Residue-Level Attributions in Protein Language Models Do Not Recover Allergen Epitopes

Jianzhou Yao [1 2] **Anxiong Song** [1 2] **Katja Baerenfaller** [1 3] **Damir Zhakparov** [1 3]

Abstract

Deep allergenicity classifiers are increasingly used in safety screening of novel foods, and recent protein language models have substantially improved protein-level allergenicity prediction. However, whether their explanations capture biologically meaningful information remains unclear. We introduce an epitope-grounded residue-level benchmark for quantitatively evaluating attribution faithfulness in protein allergenicity models. Across frozen ESM-2, multi-task ESM-2, and DeepPlantAllergy, protein-level classification was robust, yet classification-head explanation signals did not significantly exceed random in their residue-level alignment with annotated epitopes across AUROC, AUPRC, and Precision@k. Integrated Gradients identified residues that were functionally important to the model, but not overlapping annotated epitopes. Saturation mutagenesis further suggested classifiers may rely on physicochemical and compositional sequence features rather than epitope-specific mechanisms. Residue-level importance signals should therefore not be interpreted as immunological explanations for safety screening or hypoallergen design without quantitative validation. Code available here.

1. Introduction

Allergenicity predictors are increasingly used in safety screening of novel foods and recombinant proteins, complementing experimental assays and expert reviews (Fernandez et al., 2021; EFSA Panel on Genetically Modified Organisms (GMO) et al., 2022). Recent advances in protein language models have substantially improved computational protein annotation, including allergenicity prediction (Lin et al., 2023; He et al., 2023; Dhouib et al., 2025). However, predictive accuracy alone is insufficient if models rely on immunologically implausible features.

In allergy, immune recognition targets specific regions of an allergenic protein called epitopes rather than the full protein, with epitope-level recognition shaping sensitization, clinical reactivity, and cross-reactivity (Møiniche et al., 2026). We therefore ask: *do strong protein-level allergenicity classifiers produce residue-level explanations that align with experimentally annotated epitopes?*

Recent allergenicity classifiers often visualize attribution scores, such as Integrated Gradients (IG) (Sundararajan et al., 2017), as residue-level heatmaps and show qualitative agreement with epitope motifs on individual protein examples (He et al., 2023; Dhouib et al., 2025; Liu et al., 2025), but their quantitative epitope alignment remains unknown.

We distinguish **model faithfulness** - whether an explanation highlights residues that influence the model's output, from **immunological faithfulness** - whether those residues align with experimentally observed immune-recognition sites.

Quantitative ground-truth evaluation of attributions has been pursued in other domains, including pixel-level masks for visual question answering (Arras et al., 2022) and clinician-validated features in medical imaging (Makino et al., 2022; Arun et al., 2021); curated epitope databases offer an analogous source of ground truth for the immunological faithfulness of allergenicity classifiers.

We therefore adopted rank-based metrics (AUROC, AUPRC) established for evaluating supervised epitope predictor performance under strong class imbalance (Clifford et al., 2022; Høie et al., 2024; Israeli & Louzoun, 2024; Zeng et al., 2023; Cia et al., 2023), applied here to explanation signals from allergenicity classifiers rather than models trained directly for epitope prediction. We further include a multi-task learning (MTL) model with auxiliary residue-level epitope supervision as an interventional diagnostic of epitope-relevant feature use.

Our contributions are:

[1]Swiss Institute of Allergy and Asthma Research, Davos, Switzerland [2]ETH Zurich, Zurich, Switzerland [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland. Correspondence to: Jianzhou Yao <yaojia@ethz.ch>.

Mechanistic Interpretability Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea, 2026.

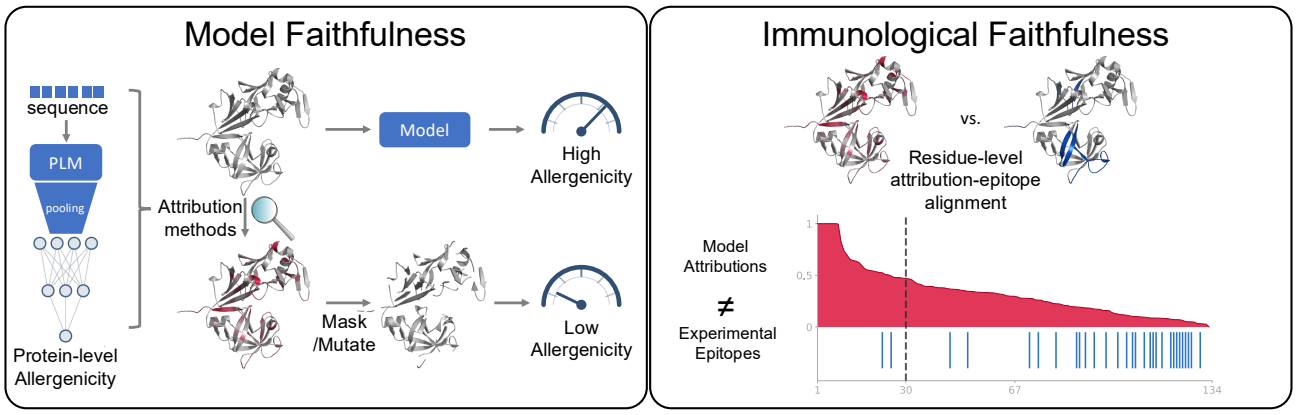

Figure 1. Graphical illustration of the two faithfulness criteria distinguished in this work. Left: Model faithfulness: residues identified by attribution methods causally influence the allergenicity prediction when masked, illustrated using the 3D structure of the allergenic protein P54958. Right: Immunological faithfulness: these residues need not coincide with experimentally annotated epitopes; blue ticks, IEDB epitopes; red profile, Integrated Gradients scores from the allergenic protein K7AKJ8 ranked from highest to lowest; dashed line, Precision@k cutoff.

1. We provide an epitope-grounded benchmark for quantitatively evaluating residue-level immunological faithfulness of allergenicity model explanations, using AUROC, AUPRC, and Precision@k.

2. We show that strong allergenicity classifiers consistently fail to produce epitope-aligned attribution signals better than random across all models and metrics.

3. We demonstrate a dissociation between model and immunological faithfulness: IG is model-faithful under masking, yet saturation mutagenesis reveals that models are sensitive to global physicochemical rather than epitope-specific features.

