DISSERTATION

STOCHASTIC MODELING TO EXPLORE THE CENTRAL DOGMA OF MOLECULAR BIOLOGY AND TO DESIGN MORE INFORMATIVE SINGLE-MOLECULE, LIVE-CELL FLUORESCENCE MICROSCOPY EXPERIMENTS

Submitted by

William Scott Raymond

School of Biomedical Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2024

Doctoral Committee:

Advisor: Brian Munsky

Asa Ben-Hur
Christopher D. Snow
Diego Krapf
Timothy J. Stasevich

ABSTRACT

STOCHASTIC MODELING TO EXPLORE THE CENTRAL DOGMA OF MOLECULAR BIOLOGY AND TO DESIGN MORE INFORMATIVE SINGLE-MOLECULE, LIVE-CELL FLUORESCENCE MICROSCOPY EXPERIMENTS

Despite being described nearly a century ago, the Central Dogma of Molecular Biology still harbors many intricacies and mysteries that scientists have yet to unravel. With the convergence of many multidisciplinary scientific advances such as stronger computing power, next-generation sequencing, machine learning, and single-cell and single-molecule experiments, cellular biologists have never had more investigative power. These complex methods often are used in tandem– necessitating a closer relationship between computational biologists, computer scientists, and bench-top experimentalists. As practice of this emerging dynamic, my corpus of work spans multiple areas within computational and quantitative biology with the goal to facilitate better computational tools to interpret and design experiment. For my main work at Colorado State University, I have developed the open source Python package "RNA sequence to Nascent protein simulator," rSNAPsim, to simulate Nascent Chain Tracking experiments and used it as a backbone for an entire experiment simulation pipeline to check experiment design feasibility. The rSNAPsim software provides start-to-finish capabilities for model design, model fitting, and model selection so that experimentalists can fit a mechanistic model to the Nascent Chain Tracking single-mRNA translation experiments. Along with this main work, I have provided computational modeling efforts on live-cell data on the first two steps of the Central Dogma, DNA transcription and mRNA translation. For the final entry in my corpus, I have used my interdisciplinary skills acquired at CSU to do machine learning based ncRNA riboswitch classification and discovery within the human genome; This work provides the broader scientific community with a starting point for searching

for this important secondary structure within humans, where it has not been described as of time of writing.

ACKNOWLEDGEMENTS

No science happens in a vacuum; I would like to thank the current and former members of the MunskyGroup lab —the camaraderie, helpfulness, selflessness, and kindness is unparalleled. I would like to give a special thanks to Dr. Zachary Fox, Dr. Lisa Weber, Dr. Huy Vo, and Mohammad Tanhaemami all of whom where instrumental to helping me when I first started in the lab as an undergraduate. They fostered such an inviting atmosphere and were always more than happy to set time aside to answer random questions from their junior researcher.

Thank you to all my collaborators, especially Dr. Linda Forero-Quintero and Dr. Luis Aguilera for allowing me to work on their manuscripts, especially Linda who trusted me to analyze her data. I would like to thank the experimentalists in Dr. Timothy J. Stasevich's who lend us their data, giving us a tangible affect on the world; Without them who knows what basement computational modelers would be sequestered in while we work in 1s and 0s.

Brian has been the par excellence of principal investigators, mentor, and friend over the past 8 years I have worked for him. Without his kindness and patience I would not have completed this Ph.D. Brian's enthusiasm for the field and his mentorship have been endless inspiration. I aspire to continue working with his zeal and would consider my self lucky to mimic a mere fraction of his ability.

To my friends who kept me sane, Rob, Janine, Isaac, Tanner, Jacob, Mark, Wes, Nick, Michelle, thank you for always being there to cheer me up and for unwavering moral support.

To my parents, thank you for the opportunity, support, and love to pursue my curiosities.

To Jenna, the hardest working person I know, your work ethic is literally one of a kind.

For conciseness, acknowledgement sections for each published manuscript included in this dissertation is moved to this section.

Acknowledgements for "Computational design and interpretation of single-RNA translation experiments"

The authors thank Kenneth Lyon and members of the Munsky lab for their feedback on the presented analyses and for their testing of the rSNAPsim software.

Acknowledgements for "Computational design and interpretation of single-RNA translation experiments"

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Acknowledgements for "Using mechanistic models and machine learning to design single-color multiplexed nascent chain tracking experiments"

WSR, and BM were supported by the NSF (1941870) and National Institutes of Health (R35GM124747). JD was supported by the National Institutes of Health (1R01AI168459-01A1). Special thanks to Dr. Hamid Chitsaz as this paper started as a student project in his CS548 - Bioinformatics class. Additional special thanks to Dr. Jeffrey Wilusz as the initial idea for this was broached as a class discussion in MIP 543 - RNA biology.

Acknowledgements for "Live-cell imaging reveals the spatiotemporal organization of endogenous RNA polymerase II phosphorylation at a single gene"

We thank all the members of the Stasevich and Munsky labs for helpful discussion and suggestions, especially Luis Aguilera for help with the spot minima analyses. T.J.S., T.M., and M.S. were supported by an award to T.J.S. from the NIH (R35GM119728). B.M. and W.R. were supported by an award to B.M. from the NIH (R35GM124747). L.S.F.Q. was supported by the W.M. Keck Foundation. T.H. was supported by an award from JSPS KAKENHI JP17K17719. H.K. and T.H. were supported by an award to H.K. from JSPS KAKENHI JP17H01417 and JP18H05527.

DEDICATION

For Marisa, who has the patience of a saint.

ETHICS STATEMENT

Science in the past centuries has propelled humanity into an era of unparalleled technological achievement, saving billions of lives from starvation, curing millions from previously intractable diseases, fostering an unprecedented era global communication and collaboration. However, this bright sun casts a long and dark shadow; Science is not divorced from its material surroundings and society it operates in. Its blessings are unevenly distributed, stolen, hoarded, hidden, and abused for ulterior motives. Each new technological advancement can and has been abused to harm people —machine learning lauded for diagnosing cancers and ordering chaotic data is often used to predict recidivism, to aggressively target marketing, or prejudicially reject resumes. Bankrupt bionic eye companies leave their early adopters with unsupported, failing technology implanted in their bodies. Fentanyl acts as a double edged sword, providing irreplaceable pain management in hospice while also killing millions as the end-stage of a manufactured for profit epidemic. Behind these inventions are scientists, scientists with the best of intentions working in an unethical system.

Those are just modern implementations in science; history is littered with famous cases of unethical experiments; I won't list them here as a comprehensive list would be as long as this dissertation. Investigation, regulation, and just recompense to victims always lags well behind the application of the actual unethical research —if justice ever comes at all. Science likes to cloak itself in objectivity, to pretend it is free from human bias as a pure truth. This perception is the mainstream, when in reality, every scientist alive has internal and external motivations beyond a pure pursuit of knowledge. Research topics are beholden to profitability to receive funding (see neglected tropical diseases). Scientists are pressured by funding requirements, competitive tenure, positive result publishing, and publish or perish mentality. These systemic pressures let unethical practices creep in to any field of Science and even with the most cognisant scientist. A vigilant investigator should not be cursorily aware of these influences, but intimately know their own biases and outside pressures. Awareness is the first step to help mitigate some of these biases and this awareness should be assessed at every step during research. Hopefully, nothing stated here is new

to any reader. My intention is to restate and reaffirm ethical considerations as a first step practice to scientific research here at the front of my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The collected papers here seem disparate at first glance, however, there is a through-line of computational, quantitative modeling of biological systems that ties these works into a cohesive set. This work spans three areas within cell biology: transcription, translation, and RNA regulation, and there is a constant need for rigorous computational effort behind the curtain to unwind these diffraction limited processes. I have had the immense pleasure to collaborate with amazing experimentalists over my time here at Colorado State University and provide modeling input to their papers —even if the consequence of this is a slight meandering in topics discussed in my dissertation. Looking back, I am quite happy with the breadth of what I studied at Colorado State University as computational modelling exposed me to many different areas of biological study where I eventually found my passion in RNA biology.

1.1 Computational modeling

When someone is lost on an unfamiliar road, they reach for a favoured map app (or reach into their car-seat pocket and pull out a battered roadside atlas pre-2007) and use it to figure out where they are. But what are they fundamentally doing? They are comparing their surroundings with someone else's previous work: comparing their observations with a previously laid consensus. If their guide (the map) is worth its salt, it has been updated with recent information (this restaurant changed its name in 2022) without straying into distracting asides (here is a map of every dandelion growing through the sidewalk). Just like the roadside maps of old, models are computational tools that simplify observations, provide baseline consensus, guide scientific investigation and are constantly revised and updated with new information by hidden computational workers. These traits—definitive consensus, updated information and streamlining—make computational models as useful to researchers as maps are to lost travelers. These tools simplify observations, provide baseline consensus, and guide scientific investigation, all while being constantly revised and up-

dated with new information by hidden computational workers. When computational modeling intersects with the diffraction limited, the microscopic, the atomic, data is not so straight forward. The need to create careful curated experiments that get interpretable data that are even capable of informing models leads to a close partnership between modelers and experimentalists that only intensifies with each passing year in the field.

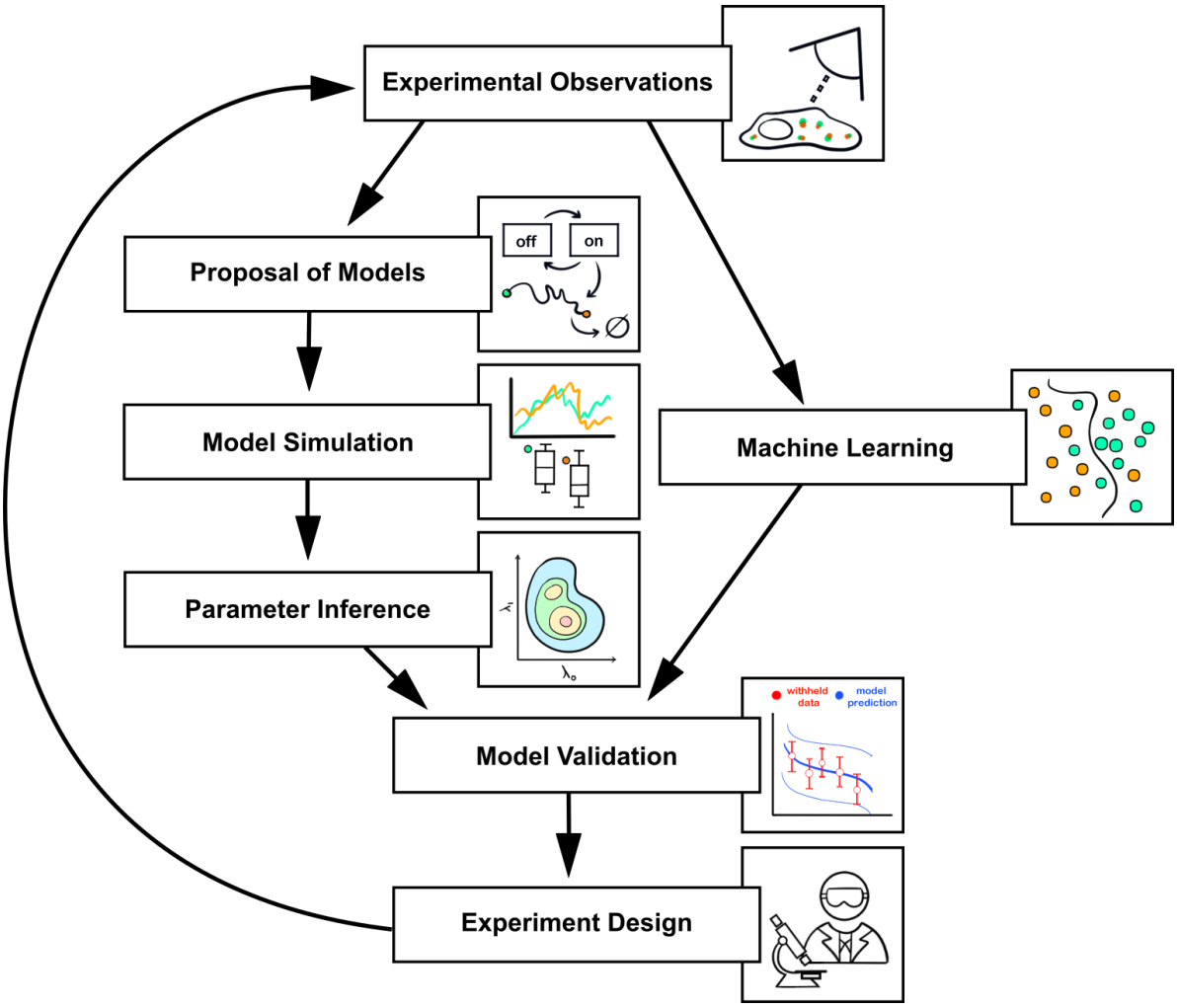


Figure 1.1: General modeling workflow

"Computational" is the operative word within computational modeling. Many of the methods within computational modeling were mathematically described prior to the mid-20th century advent of available computing. Neural networks were proposed in 1943 and 1949 by Pitts and Hebb

respectively, decades before they would be successfully implemented and nearly 80 years before the deep learning explosion enabled by better GPU technology [1, 2, 3].The foundations for Monte Carlo simulations of Markov processes were proposed by Joseph Leo Doob in 1945 long before their popular formulation by Gillespie in 1976 [4]. With the exponential growth of computing power, many of these impossibly laborious methods to do by hand were now available to the broader scientific community.

Computational modeling as a topic has a malleable implementation and varied focus, but in general, one will find some combination of the topics in the workflow to the right. Typically the computational modeling loop starts with some experimental observation or research question necessitating a deeper mathematical and mechanistic understanding that cannot be gleaned from experiments alone. The next step involves constructing a pool of models to consider for selection. Several classes of models are proposed that may replicate the data; for best practices, models ranging in complexity are used and at least one simplest case model is considered in the model pool. During this process the models will be explored and simulated intensively to observe unique phenomena. Models that do not replicate the data or provide unrealistic phenomena are often discarded at this stage (although some scientists may keep them in the model pool as a negative confirmation during parameter inference). With a group of models that can recapitulate the initial observations, the next step is to perform parameter inference to describe the most likely parameter set for each model, and their parameter ranges if possible. The most informative model out of the pool is selected with an information criterion (Bayes or Akaike). The selected model then can be validated by testing its output versus withheld data, or using it to predict experimental outcomes and comparing. A validated model only lasts as long as the next experiment supports this model, just how the map from the previous analogy is only useful as its recreation holds.

Negative confirmation with a class of simpler models during parameter inference is a crucial step to determine the evidence contained in the data for a more complex model. When a simple model recapitulates a true complex system, it suggests that a different experiment design or more data is needed if one truly believes a complex system is at play. We can preview Chapter 3.1 here,

where model selection lead to a clarifying experiment that shed light on the underlying process in more detail. A live cell transcription fluorescence reporter system was used to collect video data at a one minute frame rate —a frame rate more than sufficient to capture waves transcription. However, from the perspective of the data and model selection, the rate of RNAP2-SER5 phosphorylation of the CTD tail appeared to be instantaneous as it was happening on a scale much faster than a minute. The model selection threw away any more complex model that included a transition rate between RNAP2 and consistently selected a simpler model included in our model pool. If our team was not interested in this rate, then our model, our "map,"with the simpler would have been sufficient. However, since we were interested in this rate, this model selection outcome suggested a new experiment with a much faster frame rate on the order of seconds which was performed and gave a better clarification on this phosphorylation rate. This example is a perfect illustration of the cyclic nature of experimentation and modeling, ratcheting knowledge forward as new data updates models and models suggest new experiments until the desire for knowledge is satiated. With this brief and general explanation of computational modeling we can turn to what I modeled and why.

1.2 Central Dogma of Molecular Biology

Information processing of bio-polymers is critical to all known cellular life's persistence. Information storage, retrieval, compression, transfer, editing, and usage are all represented within life as we know it. Information in cellular life is stored in the form of nucleotide bases in nucleic acid chains, DNA or RNA. The canonical direction of "information retrieval"is for DNA bases (A, T, G, C) be read by RNA polymerases to messenger RNA (A, U, G, C) in a process called transcription. Newly transcribed mRNAs are processed, matured, and proofread and then translated by ribosomes into functional proteins during translation —these proteins are then free to perform their function(s) throughout the cell. Thus the Central Dogma of Molecular Biology is defined:

"DNA makes RNA, and RNA makes protein"- Francis Crick, 1957

While called a dogma, there are multiple exceptions and nuances to the direction of informational flow, the selection of information used, and the quality checking of information signals. With life, comes a radically large number of exceptions. Retrovirus reverse transcriptases convert RNA to DNA, myriads of viruses are solely RNA based —eschewing DNA entirely, prions pass on their errant "information" to other susceptible proteins, proteins regularly modify patterns like methylation on DNA, mRNA, and other proteins to change behavior, chromatin remodels in response to stimuli choosing which information is available, noncoding-RNA never makes proteins and instead performs their functions as RNA structures, information is compressed into editosomes (kinetoplastid mDNA) or into multiple reading frames (viral frameshifting), and sometimes genes encode noncoding introns within coding exons as a package deal. Even the final end products of information transfer, proteins, don't have such clear cut functions. Promiscuous proteins serve multiple functions. Inherently disordered proteins vex the previously established globular structure-function paradigm. Proteins come with a staggering array of isoforms and post-translational modifications as well. This is a small list of exceptions, or nuances depending who you ask, to the Central Dogma. I have listed them here to give the reader a sense of the chaos commonplace within life and how every step from DNA to protein includes some heterogeneity. Experimentalists, computational biologists, and biochemists are left to sort out the dazzling array of interactions, leaving the field just as, if not more, rich in research topics as it was in 1957.

Cellular gene expression acts in a heterogeneous manner. Low-copy number, discrete genes express through an expansion of material. For example, a single gene copy gene will have 2 copies of itself in a diploid cell; Those two genes may turn on or off in response to stimuli, creating anywhere from zero to hundreds of mRNAs, which may go on to produce zero to tens of thousands of proteins. As the Central Dogma plays out, two cells next to each with 2 identical strands of DNA describing the same gene exposed to the same concentration of stimuli in the same environment can have one cell do nothing, while the other produces tens of thousands of new proteins. The noisy single molecule interactions dictating the on/off state of two single gene copies is amplified

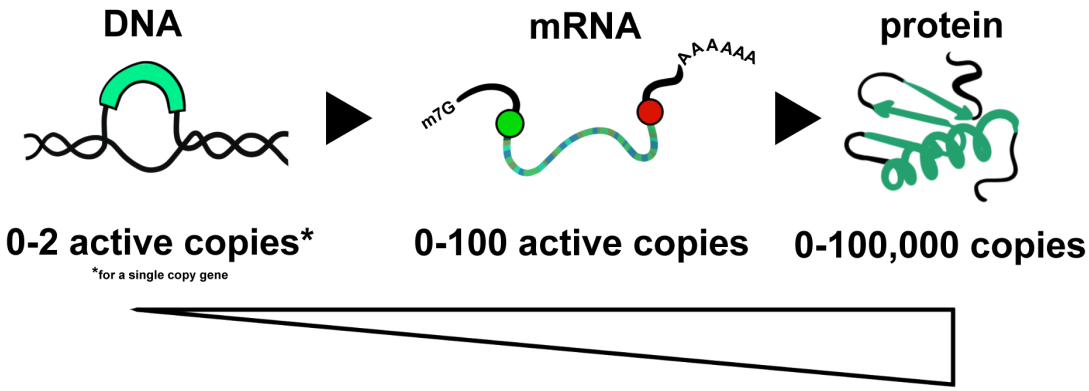


Figure 1.2: Expansion of bio-molecules

and dictates the entire cellular response. This crucial diversity of responses across cell populations helps protect cellular populations by generating divergent phenotypes in response to stimuli in their changing environment, hopefully increasing potential survival of some of the population. Trying to study these cell response distributions necessitates models that capture this variability and stochasticity that replicate the observed probabilistic behavior.

1.3 mRNA translation

The main publication produced by my dissertation work analyzed the second step of the Central Dogma, mRNA translation. mRNA translation is the process of converting the ubiquitous bio-polymer mRNA into functional proteins via cellular machines called ribosomes. Despite such a critical role in the cell, much about its inner workings are still confound the scientific field; The lack of definitive knowledge is partly due to the complexity of the process, cell heterogeneity obscuring dynamics, and partly due to a lack of technology that can provide high-quality information about mRNAs in a single molecule manner in live cells. That is not to say there are not extremely clever methods of study already employed (we will cover those methods in the next sections) the problem is these methods instead can only provide a partial picture due to inherent averaging of the methods. These methods also require copious processing, analyses, and skill to perform. To return to the complexity conundrum, mRNAs are subject to a staggering diversity of effects: mRNA circularization, tRNA abundance, codon selection, ribosome pools, subcellular location, cellular

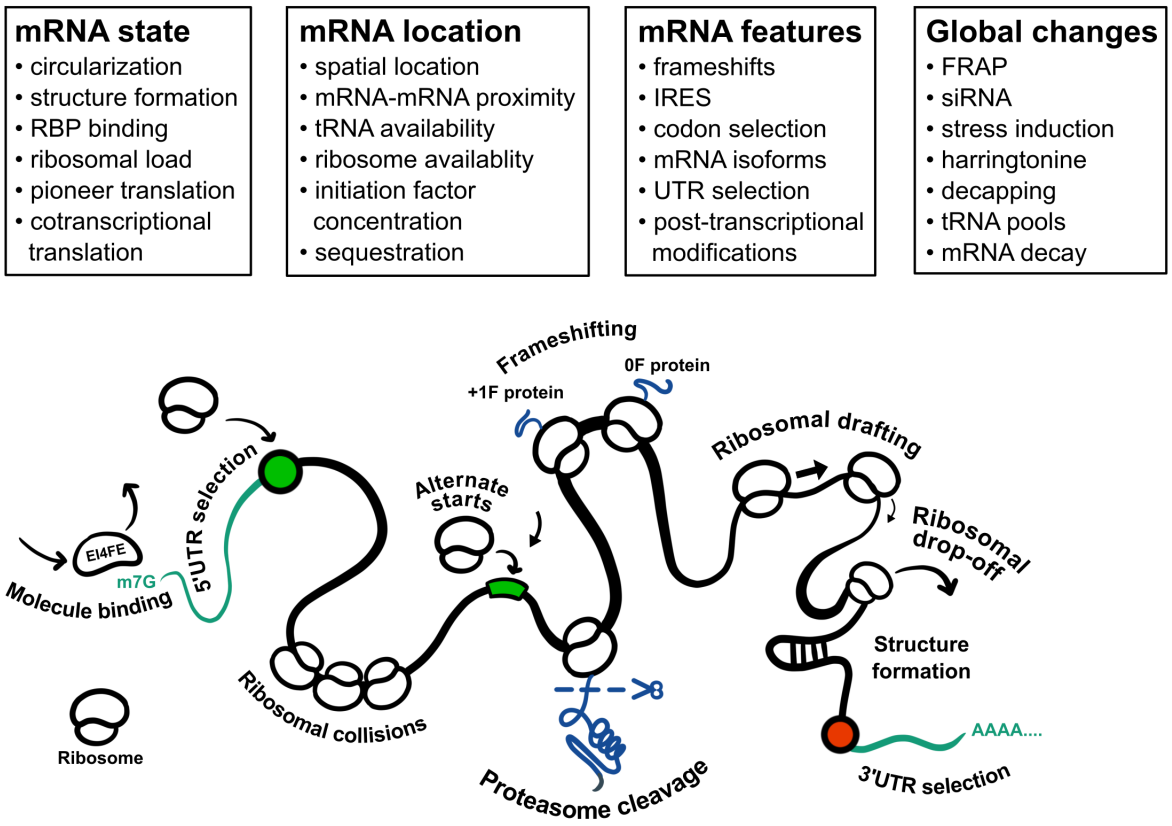


Figure 1.3: Short list of processes that affect mRNA translation dynamics

stress, RNA binding proteins, secondary structure formation, UTR selection, post-translational modifications, miRNA binding, decay processes and many, many more. This list's effect on translation can only be probed by two observations in a given experiment: how fast does an mRNA create protein? how often does an mRNA make a protein? In other words, how many ribosomes bind (initiation) and how fast do they go (elongation). To even attempt to unravel a problem such as codon selection's influence on elongation speed, these other factors have to be considered and weighed simultaneously within any experiment —an experiment that often can only quantify one facet at a given time on a given mRNA. Additionally, many of these effects are heterogeneous and noisy across species, tissue types, and subcellular localization. The insights gained in one cell line need not apply to any other much to the chagrin of scientists; leaving the use of model systems or cell lines necessary but always scrutinized.

Along with this list of mRNA dynamics, comes a concomitant and expansive group of disease states when these dynamics go awry and which scientists wish to study. As we will talk in the upcoming chapters, the main focus of my dissertation has been extending the experimental technique of Nascent Chain Tracking (NCT) through computational experiment design. This technique and my computational extensions are most applicable for examining ribosomal elongation. Deregulated elongation is the cause of many of the most devastating diseases in humans: Amyotrophic lateral sclerosis (ALS), PolyQ diseases (Huntington's Disease, some spinocerebellar ataxias, spinal and bulbar muscular atrophy), various cancers, and mitochondrial diseases such as mitochondrial encephalomyopathy lactic acidosis and stroke-like episodes (MELAS) or myoclonic epilepsy with ragged red fibers (MERRF). Each of these have some translation deregulation due to expansive repeats in coding regions, tRNA or tRNA modification loss or over-expression, or changes to elongation factors. Diseases like these could be examined in a live-cell single molecule context better with the modeling work I have done. Elongation is arguably the hardest process to study, since the changes during disease are minute and often only affect a small region of the mRNA; Detecting those changes requires specialized experimental techniques to measure and ameliorating some of those difficulties has been the my main motivation here at CSU. With that, lets discuss briefly some of the methods to measure mRNA translation before focusing on NCT.

1.3.1 Experimental methods to study mRNA translation

There is a plethora of specialized laboratory techniques for probing mRNA dynamics as of writing in 2024. The experimentalists tool box has never been larger. Broadly, the current methods can be broken into four quadrants of two options: Live cell vs fixed and direct observation vs reporter measurement.

Reporter measurement here is defined as measuring the output proteins as a proxy for mRNA activity. As illustrated by the table above, there is an abundance of bulk snapshot data from mRNAs; flow cytometry with direct mRNA tagging [5, 6], ribosomal footprinting [7, 8], RNA-seq [9, 10, 11], and proteomics [12, 13] all provide high quality bulk data of a single cellular time

Table 1.1: A non-comprehensive list of experimental methods to study mRNA as of 2024

	Live cell	Fixed cell
Direct	Nascent chain tracking (NCT) Flow cytometry (Branched DNA or MS2 GFP tagging)	smFISH MERFISH RNA-seq scRNA seq Ribo-seq / ART seq SUnSet Puro PLA RIBOmap
Reporter	Fluorescence microscopy with fluorescent proteins Flow cytometry (with fluorescent proteins)	Proteomics Pulsed Silac PUNch P BONCAT FUNCAT-PLA

point. Pulsed SILAC, PUNch-P, BONCAT / QuaNCAT allow quantification of the most recently translated proteins, and therefore the recently active state of mRNAs within a cell [14, 15]; However, they still are only providing one snapshot from a cell and cannot observe the same mRNA in time series. Similarly, FUNCAT/SUnSET tag nascently translating peptides for examination in fixed cells [16, 17], providing a readout of active translation. Two of the most powerful emerging techniques are single-cell RNA-seq (scRNA-seq) [18, 19] and RIBOmap [20]. Both of these provide single-cell spatially resolved information about RNA species (scRNA-seq) as well as active translation events (RIBOmap). Despite their immense power and widespread adoption, these bulk snapshot techniques can only collect a single time point of the a given cell's transcriptomal state as they require fixation and lysation of said cells —a limitation ameliorated by Nascent Chain Tracking (NCT).

1.3.2 Nascent chain tracking (NCT)

NCT uses a dual fluorescent labelling system where an RNA tag such as MS2 or PP7 is used outside an mRNA's coding region, and a repeat fluorescent epitope region is included within the coding region to label growing nascent polypeptide chains [21, 22, 23, 24, 25, 26, 27]. The output

of this technique provides an experimentalist with a way to track mRNAs in one channel regardless of their translation state, as well as a fluctuating signal of live-cell in-real-time protein synthesis of a single mRNA with a second channel; These signals appear in the cell as a single diffraction limited spot within a fluorescence microscope co-localized across two color channels. Particle tracking and image processing of NCT videos recaptures the local intensity fluctuations of a given mRNA and its translational state. NCT has been used to great effect in the past eight years to study dynamics such as mRNA frameshifting [28], mRNA heterogeneity [29], toxic codon repeats [30], codon usage [31], cellular stress response in mRNAs [32], mRNA decay [33, 34], IRES mediated translation [35], pioneer rounds of translation [34], stress granule formation [32, 36], mRNA translation in a spatial context [37], microRNA mediated dynamics [38] and several more dynamics of interest. A more complete and recent review of NCT applications can be found in Cialek et al [21].

Despite the being the only current technique which can allow an experimentalist to gather data from a single mRNA over self-consistent time spans in a live cell, NCT is heavily limited by its complex and demanding implementation; NCT requires specialized mRNA construct designs to be loaded or inserted endogenously into cell lines, custom DIY fluorescent microscopes, copious amounts of microscope time to acquire a small amount of data relative to bulk methods, and complicated often home-brewed computational analyses. These barriers to entry are unacceptable for a technique which is essential to unravel heterogeneous mRNA dynamics that fluctuate based on mRNA, biochemical, or cellular states.

Of the 16 papers using NCT in the literature as of Spring 2024, 88% use computational modelling as a complement to their evidence. Scopes of these models vary greatly from simple ballistic models [26], systems of Ordinary Differential Equations (ODEs) [36, 33, 24, 32], signal decomposition techniques [29], full spatial reaction-diffusion models, and codon-dependent totally asymmetric simple exclusion processes of mRNAs with multiple states [28, 35]. Across these papers, the authors have conducted their computational analyses independently. Some of these methods could be combined into a unified software to aid experimentalists. In Chapter 2.2, I will talk about the rSNAPsim (Rna Sequence to NAscent Protein SIMulator). rSNAPsim aims to provide

experimentalists using NCT with an open source software package to ease some of the analytical burden of using such an intensive method. rSNAPsim to allows its users to generate, simulate, explore, analyse, and fit mRNA translation models. This software package has been a focus of my dissertation and computational modeling work while at CSU and its the fundamental basis for the published work done in Chapter 2.1 and 2.3. Let's introduce briefly the main two models contained in the rSNAPsim.

1.4 Computational models for mRNA translation and DNA transcription

Two main classes of models are used in my dissertation work: linear systems of ODEs based models and Totally Asymmetric Simple Exclusion Processes (TASEP) models.

1.4.1 ODE based approaches

Linear systems of ODEs are the first go to mathematical model for describing species changes over time within a chosen biological processes. These systems are deterministic, easy to produce and quick to solve —but often limited in their ability to capture all the nonlinear behavior a modeler may wish to reproduce. Additionally, they rely on two key approximations when being used to describe molecular behavior. 1. Spatial effects are neglected, that is to say the model is in a well mixed environment and species have no diffusion based reactions or spatial relations. 2. Species are represented with continuous functions, that is to say there are no discrete molecule counts and the species are treated as well mixed concentrations. This species approximation is only valid in larger number of molecule counts, on the order of hundreds of species; Any lower and stochastic noise may be a large factor in molecular behaviour. Despite these two limitations, linear ODEs are invaluable mathematical modelling tools that have endless uses and are arguably the most ubiquitous mechanistic modelling tool in biology.

Linear ODEs for biological processes of $\mathbf{x}$ species are often formulated with two matrices: Stoichiometry ($\mathbf{S}$), Propensity ($\mathbf{W}_1$) and one input vector, $\mathbf{W}_0(t)$. A stoichiometry matrix can be thought as a list of rows defining the change in the species for any given reaction. Therefore the $\mathbf{S}$ matrix is defined as a N_{species} by $N_{\text{reactions}}$ matrix where each row corresponds to the a reaction.

The propensity matrix, $\mathbf{W}_1$, is defined as the rates at which reactions happen and is the inverse dimensions of the stoichiometry matrix, $N_{\text{reactions}}$ by N_{species}. $\mathbf{W}_0(t)$, the input vector is the input into the system from outside to each species, often defined in boundary conditions within biological contexts. The system of linear ODEs is then written in the following form:

$$\frac{d}{dt}\mathbf{x} = \mathbf{S}\mathbf{W}_1\mathbf{x} + \mathbf{W}_0(t) \tag{1.1}$$

The analytical solution to problems that can be written like this is readily available using a matrix exponential:

$$\mathbf{x}(t) = e^{(\mathbf{S}\mathbf{W}_1(\mathrm{t}))} \cdot x(0) + \int_0^{\mathrm{t}} e^{(\mathbf{S}\mathbf{W}_1(\mathrm{t}-\tau))\mathrm{W}_0(\tau)\mathrm{d}\tau} \tag{1.2}$$

Modeling with ODEs for modeling DNA transcription is showcased in Chapter 3.1; The model selection performed there was with 6 different mechanistic models consisting of linear systems of ODEs.

1.5 Stochastic modeling

Despite the usefulness of linear systems of ODEs, there are systems that are inherently noisy due to low molecular counts, or systems where a deterministic approach may mask bi-modalities. Unfortunately for computational biologists, most cell biology systems suffer (from our perspective, from their perspective noise is a useful tool) from randomness leaving a need for stochastic simulation approaches [39, 40, 41, 42]. Stochastic simulation requires using any of numerous Monte Carlo methods to provide statistically accurate single trajectories of a process. As a consequence of this, often thousands if not millions of simulations are needed to recapture a underlying probability distribution function. The core stochastic simulation algorithm used in my work is the direct method of Gillespie's algorithm [4, 43]. The formulation of stoichiometries and propensities allows us to instantly transition from a deterministic approach to the discrete stochastic. Starting at some initial state, $\mathbf{x}(0)$, the direct method draws two uniform random numbers, r_1 and r_2 from

$U[0,1]$; one for the time at which the next reaction happens, and one for which reaction happens. The time to the next reaction is defined as:

$$\tau = \frac{1}{\sum_{\mathrm{j}} a_{\mathrm{j}}(x)} \cdot log\left(\frac{1}{r_1}\right) \tag{1.1}$$

Where $a_0(x)$ is the current propensities of the reactions. Which reaction, i, happened is dictated by the first i that satisfies $\sum_1^{\mathrm{i}} a_{\mathrm{i}}(x) > r2 \cdot \sum_{\mathrm{j}} a_{\mathrm{j}}(x)$; In words, i is the index where the cumulative sum of the propensity vector up to i is higher than the uniform random number times the entire sum of the propensity vector. This exponential waiting time "time to next reaction" comes from a fundamental assumption that one is simulating a Poisson process. Once the time of the next reaction and which reaction occurs is generated randomly, the entire system is updated to reflect these changes and the process is repeated until a final defined time is reached. The jump from analytical solution to simulation is computationally expensive and there have been many improvements since 1976, which are reviewed in Gillespie 2007 [43]. Now we will focus on a particular stochastic simulation to approximate non-linearities in mRNA translation that would be lost by linear ODE models.

1.6 TASEP models

1.6.1 Model description

The TASEP, Totally Asymmetric Simple Exclusion Process, is a model for particles in motion in one dimension on a discretized lattice, where each particle has an exclusion, i.e., they cannot occupy the same lattice nodes. This exclusion is the "simplest"particle interaction possible (providing the SE in TASEP). The TASEP formulation was described in Macdonald 1968 for describing protein synthesis [44, 45]. The TASEP differs from a separate contemporary formulation called the ASEP; Particles in a ASEP are allowed to move in either direction on its 1D lattice ring, while in a TASEP, particles are only allowed to move in one direction, hence "totally"asymmetric vs asymmetric [44, 45, 46]. One of the simplest descriptions of a TASEP describes a lattice of N nodes, and particles occupy one node at a time, enter at the first node N_0 at rate α, leave the last node at rate

β, and step from node to node with rate k nodes/time. Particles cannot pass one another. Despite these deceptively simple rules, particles can still exhibit complex, non-equilibrium behavior and phase transitions between three domains (high density, low density, maximal flow). TASEPs with various mathematical extensions are used to describe a wide variety of phenomena like traffic flow, vesicle movement [47, 48], mRNA translation [49, 28, 50, 35], and line queuing [51]. Two recent reviews of TASEP processes in biological contexts are Zia et al. [49], and Chou et al. [52]. With the advent of stronger computing resources and newer experimental techniques, TASEPs have enjoyed a renewed interest in the mRNA translation modeling field [49]. TASEPs' main advantage over coarse-grained or ODE models is the replication of traffic jams, phase transitions, and high-density regimes that come with considering physical exclusion terms; Such terms that are neglected by other averaging models. However, deriving an analytical solution for the occupation probability of lattice nodes over time is a nontrivial task for complicated TASEPs, and may necessitate costly computational Monte Carlo simulation if no analytical solution exists. Fig. 1.4A shows a graphical representation of the simplest TASEP, and Fig. Fig. 1.4B shows an example stochastic trajectory kymograph (a visual plot of space vs time showing particle motion) of a lattice of 100 nodes and a constant stepping rate.

1.6.2 Mathematical description

The simplest TASEP can be described as follows The lattice of L nodes can be written as:

$$N = [N_0, N_1, \ldots, N_{\mathrm{L}-1}] \tag{1.1}$$

Particles from an infinite reservoir can enter N_0 and occupy one lattice node of space if N_0 is unoccupied at a rate of α units of time^{-1}:

$$\frac{dP(N_0|N_0, N_1)}{dt} = \alpha \cdot P(N_0 = 0) - k \cdot P(N_0 = 1, N_1 = 0) \tag{1.2}$$

Particles can step from the i^{th} node to the next node if the next node is unoccupied at rate k units of time^{-1}:

$$\frac{dP(N_{\text{i}}|N_{\text{i}-1}, N_{\text{i}}, N_{\text{i}+1})}{dt} = k \cdot P(N_{\text{i}} = 0, N_{\text{i}-1} = 1) - k \cdot P(N_{\text{i}} = 1, N_{\text{i}+1} = 0) \tag{1.3}$$

A particle at the final node can leave at an non-excluded rate of β units of time^{-1}.

$$\frac{dP(N_{\text{L}-1}|N_{\text{L}-1}, N_{\text{L}-2})}{dt} = -\beta \cdot P(N_{\text{L}-1} = 1) + k \cdot P(N_{\text{L}-2} = 1, N_{\text{L}-1} = 0) \tag{1.4}$$

1.6.3 TASEP domain diagram

TASEP models can exhibit nonlinear behavior and phase transitions between 3 main domains: High density (HD), low density (LD), and maximal current (MC), 1.4C and D. These domains can be observed by normalizing k, α, and β by k. High density occurs when the incoming rate, α, is greater than the outgoing lattice rate, β while β is less than 1/2. A faster incoming rate than outgoing leads to an accumulation and higher density of particles along the lattice, potentially to the point of a traffic jam where most lattice nodes are occupied. Low density occurs in the opposite case, where the outgoing rate is higher than the incoming rate and α is less than 1/2, resulting in a sparsely loaded lattice with a very low occupation probability. The maximum current domain exists when both the incoming and outgoing rates are greater than 1/2, balancing the kinematic shock of the interaction with the 2 boundary areas of the lattice. The current in the LD is dominated by the incoming rate, $J = \alpha(1 - \alpha)$ particles/s, and in the HD by the outgoing rate $J = \beta(1 - \beta)$ particles/s. In the MC, the current equals 1/4 particles/s at a steady state. One defining feature of this domain map is the sharp boundary between the LD and HD at α = β for both less than 1/2, showing that slight changes in the balance between the two rates radically alter the system behavior. For in-depth discussion and solutions, once again the following recent reviews: Zia et al. 2011 [49], Knizel et al. 2019 [53], and Matetski and Remenik 2023 [54].

1.6.4 mRNA translation TASEP

Many extensions to the canonical TASEP can capture dynamics of mRNA translation with greater accuracy. Some of the more common model extensions used for mRNA translation are described in Fig. 1.4E. The entering rate α is analogous to the ribosomal initiation rate, and termination is analogous to β. Stepping rates, k, can be adjusted to be codon-dependent, thus each "node" becomes the codon a ribosome is currently decoding. These codon-dependent rates can selected to match various models of ribosomal elongation such as codon frequency scaling, codon couplets, or tRNA usage. The size of particles can be adjusted to account for the ribosomal footprint of 9-10 codons. With these adjustments, we have a robust mRNA model that can capture many translation dynamics of interest with a realistic, mechanistic model based on ribosome location, ribosome footprints, and mRNA sequence differences. Some of the other extensions in Fig. 1.4E are used for specific mRNA dynamics, for example, multiple lattices with inter-lattice jumps can be used to model mRNA frameshifting, jumping from one open reading frame to another. Or Variable entries and exits can be used to model Internal Ribosomal Entry Sites (IRES) or bicistronic constructs.

The codon-dependent mRNA TASEP is the core to our labs software package, rSNAPsim (Rna Sequence to NAscent Protein SIMulator). Inside the open source Python package is a custom TASEP model designer that implemented, allowing for almost any TASEP model to be designed quickly and efficiently. These extended modeling capability of the rSNAPsim's TASEP model builder is showcased in Chapter 2.2.

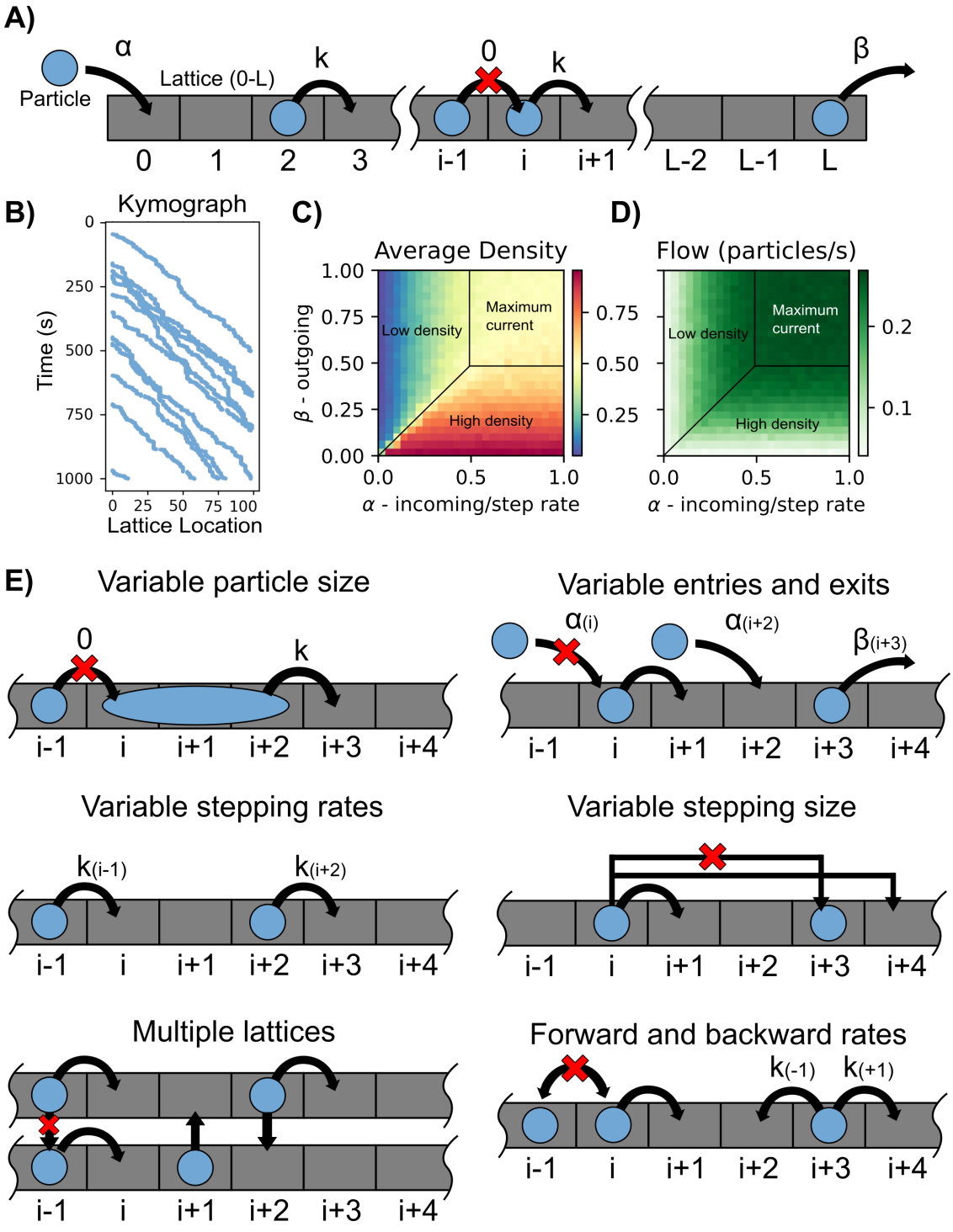


Figure 1.4: Caption on next page.

Figure 1.4: Totally asymmetric simple exclusion process (TASEP) model, its properties, and some example extensions. (A) A canonical TASEP can be described as a 1D lattice of length L, on which particles enter the first node at rate $\alpha(1/time)$, step forward one node at rate $k(1/time)$, and leave the lattice at rate $\beta(1/time)$. Particles cannot pass each other and have a volume of 1 node (exclusion). (B) Example kymograph of a TASEP simulation (C) Three domains resulting from a TASEP with stepping rate = α. Maximal current —where particles enter and leave at fast enough rates to keep a constant particle flow. High density —where incoming rates are much faster than the outgoing rate, leading to a traffic jam rate limited by the outgoing rate. Low density —where the incoming particles are slower than the rate they leave the lattice, resulting in a sparsely loaded lattice. (D) Flow in particles per second of a TASEP with stepping rate = α. (E) Example extensions to the simple TASEP. Extensions can add complexity and model different phenomena. For example, multiple lattices could be used to model multi-lane highways for traffic simulations, or particle sizes greater than one could model the ribosomal footprint of 9 codons in an mRNA translation simulation.

Chapter 2

Translation Modeling

2.1 Computational design and interpretation of single-RNA translation experiments[1]

My major contributions to the following paper were developing the open source Python package, rSNAPsim (Rna Sequence to NAscent Protein simulator) and providing figures and analyses for the paper. I was brought onto the paper in the next Chapter to convert and extend a MATLAB based TASEP model into Python for easier use, computational speed, installation, and distribution. The original TASEP model was a collaboration between Dr. Huy Vo, Dr. Luis Aguilera, and Dr. Michael May and was limited in its scope. The original version could only use human codon frequency counts to scale elongation rates and use only a single fluorescent color (FlagTag), and was in stuck in MATLAB. rSNAPsim has been extended and developed beyond the scope of the initial paper to allow a user to design a plethora of various mRNA models, model assumptions, model complexity, and arbitrary fluorescent probe design. Beyond the core models, the rSNAPsim provides an arbitrary TASEP model designer, allowing the user to add various elements such as ribosomal frameshifting, resource consumption, limited ribosome pools, arbitrary stepping rates, experimental perturbations, and mRNA states. More in-depth details are in Chapter 2.2.

2.1.1 Summary

Advances in fluorescence microscopy have introduced new assays to quantify live-cell translation dynamics at single-RNA resolution. We introduce a detailed, yet efficient sequence-based stochastic model that generates realistic synthetic data for several such assays, including Fluorescence Correlation Spectroscopy (FCS), ribosome Run-Off Assays (ROA) after Harringtonine

[1]Luis U. Aguilera, William Raymond, Zachary R. Fox, Michael May, Elliot Djokic, Tatsuya Morisaki, Timothy J. Stasevich, Brian Munsky, DOI: https://doi.org/10.1371/journal.pcbi.1007425

application, and Fluorescence Recovery After Photobleaching (FRAP). We simulate these experiments under multiple imaging conditions and for thousands of human genes, and we evaluate through simulations which experiments are most likely to provide accurate estimates of elongation kinetics. Finding that FCS analyses are optimal for both short and long length genes, we integrate our model with experimental FCS data to capture the nascent protein statistics and temporal dynamics for three human genes: KDM5B, β-actin, and H2B. Finally, we introduce a new open-source software package, RNA Sequence to NAscent Protein Simulator (rSNAPsim), to easily simulate the single-molecule translation dynamics of any gene sequence for any of these assays and for different assumptions regarding synonymous codon usage, tRNA level modifications, or ribosome pauses. rSNAPsim is implemented in Python and is available at: https://github.com/MunskyGroup/rSNAPsim.git.

2.1.2 Introduction

The central dogma of molecular biology (i.e., DNA codes are transcribed into messenger RNA, which are then translated to build proteins) has been a foundation of biological understanding since it was stated by Francis Crick in 1958. Despite their overwhelming importance to biological and biomedical science, many of the fundamental steps in the gene expression process are only now becoming observable in living cells through the application of real time single-molecule fluorescence imaging approaches. Single-molecule imaging of transcription was first achieved two decades ago through the use of the MS2 system [55], which uses bacteriophage gene sequences to encode for specific and repeated stem-loop secondary structures in the transcribed mRNAs. These stem-loops are subsequently recognized and bound by multiple fluorescently-tagged MS2 Coat Proteins (MCP), which produce bright fluorescent spots that allow for the detection and spatial tracking of single-mRNA [56]. Tracking these labeled RNA has made it possible to observe many aspects of RNA dynamics that were previously obscured using bulk RNA measurements, such as the production of RNA from different alleles [57], the movement of mRNA-protein complexes from nucleus

to cytoplasm through nuclear pores [58], and the association of RNA with different regions of the cell [59].

Even more recently, imaging single-molecule translation has also become possible through the discovery of similar approaches [23, 60, 24, 26, 25]. In this case, the mRNA is modified to encode for multiple epitopes in the open reading frame of a protein of interest (POI). As the protein is translated, these epitopes are quickly recognized and bound by fluorescent antibody fragment probes, Fig. 2.1A. By combining the MS2 approach with these epitope recognition sites, the co-localization of mRNA spots and nascent protein spots reveal single-mRNA molecules that are undergoing active translation, as shown in Fig. 2.1B. As was the case for single-RNA tracking, precise spatiotemporal imaging of translation sites within single living cells allows for multiple advances in comparison to bulk or single-cell assays [61]. For example, Morisaki and Stasevich recently reviewed three different approaches to track the translation dynamics for individual mRNA molecules over time and then use these data to infer translation rates [27]. The first design is related to Fluorescence Correlation Spectroscopy (FCS), in that the nascent protein fluorescence signal is monitored over time and used to compute the auto-covariance function ($\mathrm{G}(\tau)$, Fig. 2.1C, bottom). The time, τ_{FCS}, at which $\mathrm{G}(\tau)$ reaches zero denotes the characteristic time for a ribosome to translate the gene from the tag region to the end of the protein of interest [23]. A second approach to measure translation rate is to chemically block translation initiation (e.g., through the application of a drug such as Harringtonine, as depicted in Fig. 2.1D, top). In this Run-Off Assay (ROA) approach, the time, τ_{ROA}, at which the fluorescence signal disappears corresponds to the time for a single ribosome to translate the entire coding region, including the tag region itself [25]. A third technique, shown in Fig. 2.1E, uses Fluorescence Recovery After Photobleaching (FRAP) to optically eliminate the nascent protein fluorescence associated with a single mRNA and then record the recovery of the signal to its original level. As for the FCS approach, the time of total recovery, τ_{FRAP}, relates to the time required for a single ribosome to complete translation from the tag region to the termination codon [24, 26].

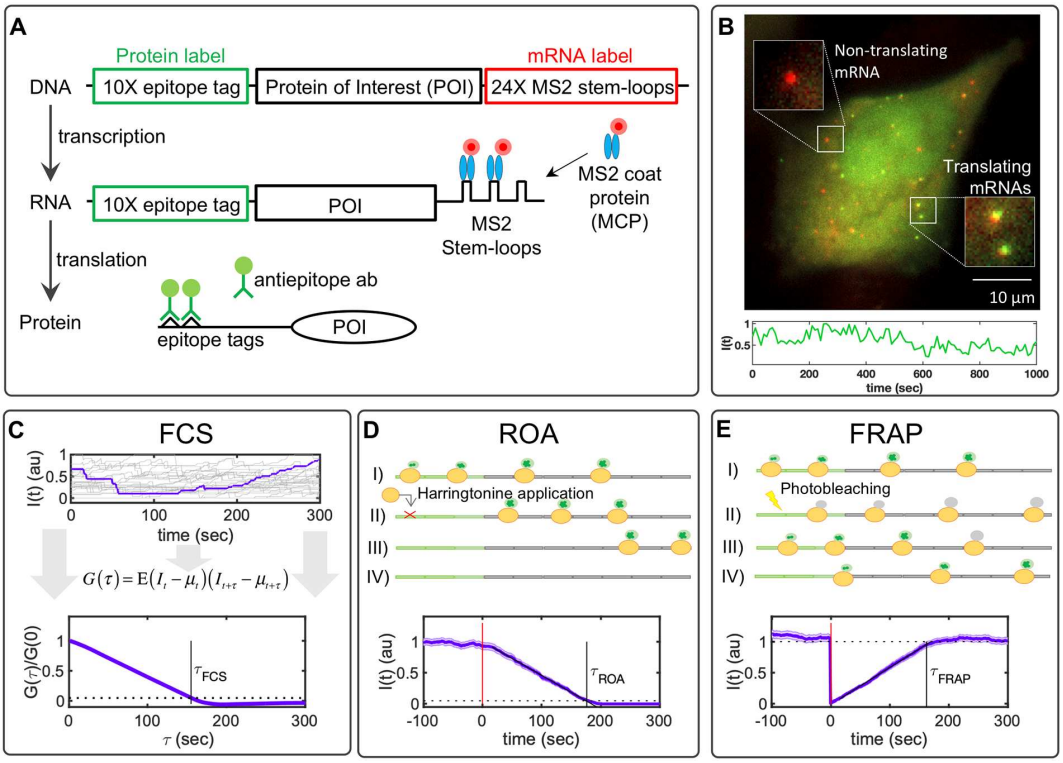


Figure 2.1: Translation studies with single-molecule resolution. A) Imaging single-molecule translation dynamics is achieved by the measurement of fluorescence spots that are produced when nascent proteins display epitopes that are recognized by antibody fragments bound to fluorescent probes. The gene construct encodes a 10X FLAG SM tag followed by a protein of interest (POI) and the 24X MS2 tag in the 3' UTR region. B) Microscopy image showing translation at single-molecule resolution; red spots represent single mRNA, and green spots represent nascent proteins. Below is a representative trace showing the intensity fluctuation dynamics of a single-transcript translating FLAG-10X-KDM5B. C) Simulated time courses representing the characteristic single-molecule fluctuation dynamics. A representative trace is selected and highlighted. At the bottom of the figure is given the normalized auto-covariance function ($G(\tau)$) calculated from simulated time courses. The time at which $G(\tau)$ hits zero represents the dwell-time (τ_{FCS}). D) Harringtonine inhibits the translation initiation step by binding to the ribosomal 60S subunit. The plot shows the average fluorescence after Harringtonine treatment. Without new initiation events, the fluctuations diminish causing the intensity to drop to zero at time τ_{ROA}. E) FRAP causes a rapid drop in the fluorescent intensity and a subsequent recovery that is proportional to the time needed by ribosomes to produce new nascent proteins with non-photobleached probes. The bottom plot shows the temporal dynamics of FRAP, where it can be observed by the abrupt decrease in intensity and a recovery time (τ_{FRAP}) correlated with the gene length. All simulations correspond to KDM5B for 100 spots and a frame rate of 1 FPS. Error bars represent the standard error of the mean.

The temporal resolution offered by live single-mRNA, nascent translation imaging makes it possible to directly visualize and quantify initiation, elongation, and termination processes in live-cells [62]. Single-molecule studies have uncovered previously unknown events and mechanisms taking place during translation, such as the presence of active and inactive transcripts in specific

locations in the cell [23, 26], different elongation rates caused by codon-optimized sequences [25], the spatiotemporal translation of specific genes in specific cellular compartments [60, 24], ribosomal frameshifting with bursty dynamics [28], and non-canonical forms of translation [63].

As these experimental techniques rapidly evolve, they induce a growing need for precise and flexible computational tools to interpret the resulting data and to design the next wave of single-RNA translation experiments. To help fill this gap, we present a versatile new set of computational design tools to estimate which specific single-mRNA translation dynamics experiments would provide the most accurate inference of model parameters. We demonstrate the generality of our analyses by simulating results for several different single-molecule experiments for a large database of human genes. We explore these different combinations of gene and experiment to ask which methodologies are better to measure specific biophysical parameters and for which types of genes. We then constrain our model by fitting it to experimental data for several genes. Finally, we describe and demonstrate the use of a new open-source and user-friendly software package: RNA Sequence to NAscent Protein Simulation (rSNAPsim), which allows the user to easily simulate the single-molecule translation dynamics of any gene. Finally, we discuss future directions and the potential limitations of the current form of this new technology.

2.1.3 Results and Discussion

Modeling single-RNA translation dynamics

To simulate translation with single-molecule resolution, we adopted a stochastic model of polymerization that is similar to those developed previously in [13, 64, 42]. We then extend this model to allow for variable ribosome sizes, codon- and tRNA-dependent translation elongation rates, and arbitrary placement of fluorescent probe binding epitopes, and we analyze these models specifically in the context of single-mRNA translation as observed using time-lapse fluorescence microscopy experiments (e.g., Fig. 2.1).

The model consists of a set of reactions where random variables $\{x_{\mathrm{i}}\}$ represent the fluctuating occupancy of ribosomes at each specific i^{th} codon along a single mRNA,

$$\emptyset \xrightarrow{w_0(x_1,\ldots,x_{\mathrm{nf}})} x_1 \xrightarrow{w_0(x_1,\ldots,x_{\mathrm{nf}+1})} \ldots x_{\mathrm{i}} \xrightarrow{w_0(x_{\mathrm{i}},\ldots,x_{\mathrm{nf}+\mathrm{i}})} \xrightarrow{w_{\mathrm{t}}} x_{\mathrm{L}}, \tag{2.1}$$

where L is the length of the gene in codons, n_{f} is the ribosome footprint, and $\mathbf{x} = [x_1, x_2, \ldots, x_{\mathrm{L}}] \in \mathbb{B}^{\mathrm{L}}$ is a binary vector of zeros and ones, known as the occupancy vector, which represents the presence (x_{i} = 1) or absence (x_{i} = 0) of ribosomes at every i^{th} codon. The initial reaction in the model describes the initiation step, where the ribosomes bind to the mRNA at the rate $w0(x_1, \ldots, x_{\mathrm{n_f}})$. Ribosomes are large bio-molecules that occupy around 20 to 30 nuclear bases (or seven to 10 codons) once bound to the mRNA [65]. This is captured in the model by specifying the ribosome footprint, n_{f} = 9, which guarantees that initiation cannot occur if another downstream ribosome is already present within the first n_{f} codons, Fig. 2.2A. This binding restriction can be written simply as:

$$w_0 = k_{\mathrm{i}} \prod_{\mathrm{j}=1}^{\mathrm{n_f}} (1 - x_{\mathrm{j}}), \tag{2.2}$$

where k_{i} is the initiation constant, and the product is equal to one if and only if there are no ribosomes within the first n_{f} codons.

Similarly, we represent the elongation reactions, where the ribosome moves along the mRNA from codon to codon in direction 5' to 3' according to:

$$w_{\mathrm{i}} = k_{\mathrm{e}}(i) \cdot x_{\mathrm{i}} \prod_{\mathrm{j}=1}^{\mathrm{nf}} (1 - x_{\mathrm{i+j}}), \text{ for } i = 1, ..., L-1; \tag{2.3}$$

where $\bar{k_{\mathrm{e}}}(i)$ is the elongation rate at the i^{th} codon, and the product again enforces ribosome exclusion. To implement the effect of codon-usage bias and tRNA availability during protein synthesis, we adopt a similar argument to that presented by Georgoni et al., [13]: rare codons are correlated with low tRNA abundance, which cause a longer waiting time for the ribosome to synthesize the given amino acid at that codon. As tRNA concentrations have been related to codon usage [66], we assume each codon's elongation rate is proportional to its usage in the human genome according to:

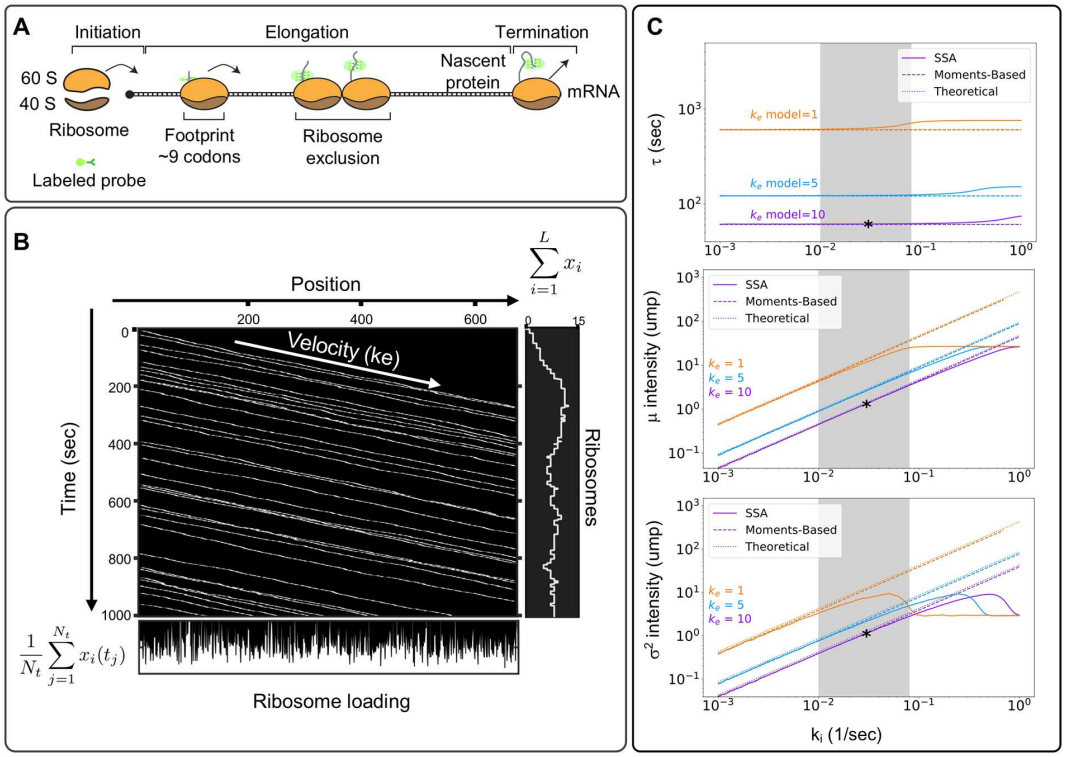


Figure 2.2: Modeling single-molecule translation. A) Translation is divided into three main processes: initiation, elongation, and termination. The ribosome footprint represents the physical space occluded by the ribosome, enforcing that no two ribosomes can occupy the same space and time. B) Kymographs represent ribosome movement as a function of time (y-axis) and position (x-axis). Each line represents a single ribosome trajectory. The average slope is proportional to the effective ribosome elongation rate. The plot to the right shows the relationship between ribosome movement and fluorescence intensity, and the plot below shows the ribosome loading at each codon position, calculated as the time-average of ribosome occupancy at the corresponding codon. C) Comparison of the average elongation time (top) and the mean (middle) or variance (bottom) of fluorescence intensity as calculated using the simplified model (Eqs 2.18 to 2.21), a linear moments-based model (Eqs 2.9 to 2.17), and a full stochastic model (Eqs 2.1 to 2.5). Gray area represents previously reported parameter values for ribosome initiation. Panels B and C correspond to simulations for the β-actin. Asterisks represent the specific parameter combination used for Table 1.

$$k_{\mathrm{e}}(i) = \bar{k}_{\mathrm{e}} \cdot (u(i)/\bar{u}), \tag{2.4}$$

where $u(i)$ denotes the codon usage frequency in the human genome (given in Table 2.1 from [67]), represents the average codon usage frequency in the human genome, and the global parameter is an average elongation constant, which can be determined through experiments.

Although simple in its specification, the above model allows for many adjustments to explore different experimental circumstances. As a few examples, (i) one can represent translation inhi-

Table 2.1: Codon usage table calculated from the Homo sapiens genome. Table is computed using 93,487 CDS (Coding DNA Sequence), that represent a total of 40,662,582 codons, Nakamura, et al., 2000.

TTT	17.6	TCT	15.2	TAT	12.2	TGT	10.6
TTC	20.3	TCC	17.7	TAC	15.3	TGC	12.6
TTA	7.7	TCA	12.2	TAA	1.0	TGA	1.6
TTG	12.9	TCG	4.4	TAG	0.8	TGG	13.2
CTT	13.2	CCT	17.5	CAT	10.9	CGT	4.5
CTC	19.6	CCC	19.8	CAC	15.1	CGC	10.4
CTA	7.2	CCA	16.9	CAA	12.3	CGA	6.2
CTG	39.6	CCG	6.9	CAG	34.2	CGG	11.4
ATT	16.0	ACT	13.1	AAT	17.0	AGT	12.1
ATC	20.8	ACC	18.9	AAC	19.1	AGC	19.5
ATA	7.5	ACA	15.1	AAA	24.4	AGA	12.2
ATG	22.0	ACG	6.1	AAG	31.9	AGG	12.0
GTT	11.0	GCT	18.4	GAT	21.8	GGT	10.8
GTC	14.5	GCC	27.7	GAC	25.1	GGC	22.2
GTA	7.1	GCA	15.8	GAA	29.0	GGA	16.5
GTG	28.1	GCG	7.4	GAG	39.6	GGG	16.5

bition analyses such as those performed in [60] by making the initiation rate, k_{i}, a function of time or external input; (ii) one can analyze effects of synonymous codon substitution by replacing codons with their more or less common relatives; (iii) one can represent codon depletion, as studied in [13] by reducing the corresponding rates $k_{\mathrm{e}}(i)$ for all i corresponding to the depleted tRNA; (iv) one could explore the effects of pausing or traffic jams at specific codons by reducing $k_{\mathrm{e}}(i)$ at specific codons, or (v) one can represent bursting kinetics by replacing the constant k_{i} with a discrete-stochastic activation/deactivation process. We will explore several of these circumstances below.

Kymograph representation of single-mRNA translation dynamics.

With our simple specification of the translation initiation, elongation and termination reactions, we can now simulate random trajectories, $\mathbf{x}$, which we collect to form binary occupancy trajectory matrices

$\mathbf{X} = [\mathbf{x}(t_1)\mathbf{T}, ..., \mathbf{x}(t_{\mathrm{N}_1})\mathbf{T}]\mathbf{T} \in \mathbb{B}^{\mathrm{N_t} \times \mathrm{L}}$, where each row refers to the i^{th} position on the gene, and each column represents a specific time t_{j}. To visualize ribosome movement trajectories, each

random $\mathbf{X}$ can be plotted in two dimensions (position v.s. time) to form kymographs similar to those extensively used to represent organelle movement [68]. For example, Fig. 2.2B shows a visualization of $\mathbf{X}$ for a case study on the β-actin gene. Each line from left to right on the kymograph corresponds to the movement of a single ribosome from initiation to termination. We note that averaging along the columns of X (i.e., in the vertical direction of the kymograph) yields the time-averaged loading of the ribosomes at each codon position, and summing across the rows of $\mathbf{X}$ (i.e., in the horizontal direction of the kymograph) yields the number of ribosomes for that mRNA at each instant in time.

Relating protein elongation dynamics to fluorescence signal intensities.

To relate our model describing ribosome occupancy to experimental measurements of translation spot fluorescence, we introduce a fluorescence intensity vector that converts the instantaneous occupancy vector, $\mathbf{x}(t)$, to the total number of translated epitopes available to bind to fluorescent markers. This intensity vector can be written as:

$$I(t) = \sum_{\mathrm{i}=1}^{\mathrm{L}} c_{\mathrm{i}} \cdot x_{\mathrm{i}}(t) = \mathbf{cxT}, \tag{2.5}$$

where $\mathbf{c} = [c_1, c_2, ..., c_{\mathrm{L}}]$ and each c_{i} is the cumulative number of fluorescent probes bound to epitopes encoded at positions (1, ..., i) along the mRNA. For example, c would be defined as $\mathbf{c} = [0, 0, 1, 1, 2, 2, 3, 3, ..., 3]$ for an RNA sequence with epitopes encoded at positions [57, 59, 60]. We note that the random occupancy matrices, $\mathbf{X}$, are easily converted to intensity time traces using the simple algebraic operation $\mathrm{I} = [I(t_1), ..., I(t_{\mathrm{N_t}})] = \mathbf{cXT}$.

Simplifications for combinatorial analyses of genes, parameters, and experiment designs

The model as defined above is sufficient to simulate fluorescence dynamics for any specified gene and for a vast range of potential time-lapse microscopy experiments. However, these simulations become computationally intensive when studying combinations of thousands of genes, using thousands of different parameters sets, and for hundreds of different experiment designs. To ame-

liorate this concern, we next introduce model simplifications that progressively remove elements from the original model, such as ribosome exclusion and single-codon resolution, while retaining effects of codon-dependent translation rates and the geometric placement of fluorescent tags. We then test under what conditions (i.e., parameters and gene lengths) these simplifications are valid, and we compare these conditions to experimentally reported values.

Approximations for the means, variances, and auto-covariances of nascent translation kinetics.

When ribosome loading is sparse (e.g., for slow initiation or fast elongation such that ($k_{\mathrm{i}}/k_{\mathrm{e}} \ll 1/n_{\mathrm{f}}$)), ribosome collisions will become negligible, and the nonlinearities in Eqs 2.2 and 2.3 have less effect on the overall ribosome dynamics. Under such circumstances, it is possible to derive a simplified linear system model for the elongation dynamics. In the linear model, the propensity of the codon-dependent elongation step (Eq 2.2) is simplified to $w_{\mathrm{i}}(x_{\mathrm{i}}) = k_{\mathrm{i}} x_{\mathrm{i}}$ such that the ability of a ribosome to add another amino acid only depends on the current position of the ribosome, and not on the footprint of other ribosomes.

We define the reaction stoichiometry matrix to describe the change in the ribosome loading vector, **x**, for every reaction as:

$$\mathbf{S}_{\mathrm{i,j}} \begin{cases} \quad 1 \text{ for all } i = j, \\ -1 \text{ for all } i = j - 1, \end{cases} \tag{2.6}$$

where i corresponds to each codon in the protein of interest. The first column of **S** corresponds to the initiation reaction, the next L - 1 columns refer to elongation steps when an individual ribosome transitions from the i^{th} to the i + 1^{th} codon, and the final column corresponds to the final elongation step and termination. Maintaining the same order of reactions, and neglecting ribosome exclusion, the propensities of all reactions can be written in the affine linear form as:

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{W}_1 \mathbf{x}, \tag{2.7}$$

where $\mathbf{w}_0$ is a column vector of zeros with the first entry k_i, and $\mathbf{W}_1$ is a matrix defined as:

$$[\mathbf{W}_1]_{\mathrm{i,j}} \begin{cases} k_\mathrm{e}(i) \text{ for all } i = j+1 \\ \quad 0 \text{ otherwise.} \end{cases} \tag{2.8}$$

Using the definition of the fluorescence intensity from Eq 2.5, the first two uncentered moments of the intensity $I(t)$ can be written in terms of the ribosome position vector $\mathbf{x}(t)$ as:

$$\mathbb{E}\{I(t)\} = \mathbb{E}\{\mathbf{c}\mathbf{x}(t)\} = \mathbf{c}\mathbb{E}\{\mathbf{x}(t)\}, \tag{2.9}$$

$$\mathbf{\Sigma}_\mathrm{I}(0) = \mathbb{E}\{(I(t) - \mathbb{E}\{I(t)\})^2\} = \mathbf{c}\mathbf{\Sigma}_\mathrm{x}(0)\mathbf{c}\mathbf{T} \tag{2.10}$$

where $\mathbb{E}\{\mathbf{x}(t)\}$ and $\mathbf{\Sigma}x(0)$ are the mean and zero-lag-time variance in the ribosome occupancy vector, respectively. For the approximate linear propensity functions in Eq 2.7, the moments of the ribosome position vector are governed by the equations [69]:

$$\frac{d\mathbb{E}\{\mathbf{x}\}}{dt} = \mathbf{S}\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{S}\mathbf{w}_0 \tag{2.11}$$

$$\frac{d\mathbf{\Sigma}\mathbf{x}}{dt} = \mathbf{S}\mathbf{W}_1\mathbf{\Sigma}\mathbf{x} + \mathbf{\Sigma}\mathbf{x}{\mathbf{W}_1}^\mathrm{T}\mathbf{S}^\mathrm{T} + \mathbf{S}\mathrm{diag}(\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{w}_0)\mathbf{S}^\mathrm{T} \tag{2.12}$$

By setting the left hand side of Eq 2.11 to zero, the steady-state mean ribosome loading vector can be found by solving the algebraic expression:

$$\mathbf{S}\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{S}\mathbf{w}_0 = 0 \tag{2.13}$$

Similarly, the steady-state covariance matrix, Σ_x, in the ribosome loading vector is given by the solution to the Lyapunov equation (from right hand side of Eq 2.12):

$$\mathbf{S}\mathbf{W}_1\mathbf{\Sigma}\mathbf{x} + \mathbf{\Sigma}\mathbf{x}{\mathbf{W}_1}^\mathrm{T}\mathbf{S}^\mathrm{T} + \mathbf{S}\mathrm{diag}(\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{w}_0)\mathbf{S}^\mathrm{T} = 0 \tag{2.14}$$

The auto-covariance dynamics of the nascent protein fluorescence intensity is defined:

$$\begin{aligned} G(\tau) &= \mathbb{E}\{(I(t) - \mathbb{E}\{I(t)\})(I(t+\tau) - \mathbb{E}\{I(t+\tau)\})\} \\ &= \mathbb{E}\{\mathbf{c}(\mathbf{x}(t) - \mathbb{E}\{\mathbf{c}\mathbf{x}(t)\})\mathbf{x}(t+\tau) - \mathbb{E}\{\mathbf{c}\mathbf{x}(t+\tau)\})^{\mathrm{T}}\mathbf{c}^{\mathrm{T}}\} \\ &= \mathbf{c}\mathbb{E}\{(\mathbf{x}(t) - \mathbb{E}\{\mathbf{x}(t)\})(\mathbf{x}(t+\tau) - \mathbb{E}\{\mathbf{x}(t+\tau)\})^{\mathrm{T}}\}\mathbf{c}\mathbf{T} \qquad (2.15) \\ &= \mathbf{c}\mathbf{\Sigma}x(\tau)\mathbf{c}T \qquad (2.16) \end{aligned}$$

where $\Sigma\mathbf{x}(\tau)$ is the cross-covariance of the ribosome occupancies at a lag time of length τ. Noting that the probe design, $\mathbf{c}$, is constant with respect to τ, it is only necessary to find the cross-covariances of the ribosome occupancy. Following the regression theorem [70], these covariances are given by the solution to the set of ODEs,

$$\frac{d\mathbf{\Sigma}\mathbf{x}(\tau)}{d\tau} = \phi\mathbf{\Sigma}\mathbf{x}(\tau), \qquad (2.17)$$

where the initial condition is provided by steady-state covariance (i.e., the solution for $\mathbf{\Sigma}\mathbf{x}(0)$ in Eq 2.14) and the autonomous matrix of the process is given by $\phi = \mathbf{S}\mathbf{W}_1$. Integrating Eq 2.17, the auto-covariance of the intensity $G(\tau)$ can be found using Eq 2.16. We reiterate the fact that this simplification relies only on the assumption of sparse loading of ribosomes on the mRNA, and the moments analyses in Eqs 2.13, 2.14 and 2.17 retain the codon-dependent rate through the definition of the matrix $\mathbf{W1}$ and the specific positions of probes through the definition of the vector $\mathbf{c}$.

Simplified algebraic expressions for nascent translation kinetics.

In the limit of low initiation events and long genes, the probe region can be further approximated by a single point, and the above model can be simplified even further to allow direct estimation of steady-state translation features. First, since the average time for a ribosome to move one

codon is $\mathbb{E}\{\delta t_{\mathrm{i}}\} = 1/k_{\mathrm{e}}(i)$, the total average time it takes a ribosome to complete translation from the start codon to the end of the mRNA is:

$$\mathbb{E}\{\tau\} = \sum_{\mathrm{i}=1}^{\mathrm{L}} \frac{1}{k_{\mathrm{e}}(i)}, \tag{2.18}$$

where L is the gene length. Using the codon-dependent translation rates from Eq 2.4, we can modify Eq 2.18 to

$$\mathbb{E}\{\tau\} = \frac{1}{\bar{k}_{\mathrm{e}}} \sum_{\mathrm{i}=1}^{\mathrm{L}} \frac{\bar{u}}{u(i)}. \tag{2.19}$$

If one could experimentally measure τ_{Exp} using one of the techniques described above, then $\bar{k}_{\mathrm{e}}$ could be estimated as:

$$\bar{k}_{\mathrm{e}} \approx \frac{1}{\tau_{\mathrm{Exp}}} \sum_{\mathrm{i}=\mathrm{n_p}}^{\mathrm{L}} \frac{\bar{u}}{u(i)}, \tag{2.20}$$

where n_{p} is the effective codon position of the fluorescent tag. In practice, the specification of n_{p} will vary depending upon the type of experiment (e.g., FCS, FRAP or ROA) used to estimate τ_{Exp}, as will be discussed in more detail below.

Given the apparent association time of a ribosome on the mRNA (τ) and the initiation rate (k_{i}), the distribution for the number of visible ribosomes on a transcript at steady state can also be estimated using this simplified model. Under the assumption that each initiation event is an independent and exponentially distributed random event, the number of ribosomes downstream from the $n_{\mathrm{b}}^{\mathrm{th}}$ codon, and therefore the fluorescence in units of mature proteins, would be approximated by a Poisson distribution with mean (and variance) equal to

$$\mu \approx \sigma^2 \approx k_{\mathrm{i}} \cdot \tau \tag{2.21}$$

For a more realistic treatment of the fluorescence intensity, one could assume that the multiple probes are spread uniformly over a finite region, such that the fluorescence will increase linearly

as ribosomes pass through the probe region. To approximate this gradual increase in fluorescence, Eq 2.21 can be corrected by a multiplicative factor (see Methods) as:

$$\mu_{\mathrm{I}} \approx k_{\mathrm{i}} \cdot \tau(1 - \frac{Lt}{2L}) \tag{2.22}$$

$$\sigma_{\mathrm{I}}^2 \approx k_{\mathrm{i}} \cdot \tau(1 - 2\frac{Lt}{3L}) \tag{2.23}$$

where Lt is the length of the tag region (e.g., Lt = 318 aa for the 10X FLAG 'Spaghetti Monster' SM-tag used in [23]).

Agreement of full and simplified models for codon-dependent translation kinetics.

To demonstrate the close agreement between the full stochastic model, the reduced linear moments model, and the simplified theoretical analysis, Table 1 compares the model generated values for each of the quantities τ, μ_{I}, and σ_{I}^2 for three different human genes H2B (L = 128aa), β-actin (L = 375aa), and KDM5B (L = 1549aa), using reported parameters of ki = 0.03 s^{-1} and s^{-1} [23]. For further comparison, Fig. 2.2C compares estimates of τ (top), μ_{I} (middle), and σ_{I}^2 (bottom) for the β-actin gene for each of the three analyses, and as a function of different initiation and elongation rates. This comparison demonstrates that, at least for fast elongation rates, the full stochastic analysis and the moments-based computation are in excellent agreement to estimate the effective time as well as the mean and variance in the level of nascent proteins per RNA. However, when the initiation rate approaches $\bar{k}_{\mathrm{e}}/nf$, ribosome collisions become more prevalent, which substantially lengthens the effective elongation time (Fig. 2.2C top), and leads to a saturation of ribosomes (Fig. 2.2C middle and bottom), and these nonlinear behaviors are not captured by the moment-based model. For longer genes, the simplified theoretical estimates from Eqs 2.18- 2.21 are also in good agreement with the complete model. For shorter genes, it becomes less realistic to approximate the tag region with a single point or a uniform distribution, and the error of this approximation leads to poorer estimates of the elongation time and the Poisson approximation over-estimates the true variance (see H2B in Table 1). However, even for short genes, the lin-

ear moments-based model, which includes the exact positions of all probes and the codon usage, provides a more accurate estimate of the true system behaviors.

