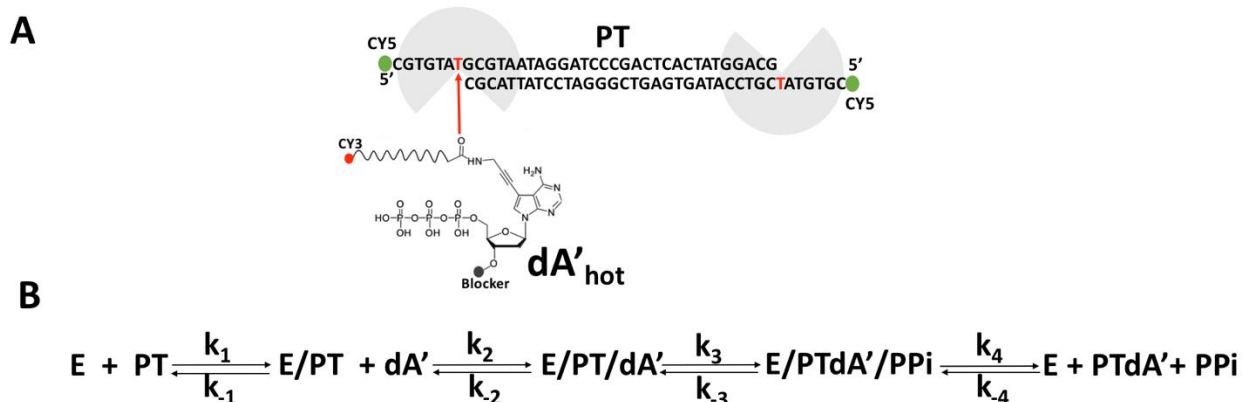


Rational evolution of a recombinant DNA polymerase for efficient incorporation of unnatural nucleotides by dual-site boosting: Supplementary Information

Supplementary Schemes



Scheme S1. Reaction scheme of incorporation of modified dATP by KOD. Panel A depicts the molecular structures of CY5-labelled DNA duplex (denoted as PT) and modified dATP (denoted as dA'_{hot}) with a reversible 3-O' blocker attached and a CY3-label attached through a long linker; Panel B shows the reaction scheme: E/PT represents the binary complex formed between the enzyme (E) and DNA (PT); E/PT/dA' represents the ternary complex formed by enzyme, DNA and modified dATP before the formation of the Pa-O3' covalent bond (chemical step); E/PTdA'/PPI represents the ternary complex formed by the enzyme, dA'-bound DNA duplex and pyrophosphate following the chemical step; PTdA' represent the dA'-bound DNA duplex which dissociates from the enzyme.

Supplementary Figures

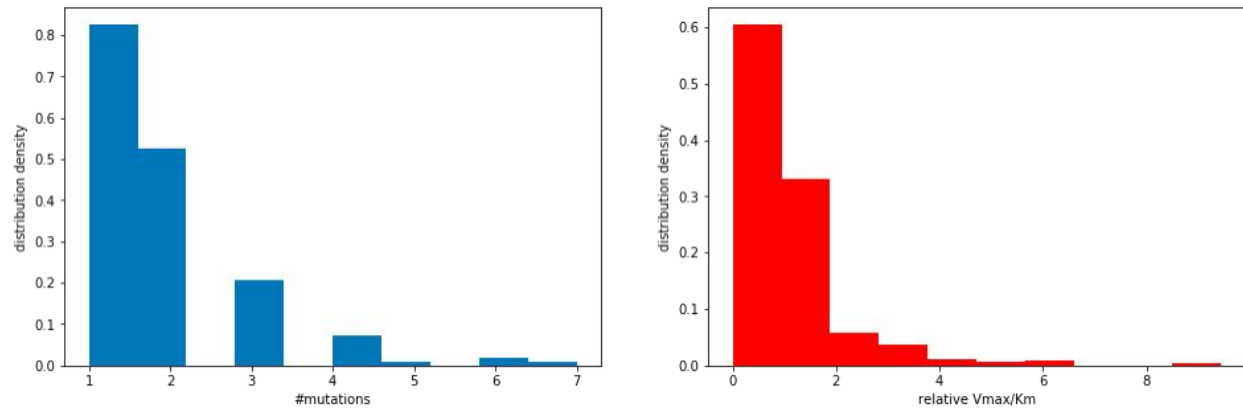


Figure S1. left panel: normalized distributions of the number of mutation sites with respect to parent KOD (RF) in training data set; right panel: normalized distribution of the experimentally measured relative $V_{\max}/K_m$ ($E(\text{mut})$) values of the training data.

