Published in final edited form as:
J Phys Chem Lett. 2024 March 21; 15(11): 3206–3213. doi:10.1021/acs.jpclett.4c00161.

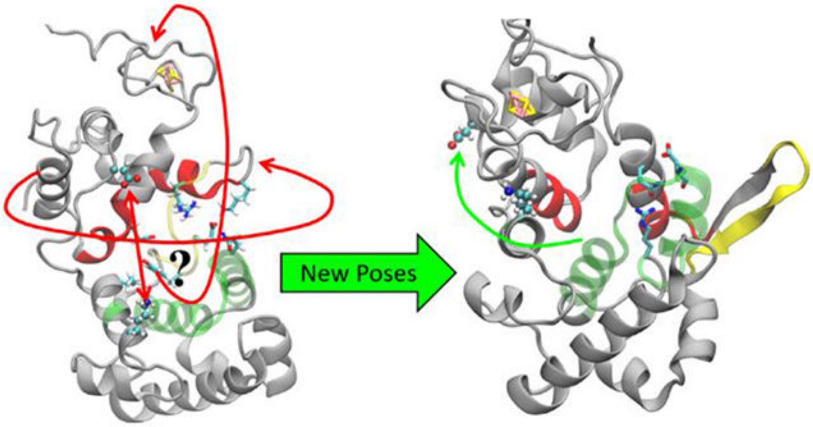
Enhanced Sampling for Conformational Changes and Molecular Mechanisms of Human NTHL1

Ryan E. Odstrcil,
Prashanta Dutta,
Jin Liu[*]
School of Mechanical and Materials Engineering, Washington State University, Pullman, Washington 99164, USA

Abstract

The functionalities of proteins rely on protein conformational changes during many process. Identification of the protein conformations and capturing transitions among different conformations are important but extremely challenging in both experiments and simulations. In this work, we develop a machine learning based approach to identify a reaction coordinate that accelerates the exploration of protein conformational changes in molecular simulations. We implement our approach to study the conformational changes of human NTHL1 during DNA repair. Our results identified three distinct conformations: open (stable), closed (unstable), and bundle (stable). The existence of this bundle conformation can rationalize the recent experimental observations. Comparison with an NTHL1 mutant demonstrates that a closely packed cluster of positively charged residues in the linker could be a factor to search for in the genes encoding when screening for genetic abnormalities. Results will lead to better modulation of the DNA repair pathway to protect against carcinogenesis.

Graphical Abstract



[*] jin.liu2@wsu.edu (Prof. J. Liu).

Supporting Information
Details of numerical methods, one table and eleven figures to demonstrate the numerical convergence and illustrate protein structures/interactions described in this paper. This information is available free of charge via the Internet at http://pubs.acs.org

As biological macromolecules age, they are subject to spontaneous decomposition and DNA is no exception. DNA can be damaged by hydrolysis, oxidation, and nonenzymatic methylation at significant rates *in vivo*, and the accumulation of the DNA decay may contribute to spontaneous mutagenesis and carcinogenesis[1]. Damaged DNA cannot be replaced and must therefore be repaired to remain intact[2–3]. In human cells, one such approach is base excision repair (BER) that actively removes non-bulky DNA lesions[4–7]. This approach has several basic steps, including lesion recognition, excision of the damaged nucleotide, and resynthesis using error-free DNA polymerases[8–10]. The BER pathway processes around 30,000 damaged base lesions per cell per day and defects in the pathway can lead to various diseases, including cancer[11–12]. Therefore, identifying the impact of BER proteins in disease progression is of paramount importance[13].

DNA glycosylases recognize and remove the specific base damages to produce an apurinic/apyrimidinic site. These molecules can thus be seen as the first barricade in the BER pathway for fending against cellular mutations that may lead to increased mutagenesis or genomic alterations[14–15]. Within this class of molecules, NTHL1 is a bifunctional DNA glycosylase that is important for the removal of oxidative pyrimidine damage[16]. The loss of NTHL1 can create high lifetime risks for adenomas and other types of tumors, such as colorectal cancer or adenomatous polyposis[11, 17–19]. It has been shown that some variants of NTHL1 have defective repair activities for BER[20–21]. Additionally, NTHL1 is upregulated in some cancers[22] and when the balance between NTHL1 expression and other DNA repair pathways is upset, the dysregulation could impact cancer progression[16, 23–24]. Identifying mechanisms of how this protein functions and what mutations or inhibitors may impact its functions, can lead to improved modulation of the BER pathway.

Efforts to understand the functions of human NTHL1 have been hampered by a lack of available crystal structures that have only recently been discovered[25]. The structure of human NTHL1 shows similarities to its bacterial homologs[26] with two globular helical domains: a six-helical bundle domain containing a helix-hairpin-helix DNA-binding motif, and a helical domain containing a [4FE-4S] cluster, referred to hereafter as the hairpin and cluster domains. During DNA binding in bacterial homologs, a lysine (K220) and aspartate (D239) in the hairpin domain and cluster domain, respectively, are involved in catalytic reactions with damaged DNA[26]. However, the human NTHL1 was crystallized in a significantly different conformation than other homologs[26–27]. Other homologs were found in a closed conformation where the distance between catalytic residues was around 5 Å while the human NTHL1 structure was captured in a novel open conformation with the catalytic residues at a distance of around 23 Å. The human NTHL1 needs to undergo large scale conformational changes in order for catalysis to occur and Carroll *et al.*[25] discovered that a flexible linker (residues 110-125) between the two domains is necessary for this conformational change. When the flexible linker was substituted with a shorter sequence from a homolog EcoNth (residues 21-28), the mutated NTHL1 (NTHL1^{m}) crystallized in a closed conformation and had reduced activity in lesion-containing DNA. Since the large scale interdomain rearrangements in NTHL1 are unprecedented and are necessary for proper functioning, insights into the molecular mechanisms could prove beneficial for identifying and mitigating possible NTHL1-related health risks.

