Updating the *in silico* human surfaceome with meta-ensemble learning and feature engineering

Daniel Bojar[1*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, 4058, Basel, Switzerland

e-mail: daniel.bojar@bsse.ethz.ch, TEL: +41 61 387 32 83

Abstract

Next to being targeted by most available drugs, human proteins located in the plasma membrane are also responsible for a plethora of essential cellular functions, ranging from signaling to transport processes. In order to target and study these transmembrane proteins, their plasma membrane location has to be established. Yet experimental validation of the thousands of potential plasma membrane proteins is laborious and technically challenging. A recent study performed machine learning to classify surface and non-surface transmembrane proteins in human cells based on curated high-quality training data from the Cell Surface Protein Atlas (CSPA) and other databases, reporting a cross-validation prediction accuracy of 93.5% (1). Here, we report an improved version of the surfaceome predictor SURFY, SURFY2, using the same training data with a meta-ensemble classification approach involving feature engineering. SURFY2 yielded predictions with an accuracy score of 95.5% on a test dataset never seen before by the classifier. Importantly, we found several high-confidence re-classifications of disease-relevant proteins among the discrepant predictions between SURFY and SURFY2. To rationalize the prediction mechanism of SURFY2 and analyze differently classified transmembrane proteins we investigated classifier feature importances and data distributions between prediction sets. SURFY2 exhibited both an increased precision as well as recall compared to SURFY and delivers the best *in silico* human surfaceome up to now. This updated version of the surfaceome will instigate further advances in drug targeting and research on cellular signaling as well as transport processes.

Keywords: machine learning, surfaceome, cell surface protein, SURFY, classification

Introduction

From receptors allowing cells to communicate with the exterior world to transporters enabling the exchange of molecules, proteins located in the plasma membrane are integral to cellular functions. In human cells, virtually all surface-exposed transmembrane proteins enter the secretory pathway comprising the ER and the Golgi prior to their arrival at the plasma membrane. This is ensured by the presence of an N-terminal signal peptide rerouting translating ribosomes to the ER. Yet not all transmembrane proteins containing a signal peptide sequence finish their journey in the plasma membrane. Non-surface transmembrane proteins can reside in the ER, the Golgi, lysosomes or other related compartments. Thus, the presence of a signal peptide is not sufficient to determine the surface exposure of a transmembrane protein. The question which features of a protein are deciding its subcellular fate is therefore of paramount importance.

Determining or accurately predicting the subcellular location of a transmembrane protein is of considerable interest. While intracellular transmembrane proteins are hard to study or manipulate, surface-exposed transmembrane proteins can be studied by probes without their internalization (e.g., antibodies or labeled ligands) (2). Additionally, pharmaceutical drugs are frequently directed against cell-surface proteins (3), making this information especially pertinent to the development of new therapies. A comprehensive catalog of surface transmembrane proteins would therefore constitute an extremely valuable repository to choose targets for research and drug development. While most resources infer or predict the subcellular localization of proteins (4,5), some databases such as the Cell Surface Protein Atlas (CSPA) experimentally determined the surface-exposed state of hundreds of human proteins by techniques such as mass spectrometry (6). Due to the low abundance of transmembrane proteins and their dependency on cell state and cell identity, the CSPA is however not including the whole surfaceome, the set of all surface proteins.

Being presented with an experimentally determined and highly curated database (CSPA and other sources) and the challenge to sort transmembrane proteins not covered by this database into two classes (surface-exposed and intracellular), is a classic set-up for machine learning. More specifically, supervised binary classification via a classifier trained on the input data could lead to the accurate prediction of surface proteins. Performed by the same research group which created the CSPA, this endeavor was undertaken with the machine learning-based surfaceome predictor SURFY (1). Rising to prominence in biology by automation of phenotype recognition (7), image classification (8), and many other applications (9), machine learning aims to find and recognize patterns in data to predict a state or make decisions. By using features from AAindex1 (10) and UniProt (11), a classifier model was built. Consisting of a random forest algorithm (12), an ensemble method utilizing an array of decision trees trained on random subsets of the features and samples, SURFY resulted in an out-of-bag accuracy of 93.5%. The authors then used their classifier to predict the plasma membrane localization of the remaining human transmembrane proteins, yielding a human surfaceome of 2886 proteins and substantially advancing the field of human membrane protein research by equipping it with this resource.

By averaging over a large number of decision trees, random forest algorithms are a popular choice in machine learning and showcase a decidedly lower variance than their base estimators (13). Yet even though they often perform surprisingly well, random forests are far from the pinnacle of machine learning. Even with optimized hyperparameters, they eventually reach a prediction accuracy limit for a given dataset. Just as the random forest algorithm utilizes decision trees as estimators for an ensemble, other classifiers such as voting classifiers can use algorithms such as random forest as estimators. Usually, using these kinds of meta-ensembles (or ensembles of ensembles) highlights the strengths of the individual classifiers while smoothening their weaknesses, resulting in superior outcomes (14). Other methods to combine different classifiers would include stacking, in which the out-of-fold predictions of individual classifiers are used as new features for a meta-classifier (15). This procedure to create new features from properties of the original features is also referred to as feature engineering. Both methods, voting classifiers and stacking classifiers profit from diverse base estimators which are not too highly correlated (16,17), as there would be no benefit to combining their predictions otherwise.

Here, we report the utilization of feature engineering by a stacking classifier which uses a voting classifier as a meta-estimator to create SURFY2, effectively constituting a meta-ensemble classifier. Making use of a richer diversity of optimized base estimators, SURFY2 is able to achieve higher recall, precision and accuracy values than SURFY on a test dataset never seen by the model before, with a clear improvement seen in a final prediction accuracy of 95.5%. We also probed the model by investigating its feature importances. Thanks to the superior attributes of SURFY2, we identified 114 cases of different classification of transmembrane proteins in the unlabeled dataset by SURFY and SURFY2. Potential reasons why these proteins were initially classified otherwise were explored by the comparison of feature importances between the two approaches, an investigation of the respective data distributions and a visualization through nonlinear dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE). We envision this updated, more accurate, version of the human surfaceome to assist and accelerate the development of knowledge about human surface proteins as well as novel pharmaceutical drugs for human diseases.

Results

Building the meta-ensemble classifier for human surfaceome prediction

The same dataset as already described in detail in the study by Bausch-Fluck et al. (1) was used to train our classifiers. Consisting of 253 features used for classification in total, it contained, among others, features regarding the presence and characteristics of transmembrane segments, glycosylation sites and cytoplasmic domains. Some features were present both as absolute counts as well as their relative fraction. The training dataset comprised 910 surface-exposed and 657 intracellular transmembrane proteins from the CSPA and other sources (Supplementary Table 1).

After splitting the training dataset into train, validation and test sets, the model was built on the train data. First, 12 classifiers from the scikit-learn Python library were trained on the train data for predicting the ‘surface’ feature. These classification algorithms comprised Random Forest, Logistic Regression, Gradient Boosting, Extra-Trees, Gaussian Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, XGBClassifier, AdaBoost, Multilayer Perceptron, Linear Discriminant Analysis and Quadratic Discriminant Analysis algorithms. After optimizing hyperparameters through cross-validation, these classifiers were used to predict class probabilities on out-of-fold samples. These generated class probabilities were then used as 24 new engineered features and merged with the initial dataset.

Using a greedy selection procedure, a soft voting classifier relying on class probabilities was trained on the original train dataset, scored via cross-validation and finalized by fixing optimized Random Forest, Logistic Regression and Gradient Boosting estimators as its constituents. Based on cross-validation scores, the weights of the estimators were optimized in an exhaustive search and fixed at (3,2,3) for the abovementioned estimators respectively.

