PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Membrane contact probability: An essential and predictive character for the structural and functional studies of membrane proteins

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

1 Center for Quantitative Biology, Academy for Advanced Interdisciplinary studies, Peking University, Beijing, China, **2** School of Life Sciences, Peking University, Beijing, China, **3** Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

* 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 the protein structure and function predictions. Here, we propose a characteristic and predictive quantity, the membrane contact probability (MCP), 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 MCP 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. As the first of many potential applications, we demonstrate that MCP can be used to systematically improve the prediction precision of the protein contact maps and structures.

Author summary

The distribution of residues on protein surfaces is largely determined by the surrounding environment. For soluble proteins, most of the residues on the outer surface are hydrophilic, and people use the quantity "solvent accessibility" to describe and predict these surface residues. In contrast, for membrane proteins that are embedded in a lipid bilayer, many of their surface residues are hydrophobic and membrane-contacting, but there is yet a widely-accepted quantity for the description or prediction of this characteristic property. Here, we propose a new quantity termed "membrane contact probability (MCP)", which can be used to describe and predict the membrane-contacting surface residues of proteins. We also propose a machine learning-based method to predict MCP from protein sequences, utilizing the dataset generated by physics-based computer simulations. We demonstrate that a quantity such as MCP is helpful for protein structure prediction, and we believe that it will find broad applications in the structure and function studies of membrane proteins.

Citation: Wang L, Zhang J, Wang D, Song C (2022) Membrane contact probability: An essential and predictive character for the structural and functional studies of membrane proteins. PLoS Comput Biol 18(3): e1009972. https://doi.org/10.1371/journal.pcbi.1009972

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: March 11, 2021

Accepted: February 25, 2022

Published: March 30, 2022

Data Availability Statement: The source code of the MCP predictor and MCP-incorporated CM predictor can be found on GitHub (https://github.com/computbiophys), and a public computation server can be found online (http://www.songlab.cn).

Funding: This work was supported by grants from the Ministry of Science and Technology of China (2021YFE0108100 to C.S.) and the National Natural Science Foundation of China (21873006 and 32071251 to C.S.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

This is a *PLOS Computational Biology* Methods paper.

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 represent a large subgroup of proteins, which are responsible for the substance transport and signal transduction across cell membranes. 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, 13–15]. 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 are then extensively used as inputs for contact map (CM) and structure predictions [6–8, 16, 17]. 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 membrane-contacting feature of a membrane protein should be explicitly considered and utilized with the same weight as SA in the structure and function studies of membrane proteins.

The SA of a given protein can be predicted with deep learning-based models to high accuracy [18]. 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 [19, 20]. In principle, one can use similar protocols to generate lipid accessibility datasets of membrane proteins with a known structure. In fact, there have been such attempts [21–24] to calculate the relative or absolutely accessible surface area for lipids. For example, mp_lipid_acc [25], 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 shape-fixed membrane slab to define the lipid environment for the calculation of the lipid accessibility of residues, which was merely a spacial property. Based on these calculations, researchers also tried to predict lipid accessibility from membrane protein sequences [26–29], but the dataset of only about 100 proteins seemed too small to get a satisfactory prediction. Therefore, there has not been a well-defined and widely accepted quantity to describe the membrane-contacting properties of amino acids of a sequence, although there are multiple methods to predict the transmembrane topology of a membrane protein [30–34] (S1 Table).

To account for the hydrophobic-surface feature of membrane proteins, we propose a new characteristic quantity in this work, the membrane contact probability (MCP), 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 [35], to predict the MCP to a good accuracy for a given protein sequence, based on the highly informative data from the MemProtMD database [36, 37]. We integrated MCP into the recently developed ResNet-based contact map predictor [6, 17, 38], and the results showed a consistent and significant improvement of the contact map and structure prediction. Therefore, we propose that the MCP is an essential property of membrane proteins, which can be predicted and used for broad applications such as the contact map and structure prediction of (membrane) proteins.

Results

MCP can be predicted with 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 MCP 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 MCP is difficult to obtain directly from experimental measurements, but it can be easily calculated from molecular dynamics (MD) simulations of membrane proteins. Stansfeld et al. performed systematic coarse-grained (CG) 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 [36, 37]. Based on this pioneering work, we extracted the MCP information of all of the simulated membrane proteins as the training dataset (termed 'MCP-Large') for our DCRNN model (please refer to the "Materials and methods" section and Fig 1 for details). With this model, we were able to predict the MCP for a given sequence to good accuracy. Please note that, while the training dataset was obtained from MD simulations for the membrane proteins with a known structure, our prediction model does not require any structural information. A protein sequence is all that is needed for the MCP prediction.

As shown in Table 1, the overall Pearson Correlation Coefficient (PCC) between our MCP prediction and the MD observation (obtained from MemProtMD, which can be viewed as the ground truth here) 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. Ideally, it would be better to use datasets of lower sequence similarity (<25%). However, due to the limited membrane protein data available, using a low-sequence-similarity dataset would lead to the under-training problem. As a test, we trained an MCP 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% (termed 'MCP-Small'). The prediction accuracy was still satisfactory with an overall PCC of about 0.65 (S2 Table), but significantly dropped compared to with the larger dataset. Therefore, in this work, we utilized the MCP predictor trained by the MCP-Large dataset, which made the method less *De Novo* but generating more accurate predictions. As a comparison, the highest prediction accuracy of SA is around 80% at present [18], after many years of development with a much larger training dataset.

We analyzed the prediction performance for the α-helix and β-sheet structures, two of the most common secondary structures of transmembrane proteins. The α/β PCC was calculated

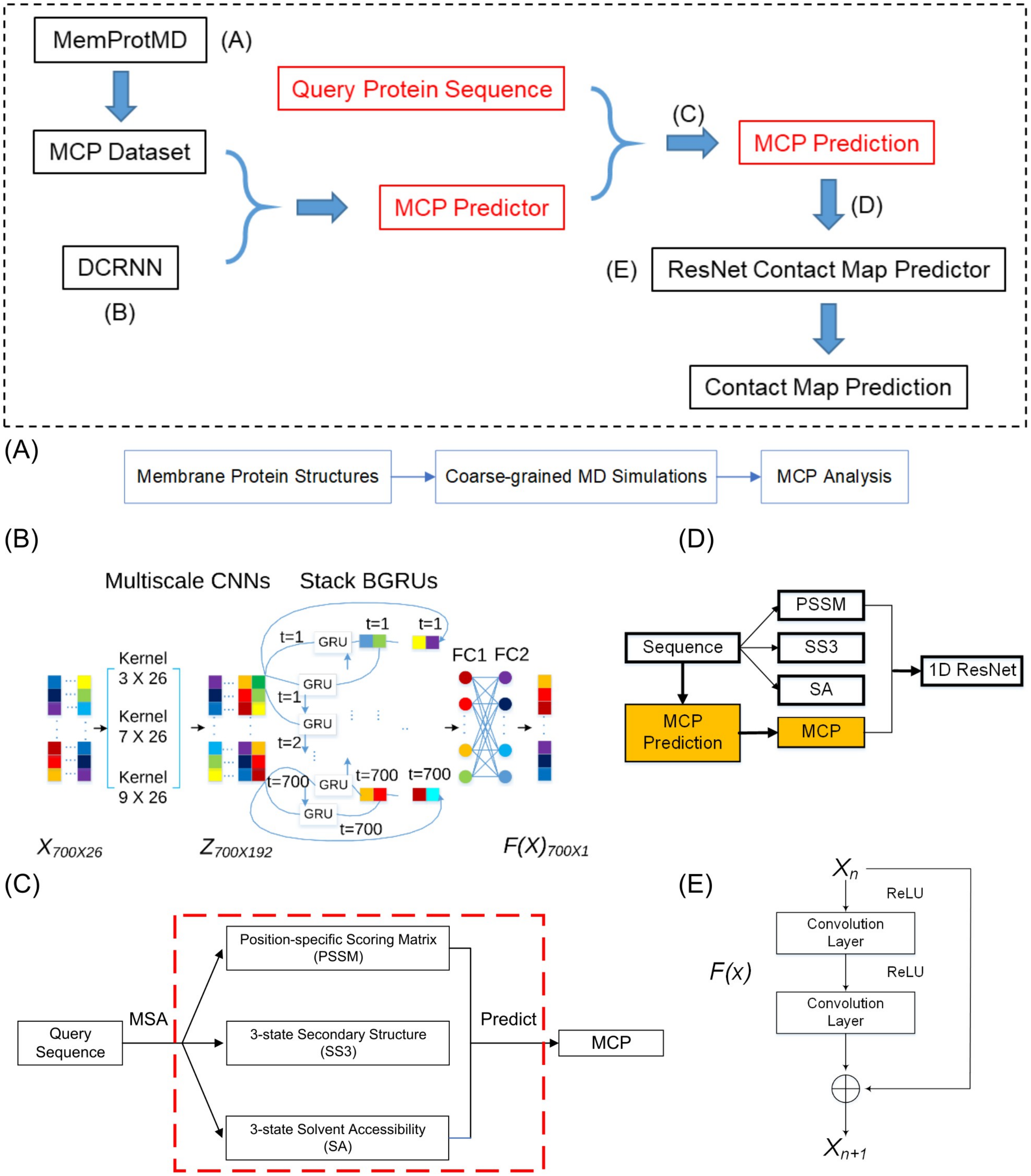

Fig 1. The method schemes. The overall scheme is presented in the black dashed rectangular in the top panel, and the key steps are described in: (A) Extraction of MCP from the MemProtMD database. (B) The DCRNN model for the MCP prediction. (C) The process of MCP prediction from a query sequence. (D) The method of MCP incorporated into the ResNet model. The MCP was used as a 1D input in the same way as SA. (E) The unit of the ResNet model. Each unit of the ResNet model contains two convolution layers and two activation layers.

https://doi.org/10.1371/journal.pcbi.1009972.g001

Table 1. The performance of our MCP predictor.

