**Research Statement**

Alex X. Lu  
Department of Computer Science  
University of Toronto

alexlu@cs.toronto.edu

In the past decade, deep learning has revolutionized the way we understand biology. Neural networks can recognize signals in biological data more sensitively than conventional approaches, leading to performance gains in bioinformatics problems across domains1. While we know how to apply deep learning to automate routine tasks where we can reliably annotate data2,3, we know less about how to effectively apply these methods when biological classes are unknown to us or otherwise cannot be labeled, or when training data is scarce.

My research focuses on addressing computational questions that emerge from applying deep learning to these more challenging datasets in biology:

1. *How do we train neural networks to detect specific biological signals in data?*  
   Many biological datasets are characterized by a low signal-to-noise ratio. For any deep learning application to be successful, models must focus on specific signals relevant to our biological question, while ignoring most of the variation, caused by noise.
2. *How can we purpose deep learning for the discovery of new hypotheses?*  
   High-throughput datasets may contain phenotypes and patterns previously unknown to us. Finding these cases may support the development of novel hypotheses. At the same time, these datasets are too big to evaluate case-by-case, so discovery benefits from computational methods.
3. *How do we design deep learning methods that generalize under data constraints?*   
   Smaller labs may only generate limited quantities of data. To make the benefits of deep learning more widespread, we need to find ways to train generalizable models with small datasets.

Currently, many of the methods we employ in biology address the first challenge by brute force: to get neural networks to learn specifically about biology, researchers manually label training examples for all the biological classes that they want the model to recognize. Each class must contain diverse examples that show the class under uninteresting variation that the biologist wants the model to ignore, requiring thousands2, if not millions of examples4. This process, supervised learning, is time-consuming, limiting deep learning in efficiently and rapidly analyzing data. Moreover, it directly limits the applicability of deep learning to the second and third questions. Since we must have prior knowledge of biology to be able to label it, supervised models are biased towards biology we already know a lot about, making them less suited for discovery. Since we require large training datasets to learn biology, deep learning cannot be easily applied to smaller datasets or less diverse datasets.

My research will focus on creating innovative learning strategies for artificial intelligences in biology. Contrary to supervised learning, which can be described as rote-learning, I design methods that learn to recognize biological signal through puzzle-solving. These methods will let us train high-performance models with less effort, less bias, and more strategic learning, allowing them to be more data efficient.

**1. Background: Self-supervised methods learn transferrable knowledge through play**

Self-supervised learning is a recent innovation which teaches neural networks transferrable skills and representations through puzzle-solving5. These methods are inspired by developmental psychology: children learn rapidly about the world around them from play and exploration6. Self-supervised methods parallel this learning by training neural networks on tasks that often resemble play. By solving tasks like colorizing greyscale images7 or guessing how much an image has been rotated by8, models learn representations that outperform previous unsupervised methods in applications like object classification, and in some cases, fully supervised methods9.

Since self-supervised tasks are designed to be autonomous, they solve the issue of needing large labeled datasets to train neural networks. Instead, self-supervised methods learn to focus on relevant concepts in input data through design of the training task. Different tasks target different signals in data: colorization results in models sensitive to the boundaries of objects in images10, while predicting rotations ignores textures11. Compared to conventional unsupervised methods, which have no way to distinguish learning about relevant variation versus noise and thus often perform poorly in biological analyses12–14, self-supervised methods allow us to target specific kinds of variation in data. While self-supervised learning can potentially address bottlenecks in biological data analysis, these methods requires creating tasks that learn relevant features for biological domains.

I pioneered one of the first self-supervised methods for biology, for fluorescent microscopy images12. Other self-supervised work is emerging for biological data15,16 (especially in protein sequences17,18), but many repurpose methods from mainstream data domains. I design methods that incorporate biological assumptions. My work in images assumes that genetically identical cells often express proteins similarly to pose an inpainting task between pairs of cells12; my current work in protein sequences assumes that homologues share functions. In the remainder of my research statement, I will outline how my methodological expertise enables me to address open biological challenges.

