Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Machine learning in computational modelling of membrane protein sequences and structures: From methodologies to applications

Jianfeng Sun [a], Arulsamy Kulandaisamy [b], Jacklyn Liu [c], Kai Hu [d], M. Michael Gromiha [b,*], Yuan Zhang [d,*]

[a] Botnar Research Centre, Nuffield Department of Orthopedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Headington, Oxford OX3 7LD, UK
[b] Department of Biotechnology, Bhupat and Jyoti Mehta School of BioSciences, Indian Institute of Technology Madras, Chennai 600 036, Tamilnadu, India
[c] UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, UK
[d] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan 411105, China

ARTICLE INFO

Article history:
Received 16 November 2022
Received in revised form 16 January 2023
Accepted 25 January 2023
Available online 28 January 2023

Keywords:
Membrane proteins
Machine learning
Databases
Prediction tools
Interaction sites
Mutations
Diseases

ABSTRACT

Membrane proteins mediate a wide spectrum of biological processes, such as signal transduction and cell communication. Due to the arduous and costly nature inherent to the experimental process, membrane proteins have long been devoid of well-resolved atomic-level tertiary structures and, consequently, the understanding of their functional roles underlying a multitude of life activities has been hampered. Currently, computational tools dedicated to furthering the structure-function understanding are primarily focused on utilizing intelligent algorithms to address a variety of site-wise prediction problems (e.g., topology and interaction sites), but are scattered across different computing sources. Moreover, the recent advent of deep learning techniques has immensely expedited the development of computational tools for membrane protein-related prediction problems. Given the growing number of applications optimized particularly by manifold deep neural networks, we herein provide a review on the current status of computational strategies mainly in membrane protein type classification, topology identification, interaction site detection, and pathogenic effect prediction. Meanwhile, we provide an overview of how the entire prediction process proceeds, including database collection, data pre-processing, feature extraction, and method selection. This review is expected to be useful for developing more extendable computational tools specific to membrane proteins.

Contents

[*] Corresponding authors.
E-mail addresses: gromiha@iitm.ac.in (M.M. Gromiha), yuanz@xtu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.csbj.2023.01.036

1. Introduction

Broadly, membrane proteins are pumps, channels, and carriers that regulate the permeability of plasma membranes [1,2]. They synergize with one another to ensure the functionality of plasma membranes. For example, pumps build electrochemical gradients across membranes that are then used for channels to permeate ions with maintained gradients taken as energy sources by carriers to transport substrates [3–6]. The subclasses within each of the three functional groups can possess different roles. For example, several types of ligand-gated ion channels are destined for controlling changes in membrane potential [7,8] and many voltage-gated channels are earmarked for mediating membrane repolarization [9]. The precise concentrations of the substrates required for life activities in various sealed cellular compartments necessitate the checkpoint role of membrane proteins in the obstruction and clearance of external molecules and ions [1]. While membrane proteins constitute approximately 20–30% gene coding proteins [10–12], the number of their available structures is less than 3% of all experimentally determined structures [13,14]. This is largely due to their long amino acid sequences and their hydrophobic surfaces, which makes experimentally determining their structures incredibly challenging compared to soluble proteins [15,16]. For example, efforts in resolving the SCN5A-encoded human cardiac voltage-gated sodium channel Nav1.5, 2016 amino acids long, have yielded only a handful of fragmented structures for a limited number of functional regions mostly at its C-terminal domain [17–24]. Crucially, the lack of structural information for this large volume of proteins renders the study of their functions incredibly challenging. Thus, there is a dire need for the development of computational approaches for the different types of sequence-, structure-, and function-related analyses of membrane proteins.

There has been a sharp increase in the number of computational applications in recent years with a notable acceleration in the development of new intelligent algorithms [25]. In particular, the immense success in predicting highly accurate 3D protein structures by AlphaFold2 has given fresh impetus in related fields [26]. The central role of deep learning has become indisputable and its use in biological applications has evolved very quickly. For example, deep residual neural networks in AlphaFold have rapidly turned to much deeper and more complex neural networks infested with multiple attention-based modules in AlphaFold2 [26,27]. Two important findings [28,29] released very recently have made an effort to collect, annotate, evaluate, and visualize the topologies and membrane localization of transmembrane (TM) protein structures predicted by AlphaFold2. Furthermore, TM protein topology prediction has been improved greatly by adopting a nature language processing (NLP) advance [30], ProtT5 [31], which is a protein language embedding technique pretraining a series of auto-regressive and auto-encoder models [32].

In addition to improvements in methodology, informative biological features also play an indispensable role in boosting the performance of many kinds of protein-related prediction problems [33,34]. A remarkable inroad has been made over the past decade into the accurate modelling of relationships between residues by taking advantage of the direct coupling analysis (DCA) method to validate the phenomenon that evolutionary strengths between residues can be regarded as their spatial proximities [35–37]. It provides a bridge between statistical constraints and residue spatial proximities in structures, which makes it highly useful in tackling structure-related prediction problems, such as residue contact prediction [36] and interaction site prediction [38]. Together with deep learning methods, the utility of these coevolutionary features has been demonstrated in predicting residue contacts (note that Alphafold2 is dominant in this area) [39] and interaction sites [38] in TM proteins.

Although the methodological improvement continues to leap year by year, performing membrane protein-related prediction tasks remains challenging due to a variety of issues, such as feature-limited membrane protein databases [40], inadequate structure data [13], and additional attention to the establishment of membrane proteins in oligomer form compared to soluble proteins [41–44]. These also play a crucial role in improving prediction performance. Recently, Zaucha et al. [45] and Kulandaisamy et al. [46] comprehensively reviewed computational databases, tools, and related analyses, which improved the understanding of mechanisms of disease-causing mutations in membrane protein structures and functions but did not expand beyond this biological context. In this review, we seek to outline the sweeping changes of both machine learning approaches and implications from recent biological findings in membrane protein structure-related studies. First, we summarize currently available membrane protein databases, residue coevolution in different prediction problems, modelling processes, and machine learning method development. Then, we systematically investigate the current states of four established membrane protein computational problems: 1) membrane protein type classification, 2) transmembrane protein topology prediction, 3) transmembrane protein interaction site prediction, and 4) pathogenic effect prediction of mutations in transmembrane protein sites mainly from a machine learning perspective. Furthermore, we recapitulate the

Table 1
Membrane protein databases.

Name	Source	Year (first)	Number of structures	Version	Citation
PDBTM	http://pdbtm.enzim.hu	2005	7692 (α-helical: 7123, β-barrel: 509)	09.16.2022	[41,62]
OPM	https://opm.phar.umich.edu	2006	6568	03.10.2022	[63,65]
mpstruc	https://blanco.biomol.uci.edu/mpstruc	2009	6415 (Coordinate)	09.25.2022	[72,73]
EncoMPASS	https://encompass.ninds.nih.gov	2018	2344	06.24.2021	[69]
MemProtMD	http://memprotmd.bioch.ox.ac.uk	2018	> 5000	-	[66,67]
PerMemDB	http://83.212.109.111:8088/permemdb	2020	231770 (UniProt or predicted)	v1.3	[68]
Membranome3.0	https://membranome.org	2022	5758	3.0	[70,71,74]

drug-target interaction prediction given the pharmaceutical significance of membrane proteins, which are targeted by more than 50% commercially available drugs. In the meantime, using the Drugbank [47] and MutHTP [48] databases, we show several basic data analyses to appreciate the knowledge about membrane protein distributions and disease-causing mutations in relation to human transmembrane protein families, respectively.

2. Data preparation, feature extraction, and methodology

2.1. Membrane proteins

Membrane proteins, comprised of integral and peripheral membrane proteins, are replete with intrinsic biochemical and biophysical properties that are crucial in manifold biological functions [49], such as signal transduction [50] and enzymatic activity [51]. Integral membrane (IM) proteins, which comprise a substantial proportion of the membrane proteome, are permanently embedded in lipid bilayers [52–54], while peripheral membrane (PM) proteins often bind transiently to the surface and are comparatively rare [55–57]. Both types of proteins morphologically interact with plasma lipids in many ways [49,58,59]. For example, some IM proteins cross the lipid bilayer (i.e., transmembrane (TM) proteins) with domains exposed to both the extracellular and intracellular space, respectively, while others contain only one exposed domain, with the remainder of the protein buried within the bilayer. TM proteins are the most common type of IM proteins and a large majority of them are characterized by a bundle of α-helices facing one another within biological membranes [60,61]. According to the PDBTM database (version: 09.02.2022) [41], 14/15 TM proteins are formed in an α-helical fashion with a small proportion demonstrating β-barrel morphology. Given this variety, there have been many methods established for predicting the different properties of α-helical TM proteins.

2.2. Structural databases

The structural databases summarized in Table 1, such as PDBTM [41,62], OPM [63], and *mpstruc*, altogether provide a plethora of information regarding different types of membrane proteins, which serve as fundamental materials for training machine learning models and help to ensure model quality. These databases are often constructed with a number of built-in algorithms to automatically and exclusively derive membrane protein structures from the protein data bank (PDB) and, possibly, refine the structures thereafter, such as by calculating the orientation of membrane planes or filling missing atoms. PDBTM is a database of α-helical and β-barrel TM protein structures with geometrically localized membrane planes determined by the TMDET algorithm [64]. Residues in TM regions of each protein are annotated with H or B if this protein belongs to an α-helical or β-barrel type. Furthermore, the database provides detailed annotations of the remaining protein structure, such as interfacial helices (i.e., parallel with the membrane) and membrane loops (i.e., re-entrant types with both ends towards the same side of the membrane). To provide a comprehensive source of membrane proteins, the OPM database developed the PPM 3.0 algorithm [65] (available at https://opm.phar.umich.edu/ppm_server3) to collect both IM and PM proteins by the orientation of membrane planes, within which all TM amino acids are annotated and determined by minimizing a transfer energy function. Besides, it gives each protein its hydrophobic thickness and coordinates of curved membrane boundaries and automatically calculates the fit of a protein to a planar or spherical bilayer, which has been updated in its latest version. Unlike these two databases, *mpstruc* stores weekly updated membrane protein structures collected manually from PDB without the calculated membrane orientation but with functional and taxonomic information. A statistical comparison of the three databases with regards to sequences, structures, and functions was performed in a recent study [40]. Furthermore, other membrane protein databases have emerged over the past decade, including MemProtMD [66,67], PerMemDB [68], and EncoMPASS [69]. MemProtMD is an archive of membrane-embedded proteins while PerMemDB is a repository of PM proteins. EncoMPASS focuses on symmetries and homologue similarities of single-chain-based membrane protein structures collected from OPM [65]. In addition, since half of TM proteins are single-pass (bitopic) and, themselves, constitute a functionally diverse group [70] researchers have started building and organizing a further proprietary database exclusive to these: Membranome 3.0 [71].

2.3. Datasets

Once raw data are fetched from the membrane protein structural databases, a highly stringent pre-processing step is undertaken to yield a dataset suitable for training machine learning models [75]. This serves to minimize data redundancy, corruption, etc. This procedure is also beneficial for downstream machine-learning optimization processes since it allows for a level of objectivity based on the used test datasets and further increases the possibility of an improved model generalization ability based on future datasets. This pre-processing often encompasses data cleaning to remove corrupted data (e.g., delete obsolete protein entries or those lacking structural annotations), select proteins satisfying a research goal of interest (e.g., the number of interaction or mutation sites per protein above a pre-defined cut-off), and reduce protein redundancy based on sequence identities or, more stringently, structural similarities, and other quality control procedures (e.g., experimental methods or resolution). The schematic illustration is shown in Fig. 1.

2.4. Feature preparation

Most features used for characterizing globular proteins can be similarly used for membrane proteins as they are commonly seen in describing the sequence, structural, physiochemical, positional, and evolutionary properties of a protein. However, considering the topological difference (i.e., membranous vs. intra- and extra-cellular) between membrane proteins and globular proteins, there are still a variety of features that need to be considered for the former. For example, since TM proteins are anchored to the membrane by α-helices and, thereby, interact with lipids [76–78], it is necessary to

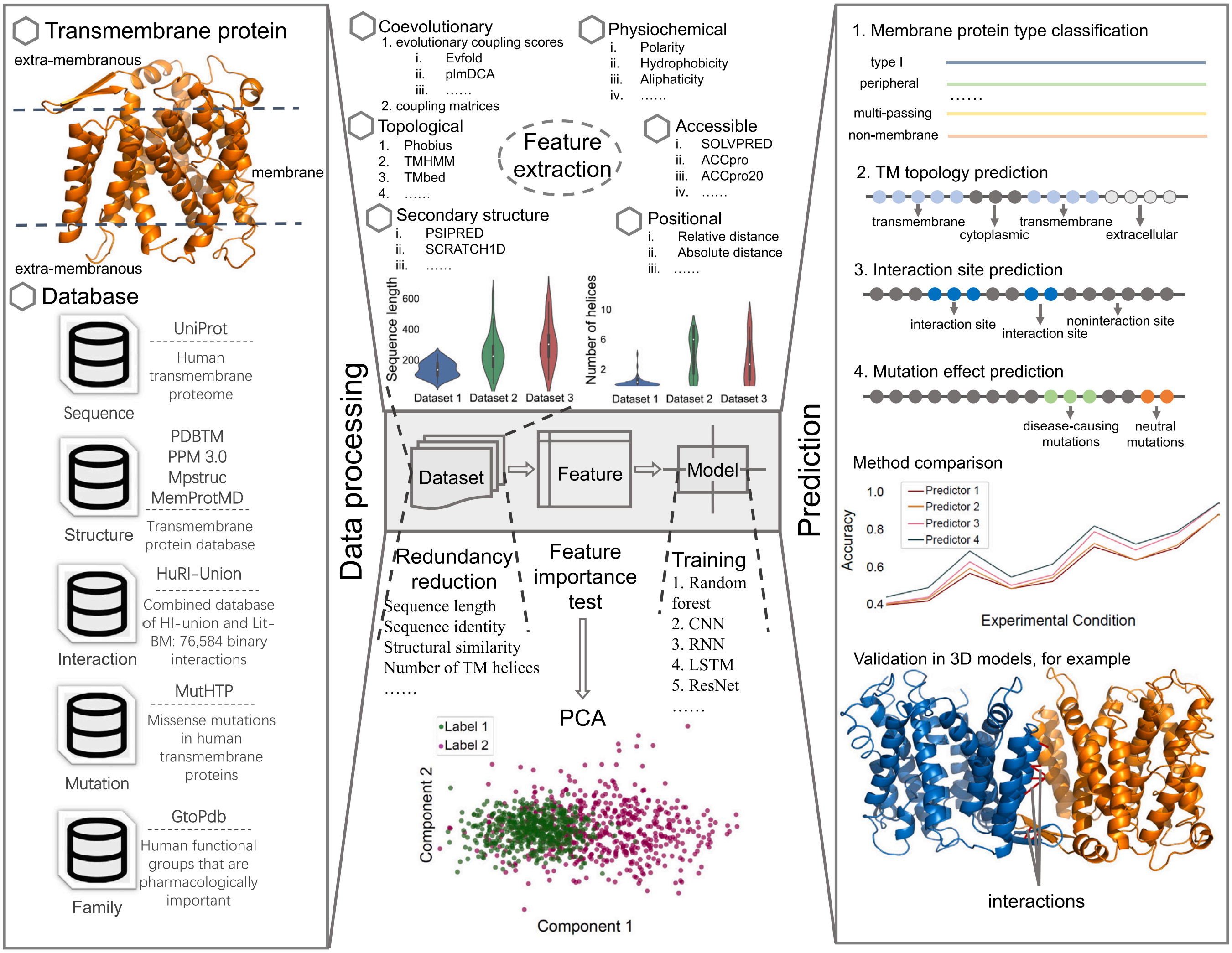

Fig. 1. Schematic illustration of the development of machine learning methods for membrane protein-related prediction problems. It includes database preparation, data pre-processing, feature extraction, problem description, and machine learning. The sequence, structural, interaction, mutation, and/or protein family databases will optionally be used depending on how to organize a prediction process. For example, the sequence database (UniProt) and/or the structural databases (e.g., PDBTM) are involved in the process of predicting structures, interactions, and mutations, while the protein type classification process may only require sequence information from UniProt. During data pre-processing, the redundancy between sequences or structures has to be removed based on sequence identity or structural similarity. Prior to model training, protein or residue features, which are biologically relevant with a prediction problem and are informative to train machine learning models, can be shortlisted by using feature importance test techniques, such as principal component analysis (PCA). The protein features mainly include amino acid compositions, while the residue features mainly include positional, topological, coevolutionary, and physiochemical properties. The feature importance test is crucial to improve the performance of the resulting models. The protein type classification task requires protein features as input while the topology prediction, interaction site prediction, and mutation effect prediction tasks require residue features as input. After model training, a variety of evaluation metrics (e.g., precision and recall) should be adopted for quality control.

study a variety of lipid-accessible protein properties in practice, such as the helix orientation relative to lipids. We detail a helix orientation prediction process below, employed by the LIPS (lipid-facing surface) method [79]. It illustrates the close link between the specialized features and some of the biological processes that the membrane proteins perform. Importantly, TM proteins are enriched for coiled coils in TM regions [80], which consist of seven periodically occurring amino acid residues, termed heptad repeats, each represented by ABCDEFG [81,82] (Fig. 2a). As shown in Fig. 2b, using a heptad repeat, seven helical faces are generated if each of the seven residues takes turns being thought of as an anchoring residue. As illustrated by Adamian and Liang [79], each anchoring residue is complemented by two residues (occurring two positions apart from the anchoring residue), which together constitute one of the seven surfaces. Therefore, starting from the first residue A, the formed 7 helical faces are ADE, BEF, CFG, DGA, EAB, FBC, and GCD. In the LIPS (lipid-facing surface) pipeline, sliding from the first position in a given TM protein sequence, each residue is serially partitioned into one of the seven helical faces and assigned an entropy and a lipophilicity score. These scores, for each helical face, are then incorporated to yield the LIPS score used to estimate its helix orientation. As the majority of residues involved in interactions between TM helices can be aligned to the heptad repeats [79], either or both the face-level LIPS scores and the residue-level lipophilicity scores may be helpful in identifying interaction sites in TM proteins. In the MBPred work, the importance of using the helical face-related scores for interaction site prediction is partly demonstrated by the mean decrease in impurity (i.e., also called Gini importance) and the leave-one-out test [38]. We have recently provided a Python interface to access the LIPS method, which will shortly be available at https://github.com/2003100127/tmkit. Furthermore, the physiochemical features of residues within transmembrane proteins should also be considered important in various TM protein-related prediction problems. For example, interactions between buried polar inter-helical residues are commonplace in TM regions [83] and hydrophobic interactions can help to judge the orientations of

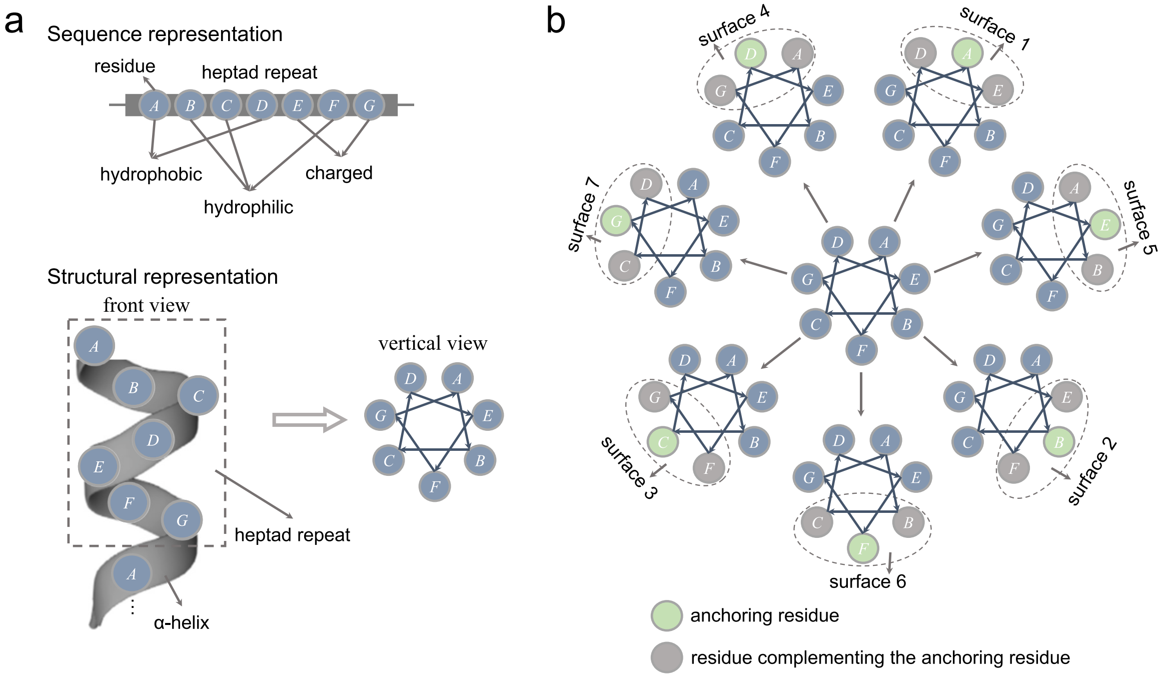

Fig. 2. Schematic illustration of helical surfaces generated using the LIPS method. (a) shows the representations of the heptad repeat (seven residues ABCDEFG) in sequence and structural contexts. (b) shows the seven canonical surfaces of transmembrane α-helices, which are generated by taking each of the seven residues as an anchoring residue. Each surface consists of an anchoring residue and two residues complementing the anchoring residue. The seven surfaces are ADE, BEF, CFG, DGA, EAB, FBC, and GCD.

peripheral proteins in membranes [84]. Finally, the different feature vectors are concatenated together into one for a residue or a TM protein to be taken as input for prediction models (Fig. 1).

2.5. Coevolution

Direct coupling analysis (DCA) is a statistical inference technique, which has notably improved the identification of residue contacts by differentiating, both directly and indirectly, coupled residues, making use of residue coevolution [85–87]. It starts with the premise that residues in spatial contact are often co-evolved [88,89]. When a residue is mutated, its contacting partner will also be mutated in order to compensate for alterations of the raw evolutionary pressure between them [33]. Such a phenomenon is recorded in multiple sequence alignment (MSA) and can, historically, be observed through their paired MSA columns [36]. This underlies the statistical deduction of spatially closed acids, with high correlation values to be used as constraints for folding, for example, TM proteins [90,91]. To our knowledge, the research in which the DCA approach was first formulated and applied to achieve a marked improvement in the residue contact prediction should date back to 2009 [92], followed by an updated version in 2011 [93]. Since then, other similar methods, but with various refinements, have been proposed in close succession, e.g., PSICOV [94], GREMLIN [95], and plmDCA [96]. The progress observed for coevolution-based inference techniques (note that we specially refer to those methods based on the DCA approach), which has occurred within less than fifteen years, has greatly benefitted the current computation-assisted research in structural biology [97–99]. For instance, it is well-believed that the incorporation of coevolutionary information into features used for training deep learning models plays a predominant role in highly accurate protein structure prediction [26,27], and doubtless, improved TM protein modelling [100]. In addition to the success in deducing spatial proximities between residue pairs, the coevolution technique also allows the estimation of the likelihood of a single residue to be involved in a contact [101], thus starting to be applied in single-site-related prediction problems [38,102]. The MBPred work made the first attempt to verify its usefulness in TM protein interaction site prediction in which the above-threshold evolutionary coupling strengths of a residue are prioritized and aggregated to evaluate the potential coevolution with multiple residues in interaction interfaces or rare residues in non-interaction interfaces [38]. This has built a link between a residue's coevolutionary property and its structural/functional importance. Our recently developed TMKit tool (available shortly at https://github.com/2003100127/tmkit) provides the corresponding module for extracting the coevolutionary information of residue pairs or single residues.

2.6. Machine learning methodology

In addition to choosing a set of informative features to improve prediction performance, methodological improvement is commonly seen as another driving force in a number of biological fields that require computational modelling techniques [103–109]. In particular, the frontier of 3D structural prediction has recently witnessed a rapid transition from conventional machine learning to deep learning methods [110,111]. An immense amount of effort in applying deep learning has been applied to structure prediction due to its crucial importance in promoting the understanding of biological mechanisms. For example, all of the top best-performing tools in structure-related prediction tasks, as presented at recent CASP events, were developed by utilizing deep neural networks [112,113], such as the deep-learning-powered AlphaFold2 method, which has achieved stunning performance in predicting protein structures [26,114]. However, structure-related prediction tasks with regards to TM proteins were not given its own category at previous CASP events due to their scarcity. For example, there have been only 3 TM proteins (T1024, T1058 and T1098) identified using the TMHMM tool compared to roughly one hundred non-TM targets, as demonstrated at CASP14. Also, deep learning has been applied for addressing many other prediction problems specialized for TM proteins at a comparatively slow speed. However, in the past two years, deep learning

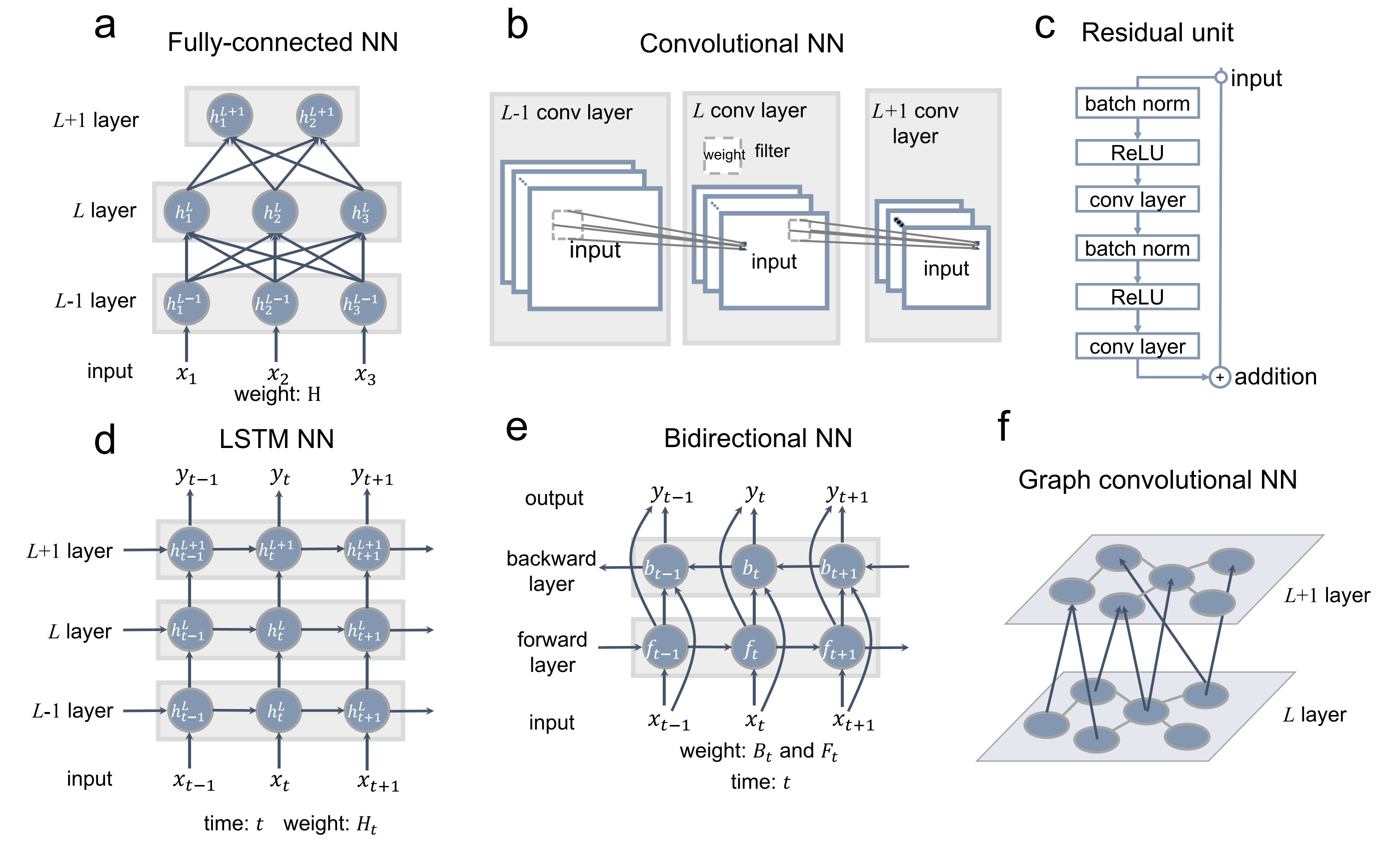

Fig. 3. Deep learning architectures. (a) Fully-connected neural networks. (b) Convolutional neural networks. (c) Residual units of residual neural networks. (d) Long short-term memory (LSTM) neural networks. (e) Bidirectional neural networks. (f) Graph convolutional neural networks. NN: neural network. Typically, a neural network consists of multiple neural layers, each set to extract features from input data. Blue circles represent neurons in neural layers. In a fully-connected neural network, neurons in one neural layer are needed to connect to every neuron in its adjacent neural layer. The edges between neurons are the weights required to be estimated during training. A convolutional neural layer is a layer in which multiple filters (i.e., weights required to be estimated) are placed to perform convolutional operations on input data that are image-like. A residual neural network consists of linearly connected residual units. But within each residual unit, every two convolutional layers are connected in a residual manner, where the final output of the residual unit is the sum of the output of the second convolutional layer and the raw input. This allows training a residual neural network at a much faster speed than training a convolutional neural network with the equal number of convolutional neural layers. The LSTM and bidirectional neural networks are two types of recurrent neural networks. In a recurrent neural layer, the input at the current time step is combined with the output from the previous time step(s). One main difference between the LSTM and bidirectional neural networks is that the former is trained in forward time directions while the latter is trained in both forward and backward time directions. Graph convolutional neural networks take as input the data that are represented by graphs and allow applying convolutional operations on the graph-based data to extract features. Taking the protein-ligand binding as an example, the structure of a ligand can be represented by a graph and the residues binding to the ligand at the binding pocket can also be conceived as being in a graph, which are together taken as input into a graph convolutional neural network.

methods have started to further enhance the TM protein topology prediction, showing a pronounced improvement [115,116].

As an important category in machine learning, deep learning methods have adopted several special neural network architectures, primarily including convolutional neural layers, recurrent neural layers, and graph convolutional neural layers (Fig. 3) [117,118]. A unique feature to distinguish deep learning methods from conventional machine learning methods is the automatic feature extraction ability from raw input data [117], which efficiently reduces the reliance on feature selection by domain experts [119]. Convolutional neural networks (CNNs) are the most commonly-used deep learning methods for a wide range of biological problems. A typical CNN architecture usually comprises a stack of multiple convolutional neural layers where the input data are required to be converted to image-like objects [120]. Compared with fully-connected networks with the size of parameters equal to the size of input data, a CNN allows for fast training through its parameter sharing setting [121] where the size of all kernels (i.e., parameters) in a layer is usually significantly smaller than the size of input data in this layer [122]. However, it remains computationally arduous to train CNNs with deep layers. Recent work has sought to address this with the development of the residual neural Network (ResNet), a variant version of CNNs, which has demonstrated the feasibility of training ultradeep neural networks [123]. The increased training speed lies in the use of a residual building block (ResBB) with a unique connection to allow the addition of the current output through the ResBB while enabling the previous input to bypass the ResBB, which regularly alternates between every several convolutional neural layers [124]. Given its transformative potential, we previously applied a 29-layer ResNet and a 38-layer ResNet for two-stage learning of interhelical residue contacts in TM proteins [39] and a 59-layer ResNet for learning interaction sites in TM proteins [102]. The ResNet approach has been widely applied to a wide range of modelling problems in structure biology. Moreover, graph neural networks (GNNs) have emerged as a useful tool to learn the characterization of geometrically structured data that can be represented by graphs [125]. The connection strengths between edges in the graphs are usually metrics for evaluating the similarity between nodes [126]. Therefore, the biological application of the GNN method is more often seen in the inference of associations between molecules, e.g., drugs [127], but rarely seen in membrane protein studies. More recently, deep learning-based structural biology fields have started to benefit from using transfer learning to improve prediction accuracy. The idea is that language models are first trained on vast amounts of protein sequences to comprehend and learn their biological features (also called protein embeddings) usually in an unsupervised manner and, based on the learned information, the pretrained models are then transferred to perform other prediction tasks [31]. Current language models that are able to complete the above tasks at high accuracy are transformers [128], which are mainly constructed using attention-based modules that enhance the extraction ability of informative features [129]. For example, a recent study has shown the improved performance in membrane protein topology prediction, which is achieved by using the ProtT5 transformer model [116].

2.7. Human transmembrane protein families

Using the GtoPdb database (https://www.guidetopharmacology.org), human transmembrane proteins are categorized into eight functional groups: *transporter*, *G-protein-coupled receptor* (GPCR), *enzyme*, *catalytic receptor*, *ligand-gated ion channel* (LGIC), *voltage-gated ion channel* (VGIC), *other ion channel* (other IC), and *other protein*, which we used to analyse disease-causing mutations in Sections 6.2 and 6.3.

3. Membrane protein type classification

There are eight types of membrane proteins documented in the literature, comprising type I, type II, type III, type IV, multi-pass, lipid-chain-anchored, GPI-anchored, and peripheral proteins [130]. Among them, type I, type II, type III, type IV are single-pass proteins and together with multi-pass proteins they are categorized as TM proteins. The remaining types are surface-bound membrane proteins. From a functional perspective, one difference between these two groups is that TM proteins can function in cytoplasmic, extracellular, and/or membrane regions, whereas surface-bound proteins mainly function on one side of the lipid bilayer. To be able to readily determine functional groups from the vast array of membrane proteins, for which only protein sequences are available, computational techniques are crucial. Apart from two classical algorithms, ProtLock [131] and MemType-2 L [130], methods published in the past decade have been summarized in Table 2. Note that several methods have not been made available for use but still provide information about the development of methodologies over time. These techniques primarily phase this classification problem by detecting membrane proteins first and, then, applying an eight-label classification procedure for the eight protein types [132–134].

As shown in Table 2**,** the salient feature that almost all methods have in common is the incorporation of the pseudo position-specific scoring matrix (PSSM) [130] and/or the pseudo amino acid composition (AAC) [135–137] into the classification of membrane protein types. Most machine learning methods require a fixed-length vector as input. However, if features of individual residues are used, proteins of different lengths are described as feature vectors of different lengths, which is due to the different number of residues within the proteins. In comparison, the PSSM- and AAC-based features address this by encoding proteins of different lengths as fixed-length feature vectors. This is advantageous when training classifiers that take as input the protein sequences of different lengths. Importantly, most classifiers were developed using conventional machine learning algorithms. Therefore, it would be interesting for future studies to explore the utility of deep learning in improving current classification techniques.

4. Transmembrane protein topology prediction

Annotations of membrane protein topology can be used to assist structure-function studies [145–147] but experimentally determining these has long been a mammoth task [148]. Therefore, the development of computational identification approaches has gained great popularity in the past few decades [149]. Topology prediction enables differentiation of TM proteins but detailed protein types (e.g., type I or type III) cannot be ascertained. Nevertheless, it can determine other biologically relevant information including localization, number of protein segments relative to the membrane in cytoplasmic/intracellular, transmembrane and extracellular regions. Moreover, predicted topology profiles have been suggested to refine the computational design of membrane proteins [150–154].

With substantial recent progress, there are now extensive webservers and standalone packages available across an assortment of computational techniques (Table 3). In particular, the past two years have seen an increase in the number of predictors developed by deep learning, including Li's work [155], Membrain3.0 [156], DeepTMHMM [115], DeepTMpred [157], and TMbed [116]. It is worth noting that many of these were developed by transfer learning approaches where transformer-based pretrained models were transferred from nature language processing (NLP) modelling [30,128,158] to topology prediction. For example, both DeepTMHMM and DeepTMpred adopted the pretrained ESM model [159], while TMbed took advantage of the ProtT5 model pretrained on a number of autoregressive and auto-encoder models [31]. These three methods represent the current state-of-the-art in this field, with TMbed having demonstrated the lowest false positive values. Nevertheless, the two canonical methods TMHMM2.0 [160] and PolyPhobius [161] developed early in the 2000 s remain programs-of-choice for topology prediction, which were widely adopted by topology-assisted studies [39,102] and public databases such as UniProt [162,163]. Current available computational approaches are generally used for determining the topology of α-helical TM proteins since the vast majority of membrane proteins are α-helix-bundled and are criss-crossed through membranes. According to PDBTM (version: 2022–08–26), α-helical TM proteins account for 92.6% of TM proteins. More recently, a few methods have begun to extend to predict the topology of β-barrel TM proteins in addition to the α-helical topology prediction. Example methods are HMM-TMv2-HNN [164], DeepTMHMM [115], and TMbed [116]. The integration of available topology information is helpful in performing multi-factor analyses; for example, it has been used to understand the biological environment in which disease-causing mutations or interaction sites are located.

5. Membrane protein interactions

Biophysical interactions between proteins are essential for a wide variety of biological processes. However, a clear understanding of the interactions at an intermolecular level has been encumbered due to high costs and the time-consuming nature of protein structure determination techniques. It is particularly difficult regarding membrane proteins as many of them are large and lipid-anchored assemblies. Computational techniques are therefore much-needed for the annotation of functionally important sites in membrane proteins [179].

5.1. Experimentally resolved binary PPI interactome maps

PPIs determined experimentally by high-throughput techniques are binary, pairing, and structure-free to allow for the analysis of the interconnections between disease-related proteins in the pathological cellular context [180,181]. Typical methods for conducting these high-throughput experiments include yeast two-hybrid (Y2H) assays and affinity purification mass spectrometry (AP-MS). The Y2H assay identifies PPIs through the activation of a transcription factor (TF) in living yeast cells once a protein of interest residing in the TF DNA-binding domain comes in physical contact with a bait protein in the activation domain [182,183]. Alternatively, AP-MS involves purifying protein complexes formed by a tag-fused bait protein (e.g., antibody) and its interaction partners, which then undergo mass spectrometry for refined characterization [184–186]. Compared to

Table 2
Type classification tools.

Method	Source	Year	Algorithm	Category	Feature	Subclass	Citation
ProtLock	-	1997	Distance-based algorithms	Statistics	ACC	Yes	[131]
MemType-2 L	http://www.csbio.sjtu.edu.cn/bioinf/MemType	2007	Ensemble of the optimized evidence-theoretic KNN algorithms	ML	Evolution, Pse-PSSM	Yes	[130]
Ali's work	-	2015	Multiple-stage training with KNN, PNN, SVM, Naïve bayes multinomial, and voting feature interval	ML	Pse-AAC	Yes	[138]
Butt's work	-	2016	Multilayer neural networks	ML	Feature extraction with statistical moments	No	[134]
MBPpred	http://bioinformatics.biol.uoa.gr/MBPpred	2016	Profile hidden Markov models	ML	Membrane binding domains	No	[139]
Guo's work1	-	2017	Stacked generalization (ensemble) of SVM, KNN, RFs, neural networks, multiple logistic regression	ML	Pse-AAC	Yes	[140]
iMem-2LSAAC	-	2018	Multiple-stage training with the KNN, PNN, SVM, generalize regression neural network, and random forest algorithms	ML	Pse-AAC	Yes	[132]
MKSVM-HSIC	https://github.com/hzwh6910/Identification-of-Membrane-Protein-Types-via-Multivariate-Information-Fusion-with-Hilbert-Schmidt	2019	Multiple kernel SVM	ML	Pse-PSSM	Yes	[141]
Guo's work2	https://github.com/DragonKnightss/MembraneProteinTypePrediction	2019	Convolutional and bidirectional long short term-memory neural networks	DL	Pse-PSSM and other PSSM-based features	Yes	[142]
TooT-M	https://github.com/bioinformatics-group/TooT-M	2020	Selective voting ensemble classifiers	ML	Pse-PSSM, Pse-AAC	No	[133]
Zhang's work	-	2021	SVM, RF, simple logistic, Naive bayes, nearest neighbors, and decision trees	ML	PseAAC	Yes	[143]
iMPT-FDNPL	https://github.com/mufei111/iMPT-FDNPL	2021	word2vector, random k-labelsets ensemble (RAkEL), and RFs	ML	Sequence	Yes	[144]

Note: SVM, support vector machine; KNN, K-nearest neighbour; RF, random forest; PNN, probabilistic neural network; Pse-PSSM, pseudo position-specific scoring matrix; Pse-AAC, pseudo amino acid composition. The subclass column represents whether the listed programs can be used to classify the subclasses of membrane proteins, including type I, type II, type III, type IV, multi-pass, lipid-chain-anchored, GPI-anchored, and/or peripheral proteins. If no, the program(s) can only be used to distinguish between membrane and globular proteins.

Table 3
Topology prediction tools.

Method	Source	Year	Algorithm	Category	Signal peptide	Note	Reference
TMHMM2.0	https://services.healthtech.dtu.dk/service.php?TMHMM-2.0	2001	Hidden Markov models	ML	No	-	[160]
HMMTOP	http://enzim.hu/hmmtop	2001	Hidden Markov models	ML	No	-	[165]
PolyPhobius	https://phobius.sbc.su.se/poly.html	2004	Hidden Markov models	ML	Yes	-	[161,166,167]
Philius	http://www.yeastrc.org/philius	2008	Dynamic Bayesian Networks	ML	Yes	-	[168]
OCTOPUS	http://octopus.cbr.su.se	2008	Hidden Markov models and artificial neural networks	ML	No	Interface, loop	[169]
SPOCTOPUS	http://octopus.cbr.su.se	2008	Hidden Markov models and neural networks	ML	Yes	-	[170]
CCTOP	http://cctop.ttk.hu	2015	Ensemble of existing multiple predictors	Consensus	Yes	-	[171,172]
TOPCONS	https://github.com/ElofssonLab/TOPCONS2	2015	Ensemble of existing multiple predictors	Consensus	Yes	-	[173]
SCAMPI2	https://scampi.cbr.su.se	2015	Parameter fine-tuning	Statistics	No	-	[174,175]
TMSEG	https://github.com/Rostlab/TMSEG	2016	Random forests	ML	Yes	-	[176]
NNME	http://genomics.fzu.edu.cn/nnme/index.html	2017	Neural networks and random forests	ML	No	-	[177]
Li's work	-	2019	Faster-RCNN	DL	No	-	[155]
Membrain3.0	http://www.csbio.sjtu.edu.cn/bioinf/MemBrain	2020	Convolutional neural networks	DL	Yes	-	[156,178]
HMM-TMv2-HNN	http://www.compgen.org/tools/PRED-TMBB2	2021	Hidden Markov models and neural networks	ML	No	Beta barrel	[164]
DeepTMHMM	https://dtu.biolib.com/DeepTMHMM	2022	Deep learning protein language modelling	DL	Yes	Beta barrel	[115]
DeepTMpred	https://github.com/ISYSLAB-HUST/DeepTMpred	2022	Deep transfer learning	DL	No	-	[157]
TMbed	https://github.com/BernhoferM/TMbed	2022	ProtT5 embeddings	DL	Yes	Beta barrel	[116]

AP-MS, the Y2H assay is more affordable and inexpensive as it averts high resource consumption such as that required for protein purification (their advantages and disadvantages are discussed in [187,188]). It is noted that the classical Y2H system is not suitable for determining those interactions involving TM proteins since it requires interacting proteins to be present in the nucleus but TM proteins cannot fold without the lipid bilayer [189]. The human binary PPI interactome map maintained by the Center for Cancer Systems Biology (CCSB) is mainly based on results of Y2H assays, with the first version possibly backdated to 2005 [190,191]. In the most recent version, HuRI, this quantity has grown to > 60,000 binary PPIs that are enriched for TM protein interactions [181]. It is estimated that human PPIs in HuRI that involve at least one TM protein make up around 40% of the whole human interactome [102].

5.2. *Experimentally resolved interaction sites*

Interaction sites residing at protein interfaces are key to understanding molecular mechanisms and cellular functions of a protein [192]. Details of interactions between inter-protein residues in protein complexes can be solved experimentally by X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryogenic electron microscopy (cryo-EM) [193,194]. X-ray crystallography is currently the most prevailing method for structure determination [195]. Comparatively, high sensitivity to the nuances of local structural changes makes NMR spectroscopy powerful in determining protein-ligand binding, showing promise for drug discovery [196]. The biggest advantage of using Cryo-EM over both methods is that Cryo-EM is more suited for determining membrane protein complexes that are large [16,193,197,198], e.g., the ACE2 protein in complex with the spike glycoprotein of SARS-CoV-2 [199].

5.3. *Prediction of TM protein interaction sites*

5.3.1. *Definition of interaction sites in membrane proteins*

Cocooned by 3D protein structures consisting of multiple subunits (i.e., chains), interaction interfaces at the inter-protein level are characterized by tightly packed regions. Interaction sites residing in the interaction interfaces are spatially close residues that can be derived computationally from the structures by setting a distance-based cut-off (e.g., below 5.5 Å or 6 Å). To further restrict the number of interaction sites identified and, thus, filter potential false positives, the relative solvent accessibility value of a residue in an interaction in the unbound state can be restricted to fluctuate within a predefined range (e.g., at least above 0.2). Subsequently, interaction and noninteraction sites can be partitioned and taken as ground truth for model training. It is noted that since TM proteins are often co-crystallized with antibodies in order to increase their solubility in the aqueous phase, the light and/or heavy chains need to be removed before the interaction sites are calculated [102,200].

5.3.2. *Biological assembly*

The structure of a protein deposited in a PDB file downloaded from https://www.rcsb.org/ is only a representation of 3D coordinates, which provides the structure as an asymmetric unit [42]. The actual biological unit of the protein (i.e., the biological assembly or biomolecule), which functions in oligomer form in the cellular context, is required for functional studies (https://pdbj.org/help/about-aubu) [201], for example, protein-protein interactions. Such a structure can be generated by applying symmetry operations (e.g., axes of rotations, translations, or a combination of both) to the asymmetric unit (https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies) [41]. Symmetry operations necessary for this type of transformation can be accessed via the BIOMOLECULE records aligned to REMARK 350 in the PDB file [43]. However, biological assemblies of TM proteins generated by

Table 4
Interaction site prediction tools.

Method	Source	Year	Algorithm	Category	Dataset	Note	Citation
Bordner's work	-	2009	Random forests	ML	α-helical and β-barrel membrane proteins	-	[44]
MBPred	https://github.com/bojigu/MBPred	2019	Random forests	ML	α-helical transmembrane proteins	full-length sequence, membrane, extracellular, and cytoplasmic predictions	[38]
DeepTMInter	https://github.com/2003100127/deeptminter	2021	Ultra-deep residual neural networks with the stacked generalization ensemble technique	DL	α-helical transmembrane proteins	full-length sequence, membrane, extracellular, and cytoplasmic predictions	[102]

directly applying these operations may be inaccurate because the BIOMOLECULE records generated by the PISA algorithm are specialized for globular proteins. To address this, the TMDET algorithm has developed a refined calculation process specialized for TM proteins based on the BIOMOLECULE records. In Bordner's work, MBPred, and DeepTMInter, all interaction sites of a protein chain are scanned and detected through their biological assemblies generated using TMDET or through manual calibration.

5.3.3. Quality control of datasets

A high-quality dataset is instrumental for generating high-performance machine learning models. Prediction of structure-derived interaction sites involves a complex data pre-processing procedure as the data used for training cannot be directly fetched from databases supplying functional site annotations. Generating such a dataset normally starts from collecting an initial database of protein structures and requires several preliminary procedures. Firstly, only well-resolved protein structures are retained by applying a few experimental determination parameters (e.g., a 3.5 resolution and the X-ray crystallography method). Secondly, the biological assemblies of the retained protein structures are generated according to the procedures mentioned in the above section. Thirdly, within each biological assembly, calculating all interacting chains with at least one interaction site is performed to ensure that computing resources necessary in other quality control steps will not be invested in protein chains that have never been used. Further data quality control processes may vary depending on downstream analyses but involve several required steps, including redundancy removal by setting sequence-identity or structure-similarity thresholds. Once these various steps are completed, interaction sites of interacting chains can then be calculated.

5.3.4. Learning algorithms

Prediction of interaction sites specific to membrane proteins is currently in its infancy and there are only a handful of methods available for use (Table 4). Bordner made the first attempt in 2009 to predict membrane protein interaction sites with a random forest algorithm based on a combination of multimeric membrane proteins. Using almost the same tree-based approach, the MBPred method developed in 2019 improved the prediction performance by comparing itself with a reimplementation of the Bordner work. The utility of introducing co-evolutionary features to interaction site prediction was first demonstrated in the MBPred work. Thereafter, a follow-up study, DeepTMInter, continued the adoption of the co-evolution strategy but offered a reinforced version by adding an additional co-evolutionary feature and learning the co-evolution representation of a residue in sequence context in which a sliding window was applied. The ResNet method has been well tried in protein structure-related prediction fields [27,202,203]. DeepTMInter is a deep learning implementation, which utilizes a hand-picked ResNet architecture prevalent in image recognition, thereby boosting its performance. The ResNet model used for building DeepTMInter has a 59-layer architecture and is assembled by following conventional practice where residual units alternate between every two max-pooling operations, similar to that used in [123]. DeepTMInter had a pronounced improvement in its performance in predicting interaction sites in full-length sequences, TM, extracellular, and intracellular regions, which might be attributed to the methodological approach, the increased size of training samples, and/or the features used. In Fig. 4 we provide the interaction sites of two example TM proteins (PDB id: 6T0B chain m and PDB id: 5B0W chain A) predicted by two TM protein-specific predictors, DeepTMInter and MBPred, and two soluble protein-specific predictors, DELPHI and GraphPPIS. As TM proteins are large biomolecular assemblies consisting of multiple subunits, the predicted interaction sites between subunits are extremely useful for the oligomerization of their quaternary structures [204–207]. Given that deep learning

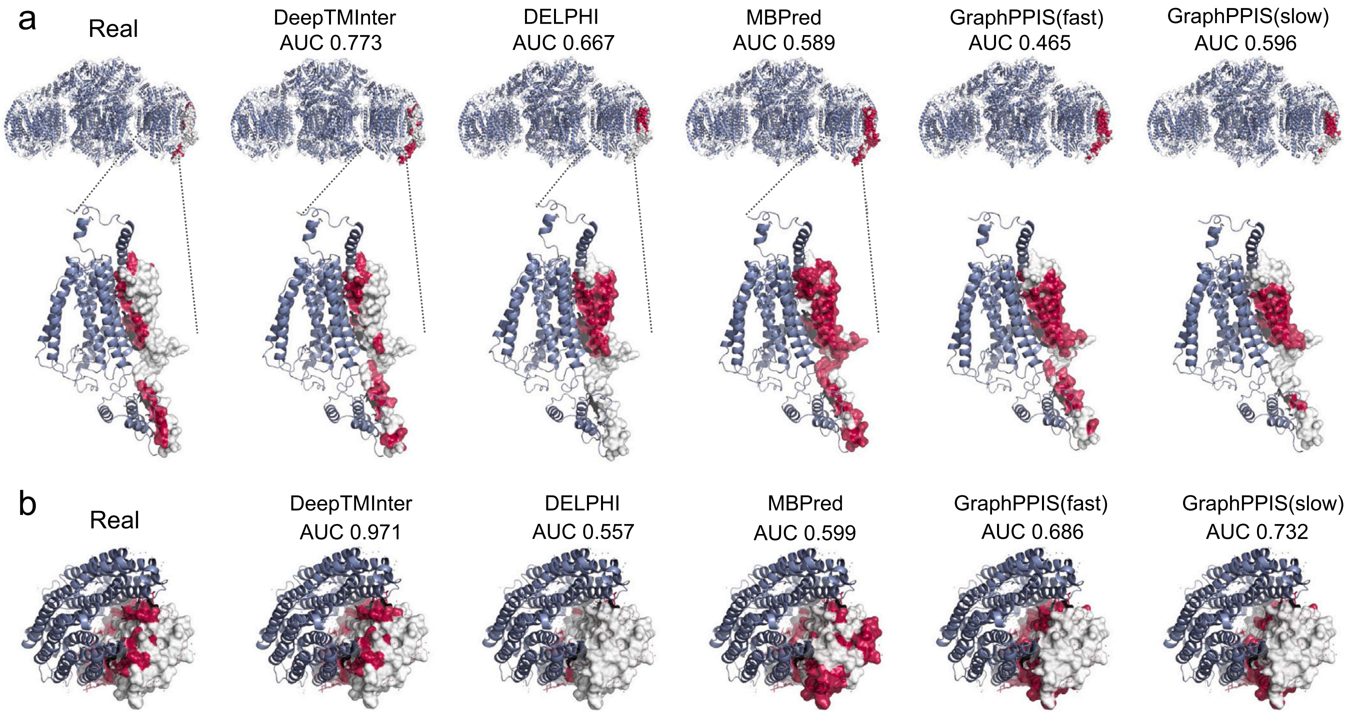

Fig. 4. Prediction of interaction sites of two example TM proteins, (a) the 11-cis isomer of pharaonis halorhodopsin (PDB id: 5B0W chain A) and (b) the hypoxic isoforms of mature mitochondrial III-IV supercomplexes (PDB id: 6T0B chain m). DeepTMInter and MBPred are two methods specialized for TM proteins, while DELPHI and GraphPPIS are two methods specialized for globular proteins. Using the GraphPPIS webserver, the prediction results can be generated in two modes. The fast mode allows predicting interaction sites based on the BLOSUM62 matrix and the DSSP secondary structures of an input amino acid sequence while the slow mode allows predicting interaction sites based on the position-specific scoring matrix, the hidden Markov model profile, and the DSSP secondary structures of the input amino acid sequence. The two protein chains 5B0WA and 6T0Bm are presented in surface form while their interacting protein chains are presented in cartoon form. In more detail, non-interaction sites in 5B0WA and 6T0Bm are shown using white surfaces while their interaction sites predicted by these tools are highlighted using red surfaces.

continues to make sweeping changes to the landscape of structural prediction fields, the future development of protein interaction site prediction should take advantage of deep learning models that will likely be more powerful in representation learning and provide improved feature extraction from the intrinsic surroundings of the protein sequence and the structural elements of interaction sites.

6. Mutation effect prediction

6.1. Membrane protein mutation databases and mutation-effect predictors

In the absence of databases comprised of experimentally-derived data, the prediction of mutation/variant effects becomes key to filling the gap between millions of mutations in proteins and their impact on human diseases (i.e., pathogenicity) [208–212]. Given that disease-causing/pathogenic mutations occur more frequently in membrane proteins than in other types of proteins [45,213], computational strategies are needed to accelerate the development of prediction tools that are able to accurately evaluate the effects of pathogenic mutations in the former [214,215]. There have been a few tools developed, and presently available for this purpose, including Pred-MutHTP[215], BorodaTM [216], mCSM-membrane[217], and MutTMPredictor [218], but this ultimately far less than the number developed for globular proteins [209,210,219–221]. Indeed, various deep learning strategies have been developed to predict the functional impact of mutations occurring in globular proteins [222–225] but have yet to be applied for membrane proteins. Most of the tools specialized for membrane proteins are made by tree-based machine learning strategies (e.g., random forest). It is unknown as to whether

Table 5
Mutation effect databases and prediction tools.

Database					
Name	Source	Year			Citation
MutHTP	http://www.iitm.ac.in/bioinfo/MutHTP	2018			[48]
TMSNP	http://lmc.uab.es/tmsnp	2021			[213]
Predictor					
Method	**Source**	**Year**	**Algorithm**	**Category**	**Citation**
Pred-MutHTP	https://www.iitm.ac.in/bioinfo/PredMutHTP	2019	Voting with neural networks, naïve bayes, tree decision methods, random forests, and/or logistic regression	ML	[215]
BorodaTM	https://www.iitm.ac.in/bioinfo/MutHTP/boroda.php	2019	XGBoost	ML	[216]
mCSM-membrane	http://biosig.unimelb.edu.au/mcsm_membrane	2020	Random forests and regression	ML	[217]
MutTMPredictor	http://csbio.njust.edu.cn/bioinf/muttmpredictor	2021	Cascade XGBoost	ML	[218]
TMSNP-predictor	https://github.com/adriangarciarecio/TMSNP	2021	Random forests, XGBoost, and SVM	ML	[213]

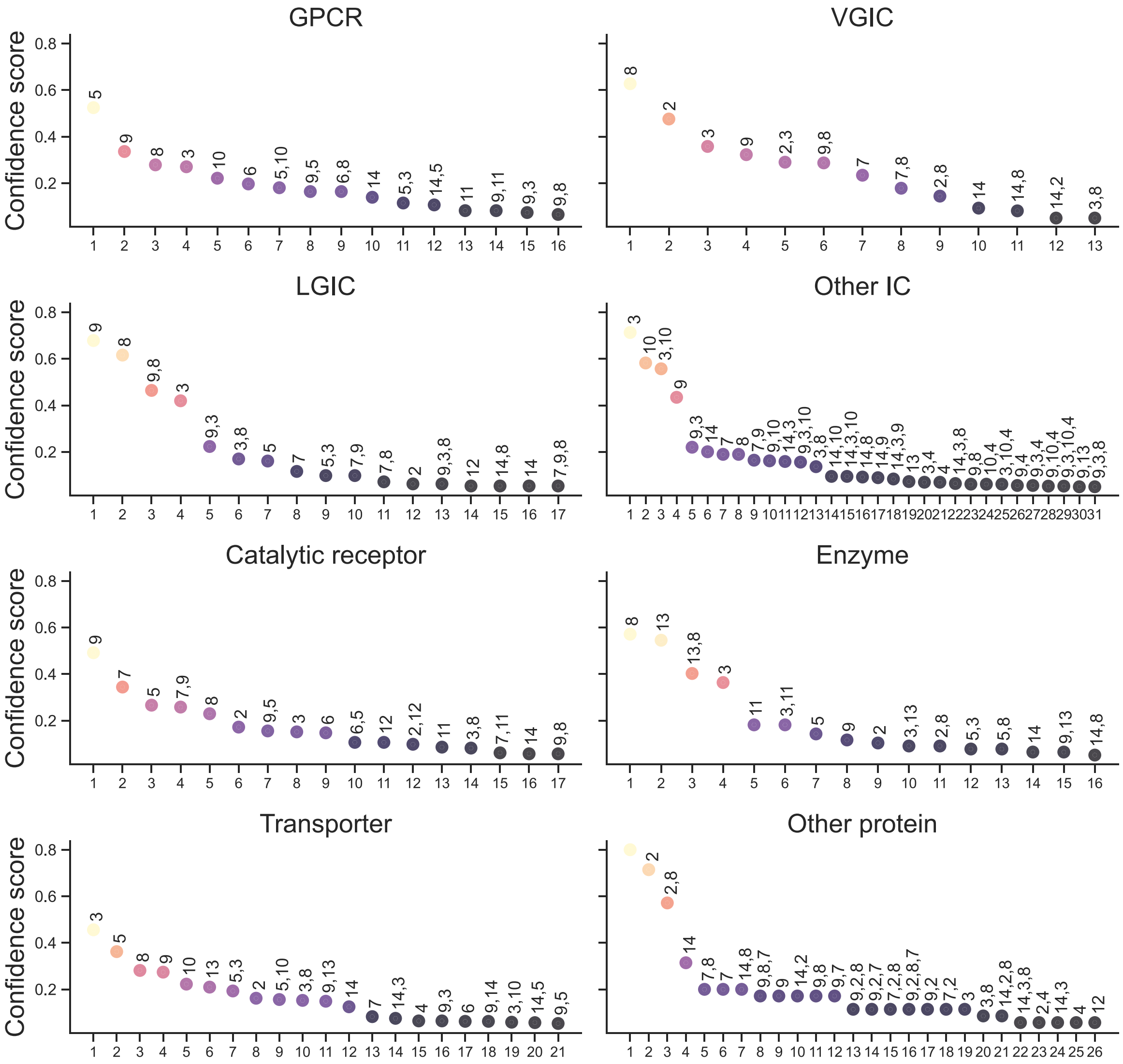

Fig. 5. Disease patterns in pathogenic mutations across eight human transmembrane protein families: *transporter*, *G-protein-coupled receptor* (GPCR), *enzyme*, *catalytic receptor*, *ligand-gated ion channel* (LGIC), *voltage-gated ion channel* (VGIC), *other ion channel* (other IC), and *other protein*. Proteins are annotated as one of the above protein families based on the GtoPdb databases (see Section 2.7). The number on the plots represents one of the following disease types appearing in the MutHTP databases: 1: cancers, 2: cardiovascular diseases, 3: congenital disorders of metabolism, 4: digestive system diseases, 5: endocrine and metabolic diseases, 6: immune system diseases, 7: musculoskeletal diseases, 8: nervous system diseases, 9: other congenital disorders, 10: urinary system diseases, 11: reproductive system diseases, 12: respiratory diseases, 13: skin diseases, and 14: other types. The confidence score is produced by the Apriori algorithm (see Section 6.2) and is used to show how confidently a disease type occurs or multiple disease types co-occur in a protein family when mutations are pathogenic.

deep learning is conducive to a better prediction ability. Nevertheless, with the available methods, it is possible to gain knowledge about the pathogenicity of amino acid substitutions in many membrane proteins with mutations highly related to a variety of human diseases. For example, twelve tools were systematically benchmarked for evaluating the pathogenic impact of the variants of a voltage-gated sodium channel, hNav1.5 [226], and determined more than 70 potential pathogenic variants in both TM and extracellular regions. In another study, several tools (e.g., Rhapsody and EVmutation) were employed to determine deleterious mutations in the renal outer medullary potassium channel [227]. Crucially, training these tools with high prediction capacity depends heavily upon the availability of high-confidence experimental data. Two presently available databases, MutHTP [48] and TMSNP [213], provide an extensive repertoire of experimentally verified membrane protein disease-causing and neutral mutations (Table 5), which opens up the possibility of developing different kinds of predictors. The MutHTP database is an exhaustive collection of pathogenic and neutral missense, insertion, and deletion mutations specific to human TM proteins [48]. These mutations were derived from the Humsavar, SwissVar, 1000 Genomes, COSMIC and ClinVar databases. TMSNP is a more recently developed database maintaining pathogenic and neutral missense mutations with information on structural and environment features, which enables their differentiation from globular proteins. Compared to MutHTP, the mutations in TMSNP appear only in TM regions.

Based on these two databases, a variety of prediction tools have been developed to assist in determining pathogenicity for various missense mutations (Table 5). Pred-MutHTP was the first work to use MutHTP to predict variant effects of missense mutation sites. To ensure high quality, all methods available in WEKA [228], such as the Bayesian network, the logistic regression model, the multilayer perceptron and the random forest were extracted to train their individual models, and perform membrane topology-wise evaluation. It then prioritized the predictions of the voting algorithm for use. Compared to Pred-MutHTP, a subsequent effort, MutTMPredictor,

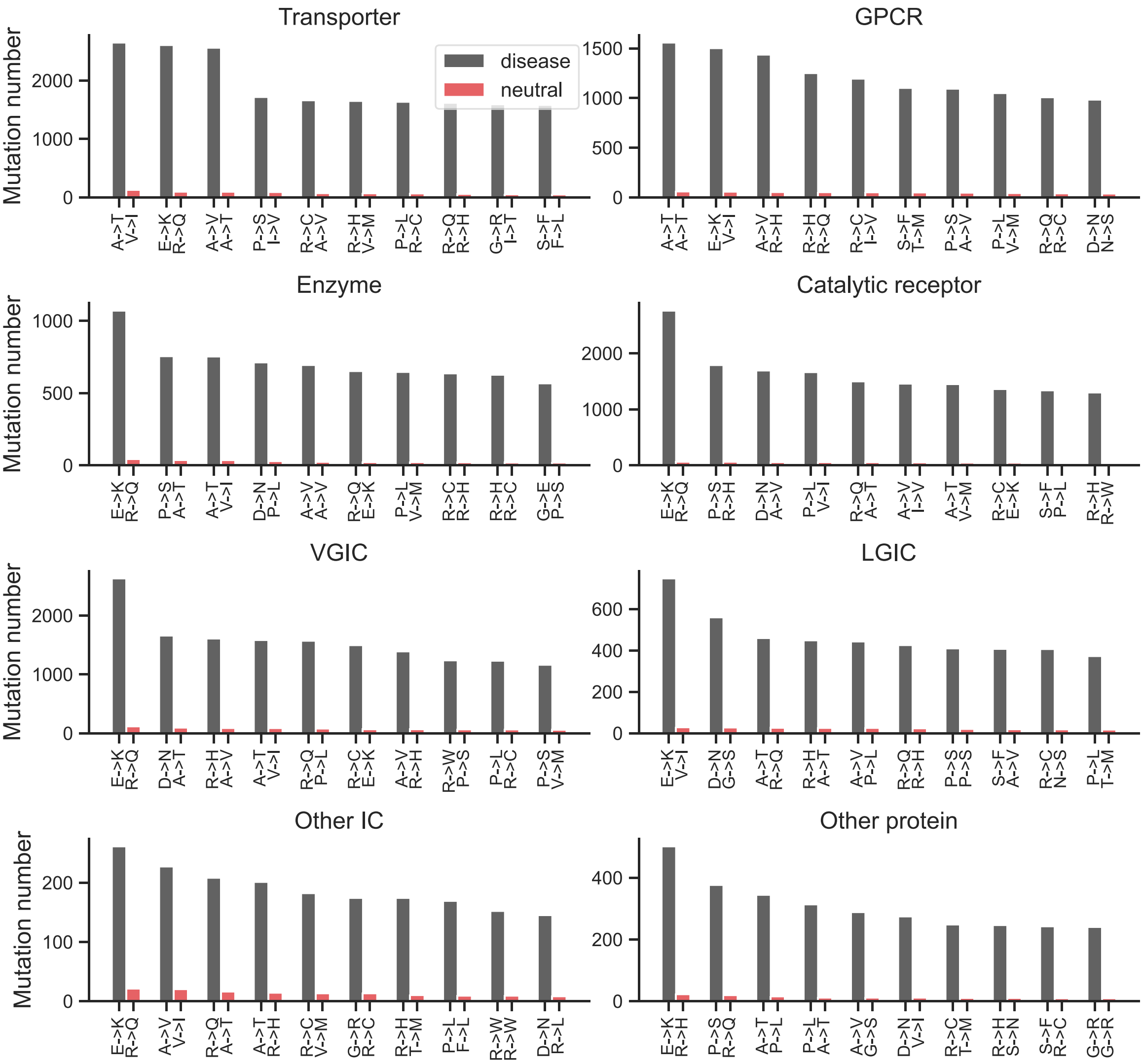

Fig. 6. Top-ranked wild-to-variant types across eight human transmembrane protein families: *transporter*, *G-protein-coupled receptor* (GPCR), *enzyme*, *catalytic receptor*, *ligand-gated ion channel* (LGIC), *voltage-gated ion channel* (VGIC), *other ion channel* (other IC), and *other protein*. E- > K, for example, represents a mutation from amino acid E (wild-type) to amino acid K (variant-type) in TM proteins. Black bars represent disease-causing/pathogenic mutations while red bars represent neutral mutations.

demonstrated improved AUC values in mutations located in the extracellular regions [218]. The utility of considering evolutionary information as features was demonstrated in these two methods. Additionally, an in-house predictor was constructed alongside the simultaneous establishment of the TMSNP database, which was trained by distilling pathogenic mutations from 358 proteins and non-pathogenic mutations from 2420 proteins. It demonstrated higher accuracy, specificity, and MCC performance but lower sensitivity and coverage performance than Pred-MutHTP [213]. Built without relying on the aforementioned two databases, two other predictors, BorodaTM [216] and mCSM-membrane [217] also employed tree-based methods to quantify the mutation effects for which protein 3D structures were available. BorodaTM demonstrated high performance in mutation effect prediction specific to TM regions using known protein structures. In order to characterize the structural signature of a mutation residue, mCSM-membrane exploited a graph-based strategy to assign a graph containing the residue of focus and their adjacent residues with features inclusive of pharmacophores and interactions. mCSM-membrane performed significantly better than a number of tools developed for globular soluble proteins.

6.2. Family-specific disease patterns in MutHTP

To better understand disease patterns resulting from pathogenic mutations specific to human transmembrane protein families, we applied the Apriori algorithm [229,230] to perform an association rule learning analysis based on the latest version of the MutHTP database, leading to a preferred order of diseases or disease combinations. Considering the overrepresentation of cancer-type pathogenic mutations in MutHTP, we removed this type when mining the disease patterns from the database. As can be seen in Fig. 5, single diseases appear in each family more frequently than a combination of multiple diseases. Using a confidence score threshold above 0.5, we identified the two most frequently co-occurring disease combinations in the *other IC* and *other protein* families, respectively. As scanned by the Apriori algorithm, disease-causing mutations in *the other protein families* can bring about the comorbidity of cardiovascular diseases and nervous system diseases, as pointed out in [231,232]. In terms of the occurrence of only single diseases, the four most frequently occurring types of diseases are congenital disorders of metabolism, endocrine and metabolic diseases, nervous system diseases, and other congenital disorders,

Table 6
Meta-property prediction tools.

Method	Source	Year	Algorithm	Category	Scope	Note	Citation
AllesTM	https://github.com/phngs/allestm	2020	Ensemble of gradient boosting, random forests, convolutional neural networks, and bidirectional long short-term memory (LSTM) neural networks	ML and DL	Z-coordinates, flexibility, topology, relative solvent accessibility, torsion angles, secondary structures	The only specialized tool for predicting z-coordinates, torsion angles, flexibility	[242]
MASSP	http://www.meilerlab.org/index.php/servers/show?s_id=26; https://github.com/computbiolgeek/massp	2021	Convolutional neural networks and LSTM neural networks	DL	Topology, secondary structures	-	[243]
TopProperty	https://cpclab.uni-duesseldorf.de/topsuite	2021	Residual neural networks	DL	Topology, secondary structures, membrane exposure, solvent accessibility	-	[244]

which top the ranking lists of the majority of the 8 protein families. Most of the relatively high-confidence scores from Apriori to support the credibility of the discovered disease patterns are seen in the three ion channel families, typically between 0.6 and 0.8. Diseases (e.g., hyperinsulinemic hypoglycemia of infancy [233]) caused by disorders of metabolism or nervous systems has been previously documented [234–238].

6.3. Family-specific amino acid substitutions in MutHTP

A missense mutation can result in 19 types of substitutions at a single amino acid position. The different types of amino acid substitutions that are pathogenic or benign may help to better understand differences between diseases and healthy systems across human transmembrane protein families. To explore this, we investigated the substitution types of all pathogenic or benign mutations from wild to variant types in the MutHTP database. As shown in Fig. 6, 'A- > T′ and 'E- > K′ are the most common types of substitutions as a result of disease-causing mutations, with the former one detected in the transporter and GPCR families and the latter detected in the rest of the protein families. The 'E- > K′ type is also confirmed as the most frequently seen substitutions in TM regions as reported in [214]. By contrast, the 'V- > I′ and 'R- > Q′ types are most frequently observed as a result of benign mutations. Note that several most common types of substitutions as a result of benign mutations, such as 'V- > I′, are more likely to be pathogenic. In addition, the 'A- > V′ and 'D- > N′ types may be worth attention as they appear as similarly frequent as the above listed types.

7. Prediction of membrane protein stability

Kulandaisamy et al. have developed a multiple linear regression-based machine learning method to predict the thermostability of membrane proteins upon the occurrence of missense mutations [239]. First, a non-redundant dataset of 929 mutations in relation to experimental thermostability was constructed based on the MPTherm database [240]. Next, these mutations were grouped according to membrane-spanning or aqueous regions and further classified by their functions, secondary structures, and the locations of mutations in protein structures, respectively. For each mutation, various sequence- and structure-based features such as conservation scores, physiochemical properties, neighboring residues, contact potentials, atomic contacts, residue depths, and topologies were then computed to predict thermostability. As reported, after 10-fold cross validation, MPTherm-pred achieved a correlation of 0.72 with a mean absolute error of 2.85 °C for mutations located in membrane-spanning regions and a correlation of 0.73 with a mean absolute error of 3.72 °C for those located in aqueous regions. The method is available as a web server at https://web.iitm.ac.in/bioinfo2/mpthermpred/.

Other than predicting pathogenicity (described in Section 6.1), mCSM-membrane can also be used to predict membrane protein stability upon the occurrence of missense mutations. Using 223 mutations from 7 proteins of known structures [241], the difference in the Gibbs free energy of folding [217] between wild and mutant types was calculated to be indicative of protein stability. This study showed a correlation of 0.72 between experimental and predicted values.

8. Prediction of multiple properties with metamethods

Rather than focusing on single property prediction, a few studies have sought to predict a number of properties in combination, such as solvent accessibility, secondary structures, and torsion angles. These methods include AllesTM [242], MASSP [243], and TopProperty [244], which all use deep learning methods to keep abreast

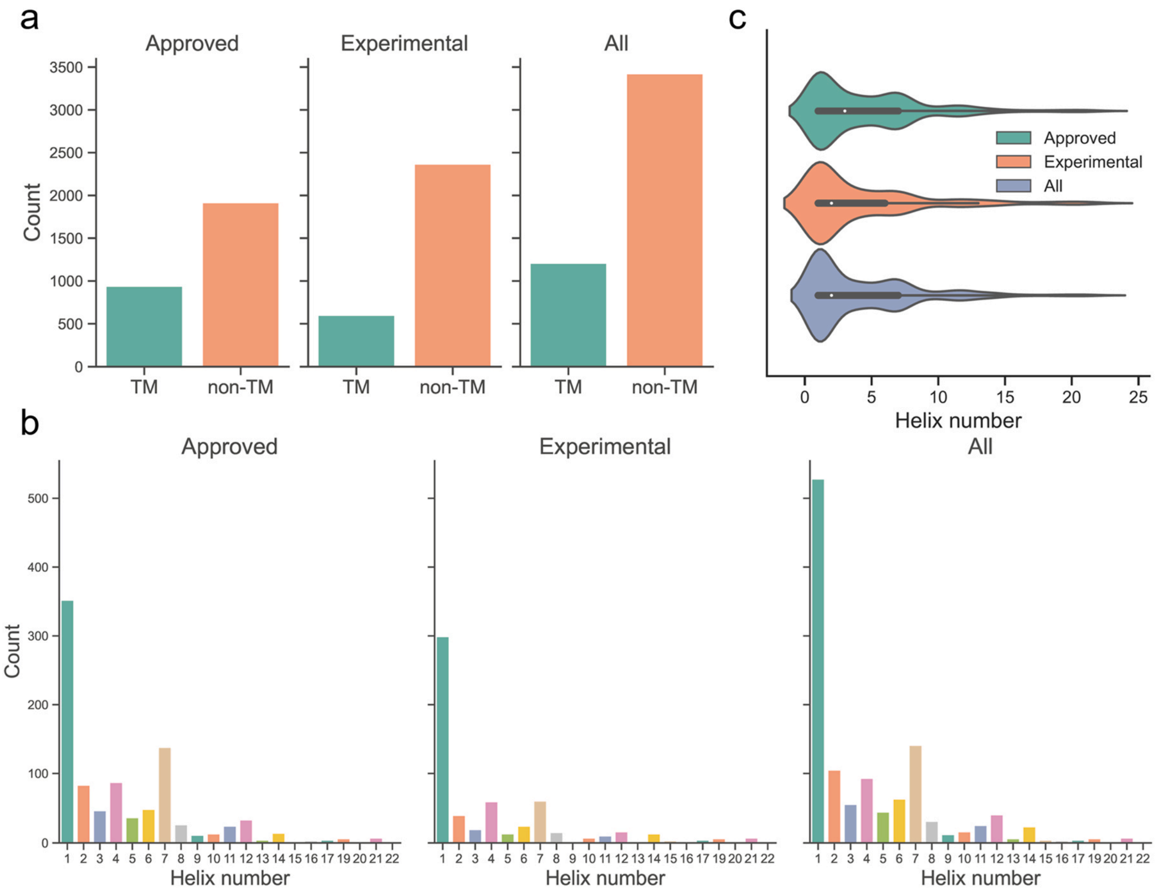

Fig. 7. Distributions of transmembrane protein targets using data from the Drugbank database. (a) Bar plots showing the number of transmembrane (TM) and non-TM proteins, which are targeted by FDA-approved, experimental, and all therapeutics in Drugbank. A TM protein target is identified from non-TM protein targets using the TMHMM tool (see Section 9) if at least one TM helix is detected in this protein target. (b) and (c) Bar and violin plots showing the number of TM proteins with respect to different TM helices, which are targeted by FDA-approved, experimental, and all therapeutics in Drugbank, respectively.

of any possible advances in prediction performance (Table 6). For example, in the AllesTM work, the ensemble of conventional machine learning methods (random forest) and deep learning methods (CNNs and bidirectional LSTM NNs) leads to superior performance in predicting Z-coordinates, flexibility, and topology, and its performance in predicting torsion angles, secondary structures, and monomer relative solvent accessibility is roughly similar to that of SPOT-1D. The ability to predict more than one property is particularly beneficial when being used for membrane protein design [150–152].

9. Membrane proteins and drug target interaction

The permeability of plasma membranes regulated by membrane proteins ensures control of the flow of ions, ligands, and other macromolecules from the extracellular to the intracellular environment. As such, membrane proteins are the targets of a variety of therapeutics [1245,246]. For example, many small molecules are used to target protein-protein interactions (discussed in [247–249]) between membrane proteins [250] or cell membrane-protein interaction interfaces [251]. It has been reported in extensive studies that membrane proteins are targeted by more than 50% of commercially available small molecule drugs [172,252–256]. Based on the Drugbank database (version: 01.04.2022) [47], we employed the TMHMM tool to identify the fraction of TM proteins, which are targeted by small molecule therapeutics and found that this comprises around half of the targets for all FDA-approved therapeutics (Fig. 7). The distribution of membrane proteins with respect to the number of helices identified bitopic proteins as the predominant target subtype across FDA-approved, experimental or all therapeutics. Polytopic proteins of 7, 2, and 4 helices were also found to be prevalent targets.

Small molecules therapeutics serve as inhibitors upon binding with aberrant membrane proteins, thus mitigating potential pathogenic effects. Therefore, the identification of drug-target interactions (DTIs) can be useful in the discovery and design of drugs. With the plethora of available sequence information and the accumulation of evidence-supporting DTI pairs, machine learning approaches have been applied for DTI prediction [257–259]. In recent years, the number of deep learning applications in this aspect has increased sharply [260], especially graph neural network applications thanks to the seamless integration of graph representation and drug structures [261]. More recently, GraphDTA, which was built using an attention-based graph neural network [262], has substantially improved prediction ability. Given that membrane proteins are greatly involved in DTIs, as indicated above, we assume that these tools will also have an acceptable ability to predict DTIs when the drug target is a membrane protein, even though the precise ratio of membrane proteins vs. globular proteins was not made clear by these aforementioned studies in their training datasets. This has been suggested in the DRUIDom work [263]. However, compared with sequence-based prediction of DTIs, the field is greatly devoid of prediction tools developed based on 3D protein structures [264]. With the increasing availability of Alphafold2-predicted structures for various membrane proteins, potential improvements in DTI prediction ability, through the incorporation of 3D structural features, will be of particular interest in the future.

10. Conclusion

In this review, we systematically evaluated the current status of machine learning applications being used to address four important membrane protein-related prediction problems: type classification, topology identification, interaction site prediction, and pathogenic effect prediction, followed by a summary of membrane proteins in drug-target interactions. We also summarized several key steps to perform such prediction tasks, including database collection, data pre-processing, feature extraction, co-evolution application, and method construction.

Membrane proteins comprise a major component of the human proteome and play significant biological roles, including that of cell gatekeepers to control the permeability of external particles as well as themselves having pharmacological potential. Ongoing progress has yielded a number of recently developed deep learning programs, which have shown potential in predicting accurate topological sites and interaction sites that are able to mirror experimental data, which open up opportunities for functional studies. Thus, building on these advances, this review can assist the development of computational strategies for future membrane protein modelling problems, which are the same or similar to the tasks discussed. Moreover, it is worth closely watching whether and how the performance of many kinds of membrane protein-related prediction tasks potentially benefits from the addition of predicted structural data in the future.

CRediT authorship contribution statement

Jianfeng Sun: Investigation, Data curation, Conceptualization, Methodology, Visualization, Validation, Writing – original draft, Writing – review & editing. **Arulsamy Kulandaisamy:** Investigation, Data curation, Methodology, Resources. **Jacklyn Liu:** Investigation, Validation, Writing – review & editing. **Kai Hu:** Methodology, Validation. **M. Michael Gromiha:** Methodology, Validation, Resources, Writing – review & editing. **Yuan Zhang:** Methodology, Project administration, Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the anonymous reviewers for their time and effort devoted to improving the quality of this work. This work is supported by the National Natural Science Foundation of China under Grants 62272404 and 61972333, in part by the Research Foundation of Education Department of Hunan Province of China under Grant 21B0172, and in part by the Natural Science Foundation of Hunan Province of China under Grant 2022JJ30571.

Appendix A. Glossary of technical terms used

Term	Description	Note
Deep learning models	Deep learning models are neural networks that can automatically extract features and learn representations from input data. To achieve this purpose, the neural networks need to be constructed with several special architectures, which mainly include convolutional neural networks, recurrent neural networks, and graph neural networks.	Deep learning models are generally applied for addressing image and speech recognition and those issues involving graph data.
Machine learning models	Machine learning models are algorithms that are used to handle classification and regression problems in an unsupervised (for unlabelled data samples) or a supervised (for labelled data samples) learning manner. All machine learning models discussed in this review are supervised learning models, which rely on a process of iteratively learning useful information from labelled input data to increase their prediction accuracy.	Deep learning models are included in machine learning models. Other commonly-seen machine learning models include the random forest and the supporting vector machine.
Feature extraction	Feature extraction is used to describe the process of capturing distinguished information from input data. For example, for classification problems, the extracted features are able to distinguish between different labels/categories of data samples.	-
Optimization processes	Training machine learning models to be intelligent to perform classification tasks relies on an optimization process in which the models iteratively learn useful information from input data. In each iteration, the models manage to minimize the difference between predicted labels and ground-truth labels, in order to improve its prediction performance.	-
Deep neural networks	Deep neural networks refer to those neural networks that belong to the deep learning models as discussed above.	The number of neural layers in deep neural networks is empirically greater than 2.
Convolutional neural networks/layers	A neural network consists of multiple neural layers, with each set to extract features from input data. A convolutional neural layer is a layer in which multiple weight matrices (or filters) are placed to perform convolutional operations on input image-like data.	-
Recurrent neural networks/layers	A recurrent neural network consists of multiple recurrent neural layers in which the input at the current time step is combined with the output from the previous time steps with feedback connections, which is suitable for modelling problems involving sequential data.	Recurrent neural networks are commonly used for speech recognition and natural language modelling.
Graph neural networks/layers	A graph neural layer in a graph neural network is the representation of a graph with nodes connected with edges, which feeds graph-structured data as input.	A typical example is that the structure of a small molecule can be converted to be a graph.
Residual neural networks	A residual neural network is a type of convolutional neural network, with every two convolutional layers connected residually, i.e., the final output of a residual neural layer is the addition of the output in the second convolutional layer and the raw input to the first convolutional layer.	The residually connected setting allows training deep neural networks at a fast speed.
Ultradeep neural networks	Ultradeep neural networks refer to those neural networks that consist of multiple neural layers. Although the number of neural layers required for this type of network is not well-defined in related fields, we empirically estimate that it should have more than 10 layers, or usually, tens or hundreds of layers.	-

Graph convolutional neural layers	A graph convolutional neural layer is a layer that allows extracting features from graph-structured data using convolutional operations.	-
Max-pooling operations	A subsampling strategy to reduce the dimension of input data	This operation is often used between convolutional layers.
Domain experts	In the context of machine learning applications in biology, domain experts are those that possess the knowledge about which biological features are informative to better perform a classification problem, without relying on the feature extraction by deep learning methods.	-
Auto-regressive and auto-encoder models	The auto-regressive model is a type of regression method to analyse a variable changing over time using time series data. The auto-encoder model is a type of neural network that learn informative features from unlabelled input data.	Auto-encoder models are widely used for image restoration.
Attention-based modules	Attention-based modules are those settings in neural layers, which can enhance the ability of extracting features from some particular parts of input data while weakening the ability in the remaining parts of the input data.	-
Tree-based approaches	The tree-based approach leverages a decision tree(s) to learn features from input data, which continuously split tree leaves to evaluate the possible consequences.	In each tree, leaves (i.e., nodes) represent the labels of training data samples and branches represent features corresponding to the labels.
Random forest algorithm	The random forest algorithm is a type of tree-based approach as discussed above.	A random forest consists of multiple decision trees to yield the final prediction result.
Mean decrease in impurity	The mean decrease in impurity, also called Gini importance, measures the total decrease in the impurity of nodes in a random forest, which is roughly calculated as the sum over the number of splits across all trees in the forest.	It helps to gauge the importance of a feature of interest in a prediction task.
Association rule learning analysis	The association rule learning analysis method is used to find associations between different data items.	It is used to make decisions about the co-occurrence of data items in a certain condition (e.g., associations between different diseases caused by pathogenic mutations).
Apriori algorithm	The Apriori algorithm is a type of association rule learning analysis method as described above.	-
Direct coupling analysis	The direct coupling analysis method is used to disentangle directly coupled residues (i.e., two residues are in physical contact) from indirectly coupled ones (i.e., the statistically associated contact of two residues but not the physical contact).	This method only takes as input a multiple sequence alignment of a protein to detect whether two residues in the protein are in physical contact.
Evolutionary coupling values	The evolutionary coupling values are yielded by the direct coupling analysis method to indicate whether two residues are in physical contact.	-
CASP events	The event of the critical assessment of techniques for protein structure prediction (CASP). The most recent CASP event is CASP15 (https://predictioncenter.org/casp15/index.cgi).	It is organized to evaluate the performance of protein structure prediction tools every two years.
Alpha helix-bundled	It is used to describe the morphology of α-helical transmembrane proteins, which are characterized by a bundle of α-helices facing one another within biological membranes.	-
Spatially close residues	A pair of residues are considered spatially close if the distance between them is within a predefined angstrom (e.g., 5.5 angstrom).	-
Biological assemblies	The biological assembly of a protein represents the actual biological unit functioning in oligomer form in the cellular context.	The structure of a protein complex downloaded from the protein data bank is an asymmetric unit of the protein complex.
TM protein interactions	A TM protein interaction refers to a physical interaction between a pair of TM proteins.	-

References

[1] Pollard T.D., Earnshaw W.C., Lippincott-Schwartz J., Johnson G.T. Cell Biology (Third Edition). 3rd ed. Elsevier; 2017.

[2] Yang Nicole J, Hinner MJ. Getting Across the Cell Membrane: An Overview for Small Molecules, Peptides, and Proteins. In: Gautier Arnaud, Hinner MJ, editors. Site-Specific Protein Labeling: Methods and Protocols New York, NY: Springer New York; 2015. p. 29–53. https://doi.org/10.1007/978-1-4939-2272-7_3

[3] Chapter 13 - Membrane Structure and Dynamics. In: Pollard TD, Earnshaw WC, Lippincott-Schwartz J, Johnson GT, editors. Cell Biology Third edition.,Elsevier; 2017. p. 227–39. https://doi.org/10.1016/B978-0-323-34126-4.00007-4

[4] Sze H. H+-Translocating ATPases: advances using membrane vesicles. Annu Rev Plant Physiol 1985;36:175–208. https://doi.org/10.1146/annurev.pp.36.060185.001135

[5] Nogueira J.J., Corry B. Ion Channel Permeation and Selectivity. The Oxford Handbook of Neuronal Ion Channels, Oxford University Press; n.d. https://doi.org/10.1093/oxfordhb/9780190669164.013.22.

[6] Zhao R, Diop-Bove N, Visentin M, Goldman ID. Mechanisms of membrane transport of folates into cells and across epithelia. Annu Rev Nutr 2011;31:177–201. https://doi.org/10.1146/annurev-nutr-072610-145133

[7] Chapter 24 - Plasma Membrane Receptors. In: Pollard TD, Earnshaw WC, Lippincott-Schwartz J, Johnson GT, editors. Cell Biology Third edition.,Elsevier; 2017. p. 411–23. https://doi.org/10.1016/B978-0-323-34126-4.00024-4. Third Edition.

[8] Hucho F, Weise C. Ligand-Gated Ion Channels. Angew Chem Int Ed 2001;40:3100–16. https://doi.org/10.1002/1521-3773(20010903)40:17<3100::AID-ANIE3100>3.0.CO;2-A

[9] Goldfarb M. Voltage-gated sodium channel-associated proteins and alternative mechanisms of inactivation and block. Cell Mol Life Sci 2012;69:1067–76. https://doi.org/10.1007/s00018-011-0832-1

[10] Wallin E, Heijne G von. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 1998;7:1029–38. https://doi.org/10.1002/pro.5560070420

[11] Sharpe HJ, Stevens TJ, Munro S. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. Cell 2010;142:158–69. https://doi.org/10.1016/j.cell.2010.05.037

[12] Guna A, Hegde RS. Transmembrane domain recognition during membrane protein biogenesis and quality control. Curr Biol 2018;28:R498–511. https://doi.org/10.1016/j.cub.2018.02.004

[13] Xia Y, Fischer AW, Teixeira P, Weiner B, Meiler J. Integrated structural biology for α-helical membrane protein structure determination. e2 Structure 2018;26:657–66. https://doi.org/10.1016/j.str.2018.02.006

[14] Thomas JMH, Simkovic F, Keegan R, Mayans O, Zhang C, Zhang Y, et al. Approaches to ab initio molecular replacement of α-helical transmembrane proteins. Acta Crystallogr Sect D 2017;73:985–96. https://doi.org/10.1107/S2059798317016436

[15] Carpenter EP, Beis K, Cameron AD, Iwata S. Overcoming the challenges of membrane protein crystallography. Curr Opin Struct Biol 2008;18:581–6. https://doi.org/10.1016/j.sbi.2008.07.001

[16] Cheng Y. Membrane protein structural biology in the era of single particle cryo-EM. Curr Opin Struct Biol 2018;52:58–63. https://doi.org/10.1016/j.sbi.2018.08.008

[17] Chagot B, Chazin WJ. Solution NMR structure of Apo-calmodulin in complex with the IQ motif of human cardiac sodium channel NaV1.5. J Mol Biol 2011;406:106–19. https://doi.org/10.1016/j.jmb.2010.11.046

[18] Chagot B, Potet F, Balser JR, Chazin WJ. Solution NMR structure of the C-terminal EF-hand domain of human cardiac sodium channel NaV1.5*. J Biol Chem 2009;284:6436–45. https://doi.org/10.1074/jbc.M807747200

[19] Wang C, Chung BC, Yan H, Wang H-G, Lee S-Y, Pitt GS. Structural analyses of Ca2+/CaM interaction with NaV channel C-termini reveal mechanisms of calcium-dependent regulation. Nat Commun 2014;5:4896. https://doi.org/10.1038/ncomms5896

[20] Gabelli SB, Boto A, Kuhns VH, Bianchet MA, Farinelli F, Aripirala S, et al. Regulation of the NaV1.5 cytoplasmic domain by calmodulin. Nat Commun 2014;5:5126. https://doi.org/10.1038/ncomms6126

[21] Sarhan MF, Tung C-C, Petegem F van, Ahern CA. Crystallographic basis for calcium regulation of sodium channels. Proc Natl Acad Sci 2012;109:3558–63. https://doi.org/10.1073/pnas.1114748109

[22] Johnson CN, Potet F, Thompson MK, Kroncke BM, Glazer AM, Voehler MW, et al. A mechanism of calmodulin modulation of the human cardiac sodium channel. e3 Structure 2018;26:683–94. https://doi.org/10.1016/j.str.2018.03.005

[23] Wang C, Chung BC, Yan H, Lee S-Y, Pitt GS. Crystal structure of the ternary complex of a NaV C-Terminal domain, a fibroblast growth factor homologous factor, and calmodulin. Structure 2012;20:1167–76. https://doi.org/10.1016/j.str.2012.05.001

[24] Gardill BR, Rivera-Acevedo RE, Tung C-C, Petegem van F. Crystal structures of Ca2+-calmodulin bound to NaV C-terminal regions suggest role for EF-hand domain in binding and inactivation. Proc Natl Acad Sci 2019;116:10763–72. https://doi.org/10.1073/pnas.1818618116

[25] Jin S, Zeng X, Xia F, Huang W, Liu X. Application of deep learning methods in biological networks. Brief Bioinform 2020. https://doi.org/10.1093/bib/bbaa043

[26] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2

[27] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577:706–10. https://doi.org/10.1038/s41586-019-1923-7

[28] Dobson L, Szekeres LI, Gerdán C, Langó T, Zeke A, Tusnády GE. TmAlphaFold database: membrane localization and evaluation of AlphaFold2 predicted alpha-helical transmembrane protein structures. Nucleic Acids Res 2022. https://doi.org/10.1093/nar/gkac928

[29] Marquet C, Grekova A, Houri L, Bernhofer M, Jimenez-Soto LF, Karl T, et al. TMvisDB: resource for transmembrane protein annotation and 3D visualization. BioRxiv 2022. https://doi.org/10.1101/2022.11.30.518551

[30] Chowdhary KR. Natural Language Processing. Fundamentals of Artificial Intelligence. New Delhi: Springer India,; 2020. p. 603–49. https://doi.org/10.1007/978-81-322-3972-7_19

[31] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2022;44:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381

[32] Bond-Taylor S, Leach A, Long Y, Willcocks CG. Deep generative modelling: a comparative review of VAEs, GANs, NOrmalizing Flows, Energy-based and Autoregressive Models. IEEE Trans Pattern Anal Mach Intell 2021:1. https://doi.org/10.1109/TPAMI.2021.3116668

[33] Golkov V, Skwark MJ, Golkov A, Dosovitskiy A, Brox T, Meiler J, et al. Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. In: Lee DD, Sugiyama M, Luxburg U v, Guyon I, Garnett R, editors. Advances in Neural Information Processing Systems 29. Curran Associates, Inc.; 2016. p. 4222–30.

[34] Zhou X, Zheng W, Li Y, Pearce R, Zhang C, Bell EW, et al. I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. Nat Protoc 2022;17:2326–53. https://doi.org/10.1038/s41596-022-00728-0

[35] Stein RR, Marks DS, Sander C. Inferring pairwise interactions from biological data using maximum-entropy probability models. PLoS Comput Biol 2015;11:e1004182.

[36] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011;6:e28766.

[37] Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proc Natl Acad Sci 2012;109:10340–5. https://doi.org/10.1073/pnas.1207864109

[38] Zeng B, Hönigschmid P, Frishman D. Residue co-evolution helps predict interaction sites in α-helical membrane proteins. J Struct Biol 2019;206:156–69. https://doi.org/10.1016/j.jsb.2019.02.009

[39] Sun J, Frishman D. DeepHelicon: accurate prediction of inter-helical residue contacts in transmembrane proteins by residual neural networks. J Struct Biol 2020;212:107574https://doi.org/10.1016/j.jsb.2020.107574

[40] Shimizu K, Cao W, Saad G, Shoji M, Terada T. Comparative analysis of membrane protein structure databases. Biochim Et Biophys Acta (BBA) - Biomembr 2018;1860:1077–91. https://doi.org/10.1016/j.bbamem.2018.01.005

[41] Kozma D, Simon I, Tusnády GE. PDBTM: protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res 2013;41:D524–9. https://doi.org/10.1093/nar/gks1169

[42] Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372:774–97. https://doi.org/10.1016/j.jmb.2007.05.022

[43] Bordner AJ, Gorin AA. Comprehensive inventory of protein complexes in the Protein Data Bank from consistent classification of interfaces. BMC Bioinforma 2008;9:234. https://doi.org/10.1186/1471-2105-9-234

[44] Bordner AJ. Predicting protein-protein binding sites in membrane proteins. BMC Bioinforma 2009;10:312. https://doi.org/10.1186/1471-2105-10-312

[45] Zaucha J, Heinzinger M, Kulandaisamy A, Kataka E, Salvádor ÓL, Popov P, et al. Mutations in transmembrane proteins: diseases, evolutionary insights, prediction and comparison with globular proteins. Brief Bioinform 2020. https://doi.org/10.1093/bib/bbaa132

[46] Kulandaisamy A, Ridha F, Frishman D, Gromiha MM. Computational approaches for investigating disease-causing mutations in membrane proteins: database development, analysis and prediction. Curr Top Med Chem 2022.

[47] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2017;46:D1074–82. https://doi.org/10.1093/nar/gkx1037

[48] Kulandaisamy A, Binny Priya S, Sakthivel R, Tarnovskaya S, Bizin I, Hönigschmid P, et al. MutHTP: mutations in human transmembrane proteins. Bioinformatics 2018;34:2325–6. https://doi.org/10.1093/bioinformatics/bty054

[49] Allen KN, Entova S, Ray LC, Imperiali B. Monotopic membrane proteins join the fold. Trends Biochem Sci 2019;44:7–20. https://doi.org/10.1016/j.tibs.2018.09.013

[50] Chang G, Roth CB. CHAPTER 19 - Signal Transduction and Integral Membrane Proteins. In: Bradshaw RA, Dennis EA, editors. Handbook of Cell Signaling Burlington: Academic Press; 2003. p. 115–8. https://doi.org/10.1016/B978-012124546-7/50379-X

[51] Wetzel RG. Extracellular Enzymatic Interactions: Storage, Redistribution, and Interspecific Communication. In: Chróst RJ, editor. Microbial Enzymes in Aquatic Environments New York, NY: Springer New York; 1991. p. 6–28. https://doi.org/10.1007/978-1-4612-3090-8_2

[52] Shen F, Huang Y-C, Tang S, Chen Y-X, Liu L. Chemical synthesis of integral membrane proteins: methods and applications. Isr J Chem 2011;51:940–52. https://doi.org/10.1002/ijch.201100076

[53] Pons M. Basic residue clusters in intrinsically disordered regions of peripheral membrane proteins: modulating 2D diffusion on cell membranes. Physchem 2021;1:152–62. https://doi.org/10.3390/physchem1020010

[54] Cymer F, von Heijne G, White SH. Mechanisms of integral membrane protein insertion and folding. J Mol Biol 2015;427:999–1022. https://doi.org/10.1016/j.jmb.2014.09.014

[55] Whited AM, Johs A. The interactions of peripheral membrane proteins with biological membranes. Chem Phys Lipids 2015;192:51–9. https://doi.org/10.1016/j.chemphyslip.2015.07.015

[56] Monje-Galvan V, Klauda JB. Peripheral membrane proteins: tying the knot between experiment and computation. Biochim Et Biophys Acta (BBA) - Biomembr 2016;1858:1584–93. https://doi.org/10.1016/j.bbamem.2016.02.018

[57] Steindorf D, Schneider D. In vivo selection of heterotypically interacting transmembrane helices: Complementary helix surfaces, rather than conserved interaction motifs, drive formation of transmembrane hetero-dimers. Biochim Et Biophys Acta (BBA) - Biomembr 2017;1859:245–56. https://doi.org/10.1016/j.bbamem.2016.11.017

[58] Larsen AH, John LH, Sansom MSP, Corey RA. Specific interactions of peripheral membrane proteins with lipids: what can molecular simulations show us. Biosci Rep 2022:42. https://doi.org/10.1042/BSR20211406

[59] Boes DM, Godoy-Hernandez A, McMillan DGG. Peripheral membrane proteins: promising therapeutic targets across domains of life. Membr (Basel) 2021:11. https://doi.org/10.3390/membranes11050346

[60] von Heijne G. Recent advances in the understanding of membrane protein assembly and structure. Q Rev Biophys 1999;32:285–307. https://doi.org/10.1017/S0033583500003541

[61] Lee AG. Biological membranes: the importance of molecular detail. Trends Biochem Sci 2011;36:493–500. https://doi.org/10.1016/j.tibs.2011.06.007

[62] Tusnády GE, Dosztányi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 2005;33:D275–8. https://doi.org/10.1093/nar/gki002

[63] Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res 2011;40:D370–6. https://doi.org/10.1093/nar/gkr703

[64] Tusnády GE, Dosztányi Z, Simon I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics 2004;21:1276–7. https://doi.org/10.1093/bioinformatics/bti121

[65] Lomize AL, Todd SC, Pogozheva ID. Spatial arrangement of proteins in planar and curved membranes by PPM 3.0. Protein Sci 2022;31:209–20. https://doi.org/10.1002/pro.4219

[66] Stansfeld PJ, Goose JE, Caffrey M, Carpenter EP, Parker JL, Newstead S, et al. MemProtMD: automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. Structure 2015;23:1350–61. https://doi.org/10.1016/j.str.2015.05.006

[67] Newport TD, Sansom MSP, Stansfeld PJ. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. Nucleic Acids Res 2018;47:D390–7. https://doi.org/10.1093/nar/gky1047

[68] Nastou KC, Tsaousis GN, Iconomidou VA. PerMemDB: A database for eukaryotic peripheral membrane proteins. Biochim Et Biophys Acta (BBA) - Biomembr 2020;1862:183076https://doi.org/10.1016/j.bbamem.2019.183076

[69] Sarti E, Aleksandrova AA, Ganta SK, Yavatkar AS, Forrest LR. EncoMPASS: an online database for analyzing structure and symmetry in membrane proteins. Nucleic Acids Res 2018;47:D315–21. https://doi.org/10.1093/nar/gky952

[70] Lomize AL, Lomize MA, Krolicki SR, Pogozheva ID. Membranome: a database for proteome-wide analysis of single-pass membrane proteins. Nucleic Acids Res 2016;45:D250–5. https://doi.org/10.1093/nar/gkw712

[71] Lomize AL, Schnitzer KA, Todd SC, Cherepanov S, Outeiral C, Deane CM, et al. Membranome 3.0: database of single-pass membrane proteins with AlphaFold models. Protein Sci 2022;31:e4318https://doi.org/10.1002/pro.4318

[72] White SH. Biophysical dissection of membrane proteins. Nature 2009;459:344–6. https://doi.org/10.1038/nature08142

[73] Bittrich S, Rose Y, Segura J, Lowe R, Westbrook JD, Duarte JM, et al. RCSB protein data bank: improved annotation, search and visualization of membrane protein structures archived in the PDB. Bioinformatics 2021;38:1452–4. https://doi.org/10.1093/bioinformatics/btab813

[74] Lomize AL, Hage JM, Pogozheva ID. Membranome 2.0: database for proteome-wide profiling of bitopic proteins and their dimers. Bioinformatics 2017;34:1061–2. https://doi.org/10.1093/bioinformatics/btx720

[75] Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2019;36:2105–12. https://doi.org/10.1093/bioinformatics/btz863

[76] van Meer G, Voelker DR, Feigenson GW. Membrane lipids: where they are and how they behave. Nat Rev Mol Cell Biol 2008;9:112–24. https://doi.org/10.1038/nrm2330

[77] Phillips R, Ursell T, Wiggins P, Sens P. Emerging roles for lipids in shaping membrane-protein function. Nature 2009;459:379–85. https://doi.org/10.1038/nature08147

[78] Lee AG. How lipids and proteins interact in a membrane: a molecular approach. Mol BioSyst 2005;1:203–12. https://doi.org/10.1039/B504527D

[79] Adamian L, Liang J. Prediction of transmembrane helix orientation in polytopic membrane proteins. BMC Struct Biol 2006;6:13. https://doi.org/10.1186/1472-6807-6-13

[80] Yano Y, Matsuzaki K. Live-cell imaging of membrane proteins by a coiled-coil labeling method—Principles and applications. Biochim Et Biophys Acta (BBA) - Biomembr 2019;1861:1011–7. https://doi.org/10.1016/j.bbamem.2019.02.009

[81] Chambers P, Pringle CR, Easton AJ. Heptad repeat sequences are located adjacent to hydrophobic regions in several types of virus fusion glycoproteins. J Gen Virol 1990;71:3075–80. https://doi.org/10.1099/0022-1317-71-12-3075

[82] Stone-Hulslander J, Morrison TG. Mutational analysis of heptad repeats in the membrane-proximal region of newcastle disease virus HN protein. J Virol 1999;73:3630–7. https://doi.org/10.1128/JVI.73.5.3630-3637.1999

[83] Mbaye MN, Hou Q, Basu S, Teheux F, Pucci F, Rooman M. A comprehensive computational study of amino acid interactions in membrane proteins. Sci Rep 2019;9:12043. https://doi.org/10.1038/s41598-019-48541-2

[84] Lomize AL, Pogozheva ID, Lomize MA, Mosberg HI. The role of hydrophobic interactions in positioning of peripheral proteins in membranes. BMC Struct Biol 2007;7:44. https://doi.org/10.1186/1472-6807-7-44

[85] Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. Mol Biol Evol 2015;33:268–80. https://doi.org/10.1093/molbev/msv211

[86] Figliuzzi M, Barrat-Charlaix P, Weigt M. How pairwise coevolutionary models capture the collective residue variability in proteins? Mol Biol Evol 2018;35:1018–27. https://doi.org/10.1093/molbev/msy007

[87] Coucke A, Uguzzoni G, Oteri F, Cocco S, Monasson R, Weigt M. Direct coevolutionary couplings reflect biophysical residue interactions in proteins. J Chem Phys 2016;145:174102https://doi.org/10.1063/1.4966156

[88] Hanson J, Paliwal KK, Litfin T, Yang Y, Zhou Y. Getting to know your neighbor: protein structure prediction comes of age with contextual machine learning. J Comput Biol 2020;27:796–814. https://doi.org/10.1089/cmb.2019.0193

[89] Singh A. Deep learning 3D structures. Nat Methods 2020;17:249. https://doi.org/10.1038/s41592-020-0779-y

[90] Hayat S, Sander C, Marks DS, Elofsson A. All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. 5413 LP – 5418 Proc Natl Acad Sci 2015;112. https://doi.org/10.1073/pnas.1419956112

[91] Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, et al. Large-scale determination of previously unsolved protein structures using evolutionary information. Elife 2015;4:e09248https://doi.org/10.7554/eLife.09248

[92] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci 2009;106:67–72. https://doi.org/10.1073/pnas.0805923106

[93] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci 2011;108:E1293–301. https://doi.org/10.1073/pnas.1111471108

[94] Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2011;28:184–90. https://doi.org/10.1093/bioinformatics/btr638

[95] Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. 15674 LP – 15679 Proc Natl Acad Sci 2013;110. https://doi.org/10.1073/pnas.1314045110

[96] Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Phys Rev E 2013;87:12707. https://doi.org/10.1103/PhysRevE.87.012707

[97] Zhou X, Zheng W, Li Y, Pearce R, Zhang C, Bell EW, et al. I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. Nat Protoc 2022. https://doi.org/10.1038/s41596-022-00728-0

[98] Hopf TA, Morinaga S, Ihara S, Touhara K, Marks DS, Benton R. Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. Nat Commun 2015;6:6077. https://doi.org/10.1038/ncomms7077

[99] Sjodt M, Brock K, Dobihal G, Rohs PDA, Green AG, Hopf TA, et al. Structure of the peptidoglycan polymerase RodA resolved by evolutionary coupling analysis. Nature 2018;556:118–21. https://doi.org/10.1038/nature25985

[100] Hegedűs T, Geisler M, Lukács GL, Farkas B. Ins and outs of AlphaFold2 transmembrane protein structure predictions. Cell Mol Life Sci 2022;79:73. https://doi.org/10.1007/s00018-021-04112-1

[101] Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 2012;149:1607–21. https://doi.org/10.1016/j.cell.2012.04.012

[102] Sun J, Frishman D. Improved sequence-based prediction of interaction sites in α-helical transmembrane proteins by deep learning. Comput Struct Biotechnol J 2021;19:1512–30. https://doi.org/10.1016/j.csbj.2021.03.005

[103] Kandathil SM, Greener JG, Jones DT. Recent developments in deep learning applied to protein structure prediction. Protein: Struct, Funct, Bioinforma 2019;87:1179–89. https://doi.org/10.1002/prot.25824

[104] Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. Protein: Struct, Funct, Bioinforma 2019;87:1149–64. https://doi.org/10.1002/prot.25792

[105] Ji Y, Lotfollahi M, Wolf FA, Theis FJ. Machine learning for perturbational single-cell omics. Cell Syst 2021;12:522–37. https://doi.org/10.1016/j.cels.2021.05.016

[106] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15:20170387. https://doi.org/10.1098/rsif.2017.0387

[107] Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform 2016;19:325–40. https://doi.org/10.1093/bib/bbw113

[108] Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nat Biotechnol 2018;36:829–38. https://doi.org/10.1038/nbt.4233

[109] Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of deep learning in biomedicine. Mol Pharm 2016;13:1445–54. https://doi.org/10.1021/acs.molpharmaceut.5b00982

[110] Laine E, Eismann S, Elofsson A, Grudinin S. Protein sequence-to-structure learning: is this the end(-to-end revolution)? Protein: Struct, Funct, Bioinforma 2021;89:1770–86. https://doi.org/10.1002/prot.26235

[111] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. Protein: Struct, Funct, Bioinforma 2021;89:1607–17. https://doi.org/10.1002/prot.26237

[112] Shrestha R, Fajardo E, Gil N, Fidelis K, Kryshtafovych A, Monastyrskyy B, et al. Assessing the accuracy of contact predictions in CASP13. Protein: Struct, Funct, Bioinforma 2019;87:1058–68. https://doi.org/10.1002/prot.25819

[113] Ruiz-Serra V, Pontes C, Milanetti E, Kryshtafovych A, Lepore R, Valencia A. Assessing the accuracy of contact and distance predictions in CASP14. Protein: Struct, Funct, Bioinforma 2021;89:1888–900. https://doi.org/10.1002/prot.26248

[114] Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Protein: Struct, Funct, Bioinforma 2021;89:1687–99. https://doi.org/10.1002/prot.26171

[115] Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatili P, Nielsen H, et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. BioRxiv 2022. https://doi.org/10.1101/2022.04.08.487609

[116] Bernhofer M, Rost B. TMbed: transmembrane proteins predicted through language model embeddings. BMC Bioinforma 2022;23:326. https://doi.org/10.1186/s12859-022-04873-x

[117] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539

[118] Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 2019;20:389–403. https://doi.org/10.1038/s41576-019-0122-6

[119] Cao Y, Geddes TA, Yang JYH, Yang P. Ensemble deep learning in bioinformatics. Nat Mach Intell 2020;2:500–8. https://doi.org/10.1038/s42256-020-0217-y

[120] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8:53. https://doi.org/10.1186/s40537-021-00444-8

[121] Dhillon A, Verma GK. Convolutional neural network: a review of models, methodologies and applications to object detection. Prog Artif Intell 2020;9:85–112. https://doi.org/10.1007/s13748-019-00203-0

[122] Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M. An introductory review of deep learning for prediction models with big data. Front Artif Intell 2020:3. https://doi.org/10.3389/frai.2020.00004

[123] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conf Comput Vis Pattern Recognit (CVPR) 2016:770–8. https://doi.org/10.1109/CVPR.2016.90

[124] He K, Zhang X, Ren S, Sun J. Leibe B, Matas J, Sebe N, Welling M, editors. Identity Mappings in Deep Residual Networks BT - Computer Vision – ECCV 2016. Cham: Springer International Publishing; 2016. p. 630–45.

[125] Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: a review of methods and applications. AI Open 2020;1:57–81. https://doi.org/10.1016/j.aiopen.2021.01.001

[126] Shuman DI, Narang SK, Frossard P, Ortega A, Vandergheynst P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process Mag 2013;30:83–98. https://doi.org/10.1109/MSP.2012.2235192

[127] Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform 2021:22. https://doi.org/10.1093/bib/bbab159

[128] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics,; 2019. p. 4171–86. https://doi.org/10.18653/v1/N19-1423

[129] Woo Sanghyun, Park J, LJ-Y, KIS. CBAM: Convolutional Block Attention Module. In: Vittorio Ferrari, Hebert M, SCWYeditors. Computer Vision – ECCV 2018. Cham: Springer International Publishing; 2018. p. 3–19.

[130] Chou K-C, Shen H-B. MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 2007;360:339–45. https://doi.org/10.1016/j.bbrc.2007.06.027

[131] Cedano J, Aloy P, Pérez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins11. In: Cohen FE, editor. J Mol Biol 1997;266:594–600. https://doi.org/10.1006/jmbi.1996.0804

[132] Arif M, Hayat M, Jan Z. iMem-2LSAAC: a two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition. J Theor Biol 2018;442:11–21. https://doi.org/10.1016/j.jtbi.2018.01.008

[133] Alballa M, Butler G. Integrative approach for detecting membrane proteins. BMC Bioinforma 2020;21:575. https://doi.org/10.1186/s12859-020-03891-x

[134] Butt AH, Khan SA, Jamil H, Rasool N, Khan YD. A prediction model for membrane proteins using moments based features. Biomed Res Int 2016;2016:8370132. https://doi.org/10.1155/2016/8370132

[135] Hayat M, Khan A. MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM. J Theor Biol 2012;292:93–102. https://doi.org/10.1016/j.jtbi.2011.09.026

[136] Wang M, Yang J, Xu Z-J, Chou K-C. SLLE for predicting membrane protein types. J Theor Biol 2005;232:7–15. https://doi.org/10.1016/j.jtbi.2004.07.023

[137] Chou K-C, Cai Y-D. Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inf Model 2005;45:407–13. https://doi.org/10.1021/ci049686v

[138] Ali F, Hayat M. Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition. J Theor Biol 2015;384:78–83. https://doi.org/10.1016/j.jtbi.2015.07.034

[139] Nastou KC, Tsaousis GN, Papandreou NC, Hamodrakas SJ. MBPpred: Proteome-wide detection of membrane lipid-binding proteins using profile Hidden Markov Models. Biochim Et Biophys Acta (BBA) - Proteins Proteom 2016;1864:747–54. https://doi.org/10.1016/j.bbapap.2016.03.015

[140] Guo L, Wang S, Cao Z. An ensemble classifier based on stacked generalization for predicting membrane protein types. 2017 10th Int Congr Image Signal Process, Biomed Eng Inform (CISP-BMEI) 2017:1–6. https://doi.org/10.1109/CISP-BMEI.2017.8302278

[141] Wang H, Ding Y, Tang J, Guo F. Identification of membrane protein types via multivariate information fusion with Hilbert–Schmidt Independence Criterion. Neurocomputing 2020;383:257–69. https://doi.org/10.1016/j.neucom.2019.11.103

[142] Guo L, Wang S, Li M, Cao Z. Accurate classification of membrane protein types based on sequence and evolutionary information using deep learning. BMC Bioinforma 2019;20:700. https://doi.org/10.1186/s12859-019-3275-6

[143] Zhang X, Chen L. Prediction of membrane protein types by fusing protein-protein interaction and protein sequence information. Biochim Et Biophys Acta (BBA) - Proteins Proteom 2020;1868:140524https://doi.org/10.1016/j.bbapap.2020.140524

[144] Chen W, Chen L, Dai Q. iMPT-FDNPL: identification of membrane protein types with functional domains and a natural language processing approach. Comput Math Methods Med 2021;2021:7681497. https://doi.org/10.1155/2021/7681497

[145] Zhou C, Zheng Y, Zhou Y. Structure prediction of membrane proteins. Genom Proteom Bioinforma 2004;2:1–5. https://doi.org/10.1016/S1672-0229(04)02001-7

[146] von Heijne G. Membrane-protein topology. Nat Rev Mol Cell Biol 2006;7:909–18. https://doi.org/10.1038/nrm2063

[147] Reeb J, Kloppmann E, Bernhofer M, Rost B. Evaluation of transmembrane helix predictions in 2014. Protein: Struct, Funct, Bioinforma 2015;83:473–84. https://doi.org/10.1002/prot.24749

[148] Tsirigos KD, Govindarajan S, Bassot C, Västermark Å, Lamb J, Shu N, et al. Topology of membrane proteins—predictions, limitations and variations. Curr Opin Struct Biol 2018;50:9–17. https://doi.org/10.1016/j.sbi.2017.10.003

[149] Adamian L, Liang J. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins1 1. (https://doi.org/https://doi.org/). In: von Heijne G, editor. J Mol Biol 2001;311:891–907. https://doi.org/10.1006/jmbi.2001.4908

[150] Vorobieva AA, White P, Liang B, Horne JE, Bera AK, Chow CM, et al. De novo design of transmembrane beta barrels. eabc8182 Science 1979;2021(371). https://doi.org/10.1126/science.abc8182

[151] Xu C, Lu P, Gamal El-Din TM, Pei XY, Johnson MC, Uyeda A, et al. Computational design of transmembrane pores. Nature 2020;585:129–34. https://doi.org/10.1038/s41586-020-2646-5

[152] Lu P, Min D, DiMaio F, Wei KY, Vahey MD, Boyken SE, et al. Accurate computational design of multipass transmembrane proteins. Science 1979;2018(359):1042–6. https://doi.org/10.1126/science.aaq1739

[153] Ananthasuresh GK. Protein Sequence Design on the Basis of Topology Optimization Techniques. Bendsøe Martin Philip and Olhoff N and SO, editor. IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials. Dordrecht: Springer Netherlands; 2006. p. 455–66.

[154] Singh A. Bottom-up de novo protein design. Nat Methods 2021;18:233. https://doi.org/10.1038/s41592-021-01097-4

[155] Li Z, Ni C, Xu J, Gao X, Cui S, Wang S. Transmembrane Topology Identification by Fusing Evolutionary and Co-Evolutionary Information with Cascaded Bidirectional Transformers. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, USA: Association for Computing Machinery,; 2019. p. 547. https://doi.org/10.1145/3307339.3343256

[156] Feng S-H, Zhang W-X, Yang J, Yang Y, Shen H-B. Topology prediction improvement of α-helical transmembrane proteins through helix-tail modeling and multiscale deep learning fusion. J Mol Biol 2020;432:1279–96. https://doi.org/10.1016/j.jmb.2019.12.007

[157] Wang L, Zhong H, Xue Z, Wang Y. Improving the topology prediction of α-helical transmembrane proteins with deep transfer learning. Comput Struct Biotechnol J 2022;20:1993–2000. https://doi.org/10.1016/j.csbj.2022.04.024

[158] Collobert R, Weston J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. Proceedings of the 25th International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery,; 2008. p. 160–7. https://doi.org/10.1145/1390156.1390177

[159] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 2021;118:e2016239118https://doi.org/10.1073/pnas.2016239118

[160] Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11. In: Cohen F, editor. J Mol Biol 2001;305:567–80. https://doi.org/10.1006/jmbi.2000.4315

[161] Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338:1027–36. https://doi.org/10.1016/j.jmb.2004.03.016

[162] The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2016;45:D158–69. https://doi.org/10.1093/nar/gkw1099

[163] Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049

[164] Tamposis IA, Sarantopoulou D, Theodoropoulou MC, Stasi EA, Kontou PI, Tsirigos KD, et al. Hidden neural networks for transmembrane protein topology prediction. Comput Struct Biotechnol J 2021;19:6090–7. https://doi.org/10.1016/j.csbj.2021.11.006

[165] Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849–50. https://doi.org/10.1093/bioinformatics/17.9.849

[166] Käll L, Krogh A, Sonnhammer ELL. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 2005;21:i251–7. https://doi.org/10.1093/bioinformatics/bti1014

[167] Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res 2007;35:W429–32. https://doi.org/10.1093/nar/gkm256

[168] Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol 2008;4:1–14. https://doi.org/10.1371/journal.pcbi.1000213

[169] Viklund H, Elofsson A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics 2008;24:1662–8. https://doi.org/10.1093/bioinformatics/btn221

[170] Viklund H, Bernsel A, Skwark M, Elofsson A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics 2008;24:2928–9. https://doi.org/10.1093/bioinformatics/btn550

[171] Dobson L, Reményi I, Tusnády GE. CCTOP: a Consensus Constrained TOPology prediction web server. Nucleic Acids Res 2015;43:W408–12. https://doi.org/10.1093/nar/gkv451

[172] Dobson L, Reményi I, Tusnády GE. The human transmembrane proteome. Biol Direct 2015;10:31. https://doi.org/10.1186/s13062-015-0061-x

[173] Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res 2015;43:W401–7. https://doi.org/10.1093/nar/gkv485

[174] Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A. Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci 2008;105:7177–81. https://doi.org/10.1073/pnas.0711151105

[175] Peters C, Tsirigos KD, Shu N, Elofsson A. Improved topology prediction using the terminal hydrophobic helices rule. Bioinformatics 2015;32:1158–62. https://doi.org/10.1093/bioinformatics/btv709

[176] Bernhofer M, Kloppmann E, Reeb J, Rost B. TMSEG: novel prediction of transmembrane helices. Protein: Struct, Funct, Bioinforma 2016;84:1706–16. https://doi.org/10.1002/prot.25155

[177] Yan R, Wang X, Huang L, Tian Y, Cai W. Transmembrane region prediction by using sequence-derived features and machine learning methods. RSC Adv 2017;7:29200–11. https://doi.org/10.1039/C7RA03883F

[178] Shen Hongbin, Chou JJ. MemBrain: improving the accuracy of predicting transmembrane helices. PLoS One 2008;3:1–6. https://doi.org/10.1371/journal.pone.0002399

[179] Lee D, Xiong D, Wierbowski S, Li L, Liang S, Yu H. Deep learning methods for 3D structural proteome and interactome modeling. Curr Opin Struct Biol 2022;73:102329https://doi.org/10.1016/j.sbi.2022.102329

[180] Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell 2014;159:1212–26. https://doi.org/10.1016/j.cell.2014.10.050

[181] Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature 2020;580:402–8. https://doi.org/10.1038/s41586-020-2188-x

[182] Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast two-hybrid, a powerful tool for systems biology. Int J Mol Sci 2009;10:2763–88. https://doi.org/10.3390/ijms10062763

[183] Shoemaker BA, Panchenko AR. Deciphering protein–protein interactions. part i. experimental techniques and databases. PLoS Comput Biol 2007;3:1–8. https://doi.org/10.1371/journal.pcbi.0030042

[184] Yugandhar K, Gupta S, Yu H. Inferring protein-protein interaction networks from mass spectrometry-based proteomic approaches: a mini-review. Comput Struct Biotechnol J 2019;17:805–11. https://doi.org/10.1016/j.csbj.2019.05.007

[185] Zhang X-F, Ou-Yang L, Hu X, Dai D-Q. Identifying binary protein-protein interactions from affinity purification mass spectrometry data. BMC Genom 2015;16:745. https://doi.org/10.1186/s12864-015-1944-z

[186] Morris JH, Knudsen GM, Verschueren E, Johnson JR, Cimermancic P, Greninger AL, et al. Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. Nat Protoc 2014;9:2539–54. https://doi.org/10.1038/nprot.2014.164

[187] Liu X, Salokas K, Weldatsadik RG, Gawriyski L, Varjosalo M. Combined proximity labeling and affinity purification−mass spectrometry workflow for mapping and visualizing protein interaction networks. Nat Protoc 2020;15:3182–211. https://doi.org/10.1038/s41596-020-0365-x

[188] Causier B, Davies B. Analysing protein-protein interactions with the yeast two-hybrid system. Plant Mol Biol 2002;50:855–70. https://doi.org/10.1023/A:1021214007897

[189] Lentze N, Auerbach D. Membrane-based yeast two-hybrid system to detect protein interactions. 19.17.1-19.17.28 Curr Protoc Protein Sci 2008;52. https://doi.org/10.1002/0471140864.ps1917s52

[190] Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature 2005;437:1173–8. https://doi.org/10.1038/nature04209

[191] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005;122:957–68. https://doi.org/10.1016/j.cell.2005.08.029

[192] Xiong D, Lee D, Li L, Zhao Q, Yu H. Implications of disease-related mutations at protein–protein interfaces. Curr Opin Struct Biol 2022;72:219–25. https://doi.org/10.1016/j.sbi.2021.11.012

[193] Shoemaker SC, Ando N. X-rays in the cryo-electron microscopy era: structural biology's dynamic future. Biochemistry 2018;57:277–85. https://doi.org/10.1021/acs.biochem.7b01031

[194] Schmidt C, Urlaub H. Combining cryo-electron microscopy (cryo-EM) and cross-linking mass spectrometry (CX-MS) for structural elucidation of large protein assemblies. Curr Opin Struct Biol 2017;46:157–68. https://doi.org/10.1016/j.sbi.2017.10.005

[195] Wang H-W, Wang J-W. How cryo-electron microscopy and X-ray crystallography complement each other. Protein Sci 2017;26:32–9. https://doi.org/10.1002/pro.3022

[196] Geraets JA, Pothula KR, Schröder GF. Integrating cryo-EM and NMR data. Curr Opin Struct Biol 2020;61:173–81. https://doi.org/10.1016/j.sbi.2020.01.008

[197] Hendrickson WA. Atomic-level analysis of membrane-protein structure. Nat Struct Mol Biol 2016;23:464–7. https://doi.org/10.1038/nsmb.3215

[198] Kermani AA. A guide to membrane protein X-ray crystallography. FEBS J 2021;288:5788–804. https://doi.org/10.1111/febs.15676

[199] Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. e6 Cell 2020;181:281–92. https://doi.org/10.1016/j.cell.2020.02.058

[200] Fuchs A, Martin-Galiano AJ, Kalman M, Fleishman S, Ben-Tal N, Frishman D. Co-evolving residues in membrane proteins. Bioinformatics 2007;23:3312–9. https://doi.org/10.1093/bioinformatics/btm515

[201] Lawson CL, Dutta S, Westbrook JD, Henrick K, Berman HM. Representation of viruses in the remediated PDB archive. Acta Crystallogr Sect D 2008;64:874–82. https://doi.org/10.1107/S0907444908017393

[202] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 2017;13:e1005324.

[203] Wang S, Li Z, Yu Y, Xu J. Folding membrane proteins by deep transfer learning. e3 Cell Syst 2017;5:202–11. https://doi.org/10.1016/j.cels.2017.09.001

[204] Dapkūnas J, Olechnovič K, Venclovas Č. Modeling of protein complexes in CASP14 with emphasis on the interaction interface prediction. Protein: Struct, Funct, Bioinforma 2021;89:1834–43. https://doi.org/10.1002/prot.26167

[205] Park T, Woo H, Yang J, Kwon S, Won J, Seok C. Protein oligomer structure prediction using GALAXY in CASP14. Protein: Struct, Funct, Bioinforma 2021;89:1844–51. https://doi.org/10.1002/prot.26203

[206] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Protein: Struct, Funct, Bioinforma 2019;87:1011–20. https://doi.org/10.1002/prot.25823

[207] Khazen G, Gyulkhandanian A, Issa T, Maroun RC. Getting to know each other: PPIMem, a novel approach for predicting transmembrane protein-protein complexes. Comput Struct Biotechnol J 2021;19:5184–97. https://doi.org/10.1016/j.csbj.2021.09.013

[208] Gunning AC, Fryer V, Fasham J, Crosby AH, Ellard S, Baple EL, et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. J Med Genet 2021;58:547–55. https://doi.org/10.1136/jmedgenet-2020-107003

[209] Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res 2018;46:7793–804. https://doi.org/10.1093/nar/gky678

[210] Zaucha J, Heinzinger M, Tarnovskaya S, Rost B, Frishman D. Family-specific analysis of variant pathogenicity prediction tools. NAR Genom Bioinform 2020:2. https://doi.org/10.1093/nargab/lqaa014

[211] Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol 2017;35:128–35. https://doi.org/10.1038/nbt.3769

[212] Ding D, Green AG, Wang B, Lite T-LV, Weinstein EN, Marks DS, et al. Co-evolution of interacting proteins through non-contacting and non-specific mutations. Nat Ecol Evol 2022;6:590–603. https://doi.org/10.1038/s41559-022-01688-0

[213] Garcia-Recio A, Gómez-Tamayo JC, Reina I, Campillo M, Cordomí A, Olivella M. TMSNP: a web server to predict pathogenesis of missense mutations in the transmembrane region of membrane proteins. NAR Genom Bioinform 2021:3. https://doi.org/10.1093/nargab/lqab008

[214] Kulandaisamy A, Priya SB, Sakthivel R, Frishman D, Gromiha MM. Statistical analysis of disease-causing and neutral mutations in human membrane proteins. Protein: Struct, Funct, Bioinforma 2019;87:452–66. https://doi.org/10.1002/prot.25667

[215] Kulandaisamy A, Zaucha J, Sakthivel R, Frishman D, Michael Gromiha M. Pred-MutHTP: prediction of disease-causing and neutral mutations in human transmembrane proteins. Hum Mutat 2020;41:581–90. https://doi.org/10.1002/humu.23961

[216] Popov P, Bizin I, Gromiha M, Kulandaisamy A, Frishman D. Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure. PLoS One 2019;14:1–13. https://doi.org/10.1371/journal.pone.0219452

[217] Pires DE v, Rodrigues CHM, Ascher DB. mCSM-membrane: predicting the effects of mutations on transmembrane proteins. Nucleic Acids Res 2020;48:W147–53. https://doi.org/10.1093/nar/gkaa416

[218] Ge F, Zhu Y-H, Xu J, Muhammad A, Song J, Yu D-J. MutTMPredictor: robust and accurate cascade XGBoost classifier for prediction of mutations in transmembrane proteins. Comput Struct Biotechnol J 2021;19:6400–16. https://doi.org/10.1016/j.csbj.2021.11.024

[219] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 2016;99:877–85. https://doi.org/10.1016/j.ajhg.2016.08.016

[220] Takeda J, Nanatsue K, Yamagishi R, Ito M, Haga N, Hirata H, et al. InMeRF: prediction of pathogenicity of missense variants by individual modeling for each amino acid substitution. NAR Genom Bioinform 2020:2. https://doi.org/10.1093/nargab/lqaa038

[221] Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 2020;11:5918. https://doi.org/10.1038/s41467-020-19669-x

[222] Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun 2021;12:510. https://doi.org/10.1038/s41467-020-20847-0

[223] Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature 2021;599:91–5. https://doi.org/10.1038/s41586-021-04043-8

[224] Won D-G, Kim D-W, Woo J, Lee K. 3Cnet: pathogenicity prediction of human variants using multitask learning with evolutionary constraints. Bioinformatics 2021;37:4626–34. https://doi.org/10.1093/bioinformatics/btab529

[225] Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods 2018;15:816–22. https://doi.org/10.1038/s41592-018-0138-4

[226] Tarnovskaya SI, Korkosh VS, Zhorov BS, Frishman D. Predicting novel disease mutations in the cardiac sodium channel. Biochem Biophys Res Commun 2020;521:603–11. https://doi.org/10.1016/j.bbrc.2019.10.142

[227] Ponzoni L, Nguyen NH, Bahar I, Brodsky JL. Complementary computational and experimental evaluation of missense variants in the ROMK potassium channel. PLoS Comput Biol 2020;16:1–20. https://doi.org/10.1371/journal.pcbi.1007749

[228] Frank Eibe, Hall M, HG, KR, PB, WIH, et al. Weka-a machine learning workbench for data mining. In: Maimon Oded, Rokach L, editors. Data Mining and Knowledge Discovery Handbook Boston, MA: Springer US; 2010. p. 1269–77. https://doi.org/10.1007/978-0-387-09823-4_66

[229] Agrawal R, Srikant R, et al. Fast algorithms for mining association rules. Proc. 20th int. conf. very large data bases vol. 1215. VLDB,; 1994. p. 487–99.

[230] Bodon F. A Trie-Based APRIORI Implementation for Mining Frequent Item Sequences. Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations. New York, NY, USA: Association for Computing Machinery,; 2005. p. 56–65. https://doi.org/10.1145/1133905.1133913

[231] Tahsili-Fahadan P, Geocadin RG. Heart–brain axis: effects of neurologic injury on cardiovascular function. Circ Res 2017;120:559–72. https://doi.org/10.1161/CIRCRESAHA.116.308446

[232] Firoz CK, Jabir NR, Khan MS, Mahmoud M, Shakil S, Damanhouri GA, et al. An overview on the correlation of neurological disorders with cardiovascular disease. Saudi J Biol Sci 2015;22:19–23. https://doi.org/10.1016/j.sjbs.2014.09.003

[233] Dworakowska B, Dołowy K. Ion channels-related diseases. Acta Biochim Pol 2000;47:685–703.

[234] Li M, Lester HA. Ion channel diseases of the central nervous system. CNS Drug Rev 2001;7:214–40. https://doi.org/10.1111/j.1527-3458.2001.tb00196.x

[235] Cooper EC, Jan LY. Ion channel genes and human neurological disease: recent progress, prospects, and challenges. Proc Natl Acad Sci 1999;96:4759–66. https://doi.org/10.1073/pnas.96.9.4759

[236] Vasconcelos LHC, Souza ILL, Pinheiro LS, Silva BA. Ion Channels in Obesity: Pathophysiology and Potential Therapeutic Targets. Front Pharm 2016:7. https://doi.org/10.3389/fphar.2016.00058

[237] Wang J, Ou S-W, Wang Y-J. Distribution and function of voltage-gated sodium channels in the nervous system. Channels 2017;11:534–54. https://doi.org/10.1080/19336950.2017.1380758

[238] Truong Aaron H, Murugesan S. and YKD and MA. Mitochondrial Ion Channels in Metabolic Disease. In: Levitan PhD I, DMDPAMeditors. Vascular Ion Channels in Physiology and Disease Cham: Springer International Publishing; 2016. p. 397–419. https://doi.org/10.1007/978-3-319-29635-7_18

[239] Kulandaisamy A, Sakthivel R, Gromiha MM. MPTherm: database for membrane protein thermodynamics for understanding folding and stability. Brief Bioinform 2020;22:2119–25. https://doi.org/10.1093/bib/bbaa064

[240] Kulandaisamy A, Zaucha J, Frishman D, Gromiha MM. MPTherm-pred: analysis and prediction of thermal stability changes upon mutations in transmembrane proteins. J Mol Biol 2021;433:166646https://doi.org/10.1016/j.jmb.2020.09.005

[241] Kroncke BM, Duran AM, Mendenhall JL, Meiler J, Blume JD, Sanders CR. Documentation of an imperative to improve methods for predicting membrane protein stability. Biochemistry 2016;55:5002–9. https://doi.org/10.1021/acs.biochem.6b00537

[242] Hönigschmid P, Breimann S, Weigl M, Frishman D. AllesTM: predicting multiple structural features of transmembrane proteins. BMC Bioinforma 2020;21:242. https://doi.org/10.1186/s12859-020-03581-8

[243] Li B, Mendenhall J, Capra JA, Meiler J. A multitask deep-learning method for predicting membrane associations and secondary structures of proteins. J Proteome Res 2021;20:4089–100. https://doi.org/10.1021/acs.jproteome.1c00410

[244] Mulnaes D, Schott-Verdugo S, Koenig F, Gohlke H. Topproperty: robust metaprediction of transmembrane and globular protein features using deep neural networks. J Chem Theory Comput 2021;17:7281–9. https://doi.org/10.1021/acs.jctc.1c00685

[245] Bennion BJ, Be NA, McNerney MW, Lao V, Carlson EM, Valdez CA, et al. Predicting a drug's membrane permeability: a computational model validated with in vitro permeability assay data. J Phys Chem B 2017;121:5228–37. https://doi.org/10.1021/acs.jpcb.7b02914

[246] Yin H, Flynn AD. Drugging membrane protein interactions. Annu Rev Biomed Eng 2016;18:51–76. https://doi.org/10.1146/annurev-bioeng-092115-025322

[247] Mabonga L, Kappo AP. Protein-protein interaction modulators: advances, successes and remaining challenges. Biophys Rev 2019;11:559–81. https://doi.org/10.1007/s12551-019-00570-x

[248] Kozakov D, Hall DR, Chuang G-Y, Cencic R, Brenke R, Grove LE, et al. Structural conservation of druggable hot spots in protein–protein interfaces. Proc Natl Acad Sci 2011;108:13528–33. https://doi.org/10.1073/pnas.1101835108

[249] Scott DE, Bayly AR, Abell C, Skidmore J. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. Nat Rev Drug Disco 2016;15:533–50. https://doi.org/10.1038/nrd.2016.29

[250] Stone TA, Deber CM. Therapeutic design of peptide modulators of protein-protein interactions in membranes. Biochim Et Biophys Acta (BBA) - Biomembr 2017;1859:577–85. https://doi.org/10.1016/j.bbamem.2016.08.013

[251] Chatzigoulas A, Cournia Z. Predicting protein–membrane interfaces of peripheral membrane proteins using ensemble machine learning. Brief Bioinform 2022:23. https://doi.org/10.1093/bib/bbab518

[252] Gromiha MM, Ou Y-Y. Bioinformatics approaches for functional annotation of membrane proteins. Brief Bioinform 2013;15:155–68. https://doi.org/10.1093/bib/bbt015

[253] Varga J, Dobson L, Reményi I, Tusnády GE. TSTMP: target selection for structural genomics of human transmembrane proteins. Nucleic Acids Res 2016;45:D325–30. https://doi.org/10.1093/nar/gkw939

[254] Rosenbaum MI, Clemmensen LS, Bredt DS, Bettler B, Strømgaard K. Targeting receptor complexes: a new dimension in drug discovery. Nat Rev Drug Disco 2020;19:884–901. https://doi.org/10.1038/s41573-020-0086-4

[255] Gulezian E, Crivello C, Bednenko J, Zafra C, Zhang Y, Colussi P, et al. Membrane protein production and formulation for drug discovery. Trends Pharm Sci 2021;42:657–74. https://doi.org/10.1016/j.tips.2021.05.006

[256] Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there. Nat Rev Drug Disco 2006;5:993–6. https://doi.org/10.1038/nrd2199

[257] Hasan Mahmud SM, Chen W, Jahan H, Dai B, Din SU, Dzisoo AM. DeepACTION: a deep learning-based method for predicting novel drug-target interactions. Anal Biochem 2020;610:113978https://doi.org/10.1016/j.ab.2020.113978

[258] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 2018;34:i821–9. https://doi.org/10.1093/bioinformatics/bty593

[259] Li F, Zhang Z, Guan J, Zhou S. Effective drug–target interaction prediction with mutual interaction neural network. Bioinformatics 2022;38:3582–9. https://doi.org/10.1093/bioinformatics/btac377

[260] Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Brief Bioinform 2020;22:247–69. https://doi.org/10.1093/bib/bbz157

[261] Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform 2021:22. https://doi.org/10.1093/bib/bbab159

[262] Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug–target binding affinity with graph neural networks. Bioinformatics 2020;37:1140–7. https://doi.org/10.1093/bioinformatics/btaa921

[263] Doğan T, Güzelcan EA, Baumann M, Koyas A, Atas H, Baxendale IR, et al. Protein domain-based prediction of drug/compound–target interactions and experimental validation on LIM kinases. PLoS Comput Biol 2021;17:1–34. https://doi.org/10.1371/journal.pcbi.1009171

[264] Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-learning-based drug–target interaction prediction. J Proteome Res 2017;16:1401–9. https://doi.org/10.1021/acs.jproteome.6b00618