

# UCSF

## UC San Francisco Previously Published Works

### Title

Constraint methods that accelerate free-energy simulations of biomolecules

### Permalink

<https://escholarship.org/uc/item/7mj6293x>

### Journal

The Journal of Chemical Physics, 143(24)

### ISSN

0021-9606

### Authors

Perez, Alberto  
MacCallum, Justin L  
Coutsias, Evangelos A  
[et al.](#)

### Publication Date

2015-12-28

### DOI

10.1063/1.4936911

Peer reviewed

## Constraint methods that accelerate free-energy simulations of biomolecules

Alberto Perez,<sup>1</sup> Justin L. MacCallum,<sup>2</sup> Evangelos A. Coutsias,<sup>1,3</sup> and Ken A. Dill<sup>1,4</sup>

<sup>1</sup>*Lauffer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, USA*

<sup>2</sup>*Department of Chemistry, University of Calgary, Calgary, Alberta T2N 1N4, Canada*

<sup>3</sup>*Department of Applied Mathematics, Stony Brook University, Stony Brook, New York 11794, USA*

<sup>4</sup>*Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, USA*

(Received 4 September 2015; accepted 18 November 2015; published online 9 December 2015)

Atomistic molecular dynamics simulations of biomolecules are critical for generating narratives about biological mechanisms. The power of atomistic simulations is that these are physics-based methods that satisfy Boltzmann's law, so they can be used to compute populations, dynamics, and mechanisms. But physical simulations are computationally intensive and do not scale well to the sizes of many important biomolecules. One way to speed up physical simulations is by coarse-graining the potential function. Another way is to harness structural knowledge, often by imposing spring-like restraints. But harnessing external knowledge in physical simulations is problematic because knowledge, data, or hunches have errors, noise, and combinatoric uncertainties. Here, we review recent principled methods for imposing restraints to speed up physics-based molecular simulations that promise to scale to larger biomolecules and motions. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4936911>]

### INTRODUCTION

An important source of insight into the structures and mechanisms of proteins and other biomolecules is atomistic computer simulations.<sup>1–7</sup> To design drugs or to create narratives of how the intricate mechanisms of proteins are encoded within their molecular structures often requires capturing fine spatial and temporal detail. Experimental science alone generally provides only information that is too coarse to tell the stories of the hundreds of thousands of structures and mechanisms of the protein universe. Metaphorically, you cannot explain the workings of a car engine from seeing a few distant snapshots. At present, molecular modeling is the only way to fill in the Angstrom-by-Angstrom and picosecond-by-picosecond details needed to provide complete narratives of biological mechanisms.<sup>8,9</sup>

In broad terms, computer simulations of biomolecules draw from two sources of insights. First, atomistic simulations draw upon knowledge of the *energetics* of the intra- and inter-molecular interactions, which is encoded within semi-empirical forcefields.<sup>10–12</sup> Second, some computer modeling of proteins draws upon *structural knowledge*, either of similar proteins, such as in the Protein Data Bank (PDB)<sup>13</sup>—a database of more than 100 000 structures known at atomic detail—or from experimental studies by NMR,<sup>14</sup> EPR,<sup>15,16</sup> or electron microscopy, or from knowledge of related amino acid sequences.<sup>17–19</sup> In the long run, it is important to develop methods that combine the advantages of energetic and structural modeling. Forcefield modeling alone is computationally expensive and does not scale well to larger systems. Structure-based methods alone do not give proper Boltzmann populations, and so are not amenable to

constructing narratives about mechanisms, or of how those mechanisms are encoded within structures and dynamics. There is great value in joining together structural and energetic approaches to biomolecule modeling. However, in practice, this has been challenging.

The field called *integrative structural biology*<sup>20–23</sup> attempts this synthesis, but falls short in one critical way: it does not attempt to provide proper free energies (said differently, it does not provide proper populations, give the proper Boltzmann distribution, or satisfy detailed balance at equilibrium). Without proper free energies, we cannot obtain physically plausible dynamics or mechanisms. The two most prominent methods satisfying detailed balance for the study of macromolecules are molecular dynamics (MD)<sup>1</sup> and Monte Carlo (MC)<sup>24</sup> techniques. The approach called *coarse-graining*<sup>25,26</sup> seeks to give proper free energies but makes other crucial trade-offs. Coarse-graining reduces the numbers of degrees of freedom, based upon choices that must be made in advance depending on what properties are to be captured in modeling and which can be sacrificed. But for many problems of biomolecules, it is not clear in advance which degrees of freedom are important and which are not.

Here, we summarize a different general approach to targeted searching. These approaches keep the fine-grained detail at all stages of the atomistic simulation, by harnessing external structural knowledge or insights. So, these methods do not lose the detailed information from the atomistic forcefield, they preserve Boltzmann populations, and they explore states of interest with high efficiency. In addition, these approaches are fairly robust against noise, errors, and uncertainties of the supplied external knowledge, and promise to scale to larger systems.

One longstanding method for restraint-based targeted sampling of conformational space is umbrella sampling,<sup>27</sup> where restraints are imposed to drive the sampling of a conformational transition that would otherwise not happen or take too long to simulate. In umbrella sampling, the simulation is guided along a path from one state to another. When the overlap of sampled populations is good enough, reweighting techniques such as the weighted histogram analysis method (WHAM)<sup>28</sup> can then be used to compute the free energy change for the overall process. Another approach is metadynamics<sup>29–32</sup>—which uses an adaptive biasing potential to guide simulations. In this method, the conformations are described by collective variables. While sampling conformational space, history-dependent Gaussian penalties are added to make conformations already visited less likely, increasing the sampling of rare events. When the sampling is converged, the free energy change can be recovered from the Gaussian penalties. These methods are successful and widely used, but they require a well-defined starting and ending state, and they can be slow if the endstate is very different from the starting state. The methods described below have somewhat broader applicability.

We describe three different restraint methods here that can speed up otherwise computationally costly searching and sampling processes.<sup>33,34</sup> One is **MELD**<sup>35</sup> (Modeling Employing Limited Data). It uses a Bayesian inference approach within MD to harness external constraint information that can be fuzzy, incomplete, or inconsistent to search different possible beginning or end states. Another is **CCR** (Confine-Convert-Release).<sup>36–38</sup> CCR is a method for computing the free energy of a conformational transition of a biomolecule from state *A* to *B*, by variable constraints, where *A* and *B* can be very different conformations. And third, we describe **AIKC** (Algebraic Inverse Kinematics Closure) for efficiently modeling loops and macrocycles and large conformational spaces of molecules that have known, but sparse, structural constraints.

## SPRING-LIKE RESTRAINTS HELP TO SPEED UP MD SIMULATIONS

Biomolecular energy landscapes are large and rugged. The energy landscape for the conformational space of a protein can be approximated by a forcefield energy,  $E_i^{ff}(\mathbf{r})$ , where  $\mathbf{r}$  represents one conformation of the chain. Boltzmann's law gives the state population  $p_i(\mathbf{r})$  and partition function  $Z$  as

$$p(\mathbf{r}) = \frac{\exp[-\beta E_i^{ff}(\mathbf{r})]}{Z}; \quad Z = \sum \exp[-\beta E_i^{ff}(\mathbf{r})], \quad (1)$$

where  $p$  is the probability of a particular microstate,  $\mathbf{r}$  is the full conformational vector,  $Z$  is the partition function, and  $\beta = 1/(k_B T)$ . Our goal is usually to find the most highly populated states, such as a protein's native structure.

