JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

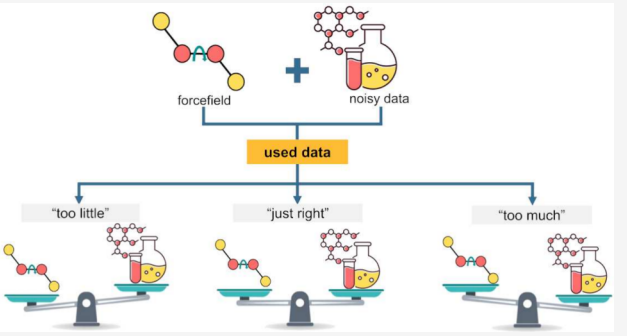
MELD-Adapt: On-the-Fly Belief Updating in Integrative Molecular Dynamics

Bhumika Singh, Arup Mondal, Kari Gaalswyk, Justin L. MacCallum,* and Alberto Perez*

Cite This: https://doi.org/10.1021/acs.jctc.4c00690 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Integrative structural biology synergizes experimental data with computational methods to elucidate the structures and interactions within biomolecules, a task that becomes critical in the absence of high-resolution structural data. A challenging step for integrating the data is knowing the expected accuracy or belief in the dataset. We previously showed that the Modeling Employing Limited Data (MELD) approach succeeds at predicting structures and finding the best interpretation of the data when the initial belief is equal to or slightly lower than the real value. However, the initial belief might be unknown to the user, as it depends on both the technique and the system of study. Here we introduce MELD-Adapt, designed to dynamically evaluate and infer the reliability of input data while at the same time finding the best interpretation of the data and the structures compatible with it. We demonstrate the utility of this method across different systems, particularly emphasizing its capability to correct initial assumptions and identify the correct fraction of data to produce reliable structural models. The approach is tested with two benchmark sets: the folding of 12 proteins with coarse physical insights and the binding of peptides with varying affinities to the extraterminal domain using chemical shift perturbation data. We find that subtle differences in data structure (e.g., locally clustered or globally distributed), starting belief, and force field preferences can have an impact on the predictions, limiting the possibility of a transferable protocol across all systems and data types. Nonetheless, we find a wide range of initial setup conditions that will lead to successful sampling and identification of native states, leading to a robust pipeline. Furthermore, disagreements about how much data is enforced and satisfied rapidly serve to identify incorrect setup conditions.



INTRODUCTION

Integrative structural biology (ISB)[1] aims to synergize experimental insights with computational methods grounded in physical or statistical principles. The goal is to unravel the intricate structures and interactions within biomolecules and their complexes. ISB is particularly invaluable when incomplete or limited structural data[2] is available that cannot, on its own, identify different metastable states of the system, or even the most relevant one. In this limited regime, the key challenges arise from the combination of sparse, ambiguous, and noisy[3] datasets and the challenge of interpreting experimental signals that are averaged over the multiple states present in the ensemble.[4−7] The goal of ISB models is to recover the different functional states that are in best agreement with such limited data. This is typically done by either reweighting ensembles produced by one technique (e.g., Molecular Dynamics) based on the data[8] or using the data itself simultaneously with a sampling strategy to generate conformational ensembles.[9,10] In this work, we focus on the latter.

To provide guiding power, the data is often transformed into some set of restraints (forward model) that the system has to satisfy, and which will incur some energy penalties when not satisfied.[6] Enforcing higher amounts of data as restraints restricts sampling to conformations compatible with the force field and these restraints, therefore focusing the search for states compatible with the data. The caveat is that noise and ambiguity prevent us from using all the data simultaneously. The correct interpretation of the data is self-consistent with a particular structure, whereas random subsets of data are inconsistent with a structural and physical/statistical model. Thus, how much data an ISB approach believes is critical in determining the structures.[9−14] When too few data points are trusted, it is easier to find states compatible with the data, but it significantly reduces guiding power (see Figure 1). On the other hand, when too much data is trusted, it increases guiding power, but it becomes incompatible with structural and physical/statistical models, producing incorrect predictions (see Figure 1). These issues become exacerbated by the

Received: May 28, 2024
Revised: September 12, 2024
Accepted: September 23, 2024

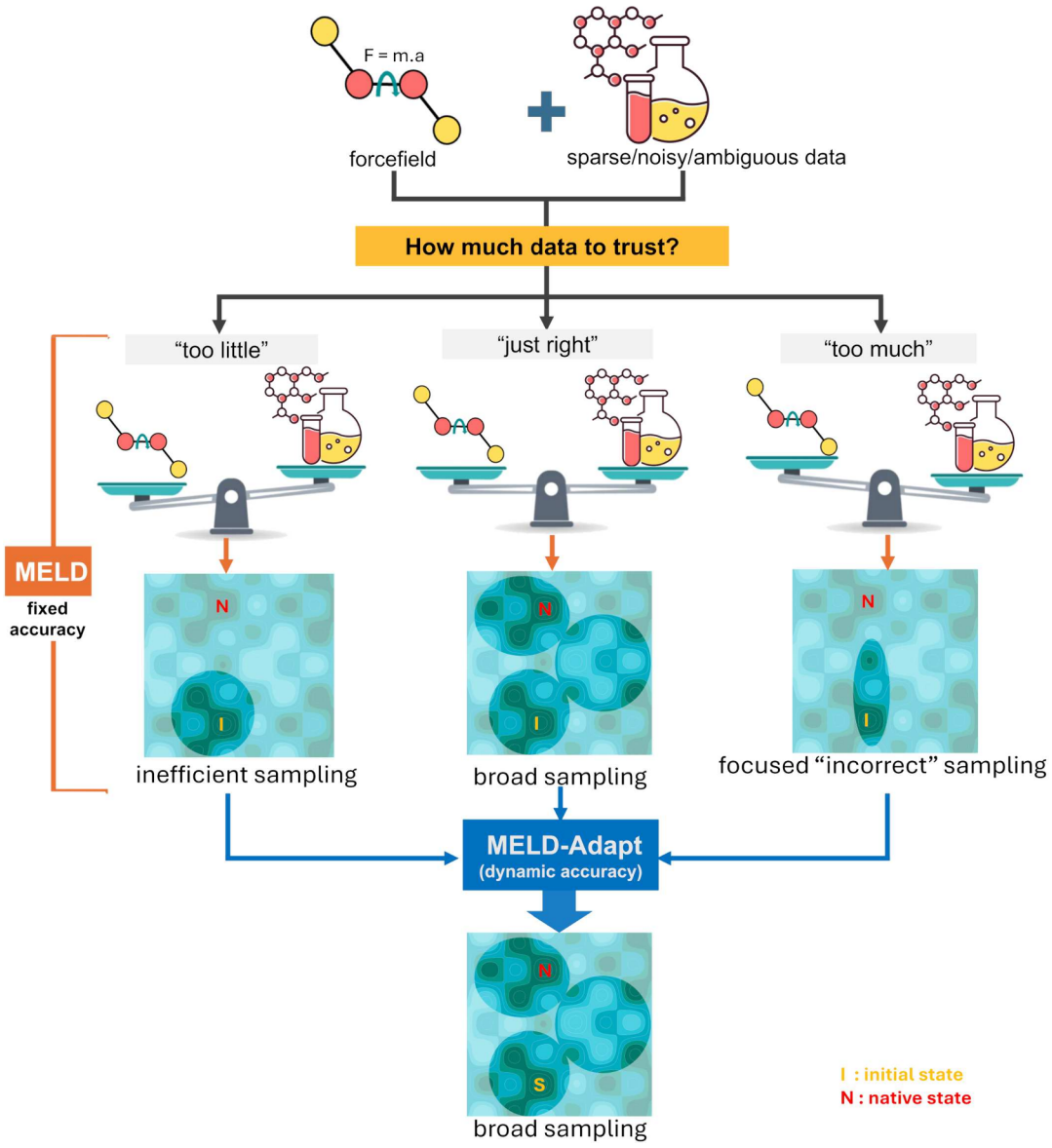


Figure 1. Balancing the quantity of semireliable data in guiding simulations. Insufficient data leads to inefficient sampling (left), while an excess of inaccurate data can result in highly focused yet incorrect predictions (right). Finding the optimal balance between data and physics-based models ensures broad and accurate sampling (middle). The letters I and N represent the initial and native states, respectively. MELD uses a fixed amount of data, which is user-defined to predict structure, potentially leading to any of the three scenarios depicted. MELD-Adapt dynamically optimizes the balance between data accuracy and force field preferences so independently of the starting trust it recovers optimal sampling.

system's internal dynamics and ensemble-averaged resolution of experimental data, leading to different possible interpretations of data, each compatible with different biologically relevant states.[1,7] Thus, setting the data belief becomes critical.

