nature communications

Article

https://doi.org/10.1038/s41467-023-43735-9

A unified Watson-Crick geometry drives transcription of six-letter expanded DNA alphabets by *E. coli* RNA polymerase

Received: 25 May 2023

Accepted: 17 November 2023

Published online: 12 December 2023

Check for updates

Juntaek Oh[1,2,9], Zelin Shan[3,9], Shuichi Hoshika[4,9], Jun Xu[1], Jenny Chong[1], Steven A. Benner[4] ✉, Dmitry Lyumkis[3,5,6] ✉ & Dong Wang[1,7,8] ✉

Artificially Expanded Genetic Information Systems (AEGIS) add independently replicable unnatural nucleotide pairs to the natural G:C and A:T/U pairs found in native DNA, joining the unnatural pairs through alternative modes of hydrogen bonding. Whether and how AEGIS pairs are recognized and processed by multi-subunit cellular RNA polymerases (RNAPs) remains unknown. Here, we show that *E. coli* RNAP selectively recognizes unnatural nucleobases in a six-letter expanded genetic system. High-resolution cryo-EM structures of three RNAP elongation complexes containing template-substrate UBPs reveal the shared principles behind the recognition of AEGIS and natural base pairs. In these structures, RNAPs are captured in an active state, poised to perform the chemistry step. At this point, the unnatural base pair adopts a Watson-Crick geometry, and the trigger loop is folded into an active conformation, indicating that the mechanistic principles underlying recognition and incorporation of natural base pairs also apply to AEGIS unnatural base pairs. These data validate the design philosophy of AEGIS unnatural basepairs. Further, we provide structural evidence supporting a long-standing hypothesis that pair mismatch during transcription occurs via tautomerization. Together, our work highlights the importance of Watson-Crick complementarity underlying the design principles of AEGIS base pair recognition.

All of natural life on Earth utilizes a 4-letter molecular “alphabet” to store and retrieve genetic information. In 1953, Watson and Crick revealed the structural basis of how natural bases (adenine (A), thymine (T), cytosine (C), and guanine (G)) form base pairs in DNA duplexes[1,2]. The specificity of the canonical Watson-Crick base pairs (A:T and G:C) arises from their complementary sizes and shapes, and specific hydrogen-bonding interactions[2].

The development of synthetic nucleotides that form bio-orthogonal unnatural base pairs (UBPs) for expanding the genetic alphabet has been a long-standing goal of synthetic biology[3–6]. An expanded genetic alphabet can greatly enhance the chemical diversity of natural nucleic acids, and therefore introduce new functions and properties. Several laboratories have developed diverse UBPs based on the presence (or absence) of H-bonding and shape similarity[3–6].

[1]Division of Pharmaceutical Sciences, Skaggs School of Pharmacy & Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA. [2]Department of Pharmacy, College of Pharmacy, Kyung Hee University, Seoul 02447, Republic of Korea. [3]The Salk Institute for Biological Studies, La Jolla, CA 92037, USA. [4]Foundation for Applied Molecular Evolution, 13709 Progress Blvd Box 7, Alachua, FL 32615, USA. [5]Department of Integrative Structural and Computational Biology, The Scripps Research Institute 10550 N Torrey Pines Road, La Jolla, CA 92037, USA. [6]Graduate School of Biological Sciences, Section of Molecular Biology, University of California San Diego, La Jolla, CA 92093, USA. [7]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093, USA. [8]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA. [9]These authors contributed equally: Juntaek Oh, Zelin Shan, Shuichi Hoshika. ✉e-mail: sbenner@ffame.org; dlyumkis@salk.edu; dongwang@ucsd.edu

Early work led to the development of a series of UBPs that are joined by alternative H-bonding patterns. These are components of an Artificially Expanded Genetic Information System (AEGIS)[7–11], and currently include an expanded eight-letter alphabet with four natural and four unnatural bases (Supplementary Fig. 1 **B:S** and **P:Z**)[5]. Such alternative H-bonded UBPs can support effective DNA replication by the *Taq* DNA polymerase and RNA transcription by the single-subunit bacteriophage T7 RNA polymerase (T7 RNAP)[12,13].

However, how multi-subunit cellular RNAPs recognize AEGIS UBPs during transcription remains unknown. To employ AEGIS in cellular systems for synthetic biology applications, it is necessary to define the mechanistic principles underlying UBP recognition and transcription using multi-subunit cellular RNAPs. Therefore, the lack of mechanistic biochemical and structural insight for how AEGIS UBPs are recognized by cellular transcription machineries represents a major knowledge gap and critical barrier to further development and implementation of AEGIS UBPs for synthetic biology applications.

Here, we investigated how two components of a synthetic AEGIS base pair: **B** and **S**, can be selectively recognized and transcribed by *E. coli* RNAP in vitro. To further define the structural basis of base pair recognition during UBP transcription, we solved structures of three RNAP elongation complexes containing template–substrate UBPs in the active site. Despite the chemical and structural differences in the functional groups of the AEGIS **B**:**S** pair in comparison with natural base pairs, we revealed that cellular RNAP recognizes d**B**:**S**TP and d**S**:**B**TP AEGIS pairs in exactly the same manner as natural pairs. The AEGIS base pairs (d**S**:**B**TP and d**B**:**S**TP) adopt a Watson-Crick geometry within RNAP active site. Strikingly, unlike other UBPs[14,15], here we found that RNAP is captured in an active state with the ordered trigger loop in a closed conformation, which is poised for chemistry step. Furthermore, we also provide direct structural evidence to support a long-standing hypothesis that bases can adopt rare tautomeric forms that enable the formation of Watson–Crick-like (WC-like) mispairs and therefore lead to misincorporation errors in DNA transcription. Taken together, our results yield mechanistic insights to define how UBPs are efficiently transcribed by a multi-subunit cellular RNAP and will guide further developments to reduce UBP misincorporation during transcription.

Results

Effective and selective recognition of B:S pair by *E. coli* RNAP during transcription

To define how AEGIS DNA is processed during bacterial transcription, we purified recombinant core *E. coli* RNAP and performed in vitro transcription assays (Fig. 1a). We observed an asymmetric transcription recognition pattern of **B**:**S** base pair with a strong strand bias. Specifically, we found that **B**TP is selectively incorporated opposite the d**S** in a template, while no other natural NTP incorporation was observed under identical conditions. In contrast, both **S**TP and UTP can be effectively incorporated opposite a d**B** in the template. Interestingly, for the UTP incorporation reaction, n + 2 extension product (second UTP incorporation opposite the n + 2 dA template) is also observed, suggesting efficient transcription extension from a B:U mismatched pair. These results provide important insights and guidelines for designing AEGIS-containing genome for transcription (for example, d**S** at the template strand and d**B** at the non-template strand).

To quantitively evaluate the substrate selectivity and incorporation efficiency of AEGIS UBPs by *E. coli* RNAP, we performed single-turnover kinetic assays for AEGIS UBP transcription (Fig. 1b and Supplementary Figs 2 and 3, and Table 1). Single turnover experiments allow direct determination of the principal kinetic parameters k_{pol} (the pseudo-first-order catalytic rate constant) and $K_{d,app}$ (the apparent equilibrium constant for dissociation of NTPs from the RNAP elongation complex) of nucleotide incorporation. The ratio $k_{pol}/K_{d,app}$ defines the catalytic efficiency for substrate incorporation.

