Monte Carlo Sampling of Protein Folding by Combining an All-Atom Physics-Based Model with a Native State Bias

Yong Wang,[†,§] Pengfei Tian,[†,‡,§] Wouter Boomsma,[¶] and Kresten Lindorff-Larsen[*,†]

Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5 DK-2200 Copenhagen N, Denmark, Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, United States, and Department of Computer Science, University of Copenhagen, 2100 Copenhagen Ø, Denmark

E-mail: lindorff@bio.ku.dk

Abstract

Energy landscape theory suggests that native interactions are a major determinant of the folding mechanism of a protein. Thus, structure-based (Gō) models have, aided by coarse-graining techniques, shown great success in capturing the mechanisms of protein folding and conformational changes. In certain cases, however, non-native interactions

[*]To whom correspondence should be addressed

[†]Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5 DK-2200 Copenhagen N, Denmark

[‡]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, United States

[¶]Department of Computer Science, University of Copenhagen, 2100 Copenhagen Ø, Denmark

[§]Contributed equally to this work

and atomic details are also essential to describe the protein dynamics, prompting the development of a variety of structure-based models which include non-native interactions, and differentiate between different types of attractive potentials. Here, we describe an all-protein-atom hybrid model, termed ProfasiGo, that integrates an implicit solvent all-atom physics-based model (called Profasi) and a structure-based Gō potential, and its implementation in two software packages (PHAISTOS and ProFASi) that are developed for Monte Carlo sampling of protein molecules. We apply the ProfasiGo model to study the folding free energy landscapes of four topologically similar proteins, one of which can be folded by the simplified potential Profasi, and two that have been folded by explicit solvent, all-atom molecular dynamics simulations with the CHARMM22* force field. Our results reveal that the hybrid ProfasiGo model is able to capture many of the details present in the physics-based potentials, while retaining the advantages of Gō models for sampling and guiding to the native state. We expect that the model will be widely applicable to study the folding of more complex proteins, or to study conformational dynamics and integration with experimental data.

Introduction

It is an essential biological fact that most,[1] though not all,[2] naturally-occurring proteins can self-organize to ordered three-dimensional structure(s). There has thus been an enormous progress in solving protein structures, as evidenced by the observation that the Protein Data Bank has collected more than 142,000 structures up to date. Despite substantial progress in combining experiments, theory and simulations to study protein folding,[3–9] there is, however, a substantial gap between the number of structures we know and the proteins for which we know the folding mechanism. In addition to the intellectual challenge involved in understanding the protein folding mechanism, modeling protein folding has potential uses in for example protein design,[10,11] in molecular drug development,[12] and in interpreting pathogenicity of genomic sequence variation.[13]

Recent advances in computer hardware, methods for enhancing sampling and protein force fields have made simulations an irreplaceable tool in the study of protein folding.[14–18] In principle, a long, equilibrium molecular dynamics (MD) simulation based on an accurate all-atom, physics-based, explicit-solvent model can not only provide spatial and temporal details on the structural ensembles of folded states, but also elucidate the mechanism of folding/unfolding transitions.[19] While this has been achieved for small fast-folding proteins,[5] and even a natural protein,[20] such work generally requires access to extensive sampling using specialized hardware (e.g. Anton[21]). Further, most proteins are not 'fast-folding',[22] and although it is possible also to reach long timescales through using e.g. Markov state models[23] or enhanced sampling techniques,[24] it will not be possible to study folding processes for many proteins using routine all-atom MD simulations in the foreseeable future.

As an alternative to the detailed all-atom physics-based models, native-structure-based models (also called Gō models[25]) have been widely applied to investigate the folding and assembly mechanisms of ordered and disordered proteins.[26–28] These models are applicable to larger sizes, complex topologies and slow kinetics, especially when aided by coarse-graining techniques.[29,30] The success of these models has been explained by the proposal that such models naturally realise a key feature of naturally-occurring proteins, that is, a minimally frustrated and ideally funnel-shaped energy landscape,[31] and indeed analyses of all-atom MD simulations reveal the central role of native contacts.[32] The principle of minimal frustration in energy landscape theory directly leads to a conclusion that protein topology is a key factor governing the folding mechanism.[32–34] Currently, there are many software tools available to build and simulate Gō-type models, including SMOG,[35] AWSEM-MD,[36] CafeMol,[37] MMTSB,[38] CHARMMing,[39] eSBMTools,[40] NAMD-Go,[41] SOP-GPU,[42] GENESIS,[43] MonteGrappa,[44] and SIMONA.[45] Most of them are based on MD simulation though the last two utilise a Monte Carlo (MC) framework. Also, nearly all previously used Gō-type models have employed a coarse grained representation of the protein.

In most Gō-type models, the native interactions are emphasised by an attractive poten-

tial, while the interactions not present in the native folded structures (non-native interactions) are usually simply treated with a repulsive potential. Nevertheless, these non-native interactions may have significant impact on folding process by adding 'roughness' to the energy landscape,[46–50] such as trapping in misfolded or intermediate states and causing aggregation and disease.[49] Residual non-native interactions, resulting in local violations of the minimal frustration principle,[51] are considered to be a consequence of the conflicting requirements of foldability and function of a protein sequence.[52] Opposite to the common view that the non-native interactions contribute primarily to the roughness of landscapes and frustrate the folding process, there are also cases that demonstrate that the non-native interactions facilitate the biological process and play an effective role in protein folding.[53–57] The potentially important role and related open questions of non-native interactions have driven the development of many enhanced structure-based models[58] by introduction of additional potentials (e.g. the Debye-Hückel-type potential to approximate electrostatic interactions at low salt concentrations,[59–61] and the Gaussian potential to model hydrophobic interactions[48,56,62]) and heterogeneous energetic parameters to distinghuish between different types of contacts.[30,60,63,64]

Inspired by previous hybrid models and multi-scale strategies,[65–71] we have developed a hybrid physics-based and structure-based model (denoted as ProfasiGo model) within the framework of both PHAISTOS[72] and ProFASi[73] simulation packages for Monte Carlo simulation of protein molecules. In our model, the physics-based term is inherited from the implicit solvent force field, denoted as Profasi, which has previously been used extensively to study protein folding, aggregation and protein structure determination.[74–78] (Note that there is both a simulation software package and an energy function called Profasi; we use the term Profasi for the energy function and ProFASi for the software package.) The physics-based term is transferable and preserves the atomistic representation (including hydrogen atoms) of the protein. We then introduce the structure-based potential (E_{Go}) as an additional term, thus 'funnelling' the underlying energy landscape further, so as to accelerate the folding

transitions. In this way the hybrid model is designed to be able to capture more complex energy landscapes. In addition, our software architecture facilitates the investigation of the driving force in protein folding (e.g. electrostatic, hydrophobic interactions and hydrogen bonds) through adjustment of the corresponding potential terms.

In this paper we focus on describing and validating the approach in studies of protein folding. In particular we study four α-helical bundles, the designed proteins α3W and α3D, and the homeodomains EnHD and UVF, that pairwise have similar topologies but differ in folding mechanism. One protein can be folded with the pure Profasi force field, and two with the all-atom CHARMM22* force field, enabling us to examine the extent to which the hybrid model may capture folding mechanisms in a force field without the structure-based term.

Models and Methods

Profasi model

The Profasi model belongs to the class of implicit solvent all-atom models (including all hydrogen atoms) designed for MC simulation[73] and has been applied in protein folding, aggregation, and determination of protein structures and ensembles with experimental restraints.[74,75,77–80] In Profasi the flexible degrees of freedom are the Ramachandran (ϕ and ψ) and side chain (χ) torsional angles, whereas bond lengths and angles, and peptide plane ω torsional angles, are fixed. The interaction potential is composed of four terms:

$$E_{\text{Profasi}} = E_{\text{local}} + E_{\text{repulsive}} + E_{\text{HB}} + E_{\text{sidechain}}.$$

The first term, E_{local}, accounts for local interactions between atoms, such as the electrostatic interactions between adjacent residues. The other three terms ($E_{\text{repulsive}}$, E_{HB} and $E_{\text{sidechain}}$) account for non-local interactions: excluded-volume effects, hydrogen-bond interactions, and

residue-specific interactions between pairs of side-chains, respectively. The precise form of these four terms can be found in the original description.[75]

Atomic Gō model

In general, the potential energy function of any Gō-type models, E_{Go}, as a function of the coordinates of the native structure, Γ_0, can be simplified into three terms:

$$\begin{aligned} E_{\mathrm{Go}}(\Gamma_0) &= E_{\mathrm{bonded}}(\Gamma_0) + E_{\mathrm{nonbonded}}(\Gamma_0) \\ &= E_{\mathrm{bonded}}(\Gamma_0) + E_{\mathrm{repulsive}}(\Gamma_0) + E_{\mathrm{attractive}}(\Gamma_0) \end{aligned}$$

The first two terms, E_{bonded} and $E_{\mathrm{repulsive}}$, maintain the correct backbone geometry, while the last term $E_{\mathrm{attractive}}$ determines the folding of a peptide chain by attractive inter-atom or coarse grained inter-residue interactions. These attractive interactions are generally defined by a pairwise contact list derived from native structure, called the native contact map. Construction of a native contact map is thus a key step to build a Gō model. In the context of a standard Gō model, the short-range forces to stabilise the native state (e.g. hydrogen bonding, salt bridges and VDW interactions), are approximately represented by the native contact map, while the long-range or nonlocal interactions, like protein-water interactions or water-mediated interactions, are considered to be averaged out and described using a mean field perspective. Currently, there are several algorithms to define the native contact map, including a cutoff-based algorithm,[29,55] Shadow Contact Map (SCM),[81] and Contacts of Structural Units.[82] The native contact maps calculated by different algorithms differ from each other to certain extents, but the resulting thermodynamic properties and folding mechanism are reasonably consistent and robust.[83–85]

For comparison with our hybrid model, we constructed a pure atomic (without hydrogen atoms) Gō model using SMOG[35] with default parameters (an all-atom contact map by SCM with a cutoff of 6.0 Å). Briefly, SCM is an algorithm to determine contacts between

interior protein surfaces without allowing unphysical or occluded contacts.[81] Here, the all-atom attractive Gō potential $E^{\mathrm{AA}}_{\mathrm{attractive}}(\Gamma_0)$ is expressed by a Lennard-Jones (LJ) potential:

$$E^{\mathrm{AA}}_{\mathrm{attractive}}(\Gamma_0) = \sum_{i<j-3} \epsilon_{\mathrm{Go}}(i,j)[(\frac{\sigma_{ij}}{r_{ij}})^{12} - 2(\frac{\sigma_{ij}}{r_{ij}})^{6}],$$

where, σ_{ij} and r_{ij} are the native and instantaneous distance between atom i and atom j, and $\epsilon_{\mathrm{Go}}(i,j)$ is the strength of pairwise attractive potential between atom i and atom j. It was homogeneously set to be 1.0 in this work, although it could be tuned to introduce sequence information,[64,86] through e.g. the Miyazawa-Jernigan matrix[63] or multi-scaling methods.[87] Aiming to represent a standard Gō model and for fair comparison, we kept the energetic parameters as general as possible.

