# Structural Predictions of Protein-DNA Binding: MELD-DNA

Reza Esmaeeli[1,*], Antonio Bauzá[2,*], and Alberto Perez[**1]

[1] Department of Chemistry, Quantum theory project, University of Florida, Gainesville, FL, 32611 USA

[2] Department of Chemistry, Universitat de les Illes Balears, Palma de Mallorca (Baleares), 07122, SPAIN

* Both authors contributed equally to this work

** To whom correspondence should be addressed. Tel: +1 (352) 3927009; Email: perez@chem.ufl.edu

## ABSTRACT

Structural, regulatory, and enzymatic proteins interact with DNA to maintain a healthy and functional genome. Yet, our structural understanding of how proteins interact with DNA is limited. We present MELD-DNA, a novel computational approach to predict the structures of protein-DNA complexes. The method combines molecular dynamics simulations with general knowledge or experimental information through Bayesian inference. The physical model is sensitive to sequence-dependent properties and conformational changes required for binding, while information accelerates sampling of bound conformations. MELD-DNA can: i) sample multiple binding modes, ii) identify the preferred binding mode from the ensembles, and iii) provide qualitative binding preferences between DNA sequences. We first assess performance on a dataset of 15 protein-DNA complexes and compare it to state-of-the-art methodologies. Furthermore, for three selected complexes, we show sequence dependence effects of binding in MELD predictions. We expect the results presented herein, together with the freely available software, will impact structural biology (by complementing DNA structural databases) and molecular recognition (by bringing new insights into aspects governing protein-DNA interactions).

## INTRODUCTION

Understanding and predicting how proteins and nucleic acids interact is key to deciphering the mechanisms regulating gene expression, genome repair, and storage; with applications in fields such as nanomedicine(1), transition metal chemistry(2), or clinical diagnosis(3) to name a few. Predicting such molecular recognition problems typically fall under molecular docking, machine learning, and molecular dynamics-based studies, which have been broadly successful in understanding how proteins recognize small molecules, peptides, and other proteins(4–6). However, the nature of protein-DNA interactions introduces several nuances that have challenged standard approaches. First, whereas proteins come in diverse shapes and sizes, the double-stranded B-DNA structure is common to most DNA sequences leading for many years to the concept of proteins “reading” DNA. Second, the highly charged interactions from the repeating phosphate backbone lead to a particular protein interacting with high affinity with many different DNA sequences. Despite the high affinity, specificity(7) to different sequences can span several orders of magnitude, leading to a preferential binding for certain sequences. Transcription factor (TF) proteins regulate gene expression by binding DNA sequences between 6 and 12 base pairs in length, statistically found thousands to millions of times along the genome. Still, most of these sites are never experimentally occupied(8, 9). TF binding has received particular attention due to its biological role; thus, we will focus on those. Understanding TF protein-DNA interactions requires answering three distinct questions : 1) what structure will a particular protein-DNA complex adopt, 2) what sequence will a particular TF preferentially bind (e.g., where along the genome will the complex form), and 3) what are the relative binding affinities to different sequences. Pipelines combining elements from structural databases, molecular dynamics, docking, and machine learning are becoming more prominent to address some (or all) of the above questions (10, 11).

While the initial views on protein-DNA recognition focused on the ability of proteins to "read" the DNA sequence and find the best binding site, current views show DNA has a much more prominent role in recognition(9, 12). Thus, some proteins interact with DNA through sequence-specific interactions, but many interact through shared DNA features (e.g., the phosphate backbone). The difference in binding patterns leads to binding mechanisms lying between two extremes: sequence readout and shape readout(9, 13–16). The first is governed by specific interactions, while the second accounts for the ability of a particular DNA sequence to adopt conformations compatible with the structure of the complex. Attempts at combining structural descriptors (e.g., average groove widths, propeller twist, or roll) combined with sequence preferences derived from genomic studies (17–19) significantly improve predictions of where TF proteins bind along the genome.  However, attempts to use detailed structural approaches, such as free energy perturbation or energy decomposition, to distinguish shape and sequence contributions to binding are not always successful, limiting their general application (20–22). Although the reason for failures are unclear, factors like the amount of DNA deformation in different transcription factors, the number of binding modes contributing to the binding free energy, and force field accuracy are suspect(22).

We require initial structures of the complex for many studies involving an understanding of binding affinity predictions, binding mechanisms, or even DNA-mediated allostery. However, there is a lack of experimental structural data and few computational methods to predict such complexes accurately. For instance, there are 6052 protein-DNA structures in the Nucleic Acid Database, compared to 171,077 protein structures in Protein Data Bank (PDB) as of October 2022(23, 24). Such small datasets pose challenges to adapting protein structure prediction approaches (e.g., AlphaFold (25)) to the problem of predicting nucleic acids and their complexes. The recent RoseTTAFoldNA(26) ML approach compensates for the smaller datasets (they report 1556 protein-nucleic acid complex clusters compared to 26128 all protein clusters) by incorporating physics-based parameters (e.g., van der Waals terms) originating from Rosetta. Despite the advances this method represents, especially in RNA structure prediction, it requires further development to correct anomalies in predicting structures of homodimers bound to DNA (prevalent in transcription factors), which leads to an overlap between monomer units and incorrect binding modes.

For many decades, docking has been the most efficient technique for predicting atomic resolution structures of macromolecular binding.(27) Docking is an invaluable tool for the virtual screening of small molecule libraries in the early stages of drug discovery(28) thanks to the balance between computational efficiency and accuracy. For larger assemblies, community efforts like the Critical Assessment of PRediction of Interactions (CAPRI)(29) have led to significant improvements in the field. However, the accuracy of docking drops rapidly for systems involving conformational changes and in charged systems, where scoring functions are less reliable. Thus, docking methods will often combine with strategies such as normal modes or a later stage using molecular dynamics to overcome limitations in scoring functions(30, 31). The most successful efforts come from the High-Ambiguity Driven Docking (HADDOCK)(32, 33) group combining docking with ambiguous data and DNA flexibility (via normal modes). They have led to helpful benchmark sets of different difficulties for assessing protein-DNA docking performance(34).

Knowing the structure of the complex opens the possibility of predicting relative binding affinities to different sequences. Alchemical free energy methods based on Molecular Dynamics simulations (MD) are the gold standard for predicting relative (and absolute) binding affinities for small molecules(35). However, they present limitations when dealing with electrostatic charge variations and flexible complexes where several binding modes can contribute to the binding free energy(36). Despite some successes for protein-DNA systems(37, 38), a recent systematic study(39) on protein-DNA complexes points to deficiencies of these approaches arising from either phase space overlap or force field issues. Thus, current methods based on physics and statistical potentials like Rosetta(40) or machine learning(41–44) far outperform traditional MD-based approaches in speed and accuracy.

