# Structural basis for expansion of the genetic alphabet by the artificial base pair dDs-dPx

Karin Betz[†], Michiko Kimoto[‡,§], Kay Diederichs[†], Ichiro Hirao[‡,§*] and Andreas Marx[†*]

[†] Departments of Chemistry and Biology, and Konstanz Research School Chemical Biology, University of Konstanz, Universitätsstrasse 10, 78457 Konstanz, Germany

[‡] Institute of Bioengineering and Nanotechnology, 31 Biopolis Way, The Nanos, #09-01, 138669 Singapore

[§] RIKEN Center for Life Science Technologies, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Supporting Information Placeholder

ABSTRACT: Hydrophobic artificial base pairs without the ability to pair via hydrogen bonds are the most promising candidates to date to expand the genetic alphabet of DNA and RNA. Pairs have been developed that are efficiently replicated and transcribed *in vitro* and already proved successful in different molecular biological applications. In spite of the success of the last years, there is still need and room for further optimization of the candidates. Structural characterization of the pairing behavior and the processes that are crucial to use the non-cognate substrates in various applications is necessary for the design of optimized structures or tailored enzymes. Here we characterize the insertion of one of the most promising hydrophobic unnatural base pairs; the **dDs–dPx** pair into a DNA strand by a DNA polymerase. We solved a crystal structure of KlenTaq DNA polymerase with a modified template/primer duplex plus the unnatural substrate. The ternary complex shows that the artificial pair adopts a planar structure just like a canonical base pair in the active site and that the enzyme is able to close and adopt a state prior to catalysis. Analyzing the enzyme environment revealed that the O-helix cannot be as tightly closed as it is the case with natural substrates which would explain the still lower incorporation efficiency and identifies further optimization potential.

Expanding the genetic alphabet with an unnatural base pair increases the functional diversity of nucleic acids and with this the biological and biotechnological scope of applications. Besides hydrogen-bonding base pairs with bonding patterns and structures different to the natural A–T and G–C pairs[1-7], hydrophobic "nucleobases" that rely on hydrophobic and packing interactions turned out to be able to selectively pair with each other and therefore are candidates to generate a third base pair[8-10]. Their main advantage is lower mispairing propensity with the four natural nucleotides due to lacking H-bonds. Most prominent in this respect are the pairs between 2-methoxynaphthalene (**NaM**) and 6-methyl-2H-isoquinoline-1-thione (**5SICS**) or the improved **NaM-TPT3** (thieno[2,3-*c*]pyridine-7(6H)-thione) pair developed in the Romesberg group[11, 12, 13, 14], and the pair between 7-(2-thienyl)imidazo[4,5-*b*]pyridine (**Ds**) and 2-nitro-4-propynylpyrrole (**Px**) pair introduced by Hirao and coworkers[15-18] (Figure 1).

The later analogues were used in the generation of high-affinity DNA aptamers containing natural and unnatural nucleotides by SELEX.[19-22] The use of the unnatural base pair surrogate **dDs–dPx** increased the chemical and structural diversity of the DNA libraries used, resulting in an improved selectivity and binding affinity. Further, a first post-transcriptional modification method for RNA transcripts using an expanded genetic alphabet was presented by using PCR-amplified DNA templates with the **dDs–dPx** pair.[23] Key for the success of these approaches is that the unnatural base pair is preserved besides the natural ones during the enzymatic synthesis.

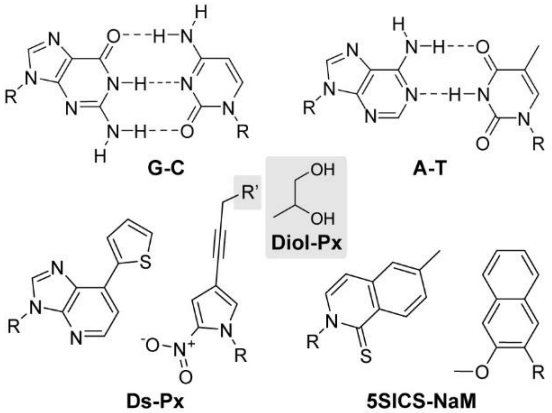



**Figure 1.** Structures of the natural base pairs and two successful hydrophobic artificial base pairs. R = 2′-deoxyribose, R' can be different functional groups like e.g. the diol functionality displayed in the grey box and used in this study.

The base pair surrogate **dDs–dPx** shows little similarity with the d**NaM**-d**5SICS** nor to the natural nucleobase pairs. It has been evolved over several screening rounds for efficiency and selectivity in replication.[15-18] The nitro group of **dPx** turned out to be crucial. First, it is believed to enable the direct or water-mediated minor groove interaction to the polymerase.[16, 18] Second, it prevents the mispairing with dA, resulting in an increase in selectivity.[16] One advantage of this base pair is that a variety of functionalities for site-specific labeling can be introduced at position 4 of the d**Px** base via a linker (grey box in Figure 1) without affecting significantly the replication efficiency.[24, 25] These groups allow site-specific introduction of larger functional groups such as dyes into DNA

strands by post-synthetic labelling. The best pairing partner of **dDs** is the dihydroxy derivative of **dPx**[24] which might allow further derivatization via Schiff base formation. The obvious structural dissimilarity of the **dDs–dPx** pair with the natural nucleobase pairs imposes the question how these analogs are processed by DNA polymerases. Up to now, only structural data of the d**NaM**–d**5SICS** base pair surrogate in the confines of a DNA polymerase (i.e., the large fragment of *Thermus aquaticus* DNA polymerase – short: KlenTaq) is available.[26, 27] This data indicates that the enzyme is forcing the unnatural base pair into a co-planar conformation when the triphosphate encounters the unnatural template thereby resolving stacking interactions that are found in a free DNA duplex.[26, 28]

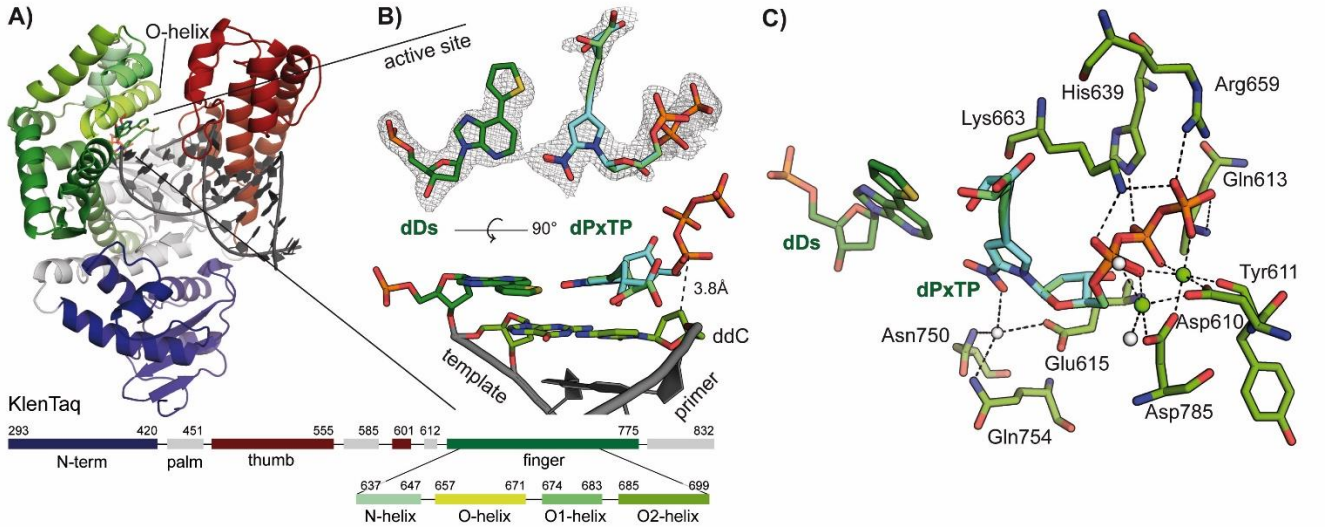



**Figure 2**. The artificial base pair **dDs–dPxTP** is pairing in an edge-to-edge manner in the active site of KlenTaq and shows the same interactions with the enzyme and catalytic ions as a natural base pair. A) Overall structure of KlenTaq (domains coloured as indicated in the figure) with bound DNA (grey) and substrate. B) **dDs–dPxTP** pair surrounded by its simulated annealing mFo-DFc omit map contoured at 3σ. The artificial templating **dDs** is shown in dark green, the R and S isomers of **dPxTP** are shown in bluecyan and greencyan, respectively C) Interactions of the triphosphate **dPxTP** with the enzyme. Magnesium ions are shown in green and water molecules are shown in white.

Here we present structural data of KlenTaq while processing the unnatural base pair **dDs–dPxTP** (of note: Diol-**dPxTP** was used in this study).

To obtain the structure, we followed a similar strategy as before.[26] For crystallization, KlenTaq was incubated with an annealed natural primer and a **dDs**-containing template as well as ddCTP to terminate the primer strand. Crystals were grown and afterwards soaked with **dPxTP**, resulting in the structure KlenTaq$_{\mathbf{dDs–dPxTP}}$. Crucial for successful soaking was the absence of high concentrations of magnesium ions. Therefore, we changed the crystallization conditions by replacing magnesium acetate or magnesium formate, which initially yielded the best crystals, with sodium acetate. We assume that at high concentrations, magnesium ions can engage complexes with the diol moiety of **dPxTP** and therefore influence the crystallization properties. The structure was solved and data up to a resolution of 1.7 Å was used in refinement (see methods and Table S1 in the supplementary information).

The overall structure of KlenTaq$_{\mathbf{dDs–dPxTP}}$ with its four domains (N-terminal, palm, thumb and finger, see Figure 2A) is similar to the natural ternary complex KlenTaq$_{\mathrm{dG–dCTP}}$ (PDB ID: 3RTV[26]) and to KlenTaq$_{\mathbf{dNaM–d5SICSTP}}$ (PDB ID: 3SV3[26]) with rmsds for Cα atoms of only 0.185Å (446 atoms aligned, 39 atoms rejected) and 0.173Å (442 atoms aligned, 91 atoms rejected), respectively. All enzyme residues could be traced, but the region of the finger domain between residues 645 and 700 was less well resolved and shows higher flexibility than the rest of the enzyme (Figure S2 A, D). The finger domain is in a closed conformation and the **dDs–dPxTP** pair is situated in the active site (Figure 2). The base pair could be unambiguously placed into the difference electron density map after initial refinement without it. All functionalities are well defined, only the electron density at the diol moiety of the triphosphate is somehow extended. This can be explained by the rotational flexibility of the terminal hydroxyl group on the one hand but also by the presence of two epimers (R or S configuration at the diol) of **dPxTP** on the other. Both isomers were modeled at the same position and refined to occupancies of 0.40 and 0.46 for the S and R isomer, respectively. Thereby the pyrrole, the ribose and the triphosphate part of the ligand superpose perfectly. The ribose moieties of the triphosphate and the templating **dDs** adopt C3′-*endo* conformations identical to the natural pair. Coordination of two magnesium ions by the triphosphate and Asp610, Asp785 and the backbone of Tyr611 characterizes an active closed complex (Figure 2C). The distance between the primer 3′-end (C3′ used for measuring as 3′-OH group is not present) and the alpha phosphate is almost identical to the natural case (3.8Å vs. 3.9Å). In addition to the metal coordination, the triphosphate is interacting with the side chains of Lys663, Arg659, His639 and the backbone of Gln613 whereby it can be well stabilized at its position. A water-mediated minor groove interaction (in the natural case mediated by O2 atoms of pyrimidines or N3 atoms of purines) is formed between the nitro group of **dPxTP** and Asn750, Gln754 as well as Glu615. Finally, the **dDs–dPxTP** pair does not intercalate but pairs in a co-planar manner strikingly similar to the cognate Watson-Crick pairs (Figure 3). Thereby even the propeller twist (relative rotation between bases within a base pair with respect to the base pairing axis) is similar between **dDs–dPxTP** and dG–dCTP (-10,5° versus -9.2°, Figure 3A and S3; as determined by use of the 3DNA server[29]).

This similarity was unexpected as the previously analyzed hydrophobic **dNaM–d5SICSTP** pair shows some differences in this regard (discussed below). Relative to the natural dG–dCTP pair, the unnatural **dDs–dPxTP** pair shows a larger overall base pair width (C1′–C1′ distance between the pairing partners is 0.7Å longer, Figure 3B). The shift is taking place only on the side of the templating nucleotide whereby the substrate triphosphate stays at its well-defined position just described. Amino acids interacting with the templating nucleotide (Arg677, Ser674 and Met673) are also slightly shifted (Figure 3D), following the movement of the template. Not only the width but also the height of the base pair is different to the cognate pairs. The thiophenyl as well as the propynyl-diol moieties point towards the O-helix of the finger domain. This causes a small shift of the overall O-helix and adjacent and different orientation of some amino acid side chains. The effect is pronounced for Thr664 which comes closest to the thiophenyl moiety and shifts upwards by 0.6Å. The extent of the displacements of the Cα- atoms along the O-helix is shown in detail in Figure S3 D. Along with these shifts, the flexibility of the entire finger domain and also the newly formed base pair is higher than in the natural case indicated by elevated B-factors (Figure S2 A–C). Arg660 makes space for the propynyl-diol moiety and adopts a position where it may interact with one of the OH groups via a hydrogen bond (Figure 3D). Its flexibility, however, indicates that the interaction cannot be strong in the present state. Similar shifts of Arg660 have already been found in other KlenTaq structures with modified substrates and this displacement seems not to diminish incorporation efficiency.[30, 31] All in all, the increased height of the pair only affects enzyme residues on the major groove side of the pair whereas the minor groove side stays unperturbed.

The previously analyzed hydrophobic **dNaM–d5SCISTP** pair also pairs edge-to-edge in the active site of KlenTaq as mentioned above. However, in contrast to the pairs dG–dCTP and **dDs–dPxTP**, the propeller twist is significantly smaller (2.4°, Figure 3A and S3) and the pair exhibits a larger relative shift of the bases along the Z-axis (stagger = -1.2Å) which is zero for dG–dCTP and only -0.12Å for **dDs–dPxTP** (Figure S3). These two base pair parameters have not been discussed previously when comparing the natural pair dG–dCTP with the hydrophobic artificial base pair **dNaM–d5SICSTP**. Together with the new structure KlenTaq$_{\text{dDs–dPxTP}}$ this unapparent difference between the two unnatural base pair surrogates in relation to the natural pair was identified.

A similarity between the two hydrophobic base pairs is their increased distance between the pairing partners (Figure 3B). Also for **dNaM–d5CISTP**, the C1′–C1′ distance is elevated (11.0Å vs. 10.6Å in the dG–dCTP case) (Figure S1A). The residues around the templating nucleotide are even shifted a bit more compared to KlenTaq$_{\text{dDs–dPxTP}}$ due to the different nucleobase structure (Figure S1B). It seems that if an artificial pair with elevated size forms in the active site of KlenTaq the enzyme allows extension to one side of the binding pocket (the template side) but not the other. This behavior might assist the enzyme to align the triphosphate substrate perfectly for a proper attack by the 3′-OH group of the primer end, although the pair shows some alterations from the natural consensus structure.

Along with the shift on the template side, a similar O-helix shift as described for KlenTaq$_{\text{dDs–dPxTP}}$ was found in the KlenTaq$_{\text{dNaM–d5SICSTP}}$ structure. In both cases, the O-helix is closing far enough so that the substrate as well as other components are well aligned for catalysis.

A recent report featuring a modelled **dDs–dPx** base pair in the active state of Deep Vent DNA polymerase, which is among the most proficient polymerases that recognize the **dDs–dPx** pair with high fidelity in replication[24], proposes the right accommodation of the unnatural base pair in the polymerase[32]. In contrast, modelling the artificial pair into the active site of a KlenTaq ternary structure revealed that here the side-chain oxygen of Thr664 clashes the templating **dDs** base.[32] Indeed, this residue in the O-helix of the finger domain is the one coming closest to the unnatural pair, but our structural data indicates that the enzyme is flexible enough to anyway accommodate the base pair and adapt to the structure. The described shift and elevated B-factors of the O-helix and other parts of the finger domain (see Figure S2), however, indicate that the O-helix does not close as tightly as with a cognate pair. This small difference could be an explanation why the two most successful hydrophobic artificial pairs are still incorporated with somewhat diminished efficiency compared to the natural counterparts. The perfect arrangement of components taking part in catalysis might be slightly changed in one or several steps during insertion. Modulating the size of the pairs towards the O-helix or mutating the enzyme in this region might provide opportunity for further optimization.

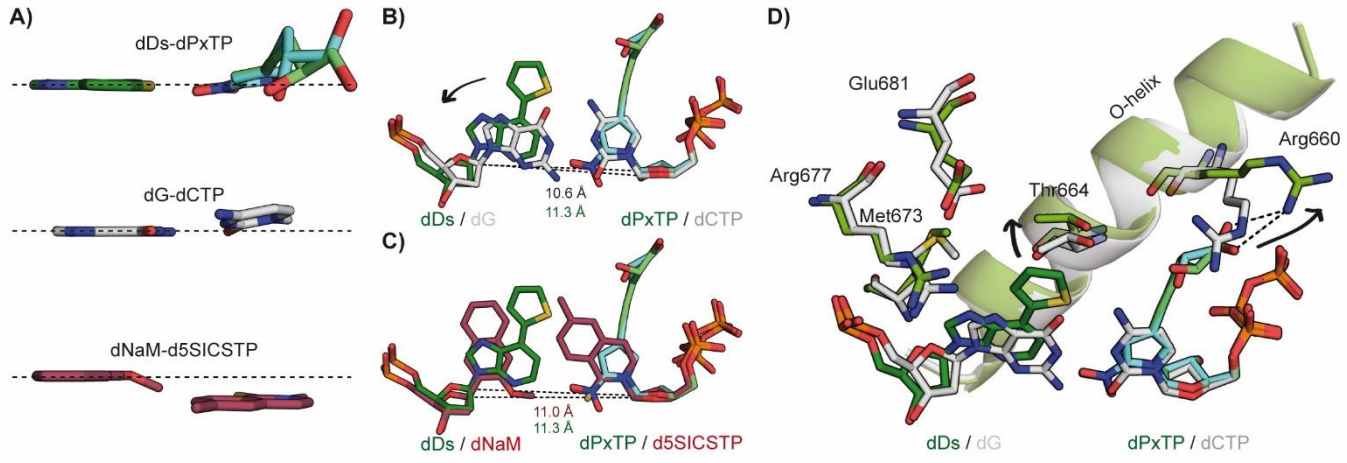


**Figure 3:** Comparison of **dDs–dPxTP** with the natural dG–dCTP pair and the unnatural **dNaM–d5SICSTP** pair in KlenTaq. A) Base pair propeller twist visualized with the templating nucleobase aligned orthogonal to the paper plane (ribose and triphosphate moieties are not shown). B) Relative position of the base pairs when the whole enzyme complex is superposed. C1'-C1' distances between the pairing partners are indicated C) Shifts of residues surrounding the artificial pair are visualized. The finger domain O-helix is shown as cartoon.

One of the features of the **dPx** nucleobase is its modifiability with different functional groups. The linker moiety of **dPxTP** is placed in a way that the modification can leave the enzyme through two possible channels as it was shown before for C5-modified deoxyuridines and C7-modified 7-deaza-adenosines in KlenTaq.[30, 33, 31] (The data in Table 2 of ref ((24) show the side-chain dependency for the replication efficiency, in which large side-chains reduced the efficiency).

In conclusion, herewith we provide the structural basis for the expansion of the genetic alphabet by the **dDs–dPxTP** pair. The data show how the artificial base pair is well accepted in the active site of KlenTaq by pairing edge-to-edge just like the natural pairs. This is in consistency with previous observations with the artificial base pair **dNaM–d5SICSTP**. Compared to **dNaM–d5SICSTP** which pairs almost co-planar, the **dDs–dPxTP** pair adopts a similar propeller twist as the natural pair. Both unnatural pairs exhibit an increased base pair width and different height compared to the natural counterparts. KlenTaq has some flexibility to adapt to the slightly increased height and width of **dDs–dPxTP** as well as **dNaM–d5SICSTP**. While the triphosphate and the minor groove side of the pair are perfectly aligned in the active site, changes in width and height are transferred to the side of the templating nucleotide and the O-helix. With both pairs, the O-helix is not as tightly closed as in the natural case, and the resulting small changes in positions may explain the still lower incorporation efficiency of the artificial substrates. Therefore, the size of the pairs in this direction as well as residues in the O-helix are suitable targets for further modification and improvement.

Our results again confirm that hydrophobic base pairs are well suited to expand the genetic alphabet as long as they can adopt similar shape as the cognate pairs and match the constrains of the active sites of the respective enzymes. Furthermore, it seems that base pair parameters (like propeller twist) can still vary between different hydrophobic base pairs candidates. We assume that the exact pairing behavior in the active site of a DNA polymerase is difficult to predict directly from the base pair architecture, rendering structural studies indispensable to further characterize future artificial base pairs for the expansion of the genomic alphabet.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

Material and Methods; additional figures (Figures S1–S3) and crystallographic data collection and refinement statistics (Table S1). The structure KlenTaq$_{\text{dDs-dPxTP}}$ has been deposited in the Protein Data Bank under the accession code 5NKL.

## AUTHOR INFORMATION


### Corresponding Author

* Andreas.Marx@uni-konstanz.de

* ichiro@ibn.a-star.edu.sg


### Notes

The authors declare no competing financial interests.

## ACKNOWLEDGMENT


We thank the beamline staff of the Swiss Light Source at the Paul Scherrer Institute for access and help at the beamline. We also thank the Konstanz Research School Chemical Biology (K.B.) and the Japan Science and Technology Agency, Precursory Research for Embryonic Science and Technology (JPMJPR13K9) (M.K.) for financial support. A part of this work was funded by the Institute of Bioengineering and Nanotechnology (Biomedical Research Council, Agency for Science, Technology and Research, Singapore).


## REFERENCES


(1) Switzer, C.; Moroney, S. E.; Benner, S. A. *J. Am. Chem. Soc.* **1989**, *111*, 8322–8323.

(2) Piccirilli, J. A.; Krauch, T.; Moroney, S. E.; Benner, S. A. *Nature* **1990**, *343*, 33–37.

(3) Ishikawa, M.; Hirao, I.; Yokoyama, S. *Tetrahedron Lett.* **2000**, *41*, 3931–3934.

(4) Fujiwara, T.; Kimoto, M.; Sugiyama, H.; Hirao, I.; Yokoyama, S. *Bioorganic & medicinal chemistry letters* **2001**, *11*, 2221–2223.

(5) Hirao, I.; Harada, Y.; Kimoto, M.; Mitsui, T.; Fujiwara, T.; Yokoyama, S. *Journal of the American Chemical Society* **2004**, *126*, 13298–13305.

(6) Yang, Z.; Hutter, D.; Sheng, P.; Sismour, A. M.; Benner, S. A. *Nucleic Acids Res.* **2006**, *34*, 6095–6101.

(7) Tarashima, N.; Komatsu, Y.; Furukawa, K.; Minakawa, N. *Chemistry - A European Journal* **2015**, *21*, 10688–10695.

(8) Kool, E. T. *Annu. Rev. Biochem.* **2002**, *71*, 191–219.

(9) McMinn, D. L.; Ogawa, A. K.; Wu, Y.; Liu, J.; Schultz, P. G.; Romesberg, F. E. *J. Am. Chem. Soc.* **1999**, *121*, 11585–11586.

(10) Mitsui, T.; Kitamura, A.; Kimoto, M.; To, T.; Sato, A.; Hirao, I.; Yokoyama, S. *J. Am. Chem. Soc.* **2003**, *125*, 5298–5307.

(11) Seo, Y. J.; Hwang, G. T.; Ordoukhanian, P.; Romesberg, F. E. *J. Am. Chem. Soc.* **2009**, *131*, 3246–3252.

(12) Lavergne, T.; Malyshev, D. A.; Romesberg, F. E. *Chem Eur J* **2012**, *18*, 1231–1239.

(13) Li, L.; Degardin, M.; Lavergne, T.; Malyshev, D. A.; Dhami, K.; Ordoukhanian, P.; Romesberg, F. E. *Journal of the American Chemical Society* **2014**, *136*, 826–829.

(14) Zhang, Y.; Lamb, B. M.; Feldman, A. W.; Zhou, A. X.; Lavergne, T.; Li, L.; Romesberg, F. E. *Proceedings of the National Academy of Sciences* **2017**, 201616443.

(15) Hirao, I.; Kimoto, M.; Mitsui, T.; Fujiwara, T.; Kawai, R.; Sato, A.; Harada, Y.; Yokoyama, S. *Nat. Methods* **2006**, *3*, 729–735.

(16) Hirao, I.; Mitsui, T.; Kimoto, M.; Yokoyama, S. *Journal of the American Chemical Society* **2007**, *129*, 15549–15555.

(17) Kimoto, M.; Kawai, R.; Mitsui, T.; Yokoyama, S.; Hirao, I. *Nucleic Acids Research* **2009**, *37*, e14.

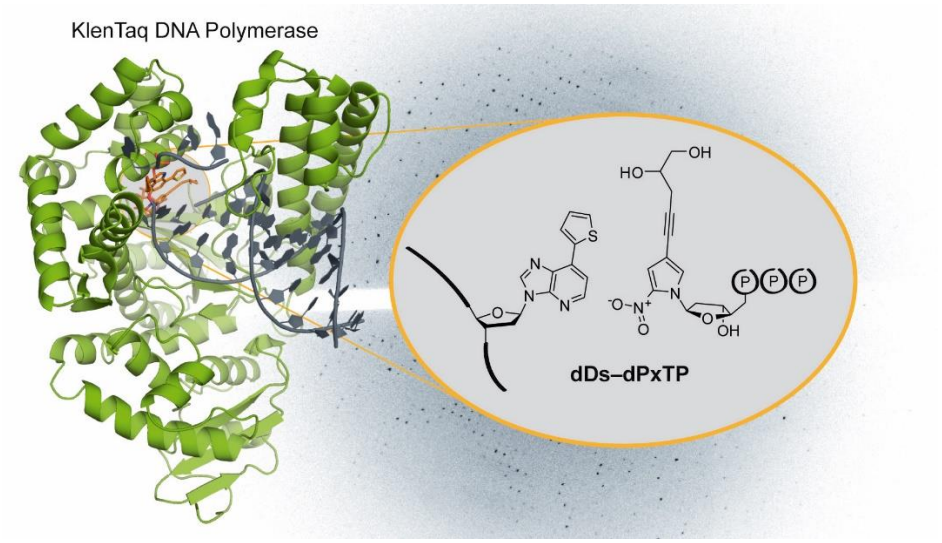
(18) Hirao, I.; Kimoto, M.; Yamashige, R. *Acc. Chem. Res.* **2012**, *45*, 2055–2065.

(19) Kimoto, M.; Yamashige, R.; Matsunaga, K.-i.; Yokoyama, S.; Hirao, I. *Nat. Biotechnol.* **2013**.

(20) Matsunaga, K.-i.; Kimoto, M.; Hanson, C.; Sanford, M.; Young, H. A.; Hirao, I. *Scientific Reports* **2015**, *5*, 18478.

(21) Kimoto, M.; Nakamura, M.; Hirao, I. *Nucleic Acids Research* **2016**, *44*, 7487–7494.

(22) Matsunaga, K.-i.; Kimoto, M.; Hirao, I. *Journal of the American Chemical Society* **2017**, *139*, 324–334.
(23) Someya, T.; Ando, A.; Kimoto, M.; Hirao, I. *Nucleic Acids Research* **2015**, *43*, 6665–6676.
(24) Yamashige, R.; Kimoto, M.; Takezawa, Y.; Sato, A.; Mitsui, T.; Yokoyama, S.; Hirao, I. *Nucleic Acids Res.* **2012**, *40*, 2793–2806.
(25) Okamoto, I.; Miyatake, Y.; Kimoto, M.; Hirao, I. *ACS Synthetic Biology* **2016**, 1220–1230.
(26) Betz, K.; Malyshev, D. A.; Lavergne, T.; Welte, W.; Diederichs, K.; Dwyer, T. J.; Ordoukhanian, P.; Romesberg, F. E.; Marx, A. *Nat. Chem. Biol.* **2012**, *8*, 612–614.
(27) Betz, K.; Malyshev, D. A.; Lavergne, T.; Welte, W.; Diederichs, K.; Romesberg, F. E.; Marx, A. *J. Am. Chem. Soc.* **2013**, *135*, 18637–18643.
(28) Malyshev, D. A.; Pfaff, D. A.; Ippoliti, S. I.; Hwang, G. T.; Dwyer, T. J.; Romesberg, F. E. *Chem Eur J* **2010**, *16*, 12650–12659.
(29) Lu, X.-J.; Olson, W. K. *Nature Protocols* **2008**, *3*, 1213–1227.
(30) Bergen, K.; Steck, A.-L.; Strütt, S.; Baccaro, A.; Welte, W.; Diederichs, K.; Marx, A. *J. Am. Chem. Soc.* **2012**, *134*, 11840–11843.
(31) Hottin, A.; Betz, K.; Diederichs, K.; Marx, A. *Chemistry - A European Journal* **2017**.
(32) Hikida, Y.; Kimoto, M.; Hirao, I.; Yokoyama, S. *Biochemical and Biophysical Research Communications* **2017**, *483*, 52–57.
(33) Hottin, A.; Marx, A. *Acc. Chem. Res.* **2016**, *49*, 418–427.




Insert Table of Contents artwork here