Having demonstrated close agreement of the simplified theoretical models with the full stochastic simulations, we can now use the much more computationally efficient theoretical analyses to explore how well different experiment designs should be expected to estimate translation parameters from single-RNA translation dynamics.

Design of experimental assays for improved quantification of translation kinetics

Using the models above, and if we could experimentally estimate the average time that ribosomes take to translate a single complete protein from a given gene, $\tau^{(\mathrm{g})}$, we could estimate $\bar{k}_{\mathrm{e}}^{(\mathrm{g})}$ using Eq 2.20. With this in mind, we next consider three approaches that have been used to estimate $\tau^{(\mathrm{g})}$ in recent experimental investigations (Fig. 2.1C-2.1E): Fluorescence Correlation Spectroscopy (FCS), Run-Off Assays (ROA), and Fluorescence Recovery After Photobleaching (FRAP). Using our full stochastic models to generate synthetic data and the simplified theoretical model to interpret these data, we ask how accurately would each of these three assays work to identify $\bar{k}_{\mathrm{e}}^{(\mathrm{g})}$ for a comprehensive list of 2,647 human genes from the PANTHER database [71] and under different imaging conditions corresponding to different frame rates or numbers of mRNA spots.

In the FCS approach, we compute the auto-covariance function, $G(\tau)$ (defined in Eq 2.15), of the simulated fluorescence intensities, and from $G(\tau)$ we estimate the time lag, τ_{FCS}, at which correlations disappear (see Fig. 2.1C and Methods). In the ROA approach, we simulate the addition of a chemical compound, such as Harringtonine, which binds the 60S ribosome subunit and prevents ribosome assembly [72], and we record the average time, τ_{ROA}, at which protein fluorescence disappears from the RNA (see Fig. 2.1D and Methods). To approximate variability in the specific time at which the drug reaches the mRNA and blocks ribosome initiation, we assume that the time of initiation blockage occurs at a normally distributed time of 60 ± 10 seconds [73]. In the FRAP analysis, we simulate an instantaneous fluorescence bleaching of all nascent proteins and then record the average time, τ_{FRAP}, at which fluorescence recovers to the average steady-state level, Fig. 2.1E [74]. To reduce the effects of stochastic sample variation in these calculations, we

applied a linear fit to ROA and FRAP experiments and determined τ_{ROA} and τ_{FRAP} when these intensities intersect defined thresholds of zero intensity for ROA or the mean recovered intensity for FRAP. For FCS, we estimate τ_{05} as the time the auto-covariance function drops below 5% of the zero-lag covariance and calculate $\tau_{\mathrm{FCS}} = \tau_{05}/0.95$.

The specific location of probes along the mRNA has different effects on the fluorescence kinetics for the three experimental analyses. The characteristic decorrelation time in FCS and recovery time in FRAP are both set by the time it takes a single ribosome to translate from the tag region to the end of the mRNA. To reflect this, we define the approximate probe location, np_{FCS} or np_{FRAP} in Eq 2.20, as the beginning of the tag region. In this case, the beginning of the tag region is at the beginning of the gene, but in general, we note that moving the fluorescent tag regions downstream toward the 3' end would shorten the effective times measured using FCS or FRAP. In contrast, for the ROA, the characteristic time is defined by how long it takes from when translation initiation is blocked until all ribosomes complete translation. Because this time depends solely on the gene length, and not on the probe placement, we assume $np_{\mathrm{ROA}} = 1$, independent of probe placement. In addition to these effects on average experiment timescale estimates, we note that placing probes as near as possible to the 5' end of the mRNA or using longer proteins increases the fluorescence signal-to-noise ratio for all three approaches and can reduce estimation uncertainties.

To generate simulated data, we assumed that all 2,647 genes in the library have a global average translation rate of $\bar{k}_{\mathrm{e}} = 10$ sec^{-1} and an initiation rate of $k_{\mathrm{i}} = 0.03$ sec^{-1}. For each experiment type and each gene, we simulated time lapse microscopy data for 100 independent RNA and for 300 frames at 1/3 frames per second (FPS). We then estimated $\tau^{(\mathrm{g})}$ from these simulations using each of the three experimental methodologies, and we estimated the corresponding average elongation rate using the specific gene sequence and Eq 2.4. Under these conditions, Fig. 2.3A-2.3C (top) show the resulting distributions of estimated $\widetilde{k}_{\mathrm{e}}$ for long genes ($>$ 1000 codons, n = 658, purple), medium length genes (500 - 1000 codons, n = 1719, blue), and short genes ($<$ 500 codons, n = 270, orange) using each of the three experimental approaches. When all genes were analyzed at the same imaging conditions (100 spots, 300 frames, 1/3 FPS), the FCS approach was the most

accurate with root mean squared (RMSE) of 0.63, 1.35, and 1.60 for short, medium and long genes, respectively. For comparison, ROA had RMSE of 2.22, 2.52, and 1.78 and FRAP had RMSE of 5.22, 4.58, and 2.68 for the same combinations of genes and imaging conditions.

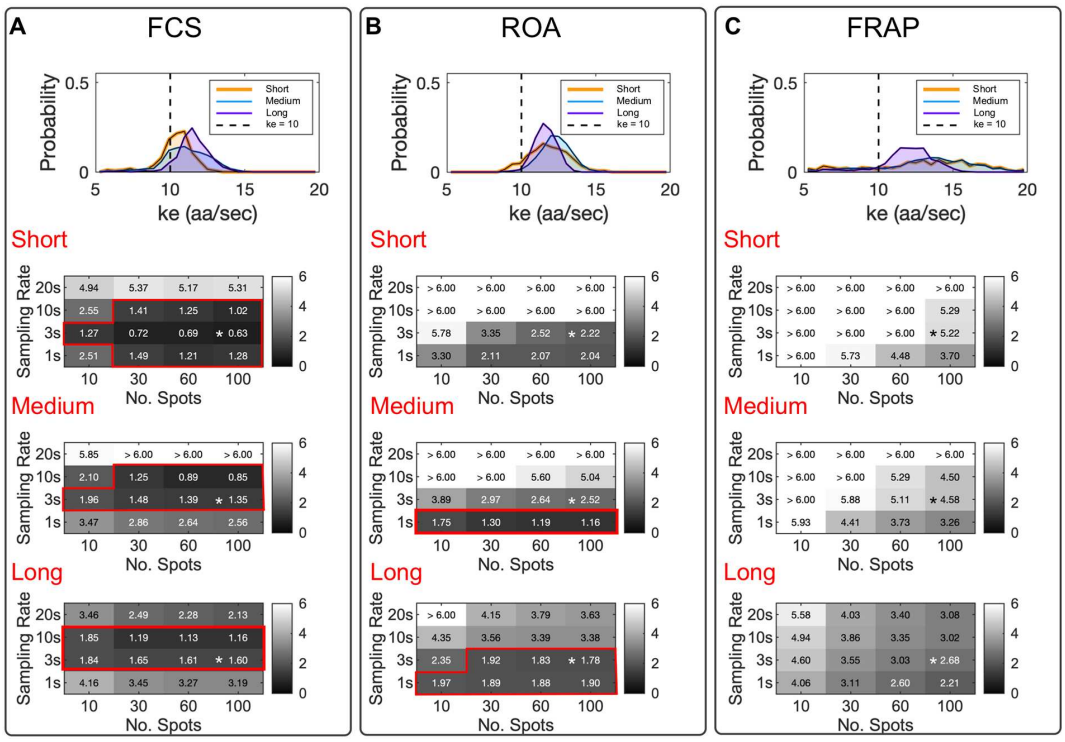


Figure 2.3: Comparing experimental methodologies to estimate ribosome elongation rates. Elongation rate estimate experiments were simulated for 2,647 human genes, using (A) Fluorescence Correlation Spectroscopy (FCS), (B) Run-Off Assays (ROA), and (C) Fluorescence Recovery After Photobleaching (FRAP). Top panels show the distributions of estimated $\tilde{k}_{\mathrm{e}}$ for long genes ($>$ 1000 codons, n = 658, purple), medium length genes (500 - 1000 codons, n = 1719, blue), and short genes ($<$ 500 codons, n = 270, orange) using 100 mRNA spots for 300 frames at 1/3 FPS. The true elongation rate is denoted by a vertical dashed line. Bottom panels show the RMSE in elongation rate estimation as a function of the number of mRNA spots and the sampling rate. Red boxes highlight all experimental designs that yield a RMSE $<$ 2.0. Asterisks represent the frame rate and number of repetitions used in panel (A). The 'true' elongation rate was set at $\bar{k}_{\mathrm{e}}$, and the initiation rate was fixed at k_{i} = 0.03 sec^{-1} for all simulations.

We next extended our analysis to consider different numbers of spots and different frame rates at which to collect the data, but under the assumption that the total number of frames would remain fixed at 300. Fig. 2.3A shows the corresponding resulting RMSE for different combinations of these experiment designs. As expected, we found the sampling rate and number of mRNA spots to directly affect the estimated $k_{\mathrm{e}}^{(\mathrm{FCS})}$. FCS was the only technique capable to estimate the true

elongation rate within a $RMSE_{\mathrm{FCS}} \leq 2.0$ sec^{-1} for short, medium and long genes. For short genes, this could be accomplished with as few as 10 spots with a frame rate of 1/3 FPS. Medium length and long genes could also be accurately quantified with 10 spots at frame rates of 1/3 FPS or 1/10 FPS.

The ROA was also capable to estimate the elongation rate to an accuracy of $RMSE_{\mathrm{ROA}}$ <2.0 sec^{-1} for medium and long genes, and for fast frame rates, the ROA approach could be more accurate than FCS. However, when applying the ROA method to short genes, we obtained $RMSE_{\mathrm{ROA}}$ >2.0 sec^{-1} under all combinations of sampling rates and repetition numbers at 100 or fewer spots (Fig. 2.3B). This effect can be explained in that the number of ribosomes actively translating each mRNA is small and highly susceptible to stochastic effects in the case of small genes. We also note that the error using ROA depends strongly on the precision of the estimate for the specific time at which translation is blocked after application of Harringtonine; if the average value of this time is unknown, or if variations exceed our assumed standard deviation of 10 seconds, then accuracy using ROA is severely diminished, especially for short genes.

We found that FRAP substantially overestimates the elongation rates for short size genes, which can be observed in Fig. 2.3C, where it is shown that recovering a $RMSE_{\mathrm{FRAP}}$ <2.0 sec^{-1} was not possible for any of the considered combinations of the number of RNA spots and sampling rates. We argue that the estimate of elongation rates using FRAP is limited by the intrinsic formulation of the fluorescent probe design. FRAP requires an intensity generating mechanism to reestablish the fluorescence to a pre-perturbation steady state. For single-molecule translation studies, this mechanism relies on ribosomal initiation events that are rare and highly susceptible to variability [23, 60, 24, 26]. This variability is reflected in the estimated τ_{FRAP} and in the final estimated elongation rate. Even for the more favorable medium and long length genes, our results indicate that for FRAP, a large number of mRNA spots (>100 mRNA spots) would be needed to achieve accurate estimates (Fig. 2.3C).

Calibration of the stochastic translation model using quantitative data from single-RNA translation experiments

Having determined that the FCS approach provides the most consistent estimate of elongation rate for genes of different lengths, we next turn to published experimental FCS data that quantified the fluctuation dynamics for three human gene constructs of different lengths: KDM5B (1549 aa), β-actin (375 aa), and H2B (128 aa) [23]. Each construct encodes for an N-terminal 10X FLAG 'Spaghetti Monster' SM-tag (318 aa) followed by the specific protein of interest (POI), and the stop codon for each POI was followed by 24 repetitions of the MS2 tag in the 3' UTR region. For each construct, the MS2 signal was used to track the mRNA motion in three dimensions, and the co-localized fluorescence intensity of the FLAG SM-tag was quantified as a function of time. These movies were collected using frame rates of 1 sec for H2B (n = 10), 3 sec for β-actin (n = 17), and 10 sec for KDM5B (n = 35), and each trajectory was tracked for up to 300 frames per mRNA. Fig. 2.4A-2.4C (left) show example time traces (in arbitrary units of fluorescence) for the nascent protein level per individual mRNA for each of the three genes. To achieve long trajectories, it is necessary to use low laser power, which introduces higher variability in signal intensities from one spot to another. Therefore, to account for variability in imaging settings between tracking experiments, all trajectories were normalized to have a variance of one prior to auto-covariance analysis.

To quantify the steady-state variability of nascent proteins per mRNA in units of mature protein (ump), we used a second, independent calibration construct that contains only a single epitope for FLAG ([23], see Methods) and which we measured using higher laser intensities. After calibration, the number of mature proteins per mRNA was rounded to the nearest integer d_{j} for a larger number of spots (1844 to 302 spots per frame for 50 imaging frames) for a total of 6435, 3973, and 751 spots for KDM5B, β-actin and H2B, respectively. The resulting data histograms were down-sampled to create an effective population of 100 translating mRNA spots for each gene, and histograms of these measurements are presented by the black lines in Fig. 2.4A-2.4C, middle.

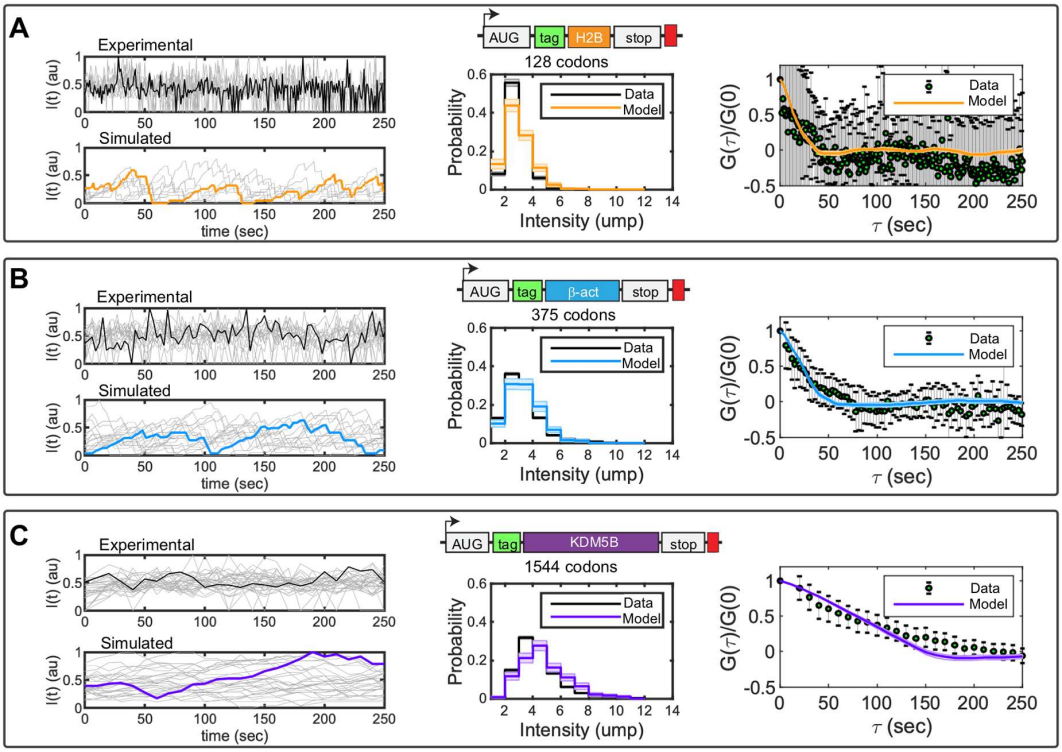


Figure 2.4: Fitting single-molecule data with the full stochastic model. Experimental data show the fluctuation dynamics of gene constructs encoding an N-terminal 10X FLAG 'Spaghetti Monster' SM-tag (green) followed by a protein of interest and finally a 24X MS2 tag (red) in the 3' UTR region. Three proteins were studied: A) H2B (orange), B) β-actin (blue) and C) KDM5B (violet). Middle figures show the simulated (colors) and measured (black) probability distributions for an mRNA to have a fluorescence intensity corresponding to i units of mature proteins (ump). Right images show the normalized auto-covariance function (G) calculated from experimentally measured (black error bars) and computationally simulated (colors) autocorrelation functions. Error bars in the experimental data and shadow bars in the simulated auto-covariance plots represent the standard errors of the mean. Elongation and initiation rates were obtained by parameter optimization, using the Hooke and Jeeves Algorithm ([75]). Optimized parameters and their uncertainties (see Methods) are provided in Eq. 2.29.

We explored if the full stochastic model could be fit to capture simultaneously the experimentally measured steady-state histogram of nascent proteins as well as the temporal dynamics of nascent protein fluctuations on single mRNA. For model comparison to the steady-state histograms, we ran 300 independent simulations per gene and parameter combination ($\wedge$) and estimated the probability to observe intensities corresponding to d = 1, 2, ... mature proteins per mRNA. We denoted resulting probability mass vector as $P(d;\wedge)$. Assuming that translation on each mRNA is independent of the rest, we could then compute the likelihood of the steady-state intensity data for each gene given the model as:

$$L_{\mathrm{Dist}}(\mathrm{Data}|\mathrm{Model}) = \prod_{\mathrm{j}=1}^{100} P(d_{\mathrm{j}}; \wedge), \tag{2.24}$$

and the log-likelihood could be computed:

$$\log L_{\mathrm{Dist}}(\mathrm{Data}|\mathrm{Model}) = \prod_{\mathrm{j}=1}^{100} \log P(d_{\mathrm{j}}; \wedge), \tag{2.25}$$

As non-translating spots could not be separated from spots below a basal FLAG intensity in the experimental data measurements, comparison between simulations and measured distributions ignore all spots with an intensity value less than 1/2 ump.

To compare temporal dynamics of the experiments to those of the model, we assumed that errors in the measurement of the average auto-covariances would be approximately normally distributed with variances equal to the measured standard error of the mean [76]. Under this assumption, the probability to measure an auto-covariance of $G_{\mathrm{D}}(\tau_{\mathrm{i}})$ at lag time τ_{i} according to a model that predicts $G_{\mathrm{M}}(\tau_{\mathrm{i}}; \wedge)$ for parameter set $\wedge$ is:

$$L_{\mathrm{AC}}(G_{\mathrm{D}}|G_{\mathrm{M}}(\wedge)) = \prod_{\mathrm{i}=1}^{\mathrm{N_t}} \frac{1}{\sqrt{2\pi\sigma(\tau_{\mathrm{i}})^2}} \exp\left(-\frac{G_{\mathrm{D}}(\tau_{\mathrm{i}}) - G_{\mathrm{M}}(\tau_{\mathrm{i}}; \wedge))^2}{2\sigma(\tau_{\mathrm{i}})^2}\right), \tag{2.26}$$

where $\sigma(\tau_{\mathrm{i}})$is approximated by the measured SEM auto-covariance at each τ_{i}. The logarithm of this likelihood function can then be written as:

$$\log L_{\mathrm{AC}}(G_{\mathrm{D}}|G_{\mathrm{M}}(\wedge)) = C - \sum_{\mathrm{i}=1}^{\mathrm{N_t}} \frac{G_{\mathrm{D}}(\tau_{\mathrm{i}}) - G_{\mathrm{M}}(\tau_{\mathrm{i}}; \wedge))^2}{2\sigma(\tau_{\mathrm{i}})^2}, \tag{2.27}$$

where C is a constant that does not depend upon the parameter set $\wedge$, and the second term is the definition of χ^2 [76] for the comparison of experimental and model-derived autocorrelation analyses.

Because the steady-state distributions and the temporal dynamics were measured using independent experiments, the total likelihood function to match both datasets is the product of the individual functions, and the total log-likelihood is the sum of the individual log-likelihoods:

$$\log L_{\text{total}}(\text{Dist}, G|M) = \sum_{\text{g}} (\log L_{\text{Dist}}(\text{Data}|\text{Model}) + \log L_{\text{AC}}(G_{\text{D}}|G_{\text{M}}(\wedge)), \tag{2.28}$$

for g = KDM5B, H2B and β-actin. Now, that we have defined a log-likelihood function to compare the data to the model under different parameter combinations, we can explore parameter space, first to maximize this likelihood and then quantify what is the uncertainty in parameters given the data.

Codon-dependent translation rates were assumed to be consistent among the three genes, as defined in Eq 2.4, but the three genes were allowed to have different initiation rates, $\{k_{\text{i}}^{(\text{g})}\}$. Under this assumption, the model has a total of four parameters. Upon fitting these parameters to maximize Eq 2.28, we found that the model could capture both the experimental distributions of nascent proteins per mRNA and the auto-covariance plots for all three genes, as shown in Fig. 2.4A-2.4C (middle and right). Optimized parameters and their uncertainties (see Methods) were found to be:

$$\begin{cases} \bar{k}_{\text{e}} = 10.6 + -0.72 sec^{-1} \\ k_{\text{i}}^{(\text{KDM5B})} = 0.022 + -0.004 sec^{-1} \\ k_{\text{i}}^{(\beta-\text{actin})} = 0.05 + -0.01 sec^{-1} \\ k_{\text{i}}^{(\text{H2B})} = 0.066 + -0.019 sec^{-1} \end{cases} \tag{2.29}$$

Exploring how translation dynamics vary with different parameters

After determining that our model was sufficient to reproduce the experimentally measured fluctuation dynamics for H2B, β-actin, and KDM5B, we next extended our analyses to consider a broader range of translation parameters. Specifically, we sought to explore the effects of variations to initiation and elongation rates as well as effects of synonymous codon substitutions or modulation of tRNA concentrations.

Ribosome collisions are rare at most experimentally observed translation initiation and elongation rates.

Previous experimental reports [23, 60, 24, 26, 25] estimated a range of values from 0.01 to 0.08 sec^{-1} for the translation initiation rate, k_{i}, and range from 3 to 13 aa/sec for the average elongation rate, $\bar{k}_{\mathrm{e}}$. Using β-actin gene as a reference, Fig. 2.5A depicts the variation in ribosome density as a function of the base parameters k_{i} and $\bar{k}_{\mathrm{e}}$, and Fig. 2.5B shows the number of times an average ribosome would collide with an upstream neighboring ribosome during a single round of translation. For most parameter combinations, ribosome loading was predicted to be very low (i.e., fewer than one ribosome per 100 codons) and collisions were rare (i.e., fewer than 10 collisions in an average round of translation). However, for slow elongation and fast initiation, such as those measured by Wang et al. [60]), a ribosome could collide with other ribosomes an average of $\sim$20 times for a gene the length of β-actin. To further illustrate the effects that these initiation and elongation rates would have on ribosome dynamics on different genes, Fig. 2.5C shows simulated kymographs for SunTag-24X-Kif18b [25], FLAG-10X-KDM5B [23], and SunTag-56X-Ki67 [26], each with their previously reported initiation and elongation rates. In addition, Fig. 2.12 and Fig. 2.13 Figs provide more detailed results of the translation elongation simulations for β-actin translation at multiple initiation rates and elongation rates, respectively. Each of these kymographs indicates that ribosome dynamics can vary from collision-free dynamics (SunTag-24X-Kif18b and FLAG-10X-KDM5B) to dynamics with multiple collisions (SunTag-56X-Ki67) and that collisions can become more prevalent at high initiation rates or low elongation rates.

Codon usage affects translation speed and ribosome loading.

Simulations of genes H2B, β-actin, and KDM5B showed that each gene's codon order influences the overall ribosome traffic dynamics, creating a non-uniform distribution of ribosomes along the mRNA (Fig. 2.6). This observation of codon dependence led us to look more deeply into possible effects that optimization could have on observable translation dynamics. Fig. 2.7 depicts simulated kymographs for the β-actin protein for three synonymous sequences containing: (i) natural codons, (ii) most frequent synonymous codon (optimized), and (iii) least frequent synonymous

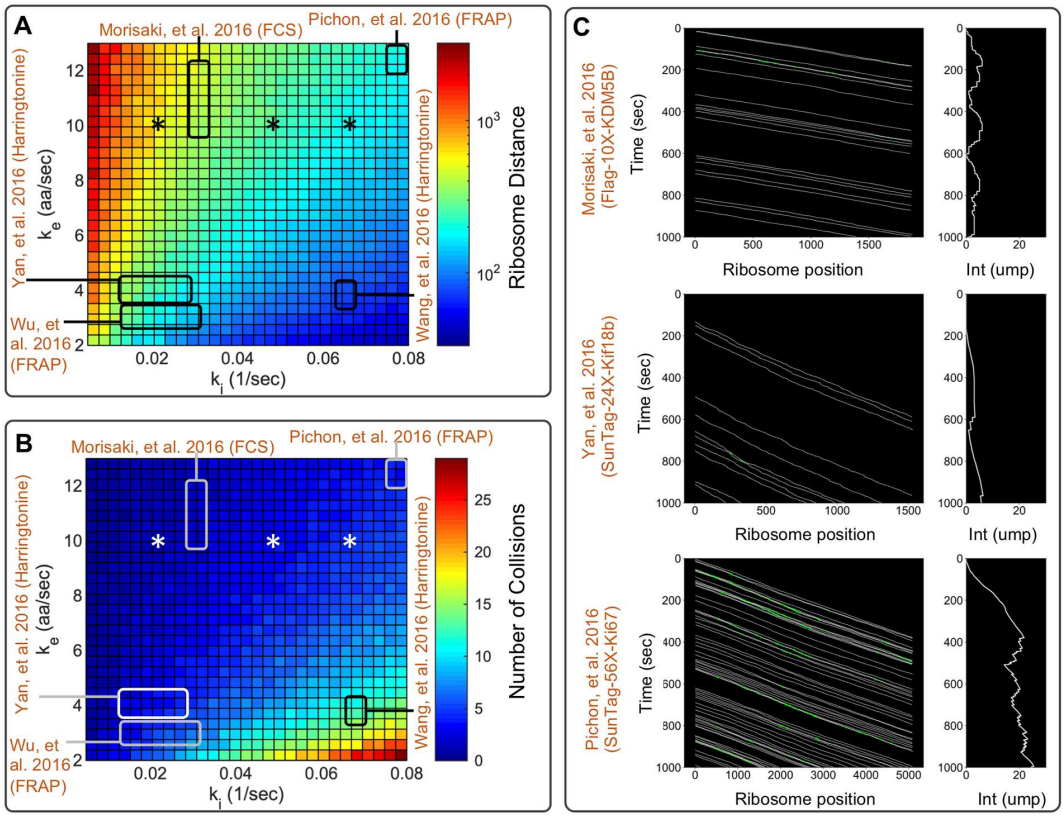


Figure 2.5: Ribosome dynamics under experimentally reported initiation and elongation rates. A) Simulated mean number of codons between ribosomes for the β-actin gene as function of initiation and elongation constants. In the plot, previous literature initiation and elongation values are highlighted by the squares [23, 60, 24, 26, 25], and values estimated in this study are denoted by asterisks. B) Simulated number of collisions per ribosome as a function of initiation and elongation constants. C) Top panel, kymograph showing the ribosomal dynamics for SunTag-24X-Kif18b using experimentally determined parameters k_{i} = 1/100 sec-1 and $\bar{k}_{\mathrm{e}}$ = 3.1 aa/sec [25]. Center panel, kymograph showing the ribosomal dynamics for FLAG-10X-KDM5B using experimentally determined parameters k_{i} = 1/30 sec-1 and $\bar{k}_{\mathrm{e}}$ = 10 aa/sec [23] Bottom panel, kymograph showing the ribosomal dynamics for SunTag-56X-Ki67 using experimentally determined parameters k_{i} = 1/13 sec-1 and $\bar{k}_{\mathrm{e}}$ = 13.2 aa/sec [26]. White lines in kymographs represent single ribosome positions, and green spots represent ribosome collisions.

codon (de-optimized). For each case, S4B Fig illustrates the corresponding ribosome loading profiles; S4C Fig shows the simulated distribution of FLAG intensities in units of mature proteins, and S4D Fig presents the corresponding simulated fluorescence auto-covariance functions. Fig. 2.8 and Fig. 2.9 show similar results for the H2B and KDM5B genes, respectively.

In all cases, optimized gene sequences speed-up ribosome dynamics, and de-optimized sequences cause a slower elongation rate that is observed in the auto-covariance plots given in S4D, Fig. 2.8D and Fig. 2.9D. Moreover, for constant initiation rates, faster elongation would lead to

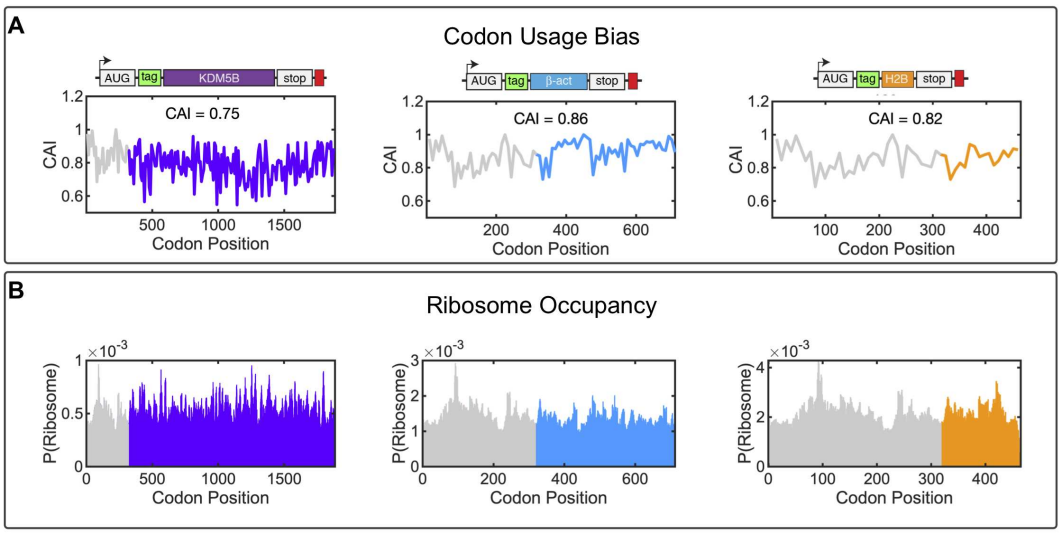


Figure 2.6: Codon usage and ribosome occupancy. Translation was simulated using the a β-actin gene, varying initiations rates from 0.03 to 0.6, a constant elongation (k_e = 10 aa/sec), and a ribosomal footprint of 9 codons. Top panels show a kymograph of the ribosome movement. Lower panels show the distribution of collisions for each k_i.

lower ribosome loading (S4B, Fig. 2.8B and Fig. 2.9B) and therefore lower fluorescence intensities, as shown in the distributions given in S4C, Fig. 2.8C and Fig. 2.9C. All three genes under consideration had natural codon usage that was enriched for the most common codons (i.e., the natural and common codon usage dynamics are very similar), such that the translation rate, ribosome loading, and fluorescence intensity could be substantially altered only by substitution to rare codons. We note that the substitution of rare codons would lead to slower elongation and substantially higher numbers of ribosome collisions.

Depletion of tRNA levels can induce ribosome traffic jams.

In addition to modulating translation speed through codon substitution, it is possible to perturb these dynamics through experimental modulation of tRNA concentrations. For example, Gorgoni et al., [13] used a mutated allele to the gene for $tRNA_{CUG}$ to reduce the concentration of the glutamine tRNA. To study how ribosome dynamics can be affected by the removal or addition of specific tRNA, we simulated the translation dynamics of H2B, β-actin, and KDM5B at several different concentrations for $tRNA_{CTC}$. Fig. 2.14 shows the effect of decreasing $tRNA_{CTC}$ concentration on the ribosome association time (left) and elongation rate (right). The simulations show

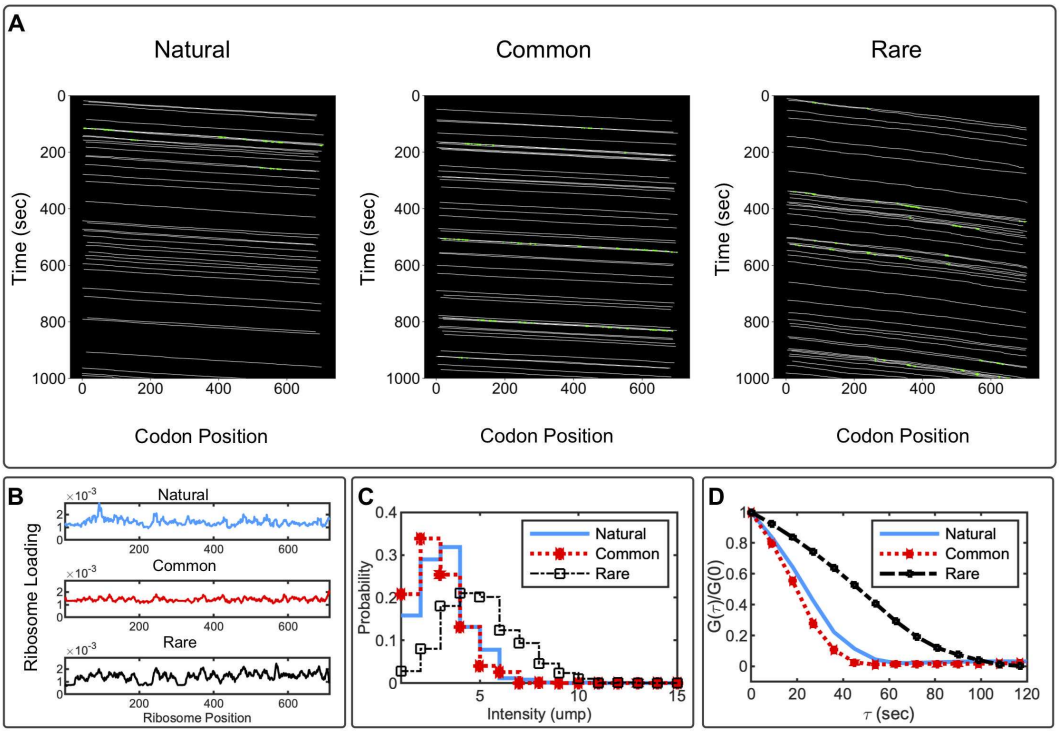


Figure 2.7: Codon optimization designs for β-actin. A) Ribosome dynamics for β-actin under different codon optimization constructs (natural sequence, using only common codons, and using only rare codons). In the kymographs, white lines represent the ribosome positions, green spots represent ribosome collisions. The average and standard deviation for the number of collisions is 3.2 ± 0.9 for the natural sequence, 2.4 ± 0.8 collisions for the optimized sequence (common codons), and 6.9 ± 1.5 collisions on the de-optimized sequence (rare codons). B) Ribosome loading for the three codon optimization constructs. D) Auto-covariances calculated for the natural gene sequence, a sequence where all codons are replaced by their most frequent synonymous codon (optimized), and a sequence where all codons are replaced by their less frequent synonymous codon (de-optimized). Simulations were performed using the optimized parameter values given in Eq. 2.29.

that ribosome dynamics are relatively unchanged provided that the tRNA_{CTC} concentration remains above approximately 10% of the native level. In contrast, depleting tRNA_{CTC} concentration below 10% of wild-type levels could lead to ribosome stalling, which was reflected in long ribosome association times and low effective elongation rates. Ribosome traffic-jams are observed under very low tRNA_{CTC} concentration as shown in Fig. 2.15 to Fig. 2.17. The prevalence of the CTC codon was found to be important in that the effect of tRNA_{CTC} depletion occurs at higher tRNA_{CTC} concentrations for the CTC codon rich KDM5B gene than for the other two constructs.

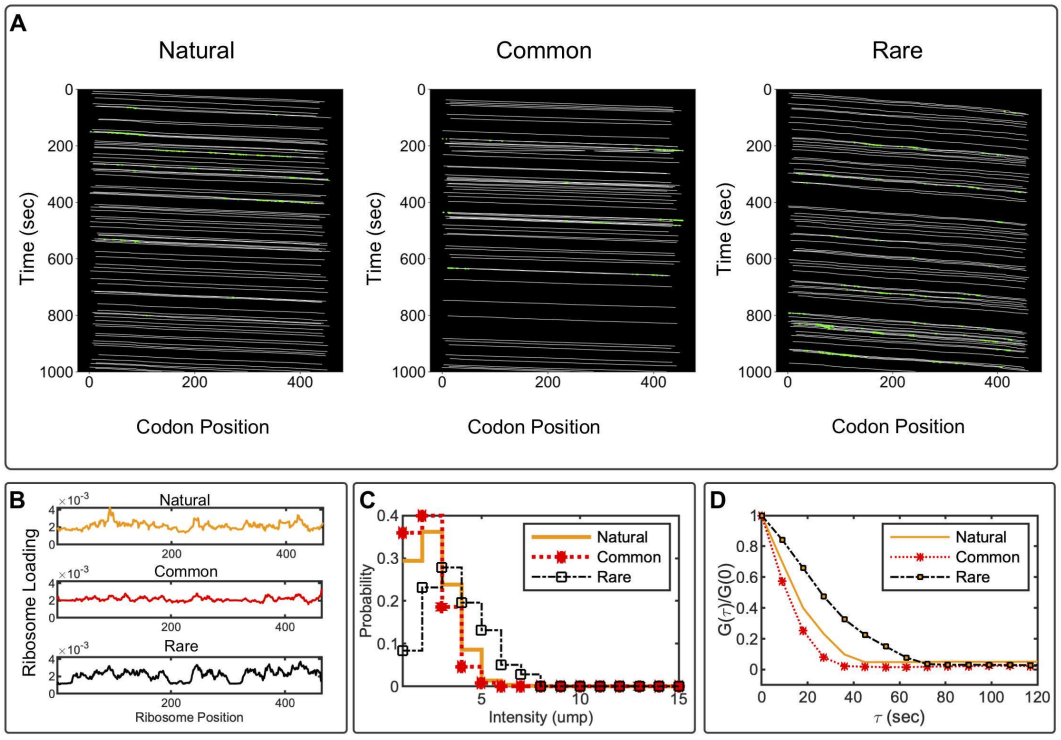


Figure 2.8: Codon optimization designs for H2B. A) Ribosomal dynamics for H2B under different codon optimization constructs (natural sequence, using only common codons, and using only rare codons). In the kymographs, white lines represent the ribosome placement, green spots represent ribosome collisions. The average and standard deviation for the number of collisions is 2.9 $\pm$ 0.7 for the natural sequence, 2.0 $\pm$ 0.6 collisions for the optimized sequence (common codons), and 6.0 $\pm$ 1.1 collisions on the de-optimized sequence (rare codons). B) Ribosome loading for the three codon optimization constructs. D) Auto-covariances calculated for the natural gene sequence, a sequence where all codons are replaced by their most frequent synonymous codon (common), and a sequence where all codons are replaced by their less frequent synonymous codon (rare). Simulations were performed using the optimized parameter values given in Eq. 2.29.

RNA sequence to NAscent protein simulation (rSNAPsim)

To facilitate the simulation of single-molecule translation dynamics, all models and analyses described above have been incorporated into a user-friendly Python toolbox, which we have called rSNAPsim. This toolbox combines a graphical user interface (GUI) divided into multiple tabs, graphical visualizations, and tables to present calculated biophysical parameters (see Fig. 2.10). This simulator performs stochastic simulations considering the widely accepted mechanisms affecting ribosome elongation, such as codon usage and ribosome interference. The toolbox is

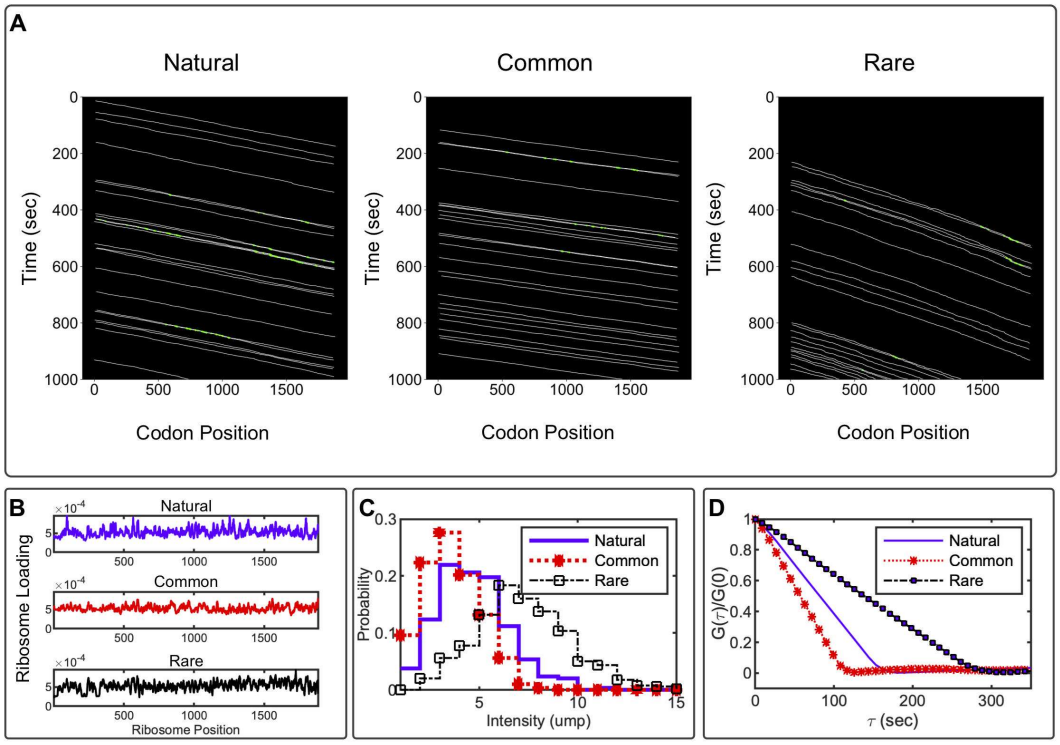


Figure 2.9: Codon optimization designs for KDM5B A) Ribosome dynamics for KDM5B under different codon optimization constructs (natural sequence, using only common codons, and using only rare codons). In the kymographs, white lines represent the ribosome placement, green spots represent ribosome collisions. The average and standard deviation for the number of collisions is 4.3 $\pm$ 2.0 for the natural sequence, 2.8 $\pm$ 1.7 collisions for the optimized sequence (common codons), and 7.8 $\pm$ 3.1 collisions on the de-optimized sequence (rare codons). B) Ribosome loading for the three codon optimization constructs. D) Auto-covariances calculated for the natural gene sequence, a sequence where all codons are replaced by their most frequent synonymous codon (common), and a sequence where all codons are replaced by their less frequent synonymous codon (rare). Simulations were performed using the optimized parameter values given in Eq. 2.29.

available in Python 2.7/3.5+ and wrappers for optimized C++ code are provided with installation instructions.

rSNAPsim takes as an input the gene sequence in Fasta format or an NCBI accession number. The user can decide on the type (FLAG, SunTag, or Hemagglutinin), number, and placement of different epitopes upstream, downstream or within the protein of interest. The toolbox provides the user with a visualization of the gene sequence and the overall gene construct including the position of the POI and the positions of the Tag epitopes. From the concatenated tags and POI sequences, rSNAPsim automatically generates a discrete single-RNA translation model with single

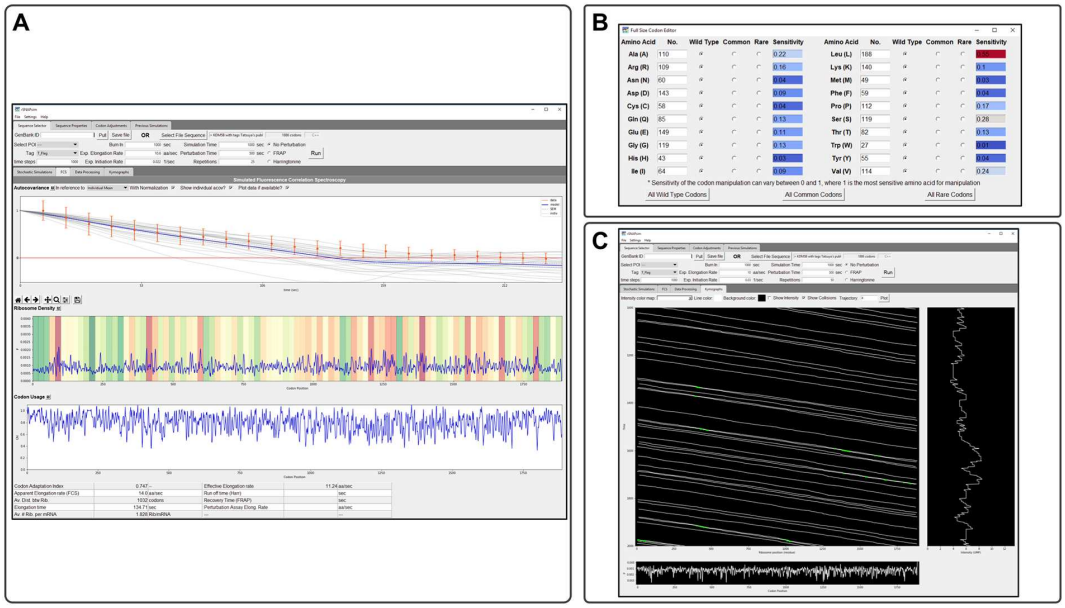


Figure 2.10: RNA Sequence to NAscent Protein Simulation (rSNAPsim). A) rSNAPsim is divided into four upper tabs and three lower tabs. Upper tabs allow the user to select and adjust sequences and then run simulations under varying conditions. Sequence selector allows the user to load a raw text file or GenBank file for their simulation needs. An option to pull from GenBank via accession number is available. All simulation parameters are also set on this tab. B) After a file is loaded, rSNAPsim allows the user to change the tRNA copy numbers and codon types under the Codon Adjustment tab. Post simulation, the lower tabs display simulation information such as average intensity over time of N simulations. C) Screen-shot of a kymograph. The kymograph tab allows the user to create their kymographs with varying display options. The Stochastic Simulation tab shows the time course data from the selected simulations. The Fluorescence Correlation Spectroscopy tab displays and compared simulated and experimental single-molecule translation dynamics, the auto-covariance function, and biophysical parameters, such as the elongation constant or ribosomal density. All functionality in the GUI is also available in a command-line module for Python included with rSNAPsim.

amino acid resolution and codon-dependent translation rates. Once generated, these models can be simulated using stochastic dynamics, and the results can be quantified in terms of predicted translation spot intensity fluctuations (i.e., single-RNA translation time traces or kymographs), ribosomal density profiles, and fluorescence signal auto-covariance. The graphical user interface also provides for easy generation of simulated results for several different experimental assays, including FCS, FRAP, and ROA. From these simulation results, biophysical parameters such as the overall elongation rate or ribosome association rate are automatically calculated and returned to the user. The toolbox provides additional interfaces for the user to design and simulate gene sequences with substitution between natural, common, or rare codons for any combination of amino acids or to

manually adjust the concentration of tRNA for specific codons. Simulations are saved automatically so that the user can compare translation dynamics for multiple different gene constructs. The toolbox allows for the user to load experimental single-mRNA fluorescence trajectories, compute auto-covariance functions with various normalization assumptions and compare these to model results. For example, the rSNAPsim screenshot in Fig. 2.10A shows a comparison of model and experimental normalized auto-covariances for KDM5B.

The open-source toolbox was tested in Mac, Windows, and Linux operating systems and is available at: `https://github.com/MunskyGroup/rSNAPsim.git`. Simulating a gene with 1567 codons for 100 repetitions of 5000 seconds each takes less than 1 minute using a laptop computer with a Core i7 and 32GB of RAM.

2.1.4 Conclusion

Imaging translation in living cells at single-molecule resolution is a new experimental technology that has been applied to only a few genes so far [23, 60, 24, 26, 25, 28, 63], but the number of such studies is expected to grow considerably in the near future [27]. Computational models can aid in this research by extracting improved biophysical understanding and parameters from single-molecule data. For example, in related analyses of transcription dynamics, Rodriguez et al., [42] used a coarse grained stochastic model to capture the polymerase elongation process and reproduce transcription dynamics for a multi-state promoter. Here, we extended that theoretical framework to include the most widely accepted mechanisms affecting nascent protein translation, including codon-dependent elongation and ribosome interference [64] and with specific attention to the placement of fluorescent probes. To complement previous models that have sought to reproduce data from earlier bulk cellular assays [13], and ribosome profiling data [77, 78], our focus has been to integrate single-mRNA stochastic dynamics models with data from *in vivo* single-RNA translation dynamics experiments.

We developed a general codon-dependent model, where nascent protein distributions and auto-covariance functions were generated by detailed stochastic simulations that tracked the positions

of ribosomes relative to their neighbors. However, in the absence of perturbations to change initiation and elongation rates, most ribosomes do not encounter others during elongation (Fig. 2.2), at least not at currently accepted elongation and initiation rates from the literature [23, 60, 24, 26, 25]. This observation justifies an assumption of sparse ribosome loading and independent ribosome motion, which allow the linear reaction rate reformulation of the codon-dependent translation model into a simplified stochastic moment model and further reduction led to analytical expressions for the steady-state mean and variance of fluorescence in units of mature protein levels per mRNA (Eq 2.21) and for the decorrelation time (Eq 2.18). For initiation rates at or below reported experimental values, the simplified analytical model and the full model are in strong agreement (Fig. 2.2). However, increasing initiation rates relative to the base elongation rate, inserting more rare codons into the sequence, or depleting tRNA levels for some codons will increase the number of ribosome collisions and violate the simplifying assumptions (Figs 2C and 5). In such circumstances, the full stochastic model predicts slower effective elongation rates, longer ribosome association times, and accumulation of more ribosomes per mRNA.

With the full and reduced models in hand, it becomes possible to predict how well three modern methodologies would estimate elongation rates from single-molecule measurements: Fluorescence Correlation Spectroscopy (FCS) [27], Fluorescence Recovery After Photobleaching (FRAP) [24, 26, 27], and Run-Off Assays (ROA) after perturbation with inhibitory drugs [60, 25]. Through simulations on 2,647 genes, we demonstrated that estimating elongation rates for long genes ($>$1000 codons) could be achieved with great accuracy using any of these methodologies, provided that a minimal number of mRNA spots are considered and with an appropriate temporal resolution as demonstrated in Fig. 2.3. However, our results suggest that FCS would be the most likely method to provide an accurate elongation rate estimate (Fig. 2.3A), especially for small and medium size genes. Although our simulation results suggest that FCS is the best single-molecule option to estimate elongation rates, it is important to remark that FCS analysis requires the tracking and measurement of intensity for single spots over long periods of time, and such measurements are susceptible to photobleaching and molecular motion. The former issue has been addressed through

the application of optical techniques such as highly inclined thin illumination microscopy [79] and the latter could be addressed through the application of molecular tethers to reduce motion [25]. On the computational side, one could potentially address concerns of bleaching or motion relative to the imaging plane by including hyper-parameters to describe these dynamics and then fit these hyper-parameters concurrently with model parameters using Bayesian analyses.

Run-off assays using Harringtonine to prevent translation initiation can give accurate estimates when genes are long, but the accuracy of such an approach is highly diminished for shorter genes (Fig. 2.3B) or if the precise time of drug action on the mRNA is not known. Our analyses suggest that run-off assays directly depend on the number of ribosomes actively translating the mRNA at the time of perturbation, and since this number is highly susceptible to stochasticity on small genes, the ROA would require analysis of a much larger number of spots to achieve accurate results.

Our analyses show that FRAP gives poor estimates for all genes of all sizes, and for all tested experimental designs, Fig. 2.3C. The recovery of the intensity after photobleaching depends heavily on the initiation rate, which has been found to be an order of magnitude smaller than the elongation rate, making the recovery a highly stochastic process as well. We directly compared the error size for the studied methods, obtaining that the error in FRAP and ROA is two times larger than in FCS, Fig. 2.11.

Using FCS data, we demonstrated that a codon-dependent translation model containing one universal average elongation rate and one gene-dependent initiation rate could capture quantitatively the distribution of nascent proteins per actively translating mRNA, as well as the temporal dynamics, for three different genes expressed in human U2OS cells (Fig. 2.4). Combining these estimates of initiation and elongation rates with reported values for the same rates identified using other methods and for other genes, we could predict ribosome dynamics and nascent protein intensities for reported gene sequences [23, 60, 24, 26, 25, 28, 63], (Fig. 2.5). Those results allowed us to conclude that relatively fast elongation rates help maintain substantial space between ribosomes on a single mRNA. As a result, these ribosomes should not often collide, and the final ribosome-mRNA association times should remain unchanged for typical initiation rates, natural

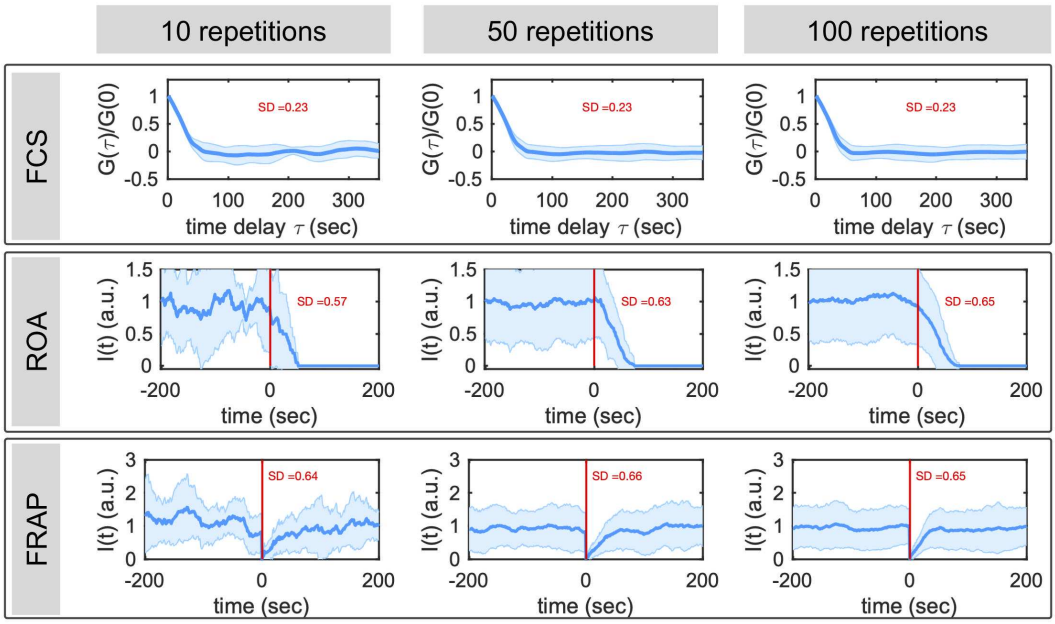


Figure 2.11: Error size for the different methodologies used to calculate elongation rates. Translation was simulated using the a β-actin gene with the optimized parameter values given in Eq. 2.29. Error bars represent the standard deviation (SD) of the number of repetitions given at the top of each plot. Vertical red lines represent the application of Harringtonine for ROA. Vertical red line represents the time of photobleaching for FRAP.

codon usage, and normal tRNA availability, as shown in Fig. 2.6. Nevertheless, ribosome dynamics may be affected by genetic or environmental perturbations, such as increased initiation rates (S1 Fig), reduction of elongation rates (S2 Fig), enrichment for rare codons (Fig. 2.7 to Fig. 2.9), or depletion of tRNA (Fig. 2.14 to Fig. 2.17).

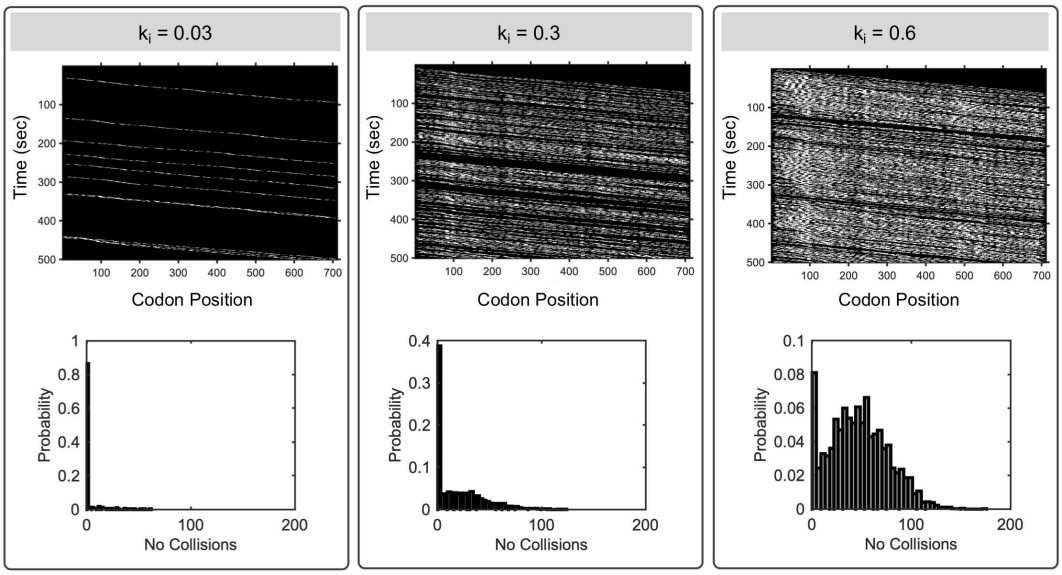


Figure 2.12: Effect of initiation rate on ribosome dynamics. Translation was simulated using the β-actin gene, varying initiations rates from 0.03 to 0.6, a constant elongation (k_{e} = 10 aa/sec), and a ribosomal footprint of 9 codons. Top panels show a kymograph of the ribosome movement. Lower panels show the distribution of collisions for each k_{i}.

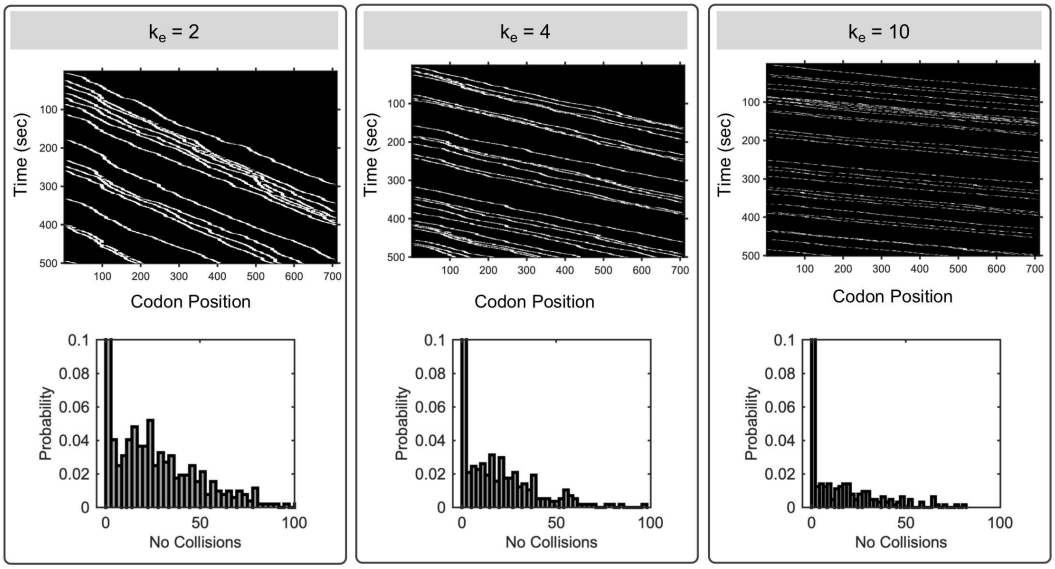


Figure 2.13: Effect of elongation rate on ribosome dynamics. Translation was simulated using the β-actin gene, varying elongation rates, a constant initiation (k_{i} = 0.06 sec^{-1}), and a ribosomal footprint of 9 codons. Top panels show a kymograph of the ribosome movement. Lower panels show the distribution of collisions per each k_{e}.

The present model and rSNAPsim toolkit have intentionally been made as general and adaptable as possible to efficiently simulate and capture the most accepted mechanisms taking place

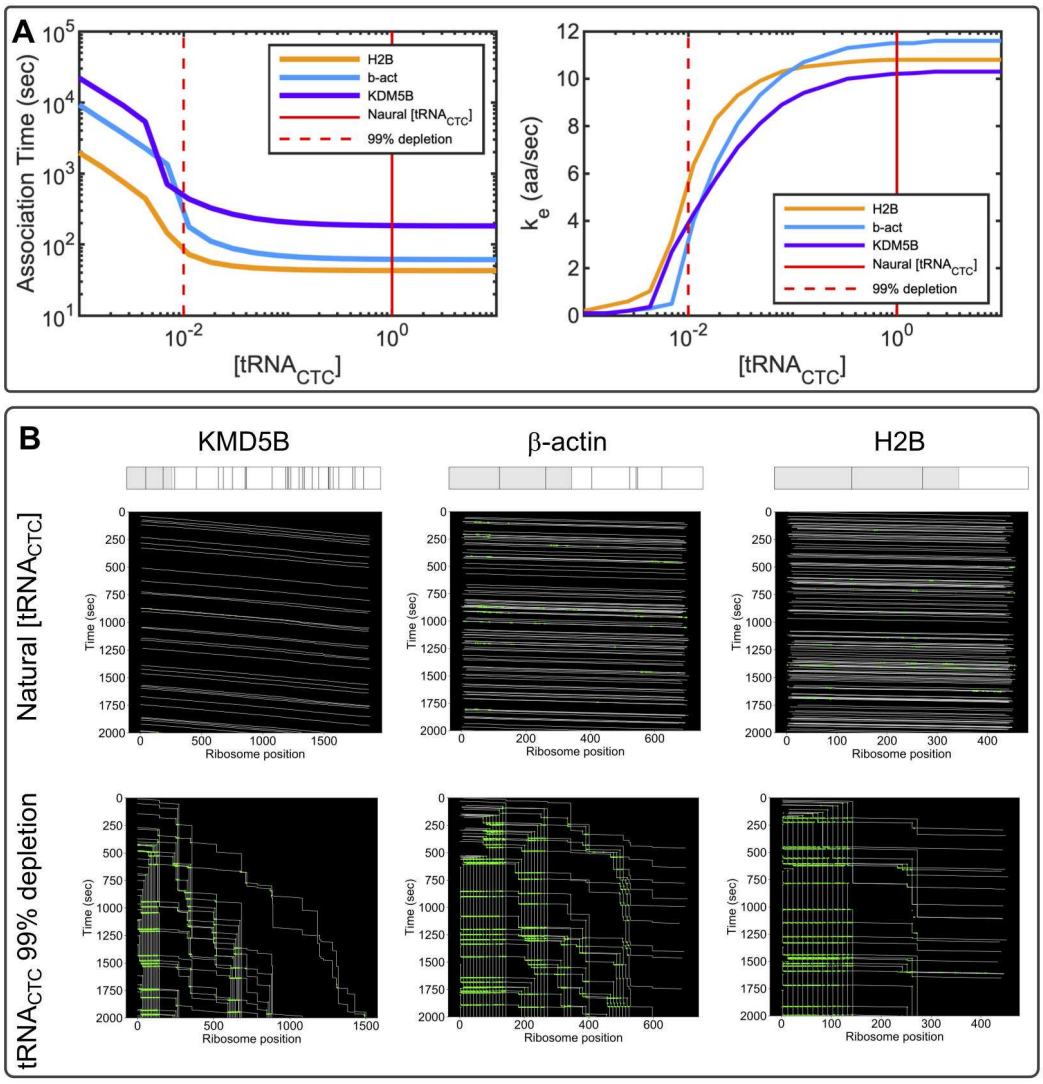


Figure 2.14: Effects of tRNA depletion on ribosomal dynamics. A) Three different genes were studied: KDM5B (magenta), β-actin (cyan) and H2B (orange). Left plot shows the ribosome association time as a function of the tRNA_{CTC} concentration. Right plot, shows the calculated elongation rates estimated by dividing the gene length by the average time needed by the ribosome to complete a round of translation. B) Kymographs show the ribosomal dynamics without depletion (upper panels) and with 99% depletion of tRNA_{CTC} (lower panels). Above the kymographs, the bar represents the studied gene, and the gray area represents the tag region, black lines denote the positions CTC codons. The frequency of the CTC codon is 29 for KDM5B, 8 for β-actin and 2 for H2B. Simulations were performed using the optimized parameter values given in Eq. 2.29.

during translation, i.e. codon-dependent elongation and ribosome interference. At present, the specific rates of codon-dependent elongation are only approximate and based on the prevalence of the corresponding tRNA in the human genome [13].

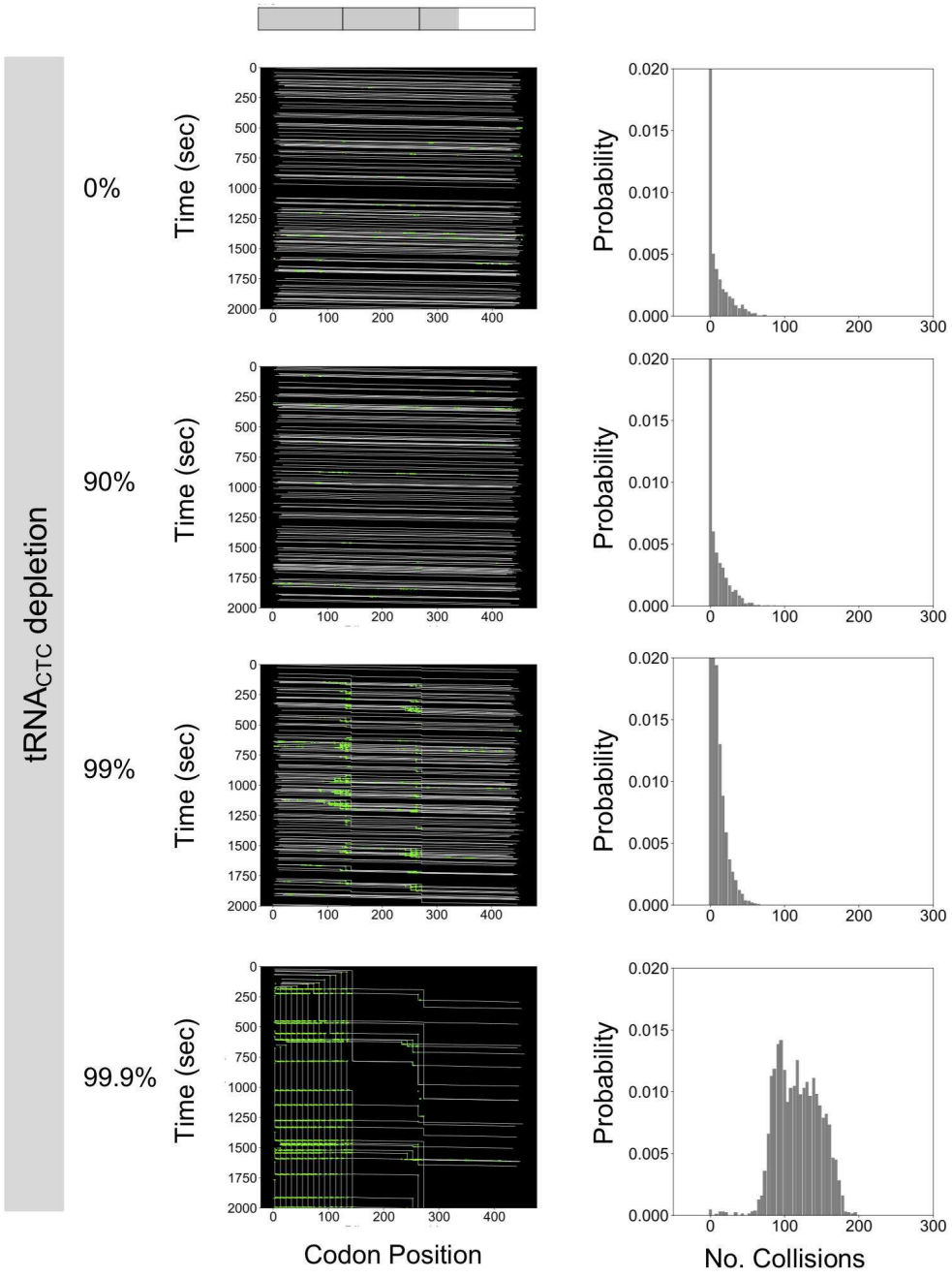


Figure 2.15: Depletion of specific tRNA$_{CTC}$ for H2B. Kymographs (left) show the simulated ribosomal dynamics under different percentages of depletion of tRNA$_{CTC}$. At the top of the kymographs, the bar represents the studied gene, the gray area represents the tag region, and black lines denote the positions of CTC codons. Histograms (right) show the probability of ribosomal collision. Simulations were performed using the optimized parameter values given in Eq. 2.29.

By modifying this assumption, it is possible to further improve fits for the elongation dynamics shown in Fig. 2.4, and one could find codon dependent rates to explain the diversity of experimen-

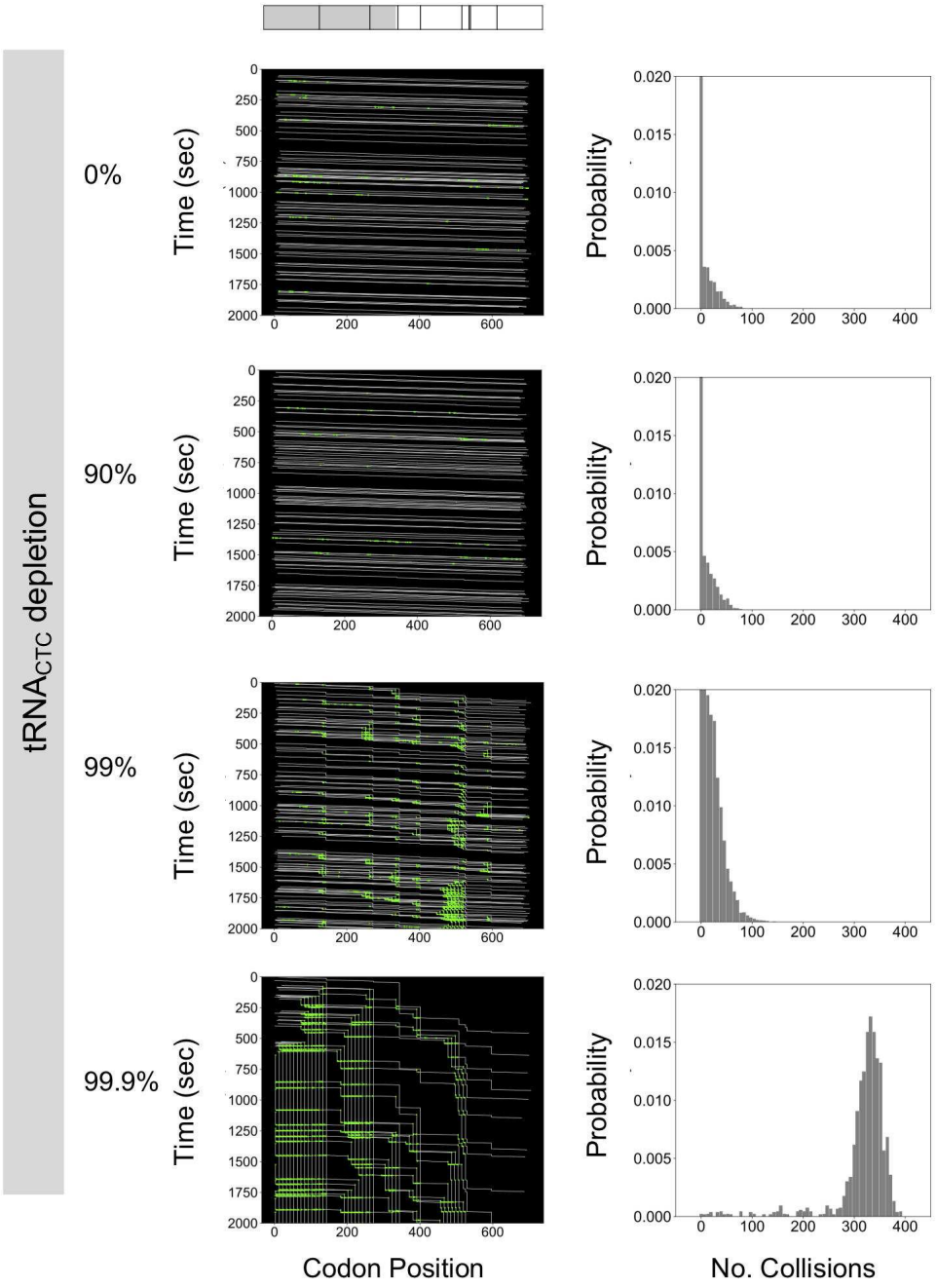


Figure 2.16: Depletion of specific tRNA_{CTC} for β-actin. Kymographs (left) show the simulated ribosomal dynamics under different percentages of depletion of tRNA_{CTC}. At the top of the kymographs, the bar represents the studied gene, the gray area represents the tag region, and black lines denote the positions of CTC codons. Histograms (right) show the probability of ribosomal collision. Simulations were performed using the optimized parameter values given in Eq. 2.29.

tally measured elongation rates depicted in Fig. 2.5. For now, we argue that data from fewer than a dozen genes (and in different cell lines) is as yet insufficient to fully constrain codon dependent rates for all 64 codons. However, as new data is collected for more and more genes, we envision that it will become possible to tune these parameters with greater precision and to capture a greater complement of genes.

In addition to variation in initiation, elongation, codon usage, and tRNA concentrations, many other factors have been described to affect ribosome dynamics. These include, but are not limited to, ribosome stalling or drop-off, pauses due to secondary structures of the specific mRNA, and the electrostatic and hydrophobic interactions between the mRNA and the ribosome [64, 78]. We expect that the increased prevalence of single-RNA translation experiments will add to the current understanding and reveal additional mechanisms taking place during translation. At the same time, such discoveries are bound to create new layers of model complexity. Although these mechanisms have not yet been implemented in our present model, they can be captured easily through modification of the set of elongation parameters, $k_{\mathrm{e}}(i)$. For example, the rSNAPsim toolbox allows for direct modification of elongation rates at a specific codon, which can be used to mimic pauses at certain locations. Furthermore, all of the computational analyses described above are easily adapted to allow for analysis of simultaneous multi-frame translation dynamics (e.g., when translation occurs on overlapping open reading frames as is the case during frame-shifted translation), as we implemented and described in [28]. Similarly, the code is easily extended to analyze translation of genes that contain more than one set of fluorescence tags in multiple colors, as has been explored experimentally in [63].

A main limitation in the experimental determination and quantification of translation mechanisms is the specific design of the experiment to make that quantification. For example, in its current form, the introduction of tag regions in the open reading frame of the gene of interest can dramatically alter the overall translation dynamics. As depicted in Fig. 2.1B, the tag region is around 300 codons in length, and this added length can substantially bias the measurement biophysical parameters, especially when quantified using FRAP or run-off assays (see Fig. 2.3). On

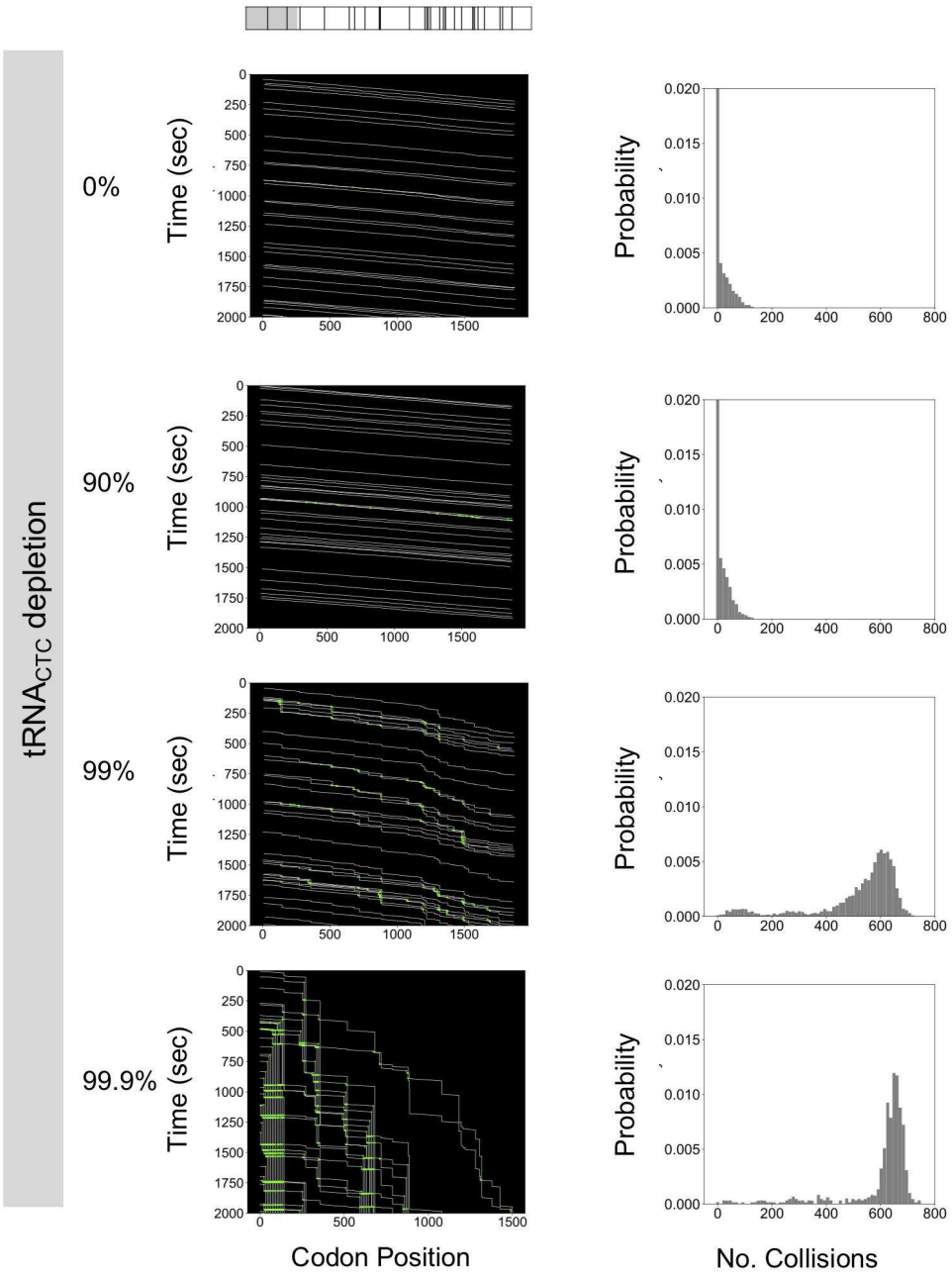


Figure 2.17: Depletion of specific tRNA$_{CTC}$ for KDM5B. Kymographs (left) show the simulated ribosomal dynamics under different percentages of depletion of tRNA$_{CTC}$. At the top of the kymographs, the bar represents the studied gene, the gray area represents the tag region, and black lines denote the positions of CTC codons. Histograms (right) show the probability of ribosomal collision. Simulations were performed using the optimized parameter values given in Eq. 2.29.

the one hand, our model can help to explain these differences (Fig. 2.11), but more importantly, the models themselves can be used to simulate and evaluate different computational designs to

determine which are more likely to reveal important biophysical mechanisms or parameters. We envision that user-friendly simulations, such as those provided by rSNAPsim, can be used to optimize combinations of probe placement, gene length, codon usage differences, video frame rates, drug-based perturbations, or specifications of movie length.

Such simulation-based designs can be conducted prior to any new experimental analysis and then used again to fit the results of those experiments, to pinpoint discrepancies that may reveal new mechanisms, and to refine model parameters and mechanisms. Such integration of experiment and computational model can help set the stage for more efficient experiments that specifically target and quantify the full complement of factors that modulate translation dynamics in living cells.