**2.** **Self-supervised learning learns unbiased representations without manual effort**

Representation methods convert data into quantitative measurements. Representations enable the numerical comparison of complex data like images, text or audio, which are difficult to compare in their raw form. Biological domains similarly benefit from representations, but in many domains, much of the variation in data is uninteresting technical or background signal. To address this, researchers engineer targeted features that measure only aspects of data relevant to their question19,20. Alternatively, neural networks can learn relevant features given large training datasets labeled with biology of interest4,21. Both options require extensive manual labour, bottlenecking our ability to analyze biological datasets.

Even in recent papers, biologists still rely on manually annotating tens of thousands of images from high-throughput experiments22–24. By creating general, high-performance, and labor-efficient methods with self-supervised learning, I hope to make computational analysis the first option that biologists use to understand their datasets, not one they arrive to after months of engineering efforts. Doing so would accelerate the speed and depth at which we discover biological insights from data.

2.1 Proposed: Creating methods for protein sequences, medical images, and drug-screening

My goal is to establish self-supervised learning as a general strategy for representation learning across biological domains. Different domains employ vastly different modalities, from sequence data to images to numerical observations. Even within the same modality, different applications require the extraction of different signals from data: in fluorescent microscopy, protein screens benefit from representations that ignore cell morphology12, but drug-screening requires sensitive representations of morphology25. My goal is to identify principles for designing self-supervised methods that target specific signals in data, so we can adapt our methods depending on the application. As few self-supervised methods for biology currently exist, a first step is developing methods across domains:

* In current research, I am developing self-supervised methods for biological sequence data, that exploit principles of conservation in evolution (see Section 3.1).
* In histopathology, autoencoders26 and neural networks pretrained on natural images27 are used to extract representations for diagnosing and subtyping cancer. Self-supervised learning may improve the quality of these representations by learning more biologically relevant features. One challenge is the large size of whole tissue slide images. I am interested in developing methods that limit computational load by learning representations using a sample of crops from these images, rather than the full image.
* In drug-screening, pharmaceuticals are prioritized based upon their effects on diseased cells in images. These applications rely on sensitive features that detect changes in morphology that may not be visible by eye25. As drugs can have incomplete penetrance, affecting only a proportion of cells, I am interested in self-supervised methods that operate on a distribution of cells.

2.2 Supporting: Existing collaborations for biomedical images and drug-screening

To support my expertise in creating self-supervised methods for biology, I have fostered collaborations that will provide me with the data required to develop methods for new domains. For medical imaging applications, I have collaborators at the [HOSPITAL INSTITUTE]. For drug-screening applications, I have a collaboration with a [BIOTECH].

**3. Discovery through interpretation and data integration**

Self-supervised learning is unbiased by prior knowledge, and models are trained on all available data (instead of just data we can reliably annotate.) Models may therefore develop features that capture both known and unknown biology within training data. By interpreting these features, and establishing correlations between biological data modalities, I hope to establish a framework for how unbiased deep learning methods can produce novel biological hypotheses.

3.1 Ongoing: Discovering functional features of biological sequences  
One test case is in applying self-supervised learning to learn features for intrinsically disordered regions (IDRs), regions of proteins that lack stable secondary or tertiary structure. IDRs carry out key functions, including protein-protein interactions and signaling28. A current research question is to identify features in the primary amino acid sequence of IDRs that are essential for function (because IDRs evolve rapidly, conservation-based methods like BLAST do not perform well.) For example, previous work has linked features like net electric charge to mitochondrial import29. As these features are identified on an ad-hoc basis20, our knowledge of features is not expected to be comprehensive.

My current work trains models using evolutionary homology. Given an IDR, the model is asked to predict which IDR is homologous from a large set where only one IDR is homologous. To solve this task, the model identifies evolutionarily conserved features that reflect conserved protein functions. I am developing interpretation methods to understand what these features are. By dropping out neurons at test time, we can identify which features affect the model’s confidence in predicting homology for IDRs. Using my interpretation methods, we can then visualize important features as motifs and as mutational scanning maps. My preliminary results suggest these models learn many features consistent with well-characterized features in literature, and I am working to identify novel features.

3.2 Proposed: Generalization across biological sequences and data integration

These self-supervised and interpretation methods are potentially general across sequence domains: an immediate extension is to apply these methods to genomic sequences. The next step is to understand how these unbiased features in proteins and DNA connect to function, which I plan to address by data integration approaches. For example, connecting features in protein sequences to microscopy images may identify protein features that drive compartmentalization in cells. Some proteins form condensates through phase separation. Defects in this behavior relate to neurological disease30; identifying protein features that drive phase-separation would let us understand how mutations disrupt these functions.

3.3 Supporting: Experience in exploratory clustering and statistics

I previously applied clustering and statistical methods to integrate nearly 400,000 microscopy images of yeast cells31, and I collaborated on work linking images of cells to drug treatments32. I expect to extend this experience in exploratory analysis, handling large datasets, and data integration to my goals in purposing unbiased deep learning for discovery through integration and integration of features.

**4. Representation learning under data constraints**

Getting deep learning models to generalize even when training datasets are small or homogeneous has practical implications in biology. Smaller labs may be limited in the amount of data that they collect: for example, as opposed to labs that employ high-throughput microscopes, smaller labs may image samples manually. I am interested in extending the benefits of deep learning to small scale experiments. Because deep learning methods are sensitive at quantifying biology, I hope to enhance the depth of insight across many experiments in many labs, not only big datasets generated by a few centralized groups.

4.1 Proposed: Identifying design principles for data-efficient representation learning

Many different methods have been proposed for more data-efficient deep learning, ranging from data augmentation33 to pre-training34 to rapid generalization through meta-learning35. My specific interest is in understanding how the way we train neural networks interacts with data constraints. Self-supervised learning has diversified the strategies employed for representation learning. Many different training methods are now available36, from discriminative classification tasks, to methods that generate synthetic data, to contrastive methods that are motivated by information theory. Generally, these methods are benchmarked with large training datasets like ImageNet. My hypothesis is that these methods will have different properties under data constraints. My preliminary results suggest that some methods are more robust for microscopy images: classification tasks become easier to “game” with less diverse datasets, where there are a few key features that models can overly rely on, whereas generative tasks still remain difficult and require attention to many features. I will explore connections with recent research in why over-parameterized models, where fewer training examples exist than parameters, can still generalize well: recent work has identified cases where more examples can lower model performance on test sets37.

4.2 Supporting: Experience in out-of-sample generalization

I have experience in designing datasets to test hypotheses about generalization. I published a microscopy image dataset designed to test if image classifiers are robust to small differences in the underlying conditions in which images are taken38, such as day-today variation in temperature, or microscope used.

1. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).

2. Kraus, O. Z. *et al.* Automated analysis of high-content microscopy data with deep learning. *Mol. Syst. Biol.* **13**, 924 (2017).

3. Lu, A. X., Zarin, T., Hsu, I. S. & Moses, A. M. YeastSpotter: Accurate and parameter-free web segmentation for microscopy images of yeast cells. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz402.

4. Sullivan, D. P. *et al.* Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* **36**, 820–828 (2018).

5. Jing, L. & Tian, Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. (2019).

6. Kosoy, E. *et al.* Exploring Exploration: Comparing Children with RL Agents in Unified Environments. *ICLR 2020* (2020).

7. Zhang, R., Isola, P. & Efros, A. A. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. in *2017 IEEE Conference on Computer Vision and Pattern Recognition* (2017).

8. Gidaris, S., Singh, P. & Komodakis, N. Unsupervised Representation Learning by Predicting Image Rotations. in *ICLR 2018* (2018).

9. Hénaff, O. J. *et al.* Data-Efficient Image Recognition with Contrastive Predictive Coding. *ICML* (2019).

10. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S. & Murphy, K. Tracking Emerges by Colorizing Videos. *ECCV* (2018).

11. Hendrycks, D., Mazeika, M., Kadavath, S. & Song, D. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. *NeurIPS* (2019).

12. Lu, A. X., Kraus, O. Z., Cooper, S. & Moses, A. M. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. *PLOS Comput. Biol.* **15**, e1007348 (2019).

13. Caicedo, J. C., McQuin, C., Goodman, A., Singh, S. & Carpenter, A. E. Weakly supervised learning of single-cell feature embeddings. *bioRxiv* 293431 (2018) doi:10.1101/293431.

14. Goldsborough, P., Pawlowski, N., Caicedo, J. C., Singh, S. & Carpenter, A. E. CytoGAN: Generative Modeling of Cell Images. *bioRxiv* 227645 (2017) doi:10.1101/227645.

15. Gildenblat, J. & Klaiman, E. Self-Supervised Similarity Learning for Digital Pathology. (2019).

16. Chen, L. *et al.* Self-supervised learning for medical image analysis using image context restoration. *Med. Image Anal.* **58**, 101539 (2019).

17. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20**, 723 (2019).

18. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. (2019).

19. Handfield, L.-F., Chong, Y. T., Simmons, J., Andrews, B. J. & Moses, A. M. Unsupervised clustering of subcellular protein expression patterns in high-throughput microscopy images reveals protein complexes and functional relationships between proteins. *PLoS Comput. Biol.* **9**, e1003085 (2013).

20. Zarin, T. *et al.* Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife* **8**, (2019).

21. Kraus, O. Z. *et al.* Automated analysis of high‐content microscopy data with deep learning. *Mol. Syst. Biol.* **13**, (2017).

22. Tkach, J. M. *et al.* Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nat. Cell Biol.* **14**, 966–76 (2012).

23. Weill, U. *et al.* Genome-wide SWAp-Tag yeast libraries for proteome exploration. *Nat. Methods* **15**, 617–622 (2018).

24. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science (80-. ).* **356**, eaal3321 (2017).

25. Caicedo, J. C., Singh, S. & Carpenter, A. E. Applications in image-based profiling of perturbations. *Curr. Opin. Biotechnol.* **39**, 134–142 (2016).

26. Ash, J. T., Darnell, G., Munro, D. & Engelhardt, B. E. Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. *bioRxiv* 458711 (2018) doi:10.1101/458711.

27. Kieffer, B., Babaie, M., Kalra, S. & Tizhoosh, H. R. Convolutional neural networks for histopathology image classification: Training vs. Using pre-trained networks. in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)* 1–6 (IEEE, 2017). doi:10.1109/IPTA.2017.8310149.

28. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).

29. Bauer, N. C., Doetsch, P. W. & Corbett, A. H. Mechanisms Regulating Protein Localization. *Traffic* **16**, 1039–1061 (2015).

30. Elbaum-Garfinkle, S. Matter over mind: Liquid phase separation and neurodegeneration. *Journal of Biological Chemistry* vol. 294 7160–7168 (2019).

31. Lu, A. X. *et al.* Integrating images from multiple microscopy screens reveals diverse patterns of change in the subcellular localization of proteins. *Elife* **7**, e31872 (2018).

32. Yang, K. *et al.* Improved Conditional Flow Models for Molecule to Image Synthesis. *arXiv* (2020).

33. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **6**, 60 (2019).

34. Erhan, D., Courville, A., Bengio, Y. & Vincent, P. Why Does Unsupervised Pre-training Help Deep Learning? 201–208 (2010).

35. Vanschoren, J. Meta-Learning: A Survey. *arXiv* (2018).

36. Liu, X. *et al.* Self-supervised Learning: Generative or Contrastive. *arXiv* (2020).

37. Nakkiran, P. *et al.* Deep Double Descent: Where Bigger Models and More Data Hurt. (2019).

38. Lu, A. X., Lu, A. X., Schormann, W., Andrews, D. W. & Moses, A. M. The Cells Out of Sample (COOS) dataset and benchmarks for measuring out-of-sample generalization of image classifiers. *NeurIPS* (2019).