2. Benchmark and Methods

Residue-level faithfulness definition. Let $x = (x_1, \ldots, x_L)$ be a protein sequence and $F(x) \in [0, 1]$ a protein-level allergenicity classifier. A residue-level scoring method produces $s = (s_1, \ldots, s_L)$, where larger s_i indicates greater importance of residue i. For allergenic proteins with experimentally annotated allergy-associated MHC class II epitopes, we define a binary mask $y = (y_1, \ldots, y_L)$, where $y_i = 1$ if residue i belongs to an annotated epitope region after mapping to the parent protein. Immunological faithfulness asks whether high-scoring residues under s correspond to positive residues under y; model faithfulness asks whether they causally influence $F(x)$.

Protein-level allergenicity data. Protein-level classifiers were trained and evaluated using the released DeepAlgPro train/test FASTA files (He et al., 2023). Exact sequence overlaps with the epitope benchmark were removed prior to training and evaluation. The resulting cleaned held-out test set contained 1,377 sequences.

Epitope-grounded benchmark curation. Because allergenicity is mediated by epitope recognition, including conformational or linear B-cell epitopes and linear peptide fragments presented to T cells by antigen-presenting cells, we curated experimentally validated allergy-associated epitopes from the Immune Epitope Database (IEDB) (Vita et al., 2019). We retained positive assays, restricted to allergic-disease contexts, and included both B-cell and T-cell epitopes under MHC class II restriction. Full-length parent proteins were retrieved through UniProt (The UniProt Consortium, 2023) and NCBI (Sayers et al., 2025). We excluded sequences with non-canonical amino acids, invalid epitope coordinates, viral or human origin, individual epitopes covering more than 25% of the parent protein, merged epitope coverage exceeding 75%, duplicate sequences, and sequences exceeding the ESM-2 context length.

Candidate negatives were retrieved from UniProtKB/Swiss-Prot reviewed entries, excluding viral and human proteins, allergen annotations, allergenic-related text, antigen- or cancer-related entries, and positive-homologous sequences. We further required evidence at protein level, annotation score 5, and annotated three-dimensional structural availability. Homology-aware splitting prevented clustered sequences from appearing in both training and evaluation splits. Full curation details are provided in Appendix A.

Models. We evaluated three classifiers.

Frozen ESM-2 used `esm2_t6_8M_UR50D` with learned attention pooling and a two-layer MLP classification head (Lin et al., 2023). The ESM-2 backbone was frozen;

only the attention pooling and classifier layers were trained.

MTL ESM-2 shares the same frozen ESM-2 backbone, initialized from the trained classification head best checkpoint, and adds an auxiliary residue-level epitope prediction head. The model was trained jointly on protein-level allergenicity labels and residue-level epitope masks:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{epi}}\mathcal{L}_{\text{epi}}. \tag{1}$$

MTL ESM-2 served as a principled intervention motivated by the MTL principle that auxiliary supervision on related tasks can improve primary task generalization through shared representation (Caruana, 1997; Ruder, 2017). If epitope-relevant residue information were accessible to the allergenicity classifier, injecting residue-level epitope supervision should improve protein-level classification or produce more epitope-aligned classification-head explanations. The auxiliary residue head was a diagnostic supervised signal, not a proposed epitope predictor or classification head explanation.

DeepPlantAllergy was included as a published external allergenicity architecture (Dhouib et al., 2025), retrained from scratch on our curated dataset to enable controlled comparison and avoid data leakage. Architecture details are in Appendix B.

Residue-level scores. We evaluated IG, occlusion, and the auxiliary MTL residue head. IG computes path-integrated sensitivity from a zero-embedding baseline to the input embedding (Sundararajan et al., 2017). Occlusion masked one residue at a time with the ESM-2 mask token and scored the resulting allergenicity decrease, $s_i^{\text{occ}} = F(x) - F(x_{\setminus i})$. Additional signals, including pooling attention weights, Gradient×Input, SmoothGrad-IG, and MTL classifier attributions, are defined and evaluated in Appendix C and E.

Metrics. Residue-level evaluation was performed on the 61 held-out splitB allergenic proteins with epitope annotations. Following epitope-prediction evaluations, we used established metrics AUROC and AUPRC to assess threshold-independent residue-level ranking under class imbalance (Clifford et al., 2022; Høie et al., 2024; Israeli & Louzoun, 2024; Zeng et al., 2023; Cia et al., 2023). For each protein, the score vector s was compared with the binary epitope mask y. We computed metrics within each protein and then averaged across proteins. Random baselines were computed identically using repeated random residue scores. AUROC measures whether epitope residues are ranked above non-epitope residues, while AUPRC emphasizes performance on sparse positive residues. We additionally reported Precision@k as a top-ranked localization metric, where k was set for each protein to their respective number of annotated epitope residues, $k = \sum_i y_i$:

$$\text{P@}k = \frac{|\text{Top}_k(s) \cap \{i : y_i = 1\}|}{k}. \tag{2}$$

For statistical testing, each non-random residue-level signal was compared against the corresponding random baseline from the same model family using a paired Wilcoxon signed-rank test across proteins. Tests were performed separately for AUROC, AUPRC, and Precision@k, and Benjamini–Hochberg correction was applied within each metric family.

Functional validation. To test model faithfulness, we ranked residues in the Frozen ESM-2 baseline by IG and simultaneously masked the top-k fraction, measuring $\Delta p = F(x) - F(x_{\text{masked}})$. The control consisted of random masking of equal size; IG-guided masking should reduce $F(x)$ more if IG is model-faithful. Full sweep details are in Appendix H.

Saturation mutagenesis. To characterize the decision basis of the Frozen ESM-2 baseline classifier, we used *in silico* saturation mutagenesis, a perturbation-based interpretability approach in which systematic single-residue substitutions are used to measure changes in model output (Lim et al., 2022; Koido et al., 2024). While IG and occlusion identify *which* residues the model relies on, saturation mutagenesis probes *why*: by exhaustively substituting each residue and measuring the resulting change in predicted allergenicity, it tests whether this reliance reflects epitope-specific residue identity or broader physicochemical properties shared across many positions. Each residue was substituted with each of the other 19 alternative amino acids, recording $\Delta p_{i,a\to b} = F(x) - F(x_{i:a\to b})$. Effects were aggregated by original amino acid and by coarse residue classes (charge, polarity, hydrophobicity, aromaticity); class definitions and transition summaries are in Appendix I.1.

