Supporting Information

The nucleotide excision repair lesion-recognition protein Rad4 captures a pre-flipped partner base in a benzo[a]pyrene-derived DNA lesion: how structure impacts the binding pathway

Hong Mu 1, Nicholas E. Geacintov 2, Jung-Hyun Min 4, Yingkai Zhang 2,3 and Suse Broyde 1,*

1 Department of Biology and 2 Department of Chemistry, New York University, New York, NY 10003, USA

3 NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

4 Department of Chemistry, University of Illinois at Chicago, Chicago, IL 60607, USA.

* To whom correspondence should be addressed. Suse Broyde, Biology Department, New York University, 100 Washington Square East, 1009 Silver Center, New York, NY 10003. Tel: +1 212 998 8231; Fax: +1 212 995 4015; Email: broyde@nyu.edu
Table of Contents

Supplementary Figures ................................................................................................ 3
Figure S1. Crystal structure of the productive open complex for Rad4-Rad23 with a CPD-containing DNA duplex (PDB ID:2QSG) \(^1\) ................................................................. 3
Figure S2. The Center of Mass (COM) distance reaction coordinates used for the preferred \textit{tightly coupled} binding pathway................................................................. 4
Figure S3. Exploration of \textit{stepwise} binding pathways from the First Capture State toward the productive open complex. ................................................................. 5
Figure S4. Clustering of the structural ensemble of the 600 ns unrestrained MD simulation starting from the initial model................................................................. 7
Figure S5. The initial binding of Rad4 to the \textit{cis}-B[a]P-dG lesion site with pre-flipped partner C: a First Capture State ensemble is achieved through unrestrained MD simulation. ................................................................. 9
Figure S6. The best representative structures for the (A) First capture, (B) Transition, (C) Intermediate, and (D) Productive Binding States. ............................................... 11
Figure S7. Structural details of Transition, Intermediate and the Productive Binding States of our preferred XPC binding pathway for the \textit{cis}-B[a]P-dG-containing duplex. See the Supplementary Movie for the structures along the dynamic pathway........ 12

Supplementary Tables ................................................................................................. 14
Table S1. DNA distortions of the 6-mer sequence between G\textsubscript{5}:C\textsubscript{5} and T\textsubscript{10}:A\textsubscript{10} that include the lesion site at G\textsubscript{8} \(^a\). .......................................................................................................................... 14
Table S2. Hydrogen bonds between BHD2-BHD3 and the lesion-containing DNA \(^a\) ..... 15

Supplementary Methods .......................................................................................... 17
Molecular modeling and structure preparation ......................................................... 17
Force field and MD simulation protocols................................................................. 17
Umbrella sampling and free energy calculation ...................................................... 18
Structural analyses ................................................................................................. 19

Supplementary References ..................................................................................... 19
Supplementary Figures

Figure S1. Crystal structure of the productive open complex for Rad4-Rad23 with a CPD-containing DNA duplex (PDB ID:2QSG).¹

Zoomed-in views of the regions denoted by colored squares (black, red and blue) are given on the right. The CPD lesion and BHD2 hairpin loop (Gly 518 to Glu 525) are not resolved in the crystal structure, and hence are represented by red and orange dashed lines. The overall structure is shown in cartoon, the side chain heavy atoms (providing van der Waals interactions) of Phe 599, Phe 597, Phe 556, Val 594, Arg 601, Pro 607, and Met 498 are in spheres, side chain heavy atoms (providing electrostatic interactions) of Arg 494, Glu 495, Asn 554, Asn 558 and Lys 606 are in sticks, and the heavy atoms of the nucleotides flanking the lesion site and of the partner nucleotides are in sticks. The TGD domain is yellow, BHD1 is blue, BHD2 is orange, BHD3 is dark green, and the DNA is light gray. The residues shown in spheres or sticks are also color-coded by atom type, with carbon in the same color as the domain, nitrogen in blue, phosphorus in orange, oxygen in red, and sulfur in yellow. Details of the opened lesion site, the binding pockets for flipped partner T bases, and the inserted BHD3 hairpin are shown in zoomed-in views from the minor groove side. Phe 599, Phe 597, and Phe 556 are essential for stabilizing the productive open complex by providing stacking interactions at the BHD3 hairpin insertion site and in the binding pockets for the flipped out partner T bases. Arg 494 and Glu 495 hydrogen bond with the DNA backbone phosphate oxygens of the partner strand (shown in dashed black lines), which forms the binding pocket for the 5’-side partner T together with Phe 597 and Met 498. Arg 601 stabilizes the base flanking the opened lesion site on the 5’ side of the lesioned strand (providing van der Waals interactions via its aliphatic portion). Lys 606 forms a hydrogen bond with the major groove side of the base flanking the opened lesion site on the 3’ side of the partner strand (shown in dashed black lines). Asn 554 and Asn 558 hydrogen bond to the 3’-side partner T, which forms the binding pocket for this base together with Phe 597, Phe 556, Val 594 and Pro 607.
The Center of Mass (COM) distance reaction coordinates used for the preferred tightly coupled binding pathway.

