JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

# Coarse-Grained Simulations of Membrane Insertion and Folding of Small Helical Proteins Using the CABS Model

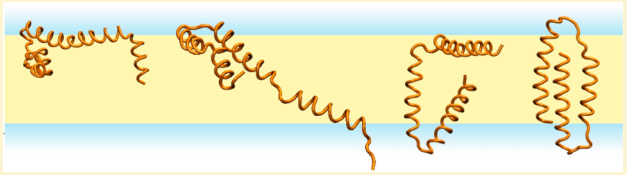
Wojciech Pulawski,[†] Michal Jamroz,[†] Michal Kolinski,[‡] Andrzej Kolinski,*[,†] and Sebastian Kmiecik*[,†]

[†]Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

[‡]Bioinformatics Laboratory, Mossakowski Medical Research Center of the Polish Academy of Sciences, Pawinskiego 5, 02-106 Warsaw, Poland

Supporting Information

**ABSTRACT:** The CABS coarse-grained model is a well-established tool for modeling globular proteins (predicting their structure, dynamics, and interactions). Here we introduce an extension of the CABS representation and force field (CABS-membrane) to the modeling of the effect of the biological membrane environment on the structure of membrane proteins. We validate the CABS-membrane model in folding simulations of 10 short helical membrane proteins not using any knowledge about their structure. The simulations start from random protein conformations placed outside the membrane environment and allow for full flexibility of the modeled proteins during their spontaneous insertion into the membrane. In the resulting trajectories, we have found models close to the experimental membrane structures. We also attempted to select the correctly folded models using simple filtering followed by structural clustering combined with reconstruction to the all-atom representation and all-atom scoring. The CABS-membrane model is a promising approach for further development toward modeling of large protein−membrane systems.



## INTRODUCTION

Membrane proteins (MPs) are important components of molecular machineries of life, and they are estimated to be encoded by up to 30% of the human genome.[1] MPs are essential for a number of key biological functions, such as signal transduction, transport of solutes, and cell communication. Many MPs are involved in human diseases; they are targets for over 50% marketed drugs[2,3] and also for numerous ongoing drug discovery projects. Despite significant efforts from both academia and industry, MP structures are vastly under-represented in the Protein Data Bank (MPs constitute only 1−2% of known structures).[4] This is because conventional structure determination techniques, such as X-ray and NMR, have serious application limitations for MPs, including protein sample preparation and other issues.[5] The poor structure understanding is paired with very limited comprehension of their structural dynamics and thus the functions they perform on the molecular level.

Since experimental structure determination of MPs is a significant challenge, various computational techniques have been developed to predict structural features of MPs; a comprehensive review of them is provided in ref 4. The other, and not less important, challenge is to simulate MP dynamics, and the most acknowledged approach is all-atom molecular dynamics (MD). Unfortunately, application of all-atom MD is limited to short simulation time scales. In practice, unguided all-atom MD simulations are restricted to small conformational changes or small protein systems.[6−8] A substantial computational speed-up may be achieved using coarse-grained modeling,[9] and a wide range of coarse-grained models have been proposed for simulations of MPs in the membrane environment.[9−18] The most popular one is definitely the MARTINI model[14] (originally developed for lipids[19,20] and subsequently extended to model proteins[14]). Although MARTINI has found a broad range of applications, it still has a number of limitations.[21] An important one with regard to modeling of protein conformational changes is that secondary structure formation or disruption cannot be modeled[21] (user-defined secondary structure has to be maintained during the simulation using secondary structure constraints). This is a result of the specifics of the MARTINI coarse-grained representation, which has rather sophisticated side-chain representation but a very simple backbone. For these reasons, implementation of different models of membrane proteins that would enable changes of the entire structure could be important for theoretical studies of protein folding induced by insertion into cell membranes.

In this work, we describe a simple modification of the CABS coarse-grained force field by a simple model of the biological membrane. To date, the CABS model has been extensively used in various modeling tasks of globular proteins and outer membrane fragments of membrane proteins (see the CABS description in Materials and Methods). As presented in recent dynamics studies of globular proteins, the CABS model allows us to model protein conformational changes (including secondary structure) during folding[22−27] or folding and binding of a disordered protein.[28] We tested the extended CABS

**Received:** June 14, 2016
**Published:** October 24, 2016





DOI: 10.1021/acs.jcim.6b00350
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

version in simulations of short $\alpha$-helical MPs (it is estimated that about 80% of MPs are $\alpha$-helical bundles[1]). These tests showed that after some modification of hydrophobic interactions between side chains, the model can be effectively used for simulations of conformational transitions during protein insertion into membranes.

## MATERIALS AND METHODS

**The CABS Model.** CABS is a high-resolution coarse-grained model of protein structure and dynamics. A detailed description of the model, including its representation, force field, and sampling scheme, was provided in 2004,[29] while its main ideas and several applications have been summarized and compared with other coarse-grained approaches in a recent review.[9] Here we provide an overview of the model necessary for the definition of its extension to coarse-grained MPs. In the CABS model, amino acid residues are represented by up to four pseudoatoms: $C_\alpha$ (**CA**), $C_\beta$ (**B**), the center of mass of the remaining portion of a side chain (**S**), and the center of the virtual $C_\alpha$–$C_\alpha$ bond (Figure 1a,b). Available positions of the $C_\alpha$ united atoms are restricted to selected points of an underlying cubic lattice. Therefore, small fluctuations of the distance between successive $C_\alpha$ atoms are allowed. The locations of the side-chain united atoms depend on the main-chain geometry and are restricted to the positions representing the most populated conformations observed in known protein structures; the locations of the side chains are discretized and stored in large tables. The resulting resolution level of the CABS representation is typical of other efficient coarse-grained protein modeling methods, including TOUCHSTONE,[30] UNRES,[31] and I-TASSER.[32] Such coarse-grained models seem to be a reasonable compromise between speed-up of the simulations and fast realistic reconstruction of the atom-level representation of the modeled structures.

