Supplementary Information

Tautomerization-dependent Recognition and Excision of Oxidation Damage in Base-excision DNA Repair

Chenxu Zhu[a,1], Lining Lu[a,1], Jun Zhang[b,c,1], Zongwei Yue[a], Jinghui Song[a], Shuai Zong[a], Menghao Liu[a,d], Olivia Stovicek[e], Yiqin Gao[b,c,2] and Chengqi Yi[a,d,f,2]

[a]State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China.

[b]Institute of Theoretical and Computational Chemistry, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China.

[c]Biodynamic Optical Imaging Center, Peking University, Beijing 100871, China.

[d]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

[e]The University of Chicago, Chicago, IL 60637, USA.

[f]Synthetic and Functional Biomolecules Center, Department of Chemical Biology, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China.

[1]C.Z., L.L. and J.Z. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: chengqi.yi@pku.edu.cn or gaoyq@pku.edu.cn

Table of Contents:

Detailed Materials and Methods

I Expression and Purification of NEIL1

Full-length NEIL1 was synthesized by GENEWIZ Inc. and cloned into pET30a(+) with C-terminal His_6-tag. Truncations and mutations were further constructed based on this plasmid. Cells were grown at 37 °C and induced with 1 mM isopropyl β-D-thiogalactoside when OD_{600nm} reached 0.8, and continued to be grown at 16 °C for an additional 18 h. Recombinant proteins were purified by Ni-NTA (GE Healthcare) in phosphate buffer (50mM Na_2HPO_4/NaH_2PO_4, 300mM NaCl, pH 8.0) and Superdex 200 (GE Healthcare) with HEPES buffer (20mM HEPES, 100mM NaCl, pH 7.5). Site-directed mutageneses were performed using the Easy Mutagenesis System (Transgen Biotech.).

II DNA Synthesis and Purification

Oligo DNA: 5'-CGT CCA XGT CTA C-3' and 5'-Biotin-TGT TCA TCT GGG TCX TCG GTA TAT CCC AT-3' (X denotes THF, abasic site and Tg) were synthesized on an Exptide 8909 DNA synthesizer using standard reagents. Oligo DNA: 5'-TAG ACY TGG ACG G-3' and 5' ATG GGA TAT ACC GAY GAC CCA TGA TGA ACA-3' (where Y denotes C, G and A) were purchased from Sangon Biotech (Shanghai). Oligo DNA was deprotected by standard methods recommended by Glen Research Corp. and purified by urea-PAGE, dissolved in H_2O, and annealed at 1 mM final concentration in annealing buffer (10 mM Tris pH 8.0 and 100 mM NaCl). All purified oligo DNAs were verified by MALDI.

III Crystallization, Data Collection and Structure Determination

Before setting up the crystallization, the protein and DNA were mixed in the ratio 1:1.2, and the mixture was incubated on ice for no less than 20 min. Protein-DNA complex (10 mg/mL) was crystallized with reservoir solution [0.1 M cacodylic acid (pH 7.0), 0.1 M NaCl, 0.05 M $MgCl_2$ and 24% (w/v) PEG 8000] at 4 °C. Diffraction-quality crystals ($P2_12_12_1$) appear after 1-2 weeks, and reflection data were collected from BL17U (SSRF) or BL19U (NCPSS) with MarCCD. The collected data were processed by HKL2000. The initial models of NEIL1-DNA complexes were solved by molecular replacement using 1DTH as model (PhaserMR). The rest of the models were built into the electron density map using Coot, followed by refinement using Refmac5. Refinement statistics are shown in Table S1. All structure figures were prepared by PyMOL (DeLano Scientific).

IV Single-Turnover Kinetics and DNA Glycosylase Activity Assay

The glycosylase activity of NEIL1 was evaluated with single-turnover cleavage experiments, and a one-phase associate model was fitted to three parallel data sets for the calculation of the catalytic turnover rate k_{obs} (min^{-1}). For the single-turnover analyses, 24 pmol of NEIL1 was mixed with 2.4 pmol substrate DNA in a reaction buffer (20 mM Tris-HCl, pH 7.6, 10 mM EDTA, 0.1 mg/mL BSA, and 150 mM NaCl) in 120 μl reaction volume and incubated at 16 °C or 37 °C. Reactions were quenched by heating at 95 °C for 5 min (for Tg) or 65 °C for 10 min (for abasic site), and were analyzed by denaturing polyacrylamide gel followed by visualization using the ChemiDoc MP imaging system. All assays were performed three times independently.

V Setting up All-Atom Models of NEIL1-DNA Complex

The 2.5 Å resolution crystal structure of NEIL-Tg (DNA) complex was used to build the simulation system. The inactive mutant NEIL1 (P2G) was used to crystallize with Tg. We replace Gly2 using Pro2 according to the corresponding residue from NEIL1-THF structure.

The heavy atoms including ions and water molecules with high density in crystal structure were preserved as in crystal structure, while those of ions and other agents (e.g. glycerol, Tris) were removed. Before reducing the model with hydrogens, we prepared force field parameters for neutral N-terminal proline and Tg (see later). The hydrogen atoms in protein and DNA were automatically recovered by tLEAP module (AMBER tools) according to AMBER-ff99SB1, and according to the defined parameters for Pro2, Glu3 and the Tg-base. Counter ions were added to neutralize the overall system. We then added extra water molecules to the system, ensuring each atom of the protein-DNA complex (and ions) could be encompassed by water shell at least 10 Å thick. SPCE water model was adopted to describe the explicit solvation. The final all-atom model contains more than 49,000 atoms (box dimension: 80*90*80 Å3).

VI QM/MM Minimization of Active Pocket with Restraints from Crystal Structure

1. Building the Initial Structure and Minimization.

We treated the box with periodic boundary, setting a 10 Å cutoff for Ewald summation accounting for

electrostatic interactions. The NEIL1-Tg2-OH system was subjected to a sequential rough (compared to the next round of minimization) QM/MM relaxation: (i) water molecules, (ii) water and protein-DNA complex except the active pocket, (iii) all the atoms. The active pocket (A.P.) consists of following residues: Pro2, Glu3, Glu6, Arg/Lys242 and Tg2-OH, all of which were treated with SCC-DFTB (see more in *section VII* "QM-region selection"), while other atoms out of the A.P. were modeled by molecular mechanics. The inter-residue distance restraints (e.g. the distance between Arg/Lys242 and Tg) were applied to the active pocket according to crystal structure, with a force constant of 50 kcal·mol·Å^{-2}, which was used to keep important contacts and interactions unperturbed (as in crystal structure). Besides, a general but weaker positional restraint of 1.0 kcal·mol·Å^{-2} on other atoms of the NEIL1-DNA complex was turned on. The final relaxation step contains a 5000-step steepest descent and a following conjugate gradient minimization (till the minimization retards due to convergence). The final structure of the NEIL1-Tg2-OH system was used to generate initial structures for NEIL1-Tg (lactam) and NEIL1-Tg4-OH systems, by replacing the Tg2-OH-base with the corresponding tautomer.

2. Modeling Tg tautomers and active pocket isomers in NEIL.

The final structure of the NEIL1-Tg2-OH system (obtained in the above step) was used to generate initial structures for NEIL1-Tg (lactam) and NEIL1-Tg4-OH systems, by replacing the Tg2-OH-base with the corresponding tautomer. Now that the structures of NEIL1-Tg tautomers were all the same (e.g. coordinates of protein and DNA, water and ions), except for the tautomerization state of Tg-base, another round of accurate minimization was begun in order to calculate the relative stability of these three tautomers in the active pocket (i.e. the tautomerization energy). In addition, the PRH2 and GLH3 residues in the original model were later replaced with N-Pro2(+) and Glu(-), respectively, in order to estimate the energy difference between the ionization states of A.P. residues.

VII QM-region Selection, Tautomerization Energy of Tg & Isomerization Energy of Active Pocket

Several selections of the QM-region were tried. In the minimal definition (QM0), only the Tg residue along with Pro2 and Glu3 were counted in the QM-region. In the moderate definition (QM1), another two charged residues, Glu6 and Arg/Lys242, were also included in the QM-region. It is worth noting that QM1 covered all the catalytically important sites in the active site. In the large definition (QM2),

residues on the other side of Tg, including (side chains of) Tyr177 and Tyr263, were also treated in QM frame. Calculations showed that for the isomerization energy, QM1 and QM2 gave rise to quantitatively comparable results, both of which were more reliable than the too small QM0. All these QM-settings were listed in Table S2. Considering that QM1 in NEIL1-242R already took 91 atoms (including 5 capping hydrogens at the QM/MM boundary; for NEIL1-242K, the number was 89) as QM-part, we chose to stop at this boundary and not to impose a much greater burden (if we chose QM2) on our computation.

In the calculation of tautomerization energy, we kept residues 6 Å beyond the QM1-region restrained during the QM/MM optimization. For the QM-region, DFT (BLYP functional with 6-31+G* basis set) was used. All the initial atomic coordinates for Tg (lactam), Tg2-OH and Tg4-OH were the same except for the protonation state of Tg. The tautomerization energy for Tg was hence estimated by the converged minimization energy difference between these three systems, and the results were listed in the main text Table 1. To keep the structural features of the active site consistent with the crystal structure, we applied interatomic distance restraints for those forming the main contacts in the crystal structure, including the unusually short N3(Tg)-N(R242) distance.