We previously introduced Modeling Employing Limited Data (MELD)[9,14] as an ISB tool based on molecular dynamics (MD), where by analyzing the ensembles using the principles of statistical mechanics[15] we can recover different metastable states compatible with data. By leveraging data, MELD explores the energy landscape more broadly than can be done with traditional MD.[16−18] The difficulty in MELD lies in knowing the amount of data to trust from the available pool that should be used.[19] It is especially important when combining sources of data that might mitigate individual limitations in each set.[20] It is possible that different substates of the system, represented by the data, satisfy a different percentage of the data. In the past, these led to running multiple simulations with different belief values, which rapidly became computationally demanding.

To address these challenges, we introduce a novel Bayesian inference approach combined with MELD, called MELD-Adapt, designed to dynamically learn and adjust the trustworthiness of data inputs, thereby enabling more accurate structure predictions. We demonstrate the effectiveness of this approach across different systems for protein and protein-peptide structure prediction, showcasing its ability to predict accurate structures even when initial assumptions are incorrect. First, we show that when the initial data accuracy is correct (e.g., has already been optimized), MELD-Adapt is able to predict accurate structures in agreement with MELD. Then we show that even when the initial assumptions are incorrect, MELD-Adapt is able to identify the correct fraction of data and the structures compatible with it, whereas traditional MELD makes incorrect predictions.

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

COMPUTATIONAL METHODS

The Modeling Employing Limited Data (MELD) Approach. The MELD methodology has been previously described[9,14] and thus only a quick summary is provided here. The MELD philosophy combines information that might be ambiguous and noisy with molecular simulations through Bayesian inference. Ambiguity refers to a source of data that might have multiple interpretations where only one is correct in the native structure of the biomolecules (e.g., atoms A and B are within 5 Å of each other or A and C are within 5 Å). By noisy data, we mean that for some data none of the possible interpretations are found in the native structure.

MELD does not use all data simultaneously to accommodate for noise and ambiguity. Rather, we provide an accuracy value for each dataset (a collection) that determines how many data points should be enforced, but not which data points. Which data is enforced dynamically changes throughout the simulation. A simulation can have multiple collections originating from different experiments or data sources. During the simulation, all of the data points in a collection are evaluated and ranked by the restraint penalty they introduce to the simulation. Then, the data points with the lowest energy are used until the next simulation step—driving the dynamics together with the force field. How many data points are used is fixed and determined by the accuracy value and number of data points in the collection. These choices provide a deterministic way for selecting which data points to enforce given a certain sampled structure.

MELD sampling at low temperatures, where restraints are strongly enforced, rarely changes which restraints are active, as this would require crossing over large energy barriers. To facilitate sampling of different subsets of data that guide to different regions of phase space, we use a Hamiltonian and Temperature Replica Exchange ladder.[21] At the highest replicas, temperature is high, and force constants for restraints vanish, allowing the efficient sampling of phase space and enforcing different subsets of data. As those structures are exchanged to lower replicas, the subset of restraints with the lowest restraint energy becomes active, guiding sampling toward regions of phase compatible with the active subset of data.

In the end, the posterior distribution agrees with both the force field and the best interpretation of the data. This can be framed as a Bayesian inference approach (eq 1):

$$p(x|D) = \frac{p(D|x)p(x)}{p(D)} \propto p(D|x)p(x) \tag{1}$$

where x represents a particular conformation at a time step given by an atomistic force field and D represents corresponding data. The prior ($p(x)$) originates from the Boltzmann distribution given a force field, and the data likelihood ($p(D|x)$) is given by the energy penalty of using some subset of data given the sampled structure. Finally, $p(x|D)$ is the posterior distribution from which we sample. It represents the probability of sampling a specific conformation given the fraction of the data enforced in the simulation.The critical step here is to determine the accuracy value for each collection. Users need to develop a deep understanding of the technique and system before selecting useful values for this—often having to compare agreements between different protocols using different accuracy values to ensure consistency and robustness of the approach. In the next section, we introduce a new implementation that allows the simulation to learn the accuracy value through the simulation using Bayesian inference.

MELD-Adapt: Inferring the Accuracy Value for a Collection. A fixed (static) accuracy value forces the decision upfront. If the user selects a value higher than the true accuracy of the data, the native basin will incur a nonzero energy restraint penalty. As a result, the ratio of Boltzmann weights between the native state and other basins will differ from that in the unbiased ensemble. Therefore, there is no guarantee that the highest population cluster in the biased ensemble will correspond to the one in the unbiased ensemble. On the other hand, choosing a number much lower than the dataset accuracy implies losing directive power in the simulations. Ideally, the accuracy can dynamically change during the simulation. This implies that throughout the simulation, in addition to selecting the most appropriate subset of data to use, there is also a process to ascertain the optimal amount of data to be utilized. To address this limitation, we have extended the Bayesian inference approach to consider the likelihood of enforcing the data (D) given a structure (x) and a number of active restraints (y). Thus, the posterior probability becomes

$$p(x, y|D) \propto p(x)p(y)p(D|x, y) \tag{2}$$

where $p(x)$ represents the prior probability over the structural variable x and is modeled using the Boltzmann distribution generated by the force field, $p(y)$ reflects the prior probability over the number of active restraints y, which crucially determines the number of active restraints associated with each collection during the simulation, and $p(D|x, y)$ quantifies the likelihood of observing a subset of the given data D under a specific structural conformation x and a particular set of active parameters y. Our primary focus for structural determination typically centers on the marginal distribution $p(x|D)$, which is obtained by integrating out the variable y to provide a more comprehensive understanding of the structural aspects of the system, all driven by the available data. Although x and y are dependent on each other ($p(D|x, y)$ is evaluated as a log-likelihood involving both variables), their priors ($p(x)$ and $p(y)$) are independent as one originates from the force field ($p(x)$) and the other from the forward-model used to implement the experimental data ($p(y)$).

Restraints are typically enforced as flat-bottom harmonic potentials, implying that their energy contribution is always greater or equal to zero. Thus, an on-the-fly increase in the number of restraints to satisfy would typically imply an increase in energy (unless a restraint is already satisfied, in which case it would not bring additional directive power). On the other hand, reducing the number of restraints will typically lead to a reduction in the overall restraint energy. Consequently, a naive approach would have a strong tendency toward satisfying zero restraints and thus lose any benefit MELD brings. We next consider what constitutes a good prior for the data ($p(y)$).

We employ a dynamic approach to update the active fraction of restraints throughout the MD simulation by utilizing Monte Carlo (MC) trials. We initialize the simulation with a prior for each collection based on the expected accuracy and a maximum and minimum range. We perform a Monte Carlo trial at given simulation intervals (e.g., every 100th MD step). During each trial, the number of active restraints can change by up to ±5. If, after the trial, the system has a number of

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

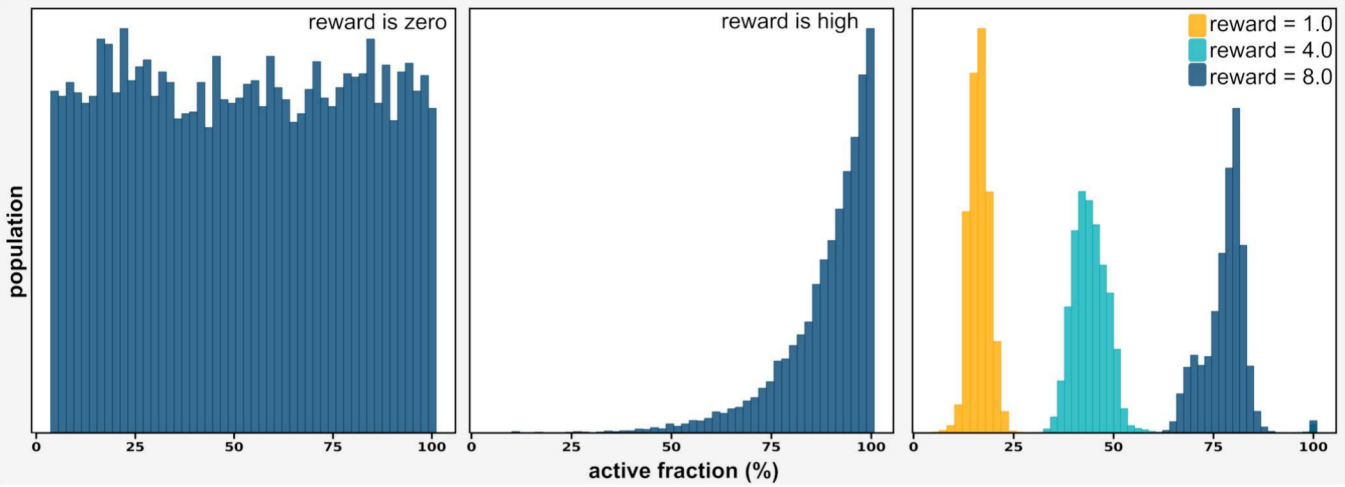


Figure 2. Effect of reward value on prior distribution. When the reward is zero, the prior is constant, and it uniformly samples all possible numbers of restraints in the absence of any forces (left). When the reward is nonzero, the prior grows exponentially, favoring a large number of active fractions in the absence of any forces (middle). In the presence of a force field, the active fractions find an upper and lower limit governed by the reward value and the force field (right).