The COM distances are shown in black dashed lines in the structures when the extruded partner C₈ is first captured (First Capture State) and upon productive binding (Productive Binding State). The structures are rendered as in Figure 1D in the main text, and the heavy atoms used to define the COMs are shown in spheres. (A) Reaction coordinate 1 (RC1) that guides the BHD3 β-hairpin tip to insert into the helix at the lesion site. The reaction coordinate is defined as the distance between the COM for sugar ring heavy atoms of the nucleotides flanking the opened DNA nucleotide steps upon productive binding (gray spheres) and of the polypeptide backbone atoms of the three amino acids (Ser 603, Thr 604, and Val 605) at the insertion leading edge of the BHD3 β-hairpin (green spheres). (B) Reaction coordinate 2 (RC2) that guides G₇ to insert into its binding pocket in BHD3. The reaction coordinate is defined as the distance between the COM for the heavy atoms of the G₇ base (cyan spheres) and of the polypeptide backbone atoms of the amino acids lining the bottom of the G₇ binding pocket (Asn 554, Asn 558, Glu 560, Val 594 and Pro 607) (green spheres). (C) Reaction coordinate 3 (RC3) that guides the B[a]P rings to extrude into the minor groove. The reaction coordinate is defined as the distance between the COM of dT₆:dA₆ and dC₉:dG₉ sugar ring heavy atoms (the nucleotides flanking the opened lesion site upon productive binding, gray spheres) and of the B[a]P benzylic ring heavy atoms (red spheres). (D) Reaction coordinate 4 (RC4) that guides G₇ to flip out of the helix. The reaction coordinate is defined as the distance between the COM for the Watson-Crick hydrogen bonding edge heavy atoms (C6, O6, N1, C2, and N2) of G₇ (cyan spheres) and of the sugar ring heavy atoms of the nucleotides dC₇:dG₇ (gray spheres). For the preferred tightly coupled pathway, a generalized distance reaction coordinate, shown in the figure, was used for guiding the structural changes from the First Capture State (32.1 Å, Figure 3 of the main text) to the Productive Binding State (-16 Å, Figure 3 of the main text) to generate initial structures for each umbrella sampling window, and was used for umbrella sampling along this pathway.
**Figure S3.** Exploration of *stepwise* binding pathways from the First Capture State toward the productive open complex.

(A) The scheme for some examples of explored stepwise high energy binding pathways shows steps and Center of Mass (COM) distance reaction coordinates (RC) used for these examples of explored pathways. RC1 and RC2 are the same as defined in Figure S2. RC5, for inserting C₈ into its binding pocket, is defined as the COM for the heavy atoms of the C₈ base (blue spheres) and that of the backbone atoms of Met 498 at the bottom of the binding pocket for C₈.
(orange spheres). Umbrella sampling of the structures along the pathways toward the productive open complex (Figure S1), in which the BHD3 hairpin inserts into the helix at the lesion site and the flipped partner bases insert into their pockets, was performed with MD simulations with restraint forces along these RCs. Since the position of the B[a]P rings was not known for productive binding, no restraints were applied to the cis-B[a]P-dG lesion during umbrella sampling. (B) Example of an explored high energy stepwise pathway. The reaction coordinates and free energy profiles are shown along the pathway: First Capture State (FC) → Intermediate 1 (INT1) → Intermediate 2 (INT2) → Intermediate 3 (INT3) → Transition State (TS*) (along the red arrows in (A)). The structures and COM distances are rendered as in Figure S2. The free energy profiles were calculated with the RC values for the last 10 ns of the MD simulations for each step. The free energy of the intermediate state for each step is labeled. The overall free energy barrier from ICS to TS* is 21.4 kcal/mol (the sum of 2.9, 3.7, 3.5, and 11.3 kcal/mol from each step), in which the flipping of G₇ into its binding pocket (the last step) contributes the most to the overall free energy barrier. Flipping of G₇ along all other explored stepwise pathways is higher than 11.3 kcal/mol. (C) Stable productively bound complexes, termed Productive Binding State 1 (PB1), obtained from the unrestrained MD simulation starting from the TS*. The structure is rendered as in Figure 1 of the main text. Zoomed-in view is from the minor groove of the lesion site to reveal the cis-B[a]P-dG lesion that is extruded into the minor groove upon the full insertion of the BHD3 hairpin. The PB1 structure is superposed to the polypeptide backbone atoms in the crystal structure of the productive open complex (PDB ID: 2QSG) and all heavy atoms of the Productive Binding State (PB) obtained from the preferred tightly coupled pathway (Figure 2 of the main text). RMSD values are shown in the figure.
Figure S4. Clustering of the structural ensemble of the 600 ns unrestrained MD simulation starting from the initial model.

