Thermodynamic model of bacterial transcription

Jin Qian,[*] David Dunlap,[†] and Laura Finzi[‡]

Physics Department, Emory University, 400 Dowman Dr., Atlanta, Georgia 30322, USA

(Received 7 March 2022; accepted 20 September 2022; published 12 October 2022)

Transcriptional pausing is highly regulated by the template DNA and nascent transcript sequences. Here, we propose a thermodynamic model of transcriptional pausing, based on the thermal energy of transcription bubbles and nascent RNA structures, to describe the kinetics of the reaction pathways between active translocation, elemental, backtracked, and hairpin-stabilized pauses. The model readily predicts experimentally detected pauses in high-resolution optical-tweezer measurements of transcription. Unlike other models, it also predicts the effect of tension and the GreA transcription factor on pausing.

DOI: 10.1103/PhysRevE.106.044406

I. INTRODUCTION

During bacterial transcription, there are frequent pauses in the forward translocation of RNA polymerase. Pauses observed *in vivo* and *in vitro* vary in durations from milliseconds to minutes [1,2]. Short pauses, which typically last less than one second and are referred to as elemental pauses, are proposed to be intermediate precursors of long pauses [3]. Long pauses, which may last tens of seconds, are classified as class I hairpin-stabilized and class II backtracked signals and have been structurally characterized and mechanistically explored [4,5]. They are thought to be regulated by the sequence of the DNA template, the structure of the nascent transcript, and the availability of transcription factors (TFs) [6–8].

Previous models of the kinetics of backtracked pauses predict some types of experimentally detected pauses [9–12] but fail to predict other types of pausing and pause duration, and do not treat external tension or TFs. Here, we propose a model based on our current biochemical understanding of transcription pausing mechanisms and optimize the parameters of the model with high-resolution transcription data. This purely thermodynamic model provides a mechanistic explanation of the effect of external tension and TFs, and after refinement accurately simulates experimentally observed pause sites and durations. Furthermore, the model accurately predicts transcription dynamics on unfamiliar DNA sequences not used for refinement and is readily extendable to incorporate the initiation and termination stages.

II. MODEL DESCRIPTION

A. Ternary transcription elongation complex configuration and state transition

Ternary transcription elongation (TEC) is described by a transcription position (m) and state (n). The position along the template (m) indicates the length of the RNA transcript. TEC can be in one of two translocation states: active ($n = 0$) or backtracked ($n < 0$), or in a conformationally distinct hairpin-stabilized state (hsp). The interconnection among these states is shown in Fig. 1(a). From an active state at position m (m, 0), a transcription complex can translocate to the next active state ($m + 1$, 0), or branch into backtracked (m, -1) or hairpin-stabilized states (m, hsp).

The energy of the TEC is estimated as the sum of four contributions: the free energy of the (i) transcription bubble, (ii) DNA-nascent RNA hybrid, (iii) nascent RNA, and (iv) RNAP-DNA:

$$G_{\mathrm{TEC}} = G_{\mathrm{bubble}} + G_{\mathrm{hybrid}} + G_{\mathrm{RNA}} + G_{\mathrm{RNAP\_binding}}. \qquad (1)$$

In this estimate, the first two terms are clearly sequence dependent, as is the secondary structure of nascent RNA (the third term). The fourth term represents interactions between the nucleic acids and RNAP subunits and is effectively constant and sequence independent, as argued previously [9–11].

To determine the configuration of a transcription bubble and the details of the energy profile of a TEC, we used an approach based on statistical mechanics, the basis of which was described by Tadigotla [10]. A transcription complex (m, n) is in a rapid equilibrium among many microstates, each defined by the parameter (b) which depends on the number of unpaired DNA bases upstream (u) and downstream (d) of the DNA-RNA hybrid inside the RNAP enzyme, the length of the hybrid (h) and the number of single-stranded RNA bases protected by RNAP (r) [Fig. 1(b)].

Equilibrium among microstates [dashed arrows in Fig. 1(b) is reached rapidly compared to the time required for state transitions. Thus, for each transcription complex (m, n), the probability of a particular microstate b is given by the Boltzmann distribution:

$$P_m^b = Z_m^{-1}\exp\left(\frac{-G_{\mathrm{TEC}}^{m,b}}{k_BT}\right), \qquad (2)$$

$$Z_m = \sum_b \exp\left(\frac{-G_{\mathrm{TEC}}^{m,b}}{k_BT}\right). \qquad (3)$$

[*] jin.qian@emory.edu
[†] ddunlap@emory.edu
[‡] lfinzi@emory.edu

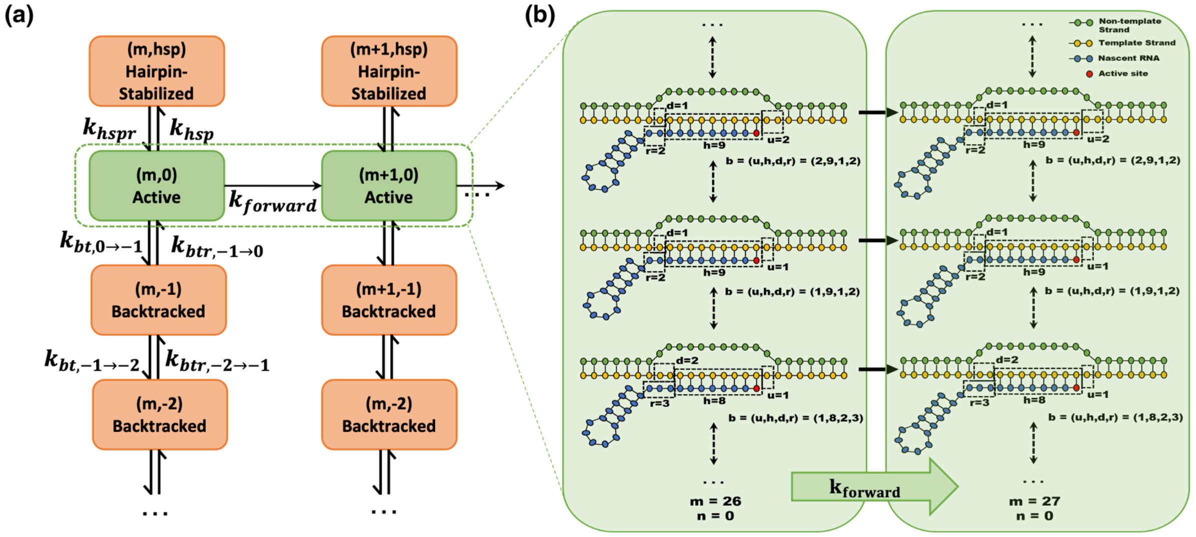

FIG. 1. State transitions in the model and the statistical approach to the transcription bubble configuration. (a) A diagram of transcriptional states considered in the model shows their interconnections. (b) An illustration of the statistical approach to characterize transcription bubble configurations including the forward translocation step. Dashed arrows indicate fast equilibrium and solid arrows indicate the allowed state transitions.