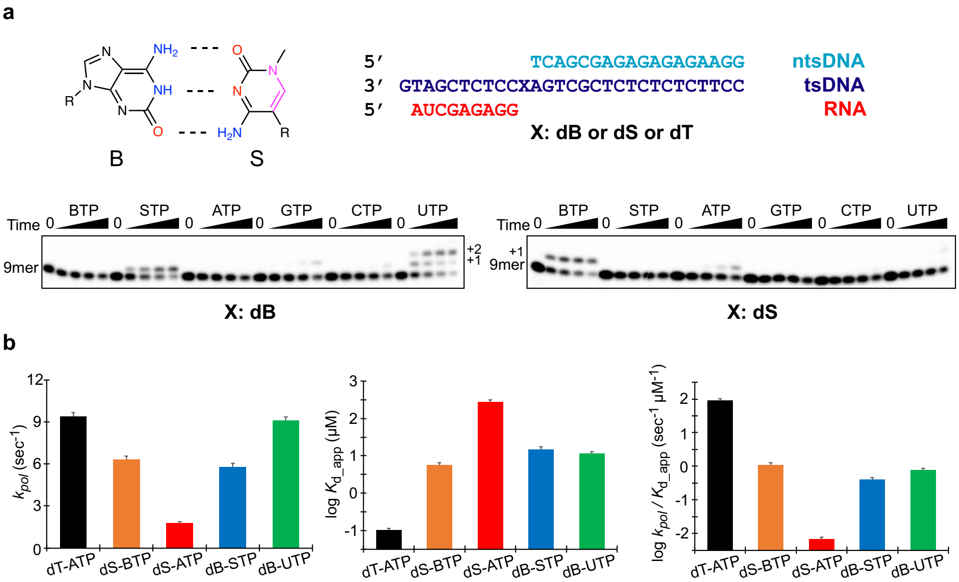

Fig. 1 | Single nucleotide incorporation and kinetic analysis of transcription recognition of B:S pair by *E. coli* RNA polymerase. a B:**S** base pair and DNA/RNA scaffold used in this study. Electron donor and accepter was colored as red and blue, respectively. Pink colored part highlights the distinct moiety from natural bases. tsDNA and ntsDNA stands for template strand DNA and non-template strand DNA, respectively. RNA, tsDNA (TS), ntsDNA (NTS) are shown in red, blue and cyan, respectively. For transcription assay, 20 µM of each nucleoside triphosphate was added, and time points were taken at 0 sec, 15 s, 1 min, 5 min and 30 min. Each reaction is repeated independently at least three times. The band of time 0 s (9mer) serves as internal molecular weight marker: p^{32}-labeled 9mer RNA (5′-AUGGAGAGG-3′). **b** Kinetic parameters for single turnover nucleotide addition. k_{pol}, K_{d_app} and k_{pol}/K_{d_app} are shown with mean and standard error (SE) of best-fit kinetic parameters determined by nonlinear regression using Michaelis-Menten equation from time-course data points (total 81-117 data points) of different concentrations (9 − 13)(Prism 8). Kinetic data of dT-ATP, d**S**-**B**TP, d**S**-ATP, d**B**-**S**TP, and d**B**-UTP are shown in black, orange, red, blue, and green columns. The underlying numerical data for this graph are shown in Table 1.

We found that the k_{pol} parameters of "matching" UBPs (d**S**:**B**TP and d**B**:**S**TP) are similar to those of dT:ATP. In terms of substrate binding, the K_{d_app} of d**S**:**B**TP, d**B**:**S**TP and d**B**:UTP were comparable and ~50–150-fold larger than K_{d_app} of natural dT:ATP incorporation. As a result, the catalytic efficiencies (k_{pol}/K_{d_app}) for d**S**:**B**TP and d**B**:**S**TP are comparable, and are only two-orders of magnitude lower than that for dT:ATP incorporation.

We also measured two misincorporation scenarios using d**B**:UTP and d**S**:ATP (Fig. 1a, b). Intriguingly, the k_{pol} *and* K_{d_app} values of d**B**:UTP misincorporation are comparable to that of matched pairs (d**S**:**B**TP and d**B**:**S**TP). In sharp contrast, d**S**:ATP misincorporation has the lowest incorporation efficiency, with a major differences in both in k_{pol} and K_{d_app} (Fig. 1a, b, and Table 1). Taken together, these findings reveal that the **B**:**S** pair can be effectively and selectively recognized by *E. coli* RNAP when d**S** is in the template strand.

The B:S pair forms a Watson-Crick base pair in the active site of *E. coli* RNAP

The **B**:**S** pair shuffles the hydrogen bonding donors and acceptors that form hydrogen bonds between the components of the base pair. Moreover, the **B**:**S** pair also changes the functional groups presented to the minor groove of the forming double helix. It is not clear how these changes will affect recognition during transcription.

Table 1 | Kinetic parameters of natural/unnatural base incorporation

	Kpol (sec^{-1})	Kd_app (μM)	Kpol / Kd_app (sec^{-1}μM^{-1})	Relative Efficiency
dT: ATP	9.4 ± 0.3	0.10 ± 0.01	92 ± 11	83
dS: BTP	6.3 ± 0.2	5.7 ± 0.9	1.1 ± 0.2	1
dS: ATP	1.8 ± 0.1	270 ± 40	(6.6 ± 1.0) * 10^{-3}	0.006
dB: STP	5.8 ± 0.3	15 ± 2	0.39 ± 0.07	0.35
dB: UTP	9.1 ± 0.2	12 ± 1	0.79 ± 0.09	0.72

The initial velocities were derived from time-course measurements conducted at various substrate concentrations. The kinetic parameters were obtained by plotting these initial velocities against substrate concentrations and fitting the data to the Michaelis-Menten equation.

To understand how this UBP is recognized by *E. coli* RNAP, we solved two cryo-EM structures of the *E. coli* RNAP elongation complex containing a site-specific UBP at the i + 1 site (d**S**:**B**TP and d**B**:**S**TP). To obtain substrate bound structures, we utilized 3′-deoxy RNA, which allows binding of incoming nucleoside triphosphate but prevents nucleotide incorporation. The two high-resolution structures of *E. coli* RNAP with d**S**:**B**TP (2.70 Å) and d**B**:**S**TP (2.65 Å) allow us to unambiguously build the UBP in the RNAP active site (Fig. 2, Supplementary Figs 4 and 5, and Supplementary Table 1).

We found that the structures of both UBPs bound in the active site of *E. coli* RNAP elongation complexes were similar to those reported previously for cognate natural substrates bound to the elongation complex (Fig. 2a)[16,17]. Despite the chemical and structural differences in the functional groups of the **B**:**S** pair (in comparison with natural base pairs), both d**S**:**B**TP and d**B**:**S**TP form Watson-Crick geometry in a manner that is effectively identical to natural Watson-Crick base pairs (G:C and A:U) (Fig. 2b, c). These results indicate that the *E. coli* RNAP active site, which has evolved to recognize natural substrates, is also competent to recognize UBPs with alternative modes of hydrogen bonding.

Several DNA polymerases, such as T7 DNA polymerase (T7 DNAP), Taq DNA polymerase (Taq DNAP), and DNA polymerase beta, make direct hydrogen bond interactions with the minor groove of incoming nucleotide and upstream DNA duplex. The H-donor residues are basic or polar residues such as Lys, Arg, Gln, and Asn[18]. In sharp contrast, multi-subunit RNAPs use non-polar, hydrophobic residues as a steric gate to detect if there are any lesions at the minor groove[19,20]. Indeed, we previously reported the Pro-gate loop that interacts with incoming NTP and template base via Van der Waals interactions[19,20]. To examine the potential minor groove interactions between the B:S pair and *E. coli* RNAP, we investigated the RNAP residues near the minor groove of

