JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

High-Throughput Simulations of Dimer and Trimer Assembly of Membrane Proteins. The DAFT Approach

Tsjerk A. Wassenaar,*[,†] Kristyna Pluhackova,[†] Anastassiia Moussatova,[‡] Durba Sengupta,[§] Siewert J. Marrink,[∥] D. Peter Tieleman,[‡,⊥] and Rainer A. Böckmann[†,⊥]

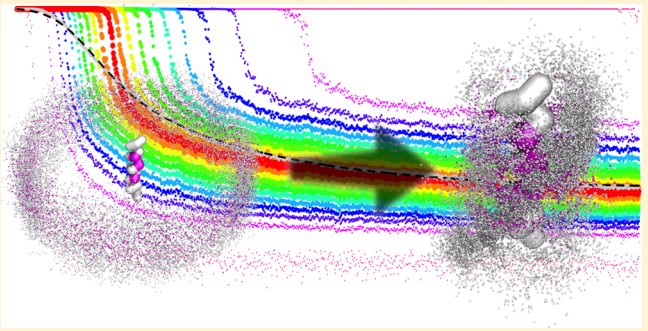
[†]Computational Biology, Department of Biology, Friedrich-Alexander University of Erlangen-Nürnberg, Staudtstrasse 5, 91058 Erlangen, Germany

[‡]Department of Biological Sciences and Institute for Biocomplexity and Informatics, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4

[§]National Chemical Laboratory, Dr. Homi Bhabha Road, Pune 411008, India

[∥]Groningen Biomolecular Sciences and Biotechnology Institute and Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands

ABSTRACT: Interactions between membrane proteins are of great biological significance and are consequently an important target for pharmacological intervention. Unfortunately, it is still difficult to obtain detailed views on such interactions, both experimentally, where the environment hampers atomic resolution investigation, and computationally, where the time and length scales are problematic. Coarse grain simulations have alleviated the later issue, but the slow movement through the bilayer, coupled to the long life times of nonoptimal dimers, still stands in the way of characterizing binding distributions. In this work, we present DAFT, a Docking Assay For Transmembrane components, developed to identify preferred binding orientations. The method builds on a program developed recently for generating custom membranes, called *insane* (INSert membrANE). The key feature of DAFT is the setup of starting structures, for which optimal periodic boundary conditions are devised. The purpose of DAFT is to perform a large number of simulations with different components, starting from unbiased noninteracting initial states, such that the simulations evolve collectively, in a manner reflecting the underlying energy landscape of interaction. The implementation and characteristic features of DAFT are explained, and the efficacy and relaxation properties of the method are explored for oligomerization of glycophorin A dimers, polyleucine dimers and trimers, MS1 trimers, and rhodopsin dimers. The results suggest that, for simple helices, such as GpA and polyleucine, in POPC/DOPC membranes series of 500 simulations of 500 ns each allow characterization of the helix dimer orientations and allow comparing associating and nonassociating components. However, the results also demonstrate that short simulations may suffer significantly from nonconvergence of the ensemble and that using too few simulations may obscure or distort features of the interaction distribution. For trimers, simulation times exceeding several microseconds appear needed, due to the increased complexity. Similarly, characterization of larger proteins, such as rhodopsin, takes longer time scales due to the slower diffusion and the increased complexity of binding interfaces. DAFT and its auxiliary programs have been made available from http://cgmartini.nl/, together with a working example.



1. INTRODUCTION

Membrane proteins form the main interface through which cells gather information about the outside environment. Their control of cell metabolism and gene transcription make them a focal point of experimental research as well as the primary target of pharmacological intervention. Membrane proteins actively modulate and are actively modulated by the lipid environment,[1−3] commonly interacting with one or more binding partners as part of complex regulatory pathways. To understand the mechanisms underlying processes such as signaling, fusion, and transport, it is necessary to get a microscopic view of the interactions between membrane proteins and what partners they bind. Unfortunately, the biomembrane environment is still challenging, not only in experiments but also in computational approaches, due to its complexity and frequently unknown exact composition.

For studying protein−protein interactions in solution, computational docking methods have been developed, which are relatively successful in the predicting and scoring of binding interfaces.[4,5] A good example is HADDOCK,[6] which performs many short simulations, starting from unbiased, maximally

Received: November 11, 2014

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

decoupled configurations. During the simulations, biasing potentials are applied, which are derived from experimental or bioinformatics data. These biasing potentials are aimed to drive the ensemble to specific interactions, identifying relative orientations matching the data.

In HADDOCK, the basic type of biasing potentials are ambiguous distance restraints between surface regions which have been designated "active" and "passive". However, it is also possible to make the complete surface passive and have no other bias than a potential to drive the components together. In this case, preferential orientation is solely caused by surface−surface interactions as described by the force field used, and the ensemble obtained will give a distribution of interfaces characteristic for the force field.

Unfortunately, docking methods are not readily applicable to heterogeneous molecular environments, such as membranes, and are unlikely to be able to include the modulating effect of lipids.[7]

Molecular dynamics (MD) simulations are another computational approach that can be used to complement experimental work and build comprehensive models of interactions on the level of molecules and atoms. MD simulations have already been used for several decades to study interactions between proteins in a membrane,[8−14] in particular also for NMR structure refinement.[15] More recently, the introduction of coarse grain (CG) models has made MD a standard approach for investigating the association of components in a membrane,[16,17] allowing simulations of larger systems and larger time scales. These studies typically report binding to occur within one to several microseconds, after which the formed complexes are found to be stable for the remainder of the simulations, up to 25 μs.[9,18−20] This observed stability may actually be a concern as it may reflect that the off-rate k_{off} is just too low to allow sufficient sampling in simulations. Indeed, it is known that even simple dimers such as glycophorin A (GpA) take significant time to exchange[21] and long-term atomistic simulations performed on a dedicated machine have recently strengthened this view.[22]

It is thus evident that the k_{off} of typical systems of interest will be too low to allow unbinding and rebinding events and hence will hamper convergence in brute force simulation approaches. On the other hand, docking approaches have proved successful in identifying bound orientations by sampling many association events, disregarding the k_{off}. Therefore, we here also focus our attention to the association of components, but using MD simulations. In addition, in contrast to the common objective of docking, where the simulations are biased toward a specific bound state, we identify potential bound states from unbiased MD simulations, i.e., relying only on a standard force field, without using biasing potentials. This keeps experimental and theoretical data available for posthoc comparison and validation and should yield a more comprehensive view on the ensemble of interactions.

The use of normal MD simulations allows a realistic, complex environment, including, e.g., lipids, solvent, and solutes. The binding of two components then depends on two factors: the probability of encounter and the probability to bind during an encounter. The first of these depends primarily on the size of the system and the diffusion rate, while the latter depends on the interactions and thus on the relative orientation. Together, these factors underly the k_{on}. So with equal *a priori* probalities of encounter, a distribution of bound orientations obtained from a set of simulations will reflect the probabilities of binding for the different orientations. We note that this need not be the same as the probability of being bound in a certain orientation, which also depends on transitions between bound states and on the existence of states that are not directly accessible, e.g., requiring a conformational change.

Against this background, we developed a method that combines a docking approach with coarse grain molecular dynamics simulations, which we call the Docking Assay For Transmembrane components (DAFT). DAFT allows setting up and performing high-throughput simulations to investigate interactions between components in a membrane, including lipid−lipid, protein−lipid, and protein−protein interactions. The number of protein components in a system can vary from none to nine, and the systems are set up to have unbiased initial configurations, to ensure equal *a priori* probabilities for encounters. Without a protein, the method allows high-throughput lipidomics screening,[23] including exploring specific interactions between lipids, and processes such as phase separation. With a single protein component, DAFT can be used to search for specific protein−lipid interactions, as an alternative to large and long time scale simulations.[24]

One particular application for which DAFT was developed is the dimerization of transmembrane helices. In this mode, our approach bears semblance with the Sidekick protocol presented recently,[25] which also performs multiple binding simulations using a coarse grain force field. Sidekick was developed as an integrated automated pipeline for simulation and analysis and can be used to assess the insertion of a transmembrane helix into a bilayer or to study the association of two helices. It offers a simple interface for use of large-scale computational resources, using a well-designed, automated back-end pipeline, with integrated analysis and handling of errors. However, Sidekick is currently limited to dimerization of helices, which are built from sequence, and to using pre-equilibrated bilayers consisting of dipalmitoylphosphatidylcholine (DPPC), dipalmitoylphosphatidylethanolamine (DPPE), and/or dipalmitoylphosphatidylglycerol (DPPG).[25]

DAFT was specifically developed to allow as input any combination of sequences, atomistic structures, or coarse grain structures. It was designed in line with the WeNMR modular software model[26] yet in a way that all options of all elements that are part of the workflow can be accessed from the master command line. This means that DAFT can harness the full power of our coarse-graining tool *martinize*,[27] the automated workflow *martinate*,[28] and the membrane/solvent builder *insane* (INSert membrANE).[23] In particular, the latter allows building arbitrary membranes and solvents for coarse grain simulations, with or without solutes, and even allows defining arbitrary lipids by specifying head, linker, and tail groups. This means that DAFT can be used to study dimerization in a complex lipid matrix, e.g., including cholesterol, sphingolipids, and/or glycolipids, and even supports bilayers as complex as a plasma membrane.[29]

Another key feature of DAFT is the specific use of periodic boundary conditions to maximally unbias simulations and optimize the dimer/multimer conformational space. This and the other features will be explained in detail later.

The DAFT machinery allows running a large number of simulations, aimed at generating an ensemble of bound states of the (protein) components, reflecting the association probabilities. The profile and the time scale of the ensemble convergence toward bound states, and the number of simulations required are the main focus points of this work.

