HHS Public Access

Author manuscript

Nat Chem Biol. Author manuscript; available in PMC 2021 December 17.

Published in final edited form as:

Nat Chem Biol. 2021 August ; 17(8): 906–914. doi:10.1038/s41589-021-00817-3.

Transcriptional processing of an unnatural base pair by eukaryotic RNA polymerase II

Juntaek Oh[1], **Ji Shin**[#1,7], **Ilona Christy Unarta**[#2], **Wei Wang**[1,8], **Aaron W. Feldman**[3], **Rebekah J. Karadeema**[3], **Liang Xu**[1,9], **Jun Xu**[1], **Jenny Chong**[1], **Ramanarayanan Krishnamurthy**[3], **Xuhui Huang**[2], **Floyd E. Romesberg**[4], **Dong Wang**[1,5,6,*]

[1]Division of Pharmaceutical Sciences, Skaggs School of Pharmacy & Pharmaceutical Sciences; University of California, San Diego, La Jolla, California 92093, United States

[2]Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

[3]Department of Chemistry, The Scripps Research Institute, La Jolla, California 92037, United States

[4]Synthorx, a Sanofi Company, La Jolla, CA 92037, United States

[5]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, United States

[6]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, United States

[7]Present Address: Department of Molecular Microbiology, Center for Advanced Laboratory Medicine (CALM), University of California, San Diego, La Jolla, California, 92093

[8]Present Address: Advanced Medical Research Institute, Shandong University, Jinan 250012, China

[9]Present Address: Department of Chemistry, Sun Yat-Sen University, Guangzhou 510275, China

[#] These authors contributed equally to this work.

Abstract

The development of unnatural base pairs (UBPs) has greatly increased the information storage capacity of DNA, allowing for transcription of unnatural RNA by the heterologously expressed T7 RNA polymerase (RNAP) in *Escherichia coli.* However, little is known about how UBPs are

[*]**Corresponding Author:** Correspondence to Dong Wang. To whom correspondence should be addressed. Tel: +1 858 822 5561; Fax: +1 858 822 1953; dongwang@ucsd.edu.

Author Contributions

F.E.R. and D.W. conceived the project. J.O., and W.W. performed structural analysis. I.C.U. and X.H. perform MD simulation. J.O., J.S., J.X., J.C. and X.L performed biochemistry experiments. A.W.F., R.J.K R.K., and F.E. R. prepared unnatural DNA templates and nucleotide triphosphate. J.O. and D.W. performed data analysis. D.W. supervised the different aspects of the work. J.O., D.W., and F.E.R. wrote the manuscript, with input from all authors.

Competing interests

All authors declare no competing interests.

transcribed by cellular RNA polymerases. Here, we investigate how synthetic unnatural nucleotides, NaM and TPT3, are recognized by eukaryotic RNA polymerase II (Pol II) and found that Pol II is able to selectively recognize UBPs with high fidelity when dTPT3 is in the template strand and rNaMTP acts as the nucleotide substrate. Our structural analysis and molecular dynamics simulation provide structural insights into transcriptional processing of UBPs in a stepwise manner. Intriguingly, we identified a novel 3'-RNA binding site after rNaM addition, termed the swing state. These results may pave the way for future studies to design transcription and translation strategies in higher organisms with expanded genetic codes.

Keywords

Unnatural base pairs; RNA polymerase II; transcription; synthetic biology; structural biology

INTRODUCTION

All organisms store their genetic information using a natural four-letter genetic alphabet system. Accurate duplication and transfer of genetic information (i.e., replication, transcription, and translation) partially rely on specific hydrogen bonding between purines and complementary pyrimidines that form the two natural base pairs. The integration of an additional unnatural base pair (UBP) that functions alongside the natural nucleotides would greatly increase the information storage capacity of DNA, allow the transcription of unnatural RNA, and enable the production of proteins containing unnatural amino acids. Pioneering work from the Benner, Romesberg, and Hirao groups reported the development of distinct UBPs that are well replicated and transcribed *in vitro*[1–5]. The family of predominantly hydrophobic UBPs identified by the Romesberg lab, represented by dNaM-d5SICS and dNaM-dTPT3, relies on hydrophobic and packing forces for their pairing (Fig. 1a)[6]. These UBPs have expanded the genetic alphabet and code in *Escherichia coli* through high fidelity replication of the UBP in DNA, which allows for the faithful retrieval of that information through the transcription of unnatural base - containing mRNAs and tRNAs (by heterologously expressed T7 RNAP), and translation via unnatural base pairing at the ribosome[7–9].

While hydrophobic dNaM-dTPT3 or dNaM-d5SICS bears little resemblance to natural base pairs, it has been shown that these UBPs support faithful replication by DNA polymerases[6,10–13]. Previous X-ray crystal structures of DNA containing dNaM-d5SICS in complex with KlenTaq DNA polymerase revealed that the dNaM-d5SICS UBP can form via co-planar edge-to-edge base pairing in a conformation similar to a natural base pair within the KlenTaq polymerase active site, despite the absence of Watson-Crick-like hydrogen bonds[6]. These structures provided insight into the mechanism for selective replication of hydrophobic UBPs.

In the *E. coli* semi-synthetic organism (SSO), the unnatural nucleotides are transcribed by the heterologously expressed single subunit RNA polymerase from T7 bacteriophage. Recent studies suggest that the fidelity of transcription by the bacteriophage polymerase in the SSO does not limit the fidelity of unnatural protein production[14], and in vitro,

transcription of a template containing d5SICS with NaM proceeded with fidelity that is comparable to a natural pair's transcription and the transcription of a template containing dNaM with 5SICS proceeded with fidelity of 92%[2,5,14,15], the mechanism of the transcription of DNA containing UBPs has not been extensively investigated as the mechanism of replication. Moreover, current efforts to optimize the *E. coli* SSO or to create semi-synthetic eukaryotes would be facilitated utilizing their cellular multi-subunit RNA polymerases, which bear limited structural or functional homology to the single subunit polymerases. Multi-subunit RNA polymerases are highly conserved in all three domains of life (i.e., archaea, bacteria, and eukarya domains). These RNA polymerases share common structural architectures and molecular mechanisms for nucleotide selection, nucleotide addition, translocation, and proofreading[16,17]. In eukaryotes, RNA polymerase II (Pol II) is responsible for recognizing protein coding genes and synthesizing messenger RNA, which serves as a template for protein translation. Pol II has high transcriptional fidelity with an error rate of about 10^{-5} to 10^{-6} [18]. The impacts of various DNA lesions on Pol II transcription have been well-documented[19–24]. By contrast, it is not known how Pol II recognizes and processes UBPs. Does Pol II recognize UBPs as a normal template and substrate? Are they transcribed faithfully as they are with T7 RNAP? Elucidating the mechanisms of UBP processing by Pol II will provide a framework for the rational design of new unnatural synthetic nucleotides that can be effectively utilized by native cellular RNA polymerases.

Here we present combined biochemical, structural and computational data revealing that Pol II is able to recognize the dNaM or dTPT3 template and utilize unnatural nucleoside triphosphates as substrates during transcription. In particular, we found that Pol II recognizes the NaM-TPT3 base pair in an asymmetric and strand-specific manner. Furthermore, we solved three crystal structures revealing the structural basis of UBP transcription by Pol II in a step-wise manner. Our structural analysis, together with molecular dynamics (MD) simulation, explains the high substrate selectivity but slow kinetics of UBP addition and extension by Pol II.

Results

Pol II recognizes NaM-TPT3 base pair in an asymmetric manner

To investigate whether Pol II can recognize the NaM-TPT3 pair, we purified 12-subunit RNA Pol II from *Saccharomyces cerevisiae* and reconstituted an elongation complex (Pol II EC) containing either a site-specific dNaM or dTPT3 in the template DNA strand (Fig. 1b). For single-nucleotide incorporation experiments, 1 mM of individual NTP (ATP, UTP, GTP or CTP) or unnatural NTP (rNaMTP or rTPT3TP) were incubated with Pol II EC harboring a template containing dTPT3 or dNaM. As shown in Figure 1c, Pol II has a strong selectivity for effectively incorporating rNaM opposite the dTPT3 template, whereas the other four canonical NTPs fail to be efficiently incorporated even after prolonged incubation up to 24 hours. Quantitatively, the K_{obs} value for rNaM incorporation is two orders of magnitude higher than the incorporation of natural NTPs (Supplementary Table 1). By contrast, this selectivity was reduced when we tested substrate incorporation opposite the dNaM template. While rTPT3TP is among the best substrate to be incorporated opposite the dNaM template,

the K_{obs} value of rTPT3TP is only about 1.4 to14-fold higher than natural NTPs (Fig. 1c,d and Supplementary Table 1). These results suggest that the recognition and selection of this UBP by Pol II is asymmetric. dTPT3 provides strong selection pressure such that only rNaMTP is preferred for incorporation, while dNaM allows incorporation of both rTPT3TP and ATP.

Transcription extension of dTPT3 template requires rNaMTP

To investigate whether Pol II is able to extend from a template containing dTPT3, we performed elongation assays with natural NTPs in the absence or presence of rNaMTP (Fig. 1e). We observed strong Pol II stalling right before the dTPT3 position (10mer), and almost no bypass products were detected in the absence of rNaMTP. This result further supports the high substrate selectivity of dTPT3. In sharp contrast, this dTPT3-induced transcriptional pausing/stalling can be resolved by adding rNaMTP to the system, where Pol II was able to extend beyond dTPT3 in the presence of rNaMTP (Fig. 1e). This result demonstrates that rNaMTP is essential for Pol II to bypass dTPT3 stalling and produce an extended transcription product. Furthermore, we did not observe transcriptional stalling at the n+1 position (11-mer RNA band), indicating that once rNaM is added, Pol II can effectively extend from the dTPT3-rNaM UBP pair (Fig. 1e). Therefore, the rate-limiting step for UBP transcription is the rNaMTP addition step, which is slower than the subsequent extension step. Taken together, our biochemical analysis further supports the ability of Pol II to selectively incorporate rNaM opposite dTPT3 in the presence of the natural NTPs and to extend beyond the dTPT3 position.

Template loading and substrate binding of dTPT3 – rNaMTP pair

To understand the mechanism of selective UBP recognition by Pol II, we solved the crystal structure of Pol II EC with a site-specific dTPT3 in the template strand, termed apo Pol II-dTPT3 EC (Fig. 2a,b). Overall, the structure of Pol II EC with dTPT3 was similar to that of other Pol II EC structures. The RMSD values between apo Pol II-dTPT3 EC and Pol II EC (with a natural DNA scaffold) at the post-translocation state (PDB ID: 6UQ2), pre-translocation state (PDB ID: 1I6H), frayed (PDB ID: 3HOZ and 3HOW) or backtracked (PDB ID: 3GTG) states were 0.51 Å, 0.69 Å, 0.74 Å, 0.72 Å, and 0.57Å, respectively [24–27]. We found that dTPT3 was fully loaded into the canonical +1 active site, despite the hydrophobic and unnatural feature of its nucleobase (Fig. 2c). Structural alignment between Pol II-dTPT3 EC and Pol II EC with dG in its +1 site (PDB 6UQ2) indicates that dTPT3 is recognized as a normal base during the template loading by Pol II (Fig. 2c)[28].

Previous studies of Pol II revealed two canonical substrate binding sites: entry site (E site) and addition site (A site) (Fig. 3). The E site is in an inverted conformation in which the base is facing away from the template strand, whereas the A site establishes Watson-Crick base pairs through hydrogen bonds and base stacking with the RNA primer (Fig. 3a). It is proposed that the nucleotide substrates first diffuse through the secondary channel and bind at the E site and then rotate into the A site[29–31]. The nucleotide substrate binding at the A site is a prerequisite for Pol II active site alignment and nucleotide addition.

To investigate how rNaMTP is accommodated in the Pol II active site, we also solved the rNaMTP-bound Pol II-dTPT3 EC structure. Interestingly, we found strong electron density at the E site instead of the A site (Fig. 3a and Extended Data Fig. 1a), indicating rNaMTP prefers the E site. We found that the hydroxyl group of ribose interacts with S1019 of Rpb2, and the triphosphate moiety of NTPs interacts with two arginine residues, R766 and R1020, of Rpb2. No direct interaction with the base of rNaMTP is observed at the E site. This rNaMTP-bound structure provides a structural explanation as to why the incorporation efficiency of rNaMTP opposite to dTPT3 template is much lower than that of natural cognate substrate to its complementary template. Because rNaMTP energetically favors the E site instead of the A site, the transition of rNaMTP from the E site to the A site is energetically unfavored, and therefore, the incorporation of rNaMTP into the 3'-RNA is slow (Fig. 3a,b).