To finalize the selection of a classifier, all base estimators including the voting classifier were tested as final models on the initial training data as well as the expanded data including the engineered features (Supplementary Table 2). The final model was then chosen based on 5-fold mean cross-validation score on the training data. One immediate observation that could be made was that nearly every classifier profited from the engineered features, resulting in an increased mean cross-validation accuracy score. This strongly supported our decision to expand the initial dataset with the predicted class probabilities as new features to facilitate an improved classification procedure.

Combining the feature engineering performed by the base estimators with the soft voting classifier in a meta-ensemble resulted in the best outcome in terms of cross-validation scores (accuracy score: 0.949). Predicting the ‘surface’ feature of the validation set with an accuracy of 95.7%, SURFY2 was then chosen as the described meta-ensemble approach (Figure 1).

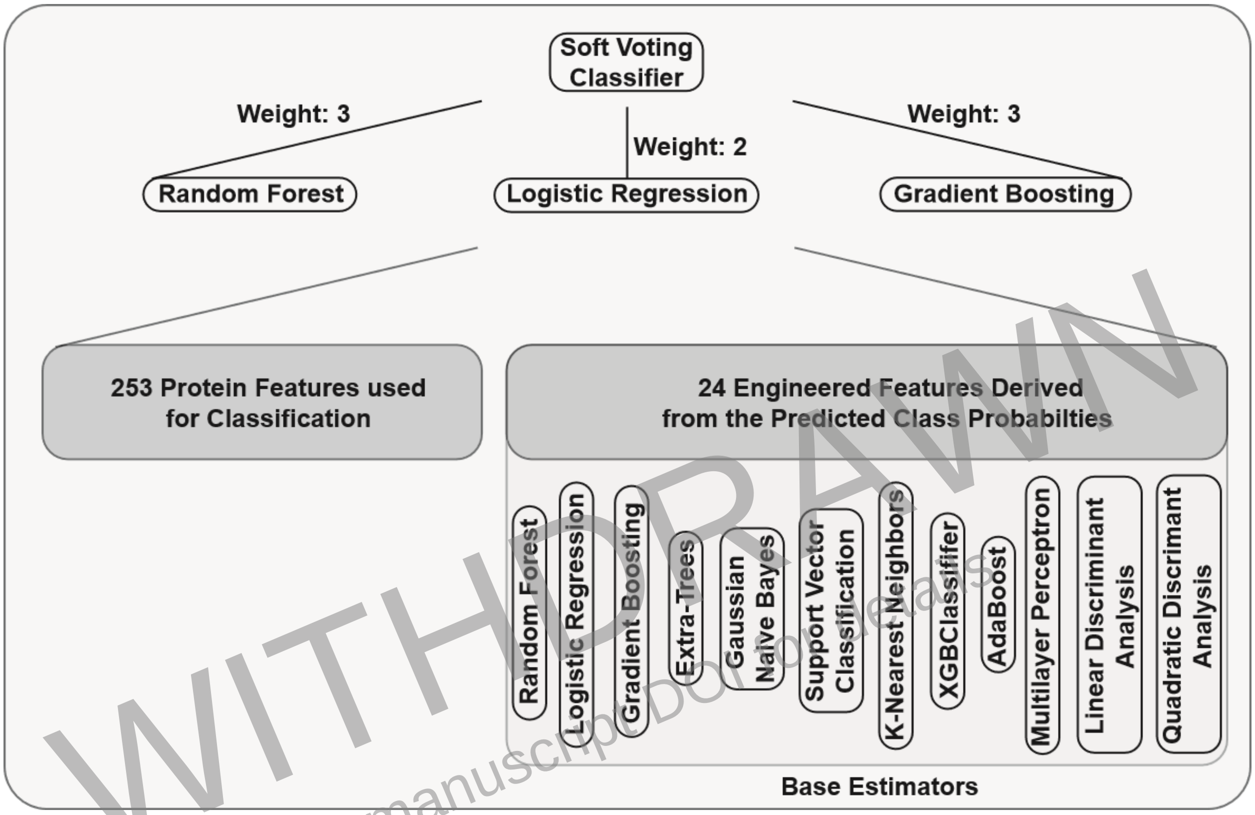

Figure 1: Meta-ensemble classifier SURFY2 relying on engineered features. The predicted class probabilities for both classes of 12 base estimators with optimized hyperparameters were appended to the 253 original features used for classification. The whole dataset was then used to train a soft voting classifier with optimized Random Forest, Logistic Regression and Gradient Boosting estimators. The resulting class probabilities achieved by these estimators were summed up according to their weight and the class with the maximum probability was chosen as the predicted class.

Finally, SURFY2 was trained on the combined train and validation dataset and used to predict the class variable in the test dataset which was not previously used for any training or hyperparameter optimization. Observing a prediction accuracy of 95.5% with accompanying excellent precision and recall values (Table 1), the robustness of SURFY2 was established. With test set precision and recall values of 94.3% and 97.6%, respectively, SURFY2 had both lower false positive (5.7%) as well as false negative (2.4%) rates than the original SURFY algorithm and was therefore expected to deliver a more representative picture of the human surfaceome.

Especially its low false negative rate endowed SURFY2 with a high confidence to not miss surface-exposed proteins and thus not miss any opportunities for research and drug development. Importantly, SURFY2 exhibited improved metrics in comparison to SURFY at every stage of the model validation process, establishing the robustness of our newly developed algorithm (Table 1). The cross-validation-derived accuracy score of our SURFY implementation on the dataset consisting of training and validation set (0.9312) approximated the value of 0.933 reported by Bausch-Fluck et al., which referred to cross-validation over all of the training data. Therefore, we believe that the implementation of SURFY which we compared SURFY2 with presented a fair comparison, strengthening our claim of an improvement in predictive accuracy. Additionally, the fact that SURFY underperformed in prediction tests on separate datasets (both with validation as well as test set) while SURFY2 was not hindered by this issue hinted at the higher generalizability and increased robustness of SURFY2 in comparison to SURFY.

Table 1: Performance metrics of SURFY and SURFY2. Accuracy, precision, recall and area under the receiver operating characteristic (ROC) curve were assessed at different stages of building SURFY2 and compared to the corresponding metrics for SURFY trained on the same data. Cross-validation results represent the mean value from 5-fold cross-validation on the indicated dataset. The other columns indicate the prediction of holdout validation and test set samples, where especially the test set never came into contact with SURFY2 during the training process. Numbers in bold indicate the superior metric value in the comparison of SURFY and SURFY2.

Metric	**CrossValidation Training Set**		**Validation Set**		**CrossValidation Train+Val Set**		**Test Set**	
Algorithm	SURFY	SURFY2	SURFY	SURFY2	SURFY	SURFY2	SURFY	SURFY2
Accuracy	0.9251	**0.9456**	0.9149	**0.9574**	0.9312	**0.9433**	0.9236	**0.9554**
Precision	0.9314	**0.9538**	0.8648	**0.9167**	0.9355	**0.9562**	0.9091	**0.9425**
Recall	0.9447	**0.9461**	0.9697	**1**	**0.9479**	0.9467	0.9524	**0.9762**
Area under ROC curve	0.9796	**0.9820**	0.9182	**0.96**	0.9814	**0.9850**	0.9214	**0.9538**

Using SURFY2 to predict the *in silico* human surfaceome

After training SURFY2 on all labeled data (comprised of train, validation and test data), the surface-exposure of the remaining 6336 human transmembrane proteins was predicted (Supplementary Table 3). For the previously unlabeled proteins this resulted in 2854 predicted surface-exposed proteins and 3482 intracellular proteins. Together with the high-confidence labeled training data, these numbers were updated to 3764 surface proteins and 4139 intracellular proteins.

Comparing the list of proteins labeled by prediction with SURFY2 with the corresponding list of SURFY led to a number of discrepancies that were labeled differently by the two algorithms (Table 2, Supplementary Table 4). With a total of 114 discrepancies, 33 proteins were now newly labeled as surface proteins by SURFY2 and 81 proteins as intracellular proteins. Having a closer look at the class probabilities predicted by SURFY and SURFY2, some candidates hovered around the classification threshold of 0.5 for both classifiers. These included for instance human vesicular acetylcholine transporter (VACHT, SURFY score: 0.5928, SURFY2 score: 0.4785) or epithelial membrane protein 2 (EMP2, SURFY score: 0.5968, SURFY2 score: 0.4662). In those cases, the confidence in the predicted class for either classifier cannot be too strong for these proteins and they may contain features not captured in the data or the classifiers.

Table 2: Most confident classification discrepancies between SURFY and SURFY2. Previously unlabeled proteins not matching in their 'surface' feature predicted by SURFY or SURFY2 were collected and sorted by the SURFY2 classification score in descending order. For each class, the ten most confident predictions of SURFY2 were listed (the full list of discrepancies can be found in Supplementary Table 4). Predicted intracellular proteins are denoted by '0' and surface-exposed proteins by '1'.

Protein	*SURFY*	*SURFY Score*	*SURFY2*	*SURFY2 Score*
TRPM4_HUMAN	0	0.3713	1	0.8725
KCNT1_HUMAN	0	0.3693	1	0.8522
ABCG1_HUMAN	0	0.3553	1	0.8469
TRPV3_HUMAN	0	0.2954	1	0.8383
SIA4A_HUMAN	0	0.3713	1	0.8344
KCMA1_HUMAN	0	0.2914	1	0.8255
PDE3B_HUMAN	0	0.1896	1	0.7763
KCNB2_HUMAN	0	0.3593	1	0.7577
KCNJ6_HUMAN	0	0.3792	1	0.7125
TMM64_HUMAN	0	0.3413	1	0.7036
PIGO_HUMAN	1	0.6387	0	0.1579
CTSRG_HUMAN	1	0.7485	0	0.1533
RNFT1_HUMAN	1	0.6327	0	0.1531
NCTR1_HUMAN	1	0.6008	0	0.1488
LIRA6_HUMAN	1	0.6347	0	0.1467
KCT2_HUMAN	1	0.6387	0	0.1417
CASD1_HUMAN	1	0.6387	0	0.1366
TMED7_HUMAN	1	0.6407	0	0.1312
CGT_HUMAN	1	0.6008	0	0.1205
LMA2L_HUMAN	1	0.5988	0	0.1039

Yet there were also a number of cases where the class probabilities predicted by SURFY2 were very clear and therefore these proteins could be confidently placed in the respective class (Table 2). These discrepant high-confidence classifications include proteins such as the Ca^{2+}-activated cation channel protein TRPM4 (Transient receptor potential cation channel subfamily M member 4). In the original SURFY algorithm, human TRPM4 received a score of 0.3713 and was therefore classified as an intracellular protein. While this score was located quite close to the classification threshold of 0.5, the classification score given to TRPM4 by SURFY2 was 0.8725. Not only did this score re-label human TRPM4 as a surface transmembrane protein located in the plasma membrane but it did so with a considerably higher confidence than SURFY. And indeed, experimental evidence collected with confocal microscopy located TRPM4 in the plasma membrane (18), validating the classification decision of SURFY2.

Analogously, the E3 ubiquitin ligase RNFT1 was classified as a surface-exposed protein by SURFY with low confidence (score: 0.6327). SURFY2 however classified RNFT1 with high confidence as an intracellular protein (score: 0.1531). Being involved in the ER-associated degradation (ERAD) pathway (19), RNFT1 would have no use at the plasma membrane and has to be located in the ER. This ER localization (and the absence of plasma membrane localization) was indeed shown by immunofluorescence staining (19), increasing the amount of confidence that can be put into predictions made by SURFY2.

Analyzing the feature importances of the estimators inside SURFY2

In order to understand the superior performance of SURFY2 we further analyzed how it made predictions. For this, the feature importances used by the estimators making up the final voting classifier were crucial as they determined the predicted class probabilities from the data. The most crucial features for the original SURFY algorithm, which was based on a Random Forest algorithm, included number of glycosylation sites, transmembrane domain counts and lengths as well as non-cytoplasmic cysteine frequency (1).

In contrast, observing the top 10 features ranked in their relative importance for the Random Forest estimator in our voting classifier, our first observation was the prominence of engineered features among the top features used for classification (Fig. 2a). All ten out of ten features for our Random Forest estimator were engineered and stemmed from our meta-ensemble approach, emphasizing its relevance. The most important engineered features originated with ensemble and boosting approaches themselves. XGBClassifier, GradientBoosting as well as AdaBoost (boosting) and RandomForest as well as ExtraTrees (ensemble) base estimators were responsible for 80% of the most important features.

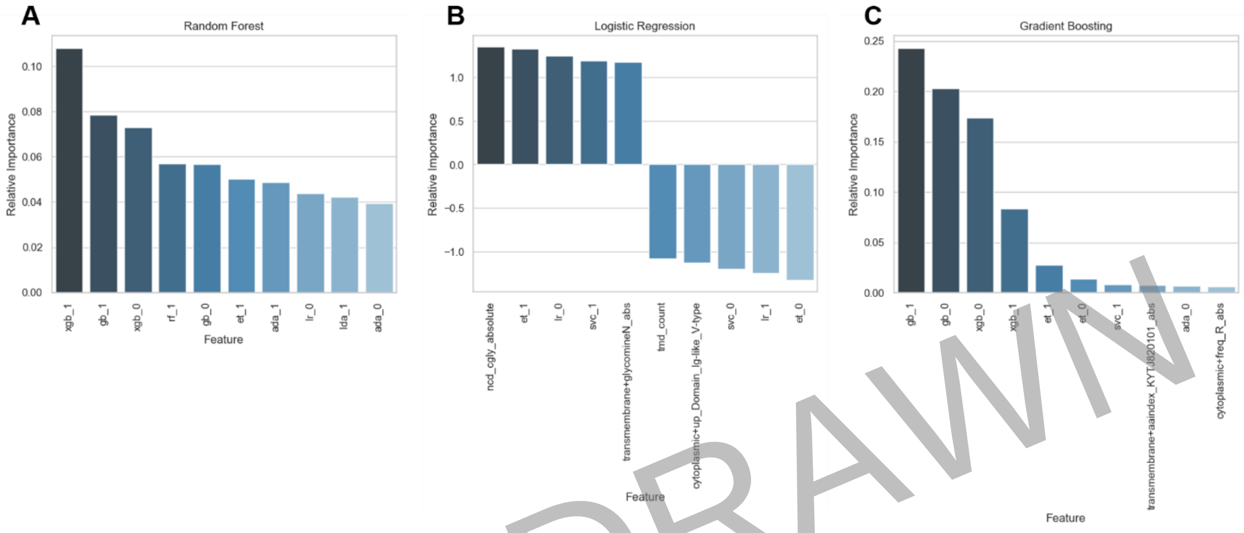

Figure 2: Feature importances for the classifiers used by the final voting classifier on the expanded training dataset. During training on the entire expanded training dataset, the feature importances were fitted and here the 10 most influential features are depicted for the RandomForest (**a**), LogisticRegression (**b**) and GradientBoosting (**c**) classifiers. For the RandomForest (**a**) and GradientBoosting (**c**) algorithms, this corresponds to the influence of these features in the construction of the base decision tree estimators. For the LogisticRegression (**b**) algorithm, the five features with the highest weight per class are shown.

Our logistic regression estimator however relied less on engineered features (only three of the five most highly-weighted features for both surface-exposed proteins and intracellular proteins). Both LogisticRegression as well as Support Vector Classification base estimators yielded highly weighted engineered features used in the prediction of both classes in the LogisticRegression classifier (Fig. 2b). While the C-glycosylation feature was also relevant for SURFY, the feature indicating predicted N-glycosylation by GlycoMine (20) in the transmembrane region was unique to SURFY2 for its prominence. Additionally, one of the most highly-weighted logistic regression features for the class of intracellular transmembrane proteins relied on Ig-like domains. This feature was also not prominently represented in SURFY and may help to explain the success of SURFY2.

Finally, while the Gradient Boosting estimator of our logistic regression mostly relied on the engineered features stemming from the XGBClassifier and GradientBoosting base estimators, its top 10 also contained features which were initially present in the dataset (Fig. 2c). Features such as the hydropathy index (AAIndex KYTJ820101, also relevant in SURFY) as well as the frequency of arginine residues in the cytoplasmic regions were present among the top features influencing the predictions of the Gradient Boosting estimator. As at least the latter feature was also not strongly represented in the original SURFY, this could further serve as part of the explanation of the performance of SURFY2.

As engineered features stemming from the XGBClassifier and GradientBoosting base estimators had the greatest influence in the upper layers of our meta-ensemble classifier, we then proceeded to analyze their respective feature importances to further understand the behavior of SURFY2 (Fig. 3). While six out of ten features were shared between the most relevant features for XGBClassifier and GradientBoosting base estimators, their relative importance differed between base estimators. From the ten most relevant features used in the original SURFY algorithm, three were present in the case of XGBClassifier and five in the case of the GradientBoosting base estimator, though again at different relative positions. Features such as the non-cytoplasmic frequency of threonine or the frequency of valine in transmembrane domains were unique to the base estimators investigated here and were not identified among the ten most influential features for SURFY. Utilization of a diverse set of features likely opened up the potential to register meaningful variance in more variables and thus potentially yielded a higher degree of generalizability.

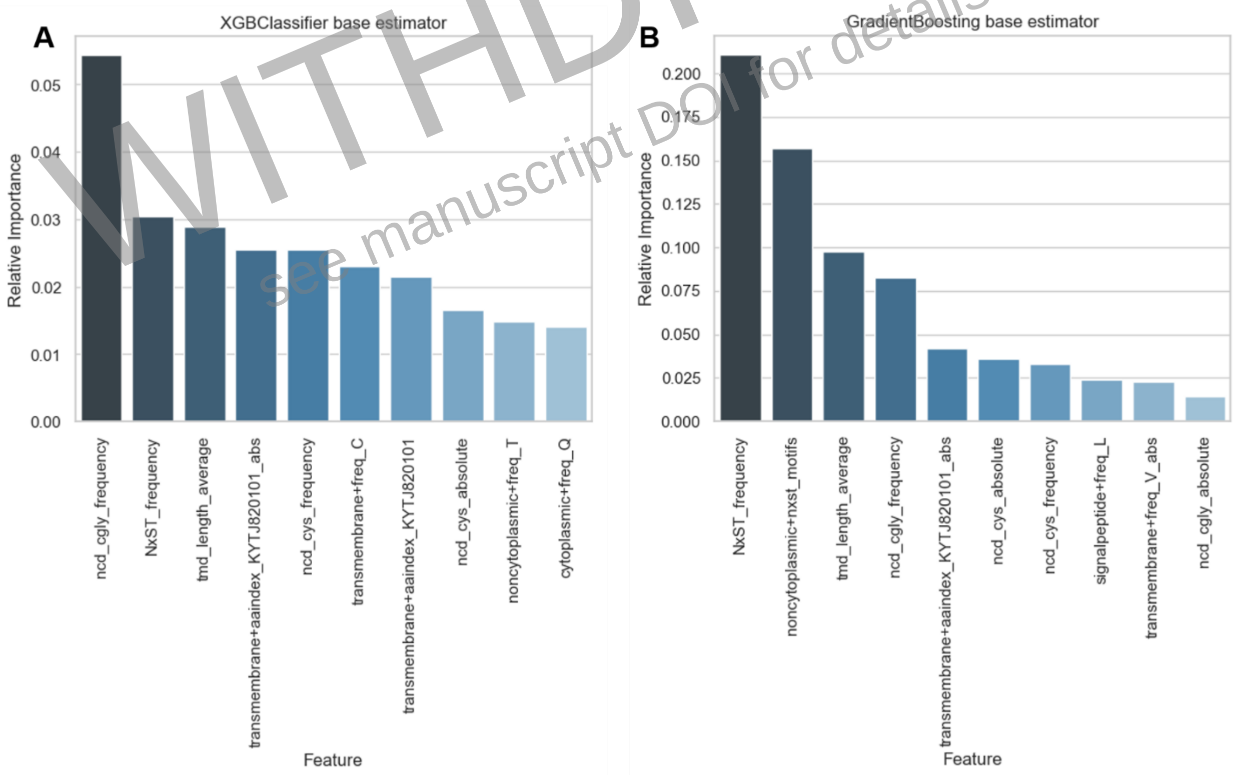

Figure 3: Feature importances for the most relevant base estimators of the meta-ensemble classifier. The XGBClassifier (**a**) and GradientBoosting (**b**) base estimators were fitted on the complete training dataset (comprising training, validation and test data). Afterwards, feature importances were extracted and the ten most influential features are shown, respectively.

Rationalizing discrepant classifications made by SURFY and SURFY2

With the gained understanding of the features important for classification in SURFY and SURFY2, we analyzed the discrepant classifications of the two classifiers to understand and evaluate why they were different. Splitting the whole dataset in two groups (‘shared predictions’ and ‘discrepant predictions’), we calculated the mean values for each feature and predicted class in both datasets. Then, we singled out the features with the largest differences (between datasets) in the mean values of predicted intracellular and surface-exposed proteins. This was done to evaluate whether the discrepant predictions were due to systematically different sample characteristics between the two datasets. The whole set of mean values of features which class mean difference changed by more than 0.5 between prediction datasets, grouped by prediction state and class label, can be found in Supplementary Table 5 and Supplementary Table 6.

As shown in Fig. 4, some features exhibited dramatically different mean values for the two classes in the discrepant prediction dataset compared to the shared prediction dataset. While the standard deviation seemed substantial, the large sample sizes resulted in statistically highly significant mean value differences between classes for both prediction datasets (using Welch's unequal variances t-test (21) with the Holm–Šidák method for correction after multiple testing (22); Supplementary Table 7). More important than the absolute mean values of the features in the classes were the ratios between classes. While for instance the presence of C-glycosylation sites in the shared prediction dataset was highly indicative of a surface-exposed protein, samples in the discrepant prediction dataset did not seem to exhibit C-glycosylation sites regardless of their surface-state. Analogously, shared predictions of surface proteins showcased an elevated count of N-glycosylation sites (N-X-S/T) compared to intracellular transmembrane proteins. In the set of discrepant predictions this trend was actually reversed, albeit less pronounced.

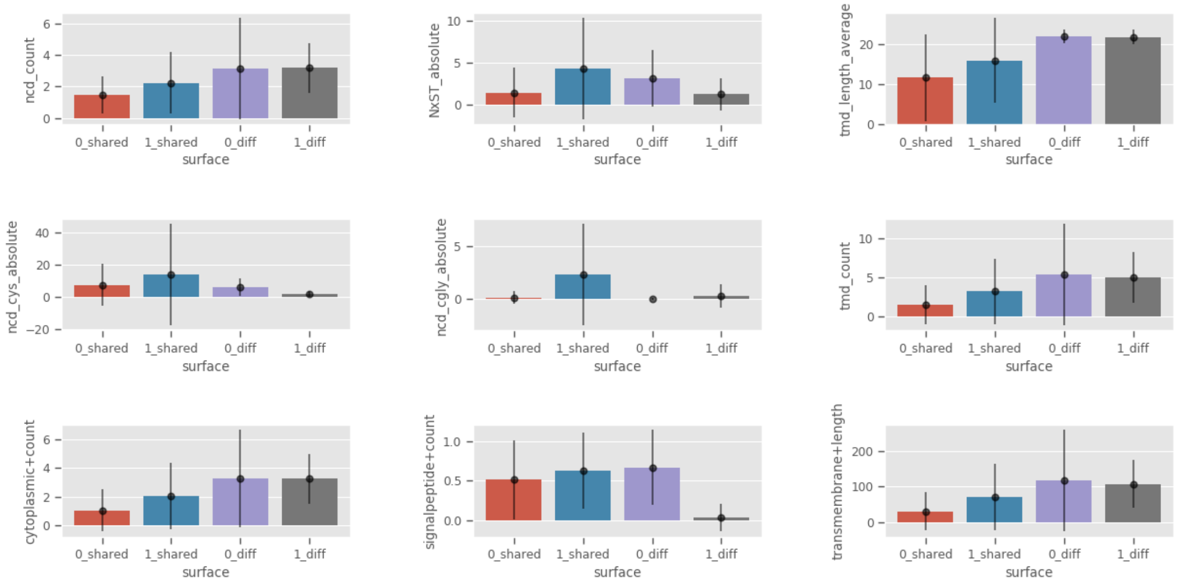

Figure 4: Mean values of selected features grouped by prediction state and class label. After the unlabeled samples were classified with SURFY2, discrepant classifications with regard to the classification performed with SURFY were identified. The dataset was then separated into shared predictions (‘shared’) and discrepant predictions (‘diff’) and the mean value of each feature

for each class was calculated in both sets. After filtering for features with mean differences differing by more than 0.5 between shared and discrepant predictions, some representative features were plotted. The whole set of filtered mean values grouped by prediction state and class label can be found in Supplementary Table 4 and Supplementary Table 5. Features are shown as means ± standard deviation.

Among the features whose ratios were the most distorted were the most highly-weighted features used by SURFY for prediction, such as glycosylation counts or transmembrane domain counts and lengths. Taking into account the inversed ratio between surface and intracellular proteins exhibited by these samples elucidated the reasons why the initial SURFY algorithm relying on these exact same features evidently misclassified them. SURFY2 however was more robust thanks to its composition of many different estimators which all relied on different features. This made it possible for SURFY2 to fit the data better and make more accurate predictions. Even for the highly unusual proteins in the discrepant prediction dataset, SURFY2 was able to make predictions with high class probabilities for many of them.

To systematically analyze the differences between the two classifiers, we used t-distributed stochastic neighbor embedding (t-SNE (23)) as a nonlinear dimensionality reduction method to visualize the proteins in the dataset in two dimensions. We performed t-SNE for all samples for which predictions from both SURFY and SURFY2 were available (Fig. 5a). The first observation that could be made was that this two-dimensional representation already visually separated a substantial portion of the intracellular transmembrane proteins from the surface-exposed proteins. Additionally, proteins in the category ‘Surface_diff’, indicating proteins now newly classified as members of the surfaceome by SURFY2, were mostly located at the edge of the purely intracellular realm and the predominantly surface-exposed areas. These potential ambiguities in their characteristics, as already discussed in the context of feature mean values (Fig. 4) could have led to the discrepancies in classification.

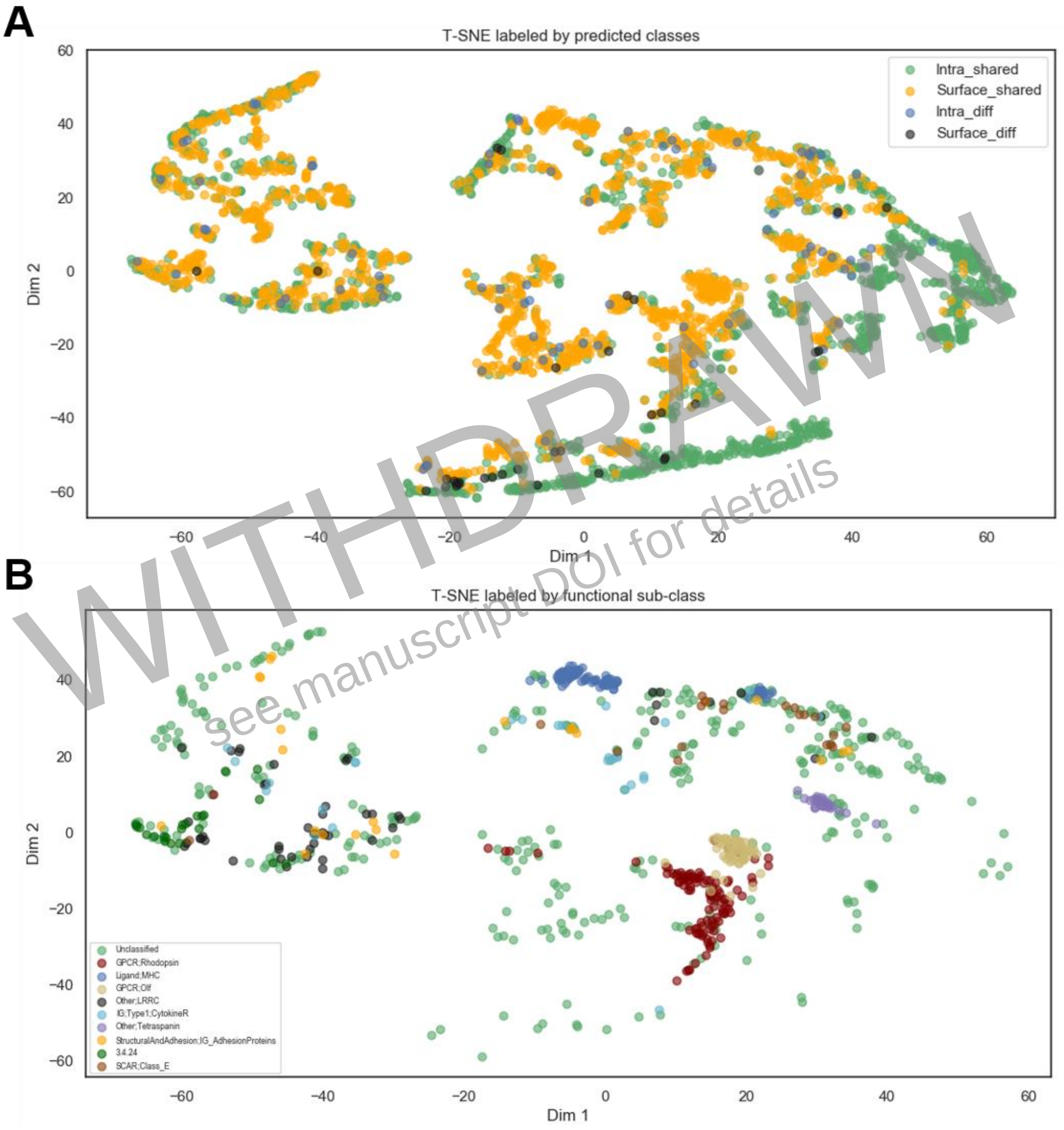

Figure 5: Functional analysis of human transmembrane proteins with t-Distributed Stochastic Neighbor Embedding (t-SNE). Samples for which predictions of both SURFY and SURFY2 were available were used for dimensionality reduction using t-SNE. The samples are shown as the two components received by the nonlinear dimensionality reduction. (**a**) Visualization of shared and discrepant predictions of SURFY and SURFY2. Proteins which 'surface' feature was predicted were colored according to four classes: proteins predicted to be intracellular by both SURFY and SURFY2 (Intra_shared), proteins predicted to be surface-exposed by both SURFY and SURFY2 (Surface_shared), proteins predicted to be intracellular by SURFY2 but not by SURFY (Intra_diff) and proteins predicted to be surface-exposed by SURFY2 but not by SURFY (Surface_diff). (**b**) Visualization of functional annotations of human transmembrane proteins. Surface-exposed proteins of the ten most prevalent functional annotations derived from Almén et al. (23) are shown as their t-SNE representation and colored according to their functional annotation.

Comparing the prediction class distribution in Fig. 5a with the corresponding distribution of the most prevalent functional annotations of surface-exposed proteins derived from Almén et al. (23) in Fig. 5b also led to some interesting observations. Dense clusters of functional protein families, such as olfactory GPCRs or MHCs, seemed to be depleted of discrepant predictions in either direction. This could stem from the fact that these families are thoroughly investigated and their characteristics were well represented in the training data, leading to a relatively straightforward classification. On the other hand, regions characterized by surface proteins of unclassified function or devoid of prevalent functional clusters were seemingly enriched in discrepant predictions. Analogously, this could reflect the fact that these types of human transmembrane proteins are not as thoroughly investigated or are not members of large families with shared characteristics and were therefore not adequately represented in the training data.

Investigating the most confident re-classifications of transmembrane proteins by SURFY2

The importance of these newly classified human transmembrane proteins by the optimized SURFY2 algorithm could be demonstrated by the analysis of disease-relevant proteins among high-confidence discrepant predictions (Table 2). We therefore analyzed two case studies involving previously misclassified disease-relevant proteins in depth and explored why the two classifiers differed in their predictions.

Potassium channel subfamily T member 1 (KCNT1) was predicted to be an intracellular transmembrane protein by SURFY (score: 0.3693) and was predicted by SURFY2, with high confidence (score: 0.8522), to be a surface-exposed protein localized in the plasma membrane. KCNT1 has been predicted to be plasma membrane-localized by the database COMPARTMENTS with the highest degree of confidence and, if expressed in CHO cells, has been demonstrated to reside in the plasma membrane by immunocytochemistry (24), validating the prediction of SURFY2. Exhibiting extremely low non-cytoplasmic cysteine counts (2 counts), the third most important feature for SURFY, compared to the average surface-exposed protein (13.8; derived from SURFY2 predictions) might explain the wrong classification by SURFY. Additionally, the absence of non-cytoplasmic C-glycosylation, unusual for a surface-exposed protein in this dataset on average, made KCNT1 possibly harder to correctly classify. KCNT1 has been implicated as the primary causal factor in early infantile epileptic encephalopathy-14 (EIEE14) when mutated (25,26) and is therefore highly disease-relevant. Reporting its correct subcellular localization and including it in the updated human surfaceome is therefore of considerable relevance and may help further investigations.

Another example would be Transient receptor potential cation channel subfamily V member 3 (TRPV3), the cause of the skin disease Olmsted syndrome when mutated (27). Incidentally, a different mutation of TRPV3 also is responsible for focal non-epidermolytic palmoplantar keratoderma 2

(FNEPPK2) (28). While being predicted to be an intracellular transmembrane protein by SURFY (score: 0.2954), SURFY2 predicted TRPV3 to be a surface-exposed transmembrane protein (score: 0.8383). Even though experimental subcellular localization data is still sparse for TRPV3, GFP-tagged TRPV3 expressed in a keratinocyte cell line has been shown to localize at the plasma membrane (29), substantially bolstering the prediction claim made by SURFY2. Here, the discrepant predictions could be explained by the absence of N-X-S/T glycosylation sites, the most relevant feature in SURFY, as well as the low non-cytoplasmic cysteine counts and absence of non-cytoplasmic C-glycosylation already encountered in the case of KCNT1.

Next to these case studies, the ten most confident discrepant predictions for each class alone harbored several more disease-relevant proteins. These included LMA2L (autosomal recessive mental retardation 52, MRT52, (30)), PIGO (Hyperphosphatasia with mental retardation syndrome 2, HPMRS2, (31)), KCNJ6 (Keppen-Lubinsky syndrome, KPLBS, (32)) and KCMA1 (Paroxysmal nonkinesigenic dyskinesia, 3, with or without generalized epilepsy, PNKD3, (33)). Inclusion (or exclusion in the cases of LMA2L and PIGO) from the human surfaceome will offer researchers a clearer picture and could substantially contribute toward research on their respective diseases.

Discussion

Since the work of Günter Blobel which earned him his Nobel Prize (34,35), protein localization in the cell has been an important topic spawning numerous areas of interest in research and applications alike. Here, we present the most accurate prediction of plasma membrane localization of human transmembrane proteins up to date. Using a meta-ensemble machine learning approach with SURFY2, we achieved a predictive accuracy of over 95% with low false negative and false positive rates and improved on all preceding prediction methods. This led to the formulation of the most reliable and comprehensive account of the human surfaceome yet which will accelerate research on proteins contained therein and potentially enable the development of new drugs. While the number of re-classified proteins is substantially smaller than the total number of human transmembrane proteins, the high expenses related to experimental research on transmembrane proteins and the high potential of waste in case of incorrect information fully justifies this endeavor to improve the available information.

Our nearest point of comparison is of course SURFY as the hitherto best prediction tool for the human surfaceome. As already stated, we achieved a clear increase in accuracy, sensitivity and specificity with our method compared to SURFY. Thanks to the diversity of our base estimators, we were able to utilize features for prediction which were not captured by the Random Forest approach used by SURFY.

This behavior became abundantly clear in the cases of discrepant classification, as these proteins presented unusual tendencies in the very features SURFY relied the most on for prediction.

Additionally, the performance estimates (including the accuracy score of 93.5%) reported by SURFY are based on cross-validation and out-of-bag estimates. Both methods have a tendency to be overly optimistic with regard to the performance of the classifier on unseen data (36). In an out-of-bag estimate, even though single decision trees are trained on a subset of the training data, the whole forest is eventually established through the entire training data. Ideally, and as done here, an external test dataset unseen by the classifier should be used to assess the performance of the final model. Therefore, the true accuracy estimates of SURFY are likely to be lower than those obtained by cross-validation and out-of-bag estimates. This is reflected in the performance estimates of the SURFY approximation reported in Table 1 and also emphasizes the strong improvement of more than 3% achieved by SURFY2 on a held back test set over previous methods in terms of accuracy score.

We believe that the predictions made by SURFY2 will fuel further research on transmembrane proteins. The most obvious route for this is of course the experimental validation of the subcellular localization of proteins predicted by SURFY2 for which no experimental data exists yet. This is especially interesting for discrepant predictions between SURFY and SURFY2. Even though we showed several examples in which the scientific literature supported the predictions made by SURFY2 over those previously made by SURFY, further experimental validation may bolster the confidence placed in SURFY2 predictions.

Cases of particular importance are proteins such as natural cytotoxicity triggering receptor 1 (NCTR1), which is classified as a plasma membrane protein by SURFY (score: 0.6008) but as an intracellular protein by SURFY2 (score: 0.1488). As an important natural killer cell receptor protein (37), NCTR1 has a high prior probability of being located on the cell surface. Yet a high-confidence prediction by SURFY2 could imply the interesting scenario in which NCTR1 is in fact not located at the cell surface which requires experimental validation.

Another curious observation is the fact that potassium ion channels are enriched in the high-confidence discrepant predictions, specifically in those predictions which were predicted as intracellular by SURFY and as surface-exposed by SURFY2. Nearly half of these high-confidence discrepant predictions (KCNT1, KCMA1, KCNB2 and KCNJ6) were identified as potassium channels, implying the exciting possibility that these transmembrane proteins share structural characteristics which made them so hard to classify for SURFY. Further research on these potassium channels, some of which are disease-relevant, might elucidate these properties.

Intriguingly, the set of transmembrane proteins newly classified as surface-exposed proteins here does not seem to possess a classic signal peptide (Fig. 4, mean value of 0.0303 compared to 0.6244 for surface proteins in the shared prediction set). As signal peptides are required for both intracellular as well as surface-exposed transmembrane proteins, this could either mean that the sequence of the signal peptide occurring in these cases is too far from the consensus sequence to be detectable or that they arrive at the plasma membrane through a nonconventional path. Both alternatives seem to be fascinating and require further research.

Given the rapid advances in machine learning techniques, a future construction of SURFY3 with even better performance seems likely. While the performance metrics reported here are already very high, further improvements might deliver an even more accurate depiction of the human surfaceome and might re-classify some of the human transmembrane proteins in the dataset. In addition to algorithmic improvements, further experimental evidence will bolster the dataset used for training and thereby might also contribute to a more accurate classifier.

If the same or similar features are measured and catalogued for non-human transmembrane proteins, a similar approach such as SURFY2 might enable the prediction of the respective species surfaceome. This might even uncover some general principles across species for the distribution of transmembrane proteins among different cellular membranes. Not only would this be an advance for basic research but it would also offer design advice to synthetic biologists to engineer proteins such as artificial receptors (38) for an exactly specified subcellular localization. Analogously, combining different classification modules might even enable an accurate prediction system for all kinds of subcellular localizations, in humans as well as in other species.

We envision the updated *in silico* human surfaceome delivered by the predictions of our meta-ensemble classifier SURFY2 to serve as a useful repository for researchers to refer to and to assist them in their research. Additionally, the questions and hypotheses stemming from these predictions, some of which were discussed here, offer a direct path forward to experimentally profit from this database. Most of all, we see the potential of SURFY2 for disease-relevant transmembrane proteins. Every piece of information regarding these proteins has the strong potential to further the efforts to master the corresponding diseases and help those in need.

Methods

SURFY2 dataset

In order to train the classifier developed here and predict the class of the unlabeled proteins, the same dataset as already described in the study of Bausch-Fluck et al. in Supplementary Table S11.4 was used (1). The features included in the original dataset are also described in the mentioned study. Briefly summarized, the positive training dataset was built by relying on a subset of high confidence entries of the CSPA (6), entries in the subcellular localization database COMPARTMENTS (39) as well as UniProtKB/Swiss-Prot (11) entries containing the keyword “Cell membrane”. Analogously, the negative training dataset was constructed from the entries of the negative benchmark set for “plasma membrane” in the COMPARTMENTS database which were not present in the CSPA. These sets were further redundancy-reduced in the original study by clustering proteins according to UniRef50 (40) and only retaining one representative member per cluster. In total, this resulted in a training dataset of 910 positive samples and 657 negative samples.

Processing data for building a classifier

All data handling, prediction and analysis was done with Python 3.7.1 (41). After extracting the training data samples with all their features from the complete dataset used in the study of Bausch-Fluck et al. (1), we constructed a new feature called ‘surface’. This binary feature simply indicated whether the sample was labeled as intracellular (0) or surface-exposed (1) in the training dataset and was used as the target feature for classification. Subsequently, randomly chosen 10% of the training dataset were split off into a separate test set. An additional randomly chosen 10% were then split off the remaining training dataset into a separate validation set. Prior to training classifiers, training and validation datasets were split into a feature matrix and a column vector containing the target variable. Then, the features ‘accession’ and ‘name’ were excluded from both feature matrixes to avoid biasing the classifier.

Building and optimizing base estimators

Supervised, binary classification was done mostly using classifiers from the scikit-learn library (version 0.20.1, (42)). The only exception from this occurred with the XGBClassifier from the XGBoost python module (version 0.81, (43)). The other classifiers used as base estimators were all implemented in scikit-learn and included RandomForestClassifier, LogisticRegression, GradientBoostingClassifier, ExtraTreesClassifier, GaussianNB, SVC, KNeighborsClassifier, AdaBoostClassifier, MLPClassifier, LinearDiscriminantAnalysis and QuadraticDiscriminantAnalysis. Hyperparameters for each classifier were optimized using GridSearchCV (scikit-learn) with the scoring function ‘balanced_accuracy’, using a 5-fold cross-validation scheme. In each round of cross-validation, the training dataset was split into 80% used for training the respective classifier with the respective hyperparameters while the ‘surface’ label of the remaining 20% was predicted and used to assess the performance of the scoring according to the scoring function.

Building a soft voting classifier

To construct a voting classifier, the VotingClassifier module from scikit-learn was used with soft voting, in which the class probabilities are summed according to their weights and the class with maximum probability is chosen. As the exhaustive search over all combinations of the 12 previously optimized base estimators would be computationally very expensive, a greedy algorithm was applied. Starting from the best single estimator, the GradientBoostingClassifier, all combinations of two estimators (with one estimator being the GradientBoostingClassifier) were tested and evaluated by 5-fold cross-validation on the training set using 'balanced_accuracy' as a scoring function. The best combination, now including the RandomForestClassifier, was then tested with all remaining estimators to find the best voting classifier that utilized three estimators. This resulted in the final voting classifier, using GradientBoostingClassifier, RandomForestClassifier and LogisticRegression as its estimators.

Subsequently, we optimized the weights the individual estimators were assigned in the soft voting classifier. For this, we exhaustively sampled all weight combinations for all three estimators (excluding redundant combinations where all estimators had identical weights), ranging from weights of 1 to 3. The performance of these different soft voting classifiers was evaluated by 5-fold cross-validation on the training set using accuracy as a scoring function. The weight combination leading to maximum accuracy was chosen as the final voting classifier.

Feature engineering with base estimators & model selection

To construct a meta-ensemble classifier, we used the StackingCVClassifier module from the mlxtend Python repository (44) to create new features from the predicted class probabilities for both classes made by the 12 base estimators. The predictions were appended to the original training dataset as 24 new features. This new expanded dataset was then used to train the estimators of the soft voting classifier and the weighted vote of their predicted class probabilities resulted in the final predicted class.

All base estimators and the final meta-ensemble classifier were trained both on the original training set as well as the expanded training set containing the novel features and evaluated to select the best model. The classifier with the highest 5-fold cross-validation accuracy was chosen as the final model and evaluated on the separate validation as well as test dataset. To compare it with the original SURFY, the scikit-learn implementation of the RandomForestClassifier using the Gini criterion and constituting 501 decision trees as base estimators was used in order to achieve a fair comparison and tested on the same data as SURFY2.

Visualization of human transmembrane protein properties using t-SNE

To visualize the distribution of human transmembrane proteins in two dimensions, the nonlinear dimensionality reduction method t-distributed stochastic neighbor embedding (t-SNE) was used. All initial features of the dataset, excluding UniProt accession code, UniProt name and the ‘surface’ label and score, were used for dimensionality reduction for all samples for which a prediction of both SURFY and SURFY2 existed to facilitate comparison. Using the scikit-learn implementation of t-SNE (TSNE), a perplexity of 40 and 1500 iterations were used to construct the first two components which were used for visualization. For visualization, as for most plots in this work, the Python library Matplotlib was used (45).

The functional sub-classes used in Fig. 5b were derived from Almén et al. (23), as already reported by Bausch-Fluck et al. (1) and paired with the corresponding data entries that were depicted in Fig. 5b. The most frequently observed functional annotations were then used as a color label.

Supplementary Information

Supplementary Tables: Description of training data, model selection, classified transmembrane proteins, discrepant predictions with SURFY, divergent feature means of both prediction sets and statistical analysis of feature means divided by classes.

Acknowledgment

The author would like to thank Christian Bock and Matteo Togninalli for their helpful feedback.

Conflicts of Interest

The author declares no conflict of interest.

Data Availability

All data and code relevant and used for this study are available at https://github.com/Bribak/SURFY2.

References

1. Bausch-Fluck D, Goldmann U, Müller S, van Oostrum M, Müller M, Schubert OT, et al. The in silico human surfaceome. Proc Natl Acad Sci. 2018 Nov 13;115(46):E10988–97.

2. Pollock SB, Hu A, Mou Y, Martinko AJ, Julien O, Hornsby M, et al. Highly multiplexed and quantitative cell-surface protein profiling using genetically barcoded antibodies. Proc Natl Acad Sci. 2018 Mar 13;115(11):2836–41.

3. Yin H, Flynn AD. Drugging Membrane Protein Interactions. Annu Rev Biomed Eng. 2016 Jul 11;18(1):51–76.

4. Savojardo C, Martelli PL, Fariselli P, Profiti G, Casadio R. BUSCA: an integrative web server to predict subcellular localization of proteins. Nucleic Acids Res. 2018 Jul 2;46(W1):W459–66.

5. Yoon Y, Lee GG. Subcellular Localization Prediction through Boosting Association Rules. IEEE/ACM Trans Comput Biol Bioinform. 2012 Mar;9(2):609–18.

6. Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, et al. A Mass Spectrometric-Derived Cell Surface Protein Atlas. PLOS ONE. 2015 Apr 20;10(4):e0121314.

7. Sommer C, Gerlich DW. Machine learning in cell biology – teaching computers to recognize phenotypes. J Cell Sci. 2013 Dec 15;126(24):5529–39.

8. Affonso C, Rossi ALD, Vieira FHA, de Carvalho ACP de LF. Deep learning for biological image classification. Expert Syst Appl. 2017 Nov;85:114–22.

9. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S. Machine Learning and Its Applications to Biology. PLoS Comput Biol. 2007;3(6):e116.

10. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2007 Dec 23;36(Database):D202–5.

11. The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015 Jan 28;43(D1):D204–12.

12. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32.

13. Genuer R. Variance reduction in purely random forests. J Nonparametric Stat. 2012 Sep;24(3):543–62.

14. Xiaoyan Mu, Watta P, Hassoun MH. Analysis of a plurality voting-based combination of classifiers. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Hong Kong, China: IEEE; 2008. p. 304–9.

15. Džeroski S, Ženko B. Is Combining Classifiers with Stacking Better than Selecting the Best One? Mach Learn. 2004 Mar;54(3):255–73.

16. Wang S, Chen H, Yao X. Negative correlation learning for classification ensembles. The 2010 International Joint Conference on Neural Networks (IJCNN). Barcelona, Spain: IEEE; 2010. p. 1–8.

17. Zanda M, Brown G, Fumera G, Roli F. Ensemble Learning in Linearly Combined Classifiers Via Negative Correlation. Multiple Classifier Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 440–9.

18. Xu X-ZS, Moebius F, Gill DL, Montell C. Regulation of melastatin, a TRP-related protein, through interaction with a cytoplasmic isoform. Proc Natl Acad Sci. 2001 Sep 11;98(19):10692–7.

19. Kaneko M, Iwase I, Yamasaki Y, Takai T, Wu Y, Kanemoto S, et al. Genome-wide identification and gene expression profiling of ubiquitin ligases for endoplasmic reticulum protein degradation. Sci Rep. 2016 Nov;6(1).

20. Li F, Li C, Wang M, Webb GI, Zhang Y, Whisstock JC, et al. GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. Bioinformatics. 2015 May 1;31(9):1411–9.

21. Welch BL. The generalization of 'Student's' problem when several different population variances are involved. Biometrika. 1947;34(1–2):28–35.

22. Sidak Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. J Am Stat Assoc. 1967 Jun;62(318):626.

23. Almén M, Nordström KJ, Fredriksson R, Schiöth HB. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biol. 2009;7(1):50.

24. Evely KM, Pryce KD, Bhattacharjee A. The Phe932Ile mutation in KCNT1 channels associated with severe epilepsy, delayed myelination and leukoencephalopathy produces a loss-of-function channel phenotype. Neuroscience. 2017 May;351:65–70.

25. Barcia G, Fleming MR, Deligniere A, Gazula V-R, Brown MR, Langouet M, et al. De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. Nat Genet. 2012 Nov;44(11):1255–9.

26. Ishii A, Shioda M, Okumura A, Kidokoro H, Sakauchi M, Shimada S, et al. A recurrent KCNT1 mutation in two sporadic cases with malignant migrating partial seizures in infancy. Gene. 2013 Dec;531(2):467–71.

27. Lin Z, Chen Q, Lee M, Cao X, Zhang J, Ma D, et al. Exome Sequencing Reveals Mutations in TRPV3 as a Cause of Olmsted Syndrome. Am J Hum Genet. 2012 Mar;90(3):558–64.

28. He Y, Zeng K, Zhang X, Chen Q, Wu J, Li H, et al. A Gain-of-Function Mutation in TRPV3 Causes Focal Palmoplantar Keratoderma in a Chinese Family. J Invest Dermatol. 2015 Mar;135(3):907–9.

29. Yadav M, Goswami C. TRPV3 mutants causing *Olmsted Syndrome* induce impaired cell adhesion and nonfunctional lysosomes. Channels. 2017 May 4;11(3):196–208.

30. Rafiullah R, Aslamkhan M, Paramasivam N, Thiel C, Mustafa G, Wiemann S, et al. Homozygous missense mutation in the *LMAN2L* gene segregates with intellectual disability in a large consanguineous Pakistani family. J Med Genet. 2016 Feb;53(2):138–44.

31. Krawitz PM, Murakami Y, Hecht J, Krüger U, Holder SE, Mortier GR, et al. Mutations in PIGO, a Member of the GPI-Anchor-Synthesis Pathway, Cause Hyperphosphatasia with Mental Retardation. Am J Hum Genet. 2012 Jul;91(1):146–51.

32. Masotti A, Uva P, Davis-Keppen L, Basel-Vanagaite L, Cohen L, Pisaneschi E, et al. Keppen-Lubinsky Syndrome Is Caused by Mutations in the Inwardly Rectifying K+ Channel Encoded by KCNJ6. Am J Hum Genet. 2015 Feb;96(2):295–300.

33. Du W, Bautista JF, Yang H, Diez-Sampedro A, You S-A, Wang L, et al. Calcium-sensitive potassium channelopathy in human epilepsy and paroxysmal movement disorder. Nat Genet. 2005 Jul;37(7):733–8.

34. Blobel G, Dobberstein B. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J Cell Biol. 1975 Dec;67(3):835–51.

35. Short B. The signal hypothesis matures with age. J Cell Biol. 2017 May 1;216(5):1207–1207.

36. Bylander T. Estimating Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates. Mach Learn. 2002 Jul;48:287–97.

37. Glasner A, Ghadially H, Gur C, Stanietsky N, Tsukerman P, Enk J, et al. Recognition and Prevention of Tumor Metastasis by the NK Receptor NKp46/NCR1. J Immunol. 2012 Mar 15;188(6):2509–15.

38. Scheller L, Strittmatter T, Fuchs D, Bojar D, Fussenegger M. Generalized extracellular molecule sensor platform for programming cellular behavior. Nat Chem Biol. 2018 Jul;14(7):723–9.

39. Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database. 2014 Feb 25;2014(0):bau012–bau012.

40. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007 May 15;23(10):1282–8.

41. van Rossum G. Python tutorial, Technical Report CS-R9526. Cent Voor Wiskd En Inform CWI Amst. 1995 May;

42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

43. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. San Francisco, California, USA: ACM Press; 2016. p. 785–94.

44. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw. 2018 Apr 22;3(24):638.

45. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9(3):90–5.