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

In the following, we describe the framework in detail and explore the convergence properties for transmembrane (TM) helix dimers of wild type (WT) GpA, the GpA mutant G83I, and polyleucine. In addition, screening of a trimer assembly is performed, comparing the trimer forming peptide MS1 to polyleucine. Finally, results are presented for rhodopsin dimers, which are compared to brute-force simulations performed previously.[18,30]

2. BACKGROUND AND IMPLEMENTATION

DAFT is a generic framework for investigating protein−protein, protein−lipid, and lipid−lipid interactions in membranes, using (MARTINI[31−33]) coarse grain simulations. The basic approach for protein−protein association, putting the proteins in a membrane and investigating the interface(s) formed, is well-established, but the DAFT implementation has unique features, offers novel approaches, and allows setting up large-scale, high-throughput projects with minimal effort. In particular, the method is developed around a number of layouts (DAFT schemes), which set the optimal arrangement and corresponding periodic boundary conditions for a given number of protein components, ranging from 0 to 9. For a single protein, or none, the optimal unit cell is a hexagonal prism, but for 2−9 optimal arrangements were designed specifically for DAFT (Figure 1). These layouts originate from a means of creating enlarged systems of semicrystalline patches, such as the purple membrane, in such a way that the periodic images would be maximally decoupled.[34] We later realized that this could be particularly useful for distributing proteins for studying association, as each layout establishes a fixed separation between any two components, corresponding to equal encounter probabilities. This is exemplified in the first two images in Figure 1, which highlight the layouts for two and three protein components. For two, the optimal arrangement is like a chessboard, while a hexagonal tiling is used for three components. In either case, the setup ensures that each chain is only directly neighboring the other chains, such that any displacement will move it in the direction of a partner. From the resulting distribution of components, the actual unit cell can be derived, which is also shown in Figure 1. The same principle applies to higher numbers of protein components. Such higher order setups have not been explored yet, though, because the properties of lower order assemblies need to be understood first.

The DAFT work flow is shown in Figure 2. The components to study can be provided as sequence, as atomistic structure, or as coarse grain structure with topology. A sequence that is provided is first built as an atomistic model. By default, it will be built as a helix, but stretched conformations are also possible and the secondary structure can be specified explicitly. The atomistic models, including those generated from sequence, are subsequently converted to MARTINI coarse grain models using *martinize*.[27] The coarse grain models are then arranged on the corresponding DAFT layout. This is repeated a specified number of times, where the distribution of the components is randomized and each one gets a random in-plane rotation. Each configuration thus obtained is then passed through *insane*,[23,35] which builds lipids and solvent according to the composition and relative quantities specified. This provides a series of starting configurations that can be run automatically or transferred to a compute cluster, GRID,[26] or cloud facility for further processing.

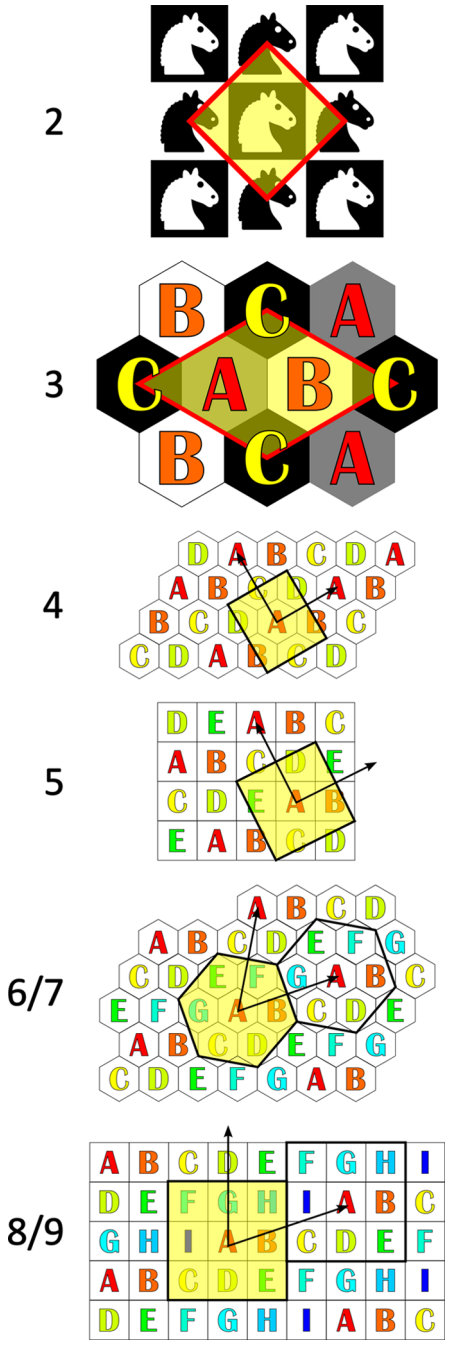


Figure 1. DAFT layouts for starting configurations. Layouts and corresponding PBC unit cells are defined for docking assays comprising between two and nine macromolecular components. For two components the setup resembles a chess board, such that each is surrounded by four copies of the partner. This yields a unit cell with a square base. With three components, a hexagonal tiling is used, in which each component is surrounded by three copies of each of the other partners. This yields a unit cell with a hexagonal base. The arrangement with four components is also based on a hexagonal tiling, but yields a unit cell with a rectangular base. With five components, an arrangement based on a square tiling is used, which also yields a unit cell with a square base. For both six and seven components the basis is again a hexagonal tiling, where one element is left unoccupied in the case of six components. The resulting unit cell also has a hexagonal base. Finally, with eight or nine components, square tiling is used for the arrangement, where one element is left unoccupied if there are eight components. The resulting unit cell has a square basis, but with skewed lattice vectors.

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

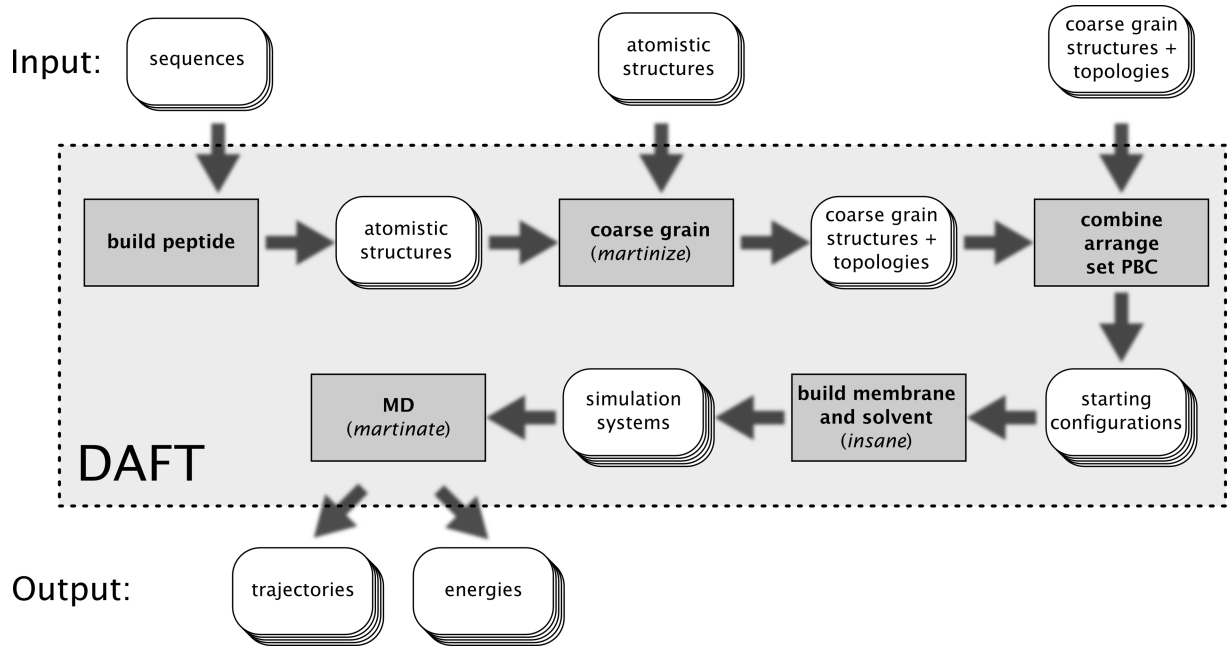


Figure 2. DAFT workflow. The DAFT protocol can take as input any combination of peptide sequences, atomistic structures, and coarse grain structures with topologies. The first are converted to atomistic structures with secondary structure as specified. The default secondary structure used is helical. All atomistic structures are converted to coarse grain structures with topologies, using *martinize*.[27] The coarse grain structures are combined, and for each combination a specified number of starting structures is generated, according to the corresponding DAFT layout. Each resulting starting configuration is processed with *insane*, which builds the membrane and solvent. The resulting systems are then run using *martinate*,[28] which performs energy minimization, position restrained *NVT* equilibration and *NpT* equilibration, followed by the production run, yielding the sets of trajectories and energies.

DAFT also allows setting up large-scale projects by providing multiple component definitions and specifying which combinations should be made. Examples are the setup of all combinations of heterodimers for a set of 10 helices or of all possible homodimers and -trimers and/or other combinations. In turn, this can be easily integrated in a work flow scanning and selecting TM helix sequences, e.g., based on a BLAST[36] search.

DAFT and the auxiliary programs, including the automated CGMD workflow *martinate* and the membrane/solvent generation tool *insane*, have been made available through the MARTINI Web site at http://cgmartini.nl/. A tutorial exemplifying the usage has also been provided there.

In the following sections we present a number of studies that were conducted to investigate the efficacy of the method and the convergence properties of the binding interfaces as a function of the number of simulations and of the time per simulation. In addition, a number of approaches are presented for analyzing the resulting ensembles of simulations.

3. METHODS

3.1. Simulations. All helix dimer simulations were performed according to the general DAFT protocol (see Figure 2). DAFT builds the sequences given as atomistic helical models and aligns them with the *z*-axis using PyMOL.[37] These models are then subsequently coarse grained using *martinize* and distributed over the corresponding DAFT layout, using a specified distance and random in-plane rotations. 1,000 starting structures were generated for both GpA systems and 500 for each of the other ones. As part of the automated work flow, the resulting configurations were processed with *insane* to generate a membrane and solvent and were subsequently energy minimized. Each system was briefly equilibrated using 10 ps of *NVT* MD, using a 2 fs time step, and 100 ps *NpT* MD, using a 20 fs time step, to bring the temperature and pressure within the range corresponding to 310 K and 1 bar. The temperature was controlled by coupling to an external heat bath using the Bussi thermostat with a coupling time of 1 ps.[38] Pressure was controlled using weak semiisotropic coupling to a reference pressure of 1 bar, with a compressibility of 3×10^{-4} bar^{-1} and a coupling time of 3 ps.[39] Production simulations of helix dimers were performed using an integration time step of 20 fs and a simulation length of 0.5 μs each. All simulations were performed using GROMACS 4.5.x.[40]

3.2. Systems. *3.2.1. Glycophorin A WT and G83I Dimers.* Glycophorin A wild type (WT) was built as helix from the sequence 72-RASL IIFGV MAGVI GTILI N-91, which corresponds to the positive control commonly used for TOXCAT assays.[41] Similarly, the negative control GpA G83I mutant[42] was built as helix from the sequence 72-RASL IIFGV MAIVI GTILI N-91. Both systems were processed according to the protocol described previously which resulted in a POPC bilayer of 56 lipids in each layer and a 1:9 lipid:CG water ratio.