The overall forward translocation rate is calculated as

$$k_{m\to m+1} = \sum_{\mathrm{b}} P_m^b k_{m\to m+1}^b. \tag{4}$$

Figure 1(b) shows the forward translocation step as an example of statistical treatment in the model. All state transitions in the model are determined according to Eq. (4), as the summation of the products of the probability and the forward translocation rate of individual microstates.

B. Forward translocation

The forward (active) translocation of RNAP is modeled by the Michaelis-Menten (M-M) equation

$$k_{\mathrm{forward}} = \frac{k_{\max}[\mathrm{NTP}]}{K_d(1+K_{\mathrm{i}})+[\mathrm{NTP}]}, \tag{5}$$

where $k_{\max}$ is the rate of NTP hydrolysis, K_d is the NTP dissociation constant, and K_{i} is the equilibrium constant between two adjacent translocation states determined by their base pairing energy. The equation is derived from the Brownian-ratchet model [13], in which forward translocation occurs in three steps: (i) a fast equilibrium between position m and position $m+1$, (ii) recruitment of NTP at active site, (iii) catalysis and release of pyrophosphate [Fig. 2(a)]. Fitting $k_{\max}$ and K_d of Eq. (5) to experimental data identifies slow translocation sites that precede the long-lived pauses. These slow translocation events are interpreted as pretranslocated, elemental pauses on the pathway of translocation. Further elaboration of the on-pathway and off-pathway characteristics of short pauses follows in the Discussion section.

C. Backtracking

Backtracking has been previously modeled using the Arrhenius Eq. (6) with an activation barrier of $40-50\ k_{\mathrm{B}}T$ for each step of backward translocation [9]. This value seems unreasonably high given that the free energy of base pairing in a transcription bubble is typically less than $-20\ k_{\mathrm{B}}T$ [10]:

$$k_{\mathrm{bt}} = k_1\ \exp\left(-\Delta G/k_{\mathrm{B}}T\right). \tag{6}$$

We take the same Arrhenius approach but treat the first step of backtracking differently from the subsequent ones [Fig. 2(b)], based on the assumption that initially the 3′ end of the nascent transcript blocks the active site and subsequently invades the secondary channel of RNAP [14], while additional backtracking stabilizes the interaction of RNA within the secondary channel.

We assume the energy barrier for an active TEC to enter the backtracked state to be

$$\Delta G_{0\to-1} = \Delta G_{\mathrm{bt}} - G_0, \tag{7}$$

where ΔG_{bt} is a fixed activation energy specific for entering a backtracked state. We can assume that ΔG_{bt} will be limited to the energy available from complete collapse of the bubble, which is estimated to be in the range $-(10\sim20)k_{\mathrm{B}}T$. ΔG_0 is the energy of a TEC at an active site. The rate constant to enter the backtracked state from the active state (0) would be

$$k_{0,\mathrm{bt}} = k_1 \exp\left(-\Delta G_{0\to-1}/k_{\mathrm{B}}T\right), \tag{8}$$

where k_1 is the prefactor of backtracking.

For any further backward translocation of RNAP, the energy barrier should relate to the energy difference between two adjacent translocation states and the backtracked distance. Thus, for $n>0$,

$$\Delta G_{-n\to-n-1} = \Delta G_{\mathrm{bt\_increment}} + 0.5(G_{-\mathrm{n}} - G_{-\mathrm{n}-1}) \tag{9}$$

and

$$k_{-n,\mathrm{bt}} = k_1 \exp(-\Delta G_{-n\to-n-1}/k_{\mathrm{B}}T), \tag{10}$$

where $\Delta G_{\mathrm{bt\_increment}}$ represents the backtracking energy barrier due to increase in the length of the transcript inserted into the secondary channel.

The model considers that the recovery from a backtracked state (k_{btr}) can be achieved by two pathways: a diffusive and

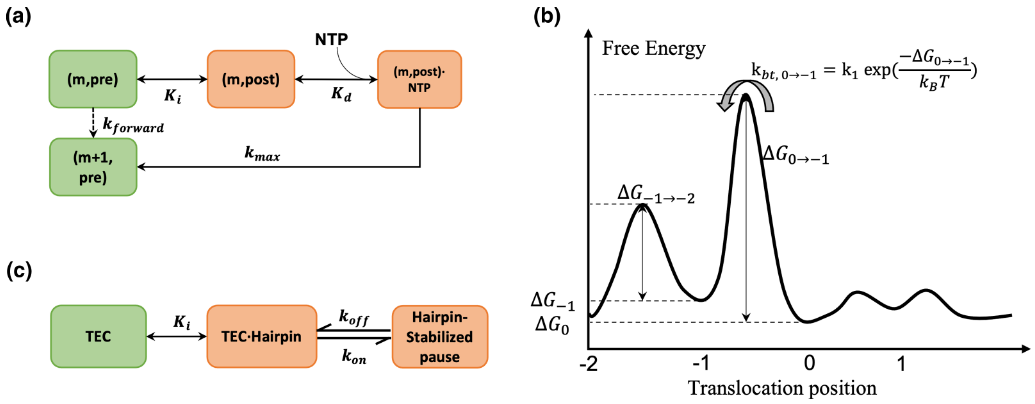

FIG. 2. Model construction. (a) An illustration of RNAP forward translocation using Michaelis-Menten equation. (b) The free energy landscape for the backtracking pathways. Note that the first backtracking step has different energy barrier than the deeper backtracking steps. (c) The proposed kinetic mechanism for the hairpin-stabilized pause.

a cleavage pathway. The former occurs through RNAP diffusion, which also follows the Arrhenius equation with the energy barriers described above:

$$k_{-n-1,\mathrm{btr}} = k_1 \exp(-\Delta G_{-n-1\to-n}/k_\mathrm{B}T) \tag{11}$$

and

$$k_{-1,\mathrm{btr}} = k_1 \exp\left(-\Delta G_{-1\to 0}/k_\mathrm{B}T\right). \tag{12}$$

The cleavage pathway occurs by cleaving nascent RNA inserted into the secondary channel to register the 3′ end of nascent RNA in the active site. This process is likely to be sequence independent, and was assumed to occur at a constant rate.

Hypertranslocation, which refers to the forward translocation of RNAP without concurrent RNA elongation at the active site, is a pausing event translocationally similar to backtracking. However, we do not include hypertranslocation in the model for two reasons. First, hypertranslocation may not be a general phenomenon during transcription [15], and it cannot be distinguished from backtracking in force spectroscopy assays. Second, hypertranslocation is never energetically favored because the extent of base pairing is reduced with respect to the active state.

D. Hairpin-stabilized pausing

To model a hairpin-stabilized pause, we take an allosteric view, in which an RNA hairpin contacts a short α helix at the tip of the RNAP flap domain that covers the RNA exit channel to induce the pause [4,16]. The pathway is modeled as a fast equilibrium between two configurational states, a state free of hairpin and a state with a hairpin positioned close to the RNAP flap domain. The equilibrium is followed by a rate-limiting catalytic step [Fig. 2(c)]. The equilibrium is considered rapid compared to the formation of chemical bonds that stabilize the inactive state.

