Hierarchical Ensembles of Intrinsically Disordered Proteins at Atomic Resolution in Molecular Dynamics Simulations

Lisa M. Pietrek,[†,¶] Lukas S. Stelzl,[†,¶] and Gerhard Hummer[*,†,‡]

[†]*Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue Straße 3, 60438 Frankfurt am Main, Germany*

[‡]*Institute for Biophysics, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany*

[¶]*These authors contributed equally*

E-mail: gerhard.hummer@biophys.mpg.de

Phone: +49 69 6303-2501

Abstract

Intrinsically disordered proteins (IDPs) constitute a large fraction of the human proteome and are critical in the regulation of cellular processes. A detailed understanding of the conformational dynamics of IDPs could help to elucidate their roles in health and disease. However the inherent flexibility of IDPs makes structural studies and their interpretation challenging. Molecular dynamics (MD) simulations could address this challenge in principle, but inaccuracies in the simulation models and the need for long simulations have stymied progress. To overcome these limitations, we adopt an hierarchical approach that builds on the "flexible meccano" model of Bernadó et al. (J. Am. Chem. Soc. 2005, 127, 17968-17969). First, we exhaustively sample small IDP fragments in all-atom simulations to capture local structure. Then, we assemble the fragments into full-length IDPs to explore the stereochemically possible

global structures of IDPs. The resulting ensembles of three-dimensional structures of full-length IDPs are highly diverse, much more so than in standard MD simulation. For the paradigmatic IDP α-synuclein, our ensemble captures both local structure, as probed by nuclear magnetic resonance (NMR) spectroscopy, and its overall dimension, as obtained from small-angle X-ray scattering (SAXS) in solution. By generating representative and meaningful starting ensembles, we can begin to exploit the massive parallelism afforded by current and future high-performance computing resources for atomic-resolution characterization of IDPs.

INTRODUCTION

Proteins with intrinsically disordered regions constitute a large fraction of the human proteome.[1] Many proteins feature disordered regions besides folded domains, while other proteins are completely unstructured. Some intrinsically disordered proteins (IDPs) transiently sample structures and some fold upon binding partners, while others remain unfolded even in an ultrahigh-affinity complex.[2] Disordered regions and IDPs play essential roles in cell-signaling,[3] where their flexibility may be vital. Assembly of IDPs in biomolecular condensates formed by liquid-liquid phase separation may be a general organizing principle in cell biology.[4] Dysregulation of liquid-liquid phase separation and aggregation of IDPs may be the pathological mechanism in many neurological diseases. The paradigm for IDPs is arguably defined by α-synuclein (aS).[5] aS was suggested to also adopt α-helical conformations throughout the whole sequence in solution in its monomeric state but recent work suggests that it is better described as a disordered random coil.[6]

Resolving the structural ensembles of IDPs in experiments is challenging due to their inherent flexibility. Nuclear magnetic resonance (NMR) studies can detect residual structures or regions in IDPs having a propensity to transiently adopt well-defined structures.[7] Chemical shifts can provide information on secondary structure elements,[8] which can be transiently populated, or the lack thereof as for aS.[6] NMR J-couplings are a sensitive probe

of local and in particular backbone structure.[9] Small-angle X-ray scattering (SAXS) can complement NMR experiments by reporting on the global structures of IDPs.[10]

Generating representative structural ensembles to interpret experiments remains difficult. Data-driven models on the basis of PDB statistics, so-called coil models, have been successfully used to study disordered proteins.[11–14] Coil models provided first insights into the molecular structure of unfolded states of proteins and IDPs. Such models, including the *flexible-meccano*[14,15] model, have been used to interpret NMR data and also solution scattering data. NMR data can be rigorously incorporated into coil models.[6,16,17] However, modeling based on the statistics of the ϕ and ψ backbone dihedral angles does not capture correlations between different degrees of freedom, e.g., between backbone and sidechain conformations. Coil models, while capturing the overall flexibility of IDPs, are essentially static and typically do not give insight into the dynamics of IDPs.

Molecular dynamics (MD) simulations can capture these correlations and hold the promise to resolve the structure and dynamics of IDPs with atomistic resolution. MD simulations are highly complementary to experiments.[18–20] However, two critical issues have stymied the full power of MD simulations: (1) Inaccuracies in the force fields and (2) the inherent slow dynamics of IDPs. A myriad of shallow free energy minima for an IDP mean that any imbalance in force field will be amplified,[21] resulting in heavily skewed conformational ensembles. Due to the countless minima, simulations will relax slowly and only a fraction of the conformational space will be visited in a typical MD simulation. Force fields for IDPs have seen a lot of development.[22–24] In particular, dispersion-corrected water models[25] have led to better solvation and more realistic simulations of IDPs. Simulations of small disordered systems demonstrated that local structure can be captured very well in all-atom molecular dynamics simulations with explicit solvent.[26,27] In comparison to the dramatic improvements in hardware and software, less progress has been made on overcoming the issue of slow relaxation of IDPs associated with their large conformational entropy. Their inherent disorder and the resulting difficulty in describing their structural states and relatively large size render

applications of enhanced sampling methods[28] such as umbrella sampling, metadynamics[29] and replica exchange molecular dynamics challenging. These methods are tailored primarily to overcome energetic barriers, and less to sample the vast and weakly structured energy landscape of an IDP.

A promising avenue to overcome the sampling limitations in molecular simulations of IDPs would be to judiciously start simulations from representative starting configurations. By choosing relevant starting configurations, rather than a single starting structure, as is typically done, one can (1) obtain much better overall sampling[30] for a given amount of computer time and (2) exploit the parallelism afforded by large-scale computing resources[31] such as Folding@Home, cloud computing and supercomputers. Running many appropriately initialized simulations rather than a single long simulation makes better use of available computing resources, by overcoming limits in the scaling of MD engines to a large number of cores. For simulations of folded proteins, automated ways to generate simulations from all available experimental structures and homology models based on experimental structures of related sequences have been developed.[32] Alternatively, enhanced sampling simulations have been used to generate useful starting configurations for MD simulations.[33] In simulations of disordered polymer melts[34,35] and biological membranes,[36] multi-scale approaches have proved to be very successful. Coarse-grained simulations are used to explore the space of possible arrangements. Equilibrated structures from coarse-grained simulations can then be used as starting points for simulations with more accurate all-atom force fields.

An efficient way of sampling possible three-dimensional arrangements of polymer chains is provided by chain-growth Monte Carlo algorithms.[37] In chain-growth algorithms, the polymer chain is assembled from structures of fragments of the full-length chain. The fragments, which are small by construction, can be sampled with accurate but computationally expensive methods. Many different full-length arrangements[38] can then be sampled by using, e.g., a coarse potential energy function. In a pioneering application of fragment assembly, Stultz et al. created ensembles of tau[39] and aS[40] biased towards NMR and SAXS data.

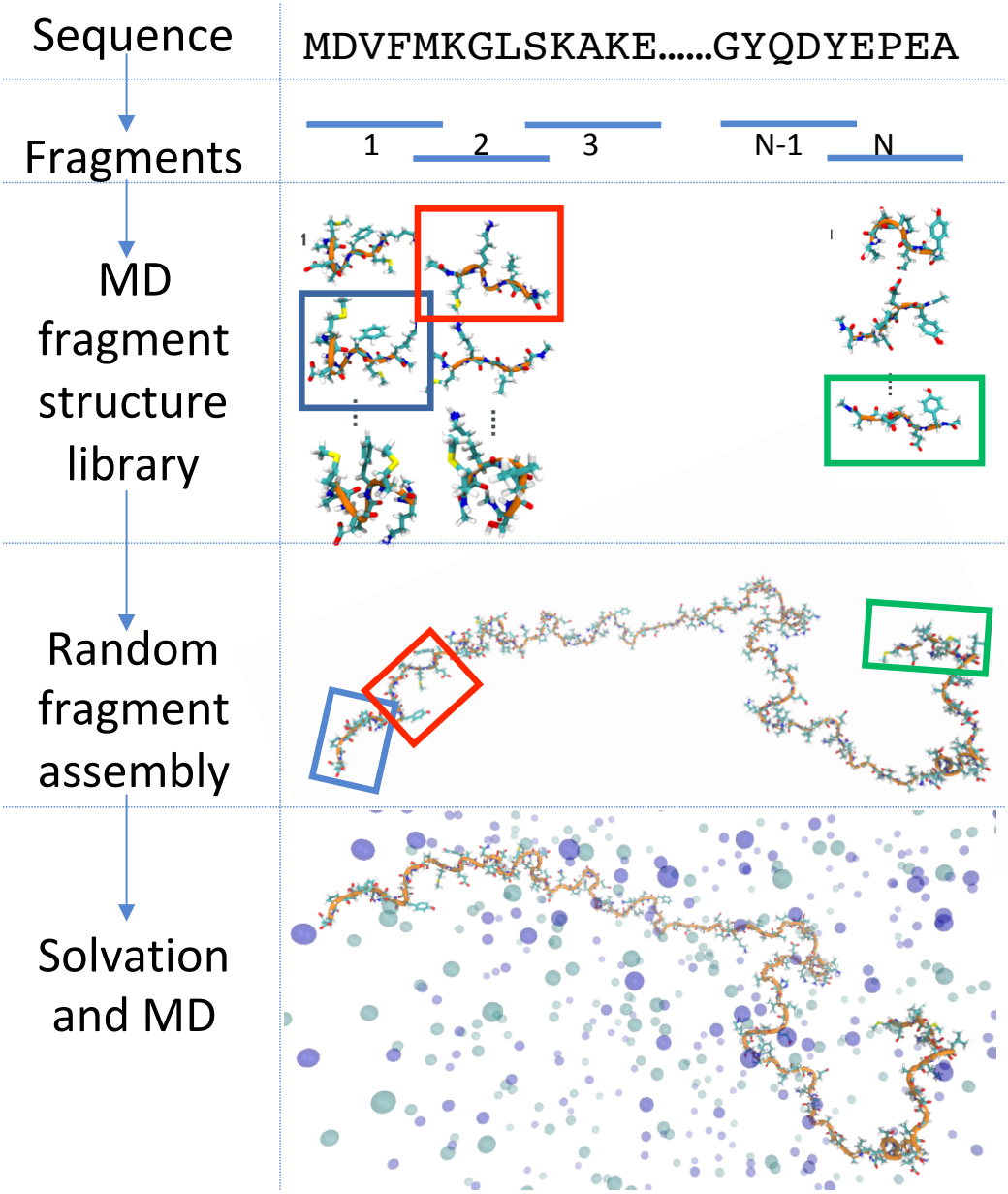

Figure 1: Overview of the hierarchical chain-growth approach to construct models of disordered proteins with atomic resolution in MD simulations. The sequence of the full-length protein is split into overlapping fragments, which can be sampled extensively. Fragment structures are assembled in an hierarchical manner by chain-growth Monte Carlo, sampling the space of possible global structures. The resulting full-length structures have atomic resolution and serve as starting points for highly parallel all-atom MD simulations.

Here, we adopt an hierarchical algorithm to create large ensembles of full-length IDP structures. These structures can be used as starting points for MD simulations and for ensemble refinement against experimental data.[17,27] We first perform all-atom MD to create extensive ensembles of fragment structures. We then merge the structures of fragments overlapping along the sequence to assemble full-length structures. The number of assembly steps grows only as the logarithm of the IDP length. This logarithmic dependence and the imposition of steric exclusion at every step of the assembly ensures a high computational efficiency of the hierarchical assembly. We show that the resulting ensembles are much more diverse than ensembles sampled in MD runs at comparable computational cost. Moreover, each structure entering the ensemble provides an excellent starting point for large scale parallel all-atom simulations on high performance computing (HPC) resources. Interestingly our ensembles agree well with high-resolution information from experiment without further refinement, emphasizing that our ensembles are useful for a direct structural analysis and as starting points for MD.

THEORY

Self-Avoiding Random Walk. Chain-growth Monte Carlo affords to connect accurate descriptions of local structure with exhaustive sampling of global structure[37,38] (Figure 1). Consider recursive growth of a heteropolymer chain with steric exclusion from a set of fragments. Note that interactions other than excluded volume could be considered,[38] but for simplicity we focus on steric exclusion. Let i_k be the index of the fragment structure at segment k of the polymer and c_k the number of such fragments. Then the structure of the polymer up to step n is uniquely described by the sequence $(i_1 i_2 \dots i_n)$ with $i_k \in \{1, \dots, c_k\}$. Note that for a heteropolymer the fragments will be different.

