Women's Hospital — Aug 2023 – Present

Supervisor:

- Dr. Li Zhou

Training:

- Large language models
- Electronic health records

Ph.D.

University of Florida, Biomedical Informatics — Aug 2017 – Aug 2023

Supervisory Committee:

- Dr. Dominick J. Lemas (Chair)
- Dr. William R. Hogan
- Dr. Timothy J. Garrett
- Dr. Mathias Brochhausen
- Dr. Mei Liu

Training:

- Mass-spectrometry
- Metabolomics
- Research software development
- Natural language processing
- Research reproducibility
- Biomedical sample handling
- Programming: Nextflow/Groovy

M.S.

University of Florida, Electrical & Computer Engineering — Aug 2015 – May 2017

Training:

- Data analysis for electronic health records
- Machine Learning
- High-performance computing
- Computer vision
- Programming: Python/R/Shell
- Computer architecture
- Computer communications

B.S.

Shandong University, School of Electrical Engineering Sep 2011 – Jul 2015
Training:

- Programming: C/C++/Assembly Language/MATLAB/Action Script
- Mathematics: Calculus, statistics, linear algebra, field theory, signal processing
- Single-chip micro computer
- Physics: electronic magnetic field
- Power electronics

RESEARCH FUNDING

UF Informatics Institute Graduate Student Fellowship 04/2021-04/2023

Support dissertation research *Enabling Reproducible Untargeted Metabolomics Research: Next Generation Untargeted Metabolomics Data Analysis Workflow* ($45,000)
Research funding given to selected graduate students nominated by their department who are working on informatics-related area at UF (awards about 5 graduate students in total each year).
University of Florida.

TRAVEL GRANTS

UF Graduate Student Council Travel Grant 2023
Support my travel to the 71st Annual Conference of American Society for Mass Spectrometry ($350)

UF HOBI Graduate Student Travel Grant 2023
Support my travel to the 71st Annual Conference of American Society for Mass Spectrometry ($500)

UF Office of the Vice President for Research Travel Grant 2023
Support my travel to the 71st Annual Conference of American Society for Mass Spectrometry ($400)

HONORS AND AWARDS

2024 Metabolomics Publication Awards: Best Review Winner 2024
First author of the best *Metabolomics* review article published in 2023

Mass General Brigham: Pillars of Excellence Awards 2024
Member of the awarded team: Language Documentation in the Electronic Health Records - Equity and Bias Team

UF Three Minute Thesis (3MT) Competition Finalist 2022
Ranked top 10 in the competition.

Alec Courtelis Award (Nomination) 2022
Award for international students who contributed significantly to the community and have outstanding academic achievement.
Each college can nominate a maximum of two students every year.

Certificate of Outstanding Merit 2022
Certificate issued to selected international students,
UF College of Medicine and International Center

Study Florida & FAIE 2022 High Education Scholarship (Nomination) 2021
Scholarship that awards international students studying in Florida.
A maximum of three international students can be nominated by each university every year.

Certificate of Outstanding Merit 2021
Certificate issued to selected international students,
UF College of Medicine and International Center

Achievement Award Scholarship for New Engineering Graduate Students 2015
Award given to selected incoming graduate student at College of Engineering,
University of Florida.

JOURNAL PUBLICATIONS

1. Lemas DJ, **Du X**, Rouhizadeh M, Lewis B, Frank S, Wright L, Spirache A, Gonzalez L, Cheves R, Magalhães M, Zapata R, Reddy R, Xu K, Parker L, Harle C, Young B, Louis-Jaques A, Zhang B, Thompson LA, Hogan WR, Modave F. Classifying early infant feeding status from clinical notes using natural language processing and machine learning. *Scientific Reports* (2024). PMID: 38570569

2. **Du X,** Dastmalchi F, Diller MA, Brochhausen M, Hogan WR, Lemas DJ. An Automated Workflow Composition System for Liquid Chromatography-Mass Spectrometry Metabolomics Data Processing. Journal of American Society of Mass Spectrometry. 2023. PMID: 37874901

3. Lemas DJ, **Du X**, Dado-Senn B, Xu K, Dobrowolski A, Magalhães M, Aristizabal-Henao JJ, Young BE, Francois M, Thompson LA, Parker LA, Neu J, Laporta J; Misra BB, Wane I, Samaan S, Garrett TJ. Untargeted Metabolomic Analysis of Lactation-Stage-Matched Human and Bovine Milk Samples at 2 Weeks Postnatal. PMID: 37686800

