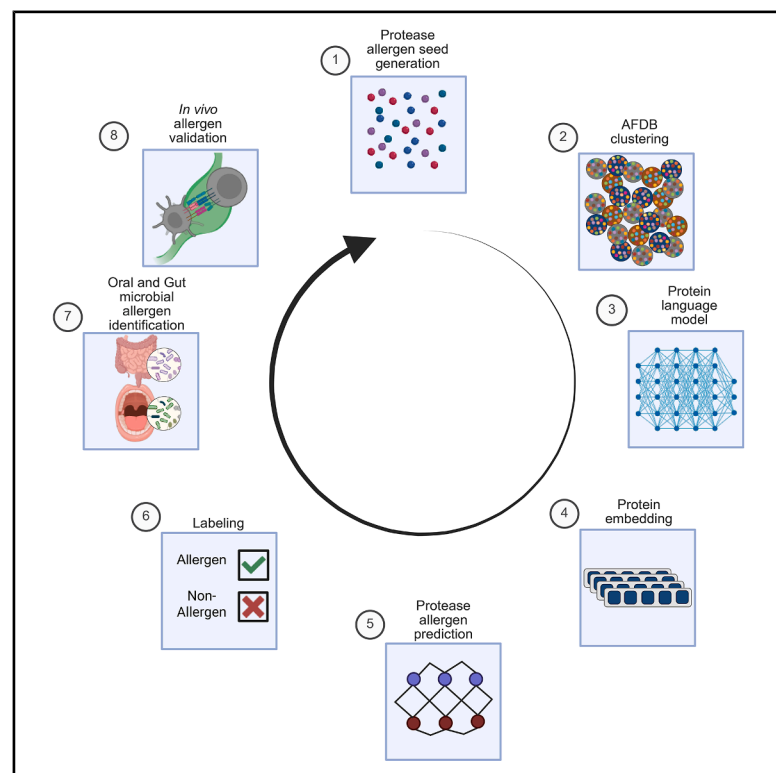


Identifying microbial protease allergens through protein language model-guided homology

Graphical abstract



Authors

Kumar Thurimella, Elena Wu, Chenhao Li, ..., Caroline L. Sokol, Ramnik J. Xavier, Sergio Bacallado

Correspondence

damian@broadinstitute.org (D.R.P.), clsokol@mgh.harvard.edu (C.L.S.), xavier@molbio.mgh.harvard.edu (R.J.X.), sb2116@cam.ac.uk (S.B.)

In brief

Thurimella et al. introduce a deep learning framework using protein language models to predict allergenic proteases in the microbiome. The model identified hundreds of candidate allergenic proteases in oral and gut microbiomes, and two predicted microbial proteases were experimentally validated as Th2-skewing allergens.

Highlights

- Deep learning framework predicts protease allergens with high accuracy
- Hundreds of candidate allergenic proteases predicted in oral and gut microbiomes
- Model generalizes beyond training on serine proteases to cysteine protease allergens
- Two microbial proteases experimentally validated as *bona fide* Th2-skewing allergens



Article

Identifying microbial protease allergens through protein language model-guided homology

Kumar Thurimella,^{1,2,3,4,8} Elena Wu,^{5,8} Chenhao Li,^{1,2} Daniel B. Graham,^{1,2} Róisín M. Owens,³ Damian R. Plichta,^{1,2,9,*} Caroline L. Sokol,^{5,7,9,*} Ramnik J. Xavier,^{1,2,7,9,*} and Sergio Bacallado^{6,9,10,*}

¹Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

²Center for Computational and Integrative Biology and Department of Molecular Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

³Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge CB3 0AS, UK

⁴School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

⁵Center for Immunology and Inflammatory Diseases, Division of Rheumatology, Allergy and Immunology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

⁶Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge CB3 0WB, UK

⁷Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA

⁸These authors contributed equally

⁹Senior author

¹⁰Lead contact

*Correspondence: damian@broadinstitute.org (D.R.P.), clsokol@mgh.harvard.edu (C.L.S.), xavier@molbio.mgh.harvard.edu (R.J.X.), sb2116@cam.ac.uk (S.B.)

<https://doi.org/10.1016/j.cels.2025.101510>

SUMMARY

Emerging research links the gut, skin, and oral microbiomes to allergies, with serine proteases (SPs) identified as potential allergens. This study leverages deep learning and pre-trained protein language models (pLMs) to uncover allergenic SPs in metagenomic data. First, we develop a model to identify the catalytic serine residue in serine hydrolases, demonstrating how pLMs capture structural information. Next, we create a deep learning framework to detect candidate SP allergens across gene catalogs, using the conserved catalytic triad to identify homologs in gut and oral sites despite low sequence identity. Our model predicts a putative SP allergen resembling V8 protease, a known trigger for protease-activated receptor 1. It also identifies a cysteine protease similar to Der f 1 from dust mites. Immunization with these proteases induced allergic responses, validating their allergenic potential experimentally. This approach uncovers candidate allergens beyond traditional methods, offering new targets for allergy research. A record of this paper's transparent peer review process is included in the supplemental information.

INTRODUCTION

The worldwide rise in allergic disease prevalence necessitates a deeper understanding of underlying mechanisms.^{1,2} Recent evidence suggests that the enzymatic activities of environmental allergens, including proteolytic activity, play a crucial role in initiating allergic responses at various tissue sites.^{3–5} Many immunodominant allergens can be categorized as serine or cysteine protease enzymes that follow a proteolytic mechanism defined by a catalytic triad in the active site.^{6–8} Despite disparate plant, animal, and fungal sources, these proteases share characteristics that suggest they have convergently evolved similar functions.⁹ This makes cysteine or serine protease activity a distinctive feature that could be exploited for allergen classification or discovery.¹⁰ The microbiome is one untapped source of possible allergenic proteases. While the gut microbiome and various microbial mechanisms have been implicated in the risk and devel-

opment of allergic diseases,^{11,12} further studies are needed to identify specific proteases involved in these pathways, including investigation of proteases, as these have previously been associated with allergic pathology.⁵

Aberrant protease reactions are a contributing factor in various inflammatory diseases.¹³ Many immunodominant allergens, such as the house dust mite allergen Der f 1 and the papaya allergen papain, are cysteine proteases, whose enzymatic functions are required for allergenic activity.^{5,14} Similarly, the proteolytic activity of serine proteases (SPs) can be linked to processes that initiate inflammation.^{15,16} SPs fall within the broader category of serine hydrolases (SHs), which collectively constitute ~1% of the human genome and are present across the domains of life.¹⁷ Some microbes, such as *Staphylococcus aureus*, encode SPs that trigger an allergy-like response, leading to the production of SP-specific immunoglobulin E (IgE) and memory T helper 2 (Th2) cells.¹⁸ One mechanism of action involves SPs



proteolytically cleaving Mas-related G protein-coupled receptors (Mrgprs) and protease-activated receptors (PAR1/PAR2). This, in addition to other possible pathways, can activate sensory nociceptors, leading to itch, Th2-skewing dendritic cell (DC) migration, Th2 development, and Th2 recruitment.^{19–27} Microbial SPs have also been shown to act as itch triggers by activating PAR1, but whether they induce allergic disease is unknown.²⁷ Other PARs are also activated by microbial proteases. Notably, PAR2 is targeted in lung tissue by *Streptococcus pneumoniae*²⁸ and in Crohn's disease by several microbes.²⁹ PAR2 has been implicated in pain pathways in patients with irritable bowel syndrome³⁰ and may contribute to nerve ending exposure through the dysregulation of tight junctions in the colonic epithelia.³¹ The diverse mechanisms by which SPs contribute to inflammatory processes, particularly in allergy-like responses, warrant further investigation.

Several allergen prediction tools have been developed, each focusing on different methodologies to uncover sequences and some potential structural homologies with known allergens. AllergenFP³² utilizes E-descriptors to capture structural properties of all allergens, while AllerHunter³³ employs the Smith-Waterman algorithm and BLOSUM62 matrices to train a machine learning model. Other tools, like AlgPred 2.0,³⁴ rely on BLAST³⁵-based similarity searches combined with multiple sequence alignments. By contrast, AllerCatPro 2.0³⁶ incorporates additional 3D structure information for epitope mapping but on a limited set of structures.

Recent advancements in deep learning methods demonstrated an impressive ability to scale and capture evolutionary and structural features of proteins.^{37–42} Advanced structural homology search tools like Foldseek⁴³ and DALI⁴⁴ further aid in incorporating new information for functional annotation. For the prediction of allergens, ProAll-D⁴⁵ uses a long short-term memory model. Other deep learning allergen prediction tools like ALLERDET⁴⁶ and DeepAlgPro⁴⁷ use custom-built deep learning models to predict allergens. Protein language models (pLMs) have been at the forefront of these advancements, demonstrating their ability to discern recurrent patterns across vast datasets of protein sequences.^{40,48–51} The success of pLMs in downstream transfer learning tasks across diverse biological topics, such as T cell receptor (TCR) epitope binding,^{52,53} protein-protein binding,^{54,55} remote homology,⁴⁸ drug and protein interactivity,⁵⁶ protein solubility,⁵⁷ and secondary structure prediction,^{40,58} highlights their potential for the identification of allergens. Indeed, AllerTrans⁵⁹ leverages pLMs for broad allergen prediction.

Despite the advancements in prediction tools, structural similarity does not necessarily equate to functional similarity, and none of the aforementioned tools specifically investigate SP allergens, in which proteolytic function, rather than structure, imparts allergenicity. We hypothesize that the per-residue granularity of pLMs^{60,61} makes them well-suited to capture the highly conserved residues of an SP catalytic triad and that the transfer learning framework can be used to identify SPs with allergenic potential, using pLM embedding representations of annotated SPs as input data in the classifier.

Traditional classification relies on clearly defined negative and positive classes, which are often difficult to establish in real-world scenarios.⁶² Creating a set of “negative” or non-

allergenic SPs can be challenging, as the set of known SPs not previously annotated as allergens may nevertheless include allergens, given the lack of systematic experimental screens of SPs for allergenic activity in the literature. Therefore, to identify SP allergens amidst other SPs and SHs, we employed a variation of positive unlabeled (PU) learning⁶³ in which the set of unannotated SPs is weighted less heavily during training than known SP allergens, reflecting their uncertain allergenic status. PU learning has previously been used to address this same problem in the context of training machine learning models to predict disease genes⁶² and has been effectively implemented using machine learning algorithms like support vector machines (SVMs).⁶⁴ In this context, SVMs optimize for the positive class while adjusting slack variables to handle outliers or negative examples.⁶² The successful application of deep learning to PU learning in fault detection⁶⁵ further motivates the application of this methodology to SP allergen classification.

To detect potential SP allergens, we employed two complementary strategies. First, we used pLM embeddings to train our models. Second, we explored the strengths of different modeling frameworks by training both a traditional SVM classifier and a deep learning model. We evaluated the generalizability of our models by applying them to microbial gene catalogs from diverse taxonomic groups, including the Unified Human Gastrointestinal Genome (UHGG)^{66,67} and Human Oral⁶⁸ microbiome datasets from MGnify. Using the diverse Foldseek clustered AlphaFold Protein Structure Database (AFDB Clusters),⁶⁹ we incorporated detailed knowledge of SP properties into our model. Our model predicted several thousand putative SP and even cysteine protease allergens in the gut and oral microbiomes. Among these, we pursued two specific candidates for experimental validation that showed structural similarity to the known allergen proteases, V8 protease and Der f 1. Using murine immunization models, we showed that these predicted SP and cysteine proteases are bona fide allergens based on their ability to activate an allergic immune response, characterized by Th2-skewing DC migration to draining lymph nodes and Th2 differentiation. Together, these approaches provide evidence that our classifier is able to predict SP-like allergens in large metagenomics datasets.

RESULTS

Catalytic serine recognition using pLMs

The highly conserved SH enzyme family is defined by a unique structural feature consisting of three amino acid residues that comprise a catalytic triad. Common motifs around the catalytic serine include GxSxG in lipases,⁷⁰ SXXK in β -lactamases,⁷¹ and others. We aimed to assess how well pLM embeddings capture catalytic SP activity with allergenic potential. Using a logistic regression (LR) classifier trained on residue embeddings from pLMs, we investigated whether these embeddings could accurately identify catalytic serines within SH sequences. Three state-of-the-art pLMs—ESM-2,⁴⁰ ProtT5,⁴⁹ and ProSE⁵¹—were used to generate protein embeddings. We manually curated a set of 232 SH sequences and used a random split of catalytic and non-catalytic serines to train and evaluate the LR classifier. Given that the embedding dimensions vary

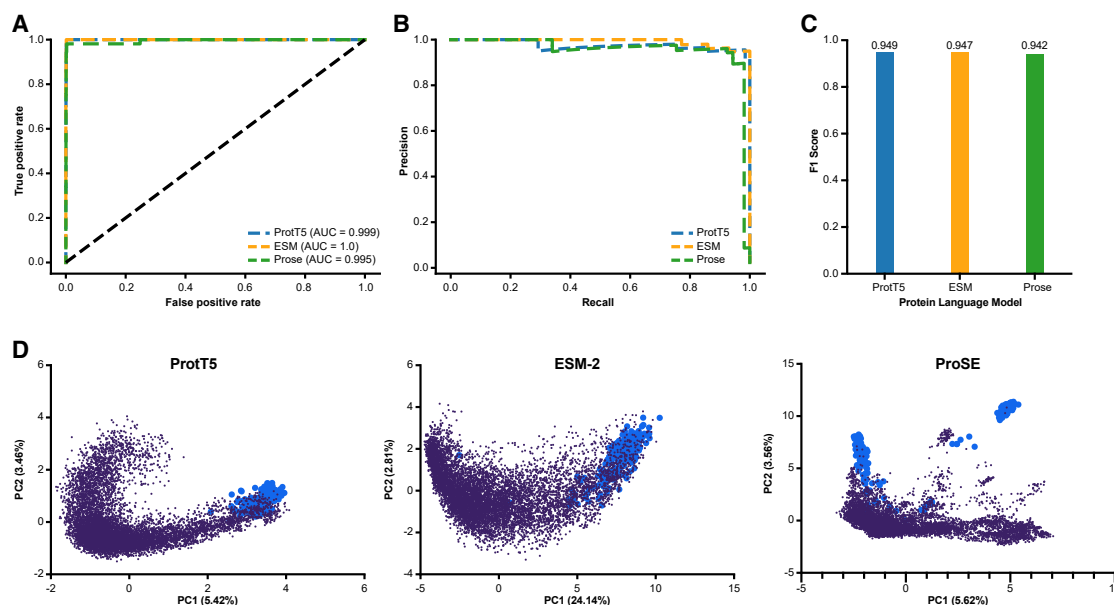


Figure 1. Different pLMs are able to identify serine residues in the catalytic site

(A and B) To benchmark the ability of LR models trained on three different pLM embeddings (ProtT5, ESM-2, and ProSE) to classify catalytic versus non-catalytic serines, we plotted (A) ROC curves and (B) precision/recall curves computed on a held-out test set.

(C) F1 scores (harmonic means of precision and recall) for each classifier with the standard threshold of 0 for the log-odds ratio.

(D) Using PCA, each pLM serine embedding was reduced from the original dimensionality of the pLM (ProtT5: 1,024, ESM-2: 1,280, and ProSE: 6,165) to two dimensions (principal components [PCs] 1 and 2). Blue points represent catalytic serines, and purple points represent non-catalytic serines.

(ESM-2: 1,280, ProtT5: 1,024, and ProSE: 6,165), we trained separate LR models for each pLM.

To benchmark the ability of each pLM to distinguish catalytic serines from non-catalytic serines in SH sequences, we used three metrics to compare the LR models trained on the pLM embeddings on a held-out test set (25%). First, we evaluated the area under the curve (AUC) of a receiver operating characteristic (ROC) plot (AUROC), in which each model's true positive rate for identifying catalytic serines was plotted against its false positive rate. The performance of all three models was greater than 0.99 (Figure 1A). Second, we computed precision (the accuracy of each model's positive predictions) and recall (the extent to which each model could capture all positive cases in the dataset) at different thresholds for each LR model. Of the three models, ESM-2 maintained maximum precision up to the highest recall value, beyond which its precision fell and continued to fall as recall increased further. Meanwhile, the precision of ProtT5 initially decreased at a lower recall value but remained relatively constant as recall increased further (Figure 1B). Third, to compute a quantitative value, we used the F1 score, which is the harmonic mean of the precision and recall values. All three pLMs had an F1 score above 0.94 (Figure 1C), ranging from 0.942 (ProSE) to 0.949 (ProtT5).

The LR models trained on the ProSE, ProtT5, and ESM-2 pLM embeddings performed similarly on each of the benchmarking tasks for catalytic/non-catalytic classification. We next used principal-component analysis (PCA) to reduce the dimensionality of the embeddings of each of the catalytic and non-catalytic serines (Figure 1D). Catalytic serines clustered in each case. In the ProSE pLM, two distinct clusters appeared to catalog all

the catalytic serines. Not only were the two classes linearly separable in the embedding space, as shown by the accuracy of LR models, but they were also visually separable when the embeddings were projected onto the top two principal components.

We chose to move forward with the ProtT5 embedding, which had the smallest embedding size (1,024), the highest F1 score, and comparable AUC and precision/recall to ESM-2. To examine whether specific features in the ProtT5 embedding could explain the LR model decision boundary, we analyzed the LR model using SHAP (Shapley additive explanations) values^{72,73} and found that no given embedding feature had an outsized impact on the model's predictions (Figure S1).

Developing classification models for SP-like allergens

We leveraged the Allergome database to identify 33 initial allergenic SPs (ASPs) from diverse taxonomies for use in our training data. We designated two negative classes for non-allergens: 116 non-allergenic serine proteases (NASPs) and 126 non-protease SHs (NPSHs). These initial proteases served as our foundational seeds—representative proteins used to collect additional training data. Collectively, we refer to all the proteins in the ASP, NASP, and NPSH sets as the SH class (STAR Methods). We utilized AFDB Clusters to supplement each initial seed in our training data with structurally similar proteins (Figure 2A). For each protein that mapped to a cluster, every cluster member was downloaded and used to augment the training dataset. This resulted in 9,004 members labeled as “allergens” and 80,841 members labeled as “non-allergens.” We included seeds from the Pfam protein families database^{74,75} (Pfam seeds, or PS) as additional non-allergen training examples. Since PS represents

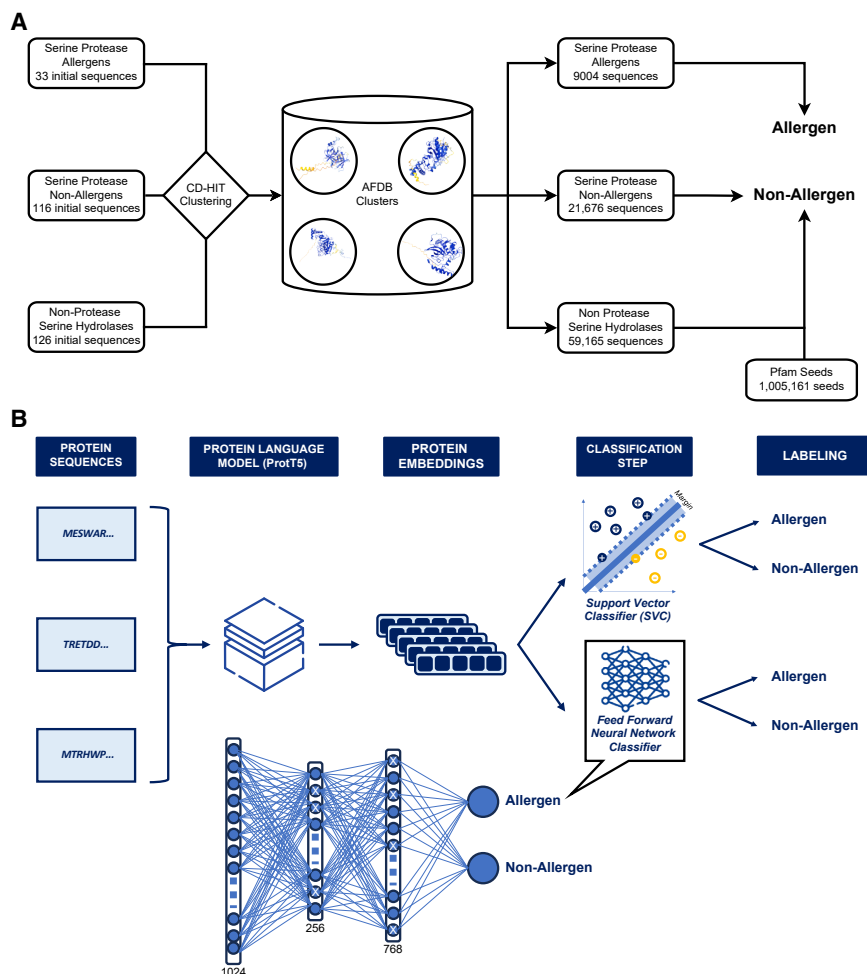


Figure 2. Curation and development of machine learning models to detect SP allergens

Using the pooled (averaged across each residue) embeddings from ProtT5, several classes of SHs and other proteins were established.

(A) A set of initial seeds for classes labeled “allergenic serine proteases,” “non-allergenic serine proteases,” and “non-protease serine hydrolases” was collected. After clustering, each initial seed was augmented with more proteins based on structural clustering defined by AlphaFold Clusters.⁶⁹ The allergenic SP class was then labeled as “allergen,” and the remaining classes and additional Pfam seeds were aggregated as “non-allergen.”

(B) ProtT5 was run on every protein in each labeled class from the structural augmentation to generate embeddings. Two models were trained in parallel on a binary classification task to label proteins as allergens or non-allergens. One model was developed as an SVM, while the other followed a deep learning paradigm using an MLP neural network architecture. The neural network had two hidden layers of size 256 and 768 that ultimately mapped to the allergen or non-allergen classes. The white X’s in the neural network indicate dropped nodes during the dropout regularization.

the entries in the NPSH class. The highest AUROC achieved on the test set was 98.5%.

To find the optimal MLP model, we employed RayTune⁷⁶ to optimize layer size, batch size, and class weights by randomly sampling hyperparameter values. We

75% of UniProt functional annotations, this exposes the classifier to a very large distribution of proteins.³⁸

We developed two binary classification models to detect allergens: an SVM, representing a classical machine learning approach, and a multilayer perceptron (MLP), a deep learning feedforward neural network (Figure 2B). Our decision to employ binary classification models was driven by the complex nature of allergen detection in the embedding space. The significant overlap between allergen and non-allergen embeddings (Figure S2A) and cosine similarity between embeddings (Figure S2B) necessitated further training beyond zero-shot classification methods. We used only the SH subclasses to train the SVM. During MLP training, we tuned the influence of different embedding subclasses (ASP, NASP, NPSH, and PS) to improve allergen detection. This weighting strategy helped the MLP prioritize learning allergen characteristics without overfitting on negative examples.

To find the optimal SVM model, we performed a randomized grid search, accounting for the kernel coefficient γ (a regularization parameter) and class weights for the SH subclasses. We selected the best of the resulting 10 models based on its AUROC. To determine this, we randomly split all the proteins in the ASP and NASP classes into an 80% training set and a 20% test set. We then augmented the training set with all of

again split the ASP and NASP classes into 80/20 train/test sets, then further subdivided the original 20% test set, creating a validation set and a final test set that each comprised 10% of the ASP and NASP data. We then augmented the training set with all of the entries in the NPSH and PS classes. We trained each model for 20 epochs, calculating the AUROC of the validation set at 20% intervals within each epoch. Upon training completion, an AUROC evaluation was performed on the designated final test set, and the model that achieved the highest AUROC (95.6%) was chosen as the optimal model (Figures S2C and S2D). The MLP classifier consists of four layers: an input of 1,024 dimensions (ProtT5 embedding size), which is projected to 256 dimensions, then to 768 dimensions, and finally to 2 dimensions representing allergen and non-allergen (Figure 2B). Among the three approaches (MLP, SVM, and zero-shot classification models), the MLP demonstrated superior performance (Figures S2E and S2F), leading us to select it as our primary model for allergen discovery.

When we first applied our MLP, trained on SH families, to the oral (791,924 proteins) and gut (4,601,771 proteins) catalogs, it labeled ~10% of each catalog as allergens, likely an overestimate due to limited negative examples. To address this, we incorporated over one million Pfam seed proteins as additional non-allergen data, expanding coverage of diverse protein domains.

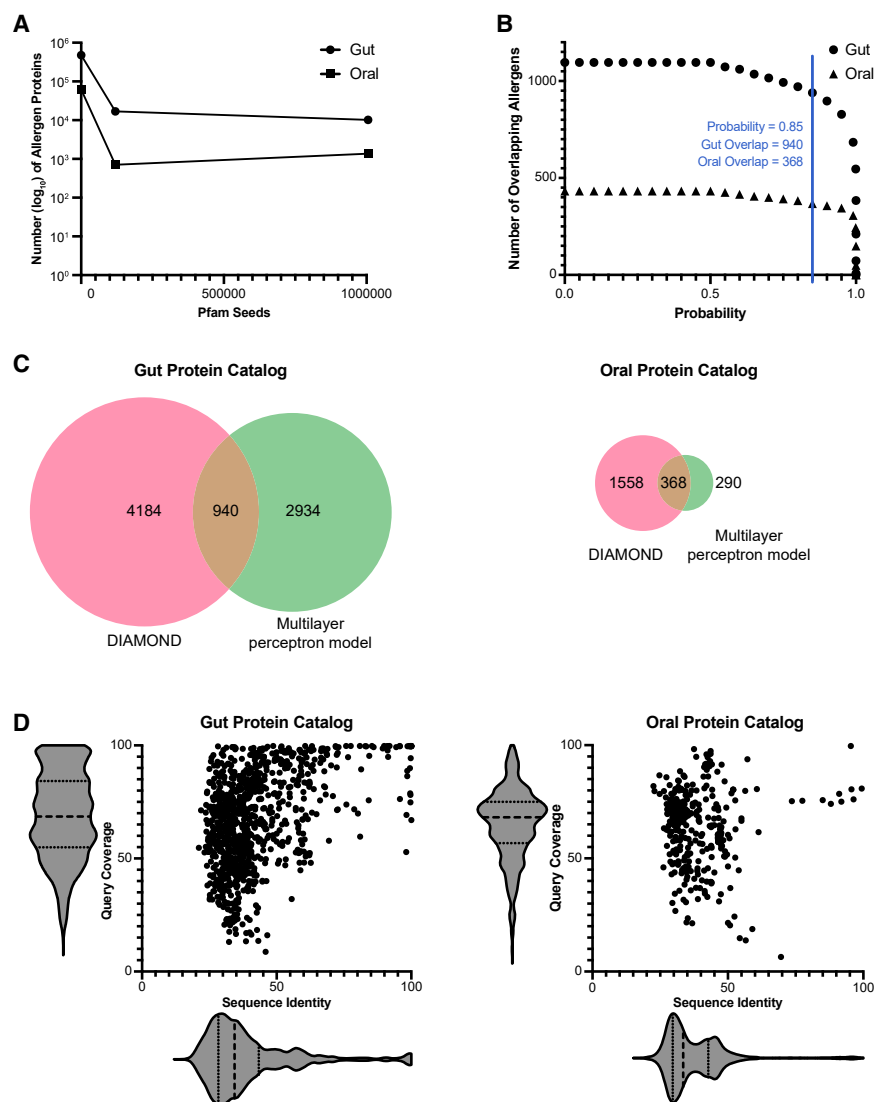


Figure 3. MLP model compared with sequence-based homology tool DIAMOND

(A) The number of labeled allergens, in log base 10 scale, in the MGnify gut and oral protein catalogs, as a function of the number of Pfam seeds in the negative (non-allergen) class.

(B) The total number of MLP model allergen predictions that overlap with DIAMOND as a function of the probability of the MLP model. The threshold of 0.85 is set to ensure high confidence for the predictions made by the model without losing many predictions.

(C) The number of allergen predictions made by DIAMOND (pink) when running sequence similarity searches on gut and oral protein catalogs over the allergen class sequences overlapped (brown) with the number of allergen predictions made by the MLP model (green).

(D) The sequence identity and query coverage of the best hit (measured by lowest *e* value) of the intersecting allergen predictions in the gut and oral protein catalog made via DIAMOND and the MLP model. The violin plots represent the density distribution of the hits along the two axes, where the dotted lines, from left to right (*x* axis) and bottom to top (*y* axis), represent the 25th, 50th, and 75th percentiles, respectively.

This step acts as a crucial regularizer. By forcing the model to distinguish proteases from a vast and diverse set of non-proteases, we prevent it from overfitting on spurious features in the small allergen set. It must instead learn the more subtle and generalizable signals of allergenicity. This adjustment reduced predicted allergens to ~0.1%–0.2% (Figure 3A) of each catalog (oral: 62,060 to 1,365; gut: 476,149 to 10,197). While some false positives may remain, these results highlight the importance of supplementing training data with diverse protein domains.

To assess the annotation quality of our classifiers, we compared our models against a traditional sequence homology method, DIAMOND⁷⁷ (STAR Methods). Our SVM model labeled 547 gut proteins and 270 oral proteins as allergens, and more than 98% of these were also identified by DIAMOND as sharing sequence similarity with known allergens. The MLP model's predictions are associated with probabilities, and as we increased the probability threshold for inclusion of the MLP classifier predictions, the overlap of the results between DIAMOND and the MLP model decreased. We set a probability threshold of 0.85 to focus on high-confidence allergen predictions while maintain-

ing overlap with DIAMOND (Figure 3B). In contrast to the SVM model, only approximately 24% of gut and 56% of oral allergens that the MLP model predicted with over 0.85 probability overlapped with DIAMOND predictions (Figure 3C). These results suggest that the MLP model can identify potential allergens distinct from those that can be detected by sequence similarity searches.

The hits that overlapped with DIAMOND were further broken down by sequence identity and query coverage between the DIAMOND queries and the reference set of allergens. Although query coverage was high, the 50th percentile of sequence identity of the overlapping hits was only 35% for the oral catalog and 37.5% for the gut catalog (Figure 3D), and overall was <47% for 75% of the hits. The MLP model's ability to train over complex non-linear relationships between embedding and function may explain its ability to recover alignments at a lower sequence identity.

Discovery of putative allergens in oral and gut protein catalogs

To determine whether the MLP model could uncover novel ASPs in the gut and oral microbial protein catalogs, we examined high-confidence MLP predictions corresponding to proteins with no similarity to previously characterized allergenic enzymes. First, of the 10,197 gut and 1,365 oral proteins that our MLP classifier had labeled as allergens, we took the 3,874 gut and 658 oral proteins that were labeled with a confidence of 0.85 or higher and folded them using ESMfold.⁴⁰ Next, using Foldseek,⁴³ we ran

the folded proteins in a structural alignment search against predicted structures in Swiss-Prot.⁷⁸ We annotated the Foldseek hits by performing a Pfam and InterPro domain scan using InterProScan⁷⁹ (STAR Methods).

Although the distribution of the Pfam annotations from this initial prioritization varied between the two protein sets, the majority of hits (37.05% in gut, 50.28% in oral) came from the subtilase family (Figure 4A). A subset of the remaining hits came from families of other protease-type enzymes, including the trypsin, peptidase-C1, and papain families, with papain being a major model allergen and its highly related counterparts, Der p 1 and Der f 1, recognized as common environmental allergens. Since our initial training seeds consisted exclusively of SPs, the labeling of cysteine proteases demonstrates our model's ability to generalize beyond its training data. We observed the same breakdown of enzyme families in the annotations provided by both Clusters of Orthologous Genes (COG) and InterPro (Figure S3). We further observed a diverse distribution of taxonomy for both sets of proteins. While most hits belonged to the Firmicutes and Actinobacteria phyla, the gut protein catalog uniquely contained hits from species within the Desulfobacterota and Verrucomicrobiota phyla (Figure 4B).

Using a priority scoring metric we developed that aggregated query coverage, target coverage, alignment template modeling (TM) score, and sequence identity from Foldseek with the *e* value of any hit also found in the InterProScan results, we filtered (STAR Methods) our set of 3,874 gut and 658 oral proteins down to 412 gut (a 9-fold reduction) and 147 oral proteins (a 4-fold reduction) (Figures 4C and 4D). Observing a positive correlation between our priority score and Foldseek's *e* value, we ranked our filtered set of proteins by $-\log(e \text{ value})$ from Foldseek. Thirteen oral allergens in the filtered set were predicted to be encoded by *Prevotella* species, including three of the top 15 proteases in the ranked list (Table S1), which aligns with *Prevotella*'s known involvement in oral inflammation⁸⁰ and potential links to asthma.⁸¹ The gut gene catalog revealed 69 hits predicted to be encoded by *Lachnospiraceae*, including 6 of the top 25 microbial proteases (Table S2). Notably, *Lachnospiraceae* members have been implicated in early-life egg and milk allergies,^{82,83} hinting at a potential role of these gut proteases in allergic pathways.

To evaluate the prevalence of these highly ranked predicted allergens in the human gut microbiome, we cross-referenced these proteins with a gut metagenomics collection of healthy individuals from the Framingham Heart Study (FHS).⁸⁴ We detected 110 of the predicted gut allergens in the FHS gut microbiome gene catalog with high homology (sequence identity > 90%, query coverage > 93%). Of those, three proteins in different *Lachnospiraceae* species were highly prevalent in the FHS population: *Roseburia hominis* (77.2% prevalent), *Eubacterium eligens* (67% prevalent), and *Eubacterium rectale* (87% prevalent). All of these putative allergens contained a subtilase Pfam domain and showed high structural and sequence similarity to known subtilisin allergens (Figure S4). Both *R. hominis* and *E. eligens* have been associated with a general lowering of inflammation^{85,86}; however, *E. rectale* was found to be enriched in pediatric patients with food or respiratory allergies.⁸⁷

Of the 412 gut and 147 oral putative allergen proteins that passed our priority scoring metric, 211 gut and 124 oral proteins

were discovered only by the MLP classifier and not by DIAMOND.⁷⁷ We focused on proteases with low sequence identity (15%–35%) to their closest Foldseek hit while maintaining high structural similarity (>0.7 TM score). As a result, we identified 52 gut and 45 oral putative proteases. We chose to structurally search these putative allergens against known crystal structures in the Protein Data Bank (PDB) and predicted structures in Swiss-Prot. We pursued two specific hits: one to a V8 protease from the gut protein catalog and another to a Der f 1 cysteine protease in the oral protein catalog. We chose these hits for their low sequence identity (in the “twilight zone”⁸⁸ of 20%–35%) and high structural homology (>0.8 TM score). The MLP model's confidence in labeling these two sequences as allergenic was consistently high, even when we altered residues within, near, and far from the catalytic triad (Figure S5).

Our first unknown allergen protease demonstrated high structural homology (TM score 0.87) and low sequence identity (31.05%) to an experimentally characterized crystal structure of a V8 protease from *Staphylococcus aureus* (Figure 5A, left). Sequence and structural alignment of the unknown allergen and the *S. aureus* V8 protease revealed complete conservation of the catalytic triad and overall high sequence and structural identity in the catalytic domain (Figure 5B, left). The putative enzyme is predicted to be from *Staphylococcus schleiferi* in the gut catalog. Using the Joint Genome Institute Integrated Microbial Genomes (JGI IMG) database, we identified several transporter proteins among the regulatory elements in the genomic neighborhood of the putative enzyme (Figure 5C, left). Consistent with this, the enzyme is predicted to contain a signal peptide. A sequence similarity search of the putative SP against other microbial genomes yielded hits mostly in the Firmicutes phylum (Figure 5D, left), many of which were from other *Staphylococcus* species. The identification of this putative V8 protease within the gut catalog highlights the potential to discover many uncharacterized SPs that may share similar allergenic functions along the PAR1 axis.

The presence of cysteine proteases (CPs) among our predicted allergens is noteworthy given that our training data consisted solely of SPs, highlighting the MLP model's ability to generalize beyond its initial scope. To validate the plausibility of CP hits, we focused on a putative CP allergen from *Tannerella forsythia* with high similarity to the allergen Der f 1 from the American house dust mite *Dermatophagoides farinae*. We found that the putative enzyme had high structural homology (TM score 0.87) and low sequence identity (25.71%) to Der f 1 (Figure 5A, right) and that the catalytic domain was highly similar, with perfect sequence and structural alignment of the catalytic triad (Figure 5B, right). The genomic neighborhood of the putative CP included a radical S-adenosyl-L-methionine (SAM) enzyme as well as transporter proteins (Figure 5C, right). We found homologous genes to this putative CP in a wider range of species than we did for the SP homologs (Figure 5D, right). In a separate sequence similarity search, we found a sequence with 99% similarity to the putative CP in a *Dialister invisus* genome recovered from metagenomic sequencing of stool samples from the FHS, in which *D. invisus* is prevalent (>85%) despite typically being found in the oral cavity.⁹¹ The presence of a Der f 1-like CP in the gut, originating from an oral microbe also known to colonize the gut,⁹² raises the possibility that allergic responses to this putative allergen could involve both the oral and gut environments.

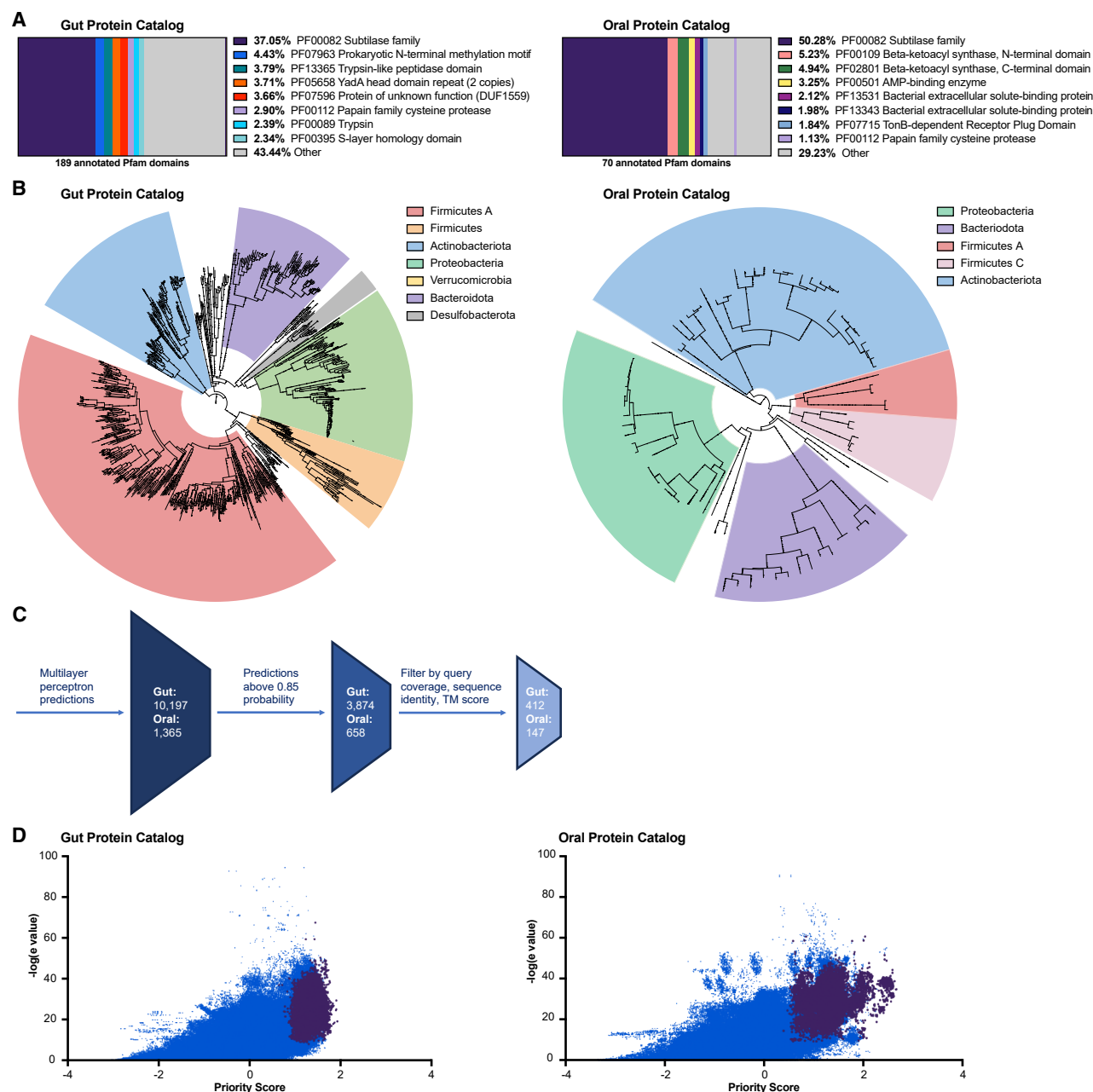


Figure 4. Prioritizing the allergen findings in gut and oral microbial protein catalogs

(A) The Pfam protein families database⁷⁵ distribution of the initial prioritized gut (left) and oral (right) hits. Only domains that were annotated in Pfam are represented.

(B) The taxonomy distribution of the prioritized hits. The gut proteins (left) exhibit broader taxonomic diversity than the oral proteins (right).

(C) The first step of our prioritization scheme was to filter the initial set of proteins that the MLP model labeled as allergens to only those in which it had high confidence (>0.85 probability). We then used InterProScan⁷⁹ and Foldseek⁴³ to stratify our hits based on query coverage (above 0.9), sequence identity (below 0.35), and structural (TM) score (above 0.7).

(D) Scatterplot of the priority scores of the initial prioritized gut (left) and oral (right) hits versus the negative log of the e value of the hits in the structural search. The subset of hits that passed this second filtering step is highlighted in purple.

Experimental validation of allergens discovered

Cutaneous exposure to protease allergens leads to the initiation of an allergic immune response, characterized in the mouse skin by the migration of allergic-skewing CD301b+ dermal DCs to the draining lymph node, where they instruct naive CD4+ T cells to

differentiate into interleukin (IL)-4-expressing Th2 cells.^{5,14,93,94} In order to determine whether our predicted SP and CP allergens were bona fide allergens, we examined the immune response downstream of papain, a well-established CP allergen from papaya, to that of recombinantly derived *S. schleiferi* SP or

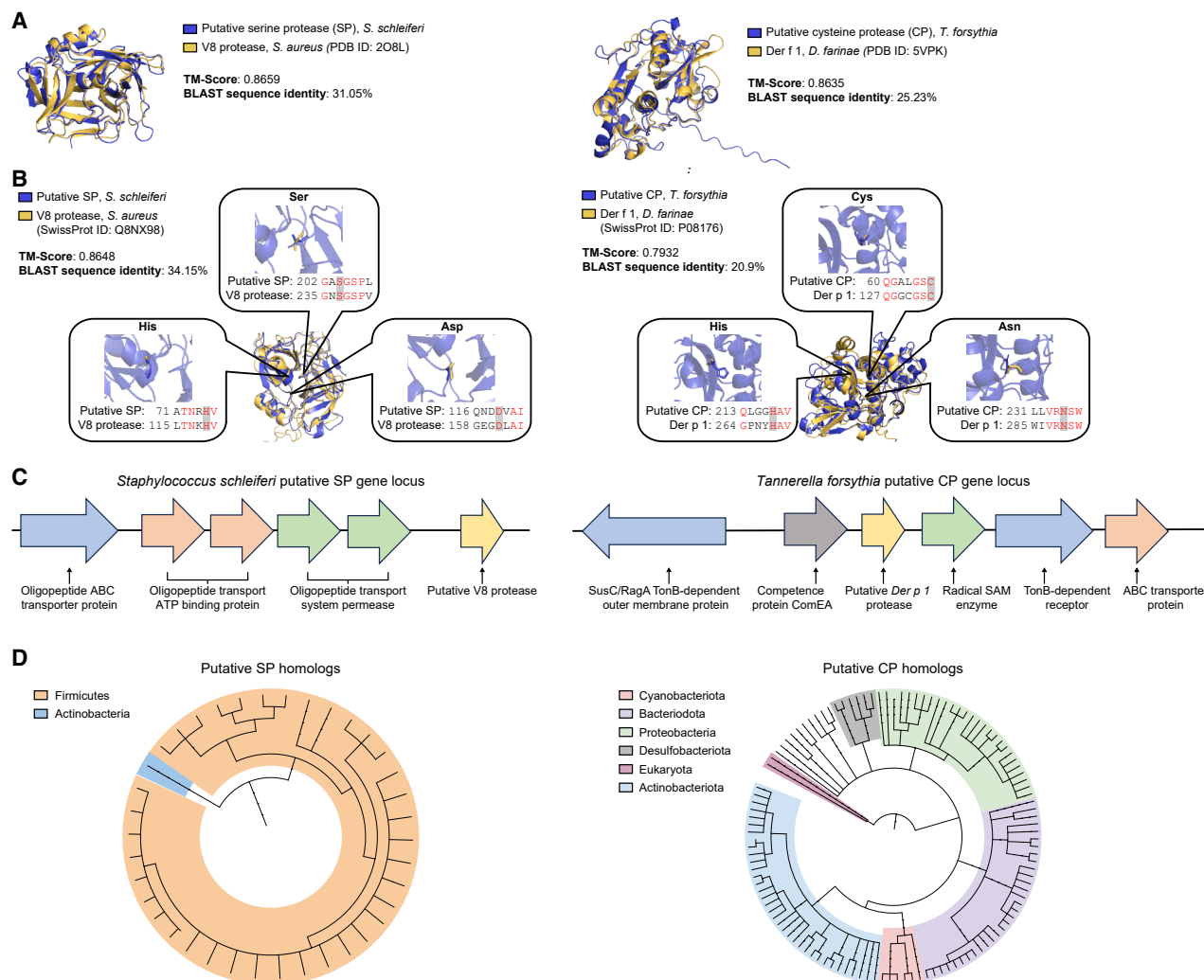


Figure 5. Beyond DIAMOND, the MLP model uncovers two putative undetected protease allergens in the gut and oral microbiome

(A) Using US-align,⁸⁹ two selected putative allergenic proteases were structurally aligned to experimentally characterized crystal structures: a SP to a known *S. aureus* V8 protease (left) and a CP to a known *Dermatophagoides farinae* (American house dust mite) Der f 1 allergen (right).

(B) A structural search was performed using Foldseek,⁴³ and the putative allergenic proteases were aligned to structures from Swiss-Prot based on TM score.⁸⁹ Structural alignments are shown with foreground elements and extraneous loops removed for better visibility of the catalytic triad region. The catalytic triad elements for a SP (left) and a CP (right) are each indicated in close-up views of the aligned structures alongside sequence alignments of the surrounding residues and are conserved between the respective structures despite low (<35%) overall sequence identity. In the sequence alignments, conserved residues are colored red, and the catalytic triad residue is highlighted in gray.

(C) The regulatory elements (genes) upstream and downstream of the gene encoding each putative protease.

(D) For each putative protein, a BLASTp search was performed against the Joint Genome Institute Integrated Microbial Genomes (JGI IMG) database.⁹⁰ Taxonomy is displayed for the top 500 homologous hits.

T. forsythia CP. As expected, papain led to the migration of activated Th2-skewing CD301b+ DCs, as measured by an increase in percentage and number of activated CD86+PDL2+CD301b+ DCs (Figures 6A–6C and S6A). This effect was largely limited to the Th2-skewing DC population, as papain had minimal effect on the total numbers of T helper 1 (Th1)-skewing XCR1+ DCs in the draining lymph node (Figures 6A and 6D). Similar to papain, both recombinantly derived *S. schleiferi* SP and *T. forsythia* CP induced the migration of activated Th2-skewing CD301b+ DCs, with minimal effect on Th1-skewing XCR1+ DCs (Figures 6A–6D). CD301b+ dermal DCs are required for the dif-

ferentiation of Th2 cells, the allergic-type CD4+ T cell characterized by the production of its hallmark cytokines IL-4 and IL-13.^{93,94} Consistent with their effects on CD301b+ DC migration, both recombinantly derived *S. schleiferi* SP and *T. forsythia* CP induced Th2 differentiation, marked by the accumulation of IL-4 but not interferon (IFN)- γ , producing CD4+ T cells in the draining lymph nodes 5 days after CP or SP immunization (Figures 6E, S6B, S6D, and S6F). This Th2-skewing effect was not shared by porcine trypsin (PT), endoprotease Lys-C from *Lysobacter enzymogenes*, or clostripain from *Clostridium histolyticum*, three proteases predicted to be non-allergenic from

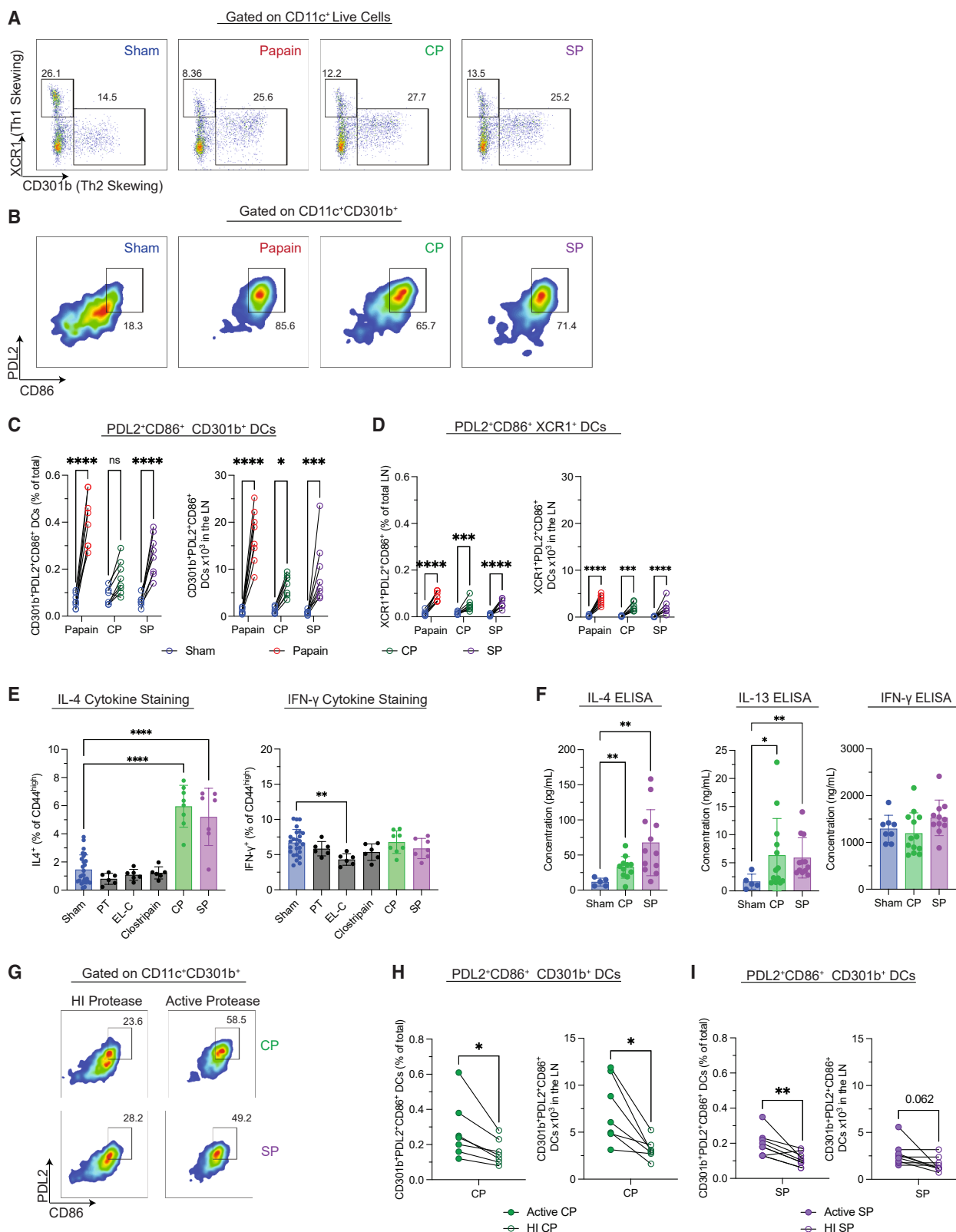


Figure 6. Deep learning predicted proteases activate the type 2 immune response

(A) Representative flow cytometry of live CD11c⁺ cells from the draining lymph node (dLN) of wild-type (WT) mice 20 h after intradermal (i.d.) immunization as indicated.

(legend continued on next page)

our model (Figures 6E, S6B, S6C, and S6E). Production of hallmark Th2 cytokines was confirmed by *in vitro* T cell restimulation, which led to the specific release of Th2 cytokines IL-4 and IL-13 but not IFN- γ (Figure 6F). These data illustrate that not only are the putative microbial allergens predicted by our pipeline capable of inducing skewed allergic immune responses *in vivo* but also that our pipeline successfully differentiates between allergic and non-allergic microbial proteases.

Our model was designed to predict enzymatically active protease allergens. The CP papain requires protease activity to initiate allergic immune responses, and irreversible disruption of papain's protease activity by either heat or pharmacological inactivation blocks its ability to induce an allergic immune response.¹⁴ To determine whether the predicted enzymatic activities of recombinantly derived *S. schleiferi* SP and *T. forsythia* CP were required for their immunostimulatory activity, we immunized mice with heat-inactivated SP or CP and analyzed downstream DC migration. Heat-inactivated recombinant SP and CP were incapable of inducing CD301b⁺ DC activation and migration, consistent with a requirement for enzymatic activity (Figures 6G–6I). Together, these data suggest that our MLP classifier is able to predict the presence of bona fide protease allergens from commensal species.

DISCUSSION

This study describes a deep learning framework using pLMs and AFDB Clusters to enhance identification of allergenic proteases in metagenomic datasets. Our approach successfully catalogs and predicts different putative allergenic proteases from gut and oral metagenomes. The conservation of the catalytic triad in typical SPs, shaped by convergent evolution,^{9,95} makes it an ideal target for deep learning and structural approaches. We demonstrated 99% accuracy of three state-of-the-art pLMs in distinguishing catalytic from non-catalytic serines.

Encouraged by pLMs' ability to identify catalytic triads in SHs, we hypothesized they could predict SP allergenicity. To address difficulties verifying negative examples, we trained our model to prioritize known allergens, penalizing negatives at varying rates. We expanded training data 10- to 100-fold using AFDB Clusters, incorporating SPs with low sequence homology but similar structure and putative function. This strategy proved effective, with our MLP classifier achieving an AUROC of 0.95, accurately

distinguishing subtle differences between allergenic and non-allergenic SPs.

When applied to large oral and gut microbiome protein catalogs, our model initially overpredicted allergens, likely due to overfitting on the small set of known allergenic proteases. This was rectified by retraining with over a million Pfam seed proteins, a step that served as a powerful regularizer. By including this vast negative set, we compelled the model to learn a simpler, more biologically meaningful decision boundary, forcing it to discard spurious correlations and instead identify the core features most predictive of allergenic function. This provided a broader proteome view, which elevated the model's performance in discriminating allergenic and non-allergenic proteases.

While these findings highlight the prevalence of predicted allergenic proteases in common gut taxa (e.g., *Lachnospiraceae* species), their presence alone does not confirm allergenicity. Using proteases predicted to be allergic or non-allergic, we validated their predicted allergic capability experimentally. This suggests that our model can be used in larger-scale efforts to determine an allergic risk score associated with colonization with microbial taxa. However, associations of certain taxa with allergies or immune-mediated effects underscore the complexity of microbial-host interactions. Additional work will need to be performed to validate the specific role of microbial proteases in promoting allergic disease.

Many predicted allergens belonged to the subtilase family, possibly reflecting its size as the second-largest SP family across life domains.^{96,97} While 48% of our initial ASP training seeds were from this family (potentially causing bias), several subtilisins are known allergens, and research links human leukocyte antigen (HLA)-DQ genotype to subtilisin hypersensitivity.⁹⁸

Our MLP model transcended primary sequence limitations, identifying critical functional motifs even within the lower sequence identity. It successfully flagged putative SP allergens sharing conserved catalytic triads with structural matches despite low pairwise identity. The model detected CPs despite their absence in training data, showcasing the model's ability to learn essential protein features. Experimental evaluation confirmed that two putative allergenic proteases directly promoted allergic immune responses dependent on their functional activity.

These implications extend beyond our specific datasets. Similar fine-tuning of embeddings from diverse enzyme classes

(B) Representative flow cytometry of CD11c+CD301b⁺ DCs from the dLN WT mice 20 h after i.d. immunization as in (A).

(C) Percent and number of PDL2+CD86+CD301b⁺ DCs were quantified by flow cytometry from mice treated contralaterally with sham protein injection, ovalbumin (OVA), or protease as in (A) and (B).

(D) Percent and number of PDL2+CD86+XCR1⁺ DCs were quantified by flow cytometry from mice treated as in (C).

(E) Percent of cytokine-positive CD4+CD44^{high} cells (gated as in Figures S6B–S6F) harvested 5 days after WT mice were immunized with sham with or without porcine trypsin (PT), endoproteinase Lys-C (EL-C), clostripain, CP, or SP.

(F) IL-4, IL-13, and IFN- γ ELISA of cell-free supernatant from the lymph node homogenate of mice immunized as in (E) after stimulation with CD3 and CD28 for 4 days *in vitro*.

(G) Representative flow cytometry of CD11c+CD301b⁺ DCs from the dLN of WT mice 20 h after contralateral i.d. immunization with active or heat-inactivated (HI) CP or SP protease.

(H and I) Percent and number of PDL2+CD86+CD301b⁺ DCs were quantified by flow cytometry from mice treated as in (G). Symbols represent individual lymph nodes (C, D, H, and I), mice or individual lymph nodes (E), or cell culture replicates (F). Bars indicate mean, and error bars indicate SEM. Statistical tests: two-way ANOVA with Sidák's multiple comparisons test (C and D), ordinary one-way ANOVA with multiple comparisons (E), two-sided unpaired Welch's *t* test (F), and two-sided paired *t* test (H and I). Data are representative of at least two independent experiments, combined in (C), (D), (F), (H), and (I), with each experiment including 3–5 mice per group. Data are representative of four independent experiments in which allergic and non-allergic proteases were tested separately and combined in (E), with each experiment including 3–4 mice per group. **p* < 0.05, ***p* < 0.01, ****p* < 0.001, *****p* < 0.0001.

has proven successful in predicting general function.^{38,99,100} While precise molecular pathways activated by protease allergens remain unclear, our work demonstrates the potential of deep learning and structural tools to identify novel allergens. With increasing metagenomic data available, this approach presents a promising avenue for discovering new microbial SPs and reshaping our understanding of microbiome-host interactions.

RESOURCE AVAILABILITY

Lead contact

Further information for resources should be directed to and will be fulfilled by the lead contact, Sergio Bacallado (sb2116@cam.ac.uk).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- This paper did not generate any new data.
- Source code, the trained model with training embeddings, and analysis scripts are available at https://gitlab.com/kthurime/serine_hydrolases.
- All of the MGnify protein catalog data are publicly available at the EBI FTP site https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/, and any of the clusters used in the initial data curation were found from the AFDB site <https://cluster.foldseek.com/>.
- Any additional information requested to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We would like to thank Martin Stražar, Sathish Subramanian, and Marie-Madlen Pust for feedback regarding the computational analysis of this project. We also thank Stephanie Aldrich and Theresa Reimels for editorial assistance with the manuscript and figures. K.T. was supported by the Gates Cambridge Trust and the Rotary Foundation. D.R.P. and R.J.X. are supported by the National Institutes of Health (NIH) DK043351, the Food Allergy Science Initiative (FASI), and the Klarman Cell Observatory at the Broad Institute. E.C. and C.L.S. are supported by NIH AI151163 and funding from FASI.

AUTHOR CONTRIBUTIONS

Conceptualization, K.T., C.L.S., R.J.X., and S.B.; methodology, K.T., D.R.P., and S.B.; software, K.T.; validation, K.T., E.W., C.L., and S.B.; formal analysis, K.T. and E.W.; data curation, K.T. and C.L.S.; writing – original draft, K.T.; writing – review and editing, K.T., E.W., C.L., D.B.G., R.M.O., D.R.P., C.L.S., R.J.X., and S.B.; visualization, K.T. and E.W.; supervision, C.L.S., D.R.P., R.J.X., and S.B.; project administration, D.R.P., R.J.X., and S.B.; resources, C.L.S., R.J.X., and S.B.

DECLARATION OF INTERESTS

R.J.X. is co-founder of Jnana Therapeutics and ConvergenceBio; scientific advisory board member at Nestlé, Arena BioWorks, and Magnet Biomedicine; and board director at MoonLake Immunotherapeutics. D.R.P. is an employee of Novonosis, Denmark. C.L.S. is a consultant for Astria Therapeutics, Bayer, and Merck and a scientific advisory board member at Granite Bio. These organizations had no roles in this study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
 - Mice

METHOD DETAILS

- Materials Availability
- Catalytic Serine Classification in SHs
- Catalytic Serine model explainability analysis
- Constructing a Serine Protease Allergen Set
- Expanding Training Data with AFDB Clusters
- Support Vector Machine training and testing
- Feature map
- Hyperparameter tuning
- Multilayer perceptron neural network
- Hyperparameter optimization and training
- Processing and applying classifiers to catalogs
- Prioritization of allergen protein sequences
- Taxonomic assignment and visualization
- Preparation of protein structure pictures
- DC migration
- T cell differentiation
- T cell in vitro stimulation
- Cytokine enzyme-linked immunosorbent assay (ELISA)

QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.cels.2025.101510>.

Received: September 26, 2024

Revised: April 10, 2025

Accepted: December 17, 2025

Published: February 20, 2026

REFERENCES

- Galli, S.J., Tsai, M., and Piliponsky, A.M. (2008). The development of allergic inflammation. *Nature* 454, 445–454. <https://doi.org/10.1038/nature07204>.
- Murrison, L.B., Brandt, E.B., Myers, J.B., and Hershey, G.K.K. (2019). Environmental exposures and mechanisms in allergy and asthma development. *J. Clin. Invest.* 129, 1504–1515. <https://doi.org/10.1172/JCI124612>.
- Carrol, C., Duval, A., Schmitt, P., Roga, S., Camus, M., Stella, A., Bulet-Schiltz, O., Gonzalez-de-Peredo, A., and Girard, J.-P. (2018). Environmental allergens induce allergic inflammation through proteolytic maturation of IL-33. *Nat. Immunol.* 19, 375–385. <https://doi.org/10.1038/s41590-018-0067-5>.
- Serhan, N., Basso, L., Sibillano, R., Petitfils, C., Meixiong, J., Bonnard, C., Reber, L.L., Marichal, T., Starkl, P., Cenac, N., et al. (2019). House dust mites activate nociceptor-mast cell clusters to drive type 2 skin inflammation. *Nat. Immunol.* 20, 1435–1443. <https://doi.org/10.1038/s41590-019-0493-z>.
- Soh, W.T., Zhang, J., Hollenberg, M.D., Vliagoftis, H., Rothenberg, M.E., Sokol, C.L., Robinson, C., and Jacquet, A. (2023). Protease allergens as initiators-regulators of allergic inflammation. *Allergy* 78, 1148–1168. <https://doi.org/10.1111/all.15678>.
- Vidal-Quist, J.C., Ortego, F., and Hernández-Crespo, P. (2021). Contribution of cysteine and serine proteases to proteolytic digestion in an allergy-eliciting house dust mite. *J. Insect Physiol.* 133, 104285. <https://doi.org/10.1016/j.jinsphys.2021.104285>.
- Teufelberger, A.R., Van Nevel, S., Hulpiau, P., Nordengrün, M., Savvides, S.N., De Graeve, S., Akula, S., Holtappels, G., De Ruyck, N., Declercq, W., et al. (2020). Mouse Strain-Dependent Difference Toward the Staphylococcus aureus Allergen Serine Protease-Like Protein D Reveals a Novel Regulator of IL-33. *Front. Immunol.* 11, 582044. <https://doi.org/10.3389/fimmu.2020.582044>.
- Kunimine, S., Takai, T., Kamijo, S., Maruyama, N., Kimitsu, T., Masutani, Y., Yoshimura, T., Suchiva, P., Shimizu, S., Ogawa, H., et al. (2021).

- Epicutaneous vaccination with protease inhibitor-treated papain prevents papain-induced Th2-mediated airway inflammation without inducing Th17 in mice. *Biochem. Biophys. Res. Commun.* 546, 192–199. <https://doi.org/10.1016/j.bbrc.2020.12.090>.
9. Gherardini, P.F., Wass, M.N., Helmer-Citterich, M., and Sternberg, M.J.E. (2007). Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.* 372, 817–845. <https://doi.org/10.1016/j.jmb.2007.06.017>.
10. Ozhelvaci, F., and Steczkiewicz, K. (2023). Identification and classification of papain-like cysteine proteinases. *J. Biol. Chem.* 299, 104801. <https://doi.org/10.1016/j.jbc.2023.104801>.
11. Sbihi, H., Boutin, R.C., Cutler, C., Suen, M., Finlay, B.B., and Turvey, S.E. (2019). Thinking bigger: How early-life environmental exposures shape the gut microbiome and influence the development of asthma and allergic disease. *Allergy* 74, 2103–2115. <https://doi.org/10.1111/all.13812>.
12. Breiteneder, H., Peng, Y.-Q., Agache, I., Diamant, Z., Eiwegger, T., Fokkens, W.J., Traidl-Hoffmann, C., Nadeau, K., O’Hehir, R.E., O’Mahony, L., et al. (2020). Biomarkers for diagnosis and prediction of therapy responses in allergic diseases and asthma. *Allergy* 75, 3039–3068. <https://doi.org/10.1111/all.14582>.
13. Canbay, V., and Auf dem Keller, U. (2021). New strategies to identify protease substrates. *Curr. Opin. Chem. Biol.* 60, 89–96. <https://doi.org/10.1016/j.cbpa.2020.09.009>.
14. Sokol, C.L., Barton, G.M., Farr, A.G., and Medzhitov, R. (2008). A mechanism for the initiation of allergen-induced T helper type 2 responses. *Nat. Immunol.* 9, 310–318. <https://doi.org/10.1038/ni1558>.
15. Heutinck, K.M., ten Berge, I.J.M., Hack, C.E., Hamann, J., and Rowshani, A.T. (2010). Serine proteases of the human immune system in health and disease. *Mol. Immunol.* 47, 1943–1955. <https://doi.org/10.1016/j.molimm.2010.04.020>.
16. Long, J.Z., and Cravatt, B.F. (2011). The metabolic serine hydrolases and their functions in mammalian physiology and disease. *Chem. Rev.* 111, 6022–6063. <https://doi.org/10.1021/cr200075y>.
17. Simon, G.M., and Cravatt, B.F. (2010). Activity-based proteomics of enzyme superfamilies: serine hydrolases as a case study. *J. Biol. Chem.* 285, 11051–11055. <https://doi.org/10.1074/jbc.R109.097600>.
18. Stentzel, S., Teufelberger, A., Nordengrün, M., Kolata, J., Schmidt, F., van Crombruggen, K., Michalik, S., Kumpfmüller, J., Tischer, S., Schweder, T., et al. (2017). Staphylococcal serine protease-like proteins are pacemakers of allergic airway reactions to *Staphylococcus aureus*. *J. Allergy Clin. Immunol.* 139, 492–500.e8. <https://doi.org/10.1016/j.jaci.2016.03.045>.
19. Perner, C., Flayer, C.H., Zhu, X., Aderhold, P.A., Dewan, Z.N.A., Voisin, T., Camire, R.B., Chow, O.A., Chiu, I.M., and Sokol, C.L. (2020). Substance P Release by Sensory Neurons Triggers Dendritic Cell Migration and Initiates the Type-2 Immune Response to Allergens. *Immunity* 53, 1063–1077.e7. <https://doi.org/10.1016/j.immuni.2020.10.001>.
20. Heuberger, D.M., and Schuepbach, R.A. (2019). Protease-activated receptors (PARs): mechanisms of action and potential therapeutic modulators in PAR-driven inflammatory diseases. *Thromb. J.* 17, 4. <https://doi.org/10.1186/s12959-019-0194-8>.
21. Buhl, T., Ikoma, A., Kempkes, C., Cevikbas, F., Sulk, M., Buddenkotte, J., Akiyama, T., Crumrine, D., Camerer, E., Carstens, E., et al. (2020). Protease-Activated Receptor-2 Regulates Neuro-Epidermal Communication in Atopic Dermatitis. *Front. Immunol.* 11, 1740. <https://doi.org/10.3389/fimmu.2020.01740>.
22. Meloun, A., and León, B. (2023). Sensing of protease activity as a triggering mechanism of Th2 cell immunity and allergic disease. *Front. Allergy* 4, 1265049. <https://doi.org/10.3389/falgy.2023.1265049>.
23. Reddy, V.B., Sun, S., Azimi, E., Elmariah, S.B., Dong, X., and Lerner, E.A. (2015). Redefining the concept of protease-activated receptors: cathepsin S evokes itch via activation of Mrgpr. *Nat. Commun.* 6, 7864. <https://doi.org/10.1038/ncomms8864>.
24. Reddy, V.B., and Lerner, E.A. (2017). Activation of mas-related G-protein-coupled receptors by the house dust mite cysteine protease Der p1 provides a new mechanism linking allergy and inflammation. *J. Biol. Chem.* 292, 17399–17406. <https://doi.org/10.1074/jbc.M117.787887>.
25. Serhan, N., Cenac, N., Basso, L., and Gaudenzio, N. (2021). Mas-related G protein-coupled receptors (Mrgprs) - Key regulators of neuroimmune interactions. *Neurosci. Lett.* 749, 135724. <https://doi.org/10.1016/j.neulet.2021.135724>.
26. Flayer, C.H., Kernin, I.J., Matatia, P.R., Zeng, X., Yarmolinsky, D.A., Han, C., Naik, P.R., Buttaci, D.R., Aderhold, P.A., Camire, R.B., et al. (2024). A $\gamma\delta$ T cell-IL-3 axis controls allergic responses through sensory neurons. *Nature* 634, 440–446. <https://doi.org/10.1038/s41586-024-07869-0>.
27. Deng, L., Costa, F., Blake, K.J., Choi, S., Chandrabalan, A., Yousuf, M.S., Shiers, S., Dubreuil, D., Vega-Mendoza, D., Rolland, C., et al. (2023). *S. aureus* drives itch and scratch-induced skin damage through a V8 protease-PAR1 axis. *Cell* 186, 5375–5393.e25. <https://doi.org/10.1016/j.cell.2023.10.019>.
28. van den Boogaard, F.E., Brands, X., Duitman, J., de Stoppelaar, S.F., Borensztajn, K.S., Roelofs, J.J.T.H., Hollenberg, M.D., Spek, C.A., Schultz, M.J., van ’t Veer, C., et al. (2018). Protease-Activated Receptor 2 Facilitates Bacterial Dissemination in Pneumococcal Pneumonia. *J. Infect. Dis.* 217, 1462–1471. <https://doi.org/10.1093/infdis/jiy010>.
29. Santiago, A., Hann, A., Constante, M., Rahmani, S., Libertucci, J., Jackson, K., Rueda, G., Rossi, L., Ramachandran, R., Ruf, W., et al. (2023). Crohn’s disease proteolytic microbiota enhances inflammation through PAR2 pathway in gnotobiotic mice. *Gut Microbes* 15, 2205425. <https://doi.org/10.1080/19490976.2023.2205425>.
30. Jimenez-Vargas, N.N., Pattison, L.A., Zhao, P., Lieu, T., Latorre, R., Jensen, D.D., Castro, J., Aurelio, L., Le, G.T., Flynn, B., et al. (2018). Protease-activated receptor-2 in endosomes signals persistent pain of irritable bowel syndrome. *Proc. Natl. Acad. Sci. USA* 115, E7438–E7447. <https://doi.org/10.1073/pnas.1721891115>.
31. Latorre, R., Hegron, A., Peach, C.J., Teng, S., Tonello, R., Retamal, J.S., Klein-Cloud, R., Bok, D., Jensen, D.D., Gottesman-Katz, L., et al. (2022). Mice expressing fluorescent PAR₂ reveal that endocytosis mediates colonic inflammation and pain. *Proc. Natl. Acad. Sci. USA* 119, e2112059119. <https://doi.org/10.1073/pnas.2112059119>.
32. Dimitrov, I., Naneva, L., Doytchinova, I., and Bangov, I. (2014). AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics* 30, 846–851. <https://doi.org/10.1093/bioinformatics/btt619>.
33. Muh, H.C., Tong, J.C., and Tammi, M.T. (2009). AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. *PLoS One* 4, e5861. <https://doi.org/10.1371/journal.pone.0005861>.
34. Sharma, N., Patiyal, S., Dhall, A., Pande, A., Arora, C., and Raghava, G.P.S. (2021). AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief. Bioinform.* 22, bbaa294. <https://doi.org/10.1093/bib/bbaa294>.
35. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
36. Nguyen, M.N., Krutz, N.L., Limviphuvadh, V., Lopata, A.L., Gerberick, G.F., and Maurer-Stroh, S. (2022). AllerCatPro 2.0: a web server for predicting protein allergenicity potential. *Nucleic Acids Res.* 50, W36–W43. <https://doi.org/10.1093/nar/gkac446>.
37. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
38. Bileschi, M.L., Belanger, D., Bryant, D.H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M.A., and Colwell, L.J. (2022). Using deep learning to annotate the protein universe. *Nat. Biotechnol.* 40, 932–937. <https://doi.org/10.1038/s41587-021-01179-w>.

