AI is transforming allergenic protein prediction

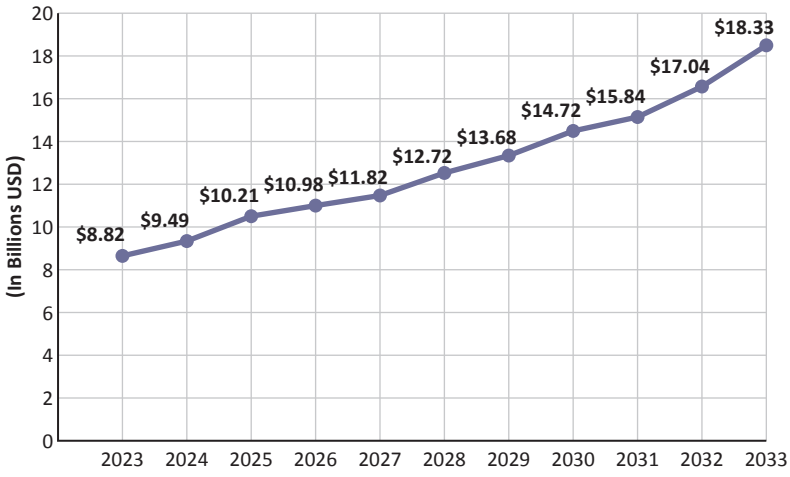
Zhenjiao Du and Yonghui Li

Concerns about hidden allergens in our food and environment are more pressing than ever, as allergies affect a growing portion of the global population. In the United States alone, over a third of the population suffers from allergic reactions. These can range from mild discomfort to severe, life-threatening events. As a result, the $9 billion allergy diagnostics and treatment market is predicted to double within ten years. The proteins and peptides responsible for these allergic responses are found everywhere—in the environment and in food and personal care products. Identifying them is a complex and critical task.

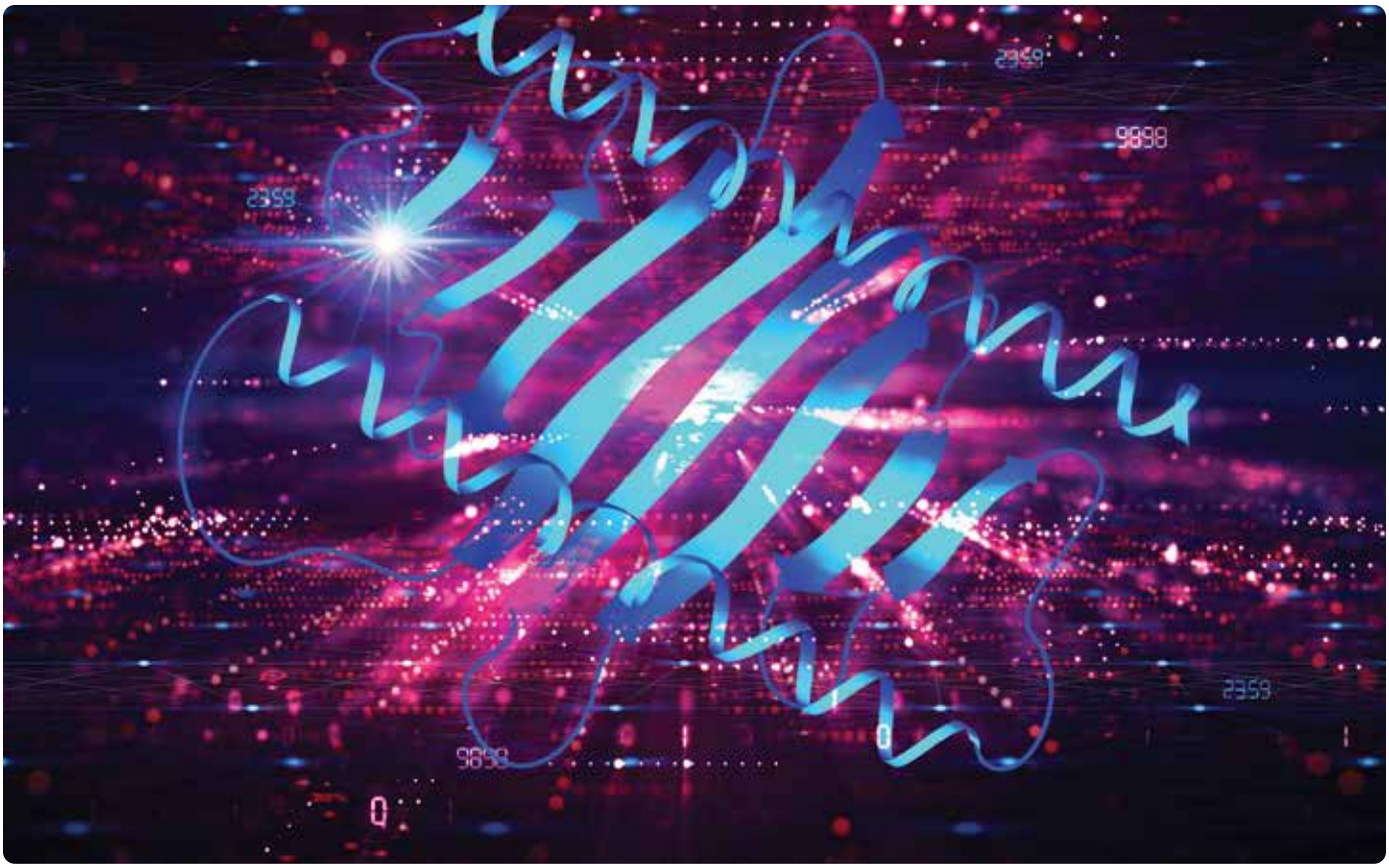
- **Artificial intelligence (AI) is revolutionizing how we predict allergenic proteins and peptides.**
- **The new tool, pLM4Alg, offers unprecedented accuracy in identifying allergenic compounds.**
- **This advancement could accelerate the development of hypoallergenic foods and ingredients.**
- **Accessible through a user-friendly web platform, pLM4Alg is freely available to researchers worldwide.**

