PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

How acidic amino acid residues facilitate DNA target site selection

Kazi Amirul Hossain[a], Mateusz Kogut[a], Joanna Słabońska[a], Subrahmanyam Sappati[a,b], Miłosz Wieczór[a,c], and Jacek Czub[a,b,1]

Edited by Robert B. Best, National Institutes of Health, Bethesda, MD; received July 20, 2022; accepted December 5, 2022, by Editorial Board Member Adriaan Bax

Despite the negative charge of the DNA backbone, acidic residues (Asp/Glu) commonly participate in the base readout, with a strong preference for cytosine. In fact, in the solved DNA/protein structures, cytosine is recognized almost exclusively by Asp/Glu through a direct hydrogen bond, while at the same time, adenine, regardless of its amino group, shows no propensity for Asp/Glu. Here, we analyzed the contribution of Asp/Glu to sequence-specific DNA binding using classical and ab initio simulations of selected transcription factors and found that it is governed by a fine balance between the repulsion from backbone phosphates and attractive interactions with cytosine. Specifically, Asp/Glu lower the affinity for noncytosine sites and thus act as negative selectors preventing off-target binding. At cytosine-containing sites, the favorable contribution does not merely rely on the formation of a single H-bond but usually requires the presence of positive potential generated by multiple cytosines, consistently with the observed excess of cytosine in the target sites. Finally, we show that the preference of Asp/Glu for cytosine over adenine is a result of the repulsion from the adenine imidazole ring and a tendency of purine–purine dinucleotides to adopt the BII conformation.

DNA–protein recognition | transcription factors | MD simulations | DNA base preference

Significance

Due to their negative charge, the role of acidic amino acids (Asp/Glu) in DNA–protein recognition is often overlooked. In this report, we show that cytosines are almost exclusively recognized by Asp/Glu and play a surprisingly important role in sequence-specific DNA–protein recognition. Importantly, we found that Asp/Glu generally act as negative selectors, helping prevent unproductive off-target binding that would slow down target search. On target sites, cytosine often occurs within cytosine tracts, preferentially recognized via Asp/Glu thanks to the accumulation of positive electrostatic potential. We further clarify at the molecular level why Asp/Glu show a strong preference for cytosine rather than adenine, even though both expose an amino group in the major groove.

Sequence-specific DNA–protein interactions are vital to cellular functions, playing a regulatory role in virtually all DNA-templated processes, including gene expression, initiation of DNA replication, and site-specific recombination (1–3). Therefore, revealing how proteins efficiently discriminate cognate DNA sites from the vast excess of noncognate sites is of critical importance; in particular, it is necessary to fully understand how transcription factors shape spatiotemporal expression patterns of their target genes (4, 5).

Because of the repetitive nature of the sugar-phosphate backbone, site-specific DNA-binding proteins (DBP) have to rely on differences in the structure and properties of the DNA bases. Specifically, different π-stacked arrays of DNA bases present distinguishing patterns of hydrogen bond donors and acceptors to DNA grooves, have different desolvation free energies, and locally affect the conformation and flexibility of the double helix (6–11). DNA-binding proteins, on the other hand, contain sequence-reading motifs that are able to exploit these often subtle differences to achieve an appropriate level of specificity for cognate sites.

Most importantly, many DBPs evolved binding surfaces composed predominantly of polar residues that are complementary to the unique pattern of functional groups on a specific base sequence. This mode of recognition relies primarily on the formation of direct or water-mediated hydrogen bonds (12–16) with DNA bases, and hence it is usually called direct (or base) readout. Since it requires intimate contact with the nucleobases, direct reading is typically carried out in the major groove where the polar base-pair edges are more accessible to protein residues (17–19).

The conformational flexibility of the DNA helix and the energetic cost of its distortion (e.g., local bending) has been shown to depend markedly on the base sequence (20–23). Thus, the affinity of many DBPs for their cognate DNA sites might also depend on the ease with which the DNA helix can be distorted into the conformation, facilitating complex stabilizing interactions with the protein (24–28). Since this sequence recognition mode does not require direct contact with the bases, it is known as indirect (or shape) readout and is particularly important for proteins binding to the minor groove (29–33).

Most of the sequence-specific DBPs combine direct and indirect interactions to fine-tune the binding affinity for their cognate DNA sites (15, 34, 35). However, quantitative decomposition of the overall affinity into contributions due to different types of interactions is challenging. After all, the ever-increasing number of high-resolution structures primarily provide information on the architecture and not energetics of

Author affiliations: [a]Department of Physical Chemistry, Gdańsk University of Technology, Gdańsk 80-233, Poland; [b]BioTechMed Center, Gdańsk University of Technology, Gdańsk 80-233, Poland; and [c]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona 08028, Spain

Author contributions: M.W. and J.C. designed research; K.A.H., M.K., J.S., and S.S. performed research; K.A.H., M.K., M.W., and J.C. analyzed data; and K.A.H. and J.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.B.B. is a guest editor invited by the Editorial Board.

Copyright © 2023 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: jacek.czub@pg.edu.pl.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2212501120/-/DCSupplemental.

Published January 12, 2023.

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

protein/DNA complexes, while binding free energy measurements usually lack structural detail. This limits our understanding of protein–DNA recognition at the molecular level and consequently makes the bottom–up design of synthetic sequence-reading motifs a nontrivial task (36–39). Notably, despite previous attempts (13, 40–42), it is generally not clear how much specific amino acid–base contacts contribute to cognate site discrimination and, consequently, how a given DNA sequence can be efficiently targeted via direct readout.

The analysis of base–amino acid contact preferences in all available high-resolution structures of protein/DNA complexes provides an important qualitative insight into the energetics of direct readout (Fig. 1 and *SI Appendix*, Fig. S1). As expected, Arg and Lys, the two basic residues which are positively charged at physiological pH, are the most common among all residues interacting with nucleobases in the major groove (Fig. 1 *A*, *Top*). In addition to providing a "nonspecific" affinity in terms of electrostatic attraction to the negatively charged DNA backbone, both Arg and Lys show preferential hydrogen bonding to the guanine (G) base and thus contribute to recognition of G-containing sequences (Fig. 1 *A*, *Bottom*). The other polar residues, particularly Gln, Asn, Ser, and Thr, are also frequently found in the major groove and, acting as both proton donors and acceptors, are more promiscuous in their interaction with nucleobases with a preference for the adenine (A) base observed for Asn and Gln. Remarkably, both acidic residues, Asp and Glu, are almost as frequently in contact with nucleobases as the other nonbasic residues, although being negatively charged, they interact unfavorably with the DNA backbone and thereby may be expected to destabilize the complex. This observation raises a question about the energetic nature of direct readout mediated by Asp and Glu. Explaining this apparent discrepancy is important because, as revealed by the amino acid propensities in Fig. 1*B*, direct recognition of the cytosine (C) base in the major groove occurs almost exclusively via Asp or Glu. This preference can at least partially be attributed to a hydrogen bond forming between Asp or Glu and the N4-amino group exposed by cytosine in the major groove of the double helix (Fig. 1*C*). However, if this is the case, why is not adenine, which also has its N6-amino group exposed in the major groove, recognized by the acidic residues too?

In a previous study on the telomeric repeat-binding factor 1 protein (TRF1), the role of a specific Asp residue in the recognition of the human telomeric sequence containing three consecutive cytosines has been explored by per-residue decomposition of the binding free energy (43). It has been found that Asp has almost no net effect on the binding affinity for the target telomeric sequence most likely because the favorable interaction with the cytosines is nearly offset by the electrostatic repulsion. In contrast, the presence of Asp can reduce the affinity for nontelomeric sequences by up to 3 kcal/mol, suggesting that the actual role of Asp is to avoid off-target binding, especially to those sequences that lack cytosine. Therefore, it could be concluded that acidic residues mostly sense the absence of cytosine and thus act as "negative" selectors, unlike basic (and other polar) residues which increase affinity for the cognate site markedly more compared to noncognate sites (hence providing "positive" selection) (43).

Here, we applied extensive classical and ab initio (∼120 μs and ∼1.7 ns of sampling time, respectively) free energy simulations to understand the role of acidic residues in base readout in energetic and structural terms. By computing the contribution of Asp/Glu to the DNA binding affinity of five selected DBPs against a systematic set of DNA sequences, we find that Asp/Glu indeed disfavor the binding to sequences that lack cytosine, consistently with the notion of negative selection. In contrast, at the target, cytosine-containing sites, the effect of Asp/Glu generally varies from net neutral to favorable with the increasing number of cytosines in the vicinity of Asp/Glu. We show that this cumulative effect arises from long-range electrostatic attraction to cytosines as well as provide molecular-level explanation of the observed strong preference of Asp/Glu for cytosine over adenine.