Similarly, the isomerization of Pro2Glu3 was calculated for the NEIL1-242R-Tg2-OH structure. The initial structures for Pro2Glu3 and Pro2+Glu3- systems were the same except for the ionization state of the two concerned residues. We also quantified the entire QM1-region with DFT(BLYP/6-31+G*) to estimate the isomerization energy for Pro2+Glu3- changing into the neutral Pro2Glu3 state. The results were shown in Table S3. It is conclusive that a neutral state for Pro2Glu3, which is able to establish extra hydrogen bonds and form a more compact active pocket structure, is more energetically favorable and thus more reasonable for our modeling.

VIII Force-Field Simulations

1. Force-field setup

The neutral non-terminal glutamate acid is available in the AMBER force field library (termed as GLH), so we only needed to prepare proper force field parameters for the neutral N-terminal Pro and Tg (similarly, for Tg2-OH and Tg4-OH). We adopted available parameters for PRH (which denotes the neutral N-terminal proline) developed by R. A. Perlow-Poehnelt et al (1). For Tg tautomers, a standard force field parameterization procedure was performed as follows: we first optimized geometry of Tg

nucleotides and obtained partial atomic charges using Hartree-Fock calculations with 6-31G* basis set, in order to keep consistency with atomic charges in the AMBER force field (2); then assigned the Tg with RESP-A1A charges using RED server (3), which required the partial atomic charges and optimized geometry from GAUSSIAN calculations; finally the force field parameters were generated according to the existing analogues in the AMBER library (e.g. DNA thymine). All the *ab initio* quantum computations were executed in GAUSSIAN 2009 (4).

2. Force-field relaxation and equilibration of crystal structures of NEIL1-DNA complex

All atomic coordinates in the initial structures of MM simulation were the same with the QM(BLYP)/MM optimized structure. The force-field system was heated up to 300 K under 1 atm, during which a positional restraint (2.0 kcal·mol·Å^{-2}) on protein and DNA was applied. SHAKE(5) algorithm was adopted to constrain all motions involving hydrogens. A uniform 2 fs step was used for all force-field MD simulations. The heating time lasted for 100 ps, followed by a 500 ps equilibration stage, during which the positional restraint was gradually tuned off (reducing 0.5 kcal·mol·Å^{-2} every 100 ps). The final structure reached at the end of the equilibration was then used for: (i) calculating the thermal fluctuations of A.P. structures and interactions using MM MD simulations (see next part); (ii) generating representing structures for umbrella sampling using QM/MM computations (see *Section IX* in SI).

3. Force-field simulations of structural fluctuations of A.P.

Unbiased force-field MD simulations were performed for NEIL1-242R-Tg (lactam), -Tg2-OH and -Tg4-OH systems, respectively. The production run, which lasted for 30 ns for each, was all put under NPT ensemble (1 atm, 300 K), with SHAKE constraint and 2 fs time step. All the simulations here were free of restraints. The data were collected every 1 ps. We calculated the interatomic distance fluctuation of those forming main contact in the optimized structure. The results were described in Fig. S4. The least structural fluctuations and the least deviation from the crystal structure were observed in the NEIL1-242R-Tg2-OH complex, and the distances between Pro2-Tg and Arg242-Tg remained short enough to trigger an enzymatic reaction (Fig. S4*A*). However, the NEIL1(R242)-Tg complex exhibited significant structural deviation from the crystal structure (Fig. S4*C*), again proving that Tg should undergo tautomerization in the active pocket. We also checked the distance between peripheral water molecules and the active site, demonstrating that the compact active pocket excludes water molecules outside the compact catalysis site (Fig. S4*D*).

IX Umbrella Sampling

1. QM/MM Steered MD

The reaction coordinate(s) characterizing each step of the enzymatic reaction was chosen as shown in Fig. 3*A* in the main text and Schemes S1, S2 and S3. The reaction coordinates were chosen to account for the bond-forming and bond-breaking in each step. We then performed steered QM(SCC-DFTB)/MM MD simulations to generate a series of structures along each reaction coordinate. The semi-empirical quantum method SCC-DFTB(6) was adopted to quantify the potential of the QM region, provided that it generally ensures satisfying accuracy, performs reliably in polar systems, and is computationally extremely efficient compared to the first-principle methods (7). The QM region was defined as QM1 (see above). For every reaction coordinate, the steered MD lasted for no less than 2000 fs (with 1 fs time step), which ensured at least 1500 fs for a change of 1 Å along the reaction coordinate so as to relax the environment, and the steering force ranged from 80 to 200 kcal·mol·Å^{-2}. The force constant used in steered MD was used as reference for the following umbrella sampling. After the umbrella sampling for one reaction coordinate was done, the next steered MD (along the next reaction coordinate) then began with the minimal free-energy structure of the previous umbrella sampling.

2. Umbrella Sampling

The umbrella sampling (8) was done using SCC-DFTB/AMBER QM/MM simulations, with QM1 region treated quantum mechanically. The harmonic restraint (used to keep the configuration locally fluctuated) was in line with the parameters in the steered MD, with the force constant ranging from 80 to 200 kcal·mol·Å^2. The window was 0.1 Å in width along most reaction coordinates, and 0.05 Å for those merely consisting of proton transfers, resulting in 70 to 90 windows for a thorough study of one mechanism. The time step was 1 fs, and simulations were executed under NPT ensemble (300 K, 1 atm). For each window, up to 80 ps QM/MM simulation was performed, and the last 70 ps trajectory was collected for data analysis. Generally, the QM/MM umbrella sampling reached a combined simulation time of ca. 6 ns for each of the four mechanisms we studied in this paper. The weighted histogram analysis method (WHAM)(9) was applied to reconstruct the unbiased free energy profiles (Figs. S5 and S6). The maxima (transition states) and minima (intermediate and stable states) along each RC were extracted and aligned to forge the illustrative overall free-energy profile throughout the entire reaction (Fig. 3 in the main text, and Fig. S6). The uncertainties of the resulting free energy profiles were assessed by a Monte Carlo bootstrapping procedure (10) (source code from: Grossfield, A, "WHAM: an implementation of the weighted histogram analysis method",

http://membrane.urmc.rochester.edu/content/wham/, version 2.0.5). We also checked the overlap of the histograms between adjacent simulation windows to validate the reliability of the reconstructed free energy surface. Extensive overlap was indeed ensured for the chosen interatomic distance coordinates (Fig. S9 *A* and *B*).

X Calculation of Tautomerization Energy Difference Between T and Tg

To explore the influence of the loss of aromaticity on the tautomerization energy, we compared the tautomerization energy of Tg with its aromatic parent, thymine. All the optimization energies, for the (neutral) nucleosides alone, were calculated in vacuum, which reflects the intrinsic chemical difference between these two nucleosides. The QM optimization was realized by DFT-B3LYP with 6-311G* basis set on GAUSSIAN 09. Indeed, the energy differences between Tg tautomers are lower than those of T tautomers (meanwhile the latter were found to be much more dipolar) (Table S4), suggesting the tautomerization requirement might serve as a possible checkpoint for NEIL1 to reject normal thymine as substrate.

We also drew the highest occupied molecular orbitals (HOMO) in the form of the electron-density diagram (Fig. S9 *C* and *D*), considering that HOMO indicates possible molecular sites that attack (or receive) electrophilic groups such as protons. Fig. S9 *C* and *D* show that the O2 site on Tg is enriched with more HOMO density than T, suggesting a stronger capability of receiving a proton from another source, in agreement with the previous finding that the tautomerization energy for Tg transforming to Tg2-OH is lower than that for T to T2-OH.

XI Determining the Difference of pK_a Value Between Tg and T Nucleoside

In our proposed mechanism, we assume that Tg adopts a neutral tautomer form rather than ionizes after it flips out from DNA helices into NEIL1 active pocket. To interrogate the possibility of Tg ionizing, we estimated its pK_a value using a classical alchemistry method, the thermodynamic integration (11) method. By designing a thermodynamic circle as shown in Scheme S4, the free energy difference between Tg and T during ionization can be calculated, which leads to a reasonable estimation of pK_a difference between the two nucleotides ($\Delta \mathrm{p}K_a = \mathrm{p}K_a(\mathrm{Tg}) - \mathrm{p}K_a(\mathrm{T})$).

For T and Tg, in their neutral nucleoside form as well as the ionic state, each one was solvated in an identical TIP3P water box with periodic boundary condition, and sodium cations were added to

neutralize the overall charges if needed (i.e. in T^- and Tg^- box). The thermodynamic integration calculations were performed on AMBER 10 package. The C5 and C6 atoms along with their attached groups were chosen as the "transformation" part during thermodynamic integration. Two independent transformations were considered: one from neutral T to neutral Tg (corresponds to $\Delta G_{\text{neutral}}$ in Scheme S4), the other from ionic T^- to Tg^- (corresponds to ΔG_{ionic} in Scheme S4). Each transformation process is composed of three basic steps: first to turn off the charges on the chosen part of the initial molecule (ΔG^1), then switch on the van der Waals parameters using softcore potential (12) method (ΔG^2), finally turn on the charges of the chosen atoms of the final molecule (ΔG^3). The total "transformation" energy, or rather, the free energy difference between the initial molecule and the final molecule, can be represented as:

$$\Delta G = \Delta G^1 + \Delta G^2 + \Delta G^3$$

And each $\Delta G^i (i = 1,2,3)$ was given by the thermodynamic integration formalism and the Gaussian numerical quadrature integration formula, where λ is the nonspatial coordinate ranging from zero to one, and $V(\lambda)$ is the mixed potential of the initial state (λ=0) and the final state λ=1:

$$\Delta G^i \equiv \Delta G^i_{\text{TI}} = \int_0^1 \left(\frac{\partial V^i(\lambda)}{\partial \lambda} \right)_\lambda d\lambda$$

$$\Delta G^i_{\text{TI}} = \sum_{k=1}^{9} w_k \left(\partial V^i(\lambda)/\partial \lambda \right)_\lambda$$

For every transformation process, we chose a series of 9 λ values (do not include 0 and 1, and the corresponding integration weight is denoted as w_i) in order to do a Gaussian quadrature integration (which is numerically more accurate than the simple trapezoid approach). For each λ value in each transformation process, the production run lasted over 10 ns (with 1 femtosecond simulation time step), which is significantly longer than the correlation time of $\partial V^i(\lambda)/\partial \lambda$; the converged (Fig. S8*E*) accumulative average of $\partial V^i(\lambda)$ was used in the final integration. For every λ, we listed the thermodynamic integration data, namely $\partial V^i(\lambda)/\partial \lambda$, and the corresponding integration weight in Table S5.

Taking the thermodynamic circle into account, the free energy difference between Tg and T during ionization is given by:

$$\Delta\Delta G = \Delta G_{\text{Tg}} - \Delta G_{\text{T}} = \Delta G_{\text{ionic}} - \Delta G_{\text{neutral}} = \sum_{i=1}^{3} (\Delta G^i_{\text{Tg}} - \Delta G^i_{\text{T}})$$

Therefore, pKa value (at temperature *T*, 300 K in this work) of Tg is different from T by:

$$\text{pK}_a(\text{Tg}) - \text{pK}_a(\text{T}) = \frac{\Delta\Delta G}{2.303RT}$$

The final results, $\Delta\Delta G$=16.8 and ΔpK_a>11 (at 300 K), indicate that Tg is unlikely to spontaneously ionize in water (given that pK_a(T) >8 at room temperature (13)). Besides, the electrostatic interactions (Step 3) contribute most substantially to the overall free energy difference.

Supplementary Tables

Table S1. Data collection and model statistics

	NEIL1 Protein	NEIL1 w.t. THF	NEIL1 P2G THF	NEIL1 242K THF	NEIL1 P2G Tg	NEIL1 P2G 242K Tg
Space Group	H3	P212121	P212121	P212121	P212121	P212121
Unit-cell Parameters	132.5, 132.5,50.8 90.0, 90.0, 120.0	73.9,108.8, 171.3 90.0,90.0,90.0	73.3,109.0, 169.8 90.0, 90.0, 90.0	73.7,109.3, 71.9 90.0, 90.0, 90.0	73.6,108.7, 169.9 90.0, 90.0, 90.0	74.0,109.3, 171.0 90.0, 90.0, 90.0
Data Collection Statistics						
Resolution	66.23-1.48 (1.48-1.52)	81.81-2.53 (2.53-2.60)	91.75-2.46 (2.46-2.52)	92.23-2.41 (2.41-2.48)	91.58-2.48 (2.48-2.54)	92.14-2.65 (2.65-2.72)
Unique reflections	52390	43613	47431	50869	46040	38781
Redundancy	5.6	3.8	5.1	5.8	4.0	6.1
R_{merge}	0.022 (0.027)	0.061 (0.081)	0.054 (0.057）	0.063 (0.078)	0.055 (0.071)	0.059 (0.087)
Completeness	99.6%	99.4	99.4%	98.85	98.5%	99.3%
Overall I/sigma	27.27	19.23	22.58	24.26	21.7	22.36
Refinement Statistics						
R_{work}	0.156	0.211	0.211	0.213	0.177	0.192
R_{free}	0.196	0.278	0.263	0.267	0.228	0.235
r.m.s.d. Values						
Bond Length (Å)	0.0223	0.0148	0.0146	0.0149	0.0145	0.0141
Bond Angles (°)	2.1958	1.8113	1.7297	1.8711	2.0283	1.8028
No. atoms		8037	7976	8118	8300	8096
Protein	2-290	2-290	2-290	2-290	2-290	2-290
Water	488	189	125	179	447	225
DNA	0	1-26	1-26	1-26	1-26	1-26
Ramachandran Plot						
Preferred (%)	94.74%	90.66%	90.18%	90.65%	91.24%	90.94%
Allowed (%)	4.89%	8.43%	8.90%	8.43%	7.84%	8.27%
Disallowed (%)	0.38%	0.91%	0.92%	0.92%	0.92%	0.79%
Accession No.	5ITQ	5ITT	5ITR	5ITU	5ITY	5ITX

Table S2. QM-region selections for QM/MM simulations of NEIL-Tg

QM-region definition	Residues in QM-region	Total QM atom numbers (excluding capping atoms)
QM0	Tg,P2,E3	59(59)[a]
QM1	Tg, P2, E3, E6, R(K)242	86(84)
QM2	Tg, P2, E3, E6, Y177, R(K)242,Y263	116(114)

a. Numbers in and out of the parentheses correspond to NEIL1-K242 and NEIL1-R242, respectively.

Table S3. Isomerization Energy of Pro2 and Glu3

Isomer[a]	Energy difference (kcal/mol)
Pro2 Glu3 [b]	0.0
Pro2+Glu3-	11.2

a. Structures are derived from NEIL1-Arg242-Tg2-OH.
b. The double neutral state (Pro2Glu3) is set as the zero-energy reference.

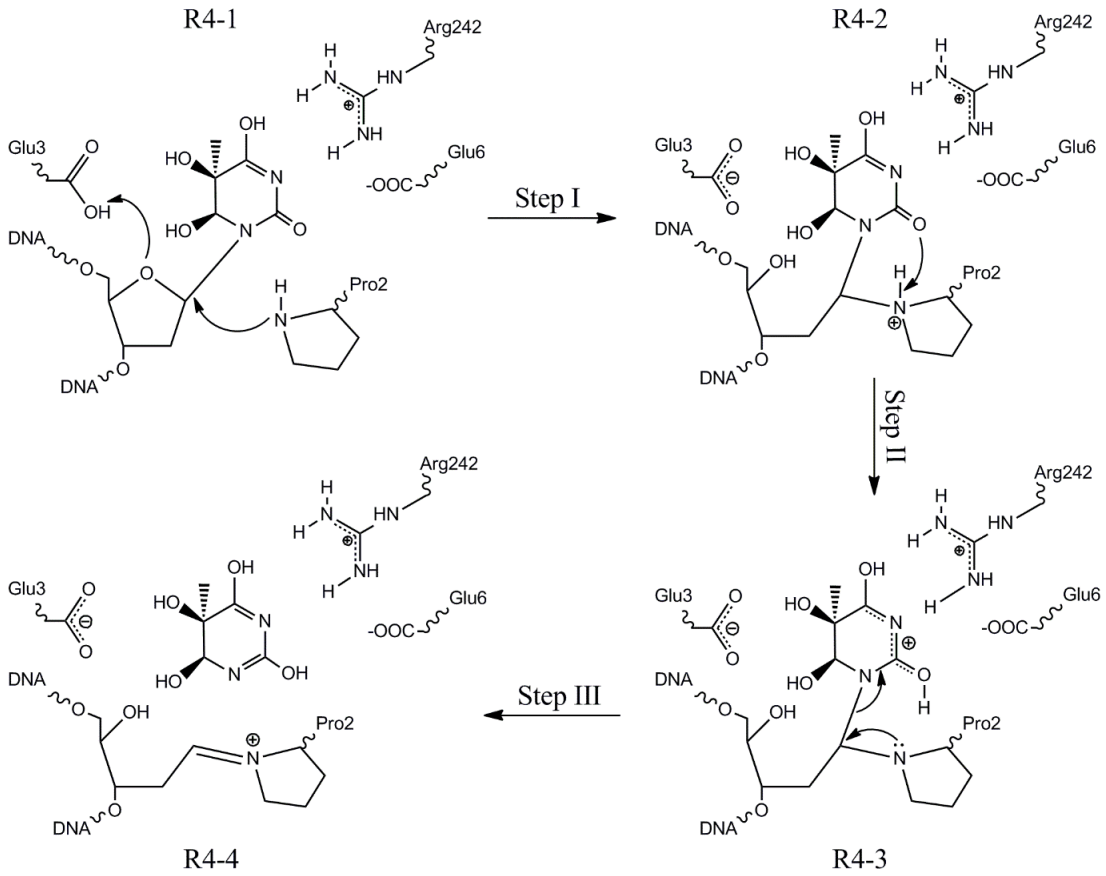
Table S4. Energy difference between tautomers of Tg and T

	ΔE/(kcal/mol)	Dipole(debye)		ΔE/(kcal/mol)	Dipole(debye)	ΔΔE/(kcal/mol)
Tg (lactam)	0	3.72	T (lactam)	0	5.74	/
Tg4-OH	23.94	5.73	T4-OH	23.08	8.57	0.87
Tg2-OH	20.46	5.54	T2-OH	22.76	7.85	-2.30

Table S5. The thermodynamic integration data and integration weights used in Gaussian quadrature integration

λ value	$\Delta\partial V^1(\lambda)/\partial\lambda$	$\Delta\partial V^2(\lambda)/\partial\lambda$	$\Delta\partial V^3(\lambda)/\partial\lambda$	Gaussian integration weight
0.01592	-8.88	-10.30	39.58	0.04064
0.08198	-9.26	-11.55	39.68	0.09032
0.19331	-9.90	-9.92	39.58	0.13031
0.33787	-10.44	-11.85	39.32	0.15617
0.5	-10.89	-10.60	38.71	0.16512
0.66213	-11.44	-10.84	37.80	0.15617
0.80669	-11.85	-8.94	35.36	0.13031
0.91802	-12.04	-8.91	36.35	0.09032
0.98408	-12.44	-10.71	34.01	0.04064

