Received: 9 December 2016 | Revised: 13 May 2017 | Accepted: 7 June 2017

DOI: 10.1002/prot.25330

CaMELS: *In silico* prediction of calmodulin binding proteins and their binding sites

Wajid Arshad Abbasi[1] | **Amina Asif**[1] | **Saiqa Andleeb**[2] | **Fayyaz ul Amir Afsar Minhas**[1]

[1]Biomedical Informatics Research Laboratory, Department of Computer and Information Sciences (DCIS), Pakistan Institute of Engineering and Applied Sciences (PIEAS), Nilore, Islamabad, Pakistan

[2]Biotechnology Laboratory, Department of Zoology, University of AJ&K, Muzaffarabad, AK, Pakistan

Correspondence
Fayyaz ul Amir Afsar Minhas, Biomedical Informatics Research Laboratory, Department of Computer and Information Sciences (DCIS), Pakistan Institute of Engineering and Applied Sciences (PIEAS), Nilore, Islamabad, Pakistan.
Email: fayyazafsar@gmail.com
or
afsar@pieas.edu.pk

Funding information
Wajid A. Abbasi is supported by a grant (PIN: 213-58990-2PS2-046) under indigenous 5000 Ph.D. fellowship scheme from the Higher Education Commission (HEC) of Pakistan.

Abstract

Due to Ca^{2+}-dependent binding and the sequence diversity of Calmodulin (CaM) binding proteins, identifying CaM interactions and binding sites in the wet-lab is tedious and costly. Therefore, computational methods for this purpose are crucial to the design of such wet-lab experiments. We present an algorithm suite called CaMELS (CalModulin intEraction Learning System) for predicting proteins that interact with CaM as well as their binding sites using sequence information alone. CaMELS offers state of the art accuracy for both CaM interaction and binding site prediction and can aid biologists in studying CaM binding proteins. For CaM interaction prediction, CaMELS uses protein sequence features coupled with a large-margin classifier. CaMELS models the binding site prediction problem using multiple instance machine learning with a custom optimization algorithm which allows more effective learning over imprecisely annotated CaM-binding sites during training. CaMELS has been extensively benchmarked using a variety of data sets, mutagenic studies, proteome-wide Gene Ontology enrichment analyses and protein structures. Our experiments indicate that CaMELS outperforms simple motif-based search and other existing methods for interaction and binding site prediction. We have also found that the whole sequence of a protein, rather than just its binding site, is important for predicting its interaction with CaM. Using the machine learning model in CaMELS, we have identified important features of protein sequences for CaM interaction prediction as well as characteristic amino acid sub-sequences and their relative position for identifying CaM binding sites. Python code for training and evaluating CaMELS together with a webserver implementation is available at the URL: http://faculty.pieas.edu.pk/fayyaz/software.html#camels.

KEYWORDS

calmodulin, gene ontology enrichment, large margin classification, multiple instance learning, protein binding site prediction, protein interaction prediction

1 | INTRODUCTION

Calmodulin (CaM) is a 149 amino acid long multifunctional calcium (Ca^{2+}) binding protein that is highly conserved across all eukaryotes.[1] CaM mediates many vital processes like immune response, muscle contraction, metabolism, nerve growth, and intracellular movement.[2] CaM is able to do all this by binding various targets in the cell including a large number of enzymes, ion channels and other proteins.[3,4] Many CaM binding proteins are mostly unable to bind Ca^{2+} directly and therefore use CaM as a signal transducer and calcium sensor.[5,6] Due to the involvement of CaM in different important biological processes, identification of proteins that bind CaM and the location of CaM binding sites within a protein can help biologists in elucidating underlying biological processes at the molecular level. Due to Ca^{2+}dependent binding and the large sequence diversity of its targets, identifying CaM interactions and binding sites in the wet lab is very costly and time-consuming.[7] Therefore, there is an utmost need for computational techniques to support wet-lab experiments by predicting CaM binding proteins and their binding sites. This work presents a highly accurate *in-silico* CaM binding site and interaction prediction method that relies only on protein sequences.

A number of algorithms have been proposed for CaM interaction and binding site prediction in the literature.[8–13] DeGrado et al.

suggested an algorithm that finds amphiphilic alpha helices in a peptide sequence for CaM binding site prediction.[13] Mruk et al. proposed a method called calmodulation meta-analysis for CaM binding site prediction by scoring the existence of canonical motifs in a given protein sequence.[12] Both the methods proposed by DeGrado et al. and Mruk et al. were designed using a limited dataset and cannot predict whether a protein will interact with CaM or not.[12,13] Furthermore, motif based analysis gives very low precision in predicting CaM binding proteins and their binding sites. Radivojac et al. and Hamilton et al. used a supervised machine learning approach for CaM binding site prediction based on classification of length-21 sequence windows in the protein.[8,9] These methods use a conventional Support Vector Machine (SVM) classifier and do not explicitly handle imprecisions in binding site annotations in the training data. Annotations of CaM binding sites in proteins available in the literature typically span more residues than the minimal set of contiguous residues responsible for the interaction.[10] Such imprecisions result from limitations of experimental procedures and time or cost considerations in identifying individual binding residues. Furthermore, all annotated binding site residues may not contribute equally to the binding energy. To address such uncertainties, Minhas and Ben-Hur formulated this problem as a Multiple Instance learning (MIL) problem.[10] Their approach, called MI-1, was designed primarily for CaM binding site prediction and offers very good accuracy for this task. However, the accuracy of MI-1 for CaM interaction prediction is very low. This is because MI-1 simply uses the predicted score of the most likely binding window in a protein as its CaM interaction propensity. However, a putative CaM-binding sequence in a protein will result in an interaction only if such an interaction is feasible in terms of the three-dimensional structure and energetics of the protein.[13] Furthermore, MI-1 uses a heuristics approach to solve the MIL problem which is computationally demanding and may not converge to the optimal solution of the problem.

In this article, we present a novel machine learning based CaM interaction and binding site prediction method called CaMELS: CalModulin intEraction Learning System. Unlike previous techniques that use binding site prediction for predicting CaM interactions, CaMELS models interaction and binding site prediction as two separate classification problems. For CaM interaction prediction, it uses global protein-level features instead of localized window-level features used in previous studies.[9,10] This has led to a significant improvement in the accuracy of interaction prediction in comparison to previous CaM interaction prediction methods. The accuracy and biological significance of CaM interaction predictions from CaMELS have been verified through cross-validation, Gene Ontology (GO) enrichment analysis,[14] evaluation on external validation data set and mutagenic studies.

CaMELS gives near perfect accuracy for predicting CaM binding sites. For this purpose, CaMELS uses a custom-built stochastic subgradient optimization based multiple instance machine learning model which is faster and more accurate than the heuristic algorithm used in MI-1. We have verified the biological relevance of our binding site predictions through cross-validation, evaluation on an external validation data set and mutagenic studies. We have also identified amino acid residues, their positions and motifs that are characteristic of CaM binding sites.[15] We have developed and deployed an easy to use webserver for CaMELS that can generate predictions for CaM interaction and binding sites based on protein sequences. We believe that CaMELS, with its state of the art accuracy, can be very useful to biologists in designing wet-lab experiments for identifying CaM binding proteins and their binding sites.

The rest of the article is organized as follows: We start off with a detailed description of the methods used for developing CaMELS and its performance validation in Section 2. Section 3 gives detailed results of the article and is largely independent of Section 2 so that the reader can easily focus on the major contributions of our work. Section 4 follow the results and discussion.

2 | METHODS

In this section, we give the detail of methodology adopted to develop and evaluate the performance of CaMELS.

2.1 | Dataset and preprocessing

CaMELS learns two separate machine learning models for predicting CaM binding proteins and their binding sites. For this purpose, CaMELS uses data of known CaM-binding and non-binding proteins as well as known CaM binding sites. Here, we give details of the data that has been used for training the CaMELS and evaluating its accuracy.

2.1.1 | Binding site dataset

For CaM binding site prediction, our dataset consists of a set of 157 CaM binding proteins taken from the CaM target database.[10,15] Each of these proteins has one or more annotated binding sites and a total of 191 binding sites were identified in these proteins. These proteins were selected in such a way that no two proteins have $>$40% sequence identity in overall or in regions annotated as binding sites. This non-redundancy in the data set ensures that our predictive model is not biased toward any specific class of protein sequences and is able to generate highly accurate predictions for unseen or novel proteins.

2.1.2 | Interaction dataset

For CaM interaction prediction, we used a set of 241 known CaM binding proteins from *Arabidopsis thaliana* as the positive set.[16] We used CD-HIT to obtain a non-redundant set of 12,217 proteins from the *Arabidopsis thaliana* proteome which is used as the negative set.[17] Keeping the sequence diversity of known CaM binding proteins into account, the proteins in the negative set share $<$30% sequence similarity with proteins in the positive set and $<$40% among themselves. Furthermore, these proteins share less than 40% sequence identity with proteins in the binding site data set.

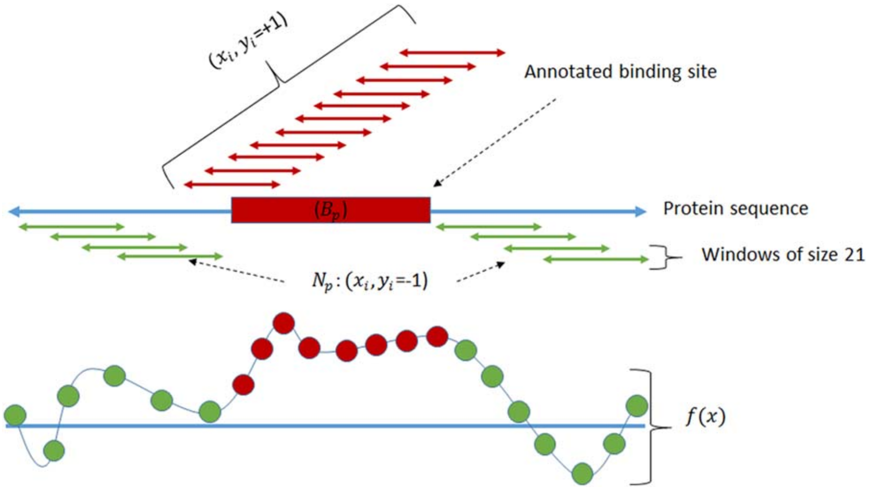


FIGURE 1 Multiple Instance Learning (MIL) Framework for CaM binding site prediction. The sequence of protein p is shown as a line whereas the annotated binding site is shown as a box. All those windows overlapping with the binding site (B_p) are positive examples (shown as lines above the protein) and the rest of the windows that do not overlap the binding site are negative examples (N_p) (shown as lines below the protein). The bottom panel illustrates characteristics of the desired discriminant function $f(x)$. The score from the trained $f(x)$ for at least one window in the binding site should be higher than the scores generated for all non-binding site windows within that protein [Color figure can be viewed at wileyonlinelibrary.com]

2.2 | Machine learning models

2.2.1 | Multiple instance learning based CaM-binding site prediction

In CaM binding site prediction, the objective is to find the region of a protein that is involved in its interaction with CaM. For this purpose, we present a novel solution based on multiple instance machine learning. It uses a sliding-window approach in which each protein sequence is divided into overlapping windows of 21 amino acids as shown in Figure 1. We represent the sequence of a window in a protein by x_i and denote its associated label by $y_i \in \{+1, -1\}$ indicating whether x_i belongs to an annotated binding site ($+1$) or not (-1). This problem can be posed as a classification problem through a discriminant function $f(x) = \mathbf{w}^T \phi(x)$, where $\phi(x)$ is a feature vector representation of window x and $\mathbf{w}$ is the weight vector. The weight vector needs to be learned based on available labeled data such that it produces higher scores for CaM binding site windows in comparison to non-binding site regions so that we can identify CaM binding sites for proteins not in the training set. Residues involved in the binding of a protein with CaM can be identified based on the values of the discriminant function $f(x)$ for the window centered at these residues (see Figure 1).

We have solved this classification problem using a conventional support vector machine (SVM)[18] as well as the multiple instance learning (MIL) framework.[19] We use the conventional SVM as a baseline for our results by taking the annotated binding site windows in a protein as positive class examples and the remaining as negatives.[9,10] As discussed in the introduction section, the annotated CaM binding sites in the binding dataset are imprecise due to limitations in experimental procedures and include residues that may not be involved in binding. A classical supervised classification approach such as an SVM cannot be used effectively with such ambiguously labeled training examples.[19] To cope with these challenges we formulated the binding site prediction problem as a MIL problem.[10]

MIL is a form of supervised learning where labels are available for bags or sets of examples and not for individual examples.[19,20] A number of methods for solving the MIL problem exist in the literature.[10,20,21] Heuristic approaches were adopted to solve the MIL problem in mi-SVM and MI-1 techniques proposed by Andrews et al. and Minhas and Ben-Hur, respectively.[10,20] In this work, we have improved the MI-1 formulation further using a stochastic sub-gradient algorithm for MIL inspired from the Pegasos solver for conventional binary SVMs by Shalev-Shwartz et al.[22] Table 1 shows the pseudocode of the proposed algorithm. The details of formulating CaM binding site prediction as a MIL problem and its solution using stochastic

TABLE 1 MIL algorithm with SSGO training for CaM binding site prediction

```
Inputs: $\lambda$ (Regularization parameter), $T$ (Maximum number of iterations)
Initialize: set $\mathbf{w}_0 = \mathbf{0}$
  For $t = 1,\ 2, \ldots,\ T$
    Select a protein p uniformly at random from the binding site dataset
    $i^* = argmax_{i \in B_p}\ \mathbf{w}_t{}^T \phi(x_i)$ //Maximum scoring binding site window
    $j^* = argmax_{j \in N_p} \mathbf{w}_t{}^T \phi(x_j)$ //Maximum scoring non-binding site window
    If $\mathbf{w}_t{}^T \phi(x_{i^*}) - \mathbf{w}_t{}^T \phi(x_{j^*}) < 1$: //in case of margin violation
      Set $\mathbf{w}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right)\mathbf{w}_t + \frac{1}{\lambda t}\left(\phi(x_{i^*}) - \phi(x_{j^*})\right)$
    else:
      Set $\mathbf{w}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right)\mathbf{w}_t$
Output: $\mathbf{w} = \ \mathbf{w}_{T+1}$
```

sub-gradient optimization (SSGO) is given in the supplementary material. The proposed algorithm is significantly simpler and faster in comparison to existing methods.

2.2.2 | Interaction prediction

In CaM interaction prediction, the objective is to predict whether a given protein interacts with CaM or not. For a protein p in the interaction dataset, we indicate its associated feature representation by $\psi(p)$. All proteins in the interaction dataset have binary labels indicating whether they interact with CaM ($+1$) or not (–1). This problem can be posed as a classification problem through a discriminant function $z(p)$ which must produce a high score for CaM binding proteins in comparison to non-interacting ones. We experimented with two different functional forms of this discriminant function:

Discriminant function scoring (DFS)

The trained classifier for CaM binding site prediction can be used for CaM interaction prediction. This can be done by using the score of the most likely binding site in the protein as the interaction propensity of that protein. The fundamental assumption behind this formulation is that the presence of a binding site within a protein is predictive of its interaction with CaM. Mathematically, the CaM interaction score for a protein p is given by $z(p)=\max_{x\in p} f(x)$ where $f(x)$ is the discriminant function used in binding site prediction. This approach was used in previous studies to predict CaM interactions of proteins in the *A. thaliana* proteome and is used here as a baseline.[9,10]

SVM with protein level features

The discriminant function scoring scheme assumes that if a protein contains a CaM binding sub-sequence then that protein interacts with CaM. We hypothesize that this may not always be the case. This is plausible especially when a protein contains a putative CaM-binding region but that region is either structurally or energetically inaccessible for interaction with CaM.[23] This led us to use features $\psi(p)$ extracted from the complete sequence of a protein p with a standard binary SVM for predicting its interaction with CaM.[18] The detail of selecting hyper-parameters for different classifiers is given in the supplementary material.

2.3 | Feature extraction

We have used two different categories of features for CaM binding site and interaction prediction.

i Window level features $\phi(\cdot)$: These features are extracted from overlapping sequence windows of length 21 in a protein. These features are used primarily for CaM binding site prediction and also for discriminant function scoring based CaM interaction prediction.

ii Whole protein features $\psi(\cdot)$: In contrast to window level features, these features are extracted from the whole of the protein sequence. These features are used for CaM interaction prediction only.

The details of each of these feature representations are given below.

2.4 | Window level features

For window level features of protein sequences, we sweep a window of length 21 across the entire length of the protein and extract features from individual windows. The following different types of window level feature representations are used.

2.4.1 | Amino acid composition features (AAC)

This feature representation captures the composition of a sequence window x by counting occurrences of different amino acids in it. The fundamental assumption underlying the use of these features is that amino acid composition of a window is predictive of its ability to interact with CaM. The amino acid composition of a given sequence widow x is a 20-dimensional vector $\phi_{AAC}(x)$ containing the counts of occurrences of the 20 amino acids in x.

2.4.2 | Position dependent composition features (PDC)

AAC captures composition of amino acids irrespective of their position in a sequence. However, the relative position of an amino acid within a sequence window can play an important role in binding site prediction. For this purpose, we use Position Dependent Composition (PDC) feature representation. PDC represents a sequence window x as a 420-dimensional vector $\phi_{PDC}(x)$ such that its component $\phi_{PDC}^{a,k}(x)$ is set to 1.0 if amino acid $a\in\{A,\ C,\ldots,Y\}$ occurs at position $k\in\{1,\ 2,\ldots,\ 21\}$ in the window.

2.4.3 | Position dependent BLOSUM-62 features (PD-blosum)

To model the substitutions of physio-chemically similar amino acids in sequence windows, we expressed each window using the BLOSUM-62 substitution matrix.[24] In contrast to AAC and PDC, this representation not only captures composition and position information amino acids but also models their physio-chemical similarity. The 420-dimensional PD-Blosum feature vector $\phi_{PD-Blosum}(x)$ for a sequence window x is obtained by stacking the columns of the BLOSUM-62 matrix corresponding to each residue in the sequence window x.

2.4.4 | Averaged blosum-62 features (blosum)

This 20-dimensional feature representation $\phi_{Blosum}(x)$ is built by averaging the columns of the BLOSUM-62 matrix corresponding to all amino acids occurring in the sequence window x.[24] The averaging across all amino acids marginalizes the effect of position in the sequence window.

2.4.5 | Propy features (propy)

To get sequence-derived structural features such as pseudo-amino acid compositions (PseAAC), autocorrelation descriptors, sequence-order-coupling number, quasi-sequence-order descriptors, amino acid composition, transition and to capture the distribution of various biophysical properties of amino acids, we used a feature extraction package called propy.[25–27] In contrast to AAC, PDC, and PD-Blosum, this feature representation not only captures composition, position and similarity but also the order of amino acids in a sequence and the correlation in their physiochemical properties. This representation gives a

$1,537$-dimensional feature vector $\phi_{propy}(x)$ for a sequence window x. Each of these features is standardized to have zero mean and unit standard deviation across all examples.

2.4.6 | Position dependent gappy triplet (PDGT)

To identify sequence motifs that are important for CaM binding site prediction, we use Position Dependent Gappy Triplet (PDGT) feature representation. This feature representation quantifies the presence of gapped-triplet motifs of the form $a[\cdot]^m b[\cdot]^n c$ at a given location k in the sequence window. Here, a, b and c are amino acids and m and n specify the number of don't-care positions or gaps in the motif. It gives a 3,800,000-dimensional vector $\phi_{PDGT}(x)$ for each window x such that its component $\phi_{PDGT_{a,b,c,k}^{m,n}}(x)$ is set to one if the sub-sequence $a[\cdot]^m b[\cdot]^n c$ occurs at position k in x and zero otherwise. Values of m and n vary from 0 to 4.

2.5 | Protein level features

Here, we describe the features extracted from the whole protein sequences.

2.5.1 | Protein level amino acid composition (AAC)

We have used AAC at the protein level to predict CaM interactions. Protein Level Amino Acid Composition (AAC) of a given protein sequence p is a 20-dimensional vector $\psi_{AAC}(p)$ containing the counts of occurrences of the 20 amino acids in p.

2.5.2 | Protein level averaged blosum-62 features (blosum)

Similar to the BLOSUM-62 features at the window level, we use these features at the protein level to model the substitutions of physiochemically similar amino acids in a protein sequence. This 20-dimensional feature representation $\psi_{Blosum}(p)$ is built by averaging the columns of the BLOSUM-62 matrix corresponding to all amino acids occurring in the given protein sequence p.[24] The averaging across all amino acids marginalizes the effect of position in the protein sequence.

2.5.3 | Protein level propy features (propy)

We have also used propy feature representation at the protein level to predict CaM interactions. This representation gives a 1,537-dimensional feature vector $\psi_{propy}(p)$ for a protein sequence p.

2.5.4 | Smith-Waterman alignment features (SW)

Proteins with similar sequences are expected to have similar functions.[28] Using this idea, we have developed a local alignment based feature representation that measures the sequence similarity of a protein with those in the CaM binding site dataset. This feature representation assumes that the alignment of a protein sequence to known CaM binders is predictive of its interaction with CaM. To capture sequence similarities of a given protein to known CaM binders, we performed local sequence alignment of a protein with the 157 proteins in our binding site dataset. For this purpose, we used SWIPE with the BLOSUM-62 substitution matrix and gap insertion and extension penalties of 10 and 0.5, respectively.[29,30] This results in a 157-dimensional feature vector $\psi_{SW}(p)$ of alignment scores for a protein sequence p.

2.6 | Performance evaluation

Here, we discuss the evaluation procedure used to assess the performance of various machine learning models. A machine learning model should generalize well on unseen data to reliably assist a biologist for the design of wet-lab experiments.[31] Therefore, we used different strategies to evaluate the generalization performance of CaMELS. The criteria used to measure the generalization power of CaMELS both for interaction and binding site prediction are given below.

2.7 | Accuracy of interaction prediction

To evaluate the performance of CaMELS for CaM interaction prediction, we used a number of different analyses. Specifically, we test the biological relevance of our results via motif-based analysis, Gene Ontology enrichment analysis, evaluation on external validation dataset and *in silico* mutation analysis.[15,32,33] We also describe a number of machine learning centric metrics for cross-validation based performance evaluation.

2.7.1 | Motif-based analysis

One commonly used approach for detecting CaM interacting proteins is to search for known CaM binding motifs from the CaM target database[12,15] in a query protein. Occurrence of such a motif in a protein is then used as evidence of its interaction with CaM. We have used this motif search based CaM interaction prediction technique as a baseline in our study.

2.7.2 | Gene ontology enrichment analysis

To check if our top predicted CaM binding proteins are functionally similar to known CaM binding protein, we performed Gene Ontology (GO) term enrichment analysis.[14] We used the GOrilla tool for Gene Ontology (GO) term enrichment analysis of known CaM binders and top scoring *A. Thaliana* proteins from different interaction prediction methods.[33] To quantify the degree of correspondence of GO terms between predicted and known CaM binding proteins, we used the Jaccard Index.[34] It measures the percentage of Gene Ontology terms common between the 240 known CaM binders and the top 240 predicted CaM binding proteins from the *A. Thaliana* proteome. The value of the Jaccard index for an ideal predictor should be equal to 100%.

2.7.3 | Evaluation on independent validation dataset

We have also evaluated the performance of CaMELS for CaM interactions prediction on novel proteins by using an independent validation dataset. This dataset contains five protein from CaM complexes in the Protein Data Bank as the positive set (PDB IDs are: 1NWD, 1SY9, 2MOK, 5DOW, and 1YRT).[35–39] As negative (non-interacting) set, we used CD-HIT to randomly select 250 non-redundant from the non-redundant PDB database.[17,40] Proteins in the positive and negative sets share <30% sequence identity with proteins in interaction and binding site datasets used in training our machine learning models. We predicted the CaM-interaction propensity for these proteins using CaMELS. We have used the complete Uniprot[41] sequence of CaM binding proteins in the positive set to validate the performance of CaMELS for CaM binding site prediction.

2.7.4 | *In silico* mutation analysis

To evaluate the performance of CaMELS for targeted mutations in CaM interacting proteins, we considered two mutagenic studies from the literature by Pley et al. and Li et al.[42,43] Pley et al. conducted a mutagenic study to identify CaM binding sites in clathrin light chains (LCa) which are responsible for transport of macromolecules between membrane-bound compartments.[43] Similarly, Li et al. performed targeted mutagenic study on Suppressor of Gene Silencing 3 (NbSGS3) of *N. benthamiana* to locate domains involved in CaM interaction.[42]

2.7.5 | Cross-validation

Cross-validation is a statistical technique to quantify the generalization performance of machine learning models by dividing the available dataset into training and testing.[32] For CaM interaction prediction, we have used 10-fold stratified cross-validation.[44] In 10-fold stratified cross-validation, the negative and positive examples from interaction dataset are randomly partitioned into 10 equal-sized subsets. From these 10 subsets, 9 subsets are used as training data for the machine learning model and the remaining subset is used as validation data to test the trained model. This process is then repeated 10 times (folds), with each of the 10 subsets used exactly once as the validation data. The fold-wise average of the following performance metrics has been used to quantify the performance of different models.

Area under ROC curve (AUC-ROC)

The Receiver Operating Characteristic (ROC) curve is obtained by plotting true positive rate versus false positive rate for different thresholds on scores generated from the machine learning model. The area under the ROC curve (AUC-ROC) is a metric used to determine the performance of the predictor. An ideal predictor will give AUC-ROC of 100% whereas random guessing will have a score of 50%.[45]

Area under 10% ROC curve (AUC-ROC$_{0.1}$)

AUC-ROC$_{0.1}$ is obtained by plotting true positive rate (TPR) in ROC curves up to the first 10% false positives. This measure gives us a sense of how many true positives are produced at low false positive rates.[10,31]

Area under precision-recall curve (AUC-PR)

The precision-recall (PR) curve is obtained by plotting precision against recall at different thresholds for the discriminant function values of a machine learning model. The area under the PR curve (AUC-PR) is a useful metric to check the performance of a predictor in cases in which the number of positive examples is much smaller than negative ones.[45] An ideal predictor will give AUC-PR of 100%.

2.8 | Accuracy of binding site prediction

To assess the performance of CaMELS for CaM binding site prediction we used motif-based analysis, cross-validation, evaluation on independent validation dataset and *In silico* mutation analysis.[15,32] For evaluation on independent validation dataset and *In silico* mutation analysis, we used the same validation dataset and mutagenic studies as the one used for analyzing interaction prediction accuracy. Therefore, we only discuss motif-based analysis and cross-validation with respect to binding site prediction.

2.8.1 | Motif-based analysis

As a baseline for comparison, we used a motif based approach in which the location of a known CaM binding site motif in a CaM binding protein is used as a prediction for its CaM binding site. For this purpose, we used motifs from CaM target database.[12,15]

2.8.2 | Cross-validation

For CaM binding site prediction, we have used Leave One Protein Out (LOPO) cross-validation. In this protocol, the classifier is tested on all residue-level windows of a protein after training it on the sequence windows from all other proteins. This process is repeated for all the proteins in the binding site dataset. In addition to AUC-ROC, AUC-ROC$_{0.1,}$ and AUC-PR, we report the following biologist-centered performance metrics as well.[31]

True hit rate (THR)

THR is the percentage of proteins in the binding site dataset in which the top scoring residue predicted by a classifier lies in an annotated binding site. A high value of THR implies that the chances of the top scoring window predicted by a machine learning model is indeed part of a CaM-binding site. An ideal classifier would have THR = 100%.

False hit rate (FHR)

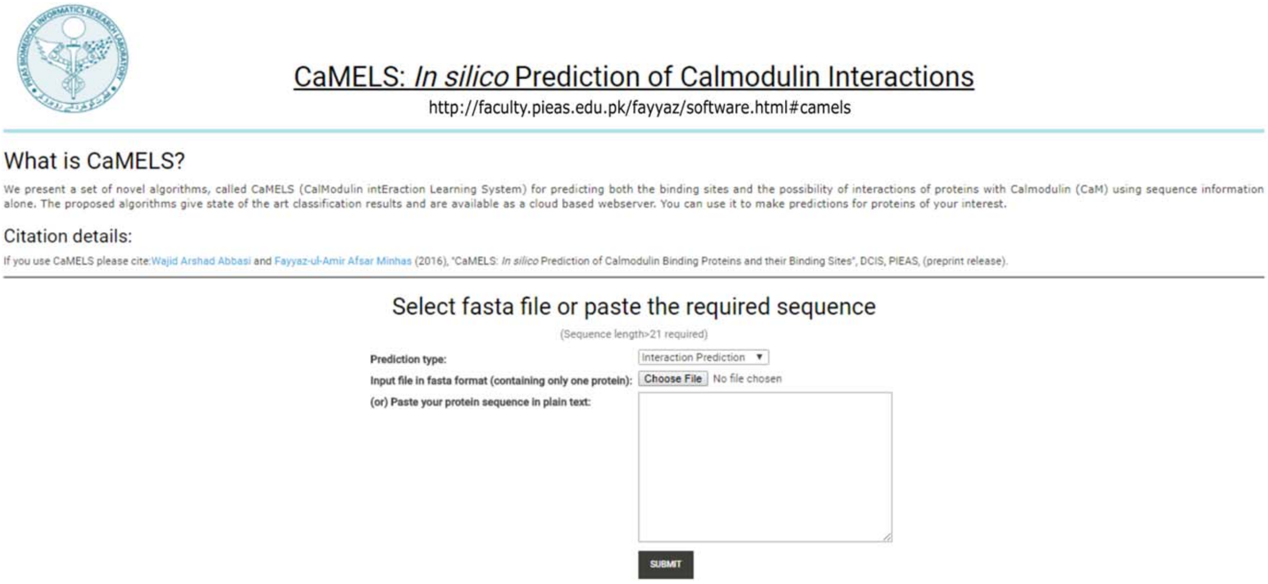
FHR is the percentage of non-binding site windows that score higher than the highest scoring window in the annotated binding site of a protein. The lower the value of FHR, the lesser is the number of non-binding windows which have a score higher than a window in the true binding site. An ideal classifier would have FHR = 0%.

Rank of the first positive prediction (RFPP)

This metric gives the distribution of the rank of the top true positive prediction across all proteins. In comparison to AUC-PR, this performance metric is more useful to biologists as it reveals the expected number of mutagenic experiments required for identifying the CaM-binding site in a protein based on the ranked list its window-level prediction scores. It is formally defined as follows: $RFPP\ (p){=}q$, if p% of proteins tested with a predictive model have at least one true binding site among the top q predictions from the model. An ideal predictor should have $RFPP\ (100)\ =1$, that is, for all proteins, the top scoring window is always a part of the true CaM binding site.[31,46]

2.9 | Webserver for CaMELS

We have developed and deployed a webserver of CaMELS which uses the optimal machine learning model for CaM binding and binding site prediction. This webserver takes a query protein sequence in plain or fasta format and performs CaM interaction and binding site prediction for it. The user interface of the CaMELS webserver is shown in



CaMELS: *In silico* Prediction of Calmodulin Interactions

http://faculty.pieas.edu.pk/fayyaz/software.html#camels

What is CaMELS?

We present a set of novel algorithms, called CaMELS (CalModulin intEraction Learning System) for predicting both the binding sites and the possibility of interactions of proteins with Calmodulin (CaM) using sequence information alone. The proposed algorithms give state of the art classification results and are available as a cloud based webserver. You can use it to make predictions for proteins of your interest.

Citation details:

If you use CaMELS please cite:Wajid Arshad Abbasi and Fayyaz-ul-Amir Afsar Minhas (2016), "CaMELS: *In silico* Prediction of Calmodulin Binding Proteins and their Binding Sites", DCIS, PIEAS, (preprint release).

Select fasta file or paste the required sequence

(Sequence length>21 required)

Prediction type: Interaction Prediction ▼

Input file in fasta format (containing only one protein): Choose File No file chosen

(or) Paste your protein sequence in plain text:

FIGURE 2 The user interface of the CaMELS webserver. A user can submit fasta file or plain sequence of a protein of interest for CaM interaction and binding site prediction [Color figure can be viewed at wileyonlinelibrary.com]

Figure 2. After the successful submission of a protein sequence, the users are redirected to a page showing CaMELS predicted scores for CaM interaction and binding site predictions. For CaM interaction prediction, the predicted score shows the interaction propensity of the submitted protein with CaM. The higher the score, more chances for a protein to be CaM interacting. Similarly, for CaM binding site prediction, residue level scores of all windows along with the predicted binding site location and its score are shown. A plot of residue level scores of all windows for binding site prediction with the location of the predicted binding site is also shown on this page. The webserver is available at the following URL: http://faculty.pieas.edu.pk/fayyaz/software.html#camels.

3 | RESULTS AND DISCUSSION

In this section, we present and discuss the results and major outcomes of our study for CaM interaction and binding site prediction.

3.1 | Interaction prediction

The results of various analyses for interaction prediction are given in Tables 2 and 3 and Figures 3 and 4. We begin by listing the major contributions of our work for CaM interaction prediction before presenting detailed results.

- Our proposed CaM binding prediction method, CaMELS, gives significant improvement in terms of accuracy of CaM interaction prediction in comparison to the previous state of the art CaM interaction predictors (AUC-PR of 58.3% vs. 14.8% for MI-1[10]).
- For predicting CaM interactions, CaMELS uses information extracted from the whole protein rather than from the most likely CaM binding site in a protein as done in previous approaches (DFS and MI-1).[9,10] We have identified that features from the whole sequence of a protein are more predictive of CaM interactions in comparison to localized widow-level features (AUC-PR of 55.0% vs. 26.0% for DFS). This seems to suggest that the whole sequence of a protein is

TABLE 2 Interaction prediction results for all models

Method	Features	AUC-PR	AUC-ROC	AUC-ROC$_{0.1}$
CaMELS	Protein level Propy (propy)	**55.0**	**86.7**	**65.1**
	Smith-Waterman Alignment (SW)	40.2	78.4	51.9
	Protein level Amino Acid Composition (AAC)	26.8	74.7	40.4
	Protein level averaged Blosum-62 (Blosum)	11.4	78.4	32.6
DFS	Position Dependent Blosum-62 (PD-Blosum)	6.0	68.0	18.0
	Averaged Blosum-62 (Blosum)	4.0	74.0	26.0
	Amino Acid Composition (AAC)	4.0	72.0	24.0
	Position Dependent Composition (PDC)	3.0	69.0	16.6
	Combination of AAC and PDC	2.6	71.0	17.0

TABLE 3 Gene Ontology term enrichment analysis results

	Jaccard Index		
Method	Process	Function	Component
CaMELS	**34.0**	**68.0**	**40.0**
MI-1	1.0	0.0	0.0

required for the prediction of protein interaction and the mere presence of a putative CaM-binding region in a protein is not a good predictor of its interaction with CaM.

- CaMELS performs significantly better than motif-based search for CaM interacting proteins. This corroborates with our earlier finding that the whole sequence of a protein is required for predicting CaM binding. Motif-based search is very ineffective for CaM interaction prediction due to the high sequence diversity of CaM binding proteins.
- Gene Ontology (GO) term enrichment analysis of predicted CaM binding proteins from CaMELS shows significant overlap of biological processes, cellular functions and localization with known CaM binders.
- Our evaluation on an external validation dataset shows that CaMELS can accurately predict CaM binding proteins even if they share no significant sequence similarity to proteins in the training data set.
- We also demonstrate that CaMELS can be used to predict CaM interacting proteins in species other than *A. Thaliana* on which it has been trained.
- Our experiments also show that changes in prediction scores from CaMELS due to *in silico* mutations in proteins correlate very well with experimentally observed behavior of CaM binding proteins in two different experimental mutagenic studies.
- We have evaluated the performance of different types of features for CaM interaction prediction and report the importance of individual features for this task. Specifically, we have found that features that quantify the correlation between physiochemical properties of different regions of a protein are highely predictive of its interaction with CaM.
- We have also developed and deployed a cloud based web srever for CaM Interaction prediction through CaMELS.

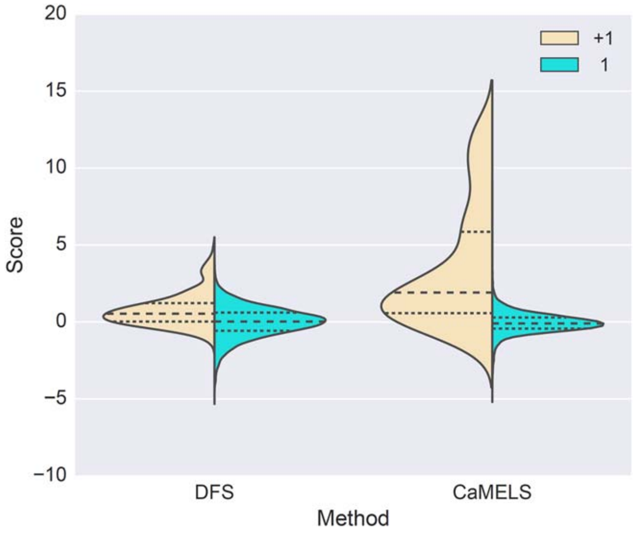


FIGURE 4 Violin plot showing the predictive performance of CaMELS and DFS. Density distributions of CaM interacting (+1) and non-interacting (−1) proteins for DFS and CaMELS scores are shown. Dotted lines show the first, second and third quartiles of these densities [Color figure can be viewed at wileyonlinelibrary.com]

In what follows, we report the results in support of our major outcomes for CaM interaction prediction.

3.1.1 | Improvement in CaM interaction prediction results

The cross-validation results of CaMELS over different feature representations are given in Table 2, Figures 3 and 4. The results are

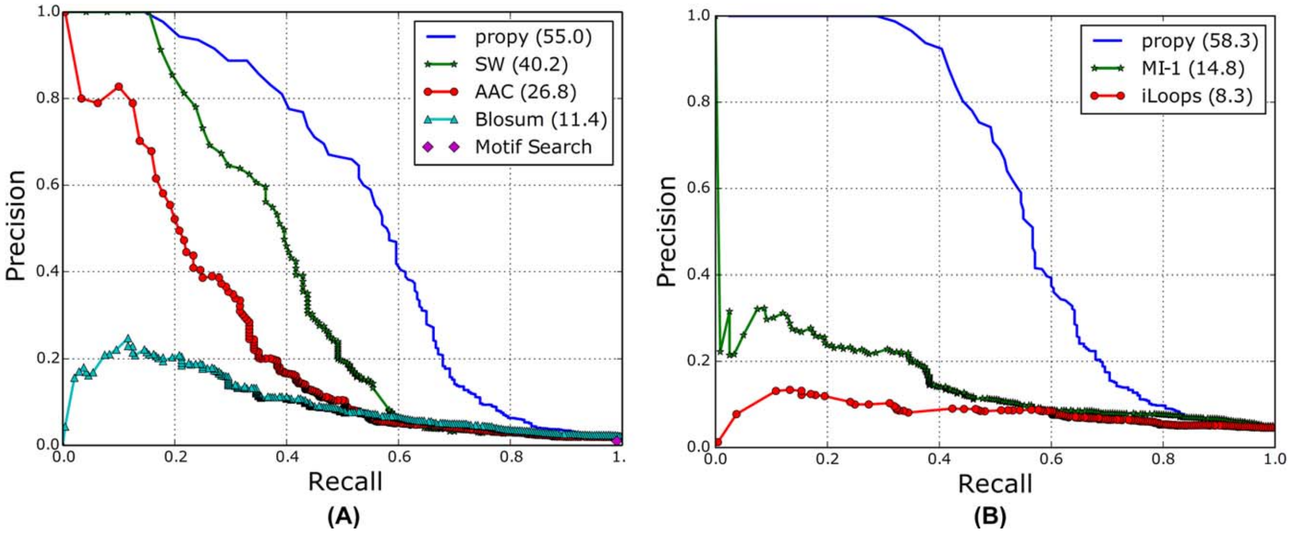


FIGURE 3 **(A)** Precision-recall (PR) curves for interaction prediction across all models; **(B)** Precision-recall curves for the comparison of CaMELS with MI-1 and iLoops. The mean area under the curve is shown in parenthesis across different folds [Color figure can be viewed at wileyonlinelibrary.com]

shown in comparison to other CaM interaction prediction techniques such as Discriminant Function Scoring (DFS), MI-1,[10] iLoops[47] and Motif based search.[15] These results show a large improvement in predictive performance of CaMELS in comparison to cascade classification in MI-1 and DFS over the same dataset and evaluation protocol.[10] By using propy features with SVM in CaMELS, we obtained a maximum AUC-PR of 55% (see Table 2 and Figure 3A). This is significantly better in comparison to the DFS approach which gives an AUC-PR of only 6.0%. These quantitative results indicate that our proposed scheme can identify CaM binding proteins with significantly less false positives in comparison to existing methods. This can help a biologist in rapidly identifying putative CaM-binding proteins at the proteome level for further analysis in the wet lab. We have also observed a similar trend of improvement in accuracy in terms of AUC-ROC where CaMELS scores an AUC-ROC of 86.7% whereas DFS gives an AUC-ROC of 74.0% (Table 2). A marked increase is also noted in $AUC_{0.1}$ from 26.0% to 65.1%. This is also evident in Figure 4 which shows the densities of the prediction scores obtained from the DFS and SVM-based methods for the interacting and non-interacting proteins in our evaluation dataset. It can be easily noticed that the degree of overlap between distributions of scores of interacting and non-interacting proteins is significantly larger for DFS in comparison to CaMELS (see Figure 4). This plot shows that, with CaMELS, >50% of CaM binding proteins in our target database score higher than the highest scoring non-interacting protein. On the other hand, with DFS, >3000 out of ~12000 non-interacting proteins score higher than 50% of known CaM binding proteins.

We have also compared the CaM interaction prediction performance of CaMELS with the previous state of the art method (MI-1) and a general purpose protein interaction predictor called iLoops.[10,47] The PR curves for this comparison using the same dataset and evaluation protocol are shown in Figure 3B. CaMELS performs significantly better than both MI-1 and iLoops on a reduced data set of 5000 randomly sampled non-interacting proteins. This reduction was done due to the limitation of the iLoops server. CaMELS gives an AUC-PR of 58.3% whereas, the AUC-PR obtained for MI-1 and iLoops are 14.8% and 8.3%, respectively (see Figure 3B). These results clearly show that CaMELS outperforms all existing methods that can be used for predicting CaM interactions.

3.1.2 | Motifs from CaM target database fail to predict CaM interactions

We can use the set of known CaM-binding motifs in the CaM target database to identify CaM binding proteins through a simple motif search. We have found that motif-based analysis gives poor performance in comparison to discriminative machine learning based models and is unable to predict CaM interacting proteins accurately. Motif-search for CaM interaction gives near-perfect recall but has the very low precision (~1.0%) as shown in Figure 3A. This shows that predicting CaM interacting proteins based on the mere presence of CaM-binding motif produces a large number of false positives. This can be explained by the presence of CaM-binding motifs in protein sequences that do not bind CaM because the motif may not be structurally or energetically accessible for interaction.[23] These findings are in agreement with the literature.[48,49] The low performance of motif-based search for CaM interaction prediction can be further explained by the high sequence variation among CaM binding proteins.[12]

3.1.3 | Whole protein sequence plays a role in CaM interaction prediction

We have experimented with different classification schemes (CaMELS and DFS) and feature representations of proteins for prediction of CaM interactions. For predicting CaM interacting proteins, CaMELS uses features extracted from the whole protein sequence. On the other hand, DFS and MI-1 first employ a machine learning to identify the most likely CaM binding window in a protein and then uses it to predict whether the protein binds CaM or not. The quantitative cross-validation results for these techniques are shown in Table 2. These results show an interesting trend: CaMELS, which uses features extracted from the whole of a protein sequence, performs much better than DFS which is based on information contained in the most likely CaM binding window in a protein. The maximum AUC-PR for the DFS approach is 6.0% whereas CaMELS gives a maximum AUC-PR of 55.0%. This holds true even when comparing the performance of the same type of features at the whole-protein and window levels. For example, amino acid composition (AAC) of the whole protein gives an AUC-PR of 26.8% with CaMELS whereas DFS gives an AUC-PR of only 4.0% with the same features. These empirical results suggest that information contained in the whole of the protein is significantly more predictive of CaM interaction than the most likely CaM binding window in a protein. This can potentially be explained by the fact that the mere presence of a CaM binding site in a protein is not predictive of its interaction with CaM as the binding site may not be structurally or energetically accessible for interaction.[13,50,51] All these findings corroborate with our hypothesis that the whole of a protein sequence plays an important role in CaM interaction prediction.

3.1.4 | Gene ontology enrichment analysis

Gene Ontology (GO) analysis gives information about the function and localization of gene products.[14] A high correspondence between gene ontology terms of predicted and known CaM binding proteins indicates the accuracy of our predictor. We verified the biological significance of our results and the ability of CaMELS to predict CaM-binding proteins at the proteome level by gene ontology (GO) term enrichment analysis of the top 241 predictions from CaMELS from the proteome of *A. Thaliana*. The ranked list of these proteins together with their binding sites is given in supplementary material (see Supporting Information Table S1). Table 3 shows the results of this analysis in terms of the Jaccard index which quantifies the degree of overlap between the gene ontology terms of known and predicted CaM binding proteins. We observed a significant overlap between the GO terms of known and top scoring CaM binding proteins for molecular function and biological process ontologies with Jaccard Index scores of 68% and 34%, respectively (see Table 3). This is a significant improvement in comparison to MI-1.[10] GO term enrichment analysis of the top predictions from DFS do not overlap with the enriched GO terms of known CaM binders (Table 3). The enriched terms include phosphorylation, signal transduction, signaling, kinase activity, and so forth which correlate with known functions of CaM binders.[52] Furthermore, proteins at ranks 3 and 6 the list have been predicted by the Gene Ontology Consortium to likely bind CaM as well. These proteins are Calcium-dependent protein kinase 10 (UniProt id: Q9M9V8) and CDPK-related kinase 8 (UniProt id: Q9FX86).[41]

3.1.5 | Performance validation on novel proteins

To assess the generalization accuracy of CaMELS for CaM interaction prediction on novel proteins, we performed a validation on an independent dataset. This dataset contains 5 known CaM binding proteins as the positive set and 250 proteins selected at random from non-redundant PDB database as the negative set. Proteins in this data set share no significant sequence similarity to proteins in our training data. The CaM-binding proteins in this set are: adenylate cyclase (UniProtKB AC: P0DKX7; PDB id: IYRT) from *Bordetella pertussis*,[35] Glutamate decarboxylase (UniProtKB AC: Q07346; PDB id: 1NWD) from *Petunia hybrid*,[36] Cyclic nucleotide-gated olfactory channel (UniProtKB AC: Q03041; PDB id: 1SY9) from *Bos Taurus*,[37] Chloride anion exchanger (UniProtKB AC: Q9WVC8; PDB id: 5DOW) from *Mus musculus*[38] and Cyclic nucleotide-gated olfactory channel (UniProtKB AC: Q00195; PDB id: 2M0K) from *Rattus norvegicus*.[39] The proteins in the negative validation set are also taken from a variety of different species. The complete list of all proteins in this data set is given in the supplementary material. We obtained the interaction prediction score for all proteins in this validation set through the CaMELS web-server. Out of the 5 CaM binding proteins, CaMELS predicts 4 proteins correctly as CaM interacting with their prediction scores higher than the highest scoring protein in the negative set. The higher scoring CaM binding proteins are: Cyclic nucleotide-gated olfactory channel, Glutamate decarboxylase, Cyclic nucleotide-gated olfactory channel and Chloride anion exchanger. As proteins in the independent validation set have very low sequence identity with any of the protein in our training set, therefore, these results clearly show the high generalization performance of CaMELS on novel proteins. Furthermore, none of the proteins in our independent validation set belongs to *A. thaliana* on which CaMELS training is based. Therefore, these results on the validation set also support the ability of CaMELS to predict CaM interactions correctly across different organisms at the proteome level.

3.1.6 | In silico mutation analysis

To assess whether CaMELS can be used for predicting the effect of mutations in CaM binding proteins, we used two mutagenic studies by Pley et al. and Li et al.[42,43] In these studies deletions were performed on CaM binding proteins clathrin light chains (LCa) and Suppressor of Gene Silencing 3 (NbSGS3) which led to inhibition of their ability to bind CaM. For *in silico* mutation analysis, we get CaM interacting scores through CaMELS before and after these mutations for both proteins. We observed that the prediction scores of mutated proteins are lower than their wild-type versions in both cases (0.77 to 0.66 for LCa and 0.5 to 0.3 for NbSGS3). These results clearly demonstrate the use of CaMELS for *in silico* mutation analysis of CaM binding proteins.

3.1.7 | Feature analysis

We have used different protein level feature representations for CaM interaction prediction such as propy, Smith-Waterman alignment scores, Amino Acid Composition, and Blosum features. We obtained the best performance using propy features in comparison to other feature representations (Table 2; Figure 3A). We expect this to be a consequence of incorporating k-mer features and correlation factors in the protein chain.[27] We tested this hypothesis by taking different combinations of propy features. With the 20-dimensional amino acid composition and 400-dimensional dipeptide frequency features, we obtained an AUC-PR of 47.0% whereas the 720-dimensional sequence correlation features (normalized Moreau-Broto autocorrelation, Moran autocorrelation, Geary autocorrelation, sequence-order-coupling number)[27] alone produced an AUC-PR of 52.3%. Please note that using all the 1,537 propy features results in an AUC-PR of 55.0%. This shows that the auto-correlation features of physiochemical properties of amino acid present in the protein sequence are responsible for the improvement in prediction accuracy. These relatively high performance of these sequence correlation features seems to suggest that CaM interaction is possible only when there are certain regions in a protein whose physiochemical properties are correlated with each other.

3.2 | Binding site prediction

We first list the major outcomes of our work for predicting CaM binding sites in proteins before presenting detailed results.

- CaMELS predicts binding sites in CaM binding proteins with near perfect accuracy.
- CaMELS offers significant performance improvement in comparison to the previous state of the art CaM binding site prediction approaches (AUC-ROC$_{0.1}$: 80.2% vs 59.0% for MI-1[10]). Consequently, the amount of time and effort required for wet-lab experiments to identify CaM binding sites can be significantly reduced.
- Our experiments show that CaMELS can identify CaM binding sites in proteins which are significantly different from the set of proteins used for its training.
- We have found that position-specific information related to substitution of physiochemically similar amino acids plays an important role in predicting CaM binding sites.
- Based on our machine learning model, we report the contribution of different amino acids and their relative positions in a protein for locating CaM binding sites.
- Using specialized gappy-triplet sequence descriptors in our machine learning model, we have identified sub-sequences or motifs which are characteristic of CaM binding sites. These motifs exhibit significant agreement with previously known CaM-binding motifs.

TABLE 4 Binding site prediction results for all models. AUC-PR was not available for MI-1, mi-SVM and SVM

Method	Features	AUC-PR	AUC-ROC	AUC-ROC$_{0.1}$	THR	FHR
CaMELS	Position Dependent Blosum-62 (PD-Blosum)	**87.0**	**99.1**	**80.2**	**77**	**1.0**
	Combination of AAC and PDC	85.6	98.9	77.6	75	1.0
	Position Dependent Gappy triplet (PDGT)	85.4	99.04	78.0	74	1.0
	Position Dependent Composition (PDC)	84.1	98.4	76.2	72	2.0
	Propy Features (propy)	81.2	98.0	74.7	68	2.0
	Amino Acid Composition (AAC)	80.7	97.8	72.3	68	2.0
MI-1	Combination of AAC and PDC	–	96.9	59.0	75	1.2
mi-SVM	Combination of AAC and PDC	–	96.2	55.6	68	1.9
SVM	Combination of AAC and PDC	–	95.9	55.1	65	2.1

- Through cross-validation experiments, we demonstrate that the improved performance of CaMELS for CaM binding prediction is a consequence of the use of multiple instance learning and our novel algorithm for solving the multiple instance machine learning problem.
- We have developed and deployed a webserver that can be used to identify CaM binding regions in proteins.

3.2.1 | Improvement in accuracy of CaM binding site prediction

Table 4 and Figure 5 show the results of CaMELS for binding site prediction across different features using leave-one-protein-out (LOPO) cross-validation. Table 4 also shows the results previous state of the art classification schemes (SVM, mi-SVM, MI-1) using the same dataset and evaluation protocol.[10] To analyze the prediction performance of

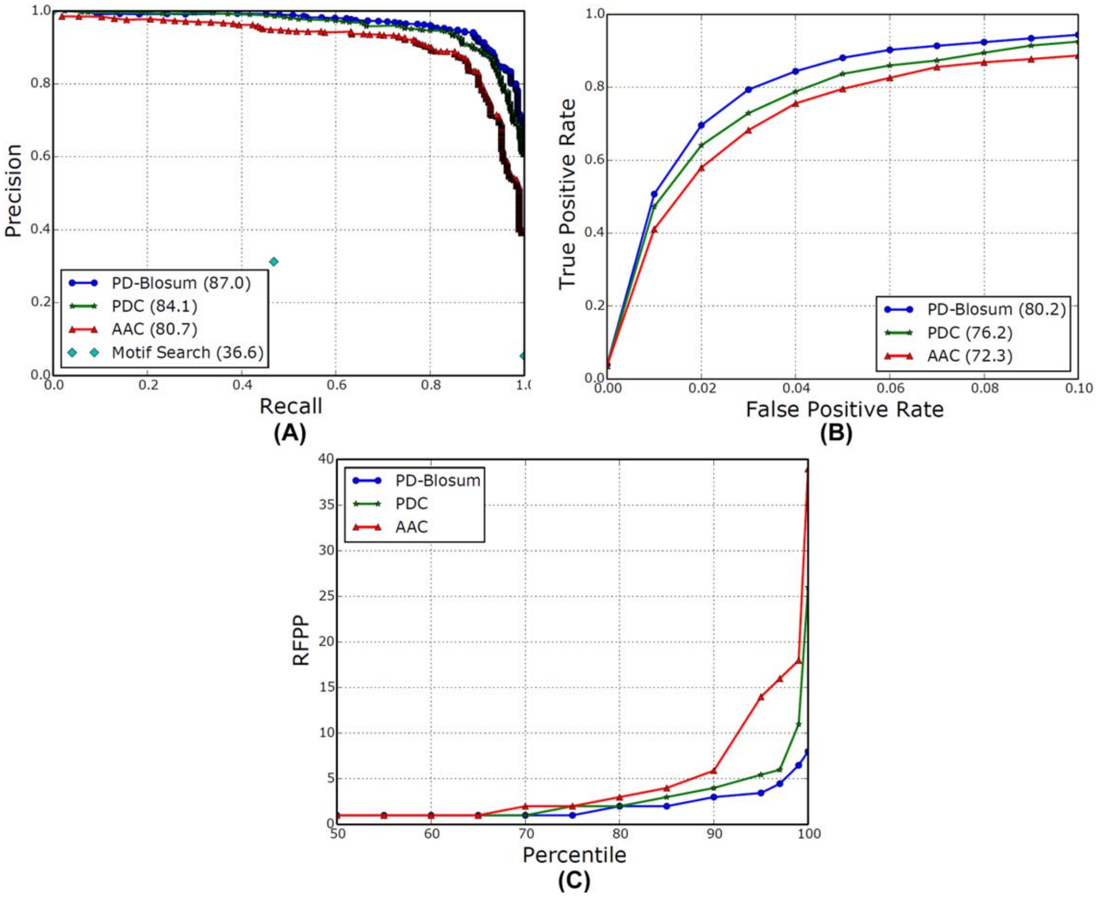


FIGURE 5 **(A)** Precision-recall (PR) curves for binding site prediction across all models; **(B)** ROC$_{0.1}$ curves for binding site prediction across all models. **(C)** Plot of the Rank of First Positive Prediction (RFPP) for binding site prediction across different models. The mean area under the curve across different folds is shown in parenthesis [Color figure can be viewed at wileyonlinelibrary.com]

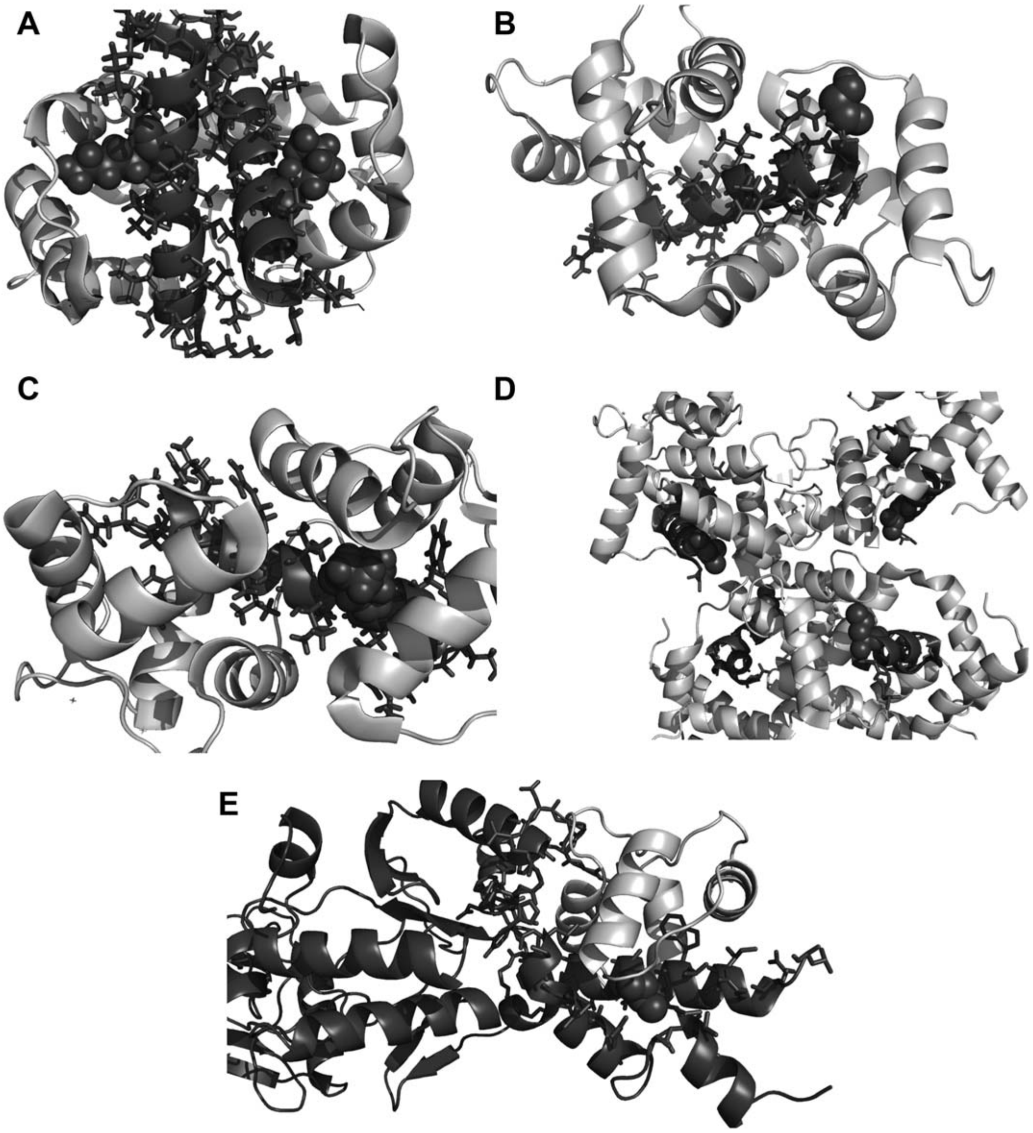


FIGURE 6 3D Structures of CaM complexes used in the independent validation dataset along with predicted binding sites from CaMELS. Calmodulin (CaM) is shown in light grey while CaM binding protein is in dark grey color. The predicted central residue of the binding site from the whole sequence of the binding protein is shown as a sphere. Residues of the CaM binding protein within 5Å of CaM are shown in stick form. **(A)** CaM bound to the C-terminal Domain of Petunia Glutamate Decarboxylase (PDB ID: 1NWD); **(B)** CaM complexed with a fragment of the olfactory CNG channel (PDB ID: 1SY9); **(C)** CaM complexed with Olfactory Cyclic Nucleotide-Gated Ion Channel (PDB ID: 2M0K); **(D)** CaM and Chloride anion exchanger (PDB ID: 5DOW) **(E)** CaM bound to the catalytic domain of adenylyl cyclase (PDB ID: 1YRT)

these methods in detail, we have used different performance metrics such as area under the precision-recall (AUC-ROC) and receiver operating characteristics curves (AUC-ROC) in addition to False and True Hit rates (FHR and THR). As shown in Figure 5A and Table 4, CaMELS gives an AUC-PR of 87.0% with AUC-ROC of 99.1%. These results show a significant improvement in comparison to previous state of the art CaM binding site prediction method MI-1 with an AUC-ROC of 96.9%. A notable increase is also seen in $AUC_{0.1}$ from 59.0% to 80.2% (Table 4; Figure 5B). This increase in $AUC_{0.1}$ shows that CaMELS is more accurate in predicting true binding sites with a very low false positive error rate. CaMELS gives a true hit-rate (THR) of 77% which indicates that in 77% of the proteins in our validation set, the top scoring window identified by CaMELS is the true binding site. The False Hit Rate (FHR) of 1.0% shows that, on average, only one out of every 100 non-binding site windows in a protein is expected to score higher than the true binding site in that protein. These results clearly show the effectiveness of CaMELS in designing wet-lab experiments for CaM binding site identification.

We have also analyzed the performance of CaMELS using the distribution of the rank of first positive prediction which quantifies how often the top predictions from CaMELS can be expected to be correct. These results are shown in Figure 5C for different features. We can see that CaMELS, using the position dependent BLOSUM features, gives the maximum RFPP of 8. This implies that for all the proteins in our dataset, the true binding site of a CaM binding protein always lies within top 8 CaMELS predictions. Thus, it can be expected that a

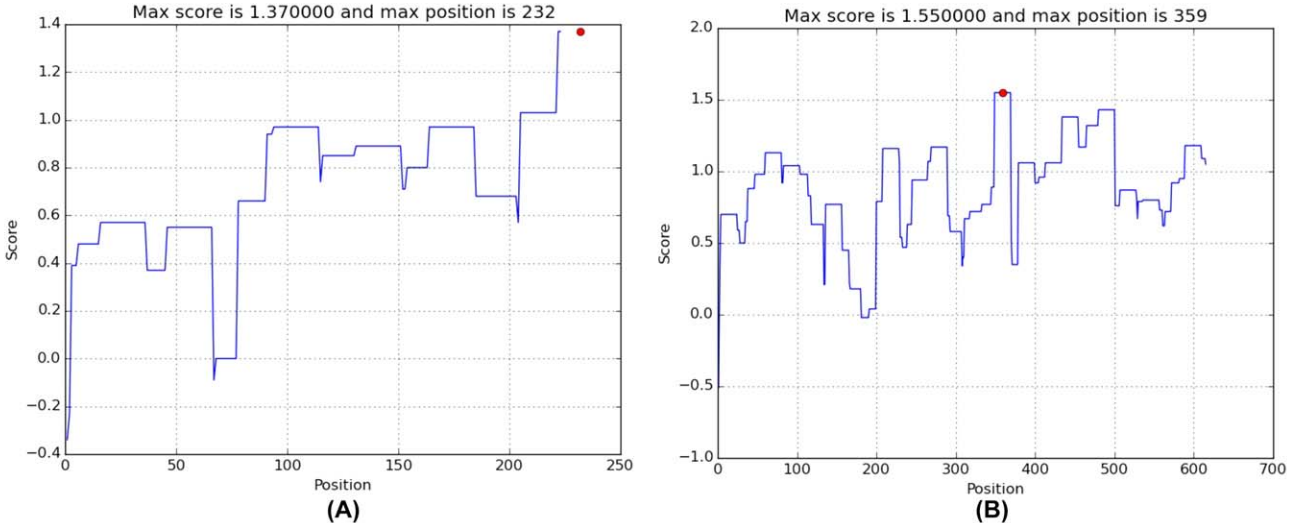


FIGURE 7 Window level scores of proteins used in mutagenic studies generated through CaMELS webserver for binding site prediction. Location of the highest scoring window (predicted binding site) is represented as a red dot. **(A)** LCa; **(B)** SGS3 of *Nicotiana Benthamiana* [Color figure can be viewed at wileyonlinelibrary.com]

biologist using CaMELS to design mutagenic experiments for identifying CaM binding sites of a protein, will have to run no >8 wet-lab experiments to locate the true binding site. Furthermore, for 90% of the proteins in our data set, the true binding site is always within the top 3 predictions by CaMELS. This is a major improvement over the previous state of the art methods.

3.2.2 | Comparison to motif-based CaM binding site search

A commonly used method of locating CaM binding sites in proteins is to search for known CaM-binding motifs in a protein. We have compared the performance of CaMELS to motif-based search for locating CaM binding sites in proteins. Motif-based search gives an AUC-PR of only 36.6% which is significantly lower in comparison to CaMELS (see Figure 5A). These results of motif-based analysis are in strong agreement with findings presented in the study by Mruk et al.[12]

3.2.3 | Testing on novel protein sequences

To further evaluate the generalization accuracy of CaMELS for CaM binding site prediction, we have used an independent dataset of 5 CaM binding proteins from the Protein Data Bank which have <40% sequence identity to proteins in the CaMELS training data set. Since the structures of these proteins in complex with CaM are known, it is easy to compare the binding site predicted by CaMELS with the true binding site. For all CaM binding proteins in this external validation dataset, a significant number of residues in the top scoring sequence window across all possible length-21 sequence windows in these proteins occur within 5 Å of CaM in the complex structure. Figure 6 shows the 3 D structures of these 5 CaM complexes along with predicted binding sites. CaM is shown in light gray color whereas the CaM binding protein is shown in a darker shade of gray. The residues predicted by CaMELS using the complete protein sequence are shown as spheres and the residues that occur within 5 Å of CaM are shown in stick representation. It is easy to see that, for all these proteins, CaMELS is able to identify the correct binding site even though these proteins are significantly different in sequence to the proteins used for training CaMELS.

We have tested CaMELS further using two additional proteins: LCa and SGS3 of *Nicotiana Benthamiana* whose binding sites were identified through mutagenic studies by Pley et al.[43] and Li et al.[42] respectively. It is important to note that these proteins also have very low sequence similarity with the proteins used in training the machine learning model in CaMELS. For the LCa protein, the top scoring windows from CaMELS occurs at the residues 224–243. This region of the protein was found to be involved in its interaction with CaM through deletion mutations by Pley et al. Similarly, in the case of NbSGS3 protein, the predicted binding site overlaps with the XS domain which has been shown to interact with CaM interacting by Li et al. These results clearly demonstrate the effectiveness of CaMELS in identifying CaM binding sites (see Figure 7).

3.2.4 | Contribution of individual amino acids and identification of motifs

It is interesting to notice in Table 4 that the position dependent composition features have higher accuracy than position independent amino acid composition. This shows that the binding of a protein to CaM is not a consequence of just the composition of amino acids in the protein and that the position of different amino acids in the binding window is also important for this purpose. One of the most useful features of CaMELS is that its machine learning model can be used to provide insight into the nature of the interaction of proteins with CaM. Specifically, the weight vector of the machine learning model can be used to assess the contribution of individual features for CaM binding site prediction. Figure 8 shows the weight vectors obtained from training CaMELS with amino acid composition

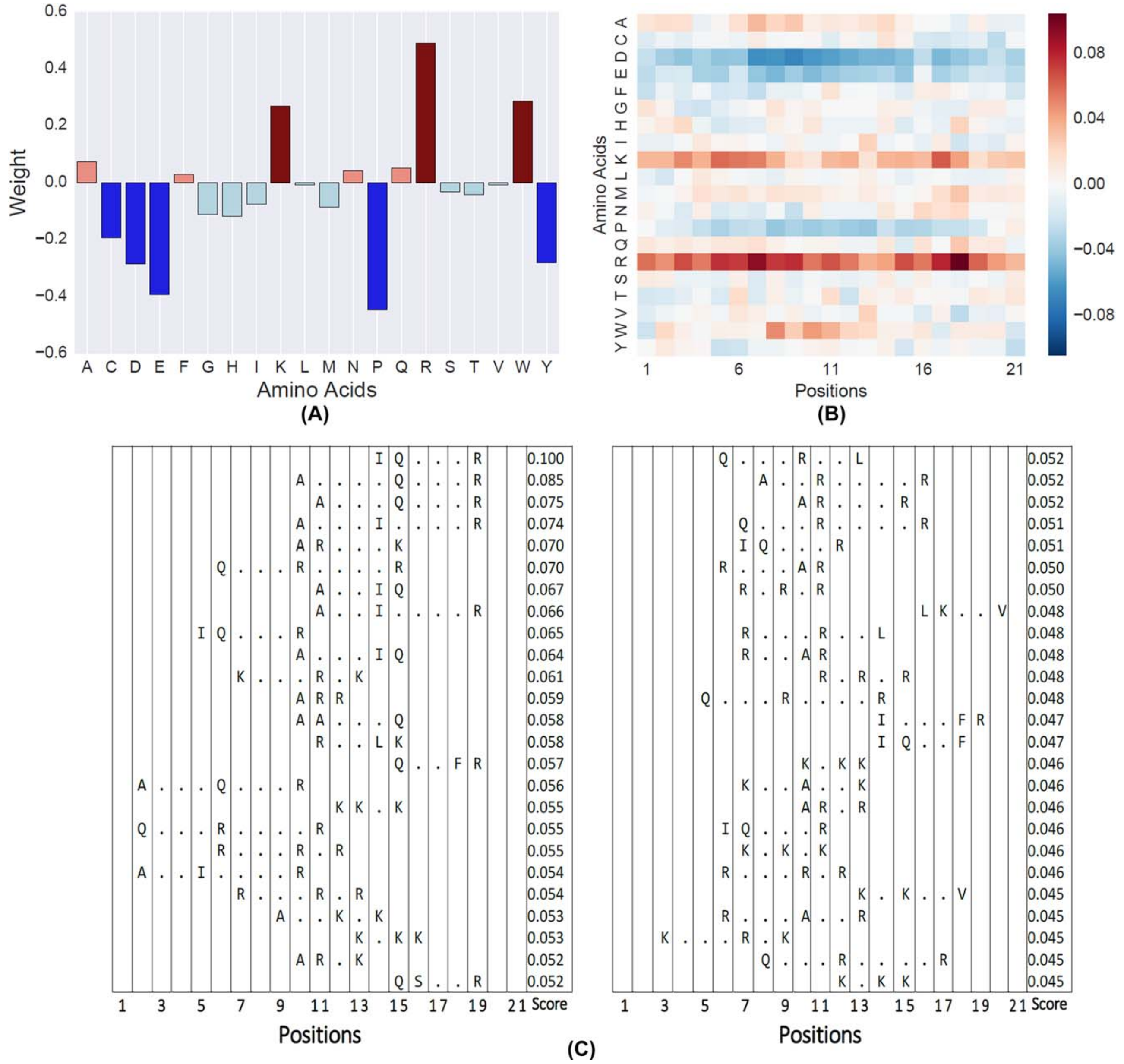


FIGURE 8 Weight vectors for position-independent (AAC) and position-dependent (PDC) amino acids composition along with learned motifs. **(A)** Weights of different amino acids in the position-independent AAC feature representation; **(B)** Heat map of the weights of different amino acids against their position from position-dependent composition (PDC) feature representation; **(C)** Top 50 motifs from the position-dependent gappy triplet feature representation. The numeric column shows actual weight values for different gappy triplet subsequences

(AAC) and position-dependent composition (PDC) features. These weight vectors show the importance of individual amino acids in determining CaM binding sites within a protein sequence. The weight vector of AAC feature representation shown in Figure 8A depicts large positive weights for positively charged amino acids Arginine (R), Lysine (K) and the hydrophobic amino acid Tryptophan (W). Amino acids such as Aspartic acid (D), Glutamic acid (E), Proline (P) and Tyrosine (Y) have large negative weights. These amino acid propensities in CaM binding sites are in close agreement with previous studies and also with known CaM-binding motifs.[9,10,15,53] The weight vector of position dependent feature representation, shown in Figure 8B illustrates the role of different amino acids in binding site prediction with respect to their positions in a given window. For example, Lysine (K) shows large positive weights at the end of the window but small in the middle; Arginine (R) shows large positive weights at positions 8 and 18; Tryptophan (W) has large positive weights at middle and negative weights at the corner of the window; Aspartic acid (D), Glutamic acid (E), Proline (P) show their negative role in CaM binding with large negative weights in the middle. This position dependent learning behavior of the classifier is in close agreement with known CaM-binding motifs and can be used to extract more biologically relevant motifs.[15]

We have also used position dependent gappy triplet (PDGT) feature representation to find motifs relevant to CaM binding. This feature representation can predict CaM binding sites very accurately (see Table 4). However, the primary purpose of this feature representation is to identify the subsequences or motifs which are responsible for prediction of binding sites. These motifs are ranked in terms of their weight values. Figure 8C shows the top 50 motifs and their position within a window of size 21. The top scoring motifs such as *IQ...R*, *A...Q...R*, *A...I....R* and *A.IQ* are components of IQ-subclass of motifs in the CaM Target Database.[15] It is also clear from Figure 8C that most of the top-ranking features correspond to a motif with 3 or 4 do not care positions. This is in agreement with the known fact that CaM binding involve alpha helices and this gap corresponds to the periodicity of the alpha helix.[10] These results indicate the ability of CaMELS to learn CaM binding motifs from protein sequences.

3.2.5 | Stochastic sub-gradient optimization method for MIL

Table 4 shows the accuracy of different machine learning models for predicting binding sites in CaM binding proteins. The difference in the prediction accuracy between a conventional SVM and multiple instance learning based methods (mi-SVM, MI-1 and CaMELS) shows the effectiveness of modeling CaM binding site prediction problem through multiple instance learning. Furthermore, the improvement in the accuracy of CaMELS with respect to other MIL based techniques (MI-1 and mi-SVM) is a consequence of solving the MIL optimization problem through the proposed stochastic sub-gradient optimization (SSGO).

3.2.6 | Analysis of features

CaMELS uses different window-level feature representations for CaM binding site perdition such as PD-Blosum, PDC, AAC, a combination of AC and PDC and propy. The PD-Blosum feature representation gives the best results in comparison to other features (Table 4, Figure 5). These features represent an amino acid in protein sequence window in terms of its corresponding column in the BLOSUM-62 substitution matrix and its relative position in the protein sequence. This clearly shows that it is important to model the position-specific nature of the binding site problem as well as the physiochemical and substitution frequency of different amino acids for accurate CaM binding site prediction.

4 | CONCLUSIONS

We have presented a set of models for CaM interaction and binding site prediction called CaMELS. CaMELS uses protein sequence information only and offers state of the art accuracy both for interaction and binding site prediction. For interaction prediction, CaMELS achieved significant improvement in performance using protein level features in comparison to earlier methods that used information derived only from the most likely CaM binding site in a protein. This shows that sequence information of the whole protein is predictive of its interaction with CaM. We have also presented a multiple instance learning model for solving the binding site prediction problem. Our results show near perfect classification accuracy for this problem with the use of a stochastic gradient solver. The proposed suite of algorithms is expected to be very helpful to biologists working on analyzing the functions and interaction behavior of CaM and its target proteins.

ACKNOWLEDGMENTS

The authors would like to thank of Dr. Asa Ben-Hur, Colorado State University, Fort Collins, USA and Dr. Imran Amin and Ms. Hira Kamal, National Institute of Biotechnology and Genetic Engineering, Pakistan for their support. We would also like to acknowledge the technical support by Mr. Naveed Akhtar, High-Performance Computing (HPC) lab, PIEAS, Pakistan. We are also thankful to the anonymous reviewer(s) and the journal editor(s) for their valuable comments. Wajid A. Abbasi is supported by a grant (PIN: 213-58990-2PS2-046) under indigenous 5000 Ph.D. fellowship scheme from the Higher Education Commission (HEC) of Pakistan.

CONFLICT OF INTEREST

None declared.

REFERENCES

[1] Bouché N, Yellin A, Snedden WA, Fromm H. Plant-specific calmodulin-binding proteins. *Annu Rev Plant Biol.* 2005;56:435–466.

[2] Chin D, Means AR. Calmodulin: a prototypical calcium sensor. *Trends Cell Biol.* 2000;10:322–328.

[3] Yamniuk AP, Vogel HJ. Calmodulin's flexibility allows for promiscuity in its interactions with target proteins and peptides. *Mol Biotechnol.* 2004;27:33–57.

[4] Reichow SL, Clemens DM, Freites JA, et al. Allosteric mechanism of water-channel gating by Ca2+-calmodulin. *Nat Struct Mol Biol.* 2013;20:1085–1092.

[5] Vogel HJ. Calmodulin: a versatile calcium mediator protein. *Biochem Cell Biol.* 1994;72:357–376.

[6] Möller W, Brown DM, Kreyling WG, Stone V. Ultrafine particles cause cytoskeletal dysfunctions in macrophages: role of intracellular calcium. *Part Fibre Toxicol.* 2005;2:7.

[7] Reddy ASN, Ben-Hur A, Day IS. Experimental and computational approaches for the study of calmodulin interactions. *Phytochemistry.* 2011;72:1007–1019.

[8] Radivojac P, Vucetic S, O'Connor TR, Uversky VN, Obradovic Z, Dunker AK. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins.* 2006;63:398–410.

[9] Hamilton M, Reddy ASN, Ben-Hur A. Kernel Methods for Calmodulin Binding and Binding Site Prediction. In: Proceedings of the 2Nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine; 2011; New York, NY, USA. ACM. p. 381–386. (Proceedings of the 2Nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine).

[10] Minhas FAA, Ben-Hur A. Multiple instance learning of Calmodulin binding sites. *Bioinformatics.* 2012;28:i416–i422.

[11] Wang L, Liu Z-P, Zhang X-S, Chen L. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Eng Des Sel.* 2012;25:119–126.

[12] Mruk K, Farley BM, Ritacco AW, Kobertz WR. Calmodulation meta-analysis: Predicting calmodulin binding via canonical motif clustering. *J Gen Physiol.* 2014;144:105–114.

[13] Degrado WF, Erickson-Viitanen S, Wolfe HR, O'Neil KT. Predicted calmodulin-binding sequence in the γ subunit of phosphorylase b kinase. *Proteins.* 1987;2:20–33.

[14] Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet.* 2000;25:25–29.

[15] Yap KL, Kim J, Truong K, Sherman M, Yuan T, Ikura M. Calmodulin Target Database. *J Struct Func Genom.* 2000;1:8–14.

[16] Popescu SC, Popescu GV, Bachan S, et al. Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays. *Proc Natl Acad Sci U S A.* 2007;104:4730–4735.

[17] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics.* 2010; 26:680–682.

[18] Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20: 273–297.

[19] Dietterich T, Lathrop RH. Solving the multiple-instance problem with axis-parallel rectangles. *Artif Int.* 1997;89:31–71.

[20] Andrews S, Tsochantaridis I, Hofmann T. *Support vector machines for multiple-instance learning*. MIT Press; 2003: 561–568.

[21] Leistner C, Saffari A, Bischof H. MIForests: Multiple-instance learning with randomized trees. In: Daniilidis K, Maragos P, Paragios N, eds. *Computer Vision – ECCV 2010., Lecture Notes in Computer Science*. Heidelberg: Springer Berlin; 2010:2942.

[22] Shalev-Shwartz S, Singer Y, Srebro N, Cotter A. Pegasos: primal estimated sub-gradient solver for SVM. *Math Program.* 2011;127:3–30.

[23] London N, Movshovitz-Attias D, Schueler-Furman O. The structural basis of peptide-protein binding strategies. *Structure.* 2010;18:188–199.

[24] Eddy SR. Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechol.* 2004;22:1035–1036.

[25] Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucl Acids Res.* 2006;34:W32–W37.

[26] Limongelli I, Marini S, Bellazzi R. PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC Bioinform.* 2015; 16:123.

[27] Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics.* 2013;29:960–962.

[28] Craig N, Green R, Greider C, Cohen-Fix O, Storz G, Wolberger C. *Molecular Biology: Principles of Genome Function*. Oxford: OUP; 2014.

[29] Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol.* 1981;147:195–197.

[30] Rognes T. Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinform.* 2011;12:221.

[31] Abbasi WA, Minhas FUAA. Issues in performance evaluation for host–pathogen protein interaction prediction. *J Bioinform Comput Biol.* 2016;14:1650011.

[32] Kohavi R. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2; 1995; San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. pp. 1137–1143. (Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2).

[33] Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinform.* 2009;10:48.

[34] Levandowsky M, Winter D. Distance between Sets. *Nature.* 1971; 234:34–35.

[35] Guo Q, Shen Y, Lee Y-S, Gibbs CS, Mrksich M, Tang W-J. Structural basis for the interaction of Bordetella pertussis adenylyl cyclase toxin with calmodulin. *EMBO J.* 2005;24:3190–3201.

[36] Yap KL, Yuan T, Mal TK, Vogel HJ, Ikura M. Structural basis for simultaneous binding of two carboxy-terminal peptides of plant glutamate decarboxylase to calmodulin. *J Mol Biol.* 2003;328:193–204.

[37] Contessa GM, Orsale M, Melino S, et al. Structure of calmodulin complexed with an olfactory CNG channelfragment and role of the central linker: Residual dipolar couplingsto evaluate calmodulin binding modes outside the kinase family. *J Biomol NMR.* 2005;31:185–199.

[38] Keller JP. Solution of the structure of a calmodulin–peptide complex in a novel configuration from a variably twinned data set. *Acta Crystallogr Sect D Struct Biol.* 2017;73:22–31.

[39] Irene D, Huang J-W, Chung T-Y, et al. Binding orientation and specificity of calmodulin to rat olfactory cyclic nucleotide-gated ion channel. *J Biomol Struct Dyn.* 2013;31:414–425.

[40] Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol.* 2003;10:980.

[41] UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* 2015;43(Database issue):D204–D212.

[42] Li F, Zhao N, Li Z, et al. A calmodulin-like protein suppresses RNA silencing and promotes geminivirus infection by degrading SGS3 via the autophagy pathway in Nicotiana benthamiana. *PLoS Pathog.* 2017;13:e1006213.

[43] Pley UM, Hill BL, Alibert C, Brodsky FM, Parham P. The interaction of calmodulin with clathrin-coated vesicles, triskelions, and light chains localization of a binding site. *J Biol Chem.* 1995;270:2395–2402.

[44] Alpaydin E. Design and analysis of machine learning experiments, in *Introduction to Machine Learning*, 2nd ed., Cambridge, MA: MIT Press; 2010, pp. 486–488.

[45] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning; 2006; New York, NY, USA. ACM. p 233–240. (Proceedings of the 23rd International Conference on Machine Learning).

[46] Minhas F, Geiss BJ, Ben-Hur A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. *Protein Struct Funct Bioinform.* 2014;82:1142–1155.

[47] Planas-Iglesias J, Marin-Lopez MA, Bonet J, Garcia-Garcia J, Oliva B. iLoops: a protein–protein interaction prediction server based on structural features. *Bioinformatics.* 2013;29:2360–2362.

[48] Schölkopf B, Tsuda K, Vert J-P. *Kernel Methods in Computational Biology*. Cambridge, Massachusetts, USA: MIT Press; 2004.

[49] Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. *Bioinformatics.* 2004; 20:467–476.

[50] Kessel A, Ben-Tal N. *Introduction to Proteins: Structure, Function, and Motion*. Boca Raton, FL, USA: CRC Press; 2010.

[51] Petsko GA, Ringe D. *Protein Structure and Function*. USA: Sinauer Associates/New Science Press; 2004.

[52] Swulius MT, Waxham MN. Ca2+/Calmodulin-dependent protein kinases. *Cell Mol Life Sci.* 2008;65:2637–2657.

[53] Rhoads AR, Friedberg F. Sequence motifs for calmodulin recognition. *FASEB J.* 1997;11:331–340.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Abbasi WA, Asif A, Andleeb S, Minhas FAA. CaMELS: *In silico* prediction of calmodulin binding proteins and their binding sites. *Proteins*. 2017;00:1–17. https://doi.org/10.1002/prot.25330