Binding energies of nucleobase complexes: Relevance to homology recognition of DNA

Sergio Cruz León,[1,2] Mara Prentiss,[3] and Maria Fyta[1,*]

[1] *Institute for Computational Physics, Universität Stuttgart, Allmandring 3, 70569 Stuttgart, Germany*

[2] *Departamento de Ciencias Naturales, Escuela Colombiana de Ingeniería Julio Garavito, AK 45 205-59, Bogotá, Colombia*

[3] *Department of Physics, Harvard University, 17 Oxford Street, Cambridge, Massachusetts 02138, USA*

(Received 26 February 2016; published 13 June 2016)

The binding energies of complexes of DNA nucleobase pairs are evaluated using quantum mechanical calculations at the level of dispersion corrected density functional theory. We begin with Watson-Crick base pairs of singlets, duplets, and triplets and calculate their binding energies. At a second step, mismatches are incorporated into the Watson-Crick complexes in order to evaluate the variation in the binding energy with respect to the canonical Watson-Crick pairs. A linear variation of this binding energy with the degree of mismatching is observed. The binding energies for the duplets and triplets containing mismatches are further compared to the energies of the respective singlets in order to assess the degree of collectivity in these complexes. This study also suggests that mismatches do not considerably affect the energetics of canonical base pairs. Our work is highly relevant to the recognition process in DNA promoted through the RecA protein and suggests a clear distinction between recognition in singlets, and recognition in duplets or triplets. Our work assesses the importance of collectivity in the homology recognition of DNA.

DOI: 10.1103/PhysRevE.93.062410

I. INTRODUCTION

Nucleic acids are continuously susceptible to damage imposed by external factors like UV radiation or chemical exposure [1–4], as well as by internal processes. Chromosomes can be disrupted during the replication and recombination processes [5]. Permanent mutations can be generated on DNA molecules leading to serious problems, often related to cancer tumors [6,7]. Such processes can also result in the formation of mismatched base pairs (bps) or even inhibit gene expression [5]. However, cells have generated an enzymatic repair mechanism, known as the SOS repair system [8]. The SOS mechanism is activated when a chromosome is broken and activates a complex machinery involving proteins, which are capable of repairing damages with a high fidelity [9]. This reconstruction mechanism compares the broken chromosome with a complete chromosome through a process called homology recognition (HR). The HR process is promoted by a class of proteins known as RecA [10]. Once the HR is complete, RecA performs homologous alignment and strand exchange [11–13]. Homology sampling occurs via Watson-Crick-type base pairing [14]. The chromosome can then be reconstructed through the polymerase chain reaction mechanism [15]. These mechanisms render cells a robust system, despite the fact that nucleic acids residing in biophysical systems are continuously prone to damage, chemical and structural modifications.

The protein RecA involved in the recognition process belongs to a family of proteins. It is a DNA-dependent ATPase for performing homology recognition, while trying to repair DNA [16]. The RecA protein has been found in virtually all bacteria [17,18], in archaea [19], in bacteriophage genomes [20], as well as in eukaryotes [21] and humans [22]. RecA is regulated within the SOS response system and functions in a network that determines under which conditions and in which way RecA is expressed [7]. All processes related with the preservation and repair of the genome are centered in the recombination mechanism that RecA protein catalyzes. In fact, many studies have proven that problems in the regulation of the recombination process can drive to instabilities within the chromosomes and carcinogenesis [6,23,24]. Recombination and its regulation involve a plethora of proteins, but are still not fully understood [7].

When a chromosome is broken, the SOS system is activated and RecA proteins invade the broken single strand of DNA (ssDNA). The role of RecA is to compare base pair (bp) complexes in order to find the homologous partner of a ssDNA molecule. This process is referred to in literature as *homology search* or *homology recognition*. Base sampling is facilitated by having an extended and underwound conformation of the DNA [25], or the secondary DNA-binding site that mediates the homology sampling reaction binds to the strands of the double-stranded DNA (dsDNA), leaving the second strand available for base sampling [26]. The atomistic mechanism by which the HR process occurs is, though, not fully understood [27]. It is believed though, that the fidelity of the search process is governed by the distance between the DNA-binding sites [28]. Crystallographic evidence shows that RecA separates a strand in sequences of three bases (triplet), and then searches for the homologous partner. Alternatively, "recognition in duplets" is suggested as a possible mechanism in recent experiments [29]. In the case of a successful HR, RecA also catalyzes a process known as strand exchange in which alignment of the homologous regions and transfer of one of the strands of the dsDNA to the ssDNA takes place. This results in the reparation of a damaged chromosome.

It has been proposed that during the recognition process, within each triplet a nearly B-form structure of the DNA (B-DNA) is preserved [30]. This stacking is sufficient and allows the homology discrimination process to exploit the energy difference between the Watson-Crick (WC) and the mismatched pairing. Between the triplets, the large rises are related to a mechanical stress with important functional roles in homology recognition and strand exchange. HR

[*] Corresponding author: mfyta@icp.uni-stuttgart.de

attempts to verify whether a dsDNA is complementary to the target ssDNA, as implied above. For a correct pairing the complex must remain stable. For mismatched pairing (even if this involves only one mismatched base pair), the complex must rapidly unbind. In thermodynamic equilibrium, the populations in different binding configurations are determined by their binding energies. These can control the rapidness and accuracy of the recognition process [30]. It has also been proposed that in sequence recognition one mismatch in the B-DNA produces collective destabilization that extends to approximately six base pairs [31]. Experiments on homology recognition for B-DNA, when abasic sites are added to divide the bases, showed that dividing into groups improved recognition in comparison with recognition in B-form [32].

Additional information, such as the biochemical details of RecA were recently reported [33,34], as well as the geometry of the DNA strands probed by the RecA protein [35]. Nevertheless, details on the HR and strand exchange are further being investigated [29,30,36–38]. Several physical arguments have been followed to unravel in detail the recognition process. For example, studies have underlined the importance of taking into account the deformation of the backbone due to the presence of RecA [30] or the entropy and enthalpic effects involved in the process [29]. Recently, a detailed mechanism for HR based on a decision tree was suggested [39]. Within that model, collective effects are a key step in the decision cascade in HR [40].