3. Results and Discussion

3.1. Strong classification does not imply epitope-relevant attribution

We first ask whether strong protein-level performance is accompanied by epitope-aligned attribution signals. Despite all evaluated models achieved strong protein-level allergenicity classification on the held-out DeepAlgPro test set (Table 1), no model-derived attribution signal exceeded its corresponding random baseline, and several were significantly lower than random, indicating that these signals did not capture epitope-relevant information and may instead reflect other model-internal or biological information (Figure 2). This pattern persisted in the full signal comparison (Figure 6). A label-scrambling negative control confirmed that the evaluation was well calibrated: scrambling epitope

Table 1. Protein-level classification performance on the held-out DeepAlgPro test set ($n = 1377$). Additional metrics in Table 2.

Model	AUROC	F1	MCC
Frozen ESM-2	0.970	0.917	0.835
MTL ESM-2	0.962	0.910	0.821
DeepPlantAllergy	0.974	0.939	0.878

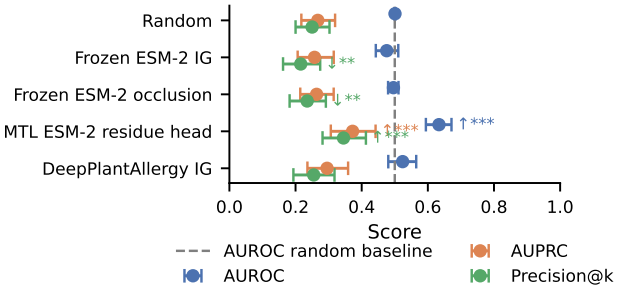

Figure 2. Residue-level alignment between model-derived scores and IEDB epitope annotations for representative methods. Error bars denote 95% bootstrap confidence intervals (CIs). Significance markers denote adjusted p-values: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; arrows indicate direction (↑ increase, ↓ decrease). Numerical values are reported in Table 3.

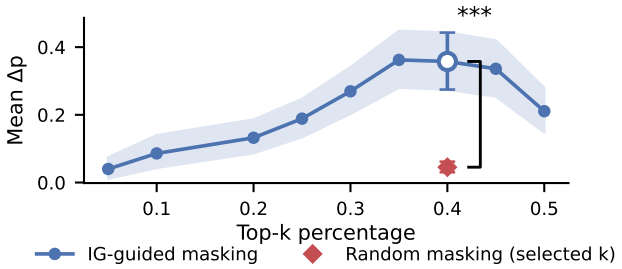

Figure 3. Functional masking analysis for Frozen ESM-2 using IG. The blue curve shows the mean change in predicted allergenicity when masking the top-k IG-ranked residues; shaded regions denote bootstrap 95% CIs. The open blue marker denotes the selected k. The bracket compares IG-guided and random masking at $k = 40\%$ using a paired two-sided Wilcoxon signed-rank test across proteins; *** indicates $p < 0.001$.

labels substantially reduced the performance of the supervised MTL residue head, while leaving the already near-random attribution methods essentially unchanged (Figure 7).

We next test whether explicit epitope supervision improves alignment. The auxiliary residue-level head in MTL ESM-2 learned directly supervised epitope signals, but incorporating this supervision improved neither protein-level classification performance nor the alignment between classification-head attributions and experimentally validated epitope positions. This null result was consistent with prior work showing that MTL improves generalization only when tasks cooperate at the representation level (Standley et al., 2020; Wu et al., 2020). Although a frozen backbone could in principle limit representation-level transfer, we additionally trained an exploratory MTL variant with the top ESM-2 layer unfrozen during joint training; despite enabling representation-level adaptation, it showed the same qualitative pattern of weak classification-head epitope alignment (Figure 6). The failure of explicit epitope supervision to transfer to the classification head therefore provided evidence that protein-level allergenicity classification can be achieved without relying on localized epitope-relevant features.

3.2. IG is model-faithful but immunologically misaligned

Poor epitope alignment could reflect either uninformative attributions or accurate identification of model-relevant residues that were not epitope positions. We distinguished these possibilities for the Frozen ESM-2 baseline using the functional masking test defined in Section 2.

Residues were ranked by IGs, and the top-k fraction was simultaneously masked to assess their influence on the model output. The primary analysis was restricted to high-confidence predictions ($F(x) > 0.70$, $n = 46$ proteins). At the selected operating point $k = 40\%$, IG-guided masking reduced predicted allergenicity significantly more than random masking of equal size (Figure 3), with the effect persisting without the confidence filter (Figure 8). Details of operating point selection are provided in Appendix H.

These results indicated that IG identified residues that causally influenced the model's predictions under this perturbation test. Yet, given the weak alignment with annotated epitopes, this implied the gap between model faithfulness and immunological faithfulness: the model relied on residue-level signals that are predictive but not epitope-specific.

3.3. Mutagenesis reveals global physicochemical rather than epitope-specific sensitivity

To further characterize the decision basis of the Frozen ESM-2 classifier, we performed saturation mutagenesis on the 37 proteins validated by the IG-masking criterion at $k = 40\%$.

Individual amino acids differed substantially in both the fraction of substitutions that reduce predicted allergenicity and the mean effect size Δp (Figure 4). Methionine and arginine showed negative mean Δp, indicating that substituting them tended, on average, to increase predicted allergenicity, while glycine exhibited the highest fraction of reducing substitutions and the largest mean Δp, followed by lysine and cysteine.

