Lipid contact probability: an essential and predictive character for the structural and functional studies of membrane proteins

Lei Wang[1], Jiangguo Zhang[2], Dali Wang[1], and Chen Song[1,3,*]

[1]Center for Quantitative Biology, Academy for Advanced Interdisciplinary studies, Peking University, Beijing 100871, China

[2]School of Life Sciences, Peking University, Beijing 100871, China

[3]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[*]To whom correspondence should be addressed; E-mail: c.song@pku.edu.cn

Abstract

One of the unique traits of membrane proteins is that a significant fraction of their hydrophobic amino acids is exposed to the hydrophobic core of lipid bilayers rather than being embedded in the protein interior, which is often not explicitly considered in protein structure prediction. Here, we propose a characteristic and predictive quantity, the lipid contact probability (LCP), to describe the likelihood of the amino acids of a given sequence being in direct contact with the acyl chains of lipid molecules. We show that LCP is complementary to solvent accessibility in characterizing the outer surface of membrane proteins, and it can be predicted for any given sequence with a machine learning-based method by utilizing a training dataset extracted from MemProtMD, a database generated from molecular dynamics simulations for the membrane proteins with a known structure. We demonstrate that LCP can be used to systematically improve the prediction accuracy of the contact maps of proteins, as well as to identify membrane-anchored soluble proteins from sequences.

Introduction

Proteins with various amino acid sequences are folded into specific structures with unique functions [1]. The relationship between the sequence, structure, and function of proteins has been extensively studied for decades. To characterize the structural and functional features of various proteins, researchers defined some essential and predictive properties, such as the secondary structure (SS) and solvent accessibility (SA) [2–5], which are widely used in the analysis and prediction of the structure and function of proteins [6–8].

Membrane proteins are composed of a large subgroup of proteins, which are responsible for the substance transport and signal transduction across cell membranes; therefore, they are essential for living organisms. Currently, over 50% of the known drugs target membrane proteins [9]. Therefore, studying the structures of membrane proteins is of high interest. The recent progress of Cryo-EM has facilitated the determination of many membrane protein structures [10], but these are still only a small fraction of the known sequences that encode into membrane proteins [11, 12], rendering the structural prediction of membrane proteins highly desirable.

Recent developments in deep learning have improved the protein structure prediction accuracy to a large extent [7, 8, 13]. However, researchers have typically focused mostly on soluble proteins, which are easier for structural determination; and therefore, a large structural dataset is available. For those soluble proteins, SS and SA can be routinely predicted to characterize their local structural features [2, 3, 5], which were extensively used as inputs for contact map and structure predictions [6–8, 14, 15]. Although membrane proteins share some common features with soluble proteins, and both SS and SA are essential and applicable to membrane proteins too, membrane proteins are distinct from soluble proteins in the sense that a significant part of their amino acids on the outer surface are in direct contact with the hydrophobic acyl chains of lipid molecules. This means that a large fraction of the outer surface of membrane proteins

is covered by hydrophobic residues, which would be considered not 'solvent exposed' and therefore embedded in the 'interior,' if one predicts in the same way as for soluble proteins. Strikingly, this remarkable difference between membrane proteins and soluble proteins has been rarely considered in the structural predictions to date, probably due to the absence of a quantitative prediction method for this lipid-exposing property from sequences. We believe that the lipid-contacting feature of a membrane protein should be explicitly considered and utilized with the same weight as solvent accessibility in the structural and functional studies of membrane proteins.

To account for this hydrophobic-surface feature of membrane proteins, we propose a new characteristic quantity in this work, the lipid contact probability (LCP), to describe the likelihood of direct contact between the protein amino acids and the hydrophobic acyl chains of lipid molecules. We show that we can use a deep learning-based method, DCRNN [16], to predict the LCP with a good accuracy, based on the highly informative MemProtMD database [17, 18]. We integrated LCP into the recently developed ResNet-based contact map predictor [6, 15, 19], and the results showed a consistent and significant improvement of the contact map prediction. Also, by analyzing the LCP predictions, we may be able to identify the soluble proteins that deeply bind into membranes with a certain interface from their sequences. Therefore, we propose that the LCP is an essential property of membrane proteins, and it can be predicted for any sequence, which can then be used for broad applications, such as the membrane contact prediction of soluble proteins, and the contact map and structure prediction of (membrane) proteins.

Results

LCP can be predicted with a good accuracy for membrane proteins

We consider an amino acid to be in direct contact with the hydrophobic core of membranes if its α carbon atom is within 6 Å of the lipid acyl chain carbon atoms, and define the LCP to be the fraction-of-time probability of a certain amino acid in direct contact with the hydrophobic acyl chains of a lipid bilayer. The LCP is difficult to obtain directly from experimental measurements, but it can be calculated from molecular dynamics (MD) simulations of membrane proteins. Stansfeld et al. performed systematic coarse-grained MD simulations for all of the membrane proteins with a known structure, and the simulation and analysis results were deposited into a database named MemProtMD [17, 18]. Based on this pioneering work, we extracted the LCP information of all of the simulated membrane proteins as the training dataset for our DCRNN model (please refer to the "Methods" section and Fig. S1a&b for details). With this model, we were able to predict the LCP of a given sequence with a good accuracy.

As shown in Table 1, the overall Pearson Correlation Coefficient (PCC) between our LCP prediction and the ground truth (the MD results from MemProtMD) of the studied membrane proteins reached 0.77 for the training set, 0.76 for the validation set, and 0.77 for the test set, respectively. In addition, we also trained a LCP predictor with a small dataset containing less-redundant sequences, in which the sequence identity between the training set and test set was less than 40% and the prediction accuracy was satisfactory too, with an overall PCC of about 0.65 (Table S1). In this work, we mainly used the LCP predictor trained by the big dataset (please refer to the method for details), which generated overall more accurate predictions. As a comparison, the highest prediction accuracy of SA is around 80% at present [20], after many years of development with a much larger training dataset. We further analyzed the prediction performance for the α-helix and β-sheet structures, two of the most common secondary structures of transmembrane proteins. For the residues in the α-helix structures, we

obtained a PCC about 0.84, and 0.71 for the residues in β-sheet between the prediction and ground truth in the test set (Table 1). Evidently, the prediction accuracy is better for the α-helix structures, probably because we had a larger dataset of α-helix structures in the training set (54%), compared to that of the β-sheet structures (23%). The coil structures were the most difficult ones, as these are the most flexible and less abundant (23%) structures in the datasets.

We picked an α-helix and a β-sheet membrane proteins from the test set as representative cases to investigate the prediction details. The PDBID of the α-helix membrane protein was 5aym [21], representing a crystal structure of a bacterial homologue of iron transporter ferroportin in the outward-facing state with soaked iron with more than 10 transmembrane helices and 440 residues. The PDBID of the β-sheet membrane protein was 4e1t [22], an X-ray crystal structure of the transmembrane beta-domain from invasion from *yersinia pseudotuberculosis* with 245 residues. The prediction results of the LCP and SA are shown in Fig. 1. As can be seen, most of the hydrophobic lipid-contacting amino acids were successfully predicted (Fig. 1a&b), with the overall prediction PCCs of 0.76 and 0.71 for the α-helix and β-sheet membrane proteins, respectively. We mapped the LCP onto the surface of the protein structures, and it is evident that the distribution of high LCP values is indeed in the transmembrane region, consistent with the ground truth (Fig. 1c-f). For the α-helix protein (5aym), most of the residues contacting the hydrophobic core of the lipid bilayer were identified by our LCP prediction (Fig. 1a&e). For the β-sheet protein (4e1t), there were some lipid-contacting residues missing in our prediction, as shown with the dashed circles in Fig. 1f. However, the predicted high-LCP residues are all in the transmembrane region and on the outer surface, indicating that the LCP prediction can reach a satisfactory accuracy in characterizing the membrane hydrophobic core-contacting residues. The results obtained from the LCP predictor trained by the small dataset were similar (Fig. S2).

LCP is complementary to SA and provides important structural information for membrane proteins

For the two specific membrane proteins discussed above, we also predicted the SA from their sequences [4] and colored the protein according to the SA values of each amino acid (Fig. 1g&h). As can be seen, the SA prediction was also quite reasonable, but the high-SA amino acids do not cover the full surface of the two membrane proteins, highlighting the deficiency of only using SA to describe the outer surface of a membrane protein. In fact, the high-SA region (blue colored surface) and high-LCP region (red colored surface) are complementary to each other on the outer surfaces of the two proteins, and together they give a complete description about the outer surface of the membrane protein structures (Fig. 1e-h).

To give a more comprehensive analysis, we plotted the predicted LCP and SA of two datasets in Fig. 2, and divided each figure into four regions: (I) Both the LCP and SA values are larger than 0.5, indicating the amino acids in this region are likely to be exposed to both water and lipid molecules; (II) LCP > 0.5 and SA < 0.5, meaning the amino acids in this region are more likely to be exposed to hydrophobic lipid molecules than to water molecules; (III) LCP < 0.5 and SA < 0.5, meaning the amino acids are not likely to be exposed to either lipid or water molecules, so they are probably embedded in the interior of proteins; and (IV) LCP < 0.5 and SA > 0.5, meaning the amino acids are more likely to be exposed to water molecules than to the lipid bilayer.

Fig. 2a was generated from the test dataset pdb25-test-500 [6], which contained 25 membrane proteins and 302 soluble proteins after removing the sequence redundancy with respect to our training set. Therefore, some outer-surface amino acids in the dataset should be exposed to water molecules, and some to lipid molecules. If one does not consider LCP and uses only the SA to predict and characterize the amino acids of the dataset, the residues would be considered to be either exposed to water or embedded in the interior, which is actually the logic of most of the common SA predictors. However,

from Fig. 2a, we can tell that a significant number of amino acids, especially those in region II, are neither exposed to water molecules nor embedded in the interior of proteins. They are exposed to lipid molecules. Therefore, this plot highlighted the necessity of taking the LCP into account while considering the outward-facing amino acids of a (membrane) protein.

Therefore, the amino acids located in regions II, III, and IV are likely to be those exposed to the hydrophobic core of the lipid bilayers, those embedded in the protein interior, and those exposed to water molecules, respectively. At a first glance, region I does not make much sense: are there residues that have a high probably to be exposed to both water and lipid molecules? Further analysis using the OPM server [23] revealed that most of the amino acids located in region I actually represent these residues of the membrane protein sitting at the water-membrane interface (Fig. S3). It should be noted that only a very small fraction of the amino acids was found to be located in this region under the cutoff of 0.5 for this dataset. To be specific, the amino acids in region I only account for 0.018% (14/79796) of the total amount of the amino acids in the dataset. These residues can be determined by the cross region of the high-LCP and high-SA amino acids, which are not of interest to us at this stage. Therefore, we do not include the amino acids in region I for further discussions in this work.

LCP can identify the potential interaction interface of soluble proteins with membranes

As a control, we tested the LCP prediction on a soluble protein dataset, composed of 102 Pfam proteins after removing the sequence redundancy with respect to our training set, which is often used as a test set for contact map and protein structure prediction [6]. The results are shown in Fig. 2b. As can be seen, most of the amino acids were predicted to be not lipid-exposed (LCP < 0.5), indicating that there are no strong false-positive problems in our LCP predictor. However, there were still about 14 amino

acids predicted to be likely having direct contact with the hydrophobic acyl chains of lipid molecules, which was a bit surprising since this dataset was believed to contain only soluble proteins.

These 14 amino acids were distributed in 7 proteins (Table S2). Considering our average prediction accuracy is around 70% (PCC), we decided to rule out the cases in which there were less than three high-LCP predictions in a protein, considering them outliers, and then there was only one protein left, whose PDBID was 1f6b. This is a crystal structure of the SAR1-GDP complex [24]. Our prediction indicated that six of its residues should be in direct contact with the hydrophobic lipid acyl chains, and these residues are in proximity in the sequence (Table S2). Two of the six residues were resolved in the X-ray structure, both of which were on the outer surface. We suspected that these residues may interact with the hydrophobic core of a membrane, although they belong to a soluble protein. We ran multiscale MD simulations to check whether it was possible that the soluble protein would have direct contact with the hydrophobic core of a lipid bilayer. Following Stansfeld's protein-lipid contacting definition [25], we determined the lipid-interacting residues as those within 6 Å of the lipid tails and calculated the lipid interacting probability of each residue in the simulations. Indeed, our self-assembly MD simulation results demonstrated that this protein would be stably anchored onto a membrane's surface, and both the coarse-grained and all-atom simulations confirmed a stable binding interface (Fig. 3a-c&S4). Also, previous experimental studies showed that this protein is responsible for the GDP transport from the endoplasmic reticulum (ER), and mutation experiments showed that the NH_2-terminal α helix (residues 13-19) is probably related to the stable membrane recruitment with GDP to the ER membrane [24], which provides additional support for our prediction and simulation results. Therefore, the 'abnormal' LCP prediction turned out to be a functionally relevant one: although some soluble proteins are not embedded in a lipid bilayer, they may bind and dip into the membrane deeply enough so that some amino acids can reach the hydrophobic core region.

The LCP values obtained from our DCRNN prediction and MD simulations are not completely identical (Fig. 3d&e), but the comparison demonstrates that they overlapped at residues M1, F3, G11, F12 and F18 (the purple regions in Fig. 3d), showing a converged membrane-interacting interface. Therefore, it appears that the LCP prediction is reasonable and can possibly be used to identify the potential interaction interfaces of soluble proteins with membranes.

LCP can improve the accuracy of the contact map prediction

SA has long been used as a fundamental input for protein structure prediction. Since the LCP is an essential character of amino acids and complementary to SA in describing the outer surface of membrane proteins as well as the membrane-binding surface of soluble proteins, it is natural to think that the LCP is probably helpful for protein structure prediction, too. Therefore, we tried to incorporate the LCP into a recently developed ResNet-based contact map predictor [6] to see whether the prediction accuracy can be improved. The contact map of a protein provides some of the most important information for protein structure prediction. For example, the well-known AlphaFold of Google uses CM information as one of its two crucial inputs for constructing 3D protein structure. In fact, the recent progress of CM prediction is one of the major reasons for the success of AlphaFold1 [8].

Notably, the prediction accuracy of CM was hugely improved in the last couple of years, especially since the introduction of ResNet into the field by Xu et al. [6, 26]. In the state-of-the-art ResNet method, as well as most if not all of the prediction methods, SA is an essential 1D input for the model. To check the usefulness of the LCP, we incorporated the LCP into the ResNet the same way as SA, considering the LCP as a 1D input in parallel with SA (please refer to the method and Fig. S1c for details), and then we checked whether the CM prediction can be improved.

We took the above dataset containing 327 test proteins as our prediction target, and

we compared the prediction results of our LCP-incorporated ResNet CM prediction with the original ResNet model. The top L/k (k=10, 5, 2, 1) results are shown in Table 2 and demonstrate that the inclusion of LCP in the ResNet predictor systematically improved the prediction accuracy. On average, the prediction accuracy can be improved by about 3%, if only the top L/k predictions considered. The results showed that the LCP is particularly useful for the improvement of long-range contact prediction, with an improvement of up to 7% in the first L predictions of the long-range tests. In order to evaluate the prediction accuracy for the whole contact maps, we calculated the PCC between the predictions and ground-truth contact maps calculated from known protein structures. The PCC is about 0.29 with the original ResNet predictor and about 0.37 with our LCP-incorporated ResNet predictor. Thus, the relative improvement of the prediction accuracy is about 28% with the incorporation of the LCP into the ResNet model, if the whole contact map considered.

As a further validation, we also calculated the prediction accuracy for a dataset composed of 495 test proteins, which was a subset of PDB25 [27] created in May 2020 (25% non-redundant sequences with resolution higher than 2.5 Å and R-factors less than 1.0). Any test proteins sharing > 25% sequence identity with any training proteins were excluded. Again, the ResNet predictor with LCP incorporated showed a consistently higher accuracy (Table S3). We separated the above dataset into two subsets, the soluble proteins and membrane proteins, and we calculated the prediction accuracies for each subset separately. As can be seen in Table S3, the prediction accuracy was improved by about 2% for the top predictions of all of the proteins in the dataset, which was not very significant. However, when we looked at the whole contact maps, we can see that the LCP is highly useful to improve the prediction of membrane proteins, with the PCC values of 0.132 for the original ResNet predictor and 0.222 for our LCP-incorporated ResNet predictor. Thus, the relative PCC improvement for membrane proteins is around 68%. The prediction for soluble proteins was improved too, with the values of 0.147 and 0.177 for the two models, respectively. Thus, the relative PCC improvement for

soluble proteins is around 20%. Therefore, the LCP can indeed improve the prediction accuracy of the whole contact map of proteins, although sometimes it is not immediately obvious if one only looks at the top predictions. The improvement is more significant for membrane proteins than soluble proteins.

Using the aforementioned 5aym and 4e1t as representative cases, we analyzed the differences of the CM predictions before and after the incorporation of LCP in the ResNet model, and compared to the results from other CM prediction tools [28–31] (Table S4). It is obvious that the LCP-incorporated ResNet predictor performed the best. In Fig. 4, we show the contact maps of the ground truth, and the ResNet predictions with and without using the LCP information for the two membrane proteins. The prediction with the LCP is better correlated with the ground truth with a PCC of 0.53 and 0.69, while the prediction without the LCP showed a PCC of about 0.35 and 0.62 with the ground truth, respectively. Looking at the six panels of Fig. 4, one can immediately recognize that the incorporation of LCP removed many false positives compared to the model without LCP. As a result, the calculated long-range contact prediction accuracy of top L/5 was improved by 10% and 6% for the two cases, respectively (Table S4).

Discussion

Protein folding is driven by both enthalpy and entropy. The distribution of the amino acids on the outer surface of a protein is mainly determined by entropy. For a soluble protein, the majority of the outer-surface amino acids are hydrophilic to form a better match with the surrounding water molecules and maximize the entropy. This phenomenon has been widely recognized and utilized in accessing the structural and functional properties of soluble proteins for many years [32]. The widely-accepted quantity describing which amino acids are more likey to be exposed to water is the solvent accessibility (SA), which is a crucial quantity for the analysis and prediction of soluble protein structures [4, 5, 8]. Amino acids with high SA values are likely

located on the outer surface, while those with low SA values are probably embedded in the interior of the protein to reduce the hydrophobic mismatch, thus providing a strong conformational constraint on the protein structure for a given sequence. However, for membrane proteins, SA is not the whole story about their outer surfaces. A significant part of the outer-surface amino acids of membrane proteins are hydrophobic, and they have direct contact with the hydrophobic acyl chains, also driven by entropy. Therefore, high-SA amino acids cannot cover the full surface of a membrane protein, and low SA values may even apply a false constraint on the distribution of the outer-surface hydrophobic amino acids by wrongly embedding them into the interior of the membrane protein. Therefore, we believe a complementary quantity to describe the lipid contact probability of amino acids is highly valuable for membrane protein studies.

The SA of a given protein can be routinely predicted with deep learning-based models to a high accuracy [20]. The dataset of SA for the training was generated by analyzing the outer surface of proteins with a known structure via rolling a water-size sphere over the surface [33, 34]. In principle, one can use similar protocols to generate lipid accessibility dataset of membrane proteins with a known structure. In fact, there have been such attempts [35–38] to calculate the relative or absolutely accessible surface area for lipids. For example, mp_lipid_acc [39], part of the Rosetta software suite, can identity lipid-accessible surface area or lipid accessible residues based on known α-helix and β-sheet membrane protein structures. However, in most of the above cases, researchers used a highly simplified and fixed membrane slab to define the lipid environment to calculate the lipid accessibility of residues, which was purely a spacial and configurational property. Based on these calculations, researchers also tried to predict lipid accessibility from membrane protein sequences [40–43], but the dataset of only about 100 proteins seemed too small to get a satisfactory prediction.

Lipid molecules are much larger and more complex than water molecules, introducing more uncertainties to the LCP calculation if the roll-a-sphere protocol is adopted as above. For example, the interface of some oligomeric transmembrane proteins can

be filled with lipid molecules, which is hard to determine from the structural viewpoint alone [44]. Therefore, here, we propose a new method to extract the LCP information for the membrane proteins with a known structure by utilizing the outcome of MD simulations (MemProtMD). The advantage of this method is that the LCP values were directly obtained from statistical analysis of the MD simulations, which contain the full details of the acyl chains moving around the membrane proteins, both spatially and temporally, and therefore provide much more accurate, dynamic, and complete information about the LCP.

As shown in Fig. 2, the LCP and SA are complementary to each other in defining the outer surface of a membrane protein. One can easily imagine that the inclusion of the LCP is beneficial to the structural analysis and prediction of proteins for two reasons. First, the incorporation of the LCP can tell which hydrophobic residues should face outward rather than be embedded in the interior (as learned from soluble proteins), and thus can reduce the false constraints forcing the hydrophobic residues to become embedded. Second, the opposite is probably true as well: the incorporation of the LCP can increase the true positives for the hydrophobic residues embedded inside and the hydrophilic residues exposed to the outside. Therefore, when we trained the LCP-incorporated ResNet model for CM prediction, the accuracy for both the soluble and membrane proteins was improved. In addition to improving the CM prediction accuracy, we believe the LCP can be helpful for 3D protein structure construction as well. For example, one can apply extra constraints to push the high-LCP residues onto the protein surface during the 3D structure modeling of membrane proteins, or take the LCP into account when scoring the modeled protein structures.

Our results also demonstrated an additional interesting and important application of the LCP: successive high-LCP residue distribution in a soluble protein may indicate that this protein can stably and deeply bind to the membrane surface. Therefore, one can use the LCP to screen the sequences of interest and identify the soluble proteins that can deeply bind to the membrane. There have been various tools that can predict the

transmembrane topology of a protein or membrane protein type, for example, TMHMM [45, 46], Octopus [47] and MemType-2L [48], but these were not designed to identify the deeply-anchored or half-inserted proteins on a bilayer. The combination of our LCP predictor with the existing trans-membrane topology predictor can tell us which proteins are deeply anchored on the membrane surface but non-transmembrane. A protein predicted to have successive high-LCP amino acids but not a transmembrane may be a membrane-anchored protein. The screening efficiency was high, about 12 sequences/hour on average, including the data pre-processing on one CPU, allowing us to screen the human proteome within two weeks on a six-core workstation. Without data pre-processing, the LCP can be calculated within 0.1 s per sequence from known sequence features. For the sequences with a length between 26 and 700, our calculation results predicted that about 10% of the non-transmembrane proteins of humans can deeply bind to membranes (please refer to the supplementary material). The anchoring residues and interaction interface can be identified as well, which may serve as a resource for future structural and functional studies.

In summary, we propose that the LCP is an essential, characteristic, and predictive quantity of proteins that should be explicitly considered in the study of membrane proteins. The usage of MD outcomes as the training dataset for deep learning may generate a more accurate prediction for the LCP and other similar properties of proteins. We believe that the LCP would be able to find a wide range of applications in various aspects of the structural and functional studies of proteins.

Materials and Methods

Extraction of LCP from the MemProtMD database

Our aim was to predict the LCP efficiently and accurately for any given sequence, which could then be used for the structural and functional analysis of proteins with

a known sequence but unknown structure. To do this, we needed a large dataset of the LCP of membrane proteins for the training using machine learning-based methods. It is difficult to obtain the LCP information directly from experiments, but molecular dynamics (MD) simulations have been proven to be a reliable method for studying the interactions between the membrane proteins with a known structure and lipid bilayers [49–51]. With MD simulations, we can get the dynamic and quantitative information regarding the protein and lipid contact. In fact, Stansfeld et al. performed extensive MD simulations for all of the membrane proteins with a known structure and deposited the relevant data into the MemProtMD database [17, 18, 25], which paved the way for the current study.

The MemProtMD database contains the information about the stable orientation of the membrane proteins with a known structure in an explicit lipid bilayer environment. The information was obtained by running sophisticated multi-scale MD simulations [49–51]. The database also contains statistical information of the contact between the membrane proteins and lipid bilayer. As it is more difficult to discriminate the headgroup contact from the solvent contact, we chose to only consider the hydrophobic acyl tail contact of each protein residue at this stage. We considered that a certain protein residue was in direct contact with the hydrophobic acyl tail if the distance between them was less than 6 Å. The contact probability was defined as the average occupancy of the selected groups in direct contact calculated over the final 800 ns of the coarse-grained (CG) MD trajectory, after each membrane protein simulation system reached equilibrium. There are more than 3500 MD simulation results in the web database (http://memprotmd.bioch.ox.ac.uk/) [18]. We downloaded the files of the atomistic structure with the acyl tail contact probability from the MemProtMD database. In the files, the contact probability for each atom replaced the value of the temperature factor. We extracted the value of each C_α as our LCP ground truth for each residue. Therefore, the LCP was extracted from the above procedure as shown in Fig. S1a.

We extracted 3604 simulation results from the web database in April 2019. There

were about 90% α-helical membrane proteins and 10% β-barrel membrane proteins in the dataset. We separated each file by the number of chains. We excluded the sequences that were longer than 700 residues or shorter than 26 residues. In the end, we got 12691 result files, removed the files of the same sequences, and extracted 5900 of them at random as the membrane protein dataset.

The dataset for the LCP model training

The values of LCP lie between 0 and 1, with 0 meaning no contact at all and 1 meaning persistent contact with the hydrophobic core of membranes throughout the simulation time. A fractional number tells in what percentage of the simulation time a direct contact between an amino acid and hydrophobic acyl chains of lipid molecules was observed. The above information extracted from MemProtMD provided the original dataset for the sequences of membrane proteins with a known structure.

To do the LCP prediction, we also considered soluble proteins. The training dataset was composed of 5000 membrane protein sequences from the MemProtMD database and 5000 soluble protein sequences. The soluble protein sequences were chosen randomly from the soluble proteins dataset with less than a 25% sequence identity [27]. Then, we set the LCP of each soluble protein residue to be zero. In addition, we divided the remaining 900 membrane protein sequences in to two subsets: 500 membrane protein sequences were used as the test set, and the other 400 membrane protein sequences were used as the validation set.

We also constructed the datasets with less-redundancy sequences, in which we only used the sequences with less than a 40% identity between any two sequences in the original dataset, resulting in 898 sequences. The 898 sequences were divided into three subsets: 718 randomly chosen membrane protein sequences formed the training set, 90 membrane protein sequences were used as the test set, and the other 90 membrane protein sequences were used as the validation set.

The DCRNN model for LCP prediction

We considered the LCP prediction as a regression problem and used a combination of deep convolutional and recurrent neural network (DCRNN) to do the prediction, which is one of the state-of-the-art models used for protein secondary structure prediction [16]. Due to the long-range dependencies in the protein sequence-based model, we referred to the bidirectional gate recurrent units (BGRUs) for the global context extraction, which contains a forward gate recurrent unit (GRU) [52] and a backward GRU. In the model, we also combined multiscale CNN layers for the local context extraction.

As illustrated in Fig. S1b, we can see an overview of the model for the LCP prediction. We defined the loss function as the residual sum of the squares between the values of the prediction and the ground truth with L_2-norms representing the residual sum of the squares for each sequence, length of a sequence, the serial number of each residue, the ground truth, and the prediction value of the LCP for each residue.

In our model, we utilized the protein features generated by RaptorX-Property [4], including the predicted three-state secondary structure (SS3), three-state solvent accessibility (SA), and PSSM. For each protein sequence, we ran HHblits 3.0.3 [53] (with E-value 0.001 and 3 iterations) to search the uniclust30 database dated in October 2017 to find its sequence homologous and then built its multiple sequence alignment (MSA).

Our code was implemented with Tensorflow (`https://www.tensorflow.org`) of Python (`https://www.python.org/`). The weights in our neural networks were initialized with the default parameters in Tensorflow. We trained all of the layers in our deep network using the Adam optimizer [54]. We set the batch size as 1. The training was conducted on a workstation with a six-core (12-thread) Intel Xeon E5-1650 CPU. It took around 24 hours to train one model with 200 epochs.

The ResNet model for the protein contact map prediction

The deep residual net (ResNet) has been widely used for image recognition and won the first place on the ILSVRC 2015 classification task [55]. Xu et al. developed and utilized the ResNet for the contact map prediction of proteins and won the first place on CASP 12 and CASP13 (RaptorX). To test if the LCP is useful for the structural prediction of membrane proteins, we integrated the LCP into the ResNet model created by Xu et al. [6] and checked whether the prediction accuracy could be improved. The revised model contained two residual neural networks. Fig. S1d shows each residual block of ResNet.

For the design of the convolution kernel, we used 17 in 1-D convolution and 5×5 in 2-D convolution like the original implemention in the RaptorX-Contact [6]. We constructed the model with 60 2D convolutional layers and two 1D convolutional layers when we combined the model with the LCP.

We used similar input features as the original ResNet model [6], plus the additional LCP predicted from the protein sequence in parallel to SA as a 1D input (Fig. S1c). In addition to the LCP predicted by our model, the input features included the PSSM, SS3, and SA generated by RaptorX-Property [4], as well as the EC information generated by CCMpred [28], pairwise potential, and mutual information generated by alnstats [56]. For each protein sequence, we ran HHblits 3.0.3 (with an E-value of 0.001 and 3 iterations) to search the uniclust30 database dated October 2017 to find its sequence homologous and then built its MSA. Then we generated sequence profiles from the MSA and predicted all of the needed features above.

The prediction of the protein CM was transformed into a binary classification problem. For each amino acid pair, we restricted the prediction result within [0, 1] through the sigmoid function, which represents the possibility of the two residues (β carbon) within a distance of 8 Å.

Our training data were a subset of PDB25 [27] created in April 2018, which only included proteins with less than a 25% sequence identity. We excluded a protein from

the training set if it met one of the following conditions: (1) sequence length shorter than 26 or larger than 700, (2) resolution worse than 2.5 Å, (3) has domains made up of multiple protein chains, or (4) has unusual amino acids other than the 20 standard ones. In the end, there were 10054 sequences in our training set, which contained around 150 membrane proteins.

We then evaluated the prediction accuracy of the top L/k (k = 10,5,2,1) predicted contacts, where L is the protein sequence length. The prediction accuracy was defined as the percentage of native contacts among the top L/k predicted contacts. We also divided the contacts into three groups according to the sequence distance of two residues. A contact is short-, medium-, and long-range when its sequence distance falls into [6, 11], [12, 23], and ≥ 24, respectively.

In the LCP-incorporated ResNet predictor, we compared the CM prediction results with the LCP predictors trained by the large dataset containing 5000 membrane proteins and small dataset containing only the less-redundancy membrane proteins. According to the results of the CM prediction, we used the LCP predictor trained by the large dataset in this work.

Our code was implemented with Tensorflow in Python. Weights in our neural networks were initialized with the default parameters in Tensorflow. We used the Adam optimizer to do the training with the batch size of 1. The training was conducted with Tesla V100 Nvidia GPU with 32 GB of GPU memory, on which it takes around 40 hours to train a model with 20 epochs.

Molecular dynamics simulations

We used MD simulations to validate whether the predicted soluble protein with multiple high-LCP amino acids was indeed interacting with and anchored into membrane. The atomistic protein structure was downloaded from Protein Data Bank (PDB). There were missing residues (residue 1-12 in 1f6b) in the N-terminus. To avoid the uncertainty

induced by the incomplete protein structure, we filled these missing residues with MODELLER [57]. According to the secondary structure prediction results generated by RaptorX [4], we constrained the residue 3-9 of 1f6b to form a helix. 15 possible structures were generated by MODELLER and the best-scored structure was picked as the initial structure for the following MD simulations.

First, coarse-grained (CG) MD simulations were performed with the MARTINI 3.0 force field [58]. The CG structure and topology files were generated with the script *martinize.py* [59, 60]. Following Stansfeld's protocol when generating the Mem-ProtMD database [17], we built the lipid-around-protein system with the self-assembly protocol, in which the protein was put into the simulation box with a random pose and 1,2-dihexadecanoyl-rac-glycero-3-phosphocholine (DPPC) molecules were placed randomly around the protein (Fig. 3a). After a tens-of-nanoseconds simulation, the lipid molecules formed a bilayer spontaneously, and the protein found its most stable orientation (Fig. 3a). The elastic network (ELN) was used to maintain the global conformation of the proteins during the simulations. To reproduce the flexibility of the loop, we removed the ELN between all of the loop regions and their neighboring residues. Before performing production MD simulations, we equilibrated the system to eliminate inappropriate contacts and reach the target conditions. After the 5000-step energy minimization procedure, 0.5-ns NVT (canonical ensemble) equilibration was performed with a time step of 20 fs. Then, we ran nine 500-ns independent simulations with a time step of 20 fs under the NPT (isothermal-isobaric) ensemble for each system. The V-rescale algorithm and the Berendsen algorithm were used to maintain the system temperature (310 K) and pressure (1.0 Bar) [61, 62], respectively. The electrostatic interactions were calculated with the reaction-field method. The Coulomb interaction and van der Waals interaction were both cut off at 1.1 nm.

We plotted the distance evolution between the center of mass (COM) of the bilayer and the COM of the protein over the simulation time in Fig. 3b, which can be used to evaluate whether the protein binds to the bilayer or not. We also calculated the

angle between the first principal axis of the protein and the Z-axis to reflect the protein orientation on the bilayer (Fig. 3c).

From Fig. 3b&c, we can see that 1f6b reached equilibrium after the first tens-of-nanoseconds simulation in all of the nine trajectories. Therefore, we considered the first 100-ns trajectories as the equilibrated procedure and calculated the protein-lipid interacting probability based on the last 400-ns trajectory of each simulation. The equilibrated structure is shown in Fig. 3a.

Then the script *backward.py* [63] was used to transform the equilibrated coarse-grained system (Fig. 3a) to all-atom system, which was utilized as the initial system (Fig. S4a) for the following all-atom simulations. After the 5000-step energy minimization, the system was equilibrated for 0.5 ns in the NVT ensemble and 1.0 ns in the NPT ensemble. Position restraints with a force constant of 1000 kJ/mol/nm^2 were applied on all heavy atoms of the protein to maintain the conformation during the equilibration process. The Berendsen algorithm [62] was used to keep the system temperature and pressure at 310 K and 1.0 Bar, respectively. The van der Waals interaction were cut off at a distance of 1.0 nm. The long-range electrostatic interactions were calculated with the Particle-Mesh Ewald (PME) method [64]. After the system was equilibrated to the desired condition, we removed the position restraints and performed a 200-ns all-atom MD simulation to evaluate the stability of the protein anchoring on the membrane surface. The temperature and pressure coupling algorithms were set to V-rescale and Parrinello-Rahman [61, 65] to maintain the system temperature at 310 K and pressure at 1.0 Bar, respectively. Both coupling constants were set to 1.0 ps. The all-atom MD simulation was performed with the *Amber99sb-ildn* force field [66] in combination with the *Slipids* force field [67, 68].

References

[1] M. Howarth. Say it with proteins: an alphabet of crystal structures. *Nature Structural &Molecular Biology*, 22:349, 2015.

[2] Zhiyong Wang, Feng Zhao, Jian Peng, and J. Xu. Protein 8-class secondary structure prediction using conditional neural fields. *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 109–114, 2010.

[3] S. Wang, J. Peng, Jianzhu Ma, and J. Xu. Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports*, 6:18962, 2016.

[4] S. Wang, Wei Li, S. Liu, and J. Xu. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research*, 44:W430 – W435, 2016.

[5] Yaoqi Zhou, Andrzej Kloczkowski, Eshel Faraggi, and Yuedong Yang, editors. *Prediction of Protein Secondary Structure*, volume 1484 of *Methods in Molecular Biology*. Springer New York, New York, NY, 2017.

[6] S. Wang, S. Sun, Z. Li, Renyu Zhang, and J. Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biology*, 13, 2017.

[7] Jianyi Yang, I. Anishchenko, Hahnbeom Park, Zhenling Peng, S. Ovchinnikov, and D. Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117:1496 – 1503, 2020.

[8] A. Senior, R. Evans, John Jumper, James R M Kirkpatrick, L. Sifre, Tim Green, C. Qin, Augustin ídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, S. Petersen, K. Simonyan, Steve Crossan, Pushmeet Kohli, D. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577:706–710, 2020.

[9] M. Uhlén, L. Fagerberg, B. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Åsa Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, Ingmarie Olsson, K. Edlund, E. Lundberg, Sanjay Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, Jenny Ottosson Takanen, S. Hober, T. Alm, P. Edqvist, Holger Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. Schwenk, M. Hamsten, Kalle von Feilitzen, Mattias Forsberg, L. Persson, Fredric Johansson, Martin Zwahlen, G. von Heijne, J. Nielsen, and F. Pontén. Tissue-based map of the human proteome. *Science*, 347:1260419, 2015.

[10] Y. Cheng. Membrane protein structural biology in the era of single particle cryo-EM. *Current opinion in structural biology*, 52:58–63, 2018.

[11] B. Rost and Christian Sander. Bridging the protein sequence-structure gap by structure predictions. *Annual review of biophysics and biomolecular structure*, 25:113–136, 1996.

[12] Tilman Flock, A. Venkatakrishnan, KR Vinothkumar, and M. Babu. Deciphering membrane protein structures from protein sequences. *Genome Biology*, 13:160, 2012.

[13] Arunima Singh. Deep learning 3D structures. *Nature Methods*, 17:249, 2020.

[14] B. Adhikari, Jie Hou, and J. Cheng. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, 34:1466 – 1472, 2018.

[15] Jack Hanson, K. Paliwal, Thomas Litfin, Yuedong Yang, and Y. Zhou. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, 34:40394045, 2018.

[16] Z. Li and Y. Yu. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. In *IJCAI*, 2016.

[17] Phillip J. Stansfeld, Joseph E. Goose, M. Caffrey, E. P. Carpenter, J. Parker, S. Newstead, and M. Sansom. MemProtMD: Automated insertion of membrane protein structures into explicit lipid membranes. *Structure(London, England:1993)*, 23:1350 – 1361, 2015.

[18] Thomas D. Newport, M. Sansom, and Phillip J. Stansfeld. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Research*, 47:D390 – D397, 2019.

[19] Y. Li, C. Zhang, Eric W. Bell, Dong-Jun Yu, and Y. Zhang. Ensembling multiple raw coevolutionary features with deep residual neural networks for contactmap prediction in CASP13. *Proteins: Structure*, 87:1082 – 1091, 2019.

[20] Buzhong Zhang, L. Li, and Qiang Lü. Protein solvent-accessibility prediction by a stacked deep bidirectional recurrent neural network. *Biomolecules*, 8:33, 2018.

[21] Reiya Taniguchi, H. Kato, J. Font, C. Deshpande, Miki Wada, K. Ito, R. Ishitani, M. Jormakka, and O. Nureki. Outward- and inward-facing structures of a putative bacterial transition-metal transporter with homology to ferroportin. *Nature Communications*, 6:8545, 2015.

[22] J. W. Fairman, Nathalie Dautin, D. Wójtowicz, W. Liu, N. Noinaj, T. J. Barnard, E. Udho, T. Przytycka, V. Cherezov, and S. Buchanan. Crystal structures of the outer membrane domain of intimin and invasin from enterohemorrhagic E. coli and enteropathogenic Y. pseudotuberculosis. *Structure*, 20 7:1233–1243, 2012.

[23] Mikhail A. Lomize, I. Pogozheva, Hyeon Joo, H. Mosberg, and A. Lomize. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Research*, 40:D370 – D376, 2012.

[24] Mingdong Huang, J. Weissman, S. Béraud-Dufour, P. Luan, C. Wang, W. Chen, M. Aridor, I. Wilson, and W. Balch. Crystal structure of Sar1-GDP at 1.7 å resolution and the role of the NH2 terminus in ER export. *The Journal of Cell Biology*, 155:937 – 948, 2001.

[25] Phillip J. Stansfeld, E. Jefferys, and M. Sansom. Multiscale simulations reveal conserved patterns of lipid interactions with aquaporins. *Structure(London, England:1993)*, 21:810 – 819, 2013.

[26] Sheng Wang, Z. Li, Yi zhou Yu, and Jinbo Xu. Folding membrane proteins by deep transfer learning. *Cell systems*, 5 3:202–211.e3, 2017.

[27] G. Wang and Roland L. Dunbrack. PISCES: a protein sequence culling server. *Bioinformatics*, 19 12:1589–1591, 2003.

[28] Stefan Seemayer, Markus Gruber, and J. Söding. CCMpredfast and precise prediction of protein residueresidue contacts from correlated mutations. *Bioinformatics*, 30:3128 – 3130, 2014.

[29] D. Jones, Daniel W. A. Buchan, Domenico Cozzetto, and M. Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28 2:184–190, 2012.

[30] D. Jones, T. Singh, T. Kosciolek, and Stuart J. Tetchner. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31:999 – 1006, 2015.

[31] Daniel W. A. Buchan and D. Jones. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins*, 86:78 – 83, 2018.

[32] A. Lesk and C. Chothia. Solvent accessibility, protein surfaces, and protein folding. *Biophysical journal*, 32 1:35–47, 1980.

[33] W. Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogenbonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[34] Jianzhu Ma and Sheng Wang. AcconPred: Predicting solvent accessibility and contact number simultaneously by a multitask learning framework under the conditional neural fields model. *BioMed Research International*, 2015:678764, 2015.

[35] T. Beuming and H. Weinstein. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics*, 20 12:1822–1835, 2004.

[36] Z. Yuan, Fasheng Zhang, M. Davis, M. Bodén, and R. Teasdale. Predicting the solvent accessibility of transmembrane residues from protein sequence. *Journal of proteome research*, 5 5:1063–1070, 2006.

[37] Kristoffer Illergård, Simone Callegari, and A. Elofsson. MPRAP: An accessibility predictor for a-helical transmem-brane proteins that performs well inside and outside the membrane. *BMC Bioinformatics*, 11:333, 2009.

[38] Chang Lu, Zhe Liu, Bowen Kan, Yingli Gong, Zhiqiang Ma, and Han Wang. TMP-SSurface: A deep learning-based predictor for surface accessibility of transmembrane protein residues. *Crystals*, 9(12):640, December 2019.

[39] Julia Koehler Leman, Sergey Lyskov, and R. Bonneau. Computing structure-based lipid accessibility of membrane proteins with mp_lipid_acc in RosettaMP. *BMC Bioinformatics*, 18:115, 2017.

[40] L. Adamian and J. Liang. Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Structural Biology*, 6:13, 2006.

[41] M. Phatak, Rafal Adamczak, Baoqiang Cao, M. Wagner, and J. Meller. Solvent

and lipid accessibility prediction as a basis for model quality assessment in soluble and membrane proteins. *Current protein & peptide science*, 12 6:563–573, 2011.

[42] Jhih-Siang Lai, Cheng-Wei Cheng, A. Lo, Ting-Yi Sung, and Wen-Lian Hsu. Lipid exposure prediction enhances the inference of rotational angles of transmembrane helices. *BMC Bioinformatics*, 14:304, 2013.

[43] T. Nugent and D. Jones. Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Computational Biology*, 6, 2010.

[44] M. Zhang, D. Wang, Y. Kang, J. Wu, Fuqiang Yao, Chengfang Pan, Zhiqiang Yan, Chen Song, and L. Chen. Structure of the mechanosensitive OSCA channels. *Nature Structural & Molecular Biology*, 25:850–858, 2018.

[45] E. Sonnhammer, G. Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 6:175–182, 1998.

[46] A. Krogh, B. Larsson, G. von Heijne, and E. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305 3:567–580, 2001.

[47] Håkan Viklund and A. Elofsson. OCTOPUS: improving topology prediction by two-track ann-based preference scores and an extended topological grammar. *Bioinformatics*, 24 15:1662–1668, 2008.

[48] K. Chou and Hong-Bin Shen. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochemical and biophysical research communications*, 360 2:339–345, 2007.

[49] P. Bond and M. Sansom. Insertion and assembly of membrane proteins via simulation. *Journal of the American Chemical Society*, 128 8:2697–2704, 2006.

[50] K. A. Scott, P. Bond, A. Ivetac, A. Chetwynd, S. Khalid, and M. Sansom. Coarse-grained MD simulations of membrane protein-bilayer self-assembly. *Structure*, 16 4:621–630, 2008.

[51] C. Arnarez, J. Mazat, J. Elezgaray, Siewert-J Marrink, and X. Periole. Evidence for cardiolipin binding sites on the membrane-exposed surface of the cytochrome bc_1. *Journal of the American Chemical Society*, 135 8:3112–3120, 2013.

[52] R. Józefowicz, W. Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *ICML*, 2015.

[53] Michael Remmert, A. Biegert, Andreas Hauser, and J. Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9:173–175, 2012.

[54] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[55] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[56] M. Betancourt and D. Thirumalai. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science*, 8:361–369, 1999.

[57] A. ali and T. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234 3:779–815, 1993.

[58] S. Marrink, H. J. Risselada, S. Yefimov, D. Tieleman, and A. H. de Vries. The MARTINI force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry. B*, 111 27:7812–7824, 2007.

[59] L. Monticelli, S. Kandasamy, X. Periole, R. Larson, D. Tieleman, and Siewert-Jan Marrink. The MARTINI coarse-grained force field: Extension to proteins. *Journal of chemical theory and computation*, 4 5:819–834, 2008.

[60] Djurre H. de Jong, G. Singh, W. Bennett, C. Arnarez, T. Wassenaar, L. Schäfer, X. Periole, D. Tieleman, and S. Marrink. Improved parameters for the martini coarse-grained protein force field. *Journal of chemical theory and computation*, 9 1:687–697, 2013.

[61] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126 1:014101, 2007.

[62] H. Berendsen, J. P. Postma, W. F. Gunsteren, A. Dinola, and J. Haak. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics*, 81:3684–3690, 1984.

[63] T. Wassenaar, Kristyna Pluhackova, R. Böckmann, S. Marrink, and D. Tieleman. Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of chemical theory and computation*, 10 2:676–690, 2014.

[64] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An $N\log(N)$ method for Ewald sums in large systems. *Journal of Chemical Physics*, 98:10089–10092, 1993.

[65] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52:7182–7190, 1981.

[66] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. Dror, and D. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78:1950 – 1958, 2010.

[67] Joakim P. M. Jämbeck and A. Lyubartsev. Derivation and systematic validation of a refined all-atom force field for phosphatidylcholine lipids. *The Journal of Physical Chemistry. B*, 116:3164 – 3179, 2012.

[68] Joakim P. M. Jämbeck and A. Lyubartsev. An extension and further validation of an all-atomistic force field for biological membranes. *Journal of chemical theory and computation*, 8 8:2938–2948, 2012.

Acknowledgements

We thank Xiao Luo and Prof. Minghua Deng at the School of Mathematical Sciences at Peking University, who discussed with us and provided instructive suggestions. This work was supported by grants from the National Natural Science Foundation of China (21873006 and 32071251) and the Ministry of Science and Technology of China (2016YFA0500401). Part of the MD simulations was performed on the Computing Platform of the Center for Life Sciences at Peking University.

Author Contributions

C.S. conceived the idea and supervised the project. L.W. designed the LCP predictor and performed the calculation and analysis. J.Z. reproduced the ResNet model. D.W. performed molecular dynamics simulations and related analysis. C.S., L.W., and D.W. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Tables

Table 1: The performance of our LCP predictor

Evaluation	Training	Validation	Test
Overall			
MSE	0.047	0.050	0.048
PCC	0.774	0.764	0.769
$\alpha-$helix (H)			
MSE	0.053	0.057	0.056
PCC	0.848	0.832	0.838
$\beta-$sheet (E)			
MSE	0.026	0.028	0.027
PCC	0.737	0.697	0.713
Coil (C)			
MSE	0.012	0.012	0.013
PCC	0.614	0.561	0.580

Table 2: The overall contact prediction accuracy of the 327 test proteins

Methods	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
CCMpred	0.46	0.33	0.19	0.13	0.53	0.40	0.24	0.16	0.65	0.58	0.42	0.29
ResNet	0.75	0.62	0.39	0.23	0.80	0.70	0.49	0.31	0.89	0.85	0.74	0.59
ResNet+LCP	0.78	0.65	0.41	0.24	0.85	0.75	0.53	0.33	0.91	0.89	0.80	0.66

Figures

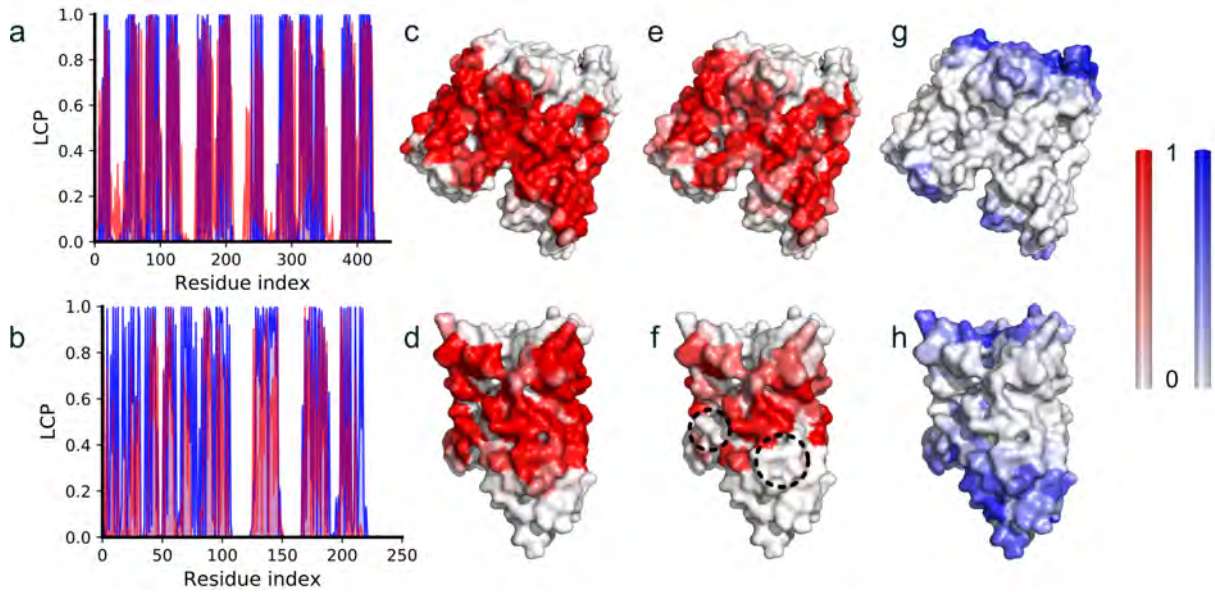

Figure 1: The LCP prediction of two representative proteins. (a-b) Comparison between the ground truth (blue) and the prediction (red) of LCP. (c-d), The colored outer surface of the two proteins according to the ground-truth LCP values obtained from MD simulations. (e-f), Similar to (c-d), but colored according to the predicted LCP values. (g-h), Similar to (c-d), but colored according to the predicted SA values by RaptorX.

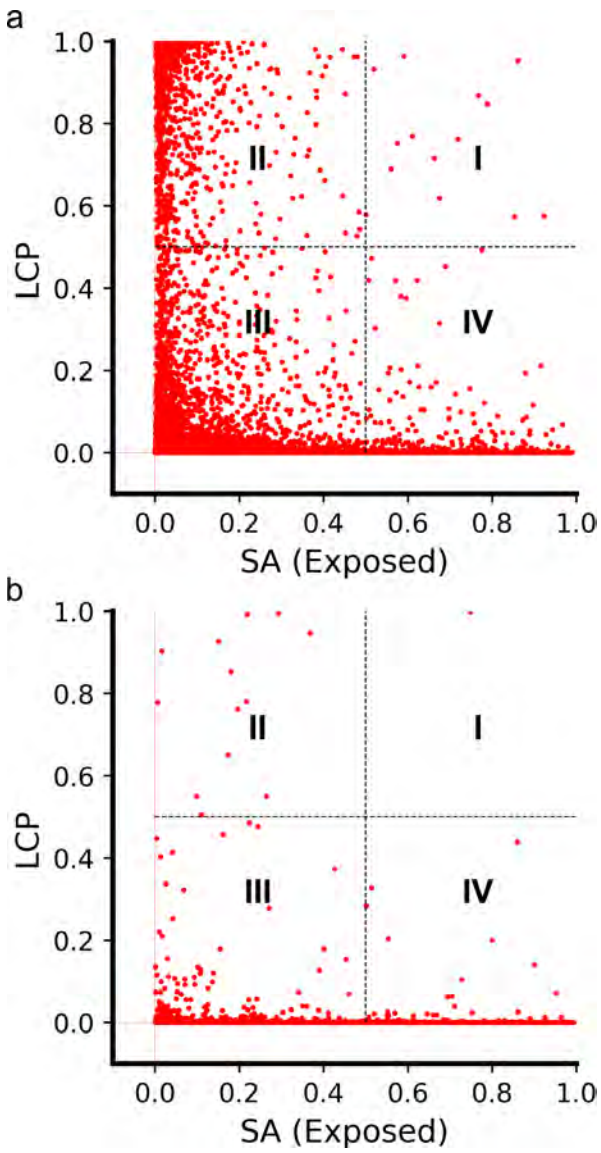

Figure 2: The 2D plot of the complementary LCP and SA predicting the likely location of amino acids in a protein, for the 327 protein dataset (a) and 102 Pfam protein dataset (b), respectively.

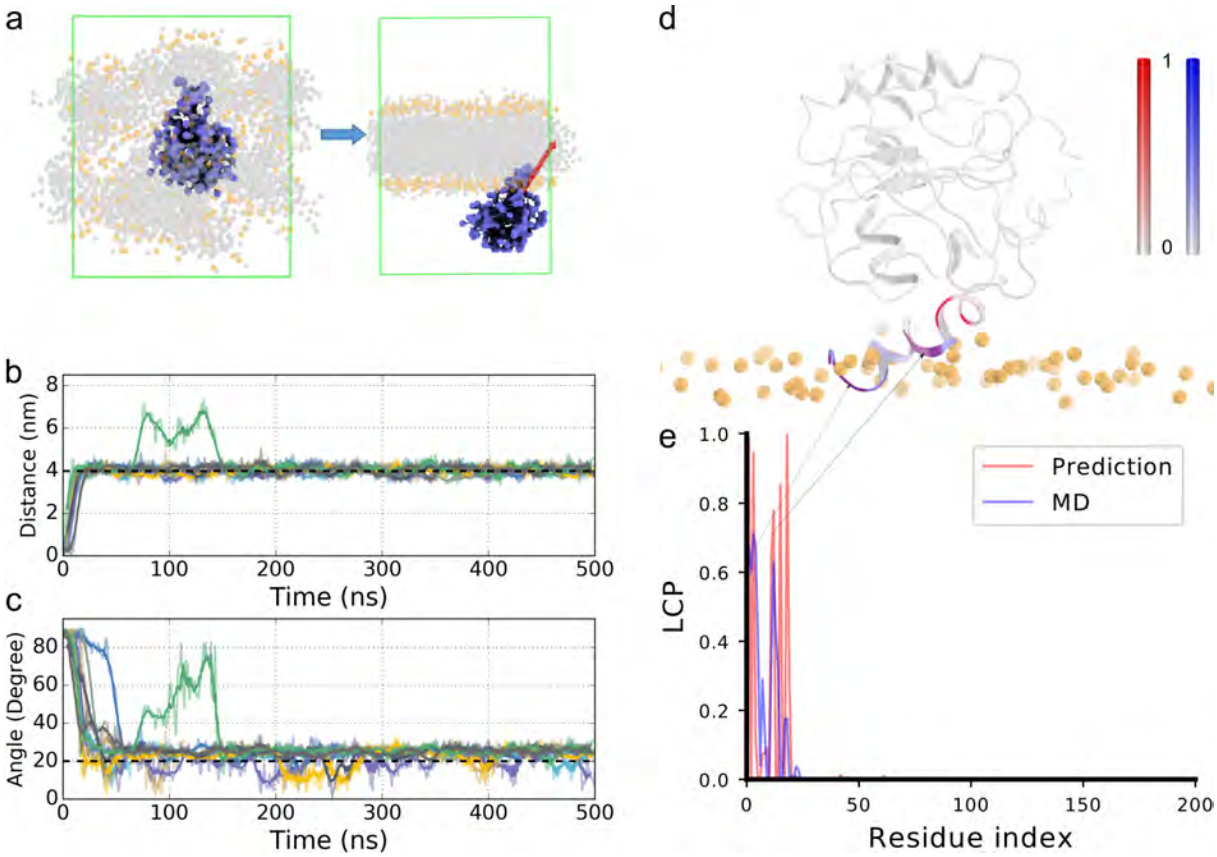

Figure 3: The coarse-grained MD simulation results for 1f6b. (a) The evolution (initial and final structures) of the 1f6b system in the coarse-grained self-assembly MD simulations. Orange spheres and grey spheres represent lipid head groups and hydrophobic tails, respectively. The red arrow represents the first principal axis of the protein. Water molecules were filled in the whole simulation box, but are not shown here for clarity. (b) The distance between the COM of the protein and the COM of the bilayer in the nine simulation trajectories of 1f6b. (c) The orientation of 1f6b in the trajectories, represented by the angle between the first principal axis of the protein structure and the Z-axis of the simulation system. (d) The orientation of 1f6b bound to the lipid bilayer obtained from MD simulations, with the membrane-interacting residues colored according to panel (e). (e) The overlaid LCP generated by our LCP predictor (red) and MD simulations (blue).

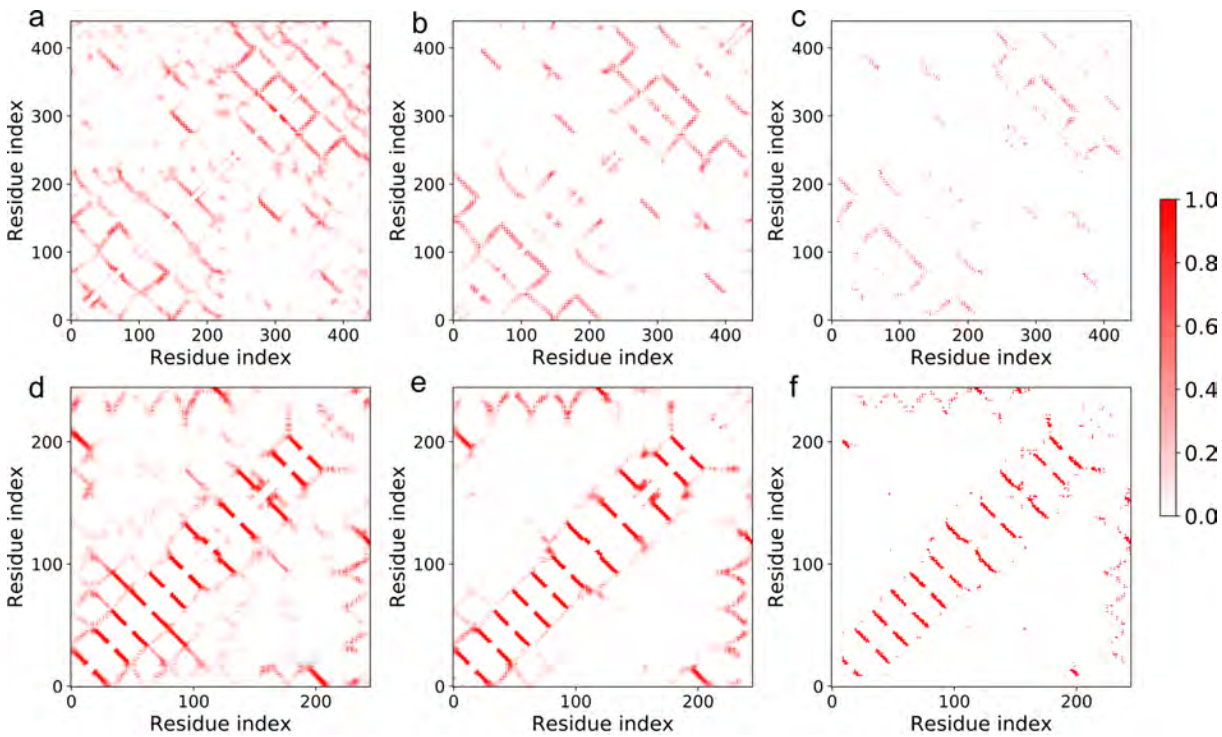

Figure 4: The details of the predicted LCP improving the CM prediction for two representative cases. (a), (b) The predicted CMs for 5aym without and with LCP incorporated, respectively. (c) The ground-truth CM calculated from the known structure 5aym. (d-f) Similar to (a-c), but for protein 4e1t.

Supplementary Information for: Lipid contact probability: an essential and predictive character for the structural and functional studies of membrane proteins

Lei Wang[1], Jiangguo Zhang[2], Dali Wang[1], and Chen Song[1,3,*]

[1]Center for Quantitative Biology, Academy for Advanced Interdisciplinary studies, Peking University, Beijing 100871, China

[2]School of Life Sciences, Peking University, Beijing 100871, China

[3]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

[*]To whom correspondence should be addressed; E-mail: c.song@pku.edu.cn

Supplementary Tables

Table S1: The performance of our LCP predictor trained by the less-redundancy dataset

Evaluation	Training	Validation	Test
Overall			
MSE	0.057	0.074	0.080
PCC	0.705	0.656	0.635
$\alpha-$helix (H)			
MSE	0.063	0.080	0.090
PCC	0.819	0.766	0.728
$\beta-$sheet (E)			
MSE	0.032	0.045	0.041
PCC	0.706	0.598	0.575
Coil (C)			
MSE	0.014	0.021	0.018
PCC	0.618	0.446	0.429

Table S2: Amino acids predicted with high LCP values in the 102 Pfam soluble protein dataset.

Sequence ID	Residue ID (LCP $\geq$ 0.5)	LCP values
1f6bA	M1	0.996
	F3	0.947
	G11	0.651
	F12	0.779
	V15	0.854
	F18	0.995
1mugA	V2	0.990
	Q126	0.550
1rybA	Y66	0.550
	L104	0.903
1gz2A	F98	0.927
1qf9A	M192	0.781
1vfyA	L186	0.762
1vhuA	A14	0.506

Table S3: The contact map prediction accuracy of the additional 495-protein dataset: overall, soluble, and membrane proteins

495 test proteins in total												
Methods	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
ResNet	0.75	0.62	0.38	0.22	0.78	0.67	0.44	0.27	0.86	0.82	0.71	0.55
ResNet + LCP	0.76	0.64	0.40	0.23	0.79	0.69	0.46	0.28	0.86	0.83	0.73	0.57
480 soluble proteins												
Methods	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
ResNet	0.75	0.62	0.38	0.22	0.78	0.68	0.45	0.27	0.86	0.82	0.71	0.56
ResNet + LCP	0.76	0.64	0.40	0.23	0.80	0.69	0.46	0.28	0.86	0.83	0.73	0.57
15 membrane proteins												
Methods	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
ResNet	0.59	0.47	0.27	0.15	0.64	0.53	0.30	0.17	0.82	0.79	0.67	0.50
ResNet + LCP	0.60	0.50	0.29	0.16	0.66	0.53	0.31	0.18	0.83	0.80	0.70	0.52

Table S4: The contact map prediction accuracy for the two representative cases: 5aym and 4e1t

Contact prediction accuracy for 5aym												
Methods	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
CCMpred	0.16	0.09	0.05	0.04	0.07	0.05	0.04	0.03	0.34	0.18	0.10	0.08
PSICOV	0.30	0.18	0.09	0.06	0.20	0.11	0.06	0.05	0.73	0.64	0.41	0.29
MetaPSICOV	0.39	0.26	0.14	0.08	0.36	0.25	0.15	0.08	0.66	0.64	0.50	0.35
MetaPSICOV2	0.41	0.27	0.15	0.08	0.43	0.28	0.16	0.10	0.77	0.68	0.56	0.38
ResNet	0.48	0.32	0.16	0.09	0.55	0.35	0.20	0.11	0.86	0.76	0.58	0.48
ResNet + LCP	0.59	0.35	0.16	0.09	0.70	0.42	0.21	0.11	0.95	0.86	0.76	0.54
Contact prediction accuracy for 4e1t												
Methods	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
PSICOV	0.33	0.31	0.20	0.17	0.46	0.29	0.20	0.17	0.38	0.27	0.20	0.13
CCMpred	0.54	0.51	0.34	0.22	0.71	0.53	0.34	0.21	0.54	0.53	0.30	0.18
MetaPSICOV	0.79	0.82	0.59	0.38	0.83	0.82	0.68	0.49	0.58	0.43	0.36	0.23
MetaPSICOV2	0.88	0.82	0.66	0.39	1.00	0.94	0.77	0.56	0.71	0.61	0.43	0.28
ResNet	1.00	0.96	0.79	0.44	1.00	0.98	0.91	0.66	1.00	0.92	0.80	0.56
ResNet + LCP	1.00	0.98	0.81	0.44	1.00	1.00	0.96	0.66	1.00	0.98	0.83	0.63

Supplementary Figures

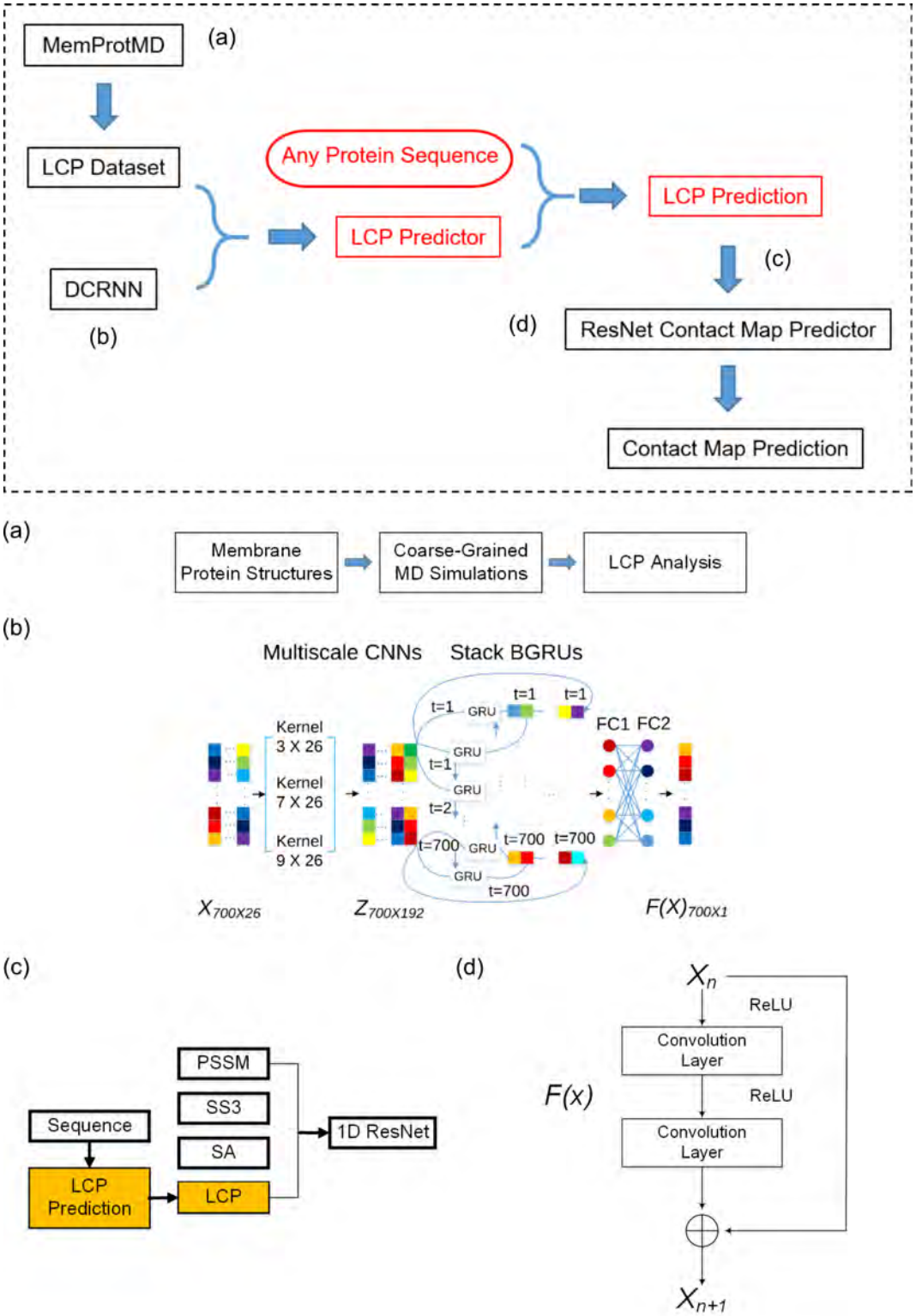

Figure S1: The method schemes. The overall scheme is presented in the dashed rectangular box, and the key steps are described in: (a) Extraction of LCP from the MemProtMD database. (b) The DCRNN model for the LCP prediction. (c) The method of LCP incorporated into the ResNet model. (d) The unit of the ResNet model.

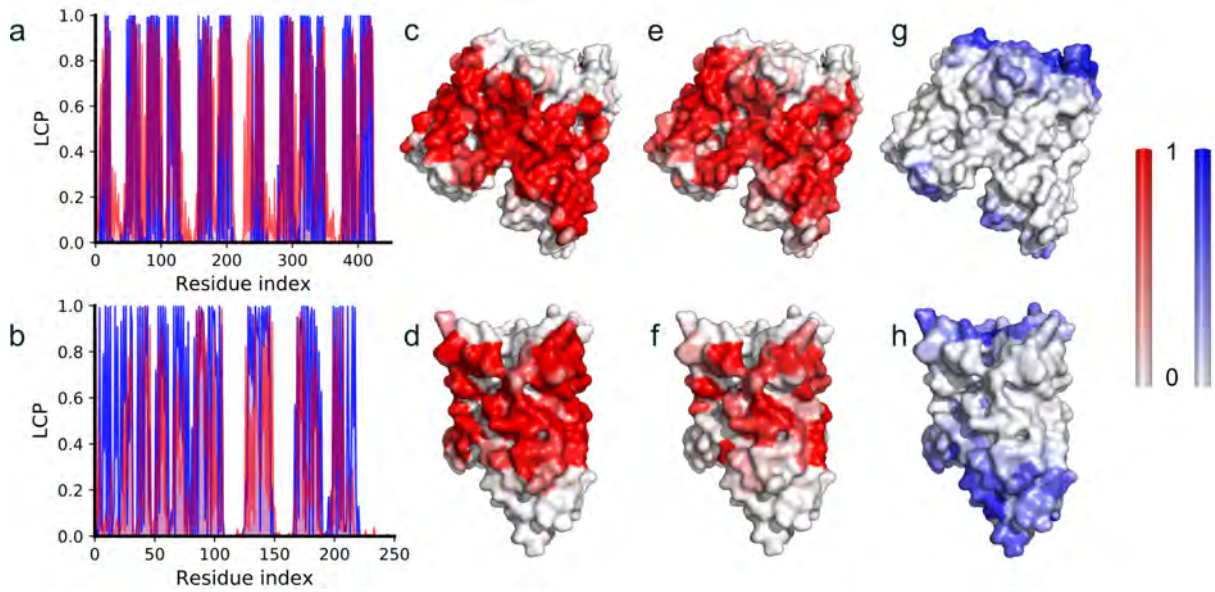

Figure S2: The LCP prediction for two representative proteins with the predictor trained by the less-redundancy dataset. (a-b) Comparison between the ground truth (blue) and the prediction (red) of LCP. (c-d) The colored outer surface of the two proteins according to the ground-truth LCP values. (e-f) Similar to (c-d), but colored according to the predicted LCP values. (g-h) Similar to (c-d), but colored according to the predicted SA values by RaptorX.

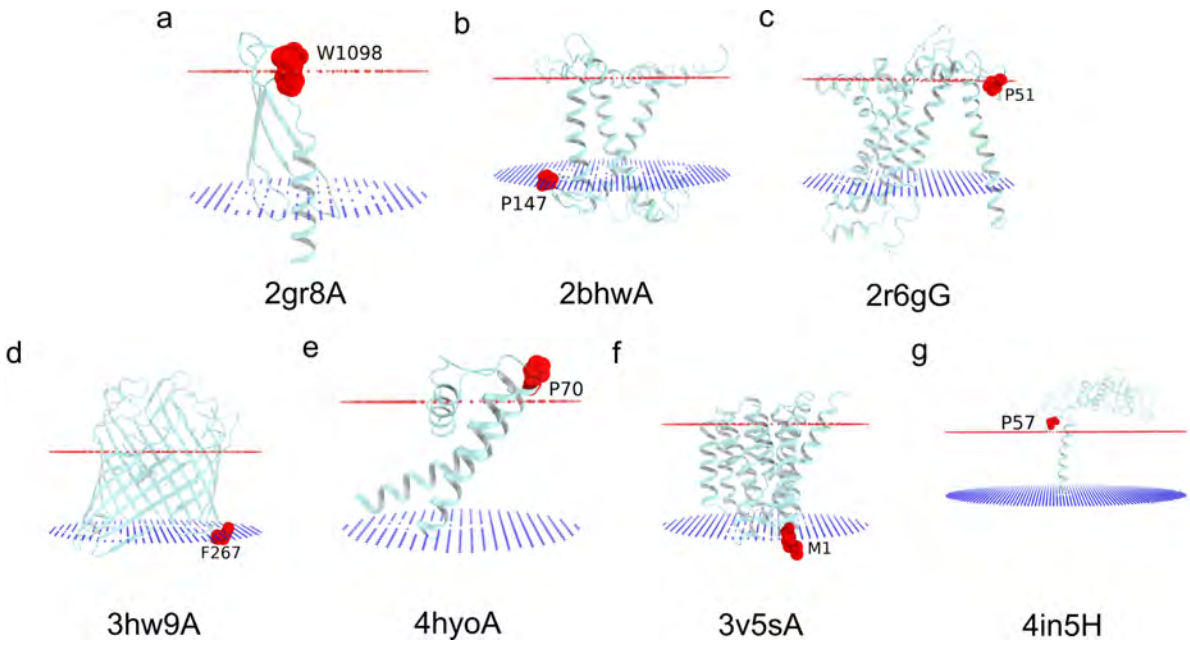

Figure S3: The structures of membrane proteins showing where the residues in the region I of Figure 2a are. The hydrophobic boundaries of the lipid bilayer are represented as the red and blue pseudo-atoms, meaning outer cell membrane and inner cell membrane, respectively.

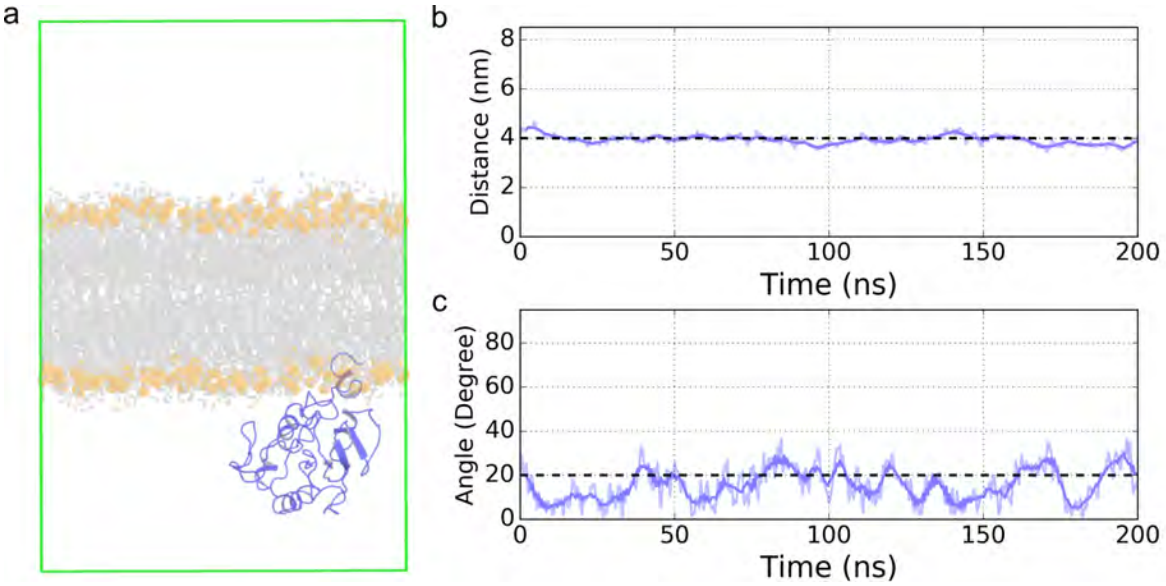

Figure S4: The all-atom MD simulation results for 1f6b. (a) The all-atom structure of 1f6b transformed from Fig.3a. (b) The distance evolution between the COM of the protein 1f6b and the COM of the bilayer, showing that the protein was stably anchored on the bilayer with a nearly constant COM distance of about 4 nm. (c) The angle between the first principal axis of the protein 1f6b and Z-axis, showing that the protein rotating on the bilayer surface within a narrow angle range.