TopProperty: Robust Meta-prediction of Transmembrane and Globular Protein Features using Deep Neural Networks

Daniel Mulnaes[1†], Stephan Schott-Verdugo[2†], Filip Koenig[1], and Holger Gohlke[1,2*]

[†] These authors contributed equally to this work.

[1]Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

[2]John von Neumann Institute for Computing (NIC), Jülich Supercomputing Centre (JSC), Institute of Biological Information Processing (IBI-7: Structural Bioinformatics), and Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Str., 52425 Jülich, Germany

Author ORCID

Daniel Mulnaes: 0000-0003-2162-5918

Stephan Schott-Verdugo: 0000-0003-0735-1404

Filip Koenig: 0000-0003-0852-440X

Holger Gohlke: 0000-0001-8613-1447

[*]Address: Universitätsstr. 1, 40225 Düsseldorf, Germany.
Phone: (+49) 211 81 13662; Fax: (+49) 211 81 13847
E-mail: gohlke@uni-duesseldorf.de

Abstract

Transmembrane proteins (TMPs) are critical components of cellular life. However, due to experimental challenges, the number of experimentally resolved TMP structures is severely underrepresented in databases compared to their cellular abundance. Prediction of (per-residue) features such as transmembrane topology, membrane exposure, secondary structure, and solvent accessibility can be a useful starting point for experimental design or protein structure prediction, but often requires different computational tools for different features or types of proteins. We present TopProperty, a meta-predictor that predicts all of these features for TMPs or globular proteins. TopProperty is trained on datasets without bias towards a high number of sequence homologs, and the predictions are significantly better than the evaluated state-of-the-art primary predictors on all quality metrics. TopProperty eliminates the need for protein type- or feature-tailored tools, specifically for TMPs. TopProperty is freely available as web server and standalone at https://cpclab.uni-duesseldorf.de/topsuite/.

Introduction

Transmembrane proteins (TMPs) play a vital role in both eukaryotic and prokaryotic cells. They are essential for several cellular processes, including signal transduction, molecular transportation, energy production, and cell adhesion [1]. Despite making up 20-30% of various genomes including the human one [2-4], TMPs constitute only a small fraction of the resolved protein structures available today [5-6] due to experimental challenges in structure determination. Interestingly, it is estimated that more than half of the currently used drug targets are TMPs [7].

TMPs generally fall into two major categories: Transmembrane α-helical bundles/anchors (TMHs) and transmembrane β-barrels (TMBs). While TMHs have seen a lot of scientific attention [1], TMBs are much more sparsely studied, as evident by both the number of experimentally resolved structures and the number of computational methods dedicated to each type [8-9]. Due to the location of TMPs in a bilayer environment and the anisotropic structure of cellular membranes, TMPs expose hydrophobic residues to the protein surface in the transmembrane region, while having hydrophilic residues exposed in water environments. As such, the Transmembrane Topology (TMT) reveals significant information regarding the way a TMP structure is folded and placed within the membrane. Thus, TMT predictions provide low-resolution information for TMPs, which can be a helpful starting point for experimental design or as constraints for protein structure prediction [1], especially when used alongside other linear features such as secondary structure (SS) [10-13], relative solvent accessibility (RSA) [11, 13-

[17], and residue-residue contacts and distances [18]. Furthermore, some methods have focussed on identifying membrane exposed (ME) residues based on lipophilicity [8, 19] and helix-helix interactions [20] to be used as modeling constraints.

While general features such as SS and RSA have frequently been predicted using deep learning techniques, TMT and ME features have seen less attention from the deep learning community due to a higher focus on residue-residue contacts and distance predictions [21-22]. Two classical issues with standard deep neural networks (DNNs) are the vanishing gradient problem [23] and different random initializations leading to different minima for identical model architectures [21, 24]. Nowadays, these issues have largely been resolved with the introduction of Residual networks [25] and the development of drop-out [26], which has become widely used in bioinformatics applications such as SS [27], residue contact prediction [21], and domain boundary prediction [28].

So far, the prediction of the different types of linear features mentioned above has often required different methods specific to the protein or feature in question. In this work, we present TopProperty, a meta-predictor that predicts TMT of both TMHs and TMBs as well as linear features such as SS, RSA, and ME using DNNs, making it applicable irrespective if a protein is a TMP or not. TopProperty improves the prediction of these features over existing methods and makes predictions easier for the user and agnostic to the type of protein submitted to the server. Finally, TopProperty shows robust results, particularly for proteins with a low number of sequence homologs, making it particularly suitable for *de novo* predictions. TopProperty shows better performance than state-of-the-art primary predictors in terms of Q3 scores, MCC values, and correlation coefficients for both TMPs and globular proteins [1].

Materials and Methods

Workflow. The goal of TopProperty is to give robust and accurate predictions of SS, RSA, TMT, and ME. The TopProperty workflow has three steps described below and shown in Figure 1:

1. **Primary Predictors.** The input sequence is submitted to 11 primary predictors to calculate 123 SS/RSA features and 16 primary predictors to calculate 129 TMT/ME features (see Input Features section below).
2. **DNNs.** Features are predicted using two ensembles of 1D residue-wise DNNs. The first predicts SS/RSA and the second TMT and ME. Each DNN in the ensemble has a different sliding window size to capture different amounts of local and non-local information.

3. **Post Processing.** The output of each of the different DNNs in a given ensemble is combined into a single score for each residue using a weighted average according to the performance of each of the different DNNs across the training dataset (MCC values for classification networks, Pearson's R^2 for RSA predictions). This weighted average is presented as the TopProperty prediction for SS, RSA, and ME. For TMT predictions, we found that, in rare cases, the DNNs directly interchanged between intracellular (I) and extracellular/luminal (O) labels, leading to predictions where they transitioned directly from one to the other without a membrane label in between. This was resolved by assigning each solvent-exposed segment the majority label for the segment and inverting minority labels.

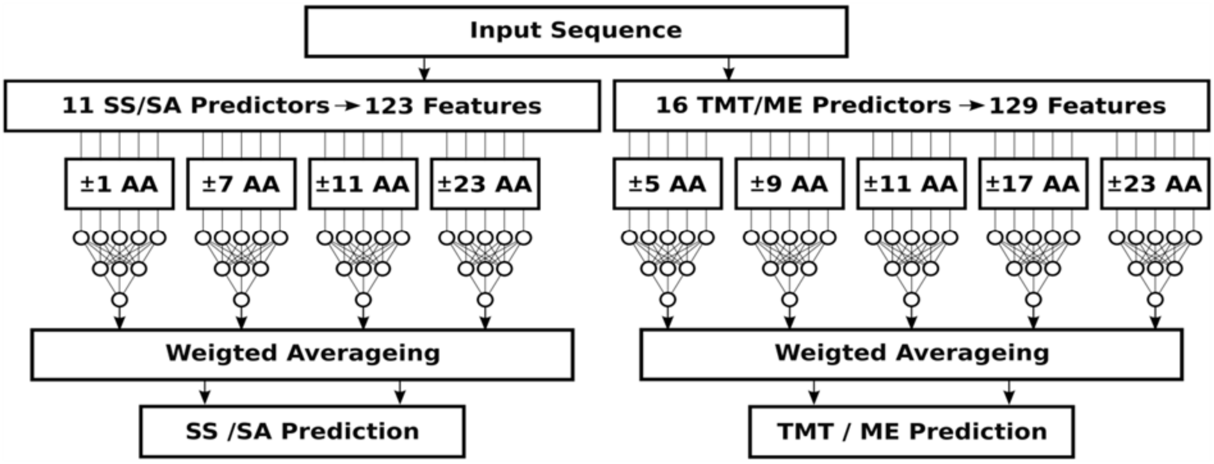


Figure 1. TopProperty Architecture. The target sequence is first submitted to 11 SS/SA primary predictors and 16 TMT/ME primary predictors to generate 123 SS/SA features and 129 TMT/ME features. These input features are used to predict SS/SA and TMT/ME using two ensembles of DNN's that use sliding windows of different sizes to capture local and non-local structural elements. Finally, the outputs of the different DNN's are combined using a weighted mean and provided to the user.

Datasets. Studies have found that many datasets used for contact prediction show a bias towards targets with many homologous sequences available [29]. This, in turn, leads to a significant overestimation of prediction performance for *de novo* cases, for which such abundance is rarely available. Thus, special care should be taken to generate unbiased datasets with respect to the number of sequence homologs.

For TopProperty, we generated four datasets (see Table S1 for the corresponding PDB IDs). The first dataset, termed the general training set, is based on the PDB and used for the training of predictors of secondary structure (SS) and solvent accessibility (SA). The second dataset is based on the NOUMENON dataset [29] and used to evaluate SS and SA predictions. This dataset is termed the NOUMENON test set. The last two datasets are based on the Orientation of Proteins in Membranes (OPM) database [30] and are used for training and evaluation of Transmembrane Topology (TMT) and Membrane Exposure (ME). These are termed the OPM training set and OPM test set, respectively.

General training set. To avoid bias towards systems with a high number of sequence homologs in our training set, the proteins of the general training set were culled from the PDB using the PISCES server [31]. The culling criteria were a pairwise sequence identity < 20%, resolution < 2 Å, and length < 500 residues. Non-terminal missing loops in the 3D structure were added using Rosetta 3.6 [32] and the loop building method in Modeller9 [33]. This repair is necessary to obtain a sequence and structure for which the true features can be determined for each residue. To measure the amount of sequence information, we use the number of effective sequences in an MSA generated by HHBLITS [34] (γ) divided by the length of the target sequence (δ). This measure (γ/δ) is termed the N_{eff} ratio. Compared to other datasets [29], no bias towards targets with a high number of sequence homologs is present in this dataset. Indeed, 23% of the targets have a N_{eff} ratio less than 1, and 67% have a N_{eff} ratio less than 10. Robustness to low amounts of sequence information is essential when predicting features for TMPs, which are likely to have shallower alignments. The final size of the general training set is 3235 proteins.

NOUMENON test set. For evaluation of TopProperty globular predictions, we used the NOUMENON dataset [29] after removing 23 proteins that overlap with the general training set. The NOUMENON dataset is designed to contain challenging protein targets with a low amount of available sequence information. Missing regions were rebuilt as for the general training set. The final size of the NOUMENON test set is 127 proteins.

OPM datasets. The number of resolved transmembrane protein structures is significantly smaller than the number of globular proteins. Therefore, we selected all chains from the OPM database [30] and clustered them at 70% sequence identity using CDhit [35]. Due to experimental difficulties when resolving structures of TMPs, many OPM entries are partially incomplete, and others are only partially transmembrane. To obtain a dataset containing complete structures with minimal information loss, we, therefore, used the following workflow:

1. Only TMPs were selected from the OPM, requiring that at least one protein residue must be within 5Å of the center of the membrane to be considered.
2. The biological sequence (without experimental artifacts such as missing or non-standard residues) of each structure was matched to the resolved sequence (which may have missing or altered residues) using MAFFT7 [36]. If the identity was over 90% and the chain ID matched the OPM, the sequence is matched to that chain; otherwise, it is matched to the chain with the highest identity.

3. Based on the matched sequences, the structures were repaired as for the general training set. During repair, unstructured termini were removed, and unresolved loops were inserted if the size is < 20 residues and replaced with 20 glycine residues otherwise.
4. DDOMAIN [37] is used to parse each structure into domains. Only TM domains, and TM-adjacent domains, were kept. This decreases the size of large, partially-TM systems, while maintaining potential interactions between TM and TM adjacent domains and improving the TMT class balance.
5. To establish if the repair procedure changed the input structures significantly, the repaired structures are aligned to their raw OPM structure [38], and the C_α atom Root Mean Square Deviation (RMSD) is calculated for non-loop residues (according to DSSP [39]). Proteins with RMSDs > 3 Å were removed.
6. Proteins were re-clustered at 70% sequence identity using CDhit, using the sequences of the repaired structures rather than the full sequence. Structures > 1000 residues or < 25 residues were removed. The final structures were superimposed to the original OPM entry to recover the membrane orientation.

The final OPM dataset contains 1165 continuous TMPs, with missing residues either repaired or replaced with poly-glycine loops. This is almost an order of magnitude larger than the dataset used for benchmarking CCTOP [1], a well-known transmembrane topology meta-predictor. While the OPM database is not an experimental database, it is a widely trusted source of membrane protein information. The OPM dataset entries used for training are single chains, since none of the primary predictors are able to handle multi-chain systems. In some cases, this single-chain limitation may alter ME labels, *e.g.*, for targets where multiple transmembrane segments interact in the biological unit. However, overall we expect this effect to be minor and opted to keep single-chain prediction to enable a fair comparison to existing methods. Because the dataset contains homologs of all TMPs, it is important to split it into training and test sets carefully. We clustered the dataset with MMSeqs2 [40] to 20% identity and separated the dataset by clusters into training and test. To do so, we used a multiple steepest descent algorithm, which calculates the distribution of different control variables and optimizes the distribution similarity between training- and test-set, while ensuring that, between the splits, no two proteins share more than 20% identity. This is done similarly as in [28]. The control variables were selected to minimize bias in the dataset splitting. We controlled for the relative proportion of TMHs and TMBs as well as distributions of protein length, N_{eff} ratio (amount of sequence information), and prediction difficulty (PHOBIUS Q3 score for TMHs and BOCTOPUS Q3 Score for TMBs).

The two datasets are denoted as the OPM training set and OPM test set, respectively. The OPM training set contains 933 proteins (80%), totaling 266880 residues (80%). The OPM test set contains 232 proteins (20%) and 66773 residues (20%). A single protein (1KQFC) was omitted from the test evaluation set, as it was also included in the general training set. The ratio of TMH to TMB proteins was 8.6:1 in both datasets. Between the training and test set, no two proteins have more than 20% identity, but the datasets are otherwise similar in terms of their difficulty, protein size, and TMH/TMB compositions.

Input Features. For the predictions of SS and RSA, the input features of TopProperty include:

1. Secondary Structure (SS) predicted by NetSurfP [14], SSpro5*ab initio* [16], DeepCNF-SS [10], PSIPRED4 [11], MUFOLD-SS [12], SPIDER3 [13], and SPIDER3-Single [41].
2. Solvent Accessibility (SA) predicted by NetSurfP [14], SANN [15], ACCpro5*ab initio* [16], AcconPred [17], SOLVPRED [11], SPIDER3 [13], and SPIDER3-Single [41].
3. Φ, Ψ, θ and τ angles predicted by SPIDER3 [13] and SPIDER3-Single [41].
4. Half-sphere Exposure (HSE) and Contact number (CN) predicted by AcconPred [17], SPIDER3 [13], and SPIDER3-Single [41], and residue disorder by DISOPRED [42].
5. Shannon entropy-based sequence conservation [43] and position-specific scoring matrix log-odds ratios calculated from an HHBLITS [34] alignment after Henikoff-Henikoff re-weighting [44].
6. Two global features are used to indicate the amount of sequence information. The N_{eff} ratio calculated as the number of effective sequences in the alignment (γ) divided by the number of residues in the target sequence (δ), and the N_{eff} score $(1+(\gamma/\delta)^{-0.5})^{-1}$ [45]. These measure the number of sequences in the alignment relative to the protein size on an absolute and normalized scale and indicate the difficulty of a prediction given the amount of sequence information.

For the predictions of TMT and ME, the input features of TopProperty include:

1. TMH topology predictions from TMHMM [3], HMMTOP[46], PHILIUS [47], PHOBIUS [48], POLYPHOBIUS [49], OCTOPUS [50], SPOCTOPUS [51], SCAMPI [52], TOPCONS [53], MEMSAT3 [54], and PROTEUS [55].
2. TMB topology predictions from BOCTOPUS [8], BETAWARE [9], and PROTEUS [55].
3. Coiled-coil predictions from COILS2 [56], helix orientation from LIPS [19], helix interaction predictions from RHYTHM [20], and four TMP potentials, three for TMHs [57-59] and one for TMBs [60].
4. As for SS and RSA, position-specific scoring matrix log-odds ratios calculated from an HHBLITS [34] alignment after Henikoff-Henikoff re-weighting [44].

As the OPM datasets contain close homologs to all TMPs, structure-based predictions from TOPCONS and PROTEUS were removed, since a structural match is always available and artificially inflates the performance of these predictors. For the same reason, the *ab initio* versions of SSpro5 and ACCpro5 were used for all datasets. The total number of input features are 123 for the SS/RSA DNNs and 129 for the TMT/ME DNNs, respectively.

Target Scores. The goal for TopProperty is to predict SS, RSA, TMT, and ME. For SS and RSA, the true values were calculated with DSSP [61], using the three-class classification α-helix, β-strand, and coil for SS. For TMT and ME labels, the true label for globular proteins cannot be obtained from databases like OPM (since these proteins do not appear in the OPM database). Thus, the prediction of these labels is trained on the entries of the OPM database for which such labels can be calculated. For TMT and ME, the true labels were calculated as follows:

TMT was defined as a three-class label for each residue: Inside the cytoplasm (label I), in the membrane slab (label M), or outside the cytoplasm (label O). The true labels are based solely on the position of the residue's C_α atom relative to the membrane slab according to OPM-defined protein orientation and membrane thickness and are independent of their SS class.

ME is defined as a binary class, where a residue is exposed to the membrane or not. Calculating the true label is more intricate since TMPs often function as pores or channels, and therefore simply using RSA does not suffice; residues inside the pore are accessible to the solvent but not the membrane. ME is thus determined for each TM residue segment, based on three variables: The SS and RSA of the residue determined by DSSP, and the angle between the C_α-C_β vector and the vector from the center of mass of the residues in the membrane to the C_α atom (denoted as the membrane angle). SS is used to distinguish TMHs from TMBs, RSA is used to determine surface residues, and the membrane angle is used to distinguish residues pointing to the inside of a channel or pore from residues pointing to the outside. The following rules are used to determine the true ME label:

1. Only residues with a TMT label M can be exposed.
2. α-helical membrane residues are exposed if they have ≥ 40 % RSA (fully exposed [17]) or have ≥ 10 % RSA (partially exposed [17]) and a membrane angle ≤ 100° and otherwise not.
3. Non-α-helical membrane residues are exposed if they have ≥10 % RSA and a membrane angle ≤ 100° and otherwise not. If the β-strand content of a TM segment is > 40%, the membrane exposure pattern is forced to fit an alternating sequence of exposed and not-exposed residues, matching the alternating ME pattern of TMBs [8].

4. If a TM segment is one residue long or does not have >40 % α-helix or > 40 % β-strand residues, all residues in the segment are assigned as exposed.

Proteins with more than four TM β-sheet segments are classified as TMBs. Otherwise, if they have any TM α-helix segment, they are classified as TMHs. If neither of the above is true, they are classified as TM-other. An example of true TMT and ME labels is shown in Figure 2.

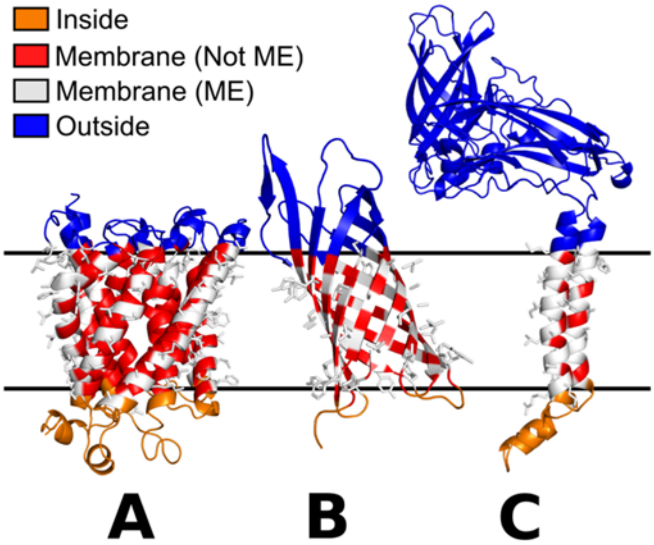


Figure 2. True TMT and ME labels according to OPM. **(A)** A TM α-helix bundle (PDB ID: 2B2H_A). **(B)** A TM β-barrel (PDB ID: 1P4T_A). **(C)** A protein anchored in the membrane by two TM α-helices (PDB ID: 3RGB_A). Residues labeled outside are shown in blue, residues labeled inside are shown in orange. Membrane residues are shown in red if they are not exposed to the membrane and white with side-chains if they are membrane exposed. Two black horizontal lines are a visual aid to indicate the membrane slab.

Quality Metrics. In all cases, the metrics were measured on a per-protein basis. For three-class SS prediction (α-helix, β-strand, coil), we use the Q3 score (fraction of correctly predicted labels) as well as the per-label and combined Matthews Correlation Coefficient (MCC) [62] as quality metrics. For RSA prediction, we use the Pearson's Correlation Coefficient (Pearson's R), as well as the Q3 score and MCCs for residue classification into buried ($RSA \leq 0.1$), medium ($0.1 < RSA \leq 0.4$), and exposed ($RSA > 0.4$) classes [17] as quality metrics. For TMT prediction, we use the Q3 score as well as the per-label and combined MCCs, while for ME prediction, we use the Q3 score as well as the binary MCC. In all per-label MCC calculations, if the predictions are equal to the true values, a value of 1 was assigned, irrespective of if the metric was poorly defined. In this way, predictions were considered as correctly classified where, for example, no β-strand or no buried residues are observed and also not predicted. Statistical significance of higher scores between TopProperty and other predictors was evaluated using the one-sided Wilcoxon signed-rank test as implemented in SciPy [63], using as an alternative hypothesis that the median TopProperty difference to the predictor with which it is compared is greater than 0.

DNNs. The neural networks used in TopProperty are implemented in python using Keras [64] with the TensorFlow [65] back-end and utilities from Sci-Kit-learn [66]. A modified ResNetv2 architecture [67] from the Keras-contrib Git-Hub repository [68] was used to compile the model, changing 2D convolution, averaging, and pooling layers for their 1D counterparts. Additionally, functions to fork the network after initial pooling and to converge the predictions into one main output were included, allowing to predict multiple outputs in the same network. The modified code can be downloaded along with the TopProperty datasets (http://dx.doi.org/10.25838/d5p-20). To prevent overtraining, a 50% drop-out ratio [26] was set for every convolution layer, evaluating for early stopping on a 20% split of the data to monitor the loss. To handle class imbalance, we use class reweighting between the different SS classes (α-helix, β-sheet, and coil) and TMT classes (inside, membrane, outside). Convergence of the training was enforced by a learning rate reduction on *plateaus* of 5 epochs, starting from 0.01 and reducing it in steps of √0.1 to a minimum of $1x10^{-5}$ for SS/SA and $1x10^{-7}$ for TMT/ME. For both networks, batches of 500 samples were fed, until all data was seen during an epoch.

The primary features are shaped as sliding window images with 123 feature channels for SS/RSA and 129 feature channels for TMT/ME DNNs and differing window sizes, where the target residue is located in the center of the window. As the predictions are made per residue, and to favor generalization across multiple proteins, each batch contained randomly ordered images from different proteins in the corresponding training set. For each window size, initial kernel sizes between 1 and the window size were evaluated, selecting the top performer for the final model.

The SS/RSA DNNs are an ensemble of four 1D DNN's trained to predict three-class SS and RSA simultaneously, with a structure equivalent to a ResNet50 and ResNet18, respectively. For SS predictions, a categorical cross-entropy loss function with a final softmax layer was used, while for the RSA predictions, a mean square error loss function and a linear final layer were implemented as is standard for regression tasks. The RSA predictions were passed to the SS predictions, setting SS as the main prediction (see below and Figure S1). A stride of 3 was used, with sliding window sizes of 1, 7, 11, and 23 residues and initial kernel sizes of 1, 7, 3, and 19, respectively. These networks thus predict SS and RSA with different amounts of local and non-local sequence information.

For the TMT/ME DNNs, an ensemble of five 1D DNNs were trained, predicting 3-class TMT and binary ME labels simultaneously. Both network forks are equivalent to ResNet50s, using categorical cross-entropy as the loss function and a softmax layer to generate the final predictions; the ME predictions were passed to the TMT dense layer, setting TMT as the main prediction (Figure S1). These networks are similar to the SS/RSA networks except they use a stride of 1 and sliding window sizes of 5, 9, 11, 17, and 23 residues with initial kernel sizes of 1, 1, 3, 1, and 23, respectively.

The output of each of the different DNNs in a given ensemble is combined into a single score for each residue using a weighted average according to the performance of each of the different DNNs across the training dataset (MCC values for classification networks, Pearson's R^2 for RSA predictions).

Results

To evaluate TopProperty performance for SS and RSA predictions, we compared TopProperty to the four best primary predictors on the NOUMENON test set and the OPM test set, respectively. The best SS primary predictors according to Q3 score were MuFold [12], SPIDER3 [13], PSIPRED4 [69], and DeepCNF-SS [10]. The best RSA primary predictors according to Pearson's R were SPIDER3 [13], SANN [70], SOLVPRED [69], and NetSurfP [71]. The same predictors but with different relative performances were identified for the OPM test set. On the OPM test set, we also compared TopProperty RSA predictions to TMP-SSurface-2 [72], a recently published method specifically trained on TMPs. Although TopProperty was not trained specifically on TMPs, it performed equivalent or better compared to TMP-SSurface-2. The performance in terms of Q3 scores and MCCs for the three SS classes are shown in Figure 3A and 3B for the NOUMENON test set and OPM test set, respectively. The RSA performance in terms of Pearson's R, as well as Q3 and MCCs for the three RSA classes, are shown in Figure 3C and 3D, respectively. Comparisons to all primary predictors can be found in Figures S2 and S3.

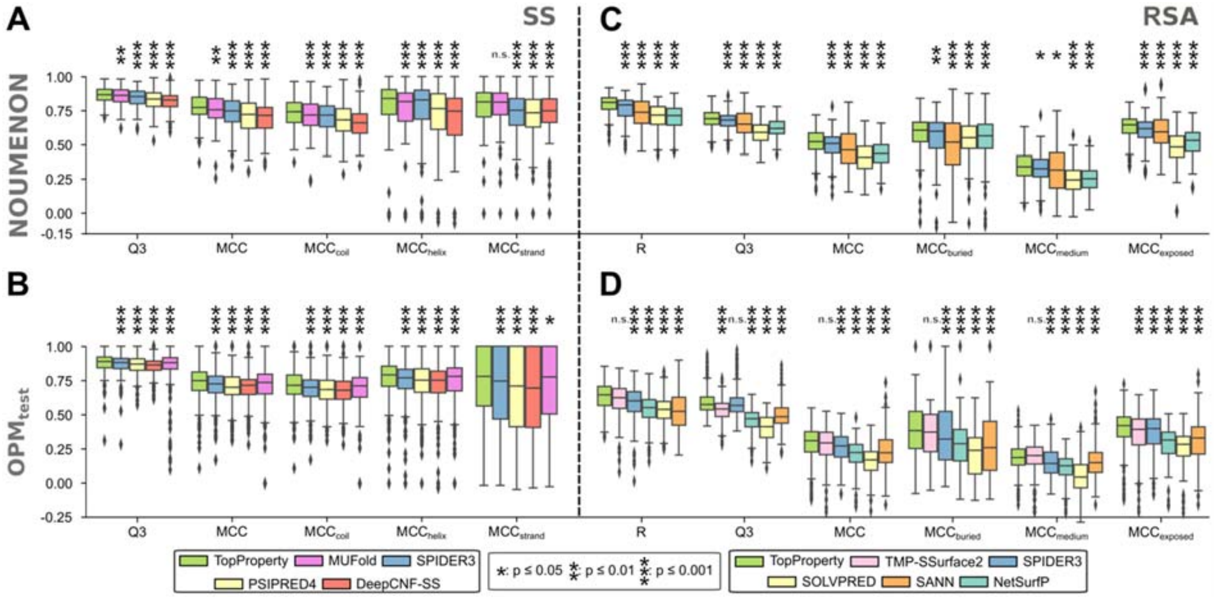


Figure 3. Performance of TopProperty SS and the four top-performing SS three-class primary predictors on the NOUMENON test set **(A)** and the OPM test set **(B)**, sorted according to the average Q3 score. Performance of TopProperty RSA and the four top-performing RSA primary predictors on the NOUMENON test set **(C)** and the OPM test set **(D)**, sorted according to the average Pearson's R. Significantly higher TopProperty scores are calculated using the Wilcoxon signed-rank test and indicated with asterisks (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$). For numerical values see Table S2.

Overall, TopProperty performs significantly better than all predictors in Q3 and MCC for SS and R for RSA, on both the NOUMENON and the OPM test sets. This result is also observed for per-label metrics, with the notable exception of MUFold predictions of β-strands on the NOUMENON data set, where no significant difference was identified, and SPIDER3 Q3 classification for per-residue buriedness. Regarding the latter, neither method was directly trained to classify buriedness, but to predict RSA directly, making the Pearson's R metric a more appropriate representation of the underlying model.

It is interesting to note that the SS prediction on the NOUMENON and OPM test sets is comparable between all predictors except for β-strands, where it has much higher variance in the predictions for TMPs. This is caused by the large TMH/TMB imbalance in the OPM datasets, where TMHs (which often have no β-strands) make up 90% of the proteins, but TMBs (which often have exclusively β-strands) make up 10%. For the RSA predictions, the performance on the OPM dataset is considerably worse than on the NOUMENON dataset. This is likely because for TMPs, solvent accessibility is more difficult to predict due to the exposure of hydrophobic residues on the surface, where they are oriented towards the membrane.

To evaluate TopProperty performance for TMT and ME predictions, we compared TopProperty to the best primary predictors for TMHs (Figure 4A) and to the TMB-specific predictors (Figure 4B), as well as to the popular meta-server CCTOP on the OPM test set. Comparisons to all primary predictors can be found in Figure S4. The best TMT primary predictors for TMHs according to Q3 score were SPOCTOPUS, OCTOPUS, and PolyPhobius. BOCTOPUS was better than BetAware as TMT primary predictor for TMBs according to Q3 score. Similarly as for SS, the TMT performance is evaluated in terms of Q3 scores and MCCs for the three TMT classes. TopProperty shows a better performance in Q3 and MCC on TMT classification for both TMHs and TMBs of the OPM test set. This is a notable result, as TMT predictors are in general specific to a certain TMP class (either TMH or TMB). Although CCTop is the second-best method for TMHs and shows comparable performance to TopProperty for the inside class, it is only on the fourth rank for TMBs with significantly lower scores. BOCTOPUS and BetAware, on the other hand, provide good predictions for TMBs, but not for TMHs (see Figure S4).

For ME predictions, only BOCTOPUS provides a classification of residues exposed to the membrane, and only for TMBs. Thus, we calculated ME performance for TMHs without any reference, and for TMBs, we compared TopProperty to BOCTOPUS (Figure 4C and 4D). Of the tested methods, TopProperty is the only one that performs well for either TM class.

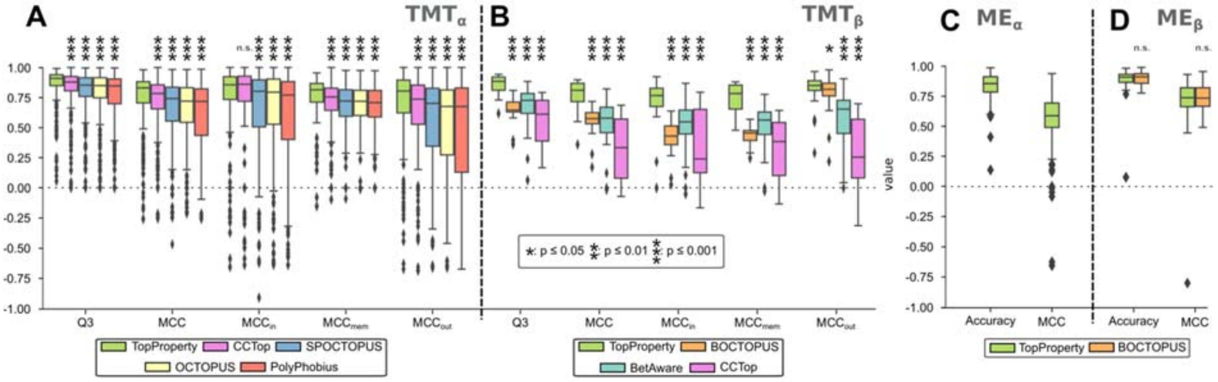


Figure 4. TopProperty TMT prediction performance for TMHs **(A)** and TMBs **(B)** and ME prediction performance for TMHs **(C)** and TMBs **(D)** on the OPM test set. Significantly higher TopProperty scores are calculated using the Wilcoxon signed-rank test and indicated with asterisks (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$). For numerical values see Table S3.

The TopProperty web server provides an easy way for users to get SS, RSA, TMT, and ME predictions for their protein of interest. It accepts input sequences between 30 and 1000 residues in length and provides the predictions in an easy-to-read alignment format and download file. TopProperty is implemented as part of the TopSuite server [73], which also has methods for domain boundary prediction [28], protein structure prediction, [74] and protein model quality estimation [74]. Figure 5 illustrates the TopProperty predictions displayed on the native structure with a subset of the output from the TopProperty web server and the true values shown for comparison. Experimental 3D structures are shown for clarity but do not reflect the output of the web server, since TopProperty does not perform protein folding.

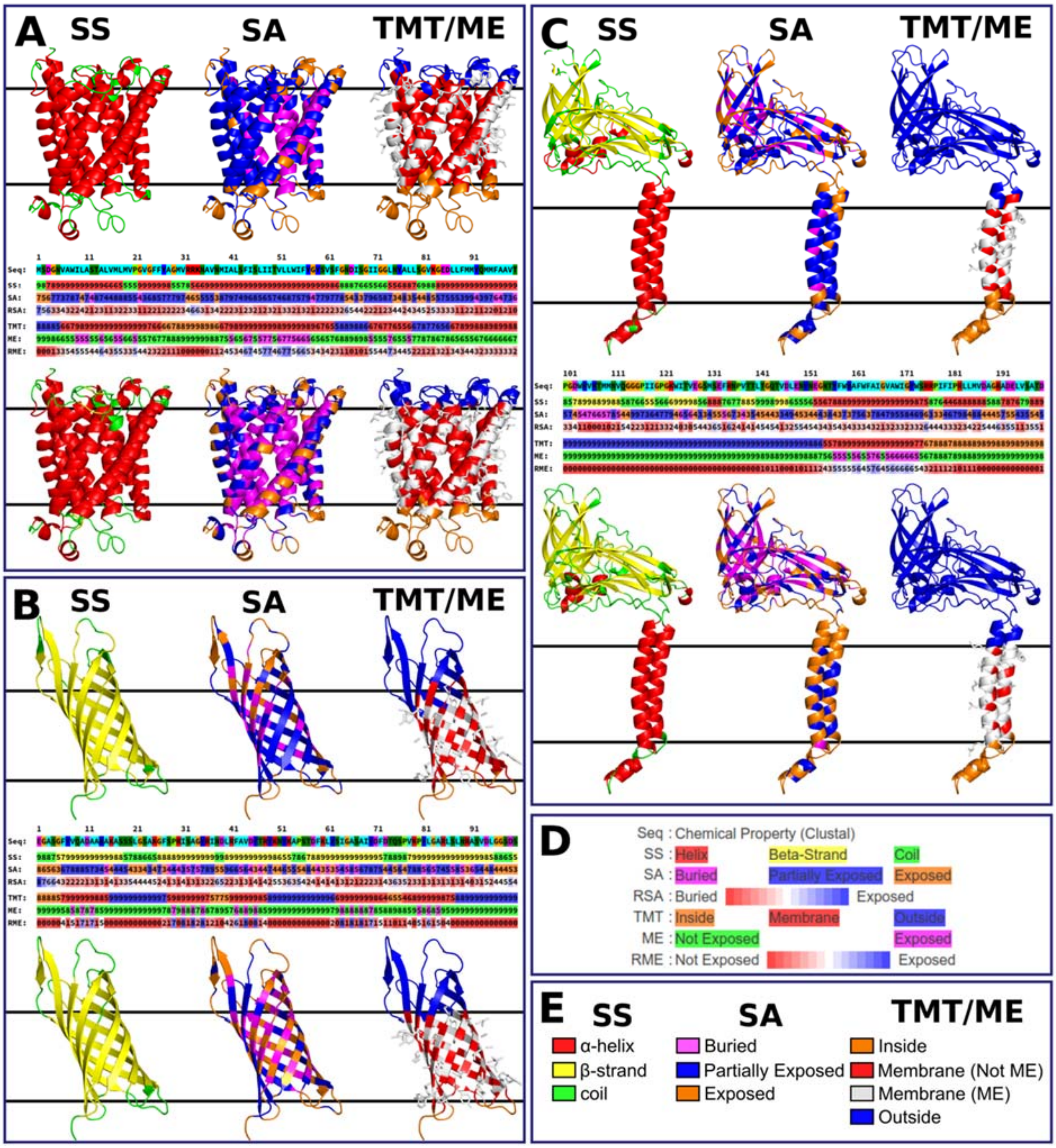


Figure 5. (A, B, C) Examples of TopProperty predictions. The native structure (top and bottom row) is colored by TopProperty-predicted (top row) or true (bottom row) properties. The color-coding on the structures is indicated in **(E)**. The two black lines are visual aids indicating the membrane slab. A subset of the TopProperty web server output for each protein is shown in the middle row indicating the predicted properties and confidences. The web server output is color-coded according to the predicted properties as indicated in **(D)**; numbers indicate the confidence of the prediction or the scale of the predicted property. RME is the relative membrane exposure and reflects the exposure probability scaled from 0 to 9. **(A)** PDB ID 2B2H_A (the last letter denotes the chain) is a TM α-helix bundle from the OPM test set. **(B)** PDB ID 1P4T_A is a TM β-barrel from the OPM test set. **(C)** PDB ID 3RGB_A from the OPM test set is anchored in the membrane by two TM α-helices.

Conclusion

We developed TopProperty, a meta-method for predicting three-class SS, RSA, TMT, and ME. The results are based on 27 primary predictors whose predictions are integrated using two ensembles of DNNs. TopProperty was trained on unbiased training dataset containing more than 4167 proteins and showed consistent improvement across all target values and quality metrics. Unlike similar methods such as CCTOP, TopProperty performance is independent of the type of input protein and works for both TMHs and TMBs. Furthermore, TopProperty provides consistently better SS and RSA predictions for both globular and transmembrane proteins and is a novel DNN method for predicting ME.

Other methods have been described for predicting directly or indirectly the membrane exposure of TMP residues, such as LIPS [19], RHYTHM [20], TMX [75], and more recently TMP-SSurface2 [72] and LCP [76]. From these predictors, LIPS and RHYTHM provide information regarding exposure per α-helix face, obscuring the per-residue signal, and lack information for TMBs; TMX is not available anymore. LCP is described in a preprint as a novel method that predicts the probability of interacting with lipids in a membrane environment, but is not available for testing. Lastly, TMP-SSurface2 predicts the RSA for residues placed in a membrane environment rather than a direct exposure classification. Interestingly, the Pearson's R in the same range as for our independent OPM test set (Figure 2D) was reported [72].

The recent release of AlphaFold [77] has fundamentally changed protein structure prediction as a field of research. While some evidence suggests that AlphaFold performs well on membrane proteins [78], the method still has limitations for proteins that are particularly big or have an oligomeric assembly. AlphaFold also does not provide information regarding the placement of a protein within the membrane bilayer. We expect that TopProperty predictions will be particularly useful in conjunction with methods such as AlphaFold to interpret and understand structure predictions of transmembrane proteins. For such targets, TopProperty can give insights into the orientation of the protein in the membrane and provide restraints for model refinement, e.g., within Rosetta [79].

The ability to predict multiple linear features with a single tool makes TopProperty a useful first step to describe and understand uncharacterized TMPs and globular proteins. TopProperty is available as a standalone and web server at https://cpclab.uni-duesseldorf.de/topsuite/, as are the datasets and DNN models (http://dx.doi.org/10.25838/d5p-20). The total run-time of a TopProperty job varies depending on the size of the protein in question and the load of the server, but generally, the user will receive a prediction within a few hours.

Acknowledgments

We are grateful to the developers of all primary programs used in this work for making their methods available as standalone to the scientific community. This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 267205415 – SFB 1208 (project A03 to HG) and by the Bundesministerium für Bildung und Forschung (BMBF) – Förderkennzeichen 031L0182 – InCelluloProtStruct to HG. We are grateful for computational support and infrastructure provided by the "Zentrum für Informations- und Medientechnologie" (ZIM) at the Heinrich Heine University Düsseldorf and the computing time provided by the John von Neumann Institute for Computing (NIC) to HG on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user ID: HKF7, VSK33).

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: XXX. The supporting information includes: All PDB IDs of proteins used for training and evaluating TopProperty. Schematic of the DNN architecture used in TopProperty, performance comparison between TopProperty and all primary predictors for SS and RSA predictions, performance comparison between TopProperty and all primary predictors for TMT predictions, numerical data for Figures 3 and 4.

Author Contributions

DM and SS jointly developed the method, implemented primary predictors, performed benchmark calculations, and wrote the manuscript. DM curated datasets and performed post-processing and method maintenance. SS implemented and trained all DNNs and performed data analysis. FK implemented the TopProperty web server. HG conceived the study, supervised and managed the project, secured funding and resources for the project, and revised the manuscript. All authors reviewed and approved the manuscript. The authors declare no competing interests.

Data and software availability

TopProperty is available as a standalone and web server at https://cpclab.uni-duesseldorf.de/topsuite/ . The datasets used and DNN models generated are available at http://dx.doi.org/10.25838/d5p-20 .

References

1. Dobson, L.; Reményi, I.; Tusnády, G. E., CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.* **2015,** *43* (W1), W408-W412.

2. Wallin, E.; Heijne, G. V., Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science* **1998,** *7* (4), 1029-1038.

3. Krogh, A.; Larsson, B.; Von Heijne, G.; Sonnhammer, E. L., Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **2001,** *305* (3), 567-580.

4. Dobson, L.; Reményi, I.; Tusnády, G. E., The human transmembrane proteome. *Biol. Direct* **2015,** *10* (1), 1-18.

5. Tusnády, G. E.; Dosztányi, Z.; Simon, I., Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* **2004,** *20* (17), 2964-2972.

6. Kozma, D.; Simon, I.; Tusnady, G. E., PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* **2012,** *41* (D1), D524-D529.

7. Rask-Andersen, M.; Almen, M. S.; Schioth, H. B., Trends in the exploitation of novel drug targets. *Nature Reviews Drug Discovery* **2011,** *10* (8), 579-590.

8. Hayat, S.; Elofsson, A., BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. *Bioinformatics* **2012,** *28* (4), 516-522.

9. Savojardo, C.; Fariselli, P.; Casadio, R., BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. *Bioinformatics* **2013,** *29* (4), 504-505.

10. Wang, S.; Peng, J.; Ma, J.; Xu, J., Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.* **2016,** *6*, 18962.

11. Mcguffin, L. J.; Bryson, K.; Jones, D. T., The PSIPRED protein structure prediction server. *Bioinformatics* **2000,** *16* (4), 404-405.

12. Fang, C.; Shang, Y.; Xu, D., MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. *Proteins: Struct. Funct. Bioinform.* **2018,** *86* (5), 592-598.

13. 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* (18), 2842-2849.

14. Petersen, B.; Petersen, T. N.; Andersen, P.; Nielsen, M.; Lundegaard, C., A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **2009,** *9* (1), 51.

15. Joo, K.; Lee, S. J.; Lee, J., Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins: Struct. Funct. Bioinform.* **2012,** *80* (7), 1791-1797.

16. Magnan, C. N.; Baldi, P., SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* **2014,** *30* (18), 2592-2597.

17. Ma, J.; Wang, S., AcconPred: Predicting solvent accessibility and contact number simultaneously by a multitask learning framework under the conditional neural fields model. *BioMed research international* **2015,** *2015*.

18. Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D., Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **2020,** *117* (3), 1496-1503.

19. Adamian, L.; Liang, J., Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Struct. Biol.* **2006,** *6* (1), 13.

20. Hildebrand, P. W.; Lorenzen, S.; Goede, A.; Preissner, R., Analysis and prediction of helix–helix interactions in membrane channels and transporters. *Proteins: Struct. Funct. Bioinform.* **2006,** *64* (1), 253-262.

21. Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J., Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biology* **2017,** *13* (1), e1005324.

22. Schaarschmidt, J.; Monastyrskyy, B.; Kryshtafovych, A.; Bonvin, A. M., Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Struct. Funct. Bioinform.* **2018,** *86* (S1), 51-66.

23. Hochreiter, S., The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **1998,** *6* (02), 107-116.

24. Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y.; Valencia, A., Accurate Prediction of Protein Contact Maps by Coupling Residual Two-Dimensional Bidirectional Long Short-Term Memory with Convolutional Neural Networks. *Bioinformatics* **2018**.

25. He, K.; Zhang, X.; Ren, S.; Sun, J. In *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp 770-778.

26. Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. R., Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* **2012**.

27. Zhang, B.; Li, J.; Lu, Q., Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics* **2018,** *19* (1), 293.

28. Mulnaes, D.; Golchin, P.; Koenig, F.; Gohlke, H., TopDomain: Exhaustive Protein Domain Boundary Metaprediction Combining Multisource Information and Deep Learning. *J. Chem. Theory Comput.* **2021**.

29. Orlando, G.; Raimondi, D.; Vranken, W., Observation selection bias in contact prediction and its implications for structural bioinformatics. *Sci. Rep.* **2016,** *6*, 36679.

30. Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I., OPM: orientations of proteins in membranes database. *Bioinformatics* **2006,** *22* (5), 623-625.

31. Wang, G.; Dunbrack Jr, R. L., PISCES: a protein sequence culling server. *Bioinformatics* **2003,** *19* (12), 1589-1591.

32. Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D., Protein structure prediction using Rosetta. In *Methods Enzymol.*, Elsevier: 2004; Vol. 383, pp 66-93.

33. Webb, B.; Sali, A., Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinf.* **2014**, 5.6. 1-5.6. 32.

34. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J., HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **2012,** *9* (2), 173-175.

35. Li, W.; Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006,** *22* (13), 1658-1659.

36. Katoh, K.; Standley, D. M., MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **2013,** *30* (4), 772-780.

37. Zhou, H.; Xue, B.; Zhou, Y., DDOMAIN: Dividing structures into domains using a normalized domain–domain interaction profile. *Protein Science* **2007,** *16* (5), 947-955.

38. Zhang, Y.; Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005,** *33* (7), 2302-2309.

39. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-2637.

40. Steinegger, M.; Söding, J., Clustering huge protein sequence sets in linear time. *Nature communications* **2018,** *9* (1), 1-8.

41. Heffernan, R.; Paliwal, K.; Lyons, J.; Singh, J.; Yang, Y.; Zhou, Y., Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *JCoCh* **2018**.

42. Ward, J. J.; Mcguffin, L. J.; Bryson, K.; Buxton, B. F.; Jones, D. T., The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **2004,** *20* (13), 2138-2139.

43. Capra, J. A.; Singh, M., Predicting functionally important residues from sequence conservation. *Bioinformatics* **2007,** *23* (15), 1875-1882.

44. Henikoff, S.; Henikoff, J. G., Position-based sequence weights. *J. Mol. Biol.* **1994,** *243* (4), 574-578.

45. Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C., Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **2011,** *6* (12), e28766.

46. Tusnady, G. E.; Simon, I., Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* **1998,** *283* (2), 489-506.

47. Reynolds, S. M.; Käll, L.; Riffle, M. E.; Bilmes, J. A.; Noble, W. S., Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Computational Biology* **2008,** *4* (11), e1000213.

48. Käll, L.; Krogh, A.; Sonnhammer, E. L., A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **2004,** *338* (5), 1027-1036.

49. Käll, L.; Krogh, A.; Sonnhammer, E. L., An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* **2005,** *21* (suppl_1), i251-i257.

50. Viklund, H.; Elofsson, A., OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* **2008,** *24* (15), 1662-1668.

51. Viklund, H.; Bernsel, A.; Skwark, M.; Elofsson, A., SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **2008,** *24* (24), 2928-2929.

52. Bernsel, A.; Viklund, H.; Falk, J.; Lindahl, E.; Von Heijne, G.; Elofsson, A., Prediction of membrane-protein topology from first principles. *Proceedings of the National Academy of Sciences* **2008,** *105* (20), 7177-7181.

53. Bernsel, A.; Viklund, H.; Hennerdal, A.; Elofsson, A., TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* **2009,** *37* (suppl_2), W465-W468.

54. Jones, D. T., Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **2007,** *23* (5), 538-544.

55. Montgomerie, S.; Sundararaj, S.; Gallin, W. J.; Wishart, D. S., Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* **2006,** *7* (1), 301.

56. Lupas, A., Predicting coiled-coil regions in proteins. *Current Opinion in Structural Biology* **1997,** *7* (3), 388-393.

57. Park, Y.; Helms, V., On the derivation of propensity scales for predicting exposed transmembrane residues of helical membrane proteins. *Bioinformatics* **2007,** *23* (6), 701-708.

58. Samatey, F. A.; Xu, C.; Popot, J.-L., On the distribution of amino acid residues in transmembrane alpha-helix bundles. *Proceedings of the National Academy of Sciences* **1995,** *92* (10), 4577-4581.

59. Pilpel, Y.; Ben-Tal, N.; Lancet, D., kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.* **1999,** *294* (4), 921-935.

60. Freeman Jr, T. C.; Wimley, W. C., A highly accurate statistical approach for the prediction of transmembrane β-barrels. *Bioinformatics* **2010,** *26* (16), 1965-1974.

61. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** *22* (12), 2577-2637.

62. Jurman, G.; Riccadonna, S.; Furlanello, C., A comparison of MCC and CEN error measures in multi-class prediction. *PLoS One* **2012,** *7* (8), e41882.

63. Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; Van Der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, I.; Feng, Y.; Moore, E. W.; Vanderplas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; Van Mulbregt, P.; Scipy, C., SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020,** *17* (3), 261-272.

64. Chollet, F., Keras. 2015.

65. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. In *TensorFlow: A System for Large-Scale Machine Learning*, OSDI, 2016; pp 265-283.

66. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011,** *12* (Oct), 2825-2830.

67. He K., Z. X., Ren S., Sun J. , Identity Mappings in Deep Residual Networks. *Lecture Notes in Computer Science* **2016,** *9908*.

68. Keras community contributions. https://github.com/keras-team/keras-contrib.

69. Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999,** *292* (2), 195-202.

70. Joo, K.; Lee, S. J.; Lee, J., Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins: Structure, Function, and Bioinformatics* **2012,** *80* (7), 1791-1797.

71. 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., NetSurfP-2.0: Improved prediction of protein

structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics* **2019,** *87* (6), 520-527.

72. Liu, Z.; Gong, Y.; Guo, Y.; Zhang, X.; Lu, C.; Zhang, L.; Wang, H., TMP-SSurface2: A Novel Deep Learning-Based Surface Accessibility Predictor for Transmembrane Protein Sequence. *Frontiers in Genetics* **2021,** *12*, 328.

73. Mulnaes, D.; Koenig, F.; Gohlke, H., TopSuite Web Server: A Meta-Suite for Deep-Learning-Based Protein Structure and Quality Prediction. *J. Chem. Inf. Model.* **2021,** *61* (2), 548-553.

74. Mulnaes, D.; Porta, N.; Clemens, R.; Apanasenko, I.; Reiners, J.; Gremer, L.; Neudecker, P.; Smits, S. H.; Gohlke, H., TopModel: Template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks. *J. Chem. Theory Comput.* **2020,** *16* (3), 1953-1967.

75. Park, Y.; Hayat, S.; Helms, V., Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC Bioinformatics* **2007,** *8*, 302.

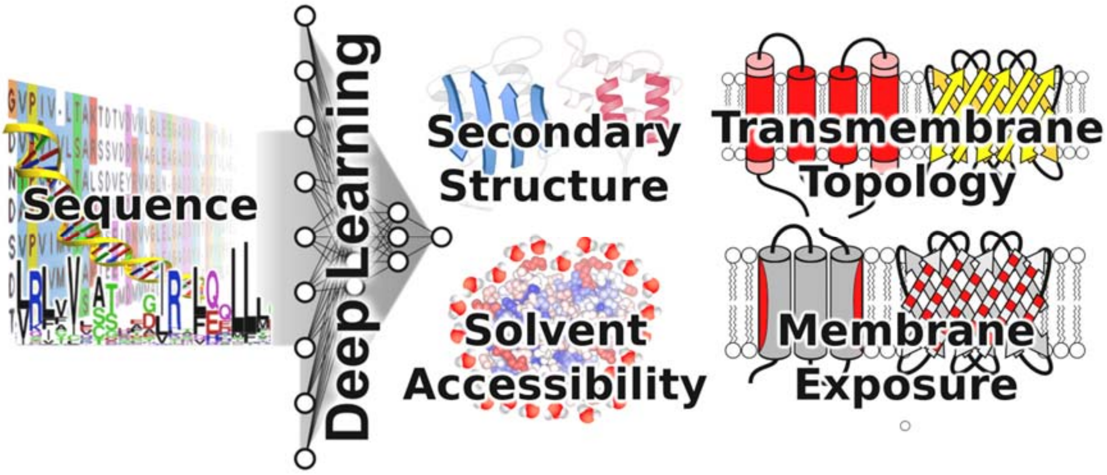
76. Wang, L.; Zhang, J.; Wang, D.; Song, C., Lipid contact probability: an essential and predictive character for the structural and functional studies of membrane proteins. *bioRxiv* **2021**.

77. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., Highly accurate protein structure prediction with AlphaFold. *Natur* **2021**, 1-11.

78. Hegedűs, T.; Geisler, M.; Lukács, G.; Farkas, B., AlphaFold2 transmembrane protein structure prediction shines. *bioRxiv* **2021**.

79. Alford, R. F.; Fleming, P. J.; Fleming, K. G.; Gray, J. J., Protein structure prediction and design in a biologically realistic implicit membrane. *Biophys. J.* **2020,** *118* (8), 2042-2055.

Table of Content Graphic


Sequence
Deep Learning
Secondary Structure
Transmembrane Topology
Solvent Accessibility
Membrane Exposure