*Published online 2 December 2022*

*NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 4* **1**
*https://doi.org/10.1093/nargab/lqac091*

# Predictive modeling of moonlighting DNA-binding proteins

**Dana Mary Varghese[1], Ruth Nussinov[2,3] and Shandar Ahmad[1,*]**

[1]School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi- 110067, India, [2]Computational Structural Biology Section, Cancer Innovation Laboratory, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA and [3]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Israel

Received June 06, 2022; Revised October 25, 2022; Editorial Decision November 01, 2022; Accepted November 11, 2022

## ABSTRACT

**Moonlighting proteins are multifunctional, single-polypeptide chains capable of performing multiple autonomous functions. Most moonlighting proteins have been discovered through work unrelated to their multifunctionality. We believe that prediction of moonlighting proteins from first principles, that is, using sequence, predicted structure, evolutionary profiles, and global gene expression profiles, for only one functional class of proteins in a single organism at a time will significantly advance our understanding of multifunctional proteins. In this work, we investigated human moonlighting DNA-binding proteins (mDBPs) in terms of properties that distinguish them from other (non-moonlighting) proteins with the same DNA-binding protein (DBP) function. Following a careful and comprehensive analysis of discriminatory features, a machine learning model was developed to assess the predictability of mDBPs from other DBPs (oDBPs). We observed that mDBPs can be discriminated from oDBPs with high accuracy of 74% AUC of ROC using these first principles features. A number of novel predicted mDBPs were found to have literature support for their being moonlighting and others are proposed as candidates, for which the moonlighting function is currently unknown. We believe that this work will help in deciphering and annotating novel moonlighting DBPs and scale up other functions. The source codes and data sets used for this work are freely available at https://zenodo.org/record/7299265#.Y2pO3ctBxPY**

## INTRODUCTION

The human genome encodes close to 20,000 protein-coding genes, which is clearly a small number when compared with the complexity and diversity of biological functions (1). Several combinatorial and versatile approaches are employed by proteins to perform the huge number of molecular functions involved in the human lifecycle. One such mechanism, called moonlighting has drawn the attention of the community (2–5). Moonlighting proteins are a subset of multitasking proteins. However, unlike many other multitasking proteins, the moonlighting function is not encoded in different functional or structural domains. They switch between their moonlighting functions by partnering with different sets of substrates, using different oligomerization states, post-translational modifications or as a response to changes in the physio-chemical environment or subcellular localization (4,6–9). Different micro-mechanisms adopted by moonlighting proteins to perform multiple functions are illustrated in Figure 1. Moonlighting functions of the same protein might be spatially or temporally differentiated, i.e. different moonlighting functions may be associated with different cellular localization of DBPs or they may perform their moonlighting functions under different cellular conditions at different time points. Indeed change in cellular localization is well-known to be a driving force for moonlighting (10). To accomplish this functional pattern in a cell, the levels, locations, and contexts of gene expression are controlled. For example, crystallins are produced in a variety of tissues, where they perform their canonical function. In some vertebrates, they have been shown to express at a higher level in the eye lens, induced by some transcription factors where they serve an additional moonlighting function as a structural protein (8,11–13). Heat shock protein Hsp90 has been identified as a secreted, cell surface, and nuclear protein. Whereas normally molecular chaperons perform their canonical cytoplasmic functions assisting protein folding (14), and inside the nucleus regulating nuclear functions (15), when secreted, Hsp90 has pro-tumorigenic properties (16,17). Most such moonlighting functions are organism specific, emphasizing the usefulness of a predictive model that focuses on a single species.

The discovery of moonlighting proteins has mostly been serendipitous. Researchers have focused on literature mining, Gene ontology (GO) mining, and other

*To whom correspondence should be addressed. Tel: +91 11 8788; Email: shandar@jnu.ac.in

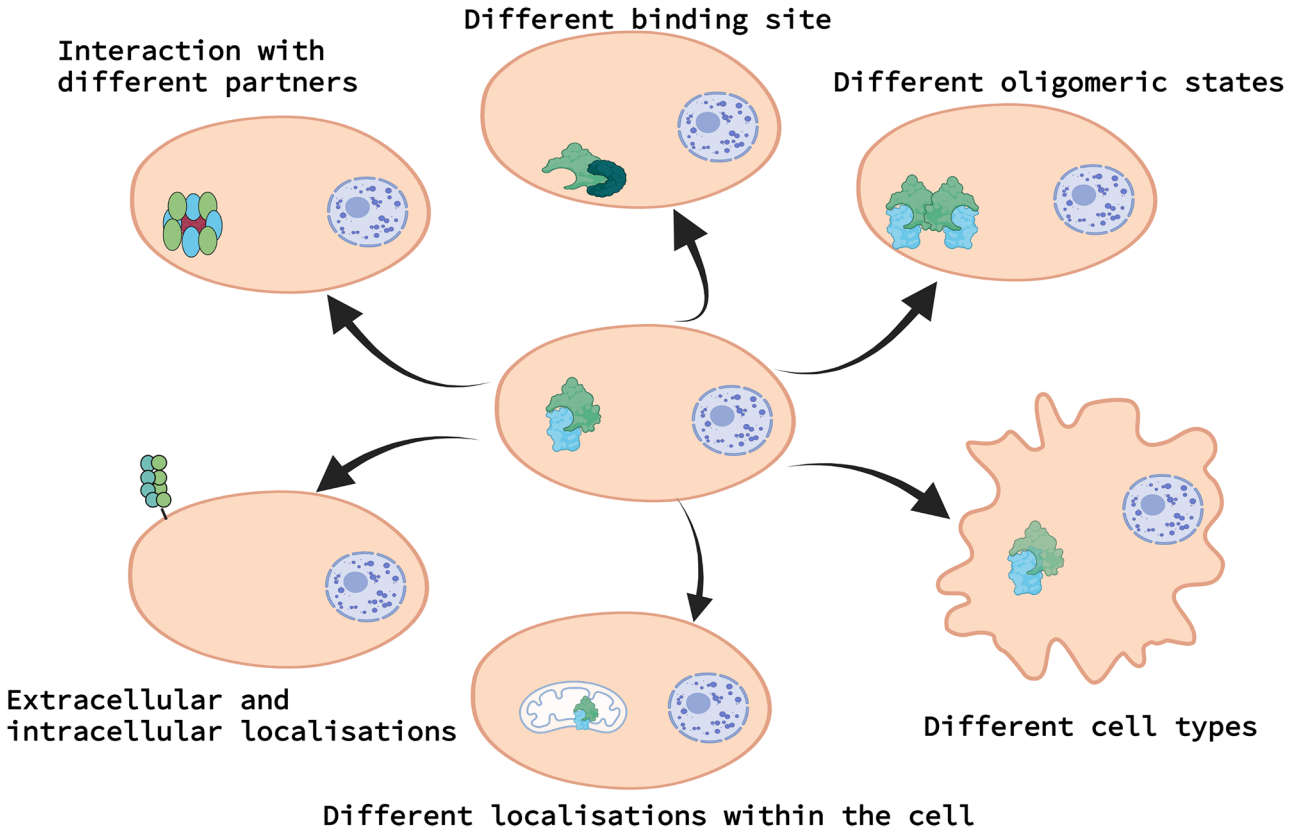


**Figure 1.** Mechanism of action of moonlighting proteins. This image was created with BioRender (https://biorender.com).

methods to label proteins as moonlighting in the first place. As a result, several databases of moonlighting proteins have been reported. Prominent among them are Moonprot3.0, which includes expert-curated protein (18), MoonDB 2.0 comprising of predicted and manually curated extreme multifunctional and moonlighting proteins (19) and MultitaskProtDB-II with curated moonlighting proteins (20). At least one study (PlantMP) has reported the manual curation of a species-oriented database of moonlighting proteins by experts (21).

Three major strategies have been adopted so far for identifying novel moonlighting proteins. The first identifies them using pre-existing annotations such as Gene Ontology and text mining. Computational methods that used this approach include MoonGO which identifies overlapping clusters in protein–protein interaction networks using overlapping cluster generator algorithm and combines it with Gene Ontology to identify candidate moonlighting proteins as a subclass of extreme multifunctional proteins (22,23) and DextMP which applied natural language processing to identify moonlighting proteins based on published literature (24). The second strategy, which goes beyond the compilation of moonlighting proteins from the literature and from ontology trees, attempts direct annotation of the moonlighting function (25,26). In general, sequence-based functional annotation has often been based on detection and identification of homologous, conserved motifs/domains. However, such an approach is difficult to apply to moonlighting proteins because the functional signatures for one of the functions may be too weak and camouflaged by the other well known function (25,27). Therefore, studies of sequence-based moonlighting annotations have attempted to exploit methods for detecting remote homologies. For example, Gomez *et al.* examined eleven approaches and determined that PSI-BLAST (28) performed relatively well at identifying moonlighting functions (29). Khan et al. compared protein function prediction (PFP) and extended similarity group (ESG) to PSI-BLAST and discovered that PFP, which derives functional information from weakly related sequences, was the most accurate in predicting the alternative functions of moonlighting proteins (26). Both studies suggest that non-canonical functions may be observed in distantly related sequences even when close homologies are absent. Gomez et al in 2011 confirmed that protein–protein interaction databases do indeed reveal moonlighting proteins and suggested that PPI databases might be beneficial for indicating multifunctionality (22,30). Hernandez *et al.* examined a collection of moonlighting proteins to determine if they are inherently disordered proteins since the latter easily fold into multiple conformations that could help perform multiple functions (4). However, their findings suggested that the majority of moonlighting proteins are not fundamentally disordered proteins. (31). Hernandaz *et al.* used PSI-BLAST to identify 42% of the 288 moonlighting proteins in MultitaskProtDB, whereas only 8% were detected by both PSI-BLAST and InterPro (27). Algorithms such as PSI-BLAST, PFP and ESG identify distant homologs by matching stretches of amino acid residues from

distinct domains to regions of a probable moonlighting protein. These searches generate a significant number of hits, from which the real positives must be extracted. Hernandez *et al.* discovered that when PSI-BLAST is paired with PPI, the best performance is obtained. Additionally, structural information and mutation correlation analysis may be used to further narrow the field (25). Finally, the third strategy to identify novel moonlighting functions aims to predict them using directly computed features, odelled with machine learning. Among them, MPFit employs many omics-based characteristics such as protein–protein interaction, gene expression, phylogenetic profiles, genetic relationships, network-based graph properties, and disordered protein regions, as well as the option to include or exclude Gene Ontology (32). IdentPMP tries to identify plant-only moonlighting proteins based on amino acid composition and content predicted using iLearn (33). MEL-MP attempts to predict proteins based on primary protein sequence information, evolutionary information, physical chemical properties and secondary structural features of the proteins (34). Shirafkhan *et al*. developed a method to predict moonlighting proteins based on 37 different feature vectors derived from amino acid sequences (35).

While these methods have shown a varied degree of predictive performance, they suffer from the fact that they treat moonlighting function as a single property without identifying specific roles of these features in moonlighting proteins of a special functional class with common functions. Such methods may have a tendency to exaggerate predictive performance due to the very general nature of the control data sets. We argue that studying moonlighting behavior for specific functional group of proteins in a specific organism will provide further clues into functional co-occurrences and help in predicting them more accurately without overestimating predictive accuracy. Thus, we selected one of the most widely investigated functional class of proteins, the human DNA-binding proteins (DBPs) and created a database of human moonlighting versus other DBPs (mDBPs versus oDBPs). Using these data sets and their sequence, structural, evolutionary, PPI network and gene expression features, we evaluated the extent to which such rigorously compiled groups can be discriminated from one another. Our results indicate that using strict controls of oDBP against mDBPs, ML models could be trained to test data AUC of ROC up to 74%. Proteins that were consistently predicted to be mDBPs are proposed to be candidate DBPs that possibly perform currently unknown moonlighting functions.

## MATERIALS AND METHODS

### Data preparation

Data sets are crucial for any predictive modeling, and updated data sets can improve predictability. However, labeling proteins as moonlighting through a manual curation process is a laborious job, partly because annotation of moonlighting is not reported with the corresponding searchable keyword. One needs to carry out substantial literature search to dig out the multifunctionality of proteins and then look at the sequence, structure and domain information to unambiguously annotate them as moonlighting. Numerous groups have systematically pursued database development, which is of a sufficiently good quality to develop trainable models. We have therefore used publicly available moonlighting databases for the current annotations of moonlighting proteins as described in the following sections. From the predictive models, candidate novel moonlighting proteins are proposed if they have multiple functions without a separate functional domain associated to each one of them. Detailed procedures involved in preparing training data sets are explained below.

### Moonlighting DNA binding proteins (mDBPs)

Figure 2 depicts a flowchart of the steps involved in preparing the dataset of mDBPs. As illustrated, we have currently focused on human moonlighting DNA binding proteins. Moonlighting annotations in our work have been taken from Moonprot 3.0 (18), MultitaskProtDB-II (20) and MoonDB 2.0 (19). However, DBP functions are not clearly annotated in these data sets. Therefore, we used a database of DNA-binding proteins in humans and their corresponding gene expression profiles from our previously published work with a database called GIGEASA (36). To combine the two complementary annotations, a list of compiled moonlighting proteins were compared to a list of 2407 DNA binding proteins listed in GIGEASA datasets (36). This resulted in the annotations of 46 mDBPs. A further set of filters was applied to exclude sequences whose similarity is higher than 25%, proteins having domains related to a secondary function documented in the literature, and those lacking a direct DNA binding annotation. This resulted in 29 high confidence non-redundant mDBPs as shown in Supplementary Table ST1.

### Other (non-moonlighting) DNA binding proteins (oDBPs)

We focus on proteins with common DBP function and characterize how an additional moonlighting function can be predicted for them. To compile control data of DBPs with no moonlighting function, we restarted from the list of 2407 DNA binding proteins from GIGEASA dataset and filtered out 46 moonlighting proteins and their known homologs. This resulted in 647 proteins, which we re-clustered at 25% sequence identity, removing three redundant proteins. An additional protein was removed due to failure in extracting all features analyzed in this work. This final list of 643 non-moonlighting (other) DNA binding proteins (oDBPs) were used as a control data of DBPs against mDBPs (Supplementary Table ST1).

### Feature sets covered

We identified different feature sets of moonlighting functions that were studied and reviewed earlier by several groups. In this work, we used five types of feature sets related to the protein sequence, structure, or gene-level expression profiles. These include (i) single protein sequence and predicted binding site features, (ii) sequence-based evolutionary features, (iii) network features based on protein–protein interactions, (iv) sequence-predicted secondary structural features and (v) global gene expression

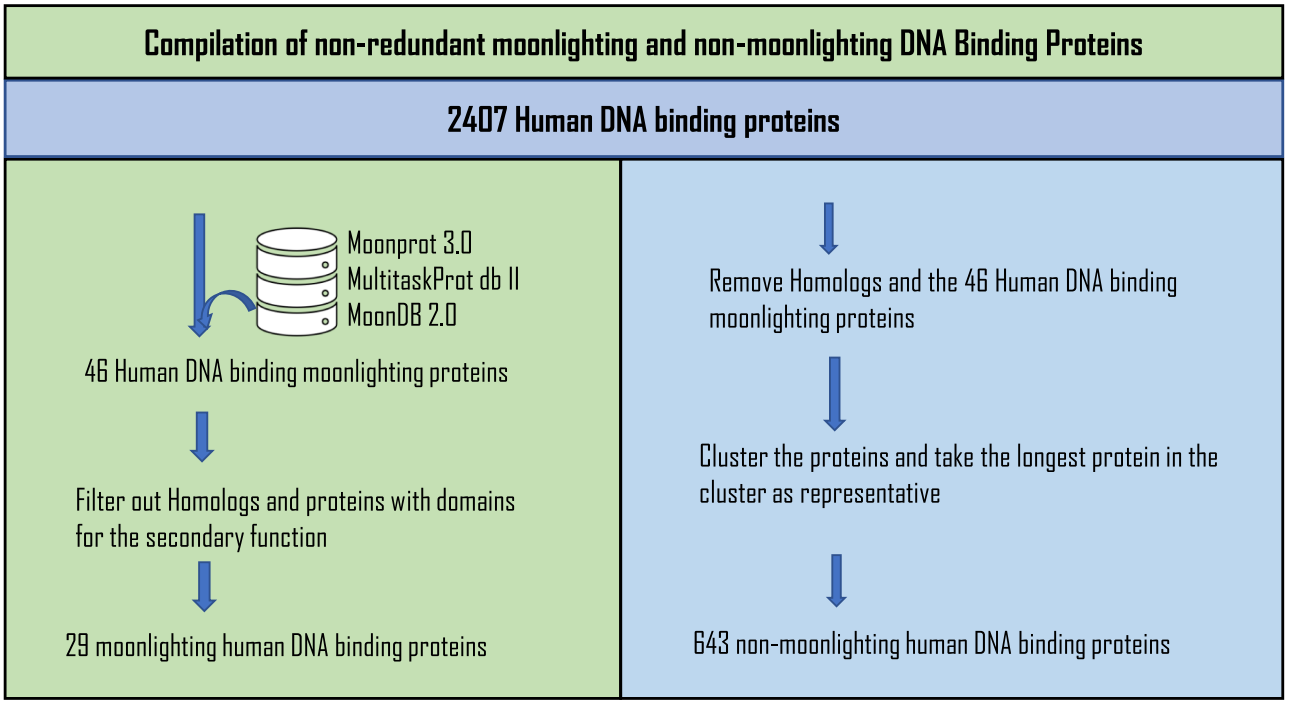


**Figure 2.** Flowchart showing the description of the data preparation to select moonlighting and non-moonlighting DNA binding proteins.

profiles. Computation of each of these features was carried out as follows.

### Single protein sequence and predicted binding site features

Single sequence features are represented by the amino acid composition of each protein, and their sequence-predicted binding sites features, taken from our previously published work (37). These so-called GIGEASA dataset features consist of the binding site prediction scores for carbohydrates, DNA, RNA, adenosine triphosphate (ATP), and protein binding sites, as well as the amino acid composition and length of the proteins (36). These features can be readily used for protein-level predictions in a function prediction program. Since the binding sites are predicted at the whole sequence level, their protein-level summaries are generated by protein-wise averages, quantiles and other representative scores as used in our previous work (36).

### Gene expression features

In our previous work, we have integrated gene expression profiles from the entire Affymetrix platform GPL570, representing >70 000 experiments. A single gene has been represented by the frequency of its occurrence in each of the 20 pre-defined bins of expression values. Frequencies of experiments (samples) in each bin is used as a 20-dimensional gene expression feature of each protein mapped to a specific gene in the database. Co-expression of a single gene with others is also vectorized in the same way. The gene expression and co-expression profiles that were binned into 20 equal-probability bins in the GIGEASA dataset were further coarse-grained to 5 bins to increase bin-wise occupancy of proteins. Coarse-graining was performed by pooling together every 4 successive bins starting from the first to make a new bin.

### Evolutionary features

Position specific scoring matrix (PSSM) (28) profiles for the 672 proteins were generated using PSI-BLAST with default parameters for three iterations against the NR database downloaded from NCBI. The log-odds score from the resulting profiles were taken and two types of PSSM features were generated for each protein; that is, the average of the log odds value for all the amino acids were taken and the average of the log odds value for each of the 20 amino acids separately (leading to $20 \times 20 = 400$ features). The concatenated feature values of 400 + 20 features were used as inputs for PSSM-based predictions.

### Predicted structural features

We predicted the secondary structure features (alpha helix, strand, coil), solvent accessibility (buried, exposed and moderate) and disorder (disordered or ordered) for each protein using a local copy of Raptor X (38). Raptor X gives both 3-state and 8-state secondary structure predictions of which we used 3-state secondary structure helix, beta-sheet, and loop to enable enough feature value diversity in each. The solvent accessibility of residues was predicted using 2 cutoffs for the three states. Those below 10% accessibility are predicted to be buried while those above 42% are considered as exposed and those between 10% and 42% are predicted to be moderate. Raptor X also gives an order or disorder prediction score at the residue level based on the probability of the residue to be in an ordered segment or not.

### PPI-network features

A protein–protein interaction (PPI) network represents all known protein interactions. When utilized with the functional annotation of the component proteins, they can aid in the discovery of moonlighting proteins, like MoonGO and OCG do (22,23). In our work, protein–protein interaction network features like the number of pathways that the protein is involved in and a binary representation of whether the protein is a hub or bottleneck extracted from Targetmine (39) are primary network features investigated for their ability to predict moonlighting.

### Model selection

Based on each feature set and their combination, a machine learning classifier was trained using an arbitrarily selected set of potential computational models. Each model attempts to predict class labels (mDBP or oDBP) by employing selected feature sets using leave-one-out cross-validation. Five different Machine learning algorithms, i.e. Random Forest (RF) (40), Balanced Random Forest (BRF) (41), Catboost (42), XGBoost (43,44) , and LightGBM (45) were compared. Catboost was the best performing model and was used for developing the final prediction model. Catboost is a gradient boosting decision tree-based algorithm which is good for dealing with categorical data because of the way it handles encoding and being less prone to overfitting, this model has become a preferred ML method for large data sets. It uses ordered boosting which makes sure that it does not evaluate a candidate tree with the examples it has used to build the tree, and eventually uses all the examples (42). Additionally, the performance of each feature set and combination of extracted features was examined. To deal with imbalanced data, classifier model overweights the minority class during training. Thus, Catboost classifier with auto_class_weights = 'Balanced' was used. This automatically calculates class weights either on the total weight or the total number of objects in each class. The values are used as multipliers for the object weights.

## RESULTS AND DISCUSSIONS

To assess if we can predict novel mDBPs directly from their computable features, we first examined the statistical distribution of representative feature sets and assessed how these features individually discriminate mDBPs from oDBPs. A binary classifier was then trained so that the cumulative impact of discriminatory features can be captured, prediction performance be evaluated and some candidate novel mDBPs could be proposed. Results of these analyses are presented below.

### Discriminatory features of moonlighting function in DBPs

We used five different groups of features for their potential to distinguish between mDBPs and oDBPs. Most considered feature sets have been compared in Figure 3A–F with additional details in Supplementary Table ST2 (explained below). Complete set of p-values for all the considered features are provided in Supplementary Table ST3. Investigation of most discriminatory members of each feature set are discussed in the following.

### Single sequence and predicted binding site differences between mDBPs and oDBPs:

Sequence features, including binding site predictions for DNA, RNA, ATP and carbohydrates were used as in our earlier works (36). CountX represents the frequency of occurrence of residue X in the protein sequence. For example, countS and countV represent the counts of the amino acids serine and valine. The top25ppi, top25dna, top25rna and top25carb represents the top 25 predicted binding scores for protein–protein interactions, DNA, RNA, and carbohydrate binding computed for each residue. Similarly, for *top5* and *top10*, *q3ppi* shows the quartile 3 of the predictions of protein–protein interaction sites. In Figure 3A, the 20 statistically most significant attributes amongst these features have been displayed. In addition, Figure 3B shows that the frequency of the number of pathways that a protein is involved in, and the nature of the topological positioning of the protein in the network like bottlenecks or hub proteins are linked to mDBPs versus oDBPs. From these Figure 3A and B, it become evident that the frequencies of Serine, Valine and Histidine are most significantly different between the two classes of proteins and these residues may play a role in allowing multiple functions. Among the binding site prediction scores, protein–protein interaction sites are found to be the most significant discriminatory features. The ability of sequence-predicted PPIs to discriminate between mDBPs and oDBPs is a novel result from the current work that supports these previous observations and may be useful in identifying moonlighting proteins other than DBPs.

### Moonlighting and evolutionary profiles

Evolutionary profiles of proteins represented by their position-specific substitution matrices (PSSMs) have been widely used to annotate proteins, including DBPs and their binding sites (46–48). It is not obvious at the outset if they will also have different patterns in mDBPs versus oDBPs. Since the PSSMs of proteins are of different sizes, a summary of PSSM is typically used as a feature for predictive models. Consistent with that practice we computed 20 overall and then 20 × 20 amino-acid wise column averages of PSSMs as protein features as described in the Methods (46) and performed a t-test of significance between mDBP and oDBPs (see Figure 3C). We do observe that many evolutionary features from the 20 × 20 representation of amino acid pairs are significantly different between mDBPs and oDBPs. Complete results are shown in Supplementary Table ST2. Specifically, the average log odd score of Ser in Pro columns and average Ala log odd scores in Val columns are found to have the highest significance. Among the overall column averages, Tyr was found to be the most significant residue column among all 20 amino acids. Other pairs such as Tyr in the Trp column, Asp in the Glu column and Cys in the Val column are the top residues in evolutionary profiles which distinguish between mDBPs and oDBPs.

### Predicted structural features

Sequence-predicted structural features have been shown to improve protein function prediction (49–53). We predicted

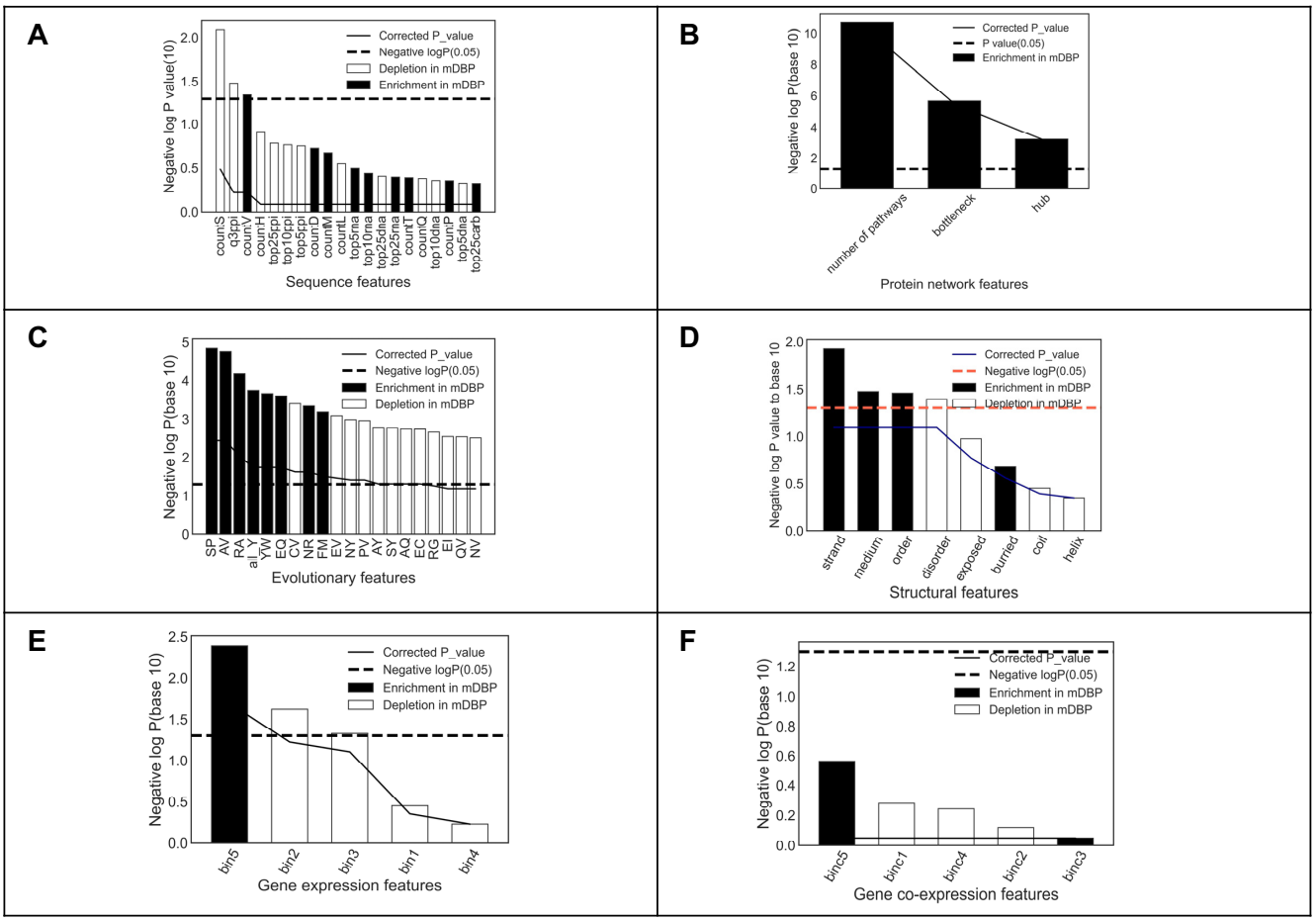


**Figure 3.** Comparison of sequence, network, predicted structural, evolutionary and global gene expression features between moonlighting and other DNA-binding proteins (mDBP vs oDBP). (**A**) Sequence features, including binding site predictions for DNA, RNA, ATP and carbohydrates as utilized in our earlier works, amino acid compositions e.g. frequency of Ser, Val shown as CountS and CountV (see Materials and Methods). (**B**) Protein-protein interaction features include the number of pathways in which the proteins participate and the attributes of the proteins within their interaction networks, such as whether they operate as bottlenecks or hub proteins. (**C**) Evolutionary characteristics represented by the average log odds value for all amino acids (like all_Y refers to the average logodd value of Tyrosine) and for individual amino acids (represented as XY where X shows all the rows where the amino acid X is found and second alphabet represents the column Y for which the average value is calculated. Likewise, SP represents all rows where Serine occurs and Proline is present in the column). (**D**) Predicted secondary structural features such as strand, coil or helix, solvent accessibility as buried, moderate, or exposed, and conformational aspects of order and disorder. (**E**) Gene expression and co-expression quantified by their frequency of occurrence in the five bins, reduced from 20 bins in our previous works as defined in Materials and Methods. (**F**) Gene co-expression quantified by their frequency of occurrence in the five bins, reduced from 20 bins in our previous works as defined in Methods.

some structural properties of mDBPs and oDBPs and compared their average values in the two groups. Figure 3D presents the results of this analysis. We observe that the average number of predicted strands is higher in mDBPs than oDBPs. DNA-binding proteins primarily interact through their recognition helices and hence the presence of a higher number of strands may itself be an indication of moonlighting function. One might wonder that keeping the helical interfaces intact for DNA-binding activity, additional strands are only used for moonlighting function. To examine this, we did compare the sizes of the mDBPS and oDBPs together with their strand counts (supplementary table ST1). We observed that even though the strands count on the whole is higher in mDBPs, it does not appear to be driven by protein sizes, as there was no sigjificant difference between the lengths of mDBPs and oDBPs. One might speculate that actual DNA interface is also composed of strands in moonlighting proteins. However, in the absence of complete protein-DNA complex structures with moonlighting substrates, drawing a general rule is difficult at this stage.

Finally, the number of residues in the intermediate solvent accessibility range is higher in moonlighting proteins. This observation is interesting as a partially exposed surface area in these residues may allow them to switch from exposed to buried states quickly, enabling moonlighting. It would also be of interest in the future to explore if the presence of partially exposed residues is a general property of all moonlighting proteins or specific to moonlighting DBPs.

## Gene expression features of mDBPs versus oDBPs

In our previous works, we showed that global gene expression profiles collected from >70 000 experiments can be used to improve the predictability of DNA-binding proteins, particularly those which have weak sequence level binding scores in prediction models (36). These gene expres-

sion profiles are actually the relative number of experiments for which a corresponding gene was found to have an expression value represented by an expression value range or bin. Similarly, co-expression profile is the number of times a gene had its co-expression (Pearson's correlation) value within the range represented by the corresponding range or bin, counted for all gene pairs involving that gene. In our previous work, we used 20 bins to represent expression values and separately, the co-expression of a gene paired with all other genes. In this work, the 20-bins were further coarse-grained into 5 bins to enable more data in each bin for the small sample size here. After merging the bins as described in Methods, we could determine the relative number of times a mDBP shows expression values in the range represented by each bin and compare it with the frequencies observed in oDBPs. Figure 3E-F shows the results from this analysis. The bins from the normalized expression values are represented by bin1, bin2 upto bin5 and those based on co-expression are binc1, binc2, onwards. We observe that absolute expression values distribution of gene expression remains predictive of mDBPs to some extent as the number of times the two frequencies differ in mDBPs from oDBPs is statistically significant. Thus, the results indicate that moonlighting proteins express at relatively higher level than their non-moonlighting counterparts, thereby increasing the frequency in bin5 at the cost other bins. Although, we do see some differences in the co-expression profiles as well, none of the individual frequencies in each bin was observed to be statistically significant. It is however possible that the different weak features are collectively predictive of mDBPs, an issue that we will look into in the next section on classifier performances.

### Classification model for mDBP versus oDBPs

In the above sections, we investigated individual features of mDBPs which might distinguish them from oDBPs. However, many of these individual features may have a synergetic effect and the same can be captured by developing predictive models using each feature set as input and then combining and assessing the predictive performance by comparisons. We would also like to assess if the available methods of predictive moonlighting proteins are outperformed by the proposed exclusive mDBP de novo prediction. Several methods for predicting moonlighting for proteins are available with varied claims of performance. For example, DextMP, IdentPMP, MEL-MP and an unnamed method by Zahiri *et al.* reported AUC values of 80% and 90%. MPFIT, on the other hand, reported 98% accuracy. However, all these methods are general in nature and their applicability to a specific group of proteins such as mDBPs was unclear. Only MPFit's source codes were available for the prediction of moonlighting proteins using the non-redundant DNA binding dataset that was compiled. MEL-MP could also be used to fetch a list of predicted mDBPs. Thus, we tested the performances of these two groups of methods for predicting mDBPs from oDBPs. We used three variants of MPFIT and MEL-MP to evaluate if general moonlighting prediction methods can also pick the moonlighting DBPs (See Table 1). We observed that these available methods could only reach an AUC of ROC close to 62% in predicting mDBPs from oDBPs, making the development of a novel method even more important.

The comparison of prediction performance with available published methods is provided only to explain that a focused study of single biological function can help in improving moonlighting in that class of proteins. We do not suggest that first principles method of predicting moonlighting proteins is better than published works when it comes to overall prediction of moonlighting, which we have deliberately not attempted in view of the scope of this work.

To use a first principles method to predict mDBPs based on the feature sets investigated above, we trained the mDBP versus oDBP data sets analyzed above using the catboost approach. (catboost was found the most suitable among the models tested on random samples; data not shown). These results could be further improved by running catboost multiple (10) times and averaging the predictions from the ensemble of these models. Results from all these experiments under leave-one-out cross-validation training are presented in a section of Table 1. Additional details are shown in ROC and PR curves in supplementary figures SF1 and SF2.

### Predictability of mDBPs from individual and cumulative feature sets

The second section of Table 1 shows that protein–protein interaction network-based features were able to better differentiate between moonlighting and non-moonlighting DNA binding proteins with a performance of 72% AUC of ROC with the best precision and recall. The combination of all features reduced the performance to AUC of ROC 69%, presumably due to over-fitting and an increase in dimensionality. Since the feature sets were not additive in their prediction performance, we performed ensemble methods to predict moonlighting proteins from their prediction scores as well as labels. The final performance was 74%.

The performance of our current model was then compared with other prediction models developed for the prediction of moonlighting proteins.

### Candidate novel moonlight DBPs

Combining all feature sets into an ensemble model, we observed an AUC of ROC at 74% in our data sets. However, false positives with high prediction scores could still be candidate novel mDBPS whose moonlighting has yet not been established. Thus, we provide the database of all final predicted scores for all the DBPs in our database (Supplementary Table ST4). From this table, we compiled a list of top scoring false positive oDBPs and searched the literature for their support as candidate novel mDBPs (see Table 2). We found that 10 out of 53 top false positives from our predictions had literature support. For example, a false positive mDBP in Table 2, Cyclin-dependent kinase 9 (Cdk9) is a subunit of the positive transcription elongation factor b (P-TEFb), which promotes the elongation of pre-mRNA. Although a kinase per say, it has been suggested that CDK9 responds to replication stress by localizing to chromatin to reduce the breakdown of stalled replication forks and promote recovery from replication arrest (54). It likely binds to promoter regions of certain

**Table 1.** Prediction performance of different predictive models in the public domain and those proposed with different feature sets. Overall 672 proteins are included in these evaluation and all prediction models are trained using leave one out method. Ten iterations of training are carried out and for our proposed models and performance scores are averaged to assess their predictions with standard deviation between models shown alongside. Final model with 74% AUC is based on prediction scores derived by averaging 10 models trained on all feature sets taken together. It may be noted that performance scores, other than AUC are threshold-dependent as the prediction output is continuous (not binary) and in our work, we have selected the thresholds which correspond to the best *F*-score

| | Feature set | AUC | MCC | F-score | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|
| **Public sourced method** | **MPFIT(Phylo + GE + GI + DOR + NET)** | 0.56 | 0.06 | 0.11 | 0.77 | 0.34 | 0.79 |
| | **MPFIT(Phylo + PPI + GE)** | 0.51 | 0.01 | 0.08 | 0.13 | 0.93 | 0.09 |
| | **MPFIT(from provided predicted proteins)** | 0.60 | 0.17 | 0.21 | 0.91 | 0.28 | 0.94 |
| | **MEL-MP(provided predicted proteins)** | 0.62 | 0.11 | 0.14 | 0.74 | 0.48 | 0.75 |
| **De novo mDBP vs oDBP prediction model performances** | **Sequence** | 0.57 ± 0.03 | -0.04 e-2 ± 0.03 | 0.04 ± 0.03 | 0.93 ± 0.00 | 0.03 ± 0.03 | 0.97 ± 0.00 |
| | **Secondary structure** | 0.49 ± 0.01 | 0.05 ± 0.02 | 0.10 ± 0.01 | 0.73 ± 0.00 | 0.35 ± 0.03 | 0.75 ± 0.01 |
| | **Gene expression** | 0.60 ± 0.01 | 0.08 ± 0.02 | 0.12 ± 0.01 | 0.75 ± 0.01 | 0.40 ± 0.04 | 0.77 ± 0.01 |
| | **Evolutionary** | 0.65 ± 0.02 | 0.11 ± 0.04 | 0.15 ± 0.04 | 0.91 ± 0.00 | 0.20 ± 0.06 | 0.94 ± 0.00 |
| | **Network** | 0.72 ± 0.01 | 0.18 ± 0.01 | 0.19 ± 0.01 | 0.80 ± 0.00 | 0.53 ± 0.02 | 0.82 ± 0.00 |
| | **All** | 0.69 ± 0.02 | 0.11 ± 0.04 | 0.15 ± 0.04 | 0.92 ± 0.00 | 0.17 ± 0.05 | 0.95 ± 0.00 |
| | **Ensemble averaged** | 0.74 ± 0.01 | 0.16 ± 0.04 | 0.19 ± 0.03 | 0.93 ± 0.00 | 0.19 ± 0.03 | 0.97 ± 0.00 |

transcription factors in cardiac muscles (https://zfin.org/ZDB-GENE-030131-321). Through an unknown mechanism, Cdk9 complexes with cyclin K, Atrip, Atr, and claspin proteins, thereby regulating single-stranded DNA interaction with replication protein A, ensuring the stability of the replication fork (55). Loss of Cdk 9-cyclin K complex activity increases DNA damage signaling in replicating cells with a diminished capacity to recover from replication arrest (55,56).

Another example in the candidate list corresponds to Inosine-5′-monophosphate dehydrogenase 2 (IMPDH2). This protein has been known for enzymatic activity before being established as a transcription factor. Somehow, the protein has not been included in the current updates of moonlighting databases

Another protein, FOS, has been already reported to be involved in lipid synthesis in neurons and based on that can be annotated as moonlighting (57). Somehow, this annotation has escaped from the current versions of moonlighting databases and our method was able to recover its annotation as mDBP. Similarly, we found moonlighting support for another DBP called NRF-1. NRF-1 is a transcription factor which has been reported to perform a role in neurite growth as well as lipid homeostasis (58,59).

Yet another candidate mDBP in Table 2, mitochondrial superoxide dismutase 2 (SOD2) protein, is an enzyme that primarily accelerates the dismutation of O2. However, further increases in SOD2 expression was reported to worsen oxidative stress, suggesting that SOD2 may play a prooxidant role (60). Also, SOD2, which normally binds manganese, can also contain iron, and generate a peroxidase-active isoform, further supporting its moonlighting function.

Overall, we found support for 10 out of 53 proteins in Table 2 for being moonlighting. These include PC, ABI2, RPS27, PARK7, and EEF1D in addition to the five specific cases (CDK9, IMPDH2, FOS, NRF-1 and SOD2) above. Thus, we believe there may be more mDBPs in the list of false positives in Table 2, whose moonlighting behavior may be established in the future.

## Cellular localization and moonlighting

Many moonlighting proteins perform their alternative functions through cellular localization. For example, a DBP which is a trans-membrane protein must translocate to the nucleus to act as a transcription factor. However, even if cellular localization is the primary driver of moonlighting, the protein needs to be transported to different locations, which is why many proteins contain a localization signal sequence and indeed importance of short sequence motifs has been documented (23,61–62). Yet, several proteins localize to different cellular compartments even when they do not have a detectable motif towards and hence the absence of a motif does not imply that a protein does not have a sequence or structure level signal. There are multiple publications which rely on sequence features for the prediction of cellular localization of proteins. For example, WoLF PSORT, YLoc, TargetP and TMHMM (63–66) have predicted the cellular localization of proteins with as high as 80% accuracy from sequence information. Thus, cellular localization information appears implicitly contained in the sequence and structural properties of the proteins and our method essentially tries to capture that indirect relationship.

In order to understand the relationship between cellular localization and moonlighting in detail, we extensively examined the predictability of mDBPs from available cellular localization annotations. We used the ‘Compartment database’ of cellular localization (67) and tried to develop a prediction model for moonlighting proteins only from compartment prediction. We did observe that cellular compartment annotation alone was able to predict moonlighting DBP function with 86% accuracy, much better than the first principles method proposed in this work. However, it also indicates that most of the current annotations of moonlighting proteins are based on their known cellular localization or their well-studied functions. This approach is good to mine for moonlighting proteins that are already noted by different names or under different contexts. However, they cannot predict novel moonlighting proteins in the way a first principles approach like the one proposed here could

**Table 2.** List of candidate novel mDBPs based on their top false positive scores

| Uniprot ID | Gene HGNC code | Name of the protein | Uniprot ID | Gene HGNC code | Name of the protein |
|---|---|---|---|---|---|
| Q09028 | RBB4 | Histone-binding protein RBBP4 | P35249 | RFC4 | Replication factor C subunit 4 |
| P27694 | RPA1 | Replication protein A 70 kDa DNA-binding subunit | Q16531 | DDB1 | DNA damage-binding protein 1 |
| P50750 | CDK9 | Cyclin-dependent kinase 9 | P41221 | WNT5A | Protein Wnt-5a |
| P12268 | IMPDH2 | Inosine-5'-monophosphate dehydrogenase 2 | Q9NYA1 | SPHK1 | Sphingosine kinase 1 |
| O15160 | POLR1C | DNA-directed RNA polymerases I and III subunit RPAC1 | P08047 | SP1 | Transcription factor Sp1 |
| P12004 | PCNA | Proliferating cell nuclear antigen | O14744 | PRMT5 | Protein arginine N-methyltransferase 5 |
| P42677 | RPS27 | 40S ribosomal protein S27 | P15927 | RPA2 | Replication protein A 32 kDa subunit |
| P07910 | HNRNPC | Heterogeneous nuclear ribonucleoproteins C1/C2 | P11387 | TOP1 | DNA topoisomerase 1 |
| Q86 × 55 | CARM1 | Histone-arginine methyltransferase CARM1 | P18848 | ATF4 | Cyclic AMP-dependent transcription factor ATF-4 |
| P11498 | PC | Pyruvate carboxylase, mitochondrial | Q6ZYL4 | GTF2H5 | General transcription factor IIH subunit 5 |
| O96019 | ACTL6A | Actin-like protein 6A | P56282 | POLE2 | DNA polymerase epsilon subunit 2 |
| Q99497 | PARK7 | Parkinson disease protein 7 | Q9NYB9 | ABI2 | Abl interactor 2 |
| Q02878 | RPL6 | 60S ribosomal protein L6 | Q14814 | MEF2D | Myocyte-specific enhancer factor 2D |
| Q9Y230 | RUVBL2 | RuvB-like 2 | P49005 | POLD2 | DNA polymerase delta subunit 2 |
| O75534 | CSDE1 | Cold shock domain-containing protein E1 | P01100 | FOS | Protein c-Fos |
| P25490 | YY1 | Transcriptional repressor protein YY1 | O60907 | TBL1X | F-box-like/WD repeat-containing protein TBL1X |
| Q16656 | NRF1 | Nuclear respiratory factor 1 | P30876 | POLR2B | DNA-directed RNA polymerase II subunit RPB2 |
| Q9UHX1 | PUF60 | Poly(U)-binding-splicing factor | Q96T60 | PNKP | Bifunctional polynucleotide phosphatase/kinase |
| P35232 | PHB1 | Prohibitin 1 | Q00403 | GTF2B | Transcription initiation factor IIB |
| P20226 | TBP | TATA-box-binding protein | P19388 | POLR2E | DNA-directed RNA polymerases I, II, and III subunit RPABC1 |
| Q13620 | CUL4B | Cullin-4B | P30044 | PRDX5 | Peroxiredoxin-5, mitochondrial |
| P29692 | EEF1D | Elongation factor 1-delta | O60869 | EDF1 | Endothelial differentiation-related factor 1 |
| P55895 | RAG2 | V(D)J recombination-activating protein 2 | P62841 | RPS15 | 40S ribosomal protein S15 |
| P35244 | RPA3 | Replication protein A 14 kDa subunit | P19838 | NFKB1 | Nuclear factor NF-kappa-B p105 subunit |
| Q9UQ80 | PA2G4 | Proliferation-associated protein 2G4 | P04179 | SOD2 | Superoxide dismutase |
| P04083 | ANXA1 | Annexin A1 | P05067 | APP | Amyloid-beta precursor protein |
| Q12824 | SMARCB1 | SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily B member 1 | | | |

do. Nor do they provide insights into the mechanistic basis of moonlighting. Further, cellular localization databases and prediction methods are available only for a few species and methods relying heavily on these annotations are not directly applicable for inferring proteins that perform a moonlighting function. Thus, a case for predicting novel mDBPs from the proposed features is made out for this work.

## CONCLUSIONS

In this work, we have investigated the moonlighting behavior of a special class of proteins, that is, DNA-binding proteins as compared to their non-moonlighting counterparts. We looked at their predictability by general methods of moonlighting prediction in the public domain and developed a thorough strategy to predict them from first principles using sequence, predicted structure, evolutionary profiles, and global gene expression profiles. Our results indicate that mDBPs can indeed be predicted from proposed feature sets with reasonable confidence. Some of the high-scoring false predictions of mDBPs were found to already have literature evidence of their being mDBPs and others are proposed to be candidate novel mDBPs and need further experimental assessment.

## DATA AVAILABILITY

The source codes and data sets used for this work are freely available at https://github.com/Sciwhylab/DNA_binding_moonlighting-protein_predictor.git and https://doi.org/10.5281/zenodo.7299265.

## SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

## ACKNOWLEDGEMENTS

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial

products, or organizations imply endorsement by the U.S. Government.

## FUNDING

National Cancer Institute, National Institutes of Health [HHSN261201500003I]; Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research (in part). The work has been supported in part by Indian Council of Medical Research Fellowships (to DMV)
*Conflict of interest statement.* None declared.

## REFERENCES

1. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G. and Fairley,S. (2012) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
2. Zhang,Z., Wang,H., Li,M., Agrawal,S., Chen,X. and Zhang,R. (2004) MDM2 is a negative regulator of p21WAF1/CIP1, independent of p53. *J. Biol. Chem.*, **279**, 16000–16006.
3. Saji,S., Okumura,N., Eguchi,H., Nakashima,S., Suzuki,A., Toi,M., Nozawa,Y., Saji,S. and Hayashi,S.-i. (2001) MDM2 enhances the function of estrogen receptor α in human breast cancer cells. *Biochem. Biophys. Res. Commun.*, **281**, 259–265.
4. Tompa,P., Szász,C. and Buday,L. (2005) Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.*, **30**, 484–489.
5. Koliadenko,V. and Wilanowski,T. (2020) Additional functions of selected proteins involved in DNA repair. *Free Radic. Biol. Med.*, **146**, 1–15.
6. Jeffery,C.J. (1999) Moonlighting proteins. *Trends Biochem. Sci.*, **24**, 8–11.
7. Huberts,D.H. and van der Klei,I.J. (2010) Moonlighting proteins: an intriguing mode of multitasking. *Biochim. Biophys. Acta (BBA) Mol. Cell Res.*, **1803**, 520–525.
8. Copley,S.D. (2012) Moonlighting is mainstream: paradigm adjustment required. *Bioessays*, **34**, 578–588.
9. Jeffery,C.J. (2014) An introduction to protein moonlighting. *Biochem. Soc. Trans.*, **42**, 1679–1683.
10. Amblee,V. and Jeffery,C. (2015) Physical features of intracellular proteins that moonlight on the cell surface. *PLoS One*, **10**, e0130575.
11. Piatigorsky,J. (2003) Gene sharing, lens crystallins and speculations on an eye/ear evolutionary relationship. *Integr. Comp. Biol.*, **43**, 492–499.
12. Sax,C.M. (1994) Expression of the α-crystallin/small heat-shock protein/molecular chaperone genes in the lens and other tissues. *Adv. Enzyme Relat. Areas Mol. Biol.*, **69**, 155–201.
13. Wistow,G. and Kim,H. (1991) Lens protein expression in mammals: taxon-specificity and the recruitment of crystallins. *J. Mol. Evol.*, **32**, 262–269.
14. Baek,S.J., Kim,K.-S., Nixon,J.B., Wilson,L.C. and Eling,T.E. (2001) Cyclooxygenase inhibitors regulate the expression of a TGF-β superfamily member that has proapoptotic and antitumorigenic activities. *Mol. Pharmacol.*, **59**, 901–908.
15. Cekanova,M., Lee,S.-H., Donnell,R.L., Sukhthankar,M., Eling,T.E., Fischer,S.M. and Baek,S.J. (2009) Nonsteroidal anti-inflammatory drug-activated gene-1 expression inhibits urethane-induced pulmonary tumorigenesis in transgenic mice. *Cancer Prev. Res.*, **2**, 450–458.
16. Baek,S.J., Wilson,L.C. and Eling,T.E. (2002) Resveratrol enhances the expression of non-steroidal anti-inflammatory drug-activated gene (NAG-1) by increasing the expression of p53. *Carcinogenesis*, **23**, 425–432.
17. Bianchi,M.E. and Agresti,A. (2005) HMG proteins: dynamic players in gene regulation and differentiation. *Curr. Opin. Genet. Dev.*, **15**, 496–506.
18. Chen,C., Liu,H., Zabad,S., Rivera,N., Rowin,E., Hassan,M., Gomez De Jesus,S.M., Llinás Santos,P.S., Kravchenko,K. and Mikhova,M. (2021) MoonProt 3.0: an update of the moonlighting proteins database. *Nucleic Acids Res.*, **49**, D368–D372.
19. Ribeiro,D.M., Briere,G., Bely,B., Spinelli,L. and Brun,C. (2019) MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. *Nucleic Acids Res.*, **47**, D398–D402.
20. Franco-Serrano,L., Hernández,S., Calvo,A., Severi,M.A., Ferragut,G., Pérez-Pons,J., Piñol,J., Pich,Ò., Mozo-Villarias,Á. and Amela,I. (2018) MultitaskProtDB-II: an update of a database of multitasking/moonlighting proteins. *Nucleic Acids Res.*, **46**, D645–D648.
21. Su,B., Qian,Z., Li,T., Zhou,Y. and Wong,A. (2019) PlantMP: a database for moonlighting plant proteins. *Database*, **2019**, baz050.
22. Becker,E., Robisson,B., Chapple,C.E., Guénoche,A. and Brun,C. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, **28**, 84–90.
23. Chapple,C.E., Robisson,B., Spinelli,L., Guien,C., Becker,E. and Brun,C. (2015) Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.*, **6**, 7412.
24. Khan,I., Bhuiyan,M. and Kihara,D. (2017) DextMP: deep dive into text for predicting moonlighting proteins. *Bioinformatics*, **33**, i83–i91.
25. Hernández,S., Franco,L., Calvo,A., Ferragut,G., Hermoso,A., Amela,I., Gómez,A., Querol,E. and Cedano,J. (2015) Bioinformatics and moonlighting proteins. *Front. Bioeng. Biotechnol.*, **3**, 90.
26. Khan,I., Chitale,M., Rayon,C. and Kihara,D. (2012) In: *BMC proceedings*. BioMed Central, Vol. **6**, pp. 1–5.
27. Hernandez,S., Ferragut,G., Amela,I., Perez-Pons,J., Pinol,J., Mozo-Villarias,A., Cedano,J. and Querol,E. (2014) MultitaskProtDB: a database of multitasking proteins. *Nucleic Acids Res.*, **42**, D517–D520.
28. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
29. Gomez,A., Domedel,N., Cedano,J., Piñol,J. and Querol,E. (2003) Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins?*Bioinformatics*, **19**, 895–896.
30. Gomez,A., Hernandez,S., Amela,I., Piñol,J., Cedano,J. and Querol,E. (2011) Do protein–protein interaction databases identify moonlighting proteins?*Mol. Biosyst.*, **7**, 2379–2382.
31. Hernandez,S., Amela,I., Cedano,J., Piñol,J., Perez-Pons,J.A., Mozo-Villarias,A. and Querol,E. (2012) Do moonlighting proteins belong to the intrinsically disordered protein class? *J. Proteom. Bioinform.*, **5**, 262–264.
32. Khan,I. and Kihara,D. (2016) Genome-scale prediction of moonlighting proteins using diverse protein association information. *Bioinformatics*, **32**, 2281–2288.
33. Liu,X., Shen,Y., Zhang,Y., Liu,F., Ma,Z., Yue,Z. and Yue,Y. (2021) IdentPMP: identification of moonlighting proteins in plants using sequence-based learning models. *PeerJ*, **9**, e11900.
34. Li,Y., Zhao,J., Liu,Z., Wang,C., Wei,L., Han,S. and Du,W. (2021) De novo prediction of moonlighting proteins using multimodal deep ensemble learning. *Frontiers in Genetics*, **12**, 254.
35. Shirafkan,F., Gharaghani,S., Rahimian,K., Sajedi,R.H. and Zahiri,J. (2021) Moonlighting protein prediction using physico-chemical and evolutional properties via machine learning methods. *BMC Bioinformatics*, **22**, 261.
36. Ahmad,S., Prathipati,P., Tripathi,L.P., Chen,Y.-A., Arya,A., Murakami,Y. and Mizuguchi,K. (2018) Integrating sequence and gene expression information predicts genome-wide DNA-binding proteins and suggests a cooperative mechanism. *Nucleic Acids Res.*, **46**, 54–70.
37. Andrabi,M., Mizuguchi,K., Sarai,A. and Ahmad,S. (2009) Prediction of mono-and di-nucleotide-specific DNA-binding sites in proteins using neural networks. *BMC Struct. Biol.*, **9**, 30.
38. Källberg,M., Wang,H., Wang,S., Peng,J., Wang,Z., Lu,H. and Xu,J. (2012) Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.*, **7**, 1511–1522.
39. Chen,Y.-A., Tripathi,L.P., Fujiwara,T., Kameyama,T., Itoh,M.N. and Mizuguchi,K. (2019) The targetmine data warehouse: enhancement and updates. *Front. Genetics*, **10**, 934.
40. Breiman,L. (2001) Random forests. *Machine Learning*, **45**, 5–32.
41. Chen,C., Liaw,A. and Breiman,L. (2004) Using random forest to learn imbalanced data. *University of California, Berkeley*, **110**, 24.
42. Hancock,J.T. and Khoshgoftaar,T.M. (2020) CatBoost for big data: an interdisciplinary review. *J. Big Data*, **7**, 94.
43. Chen,T. and Guestrin,C. (2016) In: *Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining*. pp. 785–794.

44. Le,N., Do,D., Nguyen,T. and Le,Q. (2021) A sequence-based prediction of Kruppel-like factors proteins using XGBoost and optimized features. *Gene*, **787**, 145643.
45. Ke,G., Meng,Q., Finley,T., Wang,T., Chen,W., Ma,W., Ye,Q. and Liu,T.-Y. (2017) Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process Syst.*, **30**, 3146–3154.
46. Chauhan,S. and Ahmad,S. (2019) Enabling full-length evolutionary profiles based deep convolutional neural network for predicting DNA-binding proteins from sequence. *Proteins: Struct. Funct. Bioinformatics*, **88**, 15–30.
47. Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins.. *BMC Bioinformatics*, **6**, 33.
48. Le,N. and Ou,Y. (2016) Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins. *BMC Bioinformatics*, **17**, 501.
49. Ofran,Y., Mysore,V. and Rost,B. (2007) Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**, i347–i353.
50. Yang,X., Wang,J., Sun,J. and Liu,R. (2015) SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. *PloS one*, **10**, e0133260.
51. Peng,Z. and Kurgan,L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.*, **43**, e121.
52. Taherzadeh,G., Yang,Y., Zhang,T., Liew,A.W.C. and Zhou,Y. (2016) Sequence-based prediction of protein–peptide binding sites using support vector machine. *J. Comput. Chem.*, **37**, 1223–1229.
53. Zhang,T., Zhang,H., Chen,K., Ruan,J., Shen,S. and Kurgan,L. (2010) Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Pept. Sci*, **11**, 609–628.
54. Yu,D. and Cortez,D. (2011) A role for CDK9-cyclin k in maintaining genome integrity. *Cell Cycle*, **10**, 28–32.
55. Guo,Y., Shyr,Y. and Cortez,D. (2010) Cyclin-dependent kinase 9–cyclin k functions in the replication stress response. *EMBO Rep.*, **11**, 876–882.
56. Lim,S. and Kaldis,P. (2013) Cdks, cyclins and CKIs: roles beyond cell cycle regulation. *Development*, **140**, 3079–3093.
57. Rodríguez-Berdini,L., Ferrero,G., Bustos,P.F., AM,C.G., Prucca,C., Quiroga,S. and Caputto,B. (2020) The moonlighting protein c-Fos activates lipid synthesis in neurons, an activity that is critical for cellular differentiation and cortical development. *J. Biol. Chem.*, **295**, 8808–8818.
58. Chang,W., Chen,H., Chiou,R., Chen,C. and Huang,A. (2005) A novel function of transcription factor alpha-Pal/NRF-1: increasing neurite outgrowth. *Biochem. Biophys. Res. Commun.*, **334**, 199–206.
59. Ruvkun,G. and Lehrbach,N. (2022) Regulation and functions of the ER-Associated nrf1 transcription factor. *Cold Spring Harb. Perspect. Biol.*, https://doi.org/10.1101/cshperspect.a041266.
60. Ganini,D., Santos,J., Bonini,M. and Mason,R. (2018) Switch of mitochondrial superoxide dismutase into a prooxidant peroxidase in manganese-deficient cells and mice. *Cell Chem. Biol.***25**, 413–425.
61. Henderson,B. and Martin,A. (2014) Protein moonlighting: a new factor in biology and medicine. *Biochem. Soc. Trans*, **42**, 1671–1678.
62. Zanzoni,A., Ribeiro,D.M. and Brun,C. (2019) Understanding protein multifunctionality: from short linear motifs to cellular functions.. *Cell. Mol. Life Sci*, **76**, 4407–4412.
63. Horton,P., Park,K.-J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
64. Briesemeister,S., Rahnenführer,J. and Kohlbacher,O. (2010) YLoc—an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.*, **38**, W497–W502.
65. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
66. Krogh,A., Larsson,B., Von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
67. Binder,J.X., Pletscher-Frankild,S., Tsafou,K., Stolte,C., O'Donoghue,S.I., Schneider,R. and Jensen,L.J. (2014) COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database*, **2014**, bau012.