Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Hist-i-fy: Multiple histidine function prediction based on protein sequences using deep neural network

Debashree Bandyopadhyay[1], Abhishek Jalan[1], Dibyansu Diptiman[1], Rishabh Pal[1], and Sachin Dodwani[1]

[1]Birla Institute of Technology & Science Pilani - Hyderabad Campus

June 21, 2023

Abstract

Histidine (His) is the most reactive amino acid at enzyme active sites. Multiple post-translational modifications (functions) are reported for His side chains. The high-throughput sequencing techniques produce a large number of protein sequences without functional annotations at the amino acid level. Experimental characterization of His functions in proteins is laborious and time-consuming. Computational characterization based on protein sequences may complement the need. There are only a handful of Histidine function prediction tools available and those annotate only a single function. Here we curated a dataset of active Histidine with known functions based on protein sequences obtained from UniProt database (sample size n=1584) and trained against four machine learning methods. The convolution neural network (CNN) model (" *Hist-i-fy*") performed the best with 75% overall accuracy. The external validation of Hist-i-fy on phosphorylated histidine data (sample size 34) showed 94.1% prediction accuracy. For the first time, we report multiple His function prediction, based on protein sequences using deep neural networks. The inputs to the model are i) protein sequence containing His, and ii) the His residue number. The model predicts one out of the eight histidine functions, namely, acetylation, ribosylation, glycosylation, hydroxylation, methylation, oxidation, phosphorylation, and protein splicing. The novelty of the work is, it predicts maximum number of histidine functions at a time with optimal performance. There is a scope of improvement in the model upon availability of a larger dataset. The model is available as a web application (https://histify.streamlit.app/) and a stand-alone code https://github.com/dibyansu24-maker/Histify).

Hist-i-fy: Multiple histidine function prediction based on protein sequences using deep neural network

Abhishek Jalan#, Dibyansu Diptiman#, Rishabh Pal# , Sachin Dodwani#, and Debashree Bandyopadhyay*

Department of Biological Sciences, Birla Institute of Technology and Science, Pilani, Hyderabad, 500078, Telangana, India

# equal contribution from all these authors, names are arranged alphabetically

***Correspondence:** Debashree Bandyopadhyay; Email:*banerjee.debi@hyderabad.bits-pilani.ac.in*

ABSTRACT

Histidine (His) is the most reactive amino acid at enzyme active sites. Multiple post-translational modifications (functions) are reported for His side chains. The high-throughput sequencing techniques produce a large number of protein sequences without functional annotations at the amino acid level. Experimental characterization of His functions in proteins is laborious and time-consuming. Computational characterization based on protein sequences may complement the need. There are only a handful of Histidine function prediction tools available and those annotate only a single function. Here we curated a dataset of active

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Histidine with known functions based on protein sequences obtained from UniProt database (sample size n=1584) and trained against four machine learning methods. The convolution neural network (CNN) model ("*Hist-i-fy* ") performed the best with 75% overall accuracy. The external validation of Hist-i-fy on phosphorylated histidine data (sample size 34) showed 94.1% prediction accuracy. For the first time, we report multiple His function prediction, based on protein sequences using deep neural networks. The inputs to the model are i) protein sequence containing His, and ii) the His residue number. The model predicts one out of the eight histidine functions, namely, acetylation, ribosylation, glycosylation, hydroxylation, methylation, oxidation, phosphorylation, and protein splicing. The novelty of the work is, it predicts maximum number of histidine functions at a time with optimal performance. There is a scope of improvement in the model upon availability of a larger dataset. The model is available as a web application (*https://histify.streamlit.app/*) and a stand-alone code*https://github.com/dibyansu24-maker/Histify*).

KEYWORDS:

Histidine; post-translational modifications; Artificial Neural Network (ANN); Convoluted Neural Network (CNN); Long Short-term Memory (LSTM); Logistic Regression; protein sequence; UNIPROT database, accuracy, recall, precision

1. Introduction: Enzyme functions are primarily executed through the catalytic residues. With the availability of the high-throughput sequence data, a large number of protein sequences are known without functional characterizations [1]. Computational characterization would facilitate rapid initial screening that can be verified further with experimental observations. Cysteine (Cys) and Histidine (His) are the two most important amino acid residues observed at the catalytic sites of all enzyme classes[2] [3]. The thiol group of cysteine amino acid side chain can undergo oxidation leading to various chemical and post-translational modifications that impact the structure and function of proteins in different capacities. A histidine imidazole is an electron-deficient heteroaromatic ring (pKa = 6.8) that makes it a suitable candidate for proton buffering, metal ion chelation, and antioxidant agents. Due to the similar values of the imidazole ring pKa and the physiological pH (= 7.4), His efficiently participates in enzyme catalysis. His residue is particularly important in acid base catalysis due to its amphoteric character. Apart from that, it participates in elimination-addition and redox reactions. Experimental characterizations are done for various His post-translational modifications those are involved in protein-protein interactions and catalysis. Extensive computational characterization of the cysteine functions has been done by our group [4], [5], [6]. However, the post-translational modification of His is less explored compared to that of Cys or Lys. The computational characterizations of His functions, so far, were reported for single modifications only. For example, histidine phosphorylation sites were predicted using a convoluted neural network (CNN) - based model, PROSPECT [7], and support vector machine-based model, pHisPred [8]. Transition metal-binding sites for Cys and His were predicted by exploiting position-specific evolutionary profiles using support vector machines and neural networks [9]. The CNN-based prediction model, PROSPECT, inputs a protein sequence and returns predicted histidine sites with 72% accuracy. The transition-metal-binding sites of histidine and cysteine were predicted from protein sequences with 73% precision. To the best of our knowledge, prediction of multiple His post-translational modifications is not reported. For the first time, we attempt to predict eight post-translational modifications of His from, i) protein sequence and ii) His residue position only, using deep neural networks. The convolution neural network (CNN) performed the best. The output of the models yields the most probable His modification. The internal evaluation accuracies are comparable to the single prediction methods, *albeit*, our results showed better performances than the existing ones. The model was blindly tested for external evaluation on independent phosphorylated Histidine data points.

2. METHODS AND MATERIALS

2.1. Histidine Chemical and Post-translational Modifications There are eight chemical and Post-Translational Modifications (PTM) (functions) annotated in this work (Figure 1). These His modifications, namely, acetylation, ribosylation, glycosylation, hydroxylation, methylation, oxidation, phosphorylation, and protein splicing, are discussed below.

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

FIGURE 1: *Post-Translational Modifications of Histidine Residues considered in this study*

2.1.1 Acetylation

Acetylation is a process of transferring an acetyl group (-CH3CO) to a molecule, mainly to N- or O- atoms, known as N-acetylation or O-acetylation. Acetylation of His is recently reported for Histidine-Tyrosine (HY) and Tyrosine-Histidine (YH) dyads treated with acetic anhydride [10]. His undergoes N-acetylation at its imidazole ring.

2.1.2 ADP-Ribosylation

Histidine ADP-ribosylation is a stress-induced rare phenomenon [11], unlike frequently observed serine-ADP-ribosylation. Mimetics of ADP-ribosylated Histidine was recently studied [12].

2.1.3 Glycosylation

Different amino acids may undergo different types of glycosylation, namely, C-linked, N-linked, O-linked, or S-linked [13]. Histidine undergoes N-linked glycosylation.**2.1.4 Hydroxylation** Protein hydroxylation, a post-translational modification, is carried out by 2-oxoglutarate-dependent dioxygenases. This post-translational modification can be induced by hypoxia-induced-factor alpha (HIF-a) on proline [14]. Hydroxylation may also involve protein-protein interactions and downstream signalling. Apart from proline, lysine, asparagine, aspartate, and histidine can also undergo hydroxylation modification [15].**2.1.5 Methylation** The actin and myosin proteins undergo the post-translational modification (PTM) of histidine methylation. There are two different locations where it can happen: 1-methyl histidine (1MeH) and 3-methyl histidine (3MeH) [16].**2.1.6 Oxidation** Under unusual or stressed conditions histidine undergoes oxidation to 2-oxo-histidine (2-oxo-His) (Figure 1). Photo-induced oxidation of Histidine leading to various cross-links, including intact His, Lys and Cys, was observed in high-molecular weight (HMW) fractions of monoclonal anti-bodies [17]. Oxidation of His residue is also observed in proteins from cells undergoing oxidative stress [18]. The 2-oxo-His changes the dissociation pattern of peptide ions in Mass-spectroscopy studies [19].**2.1.7 Phosphorylation** His phosphorylation is crucial step in various cellular processes, such as signal transduction, cell cycle, proliferation, differentiation, and apoptosis, Phosphorylated His contributes 6% to all the phosphorylated amino acids. However, phosphorylation of His is less explored compared to phosphorylated serine, threonine and tyrosine. Recently a consolidated database on phosphorylated His (HisPhosSite) is available [20]. Histidine Kinase (HK)s is one of the classical non-animal kingdom kinases that phosphorylate His, although, in a 2-step manner - i) transfer phosphate from ATP to His and ii) then transfer the phosphate to an aspartate residue [21].**2.1.8 Protein Splicing** Protein splicing is triggered via acid-base catalysis that involves multiple conserved His at the active site. Histidine probably plays dual role in protein splicing, first as a general base to start acyl shift splicing and next as a general acid to break the scissile bond at the N-terminal splicing junction [22].**2.2 Sequence signatures around different His post-translational modifications:**

Many of the His post-translational modifications were identified with specific sequence signatures or motifs. For example, His hydroxylation motif is a part of Hydrogen-bond (H-bond) cluster that is brought into the register by GXXG motif [23]. For His methylation, the common motif observed in short methylated peptides was GHXHXH [24]. Histidine acetylation motif deduced from mass spectrometry data based on diacetyl-fed rat lung proteins was GXPGXXGHXGXXG [25]. However, some of the Histidine post-translation modifications do not carry sequence signatures. For example, no specific sequence motif is reported for His glycosylation. For His phosphorylation, no clear sequence motif was identified [26].

2.3 Training dataset generation for Histidine post-translational modifications

There are eight His post-translational modifications (Figure 1) annotated in this work based on the availability of protein sequences from the UniProt database [27]. From the “Keyword” subsection of UniProt, category name ”PTM” was selected to track all possible post-translational modifications. The text filters (not case sensitive) – “His”, or “Histidine” were used to identify the experimentally annotated His functions from the PTM category, curated on November 2022. A total of sixteen modifications were identified, some

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

of those have very few data points. Finally, eight modifications were selected for the training dataset with a number of data points more than or equal to twenty (Table 1).

Modifications	**Number of Datapoints**
Methylation	303
Acetylation	172
Ribosylation	20
Glycosylation	105
Phosphorylation	532
Oxidation	101
Protein splicing	329
Hydroxylation	22

Table 1: Statistics of the His post-translational modifications obtained from UNIPROT database

2.4 Test dataset generation:

Test dataset was curated from the mass spectroscopy data [28], available from the Supplementary Table 1 of that reference. The UniProt ID was used to retrieve the sequence from the UniProt database and the corresponding His residue number. All these His residues are phosphorylated. This independent test dataset consists of 34 phosphorylated His.

2.5 Processing of the training dataset 2.5.1 Selection of input parameters for deep learning models: The training dataset was pre-processed by selecting a stretch of amino acids from each protein sequence with His of interest at the centre and that is flanked by amino acids with a variable window size, from three to ten. The length of the amino acid sequence will be 2(n)+1 for window size n. For example, amino acid sequence length will be seven for window size three. Hence, all the training sequences will have equal number of amino acids with His (of interest) at the centre, for a given window size. This set of sequences (per window size) were used as the input (X-parameter) for deep neural network models. The Y-parameters were the post-translational modifications. A representative input file to the deep neural network model is shown (Table 2). The relative performances of variable window sizes were tested.

S.No	*Modified sequence*	*Residue No.*	*Modifications*
0	PPGRRMGHAGAIIAG	299	Acetylation
1	RTIYLCRHGESEFNL	257	Acetylation
2	YKLIMLRHGEGAWNK	11	Acetylation
3	YKLIMLRHGEGAWNK	11	Acetylation
4	RSIYLCRHGESELNL	259	Acetylation
....	....	....	....
1579	VSFGSSCHGAGRKMS	882	protein-splicing
1580	VSFGSSCHGAGRKMS	881	protein-splicing
1581	DALFSTVHGAGRVMS	781	protein-splicing
1582	EALFSTVHGAGRVMS	790.0	protein-splicing
1583	AALRSTIHGAGRVMS	758.0	protein-splicing

Table 2: Representation of the input file used in the deep neural networks. The total number of datapoints (including all the modifications) is 1584 and 3 input parameters (2 X-Parameters and 1 Y-parameter) were used

2.5.2 Tokenization of the data: Character tokenization was performed to convert text (Table 2) into a list of characters using keras pre-processing library [29]. It builds a corpus of all characters and assigns a number to each character. After tokenization, 1584x15 dimension matrix was converted to a string of integers (Table

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

3). (15 corresponds to the sequence length, that is of window size 7). Finally, these integers were considered as X parameters, rather than the alphabetical characters (column 2 of Table 2). The Y-parameter was processed using Label Binarizer [30] which accepts categorical data as input and returns a NumPy array. The training dataset was randomly split into train and test dataset in a ratio of 2:1 using the Sklearn library function train_test_split. Thus, 1584 data points produced 1061 entries for training, and 523 for testing.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
0	15	15	1	9	9	16	1	2	4	1	4	7	7	4	1
1	9	5	7	17	6	20	9	2	1	10	3	10	13	14	6
2	17	12	6	7	16	6	9	2	1	10	1	4	19	14	12
3	17	12	6	7	16	6	9	2	1	10	1	4	19	14	12
4	9	3	7	17	6	20	9	2	1	10	3	10	6	14	16
....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....
1579	8	3	13	1	3	3	20	2	1	4	1	9	12	16	3
1580	8	3	13	1	3	3	20	2	1	4	1	9	12	16	3
1581	11	4	6	13	3	5	8	2	1	4	1	9	8	16	3
1582	10	4	6	13	3	5	8	2	1	4	1	9	8	16	3
1583	4	4	6	9	3	5	7	2	1	4	1	9	8	16	3

*Table 3: Tokenized 1584*15 matrix (1584 entries x 15 amino acids)***2.6 Algorithms** Four classification models (namely, Logistic regression and Deep learning algorithms - Convolutional Neural Networks (CNNs), Artificial Neural Networks (ANNs), and Long Short-Term Memory (LSTM) networks) were used to train and benchmark the dataset. These models can adjust the weights of the input data to produce a value for each class, either positive or negative, and can learn complex non-linear relationships between features and the target variable. Logistic regression is a simple and efficient algorithm that can be used for binary classification problems. It works well when the relationship between the features and the target variable is approximately linear. Deep learning algorithms like ANNs, LSTMs, and CNNs are more powerful than logistic regression for complex classification problems. They are designed to automatically learn hierarchical representations of the data, which can capture complex non-linear relationships between the features and the target variable. In general, deep learning algorithms require more data and computing resources than logistic regression but can produce results with higher accuracy on complex classification tasks.**2.6.1 Logistic regression** It is commonly used for prediction and classification problems. In the current study the model consists of eight classes (eight His modifications). As logistic regression is a binary classification method, One-Versus-Rest (OVR) logistic regression method was used where the model is trained separately for each class to determine if an observation is part of that class or not (making OVR a binary classification problem). The method assumes that every classification issue (whether involving class 0 or something else) stands on its own.**2.6.2 ANN** Artificial neural network (ANN) is a single hidden layer neural network (consists of input layer, hidden layer and output layer) that attempts to categorize each observation as one out of many discrete classes. The input to the model could be either categorical or numeric and the dependent variable (Y-parameter) should be categorical.**2.6.3 LSTM** Long-Short Term Memory (LSTM) recurrent neural networks are specialized recurrent neural networks. Recurrent neural networks (RNN) run in cycles that receive the input of network activations to the current time step from the previous time step. These activations are stored (for an amount of time, not fixed a priori) in the internal states of the network, known as long-short-term memory. Thus LSTM-RNNs can exploit a dynamically changing contextual information to transform an input sequence to an output sequence. LSTM has multiple hidden layers, in contrast to a single hidden layer in ANN.**2.6.4 CNN** A convolutional neural network consists of an input layer, multiple hidden layers and an output layer. At least one or more hidden layers should perform convolutions between the convolution kernel and the layer of input matrix. (Convolution is a mathematical operation, that is a dot product of two functions.) The convolution operation generates a feature map, as the convolution kernel slides along the input matrix for the layer and prepare the input of the next layer. This layer is followed by pooling layers, fully connected layers, and normalization layers those enable spatial hierarchical feature learning by backpropagation, in an

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

automated manner.***FIGURE 2*** *Workflow of CNN Model. The number of input and output features, and the number of batch normalization are shown in parenthesis. A "None" value in the shape represents any size (large than or equal to 1) in that dimension.*The workflow of the CNN model on the training dataset is shown with the number of input and output features and the batch normalization (Figure 2). The first layer, input layer, consists of neurons that are connected to individual inputs those are passed to the next layer (embedding layer), without having any weights (or biases). Embedding layer transforms each word into a vector of a pre-determined length. (The vocabulary is first encoded as a series of integers, and then the embedding layer retrieves the embedding vector for each word-index.) Convolutional layer is the core component of a CNN. It has a collection of filters (128 in this case, Figure 2) whose settings will be learnt as part of the training process. In most of the cases, the length of the sequence is less than the size of the filters. Each filter is used to generate an activation map by convolving with the input. Average pooling is used to construct a down-sampled (pooled) feature map by computing the average for each patch of the feature map. Normalization in a Neural Network is called "Batch Norm," and it takes place between the layers of the network rather than with the raw data itself. Mini-batches are processed instead of the entire dataset at once. Learning is simplified as the pace of the instruction is increased. The Global Average Pooling-1D layer is used to reduce the dimensionality of the input data after the convolutional layers. Dense Layer is a layer of neurons within which each neuron gets input from all the neurons in the preceding layer, thus becomes "dense" and fully connected. Output layer is typically the last layer in the network, and it consists of neurons that represent the different classes or categories that the network is trained to predict. Here, two dense layers were added, each with RELU activation function and batch-normalization. The first dense layer has 256 units, and the second dense layer has 128 units respectively. The final Dense layer has 8 units, representing the number of classes in the classification task. The activation function used here is softmax, which converts the output into probabilities for each class. The output layer neurons are connected to all the neurons in the previous layer, and the weights and biases of the neurons in the output layer determine the strength of the predictions for each class.

7

2.7 Evaluation meters To evaluate the efficacy of the models, the confusion matrix and the classification report were used. The confusion matrix is a compact representation of the results of a classification task prediction. It is used to evaluate the overall performance in terms of accuracy, precision, recall, and F1-score. A combination of correct (True) and incorrect (False) classifications is used to determine these measures. In this context, the projected classes may be generally referred to as "Positive" or "Negative." Total four parameters are used, True Positive (TP) (case is positive, prediction is also positive), True Negative (TN) (case is negative and prediction is also negative), False Positive (FP) (case is negative and prediction is positive) and False Negative (FN) (case is positive and false negative). Using these four parameters evaluation meters, namely, accuracy, precision, recall and F1-score, were computed.

Accuracy = (TP + TN)/ (TP + TN + FP + FN) Eq 1

Accuracy is the proportion of the true positive and false negative instances predicted correctly by the classifier with respect to all the data points.

Precision = TP/ (TP + FP) Eq 2

Precision is the proportion of the true positive instances predicted correctly by the classifier, with respect to all the data points.**Recall = TP/(TP+FN) Eq 3**Recall is the proportion of the true positive instances predicted correctly out of the true positive and false negative data points. In other words, the ability of a classifier to retrieve all appropriate examples is called recall (also known as sensitivity).**F1 Score = 2*(Recall * Precision) / (Recall + Precision) Eq 4**F1 scores include both precision and recall into their calculation, they are often regarded as inferior to accuracy metrics. It ranges from 0.0 to 1.0. To compare classifier models, it is recommended to utilise the weighted average of F1 rather than overall accuracy.**Weighted average F1-score** $=\sum_{i=1}^{8}$ **F1−score** $*$ (**no. of instances**))/**8Eq 53 RESULTS AND DISCUSSION3.1 Selection of the optimal window size based on the neural network model**

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

performances: To address the objective of the study - prediction of multiple His modifications based on a given protein sequence - we initially attempted to optimize the target amino acid sequence length (variable window size from three to ten, as described in the method section). The results were shown for CNN model. The accuracy of CNN models follows a Gaussian distribution between window sizes three to ten (Table 4). The maximum accuracy on the training dataset was observed for window size seven (that is sequence length of 15 [2x7=+1] amino acids). Hence, window size seven was as selected default for subsequent model creation, training and validation.

Window size	*Respective Accuracy of CNN Model*
3	73.60%
4	73.80%
5	72.23%
6	75.14%
7	75.47%
8	69.20%
9	70.70%
10	71.70%

*Table 4: Performance of the CNN model on the training dataset with variable window size***3.2 Selection of the optimal neural network model on the training dataset:**The training dataset was benchmarked against four different classifiers, namely, Logistic regression, ANN, LSTM and CNN. Some of these classifiers are simple and computational less expensive and two others (LSTM and CNN) are complex.**3.2.1 Logistic regression** The overall performance of logistic regression is shown (Table S1). The precision, recall and F1-score vary for different modifications. Logistic regression was unable to predict methylation and phosphorylation based on the current validation dataset. The possible reasons could be that i) the target label has no linear correlation with the features and/or ii) the sample sizes (in the validation dataset) are uneven with respect to different classes (Table S1). The accuracy from this classifier on the validation dataset was 0.67.**3.2.2 ANN** The accuracy achieved using ANN model was 70%, slightly better than that of the logistic regression. This is presumably due to the presence of three layers in ANN (an input layer, a hidden layer and an output layer) in contrast to the logistic regression. Moreover, performance of logistic regression reduces when trained on noisy data or the samples are unevenly distributed between classes. Variation was observed in the prediction results for different modifications (Table S2). ANN model was unable to predict oxidation modification from the current dataset, although, it has successfully predicted methylation and phosphorylation modifications, unlike logistic regression. To note, the modifications in validation dataset varies across the classifier, as the train to test (validation) dataset was randomly split into 2:1 ratio and each random split contains different ratio of His modifications. For example, in logistic regression, support (the number of validation data point) value for phosphorylation was only 8 in contrast to 175 in ANN model. This could presumably justify why logistic regression was unable to predict phosphorylation whereas ANN has accomplished it successfully.**3.2.3 LSTM** The accuracy obtained from LSTM was 71%, better than those obtained from logistic regression and ANN models. The results produced by LSTM were better, most likely due to feed data back while training. LSTM works the best on a known set of patterns or sequences. As mentioned above, His hydroxylation and methylation were observed with characteristic sequence motifs [23] [24]. Moreover, protein splicing involves multiple conserved His at the enzyme active sites [22]. The conserved patterns for these three His modifications lead to improved recall value (that is, high rate of true positive prediction with respect to false negative values) (Table S3). Despite of improved performance of LSTM, the classifier was unable to compute precision, recall and F1-score for ribosylation and oxidation modifications.**3.2.4 CNN** The overall accuracy obtained from CNN model was the 75.47%, best out of all the classifiers. The notable observation was that the CNN model was capable of predicting all the modifications, unlike other classifiers (Table 5). The variation in predicting different modifications also exists in CNN model as in other classifiers. The superior performance of the CNN model is most likely due to the application of convolutional layers, those automatically lowers the dimensionality of sequences, yet preserv-

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

ing the information. The logistic regression works the best with a predefined relation between input and output, that was not so explicit in the training dataset. As ANN is a simple neural network model with only one hidden layer, learning was less accurate. LSTM works the best with pattern recognition thus the model was capable of better prediction of hydroxylation, methylation and protein splicing with known patterns. However, the performance of acetylation prediction was consistently poor across all the classifiers, although sequence-specificity of His acetylation was reported in literature [10][25]. Comparing the overall performance from the above benchmarking exercise (Table S4), CNN model was selected for further His modification prediction for an unknown protein. This model, termed as Hist-i-fy, was tested on an independent dataset of His phosphorylation, obtained from mass spectroscopy.

PTM	precision	Recall	f1-score	Support#
Acetylation	0.32	0.20	0.24	46
Ribosylation	0.31	0.62	0.42	8
Glycosylation	0.73	0.77	0.75	35
hydroxylation	0.67	1.00	0.80	4
methylation	0.78	0.54	0.64	103
oxidation	0.53	1.00	0.69	45
phosphorylation	0.80	0.79	0.80	175
protein-splicing	1.00	1.00	1.00	107
macro avg	0.64	0.74	0.67	523
weighted avg	0.76	0.75	0.74	523

Table 5: Classification report for CNN model; #number of instances from validation dataset

3.3 Comparison of the results from the present study and the literature reports:

Currently, there are only a few His post-translational modification prediction tools reported in the literature, namely, pHisPred, iPhosH-PseAAC, Prospect and His-Cys metal binding prediction. All of these tools can predict one His function at a time. His-Cys metal binding prediction tool can predict metal-binding of His and Cys amino acids. The training data sets used to develop these prediction tools were small enough and the sizes were comparable to the dataset used in this study (Table 6). The internal prediction accuracies for iPhos-PseAAC, Prospect and pHisPred were 33%, 72% and 73% respectively. His-Cys metal binding sites (predicting two amino acids at a time) have reported 73% precision and 61% recall values. The best internal prediction accuracy was obtained from the current model, Hist-i-fy. However, there is a scope of improvement for the model performance upon availability of larger data sets. For external validation, we have tested the Hist-i-fy model on an independent dataset of histidine phosphorylation, generated from mass spectroscopy, sample size, 34. The prediction accuracy of the Hist-i-fy model on the test dataset was 94.1% only. To note, the training and the test datasets are independent of each other and the test dataset consists of only one modification, phosphorylation. Moreover, the training accuracy was a cumulative accuracy for all the modifications and the test accuracy was only for phosphorylation. Thus, the accuracy observed in the test dataset was higher than that in the training dataset. For comparison purpose, the same test dataset was used for histidine phosphorylation prediction using pHisPred tool. The prediction accuracy from pHisPred was 94.0, comparable to the results from Hist-i-fy. For the first-time we report prediction of eight histidine modifications from a given protein sequence, with a reasonably high accuracy.

Single Histidine Prediction	**Single Histidine Prediction**	**Single Histidine Prediction**
Name of the software	**No. of sequences used to train the model**	**Performance of the algorithms**
iPhosH-PseAAC	795	Not working
PROSPECT	172	Not working
pHispred	560	94.0%
Multiple amino acid Prediction	**Multiple amino acid Prediction**	**Multiple amino acid Predictio**
His-Cys metal binding prediction	2982	Not working

Multiple Histidine Prediction	**Multiple Histidine Prediction**	**Multiple Histidine Prediction**
Hist-i-fy	1584	94.1%

Table 6: Histidine function prediction models

3.4 Availability of Hist-i-fy:

The source code for Hist-i-fy is available at*https://github.com/dibyansu24-maker/Histify*and can be easily operated on any platform, following the instructions on GitHub. Moreover, Hist-i-fy enables users to train the model with the user-defined data.

Hist-i-fy prediction server is hosted on*https://histify.streamlit.app/*using Streamlit, a powerful Python library for building interactive web applications. Through the intuitive and user-friendly interface of Streamlit, users can predict sequence modifications in two modes: single sequence or multiple sequences (to be uploaded as a CSV file containing sequences and Histidine residue numbers). Streamlit seamlessly integrates with GitHub, facilitating the hosting process and ensuring a smooth user experience. The Hist-i-fy model processes the input data and generates predictions for the respective sequences, providing users with comprehensive insights into their sequence modifications.

4. CONCLUSION: The functional characterization of the proteins and their amino acids lag behind the protein sequence determination. Experimental characterization is time-consuming and laborious, those can be complemented by computational methods. Histidine being one of the most important amino acid at the enzyme active site, functional characterization would potentially address many biological problems. His undergoes multiple modifications, sixteen such was reported in the UNIPROT database. However, only a handful of His (single) function prediction tools are available. Objective of this study was to predict multiple histidine modifications (functions) based on protein sequences. Here we trained and validated eight histidine modifications using Convoluted Neural Network model. The training dataset was curated from UNIPORT database. The overall accuracy produced by the CNN model was 75%, although, prediction of individual modifications varies, depending on the data size, existing sequence pattern etc. The external validity of the model was tested on an independent phosphorylation dataset obtained from proteomics study. Accuracy of phosphorylation prediction (external validation) was 94.1% much higher than that of accuracy (including all the eight modifications) from internal validation. The external validity result was comparable from existing His phosphorylation prediction tool - pHisPred. The final CNN model (termed as Hist-i-fy) is publicly available as a web application and a stand-alone program.**5. REFERENCE:**

[1] M. Cui, C. Cheng, and L. Zhang, "High-throughput proteomics: a methodological mini-review," *Lab. Investig. J. Tech. Methods Pathol.*, vol. 102, no. 11, pp. 1170–1181, Nov. 2022, doi: 10.1038/s41374-022-00830-7.

[2] T. K. Harris and G. J. Turner, "Structural Basis of Perturbed pKa Values of Catalytic Groups in Enzyme Active Sites," *IUBMB Life*, vol. 53, no. 2, pp. 85–98, 2002, doi: 10.1080/15216540211468.

[3] A. Gutteridge and J. M. Thornton, "Understanding nature's catalytic toolkit," *Trends Biochem. Sci.*, vol. 30, no. 11, pp. 622–629, Nov. 2005, doi: 10.1016/j.tibs.2005.09.006.

[4] A. Bhatnagar and D. Bandyopadhyay, "Characterization of cysteine thiol modifications based on protein microenvironments and local secondary structures,"*Proteins*, vol. 86, no. 2, pp. 192–209, Feb. 2018, doi: 10.1002/prot.25424.

[5] V. Nallapareddy, S. Bogam, H. Devarakonda, S. Paliwal, and D. Bandyopadhyay, "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*, vol. 89, no. 7, pp. 745–761, Jul. 2021, doi: 10.1002/prot.26056.

[6] A. Bhatnagar, M. I. Apostol, and D. Bandyopadhyay, "Amino acid function relates to its embedded protein microenvironment: A study on disulfide-bridged cystine," *Proteins Struct. Funct. Bioinforma.*, vol.

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

84, no. 11, pp. 1576–1589, 2016, doi: 10.1002/prot.25101.

[7] Z. Chen*et al.*, "PROSPECT: A web server for predicting protein histidine phosphorylation sites," *J. Bioinform. Comput. Biol.*, vol. 18, Mar. 2020, doi: 10.1142/S0219720020500183.

[8] J. Zhao*et al.*, "pHisPred: a tool for the identification of histidine phosphorylation sites by integrating amino acid patterns and properties," *BMC Bioinformatics*, vol. 23, Sep. 2022, doi: 10.1186/s12859-022-04938-x.

[9] A. Passerini, M. Punta, A. Ceroni, B. Rost, and P. Frasconi, "Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks," *Proteins Struct. Funct. Bioinforma.*, vol. 65, no. 2, pp. 305–316, 2006, doi: 10.1002/prot.21135.

[10] S. Lc and M. M, "Using Peptide Arrays To Discover the Sequence-Specific Acetylation of the Histidine-Tyrosine Dyad," *Biochemistry*, vol. 58, no. 13, Apr. 2019, doi: 10.1021/acs.biochem.9b00022.

[11] S. Larsen*et al.*, "Mapping Physiological ADP-Ribosylation Using Activated Ion Electron Transfer Dissociation," *Cell Rep.*, vol. 32, p. 108176, Sep. 2020, doi: 10.1016/j.celrep.2020.108176.

[12] H. Minnee*et al.*, "Mimetics of ADP-ribosylated histidine through copper(I)-catalyzed click chemistry," *Org. Lett.*, vol. 24, no. 21, pp. 3776–3780, May 2022, doi: 10.1021/acs.orglett.2c01300.

[13] D. Dutta, C. Mandal, and C. Mandal, "Unusual glycosylation of proteins: Beyond the universal sequon and other amino acids," *Biochim. Biophys. Acta BBA - Gen. Subj.*, vol. 1861, no. 12, pp. 3096–3108, Dec. 2017, doi: 10.1016/j.bbagen.2017.08.025.

[14] G. Zurlo, J. Guo, M. Takada, W. Wei, and Q. Zhang, "New Insights into Protein Hydroxylation and Its Important Role in Human Diseases," *Biochim. Biophys. Acta*, vol. 1866, no. 2, pp. 208–220, Dec. 2016, doi: 10.1016/j.bbcan.2016.09.004.

[15] S. Markolovic, S. E. Wilkins, and C. J. Schofield, "Protein Hydroxylation Catalyzed by 2-Oxoglutarate-dependent Oxygenases," *J. Biol. Chem.*, vol. 290, no. 34, pp. 20712–20722, Aug. 2015, doi: 10.1074/jbc.R115.662627.

[16] M. E. Jakobsson, "Enzymology and significance of protein histidine methylation," *J. Biol. Chem.*, vol. 297, no. 4, Oct. 2021, doi: 10.1016/j.jbc.2021.101130.

[17] C.-F. Xu*et al.*, "Discovery and Characterization of Histidine Oxidation Initiated Cross-links in an IgG1 Monoclonal Antibody," *Anal. Chem.*, vol. 89, no. 15, pp. 7915–7923, Aug. 2017, doi: 10.1021/acs.analchem.7b00860.

[18] C. Schöneich, "Reactive oxygen species and biological aging: a mechanistic approach," *Exp. Gerontol.*, vol. 34, no. 1, pp. 19–34, Jan. 1999, doi: 10.1016/S0531-5565(98)00066-7.

[19] J. D. Bridgewater, R. Srikanth, J. Lim, and R. W. Vachet, "The Effect of Histidine Oxidation on the Dissociation Patterns of Peptide Ions," *J. Am. Soc. Mass Spectrom.*, vol. 18, no. 3, pp. 553–562, Mar. 2007, doi: 10.1016/j.jasms.2006.11.001.

[20] J. Zhao*et al.*, "HisPhosSite: A comprehensive database of histidine phosphorylated proteins and sites," *J. Proteomics*, vol. 243, p. 104262, Jul. 2021, doi: 10.1016/j.jprot.2021.104262.

[21] P. M. Wolanin, P. A. Thomason, and J. B. Stock, "Histidine protein kinases: key signal transducers outside the animal kingdom," *Genome Biol.*, vol. 3, no. 10, p. reviews3013.1-reviews3013.8, 2002, doi: 10.1186/gb-2002-3-10-reviews3013.

[22] Z. Du*et al.*, "Highly Conserved Histidine Plays a Dual Catalytic Role in Protein Splicing: A pKa Shift Mechanism," *J. Am. Chem. Soc.*, vol. 131, no. 32, pp. 11581–11589, Aug. 2009, doi: 10.1021/ja904318w.

Posted on 21 Jun 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.168732552.21824171/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

[23] J. R. Herrmann, J. C. Panitz, S. Unterreitmeier, A. Fuchs, D. Frishman, and D. Langosch, "Complex Patterns of Histidine, Hydroxylated Amino Acids and the GxxxG Motif Mediate High-affinity Transmembrane Domain Interactions," *J. Mol. Biol.*, vol. 385, no. 3, pp. 912–923, Jan. 2009, doi: 10.1016/j.jmb.2008.10.058.

[24] M. Lv*et al.*, "METTL9 mediated N1-histidine methylation of zinc transporters is required for tumor growth," *Protein Cell*, vol. 12, no. 12, pp. 965–970, Dec. 2021, doi: 10.1007/s13238-021-00857-4.

[25] L. D. L. Jedlicka *et al.*, "Increased chemical acetylation of peptides and proteins in rats after daily ingestion of diacetyl analyzed by Nano-LC-MS/MS," *PeerJ*, vol. 6, p. e4688, Apr. 2018, doi: 10.7717/peerj.4688.

[26] K. Terashima*et al.*, "Impurity effects on electron–mode coupling in high-temperature superconductors," *Nat. Phys.*, vol. 2, no. 1, Art. no. 1, Jan. 2006, doi: 10.1038/nphys200.

[27] The UniProt Consortium, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/nar/gkaa1100.

[28] C. M. Potel, M.-H. Lin, A. J. R. Heck, and S. Lemeer, "Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics," *Nat. Methods*, vol. 15, no. 3, pp. 187–190, Mar. 2018, doi: 10.1038/nmeth.4580.

[29] "Tokenization and Text Data Preparation with TensorFlow & Keras," *KDnuggets*. https://www.kdnuggets.com/tokenization-and-text-data-preparation-with-tensorflow-keras.html (accessed Apr. 14, 2023).

[30] "sklearn.preprocessing.LabelBinarizer," *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.preprocessing (accessed Apr. 14, 2023).