(A) The raw pairwise RMS distance matrices for the Rad4 - cis-B[a]P-dG-containing DNA complex. D1 is the RMS distance for the polypeptide backbone atoms of the Rad4 without the hairpin loops of the beta-hairpin domains. D2 is the RMS distance for heavy atoms of the 6-mer around the lesion site (G5:C5 and T10:A10, Figure 1 of the main text). D3 is the RMS distance for polypeptide backbone atoms of the BHD3 hairpin (Phe 590 to Ala 615). D4 is the RMS distance...
for polypeptide backbone atoms of the BHD2 hairpin (Lys 511 to Tyr 532). Cutoff is set at 2.5 Å to reveal the clustering based on each of these four distance matrices. (B) The pairwise reweighted distance matrix for clustering of the structural ensembles. In order to give equal weight to the conformational changes for different parts of the complex, the raw RMS distances were reweighted using the equation shown in the figure. The clustering cutoff was set at 0.5. The structure of the complex achieves a stable ensemble after 300 ns: this is the ensemble for the First Capture State (Figure S5).
Figure S5. The initial binding of Rad4 to the cis-B[a]P-dG lesion site with pre-flipped partner C: a First Capture State ensemble is achieved through unrestrained MD simulation.

(A) Achieving the First Capture State from the initial model (IM). Five representative snapshots are taken from the 600 ns trajectory to illustrate how the BHD2 hairpin inserts into the minor groove 3’ to the lesion site and how the flipped partner C is captured by the approaching BHD3 domain and Arg 494 of the BHD2 domain. The structures are rendered as in Figure 1D of the main text, with the DNA also shown in surface. The phosphate groups along the DNA backbone.
around the lesion site and the B[a]P benzylic hydroxyl group at C9 are color-coded: the phosphate group between A6 and G7 is dark blue, the phosphate group between G7 and C8 is marine, the phosphate group between C8 and G9 is pale cyan, the phosphate group between G11 and T10 is light pink, the phosphate group between T10 and C9 is hot pink, and the B[a]P benzylic hydroxyl group at C9 atom is purple. The time dependence values for the distances between the Lys 521 amino nitrogen atom and Center of Mass (COM) for each phosphate group or the benzylic hydroxyl group (Lys 521 – DNA backbone phosphate/lesion distance) are color coded accordingly, and show the steps that Lys 521 takes to guide the BHD2 hairpin insertion into the minor groove. The BHD2 - minor groove distance is measured as the distance between the COM of Val 517 and Thr 518 polypeptide backbone atoms and that of the sugar ring heavy atoms of dC8:dG9. BHD3 - TGD distance is the distance between the COM of the BHD3 domain polypeptide backbone atoms and that of the TGD domain polypeptide backbone atoms. The BHD3 hairpin tip – center of hairpin insertion site distance is defined the same as Reaction Coordinate 1 in Figure S2. The B[a]P – partner C distance is defined as the COM of the aromatic ring B (circled in black dashed lines) heavy atoms and that of the dC8 sugar ring heavy atoms. (B) Structural details of the First Capture State. The First Capture State structure is the best representative frame of the 300 – 600 ns ensemble, and is rendered as in Figure 2 of the main text. The view is into the minor groove for minor groove interactions, and is into the major groove for the major groove interactions. The amino acids, DNA bases and DNA backbone phosphate groups forming hydrogen bonds (black dashed lines) are shown in sticks, and the side chains of Val 517 and Thr 518 are shown in spheres. Black dashed lines with distances indicated are electrostatic interactions that are not within hydrogen bonding distances. These amino acids and DNA residues are color coded by atom type as in Figure 2 of the main text. The dynamics of the BHD2 hairpin tip is shown in the left panel, where BHD2 hairpins at 400ns and 450ns are shown in light orange and light maroon. The BHD2 domain has extensive interactions with the minor groove 3' to the lesion site: Arg 494, Gln 495, Arg 515, E523 and S603 hydrogen bond with the DNA backbone phosphate oxygens on the unmodified strand side of the lesion site (middle), Lys 521 and Lys 514 hydrogen bond with the DNA backbone phosphate oxygens on the lesioned strand (left), and Thr 516 and Val 517 insert into the minor groove 3' to the lesion (right) while Thr 516's hydroxyl group hydrogen bonds to O2 of T10 base (Table S2). Although the BHD3 domain does not have extensive interactions with the local lesion-containing region in this state, its β-hairpin is anchored on the major groove side by the stacking of Phe 599 with the pre-flipped partner base C8, the hydrogen bond between S603 and dG7 backbone phosphate oxygen, and the hydrogen bonds of Arg 601 with the major groove side of A4, G5, and A6 bases (Table S2). Hydrogen bond occupancies, distances and angles are given in Table S2, which also presents criteria for hydrogen bonding.
Figure S6. The best representative structures for the (A) First capture, (B) Transition, (C) Intermediate, and (D) Productive Binding States.