MD simulation further supports dTPT3 – rNaMTP selectivity

Our biochemical and structural analyses show selective but slow incorporation of the dTPT3 – rNaMTP pair. However, it is not clear why a dTPT3 template prefers rNaM incorporation over natural ribonucleotides. Based on previous structural studies of Pol II transcription, substrate selection is mainly achieved when the substrate enters the A site, where several key interactions are established between substrate-template base pairs as well as between substrate and Pol II active site residues. The allowable distance between the 3' hydroxyl group (O3') of the RNA primer and alpha phosphate group (Pα) of an A site substrate is an important prerequisite for nucleotide addition. We hypothesize that Pol II utilizes the same principle to select rNaMTP over other natural NTPs at the A site. Thus, the rNaMTP binding at the A site is more favorable to form a catalytically active conformation than that for other natural NTPs. We therefore compared the stability of rNaMTP and natural NTPs in the A site using MD simulation to obtain further understanding of the molecular basis of the high selectivity for rNaMTP (see Methods section for simulation setup[32–35]) (Fig. 4 and Extended Data Fig. 2).

Our simulation results revealed strong localization of rNaMTP in terms of base pair distance and base plane angle, whereas other canonical NTPs show widely dispersed peaks, suggesting that canonical NTPs are less stable in the A site opposite a dTPT3 template (Fig. 4a–e, left plots). We further analyzed the distance distribution between the 3' hydroxyl group of the RNA primer and the alpha phosphate group of the substrate (Fig. 4a–e, right plots). Notably, according to quantum calculation, the distance between O3' (RNA) and Pα (NTP) is around 3.5 Å or below before the chemical reaction occurs[36]. Using these results, we defined good activation geometries which satisfy both criteria: O3' – Pα distance within a range of 3 to 3.5 Å that allows chemistry (nucleotide incorporation); base pair distance within a range of 7 to 9 Å (for rNaMTP, GTP and ATP) or 6 to 8 Å (for CTP and UTP)[36,37]. Again, we confirmed that the nucleophilic attack distance in rNaMTP is clustered at a distance less than 3.5 Å, while those distances of other NTPs were highly dispersed across 3 to 8 Å (Fig. 4a–e, right plots). Among all five substrates tested, rNaMTP has the highest score, with 40% of total simulation frames adopting a catalytically active conformation while that of GTP, ATP, UTP and CTP were 14, 4, 0.4 and 3%, respectively (Fig. 4f). Importantly, these MD simulation results agree well with our biochemistry data and support

specific incorporation of rNaMTP by Pol II. We also performed MD simulation with dNaM in the template position and rTPT3TP and ATP as the incoming substrate to elucidate the mechanism of low selectivity in the dNaM template (Extended Data Fig. 3). Frames adopting good activation geometry for rTPT3TP and ATP were both around 17%. This result shows that rTPT3TP and ATP have a similar probability of being incorporated when the template DNA contains dNaM, which is in contrast with the high selectivity of dTPT3 for rNaM incorporation.

Post-incorporation structure reveals new rNaM binding site

To further investigate the conformation of dTPT3-rNaM base pairing within the Pol II active site, we solved the structure of the post-incorporation state of Pol II-dTPT3 EC with an rNaM incorporated into the 3’ end of the RNA. Interestingly, we found that the Pol II EC structure containing the dTPT3-rNaM pair was distinct from all canonical Pol II EC structures of natural scaffolds solved to date (Fig. 5 and Extended Data Fig. 1b).

Previous studies of Pol II EC with a natural scaffold revealed that upon nucleotide addition, the newly added NMP at 3’-RNA occupies the +1 position, termed the pre-translocation state (Fig. 5b)[27]. After 1 bp forward translocation, Pol II reaches the post-translocation state, in which the new 3’-RNA is translocated to the −1 position and leaves the +1 position empty for the next round of nucleotide addition[38–40]. In sharp contrast, we found that the newly incorporated 3’-rNaM is located at a novel site, termed the swing state. In this swing state, both the ribose and base moiety of rNaM retract to the position essentially halfway between the A and E sites. The rNaM at the swing state is stabilized by multiple chemical interactions, including hydrogen bonds, charge interactions, and Van der Waals interactions. The hydroxyl group of the ribose forms putative hydrogen bonds with two conserved arginine residues (Rpb2 R766 and R1020), whereas the phosphodiester backbone was stabilized by Mg^{2+} ions. The hydrophobic nucleobase is accommodated by several surrounding residues, including conserved residues Rpb1 T827, T831 from the bridge helix, Rpb1 N479, and Rpb2 Y769 (Fig. 5a, middle panel). Furthermore, the swing state is partially overlapped with the binding site of the TFIIS tip domain III (Fig. 5a, right panel). The distance between the center of the nucleobase of rNaM and the center of the −1 nucleobase is 8.1 Å, which is ~ 4.4 Å longer than regular base stacking (~ 3.7 Å). This unique swing state observed in our dTPT3-rNaM structure is structurally distinct from the canonical pre-translocation, frayed, or backtracked state (Fig. 5b). Backtracking of the Pol II elongation complex is characterized as a proofreading mechanism and can be driven by mismatch incorporation or pausing by transcription blockage[26,31,41]. In the canonical frayed state, the newly added 3’-RNA is flipped away from its original template, and both the frayed base and ribose occupy a pore site between the A and E sites. The population of the frayed state is thought to pause the complex and facilitate backtracking[25]. The orientation of the base and ribose moiety of rNaM in the swing state is completely distinct from the canonical frayed state or backtracked states (Fig. 5b).

Swing state provides structural insights into UBP extension

Previous structural studies of predominantly hydrophobic UBPs in a DNA polymerase active site and in duplex DNA revealed distinct patterns of edge-to-edge and cross-strand

intercalation base pairing of UBPs (Fig. 6a,b and Extended Data Fig. 4)[11,12]. To further understand how different enzyme active sites affect the pairing configuration of UBPs, we compared the dTPT3-rNaM pair within the Pol II active site with its closely related analogue pairs, d5SICS-dNaM in the active site of a DNA polymerase and d5SICS-dMMO2 within a free DNA duplex[11,12]. Interestingly, despite the close structural similarity among dTPT3-rNaM, d5SICS-dNaM, and d5SICS-dMMO2 pairs, the configurations and patterns of cross-strand intercalation UBPs are distinct (Fig. 6b and Extended Data Fig. 4). As shown in Fig. 6b, the nucleobase of dNaM at the terminus of the primer strand is sandwiched between the d5SCIS template and 3' neighbor template base (−1 position) in the DNA polymerase active site. The distance between dNaM and its 5' neighbor base is 4.7 Å, whereas the distance between d5SCIS and its 3' neighbor base is 6.5 Å. Interestingly, for the same DNA polymerase, d5SCISTP can form an edge-to-edge co-planar packing with a dNaM template (Extended Data Fig. S4). In the case of Pol II, however, the presence of the bridge helix rules out the possibility of a template strand shift to make space for the incoming nucleotide in the RNA primer to be accommodated between +1 dTPT3 and −1 template base (Fig. 6b). Instead, the Pol II active site has space available to allow for the movement of the newly added rNaM, producing a unique swing state of dTPT3-rNaM. The distance between rNaM and its 5' neighbor base is 8.1 Å, whereas the distance between dTPT3 and its 3' neighbor base is 4.1 Å. Interestingly, this pattern is similar to that observed with free duplex DNA where the base of d5SCIS is inserted between dMMO2 and its 5' neighbor base. The distance between dMMO2 and its 5' neighbor base is 6.0 Å, whereas the distance between d5SICS and its 3'neighbor base is 4.2 Å.

Based on our study and previous structural studies of UBPs[6,10–12], we propose a model for UBP processing during Pol II transcription (Fig. 6c). The incoming rNaMTP first binds to the E site. Equilibrium between the A and E sites then allows for the slow transition of rNaMTP into the A site. Among canonical and unnatural NTPs, matched rNaMTP has the highest probability of addition opposite the dTPT3 template, in terms of good activation geometry. After nucleotide addition and the pyrophosphate is released, the newly added rNaM retracts to the swing state. In the case of Pol II, the presence of the bridge helix restrains the position of the unnatural nucleobase (dTPT3) in the template strand, such that only rNaM is allowed to move into a cross-strand intercalated structure. Therefore, our swing state structure can be interpreted as a unique primed state, where retraction of rNaM is not for TFIIS cleavage, fraying or backtracking, but rather for preparing cross-strand intercalation during forward translocation (Fig. 6c, extension). Since Pol II residues mainly interact with upstream RNA/DNA hybrid phosphate backbone groups via positively charged residues, we did not expect a strong barrier for forward translocation of the RNA/DNA hybrid with cross-strand intercalated UBPs. Indeed, our biochemical analysis revealed that after addition, Pol II can effectively extend the RNA primer beyond the UBP site. Therefore, rNaM will intercalate between the +1 template dTPT3 and +2 base after translocation, without compromising the extension competence of Pol II. In this case, rNaM intercalation between the −1 and +1 template may not occur. In contrast, in the case of a DNA polymerase, UBP intercalation mainly depends on sequence context and not structural constraints. Therefore, the nucleobase of the newly added unnatural nucleotide in the primer can intercalate in either the −1/+1 or +1/+2 positions[12].

DISCUSSION

In this study, we showed that the TPT3-NaM base pair is selectively recognized by Pol II, albeit in an asymmetric, strand-specific manner. With dTPT3 in the template strand, Pol II has robust discrimination power and strongly prefers the incorporation of rNaMTP over other natural NTPs by at least two orders of magnitudes (Supplementary Table 1). In contrast, Pol II has modest discrimination power in selecting rTPT3TP opposite the dNaM template (Fig. 1). This asymmetric, strand-specific selection is distinct from that in DNA replication[13], as well as transcription by T7 RNAP[14]. Notably, there are several key differences between single subunit T7 RNAP and multi-subunit RNA Pol II. First, the active site architecture of single subunit T7 RNAP is dramatically different from that of Pol II[27,30,31,42,43]. Second, T7 RNAP does not have proofreading activity[44]. In sharp contrast, Pol II has a “dual mode” of proofreading activities: intrinsic cleavage and TFIIS-stimulated cleavage[45]. These differences between T7 RNAP and Pol II could account for the different transcription outcomes observed in our studies.

We solved a series of crystal structures of Pol II ECs containing a dTPT3 template and revealed three key steps of rNaMTP incorporation: template loading, substrate binding (pre-chemistry), and substrate addition (post-chemistry). The template loading step can be defined as when the template base crosses over the bridge helix and is loaded into the Pol II active site (+1 template position)[31,46]. In this state, the template is positioned to form a base pair with the incoming cognate NTP. Template loading is a critical checkpoint for Pol II to detect if there are any structural or chemical alterations in template bases, such as DNA lesions. Indeed, recent structural studies of Pol II EC with DNA lesions, such as abasic site, cyclobutane pyrimidine dimer, 8,5′-cyclo-2′-deoxyadenosine, or 5-guanidinohydantoin show that these lesions are prone to be stuck above the bridge helix and fail to be properly loaded to the +1 active site, resulting in a half-translocated state[21–24]. Intriguingly, our apo Pol II EC structure harboring dTPT3 revealed that dTPT3 is able to be fully loaded at the canonical +1 template state despite its nucleobase being hydrophobic and bearing little structural similarity to a natural nucleobase (Fig. 2c).

For the substrate binding (pre-chemistry) state, interestingly, we observed rNaMTP occupying the E site instead of the A site. These results suggest that binding at the A site is energetically unfavored relative to the E site, presumably due to a lack of hydrogen bonding between rNaMTP and dTPT3 in comparison with canonical Watson-Crick pairs. To further understand the molecular basis of the high selectivity of rNaMTP incorporation, we performed MD simulation. Intriguingly, our simulation mimics the scanning movement of each substrate to search for the most stable conformation prior to nucleotide addition. It is noteworthy that rNaMTP shows a single, highly concentrated island, and most of the frames also maintain the allowed distance of nucleophilic SN_2 attack (Fig. 4). In contrast, other NTPs are largely dispersed in the base pairing and nucleophilic SN_2 attack distance plot[47]. As a result, dTPT3-rNaM shows the highest ratio of catalytically active conformation, revealing the mechanism of substrate selectivity in the A site. Taken together, our structural and computational studies are in good agreement with biochemical results and explain the high selectivity and the relatively slow kinetics of rNaMTP incorporation opposite dTPT3 in a template.

