The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1628–1654
https://doi.org/10.1214/17-AOAS1124

EXPLORING THE CONFORMATIONAL SPACE FOR PROTEIN FOLDING WITH SEQUENTIAL MONTE CARLO

BY SAMUEL W. K. WONG*, JUN S. LIU[†,1] AND S. C. KOU[†,2]

*University of Florida** *and Harvard University*[†]

Computational methods for protein structure prediction from amino acid sequence are of vital importance in modern applications, for example protein design in biomedicine. Efficient sampling of conformations according to a given energy function remains a bottleneck, yet is a vital step for energy-based structure prediction methods. While the Protein Data Bank of experimentally determined 3-D protein structures has steadily increased in size, structure predictions for new proteins tend to be unreliable in the amino acid segments where there is low sequence similarity with known structures. In this paper we introduce a new method for building such segments of protein structures, inspired by sequential Monte Carlo methods. We apply our method to examples of real 3-D structure predictions and demonstrate its promise for improving low confidence segments. We also provide applications to the prediction of reconstructed segments in known structures, and to the assessment of energy function accuracy. We find that our method is able to produce conformations that have both low energies and good coverage of the conformational space and hence can be a useful tool for protein design and structure prediction.

1. Introduction. In his seminal work in the 1970s, the Nobel laureate Christian B. Anfinsen (1973) proposed that the stable 3-D structure of a protein is essentially determined by its amino acid sequence. Since then, the question of how a protein, consisting of a linear sequence of amino acids, acquires that stable 3-D structure has come to be known as the *protein folding problem*, and has challenged scientists of many disciplines for about a half-century; see, for example, Dill and MacCallum (2012) for a brief review. The traditional way to determine a protein's 3-D structure is by laboratory work, such as crystallography. The first known example of structure determination of a protein molecule by X-ray was done by Kendrew et al. (1958). Even today, there are substantial costs and difficulties associated with these laboratory techniques; hence while improvements in genome sequencing technologies have enabled the number of known protein sequences to expand rapidly, fewer than 1% of these sequences have a known 3-D structure from laboratory work [Lee, Redfern and Orengo (2007)].

Received May 2017; revised November 2017.

[1]Supported in part by NSF Grant DMS-1613035 and NIH R01 GM113242-01.

[2]Supported in part by NSF Grant DMS-1510446.

Key words and phrases. Protein structure prediction, particle filter, structure refinement, energy optimization.

Computational approaches for protein structure prediction from amino acid sequence have been developed since the advent of computers, yet many unsolved challenges remain [Friesner, Prigogine and Rice (2002)]. To encourage the scientific community to test the capabilities of current algorithms and methods, a set of blinded protein structure prediction experiments has been organized every two years since 1994, known as the Critical Assessment of protein Structure Prediction (CASP).[3] Participants do not know the true structures when they submit their predictions. The true structures for proteins used in CASP will have been recently determined in a laboratory but are not publicly released until the CASP experiment has concluded. The CASP experiments have documented substantial progress in prediction accuracy over the past two decades, and currently the most successful structure prediction algorithms operate with the assumption that similarities in sequence often correspond to similarities in 3-D structure [Krissinel (2007)]. However, many unsolved challenges remain, in particular when a new sequence has few similarities with the sequences of known structures [Moult et al. (2016)]. In addition, there is increasing interest in designing amino acid sequences to achieve specific 3-D structure and function, for example, in drug discovery [Khoury et al. (2014)]. As a result, further advances in the efficiency and accuracy of computational structure prediction algorithms are in critical need.

1.1. *The sequence–to–structure correspondence.* The key principle underlying protein structure prediction is that a given amino acid sequence generally has a unique 3-D structure. To illustrate, in the left panel of Figure 1 we show the length 223 amino acid sequence for 5JMU:A, a protein in Eubacterium rectale involved in carbohydrate metabolism [Tan et al. (2016)]. Each letter denotes one of

```
SVYDPAATADTVNPGNKIIYLTFDDGPGKYTQGLLDVL
YNVKATFFVTNTHPDYQNMIAEEAKRGHTVAIHSASHK
QIYTSEQAFFDDLEQMNSIIKAQTGNDASIIRFPGGSS
VSKDYCPGIMTQLVNDVTARGLLYCDWNVSSGDANPKP
TEQVVQNVISGVQSHNVSVVLQHDIKEFSVNAVEQIIQ
QANGYTFLPLTTSSPMSHHRVNN
```

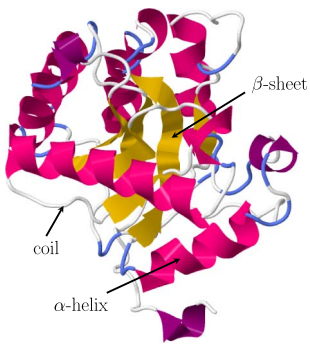

FIG. 1. *Amino acid sequence* (*left*) *and* 3*-D structure* (*right*) *as determined by X-ray crystallography of the catalytic domain of peptidoglycan N-acetylglucosamine deacetylase from Eubacterium rectale* (*Protein Data Bank ID*: 5*JU*:*A*).

[3] http://predictioncenter.org.

the 20 different amino acid types that are the building blocks of proteins. That sequence folds into the stable 3-D structure shown in the right panel, which has been determined by X-ray crystallography and was released to the public on June 29, 2016. In the figure we have drawn arrows pointing to examples of the two main types of secondary structure that occur over segments of amino acids, called the α-helix and β-sheet, that have regular angular patterns. Segments without regular secondary structure are known as coils, and an example is indicated in the figure as well.

1.2. *The energy landscape*. A *conformation* refers to a specific arrangement of the atoms of a protein in 3-D space. A powerful approach for structure prediction is to assume that the true (or *native*) conformation of a protein is the one with minimum potential energy. This is based on the energy landscape theory [Onuchic, Luthey-Schulten and Wolynes (1997)], and accordingly, many computational methods make their predictions according to an energy function. The structure with the lowest energy value found is selected as the prediction [e.g., Soto et al. (2008), Cooper et al. (2010), Tang, Zhang and Liang (2014), Liang, Zhang and Standley (2011)]. There are two main interconnected challenges associated with this approach—(i) the energy function and (ii) the search method. For the first challenge, an ideal energy function for guiding the search should assign the lowest energy to the conformation closest to the truth. However, we do not know nature's "true" energy function, so it is necessary to develop models for energy that aim to approximate it and provide good guidance for the search.[4] These models can be physics based or fitted from data; see Lazaridis and Karplus (2000) for an overview. The second challenge, and the focus of this paper, is to find the minimum energy conformation for a given energy function.

Many proteins of interest are composed of 100–600 amino acids, with corresponding geometric degrees of freedom numbering from several hundred to a few thousand. Thus given the large degrees of freedom, a deterministic energy minimization approach or a direct search for the minimum is impractical. As we next describe, a different type of approach that utilizes information from known 3-D structures, when available, has proven to be more effective for predicting the overall structures of new proteins.

1.3. *The protein data bank for building* 3*-D structure predictions*. Proteins with known 3-D structures are publicly available from the Protein Data Bank (PDB) [Bernstein et al. (1977)], which is a key source of data for training structure prediction algorithms. The PDB now contains over 110,000 structures such as the example 5JMU:A. The *homology modeling* approach was developed to leverage

[4]Such realistic energy functions will stabilize the energies of conformations in a neighborhood close to the truth. See Zhang et al. (2007) for additional discussion.

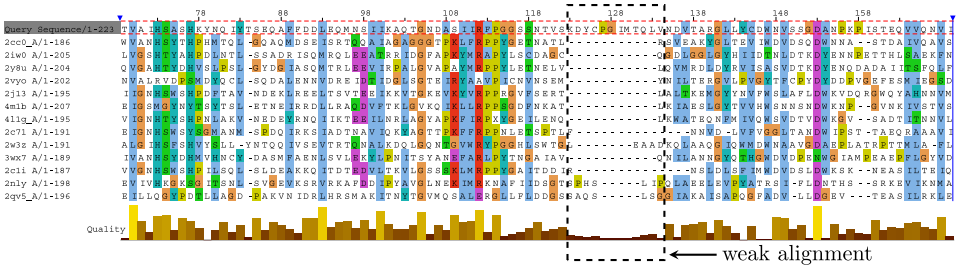

FIG. 2. *Sequence alignment for* 5*JMU:A with proteins that have known* 3*-D structures, excerpt shown for amino acid positions* 68–168. *Visualized using Jalview.*

these data and make 3-D structure predictions based on sequence alignments. The sequence of interest is aligned to sequences in the PDB to obtain structural templates, and these templates are then stitched together to build a structure prediction that satisfies geometric constraints, for example, using the MODELLER software [Fiser and Šali (2003)]. To illustrate, in Figure 2 we have shown a portion of the alignment for 5JMU:A using sequences from the PDB that were available in April, 2016. We will revisit this example in detail in the Applications section.

While homology modeling has become quite effective for overall structure predictions as the size of the PDB has grown, its accuracy tends to be lower in segments that have low sequence similarity with known structures. In Figure 2, the segment ∼120–135 of 5JMU:A has a noticeable lack of alignment. Therefore, it would be useful to consider optimizing such segments using an energy-guided search of the conformational space, after homology modeling has been performed.