4. **Du X**, Dastmalchi F, Ye H, Garrett TJ, Dillar MA, Liu M, Hogan WR, Brochhausen M, Lemas DJ. Evaluating LC-HRMS Metabolomics Data Processing Software using FAIR Principles for Research Software. *Metabolomics* (Editor-Selected Cover Article) (2023 Metabolomics Award: Best Review Article Winner). 2023. PMID: 36745241

5. **Du X**, Aristizabal-Henao JJ, Garrett TJ, Brochhausen M, Hogan WR, Lemas DJ. A checklist for reproducible computational analysis In clinical metabolomics research. *Metabolites*. 2022. PMID: 35050209

6. Lemas DJ, Wright L, Flood-Grady E, Francois M, Chen LY, Hentschel A, **Du X,** Hsiao CJ, Chen H, Neu J, Theis RP, Shenkman E, Krieger J. Perspectives of pregnant and breastfeeding women on longitudinal clinical studies that require non-invasive biospecimen collection – a qualitative study. *BMC Pregnancy and Childbirth*. 2021. PMID: 33472584

7. Hentschel A, Hsiao CJ, Chen LY, Wright L, Shaw J, **Du X,** Flood-Grady E, Harle CA, Reeder CF, Francois M, Louis-Jacques A, Shenkman E, Krieger J, Lemas DJ. (2021) Perspectives of pregnant and breastfeeding persons on participating in longitudinal mother-baby studies involving electronic health records: a qualitative study. *JMIR Pediatrics and Parenting.* 2021. PMID: 33666558

8. Lemas DJ, Loop LS, Duong M, Schleffer A, Collins C, Bowden JA, **Du X,** Patel K, Ciesielski AL, Ridge Z, Wagner J, Subedi B, Delcher C. Estimating drug consumption during a college sporting event from wastewater using liquid chromatography mass spectrometry. *Science of The Total Environment*. 2021. PMID: 33385644

9. Lure AC, **Du X.** Black EW, Irons R, Lemas DJ, Taylor JA, Lavilla O, de la Cruz D, Neu J. Using machine learning analysis to assist in differentiating between necrotizing enterocolitis and spontaneous intestinal perforation: a novel predictive analytics tool. *Journal of Pediatric Surgery*. 2020. PMID: 33342603

10. **Du X**, Min J, Shah CP, Boshnoi R, Hogan WR, Lemas DJ. Predicting in-hospital mortality of patients with febrile neutropenia using machine learning models. *International Journal of Medical Informatics*. 2020. PMID: 32325370

11. Bian J, Zhao Y, Salloum RG, Guo Y, Wang M, Prosperi M, Zhang H, **Du X,** Ramirez-Diaz LJ, He Z, Sun Y. (2017) Using social media data to understand the impact of promotional information on laypeople's discussions: a case study of lynch syndrome. *Journal of Medical Internet Research*. 2017. PMID: 29237586

12. Du C, **Du X.** (2016) Cache optimization by fully-replacement policy. *American Journal of Embedded Systems and Applications.* 2016.

MANUSCRIPTS UNDER REVIEW

1. **Du X,** Novoa-Laurentiev J, Plasaek JM, Chuang YW, Wang L, Marshall G, Mueller SK, Chang F, Datta S, Paek H, Lin B, Wei Q, Wang X, Wang J, Ding H, Manion FJ, Du J, Bates DW, Zhou L. Enhancing Early Detection of Cognitive Decline in the Elderly: A Comparative Study Utilizing Large Language Models in Clinical Notes. *medRxiv*. 2024

2. **Du X,** Zhengyang Zhou, Yifei Wang, Ya-Wen Chuang, Richard Yang, Wenyu Zhang, Xinyi Wang, Rui Zhang, Pengyu Hong, David W. Bates, Li Zhou. Generative Large Language Models in Electronic Health Records for Patient Care Since 2023: A Systematic Review. *medRxiv*. 2024

CONFERENCE PRESENTATIONS

1. **Du X,** Novoa-Laurentiev J, Plasaek JM, Chuang YW, Wang L, Marshall G, Mueller SK, Chang F, Datta S, Paek H, Lin B, Wei Q, Wang X, Wang J, Ding H, Manion FJ, Du J, Bates DW, Zhou L. Enhancing Early Detection of Cognitive Decline in the Elderly through Ensemble of NLP: A Comparative Study Utilizing Large Language Models in Clinical Notes. Annual Symposium of American Medical Informatics Association Nov 09-13, 2024. San Francisco, CA. **(Oral – Podium Abstract)**

2. **Du X,** Novoa-Laurentiev J, Plasaek JM, Chuang YW, Wang L, Marshall G, Mueller SK, Chang F, Datta S, Paek H, Lin B, Wei Q, Wang X, Wang J, Ding H, Manion FJ, Du J, Bates DW, Zhou L. Detection of Cognitive Decline in Clinical Notes Using an Ensemble of Large Language Models, Deep Learning Models, and Machine Learning Models. Apr 01, 2024. MIT-MGB AI Cures. Somerville, MA.

