JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC

Article

Multi-*e*GO: Model Improvements toward the Study of Complex Self-Assembly Processes

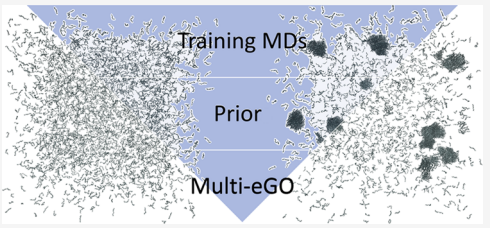
Fran Bačić Toplek,[§] Emanuele Scalone,[§] Bruno Stegani, Cristina Paissoni, Riccardo Capelli, and Carlo Camilloni*

Cite This: *J. Chem. Theory Comput.* 2024, 20, 459−468

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Structure-based models have been instrumental in simulating protein folding and suggesting hypotheses about the mechanisms involved. Nowadays, at least for fast-folding proteins, folding can be simulated in explicit solvent using classical molecular dynamics. However, other self-assembly processes, such as protein aggregation, are still far from being accessible. Recently, we proposed that a hybrid multistate structure-based model, multi-*e*GO, could help to bridge the gap toward the simulation of out-of-equilibrium, concentration-dependent self-assembly processes. Here, we further improve the model and show how multi-*e*GO can effectively and accurately learn the conformational ensemble of the amyloid β42 intrinsically disordered peptide, reproduce the well-established folding mechanism of the B1 immunoglobulin-binding domain of streptococcal protein G, and reproduce the aggregation as a function of the concentration of the transthyretin 105−115 amyloidogenic peptide. We envision that by learning from the dynamics of a few minima, multi-*e*GO can become a platform for simulating processes inaccessible to other simulation techniques.



1. INTRODUCTION

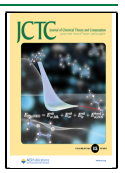
Molecular dynamics (MD) simulations, based on conventional transferable molecular mechanics force fields, have become a standard tool in biological research thanks to their ability to resolve the atomic details of many molecular processes.[1] This success is the result of a combination of increased computational resources and associated software, improved sampling methods, more accurate force fields, and better methods for integrating simulations with experimental information.[2] In addition, intrinsically disordered proteins (IDPs) or regions have highlighted the need to complement the structure with dynamics by challenging the sequence-structure paradigm.[3] Notably, the revolution in AI-based structure prediction tools has further widened the scope of simulations by allowing the study of systems whose structure has not yet been experimentally determined.[4,5]

Despite these advances, conventional atomistic MD simulations are still limited to the study of relatively few molecules on a time scale of tens of microseconds.[6−8] To overcome size and time scale limitations, one can limit the scope of the simulation technique to specific subdomains and use simplified models.[9] This strategy is exemplified by the Martini force field,[10] which focuses mainly on folded proteins and lipid membranes and allows the simulation of large protein complexes in a realistic environment. More recently, other simplified models have emerged, focusing on IDPs and their interaction processes in the context of liquid−liquid phase separation.[11−14] We are also seeing the first examples of simplified models resulting from machine-learned potentials trained on classical force fields.[15] Since the 1990s, structure-based models have played a key role in elucidating the protein folding process by learning a system-dependent potential that should have an absolute energy minimum centered on the chosen folded structure.[16]

Recently, building on the observation that the amyloid structure could be the most stable one that protein molecules can adopt under physiological conditions,[17] we have revisited structure-based models to incorporate information from multiple minima with the aim of describing the aggregation of a peptide into an amyloid fibril.[18] Our model, called multi-*e*GO, allowed us to qualitatively capture the experimental macroscopic features of protein aggregation, including kinetics and fibril morphology, and to shed light on the microscopic features of the process.[18]

Multi-*e*GO is a hybrid transferable/structure-based model defined from a combination of simulations and structures. Only the heavy atoms (nonhydrogens) are included to maintain atomic resolution. Bonds, angles, dihedrals, and default $C^{(12)}$ values are based on the GROMOS54a7 force

Received: October 25, 2023
Revised: December 16, 2023
Accepted: December 18, 2023
Published: December 28, 2023



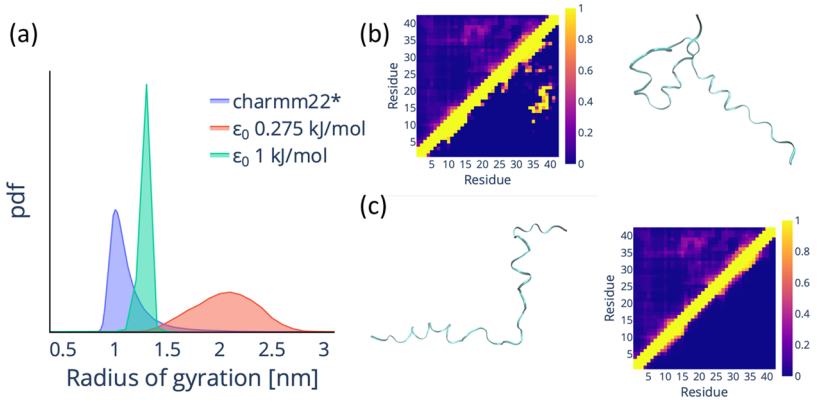


Figure 1. Original multi-*e*GO fails to reproduce the conformational dynamics of the Aβ42 monomer. (a) Probability density function (pdf) for the backbone radius of gyration of Aβ42 for a training simulation (charmm22*, blue)[22] and for an original multi-*e*GO simulation with ε_0 = 0.275 kJ/mol (red) and ε_0 = 1 kJ/mol (green). (b) Comparison of the per-residue probability contact map for the training and original multi-*e*GO simulation with ε_0 = 1 kJ/mol and a representative structure from the multi-*e*GO simulation. The colored bar represents the contact probability. (c) Comparison of the per-residue probability contact map for the training and original multi-*e*GO simulation with ε_0 = 0.275 kJ/mol and a representative structure from the multi-*e*GO simulation. The color bar represents the contact probability.

field,[19] being already optimized without nonpolar hydrogens, dihedral terms for the φ and ψ torsions and some 1−4 nonbonded interactions (i.e., pairs of atoms that are separated by three consecutive covalent bonds) are specifically reoptimized. Attractive nonbonded interactions are obtained from either a PDB structure or an MD simulation and parametrized using the Lennard−Jones (LJ) potential. The structure-based potential is defined by pairs of atoms within a 0.55 nm cutoff, with an interaction strength rescaled using contact probabilities. The $\varepsilon_{i,j}$ of the LJ potential was heuristically defined as

$$\varepsilon_{i,j} = \varepsilon_0\left(1 - \frac{\ln P_{i,j}^{MD}}{\ln P_{threshold}}\right) = -\frac{\varepsilon_0}{\ln P_{threshold}}\cdot\ln\frac{P_{i,j}^{MD}}{P_{threshold}} \quad (1)$$

where ε_0 is the maximum interaction energy provided in the input, $P_{i,j}^{MD}$ is the population for the contact between atoms i,j as obtained from a training MD simulation, and $P_{threshold}^{MD}$ is a minimum population that should be considered. Such LJ parametrization allowed an increase in protein flexibility when combined with the transferable bonded terms.