Results and Discussion

Acidic Residues Act as Negative Selectors by Helping to Avoid Cytosine-Poor Sequences. To understand the energetics of direct readout mediated by the acidic amino acid residues, and specifically to test whether it occurs according to the negative selection mechanism (43), we first evaluated contributions of the selected interfacial Asp or Glu residues to the DNA-binding affinity for a diverse set of five protein/DNA complexes (*SI Appendix*, Table S3). The contribution was defined as the difference in the DNA-binding free energy, $\Delta\Delta G$, between the wild-type protein and its mutant in which a given Asp or Glu residue was substituted by alanine (Fig. 2*A*). To calculate the contributions, we used a thermodynamic cycle in which $\Delta\Delta G$ is obtained by transforming an acidic residue into alanine either in the presence or in the absence of bound DNA (Fig. 2*A*). Mutational free energy changes corresponding to these "alchemical" transformations (ΔG_m) were computed through Hamiltonian-replica exchange molecular dynamics simulations of the examined protein/DNA complexes (*Methods* for details). To capture the dependence of $\Delta\Delta G$ on local DNA sequence, for each of the examined complexes, we sampled the sequence space by creating DNA variants by substituting either a directly H-bonded cytosine only or all bases within 5 Å of Asp or Glu (Fig. 2*B* and *SI Appendix*, *Methods* and Table S1 for details). This systematic approach resulted in 24 independent DNA/protein systems (*SI Appendix*, Fig. S3 and Table S1) for which the calculated $\Delta\Delta G$ are shown in *SI Appendix*, Table S3.

$\Delta\Delta G$ values in *SI Appendix*, Table S3 reveal that the contributions of acidic residues to DNA affinity differ markedly depending on whether Asp or Glu interact directly with cytosine (C sequences) or other canonical nucleobases (A and G/T sequences). In particular, they are found to disfavor the binding to the non-C sequences (with average $\Delta\Delta G$ of 1.10 ± 0.74 kcal/mol and 0.65 ± 0.36 kcal/mol for A and G/T, respectively; Fig. 3 *A*, *Right*), at the same time, slightly increasing the affinity for the C sites (with average $\Delta\Delta G$ of -1.41 ± 0.60 kcal/mol). This finding is consistent with the known nucleobase propensities (Fig. 1), including a strong preference for cytosine over adenine which in principle should also be capable of forming an H-bond with the carboxylic group. Since the computed $\Delta\Delta G$ are generally unfavorable for the non-C sequences and often only marginally favorable for the C sequences, our data also support the notion of a negative selection mechanism in which acidic residues prevent the protein from (off-target) binding to non-C sequences.

Regardless of these general conclusions, we note that even though at the C sequences, Asp or Glu always form a single H-bond to the closest cytosine, their contribution to the affinity for these sequences can vary in a wide range, from negligible (∼0 kcal/mol) to highly favorable (−4 kcal/mol). Also unexpectedly, for a few non-C sequences, we found negative $\Delta\Delta G$ (see e.g., C2T and C2A variants of Zif268 in *SI Appendix*, Table S3). These observations indicate that base readout by Asp and Glu

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

depends on the local sequence context and other structural characteristics affecting the direct interaction.

To clarify this dependence, we first identified associations between the simulation-derived $\Delta\Delta G$ contributions and relevant structural features of the protein/DNA complexes by calculating Pearson correlation coefficients (Fig. 3*A*, *Left*). Specifically, to understand the importance of direct H-bonds and local sequence composition, the employed set of features includes the average number of H-bonds formed by Asp/Glu with the two possible partners: cytosine (#Hb-C) and adenine (#Hb-A), as well as the number of different nucleobases in the vicinity of Asp/Glu (#A, #C, #G, and #T). Since the goal is to study the involvement of Asp/Glu in direct readout of base functional groups, the latter features were obtained as the number of exocyclic functional groups exposed in the major groove (i.e., N4-amino group of C, N6-amino group of A, O6-carbonyl group of G, and O4-carbonyl group of T) within a 5 Å cutoff of Asp/Glu, averaged over the trajectory. To incorporate the effect of (partial) dehydration of Asp/Glu at the interface and competing salt bridges with neighboring basic residues, we used the number of H-bonds with water molecules (#Hb-H_2O) and the number of contacts with Arg/Lys (#Arg/Lys) as two additional features.

Linear correlations shown in Fig. 3*A* reveal that the number of H-bond with cytosine (#Hb-C) correlates relatively well with a favorable contribution of Asp/Glu to the DNA binding affinity. However, unexpectedly, an even better predictor of $\Delta\Delta G$ seems to be the number of cytosine residues in the local sequence of Asp/Glu (#C).

To independently test this finding and include possible nonlinear correlations between $\Delta\Delta G$ and our set of features, we used a hierarchical random forest-based approach in which the features are ranked according to their importance in prediction using the Shapley values (*SI Appendix*, *Methods* for details).

As can be seen from *SI Appendix*, Fig. S4, #C again ranks highest, but in a nonlinear approach, its predictive power is as much as twice that of #Hb-C (average absolute shift from the baseline prediction of $\Delta\Delta G$ are 1.0 and 0.45 kcal/mol for #C and #Hb-C, respectively). This means that recognition of C-containing sites by acidic residues does not exclusively rely on the formation of a single H-bond but is dependent on the number of cytosine N4-amino groups in the local vicinity of Asp/Glu (Fig. 3*B*). This is also consistent with a significant statistical overrepresentation of cytosine-rich sequences among the DNA sites recognized by the acidic residues in the experimentally solved protein/DNA complexes (*SI Appendix*, Fig. S5).

One could speculate that this cumulative effect of cytosine is due to either cooperativity of H-bonds formed simultaneously with two adjacent N4-amino groups or the possibility to dynamically switch between them, which would lead to a smaller entropic penalty upon binding to DNA. However, our data indicate that Asp/Glu at any given time can only form at most one bond with the N4-amino groups even at multicytosine sites where more than one cytosine is present in the immediate vicinity of Asp/Glu (*SI Appendix*, Table S2). We also found that (with the exception of TRF1), the acidic residues exhibit a single dominant binding mode and do not dynamically switch between different cytosine H-bond donors (*SI Appendix*, Fig. S6). Thus,

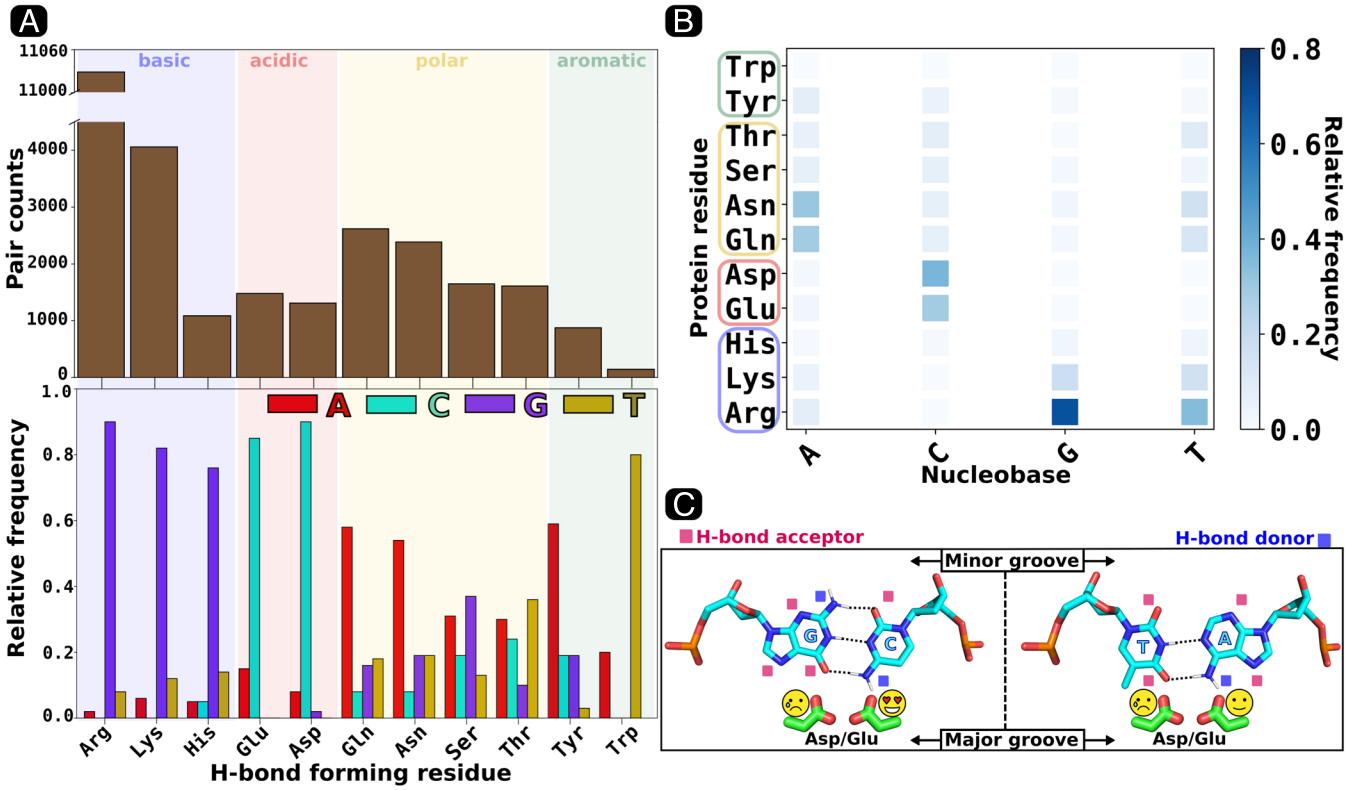

Fig. 1. Amino acid–base preferences in the DNA major groove, calculated based on 4623 protein/DNA structures deposited in the PDB database. Hydrogen bonding and contact information was retrieved from the DNAproDB database (44). (*A*, *Top*) Total number of amino acid residues of a given type (with side chains capable of forming hydrogen bonds) within 4.5 Å of any nucleobase in the major groove, across all protein/DNA complexes. (*Bottom*) Relative frequencies with which a given amino acid side chain forms a hydrogen bond with one of the four nucleobases in the major groove. (*B*) Relative frequencies with which a given nucleobase forms a hydrogen bond with specific amino acid side chains in the major groove. (*C*) Schematic representation of possible modes of interaction between the acidic residues (Asp/Glu) and the GC or AT base pairs in the major groove.

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

as will be discussed more elaborately below, we conclude that the cumulative effect of cytosine can be attributed to favorable longer-range attractive electrostatic interactions between Asp/Glu and the N4-amino groups of cytosine.

Among the remaining features, only the number of guanines close to Asp/Glu (#G) shows nonnegligible correlation with $\Delta\Delta G$ (Fig. 3*A*), which is understood given its Watson–Crick pairing with cytosine (*SI Appendix*, Fig. S7). Consistently with the known nucleobase propensities (Fig. 1), other bases (#A and #T) do not seem to have any power in explaining $\Delta\Delta G$. Also, compared to #Hb-C, the number of H-bonds with adenine (#Hb-A) turned out to be significantly less correlated with $\Delta\Delta G$, reflecting the preference for cytosine readout by acidic residues.

Even though most other considered features showed little predictive power over the entire set of DNA/protein complexes, some of them proved useful in interpreting the outliers. In particular, for C2T and C2A variants of Zif268, unexpectedly negative $\Delta\Delta G$ values (-1.6 and -1.8 kcal/mol, respectively) could be explained in terms of the formation of salt bridges with the neighboring basic residues (#Arg/Lys). Indeed, for Zif268, we observed that the favorable interaction between the Asp residue and the neighboring basic residues increases markedly upon DNA binding, with this increase being even more pronounced for non-C sequences (*SI Appendix*, Fig. S8). These observations indicate that our limited data set might not be sufficient to capture all subtle factors affecting the contribution of Asp/Glu to DNA affinity.

Positive Potential Generated by Accumulation of Cytosine Is Essential for the Favorable Binding of Asp/Glu . To understand the observed cumulative effect of cytosine on the binding free energy of Asp/Glu to DNA, we first estimated the enthalpic contribution to the obtained $\Delta\Delta G$ values and decomposed it into interactions between the acidic residue of interest and the remaining constituents of the system (Fig. 4*A*).

It is seen from Fig. 4*A* that out of four canonical bases, only cytosine helps to offset a strong electrostatic repulsion between Asp/Glu and the negatively charged sugar–phosphate backbone (BB), providing an average enthalpic stabilization of ∼10 kcal/mol per one base present in the immediate vicinity of the acidic residue. Notably, this stabilization increases with the number of cytosine amino groups in the local sequence within 8 Å of Asp/Glu (Fig. 4*B*). In fact, the attractive interactions between Asp/Glu and cytosine strengthen greatly when three or more N4-amino groups are present in a tract.

It can be thus concluded that the observed preference of Asp/Glu to interact with cytosine-rich sites arises from the long-range electrostatic attraction to the exocyclic cytosine amino groups. Consistent with this conclusion, *SI Appendix*, Fig. S9 shows that a patch of positive electrostatic potential in the major groove becomes more prominent with the increasing number of cytosines in the local sequence.

Fig. 4*A* also reveals that, because of the negatively charged carbonyl oxygens exposed to the major groove by guanine and thymine, both these bases markedly disfavor the binding of acidic residues (by 15 and 8 kcal/mol, respectively). Importantly, despite having an amino group in the major groove, adenine shows only negligible attractive interaction with Asp/Glu it is in contact with (-1 kcal/mol). At the same time, the dependence of the interaction energy on the number of adenine residues around Asp/Glu (#A) shows that longer-range electrostatic interactions with adenine are net repulsive (*SI Appendix*, Fig. S10).

The observed significant differences in the enthalpic contributions among the four nucleobases correlate with their known propensities (Fig. 1) and thus seem to provide a molecular-level

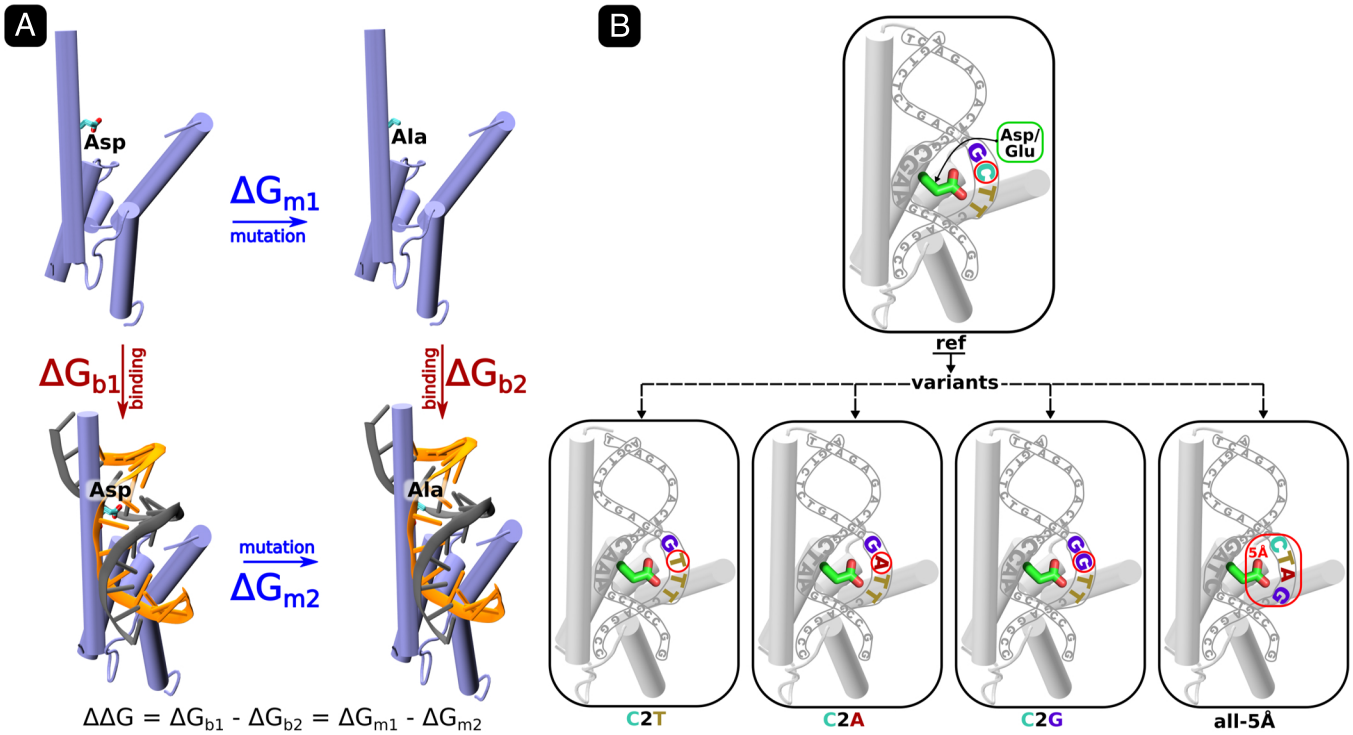

Fig. 2. (*A*) Thermodynamic cycle used to calculate the difference in the DNA-binding free energy between the wild-type protein and its Ala mutant ($\Delta\Delta G_b$), representing the contribution of Asp/Glu to the binding affinity to a given DNA site (protein in light blue; DNA in orange and gray). $\Delta\Delta G_b$ is obtained by subtracting the free energies of "alchemically" transforming Asp/Glu into Ala in the absence and in the presence of DNA (ΔG_{m1} and ΔG_{m2}, respectively). (*B*) Schematic representation of the sequence sampling procedure: Starting from the selected experimental protein/DNA complexes (ref), we obtained DNA variants either by substituting a cytosine directly H-bonded to Asp/Glu to thymine (C2T), adenine (C2A), and guanine (C2G) or by substituting all the nucleobases within the 5 Å of Asp/Glu (all-5 Å) (*SI Appendix*, Figs. S2 and S3, and Table S1 for all generated complexes and *Methods* for details).

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

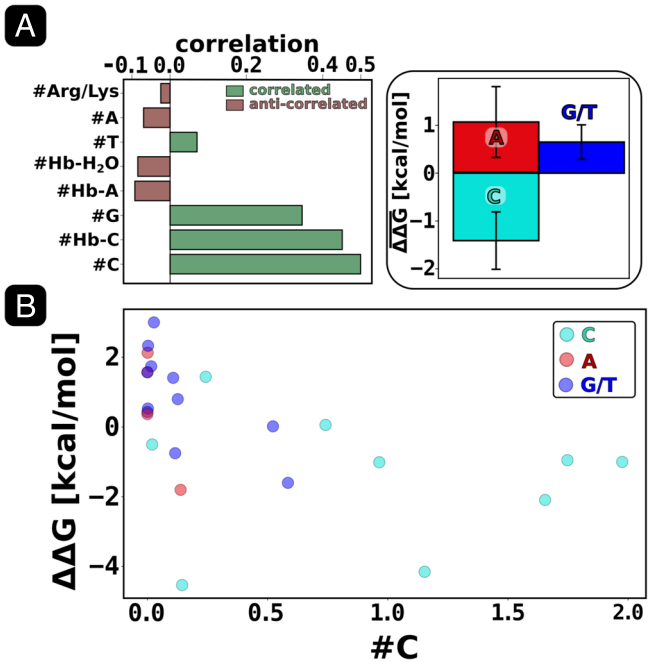

Fig. 3. (*A*) Correlations between the simulation-derived $\Delta\Delta G$ values and the relevant structural features of DNA/protein complexes, evaluated as Pearson correlation coefficients. The features tested are defined in the text (for details and numeric values of the features, *SI Appendix*, *Methods* and Table S2). Positive correlation indicates that an increase in the value of a given feature makes the contribution of Asp/Glu to the binding affinity more favorable (i.e., $\Delta\Delta G$ becomes more negative). (*Right*) The simulation-derived $\Delta\Delta G$ values averaged over the DNA sites with cytosine, adenine, or the remaining nucleobases interacting directly with Asp/Glu (C, A, or G/T sequences, respectively). (*B*) Dependence between the simulation-derived $\Delta\Delta G$ values and the number of cytosine residues in the local vicinity of Asp/Glu (#C).

explanation for the negative selection mediated by the acidic residues.

Interestingly, it can be seen from Fig. 4*A* that interactions of the acidic side chain with the solvent and, to a lesser extent, the rest of the protein also promote binding of Asp/Glu to DNA. The former contribution results from the accumulation of K^+ counterions at the DNA surface (*SI Appendix*, Fig. S11) and the latter from the stabilization of salt bridges between Asp/Glu and abundant basic residues (Arg and Lys) due to partial dehydration at the DNA/protein interface.

Repulsion from the Imidazole Ring and Propensity for BII Conformation Lead to the Low Affinity of Asp/Glu for Adenine. Next, we asked what is the molecular basis underlying a strong preference of the acidic residues for cytosine over adenine, despite both nucleobases having an amino group exposed in the major groove (Figs. 1 and 4*A*). To this end, we calculated the difference in the free energy of Asp/Glu binding to cytosine and adenine, using a model system containing a single propionic acid molecule, mimicking an acidic side-chain, competing for the interaction with cytosine and adenine on adjacent sites in a canonical B-DNA decamer (*Methods* for details). To examine the preferential interaction, cytosine and adenine nucleobases were made equally accessible in the major groove, by using the 5′-GTCAAT-3′ sequence in the middle of the decamer (Fig. 5*A* and *SI Appendix*, Fig. S14).

It has been reported that compared to pyrimidine–pyrimidine and pyrimidine–purine dinucleotide steps, purine–purine steps have a much higher propensity to deviate from the canonical BI phosphate conformation by favoring the BII conformation (*SI Appendix*, Fig. S12) (45). This BI/BII population ratio is thought to play an important role in the DNA readout by proteins (46). Thus, to determine whether BI/BII conformational dynamics affect the base preferences of Asp/Glu, we used one more sequence (5′-TGACAT-3′) in which cytosine and adenine are still equally accessible in the center of the decamer; however, the dinucleotide step involving the adenine (i.e., GpA) is known to populate the BII conformation (45, 47) (Fig. 5*A* and *Methods* for details).

The prepared systems were subject to conventional MD simulation with the propionate anion kept near the DNA surface with a flat-bottom harmonic potential to obtain its spatial distribution in the major groove (*Methods* for details). From this distribution, the relative free energy of propionate binding to cytosine and adenine, $\Delta\Delta G_{ca}$, was obtained as $-RT\ln(p_a/p_c)$, where p_a and p_c denote the probability of forming an H-bond with the central cytosine and adenine, respectively. The computed propionate distributions were well equilibrated, as indicated by the convergence of $\Delta\Delta G_{ca}$ shown in *SI Appendix*, Fig. S15.

Fig. 5*A* shows that the binding of propionate to cytosine is markedly more favorable than to adenine (by 0.8 and 1.9 kcal/mol for the first and second sequences, respectively), consistently with the propensities extracted from the structural data (Fig. 1) and our interaction analysis (Fig. 4*A*). The stronger preference for cytosine observed for the second sequence (Fig. 5*A*) can presumably be attributed to the tendency to adopt the BII conformation (*SI Appendix*, Fig. S12).

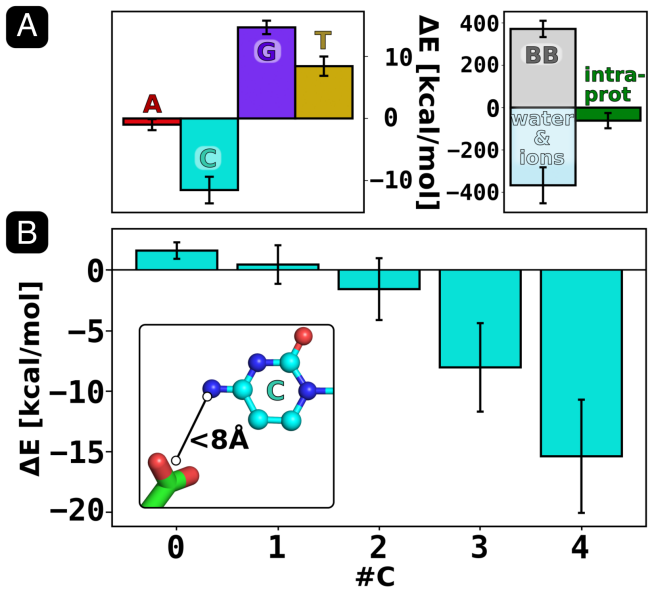

Fig. 4. (*A*) Enthalpic contributions to the binding free energy, computed as the average changes in the interaction energy (ΔE) between the examined acidic residues and other constituents of the system: DNA nucleobases (A, C, G, and T), DNA backbone (BB), solvent (water & ions), and the rest of the protein (intra-prot). The contributions from A, C, G, and T are calculated as the average interaction energy with the nucleobase of a given type present in the immediate vicinity (within 4.5 Å) of Asp/Glu (*B*) Asp/Glu–cytosine interaction energy as a function of the number of cytosine N4-amino groups present within 8 Å of the acidic residue. Since the interaction energy is averaged over all systems satisfying the cutoff criterion, including those in which Asp/Glu does not form close contact with the cytosine base, the increments associated with each additional cytosine are smaller than the per-base enthalpic contribution computed for direct Asp/Glu–cytosine interaction in *A*.

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

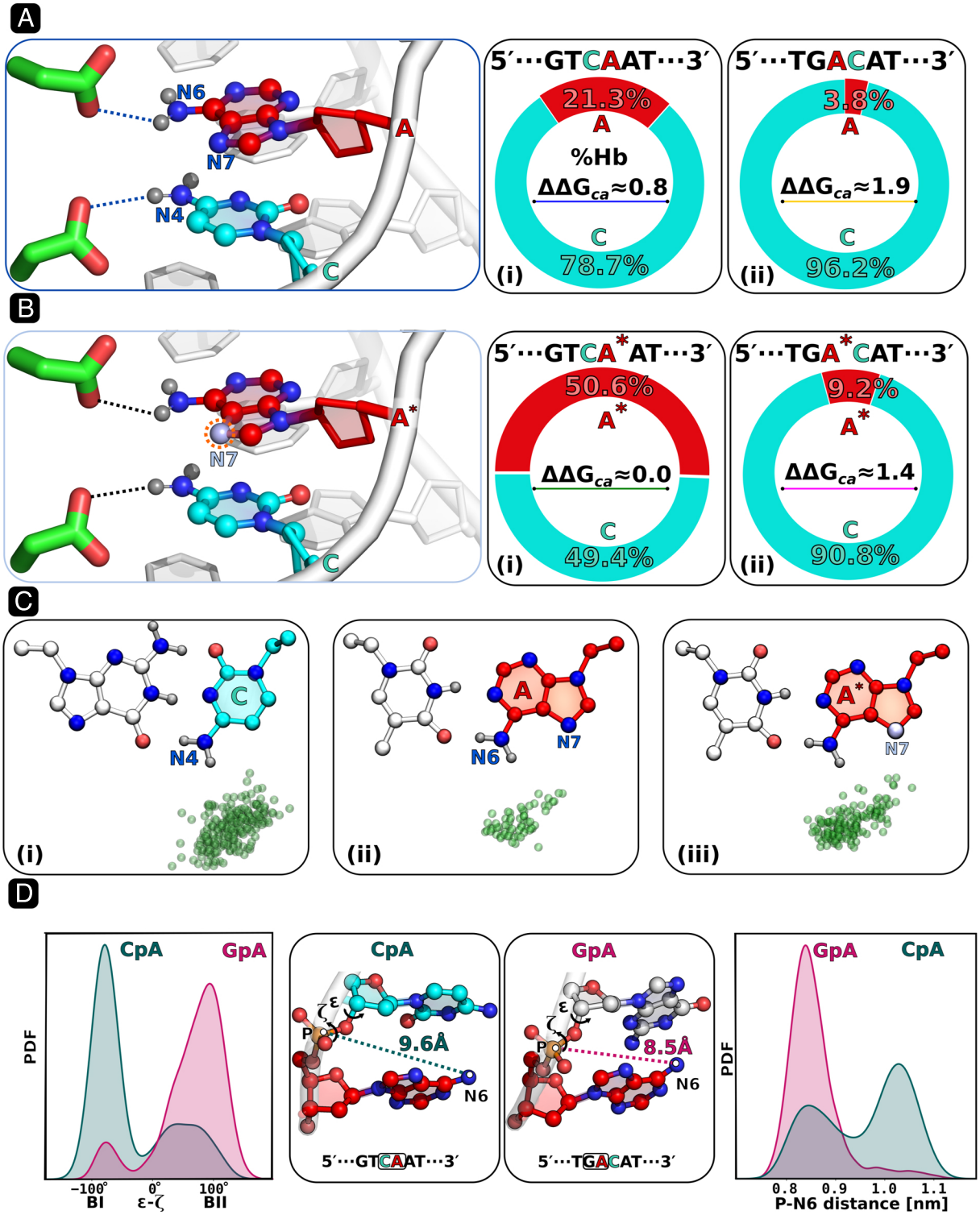

Fig. 5. (*A*) Percentage of hydrogen bonds formed by the propionate ion with the cytosine N4-amino group (cyan) and the adenine N6-amino group (red) in two different sequence contexts, shown in subpanels (*i*) and (*ii*). The corresponding differences in the free energy of propionate binding to cytosine and adenine, $\Delta\Delta G_{ca}$ (in kcal/mol), were calculated directly from the equilibrium H-bond populations by Boltzmann inversion. (*B*) Same as (*A*) except that the negatively charged N7 atom in the adenine imidazole ring was made electrically neutral in both sequences (the modified adenine is denoted as *A**). (*C*) Spatial distribution of the propionate ion hydrogen-bonded to the N4-amino group of cytosine (*C*, subpanel *i*), N6-amino group of adenine (*A*, *ii*), and modified adenine (*A**, *iii*). Green spheres represent the carboxylic carbon atoms, and their numbers are proportional to equilibrium H-bond populations. (*D*, *Left*) BI/BII population ratio for the CpA and GpA dinucleotide steps containing the central adenine in both studied sequences shown as the distributions of the difference between ϵ and ζ torsion angles (for definitions *SI Appendix*, Fig. S12). (*Right*) Distribution of the distance between the phosphate group and the adenine amino group (P–N6) for the two respective dinucleotide steps. Structural representations of the two steps are shown in the middle panel along with the average P–N6 distances. The distributions of relevant helical parameters characterizing the BI/BII equilibrium (twist, roll, and x-disp) are shown in *SI Appendix*, Fig. S13.

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

Therefore, to better understand the origin of these differences, we first compared the spatial distributions of the propionate anion around its binding partners (Fig. 5*C*). We found that the propionate clearly avoids close contact with the negatively charged N7 nitrogen in the adenine imidazole ring (Fig. 5*C*), which suggests that the repulsive interaction between them might be a major factor responsible for lowering the affinity for adenine. To test this conclusion, next, we neutralized the partial charge of the adenine N7 and recomputed $\Delta\Delta G_{ca}$ for both sequences, using the same MD approach (*Methods* for details). It can be seen in Fig. 5*B* that the neutralization of N7 indeed made the hydrogen bonding with cytosine and adenine equally probable in the first sequence context ($\Delta\Delta G_{ca} \approx 0$), confirming the critical role of the repulsion from the imidazole ring in the preference for cytosine. However, for the sequence with the inverted BI/BII ratio, even though $\Delta\Delta G_{ca}$ decreased (to 1.4 kcal/mol) by 0.5, the interaction with cytosine is still clearly preferred (Fig. 5 *B*, *ii*). The likely explanation is that in this sequence context, the propionate–adenine H-bond is disfavored by the GpA step phosphate group that in its BII conformation approaches the N6-amino group of adenine by almost 2 Å compared to the BI conformation (Fig. 5*D*). Indeed, when the conformation of the GpA* step in the 5′-TGA*CAT-3′ sequence is changed to BI by an external restraint (*SI Appendix*, Fig. S16*A*), $\Delta\Delta G_{ca}$ is further reduced from 1.4 to 0.5 kcal/mol, implying that BII contributes 0.9 kcal/mol to the cytosine preference for this particular sequence. Conversely, restraining the CpA* step in the 5′-GTCA*AT-3′ sequence to BII leads to a 0.4-kcal/mol increase in the preference for cytosine, showing that the phosphate conformation effect depends on the sequence context (*SI Appendix*, Fig. S16*B*). More generally, this result indicates that direct base sensing and indirect effects relying on sequence-dependent polymorphism are often interdependent and work hand in hand to fine-tune the affinity for a given site.

Quantum Chemical Calculations Confirm the Origin of Asp/Glu Preference for Cytosine Over Adenine. Since simple force field-based models do not incorporate electronic polarization and can suffer from poor description of H-bonding interactions, we used quantum chemical calculations to validate our findings on the origin of base preferences of acidic residues. To this end, we first calculated the free energy profiles for binding of a single propionate ion (mimicking the side chain of Asp/Glu) to cytosine or adenine in the major groove of B-DNA decamer using hybrid quantum/classical (QM/MM) ab initio molecular dynamics (AIMD) combined with umbrella sampling (*Methods* for details). The QM region, treated using DFT at the TPSS/def2SVP level, included the propionate, the central base pair with C or A, two flanking bases above and below C or A, and several water molecules in between the propionate and DNA (*SI Appendix*, Fig. S19). As a reaction coordinate, *r*, we used the distance between the selected propionate oxygen atom and either the nitrogen N4 in cytosine or N6 in adenine (Fig. 6*A*).

A well-pronounced bound-state ($r < 3.3$ Å) minimum of 2 kcal/mol in the resulting free energy profile for cytosine and only a very shallow minimum for adenine in Fig. 6*A* are clearly consistent with the observed preferential targeting of cytosine by acidic residues. To explain this preference, from the AIMD-generated bound-state ensembles, we extracted simpler subsystems consisting of the propionate bound to either GC or AT base pair and optimized them using the B3LYP/def2TZVP model chemistry (Fig. 6 *B*, *iv*). By computing binding energies using the continuum solvation model for water, we determined

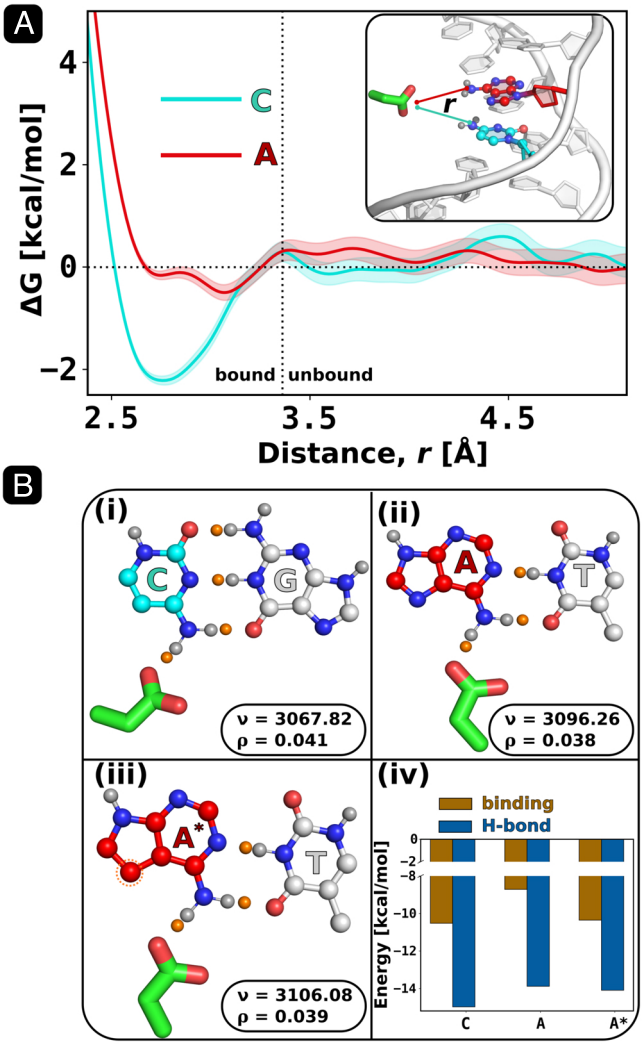

Fig. 6. (*A*) Free energy profiles for the binding of propionic acid to cytosine (C) and adenine (A) in the major groove of a B-DNA decamer, computed using QM/MM ab initio molecular dynamics. (*B*) DFT-optimized structures of the complexes formed by propionic acid with the three base pairs: GC (*i*), AT (*ii*), or A*T (*iii*), where A* is 7-deazaadenine, i.e., adenine in which N7 of the imidazole ring is replaced by a -CH group. Hydrogen bond critical points (CP) are indicated by orange spheres. The electron densities at the CPs (ρ in e/bohr^3) and N–H stretching vibrational frequencies (ν in cm^{-1}) characterizing the H-bonds to the propionate ion are shown in the insets next to the structures (the corresponding vibrational modes are depicted in *SI Appendix*, Fig. S17). (*iv*) Hydrogen bond energies and binding energies (in aqueous solution) for each of the above complexes, calculated at the B3LYP/def2TZVP level. For the dependence of binding energies on dielectric constant, *SI Appendix*, Fig. S18.

that the binding of propionate to cytosine in the GC pair is by 1.8 kcal/mol more energetically favorable than to *A* in the AT pair (Fig. 6 *B*, *Bottom Right*), consistently with our previous findings. In fact, the energetic preference for GC is even more pronounced in low dielectric media, probably more adequate for modeling a desolvated DNA/protein interface (*SI Appendix*, Fig. S18).

However, the calculated difference in the binding energies cannot be fully explained by different strengths of hydrogen bonds formed by the propionate with the cytosine or adenine amino groups. Indeed, as can be seen from Fig. 6*B*, the hydrogen bond energies estimated from the electron densities at the H-bond critical points (48) are ∼15 and 14 kcal/mol for *C* and *A*, respectively, and thus can account for only roughly half of the difference in the binding energies. Since it is well

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

known that the D–H stretching vibration frequency decreases with increasing H-bond strength (49), the same conclusion can be drawn from quite similar red shifts of the N–H stretching modes upon complex formation with the propionate (by 336.6 and 209.9 cm^{-1} for *C* and *A*, respectively; Fig. 6*B*). Notably, this finding supports our prediction (Fig. 5) that repulsion from the imidazole ring is another factor contributing to lowering the affinity for adenine.

To additionally test whether this prediction persists also in the quantum picture, we replaced adenine in the AT base pair by 7-deazaadenine in which a -CH group is substituted for the N7 atom in the imidazole ring. We found that after the modification, the propionate binding energy increased markedly almost reaching the value computed for cytosine. At the same time, hydrogen bond energy remained virtually unaffected with respect to adenine (Fig. 6*B*), confirming that unfavorable interaction with N7 is indeed a major destabilizing factor.

Conclusions

In this work, we examined the role of the acidic residues (Asp and Glu) in sequence-specific DNA–protein interactions, using classical and ab initio molecular dynamics free energy calculations. Statistical analysis of known high-resolution DNA/protein structures reveals that, despite the negative charge of the DNA backbone, acidic residues are commonly found in the DNA major groove where they actively participate in sequence recognition. Specifically, they show a strong preference for hydrogen bonding with the cytosine base, even though adenine also exposes its amino group in the major groove. In fact, the direct readout of cytosine by DNA-binding proteins occurs almost exclusively through Asp or Glu.

By computing the changes in binding free energy of selected transcription factors upon mutation of Asp/Glu to alanine ($\Delta\Delta$G) against a systematic set of DNA sequences, we found that the contribution of the acidic residues to DNA-binding affinity is a result of a fine balance between the electrostatic repulsion from the DNA backbone and specific interactions with nucleobases. In particular, at noncytosine sequences, where the repulsion is not compensated by any major attractive forces, the acidic residues generally disfavor binding ($\Delta\Delta G > 0$), thereby acting as negative selectors whose role is mainly to avoid these (usually off-target) sites. In contrast, at cytosine-containing sequences, the contribution generally varies from negligible ($\Delta\Delta G \approx 0$) to favorable ($\Delta\Delta G < 0$) with the increasing number of cytosines in the immediate vicinity of Asp/Glu. As revealed by our energetic analysis, this cumulative effect, also consistent with a significant excess of cytosine at the DNA sites recognized by acidic residues, relies on the long-range electrostatic attraction to cytosines in the major groove, and not merely on the local H-bond interaction with the amino group. We could therefore conclude that the recognition of cytosine tracts by Asp/Glu is a universal feature of the famously complex protein–DNA recognition code. We also hypothesize that the long-range nature of this interaction might accelerate target search by destabilizing transient binding complexes at off-target sites, thereby providing an evolutionary mechanism to tune binding kinetics as target sequences became increasingly sparse in larger and larger genomes.

Furthermore, our analysis of the model system containing the propionate ion interacting with DNA duplex provides explanation of the strong preference of Asp/Glu for cytosine vs adenine binding. Namely, as indicated by classical MD simulations, forming an H-bond to adenine is disfavored by electrostatic repulsion with the N7 atom of the imidazole ring and, in certain sequence context, by the tendency of purine–purine dinucleotides to adopt BII backbone conformation. In the BII state, the phosphate group approaches the amino group of adenine and renders its interaction with the negatively charged residues less energetically favorable. With quantum chemical calculations, we reproduced the large difference in the affinity of propionate to cytosine and adenine ($\Delta\Delta G_{ca} \approx 2$ kcal/mol) and found that the repulsion from N7 atom accounts for roughly half of the difference in the binding energy. The remaining half results from different strengths of hydrogen bonds, which for cytosine are ∼1 kcal/mol stronger.

While in the nearest future, protein design will most likely be dominated by complex heuristics generated by increasingly sophisticated machine learning models (50, 51), the basic recognition rules we identify and describe here might prove useful for the understanding and rational use of the underlying patterns in the study of evolution, disease, and molecular engineering. We also wish to highlight the importance of negative selection for all processes involving selectivity in molecular recognition, such as rational drug design or knowledge-based engineering of antibodies.

Materials and Methods

Molecular Systems. To understand the energetics of base readout mediated by the acidic residues, we selected from the PDB database four high-resolution structures of B-DNA duplex bound by structurally diverse, sequence-specific transcription factors that contain Asp/Glu residues involved in direct interactions with cytosine in the major groove. Specifically, the selected complexes were 1) basic helix–loop–helix (bHLH) domain of CLOCK (Circadian locomotor output cycles kaput) and BMAL1 (brain and muscle ARNT-like 1) (PDB id: 4H10) (52), 2) Zif268 zinc-finger (PDB id: 1ZAA) (53), 3) DNA-binding domain of Myb (PDB id: 1MSE) (54), and 4) erythroblast transformation–specific domain of ERG3 (PDB id: 5YBD) (55) (*SI Appendix*, Fig. S2). This set was complemented by another sequence-specific binder, i.e., the telomeric protein TRF1 (PDB id: 1W0T (56)), studied in our previous work (43) (*SI Appendix*, Fig. S20). To sample the sequence space, from each of the above (reference) DNA/protein complexes, we created 4 to 5 of their sequence variants by substituting the cytosine directly H-bonded to Asp/Glu to all possible canonical bases, i.e., thymine (C2T), adenine (C2A), and guanine (C2G) as well as by mutating all the nucleobases within the 5 Åof Asp/Glu (all-5 Å) through either transitions (purine-to-purine and pyrimidine-to-pyrimidine substitutions) or transversions (purine-to-pyrimidine and vice versa) (Fig. 2*B*, *SI Appendix*, Table S1 and Fig. S3). All the base substitutions were made using X3DNA package (57). For TRF1, consistently with the previous work (43), we considered one (off-target) variant, i.e., an inverse telomeric sequence (*SI Appendix*, Fig. S20).

To investigate the molecular basis of the strong preference of acidic residues for cytosine over adenine, we used the model system in which a single propionic acid anion interacts with a B-DNA decamer in the major groove. To make cytosine and adenine equally accessible to the propionate in the center of the decamer, we used the following DNA sequence: 5′-C·A·T·G·T·**C**·**A**·A·T·C-3′. To capture how the noncanonical BII backbone conformation of B-DNA can affect this preferential interaction, we also used a second sequence (5′-G·A·T·T·G·**A**·**C**·A·T·G-3′) in which the GpA dinucleotide step, involving the central adenine (A6), has a strong tendency to adopt the BII conformation (45, 47). The DNA decamers were built using the X3DNA package (57). To directly evaluate the effect of adenine N7 atom on the cytosine/adenine preferences, we created an additional variant of both sequences in which the partial charge on N7 of the central adenine was modified from −0.62 to −0.02 (which corresponds to the aromatic -CH group). To keep the modified adenine (denoted as A*) electrically neutral and thus avoid artificial attraction of the propionate, the compensating charge of 0.6 was uniformly distributed over the remaining atoms of A* (*SI Appendix*, Fig. S21 for comparison of the original and modified charges)

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

All the above molecular systems were solvated with TIP3P water molecules (58) in a dodecahedron box with a minimum distance of 1 nm between the solute and the box edges. K^+ and Cl^- ions were added to reach a physiological salt concentration of 0.15 M and neutralize the system. In the case of the Zif268 zinc finger, Zn^{2+} ions present in the crystal structure were preserved and bound to the coordinating residues using the bonding parameters from the zinc amber force field (ZAFF) (59).

Simulation Details. All force field-based molecular dynamics (MD) simulations were carried out in the isothermal–isobaric (NPT) ensemble using Gromacs 2018.8 (60) and the Amber-parmbsc1 force field (61). The temperature was kept constant at 300 K using the v-rescale thermostat (62) with a time constant of 0.1 ps, and the pressure was maintained at 1 bar using the isotropic Parrinello–Rahman barostat (63). Periodic boundary conditions were applied in 3D, and long-range electrostatic interactions were computed using the particle mesh Ewald (PME) method (64) with a real-space cutoff of 1.2 nm and a Fourier grid spacing of 0.12 nm. Van der Waals interactions were described by the Lennard-Jones potential with a cut-off of 1.2 nm and a switching distance of 1 nm. The default Gromacs soft-core potentials were applied to avoid singularity points in all the alchemical free energy simulations. The bond lengths for protein and DNA molecules were constrained by P-LINCS (65), and SETTLE (66) was used to constrain the geometry of water molecules. The leap-frog algorithm was used for the integration of the equations of motion with a time step of 2 fs. Prior to all production simulations, all the systems were equilibrated for at least 100 ns.

Alchemical Binding Free Energy Calculations. The contribution of Asp/Glu to DNA-binding affinity was assessed as the binding free energy difference ($\Delta\Delta G$) between the wild-type protein and its mutant in which a given Asp/Glu was substituted by alanine (*SI Appendix*, Fig. S2 specifies which interfacial Asp/Glu residues were selected for each of the considered proteins). For this purpose, we used a thermodynamic cycle (Fig. 2*A*) that allows for determination of $\Delta\Delta G$ by computing and subtracting the free energies associated with "alchemical" transformation of Asp/Glu to alanine, either in the absence (ΔG_{m1}) or in the presence (ΔG_{m2}) of DNA bound to the protein. The transformation is achieved by simulating the system independently for a set of values of the scaling parameter λ which varies between 0 and 1 to linearly interpolate between the potential energy functions of the physical end states. To speed up the free energy convergence, the neighboring λ-windows were allowed to exchange their configurations every 0.5 ps according to the Metropolis criterion, and the values of λ were optimized to achieve the acceptance rate of at least 10%, using an in-house script (https://gitlab.com/KomBioMol/converge_lambdas; for details see ref. 67). The hybrid topology for the Asp/Glu $\rightarrow$ Ala mutations was generated using the pmx web server (68). All the systems were simulated for at least 300 ns in each a λ-window until a reasonable convergence of $\Delta\Delta G$ was reached (*SI Appendix*, Fig. S22). Since for the C2A variant of the complex involving Zif268 we observed a partial dissociation of the protein from DNA and thus a deviation from the well-defined bound state, to avoid artifacts in $\Delta\Delta G$, we prevented this dissociation by keeping the initial center-of-mass distance between Zif268 and DNA using the harmonic restraint with a force constant of 119.61 kcal/(mol nm^2). For the analyses, we used only the data obtained for the restrained complex, while the structural characteristics of the unrestrained one, labeled $C2A_u$, are also included in *SI Appendix*, Table S2.

Quantum Chemical Calculations. All QM/MM ab initio molecular dynamics (AIMD) simulations of the model system containing a single propionate anion interacting with a fully solvated B-DNA decamer were performed with the NAMD 2.14 molecular dynamics engine (69) interfaced with the Orca 4.2 quantum chemistry program (70). The QM region included the propionate, the central GC or AT base pair, two flanking bases above and below C or A, and four water molecules in between the propionate and DNA (*SI Appendix*, Fig. S19). To saturate the covalent bonds between the QM and MM regions, we used hydrogen link atoms with the default charge distribution scheme (71). The QM region was treated at the DFT level by using the TPSS functional (72) in combination with def2-SVP basis set and the Grimme D3 dispersion correction (73). The resolution-of-identity (RI) approximation for coulomb integrals was used in combination with the def2/J auxiliary basis set. The MM subsystem was described with the Amber99-parmbsc1 force field (61) for DNA and ions and TIP3P for water. The main contribution to the electrostatic interaction between the QM and MM subsystems was modeled through electrostatic embedding, i.e., by passing the MM partial charges surrounding the QM region to Orca, using the default cutoff and charge shifting scheme (71). The remaining (long-range) electrostatic interactions between the MM and QM regions and the MM electrostatics were calculated using the particle mesh Ewald method (PME) (64) with a real space cutoff of 1.2 nm and the QM partial (Mulliken) charges being updated each step. The MM and QM–MM van der Waals interactions were described with the Lennard-Jones potential with a cutoff of 1.2 nm. The simulations were performed in NPT ensemble with the temperature maintained at 310 K by Langevin dynamics with a damping coefficient of 50 ps^{-1} and pressure maintained at 1 bar with the Langevin piston method with an oscillation period of 0.2 ps and a damping time scale of 0.1 ps (74). The velocity Verlet algorithm was used to integrate equations of motion with a time step of 0.5 fs.

The free energy profiles for the hydrogen bond formation between the propionate anion and cytosine (C) or adenine (A) in the context of a B-DNA decamer were obtained using AIMD-based umbrella sampling (US) simulations. The reaction coordinate, *r*, was defined as the distance between the proton donor and acceptor, i.e., the nitrogen N4 in C or N6 in A and one of the oxygen atoms in the propionate. The systems were restrained along the reaction coordinate in 27 independent US "windows" separated by 0.01 nm using the harmonic potential with a spring constant of 500 kcal/(mol nm^2), thus spanning the 0.24 to 0.50 nm range of the reaction coordinate. To produce the initial frames for the US simulations, we extracted the representative bound-state configurations from our classical MD trajectories and performed enforced dissociation AIMD simulations during which *r* was gradually increased to 0.50 nm over 5 ps, by applying a moving harmonic potential with a spring constant of 1,000 kcal/(mol nm^2). In each window, the systems were simulated for at least 30 ps, and the free energy profiles were determined from the last 15 ps by using the standard weighted histogram analysis method (WHAM) (75). Uncertainties in the free energy estimates were obtained using a bootstrap approach taking into account autocorrelation of the *r* time series (76).

Quantum chemical calculations of simpler systems composed of a single propionate molecule bound to the GC, AT, or A*T base pair, where A* is 7-deazaadenine, were performed using Gaussian 16 (77) at the B3LYP/def2TZVP level of theory with the D3 empirical dispersion correction (78), in the IEFPCM model. The structures were optimized, and the stability of the obtained geometries was confirmed by vibrational frequency analysis. Electron densities at the hydrogen bond critical points characterizing their strength were carried out by DAMQT (79). Hydrogen bond energies were obtained from the critical point densities using the empirical relation proposed by Emamian et al. (48). The binding energies of the propionate anion to the considered base pairs were calculated by subtracting the energy of the optimized constituents (propionate and a given base pair) from the energy of the optimized complex.

Data, Materials, and Software Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. This research was supported in part by PL-Grid Infrastructure. Computational resources were provided also by the TASK (Gdansk), WCSS (Wroclaw), and ICM (Warsaw) Centers.

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.

1. P. M. Dehé, P. H. L. Gaillard, Control of structure-specific endonucleases to maintain genome stability. *Nat. Rev. Mol. Cell Biol.* **18**, 315–330 (2017).
2. C. E. Ang, M. Wernig, Profiling DNA-transcription factor interactions. *Nat. Biotechnol.* **36**, 501–502 (2018).
3. J. Hörberg, K. Moreau, M. J. Tamás, A. Reymer, Sequence-specific dynamics of DNA response elements and their flanking sites regulate the recognition by AP-1 transcription factors. *Nucleic Acids Res.* **49**, 9280–9293 (2021).
4. S. Inukai, K. H. Kock, M. L. Bulyk, Transcription factor-DNA binding: Beyond binding site motifs. *Curr. Opin. Genet. Dev.* **43**, 110–119 (2017).
5. J. F. Kribelbauer, C. Rastogi, H. J. Bussemaker, R. S. Mann, Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Ann. Rev. Cell Dev. Biol.* **35**, 357 (2019).
6. A. Scipioni, C. Anselmi, G. Zuccheri, B. Samori, P. De Santis, Sequence-dependent DNA curvature and flexibility from scanning force microscopy images. *Biophys. J.* **83**, 2408–2418 (2002).
7. A. Perez, F. Lankas, F. J. Luque, M. Orozco, Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.* **36**, 2379–2394 (2008).
8. M. Fuxreiter, I. Simon, S. Bondos, Dynamic protein-DNA recognition: Beyond what can be seen. *Trends Biochem. Sci.* **36**, 415–423 (2011).
9. M. A. Öztürk, G. V. Pachov, R. C. Wade, V. Cojocaru, Conformational selection and dynamic adaptation upon linker histone binding to the nucleosome. *Nucleic Acids Res.* **44**, 6599–6613 (2016).
10. A. Balaceanu *et al*., Modulation of the helical properties of DNA: Next-to-nearest neighbour effects and beyond. *Nucleic Acids Res.* **47**, 4418–4430 (2019).
11. A. K. Jaiswal, A. Krishnamachari, Physicochemical property based computational scheme for classifying DNA sequence elements of *Saccharomyces cerevisiae*. *Comput. Biol. Chem.* **79**, 193–201 (2019).
12. G. B. Koudelka, P. Carlson, DNA twisting and the effects of non-contacted bases on affinity of 434 operator for 434 repressor. *Nature* **355**, 89–91 (1992).
13. A. C. Cheng, W. W. Chen, C. N. Fuhrmann, A. D. Frankel, Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J. Mol. Biol.* **327**, 781–796 (2003).
14. C. G. Kalodimos *et al*., Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **305**, 386–389 (2004).
15. A. Sarai, H. Kono, Protein-DNA recognition patterns and predictions. *Ann. Rev. Biophys. Biomol. Struct.* **34**, 379 (2005).
16. S. Ahmad, O. Keskin, A. Sarai, R. Nussinov, Protein-DNA interactions: Structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.* **36**, 5922–5932 (2008).
17. R. Schleif, DNA binding by proteins. *Science* **241**, 1182–1187 (1988).
18. C. Escudé, J. S. Sun, DNA major groove binders: Triple helix-forming oligonucleotides, triple helix-specific DNA ligands and cleaving agents. *DNA Binders Relat. Subj.* **253**, 109–148 (2005).
19. S. Poddar, D. Chakravarty, P. Chakrabarti, Structural changes in DNA-binding proteins on complexation. *Nucleic Acids Res.* **46**, 3298–3308 (2018).
20. A. Hospital *et al*., Naflex: A web server for the study of nucleic acid flexibility. *Nucleic Acids Res.* **41**, W47–W55 (2013).
21. M. Slattery *et al*., Absence of a simple code: How transcription factors read the genome. *Trends Biochem. Sci.* **39**, 381–399 (2014).
22. G. Da Rosa *et al*., Sequence-dependent structural properties of B-DNA: What have we learned in 40 years? *Biophys. Rev.* **13**, 1–11 (2021).
23. A. Marin-Gonzalez, J. Vilhena, R. Perez, F. Moreno-Herrero, A molecular view of DNA flexibility. *Q. Rev. Biophys.* **54**, 1–20 (2021).
24. S. Chen *et al*., Indirect readout of DNA sequence at the primary-kink site in the cap-DNA complex: Alteration of DNA binding specificity through alteration of DNA kinking. *J. Mol. Biol.* **314**, 75–82 (2001).
25. J. S. Lamoureux, J. T. Maynes, J. M. Glover, Recognition of 5-YpG-3 sequences by coupled stacking/hydrogen bonding interactions with amino acid residues. *J. Mol. Biol.* **335**, 399–408 (2004).
26. T. E. Cheatham III, Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.* **14**, 360–367 (2004).
27. R. Rohs *et al*., Origins of specificity in protein-DNA recognition. *Ann. Rev. Biochem.* **79**, 233 (2010).
28. L. A. Harris, D. Watkins, L. D. Williams, G. B. Koudelka, Indirect readout of DNA sequence by P22 repressor: Roles of DNA and protein functional groups in modulating DNA conformation. *J. Mol. Biol.* **425**, 133–143 (2013).
29. D. Bosch, M. Campillo, L. Pardo, Binding of proteins to the minor groove of DNA: What are the structural and energetic determinants for kinking a basepair step? *J. Comput. Chem.* **24**, 682–691 (2003).
30. G. B. Koudelka, S. A. Mauro, M. Ciubotaru, Indirect readout of DNA sequence by proteins: The roles of DNA sequence-dependent intrinsic and extrinsic forces. *Prog. Nucleic Acid Res. Mol. Biol.* **81**, 143–177 (2006).
31. L. Etheve, J. Martin, R. Lavery, Decomposing protein-DNA binding and recognition using simplified protein models. *Nucleic Acids Res.* **45**, 10270–10283 (2017).
32. H. Y. Alniss, Thermodynamics of DNA minor groove binders: Perspective. *J. Med. Chem.* **62**, 385–402 (2018).
33. F. Battistini *et al*., How B-DNA dynamics decipher sequence-selective protein recognition. *J. Mol. Biol.* **431**, 3845–3859 (2019).
34. S. Fujii, H. Kono, S. Takenaka, N. Go, A. Sarai, Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.* **35**, 6063–6074 (2007).
35. D. Watkins, C. Hsiao, K. K. Woods, G. B. Koudelka, L. D. Williams, P22 c2 repressor- operator complex: Mechanisms of direct and indirect readout. *Biochemistry* **47**, 2325–2338 (2008).
36. T. G. Uil, H. J. Haisma, M. G. Rots, Therapeutic modulation of endogenous gene function by agents with designed DNA-sequence specificities. *Nucleic Acids Res.* **31**, 6064–6078 (2003).
37. M. Ehrenberg, J. Elf, E. Aurell, R. Sandberg, J. Tegnér, Systems biology is taking off. *Genome Res.* **13**, 2377–2380 (2003).
38. S. Wickramaratne, S. Mukherjee, P. W. Villalta, O. D. Scharer, N. Y. Tretyakova, Synthesis of sequence-specific DNA-protein conjugates via a reductive amination strategy. *Bioconjugate Chem.* **24**, 1496–1506 (2013).
39. A. Currin *et al*., Highly multiplexed, fast and accurate nanopore sequencing for verification of synthetic DNA constructs and sequence libraries. *Synth. Biol.* **4**, ysz025 (2019).
40. Y. Mandel-Gutfreund, H. Margalit, Quantitative parameters for amino acid-base interaction: Implications for prediction of protein-DNA binding sites. *Nucleic Acids Res.* **26**, 2306–2312 (1998).
41. L. Etheve, J. Martin, R. Lavery, Dynamics and recognition within a protein-DNA complex: A molecular dynamics study of the SKN-1/DNA interaction. *Nucleic Acids Res.* **44**, 1440–1448 (2016).
42. Q. Liao *et al*., Long time-scale atomistic simulations of the structure and dynamics of transcription factor-DNA recognition. *J. Phys. Chem. B* **123**, 3576–3590 (2019).
43. M. Wieczór, J. Czub, How proteins bind to DNA: Target discrimination and dynamic sequence search by the telomeric protein TRF1. *Nucleic Acids Res.* **45**, 7643–7654 (2017).
44. J. M. Sagendorf, H. M. Berman, R. Rohs, DNAproDB: An interactive tool for structural analysis of DNA-protein complexes. *Nucleic Acids Res.* **45**, W89–W97 (2017).
45. A. Balaceanu *et al*., The role of unconventional hydrogen bonds in determining BII propensities in B-DNA. *J. Phys. Chem. Lett.* **8**, 21–28 (2017).
46. C. Tisné, M. Delepierre, B. Hartmann, How NF-κb can be attracted by its cognate DNA. *J. Mol. Biol.* **293**, 139–150 (1999).
47. S. Derreumaux, M. Chaoui, G. Tevanian, S. Fermandjian, Impact of CpG methylation on structure, dynamics and solvation of camp DNA responsive element. *Nucleic Acids Res.* **29**, 2314–2326 (2001).
48. S. Emamian, T. Lu, H. Kruse, H. Emamian, Exploring nature and predicting strength of hydrogen bonds: A correlation analysis between atoms-in-molecules descriptors, binding energies, and energy components of symmetry-adapted perturbation theory. *J. Comput. Chem.* **40**, 2868–2881 (2019).
49. E. Arunan *et al*., Definition of the hydrogen bond (IUPAC recommendations 2011). *Pure Appl. Chem.* **83**, 1637–1641 (2011).
50. M. J. Volk *et al*., Biosystems design by machine learning. *ACS Synth. Biol.* **9**, 1514–1533 (2020).
51. J. Jumper *et al*., Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
52. Z. Wang, Y. Wu, L. Li, X. D. Su, Intermolecular recognition revealed by the complex structure of human clock-bmal1 basic helix-loop-helix domains with e-box DNA. *Cell Res.* **23**, 213–224 (2013).
53. N. P. Pavletich, C. O. Pabo, Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809–817 (1991).
54. K. Ogata *et al*., Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **79**, 639–648 (1994).
55. R. Sharma, S. P. Gangwar, A. K. Saxena, Comparative structure analysis of the ETSI domain of ERG3 and its complex with the E74 promoter DNA sequence. *Acta Crystallogr. Sec. F: Struc. Biol. Commun.* **74**, 656–663 (2018).
56. R. Court, L. Chapman, L. Fairall, D. Rhodes, How the human telomeric proteins TRF1 and TRF2 recognize telomeric DNA: A view from high-resolution crystal structures. *EMBO Rep.* **6**, 39–45 (2005).
57. X. J. Lu, W. K. Olson, 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108–5121 (2003).
58. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
59. M. B. Peters *et al*., Structural survey of zinc-containing proteins and development of the zinc amber force field (ZAFF). *J. Chem. Theory Comput.* **6**, 2935–2947 (2010).
60. M. J. Abraham *et al*., Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
61. I. Ivani *et al*., Parmbsc1: A refined force field for DNA simulations. *Nat. Methods* **13**, 55–58 (2016).
62. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
63. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
64. D. M. York, T. A. Darden, L. G. Pedersen, The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the ewald and truncated list methods. *J. Chem. Phys.* **99**, 8345–8348 (1993).
65. B. Hess, P-LINCS: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* **4**, 116–122 (2008).
66. S. Miyamoto, P. A. Kollman, Settle: An analytical version of the shake and rattle algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).
67. M. Wieczór, J. Czub, Telomere uncapping by common oxidative guanine lesions: Insights from atomistic models. *Free Radical Biol. Med.* **148**, 162–169 (2020).
68. V. Gapsys, B. L. de Groot, pmx webserver: A user friendly interface for alchemistry. *J. Chem. Inf. Model.* **57**, 109–114 (2017).
69. J. C. Phillips *et al*., Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).
70. F. Neese, Software update: The ORCA program system, version 4.0. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **8**, e1327 (2018).
71. M. C. Melo *et al*., NAMD goes quantum: An integrative suite for hybrid simulations. *Nat. Methods* **15**, 351–354 (2018).
72. J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, J. Sun, Workhorse semilocal density functional for condensed matter physics and quantum chemistry. *Phys. Rev. Lett.* **103**, 026403 (2009).
73. S. Grimme, Density functional theory with London dispersion corrections. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **1**, 211–228 (2011).
74. S. E. Feller, Y. Zhang, R. W. Pastor, B. R. Brooks, Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* **103**, 4613–4621 (1995).
75. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* **16**, 1339–1350 (1995).
76. J. S. Hub, B. L. De Groot, D. Van Der Spoel, g_wham, a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.* **6**, 3713–3720 (2010).
77. Me. Frisch *et al*., Gaussian 16 (2016).
78. S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).
79. A. Kumar *et al*., DAMQT 2.1. 0: A new version of the DAMQT package enabled with the topographical analysis of electron density and electrostatic potential in molecules (2015).

Downloaded from https://www.pnas.org by 89.69.222.66 on January 12, 2023 from IP address 89.69.222.66.