The partition function of the polymer composed of N fragments is

$$Z = \sum_{i_1=1}^{c_1} \sum_{i_2=1}^{c_2} \dots \sum_{i_N=1}^{c_N} \theta(i_1 i_2 \dots i_N) \tag{1}$$

where $\theta(i_1 i_2 \dots i_N) = 1$ if the chain is sterically permitted and $\theta(i_1 i_2 \dots i_N) = 0$ otherwise. The probability of a particular configuration $i_1 i_2 \dots i_N$ is then

$$p(i_1 i_2 \dots i_N) = \frac{\theta(i_1 i_2 \dots i_N)}{Z} \tag{2}$$

Merging Chains from a Non-Overlapping Sub-Ensemble Using an Hierarchical Approach. For long chains we take advantage of the fact that the problem is hierarchical. We first define a segment partition function,

$$Z_{m:n} = \sum_{i_m=1}^{c_m} \sum_{i_{m+1}=1}^{c_{m+1}} \dots \sum_{i_n=1}^{c_n} \theta(i_m \dots i_n) \tag{3}$$

for $1 \leq m < n \leq N$ such that $Z_{1:N} \equiv Z_N$. The probabilities of two sub-chains of length k and $N-k$ with $1 < k < N$ with fragment sequences $(i_1 i_2 ... i_k)$ and $(i_{k+1} i_{k+2} ... i_N)$ can be written as

$$p_1^k(i_1 i_2 \dots i_k) = \frac{\theta(i_1 i_2 \dots i_k)}{Z_{1:k}} \tag{4}$$

and

$$p_{k+1}^N(i_{k+1} i_{k+2} \dots i_N) = \frac{\theta(i_{k+1} i_{k+2} \dots i_N)}{Z_{k+1:N}} \tag{5}$$

where $p_1^N(\dots) \equiv p(\dots)$ as defined above.

In the merge step, one conformation each from the ensembles of clash-free sub-chains is drawn with probabilities $p_1^k(i_1 i_2 \dots i_k)$ and $p_{k+1}^N(i_{k+1} i_{k+2} \dots i_N)$, respectively. The two sub-chains are merged if this does not lead to a clash, i.e., if $\theta(i_1 i_2 \dots i_k, i_{k+1} i_{k+2} \dots i_N) = 1$. The

normalized probability of the merged chain then satisfies

$$\begin{aligned} & q(i_1 \dots i_k : i_{k+1} \dots i_N) \\ & = \frac{p^{k_1}(i_1 \dots i_k) p^N(i_{k+1} \dots i_N) \theta(i_1 \dots i_k i_{k+1} \dots i_N)}{\sum_{i'_1=1}^{c_1} \dots \sum_{i'_N=1}^{c_N} p_1^k(i'_1 \dots i'_k) p_1^k(i'_{k+1} \dots i'_N) \theta(i'_1 \dots i'_k i'_{k+1} \dots i'_N)} \\ & \propto \theta(i_1 i_2 \dots i_k) \theta(i_{k+1} i_{k+2} \dots i_N) \theta(i_1 i_2 \dots i_k i_{k+1} i_{k+2} \dots i_N) \\ & \propto \theta(i_1 i_2 \dots i_k, i_{k+1} i_{k+2} \dots i_N) \end{aligned} \tag{6}$$

In deriving the final probability we exploited the fact that if $\theta(i_1 i_2 ... i_k, i_{k+1} ... i_N) = 1$ then $\theta(i_1 i_2 ... i_k) = \theta(i_k + 1 i_{k+2} ... i_N) = 1$. I.e., if there is no clash in the full-length chain, there cannot be a clash in any of its sub-chains. From eq 6 it follows that, properly normalized,

$$\begin{aligned} q(i_1 \dots i_k : i_{k+1} \dots i_N) &= \frac{\theta(i_1 \dots i_N)}{\sum_{i'_1=1}^{c_1} \dots \sum_{i'_N=1}^{c_N} \theta(i'_1 \dots \dots i'_N)} \\ &= \frac{\theta(i_1 \dots i_N)}{Z_N} \equiv p(i_1 \dots i_N) \end{aligned} \tag{7}$$

I.e., if two chains are picked at random from the $1:k$ and $k+1:N$ ensembles, respectively, and then merged without a clash, they enter the $1:N$ ensemble with uniform weight. In essence, merging two sub-chains without re-weighting is possible because $\theta \in \{0, 1\}$. If we instead had weight factors defined by a Boltzmann factor for a potential energy that varied continuously, instead of assuming only values $U \in \{0, \infty\}$, then we would need to reweight the merged chains.[38]

Hierarchical Algorithm to Generate Self-Avoiding Random Walks. The ability to merge sub-chains makes it possible to grow chains hierarchically. This procedure is particularly efficient if the number of fragments N is a power of 2, i.e., $N = 2^M$ with M integer. We can then obtain properly weighted chains of length $2L$ by merging chains of length L and checking for clashes. Starting with monomers ($L = 1$) we end up with full-length chains after M steps. We first create ensembles of sterically allowed pairs $(i_1 i_2)$, $(i_3 i_4)$, ... $(i_{N-1} i_N)$

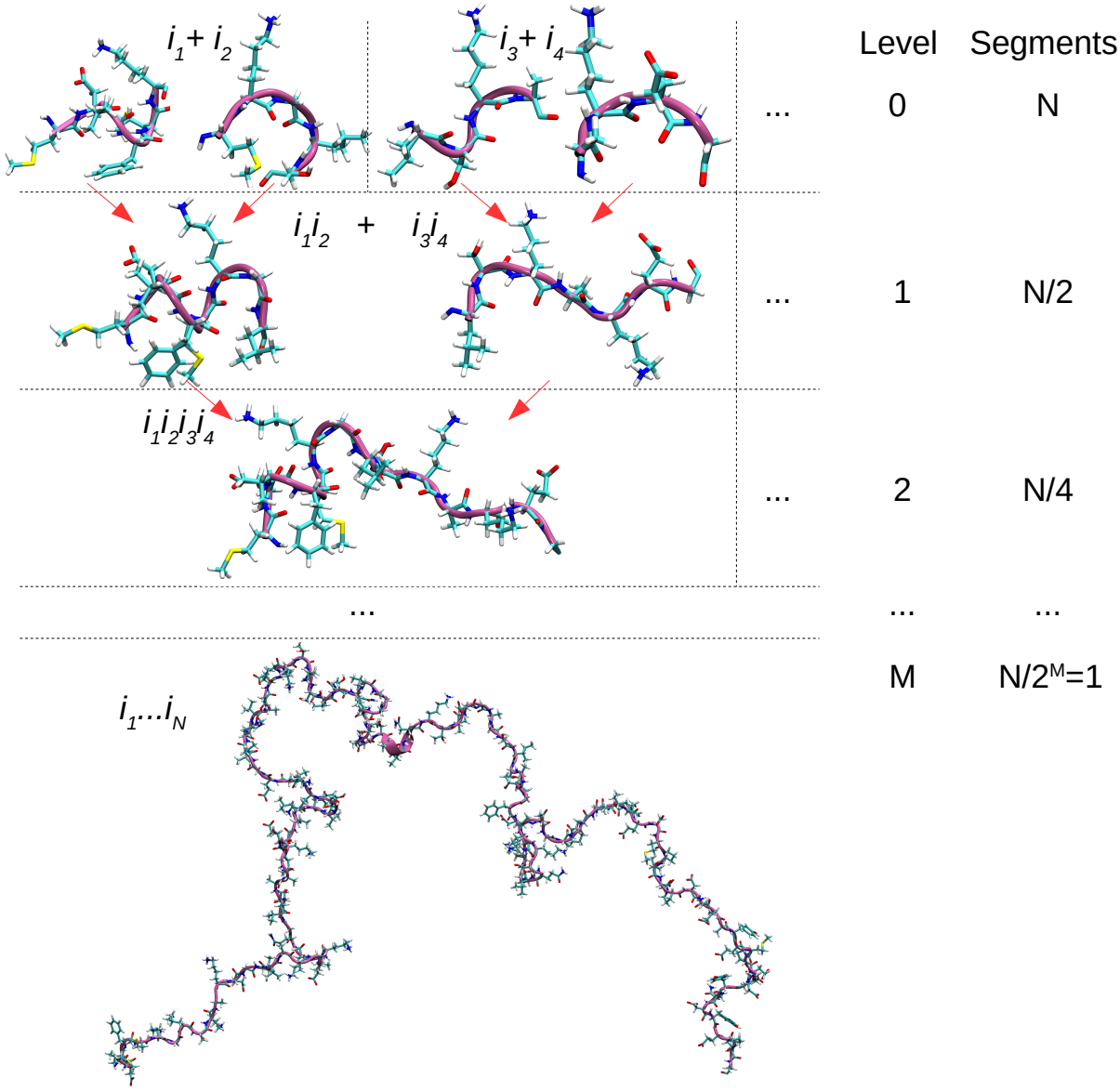

Figure 2: Generating all-atom IDP structures using an hierarchical Monte Carlo chain-growth algorithm. Level 0: To generate input for the chain growth, the full-length protein is split into N fragments, which are thoroughly sampled in all-atom MD simulations. Here, each fragment has an overlap of two residues with the subsequent fragment. Structures from the fragment libraries are picked at random to generate pairs of fragments. N depends on the chain length with $N \leq 2^M$, where M is an integer (here: $N = 2^M$). Level 1: Create ensembles of $N/2$ quadruplets. Level 2: Create ensemble of $N/4$ quadruplets and so on until, at level M, one arrives at a full-length structure of the IDP. In each fragment-assembly step, we enforce excluded-volume interactions.

(Figure 2A). Then we create ensembles of allowed quadruplets as pairs of allowed pairs, $(i_1 \dots i_4) = ((i_1 i_2), (i_3 i_4))$ etc., checked for steric clashes (Figure 2B). From these ensembles, we create ensembles of sterically allowed octuplets (Figure 2C) as pairs of sterically allowed quadruplets. At the M-th step, we obtain the final structures with the proper weight in the ensemble of sterically allowed structures (Figure 2D). Chains can also be merged hierarchically if N is not a power of 2. Instead of factorizing by larger prime numbers and merging, say, triplets, here we consistently merge pairs. The merged sub-chains can then differ in length. At every step, one merges pairs where possible and otherwise promotes the remaining singlet, as sketched graphically in Figure S1. This procedure requires M steps where $2^{M-1} < N \leq 2^M$.

SIMULATION METHODS

Implementation of Hierarchical Chain-Growth Monte Carlo Algorithm. We assembled the fragments form temperature replica-exchange molecular dynamics (REMD) simulations[41] into full-length structures, as illustrated in Figure 1, using the hierarchical chain-growth Monte Carlo algorithm (Figure 2) described above. Here, each fragment had an overlap of two residues with the subsequent fragment and capped termini (Figure 3A), but other choices are possible. Only steric interactions between fragments were considered in the chain growth. The chain-growth algorithm was implemented by building on the MDAnalysis Python library[42,43] as described below.

0. We randomly draw two conformations from the whole set of sampled conformations in the preceding hierarchy level.

1. We perform a rigid body superposition over the peptide bonds between residues $j-1$ and j of the first fragment and between residues i and $i+1$ of the subsequent fragment. Here, i designates the first and j the last residue of a fragment (excluding the end-capping groups). Thus, we align the four backbone atoms C, O, N, and H in the

peptide bonds between residues common to both fragments (Figure 3A, residues 4 and 5).

   (a) If the root-mean-square deviation (RMSD) of the superimposed region is below a given cut-off, here 0.6 Å, we accept the alignment.

   (b) Else we discard the conformations, draw new conformations and start again with step 1.

2. We check the aligned structures for clashes. The excluded volume is detected by calculating a neighbor list (as implemented in MDAnalysis[42,43]) for the residues from both fragments outside the alignment regions. I.e., all atoms from residues $j-1$ and j of the first and residues i and $i+1$ of the subsequent fragment as well as hydrogen atoms are excluded from the calculation of the neighbor list. Heavy atoms within a distance of 2.0 Å count as clash.

   (a) If no steric clash was detected we proceed to merging the fragments.

   (b) Else we discard the conformations, draw new conformations and start again with step 1.

3. We stitch the superimposed fragments together. We merge the fragments by taking backbone and sidechain atoms of residues $i : j-1$ from the first fragment and $i+1 : j$ from the subsequent fragment. In Figure 3B, the alignment, clash calculation and the stitching procedure is shown exemplary for hierarchy level 0 fragments 0 and 1.