Thanks to the advancements in artificial intelligence (AI), we can now predict which proteins might trigger allergic reactions before they reach our tables or become incorporated into products. Traditional methods for identifying allergenic proteins rely heavily on laboratory experiments and clinical trials. Although these approaches are effective, they are time-consuming, costly, and inefficient for large-scale screening. Testing every new protein through wet-lab experiments is challenging, especially given the myriad of proteins that exist.

Traditional allergen identification methods are like searching for a needle in a haystack. They require significant resources and may not keep pace with the rapid introduction of new proteins in food and consumer



Predicted US allergy diagnostics and therapeutics market size from 2023 to 2033. Source: https://tinyurl.com/2s37upt2



products. What is needed is an efficient large-scale screening method to identify potential allergens without the need for exhaustive lab work.

THE RISE OF AI IN ALLERGY RESEARCH

To address these challenges, scientists began exploring computational methods. Early efforts used knowledge-based approaches. For instance, in 2001, the Food and Agriculture Organization of the United Nations (FAO) and the World Health Organization (WHO) formed a joint consultation to evaluate the allergenicity of genetically modified foods and issued a report that included guidelines for evaluating the allergenic potential of proteins (https://tinyurl.com/2hjvdeb4). The report suggests standard methods for testing allergenicity based on a protein's amino acid sequence which requires laborious screening. An alternative approach is to use AI or machine learning which offer purely data-driven methods that can understand the traits of known allergens and apply generalized indicators to previously unseen allergens.

A major challenge for allergy prediction models has been how to effectively represent proteins and peptides numerically so that they can be analyzed by machine learning models. The early models that relied on simple features like amino acid composition struggled with complex sequences and novel proteins that did not closely resemble known allergens.