Encouraged by our previous results, we built a multi-*e*GO model to simulate the amyloid β42 peptide (Aβ42)[20,21] using previously published MD trajectories[22] for training (with the only difference that we set $P_{threshold}$ to 0.01 compared to the 0.09 value used in our previous work). As shown in Figure 1, the original multi-*e*GO parametrization does not allow us to capture the training Aβ42 conformational ensemble with any ε_0 value. Our interpretation is that this is due to the imbalance between the highly populated local contacts and the very weakly populated contacts between distant residues, which are typical IDPs such as Aβ42. Our hypothesis was supported by the impossibility of learning contacts using the smaller epsilon value and the impossibility of getting to a compact structure at a larger epsilon. These results indicate the need to improve eq 1 to better account for the polymer properties.

In what follows, we therefore present a reformulation of the multi-*e*GO model in which following Bayesian statistics, we update a prior polymer model with additional information learned from training simulations. We show how this leads to a more complete description of attractive and repulsive interactions. We demonstrate that the updated multi-*e*GO can correctly learn the dynamics of an IDP such as Aβ42, can be used to describe the folding mechanism of a small protein such as the B1 immunoglobulin-binding domain of streptococcal protein G (GB1),[23] and can still reproduce the recently published results on the aggregation of the transthyretin 105−115 amyloidogenic peptide ($TTR_{105-115}$).[18,24] We therefore propose this improved multi-*e*GO as a model that, using only the information that can be generated with conventional MD, can approximate processes on size and time scales that are orders of magnitude larger than the state of the art.

2. THEORY

2.1. Multi-*e*GO: A Bayesian Reformulation. Equation 1 introduced above suggests, in the form on the right side, that the MD contact probabilities are weighted by a uniform, uninformative, prior distribution. Proteins are polymers with a local geometry described by the Ramachandran plot, and as such, the contact probabilities between atoms along the chain are influenced by the chain geometry. A more informative prior would be that of a self-avoiding chain with local geometries as close as possible to those of proteins. We call random coil (RC) probability distribution the contact pair distribution resulting from a simulation of such a model (cf., next section). Consideration of this new prior leads to a reformulated equation for the interaction energy

$$\varepsilon_{i,j} = -\frac{\varepsilon_0^{intra}}{\ln P_{threshold}^{RC}}\cdot\ln\frac{P_{i,j}^{MD}}{\max(P_{i,j}^{RC}, P_{threshold}^{RC})} \quad (2)$$

where $P_{i,j}^{MD}$ and $P_{i,j}^{RC}$ are the fraction of frames with a native contact in the MD and RC simulations, respectively. In this new parametrization, the information on the prior model is retained until a minimum value $P_{threshold}^{RC}$ is reached. This equation is the core of the new multi-*e*GO, and it is important to note that it is meant to account in general for any prior assumptions, which means that other prior models can be used for specific problems, as will be shown later for intermolecular interactions. Another important consequence of this formula is that if $P_{i,j}^{RC} > P_{i,j}^{MD}$, the sign of the energy changes, indicating the need to introduce repulsive interactions. Notably, this is similar to approaches previously introduced to reweight statistical potentials used in protein structure prediction.[25]

https://doi.org/10.1021/acs.jctc.3c01182
J. Chem. Theory Comput. 2024, 20, 459−468

2.2. Multi-*e*GO Prior, Random Coil, Model. In the new formulation of multi-*e*GO, the prior is essential to define the nonbonded interaction strength. Therefore, we reoptimized the local geometries of the model to obtain the most informative prior possible by reparametrizing the default $C^{(12)}$ parameters of the LJ potential, the 1−4 excluded volume pairs, and the dihedral parameters. First, bonds, angles, and proper and improper dihedrals were taken from the GROMOS54a7[19] force field as before. $C^{(12)}$ values for the GROMOS54a7 atom types were scaled down so that the atomic radius is defined as the distance at which the repulsion is equal to the thermal energy at 300 K (k_BT = 2.49 kJ/mol). In the case of oxygen−oxygen interactions, we introduced a scaled-up $C^{(12)}$ (11.4-fold larger than the obtained by the procedure described above) to account for their strong electrostatic repulsion. We then introduced new 1−4 pairs, defined only by their $C^{(12)}$, to account for the correct local excluded volume potential. The newly introduced pairs include $C_{-1}-C_\beta$, $C_\beta-O$, $C_\beta-N_{+1}$, $N-N_{+1}$, $C-C_{+1}$, $C-C_\gamma$, and $N-C_\gamma$, most of which were added based on.[26,27] 1−4 $C^{(12)}$ were fine-tuned using different dipeptides to match a corresponding target Ramachandran distribution, as obtained from an explicit solvent simulation using the CHARMM22* force field.[28] Finally, the parameters corresponding to the φ and ψ backbone dihedrals were optimized to minimize the difference between the Ramachandran distributions. The dipeptides used were glycine dipeptide, proline dipeptide, alanine dipeptide, and valine dipeptide, the latter two being used to represent residues with small and bulky side chains, respectively. The small amino acids include alanine, serine, and threonine, while the valine dipeptide was chosen as a proxy for the other 15 amino acids. This distinction was necessary because bulky amino acids behave differently toward the upper left corner (extended β conformation) of the Ramachandran distribution. Bulky amino acids generally have less extended β conformations and a smoother β distribution overall, requiring separately optimized dihedral parameters.

2.3. Attractive Interactions. Nonbonded interactions in multi-*e*GO are implemented using the LJ potential, i.e., $U_{LJ} = \frac{C^{(12)}}{r^{12}} - \frac{C^{(6)}}{r^6} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$, with repulsive interactions defined only by the value $C^{(12)} = 4\varepsilon\sigma$.[12] To generate the model potential, we then set both ε and σ. As already introduced above, ε is defined as in eq 2 from the ratio between the contact probability of a pair of atoms as observed in the training MD simulation and in the corresponding RC simulation (i.e., a simulation based only on the multi-*e*GO prior model). As a rule, the RC simulation must be performed at the same temperature as the corresponding training MD. The multi-*e*GO simulation can then be performed at a different temperature, but given the simplified nature of the interactions, the extrapolation in temperature will add additional approximations. The contact probability between two atoms is defined here as $P_{i,j} = \int_0^{R_{i,j}^{cut}} P_{i,j}(x)dx$, i.e., the probability of observing the i,j pair in the distance range $[0, R_{i,j}^{cut}]$. In the original multi-*e*GO, $R_{i,j}^{cut}$ was set to 0.55 nm irrespective of the i,j pair, but it should be noted that in a conventional MD simulation, certain atom types can hardly form interactions in this distance due to their size, e.g., α carbons, while for other pairs, this distance is far too permissive, e.g., nitrogen−oxygen forming a hydrogen bond. Therefore, we have introduced a variable $R_{i,j}^{cut} = f_{cut}(C_i^{(12)}C_j^{(12)})^{1/24}$, where the cutoff factor f_{cut} = 1.45 is chosen so that the atom forming pairs are subject to 80% of the standard LJ attraction. In LJ, σ is related to the position of the minimum of the potential as $\sigma = \frac{r_{min}}{2^{1/6}}$, where r_{min} should be considered as the interaction length. To estimate r_{min} from our simulations, we use an exponential averaging with a resolution of 0.1 nm

$$\langle r_{min}\rangle_{exp} = \frac{1}{0.1}\left[\ln\left(\int_0^{R_{i,j}^{cut}} P_{i,j}(x)\exp\left(\frac{1}{0.1x}\right)dx\right)\right]^{-1} \quad (3)$$

