THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Multibody Terms in Protein Coarse-Grained Models: A Top-Down Perspective

Published as part of The Journal of Physical Chemistry virtual special issue "Jose Onuchic Festschrift".

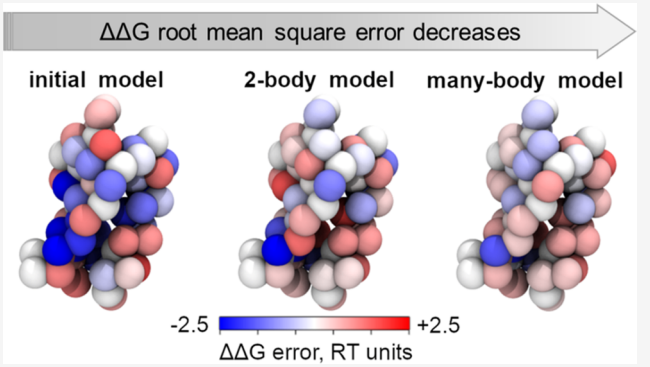
Iryna Zaporozhets and Cecilia Clementi*

Cite This: *J. Phys. Chem. B* 2023, 127, 6920–6927

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Coarse-grained models allow computational investigation of biomolecular processes occurring on long time and length scales, intractable with atomistic simulation. Traditionally, many coarse-grained models rely mostly on pairwise interaction potentials. However, the decimation of degrees of freedom should, in principle, lead to a complex many-body effective interaction potential. In this work, we use experimental data on mutant stability to parametrize coarse-grained models for two proteins with and without many-body terms. We demonstrate that many-body terms are necessary to reproduce quantitatively the effects of point mutations on protein stability, particularly to implicitly take into account the effect of the solvent.



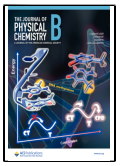
INTRODUCTION

Recent advances in software and hardware for molecular dynamics simulation and enhanced sampling techniques have significantly increased the time and length scale of biomolecular processes, which are amenable to computational investigation. Nevertheless, the characterization of many biomedically relevant molecular processes that take place on a time scale beyond milliseconds remains a daunting task, if not intractable, as computation involves time propagation of billions of degrees of freedom.

However, it is reasonable to ask whether all of the degrees of freedom are essential to elucidate slow processes in biomolecular dynamics or if some coarser representations can be used. Evidence exists that long-term conformational dynamics of a biomolecular system can be described by a small set of collective variables instead of the entire set of degrees of freedom.[1−4] Consequently, less informative degrees of freedom can be integrated out, and groups of atoms can be merged into effective beads, i.e., the system can be coarse grained. Indeed, coarse-grained models are popular tools to explore long time scale processes in biomolecular systems, well beyond what is possible with atomistic simulations.[5−9]

The design of a coarse-grained (CG) model involves two main steps. First, one must determine how the detailed, all-atom representation of the system is mapped into a set of effective interaction sites. Second, an effective interaction potential between these sites needs to be defined. Since the first protein CG model, introduced by Levitt and Warshel,[10] many protein coarse-grained models have been developed, and their design principles can broadly be categorized as "bottom up", "top down", or "knowledge based".[11] In bottom-up approaches, the coarse-grained potential of mean force (PMF) that governs the time evolution of the coarse-grained system is derived to match given properties of a reference all-atom model.[12−18] In principle, a bottom-up method provides a rigorous bridge between atomistic and coarse-grained force fields and demonstrates that integrating out degrees of freedom results in an increasingly complex multibody potential of mean force.[12] This complexity means that, in practice, a rigorous treatment is possible only for simple systems, while for practical applications, additional approximations should be made, such as a basis set for the approximation of the PMF.[11] Alternatively, top-down approaches do not rely on the underlying fine-grained representation. Instead, physicochemical considerations are used to define the interaction potential parametrized to match experimentally measured macro- or microscopic observables. Together with knowledge-based models that take advantage of the known structural information,[19,20] top-down approaches[21−23] constitute the majority of the protein coarse-grained models that are currently used in practice. Despite their simplicity, such

Received: July 4, 2023
Revised: July 12, 2023
Published: July 27, 2023



https://doi.org/10.1021/acs.jpcb.3c04493
J. Phys. Chem. B 2023, 127, 6920–6927

minimalistic models can provide valuable insights into biomolecular processes, such as protein folding and aggregation,[24−28] and can be used to understand the fundamental principles that govern them. However, one should be careful when interpreting results obtained with a coarse-grained model, as the effects of the approximations introduced are not always well understood.

One of the ways to mitigate these effects involves incorporating increasingly more experimental data in a newly designed or already existing model. Several approaches that solve this task have been suggested. A widely used approach involves minimization of information that should be added to achieve an agreement with the experimental data, i.e., maximize entropy.[29−31] This approach was used, for example, to parametrize a force field for intrinsically disordered proteins (IDPs) using the radii of gyration as experimental constraints.[32] Clementi and Matysiak developed another approach that allows refining a structure-based model by using measurements of free energy differences upon mutations (ΔΔG) as experimental data.[24] The suggested algorithm exploited the explicit dependence of ΔΔG on the model parameters and was successfully used to optimize a CG model for the investigation of the role of mutations in misfolding and aggregation of ribosomal protein S6.[25] Norgaard et al.[33] used a free energy perturbation method[34] to find optimal parameters on nonbonded interactions by maximizing the likelihood of reproducing experimental data given the model parameters. Chen et al.[35] extended this approach and suggested a framework for the observable-driven design of effective molecular models (ODEM) that allows the parametrization of models incorporating any type of thermodynamic average. A similar approach was recently used in combination with neural network potentials.[36]

Nevertheless, no amount of experimental data can "save" a model if the interaction potential used is a priori restricted to a functional form that cannot capture a process of interest, for example, multibody effects. Traditionally, CG force fields use the same functional form for the effective potential energy as in atomistic simulation, with nonbonded interactions represented only by pairwise terms. However, it is well known that CG force fields should contain many-body terms,[37−39] for instance, to adequately describe the structural properties of water[40] and the folding free energy landscape of small proteins.[41] At the same time, CG force fields for methanol[40] and biomolecular phase separation[42] demonstrate satisfactory performance with pairwise potentials only.

In this work, we use a top-down approach to demonstrate the effect of many-body interactions on a protein CG force field's ability to describe mutations' effects on protein stability. We optimize a structure-based coarse-grained force field for two proteins, ubiquitin and the B1 domain of protein G (GB1), using experimentally measured free energy differences upon mutations, ΔΔG, as reference observables. We show that while for ubiquitin the experimental data can be reproduced using only two-body nonbonded interactions, the full set of observables for protein G cannot be adequately reproduced without many-body terms. The addition of a many-body term to take into account solvation effects in the CG energy function allows us to satisfactorily recover the experimental data for both proteins.

METHODS

Optimization Procedure. Our goal is to optimize the parameters of a CG force field using experimental data as a reference. We use a previously proposed method, the observable-driven design of effective molecular models (ODEM). In the following, we briefly summarize the ODEM procedure's main steps and then discuss each one in more detail.

ODEM Algorithm. The ODEM framework is described in Figure 1.

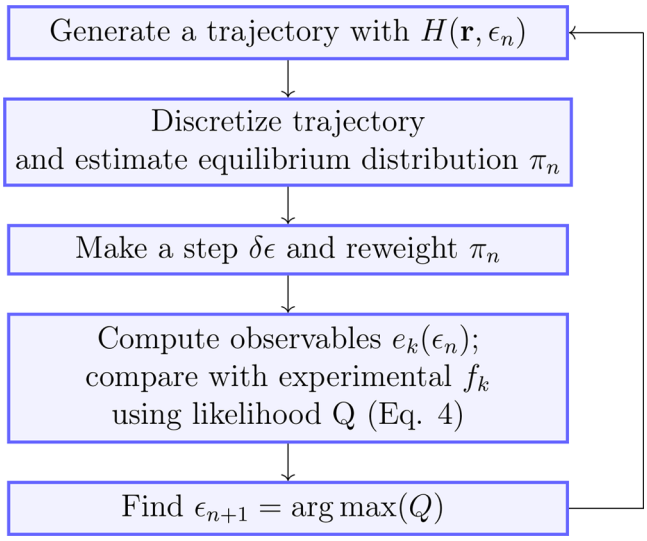


Figure 1. ODEM workflow at iteration n.

The method requires the initial definition of a trial coarse-grained Hamiltonian $H(\epsilon_0)$ with an initial set of parameters $\{\epsilon_0\}$. Molecular dynamics simulation can be cheaply performed with $H(\epsilon_0)$. The sampled conformational space is then partitioned into a set of N discrete microstates $\{S_i\}$, $i \in 1, ..., N$, and a Markov state model[43] is built that allows estimating the equilibrium probability distribution π^0 associated with $H(\epsilon_0)$.

A new set of parameters $\{\epsilon\}$ can be considered a perturbation of the original ones, i.e., $\epsilon = \epsilon_0 + \delta\epsilon$. For a model with a set of parameters $\{\epsilon\}$, the equilibrium distribution can be estimated from the knowledge of the distribution with a previous set of parameters $\{\epsilon_0\}$ as

$$\pi_i = A_N \frac{\pi_i^0}{n_i} \sum_{r \in S_i} \exp\{-\beta[H(\mathbf{r}, \epsilon) - H(\mathbf{r}, \epsilon_0)]\} \tag{1}$$

where n_i is the number of frames that belong to microstate S_i and A_N is a normalizing constant

$$A_N = \left[\sum_i \frac{\pi_i^0}{n_i} \sum_{r \in S_i} \exp\{-\beta[H(\mathbf{r}, \epsilon) - H(\mathbf{r}, \epsilon_0)]\}\right]^{-1} \tag{2}$$

In general, a macroscopic observable e_k can be estimated from simulation as an ensemble average of the corresponding quantity $g_k(\mathbf{r})$ calculated for each sampled configuration r

$$e_k = \sum_i \frac{\pi_i}{n_i} \sum_{r \in S_i} g_k(\mathbf{r}) \tag{3}$$

https://doi.org/10.1021/acs.jpcb.3c04493
J. Phys. Chem. B 2023, 127, 6920−6927

Together, eqs 1 and 3 introduce a functional dependence of the value of the observable on the model parameters. The agreement between the calculated values of the observables and the experimental data can be quantified via the likelihood Q of obtaining results consistent with experimental values given the model parameters. Assuming that the errors in the measurement of the observables are independent and normally distributed, such a likelihood can be defined as

$$Q = \prod_k \exp\left(-\frac{1}{2}\frac{(e_k - f_k)^2}{\sigma_k^2}\right) \tag{4}$$

where f_k and σ_k correspond to the experimental value of the observable e_k and the corresponding uncertainty. An updated set of parameters can be obtained by minimizing with minibatch gradient descent[44] a loss function defined as

$$L(\epsilon) = -\ln Q(\epsilon) + \alpha(||\boldsymbol{\epsilon} - \boldsymbol{\epsilon_0}||^2) \tag{5}$$

The regularization parameter α in eq 5 controls the strength of the perturbation by preventing large differences between the initial set of parameters ϵ_0 and the updated parameters ϵ in order to maintain a significant overlap between the relevant configurational space sampled by the corresponding models (see SI for a discussion on the robustness of the results in a range of values for the parameter α). Since the perturbation needs to be small, the procedure delineated above is repeated multiple times until convergence in the likelihood Q is reached.

Choice of Model. As a starting point, we use a C_α structure-based coarse-grained model, where each residue is represented by a single bead, placed in the position of C_α atom.[20]

The model Hamiltonian is defined by the equation

$$H(\mathbf{r}) = H_{\text{bonded}}(\mathbf{r}) + H_{\text{nonbonded}}(\mathbf{r}) \tag{6}$$

where the $H_{\text{bonded}}(\mathbf{r})$ term represents local interactions, namely, bonds, angles, and dihedral angles, between consecutive CG beads. In this work, the parametrization of these terms remains fixed (see SI for details). Following refs 20 and 45, $H_{\text{nonbonded}}$ is a sum over all interactions V_{ij} between every pair of residues i and j with functional form defined as

$$V_{ij}(r_{ij}, r_{ij}^0, \epsilon_{ij}) = \left(\left(\frac{\sigma_{\text{ev}}}{r_{ij}}\right)^{12} + \epsilon_{ij}\right)\left[1 - \exp\left(\frac{-(r_{ij} - r_{ij}^0)^2}{2\sigma_g^2}\right)\right] - \epsilon_{ij} \tag{7}$$

for attractive interactions and as

$$V_{ij}(r_{ij}, r_{ij}^0, \epsilon_{ij}) = \left(\frac{\sigma_{\text{ev}}}{r_{ij}}\right)^{12} - \frac{1}{2}\epsilon_{ij}\left[\tanh\left(\frac{r_{ij}^0 - r_{ij} + \sigma_t}{\sigma_t}\right) + 1\right] \tag{8}$$

for repulsive interactions. Parameters σ_{ev}, σ_g, and σ_t are kept fixed and set to the same value for all of the beads. The parameters σ_g and σ_t control the width of the repulsive or attractive interaction and are taken to be $\sigma_g = \sigma_t = 0.05$ nm. The parameter σ_{ev} represents the excluded volume and is set to 0.4 nm. The variable r_{ij} denotes the distance between beads i and j.

The strength of the interaction between beads is defined by the parameter $\epsilon_{ij} > 0$ for attractive interactions and $\epsilon_{ij} < 0$ for repulsive ones. Each pair of beads has its own independent parameter that is optimized in the ODEM framework.

To initialize the optimization procedure, all of the pairwise interactions are divided into two groups: native and nonnative interactions. To determine the native contacts, we used the "shadow contact map".[46] For native contacts, r_{ij}^0 in eqs 7 and 8 is taken to be the distance between the C_α atoms in the native structure. For nonnative pairs, r_{ij}^0 is defined as $r_{ij}^0 = \sigma_{\text{ev}} + 0.2$ nm as in previous work.[35] At the first step of the optimization, the contact strength ϵ_{ij} is set to 1 for native interactions and to 0 for non-native interactions. During the ODEM optimization, the parameters can freely adjust, and the character of the potential (attractive or repulsive) is changed accordingly.

This functional form (eqs 7 and 8) has been used in several previous studies with C_α models.[35,45] In order to rule out the limited expressivity of the functional form of the 2-body interactions as the cause of disagreement between the observables experimentally measured and those estimated from simulations, we also perform an additional optimization according to the following procedure. After a converged optimization of the parameters for the potential given by eqs 7 and 8, each nonbonded interaction is represented as

$$V_{ij} = \left(\frac{\sigma_{\text{ev}}}{r_{ij}}\right)^{12} + \sum_{n=0}^{N_{bf}} c_n B_n(r_{ij}) \tag{9}$$

where B_n is a cubic B-spline and c_n are the corresponding adjustable coefficients. Such a representation is much more flexible and allows representing 2-body interaction potentials of arbitrary complexity.

The initial values of c_n are set to reproduce the optimized potential given by eqs 7 and 8. The spline representation (eq 9) is then further optimized by adjusting the c_n coefficients[47] following the ODEM procedure (see Figure 1).

As reference structures for constructing the structure-based models, crystal structures 1UBQ[48] and 1PGB[49] are used for ubiquitin and protein G, respectively. The parameters for the bonded interactions and the contact map were generated with the SMOG Web server.[50]

To investigate the importance of multibody effects, once the two-body model is fully optimized, an additional multibody term is incorporated and an additional optimization of the 2- and many-body nonbonded interactions together is performed. In order to model the effect of the solvent, the multibody term is taken to be in the form of the burial term used in the AWSEM force field[23]

$$V_{\text{multibody}} = -\frac{1}{2}\lambda_{\text{burial}}\sum_{i=1}^{N}\sum_{\mu=1}^{3}\gamma_{\text{burial}}(a_i, \mu) f(\rho_i, \mu) \tag{10}$$

where

$$f(\rho_i, \mu) = \tanh[\eta(\rho_i - \rho_{\text{min}}^\mu)] + \tanh[\eta(\rho_{\text{max}}^\mu - \rho_i)] \tag{11}$$

and ρ_i is the number of residues in direct contact defined as

$$\rho_i = \sum_{j=1}^{N}\left[\frac{1}{4}(1 + \tanh[\eta(r_{ij} - r_{\text{min}})])(1 + \tanh[\eta(r_{\text{max}} - r_{ij})])\right] \tag{12}$$

The parameters η, r_{max}, r_{min}, ρ_{min}^μ, ρ_{max}^μ, and λ_{burial} are fixed and given in the SI, while the parameters $\gamma_{\text{burial}}(a_i, \mu)$ are adjusted during the optimization.

https://doi.org/10.1021/acs.jpcb.3c04493
J. Phys. Chem. B 2023, 127, 6920−6927

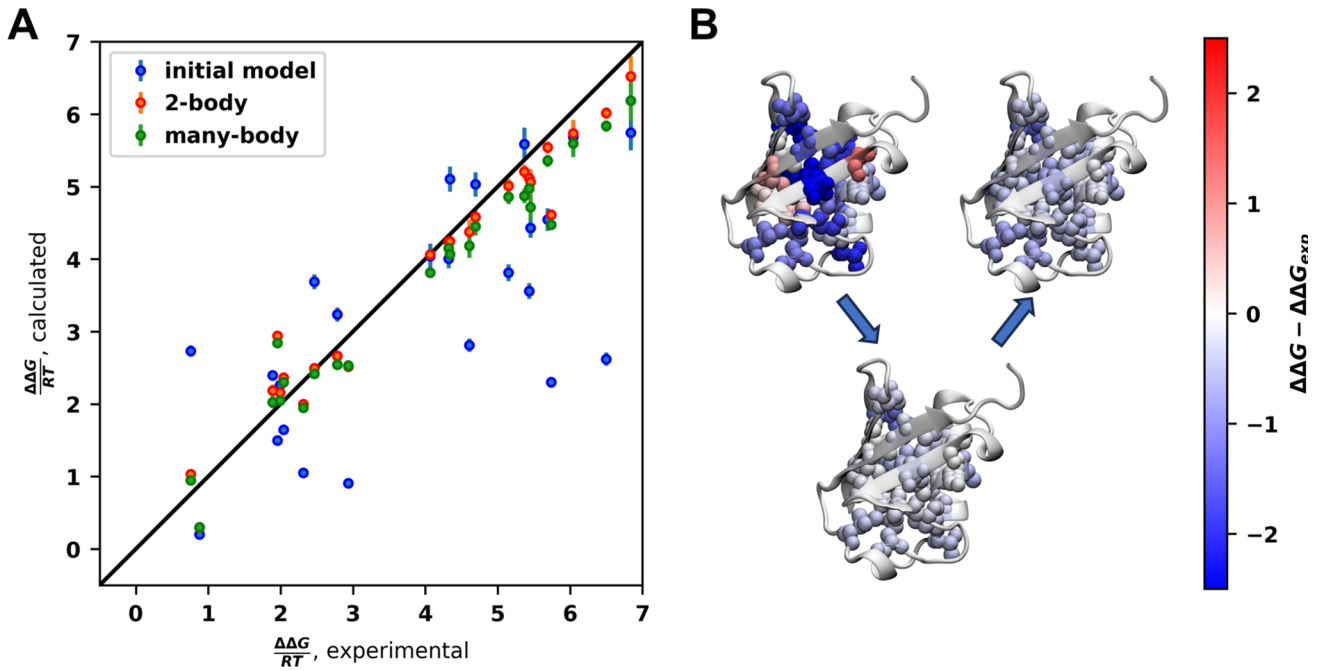


Figure 2. Optimization of the coarse-grained model of ubiquitin. (A) Correlation between the calculated and the experimental ΔΔ*G*s. (Blue) Uniform structure-based CG model, where all native contacts have strength 1 and nonnative interactions have strength 0. (Orange) Optimized model with 2-body potential only. (Green) Optimized model with 2- and many-body potential terms. Error bars represent one standard deviation over 5 independent optimization runs. (B) Absolute error in ΔΔ*G*, shown for the initial model (top left), the model optimized with 2-body interactions only (bottom middle), and the model with a many-body correction (top right).

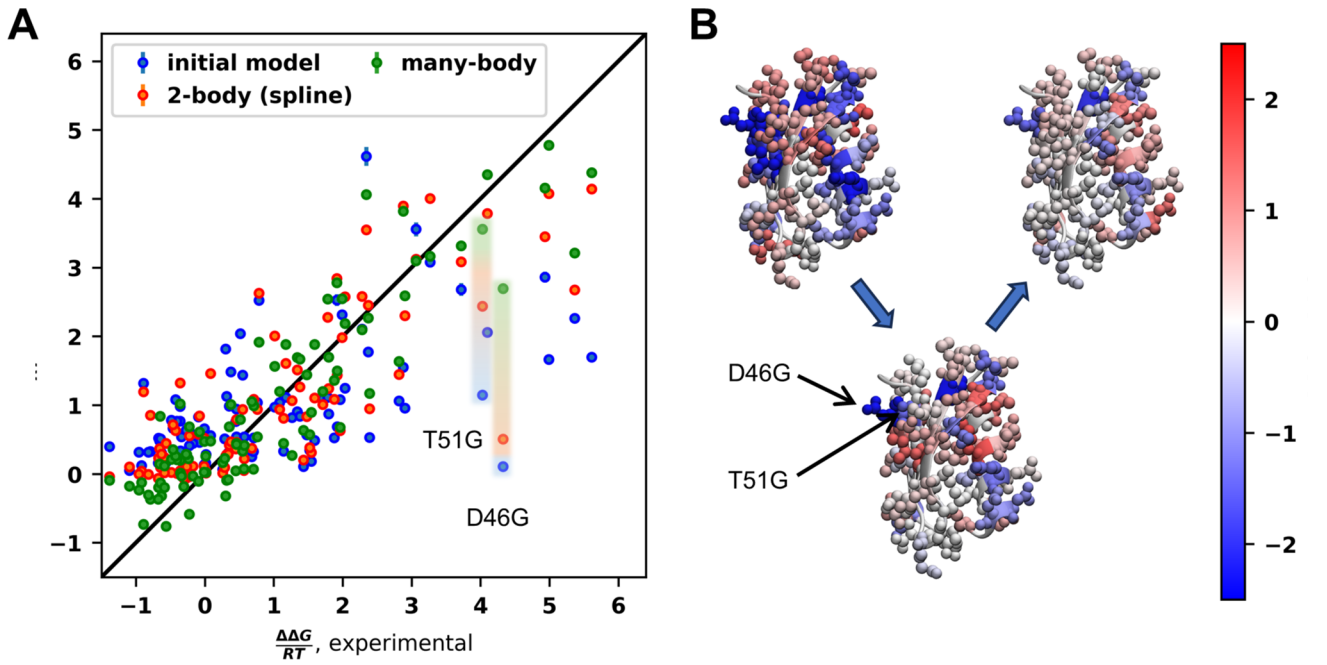


Figure 3. Optimization of the coarse-grained model of the domain B1 protein G. (A) Correlation between the calculated and the experimental ΔΔ*G*s. (Blue) Uniform structure-based CG model, where all native contacts have uniform strength and nonnative interactions have strength 0. (Orange) Optimized model with a 2-body spline-based potential. (Green) Optimized model with 2- and many-body potential. Error bars represent one standard deviation over 5 independent optimization runs. (B) Absolute error in ΔΔ*G*, shown for the initial model (top left), the model optimized with 2-body interactions only (bottom middle), and the model with a many-body correction (top right). Results for mutations D46G and T51G, which involve surface residues, are highlighted.

Reference Observables. For each protein, we use as reference observables a set of experimentally determined free energy differences upon mutation, ΔΔ*G*s, from the work by Nisthal et al. for GB1[51] and Went et al. for ubiquitin.[52] For optimization, the experimental values of ΔΔ*G*s are extrapolated to zero concentration of denaturant and rescaled to the protein folding temperature as described in ref 24 (see SI for details on data preparation).

https://doi.org/10.1021/acs.jpcb.3c04493
J. Phys. Chem. B 2023, 127, 6920−6927

Table 1. Optimization Results for 1GB1 Mutants, Where Absolute Error for 2-Body Spline Model Exceeds 1.5 RT Units

		ΔΔG, RT units (absolute error wrt experiment, RT units)			
mutant	experimental ΔΔG, RT units	initial	2 body	2 body, spline	many body
D46G	4.33	0.11 (−4.22)	0.26 (−4.07)	0.51 (−3.82)	2.70 (−1.63)
Y03L	5.37	2.26 (−3.11)	2.36 (−3.10)	2.67 (−2.69)	3.21 (−2.15)
Y03F	−0.89	1.32 (2.21)	1.25 (2.14)	1.20 (2.09)	−0.73 (0.16)
F30L	0.78	2.52 (1.74)	2.59 (1.81)	2.63 (1.84)	1.92 (1.14)
T18A	−0.36	0.98 (1.34)	1.18 (1.54)	1.32 (1.68)	1.02 (1.38)
I06V	−0.79	0.32 (1.11)	0.86 (1.65)	0.85 (1.65)	−0.37 (0.43)
T51G	4.03	1.15 (−2.88)	2.36 (−1.66)	2.44 (−1.59)	3.56 (−0.47)

From the coarse-grained simulation, values of $\Delta\Delta G$ can be calculated by treating the effects of the mutation of the kth residue as a perturbation $\delta H(k)$ to the wild-type Hamiltonian H.[24,25] If we assume that the perturbation does not significantly change the density of states, the effect of the mutation on the free energy landscape can be estimated as

$$\beta\Delta\Delta G(k) = \ln \frac{\langle \exp(-\delta H(k)) \rangle_{\mathrm{U}}}{\langle \exp(-\delta H(k)) \rangle_{\mathrm{F}}} \tag{13}$$

where the angular brackets represent the canonical average over the folded (F) or unfolded (U) state ensemble of the unperturbed system.

For the 2-body potential, the perturbation is

$$\delta \mathrm{H}^{k}_{\text{2-body}} = \sum_{i,j} V_{ij}(r_{ij}, r^{0}_{ij}, \epsilon^{k}_{ij}) - \sum_{i,j} V_{ij}(r_{ij}, r^{0}_{ij}, \epsilon_{ij}) \tag{14}$$

where the strength of the pairwise interaction ϵ^{k}_{ij} in the mutant k is obtained by rescaling the strength of the corresponding interaction parameter in the wild-type protein $\epsilon^{k}_{ij}=f^{k}_{ij}\epsilon_{ij}$, and the rescaling factor $0 \le f^{k}_{ij} \le 1$ is calculated as the fraction of native contacts deleted by the mutation in the native structure of the wild-type protein. In our analysis, we include only mutants for which $f^{k}_{ij} \neq 0$. The change in the multibody contribution to the Hamiltonian (eq 10) as a result of a point mutation is calculated for each multibody term as

$$\delta H^{k}_{\text{many-body}} = -\frac{1}{2}\lambda_{\text{burial}} \sum_{\mu=1}^{3} (\gamma_{\text{burial}}(a'_k, \mu) - \gamma_{\text{burial}}(a_k, \mu)) f(\rho_i, \mu) \tag{15}$$

Here, a_k represents the identity of the mutated residue in the wild-type protein, and a'_k represents the amino acid residue in the mutant.

RESULTS

For both ubiquitin and GB1, the optimization procedure was repeated five times, and the corresponding aggregated data are presented in Figures 2 and 3. First, the parameters of the pairwise potential are optimized until convergence, starting from the "vanilla" structure-based model with 2-body nonbonded interactions only.

For ubiquitin (Figure 2), the convergence of the root-mean-square-error (RMSE) for the observables is reached within 10−15 iterations, producing good agreement with the experimental results even with a fixed functional form (eqs 7 and 8) with a final RMSE of 0.42 ± 0.03 RT units.

In the case of GB1 (Figure 3), the optimization with a fixed-form 2-body potential takes 80−90 iterations and converges to a significantly higher value of RMSE, 1.002 ± 0.003 RT units (starting from a value of 1.276 ± 0.002 RT in the structure-based model). To account for possible limitations of the functional form used to represent 2-body nonbonded interactions, we have additionally optimized a spline-based 2-body model (eq 9) using the optimized 2-body model with fixed functional form as a starting point (as discussed in the previous section). The increased expressivity of the 2-body interaction potential provides only a minimal improvement in the agreement with experimental results and achieves an RMSE of 0.94 RT units with the absolute error of 7 out of 95 considered mutations still exceeding 1.5 RT units (see Table 1).

For both systems, the optimized two-body model producing the lowest RMSE is selected for further analysis. An additional multibody term is then added to the selected model, and the optimization is continued by tuning both the two-body and the multibody parameters together. The incorporation of multibody interactions in the model for GB1 significantly improves the overall agreement with experimental data (RMSE of 0.69 RT units) and corrects the most prominent outliers (see Figure 3).

It is worth noting that in the 2-body part of the potential an individual parameter is assigned for each pair of residues, not per pair of residue types. This choice gives a more flexible potential, suitable for our task of proving the limitation of 2-body interactions; however, it makes the model parameters depend on both the type and the position of the residues and does not allow a direct comparison of parameters between different proteins or with other knowledge-based potentials, where the parameters depend only on the residue types involved.

DISCUSSION AND CONCLUSIONS

In the case of the GB1, the largest deviation from the experimental value of $\Delta\Delta G$ is observed for the mutant D46G (Table 1): the standard structure-based model underestimates the destabilizing effect of this mutation by 4.3 RT units. The optimization using a fixed function 2-body potential reduces this error slightly to 4.1 RT units, and increasing the expressivity of the 2-body potential with the spline model only slightly decreases the error to 3.8 RT units. However, the incorporation of many-body effects decreases the error more than twice to 1.6 RT units. For each mutant, we calculated the Ω angle as a metric of side-chain orientation, as defined by Yan and Jernigan.[53] Among the seven mutations reported in Table 1, all but F30L and T18A involve residues where the side chains point outward from the surface ($\cos(\Omega) < 0$) in the crystal structure. D46 and I06 are significantly exposed to the solvent with a solvent-accessible area of 70.06 and 68.22 Å^2, respectively. Additionally, except for F30 and I06, all of the other mutations are polar or charged residues that can

https://doi.org/10.1021/acs.jpcb.3c04493
J. Phys. Chem. B 2023, 127, 6920−6927

participate in specific interactions. We interpret these results as indicating that, at least for this protein, multibody terms are needed to account for the solvent degrees of freedom that are implicit in the CG model. This is in agreement with previous results, suggesting that while for short chains without specific interactions a solvation potential can be reliably approximated as a sum of pairwise potentials,[54] multibody terms cannot be neglected for more complicated systems.[55]

Ubiquitin has a size and topology similar to GB1, and the performance of the original zeroth-order structure-based model is also similar (with RMSE 1.54 RT compared to the initial RMSE 1.28 RT in the case of GB1). The optimization yields a remarkably good agreement with experimental data already with a simple 2-body model with a final RMSE 0.42 ± 0.03 RT. Even the effects of mutation of polar/charged residues, such as K27A and Q41A, are reproduced relatively well. Further addition of many-body terms does not significantly improve the model, and the overall model performance is comparable (see Figure 2).

We speculate that this result reflects the properties of ubiquitin compared to the GB1 and the position of the considered mutations. Our results suggest that for ubiquitin, many-body effects do not play a major role in the change of stability $\Delta\Delta G$s for the available mutants. As shown in Figure 2B, most of the mutated residues are buried inside the protein, and solvent effects on the stability may be limited. An analogous result is known for CG models of liquids: While for liquid methanol two-body interactions in a CG model are sufficient to reproduce the structural features obtained with an atomistic system, a two-body-only representation is inadequate to describe the same properties of liquid water.[40] Our conclusions on the importance of multibody terms are consistent with the findings of Wang et al.,[41] who, by using a bottom-up approach, showed that multibody terms are needed for small proteins to reproduce the folding free energy landscape of atomistic simulations correctly. For this reason and to allow a more expressive functional form for the multibody terms, recent advances in the development of protein CG models make use of neural networks to represent nonbonded interaction potentials.[41,56−64]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.3c04493.

Hamiltonians; simulation details; analysis protocol; free energy surfaces for initial and optimized models; Chapman−Kolmogorov tests for Markov state models for initial and optimized models; details of $\Delta\Delta G$ calculations; effects of hyperparameters on $\Delta\Delta G$ values; description of training data sets; optimization procedure; optimization results for GB1 and ubiquitin; correlation between errors in the two-body model and physico-chemical properties of mutated residues; residue-wise properties of GB1 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Cecilia Clementi − *Department of Physics, Freie Universität, Berlin 14195, Germany; Center for Theoretical Biological Physics and Department of Chemistry, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0001-9221-2358; Email: cecilia.clementi@fu-berlin.de

Author

Iryna Zaporozhets − *Department of Chemistry and Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States; Department of Physics, Freie Universität, Berlin 14195, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.3c04493

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the members of Clementi's group for insightful discussions and comments on the manuscript. We acknowledge funding from the Deutsche Forschungsgemeinschaft DFG (SFB/TRR 186, Project A12; SFB 1114, Projects A04, B03, and B08; SFB 1078, Project C7; and RTG 2433, Project Q05), the National Science Foundation (CHE-1900374 and PHY-2019745), and the Einstein Foundation Berlin (Project 0420815101).

■ REFERENCES

(1) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.

(2) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9885−9890.

(3) Clementi, C. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10−15.

(4) Noé, F.; Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.* **2017**, *43*, 141−147.

(5) Freitas, F. C.; Fuchs, G.; de Oliveira, R. J.; Whitford, P. C. The Dynamics of Subunit Rotation in a Eukaryotic Ribosome. *Biophysica* **2021**, *1*, 204−221.

(6) Cassaignau, A. M.; Włodarski, T.; Chan, S. H.; Woodburn, L. F.; Bukvin, I. V.; Streit, J. O.; Cabrita, L. D.; Waudby, C. A.; Christodoulou, J. Interactions between nascent proteins and the ribosome surface inhibit co-translational folding. *Nat. Chem.* **2021**, *13*, 1214−1220.

(7) Dessaux, D.; Mathé, J.; Ramirez, R.; Basdevant, N. Current Rectification and Ionic Selectivity of a-Hemolysin: Coarse-Grained Molecular Dynamics Simulations. *J. Phys. Chem. B* **2022**, *126*, 4189−4199.

(8) Pak, A. J.; Yu, A.; Ke, Z.; Briggs, J. A.; Voth, G. A. Cooperative multivalent receptor binding promotes exposure of the SARS-CoV-2 fusion machinery core. *Nat. Commun.* **2022**, *13*, 1002.

(9) Jin, S.; Bueno, C.; Lu, W.; Wang, Q.; Chen, M.; Chen, X.; Wolynes, P. G.; Gao, Y. Computationally exploring the mechanism of bacteriophage T7 gp4 helicase translocating along ssDNA. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119*, No. e2202239119.

(10) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253*, 694−698.

(11) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139*, 090901.

(12) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244114.

(13) Tschöp, W.; Kremer, K.; Batoulis, J.; Bürger, T.; Hahn, O. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polym.* **1998**, *49*, 61−74.
(14) Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.
(15) Chaimovich, A.; Shell, M. S. Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.* **2011**, *134*, 094112.
(16) Lyubartsev, A. P.; Laaksonen, A. Osmotic and activity coefficients from effective potentials for hydrated ions. *Phys. Rev. E* **1997**, *55*, 5689−5696.
(17) Müller-Plathe, F. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem* **2002**, *3*, 754−769.
(18) Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.* **2003**, *24*, 1624−1636.
(19) Taketomi, H.; Ueda, Y.; Gō, N. Studies on protein folding, unfolding and fluctuation by computer simulatioin: I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **1975**, *7*, 445−459.
(20) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: What determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937−953.
(21) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.
(22) Souza, P. C.; Alessandri, R.; Barnoud, J.; Thallmair, S.; Faustino, I.; Grünewald, F.; Patmanidis, I.; Abdizadeh, H.; Bruininks, B. M.; Wassenaar, T. A.; et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat. Methods* **2021**, *18*, 382−388.
(23) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **2012**, *116*, 8494−8503.
(24) Matysiak, S.; Clementi, C. Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: How far can a minimalist model go? *J. Mol. Biol.* **2004**, *343*, 235−248.
(25) Matysiak, S.; Clementi, C. Minimalist Protein Model as a Diagnostic Tool for Misfolding and Aggregation. *J. Mol. Biol.* **2006**, *363*, 297−308.
(26) Chen, M.; Schafer, N. P.; Wolynes, P. G. Surveying the Energy Landscapes of Aβ Fibril Polymorphism. *J. Phys. Chem. B* **2018**, *122*, 11414−11430.
(27) Chen, X.; Lu, W.; Tsai, M. Y.; Jin, S.; Wolynes, P. G. Exploring the folding energy landscapes of heme proteins using a hybrid AWSEM-heme model. *J. Biol. Phys.* **2022**, *48*, 37−53.
(28) Zheng, W.; Tsai, M.-Y.; Chen, M.; Wolynes, P. G. Exploring the aggregation free energy landscape of the amyloid-β protein (1−40). *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 11835−11840.
(29) Cesari, A.; Gil-Ley, A.; Bussi, G. Combining Simulations and Solution Experiments as a Paradigm for RNA Force Field Refinement. *J. Chem. Theory Comput.* **2016**, *12*, 6192−6200.
(30) Dannenhoffer-Lafage, T.; White, A. D.; Voth, G. A. A Direct Method for Incorporating Experimental Data into Multiscale Coarse-Grained Models. *J. Chem. Theory Comput.* **2016**, *12*, 2144−2153.
(31) Latham, A. P.; Zhang, B. Improving Coarse-Grained Protein Force Fields with Small-Angle X-ray Scattering Data. *J. Phys. Chem. B* **2019**, *123*, 1026−1034.
(32) Latham, A. P.; Zhang, B. Maximum Entropy Optimized Force Field for Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **2020**, *16*, 773−781.
(33) Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.* **2008**, *94*, 182−192.
(34) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(35) Chen, J.; Chen, J.; Pinamonti, G.; Clementi, C. Learning Effective Molecular Models from Experimental Observables. *J. Chem. Theory Comput.* **2018**, *14*, 3849−3858.
(36) Thaler, S.; Zavadlav, J. Learning neural network potentials from experimental data via Differentiable Trajectory Reweighting. *Nat. Commun.* **2021**, *12*, 6884.
(37) Ejtehadi, M. R.; Avall, S. P.; Plotkin, S. S. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15088.
(38) Larini, L.; Lu, L.; Voth, G. A. The multiscale coarse-graining method. VI. Implementation of three-body coarse-grained potentials. *J. Chem. Phys.* **2010**, *132*, 164107.
(39) John, S. T.; Csányi, G. Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials. *J. Phys. Chem. B* **2017**, *121*, 10934−10949.
(40) Scherer, C.; Andrienko, D. Understanding three-body contributions to coarse-grained force fields. *Phys. Chem. Chem. Phys.* **2018**, *20*, 22387.
(41) Wang, J.; Charron, N.; Husic, B.; Olsson, S.; Noé, F.; Clementi, C. Multi-body effects in a coarse-grained protein force field. *J. Chem. Phys.* **2021**, *154*, 164113.
(42) Joseph, J. A.; Reinhardt, A.; Aguirre, A.; Chew, P. Y.; Russell, K. O.; Espinosa, J. R.; Garaizar, A.; Collepardo-Guevara, R. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat. Comput. Sci.* **2021**, *1*, 732−743.
(43) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.
(44) Ruder, S. An overview of gradient descent optimization algorithms. *arXiv (Computer Science)* DOI: 10.48550/arXiv.1609.04747 (accessed 2023−07−10).
(45) Lammert, H.; Schug, A.; Onuchic, J. N. Robustness and generalization of structure-based models for protein folding and function. *Proteins* **2009**, *77*, 881−891.
(46) Noel, J. K.; Whitford, P. C.; Onuchic, J. N. The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B* **2012**, *116*, 8692−8702.
(47) Xie, S. R.; Rupp, M.; Hennig, R. G. Ultra-fast interpretable machine-learning potentials. *arXiv (Condensed Matter)* DOI: 10.48550/arXiv.2110.00624 (accessed 2023−07−10).
(48) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **1987**, *194*, 531−544.
(49) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. Two Crystal Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein G and Comparison with NMR. *Biochemistry* **1994**, *33*, 4721−4729.
(50) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput. Biol.* **2016**, *12*, No. e1004794.
(51) Nisthal, A.; Wang, C. Y.; Ary, M. L.; Mayo, S. L. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 16367−16377.
(52) Went, H. M.; Jackson, S. E. Ubiquitin folds through a highly polarized transition state. *Protein Eng. Des. Sel.* **2005**, *18*, 229−237.
(53) Yan, A.; Jernigan, R. L. How do side chains orient globally in protein structures? *Proteins* **2005**, *61*, 513−522.
(54) Taylor, M. P.; Petersen, G. M. Solvation potentials for flexible chain molecules in solution: On the validity of a pairwise decomposition. *J. Chem. Phys.* **2007**, *127*, 184901.
(55) Chang, R.; Yethiraj, A. Solvent effects on the collapse dynamics of polymers. *J. Chem. Phys.* **2001**, *114*, 7688−7699.

(56) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Central Science* **2019**, *5*, 755−767.
(57) Husic, B. E.; Charron, N. E.; Lemm, D.; Wang, J.; Perez, A.; Majewski, M.; Krämer, A.; Chen, Y.; Olsson, S.; de Fabritiis, G.; et al. Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.* **2020**, *153*, 194101.
(58) Chen, Y.; Krämer, A.; Charron, N. E.; Husic, B. E.; Clementi, C.; Noé, F. Machine learning implicit solvation for molecular dynamics. *J. Chem. Phys.* **2021**, *155*, 084101.
(59) Majewski, M.; Pérez, A.; Thölke, P.; Doerr, S.; Charron, N. E.; Giorgino, T.; Husic, B. E.; Clementi, C.; Noé, F.; De Fabritiis, G. Machine Learning Coarse-Grained Potentials of Protein Thermodynamics. *arXiv (Quantitative Biology)* DOI: 10.48550/arXiv.2212.07492 (accessed 2023−07−10).
(60) Köhler, J.; Chen, Y.; Krämer, A.; Clementi, C.; Noé, F. Flow-Matching: Efficient Coarse-Graining of Molecular Dynamics without Forces. *J. Chem. Theory Comput.* **2023**, *19*, 942−952.
(61) Durumeric, A. E.; Charron, N. E.; Templeton, C.; Musil, F.; Bonneau, K.; Pasos-Trejo, A. S.; Chen, Y.; Kelkar, A.; Noé, F.; Clementi, C. Machine learned coarse-grained protein force-fields: Are we there yet? *Curr. Opin. Struct. Biol.* **2023**, *79*, 102533.
(62) Sahrmann, P. G.; Loose, T. D.; Durumeric, A. E. P.; Voth, G. A. Utilizing Machine Learning to Greatly Expand the Range and Accuracy of Bottom-Up Coarse-Grained Models through Virtual Particles. *J. Chem. Theory Comput.* **2023**, DOI: 10.1021/acs.jctc.2c01183.
(63) Lemke, T.; Peter, C. Neural network based prediction of conformational free energies - a new route toward coarse-grained simulation models. *J. Chem. Theory Comput.* **2017**, *13*, 6213−6221.
(64) Yao, S.; Van, R.; Pan, X.; Park, J. H.; Mao, Y.; Pu, J.; Mei, Y.; Shao, Y. Machine learning based implicit solvent model for aqueous-solution alanine dipeptide molecular dynamics simulations. *RSC Adv.* **2023**, *13*, 4565−4577.

https://doi.org/10.1021/acs.jpcb.3c04493
J. Phys. Chem. B 2023, 127, 6920−6927