REMD Simulation of Fragments. REMD simulations were run in GROMACS/2016.4[44] with the AMBER99SB*-ILDN-q force field[26,45–47] and the TIP3P water model.[48] The 46 aS penta-peptide fragments were capped at the N- and C-terminus by acetyl and N-methyl groups, respectively. The fragments were solvated in water with 150 mM NaCl, ensuring overall charge neutrality. The resulting systems contained about 1900 atoms each. For

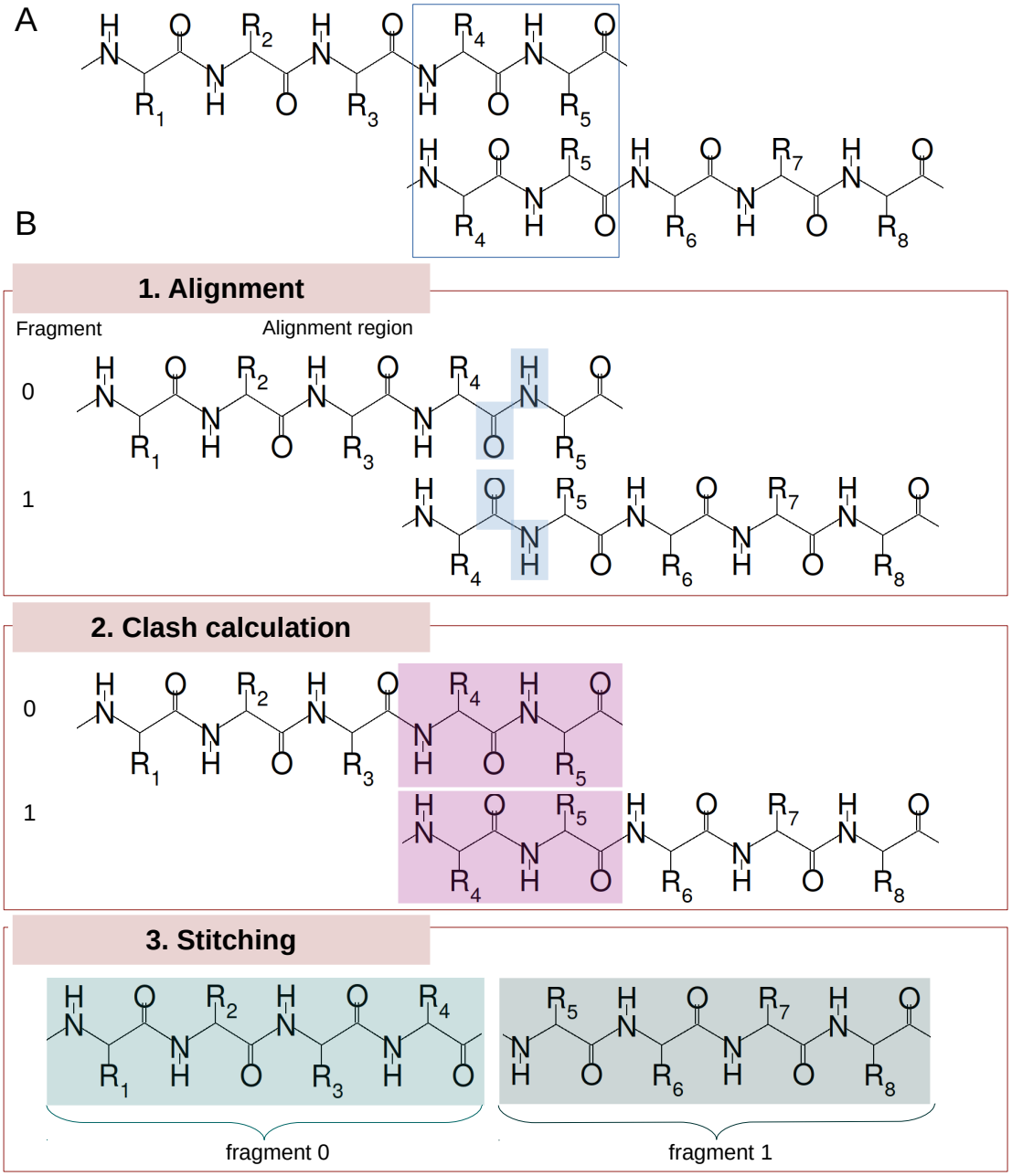

Figure 3: Implementation of the hierarchical Monte Carlo chain-growth algorithm. The algorithm is illustrated for the merging of two fragments at the beginning of a peptide chain (i.e., fragments 0 and 1 at hierarchy level 0). (A) Fragments with a sequence overlap of two residues with the subsequent fragment (blue box) are used as input. (B.1) To merge two randomly picked fragments at hierarchy level 0, the peptide bonds (blue shading) of the fragments are aligned. If the RMSD of the superimposed atoms is below a cut-off (here 0.6 Å), the aligned fragments are checked for clashes. (B.2) Steric overlap is probed with a heavy-atom cut-off distance of 2.0 Å. Residues (backbone and sidechain) at the alignment point (magenta shading), the endcapping groups (ACE and NME), and hydrogen atoms are excluded from the clash calculation. (B.3) If no steric clash is detected, the structure combining residues 1-4 from fragment 0 and residues 5-8 from fragment 1 is stored for use in the next hierarchy level.

each of the 46 aS penta-peptide fragments, REMD simulations were performed for 100 ns using 24 replicas spanning a temperature range of 288 K to 431 K at constant pressure, at temperatures set according to the algorithm by Patriksson *et al.*[49]

To maintain a pressure of 1 bar, the Parrinello-Rahman[50] barostat was used. Temperature coupling was achieved by velocity rescaling with a time constant of 0.1 ps using the Bussi-Donadio-Parrinello thermostat.[51] The P-LINCS algorithm was used to constrain all bonds.[52] Using the particle mesh Ewald method, long-range electrostatics were calculated with a cut-off of 10 Å. The van der Waals cut-off was set to 12 Å. REMD production runs were preceded by energy minimization and 1 ns equilibration in the NPT ensemble. During the production runs of 100 ns (per replica) structures were saved in intervals of 10 ps. In this way, we created a library of 10 000 fragment structures for each peptide segment at the temperature of interest, $T = 288$ K.

MD Simulations of Full-Length Models. All-atom MD simulations of full-length aS were run in Gromacs/2016.4[44] with the AMBER99SB*-ILDN-q force field[26,45–47] using the TIP4P-D water model.[25] Twenty models were chosen at random from the ensemble of models generated by hierarchical assembly. The aS chains with charged termini were each solvated in water with 150 mM NaCl, ensuring overall charge neutrality. Each system contained about 350 000 atoms. Simulations were performed at a constant temperature of 300 K using the Bussi-Donadio-Parrinello velocity-rescaling thermostat with a time constant of 0.1 ps.[51] The pressure was maintained at 1 bar using the Parrinello-Rahman barostat.[50] The P-LINCS algorithm[52] was used to constrain all bonds. To calculate long-range electrostatics the particle mesh Ewald method was used with a cut-off of 12 Å. A cut-off of 12 Å was used for van-der-Waals interactions. Energy minimization and 200 ps equilibration were performed before running production runs of 100 ns.

Calculation of Experimental Observables. We calculated NMR chemical shifts with SPARTA+.[53] Reference random coil chemical shifts were predicted using the POTENCI web

server developed by Nielsen and Mulder[54] to arrive at secondary chemical shift predictions by subtraction of the reference value. J-couplings were calculated as previously described[55] using the original Karplus parameters described by Wirmer and Schwalbe.[56] The radius of gyration $R_{G,i}$ for a saved aS structure i was calculated using FoXS,[57] taking the solvent shell into account. In addition, a geometric radius of gyration $R_{G,i}$ was computed from the protein coordinates with the MDAnalysis Python library.[42,43] The model ensemble average was calculated as the root-mean-square average, $R_G = (\sum_i^N {R_{G,i}}^2/N)^{1/2}$, where $R_{G,i}$ is the radius of gyration of the i-th member of the ensemble of size N.

RESULTS AND DISCUSSION

Testing and Validation of the Hierarchical Chain-Growth Approach. We first verified that the Monte Carlo hierarchical chain-growth algorithm works as expected and tested different practical implementations. In the hierarchical chain-growth algorithm, we assemble the full-length chain by generating possible structures of dimers of fragments (Figure 2A) and then the structures of possible quadruplets (Figure 2B) and so on until the full chain is grown (Figure 2D). As expected from its derivation, the chain-growth algorithm generates ensembles in which each structure appears with a Boltzmann weight, which we confirmed by comparing chains with up to 26 residues to chains grown by brute-force generation of self-avoiding random walks (Figure S2). We also evaluated the effect of using fragments of different length to grow 50-amino-acid long aS sub-chains. Comparing 3mer, 4mer and 5mer fragments, we found that using larger fragments resulted in somewhat more compact models (Figure S4A), with a 5mer having a radius of gyration R_G about 1 Å less than a 4mer. Larger fragments can adopt more compact conformations, with more interactions within the fragments. In terms of the end-to-end distance χ (Figure S4B) and the diversity of structures, as measured by the pairwise RMSD between structures in the ensemble (Figure S4C), the differences between 3mer, 4mer and 5mer fragments are small. We decided

to use 5mer fragments in the following to capture also less-extended structures in our initial pool of fragment structures.

Secondly, we compared the effects o aligning the fragments at the peptide bond (Figure 3) or the backbone of the residue at the merge point (Figure S3). The peptide bond is relatively rigid due its partial double-bond character compared to the backbone of the residue at the merge point with its rotatable ϕ and ψ dihedral angles. Indeed, using the backbone rather than the peptide bond for alignment results in larger differences between the assembly of 5mer, 4mer and 3mer fragments, as judged by the distributions of R_G, χ and pairwise RMSDs between structures (Figure S4). Importantly, assembling the chain via alignment of the peptide bond preserves the conformational distributions from MD simulations of the fragments. Figure 4A and B illustrate this point for the ψ angle of A11 and Y39. By contrast, assembly via the backbone introduces a bias towards extended structures with $\psi > 100°$.

Thirdly, we considered different overlap between the fragments. Using an overlap of one or two residues between the fragment makes no significant difference (Figure S5). For subsequent chain growth, we used the central three residues of 5mer fragments. The central residues of a fragment should be more representative of the local structure in the context of an IDP chain.

We conclude that the different choices one could make in implementing our chain-growth algorithm, overall, do not have drastic effects on the global structures of the generated ensembles. Local structure may be preserved better by aligning on the rigid peptide bond rather than on the more flexible backbone.

Full-length Models of aS. We generated a highly diverse ensemble of full-length all-atom aS structures with the hierarchical Monte Carlo chain-growth algorithm. For aS, we split its primary sequence into 46 fragments to produce input structures for the chain growth. For each fragment we ran exhaustive atomistic simulations with explicit solvent using REMD. This initial sampling phase already lends itself to HPC resources with a large

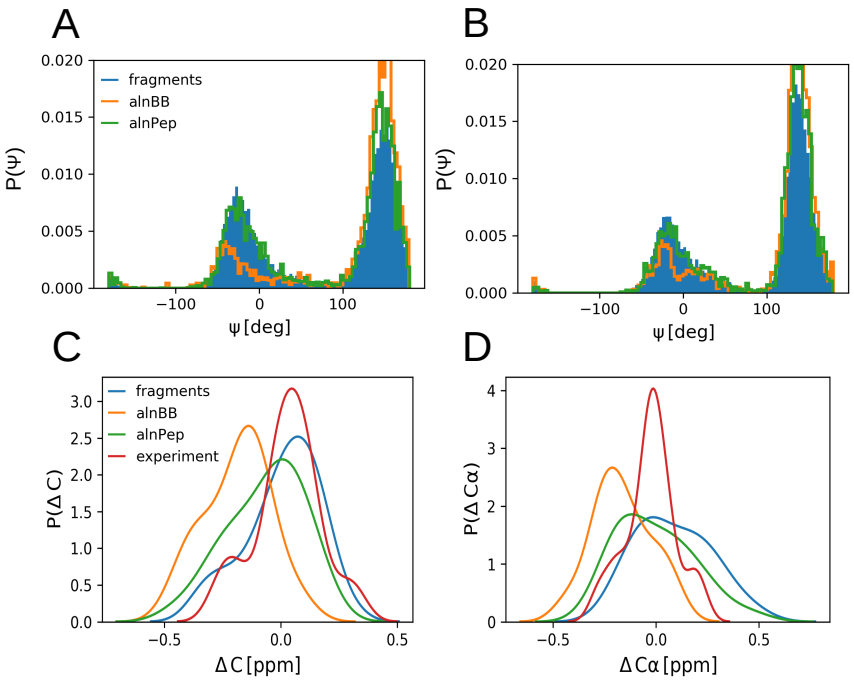

Figure 4: Comparison of different alignment and stitching approaches to implement the chain-growth algorithm. (A,B) Distribution of the ψ dihedral angle for A11 (A) and Y39 (B) before and after the assembly of the MD fragments into full-length structures. Results are shown for two different alignment approaches (orange: alignment over the backbone atoms, alnBB; green: alignment over the peptide bond, alnPep) and for the peptide fragments (blue). (C,D) Distribution of Δppm secondary chemical shifts of ^{13}C (C) and $^{13}C\alpha$ (D) from experiment (red), MD fragments (blue), and full-length models grown with backbone alignment (orange, alnBB) and peptide bond alignment (green, alnPep).

number of computing nodes as each fragment can be simulated independently from the other fragments and no overhead is incurred due to inter-node communication. During chain growth, structures are drawn from the simulation ensembles for the individual fragments.

