scientific reports

Check for updates

OPEN

Innovative data augmentation strategy for deep learning on biological datasets with limited gene representations focused on chloroplast genomes

Mohammad Ali Abbasi-Vineh[1], Shirin Rouzbahani[1], Kaveh Kavousi[2]✉ & Masoumeh Emadpour[1,3]✉

One key barrier to applying deep learning (DL) to omics and other biological datasets is data scarcity, particularly when each gene or protein is represented by a single sequence. This fundamental challenge is mainly relevant in research involving genetically constrained organisms, organelles, specialized cell types, and biological cycles and pathways. This study introduces a novel data augmentation strategy designed to facilitate the application of DL models to omics datasets. This approach generated a high number of overlapping subsequences with controlled overlaps and shared nucleotide features through a sliding window technique. A hybrid model of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers was applied across augmented datasets comprising genes and proteins from eight microalgae and higher plant chloroplasts. The data augmentation strategy enabled employing DL methods on these datasets and significantly improved the model performance by avoiding common issues such as overfitting and non-representative sequence variations. The current augmentation process is highly adaptable, providing flexibility across different types of biological data repositories. Furthermore, a complementary k-mer-based data augmentation strategy was introduced for unlabeled datasets, enhancing unsupervised analysis. Overall, these innovative strategies provide robust solutions for optimizing model training potential in the study of datasets with limited data availability.

Keywords Machine learning, Small data, Chloroplast genome, Genomics, Transcriptomics, Proteomics

Machine learning (ML) and especially deep learning (DL) have revolutionized the landscape of biological and genetic research, leading to significant breakthroughs in areas such as genomics, transcriptomics, and proteomics[1–4]. By harnessing the power of large datasets, DL/ML can uncover intricate patterns and generate predictions that surpass the capabilities of traditional computational methods. This ability is particularly transformative in fields where the complexity and scale of data often impede conventional analytical approaches[5–7].

A key requirement for training deep learning models is access to substantial amounts of data. The robustness of DL models heavily depends on the volume of data, as larger datasets help prevent overfitting—a common pitfall in machine learning where models become overly specialized to the training data[5–7]. Overfitting leads to poor generalization, meaning the model performs well on training data but fails when exposed to unseen datasets. In biological research, where datasets can be limited in size, the risk of overfitting becomes especially pronounced. Instead of learning genuine patterns, models trained on small datasets tend to memorize irrelevant details, reducing their efficacy[5,6,8].

[1]Department of Agricultural Biotechnology, Tarbiat Modares University (TMU), Tehran 1497713111, Iran. [2]Department of Bioinformatics, Laboratory of Complex Biological Systems and Bioinformatics (CBB), Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran. [3]Plant Physiology, Dahlem Centre of Plant Sciences, Freie Universität Berlin, 14195 Berlin, Germany. ✉email: kkavousi@ut.ac.ir; m.emadpour@modares.ac.ir

Particularly in the case of small, biologically constrained datasets—such as those derived from certain organisms, specific tissues, or cells—traditional DL approaches often struggle to generalize due to the limited number of sequences[9]. This issue is compounded when working with datasets containing or expressing few genes, specifically each representing a unique gene or protein. Such small datasets are common in genomics, especially when working with specialized cell types (e.g., embryonic stem cells, myeloid leukemia cells, quiescent cells), organelles (e.g., mitochondria, chloroplasts), or certain bacterial strains (Ca. *Nasuia deltocephalinicola*, Ca. *Carsonella ruddii*)[10–15]. Furthermore, in some biological systems, most cellular processes or pathways (such as cell cycle, apoptosis, signaling, and metabolic pathways) are represented by only a small subset of genes or proteins[16–19]. With the limited datasets, each representing a unique gene or protein, it is not feasible to effectively train, validate, and test the deep learning model. As a result, a significant gap has remained in the application of DL approaches to small biologically constrained datasets.

Chloroplasts, specialized organelles in plants and microalgae, play a critical role in photosynthesis by converting light energy into chemical energy[20]. Beyond this primary function, chloroplasts participate in various metabolic pathways and have become a promising platform for biotechnological applications, including genetic engineering and the production of high-value biomolecules[21–26]. The chloroplast genome, typically comprising 100 to 200 genes, is vital for photosynthesis and for maintaining the transcriptional and translational machinery essential to the organelle's function[12,20]. The limited number and diversity of unique genes or protein sequences in chloroplast genomes have constrained the application of deep learning techniques to them.

Data augmentation is a critical technique developed to address the limited data availability by artificially expanding the size and diversity of datasets[27]. Various methods for text data (i.e., sequentially continuous and coherent data) augmentation can be broadly categorized into two main types: symbolic and neural techniques. Symbolic techniques include approaches such as synonym replacement, random deletion, and word shuffling. In contrast, neural techniques encompass methods such as generative data augmentation and style modification[28]. Nevertheless, in genetics and molecular biology, especially when dealing with sequence data, these approaches are far less feasible, as even a single nucleotide alteration can significantly affect the functionality of regulatory elements or the biological activities of genes and proteins[29,30]. The unique nature of biological sequences limits the scope for generating variations without compromising the integrity of the data. In recent years, augmentation techniques have also been developed for enhancing deep learning performance on relatively small datasets in genomics[29–32]. However, these approaches could not apply to datasets where each gene or protein is represented by a single sequence.

Therefore, developing novel methodologies to integrate deep learning and machine learning with limited omics data represents a critical area of investigation. The present study aimed to address this gap by proposing an innovative approach to enable the effective application of DL and ML models to chloroplast genomes, as one of the naturally constrained datasets.

Results

Data augmentation approach to enable deep learning analyses

Augmenting nucleotide sequence datasets

An innovative augmentation strategy was implemented to address the challenges inherent to limited gene representation in deep learning applications, generating overlapping subsequences and shared nucleotide motifs to expand a dataset of the sequences without modification of any nucleotides. Indeed, each 300-nucleotide gene sequence was decomposed into overlapping k-mers of 40 nucleotides using a variable overlap range (5–20 nucleotides), with a requirement that each k-mer shared a minimum of 15 consecutive nucleotides with at least one other k-mer. This method generated variable overlaps, ensuring a high degree of data diversity without redundancy. The comprehensive coverage of overlapping subsequences across a single gene sequence has been illustrated in Fig. 1. The figure provided a visual representation, showcasing the structural integrity of the augmented sequences while also reflecting the increased complexity introduced by the overlapping strategy. According to the current configuration of the augmentation method, between 50% and 87.5% of each sequence was designated as invariant, preserving conserved regions that emphasize significant differences among subsequences to support effective model training. Conversely, a portion ranging from 12.5% to 50% at the left, right, or both ends of each sequence was treated as variable, introducing diversity into the subsequences and enriching the training dataset. This approach resulted in the generation of 261 subsequences from each original sequence, representing a substantial augmentation. The level of diversity, along with the controllable conserved regions, can be easily adjusted based on the specific characteristics of different datasets. For more detailed information on the augmentation approach, please refer to Sect. 5.2 of the Materials and Methods.

Applying this approach, the original dataset of 100 sequences— e.g., for the chloroplast of *C. reinhardtii* dataset— was transformed into a larger dataset of 26,100 subsequences (261 subsequences per gene sequence). This approach provided a robust data foundation with high-quality representatives for each original sequence, enabling subsequent model training. Each overlapping subsequence provided a slight variation from the others, which increased the amount of data available for training deep learning models without altering the fundamental information contained in the original sequences. The nucleotide sequence augmentation approach was applied to the sequence datasets from eight different chloroplast genomes from microalgae and higher plants. These diverse datasets were used for further analysis to investigate the efficiency of the approach.

CNN-LSTM model performance on the augmented nucleotide sequences

The results of the CNN-LSTM hybrid model were independently evaluated across multiple genome datasets — *C. reinhardtii*, *C. vulgaris*, *A. thaliana*, *T. aestivum*, *Z. mays*, *G. max*, *N. tabacum*, and *O. sativa*—with and without the data augmentation (Fig. 2). The model performance on non-augmented data showed no accuracy, indicating its inability to make reliable predictions. However, the model performance showed significantly higher accuracy

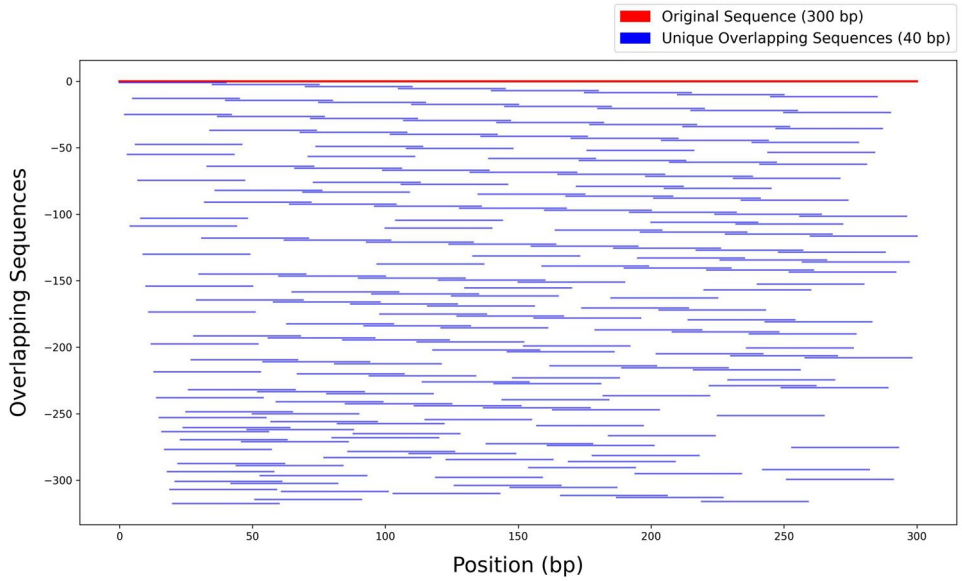

Fig. 1. Graphical visualization of the augmentation process. A single original sequence and its corresponding generated overlapping subsequences were plotted. The red line represents the original 300-bp sequence, while the blue lines indicate each unique overlapping subsequence (40 bp) and their positional relationships. The graph highlights how variable overlaps and common base constraints create a complex yet non-redundant dataset.

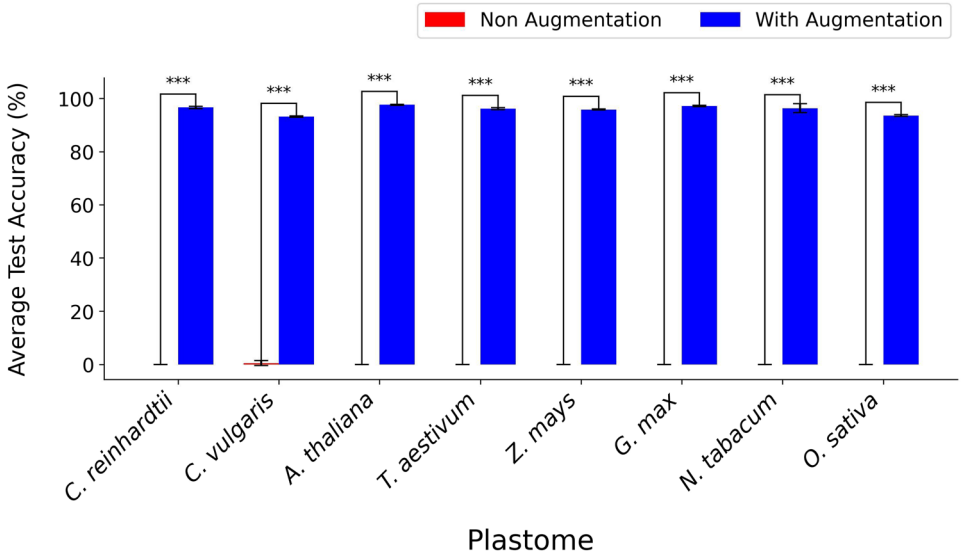

Fig. 2. Average test accuracy across eight plastomes with and without the data augmentation across diverse genomes. The bar chart displays the average test accuracy (%) of the CNN-LSTM hybrid model independently applied to eight different genomes: *Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Arabidopsis thaliana*, *Triticum aestivum*, *Zea mays*, *Glycine max*, *Nicotiana tabacum*, and *Oryza sativa*. The average test accuracies were calculated across five trials. Two conditions are compared: models without data augmentation (red bars) and models with data augmentation (blue bars). Without augmentation, the average test accuracy for *C. vulgaris* was 0.476, while the average test accuracy for the other models was 0. Therefore, as a result, the red bars indicating the performance of models without data augmentation are not visible (except for *C. vulgaris*). The Student's *t*-test analysis in mean average test accuracy for the model performance with and without the applied data augmentation was separately analyzed for each genome. The significant differences at $t < 0.001$ are marked by three asterisks (***).

with data augmentation across all datasets (Fig. 2). For instance, accuracies for augmented data were highest for *A. thaliana* (97.66%), followed closely by *G. max* (97.18%) and *C. reinhardtii* (96.62%), demonstrating the model's ability to generalize well across higher plant and algal genomes. The standard error analysis further supported the robustness of the augmented approach, as shown by the low error rates for genome datasets such as *C. vulgaris* (0.25%) and *O. sativa* (0.33%) (Fig. 2).

Furthermore, the high accuracy achieved in both training and validation, coupled with minimal discrepancies between these metrics, indicated that the model generalized well to unseen data and did not suffer from overfitting. For instance, the CNN-LSTM model achieved a training accuracy of 97.13% with a corresponding training loss of 0.0641, while the validation accuracy reached 96.37% with a validation loss of 0.0671 on the *C. reinhardtii* dataset. The final test accuracy was 96.27%, with a test loss as low as 0.0661, further demonstrating the robustness of the model (Fig. 2). To further confirm the latter, the lack of improvement in validation accuracy, coupled with a steady rise in validation loss, demonstrated that the model was unable to generalize and learn effectively without data augmentation (Fig. 3A and S1). However, with data augmentation implementation, the model exhibited remarkable signs of learning and generalization. The training loss steadily decreased throughout the training process, converging to a minimal level close to zero by the final epochs (Fig. 3B and S1). Similarly, the validation loss showed a rapid initial decline, followed by a gradual decrease, ultimately reaching a low and stable value, which indicated the effectiveness of the learning model without substantial overfitting (Fig. 3B and S1). The model's training accuracy increased continuously with each epoch, reaching over 96%. The validation accuracy followed a similar upward trajectory, achieving a comparable high accuracy, thus indicating successful generalization to unseen data. The close alignment of training and validation accuracy by the final epochs further confirmed that the model did not suffer from overfitting, affirming the robustness of the data augmentation strategy.

Further analyses to investigate the efficiency of the augmentation approach

The model's predictive efficiency was further quantified using the analysis of i) precision-recall curves and AUC scores, ii) correlation analysis of predicted vs. experimental data, and iii) feature importance analysis.

i) Evaluation of precision-recall performance

The precision-recall curves of the CNN-LSTM hybrid model were evaluated across multiple genomes, both with and without data augmentation, using the average area under the curve (AUC) as a metric. For the non-augmented datasets, the average AUC values ranged between 0.021 and 0.065 (Fig. 4). Specifically, the highest average AUC was observed for *T. aestivum* (0.0656), followed closely by *C. reinhardtii* (0.0524) and *Z. mays* (0.039) when no augmentation was applied (Fig. 4). In contrast, the results indicated significantly higher average AUC scores when the CNN-LSTM hybrid model was trained on the same datasets with data augmentation. For example, the average AUC for augmented data reached its peak for *A. thaliana* (0.991), followed closely by *T. aestivum* (0.9867) and *C. reinhardtii* (0.984). Statistical analysis revealed that the average AUC scores for each genome dataset were significantly higher ($t < 0.001$) across Augmented datasets compared to those without augmentation (Fig. 4).

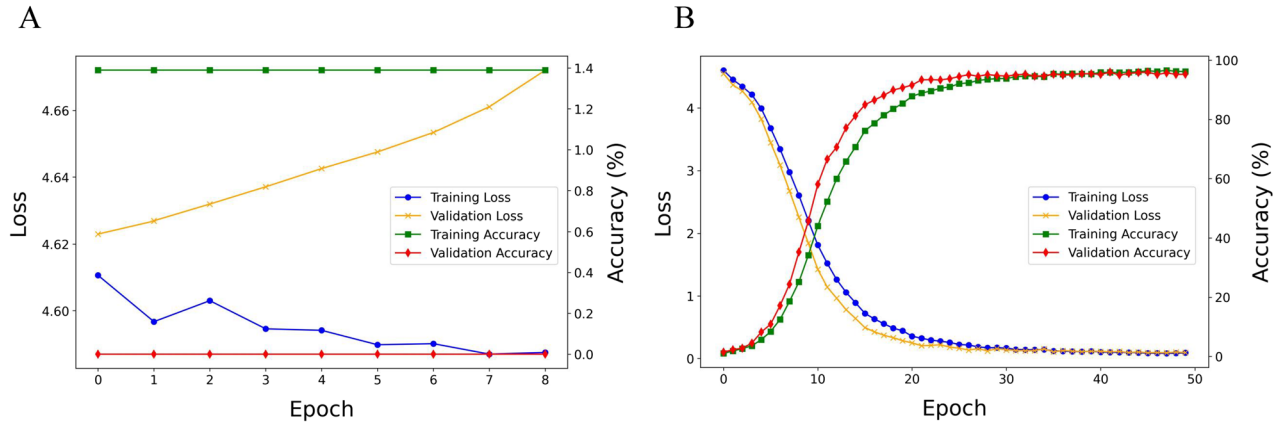

Fig. 3. Training and validation performance of the CNN-LSTM model on the *Chlamydomonas reinhardtii* dataset with and without data augmentation. The figure presents the training and validation performance of the CNN-LSTM model across multiple epochs, tracking both the loss and accuracy metrics for the dataset (**A**) without the data augmentation and (**B**) with the data augmentation. The x-axis denotes the number of epochs, capturing the progress of model training, while the primary y-axis (left) illustrates the mean loss values for both training and validation sets. Training loss is represented by a blue line with circular markers, while validation loss is indicated by an orange line with cross markers. The secondary y-axis (right) displays accuracy in percentage, with green (squares) and red (diamonds) lines representing training and validation accuracy, respectively. The same training and validation performance of the other seven genome datasets are presented in S1.

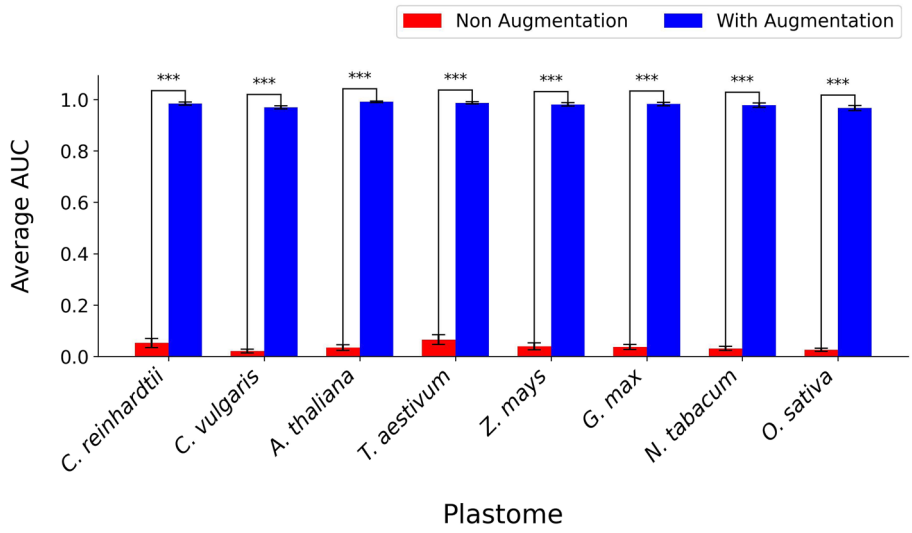

Fig. 4. Comparison of average area under the curve (AUC) mean scores for the CNN-LSTM hybrid model with and without augmentation across the multiple genome datasets. The bar chart presents the average AUC scores of the model independently evaluated on eight distinct plastid genomes: *Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Arabidopsis thaliana*, *Triticum aestivum*, *Zea mays*, *Glycine max*, *Nicotiana tabacum*, and *Oryza sativa*. Two experimental conditions are shown: model trained without augmentation (red bars) and model trained with data augmentation (blue bars). The Student's *t*-test comparing the mean AUC scores for model performance with and without the data augmentation for each genome was separately analyzed, and the mean differences ($t < 0.001$) were indicated by three asterisks (***).

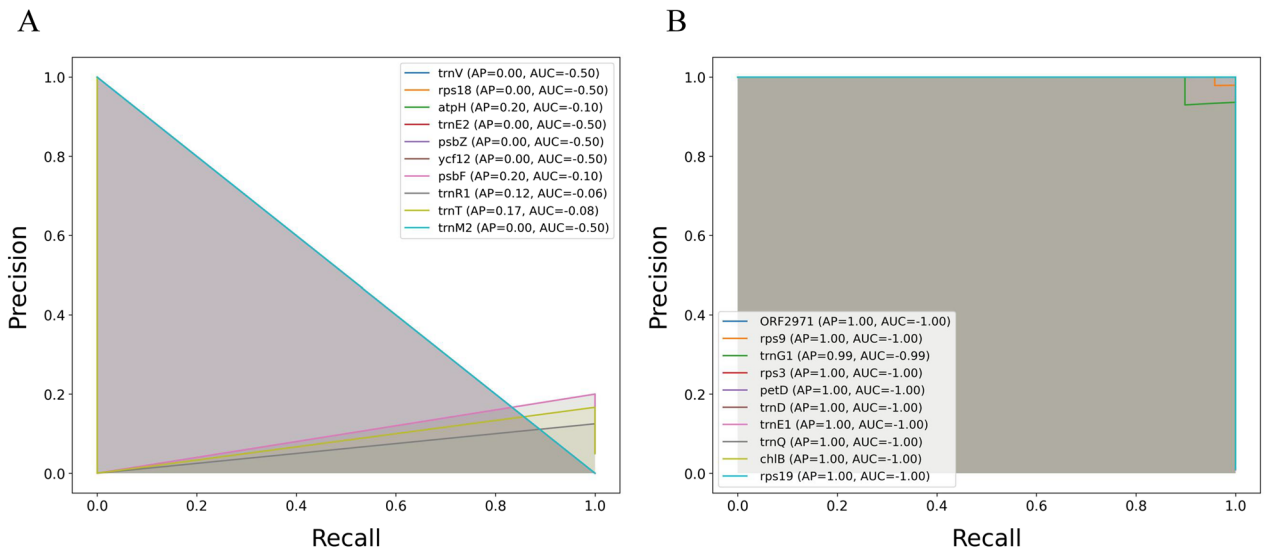

Fig. 5. Precision-recall curve analysis for the CNN-LSTM model with and without augmentation with class-specific average precision (AP) and area under the curve (AUC) metrics across the *Chlamydomonas reinhardtii* dataset. This plot presents the precision-recall curves for a subset of 10 randomly selected classes from the model's classification of the dataset (A) without the data augmentation and (B) with the data augmentation. The precision-recall curve for each class indicates the trade-off between precision and recall values at varying thresholds. The curves are accompanied by the average precision (AP) score and the area under the curve (AUC) for each selected class. The precision-recall curve analysis of the DNA datasets of the other seven plastomes is shown in S2.

Complementing the precision-recall analysis, the precision-recall curves for 10 randomly selected classes of distinct datasets were illustrated. The results demonstrated improvements in both the average precision (AP) and area under the curve (AUC) metrics when data augmentation was applied (Fig. 5 and S2). For instance, with data augmentation on *C. reinhardtii* dataset, the model achieved the highest AP and AUC values (AP = 1.00

and AUC = −1.00) across the 10 classes (Fig. 5 B), underscoring the robustness gained through introducing variability during training. In contrast, the highest AP and AUC values were 0.33 and −0.50, respectively, on the same dataset without augmentation (Fig. 5 A). More specifically, the AUC values without augmentation ranged from −0.06 to −0.50, whereas with augmentation, they remarkably increased to a range between −0.99 and −1.00 (Fig. 5).

ii) Correlation analysis of predicted versus experimental data

To further assess the impact of data augmentation on model performance, the Pearson correlation coefficient between model predictions and experimental values was analyzed across eight genome datasets, both with and without data augmentation. The data augmentation consistently demonstrated higher significant ($t < 0.001$) correlation values for all genome datasets (Fig. 6), reflecting an enhanced alignment between the model's predictions and the actual experimental outcomes. For instance, across the genome dataset from *C. reinhardtii*, the hybrid model with data augmentation achieved a Pearson correlation coefficient of 0.98 that significantly surpassed those without augmentation with a Pearson correlation coefficient of 0.00 (Fig. 6). This indicated that data augmentation notably enabled the model's ability to generalize, leading to predictions that match the empirical data more closely.

iii) The feature importance analysis

The feature importance analysis utilizing SHAP (SHapley Additive exPlanations) was used to identify key features contributing to the prediction accuracy of the CNN-LSTM on the datasets applied to the data augmentation (Figs. 7 and S3). The analysis provided valuable insights into elucidating the interactions between selections of influential features (nucleotides in specific positions) in the prediction model for some sequences of the datasets. For instance, by applying data augmentation on the *C. reinhardtii* dataset, the hybrid model demonstrated a marked ability to interpret the significance of specific nucleotides during classification tasks (Fig. 7 and S3).

The matrix-like plot of the dataset revealed that certain feature pairs, such as those contained features 15 and 90, exhibited a broad distribution across both positive and negative SHAP interaction values. The latter indicated that these features could have a significant impact on model predictions in both enhancing and suppressive capacities (Fig. 7). Conversely, interactions between features 22 and 4 displayed clustering near the baseline, suggesting a relatively minimal impact on the model's output. Furthermore, high values of feature 88 interacting with feature 67 resulted in a more pronounced positive interaction effect, as indicated by the cluster of red points to the right of the baseline (Fig. 7).

Augmenting of protein sequence datasets

To evaluate the performance of the CNN-LSTM hybrid model, protein datasets that underwent data augmentation were also employed. These datasets presented two remarkable differences from the DNA datasets, which posed

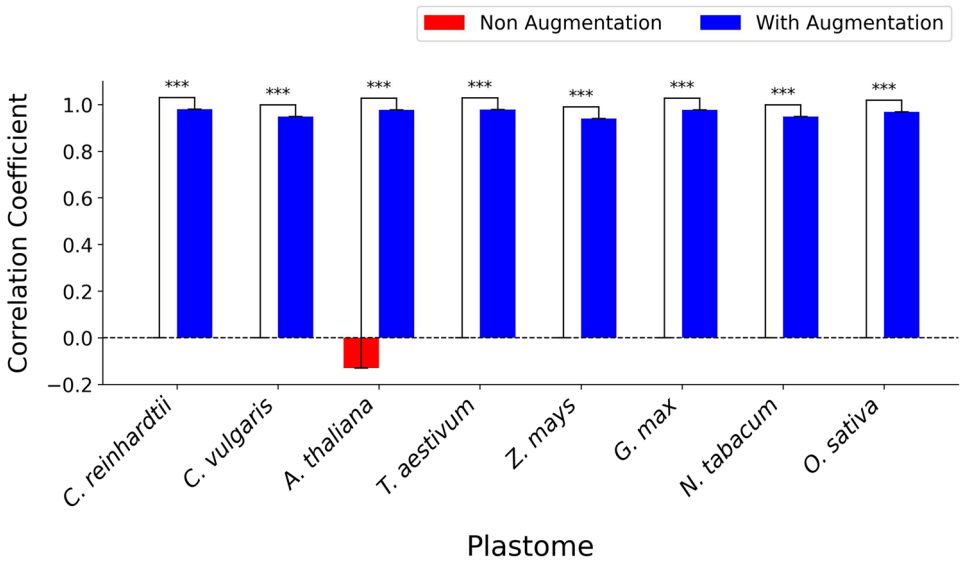

Fig. 6. Model performance via Pearson correlation coefficients between predicted and actual values across multiple genomes with and without the augmentation. The bar chart illustrates the Pearson correlation coefficients between model predictions and experimental values for eight plastomes—*Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Arabidopsis thaliana*, *Triticum aestivum*, *Zea mays*, *Glycine max*, *Nicotiana tabacum*, and *Oryza sativa*. The hybrid model for each genome dataset was evaluated under two conditions: without augmentation (red bars) and with data augmentation (blue bars). The significant mean difference ($t < 0.001$) for Pearson correlation coefficients was separately analyzed with the Student's *t*-test between the model performance for each genome with and without the data augmentation, showing by three asterisks (***).

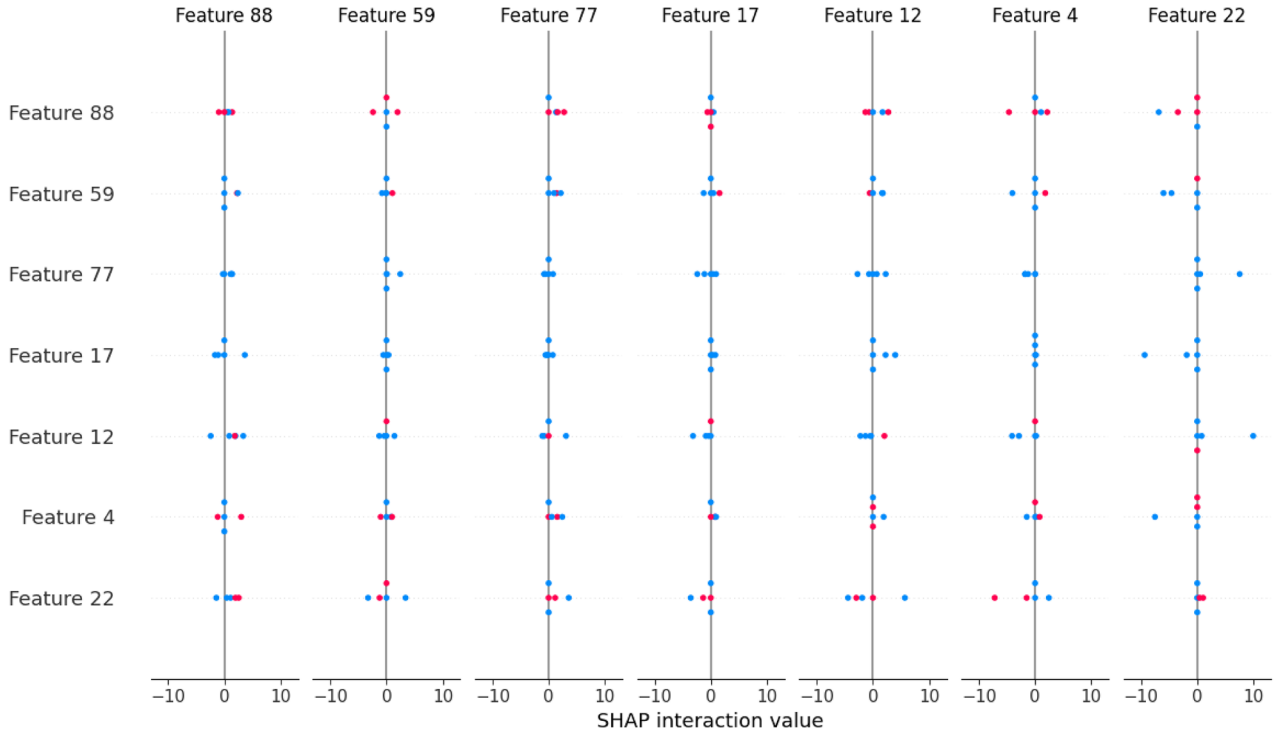

Fig. 7. Feature importance analysis of nucleotide sequences from the augmented *Chlamydomonas reinhardtii* plastome dataset was performed using SHAP within the CNN-LSTM model framework. Each feature corresponds to a specific nucleotide type (A, C, G, or T) at one of the 40 positions within each input sequence, as encoded by a one-hot representation. Therefore, each nucleotide position contributes four potential features, resulting in a total of 160 (40*4) distinct feature indices. In the SHAP interaction plot, each feature pair interaction is visualized in two symmetrical positions: where each feature is considered once as the primary and again as the secondary feature. Each feature pair position includes multiple points (five in this plot as a subset of samples), representing how the interaction varies across different samples in the dataset. Each dot's color indicates the feature value, with blue dots representing lower values and red dots representing higher values. The direction (left or right of zero) and spread of points demonstrate the interaction's influence, with points to the right indicating positive interaction contributions to the model and points to the left showing negative interactions. The same feature importance analysis of the other seven genome datasets is presented in S3.

additional challenges for applying deep learning models to identify meaningful patterns. First, a substantial portion of the genes of the plastome corresponds to non-coding RNAs, including tRNAs and rRNAs, resulting in a lower number of corresponding protein sequences in each dataset. For example, the DNA dataset for *A. thaliana* included 113 genes, while its protein dataset contained 56 sequences. Second, the variability in coding sequence (CDS) lengths leads to a highly imbalanced dataset, complicating the training of models. For instance, the varied lengths of the protein sequences in the *C. reinhardtii* dataset, ranging from 112 to 1995 aa, led to the generation of 73 subsequences from the shorter sequence (rpl20) while 1956 subsequences were generated from the longer sequence (ORF1995) (S4). Consequently, the effectiveness and efficiency of the data augmentation technique were assessed to address limitations associated with low sequence counts and imbalance.

Efficiency of the data augmentation on the model performance The performance of the hybrid model, with and without the data augmentation, was evaluated across eight genomes using four metrics: macro precision, macro recall, macro F1 score, and their corresponding weighted scores (Fig. 8). The results showed a significant distinction between the performance of the model applied to each dataset with and without data augmentation. Model performance on each dataset with the data augmentation significantly outperformed its corresponding counterparts without it. Specifically, for the macro scores, the model applied to the datasets with data augmentation demonstrated precision values ranging from 0.971 to 0.991, recall values ranging from 0.973 to 0.992, and F1 scores from 0.969 to 0.992 across the genomes, indicating generally high performance (Fig. 8). The weighted scores also revealed similar trends, with precision values ranging from 0.924 to 0.990, recall values from 0.925 to 0.989, and F1 scores from 0.924 to 0.989 (Fig. 8).

The improvements in the average precision, recall, and F1 score were statistically significant ($t < 0.001$) compared to the corresponding metrics of the model's performance on the datasets without data augmentation, which resulted in zero values for all metrics (Fig. 8). In other words, significant differences comparing the mean performance metrics between the two conditions were observed for all genomes, confirming that data augmentation significantly enhanced the model's performance across all genomes. Ultimately, the incorporation

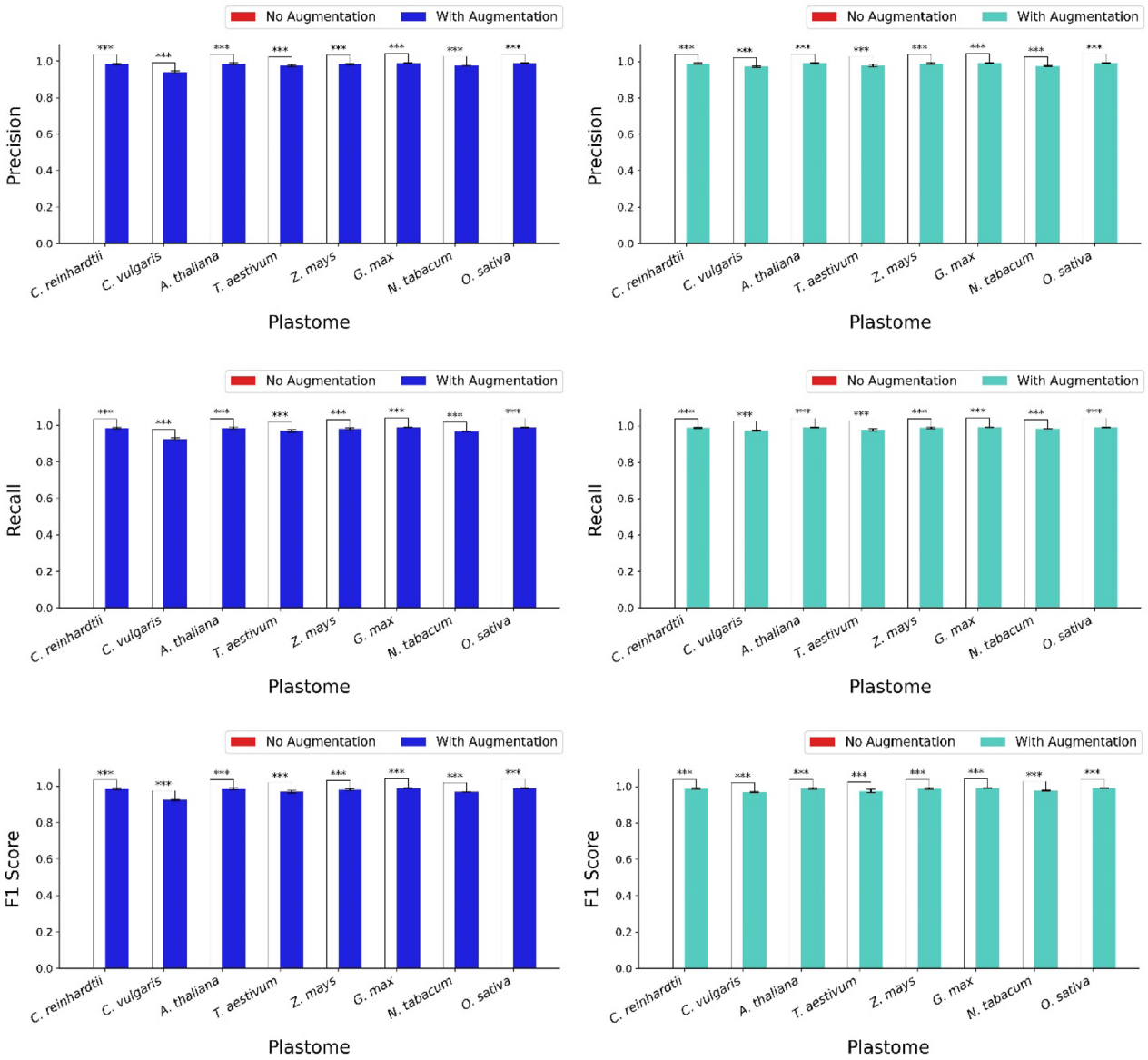

Fig. 8. Average precision, recall, and F1 score (both macro and weighted) across eight genomes with and without the data augmentation. The bar chart shows the average macro precision, recall, and F1 score, as well as weighted precision, recall, and F1 score of the CNN-LSTM hybrid model independently applied to eight different plastomes: *Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Arabidopsis thaliana*, *Triticum aestivum*, *Zea mays*, *Glycine max*, *Nicotiana tabacum*, and *Oryza sativa*. Average test accuracies were calculated across three trials and five folds. Two conditions are compared: models with data augmentation (blue bars for macro scores and turquoise bars for weighted scores) and models without the data augmentation (red bars). Since the metric values for all datasets without augmentation were 0, the red bars indicating the performance of models without data augmentation are not visible. Statistical analysis using the Student's *t*-test was conducted to compare mean test accuracies for each genome under both conditions, with significant differences at $t < 0.001$ marked by three asterisks (***).

of data augmentation substantially improved the precision, recall, and F1 score of the CNN-LSTM hybrid model, highlighting its importance in optimizing model performance across various genomes (Fig. 8).

Performance evaluation of plastome protein sequences classifications The confusion matrix provided a detailed visualization of the CNN-LSTM model's classification performance across protein sequence categories within the datasets, which were highly imbalanced (Fig. 9 and S5). The results, e.g., for *C. reinhardtii* dataset, showed that the model achieved high accuracy, as indicated by the dominant presence of higher values along the diagonal of the matrix representing correctly predicted classifications. The diagonal cells reflected strong performance, where the true and predicted labels align across most protein sequence categories. Misclassifications were minimal, as most non-diagonal cells contain zero values, indicating no false predictions (Fig. 9). However, a

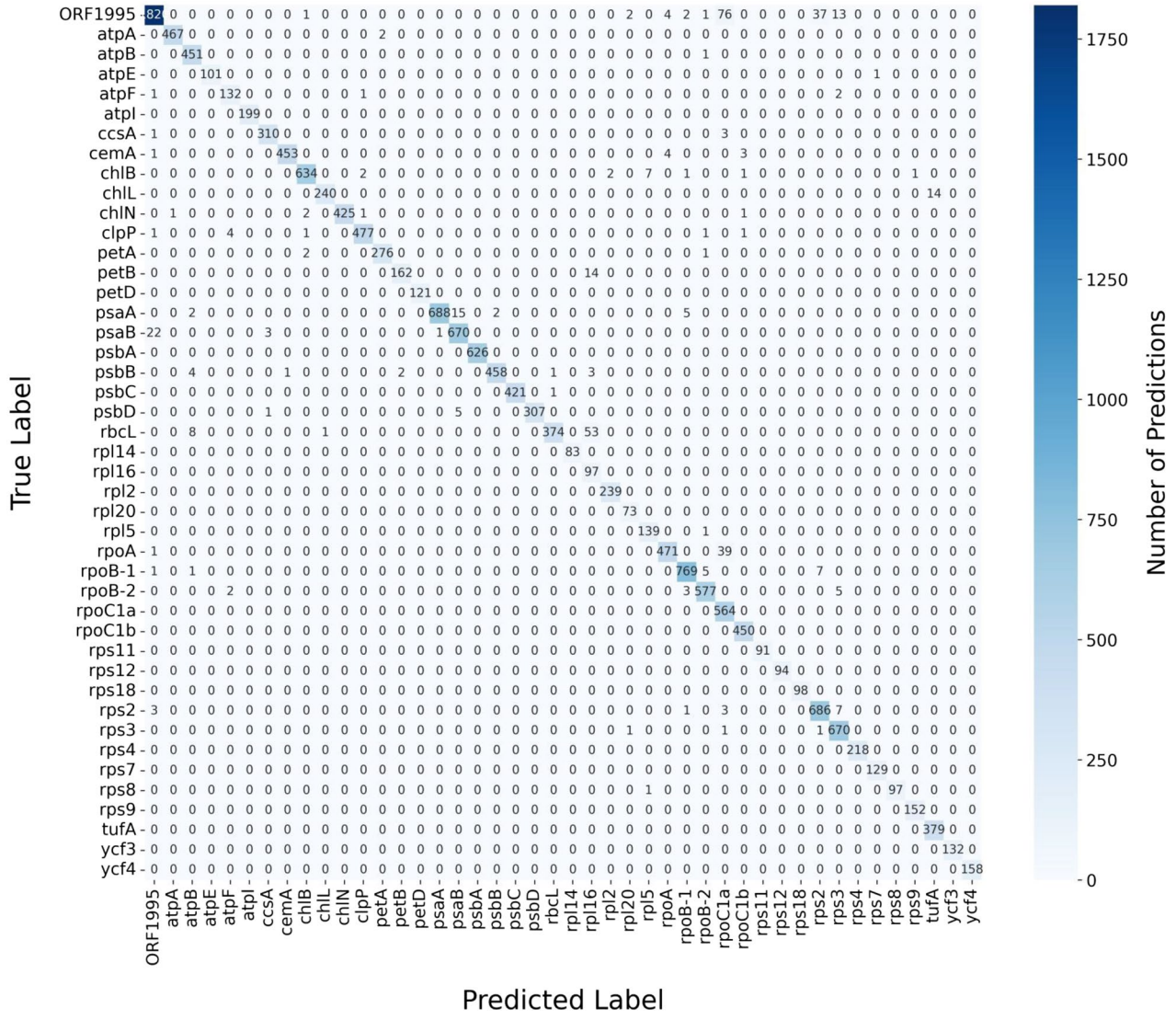

Fig. 9. Confusion matrix for classifying *Chlamydomonas reinhardtii* chloroplast genome protein sequences with data augmentation using CNN-LSTM model. The confusion matrix presented here visualizes the model's classification performance across different protein sequence categories in the dataset. Each row of the matrix corresponds to the actual class label, while each column represents the predicted label. Cells along the diagonal indicate correctly predicted classifications, where the true label matches the predicted label. Non-diagonal cells represent misclassifications, where the predicted label does not match the true label. The intensity of color in each cell corresponds to the number of instances predicted for that class, with darker shades representing higher counts. The confusion matrix for classifying the protein sequences of other seven plastomes are presented in S5.

few non-diagonal cells displayed low values, suggesting occasional misclassifications in certain protein sequence categories. These results highlighted the model's ability to make accurate predictions overall, with only a small number of misclassifications occurring for specific classes. These low values highlighted specific classes where the model occasionally misclassifies sequences, which may be indicative of class confusion or potential imbalances in the dataset (Fig. 9).

Model performance using precision-recall and Receiver Operating Characteristic (ROC) curve analysis on plastome protein datasets The performance of the CNN-LSTM model on the protein datasets with data augmentation was further evaluated using precision-recall and ROC curves (Fig. 10 and S6). In both curves, e.g., for the *C. reinhardtii* plastome protein dataset, the model demonstrated exceptionally high classification accuracy (Fig. 10). The precision-recall curve highlighted a near-perfect balance between precision and recall, with the curve consistently close to the top-right corner. This indicated that the model maintained a high precision across a wide range of recall values, minimizing false positives even as recall increased. The AUC for the precision-re-

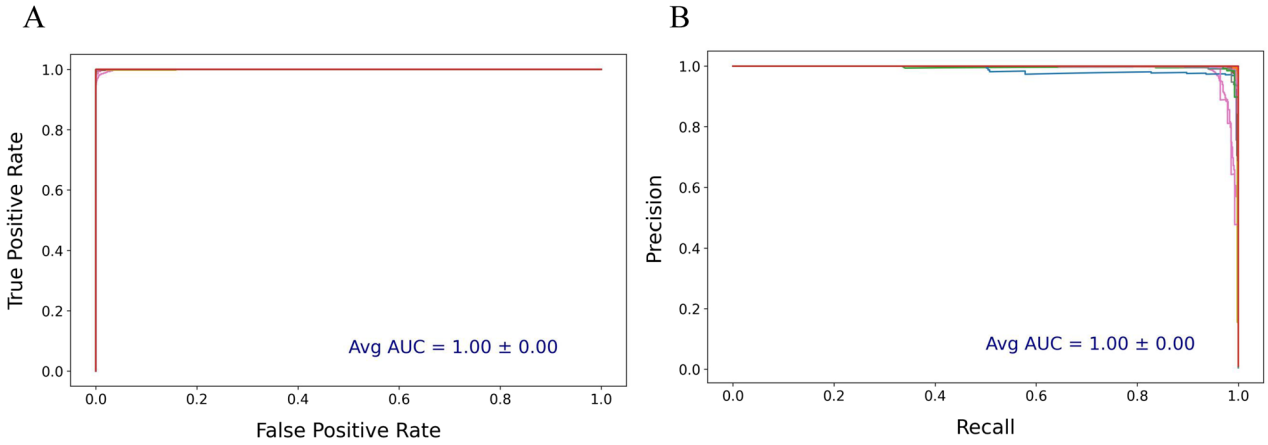

Fig. 10. Precision-recall and Receiver Operating Characteristic (ROC) curves for *Chlamydomonas reinhardtii* plastome protein sequences with the data augmentation. This presents (**A**) ROC curves and (**B**) precision-recall generated from a CNN-LSTM model applied to the dataset protein sequence classification data over 50 epochs and 5 cross-validation folds. In the precision-recall curve, the x-axis indicates recall (sensitivity), and the y-axis shows precision. In the ROC curve, the x-axis represents the false positive rate, while the y-axis shows the true positive rate. The average AUC values from all classes was annotated for each plot. The precision-recall and ROC curves for other plastome protein datasets are shown in S6.

call curve averaged 1.00 with a standard error of 0.00, underscoring the model's strong ability to correctly classify positive instances with minimal error (Fig. 10).

Similarly, the ROC curve showed an ideal shape, closely approaching the top-left corner of the plot. The false positive rate remained low as the true positive rate increased, reflecting the model's high specificity and sensitivity in distinguishing true positives from false positives (Fig. 10). The average AUC for the ROC curve was also 1.00 with a standard error of 0.00, indicating consistent performance across all cross-validation folds and minimal variability. In other words, the ROC curve's perfect diagonal placement reflected optimal performance, indicating a strong ability to discriminate between the positive and negative classes with minimal false positives and false negatives (Fig. 10).

Model performance across epochs using mean loss, F1 score, and recall on plastome protein datasets The model demonstrated substantial learning and generalization (Fig. 11 and S7). For *C. reinhardtii* with the augmentation, the mean training loss exhibited a rapid decline during the first 10 epochs, after which it decreased steadily, approaching a minimal value by the final epochs (Fig. 11). This decrease in loss indicated effective model training with diminishing errors. Likewise, the mean validation loss followed a similar pattern, showing an initial sharp decline, followed by a gradual decrease, ultimately stabilizing at a low level, which showed that the model was learning effectively without significant overfitting (Fig. 11).

Both the training and validation recall and F1 scores showed consistent improvement throughout the epochs. The training recall and F1 scores steadily increased, surpassing 95% by the later epochs. The validation scores mirrored this upward trend, reaching similarly high values, which indicated that the model was able to generalize well to unseen data. The close alignment between the training and validation performance metrics in the final epochs further confirmed that overfitting was not present, emphasizing the effectiveness and robustness of the data augmentation approach (Fig. 11).

Therefore, the developed innovative augmentation strategy effectively addressed the challenges of applying deep learning to omics and other related biological datasets with limited data availability, particularly in cases where each gene or protein is represented by a single sequence. As illustrated in Fig. 12, a graphical flowchart summarizes the existing challenges posed by limited sequence availability for enabling deep learning applications in omics and related datasets, alongside the strategy to overcome these challenges for ease of understanding. This method expanded the datasets without introducing noise or causing overfitting, while preserving the biological integrity of the sequences and ensuring comprehensive coverage of the original data.

Unveiling the power of the CNN-LSTM model for short omics sequences

The performance of the CNN-LSTM hybrid model was evaluated against other machine learning models, including Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN), on datasets containing sequences of 40 nucleotides or amino acids. These models struggled to effectively learn patterns from the short sequences, achieving less than 18% accuracy on the test datasets (unpublished data). In contrast, the optimized CNN-LSTM model, with its architecture shown in Fig. 13, demonstrated significantly improved performance on both the training and test datasets, highlighting its ability to capture the complex patterns within the input sequences. This superior performance reflects the hybrid model's capability to address the challenges posed by the short length and inherent complexity of the datasets, which could not be adequately handled by traditional models.

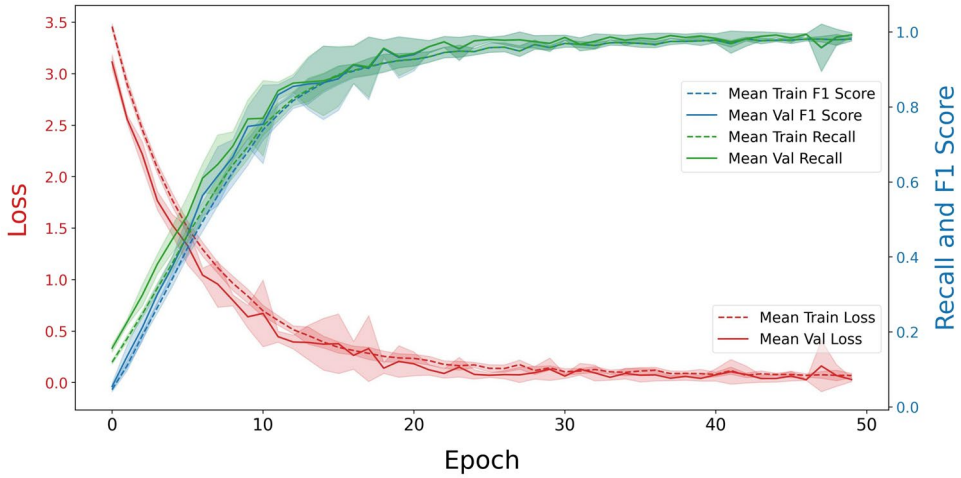

Fig. 11. Model performance across epochs via mean loss, F1 score, and recall on plastome protein dataset of *Chlamydomonas reinhardtii* with augmentation using CNN-LSTM Model. This figure illustrated the model's training and validation performance across 50 epochs, averaged over 5 cross-validation folds. The x-axis represents the number of epochs, while the y-axes display mean loss values (left y-axis) and performance metrics: recall and F1 score (right y-axis). The red, green, and blue curves depict the mean loss, recall, and F1 score for both the training (dashed line) and validation (solid line) sets. The shaded areas around these lines represent the standard deviation across folds, reflecting the variability in loss for each epoch. The same model performance for other genomes are presented in S7.

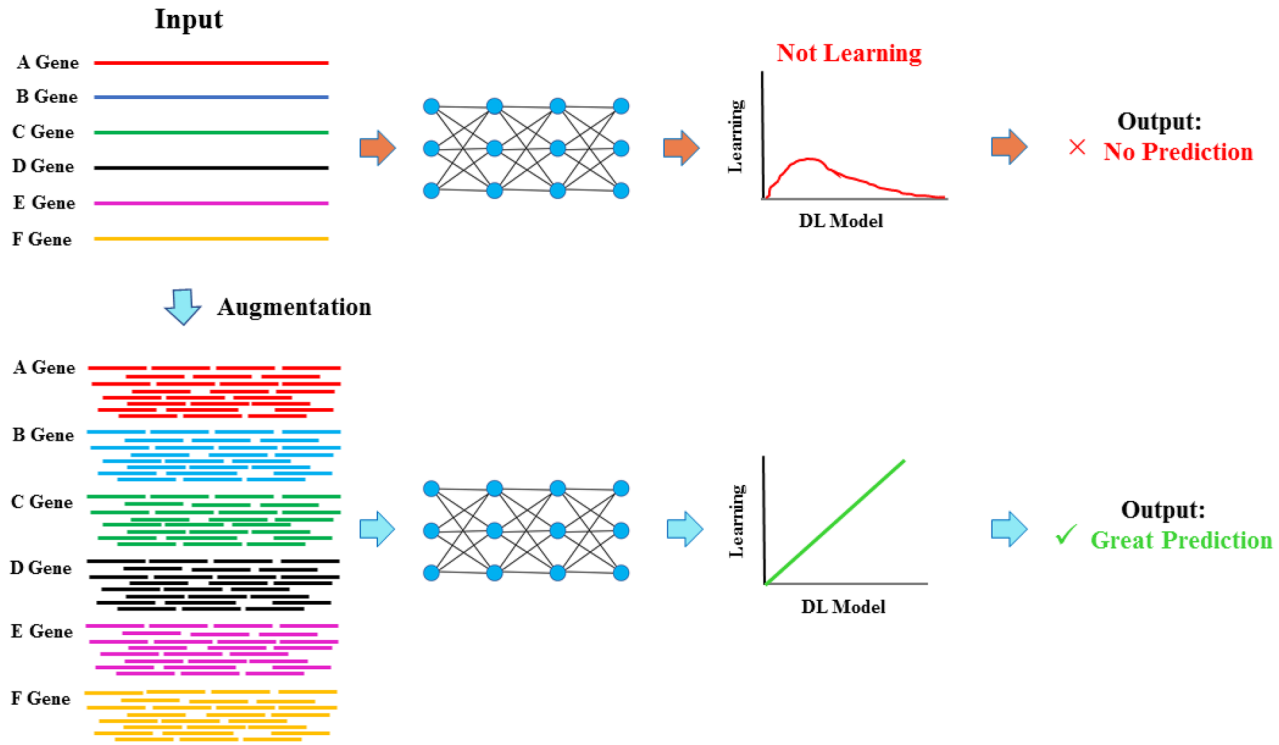

Fig. 12. Graphical flowchart illustrating data augmentation strategy to address the challenges of limited sequence availability in omics and related biological datasets, thereby enabling effective deep learning applications.

Data augmentation strategy for enabling unsupervised analysis

In unlabeled data analysis, an approach was developed to maximize feature generation from a dataset of limited genomic sequences, converting each 300-nucleotide sequence into a high-dimensional set of k-mer features (from 5 to 12 nucleotides). Using the k-mer approach, common motifs between each pair of sequences was

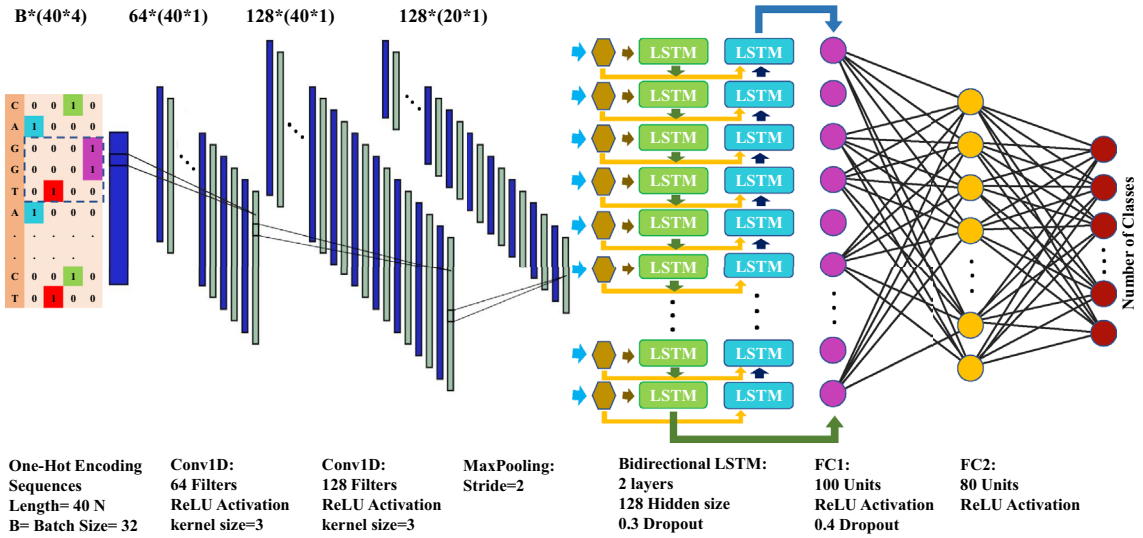

Fig. 13. A graphical illustration of the optimized CNN-LSTM model architecture designed for datasets with limited gene representations. This architecture was independently evaluated on multiple nucleotide and protein datasets when applied with and without data augmentation. One-hot encoding sequences with four channels here are for nucleotide sequences that is different from that of amino acid sequences. The architecture comprises two sequential convolutional layers followed by a max-pooling layer. The output is reshaped and passed through a bidirectional LSTM with two layers, followed by two fully connected layers. The final output layer maps the processed features to the number of classes in the dataset. Further details are provided in the material and method section.

quantified, identifying common in k-mer content even among a small dataset. This method produced thousands of pairwise metrics, capturing a high degree of sequence similarity and difference, which underscores the effectiveness of k-mer-based feature generation for producing robust datasets from limited gene samples. To determine all possible pairwise interactions among the dataset, the number of unique gene sequence pairs was calculated using the following formula:

$$C(n,k) = \frac{k!}{(n-k)!n!}$$

In this context, *n* represents the total number of gene sequences under consideration, while *k* is set to 2, indicating that the analysis focuses on pairs of gene sequences.

For example, in the dataset of *C. reinhardtii* with 100 plastome sequence genes, 4,950 unique combinations of gene pairs were obtained. The resulting dataset comprised a vast number of k-mer counts across sequences, providing meaningful variations and increasing the dataset's depth for unsupervised learning.

In this unsupervised approach, the k-mer extraction was applied to the same 300-nucleotide sequences that previously used for supervised augmentation. The goal was to capture recurring k-mers across sequence pairs, thus identifying core nucleotide patterns that may indicate biologically relevant relationships among sequences without requiring labeled data. By extracting k-mers across lengths of 5 to 12 nucleotides, the method generated a set of shared k-mers between sequence pairs, providing a framework for comparative analysis. The results revealed diverse levels of k-mer sharing between pairs, quantifying genetic similarity and establishing a foundational distance matrix for subsequent clustering and dimensionality reduction.

To assess the k-mer patterns, a distance matrix representing the inverse of common k-mers between pairs was created, which was then clustered using k-means and visualized through Principal Component Analysis (PCA). This clustering revealed distinct groups of sequences, each representing unique k-mer signatures. The process produced clusters that allowed for the visualization of sequence relationships based on k-mer composition, ultimately emphasizing how even a low count of genes can yield a feature-rich dataset appropriate for unsupervised learning. This analysis demonstrated the flexibility and utility of k-mer-based feature generation in producing sufficient, biologically relevant metrics from limited sequence data, meeting the needs of projects constrained by small gene counts (Fig. 14 and S8).

Further analysis of the identified clusters revealed biologically meaningful groupings, where each cluster represented sequences sharing characteristic k-mer motifs. For example, Cluster 1 included sequences that appeared closely related based on recurrent k-mers, potentially indicating shared functions or evolutionary related traits. In contrast, Cluster 4 housed sequences with a different k-mer profile, pointing to distinct lineage or functional properties (Fig. 14). This clustering of unlabeled data demonstrated the potential of k-mer similarity measures to identify sequence relationships, providing insight into the inherent structure of genomic data in a label-free setting.

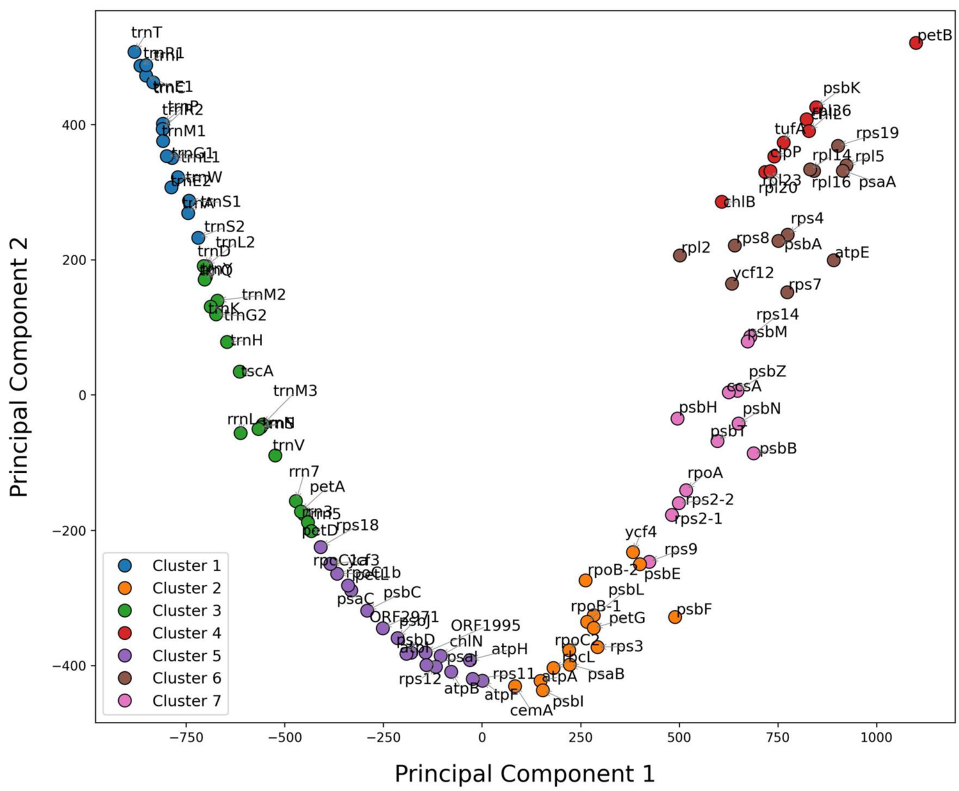

Fig. 14. Clustering of the *C. reinhardtii* plastome dataset of 300-nucleotide sequences based on the total number of common k-mers. Each data point on the scatter plot was annotated with its corresponding sequence, allowing for easy identification and interpretation of individual sequences within each cluster. Every group of sequence clustering in a group was distinguished with the same color. The clustering plots of the DNA datasets of other seven plastomes are presented in S8.

Discussion

Limited representation in small biological datasets, especially those containing approximately 100 unique gene or protein sequences, presents a major obstacle for deep learning (DL) applications. These datasets often lack the diversity and scale needed for DL models to effectively learn complex patterns and relationships. In other words, DL approaches have demonstrated significant potential for analyzing large-scale datasets in the fields of genomics, transcriptomics, and proteomics, owing to their capacity to discern intricate patterns and relationships within extensive datasets[2,29,31–33]. However, due to the inherent characteristics of the algorithms employed, their applicability is limited when applied to datasets with limited representation—notably when each gene or protein is represented by a single sequence. In the current study, a novel data augmentation strategy was developed that addressed this fundamental challenge. The approach involved generating overlapping subsequences with variable overlaps, preserving the biological relevance of each subsequence while significantly expanding the available data. A sliding window technique with controlled overlap and shared nucleotide features, enabled the creation of a large and diverse dataset covering their entire original sequences while maintaining the functional and structural integrity.

Traditional data augmentation methods, which involve altering sequences or generating synthetic data by introducing noise or irrelevant variations, are often not feasible due to the risk of introducing biologically irrelevant changes[28,29,31]. In the current study, this limitation was specifically addressed by applying a method that expanded datasets without altering the inherent biological content of the sequences. This is a significant advancement over previous methods in genomic deep learning, which often rely on random transformations or simulating evolutionary variations[29–32]. The distinguishing feature of the current approach lies in its ability to apply deep learning techniques to naturally constrained datasets efficiently. In contrast, previous data augmentation methods, despite their inherent limitations, primarily aimed to enhance the performance of models that applied to relatively small datasets. This is particularly crucial for omics data, where even a single nucleotide change can drastically impact the function of regulatory elements or the biological activity of genes and proteins[29,30].

The data augmentation was essential for expanding each original sequence without introducing non-representative changes, which allowed the model to achieve notable improvements in accuracy and generalizability. As shown in the results, the model performed poorly on non-augmented datasets, with accuracy values close to zero, indicating a lack of predictive capability in limited-data conditions. In contrast, the model achieved high accuracy scores with augmented datasets, reaching up to 97.66% for the *A. thaliana* genome,

and demonstrated low standard errors across all genome datasets. This improvement illustrated the model's adaptability and enhanced learning capability when trained on a diverse set of augmented data. The current study aligns with previous research, such as CLPr_in_ML, that demonstrates the effectiveness of advanced computational approaches in enhancing model performance by preserving critical invariant features while introducing controlled variability[34]. Moreover, the CNN-LSTM hybrid model demonstrated high performance across both DNA (e.g., utilizing 100 sequences of *C. reinhardtii*) and protein (e.g., employing 46 sequences of *C. reinhardtii*) datasets, with no signs of overfitting despite the data scarcity. This is in stark contrast to the typical overfitting observed in deep learning models trained on small datasets without sufficient augmentation[30,35–37].

The high training and validation accuracies, combined with closely aligned loss metrics, further confirmed the robustness of the model, as it successfully avoided overfitting while generalizing well to unseen data. For instance, in the *C. reinhardtii* dataset, the training and validation accuracies converged above 96%, indicating a stable and high-performing model without significant deviations between the training and validation phases. The data augmentation strategy contributed significantly to this performance by introducing variability within the training data, thereby enhancing the model's ability to recognize patterns and retain important nucleotide or protein sequence features.

Multiple analyses supported the model's effectiveness. First, the high performance of the CNN-LSTM hybrid model on augmented datasets was obtained across eight diverse chloroplast genomes from microalgae and higher plants. Second, this model's success on both DNA and protein datasets demonstrates the versatility and effectiveness of the CNN-LSTM hybrid approach. Precision-recall curves also revealed that data augmentation consistently increased the area under the curve (AUC) values across all genomes, with the augmented data achieving AUC scores up to 0.991. Correlation analysis further validated these findings, demonstrating a high Pearson correlation between the model's predictions and actual experimental values, with values reaching as high as 0.98 for augmented datasets. This strong correlation indicated that the model's predictions were aligned with empirical results, underscoring the augmentation's role in enhancing predictive accuracy.

Using SHAP values, feature importance analysis, provided additional insights into how specific nucleotide features influenced the model's predictions. For instance, certain nucleotide pairs showed positive and negative interactions, which could reflect underlying biological relevance. This ability to interpret feature significance through augmentation improved classification accuracy and added a layer of interpretability to the model's decision-making process. An advanced feature importance analysis approach also could enable the identification of critical motifs within the original sequences. Considering the point that the model has trained on overlapping subsequences, it effectively learned intricate patterns and sequence motifs that significantly enhanced predictive accuracy for classification tasks. These identified motifs serve as distinguishing markers, allowing for clear differentiation between classes. By aligning these subsequences and examining overlapping regions containing these key patterns and motifs, a comprehensive, cumulative insight into each original sequence will be obtained, highlighting sequence features that are instrumental in class separation.

The results also demonstrated the effectiveness of data augmentation in handling imbalanced protein datasets, where low sequence counts and length variability presented significant challenges[38–40]. By generating a more balanced dataset, the model's precision, recall, and F1 scores showed substantial improvements, with macro F1 scores reaching up to 0.992. Confusion matrices further highlighted the model's ability to correctly classify protein sequences, showing high values along the diagonal and minimal misclassifications. The protein dataset results were reinforced by precision-recall and ROC curve analyses, both of which showed near-perfect scores, validating the model's classification performance.

The current user-friendly augmentation process is highly adaptable, enabling researchers to modify key parameters such as the length of the subsequences, the degree of overlap, and the total number of generated subsequences, providing flexibility across different types of omics datasets. This flexibility is particularly beneficial for researchers working with diverse omics datasets, from small datasets with limited representation to larger and more comprehensive ones including whole-genome data. The adaptability is also particularly beneficial for pre-handling imbalanced datasets resulting from sequences of significantly longer lengths. Specifically, reducing the degree of overlap among these longer sequences can lead to a decrease in overcounted sequences, thereby mitigating the imbalance within the datasets. Additionally, our current approach can be applied not only to DNA and protein datasets but also to RNA datasets, making it a versatile tool for many areas of omics research.

The current approach offers notable advantages over transfer learning in biological research, particularly for tasks with limited datasets[41–43]. Unlike transfer learning, this approach does not depend on extensive external datasets for pre-training, which can be a labor-intensive, costly, and complex process. Transfer learning requires a relevant pre-trained model for each dataset, posing challenges in model availability; furthermore, if the original and new tasks diverge significantly, the performance of the pre-trained model may deteriorate. In contrast, the current method exhibits user-friendliness and adaptability across diverse DNA, RNA, and protein datasets, rendering it a versatile instrument for various omics applications while ensuring both accuracy and biological relevance. Additionally, unlike the opaque nature of transfer learning, the current approach enhances interpretability by elucidating meaningful sequence motifs and patterns.

In addition to introducing the novel data augmentation technique, the results underscored the efficacy of the CNN-LSTM hybrid model as a powerful tool for analyzing short omics sequences, such as those containing only 40 nucleotides or amino acids. Traditional models, including SVM, MLP, RNN, and CNN, failed to generalize effectively due to the limited information content in the short sequences and the complex relationships between features. By integrating CNN's strength in capturing spatial patterns with LSTM's ability to model sequential dependencies, the hybrid model provided a robust solution for learning intricate patterns from such datasets. The current study highlighted the importance of adopting hybrid architectures for tasks involving short biological sequences, where conventional models are insufficient. Furthermore, the CNN-LSTM approach offered a

promising framework for advancing sequence-based analyses, particularly in domains where data complexity and brevity present significant challenges.

Overall, the CNN-LSTM hybrid model, supported by a comprehensive data augmentation strategy, displayed high predictive accuracy, robustness, and generalization capabilities on both DNA and protein datasets across various genomes. The augmentation approach not only enabled the application of DL on biologically constrained datasets while maintaining the integrity of the original sequences but also mitigated overfitting, allowing for the effective differentiation of complex sequences. The current introduced approach enhances the applicability of deep learning to omics, offering new opportunities for researchers working with underrepresented biological data. The approach achieved by expanding small datasets without introducing noise or overfitting.

Indeed, the approach presented in the current study has broad applications in omics, molecular biology, and bioinformatics, particularly in addressing challenges associated with small or constrained datasets. For instance, it could capture intra-species variations or distinct biological functions not represented in the phylogenetic tree, support the RNA-Seq analysis of rare transcripts, and enhance the discovery of their RNA secondary structure motifs. The method could enable robust classification of protein families, identification of active domains, and rational protein design. It also addresses data imbalance, enabling the classification of rare elements such as orphan genes or unique protein domains. In addition, the approach advances predictive modeling, such as gene regulatory network reconstruction, and supports applications in evolutionary biology by tracing conserved and variable sequence motifs across species. This strategy enhances multi-omics integration by aligning features across genomics, transcriptomics, and proteomics, enabling more comprehensive systems biology analyses. Furthermore, the method has significant potential in diagnostics and therapeutics, including biomarker discovery, identification of nucleotide or protein motifs as drug targets, and development of CRISPR guides or therapeutic RNAs, making it a versatile tool for diverse biological and biomedical research applications.

In addition, the k-mer-based data augmentation strategy also offered a robust solution for enhancing unsupervised analysis in genomic research, particularly when faced with the limitations of small datasets. By extracting recurring nucleotide motifs from sequence pairs, the approach effectively amplifies the feature space, allowing for meaningful comparisons even in scenarios where sequence availability is low. The generation of pairwise k-mer metrics enriched the dataset with a multitude of similarity measures that reflect both sequence conservation and variation, thereby facilitating the derivation of biological insights without reliance on labeled data. This method not only supports exploratory analyses but also fosters hypothesis generation, making it a versatile tool applicable to DNA, RNA, and protein datasets. Importantly, to optimize the extraction process and ensure the reliability of k-mer generation, it is crucial to consider sequences of similar lengths; significant length disparities can compromise the integrity of the extracted features. This k-mer-based augmentation strategy proves to be a powerful mechanism for producing feature-rich datasets that enhance unsupervised clustering and analysis in omics studies. For instance, the current approach could serve as a versatile tool for identifying the functions of unknown genes, such as plastid *ycf* genes, and ORFs (open reading frame). These genes remain uncharacterized because their critical roles in the cell make them resistant to functional analysis through mutation studies[44–46]. By identifying common regulatory elements and functional motifs, it may be possible to infer the potential functions of these genes. Additionally, this method could be instrumental in uncovering co-functional genes or proteins and their associated networks by detecting shared motifs, thereby providing insights into their cooperative roles within cellular systems.

Conclusion

In the current study, a novel and highly effective data augmentation strategy was introduced that addresses the challenges posed by small genomic datasets. This approach enables deep learning applications for omics and other related biological datasets, even when limited sequence data is available. The method expands the datasets while preserving the biological integrity of the sequences and ensuring comprehensive coverage of the original sequences. This makes it a powerful tool for training deep learning models on small genomic datasets to recognize the intricate patterns and sequence motifs in the original sequences. The flexibility of this user-friendly approach, combined with its ability to generate biologically relevant data, makes it applicable to a wide range of biological data repositories. By enabling high-performance deep learning models to capture subtle, biologically relevant patterns in small datasets, the current approach offered a significant advancement in the field of biological data analysis. Another key advantage of the current introduced approach is its user-friendliness that all its parameters such as the overlapping length and the length of the generated sequences could easily be modified. This flexibility makes the method applicable not only to gene or regulatory sequences but also to genomic sequences. Moreover, this strategy is not only applicable to small datasets but also provides a robust framework for enhancing the performance of deep learning models in general. Using this novel augmentation technique, researchers can work with genomic datasets of various sizes and complexities, ensuring that their deep learning models can capture meaningful biological patterns without introducing noise or overfitting. Furthermore, the k-mer-based data augmentation strategy also provides a robust solution for unsupervised clustering, enabling the identification of unknown gene functions, uncovering co-functional genes and proteins, and advancing systems biology studies.

Materials and methods

Genomes and corresponding prepared DNA and protein datasets

To evaluate the efficiency and reproducibility of the data augmentation approaches for biological datasets with limited representations, two strategies were employed concurrently. The method was applied to two primary datasets: one comprising DNA sequences, characterized by four distinct nucleotide bases, and the other consisting of protein sequences, composed of 20 different amino acids. Additionally, the method was independently evaluated on eight diverse datasets representing DNA and protein sequences sourced from

chloroplast reference genomes from microalgae and higher plants to assess its robustness and generalizability. These genomes included *Chlamydomonas reinhardtii* (NC_005353.1), *Chlorella vulgaris* (NC_001865.1), *Arabidopsis thaliana* (NC_000932.1), *Nicotiana tabacum* (MZ707522.1), *Triticum aestivum* (NC_002762.1), *Oryza sativa* (NC_031333.1), *Zea mays* (NC_001666.2), and *Glycine max* (NC_007942.1).

DNA dataset preparation
The number of genes included in the chloroplast datasets was as follows: 100 for *C. reinhardtii*, 210 for *C. vulgaris*, 113 for *A. thaliana*, 113 for *N. tabacum*, 116 for *T. aestivum*, 122 for *O. sativa*, 125 for *Z. mays*, and 110 for *G. max*. The genomic regions including 300 nucleotides upstream of the transcription start sites of genes, within the selected genomes, were analyzed as distinct DNA datasets (S9).

Protein dataset preparation
To establish the protein datasets, the protein-coding sequences (CDS) from the aforementioned genomes were operated (S9). To ensure sufficient sequence length for analysis, a minimum threshold of 100 amino acids (aa) was applied to coding sequences. Consequently, the number of proteins meeting this criterion for each genome was as follows: 46 for *C. reinhardtii*, 70 for *C. vulgaris*, 56 for *A. thaliana*, 58 for *N. tabacum*, 55 for *T. aestivum*, 57 for *O. sativa*, 64 for *Z. mays*, and 56 for *G. max*. The lengths of the protein sequences varied across these datasets, ranging from 112 to 1995 aa for *C. reinhardtii*, 103 to 1720 aa for *C. vulgaris*, 101 to 2294 aa for *A. thaliana*, 101 to 2280 aa for *N. tabacum*, 101 to 1479 aa for *T. aestivum*, 101 to 1513 aa for *O. sativa*, 101 to 1527 aa for *Z. mays*, and 101 to 2287 aa for *G. max*.

Applying data augmentation for deep learning model inputs

A method for generating overlapping subsequences was employed to address the limitations posed by the datasets with few and limited gene representations of nucleotide and amino acid sequences, thereby enabling the dataset's utility for deep learning models. The original DNA datasets comprised between 100 and 210 sequences, while the protein datasets included 46 and 70 sequences (S9).

A sliding window approach was employed to ensure that sufficient subsequences were generated, producing subsequences of 40 nucleotides with overlaps ranging from 5 to 20 nucleotides. To elaborate further, at least 50% (ranging between 50% and 87.5%) of each sequence was classified as invariant. This ensured the presence of conserved regions that maintain remarkable differences between subsequences to facilitate effective model training. Conversely, 12.5% to 50% of each sequence was treated as a variable, introducing diversity into the subsequences and enriching the input dataset for model training. This method enabled the generation of overlapping subsequences by shifting the window across the sequence, ensuring that neighboring subsequences shared common nucleotides. By varying the degree of overlap, the diversity of the generated subsequences was increased, maximizing the coverage of different regions of the sequence. Additionally, overlap patterns were applied to the left side, right side, and both sides of each sequence to achieve a comprehensive representation. Consistency was maintained in the length of the subsequences, fixed at 40 nucleotides, a key requirement for downstream DL models. All overlapping subsequences derived from a single sequence were labeled with the corresponding sequence name, and the same labeling approach was applied to subsequences generated from other sequences. This process ultimately resulted in the creation of the input dataset for subsequent model training (S10).

Implementation of CNN-LSTM hybrid model
To validate the efficiency and robustness of the data augmentation approach for DNA and protein datasets, a hybrid deep learning model combining Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers was implemented in PyTorch[47,48]. This hybrid model was designed to process the one-hot encoded nucleotide sequences as well as amino acid sequences, and then classify them according to their respective labels.

The CNN component of the CNN-LSTM architecture was responsible for extracting local features from the sequences by applying two one-dimensional convolutional layers with 64 and 128 filters, respectively. These filters captured meaningful patterns within the sequences. Each filter is followed by ReLU activation[49–51] and max-pooling[52]. The max-pooling layer was included to reduce dimensionality and focus on the most relevant features, thereby improving computational efficiency. Following feature extraction, the data were passed through a two-layer bidirectional LSTM with 128 hidden units per layer, which captured the sequential relationships between nucleotides. The bidirectional configuration of the LSTM doubled the effective hidden size, allowing the model to capture both forward and reverse patterns in the sequences. The output from the LSTM layer was then processed through two fully connected layers with dropout (0.4) to prevent overfitting, and finally mapped to the output layer for classification. For more details, the model was trained for 50 epochs using cross-entropy loss and the Adam optimizer[53], with early stopping implemented to prevent overfitting by halting training once the validation loss ceased improving. Training, validation, and test sets were created by randomly splitting the data, allocating 80% for training and 20% for testing, with a subsequent 10% validation split from the training set.

Further modifications were applied to the hybrid model utilized on the protein datasets. The model was trained using cross-entropy loss, and in cases of class imbalance, the weight of the classes was calculated using the inverse frequency method to adjust the loss accordingly.

The optimization of hyperparameters for the CNN-LSTM hybrid model was performed separately for DNA and protein datasets. For the DNA datasets, the hyperparameters were initially optimized using the *C. reinhardtii* dataset, which served as the representative dataset. The optimized hyperparameter values obtained from this process were subsequently applied to the remaining DNA datasets to ensure consistency and comparability across analyses. Similarly, for protein datasets, hyperparameters were optimized using the representative dataset, and the optimized values were then applied to the other protein datasets. This approach ensured uniformity in

hyperparameter settings across similar dataset types, facilitating a systematic evaluation of model performance across all datasets.

Analyses to evaluate the effectiveness of the augmentation approach

Evaluation and performance metrics on DNA datasets The performance of the CNN-LSTM hybrid model was evaluated on eight DNA datasets using multiple metrics to ensure a comprehensive assessment of its effectiveness with and without data augmentation. The key evaluation metric was the average test accuracy (%), calculated across all datasets over multiple trials (5 trials) and epochs.

In addition to the accuracy, the average area under the precision-recall curve (AUC-PR) was computed for each dataset, and precision-recall curves were plotted for randomly selected classes (10 classes) to visualize performance trends.

To further assess the model's predictive quality, the Pearson correlation coefficient was calculated to evaluate the relationship between model predictions and experimental results for both augmented and non-augmented datasets.

Finally, feature importance analysis was performed using SHAP (SHapley Additive exPlanations) to quantify the contributions of individual nucleotide positions to the model's predictions. SHAP values were computed using a subset of the training data as a background set and applied to selected test samples. A summary plot was generated to highlight the most impactful nucleotide positions, providing insights into the model's decision-making process and biological relevance.

Evaluation and performance metrics on protein datasets The CNN-LSTM model's performance was further evaluated on eight protein datasets using a series of analyses designed to ensure a robust assessment. To address the class imbalance, a stratified fivefold cross-validation approach was implemented, ensuring balanced representation across all folds. Evaluation metrics included weighted and macro-averaged precision, recall, and F1 scores, calculated per epoch and averaged across all folds for both augmented and non-augmented datasets.

A confusion matrix was generated to provide a detailed breakdown of the model's predictions, capturing true positives, false positives, true negatives, and false negatives for each class. The confusion matrix values from each fold were aggregated to form an overall matrix, visualized as a heatmap to highlight classification accuracy and error distribution across all classes.

Additionally, precision-recall curve analysis was conducted by collecting true labels and predicted probabilities for each class across all folds. Precision-recall curves were plotted for each class, and the average AUC-PR score along with its standard error, was reported to summarize the model's overall classification ability. Receiver operating characteristic (ROC) curve analysis was also performed to evaluate the sensitivity and positive predictive value.

Finally, the model's performance was tracked using mean loss, F1 score, and recall as primary metrics. These values were computed for each epoch and aggregated across all folds during both the training and validation phases to provide a comprehensive evaluation of the model's performance.

Data augmentation strategy for unsupervised analysis

In addition to the data augmentation method designed for deep learning applications with labeled data where each sequence is associated with a corresponding label (e.g., sequence name, species name, etc.)-an alternative approach was employed to maximize feature generation from small datasets lacking labels, enabling the analysis of unlabeled data.

K-mer analysis and identification of common patterns

Each 300-nucleotide sequence was transformed into a high-dimensional set of k-mer features to enhance feature generation from the limited dataset of genomic sequences. K-mers of varying lengths (ranging from 5 to 12 nucleotides) were extracted to analyze sequence similarity and identify shared patterns. A custom Python script was utilized to systematically identify common k-mers between each pair of sequences by comparing their respective k-mer dictionaries. For every pair, the total count of shared k-mers was calculated by aggregating the minimum occurrences of each common k-mer. The output, which included the total shared k-mer counts for each sequence pair, provided critical insights into conserved sequence motifs for further analysis.

Clustering and visualization of sequence similarities

To evaluate sequence similarities and group-related sequences, a clustering and dimensionality reduction approach was employed. A distance matrix was constructed using the total shared k-mer counts between sequence pairs, with the distance defined as inversely proportional to the number of shared k-mers. Sequence pairs with no common k-mers were assigned a distance value of zero.

K-means clustering was then performed, selecting the number of clusters based on biological relevance and the dataset's characteristics. Principal Component Analysis (PCA) was applied for dimensionality reduction, projecting the high-dimensional distance matrix into a two-dimensional space for visualization. The results of PCA, combined with the cluster assignments, were visualized using the seaborn library, with data points color-coded by cluster. This PCA plot provided an intuitive representation of sequence similarities, with distinct clusters highlighting groups of sequences sharing common k-mer patterns.

Computational resource for the model training and evaluation

The CNN-LSTM model and the additional analyses, including the data augmentation methods, confusion matrix evaluation, shape feature analysis, correlation analysis, precision-recall metrics, Receiver Operating Characteristic (ROC) curve analysis, and k-means clustering were conducted. The analyses were performed

using the free version of Google Colaboratory (Google Colab, https://colab.research.google.com), running Python version 3.10.12, an online platform that provides cloud-based access to a shared computing environment. The corresponding Python scripts are publicly available at https://github.com/MAAbbasi-Vineh/Data-Augmentation.

Data availability

Data is provided within the manuscript or supplementary information files. The source code, including the data augmentation implementation, model implementation, and optimization, is released at https://github.com/MAAbbasi-Vineh/Data-Augmentation. Further inquiries are available from the corresponding author upon reasonable request.

Received: 13 February 2025; Accepted: 21 July 2025
Published online: 25 July 2025

References

1. N. Gandhewar, A. Pimpalkar, A. Jadhav, N. Shelke, and R. Jain, Leveraging deep learning for genomics analysis: Advances and applications *Genomics at the Nexus of AI, Computer Vision, Machine Learning,* 191-225 (2025).
2. Chandrashekar, K., Niranjan, V., Vishal, A. & Setlur, A. S. Integration of artificial intelligence, machine learning and deep learning techniques in genomics: Review on computational perspectives for NGS analysis of DNA and RNA seq data. *Curr. Bioinform.* **19**(9), 825–844 (2024).
3. Mohammed, M. A., Abdulkareem, K. H., Dinar, A. M. & Zapirain, B. G. Rise of deep learning clinical applications and challenges in omics data: A systematic review. *Cell Rep. Methods* **13**(4), 664 (2023).
4. J. G. Meyer, "Deep learning neural network tools for proteomics," *Cell Reports Methods* 1 2 2021.
5. K. Choudhary *et al.*, "Recent advances and applications of deep learning methods in materials science 8 1 59, 2022.
6. Alzubaidi, L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **8**, 1–74 (2021).
7. Zhao, X. et al. A comprehensive review of methods for hydrological forecasting based on deep learning. *Water* **16**(10), 1407 (2024).
8. Bansal, A., Sharma, R. & Kathuria, M. A systematic review on data scarcity problem in deep learning: Solution and applications. *ACM Comput. Surv.* **54**(10s), 1–29 (2022).
9. Dou, B. et al. Machine learning methods for small data challenges in molecular science. *Chem. Rev.* **123**(13), 8736–8780 (2023).
10. Bennett, G. M. & Moran, N. A. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biol. Evol.* **5**(9), 1675–1688 (2013).
11. Nakabachi, A. et al. The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* **314**(5797), 267–267 (2006).
12. Green, B. Chloroplast genomes of photosynthetic eukaryotes. *Plant J.* **66**(1), 34–44 (2011).
13. Chinnery, P. F. & Hudson, G. Mitochondrial genetics. *Br. Med. Bull.* **106**(1), 135–159 (2013).
14. Hipp, J. A., Hipp, J. D., Atala, A. & Soker, S. "Functional genomics: New insights into the 'function' of low levels of gene expression in stem cells," (in eng). *Curr Genomics* **11**(5), 354–358 (2010).
15. Khoa, L. T. P. et al. Quiescence enables unrestricted cell fate in naive embryonic stem cells. *Nat. Commun.* **15**(1), 1721 (2024).
16. Viner-Breuer, R., Yilmaz, A., Benvenisty, N. & Goldberg, M. The essentiality landscape of cell cycle related genes in human pluripotent and cancer cells. *Cell Div.* **14**, 1–13 (2019).
17. Dolatabadi, S. et al. Cell cycle and cell size dependent gene expression reveals distinct subpopulations at single-cell level. *Front. Genet.* **8**, 1 (2017).
18. Pucci, B., Kasten, M. & Giordano, A. Cell cycle and apoptosis. *Neoplasia* **2**(4), 291–299 (2000).
19. Carthew, R. W. "Gene regulation and cellular metabolism: an essential partnership," (in eng). *Trends Genet* **37**(4), 389–400 (2021).
20. Dobrogojski, J., Adamiec, M. & Luciński, R. The chloroplast genome: A review. *Acta Physiol. Plant.* **42**(6), 98 (2020).
21. Abbasi-Vineh, M. A., and Emadpour, M., "The first introduction of an exogenous 5′ untranslated region for control of plastid transgene expression in *Chlamydomonas reinhardtii*," *Molecular Biotechnology* 1–14, 2024.
22. Maliga, P. & Bock, R. Plastid biotechnology: Food, fuel, and medicine for the 21st century. *Plant Physiol.* **155**(4), 1501–1510 (2011).
23. Daniell, H., Lin, C.-S., Yu, M. & Chang, W.-J. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biol.* **17**, 1–29 (2016).
24. Dyo, Y. M. & Purton, S. The algal chloroplast as a synthetic biology platform for production of therapeutic proteins. *Microbiology* **164**(2), 113–121 (2018).
25. Berahmand, R., Emadpour, M., Javaran, M. J., Haji-Allahverdipoor, K. & Akbarabadi, A. Molecular dynamics simulations of ribosome-binding sites in theophylline-responsive riboswitch associated with improving the gene expression regulation in chloroplasts. *J. Bioinform. Comput. Biol.* **22**(5), 2450023 (2024).
26. Emadpour, M., Karcher, D. & Bock, R. Boosting riboswitch efficiency by RNA amplification. *Nucleic Acids Res.* **43**(10), e66–e66 (2015).
27. Mumuni, A. & Mumuni, F. Data augmentation: a comprehensive survey of modern approaches. *Array* **16**, 100258 (2022).
28. Shorten, C., Khoshgoftaar, T. M. & Furht, B. Text data augmentation for deep learning. *Journal of Big Data* **8**(1), 101 (2021).
29. Lee, N. K., Tang, Z., Toneyan, S. & Koo, P. K. EvoAug: Improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. *Genome Biol.* **24**(1), 105 (2023).
30. Duncan, A. G., Mitchell, J. A. & Moses, A. M. Improving the performance of supervised deep learning for regulatory genomics using phylogenetic augmentation. *Bioinformatics* **40**(4), btae190 (2024).
31. Montesinos-López, O. A. et al. Data augmentation enhances plant-genomic-enabled predictions. *Genes* **15**(3), 286 (2024).
32. Kircher, M. et al. Augmentation of transcriptomic data for improved classification of patients with respiratory diseases of viral origin. *Int. J. Mol. Sci.* **23**(5), 2481 (2022).
33. Wen, B. et al. Deep learning in proteomics. *Proteomics* **20**(21–22), 1900335 (2020).
34. Chen, B., Li, N. & Bao, W. CLPr_in_ML: Cleft lip and palate reconstructed features with machine learning. *Curr. Bioinform.* **20**(2), 179–193 (2025).
35. Xu, C., Coen-Pirani, P. & Jiang, X. Empirical study of overfitting in deep learning for predicting breast cancer metastasis. *Cancers* **15**(7), 1969 (2023).
36. X. Ying, "An overview of overfitting and its solutions," In *Journal of physics: Conference series* 1168 022022: IOP Publishing. 2019.
37. Alzubaidi, L. et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J. Big Data* **10**(1), 46 (2023).
38. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**(141), 20170387 (2018).
39. Sapoval, N. et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* **13**(1), 1728 (2022).

40. Yousef, M. & Allmer, J. Deep learning in bioinformatics. *Turk. J. Biol.* **47**(6), 366–382 (2023).
41. Park, Y., Hauschild, A.-C. & Heider, D. Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing. *NAR Genom. Bioinform.* **3**(4), 104 (2021).
42. Cai, C. et al. Transfer learning for drug discovery. *J. Med. Chem.* **63**(16), 8683–8694 (2020).
43. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**(7965), 616–624 (2023).
44. B. L. Nielsen and N. Ahmad, "Editorial: Advances in plastid biology and its applications, volume II," (in English), Editorial 14 2023-May-31 2023.
45. Wang, L. et al. Complete sequence and analysis of plastid genomes of two economically important red algae: Pyropia haitanensis and Pyropia yezoensis. *PLoS ONE* **8**(5), e65902 (2013).
46. Yang, X.-F. et al. PBR1 selectively controls biogenesis of photosynthetic complexes by modulating translation of the large chloroplast gene Ycf1 in Arabidopsis. *Cell discov.* **2**(1), 1–19 (2016).
47. S. Imambi, K. B. Prakash, and G. Kanagachidambaresan, "PyTorch. programming with tensorFlow: solution for edge computing applications," Berlin/Heidelberg, Germany Springer87–104 2021.
48. A. Testas, "Deep learning with PyTorch for classification," In *building scalable deep learning pipelines on AWS: Develop, train, and deploy deep learning models*: Springer 321–429 2024
49. Y. Bai, "RELU-function and derived function review," In *SHS Web of Conferences*, 144 02006: EDP Sciences 2022.
50. T. Szandała, "Review and comparison of commonly used activation functions for deep neural networks," *Bio-Inspired Neurocomputing* 203–224, 2021.
51. Schmidt-Hieber, J. Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Stat.* **48**(4), 1875–1897 (2020).
52. D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24–26, 2014, Proc. 9* 364–375: Springer. 2014
53. D. Kingma and B. Jimmy, " Adam: a method for stochastic optimization," In *Presented at International Conference on Learning Representations (ICLR)*, 2015.

Author contributions

M.A.A.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft; S.R.: Conceptualization, Data curation, Investigation, Methodology, Resources, Validation, Writing – original draft; K.K.: Conceptualization, Software, Validation, Writing – original draft; M.E.: Conceptualization, Validation, Visualization, Writing – original draft.

Funding

The current study was financially supported as part of a Doctoral thesis by the university (approved proposal No.: 86132).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-12796-9.

Correspondence and requests for materials should be addressed to K.K. or M.E.