restraints between the maximum and minimum values, the change in the number of restraints is accepted based on the Metropolis criterion. In the Metropolis criterion decreasing the energy will always result in acceptance of the change. Since the restraint energy for each restraint is always greater or equal to zero ($E_{rest} \geq 0$), reducing the number of restraints would always be favored. In essence, this corresponds to having a uniform prior that does not promote or discourage a specific number of restraints and just relies on the restraint energy. As the restraint energy with a uniform prior would lead to satisfying few restraints and therefore losing guiding power, it becomes important to establish a prior that promotes enforcing a higher number of restraints.

We decided to use an exponential prior to reward the addition of restraints over the initial value (see Figure 2). Effectively, this prior is introduced by adding a reward energy term for every restraint above an initial value, the $E_{\text{reward}} = -\lambda k_B T(N_i - N_{\text{prior}})$, where λ determines the magnitude of the reward per restraint, N_i is the number of current proposed restraints, and N_{prior} is the number of restraints we set as a prior for the simulation. Effectively, this E_{reward} is the log-prior and requires two hyperparameters: λ (reward magnitude) and N_{prior} (starting belief). When λ is set to zero, we return to the uniform distribution prior. When $\lambda > 0$, we have an exponential prior. When combining the prior with the restraint energy and force field energy, we obtain a distribution of sampled restraints (see Figure 2). As λ increases, the method would enforce more and more data−overcoming force field preferences. Thus, it is important to find a good reward magnitude that will balance between the best use of the data and force field. Below we detail the type of data used in this study and how we identified the value for λ.

Most importantly, the decision of how many restraints to satisfy is particular to each walker in the replica exchange system. As each walker samples different conditions along replicas (temperature and Hamiltonian), the distributions of how many restraints are satisfied will also change.

Benchmark Datasets. We use two benchmark sets for which we have previous experience using MELD with fixed accuracy values. The first benchmark is the folding of 12 proteins starting from sequence using coarse physical insights (CPIs) as external information. The second benchmark is the binding of a series of peptides that fold upon binding the extraterminal (ET) domain of bromo and extraterminal domain (BET) proteins using chemical shift perturbation (CSP) data for the ET domain measured in the presence/absence of the peptide.[22] Each type of data (CPIs or CSPs) has its own strengths and weaknesses.

CPIs originate from general protein principles[14] such as the presence of hydrophobic cores or pairing of beta strands based on secondary structure predictions using PSIPRED[23] protein structure prediction server.[24] This leads to many possible restraints across pairs of residues along the sequence. Since this data does not originate from an experiment, it is typically less self-consistent than that from experiments and provides less guiding power. The advantage is that this data is easy to derive starting from the protein sequence, but the amount of possible restraints and, therefore, the amount of noise increases with protein complexity (in terms of secondary structure) and size. Thus, the approach has only been successful for small proteins.

CSP datasets for peptide binding provide information about chemical shifts in the protein that change in the presence/absence of the peptide but provide no structural information about the peptide.[22] Typically, we process this data to identify residues in the protein with the largest perturbation and hypothesize that this change could be due to interactions with the peptide. However, we do not know which residues in the peptide interact, thus generating all possible interaction combinations between peptide residues and protein residues above the threshold, leading to a noisy dataset. Some of the selected protein residues might not even be involved in interactions with the peptide and change their chemical environment due to allosteric changes, further increasing the noise level in the dataset.

The major difference between CSP and CPIs is that CPIs are distributed through the 3D space of the protein, thus leading to many incompatible sets of restraints. On the other hand, CSP data tends to be localized near the binding site; thus, small conformational changes in the protein-peptide interaction easily give rise to satisfying a higher number of restraints (e.g., an extended conformation wrapping around the

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

active site) − challenging the biological relevance of the simulations.

Simulation Details. *Protein Folding Benchmark.* We conducted MELD-Adapt simulations on a dataset comprising 12 proteins (listed in Table S1). All simulations were performed using the OpenMM[25] suite of programs, starting from fully extended chains and a combination of the ff14SB[26] and ff99SB[27] atomic force fields for side chains and backbone, respectively. The simulations were executed within the GBNeck2 implicit solvent model,[28] using hydrogen mass repartitioning and a 3.5 fs time step.[29] To mitigate issues related to local energy minima, and to achieve efficient sampling, MELD uses Hamiltonian and Temperature Replica Exchange Molecular Dynamics (H, T-REMD).[21] We used 30 replicas such that each replica operates at varying temperatures within the range of 300 K to 550 K while maintaining a strong force constant (250 KJ·mol·nm^{-2}) for low-temperature replicas and gradually reducing it to zero for high-temperature replicas.

We performed simulations choosing different values for the two parameters that control how much data is enforced: the reward value and the prior belief. For the reward value, $E_{\text{reward}} = -\lambda k_B T\Delta$, choosing higher rewards promotes the activation of a higher number of restraints. We experimented with different values of λ, specifically 0.25, 0.5, 1.0, 2.0, 4.0, and 8.0. Consequently, six different simulations were conducted for each test system, each utilizing different reward values. For these systems, we chose our initial belief as we had done in our prior work,[14] which had already been optimized to account for most proteins (each hydrophobic residue is on average in contact with 2.4 other hydrophobic residues), and 45% of residues predicted to be part of β-strands will be involved in N−H···O hydrogen bonding to another strand residue. We then selected three systems (3GB1, T0769, and T0773), for which we repeated the calculation with an initial belief that was either lower (1.2 hydrophobic contacts and 25% of strand residues directed toward strand pairing) or higher than our previous work (4.8 hydrophobic contacts and 65% of strand residues directed toward strand pairings).

Protein−Peptide Binding Benchmark. We chose five different reward values to optimize: λ = [0, 0.25, 0.5, 0.75, 1.0]. Furthermore, for these systems, where intrinsically disordered peptides fold upon binding, the force field plays a crucial role. Thus, along with the reward value optimization, we tested there different force field and solvent model combinations: ff14SBside+gbNeck2,[26,28] ff14SBside+obc,[26,30] ff12SB-cMAP+obc.[30,31] For each of them we ran the traditional MELD approach with fixed number of data belief or the current MELD-Adapt protocol. Our previous experience with this systems used standard MELD with the ff14SBside +gbNeck2 combination.[22]

Simulations start from an unstructured peptide far from the protein receptor. We use the ET domain of BRD3 with peptides TP,[32] NSD3,[32] CHD4,[33] and BRG1[33] and the ET domain of BRD4 for the LANA[34] peptide based on the solved NMR structures deposited in the PDB (7JQ8, 7JYN, 6BGG, 6BGH, 2ND0).[32−34] We used available CSP data from TP and NSD3[32] and transferred the ET:TP data to the remaining systems, as previously done.[22] We set our initial trust of the CSP data at 4% for both MELD and MELD-Adapt protocols. Each simulation ran for 1 μs with 30 replicas. Temperature and restraint strength were scaled nonlinearly. Temperature ranged from 300 to 500 K, and force constant ranged from 350 kJ mol^{-1} nm^{-2} in the lower replicas to 0 kJ mol^{-1} nm^{-2} at the highest replica. More detail regarding setting up the H,T-REMD protocol can be found in our previous study.[22]

Analysis. We analyze the ensembles produced by different protocols by their ability to predict native-like confirmations (e.g., in the highest population cluster), provide faster and more robust convergence to the native state (e.g., by looking at the behavior across all replicas), and by finding the ideal fraction of the data to trust. This leads to the following three types of analysis.

