

pLM4Alg: Protein Language Model-Based Predictors for Allergenic Proteins and Peptides

Published as part of *Journal of Agricultural and Food Chemistry* virtual special issue "3rd International Symposium on Bioactive Peptides".

Zhenjiao Du, Yixiang Xu, Changqi Liu, and Yonghui Li*



Cite This: <https://doi.org/10.1021/acs.jafc.3c07143>



Read Online

ACCESS |



Metrics & More



Article Recommendations



Supporting Information

ABSTRACT: The rising prevalence of allergy demands efficient and accurate bioinformatic tools to expedite allergen identification and risk assessment while also reducing wet experiment expenses and time. Recently, pretrained protein language models (pLMs) have successfully predicted protein structure and function. However, to our best knowledge, they have not been used for predicting allergenic proteins/peptides. Therefore, this study aims to develop robust models for allergenic protein/peptide prediction using five pLMs of varying sizes and systematically assess their performance through fine-tuning with a convolutional neural network. The developed pLM4Alg models have achieved state-of-the-art performance with accuracy, Matthews correlation coefficient, and area under the curve scoring 93.4–95.1%, 0.869–0.902, and 0.981–0.990, respectively. Moreover, pLM4Alg is the first model capable of handling prediction tasks involving residue-missed sequences and sequences containing nonstandard amino acid residues. To facilitate easy access, a user-friendly web server (<https://f6wxpfd3sh.us-east-1.awsapprunner.com>) has been established. pLM4Alg is expected to become the leading machine learning-based prediction model for allergenic peptides and proteins. Its collaboration with other predictors holds great promise for accelerating allergy research.

KEYWORDS: allergenicity, allergens, proteins and peptides, protein language model, convolutional neural network, deep learning

INTRODUCTION

Allergic diseases encompass a range of disorders driven by both innate and adaptive immune responses, in collaboration with epithelial cells.¹ These interactions lead to hypersensitivity against otherwise harmless foreign substances (e.g., foods, dust mites, pollens, chemicals, etc.).² Most allergic diseases involve a type 2 immune response, including various cells like Th2 cells, type 2 B cells, type 2 innate lymphoid cells, type 2 macrophages, basophils, eosinophils, mast cells, as well as specific NK cells and NK-T cells that secrete interleukin (IL)-4, IL-5, IL-9, IL-13, IL-25, and IL-33 cytokines.^{3,4} Upon exposure, allergens cross-link immunoglobulin E (IgE) antibodies on the surface of granulocytes, which leads to the release of mediators and inflammatory molecules (e.g., histamine and leukotriene), triggering various symptoms, including allergic asthma, rhinitis, skin reactions, and even death.⁵ Over 30% of the U.S. population has been affected by these adverse reactions, and the prevalence continues to rise.^{5,6} Numerous theories have been proposed to explain the growing allergy epidemics including the hygiene hypothesis and its extensions (e.g., the old friends hypothesis and biodiversity hypothesis),⁷ the nutritional immunomodulation hypothesis (e.g., the vitamin D hypothesis),^{8,9} the epithelial barrier hypothesis,^{1,10} and the climate change hypothesis.¹¹

For a substance to be allergenic, it must induce IgE production during sensitization, involving dendritic cells, T cells, and B cells, and trigger clinical reactions upon subsequent exposures, encompassing immediate- and late-phase responses.

However, this simplistic view overlooks complexities during initial allergen exposure, such as allergen biochemistry, surrounding innate immune stimulants, allergen stability in tissues and mucosa, and dosage and duration of interaction with the immune system.¹⁰ The majority of allergens are protein/peptides, which are widely present in foods, personal care products, and the environment.^{12–14} Notably, many allergens, especially in plants and foods, belong to a few protein families.¹⁵ The reason certain proteins act as allergens remains unclear, but they often share characteristics contributing to allergenicity, including small size, water solubility, glycosylation, repetitive structures, numerous disulfide bonds, resistance to heat, acid, and proteolysis, ligand-binding property, ability to interact with membranes and lipids, and intrinsic adjuvant property.^{10,16,17}

To improve efficiency in allergen identification and accelerate allergenic risk evaluation, computational approaches have been continuously proposed and have made significant progress in the past decades. The first guideline was proposed by the consultation group of FAO/WHO in 2001, where the allergenic potential of a protein was defined by an exact match

Received: October 2, 2023

Revised: November 20, 2023

Accepted: December 8, 2023



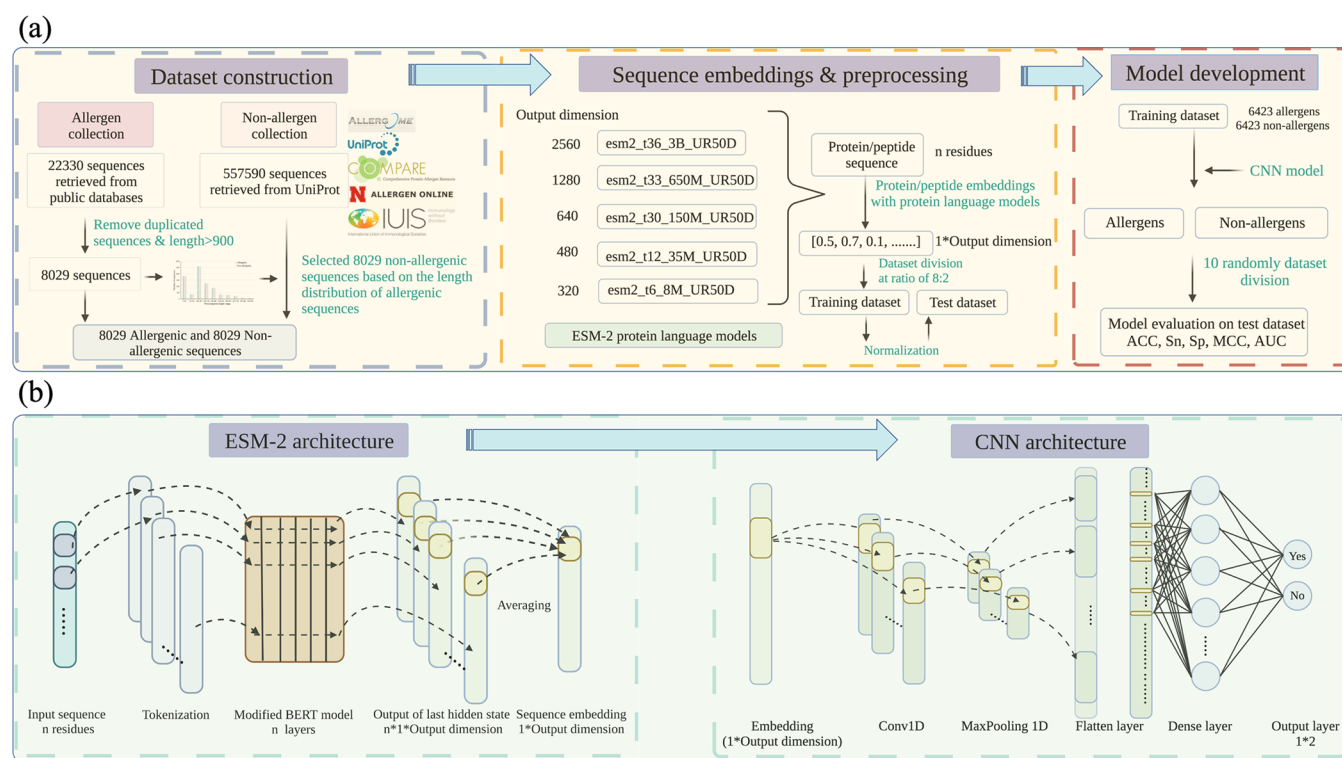


Figure 1. Schematic framework of pLM4Alg model development (a) and detailed architecture of ESM-2 protein language models and convolutional neural network (CNN) models (b).

of a stretch of six or more consecutive identical amino acids (rule 1), or more than 35% identity within any window of 80 amino acids in comparison to any known allergen (rule 2).¹⁸ Subsequently, various approaches are proposed from different perspectives for allergen prediction, such as motif-based approaches, similarity searches, machine learning-based modeling, etc. Prediction tools, such as AllerCatPro 2.0, AlgPred 2.0, have been used extensively for preliminary allergenic risk assessment and identification of new allergens.^{5,14,19–28} One representative project derived from the FAO/WHO guidelines was AllerCatPro/AllerCatPro 2.0, which implemented a hierarchical workflow employing five criteria with improved similarity-checking methods for both sequences and 3D structures.^{20,27} The criteria used in AllerCatPro were based on statistical analysis of existing allergens, representing a form of experience/knowledge-based decision-making. For instance, if a protein sequence shares >35% identity with known allergens over 90 windows, it would be classified into the group with strong evidence for allergenic potential.^{20,27} So far, this approach has achieved a great performance in the allergen benchmark data sets.²⁷ However, the thresholds set in these criteria were based on previous studies, common sense, and accumulated knowledge. There was no independent data set for validating these criteria. Additionally, the reliance on allergens in the search databases for similarity checking significantly limited model robustness when encountering unseen allergens.

Another representative technique is the machine learning (ML) approach where the classification criterion is determined by the ML model by learning from a separate training data set.^{5,14,19,21–26,28} Besides the selection of modeling method, the most challenging task in the ML approach is the representation or encoding of protein/peptide sequences into a numerical vector/matrix. Various encoding methods,

including amino acid descriptors, amino acid composition (AAC), pseudoamino acid composition (PseAAC), dipeptide composition (DPC), amino acid descriptors (AAD), position-specific scoring matrix (PSSM), physicochemical descriptors, biomedical properties, k-mer dictionary-based binary representation, etc., have been widely used in predicting allergenicity and other properties/bioactivities.^{5,14,21,22,25,28–34} However, these features may not always accurately represent protein sequences and simple combinations can cause high-dimensional problems as well as the feature redundancy.³⁵

Recently, inspired by the extraordinary performance of transformer-based models in natural language processing (NLP) tasks, a series of attempts have been made in pretrained protein language models (pLMs) (e.g., ProtTrans, ESM, UniRep, etc.) for downstream protein sequence tasks.^{22,35–40} The central concept lies in incorporating the attention mechanism during the self-supervised learning process in these pLMs. This mechanism enables the model to observe the entire sequence simultaneously to learn the representation of each residue as well as their relationship, rather than locally (as in convolutional neural network (CNN)) or sequentially (as in long short-term networks (LSTM)). As a result, the model can capture the overall contextual information on the entire sequence more effectively and efficiently.⁴¹

Evolutionary scale modeling (ESM), one of the pretrained pLMs, was updated into ESM-2 in 2022 by the Meta Fundamental AI Research Protein Team (FAIR) for protein representation.³⁸ In our previous study, it exhibited a great potential in developing a universal model architecture and achieved state-of-the-art (SOTA) performance in 15 out of 20 peptide bioactivity/function data sets by fitting the same model architecture in different peptide data sets.³⁵ Due to the limited data availability of bioactive peptides and the universal application scenarios, a small size ESM-2 pLM (320 output

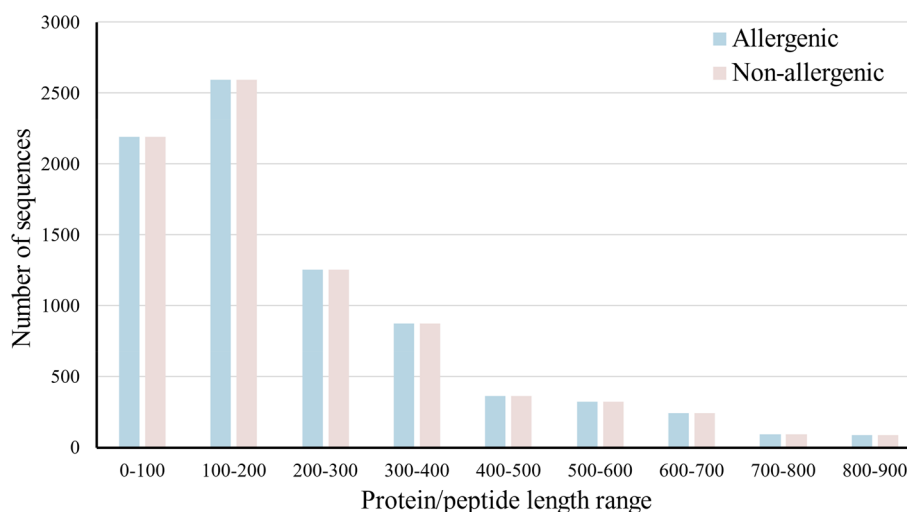


Figure 2. Data set profile of protein/peptide length distribution. (Note: There are 1816 sequences with less than 50 residues in the 0–100 range.)

dimension) and a simplified model architecture were selected.³⁵ ESM-2 includes five additional pretrained pLMs varying from 48 layers with 15 billion parameters for 5120 output embeddings dimensions to 6 layers with 8 million parameters for 320 output embeddings dimensions. The higher output dimensions corresponded to more information extracted and retained from the protein sequences, resulting in better performance in the downstream tasks.^{38,39} However, the high dimension feature space can sometimes cause the curse of dimensionality and hinder the improvement in performance when the data set size is small.²⁹ Up to now, more than 8000 allergens have been identified and archived in public databases,^{6,42–45} which allows us to fully exploit the potential of various ESM-2 pLMs with different sizes in allergen prediction.

To the best of our knowledge, ESM-2 pLMs have not been used for the numerical representation of allergenic proteins/peptides for prediction model development. This study aims to develop pLM-based high-quality models for assessing the allergenic potential of proteins/peptides. A series of pLMs of different sizes were compared regarding their performance in allergenicity prediction when combined with CNN (Figure 1). In addition, the study established a new allergenic sequence data set that includes nonstandard residues ([BOUZ]) and employed it in the model development. Based on the final performance, we deployed a user-friendly web server, accessible online at no cost, to meet the great demands for allergenicity risk assessment and allergen identification. The comparison between different pretrained pLMs will provide valuable insight into the relationship between pLM size, fine-tuning with CNN, and prediction performance in downstream protein/peptide tasks. Moreover, the study may inspire future development of pLM-based models for other physicochemical/bioactive property predictions of proteins and peptides, particularly in the areas of architecture design and hyperparameter tuning.

MATERIALS AND METHODS

Data Set Construction. The data set used in this study was compiled from five public databases: ALLERGOME, COMPARE, AllergenOnline, WHO/IUIS, and Uniprot.^{42–45} The sequences in these databases were all peer-reviewed by expert panels and/or sourced from peer-reviewed journals and were treated as high-quality

data sources. Specifically, a total of 2631 allergens were downloaded from COMPARE at <https://comparedatabase.org/> (the latest updates on January 26, 2023). Additionally, 2233, 1090, and 4981 sequences were retrieved from AllergenOnline (<http://www.allergenonline.org/>) (released on February 14, 2021), WHO/IUIS (<http://www.allergen.org/index.php>) (the latest updates on April 3, 2023), and ALLERGOME (<https://www.allergome.org/index.php>) (the latest updates on February 28, 2023), respectively. These sequences were downloaded from the NCBI database (<https://www.ncbi.nlm.nih.gov/>) by data mining using Selenium with python. The AllergenOnline database also contains 1041 peptides and 75 proteins related to celiac disease, released on April 7, 2022. Furthermore, 1279 allergenic sequences were extracted from UniProt (<https://www.uniprot.org/>) using the keyword “allergen AND reviewed: yes”. After removing duplicate sequences and a very limited number of long sequences (>900 residues), 8029 allergenic sequences as the positive data set from a total of 22330 collected allergenic sequences are finally obtained (Supporting Information, SI, Document S1). Sequences containing nonstandard amino acids ([BOUZ]) were retained in the data sets, since ESM-2 pLMs can recognize and encode them. Additionally, some sequences from ALLERGOME containing X (“unknown”) residues were retained since they can also be encoded by ESM-2 pLMs.

Constructing a nonallergens data set is challenging since there are no ready-to-use databases, specifically dedicated to nonallergenic sequences. A common approach is to retrieve random sequences through the exclusion of keywords.²⁹ In this study, we followed a previous study by using the query “NOT allergen NOT cancer NOT allergenic AND reviewed: yes” to retrieve the data set from UniProt.⁵ This process yielded 557 590 sequences (retrieved on May 12, 2023). Out of these, 8029 sequences were selected based on the same length distribution as observed in the positive data set. The length distribution of the allergenic and nonallergenic sequences is shown in Figure 2. Finally, the data set contains 8029 allergenic and 8029 nonallergenic sequences.

Protein Language Model for Embeddings and Data Set Preprocessing. ESM is a pLM project initiated by FAIR in 2019 (<https://github.com/facebookresearch/esm>). The SOTA version (ESM-2) was released in 2022 and trained on UR50/D 2021_04 data sets.³⁸ Out of the six available pLMs, five were selected for sequence embeddings due to their more manageable RAM memory consumption (Table 1).

CNN Model Architecture and Prediction Model Development. Inspired by our previous study, a convolutional neural network (CNN) was used to build the connection between the sequence embeddings and allergenicity prediction.³⁵ The overall architecture in our previous study was retained, where a convolutional layer, a dropout layer (rate = 0.15), a flatten layer, and two dense layers with a

Table 1. Available ESM-2 Pretrained Models

pretrained model name	number of layers	number of parameters	output dimensions
esm2_t48_15B_UR50D ^a	48	15 billion	5120
esm2_t36_3B_UR50D	36	3 billion	2560
esm2_t33_650M_UR50D	33	650 million	1280
esm2_t30_150M_UR50D	30	150 million	640
esm2_t12_35M_UR50D	12	35 million	480
esm2_t6_8M_UR50D	6	8 million	320

^aThis pretrained model was not selected because of the excessive computation demands in sequence embeddings.

dropout layer (rate = 0.15) were connected sequentially. Four hyperparameters (filter size, kernel size, and stride in the convolutional layer, and units in the first dense layer) were optimized through customized grid search. Filter size was selected from [16, 32, 64, 128, 256], kernel size from [32, 64, 128, 256], stride from [1, 2, 4, 8], and units from [3, 6, 9, 12, 32, 64, 128, 256, 512, 1024, 2056, 4089, 8192], respectively.

Stochastic gradient descent (SGD) was chosen as the optimizer to accelerate the model fitting and a customized learning rate scheduler was designed to update the learning rate at every epoch. Additionally, an early stop and model checkpoint were set, and the model with best performance in accuracy was stored under the consideration of validation accuracy. The model was built with Keras framework (<http://www.keras.io>) and tested on a Google Colab platform with GPU for acceleration. The detailed codes are available at <https://github.com/dzjxzyd/pLM4Alg>.

The data set was randomly divided into a training data set and a test data set at a ratio of 8:2. The test data set is independent from the training process and only used in the final model performance evaluation. A grid search was conducted initially to roughly find out the potential hyperparameters for better model performance. Then, 5-fold cross validation (CV) was used in hyperparameter tuning. The optimal hyperparameters selected for each pLM were used for the CNN training on the entire training data set without early stop and check point. The training process was stopped after it reached the epoch limit, which was observed in the CV process. Performance evaluation was conducted in the test data set ten times with a random data set division to ensure the robustness of the model performance.

Model Evaluation. Accuracy (ACC), sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and area under the curve (AUC) were adopted to evaluate the model performance. These parameters were calculated based on the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN), according to the following equations:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100,$$

$$Sn = \frac{TP}{TP + FN} \times 100,$$

$$Sp = \frac{TN}{TN + FP} \times 100,$$

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

The AUC is calculated using the sklearn "roc_auc_score" function through the portability distribution of the model prediction in the test data set.

Web Server Development. A user-friendly web server (available at <https://f6wxpfd3sh.us-east-1.awsapprunner.com>) was deployed on Amazon Web Services (AWS) App Runner based on AWS Elastic Container Registry (ECR). The Web site was designed with html and css scripts, and the model deployment was achieved with Flask (2.2.2) and Docker (24.0.5).

Considering computational needs in pLMs for sequence embeddings, available resources, and financial constraints, three pLM-based models, excluding esm2_t36_3B_UR50D (2560 output dimension) and esm2_t33_650M_UR50D (1280 output dimension), were deployed for the prediction tasks on the web server. The web server provides a large-scale processing option, which allows users to upload their peptides information through xls, xlsx, fasta, and txt formats for allergenic sequence prediction. The detailed scripts and the deployed models for web server development are provided and available at <https://github.com/dzjxzyd/pLM4Alg> and https://drive.google.com/drive/folders/1veD40uj8R7gpIo8y1niXtoKij_vZ3eiD?usp=sharing.

Besides, we also provide a user-friendly script for web server setup based on Google Colab to support the usage of all the five models in allergenic peptides/proteins prediction. Users can download the script at <https://drive.google.com/file/d/1xwlSCZhuTvRo54CzEtqMEWwIEV9xkyf/view?usp=sharing> and open it with Google Colab to run the web server locally on Colab cloud platform freely.

RESULTS AND DISCUSSION

Construction of a Comprehensive Database of Known Allergens. A high-quality data set is vital for the development of a machine learning model as it directly determines the application scenarios of the final model. In this study, we only focused on data sets that have been regularly maintained and continuously updated over the past decades and derived from reliable data sources (peer-reviewed papers and/or reviewed by experts). The data sets used in previous studies were not used since some of them were outdated or unavailable.^{5,19} The detailed allergenic sequences extracted from each database are available in [SI Document S1](#).

Though there are some overlaps in the allergenic sequences, each database has some unique features. ALLERGENOME is the largest database and fully based on allergens causing IgE-mediated reactions collected from peer-reviewed papers. It excludes low molecular weight compounds causing non-IgE-mediated allergic diseases (e.g., contact dermatitis, drug allergy) and molecules causing cell-mediated diseases (e.g., celiac disease).⁴² COMPARE and AllergenOnline are general allergenic protein databases. Recently, AllergenOnline included a separate celiac disease database containing a large number of short peptides.^{6,44} The WHO/IUIS database is more stringent, requiring allergens archived to meet the WHO/IUIS criteria and demonstrate protein-specific IgE-reactivity from at least five patients.^{20,43} The Uniprot database is not solely dedicated to allergens, but is one of the largest databases for general protein annotations. The Swiss-Prot (reviewed) database is a high-quality, manually annotated, and nonredundant protein sequence database, which consolidates experimental results, computed features, and scientific conclusions.⁴⁵

The allergenic sequence length distribution is shown in [Figure 2](#). The generalizability of machine learning models is highly dependent on their original data sets. Our data sets cover a wide range of lengths, spanning from 0 to 900 residues, with most sequences having lengths below 400 residues. Notably, there are 1816 sequences (22.5%) shorter than 50 residues, which assures the model's ability in predicting small allergenic proteins/peptides. In contrast, in the study of Yu et al., a total of 1000 sequences between 5 and 50 residues were used for developing their classification model.¹⁴ The data set constructed in our study is the largest and the most up-to-date, except for the study of Sharma et al., where 10 075 sequences were collected by compiling public databases and data sets from previous studies.^{5,14,19,21} It should be noticed that, in this

Table 2. Comparison between pLM4Alg and Machine Learning-Based Allergenic Sequence Prediction Models^a

project name ^b	data set	ACC (%)	Sn (%)	Sp (%)	MCC	AUC	web server working	year release	ref.
pLM4Alg-2560	16058	95.1 ± 0.4	94.2 ± 0.7	96.0 ± 0.4	0.902 ± 0.008	0.990 ± 0.001	yes	2023	
pLM4Alg-1280	16058	94.7 ± 0.4	93.5 ± 0.6	95.9 ± 0.4	0.894 ± 0.009	0.988 ± 0.001	yes	2023	
pLM4Alg-640	16058	94.4 ± 0.4	93.3 ± 0.6	95.4 ± 0.4	0.888 ± 0.008	0.986 ± 0.002	yes	2023	
pLM4Alg-480	16058	94.0 ± 0.4	93.0 ± 0.8	95.1 ± 0.5	0.881 ± 0.008	0.985 ± 0.002	yes	2023	
pLM4Alg-320	16058	93.4 ± 0.3	91.9 ± 0.6	94.9 ± 0.6	0.869 ± 0.006	0.981 ± 0.001	yes	2023	
AlgPred 2.0	20150	92.7	94.04	91.46	0.86	0.97	yes	2021	5
AllerStat ^c	21154	N/A	N/A	N/A	0.4495	0.878	no	2023	26
not specified-1 ^c	1183	93.1 ± 1.5	N/A	N/A	N/A	0.9578	no	2021	22
not specified-2 ^c	525745	88	91	87	0.32	N/A	no	2019	24
not specified-3	954	93	90	95	0.86	0.98	no	2023	21
not specified-4	2000	69.66	N/A	N/A	N/A	0.764	no	2023	14
not specified-5 ^d	178	88.24	83.82	90.91	0.74	N/A	no	2020	23
Allerdictor ^c	42977	N/A	N/A	N/A	N/A	0.969 ± 0.005	no	2014	25
AllerTOPv2	4854	88.7	86.7	90.7	0.775	N/A	yes	2014	19
AllergenFP	4854	87.9	86.8	89.1	0.759	N/A	yes	2014	28

^aNote: N/A: not available; ACC: accuracy; Sn: sensitivity; Sp: specificity; MCC: Matthews correlation coefficient; and AUC: Area under the curve.

^bThe number after the hyphen sign represents output dimension of the specific pretrained ESM-2 pLMs (e.g., pLM4Alg-320* represents the model that was developed based on pretrained pLM esm2_t6_8M_UR50D where the output dimension after embeddings is 320). ^cImbalanced data set (3907 allergenic sequences and 39 070 nonallergenic sequences for Allerdictor; 2248 allergenic sequences and 18 906 nonallergenic sequences for AllerStat; 583 allergenic sequences and 600 nonallergenic sequences for the Not specified-1; 1673 allergenic sequences and 524 072 nonallergenic sequences for the Not specified-2). ^dThis model is for classifying plant and animal allergens.

study, CD-HIT was not used for the nonredundancy construction since the superiority of pLMs in effective representation/embeddings of peptide/protein sequence. The pLMs-based sequence embeddings can avoid the similarity in feature space between similar sequences to a large extent, compared to traditional peptide representation methods (see details in the next subsection). Besides, we also conducted some preliminary experiments with the data set cleaned by CD-HIT under different thresholds (0.4 and 0.9). The model performance was undermined by the decreased data set size, rather than improved by the less redundant data set generated from CD-HIT (SI Document S2).

Pretrained pLMs in Sequence Embeddings. The ESM-2 pLMs are a modified version of the bidirectional encoder representations from transformers (BERT)-based architecture trained on a data set (UR50/D 2021_04 at https://huggingface.co/datasets/nferruz/UR50_2021_04) containing 48 million sequences. The data set (UR50/D 2021_04) used for developing ESM-2 pLMs also contained a proportion (around 2%) of sequences with lengths below 50 residues (<https://www.uniprot.org/uniref?query=%2A>). This assured the universality of ESM-2 pLMs in handling our allergenic protein/peptide data sets. Moreover, since the data set contains nonstandard amino acids ([BOUZ]), the resulting ESM-2 pLMs have the ability to encode protein sequences with these residues, whereas the sequences with nonstandard residues were generally removed during model development in previous studies.^{5,21,27} Furthermore, even if a given sequence has absent residues at a few positions, ESM-2 pLMs can still encode the sequence because of its inherent capacity from the masked language model strategy.^{38,39}

The output embedding matrix ($n \times \text{dim}$) for the input sequence is the last hidden state of ESM-2 models, where n represents the length of the sequences and dim depends on the selected pretrained model (e.g., the dim of esm2_t6_8M_UR50D model is 320) (Figure 1b). Each row in the embedding matrix represents the corresponding residues, and the generation of the embedding considers both the residues

and the context of the entire sequence. Subsequently, an average pooling operation is conducted to generate a $1 \times \text{dim}$ feature to represent the original sequences, thus allowing sequences with varying lengths to be unified to the same length and loaded into the model for prediction tasks.³⁸ The representation of a sequence generated by ESM-2 pLM can be considered a global descriptor of the sequence.

In previous studies where global descriptors (e.g., AAC, DPC, PSSM, etc.) were used for sequence embeddings, the sequential information on proteins/peptides was lost, thereby hindering further improvement in prediction model performance.^{5,13,21} For example, a tripeptide “ALL” encoded by AAC would be indistinguishable from the tripeptide “LLA”. It can be observed in the study of Sharma et al., where the pure machine learning models only achieved 84% accuracy with AAC as the descriptor.⁵ However, local descriptors (e.g., amino acid index, physicochemical properties, etc.) faced challenges in unifying the feature dimension of the embeddings. In a study by Yu et al., each property in AAindex of a peptide was the average of all residues, resulting in a less promising performance for allergenic peptide prediction (accuracy = 72.2%).¹⁴ In addition, amino acid descriptors such as DPPS, 3Z-scale, 5-Zscale, HESH, FASGAI, ISA-ECI, and E-scale, yielded inferior performance than AAindex.¹⁴ These encoding approaches also suffered from the lack of sequential information present in global descriptors (AAC, DPC, etc.). Another feature dimension unifying strategy is to select m N-terminus residues and n C-terminus residues (m and n are numbers). However, this is also an expedient solution since the information on the middle residues is lost.²⁹

Since pLMs could capture more feature information at both the residue-wise and the contextual levels, it is more sensitive compared to the above-mentioned traditional encoding methods. Consequently, pLMs-based embeddings can effectively differentiate proteins/peptides. For example, similar sequences (e.g., “ALL” and “LLA”, “AAL” and “ALLA”, “ALLERGY” and “ALLERGG”) can be encoded distinctly, avoiding highly similar or identical feature matrixes/vectors

between them. This distinction is critical for subsequent machine learning models to identify specific patterns with a high-resolution. Herein, ESM-2 pLMs possess such a capacity and thus do not require the popular redundancy removal approach (CD-HIT) to clean the data set. It is worth noting that in our previous study on the angiotensin-converting enzyme inhibitory peptide prediction model, ESM-based embeddings exhibited a significant superiority over traditional descriptors (local descriptors such as AAIndex, 3Z-scale, DPPS, etc., and global descriptors such as AAC, DPC, etc.).⁴⁰ Such a superiority is also observed in the study of Wang et al. where a language model, protBERT, was used for the allergenic sequence representation and exhibited better performance compared to PseAAC.²²

Performance of pLMs-Based Models. The overall architecture used in this study was based on our previous universal architecture design using pLMs.³⁵ Four key hyperparameters (filter size, kernel size, and stride in the convolutional layer, and units in the first dense layer) were selected in the CNN model development to fit the training data set after embedding with different ESM-2 pLMs. In our customized grid searching, units in the first dense layer played a dominate role in improving the final model performance as shown in SI Document S2. Due to limited computation resources, random data set divisions were not performed during the grid search. Instead, the hyperparameter combinations that had the higher performance with a specific ESM-2 pLM were further evaluated in 5-fold CV. The hyperparameters achieved promising performance in CV were used for the final model development on the entire training data set, which was evaluated on the independent test data set ten times with random data set divisions. As shown in Table 2, pLM4Alg-2560 achieved the best performance with ACC = $95.1 \pm 0.4\%$, Sn = $94.2 \pm 0.7\%$, Sp = $96.0 \pm 0.4\%$, MCC = 0.902 ± 0.008 , and AUC = 0.99 ± 0.001 .

A positive correlation between overall performance and output dimension was observed, which is consistent with previous studies on downstream tasks of pLMs.^{37,38} This can be attributed to the higher output dimension and the greater amount of contained information in the larger pretrained pLMs. The original feature dimension of a sequence is high when encoded to contain the complete sequence information (e.g., one hot encoding). As an encoder, the pLM can map the full input space to a lower dimension, while striving to retain as much meaningful information as possible. Hence, a higher output dimension means that the encoder preserved more information after the encoding process. The final model's performance was also dependent on the data set size, as enough data points are needed to allow the neural network to tune the parameters with backpropagation. Although a higher output dimension contains more information, it also requires more parameters to fit the feature space. Consequently, a sufficiently large data set size is necessary to accommodate the increase in feature dimensions. Otherwise, it would paralyze further improvement in model performance because of the curse of dimensionality.⁴⁰ This might explain why there was a logarithmic increase in the model performance with the increase in output dimension.

Comparison with Existing Methods. The performances of the existing ML-based predictors are shown in Table 2. All pLM-based models outperformed the state-of-the-art (SOTA) models regarding the evaluation matrix. The performance of a ML model is influenced by various factors, such as its data sets,

features, and modeling methods. Considering the different data sets used in the listed models, it would be unfair to simply state that one performs better than others. To ensure a fair comparison, we employed pLM-based models to fit the three original data sets of existing models (Table 2) that are available online, using the selected hyperparameters based on our data sets. The pLM-based models exhibited superior performance (SI Document S3) on the AllerTOPv2, AllergenFP, and AllerStat data sets than the existing models. ACC of our models ranged from 91.2% to 92.8%, higher than those of AllerTOPv2 (88.7%) and of AllergenFP (87.9%), using the same data set. Furthermore, our models had an AUC from 96.8% to 97.7%, comparing to that (87.8%) of the AllerStat model.^{19,26,28} These results highlight the superiority of pLM-based models over the traditional modeling methods in allergenicity prediction.

The improvement in the model performance for the pLM4Alg architecture also partially resulted from the data set size. Our data set is larger compared to the AllerTOPv2 and AllergenFP data set, which was only two-thirds of ours, and the AllerStat data set was a highly imbalanced data set with only 2248 allergenic sequences. For AlgPred 2.0, our data set source is the same as that used in its machine learning model development regarding the databases. The best model in the AlgPred 2.0 papers is a hybrid model, integrating a k-nearest neighbors (KNN) model based on traditional embeddings (AAC and PSSM), along with two approaches for mapping IgE epitopes for allergen prediction: basic local alignment search tool (BLAST)-based searching and motif-emerging and with classes-identification (MERC1).⁵ With an additional boost from epitope mapping for classification, the performance in ACC increased from 84.1% under the solely AAC-based random forest model to 92.8% and achieved the SOTA performance at that time. In this study, we used a similar data set as that in AlgPred 2.0 and achieved better performance in pLM4Alg models without the help of epitope mapping.

The primary distinction between pLM4Alg and previous approaches lies in the utilization of pLM. As discussed in the previous section, pLM demonstrated the highest prediction power based on our previous study comparing different peptide descriptors.⁴⁰ Furthermore, to the best of our knowledge, deep learning methods such as CNN have not been used in allergenicity prediction until now, which contributed to the observed improvement in performance. In the original pLM papers, a multilayer perceptron (MLP) was suggested and used in downstream tasks.^{37,38} In this study, we introduced a convolutional layer before the dense layers by employing CNN, which allows to extract more patterns from the embeddings, as this operation yielded a promising performance in our previous attempts.³⁵ Traditional ML methods (e.g., random forest, support vector machine, k-nearest neighboring, Naïve Bayes, etc.) have been the mainstream approaches in previous studies on allergenicity prediction.^{5,14,19,21,22,26,28,28} Deep learning methods generally have higher model complexity than traditional ML modeling methods. As allergy research gains increasing attention, it is anticipated that the data set sizes will continuously grow in the future. Consequently, deep learning methods are expected to be extensively exploited in predicting allergenic sequences and achieve higher performance in the future.

In addition to its superior model performance, pLM4Alg stands out as an integrated prediction model capable of handling both peptides and proteins, including sequences

containing absent residues or nonstandard amino acids, a feature unavailable in existing predictors to date. However, a limitation of pLM4Alg is the length restriction of input protein sequences. When protein sequences exceed 900 residues, the computational requirements for embeddings become enormous, surpassing the capabilities of our test setup with 80 GB RAM and 40 GB GPU memory. Meanwhile, due to the limited sequence availability, sequences larger than 900 residues were excluded in our study. Therefore, pLM4Alg is not recommended for prediction tasks which involve the long sequence (>900 residues) proteins. Since the traditional sequence embedding methods have significantly lower computational demand than pLMs, they are not limited by this constraint.^{5,14,19,21,28}

AllerCatPro 2.0 is another representative allergenicity prediction tool relying on the similarity in protein sequences and 3D structures between input sequences and its self-constructed allergenic sequence data sets.²⁷ AllerCatPro 2.0 could predict weak, strong, or no evidence of allergenic potential, while pLM4Alg are binary classification models. However, it should be noted that there was no machine learning involved in the prediction-making process of AllerCatPro 2.0. Most of the benchmark data sets in AllerCatPro 2.0 were also included in our data set construction and used for model training or evaluation. Thus, directly comparing the performance of our models with AllerCatPro 2.0 may not be entirely fair or necessary. According to the data set construction of AllerCatPro 2.0, it may not be suitable for sequence prediction shorter than 50 residues or containing nonstandard amino acids. As both tools possess unique capabilities, pLM4Alg and AllerCatPro 2.0 can be employed jointly to meet the diverse needs for identification of an unknown allergen and risk assessment.

The growing demand for identifying new allergens and allergenic risk assessment drives the development of efficient and accurate in silico prediction models. In this study, the latest pretrained protein language models are introduced to develop an allergenic sequence prediction model. A systematic comparison was made on the model performance based on different sizes of pLMs and fine-tuning with CNN. The proposed models, pLM4Alg, achieved state-of-the-art performance in the field of machine learning-based allergenic sequence prediction. A positive relationship between the prediction performance and the size of pLMs was observed. Additionally, we constructed the latest allergenic sequence data set (SI Document S1) by collecting from well-known and peer-reviewed databases. To make our newly developed pLM-based models easily accessible to the scientific and industrial communities, we deployed them on a user-friendly and free-access web server (<https://f6wxpfd3sh.us-east-1.awsapprunner.com/>).

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jafc.3c07143>.

- S1. Detailed data set containing the allergenic and nonallergenic sequences (XLSX)
- S2. Hyperparameters finetuning of the pLM4Alg models (XLSX)
- S3. Performance comparison between PLM-based models with other existing models (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Yonghui Li – Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States; orcid.org/0000-0003-4320-0806; Phone: 785-532-4061; Email: yonghui@ksu.edu

Authors

Zhenjiao Du – Department of Grain Science and Industry, Kansas State University, Manhattan, Kansas 66506, United States; orcid.org/0000-0002-8492-4328

Yixiang Xu – Healthy Processed Foods Research Unit, Western Regional Research Center, USDA-ARS, Albany, California 94710, United States; orcid.org/0000-0003-3122-6101

Changqi Liu – School of Exercise and Nutritional Sciences, San Diego State University, San Diego, California 92182, United States; orcid.org/0000-0001-6526-5024

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jafc.3c07143>

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This is contribution No. 24-055-J from the Kansas Agricultural Experimental Station. This work was supported in part by the USDA National Institute of Food and Agriculture Hatch project (Grant Accession No. 7003330) and the USDA Pulse Crop Health Initiative project (Grant Accession No. 0439200). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

■ REFERENCES

- (1) Celebi Sozener, Z.; Ozdel Ozturk, B.; Cerci, P.; Turk, M.; Gorgulu Akin, B.; Akdis, M.; Altiner, S.; Ozbey, U.; Ogulur, I.; Mitamura, Y.; Yilmaz, I.; Nadeau, K.; Ozdemir, C.; Mungan, D.; Akdis, C. A. Epithelial Barrier Hypothesis: Effect of the External Exposome on the Microbiome and Epithelial Barriers in Allergic Disease. *Allergy* **2022**, *77* (5), 1418–1449.
- (2) Ogulur, I.; Pat, Y.; Ardicli, O.; Barletta, E.; Cevhertas, L.; Fernandez-Santamaria, R.; Huang, M.; Bel Imam, M.; Koch, J.; Ma, S.; Maurer, D. J.; Mitamura, Y.; Peng, Y.; Radzikowska, U.; Rinaldi, A. O.; Rodriguez-Coira, J.; Satitsuksanoa, P.; Schneider, S. R.; Wallmann, A.; Zhakparov, D.; Ziadlou, R.; Brüggemann, M.-C.; van de Veen, W.; Sokolowska, M.; Baerenfaller, K.; Zhang, L.; Akdis, M.; Akdis, C. A. Advances and Highlights in Biomarkers of Allergic Diseases. *Allergy* **2021**, *76* (12), 3659–3686.
- (3) Han, X.; Krempski, J. W.; Nadeau, K. Advances and Novel Developments in Mechanisms of Allergic Inflammation. *Allergy* **2020**, *75* (12), 3100–3111.
- (4) Breiteneder, H.; Diamant, Z.; Eiwegger, T.; Fokkens, W. J.; Traidl-Hoffmann, C.; Nadeau, K.; O'Hehir, R. E.; O'Mahony, L.; Pfaar, O.; Torres, M. J.; Wang, D. Y.; Zhang, L.; Akdis, C. A. Future Research Trends in Understanding the Mechanisms Underlying Allergic Diseases for Improved Patient Care. *Allergy* **2019**, *74* (12), 2293–2311.
- (5) Sharma, N.; Patiyal, S.; Dhall, A.; Pande, A.; Arora, C.; Raghava, G. P. S. AlgPred 2.0: An Improved Method for Predicting Allergenic Proteins and Mapping of IgE Epitopes. *Briefings in Bioinformatics* **2021**, *22* (4), bbaa294.
- (6) Goodman, R. E.; Ebisawa, M.; Ferreira, F.; Sampson, H. A.; Van Ree, R.; Vieths, S.; Baumert, J. L.; Bohle, B.; Lalithambika, S.; Wise, J.;

- Taylor, S. L. AllergenOnline: A Peer-Reviewed, Curated Allergen Database to Assess Novel Food Proteins for Potential Cross-Reactivity. *Mol. Nutr. Food Res.* **2016**, *60* (5), 1183–1198.
- (7) Lambrecht, B. N.; Hammad, H. The Immunology of the Allergy Epidemic and the Hygiene Hypothesis. *Nat. Immunol.* **2017**, *18* (10), 1076–1083.
- (8) Vlieg-Boerstra, B.; Groetch, M.; Vassilopoulou, E.; Meyer, R.; Laitinen, K.; Swain, A.; Durban, R.; Benjamin, O.; Bottse, R.; Grimshaw, K.; Netting, M.; O'Mahony, L.; de Jong, N.; Skypala, I. J. The Immune-Supportive Diet in Allergy Management: A Narrative Review and Proposal. *Allergy* **2023**, *78* (6), 1441–1458.
- (9) Bozzetto, S.; Carraro, S.; Giordano, G.; Boner, A.; Baraldi, E. Asthma, Allergy and Respiratory Infections: The Vitamin D Hypothesis. *Allergy* **2012**, *67* (1), 10–17.
- (10) Akdis, C. A. Allergy and Hypersensitivity: Mechanisms of Allergic Disease. *Curr. Opin Immunol.* **2006**, *18* (6), 718–726.
- (11) Rothenberg, M. E. The Climate Change Hypothesis for the Allergy Epidemic. *Journal of Allergy and Clinical Immunology* **2022**, *149* (5), 1522–1524.
- (12) Sharma, N.; Patiyal, S.; Dhall, A.; Devi, N. L.; Raghava, G. P. S. ChAlPred: A Web Server for Prediction of Allergenicity of Chemical Compounds. *Computers in Biology and Medicine* **2021**, *136*, 104746.
- (13) Wang, J.; Yu, Y.; Zhao, Y.; Zhang, D.; Li, J. Evaluation and Integration of Existing Methods for Computational Prediction of Allergens. *BMC Bioinformatics* **2013**, *14* (4), S1.
- (14) Yu, X.-X.; Liu, M.-Q.; Li, X.-Y.; Zhang, Y.-H.; Tao, B.-J. Qualitative and Quantitative Prediction of Food Allergen Epitopes Based on Machine Learning Combined with in Vitro Experimental Validation. *Food Chem.* **2023**, *405*, 134796.
- (15) Radauer, C.; Bublin, M.; Wagner, S.; Mari, A.; Breiteneder, H. Allergens Are Distributed into Few Protein Families and Possess a Restricted Number of Biochemical Functions. *Journal of Allergy and Clinical Immunology* **2008**, *121* (4), 847–852.
- (16) Scheurer, S.; Toda, M.; Vieths, S. What Makes an Allergen? *Clin. Exp. Allergy* **2015**, *45* (7), 1150–1161.
- (17) Sathe, S. K.; Liu, C.; Zaffran, V. D. Food Allergy. *Annu. Rev. Food Sci. Technol.* **2016**, *7*, 191–220.
- (18) FAO/WHO. Evaluation of Allergenicity of Genetically Modified Foods: Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. *Rome, Italy* **2001**, 22–25.
- (19) Dimitrov, I.; Bangov, I.; Flower, D. R.; Doytchinova, I. AllerTOP v.2—a Server for in Silico Prediction of Allergens. *J. Mol. Model* **2014**, *20* (6), 2278.
- (20) Maurer-Stroh, S.; Krutz, N. L.; Kern, P. S.; Gunalan, V.; Nguyen, M. N.; Limvipuvadh, V.; Eisenhaber, F.; Gerberick, G. F. AllerCatPro—Prediction of Protein Allergenicity Potential from the Protein Sequence. *Bioinformatics* **2019**, *35* (17), 3020–3027.
- (21) 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), 3698–3704.
- (22) 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.
- (23) Behbahani, M.; Rabiei, P.; Mohabatkari, H. A Comparative Analysis of Allergen Proteins between Plants and Animals Using Several Computational Tools and Chou's PseAAC Concept. *International Archives of Allergy and Immunology* **2020**, *181* (11), 813–821.
- (24) Westerhout, J.; Krone, T.; Snippe, A.; Babé, L.; McClain, S.; Ladics, G. S.; Houben, G. F.; Verhoeckx, K. CM. Allergenicity Prediction of Novel and Modified Proteins: Not a Mission Impossible! Development of a Random Forest Allergenicity Prediction Model. *Regul. Toxicol. Pharmacol.* **2019**, *107*, 104422.
- (25) Dang, H. X.; Lawrence, C. B. Allerdicator: Fast Allergen Prediction Using Text Classification Techniques. *Bioinformatics* **2014**, *30* (8), 1120–1128.
- (26) Goto, K.; Tamehiro, N.; Yoshida, T.; Hanada, H.; Sakuma, T.; Adachi, R.; Kondo, K.; Takeuchi, I. Novel Machine Learning Method AllerStat Identifies Statistically Significant Allergen-Specific Patterns in Protein Sequences. *J. Biol. Chem.* **2023**, *299*, 104733.
- (27) Nguyen, M. N.; Krutz, N. L.; Limvipuvadh, V.; Lopata, A. L.; Gerberick, G. F.; Maurer-Stroh, S. AllerCatPro 2.0: A Web Server for Predicting Protein Allergenicity Potential. *Nucleic Acids Res.* **2022**, *50* (W1), W36–W43.
- (28) Dimitrov, I.; Naneva, L.; Doytchinova, I.; Bangov, I. AllergenFP: Allergenicity Prediction by Descriptor Fingerprints. *Bioinformatics* **2014**, *30* (6), 846–851.
- (29) Du, Z.; Comer, J.; Li, Y. Bioinformatics Approaches to Discovering Food-Derived Bioactive Peptides: Reviews and Perspectives. *TrAC Trends in Analytical Chemistry* **2023**, *162*, 117051.
- (30) Du, Z.; Wang, D.; Li, Y. Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. *ACS Omega* **2022**, *7*, 25760.
- (31) Du, Z.; Li, Y. Review and Perspective on Bioactive Peptides: A Roadmap for Research, Development, and Future Opportunities. *Journal of Agriculture and Food Research* **2022**, *9*, 100353.
- (32) Nedyalkova, M.; Simeonov, V. Multivariate Chemometrics as a Strategy to Predict the Allergenic Nature of Food Proteins. *Symmetry* **2020**, *12* (10), 1616.
- (33) Naneva, L.; Nedyalkova, M.; Madurga, S.; Mas, F.; Simeonov, V. Applying Discriminant and Cluster Analyses to Separate Allergenic from Non-Allergenic Proteins. *Open Chemistry* **2019**, *17* (1), 401–407.
- (34) Islam, M. R.; Kashem, M. A.; Mia, L. AllerHybrid: A Hybrid System to Predict the Allergen Using K-Mer and Physicochemical Properties. In *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*; IEEE: Dhaka, Bangladesh, 2021; pp 1–6 DOI: [10.1109/ICEEICT53905.2021.9667943](https://doi.org/10.1109/ICEEICT53905.2021.9667943).
- (35) Du, Z.; Ding, X.; Xu, Y.; Li, Y. UniDL4BioPep: A Universal Deep Learning Architecture for Binary Classification in Peptide Bioactivity. *Briefings in Bioinformatics* **2023**, *24*, 1–10.
- (36) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16* (12), 1315–1322.
- (37) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44* (10), 7112–7127.
- (38) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123–1130.
- (39) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (15), No. e2016239118.
- (40) Du, Z.; Ding, X.; Hsu, W.; Munir, A.; Xu, Y.; Li, Y. pLM4ACE: A Protein Language Model Based Predictor for Antihypertensive Peptide Screening. *Food Chem.* **2024**, *431*, 137162.
- (41) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Advances in Neural Information Processing Systems* **2017**.
- (42) Mari, A.; Rasi, C.; Palazzo, P.; Scala, E. Allergen Databases: Current Status and Perspectives. *Curr. Allergy Asthma Rep* **2009**, *9* (5), 376–383.
- (43) Pomés, A.; Davies, J. M.; Gadermaier, G.; Hilger, C.; Holzhauser, T.; Lidholm, J.; Lopata, A.; Mueller, G. A.; Nandy, A.; Radauer, C.; Chan, S. K.; Jappe, U.; Kleine-Tebbe, J.; Thomas, W. R.; Chapman, M. D.; van Hage, M.; van Ree, R.; Vieths, S.; Raulf, M.;

Goodman, R. E. WHO/IUIS Allergen Nomenclature: Providing a Common Language. *Mol. Immunol.* **2018**, *100*, 3–13.

(44) van Ree, R.; Sapiter Ballerda, D.; Berin, M. C.; Beuf, L.; Chang, A.; Gadermaier, G.; Guevera, P. A.; Hoffmann-Sommergruber, K.; Islamovic, E.; Koski, L.; Kough, J.; Ladics, G. S.; McClain, S.; McKillop, K. A.; Mitchell-Ryan, S.; Narrod, C. A.; Pereira Mouries, L.; Pettit, S.; Poulsen, L. K.; Silvanovich, A.; Song, P.; Teuber, S. S.; Bowman, C. The COMPARE Database: A Public Resource for Allergen Identification, Adapted for Continuous Improvement. *Front. Allergy* **2021**, *2* DOI: [10.3389/falgy.2021.700533](https://doi.org/10.3389/falgy.2021.700533).

(45) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480–D489.