JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2017, 13, 6275−6289

pubs.acs.org/JCTC

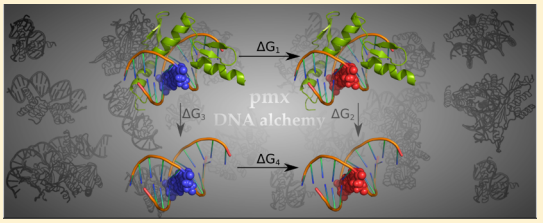
Alchemical Free Energy Calculations for Nucleotide Mutations in Protein−DNA Complexes

Vytautas Gapsys* and Bert L. de Groot*

Computational Biomolecular Dynamics Group, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

Supporting Information

ABSTRACT: Nucleotide-sequence-dependent interactions between proteins and DNA are responsible for a wide range of gene regulatory functions. Accurate and generalizable methods to evaluate the strength of protein−DNA binding have long been sought. While numerous computational approaches have been developed, most of them require fitting parameters to experimental data to a certain degree, e.g., machine learning algorithms or knowledge-based statistical potentials. Molecular-dynamics-based free energy calculations offer a robust, system-independent, first-principles-based method to calculate free energy differences upon nucleotide mutation. We present an automated procedure to set up alchemical MD-based calculations to evaluate free energy changes occurring as the result of a nucleotide mutation in DNA. We used these methods to perform a large-scale mutation scan comprising 397 nucleotide mutation cases in 16 protein−DNA complexes. The obtained prediction accuracy reaches 5.6 kJ/mol average unsigned deviation from experiment with a correlation coefficient of 0.57 with respect to the experimentally measured free energies. Overall, the first-principles-based approach performed on par with the molecular modeling approaches Rosetta and FoldX. Subsequently, we utilized the MD-based free energy calculations to construct protein−DNA binding profiles for the zinc finger protein Zif268. The calculation results compare remarkably well with the experimentally determined binding profiles. The software automating the structure and topology setup for alchemical calculations is a part of the pmx package; the utilities have also been made available online at http://pmx.mpibpc.mpg.de/dna_webserver.html.



1. INTRODUCTION

The ability of proteins to recognize and interact with DNA is vital for a multitude of cellular processes. Proteins have evolved to identify specific regions in DNA on the basis of the nucleic base sequence and shape of the molecule.[1] The interaction landscape is further complicated by the inherent flexibility of the DNA helix as well as its binding partner.[2] A plethora of experimental and computational approaches are available for the structural and thermodynamic characterization of protein−DNA complexes.

Over the past decade a number of large-scale studies have significantly expanded the knowledge of the DNA binding sites recognized by transcription factors (TFs).[3−5] In these high-throughput SELEX[6] and protein binding microarray (PBM)[7] experiments, binding profiles for hundreds of transcription factors were determined (Weirauch et al.[5] examined more than 1000 TFs). While such studies provide invaluable insight into the protein−DNA interaction specificity, naturally the experimental investigations are also labor- and resource-demanding. In parallel to the breakthroughs in the experimental techniques, computational approaches have evolved to predict protein−DNA binding specificity with increasing accuracy.[8]

Numerous machine learning algorithms have been trained on the basis of amino acid sequences or simple contact models to predict the binding profiles for specific protein motifs, e.g., zinc fingers.[9−11] A more general approach of constructing knowledge-based statistical potentials allows the generation of binding profiles for a broad range of proteins.[12,13] The physical/statistical potentials Rosetta and FoldX have also been successfully employed to calculate the free energy differences of nucleotide mutations and subsequently determine full binding profiles for the protein−DNA complexes.[14−16] Another class of approaches comprises the molecular dynamics (MD)-based calculations that rely on first principles of statistical mechanics. These approaches present a robust yet computationally more expensive access to the free energy differences upon nucleic acid mutations.

In recent years, alchemical approaches for ligand modifications[17] and amino acid mutations[18] have been shown to yield accurate results in large-scale free energy calculations. Furthermore, the previously technically demanding setup for simulations of this type has been automated, making the approaches widely applicable.[19−21] Nucleic acid mutations by means of MD-based free energy calculations, however, have received less attention. Historically, a number of small-scale studies have been performed, concentrating on one or a few systems of interest and a handful of mutations. The early studies exploring the suitability of free energy perturbation methods for biomolecular applications calculated solvation free

Received: August 10, 2017
Published: November 10, 2017

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

energies of nucleic acid bases.[22] Subsequently, the field advanced to the successful application of alchemical MD-based approaches to estimate nucleotide-mutation-induced changes in ligand−DNA interactions[23] and the stability of the DNA helix[24] and protein−DNA complexes.[25,26] A recent study made an attempt at a nucleotide mutation scan in four protein−DNA complexes by means of nonequilibrium thermodynamic integration.[27] In their investigation, however, the authors observed a drastic difference between the computed and experimentally obtained free energy differences, concluding that the force field parametrization and insufficient sampling were the causes of the poor predictive power.

In the current work, we aimed to push the limits of accuracy and scale that are possible to achieve by means of first-principles-based free energy calculations. For that purpose, we evaluated free energy differences in protein−DNA binding due to nucleic acid mutations in 16 protein−DNA complexes. Overall, 397 mutation cases were analyzed and compared to the experimental measurements, reaching an average unsigned error of 5.6 kJ/mol. We further used our computational methods to construct the consensus binding profile of a Zif268 protein and subsequently compared the computed results to the experimentally obtained profiles. Our findings demonstrate that the MD-based calculations perform on par with the well-established modeling approaches Rosetta and FoldX as well as with the machine learning techniques trained against the specific protein targets.

The utilities for the automated hybrid structure and topology generation used in this work are freely available and constitute a part of a more general pmx library dedicated to the free energy calculation setup. In addition, online generation of the hybrid structures/topologies is made available via the pmx Web server.

2. METHODS

2.1. Hybrid Stuctures/Topologies. High-throughput MD-based alchemical free energy calculations require automation of hybrid structure/topology generation. A successful single topology approach to construct hybrid nucleotides was demonstrated by Seeliger et al.[26] In the current work, we generalized this approach to be applicable to a number of molecular mechanics force fields in a similar way as has previously been done for the amino acid mutations in proteins.[19]

The DNA nucleotide mutation support was incorporated into the already established pmx workflow:[19] a simplified scheme is depicted in Figure 1. Hybrid structure/topology generation for DNA nucleotides requires mutation libraries to be pregenerated for every supported force field. The mutation libraries contain the mapping information necessary for morphing one nucleotide to any other nucleotide. We considered only the nonmodified nucleotides: adenine, thymine, cytosine, and guanine. In the current work, we created libraries for the Amber99sb*ILDN-BSC1[28−31] and Charmm36m[32,33] force fields. Generation of the new nucleotide mutation libraries was enabled via the pmx utilities.

Mapping between the nucleic base pairs follows the strategy introduced by Seeliger et al.[26] For purine-to-purine and pyrimidine-to-pyrimidine mutations, maximum common substructure atom pairing is used, effectively minimizing the perturbation needed to morph one nucleotide to another. For the purine-to-pyrimidine and pyrimidine-to-purine mutations, the whole nucleic base is created/annihilated using dummy atoms, i.e., atoms without the nonbonded interaction

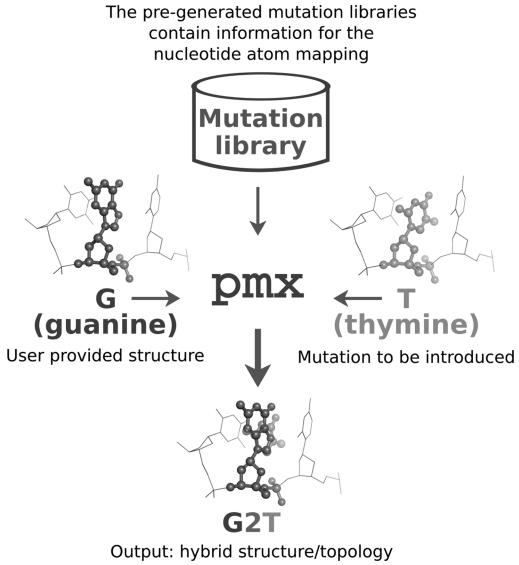


Figure 1. Schematic representation of the pmx-based nucleotide mutation procedure. The user needs to provide an input structure and select a mutation to be introduced. The pmx software utilizes the pregenerated mutation libraries to generate hybrid structures and topologies for the subsequent molecular dynamics simulations.

parameters in their inactive state. In addition, for these cases several dihedral terms are introduced to force the dummy atoms to follow the rotations of the nucleic base in the physical state, i.e., the state that is coupled to the environment. The number of dummy atoms introduced for the purine/pyrimidine mutations ranged from 26 to 29 depending on the mutation, while purine/purine (adenine/guanine) required only seven dummy atoms and pyrimidine/pyrimidine (thymine/cytosine) needed eight dummy atoms.

Provided the mutation libraries and a DNA input structure, pmx generates a structure containing all of the atoms required to represent both physical states of a hybrid nucleotide, e.g., guanine and thymine in Figure 1. In the next step, the Gromacs[34] native topology generation tool pdb2gmx together with the pmx script allow the proper topology of the hybrid construct to be obtained. The generated files are compatible with the current Gromacs topology organization (version 4.5 and higher).

2.2. pmx DNA Web Server. To facilitate the usability of the pmx-based hybrid structure/topology generation workflow, in addition to the pmx command line tools, a Web server was created. The DNA nucleotide handling extends the previously described web-based amino acid mutation infrastructure.[20] The back end of the Web server implements the hybrid structure/topology creation procedure described above. The user is allowed to interactively select the force field to be used and the mutations to perform. Mutations can be introduced in one or both complementary DNA chains. An additional functionality enables a scan to be performed with all of the nucleotide mutations possible over a DNA structure provided by the user. This feature is particularly useful for protein−DNA binding profile calculations.

2.3. Free Energy Calculations. For all of the MD-based free energy calculations carried out in this work, a non-equilibrium setup was used.[35] The initial equilibrations were performed by running a 1 ns simulation with position restraints on all heavy atoms starting from an energy-minimized structure.