Clustering. We used hierarchical agglomerative clustering to group structurally similar configurations and determine their respective populations using AMBER's *cpptraj* package.[35] The clustering process was applied to the second half of the trajectory from the five lowest temperature replicas, utilizing a cutoff distance of 2 Å and considering only Cα carbons of residues with predicted secondary structure. In cases where the secondary structure prediction covers less than 50% of the residues, the Cα of all residues were included in the clustering (listed in Table S2). For protein-peptide complexes, we used a cutoff of 1.5 Å for clustering, and unstructured peptide regions were removed from clustering. Our reported predictions include the centroid of the most populated cluster (top1) or the best centroid from the top 5 population clusters (top5) and their population and RMSD from native (interface RMSD or iRMSD[36] for protein-peptide complexes). Additionally, we assessed the fraction of frames with structures that conformed to the same number of restraints as native structures.

Convergence Checks. We conducted structure prediction convergence checks by analyzing the ensembles produced by the 30 replicas. First, we analyzed the RMSD distribution for the ensemble generated at each replica (the different Hamiltonian and Temperature). We anticipated observing RMSD distributions that would be wide at the highest replica and become narrower distributions (either native-like or misfolded) at lower replicas, in essence following a funnel-like behavior. We expect better protocols to yield a more funneled landscape toward native-like structures. Whereas bad protocols will converge quickly to non-native states.

Second, we checked convergence by comparing RMSD distributions against the native state for each individual walker as it performs a random walk in replica space. A well-converged simulation is characterized by consistent RMSD distributions among all walkers, with peaks occurring at the same values. Simulations that do not exchange properly (e.g., exchanging locally without achieving roundtrips in replica space) will present very different RMSD distributions. In a converged simulation with stable conformations, each walker should be able to visit native-like structures, with RMSD values below 4 Å. We quantified convergence using the Kullback−Leibler (KL) divergence using the Scikit-learn package.[37] A KL divergence value of 1.0 implies a significant disparity with the other individual distributions, while a value of 0.0 indicates an exact match between the observed and the other individual distributions.

Comparison of MELD-Predicted Accuracy with the Amount of Data Satisfied by the Experimental Structure. MELD-Adapt simulations sample a distribution of accuracies in a replica-dependent manner. We expect the analysis of the distribution of enforced accuracy at the lowest replica index to overlap significantly with the number of restraints satisfied by the native structure. Thus, we analyzed the number of hydrophobic and strand pairing restraints (for protein folding) and CSP-derived distance restraints (for peptide binding)

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

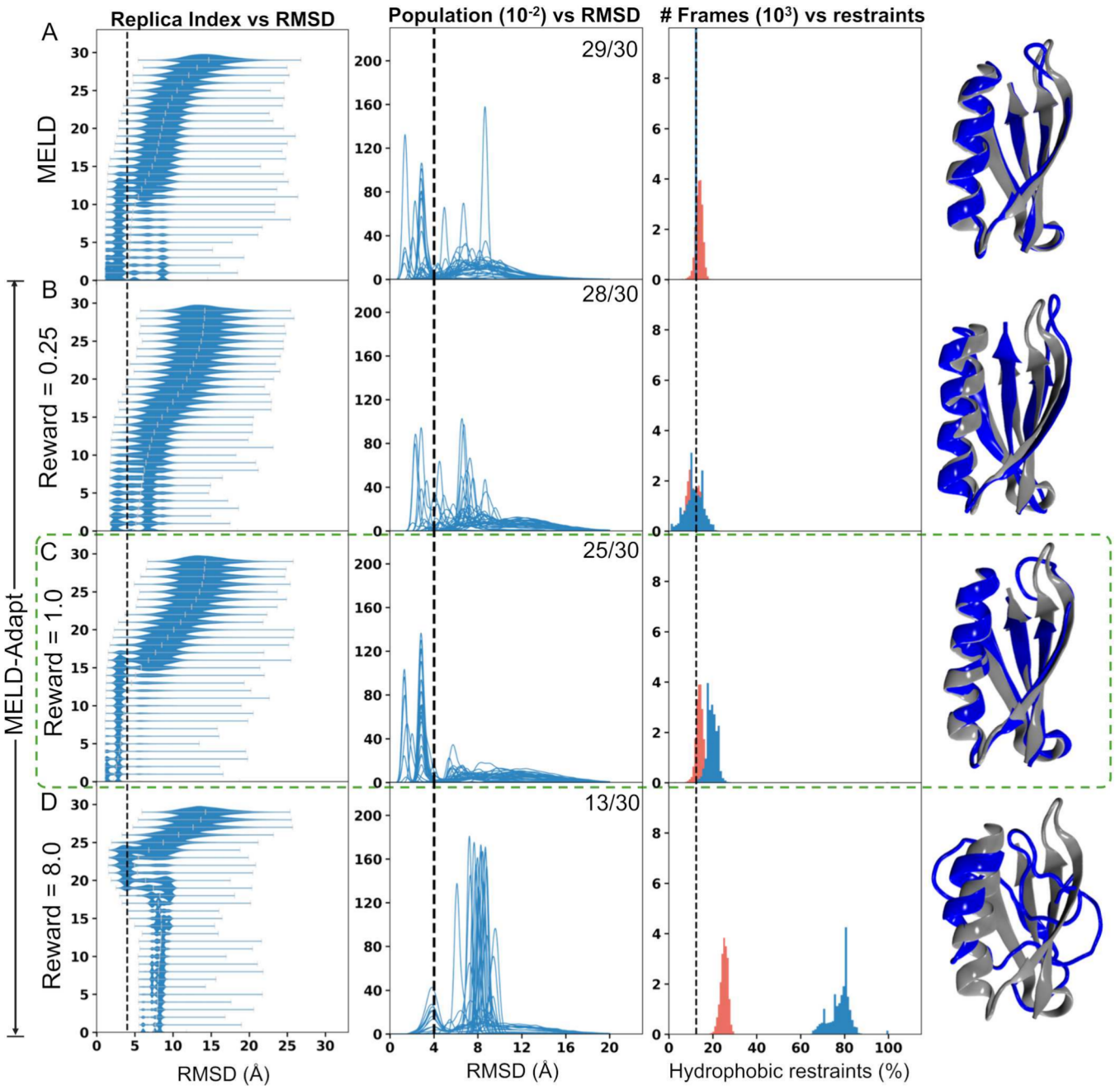


Figure 3. Comparative analysis of protocols with varying reward values for Protein G. Panel A uses a fixed belief with traditional MELD, while Panels B–D use MELD-Adapt with different λ values. The first column shows the funneling power of each protocol. The second column shows the overlap in RMSD distributions across 30 walkers; the values at the top right indicate how many of the 30 replicas have an average pairwise KL divergence lower than 0.5. The third column shows the percentage of active restraints during simulation (in blue) and the percentage of restraints satisfied (in red). The fourth column shows the centroid of the top population cluster (in blue) compared to the native structure (PDB ID 3GB1, in gray).

within the structure of each frame at the lowest temperature replica relative to the native structure using the MDTraj package.[38] The optimal reward value (λ) should yield a distribution that overlaps with the distribution obtained from satisfied restraints.

RESULTS AND DISCUSSION

MD-based approaches typically have a hard time disambiguating between sampling issues and force field issues. In this study, the addition of restraints can overpower force field preferences, and as we allow the number of restraints to change, so can the balance between force field and restraint energy contributions be skewed. This effect becomes especially pronounced in peptide–protein complexes, where the nature of the data and the marginal stability of the peptide, with the balance between intrinsically disordered proteins/peptides (IDPs) (unbound) and folded (bound), is more sensitive to changes in the number of enforced restraints. Thus, we will present and discuss the two benchmark sets separately and then provide a broader discussion.

MELD-Adapt Samples Folded States Across a Wide Range of Protocols. For each system and simulation

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

condition, we look at the ability of the approach to (1) sample the native state, (2) identify the native state as the highest population cluster, (3) identify the correct interpretation of the original data, and (4) agree between replica walkers (convergence). Figure 3 summarizes our results across different protocols for Protein G (PDB ID 3GB1). First, we look at the funneling power of different protocols, that is, the ability to narrow down sampling in low-temperature replicas to focus on the native state. Second, we look at the overlap in RMSD distributions along different replicas and calculate the KL divergence to assess convergence: if only one replica finds the native state for a long time the system is less converged than if all replicas sample the native state for a short amount of time. Third, we look at how much data (how many active restraints) the system is imposing during the simulation, and how many restraints are actually satisfied. If more restraints are enforced than satisfied, that means that we are adding restraint energy to our potential, which could alter force field preferences. On the other hand, if we are satisfying more restraints than actually enforced, this shows that the cooperative nature of restraints leads to satisfying more restrains than the ones being enforced (e.g., a contact between residues i and j trivially results in a contact between i and $j + 1$ even if there is no restraint guiding to it). Fourth, we look at the RMSD value between the centroid of the top population cluster and the native structure.