In this paper, we introduce a framework based on Molecular Dynamics that incorporates information via Bayesian inference to predict structure-sequence relationships to increase our understanding of how proteins bind nucleic acids. The approach, MELD (Modeling Employing Limited Data)(45, 46), has been previously used for predicting protein structures and their complexes with small molecules, peptides, and proteins. Here we extend the framework to DNA (MELD-DNA), with different applications depending on the type of data used. This work exemplifies three of the most common uses: 1) predicting structures of protein-DNA complexes, 2) identifying sequence sensitivity, and 3) predicting relative binding affinities. We expect the first application to be the most broadly used and thus present a generalized protocol over fifteen different proteins. We make more specific comparisons over a series of fifteen sequences for three of those systems. Finally, we exemplify the application of relative binding affinities for six sequences binding a particular TF. The current work represents an extensive simulation study with an aggregated 1 ms of sampling. Our results show that MELD-DNA successfully: i) samples multiple binding modes, ii) identifies the experimental binding mode through clustering the ensembles, and iii) is sensitive to DNA sequences and conformations.

## MATERIAL & METHODS

### General Approach

We use the MELD (Modeling Employing Limited Data) Bayesian inference approach (*$p(x|D)$ α $p(D|x)$ • $p(x)$*) to incorporate ambiguous and noisy data to enhance binding/unbinding events(45, 46). The prior distribution (*$p(x)$*) is given by the Boltzmann distribution based on the chosen force field, while the likelihood (*$p(D|x)$*) comes from the agreement of the sampled conformations (*x*) with a subset of the data (*D*, the one with the lowest restraint energy). As MELD samples the energy landscape, different subsets of data are also explored, exploiting regions compatible with some subset of data and the force field(45, 46) gives rise to the posterior distribution (*$p(x|D)$*). In practical terms, MELD uses a Hamiltonian and Temperature replica exchange molecular dynamics approach in which some replica conditions are compatible with unbound states and some with bound states. As *walkers* sample different conditions in the replica ladder, they go through cycles of binding and unbinding. We identify bound states by clustering the lowest temperature ensembles, where each cluster represents a different binding mode and is compatible with varying subsets of data. We will showcase here three protocols to address three questions: i) general binding (applied to any protein-DNA system), ii) specific binding (applied to many DNA sequences binding a particular protein where additional information is known), and iii) relative binding affinities. The type of data used to guide simulations depends on the questions we ask. Examples are accessible from Zenodo (see Data availability section). MELD simulations use 30 replicas, the parmBSC1 force field for nucleic acids(47–49) the ff14SB side force field for the protein(50, 51) and the GBneck2Nu implicit solvent model(52, 53). Throughout all protocols, we include restraints to keep the protein and DNA from unfolding at high temperatures. For proteins, we enforce secondary structure and flat-bottom harmonic restraints on native Cα-Cα contacts, the initial coordinates for simulations and to set up the restraints is based on its bound conformation. For DNA, we implement restraints that maintain hydrogen bonding patterns at each base pair to prevent DNA melting. All simulations were initialized with the protein far away (at least 30Å) from the DNA. The initial DNA conformation is generated in its canonical B-form based on the sequence (54).

**Protocol 1: Posing general knowledge in terms of ambiguous data drives protein-DNA structure prediction**

In this approach, we seek to explore multiple binding modes and rely on statistical mechanics to identify the most native-like one (e.g., the most compatible with our physical model). We presume knowledge of 1) the protein structure, 2) the DNA sequence to bind, and 3) the DNA binding domain. We generate a B-DNA structure using Chimera (54) and create a dummy atom at the N1 position of each purine base (see Fig. 1). We define the binding data as the possible interactions between Cα atoms in the binding domain and the N1 atoms (Fig. 1A). We produce a list of potential contacts, where only some might be satisfied during binding (noisy data) (Fig. 1B). We reduce the amount of possible combinatorics by taking into consideration geometric considerations (e.g., residues far away in the binding site are unlikely to interact with the same DNA base simultaneously). Clustering on the MELD ensemble we identify native-like poses in the ensemble (see Fig. 1). The current setup has two advantages: 1) by using dummy particles at the N1 site, we do not favor the protein approaching through either major or minor groove orientations, 2) the information added is not exhaustive of all possibilities – and it does not need to be, as the force field will sample the most likely conformations given the available data (Fig. 1C).

We chose fifteen protein-DNA systems to apply this approach (see Table 1). The systems include complexes with little or no deformation of the DNA from its canonical B-DNA form, others that induce moderate deformation upon binding, and complexes where the DNA is far from its canonical B-DNA conformation. The dataset also contains systems that have been solved experimentally with two different sequences, resulting in binding mode variations (e.g., different spacing between binding domains: 1R4R, 1R4O). Finally, we include two types of systems intended to challenge our approach: binding occurs through either flexible (disordered) tails (1ZME) or where large conformational changes are needed for accessing the binding site (1BGB, 2B0D).

**Table 1.** Protein-DNA systems simulated in this study, along with their DNA sequence and PDB IDs.

| System | DNA Sequence | PDB | Ref. |
|---|---|---|---|
| Nuclear Intron-Encoded Homing Endonuclease I-Ppoi | TGACTCTCTTAAGAGAGTCA | 1A74 | (55) |
| Hyperthermophile Chromosomal Protein Sac7d | GCGATCGC | 1AZP | (56) |
| 9-Cis Retinoic Acid Receptor | TAGGTCAAAGGTCAG | 1BY4 | (57) |
| Human Papillomavirus Type-18 E2 | CAACCGAATTCGGTTG | 1JJ4 | (58) |
| Phage 434 Cro | AGTACAAACTTTCTTGTAT | 3CRO | (59) |
| Fungal Transcription Factor Put3 | CGGGAAGCCAACTCCG | 1ZME | (60) |
| Murine Creb Bzip-Cre Complex | CTTGGCTGACGTCAGCCAAG | 1DH3 | (61) |
| P22 C2 Repressor | CATTTAAGATATCTTAAATA | 2R1J | (62) |
| Human Tbp Core Domain | CTGCTATAAAAGGCTG | 1CDW | (63) |
| Gcn4 Leucine Zipper | TTCCTATGACTCATCCAGTT | 1YSA | (64) |
| Gcn4 Leucine Zipper | TGGAGATGACGTCATCTCC | 2DGC | (65) |
| Glucocorticoid Receptor | TCAGAACATGATGTTCTCA | 1R4R | (66) |
| Glucocorticoid Receptor | CCAGAACATCGATGTTCTG | 1R4O | (66) |
| Ecorv Restriction Endonuclease | CGGGATATCCC | 1BGB | (67) |
| Ecorv Restriction Endonuclease | AAAGAATTCTT | 2B0D | (68) |

**Protocol 2: System-specific binding protocols tease out sequence effects**

In this scenario, we have knowledge of the bound state (and binding mode) and use this information to guide to repeating binding sites along a DNA oligomer (see Fig. 2). By using oligomers with different sequences (but the same driving information), we are interested in teasing out the drivers of binding (sequence or shape readout). When the data is too constraining, we should see the same binding mode regardless of the sequence. We compare multiple sequences for three systems involving different degrees of DNA bending upon binding (see Table 2).

**Table 2.** DNA sequences used for bZIP, TATA, and P22 complexes. DNA bases highlighted in orange represent the changes with respect to the consensus sequence (bold).

| System | Sequence | | PDB | Ref. |
|---|---|---|---|---|
| **bZIP** | Consensus | CCTTGG**CTGACGTCAG**CCAAG | 1DH3 | (61) |
| | noCG | CCTTGGCTGA**AT**TCAGCCAAG | | |
| | Random 1 | CCTTGG**ATGCTACGAT**CCAAG | | |
| | Random 2 | CCTTGG**CGTAGCTCGG**CCAAG | | |
| | Random 3 | CCTTGG**TCTATCGGTT**CCAAG | | |
| **TATA box** | Consensus | CTGC**TATAAAA**GGCTG | 1CDW | (63) |
| | Domain 1 | CTGC**CGCG**AAAGGCTG | | |
| | Domain 2 | CTGCTATA**GGG**GGCTG | | |
| | Random 1 | CTGC**CGCGGGG**GGCTG | | |
| **P22 c2 Repressor** | Consensus | CATTT**AAG**ATAT**CTT**AAATA | 2R1J | (62) |
| | Domain 1 | CATTT**CCT**ATATCTTAAATA | | |
| | Domain 2 | CATTTAAGATAT**GGC**AAATA | | |
| | Bridge 1 | CATTTAAG**CGCG**CTTAAATA | | |
| | Bridge 2 | CATTTAAG**CCGA**CTTAAATA | | |
| | Random 1 | C**GCCATTTAGGGACGATC**CA | | |

### Protocol 3: Competitive binding simulations quantify relative binding affinities

The data used in this protocol is the most constraining: we aim to distinguish the preference of a particular known binding mode to two different sequences. Thus, the data is compatible with one binding mode. Rather than defining this strictly as in docking or an alchemical free energy calculation, there is still enough freedom to sample widely inside the basin corresponding to this binding mode. To compare across systems, we have previously used a competitive binding strategy (69) in which the relative binding affinity can be determined by counting the population of the protein-DNA complex for each sequence. Converging populations to obtain statistical significance makes this protocol more computationally demanding than the previous two. We thus exemplify this protocol on the six sequences binding P22 shown in Table 2.

### HADDOCK docking predictions

We used the bio3d R module to calculate normal modes of the minimized canonical B-DNA and minimized protein for each system(70, 71). The ensemble of normal modes for each binding partner was fed into the HADDOCK web server (72, 73), specifying identical residues involved in the contact lists of MELD as *active* residues. We then analyzed the top 5 clusters and all the models generated by the docking package.

### RosettaFold2NA machine learning predictions

The RosettaFold2NA (RF2NA) program was assembled following their GitHub walkthrough(26). The FASTA sequence of each protein chain and DNA strand was provided to the program as separate files according to the instructions. Each prediction returned one structure.

## RESULTS

### MELD-DNA is successful at predicting protein-DNA complexes

We analyze our ensembles for each of the 15 proteins by asking two questions (see Fig. 3): 1) can the method sample the native state, 2) can we identify the native state with high confidence without knowing the actual structure. We find that in 13 of the 15 cases the native state is present in the ensemble, and in 11 of 15 cases it belongs to a high population cluster (present in the top five clusters by population). The ensembles represent multiple binding modes highlighting MELD's ability to explore various bound conformations. Figures S1-15 represent the distribution of structures sampled at every replica in the MELD approach for each of the 15 complexes. At high replica indexes, the system explores unbound conformations (high RMSD values). Sampling at lower replicas explores different bound states – where typically, the lowest replica index will be enriched on the state with the lowest RMSD to the experimental structure. Complex 1AZP (see Fig. S2 and S16) represents a particular case since the DNA sequence is palindromic, and our methodology captures clusters binding in either orientation (180-degree flip of the protein, see Fig. S16). Similarly, most Leucine zippers correctly identify the binding mode, with the binding domains overlapping the experimental binding mode. For these long coil-coil structures, small fluctuations near the binding site give rise to a large displacement at the ends – hence some of these structures look visually distinct from the experimental structure but retain a low interface RMSD(e.g., 2GDG in Fig. 3). Despite the use of restraints to keep the protein and DNA from unfolding, we find that both macromolecules sample a broad ensemble of conformations. Typically, the DNA and protein ensembles approach the *holo* conformation as the native binding mode is sampled (see Fig. S17).

We examined predictions from HADDOCK and RF2NA for comparison. For HADDOCK, we first tried submitting only one DNA and protein structure as input for each complex. As expected, in this scenario, we did not observe any native-like predicted complexes (which MELD also fails to predict in the absence of data). We then used several structures for both the protein and DNA sampled along their normal modes as calculated by the *bio3D* package (70, 71, 74). HADDOCK returns 200 models of the complex by default and clusters them. We observed many native-like complexes amongst the 200 generated models (Table 3 and Figs. S18-19). However, the top 5 clusters generally had higher interface RMSDs and a lower fraction of native contacts(75)(76) than MELD predictions. In our hands, the RF2NA machine learning approach produced overlaps between the two monomers in protein dimers where both bind in the same DNA region (marked as MO in Table S1). For 1BY4, it used the two monomers of the protein to form a large monomer-like protein binding only on one of the two binding sites of the DNA. We are optimistic that future versions will be available to predict structures of homodimers binding DNA.