3.2.2. Polyleucine Dimers and Trimers. The polyleucine TM helix was built from the sequence 1-RLLLL LLLLL LLLLL LLLLL LLLLR RLI-28[43] and processed according to the same protocol as used for GpA. For the dimerization study the protein surroundings were equal to those of GpA, while for the trimerization simulations the bilayer measured 84 POPC lipids per leaflet, with the same water to lipid ratio as for dimers.

3.2.3. MS1 Trimers. The model transmembrane peptide MS1 that has been used experimentally to study TM helix trimerization in micelles[44] was built as helix from the sequence 2-QLLI AVLLL IAVNL ILLIA VARLR YLVG-29. Using DAFT, three copies were embedded in a POPC bilayer with 3 nm distance to each other. Each resulting system contained 3

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

MS1 helices, 80 POPC lipids in each leaflet, and MARTINI water according to a 1:11 POPC:W ratio. The DAFT assay consisted of 500 simulations, each simulated for 3 μs.

3.2.4. Rhodopsin Dimers. An ElNeDyn[45] coarse grain model of rhodopsin, used previously to study association using a large-scale brute-force approach,[18] was kindly provided by Dr. Xavier Periole (University of Groningen) and set up using DAFT for 500 dimer simulations of 1 μs each. Each simulation was set up with an initial distance of 2.5 nm between proteins, based on the circumscribed radius of the protein in the membrane plane. The system was embedded in a DOPC bilayer, comprised of approximately 100 lipids in each leaflet, and was solvated with an approximate 17:1 coarse grain water to lipid ratio.

3.3. Analysis. A single DAFT assay yields a set of independent simulations or paths for a given number of components under specified conditions, including lipids, solvent, and temperature. Here, the analysis focuses on the relative orientations and the interactions, as well as on the convergence of the ensemble over time. Auxiliary tools were developed for analysis of contacts and of orientations, which have also been made available on http://cgmartini.nl. Posthoc statistical analysis was performed in R.[46]

3.3.1. Interaction Energy. For each simulation from an assay, the (nonbonded) interaction energy between any pair of partners was extracted and the distributions were plotted as a function of time. Convergence of the (interaction) energies over time across the simulations from one assay was analyzed by inspecting the evolution of the distribution vigintiles (5% quantiles), which give a 21-point distribution summary, including the minimum and maximum. Together with the evolution of the mean, this provides a view on the progression of the whole ensemble.

The relaxation of the mean interaction energy was further investigated by nonlinear least-squares (NLS) fitting of the data to a model of the form

$$\bar{E}_{\mathrm{I}}(t) = \frac{1 - \mathrm{e}^{-at}}{1 + \mathrm{e}^{-b(t-m)}}\bar{E}_{\mathrm{I},\infty}$$

This model combines a simple exponential relaxation with a switching function. The rationale behind the model is that the time evolution follows a two-stage process. The first stage is diffusion from the initial state to the general unbound state. This part is described with a sigmoidal function, with rate b and midpoint m. From the unbound state, the binding proceeds as a relaxation process, that is here modeled with a single-exponential decay toward a plateau value $\bar{E}_{I,\infty}$. The NLS fitting was performed in R, using the function "nls". The model was fitted progressively using different end times, to assess the convergence of the parameter estimates. Standard deviations around the plateau values were determined from the residuals, and standard errors of the mean were obtained by dividing the standard deviations by the square root of the effective sample size (statistical efficiency), determined with the R package "coda".[47]

3.3.2. Particle Distribution. Particles are projected on the Eckart (internal) frame of one of the components by performing a least-squares fit on that component. The resulting distributions are visualized in PyMOL. For each configuration, the color and the scale of the spheres is adapted to reflect the interaction energy.

3.3.3. Proximity Maps. For each frame, a contact map is determined by calculating all pairwise distances, d_{ij}, and converting those to a measure of similarity (proximity), using a kernel-like function:[48]

$$p_{ij} = \mathrm{e}^{-ad_{ij}^{2b}}$$

This measure is 1 for identity and decays smoothly to zero for larger distances, and is the same function used by *martinize*[27] to set up an elastic network (RubberBands) for MARTINI. The constants a and b determine the decay profile. The benefit of using proximities is that averaging is robust against outliers at large distances, and the average obtained is a direct relative measure of the interaction between particles across the ensemble. A p_{ij} close to 1 signifies that the particles i and j are in close proximity in most of the configurations. The proximity matrices provide a view on the interactions between sets of atoms, similar to contact maps used in other studies, but without the need to specify a threshold value for determining contacts. In this study we use the coefficients $a = 0.01$ and $b = 2$. We note that the coefficient a corresponds to distances measured in nanometers.

3.3.4. Relative Orientations Using Euler Angles. To characterize the binding configurations, the ensembles of relative orientations are analyzed. This is done by first determining the transformation required to superimpose the Eckart frames of the components. For a given component, the Eckart or internal coordinate frame E is determined from the center of mass and the principal components, which are obtained by diagonalizing the mass tensor.[49] The Eckart frame orientation $E(t)$ at time t is defined as the transformation required to superimpose the reference structure onto the structure at time t by a least-squares fit. This transformation consists of a proper rotation matrix and a displacement vector. Then for two components A and B, the relative orientation is defined as the transformation required to superimpose $E_{\mathrm{A}}(t)$ onto $E_{\mathrm{B}}(t)$. This is equivalent to stating that the relative orientation between A and B is the orientation of $E_{\mathrm{B}}(t)$ in the frame of A: $R_{\mathrm{AB}}(t) = E_{\mathrm{B}}(t)E_{\mathrm{A}}(t)^T$. To facilitate the interpretation, the orientation is subsequently decomposed into the polar coordinates of the center of mass (COM) of B, yielding the distance r, the out-of-plane shift α and the position β, and the ZYX Euler angles derived from the rotation matrix, giving the phase ϕ, the forward tilt θ, and the sideways tilt ψ, the latter of which is related directly to the crossing angle and handedness of binding.

The reference structure is usually chosen to be aligned along the membrane normal, and for homologous structures a structure alignment is performed on the references to have a consistent definition of the Eckart frame for those components. In effect, this means that a homodimer will use the same reference structure to determine the rotation, such that a symmetric binding will give a phase ϕ of 180°.

The binding position and phase are visualized as 2D density maps using a circular two-dimensional (2D) kernel-density estimate (KDE) with a Gaussian kernel function. The routine to calculate the circular 2D KDE was based on the regular 2D KDE method available in R/S-PLUS,[50] adapted to have the densities wrapped to account for periodicity of the parameters.

3.3.5. Simulation Time and Number of Simulations. The main aim of this work is determining the time scales and number of simulations required to assess the interactions of proteins in a membrane. The minimal time per simulation is determined by following distributions of several properties for the ensemble of simulations over time and determining the

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

point where the most significant binding orientations can be identified. If the aim is investigating the difference between two sets of simulations, as in the case of GpA WT and G83I, then the evolution of the difference distribution between the ensembles is followed.

The number of simulations needed for sufficient sampling was determined using bootstrapping with replacement[51] from the total distributions for different sample sizes. A total of 1,000 bootstrapped samples of each size were used to construct 95% confidence intervals of a distribution of some property or of the difference distribution between two sets. Resampling and plotting was performed in R.[46]

4. RESULTS AND DISCUSSION

4.1. Glycophorin A, Wildtype, and G83I Mutant, Polyleucine. To investigate the efficacy and the relaxation properties of the method, assays were first run for glycophorin A (GpA) wild type, for the GpA mutant G83I, and for a nonbinding polyleucine (PL) TM helix. GpA is a well-studied transmembrane helix dimer and used as positive control for dimerization in TOXCAT experiments, whereas the G83I mutant is used as negative control in such experiments. The dimerization of GpA with the MARTINI force field has been demonstrated and characterized before and was shown to agree with experimental data.[19,52]

The aim of this work was to assess the number and length of simulations required to give a consistent view of the interaction ensemble of transmembrane peptides and proteins and to contribute to understanding the statistical and convergence/relaxation properties of ensembles of simulations. To this end, we performed a total of 1,000 simulations of 512 ns for each GpA system and 500 runs for PL.

In the following sections, the results are discussed in detail, focusing on particle distributions, evolution of the interaction energy distribution, the relative orientations of dimers observed, and the dependence on the number of simulations performed.

4.1.1. Binding Distributions. The particle distributions shown in Figure 3 provide a quick view of the specificity or aspecificity of binding. The top panel of Figure 3 shows that for PL there is only a little density of the second helix around the reference, exemplifying the lack of binding. The results for GpA wild type stand in sharp contrast with this. The side view clearly shows a strong band corresponding to a right-handed helix. Apparently, a majority of structures has the same binding mode, even allowing identification of individual backbone beads. The glycine residues of the GxxxGxxG motif, indicated in dark gray, line the binding interface, and the same view is obtained from the top, where a helical wheel representation of the reference chain is added showing which residues form the interface.

The particle distributions of GpA and PL reflect the binding statistics. In the case of GpA all but one of the systems form a dimer, while in PL only 30% of pairs had any direct interaction. These first results already suggest that DAFT, in conjunction with the MARTINI 2.2 force field, can distinguish between binding and nonbinding peptides.