1.4. *Searching for minimum energy conformations.* Even when the search for the minimum energy conformation is confined to a specified continuous segment of amino acids within a protein structure, the problem is still difficult due to the large size of the conformational space. This is especially the case when the length of the segment to optimize consists of ≥ 12 amino acids [Li et al. (2011)], which will have ≥ 40 geometric degrees of freedom. For any realistic energy function, the energy landscape of the conformational space will also be both multimodal and of high dimension [Brooks, Onuchic and Wales (2001)]. It would be very ineffective to perform energy minimization routines from arbitrary starting conformations, as such a procedure would in general yield only local minima far from the global one. Obtaining conformations with low energies from an initial search is therefore very important and is the primary motivation for the methodology we present in this paper. We shall formulate this problem as the stochastic optimization of a high-dimensional distribution subject to constraints.

Since a protein is composed of a sequence of amino acids, it has been quite natural to consider sampling and optimization methods that exploit this sequential character. Segments of proteins can be built one amino acid at a time, and the

successive steps can be designed to favor low-energy conformations. This type of method has been developed previously for both simplified protein models and real 3-D structures [Wick and Siepmann (2000), Zhang, Kou and Liu (2007), Wong, Cui and Chen (1998), Tang, Zhang and Liang (2014)].

1.5. *Our method*. Our goal is to efficiently explore the conformational space for a continuous segment of amino acids within a protein, to minimize its energy. Building on these previous developments, we propose a novel approach that combines sequential construction with inspiration from sequential Monte Carlo (SMC) methods. SMC (or particle) methods are a powerful tool that can be adapted for generating proposals from any high-dimensional distribution by breaking the draws into a sequence of intermediate distributions using propagation and reweighting steps [Liu and Chen (1998)]. When the reweighting step is designed appropriately, it enables the more promising members of a particle population to survive and continue exploration of the state space [Doucet, de Freitas and Gordon (2001), Liu (2001)].

In our context, the "particles" are partially constructed conformations for the amino acid segment of interest. To achieve the goal of stochastic energy optimization, the generating of substantially identical conformations is not useful, and hence it is especially important to avoid degeneracy in the particle population. We make substantive modifications to the basic sequential importance sampling scheme for this purpose. We shall obtain, after completion of the algorithm, a particle population of substantively distinct conformations for the segment that targets the high-density regions of the corresponding Boltzmann distribution for the provided energy function, that is, the energies of the conformations constructed will be low. This particle representation is an important feature that distinguishes our method from previous ones. Additionally, we exploit the conditional structure of the geometric degrees of freedom to achieve improved efficacy for exploring the large conformational space.

Our method is also fast, requiring a typical runtime of 10 minutes for a length 12 segment on a single 3.2 GHz CPU core; further, it is simple to parallelize the particle propagation over multiple CPU cores to achieve even faster runtimes. Many previous methods that are designed to optimize the energy of protein segments require substantially more computational time, from several hours [e.g., Liang, Zhang and Zhou (2014)] to hundreds of hours [e.g., Mandell, Coutsias and Kortemme (2009)].

In Section 2 we explain the basics of protein geometry and formulate the problem in statistical terms. Our method and its important features are presented in Section 3. In Section 4 we present results obtained by applying the method to three sets of applications—sampling segments to improve structure predictions built from homology, predicting reconstructed loop segments, and assessing different energy functions. We conclude the paper with a brief discussion in Section 5. Some technical details on implementation are provided in the Appendix.

2. Statistical formulation.

2.1. *Protein geometry.* A protein structure is represented by a list of 3-D Cartesian coordinates for the positions of each atom. An equivalent representation can be given in terms of geometric degrees of freedom, which are primarily dihedral angles; bond lengths and bond angles exhibit very little variation in real structures, and so we take them to be fixed at their ideal values [Engh and Huber (1991)]. The *backbone* of a protein consists of the interconnected sequence of N, C_α, C, and O atoms for each amino acid, as follows:

$$\cdots - \mathrm{N} - \mathrm{C}_\alpha - \mathrm{C}(=\mathrm{O}) - \mathrm{N} - \mathrm{C}_\alpha - \mathrm{C}(=\mathrm{O}) - \cdots .$$

Their positions can be parameterized according to the free dihedral angles (ϕ, ψ, ω). The angle ϕ governs the distance between the C atoms of successive amino acids; in a similar way, ψ governs the distance between successive N atoms, and ω governs the distance between successive C_α atoms. The *side chain* of an amino acid that extends from its C_α atom is unique for, and thus characterizes each of the 20 different amino acid types. Positions of side chain atoms for an amino acid can likewise be parameterized in terms of the free dihedral angles χ; depending on the amino acid type, there are 0 to 4 of these. As side chains extend from the protein backbone, side chain dihedral angles can be rotated while keeping the backbone fixed.

These geometric aspects are illustrated in Figure 3, where we have magnified the heavy (i.e., nonhydrogen) atoms that compose the Glysine and Lysine amino acids in positions 28–29 of the structure of 5JMU:A. Each atom is labelled with the position of its amino acid in the sequence (27–30) and atom name. Note the backbone connectivity: the previous 27.C connects to 28.N, and likewise 29.C connects to

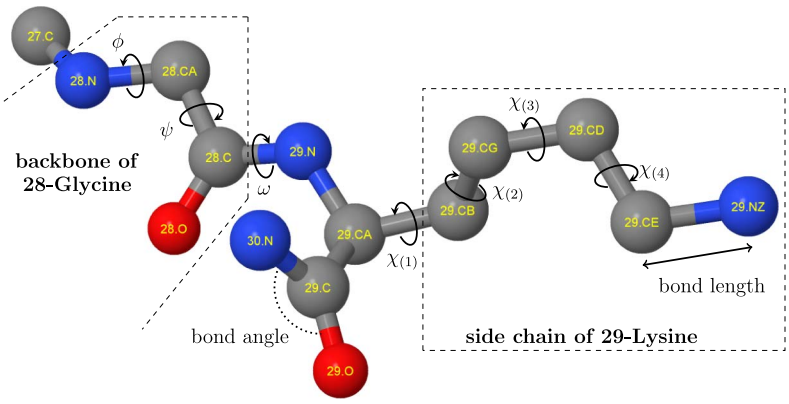

FIG. 3. *Illustration of protein geometry.*

following 30.N. The definitions of dihedral angles (ϕ, ψ, ω) for position 28 are labeled; for example, the labeled ϕ is the angle between 27.C and 28.C when viewing down the 28.N–28.CA axis. Glycine does not have a side chain, while Lysine has a long side chain with four χ dihedral angles, which we have labeled $\chi_{(1)}$ to $\chi_{(4)}$ in the diagram. An example of a bond length and a bond angle are also shown; it can be seen that the bond angle is the planar angle formed by three connected atoms.

2.2. *Representation of protein segments and energy functions.* Suppose we are given an initial 3-D structure (e.g., built from homology) for an amino acid sequence, and a continuous segment of length l amino acids within the sequence has been chosen for energy-guided optimization. Let $a_1, a_2, \ldots, a_l$ denote the sequence of amino acids of that segment with the rest of the structure held fixed. By convention we assume that the two ends of the segment are anchored by fixed positions of the C_α atom of a_1 and the C atom of a_l. Then specifying the positions of the atoms of segment between the anchors is equivalent to specifying values for the dihedral angles $(\phi_i, \psi_i, \omega_i, \chi_i)$ for $i = 1, \ldots, l-1$, and χ_l. Here, for convenience we let χ_i denote the length 0–4 vector of side chain dihedral angles of amino acid i (length depending on the amino acid type). Additionally, conformations for the segment must seamlessly connect to the two anchors with realistic bond lengths and angles; these constraints are described in Coutsias et al. (2004).

The empirical distributions of the angle pairs (ϕ_i, ψ_i) for each amino acid type have been studied extensively, and are commonly referred to in the chemistry literature as the *Ramachandran plot* [Ramachandran, Ramakrishnan and Saisekharan (1963)]. Further grouping the (ϕ_i, ψ_i) pairs according to secondary structure shows that the distributions for α-helices and β-sheets are tightly constrained due to their regular angular patterns. In contrast, an amino acid located in a coil will have a much wider range of possible dihedral angles; thus the structures of coil regions are the most difficult to predict. Note that ω is typically close to 180° and hence has less effect on the backbone shape compared to (ϕ, ψ). The probability distributions of side chain dihedrals have also been studied extensively, and these studies show that values of χ in real structures tend to cluster around a discrete set of modes, known as *rotamers*; these have been tabulated in the form of rotamer libraries for each amino acid type [e.g., Shapovalov and Dunbrack (2011)].