A standard way of introducing structural information into molecular simulations is by using “spring-like” forces—usually either parabolic or flat-bottomed restraining potentials. If we want to enforce that residues *i* and *j* are nearby to each other in space in some molecular structure being simulated, then a term is added to the standard potential energy function that drives *i* towards *j*. On the one hand,

$$\frac{F}{kT} = - \sum_{ij} p_{ij} \left[ \underbrace{\ln p_{ij}}_{\text{average}} - \underbrace{\beta \epsilon_{ij} - \lambda (\epsilon_{ij} - \bar{\epsilon})^2}_{\text{variance}} \right]$$

Distance constraint:  $k_s (d_{ij} - d_{ij_0})^2$




FIG. 1. The standard way to impose structural information within MD simulations. Spring-like potentials are imposed as restraints to enforce the knowledge that residues *i* and *j* are in close proximity, and a spring-constant imposes a variance, or uncertainty, in our knowledge.

this might seem *ad hoc*. On the other hand, it is fully in the spirit of the modern interpretations of statistical mechanics as an inference procedure for choosing an optimal model given limited information.<sup>39</sup>

Figure 1 illustrates the standard inference made in the canonical ensemble of statistical mechanics, where minimizing a free energy is equivalent to maximizing an entropy subject to satisfying an experimental observable, such as the temperature (or equivalently, the average energy,  $\langle E \rangle = \sum_{ij} p_{ij} \epsilon_{ij}$ ). Statistical physics gives a procedure for inferring the full distribution function  $p_{ij}(\mathbf{r})$ , given only a single given first-moment experimental observable. In this context, adding springs is simply equivalent to enforcing structural information. A spring enforces an estimated average distance  $\langle d_{ij} \rangle$  between *i* and *j*. At the same time, the spring constant essentially enforces a measure of our uncertainty of that knowledge, by fixing the variance ( $\langle d_{ij}^2 \rangle - \langle d_{ij} \rangle^2$ ). Knowledge of multiple constraints can readily be enforced by multiple such spring-like potentials.

A key point, however, is that enforcing a constraint (*i,j*) in a MD simulation only works if that constraint is truly present and accurately represented by the spring model, placement, and variance. But, enforcing a wrong constraint (*i,j*) will just misdirect a MD simulation to search wrong structures. A common reality is that external knowledge comes with some right information and some wrong information and does not tell us which is which. Solving this problem is the essence of the MELD method,<sup>35</sup> described below.

## MELD: COMBINING SUBSETS OF INFORMATION WITH MOLECULAR SIMULATIONS

Because the energy landscape of a folding protein is large and rugged, finding the states of high population can be prohibitively slow. Ideally, we want to sample only near high-population regions. This can be done by adding focusing restraints. Applying restraints changes the expressions for the populations and partition function,

$$p_i'(\mathbf{r}) = \frac{\exp[-\beta E_i^{ff}(\mathbf{r}) - \beta E^{rest}(\mathbf{r})]}{Z'}; \quad (2)$$

$$Z' = \sum \exp[-\beta E_i^{ff}(\mathbf{r}) - \beta E^{rest}(\mathbf{r})],$$

where  $E^{rest}(\mathbf{r})$  is the restraint energy. Conformations that are compatible with the restraints will have  $E^{rest}(\mathbf{r}) = 0$ , and those that are not compatible will have  $E^{rest}(\mathbf{r}) > 0$ . So, conformations that are compatible with the restraints will be more populated than in unrestrained simulations. The goal is to choose a set of restraints that is compatible with the native state (or any particular state of interest). Flat-bottomed potentials add no restraint energy near the targeted conformation but increase the energy everywhere else,

$$E^{rest}(\mathbf{r}) = \begin{cases} \frac{1}{2}k(r - \mathbf{r}_1)^2 & \text{if } r < \mathbf{r}_1 \\ 0 & \text{if } \mathbf{r}_1 \leq r \leq \mathbf{r}_2 \\ \frac{1}{2}k(r - \mathbf{r}_2)^2 & \text{if } r > \mathbf{r}_2 \end{cases}, \quad (3)$$

where  $E^{rest}$  is the restraint energy and  $k$  is the force constant.

But typically, the external knowledge that is supplied is corrupted in various ways, limiting our ability to convert it into inference springs that can accelerate a MD simulation. Here are three forms of corruption: (1) Information is too sparse. Some experiments, such as solid state NMR, provide accurate distances, but just not enough of them to fully determine a protein structure. (2) Information is ambiguous. Some experiments, such as EPR, indicate that residues  $i$  and  $j$  are in proximity of each other but does not tell us whether they are within 5 or 35 Å. (3) Information is uncertain. Some information, such as from databases of sequences or structures, tells us only *what might be true*, not what is exactly true. For example, in the homology modeling of one protein from another, we do not know exactly which residues of our query protein correspond to which residues of some known target protein. Recently, a method called MELD has been developed that can leverage this corrupted knowledge to speed up population-preserving MD simulations for finding target states.

To describe the MELD method, we first reformulate the problem in terms of Bayes' relationship,

$$p(\mathbf{r}|D) = \frac{p(D|\mathbf{r})p(\mathbf{r})}{p(D)} \sim p(D|\mathbf{r})p(\mathbf{r}), \quad (4)$$

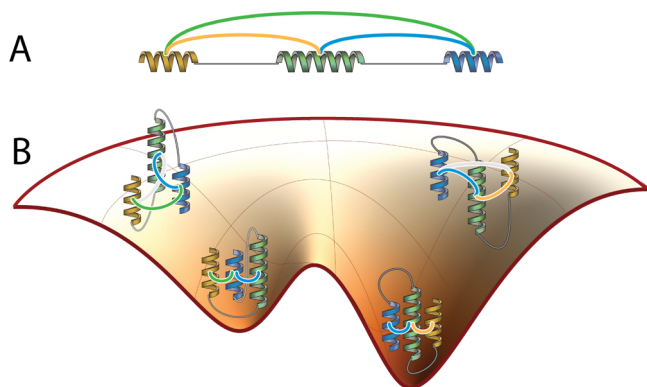


FIG. 2. The basic idea of MELD, shown on an energy landscape. Different molecular conformations “choose” which subsets of knowledge restraints are imposed within different regions of the space, indicated as two deep wells here. Adapted with permission from A. Perez, J. L. MacCallum, and K. Dill, Proc. Natl. Acad. Sci. U. S. A. 112, 11846 (2015). Copyright 2015 National Academy of Sciences, USA.

where  $\mathbf{D}$  represents the experimental data and  $\mathbf{r}$  is the chain conformation. Here,  $p(\mathbf{r})$  is the prior probability distribution of  $\mathbf{r}$ , given in our case by the Boltzmann distribution of conformations from the underlying force field.  $p(\mathbf{D}|\mathbf{r})$  is the likelihood of observing the external data, given the structure  $\mathbf{r}$ .  $P(\mathbf{D})$  is a normalization factor that can be ignored for our purposes. In MELD, rather than using the full set of data, we use subsets (see Figure 2), recognizing that some of the data might be noisy or ambiguous. We specify a degree of reliability.

For example, suppose we want to use NMR data to refine a protein structure in molecular simulations. Suppose we know that roughly 80% of the identified peaks in the experiment are correct (but we do not know which ones). Then, we take these data as having a reliability of 0.8. The problem is that different NMR peaks can be mapped to different possible contacts between residues, but we do not know the mapping. So each peak is a different “group,” one collection of all possible restraints that could explain that peak. We require one “explanation” of each peak, so one restraint from the group must be satisfied. These groups then are combined into a single collection, where 80 percent of the restraints must be satisfied.

As another example, suppose our external knowledge comes from web servers of secondary structures.<sup>41,42</sup> Such predictions are known to be about 70% accurate. So, we take the degree of reliability in this case to be 70%. It means that when MELD is choosing subsets of experimental knowledge, each subset is taken to be one particular collection of 70% of the possible secondary structure constraints. Or, if we want a computer simulation to form a hydrophobic core from all possible hydrophobic pairings, we know that only about 8% of all possible hydrophobic pairings will be correct for any given native structure. Different types of knowledge have different reliabilities.

MELD satisfies detailed balance. To do this, MELD takes the minimum restraint energy restraints for the conformation it is sampling. MELD calculates all of the possible restraint energies for a given structure, sorting the restraints in each group by energy and choosing ones having the lowest free energy, until we reach the reliability count number.

In order to hop over energy barriers, MELD samples using replica exchange (REMD) of both the Hamiltonian and the temperature; see Fig. 2. In this way, at high replica index the restraint force constant is 0 and the temperature high, allowing simulations to sample broadly. At lower replica indices, the force constant increases and the temperature decreases, so conformations are forced to sample where the springs are active. As noted above, where  $E^{rest} = 0$ , the relative populations sampled will be the same as in an unbiased simulation.

In some respects, MELD is designed to tackle problems much like those that are being explored in the new field of “computational rationality,” which considers how people and computers can make rapid choices in complex decision environments<sup>43</sup> and how people make decisions in complex games.<sup>44</sup>

## MELD IS USEFUL FOR PROTEIN-STRUCTURE REFINEMENT

Here, we summarize how MELD can be useful in protein-structure determination, for different sources of experimental

data including solid state NMR (sparse data set), EPR (ambiguous and sparse), or evolutionary restraints (noisy and sparse) on a set of proteins ranging from 56 to 166 residues.

Figure 3 shows that MELD can utilize troublesome external information of these various types, and when coupled with REMD simulations, obtain excellent native structures. Because these particular structures were also already known from independent studies, this just illustrates how MELD can find correct structures when the constraining data are problematic. The MELD method improves upon prior methods, such as X-plor, that are less robust to troublesome data.

A wide variety of types of data can be utilized within MELD for structure determination or structure prediction. Proofs of principle now exist<sup>35</sup> for solid-state NMR,<sup>14</sup> double electron-electron resonance (DEER),<sup>15,45,47</sup> and evolutionary

data.<sup>18</sup> Further development is needed to incorporate data from other experimental techniques such as SAXS,<sup>48</sup> cryoEM,<sup>49</sup> or molecular electron tomography.<sup>50</sup> MELD can be used with any kind of data that can be modeled as a set of restraints, and to which an accuracy value can be assigned. Of course, if we incorrectly instruct MELD to believe that the data reliability is better than it actually is, there is no guarantee it will give reliable structures.

### MELD + CPI (COARSE PHYSICAL INSIGHTS) IS USEFUL FOR PROTEIN-STRUCTURE PREDICTION

Here is another challenge that benefits from the MELD approach. Currently, predicting the native structure of a protein from the fully denatured state using brute-force atomistic simulations is largely out of reach of present

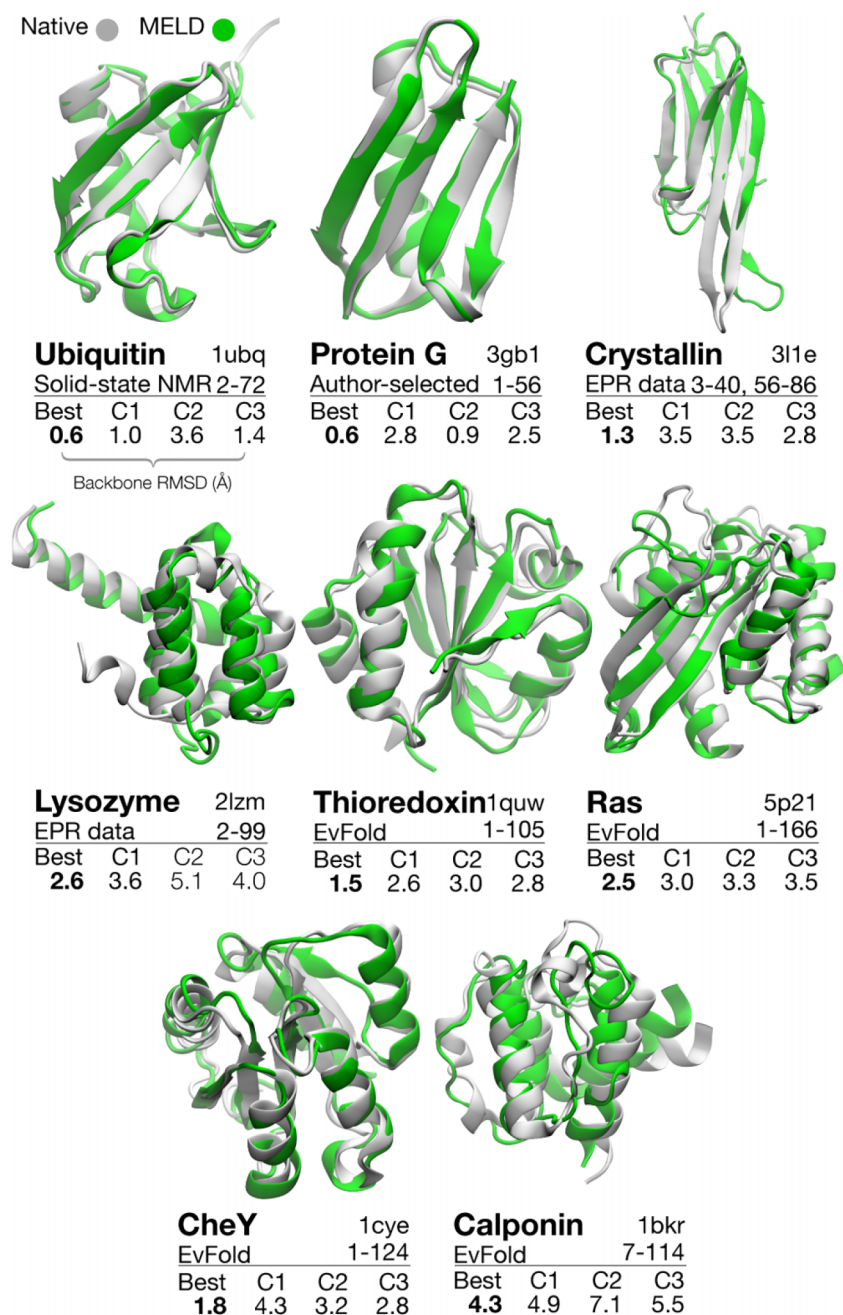


FIG. 3. MELD is useful for refining protein structures under different types of problematic information: too sparse, too ambiguous, or too uncertain. The following sources of data were used: solid state NMR for ubiquitin,<sup>14</sup> EPR data for Crystallin<sup>15</sup> and Lysozyme,<sup>45</sup> EvFold<sup>18</sup> for Thioredoxin, Ras, CheY, and Calponin, and a user-defined minimum set for protein G. Best refers to the lowest RMSD to native of the selected number of residues (below the PDB name) in the whole ensemble. C1, C2, and C3 are the centroids of the three most populated clusters in the ensemble.<sup>46</sup> Adapted with permission from J. L. MacCallum, A. Perez, and K. Dill, Proc. Natl. Acad. Sci. U. S. A. **112**, 6985 (2015). Copyright 2015 National Academy of Sciences, USA.

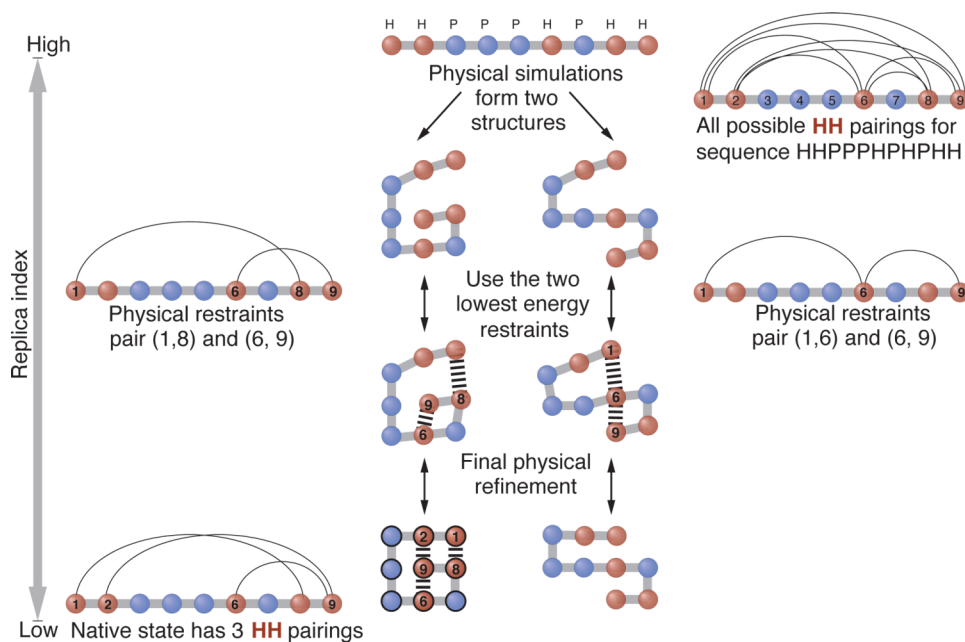


FIG. 4. The MELD method applied with Coarse Physical Insights (CPI) to find native structures. In this toy illustration, the method instructs the simulation to seek two hydrophobic contacts, within the combinatoric problem that 7 hydrophobic contacts are possible. Adapted with permission from A. Perez, J. L. MacCallum, and K. Dill, Proc. Natl. Acad. Sci. U. S. A. **112**, 11846 (2015). Copyright 2015 National Academy of Sciences, USA.

computations—with a few exceptions.<sup>5,6</sup> Imagine, instead, instructing a REMD simulation to find the native structure by finding the state of lowest free energy in an atomistic simulation at the same time as “finding a good hydrophobic core” or “finding good secondary structures.” The aim is to accelerate the discovery of the native structure, while preserving free energies. We refer to such heuristics as CPIs. MELD is a useful tool for such problems; see Fig. 4.

Figure 5 shows the predictions of MELD of the folded states of 20 small proteins,<sup>40</sup> based on 4 CPIs: (1) secondary structure web server predictions, (2) globular proteins have hydrophobic cores, (3) beta-strands pair up, and (4) globular proteins are compact. We run REMD for 500 ns with these heuristic constraints. We can measure performance in various ways. First, we ask how well MELD + forcefield gives the native structures. In 11 out of 20 cases, MELD identifies the

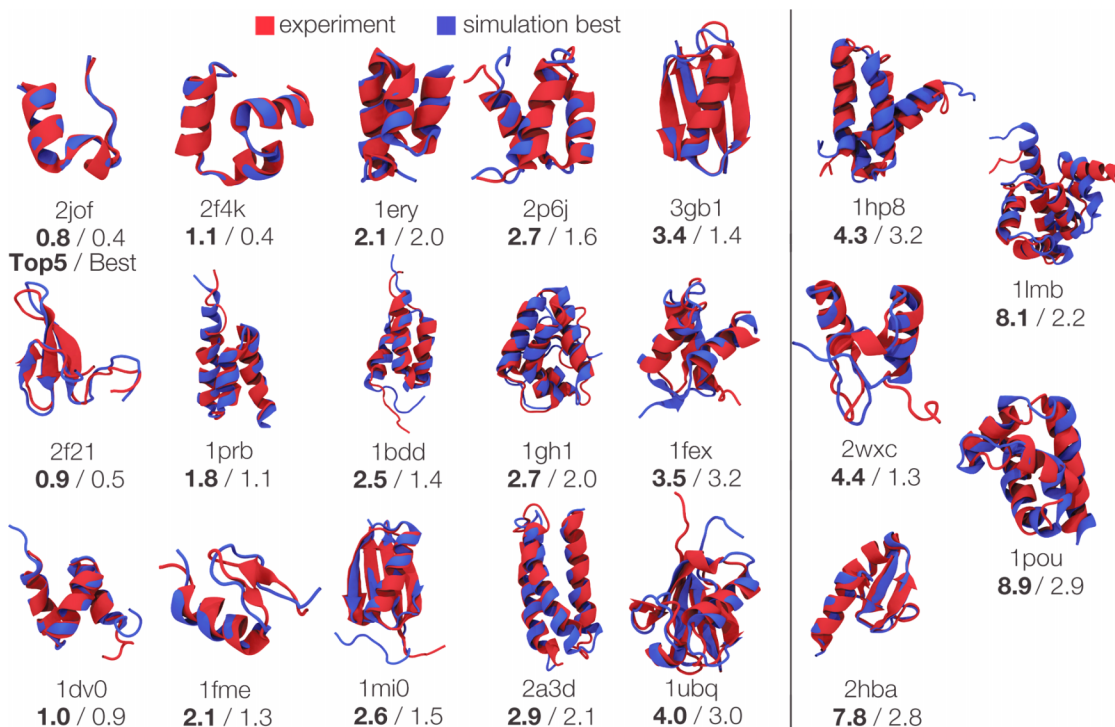


FIG. 5. MELD+CPI makes excellent predictions of the structures of 20 small proteins, starting from the fully extended state. It combines the CPI of a hydrophobic core, good secondary structures, and a compact state with REMD simulations in implicit water. All simulations were run for 500 ns, allowing to sample native states (RMSD of best structure in the ensemble below 4 Å, see non-bold numbers in the figure) in all cases. When clustering<sup>46</sup> and using the centroid of the top 5 clusters by population, the native state can be identified in 15/20 cases (left of vertical line). Adapted with permission from A. Perez, J. L. MacCallum, and K. Dill, Proc. Natl. Acad. Sci. U. S. A. **112**, 11846 (2015). Copyright 2015 National Academy of Sciences, USA.

best (highest population) structure; in 15 out of 20, the native structure is one of the top five. Some of the wrong predictions are forcefield errors. More importantly for a sampling method like MELD, we ask how well MELD samples the correct native state. We find that it samples all 20 native states well.

### MELD SCALES BETTER THAN BRUTE-FORCE MD

Figure 6 compares the performance of MELD + CPI for folding proteins, relative to brute-force MD. Two things matter for the performance of a sampling method: how closely the sampling focuses on important structures and how fast it reaches them. Both can be captured by using the performance metric,  $P = f_{\text{folded}}/t$ , where  $f_{\text{folded}}$  is the fraction of structures closer than 4 Å RMSD from the native structure, and  $t$  is the total simulation time (including all replicas). Fig. 6(a) compares the performance better performance of MELD, due to the effective harnessing of the coarse physical insights, compared to the brute-force folding simulations of Simmerling *et al.*<sup>6</sup>

Fig. 6(a) compares the scaling of  $P$  to longer chains of MELD + CPI vs. brute-force MD. Fig. 6(b) shows how the computer time needed to find native structures depends on protein chain length, for a given convergence radius. It indicates that to fold a 200-mer protein, MELD would require millisecond simulations with the current CPIs, which is nearly achievable with today's computational resources. In contrast,

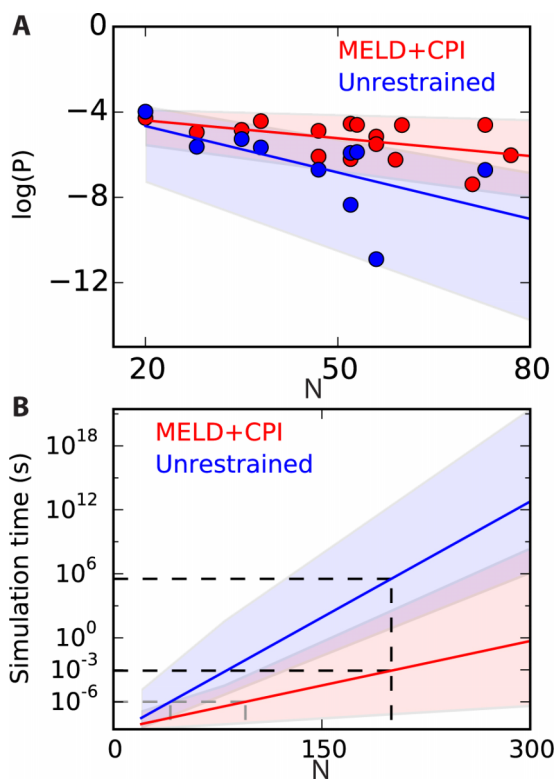


FIG. 6. The expected scaling of MELD for protein folding vs. brute-force MD. (a) Fit to data from MELD or unrestrained simulations and (b) extrapolation to larger proteins. Adapted with permission from A. Perez, J. L. MacCallum, and K. Dill, Proc. Natl. Acad. Sci. U. S. A. **112**, 11846 (2015). Copyright 2015 National Academy of Sciences, USA.

it indicates that brute-force MD would require about  $10^9$  greater computational resources, given current methods and forcefields.

### CCR COMPUTES FREE ENERGIES OF LARGE CONFORMATIONAL CHANGES

Another problem of interest is to calculate the free energy of a large change, say from conformation  $A$  to  $B$ , of a protein. A method called CCR<sup>36,38</sup> is useful. Fig. 7 shows its thermodynamic cycle. This method entails selectively tightening and weakening constraints: (1) Take the ensemble of conformation  $A$ . Tighten the constraints until you have squeezed  $A$  into  $A'$ , a single microstate that is the average structure of  $A$ . Compute the free energy of that step. (2) Perform the same process on  $B$ , converting it from an ensemble  $B$  to an average microstate  $B'$ . Compute the free energy. (3) Use normal-mode analysis and single point energies to compute microstate  $A'$  to microstate  $B'$ . Since this process is microstate-to-microstate, there is no ensemble entropy, so the free energy of this process approximately equals the normal-mode energy difference. Now, compute the full free energy change,  $\Delta G_{AB}$  using the thermodynamic cycle,

$$\begin{aligned} \Delta G_{AB} &= \Delta G_{\text{Confine}} + \Delta G_{\text{Convert}} + \Delta G_{\text{release}} \\ &= \Delta G_{AA'} + \Delta G_{A'B'} - \Delta G_{BB'}. \end{aligned} \quad (5)$$

The confinement and release steps are based on using thermodynamic integration over different spring strengths. Similar approaches have been used for ligand binding.<sup>51,52</sup>

Figure 8 shows one useful application of the CCR methodology. In the community-wide blind protein-structure prediction event called CASP (Critical Assessment of Structure Prediction),<sup>56,57</sup> different research groups use different models to try to predict the native structure of a protein. Each group is allowed to submit the 5 predicted structures it believes might best model the unknown protein structure. While sometimes, a good “needle” is found in this

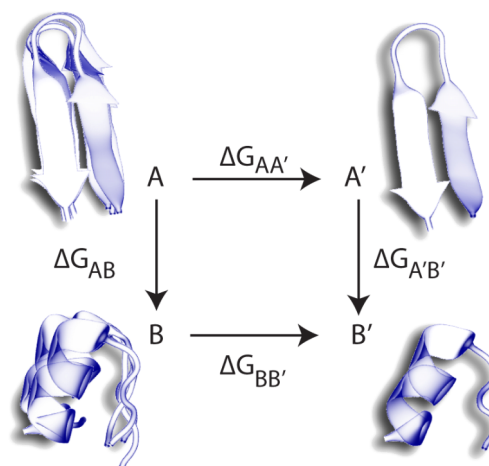


FIG. 7. The confinement convert and release thermodynamic cycle. An ensemble ( $A$ ) is confined to a single microstate ( $A'$ ). The free energy of conversion into  $B'$  is calculated based on molecular mechanics and entropic contributions from normal modes. Finally, the microstate  $B'$  is released to the ensemble of conformations  $B$ .

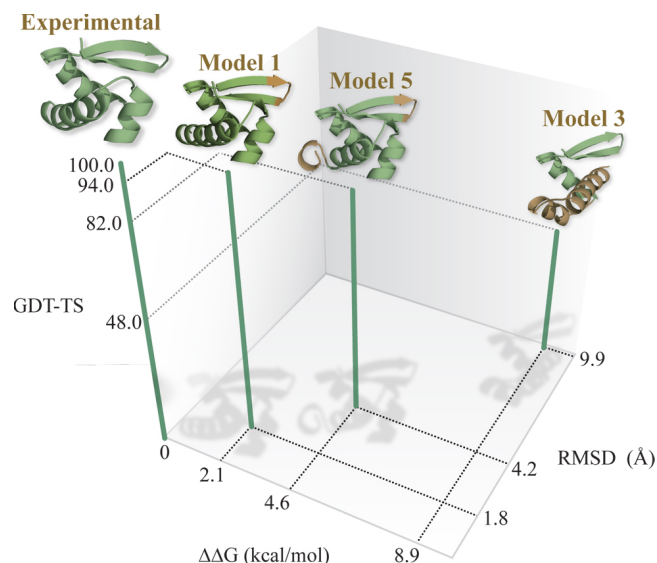


FIG. 8. CCR can order structures according to free energy. Research groups in CASP are allowed to submit 5 models for what they believe is the native structure. A common problem is that groups seldom know which of their 5 models is the best one. Here, we use CCR to compare 5 models from Rosetta server of a former CASP target. This image shows that CCR correctly picks out the best model.<sup>36</sup> CCR is also useful for other purposes, such as discriminating chameleon proteins.<sup>53–55</sup> Reprinted with permission from Roy *et al.*, *Structure* **22**, 168–175 (2013). Copyright 2013 Elsevier.

“haystack” of 5 possible options, very few research groups are typically able to know which of their 5 submissions is the closest to the true native structure.<sup>58</sup> Researchers can rarely rank order their predictions because their methods do not predict Boltzmann populations from a consistent energy model. Figure 8 shows that CCR can help. We rank-ordered predictions from different research groups using CCR-calculated free energies. In most cases, CCR correctly rank-orders by free energies the structures that are geometrically closer to the native state.<sup>36</sup>

## LOOPS CAN BE MODELED USING THE AIKC METHOD OF KINEMATIC GEOMETRY

Computer modeling of loops and flexible regions in proteins has been challenging because of the large conformational spaces that must be sampled. However, when constraints are known, such as the fixed endpoints of mobile regions, algebraic methods can provide highly effective tools for searching the important parts of conformational space.

Recent successes in loop modeling are based on Inverse Kinematics (IK) methods of robotics, which were developed for the motions of systems having some fixed (rigid) parts connected through flexible joints. An example is when a robot moves a hand through multiple arm-like linkages. Inverse Kinematic Closure (IKC) in engineering and biochemistry deals with finding all the possible conformations of the arm that are compatible with having the end points at fixed positions and orientations.<sup>33,34,59–70</sup>

Kinematic geometry methods differ from more general distance geometry methods in that the former focus on torsional degrees of freedom, rather than on distances. A

distinguishing feature of fully algebraic kinematic geometry methods is that by focusing on allowable motions through explicitly and rigorously accounting for the constraints imposed by ring closure, they produce descriptions having much lower dimensionalities. So, they can be fast and efficient. For example, conformational sampling for cyclo-octane can be reduced to a two-dimensional search.<sup>71,72</sup>

Computing protein loop ensembles is challenging because the energy surface is quite flat and the conformational spaces to be sampled are large. IK is well suited because the ends of protein loops are fixed in space (they are attached to the rest of the protein, whose structure is known) and we want to know the conformations of the loop amino acids. Previous efforts solve iteratively,<sup>63,64,67,68</sup> based on methods that seek a common zero of a set of objective functions, typically end-triad distances. As these objectives are non-convex transcendental functions of the loop torsions, it becomes very difficult to design iterative methods capable of finding all the zeros.

The advantage of AIKC methods<sup>33,60,65,66,73</sup> is that the objective has a simple analytical expression: it is a 16th degree polynomial in the half-tangent of one of the loop torsions. That can be formulated as a generalized eigenproblem of order 16 or as its characteristic polynomial whose real roots must be determined. Although these must be solved numerically, that can be done by efficient and robust algorithms. This

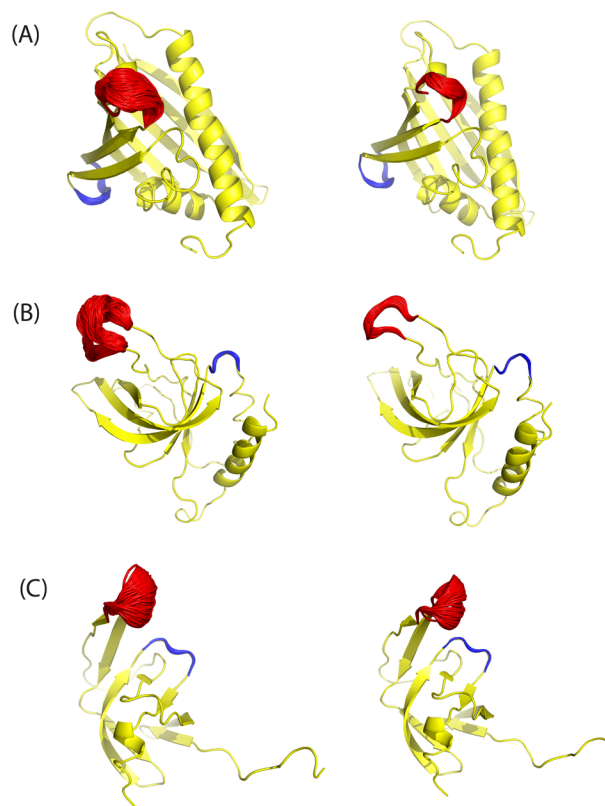


FIG. 9. Ensembles of loop structures using AIKC. Equilibrium simulations using MC sampling for proteins with PDB ID: (a) 1H2O, (b) 1XWE, and (c) 1Q9P sampled at  $T = 600$  K (left) and  $T = 300$  K (right). The sampled flexible loops, which have large fluctuation in the NMR models, are shown in red and the rigid loops with very small fluctuations are in blue. The structures in yellow are taken from MODEL 1 of the PDB file. Reprinted with permission from Nilmeier *et al.*, *J. Chem. Theory Comput.* **7**, 1564–1574 (2011). Copyright 2011 American Chemical Society.



algorithm has been incorporated into the Rosetta suite and the BRIKARD package.

Here is the basic mathematical framework for the AIKC algorithm. At the heart is the problem in inverse kinematics known as a 6R-6bar linkage. Consider a closed kinematic chain of  $N$  rotor links,  $\mathbf{b}_i$ ,  $i = 1, \dots, N$ . We fix the lengths  $b_j$  and pair-wise angles  $\theta_i$  of the links. We let  $R_i$  be the position vector of the joint (“atom”) between links  $\mathbf{b}_{i-1}$  and  $\mathbf{b}_i$  with respect to some fixed reference frame and  $\Gamma_i$  the unit vector along  $\mathbf{b}_i$ . Consider now a concerted change of all the torsions between the endpoints of a loop  $t_i \rightarrow t_i + dt_i$  that keeps the ends of the chain invariant (i.e., with respect to which the ends of the chain maintain geometrically correct attachments to rigid molecular frames). If we consider a frame attached rigidly to the first three atoms, then the effect of a change  $dt_i$  to the torsion  $t_i$  about axis  $\Gamma_i$  is to rotate all subsequent atoms at locations  $R_j$ ,  $j \geq i + 2$  by  $dR_j = \Gamma_i \times (R_j - R_i)$ . Then, at any point in space  $R$  located past the end of the chain the net effect of the concerted move cancels, and we have that

$$0 = dR = \sum_{i=1}^N \Gamma_i \times (R - R_i) dt_i$$

$$\Rightarrow \left( \sum_{i=1}^N \Gamma_i dt_i \right) \times R - \left( \sum_{i=1}^N \Gamma_i \times R_i dt_i \right) = 0.$$

Since this must be true for arbitrary  $R$ , both expressions in parentheses above must vanish independently. Basic analysis (the implicit function theorem) guarantees that six of the

variables may be expressed as differentiable functions of the remaining ones provided the  $6 \times N$  matrix  $S$  is of full rank, i.e., if  $S$  has six independent columns.

Let their indices be  $i_k$ ,  $k = 1, 2, \dots, 6$ ; the corresponding variables are called the pivots, and the remaining variables are called the drivers (indexed by  $j_k$ ,  $k = 1, \dots, N - 6$ ). Introduce  $p_k := t_{i_k}$ ,  $k = 1, \dots, 6$  for the pivots and  $q_k := t_{j_k}$ ,  $k = 1, \dots, N - 6$  for the drivers. Then, the changes in the pivots are given in terms of the changes in the drivers by

$$J \begin{pmatrix} dp_1 \\ \vdots \\ dp_6 \end{pmatrix} = - \sum_{k=1}^{N-6} P_{j_k} dq_k =: Qdq,$$

$$J := (P_{i_1} P_{i_2} \dots P_{i_6}); P_i = \begin{pmatrix} \Gamma_i \\ \Gamma_i \times R_i \end{pmatrix}.$$

The columns of the  $6 \times 6$  Jacobian  $J$  are the Plücker coordinates<sup>74,75</sup> of the six pivot axes.

The closure conditions giving the pivots in terms of the drivers are polynomials in the sines and cosines of the pivots. We introduce  $u_k = \tan(p_k/2)$ ,  $k = 1, \dots, 6$ , for the half-tangents of the pivots. The drivers  $q_k$  are sampled and set to fixed values. Then, the 6R-6bar problem may be formulated in principle as a system of polynomial equations  $F_i(u; q; \alpha) = 0$ ,  $i = 1, \dots, 6$  where  $\alpha$  is the parameter vector of bond lengths and bond angles. Using the theory of resultants,<sup>33,65,73</sup> this system may be reduced to a single polynomial of degree 16 in one of the

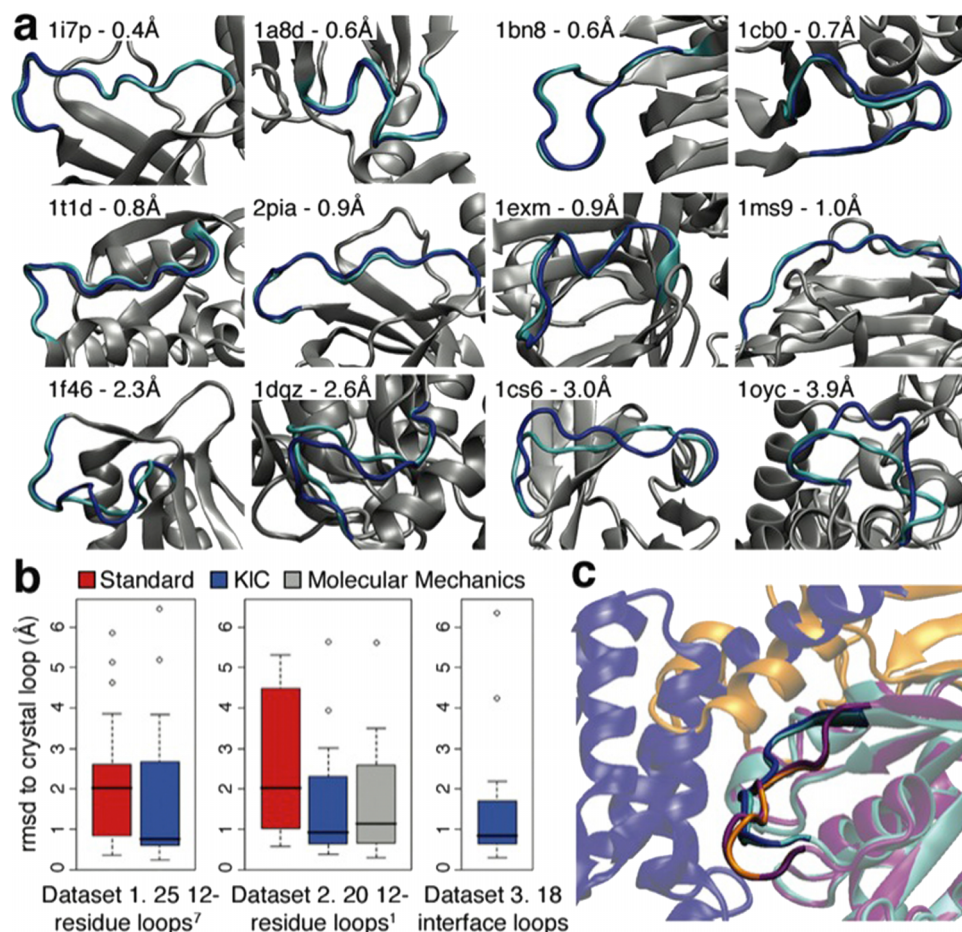


FIG. 10. Performance of KIC loop reconstruction protocol on three common datasets. (a) Representative set of 12-residue loop reconstructions (blue) (data set as in (b-middle)). PDB codes and r.m.s.d. to crystallographic loop (cyan) are shown. (b) Box-plot comparison of the standard Rosetta protocols (left), both Rosetta protocols with the molecular mechanics method (middle), and the KIC Rosetta protocol. (c) KIC reconstruction of conformational changes in the Rac switch I loop when bound to ExoS toxin. Reprinted with permission from D. J. Mandell, E. A. Coutsias, and T. Kortemme, Nat. Methods **6**, 551–552 (2009). Copyright 2009 Macmillan Publishers Ltd.

pivot variables. Multiple solutions for the pivots are possible for a given set of drivers; choosing values for the  $q_k$ , there are at most 16 real solution sets  $\{u_k^{(i)}\}_{k=1}^6$ ;  $i = 1, \dots, 16$ . The speed and robustness of modern polynomial real root finders enables the AIKC approach to explore conformational space in parallel. If the drivers are used as coordinates of the shape space, then there may be as many as 16 alternative branches, one for each set of pivots. Points on all alternative branches are generated simultaneously, in sharp contrast to trajectory-based searches that explore a single conformation at a time, and may not find all alternatives even after extensive simulation.

This equation may be written in the form

$$dp = -J^{-1}Qdq$$

giving the differential of the pivot torsions in terms of the differentials of the drivers and involving the inverse Jacobian. This expression may be used for planning concerted motions of the loop or ring. It may also be used to construct concerted Monte Carlo move sets that obey detailed balance,<sup>62,64,66,68,69</sup> Figure 9. Where the inverse Jacobian  $J^{-1}$  exists, we say that the IK problem is well posed. Where the determinant of the Jacobian vanishes, we have kinematic singularities. Move sets preserving closure for single or multiple, independent loops using more general Jacobians or other tangent space approaches including also angle and bond length perturbations are also possible.<sup>76–79</sup> However, such techniques have not been yet developed for multiply constrained systems, such as multicyclic peptides or other complex macrocycles.

## RESULTS FROM THE AIKC METHOD

Inverse kinematic closure was coupled with the structure prediction algorithm Rosetta<sup>34,80</sup> and was used via the Rosetta Kinematic Closure (KIC) protocol to predict the structures of a standard test set (Fig. 10). Although the method performed extremely well, resulting in sub-Angstrom accuracy in predicting the structures of the test set of 12-residue loops, as the loop length increases the complexity of the sampling problem grows both in dimensionality as well as in the topological complexity of the space. Several powerful algorithms have been proposed recently for reducing the size of the conformational space available to long protein loops. Typically, additional information on the torsional propensities has been incorporated and combined with either kinematics inspired or other direct methods that seek to build the loops one residue at a time while seeking to minimize a distance function<sup>81–83</sup> or Jacobian guided kinematic closure.<sup>84–86</sup> Manifold learning and dimensionality reduction<sup>72</sup> techniques can be helpful; however, such techniques must be extended to be able to handle the singularities that distinguish constrained molecular conformation spaces as algebraic varieties rather than smooth manifolds.<sup>71,72,87</sup>

## CONCLUSIONS

We have reviewed here some approaches to applying constraints within free-energy simulations of biomolecules. First, MELD is able to harness troublesome or heuristic external information to speed up free-energy simulations.

It does this by using different subsets of the external knowledge that are compatible with different stable states in conformational space. MELD is useful for determining native structures from experimental data, and when given heuristic constraints for finding native structures from unfolded states in REMD simulations. Second, CCR is a method for computing the free energy differences between conformations—even large conformational changes—in a biomolecule. And, third, we describe AIKC methods for fast and efficient sampling of the conformations of loops and flexible regions in biomolecules. There is considerable value in harnessing external information—in whatever form is available—within REMD simulations, because it can speed up and scale up free-energy simulations for targeted states and actions, ultimately for bigger biomolecules and bigger actions than can be presently simulated.

## ACKNOWLEDGMENTS

The authors appreciate the support of the Laufer Center, as well as from NIH Grant Nos. R01GM090205 and GM107104. J.L.M. is supported by the Canada Research Chairs program, the Natural Sciences and Engineering Research Council, and Compute Canada.

- <sup>1</sup>J. A. McCammon, B. R. Gelin, and M. Karplus, *Nature* **267**, 585 (1977).
- <sup>2</sup>M. Levitt and A. Warshel, *Nature* **253**, 694 (1975).
- <sup>3</sup>P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten, *Biophys. J.* **94**, L75 (2008).
- <sup>4</sup>Y. Duan and P. A. Kollman, *Science* **282**, 740 (1998).
- <sup>5</sup>K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, *Science* **334**, 517 (2011).
- <sup>6</sup>H. Nguyen, J. Maier, H. Huang, V. Perrone, and C. Simmerling, *J. Am. Chem. Soc.* **136**, 13959 (2014).
- <sup>7</sup>K. A. Beauchamp, R. McGibbon, Y.-S. Lin, and V. S. Pande, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17807 (2012).
- <sup>8</sup>S. Piana, J. L. Klepeis, and D. E. Shaw, *Curr. Opin. Struct. Biol.* **24**, 98 (2014).
- <sup>9</sup>W. Ma and K. Schulten, *J. Am. Chem. Soc.* **137**, 3031 (2015).
- <sup>10</sup>R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, *J. Chem. Theory Comput.* **8**, 3257 (2012).
- <sup>11</sup>J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, *J. Chem. Theory Comput.* **11**, 3696 (2015).
- <sup>12</sup>G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, *J. Phys. Chem. B* **105**, 6474 (2001).
- <sup>13</sup>H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res.* **28**, 235 (2000).
- <sup>14</sup>M. Huber, S. Hiller, P. Schanda, M. Ernst, A. Böckmann, R. Verel, and B. H. Meier, *ChemPhysChem* **12**, 915 (2011).
- <sup>15</sup>N. Alexander, M. Bortolus, A. Al-Mestarihi, H. Mchaourab, and J. Meiler, *Structure* **16**, 181 (2008).
- <sup>16</sup>S. J. Hirst, N. Alexander, H. S. McHaourab, and J. Meiler, *J. Struct. Biol.* **173**, 506 (2011).
- <sup>17</sup>The UniProt Consortium, *Nucleic Acids Res.* **43**, D1049 (2014).
- <sup>18</sup>D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, *PLoS One* **6**, e28766 (2011).
- <sup>19</sup>D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, *Bioinformatics* **28**, 184 (2012).
- <sup>20</sup>A. Sali and T. L. Blundell, *J. Mol. Biol.* **234**, 779 (1993).
- <sup>21</sup>D. Baker and A. Sali, *Science* **294**, 93 (2001).
- <sup>22</sup>C. Dominguez, R. Boelens, and A. M. J. J. Bonvin, *J. Am. Chem. Soc.* **125**, 1731 (2003).
- <sup>23</sup>W. Li, Y. Zhang, and J. Skolnick, *Biophys. J.* **87**, 1241 (2004).
- <sup>24</sup>N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
- <sup>25</sup>S. J. Marrink, A. H. de Vries, and A. E. Mark, *J. Phys. Chem. B* **108**, 750 (2004).
- <sup>26</sup>S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, *J. Phys. Chem. B* **111**, 7812 (2007).

- <sup>27</sup>G. M. Torrie and J. P. Valleau, *J. Comput. Phys.* **23**, 187 (1977).
- <sup>28</sup>S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *J. Comput. Chem.* **13**, 1011 (1992).
- <sup>29</sup>A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12562 (2002).
- <sup>30</sup>A. Barducci, G. Bussi, and M. Parrinello, *Phys. Rev. Lett.* **100**, 020603 (2008).
- <sup>31</sup>G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, *J. Am. Chem. Soc.* **128**, 13435 (2006).
- <sup>32</sup>S. Piana and A. Laio, *J. Phys. Chem. B* **111**, 4553 (2007).
- <sup>33</sup>E. A. Coutsias, C. Seok, M. P. Jacobson, and K. Dill, *J. Comput. Chem.* **25**, 510 (2004).
- <sup>34</sup>D. J. Mandell, E. A. Coutsias, and T. Kortemme, *Nat. Methods* **6**, 551 (2009).
- <sup>35</sup>J. L. MacCallum, A. Perez, and K. Dill, *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6985 (2015).
- <sup>36</sup>A. Roy, A. Perez, K. Dill, and J. L. MacCallum, *Structure* **22**, 168 (2013).
- <sup>37</sup>M. Spichty, M. Cecchini, and M. Karplus, *J. Phys. Chem. Lett.* **1**, 1922 (2010).
- <sup>38</sup>M. Cecchini, S. V. Krivov, M. Spichty, and M. Karplus, *J. Phys. Chem. B* **113**, 9728 (2009).
- <sup>39</sup>S. Pressé, K. Ghosh, J. Lee, and K. Dill, *Rev. Mod. Phys.* **85**, 1115 (2013).
- <sup>40</sup>A. Perez, J. L. MacCallum, and K. Dill, *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11846 (2015).
- <sup>41</sup>D. T. Jones, *J. Mol. Biol.* **292**, 195–202 (1999).
- <sup>42</sup>G. Pollastri and A. McLysaght, *Bioinformatics* **21**, 1719 (2005).
- <sup>43</sup>S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum, *Science* **349**, 273 (2015).
- <sup>44</sup>S. Gelly, L. Kocsis, M. Schoenauer, M. Sebag, D. Silver, C. Szepesvári, and O. Teytaud, *Commun. ACM* **55**, 106 (2012).
- <sup>45</sup>S. M. Islam, R. A. Stein, H. S. McHaourab, and B. Roux, *J. Phys. Chem. B* **117**, 4740 (2013).
- <sup>46</sup>X. Daura, K. Gademann, and B. Jaun, *Angew. Chem., Int. Ed.* **38**, 236 (1999).
- <sup>47</sup>B. Roux and S. M. Islam, *J. Phys. Chem. B* **117**, 4733 (2013).
- <sup>48</sup>D. Schneidman-Duhovny, S. J. Kim, and A. Sali, *BMC Struct. Biol.* **12**, 17 (2012).
- <sup>49</sup>S. Lindert, T. Hofmann, N. Wötzel, M. Karakaş, P. L. Stewart, and J. Meiler, *Biopolymers* **97**, 669 (2012).
- <sup>50</sup>V. Lučić, F. Förster, and W. Baumeister, *Annu. Rev. Biochem.* **74**, 833 (2005).
- <sup>51</sup>D. L. Mobley, J. D. Chodera, and K. Dill, *J. Chem. Theory Comput.* **3**, 1231 (2007).
- <sup>52</sup>J. Wang, Y. Deng, and B. Roux, *Biophys. J.* **91**, 2798 (2006).
- <sup>53</sup>Y. He, Y. Chen, P. Alexander, P. N. Bryan, and J. Orban, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 14412 (2008).
- <sup>54</sup>P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21149 (2009).
- <sup>55</sup>P. N. Bryan and J. Orban, *Curr. Opin. Struct. Biol.* **20**, 482 (2010).
- <sup>56</sup>J. Moulton, J. T. Pedersen, R. Judson, and K. Fidelis, *Proteins* **23**, 2 (1995).
- <sup>57</sup>J. Moulton, *Curr. Opin. Struct. Biol.* **15**, 285 (2005).
- <sup>58</sup>J. L. MacCallum, A. Perez, M. J. Schnieders, L. Hua, M. P. Jacobson, and K. Dill, *Proteins* **79**(S10), 74 (2011).
- <sup>59</sup>M. L. Husty, M. Pfurner, and H.-P. Schröcker, *Mech. Mach. Theory* **42**, 66 (2007).
- <sup>60</sup>H.-Y. Lee and C.-G. Liang, *Mech. Mach. Theory* **23**, 209 (1988).
- <sup>61</sup>D. Manocha and J. F. Canny, *IEEE Trans. Rob. Autom.* **10**, 648 (1994).
- <sup>62</sup>J. Nilmeier, L. Hua, E. A. Coutsias, and M. P. Jacobson, *J. Chem. Theory Comput.* **7**, 1564 (2011).
- <sup>63</sup>N. Gö and H. A. Scheraga, *Macromolecules* **3**, 188 (1970).
- <sup>64</sup>A. R. Dinner, *J. Comput. Chem.* **21**, 1132 (2000).
- <sup>65</sup>W. J. Wedemeyer and H. A. Scheraga, *J. Comput. Chem.* **20**, 819 (1999).
- <sup>66</sup>M. Wu and M. W. Deem, *J. Chem. Phys.* **111**, 6625 (1999).
- <sup>67</sup>A. A. Canutescu and R. L. Dunbrack, *Protein Sci.* **12**, 963 (2003).
- <sup>68</sup>L. R. Dodd, T. D. Boone, and D. N. Theodorou, *Mol. Phys.* **78**, 961 (2006).
- <sup>69</sup>S. Cahill, M. Cahill, and K. Cahill, *J. Comput. Chem.* **24**, 1364 (2003).
- <sup>70</sup>J. Cortés, T. Siméon, M. Remaud-Siméon, and V. Tran, *J. Comput. Chem.* **25**, 956 (2004).
- <sup>71</sup>S. Martin, A. Thompson, E. A. Coutsias, and J.-P. Watson, *J. Chem. Phys.* **132**, 234115 (2010).
- <sup>72</sup>W. M. Brown, S. Martin, S. N. Pollock, E. A. Coutsias, and J.-P. Watson, *J. Chem. Phys.* **129**, 064118 (2008).
- <sup>73</sup>E. A. Coutsias, C. Seok, M. J. Wester, and K. Dill, *Int. J. Quantum Chem.* **106**, 176 (2006).
- <sup>74</sup>K. H. Hunt, *Kinematic Geometry of Mechanisms* (SIAM Review, 1991), pp. 678–679.
- <sup>75</sup>A. D. Viquerat, T. Hutt, and S. D. Guest, *Mech. Mach. Theory* **63**, 73 (2013).
- <sup>76</sup>J. P. Ulmschneider and W. L. Jorgensen, *J. Chem. Phys.* **118**, 4261 (2003).
- <sup>77</sup>S. Hayward and A. Kitao, *J. Chem. Theory Comput.* **11**, 3895 (2015).
- <sup>78</sup>S. Zamuner, A. Rodriguez, F. Seno, and A. Trovato, *PLoS One* **10**, e0118342 (2015).
- <sup>79</sup>S. Bottaro, W. Boomsma, K. E. Johansson, C. Andreetta, T. Hamelryck, and J. Ferkinghoff-Borg, *J. Chem. Theory Comput.* **8**, 695 (2012).
- <sup>80</sup>A. Stein and T. Kortemme, *PLoS One* **8**, e63090 (2013).
- <sup>81</sup>M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw, and R. A. Friesner, *Proteins* **55**, 351 (2004).
- <sup>82</sup>B. D. Sellers, K. Zhu, S. Zhao, R. A. Friesner, and M. P. Jacobson, *Proteins* **72**, 959 (2008).
- <sup>83</sup>K. Tang, J. Zhang, and J. Liang, *PLoS Comput. Biol.* **10**, e1003539 (2014).
- <sup>84</sup>J. Ko, D. Lee, H. Park, E. A. Coutsias, J. Lee, and C. Seok, *Nucleic Acids Res.* **39**, W210 (2011).
- <sup>85</sup>H. Park, G. R. Lee, L. Heo, and C. Seok, *PLoS One* **9**, e113811 (2014).
- <sup>86</sup>J. Lee, D. Lee, H. Park, E. A. Coutsias, and C. Seok, *Proteins* **78**, 3428 (2010).
- <sup>87</sup>J. M. Porta and L. Jaillet, *J. Comput. Chem.* **34**, 234 (2013).