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

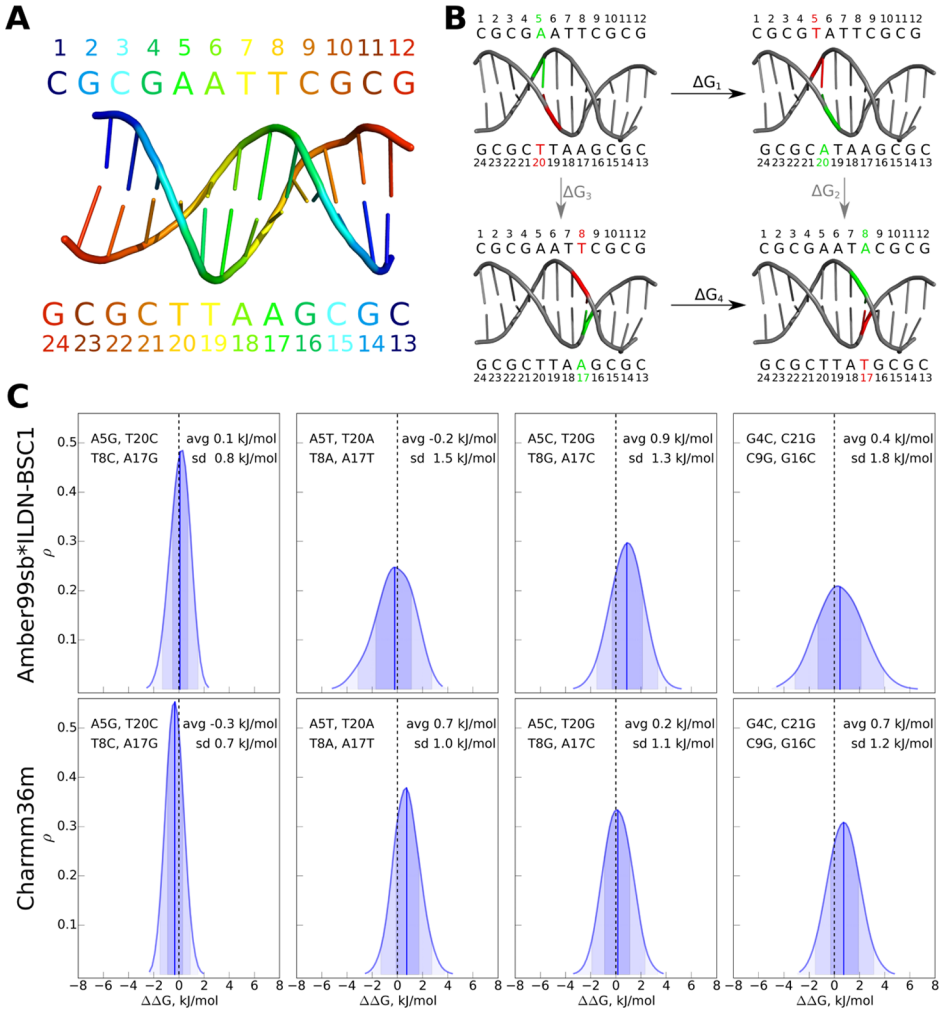


Figure 2. Validation of the free energy calculation workflow. (A) The palindromic Drew−Dickerson dodecamer was used as a model system for the validation simulations. (B) Example of the thermodynamic cycle utilizing the palindromic nature of the Drew−Dickerson dodecamer: $\Delta\Delta G = \Delta G_2 - \Delta G_3 = \Delta G_4 - \Delta G_1 = 0$ kJ/mol. (C) Results of calculations over a number of thermodynamic cycles with an expected outcome of 0 kJ/mol (see (B)). The distributions were constructed from the results of a number of independent simulations (see Methods for details).

Afterward followed a 5 ns nonrestrained simulation to further equilibrate the system. The 20 ns production runs were performed for the wild-type DNA sequence and its every mutation under consideration. From the equilibrium production simulations, 100 snapshots were extracted equidistantly in time, and for every configuration a hybrid structure/topology for the mutant was generated. Subsequently, a brief 20 ps equilibrium simulation for every snapshot was performed to equilibrate the velocities. Finally, for every configuration a 100 ps alchemical transition was started to morph the system from one physical state to another. For the double nucleotide scan of the zinc finger protein Zif268, where two mutations in each strand were performed at once, alchemical transitions of 200 ps were used. In these cases, the perturbation of the system was larger, as four nucleotides (two pairs) were modified at once; thus, the slower transitions were intended to reduce work dissipation along the alchemical path, leading to increased overlap between forward and backward work distributions, hence facilitating convergence of the free energy estimates. The transitions were performed in both directions: wild type to mutant and mutant to wild type. The work values from the nonequilibrium transitions were used to calculate free energy differences based on the Crooks fluctuation theorem[36] utilizing the maximum likelihood estimator.[37]

2.4. Validation: Closed Thermodynamic Cycle. To validate the constructed mutation libraries and the free energy calculation workflow, we constructed a thermodynamic cycle using a double-helix DNA molecule, the Drew−Dickerson dodecamer[38] (Figure 2A). The palindromic nature of the sequence of this DNA fragment allows the generation of a cycle where the free energy changes along the vertical branches are both 0 kJ/mol, as illustrated in Figure 2B. In this situation, the double free energy difference is $\Delta\Delta G = \Delta G_2 - \Delta G_3 = \Delta G_4 - \Delta G_1 = 0$ kJ/mol. We performed calculations of the double free energy differences to cover every combination of the nucleotide mutations: A2G, A2T, A2C, T2C, T2G, and C2G. In total, eight simulation setups enabled the calculation of four $\Delta\Delta G$ values that probed all of the mutation combinations (Table 1).

For every mutation set (Table 1), the hybrid structures/topologies were incorporated into the system. Ten independent equilibrium simulations of 25 ns were performed by setting the system in physical state A. Also, 10 equilibrium simulations were carried out by setting the system in physical state B. By performing transitions from state A to state B, we obtained 10 distributions of the nonequilibrium work values. Running the

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

Table 1. Mutations for the Workflow Validation Simulations

ΔG_1	ΔG_4	mutations probed
A5G, T20C	T8C, A17G	A2G, T2C
A5T, T20A	T8A, A17T	A2T
A5C, T20G	T8G, A17C	A2C, T2G
G4C, C21G	C9G, G16C	C2G

transitions in the direction B to A, another 10 distributions for the work values of the reverse process were calculated. Combinations of the forward and backward work distributions allowed 100 ΔG estimates to be obtained for one branch of the cycle. In total, combining estimates from two branches of the cycle, we were able to estimate $\Delta\Delta G$ 10^4 times, which in turn provided a distribution of the double free energy differences for every cycle considered (rows in Table 1).

All of the validation simulations and the calculations described further in this article were performed using two force fields: Amber99sb*ILDN-BSC1 and Charmm36m.

2.5. Nucleotide Mutations in Protein−DNA Complexes. The set of protein−DNA complexes used in the study comprised 16 systems assembled by Morozov et al.[14] (Figure 3). All of the experimental values converted to free energy differences and expressed in kilojoules per mole together with the respective references are collected in Table S1 in the Supporting Information. We made the following changes to the protein−DNA set of Morozov et al.: The mutation number for the Tus-*Ter* complex was extended to 38 cases, and experimental values from additional literature sources were incorporated for the AtERF1 transcription factor. For the c-Myb protein−DNA complex, an NMR structure was used[39] in which two homologous protein fragments (R2 and R3) are bound to the double-stranded DNA. It has been demonstrated that the R1 c-Myb fragment has only a small influence on the protein−DNA binding,[40] thus justifying the use of the experimental measurements on binding of the R1, R2, and R3 fragments to DNA by Morozov et al. However, where possible we updated the values for c-Myb with those for the R2 and R3 fragments interacting with DNA to retain consistency with the NMR structure used for simulations. For three systems (Zif268, Tus-*Ter*, and CAP proteins interacting with DNA), some experimental values were not determined exactly, but rather, only a lower limit for the complex destabilization upon mutation was measured. To compare our calculations to such cases, we also imposed equivalent constraints on the computed results: if a calculated $\Delta\Delta G$ exceeded the experimental lower bound, the calculated value was set to be equal to this lower-bound value.

The thermodynamic cycle depicted in Figure 4A was constructed to calculate the changes in free energy differences ($\Delta\Delta G$) in the protein−DNA complexes upon a nucleotide mutation. The equilibrium simulations for the unbound DNA

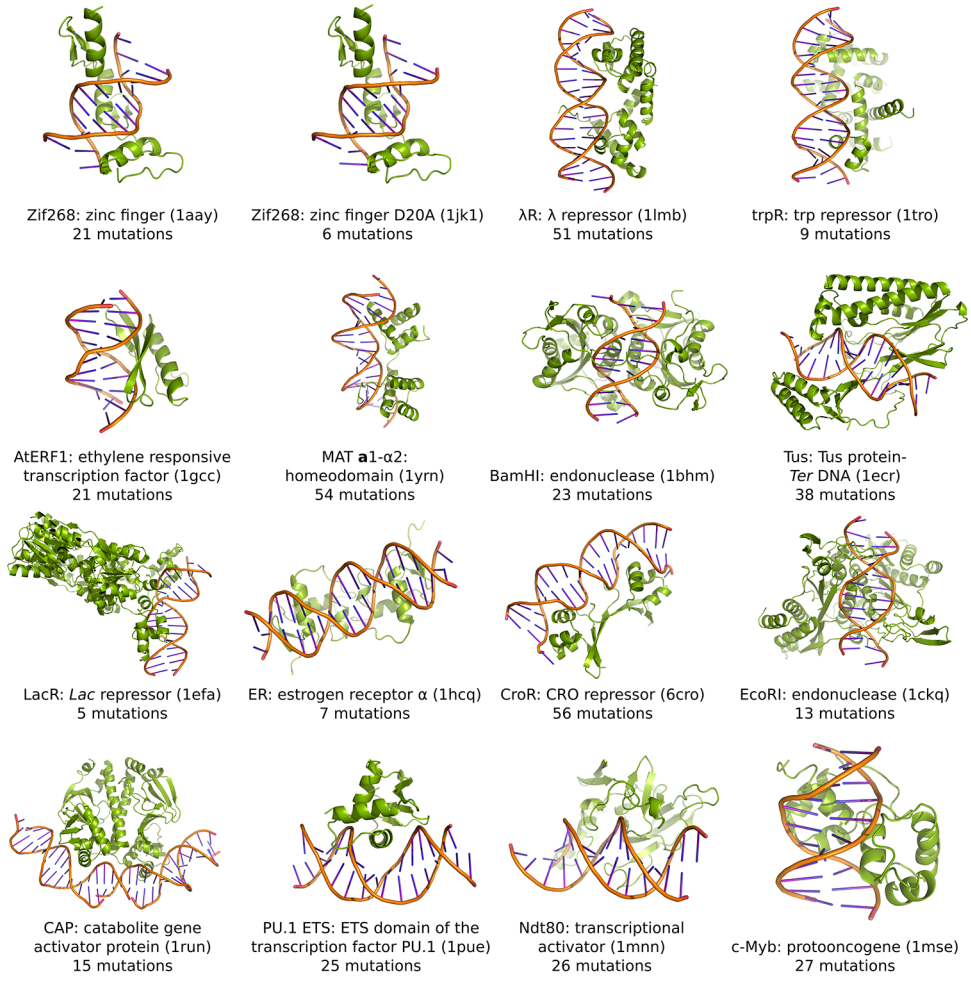


Figure 3. Systems simulated in this study. In total, 397 mutation cases in 16 protein−DNA complexes were investigated.

