JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

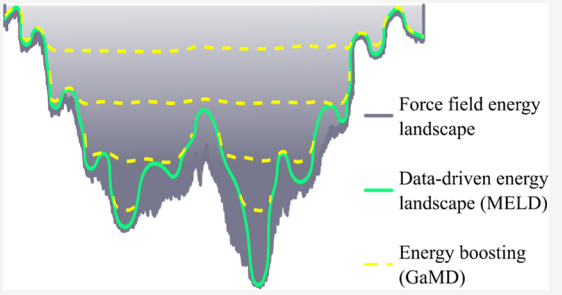
When MELD Meets GaMD: Accelerating Biomolecular Landscape Exploration

Marcelo Caparotta and Alberto Perez*

Cite This: https://doi.org/10.1021/acs.jctc.3c01019 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We introduce Gaussian accelerated MELD (GaMELD) as a new method for exploring the energy landscape of biomolecules. GaMELD combines the strengths of Gaussian accelerated molecular dynamics (GaMD) and modeling employing limited data (MELD) to navigate complex energy landscapes. MELD uses replica-exchange molecular simulations to integrate limited and uncertain data into simulations via Bayesian inference. MELD has been successfully applied to problems of structure prediction like protein folding and complex structure prediction. However, the computational cost for MELD simulations has limited its broader applicability. The synergy of GaMD and MELD surmounts this limitation efficiently sampling the energy landscape at a lower computational cost (reducing the computational cost by a factor of 2 to six). Effectively, GaMD is used to shift energy distributions along replicas to increase the overlap in energy distributions across replicas, facilitating a random walk in replica space. We tested GaMELD on a benchmark set of 12 small proteins that have been previously studied through MELD and conventional MD. GaMELD consistently achieves accurate predictions with fewer replicas. By increasing the efficacy of replica exchange, GaMELD effectively accelerates convergence in the conformational space, enabling improved sampling across a diverse set of systems.



1. INTRODUCTION

Generalized ensemble methods are an attractive enhanced sampling strategy as they forego predetermining any collective structural variable to guide the system.[1,2] However, methods such as Replica Exchange typically require many copies of the system in order to efficiently sample a range of conditions (temperature or Hamiltonian) that enhance sampling.[3] The number of replicas is determined by considering how to expand the range of conditions in such a way that the energy distributions have some overlap that will allow efficient exchange and ultimately a random walk among replica conditions.[4] REMD-based approaches have often been used for problems involving large conformational changes such as protein folding.[5] However, the folded-to-unfolded transition further limits efficient exchanges. Here we propose a methodology to improve energy distribution overlap along neighboring replicas resulting in a lower number of required replicas. We showcase the performance of the method by predicting the structures of 12 small proteins starting from their sequence.

Protein folding is the process by which a protein molecule assumes its functional three-dimensional structure from its linear amino acid sequence. Beyond its role for potentially determining structures that cannot be determined experimentally,[6,7] the approach can provide insights into other relevant states accessible to the system such as misfolded states associated with pathologies (e.g., Alzheimer's disease).[8,9] For years, the protein folding problem has been of interest for developing enhanced sampling methods.[10,11] Among its advantages is a clear way to assess success (e.g., identifying the folded state) and a generalizable strategy for applying the methodology to multiple protein sequences.

In recent years, we have shown that the MELD (modeling employing limited data)[12,13] approach is efficient at integrating experimental knowledge or general knowledge with simulations for predicting the structures of proteins[14] and their complexes.[15,16] Some of the caveats of the approach are the large computational expense due to the number of replicas needed to expand the desired simulation conditions (typically 300 to 450 K) and how information is enforced in the Hamiltonian.

Independently, GaMD (Gaussian accelerated molecular dynamics) is another enhanced sampling technique that places a bias on the energy to promote exploration of the energy surface.[17] The approach has been successful in a range of biological problems and has the advantage that is also independent of structural variables, which simplifies the transferability of the approach across molecular systems.[18,19]

Received: September 15, 2023
Revised: October 22, 2023
Accepted: November 15, 2023

By using GaMD, researchers can overcome energy barriers and explore complex and heterogeneous systems, including protein folding and ligand binding.[20] Whereas MELD typically starts from unfolded states and predicts the native state as the highest population state, GaMD often starts from the native state to bring insights into kinetics and thermodynamics of binding or folding.

In this work, we combine the virtues of GaMD with those of MELD to modulate the overlap between energies across replicas. We show this GaMELD strategy requires fewer replicas than MELD to predict the folded state of proteins while maintaining a high level of accuracy. We focus on the folding of 12 proteins with sequence lengths ranging from 28 to 92 amino acids and compare the performance against known experimental structures. We also provide benchmark results comparing GaMELD against either MELD or GaREMD (GaMD + REMD), showing that the synergy of both methods increases the accuracy and reduces the computational cost.

2. METHODS

2.1. GaMELD. We have implemented the GaMELD methodology into OpenMM,[21] by incorporating GaMD-OpenMM[22] into MELD[13] 0.6.1. Although both methods are implemented in the OpenMM simulation engine and have a Python front-end, some changes were necessary to make them work together harmoniously and accurately. The resulting method is freely available from GitHub (https://github.com/doesm/gameld).