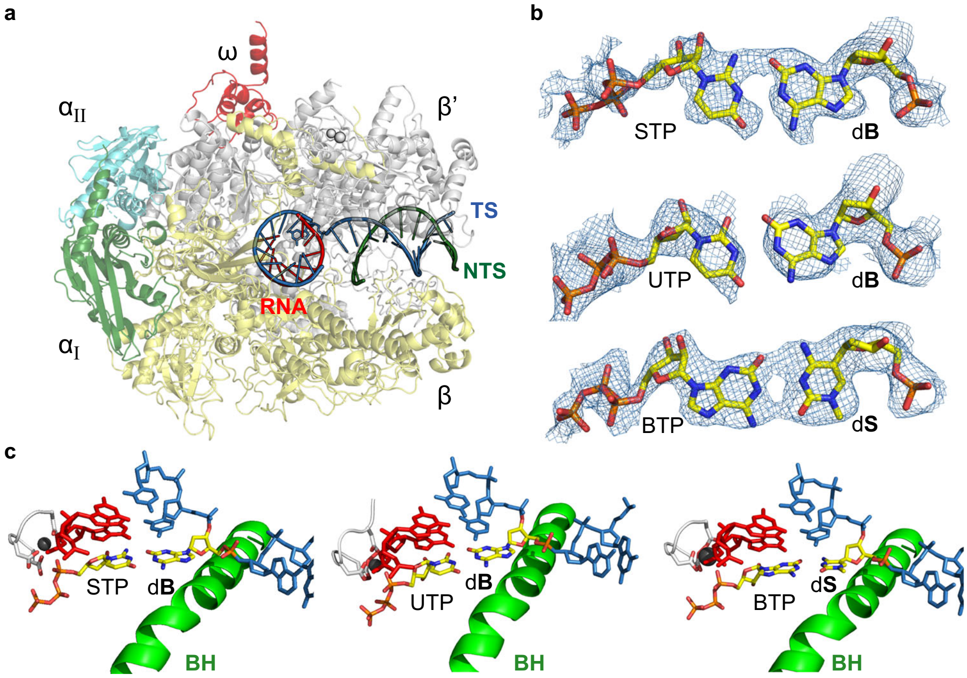

Fig. 2 | Overall structures of *E. coli* RNA polymerase elongation complex harboring unnatural base pair. a Overall structure of *E.coli* RNA polymerase EC. α_I, α_{II}, β′, β and ω subunit of RNAP were colored in green, cyan, gray, yellow, and red, respectively. RNA, template strand DNA (TS), non-template strand DNA (NTS) are shown in red, blue and dark green, respectively. **b** Model and map of three unnatural base pairs; d**B**:**S**TP, d**B**:UTP and d**S**:**B**TP are shown with Cryo-EM density maps. **c** Active site of three UBP EC structures. Bridge helix (BH) is colored in green. UBP in +1 position were colored yellow. Other parts of the structures are colored white. The nitrogen, oxygen, phosphate, and carbon atoms of the base pair at i + 1 position are colored in blue, red, orange, and yellow, respectively.

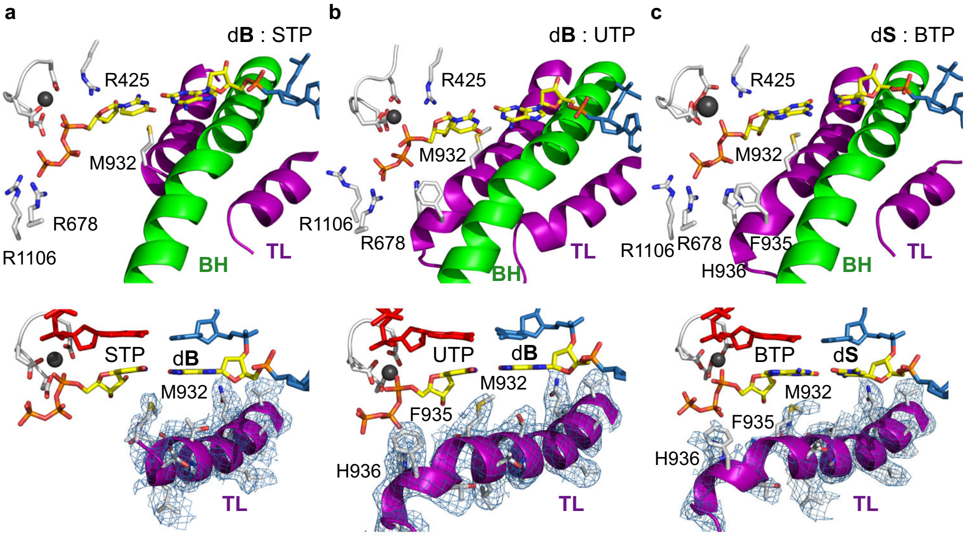

Fig. 3 | Formation of UBP induces closure of trigger loop in *E. coli* RNAP elongation complex. a–c Residues interacting with bound nucleotide triphosphate substrate are highlighted. Trigger loop (TL) and Bridge helix (BH) are colored in purple and green, respectively. Magnesium ion in black. Bottom column shows densities of TLs for (d**B**:**S**TP), (d**B**:UTP) and (d**S**:**B**TP), respectively. R678 and R1106 belongs to β subunit of RNAP. R425, M932, F935 and H936 belongs to β′ subunit of RNAP.

incoming substrate and +1 template position (Supplementary Fig. 6). We found that non-polar residues Pro427 (from Pro-gate loop) and Met1273 are close to +1 template and incoming nucleotide substrate, respectively. In addition, Ala426 (from Pro-gate loop) are close to upstream -1 base pair. These interactions are identical to natural base pair recognition by *E. coli* RNAP. These data suggest that multi-subunit RNAPs use different minor groove recognition principle than DNAPs, employing steric discrimination instead of recognition via direct hydrogen bonding interactions. Accordingly, multi-subunit RNAPs may be more tolerant to the electron donor/acceptor functional group changes at the minor groove, if they remain similar in size.

Formation of an unnatural base pair induces trigger loop closure of *E. coli* RNAP

The natural nucleotide substrate diffuses from the secondary channel of RNAP and reaches the addition site (A-site) if it can form a Watson-Crick base pair with the template base[21,22]. Cellular RNA polymerases select the correct nucleotide substrate over a mismatched substrate via a comprehensive interaction network containing template strand, RNA primer, and protein residues at the active sites.

In particular, the trigger loop (TL), a structurally conserved motif of all multi-subunit RNAPs, is important for substrate recognition and catalysis[23–25]. TL undergoes a significant conformational change in response to nucleotide binding and addition during transcription. Binding of the correct substrate induces a conformational change to an active, closed state. Closure of TL seals off the RNAP active site and provides correct positioning of the natural nucleotide substrates for incorporation in both prokaryotic and eukaryotic RNAPs[23–25]. In contrast, TL is flexible and resides in a relaxed, inactive open state in the absence of substrate or in the presence of mismatched substrate[26].

To investigate whether the binding of unnatural nucleotides leads to a conformational change in the TL, we examined the atomic structure of RNAP active site and the TL conformation. We observed well-defined density that can be attributed to the closed TL for both d**B**:**S**TP and d**S**:**B**TP structures. We built the residues of TL corresponding to the proximal TL helix (Fig. 3). Key TL residues, such as M932, F935, H936, maintained similar interactions with incoming unnatural NTPs, compared to previously reported bacterial RNAP closed elongation complex structures containing the natural scaffold (Fig. 4 and Supplementary Fig. 7)[17,24]. Therefore, we concluded that *E. coli* RNAP can recognize unnatural substrate **B**TP or **S**TP opposite their corresponding template partner using principles of Watson-Crick base-pairing, analogously to natural cognate NTPs. Furthermore, the high-resolution structures show that, when d**B**:**S**TP and d**S**:**B**TP are bound, the active site of RNAP is in the TL-closed, active state, which is poised for nucleotide addition. Taken together, these cryo-EM structures explain how and why STP and BTP can be effectively incorporated by *E. coli* RNAP.

Structural basis of transcription recognition of the dB-UTP pair by *E. coli* RNAP

Our biochemical data showed that UTP can be misincorporated opposite template d**B** in an in vitro transcription assay. This efficient misincorporation prompted us to examine how UTP interacts with the d**B** template during transcription using high-resolution structural biology.

Previous reports suggest that several different forms of **B**:U base pair can arise to allow **B**:U mismatching (Supplementary Fig. 8). The first involves a **B**:U reverse wobble base pair, generated through the major keto form. The second allows for a **B**:U Watson-Crick base pair via tautomerization of either template d**B** or incoming substrate UTP as the enol tautomer[27,28]. Free energy calculations in gas and solution suggest that the d**B**:UTP (reverse) wobble pair and the d**B** (enol):UTP pair would have similar base pairing stability[27]. To distinguish between these two scenarios, we solved the cryo-EM structure of *E. coli* RNAP containing mismatched d**B**:UTP at 3.28 Å resolution and derived an atomic model from the density (Supplementary Fig. 9 and Supplementary Table 1). We found that the configuration of d**B**:UTP base pair in the active site is consistent with Watson-Crick like geometry, but not the reverse wobble base pair geometry (Fig. 2b).

Intriguingly, we also observed strong density for the TL that allowed us to fully build intact TL (β′ residues 916-1146), including the insertion motif (β′ residues 943-1130) (Fig. 4 and Supplementary Fig. 7). The map quality of TL is even better than that from matched d**B**:**S**TP and d**S**:**B**TP (Fig. 4 and Supplementary Fig. 7). Taken together, our data revealed the structural basis of effective UTP incorporation opposite the d**B** template. The d**B**:UTP base pair adopts a Watson-Crick like geometry with an active, closed TL in a manner that is analogous to a canonical matched natural base pair.

Previous free energy calculations suggest a lack of major free energy differences for distinct conformations of d**B**:T/U base pairs[27]. This raises the question: why is the Watson-Crick geometry favored for the **B**:U pair by *E. coli* RNA polymerase? To address this question, we modeled all potential base pairs within the active site (Fig. 5). We found that the d**B** (imine):UTP pair in a wobble geometry loses a hydrogen bond as well as 2′-OH interaction between ribose and R425 of β′ in pro-gate loop (Fig. 5b). On the other hand, in the reverse wobble conformation for d**B**(keto):UTP, there would be a strong steric clash between the UTP and the pro-gate loop of *E. coli* RNAP (the conserved region of yeast Pol II (β′ 425-431, rpb1 440-460 in yeast Pol II)) (Fig. 5c)[20,29]. Thus, steric constraints promote the configuration of Watson-Crick base pairing for both natural and UBPs and discourage the reverse wobble conformation. Moreover, the Watson-Crick geometry has the strongest base stacking, with i-1 base pairs. The base stacking between -1 template DNA and d**B**, -1 RNA primer and UTP also provide further stability for Watson-Crick like pair. Taken together, these neighboring effects at RNA polymerase active site favor Watson-Crick geometry over alternative base pair conformations. Our findings reveal that Watson-Crick geometry is energetically favored during transcription. These data extend previous observations for how Watson-Crick geometry is favored during replication and translation[30–34], enhancing our understanding of the shared principles that govern the major molecular processes in the Central Dogma.

Our structure of RNAP with d**B**:UTP suggests that the d**B**:UTP base pair adopts a Watson-Crick-like geometry via tautomerization.

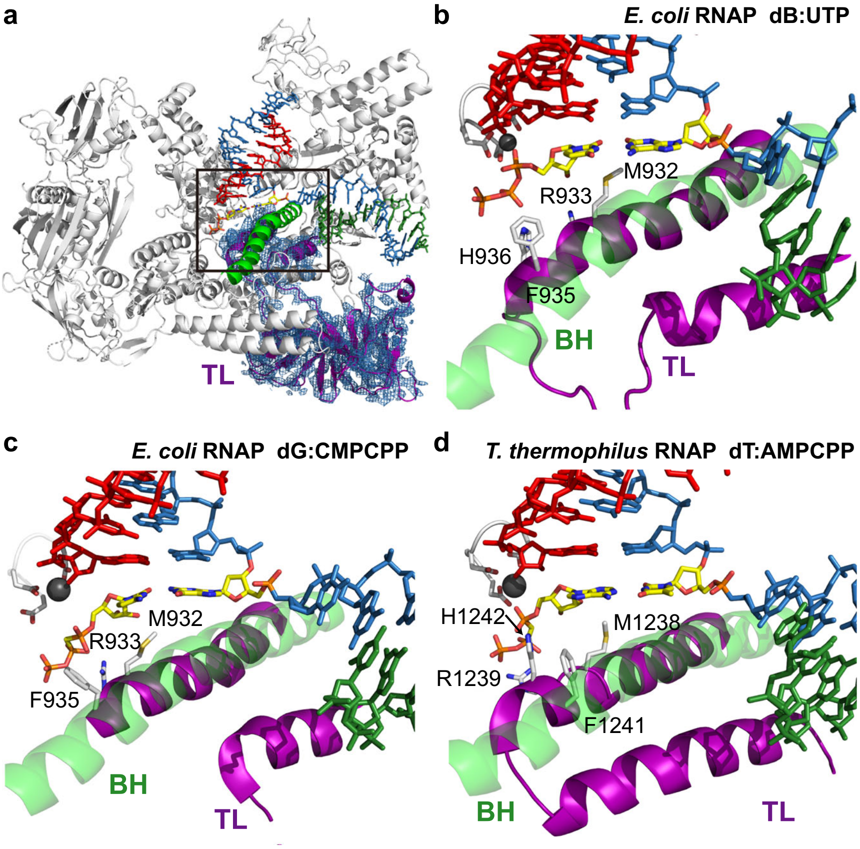

Fig. 4 | Comparison of trigger loop configurations among different RNAPs containing unnatural and natural base pairs. a *E. coli* RNAP with UBP. The trigger loop of *E. coli* RNAP d**B**:UTP (β′ residues 916-1146), including the insertion domain (β′ residues 943-1130) are highlighted in purple. **b**–**d** Comparison of three TL configurations in three RNAP EC (elongation complex) structures, including *E. coli* RNAP with d**B**:UTP (**b**), *E. coli* RNAP with dG:CMPCPP (PDB ID: 7MKO) (**c**), *T. thermophilus* RNAP with dT:AMPCPP (PDB ID: 2O5J) (**d**). All residues labeled belongs to β′ subunit of RNAP.

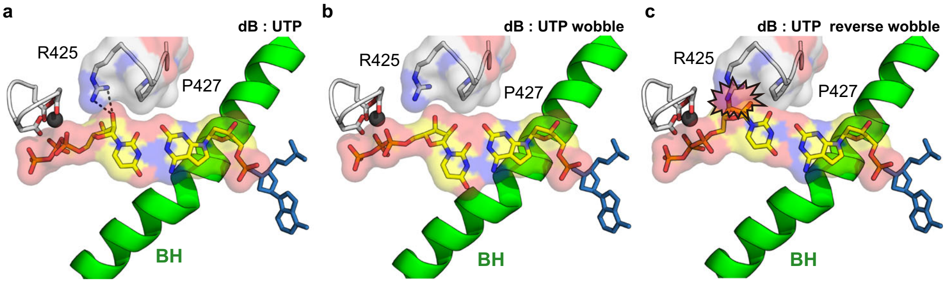

Fig. 5 | Potential mismatched dB:UTP base pair configuration within *E. coli* RNAP active site. a Surface representation of d**B**:UTP observed in our cryo-EM structure. **b** A d**B**:UTP wobble base pair model was prepared based on a G:U wobble base pair (PDB 6L0Y). **c** A d**B**:UTP reverse wobble base pair was modeled based on a B_{t1}-T reverse wobble base pair[27]. The i + 1 base pair and minor-groove edge contacting loop are highlighted in surface view. Other color code is the same as Figs. 2 and 3.

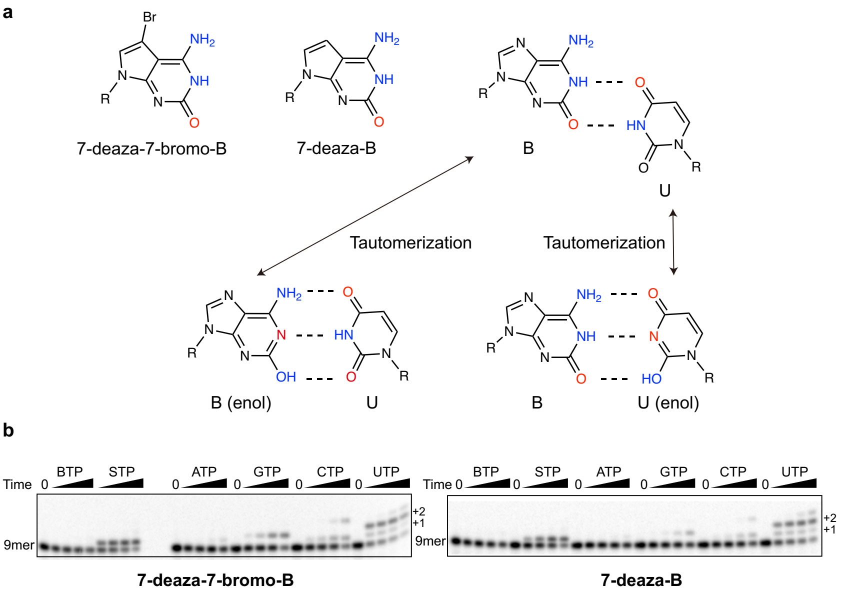

Fig. 6 | Single nucleotide incorporation assay with 7-deaza modified dB templates. a Schematic representation of modified **B** and possible **B**:U mismatch pairs. B (enol) tautomer:U pair and B:U (enol) tautomer were shown. Note that **B**:S and **B**:U (enol) pairs are modeled in our final structures. **b** Time-course of a single nucleotide incorporation assay: 100 μM of each nucleoside triphosphate was added and results were recorded in 0 s, 15 s, 1 min, 5 min and 30 min, respectively. Each reaction is repeated independently for three times ($n = 3$). The band of time 0 s (9mer) serves as internal molecular weight marker: p^{32}-labeled 9mer RNA (5′-AUGGAGAGG-3′).

However, the structure, by itself, cannot distinguish which tautomerization form will prevail when there is a **B**:U mismatch pair, because both tautomers, d**B** (enol):UTP (keto) and d**B** (keto):UTP(enol), abide by principles of Watson-Crick geometry (Fig. 6a). To determine whether the keto/enol ratios of d**B** template would affect UTP incorporation, we examined the single incorporation efficiency of nucleotide triphosphates when the two keto-dominated d**B** analogs, 7-deaza-d**B** and 7-deaza-7-bromo-d**B**, reside in the template position. Previous research suggests that K_{TAUT} (=[keto]/[enol]) of d**B** was ~10, meaning that ~10% of d**B** exists in the enol tautomer form, whereas two d**B** analogs 7-deaza-d**B** and 7-deaza-7-bromo-d**B** strongly favor the keto tautomer form, with the K_{TAUT} values of these two d**B** analogs are 10^3 and 10^4, respectively (Fig. 6a)[35–38]. In our assay condition, we found that the UTP was still incorporated effectively with two keto-dominated 7-deaza-d**B** templates, suggesting the d**B** (enol) form is unlikely the major driving force for tautomerization in d**B**:UTP Watson-Crick-like base pair (Fig. 6b). Based on this biochemical evidence, we modeled d**B** (keto):UTP (enol) in our structure. Future studies would be needed to fully understand the energy landscapes and interconversion dynamic of different tautomerization forms of d**B**:UTP within the active site of RNAP.

Discussion

Transcription recognition of unnatural B:S base pair as a third base pair

Our biochemical and structural analyses demonstrate that the **B**:**S** pair can be utilized effectively and selectively by *E. coli* RNAP as a third base pair in a manner similar to that used by natural base pairs during transcription. We show that **B**TP and **S**TP can be incorporated efficiently opposite the template d**S** and d**B**, respectively.

Strikingly, our structural studies reveal that the trigger loop can adopt a fully closed active state in all cryo-EM structures. The configuration of TL is fine-tuned to discern both matched and mismatched substrates, allowing it to play an important role in nucleotide selection, positioning, and addition. These structures suggest that *E. coli* RNAP can recognize the **B**:**S** UBPs in the same manner as natural base pairs, highlighting the potential of **B**:**S** pair as a third base pair for an expanded genetic alphabet. Our data yield mechanistic insights into how AEGIS UBPs with alternative hydrogen bonding are specifically recognized and incorporated into an elongating RNA transcript by a multi-subunit cellular RNA polymerase. Importantly, these AEGIS UBPs can be recognized just like natural base pairs, and we are able to capture the active, closed state of RNAP elongation complex. These insights will guide the design of next generation AEGIS UBPs and facilitate engineering of UBP-specific RNA polymerases. Thus, our results expand the area of synthetic biology, which will ultimately help design de novo proteins using non-canonical amino acids.

In addition to UBP with alternative hydrogen bonding developed by Benner group, other groups have developed hydrophobic UBPs without any inter-base hydrogen bonding. These include the **TPT3**:**NaM** pair by Romesberg group and the **Ds**:**Pa** pair by Hirao group[14,15]. Recent studies showed that the hydrophobic **TPT3**:**NaM** pair can be also transcribed by *E. coli* RNAP to produce mRNA with unnatural codons, which can be translated in proteins with site-specifically incorporated noncanonical amino acids[39]. However, the rate of incorporation of these hydrophobic UBPs are much slower than that of the **S**:**B** pair. Our understanding for UBP transcription is still very limited. Recently, our group reported the structural basis of transcription recognition of hydrophobic UBPs (**TPT3**:**NaM** pair by RNA Pol II and **Ds**:**Pa** pair by T7 RNAP)[14,15], none of these previous RNAP structures were captured in a closed, active state. No *E. coli* RNAP structures containing hydrophobic UBPs are yet available for a direct comparison with AEGIS UBPs. Systematic biochemical and structural studies are needed to compare transcription recognition of different UBPs by the same enzyme, or transcription recognition of the same UBPs by different RNA polymerases for future studies.

Structural evidence of mismatch base pairs via tautomerization during transcription

Mismatches of natural base pairs can arise and lead to mutations during replication, transcription, and translation. Watson and Crick proposed that spontaneous mutations may stochastically arise when an incoming base adopts an unfavorable tautomeric form[2]. The environment surrounding the base pair undoubtedly impacts the tautomeric equilibrium between keto/enol forms. In the last decade, WC-like mispairs via tautomerization have been observed in the active sites of DNA polymerases and the ribosomes in catalytically competent conformations[30–32,34,40–43]. These results provide important insights into misincorporations at replication and translation processes and support Watson and Crick's tautomer hypothesis.

However, it remained unclear whether the similar tautomerization mechanism for mismatched pairs may also arise during transcription, due to the lack of structural data. Our results demonstrate that the same tautomerization principles can be extended into the transcription process. Here, we provide evidence that UTP can tautomerize, leading to a base pair structure formed by three hydrogen bonds. Our observation that the **B**:U base pair adopts the Watson-Crick U's enol form, and not the wobble base pair U's keto form, serves as an example of base pair tautomerization during transcription. Therefore, our results provide structural evidence to support the long-standing tautomer hypothesis.

Methods

Oligonucleotides and nucleotides used in this study

RNA oligo (5′-AUCGAGAGG), 3′-deoxy RNA oligo (5′-AUCGAGAG/3′dG/), template strand (ts) DNA oligo (dT, 5′- CCTTCTCTCTCTCGCTGAT CCTCTCGATG) and non-template strand (nts) DNA oligo (5′-TCAGCGA GAGAGAGAAGG) were purchased from IDT. AEGIS ribotriphosphates (r**B**TP and r**S**TP) were from Firebird Biomolecular Sciences LLC (Alachua, FL). Standard phosphoramidites (Bz-dA, Ac-dC, dmf-dG, dT and d**B**: dmf-isodG-CE Phosphoramidite) and controlled pore glass (CPG) having standard residues were purchased from Glen Research (Sterling, VA) and AEGIS phosphoramidites (d**S**, 7-deaza-d**B** and 7-bromo-7-deaza-d**B**) were purchased from Firebird Biomolecular Sciences LLC (Alachua, FL).

All oligonucleotides containing d**S**, d**B**, 7-deaza-d**B**, and 7-bromo-7-deaza-d**B** were synthesized on an ABI 394 DNA Synthesizer following standard phosphoramidite chemistry[44,45]. Briefly, samples of CPG supporting the synthetic oligonucleotides were treated with 2.0 mL of 1 M DBU in anhydrous acetonitrile at room temperature for 24 h to remove the NPE group. Then, the CPGs were filtered, dried, and treated with concentrated ammonium hydroxide at 55 °C for 16 h. After removal of ammonium hydroxide, the oligonucleotides were purified on an ion-exchange HPLC, and then desalted using Sep-Pac® Plus C18 cartridges (Waters).

Purification of *E. coli* RNA polymerase

The expression plasmids for *E. coli* RNA polymerase (deposited by Seth Darst lab) were purchased from Addgene. We followed previous purification methods with modification[46]. pEcRNAP6 (Addgene # 128940) and pACYCDuet-LEc-rpoZ (Addgene # 128837) were co-transformed to BL21 (DE3) competent cells (Novagen). 0.25 mM of IPTG was added to cell culture when O.D. reached 0.6 and protein were expressed in 20 °C for overnight. Cells were collected and resuspended in Buffer A [50 mM Tris (pH 7.4), 500 mM NaCl, 5% (v/v) glycerol, 2 mM β-mercaptoethanol (BME) and 1X EDTA-free protease inhibitor (Sigma-Aldrich)]. After lysis and centrifugation, cell lysate was loaded to Ni-NTA resin and washed with Buffer A + 20 mM imidazole, Buffer A (300 mM NaCl) + 30 mM imidazole and eluted by Buffer A (300 mM NaCl) + 250 mM imidazole. Eluted sample was diluted with Buffer A (no salt) 2 times to reduce NaCl concentration to 150 mM. After dilution, sample was loaded to HiTrap Heparin column, and purified by increasing NaCl concentration from 150 mM to 800 mM. For anion exchange purification, heparin elute was dialyzed against Buffer B (20 mM HEPES pH 8.0, 100 mM NaCl, 5% (v/v) glycerol, 1 mM DTT). After dialysis, sample was loaded into HiTrap Q column. *E. coli* RNA polymerase ($\alpha_2\beta'\beta\omega$) was purified by increasing NaCl concentration from 100 mM to 800 mM. During concentration, buffer was changed to Buffer B with 150 mM NaCl, flash frozen in liquid nitrogen and stored in −80 °C for further use.

In vitro transcription assay and kinetic analysis

Transcription assays were performed with previously described methods with minor modifications[19]. Mini-scaffold containing 200 nM RNA (P^{32}-labeled), 400 nM tsDNA and 600 nM ntsDNA were mixed in elongation buffer (EB), 20 mM Tris (pH 8.0 at 4 °C), 40 mM KCl, 5 mM DTT and 5 mM $MgCl_2$. Scaffold was annealed by heating at 80 °C for 5 min and slow cooled down to room temperature for at least 2 h. *E. coli* RNAP elongation complex was prepared by adding Mini-scaffold in 1X EB and incubate for 20 min in 30 °C. Reaction was initiated by adding elongation complex to each nucleotide triphosphate. As a result, reaction sample has 20 nM of Mini-scaffold, 120 nM *E. coli* RNAP and 20 μM each nucleotide triphosphates in 1X EB, if not mentioned otherwise. At each timepoint, the reaction mixture was pipetted into stop buffer [90% (v/v) formamide, 50 mM EDTA, 0.05% (w/v) xylene cyanol and 0.05% (w/v) bromophenol blue]. All samples were denatured by heating 95 °C for 10 min and analyzed by 12% denaturing urea-PAGE gel analysis. All reactions were performed in room temperature. Kinetic parameters in Table 1 and Supplementary Fig. 2 were obtained by using Prism8 non-linear regression fit (Michaelis-Menten). Briefly, non-linear Michaelis-Menten regression was performed by plotting [S] vs V0 values, which obtained from time vs % incorporated substrate graph in Supplementary Fig. 2 (regression curves) and 3 (Raw gel data). Each concentration has 9 time points. Time-course data points used for kinetic parameters fitting are: total 81 data points from 9 different concentrations for dT:ATP, total 117 data points from 13 different concentrations for d**S:B**TP, total 90 data points from 10 different concentrations for d**S**:ATP, total 99 data points from 11 different concentrations for d**B:S**TP and total 81 data points from 9 different concentrations for d**B**:UTP, respectively.

Preparation of *E. coli* RNA polymerase elongation complex for electron microscopy

Mini-scaffold containing 3′-deoxy RNA, tsDNA and ntsDNA with molar ratio of 1.2:1:1.2 were annealed in 1X EB. To form elongation complex, purified *E. coli* RNAP were mixed with prepared Mini-scaffold with molar ratio 1:1.3 and incubated in ice for 1 h (pH 8.0 at 4 °C). Final 4 mM of $MgCl_2$ and **B**TP or **S**TP or UTP in 1X EB were added to elongation complex and incubated in ice for 30 min. Before sample preparation, final 4 mM CHAPSO dissolved in 1X EB were added to complex to reduce particle orientation bias[47]. The final *E. coli* RNAP concentration was 15–18 mg/ml. If dilution was needed, 4 mM substrate nucleotide triphosphate, 4 mM $MgCl_2$, and 4 mM CHAPSO in 1X EB were used.

Cryo-EM sample preparation. The procedures for grid vitrification for all three cryo-EM sample preparations were identical. 2.5 μl of 15–18 mg/ml *E. coli* RNA polymerase elongation complex was pipetted onto freshly plasma cleaned (20 s, Solarus plasma cleaner) holey grids (Quantifoil UltrAufoil R 1.2/1.3 300 mesh) at 4 °C in the cold room. The sample was incubated onto grids for 30 s before manually blotting for 6 s, then plunged into the liquid ethane using a manual plunger. The vitrified grids were transferred to liquid nitrogen for storage and data collection.

Cryo-EM data collection and processing dS. BTP and d**B:S**TP cryo-EM datasets were collected in Stanford-SLAC cryo-EM center, while d**B**:UTP cryo-EM dataset was collected in the Scripps cryo-EM facility. All datasets were collected using a Titan Krios transmission electron microscope (Thermo Fisher Scientific) operating at 300 keV, equipped with either a K2 or K3 direct electron detector (Gatan) with a GIF Quantum Filter with a slit width of 20 eV (an energy filter was not used for d**B**:UTP sample). All data collections were performed either using SerialEM, EPU or Leginon[48–51]. The software used for data collection depended on what was installed for the microscope. Some data collections were performed using stage tilt[52]. The imaging parameters for each dataset are comprehensively summarized in Supplementary Table 1.

All three cryo-EM datasets were processed using a similar workflow. For each dataset, the movies were imported into Relion 4.0-beta-2 for dose-weighted motion correction on 5-by-5 patch squares and with a B factor of 150 Å^2 using the gain reference that was generated during data collection[53,54]. The motion-corrected micrographs were then imported into Warp 1.0.9 to perform CTF estimation and particle selection[55]. The particles that scored above 0.7 were selected in Warp with a re-trained BoxNet model using constraint settings of particle diameter of 180 Å and a minimum distance between particles of 20 Å. The particle star file was made by initially extracting selected particles using a box size of 384 pixels.

The aligned micrographs in Relion and particle star file from Warp were imported into cryoSPARC V3.3.2 for the next steps of data processing[56]. The particles were re-extracted in cryoSPARC with the same box size of 384 pixels for performing 2D classification. The particles that contributed to the best classes with good particle features were selected after several rounds of 2D classification. The selected particles from 2D were subjected to multiple rounds of heterogeneous refinement to further clean the dataset until no further improvements were observed, as indicated by the Fourier shell correlation (FSC)[57].

After each round of heterogeneous refinement, particles that were assigned to the highest-resolution map were selected, whereas the remaining particles were discarded. Multiple rounds of heterogeneous refinement were performed. The selected particles from the last heterogeneous refinement were used to perform a homogeneous refinement to generate a high-resolution reconstruction. At this point, the remaining particles from homogeneous refinement were imported back to Relion for Bayesian polishing using default parameters[58]. Subsequently, the polished particles were re-imported back to cryoSPARC to perform one round of per-particle CTF refinement and one round of global CTF refinement. We repeated the procedure – Bayesian polishing in Relion followed by CTF refinement in cryoSPARC iteratively, until no further improvements in resolution were observed, as determined using the FSC. Generally, the optics parameters, per-particle defoci, and the map resolutions converged after several rounds of polishing and CTF refinement.

The final map was generated by the last homogeneous refinement step. Density modification was performed for high-resolution d**S**:**B**TP and d**B**:**S**TP maps in Phenix-1.20.-4487 prior to modeling[59,60]. The directional resolution of the map was evaluated using the 3D FSC server (3dfsc.salk.edu)[52], and the quality of the orientation distribution was evaluated using the Sampling Compensation Function (SCF)[61,62]. All images in Supplementary Figs 4, 5, and 9 were generated using UCSF Chimera[63].

Model building

We first built the atomic model for d**B**:**S**TP map, which was the highest-resolution map obtained during our experiments. The initial unhydrated atomic model of *E.coli* RNA polymerase elongation complex, resolved to 4.4 Å (PDB 6ALH was rigid body docked into the density modified d**B**:**S**TP map using UCSF Chimera[16]. Clear density in the active site of *E.coli* RNA polymerase allowed unambiguous assignment of register of the DNA template and RNA oligomer that used for assembling d**B**:**S**TP RNA polymerase elongation complex.

For unnatural nucleotide r**S**TP or r**B**TP, the coordinates and restraint files were generated by using phenix elbow, then atomic model of r**S**TP was fitted to the map[64]. The docked atomic model of RNA polymerase elongation complex was merged with the docked atomic model of r**S**TP to obtain the first intact rigid body docking model for d**B**:**S**TP map. Subsequently, the model was manually adjusted in Coot wherever discrepancies were apparent in the high-resolution map[65]. The final model was generated by iterative model adjustment using Coot and real-space refinement using phenix with restraint weights[64]. For modeling d**S**:**B**TP and d**B**:UTP maps, the same procedure was adopted, except using d**B**:**S**TP atomic model as the starting model.

d**S**:**B**TP and d**B**:**S**TP atomic models were individually hydrated to account for their high-resolution using phenix.douse. Before dousing, cryo-EM density maps were resampled and resized such that the voxel size became roughly one quarter of the resolution estimate using software available in CisTEM[66]. We next used phenix.douse in Phenix 1.20.1-4487 to programmatically add water molecules using settings (dist_min = 2.5 dist_max = 4.5 keep_input_water = TRUE sphericity_filter = false, varied in map_threshold value) to the density that did not account for protein and DNA molecules. The hydrated atomic models were individually subjected to real space refinement using phenix.real_space_refinement.

After real space refinement, the hydrated atomic models were individually inspected in Coot to evaluate agreement of each water placement within cryo-EM maps, and problematic water molecules were removed if they (I) were placed in vacant density, (II) clashed with protein and nucleotide molecules, (III) were placed in densities that belong to protein and nucleotide molecules, or (IV) were in periphery region that is far away from intact complex. We performed several iterative rounds of real space refinement in Phenix and water evaluation in Coot. The geometry of the final models and validation statistics were generated by Molprobity[67]. The refinement statistics for each structural model are summarized in Supplementary Table 1. All the structural figures were prepared using PyMOL[68].

Reporting summary

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

Data availability

The electron potential maps of RNAP elongation complexes are deposited into the electron microscopy databank as EMD-40862, EMD-40863, and EMD-40864. The models for RNAP elongation complexes have been deposited into the PDB as PDB ID 8SY5 (CryoEM structure of *E.coli* RNA polymerase elongation complex containing d**S** template and **B**TP as substrate), PDB ID 8SY6 (CryoEM structure of *E.coli* RNA polymerase elongation complex containing d**B** template and UTP as substrate), and PDB ID 8SY7 (CryoEM structure of *E.coli* RNA polymerase elongation complex containing d**B** template and **S**TP as substrate). Other cited published PDB entries are: PDB entries 6ALH (CryoEM structure of *E.coli* RNA polymerase elongation complex), 7MKO (*E.coli* RNA polymerase elongation complex), 2O5J (Crystal structure of the *T. thermophilus* RNAP polymerase elongation complex with the NTP substrate analog), 6L0Y (Structure of dsRNA with G-U wobble base pairs). Source data are provided with this paper.

References

1. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
2. Watson, J. D. & Crick, F. H. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967 (1953).
3. Kool, E. T., Morales, J. C. & Guckian, K. M. Mimicking the structure and function of DNA: Insights into DNA stability and replication. *Angew. Chem. Int. Ed.* **39**, 990–1009 (2000).
4. Hirao, I. et al. An unnatural hydrophobic base pair system: site-specific incorporation of nucleotide analogs into DNA and RNA. *Nat. Methods* **3**, 729–735 (2006).
5. Hoshika, S. et al. Hachimoji DNA and RNA: a genetic system with eight building blocks. *Science* **363**, 884–887 (2019).
6. Malyshev, D. A. & Romesberg, F. E. The expanded genetic alphabet. *Angew. Chem. Int. Ed. Engl.* **54**, 11930–11944 (2015).

7. Piccirilli, J. A., Krauch, T., Moroney, S. E. & Benner, S. A. Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **343**, 33–37 (1990).
8. Switzer, C. Y., Moroney, S. E. & Benner, S. A. Enzymatic recognition of the base pair between isocytidine and isoguanosine. *Biochemistry* **32**, 10489–10496 (1993).
9. Biondi, E. & Benner, S. A. Artificially expanded genetic information systems for new aptamer technologies. *Biomedicines* **6**, 53 (2018).
10. Sismour, A. M. et al. PCR amplification of DNA containing nonstandard base pairs by variants of reverse transcriptase from Human Immunodeficiency Virus-1. *Nucleic Acids Res.* **32**, 728–735 (2004).
11. Georgiadis, M. M. et al. Structural basis for a six nucleotide genetic alphabet. *J. Am. Chem. Soc.* **137**, 6947–6955 (2015).
12. Yang, Z., Sismour, A. M., Sheng, P., Puskar, N. L. & Benner, S. A. Enzymatic incorporation of a third nucleobase pair. *Nucleic Acids Res.* **35**, 4238–4249 (2007).
13. Leal, N. A. et al. Transcription, reverse transcription, and analysis of RNA containing artificial genetic components. *ACS Synth. Biol.* **4**, 407–413 (2015).
14. Oh, J. et al. Transcriptional processing of an unnatural base pair by eukaryotic RNA polymerase II. *Nat. Chem. Biol.* **17**, 906–914 (2021).
15. Oh, J. et al. Structural basis of transcription recognition of a hydrophobic unnatural base pair by T7 RNA polymerase. *Nat. Commun.* **14**, 195 (2023).
16. Kang, J. Y. et al. Structural basis of transcription arrest by coliphage HK022 Nun in an Escherichia coli RNA polymerase elongation complex. *Elife* **6**, e25478 (2017).
17. Qayyum, M. Z., Molodtsov, V., Renda, A. & Murakami, K. S. Structural basis of RNA polymerase recycling by the Swi2/Snf2 family of ATPase RapA in Escherichia coli. *J. Biol. Chem.* **297**, 101404 (2021).
18. Ouaray, Z., Benner, S. A., Georgiadis, M. M. & Richards, N. G. J. Building better polymerases: engineering the replication of expanded genetic alphabets. *J. Biol. Chem.* **295**, 17046–17059 (2020).
19. 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).
20. Xu, L. et al. Mechanism of DNA alkylation-induced transcriptional stalling, lesion bypass, and mutagenesis. *Proc. Natl. Acad. Sci. USA* **114**, E7082–E7091 (2017).
21. Batada, N. N., Westover, K. D., Bushnell, D. A., Levitt, M. & Kornberg, R. D. Diffusion of nucleoside triphosphates and role of the entry site to the RNA polymerase II active center. *Proc. Natl. Acad. Sci. USA* **101**, 17361–17364 (2004).
22. Westover, K. D., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. *Cell* **119**, 481–489 (2004).
23. Wang, D., Bushnell, D. A., Westover, K. D., Kaplan, C. D. & Kornberg, R. D. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127**, 941–954 (2006).
24. Vassylyev, D. G. et al. Structural basis for substrate loading in bacterial RNA polymerase. *Nature* **448**, 163–168 (2007).
25. Toulokhonov, I., Zhang, J., Palangat, M. & Landick, R. A central role of the RNA polymerase trigger loop in active-site rearrangement during transcriptional pausing. *Mol. Cell* **27**, 406–419 (2007).
26. Kaplan, C. D., Larsson, K. M. & Kornberg, R. D. The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin. *Mol. Cell* **30**, 547–556 (2008).
27. Jena, N. R. Rare tautomers of artificially expanded genetic letters and their effects on the base pair stabilities. *Chemphyschem* **23**, e202100908 (2022).
28. Kimsey, I. J. et al. Dynamic basis for dG.dT misincorporation via tautomerization and ionization. *Nature* **554**, 195–201 (2018).
29. 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).
30. Bebenek, K., Pedersen, L. C. & Kunkel, T. A. Replication infidelity via a mismatch with Watson-Crick geometry. *Proc. Natl. Acad. Sci. USA* **108**, 1862–1867 (2011).
31. Maciejewska, A. M., Lichota, K. D. & Kusmierek, J. T. Neighbouring bases in template influence base-pairing of isoguanine. *Biochem. J.* **369**, 611–618 (2003).
32. Li, P., Rangadurai, A., Al-Hashimi, H. M. & Hammes-Schiffer, S. Environmental effects on Guanine-Thymine Mispair Tautomerization explored with quantum mechanical/molecular mechanical free energy simulations. *J. Am. Chem. Soc.* **142**, 11183–11191 (2020).
33. Rozov, A., Westhof, E., Yusupov, M. & Yusupova, G. The ribosome prohibits the G*U wobble geometry at the first position of the codon-anticodon helix. *Nucleic Acids Res.* **44**, 6434–6441 (2016).
34. Wang, W., Hellinga, H. W. & Beese, L. S. Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proc. Natl. Acad. Sci. USA* **108**, 17644–17648 (2011).
35. Martinot, T. A. & Benner, S. A. Artificial genetic systems: exploiting the "aromaticity" formalism to improve the tautomeric ratio for isoguanosine derivatives. *J. Org. Chem.* **69**, 3972–3975 (2004).
36. Seela, F., Peng, X. & Li, H. Base-pairing, tautomerism, and mismatch discrimination of 7-halogenated 7-deaza-2'-deoxyisoguanosine: oligonucleotide duplexes with parallel and antiparallel chain orientation. *J. Am. Chem. Soc.* **127**, 7739–7751 (2005).
37. Eberlein, L. et al. Tautomeric equilibria of nucleobases in the Hachimoji expanded genetic alphabet. *J. Chem. Theory Comput.* **16**, 2766–2777 (2020).
38. Bande, O. et al. Isoguanine and 5-methyl-isocytosine bases, in vitro and in vivo. *Chemistry* **21**, 5009–5022 (2015).
39. Hashimoto, K., Fischer, E. C. & Romesberg, F. E. Efforts toward further Integration of an unnatural base pair into the biology of a semisynthetic organism. *J. Am. Chem. Soc.* **143**, 8603–8607 (2021).
40. Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M. & Yusupova, G. A new understanding of the decoding principle on the ribosome. *Nature* **484**, 256–259 (2012).
41. Rozov, A., Demeshkina, N., Westhof, E., Yusupov, M. & Yusupova, G. Structural insights into the translational infidelity mechanism. *Nat. Commun.* **6**, 7251 (2015).
42. Demeshkina, N., Jenner, L., Westhof, E., Yusupov, M. & Yusupova, G. New structural insights into the decoding mechanism: translation infidelity via a G.U pair with Watson-Crick geometry. *FEBS Lett.* **587**, 1848–1857 (2013).
43. Rozov, A. et al. Tautomeric G*U pairs within the molecular ribosomal grip and fidelity of decoding in bacteria. *Nucleic Acids Res.* **46**, 7425–7435 (2018).
44. Yang, Z., Hutter, D., Sheng, P., Sismour, A. M. & Benner, S. A. Artificially expanded genetic information system: a new base pair with an alternative hydrogen bonding pattern. *Nucleic Acids Res.* **34**, 6095–6101 (2006).
45. Hoshika, S., Shukla, M. S., Benner, S. A. & Georgiadis, M. M. Visualizing "alternative isoinformational engineered" DNA in A- and B-Forms at high resolution. *J. Am. Chem. Soc.* **144**, 15603–15611 (2022).
46. Twist, K. A. et al. A novel method for the production of in vivo-assembled, recombinant Escherichia coli RNA polymerase lacking the alpha C-terminal domain. *Protein Sci.* **20**, 986–995 (2011).
47. Chen, J., Noble, A. J., Kang, J. Y. & Darst, S. A. Eliminating effects of particle adsorption to the air/water interface in single-particle cryo-electron microscopy: bacterial RNA polymerase and CHAPSO. *J. Struct. Biol. X* **1**, 100005 (2019).
48. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

49. Thompson, R. F., Iadanza, M. G., Hesketh, E. L., Rawson, S. & Ranson, N. A. Collection, pre-processing and on-the-fly analysis of data for high-resolution, single-particle cryo-electron microscopy. *Nat. Protoc.* **14**, 100–118 (2019).
50. Suloway, C. et al. Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151**, 41–60 (2005).
51. Cheng, A. et al. Leginon: new features and applications. *Protein Sci.* **30**, 136–150 (2021).
52. Tan, Y. Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).
53. Kimanius, D., Dong, L., Sharov, G., Nakane, T. & Scheres, S. H. W. New tools for automated cryo-EM single-particle analysis in RELION-4.0. *Biochem. J.* **478**, 4169–4185 (2021).
54. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 A reconstruction of rotavirus VP6. *Elife* **4**, e06980 (2015).
55. Tegunov, D. & Cramer, P. Real-time cryo-electron microscopy data preprocessing with Warp. *Nat. Methods* **16**, 1146–1152 (2019).
56. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
57. Harauz, G. & van Heel, M. Exact filters for general geometry three dimensional reconstruction. *Optik* **73**, 146–156 (1986).
58. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**, e42166 (2018).
59. Terwilliger, T. C., Sobolev, O. V., Afonine, P. V., Adams, P. D. & Read, R. J. Density modification of cryo-EM maps. *Acta Crystallogr. D. Struct. Biol.* **76**, 912–925 (2020).
60. Terwilliger, T. C. et al. Improvement of cryo-EM maps by density modification. *Nat. Methods* **17**, 923–927 (2020).
61. Baldwin, P. R. & Lyumkis, D. Non-uniformity of projection distributions attenuates resolution in Cryo-EM. *Prog. Biophys. Mol. Biol.* **150**, 160–183 (2020).
62. Baldwin, P. R. & Lyumkis, D. Tools for visualizing and analyzing Fourier space sampling in Cryo-EM. *Prog. Biophys. Mol. Biol.* **160**, 53–65 (2021).
63. Pettersen, E. F. et al. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
64. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 213–221 (2010).
65. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 486–501 (2010).
66. Grant, T., Rohou, A. & Grigorieff, N. cisTEM, user-friendly software for single-particle image processing. *Elife* **7**, e35383 (2018).
67. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
68. Schrodinger, L. L. C. *The PyMOL Molecular Graphics System, Version 1.8* (2015).

Acknowledgements

This work was supported by grants from the National Institutes of Health (R01 GM102362 and R01 GM148476 to D.W., R21CA251043 to S.H., P30 CA01495 and U54 AI170855 to D.L.), from the National Science Foundation (MCB-2048095 to D.L. and MCB-2123995 to S.A.B.). D.L. also acknowledges support from the Hearst Foundations and the Margaret T. Morris Foundation. The cryo-EM data were collected at the cryo-EM S2C2 facility of SLAC National Accelerator Laboratory, which is supported by S2C2 grant U24 GM129541 and at Scripps Research (supported by S10 OD032467).

Author contributions

S.A.B., D.L. and D.W. designed research; J.O., Z.S., S.H., J.X. and J.C. performed research; J.O., Z.S., S.H., S.A.B., D.L. and D.W. analyzed data; and J.O., Z.S., S.H., S.A.B., D.L. and D.W. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-43735-9.

Correspondence and requests for materials should be addressed to Steven A. Benner, Dmitry Lyumkis or Dong Wang.

Peer review information *Nature Communications* thanks Seth Darst, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.