In general, when the data is not directive enough the funneling plots will exhibit broader distributions in the lowest temperature replica (see panels A and B in Figure 3). They will also rapidly lose native-like populations as we go to higher replicas. On the other hand, when enough data is enforced this leads to narrow RMSD histograms in the lowest temperature replica (see panel C in Figure 3). When a higher amount of data is enforced than the correct amount it can also lead to narrow distributions that are shifted with respect to the native state: the method guides to an incorrect region of phase space (see panel D in Figure 3).

Even with the fixed accuracy in MELD, we observe a distribution along the number of restraints that are satisfied in the native state centered around the enforced value (see panel A in Figure 3). This is due to the way we introduce restraints using flat-bottom harmonic restraints, where there is a quadratic term extending one Å in each direction that contributes k_BT at its maximum. Hence, thermal fluctuations can already account for small changes in satisfied restraints close to the fixed MELD value. Interestingly, with a reward value of one, we start seeing a separation between the number of restraints enforced (that is, contributing a restraint energy term to the potential, blue distribution) and those actually satisfied (red distribution, contributing an energy of 0 kcal/mol to the potential energy). Although protocols using a reward value between 0.25 and 1 kcal/mol per restraint activated above the initial belief systematically find the native state with high accuracy, the latter has a greater funneling potential and greater convergence of the results across replicas. As the difference between satisfied and enforced restraints becomes larger (λ = 4 and higher, see Figure S1), we observe convergence toward incorrect regions of phase space.

Of the ten protein systems we studied, the first nine produce excellent results with both MELD, and MELD-Adapt with rewards between 0.25 and 1. For these systems, the native state is always found as one of the top 5 population clusters, and for the 0.25 and 0.5 rewards systems also as the top population cluster (see Figures S1−S9 and Table S4). Ubiquitin, the tenth system, is the most complex protein in the dataset and a slow folder. Additionally, for this system, the accuracy of the secondary structure prediction was much lower than for the rest (see Table S3), resulting in low guiding power. When running simulations with our protocol MELD was not able to sample any native-like structures, so all clusters remain far from the native state (see Table S4). The MELD-Adapt protocol with a reward of 1 was the only to sample the native state (best RMSD in the ensemble of 3.36 Å), but not enough to get native-like clusters. It is unclear from our simulations if the failure with ubiquitin originates from a force field or sampling issue. Since MELD-Adapt was able to sample the native state, but not detect it as a high population state, it points to a possible force field deficiency.

We thus rerun ubiquitin by placing secondary structure more compatible with the native state (in accordance to the methods, this also changed which residues we clustered on; see Table S2). In these conditions, both MELD and a range of MELD-Adapt protocols were able to sample the native state and identify it through clustering (see Figure S10 and Table S4). Potentially, the secondary structure restraints are overcoming limitations in the force field that now allow us to sample the native structure in high populations. The presence of a small 3-10 helix flanked by loop regions with no secondary structure seems to be the most challenging part to get the overall topology of ubiquitin correct.

Figure S11 summarizes our results by showing improvements and failures of different reward values with respect to the original MELD approach in terms of RMSD of the top cluster and its population. In the plot, positive values for population increase and negative values for RMSD depict improvement of the MELD-Adapt protocol over MELD for a given reward value. The plot shows that reward values above λ = 2 have an overall negative effect, while lower values seem to perform similarly—a result of using good initial beliefs from our previous work.[14] This robustness across several λ values is a nice feature of the approach, since results are not overly dependent on the initial value chosen. For $\lambda > 2$, we are able to detect the protocol as a bad one (for example by looking at the disagreement between satisfied and enforced restraintsl Figures S1−S10).

As increasing λ values translate into enforcing a higher number of restraints, we thought this might affect the biased folding kinetics. Surprisingly, there are no clear trends as to the effect of different protocols on first passage time (the first time we sample the native state) (see Table S5). Indeed, Figures S1−S10 show that the distribution of satisfied restraints is very similar across protein systems for λ values between 0.25 and 1.0. However, when comparing these protocols to the original MELD results, the first passage time tends to increase.

This is expected when considering that the number of enforced restraints in MELD-Adapt is specific to each walker and recalculated at every time step. Examining the distribution of active restraints across replica conditions (varied by Temperature and Hamiltonian), we find that the belief level is replica-dependent (see Figure S12). At the highest replica index, all restraint force constants vanish and the temperature is high, so MELD-Adapt does not have preferences for how many restraints are satisfied (broad distribution of enforced restraints). At this replica, simulations sample unfolded states. As we move down the replica ladder, the force constants increase while the conformations are still largely unfolded,

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

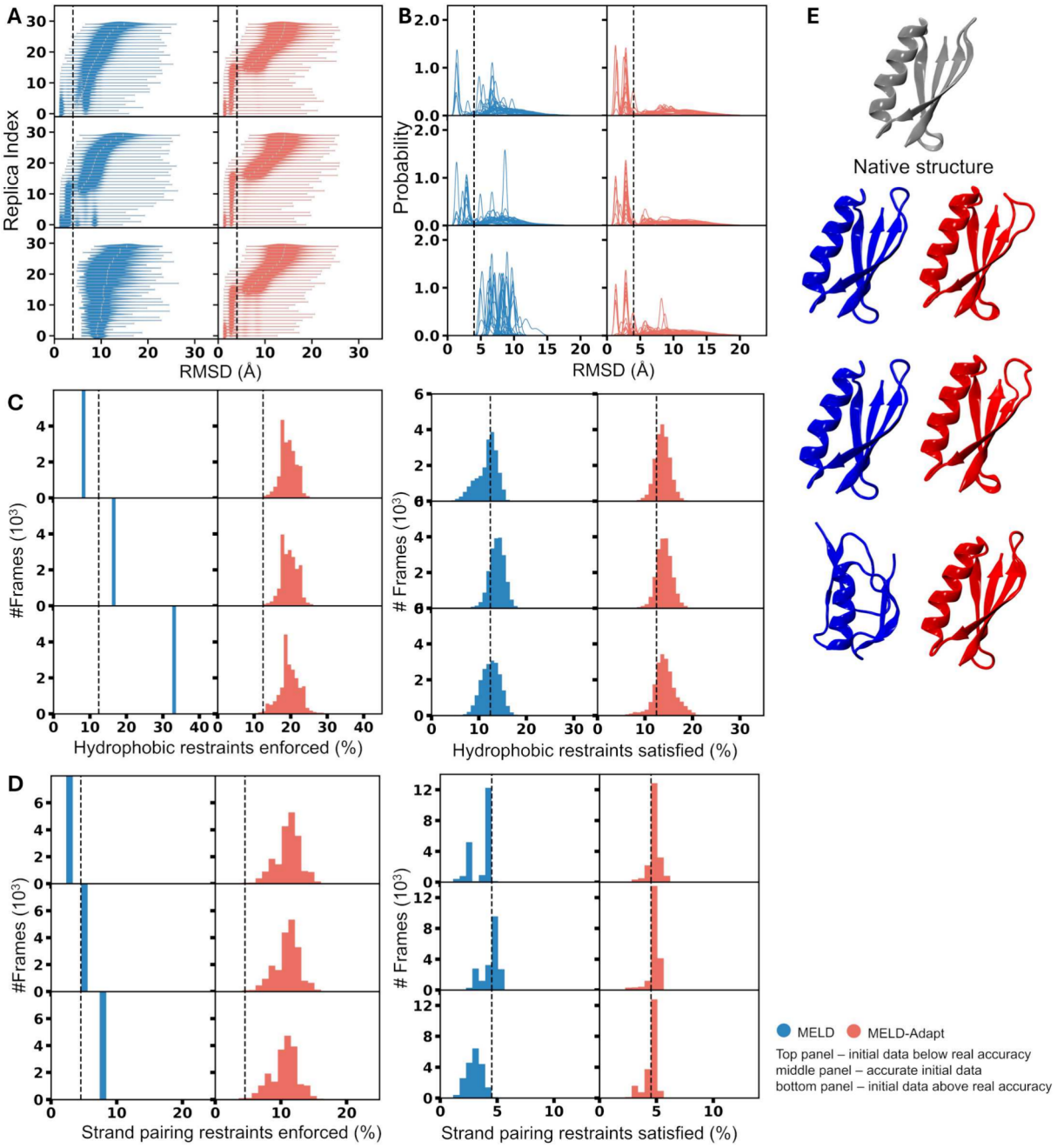


Figure 4. Folding belief analysis. Comparison of protein folding behavior between MELD (blue) and MELD-Adapt (red, with a reward value of 1.0 kcal/mol) under three scenarios: (i) initial data below real accuracy, (ii) accurate initial data, and (iii) initial data above real accuracy. The analysis includes (A) funneling protocol, (B) RMSD distribution, (C) percentage of enforced and satisfied hydrophobic restraints, and (D) percentage of enforced and satisfied strand pairing restraints, illustrating that MELD-Adapt dynamically adjusts the trust percentage and is independent from initial beliefs. (E) Best predicted structures in each protocol compared to the native structure (silver).

making it energetically costly to enforce many restraints. Consequently, MELD-Adapt typically satisfies a small fraction of restraints at these replicas, which help to initiate the folding of the protein. As the temperature continues to decrease and the restraint force constants increase further, the number of enforced restraints selected by MELD-Adapt progressively grows. At the lowest replica index, MELD-Adapt satisfies the optimal number of restraints (see Figure S12). In contrast, MELD enforces the same fraction of restraints at every step. Having more guiding information enforced at the higher replicas leads to shorter first passage times.

