Computational and Structural Biotechnology Journal 27 (2025) 1048–1059

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

SUMO-LMNet: Lossless mapping network for predicting SUMOylation sites in SUMO1 and SUMO2 using high-dimensional features

Cheng-Hsun Ho [a], Yen-Wei Chu [b,c,d,e], Lan-Ying Huang [c], Chi-Wei Chen [f,g,*]

[a] Department of Medical Laboratory Science, College of Medical Science and Technology, I-Shou University, Kaohsiung City, Taiwan
[b] Graduate Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung City, Taiwan
[c] Doctoral Program in Medical Biotechnology, National Chung Hsing University, Taichung City, Taiwan
[d] Institute of Molecular Biology, National Chung Hsing University, Taichung City, Taiwan
[e] Smart Sustainable New Agriculture Research Center (SMARTer), Taichung City, Taiwan
[f] Graduate Degree Program of Smart Healthcare & Bioinformatics, I-Shou University, Kaohsiung City, Taiwan
[g] Department of Biomedical Engineering, I-Shou University, Kaohsiung City, Taiwan

ARTICLE INFO

Keywords:
SUMOylation prediction
SUMO paralogs
Sequence features
2D convolutional neural network
Feature importance analysis

ABSTRACT

Accurate SUMOylation site prediction is crucial for deciphering gene regulation and disease mechanisms. However, distinguishing SUMO1 and SUMO2 modifications remains a major challenge due to their structural similarities. Conventional prediction models often struggle to differentiate between these paralogues, limiting their applicability in biological research. To address this, we introduce SUMO-LMNet, a deep learning-based framework for the precise prediction of SUMO1 and SUMO2 sites. Unlike previous models, SUMO-LMNet integrates a lossless mapping strategy and deep learning architectures to enhance both prediction accuracy and interpretability. Our model extracts high-dimensional features from sequences and transforms them into two-dimensional feature maps, enabling convolutional neural networks (CNNs) to effectively capture both local and global dependencies within the data. By leveraging a Lossless Mapping Network (LM-Net), this approach preserves the original feature space, ensuring that feature integrity is retained without loss of spatial information. While Grad-CAM highlights key features in individual predictions, it lacks consistency across samples and does not provide a dataset-wide evaluation of feature importance. To address this, we introduce Combined Heatmap Feature Analysis (CHFA), which systematically aggregates feature importance across multiple samples, providing a more reliable and interpretable dataset-wide assessment. Experimental results reveal distinct feature dependencies between SUMO1 and SUMO2, underscoring the necessity of paralogue-specific predictive models. Through a systematic comparison of multiple neural network architectures, we demonstrate that our model achieves over 80 % accuracy in distinguishing SUMO1 and SUMO2 modification sites. By prioritizing candidate sites for further study, our model aids experimental design and accelerates the discovery of biologically relevant SUMOylation targets. SUMO-LMNet is publicly available at https://predictor.isu.edu.tw/sumo-lmnet.

Abbreviations: Acc, Accuracy; ASDC, Adaptive Skip Dipeptide Composition; AAC, Amino Acid Composition; AAIndex, Amino Acid Index; AUC, Area Under the Curve; AC, Auto Covariance; BatchNorm, Batch Normalization; CHFA, Combined Heatmap Feature Analysis; CKSAAP, Composition of k-Spaced Amino Acid Pairs; CTDD, Composition Transition Distribution - Distribution; CTriad, Conjoint Triad; CNN, Convolutional Neural Network; DPC, Di-Peptide Composition; EAAC, Enhanced Amino Acid Composition; FN, False Negative; FP, False Positive; Grad-CAM, Gradient-weighted Class Activation Mapping; KSCTriad, k-Spaced Conjoint Triad; AESNN3, Learn from Alignments; LM-Net, Lossless Mapping Network; MCC, Matthews Correlation Coefficient; MLP, Multi-Layer Perceptron; NMBroto, Normalized Moreau-Broto; 1D-CNN, One-Dimensional Convolutional Neural Network; OPF, Overlapping Property Features; PTM, Post-translational Modification; PDB, Protein Data Bank; PseKRAAC, Pseudo K-tuple Reduced Amino Acids Composition; ROC, Receiver Operating Characteristic; ReLU, Rectified Linear Unit; Sn, Sensitivity; SUMO, Small Ubiquitin-like Modifier; Sp, Specificity; TL, Transfer Learning; TN, True Negative; TP, True Positive; Conv2D, Two-Dimensional Convolution; 2D-CNN, Two-Dimensional Convolutional Neural Network.

* Corresponding author at: Graduate Degree Program of Smart Healthcare & Bioinformatics, I-Shou University, Kaohsiung City, Taiwan.
E-mail addresses: chenghsunho@gmail.com (C.-H. Ho), ywchu@nchu.edu.tw (Y.-W. Chu), d110001762@mail.nchu.edu.tw (L.-Y. Huang), cwchen@isu.edu.tw (C.-W. Chen).

https://doi.org/10.1016/j.csbj.2025.03.005
Received 2 December 2024; Received in revised form 2 March 2025; Accepted 4 March 2025
Available online 6 March 2025

1. Introduction

Post-translational modification (PTM) is a key mechanism influencing protein function and the dynamics of biological systems, encompassing modifications such as phosphorylation, glycosylation, acetylation, and ubiquitination [1]. SUMOylation, a reversible post-translational modification involving the Small Ubiquitin-like Modifier (SUMO), bears some structural resemblance to ubiquitination but differs considerably in biological function and charge distribution [2]. Since the discovery of SUMO in the 1990s, SUMOylation has been shown to play an essential role in various biological processes, including gene expression, DNA repair, chromosome assembly, and cell signaling, with aberrant SUMOylation associated with multiple diseases, such as neurodegeneration, congenital heart disease, diabetes, and cancer [3–6]. Therefore, accurately identifying SUMOylation sites is critical to understanding their roles in biological functions and disease [7,8].

The SUMO family is highly conserved across all eukaryotes, with SUMO1 and SUMO2 playing vital roles in cellular functions. In humans, five SUMO family members have been identified (SUMO1, SUMO2, SUMO3, SUMO4, and SUMO5), of which SUMO1 and SUMO2 are the most extensively studied [9]. SUMO1 consists of 101 amino acids and typically binds to target proteins as a monomer, while SUMO2 and SUMO3 share up to 95 % homology and can form polymeric chains, facilitating more complex protein interactions. Although SUMO1 and SUMO2 have only approximately 45 % structural homology, they exhibit similar three-dimensional structures but distinct functionalities. SUMO2's ability to form polychains enhances binding flexibility and plays a critical role in cellular stress responses [10], whereas SUMO1 primarily acts in a stable monomeric form [9]. These characteristics enable SUMO1 and SUMO2 to assume essential roles in various biological processes, including cell cycle regulation, DNA repair, and protein stability maintenance [11]. Their dysregulation is linked to numerous diseases, such as cancer, neurodegenerative disorders, and diabetes [12].

Identifying SUMO1 and SUMO2 modification sites is crucial for understanding their specific biological functions and regulatory mechanisms [13]. Bouchard et al. found significant differences in SUMO1 and SUMO2 gene expression across various human tissues, with SUMO2 binding sites approximately twice as numerous as those of SUMO1, which suggests broader binding range, diversity, and higher activity in certain tissues [9]. Furthermore, while SUMO1 and SUMO2 can each bind to over 1000 distinct proteins, their specificity and functional requirements vary among tissues [14]. Accurate identification of SUMO1 and SUMO2 modification sites on various target proteins would clarify molecular differences between SUMO paralogues and provide a foundation for developing targeted SUMOylation therapies, potentially opening new avenues for treating diseases like cancer [15,16]. Despite their functional significance, experimental identification of SUMOylation sites remains challenging due to the complexity of SUMO conjugation and has led to the development of computational prediction methods as an efficient alternative.

Using computational methods to predict SUMOylation sites effectively reduces experimental time and cost, accelerating the identification of target SUMO modification sites [17,18]. While traditional approaches, such as site-directed mutagenesis and mass spectrometry-based proteomics, are precise, they are also complex and time-intensive. Most existing SUMOylation prediction tools focus on broad-spectrum SUMOylation site identification without differentiating between SUMO paralogues [19]. However, emerging evidence suggests that SUMO1 and SUMO2 have distinct biological roles. Despite this, many current tools treat SUMOylation as a single event, overlooking paralogue-specific differences. To address this gap, we developed a dedicated computational approach capable of distinguishing SUMO1 and SUMO2 modification sites, providing a more precise tool for assisting experimental design. Computational predictions like these not only reduce experimental costs but also enhance research efficiency, offering robust support for functional studies of protein modifications and disease-related validations [20].

Most SUMOylation prediction tools consider the physicochemical properties of protein sequences, conserved motif analysis, or specialized algorithms to aid in predicting SUMOylation sites [19]. For instance, PCI-SUMO employs the Parallel Cascade Identification algorithm [21], SUMOsp 1.0, while GPS-SUMO uses their uniquely developed GPS algorithm [22,23], JASSA is based on the Position Frequency Matrix [24], SUMOgo examines the crosstalk between SUMOylation and other post-translational modifications [25], SUMO-Forest combines deep learning with decision trees [26], and DeepUbiSumoPre applies deep learning [27]. However, these tools extract a limited range of features from protein sequences; even recent deep learning models like DeepUbiSumoPre utilize only seven features, including the 20 amino acids encoded in one-hot format and six groups derived from the Amino Acid Index Database (AAindex) [28].

