ProtRAP: Predicting lipid accessibility together with solvent accessibility of proteins in one run

Kai Kang,[†,¶] Lei Wang,[†,¶] and Chen Song[*,†,‡]

†*Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China*

‡*Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China*

¶*These authors contributed equally to this work.*

Received August 24, 2022; E-mail: c.song@pku.edu.cn

Abstract: Solvent accessibility has been extensively used to characterize and predict the chemical properties of surface residues of soluble proteins. In contrast, there is not yet a widely accepted quantity of the same dimension for the study of lipid accessible residues of membrane proteins. In this work, we propose that *lipid accessibility*, defined in a similar way to the solvent accessibility, can be used to well characterize the lipid accessible residues of membrane proteins. Moreover, we developed a deep learning-based method, ProtRAP (Protein Relative Accessibility Predictor), to predict the relative lipid accessibility and relative solvent accessibility of residues from a given protein sequence, which can infer which residues are likely accessible to lipids, accessible to solvent, or buried in the protein interior in one run.

Solvent accessibility (SA) was first proposed in 1971 by Lee and Richards for the characterization of surface residues of soluble proteins.[1] Ever since, SA has been used as an important one-dimensional predictive property of protein residues for the structural and functional studies of proteins. To quantitatively measure SA, Accessible Surface Area (ASA) is calculated by the rolling-a-ball algorithm.[2] Often, the relative accessible surface area (RASA) is used, which is defined as the fraction of ASA of a given amino acid residue in the polypeptide chain to that in the center of a tripeptide adjacent to glycines.[3] Generally, a residue with high SA means it is located on the surface of a protein, which can help determine protein folding and stability[4]. SA is also important for disease discovery and drug design, as residues located on the protein surface are more likely to serve as active sites or interact with ligand and drug molecules.[5–7] Therefore, there have been numerous studies focusing on the prediction of SA of proteins, which were further boosted by the rapid development of structural biology and deep learning in recent years.[8–11]

However, SA is not enough for the description of surface residues of membrane proteins. Membrane proteins are embedded or anchored in a lipid bilayer, a distinct chemical environment from solvent. As a consequence, many surface residues of membrane proteins are not solvent accessible, but lipid accessible. Strikingly, there is not yet a widely accepted quantity to characterize the lipid accessible properties of surface residues of membrane proteins, which play crucial roles in signal transduction and mass transport across membranes[12] and serve as the major targets for drug design.[13–15] Therefore, we think that something similar to SA but for the lipid accessible residues is a highly desirable yet missing quantity for the study of membrane proteins.

Recently, we proposed to use *membrane contact probability* (MCP) to predict the likelihood of proteins residues in direct contact with membranes.[16] Although showing great potential, MCP does not have the same dimension as SA, and is inherently incompatible with SA. Therefore, we sought to study *lipid accessibility* (LA), a quantity defined in the same/similar way to SA, to characterize the surface residues of membrane proteins that are accessible to lipid molecules. The ultimate purpose is to be able to combine LA and SA to describe the surface residues of membrane proteins in a more complete way.

There were very few papers studying LA of protein residues. Adamian et al. used a canonical model to predict LA of protein residues,[17] and later on, the support vector machine (SVM)[18] method was utilized as well. However, at that time, there were not much data available for the training purpose (no more than 100 membrane protein structures in their datasets), and consequently the methods did not show satisfactory performance. In the meanwhile, several methods can predict the transmembrane region of the protein, thereby indirectly predicting LA.[19–21] A recent binary prediction method was included in TopProperty, which can predict whether a residue is exposed to membranes or not, but the binary prediction is still not as compatible and informative as LA in complementing SA.[22]

To predict LA in a more direct and accurate way, in this work, we propose an attention-assisted neural network, named ProtRAP (Protein Relative Accessibility Predictor), to predict the LA of residues of membrane proteins from sequences. In fact, our model can accurately predict the relative lipid accessibility (RLA) and relative solvent accessibility (RSA) of protein residues for any given protein sequence in one run, thus can provide more complete information about the likely surface residues of both soluble and membrane proteins.

The overall method and architecture of ProtRAP is shown in Figure 1. To generate a dataset for the training purpose, as shown in Fig. 1A, we used 1362 non-redundant membrane protein structures and 7740 non-redundant soluble protein structures from the Protein Data Bank (PDB), and adopted the rolling-a-ball (of 1.4 Å radius) method to generated the RASA of each residue. Then, we used MCP, as observed in molecular dynamics (MD) simulations in the MemProtMD database,[23] to determine the RLA and RSA of each residue. We also calculated the relative buried surface area (termed "RBSA"). The detailed definitions are shown in Fig. 1A and SI. Such a definition not only generates a quantity, RLA, that has the same dimension and physical meaning as RSA and RASA, but also accounts for the more complex nature of lipid molecules than water molecules in accessing the surface residues of proteins. This dataset was then used to train the deep neural network model ProtRAP, as shown in Figs. 1B and 1C (please refer to the SI for details).

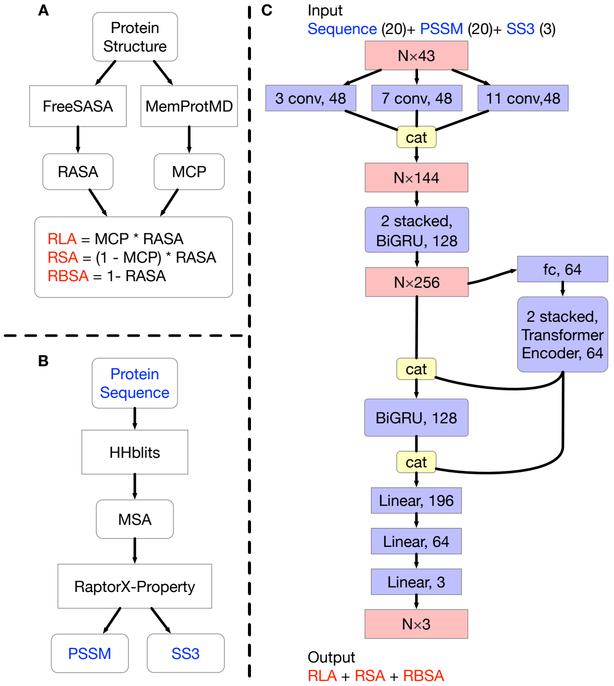


Figure 1. Our protocol for protein accessibility prediction. (A) Generation of the training dataset. Protein structures were analyzed to get the RASA and the membrane contact probability (MCP), which were further used to calculate the RLA, RSA, and RBSA of each residue for the output. (B) Preprocessing pipeline for the input. From a protein sequence, the MSA was obtained by HHblits, then the Position Specific Scoring Matrix (PSSM) and the predicted three-state secondary structure (SS3) were obtained by RaptorX-Property. (C) Architecture of the deep neural network model ProtRAP.

ProtRAP can predict the RLA and RSA of residues of a given protein sequence in one run, which can tell which residues are accessible to lipids (high RLA), to solvent (high RSA), or buried in protein interior (low RLA and low RSA, thus high RBSA). With the high-quality dataset and refined model, ProtRAP showed very satisfactory performance. Three test sets were used to evaluate the model performance, the first one containing 140 membrane proteins (Mem), the second one containing 733 soluble proteins (Sol), and a mixture of the above two (Mix). As can be seen in Table S1, the overall prediction performance, as measured by the Pearson Correlation Coefficient (PCC) with respect to the true observations, reached 0.7$\sim$0.8 for RLA and RSA. Particularly, the prediction of RLA for membrane proteins showed a high PCC of above 0.77. The ten-fold cross validation shows very low standard deviations, indicating that our model and dataset are robust.

To better understand the performance of ProtRAP, we compared it with two related methods that predict accessibility of protein residues (Table 1 and Table S2). The first method, RaptorX-Property, predicts the RSA of residues from a given protein sequence with a deep learning model DeepCNF (Deep Convolutional Neural Fields).[8] The second method, NetSurfP 2.0 is a state-of-the-art method that predicts the RASA with an architecture composed of convolutional and long short-term memory neural networks.[11] According to the data in Table 1, it appears that NetSurfP 2.0 and RaptorX-Property do a great job in predicting RASA and RSA, respectively, but none of them is able to predict RLA. Our model performed very well for the prediction of RSA of both membrane and soluble proteins, with an accuracy comparable to that of RaptorX-Property. Although our ProtRAP did not directly predict RASA, the formula $RASA = RLA + RSA$ can be used to obtain the RASA. As can be seen in Table 1, our ProtRAP is comparable to NetSurfP 2.0 in predicting the RASA of membrane and soluble proteins. Most importantly, ProtRAP can predict the RLA of membrane proteins, a function that is absent in the other two models. Therefore, from a given protein sequence, ProtRAP is able to accurately predict the RLA and RSA of residues for both soluble and membrane proteins in one run, which can then be used to infer the RASA (RLA+RSA) and RBSA (1-RASA) of residues, providing more complete and quantitative information on the exposure of amino acid residues in a protein structure.

Table 1. Comparison of protein accessibility predictions using three methods.

		RaptorX-Property	NetSurfP 2.0	ProtRAP
Mem_RASA	PCC	-	0.679	0.708
	MAE	-	0.143	0.133
	Q3	0.487	-	0.612
Mem_RSA	PCC	-	0.533	0.744
	MAE	-	0.189	0.108
	Q3	0.661	-	0.684
Mem_RLA	PCC	-	-	0.779
	MAE	-	-	0.052
Sol_RASA	PCC	-	0.765	0.755
	MAE	-	0.128	0.131
	Q3	0.640	-	0.643

To give a more intuitive picture, we selected several representative examples to showcase the ProtRAP prediction mapped onto protein structures. The first case is a soluble protein, the dextranase from *Streptococcus mutansisomaltase* (PDB ID: 3vmn).[24] From its sequence, we predicted its RLA, RSA and RBSA, then showed the results in the left of Fig. 2A (black lines), along with the ground truth (red lines) as well as the difference (*Diff*) between the prediction and ground truth (blue dotted lines, $Diff = Truth - Prediction$). The first thing to notice is that, the prediction did not exhibit any false positive for RLA. Also, it is clear that the predictions overlap very well with the ground truth, indicated by the difference (blue dotted lines) fluctuating around 0. This can also be seen in the right panel of Fig. 2A, where the predicted and true RSA and RLA are mapped onto the structure and show similar color distributions. These results indicate that our RSA prediction of soluble proteins from their sequences can indeed provide valuable information on the localization of residues in the structures, and our RLA predictions do not tend to generate false positives. Indeed, the low MAE values for the RLA prediction of soluble proteins confirmed that the soluble proteins residues are not likely to be predicted to be exposed to lipid molecules by ProtRAP (Table S1).

The second case is an α-helical dimeric transmembrane protein, the OSCA channel from *Arabidopsis thaliana*, which is a relatively large (1202 amino acids) and hom-dimeric complex membrane protein (PDB ID: 5z1f).[25] Again, the ProtRAP performed well in predicting both the RLA and RSA, as shown in Fig. 2B. The difference between the prediction and ground truth shows an overall flat line fluctu-

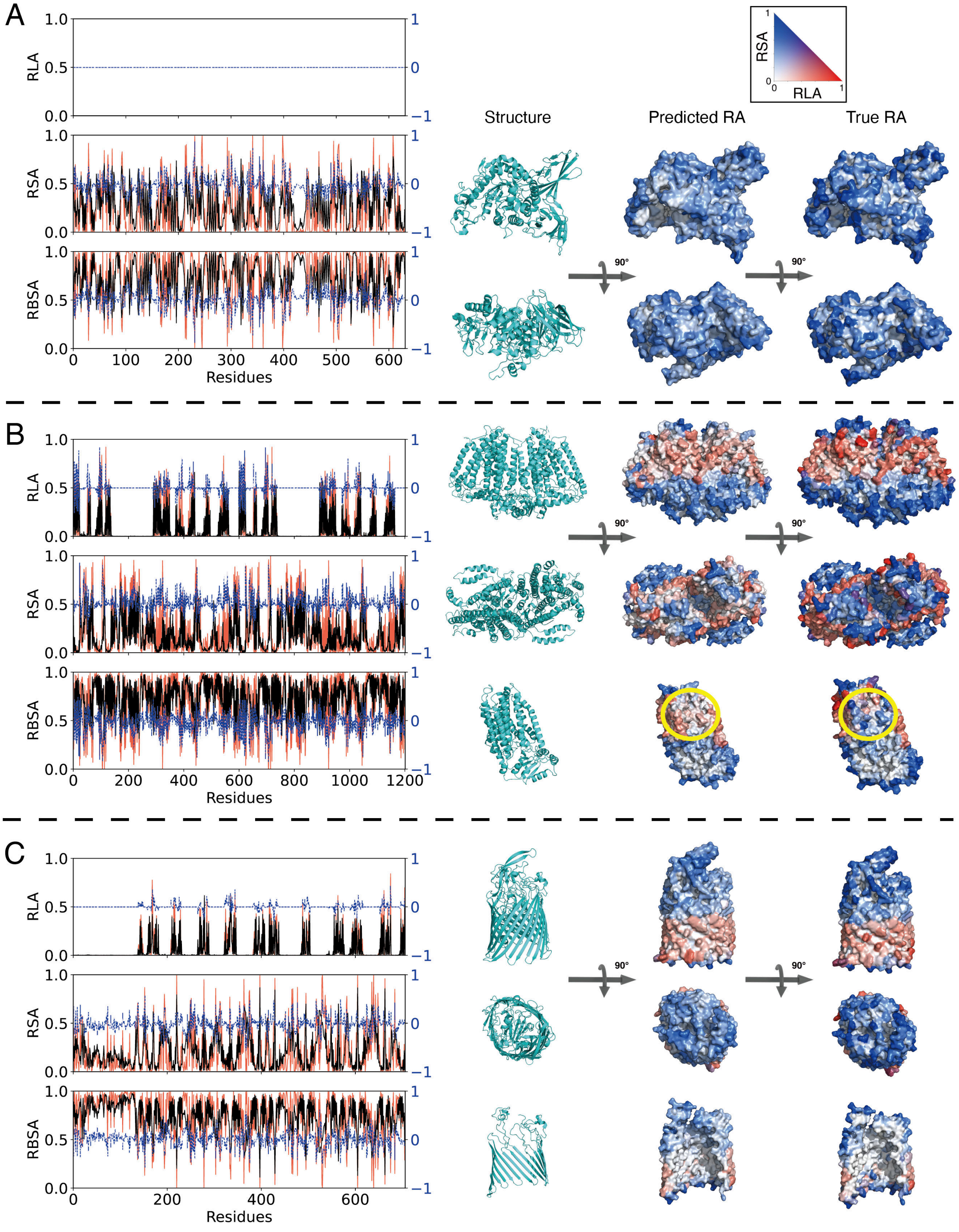


Figure 2. Accessibility prediction for three representative cases. (A) Predicted accessibility for a soluble protein. Left panels show the predicted accessibility (black line), the true accessibility (red line), and the prediction error (blue dotted line, $Truth - Prediction$) for each residue. Right panels show the structure of the protein (PDB ID: 3vmn), represented by cartoon and surface. The surface representation is colored according to the predicted relative accessibility and true relative accessibility. The color scheme is shown at top right. (B) Similar to (A) but for an α-helical transmembrane protein OSCA, whose PDB ID is 5z1f. The bottom panel on the right shows the dimeric interface. (C) Similar to (A) but for a β-barrel transmembrane protein (PDB ID: 5fr8). The bottom panel on the right shows the interior of the protein.

ating around 0 (Fig. 2B left), and a clear lipid accessible, transmembrane domain is observed on the outer surface when the prediction is mapped onto the structure, which is consistent with the ground truth (Fig. 2B right). Interestingly, the ProtRAP prediction shows a lipid accessible area on the dimer interface (yellow circle on the right bottom structure), indicating that the interface may be filled with lipid molecules. The ground truth obtained from MemProtMD shows a similar but less definitive lipid accessible area near the same spot (yellow circle). In fact, two previous structural and MD simulation studies found that the dimeric interface contains a groove that can indeed be filled by lipid molecules, which confirmed the ProtRAP prediction.[25,26] To exclude the possibility that this may represent a failure of ProtRAP in predicting the accessibility at the oligomeric interfaces, we further looked into another membrane protein, the eukaryotic CLC transporter (CmCLC), which has a similar architecture to OSCA (α-helical and dimeric) but does not have a groove at the dimeric interface (PDB ID: 3org).[27] The results are shown in Fig. S1, and it is clear that the dimeric interface is predicted to be buried in this case (yellow circle). Therefore, it appears that the ProtRAP can accurately predict the RLA of α-helical membrane proteins, and the oligomeric interface can be identified as well, no matter whether it is buried or lipid accessible in the transmembrane region. The RSA prediction is satisfactory too, without false positives of RLA for the non-transmembrane regions.

The third representative case is a wide β-barrel transmembrane protein, the siderophore receptor PirA from *Acinetobacter baumannii* (PDB ID: 5fr8).[28] Again, the prediction performance is pretty good, with an overall horizontal line fluctuating around 0 representing the difference between the prediction and ground truth, and no false positive RLA predictions for the non-transmembrane region (Fig. 2C left). A well-defined transmembrane region is observed on the outer surface, which is consistent with the ground truth (red region in Fig. 2C right). As this β-barrel transmembrane protein also has a large pore interior, we examined the accessibility of the inner surface. The right bottom of Fig. 2C shows that the interior residues are predicted to be buried or weakly solvent accessible, but not lipid accessible. This is consistent with the observation in coarse grained MD simulations (considered as ground truth in this study). This result confirms the ability of ProtRAP in predicting the surface residues of β-barrel transmembrane proteins, another large family of membrane proteins.

The sequences of the above four representative cases, colored according to the predicted accessibilities, are shown in Figs. S2. It should be noted that the above examples were not the top ranked cases in terms of prediction accuracy in our test sets, as measured by the weighted mean absolute error (wMAE, whose definition is similar to the loss function in SI) of each protein. The prediction accuracy of the dextranase ranks 570th in the 733-soluble-protein test set; the accuracy of the OSCA channel ranks 99th, the accuracy of the CLC transporter ranks 18th, and the PirA receptor ranks 24th in the 140-membrane-protein test set, respectively. Still, from these cases, we can see that ProtRAP can accurately predict lipid accessible residues that form the transmembrane regions in various situations, and does not show trend of false positive predictions of lipid accessibility for soluble proteins or water exposed regions of membrane proteins. In the meanwhile, the solvent accessible residues are accurately predicted as well. Taken together, we believe ProtRAP can be used to annotate the localization of protein residues from a given protein sequence, which can be further utilized to evaluate or even refine the predicted or modeled protein structures.

Another notable feature of ProtRAP is that it predicts RLA with high specificity, especially for soluble proteins. As the protein sequence is the only required input and the prediction for one sequence only costs about one minute (mainly cost in MSA), ProtRAP can be used for high throughput screening to discover potential membrane proteins from proteomes. Naturally, ProtRAP can also help identify the potential residues of a given protein that directly interact with membranes, which would be useful for integrative structure biology studies.

There is still room for further improvement, though. Our model is not optimized for multimer prediction. It can only take one chain as input when predicting, and cannot consider other subunits of the multimer. But the case studies of the OSCA channel and CLC transporters seem reassuring, showing that ProtRAP can capture the difference between the dimer interface and outer surface. The other known problem is that, some mutations of specific amino acids in experiments can lead to great changes in protein structure and localization, but this is difficult for our model to correctly predict at the moment. This may be attributed to the nature of deep learning model used here in exploring the multiple sequence alignment and the usage of statistical PSSM for the prediction, which can eliminate the effect of specific mutations. Therefore, an improved model that is more sensitive to specific mutations would be highly desired in the future, which will be valuable for (membrane) protein design.

Acknowledgement This work was supported by grants from the Ministry of Science and Technology of China (2021YFE0108100) and National Natural Science Foundation of China (21873006 and 32071251).

Server Availability An online computation server is available for tests and predictions at: `http://www.songlab.cn`.

References

(1) Lee, B.; Richards, F. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology* **1971**, *55*, 379–400.

(2) Shrake, A.; Rupley, J. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology* **1973**, *79*, 351–371.

(3) Chothia, C. The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology* **1976**, *105*, 1–12.

(4) Eyal, E.; Najmanovich, R.; McConkey, B. J.; Edelman, M.; Sobolev, V. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *Journal of Computational Chemistry* **2004**, *25*, 712–724.

(5) Savojardo, C.; Babbi, G.; Martelli, P. L.; Casadio, R. Functional and structural features of disease-related protein variants. *International Journal of Molecular Sciences* **2019**, *20*, 1530.

(6) Gromiha, M. M.; Ahmad, S. Role of solvent accessibility in structure based drug design. *Current Computer-Aided Drug Design* **2005**, *1*, 223–235.

(7) Ludwig, C.; Michiels, P. J. A.; Lodi, A.; Ride, J.; Bunce, C.; Gunther, U. L. Evaluation of solvent accessibility epitopes for different dehydrogenase inhibitors. *Chemmedchem* **2008**, *3*, 1371–1376.

(8) 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.

(9) Heffernan, R.; Yang, Y.; Paliwal, K.; Zhou, Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* **2017**, *33*, 2842–2849.

(10) Zhang, B.; Li, L.; Lü, Q. Protein solvent-accessibility prediction by a stacked deep bidirectional recurrent neural network. *Biomolecules* **2018**, *8*, 33.

(11) Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Soenderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B., et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics* **2019**, *87*, 520–527.

(12) Goddard, A. D.; Dijkman, P. M.; Adamson, R. J.; dos Reis, R. I.; Watts, A. *Membrane proteins - production and functional characterization*; Methods in Enzymology; 2015; Vol. 556; pp 405–424.

(13) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nature Reviews Drug Discovery* **2002**, *1*, 727–730.

(14) Dailey, M. M.; Hait, C.; Holt, P. A.; Maguire, J. M.; Meier, J. B.; Miller, M. C.; Petraccone, L.; Trent, J. O. Structure-based drug design: From nucleic acid to membrane protein targets. *Experimental and Molecular Pathology* **2009**, *86*, 141–150.

(15) Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **2015**, *347*, 1260419.

(16) Wang, L.; Zhang, J.; Wang, D.; Song, C. Membrane contact probability: An essential and predictive character for the structural and functional studies of membrane proteins. *PLoS Computational Biology* **2022**, *18*, 1–27.

(17) Adamian, L.; Liang, J. Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Structural Biology* **2006**, *6*, 1–17.

(18) Lai, J.-S.; Cheng, C.-W.; Lo, A.; Sung, T.-Y.; Hsu, W.-L. Lipid exposure prediction enhances the inference of rotational angles of transmembrane helices. *BMC Bioinformatics* **2013**, *14*, 1–16.

(19) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology* **2001**, *305*, 567–580.

(20) Jones, D. T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **2007**, *23*, 538–544.

(21) Hayat, S.; Elofsson, A. BOCTOPUS: Improved topology prediction of transmembrane β barrel proteins. *Bioinformatics* **2012**, *28*, 516–522.

(22) Mulnaes, D.; Schott-Verdugo, S.; Koenig, F.; Gohlke, H. TopProperty: Robust metaprediction of transmembrane and globular protein features using deep neural networks. *Journal of Chemical Theory and Computation* **2021**, *17*, 7281–7289.

(23) Newport, T. D.; Sansom, M. S. P.; Stansfeld, P. J. The MemProtMD database: A resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Research* **2018**, *47*, D390–D397.

(24) Suzuki, N.; Kim, Y.-M.; Fujimoto, Z.; Momma, M.; Okuyama, M.; Mori, H.; Funane, K.; Kimura, A. Structural elucidation of dextran degradation mechanism by Streptococcus mutans dextranase belonging to glycoside hydrolase family 66. *Journal of Biological Chemistry* **2012**, *287*, 19916–19926.

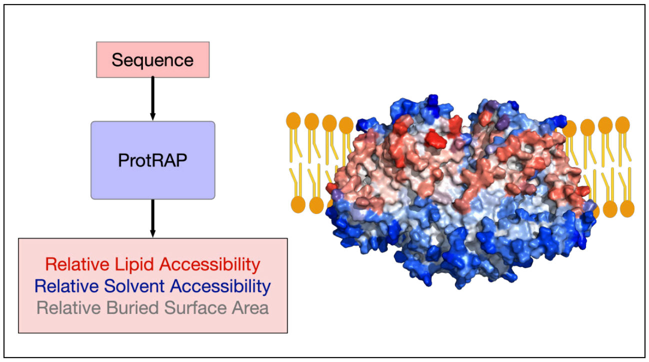
(25) Zhang, M.; Wang, D.; Kang, Y.; Wu, J.-X.; Yao, F.; Pan, C.; Yan, Z.; Song, C.; Chen, L. Structure of the mechanosensitive OSCA channels. *Nature Structural & Molecular Biology* **2018**, *25*, 850–858.

(26) Jojoa-Cruz, S.; Saotome, K.; Murthy, S. E.; Tsui, C. C. A.; Sansom, M. S.; Patapoutian, A.; Ward, A. B. Cryo-EM structure of the mechanically activated ion channel OSCA1.2. *eLife* **2018**, *7*, e41845.

(27) Feng, L.; Campbell, E. B.; Hsiung, Y.; MacKinnon, R. Structure of a eukaryotic CLC transporter defines an intermediate state in the transport cycle. *Science* **2010**, *330*, 635–641.

(28) Moynié, L.; Luscher, A.; Rolo, D.; Pletzer, D.; Tortajada, A.; Weingart, H.; Braun, Y.; Page, M. G.; Naismith, J. H.; Köhler, T. Structure and function of the PiuA and PirA siderophore-drug receptors from Pseudomonas aeruginosa and Acinetobacter baumannii. *Antimicrobial agents and chemotherapy* **2017**, *61*, e02531–16.

TOC Graphic


Sequence
ProtRAP
Relative Lipid Accessibility
Relative Solvent Accessibility
Relative Buried Surface Area