Following, these previous investigations, our work aims to evaluate the role of collective effects in the HR process from a basic physical concept: the interaction energy. Based on plain physical arguments, this research explores the difference between recognition in singlets and recognition in duplets or triplets in agreement with the decision tree model [39]. To our knowledge, such an approach has not been used before and can provide valuable insight into fine differences within DNA, which can be essential for the HR process. From our view of the relevant studies, one experimental work refers to binding energies of the two ss-DNA sites without allowing for a comparable quantitative analysis to our data [28]. In this respect, our approach provides a novel investigation path, which relies only on basic physical concepts to understand complex biophysical processes. This paper is organized as follows: in Sec. II the methodology used is presented, in Sec. III the results are analyzed and discussed in Sec. IV, while in Sec. V the conclusions are given.

II. METHODOLOGY

Quantum mechanical calculations were used to determine the binding energy of different DNA base-pair complexes as a function of the number of mismatched DNA base pairs therein. The average binding energy for each system was obtained through a density functional theory (DFT) [41,42] approach as implemented in SIESTA [43]. All structures were equilibrated until the atomic forces reached 0.04 eV/Å. An energy cutoff of 300 Ry and triple-ζ basis plus polarization orbitals were used. Efficient pseudopotentials for the simulations were considered [44], as well as two different approximations for the exchange-correlation functional. The first is the generalized gradient approximation, which can well reproduce both hydrogen and covalent bonds [45,46]. For this, the approach proposed by Perdew, Burke, and Ernzerhof (PBE) was taken [47]. For the inclusion of long-range interactions, a nonempirical long-range exchange-correlation functional (vdW-DF2) was assumed [48]. The use of both functionals, PBE and vdW-DF2, is based on two reasons. First, our benchmark simulations showed that the use of the more complex vdW-DF2 functional after a first relaxation of the structures with the PBE functional is computationally more efficient than applying directly the vdW-DF2 functional. Second, it is possible to assess the role of the long-range dispersive interactions in the DNA complexes. The simulations are performed at zero temperature, without any interaction between the strands and the RecA protein, and excluding any surrounding aqueous medium.

The simulated structures are dsDNA stacked base-pair complexes of the B-form, which are typically involved in the recognition process [33]. These include two helical strands, each having the phosphate and sugar groups and a nitrogenous nucleobase (adenine (A), thymine (T), guanine (G), and cytosine (C)) [49]. Besides the canonical WC pairs (A-T and C-G), other mismatched pairing combinations are possible [50,51], a representative fraction of which are considered here. The binding energies for singlet, duplet, and triplet complexes of one, two, and three bps, respectively, with and without the presence of mismatches are calculated. All structures were generated using the *nucleic acid builder* [52]. Hydrogen atoms were added wherever necessary to assure that the structures are kept neutral. In the following, we will use the notation ATC-TAG for a dsDNA formed by a 3′-A-T-C-5′ strand hydrogen bonded to a 5′-T-A-G-3′ strand.

All systems considered here, were structurally relaxed using the PBE approach by adding constraints to the backbone. The constraints had to be added as the structures are very small and relaxation would significantly affect their helical form. These structures were next reoptimized at the level of dispersion correction (vdW-DF2) for all possible canonical and mismatched bps. The results shown will be based on the vdW-DF2 calculations, unless otherwise stated. Due to the use of atomic orbitals in our DFT approach, the counterpoise correction for basis set superposition error (BSSE) [53] was taken into account in the energy calculations. According to this, the binding energy ΔE_{bind} is evaluated as

$$\Delta E_{\text{bind}} = E_C - E_1^{\text{ghost}} - E_2^{\text{ghost}}. \quad (1)$$

In this expression, E_C is the total energy of both weakly bonded monomers in the complex, and E_i^{ghost} $(i = 1,2)$ is the total energy of the monomer i of the complex calculated using the ghost atomic orbitals of the other monomer. Each monomer is defined as one strand in the stacked base-pair complexes. The binding energy calculated here is related to the free energy revealing a mechanism for HR.

III. RESULTS

In order to evaluate the collective effects in HR the binding energies for all nucleobase complexes are calculated based on Eq. (1). We observe in which way the binding energy is being altered due to the presence of mismatches. The difference in the binding energy of the duplets and triplets compared to the sum of the binding energies of the respective singlets is also

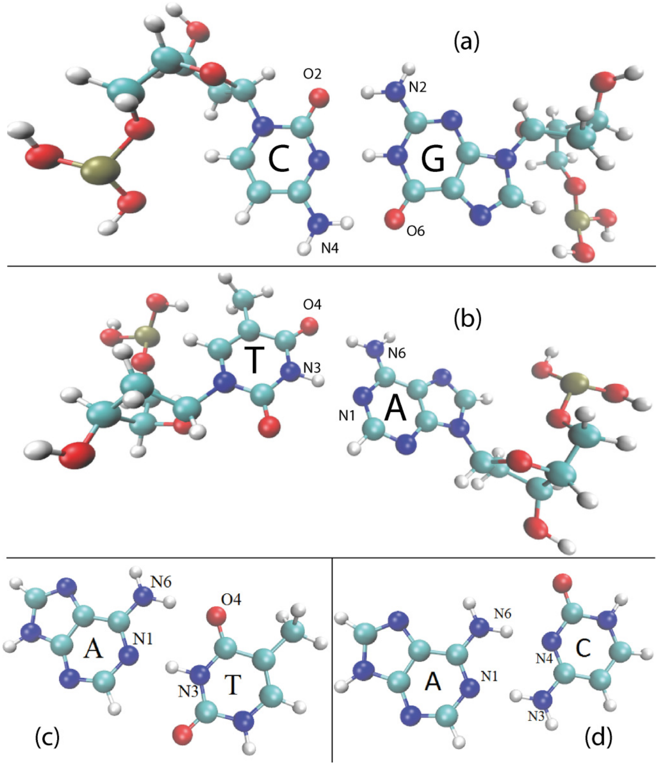

FIG. 1. Typical singlets studied in this work: (a) the A-T bp and (b) the G-C bp both including the backbone. In (c) and (d) the A-T bp and the A-C1^d bp both without the backbone are depicted. All structures were relaxed with the vdW-DF2 functional. The atom labeling shown in (c) and (d) is used in the analysis in Tables I and II. Carbon, hydrogen, oxygen, nitrogen, and phosphorous atoms are shown in cyan, white, red, blue, and gold, respectively. The same coloring scheme is used in all figures.