2.1.5 Methods

Studied gene constructs

To constrain our analyses, we use published gene sequences used on single-molecule translation studies. An initial set of sequences were obtained from Morisaki et al., [23], these constructs encode an N-terminal region with 10 repeats of FLAG-SM-tag (318aa) followed by one of three different genes of interest: KDM5B (1549 aa), β-actin (375 aa) and H2B (128 aa), the 3' UTR region contains 24 repetitions of the MS2 stem-loops. A second source of gene sequences comes from Yan, et al., [25], this gene construct encodes 24 repeats of SunTag followed by the gene of interest kif18b (1800 aa), and the 3' UTR contains 24 repeats of the PP7 bacteriophage coat protein. A sequence encoding 56 SunTag repeats, the gene of interest Ki67 (3177 aa), and the 3' UTR containing 132 repeats of MS2 stem-loops was obtained from Pichon et al., [26]. Finally, multiple gene constructs were build using 10 repeats of FLAG-SM-tag followed by a human gene. The studied human genes come from a comprehensive list of 2,647 gene sequences obtained from the PANTHER database [71].

Correction to mean and variance of fluorescence intensity for the theoretical model

Neglecting ribosome exclusion, and under an assumption of memory-less initiation with exponential rate k_{i}, the number of ribosomes to initiate translation in a fixed time, τ, is described by

a Poisson distribution with mean and variance equal to $k_{\mathrm{i}}\tau$. For a single probe site, we can fix τ as the time it takes a ribosome to move from that site to the end of the mRNA, and the mean and variance of nascent protein fluorescence can be estimated in terms of units of mature protein fluorescence according to Eq 2.21.

However, for probes that are spread out across a finite tag region, this distribution requires a slight correction to account for ribosomes within the probe region that only exhibit partial protein fluorescence. Let $\alpha(s)$ denote the intensity, scaled in units of mature protein, exhibited by a ribosome at the position, s, along the mRNA as follows:

$$\begin{cases} \alpha(s) = \begin{matrix} s/L_{\mathrm{T}} \text{ for } 0 \leq s < L_{\mathrm{T}} \\ 1 \text{ for } L_{\mathrm{T}} \leq s < L \end{matrix} \end{cases} \tag{2.30}$$

Under an assumption of uniform codon usage, a given ribosome on the mRNA has equal probability to be at any site along the mRNA. If there are an average of μ mRNA total on the mRNA, then the number at each location is approximated by a Poisson distribution with mean and variance both equal to $\mu/L \cdot ds$. Recall that the mean of the sum of two independent random variables is the sum of two means. Therefore, to find the total mean intensity contribution for all ribosomes on an average mRNA (Eq 2.22), we can integrate along the length of the mRNA to find:

$$\mu_{\mathrm{I}} = \int_0^{\mathrm{L}} \frac{\mu}{L} \alpha(s) ds, \tag{2.31}$$

$$= \left(1 - \frac{L_{\mathrm{T}}}{2L}\right) \mu \tag{2.32}$$

Similarly, we recall that the variance of a random variable with variance σ^2 and scaled by α is equal to $\alpha^2\sigma^2$ and the variance for the sum of two such variables is the sum of the corresponding variances. Therefor, by noting that $\mu = \sigma^2$, we can find the total variance of intensity on a single mRNA (Eq 2.23) as:

$$\sigma_{\mathrm{I}}^2 = \int_0^{\mathrm{L}} \frac{\mu}{L}\alpha(s)^2 ds, \tag{2.33}$$

$$= \left(1 - \frac{2L_{\mathrm{T}}}{3L}\right)\mu \tag{2.34}$$

Fluorescence Correlation Spectroscopy (FCS)

FCS is usually implemented by computing and comparing the auto-covariances (or autocorrelations) of fluorescence intensities of one or more particles within small fixed volumes [80, 81], but similar correlation analyses have been used to quantify intensity fluctuations for tracked single particles [56]. For our analysis, we compute the temporal auto-covariance times of the FLAG fluorescence signal intensity for a moving volume that is centered around the moving RNA spot.

To estimate the rate of translation elongation, we took the following approach: first, each experimental and simulated intensity time courses were centered to have zero mean by subtracting the average intensity of the time series, and then we normalize with respect to the standard deviation. Next, we computed the covariance function of the fluorescence intensity for each intensity spot according to the standard formula:

$$G(\tau) = \mathbb{E}\{(I_{\mathrm{t}} - \mu_{\mathrm{t}})(I_{\mathrm{t}+\tau} - \mu_{\mathrm{t}+\tau})\}, \tag{2.35}$$

where τ denotes the time delay and $\mathbb{E}\{v\}$ denotes the expectation of some arbitrary value v.

To reduce the effects of high-frequency shot noise and tracking errors that are not considered in the model, the zero-lag covariance $G(0)$ was removed from the analysis [82]. For simulated data, we normalize the auto-covariance function by the simulated variance, $G(0)$, which we can compute directly. For the experimental data, we cannot measure $G(0)$ directly because it is dominated by shot noise, so we instead interpolate $G(0)$ using a linear interpolation of the first four points of the measured auto-covariance function. For statistical purposes, auto-covariances for multiple intensity time courses were calculated, and their value was averaged. Final results are reported as mean values and standard error of the mean (SEM). This signal analysis allowed us to measure

the dwell time (τ_{FCS}) at which $G(\tau) = 0$, from which the average ribosome elongation rate can be calculated as:

$$k_{\mathrm{e}}^{(\mathrm{FCS})} = L/\tau_{\mathrm{FCS}} \tag{2.36}$$

Parameter uncertainty

Parameter uncertainty analyses were calculated by building parameter distributions that reproduce results within a 10% error, calculated from 1,000 independent simulations using randomly selected parameter values. Simulations were performed on the W. M. Keck High-Performance Computing Cluster at Colorado State University.

Numerical methods

For solving the model under stochastic dynamics we used the direct method from Gillespie's algorithm [4] coded in Matlab 2018b and Python 2.7. ODE models were solved in Python 2.7.

Codes and experimental data

All codes and experimental data are available at:
https://github.com/MunskyGroup/Aguilera_PLoS_CompBio_2019.git.

Contributions

Luis U. Aguilera: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review and editing

William Raymond:, Investigation, Software, Validation, Visualization

Zachary R. Fox: Investigation, Software

Michael May: Investigation

Elliot Djokic: Investigation

Tatsuya Morisaki: Data curation, Investigation, Writing - review and editing

Timothy J. Stasevich: Funding acquisition, Project administration, Supervision, Writing - review and editing

Brian Munsky: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review and editing

2.2 Generalizations to the rSNAPsim software package to allow for model building, experiment design, and multiple fluorescent colors.[2]

This chapter describes the development of the rSNAPsim as a Python software package after the 2019 paper, its current functionality, and some example uses of the package. After the 2019 paper, our lab sat down and decided what to do with the rSNAPsim and its future directions. Our lab had a pressing need for a generalized modelling and analyses software; When working with experimental collaborators, we often had to write custom code and start computational modeling from scratch each time —which screams for a need of a unifying software to save time and effort. Thus, the goals for the rSNAPsim v2.0 were and are as follows:

- Greater generalizability of mRNA models
- Usability, open source, and versatility
- Unification of multiple analyses used in our lab

rSNAPsim has been updated to allow for arbitrary probe placements, probe sequences, and as many colors as the user desires to simulate intensity trajectories for (albeit in practice 3-4 colors are resolvable in NCT). Ribosomal movement is is not locked to human codon frequency scaling, but can accommodate any arbitrary stepping rules the user desires. Lots of general functions are now included that are commonly needed for NCT analyses such as intensity covariances, sequence optimization, and sequence statistic calculations. One of the core functionalities now included is a general TASEP model builder that allows multiple models to be designed efficiently. Our model builder is indispensable since our work with experimentalist collaborators often involves a request to provide multiple models, which would have had to been hand-written each time. Over the past 5 years, the model maker has been revised three times to accommodate requests for more and more complicated mRNA models. The first version of the model maker in 2020 only provided custom ribosomal movement (arbitrary jumps, enters and exits). When presented to our collaborators in

[2]William S. Raymond, Luis U. Aguilera, Brian Munsky

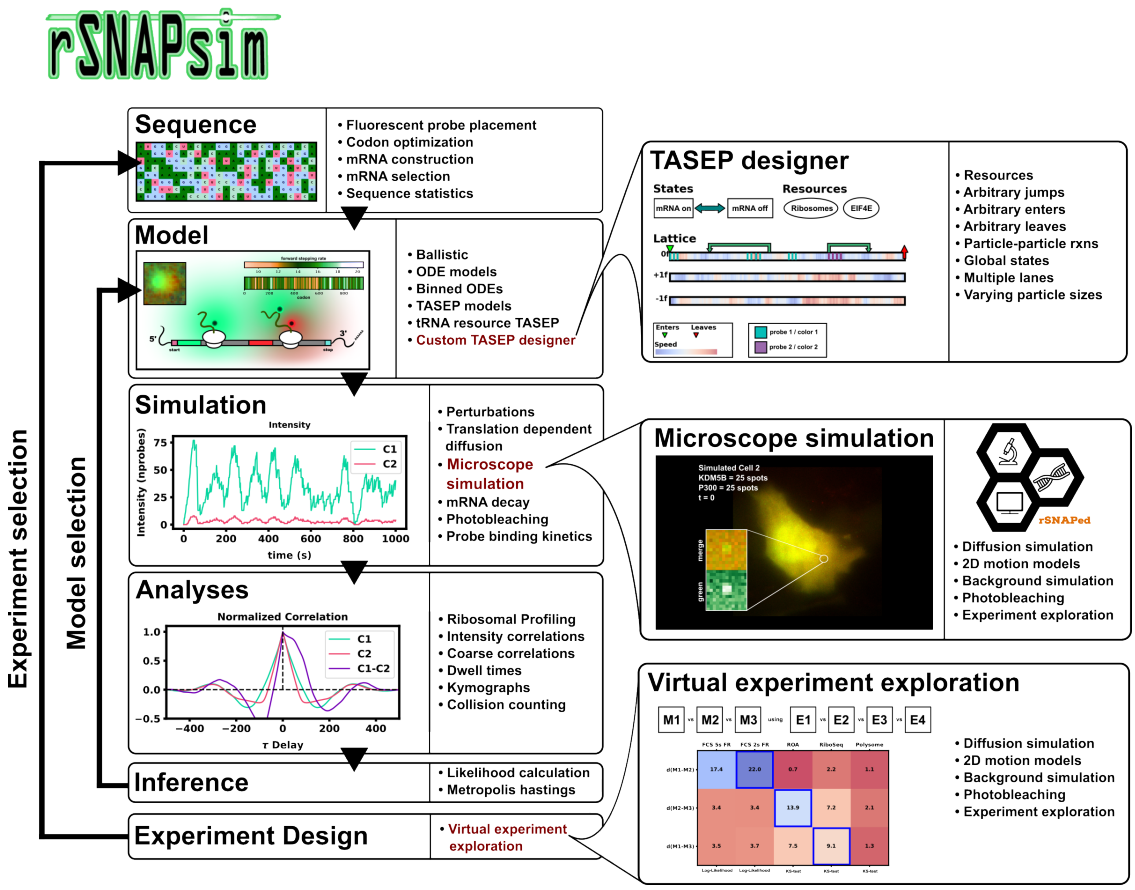


Figure 2.18: Graphical description of the rSNAPsim software environment.

the Stasevich lab, they instantly suggested a model for viral protease cleavage on nascent peptides which was not included. Rounds of back and forth coding and "focus testing" eventually lead our current general TASEP builder which is versatile enough to be used for almost any NCT experimentation (the main current limitation is if any backwards movement is included the simulation can no longer keep track of what protein was made).

In addition to model solving and analyses, rSNAPsim provides functionality to do model inference, model selection, and experiment design. The entire package is meant to be a start to finish modeling software for our lab; See Fig. 2.18 for a graphical depiction of the Python package's key modules. Another extension developed by Dr. Luis Aguilera called rSNAPed (RNA Sequence to NAscent Protein Experiment Designer) for microscope simulation will be covered in detail in Chapter 2.3.

2.2.1 rSNAPsim provides easy interface to generate sequence based mechanistic models for NCT experiments and mRNA translation.

General workflow for rSNAPsim models involves providing DNA or RNA sequence data (.gb, .txt, .dna, .fa) containing some open translation reading frames (ORFs) which can be converted to waiting times or stepping rates for ribosomes along a given mRNA. Without sequence data, a

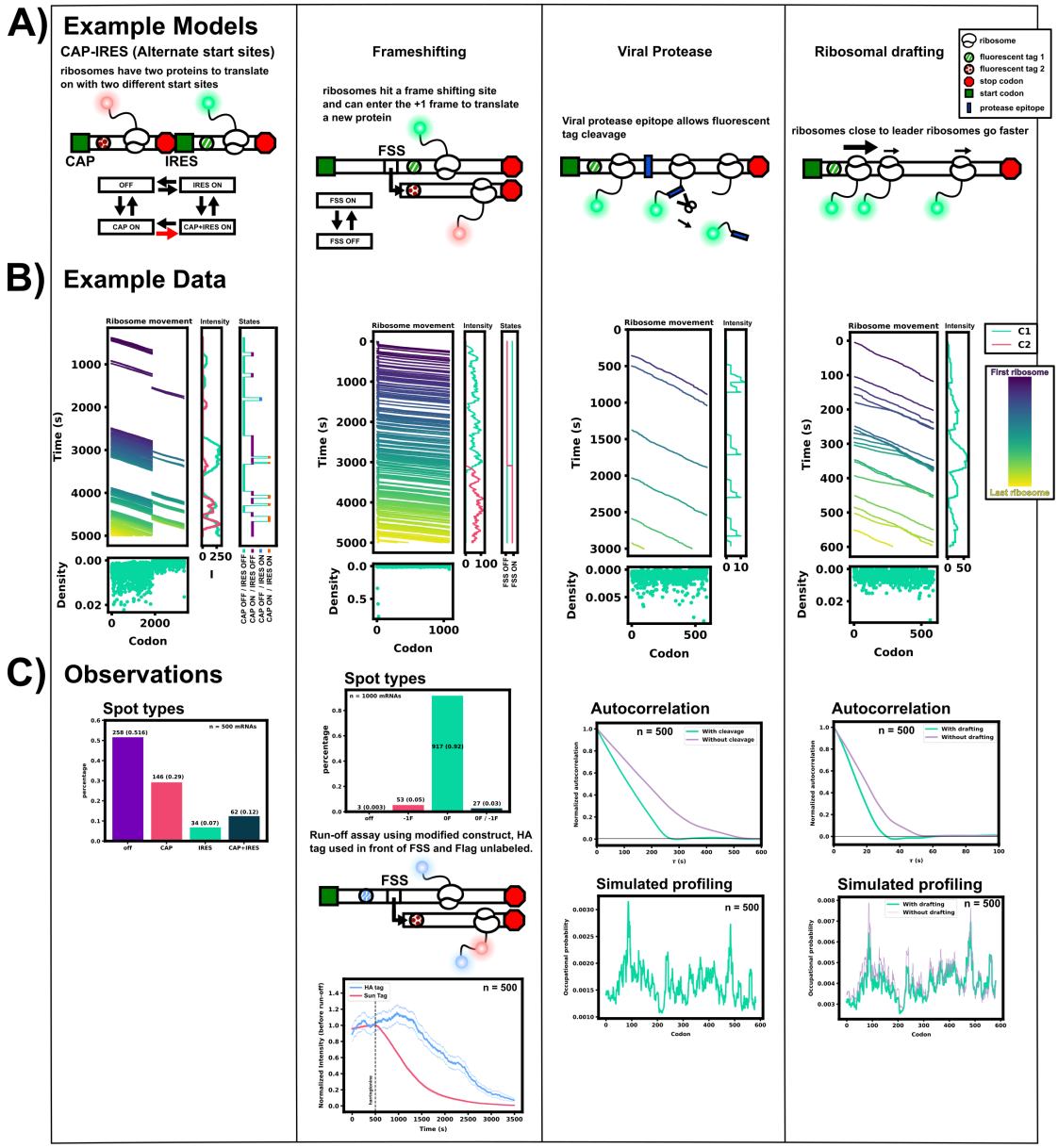


Figure 2.19: Caption on next page.

Figure 2.19: Four example models created with the rSNAPsim TASEP model maker. A) Four different models are shown to highlight the versatility of the rSNAPsim TASEP model maker: CAP-IRES, frameshifting, viral protease epitopes, and ribosomal drafting. CAP-IRES and frameshifting use the models described in [35] and [28] respectively. The CAP-IRES model features two separate open reading frames, one mediated by CAP-dependent translation initiation, the other by IRES (internal ribosomal entry site) initiation. Two open reading frames each have a separate fluorescent tag, (FLAG in 0 frame and SunTAG in -1 frame), enabling a readout of translation in both ORFs. The frameshifting model contains a Frameshift Sequence (FSS) hairpin that causes pausing at that location. Depending on the mRNA state, ribosomes either stay in frame 0 or shift to the -1 frame after the pause, subsequently translating different fluorescent color tags. The viral protease is a toy model that inserts an cleavage site which allows for enzymatic cleavage co-translationally of the nascent chain thus losing the upstream fluorescent tag. Ribosomal drafting is another toy model that speeds up downstream ribosomes if there is a leading ribosome within 100 codons. Ribosomes speed up proportionally to how far they are from the leading ribosome, resulting in trains of ribosomes all moving at a similar speed during translation. B) Example kymographs of ribosome movement, intensity trajectories, and state trajectories (if applicable) of each model. Each ribosome is colored individually by the heat-map within the kymograph. Below the kymograph is the occupation density of the mRNA over the codons. C) Example observations that can be drawn from the models. For the CAP-IRES and frameshifting model, observed spot types of 500 different spots were generated from the model to match their origin papers. For the toy models, normalized autocorrelation functions and ribosomal profiling are shown as potential data to collect from the model.

model can be built with stepping rates alone if desired. rSNAPsim detects open reading frames within the sequence data and the user selects a desired ORF to run translation simulations with; Alternatively, the user can construct a single lattice track from multiple ORF's stepping rates. This information is stored in a dedicated mRNA object. The default TASEP uses a particle size (ribosomal footprint) of 9 codons, a single initiation and termination site detected from the sequence and codon dependent stepping rates. Default stepping rates are calculated using the human Nakumura codon frequency [67] to scale an mRNA's fast or slow codons vs the average codon frequency. For this stepping rate calculation, A global elongation rate is selected by the user, $\bar{k_{\mathrm{e}}}$, and a given codon's rate, $k_{\mathrm{e,codon}}$, is calculated by:

$$k_{\mathrm{e,codon}} = \bar{k_{\mathrm{e}}} \cdot \frac{u_{\mathrm{codon}}}{\bar{u}} \tag{2.1}$$

Where u_{codon} is the codon frequency of a given codon, and $\bar{u}$ is the average codon usage across all codons in a given genome —this is the model from Chapter 2.1. The stepping rates can alternatively be provided directly to the model through any calculation as long as the length of the

stepping rates matches the number of codons in the mRNA; A different codon frequency or tRNA abundance dictionary can also used to calculate stepping rates, some of which are provided within a dictionary stored in the main module. Once the stepping rates, initiation rate, and termination rate are selected, the user can then decide simulation parameters such as simulation time or number of simulations for the TASEP before running stochastic trajectories. The core TASEP models are precompiled into a background C++ library for speed.

If the default TASEP model is not sufficient, the user can create their own models with the packages TASEP model builder. The model builder allows a multitude of models to be implemented quickly for model selection. In the simplest general terms, the model maker allows for any enter (particle arriving), jump (particle moving), leave (particle leaving), state change, or resource usage. Each of these events has a propensity function of the entire simulation: The particle (agent) array, defined rates, the current lattice state, time, the global state, the current resources, and the previous reactions.

$$\text{propensity}_\text{i} = f(\text{rates}_\text{i}, \text{ribosomes}, \text{lattice}, k_\text{elongation}, t, \text{probes}, \text{states}, \text{resources}, \#_\text{ribosomes}) \tag{2.2}$$

Each propensity pairs with a general stoichiometry that can change any counts of particles, states, probes, or resources during the simulation.

$$\text{stoichiometry}_\text{i} = [\Delta\text{ribosomes}, \Delta\text{states}, \Delta\text{resources}, \Delta\text{probes}] \tag{2.3}$$

Probe binding is dictated by any function of time or resources that ranges from 0 to 1. Whenever a ribosome encounters a probe location, a random uniform number is pulled and compared with this probe binding function and if successful (greater than the random number) a unit of fluorescence is added to that ribosome within the agent array.

$$\text{probe binding} = \max(\min(f(t, \text{resources}), 1), 0) \tag{2.4}$$

Taken together, these highly generalized propensities and stoichometries allow for a user to simulate complicated interactions such as ribosome-ribosome interactions, peptide cleavage, structure formation, frameshifts, and amino acid errors (and so much more). Figure 2.19A shows four different models generated with the model builder and their outputs. These models were selected to showcase the versatility of the current model builder software and its capability to reproduce reported literature models. Figure 2.19A shows cartoon diagrams of each model, CAP-IRES [35], frameshifting [28], a viral protease toy model, and a ribosomal drafting toy model.

CAP-IRES model

The CAP-IRES model replicated here is taken from Koch et al. [35] and uses an alternate start site NCT bicistronic construct with two separate open reading frames —one where ribosomal initiation is cap-dependent (CAP) and another mediated by an internal ribosomal entry site (IRES). Each ORF contains a different color fluoresecent tag (10xsmFLAG for CAP, 24xsmSun for IRES), allowing for investigation of the translation in each ORF of this construct simultaneously. The construct allows for investigation of translation status on local IRES elements, with the open question of "if one ORF was active does it turn off the other?" After performing model selection, the manuscript describes a best fit model of four states, S_{OFF}, $S_{\mathrm{CAP-ON}}$, $S_{\mathrm{IRES-ON}}$, $S_{\mathrm{CAP+IRES-ON}}$. CAP turning on and off and IRES turning off are both dictated by respective constant rates. However, IRES turning on has two separate rates based on whether CAP is actively translating or not, $k_{\mathrm{ON-I}}$ and $k'_{\mathrm{ON-I}}$ respectively. $k'_{\mathrm{ON-I}}$ is one magnitude larger than $k_{\mathrm{ON-I}}$, resulting in the novel dynamic where if CAP is on, IRES is much more likely to turn on and begin translating.

Fig. 2.19A (column 1) shows the model schematic and mRNA state diagram. Fig. 2.19B (column 1) shows an example stochastic trajectory of the model where all 4 states are observed. Fluorescence intensity readout displays both colors alone and simultaneously within the same mRNA over time on this example trajectory, as well as all state transitions. Fig. 2.19C (column 1) shows observations from the original manuscript recreated using the rSNAPsim model maker. Spot types were obtained by letting the simulations run to steady state and then recording the displayed intensity at a cut off time. Out of 500 mRNAs, 258 (51.6%) were off displaying no fluorescence in

NCT channels, 146 (29%) were undergoing CAP-dependent translation, 34 (7%) were undergoing IRES-dependent translation, and (62%) were translating in both open reading frames, roughly matching the reported spot type distribution from the original manuscript.

Frameshifting model

The frameshifting model is the two state bursting model used in Lyon et al. [28] to fit their frameshifting NCT construct. Two separate fluorescent tags, FLAG and SunTag, are encoded in the 0 and -1 frames downstream of a frameshift sequence (FSS) at codon 24. Ribosomes initiate at a rate $k_{\text{initiation}}$, and progress at a standard codon dependent elongation rate. The mRNA can exist in two states: FSS_{ON} where the frameshift sequence is active and folded, resulting in -1 frame proteins, and FSS_{OFF} where the frameshift sequence is unfolded and inactive resulting in 0 frame proteins. The mRNA is free to switch between the two states regardless of ribosomal load (ribosomes do not exclude secondary structure formation). Upon reaching the frameshift sequence (codon 24), ribosomes pause at two rates based on the FSS state, $k_{\text{FSS ON}}$ or $k_{\text{FSS OFF}}$. After passing the FSS, ribosomes proceed as their frame's codon dependent rates dictate. Exclusion is still enforced upon the ribosomes as different frames differ by a single nucleotide and both frames occupy the same physical space. Model rates were taken from Lyon et al. and used to reproduce their reported data using the rSNAPsim model maker.

Fig. 2.19A (column 2) shows the schematic of the model construct with FLAG tag in green and SunTag in red in the -1 frame. Fig. 2.19B (column 2) shows an example trajectory that matches the original model description. Ribosomes pause for a relatively long time at the front of the construct due to the FSS sequence, shown in the occupancy profile. The example kymograph also exhibits a state switch from $FSS\,OFF$ to $FSS\,ON$, resulting in a switch in fluorescent colors. Fig. 2.19C (column 2) shows the observations used in the original paper recreated with the rSNAPsim model. With 1000 simulated runs stopped at steady state, 917 (92%) of spots are translating in 0 frame, 53 (5%) are translating in -1 frame, 27 (3%) mRNAs recently switched states resulting in both colors present, and finally 3 (0.3%) mRNAs had intensities below a selected threshold of 5 fluorescent tags in both colors and thus were considered "Off."In the original text, the authors also used a modified

construct to investigate pausing at the frame shift site, this construct added a 10x Hemagglutinin tag in front of the FSS, moving the FSS to codon 370. This construct exhibits a "Battery"of ribosomes due to a ribosomal traffic jam built up before the slow FSS region, and when treated with harringtonine for a run-off assay, the HA tag fluorescence is retained much longer than the -1 smSun tag fluorescence. smFlag was not added to this experiment due to color resolvability limitations. rSNAPsim allows the model to be easily adjusted to this new experiment and the recreated average change in fluorescence after harringtonine application is shown for both color channels in the final panel of Fig. 2.3C (column 2).

Viral protease model

The viral protease toy model showcases a TASEP representing translation of a viral protease epitope downstream of the NCT fluorescent tag, analogous to polyprotein synthesis of many viruses (albeit with only one cleavage site on our model). Once synthesized, this epitope can be cleaved at a random rate defined by the user, removing the intensity associated with a given ribosome and lowering overall intensity of the tagged mRNA. This toy representation is highlighted here to showcase the new option of reactions that affect intensity and location based reactions that were not captured in previous rSNAPsim model maker versions. For the sequence used for the toy model, a 10xsmFLAG tag was added to the CCDS sequence from PRNP mRNA (PrP) for a total length of 591 codons with a simulated cleavage rate past codon 320. The additional propensity used is the cleavage rate times a boolean that a given ribosome is past codon 320:

$$wn_{\text{cleavage, i}^{\text{th}}\text{ ribosome}} = (x_{\text{i}} > \text{cleavage codon}) * \text{cleavage rate} \tag{2.5}$$

Where x_{i} is the location of the i^{th} ribosome. The stoichiometry for the above propensity removes 10 probes from a given ribosome since there are 10 upstream probes and a probe function of 1, guaranteeing all those upstream probes bind.

$$S_{\text{cleavage, i}^{\text{th}}\text{ ribosome}} = [0, 0, 0, ..., -10] \tag{2.6}$$

Fig. 2.19A (column 3) shows the schematic for the viral protease model. An example trajectory of the model with a high cleavage rate at codon 320 (immediately downstream of the last fluorescent epitope) is shown in Fig. 2.19B (column 3). The intensity is lost quickly for each individual nascent change once it passes the cleavage codon as this trajectory was simulated with a very high cleavage probability rate, on the same scale as the reactions of moving to the next codon. Fig. 2.19C (column 3) shows two example statistics of the model, the autocovariance and simulated ribosomal profile. The autocovariance from the fluorescence intensity exhibits a shorter dwell time due to the random loss of probes past the cleavage site. The nascent chain cleavage results in a "halved dwell time" when compared to the same model with no cleavage if the dwell time was calculated from the fluorescence correlation spectroscopy. The simulated ribosomal profiling shows no detectable differences due to the cleavage only occurring in the nascent chain.

Ribosomal drafting model

Particle-particle interactions are one of the new, powerful features of the model maker; Here we present a toy ribosomal drafting model where ribosomes speed up elongation to join a leader ribosome. Mechanistically it represents a leader ribosome that flattens or straightens mRNA and allows trailing ribosomes to translate faster, facilitating trailing ribosomes to "draft"off the leading ribosome. This creates signature tightly-packed polysomes of translating ribosomes all moving in sync with the leader ribosome.

The stepping rate propensity for ribosomal movement for the i^{th} ribosome is replaced with the following:

$$wn_{\mathrm{i^{th}\ ribosome}} = k_{\mathrm{c_i}} + f(x_{\mathrm{i+1}} - x_{\mathrm{i}}) \tag{2.7}$$

A default rate of $k_{\mathrm{c_i}}$ is adjusted with a function of distance of the i^{th} ribosome to the $i + 1^{\mathrm{th}}$ ribosome. For example, we could consider a value of ΔL codons to "lookahead" for each ribosome for the drafting influence and the closer the distance to the leading ribosome the stronger this influence is linearly. We can multiply this ratio with some scaling factor k_{scale}

$$f = k_{\text{scale}} \cdot \frac{(\Delta L - (x_{\text{i+1}} - x_{\text{i}}))}{\Delta L} \tag{2.8}$$

A value of 100 codons for ΔL was used for Fig 2.19's toy model. Fig. 2.19B (column 4) shows an example kymograph of the drafting model. The overall speed of the leading ribosomes is dictated by the stepping rate dictionary, but trailing ribosomes speed up and join the leading ribosomes if possible, resulting in close knit polysomes. Fig. 2.19C (right column) shows the drafting effect on autocovariance and ribosomal profiling vs no drafting. Simulated profiling shows a slight lower occupancy in the front of the construct due to ribosomes speeding up to join leader ribosomes as expected. Autocovariance shifts to the left representing a lower dwell time and a slightly higher average elongation speed, as the only time the ribosomes are actually faster than the default stepping rates is when they are speeding up to join the leader ribosome.

2.2.2 rSNAPsim provides multiple options to simulate experimental perturbations.

We can describe FRAP by the following per ribosome reaction for each i^{th} ribosome and j^{th} color for each tag color:

$$wn_{\text{i}^{\text{th}}\text{ ribosome, FRAP}} = k_{\text{bleach}} * (\text{probe}_{\text{j}} > 0) * (t > t_{\text{start}}) * (t < t_{\text{stop}}) \tag{2.9}$$

with a corresponding stoichiometry that removes one j^{th} probe from the i^{th} ribosome:

$$\text{probe}_{\text{j}} = \text{probe}_{\text{j}} - 1 \tag{2.10}$$

The above reaction is in words is "If this ribosome has a positive probe value between two time points, subtract a fluorescent probe."Coupling that logic with a high reaction rate will bleach probes quickly during that time range within the simulation. Fig. 2.20 left column shows an example simulation with a custom FRAP signal to bleach the cell for 200 seconds starting at 1000 seconds.

For simulating initiation inhibition, such as with harringtonine [83], a time based boolean can be added to the initiation rate in the model builder. Fig. 2.20 middle column shows an example simulation with a using the descending ramp inhibition starting at 1000 seconds.

$$wn_{\text{initiation}} = k_{\text{initiation}} \cdot (t > t_{\text{inhibition}}) \tag{2.11}$$

Different inhibitor signals could be used to fine tune inhibition or account for diffusion. For example, here is a propensity that uses a linear ramp to control initiation inhibition. At time 1000, the boolean multiplying the initiation rate takes 200 time units to transition from 1 to 0, allowing partial inhibition for a short period of time.

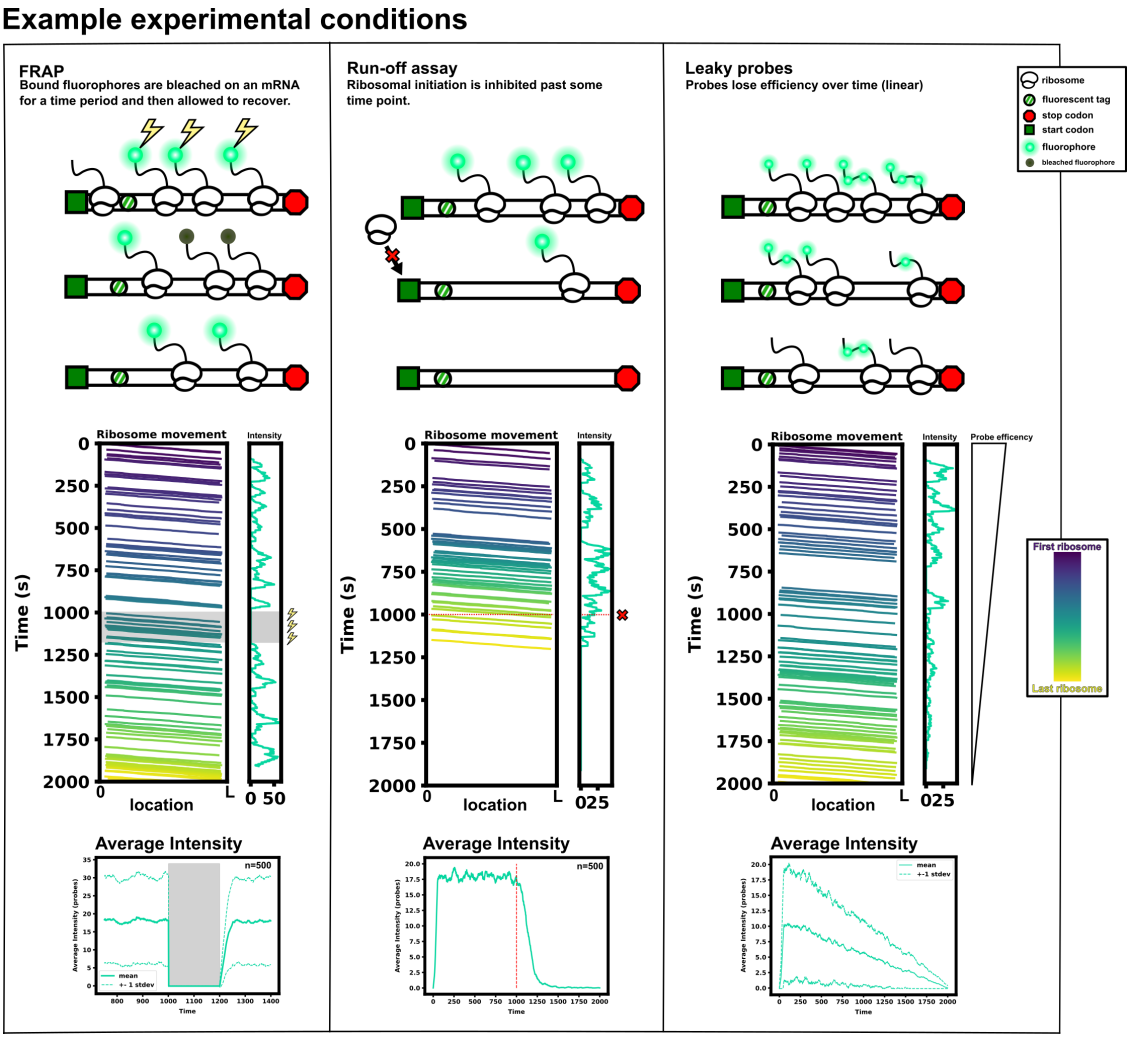


Figure 2.20: Caption on next page.

Figure 2.20: Example experimental perturbation simulations. Since rSNAPsim's model maker allows for arbitrary propensities and reactions, multiple experimental conditions can also be simulated. Before the model maker, experiments would have to be added by hand to the rSNAPsim simulation code but now we can account for almost any experimental perturbation. Fig. 2.20 (left column) depicts Fluorescence Recovery After Photobleaching (FRAP). FRAP can be simulated by adding a time dependent reaction that deletes all fluorescent probes tracked during some time range. Fig. 2.20 (middle column) depicts Run-off assays are simulated by adding a time dependent boolean to the ribosomal initiation rate. In the example ROA shown here, a time dependent ramp was used to simulate harringtonine diffusion starting at $t = 1000$. Probe efficiency can also be provided as a function if desired, Fig. 2.20 (right column). In the example, probe binding is calculated using a descending ramp function that decreases from 1 to 0 linearly over 2000 seconds.

$$wn_{\text{initiation}} = k_{\text{initiation}} \cdot \max(\min(-.005 * t + 6, 1), 0) \tag{2.12}$$

Photobleaching or probe inefficiencies can also be simulated by changing the probe binding function to any function that ranges from 0 to 1. Fig. 2.20 right column shows an example simulation with a using a descending linear ramp of probe binding efficiency, such that at 1000 seconds there's a 50% chance of a given probe binding.

2.2.3 Some example analyses the rSNAPsim can collect from simulation or calculate from experimental data.

One key functionality is using the rSNAPsim to analyze model outputs or data inputs. rSNAPsim provides two modules to calculate various statistics from intensity trajectories, inta, or ribosomal movement, riba. Fig. 2.21A shows some example intensity analyses, autocovariances and distribution calculations for 100 simulations of the frameshifting model described above. Fig. 2.21B showcases some of the ribosomal statistics that can be calculated from simulated trajectories. Polysomal profiling, and ribosomal profiling can compared with experimental data, but kymographs, collision counting, and individual ribosomal dwell times that are recorded during simulations have no experimental equivalent as of yet and are only available from simulations.

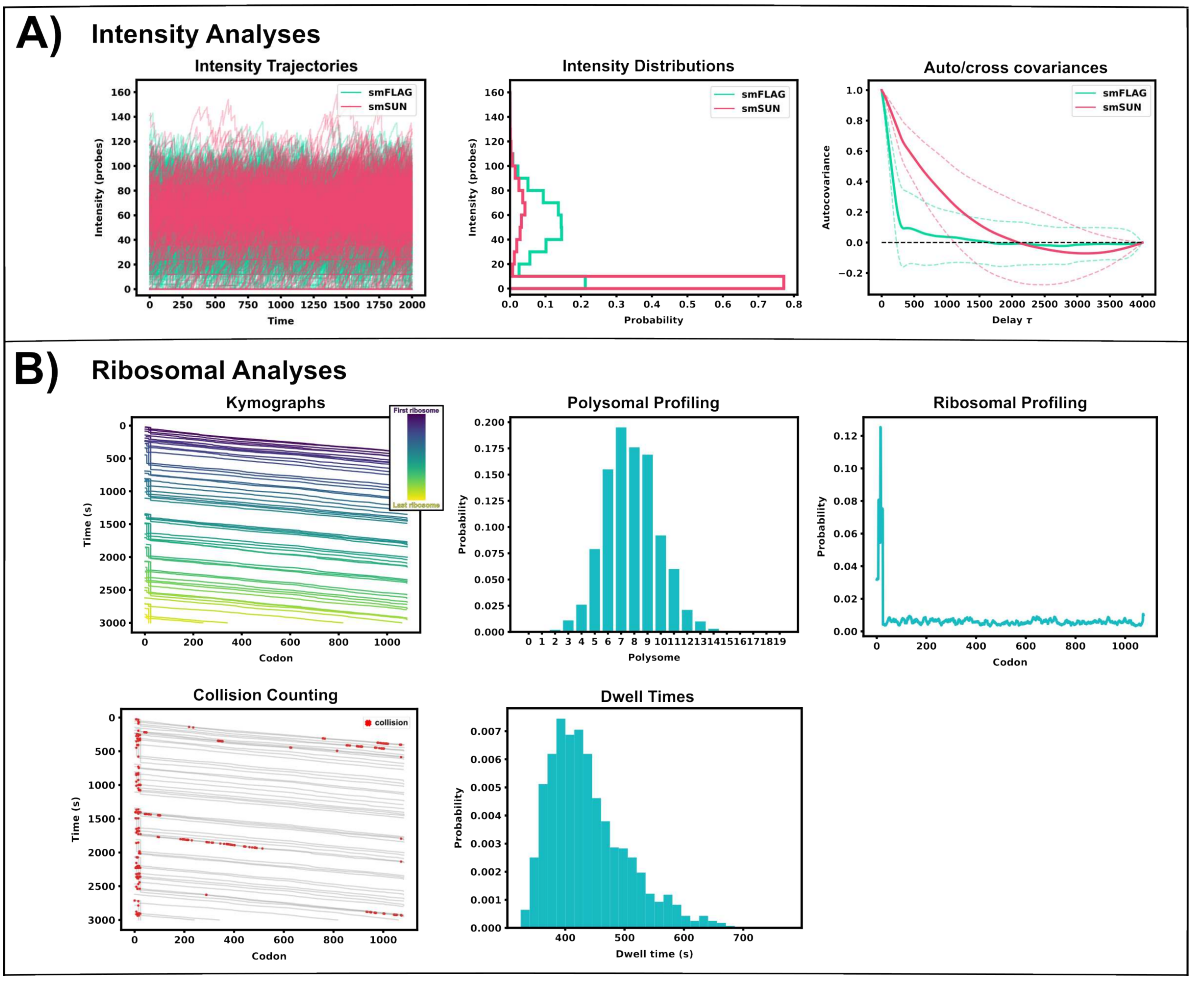


Figure 2.21: Example statistical analyses available in the rSNAPsim. A) Several intensity analyses available within the rSNAPsim. Left panel, two color intensity trajectories for the frameshifting model. Middle panel, distribution from 1000 intensity trajectories of the frameshifting model. Right panel, two autocovariances from each color in the frameshifting model from 1000 intensity trajectories. B) Ribosomal statistics available from the rSNAPsim. Left to right: Kymographs of ribosomal motion, example polysomal profile from 1000 simulations, ribosomal profiling for 1000 simulations, detected ribosomal collisions of a single simulation, and recorded dwell times per ribosome from 30 simulations.

2.2.4 rSNAPsim provides support for experiment design to improve model differentiation and hypothesis evaluation.

The final thing we will mention and showcase here is rSNAPsim's experiment design capabilities. rSNAPsim allows users to explore experiment designs computationally to find experiments that differentiate specific models at minimal laboratory material cost. Having access to a model

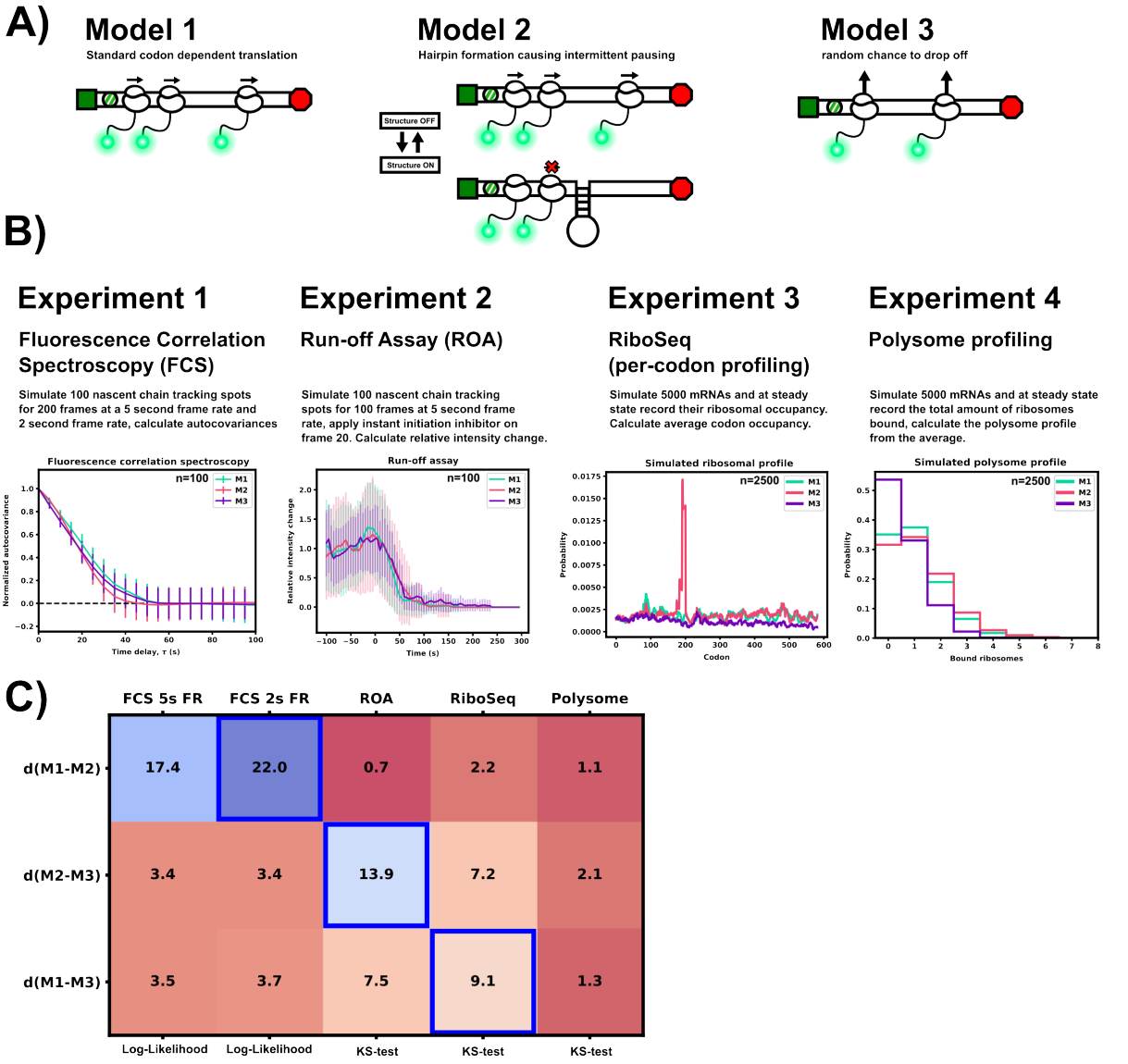


Figure 2.22: Simulated experiment design capabilities. rSNAPsim can be used to explore potential experiments for model discrimination with simulations. A) three proposed models to be explored: from left to right, codon dependent translation, intermittent pausing caused by a secondary structure, and random ribosome drop-off. B) Four proposed experiments: Fluorescence correlation spectroscopy of NCT trajectories at two different frame rates, Run-off assay with initiation inhibition of NCT trajectories, ribosomal profiling, and polysome profiling. C) normalized distance metric ratios, $\frac{d(M1-M2)}{2(d(M1-M1')+d(M2-M2'))}$, for each experiment and model pair elucidates the best experiment for differentiating each pair of models (highlighted in blue).

maker and computational tools allows modeling to leave its traditional post-facto application and be used prior to or in conjunction with preliminary experiments.

Fig. 2.22 shows a toy experiment design example using rSNAPsim to select the most informative experiments. Three different models are proposed: A codon dependent TASEP with no additional rates, a codon dependent TASEP with a transient secondary structure formation causing ribosomal pausing, and a codon dependent TASEP with a constant chance for ribosomes to drop off mid translation. For the structure formation model, the secondary structure cannot form if ribosomes are actively translating the 50 nucleotides of the structure. The experiment design goal is to select the most informative experiments for differentiating each model-model pair. This can be achieved by normalizing distance metrics comparing the model output distributions. Four different experiment types are proposed: FCS, Run-off Assay (harringtonine), ribosome profiling, and polysome profiling. Simulated FCS autocovariances were calculated using each model at steady state and collecting 200 trajectories of each model for 1000 seconds. The NCT trajectories were split into groups of 100 trajectories for calculating cross model log-likelihoods (M_{i} vs M_{j}) and log-likelihoods for 100 trajectories vs the other trajectories from the same model (M_{i} vs M_{i}'). Six log-likelihoods were calculated: Model1-Model2, Model1-Model3, Model2-Model3, Model1-Model1', Model2-Model2', Model3-Model3'.

$$LL_{\mathrm{M_i,M_j}} = \frac{1}{N_{\mathrm{pts}}} \sum_{\tau=1}^{20} \frac{(G_{\mathrm{M_i}}(\tau) - G_{\mathrm{M_j}}(\tau))^2}{\sqrt{SEM_{\mathrm{M_i}}}\sqrt{SEM_{\mathrm{M_j}}}} \tag{2.13}$$

$$LL_{\mathrm{M_i,M_i'}} = \frac{1}{N_{\mathrm{pts}}} \sum_{\tau=1}^{20} \frac{(G_{\mathrm{M_i}}(\tau) - G_{\mathrm{M_i'}}(\tau))^2}{\sqrt{SEM_{\mathrm{M_i}}}\sqrt{SEM_{\mathrm{M_i'}}}} \tag{2.14}$$

The log-likelihoods were used to calculate a balanced ratio of how informative the cross model log-likelihood was versus the self model log-likelihoods.

$$d(M_{\mathrm{i}}, M_{\mathrm{j}})_{\mathrm{NCT}} = \frac{LL_{\mathrm{M_i,M_j}}}{2(LL_{\mathrm{M_i,M_i'}} + LL_{\mathrm{M_j,M_j'}})} \tag{2.15}$$

For the run-off assays, the 6 cross and self log-likelihoods are calculated using the difference of normalized average intensities during the run-off period (til intensity reaches .05 on average) using 100 trajectories.

Table 2.2: Toy model parameters

	k_{e}	k_{on}	k_{off}	k_{dropoff}
Model 1	10	-	-	
Model 2	10	0.5 (non-excluded)	0.5	
Model 3	10	-	-	.03

$$LL_{\mathrm{M_i,M_j}} = \frac{1}{N_{\mathrm{pts}}} \sum_{\mathrm{t}=0}^{35} \frac{(\bar{I}_{\mathrm{M_i}}(t) - \bar{I}_{\mathrm{M_j}}(t))^2}{\sqrt{STD_{\mathrm{M_i}}}\sqrt{STD_{\mathrm{M_j}}}} \tag{2.16}$$

$$LL_{\mathrm{M_i,M'_i}} = \frac{1}{N_{\mathrm{pts}}} \sum_{\mathrm{t}=0}^{35} \frac{(\bar{I}_{\mathrm{M_i}}(t) - \bar{I}_{\mathrm{M'_i}}(t))^2}{\sqrt{STD_{\mathrm{M_i}}}\sqrt{STD_{\mathrm{M'_i}}}} \tag{2.17}$$

Just like the FCS distance metric ratio, the distance metric ratio for run-off assays is calculated as follows:

$$d(M_{\mathrm{i}}, M_{\mathrm{j}})_{\mathrm{ROA}} = \frac{LL_{\mathrm{M_i,M_j}}}{2(LL_{\mathrm{M_i,M'_i}} + LL_{\mathrm{M_j,M'_j}})} \tag{2.18}$$

For the ribosomal profile and polysome profile, distance metric ratios are calculated using KS-distances of the profile CDFs calculated from 2500 steady state simulations.

$$d(M_{\mathrm{i}}, M_{\mathrm{j}})_{\mathrm{polysome/profile}} = \frac{KS_{\mathrm{M_i,M_j}}}{2(KS_{\mathrm{M_i,M'_i}} + KS_{\mathrm{M_j,M'_j}})} \tag{2.19}$$

Fig. 2.22B shows representation data collected from each simulated experiment for each model. Fig. 2.22C depicts all normalized distance metric ratios for each experiment and model pair shaded by the ratio. With this quick, preliminary setup, we can inform an experimentalist of their next experiment for each model pair: To differentiate model 1 and model 2, a fluorescence correlation spectroscopy dataset should be collected from NCT, for telling model 2 and model 3 apart, a runoff assay gives the greatest distance between distributions, and for differentiating model 1 and model 3, a RiboSEQ experiment should be performed. More complicated experiment designs can be performed using rSNAPsim in conjunction with its sister package rSNAPed can simulate full NCT videos for model differentiation or dynamic exploration and classification.

The current version of rSNAPsim module is available, open source, at PyPI or our Github, https://github.com/MunskyGroup/rSNAPsim. The final part of my work here at CSU will be finalizing a descriptive paper and an updated version of the rSNAPsim expected to be released in summer 2024. The rSNAPsim formed the basis of the simulation pipeline for the paper presented in the next Chapter.

2.3 Using mechanistic models and machine learning to design single-color multiplexed nascent chain tracking experiments[3]

This paper was my primary project during my time at Colorado State University. The inception of this project came after the Aguilera 2019 paper. Now that we had a validated mechanistic model for the NCT data coming out of the Stasevich lab, we were intensely curious if we could use the model to elucidate dynamic differences across two mRNAs at the same time in the same cell. Our previous data was one mRNA per cell, we were curious if our model was still valid when both mRNAs were in the same environment – or if they had different codon usages. To tag both mRNAs in one color proved to be a very complex problem as to keep the mRNAs as similar as possible, both would need the same tag regions making them indistinguishable in the microscope. To fix this we came up with the classification from simulated NCT experiments in the following paper, providing a very nice computational simulation of any NCT experiment for our lab.

2.3.1 Summary

mRNA translation is the ubiquitous cellular process of reading messenger-RNA strands into functional proteins. Over the past decade, large strides in microscopy techniques have allowed observation of mRNA translation at a single-molecule resolution for self-consistent time-series measurements in live cells. Dubbed Nascent chain tracking (NCT), these methods have explored many temporal dynamics in mRNA translation not captured by other experimental methods such as ribosomal profiling, smFISH, pSILAC, BONCAT, or FUNCAT-PLA. However, NCT is currently restricted to the observation of one or two mRNA species at a time due to limits in the number of resolvable fluorescent tags. In this work, we propose a hybrid computational pipeline, where detailed mechanistic simulations produce realistic NCT videos, and machine learning is used to assess potential experimental designs for their ability to resolve multiple mRNA species using a single fluorescent color for all species. Through simulation, we show that with careful applica-

[3]William S. Raymond, Sadaf Ghaffari, Luis U. Aguilera, Eric Ron, Tatsuya Morisaki, Zachary R. Fox, Michael May, Timothy J. Stasevich, Brian Munsky DOI: https://doi.org/10.3389/fcell.2023.1151318

tion, this hybrid design strategy could in principle be used to extend the number of mRNA species that could be watched simultaneously within the same cell. We present a simulated example NCT experiment with seven different mRNA species within the same simulated cell and use our ML labeling to identify these spots with 90% accuracy using only two distinct fluorescent tags. The proposed extension to the NCT color palette should allow experimentalists to access a plethora of new experimental design possibilities, especially for cell signalling applications requiring simultaneous study of multiple mRNAs.

2.3.2 Introduction

mRNA translation is the process of reading messenger RNA strands to create functional proteins and is a crucial underpinning of known cellular life. With such a vital role, mRNA translation has been the subject of intense study over the past decades [84, 85, 86, 87]. Despite the focus, the effects that cellular signals have on the translation of individual mRNA molecules remains elusive due to two key factors: the staggering amount of dynamics, mechanisms, and modifications affecting the in vivo mRNA transcriptome heterogeneity and the limitations of experimental techniques that can accurately and informatively probe these dynamics molecule-by-molecule and in a time-resolved manner. Previous methods used to probe mRNA translation dynamics such as ribosomal footprinting [7, 8], RNA-seq [9, 10, 11], proteomics/protein abundances [12, 13], and smFISH [22, 88] provide snapshot bulk data in high quantity, at the detriment of obscuring the temporal dynamics of individual mRNA molecules. Pulsed SILAC, PUNch-P, BONCAT/QuaNCAT allowed mass spectrometry quantification of recently translated protein abundances via treatment with noncanonical amino acids [14, 15]. Methods such as FUNCAT/SUnSET bridged the gap of detecting active mRNA translation as well as subcellular location by labeling nascent peptide chains and imaging in fixed cells [16, 17]. FUNCAT-PLA and Puro-PLA then provided spatial resolution and imaging of the recent protein production via detection with a proximity ligation assay (PLA) [89]. Despite their innovations, these techniques require fixation or lysis of the cells of

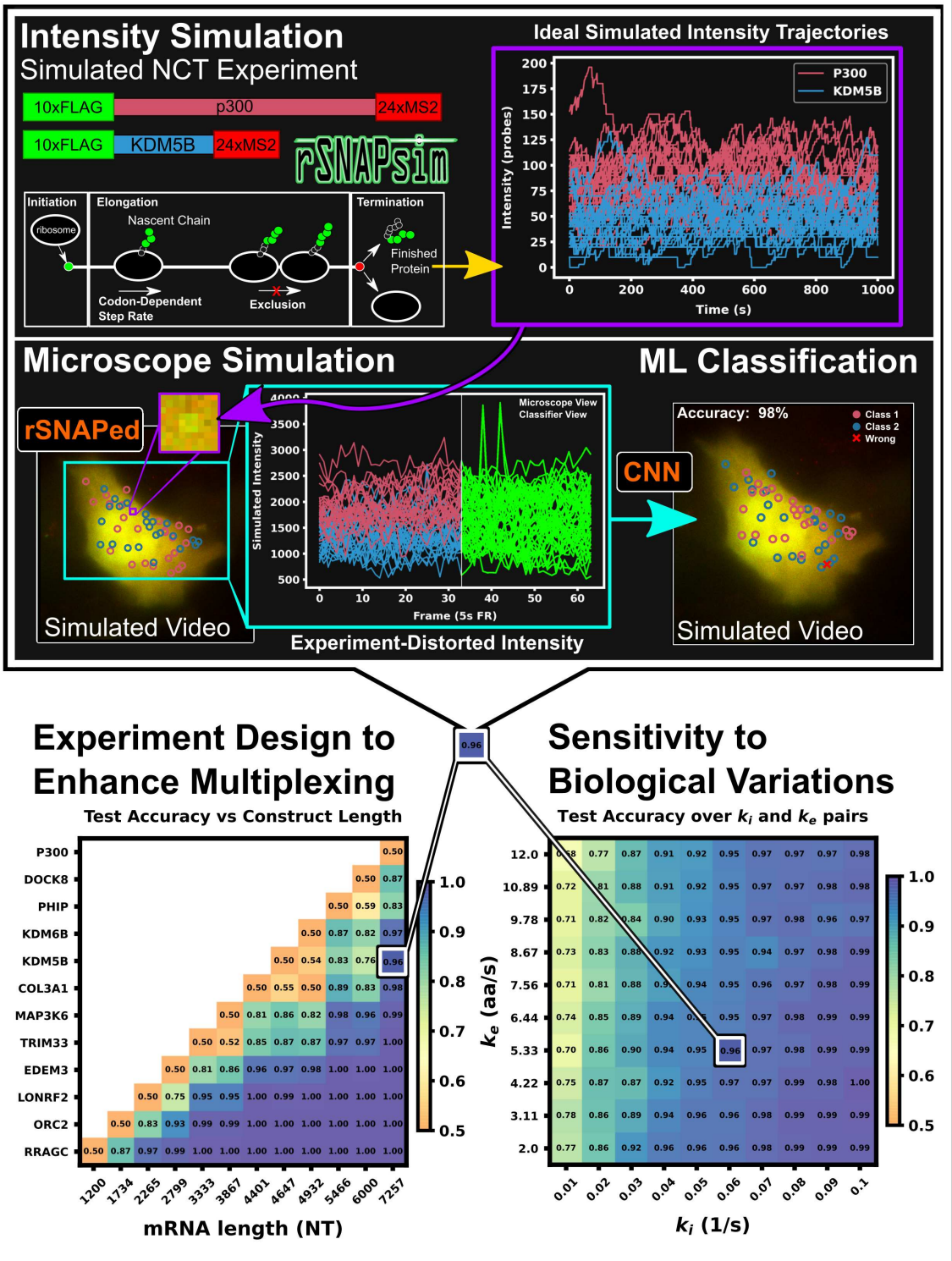


Figure 2.23: Nascent chain tracking multiplexing project graphical description.

interests. As a consequence, none of these methods are able to capture long-term temporal imaging of translation at a single-molecule level of the same mRNA molecules [90].

Since 2016, Nascent Chain Tracking (NCT) has provided experimentalists with a technique to study single, actively translating mRNA transcripts and quantify their dynamics with the use of a dual fluorescent labeling system [21, 22, 23, 24, 25, 26]. The key to this technique is multiple epitope sequences that are placed on a tag within the coding region of the studied mRNA; After this tag region is translated, fluorophore-conjugated intrabodies bind to the growing nascent polypeptide chain resulting in an amplified fluorescent spot. After translation is completed, fluorescently-tagged single proteins are free to diffuse away. The mRNA molecule is tagged in the 3' untranslated region with a hairpin loop repeat system recognized by fluorophore-conjugated MS2 or PP7 coat proteins, conferring a constant intensity in a separate color channel for tracking purposes. The combination of these two elements gives a co-localized, diffraction-limited, two-color spot trajectory denoting an mRNA location and a live nascent chain activity readout. NCT has been utilized to investigate many processes of interest such as mRNA frameshifting [28], mRNA IRES-mediated translation [35], mRNA decay [33, 34], translation suppression during cellular stress [32], and mRNA spot to spot heterogeneity [29]. NCT has also proven beneficial to extract important biophysical parameters, including elongation rates, initiation rates and ribosomal densities [23, 24, 25, 26], and microRNA mediated decay [91, 38].

Notwithstanding NCT's current adoption and importance, application of NCT to understand how different cellular signals affect translation of different mRNA is currently prevented by strong limits on the fluorophore color palette. Currently, only three or four resolvable colors exist in most microscope settings; one color is dedicated to tracking the mRNA, leaving only two or three resolvable colors to design experimental setups and constructs. Use of too many fluorescent probes with similar emission spectra leads to imaging issues such as light bleed through into each channel. Additional laser wavelengths also limits the frame rates that one can utilize due to the time required to switch laser or filters between each frame. While this has not proven detrimental to previous NCT experiments that have explored one [23, 25, 24] or two [28, 29, 35] translation products at

a time, there is an anticipation that this limitation will become a roadblock for designing future, more elaborate experiments, particularly for processes involving differential control of multiple mRNAs under cellular stimuli. Specifically, we speculate that the current form of NCT technology has already reached its peak in experiments where two different genes are correctly detected and differentiated on the same cell [28, 35, 29].

Recently, machine learning has enjoyed an explosion of applications in biological and biomedical imaging contexts [92, 93]. Convolutional neural networks have achieved the state-of-the-art performance on a wide variety of classification tasks such as speech recognition [94], computer vision [95], natural language processing [96], and in myriad biomedical contexts [97, 98]. For the purpose of our work, we utilize 1D convolutional neural networks to classify signals from two mRNA species recaptured from a realistic noise model and realistic mRNA translation model. One dimensional convolutional neural networks (CNNs) are well equipped to handle 1D signals for classification and see frequent use for applications across the biomedical field such as ECG [99] and EEG signals [100].

Applying machine learning to NCT experiments cannot be done outright as of time of writing due to a lack of large and standardized data sets, and it is not clear exactly how much data, and under which conditions, would be needed to build successful classifiers. To more efficiently explore these questions, we generate large sets of realistic NCT experiment simulations using a new pipeline, as detailed below. In brief, translation of nascent proteins and their corresponding Fluorescent intensity signals are modeled with a codon-dependent Totally Asymmetric Simple Exclusion Process (TASEP) to consider elongation rate changes due to codon selection along each mRNA transcript, as well as ribosomal collisions. In this paper, we use our previously described mRNA translation mechanistic model —a full comprehensive explanation of this model and its parametrization can be found in [50]. Fluorescent intensity signals from the mechanistic model are then combined with our realistic video rSNAPed pipeline, which applies a point spread function and adds a simulated cell background with microscope noise calibrated from real videos. mRNA molecules freely diffuse with a set diffusion rate within the simulated cell mask to simulate Brow-

nian motion. The resulting intensity from the simulated videos are then processed as if they were real images to generate "simulated nascent chain tracking" data. Through this means, we construct a controllable "experiment" whose measurement is a corrupted fluorescent signal, where we can easily control various factors such as signal-to-noise ratio (SNR), spot size, diffusion rates and mechanisms, mRNA initiation and elongation rates, and imaging conditions (frame rate / frame interval and number of frames taken).

In this paper, we use these simulations to demonstrate that high classification accuracy is achievable in principle using a machine learning approach across a large range of biophysical and experimental parameters, even if two mRNAs are utilizing identical tagging approaches and fluorescent colors. To extend NCT color palette, our computational pipeline of mRNA translation modeling, spot simulation, spot tracking, and machine learning uses as features various statistics of the spot's intensity fluctuations, such as their moments, relative intensity ratios, and decorrelation times to discriminate between different mRNA species. This type of "temporal multiplexing" could radically expand the number of mRNAs imaged in a cell. Different color fluorophores could be held in reserve for mRNAs whose characteristics are too similar to each other, or eschewed altogether to increase microscope imaging speed and decrease experiment cost. The entire pipeline can be used to explore potential NCT experimental designs without using valuable lab time and resources. We envision that the proposed model-based strategies to tag multiple mRNAs using single color tags will add new possibilities for future experimental NCT investigations.

2.3.3 Results and Discussion

To begin our exploration of the potential to use NCT signal intensity fluctuations to differentiate mRNA species, we choose baseline experiment using P300 (7257 NT) and KDM5B (4647 NT) mRNA, each with a 10xFLAG epitope tag. Both constructs are assumed to have equal initiation and elongation rates of k_{i} =0.06 1/s and k_{e} =5.33 aa/s, and both are assumed to be images for 64 frames with a rate of one frame every five seconds.

For typical mRNA and experiment designs, large training data sets may be needed to build an accurate classifier

To see how much training data are needed to build an accurate classifier, we use the baseline mRNA designs and experimental conditions (with 64 frames), and we trained our architecture with progressively increasing training data sizes. Fig. 2.34B shows the resulting accuracy on a withheld validation set of 1000 NCT spots, when the model is trained on independent training sets of the different sizes. The accuracy of the classifier levels off at about 87% when the training data set reaches 800-1000 NCT spots. Unfortunately, such a large amount of data is approaching unfeasible in current NCT laboratory settings, highlighting the need for using mechanistic simulation to supplement data when training a classifier. It is important to note that this result is for the experimental base conditions listed in Table 2.3; as we will discuss below, other parameter combinations can increase or decrease classification difficulty and result in variations for the amount of training information required.

Table 2.3: Selected base experimental conditions, statistics are calculated before microscope noise addition via rSNAPed.

	Gene 1 - KDM5B	**Gene 2 - P300**	**Imaging conditions**
Parameters	L: 4647 NT	L: 7257 NT	64 frames
	L_{tag}: 1011 NT	L_{tag}: 1011 NT	5s frame interval
	k_{e}: 5.333 aa/s	k_{e}: 5.333 aa/s	KDM5B SNR: 6.2
	k_{i}: 0.06 1/s	k_{i}: 0.06 1/s	P300 SNR: 8.9
Statistics	μ_{I}: 19.3 UMP	μ_{I}: 28.2 UMP	
	$\mathbf{\Sigma}_{\mathrm{I}}^2$: 18.7 UMP	$\mathbf{\Sigma}_{\mathrm{I}}^2$: 28.5 UMP	
	t_{dwell}: 354 s	t_{dwell}: 517 s	

* NT = nucleotides, * aa = amino acids

Realistic simulations of nascent chain tracking for KDM5b and P300 mRNA reveal that two mRNA can be distinguished using only their fluorescence intensity fluctuations.

As a proof of concept and to further explain the process of labeling spots by their behaviors rather than by their colors, Fig. 2.24 presents a example of our machine learning pipeline. Two simulated cells (Fig. 2.24A, top) are generated using the baseline conditions but with double the

amount frames (128 instead of 64, but with the same 5s interval between frames) to better highlight the ribosomal dwell-time difference between the two mRNAs. The two cells can be processed with a disk and doughnut approach (see Methods) to extract fluorescence intensity trajectories for each of the 100 spots (Fig. 2.24A, bottom).

In practice, biophysical parameters could be estimated from these NCT measurements of fluorescence intensity trajectories (i.e., from the intensity distributions and autocorrelations) as done previously in [50] and illustrated in Fig. 2.24A. For simplicity of description, we assume that these parameters are known, although we will relax this assumption later in the investigation. Using these assumed parameters, we use our computational pipeline to simulate a training data set of 5000 NCT spots from simulated NCT experiments, and we use this simulated data to train a classifier (Fig. 2.24C). With this classifier, the user can then finally label their original data artificially or use the trained classifier to label any newly collected data (Fig. 2.24D).

Simulations can reveal which aspects of experimental data are most informative for multiplexed mRNA classification

For two different NCT spots within the same cell and under the same experimental conditions, our dual input architecture (Fig. 2.34A and Methods) utilizes both relative intensity differences and signal frequency content to classify spots. Using both features allows for improved classification robustness across parameter space. There are experimental conditions and biophysical parameters where intensity information is more useful, conditions where frequency is more informative, and conditions where a mixture of both information sources is utilized by our architecture. To highlight this, we simulated three separate data sets: one with markedly different intensity distributions, one with mostly overlapping distributions, and one with nearly identical intensity distributions, Table 2.4. For each of of these conditions (which have different translation initiation rates), both P300 and KDM5b use the same average elongation rate, but because the genes have different lengths and different codon usages, they exhibit different ribosomal dwell time.

We applied our architecture to each of these data sets across a large swath of imaging conditions (i.e., different numbers of frames and frame intervals) to show our architecture's ability to self

select which features to use for classification as well as highlight which imaging conditions are ideal for these experiments conditions. Fig. 2.25A (left) shows that the first condition (different

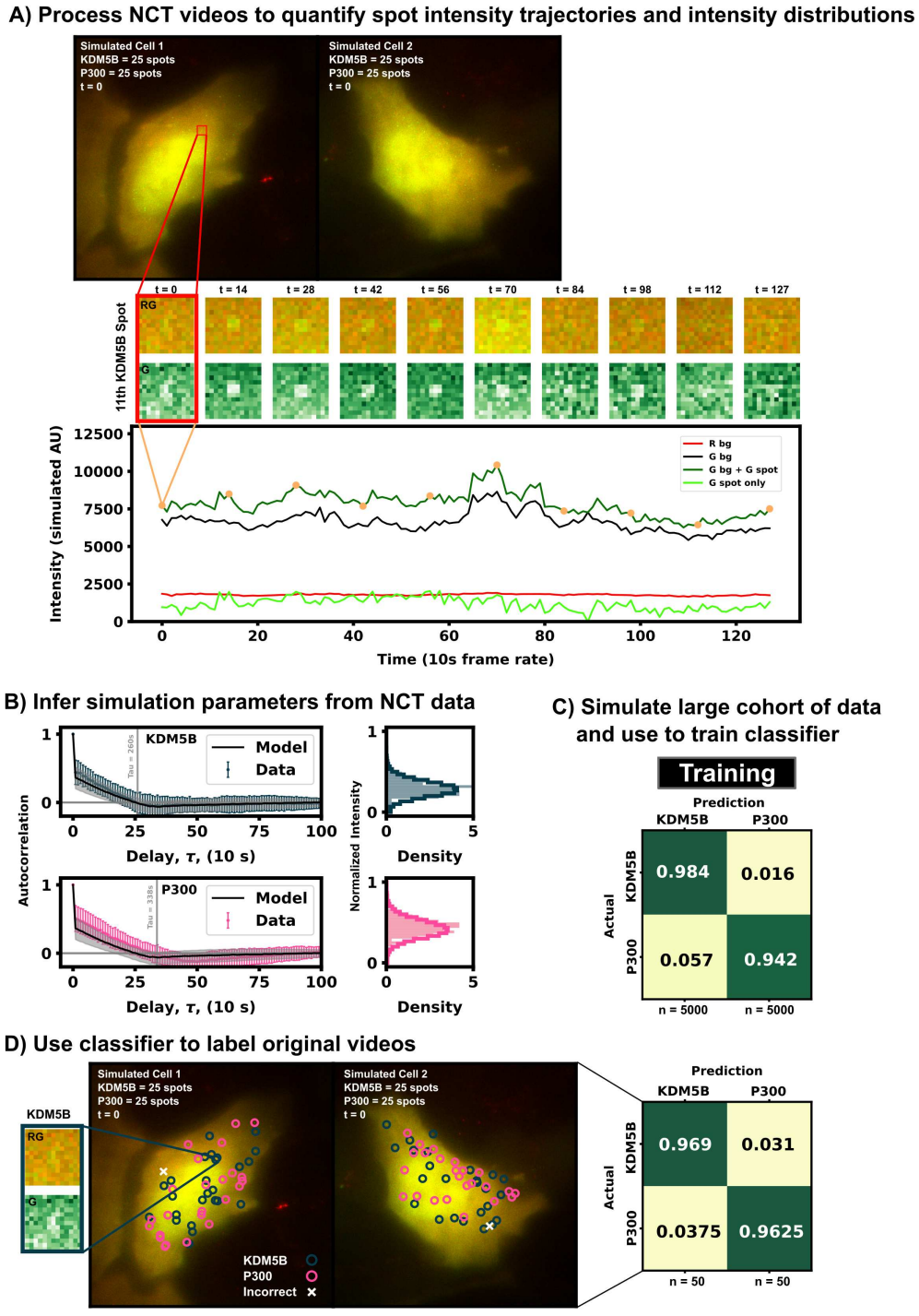


Figure 2.24: Caption on next page.

Figure 2.24: Example of labeling identically-tagged mRNAs in a simulated NCT experiment Example for labeling identically-tagged mRNAs in a simulated NCT experiment [A, top] Two simulated cells with 25 KDM5B and 25 P300 spots translating at identical biophysical parameters [A, bottom] Spot 11 in Cell 1 is highlighted in the intensity trace, showing the red background, green background, and extracted spot intensity via the disk and doughnut method B) Model parameters (k_{i} and k_{e}) are inferred by fitting auto-correlation functions and intensity distributions C) A classifier is trained with a large cohort of simulated data generated with the inferred parameters D) This classifier can then be used to label the original data or any subsequent data taken

Table 2.4: Long imaging time simulated data sets

Comparison Analysis	Variable Range	Constants
Imaging Dataset - Identical I_{μ} $I_{\mu,\mathrm{both}} \sim$ **4.7 UMP**	$k_{\mathrm{i,KDM5B}}$: 0.014139 s^{-1} $k_{\mathrm{i,P300}}$: 0.009676 s^{-1}	* Gene 1: P300 (7257 NT) * Gene 2: KDM5B (4647 NT) * Frame interval: 1 every 1 sec * Frames: 24000 * k_{e}: 5.33 $aa \cdot s^{-1}$ * D: 0.21 pixels$^2/s$ * SNR: 3
Imaging Dataset - Similar I_{μ} $I_{\mu,\mathrm{KDM5B}} \sim$ **6.0 UMP** $I_{\mu,\mathrm{P300}} \sim$ **4.7 UMP**	$k_{\mathrm{i,KDM5B}}$: 0.01860 s^{-1} $k_{\mathrm{i,P300}}$: 0.009676 s^{-1}	* Gene 1: P300 (7257 NT) * Gene 2: KDM5B (4647 NT) * Frame interval: 1 every 1 sec * Frames: 24000 * k_{e}: 5.33 $aa \cdot s^{-1}$ * D: 0.21 pixels$^2/s$ * SNR: KDM5B (3.9) & P300 (3)
Imaging Dataset - Different I_{μ} $I_{\mu,\mathrm{KDM5B}} \sim$ **13.7 UMP** $I_{\mu,\mathrm{P300}} \sim$ **4.7 UMP**	$k_{\mathrm{i,KDM5B}}$: 0.04242 s^{-1} $k_{\mathrm{i,P300}}$: 0.009676 s^{-1}	* Gene 1: P300 (7257 NT) * Gene 2: KDM5B (4647 NT) * Frame interval: 1 every 1 sec * Frames: 12000 * k_{e}: 5.33 $aa \cdot s^{-1}$ * D: 0.21 pixels$^2/s$ * SNR: KDM5B (8.9) & P300 (3)

dwell times and different intensity distributions) is trivial to classify —Simply looking at which spots are brightest and which are dimmest is sufficient to reach than 90% accuracy in just a few frames. Fig. 2.25A left) also shows that for mRNAs with such different intensity distributions, fluctuation frequencies are less informative, and the optimal frame interval should have as long a delay as possible so that each measurement is as statistically independent as possible.

Conversely, identical intensity conditions can be obtained by tuning the initiation rates such that each mRNA has the same ribosomal occupation average over time, and thus almost identical intensities (there is a slight difference due to codon dependence and ribosomal occupation probabilities per codon leading to varied occupation downstream of the tag region across each mRNA's length). In this condition, our classifier can only learn on the autocorrelation function and frequency information. Expected ribosome dwell times for P300 and KDM5B under these conditions are approximately 517 seconds and 353 seconds, respectively —a statistic that is available to the classifier through the autocorrelation of its intensity fluctuation. Fig. 2.25A (right) shows the test set classification accuracy as a function of video frame intervals and total number of frames. This plot highlights a clear region of image settings that would be sufficient to capture enough frequency information for classification, ≈20-30 seconds between frames for over 150 total frames. To be effective, the frequency-based classifier needs enough frames at the right intervals to sample the auto-correlation of ribosomal movements. If the total video time is too short, too few ribosomes will complete translation, and the NCT intensity signal would remain almost fully correlated with itself. As a result, one would have insufficient number of independent data points with which to calculate an effective autocorrelation. Conversely, if one observes the process too slowly (i.e., approaching or exceeding ribosomal dwell times), each frame would sample an independent set of ribosomes from the previous frame and any observed correlations would arise only from artifacts of imaging noise.

Fig. 2.25A (middle) shows an experimental condition where intensity distributions are similar enough such that both frequency and intensity information provide useful information for classification. The ideal imaging still uses an intermediate frame interval needed to capture the frequency differences, but classification is bolstered by the intensity information across the whole parameter space, with 10 frames at any frame interval being sufficient to provide 60% accuracy.

To further probe how intensity distributions and frequency information each contribute to ML classification, Fig. 2.25B shows each half of the architecture applied separately to 100 frames at varying frame intervals and varying amounts of frames at a 30 second frame interval (Highlighted

row and column of Fig. 2.25A, middle panel). Both individual halves Intensity distribution, (I), and Frequency, (F), have roughly a similar accuracy until frame interval grows too large to obtain

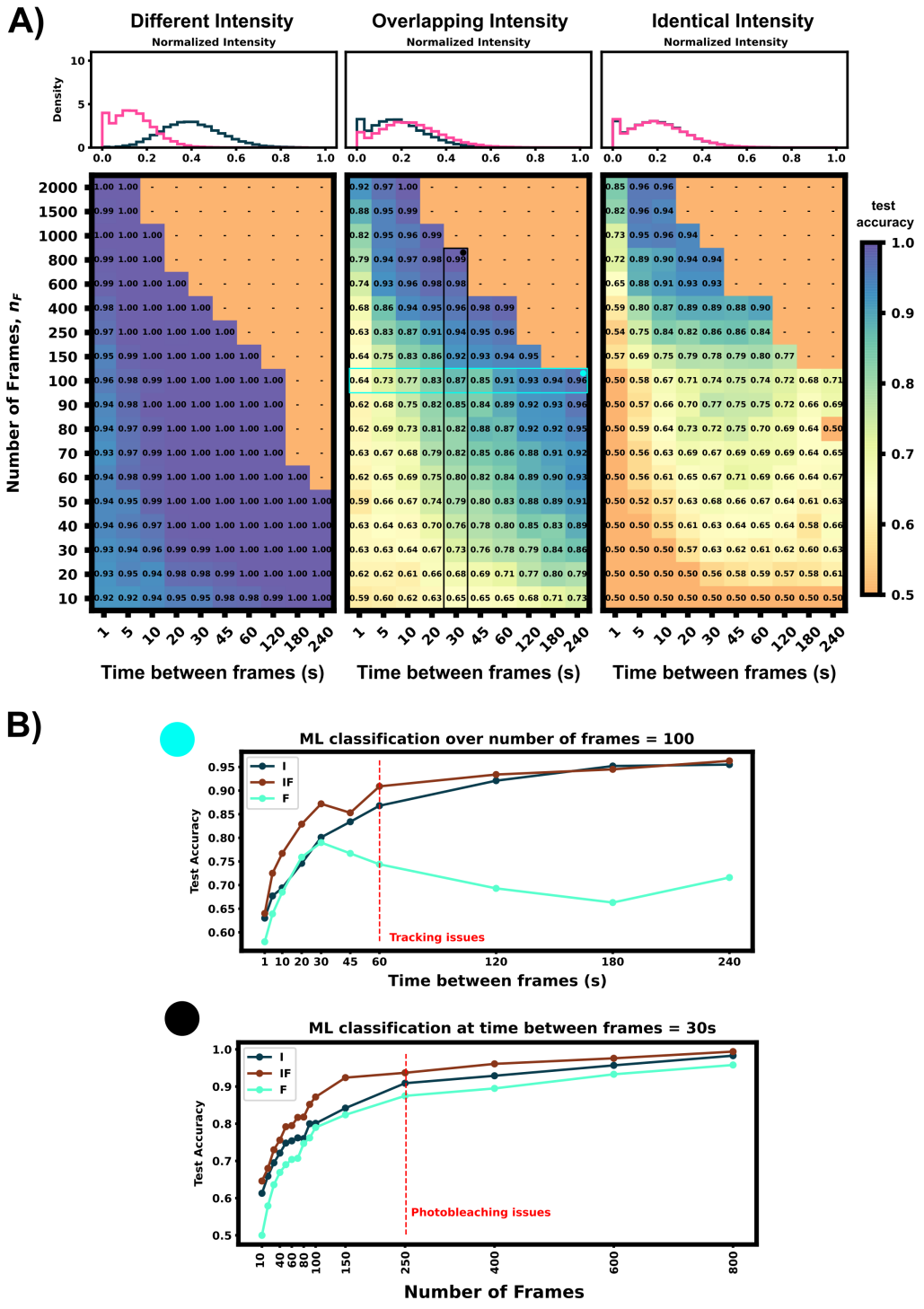


Figure 2.25: Caption on next page.