ProfasiGo model

We integrated a structure-based potential (E_{Go}) into the Profasi force field, and termed this hybrid model 'ProfasiGo': $E_{\mathrm{ProfasiGo}} = E_{\mathrm{Profasi}} + E_{\mathrm{Go}}$. We opted to use a coarse-grained version of the native contact map in which only C_α-C_α contacts are included so as to introduce minimal extra potential into Profasi force field. In other words, the Gō potential was introduced as a minimal perturbation. We implemented two variants of E_{Go} into both the ProFASi and PHAISTOS software packages (which already implemented the Profasi energy function). One is based on a 12-10 LJ-like potential:

$$E^{1210}_{\mathrm{Go}} = E^{\mathrm{CA}}_{\mathrm{attractive}}(\Gamma_0) = \sum_{i<j-3} \epsilon^{\mathrm{CA}}_{\mathrm{Go}}(i,j)[5(\frac{\sigma_{ij}}{r_{ij}})^{12} - 6(\frac{\sigma_{ij}}{r_{ij}})^{10}];$$

and the other uses a 12-10-6 potential described by

$$E^{12106}_{\mathrm{Go}} = \sum_{i<j-3} \epsilon^{\mathrm{CA}}_{\mathrm{Go}}(i,j)[13(\frac{\sigma_{ij}}{r_{ij}})^{12} - 18(\frac{\sigma_{ij}}{r_{ij}})^{10} + 4(\frac{\sigma_{ij}}{r_{ij}})^{6}].$$

They represent the potential functions used in two popular versions of the coarse-grained Gō model: the Clementi-Onuchic model[29] and the Karanicolas-Brooks model.[30] The E_{Go}^{12106} function is a modified LJ potential (Fig. S1) that incorporates a low energy barrier (a desolvation penalty) which has been shown to be able to increase the folding cooperativity of two-state folders,[88,89] and improve model prediction.[90] In both formulations, $\epsilon_{\mathrm{Go}}^{\mathrm{CA}}(i,j)$ determines the depths of the potential wells and thus sets the strength of the native bias relative to the E_{Profasi} term. To keep the different models self-consistent, we built the coarse-grained native contact map used in the ProfasiGo model from the all-atom geometric occlusion contact map used in the pure Gō model described above. In particular, we considered two residues to be in contact in the ProfasiGo model if they share at least one atomic contact in the atom-based Gō model.

MC simulations with Profasi and ProfasiGo model

The MC simulations with Profasi model and ProfasiGo model were performed by the modified versions of ProFASi[73] or PHAISTOS software,[72] both of which have implemented the Profasi force field.[75] The patches for adding the Gō functions to the ProFASi and PHAISTOS software will be available at http://github.com/XXX.

We simulated α3W using parallel tempering/replica exchange (PT) with a set of eight temperatures ranging from 279K to 394K with the same interval. To get efficient sampling of the free energy landscape, we also used MUNINN, which employs the generalized multi-histogram equations[91,92] and a nonuniform adaptive binning of the energy space, ensuring efficient scaling to large systems. We used a β (inverse temperature) ranging from 1.3 to 2.4, corresponding to a temperature range of 278K to 513K.

Three different elementary MC moves are used in the simulations: (a) biased Gaussian steps (BGS), (b) rotations of individual side-chain angles (Rot), (c) pivot-type rotations about individual backbone bonds (Pivot). The BGS move is semi-local and updates up to eight consecutive backbone degrees of freedom but keeps the ends of the segment approxi-

mately fixed.

MD simulations with pure Gō model

Simulations with the pure Gō model were performed with Gromacs 4.6.5.[93] The dynamics of the systems was simulated using the Langevin thermostat with friction coefficient of $\gamma = 1.0$. Reduced units and a time step of 0.5 fs were used. Multiple trajectories were collected at a temperature range around the folding temperature (T_f) for each protein system. The length of each simulation is 4x10^8 simulation steps to include dozens of folding/unfolding transitions. We saved conformations every 2000 integration steps.

Order parameters to characterise the folding mechanism

The fraction of native contacts, Q, has been shown to be good reaction coordinate in the study of protein folding.[32,94] To describe the folding mechanism of the three-helical bundle proteins we also employed three local order parameters for helix formation, Q_{H1}, Q_{H2}, Q_{H3}, and three order parameters that describe the pairwise assembly of helices, $Q_{\text{H1-H2}}$, $Q_{\text{H2-H3}}$, $Q_{\text{H1-H3}}$. In all cases, Q_{X} is a measure of the progress of helix formation or assembly, by quantifying how far native contacting atoms are from their respective reference distances. More precisely, Q_{X} is a summation over the native contact pairs in the list denoted as X:

$$Q_X = \frac{1}{N_X} \sum_{(i,j)\in X}^{N_X} \frac{1}{1 + e^{\beta(r_{ij} - \lambda r_{ij}^0)}},$$

where X can be H1, H2, H3, H1-H2, H1-H3 and H2-H3, which are the lists of intra-segment contacts of helix1, helix2, helix3 and inter-segment contacts between them. Here, r_{ij} is the distance between atom i and atom j in instantaneous structure (in units of nm), while r_{ij}^0 is the corresponding native distance. We set β=50 nm^{-1} and λ=1.5 in this work. Defined in this way, Q_X values fluctuate between 0 (non-native) and 1 (native).

In addition, we employed four folding order parameters: $Q_{\text{secondary}}$ to measure the frac-

tion of native contacts within the three helices, E_{HB} (backbone hydrogen bond energy) to quantify the formation of secondary structure, P_{helix} to measure the proportion of α helical content, and P_{beta} to measure the proportion of β strand content, and two assembly order parameters: Q_{tertiary} to measure the fraction of native contacts between the three helices, and E_{HP} (hydrophobic energy), to quantify the formation of hydrophobic core.

Table 1: Atomistic Models and their ability to fold α3W, α3D, EnHD and UVF

System	Size (a.a.)	Topology	Profasi	SMOG	ProfasiGo	CHARMM22*
α3W	67	left-handed	Y	Y	Y	-
α3D	73	right-handed	N	Y	Y	Y
EnHD	54	right-handed	-	Y	Y	N
UVF	52	right-handed	N	Y	Y	Y

'Y', 'N' and '-' refer to the ability of the force field or model to fold the protein. 'Y': folded successfully, 'N': failed to fold, '-': unknown.

Selection of model proteins

Our goal was to develop the hybrid ProfasiGo model, and to test its range of applicability by benchmarking against other possible methods for studying protein folding. We thus focused our work on four three-helix bundle proteins whose folding mechanisms have previously been examined. The two proteins α3W and α3D are designed proteins that consist of three α-helices connected by two turns (Fig. 1A-B). According to the arrangement of the helices, the topology of α3W and α3D are considered to be left-handed and right-handed, respectively.[95,96] In this sense they represent a pair of proteins with similar topologies but different conformational 'chiralities'. We also chose the engrailed homeodomain (EnHD) and a designed thermostabilized homeodomain (UVF)[97] (Fig. 1C-D) which are also three-helix bundle proteins. EnHD and UVF represent a pair of proteins with almost the same size and same handedness of the arrangement of the helices (Fig. 1 and Table 1). The similar topology of these four proteins allows us to use a consistent set of order parameters (as defined in the

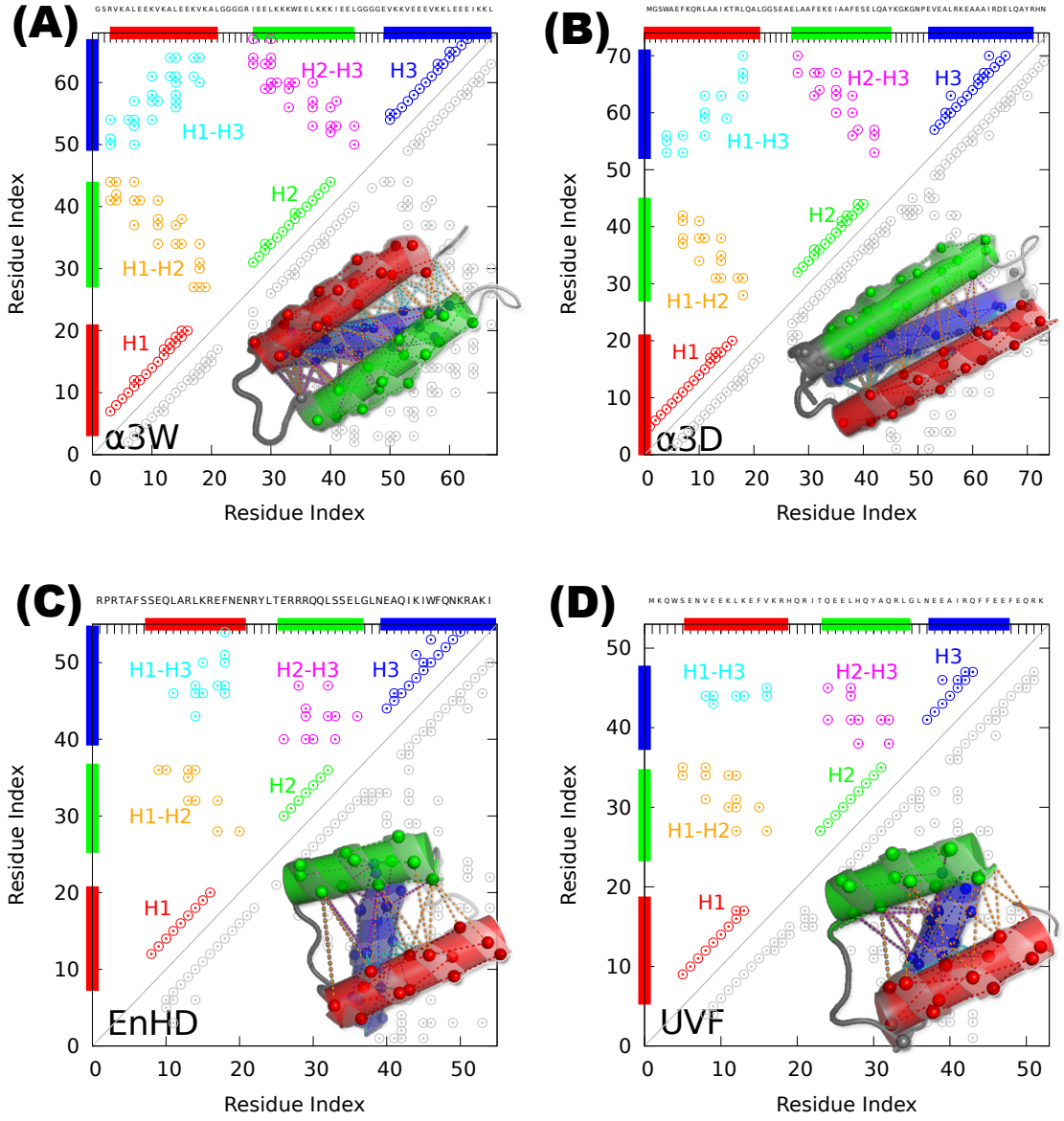


Figure 1: **Contact maps and structure of the four three-helix bundle proteins studied here.** Helices 1, 2 and 3 are coloured in red, green and blue, respectively. The inter-helix contacts between helix 1 and 2, between helix 1 and 3, and between helix 2 and 3 are coloured in orange, cyan and magenta, respectively. α3W and α3D share similar topology but different handedness of the orientation of the three helices; their sequence identity is 18%. EnHD and UVF have the same topology and handedness and a 23% sequence identity.

Method section) to characterise the folding mechanism. Because of the inclusion of a native-state bias, both the pure Gō model and the ProfasiGo model are expected to fold all four proteins to their native states. This is, however, not the case for the two physics-based force fields (Table 1). By looking across all four proteins, we can compare the folding mechanism observed in the ProfasiGo model with the results of three other models (a pure Gō model, the simple Profasi force field and an all-atom explicit solvent force field (Table 1).

The remainder of the manuscript is thus constructed as follows. (i) We first compare the folding mechanism of a single protein (α3W) in the hybrid ProfasiGo model with that of the parent Profasi model. (ii) We then compare the folding mechanism of the four proteins under the ProfasiGo model to the results in a pure Gō model, and examine the dependency of the results on model parameters. (iii) Finally, we compare the folding mechanism of α3D and UVF in ProfasiGo and the all-atom, explicit solvent CHARMM22* force field simulations.

Results and Discussion

ProfasiGo and unbiased Profasi capture similar folding mechanisms of α3W

The designed α3W protein has previously successfully been folded by simulations with both coarse-grained models[98–100] and all-atom force fields.[5,95,101–103] In particular, its folding thermodynamics has been characterized by Irbäck et al. using the pure Profasi force field,[75] making it particularly suitable to be used for testing and calibrating our ProfasiGo model.

We sampled the free energy landscape of α3W with the ProfasiGo model in the multicanonical ('flat-histogram') ensemble using the MUNINN software.[91,92] Such a generalized ensemble method can not only improve sampling efficiency, but also directly helps determine the melting temperature where folding and unfolding transitions typically occur most frequently, a time-consuming but often necessary process in MD or MC simulations.[37,64] Subsequently, the thermodynamics properties at any temperature of interest can be obtained by

reweighting techniques.[104]

Throughout this manuscript we explore protein folding through such enhanced sampling simulations, examining folding mechanisms by analysing and comparing free energy profiles. As an example, we calculated three order parameters for folding, Q_{H1}, Q_{H2} and Q_{H3} (see definitions in Models and Methods), that describe the formation of each of the three helices, and project the conformational space onto the two-dimensional free energy surfaces spanned by combinations of these coordinates (Fig. 2A). Free energy surfaces as a function of such order parameters have been widely used to elucidate protein folding and assembly mechanisms through minimum free energy pathways.[62,64] For α3W, we observe that there are no low energy pathways along the diagonal lines in these free energy surfaces (F(Q_{H1},Q_{H2}), F(Q_{H1},Q_{H3}) and F(Q_{H2},Q_{H3})), indicating that the folding of the three helices is independent of one another, without strong coupling. Furthermore, there are multiple possible folding routes, suggesting heterogeneity in the order of formation of the three helices.

In addition to examining the order of formation of the different secondary structure elements, we also analysed the relationship between formation of secondary and tertiary structure; such an analysis would be useful to distinguish between a nucleation condensation model, diffusion collision or framework model, or hydrophobic collapse model for folding. We thus calculated two additional local order parameters: $Q_{\mathrm{secondary}}$ to measure the fraction of native contacts within the three helices and P_{helix} to measure the fraction of helical content, and two order parameters aimed to capture tertiary interactions: Q_{tertiary} to measure the fraction of native contacts between the three helices and the hydrophobic energy, E_{HP}, to quantify the energy of forming the hydrophobic core. The two-dimensional free energy surfaces as a function of E_{HP} and P_{helix}, and $Q_{\mathrm{secondary}}$ and Q_{tertiary} illustrate a clear pathway that helix folding occurs before the formation of the hydrophobic core (Fig. 2A). Therefore, the thermodynamic free energy analysis suggests the folding of α3W in this model can be well described by the diffusion collision model[105] by which the native secondary structures are formed before the tertiary structures. This conclusion is consistent with previous

studies.[66,101]

Having analysed the folding free energy surfaces in our new hybrid model we proceed to compare with the surfaces obtained in the same model, but without the native bias. The basic hypothesis is that, for the proteins that can be folded by the physics-based (non-Gō) Profasi model, the hybrid model would retain most of the features observed in the less biased model. By taking the same protocol as previously applied in the work of Irbäck et al,[75] we performed replica-exchange MC simulations with the pure Profasi model, and determined the folding free energy surfaces for α3W (Fig. 2B). By comparing with the free energy surfaces obtained from the hybrid ProfasiGo model (Fig. 2A), we find overall similar shapes of the free energy landcapes, suggesting a similar mechanism for folding of α3W. In addition to additional 'roughness' of the landscape in the pure Profasi model, the major difference are two intermediate states present on the free energy surface F(E_{HP},P_{helix}) sampled by the Profasi model. Inspection of the structures of these intermediate states revealed that the consist of very long helices or a high proportion of beta strands, but without any substantial hydrophobic packing, which we consider to be artefacts of the Profasi model.

In summary, the results suggest that α3W folds by a diffusion collision mechanism in both the pure Profasi model and in the hybrid ProfasiGo model. This observation supports the idea that the introduction of the native-biased Gō potential mostly acts to smoothen the energy landscape of protein folding, but does not substantially change the folding mechanism. This in turn suggests that simulations of protein folding with the hybrid model would yield realistic folding mechanisms even in cases where folding simulations are not possible with the pure Profasi model.

Comparing the hybrid model with a pure Gō model

While the ProfasiGo model shows the ability to reproduce the folding mechanism as revealed by the unbiased Profasi model, we also examined whether the pure Gō model would suggest a similar mechanism. We performed constant temperature MD simulations using a pure

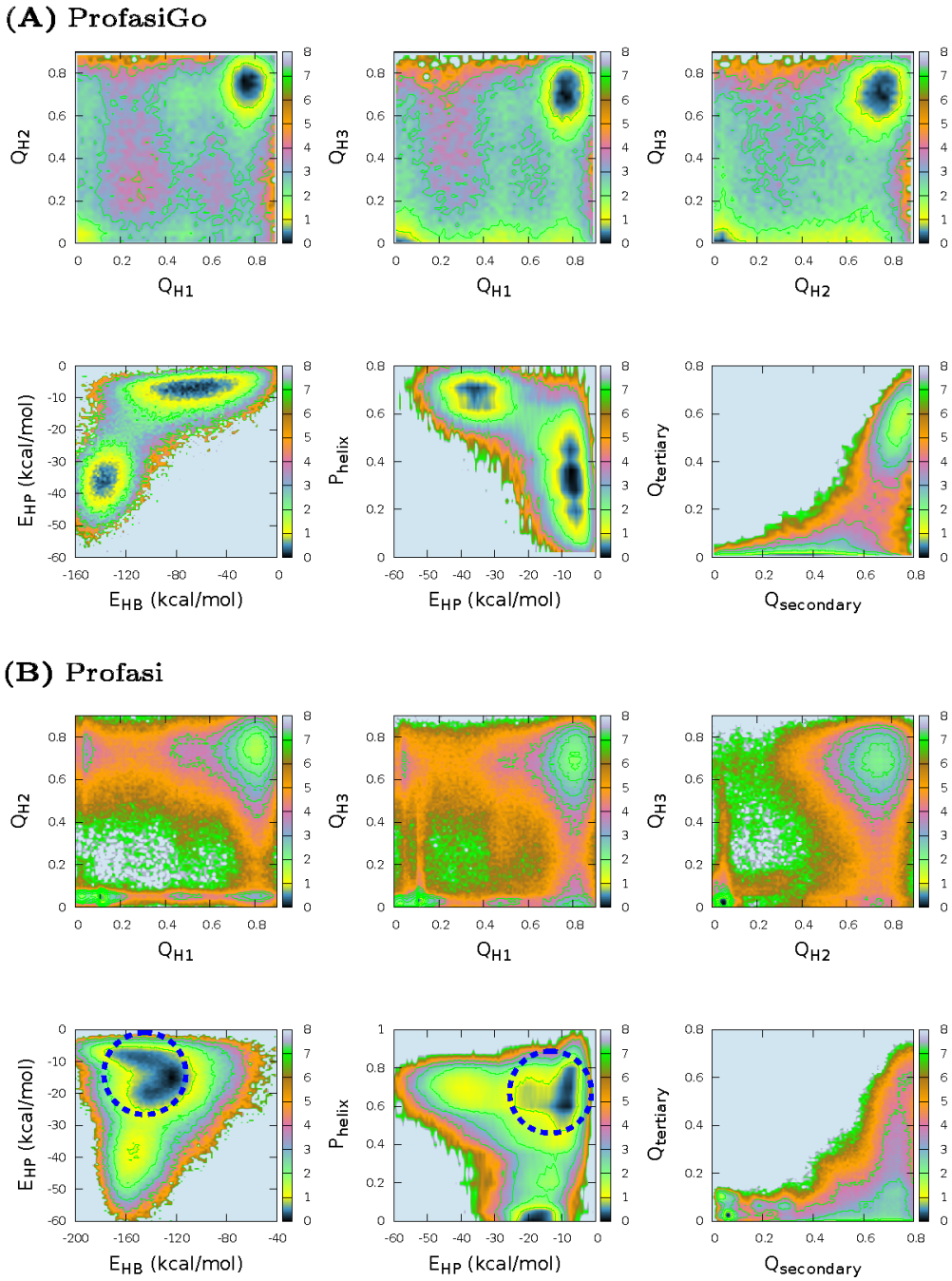


Figure 2: **Similar folding mechanisms for α3W in the Profasi and ProfasiGo models.** (A) Free energy surfaces from MC simulations with the ProfasiGo model with ϵ_{Go}=0.2 and reweighted to 346K. The top three panels show F(Q_{H1},Q_{H2}), F(Q_{H1},Q_{H3}), F(Q_{H2},Q_{H3}), where Q_{H1}, Q_{H2} and Q_{H3} are the fraction of native intra-helical contacts. The bottom three panels show F(E_{HB},E_{HP}), F(E_{HP},P_{helix}) and F($Q_{\mathrm{secondary}}$,Q_{tertiary}). Here E_{HB} and E_{HP} are the backbone hydrogen bond energy and hydrophobic energy, respectively, while P_{helix} is fraction of helix formed. Low free-energy pathways are labeled by white arrows. (B) Same plots as in A, but from MC simulations by the Profasi model and analyzed at T=303K. Possible misfolded states are highlighted by blue dashed circles. All free energies are in units of kcal/mol.

all-atom Gō model generated by SMOG server with default parameters. By performing MD simulations on different temperatures ranging from 100 to 130K, which cover the expected folding temperature for normal proteins,[35] we found the folding temperature for α3W in SMOG Gō model to be around 109K. We then projected the MD trajectories onto the same folding order parameters as we used in the ProfasiGo model (Fig. S2). Unexpectedly, the results are rather different, and indicate a more strongly coupled folding process for all three helices. Thus the results from the pure Gō model suggest a nucleation condensation mechanism, in contrast to the diffusion collision mechanism revealed by both the Profasi and ProfasiGo models. Without more detailed experimental data available for the folding of α3W it is difficult to know which model is more realistic, but our hypothesis is that the combination of the physical and Gō model in principle provides access to more complex and varied mechanisms.

Distinguishing between folding mechanisms of α3W and α3D

Next, we studied the folding mechanism of α3D, which has similar topology but different handedness of the packing the three helices. We carried out MC simulations of α3D with the same strategy as for α3W, and compared the resulting free energy surfaces of the two models. The results suggest significant difference in the free energy surfaces between α3W and α3D (Fig. 3). Not only is the folding pathway of the three helices in α3D distinct from α3W, but the order of the formation of secondary and tertiary structure is also remarkably different. For α3W, the results suggest that the secondary structure forms before the hydrophobic core, while for α3D, the formation of secondary structure and the hydrophobic core are strongly coupled. Therefore, the folding mechanisms of α3W and α3D represent two different models: diffusion collision and nucleation condensation, respectively. This conclusion is consistent with recent work carried out by Shao with integrated-tempering-sampling MD simulations.[101]

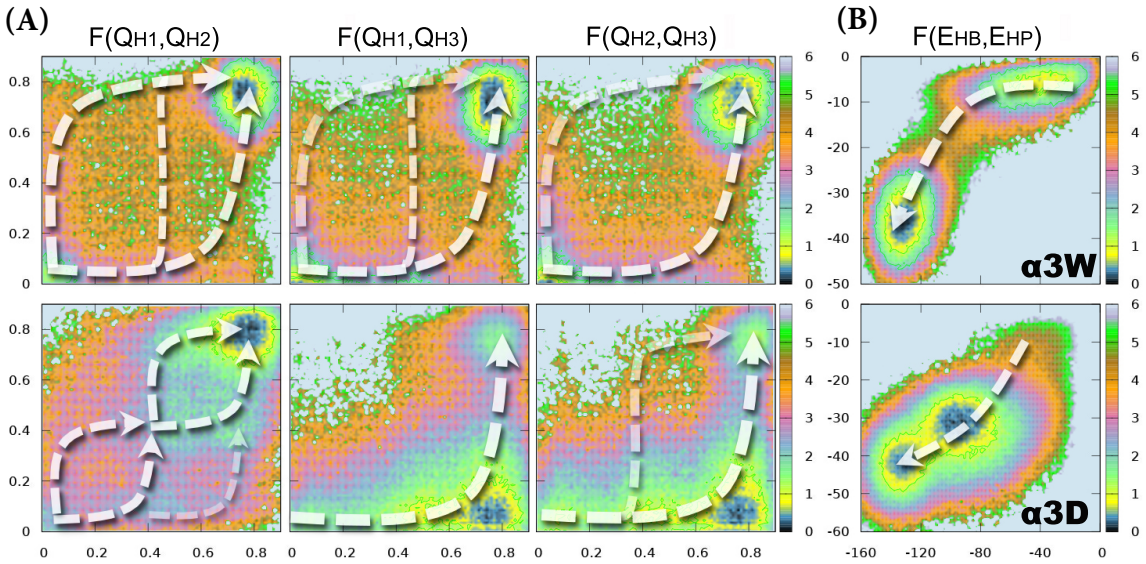


Figure 3: **Distinct folding mechanism of α3W and α3D suggested by the ProfasiGo model.** (A) F(Q_{H1}, Q_{H2}), F(Q_{H1}, Q_{H3}) and F(Q_{H1}, Q_{H2}). (B) F(E_{HB}, E_{HP}). The results for α3W and α3D are shown in the first and second row, respectively. Their free energy landscapes were sampled by MUNINN with ϵ_{Go}=0.3, and reweighted at T_f=373K for α3W and T_f=339K for α3D, respectively. Low free-energy pathways are labeled by white arrows. All free energies are in units of kcal/mol.