To assess whether this sensitivity reflects biochemical structure, the class-level transition heatmaps revealed that it was structured along physicochemical axes (residue classes defined in Table 4; Figures 9 and 10). Charge-altering tran-

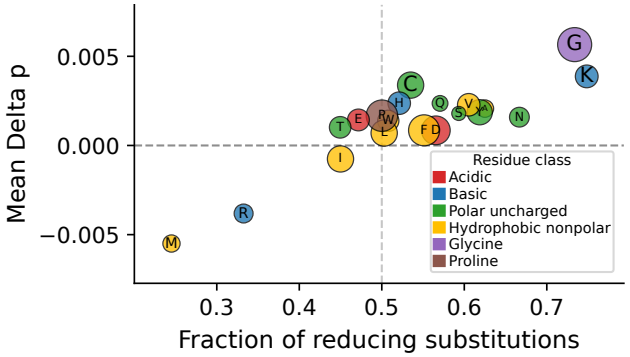

Figure 4. Saturation mutagenesis sensitivity by amino acid. Each point represents a wild-type amino acid; point size reflects its frequency across all evaluated proteins. The x-axis shows the fraction of substitutions that reduced predicted allergenicity (i.e., the fraction of substitutions originating from that amino acid with $\Delta p > 0$ across all positions and proteins). Positive values indicate a net reduction in predicted allergenicity.

sitions produced larger effects than within-class substitutions, consistent with the role of charge state in electrostatic interactions at protein surfaces and epitope recognition sites (Zhou et al., 2024; 2023). Within-aromatic and within-hydrophobic-aliphatic substitutions showed smaller effects, indicating lower model sensitivity to changes that preserved hydrophobic or aromatic character; the solvent-accessible surface area of aromatic residues correlates with IgE-binding capacity in food allergens (Zhou et al., 2023). Glycine and proline showed elevated cross-class effects in both heatmaps: glycine substitutions introduce a side chain where none existed, altering backbone turn capacity, while proline removal relieves its constrained ϕ angle and *cis*-peptide bond preference. Cysteine sensitivity reflects the multifunctional role of its sulfhydryl group, which participates in disulfide-stabilized structure, enzymatic active sites, and post-translational modification sites including lipidation (Pekar et al., 2018).

Together, these structured sensitivities were consistent with reliance on global physicochemical and compositional features rather than direct localization within annotated linear epitope regions. Mutagenesis effects across all 61 splitB proteins were significantly larger outside epitope residues than inside (Wilcoxon $p = 0.021$; Figure 11), further arguing against epitope-specific sensitivity.

4. Conclusion and Future Work

We demonstrated a consistent dissociation between predictive performance, model faithfulness, and immunological faithfulness: while the three tested models achieved strong protein-level classification performance and produced reliable attribution signals, they did not reliably recover annotated MHC class II epitopes.

The present analysis is limited by incomplete and biased IEDB annotations, restriction to linear MHC-II-derived residue masks, modest benchmark size, and the use of *in silico* perturbations that approximate rather than experimentally validate immune recognition.

These findings have direct implications for downstream use. In hypoallergen design, modifying residues highlighted by attribution methods may not reduce epitope content if these signals are not epitope-aligned. In safety screening, attribution visualizations may support model auditing but should not be interpreted as evidence of epitope-level understanding. More broadly, these results motivate biologically constrained predictors that incorporate structural and immunological priors, and call for standardized, epitope-grounded benchmarks to evaluate whether model explanations capture biologically meaningful mechanisms. They also motivate future mechanistic interpretability studies using probing and causal interventions to investigate how allergenicity-relevant information is represented within protein language models.

References

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

Arras, L., Osman, A., and Samek, W. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022.

Arun, N., Gaw, N., Singh, P., Chang, K., et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6), 2021.

Caruana, R. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.

Cia, G., Pucci, F., and Rooman, M. Critical review of conformational B-cell epitope prediction methods. *Briefings in Bioinformatics*, 24(1):bbac567, 2023. doi: 10.1093/bib/bbac567.

Clifford, J. N., Høie, M. H., Deleuran, S., Peters, B., Nielsen, M., and Marcatili, P. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Science*, 31(12):e4497, 2022. doi: 10.1002/pro.4497.

Dhouib, W., Frikha, F., Rebai, A., and Kharrat, N. DeepPlantAllergy: Deep learning for explainable prediction of allergenicity in plant proteins. *Briefings in Bioinformatics*, 26(6):bbaf605, 2025. doi: 10.1093/bib/bbaf605.

EFSA Panel on Genetically Modified Organisms (GMO), Mullins, E., Bresson, J.-L., Dalmay, T., Dewhurst, I. C.,

Epstein, M. M., Firbank, L. G., Guerche, P., Hejatko, J., Naegeli, H., Nogué, F., Rostoks, N., Sánchez Serrano, J. J., Savoini, G., Veromann, E., Veronesi, F., Fernandez Dumont, A., and Moreno, F. J. Scientific opinion on development needs for the allergenicity and protein safety assessment of food and feed products derived from biotechnology. *EFSA Journal*, 20(1):e07044, 2022. doi: 10.2903/j.efsa.2022.7044.

Fernandez, A., Mills, E. N. C., Koning, F., and Moreno, F. J. Allergenicity assessment of novel food proteins: What should be improved? *Trends in Biotechnology*, 39(1):4–8, 2021. doi: 10.1016/j.tibtech.2020.05.011. Epub 13 June 2020.

Fleri, W., Vaughan, K., Salimi, N., Vita, R., Peters, B., and Sette, A. The immune epitope database: How data are entered and retrieved. *Journal of Immunology Research*, 2017:5974574, 2017. doi: 10.1155/2017/5974574.

He, C., Ye, X., Yang, Y., Hu, L., Si, Y., Zhao, X., Chen, L., Fang, Q., Wei, Y., Wu, F., and Ye, G. DeepAlgPro: An interpretable deep neural network model for predicting allergenic proteins. *Briefings in Bioinformatics*, 24(4): bbad246, 2023. doi: 10.1093/bib/bbad246.

Høie, M. H., Gade, F. S., Johansen, J. M., Würtzen, C., Winther, O., Nielsen, M., and Marcatili, P. DiscoTope-3.0: Improved B-cell epitope prediction using inverse folding representations. *Frontiers in Immunology*, 15: 1322712, 2024. doi: 10.3389/fimmu.2024.1322712.

Israeli, S. and Louzoun, Y. Single-residue linear and conformational B cell epitopes prediction using random and ESM-2 based projections. *Briefings in Bioinformatics*, 25 (2):bbae084, 2024. doi: 10.1093/bib/bbae084.

Kapishnikov, A., Venugopalan, S., Avci, B., Wedin, B., Terry, M., and Bolukbasi, T. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5050–5058, 2021. doi: 10.1109/CVPR46437.2021.00501.

Koido, M., Tomizuka, K., and Terao, C. Fundamentals for predicting transcriptional regulations from dna sequence patterns. *Journal of Human Genetics*, 2024. doi: 10.1038/s10038-024-01256-3.

Lim, Y. W., Adler, A. S., and Johnson, D. S. Predicting antibody binders and generating synthetic antibodies using deep learning. *mAbs*, 2022. doi: 10.1080/19420862.2022.2069075.

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., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Liu, J., Negi, S. S., Yang, C., Zhou, X., Schein, C. H., Braun, W., and Kim, P. AllergenAI: A deep learning model predicting allergenicity based on protein sequence. *BMC Bioinformatics*, 26(1):279, 2025. doi: 10.1186/s12859-025-06302-1.

Makino, T., Jastrzębski, S., Oleszkiewicz, W., et al. Differences between human and machine perception in medical diagnosis. *Scientific Reports*, 12:6877, 2022.

Møiniche, M., Corneliussen, J., Johansen, K. H., Ruiz-Carrasco, A., Paulsen, C., Li, Y., Barra, C., Bangaru, S., Fernández-Quintero, M. L., Bartko, E., Blom, L., and Rivera-de Torre, E. Mapping allergen B- and T-cell epitopes: Technological advances and their role in precision allergy therapy. *Allergy*, 2026. doi: 10.1111/all.70396. URL https://doi.org/10.1111/all.70396. Early View, published online 22 May 2026.

Pekar, J., Ret, D., and Untersmayr, E. Stability of allergens. *Molecular Immunology*, 100:14–20, 2018. doi: 10.1016/j.molimm.2018.03.017.

Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., Feldgarden, M., Fine, A. M., Funk, K., Hoffman, J., Kannan, S., Kelly, C., Klimke, W., Kim, S., Lathrop, S., Marchler-Bauer, A., Murphy, T. D., O'Sullivan, C., Schmieder, E., Skripchenko, Y., Stine, A., Thibaud-Nissen, F., Wang, J., Ye, J., Zellers, E., Schneider, V. A., and Pruitt, K. D. Database resources of the national center for biotechnology information in 2025. *Nucleic Acids Research*, 53(D1):D20–D29, 2025. doi: 10.1093/nar/gkae979.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. SmoothGrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Standley, T., Zamir, A. R., Chen, D., Guibas, L., Malik, J., and Savarese, S. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9120–9132. PMLR, 2020.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

The UniProt Consortium. UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 2019. doi: 10.1093/nar/gky1006.

Wu, S., Zhang, H. R., and Ré, C. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

Zeng, Y., Wei, Z., Yuan, Q., Chen, S., Yu, W., Lu, Y., Gao, J., and Yang, Y. Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model. *Bioinformatics*, 39(4):btad187, 2023. doi: 10.1093/bioinformatics/btad187.

Zhou, X., Ren, L., Zhang, Y., Zhang, J., Li, X., Yang, A., Tong, P., Wu, Z., and Chen, H. Effect of structural targeted modifications on the potential allergenicity of peanut allergen Ara h 2. *Journal of Agricultural and Food Chemistry*, 71(1):836–845, 2023. doi: 10.1021/acs.jafc.2c06359.

Zhou, X., Zhang, M., Hu, C., Li, X., Yang, A., Tong, P., Wu, Z., and Chen, H. Effect of critical amino acids' properties on potential allergenicity of Ara h 2 epitopes. *Food and Agricultural Immunology*, 35(1), 2024. doi: 10.1080/09540105.2024.2373064.

A. Dataset Curation Details

IEDB epitope entries are compiled from two primary sources: peer-reviewed literature curated by IEDB staff and direct submissions from researchers (Fleri et al., 2017).

Positive epitope set. The initial IEDB-derived positive table contained 4,221 epitope rows after accession resolution and sequence retrieval; they consisted primarily of linear epitopes, as no conformational epitopes were available under these filters. Removing non-canonical amino acid sequences left 4,146 rows. Proteins requiring coordinate clipping were removed rather than clipped, leaving 3,991 rows. Individual epitopes spanning more than 25% of the parent protein were removed as biologically implausible for linear immune-recognition fragments. Surviving intervals were merged, and proteins whose merged epitope coverage exceeded 75% were removed as likely annotation artefacts. This yielded 416 proteins after the coverage filter, 391 after viral and human-source removal, and 374 after the maximum-length filter ($\leq$ 1022 aa, imposed by the ESM-2 context window).

Negative set. The initial UniProt-derived negative set contained 11,139 reviewed entries after applying the filters described in Section 2. Removing non-canonical sequences left 11,118 entries. Applying the maximum-length filter ($\leq$ 1022 aa) left 10,140 entries. Sequence-level redundancy was removed with `mmseqs easy-cluster` ($\geq$40% sequence identity, $\geq$80% bidirectional coverage, `--cov-mode 0`), yielding 7,572 cluster representatives. Positive-homologous negatives were identified with `mmseqs easy-search` against the full positive FASTA ($\geq$30% identity, $\geq$80% coverage, `--cov-mode 0`) and removed, leaving 7,477 clean representatives. Nearest-length greedy matching (random state 13) subsampled the 7,477 clean negative representatives to 374 negatives, matching the 374 positive proteins and producing a 1:1 positive-negative benchmark pool before splitting.

Splits. Positive proteins were grouped by their MMseqs2 cluster representative; negative proteins, already reduced to one representative per cluster, each formed their own group. A custom greedy algorithm assigned entire groups to the 80/20 split to jointly balance total samples, positives, and negatives across splits (random state 13), ensuring no homologous positive sequences appeared in both splits. This yielded 313 positive and 299 negative proteins in splitA, and 61 positive and 75 negative proteins in splitB. The residue-level faithfulness benchmark used the 61 splitB positive proteins with epitope masks. The mean positive-test epitope density was approximately 0.250.

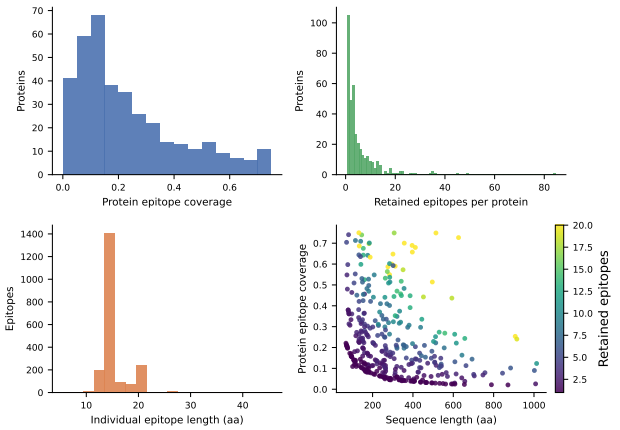

Figure 5. Positive epitope benchmark profile. The panels summarize protein-level epitope coverage, retained epitopes per protein, individual epitope lengths, and the relationship between sequence length, coverage, and retained epitope count.

B. Implementation Details

Frozen ESM-2. The frozen ESM-2 classifier used `esm2_t6_8M_UR50D`, hidden dimension 128, dropout 0.3, threshold 0.5, IG with 50 steps, and maximum sequence length 1022. The ESM-2 backbone was frozen; only learned attention pooling and the classification head were trained.

MTL ESM-2. The MTL model was initialized from the trained Frozen ESM-2 checkpoint. The ESM-2 backbone, learned attention pooling module, and protein-level classification head were reused, and a new auxiliary residue-level epitope head was added on top of the per-residue ESM-2 hidden states. In the reported frozen-MTL setting, the ESM-2 backbone remained frozen; the trainable components were the protein-level head, attention-pooling parameters, and the newly initialized auxiliary residue-level epitope head.

Training used mixed batches containing allergenic proteins with residue-level epitope masks and non-allergenic proteins without residue-level epitope supervision. Protein-level binary cross-entropy was computed for all sequences. Residue-level binary cross-entropy was computed only at valid residue positions for positive proteins with epitope annotations, using a residue mask to exclude proteins without epitope supervision from the epitope loss. The residue loss used positive-class weighting to reduce imbalance between sparse epitope residues and more abundant non-epitope residues. Specifically, for residue i,

$$\ell_i = -w_+ y_i \log \sigma(z_i) - (1 - y_i) \log(1 - \sigma(z_i)), \quad (3)$$

and the masked residue loss is

$$\mathcal{L}_{\text{epi}} = \frac{\sum_i m_i \ell_i}{\sum_i m_i}, \qquad w_+ = \frac{N_-}{N_+}. \quad (4)$$

Here z_i is the residue-level logit, m_i indicates valid supervised residue positions, and w_+ is computed from the ratio of non-epitope to epitope residues in the positive training proteins. This weighting avoided downsampling the majority non-epitope class, which is important given the limited amount of experimental epitope supervision. Protein-level class weighting was disabled because the cleaned DeepAlgPro training split was close to balanced.

Hyperparameters: batch size 24, maximum 30 epochs, early-stopping patience 5, learning rate 10^{-3}, weight decay 10^{-4}, $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{epi}} = 0.5$, epitope-head hidden dimension 128, validation fraction 0.1, 100 random attribution draws, and IG internal batch size 1. The best checkpoint was selected by validation total loss.

DeepPlantAllergy. The DeepPlantAllergy benchmark used ESM-1b embeddings with dimension 1280, convolutional filters, a three-layer bidirectional LSTM, multi-head self-attention with eight heads, adaptive average pooling, and fully connected layers. The maximum sequence length is 1000 for the ESM-1b embedding workflow.

Exploratory top-1-unfrozen MTL. As a robustness control for the frozen-backbone MTL setting, we additionally trained an exploratory MTL checkpoint in which the top ESM-2 layer is unfrozen during joint training. This checkpoint was included in the supplementary all-signal analysis but not in the primary protein-level model comparison.

C. Residue-Level Signal Definitions

The main paper reports a curated subset of residue-level scores. Here we define the full active signal inventory.

Integrated Gradients. Integrated Gradients attributes importance by integrating gradients along a path from a baseline embedding e' to the observed input embedding e (Sundararajan et al., 2017):

$$s_i^{\text{IG}} = \left\| (e_i - e_i') \int_{\alpha=0}^{1} \nabla_{e_i} F(e' + \alpha(e - e'))\, d\alpha \right\|_1 . \quad (5)$$

The integral was approximated with a finite number of interpolation steps. Path-based attributions can accumulate noisy gradients along the integration trajectory, motivating comparison with additional attribution variants (Kapishnikov et al., 2021).

Gradient×Input. For residue embedding e_i, Gradient×Input is:

$$s_i^{\text{GxI}} = \left\| e_i \odot \nabla_{e_i} F(x) \right\|_1 . \quad (6)$$

This provides a local first-order sensitivity estimate (Ancona et al., 2018).

SmoothGrad-IG. SmoothGrad-IG averages IG scores over noisy embedding perturbations:

$$s_i^{\text{SG-IG}} = \frac{1}{M} \sum_{m=1}^{M} s_i^{\text{IG}}(e + \epsilon^{(m)}), \quad \epsilon^{(m)} \sim \mathcal{N}(0, \sigma^2 I). \quad (7)$$

This tested whether epitope alignment improved when attribution scores were stabilized by noise averaging (Smilkov et al., 2017).

Occlusion. Occlusion masked one residue at a time and measured the decrease in predicted allergenicity:

$$s_i^{\text{occ}} = F(x) - F(x_{\setminus i}). \quad (8)$$

Attention weights. For models with learned attention pooling, the normalized pooling weights are used as residue-level scores. We treated these as model-internal signals rather than guaranteed explanations, since attention weights are not necessarily faithful causal explanations of model predictions.

Auxiliary MTL residue head. The auxiliary MTL residue head directly predicts a residue-level epitope probability for each position. It was reported as a diagnostic supervised signal indicating whether the injected epitope labels are learnable by an auxiliary head, not as a post-hoc explanation of the protein-level classification head.

Random baseline. The random baseline assigned random scores to residues within each protein, recomputed over repeated draws before aggregation. This controlled for protein-specific epitope density and metric behavior.

D. Additional Protein-Level Classification Metrics

Table 2. Full protein-level classification metrics on the held-out DeepAlgPro test set ($n = 1377$).

Model	AUROC	Prec.	Rec.	F1	MCC	Acc.
Frozen ESM-2	0.970	0.901	0.933	0.917	0.835	0.917
MTL ESM-2	0.962	0.886	0.936	0.910	0.821	0.910
DeepPlantAllergy	0.974	0.939	0.939	0.939	0.878	0.939

E. All-Signals' Residue-Level Faithfulness

The main paper reports a curated subset of residue-level signals to preserve readability. Here we report the full active signal inventory across all available model families. For Frozen ESM-2 and DeepPlantAllergy, this includes attention weights, Integrated Gradients, Gradient×Input, SmoothGrad-IG, occlusion, and random baselines. For MTL

ESM-2, the same classification-head signals are reported together with the auxiliary residue head. Exploratory supplementary checkpoints are included only when their probe rows are available.

Results were consistent with Section 3; see Figure 6 for the full signal breakdown.

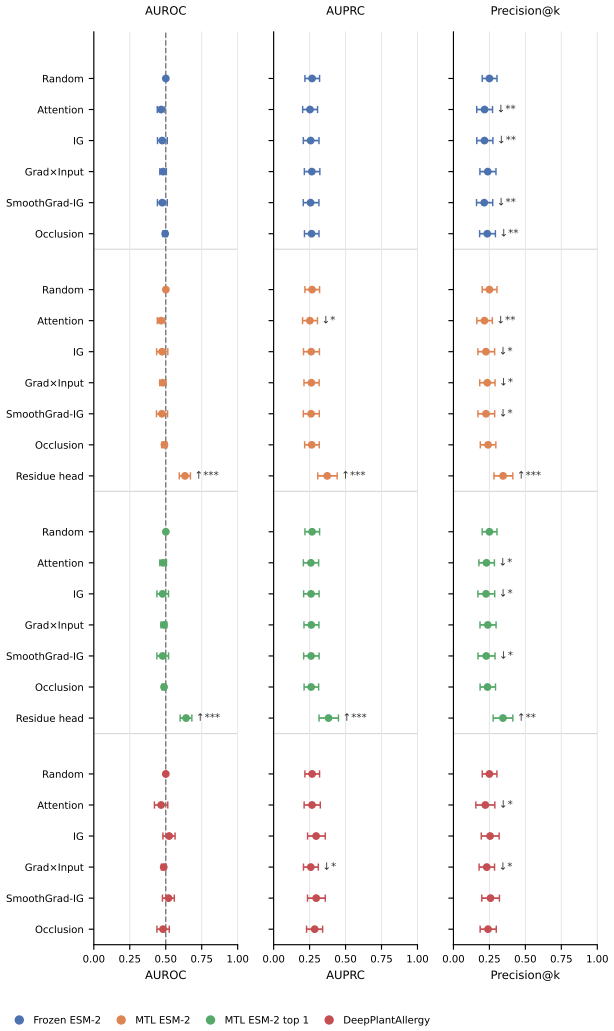

Figure 6. Supplementary all-signals faithfulness analysis. Rows are grouped by model family and show residue-level AUROC, AUPRC, and Precision@k for each available signal. Points show mean performance with bootstrap 95% confidence intervals. Significant paired Wilcoxon comparisons against the corresponding within-model random baseline are annotated after Benjamini–Hochberg correction; non-significant comparisons are omitted for readability.

Table 3. Main residue-level alignment summary on 61 held-out allergenic proteins. Values are mean per-protein metrics with bootstrap 95% CIs.

Signal	AUROC	95% CI	AUPRC	95% CI	P@k	95% CI
Random	0.501	[0.499, 0.502]	0.267	[0.218, 0.320]	0.250	[0.200, 0.303]
Frozen IG	0.476	[0.443, 0.511]	0.257	[0.206, 0.315]	0.216	[0.162, 0.275]
Frozen occlusion	0.496	[0.480, 0.511]	0.264	[0.214, 0.315]	0.235	[0.182, 0.292]
MTL auxiliary residue head	0.634	[0.594, 0.672]	0.372	[0.307, 0.442]	0.345	[0.282, 0.413]
DeepPlantAllergy IG	0.524	[0.480, 0.565]	0.295	[0.236, 0.359]	0.255	[0.194, 0.318]

F. Main Residue-Alignment Summary Table

G. Label-Scrambling Sanity Check

To verify that the residue-level evaluation is free of metric bias or label leakage, we permuted epitope labels within each protein while holding residue scores fixed. As shown in Figure 7, the MTL ESM-2 residue head suffered the largest collapse ($0.372 \rightarrow 0.271$), while post-hoc attribution methods remained within ± 0.02 of their original values. This larger drop is consistent with the residue head learning epitope-related information during MTL supervision, while scrambled AUPRC values for other methods clustered near the empirical random-ranking AUPRC baseline (~ 0.267), consistent with random behavior.

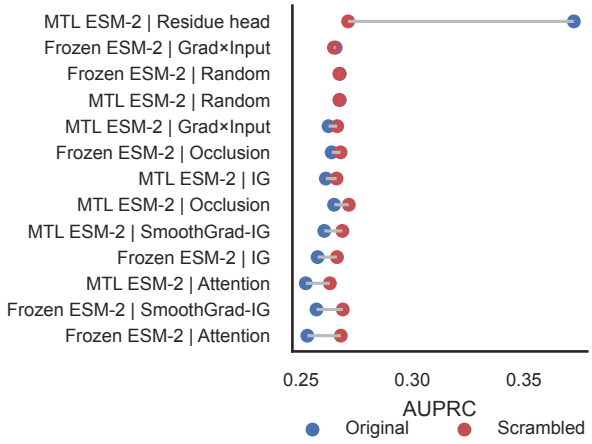

Figure 7. Label-scrambling sanity check. Epitope labels are permuted within each protein while residue scores are held fixed ($n = 61$). Rows are sorted by ΔAUPRC = original − scrambled.

H. Functional Masking Details

The masking sweep evaluated $k = \{5, 10, 20, 25, 30, 35, 40, 45, 50\}\%$ on high-confidence allergenicity predictions with baseline probability above 0.70. The selected threshold was $k = 40\%$, which maximized the fraction of validated proteins, with ties broken by larger mean Δp and then smaller k. At $k = 40\%$, 37 of 46 proteins exceeded the validation threshold $\Delta p > 0.05$. IG-vs-random masking yielded a paired

Wilcoxon $p = 1.83 \times 10^{-9}$.

This analysis was conditioned on proteins that the classifier predicted as allergens with high confidence. It tested whether IG identified residues causally important for confident positive model predictions, not whether IG masking has the same effect across all proteins.

As a sensitivity analysis, we repeated the same masking procedure without the confidence filter, using all 61 splitB allergenic proteins. At the same operating point $k = 40\%$, IG-guided masking remained substantially stronger than random masking of equal size (IG mean $\Delta p = 0.238$ vs. random -0.023; paired Wilcoxon $p = 3.7 \times 10^{-7}$), indicating that the effect was not driven by the confidence-based restriction.