A key challenge in bioinformatics is the extraction of biologically meaningful features that effectively capture the biochemical, structural, and evolutionary properties of biomolecules. In protein sequence-based learning tasks, leveraging a diverse set of high-dimensional features can enhance predictive models by capturing richer biological information and improving their ability to differentiate SUMOylation sites from non-SUMOylation sites. However, increasing feature dimensionality also introduces challenges in feature selection and interpretability, necessitating architectures capable of effectively processing and analyzing complex biological data.

Deep learning has emerged as a powerful solution to address these challenges by automatically learning hierarchical feature representations from high-dimensional sequence encodings. Recent advances in deep learning have introduced various architectures for protein sequence analysis, including multi-layer perceptrons (MLP) and convolutional neural networks (CNNs). MLP, a fully connected neural network, processes input features in a flattened format, limiting its ability to capture sequential or spatial dependencies. 1D-CNN improves upon this by applying convolutional operations along sequence positions, preserving local relationships between amino acids but lacking the ability to extract higher-dimensional feature interactions. In contrast, 2D-CNN restructures sequence-derived features into a 2D feature matrix, allowing convolutional filters to capture spatial dependencies across multiple feature dimensions. This approach is particularly beneficial for high-dimensional encoding schemes. To systematically evaluate the predictive effectiveness of these architectures, we compared MLP, 1D-CNN, and 2D-CNN in identifying SUMO1 and SUMO2 modification sites, demonstrating that 2D-CNN consistently outperforms the other models in classification accuracy.

Building on recent advancements in feature encoding, Jiangning Song and colleagues have developed a tool capable of encoding over 65 protein sequence descriptors [29,30], facilitating AI-driven applications in sequence-based biological research. This study leverages a vast and enriched array of encodings, combined with the automatic feature selection capabilities of convolutional neural networks, to construct a predictive model for human SUMO1 and SUMO2. Although protein sequences provide limited information compared to Protein Data Bank (PDB) structural models, sequence encodings can transform character combinations into high-dimensional vectors that describe SUMOylation sites. However, deep learning, despite its strong predictive performance, can sometimes compromise model interpretability. To address this challenge, we developed a Lossless Mapping Network (LM-Net) 2D-CNN named SUMO-LMNet, which retains the original size of the feature space, allowing each pixel to clearly correspond to its feature significance. We transform features into two-dimensional feature maps for SUMO paralogue identification. Compared to 1D-CNNs, 2D-CNNs excel in identifying spatial dependencies, effectively capturing spatial information, making them particularly suited to handling high-dimensional feature data. To enhance interpretability, we integrated Grad-CAM [31] which provides localized feature importance for individual

predictions, and further developed Combined Heatmap Feature Analysis (CHFA) to aggregate feature significance across all samples. By leveraging the lossless mapping architecture of LM-Net, CHFA enables a more comprehensive and consistent assessment of key features, overcoming the variability in individual Grad-CAM results.

The experimental results demonstrate that our custom-built 2D-CNN model (LM-Net) achieves an accuracy exceeding 80 % in distinguishing SUMO1 and SUMO2, outperforming pre-trained deep networks. We also observed distinct dependencies on critical features between SUMO1 and SUMO2. Through CHFA, we found that SUMO2 relies more heavily on the physicochemical properties of amino acids as well as on mid- to long-range interactions and local amino acid composition, indicating a multilayered feature requirement for its structure and function. In contrast, SUMO1 predictions depend on a more balanced combination of diverse features. These findings suggest that developing separate predictive models for SUMO1 and SUMO2 could better capture their unique feature dependencies, enhancing prediction accuracy. This underscores the necessity of dedicated models for each SUMO paralogue.

2. Materials and methods

This study aims to develop a SUMO-specific predictive model, as illustrated in Fig. 1. First, we collect the necessary SUMOylation data and categorize SUMOylation annotations according to a variety of SUMO paralogues, facilitating the construction of prediction models tailored to each SUMO type. After data preprocessing, we cluster the protein sequences based on their species origin, enabling the development of species-specific prediction models. Next, we employ encoding tools to extract 69 features from the protein sequences and apply a sliding window approach to capture amino acid information surrounding the target prediction sites. The data is divided into training, validation, and testing sets for deep learning model development. Upon completing model architecture testing, we analyze feature maps to examine characteristic differences across various SUMO types.

2.1. Dataset collection and preprocessing

Experimentally verified SUMO-specific SUMOylation sites were gathered from UniProt [32], encompassing distinct SUMO1–5 annotations. The curated dataset includes protein names, species, SUMOylation sites, SUMO types, and protein sequences, with contradictory and redundant SUMOylation sites removed. Data was organized into species-specific datasets, with Table A.1 providing a statistical summary of SUMO1–5 data for humans, showing the number of human SUMOylation sites and the corresponding protein counts. "Position" refers to lysines annotated in the database as experimentally validated SUMOylation sites. SUMO1 has 358 annotated sites, while SUMO2 has 5291. Given the limited data available for SUMO3–5, the model was developed using SUMO1 and SUMO2 data. Lysines not annotated as SUMOylation sites in the database were treated as negative data. As some lysines may bind to both SUMO1 and SUMO2, two models were constructed to predict SUMO1- and SUMO2-mediated SUMOylation sites, as shown in Fig. 2. When a protein sequence is entered, these models predict which lysines may be modified by SUMO1 or SUMO2, allowing predictions at overlapping site

The training and testing sets were created using down-sampling to balance the number of positive and negative samples, ensuring that the model was not biased toward more frequent non-SUMOylation sites. For SUMO1, as only 358 positive samples were available, we randomly selected 358 negative samples to maintain a 1:1 positive-to-negative

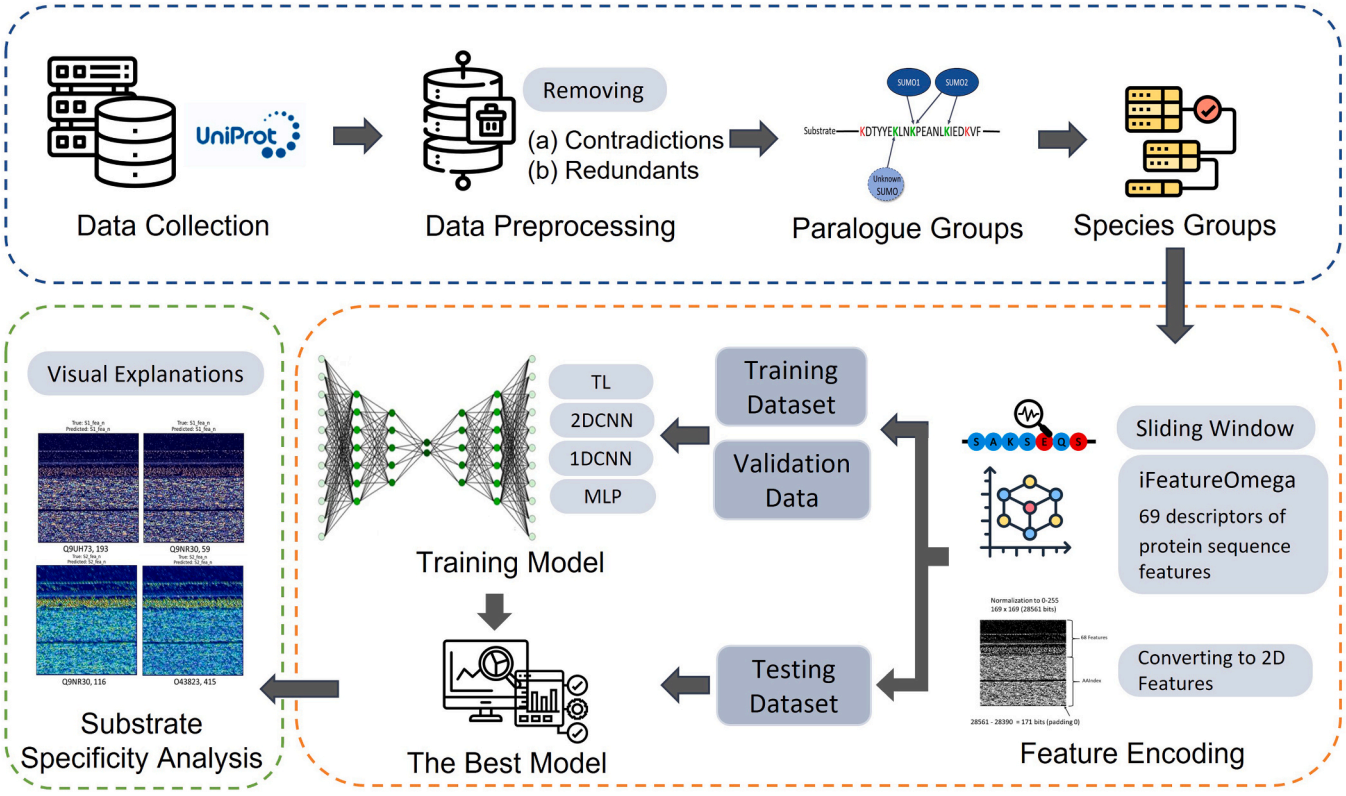

Fig. 1. Experimental flowchart. A schematic representation of the SUMO-LMNet development workflow. The process begins with the collection of experimentally verified SUMOylation site data from UniProt, followed by data preprocessing to remove contradictions and redundancies. Protein sequences are categorized into paralogue and species groups before feature extraction. Using a sliding window approach, 69 feature descriptors are encoded with iFeatureOmega and transformed into 2D feature representations. The dataset is then split into training, validation, and testing sets. Multiple deep learning models, including 1D-CNN, 2D-CNN, MLP, and transfer learning (TL) methods, are trained and evaluated to identify the best-performing predictive model. The final model undergoes substrate specificity analysis using visual explanations, providing insights into key feature dependencies.

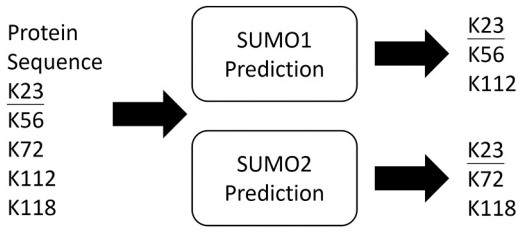

Fig. 2. Prediction model strategy. Since certain lysine residues can be modified by both SUMO1 and SUMO2 while sharing identical sequence features, an independent model construction strategy is adopted. By training separate predictive models for each paralogue, this approach ensures paralogue-specific learning and minimizes biases introduced by shared sequence characteristics. This strategy enables the identification of SUMOylation patterns unique to SUMO1 and SUMO2, improving prediction accuracy and model interpretability.

ratio. For SUMO2, we randomly down-sampled the negative dataset to match the 5204 positive samples, ensuring sequence diversity by selecting negative samples from different proteins. This approach prevents the model from being overly influenced by highly abundant negative samples from specific proteins. To further assess the robustness of the down-sampling strategy, the dataset was split into 80 % training and 20 % testing sets, with 20 % of the training data reserved for validation, as summarized in Table 1.

2.2. Sequence feature representation and encoding

To more effectively describe the differences between positive and negative samples, we not only considered the lysine site itself but also incorporated the surrounding 15 amino acids, creating a sliding window of 31 amino acid positions (Sliding Window 31), with gaps filled by the "-" symbol where necessary. As these sequence fragments consist only of characters, we used iFeatureOmega [30] to convert sequences into biologically meaningful numerical features.

From the 72 protein descriptors available in iFeatureOmega, we selected 69 features (listed in Table 2 and detailed in Table A.2) for feature encoding. This selection was based on computational efficiency to ensure that each potential SUMOylation site could be processed within one minute. Three descriptors (TPC type1, TPC type2, and KNN) were excluded because they required significantly longer computation times, which would hinder large-scale predictions.

For example, binary encoding represents each amino acid as a 20-dimensional vector, such as A: [1,0,0,…,0]. Thus, each lysine and its surrounding 30 amino acids are described by 31 × 20 (620) features. The AAIndex includes 531 methods for quantifying amino acids, assigning higher values to attributes like the short side chain of glycine. Each amino acid is converted into 531 numerical values, resulting in 531 × 31 (16,461) features for 31 amino acids. Combined with 68 additional sequence features, each site is ultimately described by a 28,390-dimensional vector.

These vectors were transformed into a 2D matrix (Fig. 3), enabling the convolutional network to better capture spatial relationships among feature regions and facilitating the observation of characteristic differences between SUMO1 and SUMO2. The matrix size is 169 × 169 (totaling 28,561 vectors), with 171 vacant positions filled with zeros. All values were normalized to a range of 0–255 and arranged according to the feature order in Table A.2, with AAIndex features occupying the latter portion.

Table 1
Training and Testing Data.

	Human SUMO1	Human SUMO2
All (Positive, Negative)	716 (358, 358)	10408 (5204, 5204)
Training (80 %)	572	8326
(Train, Validation)	(458, 114)	(6661, 1665)
Testing (20 %)	144	2082

Table 2
Characterization of 69 Protein Features.

Number	Feature	Bits	Number	Feature	Bits
1	AAC	20	…	…	…
2	EAAC	540	62	AESNN3	93
3	CKSAAP type 1	1600	63	OPF 10 bit	310
4	CKSAAP type 2	1600	64	OPF 7 bit type 1	217
24	NMBroto	24	65	OPF 7 bit type 2	217
25	AC	24	66	OPF 7 bit type 3	217
32	PseKRAAC type 1	4	67	BLOSUM62	620
33	PseKRAAC type 2	4	68	ZScale	155
34	PseKRAAC type 3 A	4	69	AAIndex	16461
…	…	…		Total	28390

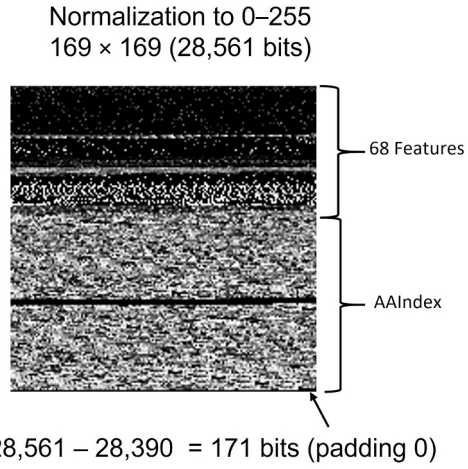

Fig. 3. Visualization of the 2D feature matrix. Each lysine residue and its surrounding sequence are encoded into a high-dimensional vector comprising 69 sequence features. These features, including physicochemical properties, amino acid compositions, and structural attributes, are transformed into a 169 × 169 two-dimensional matrix (28,561 bits) for deep learning processing. The AAIndex features are positioned in the lower portion of the matrix, while the upper region contains the remaining 68 features. To ensure matrix consistency, 171 padding bits (filled with 0 s) are appended. This transformation preserves the spatial relationships among features, enabling the 2D-CNN model to effectively capture both local and global feature interactions, thereby enhancing the accuracy of SUMOylation site prediction.

2.3. Model architecture and development

2.3.1. Lossless mapping network (LM-Net) design and implementation

This study employed a 2D convolutional neural network with five transfer learning models to evaluate their classification performance, utilizing Grad-CAM to analyze feature differences between SUMO1 and SUMO2. To preserve pixel-level feature meanings for subsequent analysis, pooling layers were excluded to retain spatial resolution and prevent feature loss. Additionally, a stride of 1 was applied to maintain the original spatial information, keeping the feature map dimensions unchanged.

The SUMO1 network architecture (Table 3, Table A.3, Fig. A.1(a)) consists of three convolutional layers followed by one fully connected layer, forming a relatively shallow network. The first convolutional layer contains 64 filters with a kernel size of 4 × 4, followed by Batch Normalization and ReLU activation. The second and third convolutional layers each contain 32 filters with a 3 × 3 kernel size, also followed by Batch Normalization and ReLU activation. After the convolutional layers, the feature maps are flattened, followed by a dropout layer (0.7) to prevent overfitting. The fully connected layer consists of 128 neurons, followed by Batch Normalization and ReLU activation, and the final

Table 3
Architecture of the Convolutional Neural Network for SUMO1.

Layer No.	Layer Type	Parameters
1	Input	shape= (169, 169, 3)
2–10	Conv2D + BatchNorm + ReLU (x3)	filters= [64, 32, 32], kernel_size= [(4, 4), (3, 3)], strides= (1,1), padding= 'same'
11	Flatten	-
12	Dropout	rate= 0.7
13–15	Dense + BatchNorm + ReLU	units= 128
16	Output (Dense)	units= 2, activation= 'softmax'

output layer includes two neurons with softmax activation for classification.

In contrast, the SUMO2 network architecture (Table 4, Table A.3, Fig. A.1(b)) is deeper, consisting of four convolutional layers and three fully connected layers. Each convolutional layer contains 64 filters with a 2 × 2 kernel size, followed by Batch Normalization and ReLU activation. Dropout layers with a rate of 0.25 are applied after each convolutional layer to mitigate overfitting. After feature extraction, the network includes three fully connected layers, each with 256 neurons, followed by Batch Normalization, ReLU activation, and an additional dropout layer (0.5). The final output layer consists of two neurons with softmax activation.

The architectures for SUMO1 and SUMO2 were determined through an iterative, layer-by-layer testing process to identify the optimal configuration for each paralogue. Various network depths and configurations were systematically evaluated based on performance metrics, including accuracy, sensitivity, and specificity, using a validation dataset. Empirical testing revealed that SUMO1 achieved optimal performance with a three-layer convolutional network, whereas SUMO2 required an additional convolutional layer and three fully connected layers to enhance feature extraction and classification performance.

2.3.2. Comparative evaluation of deep learning models

To compare the performance of other convolutional networks, we replaced the convolutional layer before flattening with MobileNetV2 [33], Xception [34], VGG16 [35], ResNet152V2 [36], and Inception-ResNetV2 [37]. Furthermore, to assess the feature extraction capability of a 1D-CNN, we substituted the 2D-CNN with a 1D-CNN while keeping the remaining network structure unchanged (Fig. A.1(c and d)). For models using only MLP, all convolutional layers before flattening were removed; the detailed structure is shown in Fig. A.1(e and f).

2.4. Model performance evaluation and validation

The performance evaluation criteria for this study are defined in Table A.5. If the predicted result and actual condition both indicate SUMOylation, it is classified as a True Positive; if the prediction is SUMOylation but the actual condition shows no SUMOylation, it is a False Positive; if the prediction shows no SUMOylation while the actual condition is SUMOylation, it is a False Negative; and if both prediction and actual condition show no SUMOylation, it is a True Negative.

Based on the definitions in Table A.5, accuracy and other performance metrics are calculated using formulae (1) to (5). Formula (1) is the Matthews Correlation Coefficient (MCC), used to evaluate the consistency between model predictions and actual results, with values ranging from [−1,1]. An MCC of 1 indicates perfect prediction accuracy, −1 indicates complete disagreement, and 0 suggests performance equivalent to random guessing. The formula is as follows:

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (1)$$

Formula (2) represents Precision, which evaluates the method's ability to correctly identify Positive cases. The formula is as follows:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Formula (3) represents Accuracy (Acc), which measures the overall prediction accuracy of the model. The formula is as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

Formula (4) represents Specificity (Sp), evaluating the model's ability to correctly identify negative cases. The formula is as follows:

$$Sp = \frac{TN}{TN + FP} \quad (4)$$

Formula (5) represents Sensitivity (Sn), assessing the model's ability to correctly recognize positive cases. The formula is as follows:

$$Sn = \frac{TP}{TP + FN} \quad (5)$$

2.5. Combined heatmap feature analysis (CHFA) method

This study employed Grad-CAM heatmap analysis to assess the significance of various features in the model's prediction results. Grad-CAM effectively visualizes the attention of convolutional neural network models toward specific regions of an image, providing valuable

Table 4
Architecture of the Convolutional Neural Network for SUMO2.

Layer No.	Layer Type	Parameters
1	Input	shape= (169, 169, 3)
2–17	Conv2D + BatchNorm + ReLU + Dropout (x4)	filters= 64, kernel_size= (2, 2), strides= (1,1), padding= 'same', activation= 'relu', kernel_regularizer=l2(0.01), bias_regularizer=l2(0.01), dropout_rate= 0.25
18	Flatten	-
19–30	Dense + BatchNorm + ReLU + Dropout (x3)	units= 256, dropout_rate= 0.5, kernel_regularizer=l2(0.01)
31	Output (Dense)	units= 2, activation= 'softmax'

interpretability. Furthermore, Grad-CAM determines feature importance using gradient information from convolutional layers, thereby circumventing the interpretive challenges that other methods may encounter owing to issues such as vanishing gradients or noise. However, because feature importance can vary across different samples, Grad-CAM alone may not provide a consistent, generalizable assessment of feature significance across the dataset. To address this limitation, we propose Combined Heatmap Feature Analysis (CHFA), which aggregates Grad-CAM results across all samples to compute statistical feature importance at each spatial location. CHFA enhances interpretability by leveraging the lossless mapping architecture of LM-Net, ensuring that feature integrity is preserved. This allows for a more comprehensive, dataset-wide analysis of feature dependencies, rather than being constrained to single-instance interpretations.

First, images were obtained from the training and testing sets, and corresponding heatmaps were generated using Grad-CAM to visualize the model's attention to a variety of regions. These heatmaps were subsequently overlaid and merged to create a comprehensive heatmap representing all images, facilitating an overall feature analysis. By aggregating feature contributions across multiple samples, CHFA reduces the influence of sample-specific variations and allows for a more stable, statistically robust evaluation of feature importance. The composite heatmap was derived by all individual heatmaps being summed and then further normalized to ensure all values fall within the range of 0–1, thereby enhancing the visual effectiveness of the heatmap. For each pixel in the composite heatmap, a one-dimensional flattening process was conducted, grouping pixels according to their original encoding order (details available in Table A.2), allowing for quantifiable analysis of each region's features. For each feature group, we calculated several metrics, i.e., Sum, Average, and Max, while also tallying the number of pixels exceeding various thresholds (0.8, 0.6, and 0.4) within that region, as detailed in Appendix B (SUMO1&2_CHFA_Statistics). These metrics provide corresponding importance indicators for each feature, enabling us to closely observe the distribution of the model's attention across various areas.

However, in datasets with high variability or sparse data, analyzing only the training set or testing set may overlook some representative features, failing to comprehensively capture the differences within the data. Combining the heatmaps from the training as well as testing sets can effectively mitigate biases introduced by data variability while also increasing the sample size, thereby enhancing the statistical stability of the analysis and the reliability of the results. This combined analysis is particularly advantageous for evaluating feature importance, as a larger sample size aids in more accurately calculating the average weights of features, providing a more holistic reflection of the model's dependence on these features.

We performed overlay and normalization operations on the Grad-CAM heatmaps generated by the model, which were subsequently used to analyze the importance of each feature. The following mathematical formula can be employed to describe this process:

Heatmap Generation: Each input image I_i is processed through the convolutional layers of the model to produce a Grad-CAM heatmap H_i, representing the model's attention to distinct regions within the image.

$$H_i = \text{GradCAM}(I_i, \text{model}) \tag{6}$$

Collection and Overlay of Heatmaps: Let us assume we have N training samples and M testing samples. The set of heatmaps for the training samples is represented as $\{H_1^{\text{train}}, H_2^{\text{train}}, \ldots, H_N^{\text{train}}\}$, while the set of heatmaps for the testing samples is represented as $\{H_1^{\text{test}}, H_2^{\text{test}}, \ldots, H_M^{\text{test}}\}$.

These heatmaps are then overlaid to form a comprehensive heatmap H_{combined}:

$$H_{\text{combined}} = \sum_{i=1}^{N} H_i^{\text{train}} + \sum_{j=1}^{M} H_j^{\text{test}} \tag{7}$$

Filling NaN Values: If the heatmap contains NaN values, we will set these values to 0:

$$H_{\text{combined}} = \text{nan\_to\_num}(H_{\text{combined}}, \text{nan} = 0) \tag{8}$$

Normalization: Finally, the comprehensive heatmap undergoes normalization, compressing its value range to [0,1]. This step ensures that the values in the heatmap maintain a consistent relative scale, facilitating visualization and subsequent analysis:

$$H_{\text{combined}} = \frac{H_{\text{combined}}}{\max(H_{\text{combined}})} \tag{9}$$

The aforementioned equations outline the process by which we overlay and normalize the sets of heatmaps from the training and testing images, enabling a comprehensive analysis of the importance of features in the classification task.

3. Results and discussion

3.1. Comparison of model architecture performance

We compared the performance of MLP, 1D-CNN, and 2D-CNN (SUMO-LMNet) in predicting SUMO1 and SUMO2 sites, as illustrated in Figs. 4 and 5. In the prediction of SUMO1 sites, the 2D-CNN for SUMO1 achieved the best result, with an accuracy of 0.875. In contrast, the accuracy for the MLP as well as 1D-CNN models for SUMO1 was only 0.792 each. For SUMO2 prediction, the 2D-CNN for SUMO2 also outperformed the other models, achieving an accuracy of 0.801.

The 2D-CNN demonstrated its advantages in feature extraction and generalization for the predictions of SUMO1 and SUMO2, significantly outperforming both MLP and 1D-CNN. This superiority can be attributed to its enhanced feature extraction capabilities compared to MLP and 1D-CNN models. MLP processes the input after flattening it, failing to preserve the spatial structure of the input; thus, its capacity to handle data with local adjacency relationships is limited. Furthermore, while the 1D-CNN processes one-dimensional sequence information and can retain some relationships among adjacent features, its information extraction is confined to a single dimension, making it challenging to capture more complex spatial patterns and dependencies among multidimensional features.

In contrast, the 2D-CNN utilizes two-dimensional convolutional kernels, allowing it to learn adjacent features simultaneously in the horizontal and vertical directions. This capability is crucial for predicting SUMO sites, as features may be interdependent within specific regions. Furthermore, the 2D-CNN typically incorporates a multi-layer convolutional structure that can extract features at various scales, ranging from low-level edge features to high-level composite features, further enhancing the model's performance. In comparison, the feature extraction capabilities of MLP and 1D-CNN models are weaker, potentially hindering their ability to fully learn the complex dependencies between SUMO sites. Thus, the multidimensional convolutional properties and hierarchical feature extraction methods of the 2D-CNN provide it with higher adaptability and generalization ability, resulting in higher accuracy and stability in SUMO site prediction tasks.

3.2. Comparison of 2D-CNN performance

3.2.1. Model performance of SUMO1

Table 5 presents the performance of the SUMO1 model on the test dataset, along with the training and validation accuracies during the

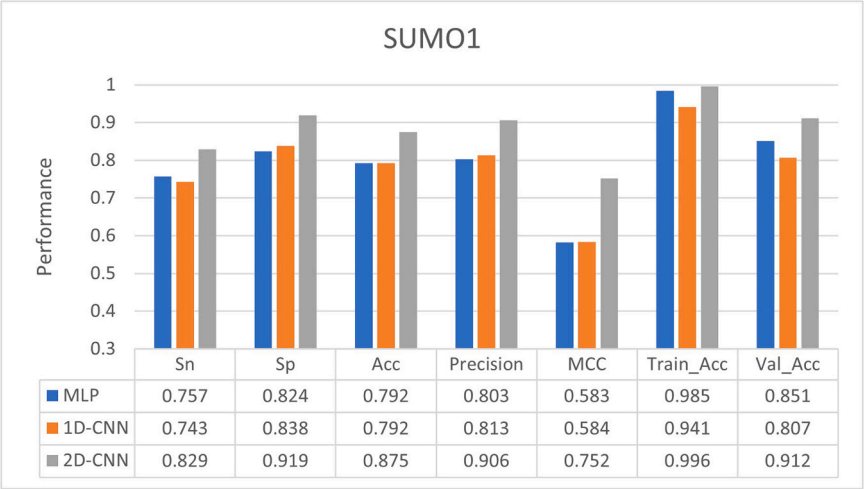

Fig. 4. Performance comparison of SUMO1 model architectures. The performance of three model architectures (MLP, 1D-CNN, and 2D-CNN) in predicting SUMO1 modification sites is evaluated across multiple metrics, including sensitivity (Sn), specificity (Sp), accuracy (Acc), precision, Matthews correlation coefficient (MCC), training accuracy (Train_Acc), and validation accuracy (Val_Acc). The 2D-CNN model achieves the highest overall performance, with 0.875 accuracy, 0.906 precision, and 0.752 MCC, outperforming both MLP and 1D-CNN. The performance of 2D-CNN is attributed to its ability to capture spatial dependencies and hierarchical feature representations, which are essential for improved model generalization and predictive accuracy in SUMOylation site identification.

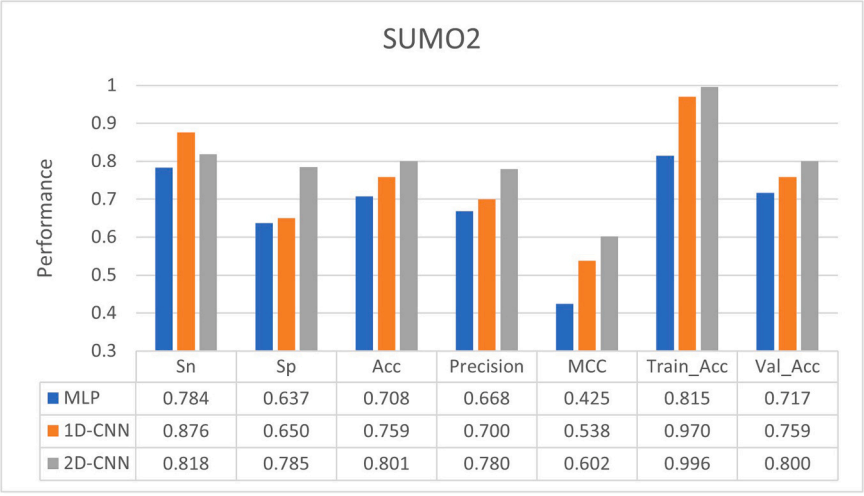

Fig. 5. Performance comparison of SUMO2 model architectures. The performance of three model architectures (MLP, 1D-CNN, and 2D-CNN) in predicting SUMO2 modification sites is evaluated across multiple metrics, including sensitivity (Sn), specificity (Sp), accuracy (Acc), precision, Matthews correlation coefficient (MCC), training accuracy (Train_Acc), and validation accuracy (Val_Acc). The 2D-CNN model achieves the highest accuracy (0.801) and MCC (0.602), outperforming both MLP and 1D-CNN. Compared to SUMO1, the larger dataset for SUMO2 requires a deeper network architecture to effectively capture complex feature dependencies. These results highlight the advantage of 2D-CNN in feature extraction and classification, reinforcing the importance of spatial feature encoding for SUMOylation site prediction.

Table 5
Performance of the Predictive Model on SUMO1.

Method	Sn	Sp	Acc	Precision	MCC	Train_Acc	Val_Acc
MobileNetV2	0.700	0.676	0.688	0.671	0.376	0.766	0.675
Xception	0.757	0.527	0.639	0.602	0.291	0.537	0.526
VGG16	0.643	0.676	0.660	0.652	0.319	0.657	0.710
ResNet152V2	0.671	0.595	0.632	0.610	0.267	0.666	0.614
InceptionResNetV2	0.857	0.203	0.521	0.504	0.079	0.714	0.597
SUMO-LMNet	0.829	0.919	0.875	0.906	0.752	0.996	0.912

training phase (Train_Acc and Val_Acc). Among the five transfer learning methods, MobileNetV2 achieved the highest training accuracy (Train_Acc = 0.766), but this advantage did not carry over to the validation dataset, where VGG16 demonstrated the highest validation accuracy (Val_Acc = 0.710). This indicates that MobileNetV2 exhibits some limitations in generalization capability.

The SUMO-LMNet (2D-CNN) model performed prominently across all metrics, with training and validation accuracies each exceeding 0.8 and the values for both phases being quite close to each other. This suggests that the model possesses good generalization ability, effectively

Table 6
Performance of the Predictive Model on SUMO2.

Method	Sn	Sp	Acc	Precision	MCC	Train_Acc	Val_Acc
MobileNetV2	0.686	0.553	0.617	0.588	0.241	0.654	0.616
Xception	0.385	0.750	0.574	0.590	0.146	0.546	0.549
VGG16	0.546	0.660	0.605	0.600	0.208	0.642	0.599
ResNet152V2	0.429	0.739	0.590	0.605	0.178	0.557	0.563
InceptionResNetV2	0.547	0.456	0.500	0.484	0.003	0.495	0.507
SUMO-LMNet	0.818	0.785	0.801	0.780	0.602	0. 996	0.800

avoiding overfitting. On the independent test dataset, SUMO-LMNet achieved a Matthews Correlation Coefficient of 0.752, significantly outperforming MobileNetV2, which had an MCC of only 0.376.

Furthermore, the CNN model achieved the best performance across all evaluation metrics, with sensitivity (Sn) = 0.829, specificity (Sp) = 0.919, accuracy (Acc) = 0.875, and precision (Precision) = 0.906. This indicates its strong ability to distinguish between positive and negative samples, as well as overall prediction accuracy. In contrast, the performances of other pretrained models were relatively weaker. For instance, although InceptionResNetV2 had the highest sensitivity (Sn = 0.857), its specificity (Sp = 0.203) was markedly low, reflecting inadequate ability to differentiate negative samples that are non-SUMOylation sites. Xception and ResNet152V2 had MCC values of 0.291 and 0.267, respectively, indicating a lack of balance in prediction results and insufficient coordination between positive and negative samples.

These results highlight the noteworthy advantages of the self-constructed 2D-CNN model, SUMO-LMNet, in predicting SUMOylation sites. Its excellent performance in sensitivity, specificity, accuracy, and precision indicates that this model's shallower convolutional layers can capture key features in the data more accurately compared to the other five transfer learning methods. Therefore, the self-built CNN architecture proves to be more suitable compared to pretrained models.

3.2.2. Model performance of SUMO2

Table 6 displays the performance of various models in predicting SUMO2. The self-constructed 2D-CNN model, SUMO-LMNet, demonstrated superiority across multiple evaluation metrics, achieving a sensitivity (Sn) = 0.818, accuracy (Acc) = 0.801, precision (Precision) = 0.780, and Matthews Correlation Coefficient (MCC) = 0.602, which outperformed the other five transfer learning models.

MobileNetV2 performed reasonably well in terms of sensitivity (Sn = 0.686); however, it exhibited lower specificity (Sp = 0.553) and accuracy (Acc = 0.617), with an MCC of only 0.241. The Xception model had a relatively high specificity (Sp = 0.750) but showed a considerable deficiency in sensitivity (Sn = 0.385) and MCC (0.146), indicating inadequate predictive capability for positive samples. VGG16 recorded an MCC of 0.208, reflecting poor overall prediction balance, while ResNet152V2 had an even lower MCC of 0.178. InceptionResNetV2 exhibited the weakest performance, with an MCC of just 0.003 and a specificity (Sp = 0.456) significantly lower than that of the other models, indicating an inability to effectively differentiate between positive and negative samples and overall weak predictive power.

The SUMO-LMNet model achieved the best results in predicting SUMO2. It should be noted that the network architectures used for SUMO1 and SUMO2 are slightly different, as listed in Tables 3 and 4. The SUMO2 model incorporates more convolutional layers, which may be attributed to a larger number of training samples compared to SUMO1. Consequently, SUMO2 requires a deeper network architecture to capture more complex features within the data. While some transfer learning models demonstrated better performance in identifying negative samples with higher specificity than CNN, they still struggled to match the performance of the self-constructed CNN in balancing positive and negative sample predictions, generalization capability, and overall predictive accuracy.

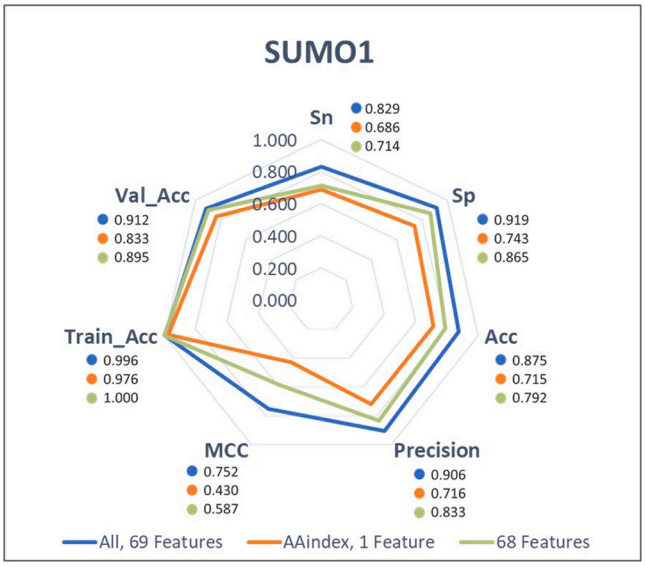

Fig. 6. Impact of feature combinations on SUMO1 prediction performance. This radar chart illustrates the effect of different feature sets on the predictive performance of the SUMO-LMNet model for SUMO1 modification sites. Three feature configurations are evaluated: (i) all 69 features (blue), (ii) only the AAIndex feature (orange), and (iii) 68 features excluding AAIndex (green). The results indicate that the full feature set (69 features) achieves the highest accuracy (0.875), precision (0.906), and MCC (0.752), highlighting the advantage of incorporating diverse sequence descriptors. In contrast, when only the AAIndex feature is used, performance declines significantly (accuracy: 0.715, MCC: 0.430), demonstrating that a single biochemical index is insufficient for accurate prediction. Meanwhile, when AAIndex is removed, the model maintains relatively high accuracy (0.792), suggesting that SUMO1 prediction is more dependent on the other 68 sequence features than on AAIndex alone.

3.3. Features and predictive performance

In the SUMO1 prediction model, a total of 28,390 vectors were used to describe SUMOylation sites, encompassing 69 features, of which the AAindex feature dominated with 16,461 vectors, accounting for 58 % of the total. To explore the impact of various feature combinations on model performance, we tested the prediction results using only the AAindex feature and only the other 68 features separately. As shown in Fig. 6, the CNN model utilizing all 69 features achieved the best performance, with an accuracy (Acc) of 0.875, a sensitivity (Sn) of 0.829, and a specificity (Sp) of 0.919. This indicates that employing the comprehensive set of features yields the best results in predicting SUMO1. When we limited the model to only the AAindex feature, the accuracy dropped to 0.715, suggesting that while AAindex is important, the absence of the other features significantly impacts the model's accuracy. Specifically, the sensitivity fell to 0.686, and the specificity decreased to 0.743 when the other 68 features were excluded. This highlights the crucial role that these additional features play in the SUMO1 prediction task, contributing to more accurate classification outcomes. In contrast, removing the AAindex feature did not significantly affect the model's performance, as it still achieved an accuracy of

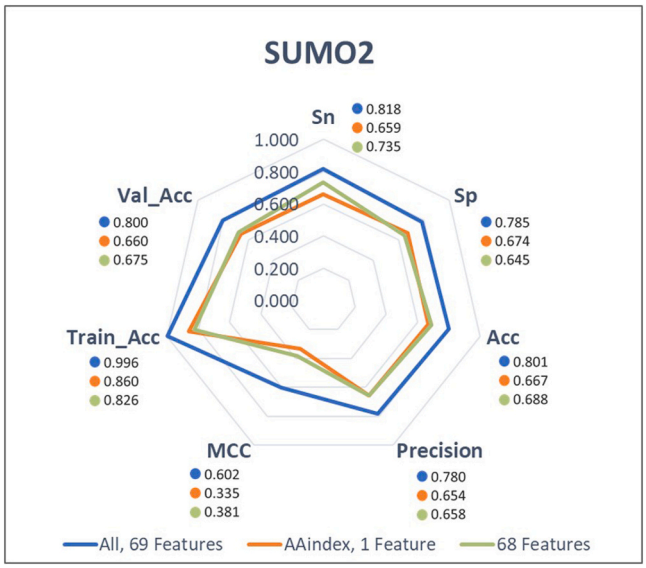

Fig. 7. Impact of feature combinations on SUMO2 prediction performance. This radar chart illustrates the effect of different feature sets on the predictive performance of the SUMO-LMNet model for SUMO2 modification sites. Three feature configurations are compared: (i) all 69 features (blue), (ii) only the AAIndex feature (orange), and (iii) 68 features excluding AAIndex (green). The results indicate that the full feature set (69 features) achieves the highest accuracy (0.801), sensitivity (0.818), and specificity (0.785), highlighting the importance of integrating diverse sequence descriptors. When only AAIndex is used, performance declines significantly (accuracy: 0.667), demonstrating that AAIndex alone is insufficient for robust prediction. Conversely, when AAIndex is removed, accuracy drops to 0.688, suggesting that SUMO2 prediction relies more heavily on AAIndex compared to SUMO1. These findings underscore the distinct feature dependencies between SUMO1 and SUMO2.

0.792. This indicates that the SUMO1 prediction heavily relies on the other 68 features. The inclusion of these features provided more information for classification, enhancing the model's predictive capability for SUMO1 sites. Thus, although AAindex holds some value in characterizing SUMO1 sites, the other 68 features have a more substantial influence on the model's performance.

In contrast, in the prediction of SUMO2, Fig. 7 illustrates the impact of various feature combinations on model performance. Compared to the results for SUMO1, the role of the AAindex feature in SUMO2 predictions is significantly more pronounced. When using all 69 features, the CNN model achieved an accuracy of 0.801, a sensitivity of 0.818, and a specificity of 0.785, indicating that a combination of multiple features is also effective in describing information relevant to SUMO2 sites. However, when only the AAindex feature was used, the model maintained an accuracy of 0.667, but when AAindex was removed, the accuracy dropped from 0.801 to 0.688. This demonstrates that AAindex plays a critical role in SUMO2 prediction and significantly affects the model's performance. In contrast, the SUMO1 prediction model showed a lower dependency on AAindex; its accuracy decreased from 0.875 to 0.792 upon removing AAindex. Thus, while AAindex serves as a key feature for SUMO2 predictions, SUMO1 relies more heavily on the remaining 68 features. This highlights the differences in feature dependencies between different SUMO paralogues, indicating that distinct characteristics may influence model performance in predicting SUMOylation sites for each paralogue.

Overall, in the SUMO1 and SUMO2 prediction tasks, the custom-built CNN model demonstrates excellent performance, particularly when leveraging all available features. The receiver operating characteristic (ROC) curves for SUMO1 and SUMO2 are shown in Appendix Fig. A.2, with the SUMO1 model achieving an Area Under the Curve (AUC) of 0.93 and the SUMO2 model attaining an AUC of 0.87. This indicates that our model architecture effectively captures the key characteristics within the data. In the prediction task for SUMO1, the other 68 features, in addition to AAindex, played a crucial role in enhancing model performance. Conversely, in the prediction of SUMO2, the larger dataset highlights the indispensable role of AAindex, with its significant contribution to improving model effectiveness. This underscores the importance of feature selection in optimizing predictive accuracy for various SUMO paralogues.

3.4. Analysis of feature importance

3.4.1. Visualization of feature importance

In the discussion of features for SUMO1 and SUMO2, the computed results of CHFA (as presented in Appendices C and D) are visualized in the feature attention heatmap shown in Fig. 8(a), illustrating the varying degrees of attention that various models pay to each feature. The pixel positions in these heatmaps correspond to features arranged in the order specified in Table A.2, allowing for an intuitive demonstration of each feature's contribution to the predictions. For instance, the prediction results for SUMO1 indicate that its significant features are primarily concentrated outside the AAindex region, while the results for SUMO2 reveal a stronger dependency on AAindex, with certain binary regions also exhibiting high importance. This suggests that, in predicting SUMO2 sites, physicochemical properties such as AAindex play a more crucial role in the model's decision-making process. There are noticeable differences in the importance assessments of the 68 different features and AAindex between SUMO1 and SUMO2, potentially attributable to variations in data volume. Furthermore, the SUMO site heatmaps during the training phase and the independent testing phase for both SUMO1 and SUMO2 are displayed in Fig. A.3. The aggregated average results in Fig. 8(a) reveal that the larger data volume for SUMO2 necessitates a more diverse array of features. Specifically, the heatmap for SUMO2 contains more light blue areas and fewer dark blue areas, indicating a higher dependence on features such as binary and AAindex compared to SUMO1. Conversely, SUMO1 demonstrates a relatively lower demand for features, exhibiting more dark blue areas, which reflects its higher reliance on a smaller number of features. Thus, from the perspective of data volume and feature demand, the vast scale of the SUMO2 dataset necessitates a richer combination of features to enhance the model's predictive performance, contrasting with the findings for SUMO1.

3.4.2. Ranking of feature importance

Through Eqs. (6) to (9), we obtained the comprehensive heatmap for training as well as testing samples (Fig. 8(a), with detailed values provided in Appendices C and D), allowing us to analyze the results of the feature regions and discuss their distribution of importance. These results illustrate the model's points of focus across various regions and the significance of various features, as detailed in Appendix B.

In this study, we compared the importance of distinct features in the SUMO1 and SUMO2 datasets, particularly focusing on the performance of each feature under the metric Count > 0.6 (which calculates the number of instances exceeding the 0.6 threshold) as listed in Table 7. The results indicate that high values in the SUMO2 dataset for Count > 0.6 are concentrated in features such as CKSAAP, KSCTriad, and EAAC. The CKSAAP (Composition of k-Spaced Amino Acid Pairs) feature encodes the frequency of amino acid pairs at various intervals, capturing interactions between non-adjacent amino acids. The KSCTriad (k-Spaced Conjoint Triad) feature reflects the distant relationships among multiple amino acids through combinations of triads spaced by k intervals, while the Enhanced Amino Acid Composition (EAAC) feature computes the amino acid composition ratio over a sliding window, emphasizing local sequence characteristics. These features provide the model with a multilayered representation of sequence information.

Specifically, the CKSAAP Type 1 and Type 2 features exhibited remarkable significance in SUMO2, with 1582 and 1580 values

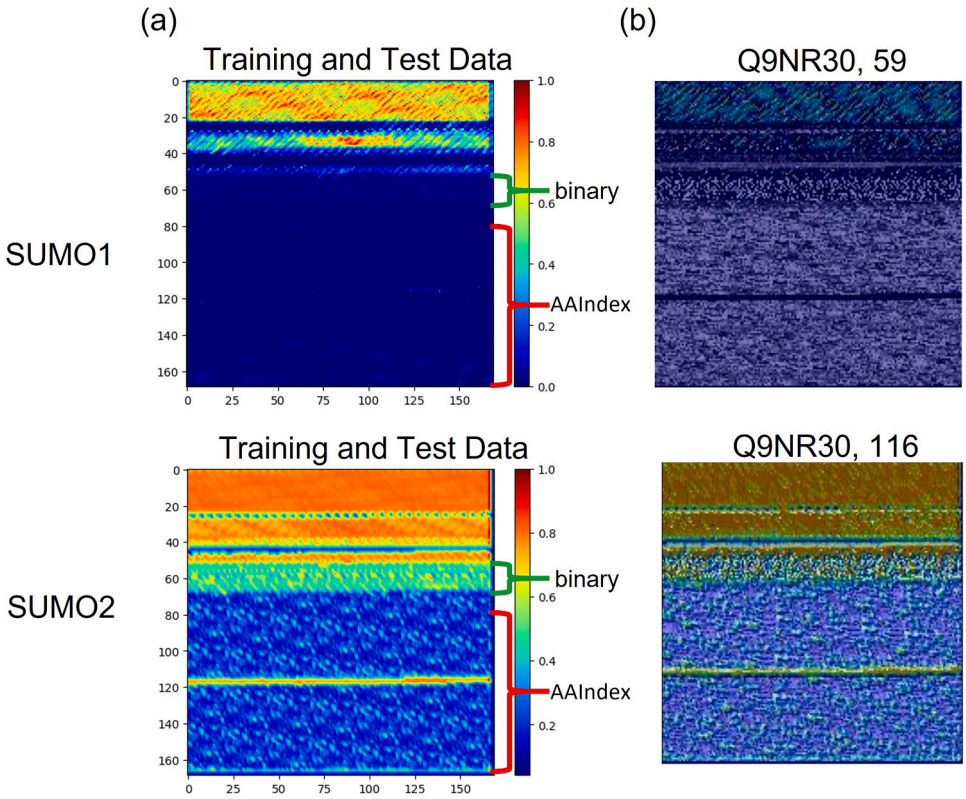

Fig. 8. Feature importance analysis via heatmap visualization. This heatmap presents the CHFA results, depicting the SUMO-LMNet model's attention distribution across different feature sets for SUMO1 and SUMO2 predictions. The pixel positions correspond to features arranged in Table A.2, enabling direct interpretation of their contributions to model predictions. The color gradient represents feature importance, with red indicating highly influential regions and blue denoting lower importance. The results reveal distinct feature dependencies between SUMO1 and SUMO2. SUMO2 predictions exhibit a stronger reliance on AAIndex and certain binary features, whereas SUMO1 predictions distribute importance more evenly across diverse feature sets.

Table 7
Feature Rankings by CountAbove0.6 and Sum.

	Ranking by Count > 0.6					Ranking by Value Sum			
	Rank	**No.**	**Feature**	**Bits**	**Count > 0.6**	**No.**	**Feature**	**Bits**	**Value Sum**
SUMO1	**1**	3	CKSAAP type 1	1600	1200	3	CKSAAP type 1	1600	1076
	2	4	CKSAAP type 2	1600	1174	4	CKSAAP type 2	1600	1073
	3	11	KSCTriad	1372	291	11	KSCTriad	1372	614
	4	2	EAAC	540	122	2	EAAC	540	218
	5	5	DPC type 1	400	39	5	DPC type 1	400	94
	6	9	CTDD	195	17	10	CTriad	343	90
	7	10	CTriad	343	5	69	AAIndex	16461	87
	8	12	ASDC	400	1	51	binary	620	67
	9	34	PseKRAAC type 3 A	4	1	12	ASDC	400	61
	10	69	AAIndex	16461	0	9	CTDD	195	33
SUMO2	**1**	3	CKSAAP type 1	1600	1582	69	AAIndex	16461	3982
	2	4	CKSAAP type 2	1600	1580	3	CKSAAP type 1	1600	1273
	3	11	KSCTriad	1372	1356	4	CKSAAP type 2	1600	1270
	4	2	EAAC	540	534	11	KSCTriad	1372	1048
	5	51	binary	620	465	2	EAAC	540	417
	6	69	AAIndex	16461	461	51	binary	620	414
	7	5	DPC type 1	400	384	5	DPC type 1	400	300
	8	10	CTriad	343	339	12	ASDC	400	266
	9	12	ASDC	400	333	10	CTriad	343	261
	10	9	CTDD	195	182	6	DPC type 2	400	215

No: The feature sequence numbers are displayed in Table A.2; Feature: Feature Name; Bits: Number of Features; Count > 0.6: Quantity of Features with Values Higher than 0.6; Value Sum: Total of Feature Values.

exceeding 0.6, respectively, out of 1600 bits. Concurrently, the counts for high values in the KSCTriad and EAAC features reached 1356 and 534, respectively. Despite the AAIndex feature having the largest representation with 16,461 bits, it contained only 461 high-value instances; nevertheless, it maintained the top rank in overall importance (sum), demonstrating its contribution as a low-density but high-value feature. While features with a larger number of bits may more readily be selected as significant given their sheer quantity, the results indicate that even features with only a few hundred bits were prominently identified as important. This observation underscores the efficacy of the

convolutional neural network constructed in this study to capture essential features while mitigating the influence of biases stemming from the bit counts of the features.

In contrast, while the top four features for the Count > 0.6 metric in SUMO1 are identical to those in SUMO2, their high-value counts are relatively lower, with CKSAAP Type 1 and Type 2 registering counts of 1200 and 1174, respectively. Other features, such as KSCTriad and EAAC, also exhibit significantly lower high-value counts in SUMO1, at 291 and 122, respectively. This result indicates a more dispersed focus on features in SUMO1, suggesting that the model requires fewer significant features for its predictions in this dataset. Furthermore, several features that are prominent in SUMO2, such as ASDC and CTriad, show exceedingly low high-value counts in SUMO1, further underscoring the limited feature demands of the model within the SUMO1 dataset. Although none of the AAIndex vectors in SUMO1 exceeded the threshold of 0.6, its overall sum still ranked seventh, demonstrating its substantial contribution to the collective feature importance. In addition, the cumulative representation of binary features denoting amino acid combinations in SUMO1 is comparatively lower, reflecting the variability in feature significance across different datasets based on their dimensionality and representation methods.

CKSAAP and KSCTriad effectively capture the medium- to long-range interactions between amino acids, while EAAC provides critical insights into local amino acid composition. AAIndex, by contrast, encompasses a comprehensive array of physicochemical properties. These features are particularly prominent in the SUMO2 dataset, likely attributable to the structural and functional demands of SUMO2, which necessitates a multifaceted understanding of non-local interactions, local composition, and physicochemical characteristics.

3.5. Computational prediction of SUMO1 modification at RanGAP1 K524: a case study

RanGAP1 (Ran GTPase activating protein 1) is a crucial regulator of nucleocytoplasmic transport and mitotic spindle formation. It was among the first identified SUMOylation substrates, with K524 serving as the primary SUMO1 modification site in human RanGAP1 (P46060, RAGP1_HUMAN, included in the test dataset) [38]. SUMOylation at this site is essential for RanGAP1 localization to the nuclear pore complex (NPC), where it interacts with RanBP2/Nup358, ensuring proper Ran GTPase cycle function [39]. Beyond nuclear transport, SUMOylation at K524 plays roles in cell cycle regulation and stress responses. Studies indicate that SUMOylated RanGAP1 modulates interactions with importins/exportins, influencing nuclear import dynamics [40]. Furthermore, under stress conditions, SUMOylation affects RanGAP1 stability and degradation, linking it to cellular stress adaptation and tumorigenesis [40,41].

Our model successfully predicted K524 as a SUMO1 modification site with a confidence score of 0.70, demonstrating its ability to recognize biologically validated SUMOylation sites. This case study highlights the application of computational SUMOylation site prediction in assisting experimental design and prioritizing candidate sites for further investigation.

4. Conclusion

This study presents SUMO-LMNet, a model designed to predict SUMOylation sites specific to SUMO1 and SUMO2. The potential structural and functional differences between SUMO1 and SUMO2 suggest that accurately identifying their modification sites through predictive modeling may aid in exploring possible links between SUMOylation abnormalities and pathological mechanisms. Most previous studies primarily focus on broad-spectrum SUMOylation site prediction without distinguishing between SUMO1- and SUMO2-specific modifications and often lack comprehensive feature extraction and sufficient model interpretability.

The SUMO-LMNet model, grounded in the concept of a "Lossless Mapping Network" (LM-Net), emphasizes distortion-free mapping of features, preserving their original dimensions as well as complete information integrity. This design prevents feature data loss, enabling the model to deliver nuanced interpretations for each feature site. By mapping high-dimensional features into a 2D feature map, SUMO-LMNet not only captures spatial dependencies but also reveals the specific biological significance of each feature point, making the distinct differences between SUMO1 and SUMO2 more apparent.

Within this framework, we employed 69 distinct sequence features, using a sliding window to capture SUMOylation sites and their surrounding amino acid contexts. These features span multiple dimensions of protein sequences, including amino acid physicochemical properties, local and long-range interactions, and compositional details. Transformed into high-dimensional vectors, these features are structured into a 2D feature map within SUMO-LMNet. The LM-Net's 2D convolutional neural network (2D-CNN) effectively captures adjacent features in horizontal as well as vertical directions, preserving spatial dependencies between features, and achieving a prediction accuracy exceeding 80 % for SUMOylation sites. Unlike traditional one-dimensional CNN models, the 2D-CNN is more suited for managing high-dimensional feature spaces, capturing distinct feature dependencies for SUMO1 and SUMO2 with higher precision.

To validate the predictive performance of SUMO-LMNet, this study compared it with five commonly used pretrained models, including MobileNetV2, VGG16, Xception, ResNet152V2, and InceptionResNetV2. The results demonstrated that SUMO-LMNet outperformed these models in sensitivity, specificity, and accuracy. For SUMO1 predictions, SUMO-LMNet achieved sensitivity, specificity, and precision all exceeding 0.8, while for SUMO2 predictions, the accuracy reached 0.801.

To deepen our understanding of SUMO-LMNet's interpretability, we developed the Combined Heatmap Feature Analysis method, integrating Grad-CAM for feature importance analysis. Through the heatmaps generated by CHFA and Grad-CAM, the model visualizes the impact of each feature on prediction outcomes. Analysis of SUMO2 highlights a strong dependence on physicochemical properties (AAIndex) and long-range amino acid interaction features (such as CKSAAP and KSCTriad). In contrast, SUMO1 shows a higher reliance on local amino acid composition. These differences underscore the necessity of constructing specialized models tailored to each SUMO paralogue.

However, there remains room for improvement in CHFA's current design. Mapping CHFA's feature dependencies to specific amino acid positions could enhance feature interpretability. Yet, owing to significant differences in bit counts across features, direct mapping may introduce redundant empty spaces in spatial distributions, while dimensionality reduction could compromise feature accuracy. Consequently, the challenges in CHFA's feature interpretation require further investigation. In addition, the LM-Net architecture not only effectively differentiates between SUMO1 and SUMO2 but could also be extended for comparative analyses of the same SUMO paralogue across different species or applied to other post-translational modifications, offering a new avenue for subtype differentiation analysis.

Although this study proposes SUMO-LMNet, which achieves strong accuracy and interpretability in predicting SUMOylation sites for SUMO1 and SUMO2, the method still has notable limitations, particularly in predicting SUMO3–5. The limited availability of experimentally validated data for SUMO3–5 hinders the effective training of data-driven deep learning models, thereby limiting accurate prediction of modification sites for these paralogues. To address this challenge, several strategies can be considered. First, transfer learning could be applied to extend models trained on SUMO1 and SUMO2 to SUMO3–5, potentially improving performance in data-scarce scenarios when combined with few-shot learning or meta-learning approaches. Second, integrating sequence and structural information—using advanced protein structure prediction tools such as AlphaFold—and transforming the entire SUMOylation complex (SUMO, Ubc9, E3, and the substrate) into a

multiscale graph representation could provide a more comprehensive view of interactions at multiple hierarchical levels, which can be leveraged through GNN-based learning frameworks. If the challenge of accurately predicting SUMO3–5 sites can be overcome, predictive tools could serve as valuable aids in supporting disease mechanism research, biomarker discovery, and the development of targeted therapies.

CRediT authorship contribution statement

Ho Cheng-Hsun: Visualization, Investigation. **Chen Chi-Wei:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Huang Lan-Ying:** Software. **Chu Yen-Wei:** Resources, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Science and Technology Council, Taiwan, under grant number 111–2222-E-214–001, 111–2314-B-214–003-MY3, 111–2221-E-005–073-MY3, 113–2321-B-006–014, 112–2634-F-005–002 and 111–2423-H-006–002-MY3; NCHU-CCH 11307 from National Chung Hsing University and Changhua Christian Hospital.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.03.005.

Data availability

Data will be made available on request.

References

[1] Wang S, Osgood AO, Chatterjee A. Uncovering post-translational modification-associated protein–protein interactions. Curr Opin Struct Biol 2022;74:102352.
[2] Celen AB, Sahin U. Sumoylation on its 25th anniversary: mechanisms, pathology, and emerging concepts. FEBS J 2020;287(15):3110–40.
[3] Princz A, Tavernarakis N. SUMOylation in neurodegenerative diseases. Gerontology 2020;66(2):122–30.
[4] Yang Y, He Y, Wang X, Liang Z, He G, Zhang P, et al. Protein SUMOylation modification and its associations with disease. Open Biol 2017;7(10):170167.
[5] Du L, Liu W, Rosen ST. Targeting SUMOylation in cancer. Curr Opin Oncol 2021;33(5):520–5.
[6] Huang C-H, Yang T-T, Lin K-I. Mechanisms and functions of SUMOylation in health and disease: a review focusing on immune cells. J Biomed Sci 2024;31(1):16.
[7] Wu Q, Jiang Y, You C. The SUMO components in rheumatoid arthritis. Rheumatology 2022;61(12):4619–30.
[8] Hua D, Wu X. Small-molecule inhibitors targeting small ubiquitin-like modifier pathway for the treatment of cancers and other diseases. Eur J Med Chem 2022;233:114227.
[9] Bouchard D, Wang W, Yang W-C, He S, Garcia A, Matunis MJ. SUMO paralogue–specific functions revealed through systematic analysis of human knockout cell lines and gene expression data. Mol Biol Cell 2021;32(19):1849–66.
[10] Zhang X-D, Goeres J, Zhang H, Yen TJ, Porter AC, Matunis MJ. SUMO-2/3 modification and binding regulate the association of CENP-E with kinetochores and progression through mitosis. Mol Cell 2008;29(6):729–41.
[11] Bhachoo JS, Garvin AJ. SUMO and the DNA damage response. Biochem Soc Trans 2024;52(2):773.
[12] Ramazi S, Dadzadi M, Darvazi M, Seddigh N, Allahverdi A. Protein modification in neurodegenerative diseases. MedComm 2024;5(8):e674.
[13] Chen L-C, Hsieh Y-L, Tan GY, Kuo T-Y, Chou Y-C, Hsu P-H, et al. Differential effects of SUMO1 and SUMO2 on circadian protein PER2 stability and function. Sci Rep 2021;11(1):14431.
[14] Suk TR, Nguyen TT, Fisk ZA, Mitkovski M, Geertsma HM, Parmasad J-LA, et al. Characterizing the differential distribution and targets of Sumo1 and Sumo2 in the mouse brain. iScience 2023;26(4).
[15] Orsini F., Pascente R., Martucci A., Palacino S., Fraser P., Arancio O., et al. SUMO2 rescues neuronal and glial cells from the toxicity of P301L mutant Tau. Frontiers in Cellular Neuroscience;18:1437995.
[16] Chen K, Shi R, Huang P, Guo S, Hu J, Han B, et al. Ginkgolic acid inhibited Tau phosphorylation and improved cognitive ability through the SUMO-1/GSK3β pathway in Aβ-Induced Alzheimer's disease model rats. J Funct Foods 2024;116:106183.
[17] Audagnotto M, Dal Peraro M. Protein post-translational modifications: in silico prediction tools and molecular modeling. Comput Struct Biotechnol J 2017;15:307–19.
[18] Khan S, AlQahtani SA, Noor S, Ahmad N. PSSM-Sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. BMC Bioinforma 2024;25(1):284.
[19] Ramazi S, Zahiri J, Arab S, Parandian Y. Computational prediction of proteins sumoylation: a review on the methods and databases. J Nanomed Res 2016;3(5).
[20] Meng L, Chan W-S, Huang L, Liu L, Chen X, Zhang W, et al. Mini-review: recent advances in post-translational modification site prediction based on deep learning. Comput Struct Biotechnol J 2022;20:3522–32.
[21] Green J., Dmochowski G., Golshani A. Prediction of protein sumoylation sites via parallel cascade identification. 29th Conference of the Canadian Medical and Biological Engineering Society. 6. 2006.
[22] Zhao Q, Xie Y, Zheng Y, Jiang S, Liu W, Mu W, et al. GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. Nucleic Acids Res 2014;42(W1). W325-W30.
[23] Xue Y, Zhou F, Fu C, Xu Y, Yao X. SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res 2006;34(_2). W254-W7.
[24] Beauclair G, Bridier-Nahmias A, Zagury J-F, Saïb A, Zamborlini A. JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMs. Bioinformatics 2015;31(21):3483–91.
[25] Chang J, Sitzmann V, Dun X, Heidrich W, Wetzstein G. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Sci Rep 2018;8(1):1–10.
[26] Qian Y, Ye S, Zhang Y, Zhang J. SUMO-Forest: a cascade forest based method for the prediction of SUMOylation sites on imbalanced data. Gene 2020;741:144536.
[27] He F., Wang R., Gao Y., Wang D., Yu Y., Xu D., et al. Protein ubiquitylation and sumoylation site prediction based on ensemble and transfer learning. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2019:117-23.
[28] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2007;36(_1). D202-D5.
[29] Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018;34(14):2499–502.
[30] Chen Z, Liu X, Zhao P, Li C, Wang Y, Li F, et al. iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. Nucleic Acids Res 2022;50(W1). W434-W47.
[31] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision. 2017:618-626.
[32] UniProt: the universal protein knowledgebase in 2023. Nucleic acids research 2023;51(D1):D523-D31.
[33] Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:4510-20.
[34] Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:1251-8.
[35] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 2014.
[36] He K., Zhang X., Ren S., Sun J. Identity mappings in deep residual networks. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer; 2016:630-645.
[37] Szegedy C., Ioffe S., Vanhoucke V., Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI conference on artificial intelligence. 31. 2017.
[38] Joseph J, Tan S-H, Karpova TS, McNally JG, Dasso M. SUMO-1 targets RanGAP1 to kinetochores and mitotic spindles. J Cell Biol 2002;156(4):595–602.
[39] Matunis MJ, Wu J, Blobel G. SUMO-1 modification and its role in targeting the Ran GTPase-activating protein, RanGAP1, to the nuclear pore complex. J Cell Biol 1998;140(3):499–509.
[40] Zhang F, Yang J, Cheng Y. Impact of RANGAP1 SUMOylation on Smad4 nuclear export by bioinformatic analysis and cell assays. Biomol Biomed 2024;24(6):1620.
[41] Zhu S, Goeres J, Sixt KM, Békés M, Zhang X-D, Salvesen GS, et al. Protection from isopeptidase-mediated deconjugation regulates paralog-selective sumoylation of RanGAP1. Mol Cell 2009;33(5):570–80.