**Table 3.** Fraction of native contacts for MELD and HADDOCK (italics) from the top five clusters and the highest value sampled in the whole ensemble. Any two residues with at least one heavy atom pair within 5Å in the experimental structure were defined as a contact. We highlight in bold instances where 70% or more of the native contacts are satisfied.

| | | **Fraction of Native Contacts** | | | | | |
|---|---|---|---|---|---|---|---|
| **System** | **# contacts** | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Best in all |
| 1A74 | 189 | 0.60<br>*0.06* | **0.85**<br>*0.06* | 0.11<br>*0.05* | 0.51<br>*0.05* | 0.577<br>*0.0* | **0.93**<br>*0.57* |
| 1AZP | 46 | 0.35<br>*0.13* | 0.31<br>*0.14* | 0.29<br>*0.15* | 0.21<br>*0.08* | 0.68<br>*0.13* | **0.95**<br>*0.67* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1BY4 | 83 | **0.73**<br>*0.01* | **0.74**<br>*0.03* | 0.24<br>*0.05* | 0.26<br>*0.02* | 0.16<br>*0.02* | **0.80**<br>*0.67* |
| 1JJ4 | 109 | 0.15<br>*0.33* | 0.31<br>*0.41* | 0.06<br>*0.23* | 0.00<br>*0.28* | 0.02<br>*0.12* | 0.34<br>*0.58* |
| 3CRO | 123 | 0.38<br>*0.06* | 0.10<br>*0.4* | 0.13<br>*0.04* | 0.08<br>*0.06* | 0.16<br>*0.1* | 0.57<br>*0.62* |
| 1ZME | 110 | 0.17<br>*0.02* | 0.27<br>*0.06* | 0.09<br>*0.04* | 0.13<br>*0.06* | 0.15<br>*0.01* | 0.62<br>*0.49* |
| 1DH3 | 55 | **0.94**<br>*0.32* | 0.04<br>*0.38* | 0.11<br>*0.40* | 0.34<br>*0.24* | 0.49<br>*0.27* | **1.00**<br>*0.44* |
| 2R1J | 99 | **0.89**<br>*0.05* | 0.18<br>*0.03* | 0.30<br>*0.04* | 0.31<br>*0.02* | 0.10<br>*0.05* | **0.92**<br>*0.62* |
| 1CDW | 84 | 0.51<br>*0.08* | 0.20<br>*0.12* | 0.07<br>*0.04* | 0.18<br>*0.04* | 0.49<br>*0.07* | 0.69<br>*0.58* |
| 1YSA | 62 | 0.64<br>*0.38* | 0.58<br>*0.44* | 0.36<br>*0.26* | 0.67<br>*0.22* | 0.52<br>*0.05* | **0.73**<br>*0.51* |
| 2DGC | 72 | 0.50<br>*0.04* | 0.62<br>*0.04* | **0.89**<br>*0.05* | 0.47<br>*0.03* | 0.68<br>*0.21* | **0.98**<br>*0.36* |
| 1R4R | 83 | 0.50<br>*0.01* | 0.27<br>*0.01* | **0.80**<br>*0.01* | 0.29<br>*0.01* | 0.42<br>*0.02* | **0.86**<br>*0.26* |
| 1R4O | 79 | **0.79**<br>*0.02* | 0.45<br>*0.04* | 0.34<br>*0.02* | 0.34<br>*0.03* | 0.31<br>*0.03* | **0.86**<br>*0.69* |
| 1BGB | 220 | 0.09<br>*0.00* | 0.08<br>*0.00* | 0.02<br>*0.00* | 0.05<br>*0.00* | 0.12<br>*0.01* | 0.19<br>*0.10* |
| 2B0D | 227 | 0.20<br>*0.00* | 0.17<br>*0.00* | 0.07<br>*0.00* | 0.17<br>*0.00* | 0.19<br>*0.01* | 0.24<br>*0.18* |

**MELD is sensitive to sequence properties**

We chose three systems representative of binding DNA in its canonical conformation, bent DNA, and significantly bent DNA (bZIP, p22, and TATA box binding proteins, respectively). For each of the systems, we simulated the consensus sequence, for which we have an experimental structure of the complex. We then generated the sequences defined in table 2, for which we have no experimental data. The consensus sequence is the highest affinity binder of all sequences. However, other sequences are also likely to bind based on charge complementarity – especially in cases with no specific interactions between protein residues and DNA nucleobases. Hence, we expect differences in the ensemble, either by binding at different sites along the sequence or affecting the populations identified by clustering with respect to the consensus sequence (see Fig. 4).

For all three cases, MELD correctly identifies the native binding mode when the consensus sequence is used -- but we differentiate two different behaviors when changing the sequence. For p22 and TATA, binding is driven by shape complementarity, and in all simulations, we observe a very similar binding pattern, with all DNA sequences adapting bent conformations. However, for bZIP, which binds DNA in its canonical B-DNA conformation, we observe dramatical shifts in cluster populations and binding preferences, with some sequences preferring to bind different sites along the sequence (see

Fig. 4A). To further study the sequence/structure relationship, we took the top two clusters from unbiased MD simulations of each protein-free DNA sequence, and we performed binding simulations restraining the conformation of each sequence to each of the possible clusters (a total of 50 simulations). Figure S20 shows the RMSD of the protein *vs.* the protein-DNA interface RMSD. While some structure/sequence combinations lead to binding, others do not. For the consensus sequence, the experimental binding mode is sampled in high populations in 9/10 simulations, with the highest population cluster falling in this region most of the time. Other sequences have a lower preference for the canonical binding site despite starting from the same conformation and using the same data. Conformation *random1-c1* (random1 sequence and cluster 1 conformation) merits a special mention: only using the sequence from random1 does this conformation allow sampling of the experimental binding mode. For this structure, we find a significant amount of binding through the minor groove in most sequences – but for *Random1* sequence, the most prevalent binding mode is through the major groove, as in the experimental structure. Thus, we conclude that the physical model used in MELD is sensitive to sequence-dependent properties. Despite this, in cases where the DNA undergoes significant conformational changes, where we expect *shape readout* to drive binding, the MELD restraints overcome sequence preferences. It is difficult to directly compare the effect of the restraints as independent MELD binding simulations are not comparable(77). We thus complement these results with relative binding affinity calculations (protocol 3).

### Relative binding affinities identify structure/sequence preferences during binding

Here two DNA duplexes and a protein are simulated together (see methods), promoting protein binding to either DNA duplex at low replicas while restricting interactions between the duplexes. During unbinding events, the protein is far from both DNA duplexes. We assess the higher affinity binders by counting how often the protein binds each duplex. In this approach, the binding mode is known; thus, MELD uses more constraining information to favor faster binding/unbinding events leading to higher statistical significance.

We first tested simulations in which we competed the consensus sequence against itself. The expected outcome is that the protein should bind equally to each sequence and is thus a test of the expected errors as well as possible systematic errors arising from the setup conditions. The observed ratio of 57/43 population was close to the expected 50/50 value (see Fig. 5b), giving us confidence that the setup conditions were not biased towards one of the two DNA structures. Competing the consensus against the other five sequences showed that the consensus sequence was preferred in all cases.