From eq 2, it is clear that ε can tend to zero if $P_{i,j}^{RC} - P_{i,j}^{MD}$, but as the value of ε decreases, the LJ potential becomes increasingly permissive with respect to distances shorter than σ, thus suggesting that ε should not become too small to prevent atoms from exploring unphysical conformations. Consequently, we can introduce a minimum fraction of ε_0^{intra} below, which we do not define as an attractive interaction. Given this minimum fraction f_ε and eq 2, an attractive interaction is defined if and only if $P_{i,j}^{MD} > (P_{threshold}^{RC})^{-f_\varepsilon} \max(P_{i,j}^{RC}, P_{threshold}^{RC})$. Here, we set f_ε = 0.2 by default so that ε is not less than $0.2\varepsilon_0^{intra}$. However, in this way, attractive interactions can still be defined for pairs of atoms with very small $P_{i,j}^{MD}$, for which the estimate of both $P_{i,j}^{MD}$ and r_{min} may be poor due to limited statistical sampling. We would therefore like to add an additional condition for learning attractive interactions, namely, $P_{i,j}^{MD} > P_{threshold}^{MD}$. Given a value for $P_{threshold}^{MD}$, it is possible to define

$P_{threshold}^{RC} = (P_{threshold}^{MD})^{1/1-f_\varepsilon}$ in such a way that the two conditions do not overlap. Now let us consider the case of two different training MD simulations, one for a system exploring a very homogeneous conformational ensemble and the other exploring a very heterogeneous one. Using a single $P_{threshold}^{MD}$ can lead to learning irrelevant contacts in the first case and discarding relevant ones in the second. To obtain an adaptive $P_{threshold}^{MD}$, we instead set P_{learn} as the fraction of the total contact population to learn from a training simulation. We sort all $P_{i,j}^{MD}$ in descending order and normalize them by the total contact population $\sum P_{i,j}^{MD}$, and then we take $P_{threshold}^{MD}$ as the $P_{i,j}^{MD}$ value associated with a cumulative sum equal to P_{learn}. Our default choice is P_{learn} = 0.9995, where a value too small would result in poor learning of the training simulations, while a value too large may result in learning of numerical noise.

2.4. Repulsive Interactions. Equation 2 can also lead to negative ε, as mentioned above, but LJ is not well-defined in this case. Instead, repulsive interactions can be implemented by setting $C^{(6)} = 0$ and $C^{(12)} > 0$. We derived a general formula to update our default $C^{(12)}$ (cf. Section 2.2) by observing the following approximate relationship between the probability of a contact and its interaction length r_{min} in the RC and MD simulations

$$P_{i,j}^{RC}\colon \exp(-C_{i,j}^{(12)}/r_{min_{RC}}^{12}) = P_{i,j}^{MD}\colon \exp(-\tilde{C}_{i,j}^{(12)}/r_{min_{MD}}^{12}) \quad (4)$$

where $\tilde{C}_{i,j}^{(12)}$ is the updated effective value we wish to set to reproduce the training MD simulation. From the above relationship, we can see that

$$\tilde{C}_{i,j}^{(12)} = -r_{min_{MD}}^{12}\ln\left(\frac{P_{i,j}^{MD}}{P_{i,j}^{RC}}\right) + C_{i,j}^{(12)}\left(\frac{r_{min_{MD}}}{r_{min_{RC}}}\right)^{12} \quad (5)$$

which we regularize as in eq 2 to obtain

https://doi.org/10.1021/acs.jctc.3c01182
J. Chem. Theory Comput. 2024, 20, 459−468

$$\tilde{C}_{i,j}^{(12)} = +\frac{\varepsilon_0^{intra}}{\ln P_{threshold}^{RC}} r_{min_{MD}}^{12} \ln\left(\frac{P_{i,j}^{MD}}{\max(P_{i,j}^{RC}, P_{threshold}^{RC})}\right) + C_{i,j}^{(12)}\left(\frac{r_{min_{MD}}}{r_{min_{RC}}}\right)^{12} \quad (6)$$

Equation 6 is applied whenever $P_{i,j}^{MD} < \max(P_{i,j}^{RC}, P_{threshold}^{RC})$ but greater than 0. Here, the second term of the sum allows the standard $C_{i,j}^{(12)}$ to be scaled up or down, while the first term is added on top. We also consider the case of $(P_{threshold}^{RC})^{-f_\varepsilon} \max(P_{i,j}^{RC}, P_{threshold}^{RC}) > P_{i,j}^{MD} > \max(P_{i,j}^{RC}, P_{threshold}^{RC})$; in this case, we use the $\tilde{C}_{i,j}^{(12)} = C_{i,j}^{(12)}\left(\frac{r_{min_{MD}}}{r_{min_{RC}}}\right)^{12}$ relationship because the first term would have a negative sign and could result in meaningless $C_{i,j}^{(12)}$ coefficients. Importantly, with respect to the case of attractive interactions, it is now possible for $P_{i,j}^{MD}$ and/or $P_{i,j}^{RC}$ to be less than $P_{threshold}^{MD}$, in which case the corresponding r_{min} is set to $R_{i,j}^{cut}$. Finally, to avoid unphysical interactions, the learned $C_{i,j}^{(12)}$ is limited to between 1/10 and 20 times the default one. 1−4 interactions are rescaled according to the same rules, but to avoid possible distortions of the Ramachandran space, their $C_{i,j}^{(12)}$ values cannot change more than a factor of 1.5. For all of the pairs of atoms not included in the attractive or repulsive cases, the prior $C_{i,j}^{(12)}$ is retained.

2.5. Intermolecular Interactions. Intramolecular geometries are generally compatible with intermolecular geometries, while the opposite is not true (i.e., some intermolecular contacts cannot be formed intramolecularly because of the constraints imposed by the polymer geometry). Therefore, in multi-*e*GO, when we learn an intramolecular interaction, it is applied intermolecularly, while the opposite is not true. If both intermolecular and intramolecular interactions are learned for the same pair of atoms, then both are retained and applied under the appropriate conditions. To estimate the intermolecular contact probabilities and interaction lengths, we analyze a training simulation containing N copies of a molecule for the probability that a pair of atoms form at least one intermolecular contact per molecule in each frame. This follows from the assumption that the contact strength should not depend on the coordination, i.e., the interactions are simply two-body. The interaction length for a pair is calculated using eq 3 over the distribution of the intermolecular pair distances. The interaction strength is then set according to eqs 2 and 6 with the possibility of setting $\varepsilon_0^{inter} \neq \varepsilon_0^{intra}$, but using an ad hoc prior model. The intermolecular prior model should estimate the probability of trivial intermolecular interactions resulting only from random collisions associated with the shape of the molecules and their concentration. In general, to avoid any issue in the entropic contributions of the model, our current approach is to run a simulation at the same concentration as the one we want to run in production, using a force field trained only for intramolecular interactions. It should be noted that an intermolecular prior model is only needed to learn intermolecular interactions specifically, not for the intramolecular model applied intermolecularly.