39. Gligorijević, V., Renfrew, P.D., Kosciółek, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., et al. (2021). Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* 12, 3168. <https://doi.org/10.1038/s41467-021-23303-9>.
40. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130. <https://doi.org/10.1126/science.ade2574>.
41. Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdriz, G., Zhang, J., Church, G.M., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* 40, 1617–1623. <https://doi.org/10.1038/s41587-022-01432-w>.
42. Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Jr., Xiong, C., Sun, Z.Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* 41, 1099–1106. <https://doi.org/10.1038/s41587-022-01618-2>.
43. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 42, 243–246. <https://doi.org/10.1038/s41587-023-01773-0>.
44. Holm, L. (2022). Dali server: structural unification of protein families. *Nucleic Acids Res.* 50, W210–W215. <https://doi.org/10.1093/nar/gkac387>.
45. Shanthappa, P.M., and Kumar, R. (2022). ProAll-D: protein allergen detection using long short term memory - a deep learning approach. *ADMET DMPK* 10, 231–240. <https://doi.org/10.5599/admet.1335>.
46. Garcia-Moreno, F.M., and Gutiérrez-Naranjo, M.A. (2022). ALLERDET: A novel web app for prediction of protein allergenicity. *J. Biomed. Inform.* 135, 104217. <https://doi.org/10.1016/j.jbi.2022.104217>.
47. He, C., Ye, X., Yang, Y., Hu, L., Si, Y., Zhao, X., Chen, L., Fang, Q., Wei, Y., Wu, F., et al. (2023). DeepAlgPro: an interpretable deep neural network model for predicting allergenic proteins. *Brief. Bioinform.* 24, bbad246. <https://doi.org/10.1093/bib/bbad246>.
48. Heinzinger, M., Littmann, M., Sillitoe, I., Bordin, N., Orengo, C., and Rost, B. (2022). Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom. Bioinform.* 4, lqac043. <https://doi.org/10.1093/nargab/lqac043>.
49. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7112–7127. <https://doi.org/10.1109/TPAMI.2021.3095381>.
50. Kaminski, K., Ludwiczak, J., Alva, V., and Dunin-Horkawicz, S. (2023). pLM-BLAST: distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics* 39, btad579. <https://doi.org/10.1093/bioinformatics/btad579>.
51. Bepler, T., and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell Syst.* 12, 654–669.e3. <https://doi.org/10.1016/j.cels.2021.05.017>.
52. Jokinen, E., Dumitrescu, A., Huuhtanen, J., Gligorijević, V., Mustjoki, S., Bonneau, R., Heinonen, M., and Lähdesmäki, H. (2023). TCRconv: predicting recognition between T cell receptors and epitopes using contextualized motifs. *Bioinformatics* 39, btac788. <https://doi.org/10.1093/bioinformatics/btac788>.
53. Yadav, S., Vora, D.S., Sundar, D., and Dhanjal, J.K. (2023). TCR-ESM: employing protein language embeddings to predict TCR-peptide-MHC binding. *Comput. Struct. Biotechnol. J.* 23, 165–173. <https://doi.org/10.1016/j.csbj.2023.11.037>.
54. Volzhenin, K., Bittner, L., and Carbone, A. (2024). SENSE-PPI reconstructs protein-protein interactions of various complexities, within, across, and between species, with sequence-based evolutionary scale modeling and deep learning. *iScience* 27, 110371. <https://doi.org/10.1016/j.isci.2024.110371>.
55. Sledzieski, S., Singh, R., Cowen, L., and Berger, B. (2021). D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst.* 12, 969–982.e6. <https://doi.org/10.1016/j.cels.2021.08.010>.
56. Singh, R., Sledzieski, S., Bryson, B., Cowen, L., and Berger, B. (2023). Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc. Natl. Acad. Sci. USA* 120, e2220778120. <https://doi.org/10.1073/pnas.2220778120>.
57. Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinform.* 20, 723. <https://doi.org/10.1186/s12859-019-3220-8>.
58. Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., et al. (2022). High-resolution de novo structure prediction from primary sequence. Preprint at bioRxiv. <https://doi.org/10.1101/2022.07.21.500999>.
59. Sarlakifar, F., Malek, H., Allahyari Fard, N., and Khotanlou, Z. (2024). AllerTrans: A Deep Learning Method for Predicting the Allergenicity of Protein Sequences. *Biol. Methods Protoc.* 10, bpaf040.
60. Schütze, K., Heinzinger, M., Steinegger, M., and Rost, B. (2022). Nearest neighbor search on embeddings rapidly identifies distant protein relations. *Front. Bioinform.* 2, 1033775. <https://doi.org/10.3389/fbinf.2022.1033775>.
61. Marquet, C., Heinzinger, M., Olenyi, T., Dallago, C., Erckert, K., Bernhofer, M., Nechaev, D., and Rost, B. (2022). Embeddings from protein language models predict conservation and variant effects. *Hum. Genet.* 141, 1629–1647. <https://doi.org/10.1007/s00439-021-02411-y>.
62. Yang, P., Li, X.-L., Mei, J.-P., Kwok, C.-K., and Ng, S.-K. (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics* 28, 2640–2647. <https://doi.org/10.1093/bioinformatics/bts504>.
63. Bekker, J., and Davis, J. (2020). Learning from positive and unlabeled data: a survey. *Mach. Learn.* 109, 719–760. <https://doi.org/10.1007/s10994-020-05877-5>.
64. Noumir, Z., Honeine, P., and Richard, C. (2012). On simple one-class classification methods. 2012 IEEE International Symposium on Information Theory Proceedings, 2012 (IEEE), pp. 2022–2026. <https://doi.org/10.1109/ISIT.2012.6283685>.
65. Qian, M., Li, Y.-F., and Han, T. (2022). Positive-Unlabeled Learning-Based Hybrid Deep Network for Intelligent Fault Detection. *IEEE Trans. Ind. Inform.* 18, 4510–4519. <https://doi.org/10.1109/TII.2021.3121777>.
66. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenoltz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 39, 105–114. <https://doi.org/10.1038/s41587-020-0603-3>.
67. Poyet, M., Groussin, M., Gibbons, S.M., Avila-Pacheco, J., Jiang, X., Kearney, S.M., Perrotta, A.R., Berdy, B., Zhao, S., Lieberman, T.D., et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* 25, 1442–1452. <https://doi.org/10.1038/s41591-019-0559-3>.
68. Gurbich, T.A., Almeida, A., Beracochea, M., Burdett, T., Burgin, J., Cochrane, G., Raj, S., Richardson, L., Rogers, A.B., Sakharova, E., et al. (2023). MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues. *J. Mol. Biol.* 435, 168016. <https://doi.org/10.1016/j.jmb.2023.168016>.
69. Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C.L.M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., and Steinegger, M. (2023). Clustering predicted structures at the scale of the known protein universe. *Nature* 622, 637–645. <https://doi.org/10.1038/s41586-023-06510-w>.
70. Holm, C., Davis, R.C., Osterlund, T., Schotz, M.C., and Fredrikson, G. (1994). Identification of the active site serine of hormone-sensitive lipase

- by site-directed mutagenesis. *FEBS Lett.* **344**, 234–238. [https://doi.org/10.1016/0014-5793\(94\)00403-x](https://doi.org/10.1016/0014-5793(94)00403-x).
71. Peitsaro, N., Polianskyte, Z., Tuimala, J., Pörn-Ares, I., Liobikas, J., Speer, O., Lindholm, D., Thompson, J., and Eriksson, O. (2008). Evolution of a family of metazoan active-site-serine enzymes from penicillin-binding proteins: a novel facet of the bacterial legacy. *BMC Evol. Biol.* **8**, 26. <https://doi.org/10.1186/1471-2148-8-26>.
72. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777.
73. Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760. <https://doi.org/10.1038/s41551-018-0304-0>.
74. Sonnhammer, E.L., Eddy, S.R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420. [https://doi.org/10.1002/\(sici\)1097-0134\(199707\)28:3<405::aid-prot10>3.0.co;2-l](https://doi.org/10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l).
75. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419. <https://doi.org/10.1093/nar/gkaa913>.
76. Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., and Stoica, I. (2018). Tune: A Research Platform for Distributed Model Selection and Training. Preprint at arXiv.
77. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368. <https://doi.org/10.1038/s41592-021-01101-x>.
78. UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531. <https://doi.org/10.1093/nar/gkac1052>.
79. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240. <https://doi.org/10.1093/bioinformatics/btu031>.
80. Könönen, E., Fteita, D., Gursoy, U.K., and Gursoy, M. (2022). Prevotella species as oral residents and infectious agents with potential impact on systemic conditions. *J. Oral Microbiol.* **14**, 2079814. <https://doi.org/10.1080/20002297.2022.2079814>.
81. Lopes Dos Santos Santiago, G., Brusselle, G., Dauwe, K., Deschaght, P., Verhofstede, C., Vanechoutte, D., Deschepper, E., Jordens, P., Joos, G., and Vanechoutte, M. (2017). Influence of chronic azithromycin treatment on the composition of the oropharyngeal microbial community in patients with severe asthma. *BMC Microbiol.* **17**, 109. <https://doi.org/10.1186/s12866-017-1022-6>.
82. Berni Canani, R., Sangwan, N., Stefka, A.T., Nocerino, R., Paparo, L., Aitoro, R., Calignano, A., Khan, A.A., Gilbert, J.A., and Nagler, C.R. (2016). Lactobacillus rhamnosus GG-supplemented formula expands butyrate-producing bacterial strains in food allergic infants. *ISME J.* **10**, 742–750. <https://doi.org/10.1038/ismej.2015.151>.
83. Fazlollahi, M., Chun, Y., Grishin, A., Wood, R.A., Burks, A.W., Dawson, P., Jones, S.M., Leung, D.Y.M., Sampson, H.A., Sicherer, S.H., et al. (2018). Early-life gut microbiome and egg allergy. *Allergy* **73**, 1515–1524. <https://doi.org/10.1111/all.13389>.
84. Li, C., Stražar, M., Mohamed, A.M.T., Pacheco, J.A., Walker, R.L., Lebar, T., Zhao, S., Lockart, J., Dame, A., Thurimella, K., et al. (2024). Gut microbiome and metabolome profiling in Framingham heart study reveals cholesterol-metabolizing bacteria. *Cell* **187**, 1834–1852.e19. <https://doi.org/10.1016/j.cell.2024.03.014>.
85. Chandrasekaran, P., Han, Y., Zerbe, C.S., Heller, T., DeRavin, S.S., Kreuzberg, S.A., Marciano, B.E., Siu, Y., Jones, D.R., Abraham, R.S., et al. (2023). Intestinal microbiome and metabolome signatures in patients with chronic granulomatous disease. *J. Allergy Clin. Immunol.* **152**, 1619–1633.e11. <https://doi.org/10.1016/j.jaci.2023.07.022>.
86. Zhu, L., Jian, X., Zhou, B., Liu, R., Muñoz, M., Sun, W., Xie, L., Chen, X., Peng, C., Maurer, M., et al. (2024). Gut microbiota facilitate chronic spontaneous urticaria. *Nat. Commun.* **15**, 112. <https://doi.org/10.1038/s41467-023-44373-x>.
87. De Filippis, F., Paparo, L., Nocerino, R., Della Gatta, G., Carucci, L., Russo, R., Pasolli, E., Ercolini, D., and Berni Canani, R. (2021). Specific gut microbiome signatures and the associated pro-inflammatory functions are linked to pediatric allergy and acquisition of immune tolerance. *Nat. Commun.* **12**, 5958. <https://doi.org/10.1038/s41467-021-26266-z>.
88. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94. <https://doi.org/10.1093/protein/12.2.85>.
89. Zhang, C., Shine, M., Pyle, A.M., and Zhang, Y. (2022). US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115. <https://doi.org/10.1038/s41592-022-01585-1>.
90. Chen, I.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S.J., Webb, C., Wu, D., et al. (2023). The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* **51**, D723–D732. <https://doi.org/10.1093/nar/gkac976>.
91. Downes, J., Munson, M., and Wade, W.G. (2003). Dialister invisus sp. nov., isolated from the human oral cavity. *Int. J. Syst. Evol. Microbiol.* **53**, 1937–1940. <https://doi.org/10.1099/ijs.0.02640-0>.
92. Rashidi, A., Ebadi, M., Weisdorf, D.J., Costalonga, M., and Staley, C. (2021). No evidence for colonization of oral bacteria in the distal gut in healthy adults. *Proc. Natl. Acad. Sci. USA* **118**, e2114152118. <https://doi.org/10.1073/pnas.2114152118>.
93. Kumamoto, Y., Linehan, M., Weinstein, J.S., Laidlaw, B.J., Craft, J.E., and Iwasaki, A. (2013). CD301b⁺ dermal dendritic cells drive T helper 2 cell-mediated immunity. *Immunity* **39**, 733–743. <https://doi.org/10.1016/j.immuni.2013.08.029>.
94. Sokol, C.L., Camire, R.B., Jones, M.C., and Luster, A.D. (2018). The chemokine receptor CCR8 promotes the migration of dendritic cells into the lymph node parenchyma to initiate the allergic immune response. *Immunity* **49**, 449–463.e6. <https://doi.org/10.1016/j.immuni.2018.07.012>.
95. Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I., and Schrag, J. (1992). The α/β hydrolase fold. *Protein Eng.* **5**, 197–211. <https://doi.org/10.1093/protein/5.3.197>.
96. Siezen, R.J., and Leunissen, J.A. (1997). Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Sci.* **6**, 501–523. <https://doi.org/10.1002/pro.5560060301>.
97. Figueiredo, J., Sousa Silva, M., and Figueiredo, A. (2018). Subtilisin-like proteases in plant defence: the past, the present and beyond. *Mol. Plant Pathol.* **19**, 1017–1028. <https://doi.org/10.1111/mpp.12567>.
98. Xue, A., Chapoval, S.P., Finn, E.S., Chowdhary, V.R., Marietta, E.V., Gaffey, T.A., Pursifull, A.C., and David, C.S. (2005). HLA-DQ8 is a predisposing molecule for detergent enzyme subtilisin BPN'-induced hypersensitivity. *Clin. Immunol.* **117**, 302–315. <https://doi.org/10.1016/j.clim.2005.08.003>.
99. Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., and Zhao, H. (2023). Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363. <https://doi.org/10.1126/science.adf2465>.
100. Thurimella, K., Mohamed, A.M., Li, C., Vatanen, T., Graham, D.B., Owens, R.M., La Rosa, S.L., Plichta, D.R., Bacallado, S., and Xavier, R.J. (2025). Protein language models uncover carbohydrate-active enzyme function in metagenomics. *BMC bioinformatics* **26** (1), 285. <https://doi.org/10.1186/s12859-025-06286-y>.
101. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011).

- Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.
102. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357.
103. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>.
104. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 8026–8037.
105. Falcon, W., The PyTorch Lightning team. (2019). PyTorch Lightning (Version 1.4) [Computer software]. <https://doi.org/10.5281/zenodo.3828935>
106. Schrödinger, LLC. (2015). The PyMOL Molecular Graphics System. Version 1.8.
107. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296. <https://doi.org/10.1093/nar/gkab301>.
108. Bachovchin, D.A., and Cravatt, B.F. (2012). The pharmacological landscape and therapeutic potential of serine hydrolases. *Nat. Rev. Drug Discov.* **11**, 52–68. <https://doi.org/10.1038/nrd3620>.
109. UniProt Consortium (2021). UniProt: the universal protein knowledge-base in 2021. *Nucleic Acids Res.* **49**, D480–D489. <https://doi.org/10.1093/nar/gkaa1100>.
110. Mari, A., Rasi, C., Palazzo, P., and Scala, E. (2009). Allergen databases: current status and perspectives. *Curr. Allergy Asthma Rep.* **9**, 376–383. <https://doi.org/10.1007/s11882-009-0055-9>.
111. Mari, A., Scala, E., Palazzo, P., Ridolfi, S., Zennaro, D., and Carabella, G. (2006). Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a model. *Cell. Immunol.* **244**, 97–100. <https://doi.org/10.1016/j.cellimm.2007.02.012>.
112. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370. <https://doi.org/10.1093/nar/gkg095>.
113. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028. <https://doi.org/10.1038/nbt.3988>.
114. Scholkopf, B., and Smola, A.J. (2018). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press).
115. Agarap, A.F. (2018). *Deep Learning using Rectified Linear Units (ReLU)*. Preprint at arXiv.
116. Loshchilov, I., and Hutter, F. (2017). Decoupled Weight Decay Regularization. ICLR 2019.
117. Li, L., Jamieson, K., Rostamizadeh, A., Gonina, K., Hardt, M., Recht, B., and Talwalkar, A. (2020). Massively Parallel Hyperparameter Tuning. In *Proceedings of the 3rd MLSys Conference*.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-mouse IL-13, Clone: eBio13A	eBioscience	Cat# 25-7133-82; RRID:AB_2573530
Anti-mouse CD11c, Clone: N418	BioLegend	Cat# 117334; RRID:AB_2562415
Anti-mouse CD172a (Sirpa), Clone: P84	BioLegend	Cat# 144012, RRID:AB_2563550
Anti-mouse CD19, Clone: 6D5	BioLegend	Cat# 115545, RRID:AB_2562136
Anti-mouse CD273 (PDL2), Clone: TY25	BioLegend	Cat# 107219, RRID:AB_2728127
Anti-mouse CD3, Clone: 17A2	BioLegend	Cat# 100210, RRID:AB_389301
Anti-mouse CD3, Clone: 17A2	BioLegend	Cat# 100227, RRID:AB_10900227
Anti-mouse CD301b (MGL2), Clone: URA-1	BioLegend	Cat# 146808, RRID:AB_2563390
Anti-mouse CD4, Clone: GK1.5	BioLegend	Cat# 100451, RRID:AB_2564591
Anti-mouse CD4, Clone: GK1.5	BioLegend	Cat# 100412, RRID:AB_312697
Anti-mouse CD4, Clone: RM4-5	BioLegend	Cat# 100555, RRID:AB_2562529
Anti-mouse CD4, Clone: RM4-5	BioLegend	Cat# 100536, RRID:AB_493701
Anti-mouse CD86, Clone: GL-1	BioLegend	Cat# 105035, RRID:AB_11126147
Anti-mouse CD8 α , Clone: 53-6.7	BioLegend	Cat# 100743, RRID:AB_2561352
Anti-mouse CD8 α , Clone: 53-6.7	BioLegend	Cat# 100734, RRID:AB_2075238
Anti-mouse CD8 α , Clone: 53-6.7	BioLegend	Cat# 100722, RRID:AB_312761
Anti-mouse CD8 α , Clone: 53-6.7	BioLegend	Cat# 100708, RRID:AB_312747
Anti-mouse FoxP3, Clone: MF-14	BioLegend	Cat# 126404, RRID:AB_1089117
Anti-mouse IFN- γ , Clone: XMG1.2	BioLegend	Cat# 505810, RRID:AB_315404
Anti-mouse IL-10, Clone: JES5-16E3	BioLegend	Cat# 505021, RRID:AB_10900417
Anti-mouse IL-4, Clone: 11B11	eBioscience	Cat# 53-7041-82, RRID:AB_469922
Anti-mouse MHCII, Clone: M5/114.15.2	BioLegend	Cat# 107636, RRID:AB_2734168
Anti-mouse NK1.1, Clone: PK136	BioLegend	Cat# 108737, RRID:AB_2562216
Anti-mouse XCR1, Clone: ZET	BioLegend	Cat# 148214, RRID:AB_2564369
Anti-mouse/human CD11b, Clone: M1/70	BioLegend	Cat# 101245, RRID:AB_2561390
Anti-mouse/human CD44, Clone: IM7	BioLegend	Cat# 103026, RRID:AB_493713
Anti-mouse/humanCD11b, Clone: RUO	BioLegend	Cat# 101206, RRID:AB_312789
Biotin anti-mouse IFN- γ Antibody, Clone: R4-6A2	BioLegend	Cat# 505704, RRID:AB_315392
Biotin anti-mouse IL-4 Antibody, Clone: BVD6-24G2	BioLegend	Cat# 504202, RRID:AB_315324
Biotin IL-13 Monoclonal Antibody, Clone: eBio1316H	eBioscience	Cat# 13-7135-85, RRID:AB_763556
eBioscience™ Fixable Viability Dye eFluor™ 780	Invitrogen	Cat# 65-0865-14
IFN gamma Monoclonal Antibody, Clone: XMG1.2	eBioscience	Cat# 14-7311-85, RRID:AB_468468
IL-13 Monoclonal Antibody, Clone: eBio13A	eBioscience	Cat# 14-7133-85, RRID:AB_763554
Purified anti-mouse IL-4 Antibody, Clone: 11B11	BioLegend	Cat# 504102, RRID:AB_315316
TruStainFcX (anti-mouse CD16/32), Clone: 93	BioLegend	Cat# 101320, RRID:AB_1574975
Chemicals, peptides, and recombinant proteins		
0.5 M EDTA, pH 8.0	Gibco	Cat # 15575-038
2-mercaptoethanol	Thermo Fisher Scientific	Cat # 21985023
BD GolgiPlug Protein Transport Inhibitor	Fisher Scientific	Cat # BDB555029
Clostripain, Clostridium histolyticum	Thermo Scientific Chemicals	Cat # AAJ61362MCR
Collagenase P	Sigma-Aldrich	Cat # 11213857001
Dispase II	Sigma-Aldrich	Cat # D4693-1G
DNase I, grade II, from bovine pancreas	Sigma-Aldrich	Cat # 10104159001

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Endoproteinase Lys-C	MedChemExpress	Cat # HY-P3208
Fetal bovine serum	Sigma-Aldrich	Cat # F4135-500ML
GlutaMAX Supplement	Fisher Scientific	Cat # 35050061
IFN-gamma (ELISA Standard) Recombinant Protein	eBioscience	Cat # 39-8311-65
Ionomycin	Sigma-Aldrich	Cat # I0634-1MG
Non-Essential Amino Acid Solution	Lonza	Cat # 536060
OPD	Sigma-Aldrich	Cat # P9187-50SET
Ovalbumin	Sigma-Aldrich	Cat # A5503
Papain from Carica papaya	Sigma-Aldrich	Cat # 76216-50MG
PBS	Corning	Cat # 21-031-CV
PMA, Molecular Biology Grade	Sigma-Aldrich	Cat # P1585-1MG
Putative cysteine protease (CP)	Genscript	N/A
Putative serine protease (SP)	Genscript	N/A
Recombinant Mouse IL-13 (ELISA Std.)	eBioscience	Cat # 29-8131-60
Recombinant Mouse IL-4 (ELISA Std.)	BioLegend	Cat # 575609
Red blood cell lysis buffer	Sigma-Aldrich	Cat # R7757
RPMI 1640, 1X	Corning	Cat # 15-040-CV
Trypsin from porcine pancreas	Sigma-Aldrich	Cat # T7409
Sodium Pyruvate (100 mM)	Thermo Fisher Scientific	Cat # 11360070

Critical commercial assays

BD Cytotfix/Cytoperm	BD Biosciences	Cat# 554714, RRID:AB_2869008
----------------------	----------------	------------------------------

Deposited data

Unified Human Gastrointestinal Genomes	MGNify - EBI	https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0.1/protein_catalogue/
Unified Human Oral Genomes	MGNify - EBI	https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-oral/v1.0/protein_catalogue/
Framingham Heart Study	Li et al. ⁸⁴	https://framinghamheartstudy.org/fhs-for-researchers/

Experimental models: Organisms/strains

Mouse: C57BL/6	Charles River Laboratory	Strain No.: 027; RRID:IMSR_CRL:027
----------------	--------------------------	------------------------------------

Software and algorithms

Allergen Annotation (analysis and tool)	10.5281/zenodo.1745964	https://gitlab.com/kthurime/serine_hydrolases
Graphpad Prism	Version 10.1.1	https://www.graphpad.com/scientific-software/prism/
ESM-2 protein language model	Lin et al. ⁴⁰	https://github.com/facebookresearch/esm
ProSE protein language model	Bepler and Berger ⁵¹	https://github.com/tbepler/prose
ProtT5 protein language model	Elnaggar et al. ⁴⁹	https://github.com/agemagician/ProtTrans
Scikit-learn	Pedregosa et al. ¹⁰¹	https://scikit-learn.org/stable/
SMOTE	Chawla et al. ¹⁰²	https://imbalanced-learn.org/stable/
SHAP	Lundberg et al. ^{72,73}	https://github.com/slundberg/shap
CD-HIT	Fu et al. ¹⁰³	https://github.com/weizhongli/cdhit
DIAMOND	Buchfink et al. ⁷⁷	https://github.com/bbuchfink/diamond
Foldseek AFDB Clusters	Barrio-Hernandez et al. ⁶⁹	https://cluster.foldseek.com/
PyTorch	Paszke et al. ¹⁰⁴	https://pytorch.org/
PyTorch Lightning	lightning.ai ¹⁰⁵	https://lightning.ai/pytorch-lightning
RayTune	Liaw et al. ⁷⁶	https://docs.ray.io/en/latest/tune/
ESMFold	Lin et al. ⁴⁰	https://github.com/facebookresearch/esm

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FlowJo	Tree Star	https://www.flowjo.com/solutions/flowjo/downloads
InterProScan	Jones et al. ⁷⁹	https://www.ebi.ac.uk/interpro/download/
US-Align	Zhang et al. ⁸⁹	https://zhanggroup.org/US-align/
PyMOL	Schrödinger, LLC ¹⁰⁶	https://pymol.org/
iTOL	Letunic and Bork ¹⁰⁷	https://itol.embl.de/
Prism v 10.0	GraphPad	https://www.graphpad.com/demos/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Mice

All animal experiments were approved by the Massachusetts General Hospital or Harvard Medical School Institutional Animal Care and Use Committee (IACUC). Mice were bred and housed under specific pathogen-free conditions at Massachusetts General Hospital in ventilated cages (≤ 5 per cage) on a 12-hour light/dark cycle at 20–25 °C and 35–65% humidity, with ad libitum access to food and water. Functionally wild-type C57BL/6 mice were either purchased from Charles River Laboratories (Wilmington, MA) or bred in-house. 6–12 week old, female and male mice were used for all experiments and numbers were matched among the experimental conditions.

METHOD DETAILS

Materials Availability

There are restrictions to the availability of the Serine Protease and Cysteine Protease used in the study because we had them both expressed and purified from a third-party vendor, Genscript, and used the entire aliquot in the above mentioned experiments.

Catalytic Serine Classification in SHs

Every SH sequence was collected from the 232 SHs in a previous study.¹⁰⁸ Each sequence was passed through the three protein language models (pLMs) ESM-2,⁴⁰ ProSE⁵¹ and ProtT5⁴⁹ and embeddings were generated on a per-residue basis. All of the serine embeddings were collected in every sequence and labeled as ‘catalytic’ or ‘non-catalytic’ using the information stored in UniProt.¹⁰⁹

To ensure deterministic results, we trained a logistic regression model using the LogisticRegression model from the sklearn.linear_model library¹⁰¹ with the random_state parameter set to 0. To address the class imbalance between the number of catalytic and non-catalytic serines, we employed the SMOTE (synthetic minority over-sampling technique)¹⁰² to create additional examples from the catalytic class, thus enhancing classification performance. After splitting the dataset into 75% for training and 25% for testing (test_train_split in sklearn.model_selection) the precision/recall and ROC curves were calculated (roc_curve, precision_recall_curve in sklearn.metrics). The F1-score was calculated on each of the held out test datasets for every pLM example as well (f1_score in sklearn.metrics).

We used principal component analysis (PCA) for the dimensionality reduction to evaluate the embeddings from the different classes (pca in sklearn.decomposition). We displayed the first and second principal components (PC) and the explained variance for each PC. Each plot was rendered in GraphPad Prism version 10.1.1 (270).

Catalytic Serine model explainability analysis

To delve into the decision-making process of our catalytic serine model, we employed SHAP (SHapley Additive exPlanations) values^{72,73} to interpret the model’s reliance on specific features within ProtT5 embeddings.

Specifically, we utilized the LinearExplainer module to analyze the contribution of each index value within the embeddings to the model’s predictions. To visualize these contributions, we generated force plots, highlighting the embedding features that exerted the most influence on the model’s output. We used the summary_plot function to visualize the specific contributions made by each feature.

Constructing a Serine Protease Allergen Set

A total of 33 allergenic serine protease (ASP) positives were downloaded from Allergome (www.allergome.com).^{110,111} The search query “serine protease” was used to filter specifically for SPs within the broader allergen database. The corresponding UniProt accession numbers¹⁰⁹ were then used to download the protein sequences of these allergens.

To compile a set of likely negative non-allergenic serine proteases (NASP), we constructed a query in Swiss-Prot¹¹² to identify bacterial proteins that were classified as serine proteases (keyword:KW-0720) but lacked evidence of allergenic activity. We restricted the query to bacterial proteases (taxonomy_name:2) smaller than 800 amino acids (length:[* TO 800]) that were not tagged as

allergens (KW-0020). The purpose of this filter was to reduce the initial NASP set size, which initially comprised 2,615 proteins, to a manageable number for further analysis. After filtering, 116 NASP serine proteases remained.

The true negative set, non-protease serine hydrolases (NPSH), was constructed using a previously curated list of 232 metabolic serine hydrolases (mSHs) and serine proteases (SPs) compiled by a previous study.¹⁰⁸ We excluded any mSHs or SPs that overlapped with the ASP or NASP sets based on their UniProt accession numbers and a protein sequence identity threshold of 100% to ensure the NPSH set consisted of distinct serine hydrolases not identified as allergens or likely negative examples. This resulted in 126 NPSH mSHs. All of the NASP and NPSH classes were run through CD-HIT¹⁰³ to ensure unique sequences at sequence identity 100%.

We define the term Serine Hydrolase SH training set as:

$$SH = ASP \cup NASP \cup NPSH \quad (\text{Equation 1})$$

We additionally trained Pfam seeds that were collected from Bileschi et al.³⁸ We performed an ultra-sensitive search between the SH dataset and all the Pfam seeds (1,296,280) using DIAMOND⁷⁷ and any seeds that overlapped between the two datasets were removed from the Pfam set, resulting in a total of 1,006,317 unique seeds.

All downloaded enzymes with their corresponding UniProt accession numbers are listed in Table S3.

Expanding Training Data with AFDB Clusters

To enrich the training data, each of the three groups was augmented with additional sequences retrieved using the Foldseek clustered AlphaFold database (AFDB Clusters).⁶⁹ The Foldseek cluster algorithm applies both sequence and structural homology to group proteins into clusters. Specifically, proteins are clustered if they share >50% sequence identity and >90% sequence coverage¹¹³ and exhibit >90% structural overlap.⁴³ Each UniProt ID in the training data was mapped to a unique cluster using the API provided on the AFDB Clusters website (<https://cluster.foldseek.com>). Then, within each cluster, only sequences associated with the “AFDB/Foldseek” clustering step were downloaded and incorporated into the augmented training set.

Utilizing AFDB cluster augmentation, the ASP, NASP, and NPSH sets were each expanded. The ASP set was increased to a total of 8,685 sequences, the NASP set to 20,708 sequences, and the NPSH set to 59,165 sequences.

Support Vector Machine training and testing

Consider a dataset $(x_i, y_i)_{i=1, \dots, n}$, where the input $x_i \in \mathbb{R}^{1024}$ is the pLM embedding of protein i , and the output y_i is a label in $\{ASP, NASP, NPSH\}$ (allergenic serine protease, non-allergenic serine proteases, and non-protease serine hydrolases). We generate a random split of the dataset into a training set $A \subset \{1, \dots, n\}$, which contains ASPs, NASPs, and NPSHs, and a validation set $B \subset \{1, \dots, n\}$, which only contains ASPs and NASPs, i.e. if $i \in B$, $y_i \in \{ASP, NASP\}$. We use the first subset A to train a weighted support vector machine with the aim of discriminating the positive class ASP from the negative classes NASP and NPSH. Recognizing that the negative class containing the *hardest* classification examples is NASP (as it consists of only SPs), the hyperparameters are tuned on the classification accuracy in a validation set B comprising only ASPs and NASPs.

The model is defined as follows. Let $\phi: \mathbb{R}^{1024} \rightarrow \mathbb{R}^d$ be a given feature map, which translates the input x into a feature vector $\phi(x)$; in the simplest version of the model, we can take the identity $\phi(x) = x$. The support vector machine discriminates positives from true negatives through a score function $f(x) = \phi(x)^T \beta$ which is linear in the features, i.e. $f(x) > 0$ indicates that an example with input x is positive. The parameter β is estimated through the following optimization:

$$\text{minimize } \sum_{i \in A} w_{y_i} \max(0, 1 - \phi(x_i)^T \beta \rho(y_i)) + \lambda \|\beta\|^2 \quad \text{subject to } \beta \in \mathbb{R}^d \quad (\text{Equation 2})$$

where, $\rho(y_i) = 1$ if $y_i = ASP$ and $\rho(y_i) = -1$ if $y_i \in \{NASP, NPSH\}$. The weights $w_{ASP}, w_{NASP}, w_{NPSH} > 0$ are hyperparameters to be learned. We learn these hyperparameters, as well as $\lambda > 0$, by cross-validation, optimizing the classification accuracy on the validation set which only contains ASPs and NASPs. In this way, the two negative classes in the training set are used only to the extent to which they help classify ASPs (positives) and NASPs (*hard* negative cases).

