Amy X. Lu

amyxlu@cs.toronto.edu
http://amyxlu.github.io
github.com/amyxlu

I am interested in **developing machine learning methods for the unique challenges in computational biology**, especially in generalizability, representation learning, and fairness. I am currently exploring these interests at Insitro, a machine learning driven drug discovery company led by Dr. Daphne Koller.

Education

- **University of Waterloo** Waterloo, Canada
  Bachelors of Science, Honours Science, Bioinformatics Option *2014 – 2018*
- **University of Toronto, Vector Institute** Toronto, Canada
  Master's in Computer Science *2019 – 2020*

Research Experience

- **Insitro** South San Francisco, USA
  Machine Learning Engineer III *2020 – Present*
  - **Research engineering, generalizability, image-based profiling:** Engineering and analyzing generalizable and biologically interpretable representations for high-throughput imaging and gene expression phenotypes.
  - **Molecular design, representation learning:** Exploring problems in ML-guided molecular design through representation learning and uncertainty quantification.
- **Stanford University** Palo Alto, USA
  Visiting Student Researcher — Domain Adaptation in Regulatory Genomics, Dr. Anshul Kundaje *2019 – 2019*
  - **Domain adaptation, regulatory genomics, interpretability:** Using domain adaptation methods to improve transcription factor binding prediction when evaluating for a different cell line.
- **University of Toronto/Vector Institute** Toronto, Canada
  Masters Student — Representation learning in genomics, Dr. Alan Moses & Dr. Marzyeh Ghassemi *2019 – 2020*
  - **Self-supervised learning, representation learning for proteins:** Developed a parameter-efficient representation for proteins using contrastive mutual information maximization (*MLCB 2020; bioRxiv*).
  - **Generalizability, microscopy imaging:** Benchmarked self-supervised computer vision methods in a microscopy image dataset with covariate shift to highlight generalization failures in machine learning (*NeurIPS 2019*).
  - **Algorithmic fairness, clinical decision support:** Quantitative and qualitative evaluation of bias in contextual word embeddings on clinical notes; fairness definitions for multi-group settings (*Spotlight, ACM CHIL 2020*).
- **Harvard Medical School/Boston Children's Hospital** Boston, USA
  Research Intern — Machine Learning in Clinical Genomics, Dr. Piotr Sliz *2018 – 2019*
  - **Genotype-phenotype studies, clinical interpretability:** Understanding disease genotype-phenotype relationships using machine learning models. Interpreted important model features to seek novel disease-associated variants from whole exome (WES) data.
  - **Variant filtering, pathway analyses:** Applied standard filtering pipelines for false-positive variants. Explored classification from pathway- and variant-level features.
  - **Statistical genetics:** Explored methods for capturing epistatic non-linearities and statistical dimensionality reduction.
- **University of Waterloo** Waterloo, Canada
  Undergraduate Thesis Student — Deep Learning in Regulatory Genomics, Dr. Andrew Doxey *2017 – 2018*
  - **Chromatin accessibility prediction, interpretability:** Trained a convolutional neural network to classify ATAC-seq accessible regions in femur growth regulation; reconstructed first-layer features as a position-weighted matrix (PWM) with statistical matches in JASPAR, a database of known motifs.
  - **Phylogenetics, metagenomic data mining:** Used various bioinformatics pipeline tools (HMMER, BLAST, etc.) to understand biochemical properties of potentially uncharacterized toxins in metagenomic data.
- **École polytechnique fédérale de Lausanne** Lausanne, Switzerland
  Research Intern — Molecular Dynamics Simulations, Dr. Matteo Dal Parero *2017*

   - **Molecular dynamics:** Used molecular dynamics (MD) and GROMACS to simulate enzyme-membrane mechanisms of antibiotics resistance.

- **University of Toronto** Toronto, Canada
  Research Intern — Data Visualization in Pharmacoepidemiology, Dr. Suzanne Cadarette *2015 – 2017*
   - **Data visualization, pharmacoepidemiology:** Analysis and visualization of the social diffusion of methodological innovation in pharmacoepidemiology.

