bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450157; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Predicting protein-membrane interfaces of peripheral membrane proteins using ensemble machine learning

Alexios Chatzigoulas[1,2,*] and Zoe Cournia[1,*]

[1]Biomedical Research Foundation, Academy of Athens, 4 Soranou Ephessiou, 11527 Athens, Greece

[2]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, 15784 Athens, Greece

*To whom correspondence should be addressed.

Abstract

Abnormal protein-membrane attachment is involved in deregulated cellular pathways and in disease. Therefore, the possibility to modulate protein-membrane interactions represents a new promising therapeutic strategy for peripheral membrane proteins that have been considered so far undruggable. A major obstacle in this drug design strategy is that the membrane binding domains of peripheral membrane proteins are usually not known. The development of fast and efficient algorithms predicting the protein-membrane interface would shed light into the accessibility of membrane-protein interfaces by drug-like molecules. Herein, we describe an ensemble machine learning methodology and algorithm for predicting membrane-penetrating amino acids. We utilize available experimental data in the literature for training 21 machine learning classifiers and a voting classifier. Evaluation of the ensemble classifier accuracy produced a macro-averaged F_1 score = 0.92 and an MCC = 0.84 for predicting correctly membrane-penetrating amino acids on unknown proteins of an independent test set. The python code for predicting protein-membrane interfaces of peripheral membrane proteins is available at https://github.com/zoecournia/DREAMM.

Introduction

Membrane proteins are topologically divided in transmembrane proteins that are permanently attached in the interior of the membrane, peripheral membrane proteins that associate non-covalently with the surface of the membrane, and lipid-anchored proteins that attach to the membrane with a covalent bond [1]. Peripheral membrane proteins are essential in cellular processes such as transporting substances across the cell membrane, activating proteins and enzymes, regulating signal transduction, and other functions [1, 2]. Abnormal protein-membrane attachment due to membrane-binding domain mutations and peripheral membrane protein over- or under activation are involved in deregulated cellular pathways and in disease [1, 3-8]. Hence, the possibility to modulate protein-membrane interactions represents a promising therapeutic strategy for many disease indications and in particular for targeting membrane proteins that have been considered undruggable such as the membrane-anchored KRAS protein, which is implicated in over 30% of cancer types [9, 10], α-synuclein, which is a main pathological hallmark of Parkinson's disease [11, 12], and lipid kinases such as PI3Kα, which is the most frequently mutated kinase and present in a variety of tumors [13] with one of its hotspot mutations, H1047R, acting on altering the protein's association with the cell membrane [14-17].

The feasibility of targeting protein-membrane interfaces is supported by the fact that peripheral membrane proteins contain a membrane-binding domain with cavities that could be potentially targeted by small molecules [18, 19]. The literature reports the feasibility of targeting the protein-membrane interface indicating that therapeutic targets binding transiently to the membrane can be targeted with small molecules, and that inhibitors of protein-membrane interactions may be identified [18, 20-25]. However, these examples are only limited compared to the overall drug design efforts of the community indicating that the accessibility of protein-membrane interfaces by small molecules has been so far unexplored possibly due to the complexity of the interface, the limited protein-membrane structural information, and the absence of tools and workflows to automate the drug design process at the protein-membrane interface. Moreover, protein-membrane interaction sites of peripheral membrane proteins are commonly undiscovered; hence, the first step into modulating the protein-membrane interface is their identification.

Several efforts towards the design of tools that detect protein-membrane regions, domains, and lipid-binding sites have appeared [26-29]; however, these are mainly applied to directly to 1D protein sequences without considering the protein structural information and in many cases the web links are outdated [26-28]. To our knowledge, only two methodologies, which predict these interaction sites from the 3D protein structure, are currently publically available: the Positioning of Proteins in Membrane (PPM) [30, 31] and the Membrane Optimal Docking Area (MODA) [32]. PPM combines an anisotropic solvent representation of the lipid bilayer, an all atom representation of a solute, and a universal solvation model, calculating rotational and translational positions of transmembrane and peripheral membrane proteins in membranes [30, 31]. MODA is based on the protein-protein interface predictor PIER [33], which builds a set of evenly distributed points at 5 Å from one another and from the protein surface, defining each patch

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450157; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as the set of all protein surface atoms. Then, it calculates a score based on atom solvent-accessible surface area (SASA) and atom type specific weights, and transfers the patch membrane propensity scores to surface amino acids, thereby predicting which amino acids contact the cell membranes.

Herein, we present an automated prediction algorithm using ensemble machine learning, which identifies membrane-binding interfaces with high accuracy (macro-averaged F_1 score = 0.92 and an MCC = 0.84) taking as input the 3D peripheral membrane protein coordinates and demonstrates better accuracy than existing methods.

Methods

Data Preparation

To construct the dataset, we use 54 peripheral membrane proteins with known 3D structures and experimentally known membrane-penetrating amino acids, retrieved from extensive literature search. For the dataset generation, protein structures were prepared by deleting unwanted chains and co-crystallized solvent atoms, adding missing side chain atoms, and converting non-standard amino acids to their standard equivalents using HTMD [34]. In case of NMR-resolved structures, the first model of the NMR ensemble was kept. Then, the dataset was split in a training set (~85% of the dataset, Table S1) and a test set (~15% of the dataset, Table S2). Finally, a dataset of 12.805 amino acids consisting of the training set samples and a dataset of 2.177 amino acids consisting of the test set were assembled. These samples were labeled in two classes, the membrane-penetrating and the non-penetrating amino acids, leading to a highly imbalanced binary classification problem (supervised learning), where the membrane-penetrating amino acids comprise ~1.3% of the total samples in the training set. The feature extraction, and feature and data selection processes are explained in detail in the SI.