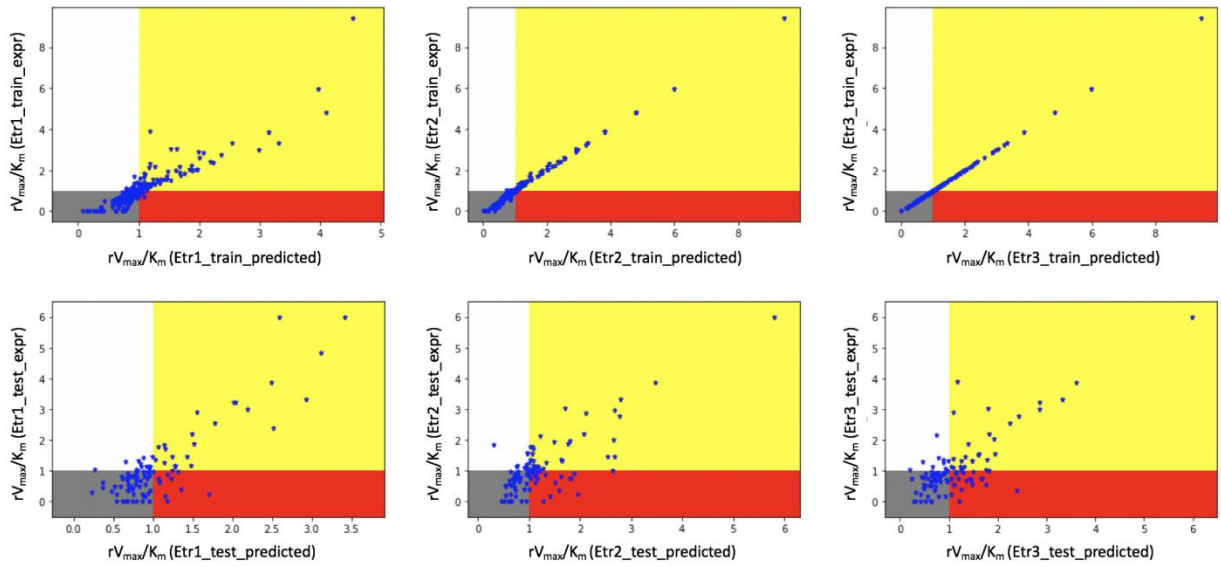


Figure S2. Extra tree regression modelling of KOD $rV_{max}/K_m(E(mut))$. For each plot, predicted and experimentally tested rV_{max}/K_m of samples were respectively shown along the X-axis and Y-axis. For each built extra tree regressor (Etr 1-3), model performances on training set and testing set were shown in the upper and lower plots arranged in the same column. Model evaluation metric R^2 score and accuracy rate on training and test samples are shown in Table S3.