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

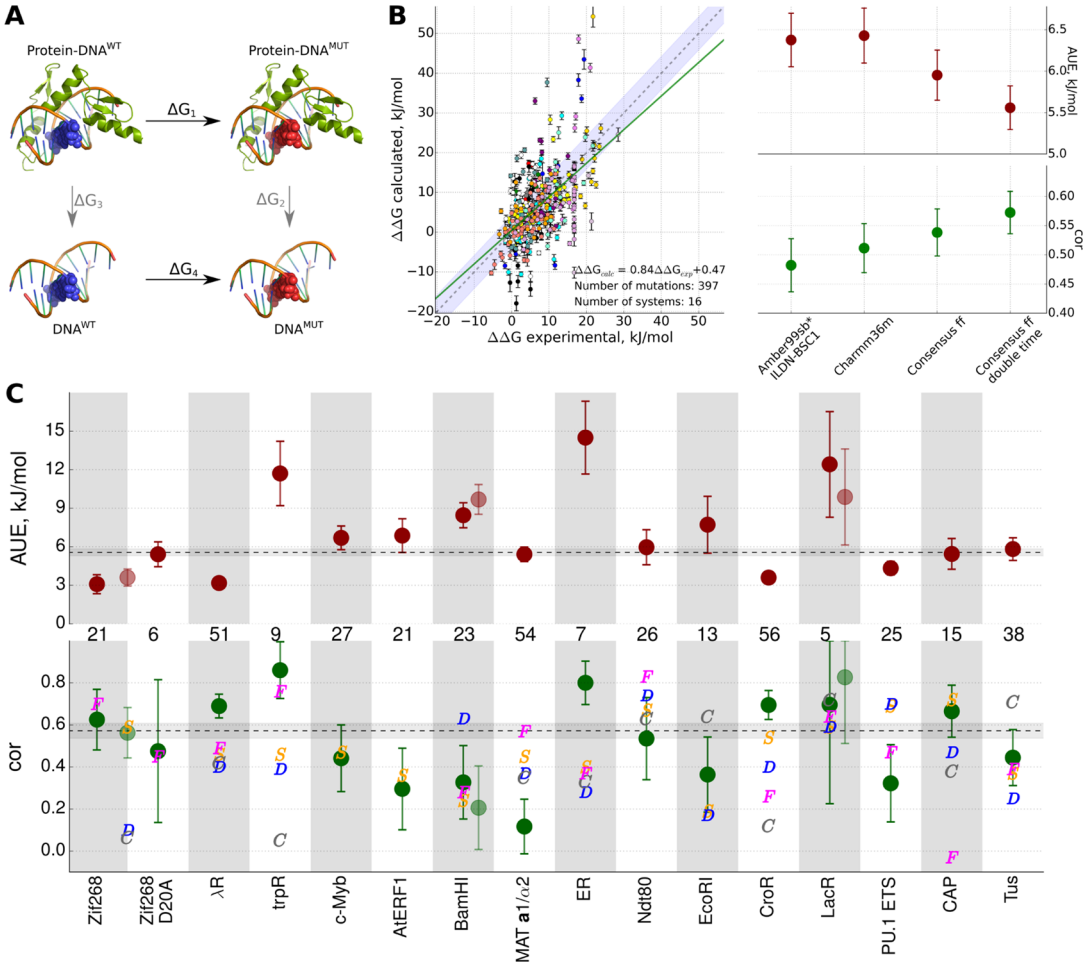


Figure 4. Results of the free energy calculations of the nucleotide mutations in the protein−DNA complexes. (A) Thermodynamic cycle used to calculate the change in the protein−DNA binding free energy upon nucleotide mutation: $\Delta\Delta G = \Delta G_2 - \Delta G_3 = \Delta G_4 - \Delta G_1$. (B) Results from the 16 systems pooled together. The left panel shows the experimental $\Delta\Delta G$ values vs the consensus force field (with the doubled simulation time) calculation results. Data points are colored according to the protein−DNA complex to which they belong. The top-right panel depicts the average unsigned error (AUE) of the calculations with respect to the experimental results, while the bottom-right panel shows the correlation coefficients between the calculations and experiments. (C) AUEs and correlation coefficients between the calculated and experimental $\Delta\Delta G$ values are shown for individual protein−DNA complexes. The calculation results come from the consensus force field with the doubled simulation time analysis. The letters in the correlation plot (bottom panel) denote the dynamic Rosetta model (*D*), static Rosetta model (*S*), contact model (*C*),[14] and FoldX model (*F*).[16] The numbers in between the top and bottom panels are the numbers of mutation cases for the protein−DNA complexes.

and protein−DNA complexes were performed without constructing the hybrid structures/topologies. The hybrids were constructed directly onto the extracted frames from the equilibrium trajectories. Prior to the start of the alchemical nonequilbrium transitions, the systems were simulated for 20 ps to equilibrate the velocities on the introduced dummy atoms.

Several protein−DNA complexes contained ligands that needed separate parametrization. For the simulations with the Amber99SB*ILDN-BSC1 force field, the bonded parameters were assigned from the general Amber force field.[41] The partial charges were obtained by the restrained fit to the electrostatic potential[42] calculated with Gaussian 09[43] at the Hartree−Fock/6-31G* level of theory. Ligand topologies for the Charmm36m force field were obtained using an automated procedure[44,45] to assign the general Charmm force field[46] parameters.

Free energy calculations were carried out in the two aforementioned force fields. Subsequently, we also employed a consensus approach[18] to reduce the force-field-imposed bias by averaging the double free energy differences obtained from the Amber99sb*ILDN-BSC1 and Charmm36m simulations. Averaging the $\Delta\Delta G$ values effectively doubles the simulation time used for the free energy calculation, thus increasing the sampling as well as reducing the force-field-related artifacts. To obtain a better understanding of the benefits coming from only combining the force fields (and not extending the sampling time), we also calculated the consensus free energies using only half of the sampling from each force field.

The calculated results were compared to the experimental measurements by means of the average unsigned error (AUE) and Pearson correlation coefficient. The standard errors for these estimates were obtained by bootstrapping.

2.6. Increasing the Sampling Time. Five systems were used to assess the influence of an increased sampling time on the free energy calculation accuracy: LacR, *Bam*HI, c-Myb, PU.1 ETS, and MAT **a**1/*α*2. For each of the cases, the equilibrium sampling time was increased 5-fold, i.e., five independent simulations, 25 ns each, were carried out for the wild-type and mutant DNA free in solution and bound to the protein. From every equilibrium simulation, 40 snapshots were extracted equidistantly in time for the subsequent non-

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

equilibrium alchemical transitions. The rest of the simulation and free energy calculation details were kept identical to those described in the previous section.

2.7. Consensus Profile for Zif268. Two approaches were explored to calculate the consensus DNA binding profile for the zinc finger protein Zif268. In the first approach, a single nucleotide scan was performed by mutating every nucleotide pair in the complementary chains one-by-one and obtaining the $\Delta\Delta G$ estimates. By setting the ΔG of the wild-type nucleotide to 0 kJ/mol, the double free energy differences for the mutants were casted to the ΔG values: $\Delta G^i_{\mathrm{mut}} = \Delta G^i_{\mathrm{wt}} + \Delta\Delta G^i_{\mathrm{mutation}}$, where i denotes the position in the DNA sequence. With the assumption that the nucleotide contributions are independent of one another, the free energies were converted to the probabilities of finding nucleotide n at sequence position i:

$$p_{n_i} = \frac{e^{-\Delta G^i_n/RT}}{\sum_j^{\mathrm{A,T,G,C}} e^{-\Delta G^i_j/RT}} \tag{1}$$

where R is the universal gas constant and T is the absolute temperature.

In the second approach, we explicitly considered the contributions of the neighboring nucleotides by calculating the $\Delta\Delta G$ values for all combinations of neighboring nucleotide pairs. To accomplish that, in addition to the already acquired single nucleotide mutations, we scanned the DNA sequence with double nucleic acid mutations. This approach provides access to the probability of finding nucleotide n at position i conditioned on the nucleotides at the neighboring positions $i-1$ and $i+1$. To enable direct comparison of the obtained results to the experimental and single nucleotide scan results, we summed over the conditional probabilities:

$$p_{n_i} = \frac{\sum_k^{\mathrm{A,T,G,C}} e^{-\Delta G^{i+1}_{n_i|k}/RT} + \sum_l^{\mathrm{A,T,G,C}} e^{-\Delta G^{i-1}_{n_i|l}/RT}}{\sum_j^{\mathrm{A,T,G,C}} \left(\sum_k^{\mathrm{A,T,G,C}} e^{-\Delta G^{i+1}_{j|k}/RT} + \sum_l^{\mathrm{A,T,G,C}} e^{-\Delta G^{i-1}_{j|l}/RT} \right)} \tag{2}$$

where $\Delta G^{i+1}_{n_i|k}$ and $\Delta G^{i-1}_{n_i|l}$ denote the free energy difference for nucleotide n at position i given nucleotide k at position $i+1$ and nucleotide l at position $i-1$, respectively. The dependence between the positions $i+1$ and $i-1$ was not considered in this case. To estimate the probabilities of the terminal nucleotides, the free energy of only one available neighboring nucleic acid was considered. The calculated probabilities were visualized as logo plots[47] and compared to the experimentally measured nucleotide frequencies for every position at the binding site.[7,48] We used the Jensen–Shannon (JS) divergence to quantify the difference between the binding profiles:

$$\begin{aligned} \mathrm{JS}(X, Y) &= \frac{1}{L}\sum_{i=1}^{L} \mathrm{JS}(x_i, y_i) \\ &= \sum_{i=1}^{L} \left[\frac{1}{2}\mathrm{KL}\left(x_i \middle\| \frac{1}{2}(x_i + y_i) \right) \right. \\ &\quad \left. + \frac{1}{2}\mathrm{KL}\left(y_i \middle\| \frac{1}{2}(x_i + y_i) \right) \right] \end{aligned} \tag{3}$$

In the above expression, X and Y represent the profiles that are being compared, L denotes the length of the sequence, and KL stands for the Kullback–Leibler divergence, which is expressed as

$$\mathrm{KL}(x_i \| y_i) = \sum_n^{\mathrm{A,T,G,C}} p^x_{n_i} \ln \frac{p^x_{n_i}}{p^y_{n_i}} \tag{4}$$

where $p^x_{n_i}$ is the probability for nucleotide n to be found at position i in sequence X. When estimating the KL divergence, we applied Laplace smoothing by adding a constant of 0.001 to all of the probabilities followed by a subsequent normalization.