Interestingly, also in the case of GpA G83I all but one of the systems formed a dimer. However, the resulting particle distribution, shown in the bottom panels of Figure 3, is more diffuse than the distribution in GpA wild type. The right-handed dimer form, binding at the GxxxGxxG side, is much less populated, showing only a faint trace in the side view. The same is seen from the top, where the density is less and of less

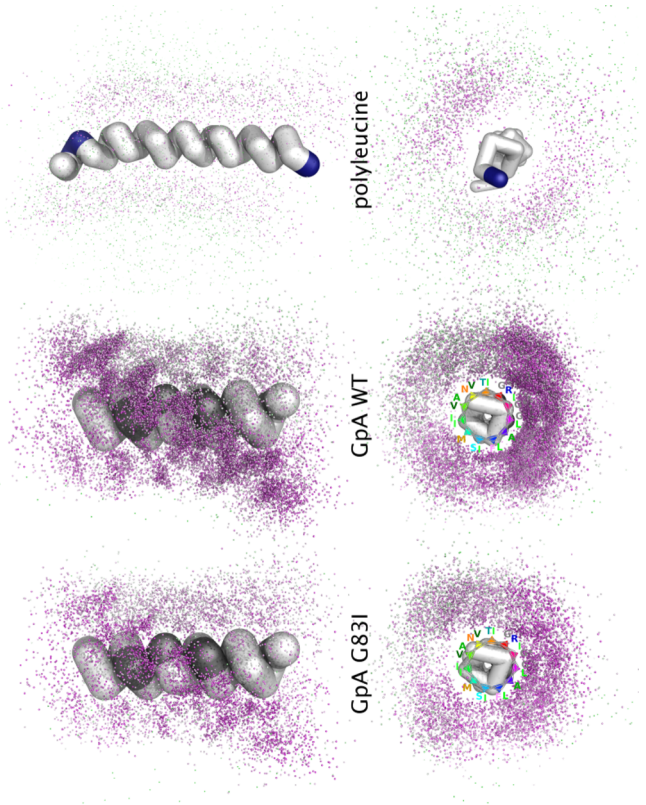


Figure 3. Particle distributions of one component around the other from final ensembles. For each assay, the last frames were collected and least-squares fit onto the first chain. Particle sizes and colors are consistent in all panels and reflect the interaction energy of the dimer. Lowest energies (approximately −300 kJ/mol) are colored purple; green particles correspond to interaction energies close to zero. The top panels show a side and top view of polyleucine, where little density is seen around the reference chain. The middle panel gives results for GpA wild type, showing a clear band of increased density corresponding to a right-handed dimer. The glycines in the GxxxGxxG motif are colored dark gray, and in the top view a helical wheel representation is added to show the locations of the side chains. The lower panels give the results for GpA G83I mutant. Here, the band corresponding to a right-handed helix is still discernible, but less distinct. The top view also shows that the binding is more diffuse and less strong, reflected in the decreased intensity of the purple color.

intense color, suggesting that there are fewer dimers with lower interaction energy. Presumably, this is caused by steric hindrance from the isoleucine side chain.

4.1.2. Interaction Energy Distribution over Time. The particle densities provide a first qualitative view on the formation of dimers. However, to investigate the dimerization and relaxation properties in particular, more quantitative measures are needed. One of the most obvious of these is the interaction energy. At the start of the simulations the components are separated and all have zero interaction energy. During the simulations, the components diffuse toward each other and bind, causing the distribution of interaction energies to shift downward. In the end, the ensemble of simulations should converge to an equilibrium. Because of the number of simulations, this process is expected to proceed as a relaxation process.

The evolution of the distributions over time is shown in Figure 4. Each panel in this figure characterizes the distribution at each time by marking the 5% points (vigintiles), which

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

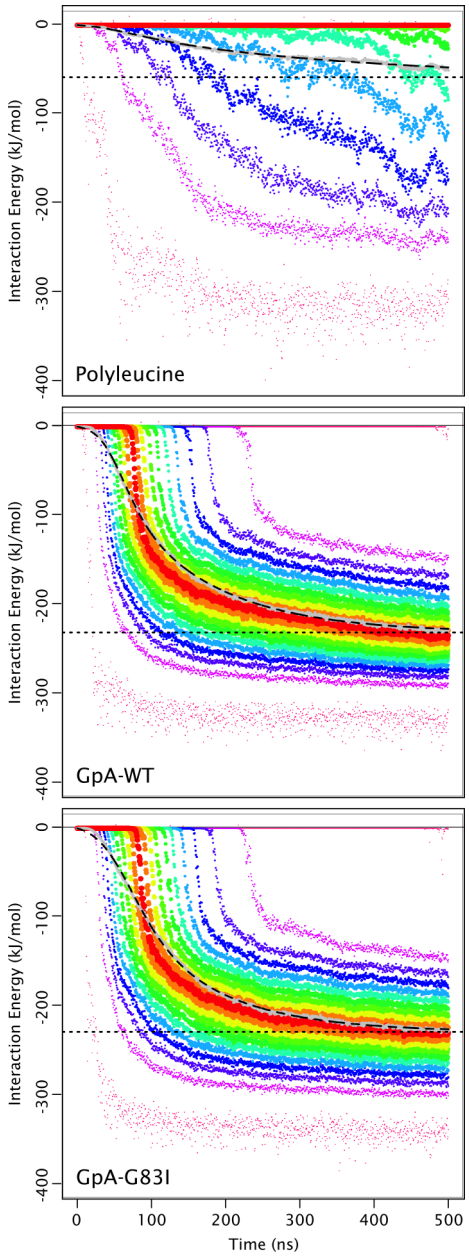


Figure 4. Interaction energy between peptides in dimers over time. The graphs show the time evolution of the interaction energy distribution polyleucine (top), GpA WT (middle), and GpA G83I (bottom). At each time, the distribution is characterized by the vigintile (5%) points. The median of the distribution is drawn in red, and vigintile points away from the median are colored using a rainbow gradient and drawn with successively smaller dots. The lowest and highest points (smallest dots, pink) indicate the minimum and maximum of the distribution. The gray line shows the progression of the mean interaction energy, the black dashed line shows the fitted model, and the horizontal black dotted line marks the estimated plateau.

together allow making inferences about the extent of the distribution, but also the shape, skewness, and the presence of multiple energy levels. In addition, the graphs show the evolution of the mean interaction energy in gray, together with the fitted mixing/relaxation model as black dashed lines. Dotted black lines indicate the estimated plateau values for the means.

In the case of PL, only 8 out of 21 vigintile bands (including the minimum) are below zero, indicating that in 65% of the simulations there is no interaction between the units. About 10% of the simulations have interaction energies below −200 kJ/mol. Interestingly, there are even a few simulations that yield dimers with an interaction energy as low as −300 kJ/mol. The mean interaction energy can be fit with the model function, giving a mixing rate and midpoint of 3.17×10^{-2} ns^{-1} and 37.0 ns, respectively, and a relaxation rate and plateau value of 3.22×10^{-3} ns^{-1} and −58.68 ± 0.95 kJ/mol. The standard error of the mean (SEM) appears to be very low. To understand this, it should be realized that this is the SEM about the mean value of the average interaction energy from 500 simulations. The distribution of averages will tend to a Student's *t*-distribution with 500 degrees of freedom, which will be narrow, even if the underlying distribution for a single observation from one simulation has a very large standard deviation (SD). To exemplify this point, we also determined the SEM for the mean average pressure from the 500 simulations. The pressure is known to have extreme fluctuations and, across simulations, was determined to have mean 1.11 and SD 42.53 bar. However, the SD of the average across simulations over time is 4.38 bar, and the SEM determined over the last 100 ns is only 0.307 bar.

The low SEM suggests a high accuracy. However, progressive fitting shows that the estimate for the plateau is not stable toward the end of the simulation. This means that no inferences can be made about equilibrium properties, despite having a good fit.

The contrast with GpA is again sharp, for both variants. Here, the first dimers are formed within 20 ns, and after 200 ns dimers were formed in more than 90% of the simulations. The lowest vigintiles, corresponding to the dimers with minimal interaction energy, decay toward plateau values of −326.86 ± 0.06 kJ/mol with a relaxation rate of 6.59×10^{-2} ns^{-1} in the case of WT, and −338.86 ± 0.04 kJ/mol with a relaxation rate of 5.1×10^{-2} ns^{-1} in the case of G83I. The mean interaction energies can be fit with the model function, yielding a mixing rate and midpoint of 5.14×10^{-2} ns^{-1} and 50.3 ns for WT and 3.57×10^{-2} ns^{-1} and 56.7 ns for G83I. The decay rate and plateau value obtained for WT are 7.94×10^{-3} ns^{-1} and −231.79 ± 0.08 kJ/mol, while for G83I the corresponding values are 8.47×10^{-3} ns^{-1} and −229.50 ± 0.03 kJ/mol. Progressive fitting shows that the estimates for the plateau values become stable after 300 ns for WT and G83I.

The lower mixing rate for G83I reflects slower diffusion, which was found to be caused by a more favorable protein−membrane interaction energy (ensemble mean difference $E_{WT/mem} - E_{G83I/mem} = 8.808 \pm 0.019$ kJ/mol for single molecules).

4.1.3. Proximity Maps. A more detailed view of the nature of the interactions and their evolution is shown in Figure 5, which shows the average proximity maps as a function of time. Each map is obtained by averaging the individual maps obtained from the corresponding time frames from one assay. A proximity measure is used, rather than distances or contacts, because averaging distances is not robust, while the use of contacts with respect to a given threshold value discards information about the underlying distribution. In a way, the use of proximities is a lazy approach, as it makes it unnecessary to filter the data for contact situations, while it still reflects features from the underlying distance distribution. Here, the proximity maps support the view obtained from the densities. First of all, the map for PL shows only faint white bands along the

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

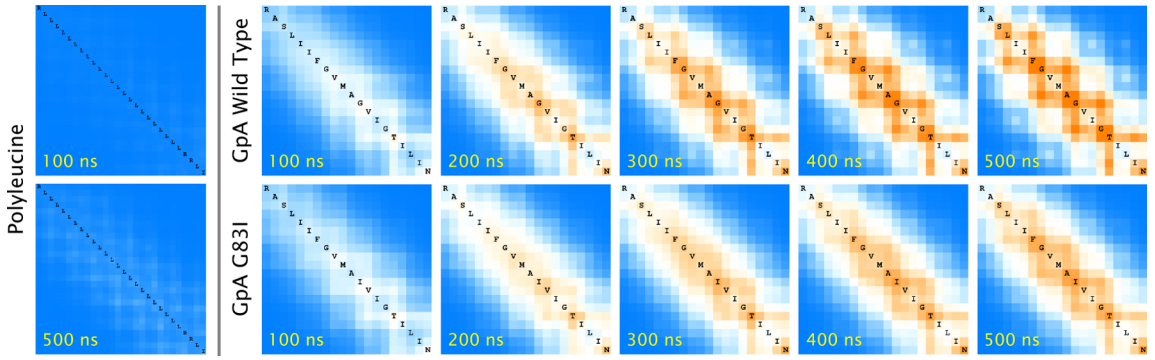


Figure 5. Convergence of proximity maps over time. Proximity maps were calculated for each pair of peptides at each time, according to the procedure described in Methods. At the left side, the proximity maps for the polyleucine dimer are shown at 0 and 500 ns, exemplifying the lack of (specific) interactions. The right top row shows the proximity maps for GpA wild type at intervals of 100 ns, demonstrating the emergence of a distinct pattern of interactions. The bottom right row shows the same for GpA mutant, in which case the pattern established is less well defined than that seen for wild type.

