Supplemental Material for:

PredMP: a web server for *de novo* prediction and visualization of membrane proteins

Sheng Wang[1,†,*], Shiyang Fei[3,†], Zongan Wang[4,†], Yu Li[1], Jinbo Xu[5], Feng Zhao[2,*] and Xin Gao[1,*]

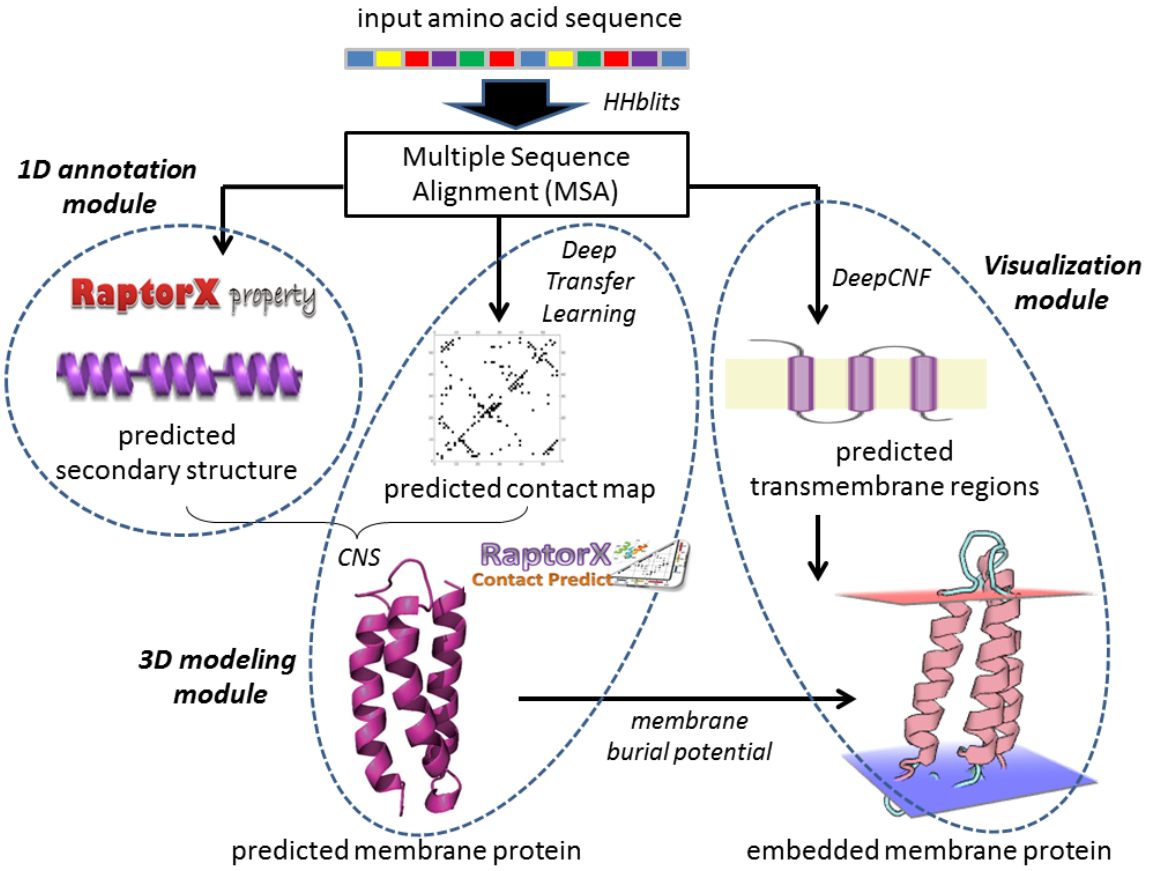
[1]Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. [2]Prospect Institute of Fatty Acids and Health, Qingdao University, China. [3]COMPASS, New York, USA. [4]Department of Chemistry, University of Chicago, USA. [5]Toyota Technological Institute at Chicago, USA.

†Contribute equally.

*To whom correspondence should be addressed: sheng.wang@kaust.edu.sa, fengzhao21c@163.com or xin.gao@kaust.edu.sa.

Contents

S1 Basic workflow of PredMP



Supplementary Figure S1. Illustration of the workflow of PredMP with three modules. Given an input membrane protein sequence, PredMP first uses HHblits [1] to generate the multiple sequence alignment (MSA). The MSA is used to (a) predict transmembrane regions by DeepCNF model, (b) predict secondary structure by RaptorX-Property server [2] through evolutionary analysis (i.e., the ***1D annotation module***), and (c) predict the contact map through Deep Transfer Learning (DTL) with co-evolutionary features [3]. The predicted secondary structure and contacts are fed into Crystallography & NMR System (CNS) suite [4] to *de novo* fold the 3D models by RaptorX-Contact server [5] (i.e., the ***3D modeling module***), which are then embedded into the bilayer membrane with the guide of predicted transmembrane regions and a depth- and residue-dependent membrane burial potential [6] in the ***visualization module***.

S2 Dataset of non-redundant membrane proteins

Supplemental Table S1. A list of 510 non-redundant membrane proteins with solved structures in Protein Data Bank (PDB) from PDBTM database [7]. The entries highlighted with the bold (bold + underline) font indicate the model with TM-score larger than 0.5 (0.6). The entries shown in blue (italic) indicate the barrel membrane proteins (single-pass helical transmembrane proteins), whereas the others are multi-pass helical transmembrane proteins. Users may refer to the link http://predmp.com/#/detail/1xxxA to check the details of the PredMP predictions, where 1xxxA is the membrane protein id (PDB ID: 1xxx plus Chain ID: A).

1a0sP	1pw4A	2evuA	2lnlA	2wpvB	3cn5A	3kvnA	3udcA	4chvA	4il3A	4or2A	4wgvA	5c8jI
1ar1B	1q16C	2f1cX	2lomA	2wsc1	3cx5C	3l11A	3ug9A	4cskA	4in5H	4p6vB	4wmzA	5cfbA
1bccE	1q90A	2f93B	2loqA	2wsc3	3d31C	3lnmB	3ukmA	4czbA	4in5L	4p6vC	4x5mA	5ctgA
1bctA	1q90B	2f95B	2lorA	2wscF	3ddlA	3lw54	3um7A	4d5bA	4j05A	4p6vD	4xk83	5d0yA
1bhaA	1qcrD	2fynB	2losA	2wscG	3dhwA	3lw5H	3uq7A	4d6tD	4j72A	4p6vE	4xnkA	5dirA
1c17M	1qd6C	2ge4A	2lotA	2wscH	3dinE	3m71A	3ux4A	4d6tG	4j7cI	4p6vF	4xnvA	5doqA
1e7pC	1qleC	2gfpA	2lp1A	2wscK	3d18C	3mk7A	3v2wA	4d6tJ	4jkvA	4p79A	4xu4A	5doqB
1ehkB	1rh5B	2gr7A	2m0qA	2wscL	3dl8E	3mk7B	3v5sA	4d6uD	4k1cA	4pgrA	4xxjA	5ee7A
1fftB	1rh5C	2gr8A	2m20A	2wswA	3dwoX	3mk7C	3vmqA	4djiA	4kjrA	4phzA	4xydB	5ek0A
1fftC	1rwtA	2h8aA	2m67A	2wwbB	3dwwA	3mktA	3vouA	4dojA	4knfA	4pirA	4y25A	5ekeA
1fw2A	1s5lB	2h8pC	2m6bA	2wwbC	3dzmA	3mp7A	3vr8C	4dveA	4kppA	4px7A	4y28G	5eulE
1fx8A	1s5lE	2hdfA	2m7gA	2x4mA	3effK	3mp7B	3vr8D	4dxwA	4kt0F	4q2eA	4y28K	5ezmA
1gzmA	1s5lX	2ibzG	2m8rA	2xq2A	3eh3A	3njtA	3vwiA	4e1tA	4kt0K	4qncA	4y28L	5f1cA
1h2sB	1sqqK	2ibzI	2mafA	2xutA	3ejzA	3nymA	3wdoA	4ea3A	4ky0A	4qndA	4y7jA	5fn2B
1h6s1	1t16A	2iubA	2mfrA	2y5yA	3emnX	3o0rB	3wmfA	4ezcA	4l6rA	4qtnA	4ymkA	5gaeh
1izlA	1tlwA	2j58A	2mgyA	2y69D	3emoA	3o7pA	3wmm1	4f35A	4l6v6	4quvA	4ymsC	5gaqA
1izlC	1tqqA	2j7aC	2mm8A	2y69G	3fhhA	3ohnA	3wmmM	4f41A	4l6v8	4r1iA	4ytpC	5garO
1jb0K	1uunA	2jafA	2mmuA	2y69I	3fidA	3orgA	3wo7A	4fqeA	4ltoA	4rdqA	4ytpD	5hk1A
1k24A	1uynX	2jlnA	2mn6A	2y69J	3g67A	3oufA	3wvfA	4fuvA	4m58A	4rfsS	4z34A	5i1mV
1kf6C	1vclA	2jo1A	2mpnA	2y69K	3gi8C	3p5nA	3wxvA	4g1uA	4m64A	4ri2A	4z3nA	5i20A
1kf6D	1vf5B	2jp3A	2mxbA	2y69L	3hd6A	3pjsK	3x29A	4g7vS	4mbsA	4rjwA	4z7fA	5i32A

1kqfB *1vf5D* 2k0lA 2n4xA *2y69M* ***3hw9A*** **3pjzA** 3x2rA **4g80I** **4meeA** **4rl8A** **4zp0A** **5i6cA**

1kqfC *1wrgA* *2k21A* 2n6lA 2yevB **3iyzA** **3pwhA** **3x3bA** **4gbyA** **4mndA** **4rl9A** **4zr0A** **5i6zA**

1kzuA **1xioA** **2k73A** ***2n7qA*** **2yevC** **3iz1A** **3q7kA** **3ze3A** **4gd3A** **4mqsA** **4rlcA** **4zr1A** 5id3A

1lghA 1xl4A 2k9pA **2nmrA** **2yiuA** **3j08A** **3qe7A** **3zevA** 4gx5A 4mt4A **4rngA** **4zw9A** **5iofA**

1m56B 1yc9A *2kluA* **2nq2A** 2ynkA **3j1zP** 3qnqA **3zjzA** *4gycB* **4n74A** 4rp8A 5a1sA 5irxA

1m56D 1yewC *2kogA* **2nr9A** **2z73A** **3j9tR** **3qraA** **3zk1A** **4h33A** 4n75A **4ryiA** **5a40A** **5ivaA**

1m57A **1yq3C** **2ks9A** *2nrgA* **2ziyA** **3jbrE** 3rbzA **3zuxA** 4he8A **4njnA** 4s0vA **5a63C** 5iwsA

1mm4A **1yq3D** 2ksdA *2o01F* ***2zjsE*** 3jcuD ***3rgwS*** 4a2nB **4he8C** 4nppA **4tkrA** **5a63D** ***5ixmB***

1mprA *1zrtE* 2kseA 2oarA *2zxeB* *3jcuH* 3rkoA **4atvA** 4he8D **4ntjA** **4tq3A** **5a6eB** **5jagA**

1n71A *1zzaA* 2ksfA **2pnoA** *2zxeG* *3jcuK* **3rkoB** **4aw6A** **4hkrA** 4nykA **4tquM** 5abbZ

1nekC 2a0lA **2ksrA** **2q67A** **3a2sX** 3jcuR **3rkoC** 4b4aA ***4hqjE*** 4o6mA **4tquN** 5araT

1nekD **2a9hA** 2kyhA **2q7mA** **3a7kA** 3jcuS **3rkoD** **4bemJ** **4httA** **4o6yA** 4twkA **5araW**

1o5wA 2akhA 2l35A **2qomA** *3anzA* 3jcuW 3rkoF 4bgnA **4huqS** 4o9pA **4u15A** **5awwG**

1occD 2akhB *2l8sA* 2r6gF **3b4rA** ***3jcuX*** **3rkoG** 4bog3 **4huqT** **4o9pB** **4u4tA** 5awwY

1oedC 2bg9A **2lckA** **2r6gG** 3b5dA **3jcuZ** 3s0xA 4bpmA 4hw9A 4o9uB **4u9lA** **5awzA**

1orsC **2bl2A** **2lhfA** **2vpwC** **3b9wA** 3jycA **3sljA** **4bwzA** 4hycA **4od4A** **4uc1A** **5aymA**

1p49A *2cpbA* 2lkgA **2w1pA** **3bryA** **3k3fA** **3sybA** **4c9jA** **4hyoA** 4ogqC **4us3A** **5azbA**

1p4tA **2d57A** **2llyA** **2wjqA** 3chxB **3kj6A** **3tijA** **4cadC** **4hzuS** **4oh3A** **4v1fA** *5bwkE*

1p7bA **2ervA** 2lmeA **2wpdJ** **3chxC** 3kp9A **3tx3A** *4cfgA* *4iffA* **4oo9A** **4wd7A** **5c6oA**

S3 Transmembrane region prediction by DeepCNF

To train a machine learning model for predicting the transmembrane region at each residue given a protein primary sequence, we performed the following procedures. We first collected 510 non-redundant transmembrane proteins (shown in Table S1) at the chain level from PDBTM [7]. To label each residue from a given transmembrane protein sequence, we used the following 9 labels extracted from PDBTM: 1 (Side1), 2 (Side2), B (Beta-strand), H (alpha-helix), C (coil), I (membrane-inside), L (membrane-loop), F (interfacial helix), and U (unknown localizations). We then trained a deep learning model, DeepCNF [8, 9], on this annotated sequence dataset.

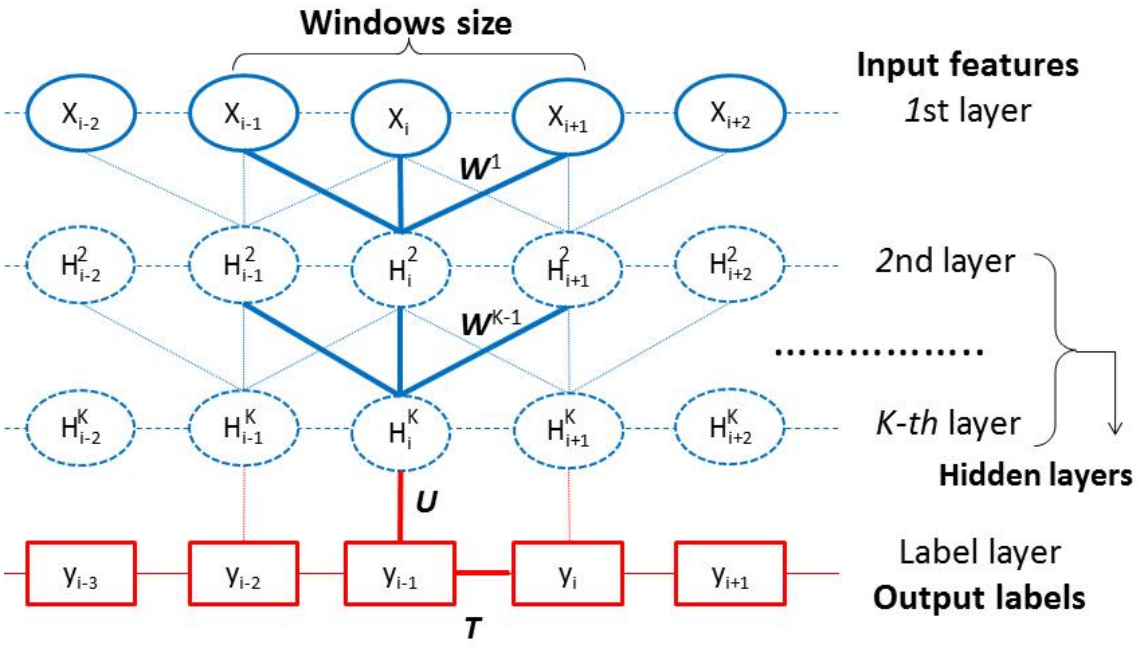
As shown in Figure S2, DeepCNF has two modules: (i) the Conditional Random Fields (CRF) [10], and (ii) the Deep Convolutional Neural Network (DCNN) [11]. DeepCNF can model not only complex relationship between the sequence and transmembrane regions by a deep hierarchical architecture, but also interdependency between adjacent transmembrane region labels [9]. To deal with the imbalanced distribution of some transmembrane region labels, such as interfacial helix and membrane-inside, we trained DeepCNF by maximizing AUC [6]. According to [9], the DCNN architecture is set as follows: it consists of five layers where each layer has 100 neurons and the window size at each layer is set to 11.

We used the following 68 input features: 20 one-hot encoding from the primary sequence, 20 position specific scoring matrix (PSSM) from PSI-BLAST [12] with E-value threshold 0.001 and three iterations to search UniRef90 [13], 20 PSSM from HHblits [1] with E-value threshold 0.001 and three iterations to search UniProt20 [13], and 8 predicted eight-state secondary structure element by DeepCNF_SS [9]. Note that although we used DeepCNF_SS to generate the predicted secondary structure features for transmembrane region prediction, the training data for DeepCNF_SS only come from non-MPs.

This method achieved 62% cross-validation predictive accuracy on classifying a residue into the nine categories of the transmembrane region. If we merged label B, H, and C as ‘transmembrane region’ label, and all other labels as ‘non-transmembrane region’ label, then this method could achieve 89% predictive accuracy, as well as AUC and AUPRC 0.94 and 0.89, respectively. Finally, using forward-backward algorithm in CRF [10], we assigned to each residue position a reliable ‘transmembrane’ or ‘non-transmembrane’ label based on the computed probability.

It should be noted that other transmembrane region (or, membrane protein topology) predictors could also be used here, such as TOPCONS [14], MEMSAT-SVM [15], PHOBIUS [16], or OCTOPUS [17], just name a few. We will add these third-party tools for predicting and visualizing transmembrane regions in the next release version of PredMP.

Last but not least, this transmembrane region prediction module will be added to RaptorX-Property [2] in the near future. Currently, users may refer to the source code at GitHub https://github.com/realbigws/RaptorX_Property_Fast.



Supplemental Figure S2. Illustration of DeepCNF. Here *i* is the position index and X_i the associated input features, H^k represents the *k*-th hidden layer, and *Y* is the output label. All the layers from the 1st to the K^{th} form a deep convolutional neural network (DCNN) with parameter $W^k\{k=1,2,...,K\}$, which is shown in blue. The K^{th} layer and the label layer form a Conditional Random Fields (CRF), which is shown in red. The parameter *U* specifies the relationship between the K^{th} layer and the label layer, and *T* the binary relationship between adjacent labels. This figure is taken from Wang S. *et. al.* [2].

S4 Blind test of membrane protein cases in CAMEO

Blind and live test in CAMEO

CAMEO [18] can be interpreted as a fully automated CASP [19], but has a smaller number (about 40) of participating servers since many CASP-participating servers are not fully automated. By ‘‘blind’’ it means that the experimentally solved structure of a test protein has not been released in PDB when it is used as a test target. By “live” it means that every weekend CAMEO releases about 20 sequences for prediction test. The test proteins used by CAMEO have no publicly available native structures before it finishes collecting models from servers. The CAMEO server ID of RaptorX-Contact (the main module in PredMP server to generate the 3D models) is Server60, and it has been fully functioning since September 2016.

Since experimentally solving the structures of membrane proteins (MPs) is challenging, starting from September 2016 and up to January 2018, we have observed 10 non-homologous MPs among all CAMEO hard targets, as shown in Table S2.

Supplemental Table S2. A list of 10 non-homologous membrane proteins among all CAMEO hard targets from Sep 2016 to Jan 2018.

5h35E (CAMEO ID: 2017-01-07_00000030_3)

5jkiA (CAMEO ID: 2017-02-18_00000075_1)

5l0wA (CAMEO ID: 2017-03-18_00000059_2)

5khnA (CAMEO ID: 2017-06-10_00000043_1)

5kymB (CAMEO ID: 2017-07-22_00000026_1)

5mm0A (CAMEO ID: 2017-08-05_00000083_1)

5gufA (CAMEO ID: 2017-10-07_00000005_1)

5ogkH (CAMEO ID: 2017-11-18_00000021_1)

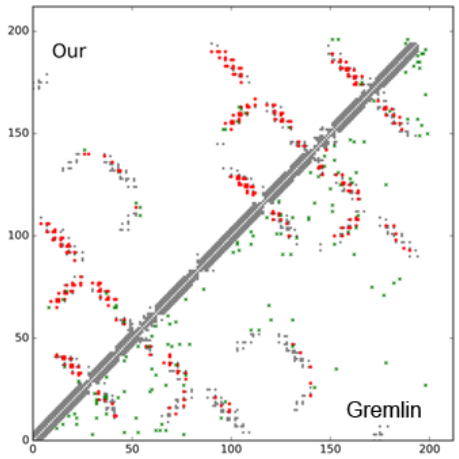
6bmsB (CAMEO ID: 2018-01-06_00000139_1)

5vkvA (CAMEO ID: 2018-01-27_00000035_1)

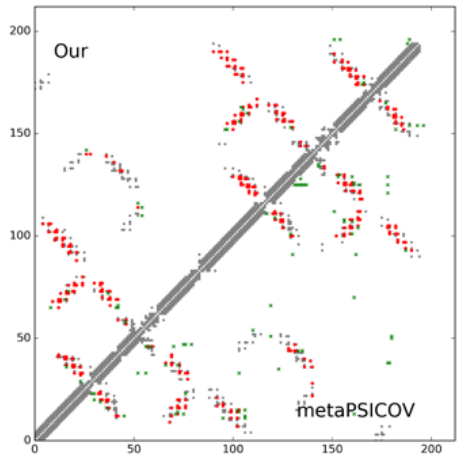
We show in the following sections that RaptorX-Contact successfully modeled all ten MPs belonging to the hard category of CAMEO.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.778	0.953	1.000	1.000	0.316	0.547	0.905	1.000
metaPSICOV	0.571	0.774	0.929	1.000	0.245	0.401	0.619	0.810
Gremlin	0.340	0.528	0.786	0.810	0.137	0.217	0.429	0.619

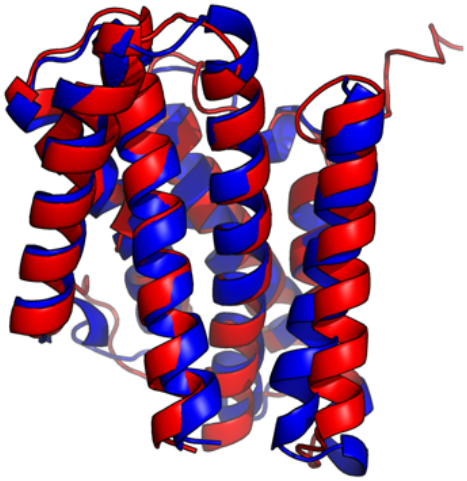
(A)



(B)



(C)



(D)

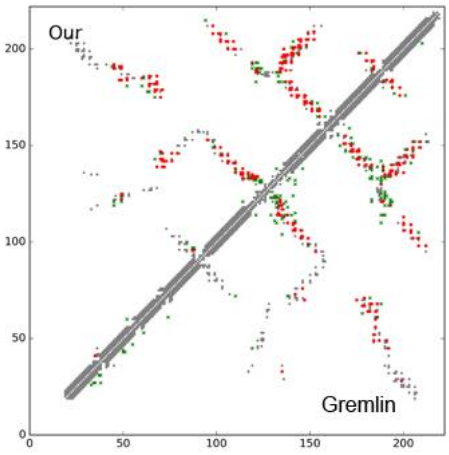
Server Name	IDDT	IDDT Cα
Server 60	62.88	72.17
Server 19	53.06	57.93
Robetta	52.70	57.53
Server 45	48.38	53.44
RaptorX	48.29	53.40

(E)

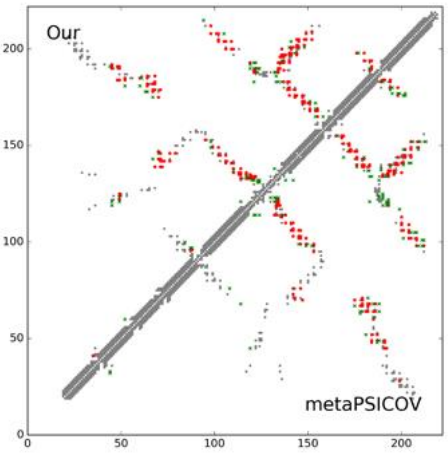
Supplemental Figure S3. Case study of CAMEO target 5h35E. This protein is an intracellular cation channel ortholog from *Sulfolobus solfataricus*. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.658	0.883	1.000	1.000	0.185	0.351	0.659	0.864
metaPSICOV	0.554	0.820	0.977	1.000	0.158	0.279	0.523	0.727
Gremlin	0.495	0.703	0.773	0.818	0.131	0.207	0.477	0.682

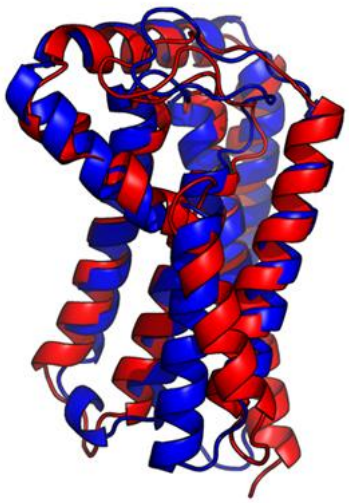
(A)



(B)



(C)



(D)

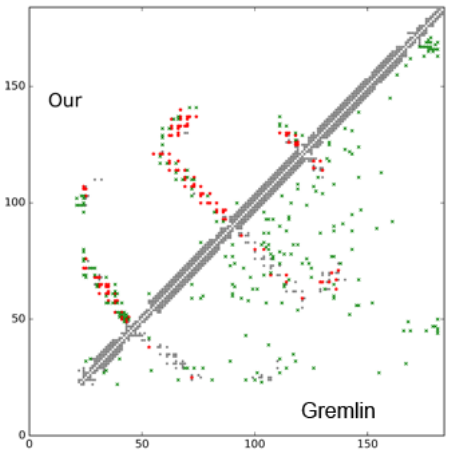
Server Name	IDDT	IDDT Cα
Robetta	60.26	70.05
Server 60	56.05	67.03
IntFOLD4-TS	53.44	63.09
Server 56	53.44	63.09
Server 57	53.40	62.11

(E)

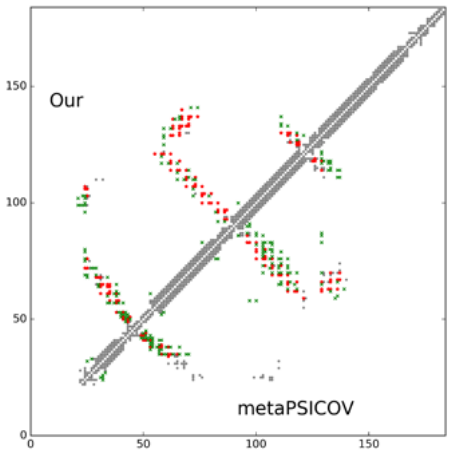
Supplemental Figure S4. Case study of CAMEO target 5jkiA. This protein is a transmembrane PAP2 type phosphatidylglycerolphosphate phosphatase from *Bacillus subtilis*. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.397	0.674	0.889	1.000	0.103	0.207	0.444	0.778
metaPSICOV	0.250	0.391	0.528	0.722	0.098	0.163	0.278	0.389
Gremlin	0.087	0.109	0.222	0.333	0.016	0.033	0.056	0.056

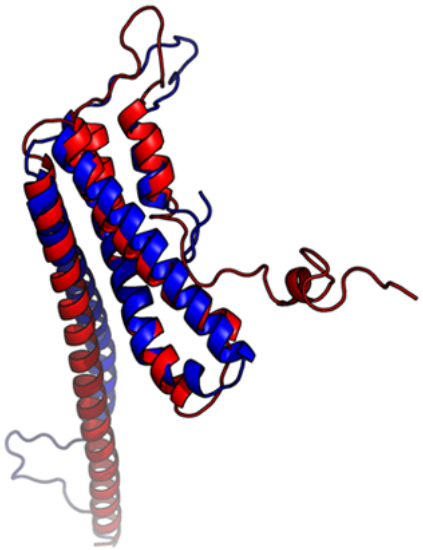
(A)



(B)



(C)



(D)

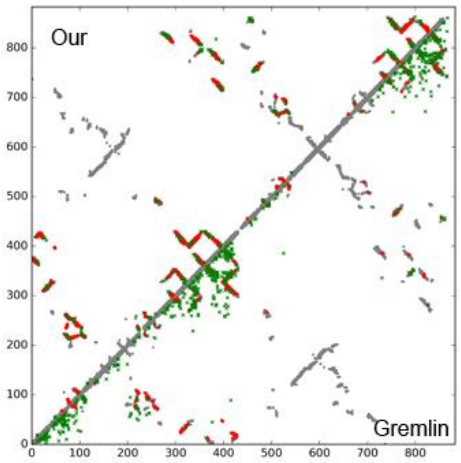
Server Name	IDDT	IDDT Cα
Server 60	60.34	74.87
Robetta	46.71	56.46
Princeton_TEMPLATE	41.32	49.82
RaptorX	38.53	46.49
Server 56	37.81	47.04

(E)

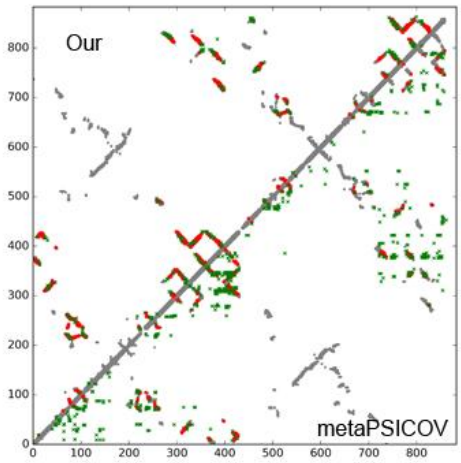
Supplemental Figure S5. Case study of CAMEO target 5l0wA. This protein is a post-translational translocation Sec71/Sec72 complex from *Escherichia coli*. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.700	0.871	0.926	0.977	0.138	0.245	0.500	0.750
metaPSICOV	0.297	0.426	0.602	0.750	0.076	0.109	0.193	0.284
Gremlin	0.254	0.388	0.585	0.704	0.063	0.122	0.256	0.454

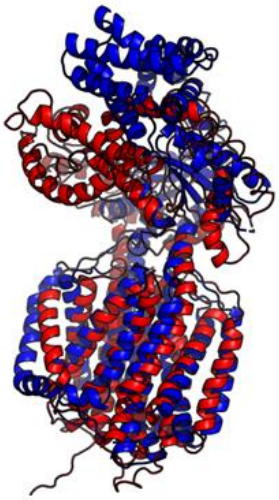
(A)



(B)



(C)



(D)

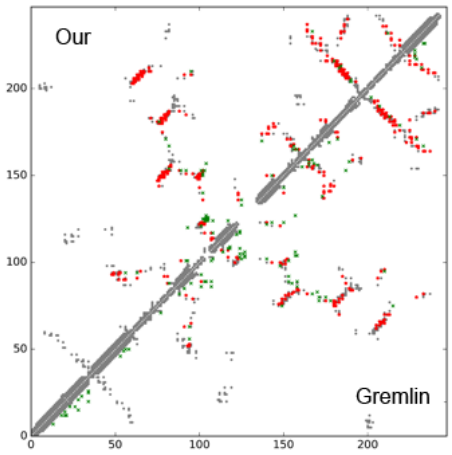
Server Name	IDDT	IDDT Cα
Server 60	41.19	48.39
Robetta	41.97	47.43
HHpredB	38.20	46.26
Server 45	38.90	44.11
RaptorX	38.71	43.87

(E)

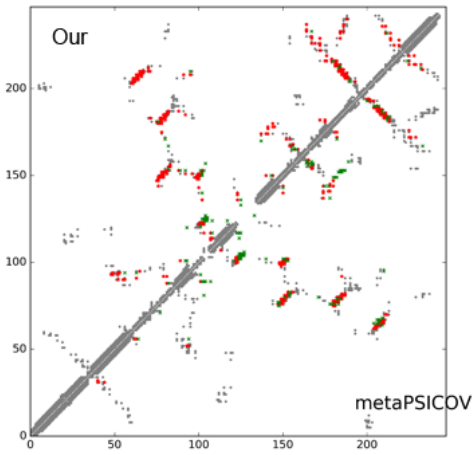
Supplemental Figure S6. Case study of CAMEO target 5khnA. This protein is the Burkholderia multivorans hopanoid transporter HpnN. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.773	0.927	0.918	1.000	0.275	0.472	0.673	0.667
metaPSICOV	0.534	0.748	0.939	0.958	0.231	0.398	0.592	0.708
Gremlin	0.563	0.780	0.837	0.917	0.142	0.260	0.551	0.625

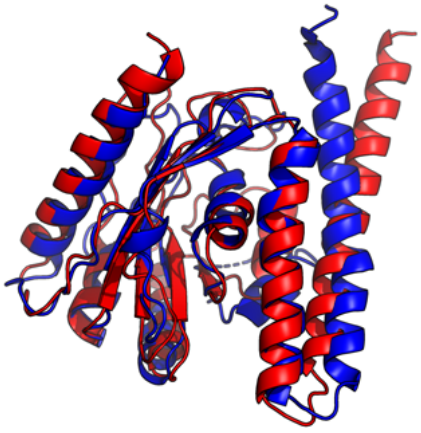
(A)



(B)



(C)



(D)

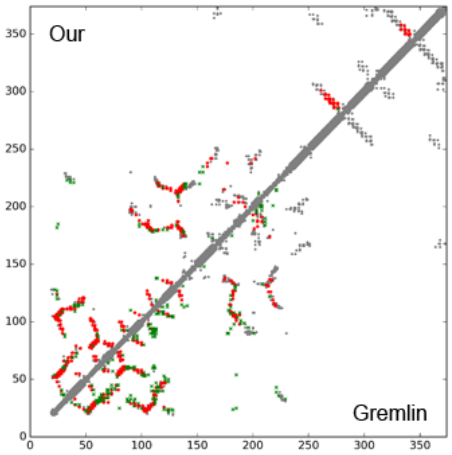
Server Name	IDDT	IDDT Cα
Robetta	57.27	66.08
Server 60	53.28	63.62
RaptorX	51.19	58.88
Server 45	51.06	59.16
IntFOLD3-TS	50.60	59.44

(E)

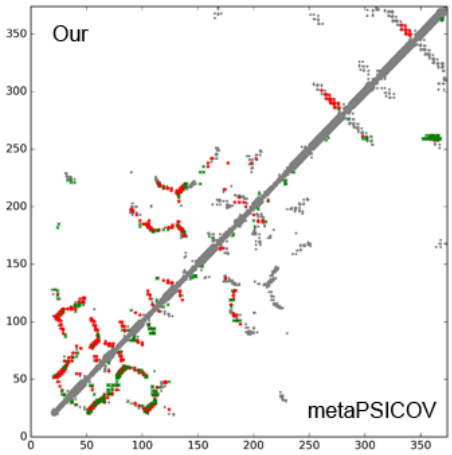
Supplemental Figure S7. Case study of CAMEO target 5kymB. This protein is the 1-acyl-sn-glycerophosphate (LPA) acyltransferase, PlsC, from *Thermotoga maritima*. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.703	0.904	1.000	1.000	0.235	0.412	0.784	0.946
metaPSICOV	0.398	0.652	0.838	0.973	0.163	0.230	0.392	0.622
Gremlin	0.436	0.604	0.784	0.838	0.174	0.278	0.419	0.541

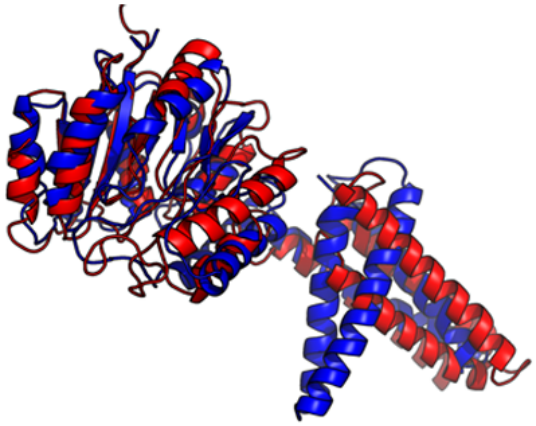
(A)



(B)



(C)



(D)

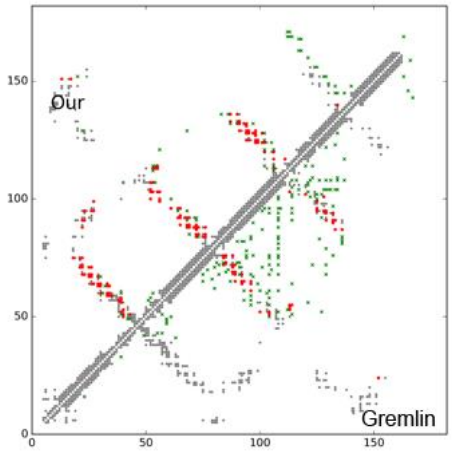
Server Name	IDDT	IDDT Cα
IntFOLD4-TS	50.99	61.45
Server 60	47.48	59.01
Robetta	50.64	58.98
HHpredB	46.49	58.53
Server 55	48.69	56.69

(E)

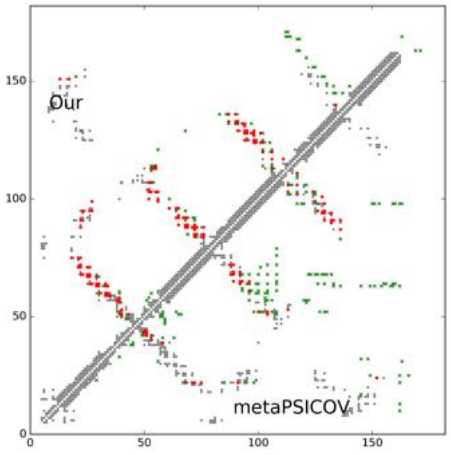
Supplemental Figure S8. Case study of CAMEO target 5mm0A. This protein is a Dolichyl phosphate mannose synthase. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.593	0.835	1.000	1.000	0.280	0.495	0.667	0.667
metaPSICOV	0.209	0.286	0.472	0.722	0.165	0.253	0.389	0.389
Gremlin	0.181	0.330	0.472	0.611	0.082	0.121	0.250	0.389

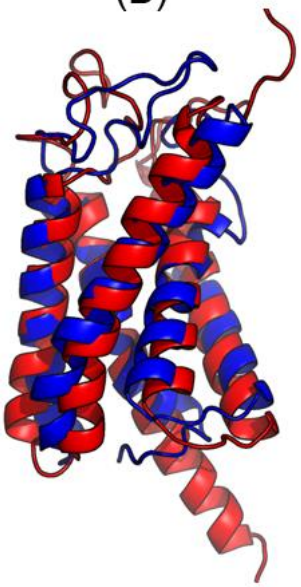
(A)



(B)



(C)



(D)

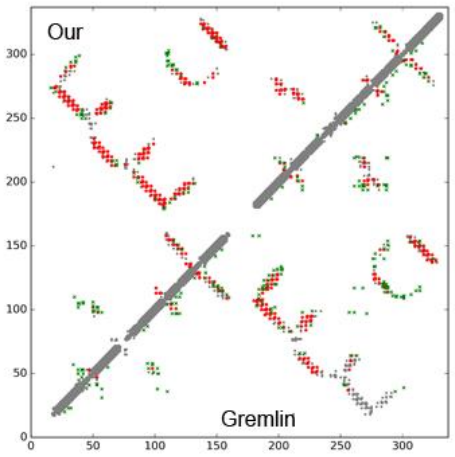
Server Name	IDDT	IDDT Cα
Server 60	53.58	63.47
IntFOLD4-TS	47.71	56.46
Robetta	46.95	56.04
SPARKS-X	43.95	53.09
Server 0	38.83	46.11

(E)

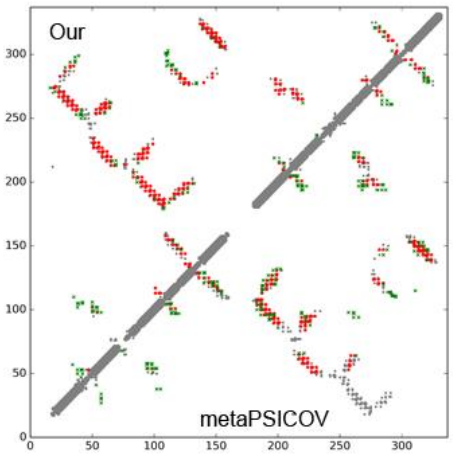
Supplemental Figure S9. Case study of CAMEO target 5gufA. This protein is a CDP-archaeol synthase (CarS). (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.697	0.881	0.970	1.000	0.042	0.083	0.209	0.424
metaPSICOV	0.487	0.696	0.910	0.970	0.045	0.089	0.194	0.364
Gremlin	0.442	0.625	0.761	0.818	0.039	0.077	0.194	0.364

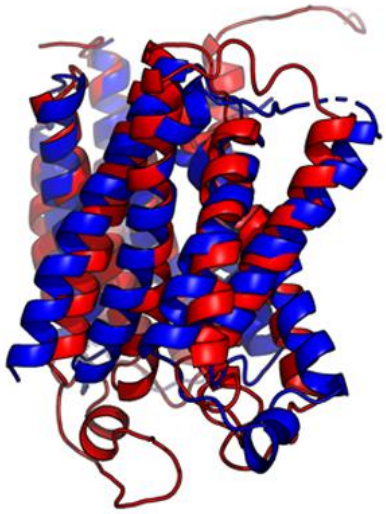
(A)



(B)



(C)



(D)

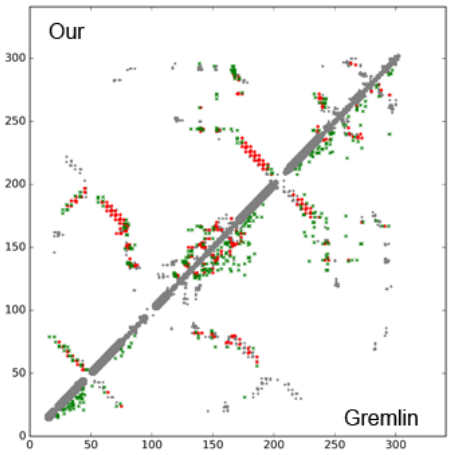
Server Name	IDDT	IDDT Cα
Robetta	61.56	69.16
IntFOLD4-TS	58.58	67.98
Server 55	59.75	67.87
Server 60	57.99	67.81
SWISS-MODEL	59.35	67.56

(E)

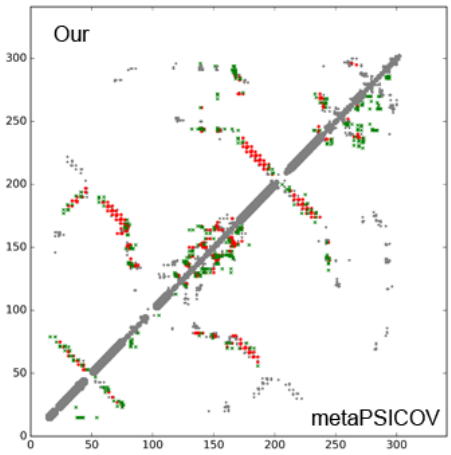
Supplemental Figure S10. Case study of CAMEO target 5ogkH. This protein is a nucleotide sugar transporter. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.408	0.612	0.912	1.000	0.205	0.288	0.456	0.647
metaPSICOV	0.252	0.394	0.647	0.824	0.117	0.188	0.353	0.471
Gremlin	0.226	0.329	0.559	0.676	0.103	0.182	0.353	0.500

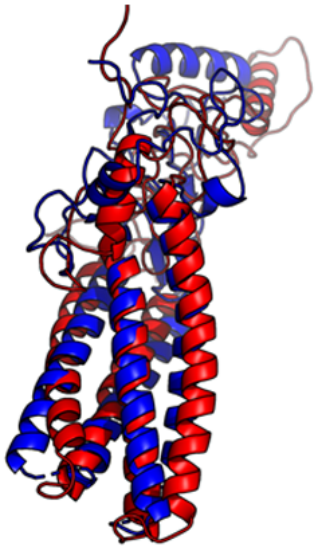
(A)



(B)



(C)



(D)

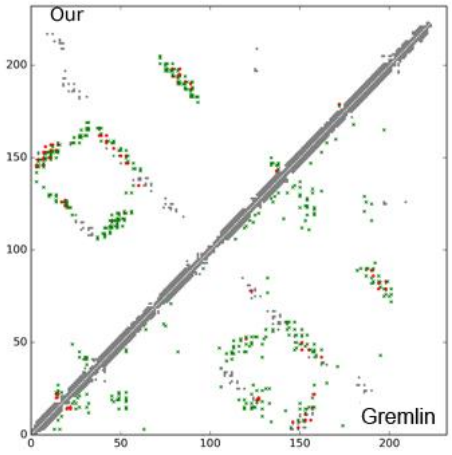
Server Name	IDDT	IDDT Cα
Server 60	41.28	49.01
Server 45	32.81	37.92
RaptorX	30.09	35.21
Robetta	30.05	35.46
Server 55	29.39	34.98

(E)

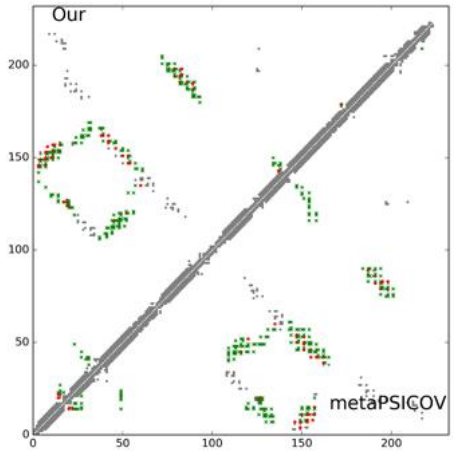
Supplemental Figure S11. Case study of CAMEO target 6bmsB. This protein is a DHHC (Asp-His-His-Cys) palmitoyltransferases. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

	Long range accuracy				Medium range accuracy			
	L	L/2	L/5	L/10	L	L/2	L/5	L/10
Our method	0.207	0.293	0.413	0.435	0.004	0.009	0.000	0.000
metaPSICOV	0.168	0.259	0.326	0.391	0.004	0.009	0.000	0.000
Gremlin	0.099	0.155	0.283	0.435	0.000	0.000	0.000	0.000

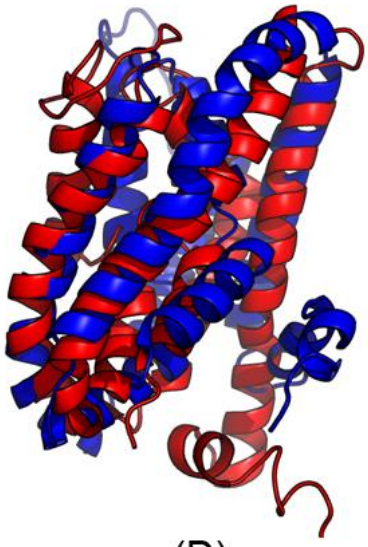
(A)



(B)



(C)



(D)

Server Name	IDDT	IDDT Cα
Server 60	44.60	50.47
Robetta	43.88	49.44
Server 71	40.06	46.20
IntFOLD4-TS	34.42	38.59
Server 45	33.36	37.87

(E)

Supplemental Figure S12. Case study of CAMEO target 5vkvA. This protein is the membrane electron transporter CcdA. (A) The long- and medium-range contact prediction accuracy of our methods, MetaPSICOV, and Gremlin. (B-C) The overlap between the native contact map and contact maps predicted by our method, Gremlin, and MetaPSICOV. Top L predicted all-range contacts are displayed. A gray, red, and green dot represents a native contact, a correct prediction, and a wrong prediction, respectively. (D) The superimposition between our predicted model (in red) and the native structure (in blue). (E) The list of top models submitted by CAMEO servers and their quality scores.

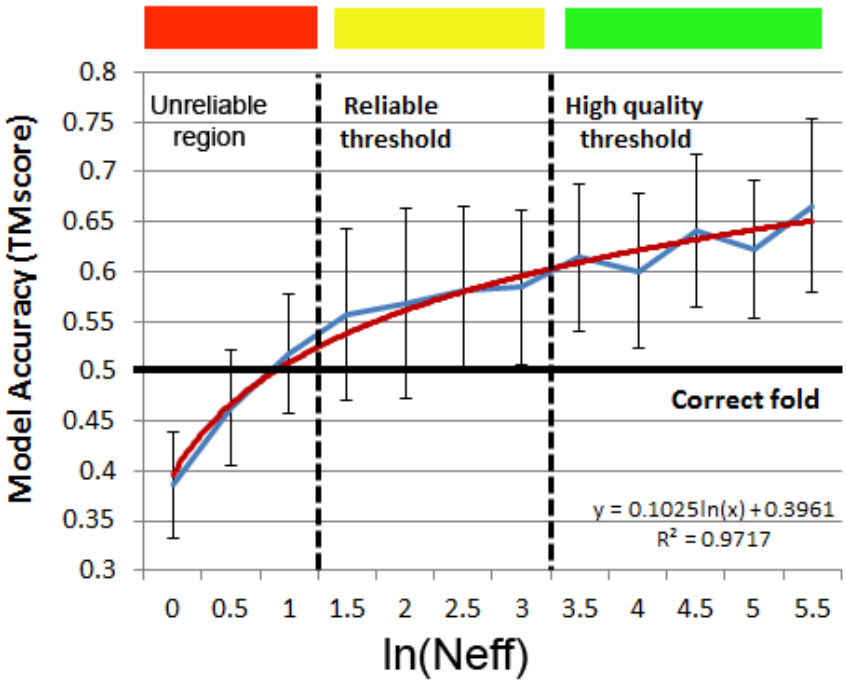
S5 Estimation of the 3D modeling accuracy

We performed a statistical study to show the relationship between 3D model quality and the number of effective sequence homologs (i.e., Meff) using 356 multi-pass helical MPs from the 510 dataset (as shown in Table S1).

We used Meff to measure the amount of homologous information in an MSA (multiple sequence alignment). It can be interpreted as the number of non-redundant (or effective) sequence homologs in an MSA when 70% sequence identity is used as cutoff [20].

We measured the quality of a 3D model by TM-score [21], which ranges from 0 to 1 indicating the worst and the best quality, respectively. A 3D model with TM-score≥0.6 is likely to have a correct fold while a 3D model with TM-score<0.5 usually does not. TM-score = 0.5 is also used by the community as a cutoff to judge if a model has a correct fold or not [22].

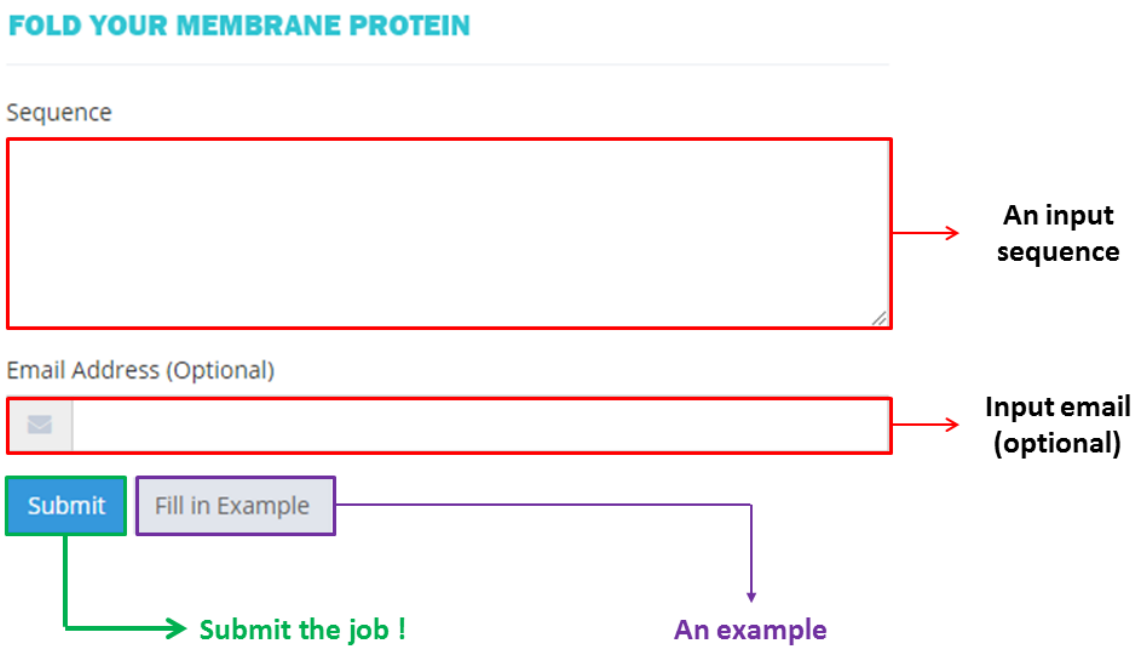
Figure S13 shows the TM-score of the 356 MPs with respect to the length-normalized Meff (or, Neff which is defined as Meff/$L^{0.7}$). When ln(Neff) is larger than 1.5 and 3.5, the predicted models on average have TM-score >0.5 and >0.6, respectively.



Supplemental Figure S13. 3D modeling accuracy of transmembrane proteins (measured by TM-score) with respect to the number of effective sequence homologs in MSA (measured by Neff defined as Meff/$L^{0.7}$). The blue curve shows the mean and standard deviation at each ln(Neff) bin at 0.5 unit, whereas the red line is a fitted curve of the blue curve.

S6 Input/output explanation of the PredMP server

Input of the PredMP server



Supplemental Figure S14. The only required input of PredMP is the membrane protein sequence. The "Job Submission" section also allows users to provide an email address to be used for notification when the job is done. An email is not required, but strongly recommended since it can be used to retrieve the results of your job.

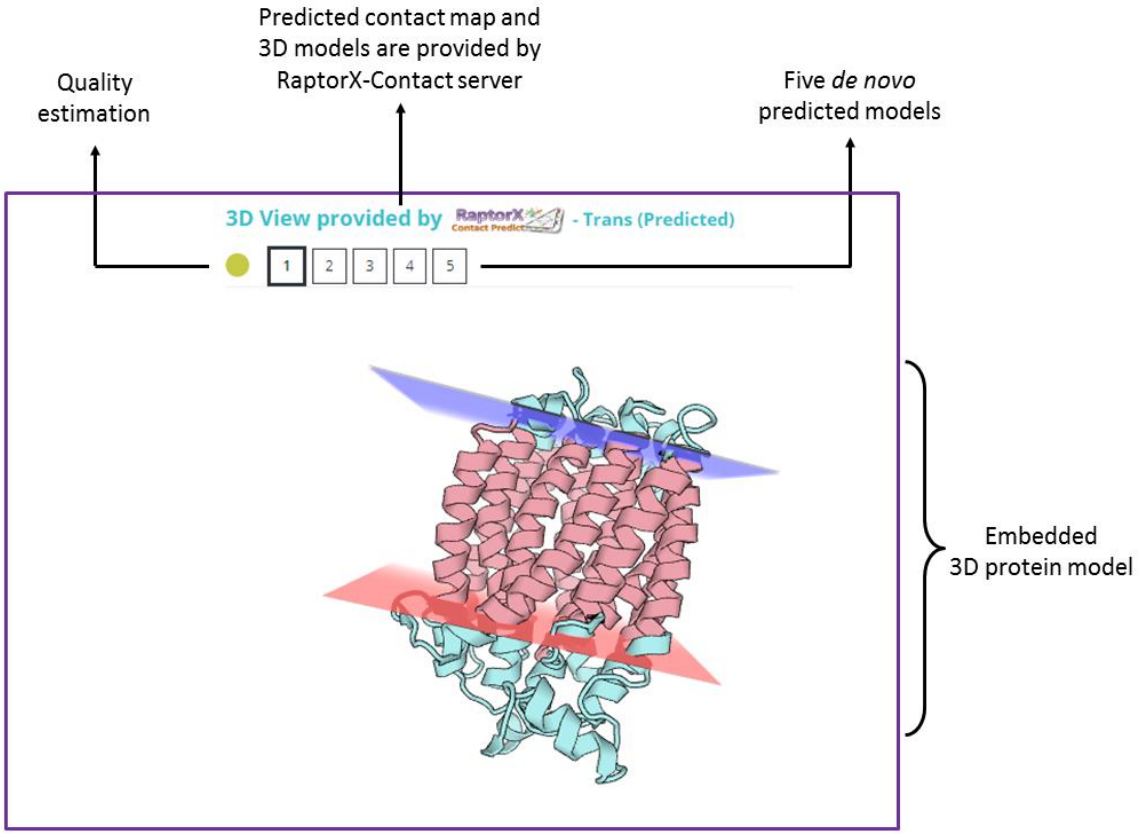
Output of the PredMP server

The outputs of the PredMP server include:

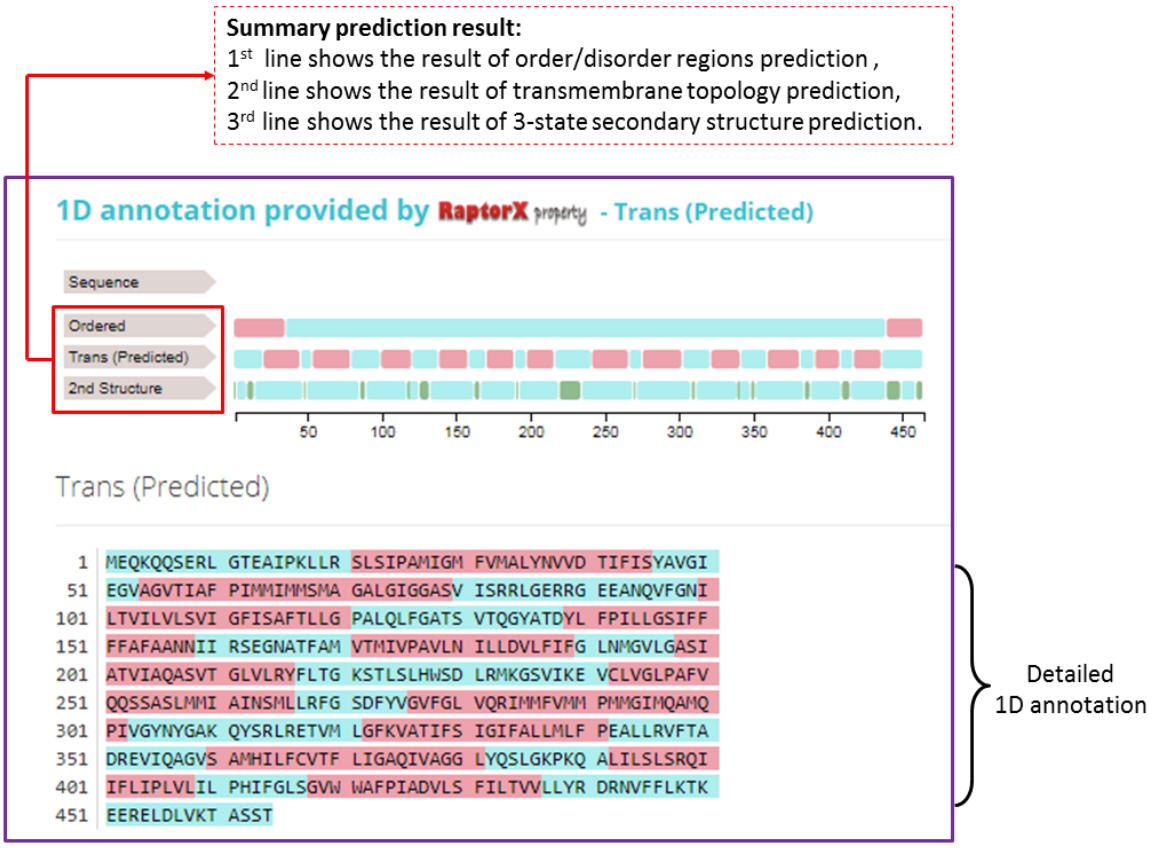
1) Five full-length *de novo* constructed 3D models of the input membrane protein sequence. These models are ranked according to the energy function of Crystallography & NMR System (CNS) [4]. These models are then embedded into the bilayer membrane by the Positioning of Proteins in Membranes (PPM) method [23], as shown in Figure S15.

2) Estimated accuracy of the predicted 3D models in three categories: high confidence, medium confidence, and low confidence. The confidence score is calculate based on Neff (defined as $Meff/L^{0.7}$) which measures the amount of homologous information in the multiple sequence alignment (MSA), as explained in Figure S13.

3) 1D annotation of local structural properties, including the predicted secondary structure, the disordered region, and the transmembrane topology, as illustrated in Figure S16.



Supplemental Figure S15. The result page of the PredMP server for the 3D model prediction followed by the embedding into the bilayer membrane. PredMP will remotely call RaptorX-Contact server to provide five full-length *de novo* constructed 3D models of the input membrane protein sequence. PredMP also estimates the accuracy of the predicted 3D models in three categories: high confidence (in green), medium confidence (in yellow), and low confidence (in red), respectively.



Supplemental Figure S16. The result page of the PredMP server for the 1D annotation of local structural properties. PredMP will remotely call RaptorX-Property server to provide these local structural properties. Specifically, the upper section shows the summary predicted results, with the first row showing the result of order/disorder regions, and the remaining rows showing the prediction of transmembrane topology, and 3-state secondary structure, respectively. By clicking on a specific summary result, such as the predicted transmembrane topology, the detailed annotation on the input sequence is shown in the lower section.

References

1. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nature methods, 2012. **9**(2): p. 173.
2. Wang, S., et al., *RaptorX-Property: a web server for protein structure property prediction.* Nucleic acids research, 2016. **44**(W1): p. W430-W435.
3. Wang, S., et al., *Folding membrane proteins by deep transfer learning.* Cell systems, 2017. **5**(3): p. 202-211. e3.
4. Brunger, A.T., et al., *Crystallography & NMR system: A new software suite for macromolecular structure determination.* Acta Crystallographica-Section D-Biological Crystallography, 1998. **54**(5): p. 905-921.
5. Wang, S., et al., *CoinFold: a web server for protein contact prediction and contact-assisted protein folding.* Nucleic acids research, 2016. **44**(W1): p. W361-W366.
6. Wang, S., J. Ma, and J. Xu, *AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields.* Bioinformatics, 2016. **32**(17): p. i672-i679.
7. Kozma, D., I. Simon, and G.E. Tusnady, *PDBTM: Protein Data Bank of transmembrane proteins after 8 years.* Nucleic acids research, 2012. **41**(D1): p. D524-D529.
8. Wang, S., et al., *DeepCNF-D: predicting protein order/disorder regions by weighted deep convolutional neural fields.* International journal of molecular sciences, 2015. **16**(8): p. 17315-17330.
9. Wang, S., et al., *Protein secondary structure prediction using deep convolutional neural fields.* Scientific reports, 2016. **6**: p. 18962.
10. Lafferty, J., A. McCallum, and F.C. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* 2001.
11. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.
12. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic acids research, 1997. **25**(17): p. 3389-3402.
13. Bairoch, A., et al., *The universal protein resource (UniProt).* Nucleic acids research, 2005. **33**(suppl_1): p. D154-D159.
14. Bernsel, A., et al., *TOPCONS: consensus prediction of membrane protein topology.* Nucleic acids research, 2009. **37**(suppl_2): p. W465-W468.
15. Nugent, T. and D.T. Jones, *Transmembrane protein topology prediction using support vector machines.* BMC bioinformatics, 2009. **10**(1): p. 159.
16. Käll, L., A. Krogh, and E.L. Sonnhammer, *A combined transmembrane topology and signal peptide prediction method.* Journal of molecular biology, 2004. **338**(5): p. 1027-1036.
17. Viklund, H. and A. Elofsson, *OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar.* Bioinformatics, 2008. **24**(15): p. 1662-1668.
18. Haas, J., et al., *The Protein Model Portal—a comprehensive resource for protein structure and model information.* Database, 2013. **2013**.
19. Schaarschmidt, J., et al., *Assessment of contact predictions in CASP12: co‐evolution and deep learning coming of age.* Proteins: Structure, Function, and Bioinformatics, 2017.
20. Wang, S., et al., *Accurate de novo prediction of protein contact map by ultra-deep learning model.* PLoS computational biology, 2017. **13**(1): p. e1005324.

21. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins: Structure, Function, and Bioinformatics, 2004. **57**(4): p. 702-710.
22. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score= 0.5?* Bioinformatics, 2010. **26**(7): p. 889-895.
23. Lomize, A.L., et al., *Positioning of proteins in membranes: a computational approach.* Protein science, 2006. **15**(6): p. 1318-1333.