Another consequence of enforcing fewer restraints at higher replicas is the ability to sample more native-like folding routes. In MELD, the same amount of data is enforced across all replicas, hence the amount of data drives toward end states, where many contacts are formed. Some subsets of the data will lead to correct protein folds and some will lead to incorrect folds. However, the intermediates found are also driven by the same amount of restraints, so the MELD folding intermediates are not necessarily meaningful. In MELD-Adapt, fewer restraints are satisfied at higher replicas, which are compatible with earlier folded intermediates. Thus, MELD-Adapt allows for more physically meaningful intermediate states.

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

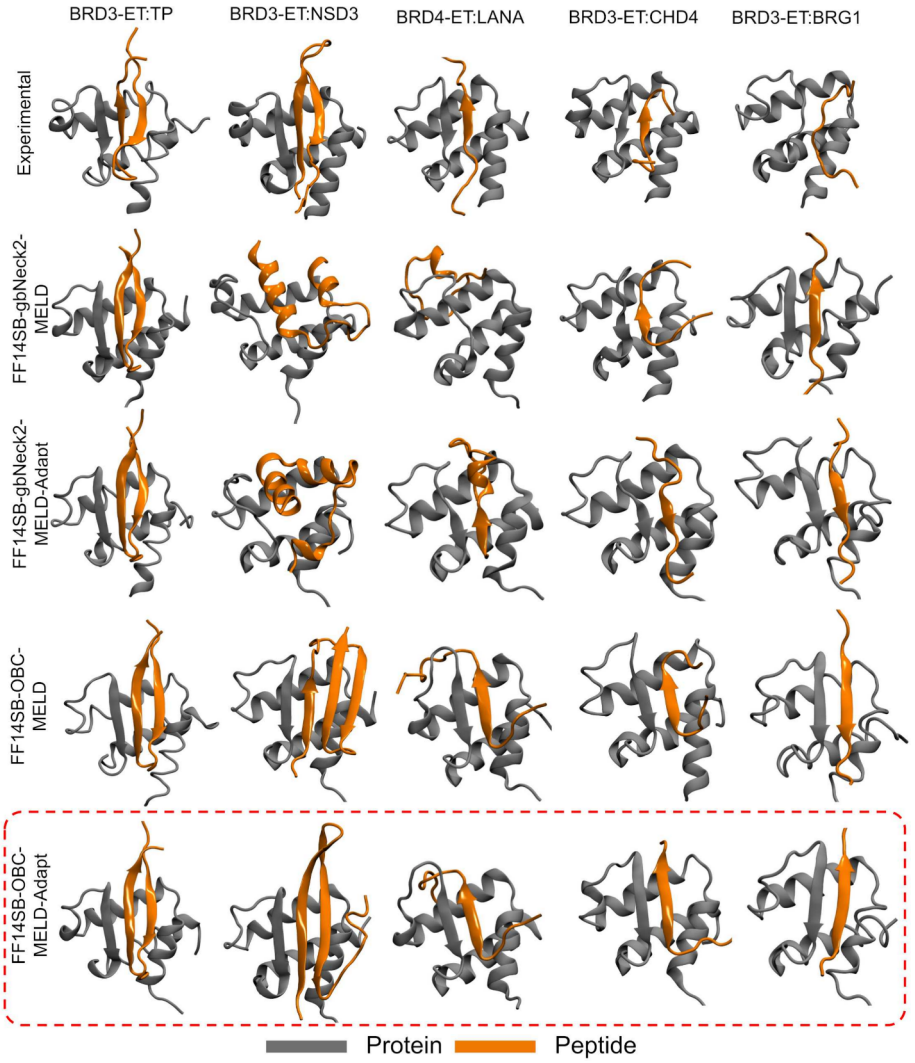


Figure 5. MELD-Adapt representative structures (centroid from the highest population cluster) for protein−peptide complexes simulated with different force fields. For each of the five simulated systems, the representative structure in each of the four simulation conditions are shown. The experimental structures are given as reference in the top row. The protein force field was fixed at the ff14SBside, with implicit solvent changing between OBC and GBneck2, and sampling using MELD or MELD-Adapt (λ = 0.25 kcal/mol). The red dashed box shows the most successful protocol.

For instance, consider protein G (3GB1), a small protein consisting of N-terminal and C-terminal hairpins that form a β-sheet through an interaction between strands 1 and 4. An α-helix then packs against the β-sheet to complete the protein's topology. The folding pathway involves first forming a non-native (registry-shifted) hairpin at the N-terminus, which stabilizes the C-termini hairpin. In this misfolded intermediate, the C-terminus is native-like, while the N-terminus is not. However, the pairing between of strands 1 (N-terminus) and 4 (C-terminus) is correct, allowing the N-terminus to partially unfold and refold into the native hairpin.

We previously showed that biasing MELD to the end state can recover the native state—but the folding pathways are incorrect.[39] Additionally, when MELD is biased to information compatible with the intermediate states, it can find the correct experimental folding pathway. However, without additional restraints to continue folding the protein, significantly longer simulation times are required to reach the fully folded state.

In MELD-Adapt simulations, the number of restraints varies across replicas, allowing compatibility with intermediate states at higher replicas and with the native state at the lowest replicas. As a result, MELD-Adapt can identify the correct folding pathway with significantly fewer resources. Indeed, Figure S13 shows the population of native-like, registry-shifted N-terminal and C-terminal hairpins present at different replica indexes for MELD and MELD-Adapt simulations. Traditional MELD simulations sample many compact states at high replica index and eventually predicts structures that are a mixture of the registry shifted and native-like N-terminal hairpin. MELD-Adapt, on the other hand, better separates the registry-shifted (populated at higher replica index) and native N-terminal states. Initially, it does not form hairpins at the highest replicas. As the registry shifted N-terminal hairpin is populated at intermediate replicas (orange), we start seeing population of the C-terminal hairpin (light blue), and as the replica index further decreases and the population of the C-terminal hairpin increases (dark blue and purple), so does the population of the correct N-terminal hairpin increase, while that of the registry-shifted hairpin decreases.

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Figure S14 further illustrates this point by looking at which data points among all possibilities are enforced for both strand pairing and hydrophobic restraints in protein G at the lowest replica. Most protocols (see Figure S14) capture the correct pairing between strands, but the first hairpin (pairing between strands 1 and 2) is off-registry for several protocols.[39] When considering these results together with hydrophobic pairings, this leads to further differences among protocols to accurately predict the native state—while the overall fold given the common core restraints is similar.

MELD-Adapt Recovers from Incorrect Beliefs. While the previous results show the importance of choosing the correct reward value, they show similar results between MELD and MELD-Adapt. The CPI initial belief was already optimized when developing the CPI protocols. Thus, MELD-Adapt converges on the MELD solution. To showcase the power of MELD-Adapt over MELD, we prepared a set of simulations using an incorrect initial belief. We chose three systems (protein G, and CASP11[40] targets T0769 and T0773),[41] where we set the initial belief to be substantially below or above the real accuracy of the data. For these three tests, we used our previously selected reward value of 1 kcal/mol. Figures 4, S15, and S16 and Tables S6−S8 summarize our results across these systems. The funneling plots show that, indeed, the MELD-Adapt method obtains similar results irrespective of the initial belief. Table S6 shows that only for two of the scenarios traditional MELD does not identify the native state in the top5 clusters, and Table S7 shows an increase in the population of the top1 cluster when using MELD-Adapt, indicating a more funneled behavior. This is a significant improvement over MELD, where enforcing too little data is not directive enough for protein G; and enforcing too much data directs protein G to an incorrect region of phase space. Surprisingly, for T0769 and T0773, we obtained good results even when satisfying incorrect data in fixed MELD. Both T0769 and T0773 were protein designs from the Baker lab,[42] which have been optimized to have very good funneling behavior and more stability than naturally occurring proteins. Thus, in this case, the balance between force field and restraint penalties favors the native state even when too much data is being enforced. In this scenario where different number of restraints are applied to the same system, it is easier to see the effect in the first passage time (see Table S8). Thus, in traditional MELD using an understimation of the data results in longer folding times. As more data is enforced, the first passage time is reduced. For the case of 3GB1 an overstimation of the data results in incorrect folding and thus cannot measure a mean first passage time. As discussed above, for the two designed proteins, adding more data just accelerates folding times. The situation is quite different when using MELD-Adapt, where simulations converge on the amount of data used. Despite this, for two of the systems (see Table S8), starting with a lower trust and increasing the number of restraints seems to be favored over starting with a high belief value.

Peptide Binding. We chose a series of five peptide systems that bind the ET domain with different binding affinities ranging from nanomolar (90 nM) to submillimolar (635 μM) (see Table S9).[22,32−34] We had previously shown that using CSP data alone we could predict the structure of the complex for the three strongest binders (TP, CHD4, and BRG1).[22] For NSD3, the structure of the complex was within the top 5 clusters, and for LANA, it was within the top 10 clusters. One of the caveats in mapping the experimental data from CSP is that the sensitivity of the data is dependent on the binding affinity; hence, we had to decide on a different threshold to optimize.[22] Furthermore, we used the CSP data from the TP and transferred it to use for BRG1, LANA, and CHD4. The threshold allows us to determine how many protein residues will likely be involved in the protein-peptide interface. Since there is no information for the peptide we have to consider that any residue in the peptide could be involved in binding—this leads to a combinatoric of possible contacts between protein and peptide. Since the peptides are of different lengths (Table S9), this also means that the number of total restraints is different for each system. In our previous work, we used a 4% accuracy after looking for self-consistency in predictions across simulations enforcing different belief values. Ideally, parameter sampling would allow us to recover the structure of the complex as the top prediction for all systems.

As in the case of folded proteins, we attempted MELD-Adapt with different reward values, first starting from the ff14SBside force field and GBneck2 implicit solvent we had used in previous work (see second row in Figure 5). However, in this case, reward values above 0.50 kcal/mol rapidly lead to incorrect predictions where the peptide stretched surrounding the active site (See Figure S17). Upon reflection, all restraining contacts in the dataset for these systems are clustered together (as we made all combinatorics from the CSP data), and the peptides are at the threshold of stability for folding upon binding. Hence, even a small reward for satisfying more restraints can easily overcome force field preferences, which are enabled by the close proximity of all candidate restraints. As the reward value increases, we observe a rapid shift in the distributions of enforced and satisfied restraints, with their overlap diminishing as λ increases (see Figure S18). For λ = 0.25 kcal/mol, predictions for each peptide sequence correctly locate the peptide in the binding site, but the peptide's internal conformation is often incorrectly predicted. The tail peptide (TP), the strongest binder in the set, is the only one correctly predicted. LANA, the weakest binder in the set, is predicted in the active site pairing as a β-strand, but the interacting residues in LANA are shifted along the sequence with respect to the experimental structure. Both CHD4 and BRG1 are predicted to bind as strands, but in a flipped conformation with respect to the experimental structure (see third row in Figure 5). Finally, NSD3 is predicted to bind adopting a helical conformation instead of the native hairpin conformation.

Acknowledging that this could be a force field deficiency, rather than an issue with the MELD-Adapt procedure, we ran simulations with different force fields (see methods) to better balance preferences between the helical and extended conformations.[31,43] Thus, we repeated our approach with the ff14SBside force field and OBC implicit solvent model. In this case, the protocol leads to accurate predictions up to a reward value of 0.50 (Figures S17 and S18). Therefore, we decided on a reward value of 0.25 for all peptide MELD-Adapt protocols. Analyzing all the simulations, we indeed observe, that top predictions using the traditional MELD approach with the OBC implicit solvent and ff14SBside with CMAP[31] correction significantly improved, capturing three of the five complexes as the top cluster (missing the orientation of BRG1 and the NSD3 peptide conformation, see the fourth row in Figure 5). The results improved with MELD-Adapt simulations, with all five predictions corresponding to the native structure of the complex. Figures S19 and S20 help rationalize the balance between the force field and parameter sampling. For instance,

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

all protocols and force fields were able to sample native-like structures of the complex (see RMSD values in Figure S20). They were just not preferred as high-population clusters in several cases. Figure S20 shows that the force field has a larger influence on the performance than whether MELD or MELD-Adapt was used. However, a head-to-head comparison between MELD and MELD-Adapt within each force field favors the later strategy. We further investigated the reason behind the protocol with ff14SBside force field and OBC solvent model being the best protocol. We analyzed all the generated restraints from the CSP data to see which and how many of them are satisfied in the top prediction from different protocols (Figure S21). We observe that protocol with the gbNeck2 solvent model shows a high number of true negative and false positive contacts irrespective of MELD or MELD-Adapt approach. The protocols with OBC satisfy most of the accurate contacts with very few true negative and false positive. MELD-Adapt shows a modest improvement with respect to MELD in this scenario (Figure S21). Overall pairs of systems, there was an insignificant improvement over the top five predictions and over the ensemble. However, the average MELD-Adapt improvement for the top cluster predictions was about 1−2 Å, showing that the approach is able to improve the ranking among force field selected top clusters, increasing its population. We repeated the simulations with an earlier force field version (ff12SB) with CMAP correction[31] which failed to reproduce most structures of the complex (even TP).

On-the-Fly Learning Dataset Accuracy Improves Modeling Predictions. Reducing human intervention and decisions in computational modeling leads to greater reproducibility. In the context of integrative approaches, the user has many choices to make, from the number of relevant states to how to model the data to how to deal with uncertainties in the data. The MELD approach starts with the idea that if we have a belief in the data that is equal to or lower than the real accuracy of the data, it can identify the best interpretation of the data as well as the structures compatible with the data. However, estimating this value is not trivial, and will change according to experiments. The ability to sample native-like structures is still conditioned by challenges like backtracking,[44] force field accuracy, and sampling efficiency—and it is hard to disambiguate the different contributions.

Here, we have shown that on-the-fly identification of data accuracy can lead to more efficient simulations, increase sampling efficiency when our initial belief is too low, and correct phase space exploration when our initial belief drives to incorrect structures. However, the method is sensitive to the internal structure of the data (local or distributed) and the balance with the force field. Thus, high reward values can lead to too much data being satisfied and overcoming force field preferences. We did not find a protocol where fixing the reward value can be uniformly transferable across datasets, but lower reward values (0.25 kcal/mol) seemed to work best overall in both cases—with low guiding power for the protein folding dataset.

We have focused our efforts on two challenging problems for MD simulations: binding and folding. Both fall under the broader category of structure prediction, where AI methods like AlphaFold (AF) have shown great success. We do not expect our method to outperform AF in cases where AF excels. However, MD-based integrative methods offer distinct advantages, such as compatibility with known experimental data and a physics model, transferability to different systems where available force fields exist, the ability to handle ensembles with relative importance determined by populations. Additionally, MD can identify intermediate and metastable states, as well as the relationships between them.

MELD-Adapt allows users to adjust the balance between how much the data is trusted versus how much the force field is relied upon. This is a decision left to the user. We recommend examining the restraint energies: if these energies are too high at the lowest replica, the force field has less influence in identifying states through clustering. On the other hand, if the restraint penalties are relatively low, they can help compensate for any shortcomings in the force field, providing additional guidance to the simulations.

CONCLUSION

This study presents an enhanced MELD approach that incorporates on-the-fly learning to dynamically calibrate the use of experimental data, improving the accuracy of biomolecular structure predictions. Our results underline the importance of carefully balancing force field influences and data restraints to avoid convergence to incorrect regions of phase space. While the method shows sensitivity to the internal structure of the data and the balance with the force field, we have demonstrated that lower reward values tend to offer a safer compromise between data guidance and depending more on force field accuracy. The ability of the MELD-Adapt technique to recover from incorrect initial beliefs showcases its potential as a powerful tool in integrative structural biology, especially when addressing the inherent uncertainties present in experimental datasets. Ultimately, our approach reduces human intervention, increasing the reproducibility of computational modeling and providing a robust framework for predicting the structure of proteins and their complexes with peptides.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00690.