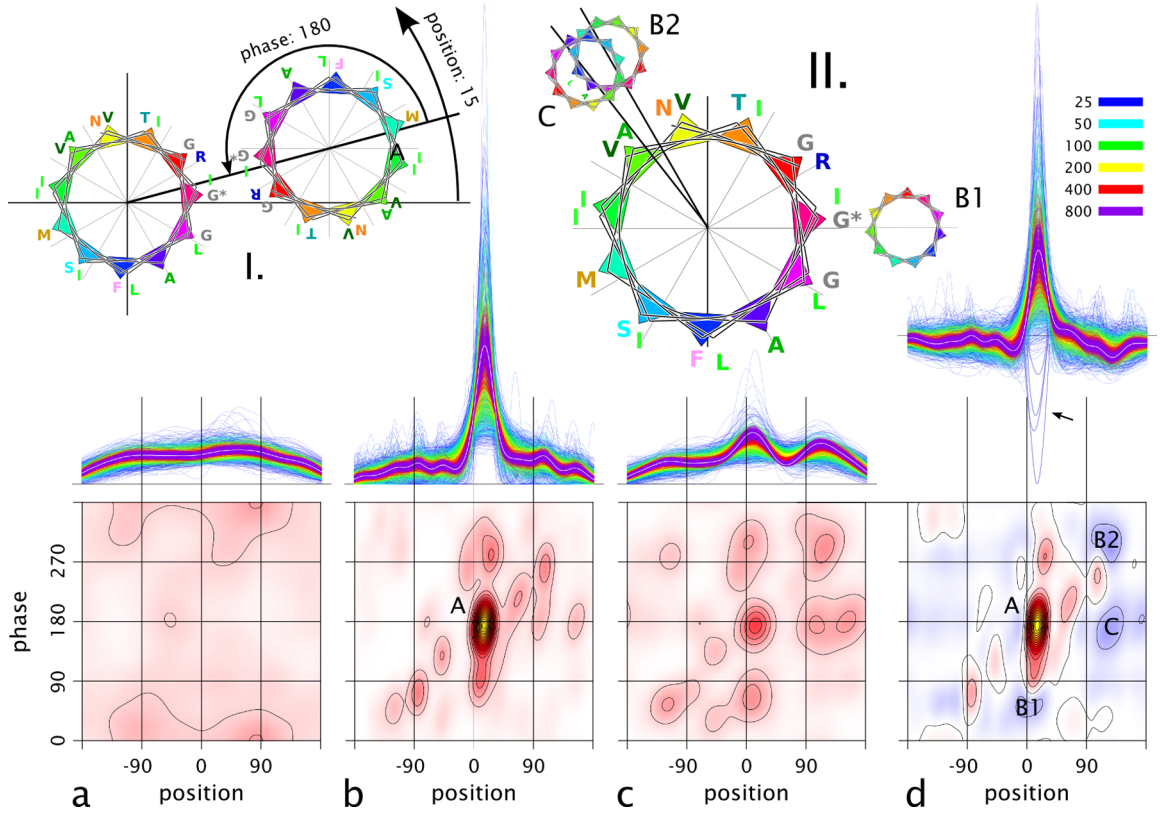


Figure 6. Convergence of orientation distributions as a function of the number of simulations. Relative orientations are summarized by binding position and binding phase, as illustrated in inset I for GpA. This inset shows a helical wheel representation of the most abundant GpA dimer, in which the glycine residues are facing each other. The lower panels show the distribution of orientations, expressed as position and phase. Polyleucine (a) shows an almost uniform distribution, exemplifying the lack of preferential orientation. In contrast, GpA wild type (b) shows a clear maximum, marked A, corresponding to the orientation depicted in inset I. The G83I mutant (c) shows less specificity than the wild type, but more than polyleucine. The difference distribution of wild type and mutant (d) shows regions of negative density (blue) illustrating alternative binding sites (B1/B2, C), which are more populated in the mutant. These alternative orientations are illustrated in inset II. Above each 2D position/phase density plot, the corresponding 1D position distribution is drawn. The central, white line shows the distribution taking all simulations together. Each set of colored lines represents a bootstrapping run, drawing samples of 25 (blue), 50 (cyan), 100 (green), 200 (yellow), 400 (red), or 800 (purple) simulations. For the difference distribution, two samples of the indicated size were drawn, one from GpA wild type and one from the mutant. An arrow marks samples with an inverted profile, indicating that orientation A was occupied more in the mutant than in wild type, opposite to the full ensemble results.

diagonal, indicative of a few nonspecific contacts, while in both GpA variants the interactions are stronger.

For GpA it is clear that the interaction between the two helices is less well defined in the mutant than in wild type. In particular, the GxxxGxxG motif yields a distinct pattern in WT, while the pattern is more diffuse in G83I. The time series in Figure 5 also shows that the characteristic proximity profiles emerge between 300 and 400 ns of simulations, while the

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

difference between the wild type and mutant can be assessed on a time scale of around 250 ns on a qualitative level. These time scales correspond to the times at which the interaction energy plateaus become estimable, suggesting that for these peptides and the membrane composition used, such time scales are sufficient for a qualitative assessment. We stress that this does not imply complete convergence of the ensemble, which would be required to allow quantitative evaluations, including mean residence times and k_{off}. This still requires sampling unbinding events.

4.1.4. Orientation Analysis. The proximity maps provide a view on the residue−residue interactions and can be used to track the evolution of the interface(s) over time. However, it is difficult to make statistical inferences about the distribution of interacting pairs. Therefore, to further characterize the interactions and assess how many simulations are required per assay, the relative orientations were investigated. To this end, we updated a method developed previously that described relative orientations between domains in terms of a displacement vector and three Euler angles, derived from the rotation matrix that superimposes the internal coordinate (Eckart) frames.[53] The alternative decomposition used here rewrites the displacement vector and the rotation matrix as a COM−COM distance and five angles, yielding an extension to polar coordinates.

The result of the orientation analysis is shown in Figure 6. Inset I at the top left shows a helical wheel representation of GpA, placed at the origin of its own internal coordinate frame. The other helical wheel is projected onto the frame of the first, with the relative orientation corresponding to the most abundant dimer species. The most important parameters of the decomposition of the orientation are indicated. The first is called the position, because it indicates where the partner binds from the viewing point of the reference. The position is defined as the angle between the line through the centers of mass and the *X*-axis. The second parameter is the phase, which corresponds to the rotation of the other chain around its long axis. A phase of 180° indicates symmetric binding, while a phase of 0° corresponds to back-to-back binding.

The panels in the bottom row show the position/phase density plots, and above each of these panels, the 1D density plot of the position is drawn. These plots are based on the last frames from all simulations. For the 1D density plots, the white lines in the center correspond to the results of taking all of these points into account, whereas each colored line corresponds to a bootstrap sample of the size indicated.

For PL (Figure 6a) it is clear that there is marginal preferential orientation. The bit of structuring observed is likely due to repulsion between the arginine anchors. The bootstrapped samples of the position show that for sample sizes of 200 or larger the total distribution is reproduced well. For lower sample sizes, the noise becomes larger, decreasing the resolution with which binding can be characterized. Yet a more problematic consequence of too small sample sizes appears to be the risk of misclassification. In particular, about 7/100 of the lines from samples of 25 simulations (blue) show increased density around a position of 60°, overestimating the specificity of binding.

In the case of GpA WT (Figure 6b), there is a single most populated orientation, labeled A, corresponding to the one shown in inset I. There are a few minor binding orientations, which is consistent with the views obtained from the particle distributions (Figure 3). The preferential symmetric binding at a position of 0°, involving the GxxxGxxG motif, is exemplified by the maximum of the 1D density plot. However, the additional positions are also reflected in the position density. Drawing subsamples from the total of 1,000 simulations shows that with a sample size of 400 simulations (red), these features can still be seen. However, with samples of 200 simulations (yellow) the resolution becomes too low to make inferences about these more subtle features. The overall shape is still reproduced for sample sizes as low as 100, but below that limit, the probability to over- or underestimate even the most significant features becomes too large. In particular, samples of size 25 are prone to be affected by outliers.

The position/phase map of the GpA G83I mutant in Figure 6c shows significantly reduced binding at site A, and more prominent binding at positions ±120° and/or with ±60° phase. The difference with WT can be seen clearly in the position/phase difference plot shown in Figure 6d. The latter plot shows that there are three main alternative locations for dimerization of the mutant. These are marked B1, B2, and C, and the relative orientations are depicted in inset II in Figure 6 as helical wheels. From the wheel representations it can be seen that B1 and B2 are the same binding mode, but with the roles of reference and partner chain switched. As a matter of fact, for phases other than 180°, there must be two equivalent binding modes.

The 1D position density plot for G83I again shows how too small sample sizes may overestimate or obscure characteristic features of the profile. Yet the consequence of this is even more clear when looking at the difference distribution, shown in Figure 6d. Each line in this plot is the difference between the density curves obtained from a sample from WT and a sample of G83I, both of the same size. This is equivalent to comparing two sets of simulations of the corresponding size. The most striking feature of the graph is that with sample sizes of 25 several lines are strongly negative at the 0° position (indicated with an arrow). Apparently there is a probability of several percent to draw the conclusion that the mutant binds there more than WT, contrary to the actual situation.

Taken together, these results suggest that for studying transmembrane helix dimer association no less than 100 simulations of 300−400 ns should be performed, while to be able to assess more subtle features in the interaction energy landscape, it is advisible to perform 200−400 simulations. Systems with lower dimerization propensity take longer to converge, and therefore, in our studies, we commonly start with assays consisting of 500 simulations of 500 ns each.

4.2. Trimers. The results on PL and GpA dimers show how DAFT allows constructing a view on the interactions of TM helix dimers within reasonable time. For assembly of trimers, the increased complexity is expected to require longer simulations, but possibly within reach. Therefore, we set out to use DAFT to explore the trimerization of the GCN4 derived peptide MS1 and compare the results to those of a polyleucine trimer (PL3) setup. The trimer simulations are based on a hexagonal arrangement, which makes the unit cell for a given distance between components 30% larger than the corresponding dimer setup, which is based on a square arrangement. The MS1 simulations were initially run for 1 μs and later extended to 3 μs, because of slower convergence. Based on the lack of interaction observed for both PL dimers and trimers, the trimer assay for polyleucine was run until 512 ns.

The evolution of the interaction energy distribution is shown in Figure 7, together with the resulting particle distributions.

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

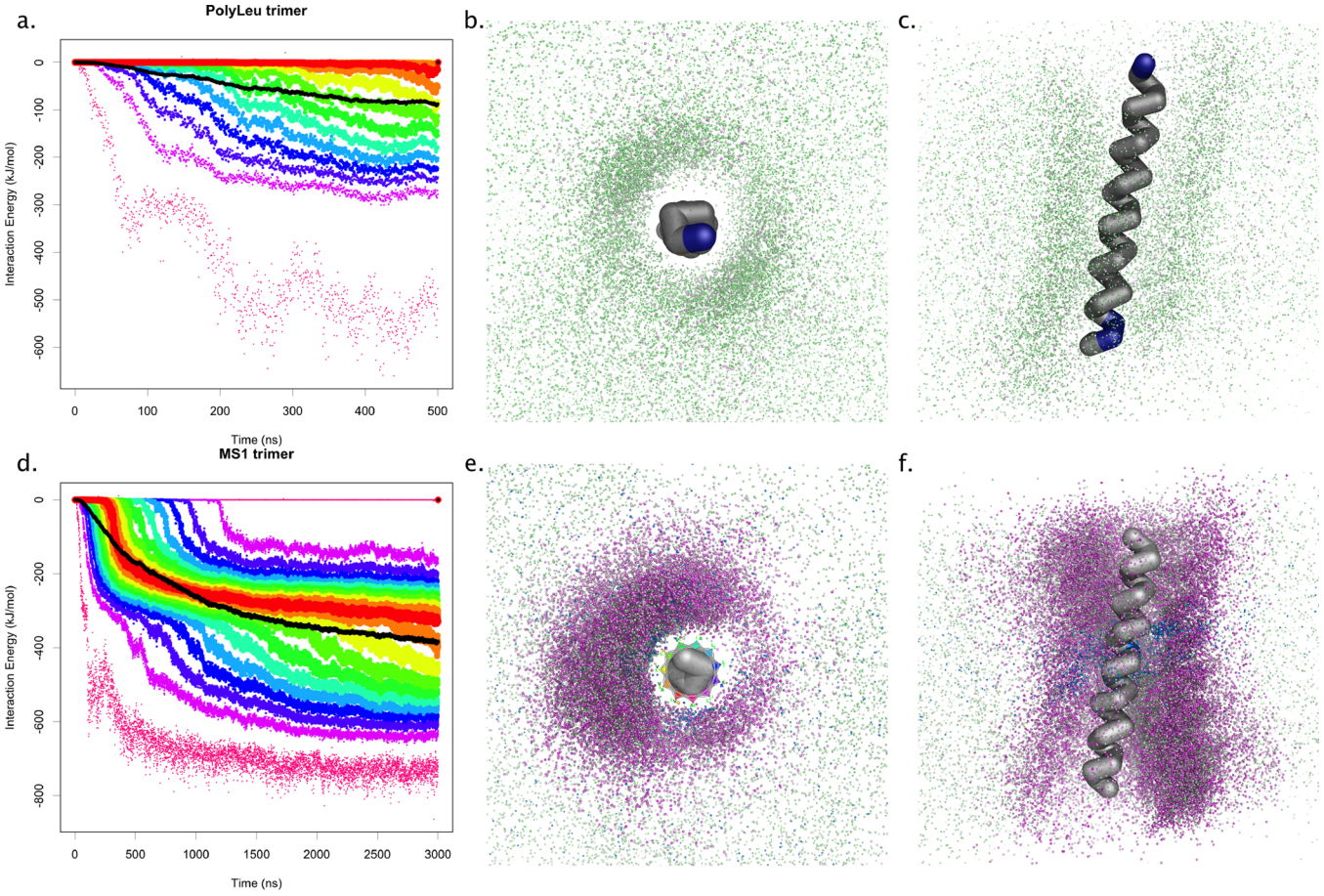


Figure 7. Convergence of interaction energy and resulting particle distributions for polyleucine and MS1 trimers. (a) Time evolution of trimer interaction energy of polyleucine. Setup and coloring of the graph is the same as in Figure 4. (b) Top view of final particle distribution of polyleucine partners around reference chain, after least-squares fit onto the latter, showing low densities corresponding to nonspecific binding. Particle sizes and colors reflect the interaction energy of the trimer, using a color range from purple (lowest) to green (highest) energy. The same color scheme is used for MS1. (c) Side view. (d) Time evolution of trimer interaction energy of MS1, showing a characteristic two-stage decrease and a resulting bimodal distribution, corresponding to the formation of dimers and trimers. (e) Top view of particle distribution of MS1 partners around reference chain after least-squares fit on the latter, showing binding mainly on one side. The purple coloring indicates that the trimers have much lower interaction energies than PL3. Blue dots are used to mark the asparagine residue. (f) Side view.

The interaction energy distribution of MS1 clearly shows two processes, with associated energy levels, corresponding to the formation of dimers (approximately −250 kJ/mol) and the formation of trimers (approximately −500 kJ/mol). The first MS1 dimer is formed after 50 ns, followed by trimerization at 100 ns. Yet it takes around 1 μs for the minimum energy to converge. After 300 ns dimers have formed in 50% of the simulations, indicated by the first drop of the median, and after 1.5 μs in 25% of simulations a trimer has formed. At the end of the simulations, trimers have formed in 45%, while in several simulations there is no association. The rest consist of dimer/monomer combinations.

The contrast with polyleucine is again sharp. PL3 shows some dimer formation and a few trimers. The latter can be explained from the probability of two chains without specific interactions to associate. The profile of the interaction energies resembles that observed for dimers shown in Figure 4, except that dimers are formed in more than 50% of the simulations, which is attributed to the higher probability of encounter due to the higher concentration. Together with the MS1 profiles, these results suggest that DAFT is also suitable for studying trimerization and can differentiate between probabilistic formation of assemblies and specific aggregation. It is evident, though, that, for specific interactions, the complexity is increased significantly, requiring much longer simulations times.

When looking at the orientation analysis in Figure 8, the most notable feature is a symmetric binding mode, labeled A, which corresponds to the dimer, shown in Figure 8b. Next to this region is another peak, labeled B1, which is shown in helical wheel representation in Figure 8c. The 270° phase suggests that there is an equivalent binding mode, B2, which is found to form a shoulder in the position/phase map. Together, these appear to suggest a possible tetrameric arrangement. However, we realized that the helices are tilted, such that the connecting triangle has positive curvature and hence should have angles larger than 60°, as shown in Figure 8. This means that the orientations B1/B2 indeed correspond to the trimer arrangement and that the density map is consistent with coexistence of dimers and trimers, in agreement with experimental observations.[44] Again, we stress that this is a qualitative characterization. It is, at present, not possible to translate the observed differences in populations into free energies, because unbinding events and shifts between binding sites are still undersampled.

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

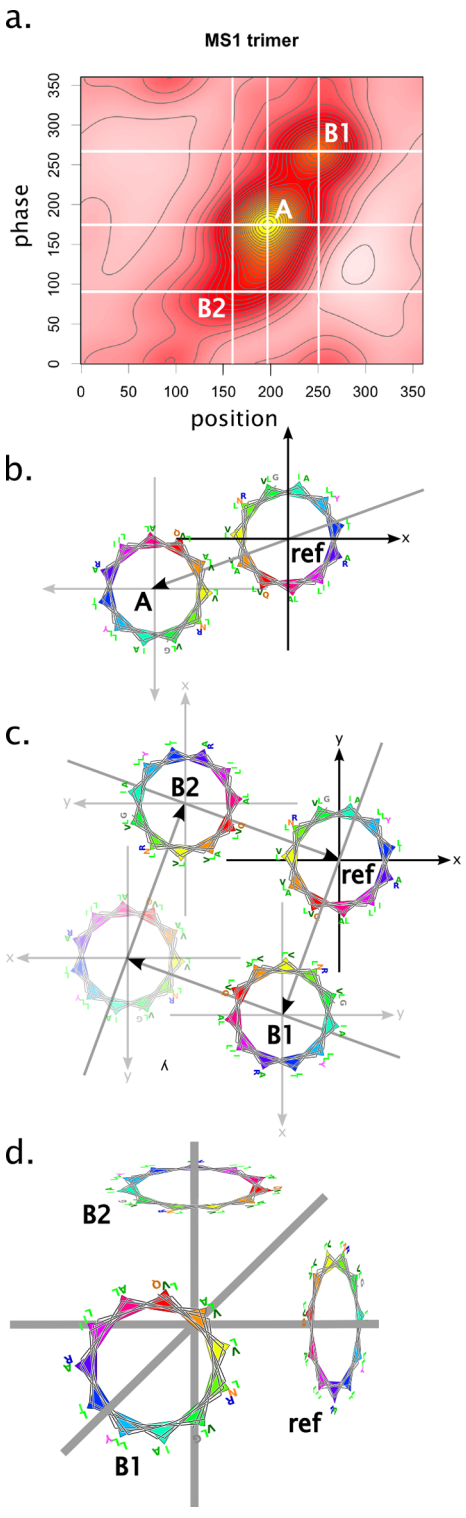


Figure 8. Orientation analysis of MS1 trimers. (a) Position/phase density map, showing the distribution of relative orientations. The most significant orientations are marked with white lines and labeled A, B1, and B2. (b) Helical wheel representation of the relative orientation corresponding to site A, showing a symmetric dimer. (c) Helical wheel representation of the relative orientation corresponding to sites B1 and B2, which are identical up to the interchange of the reference and the partner. These orientations have a phase of −90°, corresponding to a square arrangement. (d) Helical wheel representation of relative orientations corresponding to B1 and B2, with a 90° tilt angle, showing that these orientations correspond to a tilted, trimeric arrangement.

Taken together, these results demonstrate that DAFT also allows assessment of trimeric assemblies of TM helices, albeit that the increased complexity requires significantly longer simulation times. It would be interesting to explore relaxation properties of higher order oligomers, but it seems likely that the simulation times again increase significantly. Possibly, a more efficient route is to make inferences about higher order complexes from dimer and trimer studies so as to circumvent the complexity issues associated with a direct DAFT approach.

4.3. Rhodopsin Dimers. The components used in the foregoing sections are simple transmembrane helices. These are expected to have simple interaction profiles and to diffuse relatively fast, due to the size and because there is no intercalation of chains and no conformational change. To see how protein size and interaction complexity influence the assays, we also studied the dimerization of rhodopsin with DAFT. Previously, Periole and co-workers demonstrated the *in silico* dimerization of this 7TM protein,[30] followed up by a larger study[18] using 16 and 64 copies of the protein embedded in large membrane patches. Although the formation of characteristic dimers was observed, PMF calculations of selected dimers revealed discrepancies with the association simulations, exemplifying that the sampling was incomplete. The authors mentioned that the proteins stuck in initial associated states, being insufficiently able to unbind and reshuffle on the time scales used. In addition, clustering of proteins may give rise to confounding effects, as the formation of an interface affects the probability of formation of alternative dimers. The focus of DAFT on dimer association avoids such confounding effects and should allow a better view on the association propensities for different interfaces. Therefore, an attempt was made to assess the dimerization using the new protocol. To this end, 500 simulations of 1 μs were run, the results of which are summarized in Figure 9.

The top panel Figure 9a shows the time evolution of the interaction energy distribution. The profile shows a clear contrast with the results from GpA. Foremost, the association takes longer, which can be explained by the larger size and slower diffusion. Not even the minimum of the distribution has converged after 1 μs, and the mixing/decay model cannot be used on the results. This means that it is impossible to estimate the minimal and mean interaction energy plateau value using this data set. The linear decrease of the minimum over the last 800 ns suggests that the proteins involved optimize their packing after the association. This view is supported by the observation of a slow process of "buried surface area maximization" reported earlier.[30]

The results clearly show that 500 times 1 μs is not sufficient to get a converged view of association for this protein. The slower diffusion plays an important role therein. However, the progression of the lowest vigintiles also suggests that there are slow processes in play. These may involve structural rearrangements, such as side-chain packing, chain intercalation, and delipidation.

Although the relaxation is slow and incomplete, the orientation analysis shows only a small number of clear hot spots in the position/phase map (Figure 9b). In addition, comparison of the position-phase map from the end of the simulations with the profiles at 0.8 and 0.9 μs (data not shown) shows that these regions are increasing in density, at the expense of the surroundings. This suggests that the corresponding dimers, shown in Figure 9c, are the principal relative orientations, although the relative importance of these

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

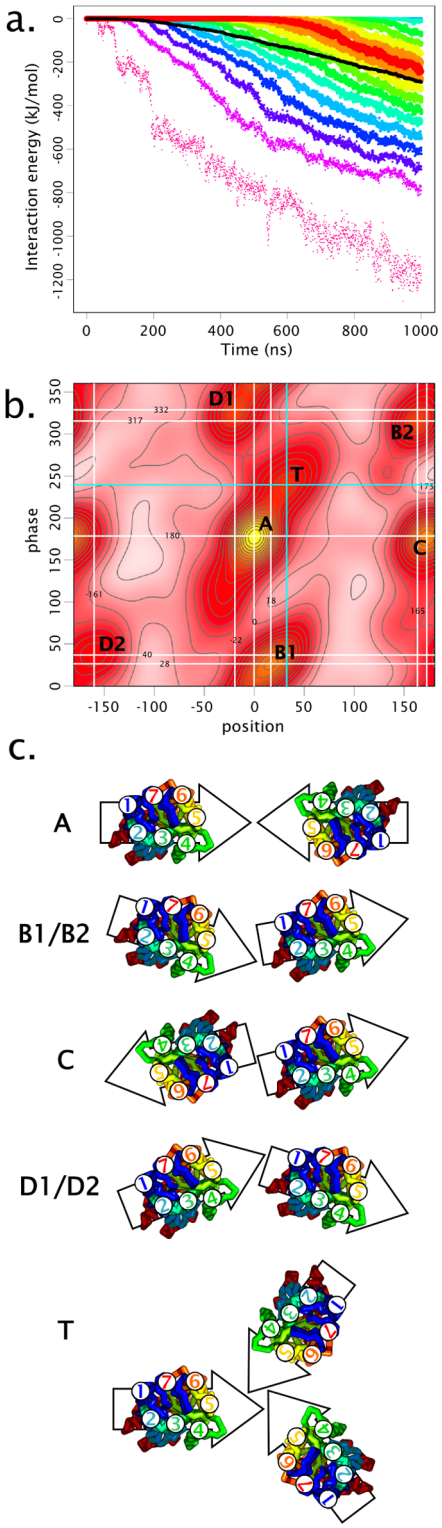


Figure 9. Summary of rhodopsin results. (a) Interaction energy distribution over time, characterized by vigintiles (5% points) and mean interaction energy (black). The vigintiles are colored from purple (minimum) to red (median) to purple (maximum). (b) Position/phase density map of relative orientations. Relative densities are calculated using a circular 2D kernel density estimate, colored from white (no density) to red to yellow (highest density). The most abundant dimers orientations are marked and indicated with white lines, while the blue lines indicate the orientation of a putative trimer. (c) Schematic representations of dimers and putative trimer corresponding to the orientations labeled.

orientations cannot be estimated. It is even possible that one of the dimer species will act as a sink and over longer time scales will cause depletion of the other regions.

Interestingly, despite the lack of convergence, the orientations observed correspond to the ones reported earlier by Periole et al.[18] Dimer A comprises both the TM5/TM5 interface and the TM4+TM5/TM4+TM5 interface, which cannot be distinguished in the position-phase map due to the KDE smoothing. Together, these configurations accounted for 11.3% of the dimers observed by Periole et al. and were the second and third strongest, respectively, in their PMF calculations.

According to the PMF calculations, the TM1/TM1 dimer, denoted C in Figure 9, was the most favorable symmetric configuration, although it only accounted for 6.9% of dimers in the large-scale self-association. Other symmetric dimers were found to be insignificant.

For the asymmetric dimers no PMF calculations were performed, but statistics from the brute force association simulations are available. The TM5/TM1+TM2 dimer we designate B1/B2 was thus found to account for 5.9% of the total dimers in the previous study, while the TM1+TM2/TM4+TM5 dimer, denoted D1/D2, formed 11.5% of the total. The latter dimer was split over two clusters constituting 9.1% (TM4+TM5/TM1+TM2) and 2.4% (TM1+TM2/TM4+TM5), respectively, where the asymmetry exemplifies the lack of convergence suggested by the authors.

An interesting feature in the position/phase map in Figure 9 is a shoulder in the distribution around A, which we designated T and marked with blue lines. This resembles the profile observed in MS1 for the trimer. Inspection of the relative orientation in Figure 9 confirms that the pairing of helices could be consistent with the formation of a trimer. The same dimer configuration was also observed by Periole et al. corresponding to two clusters accounting for 7.6% and 3.0% of the total population of their dimers.

Taken together, these results suggest that the DAFT protocol may provide clues about the interfaces that can be formed between larger proteins. However, such results will be preliminary and should be considered with care, probably within a broader framework of experimental and/or computational study. Yet the same holds for other brute force MD approaches, including large-scale self-assembly CGMD and PMF calculations.

To put this in perspective, it is worth comparing the computational resources used for the previous studies and for DAFT. To allow a direct comparison, the use of computational resources is expressed as the equivalent of seconds of simulation of a single particle with interactions, or "particle seconds", obtained by multiplying the simulation time with the number of particles simulated. Then the brute force calculations performed previously by Periole et al. represent 9.7 particle seconds, the additional PMF calculations represent 6.3 particle seconds, and the DAFT assay on rhodopsin presented here corresponds to a total of 4 particle seconds. In addition, the use of many short simulations is "embarrassingly parallel" by definition, and results can be harvested efficiently on a cluster or with Grid or cloud computing.

5. CONCLUSION

In the previous sections we have described a new approach for characterization of distributions of transmembrane proteins. Characteristic features of this approach include the use of

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

specific layouts exploiting periodic boundary conditions, and coupling to the versatile membrane generating program *insane*. This allows setting up of binding studies for dimers and trimers, as well as for higher order assemblies. The protocol integrates coarse grain molecular dynamics simulations with a docking-style approach, running a large number of simulations from unbiased starting configurations. The program is designed to be user-friendly, yet versatile, and allows setting up and running multiple assays using any combination of transmembrane helix sequences, atomistic models, or coarse grain models with a topology.

The main aim of DAFT is characterizing association and investigating the relative orientations and the interfaces formed. The use of many simulations avoids the need for unbinding and rebinding events to sample interfaces. Of course, this means that no or little information is obtained about binding life times, and it currently remains impossible to estimate the k_{off} and the association constant K_A from brute force simulations in this form or another.

The results in this study demonstrate how DAFT can be used to characterize helix−helix association and discriminate between binding (GpA) and nonbinding (polyleucine) peptides. In addition, the effect of mutations on the formation of dimers can be assessed as illustrated by the results on GpA-WT and GpA-G83I. The relaxation of the ensembles toward bound states could be modeled with a combined mixing/relaxation model, which suggested estimability of the mean interaction energy of the bound state from simulations with a minimum duration of 300 ns. A similar view emerged from inspection of interaction maps. Together these results strongly argue against the use of simulations shorter than 300 ns. Probably it is advisible to aim for more complete relaxation, suggesting that simulation times of at least 500 ns are warranted.

The number of simulations has a pronounced effect on the resolution, as was shown by bootstrapping the orientation profiles. 400 simulations per assay appear sufficient to obtain a high-resolution view of the association orientations. Extreme caution should be taken when making inferences from less than 100 simulations, as there is a significant risk of drawing wrong conclusions, even to the point of drawing opposite conclusions.

DAFT also appears suitable to explore the formation of trimers of simple peptides, although the increased complexity requires considerably longer simulations. The same is true for larger components, such as rhodopsin, where the diffusion and reorientation are much slower, and the more complex interactions cause the evolution of the ensemble to proceed much slower.