The JS divergence was used to compare the computed binding profiles to the experimental references. First, 10^4 random profiles were generated by sampling the nucleotide probabilities from a uniform distribution. The JS divergence was calculated for every random profile and the experimentally obtained binding profiles. Further, we used the free energies and the associated standard deviations ($\sigma_{\Delta G}$) calculated from the single nucleotide scan to obtain 10^4 profiles, where the probability for every nucleotide at each position was sampled from a normal distribution $\mathcal{N}(\Delta G, \sigma^2_{\Delta G})$. Each of the profiles was compared to the experimental reference by means of JS divergence. The same sampling and comparison to the experimental sequence procedure was repeated using the free energies obtained from the double nucleotide scan. This approach allowed an assessment of the significance of recovering experimental binding profiles by the single and double nucleotide scans. The following experimentally obtained binding profiles were used for the comparison: SELEX experiments[4,48] and protein microarray binding (PBM) experiments.[3,7] The binding profiles from the PBM experiments[3,7] were retrieved from the UniProbe database.[49]

To compare the profiles obtained from the single and double nucleotide scans to one another, we employed another strategy. Having obtained the profile samples for the single nucleotide scan (in the same way as in the previous comparison), we calculated the JS divergence among the profiles by randomly selecting pairs of samples from the generated pool of profiles. The same calculation was performed for the double nucleotide scan. We termed these comparisons "intra", since the JS divergence calculations were performed using the profiles sampled with the data of one nucleotide scan only. Finally, the "inter" comparison was performed by randomly sampling pairs of profiles from the single and double nucleotide scans and comparing them in terms of the JS divergence.

We also calculated the binding profile entropy normalized per nucleotide position (expressed in bits):

$$H = \frac{1}{L}\sum_{i=1}^{L} \sum_n^{\mathrm{A,T,G,C}} -p_{n_i} \log_2 p_{n_i} \tag{5}$$

To compare the free-energy-based binding profile determination to other computational methods, we generated the Zif268 binding profile using a number of established approaches. Two random forest machine learning methods based on the amino acid sequence were probed: B1H-RC[11] and ZFModels.[9] Another approach was based on the contact model using support vector machines.[10] All of these algorithms were trained on a large set of zinc finger binding motifs. In addition, we also tested a more general protein–DNA binding profile prediction algorithm that was not particularly aimed at the zinc finger analysis. This approach, termed PiDNA,[13] utilizes structural models of protein and DNA, generates mutations,

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275–6289

and calculates free energy differences using a knowledge-based potential energy function. While PiDNA was not primarily designed for the zinc finger binding motif detection, Zif268 was used in the parametrization of PiDNA's energy function.[50]

2.8. Molecular Dynamics Simulation Parameters. For all of the simulations carried out in this work, the following simulation setup was used. The system was placed in a dodecahedral box and solvated with TIP3P water.[51] Bond and angle vibrations in water molecules were constrained using the SETTLE algorithm.[52] The bond vibrations in other molecules were constrained using the LINCS algorithm.[53] Sodium and chloride ions were added to neutralize the system and reach a salt concentration of 150 mM. For the simulations with the Amber99SB*ILDN-BSC1 force field, Joung and Cheatham[54] ion parameters were used. A time step of 2 fs was used to integrate the equations of motion. The thermostating of the system was performed with the velocity rescaling thermostat[55] using a time constant of 0.1 ps and a reference temperature of 298 K. The pressure was kept at 1 bar by means of the Parrinello−Rahman barostat[56] with a time constant of 5 ps. The particle mesh Ewald method[57,58] was used to treat the electrostatic interactions with a Fourier grid spacing of 0.12 nm, interpolation order of 4, and relative interaction strength at the cutoff of 10^{-5}. A short-range electrostatic interaction cutoff of 1.1 nm was used for the equilibration simulations, while for the nonequilibrium transitions a cutoff of 1.2 nm was employed. The van der Waals interactions were switched off in the range from 1.0 to 1.1 nm. A dispersion correction for energy and pressure was applied. A soft-core function with the default parameters[59] was used for the nonbonded interactions during the nonequilibrium transitions. Equilibrium simulations were run with Gromacs 5.1, while nonequilibrium transitions were performed with Gromacs 4.6 using a specialized soft-core function.[59]

3. RESULTS

3.1. DNA Mutation Library Validation. The constructed DNA mutation libraries and overall free energy calculation workflow were validated by calculating double free energy differences over a thermodynamic cycle using the Drew−Dickerson dodecamer as a model system (Figure 2A,B). By construction the theoretical $\Delta\Delta G$ value of such a cycle is 0 kJ/mol.

The distributions of the $\Delta\Delta G$ values obtained from the alchemical calculations are shown in Figure 2C. These estimates provide insight into the magnitude of error that could be expected when using our approach for the subsequent calculation of the free energies in the protein−DNA complexes. The purine-to-purine and pyrimidine-to-pyrimidine mutations (first column in Figure 2C) show the smallest spread around the expected value. This is a natural consequence of the mutation library construction: purine/purine and pyrimidine/pyrimidine atom mappings involved minimal possible perturbation to the system. For the other mutations, the mappings follow a dual topology approach for the whole nucleic base, thus resulting in a larger perturbation and broader $\Delta\Delta G$ distributions.

On average the deviations from the theoretical value of 0 kJ/mol overall did not exceed 0.9 kJ/mol. In all of the investigated cycles, the 0 kJ/mol mark fell within the range of one standard deviation from the calculated mean $\Delta\Delta G$ value.

These calculations validate our hybrid structure/topology and subsequent free energy calculation setup procedures. In addition, the calculations of the $\Delta\Delta G$ values over a closed thermodynamic cycle quantify the maximal accuracy that can be expected in the subsequent calculations of the mutations in protein−DNA complexes.

3.2. Large-Scale Nucleotide Mutation Scan in Protein−DNA Complexes. The overall results of the nucleotide mutation scan are summarized in Figure 4. The thermodynamic cycle depicted in Figure 4A was constructed to calculate the double free energy differences upon nucleic acid mutations in 16 protein−DNA complexes. In total, 397 mutation cases were studied, and the pooled results are shown in Figure 4B. For the most accurate protocol (consensus of the force fields and doubled simulation time), we obtained an average unsigned error (AUE) of 5.6 kJ/mol and a correlation coefficient of 0.57 for the whole data set when comparing the calculated values to the experimental measurements. The results for the two force fields considered separately drop to an AUE of ∼6.5 kJ/mol and correlation coefficient of ∼0.5. The trends in comparison of the calculations to the experimental measurements are clearly visible: combination of the force fields decreases the AUE and increases the correlation. Longer sampling time further improves the agreement between the calculated and measured $\Delta\Delta G$ values.

For an in depth analysis of the performance of the MD-based alchemical free energy calculations, we dissected the AUE and correlation estimates system-wise (Figures 4C and S1−S16). In addition, as a reference value to assess the quality of the calculations, we have indicated the correlation values obtained by Morozov et al.[14] by means of Rosetta-based modeling.[60] The letter *S* in the correlation plot (Figure 4C) denotes the static model approach by Morozov et al.: here neither amino acid side chain nor DNA rearrangements were allowed, and Rosetta was used only to score the structural models. The dynamic model (marked by the letter *D*) involved minimization of the interfacial side chains and DNA torsion angles. The contact model (letter *C*) was a simplistic yet predictive approach that did not employ the Rosetta scoring but relied solely on the number of contacts that the consensus DNA sequence made with the protein. The details of the models as well as the original calculations using the Rosetta protocols are described in ref 14. Another reference value (indicated by the letter *F* in Figure 4C) denotes correlations with the experimental $\Delta\Delta G$ values obtained from calculations using the FoldX software.[16]

The calculated $\Delta\Delta G$ values for the zinc finger transcription factor Zif268 and its D20A mutant showed a remarkable agreement with experiment. MD-based alchemical calculations have already been used to investigate this protein,[26] and with our current protocol we were able to reproduce the previous observations. The direct comparison to experiment among the different prediction approaches should be considered with caution because not all of the experimental values were measured exactly—in some cases only the lower bound of destabilization could be obtained.

The λ repressor dimer (λR) is another example for which high-accuracy agreement with the experimental measurements was obtained. For this case, both the AUE and correlation coefficient are significantly better than the average values estimated over all of the systems. Only two outlier mutations for λR were predicted to destabilize the complex significantly more than was observed in the experiment (Figure S3). A different situation was observed for the *trp* repressor (*trp*R)−DNA complex. Here, in a small set of mutations (nine values), the correct trend in terms of correlation was identified very

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

accurately, but the absolute $\Delta\Delta G$ values strongly deviate from the experimental ones (AUE of 12.4 kJ/mol). Interestingly, *trp*R has contacts only with the DNA backbone and not with the nucleic bases.[61] Therefore, the ability to detect the correct correlation is remarkable, as the mutation effect must majorly be manifested via the changes in DNA geometry. The lack of contacts with the nucleic bases also explains the poor performance of the contact model.

For the transcription activator/repressor c-Myb protein−DNA complex we observed predictive power similar to that of the static Rosetta model. As described in Methods, we updated the experimental value set for this case. When the same experimental values as reported by Morozov et al. were used, the results were similar to those in Figure 4C (AUE = 7.3 kJ/mol, cor = 0.49). The c-Myb simulations, as well as those for ethylene-responsive transcription factor AtERF1, were started from the NMR models, in contrast to the rest of the systems, which were initiated with crystallographically resolved structures. This difference in the starting configuration may explain the decreased agreement between the calculated free energies and the experimental measurements.

*Bam*HI restriction endonuclease has been investigated experimentally by mutating the flanking nucleotides up- and downstream from the recognized sequence.[62] The association constants obtained from the nitrocellulose-binding experiments showed a decrease in binding affinity in a narrow range reaching only up to 8 kJ/mol (Figure S7). The calculated values, however, display a larger spread, in some cases indicating that the mutations may be favorable (Figure S7). The trend in terms of correlation coefficient for *Bam*HI is comparable to those for the static Rosetta model and FoldX calculations, whereas the dynamic relaxation allowed Rosetta to capture the mutation effects more accurately. In addition to the simulations starting from the crystallographic structure (Protein Data Bank (PDB) entry 1bhm),[63] we also performed the calculations with another structure (PDB entry 2bam).[64] Here the protein−DNA complex was resolved together with two bound calcium ions; also, more residues in the protein's N-terminus interacting with DNA were resolved in the 2bam structure. Nevertheless, the results did not change significantly with the use of the different starting structure. It is also important to note that for the simulations with both crystallographic structures, the first nucleotide in the flanking region was missing in comparison with the experimental mutation setup.

Another endonuclease analyzed in this work, *Eco*RI, while having a low sequence similarity to *Bam*HI, shares similar structural features.[65] In this case, Lesser et al.[66] mutated nucleotides in the recognition site and observed that the substitutions of the canonical sequence were highly unfavorable. The alchemical calculations captured the destabilizing effect well for all of the mutations analyzed. While the correlation coefficient in this case is lower than the average value, it is partly distorted by one outlier (Figure S11) in which a simultaneous mutation of six nucleotides was performed and the convergence of the $\Delta\Delta G$ estimate was not yet achieved (the correlation coefficient without this value reaches 0.53). In spite of the low correlation coefficient, the alchemical calculations were able to outperform both the static and dynamic Rosetta models.

The MAT**a**1 and MATα2 homeodomain proteins bind DNA to form heterodimers and act as repressors in yeast. Jin et al.[67] constructed an assay in which binding of the MAT**a**1/α2 proteins to a consensus DNA site would repress *lacZ* expression, which could be monitored by observing the repression ratio of the β-galactosidase activity. This assay allowed quantification of the effects of the nucleotide substitutions in the consensus site on the gene repression. On the other hand, the monitored quantity (repression ratio) is only indirectly related to the free energy changes in the protein−DNA interaction. Another complication in this case was the range of the free energy change values for the mutations: the experimental $\Delta\Delta G$ values did not exceed 9 kJ/mol, thus posing a difficult challenge of capturing subtle differences. The calculated double free energy differences show only a very weak correlation with the repression ratios converted to double free energy differences. Interestingly, the AUE in this case is below the average value, indicating that the absolute errors made in free energy estimation were not large. Modeling with Rosetta or FoldX was more successful for this case, with correlation coefficients ranging from 0.35 to 0.57.

The free energy calculations for the estrogen receptor (ER) bound to DNA captured the trends well: in a set of only seven mutations, weakly, intermediately, and strongly destabilizing mutations were clearly identified (Figures 4C and S9). Matching the exact free energy differences in this case is difficult to expect, since the experimental measurements provide only semiquantitative estimates of the $\Delta\Delta G$ values.[68]

For the mutations of DNA interacting with the transcriptional activator Ndt80, the performance of alchemical calculations was comparable to that of Rosetta modeling and slightly worse than that of FoldX calculations. The $\Delta\Delta G$ estimates were able to distinguish the destabilizing mutations from those having a mild effect or even stabilizing the interaction (Figure S10).

The prediction quality for the Cro repressor complex was above average in terms of both AUE and correlation coefficient. The agreement with experiment was also significantly better than those of the Rosetta and FoldX models. Interestingly, in this case the results obtained from the two force fields differed significantly (Figure S12), with Charmm36m outperforming Amber99sb*ILDN-BSC1. The consensus approach again resulted in good agreement with the experimental measurements.

In the case of the *Lac* repressor (*Lac*R)−DNA complex, the alchemical calculations captured well the trend in the double free energy differences, whereas a large error was made in predicting the absolute $\Delta\Delta G$ values. This deviation mainly comes from the calculation in the Amber99sb*ILDN-BSC1 force field (Figure S13). Upon closer inspection of these results, it appeared that the Amber free energy estimates lacked convergence: the work distributions obtained from the forward and backward transitions lacked overlap, indicating large work dissipation along the alchemical path. To improve the convergence, the Amber99sb*ILDN-BSC1 transitions were repeated three times. While this did not remove the insufficient convergence artifact completely, the increased overlap of the forward and backward work distributions allowed more accurate free energy estimates to be acquired, which in turn improved the agreement with the experimental measurements (light symbols for *Lac*R in Figures 4C and S13).

The PU.1 ETS protein, similar to the *Bam*HI endonuclease, is capable of an indirect DNA readout. Its binding affinity to DNA has been demonstrated to change with mutation of the residues flanking the core consensus binding site.[69] Similar to the case of *Bam*HI, the alchemical calculations for the PU.1

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

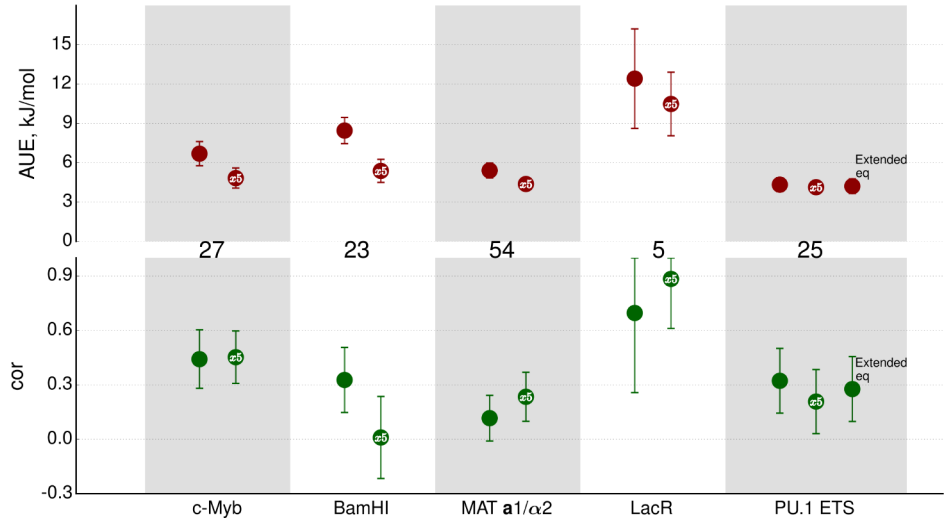


Figure 5. Results of the increased sampling simulations. The average unsigned error (AUE) and correlation coefficient between the calculated and experimental $\Delta\Delta G$ values are shown. The "x5" labels mark the results obtained from the simulations repeated five times. For the PU.1 ETS system an additional calculation was performed (marked "Extended eq") by extending a 25 ns run to reach an equilibrium simulation length of 50 ns. The numbers in between the panels are the numbers of mutation cases for the protein−DNA complexes.

ETS protein−DNA complex performed worse than average. The indirect DNA readout based on the nucleic base coupling was captured significantly better by Rosetta considering the DNA conformational energies only (Figure 4C). On the other hand, it is important to note that the AUE of the alchemical calculations for the case was significantly lower than the average over all of the systems analyzed. This is due to the fact that the calculations were able to correctly predict that most of the double free energy differences in this case were small in their absolute value, which in turn also made it difficult to capture the trends in this data set with more accuracy.

The double free energy differences for the catabolite gene activator protein (CAP)−DNA complex were predicted with above-average accuracy, outperforming the dynamic Rosetta and FoldX models and performing on par with the static Rosetta free energy calculation. The only troublesome observation from the mutations of the CAP system is the mildly stabilizing $\Delta\Delta G$ estimates for two mutation cases (Figure S15), while the experimental measurements predict destabilization of the complex. Upon closer inspection, this erroneous prediction was identified to come from the Amber99sb*ILDN-BSC1 calculations only.

The quality of the free energy estimates for the Tus protein interacting with the *Ter* DNA sequence is slightly higher than the Rosetta and FoldX predictions. For this protein−DNA complex, we expanded the experimental data set in comparison with the one used by Morozov et al. by incorporating more mutations and the associated $\Delta\Delta G$ values.[70] Compared with the original data set used by Morozov et al., the calculation quality did not change much (AUE = 4.7 kJ/mol and correlation coefficient = 0.41).

3.3. How Much Can We Improve with Increased Sampling Time? To probe whether longer sampling times would yield free energy estimates closer to the experimentally measured values, we selected five systems for an extended investigation. For each of the systems, the 25 ns equilibrium simulation was repeated five times independently. These equilibrium runs were subsequently used to start the alchemical nonequilibrium transitions (see Methods for a detailed description). The results of the extended sampling calculations are summarized in Figure 5.

Interestingly, while a consistent improvement in the absolute agreement with experiment in terms of AUE is visible in all cases, changes in capturing the trend in terms of the correlation coefficient are not homogeneous. For the c-Myb protein−DNA complex, the absolute prediction accuracy improved from 6.7 to 4.8 kJ/mol with the increased sampling time (AUE with respect to the experiment). The correlation coefficient, however, increased only marginally from 0.44 to 0.45. A very similar situation was observed for the MAT **a**1/α2 protein−DNA complex: a minor decrease in the AUE and increase in the correlation coefficient. In the case of *Lac*R bound to DNA, the increased sampling increased the accuracy both in terms of the AUE and correlation coefficient. However, for this system only five mutations were analyzed, and thus, the improvement in the agreement with experiment cannot be deemed significant, falling within the range of the large error bars.

For the cases with the nucleotide mutations in the flanking regions, *Bam*HI and PU.1 ETS, the increased sampling reduced the correlation coefficent of the calculated values with the experimental measurements.

While performing five independent equilibrium simulations for every mutation case increased the sampling time, such a setup also introduced bias by starting each of the runs from the same crystallographic or NMR structure. To investigate whether reducing this bias would improve the free energy calculation accuracy we further investigated the PU.1 ETS protein−DNA complex. For this case an additional set of equilibrium 25 ns simulations was performed by starting the simulations from the last conformation of a previous equilibrium simulation. Subsequently, the trajectory of the last 25 ns was used to spawn the nonequilibrium transitions and obtain the double free energy differences. The outcome of this calculation is illustrated in Figure 5, marked as "Extended eq". Clearly, for this protein−DNA complex prolonged equilibrium simulations had no effect on improving the free energy calculation accuracy.

3.4. $\Delta\Delta G$-Based Binding Profiles. The experimentally obtained and calculated binding profiles for the zinc finger