Figure 2.25: Classification accuracy *versus* imaging conditions and differences in mRNA intensity means. A) Accuracy *versus* frame interval and number of frames (Left) For constructs with substantially different intensities, the classifier requires only a few frames for a high classification accuracy (Middle) Overlapping, but non-identical, intensity conditions leverage both frequency and intensity information for classification (Right) Identical intensity conditions can only classify using frequency information, which requires an ideal frame interval B) Accuracy of ML using Intensity only (I), Frequency only (F), and both (IF) *versus* frame interval (top) or number of frames (bottom). Plots for (IF) correspond to the vertical and horizontal regions highlighted in Panel A, middle.

a good sampling of the autocorrelation function (60s). When both features are available, (IF), the classifier has a marked improvement in accuracy compared to the individual halves. It is important to note that we selected these specific experimental conditions such that there is partial information in both the autocorrelation and the intensity moments. If the experiment is designed such that either frequency or intensity is more informative, the proposed ML architecture adapts to rely more on that type of information and including the other will have marginal or no effect on validation accuracy (e.g., Fig. 2.25A left or right panel). Full classification heat-maps using architecture subsections are shown in Figure 2.26.

Although classification requires collection of a sufficient video length and at high enough temporal resolution, other experimental considerations are missing in this first analysis but must be taken into account to constrain imaging conditions in a real laboratory setting. Taking too many frames or a too short a frame interval may increase photobleaching effects that will require recalibration or computational correction or may necessitate the use of a lower laser power, which will reduce signal strength. Conversely, choosing too low a frame interval (i.e., longer delays between images) could lead to spot tracking issues if particles diffuse too fast in comparison to the frame rate or if there is too high a density of overlapping particles. This trade-off between too fast and too slow a frame interval can be partially ameliorated by tracking only on the RNA tag channel with a higher frame interval and imaging in the protein channel with a slower rate, but this solution requires more complex steps for image collection and processing.

Simulated data is ideal for testing different experimental and biological conditions to probe the possibilities and limitations of NCT multiplexing

Now that we have a computational pipeline to generate simulated data and a flexible classifier to train using both intensity and frequency information, we can explore multiple parameter spaces to guide experimental design toward conditions that are more conducive for accurate classification. Additionally, by generating data over a large parameter swath, we can examine multiple experiment

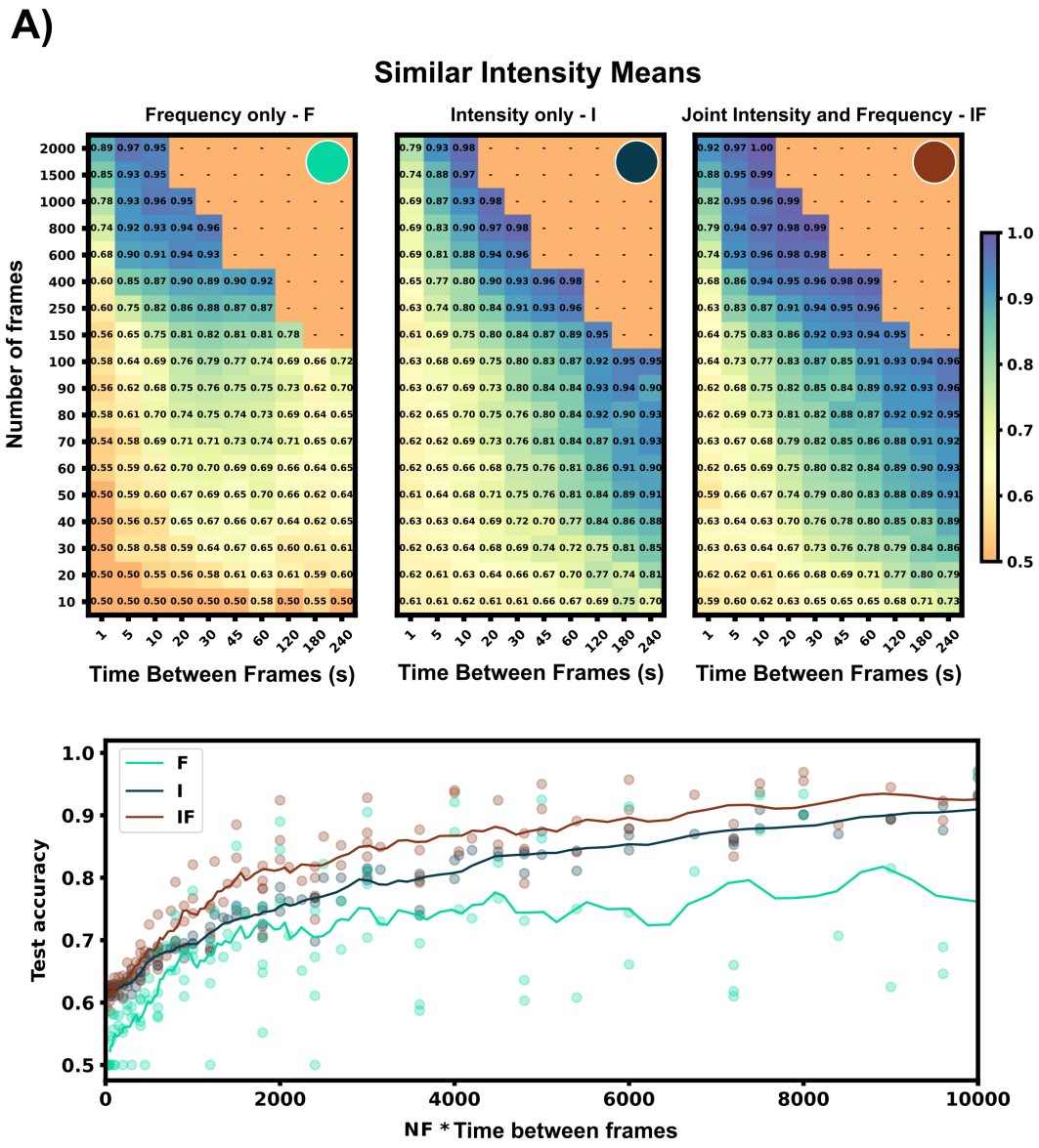


Figure 2.26: Caption on next page.

Figure 2.26: Classification accuracy for both architecture halves and the full architecture. A) Accuracy *versus* (Top) Accuracy of separate architecture halves (I,F) and joint architecture (IF) across a large range of frame rates and number of frames. Heat maps show each architecture half or the joint architecture applied on an NCT data set with different frequency content (decorrelation times τ = 354s vs τ = 517s and similar intensity means (4.7 UMP vs 6 UMP). (Bottom) Using both architecture halves increases average test accuracy across all conditions vs the individual feature architectures. All test accuracies from the heat maps above are plotted vs their video length (NF * Time between frames) for the full architecture (IF) and each half architecture (I, F). A trend-line was generated with a moving average of 100 seconds.

Table 2.5: Dynamics affected by selected variables to investigate with the NCT ML pipeline

Variable	**Dynamic impacted when increased**	**Controllable**
Frame interval (FI)	↑ information / resolution	Experimentally controllable
Number of Frames (n_{F})	↑ information / resolution	
mRNA Length (L_{mRNA})	↑ ribosomal dwell time ↑ fluorescent intensity	Semi-controllable (Gene/tag selection)
Initiation Rate (k_{i})	↑ fluorescent intensity	Semi-controllable (UTR selection)
Elongation Rate (k_{e})	↓ ribosomal dwell time ↓ fluorescent intensity	Semi-controllable (Codon selection)

setups that are unresolvable to obtain insight for what mitigation strategies an experimentalist could take to improve classification results.

In this study, we limit the scope of exploration to five key variables (Table 2.5): mRNA length (L_{mRNA}), frame interval (time in seconds between frames, FI), number of frames used (n_{F}), ribosomal initiation rate (k_{i}) and ribosomal elongation rate (k_{e}). We choose these specific parameters because they are the most experimentally relevant as they influence how long of a video to take and what types of mRNA constructs to design. Additionally, these five parameters directly influence one or both of the two statistics we are using as learning features, specifically the decorrelation times and intensity levels. The selected parameters to explore and a list of dynamics affected by each of these parameters are provided in Table 2.5. For example, an increase in mRNA length has a corresponding increase in ribosomal dwell time and therefore fluorescent intensity of a mRNA spot.

Although, for the sake of brevity, we focus on the five parameters presented in Table 2.5, we note that our proposed simulation and classification pipeline is general and can be used to

explore many other mechanisms of the translation or the imaging system. Other parameters of potential interest may improve or worsen classification accuracy. These include effects such as non-equilibrium dynamics (all mRNA simulations in this paper are at steady state), differing diffusion dynamics (spatial considerations from cell morphology or dependence on mRNA translation state). To focus our analysis on our current parameters of interest, all data in the main text are generated without photobleaching effects and are analyzed using the specific coordinates of the simulated spots within the rSNAPed module; i.e., we are not relying on a spot detection and tracking algorithm.

A preliminary exploration of photobleaching and imperfect tracking is presented in the Supplemental Material Section 1.1. We find that classification accuracy for perfectly tracked KDM5B and P300 mRNA spots remains high even with substantial levels of photobleaching (see Fig. 2.6D), at least provided that all NCT probes bleach at the same rate. In this case, the relative differences in fluorescence intensities and fluctuations (at higher frequency than the bleaching rate) can still be used for classification. However, we note that the number of correctly classified spots per cell depends heavily on the accuracy and completeness of tracking. Using the trackpy tracking pipeline [101], we can only track approximately 40% of spots for long contiguous time courses (Fig. 2.6C), although classification accuracy for those spots is only slightly below that for perfect tracking (Fig. 2.6D). We note, however, that this preliminary analysis assumes a simple exponential decay model for photobleaching and only explores one option for particle tracking; a complete analysis would require in-depth examination of different photobleaching models [102, 103] and should explore additional approaches for track linking [104]. In the current analyses, all mRNA translation models are run until they reach steady state (burn in time: 1000 seconds) before they are used to generate NCT spots, but rSNAPed allows for simulation of non-stationary conditions or experiment perturbations (e.g., Harringtonine treatment to interrupt translation, or fluorescence recovery after photo-bleaching (FRAP) to examine ribosome replacement as studied in [24, 60, 26]. Finally, in the current analyses, all spot motion is simulated using normal Brownian motion with a constant diffusion rate of 0.925 μm^2/s (0.55 pixel^2/s) but settings for rSNAPed can easily be

changed to allow for anomalous or temporal or mRNA state-dependent diffusion rates. In addition, beyond using just intensity and frequency, one could potentially improve classification using additional features for machine learning such as x, y, and z spatial positions or spot velocities.

Fig. 2.27(top) summarizes the set up for four parameter sweeps designed to explore the effects of selected parameters on ML classification: (CT1) compares pairs of mRNA with many different lengths, (CT2) compares two mRNA with many different combinations of shared elongation and initiation rates, (CT3) compares two mRNA each with different elongation rates, and (CT4) compares two mRNA each with different initiation rates. For each comparison, mRNA translation simulations use parameters selected from a survey of literature reporting experimentally measured rates [23, 60, 24, 26, 25], and all parameters are shown in Table 2.6. For CT1, the mRNA length range selected (1200 - 7257 nt) covers 53% of the lengths in the human consensus coding sequences (current CCDS nucleotide release - 11.28.2021 [105]). A baseline NCT experiment was selected as 10xFLAG-p300 vs 10xFLAG-KDM5B with an initiation rate and elongation rate of 0.06 1/s and 5.33 aa/s at imaging conditions of 64 frames with a 5s frame interval. When parameters are held constant in the comparison analysis they are held to these values. The results of the various comparison tests are shown in Figures 2.27A-D and 2.28A-D, to be discussed individually below, and all full data sets are available upon request. Data sets can also be resimulated from the provided Github repository https://github.com/MunskyGroup/Multiplexing_project, DOI: https://zenodo.org/record/7884701.

mRNAs with sufficiently different lengths can be differentiated using their fluctuation intensity signals

To explore the effect of mRNA length differences on classification, we applied our architecture to NCT experiments where the only difference between mRNAs is their length (Figures 2.27A and 2.28A). In addition to the previous P300 and KDM5B constructs, ten new genes with approximately evenly spaced nucleotide length coding regions were selected from the human consensus coding sequence database, Table 2.6 row 1. A standard 1011 nucleotide 10x-FLAG tag was added to the N-terminus end of each before simulating the NCT experiments. We assume a common set

of global cell translation parameters, that is, every mRNA has a common ribosomal initiation and elongation rate since all mRNA would be in the same cell and they have been designed to have

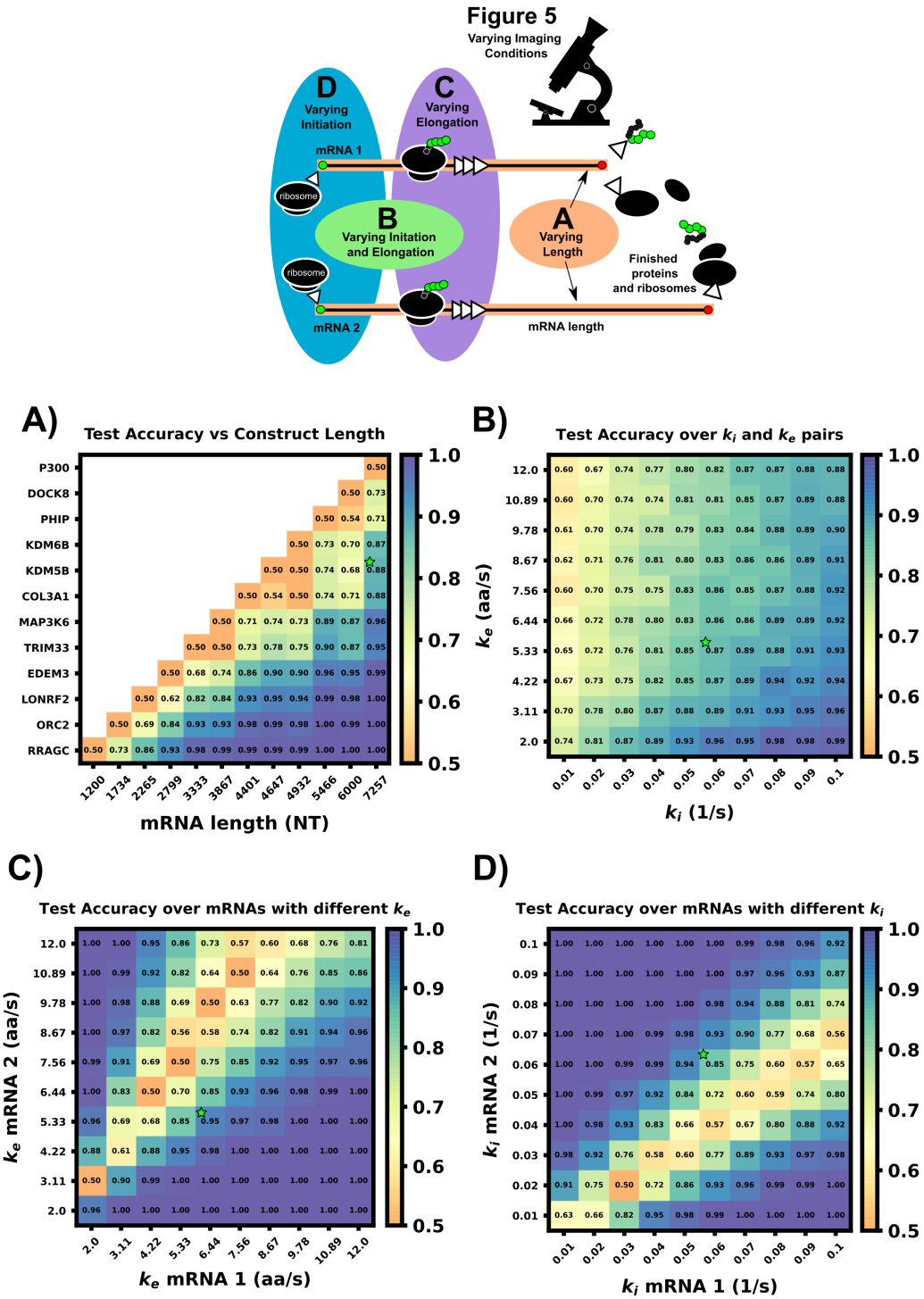


Figure 2.27: Caption on next page.

Figure 2.27: Comparison of ML test accuracy under variations in biophysical parameters. TOP: legend of which experimental parameters are changed for each panel A) Effect of construct length on classification test accuracy when trained on 4000 NCT spots and tested on 1,000 withheld spots. Imaging conditions, initiation, and elongation are held constant while mRNA lengths are swept from 1200 NT to 7257 NT using different mRNAs. All classifiers are trained on 4000 NCT spots and tested on 1000 NCT spots to get the test accuracy B) Classification accuracy for P300 and KDM5B versus shared initiation and elongation rates C) Classification of P300 and KDM5B with shared initiation rate (0.06 1/s) but with different varying elongation rates. D) Classification of P300 and KDM5B with shared elongation rates (5.33 aa/s) and varying initiation rates. The green star in each panel denotes the default P300/KDM5B experiment with 5 s frame interval, 64 frames, initiation rate of 0.06 1/s, and elongation rate of 5.33 aa/s.

identical UTRs. Specifically, for this parameter sweep, we assume that $k_{\mathrm{e}} = 5.33$ aa/s and $k_{\mathrm{i}} = 0.06$ ribosomes/s and that video is recorded for a moderate length of 64 frames at a rate of one frame every five seconds. For each NCT simulation, the longer of the two mRNA species will

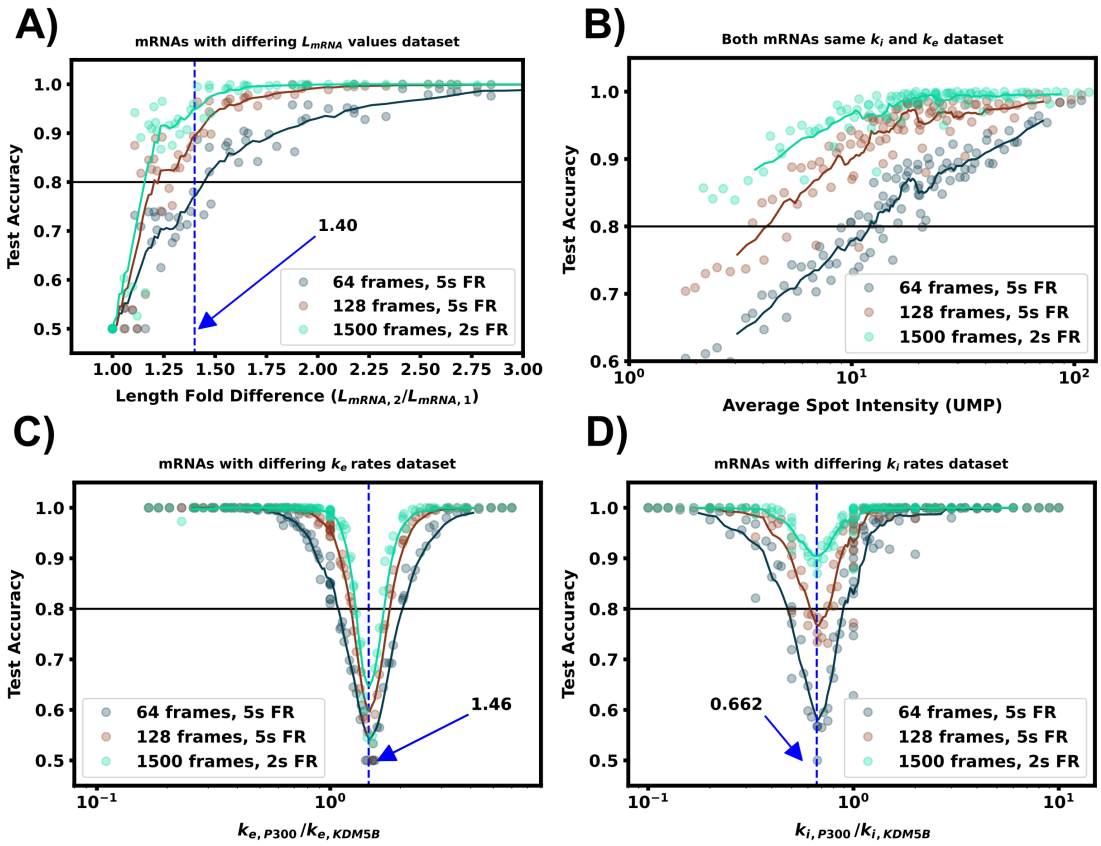


Figure 2.28: Caption on next page.

Figure 2.28: Increasing video length to resolve difficult to classify mRNA combinations. A) Classification accuracy versus mRNA length fold difference, assuming identical tag designs and parameters and videos with 64, 128, or 1,500 frames B) Classification accuracy for P300 and KDM5B with identical tags and parameters vs. average P300 intensity (proxy for signal-to-noise ratio). As SNR, video length, and resolution increase, there is a corresponding increase in classification accuracy C) Classification accuracy versus ratio of P300 and KDM5B elongation rates. As parameters approach the dotted line at $k_{\mathrm{e,P300}}/k_{\mathrm{e,KDM5B}}$ =1.46, the frequency and intensity information is identical between the two mRNAs, and increasing video length provides only marginal improvements D) Classification accuracy versus ratio of P300 and KDM5B initiation rates. As parameters approach the dotted line at $k_{\mathrm{i,P300}}/k_{\mathrm{i,KDM5B}}$ =0.648, the two mRNA attain similar intensity means, but classification can be achieved through frequency content and is improved substantially by collecting longer videos.

retain ribosomes for longer periods of time and will therefore exhibit slower decorrelation times and higher average intensities.

For convenience, the difference between mRNA lengths can be quantified by the fold change:

$$\Delta L_{\mathrm{fold}} = \frac{\max(L_{\mathrm{mRNA},1}, L_{\mathrm{mRNA},2})}{\min(L_{\mathrm{mRNA},1}, L_{\mathrm{mRNA},2})}. \tag{2.1}$$

As one should expect, Figures 2.27A and 2.28A show that as ΔL_{fold} becomes larger, the classification becomes easier for our machine learning architecture. For our simulated conditions and video length, we find that a 1.4-fold difference is sufficient to achieve greater than 80% classification accuracy of any two mRNA combinations of the 12 we selected. However, Fig. 2.28A shows that one can lower the required length fold change by increasing the number of frames given. For example, if one extends imaging to 128 frames at 5s resolution (double the frames of the original condition), then one could achieve an 80% classification for a smaller ΔL_{fold}=1.1. However, there is a diminishing benefit to adding extra video length; extending imaging to 1500 frames at 2s resolution (a barely achievable amount of data to collect with current NCT capabilities) provides only a marginal further improvement (compare teal and brown lines). This diminishing return emphasizes the need for careful consideration when designing NCT experiments with multiple mRNAs with identical tags, either to avoid designs that would require an unobtainable amount of sampling or to reduce sampling for designs that can be differentiated with fewer imaging resources.

Table 2.6: Comparison analyses parameters

Comparison Analysis	Variable Range	Constants
mRNA Length vs mRNA Length	mRNA length (without tag): 1200 NT - 7257 NT RRAGC - 1200 ORC2 - 1734 LONRF2 - 2265 EDEM3 - 2799 TRIM33 - 3333 MAP3K6 - 3867 COL3A1 - 4401 KDM5B - 4657 KDM6B - 4932 PHIP - 5466 DOCK8 - 6000 P300 - 7257	* Frame interval: 5s * Frames: 64 * k_{i}: 0.06 s^{-1} * k_{e}: 5.33 $aa \cdot s^{-1}$ D: 0.21 $pixels^2/s$ SNR: 3.7 (RRAGC) - 17.6 (P300)
$\mathbf{k_e}$ **(both mRNAs) vs** $\mathbf{k_i}$ **(both mRNAs)**	k_{i}: 0.01 - 0.1 s^{-1} k_{e}: 2 - 12 $aa \cdot s^{-1}$	* Gene 1: P300 (7257 NT) * Gene 2: KDM5B (4647 NT) * Frame interval: 5s * Frames: 64 D: 0.21 $pixels^2/s$ SNR 1- 36
$\mathbf{k_e}$ $mRNA_1$ **vs** $\mathbf{k_e}$ $mRNA_2$	k_{e}: 2 - 12 $aa \cdot s^{-1}$	* Gene 1: P300 (7257 NT) * Gene 2: KDM5B (4647 NT) * Frame interval: 5s * Frames: 64 * k_{i}: 0.06 s^{-1} D: 0.21 $pixels^2/s$ SNR KDM5B: 2.6 - 14.2 SNR P300: 3.9 - 23
$\mathbf{k_i}$ $mRNA_1$ **vs** $\mathbf{k_i}$ $mRNA_2$	$k_{\mathrm{initation}}$: 0.01 - 0.1 s^{-1}	* Gene 1: P300 (7257 NT) * Gene 2: KDM5B (4647 NT) * Frame interval: 5s * Frames: 64 * k_{e}: 5.33 $aa \cdot s^{-1}$ D: 0.21 $pixels^2/s$ SNR KDM5B: 1.3 - 9.3 SNR P300: 1.7 - 13.3

mRNA with different lengths can be distinguished for a range of different combinations of their biophysical parameters.

In the next comparison, we sought to understand how classification accuracy depends upon the biophysical parameters that govern translation dynamics. Specifically, Figures 2.27B and 2.28B explore how the accuracy to classify P300 and KDM5B constructs would depend on their shared rates of translation initiation and elongation (k_{i} and k_{e}, respectively). Because the two constructs have fixed lengths in this comparison, every combination of $(k_{\mathrm{i}}, k_{\mathrm{e}})$ yields the same ratio of intensity mean and decorrelation time. Therefore, classification should be possible across the entire parameter range, provided that one obtains a sufficient level of video sampling. Fig. 2.28B and Supplemental Figure S4 (second row) show that indeed, once there is enough video resolution, all $(k_{\mathrm{e}}, k_{\mathrm{i}})$ pairs can be classified with higher than 90% accuracy. However, when constrained to a fixed number of frames and frame interval (e.g., 64 frames with 5s frame interval as considered in Fig. 2.27B), some combinations of parameters yield signals that are brighter and are therefore easier to classify using intensity statistics. Specifically, when the initiation rate is high and the elongation rate is low, more ribosomes enter per second and remain longer on the mRNA. Conversely, "sparse" loading conditions prove harder to classify due to low ribosomal occupancy and rare ribosomal entry and thus, a lower signal to noise ratio.

In natural constructs, different 3' and 5' UTR sequences will affect the availability of initiation factors and regulatory elements such as uORFs, and as such one should expect that translation initiation rates will vary from one mRNA species to the next [106, 107, 108, 109]. To explore how differences in initiation rate would affect classification accuracy, Figures 2.27C and 2.28C compare the classification accuracy as a function of both mRNAs' unique initiation rates. In this case, it is possible for both mRNA to have very different or nearly identical intensity means depending on how close the ratio of the initiation rates compares to the inverse ratio of their lengths. For our particular case of KDM5B and P300 mRNA, if we neglect ribosomal collisions, the ratio of mean intensities can be estimated as in [50]:

$$\frac{I_{\mathrm{P300}}}{I_{\mathrm{KDM5B}}} = \frac{\left(1 - \frac{L_{\mathrm{tag}}}{2L_{\mathrm{P300}}}\right) \cdot \tau_{\mathrm{P300}} \cdot k_{\mathrm{i,P300}}}{\left(1 - \frac{L_{\mathrm{tag}}}{2L_{\mathrm{KDM5B}}}\right) \cdot \tau_{\mathrm{KDM5B}} \cdot k_{\mathrm{i,KDM5B}}} \tag{2.2}$$

By setting this intensity ratio to unity, we can rearrange to find the corresponding ratio of initiation rates as

$$\frac{k_{\mathrm{i,P300}}}{k_{\mathrm{i,KDM5B}}} = \frac{\left(1 - \frac{L_{\mathrm{tag}}}{2L_{\mathrm{KDM5B}}}\right) \cdot \tau_{\mathrm{KDM5B}}}{\left(1 - \frac{L_{\mathrm{tag}}}{2L_{\mathrm{P300}}}\right) \cdot \tau_{\mathrm{P300}}} = \frac{\left(1 - \frac{1011nt}{2 \cdot 5658nt}\right) \cdot 353s}{\left(1 - \frac{1011nt}{2 \cdot 8268NT}\right) \cdot 517s} = 0.662, \tag{2.3}$$

where we calculated the expected elongation times for the two mRNA (τ_{KDM5B} and τ_{P300}) under an assumption of sparse loading for the ribosomes (i.e., no collisions). Figures 2.27C and 2.28C show that when the two mRNAs' initiation rates approach this critical ratio, the accuracy decreases substantially. However, although these mRNAs may have similar intensities, their different lengths still result in distinct dwell times, and Figures 2.28C and S4 show that frequency information can still provide for accurate classification, especially as one increases the amount of video.

Figures 2.27D and 2.28D explore the opposite circumstance, where the two mRNA have the same initiation rate, but with two different elongation rates. In this case, it would be possible for both the intensity means and the dwell times to be identical for the two constructs if their elongation rates satisfy the ratio:

$$\frac{k_{\mathrm{e,P300}}}{k_{\mathrm{e,KDM5B}}} = \frac{\tau_{\mathrm{P300}}/L_{\mathrm{P300}}}{\tau_{\mathrm{KDM5B}}/L_{\mathrm{KDM5B}}} = \frac{2756aa}{1886aa} \cdot 1 = 1.46129. \tag{2.4}$$

Figures 2.27D and 2.28D show that NCT signals along this parameter manifold are virtually indistinguishable as the only variation between their statistics is the time ribosomes spend in the tag region. Specifically, ribosomes on P300 reach full intensity fluorescence 1.46x faster than those on KDM5B. Increasing video resolution could potentially resolve parameter sets close to this manifold, but Fig. 2.28D shows that accurate discrimination would require an unrealistic amount of NCT video. In conditions like these, it may be best to tag the mRNAs with different tag colors or use different tagging strategies as we discuss in the next section.

mRNAs with similar fluctuations and intensities can be made more classifiable via intelligent design of tag placements

Considering the comparisons in Fig. 2.27A-D, we observed that there are several conditions under which machine learning may be unable to classify NCT spots. Specifically, classification may fail when: (1) videos are too short to sample intensity distributions or have a too poorly chosen frame interval to quantify intensity frequencies (Fig. 2.25); (2) when signal intensities are too low to capture the relevant statistics; or (3) when the mRNA lengths, initiation rates, or elongation rates combine such that both mRNAs yield identical statistics for both intensity and frequencies. As discussed above, the solution to the first two "failure modes" is to collect more information (i.e., longer videos with a more appropriate temporal resolution). In contrast, for conditions where intensity and frequency statistics are nearly identical, such as in Figures 2.27C and 2.28C, no amount of extra information or imaging will be able to tell these NCT spots apart. In such a circumstance, the similarity between the two mRNA could be ameliorated by changing the mRNA constructs themselves. For example, one could alter the fluorescent signal statistics either by lengthening one of the mRNAs in question or by employing an alternate tagging design. Changing the length of the mRNA with linker or junk regions may introduce unwanted effects in the mRNA / protein targets under study, so changing the tag region is preferable. To demonstrate this possibility, Fig. 2.29A and B considers the case from above where the elongation rates of P300 ($k_{\mathrm{e}}^{\mathrm{P300}}$=11.04 aa/s) and KDM5B ($k_{\mathrm{e}}^{\mathrm{KDM5B}}$= 7.56 aa/s) differ by the critical factor of 1.46, such that the 10xFLAG-P300 and 10xFLAG-KDM5B constructs yield identical intensity fluctuations that cannot be discriminated from one another. We then propose several modifications to the tagging scheme for KDM5B to explore how different designs might affect classification accuracy as follows:

- 10x Flag Tag on the N-terminus of KDM5B (original ineffective design)
- Splitting the tag region to relocate 3 epitopes to the C-terminus
- Relocating the tag region to the C-terminus of KDM5B's CDS

- Adding 5 epitopes to the end of the 10x Flag Tag (adding in 5 'DYKDDDDK' sequences separated by two glycines each)
- Removing 5 epitopes from the 10x Flag Tag (mutating the last 5 epitopes from 'DYKDDDDK' to 'DYKDGGDK')
- Relocating the 10x Flag Tag to the C-terminus of KDM5B's CDS

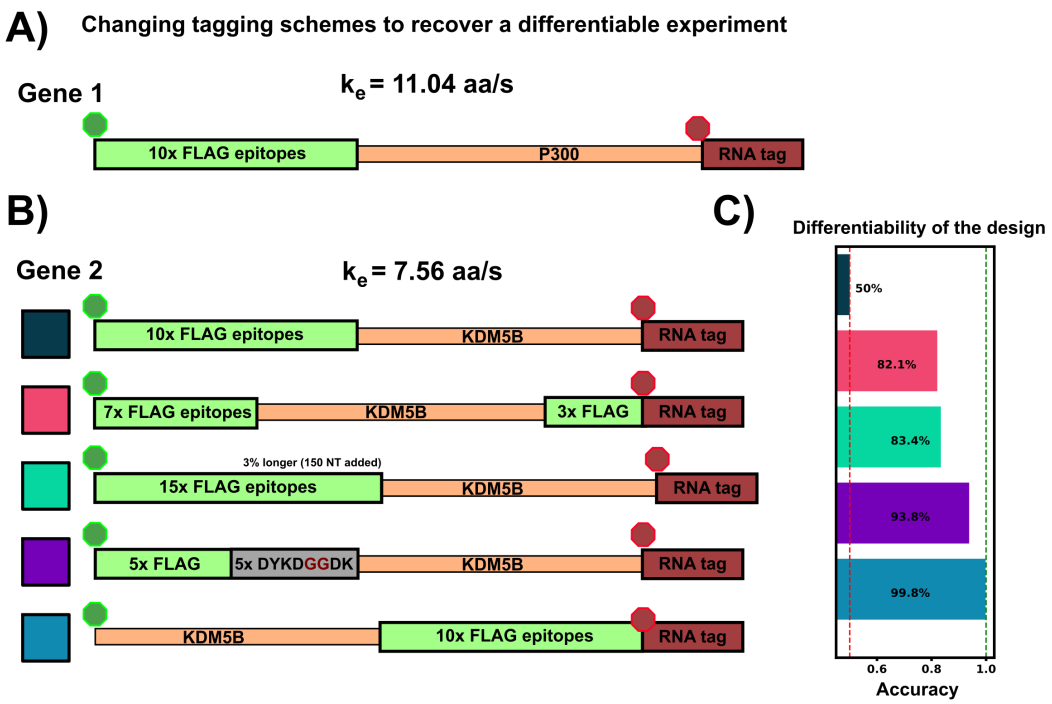


Figure 2.29: Changing tag designs to improve classification accuracy. A) Tag design for P300 construct is kept fixed B) Five different tag designs for KDM5B created by splitting the tag, increasing or decreasing the amount of epitopes, or relocating the tag region to the 3' end C) Accuracy for classification corresponding to each of the design combinations, and all assuming an elongation rate ratio of 1.46, under which the original design was non-classifiable (Figures 5D, 6D). All alternative designs would dramatically increase classification accuracy.

Fig. 2.28(C) shows that any of these permutations to the original 10x Flag tag on the 5' end of KDM5B allows the two mRNAs to be classified with greater than 85% test accuracy. Each tagging strategy changes the intensity dynamics of KDM5B spots, shifting them away from similar means and variances of the P300 spots, allowing classification without resorting to using a different color

tag. One should note that each of these strategies comes with its own potential drawbacks: Moving the 10x flag to the end doesn't allow any information about the upstream translation dynamics to be captured in the NCT experiment, and removing / moving 5 epitopes tag to the 3' end creates a dimmer spot, potentially obscuring translation dynamics under study. Adding epitopes also has its own potential drawback of needing longer plasmids for transfection.

Classifiers can retain accuracy despite uncertainties or assumption errors in biophysical parameters

In the previous sections, we explored how well classification would work in the ideal situation in which the classifier is trained on data that match (in probability) to the circumstances of the testing data (although every intensity trajectory is different due to stochastic fluctuations in translation, motion, and cellular background noise). In other words, the stochastic model used to generate the training data was the same as the model used to generate the testing data. For a more realistic test of how well one might expect a classifier trained using simulated data to work when applied to experimental data, one must acknowledge that true biophysical parameters are unknown, and they may vary from cell to cell or from one individual mRNA to the next. For example, in the analysis of KDM5B and P300, it is reasonable to assume that the mRNA designs and lengths are known, but one might only have a rough estimate for the initiation and elongation rates based on analyses for other mRNA, cells, or conditions. Ideally, the classifier should still work despite finite errors in these parameter estimates. To explore how well a simulation-trained classifier might work when parameters are incorrect, Fig. 2.31 shows the accuracy versus *unknown* rates k_{e} and k_{i} when the model is trained at three specific assumptions for those rates (denoted by red squares in Fig. 2.31). When the model is trained with a fast elongation rate and a slow initiation rate (Fig. 2.31A), classification accuracy is always poor (as discussed above, see Fig. 2.31B). However, when the model is trained in a condition that is more conducive for accurate classification (Fig. 2.31B), the accuracy is strong not only for the exact parameters under which the model was trained, but for a large range of surrounding parameter sets. The practical implication of this result is that one could in principle use an approximate model (e.g., the simulations presented in this study) to train a classifier and

then reliably trust that classifier despite unavoidable but finite errors in the underlying model or parameter assumptions. To maximize the utility of such a classifier, we performed a search over all possible sets of parameters on which to train the model and asked which training set leads to a classifier that would work best when averaged over the unknown "true" parameters. Fig. 2.31C) depicts the accuracy of this model (which is trained at $k_{\mathrm{i}} = 0.07$ s^{-1} and $k_{\mathrm{e}} = 6.44$ aa/s) as a function of all parameters, and it results in an expected average classification accuracy of 70% but with classification greater than 80% for large regions of parameter spaces.

Classification is possible with photobleaching and tracking errors.[2]

To examine the effects of photobleaching on classification, simulated videos were subjected to 11 different photobleaching rates before processing. 75 Simulated cell videos of 350 frames at

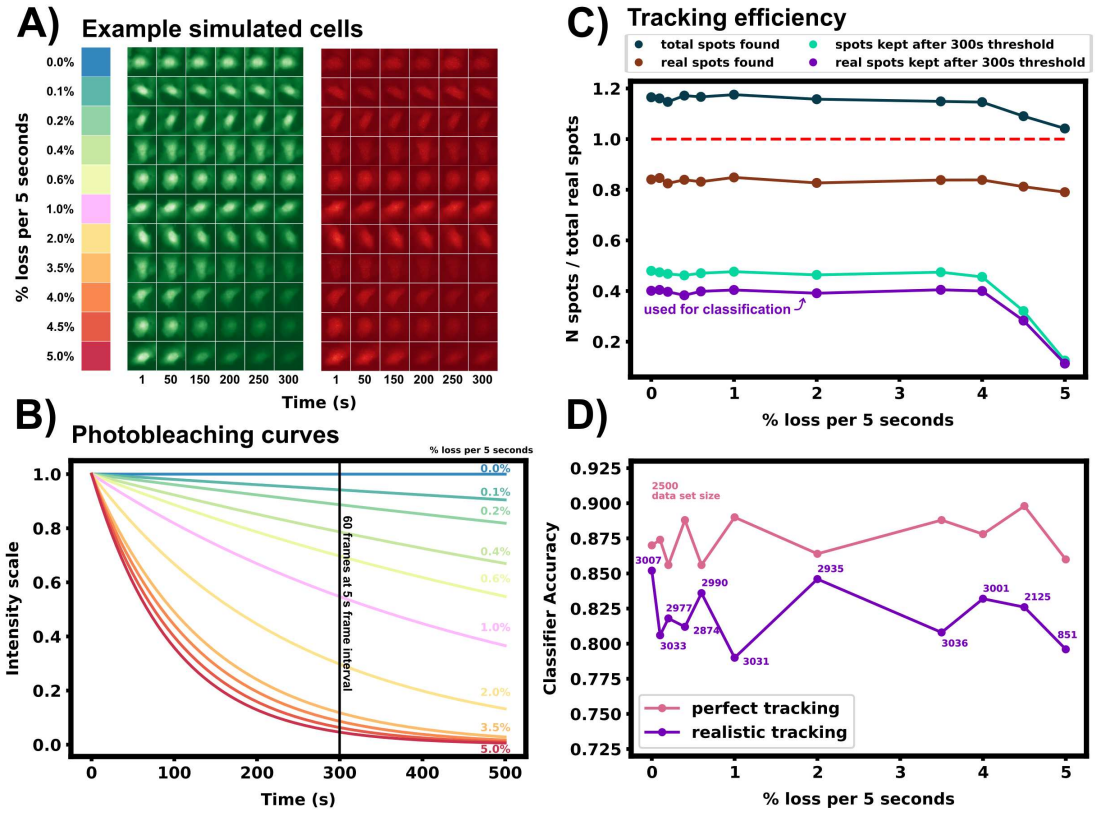


Figure 2.30: Caption on next page.

[2]This section was moved from its original location in supplemental information.

Figure 2.30: Effects of photobleaching and tracking on machine learning classification. A) Representative examples from 75 simulated cells of KDM5B and P300 at base conditions (Table 2) that were simulated at each of 11 different photobleaching rates. Tracking was performed on RNA spots in the red channel at a frame interval of 1 s. Translation was quantified in the green channel for 60 frames at a 5 s interval. B) Photobleaching intensity curves that scale video frames (red and green channels) to generate simulated cell videos (panel A). C) Efficiency for realistic tracking is reported as (# spots tracked / # total true spots). Total number of spots found by Trackpy per data set includes false positive spots, real trajectories, and real trajectories that are improperly linked or "fragmented," resulting in more spots being found than real simulated spots. When filtering with a squared error to real simulated spots, 80% of real spots are recovered in full or fragmented form (brown line). Requiring any spot, real or fake, to be tracked longer than 300 seconds results in $\sim$50% recovery until large photobleaching rates (green line). Filtering for spots with low matching error and existing longer than 300 seconds recovers $\sim$40% of true spots until larger photobleaching rates (purple line). The purple line represents the detected spot subset that was used for training in the "realistic tracking" condition. D) Classification accuracy versus photobleaching loss rate under perfect tracking (pink) or realistic tracking with image processing errors (purple). 2,500 spots were used for training in the perfect tracking case. For realistic tracking, the actual number of identified and true tracks varies and is shown in purple numbers for each photobleaching rate (out of 7,500 total generated spots). When training classifiers, 500 spots were withheld for testing classification accuracy.

1 second frame interval containing 25 spots of KDM5B and 25 spots of P300 with k_{i} = 0.06 s^{-1} and k_{e} = 5.33 $aa \cdot s^{-1}$ were simulated for each photobleaching rate ranging from 0.01% lost per 5 seconds to 5% lost per 5 seconds. Diffraction limited RNA spots were added to the channel zero (red) for simulation of particles for tracking; These RNA tag spots were co-localized with the channel one (green) spots for their corresponding NCT protein signal. Photobleaching was added post simulation by multiplying the resultant videos by the normalized photobleaching curves shown in Supplemental Fig. 2.30B. Sample videos are shown in Supplemental Fig. 2.30A for each photobleaching rate. Extrema were removed from each video channel by normalizing intensities to the 0.05$^{\mathrm{th}}$ and 99.9$^{\mathrm{th}}$ percentile of intensity. Particle tracking was performed on each video after image processing. The video's red channel was filtered for spots by applying a bandpass filter (skimage.difference_of_gaussians, low_sigma = 0.1, high_sigma = 5, truncate = 3) and Laplace of Gaussian's filter (scipy.ndimage.gaussian_laplace, sigma = 1.5). The filtered video was passed to Trackpy [101]. Trackpy was iteratively called on the filtered video's red channel with increasing until the number of spots detected levels off (approximated derivative of # spots detected / intensity threshold = 0). The location where the number of spots detected levels off is used as the final intensity threshold. Tracking efficiency results are shown in Supplemental Fig. 2.30C. This iterative

Trackpy approach on average finds 120% of the true simulated number of spots —consisting of false positive spots $\sim$33%, real spots and real spots whose trajectories are not correctly linked resulting in multiple particle tracks from one spot, $\sim$66%. To filter out false positive spots and short detected trajectories, all spots were then matched with their minimum squared error divided by the length of the tracked trajectory (time normalized squared error) to the true locations from their simulated video. Spots with an error lower than 3 pixels were considered "real"particles and kept for classification tests, $\sim$80% of real spots were recovered in some form with this error matching. Additionally, any spot not detected for $\geq$300 seconds was excluded from classification. Percentage of spots detected and kept for classification as a function of photobleaching rate is shown in the purple line in Supplemental Fig. 2.30C. Tracking is around 40% efficient at recovering long trajectory true spots unless the videos significantly photobleach before the threshold of 300 seconds. All final spots used for classification were down-sampled to 60 frames at a 5 second frame interval for classification. Classification was performed on the "perfect tracking" using 2000 training spots / 500 withheld spots. The "realistic tracking"data set was trained on the maximum possible recovered spots from the 75 total cells while withholding 500 spots for validation. Supplemental Fig. 2.30D shows the accuracy versus photobleaching rates with "perfect tracking"and with the "realistic tracking"from the Trackpy pipeline. Test accuracy stays level across photobleaching rates for both types of tracking; This would change with longer videos that can supplement their classification with the frequency information of the autocorrelation —the videos would be easier to classify with lower photobleaching rates where a better autocorrelation can be measured. However, at only 300 seconds of video, there is not enough time to acquire adequate dwell time information and the classifier is exclusively using intensity differences to tell spots apart; With a consistent photobleaching across genes, these intensity differences are also consistent at any given time point (until all intensity is lost). Greater in-depth simulations of photobleaching such as differing photobleaching rates for different elements of the simulation or other photobleaching models and their effect on classification can be explored in the future.

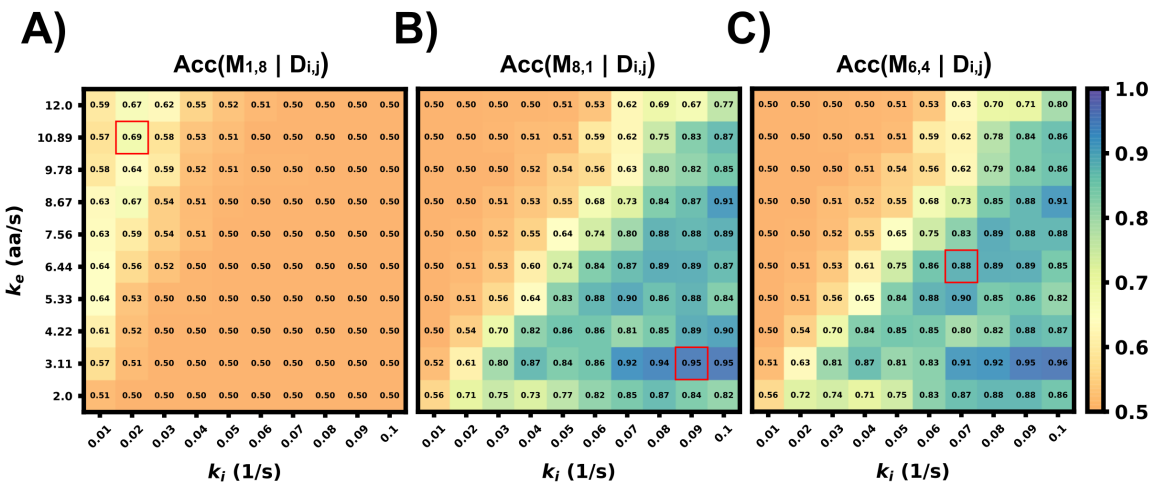


Figure 2.31: Accuracy of classifier when trained with incorrect parameter assumptions. Accuracy *versus* the actual rates k_e and k_i when the model is exclusively trained on three specific, but possibly incorrect, sets of these parameters A) (k_i =0.02 s^{-1}, k_e =10.89 aa/s), Average accuracy = 52.4% B) (k_i =0.09 s^{-1}, k_e =3.11 aa/s), Average accuracy = 70.1% C) (k_i =0.07 s^{-1}, k_e =6.44 aa/s), Average accuracy = 70.2%.

Using combinations of intensity fluctuation information and different color mRNA tags, one could design experiments to distinguish several mRNAs within the same cell

Finally, to highlight the how our proposed pipeline might be used to increase the potential of NCT experiments, we demonstrate it on the simulation of seven tagged species within a single cell under the assumption of consistent initiation and elongation rates across spots. Using the Fig. 2.27A heat map, we selected four mRNAs species that were differentiable from each other with a higher than 90% accuracy for the green channel, and three mRNA species for the blue channel with the same accuracy threshold, Table 2.7. A single simulated "multiplex" cell video was generated with the conditions described in both tables. Ten spots for each of the seven mRNAs were added to the appropriate color channel for each cell. An additional class of non-translating spots was also added by taking 2500 trajectories from the opposite (no-spot) channel, but with the same Brownian motion. The ML architecture was adjusted to account for the multiple mRNA labels and the noise label for non-translating spots, and the final layer was set to a softmax layer and an output of the number of species in each color channel (four for blue and five for green, including one the non-translating spots in each channel). A model was trained with the process described in

Table 2.7: Multiplexing simulation parameters

Comparison Analysis	Variable Range	Constants
Green Channel (4 spots)	mRNA length (without tag): * RRAGC (1200 NT) * LONRF2 (2265 NT) * MAP3K6 (3867 NT) * DOCK8 (6000 NT)	* Frame interval: 5s * Frames: 64 * k_{i}: 0.06 s^{-1} * k_{e}: 5.33 $aa \cdot s^{-1}$ D: 0.925 $\mu m^2/s$ SNR: 3.7-15.2
Blue Channel (3 spots)	mRNA length (without tag): * ORC (1734 NT) * TRIM33 (3333 NT) * PHIP (5466 NT)	* Frame interval: 5s * Frames: 64 * k_{i}: 0.06 s^{-1} * k_{e}: 5.33 $aa \cdot s^{-1}$ D: 0.925 $\mu m^2/s$ SNR: 5.2-14.9

the ML section for each color channel using the matching data from the construct length dataset, (7500 NCT spots for blue, 10000 spots for green with 80:20 train-test split and three-fold cross validation). Test Video trajectories were normalized with the same min-max scaling as the training data.

After training, the models were applied to classify the simulated trajectories in 50 new multiplexing videos. Fig. 2.32A shows representative images of the artificial ML labeling results applied to an example multiplexing video. Correctly identified spots are denoted by circles and incorrect classification results are shown with 'x'. A representative crop for each spot type is shown on the left. Fig. 2.32B shows the confusion matrix for each mRNA class and the blank trajectories. The green channel classifier had an 81% accuracy (disregarding blank trajectories) and struggled mostly with the two middle length genes LONRF2 and MAP3K6. The blue channel classifier had a 91% accuracy disregarding noise only spots. As expected, the majority of misclassified spots were to their length neighbors that have the most similar statistics. The shortest, dimmest mRNA NCT spots, ORC2 and RRAGC, were still classifiable from the non-translating spots with almost 100% accuracy. 2.2% of RRAGC spots were misclassified as noise. Overall on the test video, the classifier correctly identified 64 out of 70 spots, demonstrating the potential to label multiple species in the same cell with a correct tagging scheme and ML labeling.

In addition to the discrete labels generated by the classifier, the softmax output (shown in Fig. 2.32C, left) provides a quantification of the classification confidence for each spot. Fig. 2.32C shows the effect of sorting all classified non-noise spots by their softmax output and discarding those spots with the lowest confidence for classification. By discarding 50% of the spots with the lowest confidences yields a dramatic improvement in accuracy for both the green channel (from 81% to 92%) and blue channels (from 90% to 98%). In other words, although perfect classification may not be achievable, one can focus on confidently identified mRNA to analyse their behaviors, e.g., to determine how different mRNA types respond to subsequent cellular signals, drugs, or stress perturbations. It is important to note, our architecture is small but no softmax calibration was used; a better metric of confidence could potentially be obtained by using a softmax calibration in future works.

2.3.4 Conclusion and Future Work

Temporally- or spatially-resolved activation or repression of translation provides for a potent mechanism by which cells could rapidly alter their protein content in response to cellular signals [110, 111, 112, 113, 114]. Recent advances in Nascent Chain Tracking (NCT) experiments has made it possible to observe this regulation at the level of single mRNA molecules in living cells [28, 32, 35, 38, 91]. However, limitations on the number of distinct fluorophores prevents current NCT experiments from exploring more than one or two different mRNA species at a time.

In this work, we use computational simulations to propose a solution to circumvent this limitation. Specifically, we provide a pipeline (Fig. 2.33) to combine mechanistic models (including detailed simulation of nascent protein elongation and corrupted by fluorescence background and camera noise) with machine learning to classify mRNA species based on their fluctuating fluorescence intensity signals in NCT experiments. We show that multiple mRNAs labeled with identical fluorescent tags could be distinguished provided that the mRNAs have some variation in their intensity distributions or fluctuation frequency content, e.g., due to different lengths (Figs. 5a, 6a) or different translation parameters (Figs. 5B-D). We also demonstrate how our computational

pipeline could help to guide the design of experiments to make it easier to access these features and distinguish between different types of translating mRNA. Specifically, by changing multiple biophysical parameters or experimental design variables —e.g., changing tag design (Fig. 2.29), mRNA lengths (Fig. 2.27a), ribosomal elongation and initiation rates (Figs. 5 B-D), or the length and temporal resolution of NCT movies (Fig. 2.25) —we explored which realistic designs of NCT experiments would provide insight for classification, and which would not allow for NCT multiplexing.

Our realistic simulations show that that ML labeling accuracy higher than 80% can be achieved under reasonable NCT experimental settings (Figs. 2-9). Longer videos with appropriately-chosen

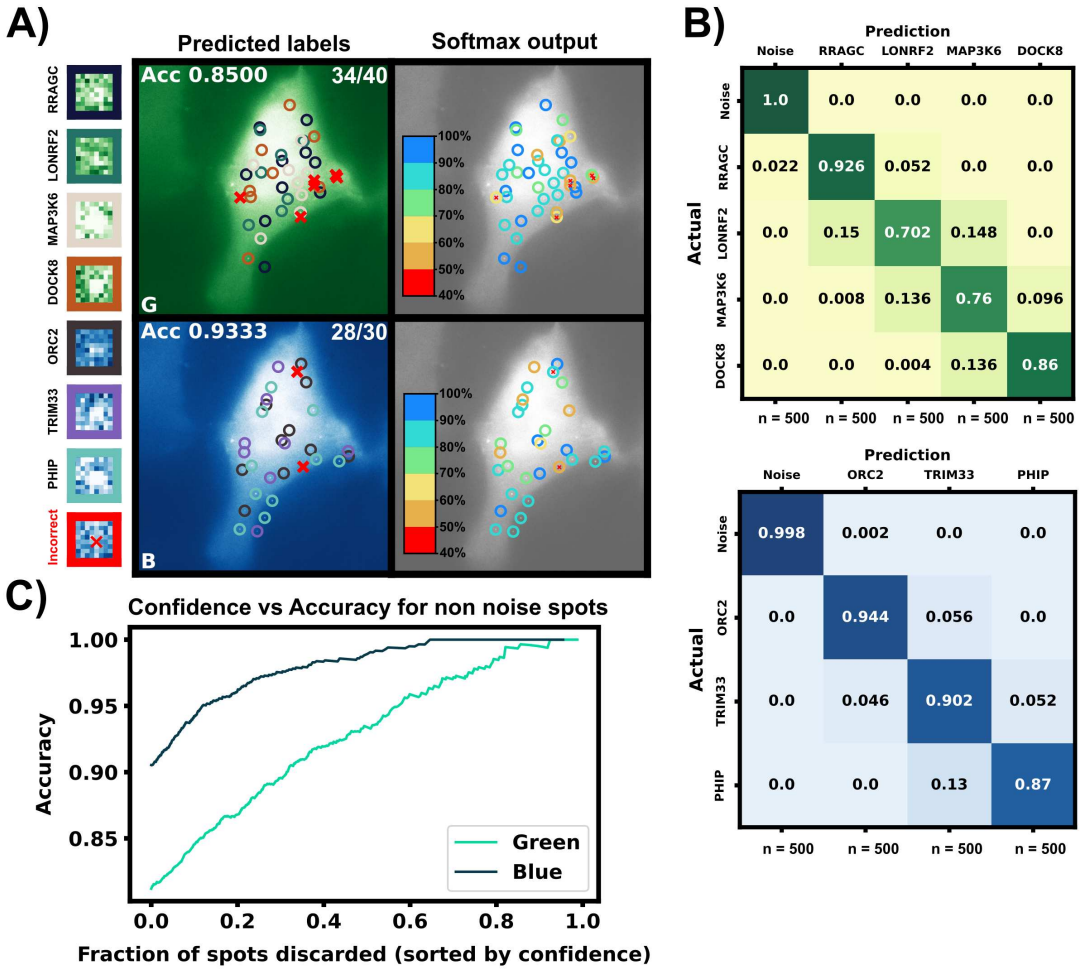


Figure 2.32: Caption on next page.

Figure 2.32: Simulated multiplexing of seven different mRNA species in a single cell. Ten mRNAs each of RRAGC, LONRF2, MAP3K6, and DOCK8 with identical were simulated in the green channel, and ten mRNAs each of ORC2, TRIM33, and PHIP were simulated in the blue channel with our pipeline, and all with identical tag designs and parameters. Our architecture was modified for multiclass labeling, and a model was trained for the green and blue channel for artificial labeling of the example video A) Example frame from video classification with seven different mRNA transcript types. Incorrectly labeled spots are marked with an X (6/70 spots). Crops of example spots are show to the left B) Confusion matrices for the green and blue channels when tested on 50 cells containing 10 spots of each mRNA C) Accuracy of the classifier versus the fraction of low-confidence spots that is discarded. If one only considers the 50% most confident spots, then accuracy rises to 93.4% and 98.9% for the blue and green channels, respectively.

frame intervals would lead to a better classification of NCT signals, albeit with diminishing returns (Figs. 4,6). A more strategic approach shows that selectively tagging pairs of mRNA species to achieve the greatest difference in expected intensity or frequency content achieves higher classification with a minimal number of frames (Fig. 2.32). We also demonstrated that different tagging strategies (Fig. 2.29) can help to separate hard to classify mRNA species, with tag design options ranging from simply adding more tag epitopes to increase one mRNA's intensity, to relocating or dividing the tag region between the 5' and 3' end of the CDS to alter the frequency content. Using these strategies, additional fluorophore colors can be held in reserve for mRNAs that are too similar in their lengths or dynamics, and we demonstrate that our multiplexing pipeline could distinguish seven different mRNAs at greater than 90% accuracy when using only two different fluorescent tag colors and only one per gene (Fig. 2.32). Additional mRNA could be considered by combining multiple colors on the same mRNA (e.g., mRNA with red and green could easily be distinguished from those with red or green alone). Based on these promising results of our detailed simulations, we envision that the next generation of NCT experiments will be able to track and differentiate multiple identically-tagged translating mRNA within the same cells (especially if these use the identified experiment designs for tag positions, gene lengths, and video frame intervals).

A limitation of our proposed use of simulations to train classifiers for experimental data, is that creating realistic simulations requires prior knowledge of system parameters. Some of these parameters are known in advance (e.g., relative mRNA lengths and codon usage), but others need to be estimated from literature values or preliminary experiments. In many cases, initial

control experiments would be needed to ensure that NCT constructs are working for individual mRNA, and these parameters could be estimated for individual mRNA before attempting multiplexing experiments. In other cases, the important rates (elongation and initiation) may remain similar for different mRNA provided they are analyzed in the same cell types. For example, in [50], three mRNA of different lengths were analyzed and elongation was found to be constant $k_{\text{elongation}} = 10.6 \pm 0.72s^{-1}$ while k_{i} was similar among different mRNA, $k_{\text{i}} \in \{0.022 \pm 0.004, 0.05 \pm 0.01, 0.066 \pm 0.019\}s^{-1}$. To evaluate the potential to transfer our model-based findings to real experiments where parameters are only partially known, we verified that models learned from one set of mechanistic parameters could correctly classify mRNA when tested on data that are generated using different parameters guesses (Fig. 2.31). This fact that classification accuracy remains high despite inexact knowledge of the model parameters offers hope that classifiers learned using approximate models could work on data from real experiments, without a need for collecting excessive training data. Even in the case where initiation rates are highly variable for different mRNA (e.g., for different regulatory elements in the 3' or 5' UTRs), there remains some hope to classify mRNA. In this case, with sufficiently long videos, nascent protein fluctuation frequencies could be used to identify mRNA if the lengths are sufficiently different (e.g., Figures 5D, 6D, and S4). However, if videos are too short, then one may still be able to differentiate between spots of different types based on their intensity distributions, but without additional information about initiation rates (e.g., by collecting a handful of longer videos to assign labels to each group of mRNA), it would be impossible to determine which mRNAs are which type, and additional experiments (e.g., direct mRNA labeling using Fluorescence in situ Hybridization to quantify ribosomal load or spatial correlations [115, 88, 116]) may be required.

Although, for the sake of brevity, this paper does not explore all mechanisms that can be analyzed by the rSNAPsim and rSNAPed computational pipeline (e.g., effects of tracking, photobleaching, variable probe-biding rates, ribosome pausing, etc.), future work could expand on these capabilities along with further exploration of different machine learning approaches (e.g., different ML architectures or different types of classifiers) or inclusion of additional features (e.g., includ-

ing mRNA diffusion rates, cell position information, fluorophore bleaching rates, etc.). Similarly, although the current manuscript has focused exclusively on supervised learning techniques that require known ground truth (e.g., labels in simulated data), one could potentially improve the application to real data through the addition of unsupervised machine learning or transfer learning approaches. For example, the simulation-based pipeline proposed here could be used to design tags and experimental conditions, while unsupervised approaches may be applied to differentiate spots in subsequent experiments. Finally, beyond the goal of multiplexing, we believe that our general approach to combine detailed mechanistic models, realistic simulations of microscopy and image analyses effects, and machine learning classifiers could help to design improved experiments that are more suited to other biological questions (e.g., to differentiate between competing hypotheses for translation mechanisms rather than to differentiate different mRNA species as explored here).

2.3.5 Methods

Simulated NCT experiment pipeline

Fig. 2.33 graphically describes the current study's computational pipeline, which combines the Rna Sequence to NAscent Protein SIMulator (rSNAPsim) and rSNAP Experiment Designer (rSNAPed) scientific libraries to generate synthetic training and testing data sets for multiple experimental conditions. rSNAPsim is a Python module that provides simulated fluorescent intensity traces from a given mRNA transcript using a codon-dependent TASEP to simulate ribosomal elongation [50]. This mechanistic model for translation assumes two parameters: the ribosomal *initiation rate*, k_{i}, is defined as the average number of ribosomes to initiate translation per second, assuming that other ribosomes are not blocking the initiation site. The *elongation rate*, k_{e} is defined as the global stepping rate (aa/s) averaged over all coding regions in the genome, again assuming no ribosome-to-ribosome exclusion. Because the actual local stepping rates depend the specific codon usage of an mRNA, and because rSNAPsim includes a 9-codon ribosome exclusion footprint that prevents two ribosomes from occupying the same site at the same time, the effective stepping rate for each mRNA is based on the specific sequence of the mRNA. For full

details on the rSNAPsim model, including calculations for the individual codon elongation rates, see [50]. We have shown previously that the rSNAPsim module can reproduce much of the fluctuation statistics observed in NCT translation experiments [50, 28, 35]. However, the movement of NCT spots across the cell background can lead to large drops in intensity when, for example, a spot leaves a bright nuclear region and enters a dimmer cytoplasmic region. These movements from areas of high to low or low to high backgrounds cause intensity fluctuations that are not due to the translation process itself but are, nevertheless, always present in real experimental data.

rSNAPed is a second python module that accounts for artifacts of microscopy, motion relative to a heterogeneous cellular background, and image processing effects. rSNAPed combines the rSNAPsim model intensity prediction with a point spread function, a controllable signal-to-noise ratio, simulated cellular background based on experimental video of non-labeled cells, and simulated motion for the NCT spots. To create a video for any length of time, rSNAPed takes 20-frame videos from one of seven unlabeled cells, randomly rotates and flips the videos and uses pixel-by-pixel statistics to formulate distributions from which to generate new simulated frames. Specifically, for each simulated frame, rSNAPed uses each pixel's empirical mean and standard deviation to draw a new Gaussian distributed value for the respective pixel[4] Supplemental Figure Fig. 2.12 provides a comparison of real cell background video and a simulated video from rSNAPed. Simulated mRNA spots are added to this cell background by simulating a point spread function on a 3x3 pixel patch, which is centered at a position that moves according to Brownian motion. After simulating the NCT experiment, videos are processed to find intensity trajectories using the "disk and doughnut" method [28], where the instantaneous signal is quantified as the difference between the average of the disk (3x3 patch centered at the spot) compared to the average of the doughnut (9x9 patch excluding the 3x3 disk patch). Units of intensity are reported as "units of mature protein"or UMP, which is calculated as the number of complete epitope tags

[4]Problem pixels that have extremely large standard deviations (e.g., for a video where a given pixel's intensity becomes extremely bright for one randomly-timed frame) are corrected to the 95th percentile of all pixels' standard deviations.

in the NCT spot (i.e., if a simulation has two ribosomes downstream of a 10xFLAG tag and one halfway through the tag, the intensity at that time is 25 epitopes or 2.5 UMP).

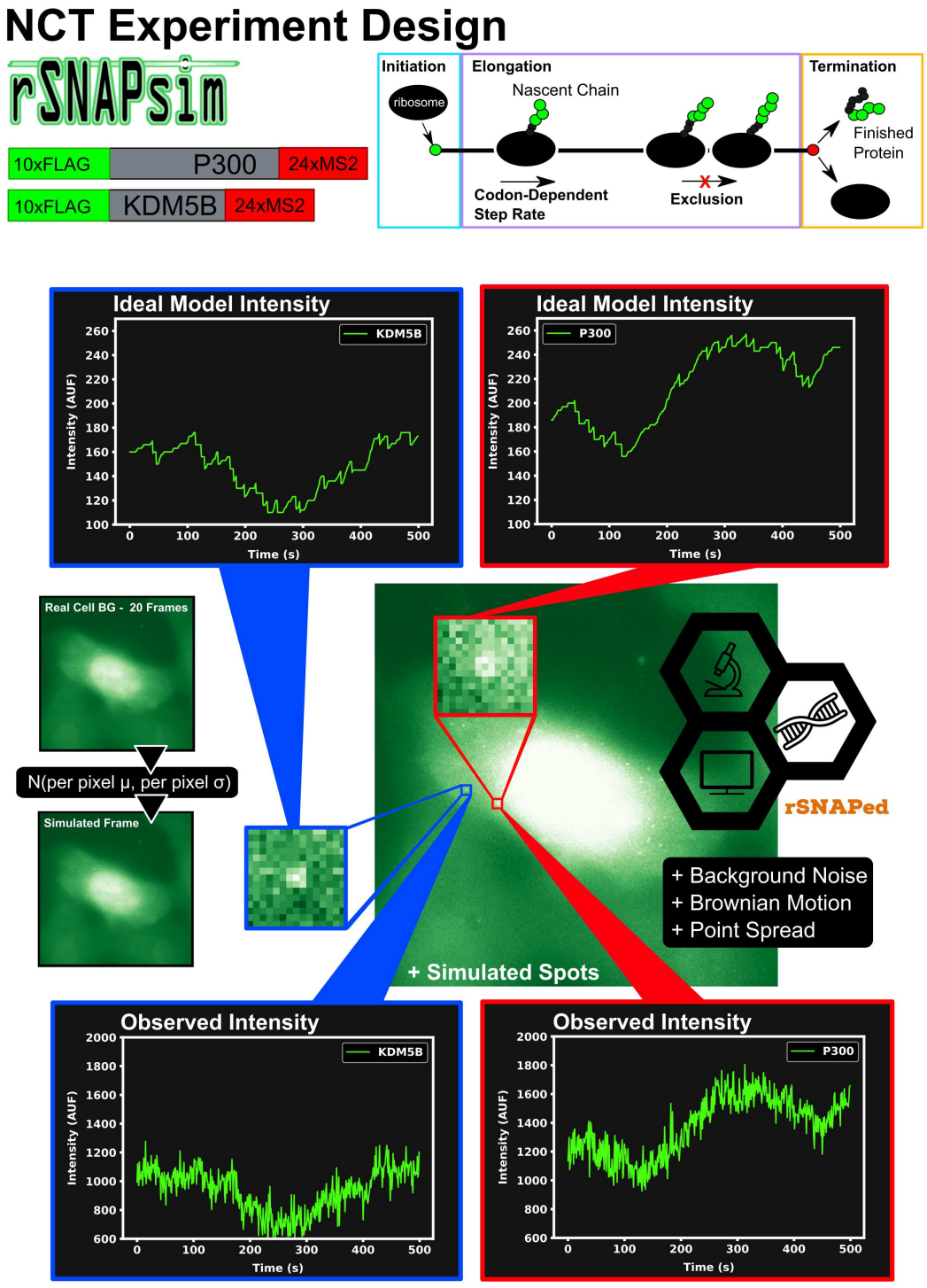


Figure 2.33: Caption on next page.

Figure 2.33: Simulated NCT experiment pipeline. Overview of approach to simulate Nascent Chain Tracking data and assign labels. rSNAPsim provides simulated NCT fluorescent intensity trajectories from a codon-dependent TASEP model for each mRNA spot. rSNAPed adds experimental spatial movement (Brownian motion) and temporal noise by introducing a point spread function for each spot. Simulated cell background frames are generated randomly from a per pixel Gaussian distribution with their means and standard deviations taken from 20 frames of real blank cell backgrounds. Spots in videos are processed with the disk and doughnut method to generate simulated NCT intensity data.

The combination of rSNAPsim and rSNAPed allows generation of vast amounts of synthetic video in different situations (e.g., for different mRNA sequences, different biophysical parameters, and different imaging conditions) that can match the translation statistics and spatial heterogeneity that an NCT spot would experience as it moves around a cell. For each mRNA in each condition, we generate 2500 simulated NCT trajectories (5000 total trajectories for two mRNA types in the same video). Theses trajectories are collated from 100 independent NCT simulated cells, each containing 25 spots of each mRNA for 3000 seconds at one second resolution (25 spots $\times$ 2 classes $\times$ 100 cells = 5000 total NCT spots for 3000 seconds). Smaller data sets can be generated as needed from these full length data sets by slicing the 1 second $\times$ 3000 frame video trajectories to the desired frame interval and number of frames. By generating such data for multiple different potential experimental conditions, we can train and test our machine learning methods to ascertain which feasible experimental conditions are most favorable to allow for successful classification.

Machine Learning

Fig. 2.34A shows the machine learning architecture used to classify mRNA spots. Given two different mRNA species in the same cell and NCT observations that rely on identical tags, one could attempt to classify the mRNA based on their intensity signals, their particle sizes, or their x, y, and z coordinates over time (with z having poorer resolution than x and y). Of these, we focus on intensity signals, which contain both statistical moments, such as the signal means and variances, as well as signal frequency content, similar to that which could be obtained with methods like spectrogram analysis or fluctuation correlation spectroscopy (FCS). We apply convolutional neural networks to classify the NCT simulation data based on one or both types of signal intensity

inputs. First, to prioritize extraction of features related to intensity statistics, we use min-max normalization of all intensity signals in the same video. Let $I_{\mathrm{i}}(t)$ denote the intensity for spot $i \in (1, 2, \ldots)$ and at frame number $t \in [0, 1, \ldots, T-1]$ that has been collected from a given NCT experiment or simulation video. Define $I_{\max}$ and $I_{\min}$ as the maximum and minimum spot intensities across all spots and all times in the video:

$$I_{\max} = \max_{\mathrm{i,t}} I_{\mathrm{i}}(t), \text{ and } I_{\min} = \min_{\mathrm{i,t}} I_{\mathrm{i}}(t). \tag{2.5}$$

The min-max normalization of each signal, $I_{\mathrm{i,norm}}$, which scales features from zero to one for later convenience in machine learning, is computed as:

$$I_{\mathrm{i,norm}}(t) = \frac{I_{\mathrm{i}}(t) - I_{\min}}{I_{\max} - I_{\min}}. \tag{2.6}$$

Second, to prioritize features related to fluctuation frequencies, we use the normalized empirical auto-correlation function for each spot, $G_{\mathrm{i}}(\tau)$, which is defined as the sample covariance between the signal fluctuation at frames t and $t+\tau$, is calculated as:

$$G_{\mathrm{i}}(\tau) = \frac{1}{T-1-\tau} \sum_{\mathrm{t}=0}^{\mathrm{T}-1-\tau} \left(\frac{I_{\mathrm{i}}(t) - \mu_{\mathrm{i}}}{\mathbf{\Sigma}_{\mathrm{i}}} \right) \left(\frac{I_{\mathrm{i}}(t+\tau) - \mu_{\mathrm{i}}}{\mathbf{\Sigma}_{\mathrm{i}}} \right), \tag{2.7}$$

where t is an index corresponding to the frame number, τ is the correlation lag time; T is the total number of frames; and μ_{i} and $\mathbf{\Sigma}_{\mathrm{i}}$ are the i^{th} signal's trajectory average and standard deviation, respectively [23, 76].

The normalized inputs from Equations 2.6 and 2.7 are each passed to their own separate convolutional 1D layer and subsequent max pooling layer for feature extraction. For convenience, the two convolutional layers have the same size filter kernels and amount of filters. The extracted feature vectors from each input are concatenated and passed into one fully connected layer with a cross-entropy objective to classify the mRNA. The entire network (2 conv1D, 2 maxpooling, 1 dense) was trained end to end, and elastic net regularization is used to reduce over-fitting [117].

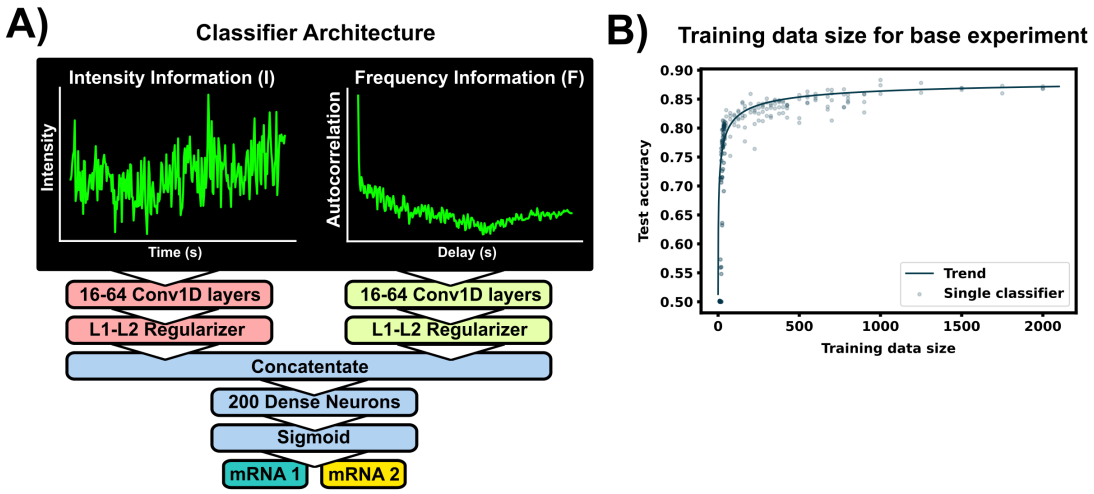


Figure 2.34: Machine learning classifier and training data size. A) The ML model consists of two separate convolutional layers–one receiving a normalized fluorescent trajectory and the other an intensity autocorrelation. The filter outputs are regularized and concatenated for a dense layer of 200 neurons for classification. Fundamentally, this architecture learns off frequency and intensity information from the NCT trajectory. B) Accuracy of the architecture for different training data sets. A total of 4,000 unique spot trajectories were split into 2–10 independent training data sets of the specified size. A classifier was trained with each training data set and tested on the same withheld validation set of 1000 NCT spots. A trend-line was added by fitting a Hill function to the test accuracy average across 15 bins in training data size $\left(y = 0.5 + \frac{0.391}{1+(6.336/x)\cdot 521}\right)$. The architecture was applied on simulated P300 and KDM5B trajectories from the selected base experimental condition (5 s frame interval, 64 frames, 0.06 1/s initiation rate, 5.33 aa/s elongation rate).

During training, data was split with an 80:20 ratio into training and testing sets, and training was performed with a 3-fold cross validation using a random search to select hyperparameters. Specifically, we searched over the possible combinations of options listed in Table 2.8 [118], and selected the hyperparameter set that maximized validation accuracy. After training and hyperparameter selection, the final architecture was tested on the 20% withheld test data to quantify the model performance on unseen data.

Hardware

All machine learning experiments were done with TensorFlow 2.2.0 on 2x NVIDIA Geforce 2080 Supers. Simulated NCT experiments were generated on an AMD Ryzen Threadripper 3970X 32-Core Processor utilizing 8 threads for each simulated NCT experiment generation. Time for

Table 2.8: Hyperparameter optimization grid

Hyperparameter	Variable range
kernel size	1x3, 1x5, 1x7
number of kernels	16, 32, 64
batch size	16, 32, 64
epochs	50, 100

generating one 5000 spot NCT experiment with base experimental conditions: $\approx$ 2h 44m. Time for generating one 50-spot cell with base experimental conditions (short output + not saving video): $\approx$ 197 seconds

Microscopy

The seven background videos used for video generation in the rSNAPed were captured using a custom-built wide-field fluorescence microscope with a highly inclined illumination regime with 488, 561, and 637 nm excitation beams (Vortran) ([79, 23]). An objective lens of 60X, NA 1.49 oil immersion, Olympus, was used. The emission signals were split by an imaging grade, ultra-flat dichroic mirror (T660lpxr, Chroma) and detected using two separate EM-CCD (iXon Ultra 888, Andor) cameras via focusing with 300 mm tube lenses ultimately producing 100X images with a 130 nm/pixel resolution. JF646 signals were detected with the 637 nm lasers and the 731/137 nm emission filter (FF01-731/137, Semrock). Cy3 signals were detected with the 561 nm lasers and the 593/46 nm emission filter (FF01-593/46, Semrock)

The lasers, the cameras, and the piezoelectric stage were synchronized via an Arduino Mega board (Arduino) and image acquisition was done with open source Micro-Manager software [119]. An imaging size of 512x512 pixels^2 was used and exposure time was set to 53.63 msec. This resulted an imaging rate of 13 Hz with 23.36 msec readout time and 13 Z-stacks were captured at 500nm step size.

U-2 OS cells were loaded with Cy3-FLAG Fab and JF646-Halo-MCP 6 to 10 hours prior to imaging. Right before imaging, cells were transferred into the stage top incubator set to a

temperature of 37°C and supplemented with 5% CO_2 (Okolab). Acquired 4D (xyzt) images were processed to 3D (xyt) maximum intensity projections for the rSNAPed image generation.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Contributions

William Raymond:, Investigation, Computational modeling, Theory, Computational simulation, Machine learning, Writing, Editing

Sadaf Ghaffari: Machine learning, Editing

Luis U. Aguilera: Computational modeling, Theory, Computational simulation, Editing

Eric Ron: Conceptualization

Tatsuya Morisaki: Conceptualization, Editing

Zachary R. Fox: Conceptualization, Computational modeling, Theory

Michael May: Computational modeling, Theory

Timothy J. Stasevich: Conceptualization, Editing, Funding acquisition

Brian Munsky: Conceptualization, Computational modeling, Theory, Writing, Editing, Funding acquisition

Funding

WSR, ER, and BM were supported by the NSF (1941870). LUA and BM were also supported by National Institutes of Health (R35GM124747). TJS and TM were supported by the NSF (1845761).