Evaluation	Training	Validation	Test
Overall			
MSE	0.047	0.050	0.048
PCC	0.774	0.764	0.769
α–helix (H)			
MSE	0.053	0.057	0.056
PCC	0.848	0.832	0.838
β–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

https://doi.org/10.1371/journal.pcbi.1009972.t001

according to the MCP values of each residue within the corresponding secondary structure types. As shown in Table 1, the prediction accuracy is better for the α-helix structures (PCC = 0.84 for the test) than for the β-sheet structures (PCC = 0.71), 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 (PCC = 0.58), as these are the most flexible and less abundant (23%) structures in the datasets. Further analysis showed that the predictor performs more reliably for multi-pass α-helix proteins than single-pass ones (S3 Table), which may also be related to the amount of proteins of different classes in the datasets (S4 Table).

We picked an α-helix and a β-sheet membrane proteins from the test set as representative cases to investigate the prediction details (Fig 2). As can be seen, most of the hydrophobic lipid-contacting amino acids were successfully predicted (Fig 2A and 2B), with the overall prediction PCCs of 0.76 and 0.71 for the α-helix and β-sheet membrane proteins, respectively. We mapped the MCP onto the surface of the protein structures, and it is evident that the distribution of high MCP values is indeed in the transmembrane region, consistent with the MD observations (Fig 2C–2H). For the α-helix protein (5aym), most of the residues contacting the hydrophobic core of the lipid bilayer were identified (Fig 2A and 2G). For the β-sheet protein (4e1t), there were some lipid-contacting residues missing in our prediction, as shown with the dashed circles in Fig 2H. However, the predicted high-MCP residues are mostly in the transmembrane region and on the outer surface (S1 Fig), indicating that the MCP prediction can reach a satisfactory accuracy in characterizing the membrane hydrophobic core-contacting residues. The results obtained from the MCP predictor trained by the MCP-Small dataset were similar (S2 Fig).

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

For the two 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 2I and 2J). As can be seen, the SA prediction was 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-MCP region (red-colored surface) are complementary to each other on

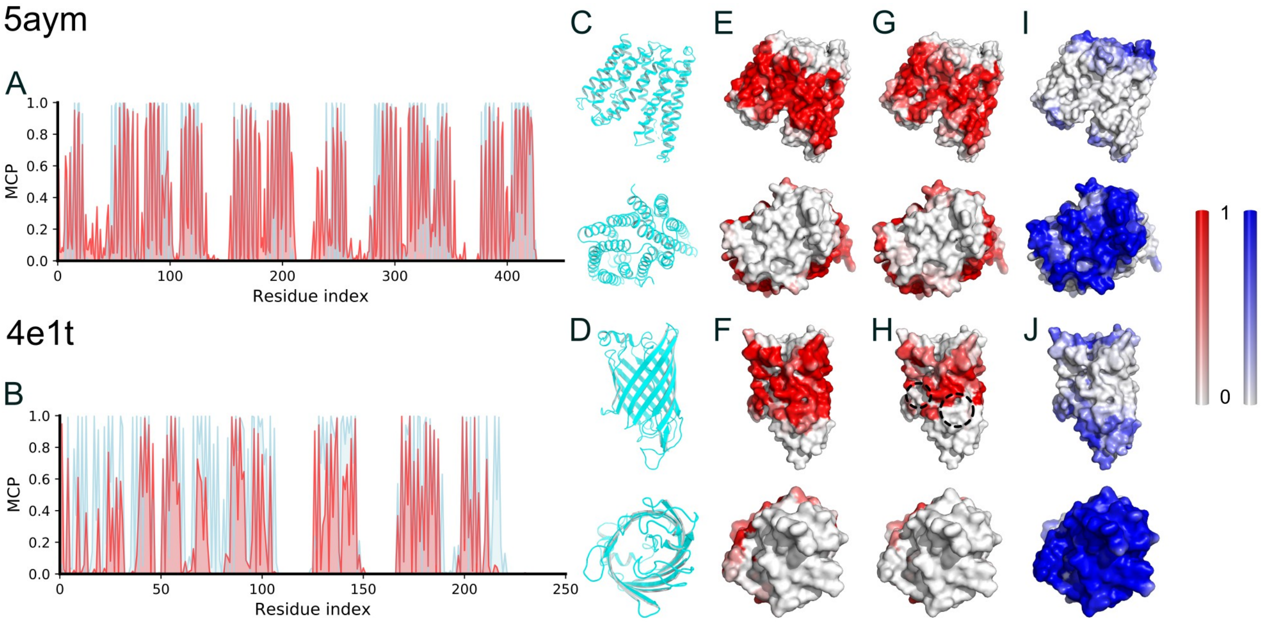

Fig 2. Membrane contact probability (MCP) of two representative membrane proteins. (A-B) Comparison between the MD observation (cyan) and the prediction (red) of the MCPs. (C-D), Side and top views of the two representative proteins. (E-F), The outer surface of the representative membrane proteins, colored according to the MCP values obtained from MD simulations. (G-H), Similar to (E-F), but colored according to the predicted MCP values. (I-J), Similar to (E-F), but colored according to the predicted SA values by RaptorX. The PDB ID of the α-helix membrane protein was 5aym [39], 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 PDB ID of the β-sheet membrane protein was 4e1t [40], an X-ray crystal structure of the transmembrane beta-domain from invasion from *Yersinia pseudotuberculosis* with 245 residues.

https://doi.org/10.1371/journal.pcbi.1009972.g002

the outer surfaces of the two proteins, and together they give a complete description of the outer surface of the membrane protein structures (Fig 2E–2J).

To give a more comprehensive analysis, we plotted the predicted MCP and SA of two datasets in Fig 3. Fig 3A 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 (327 test proteins in total). 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 MCP 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 usually the logic of many SA predictors. However, from Fig 3A, 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 MCP 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 seems puzzling: are there residues that have a high probability to be exposed to both water and lipid molecules? Further analysis using the OPM server [41] 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 (S3 Fig). 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

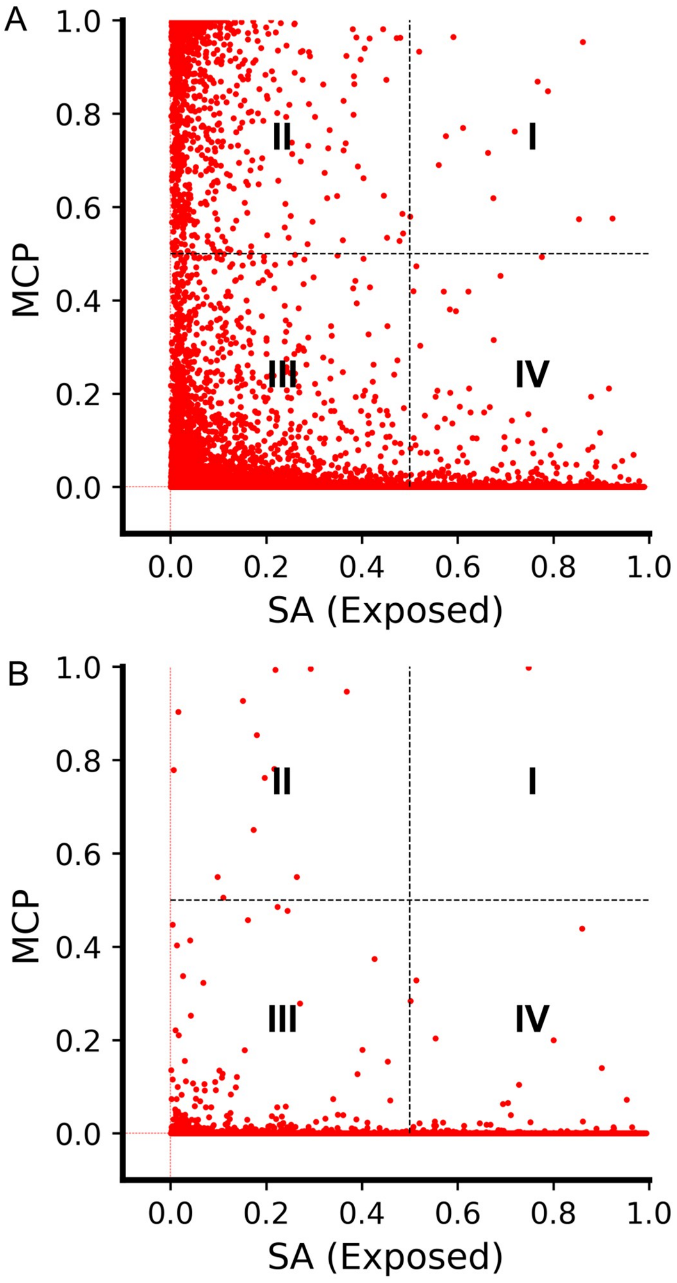

Fig 3. The 2D plot of the complementary MCP and SA predicting the likely location of amino acids in a protein, for the 327-protein dataset (A) and 102 Pfam protein dataset (B), respectively. The figures are roughly divided into four regions: (I) Both the MCP and SA values are larger than 0.5, indicating the amino acids in this region are likely to be exposed to both water and lipid molecules; (II) MCP >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) MCP <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) MCP <0.5 and SA >0.5, meaning the amino acids are more likely to be exposed to water molecules than to the lipid bilayer.

https://doi.org/10.1371/journal.pcbi.1009972.g003

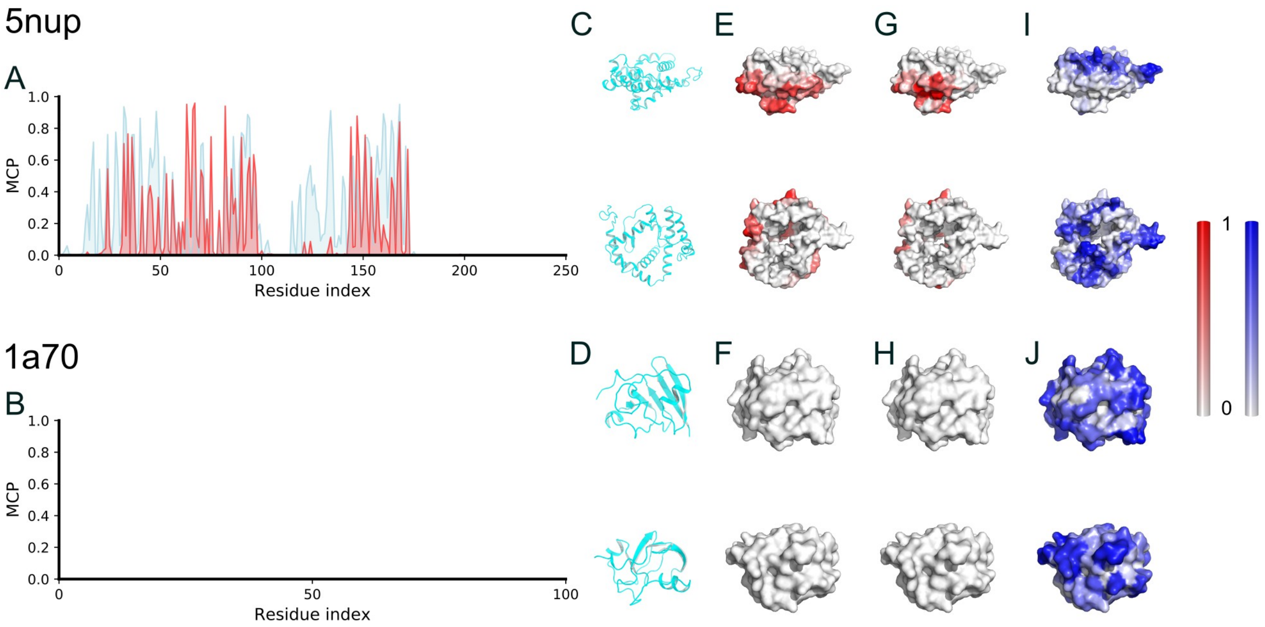

Fig 4. Membrane contact probability (MCP) of two representative non-transmembrane proteins. (A-B) Comparison between the MD observation (cyan) and the prediction (red) of the MCPs. (C-D), Side and top views of the two representative proteins. (E-F), The outer surface of the representative proteins, colored according to the MCP values obtained from MD simulations. (G-H), Similar to (E-F), but colored according to the predicted MCP values. (I-J), Similar to (E-F), but colored according to the predicted SA values by RaptorX. The PDB ID of the half-membrane-embedded protein was 5nup, which is an X-ray crystal structure of the Gram-negative bacterial α-helical outer membrane (OM) protein composed of 236 residues. The PDB ID of the soluble protein was 1a70, which is an X-ray crystal structure of ferredoxin I (Fd I) from *Spinacia oleracea* with 97 residues.

https://doi.org/10.1371/journal.pcbi.1009972.g004

amino acids in region I only account for 0.018% (14/79796) of the total amount of the amino acids in the dataset.

The MCP predictor performed well for non-transmembrane proteins too

We also examined how the predictor performs for non-transmembrane proteins. We picked another two representative proteins from the test dataset: one is a half-membrane-embedded protein, and the other is a soluble protein. As can be seen in Fig 4, most of the membrane-contacting residues of the half-membrane-embedded protein were correctly predicted, and none of the soluble protein residues were predicted to be membrane-contacting. These results indicate that our predictor may have a good performance for non-transmembrane proteins too.

As a further test, we conducted the MCP prediction for a soluble protein dataset, composed of 102 Pfam proteins after removing the sequence redundancy with respect to our training set, which was previously used as a test set for contact map and protein structure prediction [6]. The results are shown in Fig 3B. As can be seen, most of the amino acids were predicted to be not lipid-exposed (MCP <0.5), which was expected since the dataset was supposed to contain soluble proteins only. 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 unexpected.

These 14 amino acids were distributed in seven proteins (S5 Table). Considering our average prediction performance was around 70% (PCC), we ruled out the cases in which there were less than three high-MCP predictions in ten successive amino acids of a protein, considering them outliers, and then there was only one protein left, whose sequence corresponds to a crystal structure of the Sar1-GDP complex [42]. 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 (S5 Table), located on the N-terminal helix of the protein Sar1. 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. Therefore, we ran multiscale MD simulations to check whether this soluble protein would have direct contact with the hydrophobic core of a lipid bilayer. Indeed, our self-assembly CG MD and atomistic MD simulation results demonstrated that this protein would be stably anchored onto a membrane's surface (Fig 5A–5C), and both the coarse-grained and atomistic simulations confirmed a stable binding interface (Fig 5D–5G). In fact, previous experimental studies also showed that Sar1 is responsible for membrane trafficking, and its N-terminal helix probably serves as a wedge that inserts into the outer leaflet of the endoplasmic reticulum (ER) membrane and regulates the membrane curvature and fission [43]. Therefore, the 'abnormal' MCP 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.

Following Stansfeld's protein-lipid contacting definition [44], 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. The MCP values obtained from our MD simulations and DCRNN prediction were not completely identical (Fig 5H and 5I), but the comparison demonstrated that they overlapped at residues M1, F3, G11, F12 and F18 (the purple regions in Fig 5H), showing a converged membrane-interacting interface. Therefore, it appeared that the MCP predictor was good for the soluble proteins too, and perhaps could be used to identify the membrane-interacting residues of soluble or membrane-anchored proteins.

As a comparison, we used other software of relevant functions to conduct predictions for this protein (S4 Fig). From the result of BCL::Jufo9D [45], a server for the prediction of transmembrane span, the N-terminal helix of the protein Sar1 was predicted to be more likely a transition region (TR in S4 Fig), which is somewhat consistent with our results, but less quantitative and less specific when compared to the MCP prediction and the MD observations. The N-terminal helix was not predicted to be membrane-spanning by the transmembrane topology predictors OCTOPUS [32] or TMHMM [30] (S4 Fig).

MCP can improve the prediction precision of protein contact maps and structures

SA has long been used as a fundamental input for protein structure prediction. As demonstrated, MCP is an essential character of amino acids and complementary to SA in describing the outer surface of membrane proteins, so it is natural to think that the MCP would be helpful for protein structure prediction, too. Notably, the prediction precision of protein contact map was hugely improved in the last couple of years, especially since the introduction of the ResNet into the field by the XU group [6, 46]. 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 validate the usefulness of the MCP, we incorporated the MCP into the ResNet the same way as SA, considering the MCP as a 1D input in parallel with SA (please refer to the "Materials and methods" section and Fig 1C and 1D for details), and then we checked whether the contact map prediction can be improved.

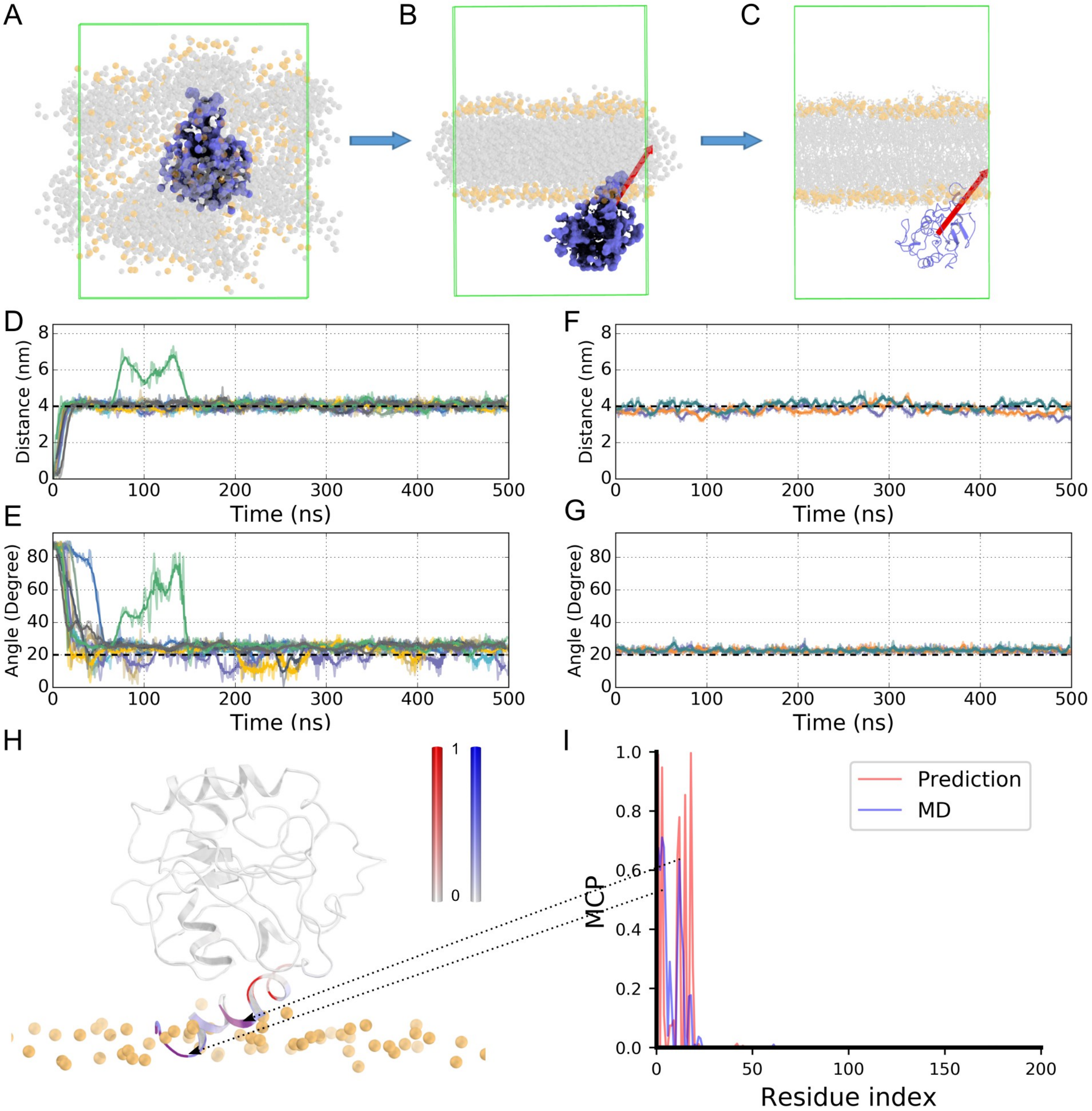

Fig 5. The Sar1 structure (1f6b) anchored on a membrane. (A) The initial and (B) final structures of the 1f6b system in the coarse-grained self-assembly MD simulations. (C) The atomistic structure of 1f6b transformed from the CG MD simulation outcome (B). The orange and grey spheres represent the lipid 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. (D) The distance between the COM of the protein and the COM of the bilayer in the nine CG simulation trajectories. (E) 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. (F) Similar to (D), (G) similar to (E), but obtained from three atomistic MD simulations. (H) The orientation of 1f6b bound to the lipid bilayer obtained from MD simulations, with the membrane-interacting residues colored according to panel (I). (I) The overlaid MCP generated by our MCP predictor (red) and MD simulations (blue).

https://doi.org/10.1371/journal.pcbi.1009972.g005

We took the above dataset containing 327 test proteins as our prediction target, and we compared the prediction results of our MCP-incorporated ResNet contact map 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 MCP in the ResNet predictor systematically improved the

Table 2. The overall contact prediction precision 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+MCP	0.78	0.65	0.41	0.24	0.85	0.75	0.53	0.33	**0.91**	**0.89**	**0.80**	**0.66**

https://doi.org/10.1371/journal.pcbi.1009972.t002

prediction precision. On average, the prediction precision can be improved by about 3%, if only the top L/k predictions are considered. The results showed that the MCP 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 precision for the whole contact maps, we calculated the PCC between the predictions and the native contact maps calculated from known protein structures of the whole dataset. The PCC is about 0.29 with the original ResNet predictor and about 0.37 with our MCP-incorporated ResNet predictor (S5 Fig). Thus, the relative improvement of the prediction precision is about 28% with the incorporation of the MCP into the ResNet model, when the whole contact map is considered.

As further validation, we calculated the prediction precision for a dataset composed of 495 test proteins, which was a subset of PDB25 [47] created in May 2020 (25% non-redundant sequences with a 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 MCP incorporated showed a consistently higher precision (S6 Table). We further separated the dataset into two subsets, the soluble proteins and membrane proteins, and we calculated the prediction precision for each subset. As can be seen in S6 Table, the prediction precision 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 prediction rather than the top predictions of the contact maps, we can see that the MCP 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 MCP-incorporated ResNet predictor when compared to the native contact maps of all the membrane protein in the dataset. Thus, the relative PCC improvement for membrane proteins was 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 was around 20%. Although the overall PCCs are still low for the entire contact maps, these results present a clear improvement in the contact map prediction when the MCP is incorporated into the ResNet model, and 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 MCP in the ResNet model and compared to the results from other CM prediction tools [48–51] (S7 Table). It is obvious that the MCP-incorporated ResNet predictor performed the best. In Fig 6, we show the native contact maps, and the ResNet predictions with and without using the MCP information for the two membrane proteins, respectively. The prediction with the MCP is better correlated with the native CMs with a PCC of 0.53 and 0.69, while the prediction without the MCP showed a PCC of about 0.35 and 0.62 compared to the native CMs, respectively. Looking at the six panels of Fig 6, one can immediately recognize that the incorporation of MCP removed many false positives compared to the model without MCP. As a result, the top L/5 long-range

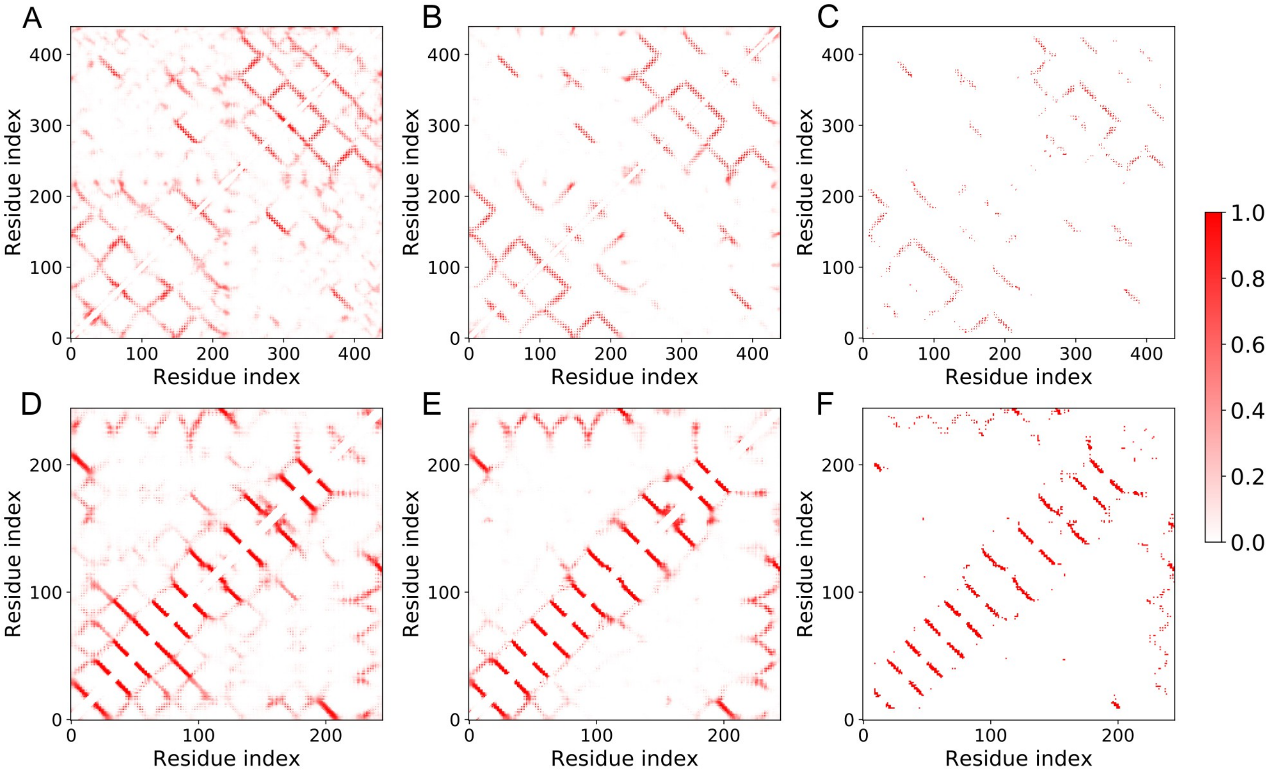

Fig 6. Incorporation of the predicted MCP improved the CM prediction for the two representative cases. The cutoff for the CM prediction was 8 Å between β carbon atoms. (A), (B) The predicted CMs for 5aym without and with MCP incorporated, respectively. (C) The native CM calculated from the known structure 5aym. (D-F) Similar to (A-C), but for the protein 4e1t.

https://doi.org/10.1371/journal.pcbi.1009972.g006

contact prediction precision was improved by 10% and 6% for the two cases, respectively (S7 Table).

Based on the predicted contact maps, we further predicted the structures of the above two representative proteins using CONFOLD2 [52]. From the top-five models, it is clear that the prediction with the MCP-incorporated contact maps yield consistently better results than with the contact maps without MCP incorporated (S8 Table). The best-predicted structures are shown in Fig 7, where the predicted structures using the MCP-incorporated contact maps are closer to the crystal structures with RMSDs of 3.25 and 3.41 Å, while the predictions without MCP have RMSDs of 4.55 and 4.12 Å compared to the crystal structures, respectively.

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 [53]. The widely-accepted quantity describing which amino acids are more likely to be exposed to water is 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

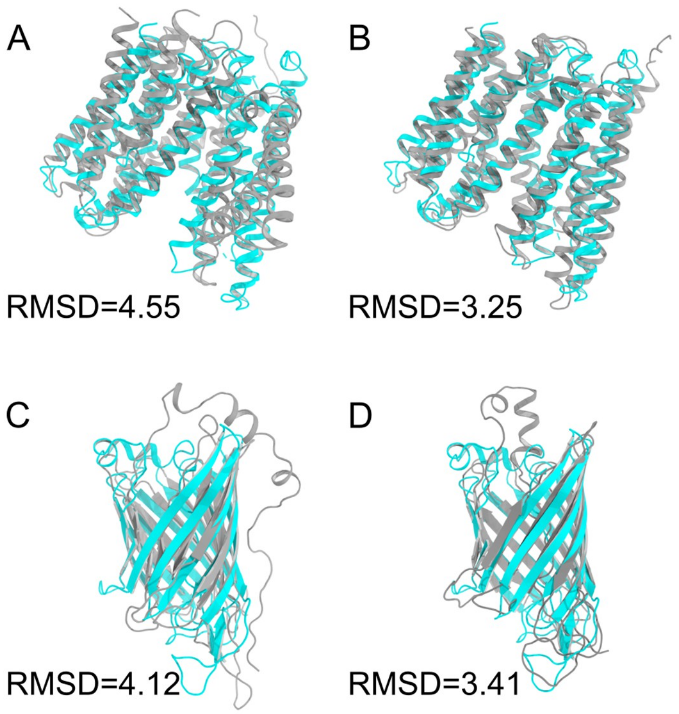

Fig 7. The MCP-incorporated CM predictions improved the protein structure prediction accuracy for the two representative cases. (A), (B) The predicted protein structures for 5aym using the predicted CMs without and with MCP incorporated, respectively. The crystal structure was shown in cyan and the predicted structures were shown in grey. (C-D) Similar to (A-B), but for the protein structure 4e1t.

https://doi.org/10.1371/journal.pcbi.1009972.g007

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 membrane contact probability of amino acids is highly valuable for membrane protein studies.

Lipid molecules are much larger and more complex than water molecules, introducing more uncertainties to the MCP calculation if the same roll-a-sphere protocol is adopted as in the SA calculations. 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 [54]. Therefore, here, we propose a new method to extract the MCP 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 MCP 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 MCP.

Hydrophobic scale is widely used to characterize the hydrophobic property of amino acids, and thus can also be used to characterize/color the outer surface of a protein. We compared the MCP prediction to the Wimley-White hydrophobicity scale [55] to check whether MCP has any advantages over simple hydrophobicity scales. As shown in S6 Fig, it is obvious that our MCP prediction gives a much more clear representation of the membrane-interacting residues than the simple hydrophobicity scale for both membrane proteins and soluble proteins. The MCP prediction also agrees much better with the MD observations than the simple hydrophobicity scale, indicating that MCP is indeed a more appropriate quantity in characterizing the outer surface of a (membrane) protein.

As shown in Fig 3, the MCP 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 MCP is beneficial to the structural analysis and prediction of proteins for two reasons. First, the incorporation of the MCP 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 MCP can increase the true positives for the hydrophobic residues embedded inside and the hydrophilic residues exposed to the outside. Therefore, when we trained the MCP-incorporated ResNet model for the contact map prediction, the precision for both the soluble and membrane proteins was improved. As a natural result, we show that MCP is helpful for 3D protein structure prediction as well (Fig 7 and S8 Table). In addition to improving contact maps, which is helpful for 3D structure prediction, we believe MCP can also be directly used for 3D structure prediction and evaluation. For example, one can apply extra constraints to push the high-MCP residues onto the protein surface during the 3D structure modeling of membrane proteins, or take the MCP into account when scoring the modeled protein structures.

As only the single-chain information was used to construct the dataset, one might wonder whether the model works for oligomeric complexes too. Our analysis showed that: 1) The model performs better for single-chain proteins than for complexes (S9 Table). 2) For a single chain within a complex, the protein-membrane and protein-protein interfaces can be distinguished from the predicted MCP (S7 Fig), where the protein-protein interface residues showed lower MCP predictions than those of the protein-membrane interface. However, one cannot tell where the protein-protein interface is from the MCP alone, as the low-MCP region may be embedded in the protein interior as well. 3) Similar to above, for a single-chain protein and a monomeric chain of a complex with similar fold but low sequence similarity, the predicted MCP can distinguish them (S8 Fig). Therefore, it appears that the MCP predictor works reasonably well for oligomeric transmembrane complexes too.

Another potential application of MCP is to determine the correct orientation of membrane proteins of known structure when embedded into membranes, similar to OPM and MemEmbed [41, 56, 57], for example. The predicted MCP can also help to position a membrane-anchored protein in close proximity to a membrane with the interface residues facing the membrane surface, which would be useful for setting up structural models for further studies such as molecular dynamics simulations. We reserve to explore these potential applications in future studies.

There are several known limitations of the method presented in this work: 1) To achieve a better performance, the datasets contain redundant sequences, which may add bias to the model and makes it not really *De Novo*. Unfortunately, with the limited amount of membrane protein structures, it is difficult to use a low-sequence-similarity dataset to achieve the best prediction performance for now. 2) The prediction for the random coiled structures is not very reliable, which is natural considering these are the most flexible and least abundant structures in membrane proteins. 3) In the datasets extracted from MemProtMD, there was only one type of lipid molecules, 1,2-dihexadecanoyl-rac-glycero-3-phosphocholine (DPPC), making it impossible to identify lipid-type-specific interactions. However, the analysis showed that a membrane protein has very similar membrane contacts in spite of the lipid molecule types in the membrane (S9 Fig), so as far as the hydrophobic core-contacting information is concerned, the prediction should be satisfactory. 4) The highly probable lipid-interacting sites can be predicted, but one can not distinguish whether it is a specific lipid binding or a non-specific binding site from the prediction alone, as the training dataset does not contain such information, so the method cannot be used to identify specific lipid-binding sites. To overcome these problems, much larger datasets with more rich information, such as the distribution of diverse lipids and the lipid residence time around membrane proteins, would be required.

In summary, we propose that the MCP 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 MCP and other similar properties of proteins. We believe that the MCP would be able to find a wide range of applications in various aspects of the structure and function studies of membrane-interacting proteins.

Materials and methods

MCP prediction

Extraction of the MCP information from the MemProtMD database. Our aim was to predict the MCP efficiently and accurately from any given protein sequence in the absence of structural information, which could then be used for structural and functional analysis of the protein. To do this, we needed a large dataset of the MCP of membrane proteins for the training using machine learning-based methods. It is difficult to obtain the MCP 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 [58–60]. With MD simulations, we can get dynamic and quantitative information regarding the protein and membrane 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 [36, 37, 44], 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 [58–60]. 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. Such a cutoff value was shown to be appropriate for discriminating the transmembrane region from water-exposed regions of membrane proteins [58, 61] (S10 Fig). 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 3,500 MD simulation results in the web database and the number continues to count (http://memprotmd.bioch.ox.ac.uk/) [37]. We downloaded the PDB files of the atomistic structure with the acyl tail contact probability from the MemProtMD database. In the PDB files, the temperature factor value (also called the B-factor) of each atom was replaced by the membrane contact probability obtained from MD analysis. Therefore, we were able to extract the membrane contact probability value of each C_α as our MCP observation for each residue. This procedure is shown in Fig 1A.

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.

During the dataset generation, the multi-chain proteins were split into single-chain sequences, and only one chain of the same sequence was adopted into the dataset. We obtained 12691 result files, removed duplicate sequences, and extracted 5900 of them at random as the membrane protein dataset in the end.

Generation of the dataset for the MCP model training. The values of MCP 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 MCP prediction, we also included soluble proteins in the dataset. 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 of the PDB25 dataset with less than a 25% sequence identity [47]. Then, we set the MCP of each soluble protein residue to be zero. In addition, we divided the remaining 900 membrane protein sequences into 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. The above dataset was termed 'MCP-Large' in this work.

We also constructed a dataset 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 of membrane proteins, resulting in 898 sequences. This smaller dataset was termed 'MCP-Small'. To ensure that the training set is relatively sufficient [24, 62], 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. Such a ~8:1:1 dataset construction ensures a relatively larger training set with reasonable validation and test sets for better convergence, which was adopted and recommended by previous work [63, 64].

The DCRNN model for the MCP prediction. We considered the MCP 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 [35]. 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) [65] and a backward GRU. In the model, we also combined multiscale CNN layers for the local context extraction.

As illustrated in Fig 1B, we can see an overview of the model for the MCP prediction. We defined the loss function as the residual sum of the squares between the values from the prediction and the MD observation plus L_2 norms of the model parameters.

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, which were concatenated to be a 1×26 array for each residue. For a protein with a sequence length of L, the resulting matrix had a dimension of $L \times 26$, which was padded with zeros to be a 700×26 matrix for the sequences shorter than 700. Then the matrix was operated by a sliding window with the kernel size of $k \times 26$ ($k = 3, 7, 9$) and a channel size of 64, as shown in Fig 1B. For each protein sequence, we ran HHblits 3.0.3 [66] (with E-value 0.001 and

3 iterations) to search the uniclust30 database dated October 2017 to find its sequence homologous and then built its multiple sequence alignment (MSA). We only consider single chains for the prediction for now.

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 the deep network using the Adam optimizer [67]. We set the batch size to be 1. The training was conducted on a workstation with a six-core (12-thread) Intel Xeon E5–1650 CPU and a GTX 1080 Nvidia GPU. It took around 24 hours to train one model with 200 epochs.

The model reached convergence after 20 epochs of training according to the MSE curves, and we stopped the training at 100 epochs when no sign of over-training was observed (S11 Fig). The 10-fold validation results are shown in S10 and S11 Tables.

Protein contact map prediction

The dataset for the protein contact map prediction. Our training data were a subset of PDB25 [47] 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 did not manually balance the soluble and membrane proteins during training, as previous studies showed that ResNet works for membrane proteins even if the training set only contained a small fraction of membrane proteins [46], meaning the learning is quite transferable. The basic statistics of the dataset are shown in S12 Table according to the database SCOPe [68, 69].

The ResNet model for the protein contact map prediction. The deep residual net (ResNet) has been widely used for image recognition and won first place on the ILSVRC 2015 classification task [70]. Xu et al. developed and utilized the ResNet for the contact map prediction of proteins and won first place on CASP 12 and CASP 13 (RaptorX). To test if the MCP is useful for the structural prediction of membrane proteins, we integrated the MCP into the ResNet model created by Xu et al. [6] and checked whether the prediction performance could be improved. The revised model contained two residual neural networks. Fig 1E 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 implementation 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 MCP.

We used similar input features as the original ResNet model [6], plus the additional MCP predicted from the protein sequence in parallel to SA as a 1D input (Fig 1D). In addition to the MCP predicted by our model, the input features included the PSSM, SS3, and SA generated by RaptorX-Property [4], as well as the evolutionary coupling (EC) information generated by CCMpred [48], pairwise potential, and mutual information generated by alnstats [71]. The pairwise potential is computed by averaging contact potential terms [72, 73] across the two alignment columns, derived from the MSA. The mutual information is calculated between positions of two different protein families in a joint alignment of sequences from the same set of organisms [74]. 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, a cutoff value widely accepted in the field. Therefore, the output of the contact map prediction was a matrix showing the probability of two residues within 8 Å.

We then evaluated the prediction precision of the top L/k (k = 10,5,2,1) predicted contacts, where L is the protein sequence length, by comparing the predictions with the native contacts calculated from known protein structures. The prediction precision 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.

For the MCP-incorporated ResNet CM predictor, we compared the CM prediction results using the MCP predictors trained by the MCP-Large dataset and the MCP-Small dataset. In the end, we used the MCP-Large predictor for the CM prediction 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 a batch size of 1. The training was conducted with a Tesla V100 Nvidia GPU with 32 GB of GPU memory, on which it takes around 40 hours to train a model with 20 epochs.

The model reached convergence after 20 epochs of training according to the precision curves (S12 Fig). The 10-fold validation results are shown in S13 Table. The normalized confusion matrix, the prediction accuracy, and the area under the curve (AUC) are shown in S14 and S15 Tables, respectively. A cutoff of 0.5 was used for the confusion matrix and prediction accuracy calculation. Although these are not commonly used to evaluate the model performance in the CM predictions as the CM matrices are highly sparse, they show that the MCP-incorporated CM predictor consistently outperforms the ResNet model without MCP.

Contact-driven protein structure prediction

With the predicted contact maps and the predicted three-state secondary structures [75], we can build protein structure models of a query sequence using CONFOLD2 [52]. With the scripts in the CONFOLD2 package, we converted the predicted secondary structures to distance, dihedral, and hydrogen bond restraints. We used the top-xL contacts as contact distance restraints, where x = 0.1, 0.2, 0.3, . . ., 4.0 and L is the length of the protein. For each predicted contact, the distance of the two corresponding residues was set in the range from 3.5 Å to 8 Å. Then we fed the processed data to the Crystallography & NMR System (CNS) [76] for model construction. For each x value, 20 structure models were constructed, and then the top-five models in each subset were selected according to the contact energy score [52], resulting in 200 models in total. Then, we ranked these 200 models using the satisfaction score according to their top L/5 long-range contacts [52]. We selected the top-50 models and clustered them into five subsets with the pairwise TM-score [77]. Finally, we selected the model closest to the centroid of each cluster and obtained five top models. Then we calculated the RMSD value of each model with regard to the crystal structure (S8 Table). In the above processes, the contact maps (with or without MCP incorporated) were the only difference for the model construction.

Molecular dynamics simulations

We used MD simulations to validate whether the predicted soluble protein with multiple high-MCP amino acids was indeed interacting with and anchored into the membrane. The atomistic protein structure was downloaded from Protein Data Bank (PDB ID: 1f6b). 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 [78]. 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 [79]. The CG structure and topology files were generated with the script *martinize.py* [80, 81]. Following Stansfeld's protocol when generating the MemProtMD database [36], 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 5A). After a tens-of-nanoseconds simulation, the lipid molecules formed a bilayer spontaneously, and the protein found its most stable orientation (Fig 5B). 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) [82, 83], 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.

Then the script *backward.py* [84] was used to transform the equilibrated coarse-grained system (Fig 5B) to all-atom system, which was utilized as the initial system (Fig 5C) 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 [83] 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 [85]. After the system was equilibrated to the desired condition, we removed the position restraints and performed three 500-ns all-atom MD simulations 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 [82, 86] 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 [87] in combination with the *Slipids* force field [88, 89].

Supporting information

S1 Fig. Comparison of the MCP prediction and the exposure (indicated by the relative solvent accessibility calculated by DSSP) for the two representative membrane proteins. The

MCP and outer exposure results are shown as red and teal lines, respectively. According to DSSP, the buried residues are defined to have an RSA value of 0-0.1, so the outer surface residues have a RSA values of 0.1-1. We calculated the percentage of membrane-contacting residues with high MCP values (>0.5) lying in the outer residues with RSA >0.1. The value is 84.0% for 5aym, and 88.9% for 4e1t. Therefore, most of the membrane-contacting residues predicted by MCP are outer-surface residues.
(TIF)

S2 Fig. Membrane contact probability (MCP) of four representative proteins with the predictor trained by the MCP-Small dataset. (A-D) Comparison between the observation (cyan) and the prediction (red) of the MCPs. (E-H), Side and top views of the four representative proteins. (I-L), The outer surface of the representative proteins, colored according to the observed MCP values obtained from MD simulations. (M-P), Similar to (I-L), but colored according to the predicted MCP values. (Q-T), Similar to (I-L), but colored according to the predicted SA values by RaptorX.
(TIF)

S3 Fig. The structures of the membrane proteins showing where the residues in region I of Fig 3A are. The hydrophobic boundaries of the lipid bilayer are represented by the red and blue pseudo-atoms, indicating the outer and inner surfaces of the bilayer, respectively.
(TIF)

S4 Fig. The prediction results of BCL::Jufo9D, OCTOPUS, and TMHMM for the Sar1 sequence. (A) The results of BCL::Jufo9D (red for membrane core (MC), sky blue for transition region (TR), and orange for solution (SO)). (B) The results of OCTOPUS (red for membrane, blue for inside, and fuchsia for outside). (C) The results of TMHMM (red for membrane, blue for inside, and fuchsia for outside).
(TIF)

S5 Fig. Comparison of the PCCs between the predictions and the native contact maps for the original ResNet predictor and our MCP-incorporated ResNet predictor. The original ResNet predictor (X-axis) vs our MCP-incorporated ResNet predictor (Y-axis) for the 327-protein dataset (A) and 495-protein dataset (B), respectively. The dashed line is the function $y = x$. Each point represents a test protein, with red points for membrane proteins.
(TIF)

S6 Fig. The protein surfaces colored according to the MCP and the Wimley-White hydrophobicity scales. (A-D), The colored outer surfaces of the four representative proteins presented in Figs 2 and 4, according to the observed MCP values from MD simulations. (E-H), Similar to (A-D), but colored according to the predicted MCP values. (I-L), Similar to (A-D), but colored according to the value of the Wimley-White hydrophobicity scales. (M-P), Similar to (I-L), but with a different color bar.
(TIF)

S7 Fig. The MCP prediction results for two membrane-embedded complexes (PDB IDs: 5mkk and 2pno). The two complex proteins were in the test dataset with the overall prediction PCCs of 0.84 and 0.61, representing one of the good and one of the poor predictions, respectively. (A-B) Comparison between the observation (light blue) and the prediction (red) of MCP. The horizontal bars on the top of the panels indicate the regions of the protein-membrane (red) and protein-protein interfaces (gray). (C-D) The outer surface of the two proteins colored according to the predicted MCP values. (E-F) Similar to (C-D), but from another view. As can be seen, the protein-protein interface residues (gray bar) overall show low MCP

values than those at the protein-membrane interfaces (red bar). The protein-membrane interface residues were defined by MCP >0.2, while the protein-protein interface residues were defined with the script 'InterfaceResidues' of Pymol.
(TIF)

S8 Fig. The MCP prediction results for two membrane proteins with similar folds but different oligo states, salmon for the monomeric protein (PDB ID: 5o0t) and olive for the oligomeric protein (PDB ID: 5ys3). (A) The structure alignment of the two proteins showed a similar fold (TM-score = 0.45, normalized by the length of 5ys3). (B) The sequence alignment of the two proteins (sequence similarity = 20.9%, calculated by MUSCLE). (C-D) The outer surface of the two proteins colored according to the predicted MCP values. (E-F) Similar to (C-D), but from another side view. (G-H) Similar to (C-D), but from the top view.(I-J) Comparison between the MD observation (light blue) and the prediction (red) of MCP. The horizontal bars on the top of the panels indicate the regions of the protein-membrane (red) and protein-protein interfaces (gray). As can be seen, the protein-protein interface residues (gray bar) show overall low MCP values than those at the protein-membrane interfaces (red bar) in the transmembrane region. The protein-membrane interface residues were defined by MCP >0.2, while the protein-protein interface residues were defined with the script 'InterfaceResidues' of Pymol.
(TIF)

S9 Fig. Comparison of the observed MCP in MD simulations with different types of lipids in membranes: cyan, orange, and purple lines for the membranes composed of DPPC (100%), POPC (100%), and mixed POPE (50%) and POPG (50%), respectively. These results show that the saturation, head group and net charge of the lipids have minor impacts on the observed MCP in the hydrophobic core region, and the differences are mostly located at the membrane-water interfaces.
(TIF)

S10 Fig. Comparison of the results of MCP analysis with different cutoff values. (A-D), The colored outer surface of the protein (PDB ID: 5aym) according to the observed MCP values obtained from MD simulations with different cutoff values of 4, 5, 6, and 8 Å, respectively. As can be seen, a cutoff value smaller than 6 Å would lead to weak signals, while a cutoff value larger than 8 Å would start to overestimate the transmembrane region. (E), Comparison between the observed MCPs with different cutoff values; black, blue, red, and cyan lines for 4, 5, 6, and 8 Å, respectively.
(TIF)

S11 Fig. MSE curves of the MCP predictor in the 10-fold cross-validation. The left panel was obtained with the MCP-Large dataset, and the right panel with the MCP-Small dataset.
(TIF)

S12 Fig. Precision curves for the test procedure of the contact map predictor in the 10-fold cross-validation: blue and orange lines for the results without and with MCP incorporated, respectively.
(TIF)

S1 File. Datasets used in this work. This file contains the datasets for the MCP model and contact map predictor in this work.
(PDF)

S1 Table. The related prediction methods for membrane protein features.
(DOCX)

S2 Table. The performance of the MCP predictor using the MCP-Small dataset.
(DOCX)

S3 Table. The performance of our MCP predictor for different transmembrane protein classes.
(DOCX)

S4 Table. The amount of proteins of different classes in the datasets.
(DOCX)

S5 Table. The amino acids predicted with high MCP values in the 102 Pfam soluble protein dataset.
(DOCX)

S6 Table. The contact map prediction precision of the additional 495-protein dataset: overall, soluble, and membrane proteins.
(DOCX)

S7 Table. The contact map prediction precision for the two representative cases: 5aym and 4e1t.
(DOCX)

S8 Table. The RMSD of the top five structure prediction models with respect to the crystal structures for the two representative cases: 5aym and 4e1t.
(DOCX)

S9 Table. The performance of our MCP predictor for different oligomeric states.
(DOCX)

S10 Table. The performance of our MCP predictor in the 10-fold cross-validation using the MCP-Large dataset.
(DOCX)

S11 Table. The performance of our MCP predictor in the 10-fold cross-validation using the MCP-Small dataset.
(DOCX)

S12 Table. The basic statistics of the datasets according to the database SCOPe.
(DOCX)

S13 Table. The performance (precision of medium- and long-range contact) of our contact map predictor in the 10-fold cross-validation.
(DOCX)

S14 Table. The normalized confusion matrix of the contact map prediction (cutoff = 0.5).
(DOCX)

S15 Table. The accuracy (cutoff = 0.5) and AUC of the contact map prediction.
(DOCX)

Acknowledgments

We thank Xiao Luo and Prof. Minghua Deng at the School of Mathematical Sciences at Peking University, and Prof. Jinbo Xu at Toyota Technological Institute at Chicago, who discussed with us and provided instructive suggestions. Part of the MD simulations was performed on the Computing Platform of the Center for Life Sciences at Peking University.

Author Contributions

Conceptualization: Chen Song.

Data curation: Lei Wang, Jiangguo Zhang, Dali Wang.

Formal analysis: Lei Wang, Jiangguo Zhang, Dali Wang.

Funding acquisition: Chen Song.

Investigation: Lei Wang, Dali Wang, Chen Song.

Methodology: Chen Song.

Project administration: Chen Song.

Resources: Chen Song.

Software: Lei Wang, Jiangguo Zhang.

Supervision: Chen Song.

Validation: Lei Wang, Jiangguo Zhang, Dali Wang, Chen Song.

Visualization: Lei Wang, Jiangguo Zhang, Dali Wang.

Writing – original draft: Lei Wang, Dali Wang, Chen Song.

Writing – review & editing: Chen Song.

References

1. Howarth M. Say it with proteins: an alphabet of crystal structures. Nature Structural &Molecular Biology. 2015; 22:349. https://doi.org/10.1038/nsmb.3011 PMID: 25945881
2. Wang Z, Zhao F, Peng J, Xu J. Protein 8-class secondary structure prediction using Conditional Neural Fields. 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2010; p. 109–114.
3. Wang S, Peng J, Ma J, Xu J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports. 2016; 6:18962. https://doi.org/10.1038/srep18962 PMID: 26752681
4. Wang S, Li W, Liu S, Xu J. RaptorX-Property: a web server for protein structure property prediction. Nucleic Acids Research. 2016; 44:W430–W435. https://doi.org/10.1093/nar/gkw306 PMID: 27112573
5. Zhou Y, Kloczkowski A, Faraggi E, Yang Y, editors. Prediction of Protein Secondary Structure. vol. 1484 of Methods in Molecular Biology. New York, NY: Springer New York; 2017.
6. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Computational Biology. 2017; 13:e1005324. https://doi.org/10.1371/journal.pcbi.1005324 PMID: 28056090
7. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences. 2020; 117:1496–1503. https://doi.org/10.1073/pnas.1914677117 PMID: 31896580
8. Senior A, Evans R, Jumper J, Kirkpatrick JRM, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020; 577:706–710. https://doi.org/10.1038/s41586-019-1923-7 PMID: 31942072
9. Uhlén M, Fagerberg L, Hallström B, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015; 347:1260419. https://doi.org/10.1126/science.1260419 PMID: 25613900

10. Cheng Y. Membrane protein structural biology in the era of single particle cryo-EM. Current Opinion in Structural Biology. 2018; 52:58–63. https://doi.org/10.1016/j.sbi.2018.08.008 PMID: 30219656

11. Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annual Review of Biophysics and Biomolecular Structure. 1996; 25:113–136. https://doi.org/10.1146/annurev.bb.25.060196.000553 PMID: 8800466

12. Flock T, Venkatakrishnan A, Vinothkumar K, Babu M. Deciphering membrane protein structures from protein sequences. Genome Biology. 2012; 13:160. https://doi.org/10.1186/gb-2012-13-6-160 PMID: 22738306

13. Singh A. Deep learning 3D structures. Nature Methods. 2020; 17:249. https://doi.org/10.1038/s41592-020-0779-y PMID: 32132733

14. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021; 373:871–876. https://doi.org/10.1126/science.abj8754 PMID: 34282049

15. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; p. 1–11.

16. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics. 2018; 34:1466–1472. https://doi.org/10.1093/bioinformatics/btx781 PMID: 29228185

17. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics. 2018; 34:4039–4045. PMID: 29931279

18. Zhang B, Li L, Lü Q. Protein Solvent-Accessibility Prediction by a Stacked Deep Bidirectional Recurrent Neural Network. Biomolecules. 2018; 8:33. https://doi.org/10.3390/biom8020033 PMID: 29799510

19. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. https://doi.org/10.1002/bip.360221211 PMID: 6667333

20. Ma J, Wang S. AcconPred: Predicting Solvent Accessibility and Contact Number Simultaneously by a Multitask Learning Framework under the Conditional Neural Fields Model. BioMed Research International. 2015; 2015:678764. https://doi.org/10.1155/2015/678764 PMID: 26339631

21. Beuming T, Weinstein H. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. Bioinformatics. 2004; 20 12:1822–1835. https://doi.org/10.1093/bioinformatics/bth143 PMID: 14988128

22. Yuan Z, Zhang F, Davis M, Bodén M, Teasdale R. Predicting the solvent accessibility of transmembrane residues from protein sequence. Journal of Proteome Research. 2006; 5 5:1063–1070. https://doi.org/10.1021/pr050397b PMID: 16674095

23. Illergård K, Callegari S, Elofsson A. MPRAP: An accessibility predictor for a-helical transmem-brane proteins that performs well inside and outside the membrane. BMC Bioinformatics. 2009; 11:333.

24. Lu C, Liu Z, Kan B, Gong Y, Ma Z, Wang H. TMP-SSurface: A Deep Learning-Based Predictor for Surface Accessibility of Transmembrane Protein Residues. Crystals. 2019; 9(12):640. https://doi.org/10.3390/cryst9120640

25. Leman JK, Lyskov S, Bonneau R. Computing structure-based lipid accessibility of membrane proteins with mp_lipid_acc in RosettaMP. BMC Bioinformatics. 2017; 18:115. https://doi.org/10.1186/s12859-017-1541-z

26. Adamian L, Liang J. Prediction of transmembrane helix orientation in polytopic membrane proteins. BMC Structural Biology. 2006; 6:13. https://doi.org/10.1186/1472-6807-6-13 PMID: 16792816

27. Phatak M, Adamczak R, Cao B, Wagner M, Meller J. Solvent and lipid accessibility prediction as a basis for model quality assessment in soluble and membrane proteins. Current Protein & Peptide Science. 2011; 12 6:563–573. https://doi.org/10.2174/138920311796957603 PMID: 21787302

28. Lai JS, Cheng CW, Lo A, Sung TY, Hsu WL. Lipid exposure prediction enhances the inference of rotational angles of transmembrane helices. BMC Bioinformatics. 2013; 14:304. https://doi.org/10.1186/1471-2105-14-304 PMID: 24112406

29. Nugent T, Jones D. Predicting Transmembrane Helix Packing Arrangements using Residue Contacts and a Force-Directed Algorithm. PLoS Computational Biology. 2010; 6. https://doi.org/10.1371/journal.pcbi.1000714 PMID: 20333233

30. Krogh A, Larsson B, von Heijne G, Sonnhammer E. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of Molecular Biology. 2001; 305 3:567–580. https://doi.org/10.1006/jmbi.2000.4315 PMID: 11152613

31. Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics. 2007; 23(5):538–544. https://doi.org/10.1093/bioinformatics/btl677 PMID: 17237066

32. Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics. 2008; 24 15:1662–1668. https://doi.org/10.1093/bioinformatics/btn221 PMID: 18474507

33. Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Computational Biology. 2008; 4(11):e1000213. https://doi.org/10.1371/journal.pcbi.1000213 PMID: 18989393

34. Feng SH, Zhang WX, Yang J, Yang Y, Shen HB. Topology prediction improvement of *α*-helical transmembrane proteins through helix-tail modeling and multiscale deep learning fusion. Journal of Molecular Biology. 2020; 432(4):1279–1296. https://doi.org/10.1016/j.jmb.2019.12.007 PMID: 31870850

35. Li Z, Yu Y. Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI); 2016.

36. Stansfeld PJ, Goose JE, Caffrey M, Carpenter EP, Parker J, Newstead S, et al. MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. Structure(London, England:1993). 2015; 23:1350–1361. https://doi.org/10.1016/j.str.2015.05.006 PMID: 26073602

37. Newport TD, Sansom M, Stansfeld PJ. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. Nucleic Acids Research. 2019; 47:D390–D397. https://doi.org/10.1093/nar/gky1047 PMID: 30418645

38. Li Y, Zhang C, Bell EW, Yu DJ, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins: Structure. 2019; 87:1082–1091. https://doi.org/10.1002/prot.25798 PMID: 31407406

39. Taniguchi R, Kato H, Font J, Deshpande C, Wada M, Ito K, et al. Outward- and inward-facing structures of a putative bacterial transition-metal transporter with homology to ferroportin. Nature Communications. 2015; 6:8545. https://doi.org/10.1038/ncomms9545 PMID: 26461048

40. Fairman JW, Dautin N, Wójtowicz D, Liu W, Noinaj N, Barnard TJ, et al. Crystal structures of the outer membrane domain of intimin and invasin from enterohemorrhagic E. coli and enteropathogenic Y. pseudotuberculosis. Structure. 2012; 20 7:1233–1243. https://doi.org/10.1016/j.str.2012.04.011 PMID: 22658748

41. Lomize MA, Pogozheva I, Joo H, Mosberg H, Lomize A. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Research. 2012; 40:D370–D376. https://doi.org/10.1093/nar/gkr703 PMID: 21890895

42. Huang M, Weissman J, Béraud-Dufour S, Luan P, Wang C, Chen W, et al. 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. 2001; 155:937–948. https://doi.org/10.1083/jcb.200106039 PMID: 11739406

43. Lee MCS, Orci L, Hamamoto S, Futai E, Ravazzola M, Schekman R. Sar1p N-Terminal Helix Initiates Membrane Curvature and Completes the Fission of a COPII Vesicle. Cell. 2005; 122(4):605–617. https://doi.org/10.1016/j.cell.2005.07.025 PMID: 16122427

44. Stansfeld PJ, Jefferys E, Sansom M. Multiscale Simulations Reveal Conserved Patterns of Lipid Interactions with Aquaporins. Structure(London, England:1993). 2013; 21:810–819. https://doi.org/10.1016/j.str.2013.03.005 PMID: 23602661

45. Leman JK, Mueller R, Karakas M, Woetzel N, Meiler J. Simultaneous prediction of protein secondary structure and transmembrane spans. Proteins: Structure, Function, and Bioinformatics. 2013; 81 (7):1127–1140. https://doi.org/10.1002/prot.24258 PMID: 23349002

46. Wang S, Li Z, zhou Yu Y, Xu J. Folding Membrane Proteins by Deep Transfer Learning. Cell Systems. 2017; 5 3:202–211.e3. https://doi.org/10.1016/j.cels.2017.09.001 PMID: 28957654

47. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19 12:1589–1591. https://doi.org/10.1093/bioinformatics/btg224 PMID: 12912846

48. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics. 2014; 30:3128–3130. https://doi.org/10.1093/bioinformatics/btu500 PMID: 25064567

49. Jones D, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012; 28 2:184–190. https://doi.org/10.1093/bioinformatics/btr638 PMID: 22101153

50. Jones D, Singh T, Kosciolek T, Tetchner SJ. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics. 2015; 31:999–1006. https://doi.org/10.1093/bioinformatics/btu791 PMID: 25431331

51. Buchan DWA, Jones D. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins. 2018; 86:78–83. https://doi.org/10.1002/prot.25379 PMID: 28901583

52. Adhikari B, Cheng J. CONFOLD2: improved contact-driven ab initio protein structure modeling. BMC Bioinformatics. 2018; 19(1):1–5. https://doi.org/10.1186/s12859-018-2032-6 PMID: 29370750

53. Lesk A, Chothia C. Solvent accessibility, protein surfaces, and protein folding. Biophysical Journal. 1980; 32 1:35–47. https://doi.org/10.1016/S0006-3495(80)84914-9 PMID: 7248454

54. Zhang M, Wang D, Kang Y, Wu J, Yao F, Pan C, et al. Structure of the mechanosensitive OSCA channels. Nature Structural & Molecular Biology. 2018; 25:850–858. https://doi.org/10.1038/s41594-018-0117-6 PMID: 30190597

55. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nature Structural & Molecular Biology. 1996; 3(10):842–848. https://doi.org/10.1038/nsb1096-842 PMID: 8836100

56. Lomize MA, Lomize A, Pogozheva I, Mosberg H. OPM: Orientations of Proteins in Membranes database. Bioinformatics. 2006; 22 5:623–625. https://doi.org/10.1093/bioinformatics/btk023 PMID: 16397007

57. Nugent T, Jones DT. Membrane protein orientation and refinement using a knowledge-based statistical potential. BMC Bioinformatics. 2013; 14(1):1–10. https://doi.org/10.1186/1471-2105-14-276

58. Bond P, Sansom M. Insertion and assembly of membrane proteins via simulation. Journal of the American Chemical Society. 2006; 128 8:2697–2704. https://doi.org/10.1021/ja0569104 PMID: 16492056

59. Scott KA, Bond P, Ivetac A, Chetwynd A, Khalid S, Sansom M. Coarse-grained MD simulations of membrane protein-bilayer self-assembly. Structure. 2008; 16 4:621–630. https://doi.org/10.1016/j.str.2008.01.014 PMID: 18400182

60. Arnarez C, Mazat J, Elezgaray J, Marrink SJ, Periole X. Evidence for cardiolipin binding sites on the membrane-exposed surface of the cytochrome bc_1. Journal of the American Chemical Society. 2013; 135 8:3112–3120. https://doi.org/10.1021/ja310577u PMID: 23363024

61. Chetwynd AP, Scott KA, Mokrab Y, Sansom MS. CGDB: a database of membrane protein/lipid interactions by coarse-grained molecular dynamics simulations. Molecular Membrane Biology. 2008; 25 (8):662–669. https://doi.org/10.1080/09687680802446534 PMID: 18937097

62. Liu Z, Gong Y, Guo Y, Zhang X, Wang H. TMP- SSurface2: A Novel Deep Learning-Based Surface Accessibility Predictor for Transmembrane Protein Sequence. Frontiers in Genetics. 2021; 12. https://doi.org/10.3389/fgene.2021.656140 PMID: 33790952

63. Jeon W, Kim D. FP2VEC: a new molecular featurizer for learning molecular properties. Bioinformatics. 2019; 35(23):4979–4985. https://doi.org/10.1093/bioinformatics/btz307 PMID: 31070725

64. Kim GB, Gao Y, Palsson BO, Lee SY. DeepTFactor: A deep learning-based tool for the prediction of transcription factors. Proceedings of the National Academy of Sciences of the United States of America. 2021; 118(2):1–5. https://doi.org/10.1073/pnas.2021171118 PMID: 33372147

65. Józefowicz R, Zaremba W, Sutskever I. An Empirical Exploration of Recurrent Network Architectures. In: In Proceedings of the 32nd International Conference on Machine Learning (ICML-15); 2015.

66. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods. 2012; 9:173–175. https://doi.org/10.1038/nmeth.1818

67. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. Computing Research Repository. 2015;abs/1412.6980.

68. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research. 2014; 42 (D1):D304–D309. https://doi.org/10.1093/nar/gkt1240 PMID: 24304899

69. Chandonia JM, Fox NK, Brenner SE. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. Nucleic Acids Research. 2019; 47(D1):D475–D481. https://doi.org/10.1093/nar/gky1134 PMID: 30500919

70. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; p. 770–778.

71. Betancourt M, Thirumalai D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Science. 1999; 8:361–369. https://doi.org/10.1110/ps.8.2.361

72. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules. 1985; 18(3):534–552. https://doi.org/10.1021/ma00145a039

73. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Science. 1999; 8(2):361–369. https://doi.org/10.1110/ps.8.2.361

74. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2008; 24(3):333–340. https://doi.org/10.1093/bioinformatics/btm604 PMID: 18057019

75. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000; 16(4):404–405. https://doi.org/10.1093/bioinformatics/16.4.404 PMID: 10869041

76. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallographica-Section D-Biological Crystallography. 1998; 54(5):905–921. https://doi.org/10.1107/S0907444998003254

77. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. 2000; 16(9):776–785. https://doi.org/10.1093/bioinformatics/16.9.776 PMID: 11108700

78. Šali A, Blundell T. Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology. 1993; 234 3:779–815. https://doi.org/10.1006/jmbi.1993.1626 PMID: 8254673

79. Marrink S, Risselada HJ, Yefimov S, Tieleman D, de Vries AH. The MARTINI force field: coarse grained model for biomolecular simulations. The Journal of Physical Chemistry B. 2007; 111 27:7812–7824. https://doi.org/10.1021/jp071097f PMID: 17569554

80. Monticelli L, Kandasamy S, Periole X, Larson R, Tieleman D, Marrink SJ. The MARTINI Coarse-Grained Force Field: Extension to Proteins. Journal of Chemical Theory and Computation. 2008; 4 5:819–834. https://doi.org/10.1021/ct700324x PMID: 26621095

81. de Jong DH, Singh G, Bennett W, Arnarez C, Wassenaar T, Schäfer L, et al. Improved Parameters for the Martini Coarse-Grained Protein Force Field. Journal of Chemical Theory and Computation. 2013; 9 1:687–697. https://doi.org/10.1021/ct300646g PMID: 26589065

82. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. The Journal of Chemical Physics. 2007; 126 1:014101. https://doi.org/10.1063/1.2408420 PMID: 17212484

83. Berendsen H, Postma JP, Gunsteren WF, Dinola A, Haak J. Molecular dynamics with coupling to an external bath. Journal of Chemical Physics. 1984; 81:3684–3690. https://doi.org/10.1063/1.448118

84. Wassenaar T, Pluhackova K, Böckmann R, Marrink S, Tieleman D. Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. Journal of Chemical Theory and Computation. 2014; 10 2:676–690. https://doi.org/10.1021/ct400617g PMID: 26580045

85. Darden T, York D, Pedersen L. Particle mesh Ewald: An *N*log(*N*) method for Ewald sums in large systems. Journal of Chemical Physics. 1993; 98:10089–10092. https://doi.org/10.1063/1.464397

86. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied Physics. 1981; 52:7182–7190. https://doi.org/10.1063/1.328693

87. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror R, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins. 2010; 78:1950–1958. https://doi.org/10.1002/prot.22711 PMID: 20408171

88. Jämbeck JPM, Lyubartsev A. Derivation and Systematic Validation of a Refined All-Atom Force Field for Phosphatidylcholine Lipids. The Journal of Physical Chemistry B. 2012; 116:3164–3179. https://doi.org/10.1021/jp212503e PMID: 22352995

89. Jämbeck JPM, Lyubartsev A. An Extension and Further Validation of an All-Atomistic Force Field for Biological Membranes. Journal of Chemical Theory and Computation. 2012; 8 8:2938–2948. https://doi.org/10.1021/ct300342n PMID: 26592132