The structures are the best representative frames of each state. The structures are the same frames and are rendered as in Figure 2 of the main text.
Figure S7. Structural details of Transition, Intermediate and the Productive Binding States of our preferred XPC binding pathway for the cis-B[a]P-dG-containing duplex. See the Supplementary Movie for the structures along the dynamic pathway.

(A) Transition State. The red arrows show the insertion directions of F599 and R601 for BHD3 hairpin insertion into the helix (viewed from the major groove). (B) Intermediate State. (C) Productive Binding State. The structures are the same frames and are rendered as in Figure 2 of the main text. The amino acids, DNA bases and DNA backbone phosphate groups that form
hydrogen bonds are also shown in sticks, and the amino acid side chains that provide van der Waals interactions with DNA are shown in spheres. These amino acids and DNA residues are color coded by atom type as in Figure 2 of the main text. Hydrogen bond occupancies, distances and angles are given in Table S2, which also presents criteria for hydrogen bonding. Black dashed lines with distances indicated are electrostatic interactions that are not with in hydrogen bonding distances.
Supplementary Tables

Table S1. DNA distortions of the 6-mer sequence between G$_5$:C$_5$ and T$_{10}$:A$_{10}$ that include the lesion site at G$_8$.$^a$

|                                | Initial model | First Capture State | Transition State | Intermediate State | Productive Binding State |
|--------------------------------|---------------|---------------------|------------------|----------------------|--------------------------|
| Minor groove width A (Å)$^b$    | 10.4          | 11.9 ± 1.0          | 10.6 ± 0.6       | 11.9 ± 0.9           | 14.6 ± 1.2               |
| Minor groove width B (Å)$^b$    | 10.3          | 6.8 ± 0.8           | 6.4 ± 0.5        | 15.7 ± 1.3           | 16.4 ± 1.0               |
| Shift (Å)$^c$                   | 0             | 0 ± 2               | 7 ± 1            | 6 ± 2                | 3 ± 2                    |
| Slide (Å)$^c$                   | 0             | -6 ± 2              | -11 ± 2          | -5 ± 2               | -3 ± 1                   |
| Rise (Å)$^c$                    | 17            | 17 ± 1              | 19 ± 1           | 22 ± 1               | 21 ± 1                   |
| Tilt (°)$^c$                    | 27            | -1 ± 8              | 16 ± 7           | 17 ± 6               | 1 ± 6                    |
| Roll (°)$^c$                    | -2            | 14 ± 10             | -26 ± 12         | -7 ± 7               | -14 ± 11                 |
| Twist (°)$^c$                   | 142           | 115 ± 9             | 99 ± 9           | 91 ± 7               | 73 ± 6                   |
| Bend (°)$^d$                    | 23            | 34                  | 44               | 32                   | 45                       |

a. The values, except for those of the bend angle, are measured for the initial model and the 30 – 60 ns structural ensemble of the umbrella sampling window for each state (First Capture State at 32.1 Å, Transition State at 16.9 Å, Intermediate State at -3.0 Å and Productive Binding State at -16.0 Å, Figures 2 and 3 of the main text). The mean value and standard deviation for each ensemble are given.

b. The minor groove widths are defined as the distances between the pairs of DNA backbone phosphorus atoms (shown in Figures 2 and 4 of the main text) minus 5.8 Å to take into account the radii of phosphate groups.

c. The Shift, Slide, Rise, Tilt, Roll and Twist parameters are defined in 3DNA.$^2$ These parameters were measured over five nucleotide steps, between G$_5$:C$_5$ and T$_{10}$:A$_{10}$ of the lesion-containing 6-mer.

