**Research Statement**

Alex X. Lu  
Department of Computer Science  
University of Toronto

alexlu@cs.toronto.edu

Advances in high-throughput microscopy technologies mean that labs can now routinely generate tens of thousands of biological images in a single day. This data creates computational challenges, arising from the practical context and motivations of big biological image data:

1. *How do we extract relevant information about biology from images?*  
   Different labs have different research questions. Even with the same images, different labs are interested in different aspects of the image, to the point that information that answers one question will hinder another by adding noise.
2. *How can we discover novel classes and behaviors of cells in big data?*  
   Some samples in high-throughput image screens may exhibit phenotypes or behaviors previously unknown to us. Finding these cases creates novel hypotheses that advance our understanding of biology and disease.
3. *How do we design methods that generalize out-of-sample and are robust to data constraints?*   
   Day-to-day variation in temperature and humidity, and inherent stochasticity in cells, means that the same experiment conducted twice will produce different distributions of images.

My goal is to create novel machine learning methods that address these questions. Specifically, I am interested in *self-supervised representation learning*, *data integration* and *interpretation*, and *robust representation* methods. In the remainder of this document, I will relate these research directions to computational challenges in biological image analysis, and outline goals and future directions.

**1. Self-supervised learning of targeted representations without manual effort**

Analyzing microscopy images requires extracting relevant information. To understand protein biology, we extract features measuring the location of proteins within imaged cells. These features benefit from ignoring noise caused by variation in morphology and size, which causes proteins contained in the same functional locations (e.g. the nucleus) to look different between cells. In contrast, other applications, that discover pharmaceuticals based on their effects on diseased cells, extract features that measure cell morphology. Because different questions require different information, researchers will engineer and curate features to target components of images relevant to their research. Alternatively, neural networks can learn relevant features given large training datasets labeled with the biology of interest. Both options require extensive manual labour, either in feature engineering or data annotation, bottlenecking image analysis.

To solve this problem, I have created self-supervised methods. Self-supervised learning teaches neural networks transferrable skills and representations through puzzle-solving. By solving autonomous tasks like colorizing greyscale images or guessing how much an image has been rotated by, models learn about the natural world. I created a method designed to teach models about protein biology using an inpainting task: given an image of a cell where a specific protein is fluorescently tagged, the model synthesizes an image of how the same protein should look like if it were expressed in a second different cell. The model learns representations that outperform previous ones in downstream analysis, suggesting that by posing the right task, neural networks can learn to extract specific biological content without any manual effort.

1.1 Future work: Self-supervised learning as a general strategy for biological data

My previous work shows that self-supervised learning can learn superior representations of biology with less manual effort. Moreover, because these models are unbiased by prior expert knowledge, and trained on all available data (instead of just data we can reliably annotate), they are useful for exploratory and discovery applications. As these properties would be useful across many biological domains, my goal is to establish self-supervised learning as a general strategy across biological data applications.

I am currently working on a self-supervised method that learns conserved properties of highly-diverged intrinsically disordered regions in protein sequences (see Section 2). In addition, I have fostered several collaborations in imaging applications related to health. For medical imaging applications, I have several collaborators at the [HOSPITAL INSTITUTE]. For drug-screening applications, I have collaborations with pharmaceutical companies, including [BIOTECH]. I expect these collaborations will provide me with the data and expertise to develop self-supervised methods for imaging in healthcare.

**2. Discovering biology with deep learning**

While deep learning methods have shown superior performance in automating routine tasks in biology for which we can reliably annotate data, like classification or segmentation, comparatively little is known about how deep learning can be used to discover unknown classes that we cannot label *a priori*. Discovering these classes is instrumental in advancing hypotheses, as they may reflect novel functions, mechanisms, or cell states.

My goal is to develop self-supervised learning as a strategy for discovering new functional features in biology. Since self-supervised learning is unbiased, models have the potential to develop features that capture both known and unknown biology. One timely test case is applying self-supervised learning to learn features in intrinsically disordered regions (IDRs), regions of proteins that lack stable secondary or tertiary structure. IDRs are widespread in eukaryotic proteomes, and carry out diverse functions related to their flexibility and exposed surface area, including protein-protein interactions and signaling. A current research question is to identify features in the primary amino acid sequence of IDRs that are essential for function (because IDR sequences are poorly conserved, classic methods like BLAST do not perform well). For example, previous work has linked properties like the net electric charge to import into the mitochondria, or the presence of certain motifs to interactions with regulatory proteins. As these features are identified on an ad-hoc basis, our knowledge of functional features in IDRs is not expected to be comprehensive, so there may be unknown features that have yet to be elucidated.

My current work uses contrastive learning with evolutionary homologues in learning self-supervised features for IDRs. Given an IDR, the model is asked to predict which IDR is homologous (from the same protein in a different species) from a set of IDRs where one IDR is homologous and the rest are randomly drawn from the proteome. To solve this task, the model needs to learn features that are evolutionarily conserved in IDRs, which are usually relevant to function. I am working on methods to interpret these features: by dropping out neurons, we can identify which features significantly affect the model’s confidence of predicting homology for any given IDR, and then we can characterize these features by aligning regions of proteins neurons are most activated for to produce motifs and by visualizing mutations that most alter the activations of neurons. My preliminary results indicate that these IDR models learn features consistent with well-characterized features in literature, like nuclear localization signals or CDK phosphorylation sites, and I am currently working to associate other potentially uncharacterized features with protein functions.

2.1 Future work: Data integration across biological modalities

Developing self-supervised learning methods (Section 1) will produce unbiased representations across biological data domains, and my work in interpretation (Section 2) provides a means to interpret features from these representations. A natural extension

**3. Biological image datasets for measuring out-of-sample robustness**

3.1 Future work: Robust representation learning methods