Using the hierarchical chain-growth algorithm described above we grew 20 000 full-length aS models. The calculation ran on 20 compute cores and we grew 1 000 full-length models per core using the same pool of fragments in the 20 runs but different random number seeds. To sample an ensemble with a large diversity in local conformations in the highest level M, we grew 10 000 structures in the levels 0 to $M-1$. The extensive sampling of local and global structures yielded highly-diverse full-length structures of aS, as judged by pairwise RMSD (Figure 5D), with an average pairwise RMSD of $\approx$56 Å between the 20 000 models.

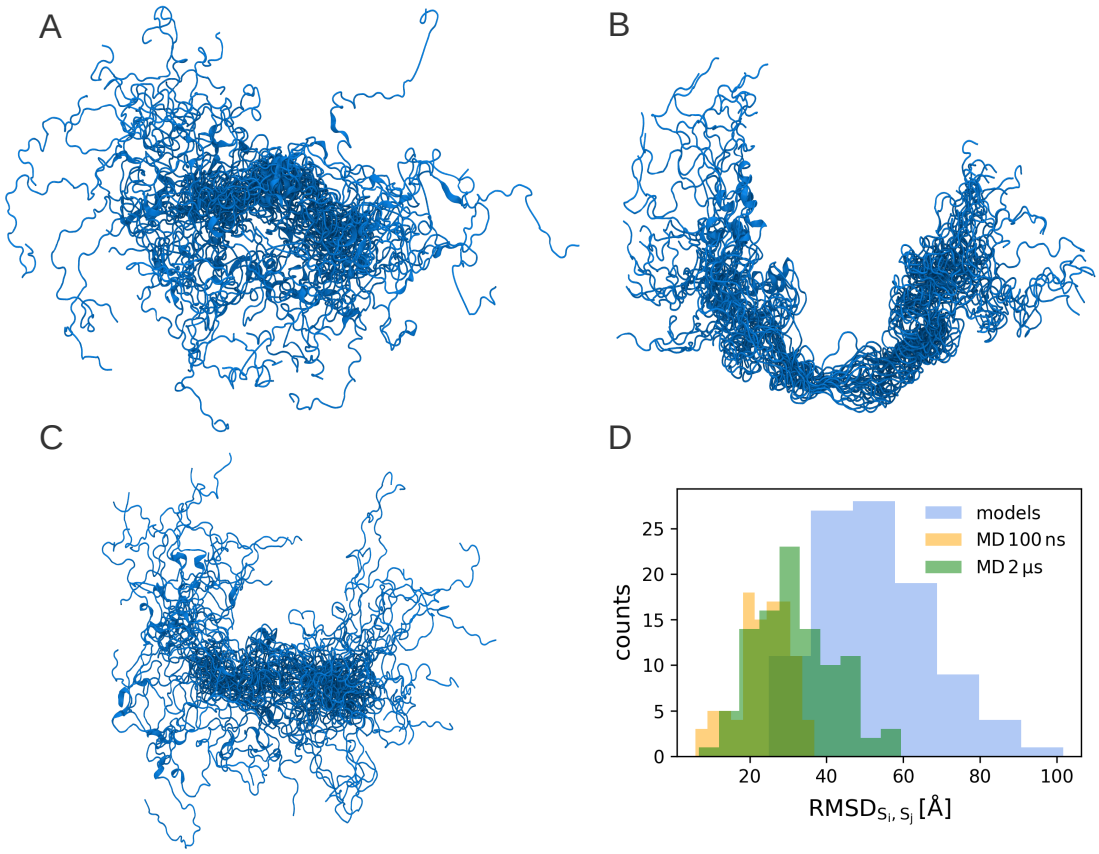

Figure 5: Conformational diversity of aS model ensembles from MD simulation and hierarchical chain growth. (A) 30 aS models from hierarchical chain growth. (B) 30 structures from 10 ns of MD. (C) 30 structures from 100 ns of MD. The structures were aligned on the central third of their sequence in (A), (B) and (C). (D) Distribution of pairwise RMSD between 100 different models obtained by MD, sampled uniformly in time from 100 ns and 2 μs, respectively, and by hierarchical chain growth.

In essence, no two structures are alike.

Sampling Efficiency of the Hierarchical Approach. Visualization makes it clear that our hierarchical approach captured a much larger conformational space than would be accessed in a typical all-atom MD simulation of an IDP. In Figure 5B the persistence of a transient hairpin conformation over the course of 10 ns of MD simulation is clearly visible. Such local structure decorrelated over 100 ns of simulation (Figure 5C). However, it is clear from comparing Figure 5A and Figure 5C that the ensemble from chain growth sampled a much larger conformational space. The structures from 100 ns of MD simulations still re-

sembled one another, unlike the structures from chain-growth, which fully explore the space of possible structures.

We compared the distribution of RMSD values between 100 models sampled with hierarchical chain growth to 100 different conformations sampled in a 100 ns and a 2 μs MD simulation. The 100 structures from the 100 ns and 2 μs MD simulation trajectories were taken at regular time intervals of 1 ns and 20 ns. For the chain growth ensemble with 100 different models we observed an average pairwise RMSD of ≈53 Å, whereas the MD ensemble after 100 ns showed an average RMSD of ≈23 Å and after 2 μs an average of ≈32 Å (Figure 5D). This demonstrates that by using the hierarchical chain-growth algorithm we were able to obtain a diverse ensemble, much more diverse than an ensemble sampled in 2-μs of MD.

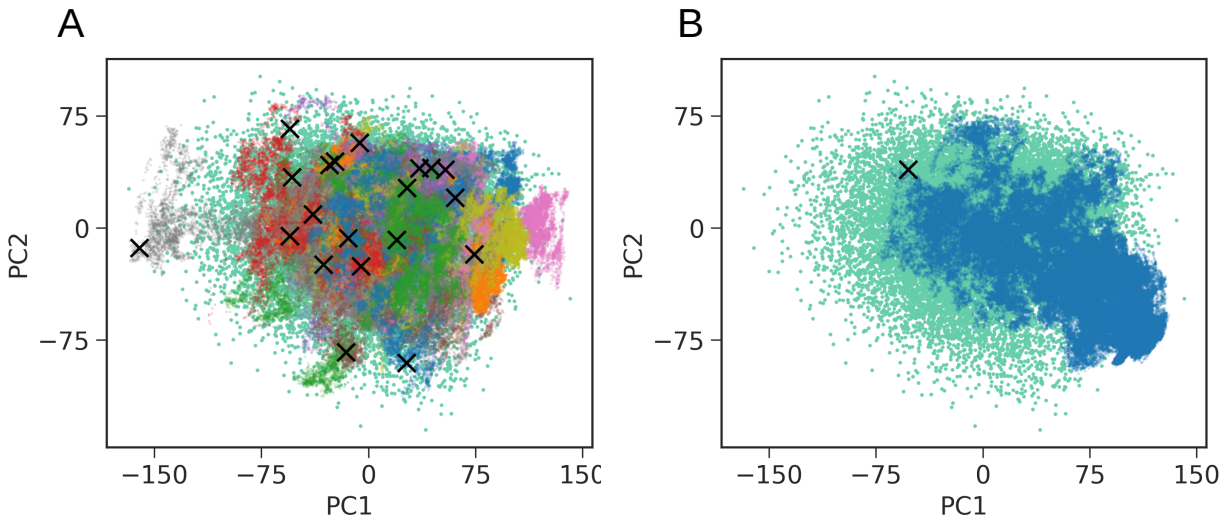

Figure 6: Conformational space of full-length models of aS projected onto the principal component axes 1 and 2, as obtained from the ensemble of hierarchical models. (A) Conformations sampled in 20 × 100 ns MD simulations with different starting structures (black crosses) projected onto the conformational space spanned by 10 000 full-length aS models (aquamarine). (B) 2 μs trajectory (blue) projected onto the conformational space sampled by the aS models (aquamarine).

Principal component analysis (PCA)[58] gives us a global view of the conformational space sampled. Figure 6 shows projections of the 20 000 models constructed with hierarchical chain growth onto the principal components 1 and 2. On top, Figure 6A shows 200 000 MD

conformations sampled in 20 × 100 ns of MD simulation started from 20 different structures (black crosses). In essence, each of the 20 runs explores a small region within the confines of the space sampled by the hierarchical models. Figure 6B shows the 2 μs MD simulation projected onto the principal components 1 and 2 of the hierarchical models. Again, the MD ensemble is contained within the hierarchical ensemble, indicating the larger conformational diversity of structures in the hierarchical ensemble.

Overall, the MD trajectories stayed within the boundaries of the conformational space defined by the ensemble from chain growth, which suggests that the hierarchical sampling covered, at least at the global level of principal axes 1 and 2, the conformation space visited in the 20 × 100 ns and μs scale MD simulations. Thus we conclude tentatively that our hierarchical approach exhaustively samples the global structures of IDPs such as aS.

All-atom MD Simulations of aS in Explicit Solvent. Our models are meaningful starting points for simulation, as shown by the overall behavior of the simulations started from our models and the conservation of their characteristics in MD simulations. As expected, the full-length structures of aS (Figure 2F) were dynamic in all-atom MD simulations in explicit solvent. In Figure 7B the starting structure for an all-atom simulation in explicit solvent is presented. The structure is extended and this particular structure features a turn close to the end of the N-terminal domain and at the beginning of the C-terminal domain of aS. Like most of our models the structure shows little in the way of well-defined secondary structure elements, featuring only a short helical segment. After 50 ns of simulation the turn at the end of the N-terminal vanished and the short helix deformed (Figure 7C). Visually, the structure expanded further. Over the next 50 ns the molecule became somewhat less expanded (Figure 7D). Preservation of their characteristics in the MD simulations suggested that the models provide meaningful starting points for simulations with the state-of-the art AMBER99SB*-ILDN-q force field[26,45–47] and TIP4P-D water model.[25] Interestingly, the simulations showed a slight compaction of the models, as judged by R_G (Figure 7A). The

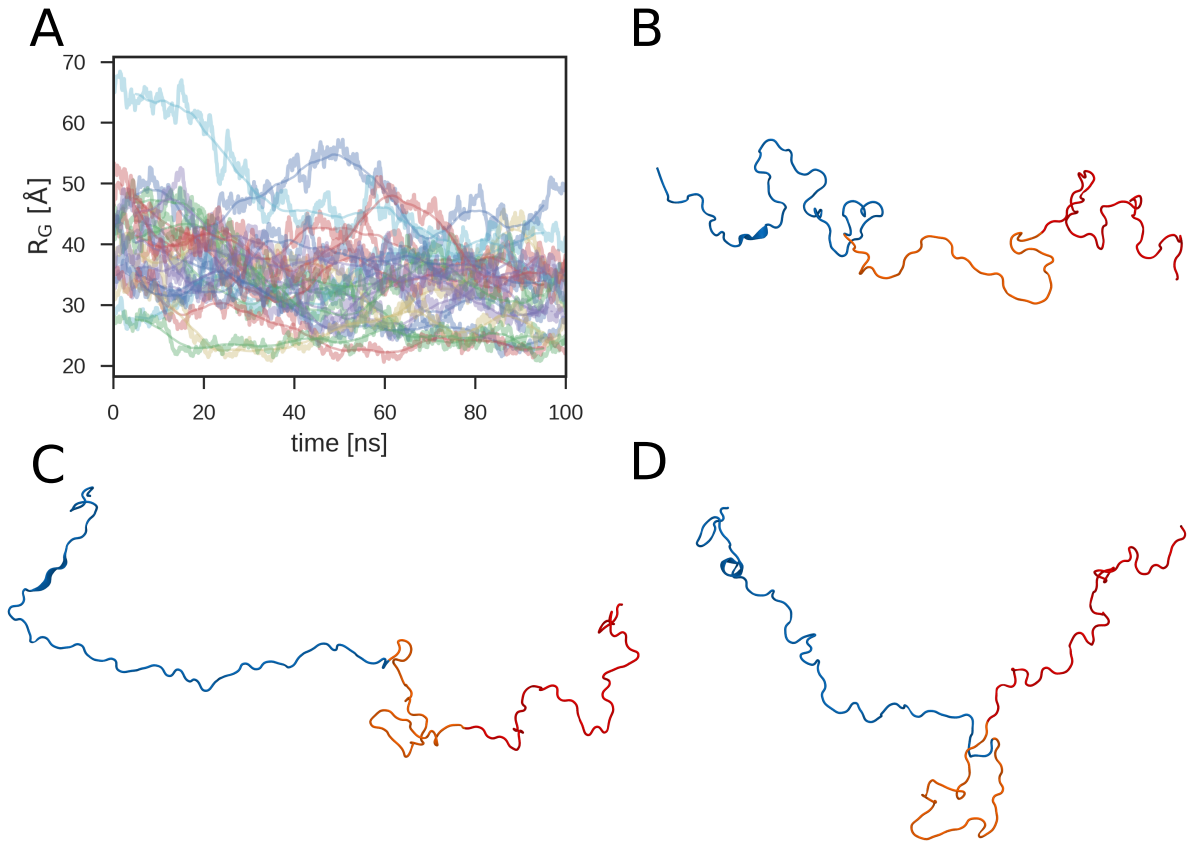

Figure 7: All-atom MD simulations of full-length models of aS. (A) fluctuations of the geometric R_G in all-atom MD (noisy curves) and the respective moving average using a window size of 10 ns (smooth curves of corresponding color). Representative snapshots are shown for a simulation run: starting structure (B), after 50 ns (C) and 100 ns of MD (D). In (B), (C) and (D) the N-terminal domain of aS is colored in blue (M1-K60), the central region with a hydrophobic motif in orange (E61-V95), and the C-terminal domain in red (K96-A140).[59]

mean square R_G dropped from ≈40 Å (which is close to experiment, as discussed below) to ≈ 33 Å during the 100 ns of MD. This compaction could be an indicator of a lack of residual structure in the models or of a poor force field and solvent model, which underestimate the solvation of the protein. However, the simulations did not access fully collapsed structures, with R_G ≈ 15 Å similar to a folded protein of this length, suggesting that the AMBER99SB*-ILDN-q force field and TIP4P-D water model[25] describe IDPs well enough.[24]