In GaMD we identify a certain threshold energy that will serve to modify the system potential. Energies below the threshold will have a quadratic bias on top of the force field energy, while those above the threshold will not be affected. Typically, in GaMD a short run of cMD is used to determine the threshold energy according to different schemes.[17] In GaMELD, we use the threshold energy, E, of the highest temperature replica and calculate the threshold of the neighboring replica below as the difference between the previous threshold and the standard deviation of the potential energies simulated during a predefined number of steps. So, if we order the replicas, r, by temperature, from highest to lowest, the first replica will maintain the same threshold, $E_{r=1} = E_{r=1}$. For the next replicas, we will apply the following equation

$$E_r = E_{r-1} - \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i - \overline{e})^2} \quad (1)$$

where N is the number of steps simulated, e_i the potential energy of the system in each step and $\overline{e}$ the mean of the e_i values. This iterative process is performed twice for each replica within the simulation, once after cMD and again after GaMD equilibration (see Figure 1).

2.2. Benchmark Systems. We used a benchmark set of 12 small proteins (Table 1) to determine the ability of our new sampling strategy to sample and identify the native state starting from an unfolded state. We used three assessment metrics based on the root-mean-square deviation (RMSD) of each protein model to the native structure. The proteins and specific residues used to measure RMSD are provided in Table 1. Unless stated otherwise, the RMSD values reported correspond to the Cα atoms in regions that are well-defined in experimentally determined structures (e.g., excluding flexible tails or loops). Whenever possible, we compare our results with previous studies on these systems for comprehensive analysis and discussion. First, we looked at whether the method was able to sample native-like structures anywhere in the ensemble *BestStruct*. This is a basic test that allows us to find issues with the sampling strategy and/or force field. We then checked the ability of the force field/sampling strategy to identify the native state within the ensemble using statistical mechanics—either as the centroid of the highest population cluster (*Best1Pop*) or as the centroid of one of the top five clusters by population (*Best5Pop*) following CASP (critical assessment of structure prediction) standards.[23] Based on previous studies,[12] we considered an RMSD value of 4 Å or lower as indicative of a successful prediction.

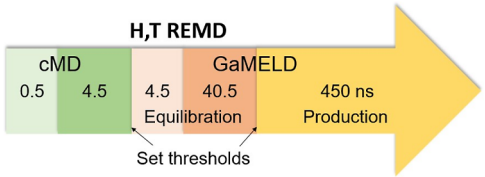


Figure 1. Schematic representation of the GaMELD method. Each replica performs three stages: conventional molecular dynamics (cMD), equilibration and production. In addition, there is a previous substage for cMD and equilibration. Energy thresholds are set after cMD and after equilibration. Times are expressed in nanoseconds.

Table 1. Protein Residues for RMSD

PDB ID	# residues	residues for RMSD	protein name
1FME[24]	28	5 to 26	structure of FSD-EY
1DV0[25]	47	4 to 42	C-terminal UBA domain of HHR23A
1FEX[26]	59	6 to 59[a]	MYB-domain of human RAP1
1HP8[27]	68	3 to 63	human P8-MTCP1
1LMB[28]	92	6 to 85[a]	lambda repressor−operator complex
1MI0[29]	65	9 to 65[a]	redesigned protein G variant NuG2
1PRB[30]	53	8 to 50[a]	albumin-binding domain
1UBQ[31]	76	1 to 72[a]	structure of ubiquitin
2HBA[32]	52	1 to 6, 17 to 52	N-terminal domain of ribosomal protein L9
2P6J[33]	52	5 to 48[a]	designed engrailed homeodomain variant UVF
2WXC[34]	47	9 to 27, 35 to 46[a]	BBL
3GB1[35]	56	1 to 56[a]	B1 domain of streptococcal protein G

[a]From ref 12.

2.3. Simulation Details. We modeled the systems using the all-atom force field ff14SB[36] to model protein side chains and the ff99SB[37] for modeling the backbone (ff14SBonlysc in AMBER)[38] with a cutoff of 1.8 nm, mbondi3 intrinsic radii and the implicit solvent model GB-Neck2.[39] Initial conformations were fully extended as generated by the tleap sequence command.[40]

For all simulations, we performed H,T-REMD[5,41] with a 2 fs time step during 500 ns. The number of replicas varied from 5 to 14 and exchange attempts occurred every 5 ps. Temperatures ranged from 300 to 450 K, increasing geometrically.

Each simulation consisted of 3 stages: cMD, GaMELD equilibration and GaMELD Production. The preparation time, where boost parameters are not updated, was 0.5 and 4.5 ns for the cMD and GaMELD equilibration respectively. The

https://doi.org/10.1021/acs.jctc.3c01019
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