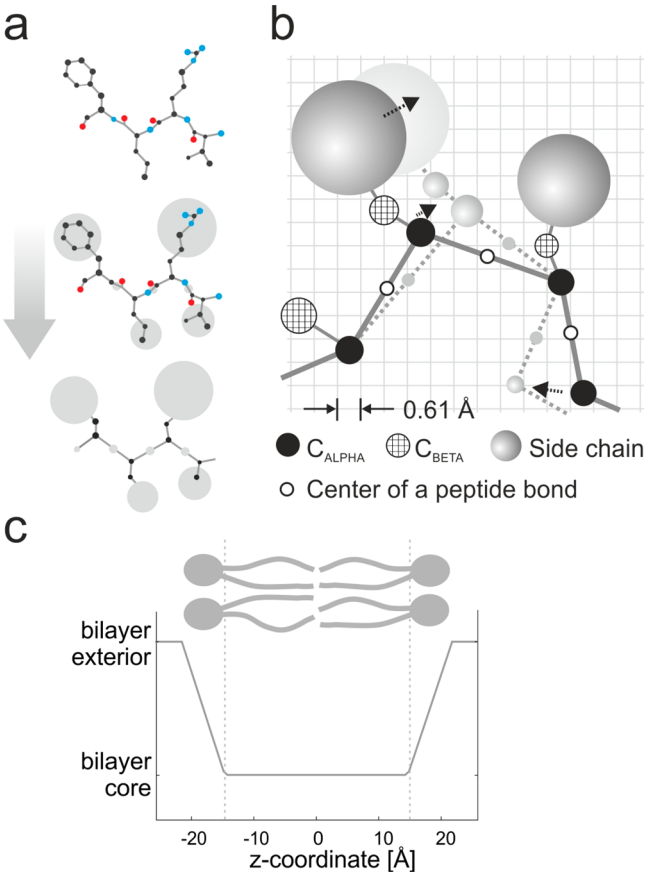



**Figure 1.** Overview of the CABS-membrane model. (a) Comparison of the CABS coarse-grained representation with its all-atom counterpart. (b) CABS representation details and an example move. (c) CABS energy term for protein residue transfer into the membrane based on the hydrophobicity of a residue ($E_{KD}$) as a function of the distance from the bilayer midplane ($z$).

In the CABS model, the interactions are described by knowledge-based statistical potentials derived from structural regularities seen in the known structures of globular proteins. These include the following: a set of conformational propensities of the main chain mimicking the polypeptide flexibility and secondary structure propensities; the excluded volume of the united atoms taken from the averaged statistics for the protein database; a model of main-chain hydrogen bonds modeled as strong directional contact interactions with weak secondary structure (statistical predictions for specific sequences) biases; and amino acid-dependent interactions between side chains. Interactions between the side chains are the most important sequence-dependent forces in the CABS model. These pairwise contact potentials are context-dependent and take into consideration the identity, size, and mutual orientation of the interacting pseudoatoms (the same concept is used in related modeling schemes,[30,32] and a detailed description of the CABS force field can be found elsewhere[29,33]). It should be pointed out that the knowledge-based statistical force field of the CABS protein model was derived for globular proteins and that the solvent is treated in strictly implicit fashion, assuming the averaged folding temperature as a reference state. In the CABS model, temperature is the parameter that controls the acceptance ratio for new conformations using the Monte Carlo (MC) method. Consequently, the simulation temperature is a dimensionless parameter that is proportional to the real temperature increments. The parameters of the CABS force field are available on our Web site (http://biocomp.chem.uw.edu.pl/tools).

The conformational space of the CABS model is sampled by a properly designed Monte Carlo scheme. A long series of small local conformational transitions (within the asymmetric Metropolis scheme) mimics the long-time dynamics of the modeled proteins. The sampling process is computationally very efficient, not only because of the coarse-grained representation of protein chains but also thanks to discrete (high-coordination lattice) positions of the $C_\alpha$ pseudoatoms. As $C_\alpha$ chains provide reference coordinates for the other pseudoatoms, the underlying lattice enables efficient storage of local moves and interactions. In this way, CABS Monte Carlo dynamics simulations are about an order of magnitude faster than otherwise equivalent continuous-space models.

The CABS model has proven to be an efficient tool for protein structure prediction, including de novo and comparative modeling schemes,[34,35] studies of protein folding mechanisms,[22−27] protein flexibility,[36−38] and modeling of protein−peptide interactions: binding mechanisms[28] and flexible protein−peptide docking.[39,40] CABS models have also been recently used in modeling nonmembrane fragments of MPs to predict the structure of extracellular loops of GPCRs[35] and the structure and functional motions of the outer membrane transporter FecA.[41] In these studies, similar to MARTINI membrane protein modeling schemes,[42] the


DOI: 10.1021/acs.jcim.6b00350
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

behavior of transmembrane fragments was controlled by distance restraints derived from known three-dimensional structures. In this paper, we present an extension of the CABS modeling scheme to possible de novo simulation of the entire membrane insertion process without any knowledge about the modeled in-membrane structures.

**Extension of CABS to Membrane Proteins.** The CABS model is extended to the simulation of membrane proteins by the addition of a new force field term mimicking the membrane environment, defined as

$$E_{KD} = -KD_{amino\ acid} \times h$$

where $KD_{amino\ acid}$ is the hydrophobicity value of a given amino acid based on the Kyte−Doolittle scale[43] and $h$ is the scaling factor (Figure 1c). The $E_{KD}$ term applies within the main membrane layer (hydrophobic core), defined on the $z$ axis in the range of −15 Å to 15 Å, which corresponds to the average size of the hydrophobic interior of a biological membrane.[44] Additionally, on both sides of the main layer surface, interface layers are defined. The layers correspond to transition regions with a thickness of 6 Å in which the environment changes from 90% nonpolar to 90% polar.[45] In these regions, $E_{KD}$ is scaled linearly by a factor corresponding to the distance to the main layer. The obtained hydrophobic profile (Figure 1c) is complementary to most previous studies using exponential functions to describe the steepness of the transition.[46,47]

The original CABS force field was derived for globular proteins, and the new force field term $E_{KD}$ provides a correction for intraprotein residue−residue interactions and accounts for the implicit membrane environment. Several simple interaction schemes mimicking the membrane environment have been tested. We found during these tests that a slight increase in attractive interaction between the side chains, as defined by the KD hydrophobicity scale,[43] is efficient to enforce membrane insertion of hydrophobic protein fragments. Other hydrophobicity scales could be considered, but when the derivation of the KD scale is taken into account,[43] it appears to be a reasonable first choice. Hydrophobicity scale corrections to the CABS force field in the membrane environment mimic the absence of implicitly treated polar solvent (water). A perhaps more straightforward way to introduce the membrane environment into CABS would be to derive a within-membrane statistical force field for amino acids similar to the CABS force field for globular proteins. Despite the relatively small number of available high-resolution MP structures, efforts toward the development of statistical potentials for MPs have already shown promising results.[48,49]

**Protein Data Set and Simulation Settings.** To test the de novo modeling capabilities of the CABS-membrane model, we chose 10 short membrane proteins with lengths of 35 to 81 residues. Larger MPs were not considered in the data set because they exceed the possibilities of purely de novo protein structure prediction, as discussed in the recent review.[9] The protein data set comprised the following proteins (PDB IDs and numbers of amino acids are given): 1IIJ (35 aa), 1N7L (53 aa), 1WAZ (46 aa), 1WU0 (72 aa), 2KSD (75 aa), 2LOP (75 aa), 1VRY (76 aa), 1A91 (79 aa), 2K9P (80 aa), and 2MOZ (81 aa). These proteins represent single-pass membrane helical proteins or multiple (two or three)-helix bundles and generally have low sequence identity with each other (see Tables S1 and S2 in the Supporting Information).