The post-chemistry structure of Pol II EC containing a newly incorporated rNaM opposite dTPT3 template also revealed a unique conformation, termed the swing state. This swing state is distinct from the canonical pre-translocation state. We did not observe an edge-to-edge base pair of dTPT3:rNaM (Anti:Syn pair)[6], suggesting that such pairing is energetically unfavored in the post-chemistry state, presumably because Pol II is no longer in a closed state (in which the trigger loop is opened and pyrophosphate is released). Instead, we observed that while dTPT3 is located at the canonical +1 template position, the nucleobase and ribose of rNaM move down toward Rpb2 R1020 and R766 (away from −1 primer), and its hydrophobic nucleobase is stabilized by Van der Waals interactions (Fig. 5a). This movement resolves a potential steric clash between rNaM and dTPT3 (both in the anti-conformation). The swing state is also different from the frayed or backtracked state, which moves further away toward the second channel (Fig. 5b). Therefore, the swing state may represent an intermediate state between the canonical state, the pre-translocation state, and the canonical frayed state. We speculate this swing state may be poised for cross-strand intercalation and favors forward translocation and extension of UBPs.

It is noteworthy that a recent study reported that rNaMTP is the most optimal ribonucleotide among nine rNaMTP derivatives for retrieval of information stored by dTPT3[3]. Therefore, dTPT3–rNaMTP represents the most promising UBP in terms of transcription and translation efficiency. Screening of four rTPT3 derivatives resulted in the discovery of rTAT1, which produces significantly more unnatural protein than tRNAs with rTPT3. Therefore, our study can pave the way for a better understanding of how these new generations of UBPs are processed by Pol II and other cellular RNA polymerases.

By combining biochemical analysis, MD simulation and X-ray crystallography, we provide mechanistic insight into UBP transcription by eukaryotic RNA polymerase II. Intriguingly, we showed that Pol II can process UBPs without significant disruption of fidelity. Moreover, our observation of asymmetric fidelity of UBP processing attests to the intrinsic differences between replication and transcription, as DNA polymerases do not show this asymmetric fidelity during replication[13]. Our structural study and MD simulation provide a molecular basis for how UBPs are recognized and accommodated by eukaryotic RNA Pol II. Furthermore, we revealed for the first time a unique swing state for RNA Pol II, which is found with a dTPT3-rNaM unnatural base pair. Our results provide key insights into the molecular basis of UBP transcription. Our demonstration that eukaryotic RNA polymerases are able to transcribe the corresponding DNA, along with the demonstrated ability of eukaryotic ribosomes to recognize the UBPs during translation[48], is expected to facilitate progress towards the creation of eukaryotic semi-synthetic organisms. The ultimate goal is to develop both prokaryotic and eukaryotic SSOs capable of processing an expanded genetic code with its native cellular machinery, including its multi-subunit RNA polymerases, which are evolutionarily distinct from single-subunit T7 RNAP. Therefore, this study sheds light on one of the key steps to be understood in this long journey: transcription of UBPs by a multi-subunit RNA polymerase.

Methods

Oligonucleotide synthesis

The sequence of the modified template strand is 5'-CTACCGATAAGCAGACGXTCCTCTCGATG-3', where X is either dTPT3 or dNaM. The modified oligonucleotide was synthesized on a 1 μmol scale using an Expedite 8909 gene synthesizer, succinyl linked LCAA-CPG (long chain alkyl amine-controlled pore glass) columns with a pore size of 500 Å, and standard protocols for the incorporation of dABz, dCBz, dGiBu and dT DNA phosphoramidites. The following hand-coupling conditions were used for the incorporation of monomer dTPT3 (15 min; Tetrazole in CH3CN; >95%). Modified phosphoramidites were used at 50-fold molar excess and 0.05 M concentration in anhydrous CH_3CN. Cleavage from solid support and removal of protecting groups was accomplished using ~30% aq. ammonia (55 °C, 16 h). Purification of the DMT-on crude oligonucleotide was performed using a PolyPak cartridge using manufacturer instructions. Purity was verified by anion-exchange HPLC on a DNAPac PA-200 column (>95%).

Other canonical DNA templates and non-template oligonucleotides were purchased from IDT. RNA primers were purchased from TriLink Biotechnologies and radiolabeled using (γ-^{32}P) ATP and T4 Polynucleotide Kinase (NEB). The sequence of the non-template DNA strand used in transcription assay is 5'-CTGCTTATCGGTAG-3'. The 10mer RNA primer used is 5'-AUCGAGAGGA-3' and the 8mer primer used is 5'-AUCGAGAG-3'.

In vitro Pol II transcription assays

Twelve-subunit wild-type RNA Pol II used for transcription was purified from *Saccharomyces cerevisiae* as described[28,31]. The Pol II elongation complexes for transcription assays were assembled with RNA/DNA scaffold using established methods. Briefly, ^{32}P-labeled RNA was annealed with tsDNA and ntsDNA with a 1:1.5:2 molar ratio to form the mini-scaffold in elongation buffer (20 mM Tris pH 7.5, 40 mM KCl, 5 mM DTT, and 5 mM $MgCl_2$). Prepared mini-scaffold was incubated with a 4-fold excess amount of Pol II at room temperature for 10 min to ensure the formation of a Pol II elongation complex. Final reaction concentrations after mixing were 25 nM mini-scaffold, 100 nM Pol II and 0.5 or 1 mM of NTP or rTPT3TP or rNaMTP in elongation buffer. Reactions were quenched at various time points by adding stop buffer which consists of 90% (v/v) formamide, 50 mM EDTA, 0.05% (w/v) xylene cyanol and 0.05% (w/v) bromophenol blue. For the TFIIS treatment, the elongation complex is mixed with 200 nM TFIIS. All samples were heated at 95 °C for 5 min and analyzed by denaturing urea/TBE PAGE, by using Quantity one and Image Lab (Bio-rad) . Nonlinear-regression data fitting was performed using Prism 6. The time dependence of product formation was fit to a one-phase association equation to determine the observed rate (k_{obs}). All the experiments were performed independently three times and shown as means with standard deviation error bars.

Crystallization and structure determination

Ten-subunit Pol II was crystallized as previously described[23,28], Briefly, tsDNA containing dTPT3 was annealed with RNA and ntsDNA, with a 1:2:2 molar ratio in elongation buffer. 12 μM of DNA/RNA hybrid was incubated with 3 μM of purified Pol II for 1 hour at 4 °C.

Excess scaffold was removed and the buffer was changed to 25 mM Tris pH 7.5, 20 mM NaCl, 5 mM DTT, 1 μM $Zn(OAc)_2$, 100 μM EDTA, by ultrafiltration. Using 6–8 mg/ml EC, crystallization trays were set up using a hanging drop vapor diffusion method with 390 mM ammonium phosphate pH 6.0, 5 mM DTT, 5 mM dioxane, and 9–13 % (w/v) PEG 6,000 at 22 °C. Before freezing, crystals were moved to cryo solution (100 mM MES pH 6.0, 350 mM NaCl, 5 mM DTT, 5 mM Dioxane, 16 % PEG 6,000, and 17 % PEG400) in a step-wise manner and incubated at 4 °C overnight. To obtain the substrate-bound Pol II-dTPT3 EC structure, 10 mM of rNaMTP and 10 mM $MgCl_2$ were added to the cryo solution and incubated overnight for the soaking experiment. The 3' end of the RNA primer was modified to prevent rNaMTP incorporation. To obtain the post-incorporation state of Pol II-dTPT3-rNaM EC structure, regular RNA primer was used in the scaffold to allow rNaM addition in crystal. X-ray datasets were collected at 100 K in BL12–2, Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory (using Blu-ice 5 and Web-ice) and BL 5.0.3 and 8.2.1, Advanced Light Source, Lawrence Berkeley National Laboratory (using Beamline Operating Software BOS/B3), respectively. For apo dTPT3 and dTPT3-rNaM (added) structure, dataset collected at 1 Å wavelength were processed by iMosflm[49]. For the dTPT3-rNaMTP complex structure, dataset collected at 0.97665 Å wavelength were processed by XDS[50]. Half correlation coefficient (CC1/2) higher than 0.3 was used to determine the high resolution cutoff[51]. The initial structure of apo dTPT3 was solved by using molecular replacement in Phenix, with non-damaged EC (with the omission of natural DNA/RNA scaffold) as a search model (PDB 2E2J)[31,52]. Other soaking structures were phased by molecular replacement by using apo dTPT3 structure as a search model. Final structures were obtained by several rounds of manual building and refinement using COOT and Phenix[52,53]. For the B-factor, we used group B-factor refinement for all structures. Comparison of group B-factor and individual B-factor refine show that using group B-factor results in better R-work/R-free values (lower R-free or smaller gap between R-work/free values). Estimated coordinate error for apo dTPT3, dTPT3-rNaMTP and dTPT3-rNaM structure was 0.49 Å, 0.59 Å, 0.67 Å, respectively. RMSD of search model vs apo dTPT3 was 0.49Å, search model vs dTPT3 + rNaMTP was 0.55Å, search model vs dTPT3 + incorporated rNaM was 0.69 Å, apo dTPT3 vs dTPT3 + rNaMTP was 0.50Å, apo dTPT3 vs dTPT3 + incorporated rNaM was 0.67Å, and dTPT3 + rNaMTP vs dTPT3 + incorporated rNaM was 0.56Å. Ramachandran outliers for each structure (dTPT3_apo, dTPT3_rNaMTP and dTPT3_rNaM) was 0.0, 0.2, 0.0 %, respectively. Data collection and refinement statistics are summarized at Supplementary Table 2. Structural figures are prepared by Pymol.

Assignment of partial charges of NTP and rNaMTP

The partial charges of ATP, GTP, UTP, and CTP were taken from published force field[54]. For rNaMTP, electrostatic potential calculation was performed by Gaussian 09[55]. For dTPT3, dNAM and rTPT3, the phosphate group is replaced by a hydrogen and 3' end is also capped with a hydrogen. Subsequently, geometric optimization and electrostatic potential calculation were performed by Gaussian 09[55]. The geometry optimizations were performed using with Becke, 3-parameter, Lee-Yang-Parr method and basis set 6-31g* (B3LYP/6-31G*). Electrostatic potential calculations were performed using Hartree-Fock method with basis set 6-31g* (HF/6-31g*). The partial charges were then fitted by RESP method

using resp module in AmberTools 13[35]. The partial charges were determined by a two-stage RESP fitting procedure. In the first stage, partial charges were fitted to the electrostatic potential while keeping the triphosphate and sugar ring the same as the published force field[54] for rNaMTP and the sugar ring the same as AMBER 99SB-ILDN force field for dTPT3, dNaM and rTPT3. At the second stage, fitting was done for methylene and methyl groups while maintaining the partial charges of all the other atoms.

MD simulations of Pol II elongation complexes

To obtain a starting model for simulation, a template strand DNA (chain T and residues 18 – 25) of apo dTPT3 structure was superposed to that of dC: GMPCPP structure (PDB 2E2J) by using LSQ superpose in COOT[56]. This gave us alignment of DNA/RNA hybrid, together with nearby Pol II residues. We then adopted GMPCPP to apo dTPT3 structure and replaced GMPCPP with each nucleoside triphosphate, by fixing the planarity of the base and position of the ribose moiety and phosphorus atom. For example, the first model prepared was dTPT3:GTP (anti), where we replaced GMPCPP with GTP and maintained planarity and other key atomic position. However, *anti*-conformation of purine base strongly collides with dTPT3. To avoid this, we rotated GTP base to adopt *syn*-conformation. For the same reason, bulky bases such as GTP, ATP and rNaMTP were modeled as *syn*-conformation. Other pyrimidine base models were prepared by replacing GMPCPP with UTP or CTP. Second magnesium ion was adopted from 2E2J after alignment. For dNaM starting model, dTPT3 in apo structure was replaced by dNaM, maintaining deoxyribose configuration and parallel base plane. Then ATP and rTPT3TP was modeled as syn-conformation as described above.

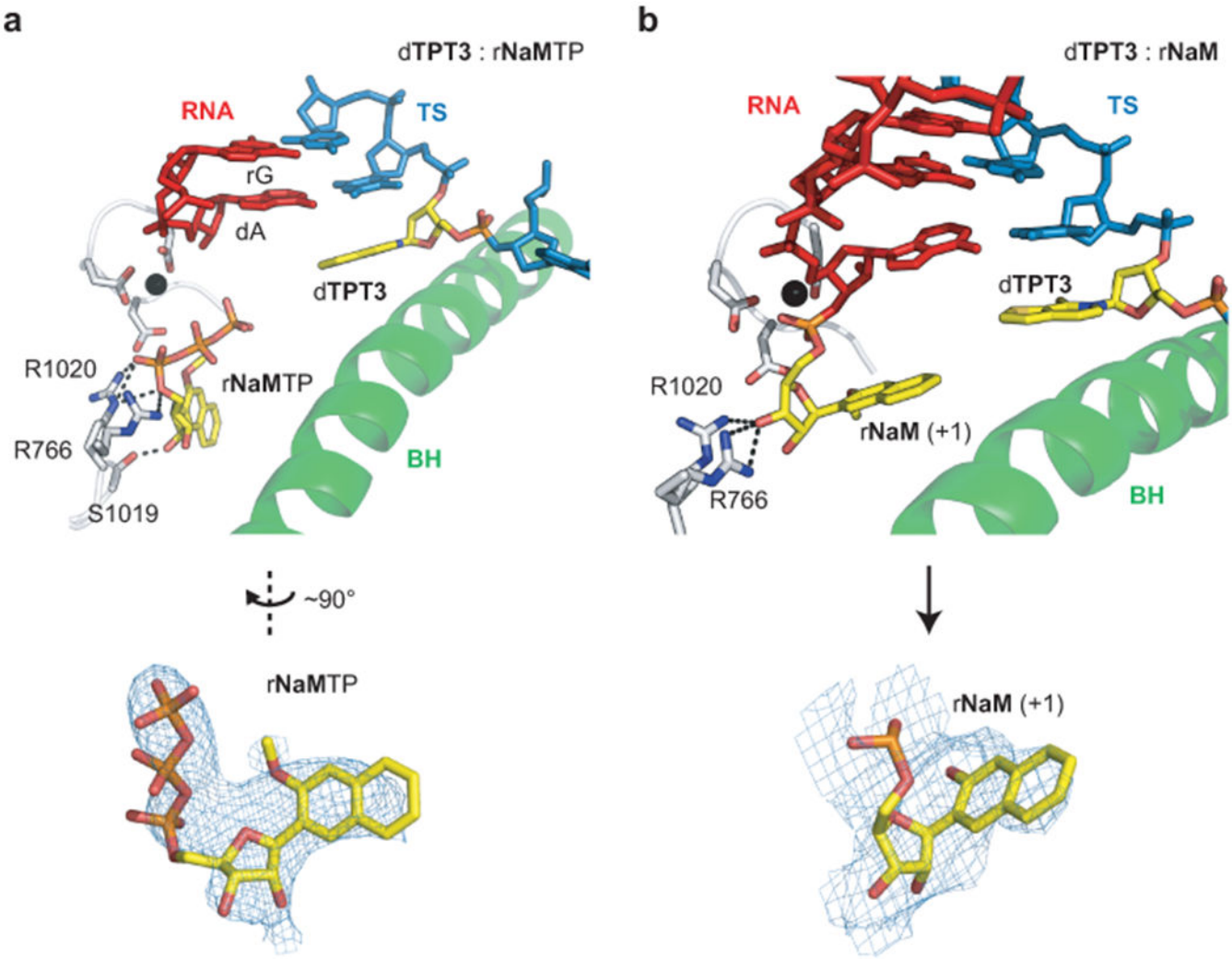
Using the model described previously as initial structures, we performed MD simulations for the elongation complex with NTPs (either rNaMTP, ATP, GTP, CTP, or UTP), and dTPT3 as the template DNA in the active site. We also performed two additional types of simulation of elongation complexes with dNaM as template DNA and either rTPT3TP or ATP in the active site. In all simulations, we modeled both Mg^{2+} ions A & B in the active site. There are seven types of Pol II elongation complex with varying pairs of template DNA and NTP substrate, i.e. dTPT3-rNaMTP, dTPT3-ATP, dTPT3-GTP, dTPT3-UTP, dTPT3-CTP, dNaM-rTPT3TP, and dNaM-ATP. For each type of simulation, the Pol II elongation complex is put in a dodecahedron box with 10 Å space between the complex and the edge of the box and solvated with TIP3P water molecules[32]. Then, Na^+ and Cl^- ions were added to neutralize the system and further reach 0.15 mol/liter salt concentration (see Supplementary Table 3 for the size and composition of each system). All MD simulations were performed using GROMACS 5.0.4 simulation package and AMBER 99SB-ILDN force field along with the partial charges obtained by RESP method for NTPs and non-conventional DNA[33,34]. Long-range electrostatic interactions were treated by Particle-Mesh Ewald (PME) method and Van der Waals short-range interactions were calculated using a cutoff of 12 Å.

For each type of simulation, the following steps of energy minimization, equilibration, and production MD simulations were performed. During energy minimization, to ensure correct coordination number of both Mg^{2+} ion A & Mg^{2+} ion B, harmonic restraints were added ($k_b = 1 \times 10^4$ kJ $mol^{-1}nm^{-2}$) between Mg^{2+} ions and their coordinating atoms. After energy minimization, LINCS algorithm was applied to constrain all bonds in the protein and water

during all of the following steps[57]. Next, the following equilibration steps were performed: i) NVT equilibration for 1 ns with position restraints on all heavy atoms (force constant = 1000 kJ mol^{-1} nm^{-2}), ii) NPT equilibration for 1 ns with position restraints on all heavy atoms (force constant = 1000 kJ mol^{-1} nm^{-2}), iii) NPT equilibration for 200 ps with harmonic restraints between the same atom pairs as the energy minimization step (k_b = 1 × 10^4 kJ $mol^{-1}nm^{-2}$), iv) NPT equilibration for 200 ps NPT equilibration harmonic restraints between the same atom pairs as the energy minimization step (k_b = 2 × 10^3 kJ $mol^{-1}nm^{-2}$), and v) NPT equilibration for 200 ps with position restraints only on the Mg^{2+} ions in the active sites. During all the equilibration steps, Berendsen thermostat was used to maintain the temperature at 310 K with coupling constant = 0.1 ps. During all NPT equilibration steps, Berendsen barostat was used with reference pressure of 1 bar and coupling constant = 0.5 ps[58]. Finally, for each type of simulation, 16 independent production MD simulations were performed (without any restraints) for 50 ns in NPT condition at 310 K with V-rescale thermostat, Parrinello-Rahman barostat and temperature annealing from 10 to 310 K in the first 2 ns[59,60]. As a result, we accumulated 800 ns simulation time for each system. For all the subsequent analysis, the first 10 ns simulation of each production MD simulation was excluded.

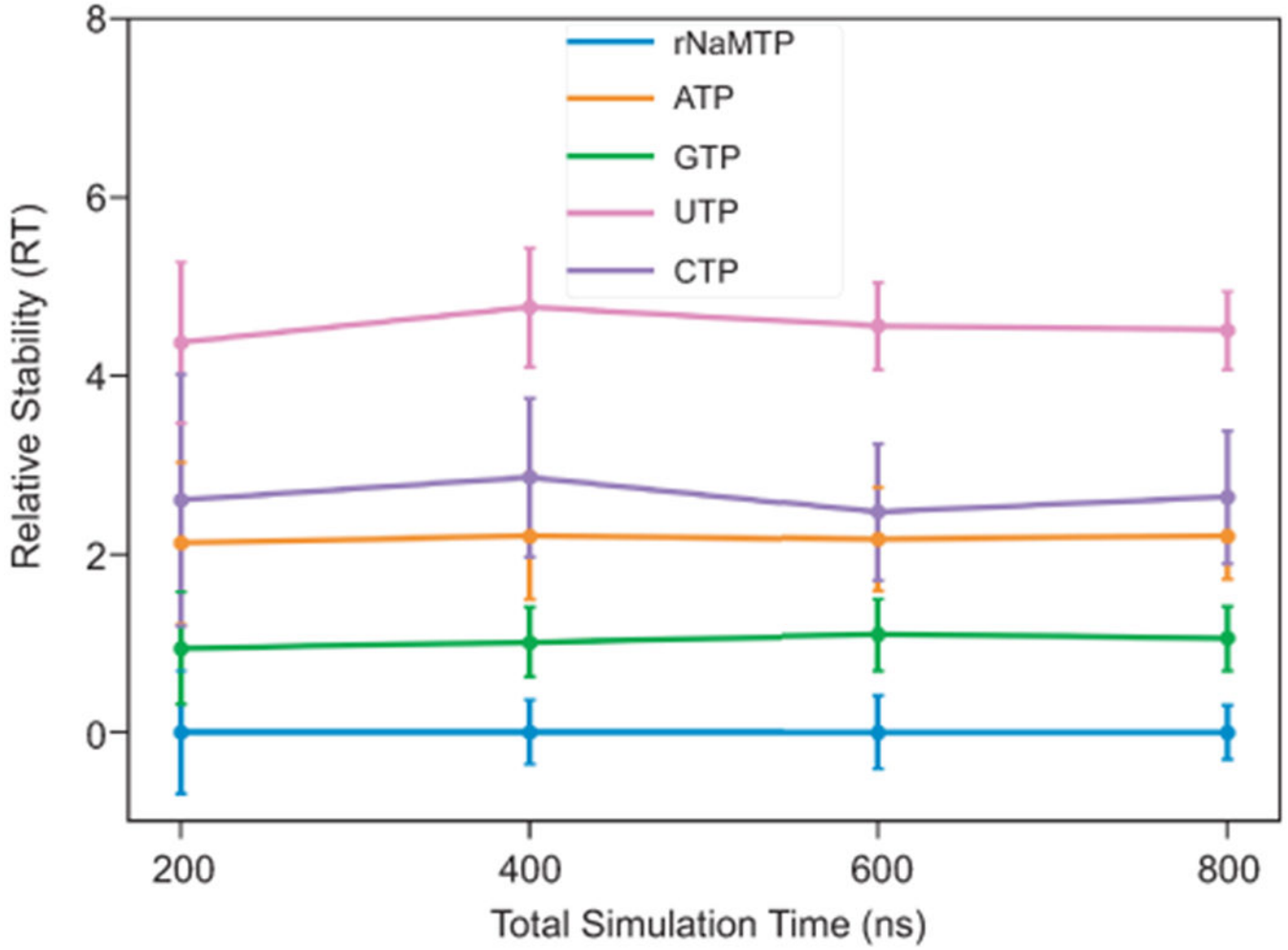
We defined good activation geometries which satisfy both criteria: O3’ – Pα distance within a range of 3 to 3.5 Å that allows chemistry (nucleotide incorporation); base pair distances calculated from the center mass of two nucleobases (7 to 9 Å for dTPT3 – rNaMTP or GTP or ATP, around 6 to 8 Å for dTPT3 – CTP or UTP and dNaM – rTPT3TP or ATP, respectively). We computed the percentage of simulation frames with catalytically active conformation (Fig. 4f and Extended Data Fig. 3).

Extended Data



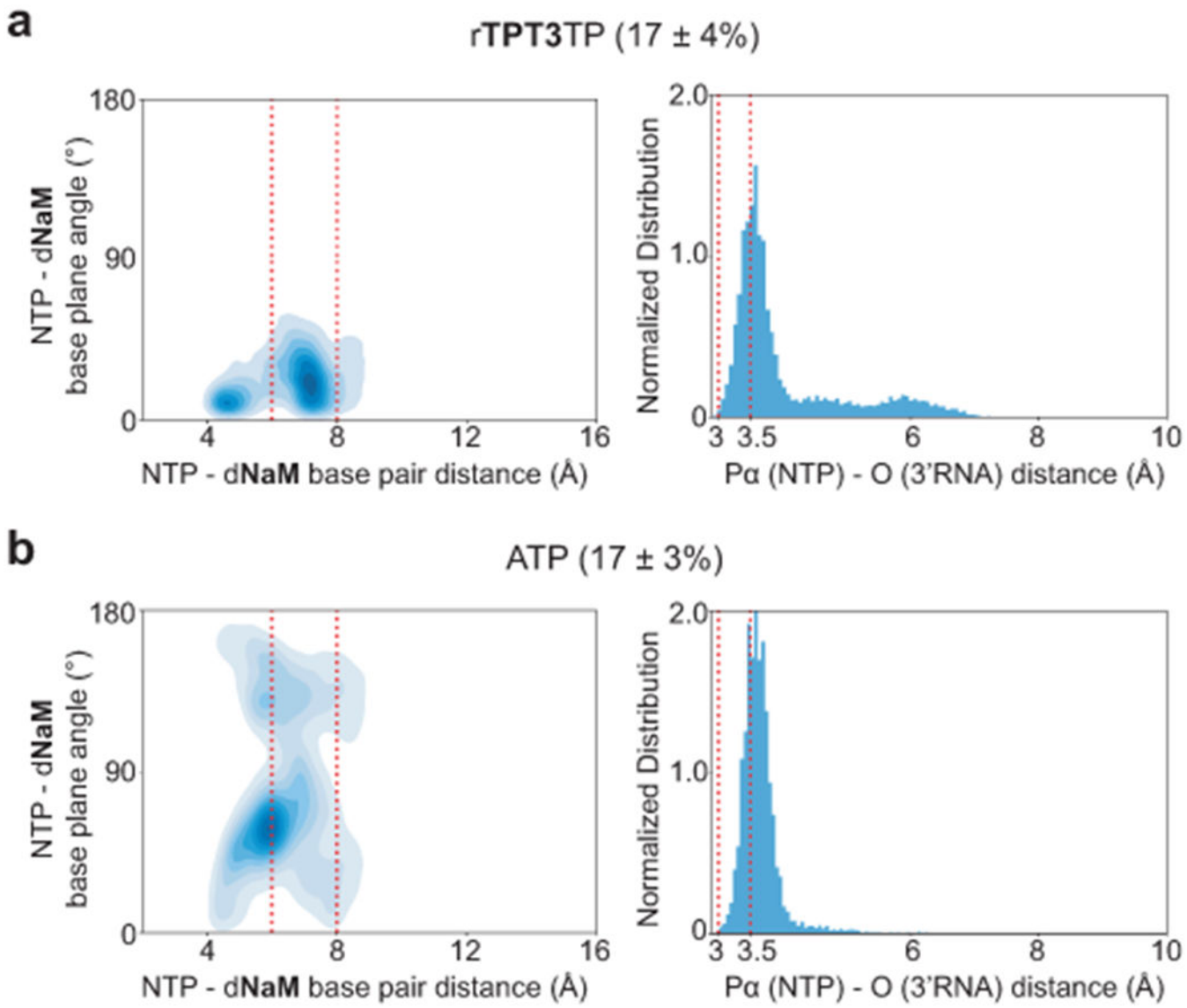
Extended Data Figure 1. Electron density map of rNaMTP or rNaM.

(a) Unbiased 2Fo-Fc omit electron density map of rNaMTP is contoured at 1.2 σ. (b) Unbiased 2Fo-Fc omit electron density map of rNaM is contoured at 1.2 σ.



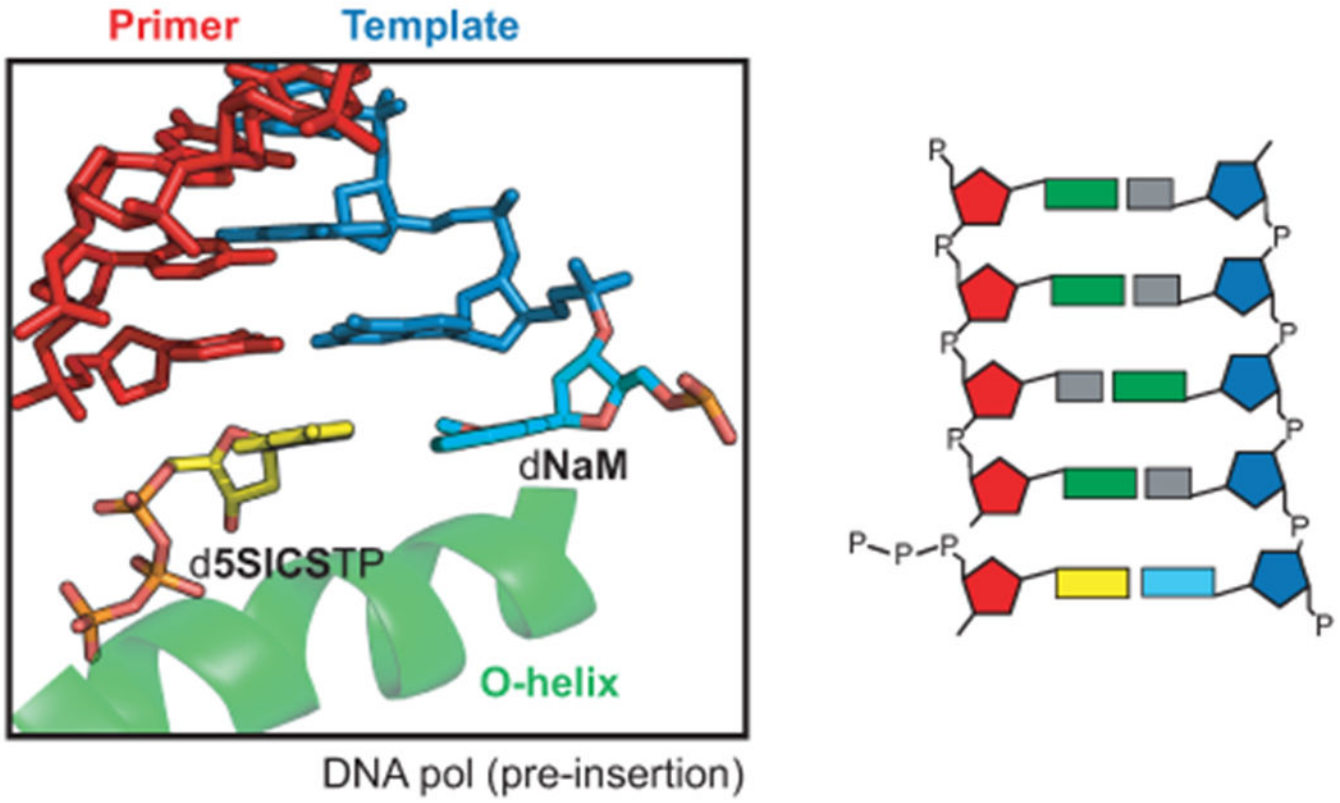
Extended Data Figure 2. The relative stability of NTPs to maintain good activation geometry is plotted as a function of total simulation time.

The relative stability is defined as the stability of NTPs in comparison with rNaMTP to maintain good activation geometry, $-\ln(p_{NTP}/p_{rNaMTP})$, in energy unit (RT). p_{NTP} and p_{rNaMTP} are the percentage of frames with good activation in NTP and rNaMTP simulations, respectively. The criteria for good activation geometry are 3.0 Å ≤ distance between O3' - Pα ≤ 3.5 Å and 7.0 Å ≤ base pair distance (rNaMTP, ATP or GTP) ≤ 9.0 Å, 6.0 Å ≤ base pair distance (CTP, UTP) ≤ 8.0 Å. The plot shows that the simulation has converged as the order of stability among NTPs remains the same regardless of the simulation time. Importantly, rNaMTP indeed is the most stable substrate when dTPT3 is the template DNA. The data are shown as mean values ± standard deviation, which were calculated by bootstrapping of N independent production MD simulations (N=4, 8, 12, 16 for data at the time point of 200, 400, 600, 800 ns in the x-axis, respectively).



Extended Data Figure 3. MD simulation of ATP and rTPT3TP at A site across dNaM (with both Mg2+ ion A & Mg2+ ion B).

(a and b) Left panel: two dimensional heatmap plot of the base pairing geometry. Base pair distance is the distance between center of mass of dNaM and NTPs. We observed significant localization of simulation frames in the dNaM-rTPT3TP pair, while ATP was highly dispersed both in distance and angle. Right panel: Distance of nucleophilic attack. Distribution of simulation frames sorted by the distance between Pα of incoming NTP and O3′ of terminal RNA is plotted. Good activation geometry (3.0 Å ≤ distance between O3′ - Pα ≤ 3.5 Å and 6.0 Å ≤ base pair distance ≤ 8.0 Å) is indicated with red dotted lines. Percentage of simulation frames with catalytically active conformation was shown as mean values ± standard deviation, which were calculated by bootstrapping of N independent production MD simulations (N=16).



Extended Data Figure 4.

UBP structure from DNA polymerase (dNaM-d5SICSTP, PDB 3SV3) shows co-planar edge-to-edge configuration.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the National Institutes of Health (R01 GM102362 to D.W, GM118178 to F.E.R. and GM128376 to R.J.K.). R.K. acknowledges NASA Exobiology (NNX14AP59G). F.E.R acknowledges support from Synthorx, a Sanofi company. X.H acknowledges the support from the Padma Harilela Endowment fund.

Data Availability

Crystal structure coordinates of apo dTPT3, dTPT3_rNaMTP and dTPT3_rNaM Pol II complexes are deposited in the Protein Data Bank database (PDB, https://www.rcsb.org) with accession numbers 7KED, 7KEE and 7KEF, respectively. Source data are provided with this paper.

References

1. Malyshev DA & Romesberg FE The expanded genetic alphabet. Angew. Chem. Int. Ed. Engl 54, 11930–44 (2015). [PubMed: 26304162]

2. Seo YJ, Matsuda S & Romesberg FE Transcription of an expanded genetic alphabet. J. Am. Chem. Soc 131, 5046–7 (2009). [PubMed: 19351201]
3. Feldman AW et al. Optimization of Replication, Transcription, and Translation in a Semi-Synthetic Organism. J. Am. Chem. Soc 141, 10644–10653 (2019). [PubMed: 31241334]
4. Piccirilli JA, Krauch T, Moroney SE & Benner SA Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. Nature 343, 33–7 (1990). [PubMed: 1688644]
5. Ohtsuki T et al. Unnatural base pairs for specific transcription. Proc. Natl. Acad. Sci. USA 98, 4922–5 (2001). [PubMed: 11320242]
6. Betz K et al. KlenTaq polymerase replicates unnatural base pairs by inducing a Watson-Crick geometry. Nat. Chem. Biol 8, 612–4 (2012). [PubMed: 22660438]
7. Malyshev DA et al. A Semi-Synthetic Organism with an Expanded Genetic Alphabet. Nature 509, 385–8 (2014). [PubMed: 24805238]
8. Zhang Y et al. A Semisynthetic Organism Engineered for the Stable Expansion of the Genetic Alphabet. Proc. Natl. Acad. Sci. USA 114, 1317–1322 (2017). [PubMed: 28115716]
9. Zhang Y et al. A Semi-Synthetic Organism that Stores and Retrieves Increased Genetic Information. Nature 551, 644–647 (2017). [PubMed: 29189780]
10. Matsuda S et al. Efforts toward expansion of the genetic alphabet: structure and replication of unnatural base pairs. J. Am. Chem. Soc 129, 10466–73 (2007). [PubMed: 17685517]
11. Malyshev DA et al. Solution structure, mechanism of replication, and optimization of an unnatural base pair. Chemistry 16, 12650–9 (2010). [PubMed: 20859962]
12. Betz K et al. Structural insights into DNA replication without hydrogen bonds. J. Am. Chem. Soc 135, 18637–43 (2013). [PubMed: 24283923]
13. Li L et al. Natural-like replication of an unnatural base pair for the expansion of the genetic alphabet and biotechnology applications. J. Am. Chem. Soc 136, 826–9 (2014). [PubMed: 24152106]
14. Zhou AX, Dong X & Romesberg FE Transcription and Reverse Transcription of an Expanded Genetic Alphabet In Vitro and in a Semisynthetic Organism. J. Am. Chem. Soc 142, 19029–19032 (2020). [PubMed: 33118814]
15. Seo YJ, Malyshev DA, Lavergne T, Ordoukhanian P & Romesberg FE Site-specific labeling of DNA and RNA using an efficiently replicated and transcribed class of unnatural base pairs. J. Am. Chem. Soc 133, 19878–88 (2011). [PubMed: 21981600]
16. Liu X, Bushnell DA & Kornberg RD RNA polymerase II transcription: structure and mechanism. Biochim. Biophys. Acta 1829, 2–8 (2013). [PubMed: 23000482]
17. Werner F & Grohmann D Evolution of multisubunit RNA polymerases in the three domains of life. Nat. Rev. Microbiol 9, 85–98 (2011). [PubMed: 21233849]
18. Gout JF et al. The landscape of transcription errors in eukaryotic cells. Sci. Adv 3, e1701484 (2017). [PubMed: 29062891]
19. Brueckner F, Hennecke U, Carell T & Cramer P CPD damage recognition by transcribing RNA polymerase II. Science 315, 859–62 (2007). [PubMed: 17290000]
20. Damsma GE, Alt A, Brueckner F, Carell T & Cramer P Mechanism of transcriptional stalling at cisplatin-damaged DNA. Nat. Struct. Mol. Biol 14, 1127–33 (2007). [PubMed: 17994106]
21. Walmacq C et al. Mechanism of translesion transcription by RNA polymerase II and its role in cellular resistance to DNA damage. Mol. Cell 46, 18–29 (2012). [PubMed: 22405652]
22. Walmacq C et al. Mechanism of RNA polymerase II bypass of oxidative cyclopurine DNA lesions. Proc. Natl. Acad. Sci. USA 112, E410–9 (2015). [PubMed: 25605892]
23. Wang W, Walmacq C, Chong J, Kashlev M & Wang D Structural basis of transcriptional stalling and bypass of abasic DNA lesion by RNA polymerase II. Proc. Natl. Acad. Sci. USA 115, E2538–E2545 (2018). [PubMed: 29487211]
24. Oh J et al. RNA polymerase II stalls on oxidative DNA damage via a torsion-latch mechanism involving lone pair-pi and CH-pi interactions. Proc. Natl. Acad. Sci. USA 117, 9338–9348 (2020). [PubMed: 32284409]
25. Sydow JF et al. Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. Mol. Cell 34, 710–21 (2009). [PubMed: 19560423]

26. Wang D et al. Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. Science 324, 1203–6 (2009). [PubMed: 19478184]
27. Gnatt AL, Cramer P, Fu J, Bushnell DA & Kornberg RD Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science 292, 1876–82 (2001). [PubMed: 11313499]
28. Oh J, Xu J, Chong J & Wang D Structural and biochemical analysis of DNA lesion-induced RNA polymerase II arrest. Methods 159-160, 29–34 (2019). [PubMed: 30797902]
29. Batada NN, Westover KD, Bushnell DA, Levitt M & Kornberg RD Diffusion of nucleoside triphosphates and role of the entry site to the RNA polymerase II active center. Proc. Natl. Acad. Sci. USA 101, 17361–4 (2004). [PubMed: 15574497]
30. Westover KD, Bushnell DA & Kornberg RD Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. Cell 119, 481–9 (2004). [PubMed: 15537538]
31. Wang D, Bushnell DA, Westover KD, Kaplan CD & Kornberg RD Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. Cell 127, 941–54 (2006). [PubMed: 17129781]
32. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW & Klein ML Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys 79, 926–935 (1983).
33. Abraham MJ et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1, 19–25 (2015).
34. Lindorff-Larsen K et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78, 1950–8 (2010). [PubMed: 20408171]
35. Case DA, D. TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts B, Hayik S, Roitberg A, Seabra G, Swails J, Goötz AW, Kolossvaá;ry I, Wong KF, Paesani F, Vanicek J, Wolf RM, Liu J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh M-J, Cui G, Roe DR, Mathews DH, Seetin MG, Salomon-Ferrer R, Sagui C, Babin V, Luchko T, Gusarov S, Kovalenko A, and Kollman PA. Amber 13. University of California, San Francisco. (2012).
36. Carvalho AT, Fernandes PA & Ramos MJ The Catalytic Mechanism of RNA Polymerase II. J. Chem. Theory Comput 7, 1177–88 (2011). [PubMed: 26606364]
37. Wang B, Opron K, Burton ZF, Cukier RI & Feig M Five checkpoints maintaining the fidelity of transcription by RNA polymerases in structural and energetic details. Nucleic Acids Res 43, 1133–46 (2015). [PubMed: 25550432]
38. Svetlov V & Nudler E Basic mechanism of transcription by RNA polymerase II. Biochim. Biophys. Acta 1829, 20–8 (2013). [PubMed: 22982365]
39. Yang W, Lee JY & Nowotny M Making and breaking nucleic acids: two-Mg2+-ion catalysis and substrate specificity. Mol. Cell 22, 5–13 (2006). [PubMed: 16600865]
40. Da LT, Wang D & Huang X Dynamics of pyrophosphate ion release and its coupled trigger loop motion from closed to open state in RNA polymerase II. J. Am. Chem. Soc 134, 2399–406 (2012). [PubMed: 22206270]
41. Cheung AC & Cramer P Structural basis of RNA polymerase II backtracking, arrest and reactivation. Nature 471, 249–53 (2011). [PubMed: 21346759]
42. Yin YW & Steitz TA The structural mechanism of translocation and helicase activity in T7 RNA polymerase. Cell 116, 393–404 (2004). [PubMed: 15016374]
43. Temiakov D et al. Structural basis for substrate selection by t7 RNA polymerase. Cell 116, 381–91 (2004). [PubMed: 15016373]
44. Huang J, Brieba LG & Sousa R Misincorporation by wild-type and mutant T7 RNA polymerases: identification of interactions that reduce misincorporation rates by stabilizing the catalytically incompetent open conformation. Biochemistry 39, 11571–80 (2000). [PubMed: 10995224]
45. Xu L et al. RNA polymerase II transcriptional fidelity control and its functional interplay with DNA modifications. Crit. Rev. Biochem. Mol. Biol 50, 503–19 (2015). [PubMed: 26392149]
46. Silva DA et al. Millisecond dynamics of RNA polymerase II translocation at atomic resolution. Proc. Natl. Acad. Sci. USA 111, 7665–70 (2014). [PubMed: 24753580]

47. Huang X et al. RNA polymerase II trigger loop residues stabilize and position the incoming nucleotide triphosphate in transcription. Proc. Natl. Acad. Sci. USA 107, 15745–50 (2010). [PubMed: 20798057]

48. Zhou AX, Sheng K, Feldman AW & Romesberg FE Progress toward Eukaryotic Semisynthetic Organisms: Translation of Unnatural Codons. J. Am. Chem. Soc 141, 20166–20170 (2019). [PubMed: 31841336]

Methods-only references

49. Battye TG, Kontogiannis L, Johnson O, Powell HR & Leslie AG iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. Acta Crystallogr. D Biol. Crystallogr 67, 271–81 (2011). [PubMed: 21460445]

50. Kabsch W Xds. Acta Crystallogr. D Biol. Crystallogr 66, 125–32 (2010). [PubMed: 20124692]

51. Evans PR & Murshudov GN How good are my data and what is the resolution? Acta Crystallogr. D Biol. Crystallogr 69, 1204–14 (2013). [PubMed: 23793146]

52. Adams PD et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr 66, 213–21 (2010). [PubMed: 20124702]

53. Emsley P, Lohkamp B, Scott WG & Cowtan K Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr 66, 486–501 (2010). [PubMed: 20383002]

54. Meagher KL, Redman LT & Carlson HA Development of polyphosphate parameters for use with the AMBER force field. J. Comput. Chem 24, 1016–25 (2003). [PubMed: 12759902]

55. Frisch MJ, T. GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Petersson GA, Nakatsuji H, Li X, Caricato M, Marenich A, Bloino J, Janesko BG, Gomperts R, Mennucci B, Hratchian HP, Ortiz JV, Izmaylov AF, Sonnenberg JL, Williams-Young D, Ding F, Lipparini F, Egidi F, Goings J, Peng B, Petrone A, Henderson T, Ranasinghe D, Zakrzewski VG, Gao J, Rega N, Zheng G, Liang W, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Throssell K, Montgomery JA Jr., Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Keith T, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Millam JM, Klene M, Adamo C, Cammi R, Ochterski JW, Martin RL, Morokuma K, Farkas O, Foresman JB, and Fox DJ. Gaussian 09, Revision A.02. Gaussian, Inc., Wallingford CT (2016).

56. Emsley P & Cowtan K Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr 60, 2126–32 (2004). [PubMed: 15572765]

57. Hess B, Bekker H, Berendsen HJC & Fraaije JGEM LINCS: A linear constraint solver for molecular simulations. J. Comput. Chem 18, 1463–1472 (1997).

58. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A & Haak JR Molecular-Dynamics with Coupling to an External Bath. J. Chem. Phys 81, 3684–3690 (1984).

59. Parrinello M & Rahman A Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method. J. Appl. Phys 52, 7182–7190 (1981).

60. Bussi G, Donadio D & Parrinello M Canonical sampling through velocity rescaling. J. Chem. Phys 126, 014101 (2007). [PubMed: 17212484]

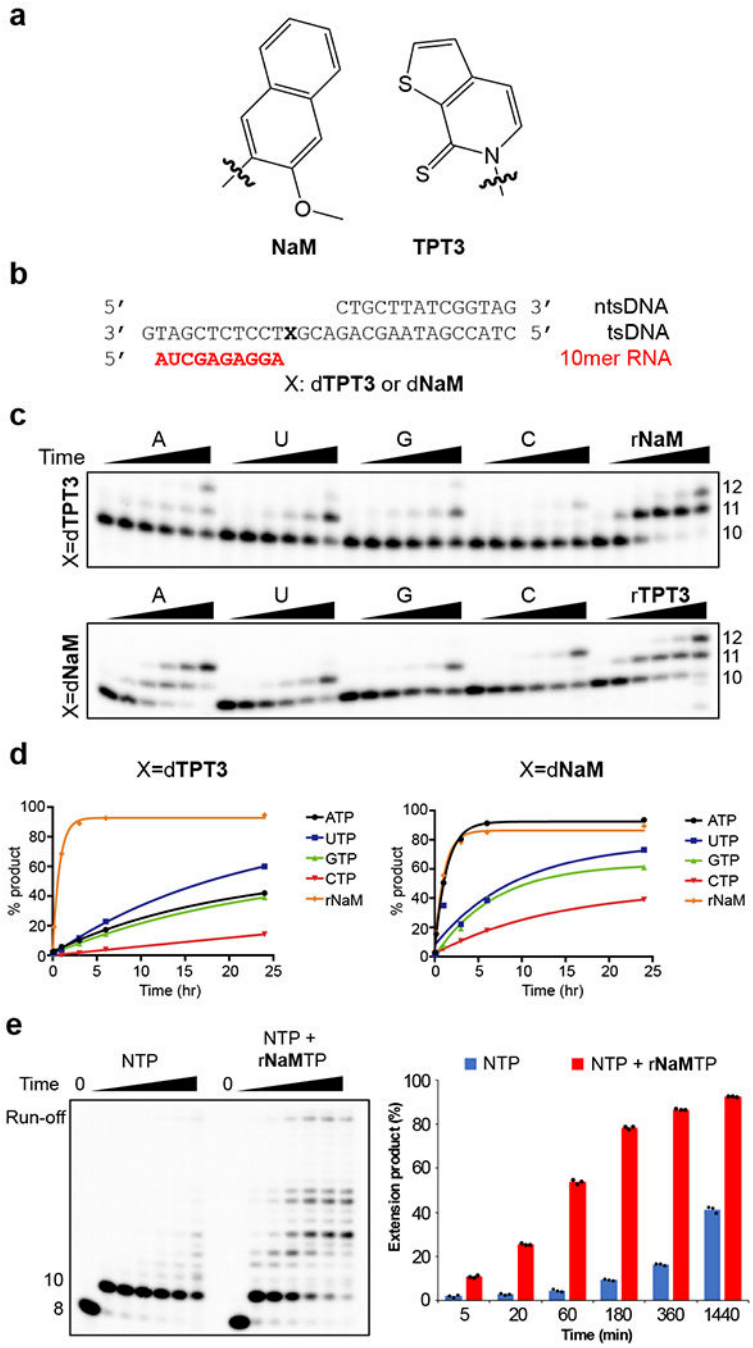


Figure 1.