3. **Du. X.**; Dastmalchi. F.; Diller. M.A.; Brochhausen. M.; Garrett. T.J.; Hogan. W.R.; Lemas. D.J. (2023). An Automated Workflow Composition System for LC-MS Metabolomics Research. Annual Conference of American Society of Mass Spectrometry. Jun 04-08, 2023. Houston, TX. **(Oral)**

4. **Du, X.;** Cardel, M.I.; Millar, D.R.; Aristizabal-Henao, J.J.; Bowden, J.A.; Lemas, D.J. Untargeted Urinary Metabolomics Analysis for An Acceptance-Based Therapy Intervention for Diverse Adolescent Girls with Overweight/Obesity. ObesityWeek, Nov.1-5, 2021. Online

5. **Du, X.**; Luran, M.; Xu, K.; Kirpich, A.; Hogan, W.R.; Garrett T.J.; Lemas, D.J. A Reproducible Pipeline for Scalable Untargeted Metabolomics Data Analysis. Annual Meeting of Metabolomics Association of North America, Nov. 15-17, 2019. Atlanta, GA.

6. **Du, X.**; Bian, J.; Prosperi, M. An Operational Deep Learning Pipeline for Classifying Life Events from Individual Tweets. *5th International Conference on Information Management and Big Data*, Sep. 03-05, 2018. Lima, Peru. **(Oral)**

SOFTWARE AND TOOLS

Nextflow4Metabolomics Suite. Reproducible pipelines for untargeted metabolomics data processing.
- Link: https://github.com/Nextflow4Metabolomics
- Role: Major contributor.

Keras. One of the most popular deep learning application programming interfaces (API) with over 375,000 users all over the world.
- Link: https://keras.io/
- Role: Contributor. My contribution enables Keras users to use Scikit-Learn to do cross-validation for deep learning models developed with Keras functional API.

INVITED TALKS

Invited Talk. Computational reproducibility in liquid chromatography-mass spectrometry-based clinical metabolomics data processing. *University of Florida Informatics Institute*. Apr. 04, 2022. Online. [video: https://bit.ly/3O8f6Rb]

TEACHING

Guest Lecturer. GMS 6804: Translational Biomedical Informatics (Dr. Dominick Lemas), University of Florida College of Medicine. Feb.-Mar. 2022. Gainesville, FL.
- Led students to visit University of Florida Research Computing Center.

- Led students to visit biomedical informatics data acquisition facilities in the University of Florida Interdisciplinary Center for Biotechnology Research.
- Introduced and discussed about research reproducibility in biomedical field, and led a discussion related to translational bioinformatics scientific papers.

Guest Lecturer. GMS 6804: Translational Bioinformatics (Dr. Dominick Lemas), University of Florida, College of Medicine. Apr. 07, 2020. Gainesville, FL.

- Introduced and discussed about a reproducible computational pipeline I developed for metabolomics data processing.

Teaching Assistant. Biomedical Informatics Summer School: Machine Learning Basics, University of Florida, College of Medicine. Jul. 23- Aug. 10, 2018. Gainesville, FL

- Taught students basic machine learning knowledges, led students to complete assignments and course projects created by myself: https://github.com/XinsongDu/Basic_ML_Practices

MENTORING

- Xinyu Chen, Harvard University, 2024
  Project: Using natural language processing to identify bilingualism from clinical notes.

- Xinyi Wang, Harvard University, 2024
  Project: Automated generation of genetic variant summarizations from genetic variant databases using large language models.

- Amanda Dobrowolski, University of Florida, 2023
  Project: Using Nextflow-based containerized workflow to process metabolomics data.

- Braeden Lewis, University of Florida, 2022
  Project: Predicting breastfeeding outcomes using machine learning approach and clinical text data.

- Ismael Wane, University of Florida, 2022
  Project: Organizing and cleaning identified metabolites using Human Metabolome Database.

- Emmanuel Elias, University of Florida, 2022
  Project: Developing a tool for automatic extraction of Human Metabolome Database ID and taxonomy information.

PROFESSIONAL TRAININGS

Biobank Portal Course
Center for Clinical Research Education, Mass General Brigham, Boston, MA.
Jun. 26, 2024
Description: This course teaches how to use the Mass General Brigham biobank portal and how to request data from it.

Mentoring Circles Program
Brigham Research Institute, Mass General Brigham, Boston, MA.
Oct. 11, 2023 – May. 24, 2024
Description: This training provided junior postdoc with guidance regarding career development.