The advantage of simulation tools is that we can decouple sequence and shape readout by freezing the DNA structure in either its *apo* or *holo* conformation – similar to a rigid docking experiment, where the DNA can only fluctuate around the initial structure. We expected all relative free energies to increase when freezing DNA structures to the *holo* conformation as the conformational free energy change to the binding free energy was “prepaid.” Interestingly, while it increased in most cases (see Fig. 5d), it decreased in others, most significantly for the random sequence. We further analyzed this sequence by fixing either the consensus or random sequence to the *holo/apo* conformation. Surprisingly, the protein recognizes the random sequence preferentially when both structures are kept in their B-DNA form but the consensus sequence when they are in their holo conformation. Thus, simulations that include flexibility to deform will have more events progressing down the replica ladder when the protein binds the random sequence. Still, once they form the complex structure, the consensus remains bound for a longer time. The approach is not currently sensitive to large free energy differences – where only one structure might be bound at the lowest replica. The method can readily be made more quantitative by using information from all replicas with proper reweighting (e.g., using the multistate Bennett acceptance ratio(78)), however this was out of the scope of the current work (79).

**DISCUSSION**

The MELD approach is designed to help answer questions involving structural and energetic considerations, drawing from DNA's sequence dependence binding preferences. Choosing the origin of the information can help in either refining structures from these methods (e.g., making predictions from these methods and generating ambiguous/noisy data to guide MELD binding), answering questions about shape/sequence readout, or even relative binding affinities. The general pipeline is available on GitHub (github.com/PDNALab/MELD-DNA) to showcase the potential of this framework. Overall, the approach successfully predicts the structures of protein-DNA complexes. Understanding sequence/structure relationships and relative binding free energies is better indicated for systems where existing available information reduces the conformational search space. The trade-off between efficient sampling and the physics model to use currently limits MELD-DNA to implicit solvent, which is known to be less accurate than simulations in explicit solvent. Nonetheless, the method samples accurate atomic structures representative of the native state for most of the systems studied, even when starting from B-DNA conformations. The physics-based nature of the ensembles readily makes this a valuable approach to obtaining structures for more detailed studies (e.g., in explicit solvent), such as using different MELD clusters as seeds for adaptive sampling simulations combined with Markov State Models(80, 81). Thus, we believe the current framework can be a powerful tool to increase our understanding of nucleic acid complexes.

We have shown a successful strategy for identifying protein-DNA complexes sensitive to the DNA sequence, exploring multiple binding modes, and samples DNA deformation during binding. The MELD approach draws on the successes of the HADDOCK docking strategy of combining ambiguous and noisy data with the search engine. It goes beyond harmonic deformations by using a molecular dynamics engine sensitive to the DNA sequence and identifying successful structures based on a statistical mechanics treatment of the generated ensemble rather than on scoring functions. It is worth noting that we employed HADDOCK as regular users versed in structural biology, and expert users might improve the performance. Regardless, while the accuracy in HADDOCK structure predictions is lower than that of MELD, it also comes at a small fraction of the computational cost: MELD simulations require 30GPUs typically running for about two days for these systems, while HADDOCK calculations take a few minutes, on a single core. HADDOCK predictions can readily be incorporated as structural aseeds for each replica in MELD, as well as a source of ambiguous/noisy information. In this regime, MELD can be used to refine HADDOCK's models while simultaneously identifying the most likely model as the one most prevalent in the lowest temperature ensemble(82).

The choice of restraints, noise, and ambiguity in the dataset will depend on how much data is available to the user. Higher accuracy and amount of data lead to faster convergence but reduced exploration of the binding landscape. For most purposes of structure prediction, protocol 1 is the most transferable and generalizable. Furthermore, users might decide to change the restraints imposed on the protein and DNA. In our current approach, the DNA has more flexibility than the protein system (see Fig. 17) to account for DNA bending during binding. However, restraints between base pairs will keep structures from sampling conformations where one base is open (e.g., for binding methyltransferases). For such cases, a user would modify the restraints affecting the desired region of DNA. Similarly, the protein restraints affect the ability to sample open/closed states, which are needed for some systems where the binding site is not accessible in the closed state. Knowing the binding site, exceptions can be made on which regions of the protein to restrain.

We see three issues related to 1) accessibility of the binding site, 2) force field accuracy, and 3) efficiency of replica exchanges. Proteins 1BGB and 2B0D require an opening event before the DNA can access the binding site – however, the standard protocol used to predict binding (protocol 1) uses distance restraints that prevent the protein from unfolding – and in this case, from accessing the open conformation state. Furthermore, force fields typically have a bias towards compact structures, especially in implicit solvent, further favoring compact closed conformations that prevent DNA binding. Some authors have also suggested a possible DNA + protein force field imbalance resulting in stronger than expected Arginine and Lysine interactions with the phosphates(83–86). Such highly charged systems challenge the accuracy of the force field in implicit model. This issue is best seen in our set of complexes for the TATA box binding protein (1CDW). The bend in DNA structure induced by TATA-binding is further accentuated in the implicit solvent (see Fig. S21), resulting in overly strong electrostatic interactions (see SI). Because of compaction and these strong electrostatic interactions, replica exchanges that favor unbinding have a lower probability than similar approaches for protein-protein and protein-peptide binding (see SI and Figs. S22-23). Thus, despite the current success, future endeavors will aim to introduce explicit solvent into the methodology.

Overall, the MELD-DNA methodology we presented herein fills a gap in computational tools that predict protein-DNA binding. We have shown that the method is sensitive to sequence and structural preferences and is thus a promising new approach to studying this type of system. The MELD code is freely available through GitHub as a plugin to openmm(87). On a diverse set of protein-DNA systems involving 15 different complexes, the method successfully predicted 10 of them as high population clusters. We believe the physics-based insights MELD-DNA can provide will advance our understanding of protein-DNA interactions and our ability to simulate events related to supramolecular chemistry. Increasing our structural knowledge and sequence binding structural preferences combined with other factors that affect *in vivo* binding (e.g., chromatin state and accessibility) can bring new understanding to the molecular mechanisms that orchestrate gene regulation.

## DATA AVAILABILITY

The MELD software is distributed freely and available through the GitHub repository (github.com/maccallumlab/meld) – a permanent Zenodo link is accessible at (https://doi.org/10.5281/zenodo.7502226). Scripts and sample data used for this report are available at our group's GitHub (github.com/PDNALab/MELD-DNA), and a permanent copy has been deposited on Zenodo (https://doi.org/10.5281/zenodo.7501938).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

## FUNDING

This work was supported by the National Science Foundation, Grant No. CHE-2235785.

## ACKNOWLEDGEMENT

The authors acknowledge startup resources to use the hipergator supercomputer center at the University of Florida.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

1. Campolongo,M.J., Tan,S.J., Xu,J. and Luo,D. (2010) DNA nanomedicine: Engineering DNA as a polymer for therapeutic and diagnostic applications. *Adv Drug Deliver Rev*, **62**, 606–616.