Feature map

We use a feature map ϕ associated with the Gaussian or RBF kernel; that is for any pair of inputs $x, x' \in \mathcal{P}$, the inner product $\langle \phi(x), \phi(x') \rangle = e^{-\|x - x'\|^2 / (2\gamma^2)}$

. The score function f for this feature map has the property that $|f(x) - f(x')|$ is small whenever the Euclidean distance $\|x - x'\| / (2\gamma^2)$ is small; i.e., we make similar predictions at inputs which are close in a Euclidean sense. Here, $\gamma > 0$ is a *scale* hyperparameter which is also learned by cross-validation. The feature vector in this case lies in an infinite-dimensional Hilbert Space, but the representer theorem allows us to solve the optimization problem (Equation 2) explicitly.¹¹⁴

Hyperparameter tuning

We used a custom-built script to randomly (uniformly) pick values for each of the 3 different class weights for the ASP, NASP, NPSH classes as well as for the gamma parameter, which corresponds to the kernel bandwidth (Table S4). Data S1 contains a log file generated by our custom script, representing the raw output of two different trials run for concurrent SVM training.

Multilayer perceptron neural network

A limitation of the support vector machine is that the score function $f(x)$ is linear in the feature vector $\phi(x)$, which is fixed up to the scaling factor σ . We thus fit a more flexible neural network classifier with a similar strategy of weighting each term in the loss according to the class y_i of the sample. Let $\psi_\theta : \mathbb{R}^{1024} \rightarrow \mathbb{R}^d$ be a multilayer perceptron (MLP) with two hidden layers and parameters θ .¹⁰⁴ We fit a classifier by minimizing the weighted cross-entropy objective:

$$\text{minimize } \sum_{i \in A} w_{y_i} \log(1 + \exp(-\rho(y_i)\psi_\theta(x_i)^T \beta)) \text{ subject to } \theta, \beta \quad (\text{Equation 3})$$

As before, the hyperparameters $w_{\text{ASP}}, w_{\text{NASP}}, w_{\text{NPSH}} > 0$ are learned by cross-validation on a validation set B containing only ASPs and NASPs. We additionally train the hyperparameter for Pfam seeds (PS), $w_{\text{PS}} > 0$ in order to expose the classifier to a larger distribution of training data.

Note that the objectives for the support vector machine and the neural network model are very similar. The former uses a hinge loss instead of a binary cross-entropy or logistic loss, and includes a penalty term $\lambda \|\beta\|^2$. However, the main difference is that in this case we are learning a complex non-linear feature map ψ_θ instead of using the one-parameter feature map ϕ of the Gaussian kernel.

The MLP architecture has three overall layers with two hidden layers. The fixed size input vectors of 1024 dimensions from ProtT5 embeddings are projected to 768 dimensions then to 1536 dimensions to a final classification output layer of 1, representing the classification of SP allergen versus non-allergen. We implemented this model using Pytorch Lightning¹⁰⁵ to create a classifier that included all of the training, validation, and testing steps. Each embedding from the ASP, NASP, NPSH and PS classes was weighted differently, utilizing the PyTorch¹⁰⁴ Binary Cross Entropy with Logits Loss function with the weights specified as a hyperparameter. We used a leaky rectified linear unit (ReLU)¹¹⁵ with dropout regularization and a weight decay in our AdamW optimizer.¹¹⁶

Hyperparameter optimization and training

To optimize our binary SP classification neural network's performance, we used RayTune,⁷⁶ a hyperparameter tuning library. We conducted a randomized grid search, evaluating 20 models concurrently with diverse hyperparameter combinations selected by RayTune (Table S5). Each model was trained for 200 epochs using the Async Successive Halving (ASHA) scheduler,¹¹⁷ which facilitates early stopping by terminating models based on minimizing training loss. After each epoch, we monitored validation accuracy, AUROC, and F1 score (from the sklearn.metrics version 1.4.0 library), with testing accuracy assessed after complete training.

For training purposes we randomly split the ASP and NASP classes into 80% training, 10% validation and 10% test using the random_split function from PyTorch with a fixed generator to ensure reproducibility. We added all the Pfam seeds and NPSH enzymes to the training set. We picked the best performing model, which had the hyperparameters:

Layer 1 Size: 256,
Layer 2 Size: 768
Learning Rate: 1.4e-4,
Batch Size: 128,
Allergenic SP Weight: 6,
Non-Allergenic SP Weight: 1,
Serine Hydrolase Weight: 2,
Pfam Seeds Weight: 1,
Dropout Probability: 24.696%

With that model, we trained the classifier on 20 epochs with validation occurring every 20% of the training and a final test accuracy assessment after completion.

Processing and applying classifiers to catalogs

We downloaded the Unified Human Gastrointestinal Genomes (UHGG) Protein Catalogs clustered at 50% sequence identity from the MGnify⁶⁸ FTP site (https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-gut/v2.0.1/protein_catalogue/). We also downloaded the Human Oral (HO) protein catalog clustered at 95% sequence identity from the MGnify FTP site (https://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-oral/v1.0/protein_catalogue/). After performing a filtering of removing any sequences that were greater than 1000 amino acids we had 4,601,771 sequences in the UHGG protein catalog and 791,924 sequences in the HO protein catalog.

Prioritization of allergen protein sequences

To obtain a probability between 0 and 1 for each protein sequence in the catalog, we applied a sigmoid function to the raw, unnormalized logits (scores) generated by the binary cross-entropy classifier. Any probability that was above 0.5 was classified as an allergen, and below 0.5 was classified as a non-allergen.

We then filtered the original set of classified proteins (10,197 gut proteins and 1,365 oral proteins) by taking everything above 0.85 probability, resulting in 3,874 proteins in the gut catalog and 658 proteins in the oral catalog. For these filtered proteins, we ran

InterProScan version 5.57-90.0⁷⁹ to find specific Pfam⁷⁵ domains in our results. We found 1,744 and 551 proteins with Pfam domain annotations in the gut and oral catalogs, respectively. We ran: `interproscan.sh -i <fasta file> -appl Pfam -iprlookup -cpu 2`

We used the ESMFold⁴⁰ version 1.0.3 command line interface `esm-fold -i <fasta file> -o output_pdb -cpu-offload -chunk-size 32` to predict protein structures of all the filtered proteins. We then used Foldseek⁴³ to perform a structural search of the proteins against the PDB and AlphaFold/Swiss-Prot databases using the `easy-search` command with an e-value set at 1. We merged the Foldseek results and InterProScan results based on a left join on the query. Following the search we devised a priority score metric that is based on the query coverage (*qcov*), the target coverage (*tcov*), the complementary sequence identity ($1 - \text{fident}$), the alignment TM-score (*alnTmscore*). If that filtered protein was found in InterProScan, the e-value from InterProScan was included. Specifically, we performed a negative $\log_{10}$ of the e-value ($-\log_{10}(ev)$) and performed a min-max normalization of the InterProScan e-values only ($ev_{\text{normalized}}$), to achieve values between 0 and 1:

$$\text{E - Value (Min - Max) Normalization : } ev_{\text{normalized}} = \frac{ev - \min(EV)}{\max(EV) - \min(EV)} \quad (\text{Equation 4})$$

Afterwards we took the arithmetic mean of the following values, where z' is the combination of the fractions:

$$\text{Arithmetic Mean : } z' = \frac{qcov + tcov + alnTmscore + (1 - \text{fident}) + ev_{\text{normalized}}}{5} \quad (\text{Equation 5})$$

We then computed a priority score in the form of z-score, where $\bar{z}$ is the mean, and $\sigma_{z'}$ is the standard deviation of the combined fractions. Z is labeled as the Priority Score:

$$Z = \frac{z' - \bar{z}}{\sigma_{z'}} \quad (\text{Equation 6})$$

Each of the protein catalogs was further prioritized based on the following criteria, where p represents the prioritized hits in each protein:

$$p = (qcov > 0.9) \cap (tcov > 0.9) \cap (\text{fident} < 0.35) \cap (\text{alnTmscore} > 0.7) \quad (\text{Equation 7})$$

This resulted in 412 and 147 proteins in the gut and oral protein catalogs. Each plot representing the priority score against the negative $\log_{10}$ of the e-value and the points were saved and rendered in GraphPad Prism version 10.1.1 (270).

The final prioritized lists with the z-score calculations are stored in [Tables S2](#) and [S3](#) for the oral and gut proteins respectively.

Taxonomic assignment and visualization

Each of the MGnify Protein Catalogs contained a metadata file for the Genome Isolate or the Metagenome Assembled Genome (MAG) with a taxonomy lineage that was previously annotated in each catalog from the same FTP site. We used the PhyloT tool (<https://phyloT.biobyte.de/>) to map the taxonomy from the metadata file and used the GTDB Release 202 mapping to provide a tree in Newick format. We visualized the resulting tree in iTOL¹⁰⁷ and represented the leaves at the species level and colored the tree by the phylum level. Any species that was not able to be mapped was discarded during the tree generation step.

Preparation of protein structure pictures

Molecular graphics were created with PyMOL.¹⁰⁶ Initially, a larger set of structural alignments were captured and initially displayed using the built-in display for Foldseek⁴³ web server. The US-Align⁸⁹ tool was used to compute the statistics of the displayed alignments.

DC migration

Internal controls were used in this experiment. C57BL/6 mice were immunized 50 μg papain (Sigma-Aldrich) or 14.2 μg CP (Genscript) or 58 μg SP (Genscript) in the right hind footpad. Mice were immunized with PBS (Sigma-Aldrich) in the left hind footpad. In other experiments, the left dorsal foot received heat inactivated protease (papain, or CP, or SP), while the right dorsal foot received active protease (papain, or CP, or SP). Then 20 hours after immunization,^{19,26} the popliteal LN was harvested and digested at 37 °C in digestion buffer: RPMI supplemented with DNase I (100 $\mu\text{g ml}^{-1}$; Roche), Dispase II (800 $\mu\text{g ml}^{-1}$; Millipore Sigma), Collagenase P (200 $\mu\text{g ml}^{-1}$; Millipore Sigma) and 1% FBS. Every 7 minutes, the digestion buffer was removed and placed into RPMI supplemented with 2 mM EDTA and 1% FBS on ice. Fresh digestion buffer was added to the LNs. Tubes were shaken before and after each addition of fresh digestion buffer. This process was repeated until no fragment of LN tissue remained. Single-cell suspensions were filtered through a 100 μm cell strainer. 2×10^6 LN cells were washed once with FACS Buffer, then cells were incubated in FACS Buffer containing fluorescent-conjugated antibodies targeting surface antigens for 30 minutes at room temperature. Cells were washed twice with DPBS, then assessed by flow cytometry on a Beckman Coulter CytoFLEX S and analyzed using FlowJo Version 10 (TreeStar).

T cell differentiation

C57BL/6 mouse footpads were bilaterally immunized with 50 μg OVA/sham (Sigma-Aldrich), or 50 μg OVA/sham (Sigma-Aldrich) and 14.2 μg CP (Genscript), or 50 μg OVA/sham (Sigma-Aldrich) and 58 μg SP (Genscript). Alternatively, C57BL/6 mice were immunized

with 50 μg OVA/sham in the left hind footpad and 50 μg OVA/sham with 14.2 μg Clostripain from *Clostridium histolyticum* (Thermo Scientific Chemicals), 14.2 μg Endoproteinase Lys-C from *Lysobacter enzymogenes* (MedChemExpress), or 14.2 μg trypsin from porcine pancreas (Sigma-Aldrich) in the right hind footpad. 5 days following immunization,^{19,26} the dLN was collected and dissociated in fully supplemented RPMI using the coarse end of a SuperFrost Plus Microscope Slide and the base of a curved forcep. Single-cell suspensions were filtered through a 100 μm cell strainer. 2×10^6 LN cells were added to a 1.5 ml Eppendorf tube (Fisher Scientific) and resuspended in 1 ml of fully supplemented RPMI with PMA (50 ng ml^{-1} ; Sigma-Aldrich), Ionomycin (1 $\mu\text{g ml}^{-1}$; Sigma-Aldrich) and Protein Transport Inhibitor (1:1,000, BD Biosciences). Samples were incubated for 3 hours in a 37 $^{\circ}\text{C}$ water bath. Samples were washed twice with FACS Buffer, then cells were incubated in FACS Buffer containing fluorescent-conjugated antibodies targeting surface antigens for 30 minutes at room temperature. Following fixation with 4% paraformaldehyde for 8 minutes at room temperature, fluorescent-conjugated antibodies targeting intracellular cytokines were added to each sample in 1 $\times$ eBioscience Permeabilization Buffer and stained overnight. Cells were washed twice with DPBS, then assessed by flow cytometry on a Beckman Coulter CytoFLEX S and analyzed using FlowJo Version 10 (TreeStar).

T cell in vitro stimulation

As previously described,^{19,26} tissue-culture-treated flat 96-well plates (Corning) were incubated with 50 μl of purified αCD3 (2 $\mu\text{g ml}^{-1}$, BioLegend) for 2 hours at 37 $^{\circ}\text{C}$, then washed three times with DPBS. Following DPBS removal, 125,000 cells in 100 μl of fully supplemented RPMI from mouse LNs as immunized in 'T cell Differentiation' were plated. Then, 100 μl of anti-CD28 (4 $\mu\text{g ml}^{-1}$, BioLegend) in fully supplemented RPMI was added. After 4 days, the cell-free supernatant was harvested, snap-frozen and stored at -80°C .

Cytokine enzyme-linked immunosorbent assay (ELISA)

Cell-free culture supernatant from stimulated T cells, was collected as described in the 'T cell *in vitro* stimulation' protocol. High binding 96 well plates (Corning) were coated with capture antibodies overnight at 4 $^{\circ}\text{C}$. IL-4, IL-13, and IFN- γ were captured with anti-mouse IL-4 (0.25 $\mu\text{g ml}^{-1}$, BioLegend), anti-mouse IL-13 (1.0 $\mu\text{g ml}^{-1}$, BioLegend), anti-mouse IFN- γ (0.5 $\mu\text{g ml}^{-1}$, eBioscience). Then, plates were washed three times with 0.1% Tween 20 in PBS and blocked with 1% (wt/vol) BSA for 4 hours at room temperature before cell-free culture supernatant was plated overnight at 4 $^{\circ}\text{C}$. Next, plates were washed three times with 0.1% Tween 20 in PBS. IL-4, IL-13, and IFN- γ were detected with biotinylated anti-mouse IL-4 (0.25 $\mu\text{g ml}^{-1}$, BioLegend), biotinylated anti-mouse IL-13 (0.25 $\mu\text{g ml}^{-1}$, eBioscience), biotinylated anti-mouse IFN- γ (1.5 $\mu\text{g ml}^{-1}$, BioLegend) for 2 hours at room temperature. Following, plates were washed four times, incubated with Streptavidin-horseradish peroxidase (BioLegend) for 30 minutes at room temperature, washed six times, and finally incubated with OPD solution (Sigma-Aldrich) for development. Development was stopped with 2M H_2SO_4 . ELISA samples were assessed on a SpectraMax iD5 microplate reader (Molecular Devices). Concentrations were determined from a standard curve.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis of results in Figure 6 were performed using GraphPad Prism version 10.1.1. Two-way ANOVA with Šídák's multiple comparisons test, ordinary one-way ANOVA with multiple comparisons, two-sided unpaired Welch' t-test, or two-sided paired t-test were applied as indicated in figure legends. Statistical significance was considered for p-values < 0.05. Details are listed in the respective figure legends.