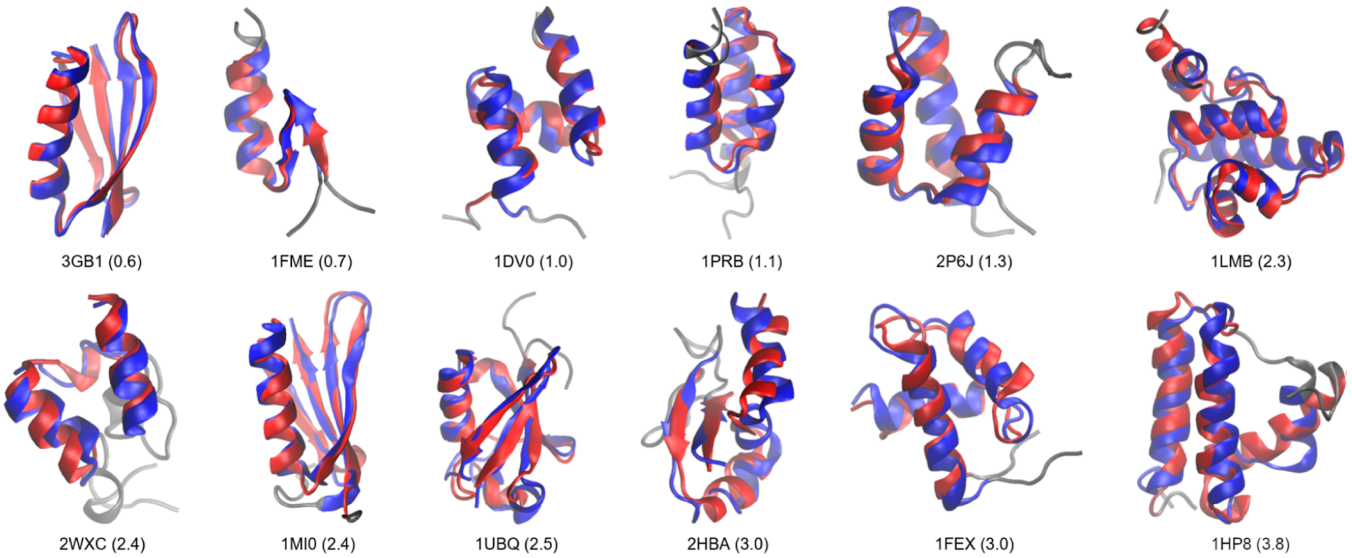


Figure 2. Superposition of the experimental structure in red and the best-predicted structure in blue, with unmeasured residues shown in gray. The RMSD in angstroms between the two structures is provided in parentheses.

Table 2. RMSD Comparison of Three Enhanced Sampling Methodologies against the Native State[a]

		BestStruc (Å)			Best1Pop (Å)			Best5Pop (Å)		
PDB ID	number of replicas	GaMELD	MELD	MELD[b]	GaMELD	MELD	MELD[b]	GaMELD	MELD	MELD[b]
3GB1	7	0.6	2.1	1.4	2.1	2.4	*7.9*	2.1	2.4	3.4
1FME	5	0.7	0.8	2.0	2.9	2.0	*7.7*	2.9	2.0	3.4
1DV0	6	1.0	4.0	0.9	3.7	*8.1*	1.0	2.6	*6.5*	1.0
1PRB	7	1.1	1.0	1.4	1.9	1.8	2.5	1.9	1.8	2.5
2P6J	7	1.3	1.6	1.7	2.7	2.3	2.7	2.7	2.3	2.7
1LMB	14	2.3	*5.2*	3.6	*10.9*	*9.5*	9.2	2.8	*8.2*	9.2
2WXC	7	2.4	3.3	2.2	*6.9*	*6.3*	*9.7*	*6.9*	*6.3*	*5.4*
1MI0	8	2.4	3.6	1.5	2.9	*8.0*	3.3	2.9	*7.9*	2.6
1UBQ	9	2.5	*4.7*	3.0	2.7	*5.4*	4.0	2.7	*5.4*	4.0
2HBA	7	3.0	3.2	0.9	3.9	3.6	*7.8*	3.9	3.6	*7.8*
1FEX	14	3.0	2.1	3.2	4.0	*6.7*	*8.9*	3.9	4.0	3.5
1HP8	8	3.8	3.7	3.2	*6.2*	*5.5*	*4.5*	*5.1*	*4.6*	*4.3*
success		100%	83%	100%	75%	42%	42%	83%	50%	67%

[a]Numbers in italics indicate the native state is not identified (Best1Pop,Best5Pop)/sampled (BestStruct). [b]From ref 12.

thresholds for GaMELD were set after 5 ns of cMD and then again after 45 ns of GaMELD equilibration (Figure 1). While GaMD allows for several options to boost the potential, here we opted for the boosting of the total potential energy only and used the upper-bound integration scheme.[17]

2.4. Coarse Physical Insights as Information. We used four types of coarse physical insights (CPI)[12] based on observations of globular proteins in the PDB (e.g., they have hydrophobic cores). We divide the four types into information originating from (i) secondary structure prediction,[42,43] (ii) hydrophobic packing, (iii) β-strand pairing, and (iv) confinement restraints for compact structures. Such information is inherently noisy, as we lack knowledge of the particular hydrophobic contacts (or strand pairs) that lead to native-like interactions. We provide such information as possible distance restraints—knowing that some will be incompatible with the system.

MELD uses Bayesian inference to identify the best interpretation of the data that is compatible with the physics of the system. This information is applied as strong distance restrictions at low-temperature replicas, while they weaken and vanish at higher replicas (higher temperatures). To configure the CPIs, we adopted the same input parameters and settings as previously described.[12]

2.5. Clustering Approach. Trajectory clustering was performed using the cluster command, a component of the CPPTRAJ program.[44] The analysis focused on the last half of the trajectory data, employing the hierarchical agglomerative (bottom-up) approach with epsilon set to 2[45] and adopting the single-linkage criteria (utilizing the shortest distance between members of two clusters). Clustering was conducted every 10 frames, and subsequently, all other frames were assigned to clusters based on their proximity to the nearest centroid.