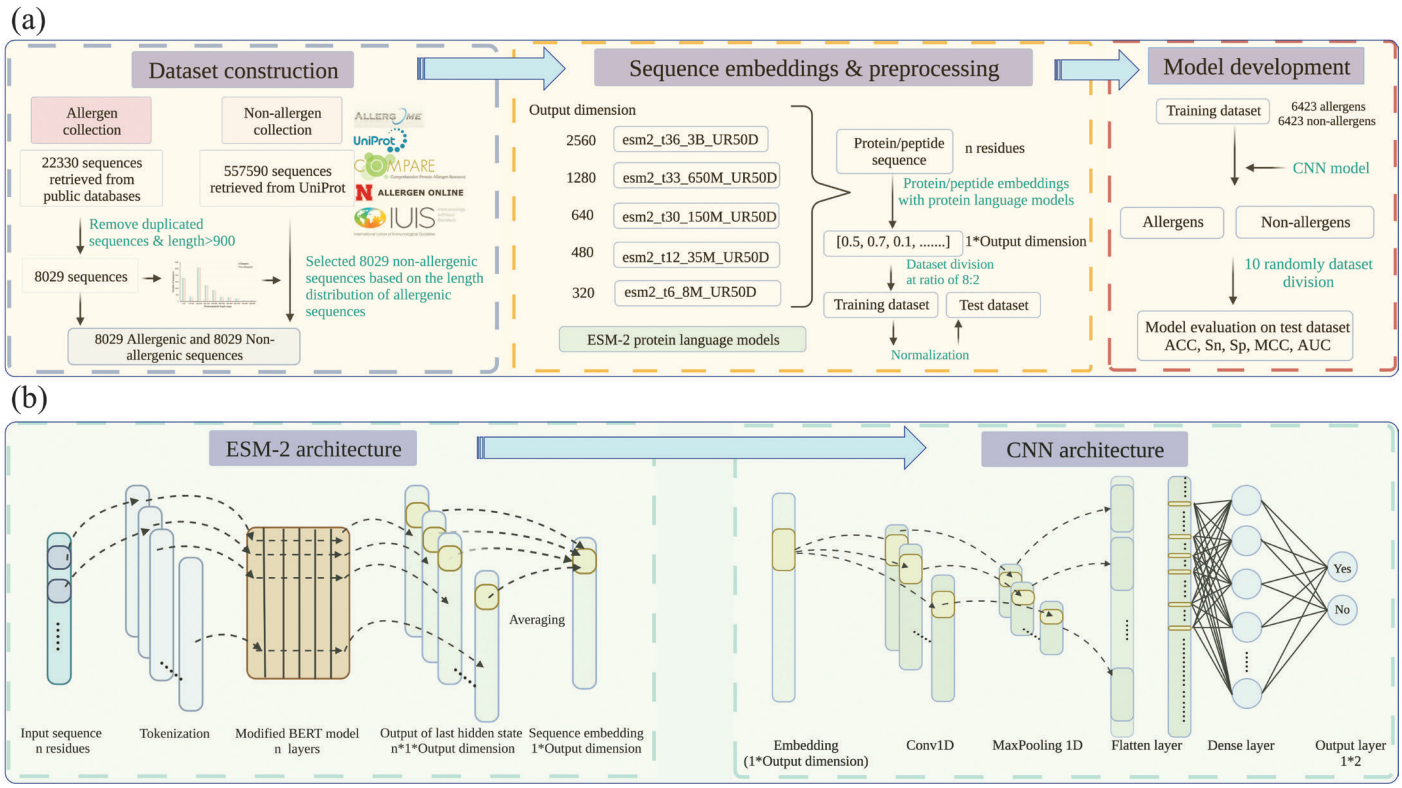
The introduction of protein language models (pLMs) marked a significant leap forward. Inspired by natural language processing, pLMs treat amino acid sequences like sentences and individual amino acids like words. This approach enables models to capture intricate patterns and contexts within protein sequences. The advancement paved the way for more accurate and efficient allergen prediction models, culminating in the development of tools like pLM4Alg (https://doi.org/10.1021/acs.jafc.3c07143).

INSPIRED BY NATURAL LANGUAGE PROCESSING, PLMS TREAT AMINO ACID SEQUENCES LIKE SENTENCES AND INDIVIDUAL AMINO ACIDS LIKE WORDS.

A REFINED MODEL

Our team developed pLM4Alg, a powerful tool that combines advanced protein language models with deep learning techniques. This method provides high-throughput predictions for determining whether a protein or peptide is likely to cause an allergic reaction.

The development of pLM4Alg began with the collec-tion of a comprehensive dataset. Our team compiled 8,029 known allergenic protein sequences and an equal number of non-allergenic sequences from reputable databases, includ-ing AllergenOnline, WHO/IUIS, ALLERGOME, COMPARE, and UniProt. This dataset provided a robust foundation for training and testing the model. Since our model was trained on high-quality, peer-reviewed data there is greater certainty about the accuracy and reliability of its predictions.



Schematic framework of pLM4Alg model development (a) and detailed architecture of ESM-2 protein language models and convolutional neural network (CNN) models (b). Source: Du, *et al.*, *J Agric Food Chem*, 72, 1, 2024.

Next, we used protein language models, specifically ESM-2, released by the Meta Fundamental AI Research Protein Team (FAIR) in 2022 (https://doi.org/10.1126/science.ade2574). The ESM-2 model captures not only the properties of individual amino acids but also the context of the entire sequence, providing a more nuanced understanding of the protein or peptide.

Finally, the team applied a convolutional neural network (CNN), a deep learning model well-suited to recognizing patterns in data. By combining the CNN with the ESM-2 model, we used transfer learning to train the system to recognize patterns indicative of allergenicity.

The results were impressive. The pLM4Alg models achieved state-of-the-art performance, with accuracy up to 95.1 percent and an area under the curve (AUC) scoring up to 0.99. The metrics indicate that pLM4Alg predicts potential allergens with a level of accuracy previously unattainable.

This level of accuracy has significant implications. It enhances our ability to assess allergenic risks efficiently and could lead to better safety measures in various industries.

WHAT SETS pLM4Alg APART

While traditional methods that compare new proteins to known allergens by looking for sequence similarities can be effective, they have limitations. They may overlook novel allergens that do not closely resemble known ones or struggle with sequences containing non-standard amino acids or missing residues.

pLM4Alg addresses these challenges. By using advanced protein language models like ESM-2, it can capture the complex relationships within protein sequences, including those with non-standard amino acids or missing residues. This makes it a more versatile and robust tool for allergen prediction, benefiting researchers and industry professionals alike.

IMPLICATIONS FOR FOOD AND PRODUCT SAFETY

The development of pLM4Alg has significant implications for the food industry, healthcare, and regulatory bodies. By providing a highly accurate method for predicting allergenicity, it can help manufacturers identify potential allergens early in the product development process. This can lead to the creation of safer foods and products, reducing the risk of allergic reactions among consumers.

For the food industry, identifying and eliminating potential allergens before products reach the market is crucial. pLM4Alg can streamline this process and save companies valuable time and resources.

For healthcare professionals, improved prediction models can assist in diagnosing allergies and developing treatments. By predicting which proteins are likely to cause allergic reactions, healthcare providers can offer better guidance to patients and develop more effective management strategies.

Regulatory agencies also stand to benefit from this advancement. Assessing the allergenic potential of new proteins, such as those introduced through genetic modification, is a critical part of ensuring public safety. pLM4Alg can reduce the burden of extensive laboratory testing, providing regulators with a more efficient, data-driven tool for decision-making.

A comparison of previous machine learning models that have been used to predict allergies, including pLM4Alg and previous representative machine learning-based allergenic sequence prediction models. Source: Du, *et al.*, *J Agric Food Chem*, 72, 1, 2024.

Project name*	Dataset size	ACC (percent)	Sn (percent)	Sp (percent)	MCC	AUC	Web server working	Year release
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	No	2021
AllerStat**	21154	N/A	N/A	N/A	0.4495	0.878	No	2023
AllerTOPv2	4854	88.7	86.7	90.7	0.775	N/A	Yes	2014
AllergenFP	4854	87.9	86.8	89.1	0.759	N/A	Yes	2014

Note: N/A: not available; ACC: accuracy; Sn: sensitivity; Sp: specificity; MCC: Matthews correlation coefficient; AUC: Area under the curve; *The 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.); **Imbalanced dataset (2248 allergenic sequences and 18906 nonallergenic sequences for AllerStat). A detailed reference list of previous models and a complete table are available in our original paper.

MAKING pLM4Alg ACCESSIBLE

Understanding the importance of accessibility, our team has made pLM4Alg available via a user-friendly web server (https://f6wxpfd3sh.us-east-1.awsapprunner.com). This platform allows researchers and industry professionals to input protein sequences and receive rapid predictions about their allergenic potential.

We aimed to make this tool accessible to everyone. By offering open access, we hope to accelerate advancements in allergy prevention and safety.

The web server supports various input formats, including .xls, .xlsx, .fasta, and .txt files for large-scale screening. This means that even those who might not be familiar with machine learning and programming can use the technology, making it a valuable resource across different sectors.

LOOKING AHEAD

The rise of allergies presents a pressing challenge, but innovations like pLM4Alg offer a promising solution. By harnessing the power of AI and protein language models, our team has developed a tool that increases our ability to predict allergenic proteins and peptides.

With its high accuracy, versatility, and accessibility, pLM4Alg is poised to become a useful resource for researchers, healthcare professionals, and the food industry. Our goal is to contribute to the development of safer products and better health outcomes.

While pLM4Alg represents a significant advancement, there are still challenges to address. One limitation is the processing of very long protein sequences—those exceeding 900 amino acid residues. Such sequences require substantial computational resources, and the dataset used to train pLM4Alg does not include sequences longer than 900 residues.

Another ongoing effort is to keep the model up-to-date with the latest data. As new allergens are discovered, incorporating them into the dataset is crucial for maintaining the model's relevance.

To provide even more comprehensive allergen assessments, our research team recommends integrating pLM4Alg with other prediction tools, especially those based on knowledge and prior experience. After potential allergenic proteins or peptides are identified through computational screening, rigorous laboratory testing and experimental validation, including immunoassays and *in vivo* assessments, are essential to confirm their allergenic potential and ensure accurate, reliable results.

Zhenjiao Du is a PhD candidate in the Department of Grain Science & Industry at Kansas State University, focusing on the application of artificial intelligence in food science. Yonghui Li is an associate professor and Du's advisor. His research explores the structure, chemistry, modification, and functionality of food proteins and bioactive peptides with the aim of developing high-quality, functional grain-based foods and ingredients. They can be contacted at yonghui@ksu.edu and zhenjiao@ksu.edu.