TABLE I. Hydrogen bond lengths (d) (in Å) within the singlets of Fig. 1 obtained through the vdW-DF2 approach. The results correspond to the singlets without the backbone. The atom labels correspond to that figure. These are compared to respective literature data (d^{ref}). The relative error (in %) obtained from the comparison is also given.

bp	Bonds	d	d^{ref}	Error
A-T	N6-O4	2.95	2.67 [55]	10.5
	N1-N3	2.88	2.79 [55]	3.2
A-C1^d	N1(A)-N4(C)	2.86	2.93 [54]	2.4
	N6(A)-N3(C)	2.86	2.91 [54]	1.7

estimated. A large difference would point to strong collectivity effects in the HR process.

A. Singlets

We begin our analysis with singlets, that is a single base pair (canonical or mismatched). Typical singlets modeled in this work are depicted in Fig. 1. For the canonical WC pairing, A-T and G-C are taken. For mismatched pairing, we choose the purine-pyrimidine mismatch A-C depicted in Fig. 1. This atomic arrangement has been found stable and was labeled A-C1^d [54]. The A-T and A-C bps also serve as a benchmark as relevant literature data are available.

The DFT energies obtained from this work are tested against available crystallographic data [56,57], previous DFT [55], and more accurate MP2 and CCSD(T) simulations [54]. Our results and a comparison to reference data on structural properties and energies are summarized in Tables I and II. The comparison is very good taking into account the level of accuracy of the DFT approach. The largest difference in the bond lengths was found equal to 0.28 Å and 0.07 Å for the WC and mismatched pairing, respectively. The binding energy calculated through Eq. (1) leads to a relative error of ≈17% for PBE and ≈14% for vdW-DF2. More accurate results for the energies are obtained from the dispersion corrected simulations compared to the PBE approach, as expected. The hydrogen bonding, though, depending on its directionality can sometimes be better captured by a PBE approach [46].

In the calculations for the singlets, we have neglected the backbone for simplicity and for allowing a comparison to available literature data mentioned above. In order to check the contribution of the backbone, we have repeated the calculations for the singlet bps including part of the backbone. Typical examples are shown in Fig. 1 for the WC bonded A-T and G-C. The binding energy from the PBE calculations for the A-T and G-C bps with the backbone were found equal to -0.72 and -1.34 eV, respectively. The vdW-DF2 binding energies are -0.70 and -1.26 eV, respectively. These energies show that the largest contribution to the binding energy comes from the base pairs and not from the backbone. Accordingly, the hydrogen bonding between the nucleobases in the pairs mostly influence the interaction of the two strands.

B. Duplets

At a second step, canonical WC and mismatched duplets are simulated. These are complexes of two stacked base pairs. In these simulations, the backbone is also included and kept fixed during relaxation. The influence of mismatched bps in the binding energies of the complexes is evaluated. Representative examples of the duplets are depicted in Fig. 2 for both WC and mismatched bp complexes. We begin with two different WC duplets: AA-TT and GG-CC. The first one corresponds to the 3′-A-A-5′ strand, which is hydrogen bonded to its complementary WC strand 5′-T-T-3′. In these WC duplets, the following mismatches are then incorporated: A-C, G-T, T-C, and T-T. (For purine-purine mismatches we have obtained unphysical results and do not further consider these. According to crystallographic evidence for such conformations, the

TABLE II. ΔE_{bind} for WC and mismatched singlets without a backbone calculated through Eq. (1). The DFT results are compared to more accurate MP2 and CCSD(T) schemes in the literature [54]. All values are in kcal/mol (eV).

bp	PBE	vdW-DF2	Ref. [54]
A-T	–18.15(−0.79)	–17.30(−0.75)	–15.4(−0.67)
G-C	–32.70(−1.42)	–29.37(−1.27)	–28.2(−1.22)
A-C1^d	–19.13(−0.83)	–18.46(−0.80)	–16.10(−0.70)

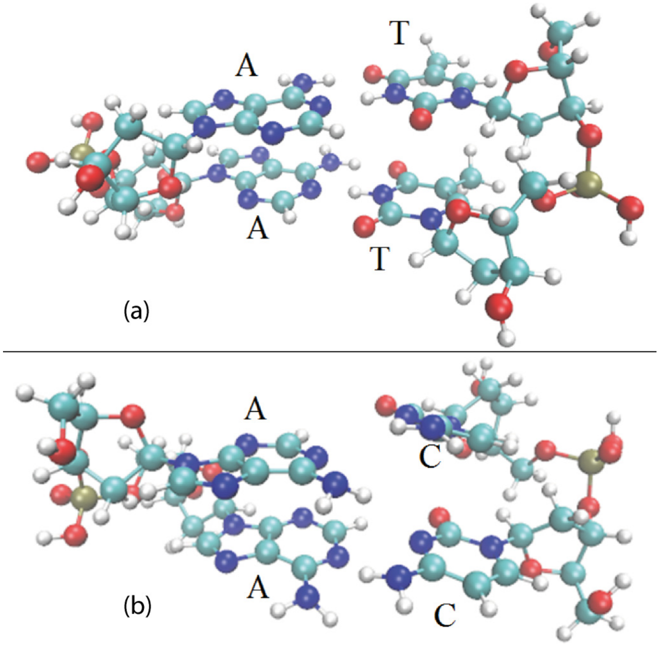

FIG. 2. vdW-DF2 relaxed atomic structures of (a) the canonical WC AA-TT and (b) the mismatched AA-CC bps. The AA-CC complex includes two mismatched bps. Both are typical duplets in our calculations.