2. Zhou,Z. and Dong,S. (2014) Protein–DNA interactions: a novel approach to improve the fluorescence stability of DNA/Ag nanoclusters. *Nanoscale*, **7**, 1296–1300.

3. Ma,X., Truong,P.L., Anh,N.H. and Sim,S.J. (2015) Single gold nanoplasmonic sensor for clinical cancer diagnosis based on specific interaction between nucleic acids and protein. *Biosens Bioelectron*, **67**, 59–65.

4. Meng,X.-Y., Zhang,H.-X., Mezei,M. and Cui,M. (2011) Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr Comput Aided-drug Des*, **7**, 146–157.

5. Weng,G., Gao,J., Wang,Z., Wang,E., Hu,X., Yao,X., Cao,D. and Hou,T. (2020) Comprehensive Evaluation of Fourteen Docking Programs on Protein-Peptide Complexes. *J Chem Theory Comput*, **16**, 3959–3969.

6. Moult,J., Fidelis,K., Kryshtafovych,A., Schwede,T. and Tramontano,A. (2016) Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins Struct Funct Bioinform*, **84**, 4–14.

7. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

8. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The Human Transcription Factors. *Cell*, **172**, 650–665.

9. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem*, **79**, 233–269.

10. Barissi,S., Sala,A., Wieczór,M., Battistini,F. and Orozco,M. (2022) DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors. *Nucleic Acids Res*, **50**, 9105–9114.

11. Ghoshdastidar,D. and Bansal,M. (2022) Flexibility of flanking DNA is a key determinant of transcription factor affinity for the core motif. *Biophys J*, **121**, 3987–4000.

12. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.

13. Rube,H.T., Rastogi,C., Kribelbauer,J.F. and Bussemaker,H.J. (2018) A unified approach for quantifying and interpreting DNA shape readout by transcription factors. *Mol Syst Biol*, **14**, e7902.

14. Yang,L., Orenstein,Y., Jolma,A., Yin,Y., Taipale,J., Shamir,R. and Rohs,R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol*, **13**, 910.

15. Schnepf,M., Reutern,M. von, Ludwig,C., Jung,C. and Gaul,U. (2020) Transcription Factor Binding Affinities and DNA Shape Readout. *Iscience*, **23**, 101694.

16. Dantas Machado,A.C., Cooper,B.H., Lei,X., Di Felice,R., Chen,L. and Rohs,R. (2020) Landscape of DNA binding signatures of myocyte enhancer factor-2B reveals a unique interplay of base and shape readout. *Nucleic Acids Res*, **48**, gkaa642-.

17. Luo,Y., Hitz,B.C., Gabdank,I., Hilton,J.A., Kagda,M.S., Lam,B., Myers,Z., Sud,P., Jou,J., Lin,K., *et al.* (2019) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res*, **48**, D882–D889.

18. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*, **489**, 57–74.

19. Auton,A., Abecasis,G.R., Altshuler,D.M., Durbin,R.M., Abecasis,G.R., Bentley,D.R., Chakravarti,A., Clark,A.G., Donnelly,P., Eichler,E.E., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

20. Etheve,L., Martin,J. and Lavery,R. (2017) Decomposing protein–DNA binding and recognition using simplified protein models. *Nucleic Acids Res*, **45**, 10270–10283.

21. Seeliger,D., Buelens,F.P., Goette,M., Groot,B.L. de and Grubmüller,H. (2011) Towards computional specificity screening of DNA-binding proteins. *Nucleic Acids Res*, **39**, 8281–8290.

22. Khabiri,M. and Freddolino,P.L. (2017) Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein–DNA Binding Free Energy Landscapes. *J Phys Chem B*, **121**, 5151–5161.

23. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235–242.

24. Narayanan,B.C., Westbrook,J., Ghosh,S., Petrov,A.I., Sweeney,B., Zirbel,C.L., Leontis,N.B. and Berman,H.M. (2014) The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res*, **42**, D114–D122.

25. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

26. Baek,M., McHugh,R., Anishchenko,I., Baker,D. and DiMaio,F. (2022) Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. *Biorxiv*, 10.1101/2022.09.09.507333.

27. Aderinwale,T., Christoffer,C.W., Sarkar,D., Alnabati,E. and Kihara,D. (2020) Computational structure modeling for diverse categories of macromolecular interactions. *Curr Opin Struc Biol*, **64**, 1–8.

28. Kitchen,D.B., Decornez,H., Furr,J.R. and Bajorath,J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*, **3**, 935–949.

29. Lensink,M.F., Brysbaert,G., Nadzirin,N., Velankar,S., Chaleil,R.A.G., Gerguri,T., Bates,P.A., Laine,E., Carbone,A., Grudinin,S., *et al.* (2019) Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins Struct Funct Bioinform*, **87**, 1200–1221.

30. Banitt,I. and Wolfson,H.J. (2011) ParaDock: a flexible non-specific DNA—rigid protein docking algorithm. *Nucleic Acids Res*, **39**, e135–e135.

31. Honorato,R.V., Roel-Touris,J. and Bonvin,A.M.J.J. (2019) MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing. *Frontiers Mol Biosci*, **6**, 102.

32. Dijk,M. van and Bonvin,A.M.J.J. (2010) Pushing the limits of what is achievable in protein-DNA docking: benchmarking HADDOCK's performance. *Nucleic Acids Res*, **38**, 5634–47.

33. Honorato,R.V., Roel-Touris,J. and Bonvin,A.M.J.J. (2019) MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing. *Frontiers Mol Biosci*, **6**, 102.

34. Dijk,M. van and Bonvin,A.M.J.J. (2008) A protein-DNA docking benchmark. *Nucleic Acids Res*, **36**, e88.

35. Wang,L., Wu,Y., Deng,Y., Kim,B., Pierce,L., Krilov,G., Lupyan,D., Robinson,S., Dahlgren,M.K., Greenwood,J., *et al.* (2015) 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*, **137**, 2695–2703.

36. Ruiter,A. de and Oostenbrink,C. (2020) Advances in the calculation of binding free energies. *Curr Opin Struc Biol*, **61**, 207–212.

37. Gapsys,V. and Groot,B.L. de (2017) Alchemical Free Energy Calculations for Nucleotide Mutations in Protein–DNA Complexes. *J Chem Theory Comput*, **13**, 6275–6289.

38. Seeliger,D., Buelens,F.P., Goette,M., Groot,B.L. de and Grubmüller,H. (2011) Towards computional specificity screening of DNA-binding proteins. *Nucleic Acids Res*, **39**, 8281–8290.

