

Comment on “Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein-DNA Binding Free Energy Landscapes”

Vytautas Gapsys,[†] Morteza Khabiri,[‡] Bert L. de Groot,^{*,‡,§} and Peter L. Freddolino^{*,‡,§}

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

[‡]Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, Michigan, United States

[§]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan, United States

J. Phys. Chem. B 2017, 121 (20), 5151–5161, DOI: 10.1021/acs.jpcc.6b12450

S Supporting Information

ABSTRACT: Sequence-specific DNA binding transcription factors play an essential role in the transcriptional regulation of all organisms. The development of reliable in silico methods to predict the binding affinity landscapes of transcription factors thus promises to provide rapid screening of transcription factor specificities and, at the same time, yield valuable insight into the atomistic details of the interactions driving those specificities. Recent literature has reported highly discrepant results on the current ability of state-of-the-art atomistic molecular dynamics simulations to reproduce experimental binding free energy landscapes for transcription factors. Here, we resolve one important discrepancy by noting that in the case of alchemical free energy calculations involving base pair mutations, a common convention used in improving end point convergence of mixed potentials in fact can lead to erroneous results. The underlying cause for inaccurate double free energy difference estimates is specific to the particular implementation of the alchemical transformation protocol. Using the Gromacs simulation package, which is not affected by this issue, we obtain free energy landscapes in agreement with the experimental measurements; equivalent results are obtained for a small set of test cases with a modified version of the AMBER package. Our findings provide a consistent and optimistic outlook on the current state of prediction of protein–DNA binding free energy interactions using molecular dynamics simulations and an important precaution for appropriate end point handling in a broad range of free energy calculations.

INTRODUCTION

We have identified a technical issue with the free energy calculations used in Khabiri and Freddolino¹ that requires reinterpretation of several findings of the paper. In the present application, namely that of estimating the change in binding free energy of a DNA-binding protein in response to DNA sequence, we had initially reported, using dual topology thermodynamic integration (TI) calculations as implemented in AMBER14, that substantial disagreement existed between the calculated and experimental changes in binding free energy, both in ranking and magnitude.¹ In developments subsequent to the paper in question, Gapsys and de Groot² applied the recently developed pmx method to the same set of proteins as in Khabiri and Freddolino¹ and obtained substantially different results in better agreement with experiment, while using the same DNA force field. Importantly, as we found through subsequent correspondence, the exchange of the equilibrium snapshots used for free energy calculations did not alter the conclusions reached with either method, demonstrating that the source of the discrepancy was in the free energy calculation itself. After investigating many potential causes for the identified discrepancy, we found that the key difference between the methods arose from the treatment of the decoupled atoms in the two approaches. As we will detail below, the convention used in Khabiri and Freddolino¹ can, in

certain cases, lead to strong coupling of the phase space of the decoupled base pair with the unmodified region of the system. In such a case, the assumption of equal contributions of the decoupled atoms implicit in the thermodynamic cycle used to analyze the calculations is thus violated, giving rise to an uncanceled contribution to the binding calculated free energy change.

The key difficulty arising from the methods used in the publication arise from the treatment of the end point states of the thermodynamic cycle (Figure 2 of Khabiri and Freddolino¹). The softcore potential implementation used in AMBER alchemical free energy calculations (and in some other implementations, for example, that in NAMD,³ unless the alchDecouple parameter is set to off) requires a dual-topology formalism and treats the decoupled atoms in each of the end-point states as noninteracting with the remainder of the system or the other end point, but fully interacting with other perturbed atoms that are part of the same end point state. Thus, even in the initial $\lambda = 0$ state, the atoms of the incoming group are subject to all standard nonbonded interactions among themselves and to both internal and external bonded terms arising from covalent bonds within the incoming group or across the incoming/common atom

Published: May 9, 2018

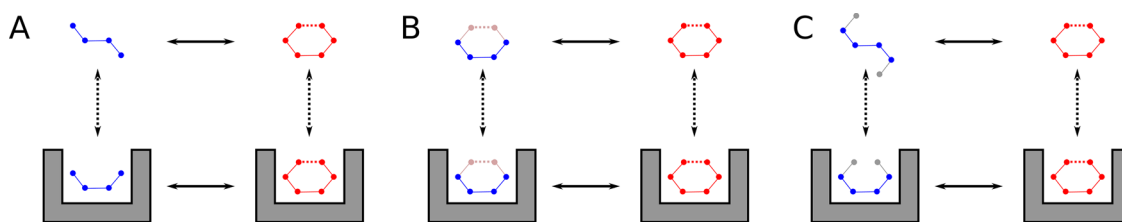


Figure 1. Schematic example of circumstances where retained nonbonded interactions can affect the outcome of alchemical transformations. (A) Hypothetical system of interest, in which a protein (gray) may bind to either of two ligands (blue or red), one of which contains a strong intramolecular nonbonded interaction (red dashed line), and the other of which is more flexible. Around the perimeter of the thermodynamic cycle, dashed arrows indicate the binding reactions for which a free energy difference is to be calculated, and solid arrows indicate the alchemical transformations that may be practically used to obtain the desired change in binding free energy. (B) Effects of retaining the internal nonbonded interactions of decoupled atoms on the system shown in (A); the conformational ensemble of the unbound blue end point (top left) is biased to maintain the red state nonbonded interaction. (C) Schematic equivalent to (B), but in which the nonbonded interactions of decoupled atoms are removed, thus, allowing the blue ligand to properly sample appropriate conformational space at the top left and bottom left thermodynamic end points without perturbation by effects from the red state.

boundary. Similar patterns are followed for atoms in the outgoing group.

The motivation behind the commonly used implementation described above is that retention of the internal nonbonded interactions and all bonded interactions of the noncommon group atoms substantially aids convergence near the simulation end points,⁴ by providing good conformational overlap between the fully “decoupled” state and states where the incoming/outgoing atoms interact with the remainder of the system. Unfortunately, a vulnerability is introduced in the process. A fundamental assumption of the thermodynamic cycle applied in all calculations described here is that the free energy contribution from the decoupled atoms of the $\lambda = 0$ state are identical in the DNA-only form and DNA–protein complex (likewise for the $\lambda = 1$ state). That equality would be trivially satisfied if the noninteracting atoms were fully “dummied” (that is, converted to either unconstrained ideal gas atoms or completely constrained noninteracting particles), but any variations, as are frequently used in practice for free energy calculations, require careful consideration. In particular, the retention of nonbonded interactions across the incoming and outgoing base pairs (that is, Watson–Crick hydrogen bonds between each base and its complement throughout the alchemical transformation), combined with normal bonded interactions between those atoms and the rest of the system, can cause a non-negligible free energy contribution from the “decoupled” base pair. This artificial contribution to free energy does not properly cancel due to differences in conformational ensembles between the DNA-only and DNA–protein states.

In a theoretical analysis of issues related to bonded contributions to free energy calculations, Boresch and Karplus⁴ distinguished between an “ideal gas atom” state, in which decoupled atoms have lost all of their bonded interactions, and an “ideal gas molecule” state, in which nonbonded interactions of the decoupled atoms are removed, but their bonded interactions are retained. Using both purely theoretical approaches and numerical calculations on model systems, Boresch and Karplus⁴ showed that the ideal gas molecule end point provides equivalent results to the ideal gas atom end point and can aid convergence.^{4,5} However, that analysis required the removal of all nonharmonic bonded terms between the decoupled atoms and the common portion of the system at the simulation end points in order to demonstrate that the ideal gas molecule state could be used (the authors did not examine the effects of deviations from that

assumption). In addition, even in the course of considering the simple case of a methanol to ethane transformation, the authors constrained the particles (denoted C1E and C1M), representing the carbon common to both topologies, on the grounds that “if C1E and C1M were free to move, there could be unphysical coupling between the two halves of the system. For example, a displacement of the ethane part that is caused by an interaction with the solvent could affect the methanol part, which should not feel this particular interaction with solvent, and vice versa”.⁵ In practice, however, modern dual topology alchemical transformations often ignore both of the warnings raised above, frequently retaining all bonded interactions between the incoming/outgoing and common atom groups and not explicitly treating the potential concern of coupling between the incoming and outgoing groups through those bonded interactions.

Upon careful consideration of the alchemical transformations used in Khabiri and Freddolino,¹ we recognize that a similar form of nonphysical coupling between the incoming and outgoing atom groups may occur in some cases where decoupled atom groups are treated in an “ideal gas molecule” state, but retain the nonbonded interactions among themselves and bonded interactions with the unchanged portion of the molecule (as is the case in the AMBER dual topology formalism). Consider the simple schematic example in Figure 1A: our system of interest consists of an arbitrary ligand molecule (blue ball/stick representation) that may be bound to a protein target (gray), and we wish to use an alchemical transformation to calculate the change in binding free energy between the blue ligand and a different potential ligand (red). Importantly for the present discussion, the red ligand contains a strong nonbonded intramolecular interaction (dashed line). Whereas Figure 1A shows an idealized representation of the system, Figure 1B shows the actual thermodynamic cycle that would arise in an (effectively single topology) FEP approach if nonbonded interactions among outgoing atoms are retained: the strong nonbonded interaction of the “red” state will remain formed even at alchemical end points where the red state is decoupled from the system, and it may thus bias the conformational ensemble of the “blue” end point states (for example, preventing the “blue” state from taking on conformations that would otherwise break the “red” state’s intramolecular interaction). On the other hand, if all nonbonded interactions of the decoupled atoms are removed (Figure 1C), the red state’s intramolecular interactions will not bias the conformations sampled by the blue state. It is

Table 1. Summary of Experimental Data Sets Used for Comparison with MD Results^a

data set name	protein	organism	accession	experiment type	ref
Human_ELK1_1	ELK1	human	M04752_2.00	HT-SELEX	Yin et al. ⁸
Human_ELK1_2		human	M5380_1.02	HT-SELEX	Jolma et al. ⁹
Human_ELK1_3		human	M09051_2.00	ChIP-Seq	Kulakovskiy et al. ¹⁰
Human_ELK1_4		human	M10488_2.00	SELEX	Matys et al. ¹¹
Human_ELK1_5		human	M04754_2.00	HT-SELEX	Yin et al. ⁸
Human_ELK1_6		human	M09531_2.00	ChIP-Seq	Heinz et al. ¹²
Mouse_ELK1		mouse	M0695_1.02	PBM	Wei et al. ¹³
Stormo_ELK1		human		HT-SELEX	Ruan et al. ⁷
Human_GCM1_1	GCM1	human	M02020_2.00	PBM	Weirauch et al. ⁶
Human_GCM1_2		human	M05868_2.00	SMiLE-seq	Isakova et al. ¹⁴
Human_GCM1_3		human	M5484_1.02	HT-SELEX	Jolma et al. ⁹
Mouse_GCM1		mouse	M0812_1.02	PBM	Badis et al. ¹⁵
Stormo_GCM1		human		HT-SELEX	Ruan et al. ⁷
Mouse_MAX	MAX	mouse	M05832_2.00	SMiLE-seq	Isakova et al. ¹⁴
Human_MAX_1		human	M04114_2.00	HT-SELEX	Yin et al. ⁸
Human_MAX_2		human	M07796_2.00	ChIP-seq	Gerstein et al. ¹⁶
Human_MAX_3		human	M3539_1.02	HT-SELEX	Matys et al. ¹¹
Human_MAX_4		human	M07791_2.00	ChIP-seq	Gerstein et al. ¹⁶
Human_MAX_5		human	M0221_1.02	PBM	Badis et al. ¹⁵
Yeast_PPR1_1	PPR1	yeast	M4340_1.02	ChIP-chip	de Boer and Hughes ¹⁷
Yeast_PPR1_2		yeast	YLR014C_1432	DNase protection	Roy et al. ¹⁸

^aAll data sets, except those titled “Stormo”, were obtained via CisBP, as described in the accompanying text.