The simulation input included the protein sequence and secondary structure assignments (in one-letter code). For the sake of simplicity, we used native assignments since the predicted secondary structure is quite accurate for membrane proteins and the exact size of the predicted helical fragment is not important for the CABS model. This is the case because the structure input in the CABS model provides a weak bias toward the local chain geometry and does not preclude different secondary structure arrangements (the role of secondary structure in the CABS force field has recently been discussed in detail[33]). Interestingly, as demonstrated in protein folding[23,50] and folding and binding[28] studies, the CABS model usually provides correct predictions of the stability of secondary structure elements during the folding process.

For each protein, we performed 20 independent CABS-membrane simulations having different combinations of temperature, $T$, and scaling factor, $h$. The following values were tested: $T$ = 1.3, 1.5, 1.7, 1.9, 2.1 and $h$ = 0.5, 1.0, 1.5, 2.0. Initial simulation analysis showed that extreme values of $T$ and $h$ gave clearly worse results in terms of the number of native-like conformations in simulation trajectories. Therefore, for further analysis we considered only six independent simulations: those with $T$ = 1.5, 1.7, 1.9 and $h$ = 1.0, 1.5. Additionally, for the same temperature set ($T$ = 1.5, 1.7, 1.9), we ran folding simulations without the membrane region ($h$ = 0). Cumulative comparison of the simulations with and without the membrane (Figure 2) showed that the introduced membrane environment provides the driving force for folding of the MPs.

Each simulation started from an extended protein conformation placed outside the membrane environment. During the simulations, the protein conformations were allowed to be fully flexible. Each simulation consisted of 10 million MC CABS microcycles (on average, each microcycle corresponded to a few attempts of protein chain micromodifications per

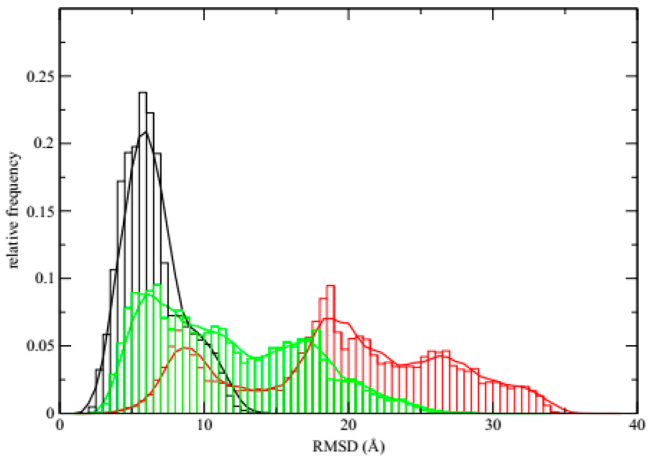



**Figure 2.** Histograms of RMSD values for simulations with and without the membrane. Three kinds of histograms are shown: for the CABS simulations without the membrane (colored in red), for CABS-membrane simulations (colored in green), and for models selected in the initial filtering step from CABS-membrane simulations (colored in black; a description of the initial filtering step is provided in the Results). The histograms show cumulative distributions of the relative frequency of RMSD values for all of the test set proteins, temperatures ($T$ = 1.5, 1.7, 1.9) and scaling factors ($h$ = 1.0, 1.5 for the CABS-membrane simulations and $h$ = 0 for the simulations without the membrane). In the presented statistics, only the second half of each simulation trajectory was taken into account. Individual distributions for each protein from the test set are presented in Figure S1.



DOI: 10.1021/acs.jcim.6b00350
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

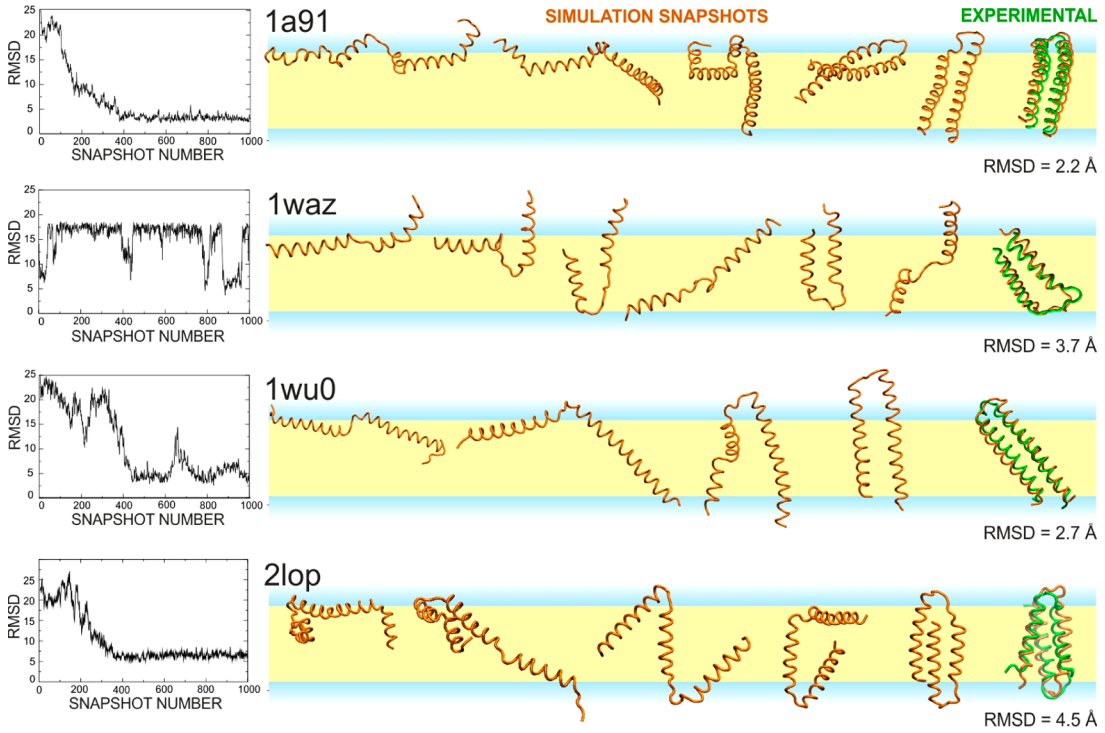



**Figure 3.** Example insertion and folding mechanisms observed in the simulations. For four proteins (PDB IDs 1A91, 1WAZ, 1WU0, and 2LOP), the figure shows the evolution of the RMSD values (reflecting similarity to the experimental structure) during example simulations. Additionally, example models illustrating the insertion and folding mechanisms are presented (colored in orange). As the last ones in the row, the lowest-RMSD models are presented and superimposed with the experimental structures (colored in green). For protein 2LOP, an example trajectory is presented in the movie provided in the Supporting Information.

single protein residue). Simulation trajectories were recorded with an interval of 10 000 MC microcycles, and 1000 snapshots were recorded in each case. A single simulation took about 3 h on a single core of an Intel Xeon 3.5 GHz processor.

**Data Analysis.** Values of the root-mean-square deviation (RMSD) between the experimental structures and predicted models based on $C_\alpha$ coordinates were calculated using the Clusco package.[51] Since experimental structures of the studied proteins are available in the form NMR ensembles of different conformational states, we calculated RMSD values with respect to all of the available states, and the lowest RMSD values are reported. It is important to note that the structural flexibility in the investigated NMR ensembles is high in some cases and likely to be underestimated.[52] Since NMR ensemble models can be characterized by a high level of uncertainty of the atomic coordinates, the reported RMSD values should not be treated as a strict measure of the difference between the experimental structure and the predicted models.

A tilt angle was defined as the angle between the membrane normal ($z$ axis) and a line fitted to the $C_\alpha$ coordinates of a helix (helix positions were predicted using TOPCONS). In the case of multiple-helix proteins, the average value of their tilt angles is reported.

The analysis of native contacts (presented in Figure S4 in the Supporting Information) was performed using $C_\alpha$ coordinates and two contact cutoffs: 6 Å for contacts close in sequence (between residues $i$ and $i + 4$ to $i + 7$) and 8 Å for contacts distant in sequence (between $i$ and $i + 8$ or more distant).

## RESULTS

**Simulations of Protein Insertion and Folding.** The CABS simulations started from completely extended protein conformations placed outside the membrane environment and allowed for full flexibility of the modeled proteins during their spontaneous insertion into the membrane (the simulation settings are described in Materials and Methods).

Visualization of the simulation trajectories allowed us to observe the common general steps of protein insertion and folding. The observed mechanisms toward the near-native structure share the following stages: (1) helix formation, (2) helix insertion into the membrane, (3) helix rearrangements and fluctuations of the secondary structure, and (4) near-native fluctuations. For four protein cases we present examples of simulation snapshots showing the evolution toward the near-native state and the best-accuracy models (Figure 3; additional analysis of the folding and insertion dynamics is provided in Figures S2–S5). These four cases represent two folding behaviors observed in our simulations after the protein insertion. The first one is almost a two-state folding mechanism in which near-native ensembles can be clearly distinguished from other inserted intermediates and the near-native state is relatively stable (proteins 1A91 and 2LOP). The second one represents the cases in which the distinction between intermediate and near-native states is somewhat blurred and the near-native state is unstable (proteins 1WAZ and 1WU0).

In all of the protein cases, most of the helical structure (about 80% or more) is formed before insertion into the membrane. This is consistent with many previous studies showing that secondary structure formation is a necessary step before insertion.[6,44,53] After insertion, the preformed helices usually undergo several conformational changes, including threading the loop from one membrane surface to the other. In more than half of the trajectories the loop stayed at the same side where it was initially formed, while in other examples the loop was


DOI: 10.1021/acs.jcim.6b00350
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

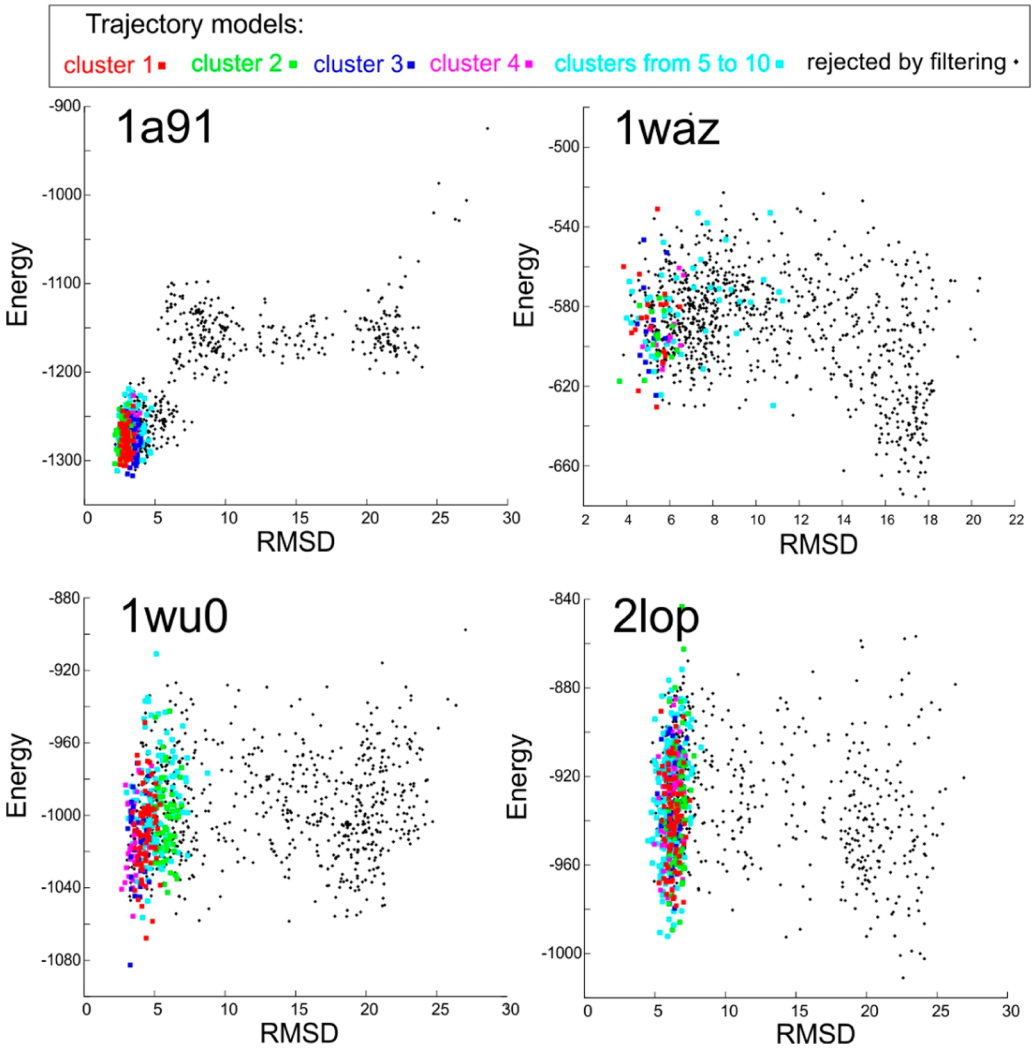



**Figure 4.** Selection of models using structural clustering. The plots present CABS energy vs similarity to the experimental structure (RMSD value) for four example trajectories of four proteins (PDB IDs 1A91, 1WAZ, 1WU0, and 2LOP). Trajectory models are colored according to the clustering procedure as belonging to clusters of the most common conformations (clusters are numbered according to the number of models, with the first one the largest) or rejected by initial filtering (see the legend).

threaded to the other side of the membrane. The conformational changes eventually resulted in a near-native topology. In all cases, the path to near-native ensembles is coupled with decreasing radius of gyration (Figure S2). The accuracy of the lowest-RMSD models was in the range of ∼2−5 Å from the experimental structure, depending on the protein (the exact RMSD value for each protein is given in the next section). Mutual helix rearrangements are usually accompanied by secondary structure fluctuations (in some cases including disruption of helix continuity and kink formation). Finally, fluctuations of near-native folds usually involve relatively small (1A91 and 2LOP) or larger (1WAZ and 1WU0) movements of unstructured fragments (loops or unstructured protein ends) and changes in the tilt angles (Figures S3 and S5). Additional analysis of the tilt angle for the specific test case 1IIJ (the simplest topology from the test set of a peptide forming a single membrane-spanning helix in the native form) showed two-state dynamics (Figure S6). The identified two states are a near-native transmembrane state and an interface state (in which the helix is located on the surface of the membrane), which is consistent with experimental[54] and theoretical[55] findings.

The observed folding and insertion mechanisms are consistent with the widely accepted two-stage model or the three-stage model for the folding of $\alpha$-helical MPs.[56] According to the two-stage model, the membrane insertion of independently stable $\alpha$-helices is followed by mutual helix packing within the membrane. For some proteins, the two-stage model was extended to the three-stage model,[57] in which the third stage may involve folding of loops or binding of prosthetic groups or other proteins (e.g., oligomer assembly).

As shown in the previous studies of the globular protein folding process,[22−27] CABS Monte Carlo dynamics provides a realistic picture of the long-time dynamics of protein systems. This is possible because the MC sampling scheme is based on a long series of very local (fast) randomly selected conformational transitions mimicking the long-time evolution of the protein structure. An analogous MC scheme was applied in this work to MPs to model the process of protein insertion into membranes. Similar simulations of protein insertion into the membrane have been also successfully performed using somewhat simpler coarse-grained models.[58,59] Providing the mechanisms of membrane insertion and structure assembly, including the formation of secondary structure elements,


DOI: 10.1021/acs.jcim.6b00350
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

tertiary structure, and the straightforward possibility of quaternary structure assembly, the CABS model enables modeling of very complex dynamic processes. This possibility distinguishes the presented model from other approaches requiring, for example, the constrained predefinition of protein secondary and higher-order structure.[21] However, there are several limitations of the CABS-membrane model besides the obvious limitations of coarse-grained force fields. In CABS, the membrane structure is highly idealized as a continuous, hydrophobic solvent environment with a smooth transition to the polar solvent environment (as illustrated in Figure 1c). The local moves of protein fragments (and their frequency) used in the MC dynamics scheme are the same within and outside the model membrane. In this way, a possible effect of coupling of local moves of a protein and lipids on the system dynamics is crudely preaveraged. Furthermore, the possible long-time fluctuations of the membrane size and shape are neglected. These limitations may somewhat distort the time scale of the observed dynamics and should be kept in mind when interpreting specific applications of the model.

**Blind Selection of Native-like Models.** Since CABS simulations start from random protein conformations placed outside the implicit membrane, the simulation trajectories consist of a significant number of transient protein conformations in the process of coupled folding and insertion into the membrane. In order to blindly filter out the protein conformations that are plausibly folded and membrane-oriented, we developed and tested a selection procedure based on the following steps: (1) initial filtering, (2) structural clustering, (3) reconstruction to the all-atom representation, and (4) minimization and scoring in the all-atom force field.

The initial filtering is based on TOPCONS,[60] a method for predicting MP topology using the protein sequence only. TOPCONS is used to assign the length and sequence position of helical fragments. This knowledge is used to check whether the predicted TM fragments adopt a proper secondary structure and, if so, to calculate angles between helices and the bilayer. Recognition of $\alpha$-helices is based on simple geometrical criteria based on $C_\alpha$ distances with some tolerance allowing for kinks and bends. Models fulfilling the following criteria were rejected: largely different secondary structure assignment than the one predicted by TOPCONS, having helices with unusual membrane angles (based on TMPad database statistics;[61] a cutoff of 45 deg was used), and poor packing models when two helices are not in contact with each other.

In the second step, structural clustering is performed on all of the initially filtered models using the Clusco package.[51] The K-means clustering method is used, with the maximum number of clusters set to 10. Figure 4 shows example results from the clustering of four trajectories (for the four proteins also presented in Figure 3).

In the third step, all of the models assigned to any of 10 clusters are reconstructed into all-atom structures using ModRefiner.[62]

Finally, in the fourth step, each model is optimized and scored using the implicit membrane model 1 (IMM1) force field.[63] The IMM1 method is a semiempirical implicit-solvent force field with all protein atoms except for nonpolar hydrogens. The IMM1 method allows computation of the energy of a protein with solvation terms; hence, it is very useful to compare structures with different conformations and different orientations in the membrane.[64] Energy evaluation included 2000 steps of minimization with frozen $C_\alpha$ atoms.

The top-ranked models are presented in Figure 5 together with superimposed experimental structures and the best-accuracy (lowest-RMSD) models obtained in CABS simulations. In some cases, the top-ranked models and best-accuracy models are of similar topology and resolution, while in other cases some distinct differences occur. The most common reason for the largest differences is that in the case of top-ranked models longer and straighter helices are energetically preferred over those with kinks (see cases 1IIJ, 1N7L, and 2KSD in Figure 5) and over shorter helices ending with irregular structure fragments (see case 2MOZ in Figure 5). The other differences concern slightly different arrangements of $\alpha$-helices with respect to the membrane in terms of their tilt angles or degree of immersion.