Journal Articles and Conference Proceedings

- C Dallago, K Schtze, M Heinzinger, T Olenyi, M Littmann, **AX Lu**, KK Yang, S Min, S Yoon, JT Morton, B Rost. Using protein sequence representations from deep learning to visualize and predict protein sets. ***Current Protocols.*** *In Revision.*
- **AX Lu**, H Zhang, M Ghassemi, AM Moses. Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization. ***Machine Learning for Computational Biology (MLCB) 2020.*** *Preprint.*
- **AX Lu**, AX Lu, AM Moses. Evolution Is All You Need: Phylogenetic Augmentation for Contrastive Learning. ***Machine Learning for Computational Biology (MLCB) 2020.*** *Preprint.*
- H Zhang*, **AX Lu***, M Abdalla, M McDermott, M Ghassemi. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. ***Spotlight, ACM Conference on Health, Inference, and Learning (CHIL) 2020.*** **Equal Contribution. Preprint.*
- AX Lu, **AX Lu**, W Schormann, M Ghassemi, DW Andrews, AM Moses. The Cells Out of Sample (COOS) dataset and benchmarks for measuring out-of-sample generalization of image classifiers. ***Neural Information Processing Systems (NeurIPS) 2019.*** *Preprint.*
- AM Moses, AX Lu, **AX Lu**, M Ghassemi. Transfer Learning vs. Batch Effects: what can we expect from neural networks in computational biology? ***Machine Learning for Computational Biology (MLCB) 2019.***
- J Ban, M Tadrous, **AX Lu**, EA Cicinelli, SM Cadarette. Diffusion of indirect comparison meta-analytic methods to study drugs: a systematic review and co-authorship network analysis. ***BMJ Open.***

Workshop Papers and Posters

- C Dallago, K Schütze, M Heinzinger, T Olenyi, M Littmann, **AX Lu**, KK Yang, S Min, S Yoon, B Rost. Streamlining value of protein embeddings through `bio_embeddings`. ***NeurIPS 2020 Workshop on Learning Meaningful Representations of Life (LMRL).***
- M Abdalla, H Zhang, **AX Lu**, I Chen, M Ghassemi. Quantifying Fairness in a Multi-Group Setting and its Impact in the Clinical Setting. ***NeurIPS 2019 Workshop on Fair ML for Health.***
- H Zhang*, **AX Lu***, M Abdalla, M McDermott, M Ghassemi. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. ***NeurIPS 2019 Workshop on Machine Learning for Healthcare (ML4H).*** **Equal contribution.*
- AX Lu, **AX Lu**, AM Moses. Paired Cell Inpainting: A Multiple-Instance Extension of Self-Supervised Learning for Bioimage Analysis. ***ICML 2019 Workshop on Self-Supervised Learning.***
- **AX Lu**, S Rockowitz, A Poduri, P Sliz. From data to precision medicine: predictive machine learning models to uncover disease-associated variants. ***Harvard Medical School BCMP Retreat 2019.***

Awards

- **NSERC Postgraduate Scholarships – Doctoral program (PGS D) Award**: Federal doctoral scholarship tenurable abroad, selected in the Committee for Computing Sciences ($63,000).
- **NSERC Canada Graduate Scholarships – Doctoral (CGS D) Award**: Federal doctoral scholarship tenurable only at a Canadian institution, selected in the Committee for Computing Sciences ($105,000) [declined].
- **Alexander Graham Bell Canada Graduate Scholarships – Master's (CGS M) Award**: Federal research grant ($17,500).
- **NSERC Michael Smith Foreign Supplement**: "Supports high-calibre Canadian graduate students in pursuing research abroad" ($6,000).
- **EPFL Scholarship of Excellence in Research**: Sponsors students for research internship at EPFL (CHF 4,500).
- **University of Waterloo**: President's Scholarship of Distinction, Arebi Family Science Scholarship.
- **Royal Conservatory of Music (RCM)**: ARCT Performer's Diploma in Piano.

Service and Activities

- **Program Committee:** Reviewer, NeurIPS Workshop on Machine Learning for Health 2020.
- **Research to the People** (*formerly SVAI*): Core Team of Research to the People, a non-profit connecting patients of rare genomic diseases to the medical/AI research community and industry partners through collaborative research initiatives.
- **Tosamaganga Hospital, Tanzania:** Supported operations at a rural Tanzanian hospital.
- **Residence Don:** Organized events, responded to crises, and established rapport with diverse students. Leader for the Velocity Residence, a spin-off program of the Velocity start-up incubator.

Teaching

- **Teaching Assistant, Genetics:** Taught weekly tutorial lectures for BIOL 239 at the University of Waterloo.
- **Piano, music theory:** Taught piano performance, ear training, RCM music history, and RCM Intermediate Rudiments.

Talks

- **Vector NLP Talks:** Quantifying and Removing Biases in Clinical Contextual Word Embeddings. *Co-presenter.*
- **Harvard Medical School BCMP Retreat 2019:** From data to precision medicine: predictive machine learning models to uncover disease-associated variants. *Lightning talk.*