We use Eq. (13) to model the entry rate to the hairpin-stabilized pause,

$$k_\mathrm{hsp} = k_\mathrm{on}/(1 + K_{i,h}), \tag{13}$$

where k_on is the catalytic rate of interaction between the RNA hairpin loop and the RNAP flap interaction, and $K_{i,h}$ is the fraction of hairpin formation. Equation (14) gives the expression for $K_{i,h}$, which represents the equilibrium among all possible RNA secondary structures. The secondary structure of RNA transcript rapidly transitions among many microstates, and the simulation of transitions among these microstates is computationally expensive. We bypass this difficulty by simplifying the equilibrium to a two-state system of the lowest energy state and the hairpin-included state:

$$K_{i,h} = \exp\left(\frac{G_\mathrm{lowest} - G_\mathrm{hairpin\_included}}{k_B T}\right). \tag{14}$$

In the absence of RNase A, which digests the nascent RNA transcript, the lowest energy state is determined by allowing all or at most a 100-nucleotide-long stretch of RNA outside of the exit channel to fold freely. A state including a hairpin is determined by first searching from the 3′ end of RNA for possible hairpin structures near the exit channel (up to 30 nt) before allowing up to 100 of the remaining ribonucleotides of the transcript to fold freely. The equilibrium between the lowest energy state and the hairpin state can be used to estimate the fraction of hairpin formation. In presence of RNase A, the length of freely folded RNA is shortened to 15 nt, which may eliminate or generate pause stabilizing hairpins (see below: Comparison of the model with experimental data).

A chemical bond between the hairpin loop and the RNAP flap is required to stabilize the hairpin-flap interaction. The catalytic rate relates to the length of stem and loop, and the fraction of G and C in the loop as shown below:

$$k_\mathrm{on} = k_2 \exp\left(-\frac{D_\mathrm{stem} * \Delta G_\mathrm{stem} + D_\mathrm{loop} * \Delta G_\mathrm{loop} + F_{GC} * \Delta G_{GC}}{kT}\right), \tag{15}$$

where k_2 is the prefactor, D_{stem} and D_{loop} are the deviation from optimal lengths of stem (3–8 bases) and loop (4–20 bases), respectively, F_{GC} is the fraction of G and C nucleotides within the loop, and ΔG_{stem}, ΔG_{loop}, and ΔG_{GC} are the energy changes due to D_{stem}, D_{loop}, and F_{GC}.

The exit rate from a hairpin-stabilized paused state (k_{hspr}) must be much slower than the entry rate, and is determined by the rate of RNAP hairpin denaturation. For simplicity, the rate is taken to be a constant.

E. The effect of tension and transcriptional factors

Tension and TFs have been reported as critical components that can affect and even determine the transcription products by adjusting the energy profile of transcription complex and/or interacting with transcription machinery [1,4]. The effects of external tension and TFs on the thermodynamics of TEC were considered in our model. For the forward translocation and backtracking pathways, we employed the idea that the equilibrium constant in forward translocation step K_i and the energy barrier of backtracking step $\Delta G_{n\to n-1}$ is modulated by the work produced by tension [17] and the presence of GreB factors,

$$K_i^* = \exp\left(\frac{G_{\text{post}} - G_{\text{pre}} - F * L_{\text{forward}}}{k_B T}\right) \qquad (16)$$

and

$$\Delta G^*_{n\to n-1} = \Delta G_{n\to n-1} + \Delta G_{\text{GreB}} + F * L_{\text{bt}}, \qquad (17)$$

where G_{pre} and G_{post} are the energy of TEC in pre- and posttranslocation states, respectively, L_{forward} and L_{bt} are the effective lengths over which external tension acts in the forward translocation step and in the backtracking step, respectively, and ΔG_{GreB} is the energy barrier change due to GreB factor.

The hairpin-stabilized pause was assumed to be unaffected by any applied tension, since it does not involve RNAP translocation, but the length of a freely folded RNA transcript can be limited by the presence of RNase A, as stated in previous sections. Since the experimental data we used to validate the model were acquired under tension of magnitude ranging from −7pN to 25pN and in presence (absence) of GreB and RNase A, we quantitatively determined the effect of tension and TFs by fitting the model with data acquired under different experimental conditions.

F. Model training

It is important to note that transcription is a process that involves only very small numbers of reactants, thus the rates cannot be determined from the chemical law of mass action. Rather, we apply two stochastic methods: (i) the continuous-time Markov chain and (ii) stochastic simulation. The continuous-time Markov chain allows us to analytically solve for the expected time spent in each state at a certain position. The stochastic simulation reveals how individual pausing events develop. The details are given in the Methods section.

The model is encapsulated in a MATLAB class object, which can generate a predicted residence time histogram with

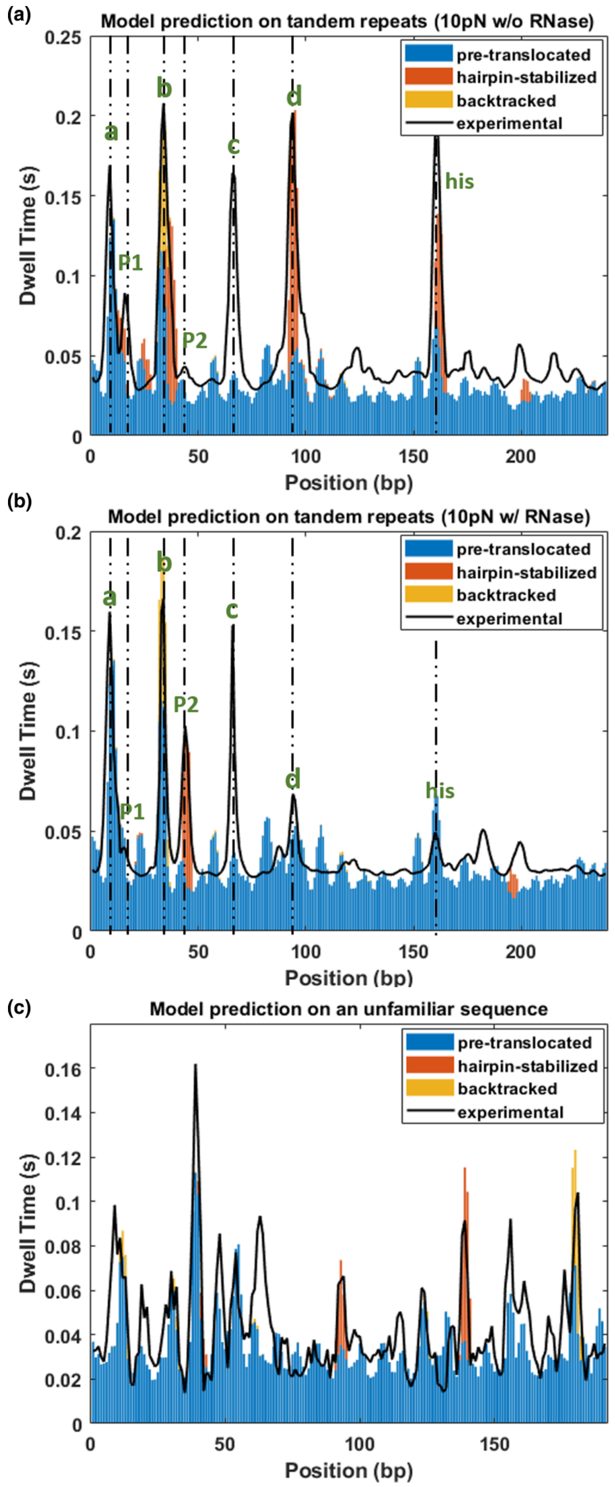

FIG. 3. Model fitting and prediction. (a) Stacked histogram produced by the model for the condition of 10 pN in presence of RNase. The residence time due to different pausing mechanisms is represented by different colors. The experimental result is shown by the black line. Goodness of fitting is 0.948 for the major pause sites except for c and 0.884 for the overall histogram. (b) Stacked histogram produced by the model for the condition of 10 pN in absence of RNase. Goodness of fitting is 0.959 for the major pause sites except for c and 0.904 for the overall histogram. (c) Predicted histogram by the model on an unfamiliar sequence. Goodness of fitting is 0.871 for the overall histogram.

TABLE I. Summary of experimental pause positions, durations, and mechansims for 10 pN under different transcriptional factor conditions.

		Averaged duration(s)			
Pause	Position of peak (bp)	WT	+GreB	+RNase	Associated state(s)
a	9	0.66	0.58	0.64	Pretranslocated
b	34	0.94	1.27	0.59	Backtracked + hairpin stabilized
c	66	0.42	0.41	0.38	Unknown
d	94	0.74	0.96	0.33	Hairpin stabilized
his	161	0.68	0.95	0.25	Hairpin stabilized
P1	16	0.41	0.40	0.25	Hairpin stabilized
P2	44	0.16	0.17	0.34	Hairpin stabilized (with RNase)

the input of a template sequence and a guess of unknown parameters. Thus, the model can be trained with the data from real-time single molecule experiments. We used the time series obtained in high-resolution optical-tweezer transcription experiments by Gabizon *et al.* [18], with or without GreB and RNase A. The transcription experiments were performed on a DNA template (8XHis) containing the T7A1 promoter followed by eight tandem repeats of a 239 bp sequence containing the *his*-leader pause site and four other known sequence-dependent pause sites [1]. The temporal resolution is high enough to detect pausing events longer than 100 ms. For transcription rates of 10–20 bp/s, this is sufficient to resolve pauses with one base-pair resolution using optical tweezers. Alignment of the traces under different forces and with different TFs generates the residence time histograms (Fig. 3) as described previously.

G. Comparison of the model with experimental data

Experimental data under different conditions with various accessory factors helped to expose the mechanisms of the pauses. Also, the analysis of the backtracking dynamics helped differentiate backtracked pauses from others. Table I summarizes the position and duration of pauses as well as their response to GreB or RNase (factors). Pauses at position "a" are likely pretranslocation, since their duration is barely affected by the addition of GreB or RNase. Pauses at position b are likely due to both backtracking and hairpin-stabilization, as their duration responds to the presence of GreB and RNase, and they are preceded by a backward RNAP translocation, as previous analysis suggests [18]. P1, d, and *his* are hairpin-stabilized pauses that almost disappear in the presence of RNase. Pause P2 is also hairpin-related, but unlike hairpin-stabilized pause P1, d, and *his*, it only appears in the presence of RNase.

We optimized the values of the model parameters (Table II) to produce a dwell time histogram that resembled the experimental data [Figs. 3(a) and 3(b)]. The model clearly reproduces the positions and lifetimes of pauses observed experimentally except for pause c We propose possible reasons why the model fails at pause c in the Discussion section. The model also successfully predicts the mechanisms of the pauses suggested by the experimental results [Fig. 4(a)]. The experimental data suggest that the presence of GreB extends

TABLE II. Values (95% confidence interval from 100 bootstrapped values) of the optimized parameters under 10 pN assisting force and WT conditions.

Parameters and descriptions		Symbol and value	Note
Forward Translocation	Rate of NTP catalysis for AUCG	$k_{\max} = [85(9), 77(5), 82(9), 41(3)]s^{-1}$	Fitted
	Equilibrium constant centerfor AUCG	$K_d = [34(3), 96(9), 15(2), 26(4)]\mu M$	
	Effective length for forward translocation	$L_{\text{forward}} = 0.56(0.07)\text{bp}$	
Backtracking	Prefactor of backtracking	$k_1 = 1000s^{-1}$	Fixed
	Energy barrier height of first base-pair backtracking	$G_{\text{bt}} = 9.8(0.8)k_BT$	Fitted with fixed $k_{\max}$ and K_d
	Energy barrier height of deeper backtracking	$G_{\text{bt\_incre}} = 1.8(0.1)k_BT$	
	Effective length for backtracking	$L_{\text{bt}} = 0.06(0.01)\text{bp}$	
Hairpin-stabilized pause	Energy change due to unlikely stem length	$\Delta G_{\text{stem}} = \text{Inf}$	Fixed values
	Energy change due to unlikely loop size	$\Delta G_{\text{loop}} = \text{Inf}$	
	Energy change due to *GC* fraction	$\Delta G_{GC} = 8.8(1.1)k_BT$	Fitted with fixed $k_{\max}$ and K_d and backtrack related parameters
	Hairpin-flap interaction rate	$k_{\text{on}} = 807(71)s^{-1}$	
	Hairpin denaturation rate	$k_{\text{hspr}} = 3.4(0.3)s^{-1}$	
	Allowed RNA-DNA hybrid length	$h = 7 \sim 9$ bp	
TEC structure	Allowed upstream spacer length	$u = 1 \sim 3$ bp	Fixed range
	Allowed downstream spacer length	$d = 1 \sim 3$ bp	
	Allowed number of single-stranded RNA protected by RNAP	$r = 1 \sim 3$ bp	

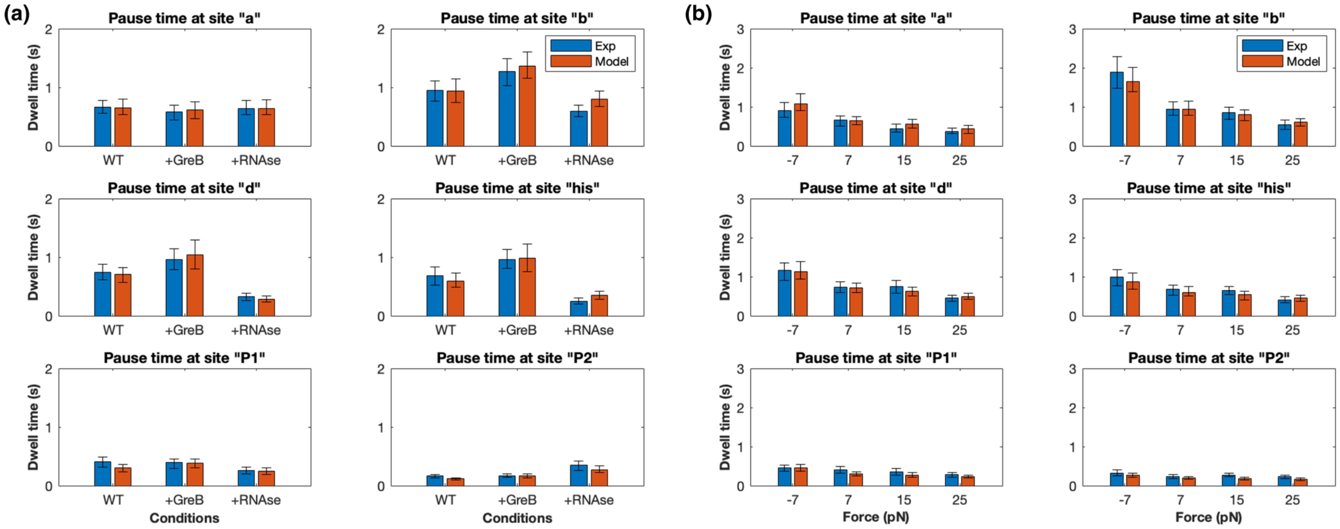

FIG. 4. Averaged dwell times from experiments (blue) and model (red) at pause sites. (a) With various transcriptional factor conditions under 10 pN assisting tension and (b) WT condition under different tensions. Error bars are the 25th and 75th percentile of 100 bootstrapped values.

the dwell time at pause b [18]. The model achieves this effect by adjusting the energy barrier of backtracking. The presence of RNase significantly decreases the dwell time at sites P1, d, and *his*, increases the dwell time at sites P2, and has little effect on the duration of pauses at other sites. Constraining the model to operate on shorter nascent RNA reproduces the observed changes in pause times by slowing or destabilizing hairpin formation at pause sites P1, d, and *his* while favoring the hairpin formation at pause site P2 (Fig. 5).

The effect of tension is modeled by introducing two different effective lengths L_{forward} and L_{bt} for forward and backtracking translocations, respectively [Fig. 4(b)]. Notice that the effective length for the forward translocation pathway is shorter than 1 base, while the external force acts on an effective length shorter than 0.1 base during backtracking (Table II). The fitted values of effective length agree with those from previous work [13,19]. These results indicate that opposing tension extends pauses by decreasing the transcription rate and accentuating the entry into backtracked pausing. It also supports the idea that the entry into long-lived pauses, such as backtracked pauses, follows short-lived pauses.

The predictive power of the model is demonstrated by the fact that it accurately predicts major pauses in the transcription of an unfamiliar 200-base sequence. This sequence preceding the repeat region of the 8XHis template was not included in the data used to optimize the model parameters. Figure 3(c) shows that the model successfully predicts the main pauses near bases 15, 45, 140, and 180 found experimentally by aligning transcription records and histogramming the dwell times.

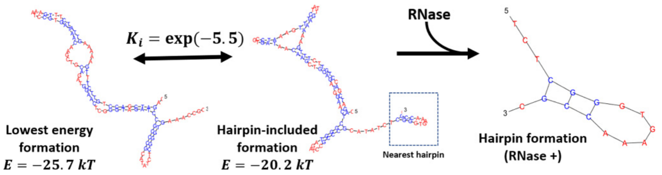

FIG. 5. Comparison of the lowest energy conformation with one including a proximal (3′) hairpin at position 44 (P2). Hairpin formation is unfavorable at this position without RNase. In the presence of RNase, the length of freely folded RNA is limited to 15 nt, so hairpin formation is favored.

To further test the validity of the model, we used Monte Carlo simulations to generate a large number of transcription traces, and we compared the dynamics of backtracking in experimental and simulated traces. The pauses at site b in simulated traces were analyzed for backtrack depth and backtrack duration (Fig. 6). The clear agreement between simulated and experimental results lends further support to the model.

III. DISCUSSION

A. Strengths and limitations of the model

Note that, in the model, short, pretranslocated pauses (also referred to as ubiquitous or elemental pauses elsewhere) are identified as positions of slow forward translocation which is on-pathway. In other reports, short pauses were identified as off-pathway events that branch off from the active translocation pathway. The dwell time data from Gabizon *et al.* [18] follows a power-law distribution up to 4–5 seconds as shown in Fig. 7. There is no indication of decay from an off-pathway elemental pause state. This led us to assume an on-pathway elemental pause state and fitting K_d and $k_{\max}$, Eq. (5) gives a good agreement at elemental pause a and predicts the slow translocation rates at other long-lived pause sites. Nevertheless, the current model, like others, cannot definitively place elemental pauses on- or off-pathway, because differentiating slow translocations from actual pausing is difficult. In fact, the model shows greater agreement with the experimental data at major, long-lived pause sites than elsewhere (see Fig. 3).

The previously reported values of K_d and $k_{\max}$ generate different pause sites from those observed in the experimental data examined here (Fig. 8) [12]. This may reflect the fact that those values originated from modeling transcription data

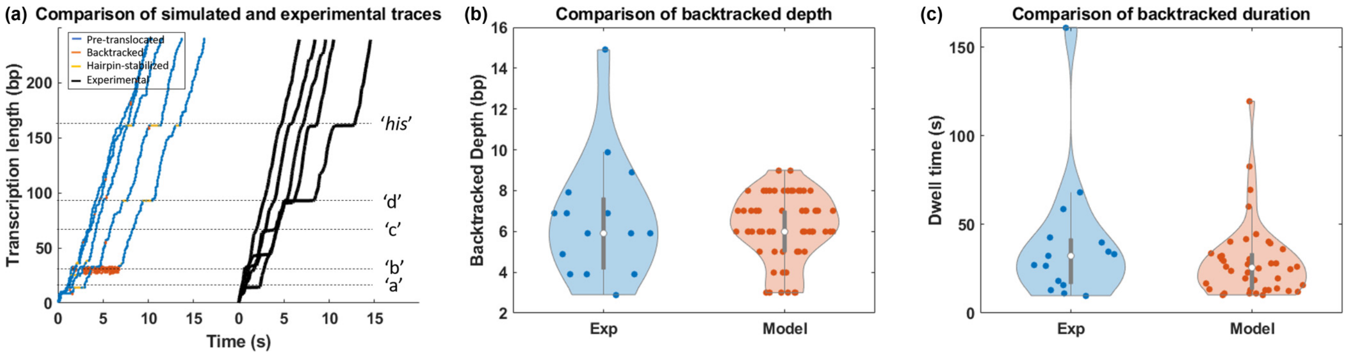

FIG. 6. Comparison between the backtracking dynamics in experimental and simulated data. (a) Examples of traces generated by Monte Carlo simulation. The simulated traces show similar pauses at sites a, b, d, and *his*, and generate comparable transcription rates to experimental data. (b) Distributions of backtrack depth observed experimentally and predicted by the model. (c) Distributions of backtrack duration observed in the experiments and predicted by the model.

without localizing pauses. Alternatively, this difference might indicate that an on-pathway state does not fully describe elemental pauses. However, the goodness of fitting and accuracy of prediction indicate that an on-pathway system patterned on the M-M equation has sufficient complexity to effectively model the elemental pauses. Given that the off-pathway events may involve unknown rearrangement of active sites of RNAP, fitting a system of suitable complexity is a good approach to bypass the difficulty in modeling off-pathway elemental pauses [20]. With our fitted values but not the previously reported values of parameters, the M-M equation predicts a slow translocation rate of 3.4 bp/s at a consensus elemental pause site identified using NET-seq [7], which lends further support for fitting K_d and $k_{\max}$ to produce correct pause sites.

The model identifies transcriptional pausing sites and correctly characterizes the mechanism of pausing. Our results support a previous theoretical analysis of transcriptional pauses which suggests that long-lived pauses develop from short-lived, more ubiquitous pauses [3]. For example, at pause site b, backtracking is favored over forward translocation because of the low forward translocation rate. Indeed, the energetic parameters of the model would predict comparable backtracking rates at the 35 bp site (pause b) and at the 190 bp site, but the fast-forward translocation rate at the 190 bp site diminishes backtracking (Fig. 9). Using the canonical Michaelis-Menten expression, we determined that the forward translocation rate along the template varies from less than 3 nt/s to 70 nt/s. This implies that a slowly transcribing complex may enter into a long-lived pause at one site, even if the backtracking energy barrier at this position is higher than the barrier height at a position where transcription is faster.

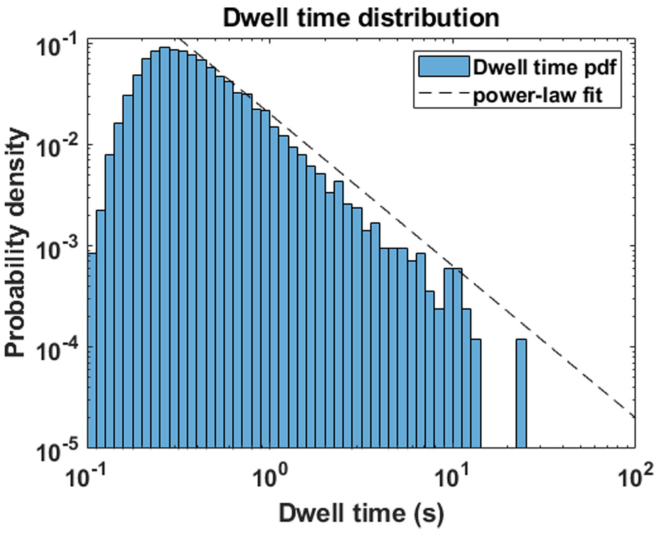

FIG. 7. The probability density distribution of dwell times in the transcription records. Dwell times ranged between 0.1 and 10 s with a power law distribution, representing a single, on-pathway state between 0.5 and 4 s and superposition of dwell times from other states between 4 and 10 s.

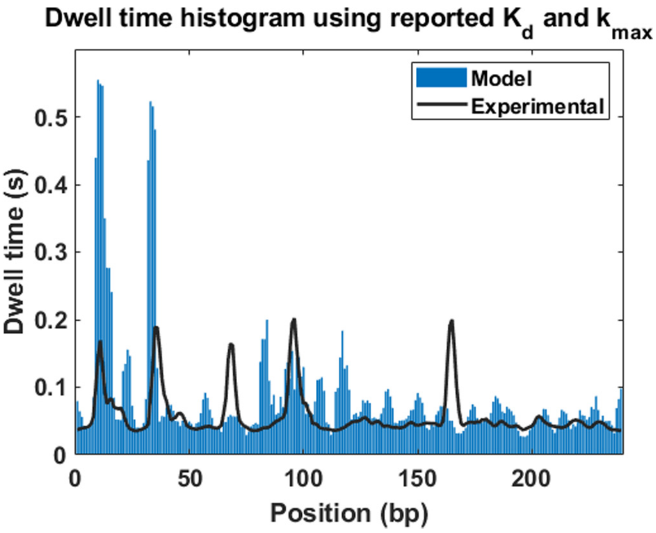

FIG. 8. Comparison between the experimental histogram and the dwell time histogram generated by fixing K_d and $k_{\max}$ using previously reported values [12]. Exceptionally long pauses are predicted at sites a (9 bp) and b (34 bp) and at 90 and 110 bp, where there are no significant, experimentally observed pauses.

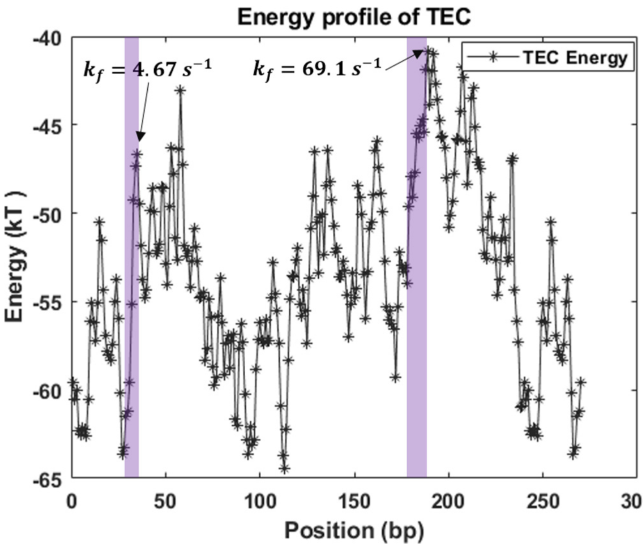

FIG. 9. Backtracking probability is highly dependent on the forward rate. The purple bands indicate two backtracking favored energy profiles. The one at 35 bp has a slow forward rate and causes pause b, while the one at 190 bp has a fast-forward rate and shows no backtracking pause in both experiment and model fitting.

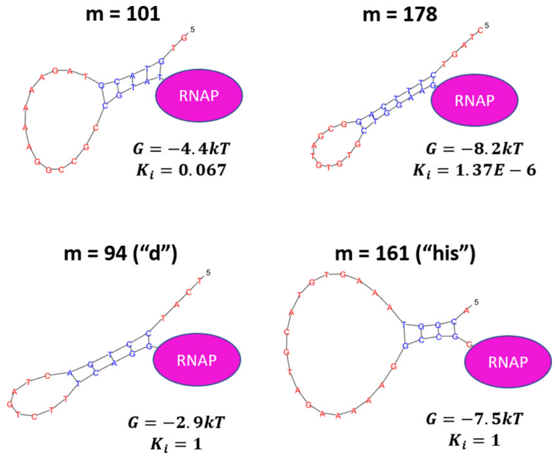

FIG. 10. Comparison of energy and $K_{i,h}$ at different positions. Hairpin formation is unfavorable at positions 101 and 178, although the hairpin structures at these positions are fairly stable, while at positions 94 and 161, the hairpin structures can readily form and induce the hairpin-stabilized pauses.

The model predicts that the effective length over which tension acts is about one-half of a base pair for the forward translocation pathway, but less than 0.1 base for the backtracking pathway. This result suggests that ordinary levels of force affecting translocation insignificantly affect the backtracking rate. During backtracking, RNAP must ratchet backward on the DNA and disrupt the RNA-DNA hybrid near the active site. We hypothesize that the rate is determined in large measure by the denaturation of the last formed base pair. Thus, external forces cannot alter this process as much as biasing the equilibrium constant in the forward translocation pathway.

The hairpin-stabilized pause requires the interaction between a transcript hairpin and the RNAP flap domain. Previous models simulated the folding of nascent transcripts using the lowest-energy method [10,13,21]. However, that method may not locate the correct positions of hairpins, since RNA folds cotranscriptionally and may not readily reach the lowest-energy configuration for RNAP at the pause site. In addition, simulation of co-transcriptional RNA folding requires enormous computational resources, so we devised a method which considers the stability difference between a structure including a hairpin and the lowest-energy structure to estimate the likelihood of hairpin formation. In this case, hairpins at position 101 and 178, although they are stable structures, are less likely to interact with RNAP than the less stable hairpins at positions 94 and 161, which correspond to pauses at sites d and *his*, respectively (Fig. 10). Our method also readily reproduced pause P2, which is significantly lengthened in the presence of RNase by favoring the proximal (3') hairpin at position 44 (P2) as illustrated in Fig. 5.

Some paused states are likely overlooked in the current model. For example, the current model cannot characterize the pauses observed at site c and at other less significant sites. The duration of the pause at site c is largely unaffected by the addition of either GreB or RNase, suggesting a mechanism distinct from backtracking or hairpin-stabilized pausing that is not captured in the current model. In a recent work by Janissen *et al.*, three interconnected paused states were extracted from long transcription assays, including an elemental paused state, a backtracked paused state, and a backtrack-stabilized state [22]. The backtrack-stabilized paused state is not included in our model, since the data for the tandemly repeated, 239 bp DNA sequence does not contain extremely long pauses (1̃00 s) that are classified as backtrack-stabilized states by Janissen *et al.*

B. Conclusion and outlook

This purely thermodynamic consideration of the transcription complex accurately reproduces transcription kinetics. By incorporating both class I and II pauses, the model refines our current understanding of active pathway and branched pathways in transcription and can be used to predict the occurrence of class I and II pauses that regulate transcription.

The model described herein significantly extends earlier efforts to model the kinetics of transcription. Bai *et al.*, Tadigotla *et al.*, and others independently proposed models in which the kinetic of transcription is treated as a competition between the active transcription pathway and a branched pathway [9,10,12,13,21]. Although their models yield results in statistical agreement with experimental results, the predicted pauses differ from those observed in single-molecule measurements, and the effects of tension and TFs were neglected. By fitting specific kinetic parameters under specific experiment conditions, our model achieves not only statistical agreement with experimental results but reveals quantitative details regarding the effects of DNA sequences, applied tension, and TFs.

Further improvements in our biochemical understanding of transcriptional pauses, in the quality of experimental data and in the model itself, could improve the predictive power. For example, the model might predict the pause at site c if the mechanism underlying this pause is determined and incorporated. Longer spans of high-resolution transcription data would also improve optimization of the model and the accuracy of predictions by providing more sequence variations.

IV. METHODS

A. Simulation of dwell time histogram using continuous-time Markov chain

With the transition rates between states, we can write the rate matrix of the Markov chain:

$$\mathbb{Q} = \begin{pmatrix} 1-\sum_n k_{\{1,n\}} & k_{0,\mathrm{bt}} & 0 & \dots & k_{\mathrm{hsp}} & k_{\mathrm{forward}} \\ k_{-1,\mathrm{btr}}+k_{\mathrm{cleavage}} & 1-\sum_n k_{\{2,n\}} & k_{-1,\mathrm{bt}} & \dots & 0 & 0 \\ k_{\mathrm{cleavage}} & k_{-2,\mathrm{btr}} & 1-\sum_n k_{\{3,n\}} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ k_{\mathrm{hspr}} & 0 & 0 & \dots & 1-\sum_n k_{\{12,n\}} & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}. \tag{18}$$

The elements of row n and column m represent the transition rate from state n to state m. The rows (columns) represent sequentially the active state, the backtracked states from 1–10 backtracking depth, the hairpin-stabilized state, and the next translocation state. The matrix is guaranteed nonsingular. Given an initial state, which is clearly $[1, 0, 0, \dots, 0, 0]$, the time spent in each state can be expressed as a matrix exponential

$$\begin{aligned} \overline{\tau} &= \alpha \exp{(\mathrm{t}\mathbb{Q})} \\ &= \alpha \sum_{n=0}^{\infty} \frac{t^n (VDV^{-1})^n}{n!} \\ &= \alpha V e^{Dt} V^{-1}, \end{aligned} \tag{19}$$

where α is the initial distribution of states, V consists of the eigenvectors of rate matrix $\mathbb{Q}$, and D is a diagonal matrix of the diagonal elements of eigenvalues of $\mathbb{Q}$ ordered like the eigenvectors in V. Thus, the expected time spent in each state is

$$\overline{\mu} = \alpha V \lambda V^{-1}, \tag{20}$$

where λ is the negative inverse of the diagonal element of D after replacing 0 eigenvalues with 1.

B. Optimizing the model with experimental data

To optimize the model parameters, we first considered only the forward translocation pathway and fitted the equilibrium parameters K_d, the kinetic parameters $k_{\max}$, and an effective length L_{forward} over which the external force acts during the forward translocation of RNAP. These parameters were optimized to generate a histogram that maximizes the goodness of fit (R), which is evaluated by the following equation:

$$R = 1 - \frac{\sum |O_i - X_i|}{\sum X_i}, \tag{21}$$

where O_i and X_i are the ith elements of the fitted and experimental histograms, respectively. In this step, parameters related to forward translocation are fitted to produce slow translocation rates at all experimentally detected pause sites. The result suggests that the pause at position "a" is a pre-translocated pause which is consistent with the experimental data showing insensitivity to GreB. However, pauses at other sites were characteristically longer than the dwell time produced by forward translocation only.

In the next step, we included the backtracked pathway, the energy barriers ΔG_{bt}, $\Delta G_{\mathrm{bt\_increment}}$, and an effective length L_{bt} for backtracking, maintaining the parameters of the forward translocation pathway set in the previous step (Table II). To reduce the complexity of the model, we fixed the prefactor k_1 as $1000s^{-1}$. The result suggests the backtracked pause is a large component of pauses at position b but not at other positions, in agreement with the analysis of backtracking dynamics.