2.6. Learning from Multiple MD Simulations. A key feature of multi-*e*GO is its ability to learn from multiple MD simulations, generally associated with different free energy minima of the system. A prototypical example is that of amyloid fibrils, where we can provide for training a simulation of the monomer protein in solution, the free energy minimum at low concentration, and that of the protein in an amyloid fibril, the free energy minimum at high concentration. When merging contacts learned from different sources, we follow the following rules: (1) among multiple contacts for a given i,j pair, we chose the one with the shortest estimated interaction length as defined above (that is, for pairs with $P_{i,j} > P_{threshold}^{MD}$, we use eq 3, otherwise $r_{min} = R_{i,j}^{cut}$); (2) among several attractive and repulsive contacts for a given i,j pair with the same r_{min}, we chose the attractive one with the largest ε; and (3) among several repulsive contacts for a given i,j pair, we chose the one with the smallest $C_{i,j}^{(12)}$. Another implemented option is to set an ensemble as the check data set. Setting an ensemble as such forces multi-*e*GO to perform a check on repulsive interactions to ensure compatibility with the check data set. The $C_{i,j}^{(12)}$ of repulsive contacts for which $r_{min}^{check} < r_{min}^{train}$ are rescaled by $(r_{min}^{check}/r_{min}^{train})^{12}$.

3. SIMULATIONS DETAILS

All MD simulations were performed using the GROMACS[29] software suite. Metadynamics[30] simulations were performed using the PLUMED2 library.[31,32] Unless explicitly stated, all simulations were performed using the same 4-step protocol consisting of (1) energy minimization using the steepest descent algorithm until the maximum force converges to a value <1000 kJ mol^{-1}nm^{-1}, (2) conjugate-gradient minimization until the maximum force converges to a value <10 kJ mol^{-1} nm^{-1}, (3) positionally restrained relaxation for 4 ns at constant pressure and temperature, and (4) the production simulation. Explicit solvent MD simulations were performed using the leapfrog algorithm with a time step of 2 fs and LINCS restraints[33] for hydrogen atoms. Nonbonded interactions are cut off at 1 nm using PME for long-range electrostatics.[34] Temperature and pressure are controlled by stochastic velocity rescaling[35] and cell rescaling[36] algorithms, respectively. Multi-*e*GO simulations were performed using stochastic dynamics integration with a time step of 5 fs and a relaxation time of 25 ps. The cutoff for the LJ interactions is set specifically for each system as $2.5\sigma_{max}$. A 10% larger radius is used for the neighbor lists, which are updated every 20 steps.

All scripts and parameters to generate a multi-*e*GO force field are publicly available on GitHub (cf. Notes). All simulations performed in this work are publicly available via Zenodo (cf. Notes).

3.1. Aβ42. The training trajectories for Aβ42 are publicly available and published in ref. 22 They include 315 μs of sampling at 278 K. An RC simulation was performed at the same temperature for 1 μs. Different ε_0 values were then tested to maximize the agreement with the target radius of the gyration probability distribution until an optimal value of 0.335 kJ/mol was found. Production multi-*e*GO simulations were run in triplicate at the same temperature for 2 μs each. Clustering analyses were performed with the cluster module of GROMACS using the gromos algorithm described in ref 37 using the root-mean-square deviation (RMSD) of the backbone atom positions as a metric and a cutoff of 0.8 nm.

3.2. Protein GB1. The explicit solvent training simulation was performed using the CHARMM22* force field[28] in conjunction with the TIP3P water model.[38] A dodecahedral box was constructed 0.7 nm from the protein surface to minimize the number of explicit water molecules. The system charge was neutralized using a NaCl concentration of 0.2 mM. After energy minimization and temperature and density

https://doi.org/10.1021/acs.jctc.3c01182
J. Chem. Theory Comput. 2024, 20, 459−468

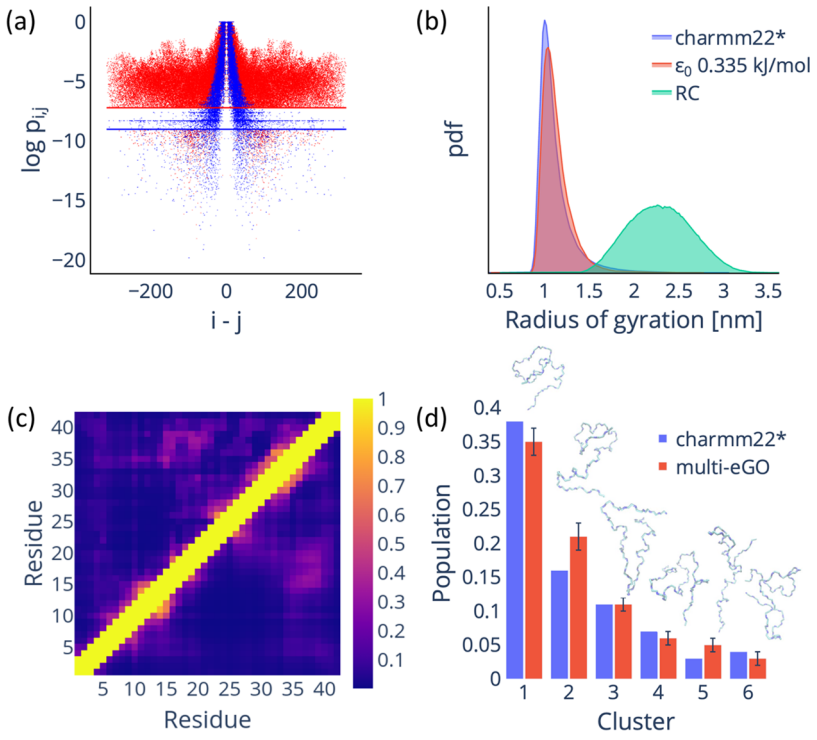


Figure 2. Reformulated multi-*e*GO model can reproduce the conformational dynamics of Aβ42 IDP. (a) Comparison of the atom-pair contact probabilities for the training (red dots) and RC (blue dots) simulations. The lines represent the $P^{MD}_{threshold}$ (red) and $P^{RC}_{threshold}$ (blue) values. (b) Comparison of the radius of gyration probability density function (calculated using the backbone atoms) for the training (blue), multi-*e*GO (orange), and RC (green) simulations. (c) The contact probability map per residue for the training (upper diagonal) and multi-*e*GO (lower diagonal) simulations. The color bar represents the contact probability. (d) Clustering analysis of multi-*e*GO and training simulations. The clusters are comparable in population size and in terms of order. A representative configuration is shown for each cluster.

equilibration, production was performed for 1 μs of simulations in the NPT ensemble (T = 300 K, P = 1 bar). The RC simulation was also performed for 1 μs at 300 K. To calibrate the multi-*e*GO energy scale ε_0, we used the melting temperature and its microscopic implications. First, we performed metadynamics simulations at the experimental melting temperature[23] T_m = 360 K to determine the ε_0 value at which the folded and unfolded states are equally populated. For these simulations, we used the all-atom RMSD with respect to the crystal structure as a collective variable, adding Gaussians every 500 steps, with an initial height of 1.2 kJ/mol, a bias factor of 15, and a width of 0.025 nm. Following the identification of an appropriate ε_0 = 0.235 kJ/mol, we prepared 200 starting configurations extracted from the RC simulation. We then ran 200 independent simulations at 300 K until the folded structure (i.e., all-atom RMSD with respect to the crystal structure of less than 0.3 nm) was reached. The data were analyzed using Biotite.[39]

3.3. $TTR_{105-115}$ Peptide. The training trajectory for $TTR_{105-115}$ was performed in our previous work[18] using the a99SB-disp force field,[40] for 1.6 μs at 300 K. The RC simulation was performed for 500 ns at the same temperature. The ε_0 for the intramolecular interactions of the multi-*e*GO simulation was tuned by maximizing the agreement of the radius of gyration probability distribution, and the best result was found for ε_0 = 0.275 kJ/mol.