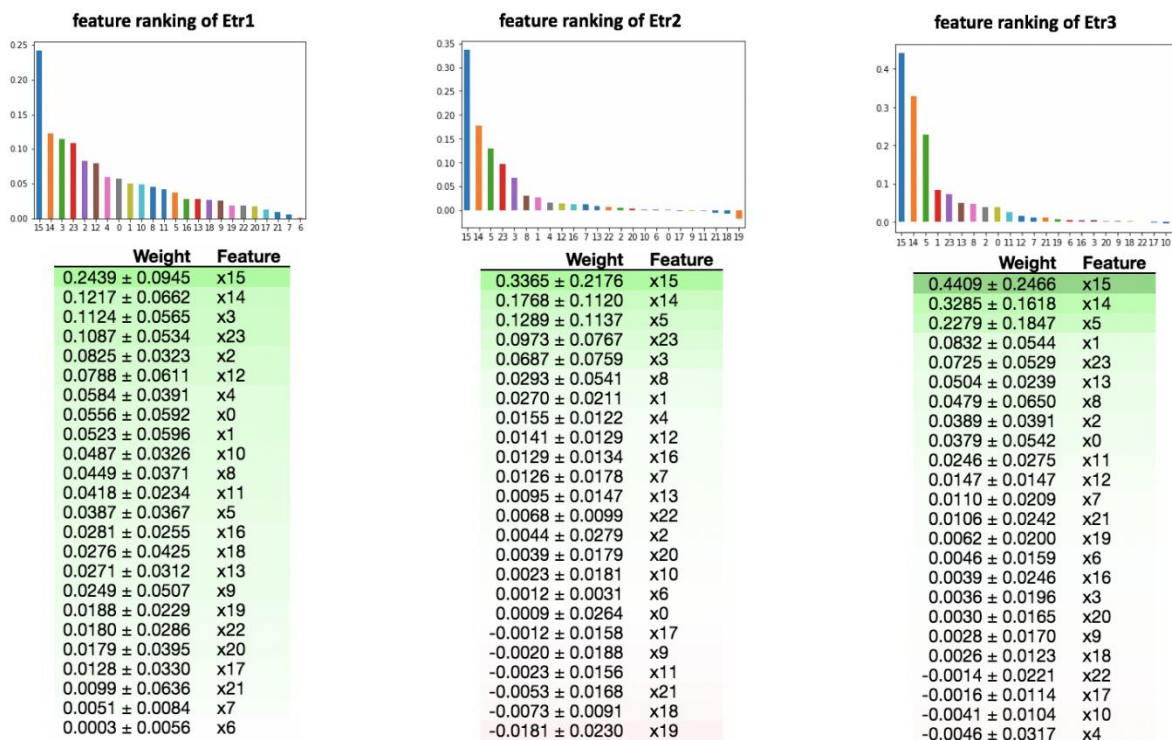


Figure S3. Feature ranking of three extra tree regressor Etr 1-3. 24 For each KOD sequence, 24 features were calculated for extra tree regressor building. They were indexed from x0 to x23, respectively corresponding to 'PI', 'Flexibility', 'Charge', 'Aliphaticness', 'TotalEntropy', 'Instability', 'aromaticity', 'gravy', 'Mutabilityscale', 'Volumescale', 'Polarizabilityscale', 'jascale', 'ASAInTripeptidescale', 'wholeseq_blosum80', 'finger_blosum80', 'thumb_blosum80', 'w_hsasa', 'f_hsasa', 't_hsasa', 'w_bf', 'f_bf', 't_bf', 'md12_pd_deltaG', 'provean_score'.

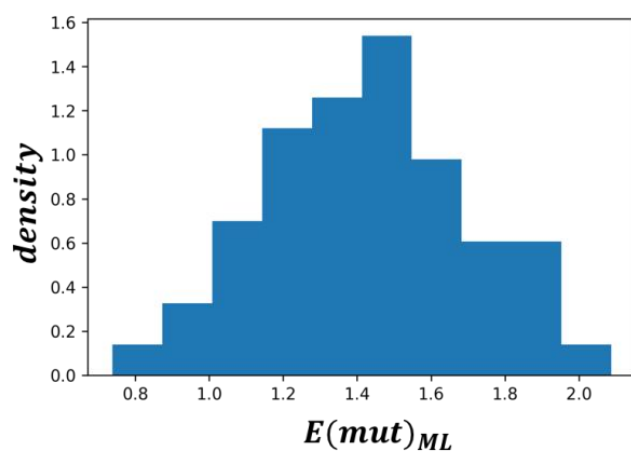


Figure S4. Distribution of predicted $E(\text{mut})$ values for the variants carrying designed double mutations located on the finger subdomain.

