Received: 15 October 2020 | Accepted: 31 January 2021

DOI: 10.1002/prot.26056

RESEARCH ARTICLE

PROTEINS WILEY

DeepCys: Structure-based multiple cysteine function prediction method trained on deep neural network: Case study on domains of unknown functions belonging to COX2 domains

Vamsi Nallapareddy | Shubham Bogam | Himaja Devarakonda | Shubham Paliwal | Debashree Bandyopadhyay

Department of Biological Sciences, Birla Institute of Technology and Science, Hyderabad, Telangana, India

Correspondence
Debashree Bandyopadhyay, Department of Biological Sciences, Birla Institute of Technology and Science, Pilani Hyderabad Campus, Hyderabad, Telangana 500078, India.
Email: banerjee.debi@hyderabad.bits-pilani.ac.in

Funding information
Department of Science and Technology (DST-SERB), Government of India, Grant/Award Number: EMR/2017/002953

Abstract

Cysteine (Cys) is the most reactive amino acid participating in a wide range of biological functions. In-silico predictions complement the experiments to meet the need of functional characterization. Multiple Cys function prediction algorithm is scarce, in contrast to specific function prediction algorithms. Here we present a deep neural network-based multiple Cys function prediction, available on web-server (DeepCys) (https://deepcys.herokuapp.com/). DeepCys model was trained and tested on two independent datasets curated from protein crystal structures. This prediction method requires three inputs, namely, PDB identifier (ID), chain ID and residue ID for a given Cys and outputs the probabilities of four cysteine functions, namely, disulphide, metal-binding, thioether and sulphenylation and predicts the most probable Cys function. The algorithm exploits the local and global protein properties, like, sequence and secondary structure motifs, buried fractions, microenvironments and protein/enzyme class. DeepCys outperformed most of the multiple and specific Cys function algorithms. This method can predict maximum number of cysteine functions. Moreover, for the first time, explicitly predicts thioether function. This tool was used to elucidate the cysteine functions on domains of unknown functions belonging to cytochrome C oxidase subunit-II like transmembrane domains. Apart from the web-server, a standalone program is also available on GitHub (https://github.com/vamsin/deepcys).

KEYWORDS
deep neural network, multiple cysteine function prediction, protein structure and sequence feature

1 | INTRODUCTION

Cysteine is a key amino acid at the catalytic site of many enzymes.[1] Unique chemical property of cysteine lies in its reactive thiol group, that can act as a nucleophile and may contribute toward various biological functions. Cysteine functions are broadly categorized into four groups observed in large number of biochemical reactions, (a) *Structural cysteines*, disulphide formation, binding to co-factors, that is, thioether formation, (b) *metal-binding cysteines*, present at enzyme active sites and involved in heavy metal scavenging (c) *catalytic cysteines* and (d) *regulatory cysteines*, involved in redox mediated various post-translational modifications.[2] Disulphide is the most common post-translational modification that facilitates the correct folding in protein structure as mentioned for the first time by Anfinsen.[3] Disulphide bond is formed between two sulfur atoms (each mentioned as half-cystine) coming from the same chain of a protein

(intra-disulphide) or from different chains of a protein (inter-disulphide), leading to native protein structure. Many metalloproteins (enzymes) involve cysteine as one of the metal-binding ligands, apart from histidine.[4,5] Thioester modifications, namely, acylation,[6,7] palmitoylation,[8-10] alkylation, and so forth, are commonly observed in fatty acid synthesis and degradation pathways. Thioether linkages[11] are often observed with ligands, especially heme (prosthetic) groups.[12] Apart from the naturally occurring post-translational modifications of cysteine many more modifications occur via reactive oxygen species[13] induced oxidative stress,[14] or reactive nitrogen species.[15] Glutathione is a cysteine-containing small molecule that can form disulphide bonds with a cysteine residue from protein. Levels of glutathionylation are often modulated by oxidative stress. However, glutathionylation may happen under normal conditions facilitating redox signaling and various other cellular activities. Persulphenylation is a modification mainly observed in plants under stress conditions.[16] Selenylation is mostly used to derivatize and protect the cysteine thiol group from oxidation, in vitro.[17] This variety of cysteine functions and their possible implication on a wide range of biological functions, make the cysteine residue an important target for function prediction in a given protein (Figure 1). Amino acid function prediction became increasingly important with the advent of the structure genomics consortium,[18] where a large number of protein crystal structures were solved with unknown functions; 3970 such structures were reported on 25 September 2020 in PDB database.[19] Prediction of functions in unknown proteins or in hypothetical proteins were attempted earlier in different species.[20,21] However, experimental determination of amino acid function is laborious, time-consuming, and expensive, hence, in-silico prediction can complement the experiments.

Most of the existing cysteine prediction methods can predict one particular type of function, termed, here as "specific cysteine function prediction," such as, disulphide prediction,[22-29] metal-binding prediction,[30-39] and sulphenylation prediction.[40-49] Besides the specific cysteine function prediction methods, four multiple cysteine function prediction methods were known, namely, diamino acid neural network application (DiANNA),[50] COPA,[51] ASP-C,[52] and Cy-preds.[53] DiANNA employed a support vector machine to predict the class of the cysteine residue in three categories, a free cysteine, a half-cysteine, or a ligand-bound cysteine. COPA was based on Cys proximity, average low pKa value and exposure of the sulfur atom; the method was capable of predicting reactive cysteines, namely, disulphide and metal binding. ASP-C was capable of predicting reactive cysteines based on active site profiling. Cy-preds was capable to predict three different types of cysteine modifications, namely, disulphide, metal-binding and post-translational modifications, based on energy components and different profiling approaches.

Earlier we have annotated four cysteine functions, namely, disulphide, thioether, metal-binding, and sulphenylation, based on protein structural properties, like, buried fraction, quantitative microenvironment descriptor (rHpy), secondary structure, and pKa values.[54-56] Only these four modifications were chosen because of their abundance in PDB crystal structures. Inspired by these functional annotations of cysteine, here we propose a deep neural network-based model, DeepCys, that exploits six different protein features, and predict any one of these four different cysteine modifications. The model was trained on high resolution protein crystal structures containing total of 108 334 cysteine residues. Along with the original training dataset, that is, without any sequence filter, two

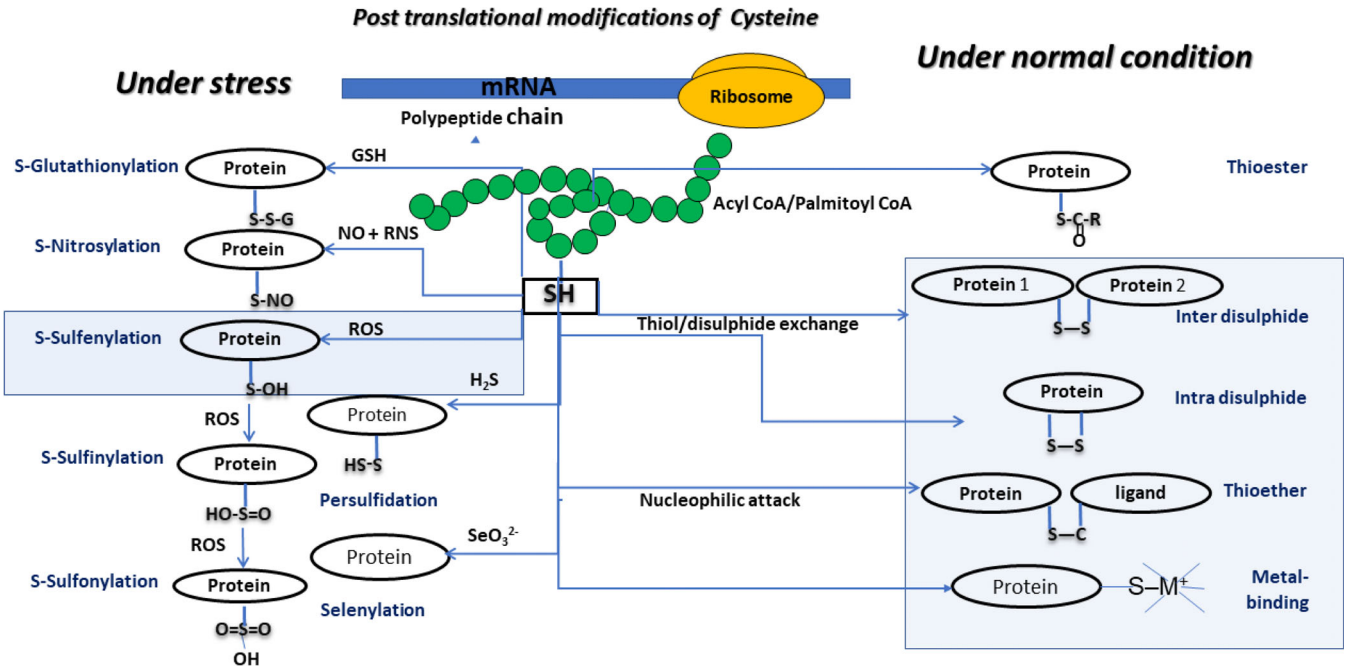

FIGURE 1 Different post-translational modifications of cysteine residues. Modifications, generally, occurring under stress are shown in the left panel and those often occur under normal conditions are shown in the right panel. The modifications studied in this work are shown within boxes with a light blue background [Color figure can be viewed at wileyonlinelibrary.com]

more nonredundant datasets, with 100% and 30% sequence identity filters, were used in order to check for overfitting of the data. Three different models, namely, DeepCys original, DeepCys 100% and DeepCys 30%, were developed based on three different training datasets and the results showed that performance of DeepCys original was the best among all. Performance of the models was evaluated based on three metrics, accuracy, specificity and sensitivity. The last parameter, sensitivity represented by number of true positives divided by total number of true positives and false negatives, was the most effective for evaluating multifunctional dataset. The DeepCys original model was tested upon an independent test data set, consisting of 126 652 number of cysteine residues from medium resolution crystal structures. A case study was performed on total 66 cysteine datapoints from domains of unknown functions (DUF) proteins belonging to cytochrome C oxidase subunit-II like trans-membrane domains. The average sensitivity values for the test and the DUF datasets were 79.25% and 87.22%, respectively. Performance of DeepCys model was compared with the existing prediction methods; DeepCys results were either comparable or better than these methods. Novelty of the current method is, for the first time a method can predict one out of four cysteine functions. Moreover, explicit prediction of thioether is also done for the first time using DeepCys.

2 | MATERIALS AND METHODS

2.1 | Extraction of experimental information for cysteine functions (modifications)

Four cysteine functions, namely, disulphide, metal-binding, thioether and sulphenylation, were either extracted from PDB entries, using distance criteria, or the information was directly extracted from the PDB header file. The information obtained from PDB file served as the experimental evidence for each modification that was compared with the results obtained from the predictive models.

2.2 | Training dataset generation

A training dataset was curated from high-resolution protein crystal structures (resolution ranges from 1.5 to 2.0 Å), deposited to PDB database,[19] dated 13 July 2020. Total 13 142 PDB entries were present in the dataset (Table S1). This dataset was termed as training dataset original.

2.3 | Nonredundant training dataset generation

The training dataset original contained all possible protein structures without any restriction on sequence identity. In addition to the training original dataset, two nonredundant datasets, training 100% and training 30%, were generated. After removal of redundancy based on sequence identity using CD-HIT.[57] This dataset contained 7188 unique PDB files and total 60 337 cysteine residues (Table S2).

Similar to training 100% dataset, training 30% dataset with sequence identity of 30% was generated to ensure no structural bias. The data size was further reduced (Table 1). A total of 3121 unique proteins and 25 435 cysteine residues were present in this dataset (Table S3).

The reason to choose three different datasets was to identify if the bias present in the training original dataset affected the overall performance. To address this question three DeepCys models were developed based on the three training datasets, namely, DeepCys original, DeepCys 100% and DeepCys 30%.

The number of PDB files and corresponding cysteines undergoing different modifications were shown for three different training datasets (Table 1). To note, that there are certain PDB files containing multiple cysteines with different modifications. Therefore, the total number of PDB files reported in Table 1 was higher than the actual number of PDB files in the dataset (Table S1).

2.4 | Identification of different cysteine modifications

2.4.1 | Disulphide

Disulphide modifications were identified in each PDB entry based on the distance criteria. The structural disulphide bond length was reported as 2.05 Å and that of reversible disuphide was 2.18 Å. Hence, 2.3 Å distance was chosen to define any disulphide bond connecting two sulfur atoms from two cysteine residues.[58] If both the cysteines belong to the same chain, the modification was considered as intrachain disulphide, in contrast to interchain disulphide where two cysteines belong to two different protein chains. The calculation was implemented by in-house python script exploiting the Biopython libraries, namely, PDB and NeighbourSearch. Each sulfur atom in disulphide modification was considered as half-cystine.

2.4.2 | Metal-binding

The metal ions identified in the training dataset were the following, Zn^{2+}, Cu^{2+}, Cd^{2+}, Fe^{2+}/Fe^{3+} and Hg^{2+}. The zinc ion was observed maximum number of times in the dataset (Table 2). It was noted earlier that the thiolate group of cysteine formed coordinate bonds with a wide range of border line to soft cations, such as Zn^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, Cd^{2+}, with maximum propensity towards zinc ion.[59] Metal populations in three different datasets were described in terms of percentage of metal ions present in the dataset. The percentage of metal ion was described by the number of specific metal ion divided by the total number of metal-binding cysteines. The distance between sulfur atom of a cysteine and the metal ion varied according to the type and the oxidation number of the metal ions. The same metal ion could maintain different distances with a cysteine sulfur atom, depending upon the function of the metalloprotein.[60] The maximum metal ion - S (Cys) distance of 2.6 Å (that was, $Cd2^{+}$ – S distance) was used, here, as search criteria. In-house python script was used to implement the calculation exploiting the Biopython libraries, namely, PDB and Neighboursearch.

TABLE 1 Four different modifications present in A, training dataset original B, training 100% dataset and C, training 30% dataset

(A) Training dataset original		
Modification	**Total number of PDB structures analyzed**	**Total number of cysteines analyzed**
Disulphide	9179	85 452[a]
Thioether	979	3244
Metal-binding	3061	18 959
Sulphenylation	373	679
(B) Training 100% dataset		
Modification	**No. of PDB files**	**No. of cysteines**
Disulphide	5015	48 138[a]
Thioether	513	1926
Metal-binding	1520	9808
Sulphenylation	218	465
(C) Training 30% dataset		
Modification	**No. of PDB files**	**No. of cysteines**
Disulphide	2017	18 292[a]
Thioether	208	926
Metal-binding	786	5910
Sulphenylation	146	307

[a]Number of half-cystines, two half-cystines constitute one cystine (containing disulphide bond).

TABLE 2 Variation of metal ion populations (described in terms of percentage, with respect to the total number of metal-binding cysteines) in different training datasets

Name of the metal ion	Training original	Training 100%	Training 30%
Zn^{2+}	74.8	72.2	76.9
Hg^{2+}	10.3	13.1	12.3
Cu^{+}/Cu^{2+}	4.7	5.4	5.1
Fe^{2+}/Fe^{3+}	8.4	7.9	5.6
Cd^{2+}	1.7	1.7	1.6

2.4.3 | Sulphenylation

Sulphenylation modification was directly extracted from the PDB header files where a cysteine residue has S-hydroxy modification, mentioned as modified S-hydroxycysteine (CSO, as per PDB nomenclature). The CSO residue name was reported as a hetero atom in PDB file. However, the current DeepCys model only considered coordinates of ATOMs and not of HETATMs. Hence, each CSO was edited to CYS and hydroxy part of CSO were removed, without changing the cysteine local microenvironment.

2.4.4 | Thioether

Thioether modification was directly extracted from the PDB header files, using following column matching criteria - the first column has "LINK," the second column has "SG," the third column has "CYS," the sixth column has "C" and the seventh column not having "CU."

2.4.5 | Test dataset generation

The test dataset was curated from medium-resolution protein crystal structures, resolutions ranging from 2.0 to 2.5 Å, reported in PDB database.[19] Total 10 864 PDB files (Table S4) with 125 652 cysteine residues were retrieved (Table 3A). The selection criteria for different modifications were the same as that of the training dataset. However, the metal ion populations varied in test dataset, compared to those in the three training datasets (Table 4). Test dataset was more populated with three heavy metal ions, namely, Hg^{2+}, Cd^{2+} and Pb^{2+}, in comparison to the training datasets. Pb^{2+} was completely absent in the training datasets and Cd^{2+} has very low population in the training datasets.

2.4.6 | DUF dataset generated for cytochrome C oxidase subunit II like trans-membrane (COX2) domains

In the current study, we focused on Cytochrome C oxidase (also known as complex IV) subunit II like transmembrane domains involved

TABLE 3 Cysteine modifications present in A) test-1 dataset and B) DUF dataset

(A) Test-1 dataset		
Modification	**Total number of PDB structures analyzed**	**Total number of cysteines analyzed**
Disulphide	8865	111 584
Thioether	421	1721
Metal-binding	1617	11 729
Sulphenylation	239	618
(B) DUF dataset		
Modification	**No. of PDB files**	**No. of cysteines**
Disulphide	5	30
Metal-binding	6	36

Abbreviation: DUF, domains of unknown function.

TABLE 4 Variation of metal ion populations (described in terms of percentage, with respect to the total number of metal-binding cysteines) in test-1 and DUF dataset

Name of the metal ion	Test 1 dataset	DUF dataset
Zn^{2+}	69.0	44.0
Hg^{2+}	13.5	—
Cu^{+}/Cu^{2+}	4.7	56.0
Fe^{2+}/Fe^{3+}	5.2	—
Cd^{2+}	7.7	—
Pb^{2+}	0.01	—

Abbreviation: DUF, domains of unknown function.

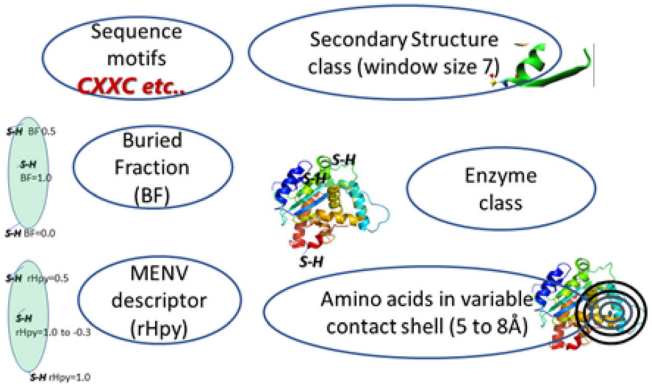

FIGURE 2 Features used for training the model. Cartoons representation of a protein with multiple cysteine thiol (—SH) groups at the center, and the features on the periphery. Feature names enclosed within blue ovals [Color figure can be viewed at wileyonlinelibrary.com]

in electron transport chain. Cytochrome c oxidase is a large integral membrane protein containing multiple chains and several metal prosthetic sites.[61] This enzyme complex accepts four electrons from four cytochrome C molecules and transfers those to two oxygen molecules. Subunit II is one of the three subunits involved in substrate binding and formation of the functional core of complex IV.[62] This subunit transfers electrons, using its binuclear copper center, from cytochrome c to the bimetallic center of subunit I. The binuclear copper center was considered as the primary acceptor of electrons in cytochrome c oxidase. However, many of the cytochrome C oxidase crystal structures contain other subunits, apart from subunit II. Those subunits contain other metal ions such as zinc and also several disulphide modifications. The cytochrome C oxidase proteins were searched in DUFs reported in SUPFAM database.[63] The keyword search resulted into one DUF ID: DUF3098, only. The corresponding SCOP family name to this DUF ID was cytochrome c oxidase subunit II-like, transmembrane region. The PDB files reported in the SCOP database were extracted (Table S5). Total number of cysteine modification and the PDB IDs were reported (Table 3B). DUF dataset comprised of Zn^{2+} and Cu^{2+} ion only (Table 4). The population of Zn^{2+} ion in DUF dataset, was significant, although it was low compared to that of the training datasets. However, the $Cu2^{+}$ ion population was very high compared to the training datasets.

2.5 | Feature generation

To determine a particular modification of a cysteine, deep learning approach was applied on a training dataset constructed from PDB database. Each cysteine was uniquely identified by the PDB identifier (ID), the chain ID, and the cysteine residue number. Six structural and functional attributes were either computed or extracted from PDB header file (Figure 2). The features computed were, (a) buried fraction, (b) rHpy, and (c) the secondary structure motif around the cysteine residue. The features extracted were, (a) amino acids present around cysteine within a variable contact shell, (b) enzyme class of the protein to which the cysteine belongs to and (c) the cysteine sequence motif.

The feature, pKa, has been identified as one of the important parameters for cysteine function predictions.[51,54] However, pKa computed using PROPKA has a predefined value of 99.99 for disulphide connectivity. This fixed pKa value for disuphide from PROPKA makes the deep learning model circular in nature. As other pKa computations were not automated like PROPKA, such as, constant pH MD simulations,[64] those cannot be used for automated feature generation.

2.6 | 1 and 2. Protein microenvironment (buried fraction and rHpy) calculation

Protein microenvironments, quantified in terms of buried fraction and rHpy, around all the 108 334 cysteines were computed using a FORTRAN program developed earlier.[56] This calculation required the following inputs: (a) three-dimensional structure of the protein, (b) CHARMM topology and parameter files,[65] and (c) Rekker's fragmental constants of individual atom types.[66] Microenvironment calculations report two outputs, (a) buried fraction and (b) rHpy. The buried fraction is defined as the normalized surface area of the cysteine thiol group buried inside the protein. The values of this parameter range from 0.0 to 1.0. Zero buried fraction indicates that the thiol group is completely exposed to the solvent and vice versa (Figure 2). The buried fraction of an amino acid (or its side chain) was computed in this FORTRAN program by calling another FORTRAN program GEPOL93.[67]

The second parameter, rHpy, termed as microenvironment property descriptor, describes the relative hydrophilic contribution of protein and the solvent toward the cysteine thiol group within its first contact shell. According to the mathematical formulation, rHpy value adopts an upper limit of one, when embedded in a pure aqueous solvent. There is no lower limit for rHpy value, which depends on the hydrophobicity of the protein interior. In our current dataset, the lower limit of rHpy for cysteine thiol group was −0.311. The buried fraction and rHpy together constituted protein microenvironment space around the cysteine thiol group.

3 | SECONDARY STRUCTURE (SS) MOTIFS

The secondary structures for all the 13 142 proteins, in the training original dataset, were calculated using the DSSP software[68] based on Kabsch and Sander algorithm.[69] The DSSP algorithm calculated the secondary structure based on the three-dimensional structure. To understand the effect of adjacent secondary structures around a cysteine, secondary structure motifs were searched with variable lengths. Variation in the length was introduced by a parameter, window size, that described the number of amino acids on either side of the central cysteine. Window size varied from 3 to 13. Multiple window sizes were tested and the window size of seven produced the highest performance on test dataset (Figure S1). Therefore, SS motif feature was generated with a window size of seven depicting the secondary structures of 15 amino acids (7*2 + 1).

4 | PROTEIN/ENZYME CLASS

The enzyme or the protein class to which the protein belongs to, was extracted from the PDB HEADER file.

5 | AMINO ACIDS PRESENT AROUND CYSTEINE WITHIN A VARIABLE CONTACT SHELL

Primary sequence of a protein folds into three-dimensional structure, assembling far off amino acid sequences within a given protein scaffold. A protein scaffold may contain more than one hydration layer, where protein atoms may interact with different layers of water molecules, within radii of 4.5 and 8 Å.[70] Analogous to hydration layers, the catalytic sites also include primary and secondary interaction shells where specific interactions were observed.[71,72] The amino acid signatures within first and second interaction shells were also considered as one of the important features by other cysteine prediction methods.[44,47] The quantitative descriptor, rHpy, (feature 2) represents the numerical value obtained from the hydrophilic contributions of each amino acid within the first contact shell, roughly 4.5 Å radius.[56] In this current feature, we have identified the amino acids present within the variable contact shells around a cysteine residue. The notion was to identify the optimal amino acid signature around a cysteine residue. An array of 20 elements (each representing 20 naturally occurring amino acids) was constructed for each of the four radius values (5-8 Å, incremented by 1 Å). Each element depicted the frequency of individual amino acids within a certain radius of the cysteine thiol group. In total there were 80 values (an array of 20 elements for each of the four different radii) in this feature.

6 | CYSTEINE SEQUENCE MOTIFS

A set of possible cysteine sequence motifs, namely, (a) CC, (b) CXC, (c) CXXC, (d) CXXXC, (e) CXXXXC, (f) CXXCXXC, (g) CXXCXXCXXC, and (h) CXXCXXCXXXC were observed earlier, mainly as a part of metal-binding, thioether, and disulphide modifications.[54] The cysteine motifs were searched from the primary sequence of each protein, within a variable window size of 3 to 13, incremented by the step size of two. The window size indicated the number of amino acids present on either side of the central cysteine residue. In some of the PDB files, few amino acids were not reported in three-dimensional coordinate (structure) although reported in the primary sequence, mentioned in the PDB header file. In those cases, amino acid residue positions of the coordinates were followed instead of the amino acid sequence, to maintain the overall consistency in this study. For each window size, a binary array of eight values was constructed, representing each of the eight cysteine motifs. "1" marked the presence of a motif, whereas "0" marked the absence in that particular stretch of amino acids. This feature has 48 values (an array of eight values for the six different window sizes).

6.1 | Metrics used to evaluate the efficiency of DeepCys model

Three different metrics, namely, accuracy, sensitivity and specificity, were used to measure the performance of the deep learning model on various datasets. Three metrices were defined using four parameters of the confusion matrix, namely, true positive (TP), true negative (TN), false positive (FP), and false negative (FN) (Equations 1-3). A confusion matrix tabulates the actual values and the predicted values, allowing one to understand the performance of the model. TP

represents the number of correctly predicted data points as proper positive class. FP represents the number of falsely predicted data points as the positive class. Similarly, TN represents the number of correctly predicted datapoints as the proper negative class and FN represents the number of falsely predicted data points as the negative class.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{FP} + \text{FN}), \quad (1)$$

$$\text{Sensitivity} = (\text{TP})/(\text{TP} + \text{FN}), \quad (2)$$

$$\text{Specificity} = (\text{TN})/(\text{TN} + \text{FP}), \quad (3)$$

As the accuracy values were expected to be more skewed toward the majority class, in case of class imbalance (ie, the case in the present dataset), it is not recommended to use the accuracy metrics to assess the model.[73] Instead, sensitivity is a better measure when focused individually on each class.[74] To measure the overall performance of the DeepCys model on the entire dataset, a new metrics was deduced based on simple arithmetic mean of sensitivity values obtained from individual classes (all the four cysteine modifications), termed here as macro-average of sensitivity.

6.2 | Description of deep-learning model

A neural network model was built using Keras, a high-level neural network application programming interface running on top of a TensorFlow backend. The model was programmed in Python, version 3.8.2. The proposed neural network model has three convolution layers followed by eight dense layers and a singular dropout layer. Each convolutional layer contained filters. These filters were applied on the input data to obtain features. Repeated application of these filters on the input generated a feature map. A feature map incorporated the necessary information detected from the input and leads to the required output. A dense layer was a regular stack of nodes. Each of these nodes received input from the nodes of the previous layer. Each dense layer was associated with a weight matrix and a bias matrix. These matrix parameters were updated during the training process. The dense layers in the neural network model started with 512 nodes, covering all the powers of two (2^n, n = 9-2), till the final layer having four nodes that represented the four cysteine modifications. A singular dropout layer with a probability of .5 was placed before the final dense layer. The dropout layer was different from other layers as it did not contain any trainable parameters. The only parameter associated with a dropout layer was a probability that determined whether a node would be randomly dropped during the training process. Dropout layers helped in reducing overfitting. Along with this dropout, regularization was done to prevent overfitting.

There was a total of three skip connections in the architecture of the neural network model (Figure 3). The data as it passed through every layer have been outlined in the architecture. The input vector has 146 values derived from the six features; one value each for buried fraction, rHpy and enzyme class; 80 values for 20 naturally occurring amino acids within four interaction shells with different radii; 48 values from eight different cysteine sequence motifs in six variable window sizes and 15 values from secondary structure folds. The Leaky ReLU activation function was employed along with batch normalization after every convolution layer and dense layer. The final dense layer used a softmax activation function. The softmax activation function resulted in four different output values (four cysteine modifications) which added up to one. The weights for all the layers were initialized using the Glorot uniform function. The loss function employed was weighted categorical cross-entropy to tackle the major class imbalance in the training dataset. The weights corresponding to each of the classes were obtained by modifying the inverse value of their frequency in the dataset (Table S6). The optimizer used was Adam.[75] A grid search algorithm was employed to figure out the optimal parameters for training the neural network. There were three optimal hyperparameters according to the grid search algorithm, namely, batch size, epochs and learning rate, having values of 256, 50, and 1e-4, respectively. Each epoch defined the process of the deep learning model being trained on the entire dataset. In this process of an epoch, the dataset was split into parts and the model was consecutively trained on these parts, termed as batches. The number of data points present in each batch represented the batch size. The learning rate was another essential hyperparameter that defined how quickly the weights of the neural network vary after each iteration of the training process.

Two model checkpoints were employed. The first checkpoint saved the best performing model after every epoch of training. The second checkpoint reduced the learning rate of the model by a factor of 0.1, if the performance did not improve for five consecutive epochs. In absence of the second checkpoint, the model failed to converge due to a very high learning rate. The failure was due to the attempts of the optimizer making greater changes to the weights and overshooting the location of the optima. A reduced learning rate helped the model converge to the optima as it steps slowly toward the optima without overshooting.

The feature generation and extraction along with the model training were carried out on a laptop, that is, MSI GF63 Thin Core i5 ninth Gen - (8 GB/512 GB SSD/Windows 10 Home/4 GB Graphics) which was equipped with an NVIDIA GeForce GTX 1650 Max Q GPU.

7 | RESULTS AND DISCUSSION

7.1 | Selection of all-feature criteria

As described above, there were total 146 values deduced from the six features. Here we attempted to identify the best set of values (and eliminate redundant values, if any) to increase the accuracy of the model using two different algorithms, namely, recursive feature elimination[76] and the genetic algorithm.[77] The set of values suggested by these two algorithms were used to train the DeepCys model on the training original dataset. These models were tested on test 1 dataset

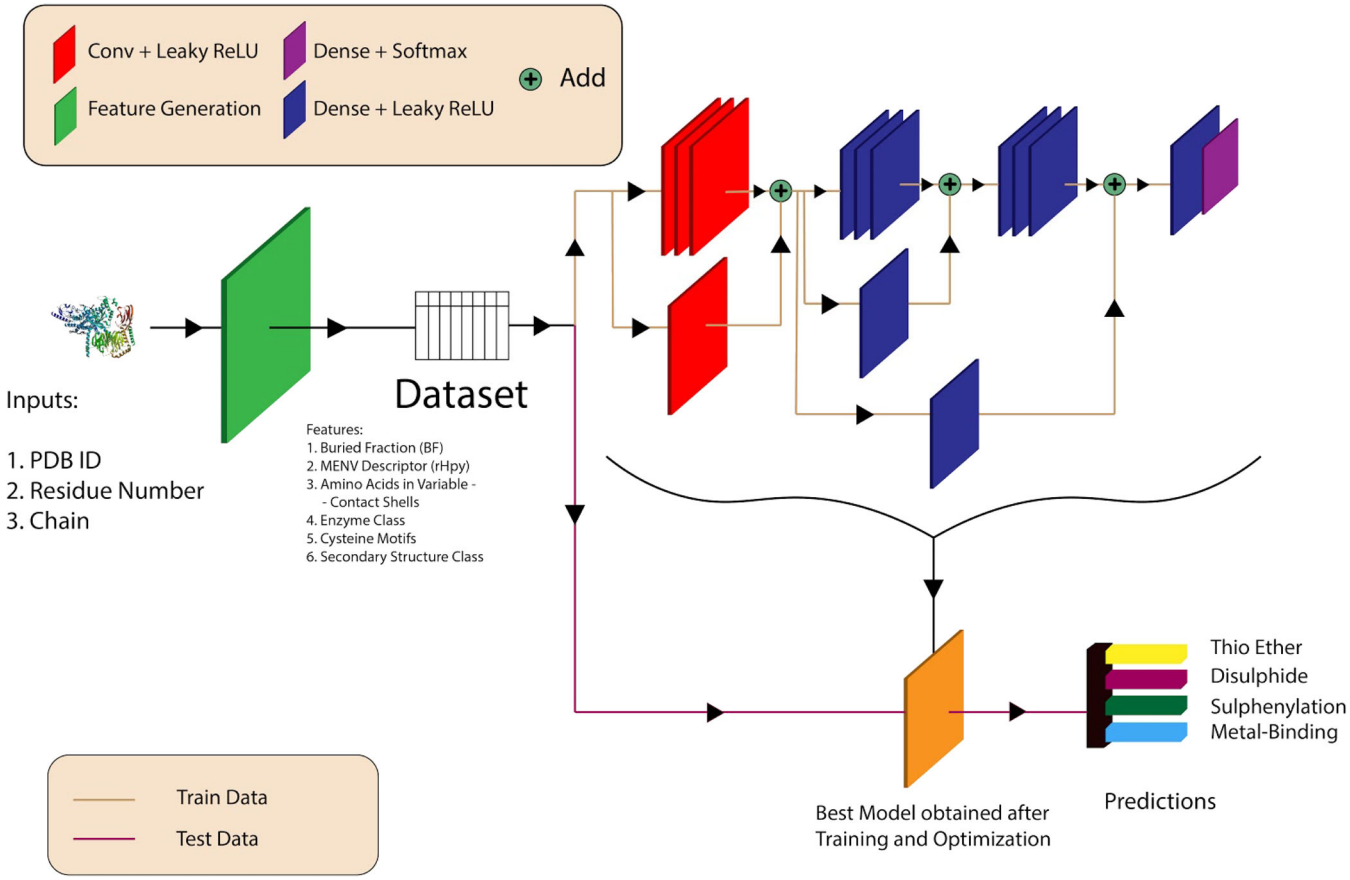

FIGURE 3 Workflow and architecture of the deep learning model for structure-based prediction

that resulted into a dip in macro-average of sensitivity values (77.1% and 76.7%, respectively) compared to the all-feature criteria (79.3%). This observation indicated that the feature selection can decrease complexity but might not, necessarily, improve accuracy.[78] Hence, we have used all-feature criteria to develop DeepCys model.

7.2 | Variation of the features across the cysteine modifications

7.2.1 | Features 1 and 2. Protein microenvironment (buried fraction and rHpy)

Prior to performing deep neural-network based cysteine function predictions, each feature was analyzed for the given dataset. The first feature, buried fraction, was used by many of the cysteine prediction functions, albeit, in slightly different way, that is, solvent accessibility.[32,44,47] The buried fraction was defined as the normalized surface area of the cysteine thiol group buried inside the protein. Whereas, solvent accessibility was the area on the protein surface that is probed accessible using a certain probe radius.[79,80] Buried fraction showed clear variation in terms of the mean and SD values across the four different modifications (Table S7). The mean buried fraction value of the disulphide modification indicated its presence in the most hydrophobic region of the protein structures. In contrast, the thioether modification was identified to be maximally exposed to the solvent, according to the mean value. The buried fraction values of metal-binding and sulphenylation modifications were comparable (Figure 4A). These observations were in accordance with our previous studies.[54,55]

The second feature, rHpy, was a measure of hydrophilicity of the microenvironment around a cysteine residue. Higher the value of rHpy, greater is the hydrophilicity of the surrounding microenvironment. The maximum limit of rHpy is one, indicating a complete aqueous environment. It is expected that the completely buried protein region will be more hydrophobic and the completely exposed protein region will be mostly hydrophilic. Based on the mean and SD of rHpy, disulphide modifications were embedded in hydrophobic microenvironment and thioether in a relatively more hydrophilic microenvironment (Table S8). The other two modifications had intermediate values, and those were comparable (Figure 4B).

7.2.2 | Feature 3: Secondary structure (SS) Motifs

Secondary structure motifs (also known as fold) are often conserved across the protein family and play an important role in various protein functions.[81,82] Many of the cysteine function prediction tools[23,32,47] used secondary structure as one of the features. To determine the

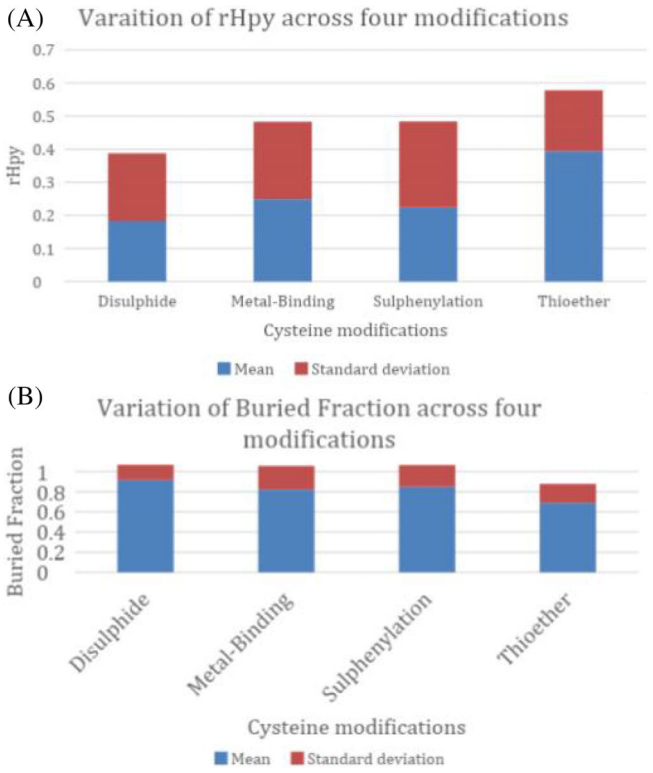

FIGURE 4 Variation in the four cysteine modifications for A, buried fraction and B, rHpy values [Color figure can be viewed at wileyonlinelibrary.com]

optimal length of the secondary structure folds around the cysteine of interest, a range of window sizes from 3 to 13 was scanned and DeepCys original model was tested on each window size. The window size of seven has produced the best result in the DeepCys model; this value of window size indicated secondary structures of 15 consecutive amino acids with the cysteine of interest at the centre (Figure S1). According to DSSP notations, each secondary structure was represented by one-letter code; alpha helix, beta bridge, strand, turn and bend and represented by H, B, E, T, and S, respectively. Thus, there were 108 334 arrays (each representing a cysteine residue in the training dataset original (Table 1a) and each array contained 15 secondary structure elements represented by one letter code. To understand the pattern of the secondary structure folds, the matrix of 108334 x 15 dimension was clustered using CD-HIT with a similarity cut off 70% (choice of this cut off was empirical), that is, each cluster has common secondary structure folds with 70% similarity. To compare across the clusters, normalized cluster size was computed using the number of secondary structure folds present in each cluster divided by the total number of instances in each cysteine modification (Tables S9). Broadly, 12 secondary structure folds were identified based on clustering (Figure 5). Preferences of the folds for each modification were shown (Table 5).

7.2.3 | Feature 4: Protein/enzyme class

Twelve major protein families and enzyme classes were noted in the training original dataset (Table 6). As we have only tabulated the major families and classes, summations of the rows are less than 100%. This analysis exhibited preferences of a specific cysteine modification toward a specific protein/enzyme class. Disulphide modifications predominantly belong to the hydrolase enzyme class and in immune systems, whereas other modifications were less frequently observed in these two protein/enzyme classes. Moreover, in toxin proteins, only modification (out the four mentioned here) observed was disulphide (Table 6). The role of disulphide linkages in toxin proteins was well established.[83,84] The metal-binding modifications were predominant in transferase, transcription factors and ligase enzyme classes (Table 6). The metal ions were reported to be essential in nucleic acid structure stabilization[59] and function, such as transcription regulation upon DNA/RNA binding.[85] Metal prosthetic groups containing iron, zinc and copper ions were identified in electron-transport chains.[61] The coordinate bond formed between cysteine thiolate and Fe(III) plays a pivotal role in the functions of various heme containing proteins, such as, P450, cytochrome C, hemoglobin.[86] The photosystems and electron transport chain proteins (cytochrome C is one such) contained thioether modifications only (Table 6). The presence of two thioether linkages in a conserved CXXCH motif[87] were well known in cytochrome C protein family involving the heme vinyl groups and the cysteine thiols.

7.2.4 | Feature 5: Amino acids within variable contact shells around a cysteine residue

The hydrophilicity around cysteine residue (denoted by rHpy) was computed within the first contact shell. However, the molecular interactions persist beyond the first contact shell.[71] Hence, contact shells with larger radii (6, 7, and 8 Å) were also considered in this feature. The frequencies of the amino acids around cysteine were defined as the number of times an amino acid appeared in a particular modification divided by the number of cysteines in that modification. The largest radius of 8 Å (describing the second contact shell[70]) was considered for comparison across the four modification (Figure 6). The most frequently occurring amino acids within other radii were also reported (Table S10). The most frequently occurring amino acid (within 8.0 Å) around disulphide modification was cysteine. Two cysteine in the proximity, happens to be half-cysteines those are involved in formation of disulphide bond. For metal-binding, the most frequently occurring amino acids within the second contact shell (8.0 Å) were cysteine and glycine. However, within the first contact shell (4.5 Å) the most populated amino acids were glycine and arginine. Occurrence of arginine-glycine rich motifs in mRNA-binding proteins and transferase enzymes were well known.[88] The highest frequencies of metal-binding cysteines, in the current dataset, were observed in transferase and transcription factors (Table 6).

Three of the modifications, namely, disulphide, thioether and suphenylation, in general, have showed higher preferences towards, beta-turn-beta secondary structure motif (Table 5) and exhibited highest content of either Cys/Ser, Cys/Gly, or Ser/Gly preferences (Figure 6). The turn and linker regions were reported earlier with high preferences toward, Cys, Ser, or Gly residues.[89]

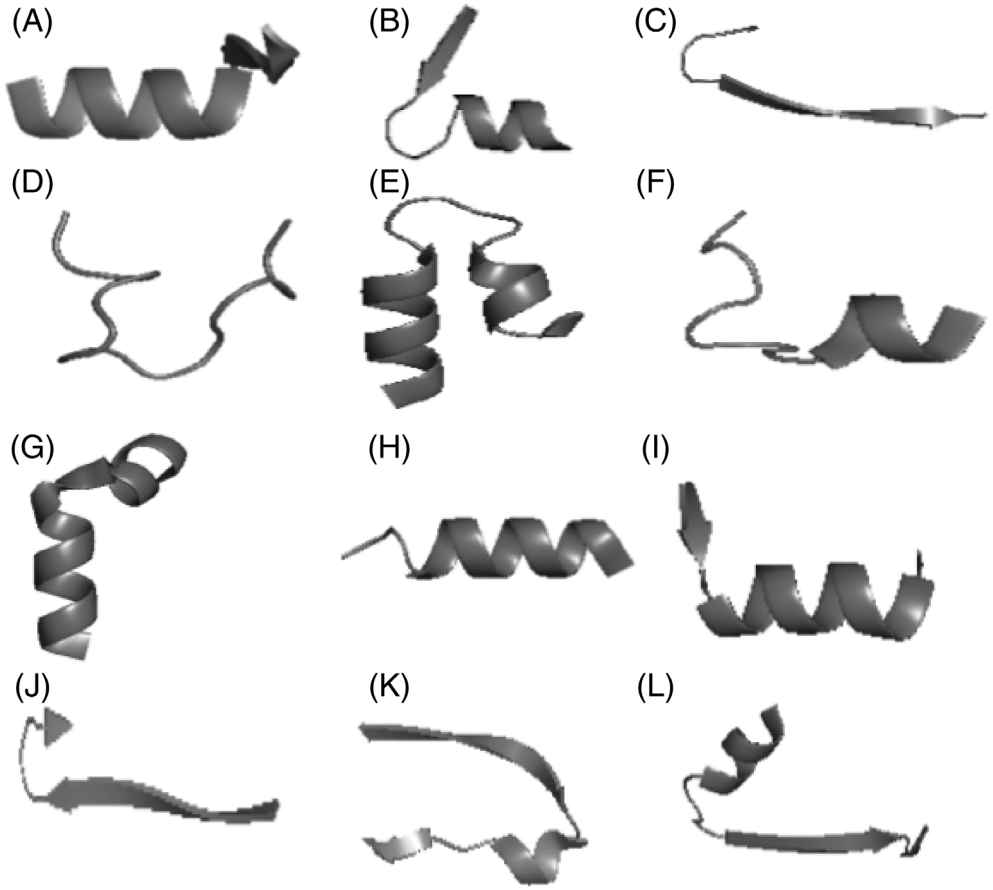

FIGURE 5 Different secondary structure folds identified from clustering analysis; A, helix beta; B, beta turn helix; C, turn beta; D, turn; E, beta helix turn helix; F, turn helix; G, helix turn helix; H, helix turn; I, beta helix; J, beta turn beta; K, beta-helix-beta; L, helix turn beta

TABLE 5 Percentage of different secondary structure motifs in four different modifications; helix, turn and beta represented by H, T, and B

Modification/ss motif	H-T-H	H-T	H-T-B	H-B	T-H	T-B	T	B-H-B	B-H-T-H	B-T-B	B-H	B-T-H
Disulphide	14.7	6.0	**28.9**	—	4.2	—	4.0	0.9	3.8	**28.6**	11.1	4.6
Metal-binding	—	—	9.3	12.8	10.4	—	—	5.8	—	**34.8**	—	13.9
Sulphenylation	19.2	**30.8**							23.1	**26.9**		
Thioether	**38.2**	—	—	—	**44.1**	8.8	—	—	—	—	—	8.8

Note: The Secondary Structure Motif with the highest percentage for each modification mentioned in bold.

TABLE 6 Occurrence of different cysteine modifications observed in different protein/enzyme classes, 1. Hydrolase, 2. Immune system, 3. Hydrolase inhibitor, 4. Oxidoreductase, 5. Toxin, 6. Transferase, 7. Transcription, 8. Ligase, 9. Lyase, 10. Sugar binding, 11. Electron transport, 12. photosynthesis

Modification	1	2	3	4	5	6	7	8	9	10	11	12
Disulphide	*31.8*	*13.6*	*8.5*	*4.9*	*3.0*	0.1	0.1	0.1	0.1	2.2	0.5	0.0
Metal-binding	10.1	0.2	0.0	**20.2**	0.1	15.1	10.1	4.2	1.9	0.1	1.9	0.0
Sulphenylation	17.4	1.2	0.0	**21.6**	0.0	14.8	0.4	0.4	7.4	4.1	0.9	2.4
Thioether	6.1	1.2	5.7	**28.4**	0.0	14.9	0.4	0.4	7.4	4.1	*18.7*	*7.2*

Note: Values reported in percentage of (number of times a modification is present in a protein/enzyme class)/total number of modifications). The enzyme class with the highest percentage for each modification is mentioned in bold. The other high percentage values per modification (along each row) are shown in italics.

7.2.5 | Feature 6: Cysteine sequence motifs

Cysteine sequence motifs were mostly associated with metal-binding and thioether modifications. Several of the cysteine metal-binding prediction tools used the metal binding sites in the proteins.[34-36] Eight sequence motifs reported earlier[54] were used to generate the sixth feature. The highest window size of 13 was employed to perform the motif search. The frequencies of these eight motifs were calculated

for each of the four modifications (Table 7). The frequency (in percentage) was defined as the number of times a motif appeared in a modification divided by the number of instances in that particular modification. It is not necessary that in all the instances, any of these eight motifs would be present. Hence, summation of the motif percentages was observed to be less than 100% for all the cysteine modifications. As per our analyses, disulphide modification was ubiquitous and has not shown preferences towards a particular motif. Presence of almost all the motifs were identified in the metal binding modification. The CXXC motif was also prevalent in thioether modification. Most of the observed thioether modifications in cytochrome C family proteins were part of CXXCH motif.[87]

7.2.6 | Evaluation of DeepCys model performance

The DeepCys model was trained on the three variants of the training datasets, namely, training original, training 100%, and training 30%. Here we have studied the performances of these three models evaluated on the test dataset to understand these two opposing effects. As per the hypothesis, the models trained on the nonredundant datasets should better perform compared to DeepCys original model. On the other hand, the significant reduction of the dataset size may reduce the performance.

All the three DeepCys models, namely, DeepCys original, DeepCys 100%, and DeepCys 30% were tested on test dataset. Overall performances of the models, measured in terms of the macro-average values (see method section for definition) monotonically decreased from DeepCys original to DeepCys 30% (79.25%, 78.25%, and 74%, respectively) indicating that the overall performance was related to the size of the dataset and not to overfitting. However, the sensitivity values for individual modifications did not follow the same trend of the overall performance (Table 8). Although sensitivity metrics was considered, other two metrics were also computed for each modification (Table S11). The sensitivity values for disulphide and thioether monotonically decreased for the three models following the trend of the overall performance, whereas that of sulphenylation remained more or less the same. In case of metal-binding modification, sensitivity values monotonically increased from DeepCys original to DeepCys 100%, exactly reverse of the overall trend. This observation could be, plausibly, explained based on the types and the frequencies of metal ions present in training datasets (Table 2) vs test dataset (Table 4). These two independent datasets have significant variation in types and frequencies of metal ions. For example, in all the three training datasets, Zn^{2+} ion propensity was more than 72%, whereas in test dataset that was only 69%. Moreover, the heavy metal ions, like, Hg^{2+} and Cd^{2+} were almost negligible, in training datasets, in contrast to a significant population of those heavy metal ions, in test dataset. The Pb^{2+} ion was present in test dataset, that was non-existent in the training dataset. On the other hand, other three modifications, namely, disulphide, thioether or sulphenylations, were trained and tested on the same data type.

The DeepCys original model has produced the highest overall performance in terms of the macro-average value of sensitivity. Hence, DeepCys original model was chosen as the optimal model on the test dataset.

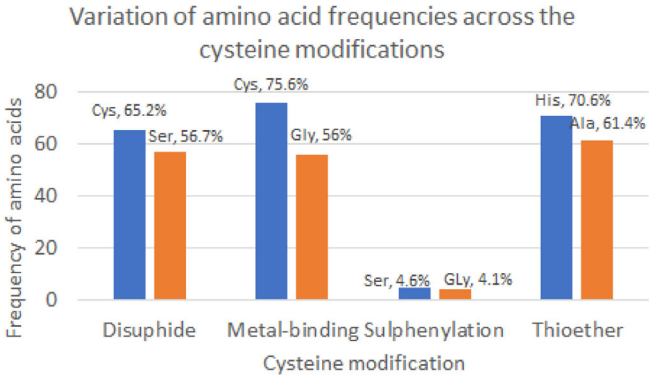

FIGURE 6 Variation in the four cysteine modifications for the most frequently observed amino acids within the contact shell of 8 Å radius around cysteine [Color figure can be viewed at wileyonlinelibrary.com]

7.3 | Comparison of the current and the existing cysteine prediction methods on "test sample dataset"

The model developed in this work (DeepCys original) was compared with the published literature, for both specific and multiple cysteine function prediction methods. Three different types of specific cysteine functions were tested, namely, metal-binding, disulphide and sulphenylation. Following prediction methods were tested - metal ion binding site prediction and docking server (MIB),[30] PSIPRED-METSITE[32] and MetalDetector v2.0,[31] for metal binding; DISULFIND[22] and Cyscon,[23] for disulphide and SulCysSite[41] and DeepCSO,[90] for sulphenylation. No explicit prediction method was known, to the best of our knowledge, for thioether, although DiANNA has implicitly indicated the predictability of thioether modification.[50]

TABLE 7 Percentage (%) of eight different motifs, 1. CC, 2. CXC, 3. CXXC, 4. CXXXC, 5. CXXXXC, 6. CXXCXXC, 7. CXXCXXXXXC, 8. CXXCXXCXXXC, in four cysteine modifications

Modification	1	2	3	4	5	6	7	8
Disulphide	7.0	7.3	4.2	9.2	9.4	0.2	0.1	0.0
Thioether	1.1	1.5	**54.4**	1.4	2.9	0.0	1.3	0.0
Metal-binding	9.7	14.0	**49.4**	15.1	17.0	3.6	2.0	0.3
Sulphenylation	6.8%	4.9	7.4	5.4	8.1	0.0	0.0	0.0

Note: The two largest percentages are indicated in bold.

TABLE 8 Performance of DeepCys original, DeepCys 100% and DeepCys 30% on test dataset

Modification	DeepCys original	DeepCys 100%	DeepCys 30%
Disulphide	**87%**	76%	71%
Thioether	**75%**	74%	62%
Metal-binding	72%	78%	**80%**
Sulphenylation	83%	**85%**	83%

Note: The performance was measured in terms of sensitivity. The model with the highest sensitivity for each modification has been mentioned in bold.

TABLE 9 Comparative analysis of DeepCys original model with cysteine function prediction methods. The performance was measured in terms of sensitivity on the sample test 1 dataset with A) specific cysteine function prediction algorithm and B) general cysteine function prediction algorithms

(A) Specific cysteine function prediction algorithm

Function	Deep-Cys	DISULFIND	Cyscon	Metal Detector V2.0	MIB	PSIPRED-METSITE	SulCysSite	DeepCSO
Disulphide	**96%**	34%	80%	—	—	—	—	—
Metal-binding	**81%**	—	—	**81%**	56%	16%	—	—
Sulphenylation	**71%**	—	—	—	—	—	28%	32%
Thioether	**67%**	—	—	—	—	—	—	—

(B) General cysteine function prediction algorithms

Modification	DeepCys	DiANNA	Cy-preds
Disulphide	96%	40% (5%) [a]	**98%**
Thioether	**67%**	0%	—
Metal-binding	81%	39% (3%) [b]	**95%**
Sulphenylation	**71%**	—	—

Note: The model with the highest sensitivity for each modification has been mentioned in bold.
Abbreviation: DiANNA, diamino acid neural network application.
[a]Within parenthesis, the half-cysteine vs free-cysteine option was used and outside parenthesis, disulphide connectivity was used.
[b]Within parenthesis, the ligand bound vs free cysteine option was used and outside parenthesis ligand bound vs half-cysteine was used.

For multiple cysteine function prediction, DiANNA[50] and Cy-preds[53] were used.

As most of the existing models were presented in terms of web servers, it was formidable task to compare all the 125 652 datapoints of test dataset, manually on webservers. Hence, a random sample of 100 datapoints were selected per modifications from the test dataset and used consistently for comparison across the methods, termed as "test sample dataset" (Table 9A). As there were no tools available explicitly for thioether prediction, DeepCys performance on thioether cannot be compared. The DeepCys original outperformed other specific cysteine function prediction methods for three other modifications on test sample dataset.

For disulphide prediction studies, the prediction tool DISULFIND was based on protein sequence only and its performance was lower compared to that of DeepCys. Cyscon, the other disulphide prediction server from Zhang's lab, exploited machine learning approach developed on protein structural information. However, it has lower performance compared to DeepCys on test sample dataset.

For metal binding studies, there were six metal ions present in the training and test datasets, and those widely varied in terms of charge, radius and chemical properties. Hence, the prediction performance was likely to vary from one metal ion to the other. In the current test sample dataset, the cations identified were either Cu^{2+} or Cd^{2+}. The DeepCys model has showed 86% and 76% sensitivity values for these two metal ions. The reported sensitivity values for Cu^{2} and Cd^{2+} ions, by MIB were 85.6% and 41.2%, respectively on their original dataset and 76% and 36%, respectively on the "test sample dataset." Hence, overall performance of MIB on the test sample dataset was poor compared to that of DeepCys. The results obtained from MetalDetectorv2.0, on test sample dataset, was comparable to that of DeepCys. The PSIPRED-METSITE only predict for Cu^{2+} and not for Cd^{2+}. The prediction accuracy by PSIPRED-METSITE for Cu^{2+} was only 16%.

In terms of sulphenylation prediction, the sensitivity value reported by SulCysSite (a sequence-based method) was only 62.89%. The newly developed, deep learning-based approach, DEEPCSO, for sulphenylation prediction yielded 32% of sensitivity value, on the test sample dataset. In the original work of DEEPCSO, the sensitivity was reported at 71.7%, that was comparable to the DeepCys original. Thus, DeepCys outperformed both the existing methods, on test sample dataset.

The multiple cysteine function prediction method, DiaNNA can predict three cysteine functions. The DiaNNA original work published in 2005, attempted to predict multiple functions, the first attempt of its kind. However, that work have separately predicted ligand-binding states and disulphide connectivity, unlike, DeepCys that is a comprehensive model to predict any of the four modifications simultaneously. There were two options in DiaNNA, (a) disulphide connectivity and (b) ternary classification. Under ternary classification, there are three options, ligand-bound vs half-cystine, ligand-bound vs free cysteine and half-cystine vs free cysteines. The ligand-bound vs half-cystine option compared the probabilities of ligand-bound state to disulphide connectivity. The other option was comparison of probabilities among ligand-bound state and free thiol state. The DiANNA original work has reported only 41.8% success in prediction of disulphide bonds in proteins, that was based on protein sequence feature only. For test sample dataset, performance of DiANNA was comparable with their original work when, disulphide connectivity option was considered. The result was very poor when half-cystine vs free cysteine option was explored for disulphide modification (Table 9B). DiaNNA has defined four ligand types, namely, Fe^{2+}/Fe^{3+}, Zn^{2+}, Cd^{2+} ions, and carbon atoms. Presumably, the last one indicated thioether formation, although, thioether was not explicitly mentioned in the original work. Both ligand-bound vs half-cystine and ligand-bound vs free cysteine options were explored, the first one performed better than the second one. Individual Cu^{2+} and Cd^{2+} ion sensitivities reported by DiaNNA were 40% and 38% respectively, using ligand-bound vs half-cystine option. The results obtained from the second option was very poor. Thioether prediction was tested for DiaNNA on test sample dataset, yielding 0% sensitivity, indicating that it was unable to capture thioether formation. The other multiple cysteine function prediction algorithm, Cy-preds, was able to predict three functions, namely, disulphide, metal-binding and post-translational modifications. Disulphide prediction performance of Cy-preds on test sample dataset was comparable to that of DeepCys, although, the former one was slightly better. For metal-binding prediction, Cy-preds exhibited better performance than that of DeepCys with respect to the test sample dataset (Table 9). However, the DeepCys was able to predict any four cysteine functions, in contrast to Cy-preds, predicting only three.

7.4 | Elucidation of cysteine function in DUF proteins belonging to cytochrome C oxidase subunit II-like transmembrane region

In cytochrome C oxidase protein, there were multiple chains containing several cysteines, involved in different functions, such as Zn^{2+} ion-binding in chains F and S (subunits VB), Cu^{2+} ion-binding in chains B and O (subunit II) and formation of disulphide bonds, in chains H and U (subunits VIB1) and so forth. The subunit II contains binuclear copper center that is the primary electron acceptor from reduced cytochrome C. There were six proteins selected from DUF ID, belonging to cyctochrome C subunit II like transmembrane region, namely, (a) catalytic core (subunits I and II) of cytochrome c oxidase from rhodobacter sphaeroides (PDB:2gsm), (b) bovine heart cytochrome C oxidase modified by dccd (PDB:2dys), (c) bovine heart cytochrome C oxidase in azide-bound state (PDB:1ocz), (d) cytochrome c oxidase from rhodobactor sphaeroides (Wild type) (PDB:1 m56), (e) The aberrant BA3-cytochrome-C oxidase from thermus thermophilus (pdb:1ehk), and (f) the paracoccus denitrificans two-subunit cytochrome C oxidase complexed with an antibody Fv fragment (PDB: 1ar1). Subunit II (COX2) were common in all these proteins. Total 66 cysteine residues were present in these pdb files, out of those 66, 36 were metal-binding and remaining 30 were disulphide modifications. DeepCys original model prediction was 80% correct for disulphide and 94.4% for metal-binding (Table S12). In case of metal-binding cysteine, only two were incorrectly predicted, the binuclear copper center, chain B:cys^{216} and chain B:cys^{220}, from pdb:1ar1. In case cys^{220}, the prediction probabilities of cysteine and thioether were equivalent, 0.41; changes occurred at the third decimal place. Analysis of the six features (high BF, 1.0, low rHpy, helix-turn-helix motif, oxidoreductase enzyme class, presence of His in the vicinity of cys and cxxc motif) around Cys^{220}, indicated equal probabilities of thioether and metal-binding modifications. Thus, it revealed that the prediction probabilities by DeepCys were highly dependent on the cysteine features.

7.5 | Web applications, backend calculations, and standalone program

A user-friendly web application, DeepCys original model, was built using the Django web framework. The flowchart of the web application was shown (Figure 7A). The web application (https://deepcys.herokuapp.com/) was deployed using Heroku, a container-based cloud platform as a service. The structure-based prediction model can be accessed by clicking the "structure prediction" button on the navigation bar. The web application has a "form" that requests three inputs, corresponding to a cysteine, namely, (a) PDB ID of the protein, (b) chain, and (c) residue ID of the cysteine (Figure 7B). These three parameters were relayed to the deep learning model. Before running the prediction model, the six features of the cysteine residue were either computed or extracted from PDB file. The deep learning model used the six features to predict the probability of a cysteine modification. The prediction outputs four probability values along with the final prediction of the cysteine modification (Figure 7C).

In addition to the web application, a batch prediction model was developed. The purpose of the batch prediction model was to make multiple predictions at a time in contrast to single predictions on the web application. The batch prediction model requires the input file containing the above three parameters corresponding to a cysteine residue. The installation and usage instruction were available on the GitHub repository. In addition, the python codes for feature generation, extraction and model training were also available on GitHub repository (https://github.com/vam-sin/deepcys).

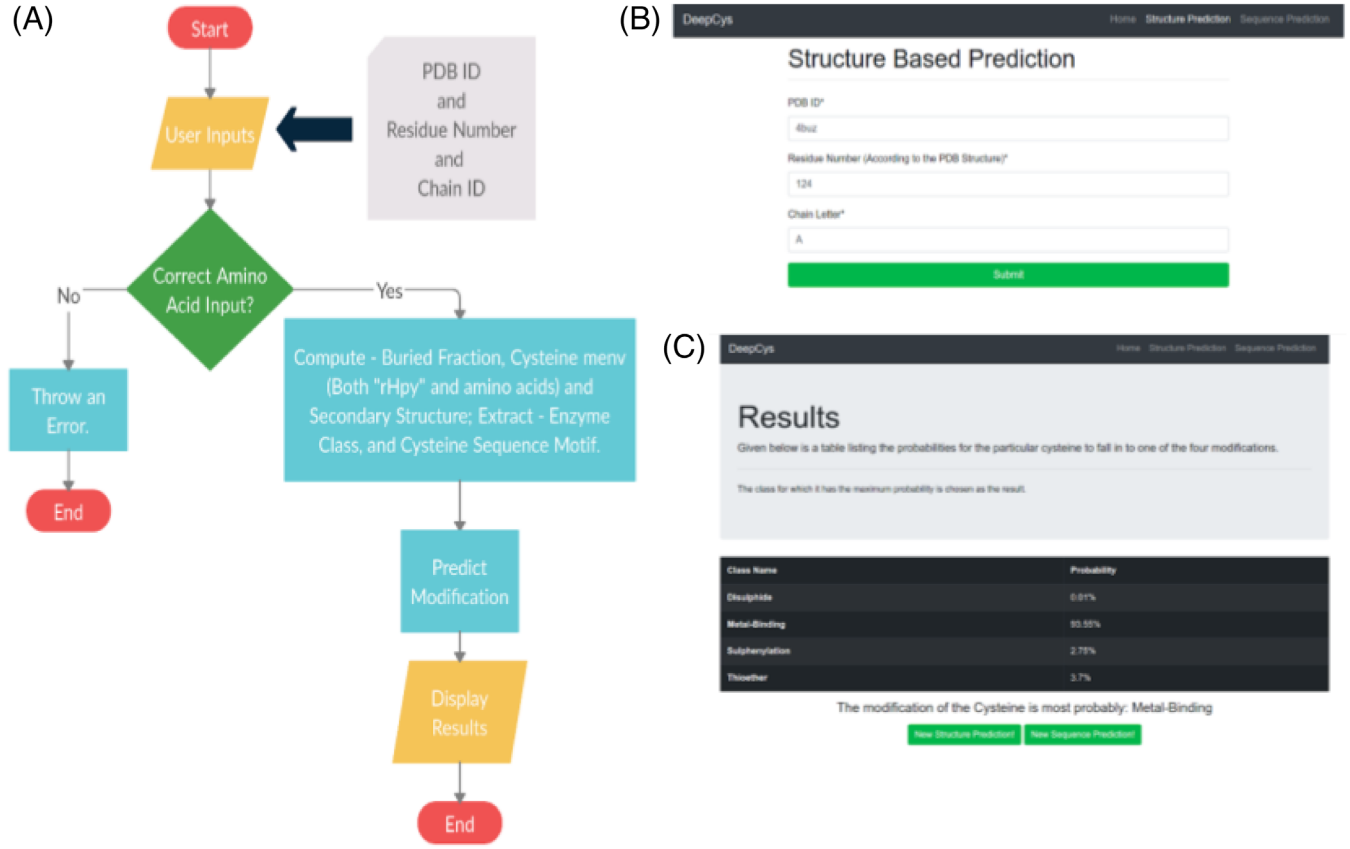

FIGURE 7 Web application for structure-based prediction A, flow chart B, input "form" C, output results [Color figure can be viewed at wileyonlinelibrary.com]

8 | CONCLUSION

Cysteine thiol group is highly reactive. It participates in different biochemical reactions leading to multiple modifications. Accurate prediction of these modifications is crucial to elucidate the cysteine functions, particularly, in proteins of unknown functions and DUFs. In this study, we present a deep learning-based approach to predict any one of the four most abundant cysteine modifications. Novelty of this work was prediction of maximum number of cysteine modifications. Moreover, thioether prediction was not attempted earlier. The DeepCys model developed in this work requires the protein structure in PDB format, the residue number and the chain identifier. Six features were either extracted or computed from PDB file those were used by deep learning approach. The final output was the probability values for four cysteine modifications, namely, disulphide, metal-binding, thioether, and sulphenylation. The modification with highest probability was reported as the predicted modification. The current prediction was benchmarked across the existing cysteine prediction tools. The DeepCys performance was better than most of the existing methods, for the given dataset. The tool is available both as a webserver and as a standalone program.

ACKNOWLEDGMENT

The authors gratefully acknowledge Professor Asis Sengupta, Indian Statistical Institute, Kolkata for his value suggestions and discussions.

DATA AVAILABILITY STATEMENT

DeepCys standalone program is available on GitHub (https://github.com/vam-sin/deepcys).

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26056.

ORCID

Debashree Bandyopadhyay https://orcid.org/0000-0003-4131-907X

REFERENCES

1. Gutteridge A, Thornton JM. Understanding Nature's catalytic toolkit. *Trends Biochem Sci*. 2005;30(11):622-629. https://doi.org/10.1016/J.TIBS.2005.09.006.
2. Marino SM, Gladyshev VN. Analysis and functional prediction of reactive cysteine residues. *J Biol Chem*. 2012;287(7):4419-4425. https://doi.org/10.1074/jbc.R111.275578.
3. Anfinsen CB. CIE: NCES folding of protein chains. *Science*. 1973;181(4096):223-230.
4. Henehan CJ, Pountney DL, Vašák M, Zerbe O. Identification of cysteine ligands in metalloproteins using optical and NMR spectroscopy: cadmium-substituted rubredoxin as a model [cd(CysS)4]2-center. *Protein Sci*. 1993;2(10):1756-1764. https://doi.org/10.1002/pro.5560021019.
5. Berg JM. Zinc finger domains: hypotheses and current knowledge. *Annu Rev Biophys Biophys Chem*. 1990;19(1):405-421. https://doi.org/10.1146/annurev.bb.19.060190.002201.

6. Kulkarni RA, Worth AJ, Zengeya TT, et al. Discovering targets of non-enzymatic acylation by thioester reactivity profiling. *Cell Chem Biol*. 2017;24(2):231-242. https://doi.org/10.1016/j.chembiol.2017.01.002.
7. Forrester MT, Hess DT, Thompson JW, et al. Site-specific analysis of protein S-acylation by resin-assisted capture. *J Lipid Res*. 2011;52(2):393-398. https://doi.org/10.1194/jlr.D011106.
8. Guan X, Fierke CA. Understanding protein Palmitoylation: biological significance and enzymology. *Sci China Chem*. 2011;54:1888-1897. https://doi.org/10.1007/s11426-011-4428-2.
9. Martin BR. Nonradioactive analysis of dynamic protein Palmitoylation. *Curr Protoc Protein Sci*. 2013;73(suppl 1):14.15.1-14.15.9. https://doi.org/10.1002/0471140864.ps1415s73.
10. Morrison E, Wegner T, Zucchetti AE, et al. Dynamic palmitoylation events following T-cell receptor signaling. *Commun Biol*. 2020;3(1):1-9. https://doi.org/10.1038/s42003-020-1063-5.
11. Uchida K, Stadtman ER. Selective cleavage of thioether linkage in proteins modified with 4-hydroxynonenal. *Proc Natl Acad Sci U S A*. 1992;89(12):5611-5615. https://doi.org/10.1073/pnas.89.12.5611.
12. Kang X, Carey J. Role of heme in structural organization of cytochrome c probed by semisynthesis †. *Biochemistry*. 1999;38:15944-15951. https://doi.org/10.1021/bi9919089.
13. Apel K, Hirt H. Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annu Rev Plant Biol*. 2004;55:373-399. https://doi.org/10.1146/annurev.arplant.55.031903.141701.
14. Farooq MA, Niazi AK, Akhtar J, et al. Acquiring control: the evolution of ROS-induced oxidative stress and redox signaling pathways in plant stress responses. *Plant Physiol Biochem*. 2019;1:353-369. https://doi.org/10.1016/j.plaphy.2019.04.039.
15. Hameister R, Kaur C, Dheen ST, Lohmann CH, Singh G. Reactive oxygen/nitrogen species (ROS/RNS) and oxidative stress in arthroplasty. *J Biomed Mater Res Part B Appl Biomater*. 2020;108(5):2073-2087. https://doi.org/10.1002/jbm.b.34546.
16. Shen J, Zhang J, Zhou M, et al. Persulfidation-based modification of cysteine desulfhydrase and the NADPH oxidase RBOHD controls guard cell abscisic acid signaling. *Plant Cell*. 2020;32(4):1000-1017. https://doi.org/10.1105/tpc.19.00826.
17. Parker WR, Brodbelt JS. Characterization of the cysteine content in proteins utilizing cysteine selenylation with 266 nm ultraviolet photodissociation (UVPD). *J Am Soc Mass Spectrom*. 2016;27(8):1344-1350. https://doi.org/10.1007/s13361-016-1405-1.
18. Williamson AR. Creating a Structural Genomics Consortium. *Nat Struct Biol*. 2000;7:953. https://doi.org/10.1038/80726.
19. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235-242.
20. Silva PFF, Novaes E, Pereira M, Soares CMA, Borges CL. In Silico characterization of hypothetical proteins from paracoccidioides lutzii. *Genet Mol Res*. 2015;14(4):17416-17425.
21. Niehaus TD, Thamm AMK, De Crécy-lagard V, Hanson AD. Proteins of unknown biochemical function: a persistent problem and a roadmap to help overcome it 1. *Plant Physiol*. 2015;169(November):1436-1442. https://doi.org/10.1104/pp.15.00959.
22. Ceroni A, Passerini A, Vullo A, Frasconi P. Disulfind: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res*. 2006;34(WEB. SERV. ISS):177-181. https://doi.org/10.1093/nar/gkl266.
23. Yang J, He B-J, Zhang R, Zhang Y, Shen H-B. Accurate disulfide-bonding network predictions improve Ab initio structure prediction of cysteine-rich proteins. *Bioinformatics*. 2015;31(23):3773-3781.
24. Craig DB, Dombkowski AA. Disulfide by design 2.0: a web-based tool for disulfide engineering in proteins. *BMC Bioinf*. 2013;14(1):6. https://doi.org/10.1186/1471-2105-14-346.
25. Chen S, Pellequer J. Identification of functionally important residues in proteins using comparative models. *Curr Med Chem*. 2004;11(5):595-605. https://doi.org/10.2174/0929867043455891.
26. Cheng J, Saigo H, Baldi P. Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins Struct Funct Bioinf*. 2005;62(3):617-629. https://doi.org/10.1002/prot.20787.
27. Lin HH, Tseng LYDBCP. A web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines. *Nucleic Acids Res*. 2010;38(suppl 2):W503-W507. https://doi.org/10.1093/nar/gkq514.
28. Laimer J, Hiebl-Flach J, Lengauer D, Lackner P. MAESTROweb: a web server for structure-based protein stability prediction. *Bioinformatics*. 2016;32(9):1414-1416. https://doi.org/10.1093/bioinformatics/btv769.
29. Yaseen A, Li Y. Dinosolve: a protein disulfide bonding prediction server using context-based features to enhance prediction accuracy. *BMC Bioinf*. 2013;14(suppl 13):S9. https://doi.org/10.1186/1471-2105-14-S13-S9.
30. Lin YF, Cheng CW, Shih CS, Hwang JK, Yu CS, Lu CHMIB. Metal ion-binding site prediction and docking server. *J Chem Inf Model*. 2016;56(12):2287-2291. https://doi.org/10.1021/acs.jcim.6b00407.
31. Passerini A, Lippi M, Frasconi P. MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. *Nucleic Acids Res*. 2011;39:W288-W292.
32. Buchan DWA, Jones DT. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res*. 2019;47(W1):W402-W407. https://doi.org/10.1093/nar/gkz297.
33. Song J, Li C, Zheng C, Revote J, Zhang Z, Webb GI. MetalExplorer, a bioinformatics tool for the improved prediction of eight types of metal-binding sites using a random forest algorithm with two- step feature selection. *Curr Bioinf*. 2016;12(6):480-489. https://doi.org/10.2174/2468422806666160618091522.
34. Putignano V, Rosato A, Banci L, Andreini C. MetalPDB in 2018: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res*. 2018;46(D1):D459-D464. https://doi.org/10.1093/nar/gkx989.
35. Valasatava Y, Rosato A, Cavallaro G, Andreini C. MetalS3, a database-mining tool for the identification of structurally similar metal sites. *J Biol Inorg Chem*. 2014;19(6):937-945. https://doi.org/10.1007/s00775-014-1128-3.
36. Valasatava Y, Rosato A, Banci L, Andreini C. MetalPredator: a web server to predict iron-sulfur cluster binding proteomes. *Bioinformatics*. 2016;32(18):2850-2852. https://doi.org/10.1093/bioinformatics/btw238.
37. Hu X, Dong Q, Yang J, Zhang Y. Recognizing metal and acid radical ion-binding sites by integrating Ab initio modeling with template-based transferals. *Bioinformatics*. 2016;32(21):3260-3269. https://doi.org/10.1093/bioinformatics/btw396.
38. He W, Liang Z, Teng M, Niu L. MFASD: a structure-based algorithm for discriminating different types of metal-binding sites. *Bioinformatics*. 2015;31(12):1938-1944. https://doi.org/10.1093/bioinformatics/btv044.
39. Sobolev V, Edelman M. Web tools for predicting metal binding sites in proteins. *Isr J Chem*. 2013;53(3–4):166-172. https://doi.org/10.1002/ijch.201200084.
40. Xu Y, Ding J, Wu L-Y. ISulf-Cys: prediction of S-Sulfenylation sites in proteins with physicochemical properties of amino acids. *PLoS One*. 2016;11(4):e0154237. https://doi.org/10.1371/journal.pone.0154237.
41. Hasan MM, Guo D, Kurata H. Computational identification of protein S-Sulfenylation sites by incorporating the multiple sequence features information. *Mol Biosyst*. 2017;13(12):2545-2550. https://doi.org/10.1039/c7mb00491e.
42. Bui VM, Lu CT, Ho TT, Lee TY. MDD-SOH: exploiting maximal dependence decomposition to identify S-Sulfenylation sites with substrate motifs. *Bioinformatics*. 2016;32(2):165-172. https://doi.org/10.1093/bioinformatics/btv558.
43. Wang X, Li C, Li F, Sharma VS, Song J, Webb GISIMLIN. A bioinformatics tool for prediction of S-Sulphenylation in the human proteome

based on multi-stage ensemble-learning models. *BMC Bioinf*. 2019; 20(1):602. https://doi.org/10.1186/s12859-019-3178-6.
44. Deng L, Xu X, Liu H. PredCSO: an ensemble method for the prediction of S-Sulfenylation sites in proteins. *Mol Omi*. 2018;14(4):257-265. https://doi.org/10.1039/c8mo00089a.
45. Wang M, Cui X, Yu B, Chen C, Ma Q, Zhou HSS-GTB. Identification of protein S-Sulfenylation sites by fusing multiple feature information and gradient tree boosting. *Neural Comput Appl*. 2020;32(17):13843-13862. https://doi.org/10.1007/s00521-020-04792-z.
46. Ju Z, Wang SY. Prediction of S-Sulfenylation sites using MRMR feature selection and fuzzy support vector machine algorithm. *J Theor Biol*. 2018;457:6-13. https://doi.org/10.1016/j.jtbi.2018.08.022.
47. Sakka M, Tzortzis G, Mantzaris MD, et al. PRESS: PRotEin S-Sulfenylation Server. *Bioinformatics*. 2016;32(17):2710-2712. https://doi.org/10.1093/bioinformatics/btw301.
48. Bui VM, Weng SL, Lu CT, Chang TH, Weng JTY, Lee TY. SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-Sulfenylation sites. *BMC Genomics*. 2016; 17(1):9. https://doi.org/10.1186/s12864-015-2299-1.
49. Butt AH, Khan YD. Prediction of S-Sulfenylation sites using statistical moments based features via CHOU'S 5-step rule. *Int J Pept Res Ther*. 2020;26(3):1291-1301. https://doi.org/10.1007/s10989-019-09931-2.
50. Ferre F, Clote P. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res*. 2005;33:W230-W232. https://doi.org/10.1093/nar/gki412.
51. Sanchez R, Riddle M, Woo J, Momand J. Prediction of reversibly oxidized protein cysteine thiols using protein structure properties. *Protein Sci*. 2008;17(3):473-481. https://doi.org/10.1110/ps.073252408.
52. Fetrow JS. Active site profiling to identify protein functional sites in sequences and structures using the deacon active site profiler (DASP). *Curr Protoc Bioinf*. 2006;Chapter 8. https://doi.org/10.1002/0471250953.bi0122s42.
53. Soylu I, Marino MS. Cy-preds: an algorithm and a web service for the analysis and prediction of cysteine reactivity. *Proteins Struct Funct Bioinf*. 2016;84(2):278-291.
54. Bhatnagar A, Bandyopadhyay D. Characterization of cysteine thiol modifications based on protein microenvironments and local secondary structures. *Proteins Struct Funct Bioinf*. 2018;86(2):192-209. https://doi.org/10.1002/prot.25424.
55. Bhatnagar A, Apostol MI, Bandyopadhyay D. Amino acid function relates to its embedded protein microenvironment: a study on disulfide-bridged cystine. *Proteins Struct Funct Bioinf*. 2016;84(11): 1576-1589. https://doi.org/10.1002/prot.25101.
56. Bandyopadhyay D, Mehler EL. Quantitative expression of protein heterogeneity: response of amino acid side chains to their local environment. *Proteins Struct Funct Bioinf*. 2008;72(2):646-659. https://doi.org/10.1002/prot.21958.
57. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22 (13):1658-1659. https://doi.org/10.1093/bioinformatics/btl158.
58. Bhattacharyya R, Pal D, Chakrabarti P. Disulfide bonds, their stereospecific environment and conservation in protein structures. *Protein Eng Des Sel*. 2004;17(11):795-808. https://doi.org/10.1093/protein/gzh093.
59. Mazmanian K, Dudev T, Lim C. How first shell–second shell interactions and metal substitution modulate protein function. *Inorg Chem*. 2018;57:14052-14061.
60. Kuppuraj G, Dudev M, Lim C. Factors governing metal-ligand distances and coordination geometries of metal complexes. *J Phys Chem B*. 2009;113(9):2952-2960. https://doi.org/10.1021/jp807972e.
61. Balsa E, Marco R, Perales-Clemente E, et al. NDUFA4 is a subunit of complex IV of the mammalian electron transport chain. *Cell Metab*. 2012;16(3):378-386. https://doi.org/10.1016/j.cmet.2012.07.015.
62. VAN KUILENBURG, P AB, Dekker HL, et al. Isoforms of human cytochrome-c oxidase: subunit composition and steady-state kinetic properties. *Eur J Biochem*. 1991;199(3):615-622. https://doi.org/10.1111/j.1432-1033.1991.tb16162.x.
63. Pandit SB, Gosar D, Abhiman S, et al. Database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res*. 2002;30(1):289-293. https://doi.org/10.1093/nar/30.1.289.
64. Radak BK, Chipot C, Suh D, et al. Constant-PH molecular dynamics simulations for large biomolecular systems. *J Chem Theory Comput*. 2017;13(12):5933-5944. https://doi.org/10.1021/acs.jctc.7b00875.
65. Brooks BR, Brooks CL, Mackerell AD, et al. The biomolecular simulation program. *J Comput Chem*. 2009;30(10):1545-1614. https://doi.org/10.1002/jcc.21287.
66. Rekker. *The Hydrophobic Fragmental Constant*. Amsterdam: Elsevier; 1977.
67. Pascual-ahuir JL, Silla E, Tuñon I. GEPOL: an improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface. *J Comput Chem*. 1994;15(10):1127-1138. https://doi.org/10.1002/jcc.540151009.
68. Joosten RP, Te Beek TAH, Krieger E, et al. A series of PDB related databases for everyday needs. *Nucleic Acids Res*. 2011;39(Database issue):D411-D419. https://doi.org/10.1093/nar/gkq1105.
69. Kabsch W, Sander C. How good are predictions of protein secondary structure? *FEBS Lett*. 1983;155(2):179-182.
70. Lábas A, Bakó I, Oláh J. Hydration sphere structure of proteins: a theoretical study. *J Mol Liq*. 2017;238:462-469. https://doi.org/10.1016/j.molliq.2017.05.038.
71. Freiberger MI, Guzovsky BA, Wolynes PG, Parra GR, Ferreiro DU. Local frustration around enzyme active sites. *Proc Nail Acad Sci U. S. A*. 2019;116(10):4037-4043.
72. Bandyopadhyay D, Bhatnagar A, Jain S, Pratyaksh P. Selective stabilization of aspartic acid protonation state within a given protein conformation occurs via specific molecular association. *J Phys Chem B*. 2020;124(26): 5350-5361. https://doi.org/10.1021/acs.jpcb.0c02629.
73. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data*. 2019;6(1). https://doi.org/10.1186/s40537-019-0192-5.
74. Luque A, Carrasco A, Martín A, De las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit*. 2019;91:216-231. https://doi.org/10.1016/j.patcog.2019.02.023.
75. Kingma DP, Ba JL. Adam: a method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings; 2015; pp. 1-15.
76. Yan K, Zhang D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sens Actuators B*. 2015; 212(March):353-363. https://doi.org/10.1016/j.snb.2015.02.025.
77. Kim HD, Park CH, Yang HC, Sim KB. Genetic algorithm based feature selection method development for pattern recognition. Paper presented at: SICE-ICASE International Joint Conference 2006; pp. 1020-1025. https://doi.org/10.1109/SICE.2006.315742.
78. Janecek A, Gansterer WNW, Demel M, Ecker G. On the relationship between feature selection and classification accuracy. Paper presented at: JMLR: Workshop and Conference Proceedings; 2008; Vol. 4, pp. 90-105.
79. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*. 1971;55(3):379-IN4. https://doi.org/10.1016/0022-2836(71)90324-X.
80. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *J Mol Biol*. 1973;79(2):351-371. https://doi.org/10.1016/0022-2836(73)90011-9.
81. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247(4):536-540. https://doi.org/10.1006/jmbi.1995.0159.

82. Brändén C-I, Tooze J. *Introduction to Protein Structure*. New York: Taylor & Francis; 1999.
83. Nowakowska-Gołacka J, Sominka H, Sowa-Rogozińska N, Słomińska-Wojewódzka M. Toxins utilize the endoplasmic reticulum-associated protein degradation pathway in their intoxication process. *Int J Mol Sci MDPI AG*. 2019;2:1307-1343. https://doi.org/10.3390/ijms20061307.
84. Botes DP. Snake venom toxins. The reactivity of the disulphide bonds of *Naja Nivea* toxin α. BBA. *Protein Struct*. 1974;359(2):242-247. https://doi.org/10.1016/0005-2795(74)90220-7.
85. Maret W, Li Y. Coordination dynamics of zinc in proteins. *Chem Rev*. 2009;109(10):4682-4707. https://doi.org/10.1021/cr800556u.
86. Shimizu T. Binding of cysteine Thiolate to the Fe(III) Heme complex is critical for the function of Heme sensor proteins. *J Inorg Biochem*. 2012;108:171-177. https://doi.org/10.1016/j.jinorgbio.2011.08.018.
87. Bushnell GW, Louie GV, Brayer GD. High-resolution three-dimensional structure of horse heart cytochrome C. *J Mol Biol*. 1990;214(2):585-595. https://doi.org/10.1016/0022-2836(90)90200-6.
88. McBride AE, Conboy AK, Brown SP, Ariyachet C, Rutledge KL. Specific sequences within arginine-glycine-rich domains affect MRNA-binding protein function. *Nucleic Acids Res*. 2009;37(13):4322-4330. https://doi.org/10.1093/nar/gkp349.
89. van Rosmalen M, Krom M, Mike Merkx M. Tuning the flexibility of glycine-serine linkers to allow RationalDesign of multidomain proteins. *Biochemistry*. 2017;56:6565-6574.
90. Lyu X, He N, Chen Z, Zhu Y, Li L. DeepCSO: a deep-learning network approach to predicting cysteine S-Sulphenylation sites. *Front Cell Dev Biol*. 2020;8:1489. https://doi.org/10.3389/fcell.2020.594587.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Nallapareddy V, Bogam S, Devarakonda H, Paliwal S, Bandyopadhyay D. DeepCys: Structure-based multiple cysteine function prediction method trained on deep neural network: Case study on domains of unknown functions belonging to COX2 domains. *Proteins*. 2021;1–17. https://doi.org/10.1002/prot.26056