The results from DAFT also raise new questions. Among the most important of these is how the ensemble relates to the Boltzmann ensemble and if it is possible to make quantitative inferences about the latter. We previously showed that it is possible to reproduce theoretically calculated PMFs for polyalanine, using DAFT.[54] Yet the specific aim in this regard is to relate DAFT results to experimentally known free energy differences.

Because of the simulation times required, the assays are run using a coarse grain force field. We are currently working on integrating DAFT with our recently developed method for resolution transformation *backward*,[55] to be able to investigate the interactions on the level of atoms and to calculate, e.g., nuclear Overhauser effects (NOEs) to compare the ensembles with NMR measurements. In addition, further studies are planned focusing on higher order oligomers.

DAFT and its auxiliary programs have been made available through http://cgmartini.nl/ and http://www.biotechnik.nat.uni-erlangen.de/research/boeckmann/downloads/DAFT. There, examples for usage of the programs have also been made available.

AUTHOR INFORMATION

Corresponding Author

*E-mail: tsjerkw@gmail.com.

Author Contributions

⊥D.P.T. and R.A.B. contributed equally to this work.

Funding

We acknowledge support by the emerging field initiative "Synthetic Biology" of the Friedrich-Alexander University of Erlangen-Nürnberg (T.A.W. and R.A.B.) and by the Research Training Group Dynamic Interactions at Biological Membranes—"from Single Molecules to Tissue" (RTG 1962; K.P. and R.A.B.) funded by the German Science Foundation (DFG). This work was also supported in part by the Canadian Institutes for Health Research (D.P.T.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

D.P.T. is an Alberta Innovates Health Solutions Scientist and Alberta Innovates Technology Futures Strategic Chair in (Bio) Molecular Simulation.

REFERENCES

(1) Phillips, R.; Ursell, T.; Wiggins, P.; Sens, P. *Nature* **2009**, *459*, 379−385.

(2) Sprong, H.; van der Sluijs, P.; van Meer, G. *Nat. Rev. Mol. Cell Biol.* **2001**, *2*, 504−513.

(3) Marsh, D. *Biochim. Biophys. Acta, Biomembr.* **2008**, *1778*, 1545−1575. (Protein Modulation of Membrane Structure)

(4) Janin, J. *Protein Sci.* **2014**, *23*, 1813−1817.

(5) Karaca, E.; Bonvin, A. M. *Methods* **2013**, *59*, 372−381. (Biophysical Methods for the Study of Protein Interactions).

(6) De Vries, S. J.; van Dijk, A. D. J.; Krzeminski, M.; van Dijk, M.; Thureau, A.; Hsu, V.; Wassenaar, T. A.; Bonvin, A. M. J. J. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 726−733.

(7) Polyansky, A. A.; Volynsky, P. E.; Efremov, R. G. *J. Am. Chem. Soc.* **2012**, *134*, 14390−14400.

(8) Castillo, N.; Monticelli, L.; Barnoud, J.; Tieleman, D. P. *Chem. Phys. Lipids* **2013**, *169*, 95−105. (Computational approaches to understanding lipid−protein interactions).

(9) Schäfer, L. V.; de Jong, D. H.; Holt, A.; Rzepiela, A. J.; de Vries, A. H.; Poolman, B.; Killian, J. A.; Marrink, S. J. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1343−1348.

(10) Cuthbertson, J. M.; Bond, P. J.; Sansom, M. S. P. *Biochemistry* **2006**, *45*, 14298−14310.

(11) Anbazhagan, V.; Schneider, D. *Biochim. Biophys. Acta, Biomembr.* **2010**, *1798*, 1899−1907.

(12) Petrache, H. I.; Grossfield, A.; MacKenzie, K. R.; Engelman, D. M.; Woolf, T. B. *J. Mol. Biol.* **2000**, *302*, 727−746.

(13) Hénin, J.; Pohorille, A.; Chipot, C. *J. Am. Chem. Soc.* **2005**, *127*, 8478−8484.

(14) Siu, S. W.; Böckmann, R. A. *J. Phys. Chem. B* **2009**, *113*, 3195−3202.

(15) Cheng, X.; Im, W. *Biophys. J.* **2012**, *102*, L27−L29.

(16) Baaden, M.; Marrink, S. J. *Curr. Opin. Struct. Biol.* **2013**, *23*, 878−886. (Catalysis and regulation/protein−protein interactions).

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(17) Schneider, A. R.; Geissler, P. L. *Front. Plant Sci.* **2013**, *4*, No. 555.
(18) Periole, X.; Knepp, A. M.; Sakmar, T. P.; Marrink, S. J.; Huber, T. *J. Am. Chem. Soc.* **2012**, *134*, 10959−10965.
(19) Sengupta, D.; Marrink, S. J. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12987−12996.
(20) Psachoulia, E.; Marshall, D. P.; Sansom, M. S. P. *Acc. Chem. Res.* **2010**, *43*, 388−396.
(21) Zuckerman, D. M. *Annu. Rev. Biophys.* **2011**, *40*, 41−62.
(22) Arkhipov, A.; Shan, Y.; Das, R.; Endres, N. F.; Eastwood, M. P.; Wemmer, D. E.; Kuriyan, J.; Shaw, D. E. *Cell* **2013**, *152*, 557−569.
(23) Wassenaar, T. A.; Ingolfsson, H. I.; Böckmann, R. A.; Tieleman, D. P.; Marrink, S. J. Submitted for publication, 2015.
(24) Arnarez, C.; Marrink, S. J.; Periole, X. *Sci. Rep.* **2013**, *3*, 1263.
(25) Hall, B. A.; Halim, K. B. A.; Buyan, A.; Emmanouil, B.; Sansom, M. S. P. *J. Chem. Theory Comput.* **2014**, *10*, 2165−2175.
(26) Wassenaar, T. A.; van Dijk, M.; Loureiro-Ferreira, N.; van der Schot, G.; de Vries, S.; Schmitz, C.; van der Zwan, J.; Boelens, R.; Giachetti, A.; Ferella, L.; Rosato, A.; Bertini, I.; Herrmann, T.; Jonker, H.; Bagaria, A.; Jaravine, V.; Güntert, P.; Schwalbe, H.; Vranken, W.; Doreleijers, J.; Vriend, G.; Vuister, G.; Franke, D.; Kikhney, A.; Svergun, D.; Fogh, R.; Ionides, J.; Laue, E.; Spronk, C.; Jurkša, S.; Verlato, M.; Badoer, S.; Dal Pra, S.; Mazzucato, M.; Frizziero, E.; Bonvin, A. M. J. J. *J. Grid Comput.* **2012**, *10*, 743−767.
(27) de Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schäfer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2013**, *9*, 687−697.
(28) Wassenaar, T. A.; Ingólfsson, H. I.; Prieß, M.; Marrink, S. J.; Schäfer, L. V. *J. Phys. Chem. B* **2013**, *117*, 3516−3530.
(29) Ingólfsson, H. I.; Melo, M. N.; van Eerden, F. J.; Arnarez, C.; Lopez, C. A.; Wassenaar, T. A.; Periole, X.; de Vries, A. H.; Tieleman, D. P.; Marrink, S. J. *J. Am. Chem. Soc.* **2014**, *136*, 14554−14559.
(30) Periole, X.; Huber, T.; Marrink, S.-J.; Sakmar, T. P. *J. Am. Chem. Soc.* **2007**, *129*, 10126−10132.
(31) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812−7824.
(32) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750−760.
(33) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.
(34) Wassenaar, T. A.; Daura, X.; Padrós, E.; Mark, A. E. *Proteins: Struct., Funct., Bioinf.* **2009**, *74*, 669−681.
(35) Pluhackova, K.; Wassenaar, T. A.; Böckmann, R. A. Methods in Molecular Biology . In *Membrane Biogenesis*, Vol. *1033*; Rapaport, D., Herrmann, J. M., Eds.; Humana Press: New York, 2013; pp 85−101.
(36) Altschul, S.; Gish, W.; Miller, W.; Meyers, E.; Lipman, D. *J. Mol. Biol.* **1990**, *215*, 403−410.
(37) *The PyMOL Molecular Graphics System*, Version 1.7.4.0 Incentive; Schrödinger: New York, 2014.
(38) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, No. 014101.
(39) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(40) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(41) Russ, W. P.; Engelman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 863−868.
(42) Conn, P. M. Optical and Spectroscopic Techniques. In *Methods in Enzymology*; Conn, P. M., Ed.; 2012; Vol. *504*, Chapter 18, p 368.
(43) Zhou, F. X.; Merianos, H. J.; Brunger, A. T.; Engelman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 2250−2255.
(44) Choma, C.; Gratkowski, H.; Lear, J. D.; DeGrado, W. F. *Nat. Struct. Biol.* **2000**, *7*, 161−166.
(45) Periole, X.; Cavalli, M.; Marrink, S.-J.; Ceruso, M. A. *J. Chem. Theory Comput.* **2009**, *5*, 2531−2543.
(46) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008; ISBN 3-900051-07-0.
(47) Plummer, M.; Best, N.; Cowles, K.; Vines, K. *R News* **2006**, *6*, 7−11.
(48) Vert, J.-P.; Tsuda, K.; Schölkopf, B. *Kernel Methods in Computational Biology*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2004; pp 35−70.
(49) Eckart, C. *Phys. Rev.* **1934**, *46*, 383.
(50) Ripley, B. D.; Venables, W. N. *Modern applied statistics with S-Plus*; Springer-Verlag: New York, NY, USA, 1994.
(51) Efron, B.; Tibshirani, R. J. *An introduction to the bootstrap*; CRC Press: Boca Raton, FL, USA, 1994; Vol. *57*.
(52) Psachoulia, E.; Fowler, P. W.; Bond, P. J.; Sansom, M. S. P. *Biochemistry* **2008**, *47*, 10503−10512.
(53) Wassenaar, T. A. *Molecular Dynamics of Sense and Sensibility in Processing and Analysis of Data*, 1st ed.; University of Groningen: Groningen, The Netherlands, 2006.
(54) Pawar, A. B.; Deshpande, S. A.; Gopal, S. M.; Wassenaar, T. A.; Athale, C. A.; Sengupta, D. *Phys. Chem. Chem. Phys.* **2015**, 1390−1398.
(55) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. *J. Chem. Theory Comput.* **2014**, *10*, 676−690.

DOI: 10.1021/ct5010092
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX