BMC Biology

RESEARCH

Open Access

Multimodal deep learning for allergenic proteins prediction

Lezheng Yu[1,2†], Yuxin Luo[3†], Shiqi Wu[4], Siyi Chen[4], Li Xue[5], Runyu Jing[1*] and Jiesi Luo[6,7*]

Abstract

Background Accurate prediction of allergens is essential for identifying the sources of allergic reactions and preventing future exposure to harmful triggers; however, the limited performance of current prediction tools hinders their practical applications.

Results Here, we present Multimodal-AlgPro, a unified framework based on a multimodal deep learning algorithm designed to predict allergens by integrating multiple dimensions, including physicochemical properties, amino acid sequences, and evolutionary information. An exhaustive search strategy for model combinations has also been introduced to ensure robust allergen prediction by thoroughly exploring every possible modality configuration to determine the most effective framework architecture. Additionally, identifying explainable sequence motifs and molecular descriptors from these models that facilitate epitope discovery is of interest. Because it leverages diverse heterogeneous features and our improved multimodal data fusion, Multimodal-AlgPro outperformed several existing methods, demonstrating its potential to significantly advance the accuracy of allergen prediction.

Conclusions Overall, Multimodal-AlgPro is a valuable tool for deciphering the mechanisms of allergic responses and offers novel insights on epitope design, with applications in both public health and industrial sectors.

Keywords Allergen, Multimodal deep learning, Epitope, SHAP, UMAP

Background

Allergy is a significant dysfunction of the immune system, marked by an exaggerated or inappropriate response to substances that are typically harmless. This dysfunction plays a central role in triggering chronic inflammation, contributing to a growing global health burden that affects millions of individuals across diverse populations [1]. Although current diagnostic and therapeutic strategies are focused on alleviating symptoms, they often fail to provide long-term relief from allergic conditions [2]. To control allergies effectively, it is crucial to identify the causative allergens, prevent re-exposure, and intervene in the immune pathways that drive allergic reactions [3]. However, this task becomes particularly challenging when allergens are rare or not yet identified. Moreover, the increasing presence of modified substances in food products, pharmaceuticals, cosmetics, and other industrial goods has made the identification and assessment

†Lezheng Yu and Yuxin Luo contributed equally to this work.

*Correspondence:
Runyu Jing
jingry@gznc.edu.cn
Jiesi Luo
ljs@swmu.edu.cn
[1] School of Mathematics and Big Data, Guizhou Education University, Guiyang 550018, China
[2] School of Chemistry and Materials Science, Guizhou Education University, Guiyang, Guizhou 550018, China
[3] School of Medical Information and Engineering, Southwest Medical University, Luzhou, Sichuan 646000, China
[4] School of Clinical Medical Sciences, Southwest Medical University, Luzhou, Sichuan 646000, China
[5] School of Public Health, Southwest Medical University, Luzhou 646000, China
[6] School of Basic Medical Science, Southwest Medical University, Luzhou, Sichuan 646000, China
[7] State Key Laboratory of Southwestern Chinese Medicine Resources, Chengdu University of Traditional Chinese Medicine, Chengdu 610075, China

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

of potential allergens even more pressing [4–7]. Considering the vast number of potential allergens and the high costs associated with experimental validation, it is impractical to test every allergen using experimental methods [8, 9]. Therefore, computational approaches for allergen prediction, conducted before clinical or experimental validation, have become indispensable.

Early allergen prediction methods relied on sequence alignment techniques, using sliding windows of fixed lengths to compare protein sequences and identify structural similarities between new proteins and known allergens [10–14]. However, these methods faced several limitations, including a high false-positive rate and inefficient resource utilization due to inaccurate allergen evaluations, which hindered their scalability for large-scale predictions [10]. Alternatively, machine learning methods such as support vector machines and random forests take advantage of labeled datasets to model the complex relationships between handcrafted allergen features and their associated properties [15–25]. These models, designed for structured, tabular data, depend on expert-derived features like amino acid or protein descriptors, motif enrichment, and epitope information. While effective, this reliance on manual feature extraction also restricts scalability and may miss crucial data interactions [26]. Furthermore, due to limitations in dataset size, the relatively shallow nature of allergen identification, and model complexity, these methods often suffer from issues related to prediction accuracy and generalization, resulting in restricted practical applicability [27].

Recent advancements in deep learning have paved the way for end-to-end prediction models that utilize convolutional neural networks, recurrent neural networks, and transformers to directly identify allergens from amino acid sequences [27–34]. Compared to traditional machine learning methods, deep learning enhances predictive accuracy by automating feature extraction and capturing complex patterns in data, thus eliminating the need for manual feature engineering [29]. However, despite these advantages, deep learning models require vast amounts of data for effective training, and the persistent issue of small dataset sizes often leads to overfitting, which undermines their generalizability and reduces their reliability in real-world applications [35]. Furthermore, deep learning-based sequence encoding methods typically neglect the evolutionary conservation of allergens, even though strong evidence links this information to their structural and functional characteristics [36]. Although previous studies have attempted to incorporate the evolutionary perspectives of allergenic proteins, most predictive methods are either based on sequence or conservation alone and do not address how these two data sources can be integrated to develop multimodal deep learning models. This oversight results in the loss of valuable information and introduces unnecessary computational complexity, diminishing the overall performance of the developed model.

The gap between the urgent need for allergen prediction and the shortcomings of existing computational methods highlights a significant research challenge. To address this, we present Multimodal-AlgPro, a deep learning model based on multimodal data fusion and ensemble model strategies, which enables comprehensive learning and integration of allergen features from multiple modalities, including sequence, composition, physicochemical properties, and evolutionary information, to increase robustness and accuracy of allergen predictions. We use a character-level convolutional network to encode amino acid sequence data and a multi-layer densely connected network to extract deeper insights from molecular features. Through an extensive evaluation of 14 distinct unimodal models and 2047 potential multimodal combinations, we systematically identified the most successful multi-modality configuration. We also applied an interpretability and quantitative analysis method to uncover novel patterns both within and across modalities, enhancing our understanding of feature contributions and prediction outcomes. Lastly, we have proposed a five-modality framework that considers complementary information from diverse data sources and validated its effectiveness on a representative independent dataset. Multimodal-AlgPro achieved state-of-the-art performance across various evaluation metrics, demonstrating a remarkable improvement over previous prediction methods. This success further confirms its potential to significantly advance the field and contribute to the development of new methodologies for allergen prediction.

Results

Deep-learning-based multimodal integration

To advance allergenicity prediction methods, we propose a deep-learning-based multimodal fusion algorithm that integrates both protein sequences and descriptors to predict and interpret allergenic potential (Fig. 1). This approach captures complementary contextual information from diverse data types, thereby enhancing predictive accuracy and interpretability. In this framework, each amino acid in the protein sequence is encoded as a 20-dimensional vector using one-hot encoding, with convolutional kernels serving as motif detectors to extract relevant sequence features. Meanwhile, proteins are represented through handcrafted descriptors, processed by a neural network to generate deep feature representations. The outputs from each modality are then merged through an early fusion operation, creating a unified protein

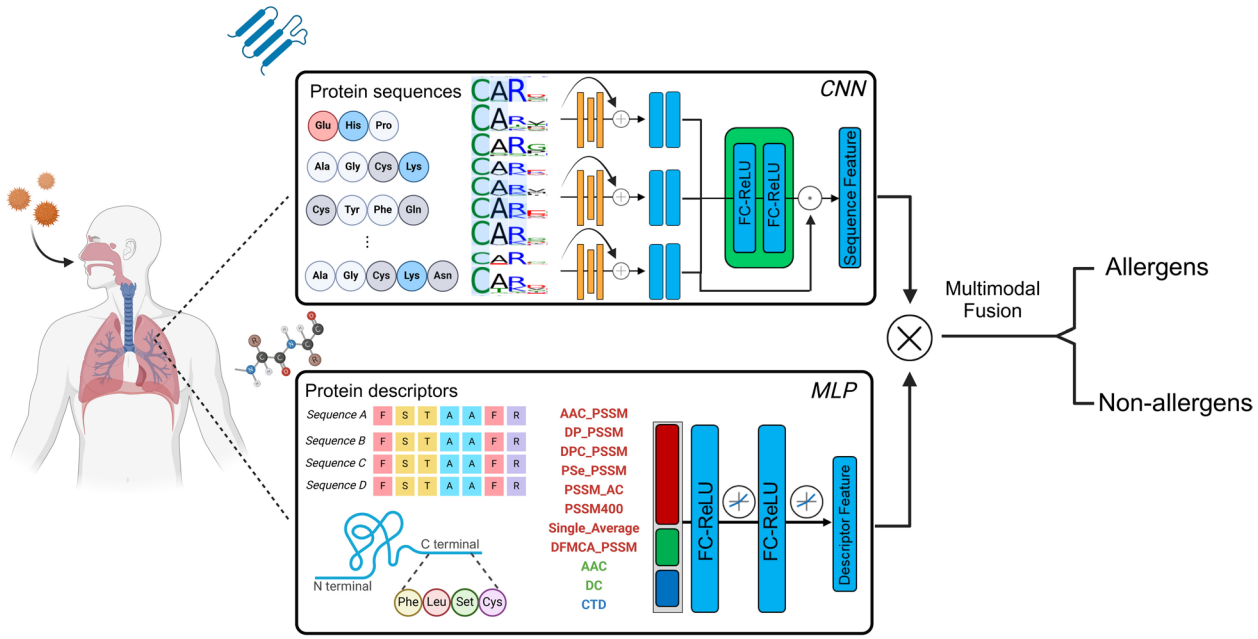


Fig. 1 The overview of multimodal deep learning algorithm. Allergens and non-allergens in the form of protein sequences with their corresponding molecular features are used as inputs in our algorithm. Our multimodal algorithm consists of three neural network modules: (1) a CNN for processing protein sequences, (2) an MLP for processing molecular features, and (3) a multimodal fusion layer that computes the fusion product to model pairwise feature interactions between protein sequences and molecular features

representation suitable for input into a classification model. This multimodal network not only supports large-scale protein prediction tasks but also explains the relationships between sequence motifs, descriptor features, and their interactions in distinguishing allergens from non-allergens. To ensure robust performance, all models were evaluated across five iterations of randomized training, validation, and testing, with ablation analyses comparing unimodal and multimodal predictions. Further analyses focused on the model combination strategy, investigating both local and global interactions between sequence and molecular features, identifying epitope patterns, and examining shifts in feature importance between unimodal and multimodal interpretations.

Benchmarking unimodal deep learning models for allergenic protein prediction

This section presents a comparative analysis of fourteen unimodal deep learning models, each trained independently on a single feature type to classify allergens and non-allergens. These models exhibit varying levels of complexity, comprising a CNN-based model and a BiLSTM-based model that both utilize one-hot encoded protein sequences, a Transformer-based model that employs dictionary-encoded sequences, and eleven MLP-based models built upon diverse descriptor features. To ensure robust and fair comparisons, each model was rigorously optimized through hyperparameter tuning (Additional file 1: Fig. S1). As summarized in Fig. 2, all three text-modality models (CNN, BiLSTM, and Transformer) were designed to extract allergenicity-related patterns directly from protein sequences, but differ in both complexity and performance. The CNN model achieved the strongest overall results across all evaluation metrics, benefiting from its simple yet effective architecture. The BiLSTM model showed intermediate performance, outperforming the Transformer. Its structure includes a reshaping layer (1000×20), a bidirectional LSTM layer with 256 units, global max pooling, a dense layer with 256 hidden units, dropout (rate 0.2), ReLU activation, and a sigmoid output. The Transformer model, although the most complex in architecture, consisting of an Embedding layer with dimension 128, a Transformer encoder with 2 attention heads and an intermediate size of 512, followed by pooling and output layers, yielded the lowest performance. These findings suggest that simpler architectures such as CNNs may generalize better in allergenicity prediction from sequence data. In contrast to the text-modality models, the feature-modality MLP models showed diverse performance depending on the selected descriptor features. Notably, the MLP architecture with PSSM-based features (PSSM_AC and

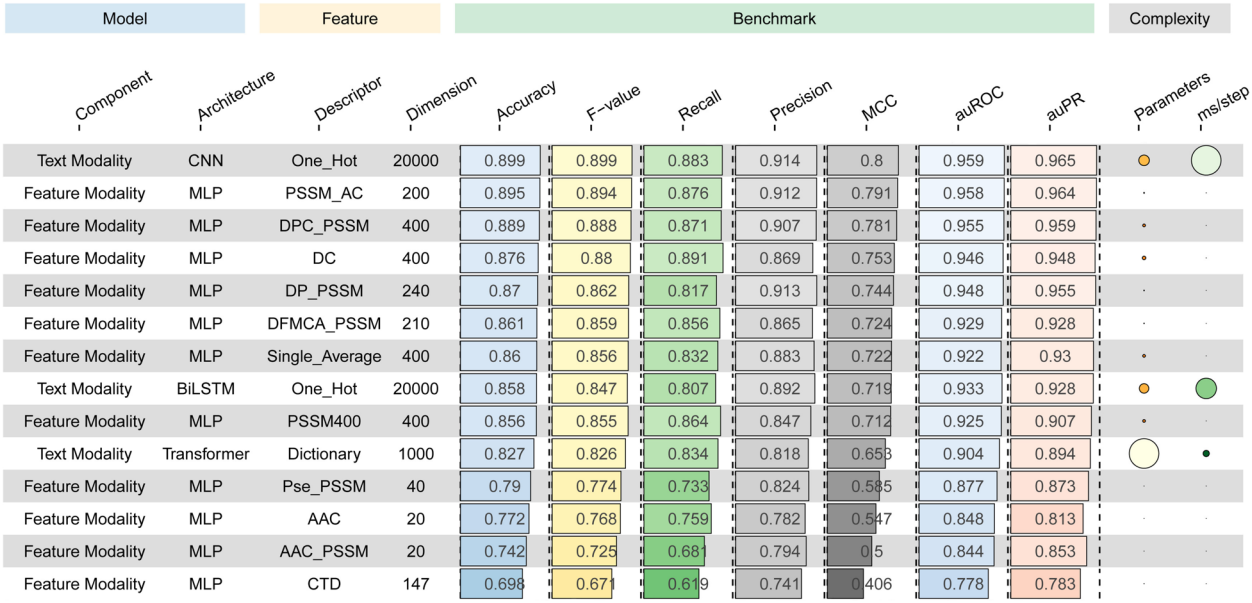


Fig. 2 Characterization and evaluation of the 14 unimodal deep learning models in this study. The unimodal models are characterized based on their modality components, network architectures, input feature types, and dimensions. The overall evaluation results are presented in terms of two criteria: classification performance, including accuracy, *F*-value, recall, precision, MCC, auROC, and auPR, with each metric represented by the length of corresponding bars indicating the relative performance levels; and computational resource requirements, quantified as normalized complexity, which includes the number of parameters and runtime in ms/step (with maximum values normalized to 1)

DPC_PSSM) achieved high accuracy, closely approaching that of the CNN model, whereas those using AAC and CTD features performed the worst among all unimodal models. Furthermore, the effect of network architecture on the feature modality was examined by comparing the predictive performance of MLP and CNN across 11 types of protein descriptors. MLP consistently outperformed CNN for most descriptors, with a few exceptions including DPC_PSSM, Pse_PSSM, and AAC_PSSM (Additional file 2: Fig. S2). This can be attributed to the tabular nature of protein descriptor data, which lacks the local spatial dependencies inherent to sequential inputs. Although CNNs are effective in automatic feature extraction, their advantage diminishes when applied to structured tabular inputs that do not exhibit spatial locality. In addition, the ROC and PR curves further illustrate these performance differences (Additional file 3: Fig. S3 and Additional file 4: Fig. S4), with the CNN model maintaining consistently high accuracy, while other models performed in line with expectations based on descriptor complexity. Overall, this benchmarking study of unimodal deep learning models establishes a clear understanding of the predictive strengths and limitations of each modality, providing a valuable baseline for the subsequent investigation into multimodal integration strategies.

Multimodal integration improves allergenic protein prediction

Following the benchmarking of unimodal models, the investigation was extended to multimodal integration methods to enhance allergenic protein prediction. The notable performance differences among unimodal models suggested that a straightforward combination of all models would be insufficient to achieve optimal accuracy. Given the fourteen unimodal models under consideration, identifying the most effective combinations presented a significant challenge. To address this, a systematic approach was adopted to examine all possible model combinations. The CNN was designated as the core of the multimodal architecture and was integrated with a series of MLP architectures, each corresponding to one of the eleven distinct protein descriptors. This strategy created a broad search space of 2047 unique combinations ($2^{11}-1$), allowing us to explore diverse model interactions and leverage each model's unique strengths to boost predictive accuracy. In Fig. 3, the top 40 combinations are ranked according to four evaluation metrics: accuracy, F-value, precision, and MCC, with detailed results for all 2047 combinations provided in Additional file 5: Table S1. Evolutionary-based descriptors contributed more significantly to model combinations compared to composition- and physicochemical-based descriptors. Among the evolutionary-based

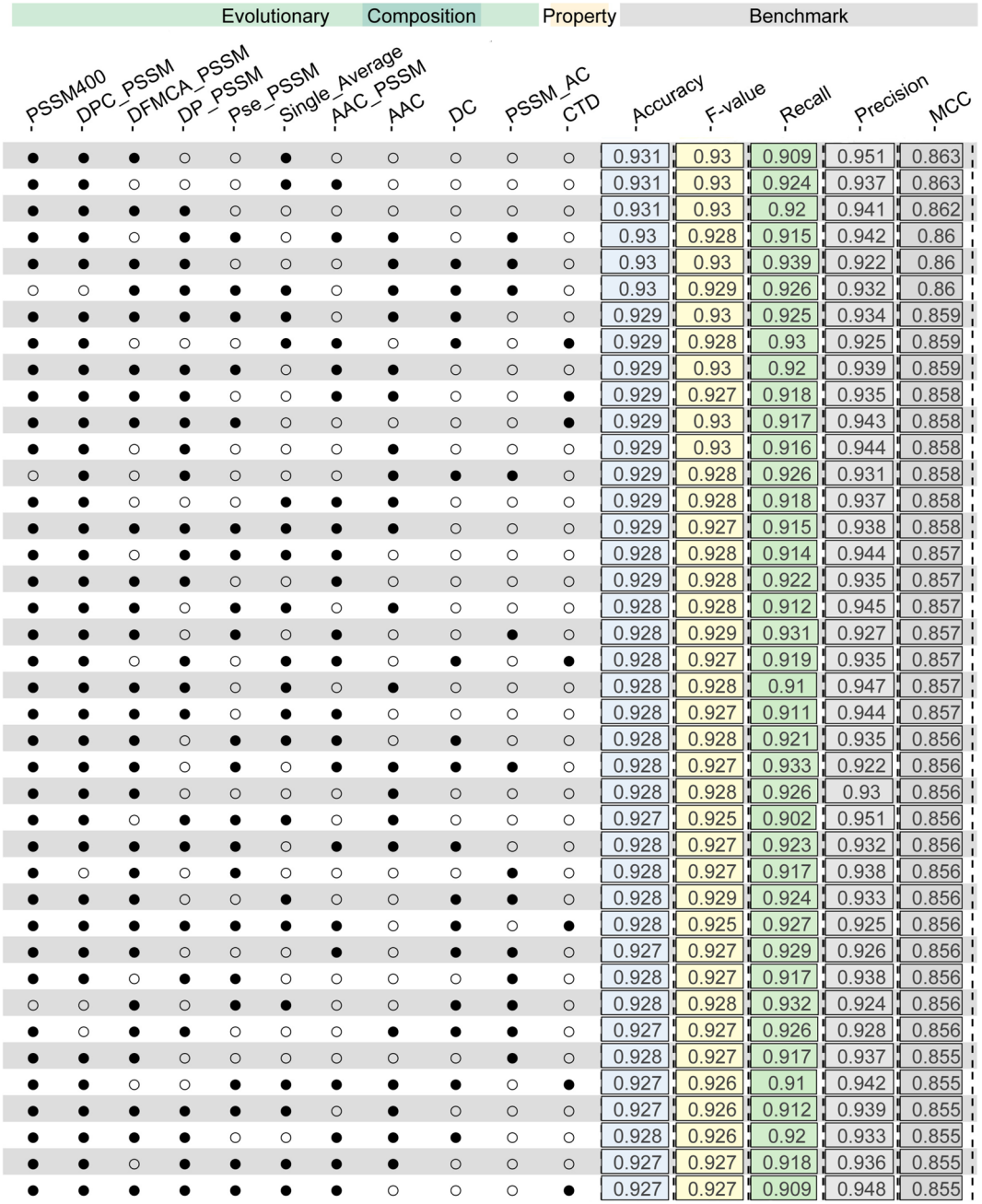


Fig. 3 Top 40 multimodal combinations and their corresponding performance metrics. Each row represents a unique combination, with performance metrics including accuracy, *F*-value, precision, and MCC. The models are categorized by their input descriptor types, with ○ and ● markers indicating the exclusion or inclusion of specific features in each combination. Metrics are averaged over 5 runs; individual data values are provided in Additional file 5: Table S1

descriptors, PSSM400 appeared in 37 of the top 40 combinations, while DPC_PSSM was included in 36. In contrast, PSSM_AC, which had the second-highest performance in the unimodal benchmarking, was present in only 13 of the top combinations. These results reveal the limitations of relying solely on unimodal performance to identify optimal multimodal combinations and demonstrate the advantages of our combinatorial approach. The best-performing multimodal combination identified was "CNN (One_Hot) + MLP (DFMCA_PSSM) + MLP (DPC_PSSM) + MLP (PSSM400) + MLP (Single_Average)," which achieved average scores of accuracy (93.1%),

F-value (92.9%), recall (90.9%), precision (95.1%), MCC (0.863), auROC (0.976), and auPR (0.977) (Table 1). The ROC and PR curves for this combination are displayed in Additional file 6: Fig. S5 and Additional file 7: Fig. S6. The effectiveness of multimodal integration in enhancing predictive performance was systematically assessed. Using the top-performing unimodal CNN model (average MCC = 0.800; Fig. 2) as a reference, predictions from unimodal and multimodal models were compared. Among the 2047 multimodal combinations evaluated, 1937 outperformed the CNN baseline, while only 110 yielded inferior performance (Fig. 4). These results provide strong evidence that multimodal integration substantially improves allergenic protein prediction.

Multimodal interpretability for model visualization and joint sequence-evolutionary pattern discovery

Allergens and non-allergens exhibit distinct structures and functions, and it is expected that these characteristics are captured within the embeddings generated by

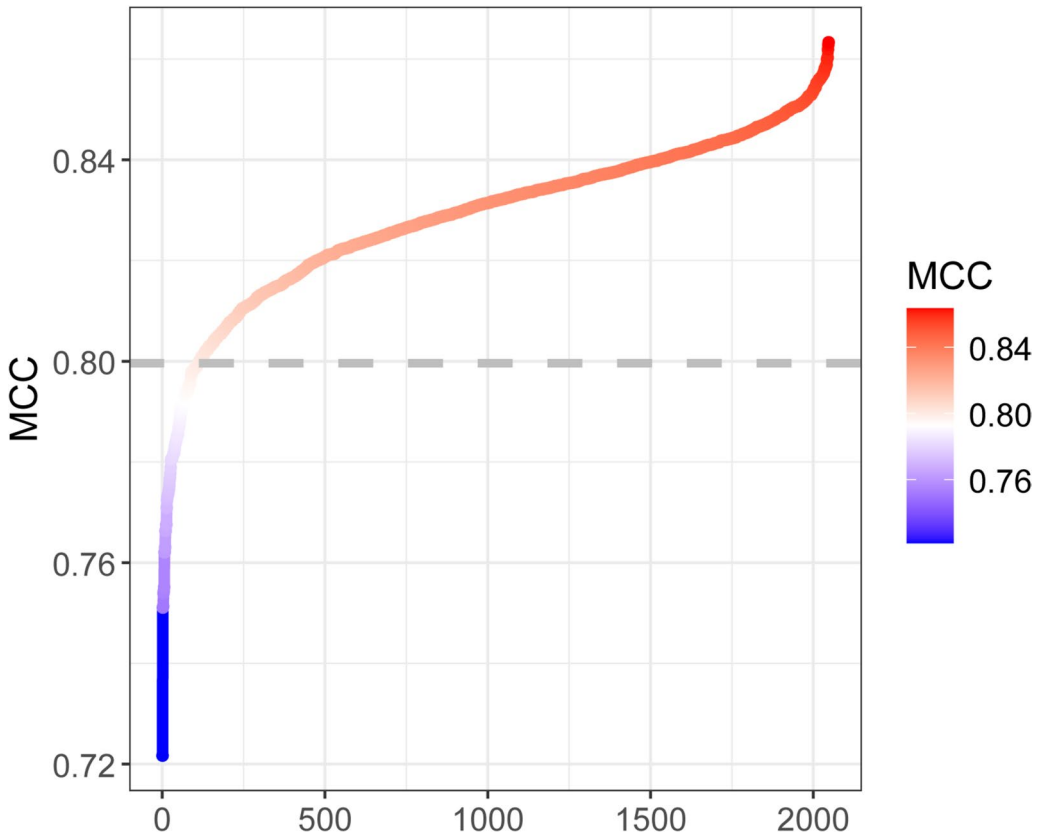


Fig. 4 Performance ranking of 2047 multimodal combinations for allergenic protein prediction. The gray dashed line represents the predictive performance of the CNN unimodal model

Table 1 Performance of the best multimodal combination architecture "CNN (One_Hot) + MLP (DFMCA_PSSM) + MLP (DPC_PSSM) + MLP (PSSM400) + MLP (Single_Average)" on the training dataset

Round	ACC (%)	*F*-value (%)	Recall (%)	PRE (%)	*MCC*	auROC	auPR
1	92.7	92.6	91.1	94.2	0.854	0.973	0.978
2	93.2	93.2	89.4	97.2	0.868	0.977	0.976
3	92.5	92.3	91.3	93.4	0.850	0.977	0.978
4	93.6	93.1	90.5	95.9	0.872	0.976	0.972
5	93.6	93.5	92.3	94.8	0.872	0.975	0.979
Average	93.1	92.9	90.9	95.1	0.863	0.976	0.977

our proposed models using raw feature inputs. To examine distributional patterns within the known allergen repertoire, feature representations from the penultimate layers of the twelve unimodal models involved in the multimodal combinations were extracted and visualized using scatter plots to depict the separation between allergens and non-allergens. Dimensionality reduction is performed using UMAP to map the embeddings onto a two-dimensional plane. To assess the impact of the learning process, both model architecture and feature type were examined through visualization. Figure 5 shows the results, where the embedding spaces for CNN (One_Hot), MLP (PSSM_AC), MLP (DPC_PSSM), MLP (DC), MLP (DP_PSSM), MLP (DFMCA_PSSM), MLP (Single_Average), and MLP (PSSM400) effectively organize protein types into distinct clusters based on their sequential, compositional, and evolutionary properties. In contrast, the embeddings for MLP (Pse_PSSM), MLP (AAC), MLP (AAC_PSSM), and MLP (CTD) struggle to create clear separations between the protein types, indicating a lack of distinguishability. This observation highlights the limitations of these features, particularly their inability to distinguish between allergens and non-allergens, providing a valuable perspective for assessing their effectiveness. Additionally, these comparative visualizations correlate with the performance of the corresponding models in unimodal benchmarking, elucidating the reasons behind their varying performance.

To interpret and validate the multimodal combination, gradient-based SHAP was applied to the trained CNN and MLP subnetworks, providing insights into how sequence and evolutionary features contribute to the prediction of allergens and non-allergens. Visualizations of individual feature contributions within the multimodal framework are referred to as local interpretability, while the analysis of contributions from each modality is termed global interpretability. Figure 6 presents both global and local interpretability results for the multimodal model combining "CNN (One_Hot) + MLP (DFMCA_PSSM) + MLP (DPC_PSSM) + MLP (PSSM400) + MLP (Single_Average)." Global interpretability first highlights the varying contributions of each

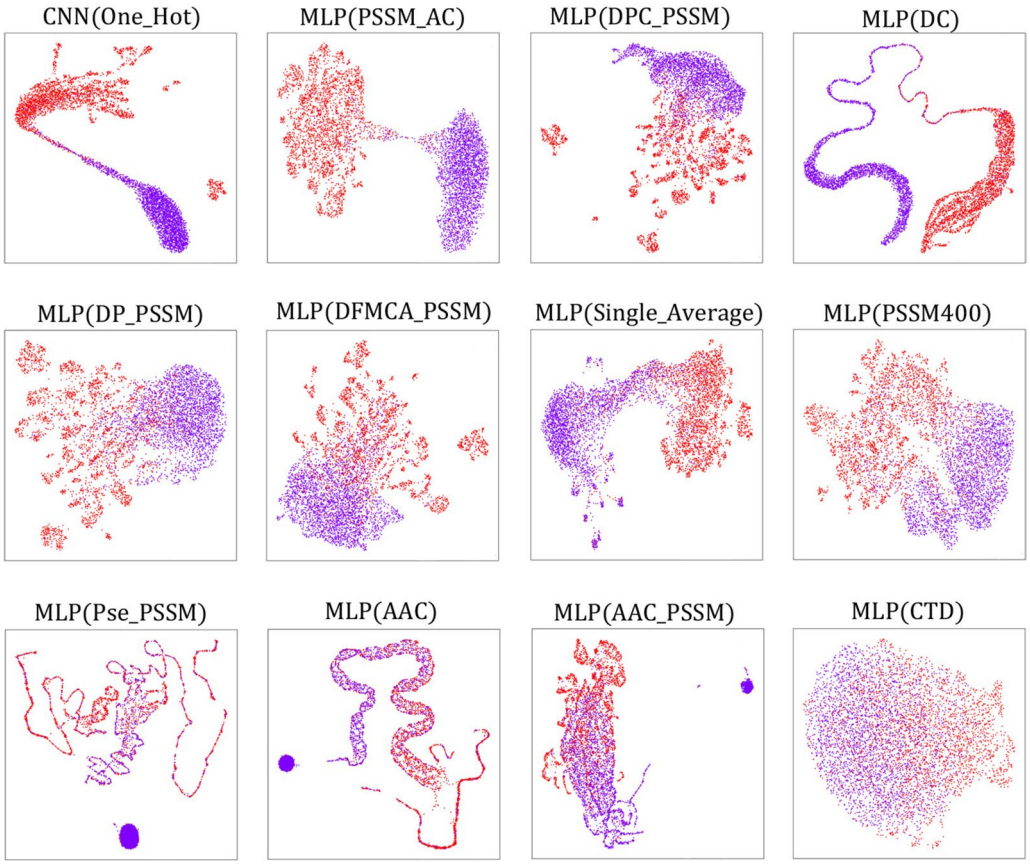


Fig. 5 UMAP visualization of the last hidden layer representations learned by various model architectures and feature types. The colored point clouds represent different protein categories, illustrating how the algorithm clusters the proteins. Allergens are denoted in red, while nonallergens are displayed in purple

module within the multimodal structure, with CNN (One_Hot) exerting the greatest effect (Fig. 6A), consistent with unimodal benchmarking results. Additionally, MLP (DPC_PSSM) demonstrates a significant effect as the second most influential modality, followed by MLP (PSSM400), MLP (Single_Average), and MLP (DFMCA_PSSM), which contribute progressively less. Therefore, local interpretability primarily focuses on the CNN (One_Hot) and MLP (DPC_PSSM) modules. The relative contribution of each evolutionary feature to model prediction in MLP (DPC_PSSM) is first quantified, as calculated by integrated gradients. These features reflect the evolutionary conservation of dipeptide pairs within the protein. Figure 6B shows that 7 out of the top 10 features correspond to hydrophobic amino acid pairs, while the remaining features belong to hydrophilic and special amino acid pairs. This suggests that the evolutionary conservation of hydrophobic amino acid pairs may be crucial for distinguishing allergens from non-allergens, as these interactions are often linked to the stable, functional structures typical of allergenic proteins. In parallel with the analysis of evolutionary features, the contribution of sequence features was also examined. SHAP attribution was employed to identify key sequence regions that significantly influence the CNN's predictions. For each allergen, a 31-amino acid subsequence was extracted, centered at the position with the highest SHAP value, including 15 residues upstream and 15 downstream, to highlight regions most critical to the convolutional decision process. These subsequences were pooled, aligned, and analyzed to identify conserved motifs across the sequences using the XSTREME tool [37]. The de novo discovered motifs, KYMVIQGEA and IPAARLFK (Fig. 6C), matched two linear epitopes from the Immune Epitope Database (IEDB) [38]: GIKYMVIQGEAGAVI and IPAARLFKAFIL, corresponding to Bet v 2 and Bet v 1, respectively. Bet v 1 is the dominant allergen in birch pollen, with IgE recognition observed in over 95% of sensitized individuals, and has also been identified as the most relevant T cell epitope involved in cross-reactivity with Api g 1 [39, 40]. The IPAARLFKAFIL epitope, corresponding to amino acid positions 13 to 24, represents a previously reported T-cell activating region within the Bet v 1 molecule, as shown in Additional file 8: Fig. S7. In contrast, Bet v 2, a profilin, is a minor allergen, with IgE reactivity reported in approximately 22% of birch pollen-allergic patients [41]. To investigate correlations among features across multiple modalities, a Top-K feature co-selection matrix was constructed by counting pairwise co-occurrences among the top 5% SHAP-ranked features in each sample. Separate matrices were generated for positive and negative samples, with SHAP values for negatives multiplied by −1 to reflect opposing contributions. This approach revealed cross-modality associations, particularly between sequence-based and evolutionary features. Strong inter-feature correlations were observed, emphasizing substantial connectivity between different modules. Notably, the first 200 amino acids of protein sequences exhibited pronounced correlations both within the sequence modality and with features from other modalities, suggesting that this region carries key information for allergen prediction (Fig. 6D). These findings highlight the significance of the first 200 amino acids in protein sequences, revealing their crucial role in allergen prediction and demonstrating the model's ability to capture key sequence-evolutionary patterns.

Comparison with other state-of-the-art methods

To further assess the classification performance of the proposed multimodal architecture, an independent dataset comprising 710 allergens and 710 non-allergens was used for evaluation. Model predictions were compared against ground-truth labels based on standard annotations provided by human experts. The class-wise prediction accuracy is displayed in the confusion matrix (Fig. 7A), showing excellent alignment with reference labels across both protein classes, with only 63 allergens and 24 non-allergens misclassified. To gain deeper insights into the internal representations learned by the multimodal architecture, UMAP was applied for dimensionality reduction (Fig. 7B). Each point represents a protein, projected from the 498-dimensional output of the multimodal fusion layer, comprising 128-, 70-, 100-, 100-, and 100-dimensional representations for CNN (One_Hot), MLP (DFMCA_PSSM), MLP (DPC_PSSM), MLP (PSSM400), and MLP (Single_Average), respectively, into

(See figure on next page.)

Fig. 6 Multimodal interpretability and association discovery. **A** Global modality contributions to allergen and non-allergen predictions based on multimodal interpretability. The contribution of each modality to the final prediction is illustrated through a bar plot of their corresponding summed SHAP values. **B** Local model explanation for the top ten features in the DPC_PSSM modality. **C** The top two sequence motif patterns revealed by the CNN modality. **D** The model can uncover associations between multimodal data interconnection, contributing to the discovery of predictive features within and across modalities. These associations are visualized in a heatmap for the corresponding allergen (left) and non-allergen (right) predictions. The color of each dot reflects the strength of the association across all features (high-to-low depicted as red-to-blue)

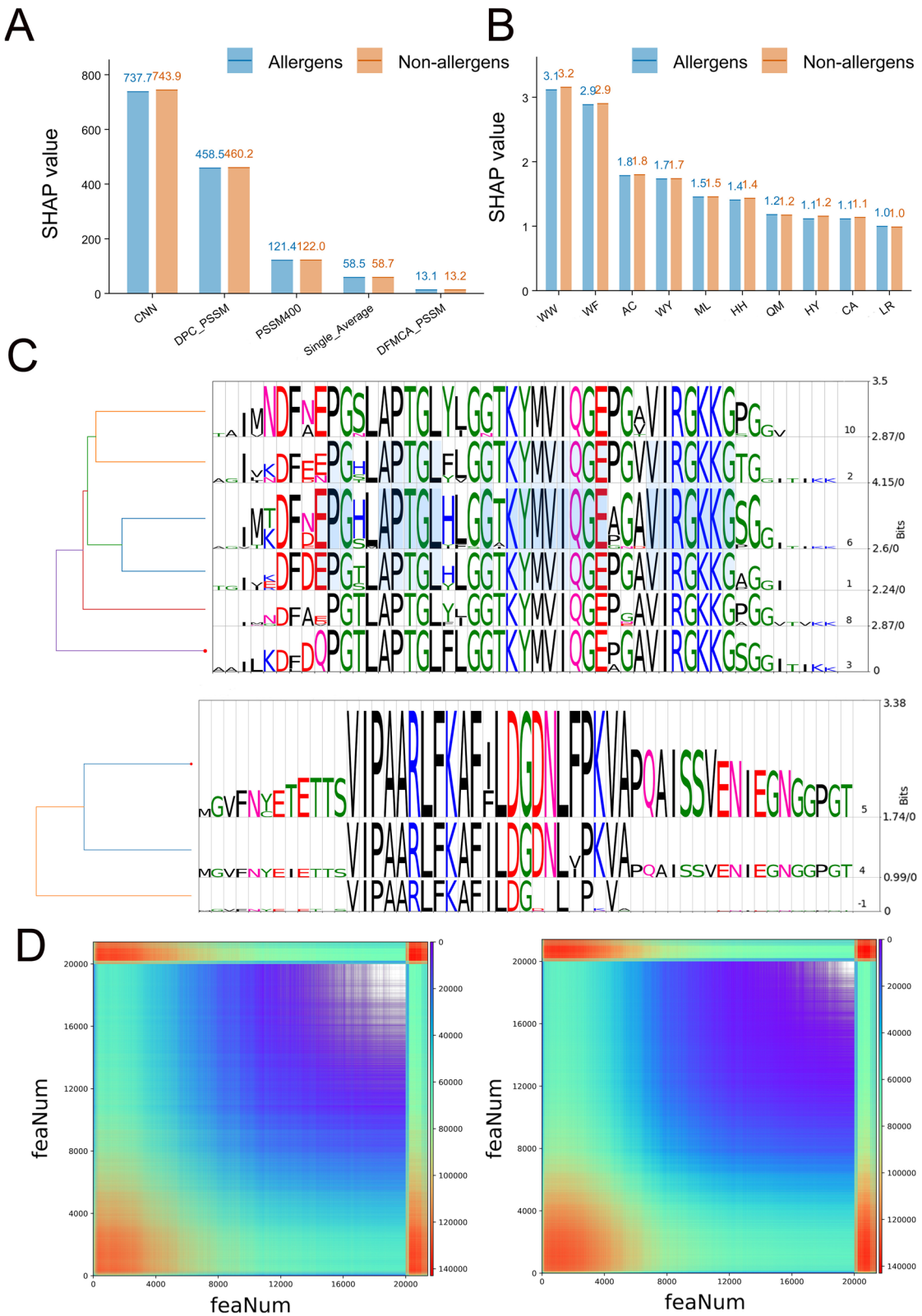


Fig. 6 (See legend on previous page.)

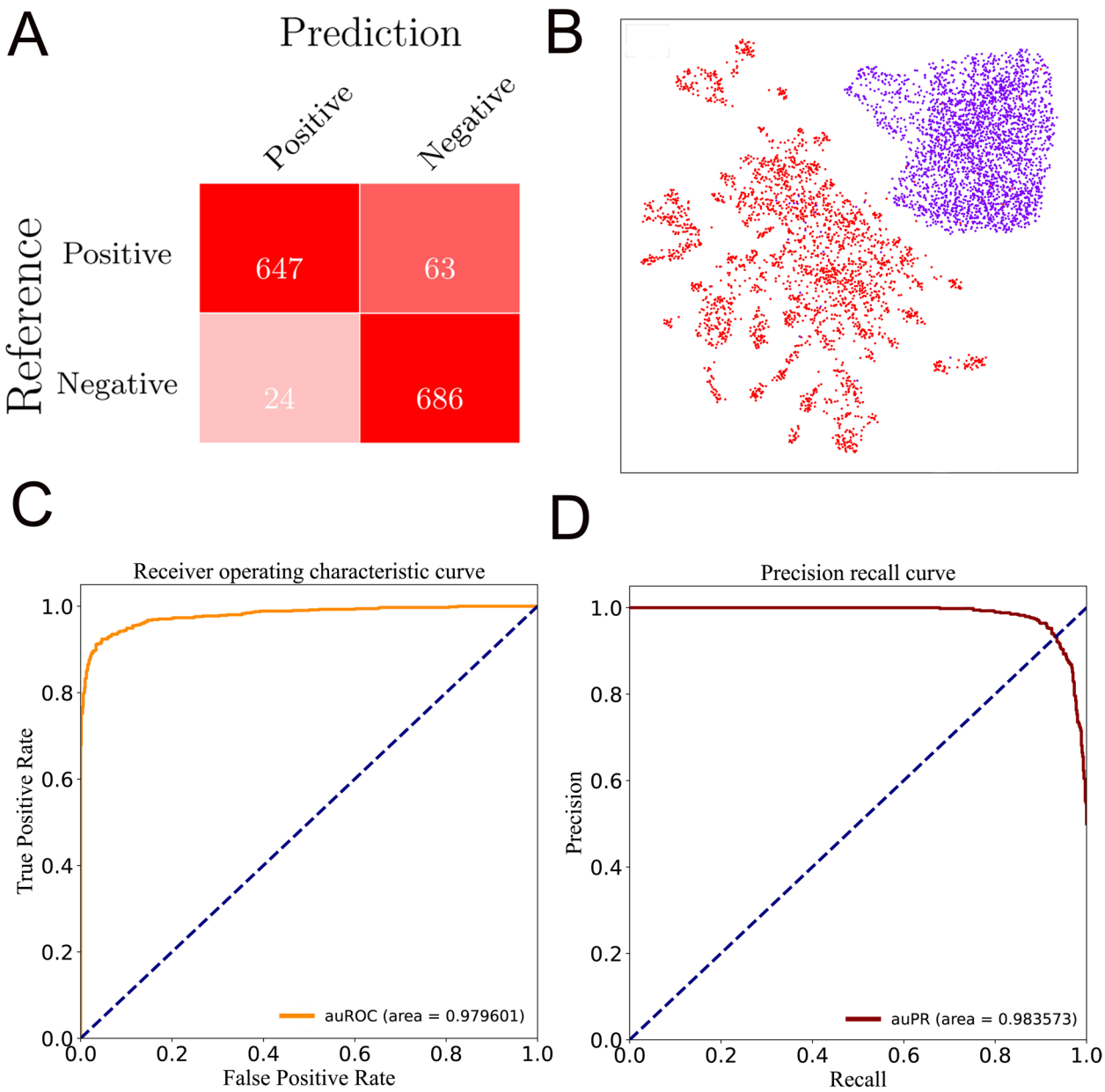


Fig. 7 Performance of Multimodal-AlgPro in differentiating 710 allergens from 710 non-allergens. **A** Confusion matrix illustrating the predictions made by Multimodal-AlgPro in comparison to the ground-truth reference labels obtained from the independent test dataset. **B** UMAP visualization of the representations learned from the multimodal fusion layer of Multimodal-AlgPro. The model achieved an area under the ROC curve (auROC) of 0.980 (**C**) and an area under the precision-recall curve (auPR) of 0.984 (**D**) for this binary classification task

two-dimensional space. Distinct clusters for allergens and non-allergens are evident, with a separation that outperforms unimodal models. The ROC and PR curves, generated by adjusting the binary classification threshold, are presented in Fig. 7C and D, respectively. The auROC and auPR scores, measured at 0.980 and 0.984, indicate the model's strong ability to differentiate allergens from non-allergens relative to the ground truth. Due to its superior performance, this multimodal architecture was selected as the final predictor for comparison with existing methods and is designated as Multimodal-AlgPro in this study.

Table 2 presents a performance comparison between the Multimodal-AlgPro and several baseline methods from existing literature, including ProAll-D [30], AllergenPF [19], AllerTOPv.2 [18], AlgPred2.0 [21], DeepAlgPro [27], PreAlgPro [33], and AllergyPred [42]. Baseline

Table 2 Benchmarking Multimodal-AlgPro performance against existing allergenicity prediction methods with independent test dataset

Method	ACC (%)	*F*-value (%)	Recall (%)	PRE (%)
AllergyPred	73.8	69.6	59.8	83.0
ProAll-D	86.0	85.8	84.9	86.8
AllergenPF	87.6	87.8	89.0	86.6
AllerTOPv.2	89.8	89.6	88.0	91.2
AlgPred2.0	90.3	90.9	**97.6**	85.1
DeepAlgPro	91.6	91.5	90.3	92.8
PreAlgPro	93.1	92.8	89.4	96.5
Multimodal-AlgPro (¯CNN)	87.8	88.1	90.6	85.8
Multimodal-AlgPro (¯Single_Average)	87.9	86.9	80.1	94.8
Multimodal-AlgPro (¯PSSM400)	88.2	88.4	89.8	86.9
Multimodal-AlgPro (¯DFMCA_PSSM)	88.4	88.2	87.0	89.4
Multimodal-AlgPro (¯DPC_PSSM)	90.0	89.7	86.8	92.8
Multimodal-AlgPro	**93.9**	**93.7**	91.1	**96.4**

Bold formatting indicates the best results on the metrics. "-" represents the models removed during the ablation experiments

values were taken from cited literature or derived through further analysis, and all methods were evaluated on the allergenicity prediction task using a consistent independent dataset. Among these methods, PreAlgPro emerges as a strong baseline, exhibiting a high accuracy score. Subsequently, the Multimodal-AlgPro and its variations, derived from a series of ablation studies designed to remove different models, were assessed and compared. The Multimodal-AlgPro exhibited superior performance, particularly excelling in accuracy, *F*-value, and precision. It achieved a peak accuracy score of 93.9%, an *F*-value of 93.7%, a precision score of 96.4%, and an impressive recall score of 91.1%, indicating excellent overall performance and discriminative power. Overall, the results indicate that Multimodal-AlgPro outperforms existing methods and its variations in the allergenicity prediction task. These findings highlight the potential of the Multimodal-AlgPro in accurately predicting allergens.

Performance evaluation of Multimodal-AlgPro on real-world imbalanced datasets

To evaluate the real-world applicability of Multimodal-AlgPro, its performance was assessed on three independent datasets: TE1, TE2, and TE3, each exhibiting varying degrees of imbalance. These datasets, curated from Zhang et al. [33], have no overlap with the training data. The allergen to non-allergen ratios range from highly skewed (for example, TE1 with a 1 to 10 ratio) to moderately imbalanced (such as TE2 at 1 to 5 and TE3 at 1 to 8). As illustrated in Additional file 9: Fig. S8, Multimodal-AlgPro consistently demonstrated strong predictive performance across all settings. Notably, it achieved over 93% accuracy and auROC values exceeding 0.94 on all three datasets. For instance, TE1 yielded 95.6% accuracy and an auROC of 0.958, while TE3 achieved 94.9% accuracy with an auROC of 0.984. These findings suggest that, despite being trained on a balanced dataset, Multimodal-AlgPro generalizes well to imbalanced distributions, underscoring its robustness in more realistic application scenarios.

Discussion

The traditional approach to identifying causative allergens and their IgE-binding epitopes through labor-intensive and time-consuming experiments significantly hampers the progress of immunological diagnostics and the development of targeted therapeutic applications [7]. To address this challenge, we present a method for interpretable, supervised multimodal deep learning that integrates protein sequences and molecular features for allergen identification, which we trained and validated on 5680 and 1420 samples, respectively, from a large and standard dataset, and compared with unimodal deep learning models as well as bioinformatics tools, achieving the highest prediction performance in a one-versus-all comparison. Our method also explores multimodal interpretability in explaining how features from sequences and molecular features contribute toward allergen recognition. To facilitate the model development process, we leverage autoBioSeqpy [43], an interactive and freely accessible tool that streamlines every aspect, from design to training, evaluation, and interpretability, offering comprehensive insights into the analyses performed during allergen prediction tasks. Our goal is to enhance the transparency and accessibility of current state-of-the-art deep learning techniques, particularly emerging

multimodal methods, making them more explainable and usable for the broader research community. By providing step-by-step guidance for each trained model, we aim to empower immunologists and researchers to develop their own models and explore the findings uncovered through deep learning.

In bioinformatics, the protein is characterized by a broad spectrum of modalities, ranging from sequence data, structural information, and molecular interactions to evolutionary context [44]. Current deep learning models operate mainly in the realm of a single modality, neglecting the broader protein-level properties, which inevitably diminishes their potential [27]. Integration of different data modalities provides opportunities to increase robustness and accuracy of computational models, bringing simulations closer to real-world scenarios [45]. Though we find that multimodal integration benefits performance improvement, our results also suggest that for certain model combinations, training unimodal algorithms using either sequences or molecular features alone can achieve comparable classification performance, as the essential characteristics of allergens can be equally captured in either modality. This raises the critical issue of selecting the suitable multimodal data fusion strategy. When the number of modalities is limited, such as in technical domains like audio, visual, and language integration for specific tasks, merging all modalities together or performing ablation tests by excluding certain features can often lead to satisfactory outcomes [46]. However, as the number of modalities increases and performance disparities widen, previous strategies may struggle to fully harness the unique advantages of each modality, resulting in suboptimal outcomes. In establishing unimodal and multimodal baselines for over 2000 diverse fusion types, our results advocate that the application of multimodal integration should be determined by exhaustively enumerating all possible combinations, which may outline the directions for exploration through multimodal data interconnections.

However, despite its strengths, the current version of Multimodal-AlgPro still faces certain limitations. Specifically, it does not take into account the local geometric structure of proteins, a crucial feature closely associated with epitope mapping and prediction [47]. Although structure-aware representations like contact maps are commonly employed by graph neural networks, these representations often fail to capture the detailed protein backbone structures required for molecular docking and dynamics simulations, particularly in the context of antigen–antibody interactions. Incorporating protein backbone structures into the model would enhance its ability to capture the geometric and dynamic aspects of protein interactions, thereby improving allergen prediction and serving as a critical area for future refinement to boost the model's performance and accuracy. Moreover, in terms of interpretability, existing methods predominantly focus on unimodal models, leaving a significant gap in multimodal deep learning [48]. To address this, multimodal interpretability must extend beyond simply explaining individual feature contributions and begin uncovering the complex, interrelated patterns across different modalities. This approach should not only elucidate how different features within a single modality influence prediction but also identify cross-modal associations, offering deeper insights into how the interactions between diverse data sources drive the model's output. By doing so, we can improve the transparency and trustworthiness of multimodal models, which is essential for their application in complex real-world scenarios such as allergen prediction.

Conclusions

In summary, Multimodal-AlgPro represents a significant step forward in the prediction of allergenic proteins, combining advanced multimodal deep learning techniques and a comprehensive range of data sources. Our approach not only overcomes existing limitations but also paves the way for new research opportunities in fields such as protein engineering, vaccine development, and other biotechnological applications. By providing a more accurate and versatile allergen prediction tool, Multimodal-AlgPro has the potential to accelerate scientific advancements and drive innovation in these areas.

Methods

Datasets

This study's dataset, based on the work of He et al. [27], includes 3550 allergenic and 3550 non-allergenic protein sequences. Allergenic data were gathered from trusted sources such as the Structural Database of Allergenic Proteins (SDAP), IUIS Allergen Nomenclature, AllergenOnline, COMPARE, and UniProt (release 2022_2), supplemented with sequences from AllerTOP v.2 [18] and AlgPred 2.0 [21] for comprehensive coverage. To ensure data quality, He et al. applied stringent preprocessing steps, including the removal of highly similar sequences using CD-HIT [49] and the exclusion of sequences exceeding 1000 amino acids or containing non-standard characters (BJOUXZ). Non-allergenic data were sourced from UniProt, with all entries linked to allergens, antigens, and cancer proteins excluded. Further filtering removed any non-allergenic sequences showing more than 40% similarity to allergenic or other non-allergenic sequences [27]. After this process, a set of non-allergens was randomly selected to match the number of allergens, maintaining balance for analysis. For model

training and evaluation, 80% of the dataset was randomly assigned to the training set, while the remaining 20% was reserved for testing, resulting in 5680 training samples and 1420 test samples.

Feature extraction algorithms

Amino acid representation

Each amino acid within a protein sequence is encoded as a 20-dimensional one-hot vector, corresponding to the 20 standard amino acids. In this representation, all elements are zero except for a single position set to one, indicating the presence of a specific amino acid based on a predefined mapping (e.g., the first position may represent alanine). Consequently, each protein sequence is transformed into a two-dimensional matrix of size (20, L), where L denotes the maximum sequence length in the dataset. For sequences shorter than L, padding with pseudo-amino acids or unknown characters (such as "X") is applied, represented by zero vectors. This padding ensures uniform input dimensions across all sequences, helping to maintain consistency and prevent bias during model training.

Composition-based descriptors

Composition serves as a straightforward yet essential descriptor for protein sequences. In this study, we applied two widely used composition-based features: amino acid composition (AAC) and dipeptide composition (DC), both effective for protein sequence analysis and prediction tasks. For sequences made up of the 20 standard amino acids, AAC computes the relative frequency of each amino acid, converting the sequence into a 20-dimensional vector where each element reflects the frequency of a particular amino acid. Dipeptide composition (DC) builds upon this by calculating the frequencies of all possible dipeptide combinations, encoding each sequence into a 400-dimensional vector, with each element representing a specific dipeptide pair.

Physicochemical-based descriptors

Composition-transition-distribution (CTD) is a widely adopted method for characterizing protein and peptide sequences by capturing their structural and physicochemical properties. CTD breaks down the sequence into three feature types: composition (C), transition (T), and distribution (D) [50]. For each physicochemical property, the 20 standard amino acids are grouped into three distinct categories. Composition calculates the overall percentage of amino acids within each group, transition measures the frequency of transitions between different groups across the sequence, and distribution assesses the positional spread of amino acids from each group throughout the sequence. In the CTD encoding, composition, transition, and distribution are represented as 21-dimensional, 21-dimensional, and 105-dimensional vectors, respectively, providing a comprehensive view of amino acid properties within the sequence.

Evolutionary-based descriptors

The position-specific scoring matrix (PSSM) is a powerful tool for capturing evolutionary information within protein sequences, effectively highlighting conserved regions across homologous sequences [51]. In constructing a PSSM for a specific protein, PSI-BLAST was utilized to perform searches against the Swiss-Prot database, employing three iterations with an E-value cutoff of 0.001. Following multiple sequence alignments, the PSSM matrix is created, where each row corresponds to an amino acid position in the sequence and each column represents one of the 20 standard amino acids. The values within the matrix are logarithmic scores that reflect the probability of each amino acid appearing at a given position, based on evolutionary conservation; positive values indicate high conservation, while negative values signify lower conservation. A variety of derived features can be extracted from the PSSM matrix through transformations of its rows, columns, or their combinations. In this study, eight PSSM-based features were selected to capture evolutionary information for proteins, including AAC_PSSM, DFMCA_PSSM, DP_PSSM, DPC_PSSM, PSe_PSSM, PSSM_AC, PSSM400, and Single_Average [52]. Detailed descriptions of these features are provided in Additional file 10: Text S1.

Multimodal deep learning architecture

The architecture of our model comprises three key components: (1) a CNN-based text modality designed for processing protein sequences, (2) an MLP-based feature modality for handling protein descriptors, and (3) a multimodal fusion layer that integrates information from both sequences and descriptors.

CNN: text modality

A convolutional neural network (CNN) was employed to predict allergens from protein sequences, as it was originally developed for image classification but has since been effectively adapted to sequence modeling tasks due to its strength in capturing local contextual features [53]. Within the CNN framework, each protein sequence is transformed into a one-hot encoded matrix, represented as a one-dimensional signal spanning 20 channels, with a categorical label corresponding to each sequence for training purposes. The input sequence length is fixed at 1000 amino acids, a size selected to ensure uniformity across the dataset and that all sequences fit within the input window.

The CNN architecture primarily consists of convolutional layers (Conv), rectified linear activation units (ReLU), dropout regularization, and fully connected layers. Each convolutional layer generates a feature map by sliding a set of filters over the input sequence, capturing local patterns within the data. The mathematical representation of this process is given by:

$$\mathrm{Conv}(X)_{if} = \mathrm{ReLU}\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W_{mn}^{f} X_{i+m,n}\right) \quad (1)$$

where X is the one-hot encoded matrix, and i and f represent the indices of the output position and filter kernel, respectively. The convolutional filter W^f is an $M \times N$ weight matrix, where M represents the window size and N corresponds to the input channels. The ReLU activation function is applied element-wise after each convolutional operation to introduce non-linearity into the network:

$$\mathrm{ReLU}(x) = \begin{cases} xifx \geq 0 \\ 0ifx < 0 \end{cases} \quad (2)$$

subsequent to the convolutional layers, ReLU activations are combined with dropout regularization layers. During training, dropout randomly sets a portion of the intermediate activations to zero, which helps mitigate overfitting and encourages robust feature learning. The fraction of activations dropped is determined by a hyperparameter known as the dropout probability $\rho \in [0,1]$; higher values of ρ result in more units being deactivated.

In this architecture, two convolutional layers with ReLU activations are interspersed with dropout regularization layers. The first convolutional layer consists of 250 one-dimensional filters, each spanning all 20 input channels with a filter width of 11 positions, designed to capture local sequence motifs. The second convolutional layer also employs 250 one-dimensional filters, but these span all 250 output channels from the previous layer, utilizing the same filter width of 11. The outputs from the second convolutional layer are then flattened into a single vector and passed to a fully connected layer containing 128 hidden ReLU units, with a dropout rate of 0.2 applied for regularization. This fully connected layer learns a higher-level representation of the protein sequences, which is subsequently integrated into the multimodal fusion layer for further processing with other data modalities.

MLP: feature modality

For the prediction of allergens using descriptors, the multilayer perceptron (MLP) algorithm was implemented, known for its effectiveness in handling discrete features [54]. However, when discrete protein descriptors encompass hundreds or even thousands of features while having relatively few training samples, a single MLP architecture may be prone to overfitting and training instabilities, even when advanced regularization techniques, such as dropout and weight decay, are applied. To address these challenges, multiple MLP architectures were developed, each specifically tailored to distinct types of protein descriptors. This strategy acknowledged the heterogeneity of various molecular features and improved model performance by ensuring that each descriptor type was processed through an optimal network design. A comprehensive range of hyperparameters was systematically explored, including the number of fully connected layers, the number of hidden nodes per layer, activation functions, and dropout probabilities. Ultimately, the final MLP architecture for protein descriptor input was constructed with carefully selected hidden layers and neurons, incorporating ReLU activation and dropout in each layer. The final fully connected layer was designed to capture a robust feature representation, which was subsequently integrated into the multimodal fusion layer with other data modalities.

Multimodal fusion layer

After constructing unimodal feature representations from the CNN and MLP subnetworks, an early fusion strategy is employed to learn a joint multimodal feature representation that captures essential interactions between the two modalities [45, 48]. The joint multimodal tensor is computed by taking the element-wise product of the protein sequence representations obtained from the CNN and the protein descriptor features from the MLP. In this process, each neuron from the text modality is multiplied by its corresponding neurons in the feature modality, allowing for the capture of all bimodal interactions. Following the early fusion step, the resulting multimodal tensor is passed through a sigmoid activation layer, which transforms the output for binary classification. The classification task is supervised using a cross-entropy loss function. The sigmoid activation function is defined as follows:

$$\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}} \quad (3)$$

This formulation effectively maps the output to a probability range between 0 and 1, which is utilized for the final classification decision.

Training details

During the training process, the training set was randomly divided into non-overlapping subsets: 70% for training, 10% for validation, and 20% for testing. These subsets were utilized to train the models, fine-tune hyperparameters, and evaluate performance. This

procedure was repeated five times, with distinct non-overlapping training, validation, and test subsets assigned for each iteration. The predictive performance was averaged across all five runs to yield the final results. The multimodal architecture was trained in an end-to-end fashion, integrating the CNN subnetwork, the MLP subnetwork, and the multimodal fusion layer. Optimization of the network was carried out using the Adam optimizer, with a learning rate set at 0.001, a b_1 coefficient of 0.9, and a b_2 coefficient of 0.999, over a span of 20 epochs. A batch size of 256 was used throughout the training process. For comparative analyses with unimodal networks, both the CNN and MLP models were trained independently, employing the same cross-entropy loss function and hyperparameter settings as those used for the multimodal network.

Model interpretability and visualization

To gain a deeper understanding of the distribution of embedded sequences and descriptor representations, and to investigate the effectiveness of deep learning, uniform manifold approximation and projection (UMAP) was employed to visualize the model's operational mechanics [55]. The embedded vectors from both positive and negative training samples were stacked and processed through UMAP, which transformed the high-dimensional data into a two-dimensional space for visualization. The algorithm was executed with optimal parameters: *n_neighbors* = *28*, *min_dist* = *0.8*, *metric* = *cosine*, chosen to effectively balance the retention of local and global structural information. The resulting two-dimensional vectors were normalized to illustrate the spatial relationships between samples, providing insights into the model's ability to capture the underlying data structure.

To evaluate the contributions of different modalities, including protein sequences and descriptors, to the model's predictions, SHAP (SHapley Additive exPlanations) attributions were computed for each training sample [56]. This process involved assigning scores to all features from each modality based on their respective contributions to the final classification. The analysis offered detailed insights into how each modality and feature influenced the model's decision-making process. After calculating SHAP attributions, the absolute values from positive and negative samples were processed separately to estimate the relative importance of each feature in the overall prediction. Further analysis of the SHAP values for individual input features within each modality allowed for the quantification of the specific contributions from both sequence and descriptor features to the model output.

Performance evaluation metrics

To evaluate the performance of the developed deep learning models, several standard metrics and visualizations were employed. These include accuracy (ACC), precision (PRE), *F*-value, recall, Matthews correlation coefficient (MCC), as well as receiver operating characteristic (ROC) curves and precision-recall (PR) curves. The metrics are computed as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$PRE = \frac{TP}{TP + FP} \tag{5}$$

$$F - value = 2 \times \frac{TP}{2TP + FP + FN} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{8}$$

In these equations, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively, based on binary classification outputs. ROC curves visualize the trade-off between the true positive rate (TPR) and false positive rate (FPR) across various classification thresholds, while PR curves illustrate precision relative to recall. The areas under the ROC and PR curves (auROC and auPR) serve as quantitative measures of the model's discriminative ability, with values ranging from 0 to 1, where 0.5 indicates random guessing and 1 indicates perfect classification. Additionally, accuracy-loss curves were utilized to monitor model performance during training, offering insights into convergence behavior and potential overfitting trends.

Computational hardware and software

All deep learning models were implemented in Python (v3.8.16) using Keras and TensorFlow, and trained with GPU acceleration on an NVIDIA platform (CUDA 11.2, cuDNN 8.1.1). Key analyses such as visualization and interpretability were conducted using tools including LayerUMAP [57], SHAP [58], Funkyheatmap [59], and MetaLogo [60]. The overall framework, including the text modality, feature modality, and multimodal fusion layer, was developed using autoBioSeqpy [43].

Abbreviations

AAC	Amino acid composition
ACC	Accuracy
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional neural network
Conv	Convolutional layers
CTD	Composition-transition-distribution
DC	Dipeptide composition
FPR	False positive rate
MCC	Matthews correlation coefficient
MLP	Multilayer perceptron
PR	Precision-recall
PRE	Precision
PSSM	Position-specific scoring matrix
ReLU	Rectified linear activation units
ROC	Receiver operating characteristic
SDAP	Structural Database of Allergenic Proteins
SHAP	SHapley Additive exPlanations
TPR	True positive rate
UMAP	Uniform manifold approximation and projection

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-025-02347-z.

Additional file 1: Fig. S1. Hyperparameter optimization for the 12 unimodal deep learning models. (A) Evaluation of CNN architectures, examining performance across variations in kernel size, filter number, and convolutional layer depths. (B) Analysis of MLP architectures, comparing performance across different protein descriptors and configurations of hidden layers.

Additional file 2: Fig. S2. Comparison of MLP and CNN architectures across 11 protein descriptors for feature-modality modeling.

Additional file 3: Fig. S3. Receiver Operating Characteristic (ROC) curves for various unimodal deep learning models evaluated in this study. (A-N) CNN (One_Hot); MLP (PSSM_AC); MLP (DPC_PSSM); MLP (DC); MLP (DP_PSSM); MLP (DFMCA_PSSM); MLP (Single_Average); BiLSTM (One_Hot); MLP (PSSM400); Transformer (Dictionary); MLP (Pse_PSSM); MLP (AAC); MLP (AAC_PSSM); MLP (CTD).

Additional file 4: Fig. S4. Precision-Recall (PR) curves for the unimodal deep learning models evaluated in this study. (A-N) CNN (One_Hot); MLP (PSSM_AC); MLP (DPC_PSSM); MLP (DC); MLP (DP_PSSM); MLP (DFMCA_PSSM); MLP (Single_Average); BiLSTM (One_Hot); MLP (PSSM400); Transformer (Dictionary); MLP (Pse_PSSM); MLP (AAC); MLP (AAC_PSSM); MLP (CTD).

Additional file 5: Table S1. Experimental results of the 2047 combinations on the training dataset.

Additional file 6: Fig. S5. Receiver Operating Characteristic (ROC) curves for the top-performing multimodal combination 'CNN (One_Hot) + MLP (DFMCA_PSSM) + MLP (DPC_PSSM) + MLP (PSSM400) + MLP (Single_Average)', evaluated across five iterations of the training and testing process.

Additional file 7: Fig. S6. Precision-Recall (PR) curves for the top-performing multimodal combination 'CNN (One_Hot) + MLP (DFMCA_PSSM) + MLP (DPC_PSSM) + MLP (PSSM400) + MLP (Single_Average)', evaluated across five iterations of the training and testing process.

Additional file 8: Fig. S7. Structural mapping of epitopes on Bet v 1, the major allergen from birch pollen (PDB: 1LLT). (A) The three-dimensional structure of Bet v 1 is shown with the backbone folding pattern and secondary structural elements rendered in pale yellow. Several representative epitopes are mapped onto the structure to illustrate their spatial distribution. These include regions corresponding to amino acid positions 13 to 27 (in red), 52 to 69 (in green), 76 to 93 (in blue), and 100 to 111 (in cyan). The region from amino acid 13 to 27, highlighted in red, corresponds to the linear peptide epitope IPAARLFKAFIL, which has been studied as part of Bet v 1 from Betula pendula and is registered under epitope ID 27784 in the Immune Epitope Database; (B) The residues forming the IPAARLFKAFIL epitope are shown mapped onto the molecular surface of Bet v 1, illustrating their surface accessibility and potential involvement in IgE binding.

Additional file 9: Fig. S8. Evaluation of Multimodal-AlgPro on three distinct independent test sets.

Additional file 10: Text S1. Evolutionary-based descriptors.

Acknowledgements

We would like to thank the anonymous reviewers and editors, whose constructive feedback has been very helpful in enhancing the presentation of this paper.

Authors' contributions

LZY and YXL collected and processed the datasets. SQW and SYC assisted with data analysis and trained the models. LX and RYJprovided guidance on the project and reviewed the manuscript. JSL drafted the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by The Science and Technology Strategic Cooperation Programs of Luzhou Municipal People's Government and Southwest Medical University (2024LZXNYDT001, China), Sichuan Science and Technology Program (2024NSFSC0889, China), Luzhou Science and Technology Program (2023SYF118, China), and the Open Research Fund of Chengdu University of Traditional Chinese Medicine State Key Laboratory of Southwestern Chinese Medicine Resources (SKLTCM2022028).

Data availability

All data generated or analysed during this study are included in this published article, its supplementary information files and publicly available repositories. The codes, architecture, parameters, dataset, functions, usage and output of the proposed model are available free of charge at Zenodo repository (https://doi.org/10.5281/zenodo.15907810) and GitHub (https://github.com/jingry/Multimodal-AlgPro).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 24 February 2025 Accepted: 17 July 2025
Published online: 31 July 2025

References

1. Yang L, Fu J, Zhou Y. Research progress in atopic march. Front Immunol. 2020;11: 1907.
2. Pramod SN. Immunological basis for the development of allergic diseases-prevalence, diagnosis and treatment strategies. In: Singh B, editor. Cell interaction-molecular and immunological basis for disease management. London: Academic; 2021.
3. Verhoeckx KCM, Vissers YM, Baumert JL, Faludi R, Feys M, Flanagan S, et al. Food processing and allergenicity. Food Chem Toxicol. 2015;80:223–40.
4. Turner PJ, Jerschow E, Umasunthar T, Lin R, Campbell DE, Boyle RJ. Fatal anaphylaxis: mortality rate and risk factors. J Allergy Clin Immunol Pract. 2017;5(5):1169–78.
5. Obermeyer G, Ferreira F. Can we predict or avoid the allergenic potential of genetically modified organisms? Int Arch Allergy Immunol. 2005;137(2):151–2.

6. Fernandez A, Mills ENC, Koning F, Moreno FJ. Allergenicity assessment of novel food proteins: what should be improved? Trends Biotechnol. 2021;39(1):4–8.
7. Durham SR, Shamji MH. Allergen immunotherapy: past, present and future. Nat Rev Immunol. 2023;23(5):317–28.
8. Agamah FE, Mazandu GK, Hassan R, Bope CD, Thomford NE, Ghansah A, et al. Computational/in silico methods in drug target and lead prediction. Brief Bioinform. 2020;21(5):1663–75.
9. Pi X, Sun Y, Deng X, Xin D, Cheng J, Guo M. Investigation of differences in allergenicity of protein from different soybean cultivars through LC/MS-MS. Int J Biol Macromol. 2022;220:1221–30.
10. Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. FASEB J. 2003;17(9):1141–3.
11. Ivanciuc O, Schein CH, Braun W. SDAP: database and computational tools for allergenic proteins. Nucleic Acids Res. 2003;31(1):359–62.
12. Li KB, Issac P, Krishnan A. Predicting allergenic proteins using wavelet transform. Bioinformatics. 2004;20(16):2572–8.
13. Riaz T, Hor HL, Krishnan A, Tang F, Li KB. WebAllergen: a web server for predicting allergenic proteins. Bioinformatics. 2005;21(10):2570–1.
14. Zhang ZH, Koh JL, Zhang GL, Choo KH, Tammi MT, Tong JC. AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. Bioinformatics. 2007;23(4):504–6.
15. Soeria-Atmadja D, Lundell T, Gustafsson MG, Hammerling U. Computational detection of allergenic proteins attains a new level of accuracy with in silico variable-length peptide extraction and machine learning. Nucleic Acids Res. 2006;34(13):3779–93.
16. Muh HC, Tong JC, Tammi MT. AllerHunter: A SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. PLoS One. 2009;4(6): e5861.
17. Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. Med Chem. 2013;9(1):133–7.
18. Dimitrov I, Bangov I, Flower DR, Doytchinova I. AllerTOP vol 2–a server for in silico prediction of allergens. J Mol Model. 2014;20(6):2278.
19. Dimitrov I, Naneva L, Doytchinova I, Bangov I. AllergenFP: allergenicity prediction by descriptor fingerprints. Bioinformatics. 2014;30(6):846–51.
20. Dang HX, Lawrence CB. Allerdictor: fast allergen prediction using text classification techniques. Bioinformatics. 2014;30(8):1120–8.
21. Sharma N, Patiyal S, Dhall A, Pande A, Arora C, Raghava GPS. Algpred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. Brief Bioinform. 2021;22(4): bbaa294.
22. Nguyen MN, Krutz NL, Limviphuvadh V, Lopata AL, Gerberick GF, Maurer-Stroh S. AllerCatPro 2.0: a web server for predicting protein allergenicity potential. Nucleic Acids Res. 2022;50(W1):W36-43.
23. Nedyalkova M, Vasighi M, Azmoon A, Naneva L, Simeonov V. Sequence-based prediction of plant allergenic proteins: machine learning classification approach. ACS Omega. 2023;8(4):3698–704.
24. Liu B, Yang Z, Liu Q, Zhang Y, Ding H, Lai H, et al. Computational prediction of allergenic proteins based on multi-feature fusion. Front Genet. 2023;14: 1294159.
25. Li Y, Sackett PW, Nielsen M, Barra C. NetAllergen, a random forest model integrating MHC-II presentation propensity for improved allergenicity prediction. Bioinform Adv. 2023;3(1):vbad151.
26. Zhou L, Zhang C, Liu F, Qiu Z, He Y. Application of deep learning in food: a review. Compr Rev Food Sci Food Saf. 2019;18(6):1793–811.
27. He C, Ye X, Yang Y, Hu L, Si Y, Zhao X, et al. Deepalgpro: an interpretable deep neural network model for predicting allergenic proteins. Brief Bioinform. 2023;24(4): bbad246.
28. Behbahani M, Rabiei P, Mohabatkar H. A comparative analysis of allergen proteins between plants and animals using several computational tools and Chou's PseAAC concept. Int Arch Allergy Immunol. 2020;181(11):813–21.
29. Wang L, Niu D, Zhao X, Wang X, Hao M, Che H. A comparative analysis of novel deep learning and ensemble learning models to predict the allergenicity of food proteins. Foods. 2021;10(4): 809.
30. Shanthappa PM, Kumar R. ProAll-D: protein allergen detection using long short term memory - a deep learning approach. ADMET DMPK. 2022;10(3):231–40.
31. Kumar A, Rana PS. A deep learning based ensemble approach for protein allergen classification. PeerJ Comput Sci. 2023;9: e1622.
32. Du Z, Xu Y, Liu C, Li Y. pLM4Alg: protein language model-based predictors for allergenic proteins and peptides. J Agric Food Chem. 2024;72(1):752–60.
33. Zhang L, Liu T. PreAlgPro: prediction of allergenic proteins with pre-trained protein language model and efficient neutral network. Int J Biol Macromol. 2024;280(Pt 3):135762.
34. Yang C, Negi SS, Schein CH, Braun W, Kim P. AllergenAI: a deep learning model predicting allergenicity based on protein sequence. bioRxiv. 2024. https://doi.org/10.1101/2024.06.22.600179.
35. Long H, Liao B, Xu X, Yang J. A hybrid deep learning model for predicting protein hydroxylation sites. Int J Mol Sci. 2018;19(9): 2817.
36. Sharma A, Sharma N, Bhalla P, Singh M. Comparative and evolutionary analysis of grass pollen allergens using *Brachypodium distachyon* as a model system. PLoS One. 2017;12(1): e0169686.
37. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43(W1):W39-49.
38. Vita R, Blazeska N, Marrama D, IEDB Curation Team Members, Duesing S, Bennett J, et al. The Immune Epitope Database (IEDB): 2024 update. Nucleic Acids Res. 2025;53(D1):D436–43.
39. Ipsen H, Løwenstein H. Isolation and immunochemical characterization of the major allergen of birch pollen (*Betula verrucosa*). J Allergy Clin Immunol. 1983;72(2):150–9.
40. Bohle B, Radakovics A, Jahn-Schmid B, Hoffmann-Sommergruber K, Fischer GF, Ebner C. Bet v 1, the major birch pollen allergen, initiates sensitization to Api g 1, the major allergen in celery: evidence at the T cell level. Eur J Immunol. 2003;33(12):3303–10.
41. Movérare R, Westritschnig K, Svensson M, Hayek B, Bende M, Pauli G, et al. Different IgE reactivity profiles in birch pollen-sensitive patients from six European populations revealed by recombinant allergens: an imprint of local sensitization. Int Arch Allergy Immunol. 2002;128(4):325–35.
42. Kemmler E, Fath EK, Preissner R, Banerjee P. AllergyPred: a web server for allergen prediction. Nucleic Acids Res. 2025;53(W1):W4-10.
43. Jing R, Li Y, Xue L, Liu F, Li M, Luo J. autoBioSeqpy: a deep learning tool for the classification of biological sequences. J Chem Inf Model. 2020;60(8):3755–64.
44. Nguyen VTD, Hy TS. Multimodal pretraining for unsupervised protein representation learning. Biol Methods Protoc. 2024;9(1):bpae043.
45. Chen RJ, Lu MY, Williamson DFK, Chen TY, Lipkova J, Noor Z, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell. 2022;40(8):865–78.
46. Rahate A, Walambe R, Ramanna S, Kotecha K. Multimodal co-learning: challenges, applications with datasets, recent advances and future directions. Inf Fusion. 2022;81:203–39.
47. Tubiana J, Schneidman-Duhovny D, Wolfson HJ. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. Nat Methods. 2022;19(6):730–9.
48. Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell. 2022;40(10):1095–110.
49. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
50. Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 2011;39(W):W385-90.
51. Yu L, Xue L, Liu F, Li Y, Jing R, Luo J. The applications of deep learning algorithms on in silico druggable proteins identification. J Adv Res. 2022;41:219–31.
52. Mohammadi A, Zahiri J, Mohammadi S, Khodarahmi M, Arab SS. PSSM-COOL: a comprehensive R package for generating evolutionary-based descriptors of protein sequences from PSSM profiles. Biol Methods Protoc. 2022;7(1):bpac008.
53. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
54. Juez-Gil M, Erdakov IN, Bustillo A, Pimenov DY. A regression-tree multi-layer-perceptron hybrid strategy for the prediction of ore crushing-plate lifetimes. J Adv Res. 2019;18:173–84.
55. Healy J, McInnes L. Uniform manifold approximation and projection. Nat Rev Methods Primers. 2024;4:82.

56. Chen H, Covert IC, Lundberg SM, Lee S-I. Algorithms to estimate Shapley value feature attributions. Nat Mach Intell. 2023;5(6):590–601.
57. Jing R, Xue L, Li M, Yu L, Luo J. layerUMAP: A tool for visualizing and understanding deep learning models in biological sequence classification using UMAP. iScience. 2022;25(12):105530.
58. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. NeurIPS. 2017;30:4765–74.
59. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37(5):547–54.
60. Chen Y, He Z, Men Y, Dong G, Hu S, Ying X. MetaLogo: a heterogeneity-aware sequence logo generator and aligner. Brief Bioinform. 2022;23(2):bbab591.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.