Experimental methods for investigating the dynamics of proteins can be expensive and resource intensive. Molecular Dynamics (MD) is an effective tool for investigating the underlying dynamics of proteins that cannot be captured in experiments. But traditional MD simulations cannot resolve many protein processes, occurring over timescales of hundreds of nanoseconds to even seconds. Enhanced sampling techniques have been developed to accelerate MD simulations by biasing the molecular system to reduce energy barriers along reaction pathways[28–29]. However, the identification of these reaction pathways – typically described in terms of a low-dimensional variable referred to as a reaction coordinate – can be difficult due to a chicken-and-egg problem: obtaining a useful reaction coordinate requires knowledge of the reaction pathway, but the reaction pathway can only be explored by accelerating sampling along the reaction coordinate. Machine learning can be leveraged to identify reaction coordinates[30–33] but these machine learning methods often rely on the observation of the reaction before predicting a reaction coordinate, which can be impractical if there is a large ensemble of conformations or large energy barriers for the molecular system. Our recently developed approach Log-Probability Estimation via Invertible Neural Networks for Enhanced Sampling (LINES)[34–35] circumvents this issue by learning the free energy surface (FES) based on molecular coordinates using a Normalizing Flow[36–37] machine learning model and then predicts reaction coordinates based on the FES gradients with respect to each molecular coordinate. LINES has foundations in local optimization methods, thus presenting an attractive alternative to other machine learning methods because LINES strives to predict reaction pathways before reactions completely occur, speeding up sampling in simulations and the convergence of reaction coordinate predictions.

In this work, we design and implement the LINES method to identify a reaction coordinate that can accelerate the sampling of complicated NTHL1 conformations during the DNA repair. An iterative process of running biased MD simulations to explore conformations, training a machine learning model to learn the FES, and then extracting a reaction coordinate is performed. The process will predict a reaction coordinate that describes and accelerates motions of the linker region and the two domains of NTHL1. Next, long biased MD simulations are run to explore the conformational spaces of NTHL1 and to identify a distinct “open”, “closed”, and “bundle” conformational state in the FES. The stability of these conformations is evaluated with unbiased MD simulation and finally the closed conformation is compared to the stable NTHL1^{m} conformation to elucidate the importance of the linker region in NTHL1. The overall algorithm flow is depicted in Fig. 1, numerical details and simulation setup can be found in Supporting Information.

The first step for accelerating sampling of NTHL1 conformations is to identify a reaction coordinate using LINES. With (un)biased MD simulations of 60 ns and the normalizing flow model as described Method section, LINES converges to the reaction coordinate within just 7 iterations of LINES. Fig. 2(a) shows the converged coefficients for all the molecular coordinates. As indicated in Figure S2, the reaction coordinate converges to > 95% similarity between iterations 5 and 6, but an additional iteration of LINES is run to ensure a consistent prediction. The converged reaction coordinate is able to improve the rate of sampling along reaction pathways, potentially between the open and closed conformations based on the cleft distance in Figure S3, demonstrating that LINES is able to identify and accelerate slow reactions involving the linker region of NTHL1. As shown in Fig. 2(a), the important molecular coordinates in the reaction(s) are $MC10 \gg MC9 > MC1 > MC4$. Inspection of the reference groups associated with these molecular coordinates reveals that MC10, MC9, and MC1 describe both linker-cluster and linker-hairpin domain interactions. These interactions stabilize the linker region in the “open” conformation identified by the crystal structure. The last molecular coordinate MC4 describes an interaction stabilizing the middle region of the linker near the center of the DNA-binding cleft of NTHL1.

With the predicted reaction coordinate, several long, biased MD simulations are run to obtain a FES. In these simulations, three simultaneous replicas are run for 600 ns each and all share the same biasing potential. The shared biasing potential accelerates the convergence of the biasing potential, leading to improved sampling. The 600 ns simulations prove sufficient in sampling based on comparisons of the FES predictions, as demonstrated in Figure S4. The converged FES, shown in Fig. 2(b), identifies 3 distinct conformations that are all separated by energy barriers of at least 3-5 $k_B T$. In these energy basins, representative conformations from each cluster are randomly selected and shown in Fig. 2(c). A comparison of these conformations reveals that there is an open, closed, and “bundle” conformation. The open conformation, centered around a cleft distance of 39 Å and RC value of −2, has the lowest energy, showing consistency with the experimental structure of NTHL1 that was crystallized in an open conformation[25]. In the open conformation, the linker region is nestled between the cluster and hairpin domains of NTHL1. For the closed conformation, the catalytic residues are within 5 Å of each other and the linker region is no longer in the cleft between the two domains. In this conformation, the values of $MC4$ change significantly, showing the importance of this molecular coordinate. Intriguingly, our simulations also identified a stable “bundle” conformation, which was unexpected by us initially. Through the cluster domain rotation, helices in this domain collapse into the DNA-binding cleft, leading to a structure very similar to the helical “bundles” with the linker region exposed to the solvent as illustrated in Fig. 2(c). In this conformation, the values of MC10, MC9, and MC1 have drastically changed, demonstrating that the reaction coordinate identified by LINES distinguishes between the open, bundle, and closed conformations. The relative free energy difference between the three energy basins and the open conformation are also calculated by integrating the FES as $\Delta FE_o^i = k_B T * log\left(\iint exp\left(-\frac{V_{bias}}{k_B T}\right) dxdy\right)$,[38] to demonstrate the relative stability of each conformation.

To demonstrate how the reaction coordinate predicted by LINES can identify reaction pathways, another set of three simultaneous replica simulations are run for 600 ns and all share the same biasing potential computed with OPES. However, the biasing potential used in these simulations is a function of the root mean square deviation (RMSD) compared to the open and closed conformations. After 600 ns of simulations, no noticeable sampling of reaction pathways was revealed. The simulations are unable to sample alternative conformations besides the open conformation shown in Fig. 2(c) and cannot capture any energy barriers between conformations, as shown in Figure S5. This indicates that RMSD reaction coordinates cannot improve sampling, but the LINES-predicted reaction coordinate greatly accelerates simulations and leads to conformation exploration.