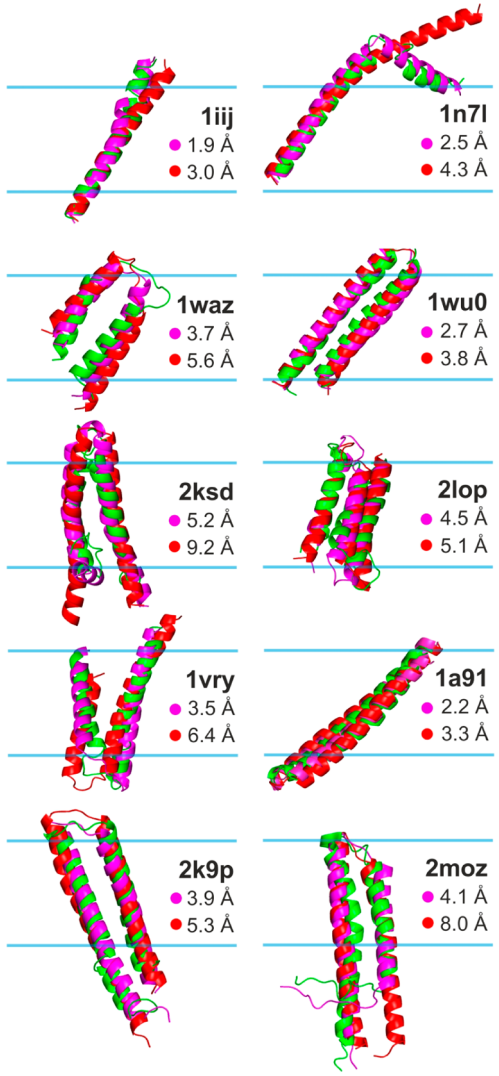



**Figure 5.** Comparison of predicted and experimentally determined models. For each protein, three kinds of structures are presented (after their superimposition): the experimentally determined structure (colored in green), the model having the lowest RMSD value (colored in magenta), and the top-ranked model (colored in red). Additionally, PDB IDs and RMSD values of the presented models are shown.

The presented prediction accuracy of the CABS-membrane method is comparable to the performance of the state-of-the-art


DOI: 10.1021/acs.jcim.6b00350
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

methods for the de novo prediction of helical MPs that have been summarized in a recent review.[4] According to the review, the best performance for small proteins (up to 100 residues) is offered by the Rosetta-membrane method,[44,45] which provides models with RMSD values below 5 Å. However, it has to be noted that the Rosetta-membrane method is based on searching of protein fragments in structural databases and therefore is not a purely de novo approach independent of the availability of protein structure data. Other successful de novo methods for predicting $\alpha$-helical MPs[4] are based on deriving evolutionary constraints from multiple sequence alignments built for the query sequence.[65] Thus, in comparison with the aforementioned state-of-the-art methods, CABS-membrane is a qualitatively different approach that does not require structural fragments of similar proteins or sequence data of homologous proteins. This opens up interesting possibilities for combining the CABS-membrane method with state-of-the-art tools in which, for example, the CABS-membrane approach can be used to enhance prediction of ambiguous protein fragments that do not exist in structural databases or are missing in comparative sequence analysis.

## CONCLUSIONS

In this work, we validated the CABS-membrane model, a simple extension of the CABS model for de novo simulations of short $\alpha$-helical MPs. Using no information about the structure of the modeled proteins or evolutionary data, we performed simulations of protein folding and insertion into a simplified membrane model. The simulations yielded protein models that are close to the experimentally determined structures.

The presented multiscale modeling pipeline based on the combination of CABS-membrane coarse-grained and atomistic simulations can be potentially improved, either in its first coarse-grained stage (e.g., by derivation of the within-membrane statistical force field) or/and in the second all-atom stage (e.g., by merging with tools for reconstruction, final model optimization, and scoring that are more accurate or better-suited to handling CABS coarse-grained models[9,66]).

As presented in the CABS modeling studies of globular proteins (see Materials and Methods), even sparse structural data (taken from experimental studies or predicted from evolutionary analysis) can be effectively utilized in the prediction of protein structure and dynamics (after translation into the form of distance restraints). Analogously, the method for simulations of short $\alpha$-helical MPs presented here can be easily extended to more sophisticated modeling schemes based on structural data from various sources (e.g., from homologous proteins with known structure, residue−residue contact predictions based on evolutionary data, or sparse experimental data), which will enable the modeling of much larger MPs.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00350.

Sequence information on the test proteins in the PIR format (Table S1); percent identity matrix of the sequences of the test proteins created using Clustal Omega (Table S2); histograms of RMSD values for simulations with and without the membrane (Figure S1); number of native-like models found in the simulations for different scaling factors ($h$) (Table S3); radius of gyration versus RMSD (Figure S2); tilt angle versus RMSD (Figure S3); native contacts versus simulation time (Figure S4); tilt angle and position of the center of the mass versus simulation time (Figure S5); and tilt angle analyses for the 1IIJ peptide forming a single membrane-spanning helix (Figure S6) (PDF)

Movie showing an example trajectory of insertion and folding of the three-helix membrane TMEM14A protein fragment (PDB ID 2LOP) using the CABS-membrane protein model. (AVI)

## AUTHOR INFORMATION


### Corresponding Authors

*Tel: +48 22 55 26 365. E-mail: kolinski@chem.uw.edu.pl.
*Tel: +48 22 55 26 364. E-mail: sekmi@chem.uw.edu.pl.


### Author Contributions

W.P. and M.J. performed the research. W.P., M.K., and S.K. analyzed the data. A.K. contributed the CABS-membrane program. W.P. and S.K. wrote the manuscript draft. All of the authors contributed to the writing of the manuscript.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

We gratefully acknowledge Dr. Szymon Niewieczerzal's help in preparing scripts for the visualization of results presented here. We also acknowledge funding from the National Science Center [Grants MAESTRO 2014/14/A/ST6/0008 and 2011/01/D/NZ2/05314] and the Foundation for Polish Science TEAM Project (TEAM/2011-7/6) cofinanced by the EU European Regional Development Fund operated within the Innovative Economy Operational Program.

## REFERENCES


(1) Fagerberg, L.; Jonasson, K.; von Heijne, G.; Uhlen, M.; Berglund, L. Prediction of the Human Membrane Proteome. *Proteomics* **2010**, *10* (6), 1141−1149.

(2) Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. Drug-Target Network. *Nat. Biotechnol.* **2007**, *25* (10), 1119−1126.

(3) Bakheet, T. M.; Doig, A. J. Properties and Identification of Human Protein Drug Targets. *Bioinformatics* **2009**, *25* (4), 451−457.

(4) Koehler Leman, J.; Ulmschneider, M. B.; Gray, J. J. Computational Modeling of Membrane Proteins. *Proteins: Struct., Funct., Genet.* **2015**, *83* (1), 1−24.

(5) Moraes, I.; Evans, G.; Sanchez-Weatherby, J.; Newstead, S.; Stewart, P. D. Membrane Protein Structure Determination - the Next Generation. *Biochim. Biophys. Acta, Biomembr.* **2014**, *1838* (1), 78−87.

(6) Im, W.; Brooks, C. L., 3rd Interfacial Folding and Membrane Insertion of Designed Peptides Studied by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (19), 6771−6776.

(7) Ulmschneider, J. P.; Ulmschneider, M. B. Folding Simulations of the Transmembrane Helix of Virus Protein U in an Implicit Membrane Model. *J. Chem. Theory Comput.* **2007**, *3* (6), 2335−2346.

(8) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334* (6055), 517−520.

(9) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116* (14), 7898−7936.

(10) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Schulten, K. Coarse Grained Protein-Lipid Model with Application to Lipoprotein Particles. *J. Phys. Chem. B* **2006**, *110* (8), 3674−3684.

(11) Shi, Q.; Izvekov, S.; Voth, G. A. Mixed Atomistic and Coarse-Grained Molecular Dynamics: Simulation of a Membrane-Bound Ion Channel. *J. Phys. Chem. B* **2006**, *110* (31), 15045−15048.




DOI: 10.1021/acs.jcim.6b00350
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(12) Bond, P. J.; Sansom, M. S. Insertion and Assembly of Membrane Proteins Via Simulation. *J. Am. Chem. Soc.* **2006**, *128* (8), 2697−2704.
(13) Lopez, C. F.; Nielsen, S. O.; Srinivas, G.; Degrado, W. F.; Klein, M. L. Probing Membrane Insertion Activity of Antimicrobial Polymers Via Coarse-Grain Molecular Dynamics. *J. Chem. Theory Comput.* **2006**, *2* (3), 649−655.
(14) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The Martini Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819−834.
(15) Hall, B. A.; Chetwynd, A. P.; Sansom, M. S. Exploring Peptide-Membrane Interactions with Coarse-Grained Md Simulations. *Biophys. J.* **2011**, *100* (8), 1940−1948.
(16) Kar, P.; Gopal, S. M.; Cheng, Y. M.; Panahi, A.; Feig, M. Transferring the Primo Coarse-Grained Force Field to the Membrane Environment: Simulations of Membrane Proteins and Helix-Helix Association. *J. Chem. Theory Comput.* **2014**, *10* (8), 3459−3472.
(17) Bereau, T.; Bennett, W. F.; Pfaendtner, J.; Deserno, M.; Karttunen, M. Folding and Insertion Thermodynamics of the Transmembrane Walp Peptide. *J. Chem. Phys.* **2015**, *143* (24), 243127.
(18) Vorobyov, I.; Kim, I.; Chu, Z. T.; Warshel, A. Refining the Treatment of Membrane Proteins by Coarse-Grained Models. *Proteins: Struct., Funct., Genet.* **2016**, *84* (1), 92−117.
(19) Marrink, S. J.; de Vries, A. H.; Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **2004**, *108* (2), 750−760.
(20) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The Martini Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111* (27), 7812−7824.
(21) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini Model. *Chem. Soc. Rev.* **2013**, *42* (16), 6801−6822.
(22) Kmiecik, S.; Kolinski, A. Simulation of Chaperonin Effect on Protein Folding: A Shift from Nucleation-Condensation to Framework Mechanism. *J. Am. Chem. Soc.* **2011**, *133* (26), 10283−10289.
(23) Kmiecik, S.; Kolinski, A. Characterization of Protein-Folding Pathways by Reduced-Space Modeling. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (30), 12330−12335.
(24) Jamroz, M.; Kolinski, A.; Kmiecik, S. Protocols for Efficient Simulations of Long-Time Protein Dynamics Using Coarse-Grained Cabs Model. *Methods Mol. Biol.* **2014**, *1137*, 235−250.
(25) Wabik, J.; Kmiecik, S.; Gront, D.; Kouza, M.; Kolinski, A. Combining Coarse-Grained Protein Models with Replica-Exchange All-Atom Molecular Dynamics. *Int. J. Mol. Sci.* **2013**, *14* (5), 9893−9905.
(26) Kmiecik, S.; Kolinski, A. Folding Pathway of the B1 Domain of Protein G Explored by Multiscale Modeling. *Biophys. J.* **2008**, *94* (3), 726−736.
(27) Kmiecik, S.; Gront, D.; Kouza, M.; Kolinski, A. From Coarse-Grained to Atomic-Level Characterization of Protein Dynamics: Transition State for the Folding of B Domain of Protein A. *J. Phys. Chem. B* **2012**, *116* (23), 7026−7032.
(28) Kurcinski, M.; Kolinski, A.; Kmiecik, S. Mechanism of Folding and Binding of an Intrinsically Disordered Protein as Revealed by Ab Initio Simulations. *J. Chem. Theory Comput.* **2014**, *10* (6), 2224−2231.
(29) Kolinski, A. Protein Modeling and Structure Prediction with a Reduced Representation. *Acta Biochim. Pol.* **2004**, *51* (2), 349−371.
(30) Zhang, Y.; Kolinski, A.; Skolnick, J. Touchstone Ii: A New Approach to Ab Initio Protein Structure Prediction. *Biophys. J.* **2003**, *85* (2), 1145−1164.
(31) Liwo, A.; He, Y.; Scheraga, H. A. Coarse-Grained Force Field: General Folding Theory. *Phys. Chem. Chem. Phys.* **2011**, *13* (38), 16890−16901.
(32) Yang, J. Y.; Yan, R. X.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-Tasser Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12* (1), 7−8.
(33) Kmiecik, S.; Kolinski, A. One-Dimensional Structural Properties of Proteins in the Coarse-Grained Cabs Model. *Methods Mol. Biol.* **2017**, *1484*, 83−113.
(34) Blaszczyk, M.; Jamroz, M.; Kmiecik, S.; Kolinski, A. Cabs-Fold: Server for the De Novo and Consensus-Based Prediction of Protein Structure. *Nucleic Acids Res.* **2013**, *41* (W1), W406−W411.
(35) Kmiecik, S.; Jamroz, M.; Kolinski, M. Structure Prediction of the Second Extracellular Loop in G-Protein-Coupled Receptors. *Biophys. J.* **2014**, *106* (11), 2408−2416.
(36) Jamroz, M.; Orozco, M.; Kolinski, A.; Kmiecik, S. Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field. *J. Chem. Theory Comput.* **2013**, *9* (1), 119−125.
(37) Jamroz, M.; Kolinski, A.; Kmiecik, S. Cabs-Flex Predictions of Protein Flexibility Compared with NMR Ensembles. *Bioinformatics* **2014**, *30* (15), 2150−2154.
(38) Jamroz, M.; Kolinski, A.; Kmiecik, S. Cabs-Flex: Server for Fast Simulation of Protein Structure Fluctuations. *Nucleic Acids Res.* **2013**, *41* (W1), W427−W431.
(39) Kurcinski, M.; Jamroz, M.; Blaszczyk, M.; Kolinski, A.; Kmiecik, S. Cabs-Dock Web Server for the Flexible Docking of Peptides to Proteins without Prior Knowledge of the Binding Site. *Nucleic Acids Res.* **2015**, *43* (W1), W419−W424.
(40) Blaszczyk, M.; Kurcinski, M.; Kouza, M.; Wieteska, L.; Debinski, A.; Kolinski, A.; Kmiecik, S. Modeling of Protein-Peptide Interactions Using the Cabs-Dock Web Server for Binding Site Search and Flexible Docking. *Methods* **2016**, *93*, 72−83.
(41) Sen, T. Z.; Kloster, M.; Jernigan, R. L.; Kolinski, A.; Bujnicki, J. M.; Kloczkowski, A. Predicting the Complex Structure and Functional Motions of the Outer Membrane Transporter and Signal Transducer Feca. *Biophys. J.* **2008**, *94* (7), 2482−2491.
(42) Periole, X.; Marrink, S. J. The Martini Coarse-Grained Force Field. *Methods Mol. Biol.* **2013**, *924*, 533−565.
(43) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157* (1), 105−132.
(44) White, S. H.; Wimley, W. C. Membrane Protein Folding and Stability: Physical Principles. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 319−365.
(45) Wiener, M. C.; King, G. I.; White, S. H. Structure of a Fluid Dioleoylphosphatidylcholine Bilayer Determined by Joint Refinement of X-Ray and Neutron-Diffraction Data 0.1. Scaling of Neutron Data and the Distributions of Double-Bonds and Water. *Biophys. J.* **1991**, *60* (3), 568−576.
(46) Lazaridis, T. Effective Energy Function for Proteins in Lipid Membranes. *Proteins: Struct., Funct., Genet.* **2003**, *52* (2), 176−192.
(47) Kessel, A.; Shental-Bechor, D.; Haliloglu, T.; Ben-Tal, N. Interactions of Hydrophobic Peptides with Lipid Bilayers: Monte Carlo Simulations with M2 Delta. *Biophys. J.* **2003**, *85* (6), 3431−3444.
(48) Schramm, C. A.; Hannigan, B. T.; Donald, J. E.; Keasar, C.; Saven, J. G.; Degrado, W. F.; Samish, I. Knowledge-Based Potential for Positioning Membrane-Associated Structures and Assessing Residue-Specific Energetic Contributions. *Structure* **2012**, *20* (5), 924−935.
(49) Nugent, T.; Jones, D. T. Membrane Protein Orientation and Refinement Using a Knowledge-Based Statistical Potential. *BMC Bioinf.* **2013**, *14* (1), 276.
(50) Kmiecik, S.; Jamroz, M.; Kolinski, A. Multiscale Approach to Protein Folding Dynamics. *Multiscale Approaches to Protein Modeling* **2011**, 281−293.
(51) Jamroz, M.; Kolinski, A. Clusco: Clustering and Comparison of Protein Models. *BMC Bioinf.* **2013**, *14*, 62.
(52) Spronk, C. A.; Nabuurs, S. B.; Bonvin, A. M.; Krieger, E.; Vuister, G. W.; Vriend, G. The Precision of Nmr Structure Ensembles Revisited. *J. Biomol. NMR* **2003**, *25* (3), 225−234.
(53) Cymer, F.; von Heijne, G.; White, S. H. Mechanisms of Integral Membrane Protein Insertion and Folding. *J. Mol. Biol.* **2015**, *427* (5), 999−1022.
(54) Wimley, W. C.; White, S. H. Designing Transmembrane Alpha-Helices That Insert Spontaneously. *Biochemistry* **2000**, *39* (15), 4432−4442.


DOI: 10.1021/acs.jcim.6b00350
*J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

(55) Ulmschneider, J. P.; Smith, J. C.; White, S. H.; Ulmschneider, M. B. In Silico Partitioning and Transmembrane Insertion of Hydrophobic Peptides under Equilibrium Conditions. *J. Am. Chem. Soc.* **2011**, *133* (39), 15487−15495.
(56) Popot, J. L.; Engelman, D. M. Helical Membrane Protein Folding, Stability, and Evolution. *Annu. Rev. Biochem.* **2000**, *69*, 881−922.
(57) Engelman, D. M.; Chen, Y.; Chin, C. N.; Curran, A. R.; Dixon, A. M.; Dupuy, A. D.; Lee, A. S.; Lehnert, U.; Matthews, E. E.; Reshetnyak, Y. K.; Senes, A.; Popot, J. L. Membrane Protein Folding: Beyond the Two Stage Model. *FEBS Lett.* **2003**, *555* (1), 122−125.
(58) Milik, M.; Skolnick, J. Insertion of Peptide Chains into Lipid Membranes: An Off-Lattice Monte Carlo Dynamics Model. *Proteins: Struct., Funct., Genet.* **1993**, *15* (1), 10−25.
(59) Milik, M.; Skolnick, J. Spontaneous Insertion of Polypeptide Chains into Membranes: A Monte Carlo Model. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (20), 9391−9395.
(60) Bernsel, A.; Viklund, H.; Hennerdal, A.; Elofsson, A. Topcons: Consensus Prediction of Membrane Protein Topology. *Nucleic Acids Res.* **2009**, *37* (Suppl.2), W465−W468.
(61) Lo, A.; Cheng, C. W.; Chiu, Y. Y.; Sung, T. Y.; Hsu, W. L. Tmpad: An Integrated Structural Database for Helix-Packing Folds in Transmembrane Proteins. *Nucleic Acids Res.* **2011**, *39* (Suppl. 1), D347−D355.
(62) Xu, D.; Zhang, Y. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. *Biophys. J.* **2011**, *101* (10), 2525−2534.
(63) Lazaridis, T. Effective Energy Function for Proteins in Lipid Membranes. *Proteins: Struct., Funct., Genet.* **2003**, *52* (2), 176−192.
(64) Feig, M. Implicit Membrane Models for Membrane Protein Simulation. *Methods Mol. Biol.* **2008**, *443*, 181−196.
(65) Hopf, T. A.; Colwell, L. J.; Sheridan, R.; Rost, B.; Sander, C.; Marks, D. S. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **2012**, *149* (7), 1607−1621.
(66) Gront, D.; Kmiecik, S.; Blaszczyk, M.; Ekonomiuk, D.; Koliński, A. Optimization of Protein Models. *WIREs Comput. Mol. Sci.* **2012**, *2* (3), 479−493.



DOI: 10.1021/acs.jcim.6b00350
*J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX