COMMENTARY

OPEN ACCESS

Revisiting computational immunogenicity prediction for novel food ingredients

Jiajun Liu,[1,7] Hua Cai,[2,7] Lingxin Luo,[1,3] Weiwei Shen,[1,3] Yixue Li,[1,4,5,6,*] Zhenni Zhu,[2,*] and Tao Huang[1,*]

[1]Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[2]Division of Health Risk Factors Monitoring and Control, National Health Commission Specialty Laboratory of Food Safety Risk Assessment and Standard Development, Shanghai Municipal Center for Disease Control and Prevention, Shanghai 201107, China

[3]School of Life Sciences, Shanghai University, Shanghai 200444, China

[4]Guangzhou National Laboratory, Guangzhou 510000, China

[5] The Guangdong-Hong Kong-Macau Joint Laboratory for Cell Fate Regulation and Diseases, Guangzhou Laboratory, Guangzhou Medical University, Guangzhou 510000, China

[6]Key Laboratory of Systems Health Science of Zhejiang Province, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

[7]These authors contributed equally

*Correspondence: li_yixue@gzlab.ac.cn (Y.L.); zhuzhenni@scdc.sh.cn (Z.Z.); huangtao@sibs.ac.cn (T.H.)

Received: February 16, 2026; Accepted: April 1, 2026; Published Online: April 7, 2026; https://doi.org/10.59717/j.xinn-nutri.2026.100008

Citation: Liu J., Cai H., Luo L., et al. (2026). Revisiting computational immunogenicity prediction for novel food ingredients. *The Innovation Nutrition* **1**:100008.

Synthetic biology now produces precision-fermented dairy proteins and engineered enzymes at rates exceeding the capacity of existing safety assessment pipelines. These novel ingredients may introduce immunogenicity risks undetectable by conventional methods. Regulatory frameworks such as the Food and Agriculture Organization of the United Nations/World Health Organization (FAO/WHO) sliding window approach classify proteins as potential allergens when sequence identity exceeds 35% over 80 amino acids, with animal testing providing supplementary validation.[1] Such homology-based rules can miss novel proteins with limited similarity to known allergens and can also over-flag proteins containing incidental sequence motifs.[1] Wet-lab validation offers definitive results, but limited throughput restricts its use in early-stage screening of large candidate libraries. In this commentary, immunogenicity is used broadly to include classical allergenicity and autoimmune cross-reactivity, whereas allergenicity refers specifically to IgE-mediated allergenicity.

Current protein allergenicity prediction faces systemic limitations: outdated reference databases underrepresent emerging protein families, feature engineering approaches generalize poorly across diverse sequences, and prediction frameworks exclude autoimmune cross-reactivity from consideration. Simply evaluating IgE-mediated allergenicity is no longer adequate for modern safety assessment. We therefore argue that future frameworks should move toward multi-class prediction distinguishing allergens, autoimmunity-related proteins, and non-allergenic proteins, coupled with residue-level outputs that generate localized immunogenicity hypotheses rather than only aggregate risk scores. Figure 1 summarizes this proposed next-generation framework.

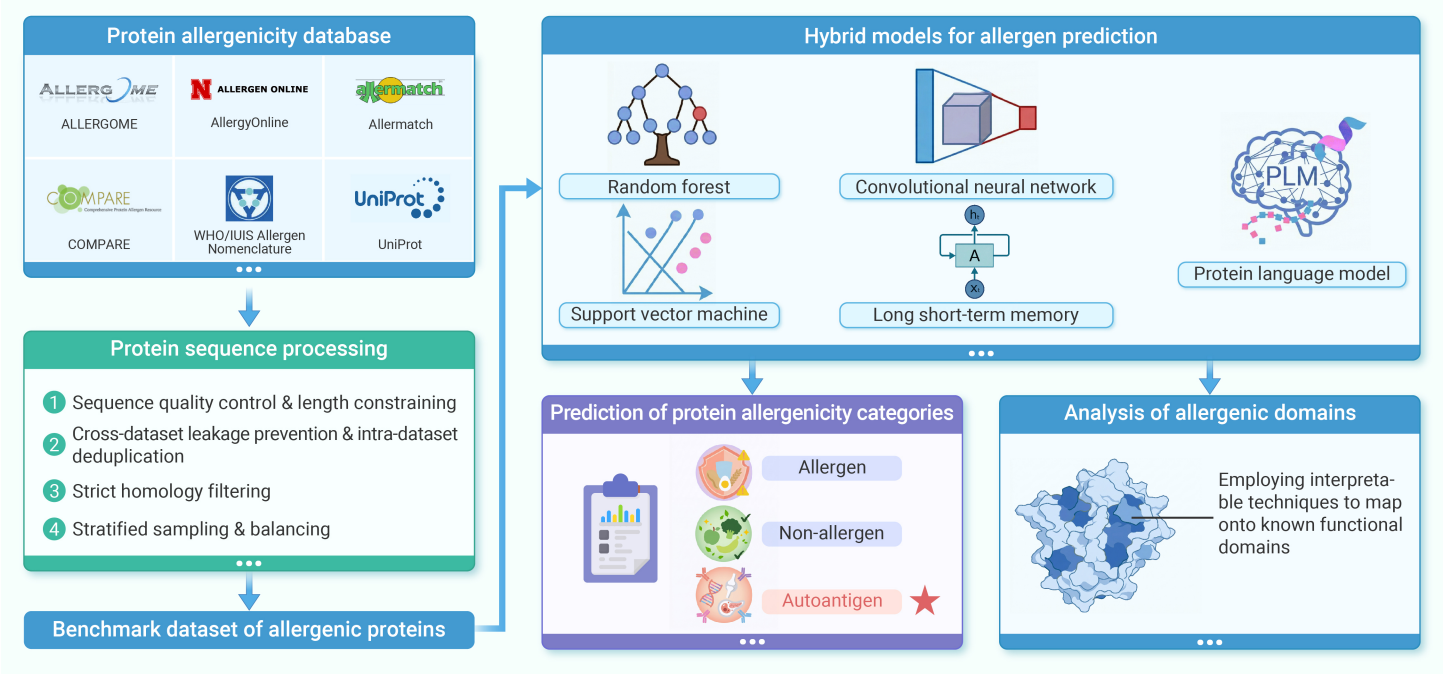

Figure 1. Conceptual framework for next-generation immunogenicity prediction

The future protein allergenicity prediction framework includes three stages: data integration, model construction, and interpretability evaluation. Multi-database data are preprocessed to build a benchmark allergenic protein dataset; one or a hybrid of algorithms (machine learning, deep learning, protein language models) is used for allergenicity classification; the framework finally outputs classification results and identifies key allergenic domains via interpretability analysis to improve the biological reliability of predictions.

DATA RESOURCE DEFECTS: STALENESS, HOMOLOGY LEAKAGE, AND LENGTH BIAS

Many recent predictors still rely on legacy datasets assembled several years earlier, introducing severe data staleness and homologous leakage into model evaluation. For example, AllergenAI (2025) still relies on AlgPred 2.0 (2021) for dataset construction.[2-3] Primary repositories like UniProt update daily, but published allergen benchmarks are typically static and are not routinely refreshed with newly curated annotations. Many benchmark

datasets also retain highly similar sequences from the same protein families, inflating apparent model accuracy through homologous leakage between training and test sets.

Sequence length filtering during preprocessing may also exclude biologically relevant proteins. AlgPred 2.0 excludes peptides shorter than 50 amino acids.[2] AllergenAI uses a fixed input length of 1,000 amino acids, padding shorter proteins and truncating longer chains.[3] Long-chain storage proteins and bioactive peptides therefore fall outside these boundaries, limiting applicability across the full spectrum of food-derived proteins. More broadly, staleness and leakage arise because benchmark datasets are often assembled as static snapshots and split randomly at the sequence level rather than by homology cluster or annotation time. As a result, newly curated allergens remain excluded, while near-identical family members can leak across training and test partitions and inflate apparent model performance. Time-stamped updates, exact deduplication, and family-aware external holdouts are therefore essential.

Current prediction frameworks classify proteins as allergens or non-allergens without considering autoimmune cross-reactivity. Gliadin illustrates this limitation: food-protein immunogenicity can extend beyond IgE-mediated allergenicity to autoimmune pathology, including celiac disease and tissue-transglutaminase-associated autoimmunity. More tentative associations have also been proposed for other food proteins, including milk proteins in type 1 diabetes-related autoimmunity, although causality remains debated.[4] These examples suggest that some food-protein risks may extend beyond classical allergenicity and are not captured by binary allergen/non-allergen frameworks. Incorporating autoimmunity-related data may therefore broaden exploratory risk assessment, although benchmark design and label definition remain challenging.

REPRESENTATION LIMITATIONS: HOMOLOGY DEPENDENCY AND STATIC FEATURE EXTRACTION

Allergenicity prediction methods span sequence alignment, handcrafted feature extraction, and deep representation learning, each with distinct trade-offs in coverage and computational requirements.

Sequence similarity approaches such as the FAO/WHO linear rule classify proteins using identity thresholds to known allergens. These methods depend on reference database completeness: proteins with low homology to catalogued allergens may escape detection, whereas proteins from the same family as known allergens may trigger false positives despite lacking immunogenic epitopes.[1]

Second-generation methods introduced machine learning algorithms. AlgPred 2.0 and related tools extract amino acid composition and physicochemical descriptors as input features for support vector machines or random forests.[2] This bypasses direct sequence alignment but requires multiple external descriptor-calculation steps, increasing computational overhead for large-scale screening. Representing proteins as fixed-length numerical vectors also discards residue positional information and long-range sequence dependencies that may contribute to allergenicity.

Third-generation methods integrate deep learning with protein language models (PLMs).[5] Recent allergenicity predictors such as AllergenAI have begun to combine PLM-derived representations with downstream neural architectures for feature extraction.[3] However, most strategies still treat PLMs largely as fixed feature extractors and train only downstream shallow heads. This setup limits task-specific adaptation and may leave immunogenicity-relevant features underused.

STRUCTURAL BLINDNESS: NEGLECT OF CONFORMATIONAL EPITOPES IN SEQUENCE MODELS

Sequence-only models retain a structural blind spot because many allergenic determinants are conformational rather than purely linear. Experimental structures deposited in the Protein Data Bank (PDB) cover only a fraction of known allergen sequences, and although AlphaFold and related tools can generate models for uncovered proteins, full-structure prediction remains computationally challenging at library scale. A practical near-term strategy is therefore to augment full-length sequence models with low-cost structural surrogates, such as predicted secondary structure, solvent accessibility, disorder, conformational epitope propensity, or coarse surface descriptors. Some models partially adopt this logic through predicted 3D similarity and surface-epitope comparison within a weight-of-evidence workflow, while AllergenAI suggests that coarse structure-derived descriptors may modestly improve performance.[3] Lightweight predictors, including ESMFold, OmegaFold, and hybrid sequence-plus-local-structure pipelines, may therefore make structural embedding more feasible for large-scale library screening.

FRAMEWORK RECONSTRUCTION: MULTI-CLASS TARGETS, FULL-LENGTH INPUTS, AND INTERPRETABILITY

Framework reconstruction should be defined at the level of prediction target, input granularity, and output interpretability rather than by model depth alone. One practical implementation would couple a long-context protein language model encoder with lightweight structure-aware features and a residue-attribution head trained jointly for three-class prediction. First, benchmark design should move from binary allergen/non-allergen labels to a three-class setting comprising allergens, autoimmunity-related proteins, and non-allergenic proteins. Such a benchmark should prioritize experimentally supported entries, apply family-aware de-redundancy and train/test separation, and use confidence tiers or manual curation to reduce label noise, particularly for autoimmunity-related assignments.

Second, models should accept full-length proteins instead of excluding short peptides or truncating long chains, because both short immunogenic fragments and multi-domain proteins may contain relevant risk signals. This can be implemented with long-context or hierarchical sequence encoders that preserve whole-chain information while remaining computationally tractable.

Third, interpretability should be residue-resolved rather than score-only. Integrated Gradients, attention rollout, or SHAP (SHapley Additive exPlanations) can generate residue-level hypotheses that can be compared with known epitopes and used to prioritize experimental redesign, for example by targeting surface-exposed high-attribution residues while preserving core functional regions. Taken together, these elements outline a testable framework for future method development in safe-by-design protein engineering.

REFERENCES

1. FAO/WHO (2001). Evaluation of allergenicity of genetically modified foods: Report of a joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology, 22–25 January 2001. Rome/Geneva: FAO/WHO. https://iris.who.int/handle/10665/340572
2. Sharma N., Patiyal S., Dhall A., et al. (2021). AlgPred 2.0: An improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief Bioinform.* **22**:bbaa294. DOI:10.1093/bib/bbaa294
3. Liu J., Negi S. S., Yang C., et al. (2025). AllergenAI: A deep learning model predicting allergenicity based on protein sequence. *BMC Bioinform.* **26**:279. DOI:10.1186/s12859-025-06302-1
4. Vojdani A., Gushgari L. R. and Vojdani E. (2020). Interaction between food antigens and the immune system: Association with autoimmune disorders. *Autoimmun. Rev.* **19**:102459. DOI:10.1016/j.autrev.2020.102459
5. Huang T. and Li Y. (2023). Current progress, challenges, and future perspectives of language models for protein representation and protein design. *The Innovation* **4**:100446. DOI:10.1016/j.xinn.2023.100446

FUNDING AND ACKNOWLEDGMENTS

This work was supported by Major Project of Guangzhou National Laboratory (GZNL2024A01003). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.