Table 2. Calculated Free Energy Changes (in Units of RT) Using Various Methods for All Possible Mutations of the CG Base Pair at Position 12 of the ELK1 Binding Site (Using the Notation of Khabiri and Freddolino¹)^a

mutation	C ₁₂ A	C ₁₂ T	C ₁₂ G
AMBER, orig	+15.14 (+11.11 – +22.66)	–24.08 (–26.13 – –22.14)	+45.87 (+41.23 – +51.30)
AMBER, mod	–0.85 (–1.81 – +0.86)	+7.33 (+5.44 – +7.79)	+0.12 (–0.59 – +0.60)
AMBER18, patched	+50.18 (+47.49 – +52.59)	–17.76 (–19.42 – –15.92)	+19.92 (+8.19 – +48.51)
Gromacs, AMBER struct	–0.56 (–2.01 – +0.88)	+5.67 (+4.07 – +7.28)	–0.89 (–2.56 – +0.79)
Gromacs, Gromacs struct	+0.41 (–0.79 – +1.61)	+8.50 (+7.40 – +9.60)	+2.31 (+1.22 – +3.40)

^aThe “orig” and “mod” AMBER free energy differences refer to the dual-topology implementation used in Khabiri and Freddolino,¹ with minor refinements (see [Methods](#)), or the modified topology implementation described in this note, respectively. “AMBER struct” and “Gromacs struct” refer to equilibrium trajectories of the end points sampled using AMBER (equivalent to those in Khabiri and Freddolino¹) or Gromacs (generated as in Gapsys and de Groot²). Estimates are given first, and then the range in parentheses indicates a 95% confidence interval. Note also that slightly different protein force fields were used for the AMBER (ff14SB) vs Gromacs (ff99SB*ildn) simulations, with the former representing an additional set of corrections on the latter;¹⁹ this likely contributes to minor differences between the calculations. The “patched” version of AMBER corresponds to a recent version in which the problematic interaction of exclusions and softcore potentials identified by the work of Loeffler et al.³⁶ has been corrected.

important to note that the problems outlined conceptually here, and in practice below, are not directly limited to either single- or dual-topology FEP approaches, but rather pertain to how dummy/decoupled atoms are treated in either formalism.

METHODS

Experimental Data Sets. To provide an expanded comparison of computed binding free energy landscapes with experiment, we have incorporated into our analysis multiple high-throughput data sets (harvested primarily from the CisBP database⁶) for each protein. The experimental data sets used are summarized in [Table 1](#); in each case, the data set with the “_1” suffix corresponds to the primary experimental data set from the host organism used in the analysis of [ref 1](#), and likewise, where relevant, the “Mouse” data sets are the same as those used in [ref 1](#). Data sets labeled “Stormo” were provided by Shuxiang Ruan and Gary Stormo via reanalysis of HT-SELEX data using their BEESEM method;⁷ for other cases, the CisBP position probability matrices were converted to estimated changes in binding free energy based on the log

ratio of the probability of each base versus the consensus sequence base at that position.

Hybrid Structure and Topology Generation. Throughout this work, we consider two sets of equilibrium MD trajectories, one generated using AMBER14 and one generated using Gromacs; as noted in the main text, the results obtained upon direct comparison were not sensitive to which set of equilibrium snapshots we used. AMBER-generated trajectories are used in all rows of [Table 2](#) except for “Gromacs, Gromacs struct”, whereas the Gromacs trajectories were used for that last row and for the generation of the full $\Delta\Delta G$ profiles.

For the AMBER-generated trajectories, a detailed description of the equilibrium MD simulation setup for DNA and Protein–DNA complex of the selected TFs including human ELK1, mouse GCM1, human MAX, and *S. cerevisiae* PPR1 and their DNA model systems in the presence of mutations is given in Khabiri and Freddolino.¹ From the equilibrium simulations, we extracted 32 evenly spaced snapshots from the last 50 ns of each trajectory for use in fast-growth thermodynamic integration (FGTI) calculations, as described below.

For the Gromacs equilibrium simulations and general parameters, all of the details match those in Gapsys and de Groot.²

AMBER Calculations. AMBER calculations were performed as in Khabiri and Freddolino,¹ except as noted below. In brief, all calculations were performed using the AMBER ff14SB protein force field¹⁹ and parmbsc1 DNA force field,²⁰ with TIP3P water, Joung-Cheatham parameters for potassium chloride,²¹ and specialized 7-site parameters for Mg²⁺.²² Simulations were performed using 2 fs timesteps with SHAKE bond constraints on all hydrogen-containing bonds; all production calculations were performed at constant volume and temperature (300 K) using a Langevin thermostat with a damping constant of 5 ps⁻¹. After generation of hybrid topology files for the snapshots of interest, the systems were equilibrated for 10 ps, with λ fixed at zero, and then we performed FGTI calculations spanning the range of $\lambda = 0.005$ to 0.995, with λ changed by 0.0001 once every 10 steps. We then calculated the work for each FGTI trajectory by numerical integration of recorded $\partial V/\partial\lambda$ values, with overall free energy changes and uncertainties calculated as in Khabiri and Freddolino.¹

Simulations designed to follow the original procedure were performed using the AMBER14 soft-core dual topology TI implementation for both electrostatic and Lennard-Jones interactions, with $\alpha = 0.2$ and $\beta = 9.0$. For the “new” method using an alternative switching function, we instead generated end point states in which all atoms of the decoupled bases at each end point were replaced with “dummy” particles that lacked all charges and Lennard-Jones interactions, thus, providing a true “ideal gas molecule” end point. In order to build dummy atoms, we removed all ordinary bonded (bond/angle/dihedral) parameters between the dummied segment of the residue and the constant atoms and replaced that deleted bond with a bond with 50 kcal mol⁻¹ Å⁻² force constant and 1.5 Å equilibrium length. In addition, dihedral angles across the newly modified bond were restrained to reference values with a force constant of 20 kcal mol⁻¹; these terms, in particular, maintain the fully dummied base pair in a conformation competent to form Watson–Crick hydrogen bonds when the nonbonded interactions of the participating atoms become active. In order to prevent major conformational changes in dummy atoms during transition and enhance convergence, a bond with force constant of 10 kcal mol⁻¹ Å⁻² and zero equilibrium length was introduced between N9 or N1 atoms of real bases to the equivalent atom in the dummy group to restrain the sugar-proximal atoms of the dummied bases. In transitions involving an incoming or outgoing thymine ring, special restraints were added for the dummied thymine methyl group: the bond between methyl carbon of the thymine ring and the carbon to which it is normally bonded in the ring system was altered to have a force constant of 100 kcal mol⁻¹ Å⁻² and a 0.5 Å equilibrium distance, to prevent end point catastrophes due to interaction of the protruding methyl group with water (we found in practice that without shrinkage of this bond, the conformations sampled by the fully decoupled thymine in the end point state could lead to severe clashes with solvent molecules upon transition to the $\lambda = 0.005$ or 0.995 states). In addition, the rotation of the methyl hydrogen atoms (with respect to the C4 and C6 atoms of the ring system) were restrained to reference values with a force constant of 20 kcal mol⁻¹, and the planarity of the methyl group with respect to the thymine ring system was restrained with a pair of

additional dihedrals, one involving C7–C5–C6–N1 atoms and one involving the C7–C5–C4–N3 atoms.

