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]
[1]Swiss Institute of Allergy and Asthma Research, Davos, Switzerland
[2]ETH Zurich, Zurich, Switzerland
[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland
† Corresponding author

Background

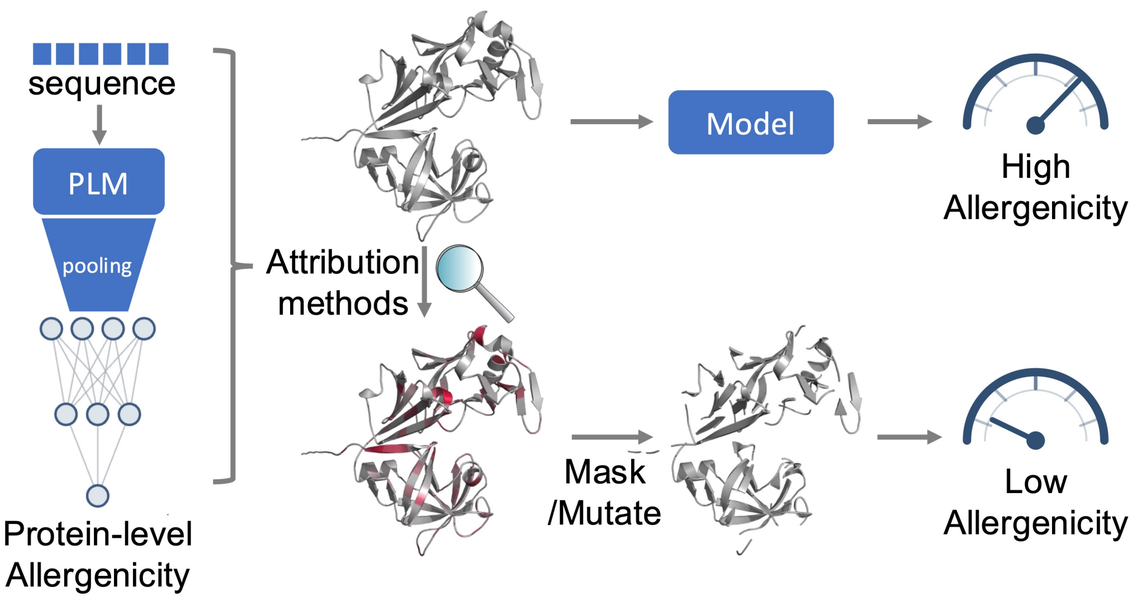
- **Protein allergenicity:** ability of a protein to trigger an allergic response.
- **Allergen epitope:** protein region recognized by the immune system.
- **Allergenicity predictors** support safety screening and hypoallergenic protein design.
- **Protein language models** (PLMs) have substantially improved allergenicity prediction.
- **Explainable AI** attribution maps are often interpreted as epitope explanations on qualitative examples by previous work.

Research Question:
Quantitatively, do PLM-based allergenicity models produce attribution maps that are both model-faithful and immunologically faithful?

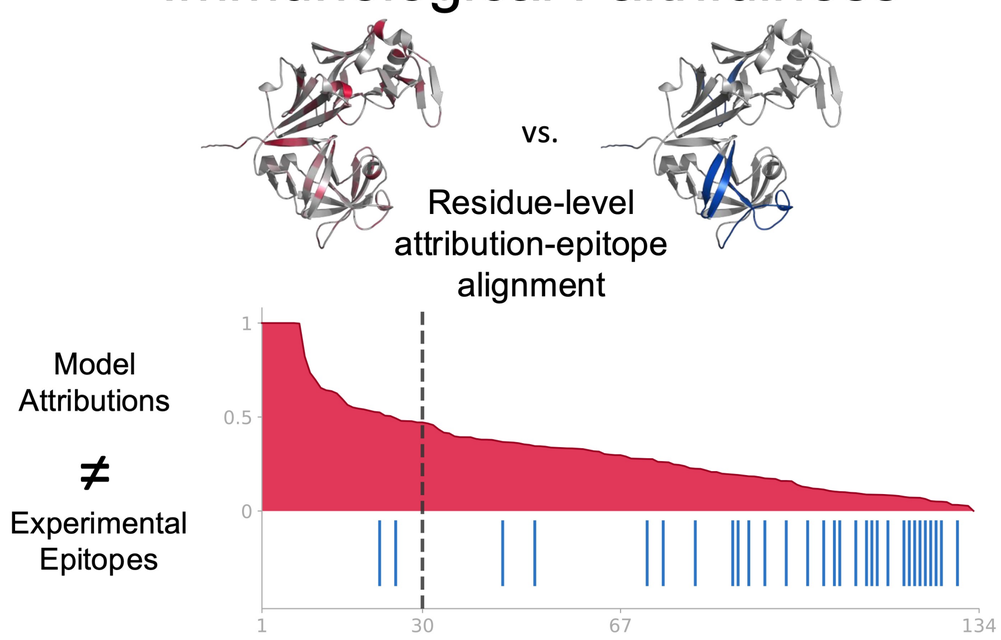
Model faithfulness: Do attributions explain the model's allergenicity classification decision?

Immunological faithfulness: Can attributions be interpreted as biological explanations for immune recognition?

Model Faithfulness



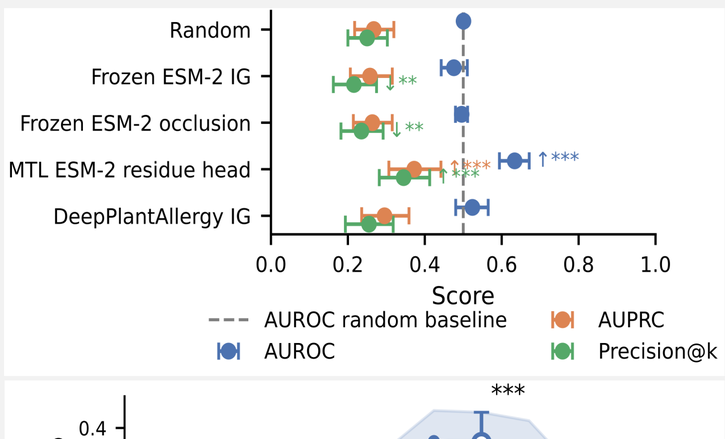
Immunological Faithfulness



Results

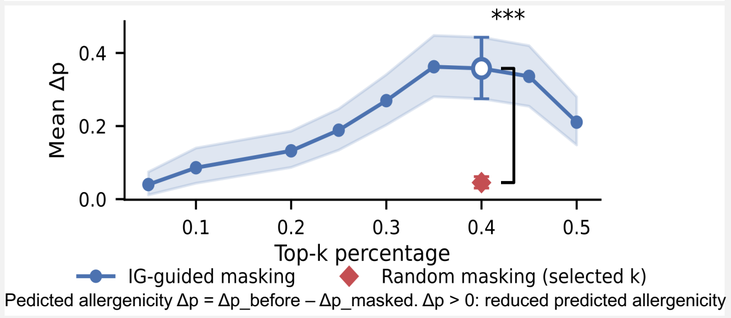
Strong classification Performance ≠ trustworthy epitope explanation

- No method produced epitope alignment significantly higher than random despite high AUROC, confirmed by labelling scrambling.
- Several alignments significantly below random.
- Representational misalignment: Epitope supervision via multi-task learning did not improve attribution alignment or classification, even with an unfrozen ESM-2 layer.



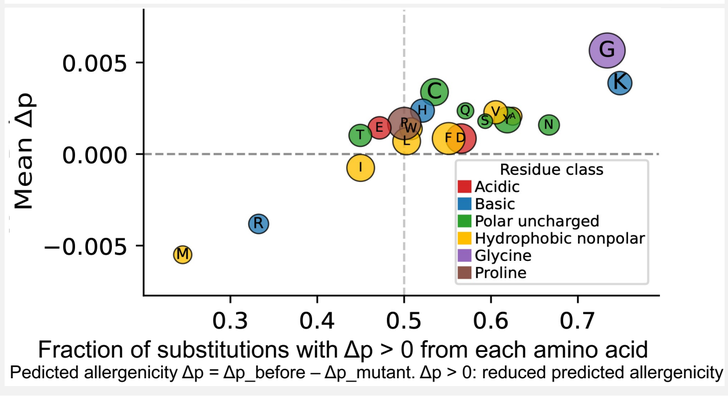
Integrated Gradient (IG) is model-faithful, but immunologically misaligned.

- Masking IG-important residues with ESM-2 <mask> token caused a significantly larger drop in predicted allergenicity compared to random masking.
- Highly attributed residues are causal for model's decision, but not epitope signals.



Mutagenesis reveals global physicochemical rather than epitope-specific sensitivity

- Mutagenesis effects significantly larger in non-epitope regions than epitope regions.
- Larger effects in charge-altering mutations, Smaller effects in within-aromatic, within-hydrophobicity mutations.
- Highest glycine sensitivity linked to small size & special properties in protein folding.
- Cysteine sensitivity aligned with disulfide-bond stabilization disruption.



Conclusion

- Strong dissociation between predictive performance, model faithfulness, and immunological faithfulness.
- Attributions should not be interpreted as epitope explanations for safety screening or hypoallergen design without quantitative validation.

Limitations: Limited IEDB-based benchmark size; no conformational epitope available; in silico validation
Future work: Incorporation of structural, immunological constraints toward trustworthy allergenicity prediction

Scan for Full Project Materials