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

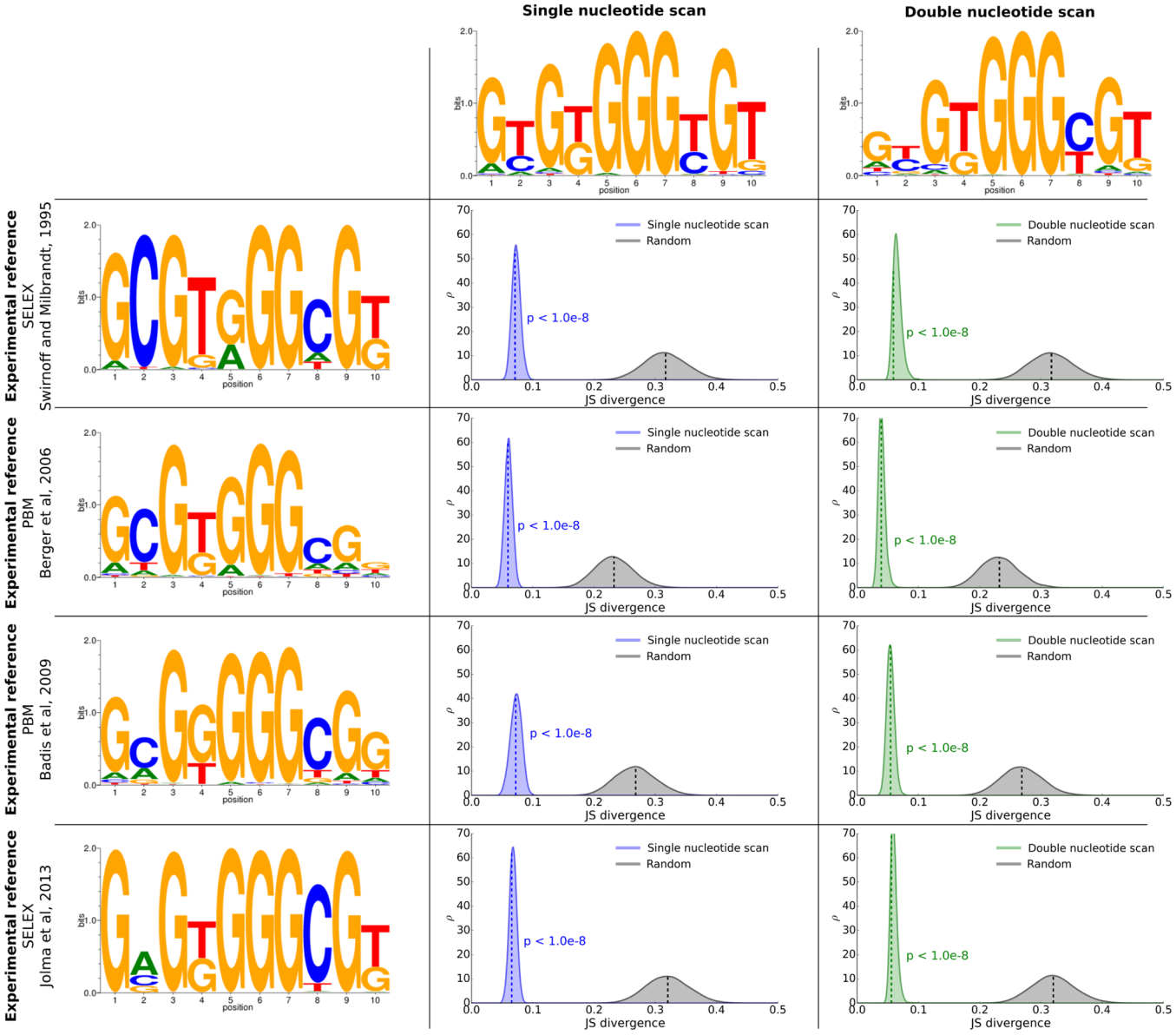


Figure 6. Calculated and experimental protein binding profiles for the Zif268 protein. The logo plots in the left column were generated from the experimentally obtained binding profiles.[3,4,7,48] The logo plots in the first row were calculated from the $\Delta\Delta G$ values obtained from the single and double nucleotide scans. The blue or green distribution in a given row and column denotes the JS divergence calculated by comparing the experimental binding profile to the calculated profile in the corresponding row and column, respectively, and the gray distribution depicts the JS divergence between the experimental profile and a randomly generated sequence.

protein Zif268 are depicted as logo plots in Figure 6. From visual comparison it appears that both the single and double nucleotide scans were able to capture the major patterns in the consensus binding sequence. This observation is further confirmed by the JS divergence analysis. The difference between the calculated profiles and any of the experimental references is significantly smaller compared with a randomly generated profile.

In comparison with the single nucleotide scan, inclusion of the neighbor effects increased the variability in the binding profiles, especially at the termini of the profile. This effect can be quantified by calculating the average information entropy for the calculated profiles. The single nucleotide scan produced a profile with an entropy of 0.6 bits per nucleotide position, while the mean entropy for the double nucleotide scan profile increased to 0.8 bits.

Including the dependence on the neighboring nucleotides into the binding profile calculation had a significant influence on the computed results (Figure 7). The blue and green distributions in Figure 7 highlight the extent to which the predicted binding profiles vary even when they are obtained from the same set of free energy calculations. Markedly, the intrinsic differences for both versions of the nucleotide scans are significantly smaller than the divergence between the scans compared with one another directly (red curve in Figure 7).

The quality of the binding profile prediction can be assessed by comparing the calculated JS divergences to the divergences among the experimentally obtained profiles (Figure 8, top row). The experimentally determined motifs show divergence comparable to that calculated from the nucleotide scans, thus indicating that predicting a profile to be even more similar to

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

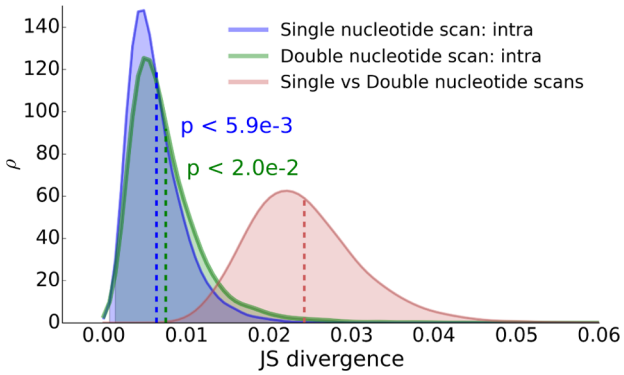


Figure 7. JS-divergence-based comparison of the binding profiles generated from the single and double nucleotide scan calculations. The blue and green distributions depict the internal divergences of the calculated distributions and illustrate the uncertainty in the determined profiles. The red distribution was generated by comparing the binding profile from the single nucleotide scan to the profile from the double nucleotide scan. The single and double nucleotide scans produce binding profiles that differ from one another by significantly more than their internal uncertainty.

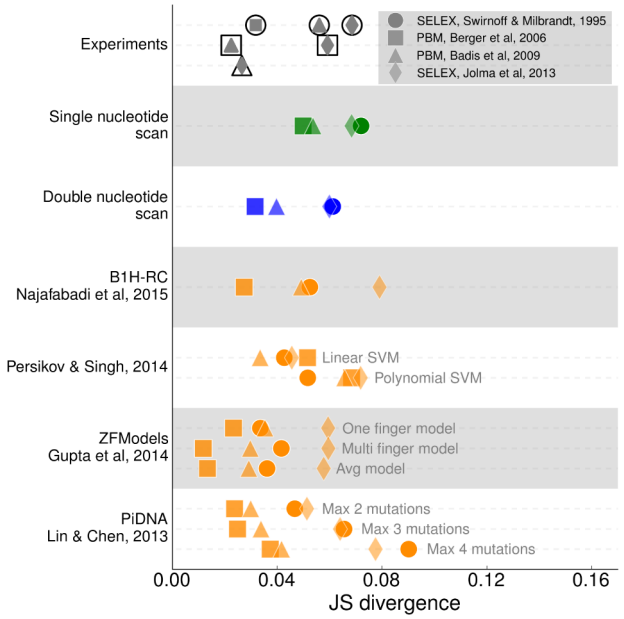


Figure 8. Comparison of the binding profiles for the Zif268 protein obtained from different computational approaches. The comparison is based on the JS divergence between the binding profiles. The experimental data were taken from refs 3, 4, 7, and 48. The first row highlights the pairwise differences between the experiments. Green and blue symbols mark comparisons using the single and double nucleotide scan calculations, respectively. The following computational approaches were used in the comparison: B1H-RC,[11] Persikov and Singh (2014),[10] ZFModels,[9] and PiDNA.[13]

any of the experimental references is hindered by the experimental uncertainty itself.

Using the same strategy of JS divergence estimation between the predicted and experimental profiles, we also evaluated a number of established computational techniques (Figure 8, orange symbols in the bottom four rows). All of the approaches showed high-quality agreement with the experimental references. Admittedly, B1H-RC,[11] ZFModels,[9] and the method of Persikov and Singh[10] were specifically designed to determine zinc finger binding profiles. The PiDNA algorithm is a more general-purpose method, but its knowledge-based energy function was tuned against the Zif268 structure. The latter method also has a feature allowing one to perform different numbers of mutations and estimate the free energy changes for the subsequent position frequency matrix generation. This approach of relaxing the nucleotide independence assumption resembles the double nucleotide scan performed in the current study. With an increasing number of tested mutations, the PiDNA-generated binding profiles diverged more from the experimental references (Figure 8, bottom row). Interestingly, the entropy of the generated motifs increased when using two, three, and four mutations (0.7, 1.0, and 1.1 bits per nucleotide position, respectively). This trend of increasing entropy matches the observation from the single and double nucleotide scans performed here.

4. DISCUSSION

4.1. pmx for DNA. The update to the pmx infrastructure presented here extends the capabilities of the software to support nucleic acid mutations in DNA. The hybrid structure/topology generation workflow closely follows the already established procedures for amino acid mutations.[19] While support in terms of the available mutation libraries is provided for two contemporary molecular mechanics force fields, the utilities are readily available to incorporate other Amber, Charmm, and OPLS family force fields. We have further validated the mutation libraries and the overall simulation setup workflow in the calculations of free energy changes across a thermodynamic cycle where the theoretical outcome is known.

Construction of the hybrid structures/topologies has also been made available via the pmx Web server interface. The online utilities allow the alchemical DNA mutation simulations to be set up without the need to install any additional software apart from the Gromacs MD engine itself.

4.2. Free Energy Estimation Accuracy. Overall, we reached an average unsigned deviation of 5.6 kJ/mol from the experimental measurements when considering all 397 analyzed nucleic acid mutations in 16 protein−DNA complexes. This AUE is larger than that observed for the amino acid mutations in the protein thermostability analysis:[18] the consensus force field approach for the protein thermostabilities allowed an AUE below 4 kJ/mol to be reached. On the other hand, the free energy estimates for the amino acid mutations in protein−protein complexes have been shown to be less accurate and yielded results comparable in quality to those observed in the current investigation.[18]

Naturally, statements regarding the accuracy in terms of agreement with experiment are highly dependent on the particular case studied. The accuracy of the experimental measurements plays an important role as well: as observed in the protein thermostability study, a difference of up to 3 kJ/mol among experimental measurements can be expected.[18] It appears that the calculated double free energy differences for the nucleic acid mutations span a wider range than the experimentally measured $\Delta\Delta G$ values (Figure 4B). There are several ways to interpret this observation: either the calculations tend to over/underestimate the actual values or the experimental measurement capabilities may be limited to obtaining values in a certain range only. In Figure 9 we compare the distributions of the calculated and experimental $\Delta\Delta G$ values used in this study (the green and cyan curves, respectively). In the background we also show a larger pool of

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

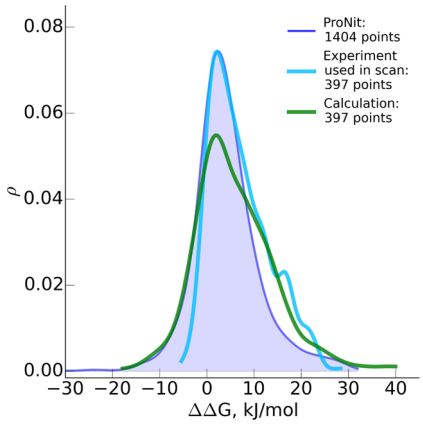


Figure 9. $\Delta\Delta G$ value ranges for the nucleotide mutations in the protein–DNA complexes. The experimental data were extracted from the ProNit database.[71,72] A total of 1404 values were considered, obtained from gel shift, filter binding, fluorescence, isothermal titration calorimetry, equilibrium competition binding, footprinting, and surface plasmon resonance experiments. The calculated value range was taken from the consensus force field with the doubled simulation time.

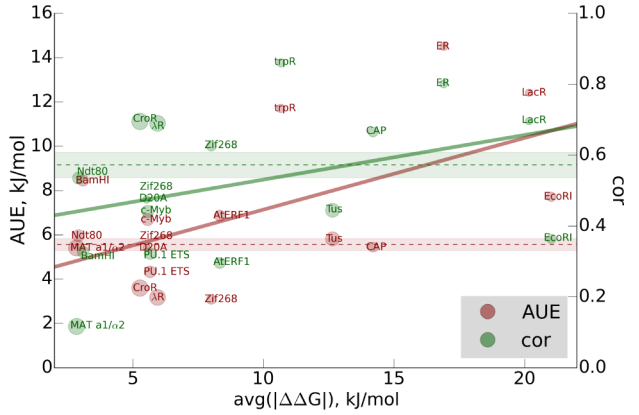


Figure 10. Evaluation of the calculation quality (based on AUE and correlation coefficient) depends on the absolute magnitude of the $\Delta\Delta G$ values in the set. With increasing absolute magnitude of the double free energy differences, the predicted free energies deviate more from the experiment in terms of AUE, but the trends in terms of correlation coefficient are captured more accurately.

data gathered from the ProNit database.[71,72] The strongest measured destabilizing mutations reach ~32 kJ/mol. This suggests a potential experimental limitation when determining highly destabilizing mutations, whereas the computational estimates do not suffer from this issue.

As for the stabilizing mutations, the calculated values match the trends expected from the large ProNit data set. The experimental data for the 397 mutation cases used in the mutation scan, however, seem to have the stabilizing mutations under-represented. This observation suggests that collecting a more representative set of mutations, including a larger number of stabilizing mutations, could further improve the agreement between the computation and experiment.

The quality assessment of the computed $\Delta\Delta G$ estimates also depends on the data set itself and the measures used to evaluate the quality of the calculations (AUE and correlation coefficient with respect to the experimental values). For example, the analyzed cases of MATa1/α2 and PU.1 ETS contained mainly mildly destabilizing mutations. For these cases, the computed $\Delta\Delta G$ values captured well the absolute changes in the binding affinity (AUE measure), whereas obtaining correct correlation coefficients within the narrow range of small values proved to be difficult (Figure 10). On the other hand, capturing the correct trends in terms of correlation coefficient was easier for the mutation sets spanning a larger range of double free energy differences. Figure 10 illustrates that the prediction quality in terms of correlation coefficient increases with the increasing absolute $\Delta\Delta G$ values in a data set. In contrast, the agreement of the computed results with the experimental measurements in terms of AUE gets worse. This highlights weak sides of the AUE and correlation coefficient measures in the current context and also emphasizes the importance of considering both measures when assessing the quality of the predicted $\Delta\Delta G$ values.

Morozov et al.[14] noted the importance of the starting structure for the quality of Rosetta modeling: the results obtained from the structures resolved by means of NMR spectroscopy agreed worse with the experimental measurements. Similarly, the MD-based free energy calculations that were started from the NMR-based structures (c-Myb and AtERF1) showed worse than average agreement with the experiments.

We also applied our simulation protocols to the four transcription factors investigated by Khabiri and Freddolino[27] and reached an average accuracy comparable to that obtained for the large-scale scan described in this work (data not shown). We are in contact with the authors of that study to investigate the larger deviations found previously.

4.3. Force Fields and Sampling. The two force fields, Amber99sb*ILDN-BSC1 and Charmm36m, performed comparably (Figure 4B). The consensus approach of averaging the results of the two force fields resulted in an improved $\Delta\Delta G$ prediction quality. A similar effect of the error cancellation between the force fields was observed in the protein thermostability calculations.[18] A more detailed analysis of the $\Delta\Delta G$ estimates from both force fields revealed that for more than 30% of the mutations analyzed in this work, Amber99sb*ILDN-BSC1 and Charmm36m make predictions pointing in the opposite directions from the experimentally measured value (Figures S17–S20 and associated text).

Calculation of the free energy changes upon nucleotide mutation may also be considered as a means to assess the quality of a force field. Currently, the DNA force field modifications are primarily validated by monitoring DNA stability and comparing structural and dynamic features obtained from simulation to the available crystallographic and NMR data.[31,32] Computing the free energy differences over a curated and well-tested mutation set with reliable experimental data offers an additional method to validate newly developed force fields.

Computationally this large-scale scan of 397 mutation cases required a combined simulation time of ~81 μs. While it is still a considerable computational challenge, such simulation times are well within the reach with the current GPU-based MD simulation engines (e.g., ~140 μs of simulation time was invested in testing the Parmbsc1 DNA force field[31]). Admittedly, the choice of sampling time used for mutations in this study was mainly based on our previous investigations of the protein thermostabilities and protein–protein interactions.[18] Investing more computational effort in sampling may improve the free energy prediction quality. In fact, we observed a better agreement with experiment once the consensus result

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275–6289

was constructed considering the whole simulations from both force fields (i.e., "Consensus ff double time" in Figure 4B).

On the other hand, the results of the increased sampling simulations for the 5 selected systems (Figure 5) are less definitive. While the increased sampling time reduced the average unsigned error from the experimental measurements, capturing the trend in terms of correlation coefficient did not improve significantly or even got worse (e.g., *Bam*HI endonuclease). Naturally, it may be of importance what approach is chosen to obtain a more representative (i.e., better-sampled) conformational distribution of the system. Starting a number of short equilibrium simulations (in this work, 25 ns runs were used) from the same starting structure may not warrant crossing higher energy barriers. Thus, in case a mutation induces a larger conformational change, this type of sampling may not suffice to observe the relevant transition in simulation. To take this concern into consideration, we also probed an approach of doubling the simulation time for the PU.1 ETS protein−DNA complex. However, for this case we observed no improvement in the $\Delta\Delta G$ estimation accuracy. It may be that the simulation time scales of 25−50 ns still do not cover relevant conformational changes upon nucleotide mutation for this system, although the Rosetta-based calculations were able to capture correlation with the experiment for PU.1 ETS much better (Figure 4C).

4.4. Determination of Binding Profiles. The method to obtain accurate free energy estimates in turn was extended to the determination of the full protein−DNA binding profiles. First, relying on the assumption of nucleotide independence in the DNA sequence, we were able to recover the binding profile for the zinc finger protein Zif268. The obtained position frequency matrix (visualized as a logo plot in Figure 6) is significantly more similar to the experimental binding profiles than randomly generated profiles.

Binding profile determination based on free energy calculations also allows testing of the validity of the nucleotide independence assumption. By performing the free energy scan mutating two neighboring nucleotides at a time, we could take into account the effects of the nearest neighbors. The binding profile generated this way was significantly different from a random selection. Including the neighbor effects also significantly altered the profile in comparison with the position frequency matrix constructed on the basis of nucleotide independence (Figure 7). In particular, the profile calculated using the double nucleotide scan had higher entropy, suggesting that the termini of the Zif268 binding site could tolerate a more diverse set of nucleotides.

The binding profiles constructed on the basis of alchemical free energy calculations perform on par with the previously established methods that were tested in this work (Figure 8). The other probed approaches have either been specifically designed to create zinc finger binding profiles or have used Zif268 to train the energy function. Therefore, it is remarkable that the first-principles-based method employed here was able to reach this level of accuracy. Obtaining a binding profile diverging even less from the experimental references is prohibited by the differences in the experimentally determined motifs themselves (as also demonstrated in ref 73).

5. CONCLUSION

This large-scale nucleotide mutation scan in the protein−DNA complexes demonstrates the readiness of molecular-dynamics-based alchemical calculations to produce high-accuracy free energy predictions. The obtained free energy differences can further be translated into binding profiles, offering a robust first-principles-based approach to contend with the already established knowledge-based potentials and machine learning algorithms. We have automated the hybrid structure/topology generation required for the alchemical calculations and provide easy-to-use access to these utilities both via a command-line implementation and as an online service (http://pmx.mpibpc.mpg.de/dna_webserver.html).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00849.

Figures S1−S16 compare the calculated and experimental free energy differences for every protein−DNA complex separately; Figures S17−S20 and the accompanying text provide details on the individual force field performance; Table S1 contains all of the experimental and calculated free energy values (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: vgapsys@gwdg.de.
*E-mail: bgroot@gwdg.de.

ORCID

Vytautas Gapsys: 0000-0002-6761-7780
Bert L. de Groot: 0000-0003-3570-3534

Funding

This work was done as part of the BioExcel CoE (www.bioexcel.eu), a project funded by the European Union (Contract H2020-EINFRA-2015-1-675728). V.G. acknowledges support by Boehringer Ingelheim Pharma GmbH.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

MD, molecular dynamics; AUE, average unsigned error; TF, transcription factor; PBM, protein binding microarray

■ REFERENCES

(1) Rohs, R.; Jin, X.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S. Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **2010**, *79*, 233−269.

(2) Rohs, R.; West, S. M.; Liu, P.; Honig, B. Nuance in the double-helix and its role in protein-DNA recognition. *Curr. Opin. Struct. Biol.* **2009**, *19*, 171−177.

(3) Badis, G.; Berger, M. F.; Philippakis, A. A.; Talukder, S.; Gehrke, A. R.; Jaeger, S. A.; Chan, E. T.; Metzler, G.; Vedenko, A.; Chen, X.; Kuznetsov, H.; Wang, C.-F.; Coburn, D.; Newburger, D. E.; Morris, Q.; Hughes, T. R.; Bulyk, M. L. Diversity and complexity in DNA recognition by transcription factors. *Science* **2009**, *324*, 1720−1723.

(4) Jolma, A.; Yan, J.; Whitington, T.; Toivonen, J.; Nitta, K. R.; Rastas, P.; Morgunova, E.; Enge, M.; Taipale, M.; Wei, G.; Palin, K.; Vaquerizas, J. M.; Vincentelli, R.; Luscombe, N. M.; Hughes, T. R.; Lemaire, P.; Ukkonen, E.; Kivioja, T.; Taipale, J. DNA-binding specificities of human transcription factors. *Cell* **2013**, *152*, 327−339.

(5) Weirauch, M. T.; Yang, A.; Albu, M.; Cote, A. G.; Montenegro-Montero, A.; Drewe, P.; Najafabadi, H. S.; Lambert, S. A.; Mann, I.; Cook, K.; Zheng, H.; Goity, A.; van Bakel, H.; Lozano, J.-C.; Galli, M.; Lewsey, M. G.; Huang, E.; Mukherjee, T.; Chen, X.; Reece-Hoyes, J. S.; Govindarajan, S.; Shaulsky, G.; Walhout, A. M.; Bouget, F.-Y.; Ratsch, G.; Larrondo, L. F.; Ecker, J. R.; Hughes, T. R. Determination

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

and inference of eukaryotic transcription factor sequence specificity. *Cell* **2014**, *158*, 1431−1443.
(6) Zhao, Y.; Granas, D.; Stormo, G. D. Inferring binding energies from selected binding sites. *PLoS Comput. Biol.* **2009**, *5*, e1000590.
(7) Berger, M. F.; Philippakis, A. A.; Qureshi, A. M.; He, F. S.; Estep, P. W.; Bulyk, M. L. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **2006**, *24*, 1429−1435.
(8) Liu, L. A.; Bradley, P. Atomistic modeling of protein−DNA interaction specificity: progress and applications. *Curr. Opin. Struct. Biol.* **2012**, *22*, 397−405.
(9) Gupta, A.; Christensen, R. G.; Bell, H. A.; Goodwin, M.; Patel, R. Y.; Pandey, M.; Enuameh, M. S.; Rayla, A. L.; Zhu, C.; Thibodeau-Beganny, S.; Brodsky, M. H.; Joung, J. K.; Wolfe, S. A.; Stormo, G. D. An improved predictive recognition model for Cys2-His2 zinc finger proteins. *Nucleic Acids Res.* **2014**, *42*, 4800−4812.
(10) Persikov, A. V.; Singh, M. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.* **2014**, *42*, 97−108.
(11) Najafabadi, H. S.; Mnaimneh, S.; Schmitges, F. W.; Garton, M.; Lam, K. N.; Yang, A.; Albu, M.; Weirauch, M. T.; Radovani, E.; Kim, J. P. M.; Greenblatt; Frey, B. J.; Hughes, T. R. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.* **2015**, *33*, 555−562.
(12) Gabdoulline, R.; Eckweiler, D.; Kel, A.; Stegmaier, P. 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. *Nucleic Acids Res.* **2012**, *40*, W180−W185.
(13) Lin, C.-K.; Chen, C.-Y. PiDNA: predicting protein−DNA interactions with structural models. *Nucleic Acids Res.* **2013**, *41*, W523−W530.
(14) Morozov, A. V.; Havranek, J. J.; Baker, D.; Siggia, E. D. Protein−DNA binding specificity predictions with structural models. *Nucleic Acids Res.* **2005**, *33*, 5781−5798.
(15) Yanover, C.; Bradley, P. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res.* **2011**, *39*, 4564.
(16) Alibes, A.; Nadra, A. D.; De Masi, F.; Bulyk, M. L.; Serrano, L.; Stricher, F. Using protein design algorithms to understand the molecular basis of disease caused by protein−DNA interactions: the Pax6 example. *Nucleic Acids Res.* **2010**, *38*, 7422−7431.
(17) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695−2703.
(18) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angew. Chem., Int. Ed.* **2016**, *55*, 7364−7368.
(19) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated protein structure and topology generation for alchemical perturbations. *J. Comput. Chem.* **2015**, *36*, 348−354.
(20) Gapsys, V.; de Groot, B. L. pmx Webserver: A User Friendly Interface for Alchemistry. *J. Chem. Inf. Model.* **2017**, *57*, 109−114.
(21) Loeffler, H. H.; Michel, J.; Woods, C. FESetup: Automating Setup for Alchemical Free Energy Simulations. *J. Chem. Inf. Model.* **2015**, *55*, 2485−2490.
(22) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. Free energy calculations by computer simulation. *Science* **1987**, *236*, 564−568.
(23) Cieplak, P.; Rao, S. N.; Grootenhuis, P. D. J.; Kollman, P. A. Free energy calculation on base specificity of drug−DNA interactions: application to daunomycin and acridine intercalation into DNA. *Biopolymers* **1990**, *29*, 717−727.
(24) Florián, J.; Goodman, M. F.; Warshel, A. Free-Energy Perturbation Calculations of DNA Destabilization by Base Substitutions: The Effect of Neutral Guanine·Thymine, Adenine·Cytosine and Adenine·Difluorotoluene Mismatches. *J. Phys. Chem. B* **2000**, *104*, 10092−10099.
(25) Beierlein, F. R.; Kneale, G. G.; Clark, T. Predicting the effects of basepair mutations in DNA-protein complexes by thermodynamic integration. *Biophys. J.* **2011**, *101*, 1130−1138.
(26) Seeliger, D.; Buelens, F. P.; Goette, M.; de Groot, B. L.; Grubmüller, H. Towards computional specificity screening of DNA-binding proteins. *Nucleic Acids Res.* **2011**, *39*, 8281−8290.
(27) Khabiri, M.; Freddolino, P. L. Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein-DNA Binding Free Energy Landscapes. *J. Phys. Chem. B* **2017**, *121*, 5151−5161.
(28) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 712−725.
(29) Best, R. B.; Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.
(30) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.
(31) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpí, J. L.; González, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D.; Orozco, M. Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **2016**, *13*, 55−58.
(32) Hart, K.; Foloppe, N.; Baker, C. M.; Denning, E. J.; Nilsson, L.; MacKerell, A. D., Jr. Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *J. Chem. Theory Comput.* **2012**, *8*, 348−362.
(33) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D., Jr CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71−73.
(34) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845−854.
(35) Gapsys, V.; Michielssens, S.; Peters, J. H.; de Groot, B. L.; Leonov, H. Calculation of binding free energies. *Methods Mol. Biol.* **2015**, *1215*, 173−209.
(36) Crooks, G. E. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1999**, *60*, 2721−2726.
(37) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Phys. Rev. Lett.* **2003**, *91*, 140601.
(38) Wing, R.; Drew, H.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E. Crystal structure analysis of a complete turn of B-DNA. *Nature* **1980**, *287*, 755−758.
(39) Ogata, K.; Morikawa, S.; Nakamura, H.; Sekikawa, A.; Inoue, T.; Kanai, H.; Sarai, A.; Ishii, S.; Nishimura, Y. Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **1994**, *79*, 639−648.
(40) Tanikawa, J.; Yasukawa, T.; Enari, M.; Ogata, K.; Nishimura, Y.; Ishii, S.; Sarai, A. Recognition of specific DNA sequences by the c-myb protooncogene product: role of three repeat units in the DNA-binding domain. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 9320−9324.
(41) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(42) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289

for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.
(43) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision C.01; Gaussian, Inc.: Wallingford, CT, 2009.
(44) Vanommeslaeghe, K.; MacKerell, A. D., Jr Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144−3154.
(45) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Jr Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155−3168.
(46) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671−690.
(47) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **2004**, *14*, 1188−1190.
(48) Swirnoff, A. H.; Milbrandt, J. DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.* **1995**, *15*, 2275−2287.
(49) Hume, M. A.; Barrera, L. A.; Gisselbrecht, S. S.; Bulyk, M. L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein−DNA interactions. *Nucleic Acids Res.* **2015**, *43*, D117−D122.
(50) Chen, C.-Y.; Chien, T.-Y.; Lin, C.-K.; Lin, C.-W.; Weng, Y.-Z.; Chang, D. T.-H. Predicting target DNA sequences of DNA-binding proteins based on unbound structures. *PLoS One* **2012**, *7*, e30446.
(51) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(52) Miyamoto, S.; Kollman, P. A. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−962.
(53) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(54) Joung, I. S.; Cheatham, T. E., III Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020−9041.
(55) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(56) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(57) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(58) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(59) Gapsys, V.; Seeliger, D.; de Groot, B. L. New soft-core potential function for molecular dynamics based alchemical free energy calculations. *J. Chem. Theory Comput.* **2012**, *8*, 2373−2382.
(60) Havranek, J. J.; Duarte, C. M.; Baker, D. A simple physical model for the prediction and design of protein−DNA interactions. *J. Mol. Biol.* **2004**, *344*, 59−70.
(61) Otwinowski, Z.; Schevitz, R. W.; Zhang, R. G.; Lawson, C. L.; Joachimiak, A.; Marmorstein, R. Q.; Luisi, B. F.; Sigler, P. B. Crystal structure of trp represser/operator complex at atomic resolution. *Nature* **1988**, *335*, 321−329.
(62) Engler, L. E.; Sapienza, P.; Dorner, L. F.; Kucera, R.; Schildkraut, I.; Jen-Jacobson, L. The energetics of the interaction of BamHI endonuclease with its recognition site GGATCC. *J. Mol. Biol.* **2001**, *307*, 619−636.
(63) Newman, M.; Strzelecka, T.; Dorner, L. F.; Schildkraut, I.; Aggarwal, A. K. Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science* **1995**, *269*, 656−664.
(64) Viadiu, H.; Aggarwal, A. K. The role of metals in catalysis by the restriction endonuclease Bam HI. *Nat. Struct. Biol.* **1998**, *5*, 910−916.
(65) Newman, M.; Strzelecka, T.; Dorner, L. F.; Schildkraut, I.; Aggarwal, A. K. Structure of Restriction-Endonuclease BamHI and Its Relationship to EcoRI. *Nature* **1994**, *368*, 660−664.
(66) Lesser, D. R.; Kurpiewski, M. R.; Jen-Jacobson, L. The energetic basis of specificity in the Eco RI endonuclease-DNA interaction. *Science* **1990**, *250*, 776−787.
(67) Jin, Y.; Zhong, H.; Vershon, A. K. The yeast a1 and α2 homeodomain proteins do not contribute equally to heterodimeric DNA binding. *Mol. Cell. Biol.* **1999**, *19*, 585−593.
(68) Boyer, M.; Poujol, N.; Margeat, E.; Royer, C. A. Quantitative characterization of the interaction between purified human estrogen receptor α and DNA using fluorescence anisotropy. *Nucleic Acids Res.* **2000**, *28*, 2494−2502.
(69) Poon, G. M. K.; Macgregor, R. B., Jr Base coupling in sequence-specific site recognition by the ETS domain of murine PU.1. *J. Mol. Biol.* **2003**, *328*, 805−819.
(70) Coskun-Ari, F. F.; Hill, T. M. Sequence-specific interactions in the Tus-Ter complex and the effect of base pair substitutions on arrest of DNA replication in Escherichia coli. *J. Biol. Chem.* **1997**, *272*, 26448−26456.
(71) Prabakaran, P.; An, J.; Gromiha, M. M.; Selvaraj, S.; Uedaira, H.; Kono, H.; Sarai, A. Thermodynamic database for protein−nucleic acid interactions (ProNIT). *Bioinformatics* **2001**, *17*, 1027−1034.
(72) Kumar, M. D. S.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. ProTherm and ProNIT: thermodynamic databases for proteins and protein−nucleic acid interactions. *Nucleic Acids Res.* **2006**, *34*, D204−D206.
(73) Stormo, G. D.; Zhao, Y. Determining the specificity of protein−DNA interactions. *Nat. Rev. Genet.* **2010**, *11*, 751−760.

DOI: 10.1021/acs.jctc.7b00849
J. Chem. Theory Comput. 2017, 13, 6275−6289