Data Availability Statement

The data sets for this study can be resimulated, classifiers can be retrained, and figures can be remade from the code in the MunskyGroup/Multiplexing_project Github Repository https:

//github.com/MunskyGroup/Multiplexing_project, DOI: https://zenodo.org/record/7884701. Large videos files to remake the figures are stored at: https://doi.org/10.5281/zenodo.7887820. Original saved classifiers, example videos, and original data set CSVs are available upon request.

Chapter 3

Transcription Modeling

3.1 Live-cell imaging reveals the spatiotemporal organization of endogenous RNA polymerase II phosphorylation at a single gene[5]

Dr. Linda Forero-Quintero and Dr. Tim Stasevich approached my PI and I in late spring 2019 and requested a computational model fit to their live-cell single-molecule transcription experiments of an endogenous reporter gene. The goal of the computational modeling was to fit the amount of RNAP2 colocalizing to the promoter region, providing additional evidence that more RNAP2 were localized than space available on the promoter region. This paradoxical over localization is consistent with a transcription phase condensate model proposed circa 2017 [120], and Dr. Forero-Quintero wished to confirm their results do indeed provide evidence for phase condensate transcriptional regulation although the core of the paper was the live-cell single-molecule application for transcription. My contribution to this paper was computationally exploring and fitting several models and ultimately performing model selection of a class of bursting models over the course of late 2019 and 2020. The final model fits did indeed suggest that more RNAP2 join the promoter cluster than space on the actual promoter. Our modeling efforts also suggested an additional experiment clarifying the transition rate of Ser5 phosphorylation as their original data could not quantify this transition rate.

To highlight my contributions to this paper, much of the relevant modeling information has been moved from the supplemental information and given more detail here. The original paper is provided as is and the added section is denoted with a disclaimer.

[5]Linda S. Forero-Quintero, William Raymond, Tetsuya Handa, Matthew N. Saxton, Tatsuya Morisaki, Hiroshi Kimura, Edouard Bertrand, Brian Munsky, Timothy J. Stasevich, DOI: https://doi.org/10.1038/s41467-021-23417-0

3.1.1 Summary

The carboxyl-terminal domain of RNA polymerase II (RNAP2) is phosphorylated during transcription in eukaryotic cells. While residue-specific phosphorylation has been mapped with exquisite spatial resolution along the 1D genome in a population of fixed cells using immunoprecipitation-based assays, the timing, kinetics, and spatial organization of phosphorylation along a single-copy gene have not yet been measured in living cells. Here, we achieve this by combining multi-color, single-molecule microscopy with fluorescent antibody-based probes that specifically bind to different phosphorylated forms of endogenous RNAP2 in living cells. Applying this methodology to a single-copy HIV-1 reporter gene provides live-cell evidence for heterogeneity in the distribution of RNAP2 along the length of the gene as well as Serine 5 phosphorylated RNAP2 clusters that remain separated in both space and time from nascent mRNA synthesis. Computational models determine that 5 to 40 RNAP2 cluster around the promoter during a typical transcriptional burst, with most phosphorylated at Serine 5 within 6 seconds of arrival and roughly half escaping the promoter in $\sim$1.5 minutes. Taken together, our data provide live-cell support for the notion of efficient transcription clusters that transiently form around promoters and contain high concentrations of RNAP2 phosphorylated at Serine 5.

3.1.2 Introduction

In eukaryotic cells, the catalytic RPB1 subunit of RNA polymerase II (RNAP2) possesses an extended carboxy-terminal domain (CTD) that consists of heptapeptide repeats (52 in humans) with a consensus sequence (Tyr_1-Ser_2-Pro_3-Thr_4-Ser_5-Pro_6-Ser_7). The CTD region is dynamically phosphorylated as RNAP2 progresses through the transcription cycle, regulating each step of transcription, from initiation to termination. In some models, RNAP2 is recruited to promoters in an unphosphorylated form (CTD-RNAP2), but is later phosphorylated at Serine 5 (Ser5ph-RNAP2) upon initiation and at Serine 2 (Ser2ph-RNAP2) during active elongation [121, 122, 123, 124]. Interest in the CTD has recently increased due to observations of highly dynamic RNAP2 clustering [124, 125, 126] that correlates with the phosphorylation status of the CTD7, [127]. In par-

ticular, recent data suggest that a transcriptional cluster forms around gene promoters early in the transcription cycle. The cluster is thought to be enriched in unphosphorylated- and Ser5ph-RNAP2 that appear to constrain chromatin movement near the transcription start site [128]. However, upon transcriptional activation, hyperphosphorylation of RNAP2 at Ser2 allows the enzyme to escape the cluster and begin active elongation [129, 128]. The dynamic clustering of RNAP2 involves many steps and a complex orchestration of multiple factors and could therefore represent a global form of transcriptional regulation [130].

RNAP2 phosphorylation throughout the transcription cycle has traditionally been studied in fixed cells using immunoprecipitation-based assays [121, 123, 131]. These studies provide precise spatial maps of the average positions of RNAP2 along the 1D genome. Unfortunately, the inherent averaging masks heterogeneity, and the procedure limits temporal resolution to timescales of tens of minutes or longer [132]. RNAP2 dynamics can instead be imaged and quantified in living cells using fluorescence microscopy, overcoming the limitations of traditional assays. Recent single-molecule tracking technologies [79, 133, 134, 135, 136] has made it possible to monitor single RNAP2 as they bind at non-specific locations throughout the genome [125, 137] as well as a specific, single-copy genes [126, 136] pre-marked with MS2 [55, 138] or PP7 [56] RNA stem-loops (that are lit up co-transcriptionally when, respectively, fluorescent MS2 or PP7 coat proteins bind to them). Each of these studies used permanent fluorescent fusion tags to track RNAP2. Fusion tags are incapable of discerning post-translational modifications to RNAP2, including transcription cycle-associated phosphorylation events.

One way to resolve post-translational modifications to RNAP2 is to use antibody-based probes that bind and light up specific modifications to residues within the CTD in vivo [139, 140, 61, 141, 142]. However, the signal-to-noise is limited with this approach because of the presence of unbound and freely diffusing probes that increase the fluorescence background (BG). Applications have therefore been restricted to large tandem gene arrays. Signal-to-noise is amplified by the multiple copies of a gene within these arrays, but heterogeneity from one gene copy to another is again

masked by averaging [143]. Therefore, the spatiotemporal dynamics of RNAP2 phosphorylation at single-copy genes remain unclear.

Here, we combine multicolor single-molecule microscopy, complementary fluorescent antibody based probes, and rigorous computational modeling to visualize, quantify, and predict endogenous RNAP2 phosphorylation dynamics at a single-copy reporter gene in living cells. This unique combination of technologies allows us to directly visualize the temporal ordering and spatial organization of RNAP2 phosphorylation and mRNA synthesis throughout the transcription cycle at the reporter gene. We find evidence for relatively high concentrations of RNAP2 near the beginning vs. end of the gene that are both spatially and temporally separate from elongating RNAP2 and nascent mRNA synthesis. Collectively, our data provide live-cell support for the existence of higher-order, phosphorylation-dependent transcriptional clusters that dynamically form and surround active genes throughout the transcription cycle.

3.1.3 Results and Discussion

Technology to visualize endogenous RNAP2 transcription cycle dynamics at a single gene

To visualize the spatiotemporal dynamics of endogenous RNAP2 phosphorylation at a single gene, we used an established HeLa cell line (H-128) harboring an MS2-tagged HIV-1 reporter gene and stably expressing both GFP-tagged MS2 coat protein (MCP) and an untagged HIV-1 trans-activator of transcription (Tat) [138]. We chose HIV-1 as our reporter gene because it is a prototypical model for RNAP2 phosphorylation [144]. The HIV-1 reporter is strongly active in our cell line due to persistent stimulation by Tat, producing a bright MCP signal that pinpoints the location of the transcription site (TS) and gauges its activity in real-time [138] (Fig. 3.1a). Consistent with this strong signal, immunostaining experiments in fixed cells revealed the TS is highly enriched in RNAP2 and relatively depleted in histones and their epigenetic modifications (Supplementary Fig. 3.2a,b). Chromatin immunoprecipitation (ChIP) experiments furthermore confirmed the presence of CTD-RNAP2 and its phosphorylated forms Ser5ph- and Ser2ph-RNAP2, respectively. In particular, we detected that CTD-RNAP2 and Ser5ph-RNAP2 signals are highest at the

transcription start site, whereas Ser2ph-RNAP2 is highest towards the end of the gene (Fig. 3.1b). However, because these data come from a population of fixed cells, whether the various forms of RNAP2 are present at the same time and place and whether or not they appear in a preferred order is difficult to extract from this assay.

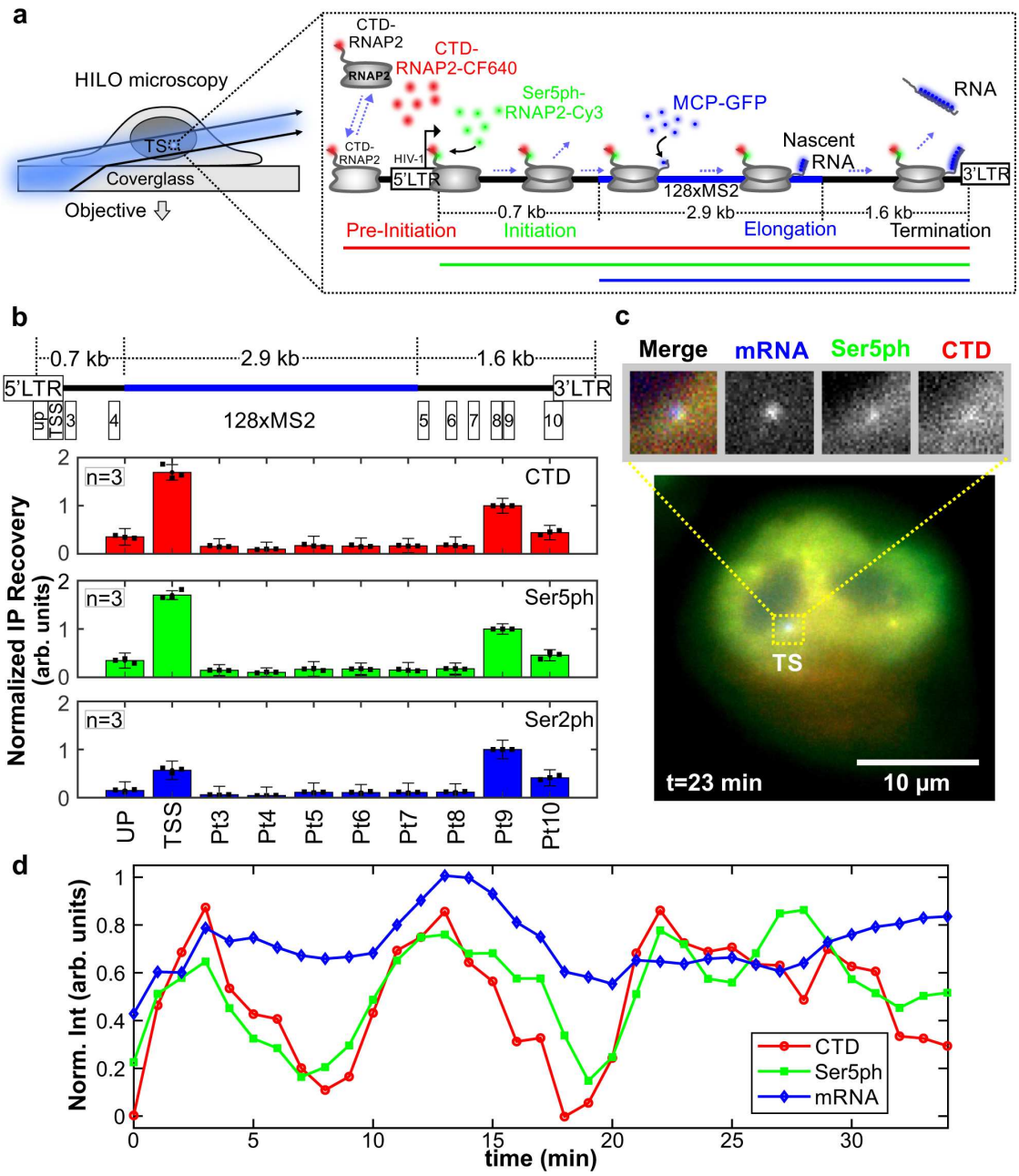


Figure 3.1: Caption on next page.

Figure 3.1: A system for imaging the endogenous RNAP2 transcription cycle at single genes. A) Schematic of the system. The reporter gene is controlled by the HIV-1 promoter and is tagged with a 128xMS2 cassette (blue bar). RNAP2 is represented in gray. RNA is marked by MCP-GFP that binds to the transcribed MS2 stem-loops (mRNA, blue). The recruited and initiated RNAP2 are labeled by Fabs (conjugated with CF640 and Cy3) that bind unphosphorylated and phosphorylated CTD RNAP2 heptad repeats (CTD, red) and Serine 5 phosphorylated repeats (Ser5ph, green), respectively. B) Average chromatin immunoprecipitation occupancy of CTD-RNAP2 (red, upper panel), Ser5ph-RNAP2 (green, middle panel), and Ser2ph-RNAP2 (blue, lower panel) across the HIV-1 reporter gene (positions 1-10 are highlighted in the cartoon above). Data are presented as mean values +− S.E.M. C) Sample live-cell showing CTD-RNAP2, Ser5ph-RNAP2, and mRNA co-localizing at the transcription site (TS), n=9 cells in 4 independent experiments. D) Normalized intensity at the TS over time from the cell in C for CTD-RNAP2 (red circles), Ser5ph-RNAP2 (green squares), and mRNA (blue diamonds).

To better characterize the spatiotemporal dynamics of single-cell RNAP2 modifications during transcription, we loaded fluorescent fragmented antibodies (Fab, generated from the same antibodies used in ChIP) [139, 145] recognizing (1) the CTD of RNAP2 (anti-CTD RNAP2) without or with residue-specific phosphorylations, and (2) heptad repeats within the CTD that are phosphorylated at Serine 5 (anti-Ser5ph RNAP2). These antibodies have previously been shown to be specific for their respective targets via Western blotting and ELISA [143] and in ChIP-seq [146] experiments. Fab generated from these antibodies has also been shown to rapidly bind and unbind their targets, making them valuable for monitoring temporal changes in RNAP2 phosphorylation [143]. Consistent with anti-CTD Fab labeling all RNAP2 and anti-Ser5ph Fab labeling a subset of RNAP2, we observed regions within the nucleus with RNAP2 enriched or depleted with Ser5ph (Supplementary Fig. 3.3). These capabilities allowed us to distinguish three distinct steps of the transcription cycle at the HIV-1 reporter gene: RNAP2 recruitment (marked by Fab against CTD-RNAP2), initiation (marked by both Fab against CTD-RNAP2 and Fab against Ser5ph-RNAP2), and elongation (marked by all Fab and MCP binding to mRNA), as depicted in Fig. 3.1a, c. Although we attempted to also visualize Ser2ph at the locus with our Fab, signal-to-noise was insufficient to detect in living cells, presumably because the antibody is not sensitive enough to recognize this modification at the single-gene level.

Nevertheless, this setup has several advantages that collectively enhance signal-to-noise at the TS. First, Fab binds endogenous RNAP2, so all RNAP2 in the cell has a high likelihood to be

labeled without having to genetically engineer a fusion knock-in tag [137, 126] and/or alpha-amanatin resistance [125]. Second, fluorescence is naturally amplified since mammalian RNAP2

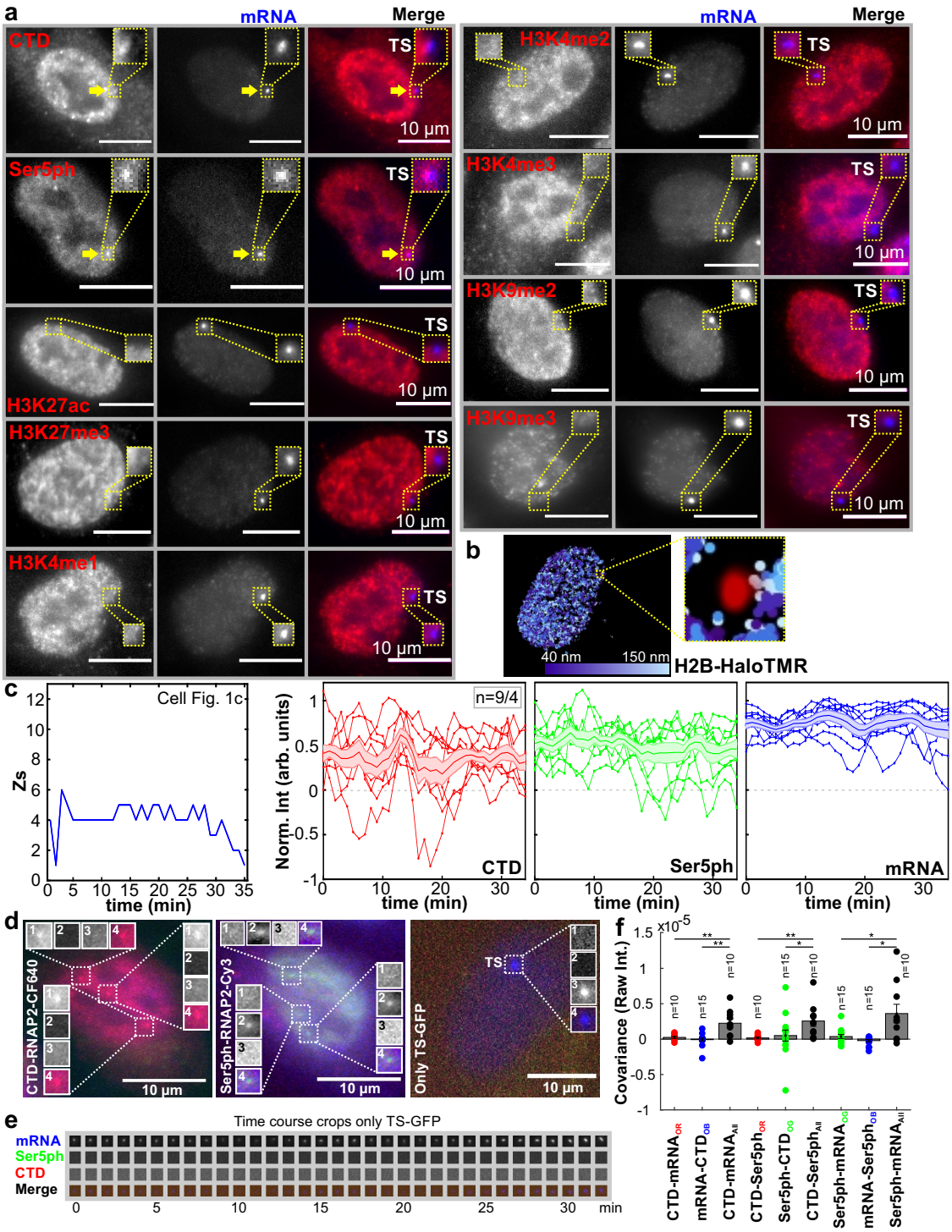


Figure 3.2: Caption on next page.

Figure 3.2: **(a)** Immunostaining (red, left panels) of CTD-RNAP2 (n=11), Ser5ph-RNAP2 (n=10), histone H3K27ac (n=12), H3K27me3 (n=14), H3K4me1-3 (n=3) and H3K9me2-3 (n=3) at the HIV-1 transcription site (TS) marked by MCP-GFP (mRNA; blue, center panels), along with a merge (right panels). **(b)** Representative cell showing a mobility map of single H2B tracks. The blue scale shows the average frame-to-frame jump size (one frame every 43.86 ms) for each tracked molecule. The track corresponding to the transcription site is shown in red. The yellow dashed box displays a zoom-in around the transcription site region, where H2B is depleted (n=17 out of 28 cells in 3 total days). Control experiments for photo-bleaching showing **(c)** left panel, "best-Z"positions of the TS over time for the exemplary cell in Fig. 3.1c right panels, normalized intensity over time for CTD-RNAP2 (red circles), Ser5ph-RNAP2 (green squares), and mRNA (blue diamonds) for all the cells recorded as in Fig. 3.1c,d. The shadow and the line in the middle represent the S.E.M and the average. **(d)** Images of cells from bleed-through control experiments. Left, a cell loaded with just Fab marking CTD-RNAP2 (CTD-RNAP2-CF640) displays endogenous puncta that are not the TS (designated Only Red "OR"spots; n=10); Middle, a cell loaded with just Fab marking Ser5ph-RNAP2 (Ser5ph-RNAP2-Cy3) displays endogenous puncta that are not the TS (designated Only-Green "OG"spots; n=15); Right, a cell without Fab in which the TS is marked solely by GFP-MCP binding mRNA (Only TS-GFP; Only Blue "OB"; n=15). Cropped images show the various "OR", "OG", and "OB"sites where the individual channels are separated and labeled as follows: (1) Red (CTD-RNAP2), (2) Green (Ser5ph-RNAP2), (3) Blue (mRNA), and (4) Merge. **(e)** Cropped images in a time course at an "OB"site demonstrates no bleed through of the mRNA channel into the other channels. **(f)** Covariance between all possible pairs of raw intensity signals is not significant at "OR", "OG", and "OB"sites, but is significant at the TS in cells containing all three signals (i.e. cells loaded with both Fab and expressing MCP-GFP; All). n=number of cells/independent experiments. Data are presented as mean values $\pm$ S.E.M. Significance was tested using a two-tailed Mann-Whitney U-test with $p \leq 0.0136$ (*) and $p \leq 0.0091$ (**).

contains 52 heptad repeats in its CTD [147], each of which can be bound by a fluorescent Fab at the TS. Third, Fab continually binds and unbinds RNAP2, mitigating the loss of fluorescence due to local photobleaching. In combination with a multi-color, single-molecule microscope [23] employing oblique HILO illumination to enhance signal-to-noise by an order of magnitude [79], these advantages allowed us to generate movies in which we monitored endogenous RNAP2 phosphorylation dynamics at the HIV-1 reporter gene in 3-colors.

As shown in Fig. 3.1c, d, movies revealed correlated fluctuations between the mRNA signal and endogenous CTD-RNAP2 and Ser5ph-RNAP2 signals at the TS. To ensure correlations were not an artifact of focusing issues, we tracked the TS in 3D (by imaging 13 z-planes per time point) to keep the MS2 signal continually in focus (Supplementary Fig. 3.2c, left panel). The correlations were also not caused by photobleaching, as signals fluctuated both up and down throughout the entire imaging time course, remaining on average constant (Supplementary Fig. 3.2c). Finally, to rule out the possibility that correlated fluctuations were caused by bleed-through from one fluo-

rescence channel to another, we re-imaged cells lacking Fab. In all cases, no bleed-through was observed (Supplementary Fig. 3.2d, e), as quantified by the covariance between channels (Supplementary Fig. 3.2f). We, therefore, conclude the correlations reflect natural bursts in endogenous transcriptional activity at the HIV-1 reporter gene, demonstrating our ability to detect and quantify endogenous RNAP2 phosphorylation dynamics at a single-copy gene.

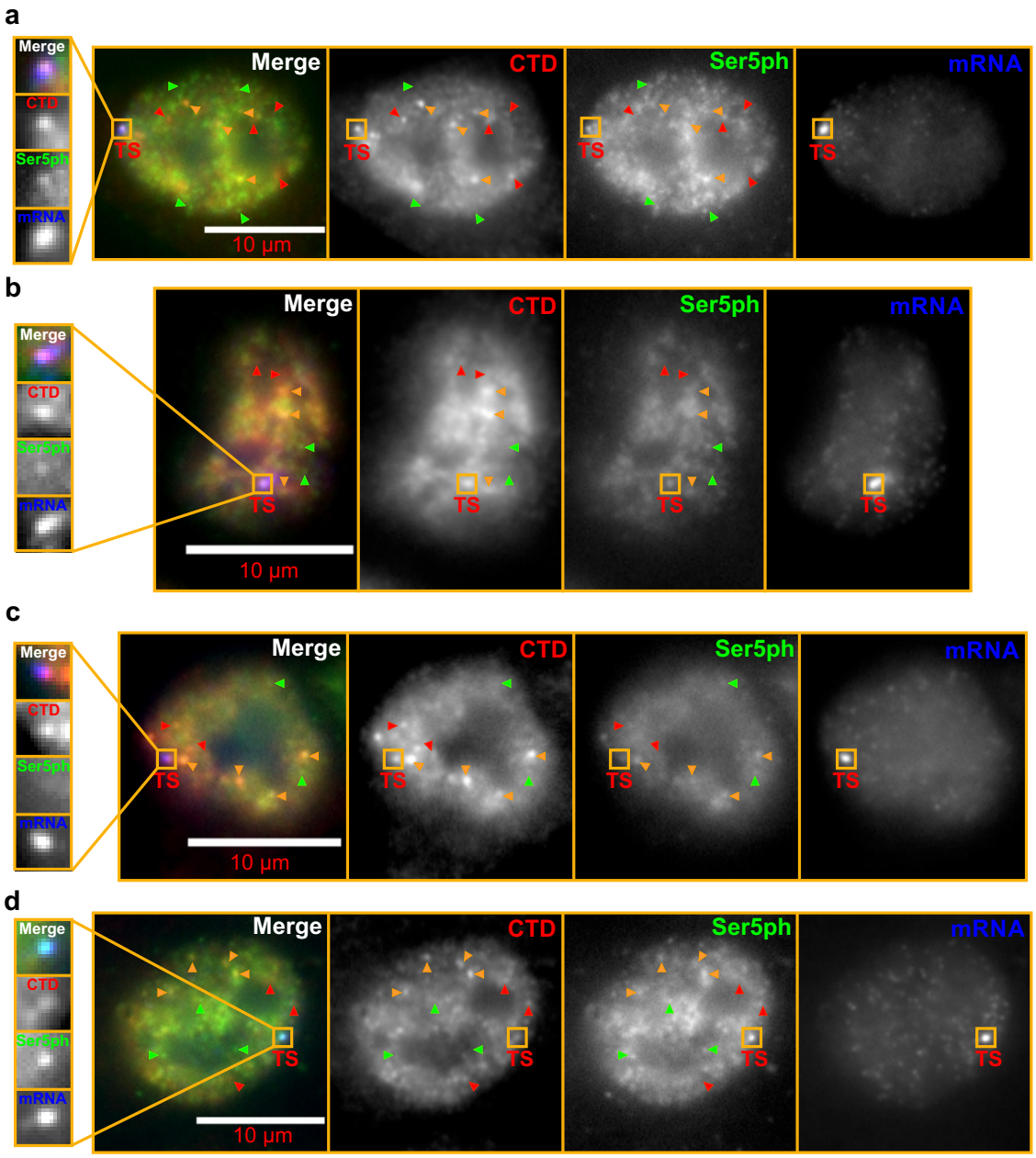


Figure 3.3: Caption on next page.

Figure 3.3: Fixed cells stained with our CTD- and Ser5ph-specific Fab. Cells with distinct staining patterns. Some areas within cell nuclei are enriched with CTD-specific Fab (red arrows), other areas are enriched with Ser5ph-specific Fab (green arrows), while still other areas are enriched with both Fabs (orange arrows). At the HIV-1 transcription site (TS), we typically see both Fabs present **(a)**, n=20 out of 29 cells. However, on occasion we can find TSs in which Ser5ph-RNAP2 staining is relatively dim **(b & c)**, n=8 out of 29 cells, or, in very rare cases, CTD-RNAP2 staining is relatively dim **(d)**, n=1 out of 29 cells. This provides evidence that the signals are fluctuating.

Long-term imaging of fluctuations at the reporter gene reveals temporal ordering of RNAP2 phosphorylation

In the majority of cells, the mRNA was steadily produced by the HIV-1 reporter gene, with strong signals persisting for hours at a time. In a few cells, the mRNA signal completely disappeared, indicating a loss of nearly all transcription activity. We were interested in capturing these rare events in a single time course to better discriminate the relative timing of our RNAP2 and mRNA signals. To accomplish this, we adjusted our imaging conditions to optimize detection of all three signals in single cells over a period of three hours (200-time points), as exemplified in Fig. 3.5a-c (also see Supplementary Movie 1 (available online), and Supplementary Fig. 3.4a-c). We again imaged in z-stacks (13 planes spaced by 0.5 μm) covering the whole nucleus at each time point throughout the entire experiment. We were therefore confident that the fluctuations were due to changes in transcription activity and not related to transcription site movement into and out of the focal plane (Supplementary Fig. 3.4d). With these imaging conditions, we found cells in which the mRNA signal turned on and off up to four times, indicating bursts of transcription and multiple complete transcription cycles. Consistent with our previous result, signals at the transcription site were highly correlated and fluctuated generally in unison, although there were distinct periods of time when one signal could be seen for multiple frames in the absence of some other signals. This again ruled out bleed-through and suggested the signals were not perfectly synchronized. To ensure the correlated fluctuations were specific to the locus and not cell-wide, we verified that covariances between the mRNA signal and the CTD or Ser5ph signals were significantly stronger when both signals were measured at the transcription site compared to when one or both signals were measured a short distance (p1) from the transcription site (Supplementary Fig. 3.4e-g).

Having established a well-controlled system to examine fluctuations at a single gene, we were confident in our ability to quantify the temporal ordering of RNAP2 and mRNA throughout the

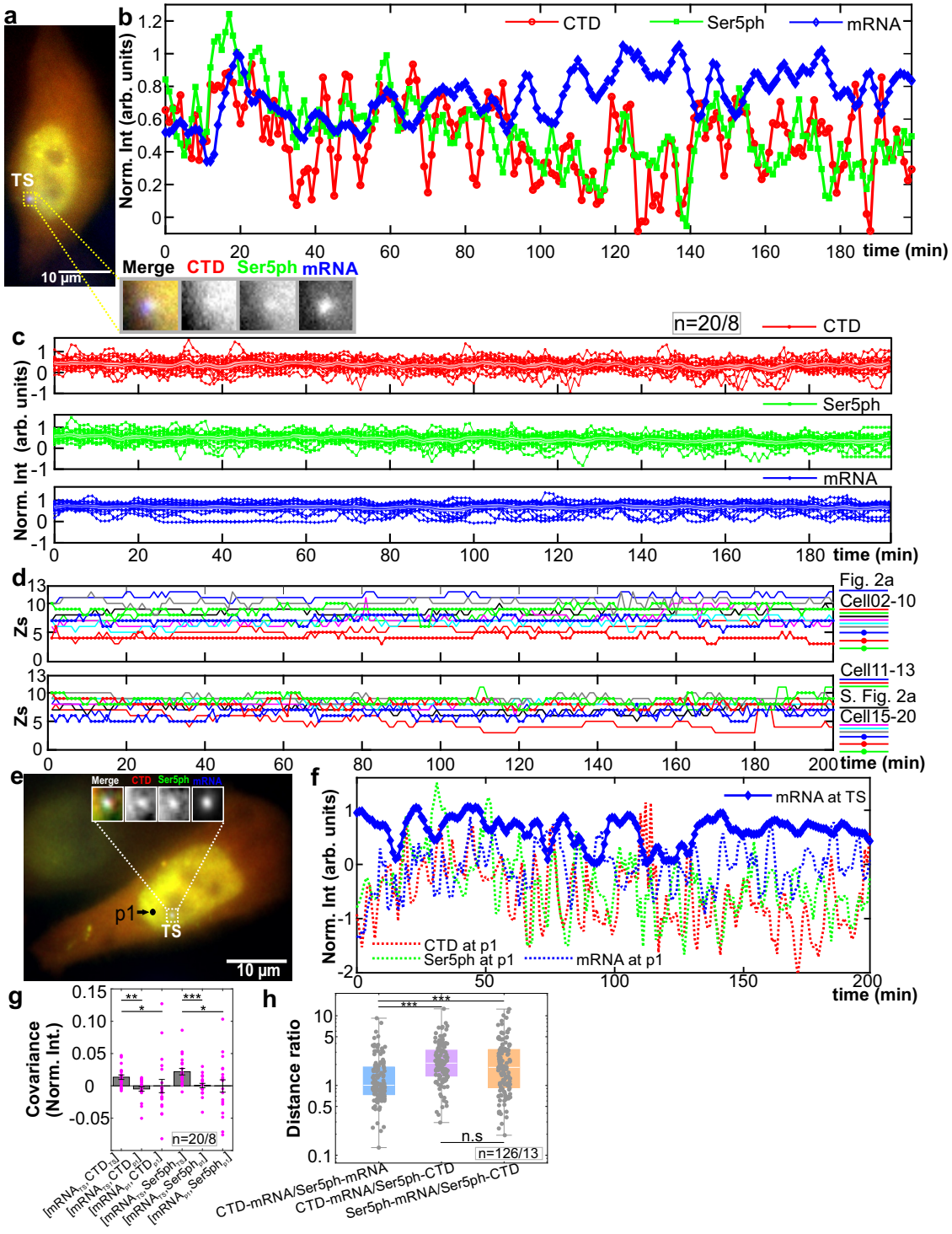


Figure 3.4: Caption on next page.

Figure 3.4: RNAP2 fluctuations at the HIV-1 reporter locus and off target.(a,b) A cell with strong and persistent transcription has co-localized CTD-RNAP2 (red circles), Ser5ph-RNAP2 (green squares), and mRNA (blue diamonds) at the transcription site (TS), n=7 out of 20 cells. **(c)** Normalized signal intensities over time for all the cells analyzed as in b. The shadow and the line in the middle represent the S.E.M and the average, respectively. **(d)** Z-positions of all the cells quantified for transcription fluctuations. Each cell is represented with a different color/symbol (legend on the right). **(e)** Exemplary cell with periods of active and inactive transcription showing a control position (p1) near the transcription site, n=13 out of 20 cells. **(f)** Normalized intensity over time-target at an off-target position near the transcription site (p1; CTD-RNAP2, dashed red; Ser5ph-RNAP2 dashed green; mRNA, dashed blue) versus the mRNA signal at the transcription site (blue diamonds). **(g)** Covariance calculation between the normalized intensities of mRNA at the transcription site against CTD-RNAP2 or Ser5ph-RNAP2 at the transcription site and at p1. Data are presented as mean values $\pm$ S.E.M. n=number of cells/number of independent experiments. **(h)** Ratiometric distribution of the euclidean distances for CTD-RNAP2 and mRNA to Ser5ph-RNAP2 and mRNA (light blue), CTD-RNAP2 and mRNA to Ser5ph-RNAP2 and CTD-RNAP2 (light purple), and Ser5ph-RNAP2 and mRNA to Ser5ph-RNAP2 and CTD-RNAP2 (light orange) in all the cells analyzed. The line in the middle of each box represents the mean. The top and the bottom of the box represent the 75% and 25% quantiles, respectively. The middle region in the error bar at the bottom and the top represent the lower and upper whiskers, respectively. n=number of events/number of cells. Significance was tested using a two-tailed Mann-Whitney U-test with $p \leq 0.05$ (*), $p \leq 0.0015$ (**), and $p \leq 9.2091 \times 10^{-4}$ (***).

transcription cycle. One thing that stood out was that peaks and troughs in the mRNA signal tended to come after the peaks and troughs in the RNAP2 signals. Although there were some exceptions due to the stochasticity of the system, in some cells this behavior was seen multiple times in even single time series (for example, see valleys at $t_{1\text{-}4}$ = 16, 75, 94, and 113 min in Fig. 3.5b, c). To better quantify this effect, we selected all events at which the mRNA signal dropped below a threshold value, extracted all three signal channels from seven minutes before to seven minutes after each event, and aligned all signals relative to these mRNA minima event times (Fig. 3.5d). This analysis revealed two important aspects of the dynamics of our system. First, the analysis confirmed the signals were strongly correlated since strong minima could be observed in all channels. These minima were significant compared to the results from unaligned signals (gray diamonds in Fig. 3.5d, p values of 1.72 x 10^{-10} for Ser5ph- and 6.39 x 10^{-4} for CTD-RNAP2). Such strong correlation between mRNA production at the HIV-1 reporter and endogenous RNAP2 would suggest the reporter is not part of a larger transcriptional unit containing multiple genes. Second, the analysis indicated a temporal ordering, with both RNAP2 signals coming before mRNA by 0.96 $+-$ 0.55 min for CTD-RNAP2 (p value 3.65 x 10^{-3}) and 0.88 $+-$ 0.24 min for Ser5ph-

RNAP2 (p value 1.28 x 10^{-5}). This delay makes sense because RNAP2 must escape the promoter and elongate 0.7 kb before it reaches the MS2 repeats. The CTD-RNAP2 signal also slightly preceded the Ser5ph-RNAP2 signal, although the delay was not significant at our sampling rate. This suggests nearly all RNAP2 at the locus either come in pre-phosphorylated or are rapidly phosphorylated at Serine 5 within a minute of arrival.

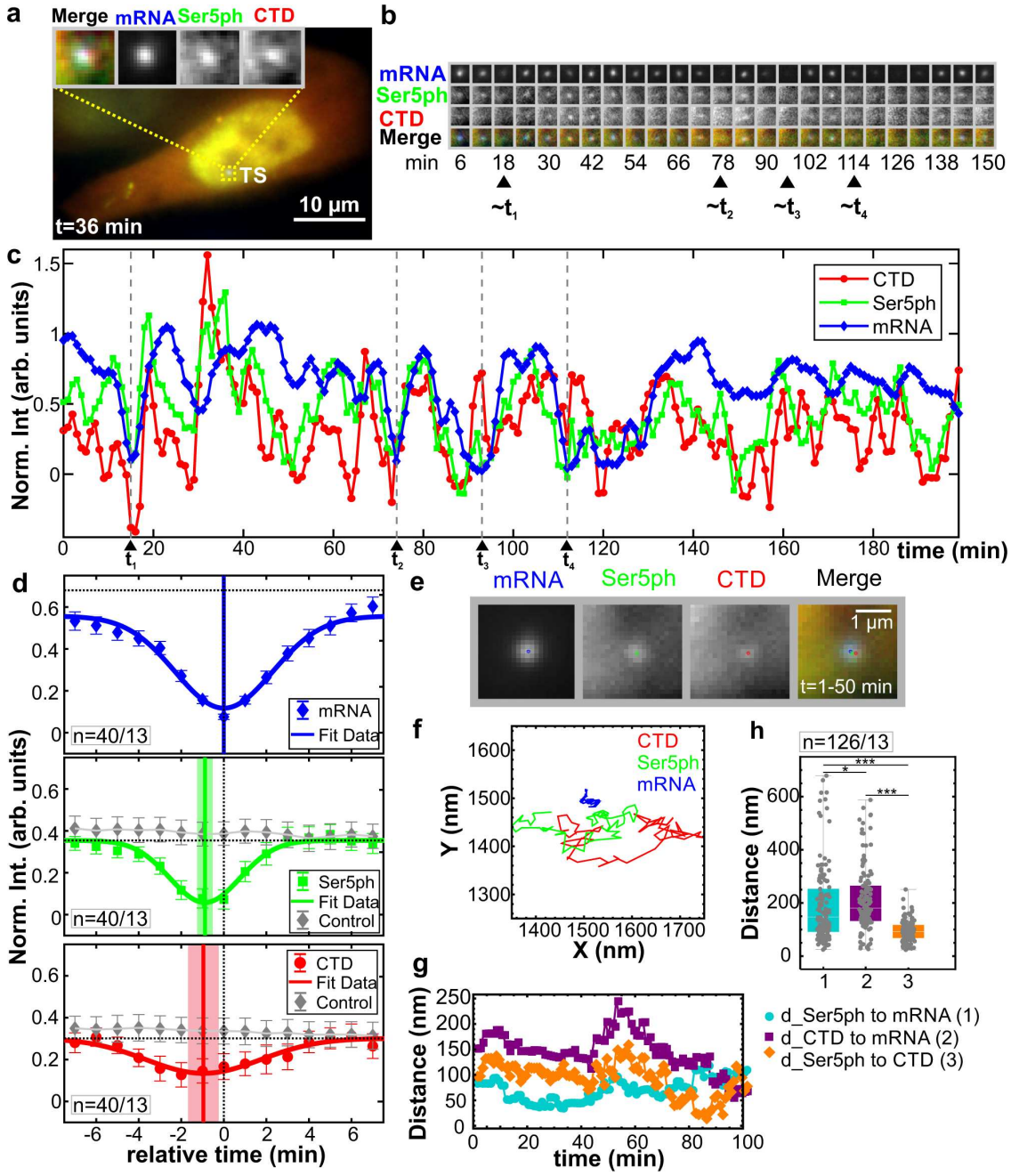


Figure 3.5: Caption on next page.

Figure 3.5: Spatiotemporal organization of the RNAP2 CTD cycle at the HIV-1 reporter gene. A) Sample cell showing co-localization of CTD-RNAP2, Ser5ph-RNAP2, and mRNA signals at the transcription site (TS), n = 13 out of 20 cells. B) The TS from (A) at select times. C) Normalized intensity fluctuations at the TS for CTD-RNAP2 (red circles), Ser5ph-RNAP2 (green squares), and mRNA (blue diamonds) vs. time. Times of minimal mRNA (less than 0.20 arb. units) are marked with dashed gray lines ($t_{1\text{-}4}$). D) The average normalized intensity of each signal surrounding times of minimal mRNA. Both the Ser5ph-RNAP2 and CTD-RNAP2 signals have deep minima well below the steady-state value (dashed horizontal line). Solid connecting lines show the Gaussian fit, and solid vertical lines mark the minima with a lighter shadow depicting the S.E.M. from the Gaussian fit. n = 40 events from 13 of 20 cells. Data are presented as mean values +− S.E.M. When the same analysis is performed at 100 random time points, no obvious minima are seen (gray diamonds). E) Cropped 50-frame (50 min total) moving-average image of the TS in (A) and the fitted center position for mRNA (blue), Ser5ph- (green), and CTD-RNAP2 (red), n = 126 events from 13 of 20 cells. F) 50-frame moving-average XY position of each signal at the TS in (A) over time. Note the mRNA signal was used as the reference signal within the crop. G) The distance between each signal in (F) over time: Ser5ph-RNAP2 to mRNA (cyan circles; (1)), CTD-RNAP2 to mRNA (Purple squares; (2)), and Ser5ph-RNAP2 to CTD-RNAP2 (orange diamonds; (3)). H) The distribution of distances measured as in (G) at all TSs in all cells analyzed (sampled every 10 min). The line in the middle of each box represents the mean. The top and the bottom of the box represent the 75% and 25% quantiles, respectively. The middle region in the error bar at the bottom and the top represent the lower and upper whiskers, respectively. Significance was tested using a two-tailed Mann-Whitney U test with $p \leq 0.0315$ (*) and $p \leq 6.969 \times 10^{-11}$(***).

Spatial organization of CTD phosphorylation at the reporter gene

RNAP2 is thought to be organized in phosphorylation-dependent clusters [129, 127]. To test this hypothesis, we measured the center position in X and Y of CTD-RNAP2, Ser5ph-RNAP2, and mRNA at the reporter gene over time (Fig. 3.5e-g). If the hypothesis is correct, we would expect to see some spatial separation in our different RNAP2 and mRNA signals. To confirm this hypothesis, we calculated the Euclidean distance between each pair of signals. As Fig. 3.5g illustrates, the distances between signals changed over time but were spatially organized such that the RNAP2 signals were significantly separated from mRNA.

Although there was considerable variation from cell to cell, this trend could be seen in the median positions from the whole population of transcription sites we tracked (Fig. 3.5h). Specifically, the median distance from mRNA to CTD-RNAP2 was ∼181 nm compared to ∼148 nm for Ser5ph-RNAP2 (p value 0.032). Likewise, the median distance between the two forms of RNAP2 was just ∼ 93 nm, significantly smaller than between either form of RNAP2 and mRNA (p value $< 6.97 \times 10^{-11}$) (Fig. 3.5h and Supplementary Fig. 3.4h). This spatial separation was consistent across

the cells we analyzed (Supplementary Fig. 3.6) and independent of the strength of transcription as gauged by CTD-RNAP2, Ser5ph-RNAP2, and mRNA signal intensities. Together these results demonstrate that RNAP2 is spatially organized within the transcription site, with active mRNA synthesis spatially distinct from clusters of CTD-RNAP2 and Ser5ph-RNAP2.

Fluctuation dynamics and statistics are captured by a simple model of transcription bursting

We wanted to obtain a more universal picture of RNAP2 phosphorylation dynamics at the HIV-1 reporter gene. We therefore performed correlation analysis [56, 82, 76] using all time points in all time series, similar to fluorescence correlation spectroscopy [148]. This technique is ideal for extracting information from noisy data provided there are a sufficient number of time series and/or time points. We began with an auto-correlation analysis, to see how long each signal remains correlated with itself given a lag time (τ) (Fig. 3.7a). The auto-correlation of each signal decays with increasing lag time and eventually flattens out near zero. We define the dwell time as the lag time at which the auto-correlation falls below 20% of its initial zero-lag value. According to this analysis, the two forms of RNAP2 had shorter average dwell times than mRNA, indicating RNAP2 was often unsuccessful in reaching the end of the gene and synthesizing an mRNA.

Next, we calculated the cross-correlation between signals. Consistent with our previous analysis aligning local minima, all possible pairs of signals were strongly correlated, as seen by large peaks in the cross-correlation curves near $\tau = 0$ (Fig. 3.7b). Measuring the precise position of each peak revealed the mRNA signal came substantially later than the CTD-RNAP2 and Ser5ph-RNAP2 signals, while the CTD-RNAP2 and Ser5ph-RNAP2 signals appeared at roughly the same time (within the 1 min sampling time of experiments). To better resolve the time delay between the CTD-RNAP2 and Ser5ph-RNAP2 signals, we re-imaged the HIV-1 TS in a single plane at a much faster frame rate (150 msec/frame) for a total of 1000 time points (150 sec). Although these higher temporal resolution experiments are much too short to capture the full auto- and cross-correlation curves, they are sufficient to resolve the short time-lag dynamics (Supplementary Fig. 3.14), and they revealed cross-correlation asymmetry with an off-center peak indicating that

the Ser5ph-RNAP2 signal comes roughly 3-6 sec after the CTD-RNAP2 signal. The various delays we measure are consistent with the temporal ordering we saw by aligning local minima of the mRNA signal (Fig. 3.5d) and provide further evidence that RNAP2 phosphorylation at Serine 5 is very rapid at the transcription site.

We next sought to find a quantitative model to unify our diverse data sets (Fig. 3.7c). We required that our model must simultaneously fit all three auto-correlation curves (Fig. 3.7a) and

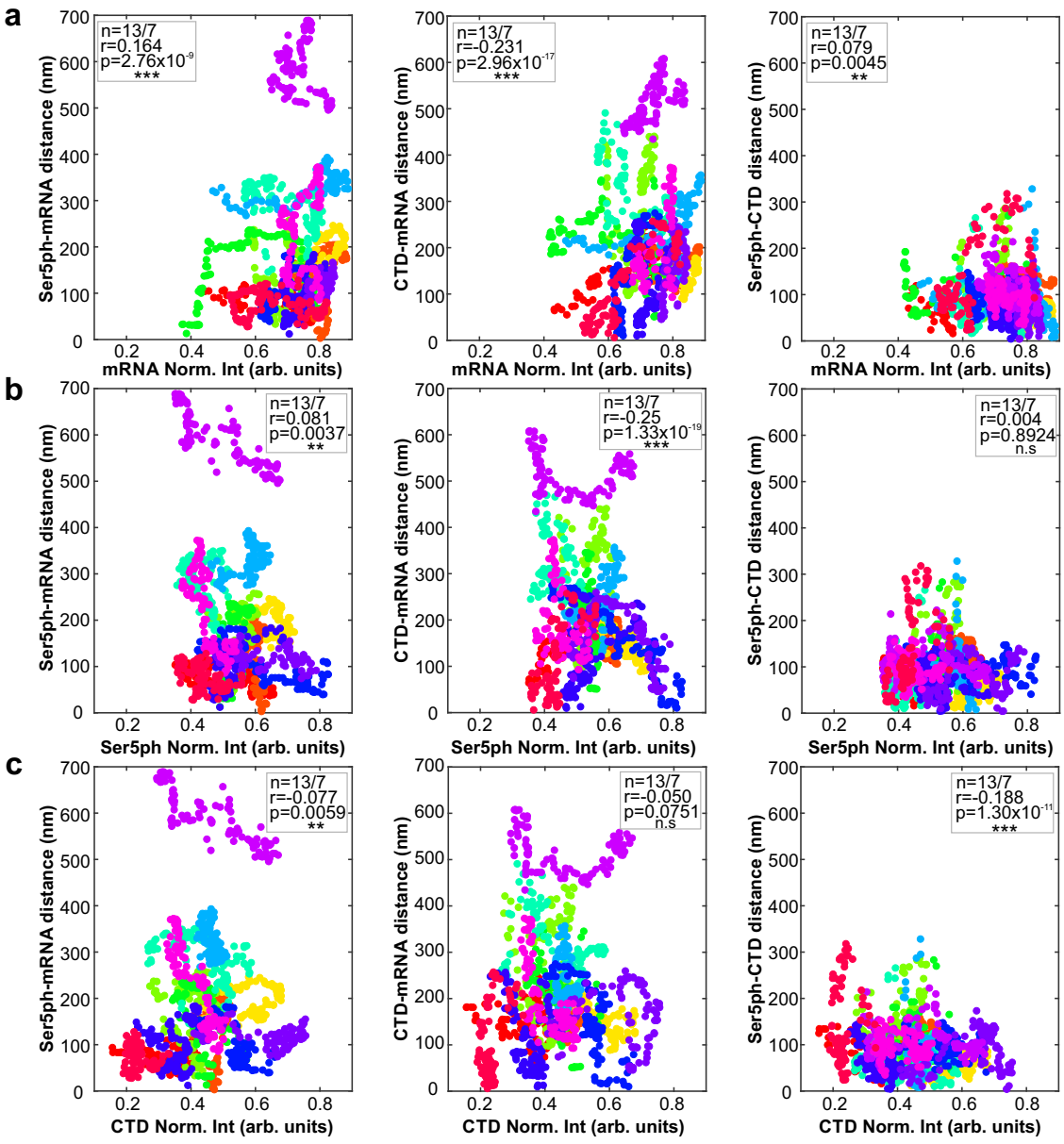


Figure 3.6: Caption on next page.

Figure 3.6: Euclidean distances distribution versus mRNA, Ser5ph-RNAP2, and CTD-RNAP2 normalized intensities. Euclidean distance between Ser5ph-RNAP2 and mRNA (left panel), CTD-RNAP2 and mRNA (middle panel), and Ser5ph-RNAP2 and CTD-RNAP2 (right panel) versus the normalized intensities of **(a)** mRNA, **(b)** Ser5ph-RNAP2, and **(c)** CTD-RNAP2 for all the cells analyzed. Each cell corresponds to one color. n=number of cells/number of independent experiments. Correlation coefficient (r) and p-values (p) as, $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***) were calculated using the "corrcoef" function in MATLAB.

all three cross-correlation curves (Fig. 3.7b). To further constrain the model, we also counted mRNA at transcription sites by comparing their intensities to single mature mRNAs using FISH-quant [149] (Fig. 3.7d, bottom). Consistent with an earlier report [138], we found the HIV-1 reporter contained an average of μ = 15.5 mRNA with a relatively large standard deviation of σ = 10.55 and Fano Factor of σ^2/μ = 7.1.

To unify our data, we posed several models with different levels of complexity (Supplementary Fig. 3.8). Each model considered a promoter with bursty expression. This was represented by specifying distinct active (ON) and inactive (OFF) promoter states with OFF-to-ON and ON-to-OFF transitions rates k_{on} and k_{off}, respectively. When the promoter is ON, RNAP2 is recruited at a rate k_{r} [40]. Upon fitting these models to our data, the fitted burst duration was much shorter than the 1-min experimental sampling time (i.e., $k_{\text{off}} \ll 1$ min). This allowed us to simplify the model to one with burst frequency $\omega = 1/(1/k_{\text{on}} + 1/k_{\text{off}}) \approx 1/k_{\text{on}}$ and geometrically distributed bursts with average size $\beta = k_{\text{r}}/k_{\text{off}}$ [41]. In all models that fit our data, RNAP2 could unsuccessfully depart the promoter at rate k_{ab} or escape at rate k_{esc}. After the escape, the RNAP2 would complete transcription at a combined rate kc that includes both elongation and processing.

In the minimal model that matched all data, CTD-RNAP2 were immediately phosphorylated upon arrival at the promoter, which was consistent with the rapid ($\ll$ 1 min) Serine 5 phosphorylation we observed (Supplementary Fig. 3.14). We also explored several more complicated models with separate steps for initiation, elongation, and processing, post-transcriptional mRNA retention [150], or with separate events describing Serine phosphorylation/initiation and de-phosphorylation/abortion (Supplementary Fig. 3.8). Each model was fit separately to maximize the likelihood for all observed data, but the inclusion of additional mechanisms and free parameters

provided only marginal improvements to the overall fit and resulted in much larger parameter uncertainties. Therefore, we used the Bayesian Information Criteria (BIC) to select our final model as the best choice given our available data (See tables in Supplementary Fig. 3.8). By simultaneously fitting all six correlation plots (Fig. 3.7a, b) as well as the nascent mRNA means and variances, we could estimate the best model's five parameters with excellent precision (Supplementary Fig. 3.9).

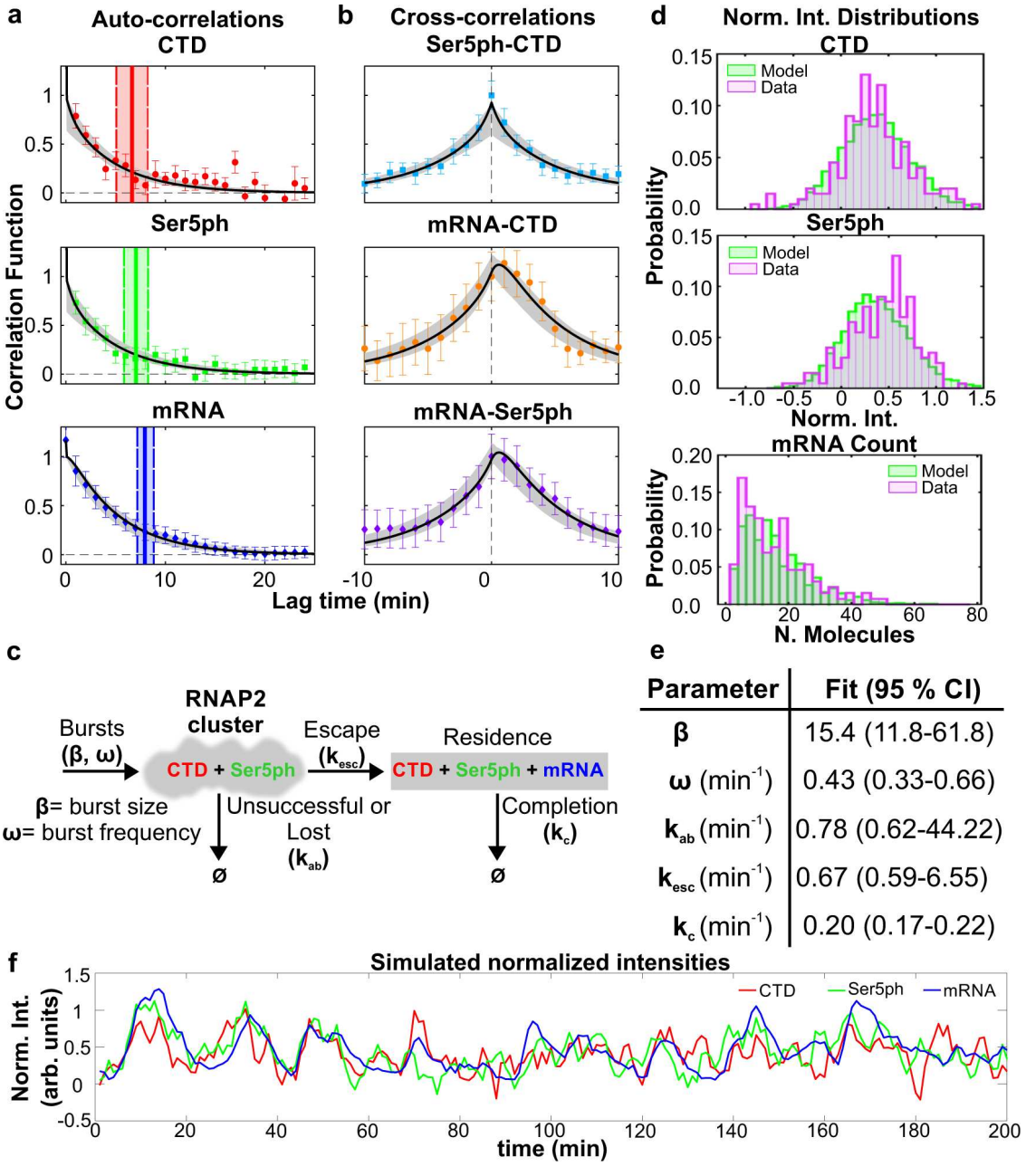
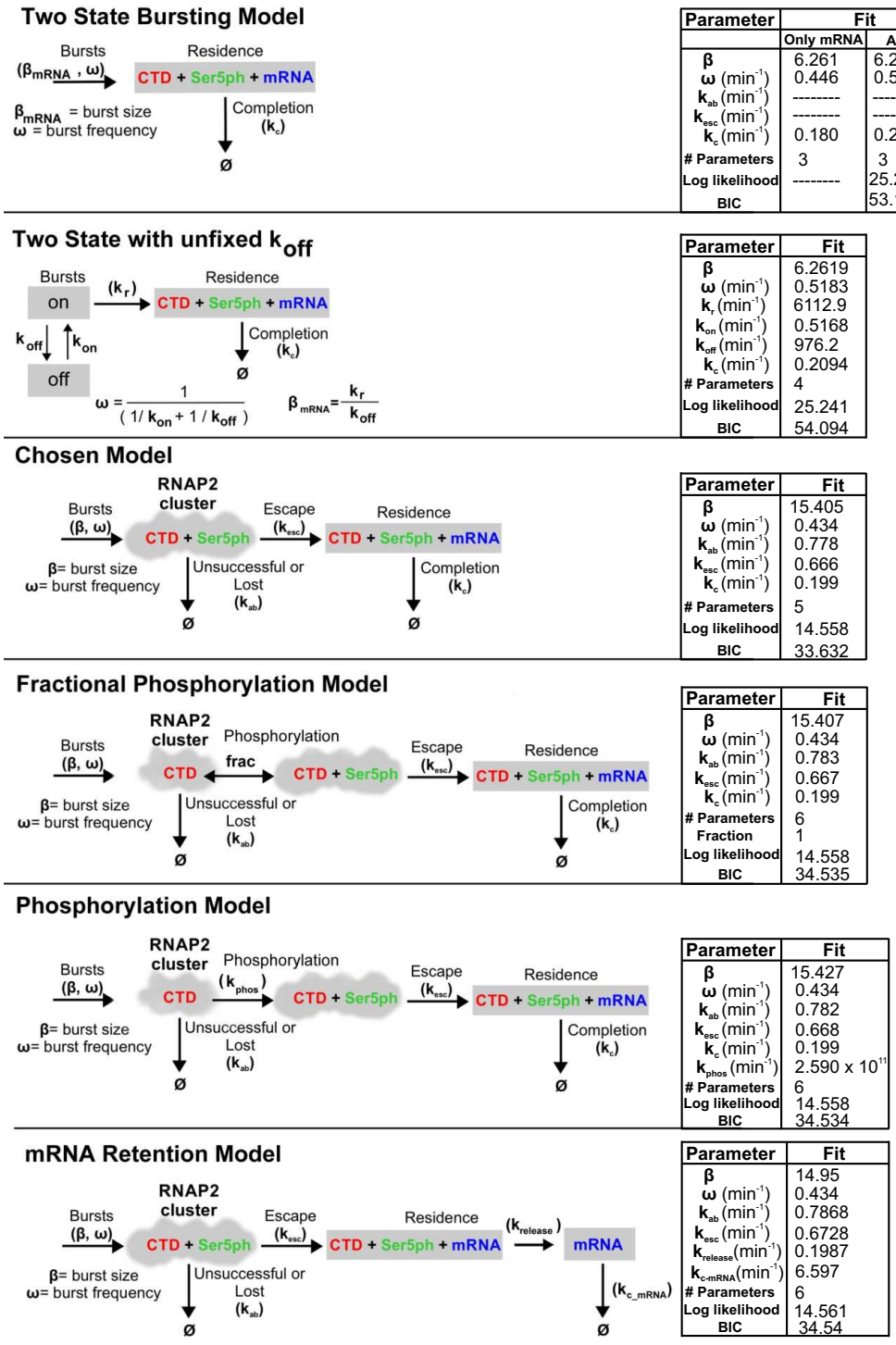


Figure 3.7: Caption on next page.

Figure 3.7: Fluorescence auto- and cross-correlations at the HIV-1 reporter gene are well fit by a unifying model of transcription. A, B) Measured and modeled (A) auto-correlation functions $AC(\tau)/G\bar{(}0)$ for each signal: CTD-RNAP2 (red circles), Ser5ph-RNAP2 (green squares), and mRNA (blue diamonds). Dwell time is defined as the time at which the autocovariance dropped below 20% of its zero-lag value (vertical full lines). Dwell time uncertainty is estimated from the model using the standard deviation from 400 simulated data sets, each with 20 cells over 200 min with 1 min simulation resolution (vertical dashed lines). B) Cross-correlation function $CC(\tau)/G\bar{(}0)$ between signal pairs: Ser5ph-RNAP2 and CTD-RNAP2 (cyan squares), mRNA and CTD-RNAP2 (orange circles), and mRNA and Ser5ph-RNAP2 (purple diamonds) at the transcription site. Model Maximum Likelihood Estimate (MLE) fit in black and sampled uncertainty in gray. C) A simple model to capture RNAP2 fluctuation dynamics at the HIV-1 reporter gene. RNAP2 enters the transcription cluster with an average geometric burst with average burst size, β, and burst frequency, ω. Phosphorylation of Serine 5 is assumed to be fast ($\ll$1 min) and/or the RNAP2 enters in a pre-phosphorylated form. RNAP2 can be lost from the cluster with rate $k_{\rm ab}$ or escape with rate $k_{\rm esc}$. RNAP2 completes transcription with rate $k_{\rm c}$. D) Probability distributions for CTD-RNAP2 and Ser5ph-RNAP2 (arbitrary units of fluorescence), and mRNA (units of mature mRNA) for experimental data (purple) and model MLE predictions (green). E) MLE parameters and 95% confidence interval (CI) range. Statistics presented for the data are the sample means $+-$ S.E.M. n=number of cells/number of independent experiments (20/8). F) Simulated trajectory (with shot noise equal to that of experiments) of CTD-RNAP2 (red), Ser5ph-RNAP2 (green), and mRNA (blue) intensities normalized to have a 95 percentile of unity.

The best-fit parameter values and their uncertainties are provided in Fig. 3.7e. According to the best fit, bursts of RNAP2 occur on average every 1/$\omega \approx$ 2.3 min and have an average size of about $\beta \approx$ 15 molecules per burst. Of the RNAP2 that arrive at the promoter, a substantial fraction f = $k_{\rm esc}$/($k_{\rm esc}$ + $k_{\rm ab}$) $\approx$ 0.46 escape the promoter and complete transcription, leading to convoys20 of about $f \cdot \beta \; approx$ 7 RNAP2 per burst. Each mRNA takes an average of 1/$k_{\rm c} \approx$ 5 min to complete elongation and processing, meaning that on average the HIV-1 reporter contains mRNA originating from $\omega/k_{\rm c} \approx$ 2 consecutive bursts. Overall, the model predicts that there is an average of $\sim$20 RNAP2 on the gene in steady-state, with an average of $\sim$5 in the cluster near the promoter in an unphosphorylated or Ser5ph form, and $\sim$15 elongating or processing near the end of the gene (see Table 1). This average picture is somewhat misleading, however, as the number of RNAP2 within the cluster fluctuates dramatically due to randomly timed bursts. According to our simulations, there are periods when as many as $\sim$90 RNAP2 come in at a time interspersed by brief and random silent periods of low RNAP2 occupancy (Supplementary Fig. 3.12a).

After fitting the model to capture the auto- and cross-correlation functions and the mean and variance of the mRNA distribution, we verified that it also correctly predicted the full probabil-



Parameter	Fit	
	Only mRNA	All
β	6.261	6.261
ω (min^{-1})	0.446	0.519
k_{ab} (min^{-1})	--------	--------
k_{esc} (min^{-1})	--------	--------
k_c (min^{-1})	0.180	0.209
# Parameters	3	3
Log likelihood	--------	25.242
BIC		53.192

Parameter	Fit
β	6.2619
ω (min^{-1})	0.5183
k_r (min^{-1})	6112.9
k_{on} (min^{-1})	0.5168
k_{off} (min^{-1})	976.2
k_c (min^{-1})	0.2094
# Parameters	4
Log likelihood	25.241
BIC	54.094

Parameter	Fit
β	15.405
ω (min^{-1})	0.434
k_{ab} (min^{-1})	0.778
k_{esc} (min^{-1})	0.666
k_c (min^{-1})	0.199
# Parameters	5
Log likelihood	14.558
BIC	33.632

Parameter	Fit
β	15.407
ω (min^{-1})	0.434
k_{ab} (min^{-1})	0.783
k_{esc} (min^{-1})	0.667
k_c (min^{-1})	0.199
# Parameters	6
Fraction	1
Log likelihood	14.558
BIC	34.535

Parameter	Fit
β	15.427
ω (min^{-1})	0.434
k_{ab} (min^{-1})	0.782
k_{esc} (min^{-1})	0.668
k_c (min^{-1})	0.199
k_{phos} (min^{-1})	2.590×10^{11}
# Parameters	6
Log likelihood	14.558
BIC	34.534

Parameter	Fit
β	14.95
ω (min^{-1})	0.434
k_{ab} (min^{-1})	0.7868
k_{esc} (min^{-1})	0.6728
$k_{release}$ (min^{-1})	0.1987
$k_{c\text{-}mRNA}$ (min^{-1})	6.597
# Parameters	6
Log likelihood	14.561
BIC	34.54

Figure 3.8: Transcription models considered for model selection Right column table shows MLE fit parameters for each model.

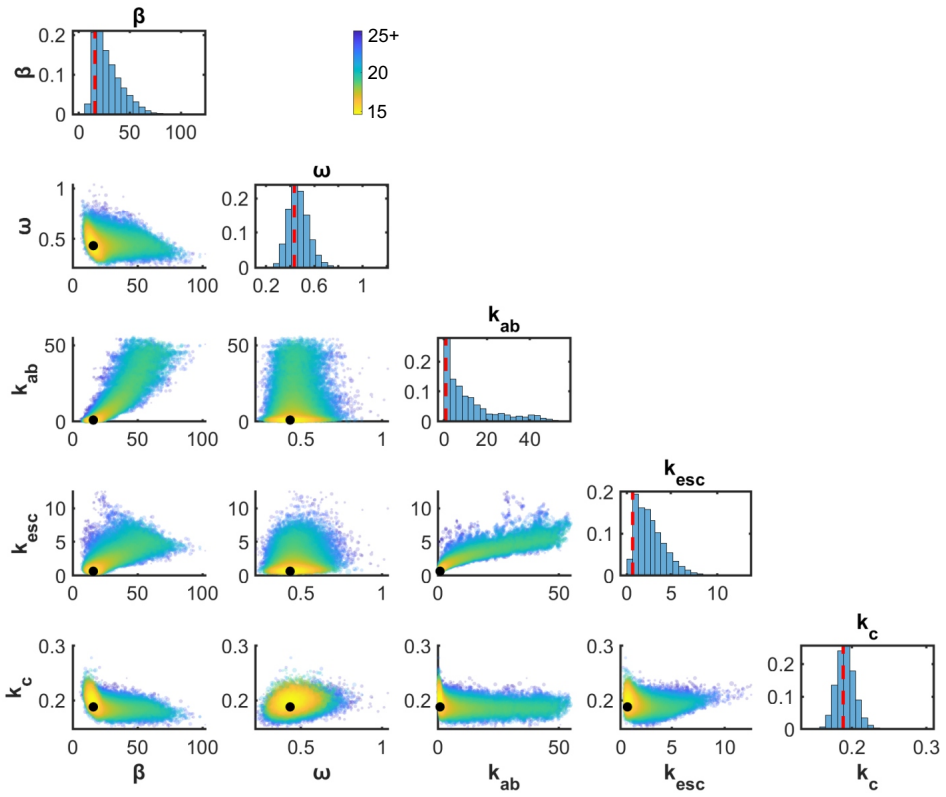


Figure 3.9: Parameter sensitivity analysis. Metropolis-Hastings algorithm was run to determine posterior uncertainty of model parameters given the experimental data. Plots on the diagonal show the marginal posterior parameter distributions for each parameter (MLE parameter estimate denoted by red dashed line) and off-diagonal plots show the joint posterior parameter distributions for all pairs of parameters (MLE parameter combination denoted by black marker. Colors denote log-likelihood value; an upper bound of the 0.5% highest log-likelihoods was selected for coloring purposes. Any log-likelihood's color above this threshold was set to that bound). A proposal distribution of a 5 dimensional Gaussian with a standard deviation of 3% of MLE parameters was used. 20 individual chains of 250000 with a thinning rate of 20 (100 million) were used to generate the posterior distributions. 40000 points of the posterior are displayed in the figure.

ity distributions for the number of nascent mRNA molecules and RNAP2 signal intensities at the HIV-1 transcription site (Fig. 3.7d). We also simulated normalized intensities including shot noise (Fig. 3.7f), and these look similar to our measured trajectories (Fig. 3.5c). The shot noise was estimated directly from the experiments by comparing the observed zero-lag covariance $G(0)$ compared to an estimate for the zero-lag autocovariance found by interpolation from the short, but nonzero time lags. These shot noise standard deviations were found to be 1.98×, 1.42×, and 0.41×

of the standard deviation for CTD, Ser5ph, and MS2 signals, respectively. Finally, we simulated ChIP data for our single-gene reporter (Supplementary Fig. 3.12b, c). To do this, we assumed an elongation rate of 4.1 kb/min (measured previously at this locus by analyzing the MS2 stochastic fluorescence fluctuations [138]) and processing rate of 0.27 min^1 (so elongation and processing times sum to our fitted $1/k_c$ completion time). With these rates, the CTD/Ser5ph-RNAP2 simulated ChIP signals from active genes displayed strong peaks at the beginning and end of the gene, as we observed in Fig. 3.1b (compare to Supplementary Fig. 3.12b, c). Overall, the excellent match between data and simulations indicates our best-fit model faithfully captures transcription dynamics at the HIV-1 reporter. To facilitate further exploration of our model, we provide a graphical user interface (GUI) [https://doi.org/10.5281/zenodo.4631141]. The GUI allows exploration of how each model parameter affects model predictions, including trajectories, auto- and cross-correlations, distributions of spot intensities, simulated ChIP data, and several derived quantities to describe the CTD-RNAP2, Ser5ph-RNAP2, and mRNA burst dynamics (Supplementary Fig. 3.10).

Additional computational modeling details[2]

Six total models were considered for model selection. The first two models were selected to be negative controls and simpler than our favored model. If model selection selected the two simpler models, we could not be confident in our chosen model to have predictive power on its added mechanisms vs the simpler bursting models. First two simpler models are shown in Fig. 3.8, each does not distinguish between Ser5 phosphorylation state, that is to say, polymerases arrive already phosphorylated on Ser5. For more complicated models, we were interested in three phenomena: Do RNAP2s dwell within the model for some time repeatedly gaining and losing Ser5ph, constantly initiating and aborting near the promoter? Do RNAP2s arrive unphosphorylated and then gain Ser5? with some time delay? Were mRNAs residing with a residence time post transcriptionally for some processing? Each of these questions corresponds to the fractional phosphorylation

[2]The following subsection is added from supplemental information and expanded upon from the original publication.

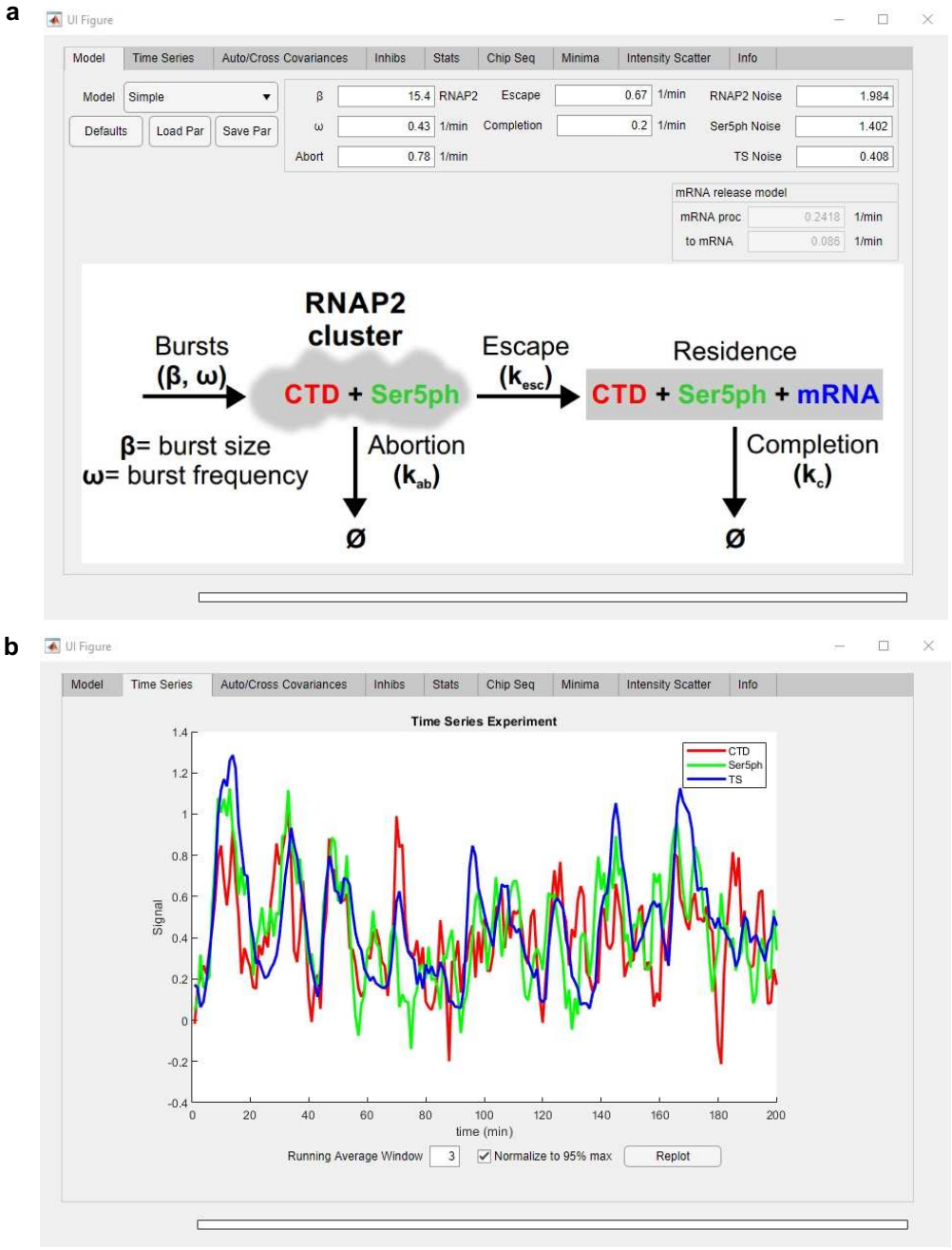


Figure 3.10: Graphical User Interface (GUI) for the transcription model. To facilitate the simulation of transcription dynamics at a single-copy gene, the model described in the main text has been incorporated into a MATLAB toolbox. **(a)** This graphical user interface (GUI) is divided into eight upper tabs, and input boxes for specification kinetic parameters. The GUI allows the simulation of intensity trajectories in each channel. **(b)** Sample display of simulated intensities normalized to the 95th percentile and running averaged with a window of three time points. The GUI also allows for display of auto-, cross-correlations, predicted minima from the experimental data previously loaded, prediction of ChIP distributions, and perturbed intensity trajectories by blocking different steps of transcription in the model.

model, the phosphorylation model, and the mRNA retention model shown in Fig. 3.8. The simpler models were readily discarded by the Bayes Information Criterion (BIC), with almost double the value of the chosen model's MLE parameters. The more complicated models had only slightly worse BICs, indicating that there was not enough evidence in the data to support these models but only just barely. This indicated that some of these mechanisms may be present but more data would need to be collected; We thought it likely that there is a rate for RNAP2 to get Ser5ph, but that its timescale was much faster than the data collection timescale of a minute. Ultimately this informed us to perform the fast imaging experiment in Fig. 3.14 which revealed a likely delay of 3-6 seconds to obtain Ser5ph making the phosphorylation model more likely. However, this new data wasn't used for a new round of model selection but can be performed in the future.

Our model also suggested two future experiments: Collecting a Larger perturbation could help verify our model predictions and allow us to validate our model further than the cursory empirical comparison we did above in Fig. 3.13, and performing ChIP experiments to validate our model's predictions in Fig. 3.12

Inhibiting distinct steps of the transcription cycle provides further evidence for the spatiotemporal organization of RNAP2 phosphorylation

So far, our collective data and modeling suggest a precise temporal ordering of transcription dynamics, beginning with the recruitment of CTD-RNAP2, followed by rapid initiation in 3-6 s (indicated by Ser5ph-RNAP2), and promoter escape and elongation within another minute or so (indicated by mRNA). Our data also provide evidence of heterogeneity in the distribution of RNAP2 along the gene, with high concentrations near the beginning and end of the gene (Fig. 3.1b). To further test our system, we perturbed it by adding three different transcription inhibitors: Triptolide (TPL), THZ1, and Flavopiridol (Flav) (Fig. 3.11). We began by inhibiting the earliest steps in the transcription cycle to attempt to prevent the formation of the RNAP2 cluster. To achieve this we added TPL, a small-molecule inhibitor that prevents promoter DNA opening and transcription initiation by inhibiting the DNA-dependent ATPase activity of the XPB subunit of TFIIH [124, 151]. TPL has also been shown to induce RNAP2 degradation on the hours timescale [152], so we im-

aged for just 30 consecutive minutes to focus on the more immediate impact of TFIIH inhibition. The addition of 5 μM TPL led to a rapid and dramatic loss of both mRNA and all RNAP2 signals at the TS within just $\sim$10 min (Fig. 3.11a-d, and Supplementary Movie 2 (available online)). Consistent with our previous findings, we observed a temporal ordering in the TPL-induced run-off of RNAP2 (Fig. 3.11c), with CTD-RNAP2 signals dropping earlier than Ser5ph-RNAP2, followed by mRNA. This ordering was observed in seven out of ten single cells we measured. Of these, four exhibited clear separation between the three traces (inset in Fig. 3.11c). Since steps that are later in the CTD cycle necessarily take longer to respond to drugs, this ordering provides further evidence that CTD-RNAP2 slightly precedes Ser5ph-RNAP2 by less than a minute, and that both RNAP2 signals come significantly earlier than mRNA. These data also demonstrate that the opening of promoter DNA by XPB is a requirement for the formation of RNAP2 clusters. This can work by at least two mechanisms: (1) All the Ser5ph-RNAP2 underwent initiation and abortion, but RNAP2 kept its Serine 5 phosphorylation; (2) Initiation of the first RNAP2 activates CDK7, which can phosphorylate many RNAP2 within the cluster.