39. Khabiri,M. and Freddolino,P.L. (2017) Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein–DNA Binding Free Energy Landscapes. *J Phys Chem B*, **121**, 5151–5161.

40. Kappel,K., Jarmoskaite,I., Vaidyanathan,P.P., Greenleaf,W.J., Herschlag,D. and Das,R. (2019) Blind tests of RNA–protein binding affinity prediction. *Proc National Acad Sci*, **116**, 8336–8341.

41. Dai,H., Umarov,R., Kuwahara,H., Li,Y., Song,L. and Gao,X. (2017) Sequence2Vec: a novel embedding approach for modeling transcription factor binding affinity landscape. *Bioinformatics*, **33**, 3575–3583.

42. Yang,W. and Deng,L. (2019) PNAB: Prediction of protein-nucleic acid binding affinity using heterogeneous ensemble models. *2019 Ieee Int Conf Bioinform Biomed Bibm*, **00**, 58–63.

43. Yang,W. and Deng,L. (2020) PreDBA: A heterogeneous ensemble approach for predicting protein-DNA binding affinity. *Sci Rep-uk*, **10**, 1278.

44. Dias,R. and Kolazckowski,B. (2015) Different combinations of atomic interactions predict protein-small molecule and protein-DNA/RNA affinities with similar accuracy. *Proteins Struct Funct Bioinform*, **83**, 2100–2114.

45. MacCallum,J.L., Perez,A. and Dill,K.A. (2015) Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc National Acad Sci*, **112**, 6985–6990.

46. Perez,A., MacCallum,J.L. and Dill,K.A. (2015) Accelerating molecular simulations of proteins using Bayesian inference on weak information. *P Natl Acad Sci Usa*, **112**, 11846–51.

47. III,T.E.C., Cieplak,P. and Kollman,P.A. (1999) A Modified Version of the Cornell et al.Force Field with Improved Sugar Pucker Phases and Helical Repeat. *J Biomol Struct Dyn*, **16**, 845–862.

48. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat Methods*, **13**, 55–58.

49. Perez,A., Marchan,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J*, **92**, 3817–3829.

50. Maier,J.A., Martinez,C., Kasavajhala,K., Wickstrom,L., Hauser,K.E. and Simmerling,C. (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput*, **11**, 3696–3713.

51. Hornak,V., Abel,R., Okur,A., Strockbine,B., Roitberg,A. and Simmerling,C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct Funct Bioinform*, **65**, 712–725.

52. Nguyen,H., Roe,D.R. and Simmerling,C. (2013) Improved Generalized Born Solvent Model Parameters for Protein Simulations. *J Chem Theory Comput*, **9**, 2020–2034.

53. Nguyen,H., Perez,A., Bermeo,S. and Simmerling,C. (2015) Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins. *J Chem Theory Comput*, **11**, 3714–3728.

54. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem*, **25**, 1605–1612.

55. Flick,K.E., Jurica,M.S., Monnat,R.J. and Stoddard,B.L. (1998) DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. *Nature*, **394**, 96–101.

56. Robinson,H., Gao,Y.-G., McCrary,B.S., Edmondson,S.P., Shriver,J.W. and Wang,A.H.-J. (1998) The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature*, **392**, 202–205.

57. Zhao,Q., Chasse,S.A., Devarakonda,S., Sierk,M.L., Ahvazi,B. and Rastinejad,F. (2000) Structural basis of RXR-DNA interactions. *J Mol Biol*, **296**, 509–20.

58. Kim,S.-S., Tam,J.K., Wang,A.-F. and Hegde,R.S. (2000) The Structural Basis of DNA Target Discrimination by Papillomavirus E2 Proteins*. *J Biol Chem*, **275**, 31245–31254.

59. Mondragón,A. and Harrison,S.C. (1991) The phage 434 complex at 2.5 Å resolution. *J Mol Biol*, **219**, 321–334.

60. Swaminathan,K., Flynn,P., Reece,R.J. and Marmorstein,R. (1997) Crystal structure of a PUT3–DNA complex reveals a novel mechanism for DMA recognition by a protein containing a Zn2Cys6 binuclear cluster. *Nat Struct Biol*, **4**, 751–759.

61. Schumacher,M.A., Goodman,R.H. and Brennan,R.G. (2000) The Structure of a CREB bZIP·Somatostatin CRE Complex Reveals the Basis for Selective Dimerization and Divalent Cation-enhanced DNA Binding*. *J Biol Chem*, **275**, 35242–35247.

62. Watkins,D., Hsiao,C., Woods,K.K., Koudelka,G.B. and Williams,L.D. (2008) P22 c2 Repressor−Operator Complex: Mechanisms of Direct and Indirect Readout ‡. *Biochemistry-us*, **47**, 2325–2338.

63. Nikolov,D.B., Chen,H., Halay,E.D., Hoffman,A., Roeder,R.G. and Burley,S.K. (1996) Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc National Acad Sci*, **93**, 4862–4867.

64. Ellenberger,T.E., Brandl,C.J., Struhl,K. and Harrison,S.C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α Helices: Crystal structure of the protein-DNA complex. *Cell*, **71**, 1223–1237.

65. Keller,W., König,P. and Richmond,T.J. (1995) Crystal Structure of a bZIP/DNA Complex at 2.2 Å: Determinants of DNA Specific Recognition. *J Mol Biol*, **254**, 657–667.

66. Luisi,B.F., Xu,W.X., Otwinowski,Z., Freedman,L.P., Yamamoto,K.R. and Sigler,P.B. (1991) Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*, **352**, 497–505.

67. Horton,N.C. and Perona,J.J. (1998) Recognition of Flanking DNA Sequences by EcoRV Endonuclease Involves Alternative Patterns of Water-mediated Contacts*. *J Biol Chem*, **273**, 21721–21729.

68. Hiller,D.A., Rodriguez,A.M. and Perona,J.J. (2005) Non-cognate Enzyme–DNA Complex: Structural and Kinetic Analysis of EcoRV Endonuclease Bound to the EcoRI Recognition Site GAATTC. *J Mol Biol*, **354**, 121–136.

69. Morrone,J.A., Perez,A., Deng,Q., Ha,S.N., Holloway,M.K., Sawyer,T.K., Sherborne,B.S., Brown,F.K. and Dill,K.A. (2017) Molecular Simulations Identify Binding Poses and Approximate Affinities of Stapled α-Helical Peptides to MDM2 and MDMX. *J Chem Theory Comput*, **13**, 863–869.

70. Grant,B.J., Rodrigues,A.P.C., ElSawy,K.M., McCammon,J.A. and Caves,L.S.D. (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinform Oxf Engl*, **22**, 2695–6.

