Characterizing highly dynamic conformational states: the transcription bubble in RNAP-promoter open complex as an example

Eitan Lerner[1,a)], Shimon Weiss[1,2,3,a)]

[1] Department of Chemistry & Biochemistry , University of California , Los Angeles , CA , USA.
[2] Molecular Biology Institute , University of California , Los Angeles , CA , USA.
[3] Department of Physiology , University of California , Los Angeles , CA , USA.

[a)] Electronic mail: eitanlerner1@g.ucla.edu; sweiss@chem.ucla.edu

ABSTRACT

Bio-macromolecules carry out complicated functions through structural changes. To understand their mechanism of action, the structure of each step has to be characterized. While classical structural biology techniques allow the characterization of a few 'structural snapshots' along the enzymatic cycle (usually of stable conformations), they do not cover all (and often fast interconverting) structures in the ensemble, where each may play an important functional role. Recently, several groups have demonstrated that structures of different conformations in solution could be solved by measuring multiple distances between different pairs of residues using single-molecule Förster resonance energy transfer (smFRET) and using them as constrains for hybrid/integrative structural modeling. However, this approach is limited in cases where the conformational dynamics is faster than the technique's temporal resolution. In this study, we combine existing tools that elucidate sub-millisecond conformational dynamics together with hybrid/integrative structural modeling to study the conformational states of the transcription bubble in the bacterial RNA polymerase (RNAP)-promoter open complex (RPo). We measured microsecond alternating laser excitation (µsALEX)-smFRET of differently labeled lacCONS promoter dsDNA constructs. We used a combination of burst variance analysis (BVA), photon-by-photon hidden Markov modelling (H2MM) and the FRET-restrained positioning and screening (FPS) approach to identify two conformational states for RPo. The experimentally-derived distances of one conformational state match the known crystal structure of bacterial RPo. The experimentally-derived distances of the other conformational state have characteristics of a scrunched RPo. These findings support the hypothesis that sub-millisecond dynamics in the transcription bubble are responsible for transcription start site selection.

I. INTRODUCTION

Biological macromolecules function through changes in their structures. To elucidate their mechanism of action, the structures at all functional step need to be determined. Atomistic structures are usually solved using X-ray crystallography[2, 3] or by nuclear magnetic resonance (NMR) spectroscopy[2]. Single-particle cryo-electron microscopy (cryo-EM) has recently been added to this toolkit for characterization and determination of multiple conformations of macromolecules in the ensemble with spatial resolution approaching the atomic level[4, 5]. Molecular mechanisms could then be inferred from structural "snapshots" attained by these methods (e.g. [6]). However, although structural snapshots can identify different conformational states (e.g. ligand-bound or unbound, folded or unfolded), they may represent just a subset of all the conformations important for biological function and also lack information on the transitions (and their time scales) between all conformational states.

Biophysical techniques such as NMR, electron paramagnetic resonance (EPR)[7] and double electron-electron resonance (DEER)[8] spectroscopies as well as fluorescence-based techniques such as fluorescence polarization[9], ensemble Förster resonance energy transfer (FRET) [10], or electron transfer[11], can provide dynamic information on macromolecular conformations and the dynamics associated with the transitions between them. However, interpreting experimental results derived from these techniques is highly model-dependent and yields limited insight on structural and mechanistic details [12, 13]. This is mainly because conformational changes in an ensemble are not synchronized, which yields averaged-out signals.

The lack of synchronicity is removed if one molecule is examined at each given time, as afforded by the single-molecule (sm) FRET (smFRET) technique[14, 15] which allows the retrieval of underlying heterogeneity and conformational dynamics. In smFRET, the efficiency of energy transfer from a donor dye to an acceptor dye is measured for each molecule, reporting on the distance between the dyes. Upon attachment of the dyes to different surface residues of the macromolecule, smFRET can report on conformational changes through a distance reaction coordinate between the dyes' attachment points. However, each pair of dyes accounts for only ***one*** distance between the attachement points (on the surface of the three-dimensional, 3D, macromolecular structure). In order to capture the full 3D structure and its conformations, multiple such one-dimensional (1D) projections need to be recorded. Indeed, several groups used this approach in conjunction with 3D structures attained from molecular simualtions (e.g. molecular dynamics, Monte Carlo)[16-21]. This approach allows identification and solving of 3D macromolecular structures that are 'too dynamic' and have not been characterized by classical structural biology techniques.

If transition rates between different conformatios are slower than the smFRET temporal resolution, they could be resolved. If, however, the conformational states are separated by a low activation barrier, yielding transition rates that are faster than the method's temporal resolution, they would be averaged-out and become indistinguishable. In smFRET of freely diffusing molecules, fluorescence bursts (of donor and acceptor photons) are generated when the molecules traverse the observation volume. These bursts typically last a few milliseconds. FRET values are calculated for each burst (molecule) with this temporal resolution. Single bursts (molecules) with different FRET efficiencies are then grouped into different subpopulations of distinct conformational states that

interconvert at time scales slower than ~10 ms. Faster conformational transitions (occuring at timescales of ~ 0.1-10 ms) are averaged out[22, 23]. Several techniques based on analysis of photon statistics were developed to study conformational dynamics occuring within this faster temporal range[24-26]. Of particular interest to the work presented here, we note the burst variance analysis (BVA) techique which is based on the analysis of the variance of FRET efficiency in within bursts[26], as well as the photon-by-photon hidden Markov modelling approach (H2MM)[27] that quantifies conformational dynamics parameters (number of conformational states involved, their FRET efficiencies, and their interconversion rate, down to ~1-10 μs[27]). To characterize the structures of different conformational states in the ensemble, it is therefore important to first identify each conformational state with the highest available temporal resolution, and to do so for multiple differently-labeled constructs.

In this study, we apply this approach to study the conformational dynamics of the transcription bubble in RNAP-promoter open complex (RPo). To initiate DNA transcription starting from a gene's promoter sequence, RNAP has to first specifically bind to the promoter and then open a transcription bubble by melting a segment of 10-12 base pairs (bp) to form the RPo[28, 29]. While doing so, RNAP positions a pair of bases from the downstream fork of the bubble in front of its active site to dictate the initial sequence of the RNA to be transcribed[30]. Therefore, the size of the transcription bubble imposes which pair of DNA bases will be in front of the active site, hence dictating the (canonical) transcription start site (TSS). We note, however, that in the presence of specific di-nucleotides, TSS can be re-programed[31-33]. Recently, Robb *et al.* have identified bubble size fluctuations in RPo and suggested that they are involved in TSS selection[34]. They performed smFRET measurements on doubly-labeled promoter dsDNA constructs sensitive of changes in the bubble. BVA analysis has indicated sub-millisecond bubble conformational dymanics, implying that DNA bases situated downstream to the transcription bubble may dynamically melt and transiently be reeled into the active site. Such hypothesized melting and reeling-in mechanism could increase the overall bubble size and is reminescent of 'DNA scrunching'[35, 36], but does not require the presence of NTPs. If validated, this mechanism could explain how transcription initiates from positions other than the canonical TSS along the promoter sequence.

Robb *et al.* have also presented results of a gel-based transcription assay for the lacCONS promoter[37], exhibiting two discrete bands (i.e. transcripts of two different distinct lengths). If these bands correlate with the transcription bubble dynamics, they may imply the existence of two conformational states, one of which is RPo, and the other is a modified RPo having a few additional DNA bases scrunched into the active site. To prove or refute this hypothesis, it is important to quantify the transcription bubble dynamics using a conformational state model and to do so for more than two positions of dyes.

We therefore performed diffusion-based smFRET measurements on a series of such donor-acceptor labeled lacCONS promoter constructs and characterized the donor-acceptor distances for each construct at each conformational state, as well as the interconversion rate constants, by employing the H2MM approach. We then used the attained distances as constraints to test which of the two identified conformational states matches the known RPo crystal structure[1] and what is the conformation of the other state. Our results support the two-state hypothesis for TSS selection in the lacCONS promoter.

II. EXPERIMENTAL

A. µsALEX-smFRET distance measurements

We performed microsecond alternating laser excitation (µsALEX) smFRET measurements[38-40] on a library of dsDNA constructs having the lacCONS promoter sequence[37]. All of the constructs had the template strand (T) labeled with the acceptor (A) ATTO647N attached to a specific DNA base (denoted here as TA) and the nontemplate strand (NT) labeled with the donor (D) Cy3B at another specific DNA base (denoted here as NTD). The identity of the labeled DNA bases is denoted relative to the promoter canonical TSS, where negative or positive values define regions that are upstream or downstream of TSS respectively. Throughout this work, we therefore use the following nomenclature to name a D-A labeled lacCONS construct: (-3)TA-(-8)NTD is one construct in the library with an A dye labeling the base at register (-3) upstream to the canonical TSS in the template strand and with a D dye labeling the base at register (-8) upstream to the canonical TSS in the nontemplate strand. Table S1 of the supplementary materials lists all constructs used for this study.

Measurements of the labeled dsDNA constructs were taken in the absence and in the presence of *Escherichia coli* (*E. coli*) RNAP in order to track conformational changes in the promoter induced by the formation of RPo. As a control, measurements were also taken in the presence of RNAP after incubation with nucleotides (NTPs) to assess promoter escape activity (Fig. S3 of the supplementary materials). All measurements were taken at T=25^{0}C, for ~30 minutes, and with the labeled dsDNA at a concentration of ≤100 pM. Additional experimental details can be found elsewhere[41, 42].

D-A distances ($\bar{r}_{D-A}$) for all constructs in RPo (and for free promoter as controls) were extracted according to the steps outlined in supplementary materials.

B. FRET-restrained positioning and screening

In this study, we measured multiple distances for dsDNA promoter in the presence of the bacterial RNAP in the RPo state, for which a crystal structure already exists (pdb code: 4XLN)[1]. After calculating $\bar{r}_{D-A}$ values for each conformational state (as identified by H2MM) for all D-A labeled constructs, distances were classified into different groups, where each group presumably defines one conformational state. As can be seen in (Fig. 4 and Table S3 of the supplementary materials) two groups were identified as RNAP-bound conformations. Since our results indicate that the transcription bubble has two conformations, we wanted to identify which of the smFRET-derived conformational states represents the solved structure (as represented in pdb code: 4XLN[1]) and to what extent and how the other conformation deviates from it.

As already demonstrated, a group of $\bar{r}_{D-A}$ distances can be used as constraints for structural determination by computation methods such as molecular dynamics (MD)[19, 43-48], Monte Carlo (MC)[20, 49] and coarse-grained (CG)[50-52] simulations as well as using structural databases[53] or distance-constrained triangulation[54]. However, since $\bar{r}_{D-A}$ do not directly report on distances between ***DNA bases*** but rather distances between the D and A ***fluorophore centers***, an additional computational step that takes into account the dye linker length and all possible dye configurations in space (dye accessible volume, or AV) is needed for this 'distance translation'. Such a tool, dubbed "the FRET-restrained

positioning and screening (FPS) approach" was developed by Kalinin et al.[19] and adapted for this study.

The FPS software calculates the AV of a dye given the atom to which it is attached to in a given pdb structure, the simplified geometrical (ellipsoid) parameters of the dye, and its linker length L_{link} and width w_{link}. In our case, L_{link} = 20 Å and w_{link} = 4.5 Å for Cy3B while L_{link} = 20.5 Å and w_{link} = 4.5 Å for ATTO647N. The ellipsoid parameters for Cy3B are {R_1 = 6.8 Å ; R_2 = 3.0 Å ; R_3 = 1.5 Å} and for ATTO647N {R_1 = 7.15 Å ; R_2 = 4.5 Å ; R_3 = 1.5 Å}. These parameters are taken from the table provided in the FPS software manual (for Cy3B, we used the Cy3 values).

The AV surface provides a set of possible positions for the fluorophore to assume in space. Each pair of such positions for D and A, makes up one possible $r_{D\text{-}A}$. The expected value $\bar{r}_{D-A}$ (and the expected standard deviation, $SD(\bar{r}_{D-A})$) are calculated by averaging over all possible pairs of dye positions. These averages are then compared to expected distances from the known structure.

We compared the experimentally-derived set of $\bar{r}_{D-A}$ values for each conformational state with the known structure of bacterial RPo[1]. The crystal structure includes the promoter T and NT strands and all the subunits that make up the RNAP holoenzyme. We separated the crystal structure into T, NT and RNAP holoenzyme, as three distinct entities. Then, using the FPS software, we tried to reconstruct the complex using the experimentally-derived distance constraints while minimizing the RMSD values between all experimental and expected D-A distances.

III. RESULTS AND DISCUSSION

A. The screening process

We measured a library of 43 dsDNA constructs doubly-labeled (D-A) at different positions of the lacCONS promoter[37]. We used all 43 constructs for µsALEX-smFRET measurements of free promoter, calibrating the correction factors *lk*, *dir* and γ (see supplementary materials – µsALEX-smFRET). However, out of these combinations, only 23 constructs had D and A dyes' positions that are expected (based on the crystal structure) to exhibit different $r_{D\text{-}A}$ in RPo as compared to the free DNA (colored rows in Table S1 of the supplementary materials). The constructs included D-A pairs positioned along two different reaction coordinates: (i) ***scrunching coordinate***, where one dye is attached to a DNA base (T or NT) upstream of the transcription bubble or at the edge of the bubble (promoter registers ≤ -8) and the other is attached to a DNA base (T or NT) downstream of the

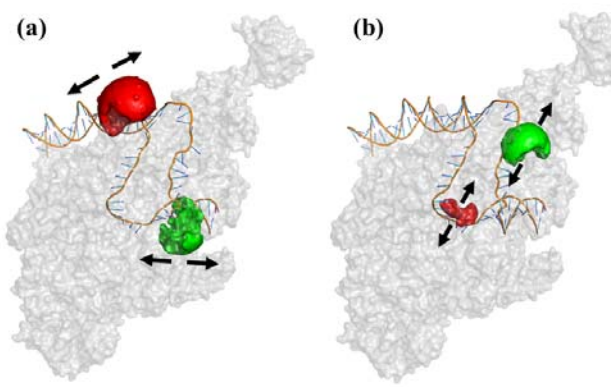


Figure 1: D-A labeled promoter constructs representing two reaction coordinates. Examples of dyes' attachment points in RPo for two reaction coordinates (pdb code: 4XLN[1]). (a) scrunching coordinates: dyes attached to bases upstream and downstream relative to the transcription bubble; (b) bubble coordinates: dyes attached to bases within the transcription bubble. The arrows show the general directions in which distance changes are expected in the two reaction coordinates. Dye AVs are also shown as green (D) and red (A) partial spheres.

transcription bubble or at the edge of the bubble (promoter registers ≥ -2; Fig. 1a and Table S1 of the supplementary materials); (ii) ***bubble coordinate***, where both dyes are attached to T and NT on registers inside the transcription bubble in RPo (promoter registers -10 to +1; Fig. 1b and Table S1 of the supplementary materials); Bubble coordinate constructs in RPo are expected to show an increase in $r_{D\text{-}A}$ due to bubble opening, while scrunching coordinate constructs in RPo are expected to show a decrease in $r_{D\text{-}A}$ due to reeling of DNA into the enzyme (as compared to free DNA). However, in some cases, degeneracies could be expected where $r_{D\text{-}A}$ values in free promoter and in RPo are similar (Fig. 4, • vs. •).

These constructs are expected to be sensitive to dynamical distance changes due to possible fluctuations in the bubble size or fluctuations in scrunching (Fig. 1). We note, however, that some of the constructs could be insensitive to these fluctuations (due to FRET limited dynamic range).

We further screened the 23 constructs according to the following criteria: (i) the *PR* histogram of free promoter yields a single population (except for cases in which quenching of FRET occurs[55]; *PR* histograms of free promoter in Figs. 2a and S2); (ii) <u>bubble opening activity</u>: *PR* histograms of RPo yield two subpopulations, one similar to that of the free promoter and the other representing the RNAP-bound fraction, with a peak *PR* value different than that of the free promoter subpopulation (comparing *PR* histograms of free promoter and RPo in Figs. 2, S2 and S3); (iii) qualitative test for conformational dynamics: a majority of sm bursts in the *PR* subpopulation representing the RNAP-bound fraction exhibit *σ(PR)* values larger than expected from shot noise (burst variance analysis, BVA, plots of RPo in Figs. 2 and S2; more information on BVA in supplementary materials); (iv) promoter escape activity: the RPo *PR* subpopulation is significantly decreased after addition of NTPs, and the resulting distribution resembles the *PR* of the free promoter (comparing PR histograms of RPo and RPo + NTPs in Fig. S3 of the supplementary materials); (v) the experimental results are converted to $r_{D\text{-}A}$ values (using Eqs. S16 and S17), and collectively compared against RPo crystal structure

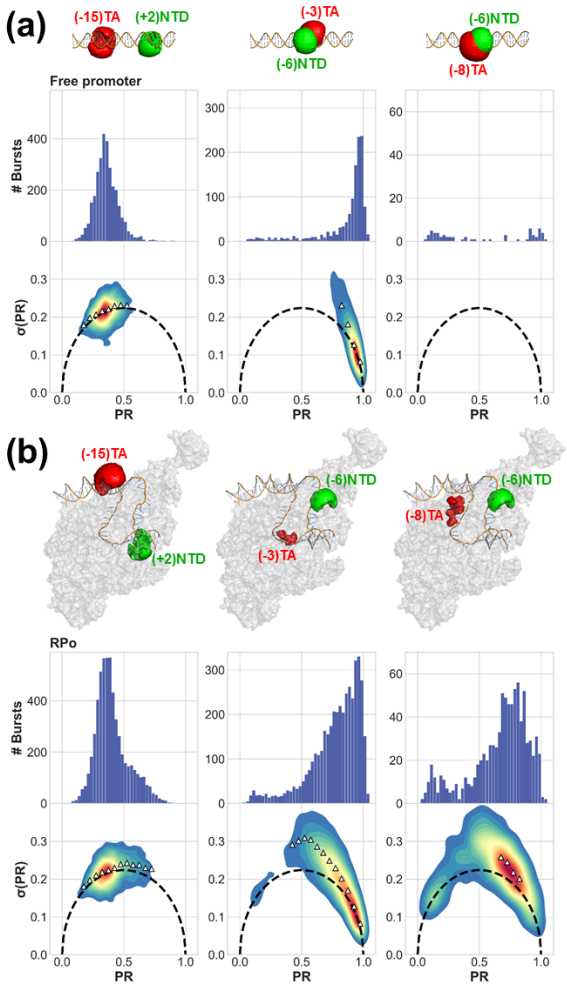


Figure 2: Qualitative assessment of smFRET dynamics using BVA. Proximity ratio (*PR*) histograms and BVA plots of *σ(PR)* versus *PR* show that: (a) free promoter is either characterized by a single *PR* population with no smFRET dynamics or with no *PR* population at all, due to quenching of FRET, and that (b) RNAP bound to promoter DNA induces a unique *PR* subpopulation that exhibits smFRET dynamics. Depiction of differently labeled D-A constructs in free promoter (a) and RPo (b) states, were generated by calculating D and A AVs (green and red surfaces, respectively) on top of DNA labeling positions in the context of RPo crystal structure (pdb code: 4XLN[1]).

(pdb code: 4XLN)[1]. This structure includes part of the promoter with registers -35 through +12, hence scrunching coordinate constructs with dyes at registers outside this range ((+15)TA-(-15)NTD and (+15)TA-(-8)NTD in our case) are excluded. 9 constructs out of 23 scrunching/bubble coordinate constructs were selected following these screening steps (yellow-shaded rows, Table S1 of the supplementary materials). Out of the 9, 6 were bubble coordinate constructs and 3 were scrunching coordinate constructs (Figs. 2, S2, S3, and yellow-shaded rows of Table S1 of the supplementary materials).

In summary: the final 9 selected constructs exhibited a unique subpopulation in the *PR* histograms that represents the RNAP-bound fraction. This subpopulation exhibited dynamics (as assessed by BVA) while the free promoter fraction did not exhibit such dynamics (Figs. 2 and S2). We therefore conclude that these bursts carry information on fast interconverting conformational states and that the number of different conformational states cannot be extracted from simple fits of *PR* histograms to a sum of Gaussians. BVA allowed us to identify the presence of fast dynamics; the next section shows how to extract quantitative dynamics parameters.

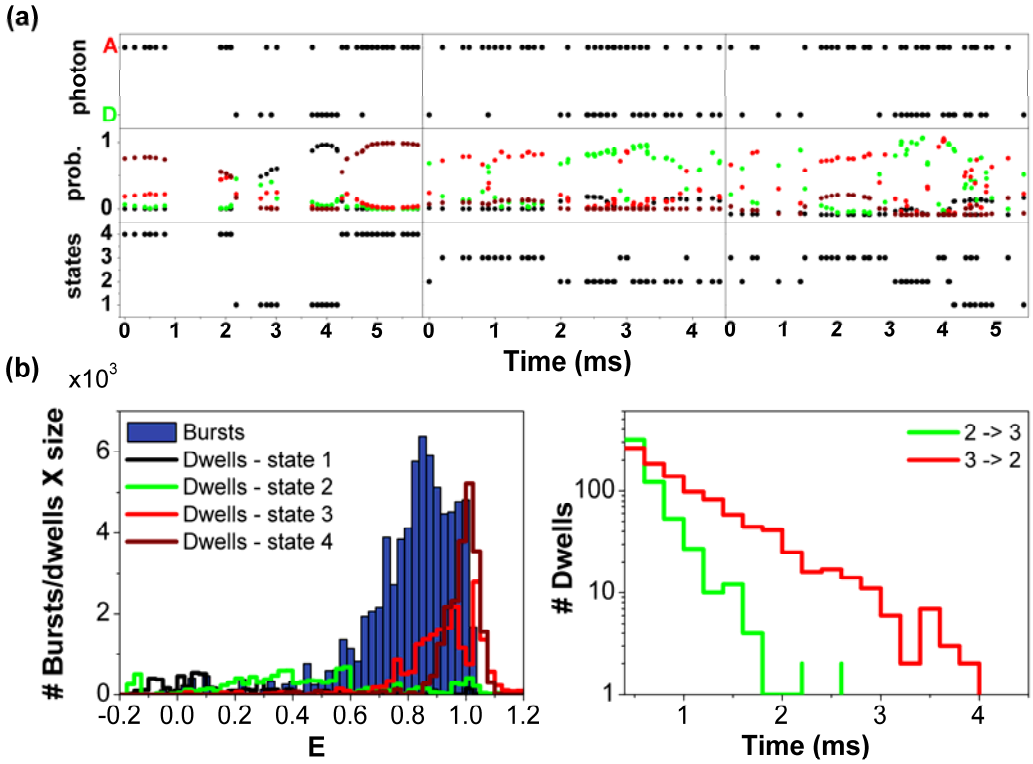


Figure 3: Results of H2MM analysis. H2MM analysis results for the construct (-3)TA-(-6)NTD in RPo. The analysis assumed the presence of 4-states and yielded the most likely *PR* values and state interconversion rate constants. Applying the Viterbi algorithm on H2MM results yields the state posterior probability for each photon in each burst. This in turn is translated into dwells of consecutive photons in each state. (a) analysis results for three representing bursts. After dwells of a given state are defined, dwell sizes dwell FRET efficiencies (E), and dwell times are extracted and plotted in histograms. (b) Dwell-derived *E* histograms for the 4 states (lines) and the burst-derived *E* histogram (filled blue bins). The histograms are weighted by the dwell/burst sizes. (c) Dwell time histograms for transitions between the two conformational states associated with the RNAP-bound fraction.

B. Quantification of underlying dynamics using H2MM

Photon-by-photon hidden Markov modelling (H2MM) analysis was applied to smFRET photons bursts of all 9 selected constructs in RPo (for more information on H2MM, see supplementary materials). Fig. 3 shows an example of analysis results for (-3)TA-(-6)NTD (bubble coordinate). In this construct, the free promoter fraction was not quenched and hence showed up as a subpopulation with high *PR* values in the *PR* histograms (Fig. 2a, center column). As expected, this peak did not exhibit dynamics in the BVA plots, and therefore was assigned a single conformational state. Upon addition of RNAP, a subpopulation with lower *PR* values emerged (RPo). This subpopulation exhibited dynamics as determined by BVA (Figure 2b, center column). It was therefore assumed to exhibit at least two conformational states (interconverting with rate constantsin the range ~10^4-10^2 s^{-1}). The model, therefore, should have at least three states (Min. *q*, Table S2 of the supplementary materials), with one free promoter state and two (or more) RNAP-bound conformational states.

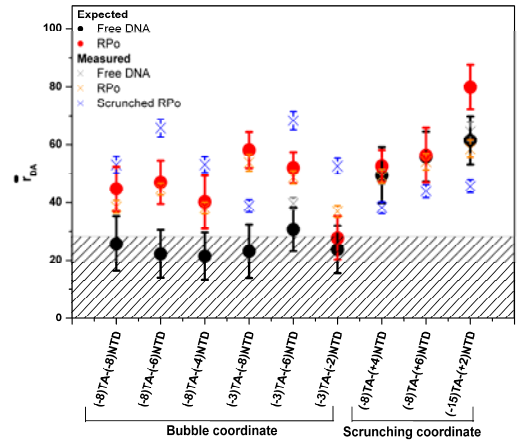


Figure 4: Comparison of measured D-A distances with expected distances. The experimentally-derived D-A distances of different conformational states for each D-A labeled lacCONS constructs are compared with the expected D-A distances from RPo crystal structure (pdb code: 4XLN[1]) in free DNA (•) and in RPo (•). Two conformational states were found (with sub-millisecond underlying interconversion dynamics) in the RNAP-bound fraction. Their derived distance values are shown (X) for the subset that defines a conformation closest to the known RPo structure. Distances for the other conformation are also shown (X). The values of the free promoter fraction are also presented (X; for the D-A labeled promoter constructs that do not exhibit quenched FRET). Quenched FRET constructs had an expected D-A distance below $0.5R_0$ (shown as shaded area).

Each photon in each burst was assigned with a state posterior probability for a given state in the model, out of which state dwells were calculated (Fig. 3a). Then, dwells-derived $\bar{E}$ and t_{dwell} were calculated and plotted as histograms (Fig. 3, b and c) and dwells-derived $\bar{E}$ histogram was compared to burst-derived $\bar{E}$ histogram. This comparison (Figure 3b) demonstrates how bursts average-out the underlying dynamics.

Using H2MM, dwells were calculated for each state (using 3 or 4 states model). Next, the state mean FRET efficiency $\bar{E}_w$ was calculated according to Eqs. S12 and S13 and the state mean interconversion rate constant, $\bar{k}$, was calculated according to Eqs. S14 and S15. Since this construct represents the bubble coordinate, we expected the $\bar{E}_w$ values of the RNAP-bound states to be smaller or equal relative to the value of the free promoter. Repeating this analysis for 3- and 4-state models yielded similar results, hence both 3 states and 4 states models could describe the underlying dynamics for this construct. Using the *BIC'* criterion (Eq. S11) requiring *BIC'<0.005*, the 4 states model was found to be a better model to describe the underlying conformational dynamics in the construct (-3)TA-(-6)NTD (Fig. S4 and Table S2 of the supplementary materials). 7 constructs were best described by a 3 states model and 2 constructs were best described by a 4 states model. Results for all constructs are summarized in Table S2 of the supplementary materials. The next section outlines an approach for classifying these results into distinct groups of $\bar{r}_{D-A}$ values that represent different conformations.

C. Classification of D-A distances to conformational states of the transcription bubble

The results of H2MM analyses (states and rate constants derived from dwells) indicate that: (i) the *PR* subpopulation of RPo (Figs. 2b and S2, and Table S2 of the supplementary materials) is an average of two dynamically interconverting states; (ii) in all constructs, except for (-15)TA-(+2)NTD, one out of the two conformational states, associated with the RNAP-bound subpopulation, had $\bar{r}_{D-A}$ values that coincided with the values expected from the crystal structure of RPo (Fig 4, X vs. •). We therefore give this state the name *RPo state*; (iii) in all bubble coordinate constructs, except for (-3)TA-(-8)NTD, the second state had larger $\bar{r}_{D-A}$ values as compared to *RPo state* $\bar{r}_{D-A}$ values (Fig. 4, X vs. X); (iv) in all scrunching coordinate constructs, the second state had smaller $\bar{r}_{D-A}$ values as compared to *RPo state* $\bar{r}_{D-A}$ values (Fig. 4, X vs. X). When taken together (except for (-3)TA-(-8)NTD), these results suggest that the second state resembles DNA scrunching: distances of scrunching coordinates decrease while distances of bubble coordinates increase (as compared to the RPo state, Fig. 4, X). We therefore name this state *scrunched RPo state*. The bubble-coordinate construct (-3)TA-(-8)NTD exhibited a decrease in distance relative to RPo state. This exception may hint to the fine details of topological changes inside (or outside of) the transcription bubble in the *scrunched RPo state*.

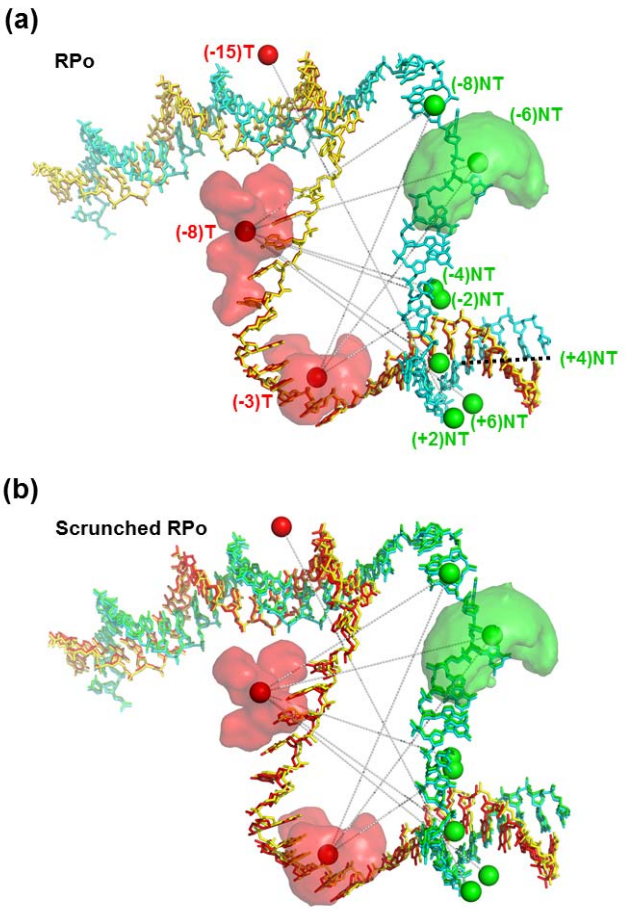


Figure 5: FPS analysis of transcription bubble conformational states. FPS analysis was performed for each conformational state against the known RPo structure (pdb code: 4XLN). (a) One set of D-A distances (*RPo state*) permits reconstruction of the template (T; red) and nontemplate (NT; green) strands in a 3D organization that matches the 3D organization of the T (yellow) and NT (cyan) strands in the crystal structure. (b), The second set of D-A distances (*scrunched RPo state*) yielded reconstruction of the template (T; red) and nontemplate (NT; green) strands in a 3D organization that does not match the 3D organization of the T (yellow) and NT (cyan) strands in the crystal structure. Solid spheres represent the mean positions of the D (green) and A (red) dyes. The set of D-A distances probed is shown by grey dashed lines connecting D and A mean positions. The mean position is always in within the dyes' AVs (3 examples of dye AVs are shown).

Next, the two sets of $\bar{r}_{D-A}$ values that were assigned to the *RPo state* and to the *scrunched RPo state* were used as distance constraints in FRET-restrained positioning and screening (FPS)[19], as described below.

D. Assessment of FRET-derived conformational states against a known RPo structure

In principle, the FRET-derived $\bar{r}_{D-A}$ values that were assigned to the *RPo state* should agree with the crystal structure of bacterial RPo[1]. Indeed, and as mentioned above, this assertion seems to be true except for (-15)TA-(+2)NTD (Fig 4, X vs. •). A more quantitative assessment was performed using the FPS approach (supplementary materials) to verify which of the conformational states' distance sets (*RPo state* and the *scrunched RPo state*) better agree with the known RPo crystal structure. FPS fitting results demonstrated that distances of the *RPo state* afforded reconstruction of the promoter DNA conformation that coincided with the known crystal structure with only minimal deviations (RMSD=5 Å; Fig. 5a), hence justifying the assignment. FPS fitting results demonstrated that distances of the *scrunched RPo state* afforded reconstruction of the promoter DNA conformation that had large deviations from the known crystal structure (RMSD=15 Å; Fig. 5b).

IV. Conclusions

A. Transcription bubble dynamics and transcription start-site selection

Robb *et al.* have previously identified sub-millisecond transcription bubble dynamics using BVA[34]. In conjunction with their smFRET studies, they also performed radioactive gel-based transcription assays and have shown that transcription with the lacCONS promoter produced two bands of different-length transcripts. The longer transcript corresponded to the expected length assuming that transcription started from the canonical TSS. The ~ 2-3 bp shorter transcript suggested that bubble dynamics occurred between RPo and another conformation in which ~2-3 downstream DNA bp were scrunched, leading to a selection of a different transcription start site. We note, however, that transcription bubble dynamics and TSS selection could, in principle, occur independently. Our study, however, further corroborate Robb *et al.* conclusion: we showed that (i) the sub-millisecond dynamics is between two conformational states for the lacCONS promoter; (ii) one conformational state matches the known RPo crystal structure; (iii) the other conformational state is suggestive of DNA scrunching permitting the start of transcription at an alternative TSS, 2-3 bp downstream to the canonical TSS.

It is still unclear, however, whether the correlated increase in bubble coordinates and decrease in scrunching coordinates can occur without the melting of downstream DNA bases. Robb *et al.*'s study was followed by several studies that examined promoter sequence determinants for TSS selection[56-60]. These groups found specific promoter sequence characteristics that affect TSS selection. The mechanism suggested by Robb *et al.* and supported by our study may explain why TSS selection is a widespread phenomenon, with different types of scrunched RPo states, corresponding to different types of TSSs in different promoter sequences. Scrunching dynamics supporting TSS selection may depend also on the efficiency of the incorporation of the first NTP, from each RPo state. Our studies identified two conformational states, the *RPo state* and *scrunched RPo state* that are stable/survive for a few hundreds of microseconds. It is

unclear, however, how efficient first NTP incorporation will be for short-lived conformations. To ultimately answer the question, an experiment that can correlate between TSS selection and bubble dynamics in real time is required.

B. Determination of dynamic structures

In smFRET of immobilized molecules, state dwells, their FRET values and their transition rates could be directly retrieved from time trajectories but limited to rate constants <100 s^{-1} [61, 62]. smFRET burst analysis of diffusing molecules can report on transition rate constants that are slower than the transit time through the observation volume (~1 ms), also effectively limited to ~100 s^{-1} [22, 23] (Recent works have shown that faster dynamics, associated with rate constants up to ~5-10x10^3 s^{-1}, could be assessed using advanced techniques[23, 25, 26]. Nevertheless, since bright dyes can produce photon detection rates of ~1 MHz, photon-by-photon analysis could provide information on faster dynamics[27]. We note that interconversion between conformational states (separated by an activation barrier larger than k_BT) could occur with even faster rate constants (~100x10^6 s^{-1}), but the amplitude of motion will be small at these very fast rates.

In this study we presented a strategy for the extraction of FRET-derived conformational states interconverting with rate constants of up to ~10x10^3 s^{-1}, that includes photon-by-photon analysis and relies on the utilization of a series of established tools[19, 26, 27, 38]. This strategy constitutes the following steps: (i) smFRET measurements are performed and the number of subpopulations in the FRET histogram is qualitatively identified. This step defines the minimal possible number of conformational states, *q*; (ii) BVA is applied to the data set to determine whether one (or more) of the FRET subpopulation(s) exhibit(s) FRET dynamics in within sm bursts. Presence of FRET dynamics hints the total number of states that may be larger than *q* (due to fast dynamics occurring while the molecule traverses the observation volume); (iii) photon-by-photon hidden Markov modelling (H2MM) analysis is applied to photon bursts with models of increasing number of states (starting from the *q*-state model). H2MM analyses results are assessed by comparing the values of a modified Bayes information criterion (*BIC'*; see Eq. S11) to determine the minimal sufficient number of states; (iv) the mean FRET efficiency values $\bar{E}_w$ are extracted for the H2MM identified states (Eqs. S12 and S13) and transformed into mean D-A distances $\bar{r}_{D-A}$ (Eqs. S16-S18); (v) steps (i)-(iv) are repeated for all labeled constructs; The more labeled constructs of more reaction coordinates are measured, the more accurate and reliable the dynamic structure determination will be; (vi) the different distances of different labeled constructs are classified and grouped to represent different conformational states; (vii) grouped $\bar{r}_{D-A}$ value sets are used as constraints in FPS to determine dynamic conformations.

In cases where crystal structures exist, FPS-determined conformations can be compared to the crystal structures and identified. FPS conformations that do not match the crystal structure might be destabilized in the crystallization process. Even when crystal structures of specific conformations do not exist, the same procedure could be used against structures computed by molecular dynamics (MD)[19, 43-48], Monte Carlo (MC)[20, 49] or coarse-grained (CG)[50-52] simulations or against database predicted structures[53].

A standardization effort for smFRET-based structure determination, led by the Hügel and Seidel groups, was initiated following the wwPDB Hybrid/Integrative methods

task force recommendations[63]. The approach presented here extends the above-mentioned efforts and outlines a general approach for assessing rapidly interconverting conformations. We demonstrated the approach for the transcription bubble at RPo with the lacCONS promoter. Due to its generality it could be used to elucidate interconverting conformations of other macromolecules.

VI. SUPPLEMENTARY MATERIALS

The supplementary materials include description of the µsALEX-smFRET approach (supplementary text, section A), description of the BVA method (supplementary text, section B), description of the H2MM approach (supplementary text, section C), description of FRET efficiencies to D-A distances conversion protocol (supplementary text, section D), steady-state fluorescence measurements results for singly-labeled (D or A) lacCONS promoter (Fig. S1), BVA results (Fig. S2), promoter opening and escape activity results (Fig. S3), H2MM state-model assessment example (Fig. S4), correction factors (Table S1), and H2MM analysis results (Table S2).

In this work we analyzed sm bursts, performed BVA and generated Figs. 2, S2 and S3 using the FRETBursts software[64]. FRETBursts Python Notebooks that include the analyses and generation of these figures are deposited in Figshare (https://figshare.com/s/59e477f31314ad16aa43, https://figshare.com/s/fd38b4e50c26c1824a12 and https://figshare.com/s/a7f97f43aab56437b746, respectively), The raw data used to generate these figures was saved in photonHDF5 format[65] and deposited in Figshare as well (https://figshare.com/s/5b9c53ec1029e18d0971, https://figshare.com/s/910547037a24331327c3 and https://figshare.com/s/51f10afcb4b58536eee6). Using the ALiX software[66], we extracted the photon stream identity and timestamp information, for all photons in the selected bursts (deposited in https://figshare.com/s/2c53c3fd40d6f755e745). These files were used as the base input for the H2MM analysis. The results of the H2MM analyses are stored in Matlab files (https://figshare.com/s/daec9ace72fb3733a7bf). In addition, the core code for the H2MM analysis was provided by Pirchi, Tsukanov *et al.*[27] Additional matlab scripts for dwell analysis (including additional Viterbi algorithm code, generously provided by Dr. Menachem Pirchi) are also deposited in Figshare, including code documentation and script usage instructions (https://figshare.com/s/cd089c26dcf04700b079). All pdb files used in FPS (https://figshare.com/s/8c4c9e21f967cac46c4d and https://figshare.com/s/9ffa0ececc19e857041b), input data (https://figshare.com/s/6469d9f5151c47676c72), analysis results (https://figshare.com/s/d9f0adada15eda2d8cd6 and https://figshare.com/s/8d684009d8d750677d1e for *RPo state* and *scrunched RPo state*, respectively) and PyMOL scripts used for preparation of Fig. 5 (https://figshare.com/s/e118297c1ca739768598), are also stored in Figshare.

VI. ACKNOWLEDGEMENTS

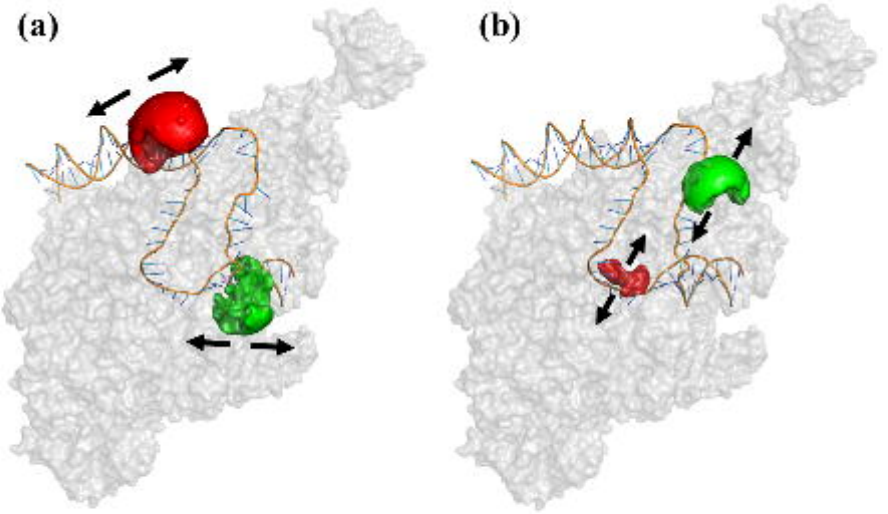
We would like to thank Dr. Xavier Michalet, Dr. Antonino Ingargiola, Dr. SangYoon Chung, Mr. Yazan Alhadid, Prof. Arieh Warshel and Dr. Raphael Alhadeff for fruitful discussions, Dr. Xavier Michalet for consultation regarding ALiX software, Dr.

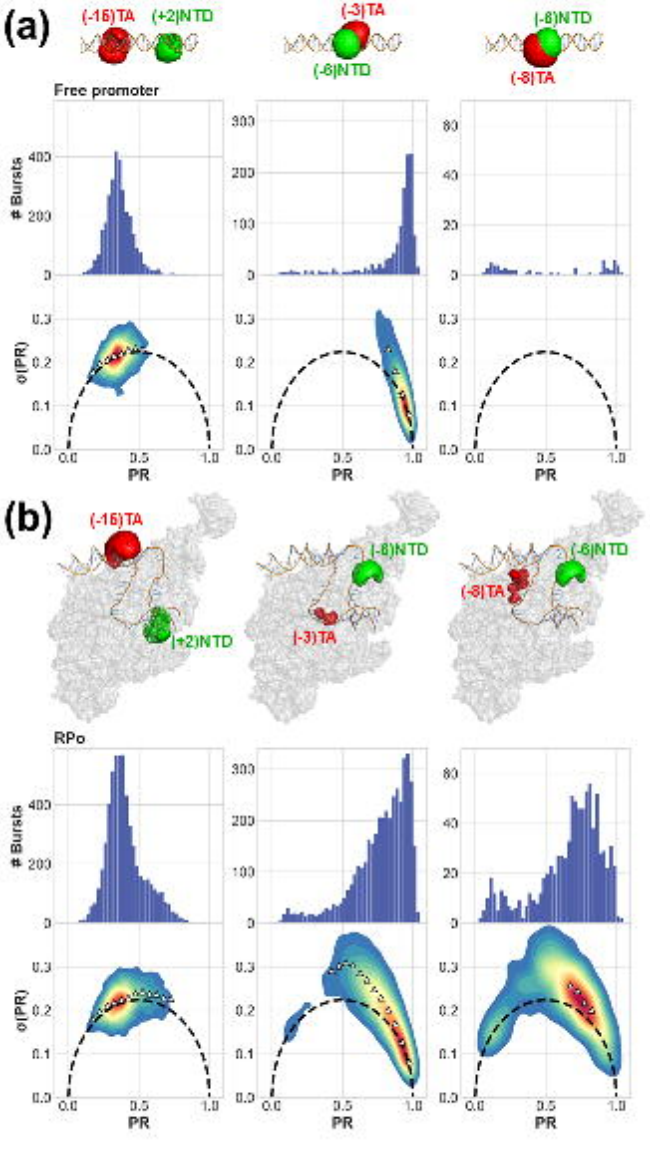
Antonino Ingargiola for consultation regarding FRETBursts software and Dr. Menachem Pirchi for consultation regarding H2MM. We thank Prof. Thorben Cordes for generously providing us with some of the labeled DNA oligos. This work was supported by the NIH under grant numbers GM069709 (to SW) and GM095904 (to XM and SW); NSF under grant number MCB-1244175 (to SW).

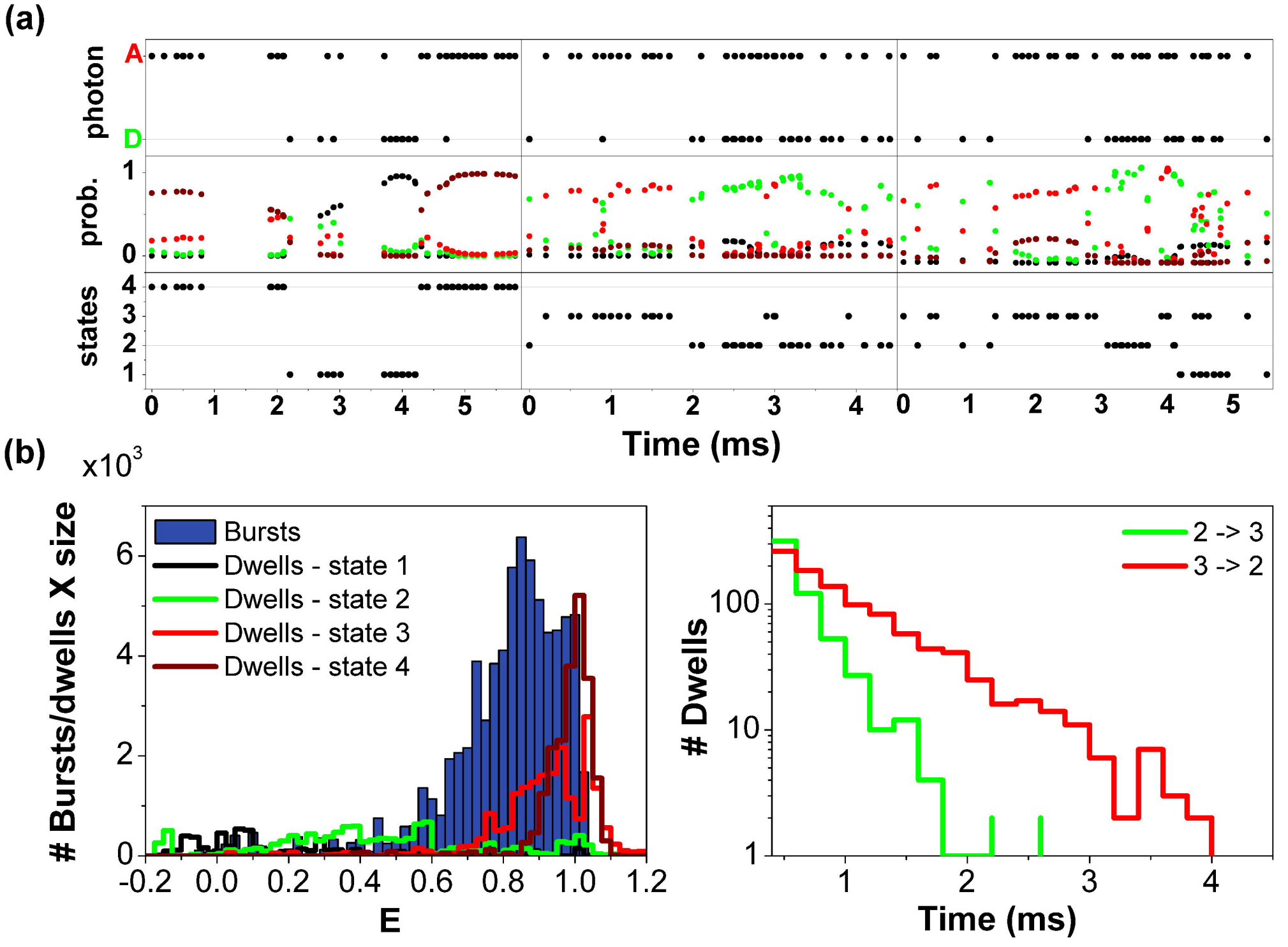
VII. REFERENCES

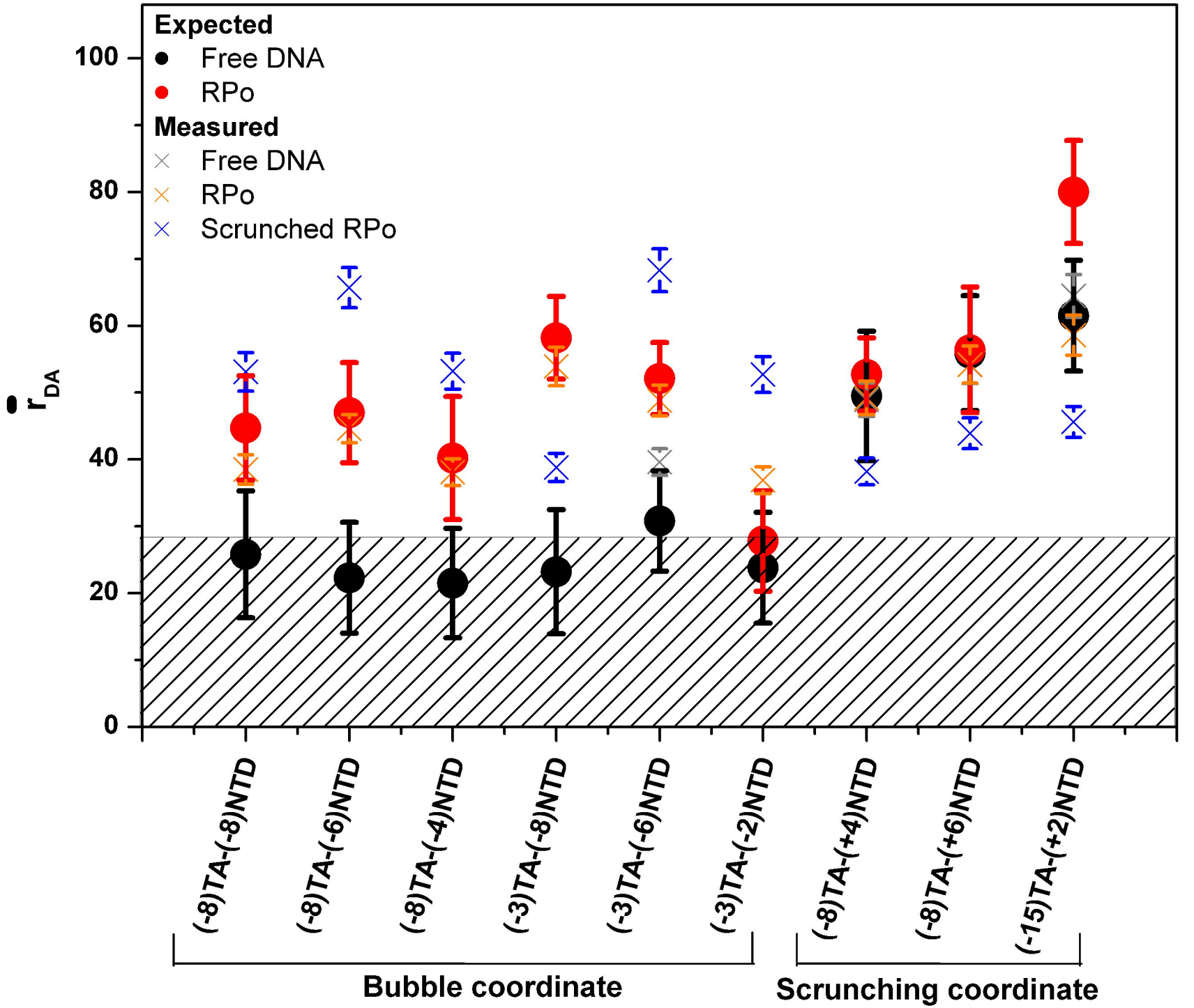
[1] B. Bae *et al.*, eLife **4** (2015).
[2] A. T. Brunger, Nat Struct Biol **4 Suppl** (1997).
[3] A. H. Zewail, Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences **363** (2005).
[4] T. A. Bharat *et al.*, Structure **23** (2015).
[5] Y. Cheng, Cell **161** (2015).
[6] M. Hennig, A. Ruf, and W. Huber, Top Curr Chem **317** (2012).
[7] H. S. Mchaourab, P. R. Steed, and K. Kazmier, Structure **19** (2011).
[8] G. Jeschke, Annu Rev Phys Chem **63** (2012).
[9] R. Rigler, and M. Ehrenberg, Quarterly Reviews of Biophysics **6** (1973).
[10] E. Haas, Methods Mol Biol **895** (2012).
[11] H. Yang *et al.*, Science **302** (2003).
[12] G. Lipari, and A. Szabo, Journal of the American Chemical Society **104** (1982).
[13] E. Meirovitch *et al.*, Prog Nucl Magn Reson Spectrosc **56** (2010).
[14] T. Ha *et al.*, Proceedings of the National Academy of Sciences of the United States of America **93** (1996).
[15] P. R. Selvin *et al.*, Biophysical journal **70** (1996).
[16] T. D. Craggs, and A. N. Kapanidis, Nature Methods **9** (2012).
[17] M. Dimura *et al.*, Curr Opin Struct Biol **40** (2016).
[18] B. Hellenkamp *et al.*, Nat Methods **14** (2017).
[19] S. Kalinin *et al.*, Nature Methods **9** (2012).
[20] A. Muschielok *et al.*, Nature Methods **5** (2008).
[21] T. O. Peulen, O. Opanasyuk, and C. A. M. Seidel, J Phys Chem B (2017).
[22] S. Kalinin *et al.*, Journal of Physical Chemistry B **114** (2010).
[23] E. Nir *et al.*, Journal of Physical Chemistry B **110** (2006).
[24] A. Hoffmann *et al.*, Physical Chemistry Chemical Physics **13** (2011).
[25] T. E. Tomov *et al.*, Biophysical journal **102** (2012).
[26] J. P. Torella *et al.*, Biophysical journal **100** (2011).
[27] M. Pirchi *et al.*, J Phys Chem B **120** (2016).
[28] R. M. Saecker, M. T. Record, Jr., and P. L. Dehaseth, Journal of molecular biology **412** (2011).
[29] E. F. Ruff, M. T. Record, Jr., and I. Artsimovitch, Biomolecules **5** (2015).
[30] G. Zhang *et al.*, Cell **98** (1999).
[31] S. P. Haugen, W. Ross, and R. L. Gourse, Nature reviews. Microbiology **6** (2008).
[32] A. Grove *et al.*, The EMBO journal **21** (2002).
[33] D. E. Lewis, and S. Adhya, Molecular microbiology **54** (2004).
[34] N. C. Robb *et al.*, Journal of molecular biology **425** (2013).
[35] A. N. Kapanidis *et al.*, Science **314** (2006).
[36] A. Revyakin *et al.*, Science **314** (2006).
[37] S. Kim *et al.*, Nat Methods **8** (2011).
[38] A. N. Kapanidis *et al.*, Proceedings of the National Academy of Sciences of the United States of America **101** (2004).
[39] A. N. Kapanidis *et al.*, Accounts of chemical research **38** (2005).
[40] N. K. Lee *et al.*, Biophysical journal **88** (2005).
[41] E. Lerner *et al.*, bioRxiv (2016).
[42] E. Lerner *et al.*, Transcription (2017).

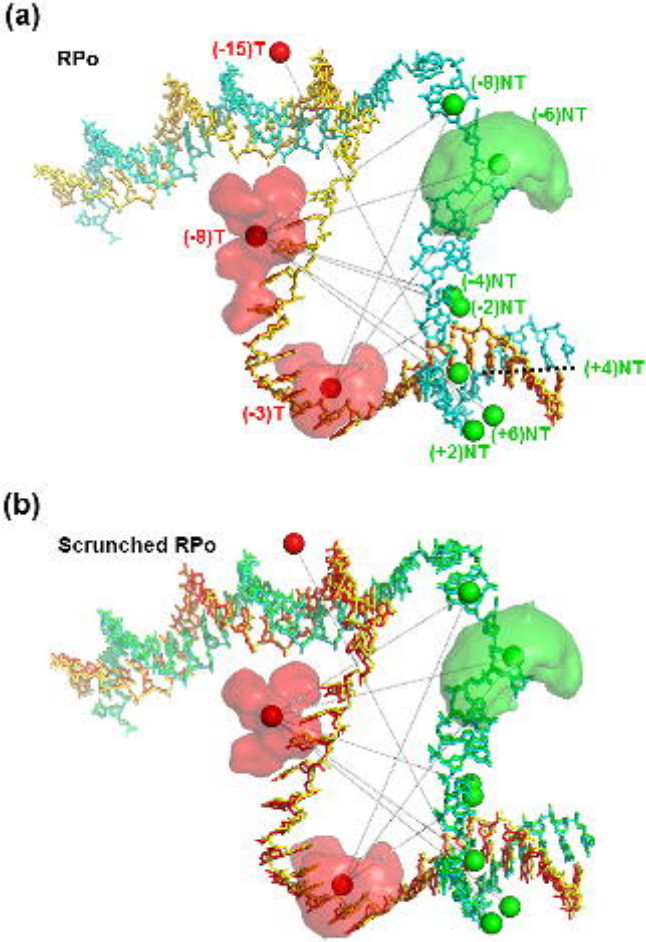
[43] J. J. McCann *et al.*, Proceedings of the National Academy of Sciences of the United States of America **109** (2012).
[44] S. Sindbert *et al.*, Journal of the American Chemical Society **133** (2011).
[45] T. Sabir *et al.*, Journal of the American Chemical Society **133** (2011).
[46] T. D. Craggs *et al.*, Nucleic Acids Research **42** (2014).
[47] L. A. Metskas, and E. Rhoades, Journal of the American Chemical Society **137** (2015).
[48] U. B. Choi *et al.*, Nat Struct Mol Biol **17** (2010).
[49] A. Nath *et al.*, Biophysical journal **103** (2012).
[50] W. Zheng, N. P. Schafer, and P. G. Wolynes, Proceedings of the National Academy of Sciences of the United States of America **110** (2013).
[51] M. B. Borgia *et al.*, Nature **474** (2011).
[52] H. Hofig *et al.*, Molecules **19** (2014).
[53] A. I. Tomescu *et al.*, Proceedings of the National Academy of Sciences of the United States of America **111** (2014).
[54] I. Rasnik *et al.*, Journal of molecular biology **336** (2004).
[55] T. Cordes *et al.*, Biochemistry **49** (2010).
[56] I. O. Vvedenskaya *et al.*, Proceedings of the National Academy of Sciences of the United States of America **113** (2016).
[57] I. O. Vvedenskaya *et al.*, Molecular Cell **60** (2015).
[58] J. T. Winkelman *et al.*, Proceedings of the National Academy of Sciences of the United States of America **113** (2016).
[59] J. T. Winkelman, and R. L. Gourse, BioEssays : news and reviews in molecular, cellular and developmental biology **39** (2017).
[60] J. T. Winkelman *et al.*, Science **351** (2016).
[61] S. E. Kim *et al.*, Journal of Physical Chemistry B **119** (2015).
[62] R. Roy, S. Hohng, and T. Ha, Nature Methods **5** (2008).
[63] A. Sali *et al.*, Structure **23** (2015).
[64] A. Ingargiola *et al.*, Plos One **11** (2016).
[65] A. Ingargiola *et al.*, Biophysical journal **110** (2016).
[66] A. Ingargiola *et al.*, bioRxiv (2016).


(a)
(b)

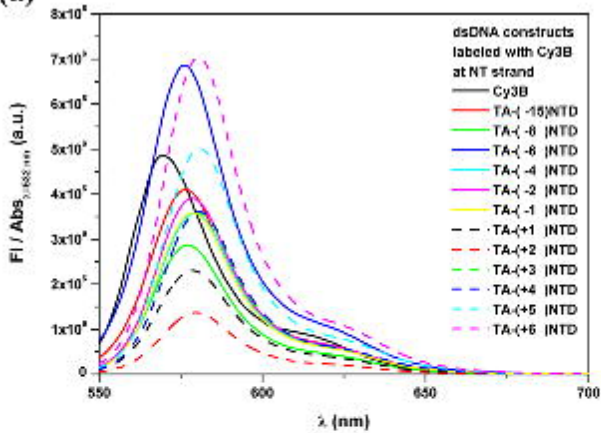

(a)
(-16)TA
(+2)NTD
(-3)TA
(-6)NTD
(-6)NTD
(-8)TA
Free promoter
Bursts
σ(PR)
PR
(b)
(-16)TA
(+2)NTD
(-6)NTD
(-3)TA
(-6)NTD
(-8)TA
RPo
Bursts
σ(PR)
PR


(a)
photon
A
D
prob.
1
0
states
4
3
2
1
0 1 2 3 4 5 0 1 2 3 4 0 1 2 3 4 5
Time (ms)
(b)
x10^3
Bursts/dwells X size
Bursts
Dwells - state 1
Dwells - state 2
Dwells - state 3
Dwells - state 4
6
4
2
0
-0.2 0.0 0.2 0.4 0.6 0.8 1.0 1.2
E
Dwells
2 -> 3
3 -> 2
100
10
1
1 2 3 4
Time (ms)

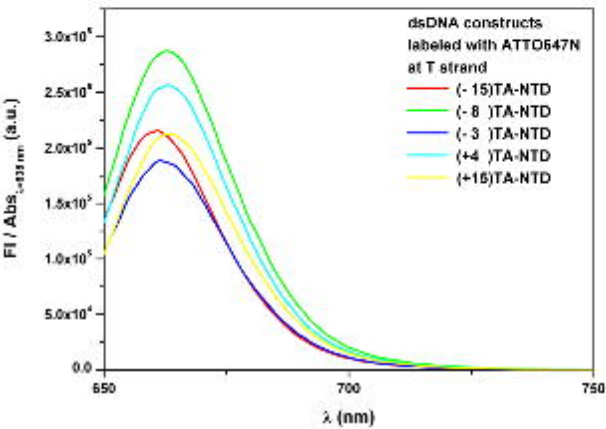

Expected
Free DNA
RPo
Measured
Free DNA
RPo
Scrunched RPo
100
80
60
40
20
0
r_{DA}
(-8)TA-(-8)NTD
(-8)TA-(-6)NTD
(-8)TA-(-4)NTD
(-3)TA-(-8)NTD
(-3)TA-(-6)NTD
(-3)TA-(-2)NTD
(-8)TA-(+4)NTD
(-8)TA-(+6)NTD
(-15)TA-(+2)NTD
Bubble coordinate
Scrunching coordinate

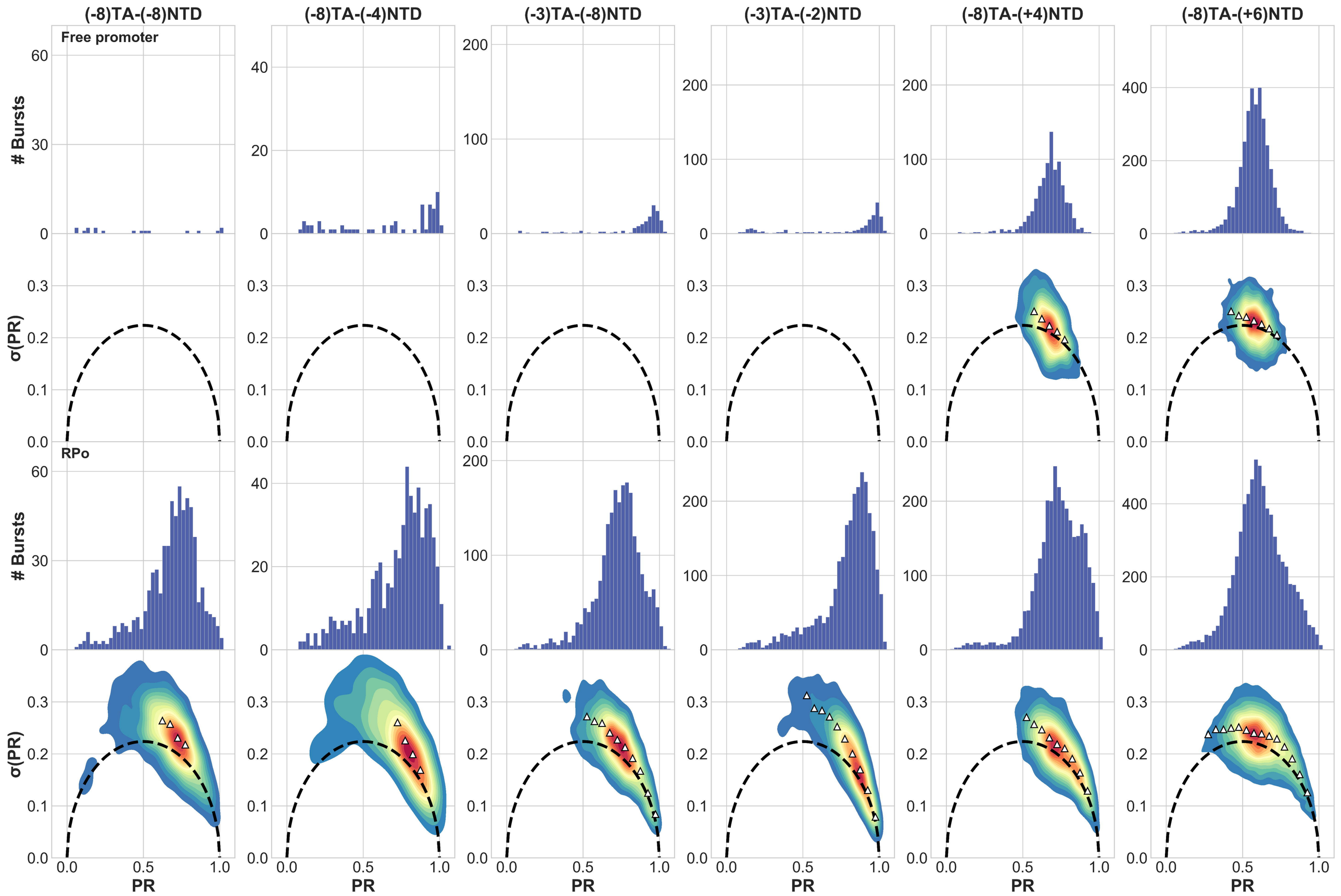

(a)
RPo
(-15)T
(-8)NT
(-6)NT
(-8)T
(-4)NT
(-2)NT
(+4)NT
(-3)T
(+6)NT
(+2)NT
(b)
Scrunched RPo

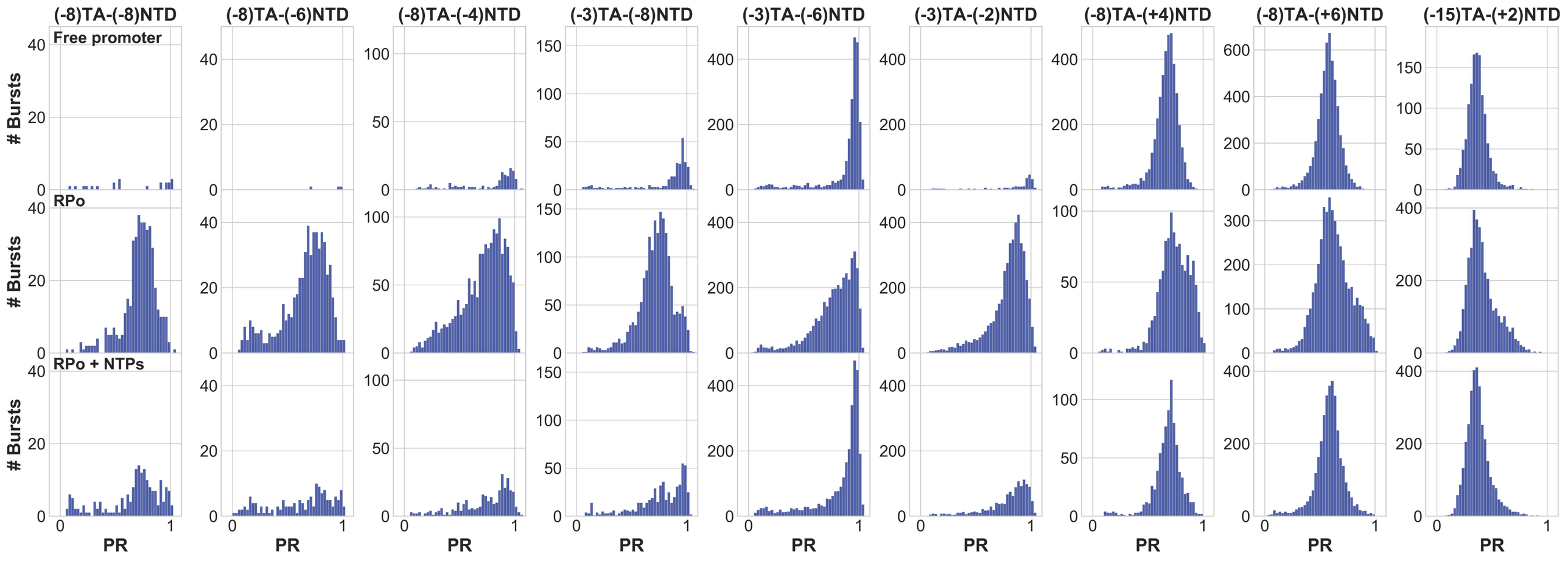
(a)

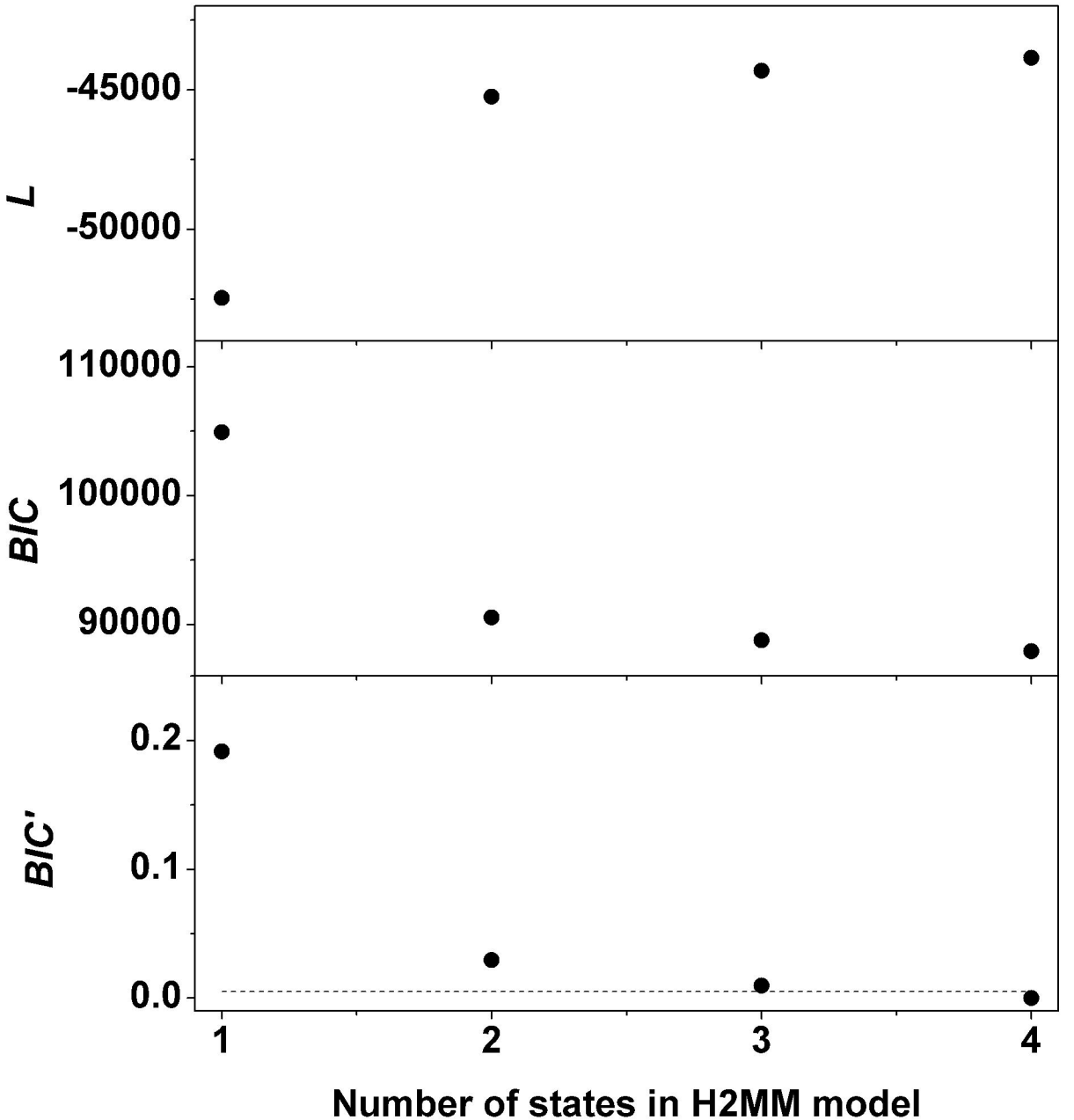

dsDNA constructs
labeled with Cy3B
at NT strand
Cy3B
TA-(-15)NTD
TA-(-8)NTD
TA-(-6)NTD
TA-(-4)NTD
TA-(-2)NTD
TA-(-1)NTD
TA-(+1)NTD
TA-(+2)NTD
TA-(+3)NTD
TA-(+4)NTD
TA-(+5)NTD
TA-(+6)NTD
Fl / Abs (a.u.)
λ (nm)

(b)


dsDNA constructs
labeled with ATTO647N
at T strand
(- 15)TA-NTD
(- 8)TA-NTD
(- 3)TA-NTD
(+4)TA-NTD
(+15)TA-NTD
Fl / Abs (a.u.)
λ (nm)


(-8)TA-(-8)NTD
(-8)TA-(-4)NTD
(-3)TA-(-8)NTD
(-3)TA-(-2)NTD
(-8)TA-(+4)NTD
(-8)TA-(+6)NTD
Free promoter
RPo
Bursts
σ(PR)
PR


(-8)TA-(-8)NTD
(-8)TA-(-6)NTD
(-8)TA-(-4)NTD
(-3)TA-(-8)NTD
(-3)TA-(-6)NTD
(-3)TA-(-2)NTD
(-8)TA-(+4)NTD
(-8)TA-(+6)NTD
(-15)TA-(+2)NTD
Free promoter
RPo
RPo + NTPs
Bursts
PR


L
-45000
-50000
BIC
110000
100000
90000
BIC'
0.2
0.1
0.0
1
2
3
4
Number of states in H2MM model