The most important features were expressing hydrophobicity and solvent exposure, but also: evolutionary conservation, secondary structure (coil loop or not), flexibility (squared fluctuations calculated with the Gaussian network model, probably associated with the flexible coil loops), dihedral angles (which again might be associated with the secondary structure), and Transferable Atom Equivalent (TAE) descriptors [35] (which are molecular properties related to electron density distribution) were determined to be critical for the accuracy of the results (more information in the SI, in Figure S1, and in Table S3). To balance the two classes several techniques were utilized leading to six different balanced datasets (see SI for more information and Figure S2).

Furthermore, a second test set of 11 peripheral membrane proteins with known 3D structures and experimentally known membrane-penetrating regions (not amino acids) was assembled. As no specific amino acids were experimentally tested, these proteins were assessed qualitatively as a second test set.

Finally, to ensure that the predictions are unbiased, the percentage of identical amino acids of the pairwise sequence alignment was calculated for all sequence pairs of the dataset, revealing high percentage identity values (more than 40%) for proteins in the training set, but not in the test sets, ensuring that the predictions in the test sets are unbiased (more information in the SI and Tables S4 and S5).

Ensemble machine learning methodology

For each one of the six training sets, 21 machine learning classifiers were trained: 19 from the scikit-learn Python package [36], the LightGBM classifier [37], and the XGBoost classifier [38]. The hyper-parameters of these classifiers were optimized to discover the best hyper-parameters that fit the data for each classifier (Table S6) [39, 40]. Specifically, for every training set and for every classifier the randomized search cross-validation technique was performed using 5 folds in a wide range of hyper-parameter values, training hundreds of thousands models (Table S6). Subsequently, iteratively exhaustive searches were performed (grid search cross-validation with 5 folds) in a small range of hyper-parameter values in the vicinity of the best hyper-parameter space determined from the randomized search cross-validation, which led to a set of optimal hyper-parameters. Notably, in the over- and under-sampled datasets, over- and under-sampling was performed in each fold separately to avoid information leak in the validation fold.

To assess the performance of the classifier models for both the randomized and grid search cross-validation procedures, the macro-averaged harmonic mean of the precision and recall, F_1 score, was chosen as a metric. Recall expresses the amount of correct predicted true positives (Eq. 1), while precision expresses the predicted true positives that are actually true (Eq. 2). The general formula of F score is derived based on a positive real variable β, where β determines the importance of recall over precision (Eq. 3). When β = 1 (F_1 score), recall and precision are weighted equally (Eq. 4), when β < 1 more weight is given in precision and when β > 1 recall is favored.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (1)$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (2)$$

$$F_\beta = (1+\beta^2)\frac{precision * recall}{(\beta^2 * precision) + recall} \quad (3)$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \quad (4)$$

For more information about the hyper-parameter tuning process and the performance metrics please refer to the SI.

Subsequently, the resulting predictions of the aforementioned classifiers were imputed as input to a meta-classifiers (second-level classifier). The voting classifier, which classifies a sample based on the majority voting of the first-level classifiers [41], and the stacking classifier, which

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450157; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

trains a classifier on the output of the first-level classifiers in order to compute the final prediction [42], were employed using the Python library mlxtend [43]. In both meta-classifiers, all possible combinations of the first-level classifiers were examined to discover the best classifier combination. Every classifier combination was tested using the independent test set with known protein-membrane amino acids (Table S2) to measure the combination with the best performance. Subsequently, considering that not every amino acids in the dataset was experimentally tested, resulting in membrane-penetrating amino acids marked as non-penetrating, the best models were manually inspected in order to assess their false positives, and the final model was chosen based on F_2 score (see Results). A schematic representation of the above procedure is illustrated in Figure 1.

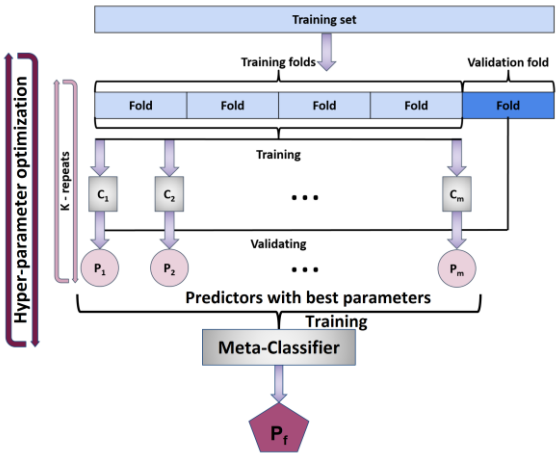

Figure 1. For each of the six datasets, we optimized the hyper-parameter space of 21 classifiers using 5-fold cross-validation on the training set. The predictions of the models with the best F1 score from these classifiers were provided as input to meta-classifiers. Given the F2 score on the test set the best meta-classifier was kept as the final predictor.