Supplementary Schemes



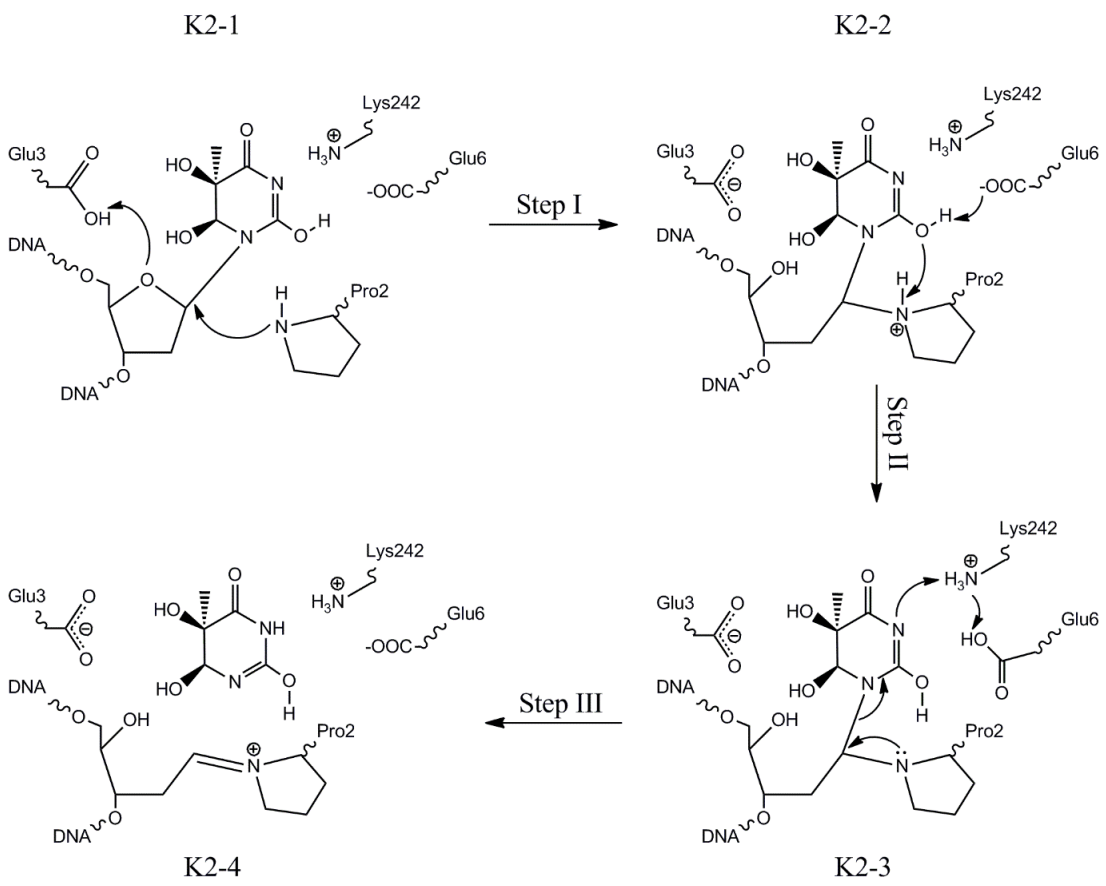
Scheme S1. Ribose-protonated pathway initiated with the NEIL1 242R-Tg4-OH structure; the reactant, intermediate and product states are termed R4-1, R4-2 and R4-3, respectively.

(1) RC1= d(C1'$_{Tg4OH}$-O4'$_{Tg4OH}$)-d(C1'$_{Tg4OH}$-N$_{Pro2}$)-d(O4'$_{Tg4OH}$ –H$_{Glu3}$) during Step I;

(2) RC2= -d(O2$_{Tg4OH}$-H$_{Pro2}$) during step II;

(3) RC3= d(C1'$_{Tg4OH}$-N1$_{Tg4OH}$) during step III;

(4) Releasing RC4= d(N$_{Arg242}$-N3$_{Tg4OH}$) after Step III.



Scheme S2. Ribose-protonated pathway initiated with the NEIL1 242K-Tg2-OH structure; the reactant, intermediate and product states are termed K2-1, K2-2 and K2-3, respectively.

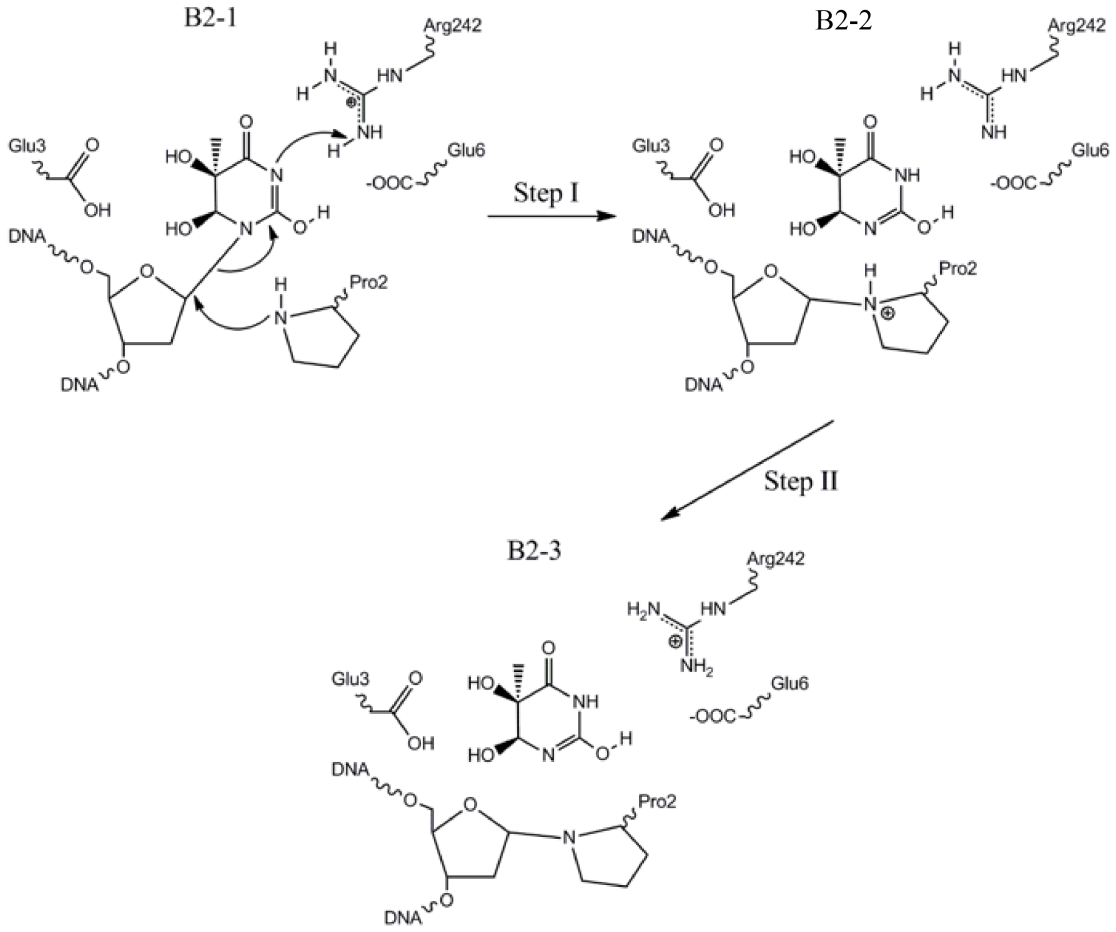
Reaction coordinates were chosen according to the mechanism:

(1) RC1=d(C1'$_{Tg2OH}$-O4'$_{Tg2OH}$)-d(C1'$_{Tg2OH}$-N$_{Pro2}$)-d(O4'$_{Tg2OH}$ –H$_{Glu3}$) during Step I;

(2) RC2=d(H$_{Pro2}$-N$_{Pro2}$)-d(H2$_{Tg2OH}$-O$_{Glu6}$) during step II;

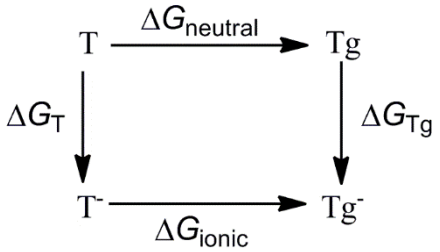
(3) RC3=d(C1'$_{Tg2OH}$-N1$_{Tg2OH}$)-d(N3$_{Tg2OH}$-H$_{Lys242}$)-d(N$_{Lys242}$-H$_{Glu6}$) during step III;

(4) Releasing RC4=d(N$_{Lys242}$-H3$_{Tg2OH}$) after Step III.



Scheme S3. Base-activated pathway initiated with the NEIL1 242R-Tg2OH structure; the reactant, intermediate and product states are termed B2-1, B2-2 and B2-3, respectively. Reaction coordinates were chosen according to the mechanism:

RC= $d(C1'_{Tg2OH}\text{-}N1_{Tg2OH})\text{-}d(C1'_{Tg2OH}\text{-}N_{Pro2})+d(N3_{Tg2OH}\text{-}H_{Arg242})$ during Step I.