Transcription assay using dTPT3 or dNaM template. (a) Chemical structure of NaM-TPT3 unnatural base pair (sugar and phosphate omitted for clarity). (b) Scaffold used in assays. X represents dTPT3 or dNaM in the template strand. 10mer RNA is colored as red. For Fig. 1e, 8mer RNA 5' AUCGAGAG was used. (c) Individual NTP incorporation into dTPT3 or dNaM template. The time points for assays from left to right were 1 min, 10 min, 1 hr, 3 hr, 6 hr and 24 hr. The positions of 10mer (start primer), 11mer (product) and 12mer (product) were labeled. (d) Quantitative analysis of single nucleotide incorporation into dTPT3 or

dNAM template. Pol II demonstrates selectivity for dTPT3 rNaMTP, while dNAM allows both ATP and rTPT3TP addition. (e) Elongation assay in the presence of NTP and rNaMTP. Fast accumulation of 10-nt RNA represents the pausing caused by dTPT3, which is subsequently resolved in the presence of rNaMTP. The time points were 5 min, 20 min, 1 h, 3 h, 6 h and 24 h. The positions of 8mer (start primer, time 0) and 10mer (major immediate extension product) were labeled at left. Ratio of UBP extension is shown in right panel. Mean values are shown as column bars with individual data points (black dots, n=3). Extension product (%) = summed intensity of bands above 10 bp / total intensity of each lane. For all transcription assays in Fig. 1, each experiment was repeated three times independently with similar results.

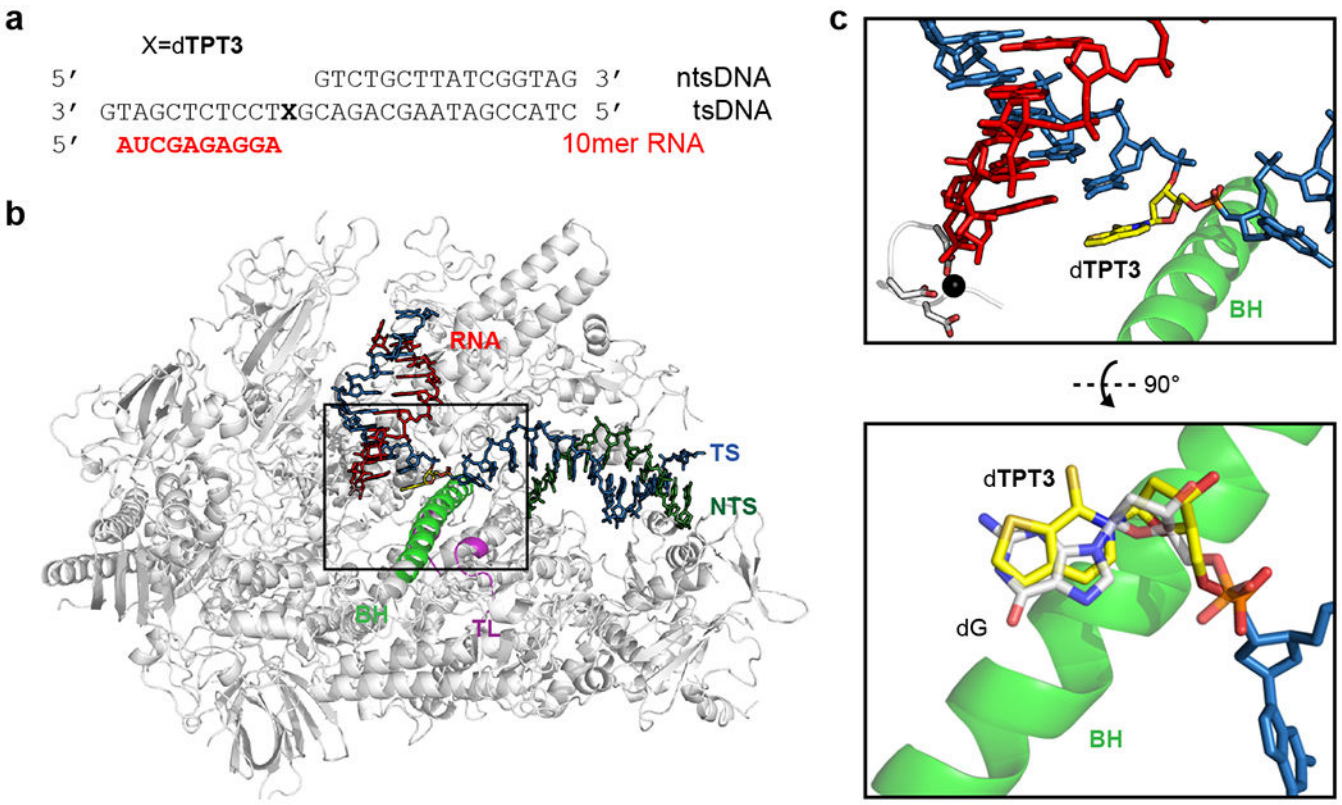


Figure 2.

Structure of Pol II-dTPT3 elongation complex. (a) Scaffold used for crystallization. (b) Overall structure of dTPT3 harboring Pol II. RNA, template strand DNA (TS) and non-template strand DNA (NTS) colored as red, blue, and deep green, respectively. Bridge helix (BH) and trigger loop (TL) are colored as green and purple, respectively. Other structure is shown as gray. For clarity, some residues of Rpb2 (20–770) are omitted. (c) Active site of Pol II-dTPT3 elongation complex indicates proper loading of dTPT3 to active site. Superposition of Pol II bridge helix between dTPT3 and non-damaged dG further supports dTPT3 is recognized as a normal template (PDB 6UQ2, bottom panel)[24]. dTPT3 and dG are colored as yellow and white, respectively.

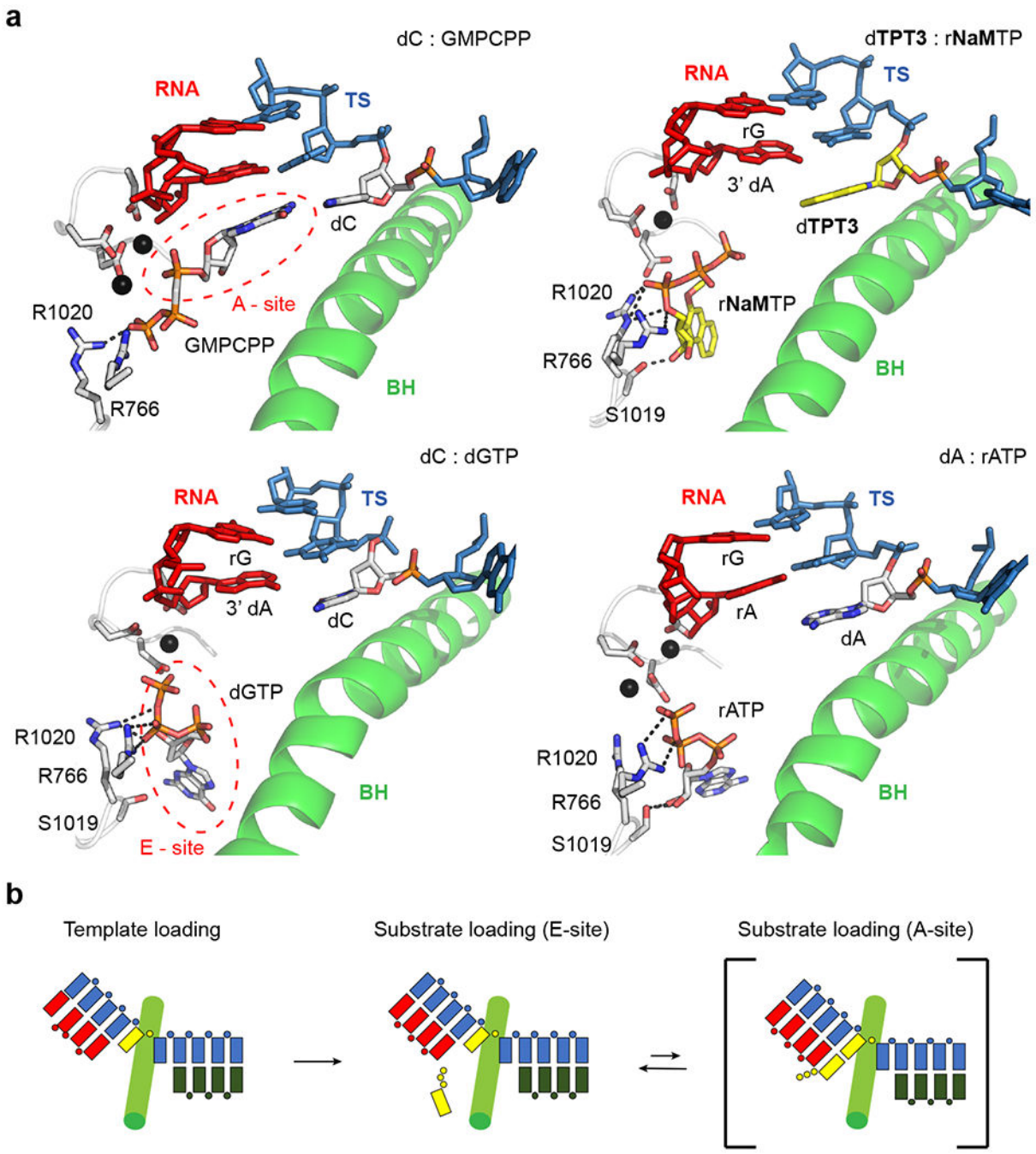


Figure 3.

Pol II-dTPT3-rNaMTP complex structure. (a) dC-GMPCPP canonical base pair, occupying A site (PDB 2E2J, left panel)[31]. dTPT3-rNaMTP complex structure (right panel). To obtain substrate bound structure of UBP Pol II, we used deoxyribose at the 3' end of an RNA primer (3' dA), which inhibits addition by incoming NTP. Binding environment is similar to that of canonical NTPs at the E site (PDB 2E2I and 1R9T, bottom panel), where two arginine residues (R766^{rpb2} and R1020^{rpb2}) interact with phosphate and S1019^{rpb2} interacts with hydroxyl group of ribose[30,31]. (b) Schematic representation of substrate binding. The

apo Pol II-dTPT3 structure represents template loading, and the Pol II-dTPT3-rNaMTP complex structure shows substrate binding at the E site. However, the rNaMTP substrate binding state at the A site was not observed (square brackets).

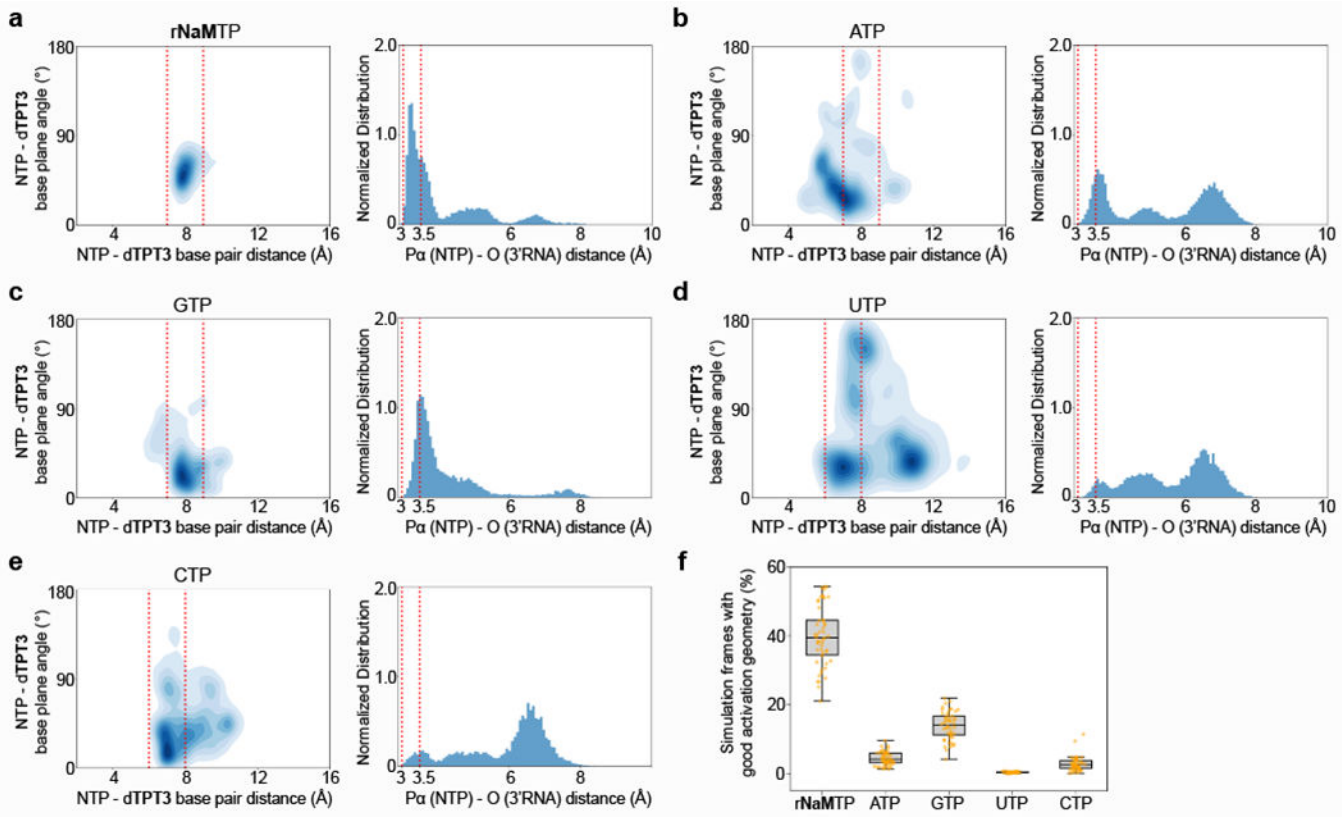


Figure 4.