Let H denote the energy function with which conformations are to be evaluated. In this paper we consider H to be given, that is, we focus on searching the conformational space using H. Let x denote the vector of all free dihedral angles of the segment. Then we wish to find the global minimum of H by stochastically searching the conformational space guided by the Boltzmann distribution

$$\pi(x) \propto \exp\{-H(x)/T\},$$

where T is the effective temperature. Without loss of generality H can be scaled such that we can take $T = 1$. Most energy functions that have been used for computational protein folding can be expressed in the generic form $H(x) =$

$H_\theta(x) + H_d(\{r_{ab}; x\})$, where H_θ is the energy of the dihedral angles x, and H_d is the energy of all pairwise distances r_{ab} between atoms (i.e., atomic interactions), whose positions have been calculated based on x. Here H_d would account for the atomic interactions within the segment, as well as atomic interactions between the segment and the rest of the protein. It should be noted that H_d is much more costly to compute than H_θ in general. When a pair of atoms are too close in space than can be possible in real structures (i.e., r_{ab} is below a certain threshold, depending on atom types[5]), they are said to have a *steric clash*; for convenience we shall simply say $H_d = +\infty$ when there is at least one steric clash. Finally, since the dimension of x will generally be ≥ 20 for any realistic scenario where the method would be applied (i.e., $l \geq 8$), the sampling becomes a difficult task due to the large space and multimodality.

3. Method. We now introduce our sequential Monte Carlo approach to construct conformations that are located in the high-density regions of $\pi(x)$. The basic idea is to build x via a sequence of incremental distributions π_i. Accordingly, we shall use the notation $\Delta H(x|x')$ to denote the incremental energy contribution of the additional atoms corresponding to x when some previous angles x' have already been sampled. For protein segments, it is sensible to let the incremental distributions correspond to adding one amino acid at a time, so that the ith incremental conditional distribution is that of $(\phi_i, \psi_i, \omega_i, \chi_i)$, conditional on $(\phi_{1:i-1}, \psi_{1:i-1}, \omega_{1:i-1}, \chi_{1:i-1})$ when $i > 1$. For notational convenience we define $x_i \equiv (\phi_{1:i}, \psi_{1:i}, \omega_{1:i}, \chi_{1:i})$ which contains the backbone and side chain angles together, and $y_i \equiv (\phi_{1:i}, \psi_{1:i}, \omega_{1:i})$ which contains the backbone angles only.

There is some flexibility to the order in which amino acids are added; the requirement is that the next amino acid to add must be adjacent to an existing amino acid with given backbone atom positions (either fixed from outside the segment, or previously added within the segment). For example, suppose we are building the segment in positions 120–130. Then the first amino acid to be added is permitted to be either position 120 or 130. A left–to–right construction would set the sequence of angles $x_1, x_2, x_3, \ldots$ to correspond to positions $120, 121, 122, \ldots$. The scheme that we adopt is to alternately add amino acids to the left and right anchors, that is, setting $x_1, x_2, x_3, x_4, \ldots$ to correspond to positions $120, 130, 121, 129, \ldots$. The intuition behind our alternating order is that the amino acids closest to the anchors are the most geometrically constrained by the fixed portion of the structure, and thus from a sequential sampling perspective it can be more efficient to sample those ones first.

We begin by providing a brief outline how a basic sequential importance sampling (SIS) scheme could be applied here. The particle population of specified size

[5]We consider a pair of atoms to be in steric clash if the Lennard–Jones 12-6 model of their Van der Waals forces exceeds 10.0 kcal/mol.

N would be initialized by sampling the first amino acid $\{x_1^{(j)}\}_{j=1}^N$ from an importance distribution η_1 and defining weights $w_1^{(j)} \propto \pi_1(x_1^{(j)})/\eta_1(x_1^{(j)})$. Then in subsequent steps, a length i amino acid segment $X_i^{(j)}$ is obtained by propagating forward the particle $x_{i-1}^{(j)}$, that is, adding one amino acid according to a proposal distribution $q_i(x_i|x_{i-1}^{(j)})$. Then by defining $\eta_i(x_i^{(j)}) = \eta_{i-1}(x_{i-1}^{(j)})q_i(x_i|x_{i-1}^{(j)})$, the weights are updated according to $w_i^{(j)} \propto \pi_i(x_i^{(j)})/\eta_i(x_i^{(j)})$. Evaluation of the angular component $\Delta H_\theta(x_i|x_{i-1})$ is much faster than that of the pairwise distance component $\Delta H_d(\{r_{ab}; x_i\}|x_{i-1})$ since the former requires only a simple density evaluation of the vector $(\phi_i, \psi_i, \omega_i, \chi_i)$, while the latter requires computing 3-D coordinates of atoms determined by x_i and all their pairwise distances with the rest of the protein. Thus it would appear that a natural choice of q_i is $\exp\{-\Delta H_\theta(x_i|x_{i-1})\}$; however, such a choice does not work well in practice since the added atoms will frequently have steric clashes with the rest of the protein. In this case a steric clash leads to $\Delta H_d(\{r_{ab}; x_i\}|x_{i-1}) = +\infty$ in the weight calculation, and those particles would have an importance weight of zero.

The aforementioned difficulty is known as the particle degeneracy problem, and has led to the proposal of various resampling schemes in the SMC literature [Liu and Chen (1998), Douc and Cappé (2005)]. However, simply replicating particles in this context does not serve the purpose of achieving wide exploration of the conformational space (i.e., the result will be an undesirably low particle diversity, hindering our goal of finding the lowest energy conformations) and thus is not a practically useful solution. Rejection control [Liu, Chen and Wong (1998)] instead discards low-weight particles and replaces them by generating new samples, which aids particle diversity but does not address the difficulty of generating promising particles over multiple propagation steps. An alternative that partially mitigates this difficulty is to make multiple trial draws from q_i before selecting one according to the incremental energy ΔH_d, such as that recommended in the configurational-bias Monte Carlo algorithm by Vlugt et al. (1998) and also known as the multiple-try method [Liu, Liang and Wong (2000)]. This increases the probability that steric clashes can be successfully avoided over the length of the segment, but since many trials are needed for each particle, the additional evaluations of ΔH_d are inefficient. Only one of those trials is selected for propagation, and the rest are wasted. Hence the multiple-try approach is also not entirely satisfactory. Fearnhead and Clifford (2003) propose a different type of resampling scheme that makes multiple propagations per particle before resampling, but where the propagations exhaustively enumerate a discrete state space for Kalman filtering. Hence we need to develop a more specialized solution.

A further challenge in designing proposals q_i concerns the role of side chains. For conformational exploration it is sensible to build the backbone and side chain of an amino acid simultaneously, as side chain atoms can occupy a large volume of 3-D space extending from the backbone (see, e.g., Lysine in position 29 on Figure 3). However, the backbone and side chain degrees of freedom have

an inherent conditional structure. Given a fixed backbone with dihedral angles $(\phi_{1:l}, \psi_{1:l}, \omega_{1:l})$, the side chain angles $\chi_{1:l}$ are free to rotate. Thus, fixing the values of χ_i along with $(\phi_i, \psi_i, \omega_i)$ during sequential growth is restrictive and can cause difficulties with side chain steric clashes in later propagation steps. Such clashes might be resolved by going back to earlier positions to rotate their side chain dihedral angles. Hence to properly handle side chains during backbone growth, we need to develop a solution that permits this type of flexibility.

We thus propose novel adaptations to the SIS scheme that are tailored for these specific challenges encountered in building protein segments. First, we design a proposal q_i for the propagation step that incorporates the necessary energy calculations to help ensure that the extensions of each particle $x_i^{(j)}$ conditional on $x_{i-1}^{(j)}$ do not become degenerate. This entails expending additional computational effort to evaluate ΔH_d systematically over a (ϕ_i, ψ_i) grid. Second, we choose to propagate a much larger number of proposals from each particle (up to 100), that is, $x_{i-1}^{(j)} \to (x_i^{(j,1)}, \ldots, x_i^{(j,100)})$, thus greatly increasing the size of an intermediate particle population similar in spirit to Fearnhead and Clifford (2003). Then N particles can be selected as representatives from this larger intermediate population. For our application this strategy permits additional flexibility in the selection of those N representatives to achieve the goals of both low energy and high particle diversity. Third, to handle the difficulty of sampling side chain angles and evaluating their energy along with $(\phi_i, \psi_i, \omega_i)$, we embed a second particle filter within the main sampling steps for handling χ_i. This provides a pool of side chain positions that is sequentially updated during backbone growth to help avoid steric clashes. Together, these novel features yield a highly efficient method for building diverse conformations with low energy for the amino acid segment of interest. Details for these three features appear in the following subsections.

3.1. *Choice of proposal distribution q_i for growth of backbone.* To evaluate ΔH_d more systematically, we first investigated the effect of discretizing the dihedral angles (ϕ, ψ) on backbone protein geometry. We found that using 5° intervals of ϕ and ψ angles provided sufficient resolution to reproduce protein backbones in the PDB with negligible error. Thus this suggests that evaluating $\Delta H_d(\{r_{ab}; \phi_i, \psi_i, \omega_i\} | y_{i-1}^{(j)})$ with (ϕ_i, ψ_i) on a 5° by 5° grid and $\omega_i = 180°$ would be suitable at the ith propagation step, in regions where (ϕ_i, ψ_i) has nonnegligible probability density, that is, $\exp[-\Delta H_\theta(\phi_i, \psi_i | x_{i-1}^{(j)})] > \varepsilon$.[6] Note that at this stage we evaluate ΔH_d conditional on $y_{i-1}^{(j)}$ rather than $x_{i-1}^{(j)}$ to allow the previous side chain atoms within the segment to remain flexible. The effect of those previous side chains will be later handled by our embedded side chain filter in Section 3.3.

[6]The amino acid Proline is an exception as its ω angle has ~0.1 probability to be ~0°, and ~0.9 probability to be ~180°. If amino acid i is Proline, we first sample ω_i to be 0° or 180° according to those probabilities.

Let (ϕ', ψ') be a (ϕ_i, ψ_i) grid point where $\Delta H_d(\{r_{ab}; \phi', \psi', \omega_i\}|y_{i-1}^{(j)}) < +\infty$. For such a (ϕ', ψ'), we next consider the side chains χ_i; let R_i denote the set of rotamer library positions for χ_i provided in Shapovalov and Dunbrack (2011). We then evaluate $\Delta H(\chi_i|\phi', \psi', \omega_i, y_{i-1}^{(j)})$ for $\chi_i \in R_i$.[7] Now, if $\min_{\chi_i \in R_i} \Delta H(\chi_i|\phi', \psi', \omega_i, y_{i-1}^{(j)}) < +\infty$ then there is at least one possible placement of the side chain for the backbone angles (ϕ', ψ').

In the SMC literature, various lookahead strategies have been developed to utilize information from future steps to improve decision making for the current particles; see Lin, Chen and Liu (2013) for an overview. Hence we also introduce that strategy here, to foresee whether (ϕ', ψ') allows for an energetically favorable placement of the next amino acid. While lookahead increases the computational burden for the current step, it can be a useful tradeoff. Nonviable (ϕ', ψ') can be eliminated at the current step, before they are potentially subjected to a full incremental evaluation at step $(i + 1)$ only to find that they are dead ends. For this purpose we opt to evaluate $\Delta H_d(\{r_{ab}; \phi_{i+1}, \psi_{i+1}\}|\phi', \psi', \omega_i, y_{i-1}^{(j)})$ where (ϕ_{i+1}, ψ_{i+1}) takes values on a very coarse 30° grid. The coarser grid is faster to evaluate and suffices to provide a rough assessment of whether backbone growth can continue successfully. If there are no pairs $(\phi'_{i+1}, \psi'_{i+1})$ such that $\Delta H_d(\{r_{ab}; \phi'_{i+1}, \psi'_{i+1}\}|\phi', \psi', \omega_i, y_{i-1}^{(j)}) < +\infty$, then we consider (ϕ', ψ') to be a dead end.

In summary, the above procedure yields a list $L_i^{(j)}$ of grid points (ϕ', ψ') that are potentially good candidates for extending the particle $x_{i-1}^{(j)}$ to the ith amino acid: they are plausible backbone angles according to ΔH_θ, the backbone atoms do not have steric clashes, they have at least one possible side chain position, and the lookahead indicates that further growth is possible. We then set $q_i(x_i|x_{i-1}^{(j)}) \propto \mathbb{1}[(\phi_i, \psi_i) \in L_i^{(j)}] p(\omega_i) p(\chi_i)$. In real structures, ω is not exactly 180° so we take $p(\omega_i)$ to be a normal distribution with SD 2.75°. $p(\chi_i)$ is taken to be uniform here, since the side chains will be handled by our embedded side chain filter.

3.2. *Selection of N particles for further backbone growth.* We incur additional computational cost in constructing q_i, as compared to the basic SIS scheme. To fully leverage the energy calculations already performed, we make multiple (up to 100) draws from q_i for each particle j, so that we propagate $x_{i-1}^{(j)} \to (x_i^{(j,1)}, \ldots, x_i^{(j,100)})$. To encourage diversity, when making draws from q_i we sample grid points (ϕ_i, ψ_i) from the list $L_i^{(j)}$ above, *without replacement*. This propagation of multiple paths per particle results in a much larger intermediate particle population $\{x_i^{(j,1)}, \ldots, x_i^{(j,100)}\}_{j=1}^N$. Then, rather than calculating weights for each particle (SIS) or resampling particles based their weights (SIR), we shall instead

[7] To permit some flexibility in the rotamers and reduce steric clashes, in practice we sample the first side-chain dihedral from a Normal centered at the rotamer position with SD 10°.

aim to obtain a particle population $x_i^{(1)}, \ldots, x_i^{(N)}$ by suitably sampling a subset of $\{x_i^{(j,1)}, \ldots, x_i^{(j,100)}\}_{j=1}^N$ that encourages low energy and high *diversity*.

Suppose that we have evaluated the energies of all the intermediate particles, $\{H(x_i^{(j,1)}), \ldots, H(x_i^{(j,100)})\}_{j=1}^N$, where we obtain each according to

$$H\big(x_i^{(j,n)}\big) = H_\theta\big(y_i^{(j,n)}\big) + H_d\big(\{r_{ab}; y_i^{(j,n)}\}\big) + H\big(\chi_{1:i}|y_i^{(j,n)}\big). \tag{1}$$

[The role of the side chains $\chi_{1:i}$ to compute the term $H(\chi_{1:i}|y_i^{(j,n)})$ will be detailed in Section 3.3.] One simple approach would be to then sample N particles from this list according to $\exp[-H(\cdot)]$. This would favor the low-energy partial conformations, which is ideal if particles currently with the lowest energies can be expected to also have the lowest energies after future propagation steps. However, within the intricate atomic environment of protein structures, growth is rather unpredictable: the eventual completed conformations with the lowest energies may not have been the lowest energy particles at the earlier stages of growth.

The developments presented in the Wang–Landau algorithm [Wang and Landau (2001)] provide some inspiration for an alternative approach that can be preferable for drawing the N particles in this case. That type of idea as we employ here is to stratify the population by energy bands and sample representatives from each band. We define two simple strata—stratum 1 as the N_0 intermediate particles with the lowest energies and stratum 2 as the remaining intermediate particles that are free of steric clashes, with N_0 appropriately chosen ($N_0 \leq N$). We then apply unequal sampling rates. We take all N_0 conformations of stratum 1, and $N - N_0$ conformations sampled uniformly at random from stratum 2. This is effective as a general strategy since it ensures that the low energy particles are well represented for further propagation, and at the same time enough representatives across the energy spectrum are included, some of which might become the most promising low-energy particles after several future steps.

To further reduce the number of particles that are very similar geometrically and encourage diversity, we choose to keep only one representative for very similar particles. We do this by calculating the root-mean-square deviation (RMSD) between each pair of particles, defined as the square root of the mean-squared 3-D Euclidean distance between the corresponding atoms of that segment with the rest of the protein fixed. If any two particles have a pairwise RMSD less than a certain threshold, only the lower energy one is kept. The discarded particles are then replaced by drawing random selections from the remaining intermediate particles in stratum 2 (i.e., those that are free of steric clashes) to return the particle population size to N.

3.3. *Embedding sequential sampling and filtering for side chains.* We now describe our solution for handling the flexibility of side chains given the backbone, while allowing χ_i to be sampled along with $(\phi_i, \psi_i, \omega_i)$ for more efficient exploration of the conformational space during propagation. Our key innovation here is

to embed a second particle filter for side chains within the main particle propagation steps for the backbone. To each intermediate particle $x_i^{(j,n)}$ we associate a set of side-chain particles, $\{s_{i,j,n}^{(k)}\}_{k=1}^{N_s}$. Each side-chain particle is a vector $\chi_{1:i}$ of side chain positions. The optimal side chain positions of $\chi_{1:i}$ for $x_i^{(j,n)}$ is then defined to be the side chain particle s that has the minimum energy for that backbone, that is, $\operatorname{argmin}_{s\in\{s_{i,j,n}^{(k)}\}_{k=1}^{N_s}} H(s|y_i^{(j,n)})$. This definition is used for the $H(\chi_{1:i}|y_i^{(j,n)})$ term to calculate the energy values $H(x_i^{(j,n)})$ in equation (1).

The pool of side chain particles is constructed as follows. At $i = 1$, the side chain particles are initialized by sampling a maximum of n_s values for χ_1 with probability $\propto \exp[-\Delta H(\chi_1|y_1^{(j,n)})]$, where $n_s < N_s$. Then for subsequent steps $i > 1$, in a similar way we first sample a maximum of n_s values for χ_i with probability $\propto \exp[-\Delta H(\chi_i|y_i^{(j,n)})]$. We remind the reader that these energies were already computed as part of the construction of the proposal q_i. Now, to propagate $s_{i-1,j,n}^{(k)} \to s_{i,j,n}^{(k)}$ means extending the existing vectors of $\chi_{1:i-1}$ with χ_i, and hence we now require additionally computing the energy of the interactions of up to $N_s \times n_s$ combinations. In this way we obtain the list of energies $\Delta H(s_{i-1,j,n}^{(k)}, \chi_i^{(k')}|y_i^{(j,n)})$, $k = 1,\ldots,N_s$ and $k' = 1,\ldots,n_s$. We select the lowest N_s energies from this list for the particles $\{s_{i,j,n}^{(k)}\}_{k=1}^{N_s}$.

The embedded side-chain filtering allows for flexibility in the overall positioning of side chains as the main particle propagation proceeds, in that their positions need not be fixed during the initial sampling of the ith amino acid by taking $N_s > 1$. By maintaining a separate set of side-chain particles of size N_s, we are more likely to continue having at least one energetically favorable vector of side chain positions $\chi_{1:i}$ as more amino acids are added. Increasing N_s thus tends to improve the energies of the completed conformations—though we find there is little practical effect beyond $N_s = 25$. Additional computational budget is better utilized for increasing N, the number of backbone particles.

An overall outline summarizing the method is given in pseudocode below.

SMC method for protein segment construction

Initialize $x_0^{(j)} = \varnothing$, $j = 1,2,\ldots,N$

For $i = 1,2,\ldots,l-3$

  For $j = 1,2,\ldots,N$

    Construct list $L_i^{(j)}$ of possible pairs (ϕ',ψ') for amino acid i

      (*refer to Section* 3.1)

    For $n = 1,2,\ldots,100$

      Propagate $x_{i-1}^{(j)} \to x_i^{(j,n)}$ as follows:

Draw one (ϕ_i, ψ_i) pair from $L_i^{(j)}$ and $\omega_i \sim p(\omega_i)$, and
sample up to n_s side chain positions for χ_i
If $i = 1$
Let $\{s_{i,j,n}^{(k)}\}_{k=1}^{N_s}$ contain the n_s positions for χ_1
If $i > 1$
Propagate $\{s_{i-1,j,n}^{(k)}\}_{k=1}^{N_s} \to \{s_{i,j,n}^{(k)}\}_{k=1}^{N_s}$ using embedded side chain filter: set $\{s_{i,j,n}^{(k)}\}_{k=1}^{N_s}$ to be the N_s vectors of $\chi_{1:i}$ with the lowest energies among the $N_s \times n_s$ combinations of $\chi_{1:i-1}$ with χ_i (*refer to Section* 3.3)
End if
End for
End for
Sample N particles from $\{x_i^{(j,1)}, \ldots, x_i^{(j,100)}\}_{j=1}^{N}$ to be $x_i^{(1)}, \ldots, x_i^{(N)}$ (*refer to Section* 3.2)
End for
Do final processing (including analytical closure for last three amino acid positions; see Appendix A.3 for details) and output conformations.

4. Applications and results. We illustrate the proposed method with three applications.

4.1. *Improving* 3*-D structure predictions from homology.* A key motivation for the development of our method is the need to improve 3-D structure predictions beyond those that can be obtained by homology modeling. This is known as the *structure refinement* problem. One important step is to generate improved conformations of interior segments with low homology confidence [Rohl et al. (2004)]. For illustrative purposes we will demonstrate the effectiveness of our method on some specific segments chosen from examples of homology-based structure predictions.

The data for these examples are obtained from the CASP website; since 2006, the CASP experiments have included structure refinement. For this task participants are given the amino acid sequence of the protein along with a starting 3-D structure. The goal is to then submit an improved prediction by modifying the

given structure. The organizers then assess whether the predictions submitted are closer to the truth than the starting structure, according to a set of standard metrics that have been adopted to measure similarity between protein structures [Modi and Dunbrack (2016)].

Here, to compare predictions with the truth (i.e., X-ray crystal structures) we use the backbone RMSD of the entire structure. For comparing entire structures, the calculation of RMSD is done after the two backbones are optimally superimposed (i.e, by rotations and translations). RMSD is a simple measure of overall similarity; an alternative metric is the Global Distance Test (GDT), which we will subsequently introduce.

We take five examples from the CASP12 experiment in May–July 2016, chosen to represent a wide range of accuracies in their given starting 3-D structures. These starting 3-D structures are shown in Figure 4 along with their CASP identifier and RMSD of the entire structure relative to the truth. We have selected one segment from each for optimization using our method; on the 3-D ribbon structures shown these are highlighted in green.

The true structure for the sequence with CASP ID TR879 was publicly released after CASP as 5JMU:A in the PDB, which we have shown in Figure 1. We use that example to describe how segments in the given structure might be selected for optimization without knowledge of the true structure. We first generated a sequence alignment for the amino acid sequence of TR879, by running HHpred [Söding, Biegert and Lupas (2005)] to search the PDB, excluding any matches that were added to the PDB after April 2016. An excerpt of the sequence alignment for TR879 was shown in Figure 2, where it can be seen that the region $\sim$123–134 has little or no alignment matches; the alignment quality histogram at the bottom summarizes this [by showing Blosum62 scores; see Eddy et al. (2004)]. Then we determined the secondary structure elements of the given 3-D structure by running Dictionary of protein secondary structure (DSSP) [Kabsch and Sander (1983)]. In particular, this showed that the given starting structure has a coil region from positions 120–130 followed by an α-helix beginning at position 131. Thus to cover most of that poor alignment region without breaking the long α-helix in the given structure, we chose the segment 120–130 for energy-guided optimization.

In a similar way we chose one segment to optimize from each of the five proteins. The exception is TR898, for which no sequence alignments could be found. It is therefore not surprising that the starting structure given for TR898 has the worst RMSD to the true structure among the proteins considered; it was likely constructed with little guidance from known structures. Hence for TR898, we simply selected the long coil segment (as identified by DSSP) in the given structure to optimize. The chosen segments, along with their energy values in the given structure as computed by the energy function we use, are summarized in Table 1. Their lengths, which range from 11 to 17 amino acids, are long and challenging for sampling methods.

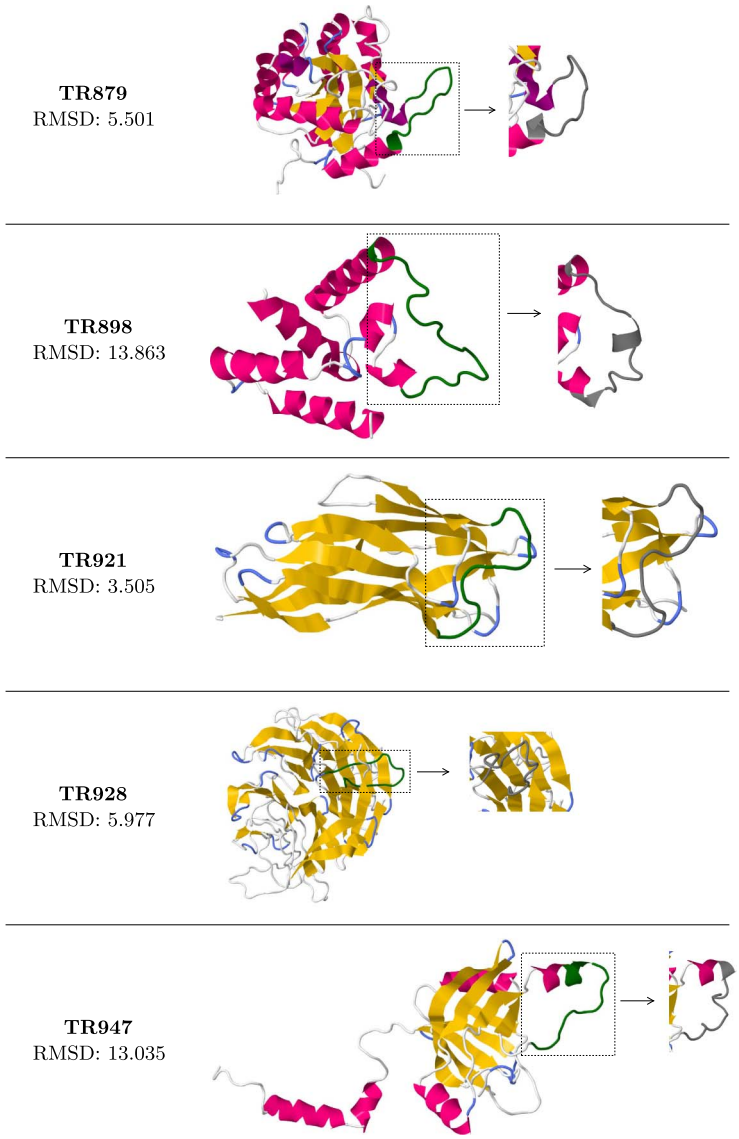

FIG. 4. *Five starting structures for refinement from CASP*12 *and their RMSD to the truth. The segment from each that we selected for optimizing is highlighted in green and displayed inside the box. Replacing each of these segments by the lowest energy conformation leads to a visible conformational change, as indicated in the close-ups to the right of the arrows.*

We applied our method to these segments, using $N = 10{,}000$ particles and an RMSD cutoff of 0.25 Angstroms for pruning conformations that are too similar, as described in Section 3.2. For each segment we output 5000 conformations, from which we selected the conformations with the lowest energy. For example, to illus-

TABLE 1
Segments selected for optimization from the five refinement examples, along with their initial energy values

	Protein				
	TR879	**TR898**	**TR921**	**TR928**	**TR947**
Segment	120–130	56–72	80–91	354–367	174–186
Length	11	17	12	14	13
Initial energy	607.4	505.6	702.8	2070.5	402.8

trate the results for TR879, in Figure 5 we have zoomed into the region of interest and plotted our five sampled conformations with the lowest energy (superimposed on the given structure which has that segment colored green). It can be seen that each of these five conformations are geometrically quite different from that of the given structure, and are also noticeably distinct geometrically among themselves. Their energy values are all within a fairly tight range ∼246–266, which indicates that we are obtaining good coverage of distinct regions of the low-energy conformational space according to this energy function. These energy values are also much lower than the 607.4 in the starting structure. Next we consider the quality of these conformations according to the RMSD measure of overall similarity to the truth. We notice that, by optimizing this length 11 segment out of the 220 amino acids in this protein, we are able to reduce the overall RMSD of the entire structure from 5.501 to the range ∼5.319–5.400, which is substantial since only 5% of the amino acids were modified. It can also be observed that the lowest energy conformation is not necessarily the best one according to this metric; in this case the fourth lowest energy conformation is better, and this phenomenon occurs since energy functions have their inaccuracies. In fact, among the sampled conformations there was one with energy 441.9, while having RMSD 5.292.

We have summarized the results for all five examples in Table 2, showing the RMSD of the lowest energy conformation sampled for each segment. The RMSD

Energy	RMSD	color
246.5	5.400	grey
247.3	5.335	turquoise
256.1	5.354	red
261.0	5.319	blue
265.9	5.391	yellow

FIG. 5. *Characteristics of the five lowest energy segment conformations sampled by our method, for TR*879 *positions* 120–130. *These segments are visualized on the right panel with their respective colors. The conformation of the segment in the given structure is colored green.*

TABLE 2
Summary of results for the five refinement examples. The RMSDs of the lowest energy conformation sampled for each segment are shown, with the changes from the given structure shown in parentheses

Protein	Length	Time (s)	Min. energy	RMSD
TR879	11	143	246.5	5.400 (↘0.100)
TR898	17	343	332.0	13.679 (↘0.184)
TR921	12	167	439.1	3.450 (↘0.055)
TR928	14	275	368.0	5.597 (↘0.380)
TR947	13	331	89.44	12.681 (↘0.354)

of the entire structure is improved in each case—relative to the given structure. More examples of energy inaccuracy are evident, for example, for TR921 the fourth lowest energy conformation has RMSD 3.426 which represents a more substantive improvement than the lowest energy conformation. Finally, we note that our method is fast as indicated in the "Time" column. On a 6-core 3.2 GHz Xeon processor the algorithm completes in under 6 minutes for each of these examples.

While RMSD is often computed on backbone atoms only, it can also be computed with side chain atoms as well (all-atom RMSD) to include an assessment of side chain accuracy. An alternative metric for measuring the similarity of two protein structures is the Global Distance Test (GDT). Like RMSD, the GDT first requires the backbones of two structures to be optimally superimposed by rotation and translation. GDT can be more robust than RMSD, as it uses a sliding window to maximize the fraction of backbone C_α atoms from the two structures that can be superimposed within different cutoff distances. "GDT total score" (GDT_TS) averages that fraction over four cutoffs (1, 2, 4, and 8 Angstroms), while "GDT high accuracy" (GDT_HA) averages over more stringent cutoffs (0.5, 1, 2, and 4 Angstroms). Thus GDT scores range from 0 to 1, where higher is better.

We also assessed the lowest energy conformations found by our method with these additional metrics. These results are summarized in Table 3. In particular, for TR898 which had the worst starting homology-based structure in this set, none of the sampled conformations had any improvement in GDT_TS and GDT_HA values; hence, when the starting structure is overall quite inaccurate relative to the truth, more substantive changes to the structure globally are needed to improve its GDT score.

We note that the fairly simple energy function we use, as detailed in Appendix A.2, is just one of many possibilities that could be used for this purpose and is by no means the most accurate. It may also be unrelated to the energy functions used (if any) by the researchers that built the given starting structure. Hence it is not unusual that we sample conformations with significantly lower energy than in the given structure, according to our energy function. Here the results on these

TABLE 3
Summary of additional RMSD and GDT metrics for the five refinement examples Quality metrics of the lowest energy conformation sampled for each segment are shown, with the changes from the starting structure shown in parentheses

Protein	RMSD	All-atom RMSD	GDT_TS	GDT_HA
TR879	5.400 (↘0.100)	5.875 (↘0.158)	0.7943 (↗0.0045)	0.6364 (↗0.0034)
TR898	13.679 (↘0.184)	14.144 (↘0.133)	0.3679 (0.0000)	0.2524 (0.0000)
TR921	3.450 (↘0.055)	4.328 (↗0.010)	0.6884 (0.0000)	0.4819 (↗0.0019)
TR928	5.597 (↘0.380)	5.812 (↘0.281)	0.6305 (↗0.0037)	0.4274 (0.0000)
TR947	12.681 (↘0.354)	13.541 (↘0.319)	0.6729 (↗0.0115)	0.5214 (↗0.0057)

examples do indicate that our current energy function is generally realistic enough to yield improvements on the overall quality of the structure, if we simply use the lowest energy conformations found to replace the corresponding segments in the given structure. In summary, these results demonstrate the speed and utility of our method for tackling an important step within structure refinement problems. Segment optimization is a very powerful technique when the rest of the protein largely resembles the truth. With the implementation of more accurate energy functions, further improvements would be expected.

4.2. *Predicting reconstructed loop segments*. The second application we present concerns testing the efficacy of the method in a controlled setting, where direct comparisons can be made with the ground truth. In the reconstruction problem, a segment is deleted from a true structure in the PDB, and a sampling method is tasked with generating conformations for the missing segment with the rest of the true structure held fixed. Note that this differs from the refinement application, where the rest of the given structure may at best only approximately resemble the truth. Thus this reconstruction problem is useful for evaluating sampling methods.

For this purpose we evaluate our method using 20 segments from a data set first introduced in Canutescu and Dunbrack (2003), 10 each of length 8 and 12. These were originally selected specifically for study as examples of coil segments that connect α-helices and β-sheets, which are known as *loops* in the bioinformatics literature, and hence this is known as the *loop reconstruction* problem [e.g., Rohl et al. (2004), Coutsias et al. (2004), Soto et al. (2008), Wong, Liu and Kou (2017)]. Most methods that have been proposed in the literature for loop reconstruction do not attempt to minimize energy as part of sampling. So in this section, our comparisons will be made with the state–of–the–art DiSGro method, which has a similar computational speed [Tang, Zhang and Liang (2014)].

The comparison is a useful test for the efficacy of our methodological innovations and in particular how the energies of the sampled conformations compare when evaluated according to the same energy function. DiSGro incorporates dihedral angle probabilities in an ad hoc manner to guide sampling, and then evaluates

final conformations according to their pairwise distance energy function only. In contrast, we use both a dihedral angle energy term and a distance-based energy term during sampling, where we have adopted the DiSGro pairwise distance energy for our distance-based energy term. Thus to obtain the most comparable energy results between the two methods, we shall evaluate final conformations using the DiSGro pairwise distance energy only.

For the loop reconstruction application, the RMSD accuracy of the sampled conformations compared to known conformation in the true structure is also important. While the RMSD accuracy will necessarily depend on the accuracy of the energy function used, it is interesting to compare how the two methods perform in this regard as well. For this purpose, the methods will select one loop conformation as the prediction for each test case. The DiSGro method selects the lowest energy conformation according to their energy function as the prediction; hence, likewise we select our lowest energy conformation (i.e., based on the combined dihedral and distance terms) as our prediction.

To obtain results from the DiSGro method, we ran the program provided by the authors. With default settings it generates 5000 backbone conformations for the computation; to obtain the best possible energies from the DiSGro method we increased that setting to 100,000 conformations. As before, we ran our method using $N = 10{,}000$ particles for the sampler and a RMSD cutoff of 0.25 Angstroms for pruning similar conformations. For both methods we output 5000 final conformations for each test case.

The results are summarized in Table 4 and organized by loop length. We find that for all 20 cases, the conformation with the lowest DiSGro pairwise distance energy is sampled by our method. We also note that the average gap between our lowest energy and the DiSGro method's lowest energy is substantially wider for length 12 loops compared to length 8 loops. Thus the advantages of our method are particularly apparent for longer loops, which is significant since the longer loops have a much larger conformational space to explore. Finally, the rightmost columns compare the RMSD values of the lowest energy conformation—as selected by the respective energy functions of the two methods. As expected, the results are somewhat noisier here due to energy function inaccuracy. Nonetheless, the energy functions are good enough for our method to achieve the lower RMSD in the majority of cases, with a length 8 average of 1.55 vs. 2.12 for DiSGro and a length 12 average of 2.41 vs. 3.45.

4.3. *Assessing different energy functions for sampling and prediction.* The third application concerns assessing the accuracy of different energy functions for conformational sampling and prediction. For a given sequence or segment, the ideal energy function should assign the lowest energy values to conformations that are closest to the truth. Thus with an effective sampling method and the guidance of a good energy function, we ought to find low-energy conformations that should

TABLE 4
Loop reconstruction comparison between our method and DiSGro

	Minimum H_d		RMSD of prediction	
Loop ID	**Ours**	**DiSGro**	**Ours**	**DiSGro**
1cru_85_92	−450.7	−326.3	2.99	5.24
1ctq_144_151	−400.6	−337.4	1.53	1.51
1d8w_334_341	−537.3	−432.5	2.23	2.36
1ds1_20_27	−416.6	−403.5	2.57	0.95
1gk8_122_129	−603.8	−527.5	0.98	1.53
1i0h_145_152	−520.8	−375.4	0.25	1.00
1ixh_106_113	−628.1	−536.6	0.68	0.53
1lam_420_427	−388.8	−318.6	2.05	2.29
1qop_14_21	−614.0	−442.0	0.71	2.67
3chb_51_58	−275.9	−213.6	1.55	1.74
Length 8 average	**−483.7**	**−391.3**	**1.55**	**1.98**
1cru_358_369	−619.5	−451.6	3.25	3.02
1ctq_26_37	−455.3	−310.1	2.24	1.73
1d4o_88_99	−702.7	−419.1	1.40	2.50
1d8w_46_57	−1030.6	−696.4	4.86	4.34
1ds1_282_293	−637.1	−293.5	1.03	5.89
1dys_291_302	−679.5	−464.2	1.33	2.17
1egu_508_519	−709.5	−397.2	1.52	3.02
1f74_11_22	−729.2	−586.1	1.45	1.66
1qlw_31_42	−453.3	−329.3	4.88	4.55
1qop_178_189	−378.7	−267.9	2.16	4.18
Length 12 average	**−639.5**	**−421.6**	**2.41**	**3.31**

also be close to the truth. Among the set of sampled conformations a good energy function should then be able to identify the most truth-like conformation. In practice, energy functions are developed by different research groups with various considerations. Therefore it is natural to ask how different energy functions compare for sampling low-energy conformations. As our method is quite effective for sampling low-energy conformations, independent of the specific choice of energy function, we now demonstrate how we may use it to assess energy function accuracy as well.

Our energy function used for the results thus far is a modified DiSGro energy function, as detailed in Appendix A.2. We adopted the DiSGro energy model for H_d, the energy component for atomic interactions. To make comparisons for different H_d, we also implemented two other energy models for atomic interactions. The first is DFIRE [Zhou and Zhou (2002)], which was designed as a statistical approximation of free energy. The second is the Lennard–Jones potential [Jones (1924)], which is a simple model for atomic Van der Waals attractive and repulsive

forces, and we have implemented the commonly used "12-6" form. We shall use each of these three models in turn for H_d in our energy function and assess their accuracies.

As a test data set, we used the same set of structures from the PDB used to construct our empirical H_θ (see Appendix A.2). Among all coil segments in those structures with lengths 8, 10, or 12, we randomly selected 1000; thus we obtained 560 length 8 segments, 308 length 10 segments and 132 length 12 segments. Similar to the loop reconstruction application, when sampling conformations for each of these segments, we shall fix the rest of the structure at the truth. For each of the three choices of H_d, we run our method with $N = 10{,}000$ particles, RMSD cutoff of 0.25 Angstroms to eliminate particles that are too similar (as described in Section 3.2) and output 5000 final conformations.

We consider two evaluation metrics for each of the three energy functions—(A) the smallest RMSD among the 5000 sampled conformations, and (B) the RMSD of the lowest energy conformation. The first metric assesses how the energy function helps guide the construction of the particle population, in the sense of whether particles with low RMSD to the truth can be retained under our method as particles propagate. The second metric assesses how the energy function performs when selecting the lowest energy conformation from the 5000 as the prediction. For a given test case, the best possible outcome for metric (B) is to be equal to metric (A), that is, the energy function selects the closest conformation to the truth available as the prediction, but metric (B) can be substantially higher due to energy function inaccuracy.

The results are summarized in Table 5, where we have computed the averages of the two metrics over the test cases for each segment length (8, 10, and 12). Naturally, overall RMSD accuracy is lower for longer segments. Similar trends are evident at each length. Sampling with the DiSGro-based energy function yields, on average, the conformation with the smallest RMSD among the particle population. In this regard the simple Lennard–Jones also performs similar or better than DFIRE. Next, when considering prediction performance we notice that all the energy functions have substantial inaccuracy, as the RMSDs of the lowest energy are

TABLE 5
Assessing energy functions using three different energy models of H_d: DiSGro, DFIRE, and Lennard–Jones

		A. Smallest RMSD sampled			**B. RMSD of lowest energy**		
Length	**Cases**	**DiSGro**	**DFIRE**	**L–J**	**DiSGro**	**DFIRE**	**L–J**
8	560	0.637	0.706	0.711	1.701	1.938	2.493
10	308	0.982	1.237	1.155	2.535	2.971	3.679
12	132	1.342	1.804	1.605	3.400	4.063	4.243

on average much larger than the smallest RMSDs sampled. Here, for each segment length the DiSGro-based energy function has again the best performance, followed by DFIRE, and then Lennard–Jones.

Thus in this way we can use our method to help systematically quantify the accuracy of different energy functions. The inaccuracies identified (i.e., low energy conformations with large RMSD to the truth) could be valuable data to use for fitting improved energy functions. We plan to pursue these directions in future research.

5. Conclusion. In this paper we have presented the conformational exploration problem within protein folding from a statistical perspective. We then introduced a new sequential method specifically for stochastically optimizing the energy of protein segments that can find utility in a variety of applications—with improving protein structure predictions being the focal point. That method was inspired by sequential Monte Carlo, where we have made important adaptations to effectively explore and generate diverse samples from the low-energy conformational space. Promising results are obtained in all three applications considered.

The work reported in this paper leads to at least two interesting directions for continued research. First, as the method presented is a useful tool for tackling an important aspect of structure refinement, it is natural to consider how we might use it to improve different segments of the same protein structure in an automatic fashion. This might involve developing algorithms to select plausible segments to refine and then applying our method iteratively on those segments. To achieve further energy optimization, the conformations constructed by our method could then be used as starting points for subsequent local minimization routines. Second, by simply modifying some off-the-shelf energy models to use as our energy function, we found that energy inaccuracies often hindered prediction accuracy. It would be useful to integrate other, perhaps more sophisticated, energy functions with our method. The conformational samples generated by our method could be used for fitting more accurate energy functions in future work as well.

APPENDIX: IMPLEMENTATION DETAILS

Full implementation of the method requires some specifics. First, values need to be specified for each of the adjustable parameters in our method. Second, a specific energy function must be chosen to use with the method. Third, the last propagation step that outputs the completed conformations requires some additional attention to ensure that the segment seamlessly connects the two anchors and to finalize the side chain positions; this incurs some additional computational cost, so it makes sense to choose only a subset of the N particles (i.e., the low energy ones) for processing. These details appear below.

A.1. Specification of method parameters. Here we list the settings we selected for the implementation of our method.

- The minimum backbone dihedral (ϕ, ψ) probability, $\varepsilon = 0.00002$. This choice of ε is simply a device for computational efficiency. The cutoff is set to minimize the computational resources spent on evaluating H_d for (ϕ_i, ψ_i) that have very low dihedral angle probability.
- Possible rotamer positions χ are those provided in the rotamer library Shapovalov and Dunbrack (2011).
- The backbone dihedral angle $\omega_i \sim N(\mu, 2.75°)$, where $\mu = 180°$ for all amino acid types except Proline. For Proline, we sample $\mu = 180°$ with probability 0.9, and $\mu = 0°$ with probability 0.1.
- The size of stratum 1 (the lowest energy particles) is set to $N_0 = 0.9N$.
- Rotamers kept for one amino acid $n_s = 20$, size of side chain particle population $N_s = 25$.

A.2. Energy function. Since the focus of the paper is on sampling, we opted for a simple construction of H_θ, along with using H_d provided by other researchers.

To create a simple probability mass function for (ϕ, ψ) over our 5° by 5° resolution grid, we computed the empirical probabilities of (ϕ, ψ) in each grid cell using structures from the PDB. The structures we used are from the CulledPDB list by PISCES [Wang and Dunbrack (2003)] on March 14, 2015, with these settings—no greater than 20% sequence similarity, resolution 2.0 Å, R-factor cutoff 0.25. We constructed one empirical PMF for each combination of amino acid type (20) and secondary structure type (according to DSSP: helix, sheet, coil). We then compute an energy using $H_\theta(\phi_i, \psi_i) = -\log(\hat{P}(\phi_i, \psi_i))$, where $\hat{P}$ is the appropriate empirical PMF for a given (ϕ_i, ψ_i) pair. For simplicity we did not model any (ϕ, ψ) dependence from one amino acid to the next.

For the H_d component we adopted the DiSGro energy developed in Tang, Zhang and Liang (2014). It is an energy model for pairwise atom interactions, optimized for scoring coil regions of protein structures. In the energy assessment section, we also consider using DFIRE and Lennard–Jones for H_d. Since these energy models do not account for bond lengths and angles, we perform an additional check by setting $H_d = +\infty$ to eliminate particle propagations that will fail to have realistic bond lengths and angles when the segment backbone is completed.

To then weight the contributions of H_θ and H_d, we placed a coefficient of 10 on our H_θ in combining the two components based on some empirical experiments:

$$H(x) = 10H_\theta(x) + H_d\big(\{r_{ab}; x\}\big).$$

These weights achieved good results with our sampling method as demonstrated, but we note that they have not been carefully optimized. Tuning the weights and energy models may further improve the results and is an interesting avenue of future work.

A.3. Final propagation step. As mentioned in Section 2, two pairs of (ϕ, ψ) in the segment are essentially deterministic in order to ensure that the segment backbone connects properly with realistic bond lengths and angles. Thus for a length l segment, the remaining two pairs of (ϕ, ψ) are determined after $(l-3)$th propagation steps. So we begin with $l-4$ regular propagation steps, and for the $(l-3)$th step the lookahead criterion is replaced by a check for whether the segment backbone can connect properly, by solving the polynomial equations in Coutsias et al. (2004). If it cannot, we set $H_d = +\infty$. Otherwise, the ΔH_d contribution of these remaining backbone atoms is added.

Segments that can connect properly will now have a complete segment backbone. We also have N_s side chain particles for $\chi_{1:(l-3)}$ from the $(l-3)$ propagation steps. We briefly describe how we finalize the side chain positions $\chi_{1:l}$ for the entire segment. First, we now fix $\chi_{1:(l-3)}$ to the minimum energy side chain particle in the set $\{s_{l-3,j,n}^{(k)}\}_{k=1}^{N_s}$ and discard the remaining side chain particles. Then we add χ_{l-2}, χ_{l-1}, χ_l incrementally by evaluating the energies of their possible rotamers; with only three side chains to add, simply selecting the minimum energy rotamer for each is fast and often adequate for avoiding steric clashes. Finally, we use a local energy minimization routine that rotates each of the side chains of the segment in turn to stabilize their energy, with the backbone fixed.

REFERENCES

ANFINSEN, C. (1973). Principles that govern the folding of protein chains. *Science* **181** 223–230.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J., MEYER, E. F., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. and TASUMI, M. (1977). The protein data bank. *Eur. J. Biochem.* **80** 319–324.

BROOKS, C. L., ONUCHIC, J. N. and WALES, D. J. (2001). Taking a walk on a landscape. *Science* **293** 612–613.

CANUTESCU, A. and DUNBRACK, R. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12** 963–972.

COOPER, S., KHATIB, F., TREUILLE, A., BARBERO, J., LEE, J., BEENEN, M., LEAVER-FAY, A., BAKER, D., POPOVIĆ, Z. et al. (2010). Predicting protein structures with a multiplayer online game. *Nature* **466** 756–760.

COUTSIAS, E., SEOK, C., JACOBSON, M. and DILL, K. (2004). A kinematic view of loop closure. *J. Comput. Chem.* **25** 510–528.

DILL, K. A. and MACCALLUM, J. L. (2012). The protein-folding problem, 50 years on. *Science* **338** 1042–1046.

DOUC, R. and CAPPÉ, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis*, 2005. *ISPA* 2005. *Proceedings of the* 4*th International Symposium on* 64–69. IEEE, New York.

DOUCET, A., DE FREITAS, N. and GORDON, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*. 3–14. Springer, New York. MR1847784

EDDY, S. R. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.* **22** 1035–1036.

ENGH, R. and HUBER, R. (1991). Accurate bond and angle parameters for X-ray protein-structure refinement. *Acta Crystallogr. Sect. A* **47** 392–400.

FEARNHEAD, P. and CLIFFORD, P. (2003). On-line inference for hidden Markov models via particle filters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 887–899. MR2017876

FISER, A. and ŠALI, A. (2003). Modeller: Generation and refinement of homology-based protein structure models. *Methods Enzymol.* **374** 461–491.

FRIESNER, R. A., PRIGOGINE, I. and RICE, S. A. (2002). *Computational Methods for Protein Folding*. Wiley, New York.

JONES, J. E. (1924). On the determination of molecular fields. II. From the equation of state of a gas. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **106** 463–477.

KABSCH, W. and SANDER, C. (1983). Dictionary of protein secondary structure—pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22** 2577–2637.

KENDREW, J. C., BODO, G., DINTZIS, H. M., PARRISH, R., WYCKOFF, H. and PHILLIPS, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181** 662–666.

KHOURY, G. A., SMADBECK, J., KIESLICH, C. A. and FLOUDAS, C. A. (2014). Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol.* **32** 99–109.

KRISSINEL, E. (2007). On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* **23** 717–723.

LAZARIDIS, T. and KARPLUS, M. (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struck. Biol.* **10** 139–145.

LEE, D., REDFERN, O. and ORENGO, C. (2007). Predicting protein function from sequence and structure. *Nat. Rev., Mol. Cell Biol.* **8** 995–1005.

LI, J., ABEL, R., ZHU, K., CAO, Y., ZHAO, S. and FRIESNER, R. A. (2011). The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins* **79** 2794–2812.

LIANG, S., ZHANG, C. and STANDLEY, D. M. (2011). Protein loop selection using orientation-dependent force fields derived by parameter optimization. *Proteins* **79** 2260–2267.

LIANG, S., ZHANG, C. and ZHOU, Y. (2014). LEAP: Highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. *J. Comput. Chem.* **35** 335–341.

LIN, M., CHEN, R. and LIU, J. S. (2013). Lookahead strategies for sequential Monte Carlo. *Statist. Sci.* **28** 69–94. MR3075339

LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR1842342

LIU, J. S. and CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* **93** 1032–1044. MR1649198

LIU, J. S., CHEN, R. and WONG, W. H. (1998). Rejection control and sequential importance sampling. *J. Amer. Statist. Assoc.* **93** 1022–1031. MR1649197

LIU, J. S., LIANG, F. and WONG, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.* **95** 121–134. MR1803145

MANDELL, D. J., COUTSIAS, E. A. and KORTEMME, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6** 551–552.

MODI, V. and DUNBRACK, R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins* **84** 260–281.

MOULT, J., FIDELIS, K., KRYSHTAFOVYCH, A., SCHWEDE, T. and TRAMONTANO, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins* **84** 4–14.

ONUCHIC, J. N., LUTHEY-SCHULTEN, Z. and WOLYNES, P. G. (1997). Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48** 545–600.

RAMACHANDRAN, G., RAMAKRISHNAN, C. and SAISEKHARAN, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7** 95–99.

ROHL, C. A., STRAUSS, C. E., CHIVIAN, D. and BAKER, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55** 656–677.

SHAPOVALOV, M. V. and DUNBRACK, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19** 844–858.

SÖDING, J., BIEGERT, A. and LUPAS, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*. **33** W244–W248.

SOTO, C. S., FASNACHT, M., ZHU, J., FORREST, L. and HONIG, B. (2008). Loop modeling: Sampling, filtering, and scoring. *Proteins* **70** 834–843.

TAN, K., GU, M., CLANCY, S. and JOACHIMIAK, A. (2016). The crystal structure of the catalytic domain of peptidoglycan N-acetylglucosamine deacetylase from Eubacterium rectale ATCC 33656 (CASP target). PDB ID: 5JMU. DOI:10.2210/pdb5jmu/pdb.

TANG, K., ZHANG, J. and LIANG, J. (2014). Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte Carlo method. *PLoS Comput. Biol.* **10** e1003539.

VLUGT, T., MARTIN, M., SMIT, B., SIEPMANN, J. and KRISHNA, R. (1998). Improving the efficiency of the configurational-bias Monte Carlo algorithm. *Mol. Phys*. **94** 727–733.

WANG, G. and DUNBRACK, R. L. (2003). PISCES: A protein sequence culling server. *Bioinformatics* **19** 1589–1591.

WANG, F. and LANDAU, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett*. **86** 2050–2053.

WICK, C. and SIEPMANN, J. (2000). Self-adapting fixed-end-point configurational-bias Monte Carlo method for the regrowth of interior segments of chain molecules with strong intramolecular interactions. *Macromolecules* **33** 7207–7218.

WONG, W., CUI, Y. and CHEN, R. (1998). Torsional relaxation for biopolymers. *J. Comput. Biol*. **5** 655–665.

WONG, S. W. K., LIU, J. S. and KOU, S. C. (2017). Fast *de novo* discovery of low-energy protein loop conformations. *Proteins* **85** 1402–1412.

ZHANG, J., KOU, S. C. and LIU, J. S. (2007). Biopolymer structure simulation and optimization via fragment regrowth Monte Carlo. *J. Chem. Phys*. **126** 225101. DOI:10.1063/1.2736681.

ZHANG, J., LIN, M., CHEN, R., LIANG, J. and LIU, J. S. (2007). Monte Carlo sampling of near-native structures of proteins with applications. *Proteins* **66** 61–68.

ZHOU, H. and ZHOU, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*. **11** 2714–2726.

S. W. K. WONG
DEPARTMENT OF STATISTICS
GRIFFIN-FLOYD HALL
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32608
USA
E-MAIL: swkwong@stat.ufl.edu

J. S. LIU
S. C. KOU
DEPARTMENT OF STATISTICS
1 OXFORD ST 7TH FL
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: jliu@stat.harvard.edu
kou@stat.harvard.edu