HHS Public Access

Author manuscript

*J Chem Inf Model*. Author manuscript; available in PMC 2017 August 01.

Published in final edited form as:

*J Chem Inf Model*. 2016 February 22; 56(2): 423–434. doi:10.1021/acs.jcim.5b00517.

# Accurate Prediction of Contact Numbers for Multi-Spanning Helical Membrane Proteins

**Bian Li**[†,‡], **Jeffrey Mendenhall**[†,‡], **Elizabeth Dong Nguyen**[‡], **Brian E. Weiner**[†,‡], **Axel W. Fischer**[†,‡], and **Jens Meiler**[*,†,‡]

[†]Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37232, United States

[‡]Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37232, United States

## Abstract

Prediction of the three-dimensional (3D) structures of proteins by computational methods is acknowledged as an unsolved problem. Accurate prediction of important structural characteristics such as contact number is expected to accelerate the otherwise slow progress being made in the prediction of 3D structure of proteins. Here, we present a dropout neural network-based method, TMH-Expo, for predicting the contact number of transmembrane helix (TMH) residues from sequence. Neuronal dropout is a strategy where certain neurons of the network are excluded from back-propagation to prevent co-adaptation of hidden-layer neurons. By using neuronal dropout, overfitting was significantly reduced and performance was noticeably improved. For multi-spanning helical membrane proteins, TMH-Expo achieved a remarkable Pearson correlation coefficient of 0.69 between predicted and experimental values and a mean absolute error of only 1.68. In addition, among those membrane protein–membrane protein interface residues, 76.8% were correctly predicted. Mapping of predicted contact numbers onto structures indicates that contact numbers predicted by TMH-Expo reflect the exposure patterns of TMHs and reveal membrane protein–membrane protein interfaces, reinforcing the potential of predicted contact numbers to be used as restraints for 3D structure prediction and protein–protein docking. TMH-Expo can be accessed via a Web server at www.meilerlab.org.

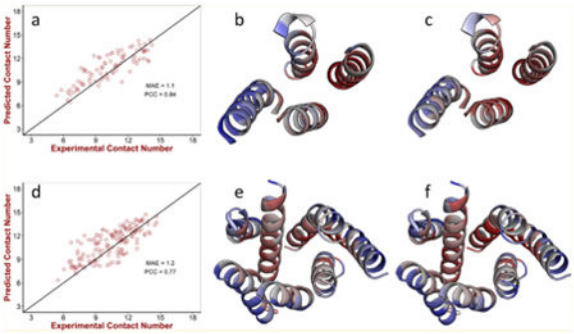
## Graphical abstract

[*]Corresponding Author: jens.meiler@vanderbilt.edu. Phone: +1 (615) 936-5662. URL: www.meilerlab.org.

Supporting Information: The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00517.

**Author Contributions**: The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript

**Notes**: The authors declare no competing financial interest.




## Introduction

Helical membrane proteins (HMPs) play essential roles in various biological processes, including signal transduction, ionic and molecular transportation across the membrane, and energy generation. Due to their pharmacological relevance, about 50% of the drugs in the market target HMPs.[1] It was estimated that HMPs constitute about 20% to 30% of the human genome.[2] In spite of their prevalence in the genome, a very small portion of structures in the Protein Databank is HMPs due to the experimental difficulties in determining the structures of HMPs. Therefore, accurate and efficient computational methods would be valuable tools to complement existing experimental techniques. One of the challenges in computational prediction of the three-dimensional (3D) structure of HMPs is to predict helix-helix packing in which a transmembrane helix (TMH) either faces the lipids or is buried in the protein core. Knowing *a priori* whether an amino acid residue is exposed to the membrane lipid or buried inside the protein core provides valuable restraint information that can be incorporated to reduce the sampling space of helix–helix packing. As an intermediate step to the prediction of the 3D structure of HMPs, it is worthwhile to develop reliable methods for predicting residue exposure.

Solvent accessibility is the most commonly used structural feature for characterizing the exposure environment of a residue.[3] However, the applicability of solvent accessibility in helix–helix packing, or *de novo* 3D structure prediction, where an astronomical conformational space needs to be sampled, is limited. Accurate computation of solvent accessibility is computationally demanding because it requires full-atom representation of amino acid side chains. Residue contact number, defined as the number of contacting residues of the residue of interest is another structural feature that reflects the exposure of a residue.[4] Computation of the contact number does not require a full-atom representation of amino acid side chains and is numerically fast. Thus, contact number is more suitable for being incorporated into the 3D structure prediction either in the form of restraints or knowledge-based potential. In addition, as contact number is negatively correlated with solvent accessibility,[5] it may provide insights into a spectrum of biological problems in which solvent accessibility has been applied, such as epitope mapping,[6] hot spots detection,[7,8] understanding of protein–protein interactions,[9–11] quality assessment of structural models,[12] and modeling of amino acid residue side chain conformation.[13]

Traditionally, prediction of the contact numbers for soluble proteins is treated as a two-state (higher or lower than the average contact number) or three-state (much higher, much lower, or close to the average contact number) classification problem.[14–16] However, the applicability of classification approach is limited as it is difficult to use discrete exposure status for scoring in 3D structure prediction. Furthermore, subdividing residues into different states requires an arbitrary selection of a specific contact number as a cutoff. Therefore, real-value predictions should be preferred.[17] The problem of predicting contact numbers for soluble proteins has been studied for more than a decade, and promising results have been achieved.[18,19] Even though a few attempts have been made to predict the burial status or real-value solvent accessibility of TMH residues,[20–23] given the fact that 3D structures of HMPs have long been desirably pursued, it is remarkable to notice that no work has been reported on predicting contact numbers for HMPs.

Here, we present a dropout neural network-based method, termed TMH-Expo, for predicting contact numbers for HMPs. We first curated a large nonredundant data set of HMPs with known structure based on which experimental contact numbers were computed. Thereafter, we examined a set of feature vectors containing local sequence or evolutionary information for contact number prediction. Subsequently, a detailed analysis of the performance of TMH-Expo was conducted. Finally, we showed that the predicted contact number reveals exposure patterns of TMHs and discussed the application of the predicted contact number to 3D structure prediction and protein–protein docking.

## Methods

### Generation of Data Set

The data set of HMPs with known structures used in the current study was retrieved from the OPM (Orientation of Proteins in the Membrane) database.[24] Peripheral HMPs and peptides were removed to obtain a set of “true” HMPs. Further refinement was carried out by removing thylakoid HMPs as they have extreme topological complexity.[25] The protein culling server PISCES[26] was used to obtain a list of HMP chains that have a sequence length between 40 and 10,000 residues, and pairwise sequence identity of 25% or less. Non-X-ray structures and $C_\alpha$-only structures, as well as X-ray structures with a resolution > 3.0 Å or an R-factor > 0.3, were excluded. This culminated the final data set that consisted of 90 chains from 71 proteins from 33 OPM superfamilies. The complete list of protein chains used in this study are listed in Table S1 of the Supporting Information. The transmembrane region for each protein chain was provided by OPM. The membrane normal aligns with the $z$-axis, and the membrane center is positioned at $z = 0$. A secondary structure type was assigned to each residue from the consensus identification of DSSP,[27] Stride,[28] and PALSSE.[29] A residue is considered as a TMH residue if it sits inside the membrane, and the residue is part of a helical conformation.

### Computation of Contact Number

The contact number of a residue $i$ was originally defined as the number of $C_\alpha$ atoms of other residues inside the sphere of radius $d$ centered at the $C_\alpha$ atom of residue $i$.[30] While this definition is straightforward, it has the disadvantage that each residue inside the sphere is

assigned an equal contribution to the total contact number. This is rather unrealistic as both van der Waals and electrostatic interactions are distance dependent. To achieve a more physically realistic approximation, we used a refined algorithm developed for contact number computation. This algorithm is similar to that of Kinjo et al.,[19] where $C_\beta$ atoms are used instead of $C_\alpha$ atoms and the boundary of the sphere is smoothed. Contribution to the total contact number is assigned to each residue inside the sphere in a distance-dependent way such that short-range contacting residues have higher contribution than long-range contacting ones. Residues whose $C_\beta$ atom is within 4.0 Å to the $C_\beta$ atom of the residue of interest are assigned a weight of 1.0; those with a distance longer than 11.4 Å are assigned a weight of 0. Any residue in between is assigned a weight between 0.0 and 1.0 according to a smooth transition function. This scheme can be summarized into the following function:

$$W_{ij}=\begin{cases} 1, d_{ij}<0 \\ \frac{1}{2}\cos\left(\frac{d_{ij}-l}{u-l}\times\pi\right)+\frac{1}{2}, l<d_{ij}<u \\ 0, d_{ij}\geq u \end{cases}$$

where $w_{ij}$ is the contribution made by residue $j$ to the total contact number of residue $i$, $d_{ij}$ is the distance between the $C_\beta$ atoms of residue $i$ and residue $j$, $l$ is the lower bound of $d_{ij}$ within which $w_{ij}$ = 1.0, and $u$ is the upper bound of $d_{ij}$ beyond which $w_{ij}$ = 0. For glycine, $H_{a2}$ is used in place of $C_\beta$ atom. The lower and upper bound are optimized values such that the correlation between the contact number and solvent accessible surface area (SASA) is maximized. Only residues separated by more than three residues along the sequence are considered in the calculation to reduce the bias due to sequence proximity. The total contact number of residue $i$ was computed by summing $w_{ij}$ over the entire protein:

$$\mathrm{CN}_i=\sum_{j\in|j-i|>3}^{n} w_{ij}$$

where $n$ is the length of the protein chain in the case of computing the monomeric contact number or the total number of residues in the protein for computing the oligomeric contact number. All nonprotein molecules were removed before computing the contact numbers. Nonprotein molecules such as coenzymes, ligands, and internal waters play important roles in the function of membrane proteins. However, the biochemical identity of the interface between these molecules and membrane proteins requires detailed analysis and is beyond the scope of this study.

## Computation of Relative Solvent Accessibility

The relative solvent accessibility (RSA) of a residue was computed as the ratio between the absolute solvent accessibility (ASA) observed in the native structure and that in an extended tripeptide conformation (A-X-A). The ASA values were computed based on the oligomeric states provided by OPM using DSSP with a probe radius of 1.4 Å[27] as with previous studies. [15,17,31,32] No further exploration on probe sizes was conducted because it has been

shown that probe size has little or no effect on the performance of RSA predictors.[23] The ASA value of each amino acid type in an extended tripeptide conformation was adopted from a similar study.[17]

## Computation of Feature Vectors

The multiple sequence alignment (MSA) for each protein sequence in the data set was obtained by searching the UniRef50[33] nonredundant sequence database with PSI-BLAS or five iterations.[34] The E-value inclusion threshold was set to 0.01. A floating point-valued position-specific scoring matrix (PSSM) was generated from PSI-BLAST checkpoint files using the source code (chkparse.c) adapted from PSIPRED.[35] Floating point-valued PSSM was preferred over integer-valued PSSM as the former provides higher precision. PSSM is an $L \times 20$ matrix, where $L$ denotes sequence length. For each sequence position $i$, there are 20 entries, each corresponding to the score of one of the 20 naturally occurring amino acids. The BLAST probability profile (BPP) for amino acid $j$ at sequence position $i$ was computed by transforming each PSSM entry $m_{ij}$ using the following equation:

$$P_{ij}=\frac{10^{m_{ij}/10}}{\sum_{j}^{20}10^{m_{ij}/10}}$$

where $j$ runs from 1 to 20. The variance-based conservation index $CI_i$ is one of the commonly used conservation indices and is defined by the following formula:

$$\mathrm{CI}_i=\sqrt{\sum_{j}^{20}(p_{ij}-p_j)^2}$$

where the summation is carried out over 20 amino acids, $p_{ij}$ is the BLAST probability of amino acid $j$ at position $i$ such that $\sum_{j}^{20}P_{ij}=1$, and $p_j$ is the average BLAST probability of amino acidj and is defined as $1/L\sum_{i}^{L}P_{ij}$. The amino acid type at each sequence position is encoded by a vector with 20 binary entries (or 20 bits). When considering a window size of $w$ centered at the residue whose contact number is to be predicted, the feature vector computed based on PSSM, BPP, or local sequence composition (LSC) has a total of $w \times 20$ components, whereas the feature vector computed based on CI has a total of $w \times 1$ components (Figure S1, Supporting Information).

## Training of Dropout Neural Networks with Back-Propagation of Errors

The support vector machine (SVM)algorithm has been applied to various bioinformatics tasks,especially solvent accessibility and contact number prediction.[18,21-23] It has the benefit of being less prone to overfitting than neural networks. Indeed, our preliminary test showed that neural networks trained without dropout (learning rate $\eta = 0.1$, momentum factor $\alpha = 0.1$, number of hidden layer neurons = 64, and number of epochs = 500) had a MAE (mean absolute error, see Performance Measures for details) of 2.70, whereas an optimized SVM

(radial basis function kernel, $\gamma$ = 0.025, cost = 0.1) had a MAE of 1.76. However, neural networks trained with dropout (learning rate $\eta$ = 0.1, momentum factor $\alpha$ = 0.1, number of hidden layer neurons = 64, number of epochs = 500, dropout rate in input layer = 0.05, and dropout rate in output layer = 0.5) had a MAE of 1.69. As dropout neural networks had a smaller MAE, we thus chose dropout neural networks as the learning algorithm in the current study.

The dropout neural networks trained in this study were fully connected three-layer feed-forward networks with a sigmoid activation function (Figure 1a). The input layer contained one unit for each component in the feature vector. Inputs to the network are either local sequence information or evolutionary information derived from PSI-BLAST computed MSAs. The window size used for computing feature vectors was set to 15, an optimal value for contact number prediction found in our preliminary testing. The output layer was composed of a single node for the residue-specific contact number or RSA The hidden layer was composed of 64 neurons. A random of 5% of units in the input layer and 50% of neurons in the hidden layer were dropped during each presentation of each training case. The networks were trained with resilient back-propagation of errors[36] with the learning rate $\eta$ set to 0.1 and momentum factor $\alpha$ set to 0.1. Weights were updated after presentation of each residue to the network. A maximum of 2000 epochs were applied.

## Jackknife Cross-Validation

A relatively low sequence identity (25%) was used in the current study; however, such low sequence identity alone might not be sufficient to exclude homology among protein chains. In fact, substantial remote homology could still exist at this level placing HMPs in the same structural superfamily.[37] Such remote homology between proteins in the training set and proteins in the validation set for testing the model can lead to an overoptimiztic estimate of the performance of the network for new structural families. As a way of preventing such overoptimism, the data set was partitioned such that each OPM superfamily forms its own subset that contains all its members and no members from other OPM superfamilies. Cross-validation of the networks was done in a jackknife manner with respect to a OPM superfamily. Of the 33 OPM superfamilies, one single OPM superfamily was withheld as the validation set for evaluating the neural networks. Then, a 5-fold cross-validation protocol adopted for our transmembrane span and secondary structure prediction algorithm[38] was carried out on the remaining 32 superfamilies (Figure 1b). This process was then repeated 33 times, with each of the 33 OPM superfamilies used exactly once as the validation set. Predictions for the 33 validation sets were combined to give the final estimate of the performance of the neural networks.

## Performance Measures

A set of performance measures were adopted to evaluate the performance of the neural networks. The primary measure was the Pearson correlation coefficient (PCC) between experimental and predicted contact numbers and RSA For a set of $n$ data points $(x_i, y_i)$, the PCC was computed as follows:

$$\mathrm{PCC}=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}}$$

For comparing our results to that from previous studies, we incorporated the following measures that are commonly used to evaluate classifiers:

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}$$

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

where MCC is the Matthews correlation coefficient,[39] FPR is the false positive rate, TP is the number of correctly predicted buried residues, TN is the number of correctly predicted exposed residues, FP is the number of incorrectly predicted buried residues, and FN is the number of incorrectly predicted exposed residues. The real-value contact number and RSA were transformed to binary states using the median as a cutoff such that the data set is equally partitioned. The mean absolute error (MAE), which is defined as the per-residue absolute difference between the experimental and predicted contact number and RSA, was used to evaluate prediction errors:

$$\mathrm{MAE}=\frac{\sum|v_{\mathrm{experimental}}-v_{\mathrm{predicted}}|}{n}$$

where $v$ is either RSA or contact number, and $n$ is the total number of residues to be predicted. The summation is carried out over all predicted residues.

## Results and Discussion

### Statistics of the Data Set

The repository of HMPs with known structures has expanded tremendously in recent years. It was reported that the latest number of unique membrane protein structures deposited in the Protein Databank is 535 (http://blanco.biomol.uci.edu/mpstruc/) compared to about 150 in 2005.[40] Curation of a data set that is representative of the population is an essential step in producing a model with high predictive accuracy. We compared the data set used to train TMH-Expo to those used in two related works, namely, $\mathrm{ASAP}_{\mathrm{mem}}$[21] and MPRAP.[23] In terms of the size of data sets, the TMH-Expo data set consists of 71 HMPs (90 unique

chains), significantly larger than the $ASAP_{mem}$ data set (also known as the Beuming—Weinstein or BW data set[20]), which has 28 HMPs (59 unique chains). The MPRAP data set has 52 HMPs (80 unique chains). Interestingly, PISCES returned only 34 HMPs (60 unique chains) from the MPRAP data set using the same criteria applied to cull the TMH-Expo data set.

Table 1 lists the frequency and mean contact number, as well as standard deviation of contact number, for each amino acid residue type. Similar to observations made by Ulmschneider and co-workers,[41] residues with nonpolar side chain such as Ala, Phe, Ile, Leu, and Val are dominantly abundant. In addition, except in the case of Ala, their mean contact numbers are not significantly higher than that of other amino acid residues. In fact, the mean contact numbers for Phe, Ile, Leu, and Val are among the lowest, an expected observation given the fact that the membrane provides an environment that is more hydrophobic than the protein interior. On the other end, the mean contact numbers for Ala, Cys, Gly, and Ser are among the highest, suggesting that on average helices enriched with these residues are more densely packed. In fact, Ala, Gly, and Ser are known to form the sequence motifs of the type AxxxA, GxxxG, and SxxxS that are believed to promote close helical packing.[42]

## Relevance of Input Features

The performance of a data-trained machine learning method depends crucially on the judicious choice of the feature vector. For solvent accessibility prediction, feature vectors containing primary sequence information or evolutionary information have been tested.[17,18,22,32] Four feature vectors, CI, LSC, BPP, and PSSM, were investigated in this study. CI, BPP, and PSSM can be considered as evolutionary information-containing feature vectors as they are derived based on MSA, whereas LSC contains purely primary sequence information. We initially examined the correlation coefficient of all features computed considering a window size of 41 (residues from $i - 20$ to $i + 20$, where $i$ is the position of the residue of interest, inclusive) with contact numbers. This resulted in 41 × 1, 41 × 20, 41 × 20, and 41 × 20 entries for feature vectors of CI, LSC, BPP, and PSSM, respectively (Figure S1, Supporting Information). Figure 2 plots the correlation coefficients of entries in each feature vector with contact numbers. For sequence-based prediction, it is well known that the use of evolutionary information derived from MSA improves prediction performance. In fact, on average, CI, BPP, and PSSM show stronger correlation with contact numbers than local sequence composition does (compare Figure 2a, c, and d with b). It is also interesting to note that PSSM generally has more strongly correlated entries than either of the other two evolutionary information-containing feature vectors do (compare Figure 2d with a and c).

## Choosing the Optimal Window Size

One further observation made from Figure 2 is that features computed from neighboring residues are substantially correlated with the contact number of the central residue, and the correlation is dependent on sequence separation. Correlation coefficient decays gradually from very strong at the central residue to very weak at a separation of 15 or more residues. This suggests that there should be an optimal window size such that the signal-to-noise ratio is maximized. For solvent accessibility or contact number prediction, window sizes of 7,[17]

9,[23] 11,[43] 15,[18,22] 17,[44] and 21[19] have been used in previous studies. These window sizes are either arbitrarily chosen or obtained by optimization over a relatively short-range. We tested a wide spectrum of window sizes ranging from 1 to 41 with a step size of 2. In this preliminary testing, the feature vector was PSSM, and the architecture of the networks was kept the same across all window sizes.

Figure 3a shows the effect of windows sizes on the performance of neural networks. As window size increases from 1 to 9, MAE decreases drastically from above 2.0 to below 1.8, a trend similar to the observation made by Park et al.[22] As the window size increases from 9 to 15, MAE follows a decreasing trend that is slight but noticeable. MAE rises gradually as the window size is further extended to beyond 21. Interestingly, MAEs for window sizes from 15 to 21 remain essentially identical. It was previously proposed that the identities of the residues lying just above ($i + 4$) and below ($i - 4$) the target residue on the same helix face are most indicative of the burial status of the central residue.[22] However, our observation suggests that including up to seven neighboring residues on either side of the central residue consistently improves the performance of the neural network (Figure 3b). The fact that MAE reaches its lowest value when the window size is 15 is especially intriguing given that heptad repeat is one of the signature patterns in helix–helix interactions.[45] In fact, Adamian et al. developed a highly accurate method for predicting helix–lipid interfaces using heptad motifs as a structural template to assign helical faces of each helical residue.[46] However, whether the optimal window size arises from heptad repeat needs further investigation.

## Dropout Prevents Overfitting and Improves Performance

Neuronal dropout is a technique developed for addressing the overfitting problem in neural networks where a large number of parameters are optimized. The key idea is to randomly drop neurons along with their connections from the neural network during each presentation of each training case (Figure 1a).[47] With this training feature, hidden neurons are prevented from co-adapting too much and forced to build a relatively independent mapping from feature space onto output space. It has been demonstrated that dropout reduces overfitting and improves performance of neural networks on classification tasks in speech recognition and handwritten digit classification.[47–49]

In order to confirm that dropout reduces overfitting and improves the performance of neural networks for contact number prediction, we compared performances of networks trained with and without dropout. As shown in Figure 4, compared to the performance of networks trained with dropout, the performance of networks trained without dropout is drastically worse. MAE for networks trained with dropout converges to a value below 1.8 after about 500 epochs of training, whereas MAE for networks trained without dropout reaches its lowest value at slightly above 1.8 after only a few epochs of training before it increases almost logarithmically. This observation mirrors the result obtained from applying dropout to speech and image recognition,[47] confirming that overfitting of the networks for contact number prediction was prevented and performance was improved by using dropout.

## Performances of Networks on Polytopic HMPs

In light of the investigation on the effects of window sizes, we first examined the performance of the networks for polytopic HMPs using each of the feature vectors separately considering a window size of 15. The performance measures of the networks were averaged over the validation sets. Table 2 summarizes our findings. When using CI as the feature vector, only a moderate PCC of 0.23 was achieved. Switching from CI to LSC increased the performance from PCC = 0.30 to PCC = 0.41. Consistent with the previous conclusion that entries in PSSM generally show stronger correlation with contact numbers, the networks achieved a significantly higher PCC (0.69) with PSSM. It is interesting to note that BP gave a worse performance (PCC = 0.65) than PSSM despite the fact that it is derived from PSSM. The result of MAE mirrors the observation made on PCC with lower MAE corresponding to higher PCC.

Traditionally, prediction of the contact number is treated as a classification problem in which a residue is categorized as either exposed or buried. It is also interesting to see the performance of the current method regarding classification of residue burial status. For computing accuracy and MCC for polytopic HMPs, the median contact number of 11.44 in the subset of polytopic HMPs was used as the cutoff. The cutoff was set in this way so that the data set is class-balanced (number of exposed residues equals that of buried residues), and the accuracy of a classifier that assigns all residues to one particular class is at most 50%. As shown in Table 2, both accuracy and MCC follow the trend found in the previous section in the sense that PSSM gives the highest accuracy (75.8%) and MCC (0.52), whereas CI gives the lowest. The final networks were trained with dropout, using PSSM with a window size of 15 as the input feature vector. All results and discussions in the rest of the paper refer to the final networks, which was termed TMH-Expo.

## Contact Numbers for Bitopic HMPs Are Difficult To Predict

By comparing the performance of the networks on polytopic HMPs to that on bitopic HMPs, we observed that the performance on bitopic HMPs are substantially worse (Table 2 and Figure S2, Supporting Information). MAEs on bitopic HMPs are considerable higher than those on polytopic HMPs (2.51 versus 1.68). PCC, accuracy, and MCC (using a cutoff of 8.50, which is the median contact number for bitopic HMPs) on bitopic HMPs are significantly lower than those on polytopic HMPs. In fact, 11 out of 12 protein chains with MAE greater than 2.5 are bitopic (Table S2, Supporting Information). The reason why contact numbers for bitopic HMPs are more difficult to predict is still unclear. One potential explanation could be that the distribution of contact numbers for bitopic HPMs is significantly different from that for polytopic HMPs (Figure 5). Using relative conservation analysis, Zviling et al. recently proposed that bitopic HPMs have various interaction modes.[50] If this is the case, the interaction modes for bitopic HMPs observed in the data set might only represent one of multiple possible modes (e.g., buried face of the helix of a bitopic HMP in one complex might be instead the exposed face when being part of another complex). Therefore, contact numbers for bitopic HMPs computed based on complex structures observed in the current data set might be biased.

### Contact Numbers for Very Exposed or Very Buried TMHs Are Difficult To Predict

One reason why the distribution of contact numbers for bitopic HMPs is drastically different from that for polytopic HMPs is that most bitopic HMPs are docked to the surface of large HMP complexes, leading to fewer interacting TMHs then a TMH at the center of a large HMP. In fact, out of the 20 bitopic HMPs in the data set, 17 are localized on the surface of a HMP complex. The fact that the contact number for bitopic HMPs are difficult to predict poses an interesting question: Is it a general feature that contact numbers for TMHs with fewer interacting TMHs are difficult to predict? In order to answer this question, we computed the MAE for each TMH. We also binned TMHs into groups according their average contact number, assuming that the average contact number is a scaled indicator of the number of interacting TMHs. Figure 6 shows that TMHs with very few interacting partners have an increased group-averaged MAE. Interestingly, Figure 6 also shows that completely buried TMHs have the highest group-averaged MAE.

### Contact Numbers of Extremely Exposed or Buried Residues Are Difficult To Predict

In addition to the overall performance, the distribution of MAE was analyzed. The positive skewness of the unimodal density curve for the distribution MAE (Figure 7a) indicates that the model was able to accurately predict the contact number for most residues. In fact, 53.5% of the residues were predicted with an absolute error of less than 1.5, and 66.6% of the residues were predicted with an absolute error of less than 2. Knowing whether the performance of the networks differs for different ranges of contact number is helpful as it indicates how reliable the result is when interpreting a prediction. We grouped residues using the same grouping scheme applied in the previous section and computed the group-averaged MAEs. Similar to the situation with TMHs, Figure 7b shows that MAE is higher toward either end of the residue groups than in the middle. This relationship implies that contact numbers for residues in the most buried groups (highest contact number) or the most exposed groups (lowest contact number) are the most difficult to predict.

### Amino Acid Bias in Prediction Error

In order to examine whether there are amino acid types for which the contact number is more difficult to predict, we computed amino acid residue-specific MAEs. Figure 8a shows the MAE for each amino acid type. In general, amino acids with charged side chains (Lys, Glu, His, Asp, and Arg) have lower MAEs than those with uncharged side chains. This is likely because of the fact that these charged residues are functionally important and are often employed by membrane proteins to bind ligands,[51] thus having similar burial status. In fact, the standard deviations of contact numbers of these charged residues are among the lowest (Table 1 and ref 52). MAEs for Pro, Ala, and Gly are among the highest and are significantly higher than those of the other residues. Prolines introduce kinks or $\pi$-bulges to TMHs.[53] Alanines and glycines form the sequence motifs of type AxxxA and GxxxG that are believed to promote close helical packing.[42] These residues have a highly variable exposure environment as indicated by the high standard deviations of the contact numbers (Table 1 and ref 52). The correlation between MAEs and the standard deviation of contact numbers of amino acid types is 0.84 (Figure 8b), suggesting that increased variability of exposure is an important determining factor for reduced prediction quality.

## Predicted Contact Numbers Reveal Exposure Pattern

An important application of contact number predictors is that they can be incorporated into scoring functions for evaluating *de novo* predicted or homology-modeled 3D protein structures. However, the possibility of this application depends on whether predicted contact numbers are accurate enough to reflect the exposure pattern of TMHs. For illustrative purposes, we mapped the experimental and predicted contact numbers onto the native structure for two protein chains (3tlwA, 4buoA). 3tlwA is one of the five subunits of the GLIC homopentameric ligand-gated ion channel[54] and is among the cases for which the networks achieved the lowest MAE and highest PCC (Figure 9a). Protein chain 4buoA is a structure of the thermostable agonist-bound G-protein-coupled receptor neurotensin receptor 1[55] for which the networks also achieved good prediction (Figure 9d). Comparing Figure 9b with c and e with f shows that contact numbers predicted by the networks correctly reflect exposure patterns for membrane-facing as well as buried TMHs. The two-phases of membrane-facing TMHs are differentiated by the alternating nature of predicted contact numbers. Thus, predicted contact numbers can be used to eliminate incorrectly predicted 3D structure models where buried TMHs are placed facing the membrane or vice versa.

## Predicting Membrane Protein–Membrane Protein Interface

Oligomerization is an essential mechanism by which many membrane proteins function.[56] In fact, 49 out of 71 HMPs in the TMH-Expo data set are oligomers. Interaction between membrane protein and membrane protein is a research area that has gained increasing attention from the biochemical community.[57,58] Given a monomer HMP with a known structure, it is desirable to identify interface-forming residues with a reasonable accuracy. As experimental contact numbers were calculated from structures where all trans-membrane subunits are included, we hypothesized that predicted contact numbers will be generally higher for interface residues than for non-interface lipid-exposed residues. If our hypothesis proved correct, then interface-forming TMHs can be identified. For evaluating the performance of TMH-Expo on identifying interface residues, we defined a residue as an interface residue if $CN_O - CN_m \geq 1$, where $CN_O$ is the contact number in oligomeric state and $CN_m$ is that in monomeric state. A residue is predicted as an interface residue if $CN_p - CN_m \geq 1$, where $CN_p$ is the predicted contact number. The cutoff value of 1 was chosen to reduce the chance of including residues on the protein core-buried face of a TMH as interface residues. A total of 16.3% residues in the data set satisfied this definition. For classifying interface residues (Table S3, Supporting Information), TMH-Expo achieved an overall accuracy of 68.6% and a sensitivity of 76.8%, significantly better than the performance reported in a similar study.[23] One should be aware of the high FPR of TMH-Expo (33.0%), a complication that could be accounted for by the fact that the oligomeric state of many HMPs is not unambiguously defined.[59]

As an example of predicting membrane protein–membrane protein interface residues, we investigated the performance of TMH-Expo for the subunit (4al0A) of the homotrimeric microsomal prostaglandin E2 synthase;[60] 4al0A has a similar FPR (32.1%) to the overall FPR of TMH-Expo. As shown in Table 3, out of the 85 TMH residues, 66 were correctly classified, giving an overall accuracy of 77.6%. Among these 32 interface residues, 30 were identified, giving a sensitivity of 93.8%. To visualize the prediction, we highlighted interface

residues identified with experimental contact numbers (Figure 10a) and those identified with predicted contact numbers (Figure 10b) on the native structure. Despite the high FPR, most false positives can be reasonably eliminated if we only consider residues on the exposed face of a TMH.

Comparison with Other Contact Number Predictors. To the best of our knowledge, TMH-Expo is the first attempt that has been made to predict contact numbers for membrane proteins. Therefore, a direct comparison of TMH-Expo with any of the other existing methods is not possible. To give an approximate sense of the performance of TMH-Expo, we compared TMH-Expo with two notable contact number predictors developed for soluble proteins. Using linear regression analysis, Kinjo et al. developed a real-valued contact number predictor with a PCC of 0.63[19] that was outperformed by TMH-Expo. Yuan developed a support vector regression-based predictor with a PCC of 0.70,[18] slightly better than TMH-Expo. However, it should be noted that the performance of Yuan's method might be favorably biased since the data set was not split in a way such that proteins in the same superfamily stay within the same subset. In addition, the structural repository of soluble proteins is significantly bigger than that of HMPs, making the training set for soluble proteins more informative.

We also trained neural network models for RSA prediction using PSSM as feature vector and the same training parameters as with training networks for contact number prediction. For RSA prediction, TMH-Expo achieved a PCC of 0.58 for polytopic HMPs. Since both accuracy and MCC are dependent on the cutoff value applied, it is rather arbitrary to make comparisons based on these two performance measures. We therefore approximately (since the data set employed in different study varies) compared our method to predictors for which PCC was reported. Yuan et al. developed a support vector regression-based predictor termed $ASAP_{mem}$ with a PCC of 0.66 for TM helical residues.[21] A random forest-based method recently reported by Wang et al. achieved a PCC of 0.68.[61] Although these two methods reportedly have better performance than TMH-Expo on RSA prediction, it should be pointed out that the cross-validation scheme employed in these studies might have favorably biased the performance. In fact, using the same cross-validation scheme, Illergård et al. trained a RSA predictor MPRAP which achieved the same PCC as TMH-Expo.[23]

## Limitations

In the current implementation of the algorithm, total contact number is computed by summing over contributions made by residues inside a sphere centered at the $C_\beta$ atom of the residue of interest. The contribution is assigned to each residue in a distance-dependent way such that close neighbors have an increased weight when compared to distant neighbors. This approach mirrors the distance-dependence of van der Waals and electrostatic interactions and is superior to the use of a single cutoff distance. However, it should be pointed out that one of the shortcomings of the current implementation is that the spatial distribution of neighboring residues is not taken into account.[5] Another limitation comes from the coarse-grained $C_\beta$ representation of the side chains in which size and bulkiness of side chains is ignored. While representing side chain atoms as a single "superatom" improves computational efficiency and is necessary in early stages of *de novo* 3D structure

prediction, it could result in loss of important structural information and leads to biased estimate of contact numbers. For instance, residues with a bulky side chain have longer $C_\beta$-$C_\beta$ distances than small residues. Thus, the average contact numbers for bulky residues might be underestimated.[62] When the information about the spatial distribution of neighboring residues is needed, a computationally slightly more demanding quantity called "neighbor vector" could be employed.[5] The neighbor vector is a vector associated with each residue whose direction and magnitude not only depend on the number of neighboring residues but also on the spatial distribution.

## Conclusion

We have developed a dropout neural network-based contact number and RSA predictor, TMH-Expo, for HMPs. TMH-Expo is the first work that reports contact number prediction for HMPs. Trained on an expanded nonredundant data set of HMPs with 5-fold cross-validation, TMH-Expo achieved an unprecedented PCC of 0.69 between experimental and predicted contact numbers. We have also shown that the training was benefitted from using neuronal dropout. With neuronal dropout, overfitting was significantly reduced, and the performance was improved. Detailed examination of MAEs and PCCs indicated that it is generally easier to predict contact numbers for polytopic HMPs than for bitopic HMPs. Mapping of predicted contact numbers onto structures demonstrated that contact numbers predicted by TMH-Expo reflect exposure patterns of TMHs and reveal interface-forming TMHs. This reinforces the idea of incorporating predicted contact numbers for predicting helix–helix packing and protein–protein docking. *De novo* protein folding and protein-protein docking studies leveraging contact numbers predicted by TMH-Expo are currently ongoing.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 GM099842, R01 DK097376, R01 HL122010, R01 GM073151) and NSF (CHE 1305874). We thank Louesa Akin for testing C++ code written as part of this study. TMH-expo has been integrated into the Biochemical Library (BCL) software suite that is being actively developed. It is also available via a Web server at http://www.meilerlab.org/index.php/servers. The BCL software suite is available at http://www.meilerlab.org/bclcommons under academic and business site licenses. The BCL source code is published under the BCL license and is available at http://www.meilerlab.org/bclcommons.

## References

1. Overington JP, Al-Lazikani B, Hopkins AL. How Many Drug Targets Are There? Nat Rev Drug Discovery. 2006; 5:993–996. [PubMed: 17139284]
2. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. J Mol Biol. 2001; 305:567–580. [PubMed: 11152613]
3. Lee B, Richards FM. The Interpretation of Protein Structures: Estimation of Static Accessibility. J Mol Biol. 1971; 55:379–400. [PubMed: 5551392]
4. Dill KA. Polymer Principles and Protein Folding. Protein Sci. 1999; 8:1166–1180. [PubMed: 10386867]

5. Durham E, Dorr B, Woetzel N, Staritzbichler R, Meiler J. Solvent Accessible Surface Area Approximations for Rapid and Accurate Protein Structure Prediction. J Mol Model. 2009; 15:1093–1108. [PubMed: 19234730]
6. Haste Andersen P, Nielsen M, Lund O. Prediction of Residuesin Discontinuous B-Cell Epitopes Using Protein 3d Structures. Protein Sci. 2006; 15:2558–2567. [PubMed: 17001032]
7. Martins JM, Ramos RM, Pimenta AC, Moreira IS. Solvent-Accessible Surface Area: How Well Can Be Applied to Hot-Spot Detection? Proteins: Struct, Funct, Genet. 2014; 82:479–490. [PubMed: 24105801]
8. Munteanu CR, Pimenta AC, Fernandez-Lozano C, Melo A, Cordeiro MN, Moreira IS. Solvent Accessible Surface Area-Based Hot-Spot Detection Methods for Protein-Protein and Protein-Nucleic Acid Interfaces. J Chem Inf Model. 2015; 55:1077–1086. [PubMed: 25845030]
9. Jones S, Thornton JM. Prediction of Protein-Protein Interaction Sites Using Patch Analysis. J Mol Biol. 1997; 272:133–143. [PubMed: 9299343]
10. Jones S, Thornton JM. Analysis of Protein-Protein Interaction Sites Using Surface Patches. J Mol Biol. 1997; 272:121–132. [PubMed: 9299342]
11. Marsh JA, Teichmann SA. Relative Solvent Accessible Surface Area Predicts Protein Conformational Changes Upon Binding. Structure. 2011; 19:859–867. [PubMed: 21645856]
12. 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. Curr Protein Pept Sci. 2011; 12:563–573. [PubMed: 21787302]
13. Eyal E, Najmanovich R, McConkey BJ, Edelman M, Sobolev V. Importance of Solvent Accessibility and Contact Surfaces in Modeling Side-Chain Conformations in Proteins. J Comput Chem. 2004; 25:712–724. [PubMed: 14978714]
14. Pollastri G, Baldi P, Fariselli P, Casadio R. Improved Prediction of the Number of Residue Contacts in Proteins by Recurrent Neural Networks. Bioinformatics. 2001; 17(Suppl 1):S234–S242. [PubMed: 11473014]
15. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of Coordination Number and Relative Solvent Accessibility in Proteins. Proteins: Struct, Funct, Genet. 2002; 47:142–153. [PubMed: 11933061]
16. Fariselli P, Casadio R. Prediction of the Number of Residue Contacts in Proteins. Proc Int Conf Intell Syst Mol Biol. 2000; 8:146–151. [PubMed: 10977075]
17. Ahmad S, Gromiha MM, Sarai A. Real Value Prediction of Solvent Accessibility from Amino Acid Sequence. Proteins: Struct, Funct, Genet. 2003; 50:629–635. [PubMed: 12577269]
18. Yuan Z. Better Prediction of Protein Contact Number Using a Support Vector Regression Analysis of Amino Acid Sequence. BMC Bioinf. 2005; 6:248.
19. Kinjo AR, Horimoto K, Nishikawa K. Predicting Absolute Contact Numbers of Native Protein Structure from Amino Acid Sequence. Proteins: Struct, Funct, Genet. 2005; 58:158–165. [PubMed: 15523668]
20. Beuming T, Weinstein H. A Knowledge-Based Scale for the Analysis and Prediction of Buried and Exposed Faces of Trans-membrane Domain Proteins. Bioinformatics. 2004; 20:1822–1835. [PubMed: 14988128]
21. Yuan Z, Zhang F, Davis MJ, Boden M, Teasdale RD. Predicting the Solvent Accessibility of Transmembrane Residues from Protein Sequence. J Proteome Res. 2006; 5:1063–1070. [PubMed: 16674095]
22. Park Y, Hayat S, Helms V. Prediction of the Burial Status of Transmembrane Residues of Helical Membrane Proteins. BMC Bioinf. 2007; 8:302.
23. Illergard K, Callegari S, Elofsson A. Mprap: An Accessibility Predictor for a-Helical Transmembrane Proteins That Performs Well inside and Outside the Membrane. BMC Bioinf. 2010; 11:333.
24. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. Opm: Orientations of Proteins in Membranes Database. Bioinformatics. 2006; 22:623–625. [PubMed: 16397007]
25. Dekker JP, Boekema EJ. Supramolecular Organization of Thylakoid Membrane Proteins in Green Plants. Biochim Biophys Acta, Bioenerg. 2005; 1706:12–39.

26. Wang G, Dunbrack RL Jr. Pisces: A Protein Sequence Culling Server. Bioinformatics. 2003; 19:1589–1591. [PubMed: 12912846]
27. Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]
28. Heinig M, Frishman D. Stride: A Web Server for Secondary Structure Assignment from Known Atomic Coordinates of Proteins. Nucleic Acids Res. 2004; 32:W500–502. [PubMed: 15215436]
29. Majumdar I, Krishna SS, Grishin NV. Palsse: A Program to Delineate Linear Secondary Structural Elements from Protein Structures. BMC Bioinf. 2005; 6:202.
30. Nishikawa K, Ooi T. Radial Locations of Amino Acid Residues in a Globular Protein: Correlation with the Sequence. J Biochem. 1986; 100:1043–1047. [PubMed: 3818558]
31. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A Generic Method for Assignment of Reliability Scores Applied to Solvent Accessibility Predictions. BMC Struct Biol. 2009; 9:51. [PubMed: 19646261]
32. Chang DT, Huang HY, Syu YT, Wu CP. Real Value Prediction of Protein Solvent Accessibility Using Enhanced Pssm Features. BMC Bioinf. 2008; 9(Suppl 12):S12.
33. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. Uniref: Comprehensive and Non-Redundant Uniprot Reference Clusters. Bioinformatics. 2007; 23:1282–1288. [PubMed: 17379688]
34. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped Blast and Psi-Blast: A New Generation of Protein Database Search Programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
35. McGuffin LJ, Bryson K, Jones DT. The Psipred Protein Structure Prediction Server. Bioinformatics. 2000; 16:404–405. [PubMed: 10869041]
36. Rumelhart DE, Hinton GE, Williams RJ. Learning Representations by Back-Propagating Errors. Nature. 1986; 323:533–536.
37. Jaakkola, T., Diekhans, M., Haussler, D. Using the Fisher Kernel Method to Detect Remote Protein Homologies. Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology; AAAI Press. 1999. p. 149-158.
38. Leman JK, Mueller R, Karakas M, Woetzel N, Meiler J. Simultaneous Prediction of Protein Secondary Structure and Trans-membrane Spans. Proteins: Struct, Funct, Genet. 2013; 81:1127–1140. [PubMed: 23349002]
39. Matthews BW. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. Biochim Biophys Acta, Protein Struct. 1975; 405:442–451.
40. White SH. The Progress of Membrane Protein Structure Determination. Protein Sci. 2004; 13:1948–1949. [PubMed: 15215534]
41. Ulmschneider MB, Sansom MS. Amino Acid Distributions in Integral Membrane Protein Structures. Biochim Biophys Acta, Biomembr. 2001; 1512:1–14.
42. Russ WP, Engelman DM. The Gxxxg Motif: A Framework for Transmembrane Helix-Helix Association. J Mol Biol. 2000; 296:911–919. [PubMed: 10677291]
43. Ma J, Wang S. Acconpred: Predicting Solvent Accessibility and Contact Number Simultaneously by a Multitask Learning Framework under the Conditional Neural Fields Model. BioMed Res Int. 2015; 2015:678764. [PubMed: 26339631]
44. Lai JS, Cheng CW, Lo A, Sung TY, Hsu WL. Lipid Exposure Prediction Enhances the Inference of Rotational Angles of Transmembrane Helices. BMC Bioinf. 2013; 14:304.
45. Walters RF, DeGrado WF. Helix-Packing Motifs in Membrane Proteins. Proc Natl Acad Sci U S A. 2006; 103:13658–13663. [PubMed: 16954199]
46. Adamian L, Liang J. Prediction of Transmembrane Helix Orientation in Polytopic Membrane Proteins. BMC Struct Biol. 2006; 6:13. [PubMed: 16792816]
47. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014; 15:1929–1958.
48. Krizhevsky, A., Sutskever, I., Hinton, GE. Imagenet Classification with Deep Convolutional Neural Networks; the Twenty-sixth Annual Conference on Neural Information Processing Systems (NIPS); Lake Tahoe, NV. December 3-8, 2012; NIPS; 2012. p. 1097-1105.

49. Deng, L., Hinton, G., Kingsbury, B. New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview; Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Vancouver, BC, Canada. May 26-31, 2013; IEEE; 1998. p. 8599-8603.
50. Zviling M, Kochva U, Arkin IT. How Important Are Transmembrane Helices of Bitopic Membrane Proteins? Biochim Biophys Acta, Biomembr. 2007; 1768:387–392.
51. Illergard K, Kauko A, Elofsson A. Why Are Polar Residues within the Membrane Core Evolutionary Conserved? Proteins: Struct, Funct, Genet. 2011; 79:79–91. [PubMed: 20938980]
52. Adamian L, Liang J. Helix-Helix Packing and Interfacial Pairwise Interactions of Residues in Membrane Proteins. J Mol Biol. 2001; 311:891–907. [PubMed: 11518538]
53. Senes A, Engel DE, DeGrado WF. Folding of Helical Membrane Proteins: The Role of Polar, Gxxxg-Like and Proline Motifs. Curr Opin Struct Biol. 2004; 14:465–479. [PubMed: 15313242]
54. Tiefenbrunn T, Liu W, Chen Y, Katritch V, Stout CD, Fee JA, Cherezov V. High Resolution Structure of the Ba3 Cytochrome C Oxidase from Thermus Thermophilus in a Lipidic Environment. PLoS One. 2011; 6:e22348. [PubMed: 21814577]
55. Egloff P, Hillenbrand M, Klenk C, Batyuk A, Heine P, Balada S, Schlinkmann KM, Scott DJ, Schutz M, Pluckthun A. Structure of Signaling-Competent Neurotensin Receptor 1 Obtained by Directed Evolution in Escherichia Coli. Proc Natl Acad Sci U S A. 2014; 111:E655–662. [PubMed: 24453215]
56. Kawano K, Yano Y, Omae K, Matsuzaki S, Matsuzaki K. Stoichiometric Analysis of Oligomerization of Membrane Proteins on Living Cells Using Coiled-Coil Labeling and Spectral Imaging. Anal Chem. 2013; 85:3454–3461. [PubMed: 23427815]
57. Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS, Fields S. Large-Scale Identification of Yeast Integral Membrane Protein Interactions. Proc Natl Acad Sci U S A. 2005; 102:12123–12128. [PubMed: 16093310]
58. Babu M, Vlasblom J, Pu S, Guo X, Graham C, Bean BD, Burston HE, Vizeacoumar FJ, Snider J, Phanse S, Fong V, Tam YY, Davey M, Hnatshak O, Bajaj N, Chandran S, Punna T, Christopolous C, Wong V, Yu A, Zhong G, Li J, Stagljar I, Conibear E, Wodak SJ, Emili A, Greenblatt JF. Interaction Landscape of Membrane-Protein Complexes in Saccharomyces Cerevisiae. Nature. 2012; 489:585–589. [PubMed: 22940862]
59. Duarte JM, Biyani N, Baskaran K, Capitani G. An Analysis of Oligomerization Interfaces in Transmembrane Proteins. BMC Struct Biol. 2013; 13:21. [PubMed: 24134166]
60. Sjogren T, Nord J, Ek M, Johansson P, Liu G, Geschwindner S. Crystal Structure of Microsomal Prostaglandin E2Synthase Provides Insight into Diversity in the Mapeg Superfamily. Proc Natl Acad Sci U S A. 2013; 110:3806–3811. [PubMed: 23431194]
61. Wang C, Xi L, Li S, Liu H, Yao X. A Sequence-Based Computational Model for the Prediction of the Solvent Accessible Surface Area for Alpha-Helix and Beta-Barrel Transmembrane Residues. J Comput Chem. 2012; 33:11–17. [PubMed: 21935968]
62. Gimpelev M, Forrest LR, Murray D, Honig B. Helical Packing Patterns in Membrane and Soluble Proteins. Biophys J. 2004; 87:4075–4086. [PubMed: 15465852]


## Abbreviations

| | |
|---|---|
| **HMP** | helical membrane protein |
| **TMH** | transmembrane helix |
| **BCL** | biochemical library |
| **PSSM** | position-specific scoring matrix |
| **PCC** | Pearson’s correlation coefficient |
| **MCC** | Matthew’s correlation coefficient |

| | |
|---|---|
| **MAE** | mean absolute error |
| **MSA** | multiple sequence alignment |
| **RSA** | relative solvent accessibility |
| **ASA** | absolute solvent accessibility |

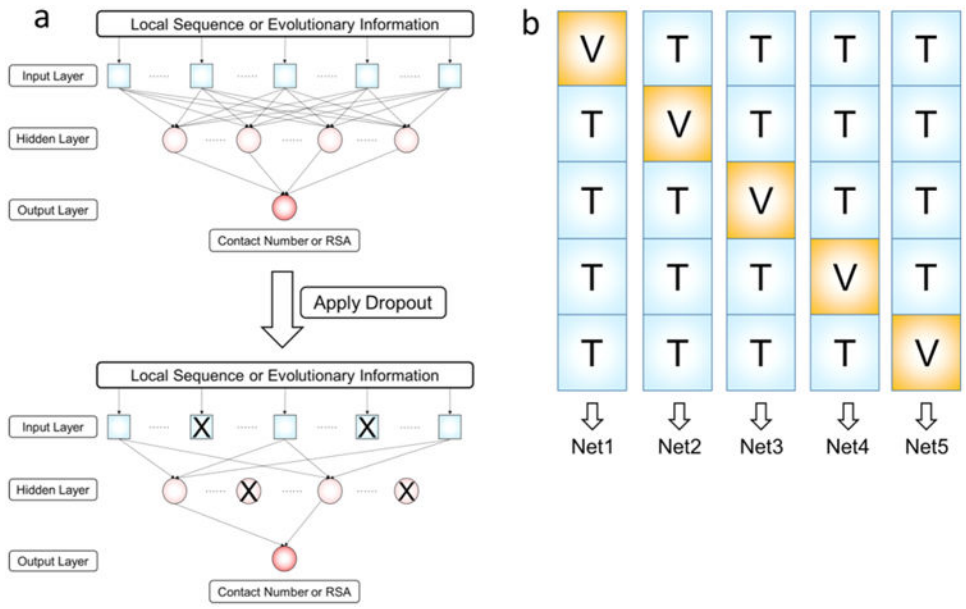



**Figure 1.**

Training of dropout neural networks with 5-fold cross-validation: (a) neural network architectures before and after applying dropout (neurons randomly dropped out are crossed), (b) 5-fold cross-validation training protocol (T, training set; V, validation set).

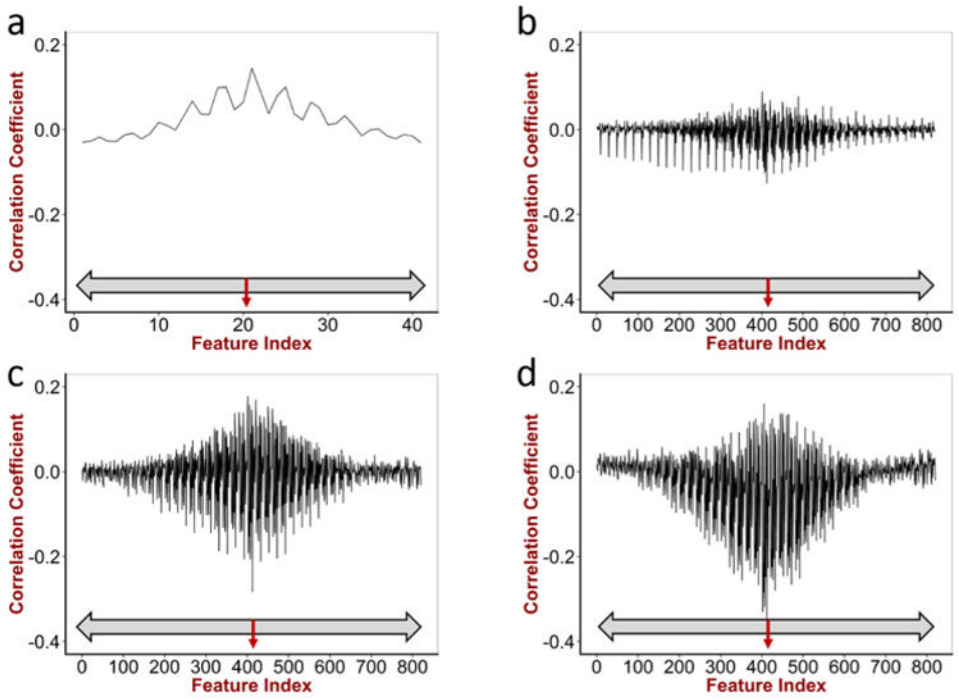



**Figure 2.**

Correlation of features with contact numbers: (a) correlation of entries in CI feature vector with contact numbers, (b) correlation of entries in LSC feature vector with contact numbers, (c) correlation of entries in BPP feature vector with contact numbers, and (d) correlation of entries in PSSM feature vector with contact numbers. Each entry in the feature vector is assigned a feature index sequentially such that it starts with 0 for the leftmost residue and ends with 40 (CI) or 820 (other feature vectors) for the rightmost residue (double-headed arrow bar). The red arrow from the arrow bar points to the central residue.

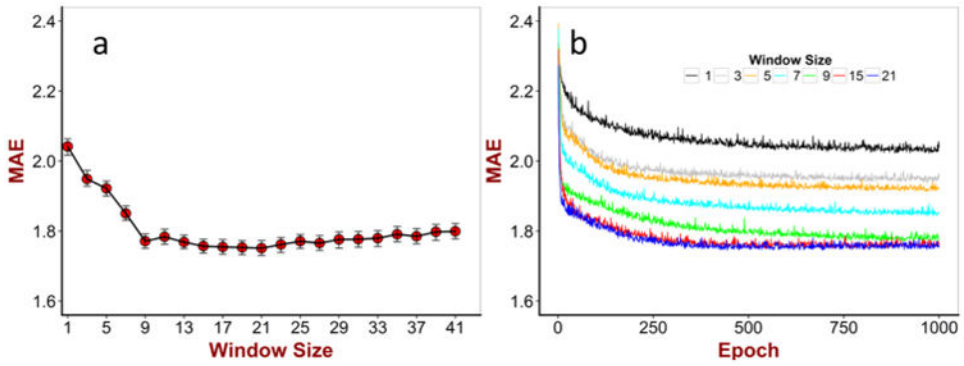



**Figure 3.**

Effect of window size on the performance of the neural networks: (a) final MAE on validation sets averaged over cross-validated neural networks and (b) MAEs averaged over cross-validated neural networks as the neural networks were being iteratively trained.

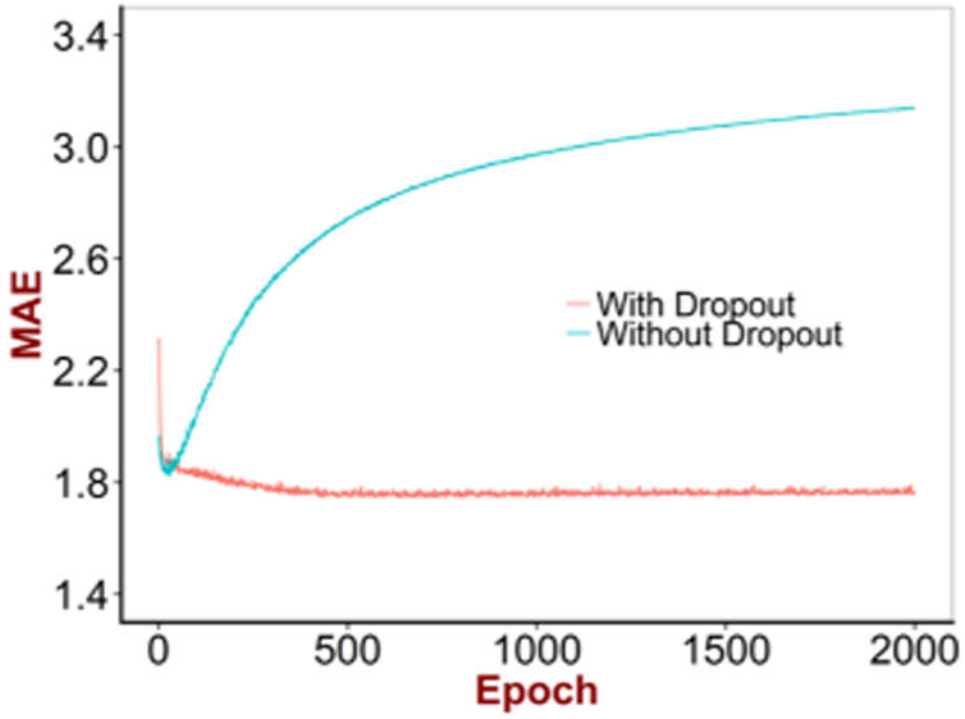



**Figure 4.**
MAE on validation sets for neural networks trained with or without dropout as learning progresses.

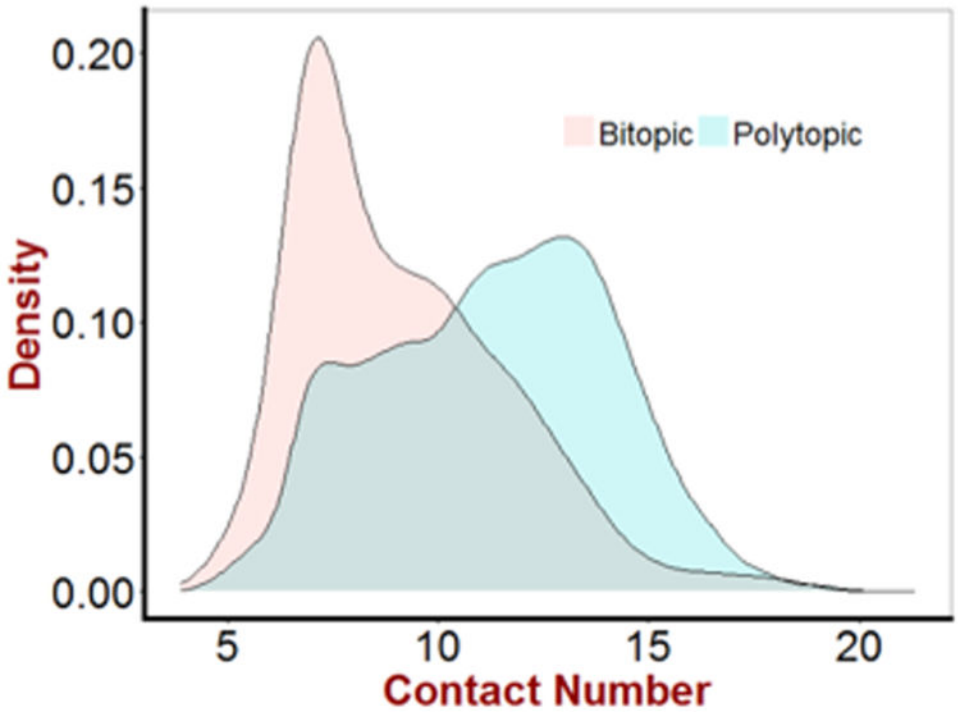



**Figure 5.**
Distribution of contact numbers of bitopic and polytopic HMPs.

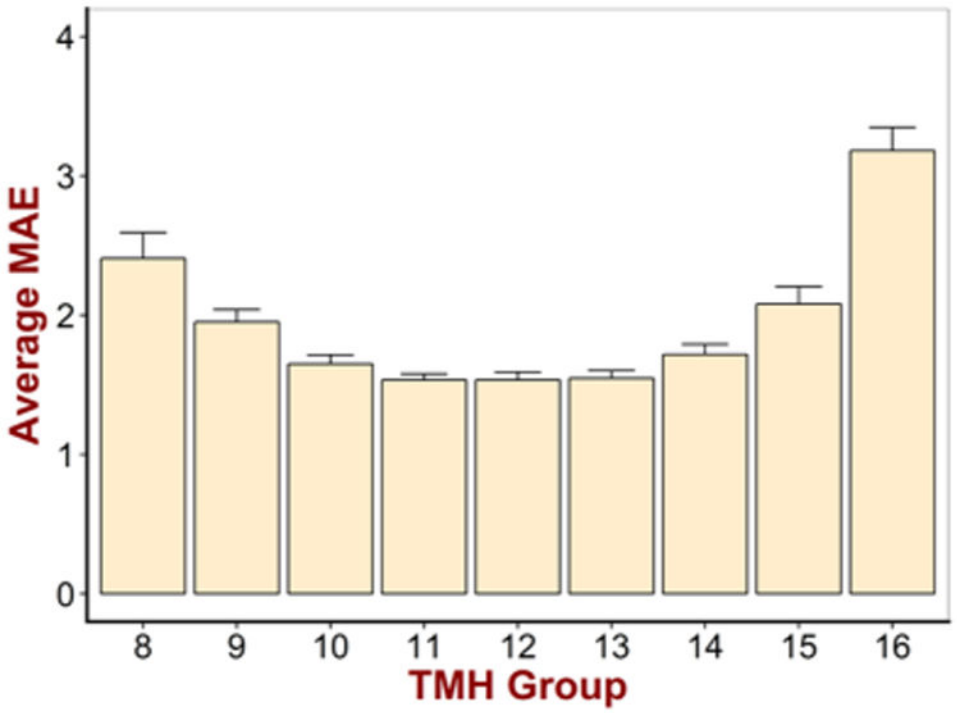



**Figure 6.**

Group-averaged MAEs for TMHs grouped according to their average contact numbers. The *x*-axis denotes average contact number of a TMH group. For instance, 10 means the group of TMHs that have average contact numbers between 9 and 10.

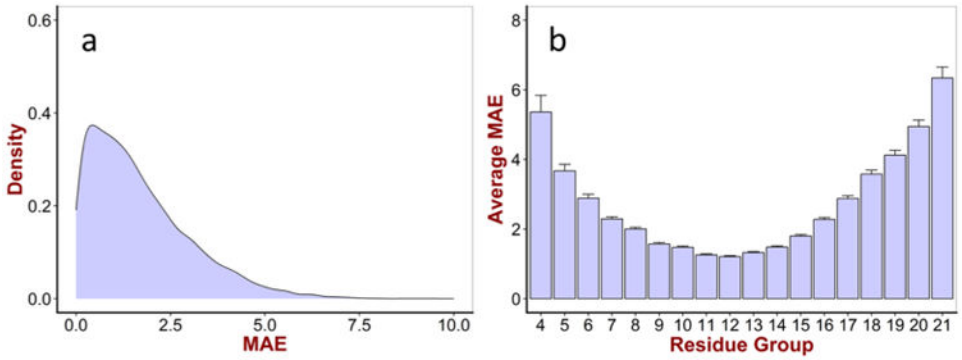



**Figure 7.**

Groupe-averaged MAEs for residues grouped according to their contact numbers. The x-axis denotes the contact number of a residue group. For instance, 10 means the group of residues that have contact numbers between 9 and 10.

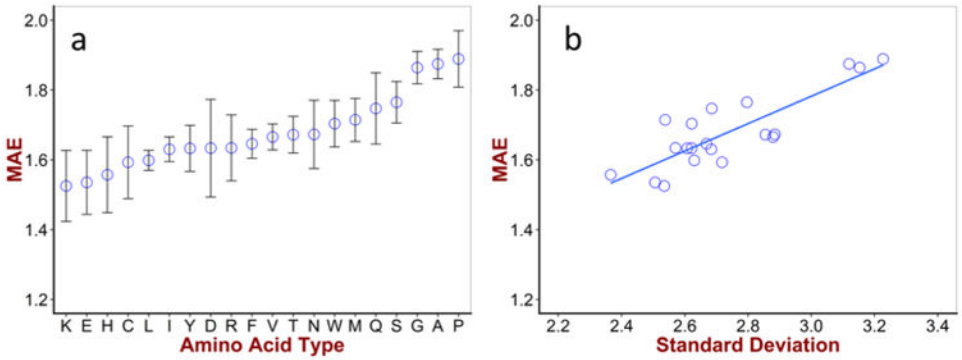



**Figure 8.**

Amino acid type-specific MAEs and dependence of MAE on standard deviation of contact numbers: (a) amino acid type-specific MAEs and (b) dependence of MAE on standard deviation of contact numbers.

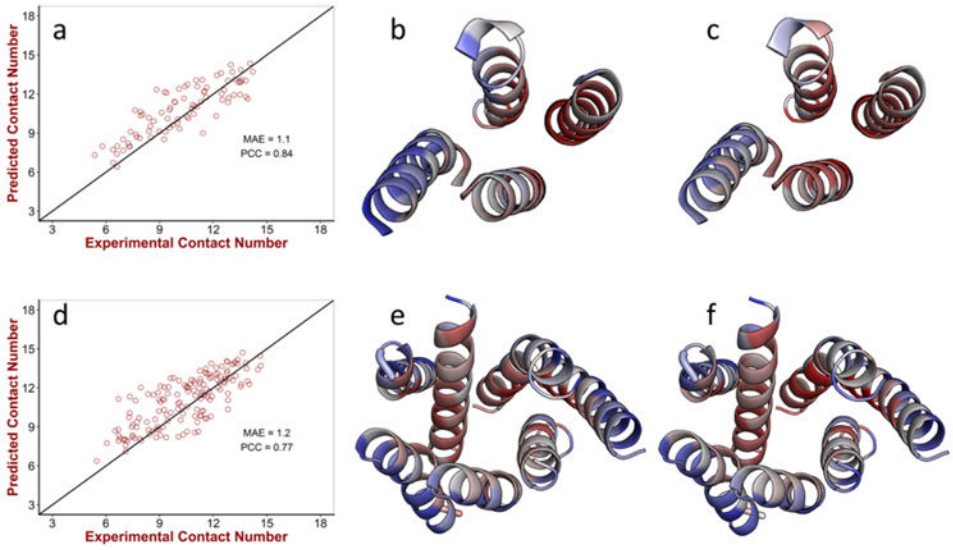



**Figure 9.**

Predicted contact numbers reveal exposure pattern of TMHs: (a) correlation between experimental and predicted contact numbers for 3tlwA, (b) mapping of experimental contact numbers onto the crystal structure of 3tlwA, ( c) mapping of predicted contact number to the crystal structure of 3tlwA, (d) correlation between experimental and predicted contact numbers for 4buoA, (e) mapping of experimental contact numbers onto the crystal structure of 4buoA, and (f) mapping of predicted contact number to the crystal structure of 4buoA. Color scheme: as contact number increases, color changes gradually from blue to red. Only TMHs are shown.

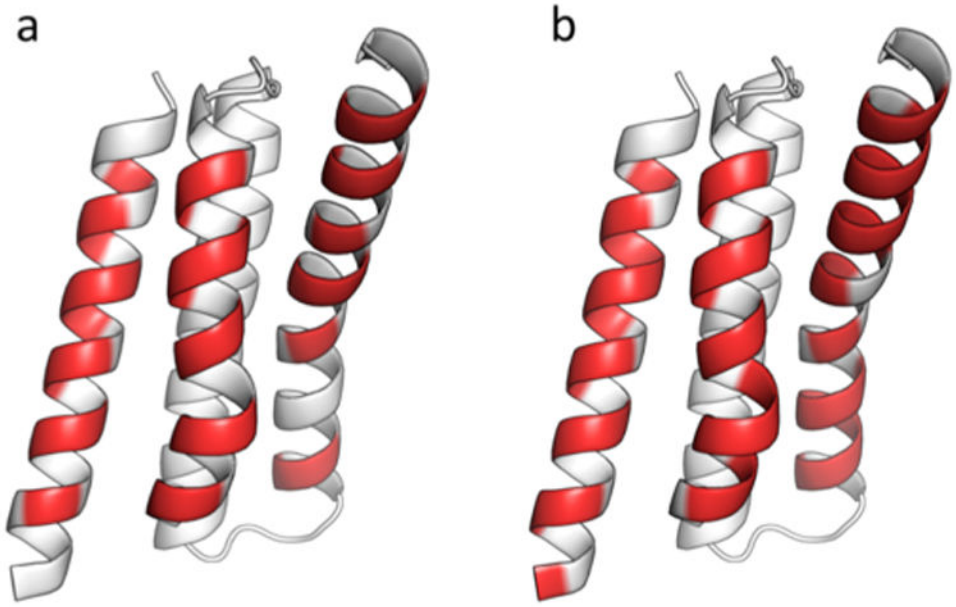



**Figure 10.**

Predicted contact numbers reveal interface-forming residues of 4al0A: (a) mapping of interface residues (colored in red) identified with experimental contact numbers onto the crystal structure of 4al0A and (b) mapping of predicted interface residues (colored in red) onto the crystal structure of 4al0A.

## Table 1

### Summary of TMH-Expo Data Set

| amino acid residue | frequency | mean contact number | standard deviation of contact number |
|---|---|---|---|
| A | 1282 | 12.09 | 3.12 |
| C | 131 | 12.46 | 2.72 |
| D | 93 | 11.34 | 2.62 |
| E | 151 | 11.10 | 2.51 |
| F | 953 | 10.58 | 2.67 |
| G | 1008 | 12.76 | 3.15 |
| H | 134 | 11.31 | 2.36 |
| I | 1242 | 10.46 | 2.68 |
| K | 156 | 9.22 | 2.53 |
| L | 1938 | 10.59 | 2.63 |
| M | 437 | 11.65 | 2.54 |
| N | 204 | 11.84 | 2.88 |
| P | 329 | 10.79 | 3.23 |
| Q | 161 | 11.02 | 2.69 |
| R | 184 | 9.94 | 2.57 |
| S | 598 | 12.20 | 2.80 |
| T | 604 | 11.83 | 2.85 |
| V | 1256 | 10.91 | 2.88 |
| W | 323 | 10.08 | 2.62 |
| Y | 381 | 11.02 | 2.61 |

**Table 2**

**Summary of Performance Measures for Contact Number Prediction**

| | | MAE | | PCC | | accuracy (%) | | MCC | |
|---|---|---|---|---|---|---|---|---|---|
| feature vector | size | P[a] | B[a] | P | B | P | B | P | B |
| CI | 15 × 1 | 2.33 | 2.63 | 0.23 | 0.33 | 58.4 | 50.1 | 0.18 | 0.00 |
| LSC | 15 × 20 | 2.18 | 2.79 | 0.41 | 0.15 | 63.9 | 50.1 | 0.28 | 0.00 |
| BPP | 15 × 20 | 1.79 | 2.62 | 0.65 | 0.28 | 73.1 | 51.5 | 0.47 | 0.05 |
| PSSM | 15 × 20 | 1.68 | 2.51 | 0.69 | 0.38 | 75.8 | 54.2 | 0.52 | 0.13 |

[a]P, polytopic; B, bitopic.

**Table 3**

**Performance of TMH-Expo on 4al0A**

| | | predicted | | |
|---|---|---|---|---|
| | | **interface** | **non-interface** | **total** |
| experimental | interface | 30 | 2 | 32 |
| | non-interface | 17 | 36 | 53 |
| | total | 47 | 38 | 85 |