Results

The first-level classifiers providing the most accurate results were those trained in the initial dataset using weights. For these classifiers and initial dataset, several second-level classifiers exhibited better performance than the individual classifiers in terms of F_1 score, precision/recall area under the curve (PR AUC), Matthews correlation coefficient (MCC), and other metric scores. The Receiver Operating Characteristics (ROC) AUC, which is regularly used in literature, was not considered as it may be misleading for highly imbalanced classification problems such as our case (see the SI) [44, 45]. Then, results from the top second-level classifiers were subject to manual inspection and the best was selected according to the F_2 score to emphasize on recall. Although, seemingly, it is natural to prioritize on precision as in our case false positives are more critical than false negatives, manual inspection of the false positive results of the top second-level meta-classifiers indicated that these could actually be true positive membrane-penetrating amino acids as they are adjacent to amino acids that are membrane-penetrating or aligned with them to adjacent loops (see below). Finally, the best performing second-level classifier was the voting classifier for a combination consisting of five classifiers, the linear discriminant analysis, the logistic regression, the linear support vector classifier, the decision tree classifier, and the light gradient boosting machine. Various metric scores of the 21 first-level classifiers and the chosen second-level classifier for the initial dataset using weights can be viewed in Table S7.

The test set predictions can be viewed in Figure 2 and Table S8, where ~2/3 of the false positive amino acids are in fact correct predictions as they are located in the protein-membrane interface adjacent to true positives or on adjacent loops. For example, in retinoid isomerohydrolase, amino acid F262 is in a 4 Å distance from the experimentally confirmed membrane-penetrating amino acids; for the glycolipid transfer protein, amino acids I143 and Y153 are next to and aligned with W142. In other examples, i.e. the cholesterol-regulated Start protein 4, amino acid M196 although it is located in different loop, it is aligned with L124, as well as for the phosphatidylinositol transfer protein beta isoform, where M74 is in a different loop but aligned with the experimentally-confirmed membrane-penetrating amino acids W202 and W203; for the PH domain of the ceramide transfer protein, I37 and W40 are next to W33 and Y36, and F81 is aligned with them in an adjacent loop. Considering that these predictions are in fact located at the protein-membrane interface, they can be considered as true positives and the macro-averaged F_1 score increases from 0.86 to 0.92 and MCC from 0.71 to 0.84.

Moreover, the ensemble classifier was applied in the test set keeping all amino acid types. To reduce false positive non-hydrophobic amino acids, only those amino acids with center of mass (COM) distance of 14 Å from at least one of the predicted hydrophobic amino acids were kept (Table S9).

Furthermore, the performance of the classifier was compared to two computational tools that predict protein-membrane interfaces from 3D structures, the PPM web-server [46], which additionally predicts the orientation of proteins in membranes, and the MODA web-server [32], in the test set without performing data selection (Table S9). Generally, the predictions of every tool was fairly accurate in predicting the protein-membrane regions with our classifier outperforming them in some cases. Specifically, for the retinoid isomerohydrolase homodimer, PPM falsely predicted the orientation (probably affected by the missing chains) and placed the protein in an orientation in which only one monomer was in contact with the membrane instead of both, while our classifier and MODA correctly predicted the protein-membrane regions in both chains, but with a few false positive predictions for MODA (Figure S3). For the VSTx1 toxin, every tool predicted the protein-membrane region but falsely predicted W25. Moreover, our classifier falsely predicted as membrane-penetrating the amino acids near the N- and C-terminus and MODA falsely predicted the beta sheet on the opposite side of the protein-

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450157; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

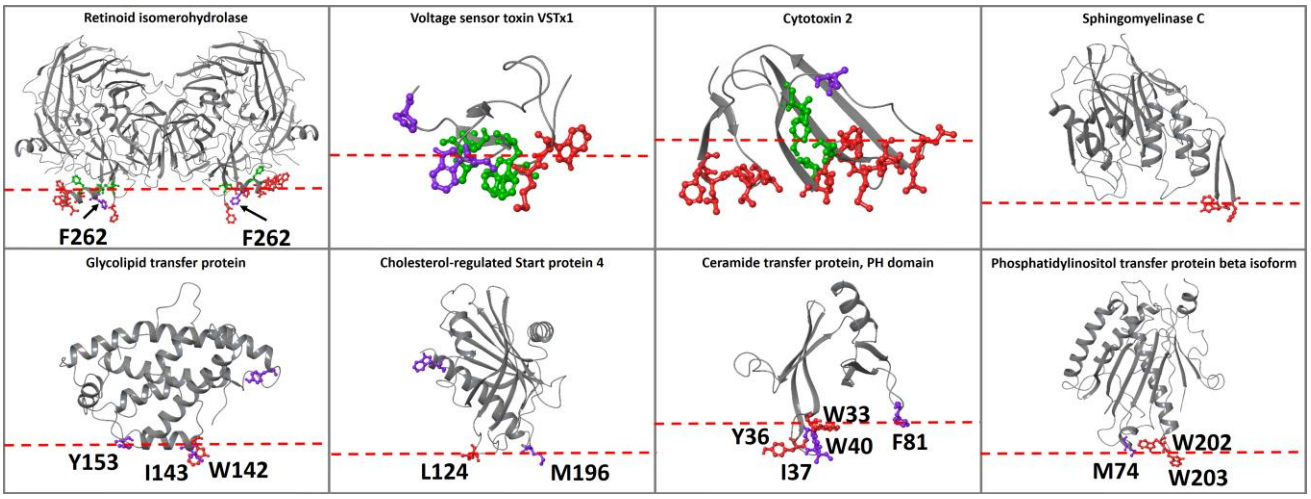

Figure 2. Proteins of the test set. The experimental membrane-penetrating amino acids predicted from the ensemble clas-sifier are depicted in red, the experimental membrane-penetrating amino acids not predicted from the classifier are depicted in green, and the amino acids predicted from the classifier that have not been experimentally-verified are depicted in purple.

membrane interface and the C-terminus region (Figure S4). For cytotoxin 2, all tools performed the same with a few false positives for our classifier and MODA in the 41-46 region (Figure S5). Similarly, for sphingomyelinase C all tools recognized the experimentally-verified membrane-penetrating amino acids W284 and F285, with PPM and MODA recognizing amino acids in distant loops that are aligned with the experimentally-verified protein-membrane region suggesting a multiregional interaction with the membrane, which is probably true, according to the proposed membrane binding model (Figure S6). For the glycolipid transfer protein, our classifier and MODA provided similar results correctly identifying the membrane-penetrating α-helix, with MODA falsely predicting the C-terminus and our classifier amino acid Y81 to be membrane-penetrating amino acids. PPM also suggested the insertion of the membrane-penetrating α-helix, but with the addition of the P40-P44 region (Figure S7). Likewise for the cholesterol-regulated Start protein 4, all tools predicted correctly

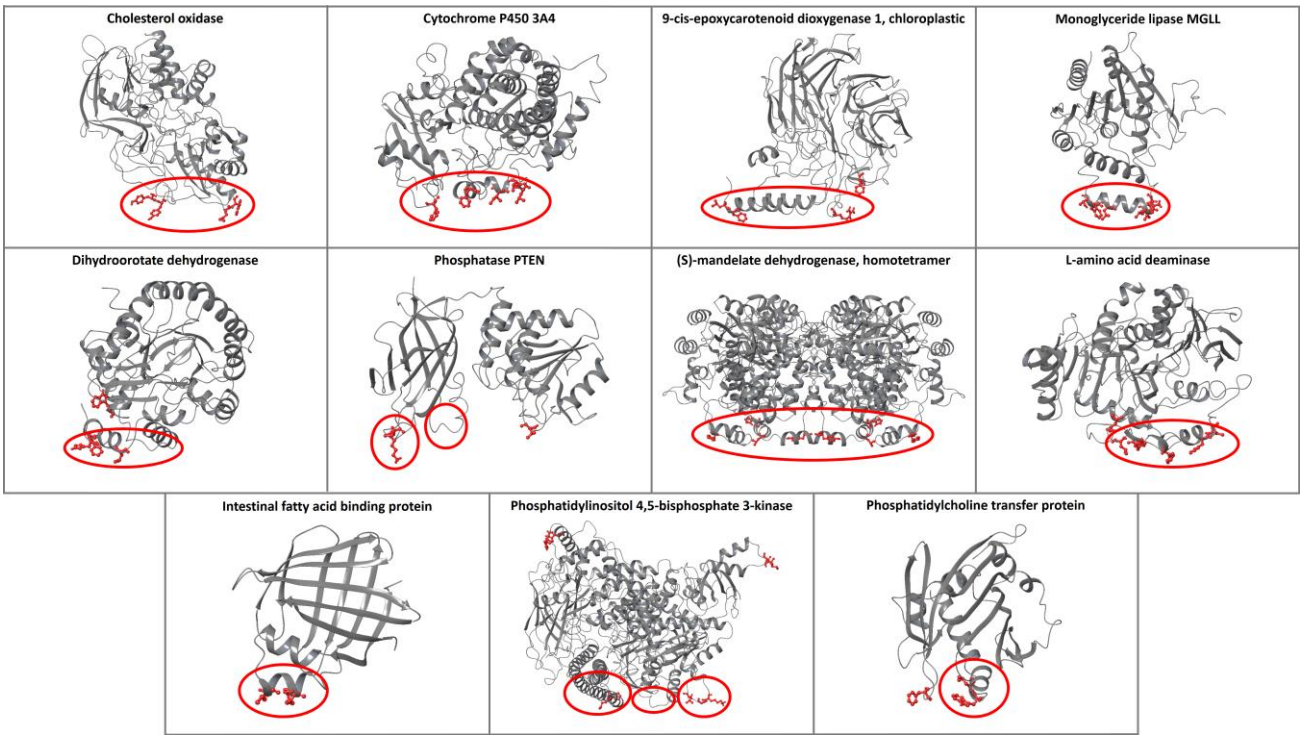

Figure 3. The predictions of the ensemble classifier for proteins with known protein-membrane interface regions. The membrane-penetrating amino acids predicted from the classifier are depicted in red, and the experimental membrane-penetrating regions are denoted with red circles.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450157; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the experimentally-verified amino acid L124 with our classifier and PPM additionally predicting the 196-200 region, MODA falsely predicting the C-terminus, and our classifier falsely predicting amino acid W91 (Figure S8). Finally, for the PH domain of the ceramide transfer protein and the phosphatidylinositol transfer protein beta isoform, the outcome was similar and correct for all tools (Figures S9-S10).

The performance of the ensemble classifier was tested with additional protein use cases with known membrane-penetrating regions (second test set, Figure 3), and the results were compared with PPM and MODA (Table S10). For the cases of cholesterol oxidase, cytochrome P450 3A4, monoglyceride lipase MGLL, L-amino acid deaminase, and intestinal fatty acid binding protein, all tools correctly identified the protein-membrane regions (Figures S11, S12, S14, S19, and S20). For the 9-cis-epoxycarotenoid dioxygenase 1, chloroplastic, all tools predicted the protein-membrane regions, with the exception of our classifier, which predicted the insertion of one of the two parallel amphipathic helices, instead of both (Figure S13). Similarly, for the dihydroorotate dehydrogenase, all tools predicted the protein-membrane regions, with our classifier falsely identifying the amino acid W362 and MODA the region 245-247 (Figure S15). For phosphatase PTEN, our classifier successfully identified the protein-membrane region 263-269 of the C2 domain and the region around the L42 amino acid of phosphatase in the same membrane plane. MODA also identified the same phosphatase region, however it falsely identified the opposite side of the C2 domain as a protein-membrane region. PPM also falsely identified the opposite side of the C2 domain suggesting an orientation, which is opposite to the actual membrane orientation (Figure S16). For (S)-mandelate dehydrogenase, the protein-membrane region was correctly identified by all tools, but our classifier and MODA also identified amino acids 53-56 to be membrane-penetrating (Figure S17); these amino acids are actually at the protein-protein interaction interface in the homotetramer of (S)-mandelate dehydrogenase and are not misclassified if we perform the predictions in the homotetramer biological assembly (Figure S18). For the phosphatidylinositol 4,5-bisphosphate 3-kinase alpha (PI3Kα), all tools predicted amino acids 232-233 as a protein-membrane interface, which is not in agreement with the experimental results, but is in fact the region where PI3Kα binds to K-RAS (Ras binding domain). Additionally, our classifier and MODA successfully identified the p110α 863-872 and the iSH2 512-525 regions, but falsely identified the 498-508 region, which links the C2 domain with the helical domain (Figure S21). The membrane orientation resulting from PPM is different from the one proposed through mutagenesis experiments [14]. Finally, for the phosphatidylcholine transfer protein, all tools provided the same results identifying the experimentally proven membrane-interacting region 184-193 and an adjacent loop, while MODA additionally predicted loop 147-148 to be membrane binding, which is in the same plane with the other two membrane binding regions (Figure S22).

Furthermore, the ensemble classifier was applied in the full structure of prothrombin (PDB: 5EDM [47]) and the results of identified membrane-protein interfaces were compared with PPM and MODA (Figure 4). All tools predicted the GLA domain 3-5 region as a membrane contacting region, which is natural for the classifier because the GLA domain of prothrombin was in the training set. Additionally, the classifier predicted the amino acids Y93, W398, and V458 and MODA predicted the amino acids Y93, Y377, R379, and R484, suggesting an orientation parallel to the actual membrane orientation, opposed to the perpendicular suggested by PPM. Y93 is a key prothrombin amino acid, which is essential for stabilizing the closed form and protects the active site pocket of the protease domain [48]. In the prothrombin closed form (PDB: 6BJR [48]), Y93 inserts its aromatic side chain into the binding pocket of the protease domain engaging W547 (W533 of 5EDM) and forms pi-pi interactions (Figure S23). The results provided by our classifier and MODA suggest that Y93 penetrates into the membrane, further indicating that when prothrombin engages the membrane the open form is favored with Y93 anchoring the membrane and opening the active site.

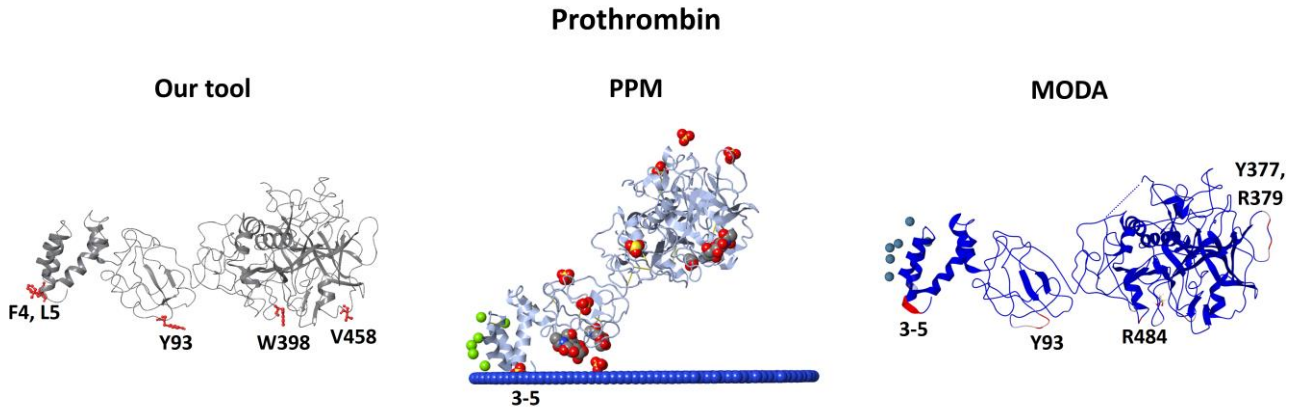

Figure 4. Comparison of the predictions provided from our classifier, PPM, and MODA for the open form of the pro-thrombin protein. Our classifier and MODA propose a parallel to the membrane orientation, suggesting the in-sertion of Y93 to the membrane, which in turn opens the active site of the protease domain (Figure S23).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450157; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, the ensemble classifier was also tested for the prediction of the protein-membrane interfaces of nine transmembrane enzymes described in Ref [49], which include a soluble domain performing extracellular catalysis. In agreement with experimental results, our classifier predicted numerous amino acids that lie in the hydrophobic lipid bilayer core, along with membrane-interacting extracellular amino acids (Figure S24).

Discussion and Conclusions

Drug design of protein-membrane interfaces for peripheral membrane proteins has been so far neglected due to the complexity of the interface and the lack of a suitable workflows and simulation technology capable of implementing this drug design strategy. Furthermore, protein-membrane interaction regions of peripheral membrane proteins are commonly unknown, and only a few rational methodologies exist that predict these regions from the 3D protein structure. To assist in protein-membrane interface recognition, a novel ensemble machine learning classifier is described trained in experimental data retrieved from extensive literature search.

The ensemble classifier results are accurate in predicting correctly the membrane-penetrating amino acids in the test set, providing with a macro-averaged F_1 score = 0.92 and an MCC = 0.84. Additionally, in a different independent dataset with experimentally known protein-membrane regions, our classifier correctly identified membrane-penetrating amino acids in these regions with a few false positive predictions. In addition, comparative results demonstrated that our classifier performed similarly, and in some cases better, than the only two available web-servers that predict protein-membrane interaction sites from the 3D protein structure, PPM and MODA. Moreover, our classifier successfully predicted the membrane-penetrating amino acids and the amino acids that lie in the hydrophobic core of the lipid bilayer in transmembrane proteins containing a soluble catalytic domain.

The features used in this study have a significant impact on the performance of the ensemble classifier. The fact that except from the hydrophobicity and the solvent exposure, other features—such us evolutionary conservation, secondary structure, flexibility, dihedral angles, and TAE descriptors—are important in our model decision-making process, offers novel physicochemical insights in the way that peripheral membrane proteins attach to the membrane.

Commonly, during development of computational tools several obstacles may emerge. In this case the first obstacle was the low number of the peripheral membrane proteins with experimentally known membrane-penetrating amino acids described in literature. The second and more crucial constraint was the small number of amino acids that were tested experimentally resulting in membrane-penetrating amino acids being marked as non-penetrating, which in turn resulted in misinforming the classifiers during the training process and rendering the selection of the best ensemble classifier strenuous. Moreover, based on the fact that membrane-penetrating amino acids are labeled as non-penetrating, the performance metrics (e.g. F scores) do not reflect the actual accuracy of the ensemble classifier, which is higher, and therefore, a direct numerical comparison of our classifier, PPM, and MODA results is not possible.

Manual inspection of false positive results revealed that several amino acids were located near the N- or C-terminus, or near missing loops, probably because the area is more solvent exposed. Intriguingly, other amino acids falsely predicted as membrane-penetrating are found to be implicated in protein-protein interactions. For example, in the case of (S)-mandelate dehydrogenase our model correctly classifies membrane-penetrating amino acids in the homotetramer form, while in the homodimer form we find that amino acids that are implicated in protein-protein interactions are also classified as membrane-penetrating. Hence, it is advisable to use the complex protein structure if it is available. The assumption that protein-membrane interactions are similar to protein-protein interactions was also deduced by [32], who adapted their protein-protein interaction interface prediction PIER algorithm [33] in MODA.

Also, it should be noted that the predictions depend on structural information, therefore, in the case where in the 3D protein structure the membrane-penetrating amino acids are in the bulk of the protein and a conformational change is necessary to face them towards the membrane, or the protein is intrinsically disordered, the ensemble classifier would not be able to predict them. Furthermore, it should be noted that neither our model, nor PPM or MODA are suitable tools for classifying if a protein is a peripheral membrane protein or not. A sequence-/evolutionary-based deep learning classifier would be more appropriate for this purpose [50]. Also, with the recent advancements in protein structure predictions, i.e., AlphaFold2 [51] and RoseTTAFold [52], the structure of unresolved proteins can be predicted with high accuracy, but in many cases, these models fail to fold the N- or C- terminus, or various protein segments. It is recommended to remove these regions, i.e., amino acids with confidence score (pLDDT) less than 70, before applying our model, since these unfolded regions are going to affect the prediction accuracy.

In the future, we plan to devise methods that orient peripheral membrane proteins in the membrane. Such an approach could include (1) placing the protein in a model membrane in all possible orientations based on the ensemble classifier predictions, (2) measuring the protein-membrane interface energy with an energy function specific for this purpose, and (3) retain the protein-membrane orientation with the lowest energy. When an appropriate method to define the membrane-protein orientation is devised, designing a web-database with our model predictions, similar to PPM and OPM [46] could also be envisaged.

Moreover, modifying the labeling system by dissecting the structures by a well-defined plane into membrane penetrating and not penetrating parts, or by labeling any amino acid from a specific distance from the experimentally known membrane penetrating amino acids as a true label warrants further investigation. Such a consideration would improve the class imbalance problem, although a defined plane and distance may be subjective to generate.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450157; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, it is noteworthy to mention that the membrane-penetrating amino acids are in many cases significant for the allosteric control of binding sites similar to prothrombin case. It is suggested that in the protein-membrane interfaces a binding pocket exists [18, 20-25]. We strongly believe that these binding pockets could act allosterically being connected with the active site and could be able to modulate the open/closed form equilibrium [17, 53] or even block protein function by disrupting the protein-membrane interactions.

Key points

- A dataset of peripheral membrane proteins with experimentally known membrane-penetrating amino acids was assembled.
- An ensemble machine learning classifier was trained utilizing thermodynamic, topographic, and property-based features.
- The ensemble machine learning classifier displayed a macro-averaged F_1 score = 0.92 and an MCC = 0.84 in identifying membrane-penetrating amino acids.
- The python code is publically available for usage at https://github.com/zoecournia/DREAMM.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported by the State Scholarships Foundation (IKY) [MIS-5000432 to A.C.]; the Hellenic Foundation for Research and Innovation (H.F.R.I.) [1780 to Z.C.]; and the high-performance computing resources provided by the National Infrastructures for Research and Technology S.A. (GRNET S.A.) [pr008033_gpu-Mem-Surf, pr008033_thin-Mem-Surf, and pr009022_gpu-Mem-Surf].

Conflict of Interest

The authors declare no conflict of interest.

References

1. Boes DM, Godoy-Hernandez A, McMillan DGG. Peripheral Membrane Proteins: Promising Therapeutic Targets across Domains of Life. *Membranes (Basel)*. 2021;**11**:346.
2. Monje-Galvan V, Klauda JB. Peripheral membrane proteins: Tying the knot between experiment and computation. *Biochim Biophys Acta*. 2016;**1858**:1584-93.
3. Segers K, Dahlback B, Nicolaes GA. Coagulation factor V and thrombophilia: background and mechanisms. *Thromb Haemost*. 2007;**98**:530-42.
4. Lashuel HA, Overk CR, Oueslati A, et al. The many faces of alpha-synuclein: from structure and toxicity to therapeutic target. *Nat Rev Neurosci*. 2013;**14**:38-48.
5. Mirsaeidi M, Gidfar S, Vu A, et al. Annexins family: insights into their functions and potential role in pathogenesis of sarcoidosis. *J Transl Med*. 2016;**14**:89.
6. Hobbs GA, Der CJ, Rossman KL. RAS isoforms and mutations in cancer at a glance. *J Cell Sci*. 2016;**129**:1287-92.
7. Costeira-Paulo J, Gault J, Popova G, et al. Lipids Shape the Electron Acceptor-Binding Site of the Peripheral Membrane Protein Dihydroorotate Dehydrogenase. *Cell chemical biology*. 2018;**25**:309-17 e4.
8. Mirza FJ, Zahid S. The Role of Synapsins in Neurological Disorders. *Neurosci Bull*. 2018;**34**:349-58.
9. Cox AD, Fesik SW, Kimmelman AC, et al. Drugging the undruggable RAS: Mission possible? *Nat Rev Drug Discov*. 2014;**13**:828-51.
10. Kessler D, Gmachl M, Mantoulidis A, et al. Drugging an undruggable pocket on KRAS. *Proc Natl Acad Sci U S A*. 2019;**116**:15823-9.
11. de Oliveira GAP, Silva JL. Alpha-synuclein stepwise aggregation reveals features of an early onset mutation in Parkinson's disease. *Communications biology*. 2019;**2**:374.
12. Hijaz BA, Volpicelli-Daley LA. Initiation and propagation of alpha-synuclein aggregation in the nervous system. *Mol Neurodegener*. 2020;**15**:19.
13. Yang J, Nie J, Ma X, et al. Targeting PI3K in cancer: mechanisms and advances in clinical trials. *Mol Cancer*. 2019;**18**:26.
14. Gabelli SB, Huang CH, Mandelker D, et al. Structural effects of oncogenic PI3Kalpha mutations. *Curr Top Microbiol Immunol*. 2010;**347**:43-53.
15. Gkeka P, Evangelidis T, Pavlaki M, et al. Investigating the structure and dynamics of the PIK3CA wild-type and H1047R oncogenic mutant. *PLoS Comput Biol*. 2014;**10**:e1003895.
16. Gkeka P, Papafotika A, Christoforidis S, et al. Exploring a non-ATP pocket for potential allosteric modulation of PI3Kalpha. *J Phys Chem B*. 2015;**119**:1002-16.
17. Cournia Z, Chatzigoulas A. Allostery in membrane proteins. *Curr Opin Struct Biol*. 2020;**62**:197-204.
18. Segers K, Sperandio O, Sack M, et al. Design of protein–membrane interaction inhibitors by virtual ligand screening, proof of concept with the C2 domain of factor V. *Proceedings of the National Academy of Sciences*. 2007;**104**:12697-702.
19. Sudhahar CG, Haney RM, Xue Y, et al. Cellular membranes and lipid-binding domains as attractive targets for drug development. *Curr Drug Targets*. 2008;**9**:603-13.
20. Spiegel PC, Kaiser SM, Simon JA, et al. Disruption of protein-membrane binding and identification of small-molecule inhibitors of coagulation factor VIII. *Chem Biol*. 2004;**11**:1413-22.
21. Liu Z, Lin L, Yuan C, et al. Trp2313-His2315 of factor VIII C2 domain is involved in membrane binding: structure of a complex between the C2 domain and an inhibitor of membrane binding. *J Biol Chem*. 2010;**285**:8824-9.
22. Nicolaes GA, Kulharia M, Voorberg J, et al. Rational design of small molecules targeting the C2 domain of coagulation factor VIII. *Blood*. 2014;**123**:113-20.
23. Chen L, Du-Cuny L, Moses S, et al. Novel inhibitors induce large conformational changes of GAB1 pleckstrin homology domain and kill breast cancer cells. *PLoS Comput Biol*. 2015;**11**:e1004021.
24. Nawrotek A, Benabdi S, Niyomchon S, et al. PH-domain-binding inhibitors of nucleotide exchange factor BRAG2 disrupt Arf GTPase signaling. *Nat Chem Biol*. 2019;**15**:358-66.
25. Li Z, Buck M. Computational Design of Myristoylated Cell-Penetrating Peptides Targeting Oncogenic K-Ras.G12D at the Effector-Binding Membrane Interface. *J Chem Inf Model*. 2020;**60**:306-15.
26. Scott DL, Diez G, Goldmann WH. Protein-lipid interactions: correlation of a predictive algorithm for lipid-binding sites with three-dimensional structural data. *Theor Biol Med Model*. 2006;**3**:17.
27. Bhardwaj N, Stahelin RV, Langlois RE, et al. Structural bioinformatics prediction of membrane-binding proteins. *J Mol Biol*. 2006;**359**:486-95.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.28.450157; this version posted October 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Sharikov Y, Walker RC, Greenberg J, et al. MAPAS: a tool for predicting membrane-contacting protein surfaces. *Nat Methods*. 2008;**5**:119.
29. Nastou KC, Tsaousis GN, Papandreou NC, et al. MBPpred: Proteome-wide detection of membrane lipid-binding proteins using profile Hidden Markov Models. *Biochim Biophys Acta*. 2016;**1864**:747-54.
30. Lomize AL, Pogozheva ID, Lomize MA, et al. Positioning of proteins in membranes: a computational approach. *Protein Sci*. 2006;**15**:1318-33.
31. Lomize AL, Pogozheva ID, Mosberg HI. Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes. *J Chem Inf Model*. 2011;**51**:930-46.
32. Kufareva I, Lenoir M, Dancea F, et al. Discovery of novel membrane binding structures and functions. *Biochem Cell Biol*. 2014;**92**:555-63.
33. Kufareva I, Budagyan L, Raush E, et al. PIER: protein interface recognition for structural proteomics. *Proteins*. 2007;**67**:400-17.
34. Doerr S, Harvey MJ, Noe F, et al. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J Chem Theory Comput*. 2016;**12**:1845-52.
35. Whitehead CE, Breneman CM, Sukumar N, et al. Transferable atom equivalent multicentered multipole expansion method. *J Comput Chem*. 2003;**24**:512-29.
36. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;**12**:2825-30.
37. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Sys*. 2017:3146-54.
38. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA. 2939785: ACM; 2016. p. 785-94.
39. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. *J Mach Learn Res*. 2012;**13**:281-305.
40. Claesen M, De Moor B. Hyperparameter search in machine learning. *arXiv preprint arXiv:150202127*. 2015.
41. Littlestone N, Warmuth MK. The Weighted Majority Algorithm. *Information and Computation*. 1994;**108**:212-61.
42. Wolpert DH. Stacked Generalization. *Neural Networks*. 1992;**5**:241-59.
43. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of open source software*. 2018;**3**:638.
44. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning; Pittsburgh, Pennsylvania, USA: Association for Computing Machinery; 2006. p. 233–40.
45. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 2015;**10**:e0118432.
46. Lomize MA, Pogozheva ID, Joo H, et al. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res*. 2012;**40**:D370-6.
47. Pozzi N, Chen Z, Di Cera E. How the Linker Connecting the Two Kringles Influences Activation and Conformational Plasticity of Prothrombin. *J Biol Chem*. 2016;**291**:6071-82.
48. Chinnaraj M, Chen Z, Pelc LA, et al. Structure of prothrombin in the closed form reveals new details on the mechanism of activation. *Sci Rep*. 2018;**8**:2945.
49. Dufrisne MB, Petrou VI, Clarke OB, et al. Structural basis for catalysis at the membrane-water interface. *Biochim Biophys Acta Mol Cell Biol Lipids*. 2017;**1862**:1368-85.
50. Guo L, Wang S, Li M, et al. Accurate classification of membrane protein types based on sequence and evolutionary information using deep learning. *BMC Bioinf*. 2019;**20**:700.
51. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;**596**:583-9.
52. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021;**373**:871-6.
53. Chatzigoulas A, Cournia Z. Rational design of allosteric modulators: Challenges and successes. *WIREs Comput Mol Sci*. 2021;**11**:e1529.