Understanding *how* assembly dynamics of the cleavage and polyadenylation machinery ensures efficient and regulatable transcription termination.

I propose to develop a mathematical modeling framework for the dynamic assembly of the cleavage and polyadenylation (CPA) machinery. This framework will predict the time that it takes the CPA machinery to assemble and terminate transcription to produce RNAs with poly-adenosine (A) tails. Developed models will provide mechanistic insights for the CPA process and generate experimentally testable hypotheses for essential mechanisms that ensure proper CPA function. Outcomes of this study will guide future experimental studies for CPA and related mechanisms. Ultimately, the modeling framework will be updated with future experimental observations to form an iterative cycle of modeling and experimentation that unveils this mystery of gene expression.

The CPA mechanism is a necessary molecular process that mediates transcription termination and adds a poly(A) tail to newly transcribed RNAs. Excluding histone RNA transcripts, the CPA process occurs to all other known eukaryotic RNA transcripts that are manufactured by RNA polymerase II (Pol II)[1]. Importantly, the CPA mechanism terminates transcription at designated sites, which is crucial for gene expression control. Particularly, the CPA machinery is assembled at the polyadenylation site (PAS) in newly transcribed RNAs. The machinery comprises multiple CPA factors that cooperatively bind to the PAS sequence as well as Pol II's C-terminal domain (CTD) (Fig. 1). The CPA machinery must be assembled within a reasonable timeframe for cleavage and polyadenylation to successfully occur at a cleavage site located near the PAS sequence. If any necessary CPA factors are not immediately available, assembly and function of the CPA machinery can be drastically slowed or completely fail altogether. That is, transcription termination fails. Consequently, RNA transcriptional readthrough (i.e. continued transcription of a successive gene) can occur. Furthermore, this process is regulatable because most RNA transcripts contain multiple PAS sequences. This allows cleavage and polyadenylation to occur at different locations on an RNA transcript, which produces different RNA isoforms. This process is called alternative polyadenylation (APA). *In sum, time-constrained dynamic interactions between CPA factors, Pol II's CTD, and PAS sequences on newly transcribed RNAs guide the CPA and APA mechanisms to manufacture specific mRNA transcripts. To ensure its biological function, the CPA machinery must be assembled with efficiency to prevent transcriptional readthrough and with proper selectivity of PAS sequences to allow desired APA regulation under various conditions.* Although the molecular components of the CPA machinery are largely clear[2], the dynamic interactions among these components remain poorly understood. This hinders the coherent understanding of *how* the CPA mechanism functions with efficiency and regulatability. *I propose to resolve this knowledge gap by developing mathematical models for the dynamics of CPA assembly based on experimental data and fundamental physical and chemical principles.*

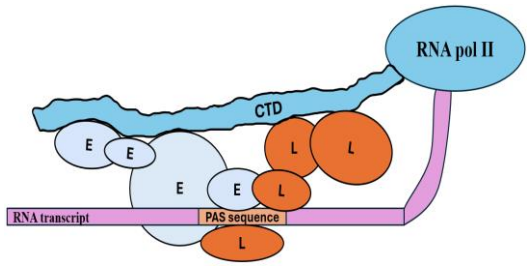

Figure 1: Cartoon of the CPA machinery's structure. E-factors can bind to the CTD before the PAS sequence is manufactured (blue). L-factors bind to PAS after the PAS sequence is manufactured (orange). E and L-factors cooperatively bind to PAS form a functional CPA machinery. *See Aim 1 for details.*

Aim 1: Develop mathematical models for CPA machinery assembly and identify possible mechanisms that ensure efficient CPA machinery assembly and function. The CPA machinery is estimated to assemble within 10 seconds after the PAS sequence on an RNA transcript is manufactured[3]. Yet, it is unclear how this timeframe is achieved when assembling so many components into a large machinery. In a preliminary study, we searched the proteomics database[4] and found that key CPA factors are approximately equal in abundance to actively transcribing Pol IIs within eukaryotic cells. Additionally, Pol II's CTD, which is crucial for CPA factor recruitment, consists of 26 (yeast) ~ 52 (vertebrates) tandem repeats of seven amino acids. Hence, each Pol II may contain multiples of binding sites for CPA factors. This could deplete CPA factors from the nucleoplasm, which could restrict the efficiency of CPA machinery assembly and function. *To evaluate the kinetics of CPA machinery assembly, I will develop a core model* (Fig. 1) *with the following experimental observation-based assumptions about the CPA machinery:* (i)

Some CPA factors are associated with Pol II throughout most of the RNA transcription process, while others are only associated after a PAS sequence appears[5]. Accordingly, I plan to categorize the CPA factors into two groups: early binders (E) and late binders (L). That is, E-factors bind to Pol II's CTD before the PAS sequence is manufactured and L-factors bind to PAS after the PAS sequence is manufactured. The E-Pol II and L-PAS complexes further bind to form the stable and functional CPA machinery. (ii) Each Pol II contains *N* binding sites for E-factors and competes for the nucleoplasmic unbound E-factors. *I will formulate a set of ordinary differential equations to depict the above biophysical processes.* These models will be able to predict how long it takes the CPA machinery to assemble given the physical and chemical constraints in the nucleus. Optional, mechanistic details, like multiple binding steps and cooperative binding, and variations in model parameters, like protein concentrations, will be added to the core model to test their effects. *I will use the model's predictions to identify possible mechanisms that are necessary to ensure efficient CPA assembly and prevention of transcriptional readthrough.* Proposed mechanisms can be tested by future experiments.

Aim 2: Develop mathematical models for the APA process and elucidate principles of PAS selection during APA. Up to 70% of all known mammalian mRNA-encoding genes contain multiple PAS sequences that can be selected by the APA mechanism to produce different RNA isoforms[6]. Genome-wide regulation of APA plays important roles in tissue specific gene expression, cell differentiation, cancer development, etc.[6] Mechanistically, CPA factors have different binding affinities to different PAS sequences, which is believed to be a major factor driving PAS selection[7]. However, proximal PAS sites could induce transcription termination before distal sites are even manufactured, thereby connecting PAS selection to transcription kinetics. Yet, a mechanistic framework that can predict PAS selection is unavailable. *I will develop a predictive mechanistic framework by incorporating into Aim 1's CPA model multiple PAS sites, where inter-PAS distances and CPA factor-PAS binding affinities will be informed by the experimental data in the literature*[7]. The APA framework will be able to predict the probability by which each PAS sequence is selected. Predictions will be compared to experimentally observed PAS selection frequency. PAS selection frequency data for different genes will be used for parameter fitting and model validation. Fitting data for different cells states to the model will reveal *how* APA is globally regulated in each cell state and *how* PAS sequences are arranged to ensure regulatability of APA. The findings will elucidate the general principles of APA and provide testable predictions for future experimental studies of APA regulation.

Intellectual Merit: The proposed research will provide a predictive mathematical modeling framework for the CPA and APA mechanisms and advance fundamental knowledge on the functionality and regulation of transcription termination. The model will provide a predictive tool for the outcomes of CPA and APA and reveal the design principles for transcription termination regulation at both single-gene and genomic levels. Additionally, the kinetics and mechanisms of the CPA machinery assembly learned from the proposed study can be generalized to other macromolecular machineries, which are prevalent in other cellular processes. Overall, findings from this project will enrich coherent understandings of the diverse regulation of RNA expression and highlight the regulatory steps of transcription termination.

Broader Impacts: The proposed project will have broader impacts in technological developments and STEM education. Technology-wise, outcomes of the research will enable innovations in bioengineering and synthetic biology tools, novel PAS site designs that allow desired APA regulations. Education-wise, I will utilize my academic background and training through the proposed research in my educational activities. Mathematical modeling has become increasingly crucial for modern biology research, but workforces with necessary skill sets do not measure up to the rising demands. I will actively engage in opportunities to teach biological modeling and cultivate interest in the subject, e.g., co-teaching the newly launched summer course on "mechanistic modeling of biological systems" at Virginia Tech. I will use my research as examples to demonstrate how mathematics, physics, chemistry, and biology interconnect, particularly *how* to make sensible models and *how* to make sense of model results.

[1]Millevoi, S. & Vagner, S. *Nuc. Acid. Res.* (2010). [2]Mitschka, S. & Mayr, C. *Nat. Rev. Mol. Cel. Bio.* (2022). [3]Chao, L. C. et al. *Mol. and Cel. Bio.* (1999). [4]Huang, Q. et al. *Mol. & Cel. Prot.* (2023). [5]Martin, G. et al. *Cel. Rep.* (2012). [6]Tian, B. & Manley, J. L. *Nat. Rev. Mol. Cel. Bio.* (2017). [7]Hamilton, K. et al. *RNA* (2019).