PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

PCCP

PAPER

A Computational Investigation on the Substrate Preference of Ten-Eleven-Translocation 2 (TET2)

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

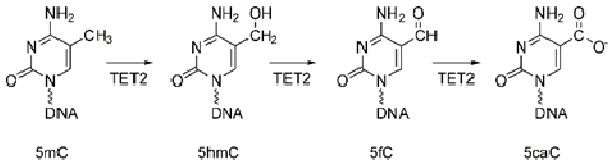
Junyan Lu[a#], Lulu Hu[b,c#], Jingdong Cheng[b#], Dong Fang[d], Chen Wang[a], Kunqian Yu[a], Hualiang Jiang[a], Qiang Cui[d], Yanhui Xu[b,c*] and Cheng Luo[a*]

TET proteins iteratively convert 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) in an Fe(II)/α-ketoglutarate-dependent manner. Our previous biochemical studies revealed TET proteins are more active on 5mC than 5hmC and 5fC. However, the source of the substrate preference of TET proteins still largely remain elusive. Here, we investigated the substrate binding and catalytic mechanisms of oxidation reactions mediated by TET2 on different substrates through computational approaches. In accordance with previous experimental reports, our computational results suggest the TET2 can bind to different substrates with comparable binding affinities and the hydrogen abstraction step in the catalytic cycle acts as the rate-limiting step. Further structure characterizations of the intermediate structures revealed that the 5-substitution groups on 5hmC and 5fC adopt an unfavorable orientation for hydrogen abstraction, which leads to higher energy barrier for 5hmC and 5fC (compared to 5mC) and thus lower catalytic efficiency. In summary, our mechanical insights demonstrate that substrate preference is the intrinsic property of TET proteins and our theretical calculation results can guide further dry-lab or wet-lab studies for the catalytic mechanism of TET proteins as well as other Fe(II)/α-ketoglutarate(KG)-dependent dioxygenases.

1 Introduction

DNA methylation at the 5 position of cytosine (5mC) is an important epigenetic modification, which plays a role in many key biological processes[1, 2]. As the opposite procedure of DNA methylation, DNA demethylation has long been considered as a passive process until the recent identification of a few kinds of enzymes, such as activation-induced cytidine deaminase (AID), ten-eleven translocation (TET) proteins and thymine-DNA glycosylase (TDG), that take part in active DNA demethylation.[3-6] As a newly discovered family of DNA-modifying enzymes, TET proteins catalyse the conversion of cytosine-5 methylation (5mC) to 5-hydroxymethyl-cytosine(5hmC) and recent reports have shown that TET proteins are key enzymes responsible for the presence of 5hmC in mouse ES and iPSCs.[7-9] TET proteins can further oxidize 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) in DNA (Scheme 1).[10] These two oxidation products can then be excised by TDG, which eventually results in the removal of DNA methylation markers through the base excision repair (BER) pathway.[3, 11] A recent study has provided evidence that 5hmC is actually stable in mammalian genome DNA, rather than merely acting as an intermediate product during the active DNA demethylation process.[12] Therefore, 5hmC may have a more direct role in transcription activation and this finding further emphasize the importance of TET proteins in epigenetic regulation.

Scheme 1. The continuous oxidation of 5mC by TET2



An interesting feature of TET2 is that although it can iteratively oxidize 5mC into 5hmC, 5fC and 5caC, its enzymatic activities on different substrates are different. Human TET1/2, mouse Tet2 and Naegleria Tet-like protein were all found to be more active on 5mC-DNA than 5hmC/5fC-DNA[10, 13, 14]. As 5hmC is considered to be a stable epigenetic marker and is significantly more prevalent than 5fC and 5caC, the observed substrate preference of TET proteins may play a key role in the

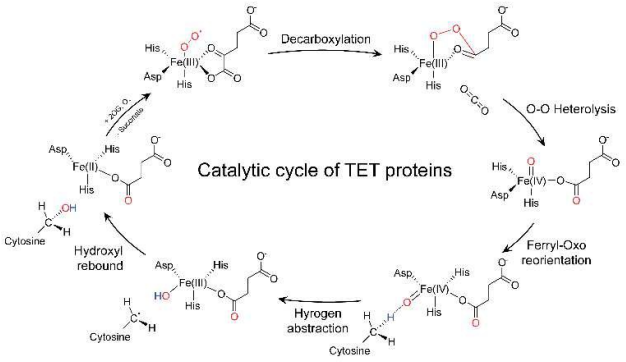
[a] Drug Discovery and Design Centre, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.
[b] Fudan University Shanghai Cancer Centre, Institute of Biomedical Sciences, Shanghai Medical College of Fudan University, Shanghai 200032, China.
[c] State Key Laboratory of Genetic Engineering, Collaborative Innovation Centre of Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200433, China.
[d] Department of Chemistry and Theoretical Chemistry Institute, University of Wisconsin-Madison, 1101 University Ave, Madison, WI 53706, USA.

†Electronic supplementary information (ESI) available: Further details of computational methods of "Alchemical" free energy calculations as well as the energy profile of the alternative pathway of hydrogen abstraction on 5hmC are given. See DOI: 10.1039/x0xx00000x.

#These authors contributed equally. The authors declare no competing financial interests.

maintaining the genome-wide 5hmC abundance and regulating the balance between 5hmC related downstream signals and DNA demethylation. Previous biochemical results suggested the substrate preference of human TET2 did not result from the substrate binding nor the recognition process because TET2 was found to bind different substrate DNAs with very similar binding affinities and crystallography studies also showed the substrate binding pocket of TET2 is well accommodated to 5mC, 5hmC and 5fC.[14, 15] On the other hand, this preference may result from the catalytic process after the initial binding.[14]

Scheme 2. Schematic model for the oxidative reactions catalysed by TET proteins.



All TET proteins possess a C-terminal catalytic domain (CD), which can be categorized into the Fe(II)/α-ketoglutarate(KG)-dependent dioxygenase superfamily. This dioxygenase superfamily also contains several other DNA and RNA base modifying enzymes, including the Escherichia coli AlkB[16] protein and its two human homologues, hABH2 and hABH3[17], the N6-methyladenosine RNA demethylase FTO[18], and the thymine 7-hydroxylase (THase)[19]. The catalytic domains of these enzymes share a similar β-strand "jellyroll" structural fold and contain a non-heme Fe(II) ion, which is ligated by a facial triad of residues consisting of two histidines and one aspartate.[20] They also share a similar catalytic mechanism that involves the activation of dioxygen by forming a high-valent ferryl-oxo (Fe(IV)-oxo) intermediate through conversion of α-KG into succinate (Scheme 2).[20]

Although previous reports suggested the substrate preference of TET2 may result from the catalytic process rather than substrate binding, the structure basis for the substrate preference is still missing, because only the reactant structures for TET2-DNA complexes are currently available and the whole catalytic cycle of TET2 involves multiple reaction steps and intermediate states, which are difficult to capture experimentally. In order to gain further insight into the catalytic cycle of TET proteins as well as to specify the source of substrate preference, we investigated mechanisms of TET2 catalyzed oxidation of 5mC, 5hmC and 5fC using molecular dynamics (MD) simulations and hybrid quantum mechanical/molecular mechanical (QM/MM) approaches. In accordance with previous biochemical experimental results, MD simulation results indicate TET2 binds to different DNA substrates with similar affinities. QM/MM calculations suggest the energy barriers for the third step (hydrogen abstraction) are significantly higher than the first two steps (decarboxylation and O-O heterolysis) and follows the order of 5fC > 5hmC > 5mC. In addition, we found that differences in hydrogen abstraction barrier may result from the different orientations of the substrate bases and different configurations of the iron centre in the Fe(IV)-oxo intermediate states: the abstractable hydrogen on 5mC situates in an suitable position while the abstractable hydrogen on 5hmC or 5fC is restrained and relatively far from the oxo group thus resulting in higher abstraction energy. We also identified some unique features in the catalytic process of TET2, compared with other Fe(II)/α-KG-dependent dioxygenases, during the calculations. These findings provide the atomic details of the substrate preference and further insights into the molecular mechanism of TET2.

2 Methodology and computational details

2.1 Model construction and preparation

The crystal structures of TET2-5mC-DNA (PDB:4NM6), TET2-5hmC-DNA (PDB:5DEU) and TET2-5fC-DNA (PDB:5D9Y) were used as initial coordinates to build the pre-catalytic TET2-5mC, TET2-5hmC and TET2-5fC model. The N-terminal missing region in the crystal structure was built using Build Homology Models and Loop Refinement protocols in Discovery Studio, version 3.0.[21] The missing regions (1464–1481 and the following 15-GS linker) in the crystal structures were too long to be built by homology modeling method. Instead, a 6-residue glycine linker was used to fill up this region to maintain the conformations of its connecting residues and the coordinates of this linker were fixed during MD simulations. As this region was far away from the catalytic pocket of TET2, such replacement would not affect the dynamic features of the catalytic pocket nor the QM/MM calculation results. The NOG molecule in the crystal structures was replaced by α-KG through changing the nitrogen atom to a carbon atom.

2.2 Molecular dynamic simulation setup

The protonated states of the titratable residues under physical conditions were predicted by H++ 3.0[22]. Each of the TET2-DNA complexes was solvated by a cubic water box with its boundary extended 10 Å away from the protein-DNA complex on all side. 40 Na+ and 20 Cl- were added to neutralize the simulations system and to make the salt concentration close to 0.1 mM. AMBER10 force field (AMBER99SB+ parmbsc0)[23, 24] was used for the protein and DNA part. Water molecules were described as the TIP3P model[25]. For the modified cytosines (5mC, 5hmC and 5fC) and α-KG, the general Amber force field (GAFF) was used[26]. The RESP charges of the modified cytosines and small molecules were derived using R.E.D server[27]. The parameters for non-heme Fe(II) developed by Cui et al. were used for the iron in TET2 active site[28]. As suggested by Cui et al., a weak harmonic restraint was applied to the Asp residue (ASP1384) to keep its coordination mode stable. Zinc AMBER

Force Filed (ZAFF) parameters were used for the three zinc binding sites in TET2[29].

All MD simulations were performed using the Gromacs 4.5.5 package with standard periodic boundary conditions[30]. Long range electrostatic interactions were treated using the particle mesh Ewald method[31]. Covalent bonds involving hydrogens were restrained with the LINCS algorithm[32]. A cut-off distance of 10 Å was applied for the Lennard-Jones interactions and short-range electrostatic interactions. After a brief energy minimization using the conjugate gradient algorithm, each system was heated up to 300K gradually in 100 ps using the NVT ensemble, with a position restraint of 2 kcal/(mol·Å2) on all the atoms of the protein-DNA complex. Waters, ions and the protein-DNA complex were separately coupled in a thermal bath using a modified Berendsen thermostat method with a coupling time of 0.1 ps[33]. Another 100 ps MD simulation using the NPT ensemble at 1 bar was performed to equilibrate the density of the system. Berendsen pressure coupling approach with a coupling time of 1 ps was used for the equilibration and production runs[33]. Finally, 100 ns canonical MD simulations with a time step of 2 fs were performed for each TET2-DNA complex model to relax the system. Coordinates were saved every 10 ps throughout the production runs.

2.3 QM/MM calculations

The TET2-substrate configurations used for QM/MM calculations were derived from the average structure of equilibrated MD trajectories (from 10 ns to 100 ns) of three TET2-DNA complex models. To define the averaged conformation, structures of the catalytic center residues (H1382, D1384, H1881, H1386, N1387, H1904, α-KG, iron and substrate bases) were clustered using average-linkage algorithm with a RMSD cut-off value of 0.5 Å. As the catalytic center residues remained stable during 100ns MD simulations of all three TET2-DNA models (Fig. 1), the largest cluster comprised the majority of the conformations in each MD trajectory (99% for TET2-5mC, 95% for TET2-5hmC and 81% for TET2-5fC). Therefore, the cluster center of the largest cluster in each TET2-substrate model was selected to represent the average conformation in MD simulation. The water molecule that coordinates to the Fe atom in each crystal structure was replaced by an O2 molecule. All the QM/MM calculations were performed using a two-layered ONIOM scheme encoded in the Gaussian09 program[34]. For each complex model, the quantum mechanical (QM) part contains the Fe atom, the side-chains of the residues that coordinate to the iron (H1382, D1384 and H1881), the O2 molecule, the α-KG molecule and the substrate base. As residue R1261 was found to form a conserved salt bridge with α-KG in all the three crystal structures, most of its side-chain were also included in the QM region. Link hydrogen atoms[35] were employed to saturate the dangling covalent bonds. The QM region was described in terms of the density functional theory (DFT) with the UB3LYP functional and 6-31G(d,p) basis set. For the hydrogen abstraction step, we additionally performed geometry optimizations and energy calculations with the QM region described in UωB97XD/6-31G(d,p), because the ωB97XD functional includes dispersion corrections, which have been shown to be important in estimating the barrier for the hydrogen abstraction catalyzed by cytochrome P450[36] and Fang et al. suggested ωB97XD was the most suitable functional to describe Fe(IV)-oxo catalytic intermediate[37]. The remainder of the system (MM region) was treated using the AMBER Parm99 force field.

As the whole TET2 protein was too large for the frequency analysis, we truncated the system to an area containing all amino acids with at least one atom within a 15 Å radius of the iron centre. To prevent the system from undergoing unnatural changes during the geometry optimizations, only the waters and residues within 10 Å of the iron centre were allowed to move during QM/MM optimization. The electrostatic interactions between the QM and MM regions were calculated by an electronic embedding scheme implemented in Gaussian09 program. The partial charges of the MM region were incorporated into QM Hamiltonian, which provides a better description of the electrostatic interaction between the QM and MM regions and allows the QM wave function to be polarized. All calculations were performed assuming quintet electronic states since previous spectroscopic studies found mononuclear iron enzymes were in high-spin (S=2) electronic ground state configurations[38] and theoretical studies suggested quintet Fe(IV)-oxo species was the most reactive toward C-H bond activation[39, 40].

After each reactant model was fully optimized by QM/MM method, relaxed potential energy surface scans were performed to obtain the initial approximate reaction pathways. For each reaction step, the stable structures obtained by the PES were fully optimized and the highest point was taken as the initial guess for the transition state (TS) optimization using QST3. Frequency analyses were performed for all critical structures to confirm no imaginary vibrational frequencies for reactant, product and intermediate structures, and only one imaginary frequency for the TS structures. Intrinsic reaction coordinate (IRC) calculations were also performed to insure that the transition states connect correctly the reactants and the products. Zero point energy (ZPE) corrections and thermal dynamic corrections (at 298.15K) were obtained by performing harmonic vibrational frequency calculations.

3 Results and discussions

3.1 Modelling and molecular dynamic simulations of the TET2-substrates complexes

To examine the mechanism of the substrate preference of TET2 for different modified cytosines, we firstly used molecular dynamics (MD) simulations to model the binding of TET2 with 5mC, 5hmC and 5fC containing DNAs. The initial coordinates of TET2 binding to 5mC, 5hmC and 5fC containing DNAs were directly derived from crystal structures (see Materials and Methods). These systems were labelled as TET2-5mC, TET2-5hmC, and TET2-5fC. We then carried out 100 ns MD simulations for each complex model and analysed the structural features of TET2 protein, substrates and their

dynamic interactions. During the simulations, the overall complex structures as well as the active site structures in the three models were stable and remained closed to the crystal structures, according to the time evolution of backbone root mean square deviation (RMSD) (Fig. 1a and 1b). The root mean square fluctuation (RMSF) calculated from MD trajectory of each complex model resembled the experimental B-factors in each TET2-substrate crystal structures (Fig. S1), which indicates the reliability of our modelled structures. The RMSD of the modified cytosines in the TET2 active site shows all the three bases could bind stably to the active site (Fig. 1c). The key hydrogen bonding interactions in the crystal structures of TET2 with 5mC, 5hmC and 5fC were well maintained during the MD simulations (Fig. 1d-f). Although previous fluorescence polarization (FP) and surface plasmon resonance (SPR) assays showed TET2 could bind to 5mC/5hmC/5fC-containing DNA with the similar overall binding affinities[14], to explore the possibility that the different catalytic efficiency might result from the subtle binding affinity differences for modified bases, which could not be captured by FP and SPR experiment, we carried out "alchemical" molecular dynamic simulations to evaluate relative free energies for TET2 to flip a modified base from DNA double helix into its catalytic pocket. A thermodynamic cycle was designed to calculate relative binding free energies (Fig. S2) and the detailed process was described in the Materials and Methods section. As a result, the simulations indicated a slightly stronger binding for 5mC over 5hmC (0.39kcal/mol) and 5fC (1.85kcal/mol) (Table S1). The dynamic properties of the three complex models and free energy calculation results all suggests that the TET2 catalytic pocket is well suited for binding of either 5mC, 5hmC or 5fC, which is in consist with the experimental observations that TET2 binds to 5mC, 5hmC or 5fC containing DNA with overall similar affinities. The subtle differences in binding free energies may result in the slightly lower Km value for 5mC hydroxylation observed in the enzymatic kinetic assays[14] but is unlikely to be the main cause of the substrate preference of TET2.

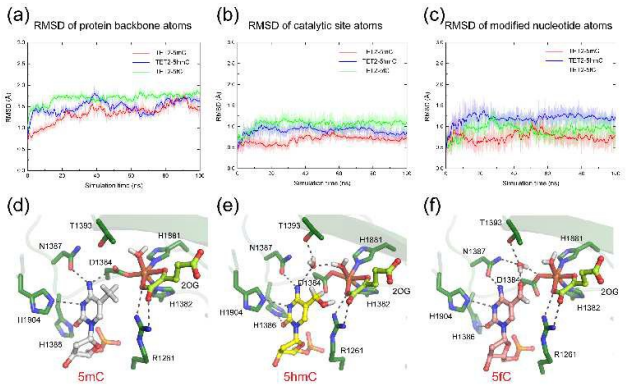


Fig. 1. Conformational stability of TET2-DNA complexes during MD simulations. The time evolution of RMSD values of protein backbone atoms (a), catalytic site (H1382, D1384, H1881, R1261, Fe and α-KG) atoms (b) and modified nucleotide (5mC, 5hmC and 5fC) atoms (c) in TET2-5mC, TET2-5hmC and TET2-5fC simulations. The low average RMSD value (less than 2 Å) indicated that both protein and DNA remained stable in each model. (d, e, f) The equilibrated active site conformation of each complex model. The hydrogen bonds were shown as dashed lines.

3.2 Optimization of the reactant structures.

As the above computational results and previous experimental data suggest the catalysis rather than the substrate binding step may be responsible for the substrate preference of TET2, we then used density function theory (DFT) methods to investigate the catalytic mechanism of TET2 mediated oxidation of different substrate bases. DFT methods have been successfully used previously to characterize the catalytic mechanisms of several Fe(II)/α-KG-dependent dioxygenases, such as alkB protein[37, 41], Clavaminic acid synthase (CAS)[42] and taurine/α-ketoglutarate dioxygenase (TauD)[43]. A hybrid quantum mechanical/molecular mechanical (QM/MM) approach, which has been widely used in modelling enzymatic reactions[44, 45], was used to consider the influence of the protein environment on the reaction centre.

Because the active site structures for different TET2-DNA complexes remained stable during MD simulations, we used the averaged structure of each MD trajectory (from 10 to 100ns) as the initial structure for subsequent QM/MM calculations. All calculations were performed assuming quintet electronic states because we are focused on the comparison of the reaction pathways for different substrates and previous spectroscopic and computational studies found mononuclear iron enzymes were in high-spin (S=2) electronic ground state, which also possesses the highest reactivity[38, 40, 41]. As shown in Fig. 2, the QM/MM optimized geometries of the active sites of TET2-5mC/5hmC/5fC complex are almost identical to the crystal structures, except that the Fe(II)-coordinating water molecule was replaced by an O2 molecule in each complex. In the crystal structures of TET2 in complex with 5hmC and 5fC, a water molecule was observed between the cytosine base and D1384, and this water molecular also remained stable during MD simulation of each complex (Fig. 1). Therefore, this water molecule observed in the crystal structures of TET2 in complex with 5hmC and 5fC was included in the computational model and positioned in the QM region. While in the crystal structure of TET2 in complex with 5mC, D1384 adopts a different orientation compared with D1384 in TET2 in complex with 5hmC and 5fC, and no water molecule was observed between D1384 and 5mC. The lack of water in the active site of TET2-5mC complex is qualitatively consistent with the more hydrophobic nature of the methyl group, compared with the hydroxymethyl group and formyl group on 5hmC and 5fC.

Fig. 2. The MD relaxed and QM/MM optimized reactant structures. (a) TET2-5mC-DNA, (b) TET2-5hmC-DNA, and (c) TET2-5fC-DNA complex models. The atoms in QM region are shown as sticks. The active site residues in MM region are shown as lines. The hydrogen bonding networks within active sites are shown as dashes.

3.3 Dioxygen activation and decarboxylation.

Based on the optimized reactant structures, the potential energy surfaces of the whole reaction pathway of TET2 mediated oxidation on different substrates were then obtained and compared. According to the conserved reaction mechanism of Fe(II)/α-KG-dependent dioxygenases46, the first step is the attack of the Fe-bound dioxygen to the carbonyl carbon (C2) of pyruvate and the formation of an Fe(II)-OO-R (peroxy bridge) intermediate (Scheme 2). We found this process also occurred concomitantly with decarboxylation of the pyruvate. The potential energy surfaces and the geometries of key structures along the reaction pathways are shown in Fig. 3. According to the energy profiles, the reaction barriers for the three substrates are very close. The potential energy barriers for 5mC, 5hmC and 5fC are 7.6 kcal/mol, 9.3 kcal/mol, 7.8 kcal/mol respectively and this reaction step is highly exoergic for all the substrates. The optimized transition state (TS) structures are also very similar. This is conceivable since the substrate bases were actually not involved in the dioxygen activation step.

However, structural differences were found in the first reaction intermediates (I1) after the decarboxylation step. Initially, the dioxygen molecule was bound to the axial position to the equatorial plane of the iron centre (trans to H1382) in each complex model. In concomitant with the decarboxylation, which freed up the equatorial position, the peroxy bridge reoriented to the equatorial position, which lies trans to H1881, in the TET2-5mC and TET2-5fC model. This reorientation process shortened the distance between the reactive oxygen and the 5-substitution groups on substrate bases (Fig. 3). A previous computational study of the AlkB reaction pathway suggested this reorientation occurred after the breakage of the peroxy bridge and the formation of the Fe(IV)-oxo intermediate, and a reorientation barrier of about 11 kcal/mol was reported[41]. By observing the optimized

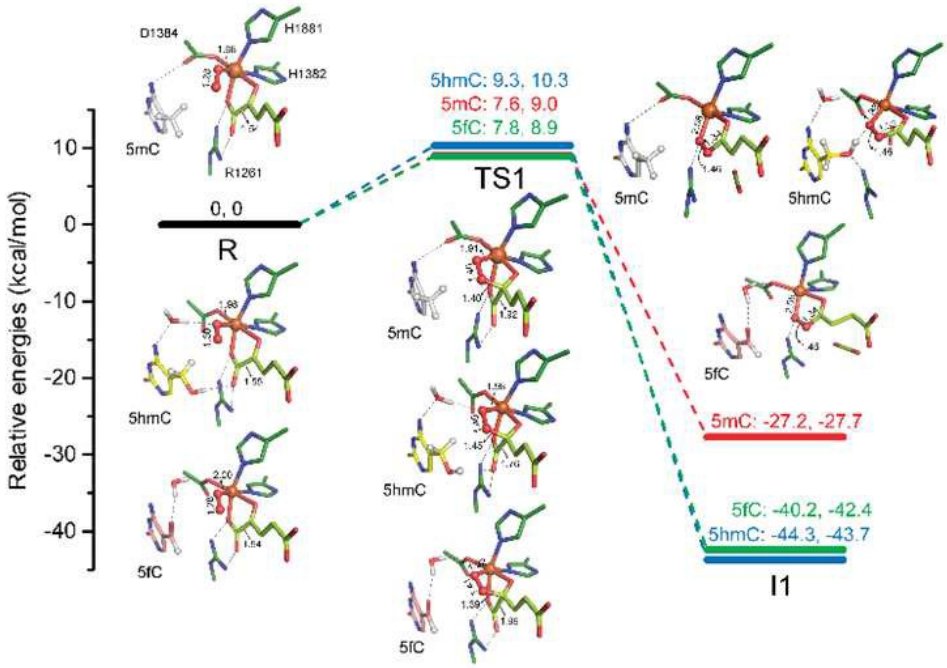


Fig. 3. Pathways for the decarboxylation reactions of 5mC, 5hmC and 5fC. Relative potential energies (ΔE) and free energies (ΔG) for reactants (R), the first transition states (TS1) and the first intermediate states (I1) are given in kcal/mol (represented as ΔE, ΔG). The geometry parameters (with distances in Å and angles in degree) for the bonds involved in the reactions are labelled. Hydrogen-bonding interactions are shown as dashes.

structures of the reactants and the I1, we suggested the spontaneous reorientation process observed in TET2 may be mainly caused by the electrostatic attraction between the peroxy bridge and R1286. An arginine residue also exists in the catalytic site of AlkB protein, however, this residue lies near the axial position and therefore hinders the reorientation of the peroxy bridge or the Fe(IV)-oxo species. On the other hand, the hydroxyl group on 5hmC forms hydrogen bonds simultaneously with R1261 and the peroxy bridge, which prevents the completely reorientation of the peroxy bridge in TET2-5hmC model (Fig. 3). Instead, the iron centre in TET2-5hmC model adopted a trigonal bipyramidal coordination geometry, rather than the square pyramidal geometry observed in the I1 of TET2-5mC and TET2-5fC model.

In addition, we found the energy of the optimized I1 structure of TET2-5mC is significantly higher than that of the TET2-5hmC and TET2-5fC. As the substrate bases were actually not involved in the decarboxylation step, the energies of this intermediate states of different substrates should be similar. We noticed the positions of the generated CO2 molecule were different in the optimized structures of the intermediate states of TET2-substrate complexes, which may result in the energy differences. As the CO2 molecule was away from the iron centre and its exact position was hard to predict using the current QM/MM method, we eliminated the CO2 molecules from our computational models in the calculations of the following reaction steps.

3.4 O-O heterolysis reaction

The reaction step following the decarboxylation is O-O heterolysis, which involves the break of the peroxy bridge and formation of the highly reactive Fe(IV)-oxo species. Because the substrate bases do not directly take part in this step as the previous step, the reaction barriers and TS structures were also found to be similar among 5mC, 5hmC and 5fC (Fig. 4). As the calculated reaction barriers were all very low (1.2~1.9 kcal/mol) and the reactions were highly exoergic, the O-O heterolysis should happen very quickly and is also irreversible. It worth noticing that, after the O-O heterolysis, the oxo group in TET2-5mC and TET2-5fC complex still lied trans to H1881, and therefore the oxo groups are close to the 5-substitute groups on the substrate bases. However, in the TET2-5hmC complex, the coordination geometry of the iron centre changed to a square pyramidal geometry similar to TET2-5mC and TET2-5fC complex, except for that the oxo group now lies trans to H1382 and formed a hydrogen bond to the water molecule in the catalysis pocket. Therefore, the oxo group is relatively far away from the 5-substitution group on 5hmC than 5mC/5fC. In all, because of the similar and low energy barrier for all the three substrates, the first two steps (decarboxylation and O-O heterolysis) are unlikely the rate limiting steps of TET2 catalysed hydrolysis or responsible for the substrate preference. These calculation results are also in accordance with the experimental observations that the time and rates of the formation of the Fe(IV)-oxo group did not show significant differences during the catalysis of the three substrates[14].

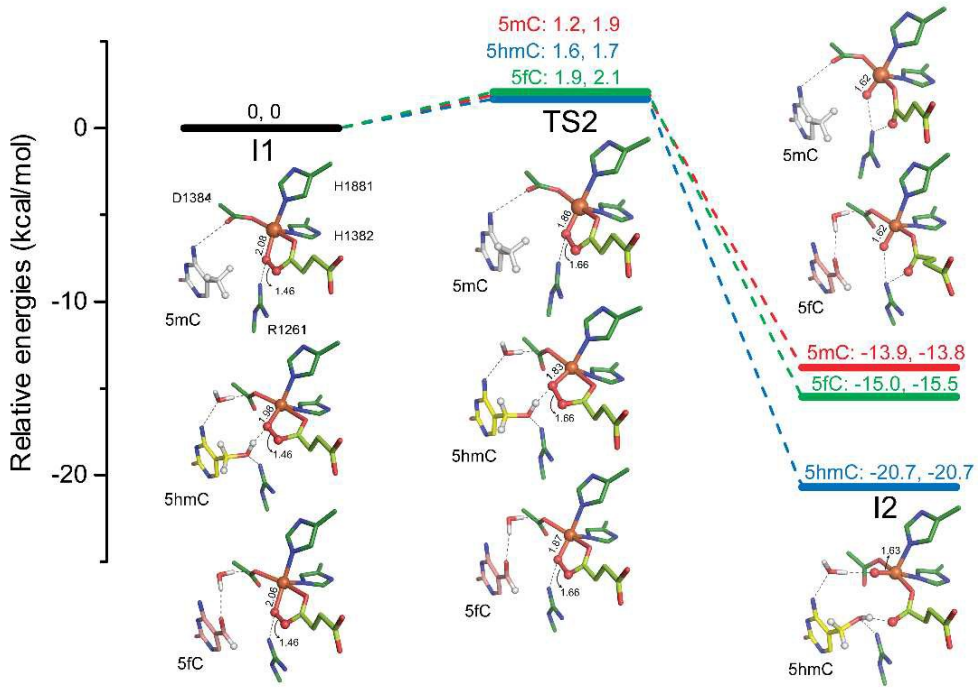


Fig. 4. Pathways for the O-O heterolysis reactions of 5mC, 5hmC and 5fC. Relative potential energies (ΔE) and free energies (ΔG) for the first intermediate states (I1), the second transition states (TS2) and the second intermediate states (I2) are given in kcal/mol (represented as ΔE, ΔG). The geometry parameters (with distances in Å and angles in degree) for the bonds involved in the reactions are labelled. Hydrogen-bonding interactions are shown as dashes.

3.5 Hydrogen abstraction

The Fe(IV)-oxo intermediate (I2), which is a critical intermediate state in the reaction pathway of Fe(II)/α-KG-dependent dioxygenases, was generated after O-O heterolysis. The high-valent Fe(IV)-oxo species is highly reactive and can therefore activate the C-H bond of substrate and mediate the hydrogen abstraction reaction. The hydrogen abstraction step is also considered as the rate-limiting step in the catalysis of an Fe(II)/alpha-ketoglutarate acid dependent dioxygenase, AlkB[41]. In the QM/MM optimized I2 structures for the TET2-5mC and TET2-5hmC complexes, the oxo group lies trans to H1881 and forms hydrogen bond with R1261. According to the QM/MM calculation results, the potential and free energy barriers of the hydrogen abstraction step for 5fC is significantly higher than 5mC (Fig. 5). Previous experimental and computational studies indicated that the energy barrier for hydrogen abstraction reaction is tightly correlated with the homolytic C–H bond dissociation energy (BDE) of substrate[46, 47]. In our previous study, the C-H BDE of the 5-substitution group on 5fC was estimated to be only a slightly higher (~1 kcal/mol) than that on 5mC[14]. However, the hydrogen abstraction energy barrier for 5fC is about 10 kcal/mol higher than that for 5mC. By comparing the optimized I2 structures of 5mC and 5fC, we suggested the significantly elevated energy barrier for 5fC hydrogen abstraction resulted from the unsuitable position of the hydrogen on 5-formyl group. As the methyl group on 5mC is free to rotate, one abstractable hydrogen on the methyl group was directly pointed to the activated oxygen atom in the Fe(IV)-oxo intermediate (Fig. 5). However, in 5fC, because of the conformation of the formyl group is restrained by the conjugation system formed between itself and the cytosine ring, the abstractable hydrogen on 5fC pointed toward another direction and is relatively far away from the activation oxygen atom (Fig. 5). Previous computational studies suggested The H-atom abstraction on the quintet surface favours the σ-pathway, thus requiring an essentially linear attack[40]. We indeed observed a linear conformation of the C5-H-oxo atoms in the transition states of hydrogen abstraction reactions of 5mC and 5fC (Fig. 5). As the formyl group on 5fC is retrained, it is more difficult for 5fC to achieve this linear conformation and therefore, elevated the abstraction barrier.

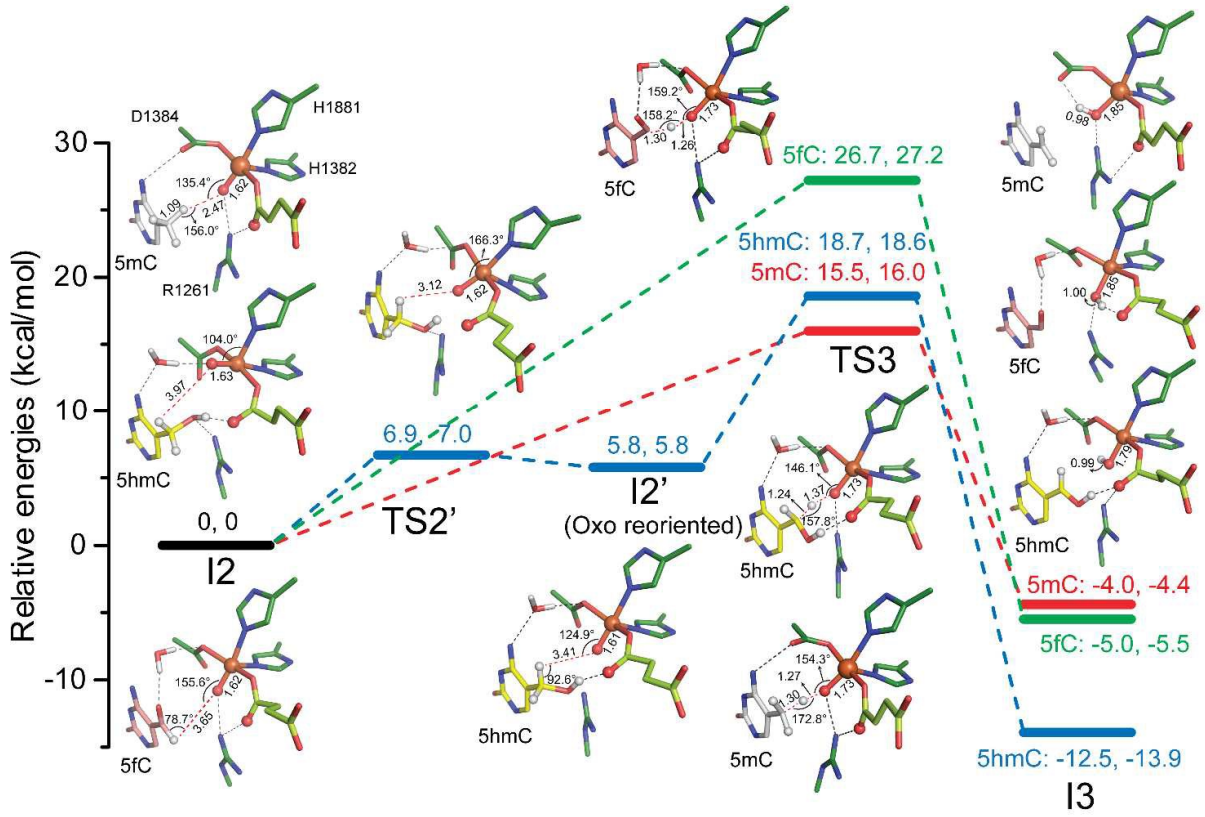


Fig. 5. Pathways for the hydrogen abstraction reactions of 5mC, 5hmC and 5fC. Relative potential energies (ΔE) and free energies (ΔG) for the second intermediate states (I1), the third transition states (TS2) and third intermediate states (I3) are given in kcal/mol (represented as ΔE, ΔG). For 5hmC, the energy and geometry of the transition state for oxo reorientation (TS2') and reoriented intermediate (I2') are also given. The geometry parameters (with distances in Å and angles in degree) for the bonds involved in the reactions are labelled. Hydrogen-bonding interactions are shown as dashes.

Different to the I2 structures of the TET2-5mC and TET2-5fC model, R1261 forms a hydrogen bond with the hydroxyl group of 5hmC and the oxo group forms a hydrogen bond with the water molecule observed in the crystal structure, which stabilized the oxo group in the axial position. As the oxo group in the axial position is far from the abstractable hydrogen on the substrate base, a reorientation process of the Fe(IV)-oxo species may needed before the hydrogen abstraction reaction, similar to the catalytic process of AlkB protein[41]. The potential energy barrier for Fe(IV)-oxo reorientation is found to be 6.9 kcal/mol and the potential energy reoriented product (I2') was also higher than that before the reorientation (Fig. 5). The abstraction energy for 5hmC from the I2' was about 13 kcal/mol, which is in accordance with that the C-H BDE for 5hmC is smaller than that for 5mC and 5fC[14]. However, according the transition state theory, as the energy of I2' is higher than that of I2, the overall energy barrier for the hydrogen abstraction barrier should be 18.7 kcal/mol, which is higher than 5mC and lower than 5fC (Fig. 5). The hydrogen atom on the methylene group of 5hmC can also be directly abstracted by the oxo group that lies in the axial position with a free energy barrier of 21.8 kcal/mol, which is higher than the two step abstraction process (Fig. S3).

In all, based on the QM/MM calculation results, we suggest the energy barriers for hydrogen abstraction reaction of 5mC, 5hmC and 5fC are significantly higher and those of the first two steps and follows the order of 5mC<5hmC<5fC. Compared with previous report, the hydrogen abstraction barriers do not strictly follow the C-H BDEs of substrate bases, and we suggest this may mainly because 5mC adopts a favourable conformation and results in lower energy barrier, whereas 5hmC and 5fC adopt unfavourable conformations restrained by inter/intramolecular hydrogen bonds and lead to higher energy barriers.

3.6 Hydroxyl rebound

The last step in the catalytic cycle of most Fe(II)/α-KG-dependent hydroxylases is the rebound of the hydroxyl radical to the substrate and the formation of the final product[20]. According to our QM/MM calculations, the reaction pathway for hydroxyl rebound of 5mC and 5fC were similar to previous reported hydroxyl rebound pathway for other Fe(II)/α-KG-dependent hydroxylases. The calculated potential energy barrier for hydroxyl rebound of 5mC was 16.9 kcal/mol, which is similar to that of the hydrogen abstraction barrier and the potential energy barrier of this step for 5fC was 20.7 kcal/mol, which was lower than the hydrogen abstraction barrier (Fig. 6). However, during the potential energy surface scan of the hydroxyl rebound process of 5hmC, we found the hydrogen on the hydroxyl group of 5hmC could directly transferred to the hydroxyl radical bound to Fe(III), which generated the final product (5fC and H2O) without hydroxyl rebound and formation of the hemiacetal intermediate (Fig. 6). As a result, the energy barrier of the final step was found to be significantly lower in TET2-5hmC model compared with TET2-5mC model and TET2-5fC model. Because there is not any hydroxyl group on 5mC and 5fC base, the direct transfer of the hydrogen on the hydroxyl group to the hydroxyl radical should be unique to the TET2 mediated oxidation on 5hmC. In addition, as the highest energy barrier determines the overall reaction rate for multi-step reactions, the overall catalytic efficiency of 5hmC should still be lower than that of 5mC, since the abstraction barrier for 5hmC is still higher than the abstraction barrier or hydroxyl rebound barrier of 5mC.

4 Conclusions

In this study, we thoroughly investigated the TET2 mediated oxidation on its three substrates—5mC, 5hmC and 5fC, in order to characterize the source of the substrate preference of TET2. Molecular dynamics (MD) simulations and "Alchemical" free energy calculations were performed based on the three TET2-substrate complex structures and the results indicated that 5mC, 5hmC and 5fC could all stably bind to the active site of TET2 with similar binding free energies, which is in accordance with previous biochemical experimental reports that TET2 binds to 5mC, 5hmC and 5fC containing DNA with comparable binding affinities. Subsequently, the hybrid quantum mechanical/molecular mechanical (QM/MM) calculations of the reaction pathways were the performed and the similarities and differences in the reaction pathways of 5mC, 5hmC and 5fC were compared (summarized in Table 1). The overall reaction pathway of TET2 mediated oxidation was found to be similar to several other Fe(II)/α-KG-dependent dioxygenases, which mainly involves four steps of reaction: decarboxylation, O-O heterolysis, hydrogen abstraction and

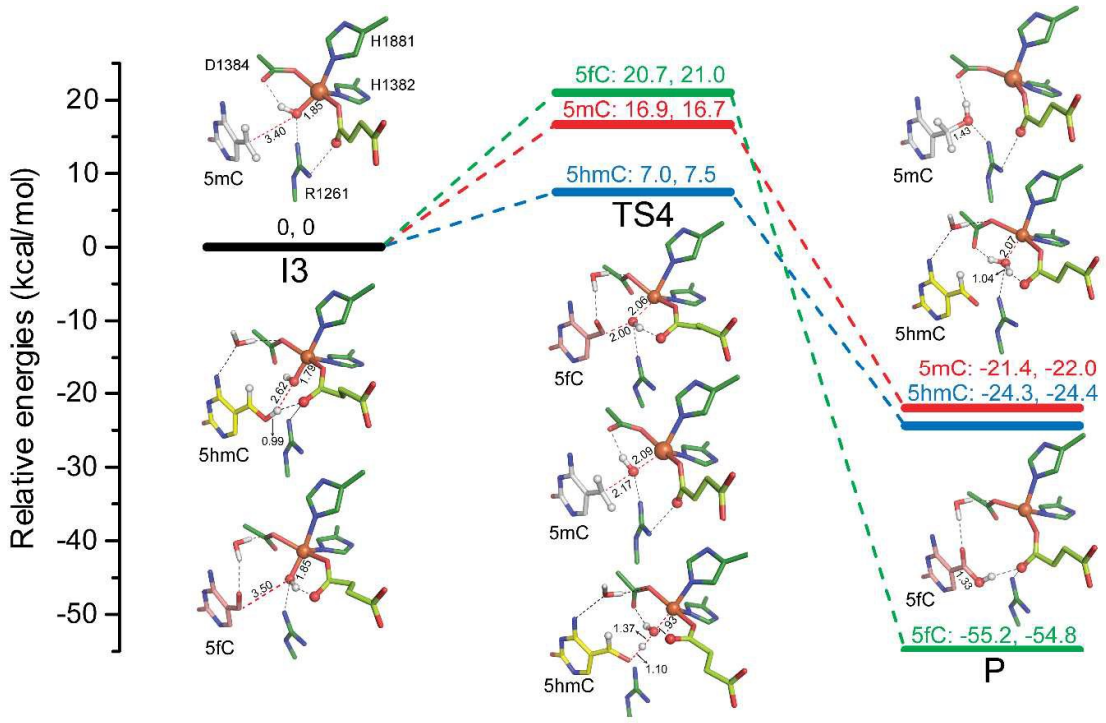


Fig. 6. Pathways for the hydroxyl rebound reactions of 5mC, 5hmC and 5fC. Relative potential energies (ΔE) and free energies (ΔG) for the third intermediate states (I3), the fourth transition states (TS4) and the final products (P) are given in kcal/mol (represented as ΔE, ΔG). The geometry parameters (with distances in Å and angles in degree) for the bonds involved in the reactions are labelled. Hydrogen-bonding interactions are shown as dashes.

hydroxyl rebound. The changes of electronic configurations of the catalytic centre atoms during the catalytic cycle were also found to be similar among the three substrates (Table S2). The energy barriers of the first two steps were relatively low and did not show significant differences among TET2-5mC/5hmC/5fC models. Therefore, the first two steps are unlikely the rate limiting steps of TET2 catalysed hydrolysis or responsible for the substrate preference.

On the other hand, the energy barriers for the third step, which is hydrogen abstraction, were significantly higher than the first two steps and also followed the order of the experimentally observed catalytic efficiencies. According to the QM/MM optimized intermediate and transient state structures, we suggest the differences in the estimated energy barriers are mainly resulted from the different orientations of the 5-substituation groups as well as the different configurations of the iron centres in the catalytic intermediate states. As the methyl group on 5mC can rotate freely, there is always a hydrogen pointed to the activated oxygen and can be directly abstracted from its position. However, the intramolecular hydrogen bond formed between the formyl group and N4 nitrogen on 5fC as well as its planner conformation hinder the approach of the hydrogen on the formyl group to the activated oxygen, and therefore the hydrogen abstraction barrier is elevated. As for 5hmC, the water molecule in the active site of TET2 and interaction between the hydroxyl group and R1261 prevent the spontaneous orientation of the oxo group and thus lengthen the distance between the abstractable hydrogen on 5hmC and the activated oxygen. An energy barrier is met when the oxo group in TET2-5hmC reorients to a more suitable position for hydrogen abstraction. In addition, the reoriented conformation of Fe(IV)-oxo species was also found to be less energetically favoured than the unreoriented conformation. Therefore, the energy barrier for 5hmC was found to be higher than that of 5mC. The energy barriers for the final step of TET2 mediated oxidation also showed large differences in different substrates. As the hydrogen on the hydroxyl group on 5hmC could directly transfer to the hydroxyl radical that bind to the iron, the energy barrier of the final step was found to be lower in TET2-5hmC model compared with TET2-5mC model and TET2-5fC model. Nevertheless, according to the transition state theory, as the overall catalytic efficiency is determined by the reaction step with this highest energy barrier, the energy barriers for the rate-determining step still follows the order of 5fC>5hmC>5mC.

In summary, our computational studies suggest the difference in the energy barriers of the hydrogen abstraction reaction step during the whole catalytic cycle of TET2 mediated oxidation contributes to the substrate preference of TET2 and this difference mainly resulted from the different orientations of the 5-substitution groups on substrate bases in the Fe(IV)-oxo intermediate states. As TET proteins are highly conserved in the catalytic site, the substrate preference is potentially an intrinsic feature of TET proteins. The relatively low oxidation efficiency for 5hmC and 5fC may be critical to support 5hmC as a stable epigenetic marker in the mammalian genome. Our simulated models of TET2-5mC/5hmC/5fC complexes in the Fe(IV)-oxo intermediate states, which have not been captured by crystal structures so far, may facilitate the mechanism-based design of covalent inhibitors of TET2 that could specifically block the continuous oxidation of 5mC in a certain stage and therefore providing chemical tools for epigenetic studies. In addition, a similar substrate preference towards substrate bases with different oxidation states has also been found in another Fe(II)/α-KG-dependent nucleotide modifying enzyme—FTO[48]. It will also be of interest to investigate whether the mechanism utilized by TET2 to distinguish different substrate is used by FTO or other nucleotide modifying enzymes to achieve substrate preference related to their biological functions.

Table 1. Summary of free energy barriers for all reaction steps of TET2 mediated hydroxylation on 5mC, 5hmC and 5fC.

Reaction Step	Free Energy Barrier (kcal/mol)		
	5mC	5hmC	5fC
Step 1 Decarboxylation	9.0	10.3	8.9
Step 2 O-O heterolysis	1.9	1.7	2.1
Step 3 Hydrogen abstraction	16.0	18.6	27.2
Step 3 (ωB97XD) Hydrogen abstraction	15.7	18.0	26.5
Step 4 Hydroxyl rebound	16.7	7.5	21.0

The energy barriers were estimated at UB3LYP/6-31G(d,p) level of theory. For step 3 (hydrogen abstraction), the barriers were additionally estimated at UωB97XD/6-31G(d,p) level, because ωB97XD functional includes dispersion corrections, which have been shown to be important in estimating the barrier for the hydrogen abstraction. Thermal dynamic corrections (at 298.15K) were obtained by performing harmonic vibrational frequency calculations.

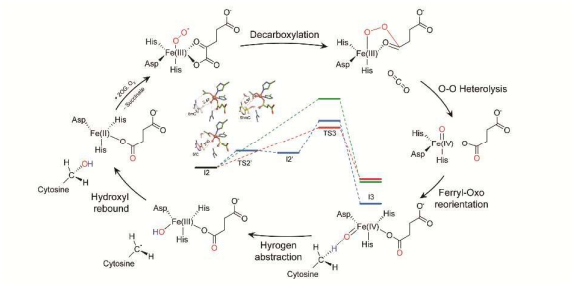
Acknowledgements

The computation resources were supported by Computer Network Information Centre, Chinese Academy of Sciences and Shanghai Supercomputing Centre. This work was supported by grants from the National Basic Research Program of China (2015CB910304 and 2011CB965300), the National Science & Technology Major Project “Key New Drug Creation and Manufacturing Program” of China (2014ZX09507-002), the National Natural Science Foundation of China (U1432242, 81430084, 21472208, 31425008, 91419301, 21210003), Basic Research Project of Shanghai Science and Technology Commission (12JC1402700), the Program of Shanghai Subject Chief Scientist (14XD1400500), the Hi-Tech Research and Development Pro-gram of China (2012AA020302), the Fund of State Key Laboratory of Toxicology and Medical Countermeasures, Academy of Military Medical Science (TMC201505), and “ShuGuang” project (11SG06) supported by Shanghai Municipal Education Com-mission and Shanghai Education Development Foundation.

References

1. A. P. Bird, *Nature*, 1986, **321**, 209-213.
2. E. Li, C. Beard and R. Jaenisch, *Nature*, 1993, **366**, 362-365.
3. Y. F. He, B. Z. Li, Z. Li, P. Liu, Y. Wang, Q. Tang, J. Ding, Y. Jia, Z. Chen, L. Li, Y. Sun, X. Li, Q. Dai, C. X. Song, K. Zhang, C. He and G. L. Xu, *Science*, 2011, **333**, 1303-1307.
4. V. P. S. Rawat, F. Mohr, K. Dohner and C. Buske, *Exp. Hematol.*, 2011, **39**, 272-281.
5. K. P. Koh, A. Yabuuchi, S. Rao, Y. Huang, K. Cunniff, J. Nardone, A. Laiho, M. Tahiliani, C. A. Sommer, G. Mostoslavsky, R. Lahesmaa, S. H. Orkin, S. J. Rodig, G. Q. Daley and A. Rao, *Cell Stem Cell*, 2011, **8**, 200-213.
6. N. Bhutani, J. J. Brady, M. Damian, A. Sacco, S. Y. Corbel and H. M. Blau, *Nature*, 2010, **463**, 1042-1047.
7. G. Ficz, M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. A. Hore, C. J. Marques, S. Andrews and W. Reik, *Nature*, 2011, **473**, 398-402.
8. S. Kriaucionis and N. Heintz, *Science*, 2009, **324**, 929-930.
9. Y. Xu, F. Wu, L. Tan, L. Kong, L. Xiong, J. Deng, A. J. Barbera, L. Zheng, H. Zhang, S. Huang, J. Min, T. Nicholson, T. Chen, G. Xu, Y. Shi, K. Zhang and Y. G. Shi, *Mol. Cell*, 2011, **42**, 451-464.
10. S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He and Y. Zhang, *Science*, 2011, **333**, 1300-1303.
11. S. Cortellino, J. Xu, M. Sannai, R. Moore, E. Caretti, A. Cigliano, M. Le Coz, K. Devarajan, A. Wessels, D. Soprano, L. K. Abramowitz, M. S. Bartolomei, F. Rambow, M. R. Bassi, T. Bruno, M. Fanciulli, C. Renner, A. J. Klein-Szanto, Y. Matsumoto, D. Kobi, I. Davidson, C. Alberti, L. Larue and A. Bellacosa, *Cell*, 2011, **146**, 67-79.
12. M. Bachman, S. Uribe-Lewis, X. P. Yang, M. Williams, A. Murrell and S. Balasubramanian, *Nat. Chem.*, 2014, **6**, 1049-1055.
13. H. Hashimoto, J. E. Pais, X. Zhang, L. Saleh, Z. Q. Fu, N. Dai, I. R. Correa, Jr., Y. Zheng and X. Cheng, *Nature*, 2014, **506**, 391-395.
14. L. Hu, J. Lu, J. Cheng, Q. Rao, Z. Li, H. Hou, Z. Lou, L. Zhang, W. Li, W. Gong, M. Liu, C. Sun, X. Yin, J. Li, X. Tan, P. Wang, Y. Wang, D. Fang, Q. Cui, P. Yang, C. He, H. Jiang, C. Luo and Y. Xu, *Nature*, 2015, DOI: 10.1038/nature15713.
15. L. Hu, Z. Li, J. Cheng, Q. Rao, W. Gong, M. Liu, Y. G. Shi, J. Zhu, P. Wang and Y. Xu, *Cell*, 2013, **155**, 1545-1555.
16. S. C. Trewick, T. F. Henshaw, R. P. Hausinger, T. Lindahl and B. Sedgwick, *Nature*, 2002, **419**, 174-178.
17. D. H. Lee, S. G. Jin, S. Cai, Y. Chen, G. P. Pfeifer and T. R. O'Connor, *J. Biol. Chem.*, 2005, **280**, 39448-39459.
18. G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang, C. Yi, T. Lindahl, T. Pan, Y. G. Yang and C. He, *Nat. Chem. Biol.*, 2011, **7**, 885-887.
19. E. Holme, *Biochemistry*, 1975, **14**, 4999-5003.
20. R. P. Hausinger, *Crit. Rev. Biochem. Mol. Biol.*, 2004, **39**, 21-68.
21. Discovery Studio, version 3.0; Accelrys Software Inc.; San Diego, CA, 2010.
22. R. Anandakrishnan, B. Aguilar and A. V. Onufriev, *Nucleic Acids Res.*, 2012, **40**, W537-541.
23. A. Perez, I. Marchan, D. Svozil, J. Sponer, T. E. Cheatham, 3rd, C. A. Laughton and M. Orozco, *Biophys. J.*, 2007, **92**, 3817-3829.
24. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712-725.
25. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926-935.
26. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157-1174.
27. E. Vanquelef, S. Simon, G. Marquant, E. Garcia, G. Klimerak, J. C. Delepine, P. Cieplak and F.-Y. Dupradeau, *Nucleic Acids Res.*, 2011, **39**, W511-W517.
28. X. Pang, K. Han and Q. Cui, *J. Comput. Chem.*, 2013, **34**, 1620-1635.
29. M. B. Peters, Y. Yang, B. Wang, L. Fusti-Molnar, M. N. Weaver and K. M. Merz, Jr., *J. Chem. Theory Comput.*, 2010, **6**, 2935-2947.
30. S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson and D. van der Spoel, *Bioinformatics*, 2013, **29**, 845-854.
31. T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089-10092.
32. B. Hess, H. Bekker, H. J. Berendsen and J. G. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463-1472.
33. H. J. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. Haak, *J. Chem. Phys.*, 1984, **81**, 3684-3690.
34. S. Dapprich, I. Komáromi, K. S. Byun, K. Morokuma and M. J. Frisch, *Journal of Molecular Structure: THEOCHEM*, 1999, **461**, 1-21.
35. B. Wang, K. R. Yang, X. Xu, M. Isegawa, H. R. Leverentz and D. G. Truhlar, *Acc. Chem. Res.*, 2014, DOI: 10.1021/ar500068a.
36. R. Lonsdale, J. N. Harvey and A. J. Mulholland, *J. Phys. Chem. Lett.*, 2010, **1**, 3232-3237.
37. D. Fang, R. L. Lord and G. A. Cisneros, *J. Phys. Chem. B*, 2013, **117**, 6410-6420.
38. C. Krebs, D. Galonic Fujimori, C. T. Walsh and J. M. Bollinger, Jr., *Acc. Chem. Res.*, 2007, **40**, 484-492.
39. D. R. Liu, M. Tahiliani, K. P. Koh, Y. H. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, L. Aravind and A. Rao, *Science*, 2009, **324**, 930-935.
40. S. Ye and F. Neese, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 1228-1233.
41. H. Liu, J. Llano and J. W. Gauld, *J. Phys. Chem. B*, 2009, **113**, 4887-4898.
42. T. Borowski, S. de Marothy, E. Broclawik, C. J. Schofield and P. E. Siegbahn, *Biochemistry*, 2007, **46**, 3682-3691.
43. S. P. de Visser, *J. Am. Chem. Soc.*, 2006, **128**, 9813-9824.
44. L. W. Chung, X. Li and K. Morokuma, *Quant. Biochem.*, 2010, 85-130.
45. T. Vreven and K. Morokuma, *Annu. Rep. Comput. Chem.*, 2006, **2**, 35-51.
46. J. Kaizer, E. J. Klinker, N. Y. Oh, J. U. Rohde, W. J. Song, A. Stubna, J. Kim, E. Munck, W. Nam and L. Que, Jr., *J. Am. Chem. Soc.*, 2004, **126**, 472-473.
47. S. P. de Visser, D. Kumar, S. Cohen, R. Shacham and S. Shaik, *J. Am. Chem. Soc.*, 2004, **126**, 8362-8363.
48. Y. Fu, G. Jia, X. Pang, R. N. Wang, X. Wang, C. J. Li, S. Smemo, Q. Dai, K. A. Bailey, M. A. Nobrega, K. L. Han, Q. Cui and C. He, *Nat. Commun.*, 2013, **4**, 1798.

Table of Contents



Computational modelling reveals the key step and atomic details for the substrate preference of TET2.