Distinguishing between folding mechanisms of EnHD and UVF

After demonstrating that the ProfasiGo model can distinguish the folding mechanism of a pair of proteins with similar topology but different handedness of the packing, we proceeded to a more challenging case: to capture the differences in the folding mechanism of a pair of proteins with the same topology; such differences are generally difficult to capture within a pure Gō model.[30]

We chose the engrailed homeodomain (EnHD) and its thermostabilized variant (UVF)[97] as our target systems (Fig. 1C-D and Table 1), and compared the global free energy landscape by projecting the conformational space onto a few global order parameters, including E_{HB} (backbone hydrogen bond energy) and E_{HP} (hydrophobic energy). The free energy surfaces are quite different between EnHD and UVF (Fig. 4), suggesting the presence of folding intermediate states, which previously have been proposed by both simulation and experimental studies.[106,107] The two-dimensional free energy surfaces of F(E_{HB},E_{HP}) suggest that EnHD has a tendency to form secondary structure before the formation of hydrophobic core, while UVF tends to form the hydrophobic core coupled with the formation of secondary structures. In addition, the conformational distribution in the free energy surfaces suggests that UVF can sample conformational regions with lower hydrophobic energy. This may be explained by the fact that UVF has higher percentage of hydrophobic residues.[101] In any case, our results suggest that the global folding mechanism of the two proteins can be distinguished by the ProfasiGo model despite the fact that they share almost exactly the same topology.

Effects of the strength and shape of the Gō potential

The strength of the Gō potential relative to the physical potential, determined by ϵ_{Go}, is a free parameter in the ProfasiGo model. To assess how sensitive the results are to the choice of this value, we performed MC simulations of α3W with the same generalized ensemble method but different values for ϵ_{Go}. The resulting free energy landscapes at their corresponding T_f

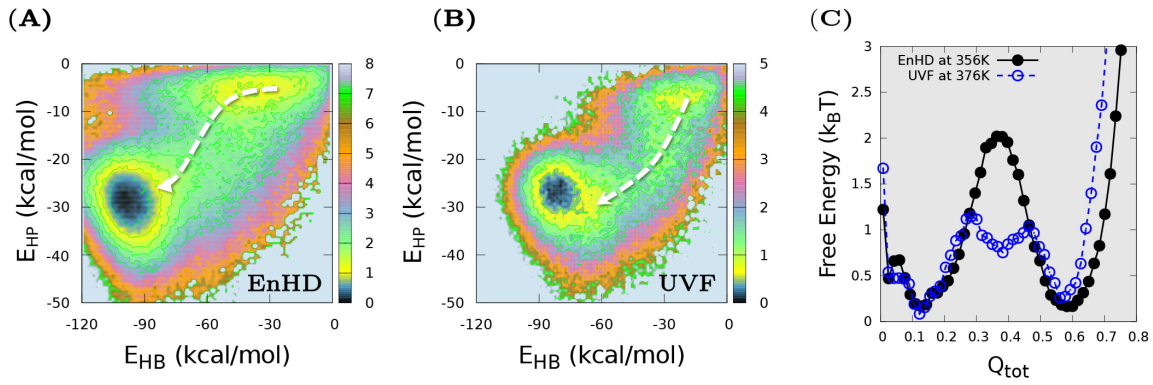
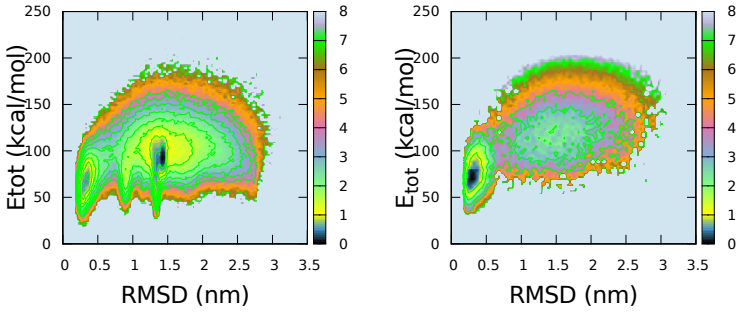


Figure 4: **The ProfasiGo model suggests distinct free energy profiles for two proteins with the same topology (EnHD and UVF).** (A-B) F(E_{HB},E_{HP}) for EnHD and UVF, respectively. (C) F(Q_{tot}) for EnHD and UVF. The results are from multicanonical MC simulations with the ProfasiGo model and reweighted to their corresponding folding temperatures (356 K and 376 K for EnHD and UVF, respectively). Low free-energy pathways are labeled by white arrows.

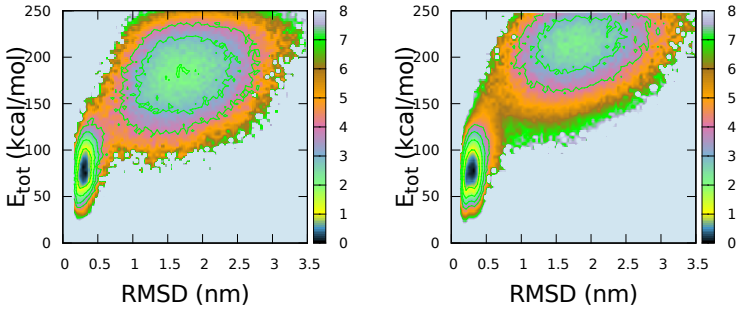
show very similar free energy surfaces for different Gō strengths, indicating the same folding mechanism. Thus, our results suggest the folding mechanism predicted by the ProfasiGo model is quite robust to the variety of Gō strength with this range. This conclusion is also supported by comparison of the free energy surfaces projected onto other order parameters, and by the corresponding simulations on UVF (Fig. S3). Unsurprisingly, we find that the free energy surfaces with different ϵ_{Go}, e.g. as a function of total potential energy (E_{tot}) and RMSD, show that the energy landscapes become more 'funnelled' as the strength of Gō potential increases (Fig. 5 and Fig. S3).

We also tested two types of Gō potential (12-10 and 12-10-6 forms as described in the Methods section), again using α3W as the test case, and found that the folding mechanism predicted by the ProfasiGo model is not sensitive to the shape of Gō potential, though we did find a small increase of the free energy barrier (Fig. S4) when using the 12-10-6 form. Previous simulations based on the coarse-grained Gō models have shown that the introduction of the desolvation barrier can help rationalise the diversity in the protein folding rates as well as the experimentally observed folding cooperativity.[88,89,108] Our observation that the mechanism is less sensitive to the choice of the functional form in the context of



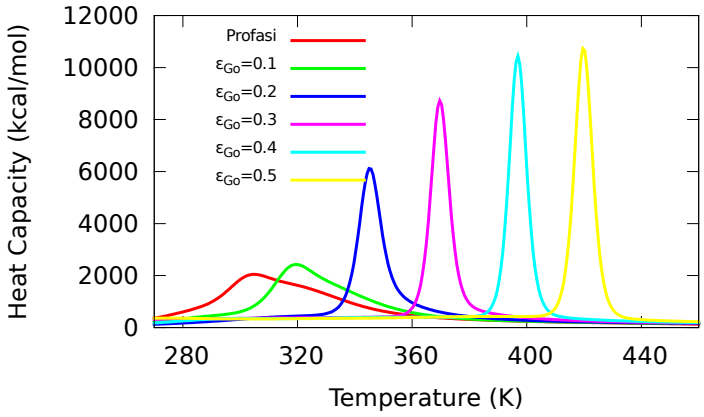
(**A**) Profasi

(**B**) ProfasiGo,ϵ_{Go}=0.1



(**C**) ProfasiGo,ϵ_{Go}=0.2

(**D**) ProfasiGo,ϵ_{Go}=0.3



(**E**) Heat Capacity

Figure 5: **The Gō potential makes the energy landscape more funnelled.** (A) Profasi model of α3W (equal to a ProfasiGo model with $\epsilon_{\mathrm{Go}} = 0.0$) at $T_f = 303K$; (B) ProfasiGo model with $\epsilon_{\mathrm{Go}} = 0.1$ at $T_f = 315K$; (C) ProfasiGo model with $\epsilon_{\mathrm{Go}} = 0.2$ at $T_f = 346K$; (D) ProfasiGo model with $\epsilon_{\mathrm{Go}} = 0.3$ at $T_f = 370K$. (E) The curves of heat capacity in ProfasiGo models with variant ϵ_{Go} ranging from 0.0 to 0.5.

the ProfasiGo model implies that the non-native interactions are fairly well captured by the physics-based term, thus alleviating this responsibility from the structure-based term.

Comparison of ProfasiGo, Gō and all-atom MD simulations

The availability of long time-scale, unbiased MD simulations of both α3D and UVF performed with ANTON[5] allowed us to conduct a final experiment, comparing the free energy landscapes and protein folding mechanisms obtained by different models, spanning from a pure Gō model (SMOG), the hybrid ProfasiGo model, to an explicit solvent, all-atom force field (CHARMM22*[109] with TIP3P water[110]).

To examine the folding mechanisms obtained from different models, we projected the folding trajectories of α3D and UVF onto the two-dimensional free energy surfaces arising as combinations of Q_{H1}, Q_{H2} and Q_{H3} (Fig. S5 and Fig. S6). The results suggest that the three helices fold independently in the ProfasiGo model, while they are strongly coupled in the pure Gō model. The folding of the three helices is more complex in CHARMM22*, but is consistent with ProfasiGo in the sense that it also finds the helices to form independently.

To examine the global free energy landscape, we further projected the folding trajectories onto two-dimensional free energy surfaces of $\mathrm{F}(Q_{\mathrm{secondary}}, Q_{\mathrm{tertiary}})$ for α3D (Fig. 6) and UVF (Fig. S7). We find that the free energy landscape from the ProfasiGo model is more similar to those of CHARMM22* than those from the pure Gō model.

Conclusions

Here we introduce a model that serves as a hybrid between an atomistic physics-based potential and a residue-level structure-based (Gō) model, in the context of a Monte Carlo simulation framework. We demonstrate that our model has the ability to successfully capture the protein folding mechanism at a level similar to a pure physics-based model. The model provides features not available with traditional structure-based approaches; for example,

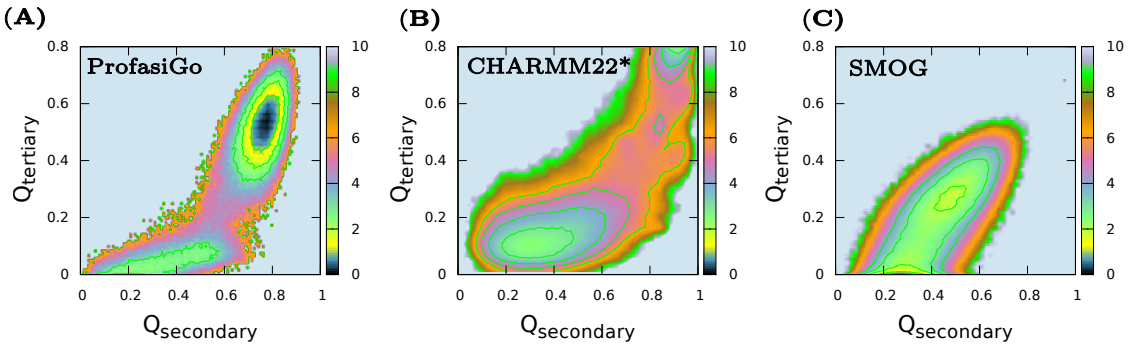


Figure 6: **Comparison of the global free energy landscapes of α3D obtained from a pure Gō model, the hybrid ProfasiGo model and an explicit solvent force field.** The results show the free energy surfaces of α3D as a function of $Q_{\text{secondary}}$ (the fraction of native contacts within the three helices) and Q_{tertiary} (the fraction of native contacts between the three helices) obtained from the ProfasiGo model with $\epsilon_{\text{Go}} = 0.5$ (A), SMOG (B) and CHARMM22* (C). All free energies are in units of kcal/mol.

it is capable of distinguishing between different folding pathways for topologically similar proteins. Finally, the procedure is complementary to physics based models in cases where these fail to fold to the native state (or do so excessively slowly).

Our procedure has a some limitations. Like for most force fields, the experimental folding temperature cannot be perfectly reproduced, and the folding temperature must therefore be located by scanning a range of temperatures. Secondly, the fact that we have chosen Monte Carlo as the basis for our approach makes it difficult to obtain realistic kinetic information. This could potentially be be mitigated by careful selection of moves, or restricting the analysis to longer time-scales.[111,112]

The Monte Carlo approach, however, provides substantial benefits in terms of computational efficiency. Our procedure requires only a few weeks of computation on a single CPU to obtain converged simulations on modest size proteins (with 40-80 residues), a dramatic improvement over comparable explicit solvent force field simulations. This makes it attractive for rapidly probing structure-mechanism relationships, taking input either directly from native structures or from indirect structural information derived from e.g. NMR spectroscopy or co-evolutionary analysis.[77,113] The ProfasiGo model may also serve as an efficient

atomistic model for sampling conformational space of large proteins, which can be refined a posteriori by reweighting with available experimental data.[114,115] Indeed, while we have here used the ProfasiGo model in the context of protein folding, we also expect it to be useful in providing access to conformational dynamics within folded states.

Overall, our results suggest that the ProfasiGo model can serve as a useful middle ground that combines the simplicity and efficiency of the Gō-type models and the accuracy and high computational cost of explicit solvent MD simulations.

Acknowledgement

K.L.-L. acknowledges funding by a Hallas-Møller Stipend from the Novo Nordisk Foundation and the BRAINSTRUC initiative from the Lundbeck Foundation. W.B. acknowledges funding from VILLUMFONDEN.

Supporting Information Available

Figure S1, Two popular functional forms of Gō potentials. Figure S2, A standard Gō model suggests a different folding mechanism of α3W compared to the Profasi and ProfasiGo models. Figure S3, Evidence that the Gō potential can funnel the free energy landscape by reducing the population of misfolded states of UVF. Figure S4, Predicted folding mechanism by the ProfasiGo model is not sensitive to the mathematical form of the Go potential in the case of α3W. Figure S5, Comparison of the folding mechanisms of α3D obtained from a pure Gō model, a hybrid ProfasiGo model and an explicit solvent force field. Figure S6, Comparison of the folding mechanism of UVF obtained from a pure Gō model, a hybrid ProfasiGo model and an explicit solvent force field. Figure S7, Comparison of the global free energy landscapes of UVF obtained from a pure Gō model, a hybrid ProfasiGo model and an explicit solvent force field. This material is available free of charge via the Internet at http://pubs.acs.org.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Anfinsen, C. B. *Science* **1973**, *181*, 223–230.

(2) Dunker, A. et al. *J Mol Graph Model* **2001**, *19*, 26 – 59.

(3) Fersht, A. R.; Daggett, V. *Cell* **2002**, *108*, 573–582.

(4) Dill, K. A.; MacCallum, J. L. *Science* **2012**, *338*, 1042–1046.

(5) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517–520.

(6) Gelman, H.; Gruebele, M. *Q Rev Biophys* **2014**, *47*, 95–142.

(7) Lapidus, L. J.; Acharya, S.; Schwantes, C. R.; Wu, L.; Shukla, D.; King, M.; DeCamp, S. J.; Pande, V. S. *Biophys J* **2014**, *107*, 947–955.

(8) Sborgi, L.; Verma, A.; Piana, S.; Lindorff-Larsen, K.; Cerminara, M.; Santiveri, C. M.; Shaw, D. E.; de Alba, E.; Muñoz, V. *J Am Chem Soc* **2015**, *137*, 6506–6516.

(9) Eaton, W. A.; Wolynes, P. G. *P Natl Acad Sci* **2017**, *114*, E9759–E9760.

(10) Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; Arrowsmith, C. H.; Baker, D. *Science* **2017**, *357*, 168–175.

(11) Johansson, K. E.; Lindorff-Larsen, K. *Curr Opin Struct Biol* **2018**, *48*, 157–163.

(12) Dobson, C. M. Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol. 2004; pp 3–16.

(13) Nielsen, S. V.; Stein, A.; Dinitzen, A. B.; Papaleo, E.; Tatham, M. H.; Poulsen, E. G.; Kassem, M. M.; Rasmussen, L. J.; Lindorff-Larsen, K.; Hartmann-Petersen, R. *PLoS Genet* **2017**, *13*, e1006739.

(14) Tozzini, V. *Curr Opin Struct Biol* **2005**, *15*, 144–150.

(15) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

(16) Saunders, M. G.; Voth, G. A. *Annu Rev Biophys* **2013**, *42*, 73–93.

(17) Eaton, W. A.; Muñoz, V. *Bioinformatics* **2014**, *22*.

(18) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. *Chem Rev* **2016**, *116*, 7898–7936.

(19) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr Opin Struct Biol* **2009**, *19*, 120–127.

(20) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *P Natl Acad Sci* **2013**, *110*, 5915–5920.

(21) Dror, R. O.; Young, C.; Shaw, D. E. *Encyclopedia Parallel Comput*; Springer, 2011; pp 60–71.

(22) Piana, S.; Klepeis, J. L.; Shaw, D. E. *Curr Opin Struct Biol* **2014**, *24*, 98–105.

(23) Husic, B. E.; Pande, V. S. *J Am Chem Soc* **2018**, *140*, 2386–2396.

(24) Valsson, O.; Tiwary, P.; Parrinello, M. *Annu Rev Phys Chem* **2016**, *67*, 159–184.

(25) Go, N. *Annu Rev Biophys and Bioeng* **1983**, *12*, 183–210.

(26) Rao, V. H. G.; Gosavi, S. *P Natl Acad Sci* **2018**, 201708173.

(27) Tian, P.; Best, R. B. *PLoS Comput Biol* **2016**, *12*, e1004933.

(28) Wang, Y.; Chu, X.; Wang, J. *Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods* **2014**, 257.

(29) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J Mol Biol* **2000**, *298*, 937–953.

(30) Karanicolas, J.; Brooks, C. L. *Protein Sci* **2002**, *11*, 2351–2361.

(31) Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. *Q Rev Biophys* **2014**, *47*, 285–363.

(32) Best, R. B.; Hummer, G.; Eaton, W. A. *Proc Natl Acad Sci USA* **2013**, *110*, 17874–17879.

(33) Levy, Y.; Wolynes, P. G.; Onuchic, J. N. *Proc Natl Acad Sci USA* **2004**, *101*, 511–516.

(34) Wang, J.; Oliveira, R. J.; Chu, X.; Whitford, P. C.; Chahine, J.; Han, W.; Wang, E.; Onuchic, J. N.; Leite, V. B. *Proc Natl Acad Sci USA* **2012**, *109*, 15763–15768.

(35) Noel, J. K.; Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. *Nucleic Acids Res* **2010**, *38*, W657–W661.

(36) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. *J Phys Chem B* **2012**, *116*, 8494–8503.

(37) Kenzaki, H.; Koga, N.; Hori, N.; Kanada, R.; Li, W.; Okazaki, K.-i.; Yao, X.-Q.; Takada, S. *J Chem Theo Compt* **2011**, *7*, 1979–1989.

(38) Feig, M.; Karanicolas, J.; Brooks III, C. L. *J Mol Graph Model* **2004**, *22*, 377–395.

(39) Pickard IV, F. C.; Miller, B. T.; Schalk, V.; Lerner, M. G.; Woodcock III, H. L.; Brooks, B. R. *PLoS Comput Biol* **2014**, *10*, e1003738.

(40) Lutz, B.; Sinner, C.; Heuermann, G.; Verma, A.; Schug, A. *Bioinformatics* **2013**, btt478.

(41) Chen, K.; Eargle, J.; Lai, J.; Kim, H.; Abeysirigunawardena, S.; Mayerle, M.; Woodson, S.; Ha, T.; Luthey-Schulten, Z. *J Phys Chem B* **2012**, *116*, 6819–6831.

(42) Zhmurov, A.; Dima, R.; Kholodov, Y.; Barsegov, V. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2984–2999.

(43) Kobayashi, C.; Jung, J.; Matsunaga, Y.; Mori, T.; Ando, T.; Tamura, K.; Kamiya, M.; Sugita, Y. *J Comput Chem* **2017**, *38*, 2193–2206.

(44) Tiana, G.; Villa, F.; Zhan, Y.; Capelli, R.; Paissoni, C.; Sormanni, P.; Heard, E.; Giorgetti, L.; Meloni, R. *Comput Phys Commun* **2014**,

(45) Strunk, T.; Wolf, M.; Brieg, M.; Klenin, K.; Biewer, A.; Tristram, F.; Ernst, M.; Kleine, P. J.; Heilmann, N.; Kondov, I.; Wenzel, W. *J Comput Chem* **2012**, *33*, 2602–2613.

(46) Cho, J.-H.; Sato, S.; Raleigh, D. P. *J Mol Biol* **2004**, *338*, 827–837.

(47) Kristjansdottir, S.; Lindorff-Larsen, K.; Fieber, W.; Dobson, C. M.; Vendruscolo, M.; Poulsen, F. M. *J Mol Biol* **2005**, *347*, 1053–1062.

(48) Zarrine-Afsar, A.; Wallin, S.; Neculai, A. M.; Neudecker, P.; Howell, P. L.; Davidson, A. R.; Chan, H. S. *Proc Natl Acad Sci USA* **2008**, *105*, 9999–10004.

(49) Zheng, W.; Schafer, N. P.; Wolynes, P. G. *Proc Natl Acad Sci USA* **2013**, *110*, 1680–1685.

(50) Hu, J.; Chen, T.; Wang, M.; Chan, H. S.; Zhang, Z. *Phys Chem Chem Phys* **2017**, *19*, 13629–13639.

(51) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struct., Funct., Bioinf.* **1995**, *21*, 167–195.

(52) Ferreiro, D. U.; Hegler, J. A.; Komives, E. A.; Wolynes, P. G. *Proc Natl Acad Sci USA* **2007**, *104*, 19819–19824.

(53) Clementi, C.; Plotkin, S. S. *Protein Sci* **2004**, *13*, 1750–1766.

(54) Faisca, P. F.; Nunes, A.; Travasso, R. D.; Shakhnovich, E. I. *Protein Sci* **2010**, *19*, 2196–2209.

(55) Wang, J.; Wang, Y.; Chu, X.; Hagen, S. J.; Han, W.; Wang, E. *PLoS Comput Biol* **2011**, *7*, e1001118.

(56) Shental-Bechor, D.; Smith, M. T.; MacKenzie, D.; Broom, A.; Marcovitz, A.; Ghashut, F.; Go, C.; Bralha, F.; Meiering, E. M.; Levy, Y. *Proc Natl Acad Sci USA* **2012**, *109*, 17839–17844.

(57) Shi, J.; Nobrega, R. P.; Schwantes, C.; Kathuria, S. V.; Bilsel, O.; Matthews, C. R.; Lane, T.; Pande, V. S. *Sci Rep* **2017**, *7*, 44116.

(58) Yadahalli, S.; Hemanth Giri Rao, V.; Gosavi, S. *Isr J Chem* **2014**, *54*, 1230–1240.

(59) Azia, A.; Levy, Y. *J Mol Biol* **2009**, *393*, 527–542.

(60) Kim, Y. C.; Hummer, G. *J Mol Biol* **2008**, *375*, 1416–1433.

(61) Wang, Y.; Gan, L.; Wang, E.; Wang, J. *J Chem Theo Compt* **2012**, *9*, 84–95.

(62) Wang, Y.; Tang, C.; Wang, E.; Wang, J. *PLoS Comput Biol* **2014**, *10*, e1003691.

(63) Miyazawa, S.; Jernigan, R. L. *J Mol Biol* **1996**, *256*, 623–644.

(64) Wang, Y.; Chu, X.; Suo, Z.; Wang, E.; Wang, J. *J Am Chem Soc* **2012**, *134*, 13755–13764.

(65) Pogorelov, T. V.; Luthey-Schulten, Z. *Biophys J* **2004**, *87*, 207–214.

(66) Meinke, J. H.; Hansmann, U. H. *J Comput Chem* **2009**, *30*, 1642–1648.

(67) Wang, Y.; Chu, X.; Longhi, S.; Roche, P.; Han, W.; Wang, E.; Wang, J. *Proc Natl Acad Sci USA* **2013**, *110*, E3743–E3752.

(68) Sikosek, T.; Krobath, H.; Chan, H. S. *PLoS Comput Biol* **2016**, *12*, e1004960.

(69) Bernhardt, N. A.; Xi, W.; Wang, W.; Hansmann, U. H. *J Chem Theo Comput* **2016**, *12*, 5656–5666.

(70) Kar, P.; Feig, M. *J Chem Theo Comput* **2017**, *13*, 5753–5765.

(71) Poma, A. B.; Cieplak, M.; Theodorakis, P. E. *J Chem Theo Comput* **2017**, *13*, 1366–1374.

(72) Strunk, T.; Wolf, M.; Brieg, M.; Klenin, K.; Biewer, A.; Tristram, F.; Ernst, M.; Kleine, P. J.; Heilmann, N.; Kondov, I.; Wenzel, W. *J Comput Chem* **2012**, *33*, 2602–2613.

(73) Irbäck, A.; Mohanty, S. *J Comput Chem* **2006**, *27*, 1548–1555.

(74) Li, D.-W.; Mohanty, S.; Irbäck, A.; Huo, S. *PLoS Comput Biol* **2008**, *4*, e1000238.

(75) Irbäck, A.; Mitternacht, S.; Mohanty, S. *BMC Biophysics* **2009**, *2*, 2.

(76) Tian, P.; Lindorff-Larsen, K.; Boomsma, W.; Jensen, M. H.; Otzen, D. E. *PloS one* **2016**, *11*, e0146096.

(77) Tian, P.; Boomsma, W.; Wang, Y.; Otzen, D. E.; Jensen, M. H.; Lindorff-Larsen, K. *J Am Chem Soc* **2014**, *137*, 22–25.

(78) Kassem, M. M.; Wang, Y.; Boomsma, W.; Lindorff-Larsen, K. *Biophys J* **2016**, *110*, 2342–2348.

(79) Boomsma, W.; Tian, P.; Frellsen, J.; Ferkinghoff-Borg, J.; Hamelryck, T.; Lindorff-Larsen, K.; Vendruscolo, M. *Proc Natl Acad Sci USA* **2014**, *111*, 13852–13857.

(80) Valentin, J. B.; Andreetta, C.; Boomsma, W.; Bottaro, S.; Ferkinghoff-Borg, J.; Frellsen, J.; Mardia, K. V.; Tian, P.; Hamelryck, T. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 288–299.

(81) Noel, J. K.; Whitford, P. C.; Onuchic, J. N. *J Phys Chem B* **2012**, *116*, 8692–8702.

(82) Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M. *Bioinformatics* **1999**, *15*, 327–332.

(83) Sułkowska, J. I.; Cieplak, M. *Biophys J* **2008**, *95*, 3174–3191.

(84) Noel, J. K.; Onuchic, J. N. *Computational Modeling of Biological Systems*; Springer, 2012; pp 31–54.

(85) Wołek, K.; Gómez-Sicilia, À.; Cieplak, M. *J Chem Phys* **2015**, *143*, 243105.

(86) Cho, S. S.; Levy, Y.; Wolynes, P. G. *P Natl Acad Sci* **2009**, *106*, 434–439.

(87) Chng, C.-P.; Yang, L.-W. *Bioinform Biol Insights* **2008**, *2*, 171–185.

(88) Cheung, M. S.; García, A. E.; Onuchic, J. N. *Proc Natl Acad Sci USA* **2002**, *99*, 685–690.

(89) Kaya, H.; Chan, H. S. *J Mol Biol* **2003**, *326*, 911–931.

(90) Zhang, Z.; Chan, H. S. *Proc Natl Acad Sci USA* **2010**, *107*, 2920–2925.

(91) Ferkinghoff-Borg, J. *Eur Phys J B* **2002**, *29*, 481–484.

(92) Frellsen, J.; Winther, O.; Ghahramani, Z.; Ferkinghoff-Borg, J. Bayesian generalised ensemble Markov chain Monte Carlo. Artif Intell Stat. 2016; pp 408–416.

(93) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845–854.

(94) Cho, S. S.; Levy, Y.; Wolynes, P. G. *Proc Natl Acad Sci USA* **2006**, *103*, 586–591.

(95) Zhang, C.; Ma, J. *Proc Natl Acad Sci USA* **2012**, *109*, 8139–8144.

(96) Glyakina, A. V.; Pereyaslavets, L. B.; Galzitskaya, O. V. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 1527–1541.

(97) McCully, M. E.; Beck, D. A.; Daggett, V. *Protein Eng Des Sel* **2013**, *26*, 35–45.

(98) Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc Natl Acad Sci USA* **2005**, *102*, 2362–2367.

(99) Bereau, T.; Deserno, M. *J Chem Phys* **2009**, *130*, 235106.

(100) Kapoor, A.; Travesset, A. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 505–516.

(101) Shao, Q. *J Phys Chem B* **2014**, *118*, 5891–5900.

(102) Jiang, F.; Wu, Y.-D. *J Am Chem Soc* **2014**, *136*, 9536–9539.

(103) Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. *J Am Chem Soc* **2014**, *136*, 13959–13962.

(104) Tian, P.; Jónsson, S. Æ.; Ferkinghoff-Borg, J.; Krivov, S. V.; Lindorff-Larsen, K.; Irba ck, A.; Boomsma, W. *J Chem Theo Compt* **2014**, *10*, 543–553.

(105) Liu, S.-Q.; Ji, X.-L.; Tao, Y.; Tan, D.-Y.; Zhang, K.-Q.; Fu, Y.-X. *Protein Eng* **2012**, 207–252.

(106) DeMarco, M. L.; Alonso, D. O.; Daggett, V. *J Mol Biol* **2004**, *341*, 1109–1124.

(107) Huang, F.; Settanni, G.; Fersht, A. R. *Protein Eng Des Sel* **2008**, *21*, 131–146.

(108) Chen, T.; Chan, H. S. *Phys Chem Chem Phys* **2014**, *16*, 6460–6479.

(109) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys J* **2011**, *100*, L47–L49.

(110) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J Chem Phys* **1983**, *79*, 926–935.

(111) Shimada, J.; Kussell, E. L.; Shakhnovich, E. I. *J Mol Biol* **2001**, *308*, 79–95.

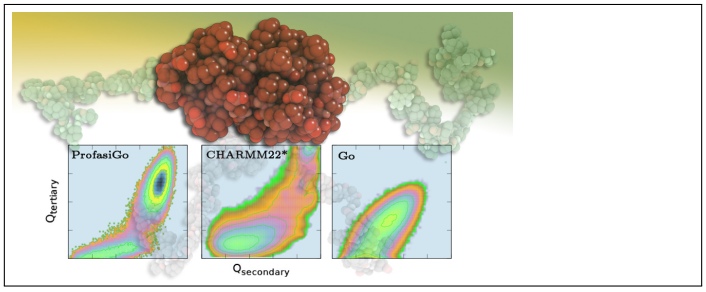
(112) Nilsson, D.; Mohanty, S.; Irbäck, A. *J Chem Phys* **2018**, *148*, 055101.

(113) Jana, B.; Morcos, F.; Onuchic, J. N. *Phys Chem Chem Phys* **2014**, *16*, 6496–6507.

(114) Bottaro, S.; Bussi, G.; Kennedy, S. D.; Turner, D. H.; Lindorff-Larsen, K. *Sci Adv* **2018**, *4*.

(115) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. *Curr Opin Struct Biol* **2017**, *42*, 106–116.

Graphical TOC Entry



Supporting Material for Monte Carlo Sampling of Protein Folding by Combining an All-Atom Physics-Based Model with a Native State Bias

Yong Wang,[†,§] Pengfei Tian,[†,‡,§] Wouter Boomsma,[¶] and Kresten Lindorff-Larsen[*,†]

Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5 DK-2200 Copenhagen N, Denmark, Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, United States, and Department of Computer Science, University of Copenhagen, 2100 Copenhagen Ø, Denmark

E-mail: lindorff@bio.ku.dk

[*]To whom correspondence should be addressed

[†]Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5 DK-2200 Copenhagen N, Denmark

[‡]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, United States

[¶]Department of Computer Science, University of Copenhagen, 2100 Copenhagen Ø, Denmark

[§]Contributed equally to this work

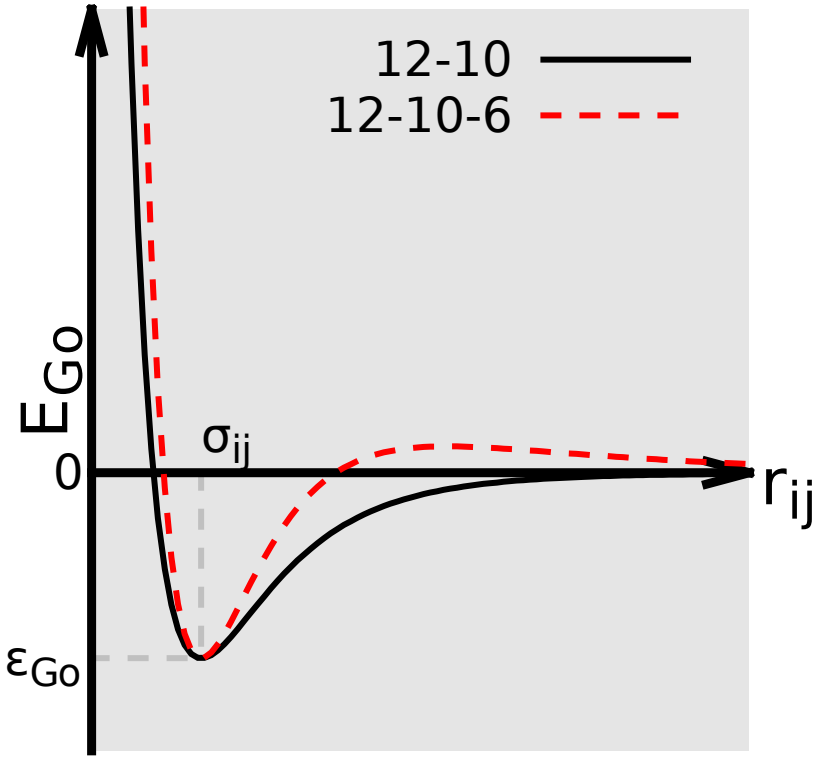


Figure S1: **Two popular functional forms of the Gō potential.** The two curves show the 12-10 Lennard-Jones-like potential, and the modified 12-10-6 potential, as black solid and red dashed lines, respectively. The 12-10-6 potential has a low energy barrier designed to mimic the desolvation effect.

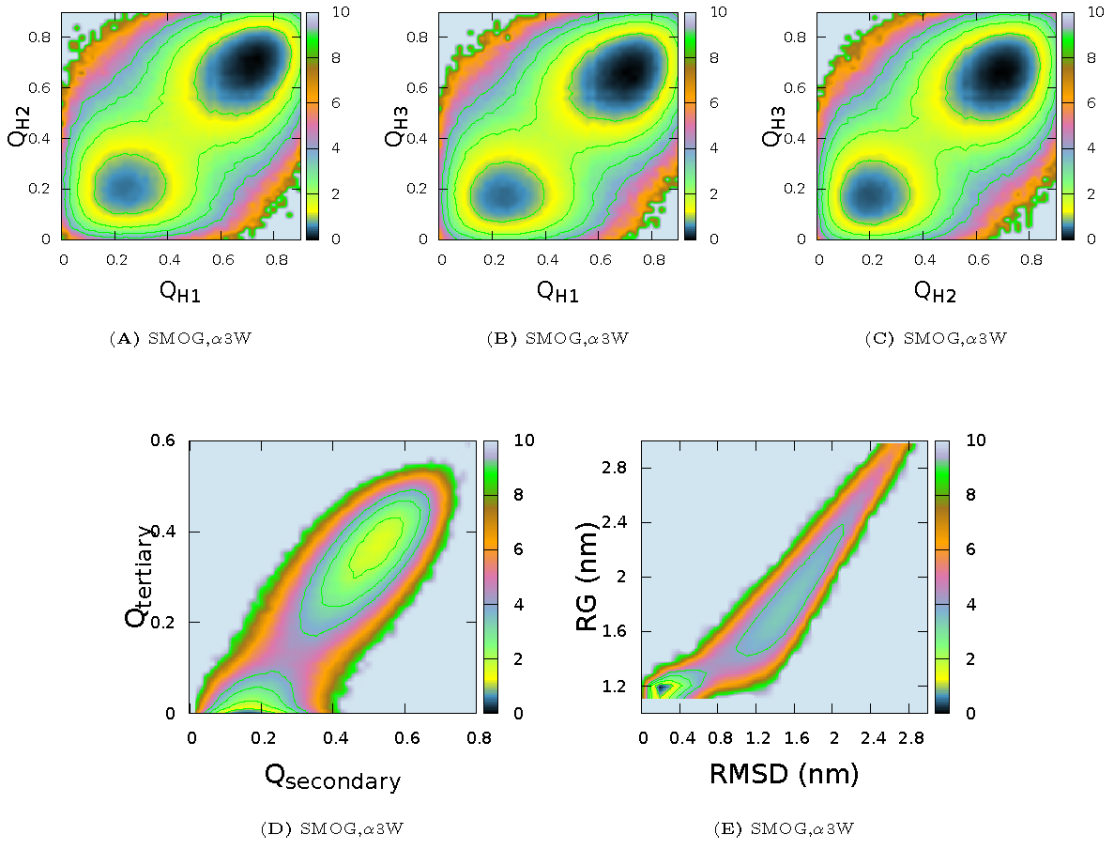


Figure S2: **Free energy landscape of α3W in the pure Gō model.** (A) F(Q_{H1},Q_{H2}). (B) F(Q_{H1},Q_{H3}). (C) F(Q_{H2},Q_{H3}). (D) F($Q_{\mathrm{secondary}}$,Q_{tertiary}). (E) F(RMSD,RG). Comparison with the Profasi and ProfasiGo models (Fig. 2 in the main text) reveals a substantially different mechanism.

Increasing the 'foldability' of a physics-based force field by adding a native bias

The free energy landscape, F(RMSD,$E_{\rm tot}$), in simulations of UVF with $\epsilon_{\rm Go}$=0.3 reveals a non-native state (RMSD=10.0) with a comparable free energy and internal energy as the native state (Fig. S3A). Note that these results were based on $\epsilon_{\rm Go}$=0.3 (Fig. S3A). In this case, increasing the native bias ($\epsilon_{\rm Go}$=0.4) substantially decreases the stability of this state (Fig. S3B). Similar results were obtained for α3W, when comparing the free energy landscape in the absence of the native bias with those of the ProfasiGo model (Fig. 5 in the main text).

There are two key ingredients that help determine the whether an energy landscape is 'well funnelled': the energy gap between the native and nonnative state and the energetic fluctuations in the non-native states.[1] The former determines the steepness of the energy funnel, while the later controls the roughness of the energy funnel. Maximization of their ratio has been used to guide the optimization of the force field parameters by following the minimal frustration principle and maximize the 'funnelledness' of the protein energy landscape.[2] Here, we provided evidence that the introduction of native structure-based information into a physics-based force field can funnel the energy landscape not only by increasing the steepness (native state becomes more energetically favourable), but also by decreasing the roughness (the energy fluctuation or the width of the energy distribution of the non-native states becomes more narrow).

The results in Fig. S3 also suggests that the pure Profasi model (equivalent to ProfasiGo with $\epsilon_{\rm Go}$=0) will not have the correct native state of UVF as its free energy minimum, since the non-native state becomes progressively more populated as the native bias is decreased. We thus suggest that analyses such as these could be useful to identify force field problems in cases where sampling the folding landscape is prohibitively difficult.

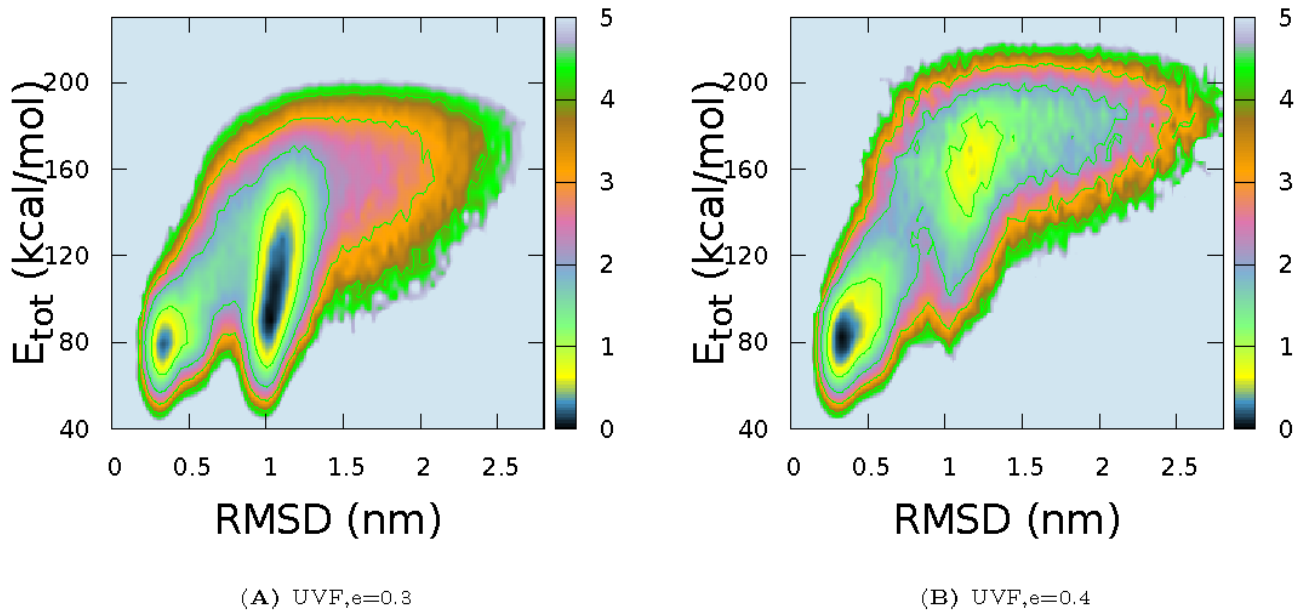


Figure S3: **Evidence that Go potential can funnel the free energy landscape by reducing the population of misfolded states in the case of UVF.** (A) Two-dimensional free energy surfaces as a function of RMSD and $E_{\rm tot}$ at $\epsilon_{\rm Go}$=0.3. (B) Two-dimensional free energy surfaces as a function of RMSD and $E_{\rm tot}$ at $\epsilon_{\rm Go}$=0.4. Note that the results were from multicanonical MC simulations by ProfasiGo model of UVF and reweighted to corresponding folding temperature.

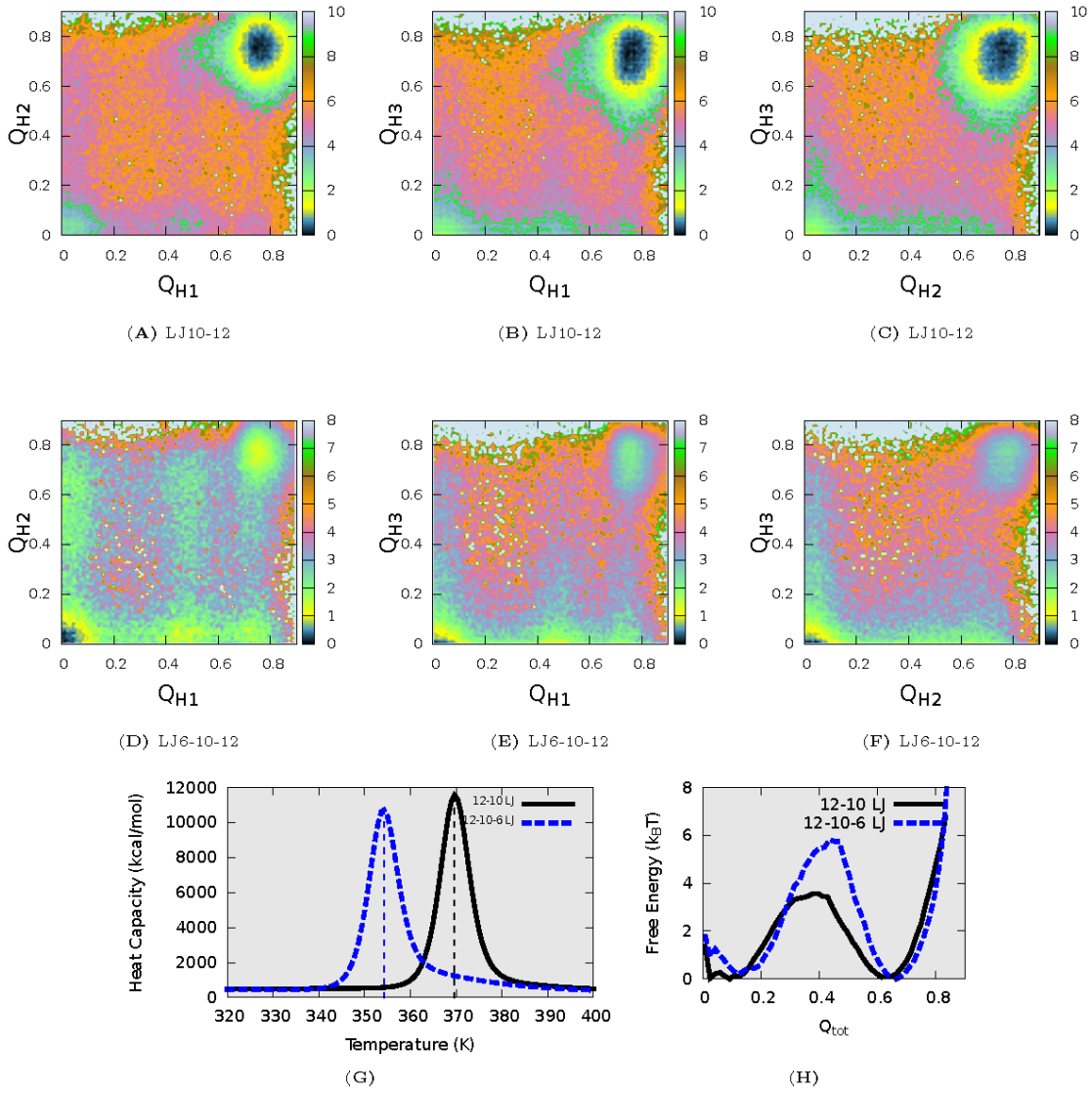


Figure S4: **The free energy landscape in the hybrid ProfasiGo is not sensitive to the mathematical form of the Gō potential.** (A–F) Free energy landscape of α3W using the 12-10 potential (A–D) or 12-10-6 potential. (A–C) F(Q_{H1}, Q_{H2}), F(Q_{H1}, Q_{H3}), F(Q_{H1}, Q_{H2}) and F(Q_{all}) with the 12-10 potential. (D–F) F(Q_{H1}, Q_{H2}), F(Q_{H1}, Q_{H3}), F(Q_{H1}, Q_{H2}) and F(Q_{all}) for 12-10-6 potential. (G) The heat capacity curves from the ProfasiGo models with the two potentials. (H) The free energy profiles of F(Q_{tot}) at corresponding T_f. The results are from multicanonical MC simulations with the ProfasiGo model with ϵ_{Go}=0.3. For ProfasiGo model with 12-10-6 potential we found $T_f^{12\text{-}10\text{-}6}$=354K, while for ProfasiGo model with 12-10 potential, $T_f^{12\text{-}10}$=370K.

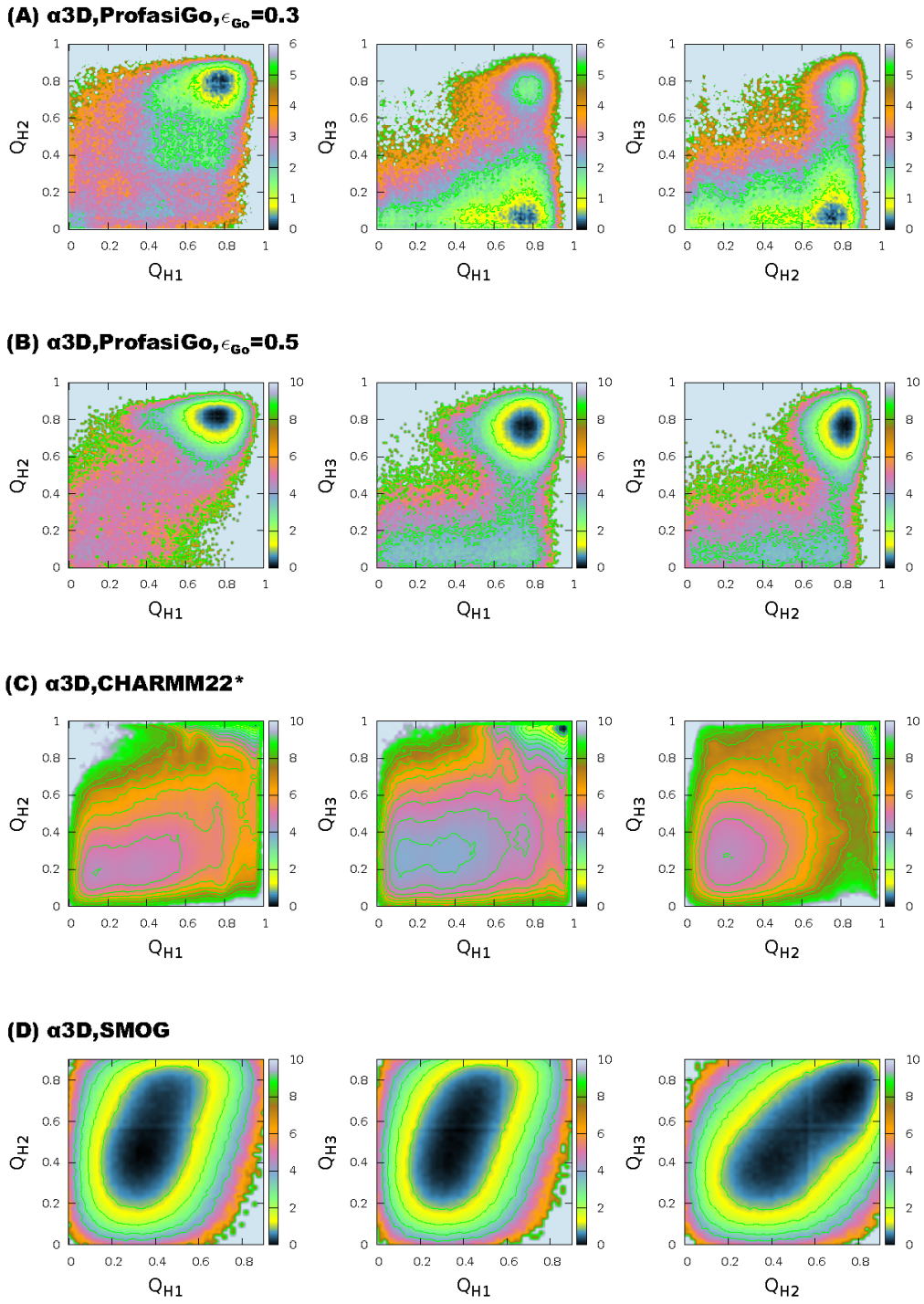


Figure S5: **Comparison of the free energy landscapes of α3D obtained from the pure Gō model, the hybrid ProfasiGo model and an explicit solvent force field.** (A) Free energy surfaces of α3D as a function of Q_{H1}, Q_{H2} and Q_{H3} obtained from the ProfasiGo model with ϵ_{Go}=0.3 at T_f=339K. (B) Free energy surfaces of α3D obtained from the ProfasiGo model with ϵ_{Go}=0.5 at T_f=373K. (C)The same free energy surfaces of α3D obtained from a previously published 707μs long MD simulation with CHARMM22* at T=370K. (D) The same free energy surfaces of α3D obtained from MD simulations with a pure Gō model (SMOG) at T_f=124K.

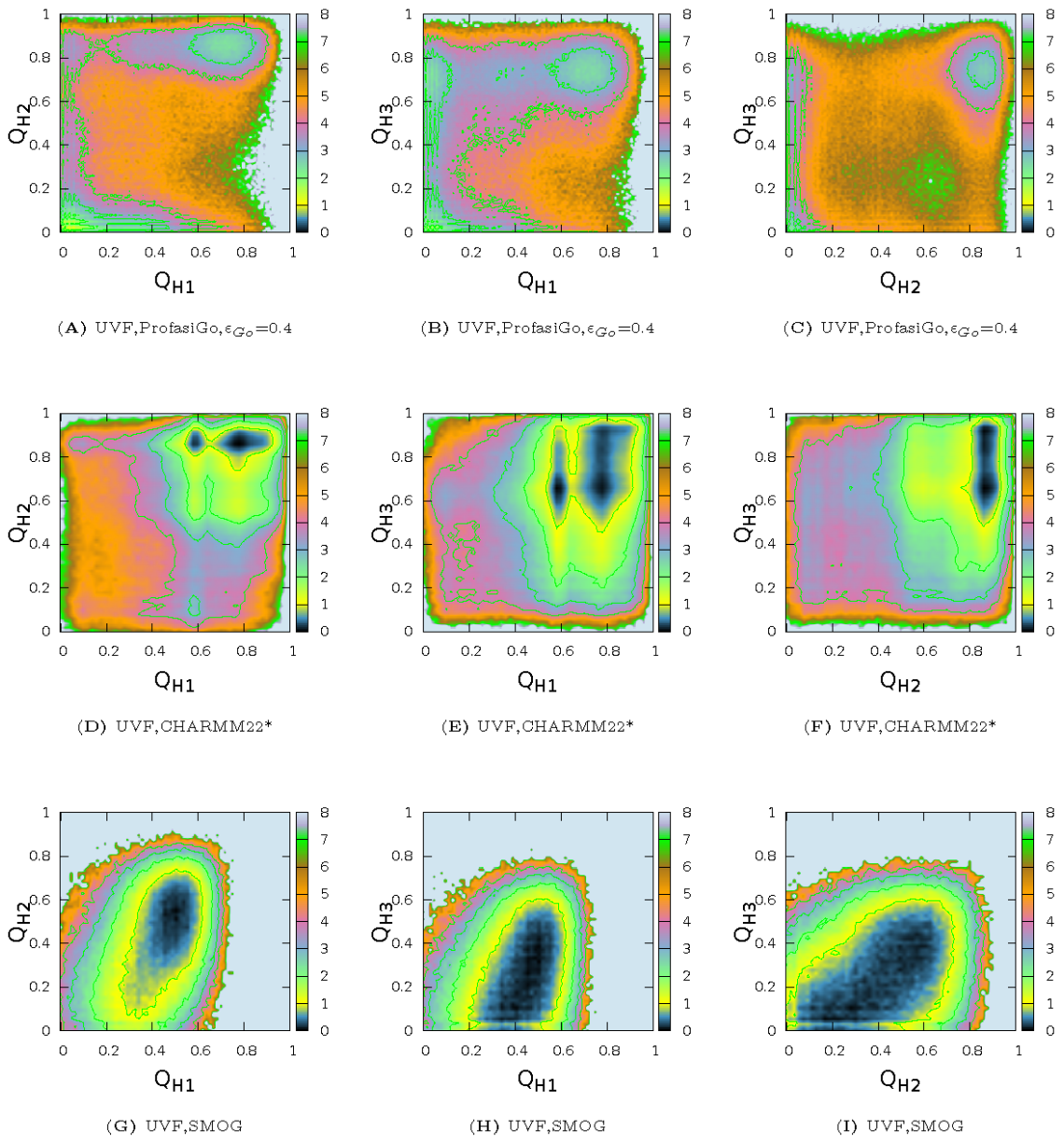
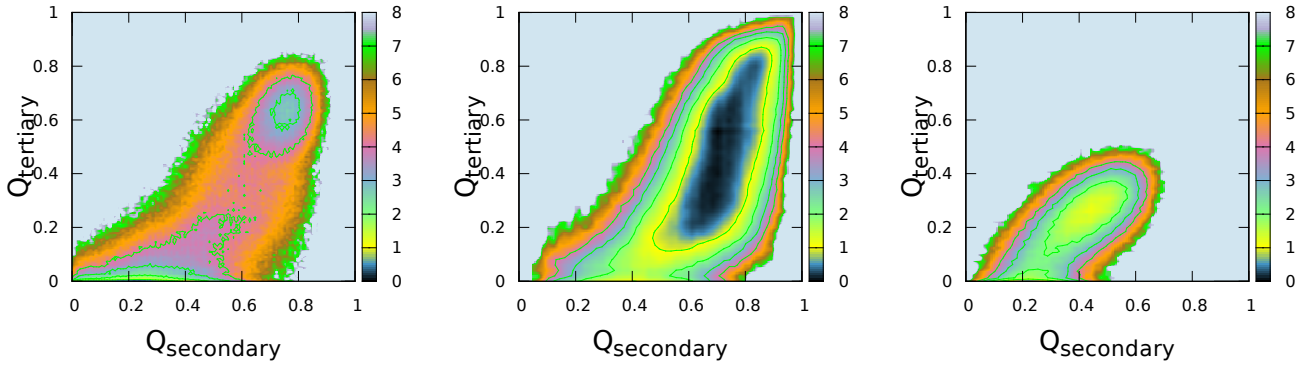


Figure S6: **Comparison of the free energy landscapes of UVF obtained from the pure Gō model, the hybrid ProfasiGo model and an explicit solvent force field.** (A–C) Free energy surfaces obtained using the ProfasiGo model with ϵ_{Go}=0.4 at T_f. (D–F) Free energy surfaces obtained using all-atom MD with the CHARMM22* force field at 360 K, somewhat below the melting temperature in this force field (390 K). (G–I) Free energy surfaces obtained using the the pure Gō model (SMOG) at 124 K.



(**A**) UVF,ϵ_{Go}=0.4 (**B**) UVF,CHARMM22* (**C**) UVF,SMOG,124K

Figure S7: **Comparison of the free energy landscapes of UVF obtained from the pure Gō model, the hybrid ProfasiGo model and an explicit solvent force field.** The figure shows the free energy surfaces of UVF as a function of $Q_{\text{secondary}}$ (the fraction of native contacts within the three helices) and Q_{tertiary} (the fraction of native contacts between the three helices) obtained from (A) the ProfasiGo model with $\epsilon_{\text{Go}} = 0.4$, (B) CHARMM22* with TIP3P water, and (C) the pure Gō model (SMOG).

References

(1) Wang, J.; Oliveira, R. J.; Chu, X.; Whitford, P. C.; Chahine, J.; Han, W.; Wang, E.; Onuchic, J. N.; Leite, V. B. *Proc Natl Acad Sci USA* **2012**, *109*, 15763–15768.

(2) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. *J Phys Chem B* **2012**, *116*, 8494–8503.