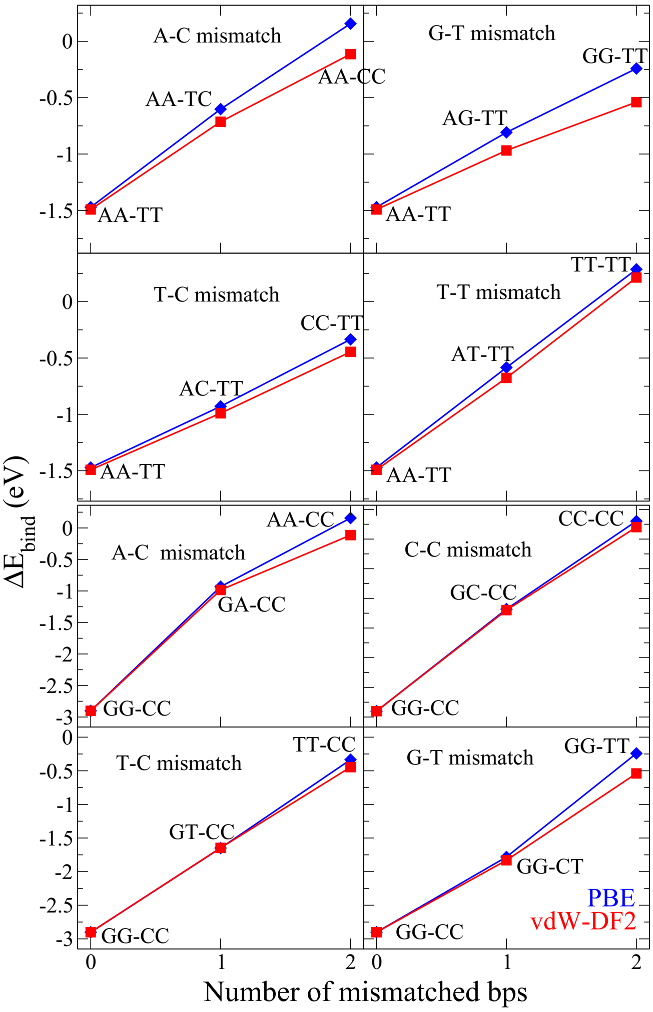

FIG. 3. Binding energies (ΔE_{bind}) for the WC duplets AA-TT (top) and GG-CC (bottom) with their respective mismatched combinations. The blue diamonds and red squares represent the PBE and vdW-DF2 results, respectively. In the x axes, "0" corresponds to the canonical WC pairing without mismatches, while "1" and "2" are pairings with one and two mismatches, respectively.

interstrand C1′-C1′ distance is farther apart than in a standard B-DNA WC bp [50,58–62].) In the case of the A-C mismatch, we have simulated AA-TC, which is a duplet with one purine-pyrimidine mismatch (A-C) and one WC bp (A-T). The AA-CC duplet, which contains two purine-pyrimidine mismatches, is also modeled. Within the same concept, the GG-CC WC duplet is modeled and the respective mismatched complexes GG-CT (a duplet with one G-T mismatch) and GG-TT, a duplet with two G-T mismatches. We proceed in the same way for a number of one and two mismatches in the WC AA-TT and GG-CC stacked bp complexes as summarized in the graphs of Fig. 3. In these the binding energies are presented as a function of the number of mismatched bps. Results are shown for both exchange-correlation functionals for comparison.

Inspection of Fig. 3 reveals a more or less linear behavior in the binding energy as the number of the mismatched base pairs increases. This trend will be discussed in the following. The linearity holds for all cases, with small deviations seen in the A-C and G-T mismatches for the GG-CC complex. Since for the A-C mismatch in the AA-TT WC pair the linearity holds, the small deviation from linearity for the A-C mismatch in GG-CC points to a possible issue with the GA-CC mismatch. This GA-CC pair seems to have a larger energy than expected for linearity denoting that the A-C mismatch in the GA-CC bp makes the pair more unstable than within other complexes having an A-C mismatch (AA-TC in the upper left panel of Fig. 3). This will be further discussed in the triplet cases below. The smaller deviation from linearity for the G-T mismatch in the GG-CC bp in the PBE case is almost negligible for the more accurate vdW-DF2 calculations. Further focus on these vdW-DF2 energy values reveals that a positive binding energy corresponds to TT-TT (T-T mismatch of the AT-AT bp) and CC-CC (C-C mismatch of the GC-GC bp). These bps are the most unfavorable ones according to our calculations.

Further observation of the energy values reveals that the two exchange-correlation functionals lead to very similar results, which deviate more as more mismatches are added. In all cases, we find that the PBE and vdW-DF2 results coincide for the WC pairing. They are very close when one mismatch is added and differ between 0.07 and 0.30 eV (for TT-TT and TT-GG, respectively) for two mismatches, that is the fully mismatched complex. Note that the same mismatches generated from different WC pairings may have slightly different binding energies. Specifically, the AA-CC mismatch can be generated either from the AA-TT WC bps or from the GG-CC bps by introducing two mismatches as implied above. Since the initial bps are structurally different, the same mismatches need to be accommodated within a different chemical environment in the AA-TT or GG-CC cases, resulting in slightly different binding energies. One should also have in mind that in our

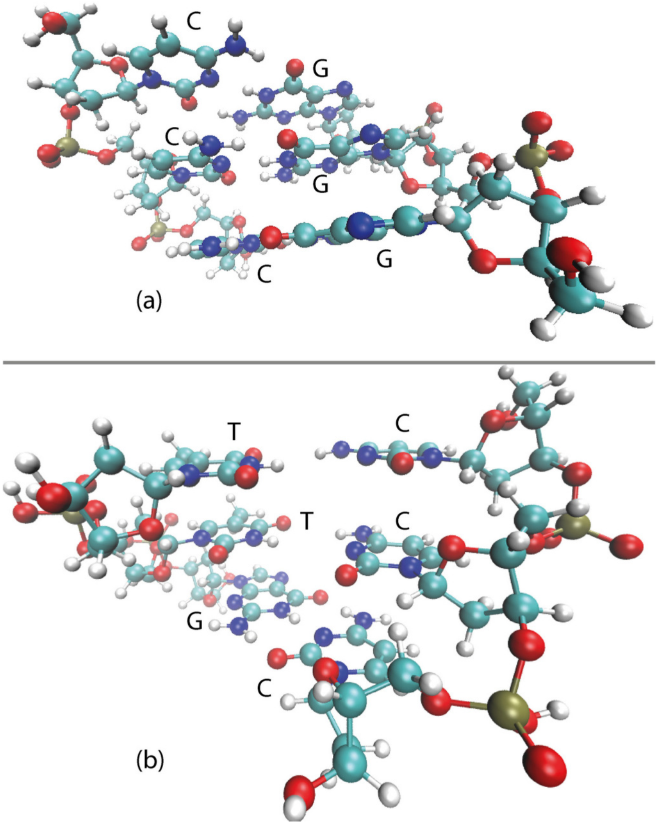

FIG. 4. (a) The CCC-GGG triplet modeled in this work. According to our convention, this complex is formed by a 3′-C-C-C-5′ strand bonded to a 5′-G-G-G-3′ strand. In (b) a TTG-CCC triplet with two mismatched bps is shown.

equilibration process the backbone has been kept fixed, not allowing for a completely free relaxation, which could also affect the respective energies.

C. Triplets

Following the same procedure described in Sec. III B, we move to triplets, that is bp complexes with three base pairs. Typical WC and mismatched triplets studied in this work are represented in Fig. 4. A WC AAA-TTT triplet includes a 3′-A-A-A-5′ strand paired with its complementary 5′-T-T-T-3′ strand. For example, representative triplet complexes with up to three A-C mismatches are AAA-TTT, AAA-TTC, AAA-TCC, and AAA-CCC. Similarly, for up to three T-T mismatches the respective triplets are AAA-TTT, AAT-TTT, ATT-TTT, and TTT-TTT, while for the G-T mismatches we model the bp complexes AAA-TTT, AAG-TTT, AGG-TTT, and GGG-TTT. Within the same spirit, the complexes for the C-T mismatch are AAA-TTT, AAC-TTT, ACC-TTT, and CCC-TTT. For the GGG-CCC bp the incorporation of mismatches is done in the same way. All binding energies for the triplets obtained using the vdW-DF2 exchange-correlation functional are summarized in Fig. 5.

This figure reveals that similar to the duplet cases, in the triplet cases the binding energy varies linearly with the number of mismatched bps. A deviation is observed in the A-C mismatch of the GGG-CCC bp, which again involves the A-C pair as in the duplet A-C mismatch discussed in the previous section. The calculations for the triplets confirm that the A-C mismatch is indeed unfavorable compared to other mismatches. Positive binding energies were found in all fully mismatched complexes (i.e., bp complexes with three mismatches) denoting unfavorable mismatching. An exception was found for the T-C mismatch for which the CCC-TTT complex has a negative binding energy. Regarding the AAA-TTT triplets, we observe a trend with respect to the type of mismatch. Specifically, the T-C mismatches have overall lower binding energies, followed by the T-G and the A-C mismatches. The case of the T-T mismatch is not clear, but has the highest energy for the triple mismatch (TTT-TTT). In the GGG-CCC triplet, the T-C mismatched cases also correspond to a lower binding energy followed by the G-T and the A-C mismatches, similar as in the AAA-TTT triplets. These trends point to a higher stability when

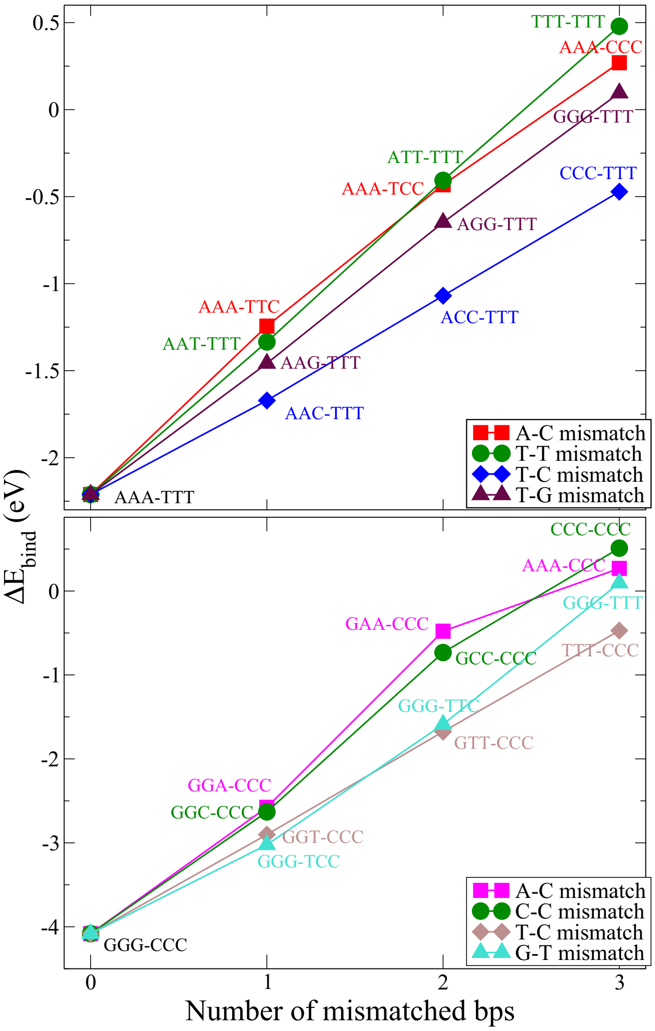

FIG. 5. Binding energy (ΔE_{bin}) for the WC triplets AAA-TTT (top) and GGG-CCC (bottom) with their respective mismatched combinations as a function of the number of mismatched base pairs. In the x axes, "0" corresponds to the canonical pairing without mismatches, while "1," "2," and "3" denote one, two, and three mismatches, respectively. All results are based on the vdW-DF2 functional.

a T-C mismatch is present compared to other mismatches. According to these results, the T-C mismatch can better accommodate in the bp complexes decreasing the binding energy.

IV. DISCUSSION

In order to understand in more detail the mechanism behind homology recognition of DNA, we investigated the issue of collectivity. For this, we have evaluated whether the comparison of the energetics of singlets, duplets, and triplets denote a collective behavior. We should first underline that the binding energies shown in Figs. 3 and 5 are large compared to experimental data [63]. For example, for the AA-TT and GG-CC canonical WC pairs, we find binding energies of -1.49 and -2.90 eV, respectively, while the experimental data correspond to -0.34 and -0.35 eV, respectively. Note, though, that the linear trends in the energies with respect to the addition of mismatches are observed here and in the experiments [63] as well. However, we attribute the discrepancy between theoretical and experimental values mainly to the fact that the sugar-phosphate backbone was kept fixed throughout the relaxations not allowing for a completely free relaxation of the structures. This was done due to the very small size of the DNA strands, which would otherwise not keep their helical conformation during a full relaxation. In addition, our simulations are static and are performed at 0K. Note that quantitatively the DFT results are not as accurate as higher order *ab initio* calculations but can qualitatively capture the trends presented in the figures. These trends are of interest in this work and not the exact binding energy values. The qualitative comparison of the binding energies for all bp complexes is sufficient in providing additional details in the HR process.

First, we assess the role of the backbone in the complexes following up the comments regarding the singlets with and without a backbone. An inspection of the binding energies of the A-T (with the backbone) ($E_{\text{bind}} = -0.70$ eV), AA-TT ($E_{\text{bind}} = -1.49$ eV), and AAA-TTT ($E_{\text{bind}} = -2.21$ eV) cases, reveals that there is a similar decrease of about -0.7 eV in the binding energy starting from A-T and adding one and two A-T pairs for constructing the duplet and triplet, respectively. The fact that each addition of one A-T unit to an A-T based complex (singlet or duplet) adds the same amount of energy to the binding energy of the resulting complex (duplet or triplet), which is almost the same as the binding energy of the single A-T unit, clearly denotes that the contribution of the backbone to the binding energy of the complexes is indeed very small. Accordingly, at a first approximation, the backbone can be neglected in the energetics analysis.

Collectivity, effects within the bp complexes will be first assessed by comparing the energy of a duplet (or a triplet) to the sum of the singlet energies composing the duplet (or triplet). For a WC pairing, our results point to a minor effect in considering the complexes as a whole or as a sum of singlets. As an example, we look at the difference of the binding energy of the WC duplet AA-TT (collective case) from the binding energies of two A-T singlets (noncollective case). Taking into account the vdW-DF2 binding energies given in Table II and Figs. 3 and 5, this difference is very small, about $-0.09(0.01)$ eV. For the triplet WC AAA-TTT, the difference in the binding energy of the complex from the respective value for three A-T singlets is $-0.11(0.04)$ eV for the singlet with(without) the backbone, respectively. The results for the G-C pairing are quite similar, with a difference between the collective and noncollective cases of $-0.38(-0.36)$ eV and $-0.30(-0.27)$ eV for the WC duplet and triplet, respectively. The numbers above are based on the singlets with the backbone. In parentheses are the differences in the binding energies based on the singlets without the backbone. Our results show first that when accounting for the backbone in the singlets, the differences in the binding energies of the collective from the noncollective cases are all negative, denoting higher binding energies for the latter cases. A second observation based on the energy analysis above indeed links to a collectivity effect [40], which is though expected to play a moderate role in the recognition process. In the AT complexes, the binding energy differences between the collective cases are small. In the GC complexes, these differences are higher, but still the relative differences are between 7 and 15% and seem to decrease as more bps are added. As discussed above, the backbone slightly contributes to the binding energies of the bp complexes. For this reason, the fact that the complexes include backbones of different sizes was not considered as it should only slightly alter the results.

For noncanonical pairing, that is for mismatches, the binding energy increases with the number of mismatches, as expected. For both duplets and triplets this variation in energy is linear, though the slopes of the lines change for different mismatches as evident from Figs. 3 and 5. This increase in the binding energy denotes less favorable binding than the WC pairing, pointing to a lower probability in having mismatches. Based on experimental data, it was proposed that the addition of a mismatched base pair into a WC bp complex does not significantly influence the free energy of the WC bps [63]. The results of our study can confirm this finding. Based on the values of the binding energies in Figs. 3 and 5 we summarize the binding energies of duplets and triplets with one mismatch and show these in the bar chart of Fig. 6. All AA-TT or GG-CC duplets with one mismatch are compared to the respective singlets, A-T or G-C (black lines in the figure). Similarly the AAA-TTT and GGG-CCC triplets with one mismatch are compared to the duplet cases, AA-TT and GG-CC (again the black lines), respectively. The comparison reveals that the relative difference in the binding energies of the duplets with respect to the WC singlets with one mismatch are in the range 1–45%. The fact, that the energy differences are larger than expected and some of our results seem to deviate from previous experimental observations about the influence of a single mismatch on a collective destabilization [31] are most probably related to temperature and solvent effects, neglected here. However, the results of Fig. 6 are not necessarily inconsistent with the experimental data when considering the duplet case, which was proposed to be the case for the first 9 bps tested in RecA [32].

The binding energies of the triplets with one mismatch are even closer to the respective energies of the WC duplets. The relative differences decrease to 0–17% for the triplets and are the smallest for the GGG-CCC cases.

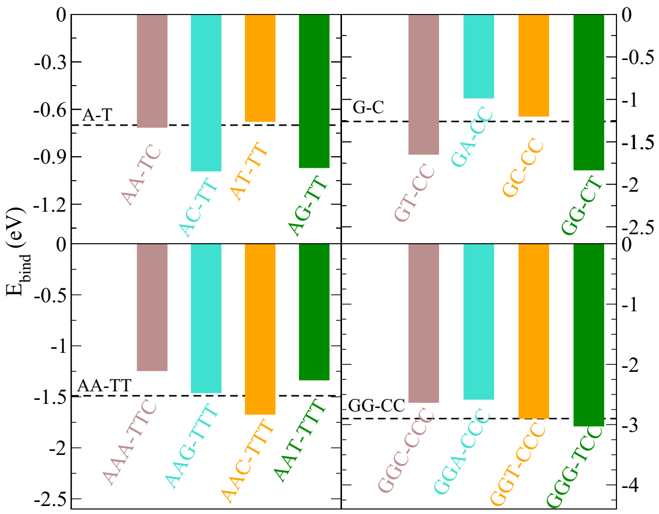

FIG. 6. Bar chart of the binding energies of duplets and triplets with one mismatch. The energies are compared to the energies for the singlets and duplets, respectively, denoted through the black dashed line (see text). The colored labels denote the bp complexes including the mismatch.

Accordingly, projection of our results to longer bp complexes implies that the influence of the mismatch added in a canonical pairing becomes even smaller. Specifically, consider a DNA strand with N (with $N > 3$) WC bps and a binding energy of E_{bind}^{N}. Addition of one mismatched pair would increase the DNA bps to $N + 1$ having a binding energy of E_{bind}^{N+1}. According to the above results, for these cases it is expected that $E_{\text{bind}}^{N} \approx E_{\text{bind}}^{N+1}$ for the addition of one mismatch in the WC bonded strands. These results indeed suggest that the main contribution to the binding (or free) energy of a bp complex originates from the WC pairing [39] and reproduces linear trends as a function of the number of WC pairs and mismatches [63].

V. CONCLUSIONS

Quantum mechanical simulations of canonical WC and mismatched DNA base pairs show that the binding energy of these bp complexes within dsDNA is a linear function of the number of mismatched bps. We have chosen to use DFT simulations to calculate the energetics of nucleobase complexes at a more accurate level compared to common atomistic simulations and in a computationally more efficient way than higher-order *ab initio* schemes. For this, we have neglected the RecA-DNA interactions and the surrounding fluid environment considering systems that are simplified with respect to the real structures but can produce reliable results on the qualitative energy differences of bp complexes. Accordingly, the trends we have observed are expected to be representative of the real systems and have the potential to serve as a basis for a further evaluation of collectivity effects through atomistic dynamical simulations [64]. In such investigations, the temperature and solvent effects could be quantitatively evaluated. It is expected that canonical conditions would impose fluctuations to the nucleobases, thus providing additional insight into their energetics.

The presence of a solvent, which can screen the charge of the DNA or interact with it by also entering into its major grooves, would also allow the investigation of a phase space supplementary to what was done here. For example, the indication that the energy differences decrease (Fig. 6) with the number of bps moving these closer to the experimental data is also related to the larger flexibility of the biomolecule in our calculations compared to the case where a solvent is present. With the inclusion of additional factors, such as temperature and solvent, the qualitative differences of the ground-state energetics of complexes of DNA bps compared to these systems in real conditions can be evaluated and serve a more complete physical understanding of the recognition process. Note, though, that the use of an implicit solvent to account for the effects of an aqueous environment would not be that efficient for two main reasons: (a) the background dielectric constant imposed by the implicit solvents will only add enthalpic effects to the enthalpic energies calculated here at 0K. The effects of a solvent are entropic and these are captured only by an explicit solvent, which could possibly also unveil additional stable conformations to those given by the enthalpic effects [65]. (b) An implicit solvent model neglects important effects, such as the charge hydration asymmetry, and would for this also fail to introduce the correct solvent influence [66]. Accordingly, the next step would be to include the solvent environment in an explicit way to fully account for the entropic effects and more closely resemble the setup of the experiments.

Using arguments based on the energetics of singlets, duplets, and triplets of canonical WC and mismatched nucleobase pairings, we have concluded that the presence of mismatched bps has a very small contribution to the binding energy of canonical pairing. This is in close agreement with recent results on other scales [39] and relevant experimental data [63]. Depending on the simple systems modeled here, this agreement highlights the very large importance of the base-pairing in the energy profiles in DNA rather than the influence of the environment. We have found that the binding energies increase linearly with the amount of mismatching. Accordingly, the bp complexes become less stable as additional mismatched bps are included in the complex. The comparison of the binding energies of duplets and triplets (collective case) to the binding energies of the number of singlets composing them (noncollective case) reveals that collectivity should play a moderate role in defining the binding energies of bp complexes. An interpretation of this statement is that a collective test of base pairs would not considerably improve the HR process. This is related to the fact that the presence of even one mismatched bp must rapidly unbind the complex, meaning that a collectivity search would delay this unbinding in the presence of mismatches. Overall, this work has attempted to analyze the energetics of base-pair complexes. The aim was to give another perspective of the way DNA is being recognized by resorting to the binding energies of single base pairs The basis of this study was to evaluate the influence of mismatched pairing and the trends of their binding energies in comparison to canonical WC pairing. It further needs to be checked whether our finding that mismatches in longer DNA strands become even less important is still valid. Force field approximations or hybrid approaches also involving the

RecA-DNA interactions can project the results of this work to a larger scale and also account for environmental effects.

ACKNOWLEDGMENTS

The authors thank C. Prévost for her correspondence on this work and for critically reading through the manuscript and J. Smiatek for insightful discussions on the revised version of this study. Funding from the German Funding Agency Deutsche Forschungsgemeinschaft (DFG) within the framework of the SFB716 collaborative network "Dynamic simulation of systems with large particle numbers" is greatly acknowledged. This work was performed on the computational resource bwUniCluster funded by the Ministry of Science, Research, and the Arts Baden-Württemberg (MWK) and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC. M.F. acknowledges support from the Juniorprofessorenprogramm funded by the Ministry of Science, Research, and the Arts Baden-Württemberg (MWK).

[1] J. Cairns, Nature **289**, 353 (1981).
[2] H. Soehnge, A. Ouhtit, and O. Ananthaswamy, Front Biosci. **2**, D538 (1997).
[3] R. Rundel and D. Nachtwey, Photochem. Photobiol. **28**, 345 (1978).
[4] F. De Gruijl, Eur. J. Cancer **35**, 2003 (1999).
[5] P. Hanawalt, *DNA Repair Mechanisms* (Elsevier, New York, 2012).
[6] P. E. Cohen and J. W. Pollard, Bioessays **23**, 996 (2001).
[7] M. M. Cox, Crit. Rev. Biochem. Mol. Biol. **42**, 41 (2007).
[8] M. Radman, in *Molecular Mechanisms for Repair of DNA* (Springer, Berlin, 1975), pp. 355–367.
[9] K. Ragunathan, C. Joo, and T. Ha, Structure **19**, 1064 (2011).
[10] M. Cox, Mol. Microbiol. **5**, 1295 (1991).
[11] S. C. Kowalczykowski and A. K. Eggleston, Annu. Rev. Biochem. **63**, 991 (1994).
[12] A. Roca and M. Cox, Crit. Rev. Biochem. Mol. Biol. **25**, 415 (1990).
[13] Z. Chen, H. Yang, and N. P. Pavletich, Nature **453**, 489 (2008).
[14] E. Folta-Stogniew, S. O'Malley, R. Gupta, K. S. Anderson, and C. M. Radding, Mol. Cell **15**, 965 (2004).
[15] J. M. Bartlett and D. Stirling, in *PCR Protocols* (Springer, Berlin, 2003), pp. 3–6.
[16] W. Selbitschka, W. Arnold, U. B. Priefer, T. Rottschäfer, M. Schmidt, R. Simon, and A. Pühler, Mol. Gen. Genet. **229**, 86 (1991).
[17] V. Brendel, L. Brocchieri, S. J. Sandler, A. J. Clark, and S. Karlin, J. Mol. Evol. **44**, 528 (1997).
[18] A. I. Roca and M. M. Cox, Prog. Nucl. Acid Res. Mol. Biol. **56**, 129 (1997).
[19] S. Yang, X. Yu, E. M. Seitz, S. C. Kowalczykowski, and E. H. Egelman, J. Mol. Biol. **314**, 1077 (2001).
[20] A. Lopes, J. Amarir-Bouhram, G. Faure, M.-A. Petit, and R. Guerois, Nucl. Acids Res. **38**, 3952 (2010).
[21] T. Ogawa, A. Shinohara, A. Nabetani, T. Ikeya, X. Yu, E. Egelman, and H. Ogawa, in *Cold Spring Harbor Symposia on Quantitative Biology* (Cold Spring Harbor Laboratory Press, New York, 1993), Vol. 58, pp. 567–576.
[22] P. Baumann and S. C. West, Trends Biochem. Sci. **23**, 247 (1998).
[23] B. de Massy, Trends Genet. **19**, 514 (2003).
[24] K. J. Hillers and A. M. Villeneuve, Curr. Biol. **13**, 1641 (2003).
[25] C. Danilowicz, E. Feinstein, A. Conover, V. W. Coljee, J. Vlassakis, Y.-L. Chan, D. K. Bishop, and M. Prentiss, Nucl. Acids Res. **40**, 1717 (2012).
[26] O. N. Voloshin and R. D. Camerini-Otero, Mol. Cell **15**, 846 (2004).
[27] A. Barzel and M. Kupiec, Nat. Rev. Genet. **9**, 27 (2008).
[28] I. De Vlaminck, M. T. van Loenhout, L. Zweifel, J. den Blanken, K. Hooning, S. Hage, J. Kerssemakers, and C. Dekker, Mol. Cell **46**, 616 (2012).
[29] L. Jiang and M. Prentiss, Phys. Rev. E **90**, 022704 (2014).
[30] J. Vlassakis, E. Feinstein, D. Yang, A. Tilloy, D. Weiller, J. Kates-Harbeck, V. Coljee, and M. Prentiss, Phys. Rev. E **87**, 032702 (2013).
[31] I. I. Cisse, H. Kim, and T. Ha, Nat. Struct. Mol. Biol. **19**, 623 (2012).
[32] A. Peacock-Villada, V. Coljee, C. Danilowicz, and M. Prentiss, PLos ONE **10**, e0130875 (2015).
[33] M. M. Cox, in *Molecular Genetics of Recombination* (Springer, Berlin, 2007), pp. 135–167.
[34] M. M. Cox, Nat. Rev. Mol. Cell. Biol. **8**, 127 (2007).
[35] C. Prévost and M. Takahashi, Q. Rev. Biophys. **36**, 429 (2003).
[36] A. Peacock-Villada, D. Yang, C. Danilowicz, E. Feinstein, N. Pollock, S. McShan, V. Coljee, and M. Prentiss, Nucl. Acids Res. **40**, 10441 (2012).
[37] Y. Savir and T. Tlusty, Mol. Cell **40**, 388 (2010).
[38] K. Klapstein, T. Chou, and R. Bruinsma, Biophys. J. **87**, 1466 (2004).
[39] D. Yang, B. Boyer, C. Prévost, C. Danilowicz, and M. Prentiss, Nucl. Acids Res. **43**, 10251 (2015).
[40] M. Prentiss, C. Prévost, and C. Danilowicz, Crit. Rev. Biochem. Mol. Biol. **50**, 453 (2015).
[41] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[42] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[43] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, J. Phys.: Condens. Matter **14**, 2745 (2002).
[44] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).
[45] I. A. Filot, A. R. Palmans, P. A. Hilbers, R. A. van Santen, E. A. Pidko, and T. F. de Greef, J. Phys. Chem. B **114**, 13667 (2010).
[46] J. Ireta, J. Neugebauer, and M. Scheffler, J. Phys. Chem. A **108**, 5692 (2004).
[47] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[48] K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, Phys. Rev. B **82**, 081101 (2010).
[49] J. D. Watson, F. H. Crick *et al.*, Nature **171**, 737 (1953).
[50] S. Neidle, *Principles of Nucleic Acid Structure* (Academic Press, San Diego, 2010).

[51] R. E. Kelly and L. N. Kantorovich, J. Phys. Chem. C **111**, 3883 (2007).

[52] "BioMaPS Institute in the Rutgers University make-na server", http://structure.usc.edu/make-na/server.html, accessed: 2015-08-19.

[53] S. F. Boys and F. d. Bernardi, Mol. Phys. **19**, 553 (1970).

[54] P. Jurečka, J. Šponer, J. Černý, and P. Hobza, Phys. Chem. Chem. Phys. **8**, 1985 (2006).

[55] C. Espejo and R. R. Rey-González, Revista Mexicana De Fisica S **53**, 212 (2007).

[56] L. Clowney, S. C. Jain, A. Srinivasan, J. Westbrook, W. K. Olson, and H. M. Berman, J. Am. Chem. Soc. **118**, 509 (1996).

[57] A. Gelbin, B. Schneider, L. Clowney, S.-H. Hsieh, W. K. Olson, and H. M. Berman, J. Am. Chem. Soc. **118**, 519 (1996).

[58] G. G. Prive, U. Heinemann, S. Chandrasegaran, L.-S. Kan, M. L. Kopka, and R. E. Dickerson, Science **238**, 498 (1987).

[59] T. Brown, W. N. Hunter, G. Kneale, and O. Kennard, Proc. Natl. Acad. Sci. USA **83**, 2402 (1986).

[60] G. A. Leonard, E. D. Booth, and T. Brown, Nucl. Acids Res. **18**, 5617 (1990).

[61] G. D. Webster, M. R. Sanderson, J. V. Skelly, S. Neidle, P. F. Swann, B. F. Li, and I. J. Tickle, Proc. Natl. Acad. Sci. USA **87**, 6693 (1990).

[62] J. V. Skelly, K. J. Edwards, T. C. Jenkins, and S. Neidle, Proc. Natl. Acad. Sci. USA **90**, 804 (1993).

[63] J. SantaLucia, Proc. Natl. Acad. Sci. USA **95**, 1460 (1998).

[64] B. Boyer, J. Ezelin, P. Poulain, A. Saladin, M. Zacharias, C. H. Robert, and C. Prévost, PLoS ONE **10**, e0116414 (2015).

[65] J. Smiatek, C. Chen, D. Liu, and A. Heuer, J. Phys. Chem. B **115**, 13788 (2011).

[66] A. Mukhopadhyay, A. T. Fenley, I. S. Tolokh, and A. V. Onufriev, J. Phys. Chem. B **116**, 9776 (2012).