We next used THZ1, which inhibits RNAP2 CTD phosphorylation at Serine 5 by targeting the TFIIH kinase CDK7, thereby preventing promoter pausing, mRNA capping, and productive elongation [124, 137, 153]. In contrast to TPL, THZ1 has a slower action, so a higher concentration and longer exposure to this drug were needed to see an effect in real-time. Treatment with 15 μM THZ1 led to a reduction in the mRNA signal at the HIV-1 reporter within 25 min (Fig. 3.11e). Likewise, both CTD-RNAP2 and Ser5ph-RNAP2 levels were on average reduced. Interestingly, in some single cells, we observed large, temporally ordered bursts in the levels of CTD-RNAP2 and Ser5ph-RNAP2, despite continued inhibition and overall loss of mRNA. These large bursts could even achieve RNAP2 levels that were as high as pretreatment levels (the thicker black curve in Fig. 3.11e highlights one example). Presumably, these bursts occur because there is residual TFIIH left in the cell that are not yet inhibited by THZ1, or because recently aborted RNAP2 retain their Ser5ph within the cluster. Since mRNA levels did not burst to the same degree, we conclude the bursts arise from clusters of RNAP2 near the promoter that initiate but fail to escape.

These transient clusters near the beginning of the gene are consistent with the high concentration of RNAP2 near the promoter we observed by ChIP (Fig. 3.1b) and are also consistent with the ChIP predictions of our best-fit model (Supplementary Fig. 3.12b, c).

We next blocked a later step in the transcription cycle using 1 μM Flav, a drug that prevents transcription elongation and RNAP2 CTD phosphorylation at Serine 2 by inhibiting the CDK9 activity of P-TEFb [137, 154]. Like THZ1, Flav also reduced the intensity of the mRNA signal, this time within $\sim$15 min (Fig. 3.11f). However, CTD-RNAP2 and Ser5ph-RNAP2 signals remained relatively unchanged, exhibiting large fluctuations and a slight overall reduction on average. This difference from THZ1 can be attributed to the later action of Flav in the transcription cycle. The high levels of CTD-RNAP2 and Ser5ph-RNAP2 signals that remained post-Flav again support a dynamic clustering model [124, 125, 129, 127, 128] in which most RNAP2 are already phosphorylated at Serine 5 and presumably make repeated attempts at initiation and promoter escape.

Finally, we attempted to qualitatively recapitulate these perturbations using our best-fit model. To do so, we evaluated several hypothetical mechanisms in which transcription is inhibited by reducing one or more of the rates, including burst statistics (ω or β), the promoter escape rate k_{esc}, or the completion rate k_{c}. According to simulations, inhibiting earlier steps (ω or β) in the transcription cycle led to the sequential loss of all RNAP2 and mRNA signals at the transcription site at a rate governed by the time scale of mRNA elongation and processing (Supplementary Fig. 3.13a), reminiscent of our TPL experiments. In contrast, inhibiting a later step (k_{esc}), led to a retention of large numbers of RNAP2 in the cluster that undergoes relatively large and rapidly changing fluctuations (Supplementary Fig. 3.13b), reminiscent of our THZ1 experiments. Blocking (k_{esc}) and reducing k_{c} by 30% led to a slight reduction in the mRNA signal and even less decrease in the RNAP2 signals with relatively large fluctuations (Supplementary Fig. 3.13c), reminiscent of our Flav experiments. We also blocked bursts (either ω or β) and reduced k_{c} by 30% and obtained an overall reduction of all the signals (Supplementary Fig. 3.13d) that do not represent any of the inhibitors tested here. The similarity between these simulations and our experimental perturbations

provides further support for our model and also provides evidence that the tested inhibitors act on distinct stages of the RNAP2 transcription cycle.

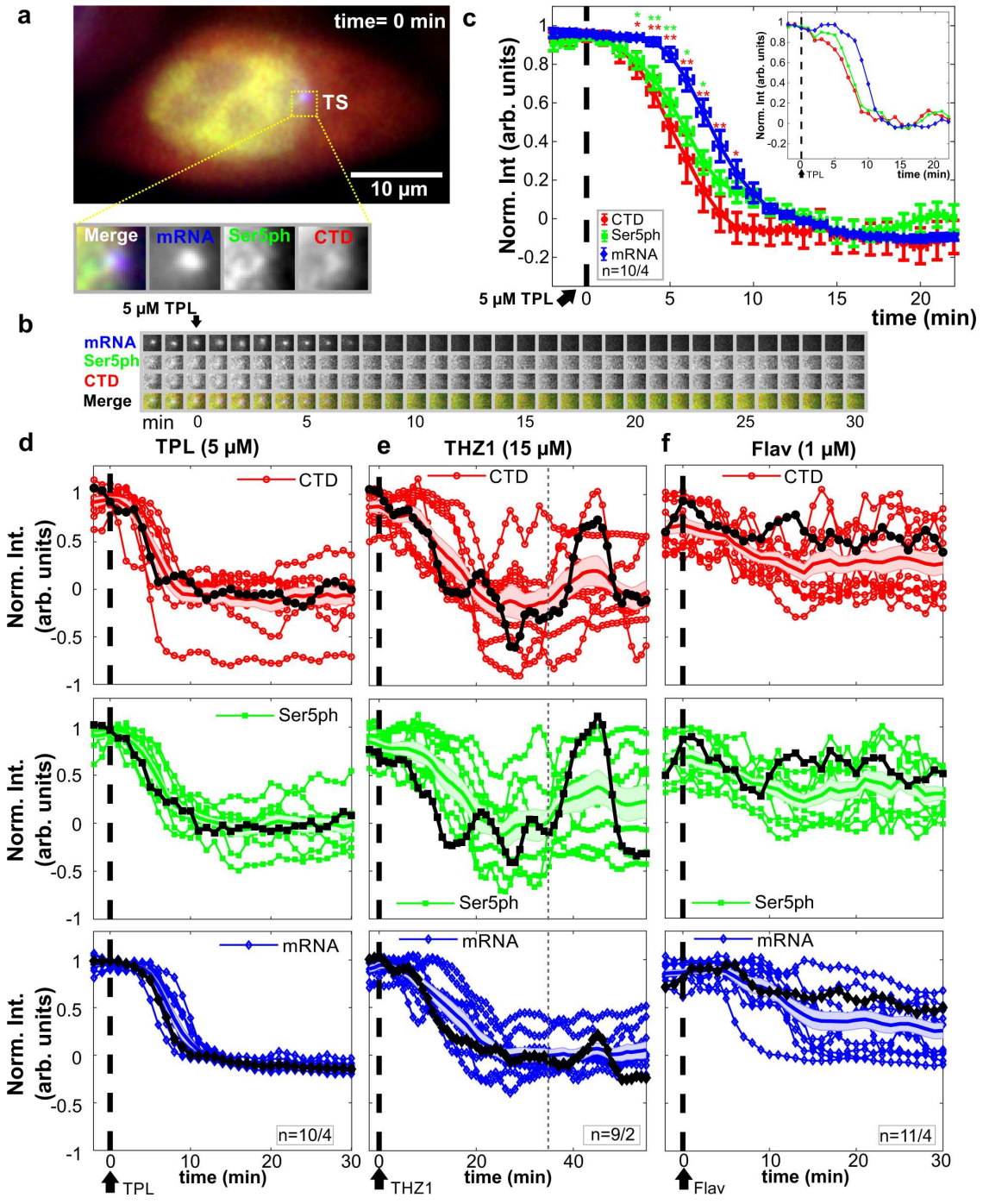


Figure 3.11: Caption on next page.

Figure 3.11: Intensity fluctuations of CTD-RNAP2, Ser5ph-RNAP2, and mRNA in the presence of transcription inhibitors. A) Sample cell before the addition of Triptolide (TPL). The transcription site (TS) is shown in the dotted box, and the inset shows a zoom-in (n = 10/4). B) The TS from (A) at all times before and after the addition of TPL. C) Normalized average TS intensity over time of all the quantified cells for CTD-RNAP2 (red circles), Ser5ph-RNAP2 (green squares), and mRNA (blue diamonds) before and after application of TPL (vertical dashed line). The inset shows the signals in a representative cell. Data are presented as mean values +− S.E.M. Significance was tested using a two-tailed Mann-Whitney U test with $p \leq 0.05$ (*) and $p \leq 0.01$ (**). D-F) Normalized intensity signals after application of various transcription inhibitors, including d TPL (5 μM), E) THZ1 (15 μM), and F Flavopiridol (Flav, 1 μM) for all the cells analyzed. Signals highlighted in black correspond to a sample single cell, the colored shadow and the full line in the middle of it correspond to the S.E.M. and the mean in each channel, respectively. The vertical gray dashed line in (E) highlights the time point at which a burst of RNAP2 is observed in the sample cell in the presence of THZ1. n = number of cells/number of independent experiments.

3.1.4 Conclusion

In this study, we measured the dynamics of the RNAP2 CTD transcription cycle at the single-gene level in living cells. By combining complementary antibody-based imaging probes with multi-color single-molecule microscopy and computational modeling, we were able to detect organization in both the temporal ordering and spatial distribution of endogenous RNAP2 phosphorylation along a single HIV-1 reporter gene.

We find that a large number of RNAP2 at the HIV-1 transcription site are clustered around the promoter in a region that is spatially distinct from elongating RNAP2 and mRNA synthesis (as depicted in Fig. 3.15). This spatial organization supports the notion of dynamic RNAP2 clusters that form transcriptional hubs [155] or factories [156, 157] that contain high concentrations of the transcription machinery. In steady-state, we estimate there is an average of ∼20 RNAP2 at the HIV-1 gene. This total number of RNAP2 is in between recent estimates of ∼80 RNAP26 clustered at the constitutively expressed beta-actin locus, ∼17 RNAP2 at an exogenous mini-gene [136], and ∼7.5 RNAP2 at the Pou5f1 locus [136]. Of the ∼20 RNAP2 at our HIV-1 reporter gene, we estimate on average ∼5 are at or near the promoter, awaiting initiation or promoter escape. During frequent bursts, however, this number can dramatically increase to as high as 90 RNAP2, with most either coming in with Serine 5 phosphorylation or rapidly acquiring Serine 5 phosphorylation within seconds (Supplementary Fig. 3.14). Given the limited amount of space at the promoter, it

is hard to imagine all of these RNAP2 are promoter bound. Instead, we believe many are unbound and collectively this fraction helps form the transcription cluster, which remains spatially distinct from mRNA synthesis.

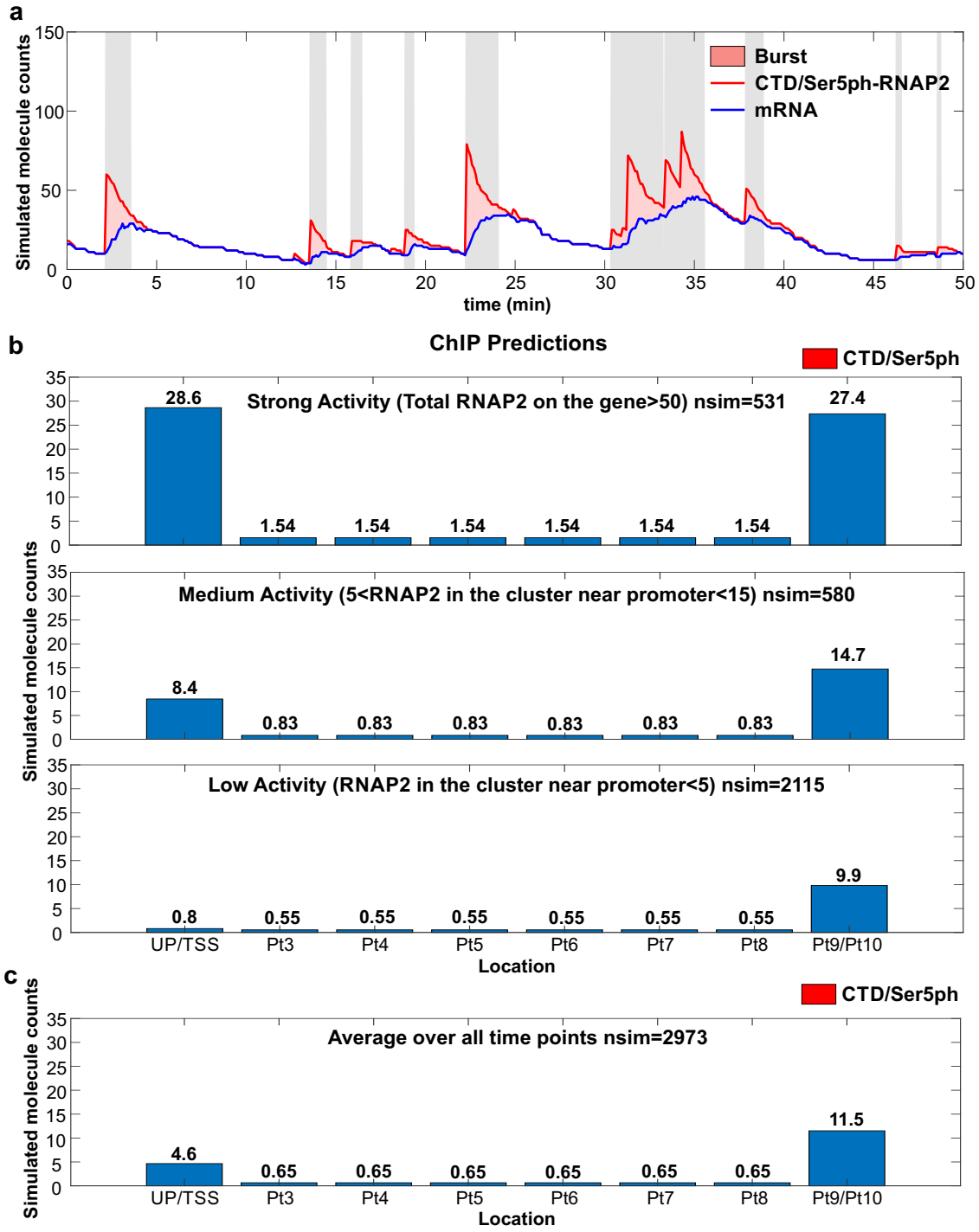


Figure 3.12: Caption on next page.

Figure 3.12: Simulated trajectories and ChIP predictions (a) Stochastic simulation for the number of nascent mRNA per transcription site (blue line), total number of RNAP2 at transcription site (red line), and number of RNAP2 in cluster near transcription site promoter (red shading). Periods with $\geq$ 10 RNAP2 at the transcription site cluster (gray shading) are classified as 'ON' (14.0% of total time); periods with no RNAP2 at the cluster are classified as 'OFF' (42.9% of time); and periods with intermediate levels of RNAP2 in the cluster are classified as 'transient' (43.1% of time). Note that for clarity these simulations do no include the experimental shot noise used to simulate actual measurements (as in Fig. 3.7f, for example). (b) Simulated ChIP data as predicted using the model for: (Top; Strong Activity) average spot during an ON period; (Middle; Medium Activity) average spot during a transient period; and (Bottom; Low Activity) average spot during an OFF period. Each stochastic simulation was run for 120,000 min and sampled at 40 min intervals to ensure de-correlated points. To estimate RNAP2 loading at the inner bins, an elongation rate of 4.1 k_{b}/min was assumed and used to get the fraction of time spent elongating versus processing of the total RNAP2 residence time. This fraction of elongation time was then distributed from the final bin uniformly to the middle bins and is represented by the middle numbers of bins Pt3-8. (c) Average simulated RNAP2 ChIP over all times including all ON, OFF and transient periods.

A major unresolved question is how RNAP2 is retained in clusters. One possibility is that RNAP2 is trapped by repeated interactions with other transcription machinery in the region. Alternatively, clusters could represent phosphorylation-dependent condensates. As others have recently shown, phase separation can be driven by phosphorylation of the unstructured RNAP2 CTD [129] and by the histidine-rich tail of P-TEFb [127]. Since Tat directly interacts with P-TEFb [144, 158], it could enhance RNAP2 recruitment and clustering at the HIV-1 reporter gene.

One possible advantage of the cluster is it retains recently aborted RNAP2 near the transcription start site so they can rapidly re-initiate. This follows from our rapid imaging experiments, which indicate initiation is very rapid (3-6 s; Supplementary Fig. 3.14) compared to promoter escape (fitted $1/k_{\mathrm{esc}}$ 1.5 min). The distinct timescales imply two hypotheses: First, most promoter escape attempts fail. This is consistent with earlier measurements based on FRAP that demonstrated successful promoter escape is a rare event [137, 159]. Second, a large fraction of RNAP2 in the cluster are inactive at any given time [125, 160]. Such a large fraction of inactive RNAP2 could arise from recently aborted molecules that retain their Ser5ph. Evidence for the retention of Ser5ph on RNAP2 after transcription abortion was seen in an earlier study [137], where Ser5ph-RNAP2 was detected in the soluble fraction of cells after transcription was globally inhibited via flavopiridol. The retention of RNAP2 also helps explain our model prediction that nearly half

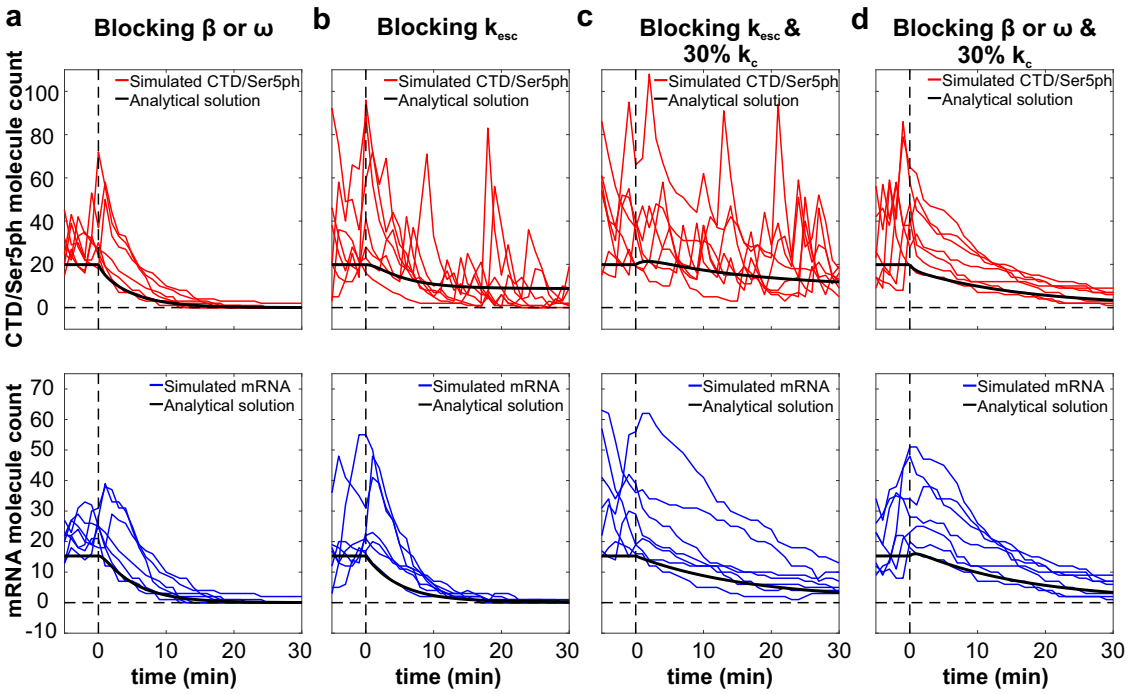


Figure 3.13: Predicted CTD/Ser5ph-RNAP2, and mRNA signals after perturbing different steps in the mathematical model. Simulated molecule counts for CTD/Ser5ph-RNAP2 (red, upper panels), and mRNA (blue, bottom panels) after blocking: (a) β or ω, (b) k_{esc}, (c) k_{esc} and 30% k_{c}, and (d) β or ω and 30% k_{c}, and their respective analytical solution in each plot (black curve). Simulated trajectories with mRNA molecule counts above the analytical solution at time of inhibition are shown with colored lines. This was done to simulate the experimental procedure of choosing transcription sites at the beginning of an experiment where all three signals could be seen. Blocking is defined as multiplying the best fit parameter by 0.01 (99% reduction), similarly blocking 30% refers to multiplying the best fit parameter by 0.3 (70% reduction). For blocking β and ω, k_{off} was defined by setting k_{off} to 1000, effectively turning off bursting dynamics.)

of the RNAP2 in the cluster ($k_{\text{esc}}/(k_{\text{esc}} + k_{\text{ab}})$ ~46%) eventually does escape the promoter and produce a full-length transcript. Thus, local recycling of transcription machinery within clusters may play a role in HIV-1 biogenesis, where Tat expression provides a positive feedback loop to amplify transcription and facilitate the rapid production of viral proteins in host cells [161].

While the overall efficiency of transcription is relatively high at the HIV-1 reporter gene compared to other genes studied, the various kinetic rates we quantified are fairly consistent with earlier work. In particular, we found RNAP2 takes around five minutes to complete transcription after promoter escape ($1/_{\text{k}}c$ in Fig. 3.7). This places an upper bound on the RNAP2 elongation and processing time. If we constrain the elongation rate to be 4.1 kb/min [138] (~1 min for the

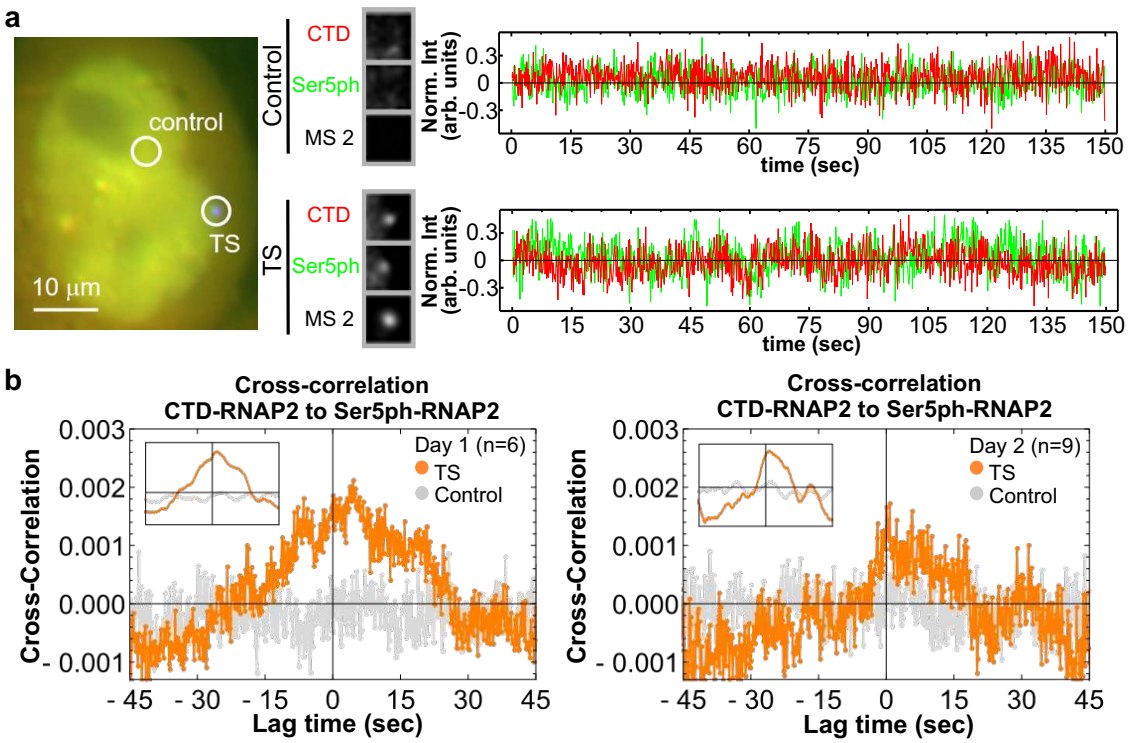


Figure 3.14: Fast-imaging experiments revealed a 3-6 sec time delay between CTD-RNAP2 and Ser5ph-RNAP2. (a) (Left) Exemplary cell for fast imaging (150 msec/frame) for a total of 1000 time points (150 sec) in a single plane. Two positions are highlighted: the HIV-1 TS and a control nonspecific spot. (Center) Crops showing the CTD- (red), Ser5ph-RNAP2 (green), and MS2 mRNA (black) signals at the TS and control positions within the exemplary cell. (Right) Normalized intensity at the TS (bottom) and control (top) positions over time from the exemplary cell for CTD-RNAP2 (red), Ser5ph-RNAP2 (green). (b) Measured cross-correlation function CC(τ) between CTD-RNAP2 and Ser5ph-RNAP2 at the TS (orange circles) and control (gray circles) positions separated by experimental day. The inset shows a 50-frame rolling average to more easily identify the peak time delay between the two signals. In both cases, the cross-correlation peaks at a lag time of roughly 3-6 seconds.

full gene), then we can assign the remaining time (∼4 min) to RNA processing at the 3' ends. Under these conditions, the model predicts a build-up in RNAP2 at the 3' end of the gene because processing takes longer than elongation. This buildup is consistent with our ChIP data in Fig. 3.1b. The estimated 4 min processing time is also consistent with an earlier estimate at this HIV-1 locus [138], although such relatively long processing may not be representative of other genes. Similarly, the RNAP2 initiation and promoter escape rates we quantified are consistent with earlier reports, taking between a minute and a few minutes [143, 159]. Finally, we also detected bursts in transcription that result in convoys of RNAP2, as previously reported [138], and consistent with widespread bursting observed across the genome [40, 162]. The global agreement

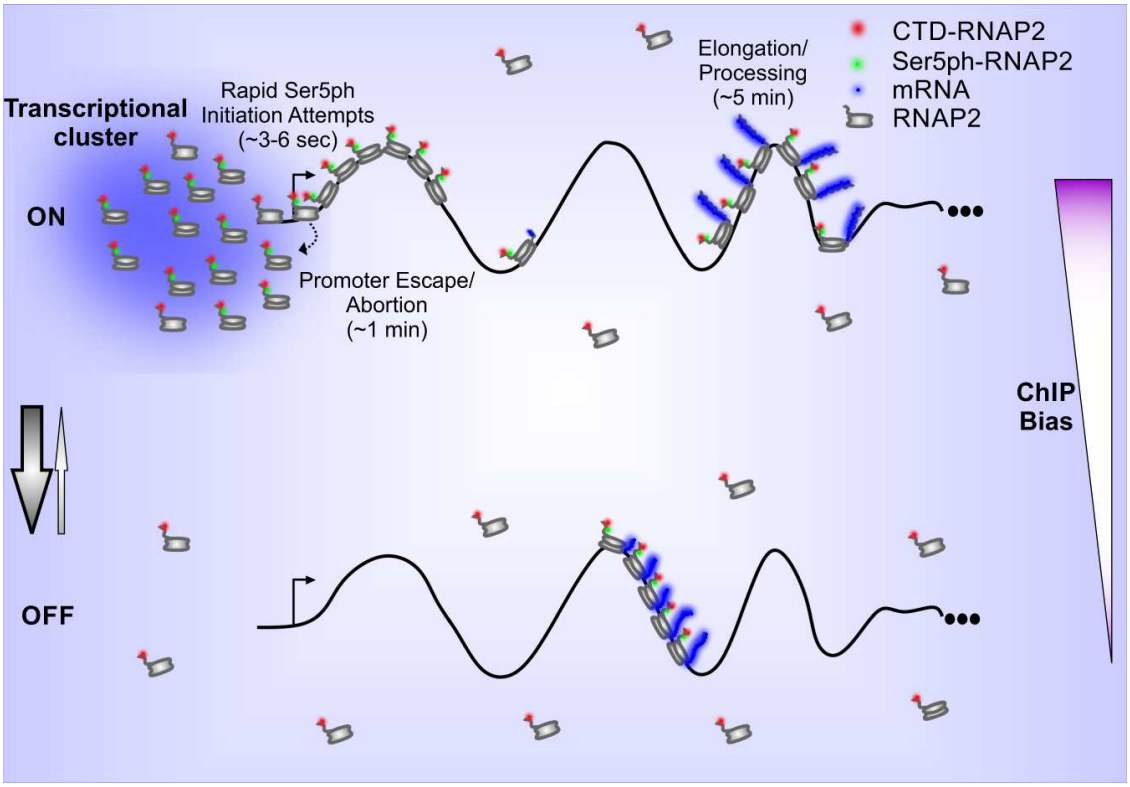


Figure 3.15: Model depicting RNAP2 transcription dynamics in a single-copy gene. During extremely short ($\ll$1 min) periods of the ON state, RNAP2 is recruited in bursts ($\sim$15 RNAP2) to the HIV-1 reporter gene, creating transient ($\sim$1 min) clusters of CTD-RNAP2 and Ser5ph-RNAP2 at the gene promoter, and initiating transcription in RNAP2 convoys ($\sim$7 RNAP2/convoy). The middle of the gene remains mostly empty due to rapid transcription, while a large number of RNAP2 ($\sim$15) concentrate at the end of the gene during processing ($\sim$4-5 min). In OFF periods ($\sim$2.3 min), RNAP2 convoys that escaped the promoter during the ON state quickly elongate and complete transcription. The gene rapidly transitions back to the OFF state when ON (denoted by arrows). ChIP assays enrich for genes with lots of RNAP2, which will bias the assay towards genes with RNAP2 clusters near the promoter.

between studies suggests some convergence in the field, particularly given the uniqueness of our dataset, which is based on fluctuations of both MS221 and RNAP2 Fab signals [143].

The ability to image by fluorescence microscopy endogenous RNAP2 phosphorylation dynamics at single-copy genes now makes it possible to estimate the RNAP2 distributions predicted by ChIP. ChIP studies of the RNAP2 CTD transcription cycle typically display heterogeneous distributions of RNAP2 that have distinct peaks of Ser5ph-RNAP2 near the promoter and Ser2ph-RNAP2 at the ends of genes [121, 123, 131]. However, based on ChIP alone, it is not clear if peaks represent the distribution of RNAP2 along single genes or instead represent a population of genes. For example, it could be that half of the genes have Ser5ph-RNAP2 paused at the beginning of

the gene, while the other half have Ser2ph-RNAP2 being processed near the end of the gene. In this extreme example, no single gene would have RNAP2 at both ends. According to our best-fit model, the situation for HIV-1 is not this extreme, but the distribution of RNAP2 does depend sensitively on the timing of bursts. For example, early in a burst RNAP2 occupancy is heavily front-loaded, with all or nearly all RNAP2 at or around the promoter in a Serine 5 phosphorylated form. Since RNAP2 ChIP by design is biased towards genes with high levels of RNAP2 at the time of assay, genes that have recently burst are likely to be overrepresented in the data (Supplementary Fig. 3.12b, c). As our model demonstrates, soon after a burst, genes tend to have far more RNAP2 clustered around the promoter than the average gene (which has just five) (Fig. 3.15). According to this interpretation, the large Ser5ph-RNAP2 ChIP peak we observe near the promoter could arise from rapid and repeated promoter-proximal initiation and/or pausing. Given the nature of ChIP, it is also possible the peak arises from RNAP2 within clusters that are non-specifically cross-linked during the fixation step. However, this latter possibility seems unlikely as promoter-proximal peaks are also observed using techniques that detect and sequence nascent mRNAs, such as GRO-seq, PRO-seq, and mNET-seq [163]. In the future, it will be interesting to see to what extent dynamic clustering observed in living cells correlates with promoter-proximal RNAP2 peaks observed across the genome in populations of fixed cells [164].

Aside from HIV-1, our technology can now be used to examine RNAP2 phosphorylation dynamics at other single-copy genes. Given the high correlation between MS2 (mRNA) and RNAP2 (Fabs), in the future MS2 may not even be required. For example, by combining Fab and CARGO [165], RNAP2 phosphorylation dynamics at an endogenous gene could be visualized without extensive genome editing. Alternatively, Fab could be combined with other labeling technologies such as lacO/lacI [166, 167], ROLEX [168], ANCHOR [169], or post-fixation via DNA FISH [170] or CasFISH [171]. Beyond RNAP2, post-translational modifications to other proteins involved in transcription could also be studied in this way, including histones [140, 172]. However, a few important caveats of Fab- or intrabody-based imaging should be kept in mind: First, if Fabs bind their targets with too low affinity, then there will be a large unbound fraction that will decrease

signal-to-noise. For the CTD-RNAP2 and Ser5ph-RNAP2 Fab, the bound fraction was determined to be greater than 80% [143]. Second, if Fabs take too long to bind their targets, then very rapid processes can be entirely missed or their timescales will appear erroneously slow. According to FRAP, the vast majority of CTD-RNAP2 and Ser5ph-RNAP Fabs used in this study bind and rebind their targets in well under 10 s [143], meaning processes on the seconds time scale can be discerned, but anything shorter may be missed. Third, if Fabs are too numerous in a cell, they may compete with one another for binding, and Fab targets could become saturated, both of which could interfere with the underlying biology. We introduce 1-3 $\times$ 106 Fab per cell [139], far less than the $\sim$1.5 $\times$ 107 RNAP2 heptad repeats [143, 173]. We therefore do not expect Fabs to compete or interfere. Together these three caveats place considerable constraints on experiments, but they are not prohibitive. With the continued development of Fab [140], scFv [61], and nanobodies [174] for live-cell imaging, finding a suitable intrabody has become significantly easier. We therefore anticipate our technology will become a valuable tool to study transcription dynamics at the single-gene level.

3.1.5 Methods

Cell culture

Transcription dynamics experiments were performed in HeLa Flp-in H9 cells (H-128). The H-128 cell line generation was described previously [138]. Briefly, H-128 cells harbor an HIV-1 reporter gene tagged with an MS2X128 cassette, controlled by Tat expression. The HIV-1 reporter comprises the 5' and 3' long terminal repeats containing the viral promoter, polyA sites, as well as HIV-1 splice donor (SD1), splice acceptor (SA7), and Rev-responsive element. H-128 cells also stably express MCP tagged with GFP (MCP-GFP), which binds to MS2 repeats when they are transcribed into mRNA. Cells were maintained in a humidified incubator at 37 degC with 5% CO2 in Dulbecco's modified Eagle medium (DMEM, Thermo Fisher Scientific, 11960-044) supplemented with 10% fetal bovine serum (FBS, Atlas Biologicals), 10 U/mL penicillin/streptomycin (P/S, In-

vitrogen), 1 mM L-glutamine (L-glut, Invitrogen) and either 400 μg/mL Neomycin (Invitrogen) or 150 μg/mL Hygromycin (Gold Biotechnology).

Chromatin immunoprecipitation and quantitative-polymerase chain reaction (ChIP-qPCR)

ChIP was performed as described previously [175] with minor modifications. Briefly, H-128 cells grown in a 10 cm dish were fixed with 1% PFA in DMEM at room temperature (RT) for 5 min, neutralized in DMEM containing 200 mM glycine for 5 min, and washed with PBS and NP-40 buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl and 0.5% NP-40). Fixed cells were lysed with 360 μL sodium dodecyl sulfate (SDS) dissolution buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA and 1% SDS) and diluted with 1440 μL ChIP dilution buffer (50 mM Tris-HCl, pH 8.0, 167 mM NaCl, 1.1% Triton 100× and 0.11% sodium deoxycholate), supplemented with a proteinase inhibitor cocktail. After shearing chromatin using a Bioruptor UCD-200 (Diagenode) at sonications of 40 sec with 50 sec intervals, eight times at high level, the median size of fragmented DNA was 200 base pairs with a range of 50-500 base pairs. The supernatant, cleared by centrifugation at 20,000×g for 10 min at 4 degC, was diluted with 5.4 mL ChIP dilution buffer and then incubated with 40 μL sheep anti-mouse IgG magnetic beads pre-incubated with 1 μg mouse anti-CTD-RNAP2 (MABI 0601), anti-Ser5ph-RNAP2 (MABI 0603), and anti-Ser2ph-RNAP2 (MABI 0602) monoclonal antibodies (Cosmo Bio USA) at 4 degC overnight with rotation. The immune complexes were washed with low-salt RIPA buffer (50 mM Tris-HCl, pH 8.0, 1 mM EDTA, 150 mM NaCl, 0.1% SDS, 1% Triton 100× and 0.1% sodium deoxycholate), high-salt RIPA buffer (50 mM Tris-HCl, pH 8.0, 1 mM EDTA, 500 mM NaCl, 0.1% SDS, 1% Triton 100× and 0.1% sodium deoxycholate) and then washed twice with TE buffer (10 mM Tris-HCl, pH 8.0, and 1 mM EDTA). DNA was eluted with ChIP elution buffer (10 mM Tris-HCl, pH 8.0, 300 mM NaCl, 5 mM EDTA and 0.5% SDS). After incubation at 65 degC overnight to reverse the cross-links, DNA was purified by RNase A and proteinase K treatments and recovered using a DNA purification kit (Qiagen). For ChIP-qPCR, the immunoprecipitated DNA and total DNA were quantified by Power

SYBR Green PCR Master Mix in an Mx3000P Real-Time qPCR System (Agilent Technologies). The primers used for qPCR are listed in Supplementary Table 1.

Antigen-binding fragment (Fab) generation and fluorescence conjugation

Fab preparation was performed using the same monoclonal antibodies used in ChIP experiments and the Pierce Mouse IgG1 Fab and F(ab')2 Preparation Kit (Thermo Scientific), as described before [143]. In brief, ficin resin was equilibrated with 25 mM cysteine (in HCl, pH 5.6) to digest the antibodies (CTD-RNAP2 or Ser5ph-RNAP2) into Fab. The IgG concentration used was 4 mg, and the digestion reaction was incubated for 5 h. Fab and Fc regions were separated using a Nab Protein A column (Thermo Scientific). Fabs were concentrated up to $\sim$1 mg/mL using an Amicon Ultra 0.5 filter (10 k cut-off, Millipore) and conjugated with CF640 or Cy3 (Invitrogen) dyes. For labeling Fab, 100 μg of purified Fab and 10 μL of 1M NaHCO were mixed to a final volume of 100 μL, then 2 μL of CF640 or 2.66 μL of Cy3 was added, and the mixture was incubated at RT for 2 h in a rotator protected from the light. The labeled Fab sample was passed through a PD-mini G-25 desalting column (GE Health care), previously equilibrated with PBS, to remove unconjugated Fab, and then the dye-conjugated Fab was concentrated up to $\sim$ 1 mg/mL with an Amicon Ultra filter 0.5 (10 k cut-off). The degree of labeling (DOL) was calculated using Eq. 3.1, where ϵ_{IgG} and ϵ_{dye} are the extinction coefficients of IgG at 280 nm and the dye (provided by the manufacturer), A_{Fab} and A_{dye} are the absorbances determined at 280 and 650 or 550 nm, and CF is the correction factor for the dye at 280 nm (provided by the manufacturer). In this study, only Fabs with a DOL between 0.75 and 1 were used for live-imaging experiments.

$$DOL = \frac{\epsilon_{\mathrm{IgG}}}{\epsilon_{\mathrm{dye}}} * \frac{1}{\frac{A_{\mathrm{fab}}}{A_{\mathrm{dye}}}^{-1} - CF} \tag{3.1}$$

Loading fluorescent Fabs into living cells

Cells were cultured in glass-bottom dishes (35 mm, 14 mm glass, Mat-Tek). The next day dye-conjugated Fabs were loaded into the cells through bead-loading [139, 143, 145, 176, 177], as follows: First, the fluorescent Fabs (CTD-RNAP2-CF640 and Ser5ph-RNAP2-Cy3, $\sim$1 mg/mL,

each) were mixed with PBS up to 4 μL in the cell culture hood. Second, the medium was removed completely from the dish and stored, and the Fab mixture was added to the center of the dish. Third, glass beads (106 μm, Sigma-Aldrich, G-4649) were immediately sprinkled on top before cells dried up and the dish was tapped $\sim$10 times against the bench. This tapping causes the beads to roll over cells and induce small tears into which the Fab can diffuse in. Fourth, the stored medium was quickly added back to the cells, again to prevent cells from drying out. Cells were then placed in the incubator to recover for 1–2 h. Post-recovery, the glass beads were gently washed out with phenol red-free DMEM (DMEM-, Thermo Fisher Scientific, 31053-028), and the cells were stored in DMEM+ medium (DMEM- supplemented with 10% FBS, 10 U/mL P/S, and 1 mM L-glut) for live-imaging experiments.

Chemicals

The transcription inhibitors, Triptolide (TPL, Sigma Aldrich), Flavopiridol (Flav, Selleck Chemicals), THZ1, (Selleck Chemicals), fluorescence dyes, Cy3 (Invitrogen), CF640 (Invitrogen), and HaloTag TMR Ligand (5mM) (Promega) were dissolved in DMSO (Sigma-Aldrich) and stored at -20 degC until use. RNAP2 inhibitors were added to the DMEM+ medium to reach the desired final concentration in the cells.

Microscopy

A custom-built widefield fluorescence microscope with highly inclined illumination was used in all experiments [79, 23]. The microscope has three excitation beams: 488, 561, and 637 nm solid-state lasers (Vortran) that are coupled and focused on the back focal plane of the objective (60×, NA 1.48 oil immersion objective, Olympus). The emission signals were split by an imaging grade, ultra-flat dichroic mirror (T6601pxr, Chroma) and detected with two aligned EM-CCD (iXon Ultra 888, Andor) cameras by focusing with a 300 mm tube lens (generating 100× images with 130 nm/pixel). Cell chambers were mounted in a stage-top incubator (Okolab) at 37 degC with 5% CO2 on a piezoelectric stage (PZU-2150, Applied Scientific Instrumentation). The focus was maintained with the CRISP Autofocus System (CRISP-890, Applied Scientific Instrumenta-

tion). The cameras, lasers, and piezoelectric stage were synchronized with an Arduino Mega board. Image acquisition was performed with Micro-Manager software (1.4.22) [119]. Unless otherwise stated, the imaging size was set to 512 $\times$ 512 pixels^2 (66.6 $\times$ 66.6 μm^2), and the exposure time set to 53.64 msec. The readout time of the cameras from the combination of the imaging size and the vertical shift speed was 23.36 msec, which resulted in an imaging rate of 13 Hz (77 msec per image).

For three-color imaging, far-red fluorescence (e.g., CF640 or Alexa Fluor 647) was imaged on one camera with an emission filter (FF01-731/137/25, Semrock), while red fluorescence (e.g., Cy3 or TMR) and green fluorescence (e.g., GFP) were alternately imaged on the other camera via a filter wheel (HS-625 HSFW TTL, Finger Lakes Instrumentation) with an emission filter for red fluorescence (593/46 nm BrightLine, Semrock) and green fluorescence (510/42 nm BrightLine, Semrock). The filter wheel position was rapidly switched during the 23.36 msec camera read-out time by the Arduino Mega board. For two-color imaging, far-red fluorescence was simultaneously imaged on one camera while red or green fluorescence was imaged on the other camera with the appropriate emission filters.

Immunofluorescence

Cells grown on glass-bottom dishes (35 mm, 14 mm glass, uncoated, Mat-Tek Corporation) were fixed with 4% Paraformaldehyde (Electron Microscopy Sciences) in 1 M HEPES (Sigma-Aldrich) with or without 10% Triton 100× (Fisher Scientific) (pH 7.4) for 10 min at RT, and washed with PBS (3×). Permeabilization (1% Triton 100× in PBS) and blocking (100% blocking One-P, Nacalai-USA) were performed individually, for 20 min at RT, gently rocking, and rinsing with PBS (3×) after each step. The cells were incubated for 2 h at RT with 1 mL of antibody solution (10% blocking One-P:90% PBS) containing 2 μg/mL of mouse monoclonal primary antibody (CTD-RNAP2 (MABI 0601), Ser5ph-RNAP2 (MABI 0603), Ser2ph-RNAP2 (MABI 0602), as described in ref. [143] and now available from Cosmo Bio USA, H3K27ac (MABI0309), H3K27me (MABI0321), H3K4me1–3 (MABI0302-0304), H3K9me2 (MABI0317), and H3K9me3 (MABI0318), purchased from Cosmo Bio USA). After rinsing with PBS (3×), the

cells were incubated for 1 h at RT with 1 mL of antibody solution containing 1.5 μg/mL of Alexa Fluor 647 Donkey Anti-Mouse IgG (Jackson ImmunoResearch) and washed with PBS (3×). Then, the cells were mounted using Aqua-Poly/Mount (Fischer Scientific) for imaging. Single images were acquired with laser powers at the back focal plane set to 86 and 51.2 μW for 488 and 637 nm, respectively.

In addition, to show that CTD- and Ser5ph-RNAP2 Fabs stain cells distinctively, immunostaining using pre-labeled Fabs against CTD-RNAP2-CF640 and Ser5ph-RNAP2-Cy3 was performed. For this type of experiment, cells were fixed, permeabilized, and blocked as described above, and then incubated in an antibody solution containing 2 μg/mL of each pre-labeled Fab for 1 h at RT. Post Fab incubation, cells were rinsed with PBS and mounted in Aqua-Poly/Mount. The images were collected using the following laser powers at the back focal plane: 123, 750, and 230 μW for 488, 561, and 637 nm, respectively.

Single-molecule experiments using H2B-Halo

Cells were plated in glass-bottom dishes at a seeding density of $\sim$104 cells/cm2. The next day, cells were transfected with 2.5 μg of H2B-Halo in a 1:1 (mass) ratio using Lipofectamine LTX (Thermofisher Scientific, 15338-100). Twenty-four-hour post-transfection, cells were stained with 5 nM Halo-Ligand TMR pretreated with 30 mM NaBH4 for 30 min in the CO2 incubator (Acros Organics) to reduce the fluorophore and induce stochastic photoblinking in live-cells [178]. After staining, the cells were washed three times total. Each wash consisted of $\times$ mL DMEM-, and 1 mL DMEM+ with a 5-min interval between washes. Cells were imaged immediately after staining and washing. For this, the imaging size was set to 256 $\times$ 256 pixels2 (33.3 $\times$ 33 3 μm2), and the exposure time set to 30 msec. This resulted in an imaging rate of 22.8 Hz (30 msec exposure + 13.86 camera readout = 43.86 msec per frame). Single z-planes were acquired for 10,000 frames total with laser powers at the objective's back focal plane set to 125 μW, and 9.93 mW for 488 and 561 nm, respectively. To minimize photobleaching, the 488 nm laser fired once every ten frames (to track the TS), while the 561 nm laser fired every frame (for tracking individual H2B).

Single-molecule tracks were identified using TrackMate 3.8 with the following parameters: LoG Detector; Estimated Blob Diameter: 5.0; Pixel Threshold: 100; Sub-Pixel Localization: Enabled; Simple LAP Tracker; Linking Max Distance: 3 pixels; Gap-Closing Max Distance: 2 pixels, Gap-closing Max-Frame Gap: 1 frame. Custom Mathematica code was used to calculate the average Euclidean displacement for each track longer than 5 frames. Tracks were plotted with a blue–purple color distribution based upon their average Euclidean displacement. The transcription site was identified using TrackMate and plotted in red.

Live-cell imaging of transcription at the HIV-1 reporter gene

To cover the entire cell nucleus, all movies were taken using 13 z-stacks with 0.5 μm spacing. The z position was moved only after all three colors were imaged in each plane. This resulted in a total cellular imaging rate of 0.5 Hz (2 s per volume). Note that the color scheme of the signals described in the text and figures is based on the color of the excitation lasers, CTD-RNAP2 in red (CF640), Ser5ph-RNAP2 in green (Cy3), and mRNA in blue (GFP). For shorter live-cell imaging as in Fig. 3.1, each cell was scanned every 1 min for 30 min with the laser power at the objective's back focal plane set to 21.4, 60.5, and 21.74 μW for 488, 561, and 637 nm, respectively, and the exposure time was 53.64 msec. For longer live-cell imaging as in Fig. 3.5, cells were imaged every 1 min for 200-time points, using weaker laser powers (1.15, 15.7, and 5.2 μW for 488, 561, and 637 nm, respectively) and longer exposure times (200 msec exposure). For faster live-cell imaging as in Supplementary Fig. 3.14, each cell was scanned at a much faster frame rate (150 msec/frame) for a total of 1000 time points (150 sec) in a single plane. For this, the imaging size was set to 256 $\times$ 256 pixels2 (33.3 $\times$ 33.3 μm^2) and the exposure time set to 53.64 msec, with the laser power at the objective's back focal plane set to 1.2 mW, 335 μW, and 77.5 μW for 488, 561, and 637 nm, respectively.

Calibrating the number of mRNA per transcription site

To count the number of nascent mRNAs at the transcription site, cells were imaged for a single time point using a higher laser power for 488 nm (230 μW at the back focal plane) and a lower

camera gain. These conditions allowed us to visualize both a single transcription site and single mature mRNAs. To calculate the number of mRNA per transcription site (see Fig. 3.7d, bottom panel): (1) Several cells were imaged on independent days. To avoid bias, cells were chosen with the same imaging conditions used for longer live-cell experiments; (2) Images were analyzed using FISH-quant V3 [149]. Mature mRNAs were detected, localized in 3D with a Gaussian fit, and then a point-spread function was applied to discard spots that were larger than diffraction-limited spots. An image showing the average intensity of the mature mRNAs was created and compared to that of the transcription site. This ratio of these gave the number of nascent mRNA at each transcription site, from which the distribution shown in Fig. 3.7d (bottom panel, purple distribution) was computed.

Quantifying signal intensities at the transcription site from live-cell imaging movies

Images were pre-processed using either Fiji [179] or custom-written batch processing Mathematica code (Wolfram Research 11.1.1) to create 2D maximum intensity projections from 3D movies. Using Mathematica code, the 3D images were corrected for photobleaching and laser fluctuations, z-stack by z-stack, by dividing the movie by the mean intensity of the whole cell or the nucleus in each channel. The offset between the two cameras was registered using a built-in Mathematica routine FindGeometricTransform, which finds a transformation function that aligned the best-fitted positions of 100 nm diameter Tetraspeck beads evenly distributed across the image field of view. 2D maximum projections and 3D image sequences from the images corrected for bleaching and laser fluctuations were then analyzed with a custom-written code in Mathematica to detect and track the transcription site. Briefly, thresholds were selected in each channel to visualize spots at the transcription site and a bandpass filter was used to highlight just the transcription site in the mRNA channel. The resulting image was binarized and used to create two masks for each time point: one marking the transcription site (TS mask: a mask semi-manually thresholded to cover just the transcription site within the image) and one marking the BG (mask: a ring of width one pixel that surrounds the transcription site and is separated from the site by two pixels). The built-in Mathematica routine ComponentMeasurements-IntensityCentroid was used to find the coordinates

of the transcription site in XY through time. The Z coordinate was determined by selecting the z-stack at which the particle in the XY coordinate had its maximum brightness ("best z"). If the transcription site disappeared (due to transcription turning off or inhibition), the Z was replaced by the Z coordinate of the last visible position. From the XYZ coordinates at each time point, a new 2D maximum projection was created considering the "best z" at each time point. From this, the pixel intensity values were recorded for each TS and BG mask, representing the mean intensity values over time at the transcription site and the BG, respectively. The raw and normalized intensity vectors were calculated per channel and a moving average of three-time points was used to display the intensity RawIntCh as a function of time, as shown in Eq. 3.2:

$$RawInt_{\mathrm{Ch}} = \langle I_{\mathrm{TS}}(t) - I_{\mathrm{BG}}(t) \rangle \,|_3 \tag{3.2}$$

where, I_{TS} is the intensity measured in the TS mask in each channel (Ch), I_{BG} is the intensity measured in the BG mask, and $\langle \cdot \rangle |3$ represents a three-time point moving average. The normalized intensity (as in Fig. 3.5c) was calculated by dividing RawIntCh by the average 95% intensity from all transcription sites. Occasionally, normalized intensities for CTD-RNAP2 and Ser5ph-RNAP2 dip below zero. This can be caused either by RNAP2 signals being temporally depleted at the transcription site relative to the BG or by bright signals in the BG due to nearby transcription in the local vicinity. To display transcription sites over time (as in Figs. 1c and 2b and Supplementary Movie 1 (available online)), 3-time point moving-average trims from the "best z" were created in each channel (showing CTD-RNAP2, Ser5ph-RNAP2, mRNA, and the merge). Each trim was centered on the intensity centroid of the mRNA.

Covariance analysis in Supplementary Figs. 1f and 3g

To test for covariance between intensity signals from control spots and the transcription site, signal covariance was calculated using the "cov" function in MATLAB. For quantification of bleed-through, the covariance was calculated between all possible pairs of raw intensities (CTD-mRNA, CTD-Ser5ph, and Ser5ph-mRNA) in normal vs. bleed-through control conditions. For quantifica-

tion of signals off-target, the covariance was calculated between all possible pairs of normalized intensities on-target at the transcription site vs. off-target at a random site p1. Significance was calculated using the Mann–Whitney U test.

Analysis of minima signals in Fig. 3.5d

The local minima in the mRNA signal of each cell were detected using the "islocalmin" function in MATLAB. The cells that exhibited minimas below a threshold (normalized intensity ≤ 0.20 arb.units) were selected by the algorithm. Then seven-time points before and after the mRNA valley were considered, including the minimum, in each channel. All the traces in each channel were averaged and fitted with a Gaussian using a 95% confidence interval to determine the minima and maximum steady state of the average trace in each channel.

To confirm that the minima were true and not an artifact of our analysis, the analysis was repeated at hundreds of random time points. Significance was calculated using the Mann–Whitney U test. The p values for the magnitude of the minima and their time delays were calculated by comparing the magnitude of the minima to the control and the time lag to minute zero in each signal, respectively.

Analysis of transcription site spatial organization in Figs. 2e–h and Supplementary Fig. 3.4h

Moving average (50 time points) movies were generated to accurately determine the mean XY position of the transcription site in each channel. As described in Quantifying signal intensities at the transcription site from live-cell imaging movies subsection, the built-in Mathematica routine "ComponentMeasurements-IntensityCentroid" was used. Once the XY positions for each signal were obtained, the Euclidean distance between each pair over time was calculated, from which distributions were calculated. Significance between signals was calculated using the Mann–Whitney U test.

Auto- and cross-correlation analysis

The auto- and cross-correlation functions were calculated for each time trace obtained from the longer movies (like in Fig. 3.5c, but without performing a 3-time point moving average), as previously described [82, 76]. The covariance function is defined as:

$$G(\tau) = \langle \delta a(t) \delta b(t+\tau) \rangle, \tag{3.3}$$

where $\langle \cdot \rangle$ indicates the temporal mean, and $\delta a(t)$ denotes the deviation about the mean, i.e., $\delta a(t) = (a(t) - \langle a(t) \rangle)$. Signals $a(t)$ and $b(t)$ can be the same signal or two different signals. In the first case, $a = b$ and $G(\tau)$ represent the auto-covariance, which is symmetric about $\tau = 0$; In the second case, $G(\tau)$ represents the cross-covariance and may be asymmetric. To calculate the cross-correlation between the CTD-RNAP2 and Ser5ph-RNAP2 signals in the fast imaging experiments in Supplementary Fig. 3.14, the intensities of tracked transcription sites through time were quantified as described above, with a couple of minor modifications: First, because imaging was in a single plane, the rate of photobleaching in the plane was not captured by the rate of photobleaching in the cell. For this reason, each signal exponentially decayed. This was corrected by dividing out a single-exponential fit to each curve. Second, we did not perform any moving average on the signals to maintain the highest possible temporal resolution.

For fitting and data analysis, the normalized covariances, $G(\tau)/\bar{G(0)}$, were used for all signals, where $\bar{G(0)}$ denotes the zero-lag auto- or cross-covariance averaged over all time points and all biological replicas. To quantify and remove shot noise from the zero-lag auto-covariances, $G(0)$ was estimated for each biological replica assuming a linear interpolation from the three shortest non-zero lag times (1, 2, 3 min) prior to averaging over all replicas. The standard error of the mean normalized covariance functions, denoted $SEM_{\bar{G}}(\tau)$, was computed as the standard deviation of $G(\tau)/\bar{G(0)}$ divided by $\sqrt{N}$.

A quantitative model for transcription

The derivation of the bursting model for RNAP2 recruitment and nascent transcription simple model begins with the specification of three variables: $x_1(t)$ describes the promoter state, $x_2(t)$ describes the number of RNAP2 in the cluster, and $x_3(t)$ describes the number of RNAP2 engaged in active transcription. Six reactions can occur: (1) a promoter can become temporarily active with propensity equal to the burst frequency, $\omega \cdot k_{\text{on}}$; (2) the active promoter can deactivate at a rate k_{off}; (3) the active promoter can recruit and phosphorylate RNAP2 at Serine 5 (Ser5ph-RNAP2) at a rate $\beta \cdot k_{\text{off}}$; (4) Ser5ph-RNAP2 can be lost from the cluster at rate k_{ab}; (5) Ser5ph-RNAP2 can escape at rate k_{esc}; and (6) escaped RNAP2 can complete transcription with rate k_{c}. We solve the model for the first and second-order statistical moments as previously described [50]. First, we combine the stoichiometry vectors for all six reactions into the stoichiometry matrix, **S** as follows:

$$\mathbf{S} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \tag{3.4}$$

and we write the linear propensity functions in vector form as

$$\mathbf{w} = \begin{bmatrix} -k_{\text{on}} & 0 & 0 \\ k_{\text{off}} & 0 & 0 \\ \beta \cdot k_{\text{off}} & 0 & 0 \\ 0 & k_{\text{ab}} & 0 \\ 0 & k_{\text{esc}} & 0 \\ 0 & 0 & k_{\text{c}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} k_{\text{on}} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{3.5}$$

$$= \mathbf{W}_1 x + w_0 \tag{3.6}$$

with this notation, the expected mean dynamics of $\mathbb{E}\{\mathbf{x}\}$ are described by the ordinary differential equation:

$$\frac{d\mathbb{E}\{\mathbf{x}\}}{dt} = \mathbf{S}(\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{w}_0). \tag{3.7}$$

From this expression, the steady state expected mean can be calculated as the solution to the algebraic expression:

$$\mathbf{S}\mathbf{W}_1\mathbb{E}_{\mathrm{SS}}\{\mathbf{x}\} + \mathbf{S}\mathbf{w}_0 = 0; \tag{3.8}$$

the steady state co-variance, Σ_{SS}, can be calculated as the solution of the algebraic Lyapunov equation:

$$\mathbf{S}\mathbf{W}_1\Sigma_{\mathrm{SS}} + \Sigma_{\mathrm{SS}}\mathbf{W}_1^{\mathrm{T}}\mathbf{S}^{\mathrm{T}} + \mathbf{S}\mathrm{diag}(\mathbf{W}_1\mathbb{E}_{\mathrm{SS}}\{\mathbf{x}\} + \mathbf{w}_0)\mathbf{S}^{\mathrm{T}} = 0; \tag{3.9}$$

and the auto- and cross-covariance functions vs. time lag, $\Sigma(\tau)$ can be calculated as the solution of the ODE:

$$\frac{d\Sigma_{\mathrm{x}}(\tau)}{d\tau} = \mathbf{S}\mathbf{W}_1\Sigma_{\mathrm{x}}(\tau), \tag{3.10}$$

with initial condition $\Sigma_{\mathrm{x}}(0) = \Sigma_{\mathrm{SS}}$ given as the solution to (9).

To convert these above expressions, which are in terms of x_1–x_3, into quantities reflecting the total RNAP2 at the transcription site $(y1 = x2 + x3)$ and number of transcribing RNAP2 $(y2 = x3)$, we define a simple linear transformation

$$\mathbf{y} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x}, \tag{3.11}$$

$$= \mathbf{c}\mathbf{x}, \tag{3.12}$$

Under this transformation, $\mathbb{E}\{\mathbf{y}\} = \mathbf{c}\{\mathbf{x}\}$ and $\Sigma_{\mathrm{y}}(\tau) = \mathbf{c}\Sigma_{\mathrm{x}}(\tau)\mathbf{c}^{\mathrm{T}}$.

We note that this version of the model does not distinguish between RNAP2 and Ser5ph-RNAP2. These two distinct forms as well as other configurations are easily incorporated by extending **x** to include a fourth or more states. In such cases, each new state adds two reaction stoichiometry vectors to Eq. 3.4, two reaction terms to Eq. 3.5, and one additional column to the output matrix c in Eq. 3.11, but the rest of the analysis remains unchanged.

Using this model formulation, it is straightforward to solve for the steady-state moments (Eqs. 3.8 and 3.9) and the auto- and cross-correlations (Eq. 3.10) for any combination of parameters. However, upon fitting this model to the data, we observed that estimates for k_{off} tended to very large values ($k_{\text{off}} \gg \omega$) and with substantial estimation uncertainty. Under these excessively large rates for k_{off}, each "on" period is extremely short-lived and attracts a geometric number of RNAP2 with mean β (this model reduction is equivalent to the strategy in models that use geometric bursts of protein to replace translation of short-lived mRNA as described previously elsewhere [41]). Therefore, to reduce the number of free parameters required by the model, we fixed k_{off} at 1000min^{-1} such that each burst would be very short-lived on the time scale of the experimental measurements. This choice led to a simpler model but had no discernible effect on the fit of the model to the data.

All codes and the GUI are available at Zenodo: https://doi.org/10.5281/zenodo.4631141.

Model parameter search

Parameters were found using maximum likelihood estimation considering several data types as follows. First, errors in the measurement of the normalized auto- and cross-covariances were assumed to be normally distributed with the measured standard error, $SEM_{\bar{\text{G}}}(\tau)$, such that their log-likelihood functions are written

$$\log L_{\text{G}}(\theta) = C_{\text{G}} - \frac{1}{2}\frac{\sum_{\text{n}=1}^{\text{N}_\text{t}}(\bar{G}_{\text{D}}(\tau_{\text{n}}) - \bar{G}_{\text{M}}(\tau_{\text{n}},\theta))^2}{(SEM_{\bar{\text{G}}_{\text{D}}}(\tau_{\text{n}}))^2} \tag{3.13}$$

where θ is the set of parameters, $\bar{G}_{\text{D}}(\tau_{\text{n}})$ is the measured covariance function in the data (D), $\bar{G}_{\text{M}}(\tau_{\text{n}},\theta)$ is the predicted covariance function of the model (M) at a time lag of τ_{n}, and C_{G} is a

normalization constant that does not depend on the parameters. The summation is over the first 15 lag times for the three auto-covariance functions and the 21 smallest lag times (i.e., -10 to 10 min) for the three cross-covariance functions.

The model was further constrained to match the mean and variance for the measured number of mRNA per transcription site as estimated in units of mature mRNA as calibrated using FISH-quant. Assuming the central limit theorem, the log-likelihood of matching the observed sample mean was estimated as:

$$\log L_{\mu}(\theta) = C_{\mu} - \frac{1}{2}\frac{((\mu_{\mathrm{D}} - \mu_{\mathrm{M}}(\theta))^2}{SEM_{\mathrm{D}}^2}, \tag{3.14}$$

where μ_{D} is the sample mean levels of mRNA from the data, $\mu_{\mathrm{M}}(\theta)$ is the mean number of mRNA predicted by the model, and SEM_{D} =0.93 is the standard error of mean level of mRNA from the data. Similarly, the log-likelihood of the measured variance, σ_{D}^2 given the model was estimated as

$$\log L_{\sigma^2}(\theta) = C_{\sigma^2} - \frac{1}{2}\frac{((\sigma_{\mathrm{D}}^2 - \sigma_{\mathrm{M}}^2(\theta))^2}{SEM_{\sigma^2}^2}, \tag{3.15}$$

where $\sigma_{\mathrm{M}}^2(\theta)$ is the mRNA variance predicted by the model, SEM_{σ^2} is the standard error for the mRNA variance, and C_{sigma^2} is a constant that does not depend on the parameters. The standard error of the sample variance was estimated using a Gaussian approximation such that

$$SEM_{\sigma^2} = SEM_{\mathrm{D}}^2\sqrt{\frac{2}{N-1}} = 14 \tag{3.16}$$

Under the assumption of independence between the different data types, the total log likelihood to match all data was the sum of the individual likelihoods

$$\log L_{\mathrm{Total}}(\theta) = \log L_{\mathrm{G}}(\theta) + \log L_{\mu}(\theta) + \log L_{\sigma^2}(\theta). \tag{3.17}$$

Maximum likelihood estimates were found using iterated rounds of MATLAB's "fminsearch" until convergence.

To compare multiple models with different numbers of mechanisms and parameters, we computed the BIC as

$$\mathrm{BIC} = k\log(n) - 2\log L_{\mathrm{Total}}(\theta) \tag{3.18}$$

In this formulation, the value for the number of independent experiments, n, was estimated at $n = 8$, which conservatively assumes one data degree of freedom for each of the six different auto- and cross-correlation signals estimated from the time-lapse experiments, and one each for the measurement of the mean and variance of mRNA per transcription site as estimated imaging a single frame using higher laser power to visualize single mature mRNAs. The number of parameters, k, disregards the directly measured shot noise magnitudes and any parameter that was fixed at a large value (e.g., k_{out} when that value was fixed to 1000 sec^{-1}). This leaves $k = 5$ parameters for the selected model: β, ω, k_{ab}, k_{esc}, and kc. The fractional phosphorylation or phosphorylation models each have one additional parameter, fraction or k_{phos}, respectively. The mRNA retention model has one additional parameter (i.e., k_{c} was replaced with $k_{\mathrm{c-mRNA}}$, and k_{release} was added). The numbers of parameters, maximum likelihood values, and parameter estimates, and BIC results for all examined models are listed in Supplementary Fig. 3.8. We note that our low conservative estimate for $n = 8$ (rather than basing n on the much larger number of independent experiments) was chosen to avoid biasing the model selection toward simpler models—larger choices of n would result in much stronger rejection of the more complex models.

Transcription inhibition experiments

For the transcription inhibition experiments in Fig. 3.11, cells were imaged every 1 min for 5-time points before applying the inhibitor ($t = 0$), TPL (5 μM), THZ1 (15 μM), or Flav (1 μM). Cells were then imaged every 1 min for 30 min total after addition of TPL or Flav, and for 55 min

total after addition of THZ1. Here, laser power at the objective's back focal plane was set to 21.4, 60.5, and 21.74 μW for 488, 561, and 637 nm, respectively, and the exposure time was 53.64 msec.

To quantify time delays in the TPL-runoff assay, TPL signals were further analyzed as follows: (1) To account for cell variability and experimental conditions, the decays curves from each cell were aligned. This was achieved by subtracting the time at which each cell reached half of the decay after TPL addition. This time was obtained by an inverse hyperbolic tangent fit applied to each channel in every cell (Fig. 3.11c); (2) after the alignment, all the traces in each channel were averaged together, and the standard error of the mean (S.E.M.) was calculated. Finally, to determine the time delays between CTD-RNAP2, Ser5ph-RNAP2, and mRNA, an inverse tanh fit was applied and weighted with respect to the variance of each signal.

Software

All images were acquired with Micro-Manager software (1.4.22). Image pre-processing was made using ImageJ (2.0.0 - rc - 67/1.52e Java 1.8.0_66, 64 - bit). Images were analyzed with a custom Wolfram Mathematica (11.1.1) code. For the fast movies, tracking of the spots was performed using the ImageJ plugin, TrackMate (3.8.0). Final plots, modeling, and the GUI were made using MATLAB R2019b; Figures were assembled together using CorelDraw 2020 (64 - bit).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The source data underlying the figures are provided as a Source Data file in the original publication online. The raw images/movies were deposited to Figshare https://doi.org/10.6084/m9.figshare.14187011. All other data included in the manuscript and/or supplemental materials are available from the corresponding authors upon reasonable request. Source data are provided with this paper.

Code availability

All codes used for the mathematical model and GUI are available at Zenodo https://doi.org/10.5281/zenodo.4631141. The Mathematica codes employed for the image processing in this paper are available from the authors upon request.

Contributions

Linda S. Forero-Quintero: Conceptualization, performed experiments, collected data, Fab preparation, software implementation and development, formal analysis, original draft, review and editing

William Raymond: Software implementation and development, computational modeling, original draft, review and editing

Tetsuya Handa: Performed experiments, collected data, provided antibodies, review and editing

Matthew N. Saxton: Performed experiments, collected data, review and editing

Tatsuya Morisaki: Software implementation and development, review and editing

Hiroshi Kimura: Provided antibodies, Fab preparation, review and editing

Edouard Bertrand: Provided H-128 cells, review and editing

Brian Munsky: Conceptualization, Software implementation and development, computational modeling, original draft, review and editing, resources, funding, and supervision.

Timothy J. Stasevich: Conceptualization, Fab preparation, Software implementation and development, original draft, review and editing, resources, funding, and supervision.

3.2 Visualization, quantification, and modeling of endogenous RNA polymerase II phosphorylation at a single-copy gene in living cells - BioProtocol[6]

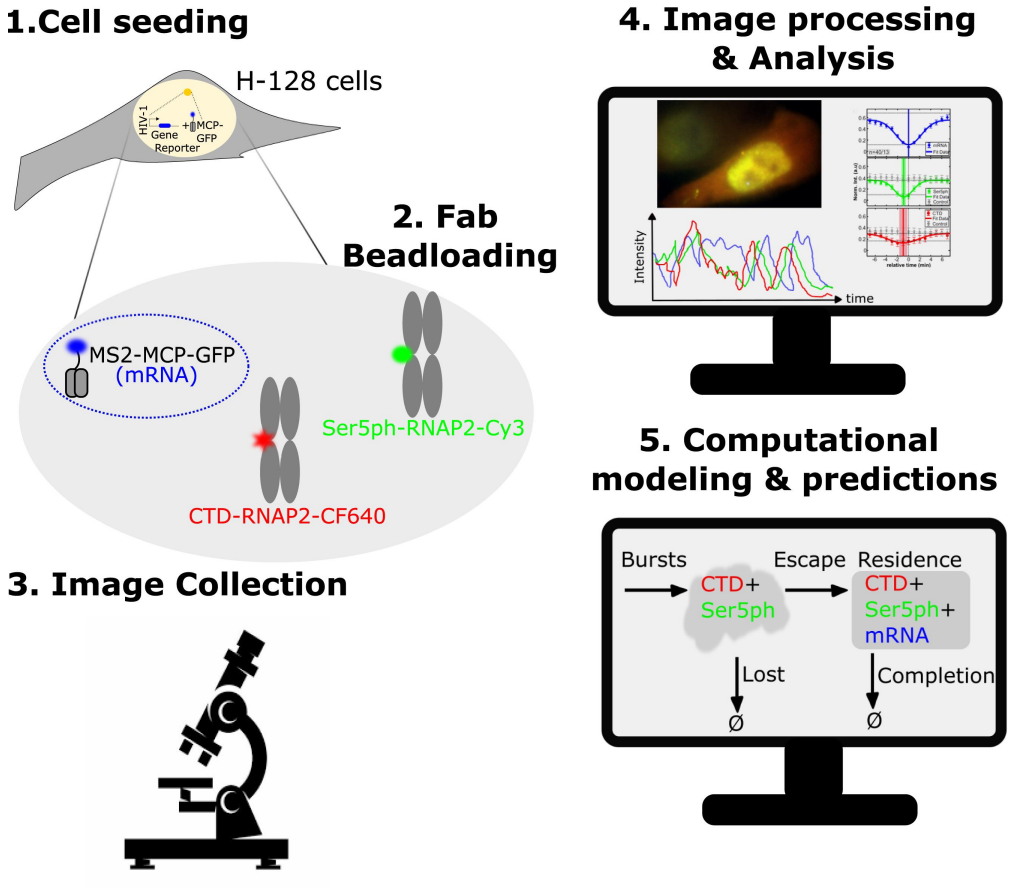


Figure 3.16: Bio-protocol graphical description

The Nature Communications paper was adapted into a Bio-protocol paper in summer 2022. We were invited to convert our manuscript into a usable protocol. To this end, I converted all the codes used in the previous paper into a general MATLAB pipeline to allow for all the analyses shown in the original manuscript on a given users data set.

[6]Linda S. Forero-Quintero, William Raymond, Brian Munsky, Timothy J. Stasevich DOI: https://doi.org/10.1101/2023.11.23.568398

Chapter 4

non-coding RNA Machine Learning

4.1 Identification of potential riboswitch elements in *Homo-Sapiens* mRNA 5'UTR sequences using Positive-Unlabeled machine learning[7]

During the course of my studies at Colorado State University, I have focused on the nucleotides relegated to central dogma of biology - DNA, tRNA, and mRNA. I took a wonderful RNA biology class taught by Dr. Jeffery Wilusz which introduced me to the insane diversity of the peripheral yet crucial non-coding RNA species: rRNA, snoRNA, snRNA, miRNA, circRNA and almost 40 other types. I credit this class with my current interest in non-coding RNA. There is a wild west of unexplored RNA species harbored within cells whose current functional understanding surmounted to "We know that it's bad if it's deleted!"I found several open questions in RNA biology intriguing, and when it came time to pursue a project for another class (bioinformatics) I opened my list of open questions to select one to work on. I eventually elected to search for Riboswitches within the human UTR, the search basically boils down to a single machine learning classification plus some computational exploration which the first pass was enough for the class project —despite the relative "ease" of the project, there were several challenges to overcome before it was a full, publishable manuscript. I won't linger every pothole, but a short list included wrangling databases with contradictory information, web-scraping multiple elements not included in downloads, databases migrating and updating, testing several different machine learning algorithms, and building a website when I have never attempted that before. I kept working slowly on this project in downtime during the main multiplexing project. In November 2023 the following manuscript was completed and submitted for publication and I am proud to say it is currently under review at PLOS computational biology.

[7]William S. Raymond, Jacob DeRoo, Brian Munsky DOI: https://doi.org/10.1101/2023.11.23.568398

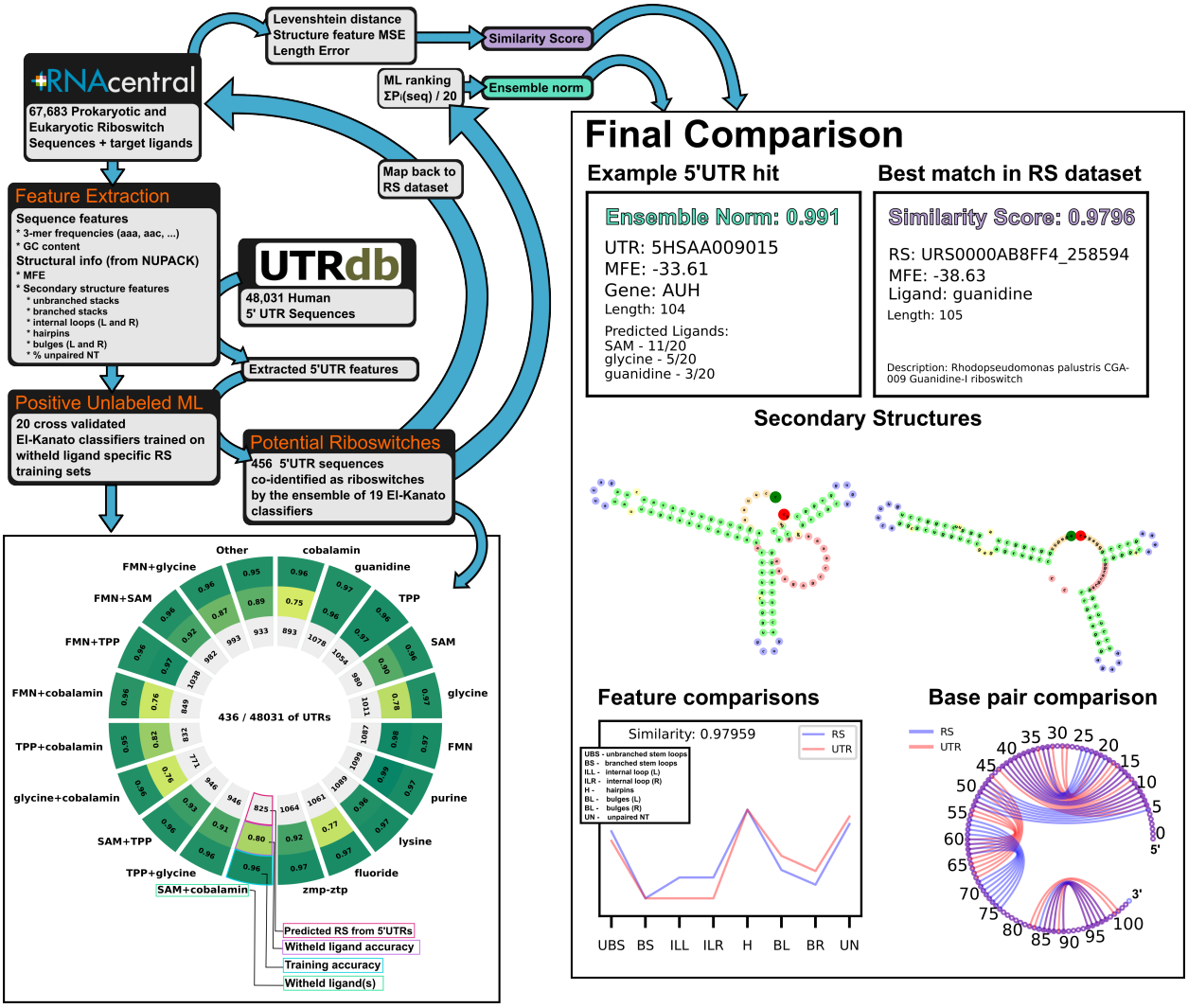


Figure 4.1: Human 5 prime UTR riboswitch ML search project graphical description

4.1.1 Summary

Riboswitches are a class of noncoding RNA structures that interact with target ligands to cause a conformational change that can then execute some regulatory purpose within the cell. Riboswitches are ubiquitous and well characterized in bacteria and prokaryotes, with additional examples also being found in fungi, plants, and yeast. To date, no purely RNA-small molecule riboswitch has been discovered in *Homo Sapiens*. Several analogous riboswitch-like mechanisms have been described within the *H. Sapiens* translatome within the past decade, prompting the question: Is there a *H. Sapiens* riboswitch dependent on only small molecule ligands? In this work, we set out to train positive unlabeled machine learning classifiers on known riboswitch sequences and

apply the classifiers to *H. Sapiens* mRNA 5'UTR sequences found in the 5'UTR database, UTRdb, in the hope of identifying a set of mRNAs to investigate for riboswitch functionality. 67,683 riboswitch sequences were obtained from RNAcentral and sorted for ligand type and used as positive examples and 48,031 5'UTR sequences were used as unlabeled, unknown examples. Positive examples were sorted by ligand, and 20 positive-unlabeled classifiers were trained on sequence and secondary structure features while withholding one or two ligand classes. Cross validation was then performed on the withheld ligand sets to obtain a validation accuracy range of 75%-99%. The joint sets of 5'UTRs identified as potential riboswitches by the 20 classifiers were then analyzed. 15333 sequences were identified as a riboswitch by one or more classifier(s) and 436 of the *H. Sapiens* 5'UTRs were labeled as harboring potential riboswitch elements by all 20 classifiers. These 436 sequences were mapped back to the most similar riboswitches within the positive data and examined. An online database of identified and ranked 5'UTRs, their features, and their most similar matches to known riboswitches, is provided to guide future experimental efforts to identify *H. Sapiens* riboswitches.

4.1.2 Introduction

A riboswitch (RS) is a non-coding RNA sequence harboring a structure with two distinct conformations. Conformational changes are induced when an aptamer region interacts with a target small molecule, revealing or occluding functional parts of the RNA. This inducible change in structure allows for broad regulation of various cellular processes via modification of protein production, mainly by affecting transcription termination/continuation, translation inhibition/activation, mRNA splicing, or mRNA stability [180, 181, 182, 183]. Whether a particular riboswitch acts in a positive or negative regulatory manner when in contact with its ligand strongly depends on the expression platform and aptamer location in relation to other elements, such as the ribosomal binding site. These regulatory effects could conceivably have disease-related implications due to under- or over-expression of a regulatory biomolecule or to a genetic mutation of the riboswitch in question. The current set of described riboswitches is predicted to be a small subset of all existing ri-

boswitches —leading to open questions such as to "which riboswitch classes are uncharacterized?" or "why do some life-essential molecules lack known riboswitch aptamers?" Riboswitch discovery has been an active area of research since their first description in 2002, with many computational and experimental efforts undertaken to elucidate new riboswitch classes [184]. Riboswitches occur ubiquitously in prokaryotes, where they enjoy a rich diversity of around 40 molecular targets [182]. In contrast, nearly every example of a eukaryotic riboswitches in the current literature was found in lower eukaryotes, such as mold, yeast, and fungi, and they leverage thiamine pyrophosphate (TPP) as their target ligand [181, 185, 184]. Among higher eukaryotes, several species of plants have a single TPP riboswitch in the 3'UTR of the conserved gene THIC —the riboswitch acts to regulate gene expression by creating an unstable mRNA product in the presence of TPP [186]. Interestingly, multiple analogous "pseudo-riboswitches" (riboswitch like elements that are stabilized by proteins and small molecules) have been located within the untranslated region (UTR) of human (*Homo Sapiens*, *H. Sapiens*) translatome [187, 188]. The existence of some analogous mechanisms in *H. Sapiens* and a described higher eukaryotic riboswitch prompts the question: "do riboswitches have an unknown niche in all higher eukaryotes, or are they simply missing?" Indeed this question is one of the largest open questions in the field today [184]. From a disease perspective, do riboswitches exist within in *H. Sapiens*, and if so, where and with what implications on human health?