MD simulation of individual NTPs and rNaMTP at the A site across the dTPT3 template (with both Mg^{2+} ion A & Mg^{2+} ion B). (a-e) Left plot panels: two dimensional heatmap plot of the base pairing geometry. Base pair distance is the distance between the center of mass of dTPT3 and NTPs. We observed significant localization of simulation frames in dTPT3-rNaMTP pair, while other NTPs are highly dispersed both in distance and angle. Right plot panels: Distance of nucleophilic attack. Distribution of simulation frames sorted by distance between Pα of incoming NTP and O3' of terminal RNA is plotted. Good activation geometry used in (f) is indicated with red dotted lines. (f) Percentage of simulation frames with catalytically active conformation, which is defined as 3.0 Å ≤ distance between O3' - Pα ≤ 3.5 Å and 7.0 Å ≤ base pair distance (rNaMTP, ATP or GTP) ≤ 9.0 Å, 6.0 Å ≤ base pair distance (CTP, UTP) ≤ 8.0 Å. box limits: interquartile range, whiskers: minimum to maximum, centre line: median, dots: individual data points are shown. The dots beyond the whiskers are considered outliers. The results were computed by bootstrapping of n independent production MD simulations (n = 16) 50 times.

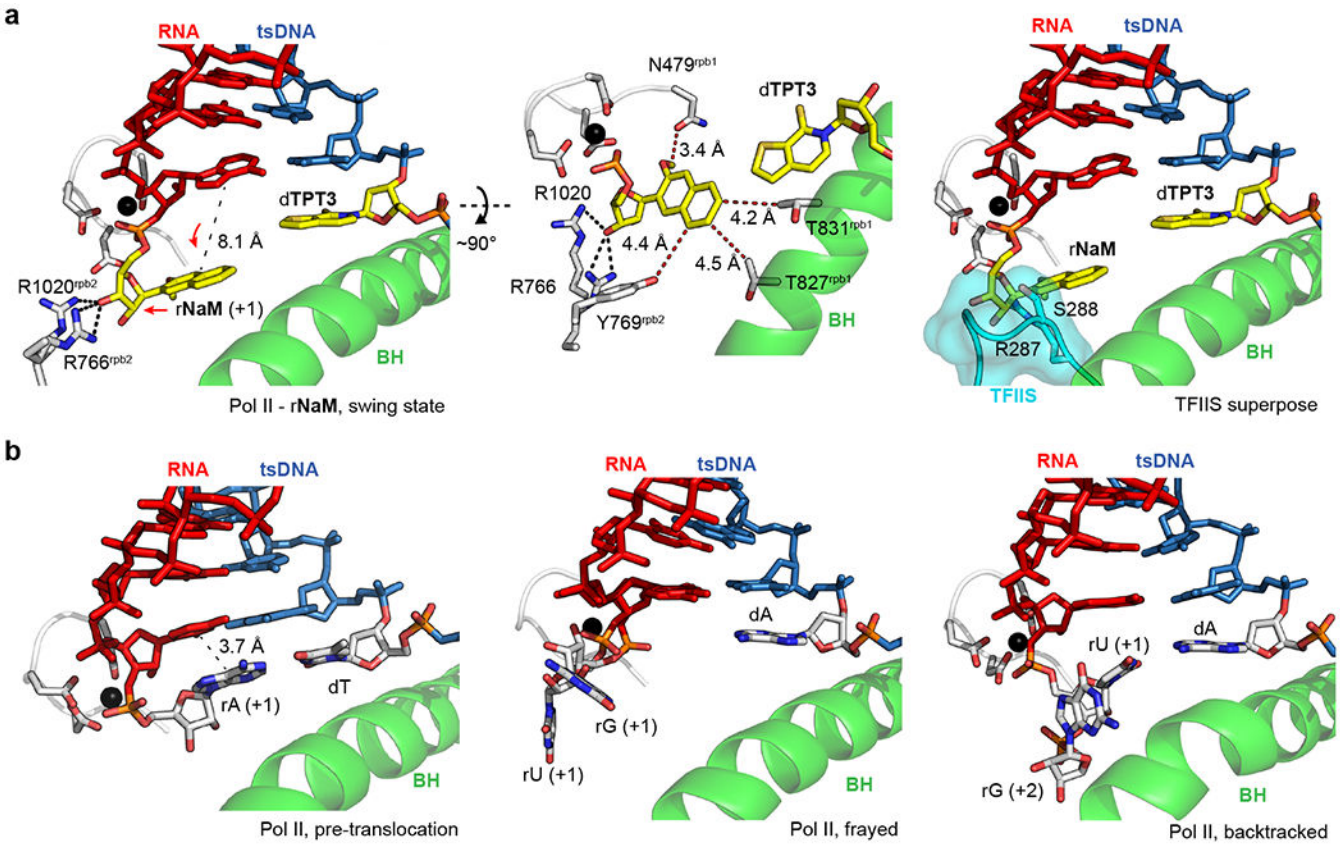


Figure 5.

Structural comparison of the dTPT3-rNaM swing state with other Pol II structures. (a) Active site of Pol II-dTPT3-rNaM. Both ribose and rNaM moiety retracted to an empty space between the A and E sites (left panel). rNaM in the swing state is stabilized by nearby elements (middle panel). Putative hydrogen bonding is indicated with bold black dash, while potential Van der Waals interactions are indicated with bold red dash. Superposition of Pol II-TFIIS structure to Pol II-rNaM indicates retracted rNaM that occupies binding position for domain III tip of TFIIS (PDB 3PO3, right panel, cyan color)[41]. (b) Pol II pre-translocation (PDB ID: 1I6H), frayed (PDB ID: 3HOZ, 3HOW) or backtracked (PDB ID:3GTG) states are aligned to the dTPT3 swing state [25,26]. The dTPT3-rNaM swing state is distinct from any other substrate binding site occupying structures.

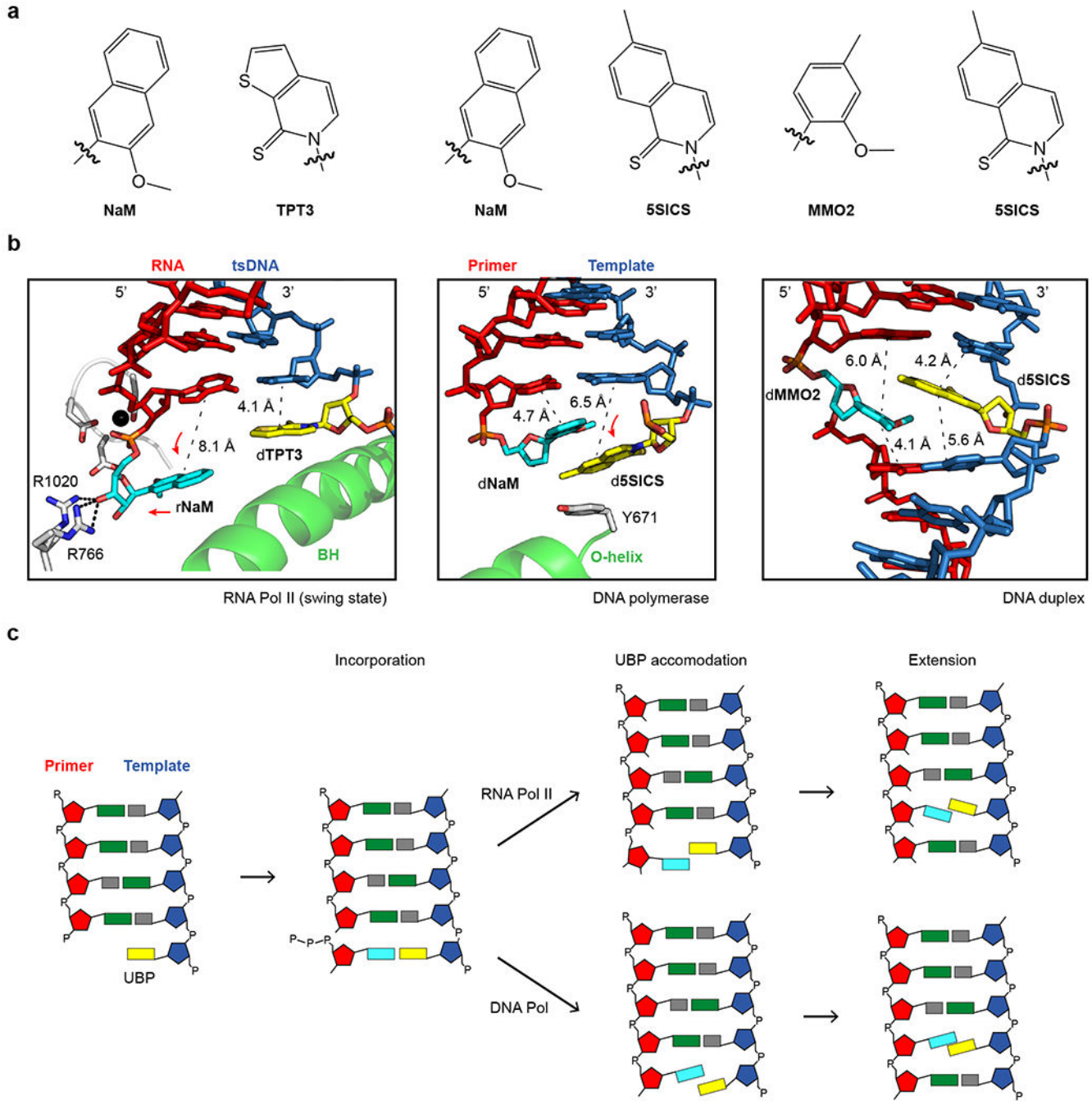


Figure 6.

Transcriptional processing of UBP by RNA polymerase II. (a) Chemical structure of NaM, TPT3, MMO2, and 5SICS, respectively. (b) UBP pairs in RNA Pol II (swing state), DNA polymerase, or duplex DNA. Bulky bridge helix of Pol II eliminates template strand UBP shift. UBP structure from DNA polymerase (d5SICS-dNaM, PDB ID: 4C8M) shows movement of template strand UBP (middle panel)[12]. UBP base pair in DNA duplex (d5SICS-dMMO2, PDB ID: 2LHO) showing cross-strand intercalation base pairing (right panel)[11]. (b) Proposed mechanism of replication/transcription processing of UBP. Our

structure of Pol II EC and previous studies with DNA polymerase suggests the UBP does not compromise the template loading step. Our MD simulation supports the highest preference for rNaMTP, with the criteria of good activation geometry. After incorporation, the dTPT3-rNaM swing state poses unique structural features in which both the ribose and base moiety of rNaM retract between the A site and the E site, termed the swing state. This swing state is poised for the formation of cross-strand intercalated state or edge-to-edge pair as Pol II translocates beyond the UBP site. DNA and RNA are colored as blue and red. Bridge helix and O-helix are colored as green. Purine and pyrimidine bases are colored as dark green and gray. TPT3 and 5SICS bases are colored as yellow. NaM and MMO2 bases are colored as cyan.