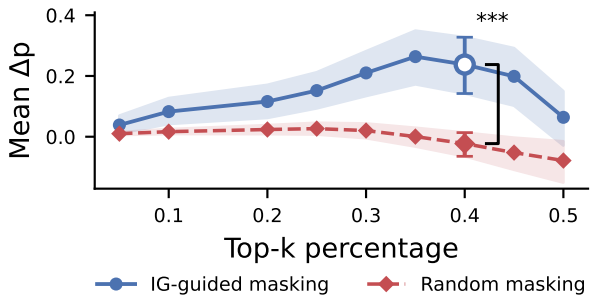

Figure 8. Unfiltered sensitivity analysis of IG-guided masking over all 61 splitB allergenic proteins. The confidence filter $F(x) > 0.70$ is removed; all other masking and random-baseline procedures match the main analysis. IG-guided masking remains stronger than random masking of equal size at the selected operating point $k = 40\%$.

Table 4. Residue classes used for mutagenesis summaries. Glycine and proline are listed individually because their effects on backbone geometry are not reducible to a shared physicochemical class. The remaining classes are coarse biochemical interpretation aids rather than mutually exclusive mechanistic claims.

Grouping	Class	Residues
Charge/polarity	Acidic (negatively charged)	D, E
	Basic (positively charged)	K, R, H
	Polar, uncharged	S, T, N, Q, Y, C
	Hydrophobic, nonpolar	A, V, L, I, M, F, W
	Glycine	G
	Proline	P
Hydrophobicity/aromaticity	Hydrophobic, aliphatic	A, V, L, I, M
	Aromatic	F, W, Y
	Polar or hydrogen-bonding	S, T, N, Q, C
	Charged	D, E, K, R, H
	Glycine	G
	Proline	P

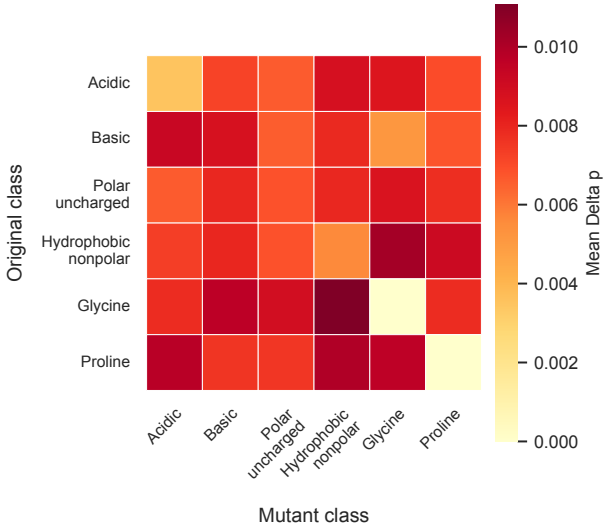

Figure 9. Charge/polarity transition heatmap for saturation mutagenesis. Each cell reports mean Δp for substitutions from an original residue class to a mutant residue class.

I. Saturation Mutagenesis Details

Saturation mutagenesis was run on the 37 proteins validated by IG masking at $k = 40\%$ (Section H). Each residue was replaced with all 19 alternative amino acids. Effects were summarized by original residue identity (Figure 4) and residue-class transitions (Figures 9 and 10).

I.1. Residue-Class Definitions

We aggregate substitutions using two coarse biochemical groupings. These categories were not used during training; they were used only to summarize whether model sensitivity follows interpretable biochemical axes.

I.2. Epitope vs. Non-Epitope Mutagenesis Sensitivity

To directly test whether the model assigns greater functional weight to epitope residues, we ran full-sequence saturation mutagenesis on all 61 splitB positive proteins and compared mean $|\Delta p|$ averaged over epitope versus non-epitope residues within each protein. Non-epitope residues showed marginally but significantly larger sensitivity than epitope residues (paired Wilcoxon $p = 0.021$, two-sided), with no significant difference in signed effect or fraction of reducing substitutions ($p > 0.87$ for both). This rules out the possibility that the model assigns greater functional weight to epitope positions and provides a third independent line of evidence for global rather than epitope-localized sensitivity.

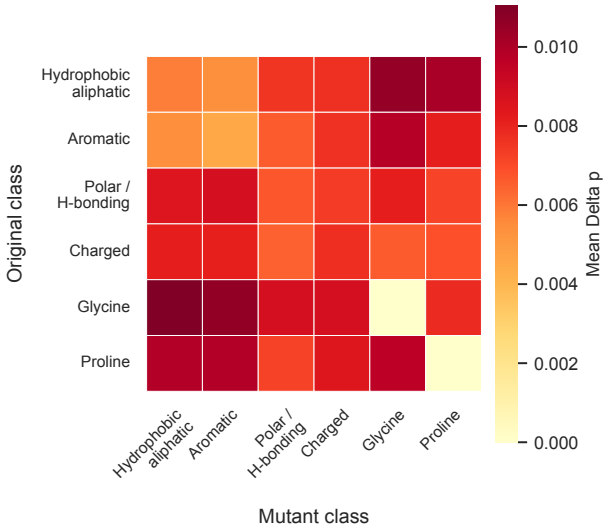

Figure 10. Hydrophobicity/aromaticity transition heatmap for saturation mutagenesis. Each cell reports mean Δp for substitutions from an original residue class to a mutant residue class.

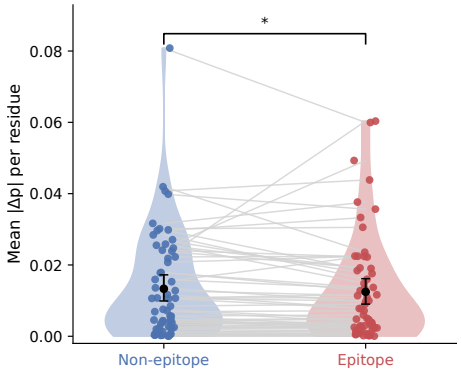

Figure 11. Paired comparison of mean $|\Delta p|$ per residue between non-epitope and epitope regions across all 61 splitB allergenic proteins. Each point is one protein; lines connect paired values. Black markers show bootstrap mean $\pm$ 95% CI. Non-epitope residues show marginally larger sensitivity than epitope residues (Wilcoxon $p = 0.021$, two-sided; mean paired difference $= -0.0009$, 95% CI $[-0.0027, 0.0012]$).