71. Grant,B.J., Skjærven,L. and Yao,X. (2021) The Bio3D packages for structural bioinformatics. *Protein Sci*, **30**, 20–30.

72. Zundert,G.C.P. van, Rodrigues,J.P.G.L.M., Trellet,M., Schmitz,C., Kastritis,P.L., Karaca,E., Melquiond,A.S.J., Dijk,M. van, Vries,S.J. de and Bonvin,A.M.J.J. (2016) The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol*, **428**, 720–725.

73. Honorato,R.V., Koukos,P.I., Jiménez-García,B., Tsaregorodtsev,A., Verlato,M., Giachetti,A., Rosato,A. and Bonvin,A.M.J.J. (2021) Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Frontiers Mol Biosci*, **8**, 729513.

74. Kurkcuoglu,Z. and Bonvin,A.M.J.J. (2020) Pre- and post-docking sampling of conformational changes using ClustENM and HADDOCK for protein-protein and protein-DNA systems. *Proteins Struct Funct Bioinform*, **88**, 292–306.

75. Méndez,R., Leplae,R., Maria,L.D. and Wodak,S.J. (2003) Assessment of blind predictions of protein–protein interactions: Current status of docking methods. *Proteins Struct Funct Bioinform*, **52**, 51–67.

76. Best,R.B., Hummer,G. and Eaton,W.A. (2013) Native contacts determine protein folding mechanisms in atomistic simulations. *Proc National Acad Sci*, **110**, 17874–17879.

77. Morrone,J.A., Perez,A., MacCallum,J. and Dill,K.A. (2017) Computed Binding of Peptides to Proteins with MELD-Accelerated Molecular Dynamics. *J Chem Theory Comput*, **13**, 870–876.

78. Shirts,M.R. and Chodera,J.D. (2008) Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys*, **129**, 124105.

79. Shirts,M.R. and Chodera,J.D. (2008) Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys*, **129**, 124105.

80. Perez,A., Sittel,F., Stock,G. and Dill,K. (2018) MELD-Path Efficiently Computes Conformational Transitions, Including Multiple and Diverse Paths. *J Chem Theory Comput*, **14**, 2109–2116.

81. Chang,L. and Perez,A. (2022) Deciphering the Folding Mechanism of Proteins G and L and Their Mutants. *J Am Chem Soc*, **144**, 14668–14677.

82. Liu,C., Brini,E., Perez,A. and Dill,K.A. (2020) Computing Ligands Bound to Proteins Using MELD-Accelerated MD. *J Chem Theory Comput*, **16**, 6377–6382.

83. Steinbrecher,T., Latzer,J. and Case,D.A. (2012) Revised AMBER Parameters for Bioorganic Phosphates. *J Chem Theory Comput*, **8**, 4405–4412.

84. You,S., Lee,H.-G., Kim,K. and Yoo,J. (2020) Improved Parameterization of Protein–DNA Interactions for Molecular Dynamics Simulations of PCNA Diffusion on DNA. *J Chem Theory Comput*, **16**, 4006–4013.

85. Bergonzo,C. and Cheatham,T.E. (2015) Improved Force Field Parameters Lead to a Better Description of RNA Structure. *J Chem Theory Comput*, **11**, 3969–3972.

86. Esadze,A., Chen,C., Zandarashvili,L., Roy,S., Pettitt,B.M. and Iwahara,J. (2016) Changes in conformational dynamics of basic side chains upon protein–DNA association. *Nucleic Acids Res*, **44**, 6961–6970.

87. Eastman,P., Friedrichs,M.S., Chodera,J.D., Radmer,R.J., Bruns,C.M., Ku,J.P., Beauchamp,K.A., Lane,T.J., Wang,L.-P., Shukla,D., *et al.* (2013) OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J Chem Theory Comput*, **9**, 461–469.

**Figure Captions**

**Figure 1. Schematic of the MELD approach. A)** Each system starts from an unbound state (minimum of 30Å distance between the protein (blue) and DNA(orange)). Information for the simulation is derived from possible interactions between the known interface Cα atoms (yellow) and dummy particles (green) projected on purine N1 atoms. Clustering yields the representative structure on the right. **B)** Representation of the data in a circular plot where each residue is a dot on the circle, with black lines representing the information used to direct binding. MELD simulations satisfy only a fraction of all the possibilities at any given time of the simulation. **C)** Schematic of the MELD Bayesian inference approach, adapted from ref (45).

**Figure 2.** Representation of ambiguity for bZIP binding. (a) Phosphate sites along the DNA sequence are combinatorically paired with $C_\beta$ atoms in the binding site of the bZIP protein (highlighted in orange). (b) The native interactions from the experimental structure are highlighted with a blue halo.

**Figure 3.** Superposition of best in top 5 clusters of MELD simulations against the experimental structure. We report the iRMSD of the top cluster, the best cluster amongst the top 5 clusters, and the best structure in the ensemble. A prediction is marked as a success if we can find a <5Å conformation in the top 5 clusters.

**Figure 4.** DNA sequence differences drive preferences in binding patterns. **A)** Bzip binding patterns seen in the top 4 clusters (c0-c3) binding to either the consensus or a random sequence, ordered by population (in %). The plot shows multiple binding modes, with the consensus having a high population (22.8%) of the experimental binding mode compared to 8.3% for a random sequence. **B)** Shape complementarity leads to similar binding modes across different sequences. Despite this, there are subtle differences along the number of base pairs between the two binding sites (denoted by the red dotted lines).

**Figure 5. P22-DNA competitive binding**. **A)** MELD simulations in which the protein is directed to bind to two different DNA sequences (blue and orange). Protein is shown as Cα dots superposing all frames binding to either sequence. The ratio of populations is related to the relative binding affinity. **B)** When the DNA is allowed to change its conformation during competitive binding freely, the consensus sequence has a marked preference for binding over the mutant (population in each mode in the bar, with grey depicting non-native binding modes. **C)** Restraining DNA conformations to their *holo* conformation removes the conformational free energy to adopt bound conformations. Despite this, the consensus sequence remains the most likely to bind. **D)** Comparison of random and consensus sequences allowing complete DNA flexibility or restraining either sequence to the *holo/apo* conformations.

## Table Captions

**Table 1.** Protein-DNA systems simulated in this study, along with their DNA sequence and PDB IDs.

**Table 2.** DNA sequences used for bZIP, TATA, and P22 complexes. DNA bases highlighted in orange represent the changes with respect to the consensus sequence (bold).

**Table 3.** Fraction of native contacts for MELD and HADDOCK (italics) from the top five clusters and the highest value sampled in the whole ensemble. Any two residues with at least one heavy atom pair within 5Å in the experimental structure were defined as a contact. We highlight in bold instances where 70% or more of the native contacts are satisfied.