d. The helix axes of nucleotide steps 4 - 8 and 10 – 21 (Figure 1 of the main text) were obtained using Curves+.$^3$ The best fitted lines (orthogonal distance regression lines) for the helix axes (black dashed lines in Figure 2 of main text) were computed using MATLAB 7.10.0 (The MathWorks, Inc.). The values of bend angles around the lesion site are measured between the fitted lines of the helical axes; for the initial model the coordinates of the DNA were used, and for the other cases the DNA structures of the best representative frames from the ensemble were employed (Figure 2 of main text).
Table S2. Hydrogen bonds between BHD2-BHD3 and the lesion-containing DNA.  

| Acceptor \ Donor Hydrogen | Donor b | Occupancy (\%) | Average Distance (Å) | Average Angle (˚) |
|---------------------------|---------|----------------|----------------------|------------------|
| **First Capture State**   |         |                |                      |                  |
| C₈@OP1                    | R494@HH22 | 98             | 2.8                  | 151              |
| C₈@OP1                    | R494@HH12 | 94             | 2.9                  | 149              |
| C₈@OP1                    | R494@HH11 | 53             | 3.0                  | 157              |
| G₉@OP1                    | Q495@HE22 | 48             | 3.0                  | 152              |
| A₁₀@O3'                   | Q495@HE22 | 32             | 3.1                  | 148              |
| A₁₂@OP1                   | K514@HZ3  | 19             | 2.9                  | 155              |
| A₁₂@OP1                   | K514@HZ1  | 18             | 2.9                  | 153              |
| A₁₂@OP1                   | K514@HZ2  | 17             | 2.8                  | 155              |
| C₈@OP2                    | R515@HH21 | 49             | 2.9                  | 153              |
| C₈@OP2                    | R515@HE   | 42             | 3.0                  | 148              |
| T₁₀@O2                    | T516@HG1  | 55             | 2.7                  | 157              |
| T₁₀@OP2                   | K521@HZ3  | 21             | 2.8                  | 154              |
| T₁₀@OP2                   | K521@HZ1  | 20             | 2.8                  | 154              |
| T₁₀@OP2                   | K521@HZ2  | 19             | 2.8                  | 154              |
| A₆@OP2                    | E523@H    | 66             | 2.9                  | 158              |
| R₆₀¹@O                    | A₆@H62    | 59             | 3.0                  | 158              |
| A₆@N7                     | R₆₀¹@HH11 | 43             | 3.0                  | 147              |
| G₉@O6                     | R₆₀¹@HH11 | 30             | 3.0                  | 146              |
| A₆@N7                     | R₆₀¹@HH12 | 29             | 3.1                  | 140              |
| G₇@OP1                    | S₆₀³@H    | 44             | 3.0                  | 144              |
| G₇@OP1                    | S₆₀³@HG   | 35             | 2.8                  | 160              |
| **Transition State**      |         |                |                      |                  |
| G₉@OP2                    | S₄₉₂@HG   | 38             | 2.7                  | 164              |
| C₈@OP1                    | R₄₉₄@HH22 | 84             | 2.9                  | 162              |
| C₈@OP2                    | R₄₉₄@HH21 | 69             | 3.1                  | 143              |
| G₉@OP1                    | R₄₉₄@HE   | 52             | 2.8                  | 156              |
| G₉@OP2                    | Q₄₉₅@HE22 | 64             | 3.0                  | 158              |
| A₁₀@O3'                   | Q₄₉₅@HE22 | 42             | 3.0                  | 155              |
| C₈@OP2                    | R₅₁₅@HH21 | 16             | 2.8                  | 159              |
| C₈@OP2                    | R₅₁₅@HH22 | 10             | 2.9                  | 155              |
| T₁₀@OP2                   | K₅₂₁@HZ1  | 28             | 2.8                  | 155              |
| T₁₀@OP2                   | K₅₂₁@HZ2  | 23             | 2.8                  | 154              |
| Acceptor   | Donor Hydrogen | Donor       | Occupancy (%) | Average Distance (Å) | Average Angle (°) |
|------------|----------------|-------------|---------------|----------------------|-------------------|
| T_{10}@OP2 | K521@HZ3       | K521@NZ     | 21            | 2.8                  | 155               |
| E600@N    | G_{7}@H22      | G_{7}@N2    | 82            | 3.2                  | 140               |
| E600@O    | G_{7}@H22      | G_{7}@N2    | 50            | 3.0                  | 136               |
| T604@O    | G_{7}@H1       | G_{7}@N1    | 90            | 3.0                  | 155               |

**Intermediate State**

| C_{8}@OP2 | R515@HH12     | R515@NH1    | 56            | 2.8                  | 161               |
| C_{8}@OP2 | R515@HH11     | R515@NH1    | 53            | 2.9                  | 159               |
| C_{8}@OP1 | R494@HH22     | R494@NH2    | 45            | 2.8                  | 153               |
| G_{9}@OP1 | R494@HH11     | R494@NH1    | 44            | 2.9                  | 148               |
| C_{8}@OP1 | R494@HH12     | R494@NH1    | 38            | 2.8                  | 149               |
| G_{9}@OP2 | Q495@HE22     | Q495@NE2    | 35            | 2.9                  | 161               |
| C_{7}@N3  | R601@HH11     | R601@NH1    | 43            | 3.2                  | 147               |
| T_{6}@O2  | S603@HG       | S603@OG     | 35            | 3.0                  | 150               |

**Productive Binding State**

| C_{8}@OP1 | R494@HH22     | R494@NH2    | 98            | 2.8                  | 153               |
| C_{8}@OP1 | R494@HH12     | R494@NH1    | 93            | 2.9                  | 149               |
| G_{9}@OP2 | R494@HH21     | R494@NH2    | 93            | 2.9                  | 158               |
| C_{8}@O5' | R494@HH12     | R494@NH1    | 13            | 3.4                  | 143               |
| C_{8}@OP2 | R519@HH11     | R519@NH1    | 95            | 2.8                  | 160               |
| G_{7}@O6  | N554@HD21     | N554@ND2    | 95            | 3.0                  | 160               |
| G_{7}@O6  | N558@HD22     | N558@ND2    | 98            | 2.9                  | 163               |
| A_{12}@OP1| T516@HG1      | T516@OG1    | 34            | 3.2                  | 150               |
| G_{11}@O3'| T516@HG1      | T516@OG1    | 17            | 3.1                  | 148               |
| C_{7}@O2  | R601@HH11     | R601@NH1    | 42            | 3.0                  | 160               |
| C_{7}@N3  | R601@HH11     | R601@NH1    | 16            | 3.1                  | 147               |
| A_{6}@O4' | T604@HG1      | T604@OG1    | 58            | 2.8                  | 155               |
| A_{6}@N7  | K606@HZ1      | K606@NZ     | 23            | 3.0                  | 152               |
| A_{6}@N7  | K606@HZ2      | K606@NZ     | 20            | 3.0                  | 152               |
| A_{6}@N7  | K606@HZ3      | K606@NZ     | 18            | 3.0                  | 153               |

a. Criteria for hydrogen bond formation: donor-acceptor distance < 3.5 Å and donor-hydrogen-acceptor angle > 135 °.
b. The acceptor, donor hydrogen and donor atoms for each hydrogen bond are labeled as "residue name@atom name".
Supplementary Methods

Molecular modeling and structure preparation

The initial model. Previously we obtained an MD equilibrated initial model (termed ‘docking complex’) for Rad4 with a CPD containing DNA duplex (Figure 4 of the main text), which represented the state of XPC translocated to the damage but prior to initial binding by the BHD2 and BHD3 domains. In brief, an apo Rad4 model, based on the crystal structure of apo Rad4 (PDB ID: 2QSF), was docked on to the damaged DNA with its TGD and BHD1 domains positioned as in the crystal structure of the productive open complex (PDB ID: 2QSG). The docked complex was subjected to MD simulation with restraint on the distance between BHD2 and BHD3 domains and the lesion site, which provided an equilibrated ‘docking complex’ of the Rad4 with CPD-containing DNA (Figure 4 of the main text). For the present study, the initial model of the Rad4 with the cis-B[a]P-dG-containing DNA was based on the equilibrated Rad4—CPD-containing DNA ‘docking complex’: the DNA duplex was modified to a cis-B[a]P-dG-containing DNA by replacing a 5-mer sequence around the CPD with the central 5-mer sequence from the NMR solution structure of a cis-B[a]P-dG-containing DNA with normal partner C (Figure 1A of the main text). We first performed 250 ns of MD simulation for the NMR solution structure and selected the most representative structure from the ensemble to replace the 5-mer sequence around the CPD. We extended the 5-mer sequence for five base pairs to the 5’-side of the lesioned strand (base pairs 1-5 in Figure 1C of the main text) using standard B-DNA in Discovery Studio 2.5 (Accelrys Software Inc.); the side chain of Arg 601 at the BHD3 β-hairpin tip was adjusted to avoid collision with the DNA, and the phenyl ring of Phe 599, the first residue along the ‘flipping path’ for the partner bases, was rotated toward the pre-flipped partner C to facilitate initial binding.

Protonation. The protonation states of charged residues were determined by the H++ and pdb2pqr methods. When the two methods disagreed, protonation states were assigned based on the following criteria: the potential H-bonding network, solvent exposure of the ionizable residues, potential steric clashes if the proton was added, and preservation of the crystal structure. The results are given in Mu et al.

Water box and counterions. The protein-DNA complex was neutralized with Na+ counterions and solvated with explicit TIP3P water in a cubic periodic box with side length of 125.0 Å using the tLEAP module of the AMBER14 suite of programs.

Force field and MD simulation protocols

All MD simulations were carried out using the AMBER14 suite of programs. We used the ff14SB force field for all MD simulations. Partial charges and parameters for the cis-B[a]P-dG lesion were obtained from Mocquet et al. The Particle-Mesh Ewald method with 9.0 Å cutoff for the non-bounded interactions was used in the energy minimizations and MD simulations. Minimizations were carried out in three stages. First, 500 steps of steepest descent minimization followed by 500 cycles of conjugate gradient minimization were conducted on the water molecules and counterions with a restraint force constant of 50 kcal/(mol·Å²) on the solute
molecules (DNA or DNA and protein complex). Then, 500 steps of steepest descent minimization followed by 500 cycles of conjugate gradient minimization were carried out on the water molecules and counterions with a restraint force constant of 10 kcal/(mol·Å²) on the solute molecules. In the last round, 500 steps of steepest descent minimization followed by 500 cycles of conjugate gradient minimization were carried out on the whole system without restraints. The minimized structure was then subjected to three rounds of equilibration. First, each system was equilibrated at constant temperature of 10K for 30ps with the solute molecules fixed with a restraint force constant of 25 kcal/(mol·Å²). Then the system was heated from 10 K to 300 K over 30 ps with the solute molecules fixed with a restraint force constant of 10 kcal/(mol·Å²) at constant volume. In the last round of equilibration, the restraint force constant on the solute was reduced through three steps: at 10 kcal/(mol·Å²) for 30 ps, at 1 kcal/(mol·Å²) for 40 ps, then at 0.1 kcal/(mol·Å²) for 50 ps with constant pressure at 300K. Following equilibration, production MD simulations for each system were carried out in an NPT ensemble at 300K and constant pressure of 1 Atm. The temperature was controlled with a Langevin thermostat with a 5 ps⁻¹ collision frequency. The pressure was maintained with the Berendsen coupling method. For all unrestrained MDs, the production MDs were carried out with 1 kcal/(mol·Å²) restraints on the distances between each pair of hydrogen bonded heavy atoms at the end base pairs, but not on any other residues of the complex. Details of the restrained MDs are given below. A 2.0 fs time step and the SHAKE algorithm were applied in all MD simulations.

Umbrella sampling and free energy calculation

The steps of our explorations of Rad4 binding pathways are summarized in Figure S3A. Structural changes were directed from the First Capture State (Figure S5) toward the productive open complex using MD simulations with restraints along the selected reaction coordinates (Figures S2, S3A and S3B) in small steps (umbrella windows). The reaction coordinates were designed to guide partner bases to insert into their binding pockets, the BHD3 β-hairpin to insert into the DNA helix at the lesion site, and the B[a]P to extrude into the minor groove, in order to achieve the productive open complex (Figures S2 and S3C).

Since the positioning of the cis-B[a]P-dG lesion for the productive binding complex is unknown, we initially explored the binding pathway with many different sequences of structural events, such as flipping partner base G₇, inserting partner bases into their pockets, and inserting the BHD3 β-hairpin into the DNA helix; however there were no restraints on the cis-B[a]P-dG lesion’s position (Figures S3). During this exploration, the final productive binding state was obtained through unrestrained MDs from the high energy transition state, revealing that the B[a]P ring system is extruded into the minor groove in the productive binding state (Figure S3C), but that stepwise pathways were not energetically feasible. Therefore, we designed a generalized distance reaction coordinate to explore the binding pathway with tightly coupled partner G₇ flipping, G₇ insertion into its binding pocket, BHD3 β-hairpin insertion and B[a]P rings extrusion (Figure S2).

For each step along the explored stepwise binding pathway, a center of mass (COM) distance reaction coordinate was defined (Figure S3). For the tightly coupled binding pathway, a generalized distance reaction coordinate was defined with four COM distances (Figure S2). To
generate initial structures for sampling each window, we carried out 200 - 500 ps MD simulations with restraint force constant in the range of 10-100 kcal/(mol × Å²) along the reaction coordinate through continuous umbrella windows at 0.2-1.0 Å intervals. Then we carried out the sampling for each window using 60 ns MD simulation starting from the initial structure of each window; restraint force constants in the range of 1-10 kcal/(mol × Å²) along the reaction coordinate were used to ensure sufficient overlap between windows. The free energy profiles were calculated using vFEP, with reaction coordinate data collected every 1 ps from the 30 - 60 ns ensemble of the sampling MD simulations.17

Structural analyses

The First Capture State was obtained through unrestrained MD simulation started from the initial model (Figures S4-S6 and Figure 1B of the main text). The trajectory was clustered using a combined parameter (Figure S4), which weighs the structural change of the Rad4, lesion-containing 6-mer sequence (G₅:C₅ and T₁₀:A₁₀), the BHD2 β-hairpin, and the BHD3 β-hairpin equally. A stable cluster was obtained from the 300 ns to 600 ns trajectory and was defined as the First Capture State ensemble (Figure S5). The structural properties of each umbrella sampling window along the preferred binding pathway were obtained from the structural ensemble of the 30 ns to 60 ns trajectory. The best representative structure is defined as the one frame that has the shortest distance (the combined parameter, Figure S4) to all other frames in each ensemble. The structural properties were calculated using the cpptraj module of AMBER14, Curves+ and MATLAB 7.10.0 (The MathWorks, Inc.), and plotted using MATLAB 7.10.0 (The MathWorks, Inc.).3, 10 All molecular structures were rendered using PyMOL 1.3.x (Schroddinger, LLC.).

Supplementary References

(1) Min, J. H., and Pavletich, N. P. (2007) Recognition of DNA damage by the Rad4 nucleotide excision repair protein. Nature 449, 570-575.
(2) Lu, X. J., and Olson, W. K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res. 31, 5108-5121.
(3) Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D., and Zakrzewska, K. (2009) Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Res. 37, 5917-5929.
(4) Cosman, M., de los Santos, C., Fiala, R., Hingerty, B. E., Ibanez, V., Luna, E., Harvey, R., Geacintov, N. E., Broyde, S., and Patel, D. J. (1993) Solution conformation of the (+)-cis-anti-[BP]dG adduct in a DNA duplex: intercalation of the covalently attached benzo[a]pyrenyl ring into the helix and displacement of the modified deoxyguanosine. Biochemistry 32, 4145-4155.
(5) Mu, H., Kropachev, K., Chen, Y., Zhang, H., Cai, Y., Geacintov, N. E., and Broyde, S. (2013) Role of structural and energetic factors in regulating repair of a bulky DNA lesion with different opposite partner bases. Biochemistry 52, 5517-5521.
(6) Anandakrishnan, R., Aguilar, B., and Onufriev, A. V. (2012) H++3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic acids research 40, W537-W541.
(7) Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic acids research* 32, W665-W667.

(8) Mu, H., Geacintov, N. E., Zhang, Y., and Broyde, S. (2015) Recognition of Damaged DNA for Nucleotide Excision Repair: A Correlated Motion Mechanism with a Mismatched cis-syn Thymine Dimer Lesion. *Biochemistry* 54, 5263-5267.

(9) Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* 79, 926-935.

(10) Case, D. A., Babin, V., Berryman, J. T., Betz, R. M., Cai, Q., Cerutti, D. S., Cheatham, T. E., 3rd, Darden, T. A., Duke, R. E., Gohlke, H., Goetz, A. W., Gussarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossváry, I., Kovalenko, A., Lee, T. S., LeGrand, S., Luchko, T., Luo, R., Madej, B., Merz, K. M., Paesani, F., Roe, D. R., Roitberg, A., Sagui, C., Salomon-Ferrer, R., Seabra, G., Simmerling, C. L., Smith, W., Swails, J., Walker, R. C., Wang, J., Wolf, R. M., Wu, X., and Kollman, P. A. (2014) AMBER14, University of California, San Francisco.

(11) Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* 11, 3696-3713.

(12) Mocquet, V., Kropachev, K., Kolbanovskiy, M., Kolbanovskiy, A., Tapias, A., Cai, Y., Broyde, S., Geacintov, N. E., and Egly, J. M. (2007) The human DNA repair factor XPC-HR23B distinguishes stereoisomeric benzo[a]pyrenyl-DNA lesions. *EMBO J.* 26, 2923-2932.

(13) Darden, T., York, D., and Pedersen, L. (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* 98, 10089-10092.

(14) Loncharich, R. J., Brooks, B. R., and Pastor, R. W. (1992) Langevin Dynamics of Peptides - the Frictional Dependence of Isomerization Rates of N-Acetyllalanyl-N'-Methylamide. *Biopolymers* 32, 523-535.

(15) Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A., and Haak, J. R. (1984) Molecular-Dynamics with Coupling to an External Bath. *J Chem Phys* 81, 3684-3690.

(16) Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J Comput Phys* 23, 327-341.

(17) Lee, T. S., Radak, B. K., Pabis, A., and York, D. M. (2013) A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations. *Journal of chemical theory and computation* 9, 153-164.