Lastly, we included the hairpin-stabilized pause pathway with the rest of the parameters in the model fixed at the values identified in the preceding two steps (Table II). For any hairpins with stems or loops exceeding 3–8 or 4–20 bases, respectively, ΔG_{stem} and/or ΔG_{loop} were set to infinity as they were unlikely to stabilize pauses. The result gives good agreement with pauses at other sites.

We repeated the procedure above to fit the experimental data with GreB and RNase. The goodness of fit is evaluated for major pause sites (dwell time $>$ 0.05 s) and overall histogram separately using Eq. (21). Overall, the model faithfully reproduces pause times at all pause sites under all conditions with the exception of pause c. Pauses at c might originate from a different mechanism. Table II gives a list of the values of the fitted model parameters.

We applied the tuned model from previous steps on a new sequence. Since this new sequence precedes the repeat region in the transcription experiment, we have fewer experimental data on this sequence and the aligned histogram shows more minor peaks than the histogram of the well-aligned repeated region. Nonetheless, the tuned model successfully reproduced the major pauses, as shown in Fig. 3(c), and the goodness of fit on this unfamiliar sequence indicates the tuned model is not an overfit.

C. Monte Carlo simulation and analysis on backtracking dynamics

To further validate the model, we generated transcription data using Monte Carlo simulation and compared the dynamics of backtracking in the experimental and simulated traces. Figure 6(a) shows the example traces generated by using the optimized parameters shown in Table II under 10 pN of assisting tension. The simulation was performed by calculating the probability of state transitions from the transition rates every 0.001 s. We collected the backtrack depth and duration from the simulated traces at pause site b and compared them to the experimental results. Figures 6(b) and 6(c) show that the simulated backtracking depth and duration were similar to those observed experimentally.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIH) Grant No. R01 GM084070 to L.F. We are grateful to Carlos Bustamante and Alex Tong for generously providing high-resolution transcription data.

[1] K. M. Herbert, A. Porta, B. J. Wong, R. A. Mooney, K. C. Neuman, R. Landick, and S. M. Block, Sequence-resolved detection of pausing by single RNA polymerase molecules, Cell **125**, 1083 (2006).

[2] E. A. Abbondanzieri, W. J. Greenleaf, J. W. Shaevitz, R. Landick, and S. M. Block, Direct observation of base-pair stepping by RNA polymerase, Nature **438**, 460 (2005).

[3] I. Artsimovitch and R. Landick, Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals, Proc. Natl. Acad. Sci. U.S.A. **97**, 7090 (2000).

[4] I. Toulokhonov, I. Artsimovitch, and R. Landick, Allosteric control of RNA polymerase by a site that contacts nascent RNA hairpins, Science **292**, 730 (2001).

[5] J. Y. Kang, T. V. Mishanina, M. J. Bellecourt, R. A. Mooney, S. A. Darst, and R. Landick, RNA polymerase accommodates a pause RNA hairpin by global conformational rearrangements that prolong pausing, Mol. Cell **69**, 802 (2018).

[6] M. M. Abdelkareem, C. Saint-André, M. Takacs, G. Papai, C. Crucifix, X. Guo, J. Ortiz, and A. Weixlbaumer, Structural basis of transcription: RNA polymerase backtracking and its reactivation, Mol. Cell **75**, 298 (2019).

[7] M. H. Larson, R. A. Mooney, J. M. Peters, T. Windgassen, D. Nayak, C. A. Gross, S. M. Block, W. J. Greenleaf, R. Landick, and J. S. Weissman, A pause sequence enriched at translation start sites drives transcription dynamics in vivo, Science **344**, 1042 (2014).

[8] K. M. Herbert, J. Zhou, R. A. Mooney, A. L. Porta, R. Landick, and S. M. Block, E. coli NusG inhibits backtracking and accelerates pause-free transcription by promoting forward translocation of RNA polymerase, J. Mol. Biol. **399**, 17 (2010).

[9] L. Bai, A. Shundrovsky, and M. D. Wang, Sequence-dependent kinetic model for transcription elongation by RNA polymerase, J. Mol. Biol. **344**, 335 (2004).

[10] V. R. Tadigotla, D. O. Maoiléidigh, A. M. Sengupta, V. Epshtein, R. H. Ebright, E. Nudler, and A. E. Ruckenstein, Thermodynamic and kinetic modeling of transcriptional pausing, Proc. Natl. Acad. Sci. U.S.A. **103**, 4439 (2006).

[11] T. D. Yager and P. H. Von Hippel, A thermodynamic analysis of RNA transcript elongation and termination in Escherichia coli, Biochemistry **30**, 1097 (1991).

[12] L. Bai, R. M. Fulbright, and M. D. Wang, Mechanochemical kinetics of transcription elongation, Phys. Rev. Lett. **98**, 068103 (2007).

[13] D. O. Maoiléidigh, V. R. Tadigotla, E. Nudler, and A. E. Ruckenstein, A unified model of transcription elongation: What have we learned from single-molecule experiments? Biophys. J. **100**, 1157 (2011).

[14] B. E. Nickels and A. Hochschild, Regulation of RNA Polymerase through the secondary channel, Cell **118**, 281 (2004).

[15] M. H. Larson, W. J. Greenleaf, R. Landick, and S. M. Block, Applied force reveals mechanistic and energetic details of transcription termination, Cell **132**, 971 (2008).

[16] A. Chauvier, J. F. Nadon, J. P. Grondin, A. M. Lamontagne, and D. A. Lafontaine, Role of a hairpin-stabilized pause in the Escherichia coli thiC riboswitch function, RNA Biology **16**, 1066 (2019).

[17] I. Tinoco Jr. and C. Bustamante, The effect of force on thermodynamics and kinetics of single molecule reactions, Biophys. Chem. **101-102**, 513 (2002).

[18] R. Gabizon, A. Lee, H. Vahedian-Movahed, R. H. Ebright, and C. J. Bustamante, Pause sequences facilitate entry into long-lived paused states by reducing RNA polymerase transcription rates, Nat. Commun. **9**, 2930 (2018).

[19] K. C. Neuman, E. A. Abbondanzieri, R. Landick, J. Gelles, and S. M. Block, Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking, Cell **115**, 437 (2003).

[20] I. Toulokhonov, J. Zhang, M. Palangat, and R. Landick, A central role of the rna polymerase trigger loop in active-site rearrangement during transcriptional pausing, Mol. Cell **27**, 406 (2007).

[21] A. Bochkareva, Y. Yuzenkova, V. R. Tadigotla, and N. Zenkin, Factor-independent transcription pausing caused by recognition of the RNA-DNA hybrid sequence, EMBO J. **31**, 630 (2012).

[22] R. Janissen, B. Eslami-Mossallam, I. Artsimovitch, M. Depken, and N. H. Dekker, High-throughput single-molecule experiments reveal heterogeneity, state switching, and three interconnected pause states in transcription, Cell Reports **39**, 110749 (2022).