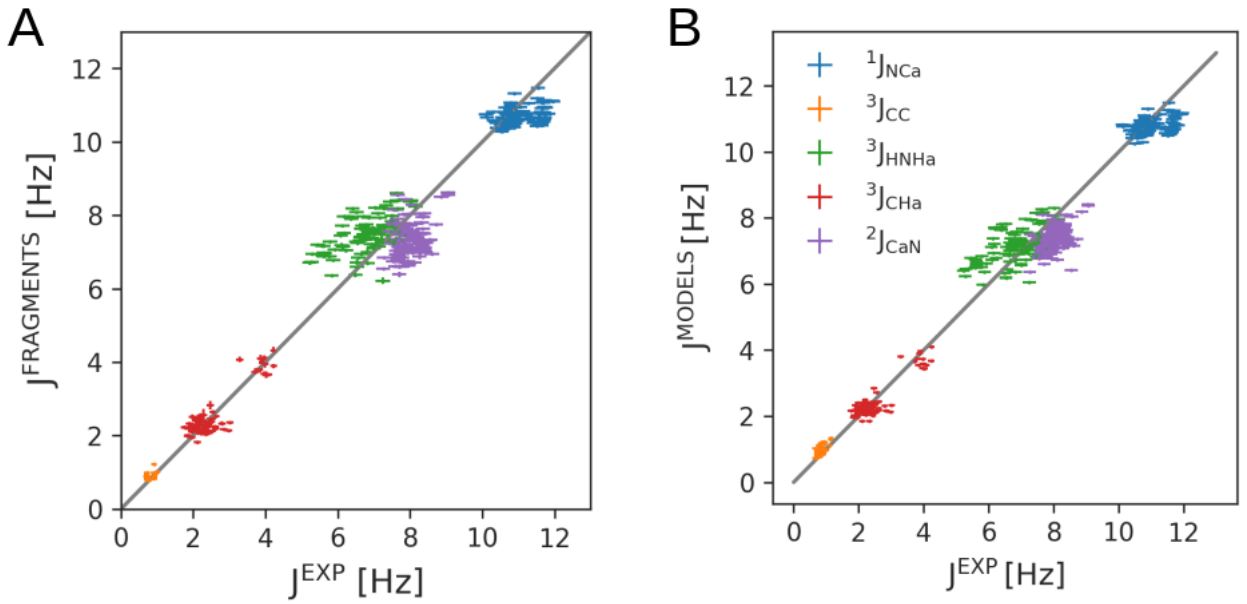

Figure 8: Comparison of calculated NMR J-couplings (y-axis) to experiment[6] (x-axis). J-couplings were obtained (A) from MD simulations of peptide fragments and (B) for full-length aS chains built by hierarchical chain growth. Each point corresponds to the ensemble average of a single residue. Standard errors of the mean were estimated by block averaging.

Comparison of aS Ensembles to NMR Experiments probing Local Structure. The aS ensembles obtained by hierarchical chain growth compare well to NMR experiments probing the local structure through J-couplings[6] and chemical shifts.[6,16,60] Without reweighting, the $^3J_{CC}$ and $^3J_{CH\alpha}$ couplings probing ψ and ϕ dihedral angles agree well with experiment (Figure 8A). The magnitude of $^1J_{NC\alpha}$, $^2J_{C\alpha N}$ and $^3J_{HNH\alpha}$ couplings, which report on the ϕ dihedral angles, were captured by our ensemble but small systematic offsets were observed. Values for $^1J_{NC\alpha}$ and $^2J_{C\alpha N}$ tend to be somewhat lower in our ensemble than in experiment, whereas $^3J_{HNH\alpha}$ values tend to be slightly overestimated. The amide nuclear spin involved in these couplings is affected by many processes such as hydrogen-bonding or geometric distortion. These processes are not well described by current force fields, which may explain the small systematic deviations.[61] The agreement with experiment is equally good for J-couplings calculated from full-length models (Figure 8B) and from the initial fragment library generated of short aS fragments (Figure 8A), highlighting that the chain-growth algorithm preserves the local structure sampled in REMD.

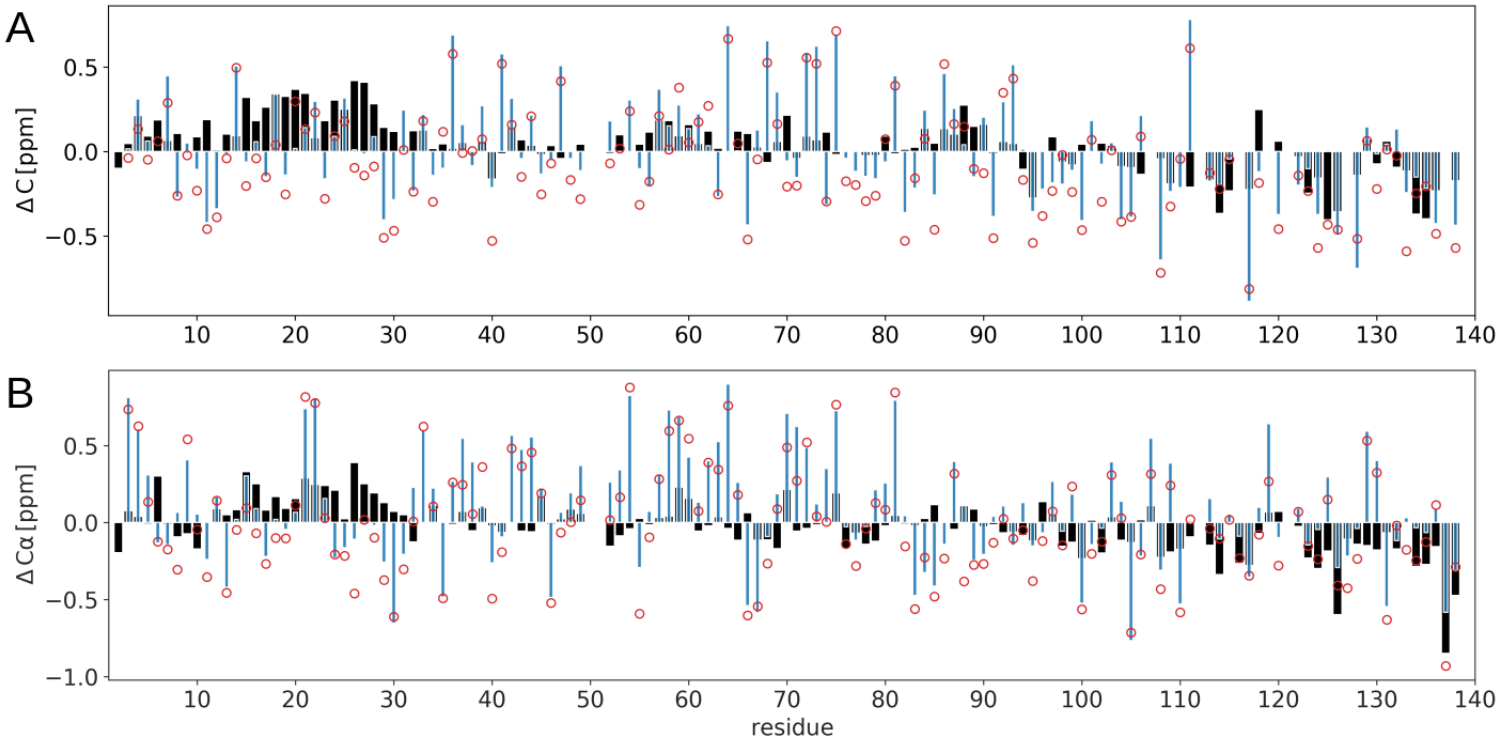

Figure 9: NMR chemical shift analysis for fragment simulations. (A) ^{13}C and (B) $^{13}C_{\alpha}$ secondary shift values from fragment all-atom MD simulations (blue bars), assembled full-length models (red circles) and experiments[6,16,60] (black bars).

The ^{13}C and $^{13}C_{\alpha}$ secondary shifts calculated from models were around zero for all residues, suggesting a local backbone conformation closely resembling random coil, in agreement with experiment[6,16,60] (Figure 9). Deviations from zero in the secondary chemical shifts (Δppm) report on (residual) structure, but these deviations were quite small (< 1 ppm) for models compared to estimates of the expected errors in calculating chemical shifts. For empirical chemical shift prediction of ^{13}C and $^{13}C_{\alpha}$ shifts, RMSD to experiment of about 1 ppm were found for a validation set of 11 proteins.[53] Here, the RMSD to experiment is 0.32 and 0.37 ppm for ^{13}C and $^{13}C_{\alpha}$ shifts, respectively, as predicted for the fragments, and 0.33 and 0.35 ppm as predicted for the full-length models. Indeed, for most residues the chemical shift predicted after assembly agrees with these predicted for the MD fragments before assembly. Still for few residues the experimental secondary shifts were captured better by the MD fragments before assembly to full-length models, e.g., the carbonyl secondary shifts reported

for A18, Figure 9A or the Cα secondary shifts reported for V16, Figure 9B. Nevertheless for some residues the agreement with experiment was better after the assembly (carbonyl secondary shifts reported for E20, Figure 9A or the Cα secondary shifts reported for V74, Figure 9B).

In Figure 4C and D, we compared the distribution of the experimental Δppm secondary chemical shifts for the MD fragments and for full-length chains grown with different alignment procedures (alignment of the flexible backbone versus alignment of the rigid peptide bond). Growing IDP chains by aligning the backbone of the fragments at the merge point results in a shift in the chemical shift distributions, which mirrors the shift towards β-strand like conformations we found with this growth procedure (Figure 4A and B). By contrast, alignment via the peptide bond results in a much smaller shift away from the distributions from the fragment predictions (Figure S6) and experiment. Alignment via the peptide bond largely preserves the structures of the fragments as judged by chemical shift predictions, but not exactly. Naturally, full-length structures will be at least subtly different from fragment structures, e.g., some fragment structures will be sterically impossible. Taken together, the analysis of the chemical shifts demonstrates that our implementation of the chain-growth procedure leads to good models of IDP structure and conversely, that chemical shifts are useful indicators in the modeling of IDPs.

