# Single-Cell Analytics via Machine Learning Language Models

**John Kalantari** and **Michael A. Mackey**
Department of Biomedical Engineering
University of Iowa, Iowa USA

## Abstract

In this paper we propose the application of language models based on various machine learning algorithms for the purpose of classifying the composition of time-lapse microscopy image sequences, aiming to contribute a new perspective to analyzing the spatio-temporal evolution of single-cell properties in an incremental manner. The obtained results encourage further investigation and suggest that language models may be beneficial for a range of biological sequence analysis problems.

## 1 Unsupervised Learning for Single-Cell Analytics

Elucidating the relationships between molecular states and macroscopic properties of biological processes is one of many ambitions in *computational systems biology*. Despite substantial knowledge of cellular dynamics obtained from bulk population-level studies, the heterogeneous nature of many cell-lines requires single-cell profiling techniques to quantify the dynamic processes from which further insights can be derived. In order to close the computational gap in computational systems biology, we must strive to build modular, data-driven models of complex biological processes from the molecular level up to the entire organism. This requires an integrative approach that focuses on the interplay between gene expression and the higher-order properties associated with cellular development and growth. To achieve this, we explore the notion of building compositional models using unsupervised learning language models. We evaluate the feasiblity of this compositional approach by constructing computational models at the single-cell level using data obtained from live-cell imaging experiments. Live-cell imaging is an essential tool for deciphering the dynamics of individual cells and cell populations by capturing and correlating cell morphology and gene-expression profiles at multiple scales of resolution. With this in mind, we aim to draw attention to a set of deep learning language models that were developed for *sequence learning* and repurposed for analyzing and describing single-cell dynamic behavior from time series data in an incremental manner. Ultimately, insights obtained from these methods can be exploited to refine the description of cellular processes, leading to integrated cellular/gene-expression models.

## 2 Experiments/Results

Many methodologies orignating from linguisitics and information theory have found their application in biological knowledge discovery. Many popular language modeling approaches utilize the $n$-gram approximation where the probability $p(w_1^N)$ of a sequence of words $w_1^N$ is factorized as $p(w_1^N) = \prod_{i=1}^{N} p(w_i|c_i)$ so that only the $n-1$ preceeding words, or context $c_i := w_{i-n+1}^{i-1}$, are used to estimate the probability of the word at position $i$. In the case of live-cell imaging data, the probability distribution for the sequence of discrete cellular events approximating the spatio-temporal dynamics of a single cell is considered an extended analogy to the distribution for a sequence of words/symbols used in formal language modeling. Here, we demonstrate the use of a new Artificial General Intelligence (AGI) framework called SYNACX for predictive and adaptive modeling of complex biological system behavior directly from time-lapse microscopy data sequences. The performance of this computational framework is subsequently evaluated against language models based on classical neural network architectures including, multi-layer Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) as well as more recent deep learning architectures including the Neural Turing Machine and End-To-End Memory Networks. The proposed framework uses an underlying statistical model based on Bayesian Nonparametric priors in which the multi-scale nature of observed spatial and temporal patterns are abstractly and concisely represented as topological structures adopted from combinatorial topology. We present initial results of two autonomous tasks performed on unlabeled live-cell imaging data from experiments on MDA-MB231 metastatic cells performed on the Large Scale Digital Cell Analysis System (LSDCAS), namely cellular event identification and large-scale temporal behavior recognition. We demonstrate the increase in accuracy and precision over current expert methods, the efficient asymptotic computational complexity of the proposed learning algorithm and its suitability for real-time predictive analytics compared to other deep learning approaches.