Scheme S4. Alchemistry of T-Tg transformation and the thermodynamic cycle of Tg ionization.

Supplementary Figures

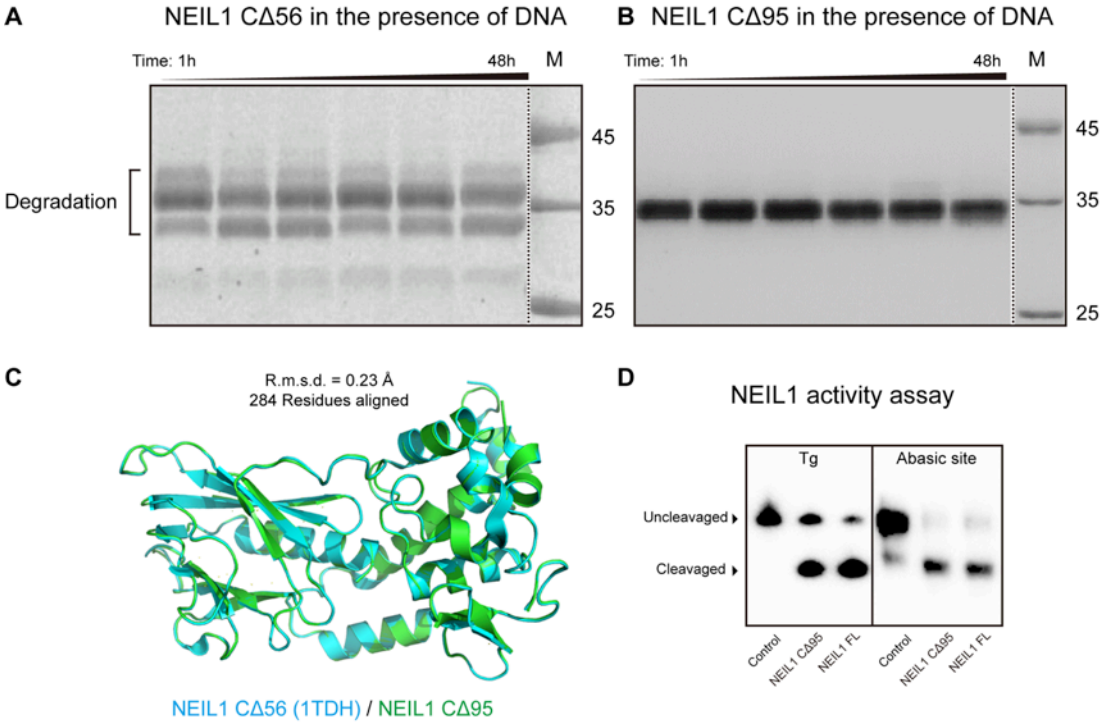


Fig. S1. Crystallization of NEIL1 CΔ95.

(A) NEIL1 CΔ56 was used previously for crystallization. However, it appeared unstable in the presence of dsDNA.

(B) Further truncation at the CTD (NEIL1 CΔ95) is stable in the presence of dsDNA.

(C) Secondary-structure matching (SSM) superposition of NEIL1 CΔ95 with CΔ56 (PDB: 1TDH) indicates further truncation at CTD introduced minimal perturbation to NEIL1 glycosylase domain.

(D) NEIL1 CΔ95 retains comparable glycosylase and lyase activity compared to full length NEIL1. All activity assays were duplicated.

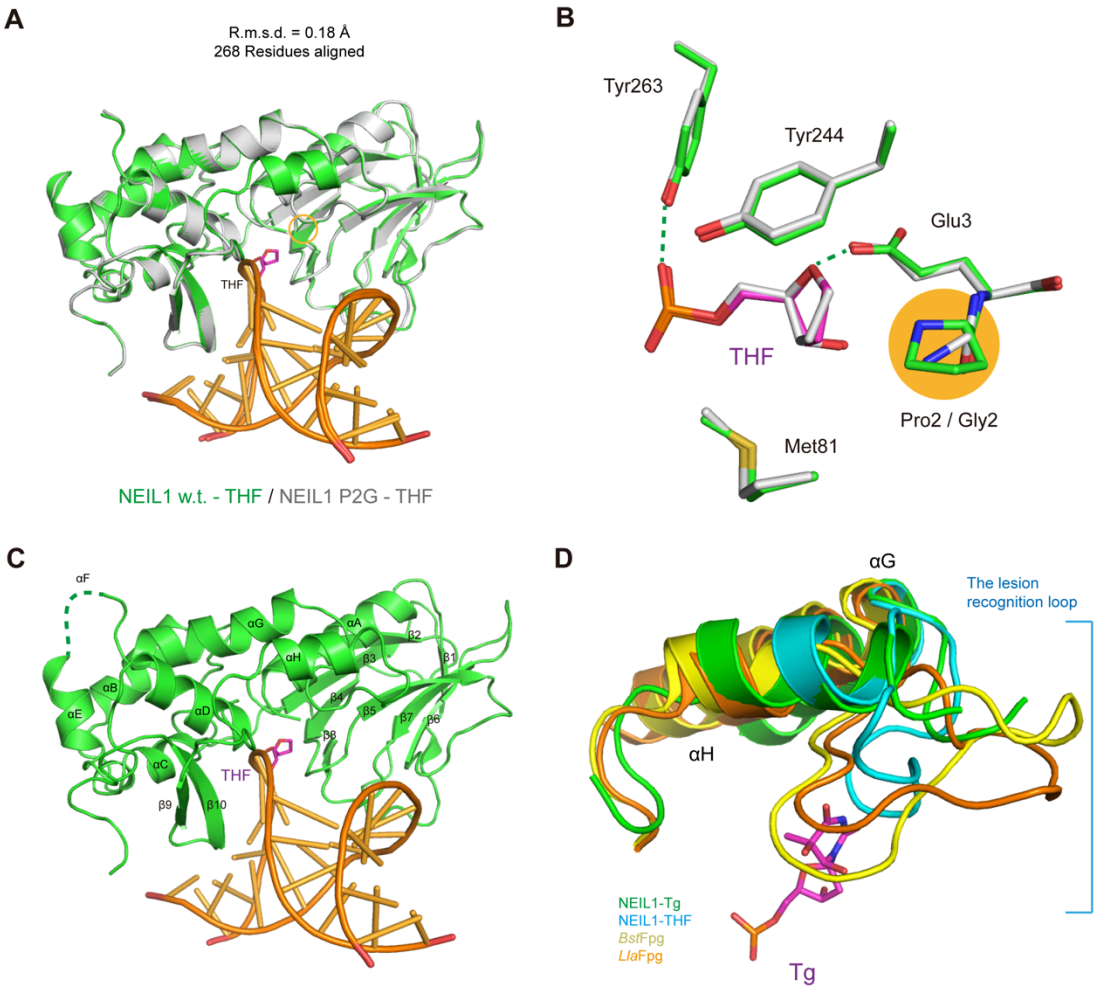


Fig. S2. The crystal structures of NEIL1 bound to duplex DNA containing substrates.

(A) SSM superposition of NEIL1 w.t. and P2G bound to THF-containing dsDNA shows that the two structures are almost identical. The Pro2/Gly2 residue is marked with a yellow circle.

(B) The Pro2-to-Gly2 mutation introduced minimal perturbation to the recognition of THF by NEIL1.

(C) The secondary structure elements of NEIL1 are marked according to ref (14).

(D) Superposition of NEIL1 with *Bst*Fpg (PDB: 1R2Y) and *Lla*Fpg (PDB: 1PJJ) shows that the loop between helix αG and αH corresponds to the *α*-F-*β*9/10 loop (lesion recognition loop) of Fpg.

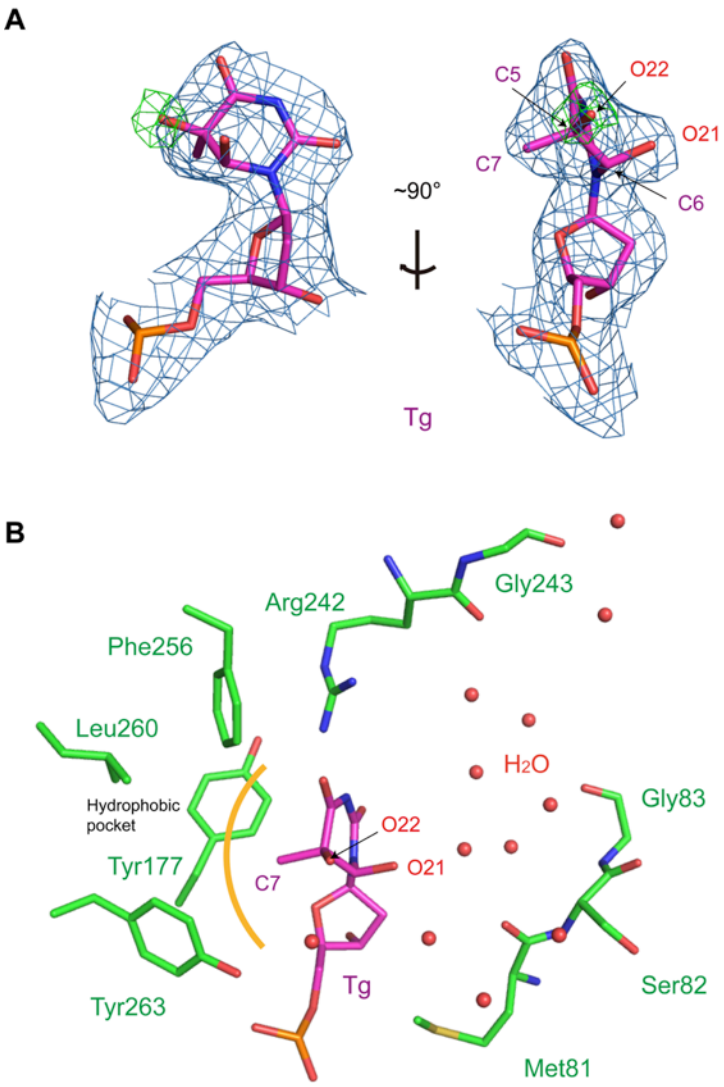


Fig. S3. The crystal structure of the Tg-bound complex.

(A) The flipped Tg fits well with the electron density map. The blue $2F_{obs}$-F_{cal} map is contoured at 1.2σ; the green F_{obs}-F_{cal} omit map was computed by removing 5-hydroxyl group (O22) of Tg and is contoured at 2.5σ.

(B) The hydrophobic pocket around the 5-methyl group and the hydrophilic side around the two hydroxyl groups of Tg.

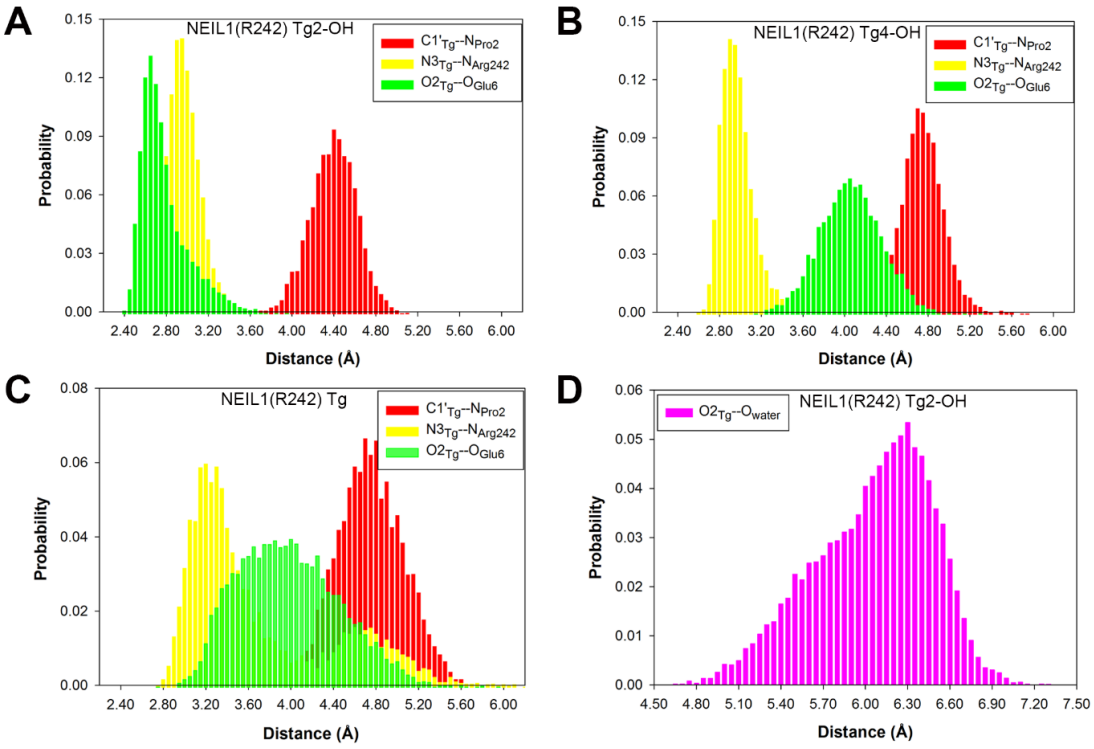


Fig. S4. Structural fluctuations of active site pocket in different NEIL1-Arg242-Tg tautomer complexes: (A) Tg2-OH, (B) Tg4-OH and (C) Tg (lactam).

(D) The distribution of water molecules around the catalysis site was also collected for the NEIL1 Arg242 Tg2-OH complex.

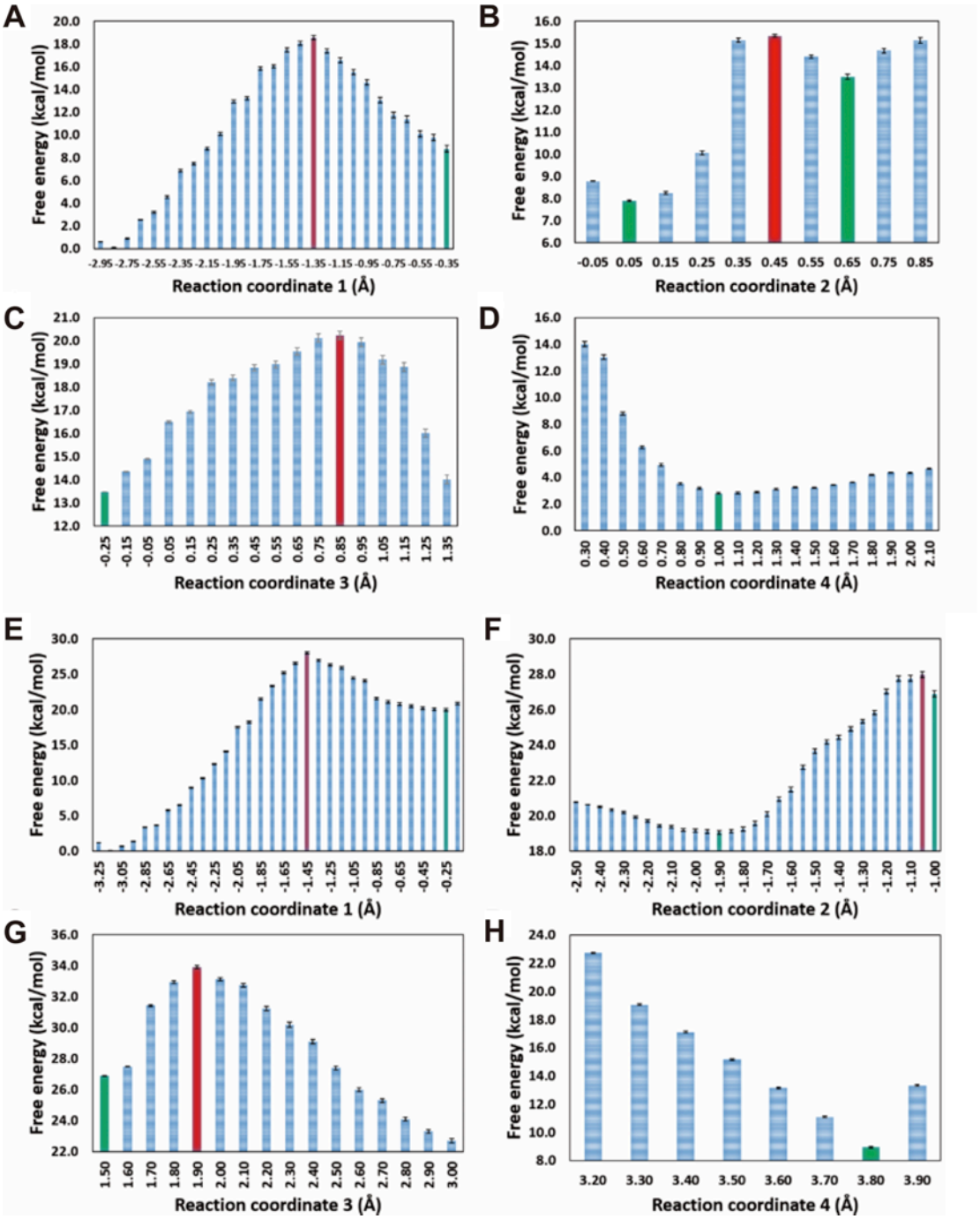


Fig. S5. Free energy profile calculated by umbrella sampling along each reaction coordinate during ribose-protonated catalysis pathway for NEIL1-Arg242-Tg2-OH and NEIL1-Arg242-Tg4-OH complex.

(A), (B), (C) and (D) panels correspond to R2-1, R2-2, R2-3 and R2-4 defined in Fig. 3*A*.

(E), (F), (G) and (H) panels correspond to R4-1, R4-2, R4-3 and R4-4 defined in Scheme S1.

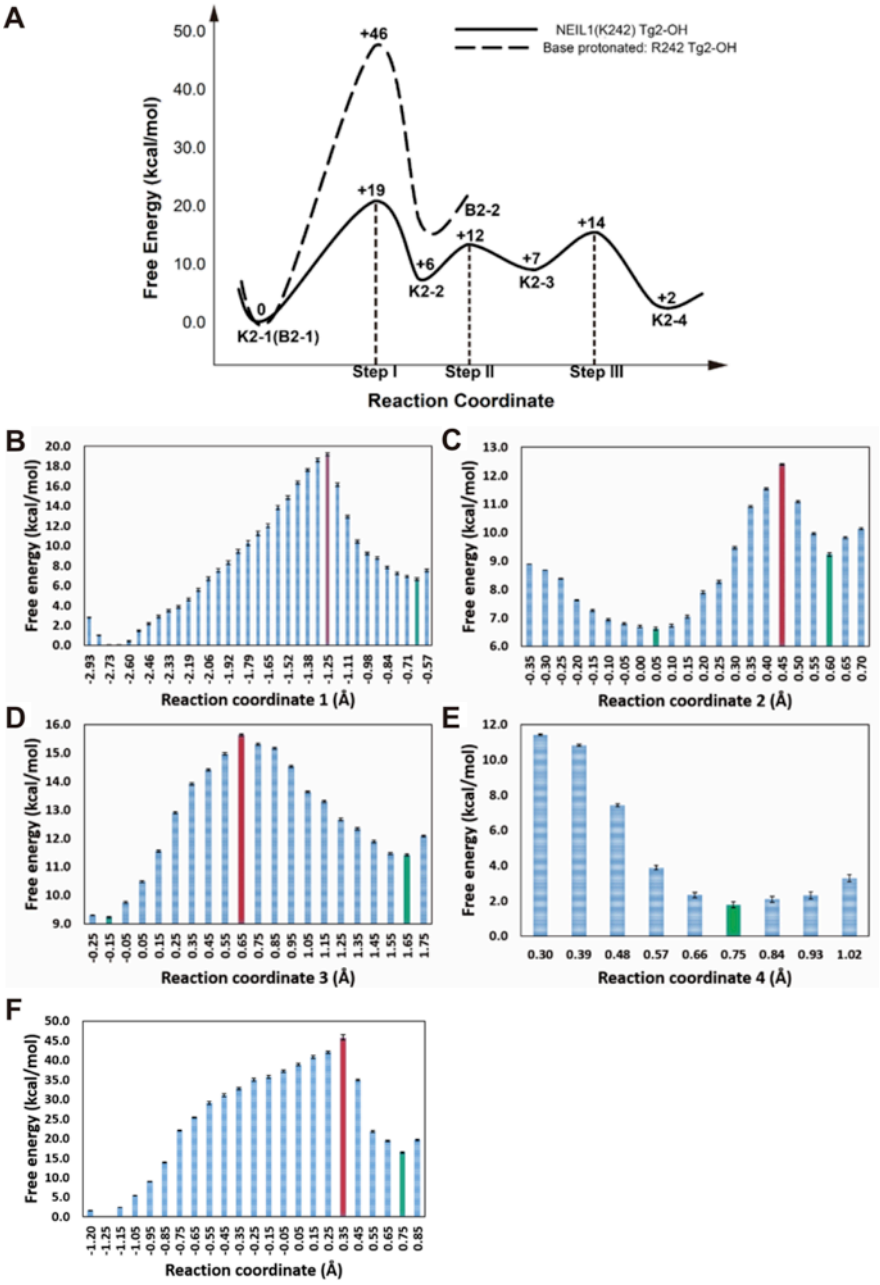


Fig. S6. (A) The merged free energy profile for the proposed mechanisms: solid line corresponds to ribose-protonated pathway in NEIL1-Lys242-Tg2-OH complex (as shown in Scheme S2), long dashed line to base-activated (protonated) pathway in NEIL1-Arg242-Tg2-OH complex (as shown in Scheme S3), short dashed line indicates the intermediate states corresponding to Scheme S2 and Scheme S3. All the intermediate and stable states are denoted by the same notation defined in Scheme S2 and Scheme S3, of which the free energies (along with those of transition states) are shown in corresponding numbers.

Free energy profile calculated by umbrella sampling along each reaction coordinate: (B), (C), (D) and (E) correspond to K2-1, K2-2, K2-3 and K2-4 defined in Scheme S2; (F) corresponds to K2-1 in Scheme S3.

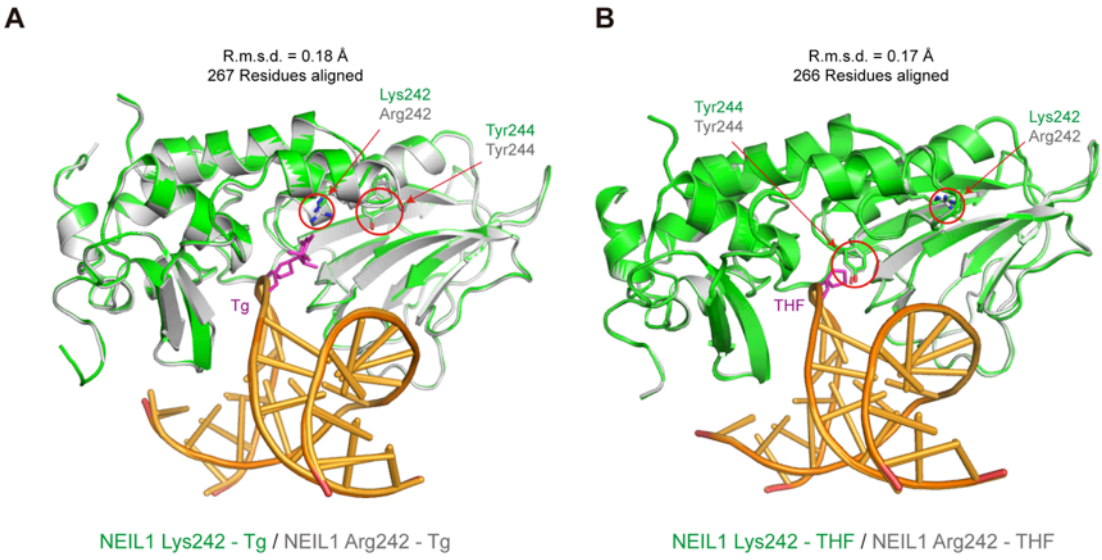


Fig. S7. Crystal structures of NEIL1 Lys242 bound to duplex DNA.

(A) The structures of NEIL1 Lys242-Tg (green) and NEIL1 Arg242-Tg (grey) are nearly identical. Both the side chains of Lys242 and Arg242 are pointing to the flipped Tg base.

(B) The structures of NEIL1 Lys242-THF and NEIL1 Arg242-THF. In both structures, the Lys/Arg242 are pointing to a solvent-exposed surface, away from THF.

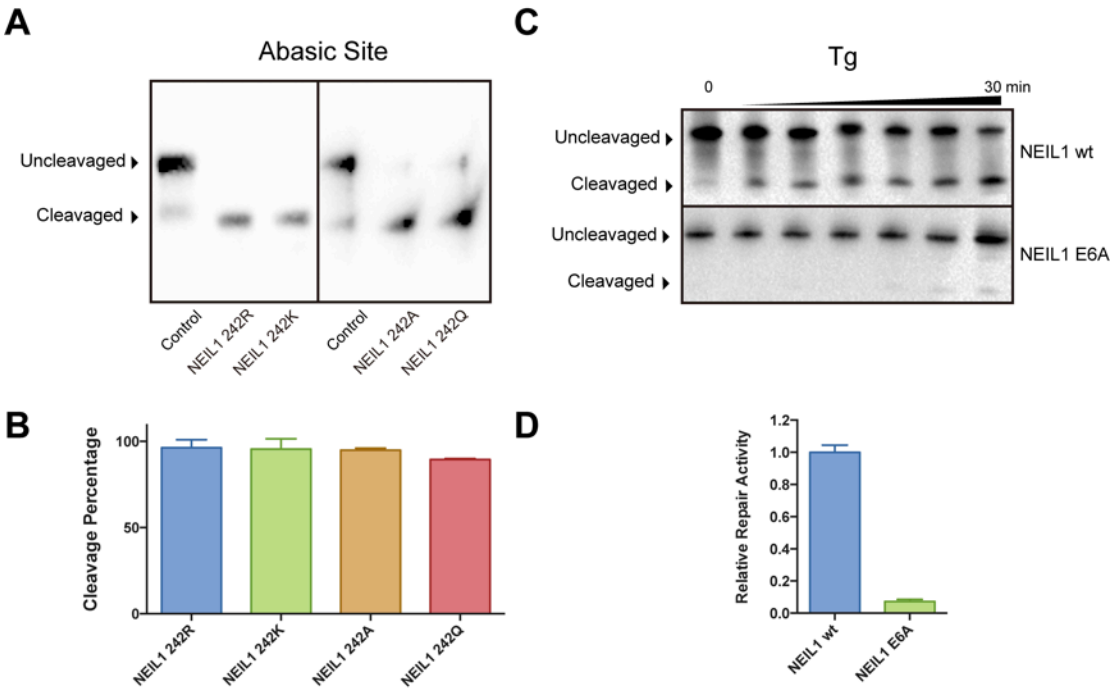


Fig. S8. Biochemical analysis of key residues in NEIL1.

Mutations at the editing site introduce minimal impact to the lyase activity of NEIL1. Denaturing PAGE analysis of lyase activity of NEIL1 to abasic site is shown in (A), and normalized cleavage percentage is shown in (B). All experiments were duplicated.

The glycosylase activity of NEIL1 to Tg is severely hindered by E6A mutation. Denaturing PAGE analysis of glycosylase/lyase NEIL1 to Tg:A is shown in (C), and normalized relative repair activity is shown in (D). All experiments were duplicated.

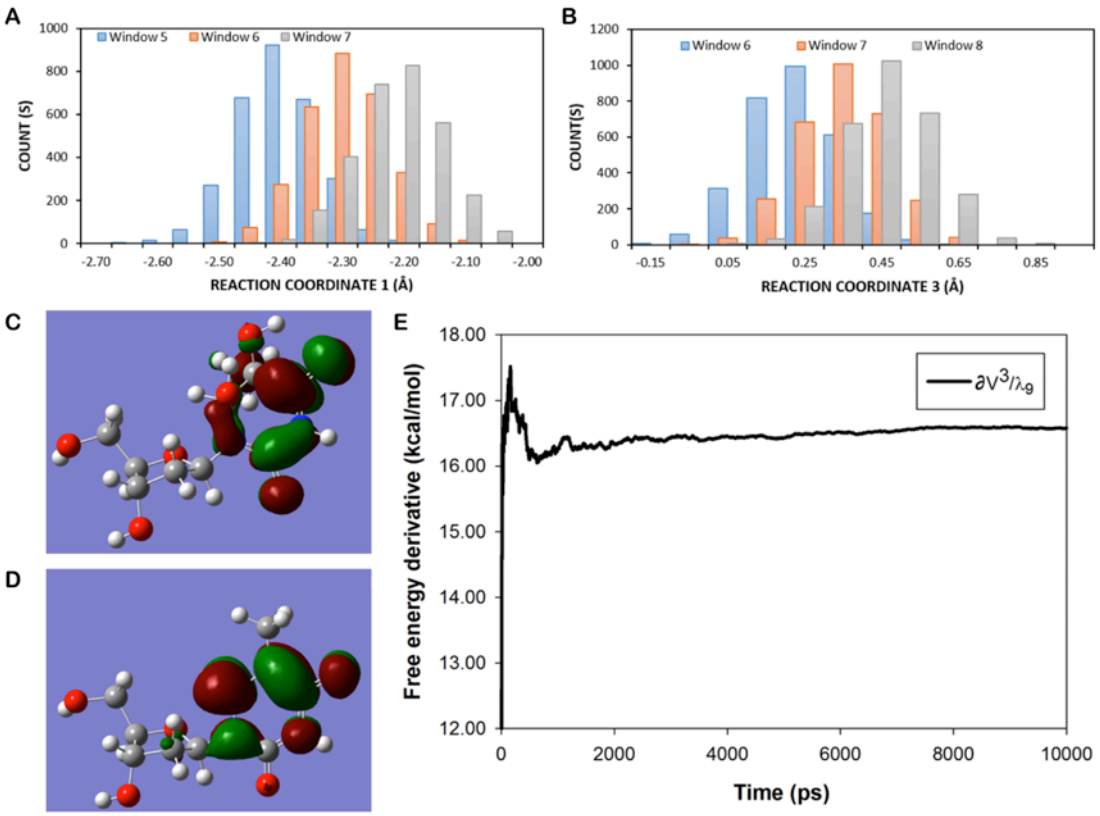


Fig. S9. The overlapping between adjacent windows in umbrella sampling data of NEIL1-Arg242-Tg2-OH ribose-protonated pathway. (A) Overlapping between windows #5, #6 and #7 in Fig. S5*A,* which are used in WHAM. (B) Overlapping between windows #6, #7 and #8 in Fig. S5*C,* which are used in WHAM.

HOMO of Tg molecule (C) and T nucleoside (D), drawn using GaussianView program with respect to the same isovalue. Obviously, being the oxidized product of T, Tg possesses more HOMO electron density at the O2 atom compared to normal T base, leading to O2 position being a better potential proton acceptor.

(E) Example of accumulative average of the potential derivative during thermodynamic integration calculation. The simulation time is clearly long enough to assure a converged free energy calculation with limited statistical uncertainties.

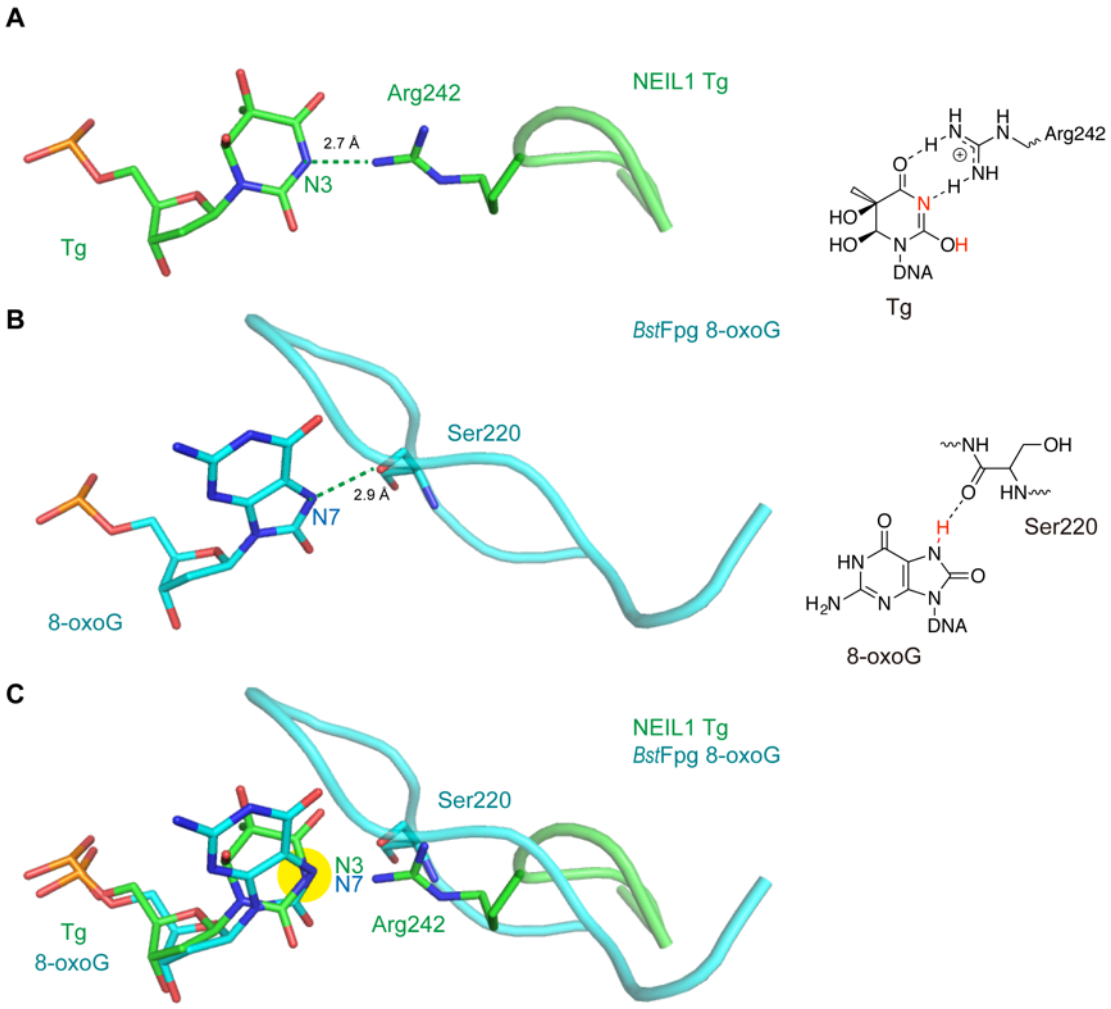


Fig. S10. Tautomerization-dependent substrate readout of NEIL1 is reminiscent of oxoG recognition by Fpg.

(A) Although the keto form of Tg is more stable in solution, NEIL1 promotes tautomeric shifts of substrate to achieve optimal binding.

(B) *Bst*Fpg (PDB: 1R2Y) detects the oxidation state of a substrate by sensing the protonation state of N7 of 8-oxoG.

(C) Superposition of NEIL1-Tg (green) with *Bst*Fpg-8-oxoG (blue) shows that the N3 of Tg overlaps well with N7 of 8-oxoG. The overlapped nitrogen atoms are highlighted in yellow.

References:

1. Perlow-Poehnelt RA, Zharkov DO, Grollman AP, & Broyde S (2004) Substrate discrimination by formamidopyrimidine-DNA glycosylase: distinguishing interactions within the active site. *Biochemistry* 43(51):16092-16105.
2. Cornell WD*, et al.* (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117(19):5179-5197.
3. Dupradeau F-Y*, et al.* (2010) The RED. tools: advances in RESP and ESP charge derivation and force field library building. *Phys. Chem. Chem. Phys.* 12(28):7821-7839.
4. Frisch M*, et al.* (2009) Gaussian 09, Revision A. 02, Gaussian. *Inc., Wallingford, CT* 200.
5. Ryckaert J-P, Ciccotti G, & Berendsen HJ (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23(3):327-341.
6. Elstner M, Jalkanen KJ, Knapp-Mohammady M, Frauenheim T, & Suhai S (2001) Energetics and structure of glycine and alanine based model peptides: Approximate SCC-DFTB, AM1 and PM3 methods in comparison with DFT, HF and MP2 calculations. *Chem. Phys.* 263(2):203-219.
7. Gaus M, Cui Q, & Elstner M (2014) Density functional tight binding: application to organic and biological molecules. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 4(1):49-61.
8. Torrie GM & Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* 23(2):187-199.
9. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, & Kollman PA (1995) Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* 16(11):1339-1350.
10. Efron B & Tibshirani RJ (1994) An introduction to the bootstrap (chapman & hall/crc monographs on statistics & applied probability).
11. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics* 3(5):300-313.
12. Steinbrecher T, Mobley DL, & Case DA (2007) Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *The Journal of chemical physics* 127(21):214108.
13. Camargo APM (2004) An electrochemical study of the adsorption and coadsorption behavior of selected purines, pyrimidines and nucleosides on Au (111). (Freie Universität Berlin).

14. Doublie S, Bandaru V, Bond JP, & Wallace SS (2004) The crystal structure of human endonuclease VIII-like 1 (NEIL1) reveals a zincless finger motif required for glycosylase activity. *Proceedings of the National Academy of Sciences of the United States of America* 101(28):10284-10289.