A training trajectory for the $TTR_{105-115}$ fibril was performed using the 2M5M PDB structure,[41] consisting of 84 monomers, in a box containing 23,000 water molecules. The system was parametrized using the CHARMM22* force field[28] and the TIP3P water model.[38] The fibril was found to be unstable, so the simulation was run with a position restraint on all of the backbone and Cβ carbons of the system for 150 ns. To weight the intermolecular interactions, 2 μs multi-*e*GO simulations of 80 monomers at concentrations of 13, 10, and 7 mM were performed, trained only on the monomer MD, with ε_0 = 0.275 kJ/mol. The ε_0 for the intermolecular interactions was also set at 0.275 kJ/mol after a stable fibril structure was verified. Aggregation kinetics simulations were set up to generate boxes of 4000 monomers at concentrations of 13, 10, and 7 mM. Three initial configurations were generated for each concentration and first equilibrated using the monomer-only force field. Simulations were then run at 310 K and followed until aggregation.

4. RESULTS

4.1. Multi-*e*GO Can Reproduce the Conformational Ensemble of an Intrinsically Disordered Protein. In Figure 1, we have shown how our first multi-*e*GO implementation was unable to learn the heterogeneous conformational ensemble of Aβ42, as represented by its radius of gyration probability distribution and per-residue average contact map. We attribute this limitation to the imbalance between local and long-range interactions. In fact, the contact probability weighting equation heuristically introduced in our previous publication, eq 1 in the Introduction, can be rewritten as the ratio between a training probability and an uninformative uniform prior. By introducing a polymer-informed prior, cf. Section 2, we can weight each contact by its probability of forming as a consequence of the polymer geometry alone. As shown in Figure 2a, the distribution of the contact probabilities in the training simulation peaks for atoms

https://doi.org/10.1021/acs.jctc.3c01182
J. Chem. Theory Comput. 2024, 20, 459−468

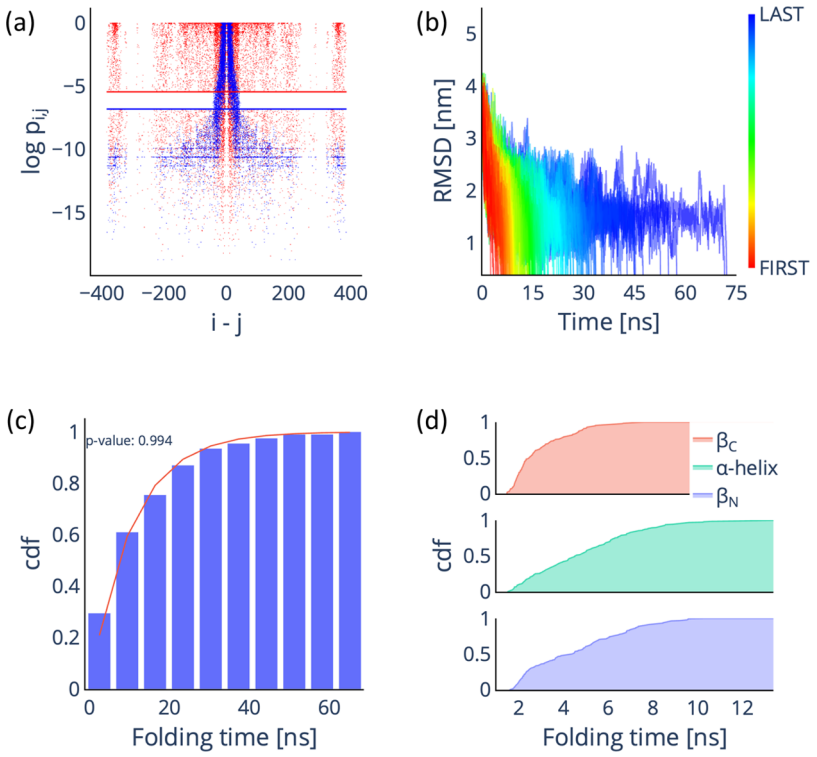


Figure 3. Multi-*e*GO can reproduce the folding mechanism of GB1. (a) Comparison of the atom-pair contact probabilities for the training (red dots) and RC (blue dots) simulations. The lines represent the $P^{MD}_{threshold}$ (red) and $P^{RC}_{threshold}$ (blue) values. (b) Time evolution of the RMSD with respect to the folded state for the 200 multi-*e*GO folding trajectories. (c) Cumulative distribution function (cdf) of the folding time distribution for the 200 multi-*e*GO folding trajectories (blue bars), fitted with the cdf of the Poisson distribution (red line) and associated *p*-value. (d) Cumulative distribution function for the folding times of the three GB1 secondary structure elements.

close in sequence, decreasing with distance, but with some regions still showing low but non-negligible values with respect to $P^{MD}_{threshold}$ (e.g., 0.0007). On the contrary, the contact probabilities in the RC simulation are equally peaked for atoms close in sequence but systematically drop to zero for atoms further apart, at some point becoming smaller than the $P^{RC}_{threshold}$ value (e.g., 0.0001). The comparison of the two distributions allows a better understanding of how the multi-*e*GO model works by immediately highlighting that some distant contacts are very important compared to the probability of their formation by chance.

In Figure 2b, we show how this multi-*e*GO reformulation allows us to improve the agreement of the radius of gyration distribution not only to overlap at the maximum, where we find most of the closed conformations, but also to show how the simulations now match the tail, i.e., the open conformations with fewer contacts. The contact map analysis shown in Figure 2c further confirms the accuracy of the model, showing a strong resemblance to the training, with an average error of around 3.5%. Finally, we performed a clustering analysis[37] on the combined trajectories of the multi-*e*GO simulation and the training to understand how the individual states are distributed. Remarkably, our analysis revealed that the six most representative clusters are equally represented in the training and multi-*e*GO conformational ensembles, as shown in Figure 2d.

Taking all of the results together, it is safe to say that the conformational ensemble sampled by multi-*e*GO is in quantitative agreement with the training ensemble used.

4.2. Multi-*e*GO Can Simulate the Folding Mechanism of a Small Protein. Having shown that multi-*e*GO can learn the conformational ensemble of an IDP, one can ask how it performs in a conventional structure-based modeling task, i.e., describing the folding mechanism of a folded protein. A system often studied by Gō models is protein GB1.[42−44] The folding of this protein has been well characterized experimentally, showing that its C-terminal hairpin folds first, followed by its N-terminal one and the α-helix.[45] To set up the model, we ran a training simulation of the folded protein and an RC simulation. Comparing the probability distribution of contact pairs in Figure 3a, we observe high probability contacts both close and farther apart in the sequence for the training simulation, as expected for a stable folded protein, while the RC simulation shows an identical behavior as previously shown for Aβ42, with highly probability contacts found only close in the sequence. After training a multi-*e*GO model to recover the experimental melting temperature of 360 K and setting ε_0 to 0.235 kJ/mol, we ran 200 independent folding simulations, starting from RC configurations, see Figure 3b. The cumulative distribution function of the folding times in Figure 3c shows a typical Poisson distribution (*p*-value of 0.994 from a Kolmogorov−Smirnov test) with an average folding time of 17.4 ns. This time is nominal and when compared with the experimental time scale of 10 ms[45] gives an idea of the speed-up achieved by multi-*e*GO.

To analyze the GB1 folding mechanism, we evaluated the folding time for each secondary structure element, i.e., the N- and C-terminal β-hairpins and the central α-helix, as the time at which each secondary structure was stably formed in the folding simulation; see Figure 3d. The analysis clearly highlights the C-terminal hairpin as the element that generally folds first. This is a way to measure the progress of the folding

https://doi.org/10.1021/acs.jctc.3c01182
J. Chem. Theory Comput. 2024, 20, 459−468

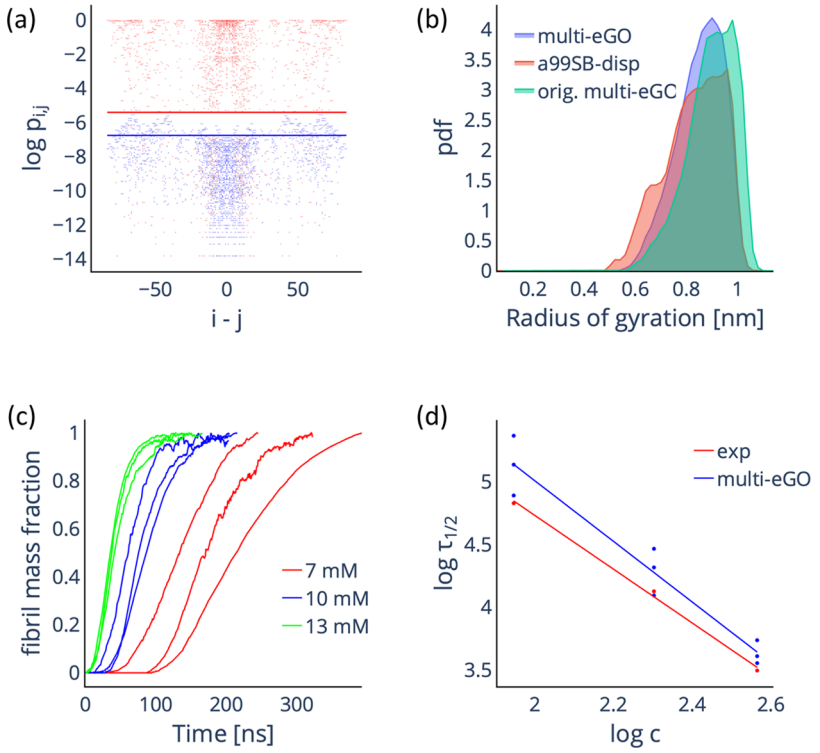


Figure 4. Multi-*e*GO simulations of $TTR_{105-115}$ aggregation as a function of the initial monomer concentration. (a) Comparison of the intermolecular contact probabilities of atom pairs for the training (red dots) and 13 mM RC (blue dots) simulations. The lines represent the $P^{MD}_{threshold}$ (red) and $P^{RC}_{threshold}$ (blue) values. (b) Comparison of the radius of the gyration probability density function (calculated using the backbone atoms) for the training monomer (red), reformulated multi-*e*GO (blue), and original multi-*e*GO (green) simulations. (c) Simulated aggregation kinetics. Curves represent the normalized number of monomers involved in an aggregate of at least 10 monomers as a function of nominal simulation time. (d) Log−log plot of the half-times as a function of the initial monomer concentration, experimental values (red) are taken from ref. 18 Both sets can be fitted by a straight line with slopes $\gamma = -2.1 \pm 0.1$ and $\gamma = -2.4 \pm 0.2$ for the experimental and simulated data, respectively.

process and clearly indicates an asymmetric folding that starts preferentially from the C-terminal and ends with either the α-helix or the N-terminal hairpin, as previously reported.[43,44,46] This result suggests that multi-*e*GO can correctly learn the native state energy of a folded protein and extrapolate about the folding mechanism.

4.3. Multi-*e*GO Can Still Qualitatively Describe the Aggregation Kinetics of $TTR_{105-115}$. In our previous work,[18] we showed that multi-*e*GO could simulate the aggregation kinetics of $TTR_{105-115}$ as a function of the initial monomer concentration, qualitatively reproducing the expected kinetics and structural features. After reformulating multi-*e*GO, we replicated these simulations. To train the model, we used the previously generated simulation of the monomer, a simulation of the fibril (with the caveat that having observed the fibril to be unstable in solution, we ran the simulation using positional restraints). We also ran an RC simulation of the monomer and three intermolecular prior simulations (at concentrations of 13, 10, and 7 mM), which are required to weight the intermolecular contacts (cf. Section 2). In Figure 4a, we compare the probabilities for the intermolecular contact pair from the training and the prior simulations. The training simulation showed many highly probable intermolecular contacts due to the stable fibril conformation. On the contrary, the prior simulation displayed low probability contacts resulting from the random collisions.

From the training and RC simulations, we set ε_0^{intra} to 0.275 kJ/mol by maximizing the agreement between the radius of gyration probability distribution for the monomer training and multi-*e*GO simulations. In Figure 4b, we show the overlap between the radius of gyration probability distributions for the training, original, and reformulated multi-*e*GO simulations. It is apparent how the reformulated multi-*e*GO better reproduces the training simulation. Next, we verified that the same value can be used for ε_0^{inter}, resulting in a stable fibril conformation. Finally, aggregation kinetics were simulated in triplicate with starting monomer concentrations of 13, 10, and 7 mM using 4000 monomers. In Figure 4c, we plotted the fraction of fibril mass as the normalized number of monomers involved in an aggregate of at least 10 monomers as a function of the simulation time. The curves showed the expected sigmoidal shape with an increasing lag time with a decreasing concentration. To compare simulations and experiments, we calculated the log−log of both aggregation half-times and concentrations. Both the experimental and simulation data show a linear trend with comparable slopes of -2.1 ± 0.1 and -2.4 ± 0.2, indicating macroscopically comparable kinetics. As in our previous work, the resulting fibrils lack the central cavity while exhibiting correct antiparallel stacking and head-to-tail lateral growth, allowing us to confirm that the reformulated model can still qualitatively describe the aggregation of the $TTR_{105-115}$ peptide.[18,41]

Simplified models for biomolecular simulations have been developed to overcome the time scale and size limitations of conventional molecular mechanics MD. Structure-based models, often at α-carbon resolution, have been used mainly to study not only protein folding[16] but also large conformational changes,[47−49] metamorphic proteins,[50,51] and the folding

https://doi.org/10.1021/acs.jctc.3c01182
J. Chem. Theory Comput. 2024, 20, 459−468

upon binding of disordered proteins with different partners.[52,53] With multi-*e*GO, we aim to develop a platform that, building on the increasing availability of high-quality MD simulations (see, e.g., ref 54), can then be used to study processes involving multiple molecules and long time scales while maintaining atomistic resolution and, indirectly, some chemical specificity. This work, by introducing a well-defined theoretical framework for learning from both homogeneous and heterogeneous conformational ensembles, is our second step in this direction.

AUTHOR INFORMATION

Corresponding Author

Carlo Camilloni − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy;* orcid.org/0000-0002-9923-8590; Email: carlo.camilloni@unimi.it

Authors

Fran Bačić Toplek − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy;* orcid.org/0000-0002-8331-0885

Emanuele Scalone − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy; Department of Chemistry, Dartmouth College, Hanover, New Hampshire 03755, United States;* orcid.org/0000-0003-4271-4856

Bruno Stegani − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy*

Cristina Paissoni − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy*

Riccardo Capelli − *Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy;* orcid.org/0000-0001-9522-3132

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c01182

Author Contributions

[§]F.B.T. and E.S. contributed equally to this work. All authors contributed to the development of the model theory. F.B.T., E.S., and B.S. wrote the code to generate the multi-*e*GO models. F.B.T., E.S., and C.C. performed and analyzed the simulations. F.B.T. and C.C. wrote the manuscript with contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

Simulations data are publicly available via Zenodo with record 10.5281/zenodo.10033341. The multi-*e*GO code and parameters are publicly available on GitHub at https://github.com/multi-ego/multi-eGO. The beta1 tag is used for a snapshot of the repository on the date of paper submission.

ACKNOWLEDGMENTS

Funding was provided to C.C. by the University of Milano—Linea 1; C.C. is also supported by Fondazione Telethon (Grant GGP19134). The authors acknowledge CINECA for an award under the ISCRA initiative for the availability of high-performance computing resources and support. The authors acknowledge Thomas Löhr, Guido Tiana, and Michele Vendruscolo for their suggestions and insights.

ABBREVIATIONS

MD, molecular dynamics; LJ, Lennard−Jones; RC, random coil; Aβ42, amyloid β42 peptide; GB1, B1 immunoglobulin-binding domain of streptococcal protein G; TTR, transthyretin; PDB, Protein Data Bank; RMSD, root-mean-square deviation; pdf, probability density function; cdf, cumulative distribution function

REFERENCES

(1) Macuglia, D.; Roux, B.; Ciccotti, G. The Emergence of Protein Dynamics Simulations: How Computational Statistical Mechanics Met Biochemistry. *Eur. Phys. J. H* **2022**, *47* (1), No. 13.

(2) Camilloni, C.; Pietrucci, F. Advanced Simulation Techniques for the Thermodynamic and Kinetic Characterization of Biological Systems. *Adv. Phys.: X* **2018**, *3* (1), No. 1477531.

(3) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114* (13), 6589−6631.

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

(5) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373*, No. eabj8754.

(6) Robustelli, P.; Ibanez-de-Opakua, A.; Campbell-Bezat, C.; Giordanetto, F.; Becker, S.; Zweckstetter, M.; Pan, A. C.; Shaw, D. E. Molecular Basis of Small-Molecule Binding to A-Synuclein. *J. Am. Chem. Soc.* **2022**, *144* (6), 2501−2510.

(7) Galvanetto, N.; Ivanović, M. T.; Chowdhury, A.; Sottini, A.; Nüesch, M. F.; Nettels, D.; Best, R. B.; Schuler, B. Extreme Dynamics in a Biomolecular Condensate. *Nature* **2023**, *619* (7971), 876−883.

(8) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **2020**, *6* (10), 1722−1734.

(9) Borges-Araújo, L.; Patmanidis, I.; Singh, A. P.; Santos, L. H. S.; Sieradzan, A. K.; Vanni, S.; Czaplewski, C.; Pantano, S.; Shinoda, W.; Monticelli, L.; Liwo, A.; Marrink, S. J.; Souza, P. C. T. Pragmatic Coarse-Graining of Proteins: Models and Applications. *J. Chem. Theory Comput.* **2023**, *19*, 7112−7135.

(10) Souza, P. C. T.; Alessandri, R.; Barnoud, J.; Thallmair, S.; Faustino, I.; Grünewald, F.; Patmanidis, I.; Abdizadeh, H.; Bruininks, B. M. H.; Wassenaar, T. A.; Kroon, P. C.; Melcr, J.; Nieto, V.; Corradi, V.; Khan, H. M.; Domański, J.; Javanainen, M.; Martinez-Seara, H.; Reuter, N.; Best, R. B.; Vattulainen, I.; Monticelli, L.; Periole, X.; Tieleman, D. P.; de Vries, A. H.; Marrink, S. J. Martini 3: A General Purpose Force Field for Coarse-Grained Molecular Dynamics. *Nat. Methods* **2021**, *18* (4), 382−388.

(11) Tesei, G.; Schulze, T. K.; Crehuet, R.; Lindorff-Larsen, K. Accurate Model of Liquid−Liquid Phase Behavior of Intrinsically

Disordered Proteins from Optimization of Single-Chain Properties. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (44), No. e2111696118.
(12) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. Sequence Determinants of Protein Phase Behavior from a Coarse-Grained Model. *PLoS Comput. Biol.* **2018**, *14* (1), No. e1005941.
(13) Benayad, Z.; von Bülow, S.; Stelzl, L. S.; Hummer, G. Simulation of FUS Protein Condensates with an Adapted Coarse-Grained Model. *J. Chem. Theory Comput.* **2021**, *17* (1), 525−537.
(14) Dannenhoffer-Lafage, T.; Best, R. B. A Data-Driven Hydrophobicity Scale for Predicting Liquid−Liquid Phase Separation of Proteins. *J. Phys. Chem. B* **2021**, *125* (16), 4046−4056.
(15) Majewski, M.; Pérez, A.; Thölke, P.; Doerr, S.; Charron, N. E.; Giorgino, T.; Husic, B. E.; Clementi, C.; Noé, F.; Fabritiis, G. D. Machine Learning Coarse-Grained Potentials of Protein Thermodynamics. *Nat. Commun.* **2023**, *14* (1), No. 5739.
(16) Takada, S. Gō Model Revisited. *Biophys. Physicobiol.* **2019**, *16* (0), 248−255.
(17) Baldwin, A. J.; Knowles, T. P. J.; Tartaglia, G. G.; Fitzpatrick, A. W.; Devlin, G. L.; Shammas, S. L.; Waudby, C. A.; Mossuto, M. F.; Meehan, S.; Gras, S. L.; Christodoulou, J.; Anthony-Cahill, S. J.; Barker, P. D.; Vendruscolo, M.; Dobson, C. M. Metastability of Native Proteins and the Phenomenon of Amyloid Formation. *J. Am. Chem. Soc.* **2011**, *133* (36), 14160−14163.
(18) Scalone, E.; Broggini, L.; Visentin, C.; Erba, D.; Toplek, F. B.; Peqini, K.; Pellegrino, S.; Ricagno, S.; Paissoni, C.; Camilloni, C. Multi-EGO: An in Silico Lens to Look into Protein Aggregation Kinetics at Atomic Resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119* (26), No. e2203181119.
(19) Huang, W.; Lin, Z.; van Gunsteren, W. F. Validation of the GROMOS 54A7 Force Field with Respect to β-Peptide Folding. *J. Chem. Theory Comput.* **2011**, *7* (5), 1237−1243.
(20) Glenner, G. G.; Wong, C. W. Alzheimer's Disease: Initial Report of the Purification and Characterization of a Novel Cerebrovascular Amyloid Protein. *Biochem. Biophys. Res. Commun.* **1984**, *120* (3), 885−890.
(21) Hamley, I. W. The Amyloid Beta Peptide: A Chemist's Perspective. Role in Alzheimer's and Fibrillization. *Chem. Rev.* **2012**, *112* (10), 5147−5192.
(22) Löhr, T.; Kohlhoff, K.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. A Kinetic Ensemble of the Alzheimer's Aβ Peptide. *Nat. Comput. Sci.* **2021**, *1* (1), 71−78.
(23) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. A Novel, Highly Stable Fold of the Immunoglobulin Binding Domain of Streptococcal Protein G. *Science* **1991**, *253* (5020), 657−661.
(24) Gustavsson, Å.; Engström, U.; Westermark, P. Normal Transthyretin and Synthetic Transthyretin Fragments from Amyloid-like Fibrils in Vitro. *Biochem. Biophys. Res. Commun.* **1991**, *175* (3), 1159−1164.
(25) Hamelryck, T.; Borg, M.; Paluszewski, M.; Paulsen, J.; Frellsen, J.; Andreetta, C.; Boomsma, W.; Bottaro, S.; Ferkinghoff-Borg, J. Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized. *PLoS One* **2010**, *5* (11), No. e13714.
(26) Ho, B. K.; Thomas, A.; Brasseur, R. Revisiting the Ramachandran Plot: Hard-sphere Repulsion, Electrostatics, and H-bonding in the A-helix. *Protein Sci.* **2003**, *12* (11), 2508−2522.
(27) Ho, B. K.; Brasseur, R. The Ramachandran Plots of Glycine and Pre-Proline. *BMC Struct. Biol.* **2005**, *5* (1), No. 14.
(28) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47−L49.
(29) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1*−*2*, 19−25.
(30) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12562−12566.
(31) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604−613.
(32) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; Capelli, R.; Carloni, P.; Ceriotti, M.; Cesari, A.; Chen, H.; Chen, W.; Colizzi, F.; De, S.; Pierre, M. D. L.; Donadio, D.; Drobot, V.; Ensing, B.; Ferguson, A. L.; Filizola, M.; Fraser, J. S.; Fu, H.; Gasparotto, P.; Gervasio, F. L.; Giberti, F.; Gil-Ley, A.; Giorgino, T.; Heller, G. T.; Hocky, G. M.; Iannuzzi, M.; Invernizzi, M.; Jelfs, K. E.; Jussupow, A.; Kirilin, E.; Laio, A.; Limongelli, V.; Lindorff-Larsen, K.; Löhr, T.; Marinelli, F.; Martin-Samos, L.; Masetti, M.; Meyer, R.; Michaelides, A.; Molteni, C.; Morishita, T.; Nava, M.; Paissoni, C.; Papaleo, E.; Parrinello, M.; Pfaendtner, J.; Piaggi, P.; Piccini, G.; Pietropaolo, A.; Pietrucci, F.; Pipolo, S.; Provasi, D.; Quigley, D.; Raiteri, P.; Raniolo, S.; Rydzewski, J.; Salvalaglio, M.; Sosso, G. C.; Spiwok, V.; Šponer, J.; Swenson, D. W. H.; Tiwary, P.; Valsson, O.; Vendruscolo, M.; Voth, G. A.; White, A. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16* (8), 670−673.
(33) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput* **2008**, *4* (1), 116−122.
(34) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577−8593.
(35) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), No. 014101.
(36) Bernetti, M.; Bussi, G. Pressure Control Using Stochastic Cell Rescaling. *J. Chem. Phys.* **2020**, *153* (11), No. 114107.
(37) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem., Int. Ed.* **1999**, *38* (1−2), 236−240.
(38) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926−935.
(39) Kunzmann, P.; Hamacher, K. Biotite: A Unifying Open Source Computational Biology Framework in Python. *BMC Bioinf.* **2018**, *19* (1), No. 346.
(40) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115* (21), E4758−E4766.
(41) Fitzpatrick, A. W. P.; Debelouchina, G. T.; Bayro, M. J.; Clare, D. K.; Caporini, M. A.; Bajaj, V. S.; Jaroniec, C. P.; Wang, L.; Ladizhansky, V.; Müller, S. A.; MacPhee, C. E.; Waudby, C. A.; Mott, H. R.; Simone, A. D.; Knowles, T. P. J.; Saibil, H. R.; Vendruscolo, M.; Orlova, E. V.; Griffin, R. G.; Dobson, C. M. Atomic Structure and Hierarchical Assembly of a Cross-β Amyloid Fibril. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (14), 5468−5473.
(42) Karanicolas, J.; Brooks, C. L. Improved Gō-like Models Demonstrate the Robustness of Protein Folding Mechanisms Towards Non-Native Interactions. *J. Mol. Biol.* **2003**, *334* (2), 309−325.
(43) Karanicolas, J.; Brooks, C. L. The Origins of Asymmetry in the Folding Transition States of Protein L and Protein G. *Protein Sci.* **2002**, *11* (10), 2351−2361.
(44) Sutto, L.; Tiana, G.; Broglia, R. A. Sequence of Events in Folding Mechanism: Beyond the Gō Model. *Protein Sci.* **2006**, *15* (7), 1638−1652.
(45) McCallister, E. L.; Alm, E.; Baker, D. Critical Role of β-Hairpin Formation in Protein G Folding. *Nat. Struct. Biol.* **2000**, *7* (8), 669−673.
(46) Camilloni, C.; Broglia, R. A.; Tiana, G. Hierarchy of Folding and Unfolding Events of Protein G, CI2, and ACBP from Explicit-Solvent Simulations. *J. Chem. Phys.* **2011**, *134* (4), No. 045105.
(47) Okazaki, K.-i.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. Multiple-Basin Energy Landscapes for Large-Amplitude Conformational Motions of Proteins: Structure-Based Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (32), 11844−11849.

https://doi.org/10.1021/acs.jctc.3c01182
J. Chem. Theory Comput. 2024, 20, 459−468

(48) Best, R. B.; Chen, Y.-G.; Hummer, G. Slow Protein Conformational Dynamics from Multiple Experimental Structures: The Helix/Sheet Transition of Arc Repressor. *Structure* **2005**, *13* (12), 1755−1763.
(49) Sutto, L.; Mereu, I.; Gervasio, F. L. A Hybrid All-Atom Structure-Based Model for Protein Folding and Large Scale Conformational Transitions. *J. Chem. Theory Comput* **2011**, *7* (12), 4208−4217.
(50) Sutto, L.; Camilloni, C. From A to B: A Ride in the Free Energy Surfaces of Protein G Domains Suggests How New Folds Arise. *J. Chem. Phys.* **2012**, *136* (18), No. 185101.
(51) Camilloni, C.; Sutto, L. Lymphotactin: How a Protein Can Adopt Two Folds. *J. Chem. Phys.* **2009**, *131* (24), No. 245105.
(52) Ganguly, D.; Chen, J. Topology-based Modeling of Intrinsically Disordered Proteins: Balancing Intrinsic Folding and Intermolecular Interactions. *Proteins: Struct., Funct., Bioinf.* **2011**, *79* (4), 1251−1266.
(53) Knott, M.; Best, R. B. Discriminating Binding Mechanisms of an Intrinsically Disordered Protein via a Multi-State Coarse-Grained Model. *J. Chem. Phys.* **2014**, *140* (17), No. 175102.
(54) Tiemann, J. K. S.; Szczuka, M.; Bouarroudj, L.; Oussaren, M.; Garcia, S.; Howard, R. J.; Delemotte, L.; Lindahl, E.; Baaden, M.; Lindorff-Larsen, K.; Chavent, M.; Poulain, P. MDverse: Shedding Light on the Dark Matter of Molecular Dynamics Simulations *ChemRxiv* 2023, DOI: 10.7554/elife.90061.

https://doi.org/10.1021/acs.jctc.3c01182
J. Chem. Theory Comput. 2024, 20, 459−468