The stabilities of each of the conformations in Fig. 2(c) were then evaluated with 100 ns unbiased MD simulations. As shown in Fig. 2(d), the time evolution of the root-mean-square-deviation (RMSD) of the C_α backbone atoms revealed that both the open and bundle conformations are stable while the closed conformation is unstable. Closer inspection of the trajectories reveals that the linker region in the open conformation, shown in Figure S6, is stabilized by a variety of polar and hydrophobic interactions. For example, Y119 experiences hydrophobic interactions with V228 but also forms hydrogen bonds with E277; E116 forms salt bridges with R103, K106, and Q145. These interactions, illustrated in Figure S7, along with other interactions hold the linker in place throughout the duration of the 100 ns MD simulations. The bundle conformation, shown in Figure S8, removes the interactions of the linker region with the cluster and hairpin domains, and allows for strong interactions between the two domains to form directly. As illustrated in Figure S9, Q94 forms strong interactions with D144 and R155, R128 forms a salt bridge with E277, and hydrogen bonds occur between the backbone atoms of K105 and Y119. Attempts were made to create a stable closed conformation, such as applying backbone restraints to induce a similar conformation to NTHL1^m and create stabilizing interactions identified by Carroll *et al.*[25], or running multiple replicas, but no stable closed conformation has been achieved. The findings from our simulations are consistent with the experiments[25], where the lowest energy conformation for human NTHL1 had been determined as the open rather than a closed conformation. It is possible that the presence of damaged DNA in the binding cleft would increase the stability of the closed conformation of NTHL1 and prevent a collapse of the cluster domain into the binding cleft.

Since the open and bundle conformations are very stable from our simulations, capturing the conformational transitions among open, closed, bundle or other conformations would be computationally infeasible. These conformations can be identified and accessed through enhanced sampling along a reaction coordinate predicted by our LINES method, demonstrating the power of the method for accelerating MD simulation.

In the experiments from Carroll *et al.*[25], the wild type human NTHL1 was crystallized in an open conformation. However, mutating the linker region with a shorter sequence from a homologous structure into NTHL1^m altered the dynamics of the enzyme, preventing an opening of the DNA binding cleft and causing the mutant to crystallize in the closed conformation. Similar to the unbiased MD simulations previously used to evaluate conformational stabilities, the NTHL1^m conformation crystal structure was simulated for 100 ns and compared to the wild type closed conformation obtained from the FES in Fig. 2. As shown in Fig. 3(a), simulations started from very similar structures for both wild type and mutant NTHL1. After 100 ns, though, the wild type structure fails to stabilize, as indicated in Fig. 3(b), whereas the mutant structure is stabilized by the formation of salt bridges between D107^m and E112^m in the linker region with R289^m and R148^m in the cluster and hairpin domains, respectively.

Even though the wild type NTHL1 linker region contains the same residues needed for the salt bridge formation as the mutant, the wild type residues are dispersed over a longer segment of the linker. As a result, wild type NTHL1 does not create a dense positively charged region to attract the cluster and hairpin domains together that would otherwise cause

NTHL1 to favor the closed conformation. Attempts were made to artificially impose this salt bridge formation for the wild type NTHL1 by applying harmonic restraints between the linker residues D111 and E116 with R153 and R296 using PLUMED[39]. For the restraints, a spring constant of 20 kJ/mol was used and the target equilibrium distance was set to 8 Å. The restrained system was simulated for 100 ns to ensure that the protein had reached an equilibrium, at which point the restraints were released. A subsequent 100 ns unbiased MD simulation revealed that imposing the restraints to create a salt bridge caused the flexible linker region to interact with the DNA binding cleft, and drove the NTHL1 system towards the open conformation where the catalytic residue distance was 15 Å, as shown in Figure S10. Our results indicate that genetic mutations of NTHL1 that create closely packed clusters of positively charged residues in the linker would likely shift the equilibrium from the open to the closed conformation and decrease the effectiveness of NTHL1 for identifying damaged DNA bases and initiating the BER pathway.

To determine if there are any other conformations in NTHL1^{m}, a machine learning analysis analogous to the wild type NTHL1 is performed. The molecular coordinates and reaction coordinate coefficients are shown in Figure S11(a–c). The molecular coordinates were created from an analogous subset of the molecular coordinates used in the wild type NTHL1 analysis. For example, the preserved residues in the mutated linker region T111^{m}, E112^{m}, and S116^{m} are also present in the wild type NTHL1 structure as T115, E116, and S121. Since some of the reference groups in the wild type NTHL1 are not present in NTHL1^{m}, the number of molecular coordinates is reduced from 14 to 8. The LINES simulations converge very rapidly and show >95% similarity across 5 iterations of LINES, as shown in Figure S11(d). A set of three replica simulations with an OPES bias potential is also performed, with the resulting FES from 600 ns of simulation shown in Figure S11(e). The FES shows two stable conformations at cleft distances of 1 nm and 2.4 nm – which correspond to the closed and open conformations, respectively. The closed conformation has the lowest free energy. There is a clear absence of the "bundle" conformation in the FES, that was observed for the wild type NTHL1 structure. Carroll *et al.*[25] proposed that a conformation change occurs in the wild type following lesion recognition yet this conformation change was not observed in the NTHL1^{m} system; the increased closed conformation stability and absence of a "bundle" conformation from these NTHL1^{m} simulations demonstrate the impact of the linker region on the energetics and dynamics of the NTHL1 system for shifting the equilibrium towards the closed conformation.

Based on the increased stability of the mutated closed conformation and the wild type open/bundle conformations, a possible mechanism for the reaction pathway of NTHL1 is proposed. In this mechanism, as illustrated in Fig. 4, the low energy open conformation is stabilized by charged, polar, and nonpolar interactions between the linker region and the cluster/hairpin domains. Among them, E116 forms salt bridges and Y119 experiences polar/nonpolar interactions. During the first part of the reaction, the linker region leaves the DNA binding cleft between the cluster and hairpin domains, causing the cleft to narrow and the catalytic residue distance to decrease as the domains come together. While the interactions with DNA were not explicitly modeled in this work, it would be during this conformation change that the damaged DNA enters the binding cleft, allowing the catalytic residues to react and bind to oxidative damage to the DNA (e.g., pyrimidine).

Due to the stability of the "bundle" conformation and instability of the closed conformation, we propose an additional conformational change to occur. As shown in Fig. 4, the cluster domain rotates to form a "bundle" conformation, separating the catalytic residues and leaving the D239 residue exposed to the solvent. The "bundle" conformation is stabilized by an abundance of helix-helix interactions as shown in the figure: Q94 in the cluster domain forms polar interactions with residues in the hairpin domain. The exact function of the "bundle" conformation is not clear currently, but the presence of this conformation could rationalize findings by Carroll *et al.*[25] Based on the experiments, it was proposed that a large, conformational change occurs at some point after lesion recognition to possibly protect undamaged bases from being erroneously cleaved. It is worth noting that our comparison of wild type and mutated linker systems can draw similar conclusions: the mutated system of NTHL1^{m} has a shortened linker that stabilizes the closed conformation with salt bridges and prevents the transition to the "bundle" conformation. Experiments also reported that conformation changes following lesion recognition were not observed for NTHL1^{m} but were observed for NTHL1. Therefore, further studies on NTHL1 that unravel reactions following lesion recognition should be beneficial for better understanding the functions and dynamics of NTHL1.

During the various stages of the BER pathway for DNA repair, a better understanding of each process and the proteins involved can lead to better modulation of the pathway and help protect against carcinogenesis[1]. NTHL1 is a glycosylase involved with the recognition and initial processing of DNA damage, but the structure of the protein was recently captured in a very stable open conformation that is novel for its sub-group of helix-hairpin-helix glycosylase proteins. The functionalities of NTHL1 during BER pathway rely on the conformational changes of the protein and associated molecular interactions. Unbiased MD simulations revealed that the open conformation is stable, indicating that traditional simulations of a long enough duration to observe transitions between the open and closed states would be computationally unrealistic. In this paper, we design and implement the machine-learning based LINES method to identify a reaction coordinate to accelerate conformational sampling of NTHL1. The predicted reaction coordinate was proven to improve sampling of the open-closed conformation pathway. Also, we were able to identify an unexpected, intriguing but stable "bundle" conformation. This bundle conformation was characterized by a removal of the linker region from the DNA binding cleft, such that the cluster domain rotates and collapses into the binding cleft. A comparison of the wild type and mutant closed NTHL1 conformations revealed that the closed mutant conformation was stabilized through the formation of salt bridges between the linker region and the cluster/hairpin NTHL1 domains. The shortening of the linker due to the mutation created a densely packed cluster of positively charged residues that could close the DNA binding cleft. Therefore, a shortening between the positively charged residues could be a factor to search for in the genes encoding NTHL1 when screening for genetic abnormalities that increase the mutagenesis and carcinogenesis in cells. Additional work remains to be done for discovering the functions of the bundle conformation, but one potential function could be to protect undamaged DNA from being erroneously cleaved.

In this work, the learned reaction coordinate was composed of a linear combination of distances between pseudo-randomly chosen residues in or around the linker region. The reaction coordinate predicted by LINES improved the rate of sampling of conformations, but the reaction coordinate and sampling could be further improved by refining the reaction coordinate to isolate specific strong interactions that correspond to the reaction pathways. This could be accomplished by identifying residues associated with large coefficients in the reaction coordinate definition and then adding new molecular coordinates incorporating residues in the vicinity of the large-coefficient residues. The refinement process could be performed iteratively through a series of reaction coordinate predictions via LINES, isolating large-coefficient residues, and expanding the list of molecular coordinates to include residues close to those important residues. The optimality of the new reaction coordinates can be evaluated through calculations of the cut-based free energy profiles, as suggested from Refs.[40–41] In another direction for improvement, LINES could be combined with collective variable-free enhanced sampling methods to help with sampling reactions orthogonal to the reaction coordinate[42]. For example, Gaussian Accelerated Molecular Dynamics can be used to identify distinct low-energy states of biomolecules[43] and could theoretically help LINES converge to a reaction coordinate more rapidly by expanding the ensemble of sampled conformations during a simulation and is thus a promising direction for future research. While these types of enhanced sampling methods do not require a reaction coordinate, their sampling is typically not as efficient as if an accurate reaction coordinate would be used to bias simulations; a combination of these collective variable-free and reaction coordinate-based techniques could work together synergistically to improve the convergence rate during machine learning and accelerate sampling in MD simulations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM152745. Computational resources were provided by Expanse CPU at SDSC through allocation MCB170012 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1. Lindahl T, Instability and Decay of the Primary Structure of DNA. Nature 1993, 362, 709–715. [PubMed: 8469282]
2. Hakem R, DNA-Damage Repair; the Good, the Bad, and the Ugly. EMBO J. 2008, 27, 589–605. [PubMed: 18285820]
3. Chatterjee N; Walker GC, Mechanisms of DNA Damage, Repair, and Mutagenesis. Environ. Mol. Mutagen 2017, 58, 235–263. [PubMed: 28485537]
4. Krokan HE; Bjørås M, Base Excision Repair. Cold Spring Harb. Perspect. Biol 2013, 5, a012583. [PubMed: 23545420]
5. Hegde ML; Hazra TK; Mitra S, Early Steps in the DNA Base Excision/Single-Strand Interruption Repair Pathway in Mammalian Cells. Cell Res. 2008, 18, 27–47. [PubMed: 18166975]
6. Seeberg E; Eide L; Bjørås M, The Base Excision Repair Pathway. Trends Biochem. Sci 1995, 20, 391–397. [PubMed: 8533150]
7. Wallace SS, Base Excision Repair: A Critical Player in Many Games. DNA Repair 2014, 19, 14–26. [PubMed: 24780558]

8. Lee TH; Kang TH, DNA Oxidation and Excision Repair Pathways. Int. J. Mol. Sci 2019, 20, 6092. [PubMed: 31816862]

9. Kim YJ; Wilson DM 3rd, Overview of Base Excision Repair Biochemistry. Curr. Mol. Pharmacol 2012, 5, 3–13. [PubMed: 22122461]

10. Wilson SH; Kunkel TA, Passing the Baton in Base Excision Repair. Nat. Struct. Biol 2000, 7, 176–178. [PubMed: 10700268]

11. Wallace SS; Murphy DL; Sweasy JB, Base Excision Repair and Cancer. Cancer Lett. 2012, 327, 73–89. [PubMed: 22252118]

12. Dianov GL; Hübscher U, Mammalian Base Excision Repair: The Forgotten Archangel. Nucleic Acids Res. 2013, 41, 3483–3490. [PubMed: 23408852]

13. Wilson DM; Bohr VA, The Mechanics of Base Excision Repair, and Its Relationship to Aging and Disease. DNA Repair 2007, 6, 544–559. [PubMed: 17112792]

14. Jacobs AL; Schär P, DNA Glycosylases: In DNA Repair and Beyond. Chromosoma 2012, 121, 1–20. [PubMed: 22048164]

15. Krokan HE; Standal R; Slupphaug G, DNA Glycosylases in the Base Excision Repair of DNA. Biochem. J 1997, 325 1–16. [PubMed: 9224623]

16. Das L; Quintana VG; Sweasy JB, NTHL1 in Genomic Integrity, Aging and Cancer. DNA Repair 2020, 93, 102920. [PubMed: 33087284]

17. Weren RD; Ligtenberg MJ; Geurts van Kessel A; De Voer RM; Hoogerbrugge N; Kuiper RP, NTHL1 and MUTYH Polyposis Syndromes: Two Sides of the Same Coin? J. Pathol 2018, 244, 135–142. [PubMed: 29105096]

18. Grolleman JE; de Voer RM; Elsayed FA; Nielsen M; Weren RDA; Palles C; Ligtenberg MJL; Vos JR; ten Broeke SW; de Miranda NFCC, et al. , Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-Tumor Phenotype. Cancer Cell 2019, 35, 256–266.e5. [PubMed: 30753826]

19. Weren RDA; Ligtenberg MJL; Kets CM; de Voer RM; Verwiel ETP; Spruijt L; van Zelst-Stams WAG; Jongmans MC; Gilissen C; Hehir-Kwa JY, et al. , A Germline Homozygous Mutation in the Base-Excision Repair Gene NTHL1 Causes Adenomatous Polyposis and Colorectal Cancer. Nat. Genet 2015, 47, 668–671. [PubMed: 25938944]

20. Shinmura K; Kato H; Kawanishi Y; Goto M; Tao H; Yoshimura K; Nakamura S; Misawa K; Sugimura H, Defective Repair Capacity of Variant Proteins of the DNA Glycosylase NTHL1 for 5-Hydroxyuracil, an Oxidation Product of Cytosine. Free Radic. Biol. Med 2019, 131, 264–273. [PubMed: 30552997]

21. Galick HA; Kathe S; Liu M; Robey-Bond S; Kidane D; Wallace SS; Sweasy JB, Germ-Line Variant of Human NTHL1 DNA Glycosylase Induces Genomic Instability and Cellular Transformation. Proc. Natl. Acad. Sci. U. S. A 2013, 110, 14314–14319. [PubMed: 23940330]

22. Limpose KL; Trego KS; Li Z; Leung SW; Sarker AH; Shah JA; Ramalingam SS; Werner EM; Dynan WS; Cooper PK, et al. , Overexpression of the Base Excision Repair NTHL1 Glycosylase Causes Genomic Instability and Early Cellular Hallmarks of Cancer. Nucleic Acids Res. 2018, 46, 4515–4532. [PubMed: 29522130]

23. Yang N; Chaudhry MA; Wallace SS, Base Excision Repair by hNTH1 and hOGG1: A Two Edged Sword in the Processing of DNA Damage in Γ-Irradiated Human Cells. DNA Repair 2006, 5, 43–51. [PubMed: 16111924]

24. Guay D; Garand C; Reddy S; Schmutte C; Lebel M, The Human Endonuclease III Enzyme Is a Relevant Target to Potentiate Cisplatin Cytotoxicity in Y-Box-Binding Protein-1 Overexpressing Tumor Cells. Cancer Sci 2008, 99, 762–769. [PubMed: 18307537]

25. Carroll BL; Zahn KE; Hanley JP; Wallace SS; Dragon JA; Doublié S, Caught in Motion: Human NTHL1 Undergoes Interdomain Rearrangement Necessary for Catalysis. Nucleic Acids Res. 2021, 49, 13165–13178. [PubMed: 34871433]

26. Kuo CF; McRee DE; Fisher CL; O'Handley SF; Cunningham RP; Tainer JA, Atomic Structure of the DNA Repair [4Fe-4S] Enzyme Endonuclease III. Science 1992, 258, 434–40. [PubMed: 1411536]

27. Fromme JC; Verdine GL, Structure of a Trapped Endonuclease III-DNA Covalent Intermediate. EMBO J. 2003, 22, 3461–3471. [PubMed: 12840008]

28. Torrie GM; Valleau JP, Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. J. Comput. Phys 1977, 23, 187–199.

29. Laio A; Parrinello M, Escaping Free-Energy Minima. Proc. Natl. Acad. Sci. U. S. A 2002, 99, 12562. [PubMed: 12271136]

30. Bonati L; Piccini G; Parrinello M, Deep Learning the Slow Modes for Rare Events Sampling. Proc. Natl. Acad. Sci. U. S. A 2021, 118, e2113533118. [PubMed: 34706940]

31. Ribeiro JML; Bravo P; Wang Y; Tiwary P, Reweighted Autoencoded Variational Bayes for Enhanced Sampling (RAVE). J. Chem. Phys 2018, 149, 072301. [PubMed: 30134694]

32. Ribeiro JML; Tiwary P, Toward Achieving Efficient and Accurate Ligand-Protein Unbinding with Deep Learning and Molecular Dynamics through RAVE. J. Chem. Theory Comput 2019, 15, 708–719. [PubMed: 30525598]

33. Mendels D; Piccini G; Brotzakis ZF; Yang YI; Parrinello M, Folding a Small Protein Using Harmonic Linear Discriminant Analysis. J. Chem. Phys 2018, 149, 194113. [PubMed: 30466286]

34. Odstrcil RE; Dutta P; Liu J, LINES: Log-Probability Estimation Via Invertible Neural Networks for Enhanced Sampling. J. Chem. Theory Comput 2022, 18, 6297–6309. [PubMed: 36099438]

35. Odstrcil RE; Dutta P; Liu J, Prediction of the Peptide–TIM3 Binding Site in Inhibiting TIM3–Galectin 9 Binding Pathways. J. Chem. Theory Comput 2023, 19, 6500–6509. [PubMed: 37649156]

36. Dinh L; Sohl-Dickstein J; Bengio S, Density Estimation Using Real NVP. ArXiv 2017, abs/1605.08803.

37. Müller T; Brian M; Rousselle F; Gross M; Novák J, Neural Importance Sampling. ACM Trans. Graph 2019, 38, 145.

38. Patil K; Wang Y; Chen Z; Suresh K; Radhakrishnan R, Activating Mutations Drive Human MEK1 Kinase Using a Gear-Shifting Mechanism. Biochem. J 2023, 480, 1733–1751. [PubMed: 37869794]

39. Bonomi M; Bussi G; Camilloni C; Tribello GA; Banáš P; Barducci A; Bernetti M; Bolhuis PG; Bottaro S; Branduardi D, et al. , Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. Nat. Methods 2019, 16, 670–673. [PubMed: 31363226]

40. Banushkina PV; Krivov SV, Optimal Reaction Coordinates. WIREs Comput. Mol. Sci 2016, 6, 748–763.

41. Krivov SV, Protein Folding Free Energy Landscape Along the Committor - the Optimal Folding Coordinate. J. Chem. Theory Comput 2018, 14, 3418–3427. [PubMed: 29791148]

42. Wang J; Bhattarai A; Do HN; Miao Y, Challenges and Frontiers of Computational Modelling of Biomolecular Recognition. QRB Discovery 2022, 3, e13. [PubMed: 37377636]

43. Miao Y; Feher VA; McCammon JA, Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. J. Chem. Theory Comput 2015, 11, 3584–3595. [PubMed: 26300708]

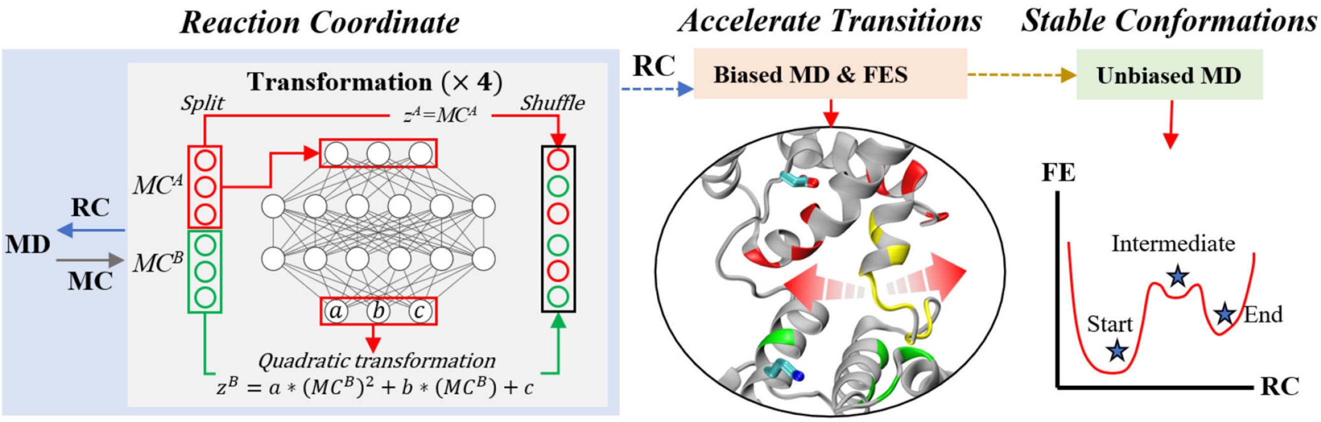


Fig. 1: Algorithmic flow.

In the reaction coordinate discovery phase, cycles of biased MD simulations and machine learning are performed to iteratively improve reaction coordinate (RC) predictions. The MD simulations save molecular coordinates (MCs). During RC discovery, the RC data distribution is transformed with a quadratic transformation and a reshuffling of the data dimensions. The transformation is a piecewise-quadratic function of the form $y(x; a, b, c) = ax^2 + bx + c$, where the parameters a, b and c are vector outputs from the neural network. The outputs are vectorized so that a different transformation can be applied to partition B (MC^B) based on the value of MC^B – hence the piecewise component of the transformation. After the reaction coordinate has converged, a long, biased MD simulation is run to estimate the converged free energy surface (FES). The reaction coordinate identifies slow reaction pathways and accelerates the pathway by amplifying the protein motions through the biasing potential. Following the FES convergence, distinct protein conformations are identified, and the stability of each conformation is evaluated with unbiased MD simulations.

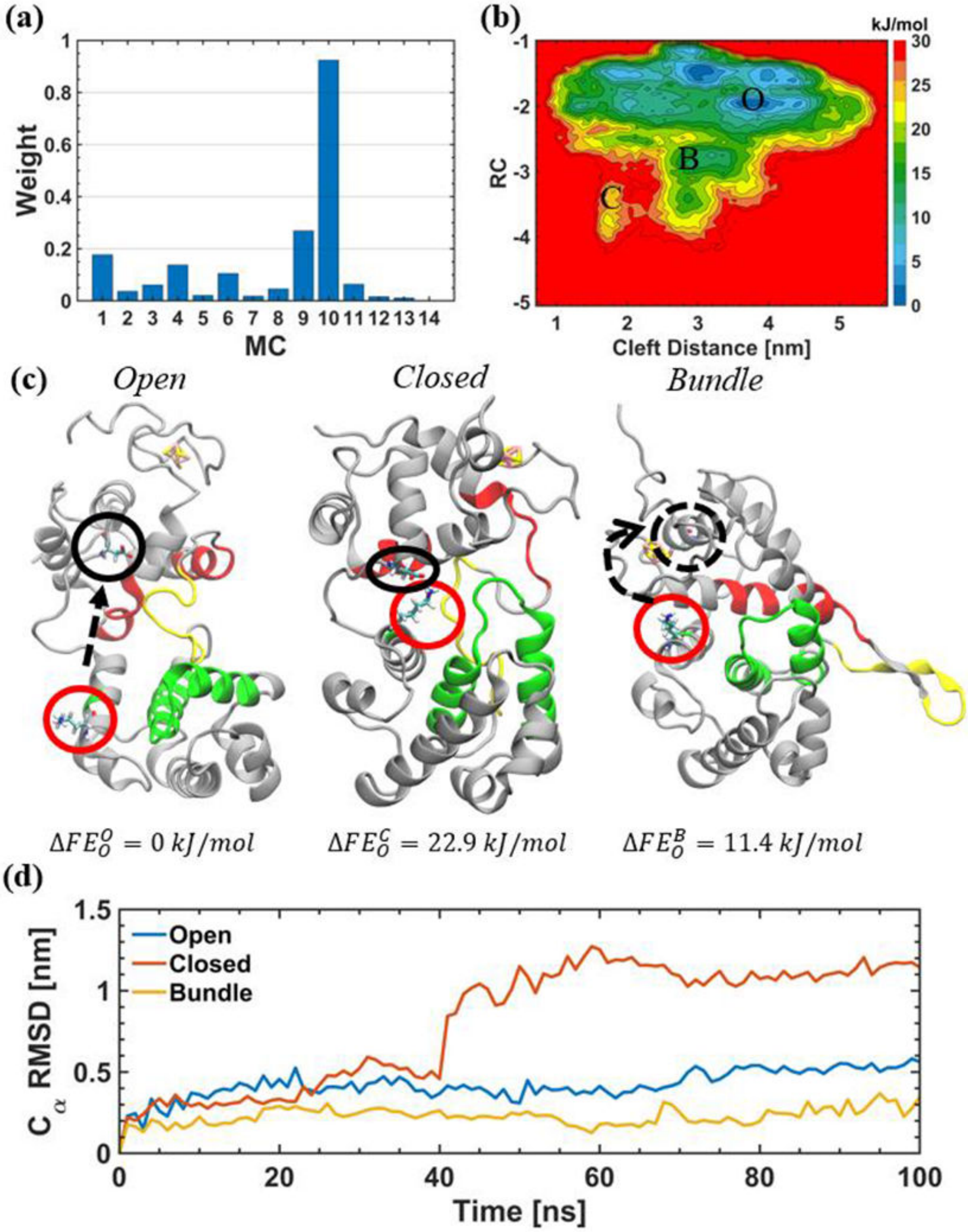


Fig. 2: NTHL1 conformations.

(a) The magnitude (weight) of the molecular coordinate coefficients in the reaction coordinate predicted from the 7th iteration of LINES. **(b)** The converged free energy surface (FES) from the 600 ns biased MD simulation. Three distinct conformations are identified: open (O), closed (C), and a bundle (B) state. **(c)** The molecular structures of the three distinct conformations. The circles mark the locations of the catalytic residues K220 (red) and D239 (black). Compared to the closed conformation, the open conformation has the residues located farther apart with the DNA-binding cleft open, while the bundle conformation has the cluster domain rotated and several helices collapse into the binding cleft to leave D239 expose to the solvent. The free energy differences between the three energy basins and the open conformation are calculated and shown below each conformation. **(d)** The stability of each distinct conformation was evaluated with the root mean square deviation (RMSD) of the C_{α} backbone atoms from 100 ns unbiased MD simulations. The open and bundle conformations are stable but the closed conformation becomes unstable after 40 ns.

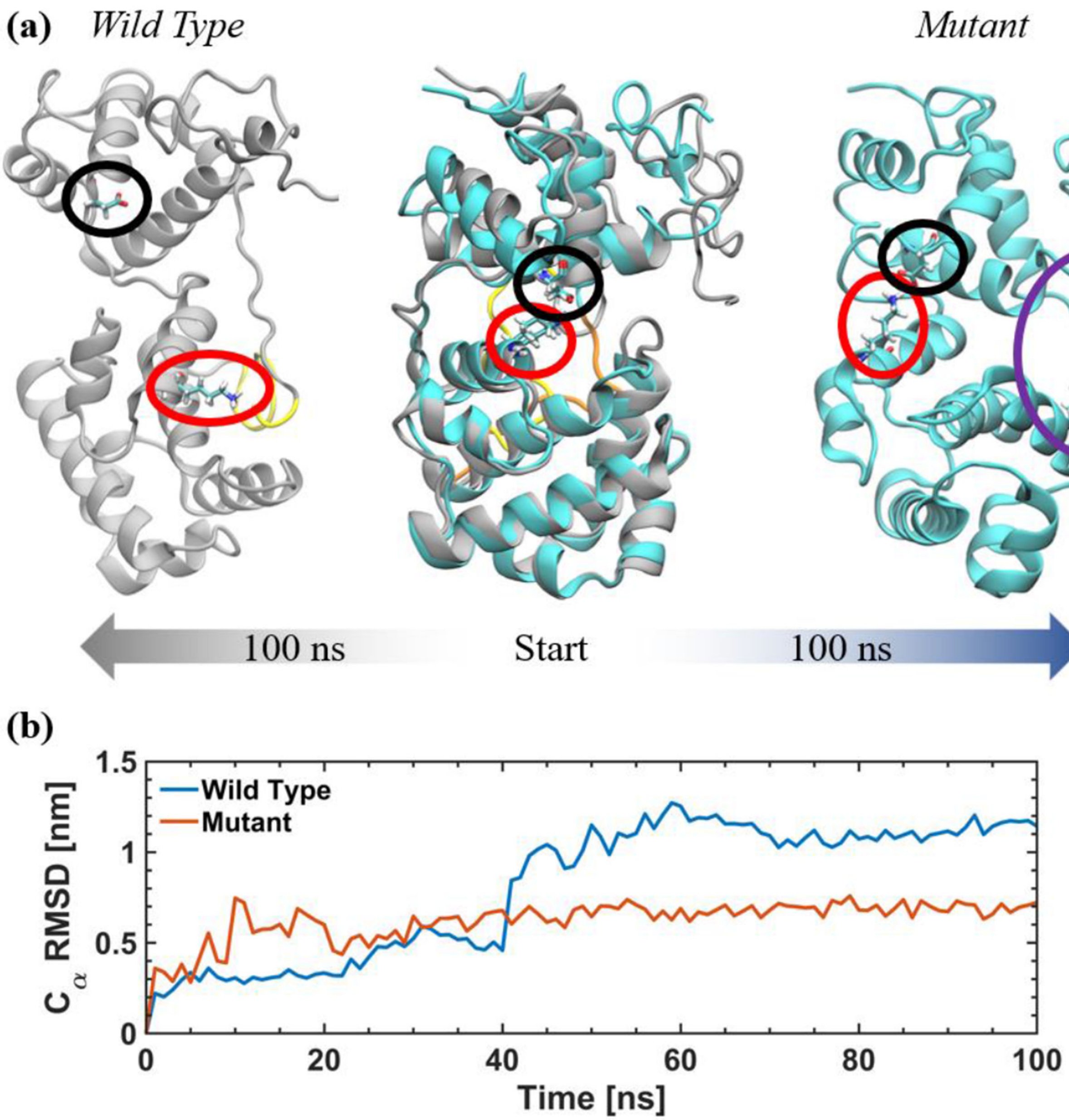


Fig. 3: Comparison of closed conformation stabilities between wild type and mutant NTHL1.

(a) Starting from very similar initial conformations, both the wild type (gray) and mutant (cyan) NTHL1 were subjected to 100 ns of unbiased MD simulations. The wild type linker region is shown in yellow and the mutant linker region in orange. The catalytic residues K220 and D239 are located by the red and black circles, respectively. The mutant conformation is stabilized by salt bridges in the purple circle. **(b)** The stability of the conformations is evaluated using the root mean square deviation (RMSD) of the C_α atoms during the 100 ns simulations. The mutant conformation stabilizes once the salt bridges between the linker and the protein domains form while the wild type protein fails to stabilize.

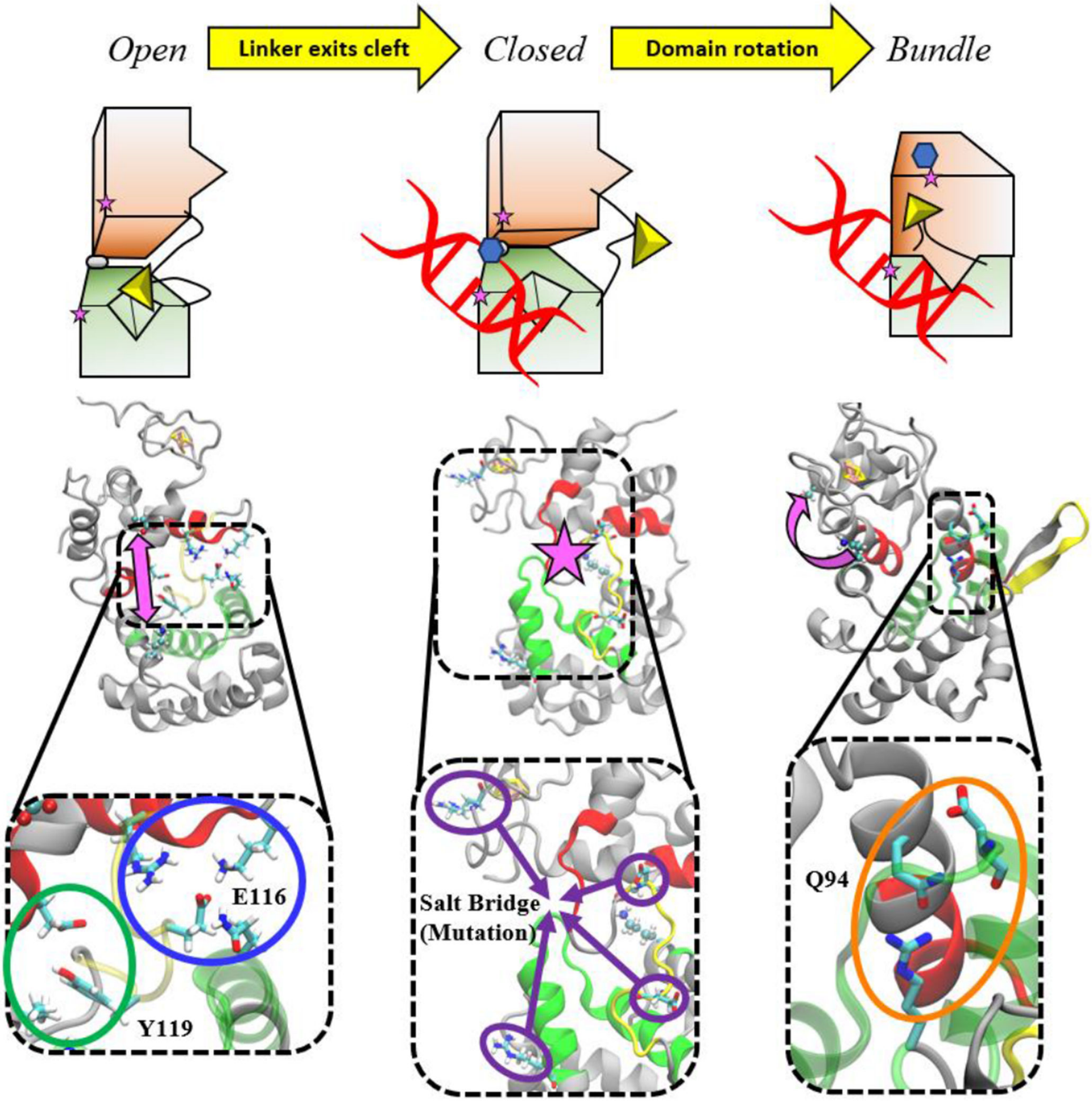


Fig. 4: NTHL1 conformations and interactions.

The general features of the conformation changes from the open to the closed, and to the bundle states. Increasing levels of detail are shown from the top to the bottom in the form of a proposed schematic representation (top), entire protein structure (middle), and the stabilizing interactions (bottom) for the open (left), closed (center), and bundle (right) conformations. For the schematic representations, the hairpin domain (green), cluster domain (orange), and linker region (yellow) are shown. The catalytic residues are represented by the pink stars in each domain. Damaged DNA is represented by the red helix, with the oxidative damage (i.e., pyrimidine) indicated by the blue hexagon. In the open conformation (left), the catalytic residues K220 and D239 are separated (pink arrow); the inset shows how a salt bridge containing E116 (blue) and polar/nonpolar interactions with Y119 (green) stabilize the open conformation. Between the open and closed conformation, the linker region (yellow backbone) exits the DNA binding cleft and the DNA binds to NTHL1. In the closed conformation (middle), the catalytic residues are close together (pink star) and the pyrimidine on the damaged DNA can react with the catalytic residues. Between the closed and bundle conformations, the cluster domain rotates and may halt catalysis. For the system with the mutated linker region, the inset illustrates a salt bridge forming that stabilizes the closed conformation and prevents the transition to the bundle conformation. In the bundle conformation (right), the catalytic residue D239 is exposed to the solvent (pink curved arrow); the inset shows how Q94 polar interactions (orange) are one example of a stabilizing interaction.