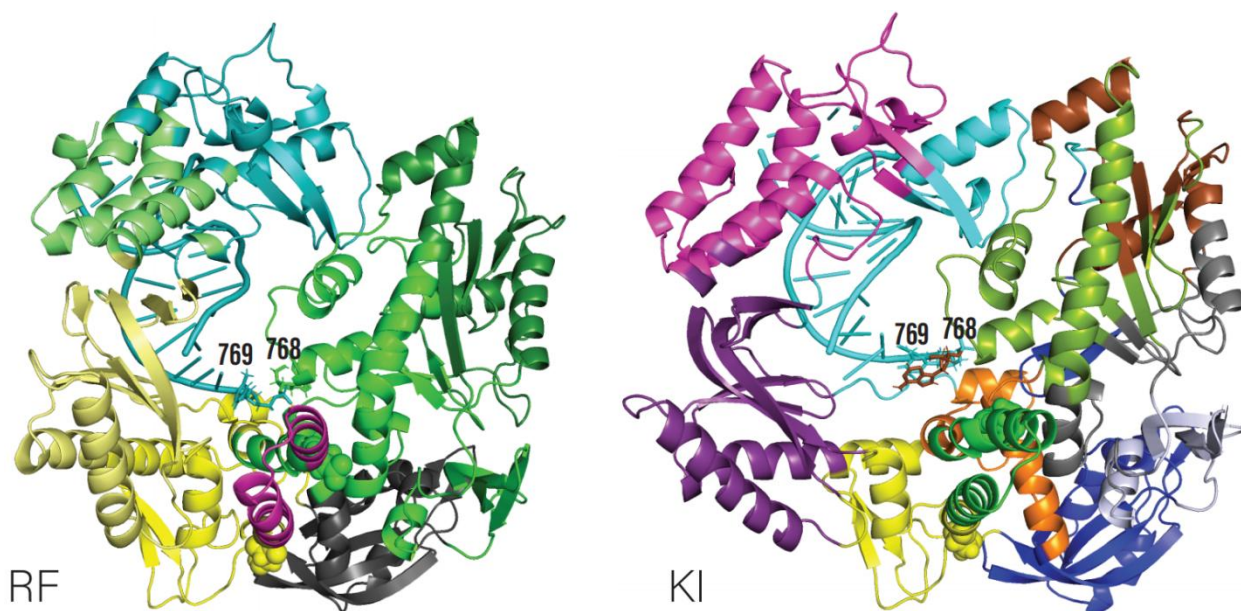


Figure S5. The dynamical cooperative groups (rendered by different colors) of binary RF (left) and KI(right) show that the nucleotide 768 of template DNA is uncooperative to other nucleotides including the core of the binding pocket (769). The two mutation sites 451 and 485 are highlighted by sphere mode.

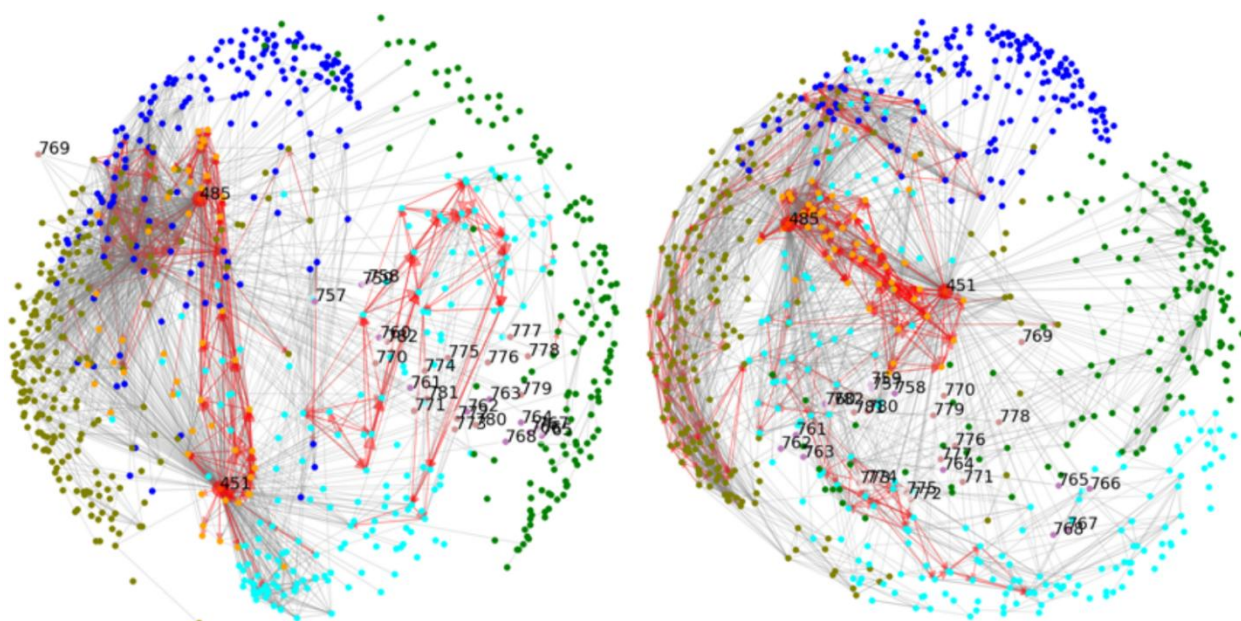


Figure S6. Superimposed pathways of all maxima flows steaming from a.a. 451 and 485 (labeled) to the DNA duplex (labeled) of RF (left panel) and KI (right panel). Here each vertex represents a residue or nucleotide, and each edge represents the mean flow between two vertexes. The edges with weightings greater than 0.5 are highlighted in red, between 0.5 and 0.1 are rendered in grey, otherwise are not displayed for clarity. Note the relative position of the vertexes are assigned according to the reciprocal of pairwise correlation *via* multidimensional scaling.

Supplementary Tables

Table S1. Comparison of substrate-specific kinetic parameters among RF, KI and KH.

Pure enzyme	V_{max}	K_m	$E(mut)$
RF	476.00	2.59	1.00
KI	750.00	1.55	2.63
KH	712.60	0.21	18.46

Table S2. Enzyme parameters of KOD variants tested in the second-round engineering. For each tested variant, the mutation contents relative to reference KOD (RF), E(mut) predicted using regressor (E(mut)_{ML}), E(mut) characterized with kinetic assay (E(mut)_{expr}) were listed.

variant	mutation_1	mutation_2	E(mut) _{ML}	E(mut) _{expr}
mut_1	E485C	I486R	1.666	0
mut_2	E485I	R484D	1.812	0.79
mut_3	E485F	I486K	1.754	1.44
mut_4	E485I	R484K	1.661	1.74
mut_5	E485I	S451P	1.725	2.29
mut_6	E485I	D480H	1.68	2.34
mut_7	E485I	D480F	1.731	2.42
mut_8	E485I	D480P	1.942	2.55
mut_9	E485I	D480G	1.77	2.7
mut_10	E485I	D480C	2.057	2.89
mut_11	E485I	L453T	1.7	2.92
mut_12	E485I	S451A	1.702	3.31
mut_13	E485I	D480E	1.668	4.19
mut_14	E485I	S451R	1.722	4.96
mut_15	E485I	D480K	1.99	5.03
mut_16	E485I	D480W	2.087	5.44
mut_17	E485I	S451M	1.785	5.98
mut_18	E485I	L453N	1.912	6.09
mut_19	E485I	D480R	1.896	6.09
mut_20	E485I	S451K	1.678	6.15
mut_21	E485I	D480Q	1.932	6.18
mut_22	E485I	D480M	1.925	6.18
mut_23	E485I	D480S	1.923	7.65
mut_24	E485C	R484I	1.71	9.54
mut_25	E485I	S451H	1.813	9.7
mut_26	E485I	S451N	1.936	9.89
mut_27	E485C	R484V	1.669	10
mut_28	E485I	L453A	1.684	12.11
mut_29	E485I	S451G	1.854	12.38
mut_30	E485I	D480T	1.853	14.72
mut_31	E485I	R484C	1.887	22.84
mut_32	E485I	S451T	1.886	23.28
KOD_KH	E485I	S451L	1.651	31.05

Table S3. Performances of Extra tree regressor (Etr 1-3), Random Forest regressor (RF), and XGBoost regressor (XGB) on the KOD_train dataset. The dataset was divided into three different training/test pairs, and each dataset pair was used to train a regressor with Extra tree, Random Forest and XGBoost algorithms implemented within Scikit-Learn. Evaluation metrics including R^2 score and accuracy on training and testing subsets under each data split were shown.

	$R^2_{\text{train_Etr}}$	$R^2_{\text{test_Etr}}$	$R^2_{\text{train_RF}}$	$R^2_{\text{test_RF}}$	$R^2_{\text{train_XGB}}$	$R^2_{\text{test_XGB}}$
train_test_split pair 1	0.74	0.61	0.82	0.49	0.98	0.39
train_test_split pair 2	0.99	0.57	0.90	0.69	0.98	0.36
train_test_split pair 3	1.00	0.60	0.89	0.56	0.98	0.56
	$\text{Acc}_{\text{train_Etr}}$	$\text{Acc}_{\text{test_Etr}}$	$\text{Acc}_{\text{train_RF}}$	$\text{Acc}_{\text{test_RF}}$	$\text{Acc}_{\text{train_XGB}}$	$\text{Acc}_{\text{test_XGB}}$
train_test_split pair 1	0.87	0.75	0.86	0.66	0.96	0.64
train_test_split pair 2	0.97	0.71	0.90	0.50	0.94	0.63
train_test_split pair 3	1.00	0.71	0.56	0.70	0.92	0.70