Comparison of aS Ensembles to SAXS Experiments probing Global Structure. The global structure of the hierarchically grown aS models is also consistent with experiment, as probed by SAXS measurements of the radius of gyration (R_G). The estimated root-mean-square radius of gyration for the ensemble of full-length aS models is $R_G \approx 40$ Å (Figure 10). In calculating the $R_{G,i}$ values of individual structures i, we took the solvent shell into account using FoXS,[57] but we obtained essentially the same R_G by calculating $R_{G,i}$ of individual structures i directly from the protein coordinates. In SAXS experiments, R_G values of 40 Å (Binolfi et al.[62]) and 45 Å (Curtain et al.[63]) have been reported, bracketing our value. The

value expected for a 140 amino acid random coil is 45 Å, using the parameters determined by Sosnick et al. considering nearest neighbor effects.[64] We find it encouraging that the chain growth captures the overall dimensions of the disordered chain. Spurious compaction due to force field issues would require the imposition of a bias during chain growth or simulation to steer the ensemble away from overly compact structures that underestimate R_G,[39,40] or a reweighting of the ensemble.[17,27]

The observed compaction during MD, with the geometric R_G dropping from 40 to 33 Å in 100 ns, suggests that the chain "as grown" before MD may be a better representation of the extended structures seen in SAXS (Figure 10), with measured R_G values between 40 and 45 Å. It is not unexpected that the long chains predicted to be extended undergo some kind of collapse. The factors driving the collapse, e.g., the hydrophobic effect and the potential overstabilization of protein-protein interactions, are not prominent for the very short fragments we simulated. In turn, this may indicate that the predicted chains are probably closer to the real ensemble than the MD refined ones, and that this type of set-up gives us a handle to test and optimize balanced force fields for IDPs. In any case, the simulations demonstrated that AMBER99SB*-ILDN-q force field[26,45–47] and TIP4P-D water model[25] describe IDPs well enough, but we note that other IDP force fields may work equally well or better.[24]

Overall our hierarchically grown IDP models captured both local conformations, as reported by NMR J-couplings and chemical shifts, and the overall dimension reported by SAXS experiments, remarkably well, without any refinement. This result encourages the use and further development of our approach for generating starting conformations of highly-parallel MD simulations of IDPs and the testing of MD force fields.

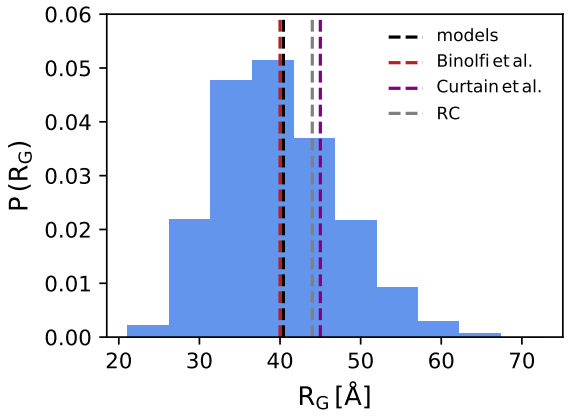

Figure 10: Distribution of the radius of gyration $R_{G,i}$ in the ensemble of aS structures obtained from hierarchical chain growth. Vertical lines indicate the root-mean-square R_G value of the ensemble (black; models), R_G values from SAXS experiments by Binolfi et al.[62] (red) and Curtain et al.[63] (magenta), and an R_G estimate for a random coil (RC; gray).[64]

CONCLUDING REMARKS

Our hierarchical IDP chain-growth approach is perfectly suited to the exascale high-performance computing resources which are becoming available. We generated highly diverse structures, much more diverse than what one would sample in a typical MD simulation. Our structures capture both NMR data probing local structure and and SAXS data probing global structure, without any refinement, emphasizing again that the structures should be excellent starting points for MD simulations. By generating starting configurations that closely follow the Boltzmann distribution, we can launch a large number of independent simulations and this swarm of simulations can fully explore the conformational space of an IDP. Simulations setup in this way may help identify and rectify force field issues for IDPs.

It would be computationally feasible to create exhaustive fragment libraries for, say, the $20^3 = 8000$ distinct amino-acid trimers with generic flanking residues. Considering the speed of assembly, a web service for generic IDP assembly is thus feasible. It would also be possible to include post-translational modifications such as phosphorylation.

Deviations from experiment can be taken into account in a Bayesian framework. Ensemble refinement by reweighting[17] can be applied very efficiently to large ensembles.[27] Bayesian

analysis of all-atom simulations of a disordered peptide[27] showed that quantitative agreement with high-resolution NMR experiments can be achieved and led to system-specific correction, which may also be important when modeling large IDPs. Stultz et al. have already shown how structural ensembles from fragment assembly can be refined against NMR and SAXS data within a Bayesian framework.[39,40]

Our approach can be extended to simulations of other flexible biomolecules and their assemblies. For instance, it could be used to model long non-coding RNAs and other single-stranded nucleic acids. For single-stranded nucleic acids, sampling is a bottleneck in force field evaluations,[65] which could be addressed by an extension of our approach. We envisage extensions of our approach to simulate dense solutions of IDPs as in biomolecular condensates formed via liquid-liquid phase separation.[4] Our chain-growth algorithm is valid whether individual chains or assemblies of chains are modeled. For modeling dense biomolecular condensates variants of the chain-growth Monte Carlo algorithm we employed here may prove advantageous. For more dilute condensates the current approach should lead to reasonable starting conformations for large-scale MD simulations.

SUPPORTING INFORMATION

Implementation of hierarchical chain-growth algorithm and comparison of chain-growth algorithms, consistency checks for hierarchical chain-growth algorithm and additional analysis of predicted chemical shifts for fragments and full-length models of aS.

Acknowledgement

We acknowledge financial support from the German Research Foundation (CRC902: Molecular Principles of RNA Based Regulation) and by the Max Planck Society. We thank Profs. Markus Zweckstetter and Harald Schwalbe and Drs. Adriaan Bax and Jürgen Köfinger for insightful discussions.

References

(1) Oates, M. E. et al. D^2P^2: database of disordered protein predictions. *Nucleic Acids Res.* **2013**, *41*, D508–D516.

(2) Borgia, A. et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **2018**, *555*, 61–66.

(3) Wright, P. E.; Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29.

(4) Banani, S. F. et al. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 285–298.

(5) Lashuel, H. A. et al. The many faces of α-synuclein: from structure and toxicity to therapeutic target. *Nat. Rev. Neurosci.* **2013**, *14*, 3848.

(6) Mantsyzov, A. B. et al. A maximum entropy approach to the study of residue-specific backbone angle distributions in α-synuclein, an intrinsically disordered protein. *Protein Sci.* **2014**, *23*, 1275–1290.

(7) Meier, S.; Blackledge, M.; Grzesiek, S. Conformational distributions of unfolded polypeptides from novel NMR techniques. *J. Chem. Phys.* **2008**, *128*, 052204.

(8) Shen, Y. et al. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **2009**, *44*, 213–223.

(9) Mukrasch, M. D. et al. Highly Populated Turn Conformations in Natively Unfolded Tau Protein Identified from Residual Dipolar Couplings and Molecular Simulation. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.

(10) Mylonas, E. et al. Domain Conformation of Tau Protein Studied by Solution Small-Angle X-ray Scattering. *Biochemistry* **2008**, *47*, 10345–10353.

(11) Fiebig, K. M. et al. Toward a Description of the Conformations of Denatured States of Proteins. Comparison of a Random Coil Model with NMR Measurements. *J. Phys. Chem.* **1996**, *100*, 2661–2666.

(12) Schwalbe, H. et al. Structural and Dynamical Properties of a Denatured Protein. Heteronuclear 3D NMR Experiments and Theoretical Simulations of Lysozyme in 8 M Urea. *Biochemistry* **1997**, *36*, 8977–8991.

(13) Feldman, H. J.; Hogue, C. W. A fast method to sample real protein conformational space. *Proteins: Struct., Funct., Bioinf.* **2000**, *39*, 112–131.

(14) Bernadó, P. et al. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17002–17007.

(15) Bernadó, P. et al. Defining Long-Range Order and Local Disorder in Native α-Synuclein Using Residual Dipolar Couplings. *J. Am. Chem. Soc.* **2005**, *127*, 17968–17969.

(16) Mantsyzov, A. B. et al. MERA: a webserver for evaluating backbone torsion angle distributions in dynamic and disordered proteins from NMR data. *J. Biomol. NMR* **2015**, *63*, 85–95.

(17) Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **2015**, *143*, 243150.

(18) Lindorff-Larsen, K. et al. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* **2012**, *134*, 3787–3791.

(19) Martin, E. W. et al. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **2016**, *138*, 15323–15335.

(20) Rezaei-Ghaleh, N. et al. Local and Global Dynamics in Intrinsically Disordered Synuclein. *Angew. Chem., Int. Ed.* **2018**, *57*, 15262–15266.

(21) Yeh, I. C.; Hummer, G. Peptide Loop-Closure Kinetics from Microsecond Molecular Dynamics Simulations in Explicit Solvent. *J. Am. Chem. Soc.* **2002**, *124*, 6563–6568.

(22) Rauscher, S. et al. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513–5524.

(23) Huang, J. et al. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73.

(24) 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*, E4758–E4766.

(25) Piana, S. et al. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113–5123.

(26) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the HelixCoil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004–9015.

(27) Köfinger, J. et al. Efficient Ensemble Refinement by Reweighting. *J. Chem. Theory Comput.* **2019**, *15*, 3390–3401.

(28) Camilloni, C.; Pietrucci, F. Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems. *Adv. Phys. X* **2018**, *3*, 1477531.

(29) Löhr, T.; Jussupow, A.; Camilloni, C. Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide. *J. Chem. Phys.* **2017**, *146*, 165102.

(30) Hummer, G.; Kevrekidis, I. G. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.* **2003**, *118*, 10762–10773.

(31) Shirts, M.; Pande, V. S. Screen Savers of the World Unite! *Science* **2000**, *290*, 1903–1904.

(32) Parton, D. L. et al. Ensembler: Enabling High-Throughput Molecular Simulations at the Superfamily Scale. *PLoS Comput. Biol.* **2016**, *12*, 1–25.

(33) Huang, X. et al. Rapid equilibrium sampling initiated from nonequilibrium data. *Proc. Natl. Acad. Sci.* **2009**, *106*, 19765–19769.

(34) Auhl, R. et al. Equilibration of long chain polymer melts in computer simulations. *J. Chem. Phys.* **2003**, *119*, 12718–12728.

(35) Zhang, G. et al. Equilibration of High Molecular Weight Polymer Melts: A Hierarchical Strategy. *ACS Macro Lett.* **2014**, *3*, 198–203.

(36) Stansfeld, P. J.; Sansom, M. S. From Coarse Grained to Atomistic: A Serial Multiscale Approach to Membrane Protein Simulations. *J. Chem. Theory Comput.* **2011**, *7*, 1157–1166.

(37) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press, Inc.: Orlando, FL, USA, 2001.

(38) Mamonov, A. B. et al. General Library-Based Monte Carlo Technique Enables Equilibrium Sampling of Semi-atomistic Protein Models. *J. Phys. Chem. B* **2009**, *113*, 10891–10904.

(39) Fisher, C. K.; Huang, A.; Stultz, C. M. Modeling Intrinsically Disordered Proteins with Bayesian Statistics. *J. Am. Chem. Soc.* **2010**, *132*, 14919–14927.

(40) Ullman, O.; Fisher, C. K.; Stultz, C. M. Explaining the Structural Plasticity of α-Synuclein. *J. Am. Chem. Soc.* **2011**, *133*, 19536–19546.

(41) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(42) Michaud-Agrawal, N. et al. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32*, 2319–2327.

(43) Gowers, R. J. et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. Proceedings of the 15th Python in Science Conference. 2016; pp 98–105.

(44) Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19 – 25.

(45) Hornak, V. et al. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–725.

(46) Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.

(47) Best, R. B.; De Sancho, D.; Mittal, J. Residue-Specific α-Helix Propensities from Molecular Simulation. *Biophys. J.* **2012**, *102*, 1462–1467.

(48) Jorgensen, W. L. et al. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(49) Patriksson, A.; van der Spoel, D. A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2073–2077.

(50) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(51) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(52) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 116–122.

(53) Shen, Y.; Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **2010**, *48*, 13–22.

(54) Nielsen, J. T.; Mulder, F. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J. Biomol. NMR* **2018**, *70*, 141–165.

(55) Best, R. B.; Buchete, N.-V.; Hummer, G. Are Current Molecular Dynamics Force Fields too Helical? *Biophys. J.* **2008**, *95*, L07 – L09.

(56) Wirmer, J.; Schwalbe, H. Angular dependence of $^{1}J(N_i,C_{\alpha i})$ and $^{2}J(N_i,C_{\alpha(i-1)})$ couplings constants measured in J-modulated HSQCs. *J. Biomol. NMR* **2002**, *23*, 47–55.

(57) Schneidman-Duhovny, D. et al. Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophys. J.* **2013**, *105*, 962–974.

(58) García, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699.

(59) Eliezer, D. et al. Conformational Properties of α-Synuclein in its Free and Lipid-associated States. *J. Mol. Biol.* **2001**, *307*, 1061–1073.

(60) Maltsev, A. S.; Ying, J.; Bax, A. Impact of N-Terminal Acetylation of α-Synuclein on Its Random Coil and Lipid Binding Properties. *Biochemistry* **2012**, *51*, 5004–5013.

(61) Vögeli, B. et al. Limits on Variations in Protein Backbone Dynamics from Precise Measurements of Scalar Couplings. *J. Am. Chem. Soc.* **2007**, *129*, 9377–9385.

(62) Binolfi, A. et al. Interaction of α-Synuclein with Divalent Metal Ions Reveals Key Differences: A Link between Structure, Binding Specificity and Fibrillation Enhancement. *J. Am. Chem. Soc.* **2006**, *128*, 9893–9901.

(63) Curtain, C. C. et al. Alpha-synuclein oligomers and fibrils originate in two distinct conformer pools: a small angle X-ray scattering and ensemble optimisation modelling study. *Mol. BioSyst.* **2015**, *11*, 190–196.

(64) Jha, A. K. et al. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13099–13104.

(65) Grotz, K. K. et al. Dispersion Correction Alleviates Dye Stacking of Single-Stranded DNA and RNA in Simulations of Single-Molecule Fluorescence Experiments. *J. Phys. Chem. B* **2018**, *122*, 11626–11639.

Supporting Information: Hierarchical Ensembles of Intrinsically Disordered Proteins at Atomic Resolution in Molecular Dynamics Simulations

Lisa M. Pietrek,[†,¶] Lukas S. Stelzl,[†,¶] and Gerhard Hummer[*,†,‡]

†*Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue Straße 3, 60438 Frankfurt am Main, Germany*

‡*Institute for Biophysics, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany*

¶*These authors contributed equally*

E-mail: gerhard.hummer@biophys.mpg.de

Phone: +49 69 6303-2501

This document contains

1. Supplementary Text
2. Supplementary Figures (S1 to S6) with captions

Supplementary Text

Self-Avoiding Random Walk

Implementation of Hierarchical Chain Growth for N not a Power of Two. As described in the main text, the hierarchical chain-growth algorithm is also applicable if N, the number of fragments composing the chain, is not a power of two. In Figure S1, the

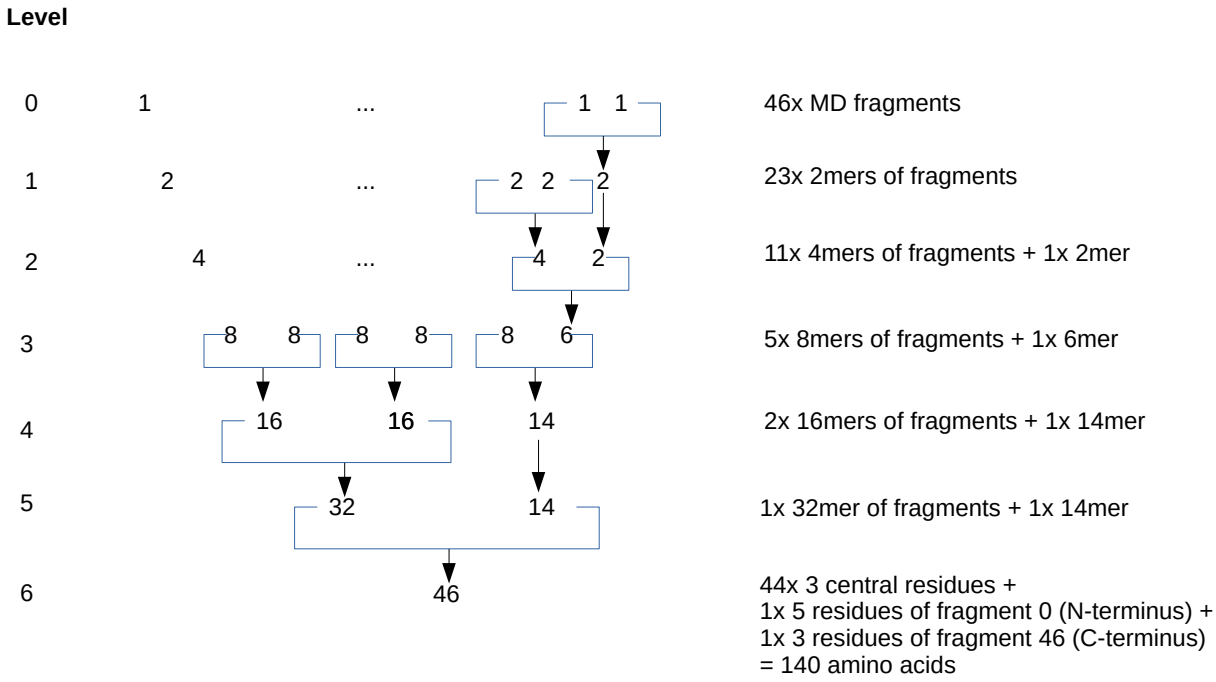

Figure S1: Implementation of hierarchical chain growth for aS. Here N, the number of fragments, is not a power of two, and the C-terminal fragment is shorter (3mer versus 5mers). At each hierarchy level, one merges fragment pairs where possible and otherwise promotes the remaining singlet to the next hierarchy level.

application of the hierarchical chain-growth algorithm to α-synuclein (aS) is shown, where 46 MD fragments are used to grow the full-length 140 amino-acid protein.

Comparison of Chain-Growth Algorithms. For reference, we also implemented a "naïve algorithm" to grow chains from fragments. In this algorithm, we draw enumerations at random and reject them if there is a clash involving any of its fragments:

1. Randomly pick a first element i_1. Set $n = 1$.
2. If $n < N$, randomly pick a new element i_{n+1}; otherwise enter $i_1 \dots i_N$ into the ensemble and return to step 1 (until the ensemble has reached a certain size).
3. Check for a clash of the new element with the rest of the chain.
   (a) If i_{n+1} does not clash with $i_1 \dots i_n$, then accept the addition and increase n by one, $n \mapsto n + 1$, and go to step 2.
   (b) Otherwise, go to step 1 and restart.

For long chains this algorithm has a very low acceptance rate, i.e., many restarts are required to build any new allowed configuration.

We compare the results of the hierarchical and naïve algorithms in Figure S2. For chains with different lengths of 8, 14, or 26 residues, we grew 10 000 chains each with the two algorithms. The end-to-end distance distributions obtained in this way are indistinguishable, supporting the theoretical arguments for the hierarchical algorithm introduced in the main text.

Consistency Checks for Hierarchical Chain Growth

To test the hierarchical algorithm for consistency, we grew short chains of 50 amino acids using different procedures. We explored the effects of varying (1) the fragment length, (2) the alignment and stitching region (Figure S3), and (3) the residue overlap between the fragments.

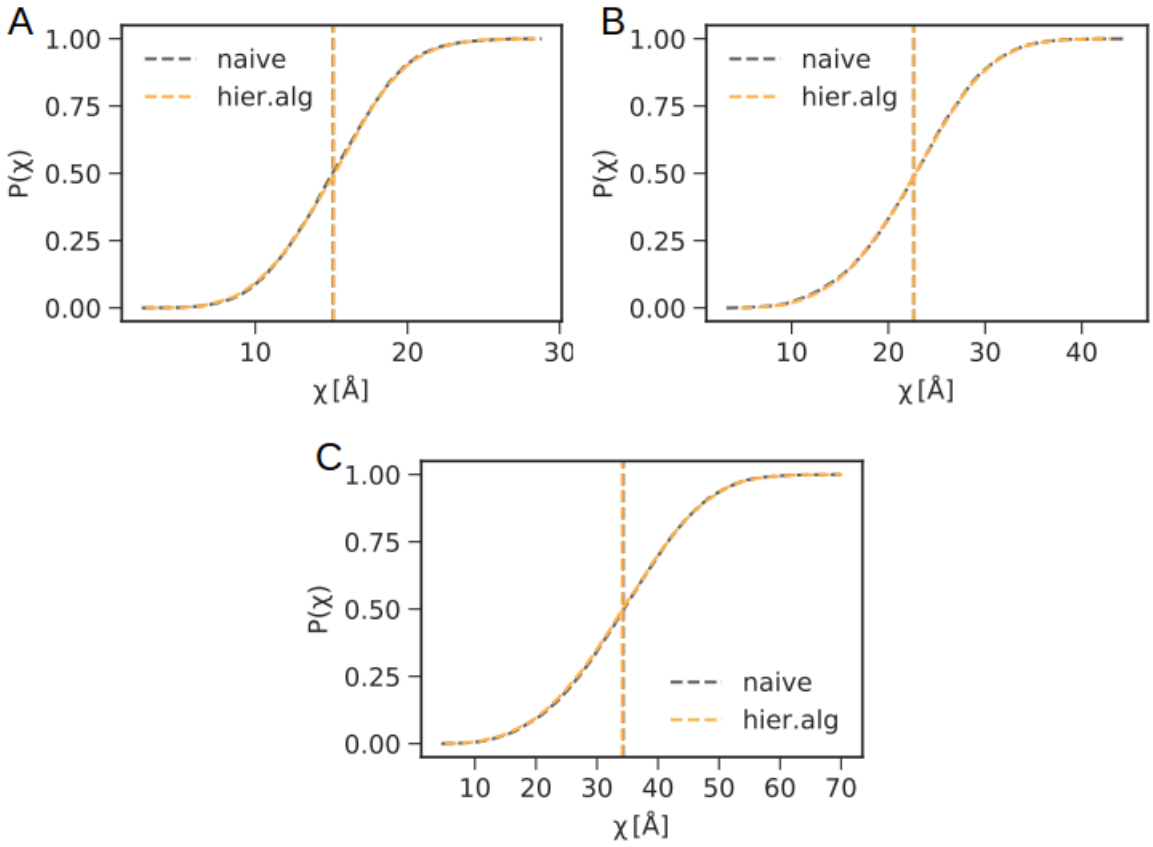

Figure S2: Cumulative distributions of the end-to-end distances χ for chains of different length grown with the naïve algorithm (blue) and the hierarchical algorithm (orange). Shown are the distributions of χ for chains with (A) 8 residues, (B) 14 residues, and (C) 26 residues grown with the the naïve algorithm (blue) and the hierarchical algorithm (orange).

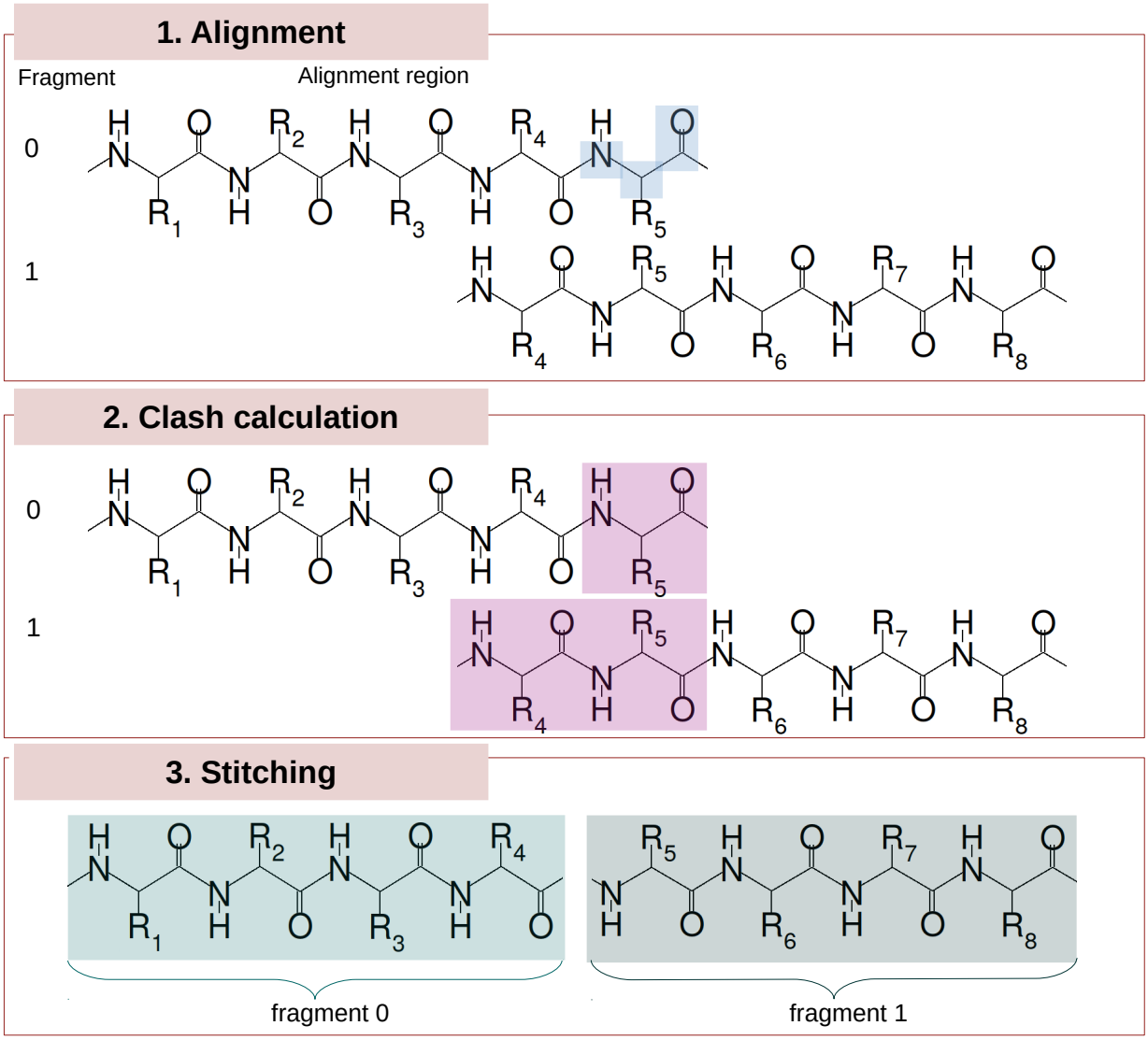

Figure S3: Implementation of the hierarchical Monte Carlo chain-growth algorithm with alignment of the backbone atoms. The growth procedure is shown for hierarchy level 0. To combine two randomly picked fragments, (1) the backbone atoms of the first overlapping residue in the adjacent fragments are aligned (blue shading). If the RMSD of the superimposed atoms is below a cut-off (here 0.6 Å) the aligned fragments are checked for clashes. (2) Steric overlap is probed with a cut-off distance of 2.0 Å. Atoms immediately before and after the alignment point and hydrogen atoms are excluded from the clash calculation (magenta shading). (3) If no steric clash is detected residues 1-4 from fragment 0 and residues 5-8 from fragment 1 are stored for use in the next hierarchy level.

Fragment Length. As shown in Figure S4, the conformation of the hierarchical chain models depends somewhat on the choice of the length of the fragments from which the models are grown. Chains grown from shorter fragments tend to be more extended according to the distributions of the radius of gyration R_G and of the end-to-end distance χ. This dependence is as a consequence of the simplifying assumption of considering only sterics in chain growth from fragments. As a result, favorable inter-fragment interactions are not accounted for, which could favor more compact structures.

Alignment and Stitching Procedure. As shown in Figure 4A-D of the main text, the models grown by the hierarchical chain-growth approach are somewhat dependent on the alignment procedure and the stitching region chosen to merge fragments together. Aligning the backbone atoms around C_α atoms via rigid body superposition, as illustrated in Figure S3, does lead to a disruption of the dihedral angle distribution right at the stitching site (compare Figure 4A and B in the main text). This in turn leads to a bias against α-helical conformations ($\psi < 0$). This bias is largely removed by merging the fragments at the peptide bond and performing the alignment and stitching as shown in Figure 3B in the main text. As shown in Figure S4, alignment of the peptide bond also results in a slight improvement of the fragment-length dependence of R_G and χ distributions.

Fragment Overlap. Aligning the peptide bond between the residues present in adjacent fragments as shown in Figure 3B in the main text, we tested whether the extent of the residue overlap between merged fragments influences the overall properties of the assembled models. Figure S5 shows only small differences between an overlap of one or two residues in the radius of gyration (A), the end-to-end distance (B), and the pairwise RMSD (C) calculated from ensembles consisting of 10 000 models for a chain with 50 amino acids.

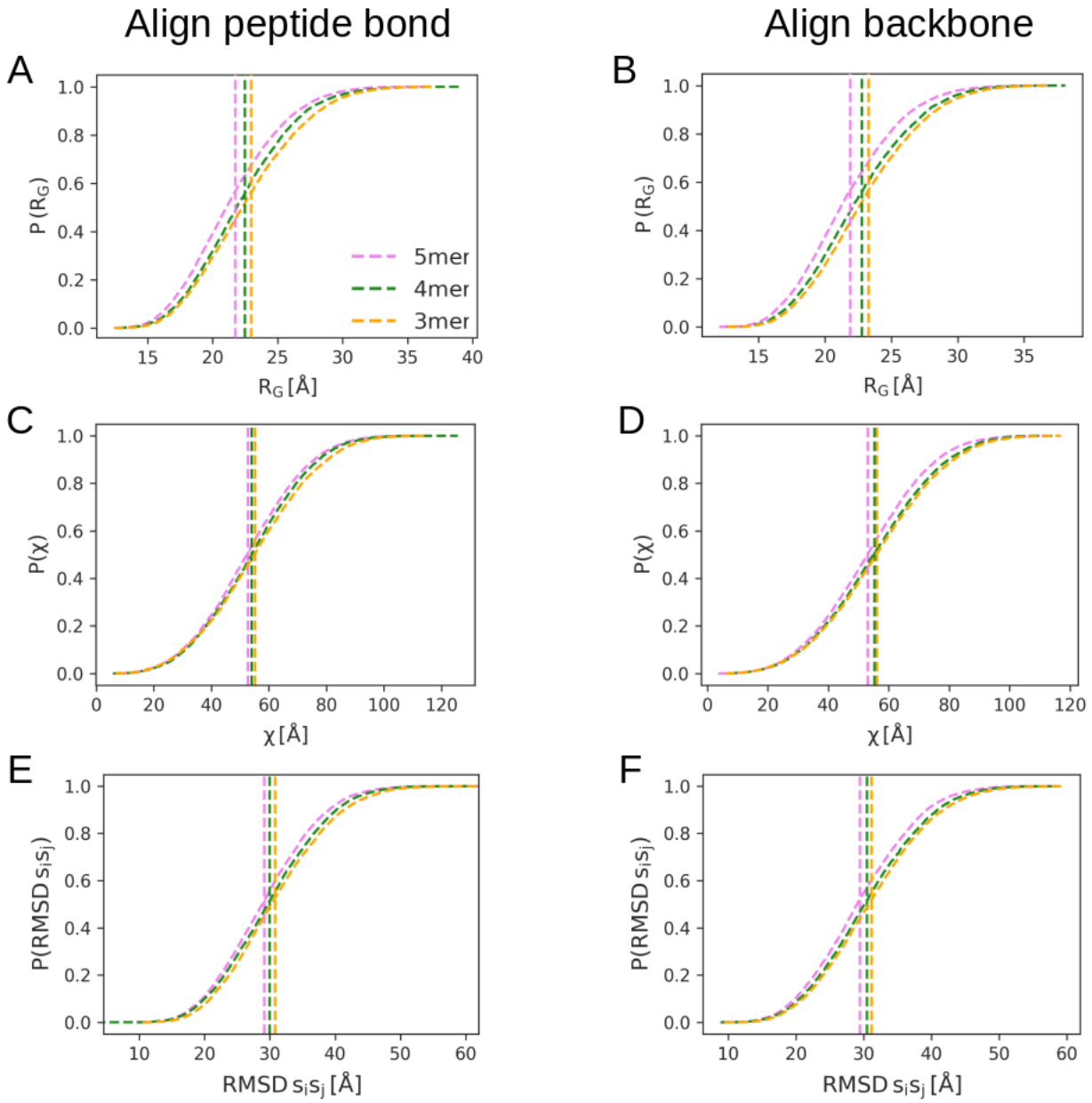

Figure S4: Comparison of different alignment procedures for models grown from fragments of different length. Cumulative distributions of the radius of gyration (A, B), end-to-end distances (here χ; C and D) and pairwise RMSD (E, F) of 10 000 models of a chain with 50 amino acids, each grown from pentamer (pink), tetramer (green), and trimer fragments (orange). Panels on the left and on the right show the distributions for models grown via alignment of the peptide bond and via alignment of the backbone atoms around C_α, respectively. Vertical dashed lines indicated the respective mean values.

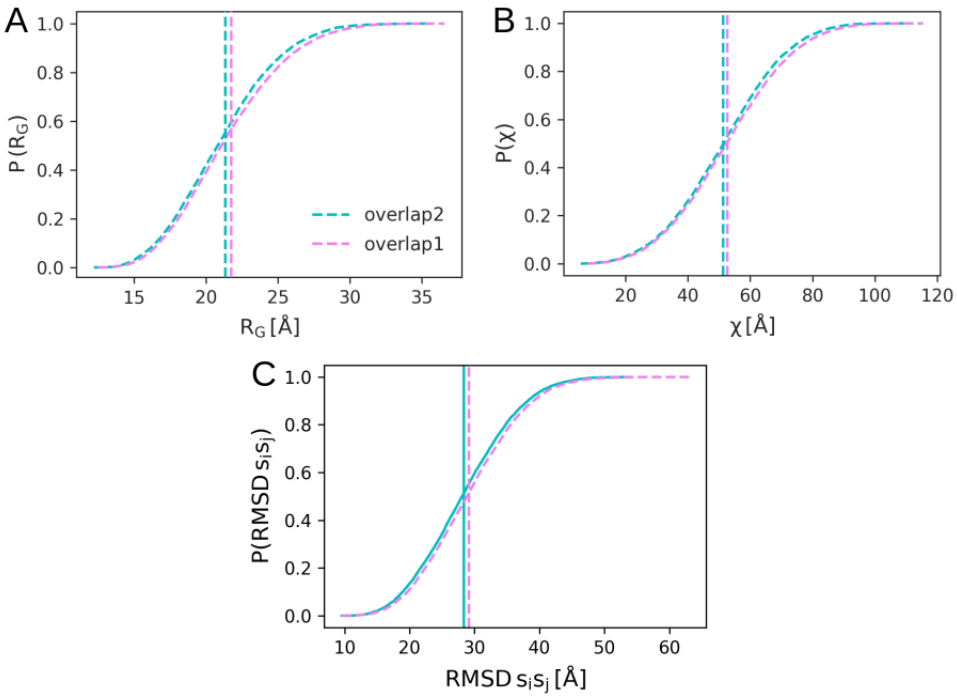

Figure S5: Effect of residue overlap between pentamer fragments on the structure of assembled chains. Cumulative distributions of the radius of gyration (A), the end-to-end distance (B), and the pairwise RMSD (C) of 10 000 models of a chain with 50 amino acids, each grown from pentamers with either 1 or 2 overlapping residues in adjacent fragments. Vertical dashed lines indicated the respective mean values.

Chemical Shifts for Fragments and Full-Length Models

Assembling full-length chains by aligning on the rigid peptide bond rather than the more flexible backbone at the merge point improves the NMR chemical shifts, as predicted for the MD fragments and the full-length model after the assembly. Figure S6A-D demonstrates that by growing the full-length models via backbone alignment (right column) a bias is introduced, shifting the Δppm chemical shifts predicted for the models towards more negative values relative to the fragment values (compare main text Figure 4C and D). By contrast, when growing full-length chains via alignment on the rigid peptide bond, Δppm chemical shifts of the assembled models agree well with those of the MD fragments (Figure S6 left column).

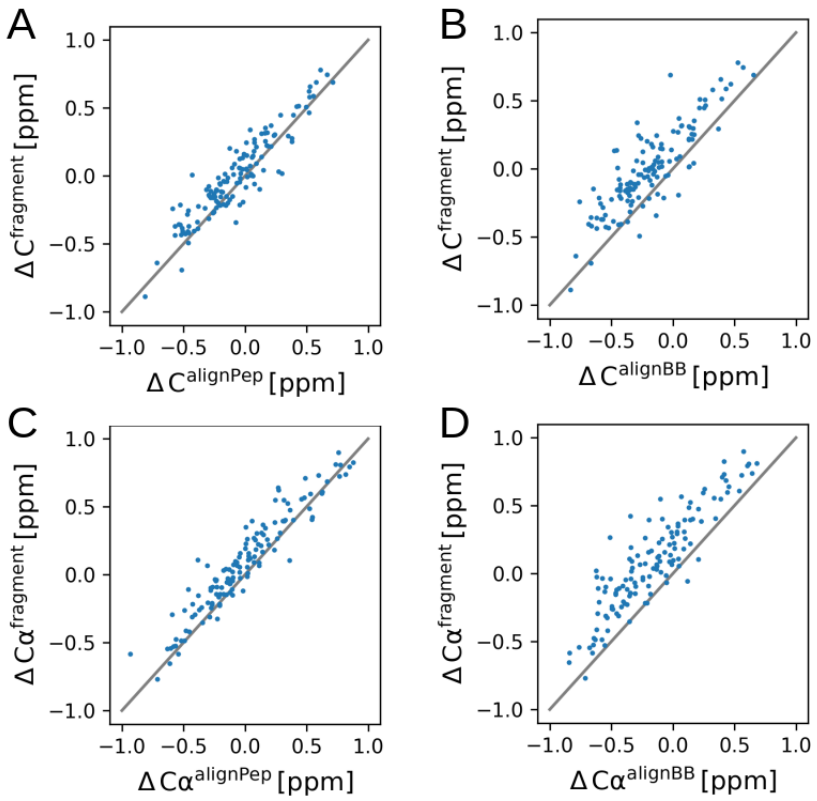

Figure S6: Correlation of secondary chemical shifts predicted for MD fragments and 10 000 full-length models. Left column: models grown via alignment of the peptide bond. Right column: models grown via alignment of the backbone.