Tables S1−S9 and Figures S1−S21 (PDF)

AUTHOR INFORMATION

Corresponding Authors

Justin L. MacCallum − *Department of Chemistry, University of Calgary, Calgary, Alberta T2N 1N4, Canada;* orcid.org/0000-0001-7917-7068; Email: justin.maccallum@ucalgary.ca

Alberto Perez − *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611-7011, United States;* orcid.org/0000-0002-5054-5338; Email: perez@chem.ufl.edu

Authors

Bhumika Singh − *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611-7011, United States*

Arup Mondal − *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611-7011, United States;* orcid.org/0000-0002-8970-3380

Kari Gaalswyk − *Department of Chemistry, University of Calgary, Calgary, Alberta T2N 1N4, Canada*

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jctc.4c00690

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.P thanks NSF CAREER CHE-2235785 for funding.

■ REFERENCES

(1) Schneidman-Duhovny, D.; Pellarin, R.; Sali, A. Uncertainty in integrative structural modeling. *Curr. Opin. Struct. Biol.* **2014**, *28*, 96−104.
(2) Ward, A. B.; Sali, A.; Wilson, I. A. Integrative structural biology. *Science* **2013**, *339*, 913−915.
(3) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106−116.
(4) Orioli, S.; Larsen, A. H.; Bottaro, S.; Lindorff-Larsen, K. How to learn from inconsistencies: Integrating molecular simulations with experimental data. *Prog. Mol. Biol. Transl. Sci.* **2020**, *170*, 123−176.
(5) Bonomi, M.; Vendruscolo, M. Determination of protein structural ensembles using cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **2019**, *56*, 37−45.
(6) Gaalswyk, K.; Muniyat, M. I.; MacCallum, J. L. The emerging role of physical modeling in the future of structure determination. *Curr. Opin. Struct. Biol.* **2018**, *49*, 145−153.
(7) Mondal, A.; Lenz, S.; MacCallum, J. L.; Perez, A. Hybrid computational methods combining experimental information with molecular dynamics. *Curr. Opin. Struct. Biol.* **2023**, *81*, 102609.
(8) Bottaro, S.; Bengtsen, T.; Lindorff-Larsen, K. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. *Methods Mol. Biol.* **2020**, *2112*, 219−240.
(9) MacCallum, J. L.; Perez, A.; Dill, K. A. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 6985−6990.
(10) Kuenze, G.; Bonneau, R.; Leman, J. K.; Meiler, J. Integrative protein modeling in RosettaNMR from sparse paramagnetic restraints. *Structure* **2019**, *27*, 1721−1734.
(11) Dominguez, C.; Boelens, R.; Bonvin, A. M. HADDOCK: a protein- protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **2003**, *125*, 1731−1737.
(12) De Vries, S. J.; Van Dijk, M.; Bonvin, A. M. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **2010**, *5*, 883−897.
(13) DiMaio, F.; Song, Y.; Li, X.; Brunner, M. J.; Xu, C.; Conticello, V.; Egelman, E.; Marlovits, T. C.; Cheng, Y.; Baker, D. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* **2015**, *12*, 361−365.
(14) Perez, A.; MacCallum, J. L.; Dill, K. A. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 11846−11851.
(15) Hollingsworth, S. A.; Dror, R. O. Molecular dynamics simulation for all. *Neuron* **2018**, *99*, 1129−1143.
(16) Schlick, T.; Portillo-Ledesma, S.; Myers, C. G.; Beljak, L.; Chen, J.; Dakhel, S.; Darling, D.; Ghosh, S.; Hall, J.; Jan, M.; et al. Biomolecular modeling and simulation: a prospering multidisciplinary field. *Annu. Rev. Biophys.* **2021**, *50*, 267−301.
(17) Shaw, D. E.; Adams, P. J.; Azaria, A.; Bank, J. A.; Batson, B.; Bell, A.; Bergdorf, M.; Bhatt, J.; Butts, J. A.; Correia, T. Anton 3: twenty microseconds of molecular dynamics simulation before lunch. In *SC '21: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*; ACM, 2021; pp 1−11..
(18) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; et al. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **2008**, *51*, 91−97.
(19) Mondal, A.; Perez, A. Simultaneous assignment and structure determination of proteins from sparsely labeled NMR datasets. *Front. Mol. Biosci.* **2021**, *8*, 774394.
(20) Gauto, D. F.; Estrozi, L. F.; Schwieters, C. D.; Effantin, G.; Macek, P.; Sounier, R.; Sivertsen, A. C.; Schmidt, E.; Kerfah, R.; Mas, G.; et al. Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. *Nat. Commun.* **2019**, *10*, 2697.
(21) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(22) Mondal, A.; Swapna, G.; Lopez, M. M.; Klang, L.; Hao, J.; Ma, L.; Roth, M. J.; Montelione, G. T.; Perez, A. Structure Determination of Challenging Protein-Peptide Complexes Combining NMR Chemical Shift Data and Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2023**, *63*, 2058−2072.
(23) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.
(24) McGuffin, L. J.; Bryson, K.; Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **2000**, *16*, 404−405.
(25) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.
(26) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(27) 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., Bioinf.* **2006**, *65*, 712−725.
(28) Nguyen, H.; Roe, D. R.; Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **2013**, *9*, 2020−2034.
(29) Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.* **2015**, *11*, 1864−1874.
(30) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 383−394.
(31) Perez, A.; MacCallum, J. L.; Brini, E.; Simmerling, C.; Dill, K. Grid-based backbone correction to the ff12SB protein force field for implicit-solvent simulations. *J. Chem. Theory Comput.* **2015**, *11*, 4770−4779.
(32) Aiyer, S.; Swapna, G. V.; Ma, L.-C.; Liu, G.; Hao, J.; Chalmers, G.; Jacobs, B. C.; Montelione, G. T.; Roth, M. J. A common binding motif in the ET domain of BRD3 forms polymorphic structural interfaces with host and viral proteins. *Structure* **2021**, *29*, 886−898.
(33) Wai, D. C.; Szyszka, T. N.; Campbell, A. E.; Kwong, C.; Wilkinson-White, L. E.; Silva, A. P.; Low, J. K.; Kwan, A. H.; Gamsjaeger, R.; Chalmers, J. D.; Patrick, W. M.; Lu, B.; Vakoc, C. R.; Blobel, G. A.; Mackay, J. P. The BRD3 ET domain recognizes a short peptide motif through a mechanism that is conserved across chromatin remodelers and transcriptional regulators. *J. Biol. Chem.* **2018**, *293*, 7160−7175.
(34) Zhang, Q.; Zeng, L.; Shen, C.; Ju, Y.; Konuma, T.; Zhao, C.; Vakoc, C.; Zhou, M.-M. Structural Mechanism of Transcriptional Regulator NSD3 Recognition by the ET Domain of BRD4. *Structure* **2016**, *24*, 1201−1208.
(35) Roe, D. R.; Cheatham, T. E., III PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **2013**, *9*, 3084−3095.

(36) Armougom, F.; Moretti, S.; Keduas, V.; Notredame, C. The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics* **2006**, *22*, e35−e39.
(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(38) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.
(39) Chang, L.; Perez, A. Deciphering the Folding Mechanism of Proteins G and L and Their Mutants. *J. Am. Chem. Soc.* **2022**, *144*, 14668−14677.
(40) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Struct., Funct., Bioinf.* **2016**, *84*, 4−14.
(41) Perez, A.; Morrone, J. A.; Brini, E.; MacCallum, J. L.; Dill, K. A. Blind protein structure prediction using accelerated free-energy simulations. *Sci. Adv.* **2016**, *2*, e1601274.
(42) Lin, Y.-R.; Koga, N.; Tatsumi-Koga, R.; Liu, G.; Clouser, A. F.; Montelione, G. T.; Baker, D. Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E5478−E5485.
(43) Shell, M. S.; Ritterson, R.; Dill, K. A. A Test on Peptide Stability of AMBER Force Fields with Implicit Solvation. *J. Phys. Chem. B* **2008**, *112*, 6878−6886.
(44) Capraro, D. T.; Roy, M.; Onuchic, J. N.; Jennings, P. A. Backtracking on the folding landscape of the beta-trefoil protein interleukin-1beta? *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 14844−14848.

https://doi.org/10.1021/acs.jctc.4c00690
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX