SCIENTIFIC REPORTS
nature research

OPEN

Characterization and Identification of Lysine Succinylation Sites based on Deep Learning Method

Kai-Yao Huang[1], Justin Bo-Kai Hsu[2] & Tzong-Yi Lee[3,4]*

Succinylation is a type of protein post-translational modification (PTM), which can play important roles in a variety of cellular processes. Due to an increasing number of site-specific succinylated peptides obtained from high-throughput mass spectrometry (MS), various tools have been developed for computationally identifying succinylated sites on proteins. However, most of these tools predict succinylation sites based on traditional machine learning methods. Hence, this work aimed to carry out the succinylation site prediction based on a deep learning model. The abundance of MS-verified succinylated peptides enabled the investigation of substrate site specificity of succinylation sites through sequence-based attributes, such as position-specific amino acid composition, the composition of *k*-spaced amino acid pairs (CKSAAP), and position-specific scoring matrix (PSSM). Additionally, the maximal dependence decomposition (MDD) was adopted to detect the substrate signatures of lysine succinylation sites by dividing all succinylated sequences into several groups with conserved substrate motifs. According to the results of ten-fold cross-validation, the deep learning model trained using PSSM and informative CKSAAP attributes can reach the best predictive performance and also perform better than traditional machine-learning methods. Moreover, an independent testing dataset that truly did not exist in the training dataset was used to compare the proposed method with six existing prediction tools. The testing dataset comprised of 218 positive and 2621 negative instances, and the proposed model could yield a promising performance with 84.40% sensitivity, 86.99% specificity, 86.79% accuracy, and an MCC value of 0.489. Finally, the proposed method has been implemented as a web-based prediction tool (CNN-SuccSite), which is now freely accessible at http://csb.cse.yzu.edu.tw/CNN-SuccSite/.

Post-translational modifications (PTMs), which are biochemical reactions occurring on proteins, have been known to have crucial roles in cellular processes such as DNA repair, transcriptional regulation, signaling pathways, protein–protein interactions, apoptosis, cell death, and metabolic pathways[1]. Protein succinylation is a type of PTM involving the attachment of a succinyl group (-CO-CH_2-CH_2-CO-) to a specific lysine residue of a protein[2]. Protein lysine succinylation, mediated by succinyl-coenzyme A (succinyl-CoA), has been identified to play crucial roles in regulating a variety of cellular processes[3,4]. In recent years, high-throughput mass spectrometry (MS) has been widely adopted to identify large-scale datasets of site-specific succinylation peptides[5–8]. Proteome-wide profiling analyses have revealed the involvement of succinylation in multiple metabolic pathways[8] and cellular physiology[9], especially for thermophilic and mesophilic bacteria[7]. In addition, a quantitative succinylome analysis in breast cancer expanded our understanding of mechanisms of tumorigenesis and provided further characterization of the pathophysiological roles of succinylation in breast cancer progression, which can enable innovative therapies for breast cancer patients[10]. However, the functions of protein succinylation in diseases and cancer are still not well studied. The limited number of studies involving functional investigations of protein succinylation has motivated us to provide a functional enrichment analysis for all succinylated proteins.

Due to the quantitative succinylome data obtained from MS-based proteomics techniques, a variety of bioinformatics tools have been developed for predicting lysine succinylation sites based on protein sequences. A list of previously proposed approaches concerning computational annotation of succinylated sites is given in

[1]Department of Medical Research, Hsinchu Mackay Memorial Hospital, Hsinchu city, 300, Taiwan. [2]Department of Medical Research, Taipei Medical University Hospital, Taipei city, 110, Taiwan. [3]Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, 518172, China. [4]School of Life and Health Sciences, The Chinese University of Hong Kong, Shenzhen, 518172, China. *email: leetzongyi@cuhk.edu.cn

Supplementary Table S1. SucPred[11] is a succinylation site prediction tool designed by Zhao *et al*. based on support vector machine (SVM) with the consideration of multiple feature-encoding algorithms. SuccFind[12] was developed based on sequence-derived features and evolutionary-derived sequence information with an enhanced feature optimization strategy. The iSuc-PseAAC was proposed by incorporating the peptide position-specific propensity into the general form of pseudo amino acid composition[13]. Hasan *et al*. have proposed a web server, named SuccinSite, that incorporates three sequence encodings — k-spaced amino acid pairs and binary and amino acid index properties — for predicting succinylated lysine sites[14]. A computational tool termed iSuc-PseOpt has been developed to predict protein succinylation sites by incorporating the sequence-coupling effects into the general pseudo amino acid composition and using K-nearest neighbors cleaning (KNNC) treatment and inserting hypothetical training samples (IHTS) treatment to optimize the training dataset[15]. In 2017, Hasan *et al*. further proposed the SccinSite 2.0 for a systematic identification of species-specific protein succinylation sites by using joint element features information[16]. Recently, a new method called Success was developed by integrating evolutionary and structural characteristics to provide accurate predictions of protein succinylation sites[17]. Lopez *et al*. also published another succinylation site prediction tool namely SSEvol-Suc[18] in 2018. In October 2018, Hasan *et al*. published the GPSuc for a global prediction of generic and species-specific succinylation sites by aggregating multiple sequence features[19]. In 2019, Hasan *et al*. further proposed a large-scale assessment of prediction tools for lysine succinylation sites[20].

Although many succinylation site prediction tools have been proposed, the performance of those approaches can still be improved. Moreover, the recent advancements of high-throughput techniques in biotechnology have identified more and more experimentally verified data of succinylation sites. The lack of deep learning-based approaches for identifying succinylation sites needed to be addressed. Therefore, we aimed to develop a new method for identifying protein succinylation sites based on a deep neural network[21]. In this work, four sequenced-based attributes, such as position-specific amino acid composition[22,23], amino acid pairs composition[24–26], position-specific scoring matrix (PSSM)[27], and *k*-spaced amino acid pairs[28,29], were considered for identifying protein succinylation sites. According to cross-validation evaluation, the model with the best cross-validation performance was further measured with an independent testing dataset.

Material and Methods

A flow chart of the proposed method is presented in Fig. 1, including (1) construction of the training dataset, (2) feature investigation and encoding, (3) model construction and performance evaluation, and (4) independent testing. First, the experimental data of known succinylated sites was mainly obtained from PLMD 3.0[30]. After constructing positive and negative training datasets, four different types of sequence-based encoding schemes were adopted to transform the sequences to multi-dimensional vectors. Then, ten-fold cross-validation was utilized to evaluate the performances of the predictive models trained based on deep convolutional neural networks. Finally, the model with the best predictive performance was further evaluated by an independent testing dataset, which was truly blind to the training dataset used for model construction. The detailed procedures are described in the following sections.

Construction of positive and negative training datasets. In this study, the dataset of experimentally verified lysine succinylation sites was mainly extracted from PLMD[30], which has accumulated 284,780 lysine modification sites from 53,501 proteins among 20 different types of PTMs. When considering only experimentally confirmed lysine succinylation sites, a total of 18,593 sites were obtained from 6,377 unique proteins. After the removal of homologous protein sequences, a $(2n+1)$-mer window size was adopted to extract fragmented sequences centered on modified sites with n left-hand and n right-hand neighboring amino acids. Given a specific number of succinylated proteins, the negative dataset was generated from non-succinylated sites which are those fragmented sequences centered on lysine residues that lack succinylation annotation. By evaluating different values of window size ($2n+1$, n is ranging from 5 to 20), the 31-mer window size ($n=15$) performed best for predicting lysine succinylation sites (Supplementary Fig. S1), based on the basic feature—position-specific amino acid composition. After filtering out the fragmented sequences with sequence lengths less than 31 amino acids, a total of 17,637 and 137,974 fragmented sequences were retained for positive and negative training datasets, respectively. A training dataset with high sequence similarity will overestimate cross-validation performance[26,31–33]. Therefore, after the removal of the duplicated and homologous sequences by employing the CD-HIT program[34] with 30% sequence identity, we obtained a total of 1,268 non-homologous succinylated proteins, which comprise 3,216 succinylated and 16,412 non-succinylated lysine residues for positive and negative training datasets, respectively. Table 1 provides the detailed statistics of positive and negative instances in accordance with various sequence identity thresholds of CD-HIT.

Feature investigation and encoding. This study aimed at the sequence-based characterization of protein succinylation site specificity. Not only the position-specific amino acid composition (PspAAC) but also the composition of *k*-spaced amino acid pairs (CKSAAPs) and position-specific scoring matrix (PSSM) were considered for use as the training attributes for constructing predictive models as well as measuring discriminating powers.

Position-specific Amino acid composition (PspAAC). Amino acid composition (AAC) has been regarded as a typical attribute for examining substrate site motifs on a variety of PTMs[35–41]. AAC was defined to determine the probability of amino acids occurring in the flanking region of PTM sites. Since a training sequence x has a length of 31 amino acids, the probability $P_x(k)$ of a specific amino acid k was elaborated as[42]

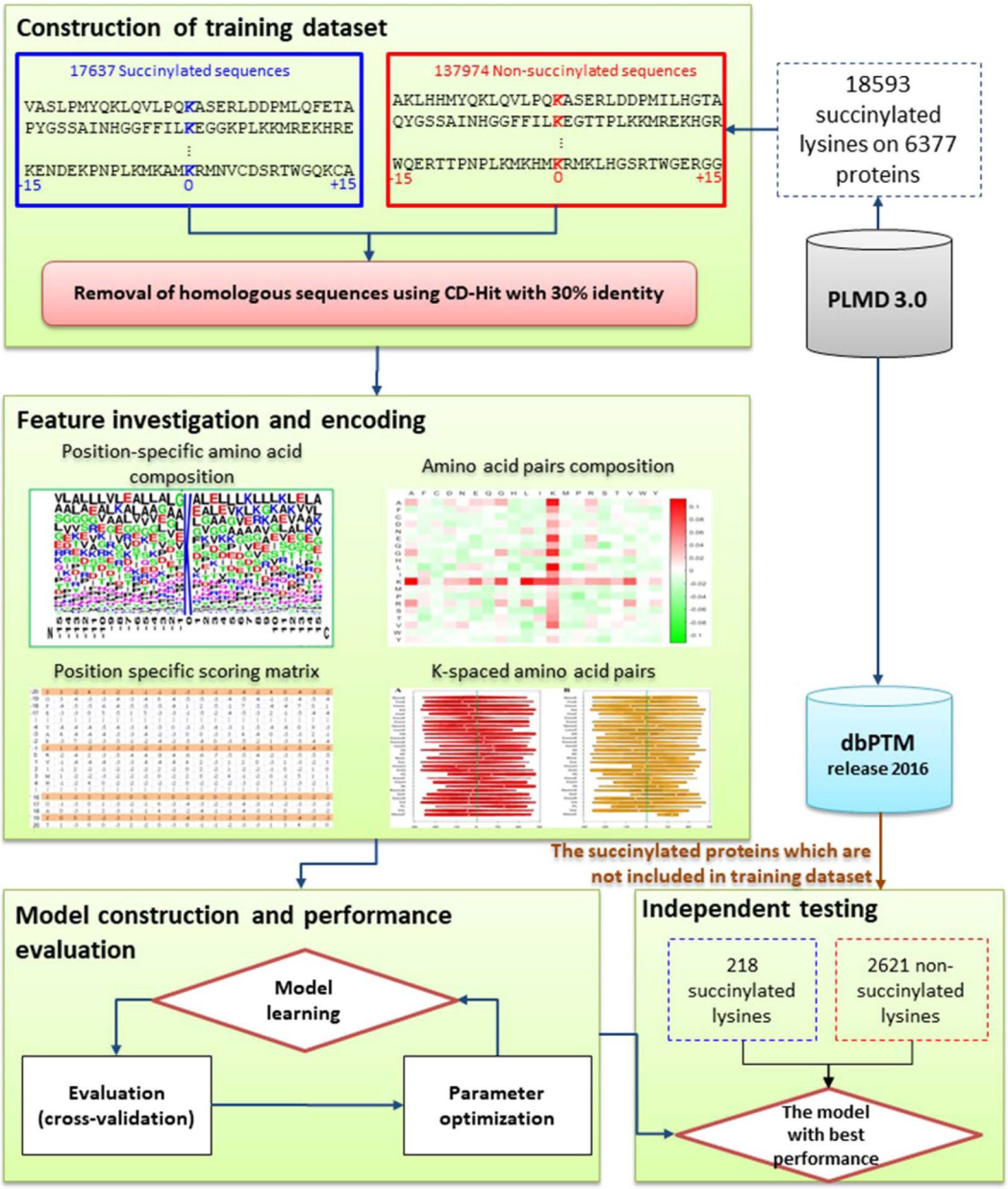


Figure 1. Flow chart of the proposed method. Four major steps were involved such as construction of training dataset, feature investigation and encoding, model construction and performance evaluation, and independent testing.

$$P_x(k) = \frac{n_x(k)}{\sum_{k=1}^{20} n_x(k)} \quad k = 1, 2, \ldots, 20 \tag{1}$$

where $n_x(k)$ represents the number of occurrences of a specific amino acid k. Refer to the method of positional weighted matrix (PWM) of amino acids around sulfation sites[43], the position-specific amino acid composition (PspAAC) around the succinylated sites was determined using non-homologous training datasets. The PspAAC specified the relative frequency of twenty amino acids of each position that surrounded the succinylation sites, and was utilized in encoding the fragment sequences. A matrix of $m \times w$ elements was used to represent the PspAAC of a training dataset, where m stands for 20 types of amino acids and w is the window size ranging from -15 to $+15$. The matrix with 20×30 features was represented as:

$$PspAAC = \begin{bmatrix} P_{-15}(1) & \cdots & P_{+15}(1) \\ \vdots & \ddots & \vdots \\ P_{-15}(m) & \cdots & P_{+15}(m) \end{bmatrix} \tag{2}$$

Sequence identity threshold	Number of succinylated proteins	Number of succinylated lysine sites	Number of non-succinylated lysine sites
Full data	6,034	17,637	137,974
100%	5,539	15,691	117,813
90%	4,924	13,656	97,629
80%	4,422	12,031	86,192
70%	3,812	10,908	69,656
60%	3,105	8,855	51,456
50%	2,517	6,201	33,904
40%	1,869	4,509	23,577
30% (**Training data**)	1,268	3,216	16,412

Table 1. Data statistics of positive and negative training datasets using CD-HIT with various values of sequence identity threshold.

Composition of k-spaced amino acid pairs (CKSAAP). The composition of k-spaced amino acid pairs (CKSAAP) has been extensively applied in analyses of protein functions[28,33,41,44–49]. This study transformed all training sequences into numeric vectors based on the encoding method of CKSAAP. Given k values ranging from zero to five, the number of occurrence of each k-spaced AAP can be determined from target sequences. If k is set as one, $\left[A_i x A_j\right]$ was used to represent the pair of amino acids A_i and A_j (i and $j = 1, \ldots, 20$, corresponding to 20 amino acids) which are separated by one residue of any amino acid x. If k is set as two, $[A_i x x A_j]$ represented the pair of amino acids A_i and A_j that are separated by two amino acids xx. The occurring count of a one-spaced AAP $[A_i x A_j]$ was represented by $N([A_i x A_j])$ and its conditional probability $P[A_i x A_j]$ was defined as:

$$P[A_i x A_j] = \frac{N([A_i x A_j])}{N([A_i x A_*])} \tag{3}$$

where $N([A_i x A_*]) = \sum_{j=1,\ldots,20} N([A_i x A_j])$. In order to identify the difference of occurring frequency of a KSAAP between positive and negative sequences, for instance, the diversity of a one-spaced AAP $[A_i x A_j]$ can be obtained from:

$$C\left[A_i x A_j\right] = \log \frac{P^+\left[A_i x A_j\right]}{P^-\left[A_i x A_j\right]} \tag{4}$$

where $P^+\left[A_i x A_j\right]$ and $P^-\left[A_i x A_j\right]$ are the conditional probabilities of a one-spaced AAP $\left[A_i x A_j\right]$ in positive and negative training sequences, respectively. In this investigation, a higher positive value of $C\left[A_i x A_j\right]$ indicated that the one-spaced AAP $\left[A_i x A_j\right]$ is a more significant attribute in the positive dataset; otherwise, a smaller negative value of $C\left[A_i x A_j\right]$ revealed it is a more abundant attribute in negative dataset. Among a total of 2400 KSAAPs, we utilized a feature selection approach, minimum redundancy–maximum relevance (mRMR), to generate an index score for each KSPAAP[50]. A KSAAP with minimum redundancy and maximum relevance was regarded as the best attribute for classifying succinylated and non-succinylated sequences. The scoring function of mRMR was described as:

$$score_j = M\left(f_j, c\right) - \frac{1}{m}\sum_{i=1}^{m} M(f_i, f_j), \tag{5}$$

where in $f_j \subset S_n$, $f_i \subset S_m$, $S_m = S - S_n$, in which S_m, S_n, and S were the attribute sets (m and n were the attribute sizes), and c is a classification variable with two possible classes. Additionally, the mutual information $M(x, y)$ was defined as:

$$M(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy, \tag{6}$$

where $p(x, y)$, $p(x)$, and $p(y)$ were regarded as the probabilistic density functions between attributes x and y. In addition, the sequential forward selection (SFS) was employed to a final set of 400 most discriminating KSAAPs according to the ranking of mRMR index scores.

Position specific scoring matrix (PSSM). From a structural viewpoint, several amino acid residues can be mutated without changing a protein's tertiary structure, and two proteins may have similar structures with different amino acid compositions[51]. PSSM profiles, which have been extensively utilized in protein secondary structure prediction, subcellular localization, and other bioinformatics analyses[51–54], were adopted herein with significant improvement. As presented in Fig. 2, the PSSM profile of each training sequence was generated by performing PSI-BLAST[55] against the database of non-homologous succinylated peptides. The PSSM profile was composed of

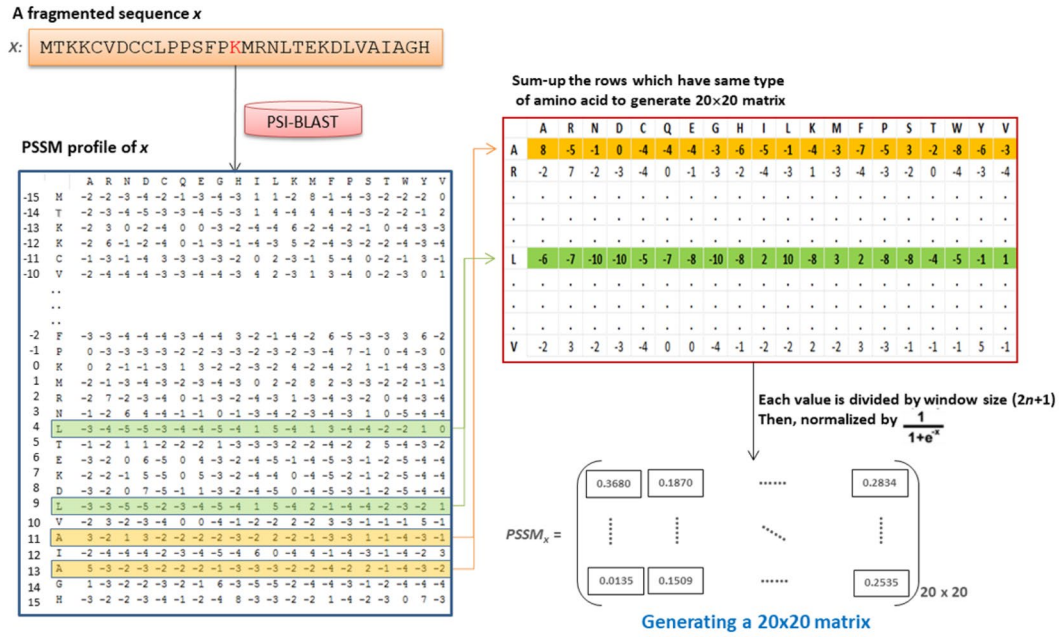


Figure 2. Flow chart of generating a 20 × 20 matrix based on the PSSM profile obtained from PSI-BLAST.

a matrix with $w \times m$ elements, where w stands for the sequence length (ranging from −15 to +15) and m represents 20 types of amino acids, which is row-centered at modified site.

$$Profile_x = \begin{bmatrix} p_{x,-15}(1) & \cdots & p_{x,-15}(m) \\ \vdots & \ddots & \vdots \\ p_{x,+15}(1) & \cdots & p_{x,+15}(m) \end{bmatrix} \tag{7}$$

Then, the $w \times m$ matrix was transformed into a matrix with 20 × 20 features $S_x(i, j)$, where i and j range from 1 to 20, by summing up the rows that were involved in the same type of amino acid i.

$$PSSM_x = \begin{bmatrix} S_x(1,\ 1) & \cdots & S_x(1,\ 20) \\ \vdots & \ddots & \vdots \\ S_x(20,\ 1) & \cdots & S_x(20,\ 20) \end{bmatrix} \tag{8}$$

Finally, each element in the 20 × 20 matrix was divided by the window length w and then normalized using a sigmoid function:

$$N_x(i, j) = \frac{1}{1 + e^{\frac{-S_x(i,j)}{w}}} i, j = 1, \ldots, 20 \tag{9}$$

Characterization of substrate site signatures. To investigate into the substrate-site specificity of succinylated sites, the maximal dependence decomposition (MDD)[40] was employed to divide positive training sequences into several groups with potentially conserved motifs. The MDD has been reported to having the ability to enhance the predictive effectiveness of computationally identifying substrate sites on different PTM types[31,35,56]. For reaching this purpose, a chi-squared test $\chi^2(P_i, P_j)$ is adopted to examine the intrinsic interdependence between two positions, P_i and P_j, which are in the neighboring upstream and downstream regions of succinylation sites. Amino acids, 20 in total, are categorized into five groups, based on their physicochemical properties: polar, acidic, basic, hydrophobic, and aromatic. Given two positions P_i and P_j, the occurring frequency of the presence of each amino acid group is determined for the elements of a contingency table. The chi-squared test is defined as:

$$\chi^2(P_i, P_j) = \sum_{m=1}^{5}\sum_{n=1}^{5} \frac{(K_{mn} - Q_{mn})^2}{Q_{mn}} \tag{10}$$

where K_{mn} is the number of positive training sequences containing amino acids in group m at position P_i and containing amino acids in group n at position P_j, for each pair (P_i, P_j) and $i \neq j$. The expectation value Q_{mn} is obtained from

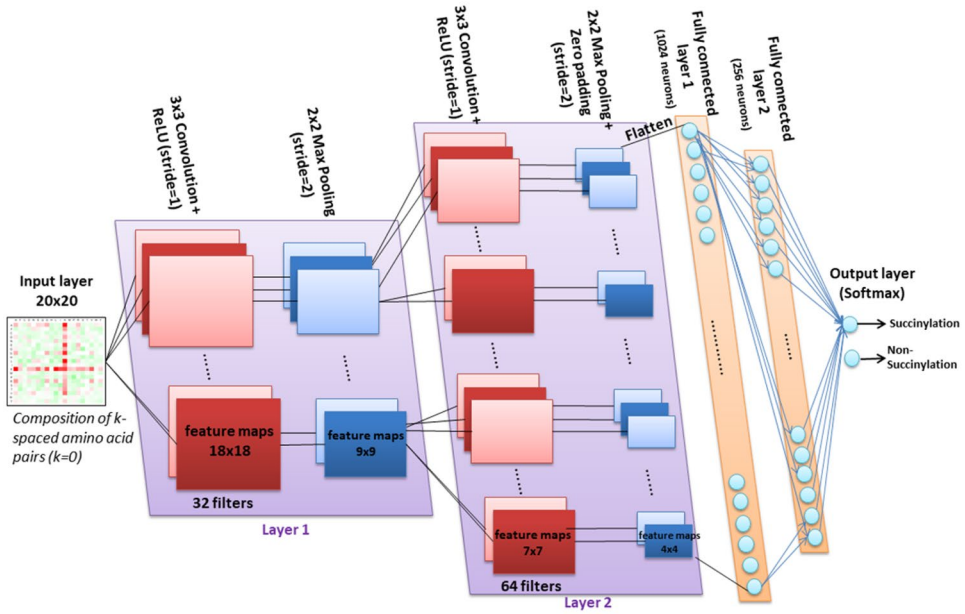


Figure 3. Schematic diagram of incorporating deep convolutional neural network with CKSAAP attribute ($K = 0$) to learn a predictive model with two-node output layer. A total of eight layers were implemented in this work, such as one input layer, two convolution layers, two max pooling layers, two fully connected layers, and one output layer. For each dense layer, the ReLU activation function was applied to avoid gradient diffusion. In addition, the dropout step was conducted in the hidden layers with an attempt to reduce overfitting. Finally, the output layer is composed of two nodes corresponding to the classifying results based on a softmax function.

$$Q_{mn} = \frac{K_{nR} - K_{Cn}}{K} \tag{11}$$

where $K_{mR} = K_{m1} + \ldots + K_{m5}$, $K_{Cn} = K_{1n} + \ldots + K_{5n}$ and K stands for the total number of positive training sequences. The $\chi^2(P_i, P_j)$ is a significant dependence if its value is larger than 34.3, based on the p-value of 0.01 with degree of freedom at 16[57]. When performing MDD on the dataset of all positive training sequences, the parameter of maximum cluster size should be specified with an appropriate cutoff value. The MDD clustering process will be terminated when all the group sizes are less than the specified value of maximum cluster size.

Construction of deep neural networks. This study involved a binary classification of lysine residues into succinylated and non-succinylated sites. Due to the emergence of applying deep learning methods in bioinformatics[58], we utilized a deep convolutional neural network (CNN), which is an extension of an artificial neural network (ANN) with multiple hidden layers between input and output layers (Supplementary Fig. S2). With the increasing number and complexity of high-throughput biological datasets, CNNs can decipher more complicated patterns and relationships within the investigated attributes than a traditional ANN, which only includes one hidden layer. A significant increase in the data count of MS/MS-identified protein succinylation would enable the number of neurons required in each layer to increase exponentially along with the potential patterns. Hence, this work exploited a CNN to learn the predictive models using various types of sequence-based attributes. In recent years, CNNs have been extended to incorporate convolution and pooling strategies in hidden layers to reduce the quantity of weights and complexity of calculations, respectively, when generating network structure. When implementing a CNN model, it is necessary to determine the number of convolution and pooling layers and choose a classification function for the output layer. As presented in Fig. 3, the first layer of CNN is the input layer. The AAPC attribute, represented as a matrix with 20×20 elements, was used as an example for constructing the CNN model.

When developing a CNN model, the convolution layer is the core layer, which functions as a pattern scanner and contains two major parameters: filters (or kernels) and stride. Each filter, which can be regarded as a small pattern with specified matrix size (e.g. 3×3 used in this work), is convolved across the width (20) and height (20) of the input data, based on the dot product between the elements of the filter and the input data in order to create new feature maps. We specified the value of stride as 1, then moved the filter one pixel at a time, and the input data with a 20×20 matrix size can be transformed into a new feature map with a matrix size of $(20-3+1) \times (20-3+1)$. The number of filters controls the depth (the number of neurons) in the convolution layer that may detect a specific type of pattern connecting to the input data. In addition to filters and stride, zero padding is a convenient approach to pad the input with zeros on the border of the matrix. Zero padding can be used to control the matrix size of input data.

The pooling layers, which comprise another critical part of a CNN model, usually immediately follow the convolution layers. Max pooling is a sort of non-linear down-sampling strategy used frequently for CNN construction. Typically, the max pooling layer can split the input matrix into a set of non-overlapping rectangles and can form a smaller matrix containing maximal outputs of each sub-region. Two major parameters used in max pooling are kernel size and stride, which are usually set as 2×2 and 2, respectively, for moving the 2×2 kernel along width or height 2 pixels at a time, discarding 75% of the activations. For instance, a feature map with matrix size of 18×18 in the convolution layer can be transformed into a smaller feature map with matrix size of 9×9 in the following max pooling layer. The function of max pooling is to reduce the amount of computing time in a CNN model and examine if the patterns extracted from the corresponding convolution layer exist in the input data or not[59].

After two convolution and max pooling layers, the highly complicated CNN modeling was accomplished by fully connected layers. Before getting into the fully connected layer, the flattening step (flatten layer) is a necessity that can be used to convert the matrix of input data into a vector. The flattening process is typically used prior to the fully connected layer. In a general CNN model, neurons in a fully connected layer have full links to all activations in the previous layer, as shown in Fig. 3. Thus, all the activations in the previous layer can be summarized by matrix multiplication along with a set of weight values on the links. Due to the occupation of most neurons in fully connected layers, an over-fitting problem might easily occur during CNN model construction. Herein, the dropout layer has been adopted to randomly mask a specified portion of its neurons in order to prevent CNN model construction from an over-fitting problem[60]. The dropout layer is carried out by dropping out the neurons with a specified probability P and retaining the neurons with probability $1 - P$. The value of probability P ranges from 0 to 1 with an attempt to determine the best P value for optimizing predictive performance. After that, we obtained a reduced network, in which the incoming and outgoing links to the dropped-out neurons are also eliminated.

As for the binary classification between succinylated and non-succinylated sites, the output layer comprised two neurons corresponding to the classification results based on a softmax function. The two nodes in the output layer were fully connected to the neurons of the previous layer. The softmax function could be regarded as a loss function by specifying how to penalize the difference between the predicted and true classes. The softmax function (or normalized exponential function) is a kind of logistic function that can be used to represent a probability distribution over K different categories. In this work, the value of K was set as two for the succinylated and non-succinylated datasets. Given a sample vector x and a weight vector w, the predicted probability for j-th class by the softmax function is defined as

$$\mathrm{P}(\mathrm{class} = \mathrm{j}/x) = \frac{e^{x^T w_j}}{\sum_{i=1}^{K} e^{x^T w_i}}, j = 1 \text{ or } 2 \tag{12}$$

This can be regarded as the probability of x for the j-th class against the composition of K linear functions: $x^T w_i$, $i = 1, \ldots, K$. Additionally, the ReLU is frequently used as the activation function when generating a CNN model with an enhanced nonlinear property but without a significant penalty for generalization accuracy[61]. In this work, the ReLU function was also employed to avoid gradient diffusion during the process of CNN construction. The ReLU function is defined as: $\mathrm{ReLU}(x) = \max(0, x)$. Another two activation functions are the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ and the hyperbolic tangent function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Performance evaluation of predictive models. In the generation of CNN models, the k-fold cross-validation was employed to evaluate their predictive performances. When implementing k-fold cross-validation, all the training data, including positive and negative sequences, were randomly clustered into k equal-sized subgroups. After having k subgroups, k-1 of them shall be regarded as the training sample and the remaining one subgroup was considered as the validation sample. In a round of k-fold cross-validation, each of the k subgroups should be considered as the validation sample once in turn. Sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC) have been used as the metrics to determine the performance of the generated models. The four metrics are defined as:

$$Sn = \frac{TP}{TP + FN} \tag{13}$$

$$Sp = \frac{TN}{TN + FP} \tag{14}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{16}$$

where TP, FN, TN, and FP denote the instances of true positive, false negative, true negative, and false positives, respectively. Due to the unbalanced positive and negative training datasets in this work, we have decided to choose MCC value as a major benchmark for achieving a relatively balanced sensitivity and specificity. After evaluating the of k-fold cross-validation, the CNN model reaching the best predictive performance was further evaluated by an independent testing dataset that was not included at all in the training dataset.

Independent testing. Due to the potential over-fitting issue originating from the training dataset, the predictive power of the generated models might be overestimated. Thus, the use of an independent testing dataset was necessary to further evaluate of the real case. In this study, the independent testing dataset was mainly collected from dbPTM[44,45,62]. Before the extraction of positive and negative testing sequences, the experimentally verified succinylated proteins in testing dataset were compared with training dataset in order to eliminate the homologous protein sequences between the two datasets. When extracting sequence fragments using the same window length as used in constructing the training dataset, the fragmented sequences might be overlapped between the two datasets. Hence, CD-HIT software was used again to delete fragmented sequences with 30% similarity. After that, the final dataset for independent testing contained 218 succinylated and 2621 non-succinylated entries. Moreover, the testing dataset was utilized to make a comparison between the proposed deep-learning models and other machine learning schemes in terms of predictive performance. Another cause of over-fitting might be due to the training process of the CNN. To avoid the over-fitting problem, we only used two convolution layers with lower filters to reduce the complexity of our model by minimizing the possible training parameters[59].

Results and Discussion

Substrate site signatures of lysine succinylation. The amino acid composition (AAC) was a feasible scheme to explore the potential motif of conserved residues around the succinylation sites based on the fragments with 31-mer sequence length. Since comparing the AAC between positive and negative datasets, the residues having significant differences could be regarded as useful attributes for succinylated sites prediction. Supplementary Fig. S3 showed that, for succinylated sites, the positively charged lysine (K) residue appeared to have the highest frequency around the substrate sites. In addition to AAC, the position-specific AAC neighboring the succinylation sites can be displayed by frequency plots of WebLogo[63]. As illustrated in Fig. 4A, there is no any amino acid having significantly high frequency near the succinylation sites, but the slightly prominent amino acid residues included Leucine (L), Lysine (K), Alanine (A), and Valine (V). Without conserved motifs observed in frequency plot, the TwoSampleLogo[64] program was further applied to compare the differences of position-specific AAC between succinylated and non-succinylated sequences. As displayed in Fig. 4B, when comparing with the sequence logo of non-succinylated sites (Fig. 4C), the most conserved motifs appeared to be associated with charged residues, in particular the positively charged K and arginine (R) residues on positions −11~−4 and +3~+12. Additionally, the negatively charged amino acids, such as aspartic acid (E), located at positions −2, +1 and +2.

A hierarchical clustering analysis was performed on the detection of motif signatures by categorizing all positive training sequences into seven subgroups that possess statistically significant dependencies of amino acid composition around the substrate sites. The MDD-clustered subgroups with motif signatures for the 5842 non-homologous succinylated sites are presented in Fig. 5 based on a tree-like structure. The motif in Group1 (933 sequences) is the significant occurrence of basic amino acids (K, R, and H) at position −5, with the highest dependence value among all subgroups. In the meantime, the remaining 4909 sequences are further analyzed based on the maximal dependency in the occurrence of amino acids neighboring the substrate sites. The Group2 (466 sequences) possesses a similar motif of basic amino acids at position −4. Additionally, the Group3 (398 sequences) and Group4 (832 sequences) also have the motif of basic amino acids at position +4 and +1, respectively. This investigation demonstrates that the detected motif signatures are consistent with the observation in two-sample logo, which having positively charged residues conserved in the upstream and downstream regions of succinylated sites. On the other hand, the Group5 (905 sequences) has the conserved motif of acidic residues at position +1. The Group6 also reveals that the position +1 is potent that contains the motif signature of polar and uncharged amino acids. The remaining data in the Group7 contain a slightly significant character in position +1.

Performance evaluation of CNN models trained with single attributes. In an attempt to examine the optimal window size for yielding the best performance, various window size values were adopted to extract the training sequences for model construction. After comprehensive analyses of performance comparisons, the window size of 31 (−15 to +15; with the succinylated residue in the center) achieved the best prediction performance, which is consistent with the difference of position-specific AACs between positive and negative training sequences. Based on the investigated features, their corresponding CNN models were built to determine the effectiveness of those features in identifying succinylation sites. As shown in Table 2, the CNN model trained with PspAAC reached an accuracy of 73.36% and an MCC value of 0.371. The AAPC model performed slightly better than the PspAAC model, which yielded an accuracy of 76.48% and an MCC value of 0.428. In the investigation of k-spaced amino acid pairs, the CNN model trained with the composition of one-spaced amino acid pairs (K = 1) provided the best performance at 77.95% sensitivity, 76.63% specificity, 76.85% accuracy, and MCC value of 0.432. After extracting the top 400 k-spaced amino acid pairs (K = 1−5) based on mRMR, the performance of the CNN model trained with the selected CKSAAP (top400) showed remarkable improvement, reaching a sensitivity of 85.35%, specificity of 83.49%, accuracy of 83.79%, and MCC value of 0.569. Among these CNNs, the model trained with the PSSM feature performed best for discriminating between succinylated and non-succinylated lysine residues. The PSSM model yielded a sensitivity, specificity, accuracy, and MCC value of 85.51%, 84.16%, 84.38%, and 0.579, respectively. Additionally, the ROC curve was generated to compare the predictive performance and stability of different CNN models (Supplementary Fig. S4). Regarding to the comparison among single features, the CNN model trained from the PSSM feature gave the best predictive power, which is consistent with the results reported in PSSM-Suc[65]. The area under ROC curve (AUC) of the CNN model trained with PSSM is 0.858. However, our investigation found that the CNN model trained with the composition of selected *k*-spaced amino acid pairs is comparable to that trained with the PSSM attribute.

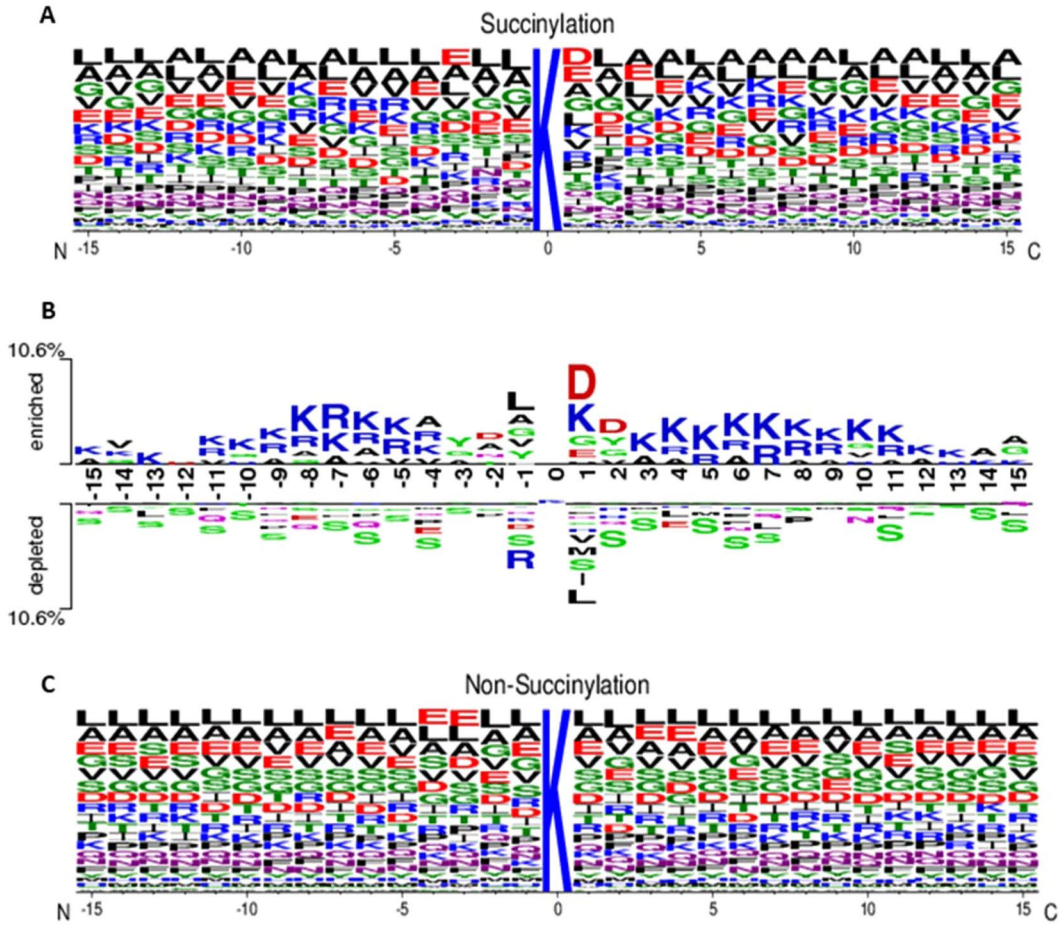


Figure 4. Position-specific amino acid composition of succinylated sites. (**A**) Position-specific amino acid composition of succinylated sequences based on the frequency plot of WebLogo. (**B**) Comparison of position-specific amino acid composition between succinylated and non-succinylated sequences based on TwoSampleLogo analysis. (**C**) Position-specific amino acid composition of non-succinylated sequences based on the frequency plot of WebLogo.

Performance evaluation of CNN models trained with hybrid attributes. In addition to the comparison of predictive powers among single attributes, we also consider a hybrid of multiple attributes to generate the predictive model. Based on the results of performance testing of single attributes, the PSSM, which can yield the best performance, was selected as the principal attribute for the combination with other single attributes. Consequently, a total of three hybrids, such as PSSM + PspAAC, PSSM + CKSAAP(top400), and PSSM + PspAAC + CKSAAP(top400), were further evaluated for uncovering their predictive capabilities in the succinylation site identification. As presented in the Table 2, the CNN model trained using the hybrid of PSSM and PspAAC attributes can reach a comparable performance with that trained using single PSSM attribute. In this investigation, the CNN model trained using the hybrid of PSSM and CKSAAP (top400) could perform best with the sensitivity of 86.94%, the specificity of 85.43%, the accuracy of 85.68%, and the MCC value of 0.608. However, the CNN model trained with the combination of all features performs slightly worse than that trained with the hybrid of PSSM and CKSAAP (top400). Additionally, Supplementary Fig. S4 revealed that the CNN model trained using the hybrid of PSSM and CKSAAP (top400) can outperform other CNN models in terms of ROC curves comparison. The AUC value of the CNN model trained with PSSM and CKSAAP (top400) is 0.886.

Performance comparison between CNN and other machine learning methods. To demonstrate the effectiveness of the deep learning method in PTM prediction, the predictive performance of this CNN model was compared with that of three popular machine learning methods: decision tree (DT), support vector machine (SVM), and random forest (RF). As summarized in Supplementary Table S1, the SVM and RF algorithms have been widely utilized to identify protein succinylation sites. In this work, the Classification and Regression Trees (CART) was employed to generate binary DTs for classifying between positive and negative instances. Based on the scikit-learn package[66], the function '*DecisionTreeClassifier*' was used to construct a classification tree by a top-down recursion. During the construction process, the 'best' feature set was selected to classify the training tuples that make a split in the tree. In addition, the CART program specified the 'Gini index' as the feature set selection approach. For the construction of RFs, the CART was again adopted to generate multiple trees with the

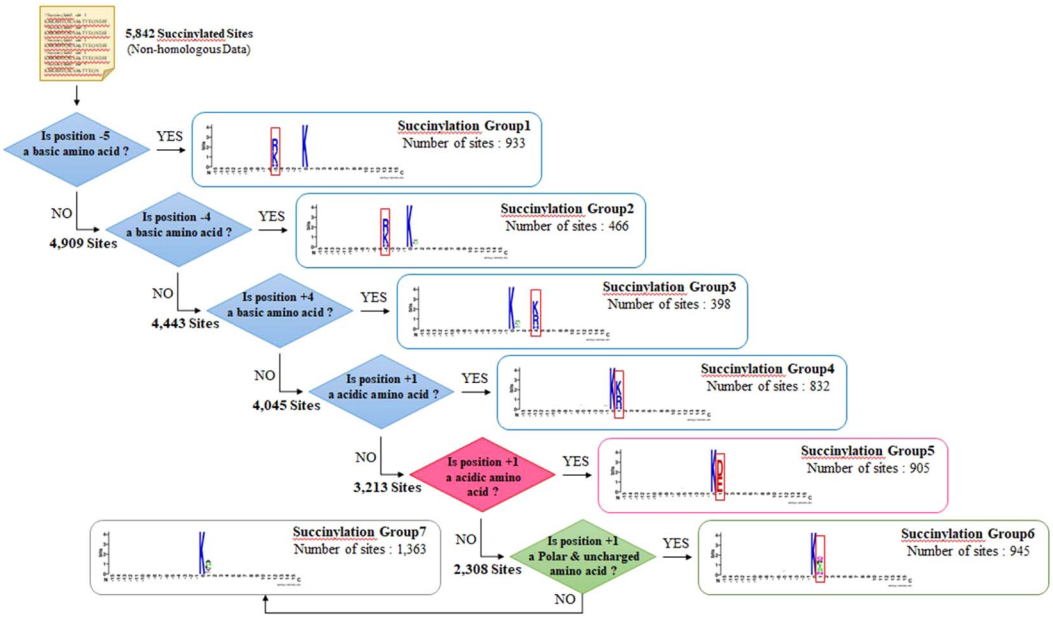


Figure 5. A hierarchical MDD-clustering process on the detection of motif signatures from 5842 succinylated sequences.

Attribute	Number of true positives	Number of false positives	Number of true negatives	Number of false negatives	Sensitivity	Specificity	Accuracy	MCC
PspAAC	2400	4411	12001	816	74.63%	73.12%	73.36%	0.371
CKSAAP (K = 0)	2512	3912	12500	704	78.11%	76.16%	76.48%	0.428
CKSAAP (K = 1)	2507	3835	12577	709	77.95%	76.63%	76.85%	0.432
CKSAAP (K = 2)	2501	3832	12580	715	77.77%	76.65%	76.83%	0.431
CKSAAP (K = 3)	2512	3912	12500	704	78.11%	76.16%	76.48%	0.428
CKSAAP (K = 4)	2494	3890	12522	722	77.55%	76.30%	76.50%	0.425
CKSAAP (K = 5)	2489	4079	12333	727	77.39%	75.15%	75.51%	0.412
CKSAAP (top400)	2745	2710	13702	471	85.35%	83.49%	83.79%	0.569
PSSM	2750	2600	13812	466	85.51%	84.16%	84.38%	0.579
PSSM + PspAAC	2759	2560	13852	457	85.79%	84.40%	84.63%	0.584
PSSM + CKSAAP (top 400)	2796	2391	14021	420	86.94%	85.43%	85.68%	0.608
PSSM + PspAAC + CKSAAP (top 400)	2789	2454	13958	427	86.72%	85.05%	85.32%	0.600

Table 2. Evaluation of ten-fold cross-validation on deep learning models trained with various types of sequence-based attributes.

'bootstrap aggregation' (bagging) of data sampling. In scikit-learn package, the function '*RandomForestClassifier*' was applied to measure the importance of training features and to generate the RF models. More specifically, Gini importance is the average decreased impurity of each feature across all trees; this impurity was the least-randomness of the given data. Moreover, the function '*svm.SVC*' in the scikit-learn package was used to train the binary SVM classifiers. The 'radial basis function' (RBF) was selected as the kernel function of SVM to transform the training data into a higher-dimensional vector space, with an attempt to search for a linearly optimal separating hyperplane.

According to the predictive performance of previous studies that have incorporated SVM or RF into their model construction, the SVM or RF models trained with combinatorial attributes could perform with reliable prediction accuracies. Based on the evaluation of ten-fold cross-validation, among the sequence-based attributes, this investigation has revealed that the DT model trained with PspAAC performed better than other attribute types (Supplementary Table S2). Instead of the PSSM attribute, both SVM and RF methods could reach a better performance by using the composition of the top 400 *k*-spaced amino acid pairs. Herein, the RF model performs slightly better than the SVM model in terms of MCC value. In addition to the comparison of different models trained using single attribute type, a hybrid of multiple attribute types was further considered into the generation of predictive models. Table 3 shows the comparison of ten-fold cross-validation between deep learning method and other three learning methods, on the basis of combining various attributes. Based on these sequence-based features, this investigation revealed that the CNN model trained using PSSM and CKSAAP(top400), which can

Method	Attribute	Number of true positives	Number of false positives	Number of true negatives	Number of false negatives	Sensitivity	Specificity	Accuracy	MCC
Decision tree	PspAAC + CKSAAP(top 400)	2282	4435	11977	934	70.96%	72.98%	72.64%	0.343
Support vector machine	PSSM + CKSAAP(top 400)	2622	3211	13201	594	81.53%	80.44%	80.61%	0.502
Random forest	PspAAC + PSSM + CKSAAP(top 400)	2605	2710	13702	611	81.00%	83.49%	83.08%	0.537
Deep learning	PSSM + CKSAAP(top 400)	2796	2391	14021	420	86.94%	85.43%	85.68%	0.608

Table 3. Comparison of ten-fold cross-validation between deep learning method and other machine learning methods.

Method	Number of true positives	Number of false positives	Number of true negatives	Number of false negatives	Sensitivity	Specificity	Accuracy	MCC
iSuc-PseAAC	31	310	2311	187	14.22%	88.17%	82.49%	0.019
SuccFind	101	921	1700	117	46.33%	64.86%	63.44%	0.062
pSuc-Lys	112	421	2200	106	51.38%	83.94%	81.44%	0.241
SuccinSite	98	221	2400	120	44.95%	91.57%	87.99%	0.308
SuccinSite 2.0	131	230	2391	87	60.09%	91.22%	88.83%	0.410
GPSuc	159	380	2241	59	72.94%	85.50%	84.54%	0.397
Our method	184	341	2280	34	84.40%	86.99%	86.79%	0.489

Table 4. Performance comparison between our method and six existing available prediction tools based on the independent testing dataset.

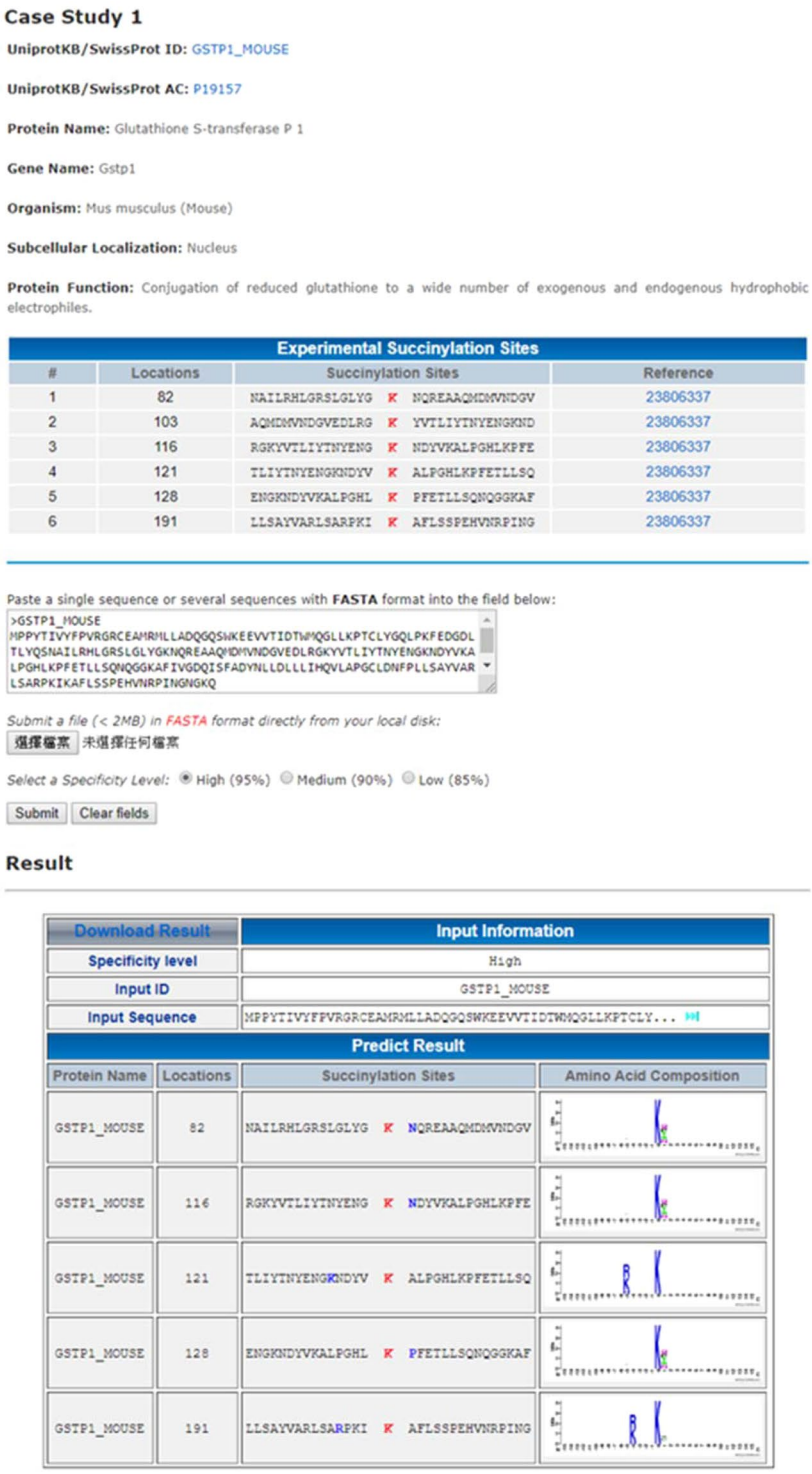
yield the sensitivity, specificity, accuracy, and MCC values at 86.94%, 85.43%, 85.68%, and 0.608, respectively, can outperform other three learning methods. However, it is noteworthy that the RF model trained using the hybrid of PspAAC, PSSM, and CKSAAP(top400) attributes can yield a comparable performance (83.08% accuracy) to the CNN model. In conclusion, the proposed CNN model can outperform other three popular machine-learning methods, with reference to the comparison of predictive performances based on the evaluation of ten-fold cross-validation.

Performance evaluation using an independent testing dataset. When discriminating between succinylated and non-succinylated sequences, it is possible to generate a predictive model whose prediction accuracy is over-estimated due to an over-fitting problem. To avoid presenting an over-estimating performance, this work compiled a dataset for independent testing. These independent testing instances, which are not present in the training dataset, were used to measure the real ability of the proposed model. The independent testing dataset comprised a total of 218 positive and 2621 negative instances. The CNN model trained using the PSSM and CKSAAP(top400) attributes can yield a promising performance with a sensitivity of 84.40%, specificity of 86.99%, accuracy of 86.79%, and MCC value of 0.489. Additionally, to judge the practicality of the proposed model, the comparison between our model and six existing prediction tools was performed using the testing dataset. As displayed in Table 4, our proposed model achieved the highest MCC value, reaching 0.489. In this comparison, the SuccinSite 2.0 can provide the best predictive accuracy (88.83%), while its specificity (91.22%) was much higher than its sensitivity (60.09%). However, the overall performance of SuccinSite 2.0 did not outperform our method in terms of MCC value. Interestingly, as presented in Supplementary Fig. S5, most of the existing prediction tools can provide much better specificity values than sensitivity values. This might be because their models were generated by using the unbalanced positive and negative datasets. In an overall evaluation, the testing results have indicated that the proposed method can provide a more reliable and stable prediction capability than other existing prediction tools, in terms of balanced sensitivity and specificity.

Implementation of web-based prediction tool. To facilitate the functional analyses of protein succinylation, the proposed method has been utilized to implement a web-based tool, named CNN-SuccSite, for classifying between succinylated and non-succinylated sites. After submitting protein sequences in the FASTA format, the CNN-SuccSite will return the prediction results, including succinylated sites, their flanking amino acids, and the corresponding substrate motif signatures. A case study of succinylation site prediction on mouse Glutathione S-transferase P 1 (Gstp1) was utilized to demonstrate the effectiveness of CNN-SuccSite. The Gstp1 contains six verified succinylation sites at Lys-82, Lys-103, Lys-116, Lys-121, Lys-128, and Lys-191[67]. As presented in Fig. 6, the CNN-SuccSite can achieve an accurate prediction at five validated succinylaion sites, according to the corresponding motif signatures.

Conclusion

Due to the abundance of experimentally verified succinylation data obtained from public resources, we were motivated to develop a new method to predict protein succinylation sites based on a deep learning strategy. Systematic investigation of various attributes in the neighborhood of substrate sites were performed on large-scale succinyl-proteome data. In accordance with the results of 10-fold cross-validation, the CNN model trained with



Case Study 1

UniprotKB/SwissProt ID: GSTP1_MOUSE

UniprotKB/SwissProt AC: P19157

Protein Name: Glutathione S-transferase P 1

Gene Name: Gstp1

Organism: Mus musculus (Mouse)

Subcellular Localization: Nucleus

Protein Function: Conjugation of reduced glutathione to a wide number of exogenous and endogenous hydrophobic electrophiles.

Experimental Succinylation Sites

#	Locations	Succinylation Sites	Reference
1	82	NAILRHLGRSLGLYG K NQREAAQMDMVNDGV	23806337
2	103	AQMDMVNDGVEDLRG K YVTLIYTNYENGKND	23806337
3	116	RGKYVTLIYTNYENG K NDYVKALPGHLKPFE	23806337
4	121	TLIYTNYENGKNDYV K ALPGHLKPFETLLSQ	23806337
5	128	ENGKNDYVKALPGHL K PFETLLSQNQGGKAF	23806337
6	191	LLSAYVARLSARPKI K AFLSSPEHVNRPING	23806337

Paste a single sequence or several sequences with FASTA format into the field below:

>GSTP1_MOUSE
MPPYTIVYFPVRGRCEAMRMLLADQGQSWKEEVVTIDTWMQGLLKPTCLYGQLPKFEDGDL
TLYQSNAILRHLGRSLGLYGKNQREAAQMDMVNDGVEDLRGKYVTLIYTNYENGKNDYVKA
LPGHLKPFETLLSQNQGGKAFIVGDQISFADYNLLDLLLIHQVLAPGCLDNFPLLSAYVAR
LSARPKIKAFLSSPEHVNRPINGNGKQ

Submit a file (< 2MB) in FASTA format directly from your local disk:

瀏覽檔案 未選擇任何檔案

Select a Specificity Level: High (95%) Medium (90%) Low (85%)

Submit Clear fields

Result

Download Result | Input Information

Specificity level	High
Input ID	GSTP1_MOUSE
Input Sequence	MPPYTIVYFPVRGRCEAMRMLLADQGQSWKEEVVTIDTWMQGLLKPTCLY...

Predict Result

Protein Name	Locations	Succinylation Sites	Amino Acid Composition
GSTP1_MOUSE	82	NAILRHLGRSLGLYG K NQREAAQMDMVNDGV	
GSTP1_MOUSE	116	RGKYVTLIYTNYENG K NDYVKALPGHLKPFE	
GSTP1_MOUSE	121	TLIYTNYENGKNDYV K ALPGHLKPFETLLSQ	
GSTP1_MOUSE	128	ENGKNDYVKALPGHL K PFETLLSQNQGGKAF	
GSTP1_MOUSE	191	LLSAYVARLSARPKI K AFLSSPEHVNRPING	

Figure 6. Case study of succinylation site prediction on Glutathione S-transferase P 1 (Gstp1).

the hybrid of PSSM and CKSAAP(top400) attributes can outperform that trained with other attributes. Besides, this investigation also demonstrated that the CNN model could provide a better performance than three popular shallow machine learning methods, including DT, SVM, and RF. Moreover, the independent testing was performed and the results demonstrated that the selected CNN model could outperform other existing prediction tools. Based on the usage of the independent testing dataset, the CNN model trained with the hybrid of PSSM and CKSAAP(top400) attributes could yield a promising performance. We truly believe that our proposed approach will help facilitate the determination of succinylated lysine residues of proteins. In the future, the physicochemical properties, such as solvent accessibility[68], hydrophobicity[69], and side-chain orientation[70], can be considered for obtaining a better predictive performance. Additionally, the tertiary structures of succinylated proteins can be

used to extract more useful information for the characterization of succinylated substrate sites. A stand-alone software will be developed for providing a practical means to facilitate the determination of succinylated targets from a large-scale proteome data.

Received: 9 July 2019; Accepted: 18 October 2019;
Published online: 07 November 2019

References

1. Huang, H. *et al*. iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic acids research* **46**, D542–D550, https://doi.org/10.1093/nar/gkx1104 (2018).
2. Lenard, J. & Singer, S. J. Succinylation of gamma globulin. *Nature* **210**, 536–537 (1966).
3. Zhang, Z. *et al*. Identification of lysine succinylation as a new post-translational modification. *Nat Chem Biol* **7**, 58–63, https://doi.org/10.1038/nchembio.495 (2011).
4. Benit, P. *et al*. Unsuspected task for an old team: succinate, fumarate and other Krebs cycle acids in metabolic remodeling. *Biochimica et biophysica acta* **1837**, 1330–1337, https://doi.org/10.1016/j.bbabio.2014.03.013 (2014).
5. Ong, S. E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1**, 252–262, https://doi.org/10.1038/nchembio736 (2005).
6. Xie, Z. *et al*. Lysine succinylation and lysine malonylation in histones. *Mol Cell Proteomics* **11**, 100–107, https://doi.org/10.1074/mcp.M111.015875 (2012).
7. Okanishi, H. *et al*. Proteome-wide identification of lysine succinylation in thermophilic and mesophilic bacteria. *Biochimica et biophysica acta* **1865**, 232–242, https://doi.org/10.1016/j.bbapap.2016.11.009 (2017).
8. Shen, C. *et al*. Succinyl-proteome profiling of a high taxol containing hybrid Taxus species (Taxus x media) revealed involvement of succinylation in multiple metabolic pathways. *Scientific reports* **6**, 21764, https://doi.org/10.1038/srep21764 (2016).
9. Xie, L. *et al*. First succinyl-proteome profiling of extensively drug-resistant Mycobacterium tuberculosis revealed involvement of succinylation in cellular physiology. *Journal of proteome research* **14**, 107–119, https://doi.org/10.1021/pr500859a (2015).
10. Liu, C. *et al*. Quantitative proteome and lysine succinylome analyses provide insights into metabolic regulation in breast cancer. *Breast cancer*. https://doi.org/10.1007/s12282-018-0893-1 (2018).
11. Zhao, X., Ning, Q., Chai, H. & Ma, Z. Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique. *Journal of theoretical biology* **374**, 60–65, https://doi.org/10.1016/j.jtbi.2015.03.029 (2015).
12. Xu, H. D., Shi, S. P., Wen, P. P. & Qiu, J. D. SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. *Bioinformatics* **31**, 3748–3750, https://doi.org/10.1093/bioinformatics/btv439 (2015).
13. Xu, Y. *et al*. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. *Scientific reports* **5**, 10184, https://doi.org/10.1038/srep10184 (2015).
14. Hasan, M. M., Yang, S., Zhou, Y. & Mollah, M. N. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. *Molecular bioSystems* **12**, 786–795, https://doi.org/10.1039/c5mb00853k (2016).
15. Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K. C. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Analytical biochemistry* **497**, 48–56, https://doi.org/10.1016/j.ab.2015.12.009 (2016).
16. Hasan, M. M., Khatun, M. S., Mollah, M. N. H., Yong, C. & Guo, D. A systematic identification of species-specific protein succinylation sites using joint element features information. *International journal of nanomedicine* **12**, 6303–6315, https://doi.org/10.2147/IJN.S140875 (2017).
17. Lopez, Y. *et al*. Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. *BMC genomics* **19**, 923, https://doi.org/10.1186/s12864-017-4336-8 (2018).
18. Dehzangi, A. *et al*. Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. *PloS one* **13**, e0191900, https://doi.org/10.1371/journal.pone.0191900 (2018).
19. Hasan, M. M. & Kurata, H. GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by aggregating multiple sequence features. *PloS one* **13**, e0200283, https://doi.org/10.1371/journal.pone.0200283 (2018).
20. Hasan, M. M., Khatun, M. S. & Kurata, H. Large-Scale Assessment of Bioinformatics Tools for Lysine Succinylation Sites. *Cells* **8**, https://doi.org/10.3390/cells8020095 (2019).
21. Xie, Y. B. *et al*. DeepNitro: Prediction of Protein Nitration and Nitrosylation Sites by Deep Learning. *Genom Proteom Bioinf* **16**, 294–306, https://doi.org/10.1016/j.gpb.2018.04.007 (2018).
22. Sahu, S. S. & Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Computational biology and chemistry* **34**, 320–327, https://doi.org/10.1016/j.compbiolchem.2010.09.002 (2010).
23. Chang, W. C. *et al*. Incorporating support vector machine for identifying protein tyrosine sulfation sites. *Journal of computational chemistry* **30**, 2526–2537, https://doi.org/10.1002/jcc.21258 (2009).
24. Park, K. J. & Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* **19**, 1656–1663 (2003).
25. Lu, C. T., Lee, T. Y., Chen, Y. J. & Chen, Y. J. An intelligent system for identifying acetylated lysine on histones and nonhistone proteins. *BioMed research international* **2014**, 528650, https://doi.org/10.1155/2014/528650 (2014).
26. Lee, T. Y., Chen, Y. J., Lu, T. C., Huang, H. D. & Chen, Y. J. SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. *PloS one* **6**, e21849, https://doi.org/10.1371/journal.pone.0021849 (2011).
27. Weng, S. L. *et al*. Investigation and identification of protein carbonylation sites based on position-specific amino acid composition and physicochemical features. *BMC bioinformatics* **18**, 66, https://doi.org/10.1186/s12859-017-1472-8 (2017).
28. Hasan, M. M. *et al*. Computational Identification of Protein Pupylation Sites by Using Profile-Based Composition of k-Spaced Amino Acid Pairs. *PloS one* **10**, e0129635, https://doi.org/10.1371/journal.pone.0129635 (2015).
29. Su, M. G., Huang, C. H., Lee, T. Y., Chen, Y. J. & Wu, H. Y. Incorporating amino acids composition and functional domains for identifying bacterial toxin proteins. *BioMed research international* **2014**, 972692, https://doi.org/10.1155/2014/972692 (2014).
30. Xu, H. *et al*. PLMD: An updated data resource of protein lysine modifications. *Journal of genetics and genomics = Yi chuan xue bao* **44**, 243–250, https://doi.org/10.1016/j.jgg.2017.03.007 (2017).
31. Bui, V. M., Lu, C. T., Ho, T. T. & Lee, T. Y. MDD-SOH: exploiting maximal dependence decomposition to identify S-sulfenylation sites with substrate motifs. *Bioinformatics* **32**, 165–172, https://doi.org/10.1093/bioinformatics/btv558 (2016).
32. Chen, Y. J. *et al*. GSHSite: exploiting an iteratively statistical method to identify s-glutathionylation sites with substrate specificity. *PloS one* **10**, e0118752, https://doi.org/10.1371/journal.pone.0118752 (2015).
33. Wong, Y. H. *et al*. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic acids research* **35**, W588–594, https://doi.org/10.1093/nar/gkm322 (2007).
34. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682, https://doi.org/10.1093/bioinformatics/btq003 (2010).

35. Weng, S. L., Kao, H. J., Huang, C. H. & Lee, T. Y. MDD-Palm: Identification of protein S-palmitoylation sites with substrate motifs based on maximal dependence decomposition. *PloS one* **12**, e0179529, https://doi.org/10.1371/journal.pone.0179529 (2017).
36. Bui, V. M. *et al*. SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. *BMC genomics* **17**(Suppl 1), 9, https://doi.org/10.1186/s12864-015-2299-1 (2016).
37. Lee, T. Y., Chang, C. W., Lu, C. T., Cheng, T. H. & Chang, T. H. Identification and characterization of lysine-methylated sites on histones and non-histone proteins. *Computational biology and chemistry* **50**, 11–18, https://doi.org/10.1016/j.compbiolchem.2014.01.009 (2014).
38. Bretana, N. A. *et al*. Identifying protein phosphorylation sites with kinase substrate specificity on human viruses. *PloS one* **7**, e40694, https://doi.org/10.1371/journal.pone.0040694 (2012).
39. Lee, T. Y. *et al*. Investigation and identification of protein gamma-glutamyl carboxylation sites. *BMC bioinformatics* **12**(Suppl 13), S10, https://doi.org/10.1186/1471-2105-12-S13-S10 (2011).
40. Lee, T. Y., Lin, Z. Q., Hsieh, S. J., Bretana, N. A. & Lu, C. T. Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics* **27**, 1780–1787, https://doi.org/10.1093/bioinformatics/btr291 (2011).
41. Huang, H. D., Lee, T. Y., Tzeng, S. W. & Horng, J. T. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic acids research* **33**, W226–229, https://doi.org/10.1093/nar/gki471 (2005).
42. Sahu, S. S. & Panda, G. A novel featue representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Computational Biology and Chemistry* **34**, 320–327 (2010).
43. Chang, W. C. *et al*. Incorporating support vector machine for identifying protein tyrosine sulfation sites. *J Comput Chem* (2009).
44. Huang, K. Y. *et al*. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res* **44**, D435–446, https://doi.org/10.1093/nar/gkv1240 (2016).
45. Lee, T. Y. *et al*. dbPTM: an information repository of protein post-translational modification. *Nucleic acids research* **34**, D622–627, https://doi.org/10.1093/nar/gkj083 (2006).
46. Lu, C. T. *et al*. DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic acids research* **41**, D295–305, https://doi.org/10.1093/nar/gks1229 (2013).
47. Zien, A. *et al*. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* **16**, 799–807 (2000).
48. Byvatov, E. & Schneider, G. Support vector machine applications in bioinformatics. *Applied bioinformatics* **2**, 67–77 (2003).
49. Dennis, G. Jr *et al*. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome biology* **4**, P3 (2003).
50. Lv, H. *et al*. Carspred: a computational tool for predicting carbonylation sites of human proteins. *PLoS One* **9**, e111478 (2014).
51. Lee, T. Y., Chen, S. A., Hung, H. Y. & Ou, Y. Y. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *PloS one* **6**, e17331, https://doi.org/10.1371/journal.pone.0017331 (2011).
52. Hsu, J. B., Bretana, N. A., Lee, T. Y. & Huang, H. D. Incorporating evolutionary information and functional domains for identifying RNA splicing factors in humans. *PloS one* **6**, e27567, https://doi.org/10.1371/journal.pone.0027567 (2011).
53. Xie, D., Li, A., Wang, M., Fan, Z. & Feng, H. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic acids research* **33**, W105–110, https://doi.org/10.1093/nar/gki359 (2005).
54. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* **292**, 195–202, https://doi.org/10.1006/jmbi.1999.3091 (1999).
55. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402 (1997).
56. Kao, H. J. *et al*. MDD-carb: a combinatorial model for the identification of protein carbonylation sites with substrate motifs. *BMC systems biology* **11**, 137, https://doi.org/10.1186/s12918-017-0511-4 (2017).
57. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78–94, https://doi.org/10.1006/jmbi.1997.0951 (1997).
58. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Briefings in bioinformatics* **18**, 851–869, https://doi.org/10.1093/bib/bbw068 (2017).
59. White, C., Ismail, H. D., Saigo, H. & Kc, D. B. CNN-BLPred: a Convolutional neural network based predictor for beta-Lactamases (BL) and their classes. *BMC bioinformatics* **18**, 577, https://doi.org/10.1186/s12859-017-1972-6 (2017).
60. Baldi, P. & Sadowski, P. The Dropout Learning Algorithm. *Artificial intelligence* **210**, 78–122, https://doi.org/10.1016/j.artint.2014.02.004 (2014).
61. Yarotsky, D. Error bounds for approximations with deep ReLU networks. *Neural networks: the official journal of the International Neural Network Society* **94**, 103–114, https://doi.org/10.1016/j.neunet.2017.07.002 (2017).
62. Huang, K. Y. *et al*. dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic acids research* **47**, D298–D308, https://doi.org/10.1093/nar/gky1074 (2019).
63. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188–1190, https://doi.org/10.1101/gr.849004 (2004).
64. Vacic, V., Iakoucheva, L. M. & Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22**, 1536–1537, https://doi.org/10.1093/bioinformatics/btl151 (2006).
65. Dehzangi, A. *et al*. PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction. *Journal of theoretical biology* **425**, 97–102, https://doi.org/10.1016/j.jtbi.2017.05.005 (2017).
66. Pedregosa, F. *et al*. Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825–2830 (2011).
67. Park, J. *et al*. SIRT5-mediated lysine desuccinylation impacts diverse metabolic pathways. *Molecular cell* **50**, 919–930, https://doi.org/10.1016/j.molcel.2013.06.001 (2013).
68. Lu, C. T., Chen, S. A., Bretana, N. A., Cheng, T. H. & Lee, T. Y. Carboxylator: incorporating solvent-accessible surface area for identifying protein carboxylation sites. *Journal of computer-aided molecular design* **25**, 987–995, https://doi.org/10.1007/s10822-011-9477-2 (2011).
69. Lee, T. Y. *et al*. N-Ace: using solvent accessibility and physicochemical properties to identify protein N-acetylation sites. *Journal of computational chemistry* **31**, 2759–2771, https://doi.org/10.1002/jcc.21569 (2010).
70. Chen, Y. J. *et al*. dbSNO 2.0: a resource for exploring structural environment, functional and disease association and regulatory network of protein S-nitrosylation. *Nucleic acids research* **43**, D503–511, https://doi.org/10.1093/nar/gku1176 (2015).

Author contributions

K.Y.H. and T.Y.L. carried out the data collection and curation, participated in the bioinformatics analyses, and drafted the manuscript. K.Y.H. carried out the web tool implementation. K.Y.H. and J.B.K.H. participated in the design of the study and performed the draft revision. T.Y.L. conceived of the study, and participated in its design and coordination and helped to revise the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-52552-4.

Correspondence and requests for materials should be addressed to T.-Y.L.