pmx/Gromacs Based Free Energy Calculations. Non-equilibrium alchemical free energy calculations based on the pmx/Gromacs framework followed the protocol described in Gapsys and de Groot.² To briefly summarize the approach, first, equilibrium simulations of 25 ns for the wild-type DNA and its mutated variants were performed considering separately the solvated DNA helix and DNA helix complexed with a protein. From the last 20 ns of each trajectory, 100 frames were extracted and subsequently used to initialize rapid 100 ps alchemical transitions driving the system from one physical state (wt/mut) to the other (mut/wt). Hybrid structures and topologies for the DNA nucleotides were generated with pmx.^{2,23} Equilibrium simulations were performed with the Gromacs version 5,²⁴ while for the alchemical transitions, Gromacs version 4.6, with a specialized soft-core function,²⁵ was used. Free energy differences were obtained from the forward/reverse work distributions based on the Crooks Fluctuation Theorem²⁶ using the maximum likelihood estimator.²⁷ The simulations were performed using the Amber99SB*ILDN-BSC1 force field.^{20,28–30} All the simulations were performed by retaining a constant temperature of 298 K by means of the velocity rescaling thermostat³¹ (time constant of 0.1 ps) and a constant pressure of 1 bar using the Parrinello–Rahman barostat³² (time constant of 5 ps). All bonds were constrained using LINCS algorithm.³³

RESULTS

To illustrate the effects of the decoupled group treatment on the protein–DNA binding free energy calculations that are of present interest, we implemented an alternative alchemical transformation in AMBER14, in which the decoupled groups lose all nonbonded interactions and have additional restraints applied to improve end point convergence. In the approach that we apply here, all atoms in the incoming/outgoing groups are still present at both end points, but the decoupled atoms have been converted to dummy particles. Because the AMBER softcore implementation can only be used with the pseudogas phase end point representation defined above, and the normal AMBER switching function for nonsoftcore calculations can be used only with outgoing, but not incoming, atoms,³⁴ we implemented an alternative switching function proposed in Steinbrecher et al.³⁵ in the sander module of AMBER14 to permit a TI calculation with complete dummied (removing all the nonbonded interactions) of incoming/outgoing atoms at the appropriate end points. The switching rule thus defined, following eq 7 of Steinbrecher et al.,³⁵ is

$$f(\lambda) = (1 - \lambda)^k \sum_{i=0}^{k-1} g(k - 1 + i, i) \lambda^i$$

with

$$g(x, y) = \frac{x!}{(x - y)!y!}$$

with $k = 6$ for all work described here. It is important to note that in this implementation many bonded contributions of the incoming/outgoing groups are retained, but the removal of all of their nonbonded interactions (as well as most angle and dihedral terms spanning the junction between the common and dummied atom groups, as detailed in Methods) breaks the coupling of their dynamics across the base pair.

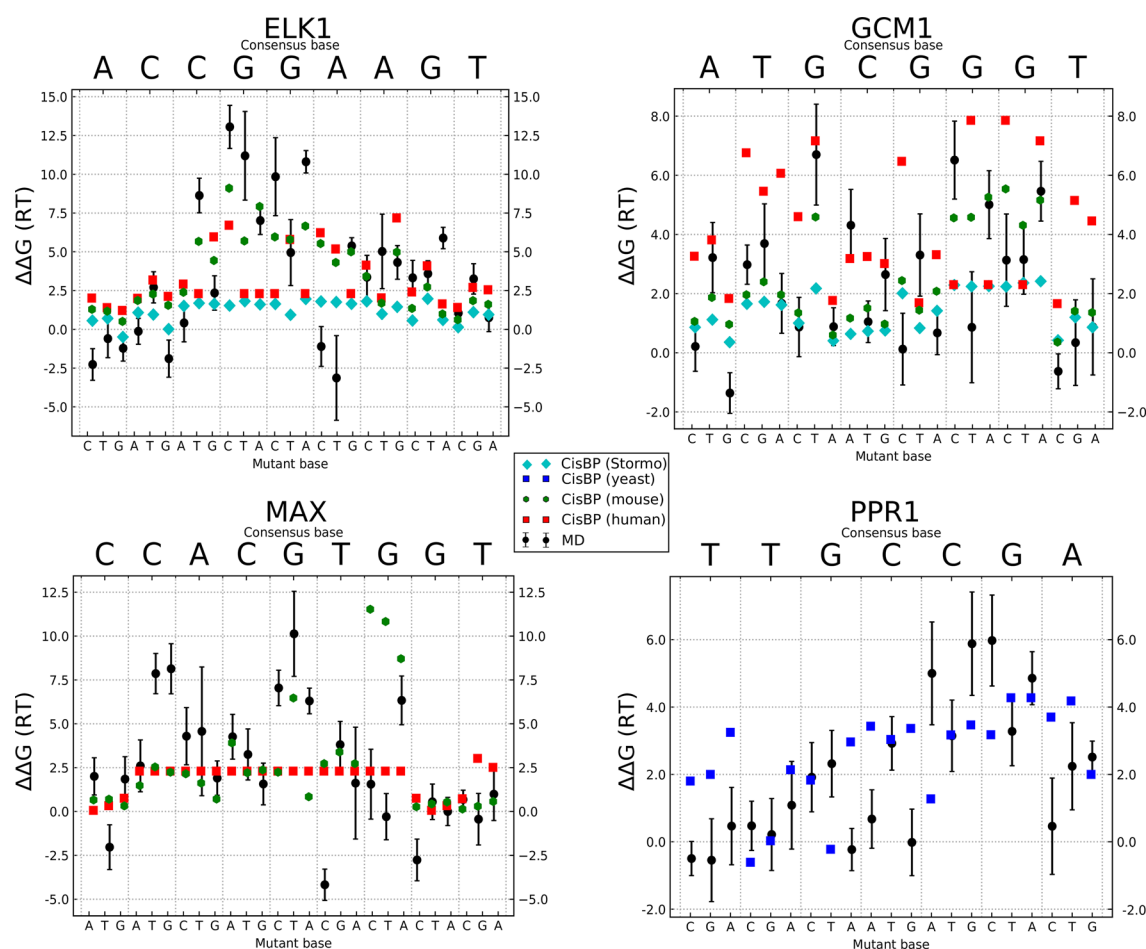


Figure 2. Comparison of calculated and experimental binding free energy changes ($\Delta\Delta G$) for all single base pair perturbations of the transcription factors considered here, with MD data obtained using the pmx approach. Values for comparison are taken from human, mouse, or yeast experimental data, as described in [Methods](#). Error bars for the MD data show an approximate 95% confidence interval based on bootstrap estimates of the standard error of the mean and a normality approximation.

We show in [Table 2](#) the binding free energy results obtained via nonequilibrium TI calculations using either the standard AMBER dual-topology softcore implementation (as in Khabiri and Freddolino¹), the Gromacs implementation of Gapsys and de Groot,² or our new AMBER implementation. We observe quantitative agreement between the new AMBER calculations and the Gromacs single topology implementation, regardless of which set of equilibrium trajectories was used; on the other hand, any implementation of alchemical transformation which permits nonbonded interactions across the base pair gives erroneous results. Results shown in [Table 2](#) demonstrate the unphysically large values for binding free energy changes observed using the protocol from Khabiri and Freddolino;¹ we also observed similarly unreliable results whether the standard AMBER formalism was used as implemented in the sander or pmemd programs, whether the entire nucleotides or only the bases were part of the perturbed group. Indeed, even the use of a single-topology procedure where the “dummied” atoms were decoupled according to the AMBER softcore formalism yielded unphysically large values for the changes in binding free energy (data not shown). On the other hand, use of the newly developed fully decoupled AMBER implementation yielded results in agreement with those obtained using the pmx/Gromacs implementation, irrespective of whether the

perturbed groups consisted of only the bases or the entire nucleotides.

Recent work by Loeffler et al.³⁶ lead to the identification of a bug in the AMBER softcore implementation, in which some 1–4 nonbonded interactions were not properly being scaled with the coupling parameter over the course of an alchemical transformation. We thus repeated our fast-growth thermodynamic integration calculations using a patched version of AMBER18 (provided by Taisung Lee and Darrin York) with the softcore interactions corrected; the results are shown on the “AMBER18, patched” line of [Table 2](#). While the corrections to softcore interaction handling clearly have an impact, the agreement with experimental results in the patched version is not substantively improved; we thus suspect that the deeper issue of strong coupling of the dynamics of the decoupled atoms with the common atoms that we have outlined above is more to blame, given that removal of all nonbonded interactions of the decoupled atoms (our “AMBER, mod” calculation) appears to correct the issue.

As a result of the methodological issues identified here, the sections of the original paper (Khabiri and Freddolino¹) pertaining to the binding free energy changes calculated using the old method (that is, Figures 3 and 4 in ref 1) should be considered unreliable, as they suffer from the technical issues outlined in this Comment. On the other hand, all insights

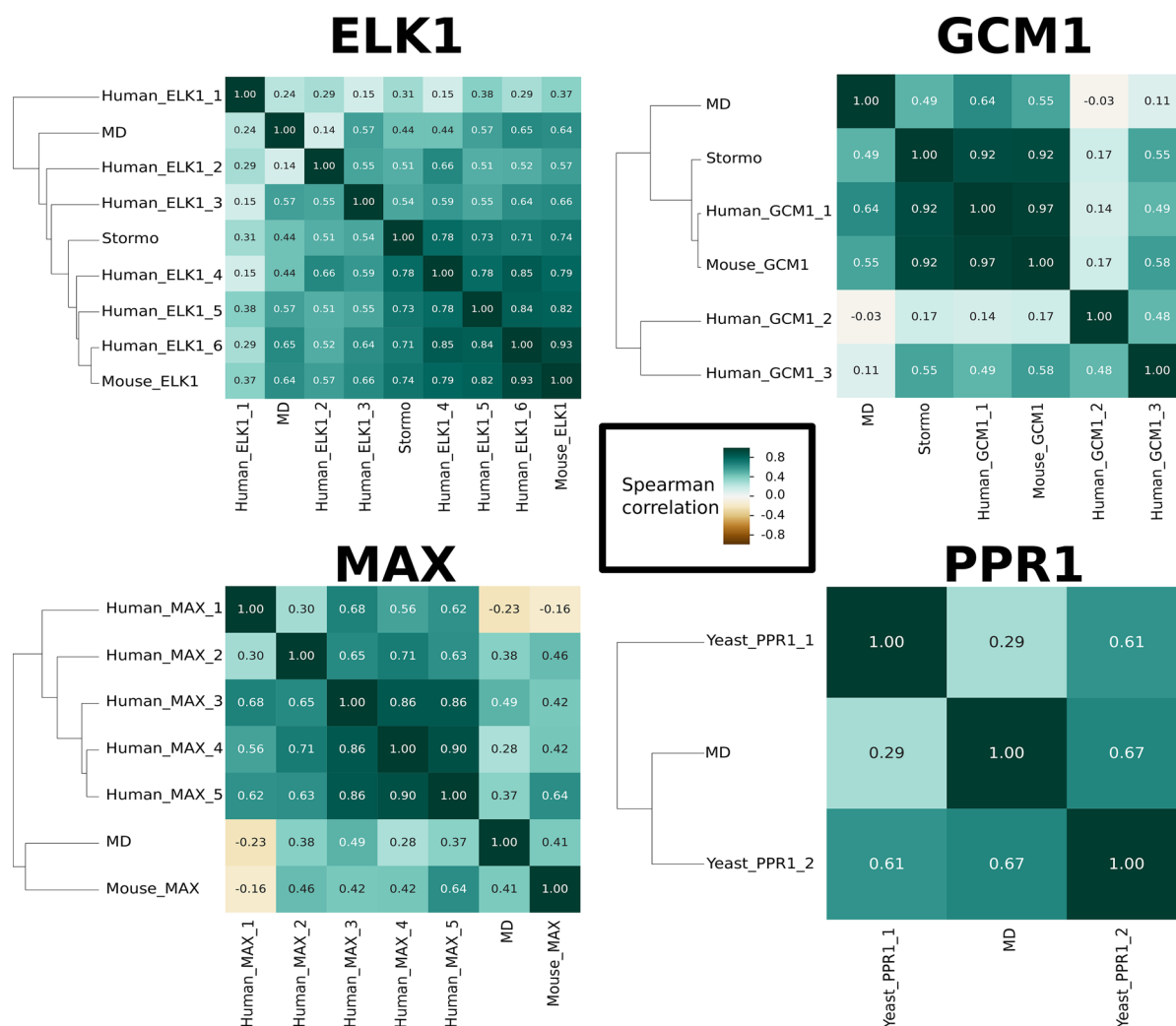


Figure 3. Matrices showing the Spearman correlations between the binding free energy differences obtained using the pmx-based simulations (MD) and a variety of experimental data sets (see Table 1 for details). Rows and columns are hierarchically clustered using UPGMA clustering.

pertaining to the results from long equilibrium MD simulations obtained in Khabiri and Freddolino¹ are still valid.

Full Results Obtained Using a Corrected Binding Free Energy Implementation. We have demonstrated above that when end point states are properly treated, consistent results for mutation-dependent binding free energy changes can be obtained using multiple distinct computational pipelines (e.g., either the corrected AMBER-based workflow described above, or the Gromacs-based pmx implementation of Gapsys and de Groot²). We have thus recalculated all of the binding free energy changes originally reported in ref 1 using the pmx workflow, in order to observe both the level of performance that can be obtained for the target transcription factors using a state-of-the-art implementation and determine whether additional insight can be obtained by cross-referencing the observed binding free energy profiles with characteristics of the protein–DNA complex, or the DNA alone, from equilibrium simulations (as we attempted to do in ref 1).

We show in Figure 2 the complete landscapes of sequence-dependent binding free energy changes ($\Delta\Delta G$) for the four proteins considered here, analogous to Figure 3 of ref 1. Compared with the results using the (flawed) dual-topology TI approach, it is immediately apparent that the unreasonably large binding free energies previously calculated¹ have been

replaced with distributions match the ranges of double free energy differences reconstructed from experimental profile measurements. In establishing expectations for correlation with different experimental data sets from different species, it is important to consider the species of origin and precise fragment of each simulated protein as well. In all cases, the simulated proteins match the crystal structures used (Table 2 of Khabiri and Freddolino¹) and, thus, are typically truncated relative to experimentally used constructs, but comprise at least the entire DNA binding domain in each case. The simulated ELK1 crystal structure was taken from the human protein, but human and mouse ELK1 proteins are identical within the region of the crystal structure. For GCM1, the mouse protein was used in simulations and is 94% identical to the human protein within the region used in simulations (with no difference in the region directly contacting the DNA). In the case of MAX, the crystal structure has an N-terminal extension that is slightly different from the original protein sequence, but is otherwise identical to both the human and the mouse proteins over the region included in simulations. For the *S. cerevisiae* protein PPR1, the fragment used in MD simulations directly matches the DNA binding domain of the yeast protein.

Prior to comparing the calculated and experimental $\Delta\Delta G$ values, it is important to note that the experimental double free

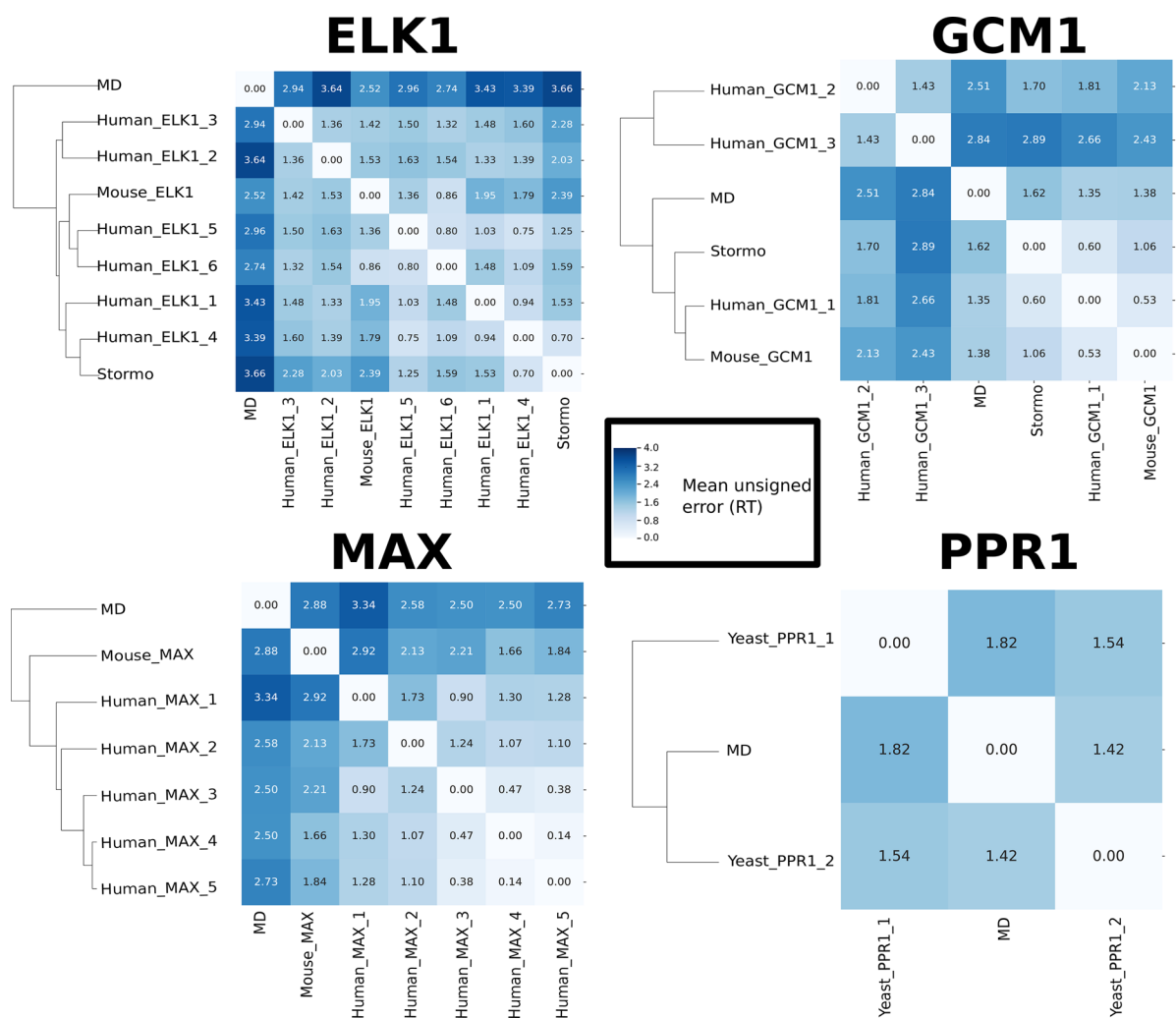


Figure 4. Matrices showing the mean unsigned error (in units of RT) between the binding free energy differences obtained using the pmx-based simulations (MD) and a variety of experimental data sets (see Table 1 for details). Rows and columns are hierarchically clustered using UPGMA clustering.

energy differences were reconstructed from the position probability matrices of the DNA–protein binding profiles. Naturally, free energy change for a mutation of a base pair at a given position integrates the information on that particular mutation in the context of all possible nucleotide neighbors of that base pair. In the current calculation setup, the neighbors of the nucleotides being mutated were kept fixed; therefore, the comparison between computation and experiment can only be viewed as approximate.

By considering the direct correlations of calculated and experimental free energies (Figure 3), we observe generally strong correlations between the computationally obtained values and experimental data sets for all cases; in particular, for the cases of GCM1, PPR1, and ELK1, the MD results cluster with the bulk of experimental data sets to the exclusion of at least one outlier experimental data set. We reach similar conclusions considering the average (mean) unsigned error across all sites as a metric to compare data sets; as shown in Figure 4, where performance is particularly strong for the GCM1 and PPR1 cases (where the MD results cannot be distinguished from the distribution of values observed across different experimental data sets). By the unsigned error measure, the MD results do appear as outliers relative to the

experimental data sets for ELK1 and MAX, apparently because the magnitudes of the calculated $\Delta\Delta G$ values are somewhat exaggerated in these two cases. Nevertheless, both the ranges of correlation coefficients and mean unsigned errors observed for the four cases considered here are well within the range of those noted across the 16 systems studied in Gapsys and de Groot² (e.g., Figure 4 of that paper). We see no notable species-specific differences between the agreement of the MD results with experimental data obtained on the human versus mouse proteins, consistent with the very high identity in the simulated regions in each case. The relatively poor agreement for MAX could potentially indicate that the approximation required for the direct comparison of the computed $\Delta\Delta G$ and the values reconstructed from the experimental binding profiles is violated. It is likely that, for this particular case, a reconstruction of the binding profile from the calculated data might be required considering conditional probabilities of the neighboring nucleotides, as demonstrated by Gapsys and de Groot.²

We further performed an analysis of the correlations of binding free energy differences with the structural characteristics of the equilibrium simulation, including the structural characteristics considered in ref 1 as well as additional features

such as the hydrogen bonding frequencies between the protein, bases, and DNA backbone. We observed no significant, systematic correlations; thus, while we are able to empirically reproduce experimentally observed binding free energy landscapes to within the bounds of uncertainty between experimental methods, no simple conceptual framework emerges to explain the observed binding free energy landscapes on the basis of structural observables.

CONCLUSIONS

While calculations described in Khabiri and Freddolino¹ were performed using fast-growth thermodynamic integration, the vulnerability in free energy calculations that we have identified could, in principle, arise in any circumstance in an alchemical free energy calculation where conformational coupling of incoming/outgoing atoms to the dynamics of the overall system causes a deviation from the assumed cancellation of free energy contributions from those atoms in different states, depending on the specific implementation of alchemical transformation. We have further shown that in one particular pathological case, where the entering/leaving atoms constitute base-paired nucleotides, the conformational coupling thus induced is sufficient to severely perturb binding free energy calculations. In any case, where problematic coupling might occur, an approach that removes all nonbonded contributions of decoupled atoms appears likely to provide improved accuracy, as is available in the Gromacs simulation package (demonstrated here) or the NAMD simulation package (with appropriate settings, as noted above) or as can be achieved with suitable modification of the nonbonded interaction treatment in AMBER.

Using appropriate methods for calculating binding free energy changes, we have shown that the computed double free energy differences can be correlated to the $\Delta\Delta G$ values reconstructed from the experimental binding profiles. Even though, in the current investigation, the experimental free energy differences were approximated from the position probability matrix without explicitly taking into account neighbor contribution, for two out of the four cases considered here (GCM1 and PPR1), the calculated free energy differences accurately matched the approximated experimental values.

In the other two cases (ELK1 and MAX), the calculated binding free energy profiles showed reasonable rank correlations with most experimental data sets, but had mean unsigned errors on the order of $\sim 3 RT$, placing them somewhat outside the bounds of the observed variations between experimental data sets (possibly indicating a failure of the approximation in casting the experimental position probability matrix into $\Delta\Delta G$ values without explicitly considering the contributions from neighboring nucleotides). Thus, consistent with the conclusions of Gapsys and de Groot,² but contrary to those expressed in Khabiri and Freddolino,¹ we find that MD simulations are now able to reproduce the binding free energy landscapes of DNA-binding proteins to the realm of uncertainties between experimental methods.

Admittedly, there is room for improvement in the reliability of such comparisons between calculations and experiment. One way to assess the current accuracy achievable by MD based calculations would be to compare the computed $\Delta\Delta G$ values to the experimental measurements of free energy differences, as shown by Gapsys and de Groot.² Another possibility, is to recalculate full binding profiles, preferably

considering effects of neighboring nucleotides, and comparing them to the experimental binding profiles. The latter approach, however, requires substantial computational effort, as the number of combinations to probe increases when considering more neighboring nucleotides.²

Moreover, computational accuracy can further be increased. As demonstrated previously,^{2,37} combining $\Delta\Delta G$ estimates from the newest generation of molecular mechanics force fields from the Amber and CHARMM families, could improve agreement with experimental measurements.

As suggested in Gapsys and de Groot,² further refinement of calculation capabilities may provide a useful benchmark for ongoing refinements of modern force fields; in addition, there may be as-yet un-noticed aspects of the relatively poorly performing cases (e.g., structural quality, degree of conformational changes in response to mutation, precise choice of the portion of the protein of interest being simulated relative to that used in experiments, etc.) that might suggest either refinements to simulation protocols, or aspects of the starting structure to provide sufficient quality for accurate calculation of the binding free energy landscape.

On May 18, 2018 an Expression of Concern was posted on the original article by Khabiri and Freddolino (ref 1). The Expression of Concern is considered resolved upon publication of this Comment. The Expression of Concern has been appended here as [Supporting Information](#).

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jpcc.8b04187>.

A copy of the "Expression of Concern" document published for the original article by Khabiri and Freddolino (ref 1; [PDF](#))

AUTHOR INFORMATION

Corresponding Authors

*E-mail: bgroot@gwdg.de.

*E-mail: petefred@umich.edu.

ORCID

Vytautas Gapsys: 0000-0002-6761-7780

Bert L. de Groot: 0000-0003-3570-3534

Peter L. Freddolino: 0000-0002-5821-4226

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

M.K. and P.L.F. were supported by NIH R35-GM128637, with computer time provided by the Extreme Science and Engineering Discovery Environment³⁸ (XSEDE Project TG-MCB140220), which is supported by National Science Foundation Grant No. ACI-1548562. V.G. and B.d.G. were supported by BioExcel CoE (www.bioexcel.eu), a project funded by the European Union Contracts H2020-INFRAEDI-02-2018-823830 and H2020-EINFRA-2015-1-675728. P.L.F. is grateful to Taisung Lee and Darrin York for providing a patched version of AMBER18 to enable testing of the effect of recent AMBER updates on the calculations described here.

REFERENCES

- (1) 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.
- (2) Gapsys, V.; de Groot, B. L. Alchemical Free Energy Calculations for Nucleotide Mutations in Protein–DNA Complexes. *J. Chem. Theory Comput.* **2017**, *13*, 6275–6289.
- (3) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (4) Boresch, S.; Karplus, M. The Role of Bonded Terms in Free Energy Simulations: 1. Theoretical Analysis. *J. Phys. Chem. A* **1999**, *103*, 103–118.
- (5) Boresch, S.; Karplus, M. The Role of Bonded Terms in Free Energy Simulations. 2. Calculation of their Influence on Free Energy Differences of Solvation. *J. Phys. Chem. A* **1999**, *103*, 119–136.
- (6) 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.; et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **2014**, *158*, 1431–1443.
- (7) Ruan, S.; Swamidass, S. J.; Stormo, G. D. BEESEM: Estimation of Binding Energy Models Using HT-SELEX Data. *Bioinformatics* **2017**, *33*, 2288–2295.
- (8) Yin, Y.; Morgunova, E.; Jolma, A.; Kaasinen, E.; Sahu, B.; Khund-Sayeed, S.; Das, P. K.; Kivioja, T.; Dave, K.; Zhong, F.; et al. Impact of Cytosine Methylation on DNA Binding Specificities of Human Transcription Factors. *Science* **2017**, *356*, eaaj2239.
- (9) Jolma, A.; Yan, J.; Whittington, T.; Toivonen, J.; Nitta, K. R.; Rastas, P.; Morgunova, E.; Enge, M.; Taipale, M.; Wei, G.; et al. DNA-binding Specificities of Human transcription Factors. *Cell* **2013**, *152*, 327–339.
- (10) Kulakovskiy, I. V.; Medvedeva, Y. A.; Schaefer, U.; Kasianov, A. S.; Vorontsov, I. E.; Bajic, V. B.; Makeev, V. J. HOCOMOCO: A Comprehensive Collection of Human Transcription Factor Binding Sites Models. *Nucleic Acids Res.* **2013**, *41*, D195–D202.
- (11) Matys, V.; Kel-Margoulis, O. V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; et al. TRANSFAC and its Module TRANSCOMP: Transcriptional Gene Regulation in Eukaryotes. *Nucleic Acids Res.* **2006**, *34*, D108–D110.
- (12) Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y. C.; Laslo, P.; Cheng, J. X.; Murre, C.; Singh, H.; Glass, C. K. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **2010**, *38*, 576–589.
- (13) Wei, G.-H.; Badis, G.; Berger, M. F.; Kivioja, T.; Palin, K.; Enge, M.; Bonke, M.; Jolma, A.; Varjosalo, M.; Gehrke, A. R. Genome-wide Analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.* **2010**, *29*, 2147–2160.
- (14) Isakova, A.; Groux, R.; Imbeault, M.; Rainer, P.; Alpern, D.; Dainese, R.; Ambrosini, G.; Trono, D.; Bucher, P.; Deplancke, B. SMiLE-seq Identifies Binding Motifs of Single and Dimeric Transcription Factors. *Nat. Methods* **2017**, *14*, 316–322.
- (15) 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.; et al. Diversity and Complexity in DNA Recognition by Transcription Factors. *Science* **2009**, *324*, 1720–1723.
- (16) Gerstein, M. B.; Kundaje, A.; Hariharan, M.; Landt, S. G.; Yan, K.-K.; Cheng, C.; Mu, X. J.; Khurana, E.; Rozowsky, J.; Alexander, R.; et al. Architecture of the Human Regulatory Network Derived from ENCODE Data. *Nature* **2012**, *489*, 91–100.
- (17) de Boer, C. G.; Hughes, T. R. YeTFaSCO: A Database of Evaluated Yeast Transcription Factor Sequence Specificities. *Nucleic Acids Res.* **2012**, *40*, 169–179.
- (18) Roy, A.; Exinger, F.; Losson, R. cis- and trans-acting Regulatory Elements of the Yeast URA3 Promoter. *Mol. Cell. Biol.* **1990**, *10*, S257–S270.
- (19) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ffl4SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.
- (20) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; et al. Parmbsc1: A Refined Force Field for DNA Simulations. *Nat. Methods* **2016**, *13*, 55–58.
- (21) 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.
- (22) Saxena, A.; Sept, D. Multisite Ion Models that Improve Coordination and Free Energy Calculations in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 3538–3542.
- (23) 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.
- (24) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1*, 19–25.
- (25) 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.
- (26) 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.
- (27) 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.
- (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) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
- (32) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.
- (33) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINC: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.
- (34) Case, D.; Babin, V.; Berryman, J.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham, T., III; Darden, T.; Duke, R. et al. AMBER 14; University of California, San Francisco.
- (35) Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear Scaling Schemes for Lennard-Jones Interactions in Free Energy Calculations. *J. Chem. Phys.* **2007**, *127*, 214108.
- (36) Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of Free Energy Calculations Across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14*, 5567–5582.
- (37) 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.
- (38) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazelwood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; et al. XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* **2014**, *16*, 62–74.