Machine learning is routinely used in Bioinformatics for a wide range of RNA related tasks, including parsing RNA-seq data, performing RNA secondary structure prediction, and providing discovery based approaches for RNA sequences, splice sites, and genome wide functional RNA elements [189, 190]. For the task of computational riboswitch prediction, previous methods leverage hidden Markov models (HMMER, RiboSW, Riboswitch Scanner [191, 192, 193]), covariance models and context free grammar (Infernal [194]), and sequence alignment + computational folding (Riboswitch Finder, RibEx, RNAConSLOpt [195, 196, 197]). Other more recent software such as Riboflow utilizes deep learning classifiers such as RNN-LSTM or convolutional neural networks (CNN) for their riboswitch identification [198]. Computational methods available to the public up

until 2018 are reviewed in Antunes et al. extensively [199]. Notably, many of these have difficulty extrapolating to unknown riboswitches and rely heavily on a previous knowledge base. Recent breakthroughs have been achieved via reverse homology searching approaches (mutate a sequence without disturbing secondary structure, search for the new mutant in a genome wide fashion), which recently helped to identify a list of potential purine riboswitches in fungi [200] —however this approach once again suffers from a lack of extrapolation and requires a known starting point structure.

Positive unlabeled learning (PU-learning) is a subclass of binary machine learning classification that attempts to learn from data that only contain positive and unknown examples. In other words, they are used to classify data where the labels of one class are known (label 1) or known and incorrect (labeled 1, truly 0), and the other class are unknown and could either be true examples (label 1), unclassified examples (label 0.5), or negative examples (label 0) [201, 202, 203, 204]. Situations producing unlabeled-positive data sets are prevalent in fields such as medical diagnosis (e.g., people with a diagnosis vs. people without a condition vs. people with a condition and no diagnostic confirmation), interest-based applications (e.g., people who engaged with an ad vs. those who did not, since not engaging could be a negative or a neutral reaction), and biology (e.g., a class of known proteins vs. proteins with unclassified but similar function vs. proteins without the same function). PU-learning is routinely applied in molecular biological discovery applications since the advent of big data approaches [205, 206, 207]; Proteomics, RNA-seq, or whole genome sequencing quantify virtually all species within a sample whether or not the molecules are characterized, creating a tranche of unlabeled data along with its positive examples [208]. Within the context of RNA, PU-learning has also been used to identify non-coding RNA genes [209], predict circular-RNA and piRNA disease associations [210, 211], to predict RNA secondary structures [212], and to classify metastasis potential from cancer cell RNA-seq data [213].

In this work, we set out to use PU-learning to identify a group of potential sequences in the *H. Sapiens* mRNA 5'UTR that may contain riboswitches —with the hope of providing a first-pass reduced list for future, targeted laboratory investigations. 67,683 sequences tagged with

"riboswitch"from the non-coding RNA database, RNAcentral, were used as positive examples. 48,031 *H. Sapiens* 5'UTR sequences were obtained from the untranslated region (UTR) database, UTRdb, and used as unlabeled examples. Sequences were sanitized and structural- and sequence-based features were extracted. 20 PU-classifiers were trained on the RS-5'UTR feature sets and validated on single or double holdouts of specific RS ligand classes. The resulting ensemble of classifiers was then examined for the overlap of 5'UTR sequences that were considered as riboswitches (positively labeled). 436 sequences were found to be potential 5'UTR hits across all 20 classifiers. These positively labeled 5'UTR hits were then compared with their most similar sequences within the riboswitch data set via metrics comparing length, dot structure differences, and structural feature similarities, and all results are presented in an interactive display website. GO analysis was also preformed to examine fold enrichment of cellular processes and functions. Further verification of the classifier ensemble was performed by applying the classifier to a set of 25 synthetic riboswitches, of which 56% were correctly discovered as riboswitches despite having no representation of similar synthetic riboswitches in any training data. Using our computationally validated ensemble, we provide a minimal list of *H. Sapiens* 5'UTRs that appear most likely to harbor riboswitch sequences in hopes that these hits could be corroborated with future experimental validation.

4.1.3 Results and Discussion

67,683 known riboswitch sequences and 43,081 *H. Sapiens* 5'UTRs were collected and sanitized for subsequent PU classification.

Two RNA databases were selected for training data: RNAcentral and UTRdb. RNAcentral is a meta collection of many databases of all types of non-coding RNA, and was utilized as the source of riboswitch sequences [214]. JSON information of all entries containing the tag "riboswitch" were queried from RNAcentral on 8.19.22 and filtered to remove duplicate sequences. Ligands were parsed from the entry descriptions and any missing ligands were obtained from the corresponding entry's RFAM data [215]. Ligands were further filtered to combine names referring to the same

ligand (e.g. “mn”, “manganese”, “Mn2+” all renamed to “Mn2+”). Cobalamin sub-types such as Adenosylcobalamin were combined under the umbrella of “cobalamin” for ligand labelling. Any protein specific ligand was renamed to “protein”(1 total) and all tRNA ligands were lumped to “tRNA”(3 total). Speculative or synthetic riboswitches (nhA-I motif, duf1646, raiA, synthetic, sul1, blank) were relabeled with ‘unknown’ as their ligand (1130 total). After ligand relabeling, 73,119 riboswitch sequences remained in the data set. After removing identical sequences, 67,683 penultimate riboswitch sequences were stored for machine learning. Riboswitches targeting cobalamin(s), TPP, S-Adenosyl methionine (SAM), glycine, FMN, purine, lysine, fluoride, and guanidine made up 82% of the riboswitch data set. Other ligand labels such as unknown, molybdenum, GMP, or nickel/cobalt made up less than 2% of the data set each (Fig. 4.2A). A full list of ligands represented in the data set can be found in Table 4.1.

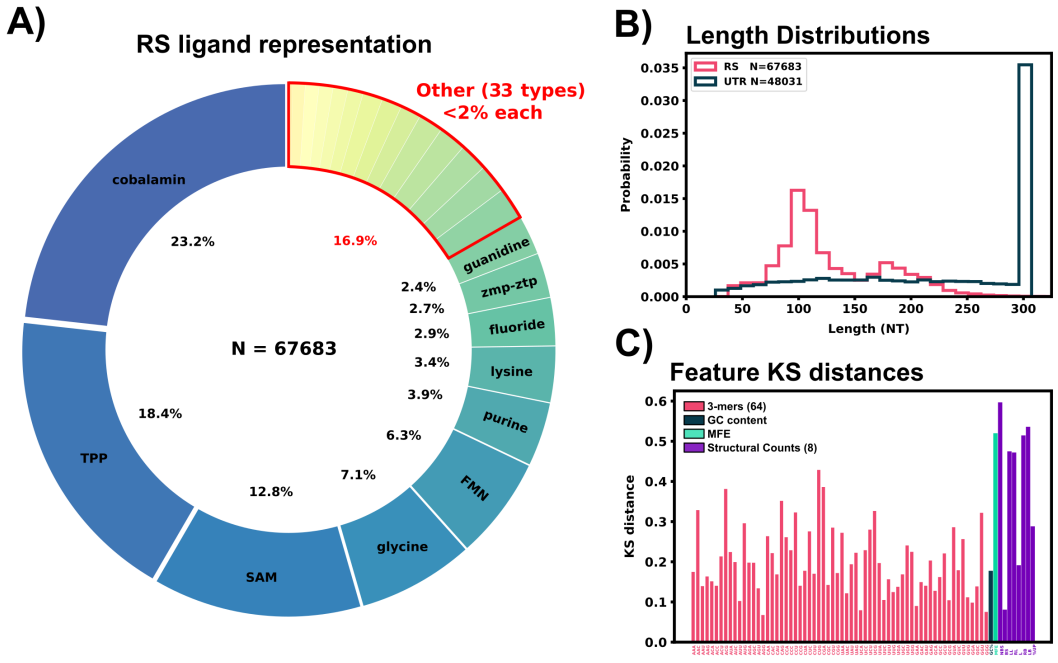


Figure 4.2: Riboswitch ligand representation, training data comparisons, and feature extraction of sequence data. A) Ligand representation within the riboswitch training data (RS). 43 different ligands are represented with 10 ligands having greater than 2% representation in the data set. B) Length distributions for the sanitized 5’UTR and RS data set. C) KS distances between the 5’UTR and RS data set for all extracted features are shown in the bottom panel.

H. Sapiens 5'UTR sequences were pulled from UTRdb, a UTR database of multiple organisims' mRNA [216] on May 2022, and all analyses and computations use these data. UTRdb has since updated the original database to add additional curated functional annotations and UTRs [217]. This update does not substantially affect the sequence data and is not expected to affect any results presented here. For the readers' convenience, we reproduce the and make these data available through the GitHub repository, https://github.com/Will-Raymond/human_riboswitch_hits. Sequences were filtered for identical sequences and stored in a data set. 5'UTR sequences were matched with their corresponding coding regions from the *H. Sapiens* consensus coding region (CCDS) release 22 (accessed 11.28.2021). 5'UTR sequences missing CCDS information were discarded from the data set. 5'UTR sequences were appended with 22 nucleotides downstream from the start codon of the mRNA, and trimmed to the last 300 nucleotides in the 5' to 3' direction if the full 5'UTR sequence was over that limit. This $\min(5'\mathrm{cap}, 275\mathrm{NT})$ to start codon to 22 NT region was selected as the area to search for potential riboswitches as regulatory conformational changes in this region could directly block or expose the ribosomal initiation site. After the sanitation, CCDS matching, and length trimming, 48,031 5'UTR sequences were stored for subsequent examination. Fig. 4.2B shows the length distribution of both data sets, and Fig. 4.3A shows an example of a 5'UTR + 25 NT sequence.

For both the known riboswitch data set and the *H. Sapiens* 5'UTR sequences data set, sequences with incomplete or multiple base pair specifications were renamed to the first matching base pair out of the order A, C, G, T/U for the unknown character according to IUPAC naming conventions, Table 4.2 [218].

74 structure and sequence based features were extracted from the 5'UTR and RS data sets

In an effort to collect a broad spectrum of information for machine learning purposes, each sequence was processed to quantify 74 features in two groups: 65 sequence features and 9 predicted structural features.

The 65 sequence features include 64 length-normalized 3-mer (AAA, AAG, AAU ... CCC) frequencies, and the GC content. To define these, the 4^k sequence *k*-mers were generated for each

transcript, and the resulting 64-element vector was normalized by the total number of k-mers (i.e., length - 2) of the corresponding transcript [219, 220]:

$$[S_1, \ldots, S_{64}] = \left[\frac{N_{\mathrm{AAA}}}{L_{\mathrm{seq}} - 2}, \frac{N_{\mathrm{AAC}}}{L_{\mathrm{seq}} - 2}, \frac{N_{\mathrm{AAU}}}{L_{\mathrm{seq}} - 2}, \frac{N_{\mathrm{AAG}}}{L_{\mathrm{seq}} - 2}, \frac{N_{\mathrm{ACA}}}{L_{\mathrm{seq}} - 2} \ldots \frac{N_{\mathrm{GGU}}}{L_{\mathrm{seq}} - 2}, \frac{N_{\mathrm{GGG}}}{L_{\mathrm{seq}} - 2} \right] \tag{4.1}$$

In addition, the GC content is defined as the count of G and C within the sequence normalized by the sequence length:

$$S_{65} = \frac{N_{\mathrm{G}} + N_{\mathrm{C}}}{L_{\mathrm{seq}}} \tag{4.2}$$

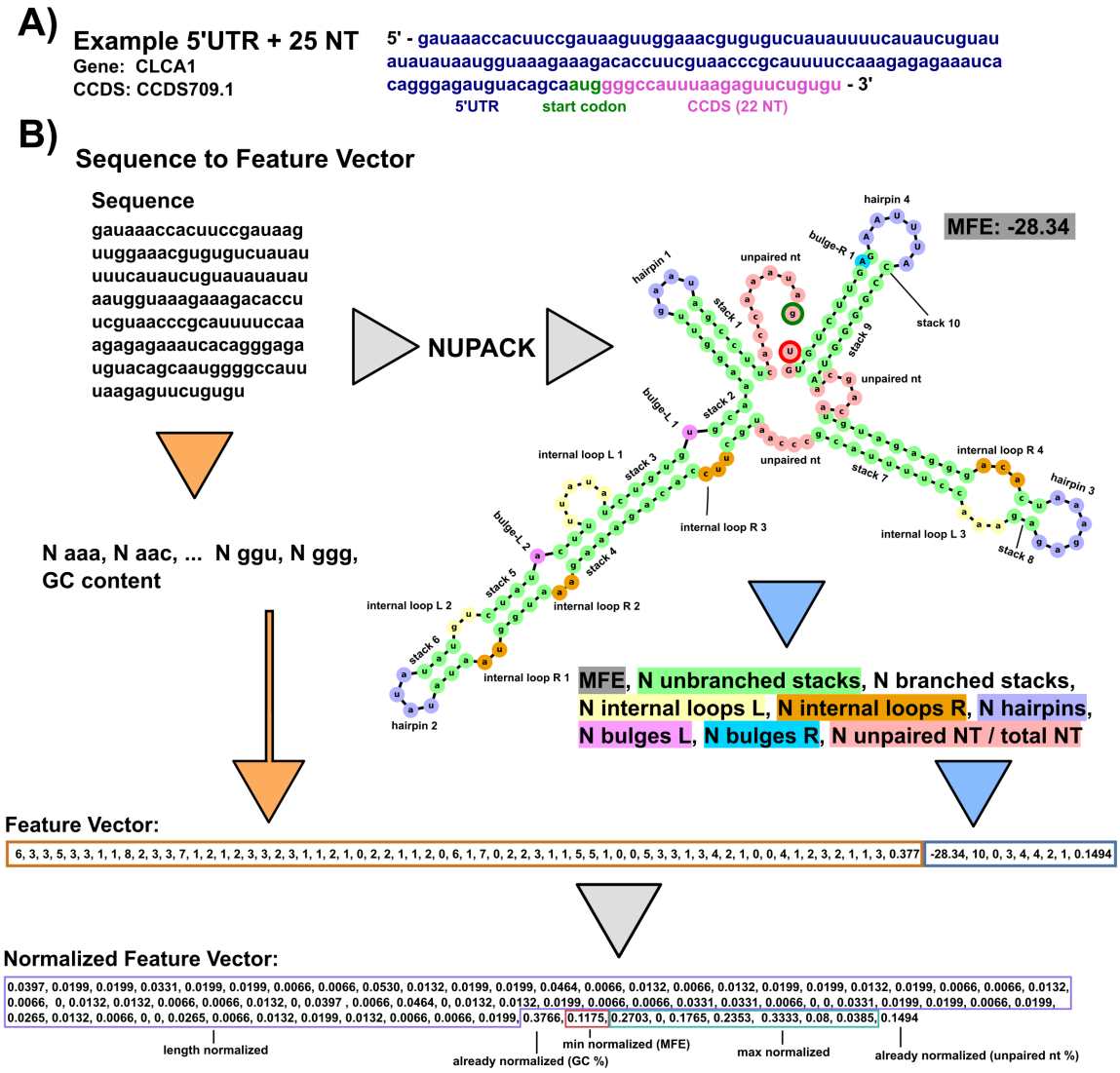


Figure 4.3: Caption on next page.

Figure 4.3: Example feature extraction of sequence data. A) Example 5'UTR sequence from the data set containing the start codon and 22 downstream nucleotides (25 total). B) Annotated example of taking an RNA sequence and converting it to a normalized feature vector for our positive-unlabeled learning. For sequence-based features, the sequence is converted into a 3-mer frequency and GC content is calculated. 3-mer frequency is normalized by the number of 3-mer subsets in the sequence (sequence length - 2). Secondary structure based features are generated by passing the sequence through NUPACK. MFE and structural features are extracted from the dot structure. Counts of hairpins, internal loops, bulges, and contiguous stacks (with and without branches) are extracted and max normalized across all the entire data set. Left (L) and right (R) designation corresponds to the 5' to 3' direction and 5' to 3' direction within a base pair stack respectively. MFEs are min-normalized across the data set. The final structural feature considered for learning is the percentage of unpaired nucleotides in the structure. The final output is a vector of length 74 normalized from 0-1.

Structural features were obtained by passing each sequence through the computational folding algorithm, NUPACK 4.0.0.23, [221, 222] to obtain a minimum free energy (MFE) secondary structure. NUPACK allows the user to specify a number of RNA strands within one set of "test tube" conditions and provides a list of commonly solved secondary structures and their mean free energies using a computational RNA model. Default NUPACK model settings were used when folding all sequences, "Model(material='rna', ensemble='stacking', celsius=37, sodium=1.0, magnesium=0.0)." For each sequence, the dot structure and the MFE of the most commonly folded non-complexed (no A:A) structure out of 100 RNA strands was saved and recorded as a sequence's secondary structure for feature extraction. The NUPACK MFE value, unpaired base pair percentage, and counts of how many consecutive stem base pairs in a branching stem or non-branching stem, number of stem loops, number of internal loops left and right, and numbers of left and right bulges were extracted and used as a "structural feature vector" defined as:

$$[S_{66}, \ldots, S_{74}] = [MFE, N_{\text{unbranched stacks}}, N_{\text{branched stacks}}, N_{\text{loops L}}, N_{\text{loops R}}, \cdots \\ N_{\text{hairpins}}, N_{\text{bulges L}}, N_{\text{loops R}}, \frac{N_{\text{unpaired NT}}}{L_{\text{seq}}}]$$

We note that psuedoknot features are not extracted or used for classification in this project as NUPACK dose not predict these structures.

Left and right for the bulges and loops were defined as "left"when residing on the 5' to 3' direction of a stem loop stack and as "right"when residing on the 3' to 5' direction of a stack. A bulge was defined as a one nucleotide unpaired on either direction interrupting a contiguous stack, loops were defined as two or more unpaired nucleotides interrupting a stack. This naming convention comes from the location when reading the dot structure left to right of the feature: "...((.(.....)))..." has one left bulge and "...((..((.....))...))..." has a left internal loop of two and a right internal loop of three. Unpaired nucleotides are defined as base pairs not within any paired stack, for example, "...(((....)))...(((....)))..." has 9 total unpaired nucleotides as the 8 unpaired nucleotides reside within a stack. The unpaired nucleotides inside stacks are instead labeled as hairpins. A branching stack is defined as one that has multiple distinct substacks within its stack, e.g. "((...((...))...((...))...))" is one branching stack containing two non-branching stacks. Counts of structural features were max-normalized by the entire combined RS and 5'UTR data set. Fig. 4.3B visually describes the process of taking an example sequence and converting it to its representative feature vector.

A two sample Kolmogorov-Smirnov distance was calculated to compare differences between the known RS features and *H. Sapiens* 5'UTR features. According to the KS distance, most structural features showed a marked disparity between the RS and 5'UTR data set, Fig. 4.2C, while sequence features ranged from 0.05 - 0.4 in their KS distance.

PU learning and cross validation achieves 75% to 99% accuracy to identify held out known riboswitches.

To assess performance of our positive-unlabeled classifiers, we generated 20 separate subsets of training and validation data by withholding specific subsets of the known riboswitches based on their class of ligand. The first ten validation subsets were generated by selecting each of the ten most represented ligand classes (each comprising 2% or more of the overall data) and leaving each one out: Cobalamin, guanidine, TPP, SAM, glycine, FMN, purine, lysine, fluoride, zmp-ztp. The next nine subsets were generated by leaving out pairs of the most commonly represented ligands: FMN+glycine, FMN+SAM, FMN+TPP, FMN+cobalamin, TPP+cobalamin, TPP+glycine, TPP+SAM, cobalamin+SAM, cobalamin+TPP. The final (and most diverse) validation set was cre-

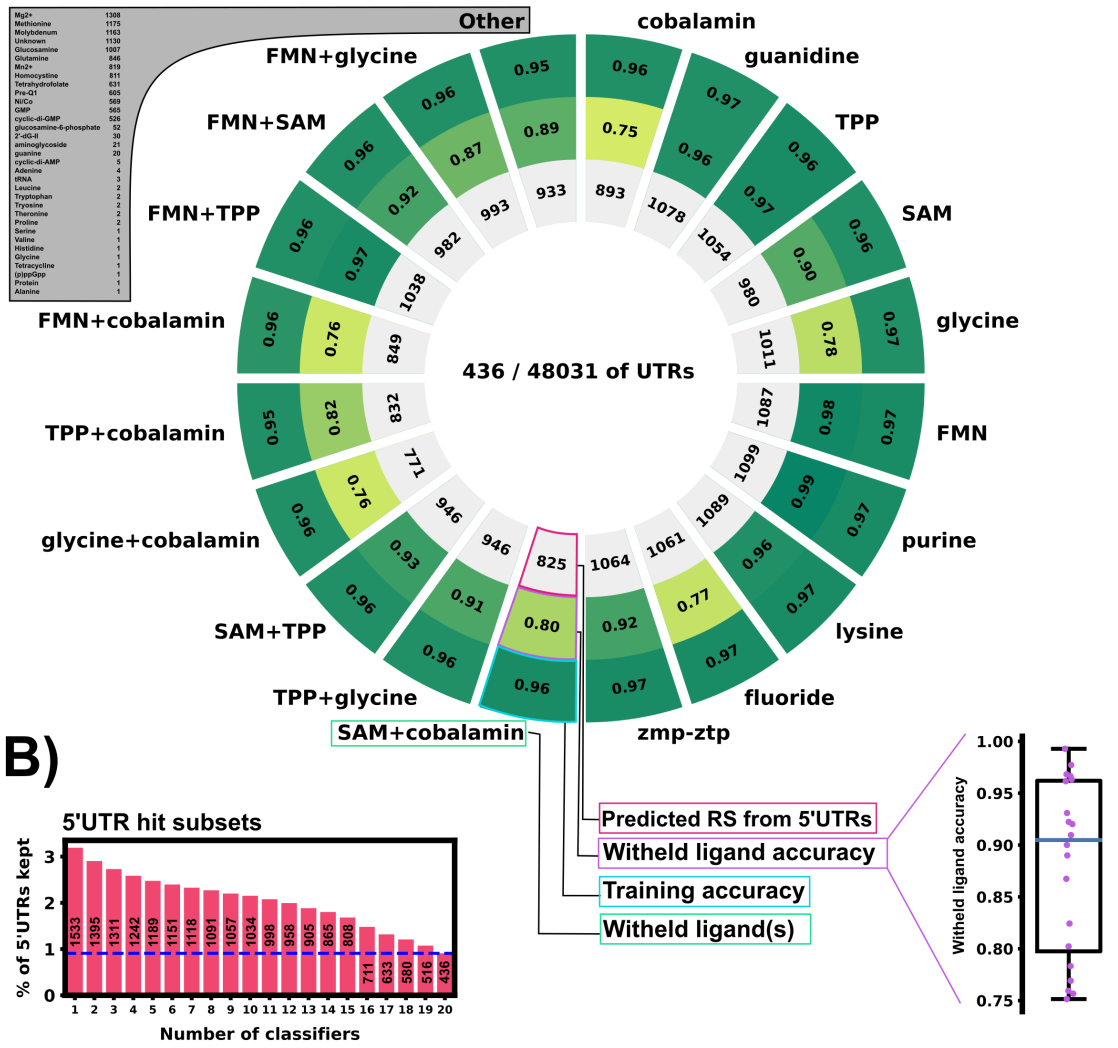


Figure 4.4: Training and validation results of 20 PU classifiers. A) Training and validation results. Each slice represents one PUlearn Elkanoto-classifier trained on a data set withholding one or two ligand-specific riboswitches. The outer ring shows the training accuracy on only positive examples (RS). The middle ring is the validation accuracy on the withheld riboswitch(es) of a particular ligand(s) class. The inner ring shows number of the predicted positive labeled 5'UTR sequences out of the 48,031 5'UTR sequences. The sub-panel on the bottom right shows the withheld validation accuracy (rounded to 2 digits) in a box plot. 436 5'UTRs were selected by all 20 classifiers as positive labeled —potentially harboring riboswitch-like features. B) 5'UTR hit subsets detected by varying numbers of classifiers (1 - 20, full sequences).

ated by selecting all riboswitch ligand classes with less than 2% representation in the full RS data set (11305 sequences, 34 ligand classes, 16.9% of the entire RS data set).

Validating each classifier on structural classes that were not provided gives a reasonable confidence that the classifier can extrapolate to riboswitches that are not included in the training set —the target task for eukaryotic riboswitch discovery. 20 Unweighted Elkan & Noto classifiers were trained on the 20 data subsets. Fig. 4.4A shows positive example training accuracy (outer ring), validation accuracy on withheld ligand sets (middle ring) and the positively labeled 5'UTR count (inner ring) of all 20 classifiers. Validation accuracy ranged from 75% to 99% across the classifiers. The classifier validated with withheld TPP riboswitches had high validation accuracy (97%), which is an encouraging sign as all currently described eukaryotic riboswitches use TPP as their ligand [223, 224, 225, 226, 227, 228]. The "other" classifier trained the diverse validation set of 34 ligand classes achieved an 89% validation accuracy, demonstrating a surprising ability to extrapolate to examples that are dissimilar from the core classes of riboswitches and are subsequently underrepresented in the training data. This classifier is the most likely to extrapolate correctly for new riboswitch discovery.

PU learning ensemble correctly identifies more than 50% of previously unseen synthetic theophylline riboswitches

As an additional verification step of our machine learning approach, we applied our ensemble of 20 classifiers to a wholly synthetic riboswitch data set, that is not represented anywhere within the training set. The training set used to train the ensemble included 14 total synthetic riboswitch sequences, none of which use theophylline as a target ligand. 25 current theophylline riboswitch sequences were obtained from Wang et al. [229] and were passed through our feature extraction and ensemble classification, Fig. 4.5. Fig. 4.5A shows the classification of the 25 theophylline riboswitches vs. a selection threshold on the ensemble probability output (ranging from .01 to .99). Upon testing, 56% (14/25) sequences were correctly identified with an ensemble probability over 50%, with 9 riboswitches classified higher than 90%. For a comparison, 300 completely random 35-250 nucleotide length sequences were generated and classified with the ensemble. 7.8% (39/300) of the sequences were falsely identified as riboswitch sequences. At very strict thresholds (requiring a $\geq$0.98 or $\geq$0.99 ensemble output), 25% of theophylline riboswitches are detected by

the ensemble, where as only 6.6% of random sequences are false positives. A two-sided binomial test was performed with the random sequence false positive rate to show the detection rate of theophylline riboswitches was significantly higher than random chance, Fig. 4.5B.

This finding indicates that our ensemble can correctly extrapolate to the synthetic riboswitches, although with an accuracy of 56%, it misses more of the synthetic riboswitches than it did for the cross validation performed above on natural sequences. One potential explanation for this loss in selectivity is that the aptamer for theophylline was discovered via directed evolution (SELEX) [230], potentially introducing a novel mechanism of action that is not captured by our original training data set. Overall, we consider the ability to find 56% hit rate for synthetic riboswitches to be another successful demonstration that the approach can identify potential riboswitches with novel structures or evolutionary origins.

PU learning with the trained ensemble model identifies and ranks a set of 1533 potential 5'UTR riboswitch hits

Now that the ML ensemble has been verified through cross-validation, it is instructive to examine which 5'UTR sequences have been identified as potential riboswitches. Among the classifiers, 436 5'UTRs were identified as harboring potential riboswitch elements by all 20 of the classifiers using a selection criteria of $\geq$0.95 classifier output. Fig. 4.4B shows the relative overlap of all 5'UTR sequences identified by one or more classifier with the same selection threshold. By contrast, the amount of 5'UTR sequences identified as riboswitches by one or more classifiers was 1533. The existence of an overlap when using all 20 classifiers instills confidence in our ensemble approach. If there was a precipitous drop in identified sequences when using more and more classifiers, that would imply that classifiers are individually identifying completely different subsets of the 5'UTR data set to consider as riboswitches. A drop from 1533 hits to 436 hits when increasing the amount of classifier agreement is substantial, but still leaves a large overlap found by all 20 trained classifiers.

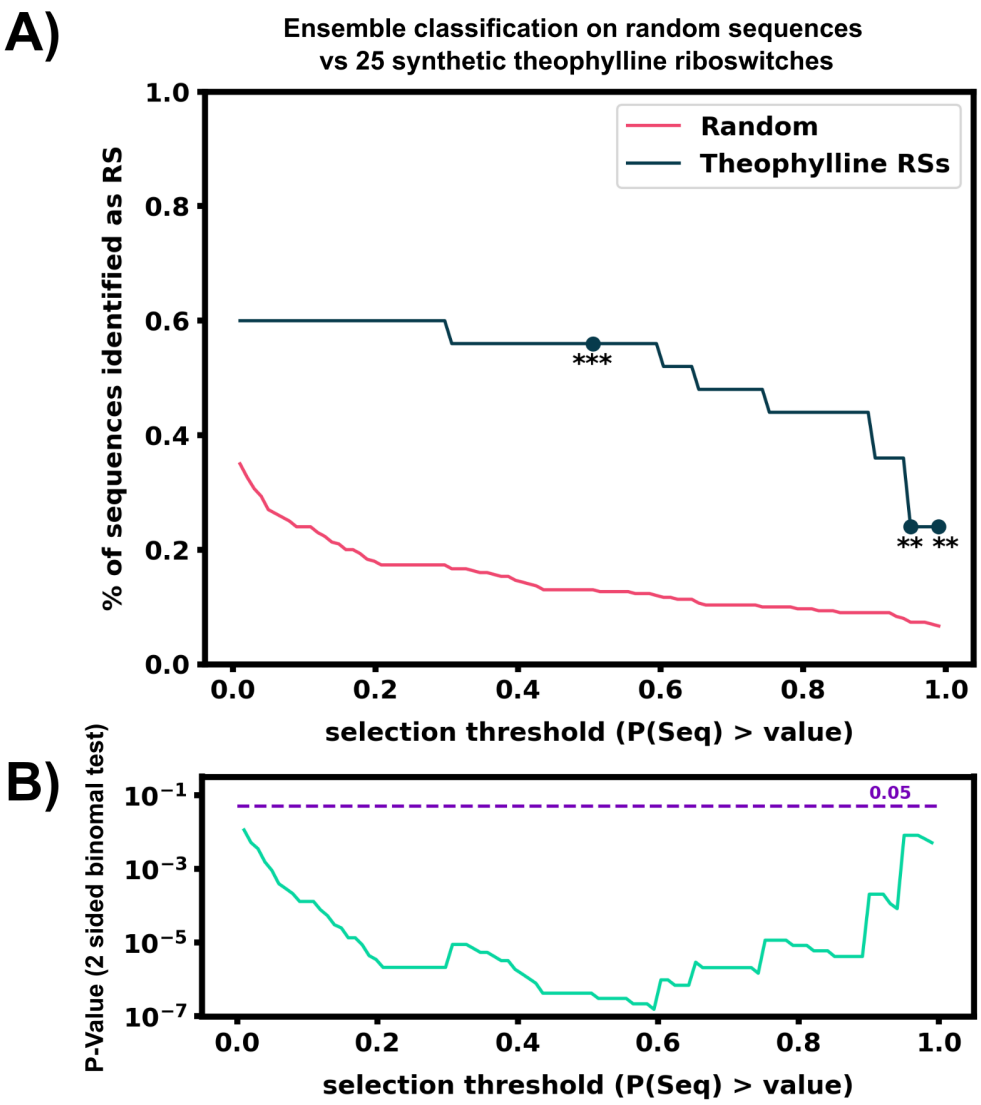


Figure 4.5: Application of the Ensemble model to a class of wholly synthetic riboswitches. A) 25 recently reviewed theophylline riboswitch sequences from [229] were obtained and classified using our 20-classifier ensemble. 300 random nucleotide sequences ranging from 30-300 nucleotides were also generated and classified using the ensemble. The percentage of sequences for both the 300 random sequences and 25 synthetic riboswitches are compared against each other for different positive ensemble probability thresholds. The ensemble incorrectly classifies 40% of the theophylline as non-riboswitches at any threshold. B) A p-value was obtained for every threshold using a two-sided binomial test of the amount theophylline hits vs. the false positive hits of the random sequences. The theophylline identified amount is significantly higher than the false positive rate (FPR) of the random sequence classification at every threshold. Select thresholds are highlighted in A:, .5, .98, and .99.

To provide a method to rank the 5'UTR hits based on the ensemble of the classifiers, we average the normalized outputs of the 20 PU classifiers to compute an ensemble similarity score, J_{Ensemble}:

$$J_{\text{Ensemble}}(UTR) = \frac{1}{20}\sum_{191}^{20} \frac{PU_{\text{i}}(UTR)}{\max(PU_{\text{i}}(All\ RS))} \tag{4.3}$$

where $PU_{\mathrm{i}}(sequence)$ is the positive-unlabeled classifier with the i^{th} withheld ligand set, which is normalized by its maximum value over all true RS.

The majority of top 5'UTR hits are consistently selected despite substantial truncation to their 5' ends.

Some potential 5'UTR hits are discarded in our analysis because we are using the full sequence data for the 5'UTR. Because many 5'UTRs can be large with multiple regulatory elements, such an approach could miss smaller riboswitch elements that are a sub-sequence and likely nearer to the start codon. To evaluate how many potential hits may be lost by not considering sub-sequences of the 5'UTR during classification, we took our 5'UTR data set and for each sequence, we generated 20 sub-sequences for each 5'UTR by truncating the mRNA at 20 evenly spaced locations upstream of the start codon, starting 30 NT from 5' end. For each sub-sequence, we extracted the new features and applied the ensemble classifier. Fig. 4.6A shows the resulting probability of selection as a riboswitch versus the fraction of the sequence used for all (48,031) 5'UTRs; Fig. 4.6B shows the same result but only for the subset of 436 5'UTRs (0.91%) that were previously identified as likely riboswitches using the full length sequence; and Fig. 4.6C shows the same result but for a distinct subset of 1210 5'UTRs (2.5%) that would have been identified as a riboswitch by five or more partial-length sub-sequences, but *not* using the full sequence. Although the probability that a given sequence being a riboswitch increases when using sub-sequences, the vast majority of 5'UTRs (97%) are still discarded as unlikely to be riboswitches. Moreover, for the 5'UTRs that were identified as a riboswitch using their full sequences (n=436), nearly half (45.5%) of these 5'UTRs are still detected as a riboswitch even when 85% of their sequence is discarded. For example, AUH is consistently detected as a riboswitch with $\geq$95% confidence for nearly every sub-sequence Conversely, a small fraction of 5'UTRs, such as ATF1, is only identified as a riboswitch when the sequence is 90% or more intact.

From a practical perspective, using the full ensemble and full sequences down-selects to more manageable number of potential hits, and given the ultimate goal of reducing the potential sequence

space to an experimentally viable number, 436 is considered to be acceptable amount for future experimental validation. However, the remaining hits from sub-sequences and ensemble agreement can be revisited and examined as needed and are provided within the supplemental data.

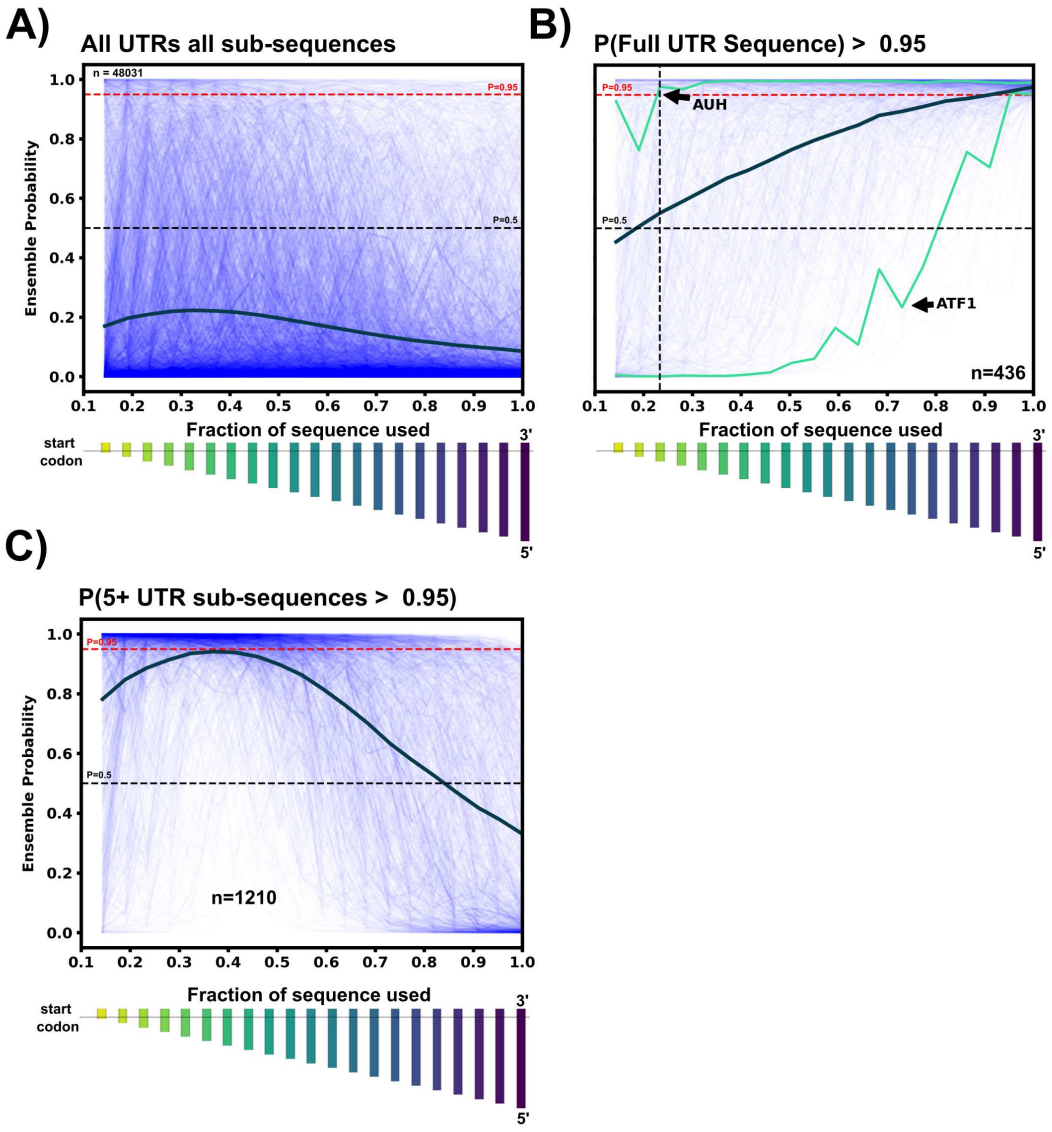


Figure 4.6: Caption on next page.

Figure 4.6: 5'UTR Sub-sequence exploration. A) For each 5'UTR sequence, 20 evenly-spaced sub-sequences were generated after the first 30 nucleotides in the 3'-5' direction, ensuring the start codon is in all sub-sequences. The relative size of each sub-sequence as a bar chart below the x-axis. For all 5'UTRs in the data set, variable-length sub-sequences were passed through the ensemble classifier to obtain the riboswitch probability. The riboswitch ensemble probability is plotted for each 5'UTR sub-sequence vs. the fraction of the sub-sequence to total 5'UTR length (thin blue lines). The thick dark blue line represents the average ensemble probability for that particular sub-sequence bin. C) Same as A, but only for the 5'UTRs whose full sequences were classified as $\geq$95% riboswitch by the ensemble. Many 5'UTR sequences such as AUH are classified as a riboswitch until almost 80% of the original sequence is removed. In contrast, some sequences such as ATF1 are no longer considered a riboswitch once 10% of the sequence is removed from the 5' end. Once again the thick dark line represents the average probability of each sub-sequence bin. C) To find sub-sequences not included in the 436 hits, 5'UTR sequences not detected as a riboswitch by the full sequence but were detected as $\geq$95% riboswitch in 5 or more sub-sequence bins were selected. These 1210 5'UTR sequences and their sub-sequence ensemble probabilities are plotted vs sub-sequence fraction. 1210 sequences could be included as potential riboswitch hits by removing some amount of 5' end nucleotides.

Identified 5'UTR hits share remarkable feature similarities to known riboswitches.

Now that our ensemble classifier has identified a subset of 5'UTRs that may harbor potential riboswitches, it is illustrative to examine the properties of these hits. It is infeasible to manually compare all 436 hits to the RS data set, so to aid with comparison, a GitHub page (https://will-raymond.github.io/human_riboswitch_hits_gallery/about/) was created to display, rank, and compare each 5'UTR to its most similar matches in the RS data set. The website contains the subset of 5'UTRs identified by all 20 classifiers as potential riboswitches, as well as any 5'UTR identified as a riboswitch by any individual classifier.

5'UTR to RS comparisons can be calculated several different ways, but for the purpose of the website, each 5'UTR was compared to each RS with a using a combination of three metrics: sequence length difference, structural feature vector mean-squared difference, and the predicted 5'UTR dot structure to RS dot structure Levenshtein distance (edit distance). Length mean squared distance was calculated as:

$$D_{\mathrm{L}} = |L_{\mathrm{UTR}} - L_{\mathrm{RS}}| \tag{4.4}$$

Likewise, the structural feature metric is also measured by the squared difference between any two extracted feature sets:

$$D_{\text{struct}} = \sum_{\text{i}}^{\text{74 features}} ([S_{5'\text{UTR}}]_{\text{i}} - [S_{\text{RS}}]_{\text{i}})^2 \tag{4.5}$$

Finally, the dot structure metric is measured by the Levenshtein distance or "edit" distance between two strings —in other words, how many edits (insertions, deletions, substitutions) to convert one string into another? Eq. 4.6 shows the recursive letter by letter formulation of the Levenshtien distance, where $\text{tail}(\text{string})$ refers to everything but the first letter of any given string. If we define *a* as the UTR sequences and *b* as the RS sequence, the Levenshtien distance can be computed as:

$$D_{\text{Lev}} = \begin{cases} \text{length}(a) & \text{if length}(b) = 0 \\ \text{length}(b) & \text{if length}(a) = 0 \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b) \end{cases} & \text{otherwise} \end{cases} \tag{4.6}$$

The combined similarity between the UTR and RS is denoted as J_{Sim} and is computed by normalizing each of the above-described distances by the corresponding maximum value for that metric over all RS in the data set. For length and Levenshtein distances, the maximum distances are: $D_{\text{Lev}} = \max((476 - L_{5'\text{UTR}}), L_{5'\text{UTR}})$, and $D_{\text{L}} = 476$, where 476 is the length of the largest RS sequence in the training data (all UTR sequences are truncated to 300 or less). For the structure feature distance, the normalization factor is obtained by comparing the 5'UTR in question to the entire RS data set and finding the maximum. The combined similarity score is then defined on a scale from 0 (no similarity) to 1 (perfect similarity) according to:

$$J_{\text{Sim}}(UTR, RS) = 1 - \frac{1}{3}\left(\frac{D_{\text{L}}}{\max(D_{\text{L}})} + \frac{D_{\text{Lev}}}{\max(D_{\text{Lev}})} + \frac{D_{\text{struct}}}{\max(D_{\text{struct}})}\right) \tag{4.7}$$

Each 5'UTR entry is displayed on the website alongside its top three J_{Sim} matches within the RS data set. The ligands of the top 20 5'UTR-RS J_{Sim} matches are also displayed as a preliminary prediction for the potential ligands for that structure. However, future experimental validation would be necessary to ascertain if these hypothetical matches are correct; the potential ligands are presented here more as an indicator of which ligand class from the prokaryotic data is most represented in J_{Sim} matches to the 5'UTR sequence.

For the reader's convenience, the website table displaying all hits can be sorted by its similarity to known RS (J_{Sim}) or by the ensemble probability from the PU classifier (J_{Ensemble}, see section 2.5). An example website page is presented in Fig. 4.7. Each column represents a 5'UTR hit (far left) or an RS entry from the training data (next 3 columns). For each sequence, the predicted secondary structure from NUPACK is displayed for visual comparison. Base pair comparisons are also shown for each 5'UTR to RS pair with a circle plot of both predicted structures overlayed. Below that is a comparison of each sequence's secondary structure features (stacks, loops, hairpins, etc). The counts of each of these secondary structure features and the sequence dot structures are provided in tables on the page as well. The goal of the website is to provide the reader with an instant visualization of each 5'UTR pair hit.

Gene ontology points to enrichment of downstream *H. Sapiens* proteins associated with small molecules and transcription / translation regulation

A predominant function of bacterial riboswitches is to regulate the proteins directly related to the riboswitch's target ligand. For example, a fluoride riboswitch may turn on genes useful for processing or mitigating fluoride for an organism [183, 180]. With this in mind, it is informative to examine the downstream proteins from the *H. Sapiens* 5'UTR hits for correlations in protein function, looking for genes associated with processing or synthesizing small molecules. Gene ontology (GO) analysis was preformed on the list of 5'UTR hits to look for any cellular function or process enrichment using the PANTHER database [231, 232, 233]. GO process results are shown in Fig. 4.8. The process ontology with significant fold enrichment fell into the following categories: Chromatin remodeling, transcription / translation regulation, mRNA splicing, mRNA and

rRNA modification, and mitochondrial ubiquinone synthesis. These enrichment results suggest a potential for small molecules to play a regulatory role in gene regulation, even if we cannot comment fully on our 5'UTR hit list without experimental validation. Interestingly, proteins directly involved in chemical stimulus detection were "unenriched"with no proteins found at all. This observation of mutual exclusion between potential riboswitches and sensing proteins is also reasonable - if there are already proteins capable to sense and respond to their intended stimulus, then there is no need to execute redundant functions in riboswitches.

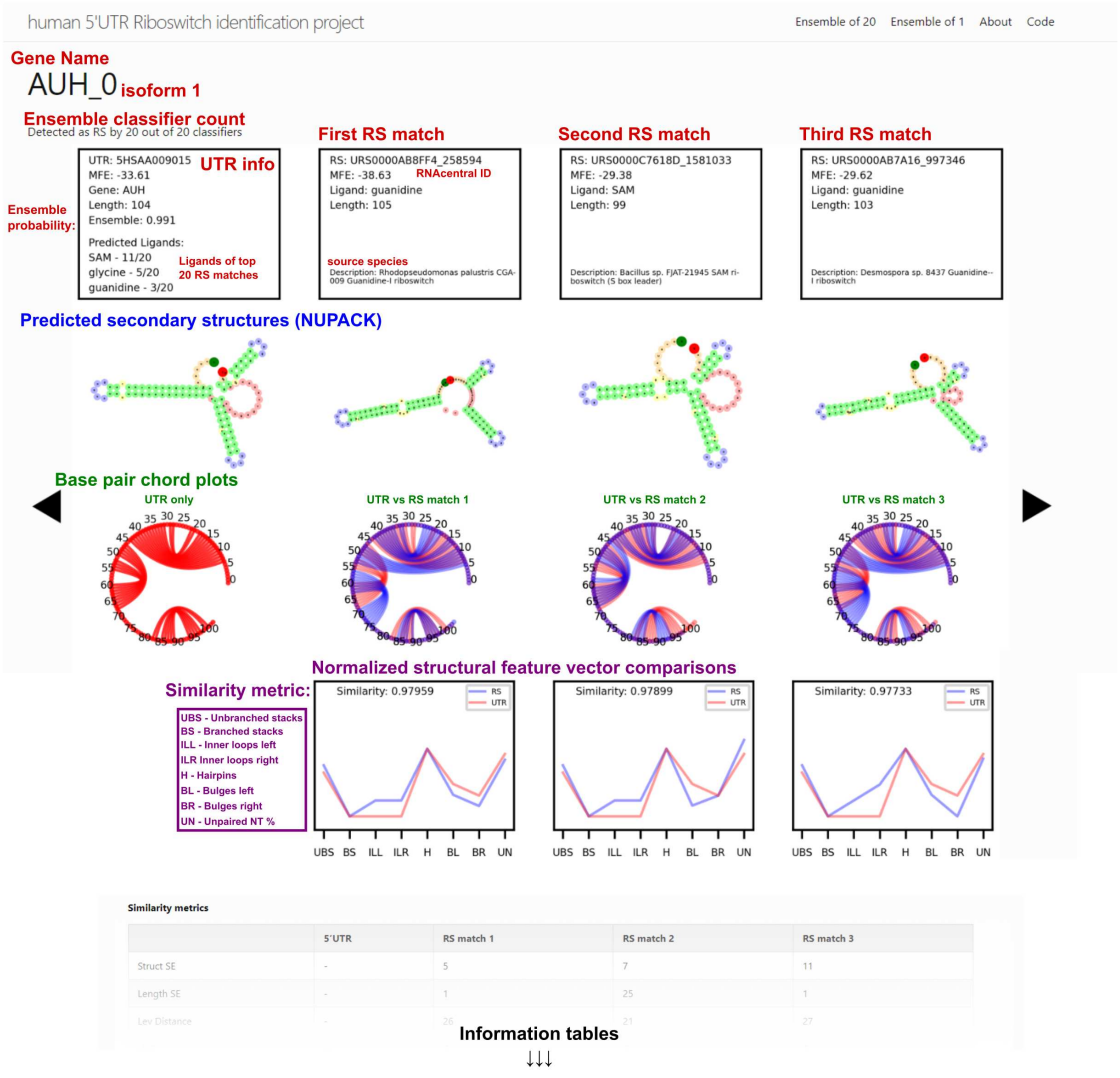


Figure 4.7: Caption on next page.

Figure 4.7: Example 5'UTR hit display from the Github Pages website. The RS-UTR display website can be found at https://will-raymond.github.io/human_riboswitch_hits_gallery/_mds/AUH_0/. The website provides information on a given 5'UTR detected by the ensemble as riboswitch. Alongside each 5'UTR sequence, information on the top three riboswitch J_{Sim} matches to the 5'UTR are displayed in each column. First row provides information on a given sequence, UTRdb or RS id, source species, MFE of the predicted structure, and ensemble prediction probability for the 5'UTR. The next row displays the NUPACK predicted secondary structure for each sequence. Below that are chord plots representing the bonded base pairs for each RS sequence overlapping the 5'UTR chord plot. The next row shows the normalized structural feature vector comparison for structure counts for the 5'UTR and a given RS. J_{Sim} is reported in these plots. Additional information such as the dot structure, origin sequence, and counts of structural features are presented in the table below the comparison plots.

GO function analysis showed a significant enrichment of downstream proteins implicated in binding various small molecules: nucleotides, nucleosides, ubiquinone, and various cyclic compounds, Fig. 4.9. RNA binding and nucleotide binding molecules were extremely enriched with p-values ranging from 10^{-7} to 10^{-17}. Notably there is a negative enrichment of G protein-coupled receptors, this could be explained by riboswitches having direct signalling activity to a cell, bypassing typical trans-membrane signalling pathways.

Some potential mRNA with potential riboswitches encode proteins with direct involvement in small molecule functions

Although we have no experimental validation for our computationally discovered 5'UTR hits, it is illustrative to highlight and comment several interesting matches found in our computational analysis. Several computational hits such as AUH and FTSJ1 have roles in processing already known ligands. AUH plays a critical role in leucine degradation, hydrating 3-methylglutaconyl-CoA to 3-hydroxy-3-methyl-glutaryl-CoA [234]. FTSJ1 is known to bind directly to S-adenosylmethionine (SAM) with a potential role in modifying ribosomal RNA and tRNAs [235].

Ubiquinone (CoQ10, CoQ) is an essential antioxidant component of the mitochondria [236, 60]. Many proteins directly related to Ubiquinone synthesis or binding were selected as harboring riboswitches: Biosynthetic proteins COQ3, COQ9, COQ7; Several respiratory complex I proteins: NDUFA2, NDUFA8, NDUFB6, NDUFB9, NDUFS1, and a respiratory complex III protein: UQCRQ. No ubiquinone binding riboswitches are currently described in the literature, although

riboswitches binding other critical components of the electron transport chain such as NAD+ have been described [237, 238]. Our overrepresentation of riboswitches within the mammalian mito-

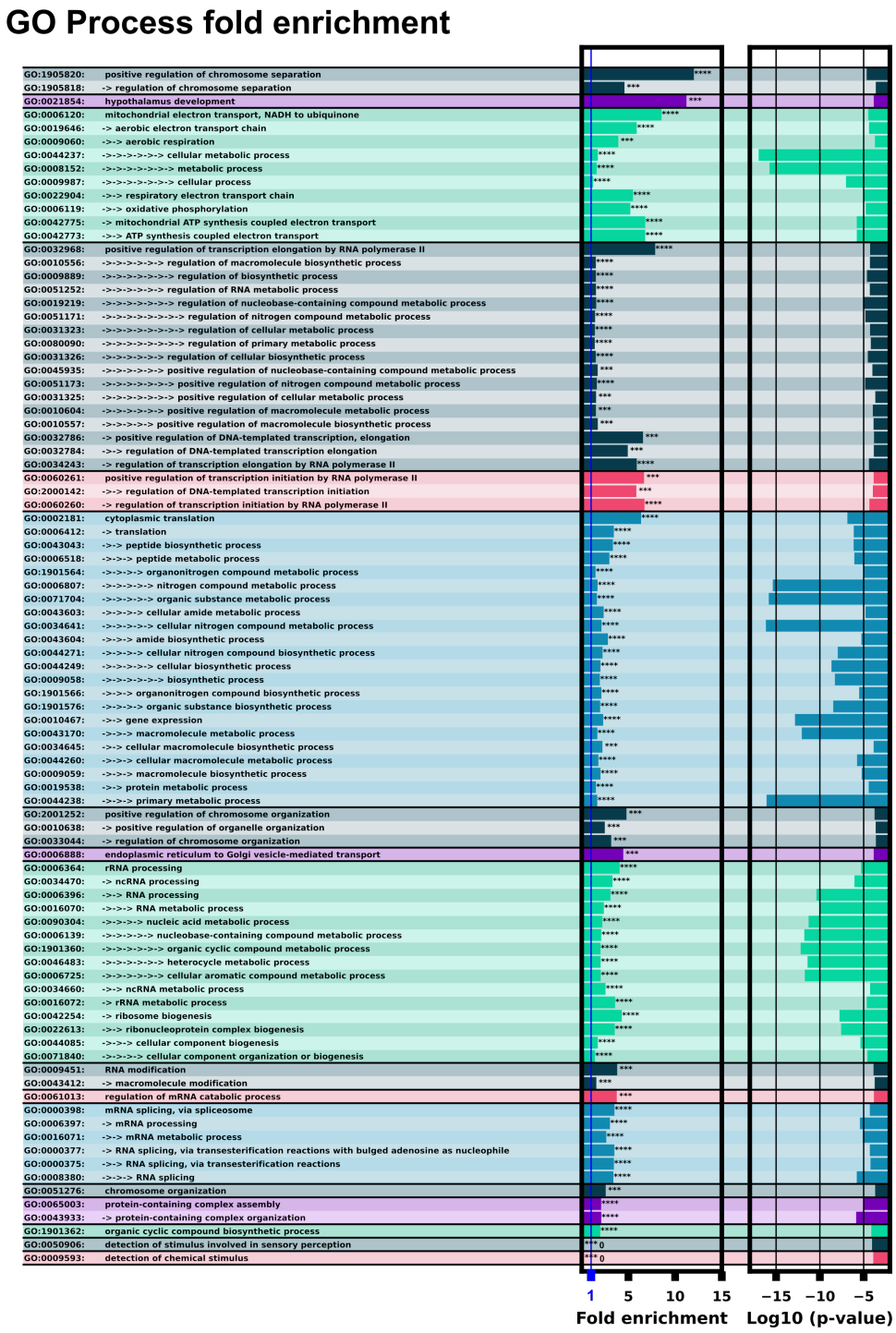


Figure 4.8: Caption on next page.

Figure 4.8: GO process analysis with ID's and terms. The left column lists the GO ID and term. Multiple arrows indicate GO term sub-levels. The left bar chart shows fold enrichment for that GO term with significance indicator. The second bar chart shows the log space of P-value significance for each enrichment.

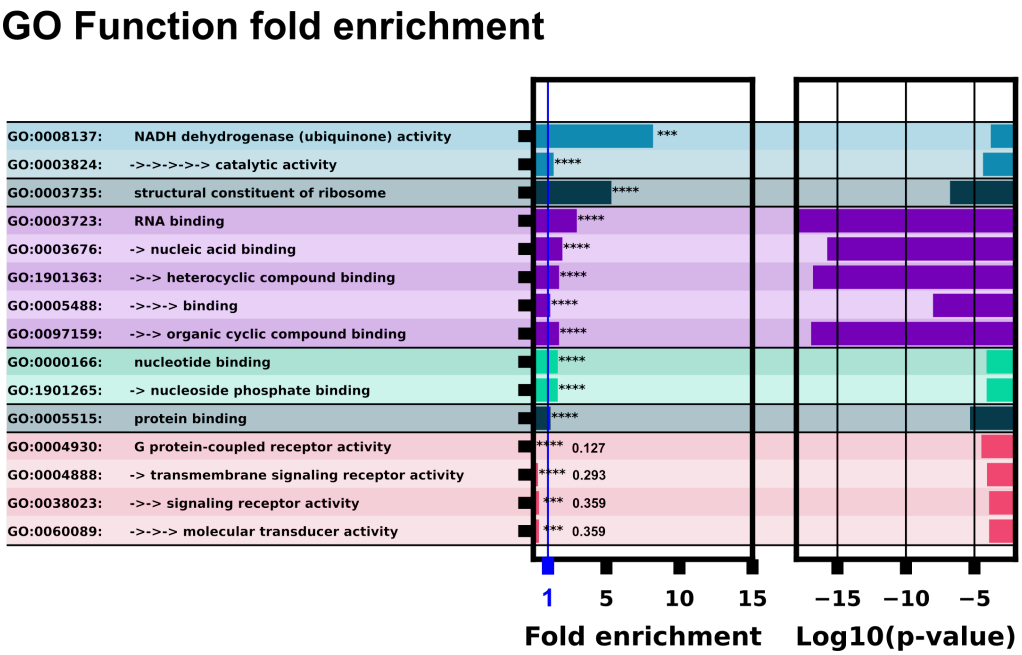


Figure 4.9: GO function analysis with ID's and terms. The left column lists the GO ID and term. Multiple arrows and indents indicate GO term sub-levels. The left bar chart column shows fold enrichment for that GO term with significance indicator. The second bar chart shows the log space of p-value significance for each GO term.

chondria aligns with theories put-forth within Venkata Subbaiah et al. [188], where they speculate that the mitochondria's reduced genome may still harbor RNA switches as a mechanism as translational control.

Some proteins represented in the hit list are implicated in small molecule or amino acid synthesis or processing. GSS is responsible for the second step of glutathione synthesis [239], PNPO directly converts vitamin B6 into its active form [240].

Finally, several close matches in predicted structure and feature vectors should also be noted: ZNF480, SPAG11B, UBAP2L-0, and TTPAL; However, to our knowledge, these proteins have no clear relation to small molecule processing.

4.1.4 Conclusion

We have trained an ensemble of machine learning riboswitch classifiers using leave-one-out cross validation of ligand classes consisting of 20 individual classifiers using sequence and predicted RNA structural features. Using this ensemble classifier, we identified a subset of the *H. Sapiens* 5'UTR predicted to harbor riboswitch-like elements (Fig. 4.4). This subset provides experimentalists with a prioritized list of sequences and genes to examine first when designing exploratory experiments. This 436 sequence subset additionally shows positive GO fold enrichment results for downstream genes in many processes with direct small molecule involvement (Fig. 4.8 and 4.9). Our approach provides a complementary strategy to the that taken in Mukherjee et al. [200], where the authors began with a known riboswitch sequence and selectively mutated nucleotides while preserving structure, and then searched genomics data for a sequence match. In contrast, our approach starts with genomics data to learn our classifier and any given sequence can be assessed for riboswitch probability. While our approach may be less targeted to the discovery of a riboswitch with a particular structure, it holds the potential to extrapolate beyond single specific structures, as exemplified by its identification of known synthetic riboswitches (Fig. 4.5). Searching for a riboswitch structure in a branch of life vastly different than where riboswitches are previously described likely needs this extrapolation ability. In future work, our approach could be replicated using the *H. Sapiens* 3'UTR, where described *H. Sapiens* pseudoriboswitches have been found. However, for the purposes of this paper, we have limited our search to the 5'UTR because the bulk of our training data (Bacterial riboswitches) act near their ribosomal binding site, equivalent to the 5'UTR, and looking in the *H. Sapiens* 5'UTR first gives the best chance for efficient machine learning extrapolation. Our approach could also be applied to other eukaryotic UTRs when experimentally validated. In this paper, we also briefly explored the how the detection of potential 5'UTRs riboswitches could be expanded by varying how much of the 5'UTR sub-sequence is used for the identification. This extended list from sub-sequences could could be useful for finding better candidate sequences should the 436 5'UTR hits not bare fruit. Because this is at present an entirely computational investigation, we are unable to conduct experimental

validation for any of the detected hits, but we provide this list to the greater scientific community in the hope that these potential targets could kick start the discovery of a riboswitch within the *H. Sapiens* translatome and beyond.

4.1.5 Materials and methods

Computation

All processing was done in Python 3.8 with Biopython [241], NumPy [242], and NUPACK 4.0.0.23. Final data was stored in .csv, .npy, and .json files and large files can be recomputed by the reader with the analysis notebook. The data files are available at `https://github.com/Will-Raymond/human_riboswitch_hits` and the entire project computation (sanitation, feature extracting, training, analyses) notebook is available at: `https://colab.research.google.com/drive/17zmKJh8iHAC2tImNNSyBrwUpU0uYKefx?usp=sharing`.
A modified BEAR encoding in Python was used for structural feature counting from dot structure strings [243].

Positive Unlabeled Machine Learning

Unweighted Elkan & Noto classifiers were used for our machine learning classifiers. In brief, this is an extension of a generalized probability classifier to train on unknown / known labels as an approximation of class labels. Each data point x has a label y which is either *0* or *1*. Along with the label pair each data point has a known or unknown flag, s, where $s = 1$ if the data point is known, and $s = 0$ when unknown. Therefore, when $s = 1$, $y = 1$ and when $s = 0$, $y = \{0, 1\}$ Any binary classifier is then used to estimate $p(s = 1|y = 1, x)$ instead of the classical estimate of $p(y, x)$. For our paper we used a SVC classifier from sci-kit learn with the following options: SVC(C=10, kernel=rbf, gamma=0.4, probability=True). All PU classifiers were made using an implementation from the Python package PUlearn. For full details refer to the original paper by Elkan & Noto [203].

GO Analysis

GO analysis was performed with the PANTHER overrepresentation test (release 10.13.2022) using the 07.01.2022 release of the PANTHER database (10.5281/zenodo.6799722Released2022-07-01) using the Fisher's Exact test with False Discovery Rate correction. The reference list used for comparison analysis was the *Homo Sapiens* gene list. Overrepresentation test was performed for the GO biological process complete and GO biological function complete annotated data sets.

Conflict of Interest Statement

The authors declare the absence of any commercial or financial relationships that could be construed as a conflict of interest for this research.

Contributions

William S. Raymond: Conceptualization, Implementation, Computation, Writing, Editing

Jacob DeRoo: Conceptualization, Editing

Brian Munsky: Supervision, Editing

Data Availability

All original data files, sanitized data files, processed feature data, PU classifiers, Figure files, and ensemble results used to create this manuscript are stored at: https://github.com/Will-Raymond/human_riboswitch_hits. Any files too large to store on the manuscript repository can be regenerated with the analysis notebook. Upon final acceptance for publication, a release containing all data and computational analyses for this manuscript is available at figshare at 10.6084/m9.figshare.24587334.

4.1.6 Tables

Table 4.1: Ligand representation within the data set

Ligand	Count	%	Ligand	Count	%	Ligand	Count	%	Ligand	Count	%
Cobalamin	15718	23.2	Molybdenum	1163	1.7	2'-dG-II	30	0.044	Histidine	1	1.5e-5
TPP	12459	18.4	Unknown	1130	1.7	aminoglycoside	21	0.031	Protein	1	1.5e-5
SAM	8686	12.8	Glucosamine	1007	1.5	guanine	20	0.029	Glycine	1	1.5e-5
Glycine	4835	7.1	Glutamine	846	1.2	cyclic-di-AMP	5	7.4e-5	Tetracycline	1	1.5e-5
FMN	4255	6.3	Mn2+	819	1.2	Adenine	4	5.9e-5	(p)ppGpp	1	1.5e-5
Purine	2648	3.9	homocysteine	811	1.2	tRNA	3	4.4e-5	Alanine	1	1.5e-5
Lysine	2318	3.4	tetrahydrofolate	631	0.9	Leucine	2	3.0e-5	Serine	1	1.5e-5
Fluoride	1975	2.9	Pre-Q1	605	0.9	Tryptophan	2	3.0e-5			
zmp-ztp	1841	2.7	Ni/Co	569	0.8	Tyrosine	2	3.0e-5			
Guanidine	1640	2.4	GMP	565	0.8	Proline	2	3.0e-5			
Mg2+	1308	1.9	cyclic-di-GMP	526	0.8	Theronine	2	3.0e-5			
Methionine	1175	1.7	glucosamine-6-phosphate	52	0.078	Valine	1	1.5e-5			

Table 4.2: Nucleotide substitution for data sanitation

IUPAC nucleotide code Base(s) Converted to	Base(s)	Converted to
A	Adenine	A
C	Cytosine	C
G	Guanine	G
T / U	Thymine or Uracil	U
R	A or G	A
Y	C or T	C
S	G or C	G
W	A or T	A
K	G or T	G
M	A or C	A
B	C or G or T	C
D	A or G or T	A
H	A or C or T	A
N	Any	A

Chapter 5

Concluding remarks and extracurriculars

This concludes the corpus of published work I have done here at Colorado State University within Dr. Brian Munsky's lab. I have contributed to five substantial papers, three second authorship papers, one first authorship and one first authorship out for review. If I had infinite time at Colorado State University, some of the future directions my research would head are expanded upon below. Hopefully, a future graduate student will extend this work.

5.1 Extensions from simulated NCT gene classification to model classification

The experiment design pipeline in Chapter 2.3 has not been used to classify real experimental data yet. In the future we wish to verify some of its capabilities by actually using it in a lab setting. There's also a never-ending list of features we would like to add if given time. One of the key features we wish to add was to switch from mRNA species classification, to model based classification. We would like to generate simulated data from two NCT experiments where the underlying mRNAs had differing mechanistic models, train classifiers on the simulated data, and see which models are differentiable and with which experimental conditions. We are curious how the "model classifiers"would do when applied to real NCT data. We are intensely interested in training a cohort of various codon dependent model classifiers and applying that cohort to real NCT traces. The classifier cohort could then tell us which codon dependent behaviour it thinks the real data matches the best. The models selected and rejected would give us information on future experiments regarding codon dependence and selection.

5.2 Extension of rSNAPed to examine other modes of mRNA diffusion

We greatly desire the simulated cells from the rSNAPed pipeline to be more realistic in the future. One of the larger extensions we have upcoming is the ability to change the diffusion be-

haviour of the simulated mRNAs within the cell. All work shown in Chapter 2.3 was done using a default Brownian motion limited within the cell masks. However, we have not verified this motion captures what we see in our microscopes. Upcoming rSNAPed versions have several different options implemented for changing diffusion coefficients over time for each mRNA molecule. Some example diffusion modes we could have used for the rSNAPed are shown in Fig. 5.1. mRNAs could have set diffusion coefficients drawn from some underlying distribution, like the power law distribution or Gaussian distribution shown in Fig. 5.1A and B. This imparts a heterogeneity to spots' movement across the cell background environment. More complicated functions of time can also be used to change the diffusion coefficient over time such as in Fig. 5.1C and D. mRNAs in Fig. 5.1C diffuse in two states, a heavily limited state with a diffusion coefficient of 0.1 pixels^2/second, and a super-diffusive state of $D = 4$ pixels^2/second. Each mRNA can randomly state transition from either state, leading to "jumps"across the cell with transient periods where their motion is limited. Diffusion can also be made a function of the actual translational state of the mRNA. In Fig. 5.1D, mRNAs calculate their molecular mass at every time point in the simulation; As a ribosome binds its mass is added, probe + fAB mass is added past every epitope location, and the amino acid incorporation continuously adds a small amount of mass per codon. For this simulated cell the following formula was used to convert from molecular mass to diffusion coefficient:

$$D = \frac{k_{\mathrm{B}} * T}{6\pi MW^{.333}\eta} \tag{5.1}$$

The Hydraulic radius of the mRNA is approximated with $MW^{.333}$ and a literature value of viscosity, η, 4.4x10^2 $Pa{\cdot}s$ [244] is used. Note that this is an approximation of the cell's cytoplasmic viscosity and assumes the same viscosity across the entire cell. The example provided here does not change its diffusion coefficient substantially over time, however, when coupled with mRNA bursting on and off an extreme variation can be simulated with a particular mRNA's diffusion coefficient as the majority of the molecular weight added is with the binding ribosomes. We can also extend even further to include different spatial viscosities to account for cellular compartments

and organelles such as the ER. We will likely have to move to more complicated diffusion models if we wish to train models for identifying various mRNA states such as ER bound or free diffusing using rSNAPed.

5.3 Extended modeling capabilities of rSNAPsim

One of the most appealing future directions using the capabilities of the rSNAPsim's model maker is utilizing it to explore translation under non-canonical tRNA conditions, such as over or under expression of specific tRNA isodecoders, loss of particular tRNA species, or tRNA modifications. Non-canonical tRNA states have been shown to have prolific, varied, and poorly understood mechanisms that result in disease states and deregulation of the cell, and rSNAPsim is capable of modeling these dynamics at a single codon resolution and providing these models quickly for a user. Upregulation of particular codon isodecoders has been linked to metastatic cancers [245]. Upon oxidative stress, specific tRNA will fragment into tRFs (tRNA-derived small RNA fragments) signalling tRNA precursor depletion and subsequent down regulation of proliferation related genes [246]. Total tRNA sequencing methods have been developed within the past few decade, but many still suffer from limitations due to the difficulty of the sequencing required [247, 248]. Translation errors during PolyQ diseases may be due to tRNA species depletion during their characteristic long repeat regions [249, 250, 251, 252]. tRNAs are encoded by a large amount of repeate genes which may create identical or similar tRNA species inside a given cell - which are then heavily post transcriptionally modified. Additionally there are tRNA derived fragments such as tiRNA and tRF, as well as distinctly different precursor tRNAs from fully mature active tRNA, both of which add to the difficulty of quantification. Given the tRNAomes importance, the rSNAPsim can be used to provide models to interrogate some of these misregulated disease states.

We can propose the following "tRNA pool TASEP" model. We can take the model described in Chapter 2.1 and Aguilera 2019 [50] and adjust it where instead of a variable elongation rate for each codon proportional to tRNA gene copy number. This new model directly considers a

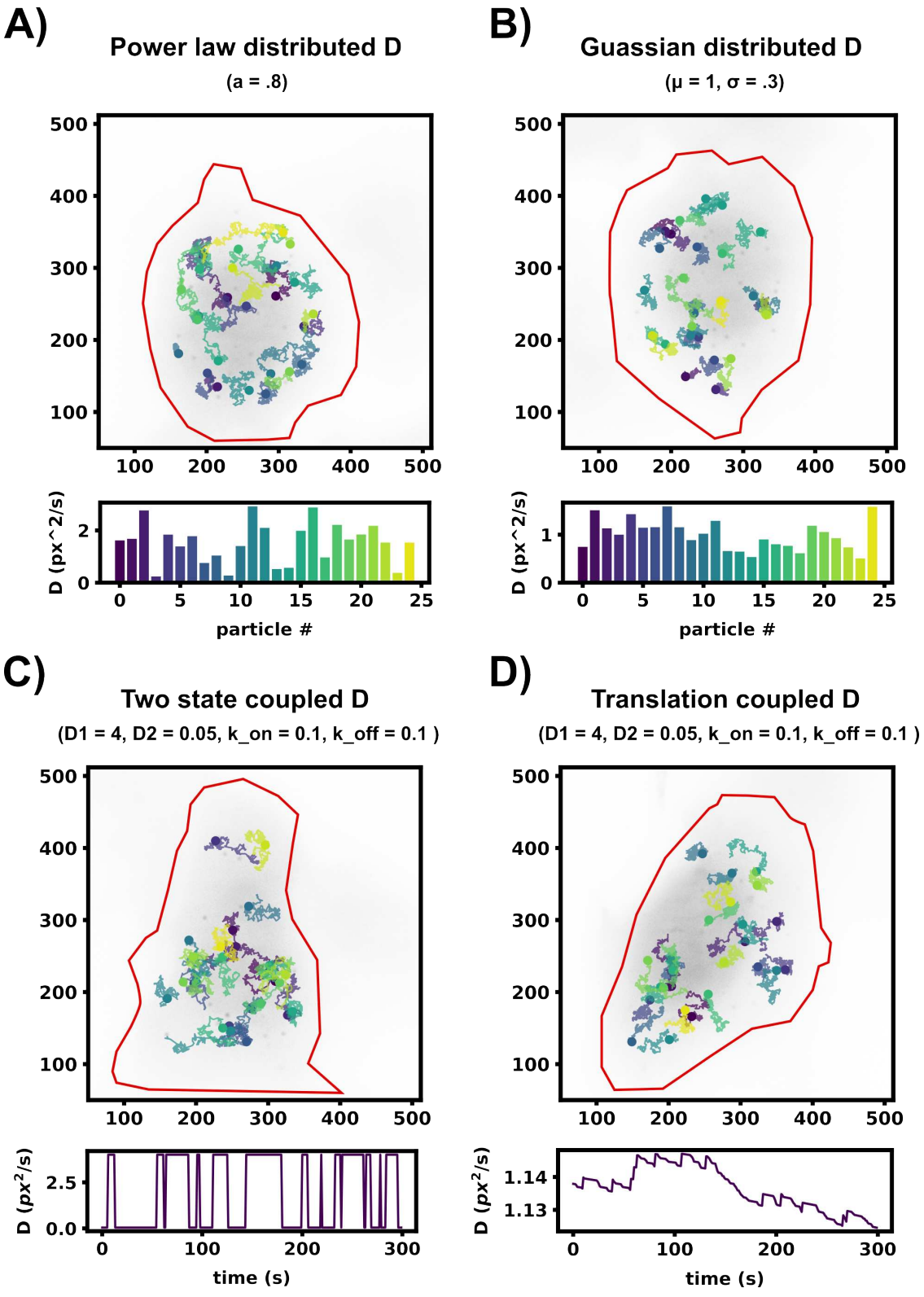


Figure 5.1: Caption on next page.

Figure 5.1: Examples of various diffusion modes in rSNAPed. rSNAPed can simulate multiple diffusion modes or models for mRNA motion. Diffusion coefficients can be passed to rSNAPed for each individual mRNA over time. A) an mRNA diffusion coefficient of 3 was scaled by a power law distributed random number ($ax^{(\mathrm{a}-1)}$ for $a = 0.8$) for each mRNA spot. Highlighted spots' diffusion coefficients are shown in the bar plot below the simulated cell. A) Each mRNA's diffusion coefficient was drawn from a Gaussian distribution for this simulated cell ($\mu = 1$, $\sigma = 0.3$). Highlighted spots' diffusion coefficients are shown in the bar plot below the simulated cell. C) Diffusion coefficients can be a function of time, in this simulated cell each mRNA spot has two possible states, one with a high diffusion coefficient of 4, and one with a low diffusion coefficient of 0.1. Each mRNA transitions from each state with an equal k_{on} and k_{off}. A single spot's diffusion coefficient over time is plotted in the plot below the simulated cell. D) Complicated functions of time can also be incorporated. In this simulated cell, mRNA diffusion is coupled to its translational state, each ribosome adds the weight of itself and its growing nascent peptide and fluorophores as it translates, corresponding to a slower in diffusion coefficient when mRNAs are heavily translated. A single spot's diffusion coefficient over time is plotted in the plot below the simulated cell.

local "resource pool"of tRNA molecules. Ribosomes will now elongate at a proportional rate to the amount of tRNA isodecoders locally available for the current codon they are reading. Moving forward one codon will consume the corresponding tRNA from the tRNA pool. The tRNA pool is replenished with charged tRNA molecules at controllable rates (proportional to human tRNA gene copy number as a default) and leaving the vicinity of the translating mRNA at a set diffusion rate. Ribosomal exclusion is still considered in this model, and for now, decoding errors are not considered (but can and will be added later!). This robust model can now account for mRNA dynamics such as repeat regions that may deplete localized tRNA resources and introduces long range ribosome interactions across the mRNA where translation at one end of a ribosome may deplete resources for far afield ribosomes.

5.3.1 Model Description

The mRNA transcript is modeled as a set of reactions from 1 to L for length of the with a ribosomal occupancy at each i^{th} position of x_{i}.

$$\emptyset \xrightarrow{w_0(x_1,\ldots,x_{\mathrm{nf}})} x_1 \xrightarrow{w_1(x_1,\ldots,x_{\mathrm{nf}+1})} \ldots x_{\mathrm{i}} \xrightarrow{w_{\mathrm{t}}} x_{\mathrm{L}} \tag{5.1}$$

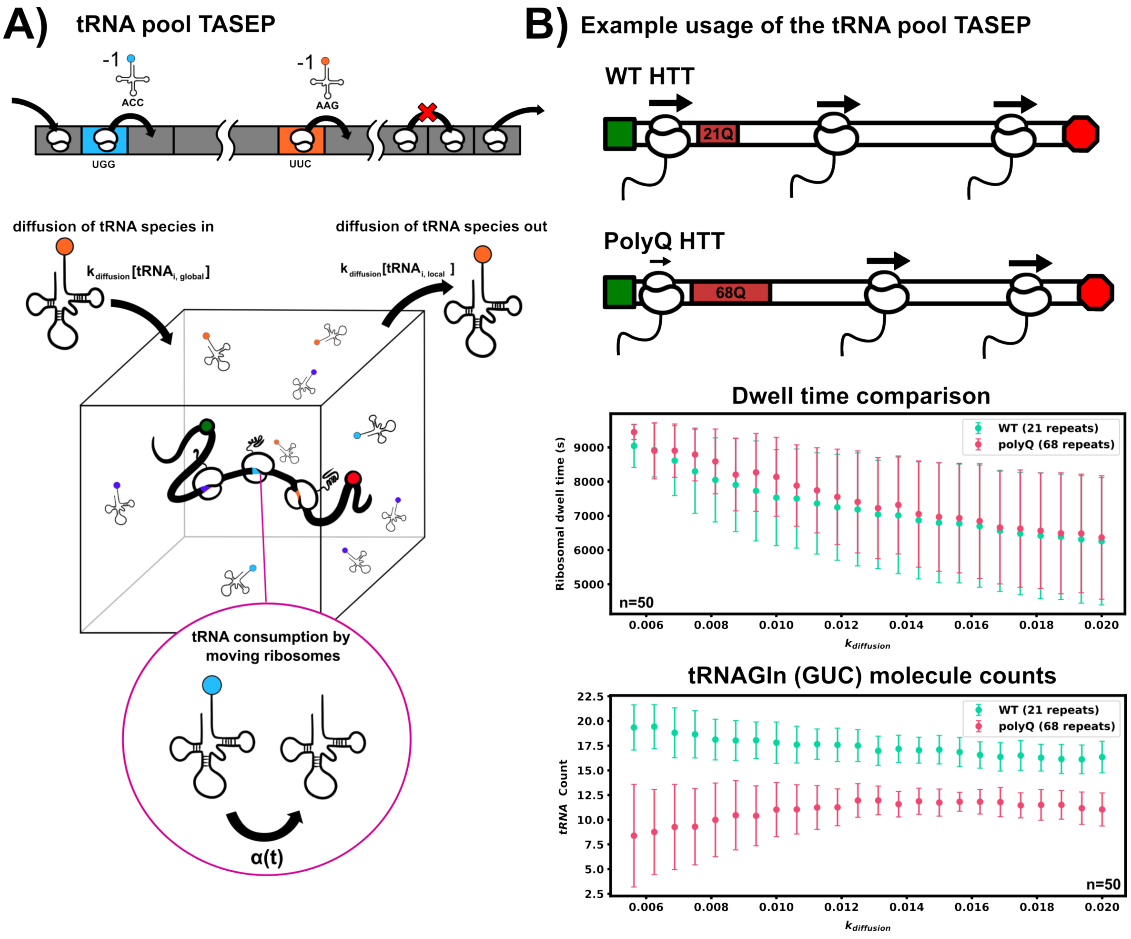


Figure 5.2: tRNA pooling model and example usage. A) Diagram describing the tRNA pooling TASEP. Ribosomes consume cognate tRNA resources as they elongate, tRNAs are free to diffuse in and out dictated by global and local concentrations and a diffusion rate. B). Example usage of the tRNA pool model for simulating a polyQ (CAG expansion) mRNA. WT Huntington's protein (21 CAG repeats) and PolyQ (68 PolyQ CAG repeats) were simulated at varying values of the diffusion scaling rate, $k_{\text{diffusion}}$. As diffusion slows, the WT HTT shows a lower dwell time and a higher local concentration of tRNA resources vs PolyQ HTT.

The ribosomal occupancy vector is therefore described as a binary vector of length L where x_{i} = 1 corresponds to a ribosome occupying the i^{th} position.

$$\mathbf{x} = x_1, x_2, ..., x_{\text{L}} \in \mathbb{B}^{\text{L}} \tag{5.2}$$

Ribosomal movement along the transcript is described by three primary steps: initiation (w_0) - where the ribosome binds at the start codon, elongation (w_{i}) - movement of the ribosome from

position 1 to L, and termination (w_{t}) - the rate at which ribosomes at position L unbind and complete translation.

mRNA transcript codons are converted to a vector of length L of index numbers 0-61, corresponding to the maximum 61 tRNA species used for elongation. While tRNA the total tRNA species are less than the total codon combinations, for the purposes of these equations we will use 61 as our maximum for the 61 sense codons in Eukarya [253].

$$\mathbf{n} = n_1, n_2, ..., n_{\mathrm{L}-1} \in (0 < \mathbb{R} < 60)^{\mathrm{L}-1} \tag{5.3}$$

Ribosomal binding is described by Eq. 5.4.

$$w_0 = k_{\mathrm{initiation}} \prod_{\mathrm{j}=1}^{\mathrm{nf}} (1 - x_{\mathrm{j}}) \tag{5.4}$$

where nf is the ribosomal footprint to account for physical exclusion if the binding site + nf is currently occupied. Ribosomal elongation is now dependent on the local concentration of the cognate tRNA species, $[tRNA_{\mathrm{n(i)}}]$, the global elongation scaling factor, $k_{\mathrm{elongation}}$, and the next codon not being excluded by other ribosomes.

$$w_{\mathrm{i}} = k_{\mathrm{elongation}} \cdot [\mathrm{tRNA}_{\mathrm{n(i)}}] \cdot x_{\mathrm{i}} \prod_{\mathrm{j}=1}^{\mathrm{nf}} (1 - x_{\mathrm{i+j}}) \text{ for } i = 1, ..., L - 1; \tag{5.5}$$

$$tRNA_{\mathrm{n(i)}} = \mathrm{tRNA}_{\mathrm{n(i)}} - 1 \tag{5.6}$$

Ribosomal termination is dictated by a termination rate at the stop codon (last lattice node).

$$w_{\mathrm{t}} = k_{\mathrm{termination}} * x_{\mathrm{L}} \tag{5.7}$$

The tRNA pool can be thought of as an arbitrary vicinity localized around the mRNA transcript where diffusion of tRNA in and out of the element is a birth death process (disregarding the additional consumption from the mRNA translation). tRNA species arrive at a defined consistent rate

for each of the 61 species of tRNA, $k_{\text{trna(id)}}$, where id refers to the tRNA id 0 through 60. tRNA leaves at a rate proportional to the current amount of that tRNA species and the defined diffusion rate $k_{\text{diffusion}}$.

$$\frac{d(\text{tRNA}_{\text{(id)}})}{dt} = k_{\text{diffusion}} \cdot [\text{tRNA}_{\text{(id)}}]_{\text{bulk}} - k_{\text{diffusion}} \cdot [\text{tRNA}_{\text{(id)}}]_{\text{local}} - \alpha(t) \tag{5.8}$$

$\alpha(t)$ can be conceptualized as the additional nonlinear consumption caused by any actively elongating ribosomes on the mRNA.

At high values of $k_{\text{diffusion}}$ and negligible mRNA consumption, $\alpha(t)$, the defined bulk and local concentrations of tRNA resources have the same concentration at a Quasi-steady state assumption.

$$[\text{tRNA}_{\text{(id)}}]_{\text{bulk}} \underset{k_{\text{diffusion}}}{\overset{k_{\text{diffusion}}}{\rightleftharpoons}} [\text{tRNA}_{\text{(id)}}]_{\text{local}} \quad \therefore \quad [\text{tRNA}_{\text{(id)}}]_{\text{bulk}} = [\text{tRNA}_{\text{(id)}}]_{\text{local}} \tag{5.9}$$

5.3.2 Novel dynamics arising from the tRNA pooling model

With the tRNA pooling model, we can explore some new mRNA dynamics. Specifically, we can examine at low $k_{\text{diffusion}}$ regimes where tRNA resources are limited, as, in the regime of high diffusion, this model recreates the Aguilera 2019 model when $k_{\text{elongation}}$ is set to one. The first thing to examine is translation of repeated codons. For this we can simulate translation on Huntington's protein (HTT). WT HTT has an exon1 polyQ region with a 8-35 repeat of the codon CAG. Expansion of this polyQ repeat past 35 codons creates a propensity for Huntington's Disease (HD), a fatal neurodegenerative condition. Length of the polyQ region is strikingly correlated with symptom onset age and disease progression. Numerous studies speculate on the cause of the disease to be aberrant protein folding, emergent RNA structures, codon mis-sense errors, slowed translation elongation and ribosomal stalling due to the dense repeated codon region [249, 250, 251, 252]. Recently, greater attention has been turned towards tRNA mis-sense errors as a predominant cause [254]. We can take our tRNA pooling model here and compare the mRNA consumption of $\text{tRNAGln}_{\text{GUC}}$ with a 21 repeat HTT and 68 repeat HTT. Fig. 5.2B shows the dwell times of 50 simulated trajectories of ribosomes across varying values of $k_{\text{diffusion}}$. As the rate of diffusion into the local compartment

is limited, dwell times between WT and PolyQ start to diverge with polyQ HTT exhibiting slower ribosomal dwell times on average. The bottom plot shows a striking difference in the local amount of tRNAs with the same tRNA diffusion parameters. The PolyQ HTT's steady state of cognate tRNAs is almost half the WT mRNA's even with high diffusion.

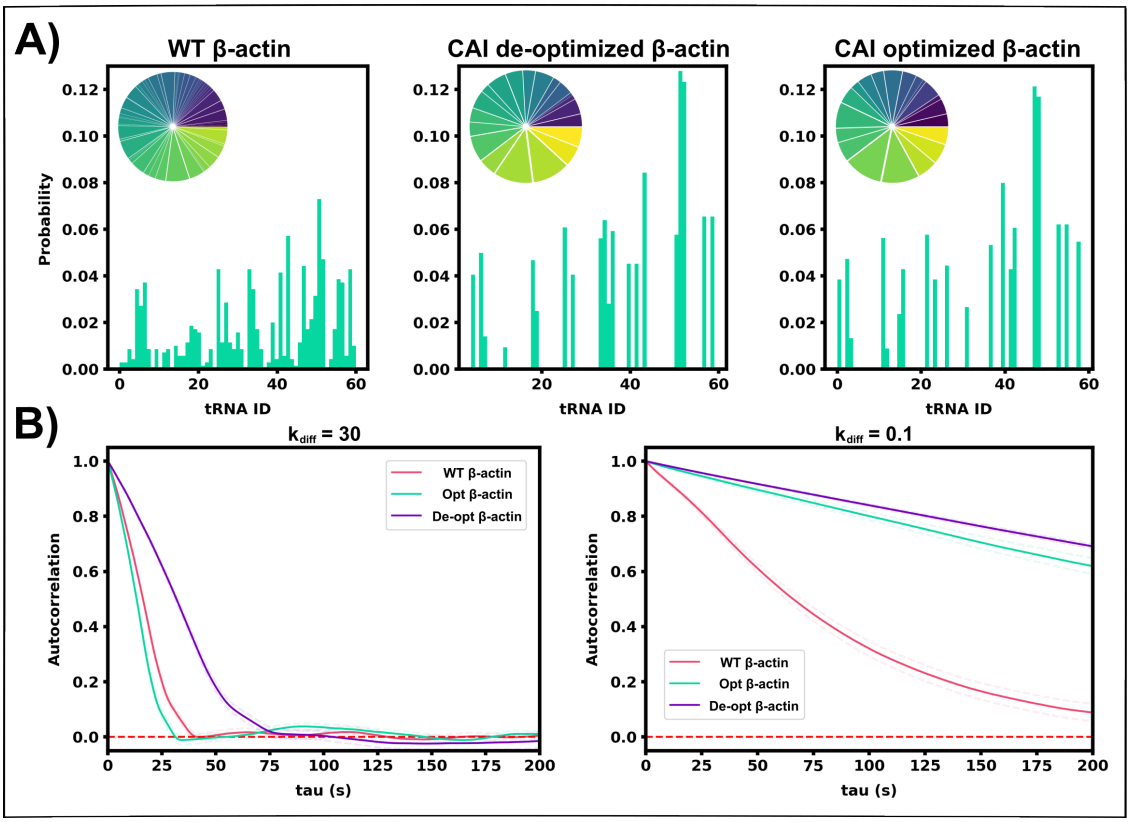


Figure 5.3: Common codon optimizations fail under simulated diffusion limited regimes. A) Three histograms of codon usage across WT, CAI de-optimized, and CAI optimized β-actin mRNAs. B) Auto-correlations of the constructs tagged with 10x-TFLAG across two different diffusion regimes: unlimited ($k_{\text{diffusion}}$ = 30) and limited ($k_{\text{diffusion}}$ = 0.1). The fastest dwell time switches from optimized codons to the WT when resources are limited as the WT has a higher diversity of codons and broader resource pool vs optimized or de-optimized.

Another dynamic that we can examine with this model is the failure of traditional optimization regimes such as CAI (codon adaptation index) or tAI (tRNA adaptation index) at low diffusion regimes. For this we took WT β-actin and de-optimized and optimized it according to the CAI and Nakamura human codon frequency [67]. Under this optimization the most frequent codons within a reference genome are used for optimization, and vice-versa for de-optimization. Fig. 5.3A

shows the relative distribution of all 61 sense codons for WT, optimized and de-optimized β-actin. Under high diffusion rates, $k_{\text{diffusion}}$ = 30, the model shows characteristic dwell times as one would expect: Slowest translation is undergone by de-optimized β-actin, fastest by the optimized, and somewhere in the middle is the WT β-actin (Fig. 5.3B left). However, when tRNA resources are

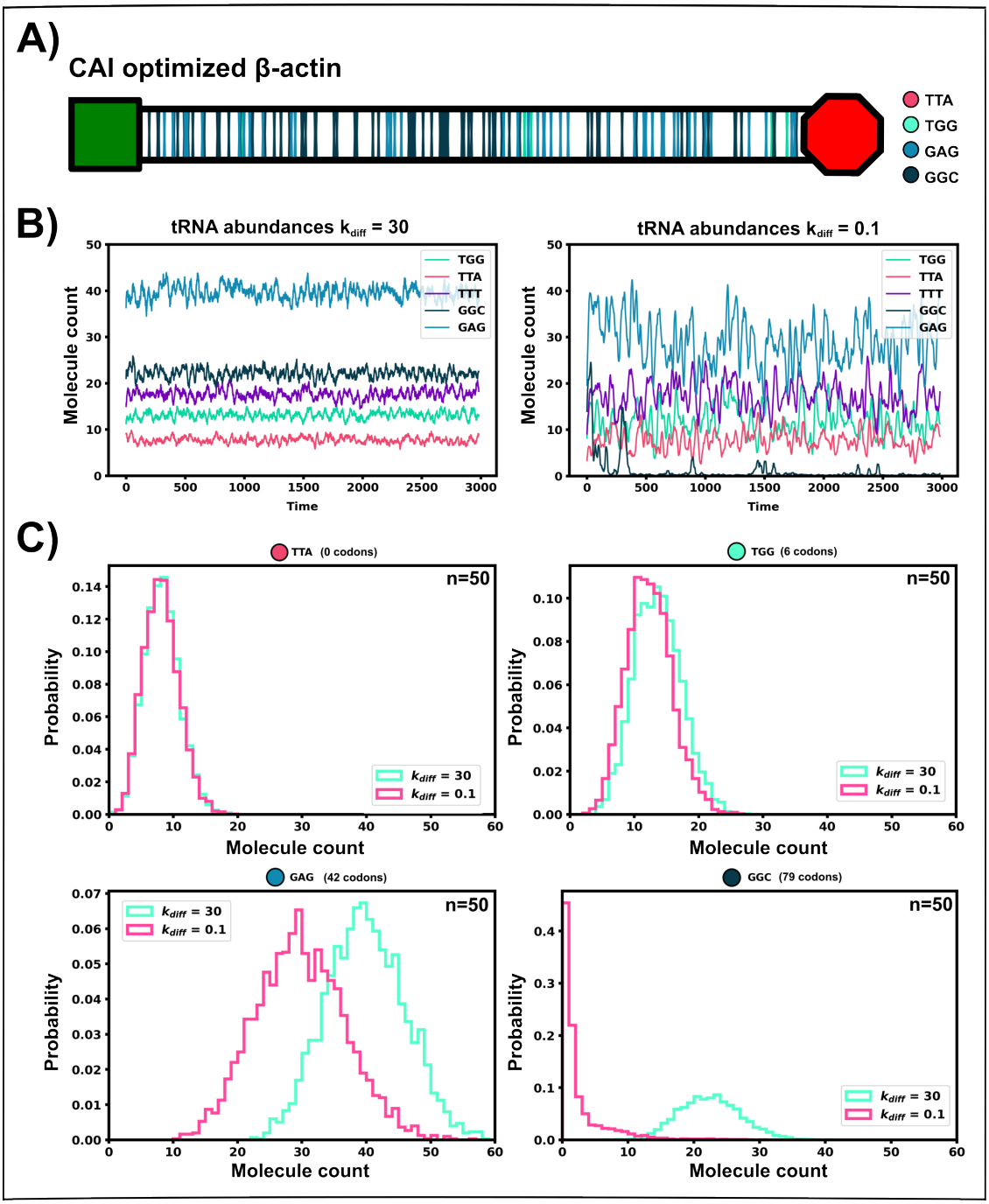


Figure 5.4: Caption on next page.

Figure 5.4: Optimized constructs deplete local tRNA resources in simulated diffusion limited regimes. A), Codon adaptation index (CAI) optimized β-actin with 4 different codons highlighted: TTA (0 codons), TGG (6 codons), GAG (42 codons), GGC (79 codons). B) tRNA abundances with two different $k_{\text{diffusion}}$ rates, one with no diffusion limitation (30) and one with a heavily diffusion limited regime (0.1). Moving averages (window=10) abundances of various tRNA resources are plotted over time from 50 simulated trajectories. C) individual distributions of the four highlighted tRNAs with under both diffusion regimes. TTA shows no change due to having no $\alpha(t)$ mRNA consumption as there are no TTA codons on the CAI β-actin. GGC shows a marked difference in the amount of available tRNA, to the point that when diffusion limited there is on average less than 10 available tRNAs for the mRNA to use.

diffusion limited or low, $k_{\text{diffusion}}$ = 0.1, the fastest translation is undergone by WT HTT since it has the largest diversity of codons, compared to both optimizations which enrich the same codons and thus have heavier, specific resource usage (Fig. 5.3B right).

We can zoom in on the local tRNA pool during translation of the CAI optimized β-actin. Fig. 5.4A shows the location of 4 different codons along the mRNA: TTA, 0 codons; TGG, 6 codons; GAG, 42 codons; GGC, 79 codons. Fig. 5.4B shows a moving average of 50 simulated tRNA pool model trajectories at $k_{\text{diffusion}}$ = 30 and 0.1. There is an expected higher randomness and consumption of tRNA species with lower diffusion rate, with GGC tRNAs almost completely consumed. Fig. 5.4C highlights the distributions of molecule counts within the 50 simulations in both diffusion regimes. TTA has no difference across diffusion rates as it is not consumed by the optimized β-actin; As more and more codons are represented on the mRNA, the larger the effect of the diffusion rate on that particular tRNA abundance.

These are just two brief examples of dynamics this model can exhibit —more in depth models could be created with the rSNAPsim that use cross talk between two mRNAs with their localized resources, or to use more accurate tRNA modification species and wobble decodings. However as we have no experimental system to compare some of these observations, this model has been placed on the proverbial shelf. We were hoping to eventually conduct NCT experiments with PolyQ constructs to observe some of the effects of repeat codons, especially within morphologies that may limit diffusion. Such studies would represent the first single-cell live cell data observing PolyQ translation.

5.4 Teaching and outreach

5.4.1 Undergraduate Quantitative Biology Summer School

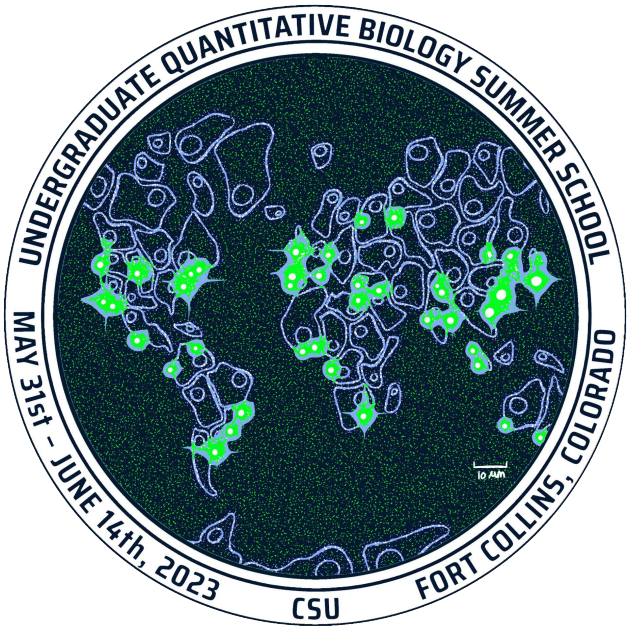


Figure 5.5: UQ-Bio summer school logo 2023

During my time at Colorado State University, I had the opportunity to participate in and facilitate the Undergraduate Quantitative Biology Summer School (and its predecessor, Quantitative Biology Summer School). The goal of the summer school is to introduce undergraduates and early graduate students to data analyses, computational modeling, and model inference all in relation to dynamic biological systems. Students often come to our school with their own biological system or data and leave with appropriate tools to conduct their own projects. While the students work on a presented challenge problem and learn the material, we also aim to introduce them to several research topics in the field with invited talks and general career advice with industry and academia panels. Each year since 2021, I have generated lecture materials, interactive Python notebooks, and lectured several of the lectures over the few week course (and designed the t-shirts). In the past two years, the focus of the program has shifted and streamlined to providing undergraduate students a 2 week condensed course starting with Python basics and ending with model inference.

I provided support to streamlining and presenting a cohesive problem for the students to work on over their course. It has been an immense pleasure to help facilitate undergraduate learning in this context, the program is a lot of work. But always worth it, and I look forward to the UQ-BIO "cycle"within our lab each year.

5.4.2 BIOM 421, Transport Phenomena

In addition to our summer school, I had the pleasure of teaching a core course of the undergraduate biomedical engineering program here at CSU. I taught Transport Phenomena (BIOM 421) to a the junior and senior undergrads in Spring 2023. As part of my duties I lectured, wrote exams, and designed course materials. This was my first teaching experience where I was the main lecturer, and I found it an experience that was challenging and rewarding. As the students were late in their program, they were all dedicated and a pleasure to teach over the semester.

Bibliography

[1] Y LeCun, Y Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, May 2015.

[2] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33, 2019.

[3] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8, March 2021.

[4] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, December 1976.

[5] Kah Teong Soh and Paul K. Wallace. *RNA flow cytometry using the branched DNA technique*, volume 1678, pages 49–77. Humana Press, 2018.

[6] Mohamed N.M. Bahrudeen, Vatsala Chauhan, Cristina S.D. Palma, Samuel M.D. Oliveira, Vinodh K. Kandavalli, and Andre S. Ribeiro. Estimating RNA numbers in single cells by RNA fluorescent tagging and flow cytometry. *Journal of Microbiological Methods*, 166, November 2019.

[7] Gloria A. Brar and Jonathan S. Weissman. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology 2015 16:11*, 16:651–664, October 2015.

[8] Nicholas T. Ingolia. Ribosome footprint profiling of translation throughout the genome. *Cell*, 165:22–23, March 2016.

[9] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 2008 10:1*, 10:57–63, January 2009.

[10] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics 2019 20:11*, 20:631–656, July 2019.

[11] Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics*, 10:317, April 2019.

[12] Bilal Aslam, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. Proteomics: Technologies and their applications. *Journal of Chromatographic Science*, 55:182–196, February 2017.

[13] Barbara Gorgoni, Luca Ciandrini, Matthew R. McFarland, M. Carmen Romano, and Ian Stansfield. Identification of the mRNA targets of tRNA-specific regulation using genome-wide simulation of translation. *Nucleic Acids Research*, 44:9231–9244, November 2016.

[14] Andrew J.M. Howden, Vincent Geoghegan, Kristin Katsch, Georgios Efstathiou, Bhaskar Bhushan, Omar Boutureira, Benjamin Thomas, David C. Trudgian, Benedikt M. Kessler, Daniela C. Dieterich, Benjamin G. Davis, and Oreste Acuto. QuaNCAT: quantitating proteome dynamics in primary cells. *Nature methods*, 10:343–346, April 2013.

[15] Katrin Eichelbaum, Markus Winter, Mauricio Berriel Diaz, Stephan Herzig, and Jeroen Krijgsveld. Selective enrichment of newly synthesized proteins for quantitative secretome analysis. *Nature Biotechnology*, pages 984–990, October 2012.

[16] Susanne Tom Dieck, Anke Müller, Anne Nehring, Flora I Hinz, Ina Bartnik, Erin M Schuman, and Daniela C Dieterich. Metabolic labeling with noncanonical amino acids and visualization by chemoselective fluorescent tagging. *Current Protocols in Cell Biology*, September 2012.

[17] Alexandre David, Brian P. Dolan, Heather D. Hickman, Jonathan J. Knowlton, Giovanna Clavarino, Philippe Pierre, Jack R. Bennink, and Jonathan W. Yewdell. Nuclear translation visualized by ribosome-bound nascent chain puromycylation. *Journal of Cell Biology*, 197:45–57, April 2012.

[18] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*, 50:1–14, August 2018.

[19] Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9, August 2017.

[20] Hu Zeng, Jiahao Huang, Jingyi Ren, Connie Kangni Wang, Zefang Tang, Haowen Zhou, Yiming Zhou, Hailing Shi, Abhishek Aditham, Xin Sui, Hongyu Chen, Jennifer A. Lo, and Xiao Wang. Spatially resolved single-cell translatomics at molecular resolution. *Science*, 380, June 2023.

[21] Charlotte A. Cialek, Amanda L. Koch, Gabriel Galindo, and Timothy J. Stasevich. Lighting up single-mRNA translation dynamics in living cells. *Current Opinion in Genetics and Development*, 61:75–82, April 2020.

[22] Xavier Pichon, Mounia Lagha, Florian Mueller, and Edouard Bertrand. A growing toolbox to image gene expression in single cells: Sensitive approaches for demanding challenges. *Molecular cell*, 71:468–480, August 2018.

[23] Tatsuya Morisaki, Kenneth Lyon, Keith F. DeLuca, Jennifer G. DeLuca, Brian P. English, Zhengjian Zhang, Luke D. Lavis, Jonathan B. Grimm, Sarada Viswanathan, Loren L. Looger, Timothee Lionnet, and Timothy J. Stasevich. Real-time quantification of single RNA translation dynamics in living cells. *Science*, 352:1425–1429, June 2016.

[24] Bin Wu, Carolina Eliscovich, Young J. Yoon, and Robert H. Singer. Translation dynamics of single mRNAs in live cells and neurons. *Science (New York, N.Y.)*, 352:1430, June 2016.

[25] Xiaowei Yan, Tim A. Hoek, Ronald D. Vale, and Marvin E. Tanenbaum. Dynamics of translation of single mRNA molecules in vivo. *Cell*, 165:976–989, May 2016.

[26] Xavier Pichon, Amandine Bastide, Adham Safieddine, Racha Chouaib, Aubin Samacoits, Eugenia Basyuk, Marion Peter, Florian Mueller, and Edouard Bertrand. Visualization of single endogenous polysomes reveals the dynamics of translation in live human cells. *Journal of Cell Biology*, 214:769–781, September 2016.

[27] Tatsuya Morisaki and Timothy J. Stasevich. Quantifying single mRNA translation kinetics in living cells. *Cold Spring Harbor Perspectives in Biology*, 10, November 2018.

[28] Kenneth Lyon, Luis U Aguilera, Tatsuya Morisaki, Brian Munsky, and Timothy J Stasevich. Live-cell single RNA imaging reveals bursts of translational frameshifting in brief. *Molecular Cell*, 75:172–183, 2019.

[29] S. Boersma, Deepak Khuperkar, Bram M.P. Verhagen, S. Sonneveld, Jonathan B. Grimm, Luke D. Lavis, and Marvin E. Tanenbaum. Multi-color single-molecule imaging uncovers extensive heterogeneity in mRNA decoding. *Cell*, 178:458–472.e19, July 2019.

[30] Malgorzata J. Latallo, Shaopeng Wang, Daoyuan Dong, Blake Nelson, Nathan M. Livingston, Rong Wu, Ning Zhao, Timothy J. Stasevich, Michael C. Bassik, Shuying Sun, and Bin Wu. Single-molecule imaging reveals distinct elongation and frameshifting dynamics between frames of expanded RNA repeats in C9ORF72-ALS/FTD. *Nature Communications 2023 14:1*, 14:1–18, September 2023.

[31] Chloe L. Barrington, Gabriel Galindo, Amanda L. Koch, Emma R. Horton, Evan J. Morrison, Samantha Tisa, Timothy J. Stasevich, and Olivia S. Rissland. Synonymous codon usage regulates translation initiation. *Cell Reports*, 42, December 2023.

[32] Stephanie L. Moon, Tatsuya Morisaki, Anthony Khong, Kenneth Lyon, Roy Parker, and Timothy J. Stasevich. Multicolor single-molecule tracking of mRNA interactions with RNP granules. *Nature cell biology*, 21:162, February 2019.

[33] Ivana Horvathova, Franka Voigt, Anna V. Kotrys, Yinxiu Zhan, Caroline G. Artus-Revel, Jan Eglinger, Michael B. Stadler, Luca Giorgetti, and Jeffrey A. Chao. The dynamics of mRNA turnover revealed by single-molecule imaging in single cells. *Molecular Cell*, 68:615–625.e9, November 2017.

[34] Tim A. Hoek, Deepak Khuperkar, Rik G.H. Lindeboom, Stijn Sonneveld, Bram M.P. Verhagen, Sanne Boersma, Michiel Vermeulen, and Marvin E. Tanenbaum. Single-molecule imaging uncovers rules governing nonsense-mediated mRNA decay. *Molecular Cell*, 75:324–339.e11, July 2019.

[35] Amanda Koch, Luis Aguilera, Tatsuya Morisaki, Brian Munsky, and Timothy J. Stasevich. Quantifying the dynamics of IRES and cap translation with single-molecule resolution in live cells. *Nature Structural and Molecular Biology 2020 27:12*, 27:1095–1104, September 2020.

[36] Johannes H. Wilbertz, Franka Voigt, Ivana Horvathova, Gregory Roth, Yinxiu Zhan, and Jeffrey A. Chao. Single-molecule imaging of mRNA localization and regulation during the integrated stress response. *Molecular cell*, 73:946–958.e7, March 2019.

[37] Jean Michel Cioni, Julie Qiaojin Lin, Anne V. Holtermann, Max Koppers, Maximilian A.H. Jakobs, Afnan Azizi, Benita Turner-Bridger, Toshiaki Shigeoka, Kristian Franze, William A. Harris, and Christine E. Holt. Late endosomes act as mRNA translation platforms and sustain mitochondria in axons. *Cell*, 176:56, January 2019.

[38] Charlotte A. Cialek, Gabriel Galindo, Tatsuya Morisaki, Ning Zhao, Taiowa A. Montgomery, and Timothy J. Stasevich. Imaging translational control by argonaute with single-molecule resolution in live cells. *Nature Communications 2022 13:1*, 13:1–14, June 2022.

[39] Jonathan M. Raser and Erin K. O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, February 2005.

[40] Brian Munsky, Gregor Neuert, and Alexander Van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336:183–187, April 2012.

[41] Niraj Kumar, Abhyudai Singh, and Rahul V. Kulkarni. Transcriptional bursting in gene expression: Analytical results for general stochastic models. *PLoS Computational Biology*, 11, October 2015.

[42] Joseph Rodriguez, Gang Ren, Christopher R. Day, Keji Zhao, Carson C. Chow, and Daniel R. Larson. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell*, 176:213–226, January 2019.

[43] Daniel T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[44] Carolyn T Macdonald and Julian H Gibbs. Kinetics of biopolymerization on nucleic acid templates. *Biopolymers*, 6:1–25, 1968.

[45] Carolyn T. MacDonald and Julian H. Gibbs. Concerning the kinetics of polypeptide synthesis on polyribosomes. *Biopolymers*, 7:707–725, 1969.

[46] Frank Spitzer. Interaction of markov processes. *Advances in Mathematics*, 5:246–290, 1970.

[47] Sudipto Muhuri, Lenin Shagolsem, and Madan Rao. Bidirectional transport in a multi-species totally asymmetric exclusion-process model. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 84:031921, September 2011.

[48] John O Wilson, Arturo D Zaragoza, Jing Xu, Marcus-Alexander Assmann, and Peter Lenz. Collective vesicle transport on biofilaments carried by competing molecular motors. *Europhysics Letters*, 84:58009, December 2008.

[49] R. K.P. Zia, J. J. Dong, and B. Schmittmann. Modeling translation in protein synthesis with TASEP: A tutorial and recent developments. *Journal of Statistical Physics 2011 144:2*, 144:405–428, April 2011.

[50] L.U. Aguilera, W. Raymond, Z.R. Fox, M. May, E. Djokic, T. Morisaki, T.J. Stasevich, and B. Munsky. Computational design and interpretation of single-rna translation experiments. *PLoS computational biology*, 15, October 2019.

[51] Chikashi Arita and Andreas Schadschneider. The dynamics of waiting: The exclusive queueing process. *Transportation Research Procedia*, 2:87–95, January 2014.

[52] Ethan Levien, Jiseon Min, Jane Kondev, al, Nasim Golafshan, Hamidreza Gharibi, Mahshid Kharaziha, T Chou, K Mallick, and R K P Zia. Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport. *Reports on Progress in Physics*, 74:116601, October 2011.

[53] Alisa Knizel, Leonid Petrov, and Axel Saenz. Generalizations of TASEP in discrete and continuous inhomogeneous space. *Communications in Mathematical Physics*, 372:797–864, December 2019.

[54] Konstantin Matetski and Daniel Remenik. Probability theory and related fields TASEP and generalizations: method for exact solution. *Probability Theory and Related Fields*, pages 615–698, April 2023.

[55] Edouard Bertrand, Pascal Chartrand, Matthias Schaefer, Shailesh M. Shenoy, Robert H. Singer, and Roy M. Long. Localization of ASH1 mRNA particles in living yeast. *Molecular Cell*, 2, October 1998.

[56] Daniel R. Larson, Daniel Zenklusen, Bin Wu, Jeffrey A. Chao, and Robert H. Singer. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science*, 332:475–478, April 2011.

[57] Sami Hocine, Pascal Raymond, Daniel Zenklusen, Jeffrey A. Chao, and Robert H. Singer. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nature Methods*, 10, 2013.

[58] Amir Mor, Shimrit Suliman, Rakefet Ben-Yishay, Sharon Yunger, Yehuda Brody, and Yaron Shav-Tal. Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nature Cell Biology*, 12:543–552, June 2010.

[59] Adina R. Buxbaum, Bin Wu, and Robert H. Singer. Single β-actin mRNA detection in neurons reveals a mechanism for regulating its translatability. *Science*, 343:419–422, January 2014.

[60] Chong Wang, Boran Han, Ruobo Zhou, and Xiaowei Zhuang. Real-time imaging of translation on single mRNA transcripts in live cells. *Cell*, 165:990–1001, May 2016.

[61] Kenneth Lyon and Timothy J. Stasevich. Imaging translational and post-translational gene regulatory dynamics in living cells with antibody-based probes. *Trends in Genetics*, 33:322–335, May 2017.

[62] Jeetayu Biswas, Yang Liu, Robert H. Singer, and Bin Wu. Fluorescence imaging methods to investigate translation in single cells. *Cold Spring Harbor Perspectives in Biology*, 11, April 2019.

[63] S. Boersma, Deepak Khuperkar, Bram M.P. Verhagen, S. Sonneveld, Jonathan B. Grimm, Luke D. Lavis, and Marvin E. Tanenbaum. Multi-color single-molecule imaging uncovers extensive heterogeneity in mRNA decoding. *Cell*, 178:458–472.e19, July 2019.

[64] Pierre Bonnin, Norbert Kern, Neil T. Young, Ian Stansfield, and M. Carmen Romano. Novel mRNA-specific effects of ribosome drop-off on translation rate and polysome profile. *PLoS Computational Biology*, 13, May 2017.

[65] Liana F Lareau, Dustin H Hite, Gregory J Hogan, and Patrick O Brown. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mrna fragments. *eLife*, 3:e01257, May 2014.

[66] Shelly Mahlab, Tamir Tuller, and Michal Linial. Conservation of the relative tRNA composition in healthy and cancerous tissues. *RNA*, 18:640–652, April 2012.

[67] Yasukazu Nakamura, Takashi Gojobori, and Toshimichi Ikemura. Codon usage tabulated from international DNA sequence databases: Status for the year 2000. *Nucleic Acids Research*, 28:292, January 2000.

[68] H. Mimi Zhou, Ingrid Brust-Mascher, and Jonathan M. Scholey. Direct visualization of the movement of the monomeric axonal transport motor UNC-104 along neuronal processes in living caenorhabditis elegans. *Journal of Neuroscience*, 21:3749–3755, June 2001.

[69] Abhyudai Singh and João P. Hespanha. Approximate moment dynamics for chemically reacting systems. *IEEE Transactions on Automatic Control*, 56:414–418, February 2011.

[70] C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, volume 13 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, third edition, 2004.

[71] Huaiyu Mi, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, 8:1551–1566, August 2013.

[72] Kseniya A. Akulich, Dmitry E. Andreev, Ilya M. Terenin, Victoria V. Smirnova, Aleksandra S. Anisimova, Desislava S. Makeeva, Valentina I. Arkhipova, Elena A. Stolboushkina, Maria B. Garber, Maria M. Prokofjeva, Pavel V. Spirin, Vladimir S. Prassolov, Ivan N. Shatsky, and Sergey E. Dmitriev. Four translation initiation pathways employed by the leaderless mRNA in eukaryotes. *Scientific Reports*, 6, November 2016.

[73] Nicholas T. Ingolia, Sina Ghaemmaghami, John R.S. Newman, and Jonathan S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324:218–223, April 2009.

[74] D. Axelrod, D. E. Koppel, J. Schlessinger, E. Elson, and W. W. Webb. Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. *Biophysical Journal*, 16:1055–1069, September 1976.

[75] Robert Hooke and T. A. Jeeves. “direct search” solution of numerical and statistical problems. *Journal of the ACM (JACM)*, 8, 1961.

[76] A. Coulon and D. R. Larson. Fluctuation analysis: Dissecting transcriptional kinetics with signal theory. *Methods in Enzymology*, 572:159–191, January 2016.

[77] Alexey A. Gritsenko, Marc Hulsman, Marcel J.T. Reinders, and Dick de Ridder. Unbiased quantitative models of protein translation derived from ribosome profiling data. *PLoS Computational Biology*, 11, August 2015.

[78] Khanh Dao Duc and Yun S. Song. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLoS Genetics*, 14, January 2018.

[79] Makio Tokunaga, Naoko Imamoto, and Kumiko Sakata-Sogawa. Highly inclined thin illumination enables clear single-molecule imaging in cells. *Nature Methods 2008 5:2*, 5:159–161, January 2008.

[80] Oleg Krichevsky and Grégoire Bonnet. Fluorescence correlation spectroscopy: The technique and its applications. *Reports on Progress in Physics*, 65, November 2002.

[81] Sina Jazani, Ioannis Sgouralis, and Steve Pressé. A method for single molecule tracking using a conventional single-focus confocal setup. *Journal of Chemical Physics*, 150, March 2019.

[82] Antoine Coulon, Matthew L Ferguson, Valeria de Turris, Murali Palangat, Carson C Chow, and Daniel R Larson. Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife*, 3, October 2014.

[83] Manuel Fresno, Antonio Jiménez, and David Vázquez. Inhibition of translation in eukaryotic systems by harringtonine. *European Journal of Biochemistry*, 72, 1977.

[84] Nagammal Neelagandan, Irene Lamberti, Hugo JF Carvalho, Cédric Gobet, and Felix Naef. What determines eukaryotic translation elongation: recent molecular and quantitative analyses of protein synthesis. *Open biology*, 10(12):200292, December 2020.

[85] Sulagna Das, Maria Vera, Valentina Gandin, Robert H. Singer, and Evelina Tutucci. Intracellular mRNA transport and localized translation. *Nature Reviews Molecular Cell Biology*, 22:483–504, July 2021.

[86] John R.P. Knight, Gavin Garland, Tuija Poyry, Emma Mead, Nikola Vlahov, Aristeidis Sfakianos, Stefano Grosso, Fabio De-Lima-Hedayioglu, Giovanna R. Mallucci, Tobias Von Der Haar, C. Mark Smales, Owen J. Sansom, and Anne E. Willis. Control of translation elongation in health and disease. *DMM Disease Models and Mechanisms*, 13, March 2020.

[87] Eugenia Basyuk, Florence Rage, and Edouard Bertrand. RNA transport from transcription to localized translation: a single molecule perspective. *RNA Biology*, 18:1221–1237, September 2020.

[88] Chenghua Cui, Wei Shu, and Peining Li. Fluorescence in situ hybridization: Cell-based genetic diagnostic and research applications. *Frontiers in Cell and Developmental Biology*, 4:89, September 2016.

[89] Susanne Tom Dieck, Lisa Kochen, Cyril Hanus, Maximilian Heumüller, Ina Bartnik, Belquis Nassim-Assir, Katrin Merk, Thorsten Mosler, Sakshi Garg, Stefanie Bunse, David A. Tirrell, and Erin M. Schuman. Direct visualization of identified and newly synthesized proteins in situ. *Nature methods*, 12:411, April 2015.

[90] Marina Chekulaeva and Markus Landthaler. Eyes on translation. *Molecular Cell*, 63:918–925, September 2016.

[91] Hotaka Kobayashi and Robert H. Singer. Single-molecule imaging of microRNA-mediated gene silencing in cells. *Nature Communications 2022 13:1*, 13:1–14, March 2022.

[92] Christian Tchito Tchapga, Thomas Attia Mih, Aurelle Tchagna Kouanou, Theophile Fozin Fonzin, Platini Kuetche Fogang, Brice Anicet Mezatio, and Daniel Tchiotsop. Biomedical image classification in a big data architecture using machine learning algorithms. *Journal of Healthcare Engineering*, 2021.

[93] Intisar Rizwan I Haque and Jeremiah Neubert. Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, 18, January 2020.

[94] Dimitri Palaz, Mathew Magimai Doss, and Ronan Collobert. Convolutional neural networks-based continuous speech recognition using raw speech signal. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4295–4299. IEEE, 2015.

[95] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009.

[96] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng. CNN for situations understanding based on sentiment analysis of twitter data. *Procedia Computer Science*, 111:376–381, 2017. The 8th International Conference on Advances in Information Technology.

[97] Qile Zhu, Xiaolin Li, Ana Conesa, and Cécile Pereira. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*, 34(9):1547–1554, December 2018.

[98] Albert K. Feeny, Mina K. Chung, Anant Madabhushi, Zachi I. Attia, Maja Cikes, Marjan Firouznia, Paul A. Friedman, Matthew M. Kalscheur, Suraj Kapa, Sanjiv M. Narayan, Peter A. Noseworthy, Rod S. Passman, Marco V. Perez, Nicholas S. Peters, Jonathan P. Piccini, Khaldoun G. Tarakji, Suma A. Thomas, Natalia A. Trayanova, Mintu P. Turakhia, and Paul J. Wang. Artificial intelligence and machine learning in arrhythmias and cardiac electrophysiology. *Circulation: Arrhythmia and Electrophysiology*, pages 873–890, 2020.

[99] Fatma Murat, Ozal Yildirim, Muhammed Talo, Ulas Baran Baloglu, Yakup Demir, and U. Rajendra Acharya. Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. *Computers in Biology and Medicine*, 120:103726, May 2020.

[100] Yu Xie and Stefan Oniga. A review of processing methods and classification algorithm for eeg signal. *Carpathian Journal of Electronic and Computer Engineering*, 13:23–29, 2020.

[101] Daniel B. Allan, Thomas Caswell, Nathan C. Keim, Casper M. van der Wel, and Ruben W. Verweij. soft-matter/trackpy: Trackpy v0.5.0. April 2021.

[102] Kota Miura, Fabrice P Cordelières, and Anna H Klemm. Bleach correction ImageJ plugin for compensating the photobleaching of time-lapse sequences. *F1000Research*, December 2020.

[103] Daniel Wüstner, Tanja Christensen, Lukasz M Solanko, and Daniel Sage. molecules photobleaching kinetics and time-integrated emission of fluorescent probes in cellular membranes. *Molecules*, 19:11096–11130, July 2014.

[104] Hao Shen, Lawrence J Tauzin, Rashad Baiyasi, Wenxiao Wang, Nicholas Moringo, Bo Shuang, and Christy F Landes. Single particle tracking: From theory to biophysical applications. *Chemical Reviews*, pages 7331–7376, May 2017.

[105] Shashikant Pujar, Nuala A. O'Leary, Catherine M. Farrell, Jane E. Loveland, Jonathan M. Mudge, Craig Wallin, Carlos G. Girón, Mark Diekhans, If Barnes, Ruth Bennett, Andrew E. Berry, Eric Cox, Claire Davidson, Tamara Goldfarb, Jose M. Gonzalez, Toby Hunt, John Jackson, Vinita Joardar, Mike P. Kay, Vamsi K. Kodali, Fergal J. Martin, Monica McAndrews, Kelly M. McGarvey, Michael Murphy, Bhanu Rajput, Sanjida H. Rangwala, Lillian D. Riddick, Ruth L. Seal, Marie Marthe Suner, David Webb, Sophia Zhu, Bronwen L. Aken, Elspeth A. Bruford, Carol J. Bult, Adam Frankish, Terence Murphy, and Kim D. Pruitt. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Research*, 46:D221, January 2018.

[106] Nicola K. Gray and Marvin Wickens. Control of translation initiation in animals. *Annual Review of Cell and Developmental Biology*, 14:399–458, November 2003.

[107] Kathrin Leppek, Rhiju Das, and Maria Barna. Functional 5' utr mRNA structures in eukaryotic translation regulation and how to find them. *Nature Reviews molecular cell biology*, pages 158–174, November 2017.

[108] Christopher S. Fraser. Quantitative studies of mRNA recruitment to the eukaryotic ribosome. *Biochimie*, 114:58–71, July 2015.

[109] Christine Mayr. Evolution and biological roles of alternative 3'UTRs. *Trends in Cell Biology*, 26:227–237, March 2016.

[110] Nahum Sonenberg and Alan G. Hinnebusch. Regulation of translation initiation in eukaryotes: Mechanisms and biological targets. *Cell*, 136:731–745, February 2009.

[111] Marc Robert Fabian, Nahum Sonenberg, and Witold Filipowicz. Regulation of mRNA translation and stability by microrna. *Annual Reviews*, 2010.

[112] John W.B. Hershey, Nahum Sonenberg, and Michael B. Mathews. Principles of translational control: An overview. *Cold Spring Harbor Perspectives in Biology*, 4:a011528, December 2012.

[113] Jing Zhao, Bo Qin, Rainer Nikolay, Christian M T Spahn, and Gong Zhang. Molecular sciences translatomics: The global view of translation. *International Journal of Molecular Sciences*, January 2019.

[114] Eyal Peer, Sharon Moshitch-Moshkovitz, Gideon Rechavi, and Dan Dominissini. The epitranscriptome in translation regulation. *Cold Spring Harbor Perspectives in Biology*, 11:a032623, August 2019.

[115] Sunjong Kwon. Single-molecule fluorescence in situ hybridization: Quantitative imaging of single RNA molecules. *BMB reports*, pages 65–72, February 2013.

[116] Kelly S Burke, Katie A Antilla, and David A Tirrell. A fluorescence in situ hybridization method to quantify mRNA translation by visualizing ribosome-mRNA interactions in single cells. *ACS Central Science*, 3:425–433, May 2017.

[117] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67:301–320, March 2005.

[118] James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

[119] Arthur D Edelstein, Mark A Tsuchida, Nenad Amodaj, Henry Pinkard, Ronald D Vale, and Nico Stuurman. Advanced methods of microscope control using μmanager software. *Journal of Biological Methods*, 1:e10, November 2014.

[120] Justin Demmerle, Siyuan Hao, and Danfeng Cai. Transcriptional condensates and phase separation: condensing information across scales and mechanisms. *Nucleus*, 14, December 2023.

[121] Stephen Buratowski. Progression through the RNA polymerase II CTD cycle. *Molecular Cell*, 36:541–546, November 2009.

[122] Roland Schüller, Ignasi Forné, Tobias Straub, Amelie Schreieck, Yves Texier, Nilay Shah, Tim Michael Decker, Patrick Cramer, Axel Imhof, and Dirk Eick. Heptad-specific phosphorylation of RNA polymerase II CTD. *Molecular Cell*, 61:305–314, January 2016.

[123] Kevin M. Harlen and L. Stirling Churchman. The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews Molecular Cell Biology*, 18:263–273, March 2017.

[124] Patrick Cramer. Organization and regulation of gene transcription. *Nature*, 573:45–54, September 2019.

[125] Ibrahim I. Cissé, Ignacio Izeddin, Sebastien Z. Causse, Lydia Boudarene, Adrien Senecal, Leila Muresan, Claire Dugast-Darzacq, Bassam Hajj, Maxime Dahan, and Xavier Darzacq. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science*, 341:664–667, August 2013.

[126] Won Ki Cho, Namrata Jayanth, Brian P. English, Takuma Inoue, J. Owen Andrews, William Conway, Jonathan B. Grimm, Jan Hendrik Spille, Luke D. Lavis, Timothée Lionnet, and Ibrahim I. Cisse. RNA polymerase II cluster dynamics predict mRNA output in living cells. *eLife*, 5, May 2016.

[127] Huasong Lu, Dan Yu, Anders S. Hansen, Sourav Ganguly, Rongdiao Liu, Alec Heckert, Xavier Darzacq, and Qiang Zhou. Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature*, 558:318–323, May 2018.

[128] Ryosuke Nagashima, Kayo Hibino, S. S. Ashwin, Michael Babokhov, Shin Fujishiro, Ryosuke Imai, Tadasu Nozaki, Sachiko Tamura, Tomomi Tani, Hiroshi Kimura, Michael Shribak, Masato T. Kanemaki, Masaki Sasai, and Kazuhiro Maeshima. Single nucleosome imaging reveals loose genome chromatin networks via active RNA polymerase II. *Journal of Cell Biology*, 218:1511–1530, March 2019.

[129] Marc Boehning, Claire Dugast-Darzacq, Marija Rankovic, Anders S. Hansen, Taekyung Yu, Herve Marie-Nelly, David T. McSwiggen, Goran Kokic, Gina M. Dailey, Patrick Cramer, Xavier Darzacq, and Markus Zweckstetter. RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature Structural and Molecular Biology*, 25:833–840, August 2018.

[130] Benjamin R. Sabari, Alessandra Dall'Agnese, Ann Boija, Isaac A. Klein, Eliot L. Coffey, Krishna Shrinivas, Brian J. Abraham, Nancy M. Hannett, Alicia V. Zamudio, John C. Manteiga, Charles H. Li, Yang E. Guo, Daniel S. Day, Jurian Schuijers, Eliza Vasile, Sohail Malik, Denes Hnisz, Tong Ihn Lee, Ibrahim I. Cisse, Robert G. Roeder, Phillip A. Sharp, Arup K. Chakraborty, and Richard A. Young. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361, June 2018.

[131] Martin Heidemann, Corinna Hintermair, Kirsten Voß, and Dirk Eick. Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1829:55–62, January 2013.

[132] Antoine Coulon, Carson C. Chow, Robert H. Singer, and Daniel R. Larson. Eukaryotic transcriptional dynamics: From single molecules to cell populations. *Nature Reviews Genetics*, 14:572–584, August 2013.

[133] Davide Mazza, Alice Abernathy, Nicole Golob, Tatsuya Morisaki, and James G. McNally. A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Research*, 40, August 2012.

[134] Bi Chang Chen, Wesley R. Legant, Kai Wang, Lin Shao, Daniel E. Milkie, Michael W. Davidson, Chris Janetopoulos, Xufeng S. Wu, John A. Hammer, Zhe Liu, Brian P. English,

Yuko Mimori-Kiyosue, Daniel P. Romero, Alex T. Ritter, Jennifer Lippincott-Schwartz, Lillian Fritz-Laylin, R. Dyche Mullins, Diana M. Mitchell, Joshua N. Bembenek, Anne Cecile Reymann, Ralph Böhme, Stephan W. Grill, Jennifer T. Wang, Geraldine Seydoux, U. Serdar Tulu, Daniel P. Kiehart, and Eric Betzig. Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. *Science*, 346, October 2014.

[135] Jiji Chen, Zhengjian Zhang, Li Li, Bi Chang Chen, Andrey Revyakin, Bassam Hajj, Wesley Legant, Maxime Dahan, Timothée Lionnet, Eric Betzig, Robert Tjian, and Zhe Liu. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell*, 156:1274–1285, March 2014.

[136] J. Li, A. Dong, Kamola Saydaminova, Hill Chang, Guanshi Wang, Hiroshi Ochiai, Takashi Yamamoto, and Alexandros Pertsinidis. Single-molecule nanoscopy elucidates RNA polymerase II transcription at single genes in live cells. *Cell*, 178:491–506, July 2019.

[137] Barbara Steurer, Roel C. Janssens, Bart Geverts, Marit E. Geijer, Franziska Wienholz, Arjan F. Theil, Jiang Chang, Shannon Dealy, Joris Pothof, Wiggert A. Van Cappellen, Adriaan B. Houtsmuller, and Jurgen A. Marteijn. Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America*, 115, April 2018.

[138] Katjana Tantale, Florian Mueller, Alja Kozulic-Pirher, Annick Lesne, Jean Marc Victor, Marie Cecile Robert, Serena Capozi, Racha Chouaib, Volker Bäcker, Julio Mateos-Langerak, Xavier Darzacq, Christophe Zimmer, Eugenia Basyuk, and Edouard Bertrand. A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nature Communications*, 7, July 2016.

[139] Yoko Hayashi-Takanaka, Kazuo Yamagata, Naohito Nozaki, and Hiroshi Kimura. Visualizing histone modifications in living cells: Spatiotemporal dynamics of h3 phosphorylation during interphase. *Journal of Cell Biology*, 187:781–790, December 2009.

[140] Hiroshi Kimura, Yoko Hayashi-Takanaka, Timothy J. Stasevich, and Yuko Sato. Visualizing posttranslational and epigenetic modifications of endogenous proteins in vivo. *Histochemistry and Cell Biology*, 144:101–109, August 2015.

[141] Sascha Conic, Dominique Desplancq, Alexia Ferrand, Veronique Fischer, Vincent Heyer, Bernardo Reina San Martin, Julien Pontabry, Mustapha Oulad-Abdelghani, N. Kishore Babu, Graham D. Wright, Nacho Molina, Etienne Weiss, and László Tora. Imaging of native transcription factors and histone phosphorylation at high resolution in live cells. *Journal of Cell Biology*, 217:1537–1552, February 2018.

[142] Yuko Sato, Lennart Hilbert, Haruka Oda, Yinan Wan, John M. Heddleston, Teng Leong Chew, Vasily Zaburdaev, Philipp Keller, Timothee Lionnet, Nadine Vastenhouw, and Hiroshi Kimura. Histone H3K27 acetylation precedes active transcription during zebrafish zygotic genome activation as revealed by live-cell analysis. *Development (Cambridge)*, 146, September 2019.

[143] Timothy J. Stasevich, Yoko Hayashi-Takanaka, Yuko Sato, Kazumitsu Maehara, Yasuyuki Ohkawa, Kumiko Sakata-Sogawa, Makio Tokunaga, Takahiro Nagase, Naohito Nozaki, James G. McNally, and Hiroshi Kimura. Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature*, 516:272–275, December 2014.

[144] Ran Taube, Xin Lin, Dan Irwin, Koh Fujinaga, and B. Matija Peterlin. Interaction between P-TEFb and the C-terminal domain of RNA polymerase II activates transcriptional elongation from sites upstream or downstream of target genes. *Molecular and Cellular Biology*, 22:321–331, January 2002.

[145] Yoko Hayashi-Takanaka, Kazuo Yamagata, Teruhiko Wakayama, Timothy J. Stasevich, Takashi Kainuma, Toshiki Tsurimoto, Makoto Tachibana, Yoichi Shinkai, Hitoshi Kurumizaka, Naohito Nozaki, and Hiroshi Kimura. Tracking epigenetic histone modifications in single cells using fab-based live endogenous modification labeling. *Nucleic Acids Research*, 39, 2011.

[146] Takayuki Nojima, Tomás Gomes, Ana Rita Fialho Grosso, Hiroshi Kimura, Michael J. Dye, Somdutta Dhir, Maria Carmo-Fonseca, and Nicholas J. Proudfoot. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell*, 161:526–540, April 2015.

[147] Justyna Zaborowska, Sylvain Egloff, and Shona Murphy. The pol II CTD: New twists in the tail. *Nature Structural and Molecular Biology*, 23:771–777, September 2016.

[148] Kirsten Bacia, Sally A. Kim, and Petra Schwille. Fluorescence cross-correlation spectroscopy in living cells. *Nature Methods*, 3:83–89, January 2006.

[149] Florian Mueller, Adrien Senecal, Katjana Tantale, Hervé Marie-Nelly, Nathalie Ly, Olivier Collin, Eugenia Basyuk, Edouard Bertrand, Xavier Darzacq, and Christophe Zimmer. FISH-quant: Automatic counting of transcripts in 3D FISH images. *Nature Methods*, 10:227–228, April 2013.

[150] Yehuda Brody, Noa Neufeld, Nicole Bieberstein, Sebastien Z. Causse, Eva Maria Böhnlein, Karla M. Neugebauer, Xavier Darzacq, and Yaron Shav-Tal. The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biology*, 9, January 2011.

[151] Denis V. Titov, Benjamin Gilman, Qing Li He, Shridhar Bhat, Woon Kai Low, Yongjun Dang, Michael Smeaton, Arnold L. Demain, Paul S. Miller, Jennifer F. Kugel, James A. Goodrich, and Jun O. Liu. XPB, a subunit of TFIIH, is a target of the natural product triptolide. *Nature Chemical Biology*, 7:182–188, January 2011.

[152] Ying Wang, Jin jian Lu, Li He, and Qiang Yu. Triptolide (TPL) inhibits global transcription by inducing proteasome-dependent degradation of RNA polymerase II (Pol II). *PLoS ONE*, 6, September 2011.

[153] Nicholas Kwiatkowski, Tinghu Zhang, Peter B. Rahl, Brian J. Abraham, Jessica Reddy, Scott B. Ficarro, Anahita Dastur, Arnaud Amzallag, Sridhar Ramaswamy, Bethany Tesar, Catherine E. Jenkins, Nancy M. Hannett, Douglas McMillin, Takaomi Sanda, Taebo Sim,

Nam Doo Kim, Thomas Look, Constantine S. Mitsiades, Andrew P. Weng, Jennifer R. Brown, Cyril H. Benes, Jarrod A. Marto, Richard A. Young, and Nathanael S. Gray. Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature*, 511:616–620, July 2014.

[154] Sheng Hao Chao, Koh Fujinaga, Jon E. Marion, Ran Taube, Edward A. Sausville, Adrian M. Senderowicz, B. Matija Peterlin, and David H. Price. Flavopiridol inhibits P-TEFb and blocks HIV-1 replication. *Journal of Biological Chemistry*, 275:28345–28348, September 2000.

[155] Won Ki Cho, Jan Hendrik Spille, Micca Hecht, Choongman Lee, Charles Li, Valentin Grube, and Ibrahim I. Cisse. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361:412–415, June 2018.

[156] Lucas Brandon Edelman and Peter Fraser. Transcription factories: Genetic programming in three dimensions. *Current Opinion in Genetics and Development*, 22:110–114, April 2012.

[157] Alexander Feuerborn and Peter R. Cook. Why the activity of a gene depends on its neighbors. *Trends in Genetics*, 31:483–490, September 2015.

[158] Tahir H. Tahirov, Nigar D. Babayeva, Katayoun Varzavand, Jeffrey J. Cooper, Stanley C. Sedore, and David H. Price. Crystal structure of HIV-1 Tat complexed with human P-TEFb. *Nature*, 465:747–751, June 2010.

[159] Xavier Darzacq, Yaron Shav-Tal, Valeria De Turris, Yehuda Brody, Shailesh M. Shenoy, Robert D. Phair, and Robert H. Singer. In vivo dynamics of RNA polymerase II transcription. *Nature Structural and Molecular Biology*, 14:796–806, August 2007.

[160] Yang Eric Guo, John C. Manteiga, Jonathan E. Henninger, Benjamin R. Sabari, Alessandra Dall'Agnese, Nancy M. Hannett, Jan Hendrik Spille, Lena K. Afeyan, Alicia V. Zamudio, Krishna Shrinivas, Brian J. Abraham, Ann Boija, Tim Michael Decker, Jenna K. Rimel, Charli B. Fant, Tong Ihn Lee, Ibrahim I. Cisse, Phillip A. Sharp, Dylan J. Taatjes, and Richard A. Young. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*, 572:543–548, August 2019.

[161] Matjaz Barboric and B. Matija Peterlin. A new paradigm in eukaryotic biology: HIV tat and the control of transcriptional elongation. *PLoS Biology*, 3, February 2005.

[162] Timothée Lionnet and Robert H. Singer. Transcription goes digital. *EMBO Reports*, 13:313–321, April 2012.

[163] Leighton Core and Karen Adelman. Promoter-proximal pausing of RNA polymerase II: A nexus of gene regulation. *Genes and Development*, 33:960–982, August 2019.

[164] Karen Adelman and John T. Lis. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics*, 13:720–731, October 2012.

[165] Bo Gu, Tomek Swigut, Andrew Spencley, Matthew R. Bauer, Mingyu Chung, Tobias Meyer, and Joanna Wysocka. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science*, 359:1050–1055, March 2018.

[166] Aaron F. Straight, Andrew S. Belmont, Carmen C. Robinett, and Andrew W. Murray. GFP tagging of budding yeast chromosomes reveals that protein-protein interactions can mediate sister chromatid cohesion. *Current Biology*, 6:599–608, December 1996.

[167] Patrick H. Viollier, Martin Thanbichler, Patrick T. McGrath, Lisandra West, Maliwan Meewan, Harley H. McAdams, and Lucy Shapiro. Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proceedings of the National Academy of Sciences of the United States of America*, 101:9257–2962, June 2004.

[168] Hiroshi Ochiai, Takeshi Sugawara, and Takashi Yamamoto. Simultaneous live imaging of the transcription and nuclear position of specific genes. *Nucleic Acids Research*, 43, October 2015.

[169] Bernard Mariamé, Sandrine Kappler-Gratias, Martin Kappler, Stéphanie Balor, Franck Gallardo, and Kerstin Bystricky. Real-time visualization and quantification of human cytomegalovirus replication in living cells using the ANCHOR DNA labeling technology. *Journal of Virology*, 92, August 2018.

[170] Yodai Takei, Sheel Shah, Sho Harvey, Lei S. Qi, and Long Cai. Multiplexed dynamic imaging of genomic loci by combined CRISPR imaging and DNA sequential FISH. *Biophysical Journal*, 112:1773–1776, May 2017.

[171] Wulan Deng, Xinghua Shi, Robert Tjian, Timothée Lionnet, and Robert H. Singer. CASFISH: CRISPR/cas9-mediated in situ labeling of genomic loci in fixed cells. *Proceedings of the National Academy of Sciences of the United States of America*, 112:11870–11875, September 2015.

[172] Yuko Sato, Masanori Mukai, Jun Ueda, Michiko Muraki, Timothy J. Stasevich, Naoki Horikoshi, Tomoya Kujirai, Hiroaki Kita, Taisuke Kimura, Seiji Hira, Yasushi Okada, Yoko Hayashi-Takanaka, Chikashi Obuse, Hitoshi Kurumizaka, Atsuo Kawahara, Kazuo Yamagata, Naohito Nozaki, and Hiroshi Kimura. Genetically encoded system to track histone modification in vivo. *Scientific Reports*, 3, August 2013.

[173] Hiroshi Kimura, Yong Tao, Robert G. Roeder, and Peter R. Cook. Quantitation of RNA polymerase II and its transcription factors in an hela cell: Little soluble holoenzyme but significant amounts of polymerases attached to the nuclear substructure. *Molecular and Cellular Biology*, 19:5383–5392, August 1999.

[174] Ulrich Rothbauer, Kourosh Zolghadr, Sergei Tillib, Danny Nowak, Lothar Schermelleh, Anja Gahl, Natalija Backmann, Katja Conrath, Serge Muyldermans, M. Cristina Cardoso, and Heinrich Leonhardt. Targeting and tracing antigens in live cells with fluorescent nanobodies. *Nature Methods*, 3:887–889, November 2006.

[175] Hiroshi Kimura, Yoko Hayashi-Takanaka, Yuji Goto, Nanako Takizawa, and Naohito Nozaki. The organization of histone H3 modifications as revealed by a panel of specific monoclonal antibodies. *Cell Structure and Function*, 33:61–73, 2008.

[176] P. L. McNeil and E. Warder. Glass beads load macromolecules into living cells. *Journal of Cell Science*, 88, 1988.

[177] Erik M.M. Manders, Hiroshi Kimura, and Peter R. Cook. Direct imaging of DNA in living cells reveals the dynamics of chromosome formation. *Journal of Cell Biology*, 144:813–21, March 1999.

[178] Lina Carlini, Alexander Benke, Luc Reymond, Gražvydas Lukinavičius, and Suliana Manley. Reduced dyes enhance single-molecule localization density for live superresolution imaging. *ChemPhysChem*, 15:750–755, March 2014.

[179] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: An open-source platform for biological-image analysis. *Nature Methods*, 9:676–682, June 2012.

[180] Catherine E. Scull, Shiba S. Dandpat, Rosa A. Romero, and Nils G. Walter. Transcriptional riboswitches integrate timescales for bacterial gene expression control. *Frontiers in Molecular Biosciences*, 7:480, January 2021.

[181] A. Serganov and E. Nudler. A Decade of Riboswitches. *Cell*, 152(1-2):17–24, Jan 2013.

[182] Phillip J Mccown, Keith A Corbino, Shira Stav, Madeline E Sherlock, and Ronald R Breaker. Riboswitch diversity and distribution. *RNA*, pages 995–1011, July 2017.

[183] Maumita Mandal and Ronald R. Breaker. Gene regulation by riboswitches. *Nature Reviews Molecular Cell Biology*, 5:451–463, June 2004.

[184] Kumari Kavita and Ronald R. Breaker. Discovering riboswitches: the past and the future. *Trends in Biochemical Sciences*, 48:119–141, February 2023.

[185] A. Wachter. Riboswitch-mediated control of gene expression in eukaryotes. *RNA Biology*, 7(1):67–76, February 2010.

[186] Samuel Bocobza, Avital Adato, Tali Mandel, Michal Shapira, Evgeny Nudler, and Asaph Aharoni. Riboswitch-dependent gene regulation and its evolution in the plant kingdom. *Genes & Development*, 21:2874–2879, November 2007.

[187] P. S. Ray, J. Jia, P. Yao, M. Majumder, M. Hatzoglou, and P. L. Fox. A stress-responsive RNA switch regulates VEGFA expression. *Nature*, 457(7231):915–919, Feb 2009.

[188] Kadiam C. Venkata Subbaiah, Omar Hedaya, Jiangbin Wu, Feng Jiang, and Peng Yao. Mammalian RNA switches: Molecular rheostats in gene regulation, disease, and medicine. *Computational and Structural Biotechnology Journal*, 17:1326, January 2019.

[189] Raphael Petegrosso, Zhuliu Li, and Rui Kuang. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Briefings in Bioinformatics*, 21:1209–1223, July 2020.

[190] Noorul Amin, Annette McGrath, and Yi Ping Phoebe Chen. Evaluation of deep learning in non-coding RNA classification. *Nature Machine Intelligence*, 1:246–256, May 2019.

[191] T. J. Wheeler and S. R. Eddy. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29(19):2487–2489, Oct 2013.

[192] T. H. Chang, H. D. Huang, L. C. Wu, C. T. Yeh, B. J. Liu, and J. T. Horng. Computational identification of riboswitches based on RNA conserved functional sequences and conformations. *RNA*, 15(7):1426–1430, May 2009.

[193] Sumit Mukherjee and Supratim Sengupta. Riboswitch scanner: an efficient pHMM-based web-server to detect riboswitches in genomic sequences. *Bioinformatics*, 32:776–778, March 2016.

[194] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, Nov 2013.

[195] P. Bengert and T. Dandekar. Riboswitch finder–a tool for identification of riboswitch RNAs. *Nucleic Acids Research*, 32(Web Server issue):W154–159, Jul 2004.

[196] Cei Abreu-Goodger and Enrique Merino. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Research*, 33:W690, July 2005.

[197] Yuan Li, Cuncong Zhong, and Shaojie Zhang. Finding consensus stable local optimal structures for aligned RNA sequences and its application to discovering riboswitch elements. *International Journal of Bioinformatics Research and Applications*, 10:498–518, September 2014.

[198] Keshav Aditya R. Premkumar, Ramit Bharanikumar, and Ashok Palaniappan. Riboflow: using deep learning to classify riboswitches with $\tilde{9}9\%$ accuracy. *bioRxiv*, 2019.

[199] Deborah Antunes, Natasha A.N. Jorge, Ernesto R. Caffarena, and Fabio Passetti. Using RNA sequence and structure for the prediction of riboswitch aptamer: A comprehensive review of available software and tools. *Frontiers in Genetics*, 8:231, January 2018.

[200] Sumit Mukherjee, Matan Drory Retwitzer, Sara M Hubbell, Michelle M Meyer, and Danny Barash. A computational approach for the identification of distant homologs of bacterial riboswitches based on inverse RNA folding. *Briefings in Bioinformatics*, March 2023.

[201] F. Denis. PAC learning from positive statistical queries? *Lecture Notes in Computer Science*, 1501:112–126, January 1998.

[202] François Denis, Rémi Gilleron, and Fabien Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348:70–83, December 2005.

[203] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. 2008.

[204] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109:719–760, April 2020.

[205] Peng Yang, Xiaoli Li, Hon Nian Chua, Chee Keong Kwoh, and See Kiong Ng. Ensemble positive unlabeled learning for disease gene identification. *PLoS ONE*, 9:e97079, May 2014.

[206] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. *Computer Methods and Programs in Biomedicine*, 166:99–105, November 2018.

[207] Zhe Ju and S. Wang. Computational identification of lysine glutarylation sites using positive-unlabeled learning. *Current Genomics*, 21:204–211, May 2020.

[208] Fuyi Li, Shuangyu Dong, Andre Leier, Meiya Han, Xudong Guo, Jing Xu, Xiaoyu Wang, Shirui Pan, Cangzhi Jia, Yang Zhang, Geoffrey I. Webb, Lachlan J.M. Coin, Chen Li, and Jiangning Song. Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Briefings in Bioinformatics*, 23, January 2022.

[209] Chunlin Wang, Chris Ding, Richard F. Meraz, and Stephen R. Holbrook. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics*, 22:2590–2596, November 2006.

[210] Xiangxiang Zeng, Yue Zhong, Wei Lin, and Quan Zou. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Briefings in Bioinformatics*, 21:1425–1436, July 2020.

[211] Syed Danish Ali, Hilal Tayara, and Kil To Chong. Identification of piRNA disease associations using deep learning. *Computational and Structural Biotechnology Journal*, 20:1208–1217, January 2022.

[212] Congzhe Su, Jeffery D. Weir, Fei Zhang, Hao Yan, and Teresa Wu. ENTRNA: A framework to predict RNA foldability. *BMC Bioinformatics*, 20:1–11, July 2019.

[213] Junyi Zhou, Xiaoyu Lu, Wennan Chang, Changlin Wan, Xiongbin Lu, Chi Zhang, and Sha Cao. PLUS: Predicting cancer metastasis potential based on positive and unlabeled learning. *PLOS Computational Biology*, 18:e1009956, March 2022.

[214] A. I. Petrov, S. J. E. Kay, I. Kalvari, K. L. Howe, K. A. Gray, E. A. Bruford, P. J. Kersey, G. Cochrane, R. D. Finn, A. Bateman, A. Kozomara, S. Griffiths-Jones, A. Frankish, C. W. Zwieb, B. Y. Lau, K. P. Williams, P. P. Chan, T. M. Lowe, J. J. Cannone, R. Gutell, M. A. Machnicka, J. M. Bujnicki, M. Yoshihama, N. Kenmochi, B. Chai, J. R. Cole, M. Szymanski, W. M. Karlowski, V. Wood, E. Huala, T. Z. Berardini, Y. Zhao, R. Chen, W. Zhu, M. D. Paraskevopoulou, I. S. Vlachos, A. G. Hatzigeorgiou, L. Ma, Z. Zhang, J. Puetz, P. F. Stadler, D. McDonald, S. Basu, P. Fey, S. R. Engel, J. M. Cherry, P. J. Volders, P. Mestdagh, J. Wower, M. B. Clark, X. C. Quek, and M. E. Dinger. RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Research*, 45(D1):D128–D134, Jan 2017.

[215] Ioanna Kalvari, Eric P Nawrocki, Joanna Argasinska, Natalia Quinones-Olvera, Robert D Finn, Alex Bateman, and Anton I Petrov. Non-coding RNA analysis using the Rfam database. *Current Protocols in Bioinformatics*, June 2018.

[216] G. Grillo, A. Turi, F. Licciulli, F. Mignone, S. Liuni, S. Banfi, V. A. Gennarino, D. S. Horner, G. Pavesi, E. Picardi, and G. Pesole. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research*, 38(Database issue):75–80, Jan 2010.

[217] Claudio Lo Giudice, Federico Zambelli, Matteo Chiara, Giulio Pavesi, Marco Antonio Tangaro, Ernesto Picardi, and Graziano Pesole. UTRdb 2.0: a comprehensive, expert curated catalog of eukaryotic mRNAs untranslated regions. *Nucleic Acids Research*, 51:D337–D344, January 2023.

[218] A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021–3030, May 1985.

[219] Phillip E.C. Compeau, Pavel A. Pevzner, and Glenn Tesler. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991, November 2011.

[220] Natapol Pornputtapong, Daniel A. Acheampong, Preecha Patumcharoenpol, Piroon Jenjaroenpun, Thidathip Wongsurawat, Se Ran Jun, Suganya Yongkiettrakul, Nipa Chokesajjawatee, and Intawat Nookaew. KITSUNE: A tool for identifying empirically optimal K-mer length for alignment-free phylogenomic analysis. *Frontiers in Bioengineering and Biotechnology*, 8:1080, September 2020.

[221] Mark E. Fornace, Nicholas J. Porubsky, and Niles A. Pierce. A unified dynamic programming framework for the analysis of interacting nucleic acid strands: Enhanced models, scalability, and speed. *ACS Synthetic Biology*, 9:2665–2678, October 2020.

[222] Mark E. Fornace, Jining Huang, Cody T. Newman, Nicholas J. Porubsky, Marshall B. Pierce, and Niles A. Pierce. NUPACK: Analysis and design of nucleic acid structures, devices, and systems. *Theoretical and Computational Chemistry*, November 2022.

[223] Darcy McRose, Jian Guo, Adam Monier, Sebastian Sudek, Susanne Wilken, Shuangchun Yan, Thomas Mock, John M. Archibald, Tadhg P. Begley, Adrian Reyes-Prieto, and Alexandra Z. Worden. Alternatives to vitamin B1 uptake revealed with discovery of riboswitches in multiple marine eukaryotic lineages. *The ISME journal*, 8:2517–2529, January 2014.

[224] Sunita Yadav, D. Swati, and Hariharan Chandrasekharan. Thiamine pyrophosphate riboswitch in some representative plant species: a bioinformatics study. *Journal of Computational Biology*, 22:1–9, January 2015.

[225] Andreas Wachter, Meral Tunc-Ozdemir, Beth C. Grove, Pamela J. Green, David K. Shintani, and Ronald R. Breaker. Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs. *The Plant Cell*, 19:3437–3450, November 2007.

[226] Ming T. Cheah, Andreas Wachter, Narasimhan Sudarsan, and Ronald R. Breaker. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, 447:497–500, April 2007.

[227] Sanshu Li and Ronald R. Breaker. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic Acids Research*, 41:3022–3031, March 2013.

[228] Sumit Mukherjee, Matan Drory Retwitzer, Danny Barash, and Supratim Sengupta. Phylogenomic and comparative analysis of the distribution and regulatory patterns of TPP riboswitches in fungi. *Scientific reports*, 8, December 2018.

[229] Xun Wang, Can Fang, Yifei Wang, Xinyu Shi, Fan Yu, Jin Xiong, Shan-Ho Chou, and Jin He. Systematic comparison and rational design of theophylline riboswitches for effective gene repression. *Microbiology Spectrum*, 11, February 2023.

[230] Robert D. Jenison, Stanley C. Gill, Arthur Pardi, and Barry Polisky. High-resolution molecular discrimination by RNA. *Science*, 263(5152):1425–1429, March 1994.

[231] Huaiyu Mi, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D Thomas. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47:419–426, January 2018.

[232] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology NIH public access author manuscript. *Nature Genetics*, 25:25–29, May 2000.

[233] The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 2021.

[234] Matthias Mack, Ute Schniegler-Mattox, Verena Peters, Georg F. Hoffmann, Michael Liesert, Wolfgang Buckel, and Johannes Zschocke. Biochemical characterization of human 3-methylglutaconyl-CoA hydratase and its role in leucine metabolism. *The FEBS Journal*, 273:2012–2022, May 2006.

[235] Jing Li, Yan-Nan Wang, Bei-Si Xu, Ya-Ping Liu, Mi Zhou, Tao Long, Hao Li, Han Dong, Yan Nie, Peng R Chen, En-Duo Wang, and Ru-Juan Liu. Intellectual disability-associated gene FTSJ1 is responsible for 2'-O-methylation of specific tRNAs. *EMBO reports*, 21:e50095, August 2020.

[236] Agustín Hidalgo-Gutiérrez, Pilar González-García, María Elena Díaz-Casado, Eliana Barriocanal-Casado, Sergio López-Herrador, Catarina M. Quinzii, and Luis C. López. Metabolic targets of coenzyme Q10 in mitochondria. *Antioxidants*, 10, April 2021.

[237] Shanker S.S. Panchapakesan, Lukas Corey, Sarah N. Malkowski, Gadareth Higgs, and Ronald R. Breaker. A second riboswitch class for the enzyme cofactor NAD+. *RNA*, 27:99–105, January 2021.

[238] Sarah N. Malkowski, Tara C.J. Spencer, and Ronald R. Breaker. Evidence that the nadA motif is a bacterial riboswitch for the ubiquitous enzyme cofactor NAD+. *RNA*, 25:1616–1627, December 2019.

[239] Runa Njälsson and Svante Norgren. Physiological and pathological aspects of GSH metabolism. *Acta Paediatrica*, 94:132–137, January 2005.

[240] Faik N. Musayev, Martino L. Di Salvo, Tzu-Ping Ko, Verne Schirch, and Martin K. Safo. Structure and properties of recombinant human pyridoxine 5'-phosphate oxidase. *Protein Science : A Publication of the Protein Society*, 12:1455, July 2003.

[241] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, March 2009.

[242] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[243] Eugenio Mattei, Gabriele Ausiello, Fabrizio Ferrè, and Manuela Helmer-Citterich. A novel approach to represent and compare RNA secondary structures. *Nucleic Acids Research*, 42:6146, June 2014.

[244] Tomasz Kalwarczyk, Natalia Ziębacz, Anna Bielejewska, Ewa Zaboklicka, Kaloian Koynov, Jędrzej Szymański, Agnieszka Wilk, Adam Patkowski, Jacek Gapiński, Hans Jürgen Butt, and Robert Hołyst. Comparative analysis of viscosity of complex liquids and cytoplasm of mammalian cells at the nanoscale. *Nano Letters*, 11:2157–2163, April 2011.

[245] Hani Goodarzi, Hoang C.B. Nguyen, Steven Zhang, Brian D. Dill, Henrik Molina, and Sohail F. Tavazoie. Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell*, 165(6):1416–1427, June 2016.

[246] Doowon Huh, Maria C Passarelli, Jenny Gao, Shahnoza N Dusmatova, Clara Goin, Lisa Fish, Alexandra M Pinzaru, Henrik Molina, Zhiji Ren, Elizabeth A McMillan, Hosseinali Asgharian, Hani Goodarzi, and Sohail F Tavazoie. A stress-induced tyrosine-tRNA depletion response mediates codon-based translational repression and growth suppression. *The EMBO Journal*, 40(2):e106696, December 2021.

[247] Anne Hoffmann, Jörg Fallmann, Elisa Vilardo, Mario Mörl, Peter F Stadler, and Fabian Amman. Accurate mapping of tRNA reads. *Bioinformatics*, 34(7):1116–1124, April 2018.

[248] Yulong Wei, Jordan R. Silke, and Xuhua Xia. An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Scientific Reports*, 9(1):1–11, December 2019.

[249] Haifei Xu, Juan Ji An, and Baoji Xu. Distinct cellular toxicity of two mutant huntingtin mRNA variants due to translation regulation. *PLoS ONE*, 12, 2017.

[250] Annika Heinz, Deepti Kailash Nabariya, and Sybille Krauss. Huntingtin and its role in mechanisms of RNA-mediated toxicity. *Toxins*, 13:487, July 2021.

[251] Sarah J. Tabrizi, Michael D. Flower, Christopher A. Ross, and Edward J. Wild. Huntington disease: new insights into molecular pathogenesis and therapeutic opportunities. *Nature Reviews Neurology*, 16:529–546, October 2020.

[252] Peggy C. Nopoulos. Huntington disease: A single-gene degenerative disorder of the striatum. *Dialogues in Clinical Neuroscience*, 18:91–98, March 2016.

[253] Ricardo Ehrlich, Marcos Davyt, Ignacio López, Cora Chalar, and Mónica Marín. On the track of the missing tRNA genes: A source of non-canonical functions? *Frontiers in Molecular Biosciences*, 8:84, March 2021.

[254] Jeremy T. Lant, Rashmi Kiri, Martin L. Duennwald, and Patrick O'Donoghue. Formation and persistence of polyglutamine aggregates in mistranslating cells. *Nucleic Acids Research*, 49:11883–11899, November 2021.