Metabolomics Winter School
Southeast Center for Integrated Metabolomics, University of Florida, Gainesville, FL.
Jan. 27 – Jan. 29, 2020.
Description: This workshop covered cutting-edge technologies about metabolomics including sample handling, sample processing, instruments, pick picking, data analysis, etc.

Bits & Bites: Short Course Series 2021 (Online)
West Coast Metabolomics Center, UC Davis, Davis, CA.
Feb. 04 – Dec.02, 2021.
Description: This series of courses covered latest computational techniques related to metabolomics research including signal processing, metabolite annotation, statistical analysis, and data interpretation.

Focus on Mentoring Series (Online),
Office of Graduate Professional Development, University of Florida, Gainesville, FL.
Feb. 2022-Mar.2022
Description: This series of courses covered the detailed explanation of student mentoring, issues in mentoring, and ways to solve those issues. It also covered topics related to research integrity.

Coursera Certifications

- Database and SQL for Data Science with Python. Nov 2023
- AI for Medical Diagnoses. Jun 2020
- Introduction to HTML 5. Apr 2020
- Practical Reinforcement Learning. May 2019
- Natural Language Processing. Jul 2018
- Mathematics for Machine Learning: Linear Algebra. Jun 2018
- Mathematics for Machine Learning: Multivariate Calculus. Jun 2018

PROFESSIONAL SERVICES

Editorial Board

- Associate Editor, Journal of Translational Medicine (2023-Present)
- Student Editor, International Journal of Medical Informatics (2024-Present)

Reviewer for Journals:

- International Journal of Medical Informatics (2023, 2024)
- Journal of Translational Medicine (2022, 2023, 2024)
- Journal of Biomedical Informatics (2020, 2023, 2024)
- Scientific Reports (2022, 2023, 2024)
- Journal of Big Data (2023)
- BMC Pregnancy and Childbirth (2023)

- Pharmacoepidemiology and Drug Safety (2023)
- Trends in Computer Science and Information Technology (2022)
- Journal of Biomolecular Techniques (2021)
- BMC Medical Informatics and Decision Making (2017)

Reviewer for Conferences:

- American Medical Informatics Association (AMIA) Annual Symposium (2018-2024)
- AMIA Summit (2023)
- AMIA Clinical Informatics Conference (2023)
- International Conference on Bioinformatics and Biomedicine (2017)
- ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (2017)

Conference Program Committee:

- Service Computation 2022, Barcelona, Spain. Apr. 24 – Apr. 28, 2022 (https://www.iaria.org/conferences2022/ComSERVICECOMPUTATION22.html)

Participant of:

- The survey for producing *Times Higher Education World University Rankings*. 2021, 2022.
- The research study *Exercise as Medicine: Evaluation of a College Multidisciplinary Fitness Intervention Strategy on Perceived Wellness, Adherence, Resting Heart Rate & Blood Pressure for Sedentary Individuals*. 2022

LANGUAGES

English: Full professional proficiency

Chinese: Native proficiency

TECHNICAL SKILLS

Programming: Python, Shell, R, Groovy, Nextflow

Applications: Machine Learning, Research Software Development, High-Performance Computing, Software Containerization

Platforms: Amazon Web Services, Amazon Mechanical Turk, HiPerGator, Jupyter Notebook, Nextflow, GitHub, Docker/Singularity.

REFERENCES

Dr. Li Zhou, MD, Ph.D., Harvard Medical School. 399 Revolution Drive, Somerville, MA 02145. Email: lzhou@bwh.harvard.edu. Phone: (617) 640-2407

Dr. Dominick J. Lemas, Ph.D., University of Florida College of Medicine, Department of Health Outcomes and Biomedical Informatics. 2004 Mowry Road-Clinical and Translational Research Building, Gainesville, FL 32610. Email: djlemas@ufl.edu. Phone: (352) 294-5971

Dr. Mathias Brochhausen, Ph.D., University of Arkansas for Medical Sciences, Department of Biomedical Informatics. 4301 West Markham Street, Little Rock, AR 72205. Email: mbrochhausen@uams.edu. Phone: (501) 686-7000

Dr. Timothy Garrett, Ph.D., University of Florida, Department of Pathology, 1395 Center Dr, Room M641c, Gainesville, FL 32610. Email: tgarrett@ufl.edu. Phone: (352) 273-5050

Dr. William R. Hogan, M.D., University of Florida College of Medicine, Department of Health Outcomes and Biomedical Informatics. 2004 Mowry Road-Clinical and Translational Research Building, Gainesville, FL 32610. Email: hoganwr@ufl.edu. Phone: (352) 294-4197