3. RESULTS AND DISCUSSION

In the initial MELD simulations, as well as through different work over the years, the number of replicas employed in the one-dimensional H,T-REMD ladder was maintained at 30 as a compromise between sampling efficiency and computational cost. However, this number is significantly higher than the number one would use to expand the same temperature range using T-REMD.[38] This computational expense has long been a

https://doi.org/10.1021/acs.jctc.3c01019
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

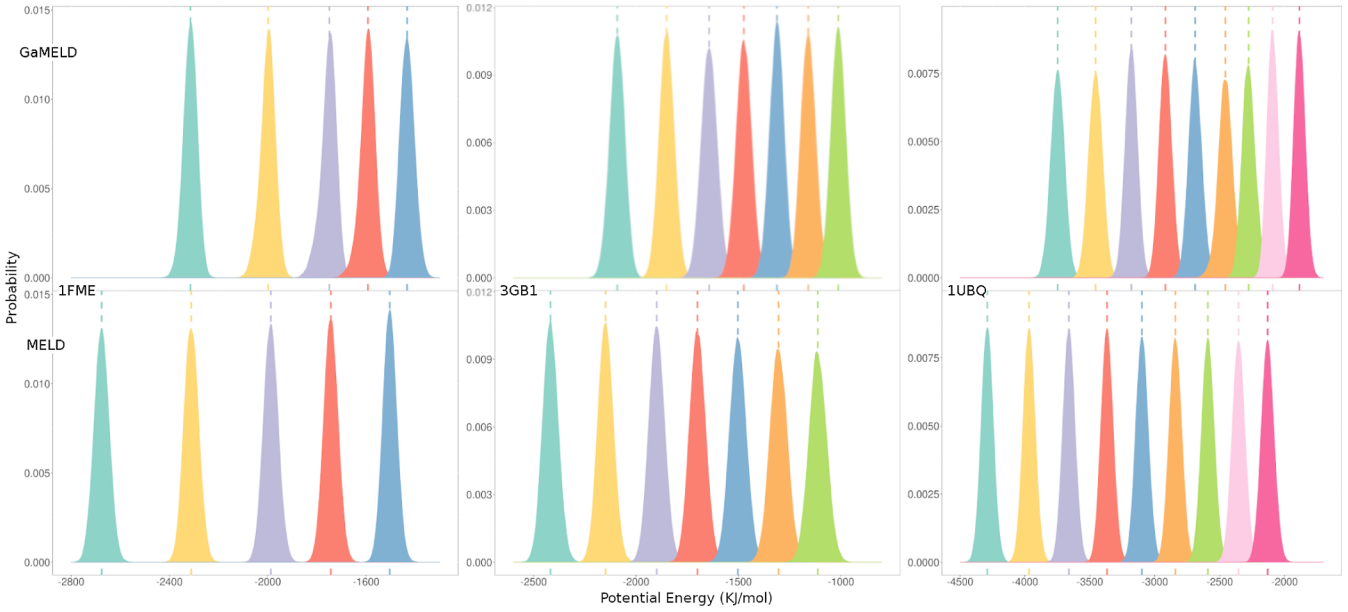


Figure 3. Kernel density estimates of Potential Energy for proteins ID 1FME (five replicas), 3GB1 (seven replicas) and 1UBQ (nine replicas). Replicas with the same number have the same color. The remaining graphs for the proteins can be found in the Supporting Information (Figure S1).

barrier to entry for the community to use MELD. For example, many accessible GPU-based clusters (e.g., through the ACCESS consortium) will rarely allow jobs with more than 16 GPUS. While MELD has the capacity of asynchronous REMD, allocating 15 GPUs for a 30-replica job doubles the amount of human time needed for calculations to complete. Thus, it would be desirable to identify protocols that maintain high accuracy while reducing the number of replicas and associated computational resources.

GaREUS[46] was recently developed as an enhanced sampling technique that merges replica-exchange umbrella sampling (REUS)[4] with Gaussian accelerated molecular dynamics (GaMD). REUS accelerates sampling along predefined reaction coordinates, while GaMD boosts the conformational sampling thereby accelerating the convergence of free-energy calculations while utilizing the same computational resources as REUS. In our current implementation, we use GaMELD to improve energy overlap across MELD replicas, thus reducing the overall computational cost of the simulations while improving the convergence and quality of the predictions.

The *BestStruc* results obtained for GaMELD indicate that our methodology is capable of sampling the correct structure in all cases. Figure 2 shows the *BestStruc* superposed with the experimental structure where most α-helix and β-sheet topologies are in perfect agreement. Furthermore, a comparison with MELD results using 30 replicas (see Table 2) shows that GaMELD is able to maintain an excellent performance while reducing the number of replicas. Both traditional MELD and GaMELD sample native states in the ensemble for all 12 systems in the study, but GaMELD is able to identify the native state through clustering more often—especially as the top cluster by population (*Best1Pop*). The average improvement over the 12 systems for *Best1Pop* is 1.5 Å, whereas the best structures in the ensemble remain similar, with a *BestStruct* average improvement of only 0.08 Å. Thus, MELD's ability to efficiently sample the native state remains similar, while the convergence rate that allows us to identify native states through clustering improves substantially. Table 2 also shows a greater number of proteins (nine systems) sampling the native state *Best1Pop* compared to vanilla MELD (five systems). Notably, GaMELD successfully predicted the structure of 1UBQ (Best1Pop = 2.7 Å), a slow-folding protein.[47] We additionally ran MELD simulations with the same number of replicas as GaMELD. In these conditions, GaMELD shows a surprising improvement in prediction accuracy over traditional MELD. In particular, MELD is no longer able to sample native conformations in the ensemble for all 12 systems, and the ability to identify a *Best5Pop* decreases to six of the 12 systems.

Initially, our goal was to set the total number of replicas as the square root of the number of residues for general applicability. This approach proved successful for nine out of the 12 proteins in our data set. However, upon a detailed analysis of the results, it became apparent that proteins 1DV0, 1LMB, and 1FEX required a different number of replicas to achieve optimal results. Surprisingly, there was no discernible correlation between the number of replicas and metrics such as the number of total round trips or local exchange probability. Nevertheless, we observed that the total number of round trips emerged as a valuable indicator for adjusting the number of replicas. Specifically, when the number of round trips fell below 100 for a 500 ns simulation, increasing the number of replicas was advantageous. Conversely, when the number of round trips exceeded 100, decreasing the number of replicas was more beneficial. This adaptive approach yielded enhancements in *Best1Pop* for both 1DV0 and 1FEX, as well as improvements to *Best5Pop* for 1LMB. These outcomes underscore the intricate interplay between the physics of the system and the data employed, resulting in complex folding routes across the folding transition—and the relevance of counting round trips to quantify the ease with which the current protein-data combination explores transitions above or below the folding point.

In our simulations, we consistently applied identical coarse physical insights (CPI) and secondary structure predictions, as previously detailed[12] (see Methods). However, the amount of information and its accuracy varied depending on the specific

https://doi.org/10.1021/acs.jctc.3c01019
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

protein sequence. Interestingly, when conducting simulations with an equivalent number of replicas, we observed that GaMELD consistently exhibited a higher exchange probability and a greater number of round trips across all considered systems (see Table S1). The number of round trips is especially important in REMD as it ensures random walks across the different conditions in the replica ladder, favoring convergence and overall exploration of the conformational space, as seen by the number of clusters obtained via identical clustering protocols (see Methods and Table S2).

As expected, the new approach is more efficient at sampling diverse sets of structures, resulting in a higher number of clusters. Surprisingly, a higher number of clusters does not necessarily lead to a better overall structure. For example, protein 2HBA sampled only 39 clusters in MELD for 1255 in GaMELD, yet the *Best1Pop* MELD result was marginally better than GaMELD (3.6 vs 3.9 Å respectively). Conversely, some systems for which both protocols identify a large number of clusters (the BBL protein, PDB ID: 2WXC) perform poorly in terms of finding the native state—which could point to a force field imbalance. Previous simulations with this protein identified similar issues using a different combination of force field and solvent model.[48] Although BBL is a fast folder, it has low stability compared to other proteins,[34] and previous reports indicate that differences in experimental conditions could affect its conformation.[49]

The combination of MELD and GaMD acts as a way to sculpt the energy landscape, with different emphasis. GaMD will increase the energy in regions below the threshold, while MELD will increase the energy in regions that are incompatible with any subset of data. Thus, effectively MELD focuses exploration along a few minima that might contain the native state, and GaMD flattens the energy of those regions, effectively shifting the energy distribution to higher energies. By having different thresholds for each replica (see Methods) we effectively decrease the energy gap between neighboring replicas, increasing the probability of exchange (see Figures 3 and S1−S2). Consequently, the number of round trips and the exchanges also increase (see Table S1). Notably, when GaMD is not used, the higher number of replicas does not always improve the results. For example, the MELD *Best1Pop* of 3GB1 with seven replicas and 1FME with five is substantially lower than the previous results with 30 replicas. This is in agreement with previous results showing how the number of replicas could impact differently depending on the nature of the process.[50−53]

A higher number of round trips should also lead to better convergence properties in the REMD approach. When selecting representative structures we rely on clustering populations. Typically, cluster populations over 30% give the user confidence that the system is predicting the native state. However, there are cases where clusters with high populations are not native-like (false positives) and cases where the top cluster population is below 30% (false negatives). While false negatives still capture the correct structure, the user would typically change the simulation protocol to increase convergence. Figure 4 shows that indeed GaMELD reduces the number of false positives while increasing the number of false negatives and true positives. One might think that the higher number of true positives might be due to lower overall exploration of conformational space. However, counting the number of clusters predicted by each approach (see Table S2) shows that GaMELD is able to sample a higher number of

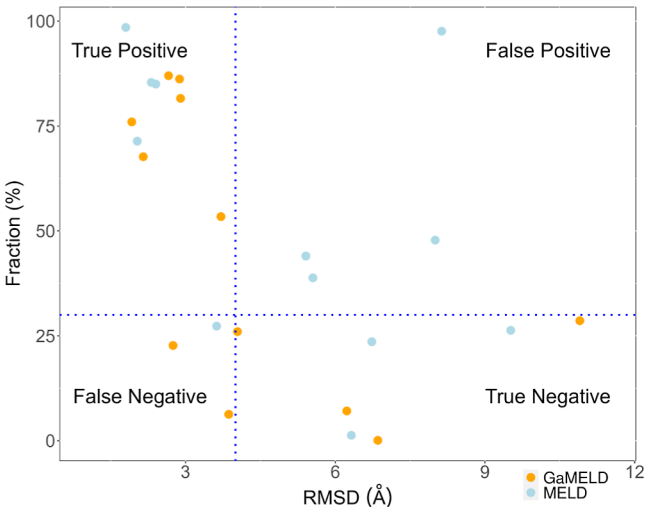


Figure 4. Most populated cluster for each protein utilizing MELD (light blue) and GaMELD (orange) is highlighted. Blue dotted lines at 4 Å and 30% signify the prediction's success.

distinct clusters. Similarly, Figure S3 shows the RMSD distribution over all replicas over time, again showing greater diversity and width of sampling using GaMELD. This demonstrates that GaMELD surpasses energy barriers more swiftly, preventing the sampling from getting trapped in an energy basin. Despite these advances, most simulations are not converged for both methodologies as shown by the RMSD density plots in Figure S4. One could argue that GaMELD simulations are closer to convergence as shown by greater overlap over independent walkers. However, proteins like protein G (3gb1 and 1mi0) or ubiquitin (1UBQ) show that only a few walkers have sampled the native state.

Finally, there is a significant difference between MELD and GaMD. MELD guarantees that the relative Boltzmann weight between different clusters that satisfy different subsets of data is the same as in the unbiased force field. On the other hand, GaMD boosts the potential of regions below the threshold (the different minima we are interested in). Since the biasing potential is quadratic, it could alter the weight between different minima compatible with MELD data. Hence, we used GaMD-based toolkits[54] to unbias the lowest temperature replica from GaMELD simulations (see Figure 5). Indeed the boosting potential follows a Gaussian distribution, which enabled accurate reweighting of the simulations using cumulant expansion to the second-order (Figure 5b). By reweighting the GaMELD simulations, we verified that indeed the native conformations further increased in population once the bias was removed (as depicted in Figure 5a). GaMELD uses the lowest replica to identify the native state, but one could look at the ensemble of replicas to identify the computational folding pathways. However, the GaMELD folding pathways are not necessarily related to the native folding pathways as MELD biases are compatible with native-like structures, and not necessarily with intermediates.[55]

While we set out to integrate GaMD and MELD to improve MELD predictions, we were also interested in understanding how MELD improves GaMD. Thus, we compared GaMD REMD (GaREMD) simulations (no CPIs) with those using GaMELD in a subset of three proteins for which REMD data was also available.[38] We ran GaMELD simulations for 0.5 μs and GaREMD simulations for 2 μs. Table 3 summarizes our

https://doi.org/10.1021/acs.jctc.3c01019
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

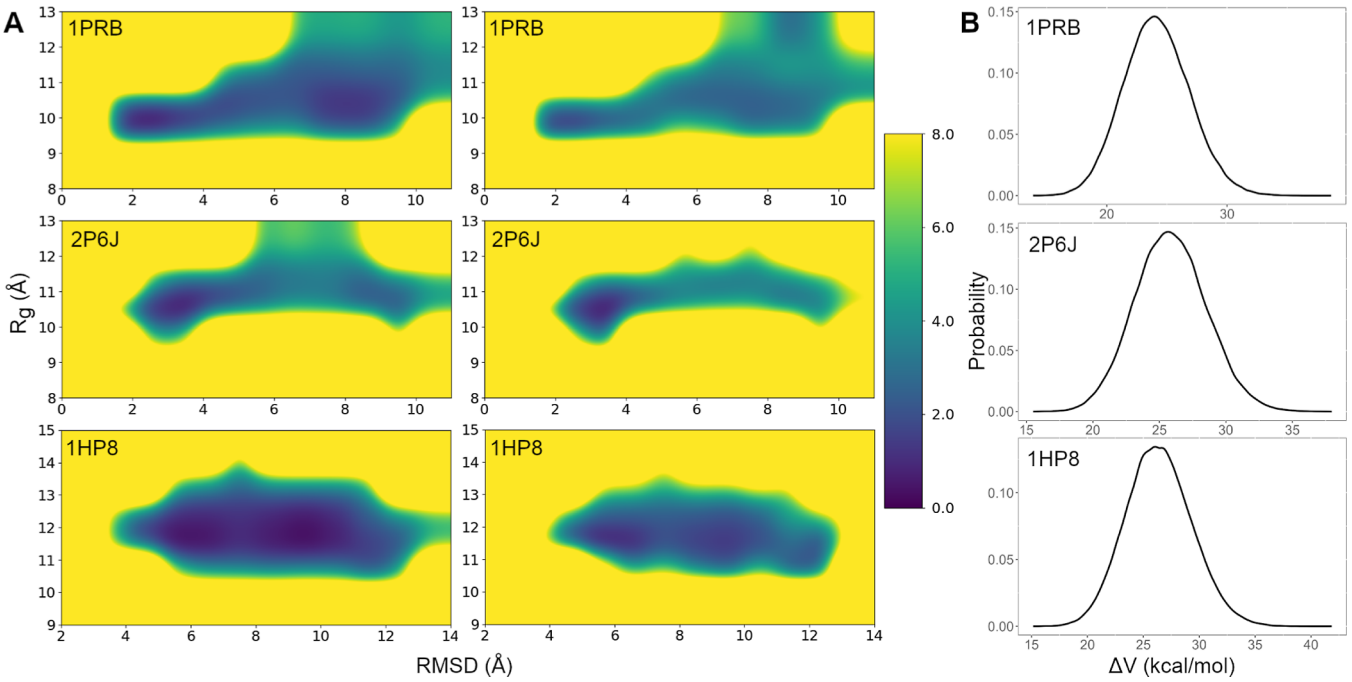


Figure 5. (A) 2D (RMSD, Rg) PMF. Left: biased. Right: unbiased (reweighted). (B) Gaussian distribution of the boost potential ΔV.

Table 3. Comparison of GaMELD with Replica Exchange with and without GaMD

	number of replicas			time (μs)			BestStruc (Å)			Best1Pop (Å)		
PDB ID	GaMELD	GaREMD	REMD[a]	GaMELD	GaREMD	REMD[a]	GaMELD	GaREMD	REMD[a]	GaMELD	GaREMD	REMD[a]
2P6J	7	7	16	0.5	2.0	3.5	1.3	1.7	1.6	2.7	3.2	3.2
2WXC	7	7	16	0.5	2.0	2.2	2.4	3.3	2.1	6.9	7.2	8.3
2HBA	7	7	10	0.5	2.0	21.2	3.0	5.6	1.6	3.9	8.1	6.0

[a]From ref 38.

results, revealing that even running shorter simulations (and a lower number of replicas than REMD) resulted in improved RMSD for the *Best1Pop* cluster. For 2WXC, which we mentioned as problematic before, none of the systems captured it through clustering—but, GaMELD sampled the native state in the ensemble in a fraction of the time.

4. CONCLUSIONS

This study introduces GaMELD, a novel method that merges GaMD with MELD, offering an improved strategy for exploring the intricate energy landscapes of biomolecules. We have benchmarked the method against a set of 12 proteins, showing that GaMELD consistently delivered accurate predictions with fewer replicas, substantially reducing computational resources. While the computational savings will depend on the particular system, the average saving for the 12 proteins studied in this work is 22 GPUs per simulation. By introducing Gaussian biases to modulate the overlap in energy distributions along replicas, the method increases the number of round trips in replica space, increases the total number of clusters identified in the ensemble, and reduces the amount of false positive clusters identified.

Like data, software should be in accordance with FAIR principles[56] (findable, accessible, interoperable, reusable). MELD has been freely accessible through GitHub for several years (https://github.com/doesm/gameld), addressing various biological challenges.[14,15,57−59] Our current research demonstrates its interoperability with other enhanced sampling methods, surpassing the individual efficacy of each method. Although we integrated it with GaMD in this study, the potential exists to combine MELD with alternative enhanced sampling techniques, such as metadynamics,[60] which has already been adapted for use with multiple parallel replicas.[61] These future possibilities open exciting avenues for tackling complex biomolecular systems.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c01019.

Comparison of averaged round trips and exchanges between GaMELD a MELD, Number of clusters, Clustering and FPT, Kernel density estimates of Potential Energy, Acceptance probability, Comparison of RMSD overlap of all replicas and RMSD density (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

https://doi.org/10.1021/acs.jctc.3c01019
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Author

Marcelo Caparotta − *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c01019

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.P. acknowledges support from a CAREER award from the National Science Foundation (CHE-2235785) to develop the GaMELD methodology.

■ REFERENCES

(1) Mitsutake, A.; Mori, Y.; Okamoto, Y. Enhanced sampling algorithms. *Biomol. Simulat. Methods Protoc.* **2013**, *924*, 153−195.

(2) Miao, Y.; McCammon, J. A. Unconstrained enhanced sampling for free energy calculations of biomolecules: a review. *Mol. Simul.* **2016**, *42*, 1046−1055.

(3) Sugita, Y.; Kamiya, M.; Oshima, H.; Re, S. Replica-exchange methods for biomolecular simulations. *Biomol. Simulat. Methods Protoc.* **2019**, *2022*, 155−177.

(4) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058−9067.

(5) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(6) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(7) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(8) Huang, K.-F.; Dong, S.-H.; Zhong, S.-S.; Li, H.; Duan, L.-L. Study on the amyloid Aβ42 with accelerated molecular dynamics simulations. *Commun. Theor. Phys.* **2019**, *71*, 1121.

(9) Grasso, G.; Danani, A. Molecular simulations of amyloid beta assemblies. *Adv. Phys.: X* **2020**, *5*, 1770627.

(10) Dill, K. A.; MacCallum, J. L. The protein-folding problem, 50 years on. *science* **2012**, *338*, 1042−1046.

(11) Kuhlman, B.; Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 681−697.

(12) 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.

(13) 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.

(14) 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*, No. e1601274.

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

(16) 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.

(17) Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J. Chem. Theory Comput.* **2015**, *11*, 3584−3595.

(18) An, X.; Bai, Q.; Bing, Z.; Liu, H.; Yao, X. Insights into the molecular mechanism of positive cooperativity between partial agonist MK-8666 and full allosteric agonist AP8 of hGPR40 by Gaussian accelerated molecular dynamics (GaMD) simulations. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3978−3989.

(19) Ricci, C. G.; Chen, J. S.; Miao, Y.; Jinek, M.; Doudna, J. A.; McCammon, J. A.; Palermo, G. Deciphering off-target effects in CRISPR-Cas9 through accelerated molecular dynamics. *ACS Cent. Sci.* **2019**, *5*, 651−662.

(20) Wang, J.; Arantes, P. R.; Bhattarai, A.; Hsu, R. V.; Pawnikar, S.; Huang, Y.-m. M.; Palermo, G.; Miao, Y. Gaussian accelerated molecular dynamics: Principles and applications. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e1521.

(21) 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*, No. e1005659.

(22) Copeland, M. M.; Do, H. N.; Votapka, L.; Joshi, K.; Wang, J.; Amaro, R. E.; Miao, Y. Gaussian accelerated molecular dynamics in OpenMM. *J. Phys. Chem. B* **2022**, *126*, 5810−5820.

(23) Mirjalili, V.; Feig, M. Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. *J. Chem. Theory Comput.* **2013**, *9*, 1294−1303.

(24) Sarisky, C. A.; Mayo, S. L. The $\beta\beta\alpha$ fold: explorations in sequence space. *J. Mol. Biol.* **2001**, *307*, 1411−1418.

(25) Withers-Ward, E. S.; Mueller, T. D.; Chen, I. S.; Feigon, J. Biochemical and structural analysis of the interaction between the UBA (2) domain of the DNA repair protein HHR23A and HIV-1 Vpr. *Biochemistry* **2000**, *39*, 14103−14112.

(26) Hanaoka, S.; Nagadoi, A.; Yoshimura, S.; Aimoto, S.; Li, B.; De Lange, T.; Nishimura, Y. NMR structure of the hRap1Myb motif reveals a canonical three-helix bundle lacking the positive surface charge typical of Myb DNA-binding domains. *J. Mol. Biol.* **2001**, *312*, 167−175.

(27) Barthe, P.; Yang, Y.-S.; Chiche, L.; Hoh, F.; Strub, M.-P.; Guignard, L.; Soulier, J.; Stern, M.-H.; van Tilbeurgh, H.; Lhoste, J.-M.; et al. Solution structure of human p8MTCP1, a cysteine-rich protein encoded by the MTCP1 oncogene, reveals a new α-helical assembly motif. *J. Mol. Biol.* **1997**, *274*, 801−815.

(28) Beamer, L. J.; Pabo, C. O. Refined 1.8 Å crystal structure of the λ repressor-operator complex. *J. Mol. Biol.* **1992**, *227*, 177−196.

(29) Nauli, S.; Kuhlman, B.; Le Trong, I.; Stenkamp, R. E.; Teller, D.; Baker, D. Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2. *Protein Sci.* **2002**, *11*, 2924−2931.

(30) Johansson, M. U.; de Château, M.; WikstroÈm, M.; Forsén, S.; Drakenberg, T.; Björck, L. Solution Structure of the Albumin-Binding GA Module: A Versatile Bacterial Protein Domain. *J. Mol. Biol.* **1997**, *266*, 859.

(31) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of ubiquitin refined at 1.8 Åresolution. *J. Mol. Biol.* **1987**, *194*, 531−544.

(32) Cho, J.-H.; Meng, W.; Sato, S.; Kim, E. Y.; Schindelin, H.; Raleigh, D. P. Energetically significant networks of coupled interactions within an unfolded protein. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 12079−12084.

(33) Shah, P. S.; Hom, G. K.; Ross, S. A.; Lassila, J. K.; Crowhurst, K. A.; Mayo, S. L. Full-sequence computational design and solution

structure of a thermostable protein variant. *J. Mol. Biol.* **2007**, *372*, 1−6.
(34) Neuweiler, H.; Sharpe, T. D.; Rutherford, T. J.; Johnson, C. M.; Allen, M. D.; Ferguson, N.; Fersht, A. R. The folding mechanism of BBL: Plasticity of transition-state structure observed within an ultrafast folding protein family. *J. Mol. Biol.* **2009**, *390*, 1060−1073.
(35) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.
(36) 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.
(37) 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.
(38) Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* **2014**, *136*, 13959−13962.
(39) Nguyen, H.; Roe, D. R.; Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **2013**, *9*, 2020−2034.
(40) Case, D.; Belfon, K.; Ben-Shalom, I.; Brozell, S.; Cerutti, D.; Cheatham, T.; Cruzeiro, V.; Darden, T.; Duke, R.; Giambasu, G.; et al. *AMBER 2020*; University of California: San Francisco, 2020.
(41) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **2000**, *113*, 6042−6051.
(42) Buchan, D. W.; Jones, D. T. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **2019**, *47*, W402−W407.
(43) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne. *J. Mol. Biol.* **1999**, *292*, 195−202.
(44) 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.
(45) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.
(46) Oshima, H.; Re, S.; Sugita, Y. Replica-exchange umbrella sampling combined with Gaussian accelerated molecular dynamics for free-energy calculation of biomolecules. *J. Chem. Theory Comput.* **2019**, *15*, 5199−5208.
(47) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 5915−5920.
(48) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517−520.
(49) Settanni, G.; Fersht, A. R. Downhill versus Barrier-Limited Folding of BBL: 3. Heterogeneity of the Native State of the BBL Peripheral Subunit Binding Domain and Its Implications for Folding Mechanisms. *J. Mol. Biol.* **2009**, *387*, 993−1001.
(50) Nymeyer, H. How Efficient Is Replica Exchange Molecular Dynamics? An Analytic Approach. *J. Chem. Theory Comput.* **2008**, *4*, 626−636.
(51) Liwo, A.; Czaplewski, C.; Ołdziej, S.; Scheraga, H. A. Computational techniques for efficient conformational sampling of proteins. *Curr. Opin. Struct. Biol.* **2008**, *18*, 134−139.
(52) Iwai, R.; Kasahara, K.; Takahashi, T. Influence of various parameters in the replica-exchange molecular dynamics method: Number of replicas, replica-exchange frequency, and thermostat coupling time constant. *Biophys. Physicobiol.* **2018**, *15*, 165−172.
(53) Knapp, B.; Ospina, L.; Deane, C. M. Avoiding false positive conclusions in molecular simulation: the importance of replicas. *J. Chem. Theory Comput.* **2018**, *14*, 6127−6138.
(54) Miao, Y.; Sinko, W.; Pierce, L.; Bucher, D.; Walker, R. C.; McCammon, J. A. Improved reweighting of accelerated molecular dynamics simulations for free energy calculation. *J. Chem. Theory Comput.* **2014**, *10*, 2677−2689.
(55) Chang, L.; Perez, A. Deciphering the Folding Mechanism of Proteins G and L and Their Mutants. *J. Am. Chem. Soc.* **2022**, *144*, 14668−14677.
(56) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; t'Hoen, P. A.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
(57) Esmaeeli, R.; Bauzá, A.; Perez, A. Structural predictions of protein−DNA binding: MELD-DNA. *Nucleic Acids Res.* **2023**, *51*, 1625−1636.
(58) Liu, C.; Brini, E.; Perez, A.; Dill, K. A. Computing Ligands Bound to Proteins Using MELD-Accelerated MD. *J. Chem. Theory Comput.* **2020**, *16*, 6377−6382.
(59) Brini, E.; Kozakov, D.; Dill, K. Predicting Protein Dimer Structures Using MELD × MD. *J. Chem. Theory Comput.* **2019**, *15*, 3381−3389.
(60) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12566.
(61) Invernizzi, M.; Piaggi, P. M.; Parrinello, M. Unified approach to enhanced sampling. *Phys. Rev. X* **2020**, *10*, 041034.

https://doi.org/10.1021/acs.jctc.3c01019
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX