OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

AI and Bioinformatics Driven Identification of CRISPR guide RNAs for Allergy Associated Gene Targets

Dhyan Murthy
dhyan.k.murthy@googlemail.com

ABSTRACT

Allergic diseases affect a substantial and growing share of the global population, yet the identification of reliable genetic targets and the development of precise therapeutic strategies remain limited. Here we integrate gene expression profiling, network-based analyses, and machine learning to prioritize candidate genetic determinants of allergic responses. Differentially expressed genes (DEGs) were identified from NCBI GEO datasets and characterised through interaction network analysis (STRING, Cytoscape) and pathway enrichment (Reactome). Genome-wide association study (GWAS) data further refined target selection. Four classifiers integrating expression, GWAS, pathway and network features were used to rank candidate genes. Under stratified cross-validation none of them discriminated above chance (ROC AUC 0.49–0.53 against a positive-class base rate of 0.204), so the ranking is reported as a transparent weighting of pre-specified evidence rather than as a predictive model. The six highest-ranked genes were taken forward for CRISPR guide RNA design. Guide RNAs were designed with CHOPCHOP and assessed computationally for predicted specificity and off-target effects using CRISPOR, yielding candidates with high predicted efficiency and low predicted off-target risk. This approach narrows a large candidate space to a small set of genes and gRNAs that can be taken into the laboratory, and documents the thresholds by which that narrowing was done. All findings are computational; no experimental validation was performed, and no causal or therapeutic role for any candidate is established here.

Keywords: *Computational Biology and Bioinformatics, Computational Biomodelling, Allergies, CRISPR-Cas9, Network biology*

INTRODUCTION

Allergic diseases are among the most common chronic conditions worldwide. The Global Burden of Disease Study 2021 estimated that 260 million people of all ages were living with asthma, and although age-standardized prevalence rates for asthma and atopic dermatitis have declined since 1990, the total number of affected individuals is projected to continue rising through 2050.[1] The mechanisms underlying these conditions and their implications for treatment have been reviewed in detail.[2] In India, allergic rhinitis affects 20–30% of the population, contributing to reduced quality of life and increased healthcare costs.[3] Infants and young children under the age of five are particularly susceptible, with asthma and

atopic dermatitis among the most frequent allergic conditions in this group.[4] These trends are expected to intensify with rising air pollution and climate change, which exacerbate allergic sensitization and disease burden.[5] Allergic diseases, including asthma, allergic rhinitis, and atopic dermatitis, are characterized by immunoglobulin E (IgE)-mediated inflammatory responses, producing symptoms such as nasal congestion, sneezing, and pruritus.

Epidemiological and genetic studies confirm that allergy prevalence is increasing worldwide.[6] Genome-wide analyses have identified several candidate genes associated with allergic responses, yet important knowledge gaps remain regarding their interactions with environmental triggers.[7] For instance, analysis of 281,104 UK Biobank samples revealed 354 exome variants linked to asthma, allergic rhinitis, and atopic dermatitis, clustering primarily on chromosomes 6 and 17.[8] However, many SNP-based associations have not been replicated, highlighting the complexity of the mutational spectrum underlying allergic diseases. Similarly, transcriptomic studies using microarray data have identified differentially expressed genes (DEGs) in allergic rhinitis, uncovering key gene signatures, but the reliance on single datasets limits generalizability across populations.[9]

Recent advances in genome editing and artificial intelligence offer opportunities to overcome these challenges. CRISPR-Cas9 has shown promise in allergen gene editing, exemplified by studies targeting the major cat allergen *Fel d 1*, which achieved editing efficiencies exceeding 90% with customizable single-guide RNAs and multi-site editing.[10] However, concerns about off-target mutations and immune responses emphasize the need for improved specificity and safety. Complementary machine learning approaches have been applied to allergy and gene-editing research. PreAcrs, an ensemble model integrating support vector machines and logistic regression, has accurately predicted anti-CRISPR proteins,[11] while iAller, a random forest model, classified allergenic proteins with 91.4% accuracy.[12] Both approaches demonstrate potential, but challenges with dataset quality and interpretability limit their translational utility.

CRISPR has also been explored for direct therapeutic applications. Gene editing has been investigated in peanut allergy and asthma models, combining CRISPR-Cas9 with short hairpin RNA (shRNA) to reduce allergenicity.[13] Such studies underscore the transformative potential of gene editing for hypoallergenic food production and targeted therapies, though barriers such as off-target effects, ethical considerations, and delivery strategies persist. Biomarker discovery also remains crucial. For example, Wang et al. (2020)[14] identified elevated CYP11A1 enzyme levels in $CD4^+$ T cells of allergic children, suggesting a potential therapeutic target, though larger and more diverse validation cohorts are needed. Together, these findings reinforce the importance of bioinformatics in deepening our understanding of the genetic architecture of allergies and advancing precision medicine.

Despite these advances, critical gaps remain. The polygenic nature of allergic diseases, coupled with complex gene–environment interactions, limits the predictive power of existing genetic studies. Furthermore, while bioinformatics has identified numerous candidate biomarkers, translating these findings into clinically actionable targets remains a significant challenge. In this study we integrate publicly available transcriptomic and genomic datasets to prioritize candidate genes associated with allergic responses, rank them using machine learning, and design candidate CRISPR guide RNAs in silico

with computational off-target assessment. The aim is narrow: to reduce a large candidate space to a small, documented set of genes and guide RNAs that could be tested experimentally. This is a computational prioritization study and does not assess the feasibility of CRISPR-based interventions as preventive or therapeutic strategies, which would require laboratory work beyond its scope.

METHODS

Data Collection

Gene expression datasets relevant to allergic conditions were obtained from the NCBI Gene Expression Omnibus (GEO) under accession numbers GSE169376 and GSE245608. Single nucleotide polymorphism (SNP) data were retrieved from the GWAS Catalog for allergy-associated traits (MONDO_0005271 and EFO_0010638) to incorporate genetic variant information and corroborate expression-based findings.

Differential expression analysis

Differentially expressed genes (DEGs) were identified using GEO2R. Test groups (allergic samples) and control groups (non-allergic samples) were defined for each dataset. Genes with adjusted $p < 0.05$ and $|\log_2 FC| > 1.5$ were considered significant. Volcano plots were generated for visualization. To derive a non-redundant gene set, DEGs from both datasets were concatenated, yielding a consolidated list of allergy-associated genes.

Network construction and analysis

Protein–protein interaction (PPI) networks were constructed using STRING, restricted to high-confidence interactions (interaction score > 0.7). Networks were imported into Cytoscape for topological analysis using the "Analyze Network" tool. Parameters included degree, betweenness centrality, closeness centrality, and clustering coefficient. Hub genes were defined based on centrality metrics.

Pathway enrichment analysis

To evaluate functional relevance, pathway enrichment analysis was performed using the Reactome database. The DEG network was uploaded, and enriched pathways were visualized. The top 10 hub genes were further examined for their occurrence across the 25 most enriched pathways.

Integration of GWAS-SNP data and dataset construction for AI models

Significant DEGs were mapped to SNP data from the GWAS Catalog to identify SNP-associated genes. A multi-feature dataset was constructed integrating: (i) $\log_2$ fold-change values from differential expression, (ii) combined centrality scores from network analysis, (iii) pathway involvement scores from Reactome enrichment, and (iv) GWAS-derived p-values and odds ratios. This dataset formed the basis for machine learning–based gene prioritization.

Machine learning model development

The unit of analysis was the gene. The dataset comprised the 1,000 genes obtained by mapping differentially expressed genes to GWAS Catalog entries for the allergy trait terms MONDO_0005271 and EFO_0010638 (retrieved 7 April 2025). The target variable, Known_Allergy_Association, was assigned as 1 where a gene had at least one qualifying GWAS Catalog entry at genome-wide significance ($p \leq 5\times10^{-8}$) and 0 otherwise. This produced 204 positive and 796 negative genes, a base rate of 0.204 and a majority-class share of 0.796. The dataset included four predictive features (Expression_Level, GWAS_Significance, Pathway_Involvement_Score, Network_Centrality), with *Known_Allergy_Association* as the target variable. Because the target was derived from the GWAS Catalog and the GWAS_Significance feature is drawn from the same source, the two are not independent by construction. This was tested directly by refitting the Random Forest with that feature removed: cross-validated ROC AUC was 0.53 ± 0.057 with the feature included and 0.46 ± 0.032 with it excluded. Both intervals span 0.50, indicating that the feature contributed no detectable discriminative information and that model performance was not inflated by dependence between feature and target.Four models were trained: Random Forest (RF), Gradient Boosting (XGB), Support Vector Machine (SVM), and Deep Neural Network (DNN). The original analysis used a single 80%/20% train–test split evaluated by overall accuracy. Because the classes are imbalanced, this protocol was replaced with stratified 5-fold cross-validation, and performance is reported as ROC AUC, precision–recall AUC, balanced accuracy and positive-class F1, each as the mean across folds. A majority-class baseline classifier was evaluated under an identical protocol so that model performance can be read against the performance obtainable without any predictive information. Candidate prioritization ranked all 1,000 genes by their out-of-fold predicted probability of a positive label, and the six highest-ranked genes were taken forward to gRNA design. The number six was chosen by the authors as a practical limit on how many targets could be carried into guide design; it is not a threshold derived from the data, and a different cut-off would yield a different candidate set. The top-10 network hub list reported in Table 1 was not used to gate this selection; network centrality enters the procedure only as one of the four input features. To establish whether the observed performance depended on the class balance of the dataset, the analysis was repeated on versions of the dataset resampled to base rates of 0.10, 0.21, 0.40 and 0.65.

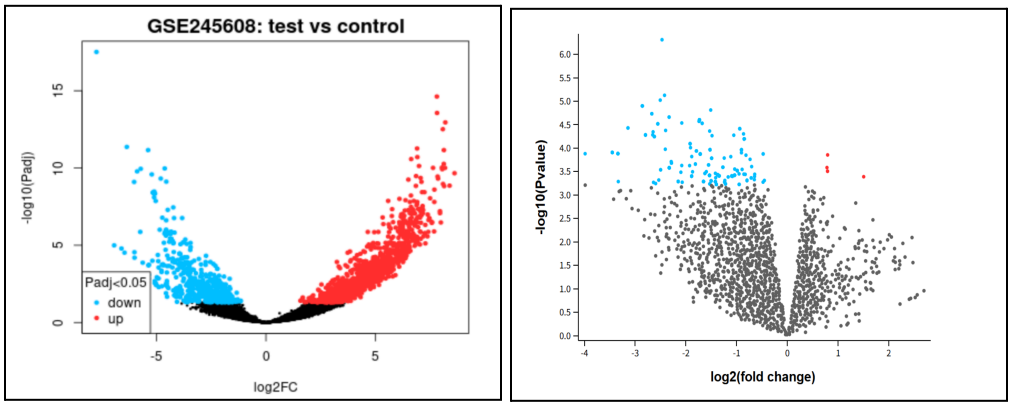
CRISPR guide RNA design and computational assessment

The top-prioritized genes were subjected to CRISPR guide RNA (gRNA) prediction using CHOPCHOP. Candidate gRNAs were filtered by efficiency (>0.40) and mismatch tolerance (≤3 mismatches, maximum of five gRNAs per gene). Selected gRNAs were further evaluated using CRISPOR to compute MIT and CFD scores, which estimate specificity and off-target potential from sequence context. Both tools generate computational predictions; no experimental assay of editing activity or specificity was performed at any stage, and the two-tool procedure is therefore described here as a dual computational assessment rather than as validation.

RESULTS

To identify genetic targets associated with allergic conditions, we analyzed two publicly available transcriptomic datasets (GSE169376 and GSE245608) comprising 41 samples in total. GSE169376 contributed 9 samples (6 allergic, 3 non-allergic control) profiled from nasal mucosa on platform

GPL21572, and GSE245608 contributed 32 samples (20 allergic, 12 non-allergic control) profiled from skin tape-strips of children with allergic asthma on platform GPL24014. Because the two series differ in tissue of origin, patient population and experimental design, differential expression was performed separately within each dataset in GEO2R and the resulting DEG lists were concatenated; the expression matrices were not merged and no cross-dataset batch correction was applied. Effect sizes are therefore not directly comparable between the two datasets, and the consolidated gene list should be read as the union of two independent analyses rather than as a single integrated cohort. Differential expression analysis (adjusted $p < 0.05$, $|\log_2 FC| > 1.5$) identified 5,938 unique differentially expressed genes (DEGs) across the datasets (Figure 1). This consolidated gene set formed the basis for downstream analyses.



(A) (B)

Figure 1. Volcano plots of differential gene expression for GSE245608 (A) and GSE169376 (B), generated in GEO2R and plotted from the GEO2R output. Each plot compares allergic samples with non-allergic control samples. The x-axis represents the log2 fold change (log2FC) in gene expression, while the y-axis shows the negative log10 of the adjusted p-value (-log10(Padj)). The DEGs which are under expressed, with a log2 fold change less than 0, are marked with blue, while DEGs which are overexpressed are marked with red. The DEGs with no significance are in black.

Network construction and hub gene identification

A protein–protein interaction (PPI) network constructed using STRING revealed 7,900 high-confidence interactions (score > 0.7) among 1,304 nodes (Figure 2). Centrality analysis in Cytoscape highlighted highly connected hub genes, including *UBA52, UBB, HSPA8, SRC,* and *RPS2* (Table 1). These hub genes are highly connected within the network. Network centrality describes topological position and is reported here as a structural property of the interaction network; it was not used as a selection filter for the candidate genes taken forward to gRNA design, which were prioritized separately as described below.

S.No.	Name	Betweenness Centrality	Closeness Centrality	Clustering Coefficient	Degree
1	UBA52	0.07414765915	0.4175895765	0.1006622517	151
2	UBB	0.05576932235	0.41024	0.09039952296	130
3	HSPA8	0.05878068347	0.4073721004	0.1169955002	108
4	SRC	0.07314922113	0.4023854363	0.06853038245	102
5	RPS2	0.02301863259	0.3854479856	0.2410087719	96
6	EPRS1	0.0184219204	0.3803025808	0.2416666667	81
7	RPL13A	0.01026950646	0.3772807534	0.3106630212	77
8	RPS3A	0.008644861768	0.3740881237	0.3098245614	76
9	RPL23A	0.009096468349	0.3681792074	0.3037837838	75
10	HSPA9	0.02466895666	0.3759530792	0.1532765642	74

Table 1. Top 10 Hub Genes after Network Analysis with highest centrality measures

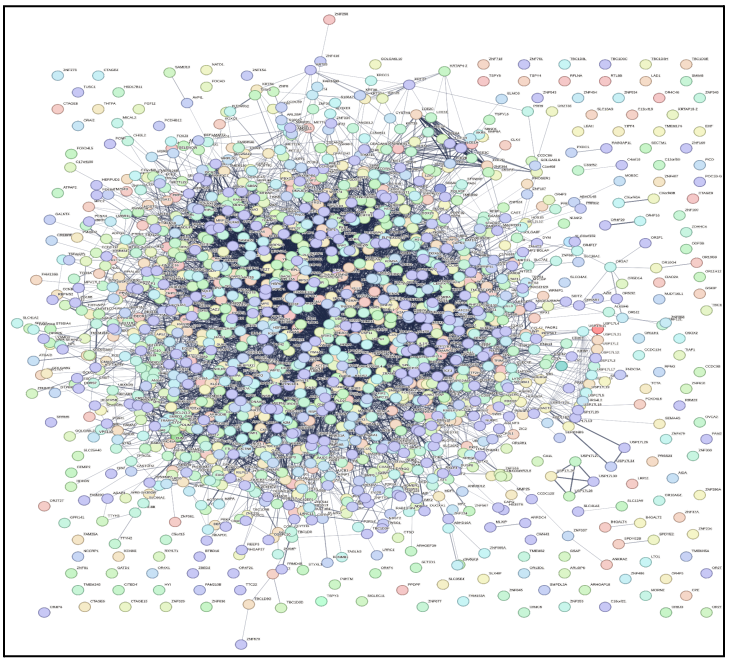


Figure 2. Protein–protein interaction (PPI) network of the consolidated DEG set, generated in STRING at a confidence threshold of 0.7 and visualised in Cytoscape. Nodes are proteins and edges are predicted or reported interactions. Node colours reflect default STRING/Cytoscape visual styling and were not generated by a clustering algorithm such as MCODE; they provide visual differentiation only and do not denote functional clusters or any other biologically meaningful grouping.

Pathway enrichment analysis

Reactome analysis linked hub genes to key biological pathways such as cornified envelope formation, deubiquitination, and interferon α/β signaling, processes known to regulate immune activation and inflammatory responses. All 10 hub genes mapped to one or more of the 25 most enriched pathways. This overlap reflects statistical enrichment and the high connectivity of these genes rather than evidence of a functional or causal role in allergic disease. Six of the hubs (UBA52, UBB, RPS2, RPL13A, RPS3A and RPL23A) are ubiquitin or ribosomal protein genes, which are highly connected in protein–protein interaction networks and recurrently enriched across pathway databases irrespective of disease context; their appearance here should not be interpreted as allergy-specific.

Integration with GWAS data and AI-driven prioritization

Mapping DEGs to GWAS catalog entries identified 1,000 SNP-associated genes, which were integrated with expression, pathway, and network features to construct a gene prioritization dataset. Machine learning models were applied to the resulting dataset, in which 204 of the 1,000 genes (20.4%) carried a positive label. The overall accuracy of approximately 81% reported in the original analysis is close to the

79.6% obtainable by assigning the majority class to every gene, and the identical accuracy, precision and recall values returned by three structurally different classifiers are consistent with models that assign the majority class to nearly all samples. Accuracy is therefore not an informative measure on this dataset and is not reported below. Under stratified 5-fold cross-validation all four classifiers performed at or near chance (Table 2): ROC AUC ranged from 0.49 ± 0.032 for the SVM to 0.53 ± 0.057 for the Random Forest, balanced accuracy from 0.50 to 0.52, and precision–recall AUC from 0.21 to 0.23 against a base rate of 0.204. The ensemble classifier reached a ROC AUC of 0.56 (Figure 4). No model exceeded the others meaningfully, indicating that the limitation lies in the features rather than in the choice of algorithm. The sensitivity analysis (Table 3) shows that this is not a consequence of class imbalance alone: across base rates from 0.10 to 0.65, ROC AUC remained between 0.47 and 0.56 and balanced accuracy between 0.48 and 0.55, while precision–recall AUC tracked the base rate at every level, which is the behaviour expected of a non-informative classifier. The four features as constructed therefore carry no detectable signal regarding the target at any class balance tested, and the models should be understood as integrating pre-specified evidence rather than as discovering new associations. Feature importances (Figure 3) rank the contribution of features within a model whose overall discrimination is close to chance and should not be read as evidence of biological importance. Ranking all 1,000 genes by out-of-fold predicted probability and taking the six highest-ranked, all six carry a positive label. Under random selection at the observed base rate of 0.204 the probability of six such draws all being positive is approximately 7×10^{-5}, so the ranking is informative in its extreme upper tail even though overall discrimination is at chance (Table 2). This is not a contradiction: ROC AUC averages performance over all decision thresholds and can remain near 0.50 while the few highest-scoring cases are accurate. The estimate rests on six observations and is correspondingly imprecise, and because six was chosen by the authors rather than derived from the data, it describes the top of the ranking rather than establishing a usable operating point. The six candidates are accordingly best understood as the genes scoring highest on pre-specified, reproducible evidence — differential expression, pathway involvement and network centrality, combined with GWAS significance — rather than as the output of a validated predictive model. The machine-learning ranking is reported for transparency and is not offered as independent validation of these genes. The six genes are *EPRS1, RPL23A, FAU, SIRT1, H2AC20,* and *MRPS10*.

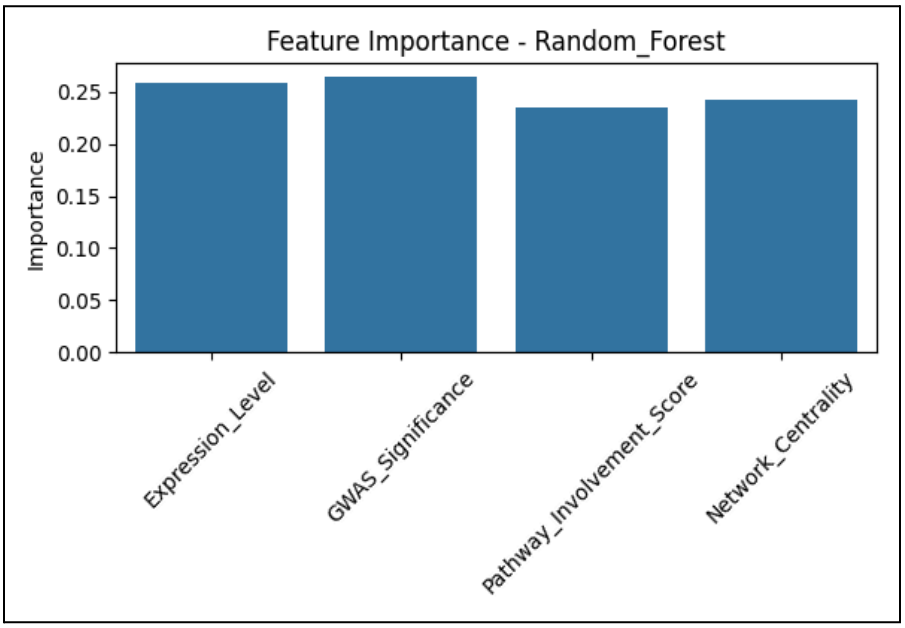


Figure 3. Random Forest feature importances, generated from the fitted model. Bars show the four input features (Expression Level, GWAS Significance, Pathway Involvement Score, Network Centrality), with GWAS Significance ranked highest.

Model	**ROC AUC**	**PR AUC**	**Balanced accuracy**	**Positive-class F1**
Random Forest	0.53 ± 0.057	0.23 ± 0.041	0.52 ± 0.044	0.29 ± 0.048
Gradient Boosting	0.51 ± 0.048	0.22 ± 0.036	0.50 ± 0.031	0.28 ± 0.037
SVM	0.49 ± 0.032	0.21 ± 0.024	0.50 ± 0.021	0.28 ± 0.024
Neural Network	0.52 ± 0.061	0.22 ± 0.045	0.51 ± 0.039	0.28 ± 0.043

Table 2. Cross-validated performance of each classifier on the full dataset (204 positives, 796 negatives). Values are mean ± standard deviation across stratified 5-fold cross-validation. Positive-class precision and recall were 0.20-0.21 and 0.49-0.51 respectively for all four models. The positive-class base rate is 0.204, which is the precision obtainable without any predictive information.

Base rate	ROC AUC	PR AUC	Balanced accuracy	Positive-class F1
0.10	0.56	0.10	0.53	0.15
0.21	0.51	0.21	0.48	0.30
0.40	0.47	0.40	0.55	0.50
0.65	0.53	0.65	0.55	0.72

Table 3. Sensitivity analysis of classifier performance across class balances. The dataset was resampled to the base rates shown; the 0.21 row is the resampled condition closest to the base rate of the actual dataset, which is 0.204 (204 positives of 1,000 genes); no row reports the unmodified dataset. Each resampled dataset was evaluated using repeated random train–test splits, and values are reported as proportions, averaged across 500 repetitions. This differs from the stratified 5-fold cross-validation used for the full-dataset results in Table 2, so values are comparable in magnitude between the two tables but were not produced by an identical protocol. Overall accuracy is omitted because, at the observed base rate, a majority-class classifier achieves 0.796 without using any predictive information.

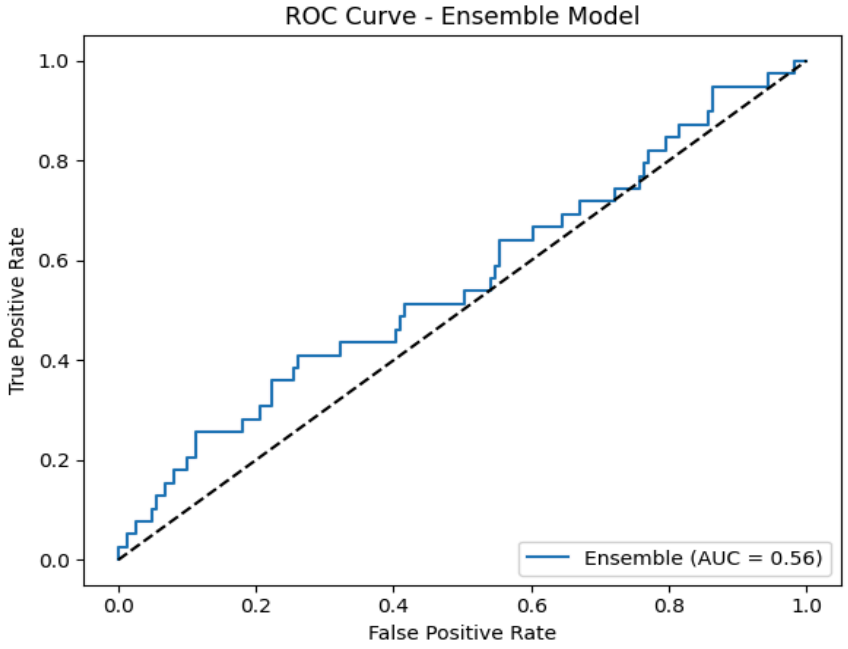


Figure 4. Receiver operating characteristic (ROC) curve for the ensemble classifier, generated from the held-out test predictions. The area under the curve (AUC) is 0.56, indicating discrimination close to chance.

CRISPR guide RNA prediction and computational assessment

For these six priority genes, CHOPCHOP predicted 232 candidate gRNAs. After filtering for efficiency (>0.40) and mismatch tolerance (≤3), 30 gRNAs were shortlisted. CRISPOR assessment returned high MIT and CFD scores, indicating strong predicted specificity and low predicted off-target potential (Table 4). The six gRNAs reported in Table 4 were selected as the highest-scoring guide for each gene from among these 30 candidates, which had already passed the efficiency and mismatch-tolerance filters; the uniformly high predicted efficiency and low predicted off-target risk across the six therefore reflect this prior filtering rather than a threshold applied after the fact. Both CHOPCHOP and CRISPOR scores are sequence-based predictions, so agreement between them indicates consistency between two computational methods rather than confirmation of activity; the six gRNAs are reported as candidates for experimental testing.

S.No.	Gene Name	Guide RNA (20 nt protospacer + PAM)	CHOPCHOP Efficiency (%)	Predicted Efficiency (CHOPCHOP)	CRISPOR MIT Score	CRISPOR CFD Score	Predicted Off-Target Risk (CRISPOR)
1	EPRS1	GTAGCCGTCAGACGAGATACTGG	87%	High	82	0.91	Low
2	RPL23A	GCGCCGAAAGCGAAGAAGGAAGG	84%	High	78	0.88	Low
3	SIRT1	TAGCAAGCGGTTCATCAGCTGGG	91%	High	86	0.94	Low
4	H2AC20	AGGCAACTACGCGGAGCGGGTGG	79%	High	73	0.81	Low
5	FAU	GAGCACGACTTGATCTTCCGGGG	88%	High	89	0.95	Low
6	MRPS10	AGGTCACGATAAGGCTGTATTGG	82%	High	76	0.85	Low

Table 4. Predicted on-target efficiency and predicted off-target risk for the selected candidate gRNAs. Sequences are shown as the 20 nt protospacer followed by the NGG PAM. Guide RNAs were categorized using predefined study-specific thresholds: a CHOPCHOP predicted efficiency of ≥70% was considered high, and a CRISPOR MIT specificity score of ≥60 was considered indicative of lower predicted off-target risk. CRISPOR CFD scores were evaluated alongside the predicted off-target sites rather than interpreted against a universal numerical cutoff. All values are computational predictions and have not been experimentally verified.

Our integrative approach prioritizes six candidate genes for CRISPR-based investigation in allergic diseases. These are not the same set as the network hub genes reported in Table 1: only EPRS1 and RPL23A appear in both, and FAU, SIRT1, H2AC20 and MRPS10 were prioritized by the evidence-integration step alone. The two analyses are reported here as parallel lines of evidence rather than as a single selection funnel, and the six candidates are therefore described as prioritized genes rather than as hub genes. By combining differential expression, network topology, pathway enrichment, GWAS integration and evidence integration across these sources, this study brings several complementary analyses to bear on a problem more often approached through a single modality. Genes such as SIRT1 and

EPRS1 have been implicated in immune regulation, which offers some independent support for their candidacy, although that support comes from the existing literature rather than from the present analysis.

The machine-learning component did not achieve discrimination above chance on this dataset. Cross-validated ROC AUC remained between 0.47 and 0.56 across all class balances tested, and precision–recall AUC tracked the base rate throughout, indicating that the four features as constructed carry no detectable information about the target. The most likely explanation lies in feature construction rather than in model choice or class imbalance: the features summarise evidence at a coarse per-gene level, and the target is derived from the same GWAS Catalog data as one of them without the two being informative about each other. The component should therefore be understood as a transparent weighting of pre-specified evidence rather than as a predictive model, and richer feature representations would be required before machine learning could contribute independent discriminative power to gene prioritization in this setting.

CRISPR-Cas9 represents a promising strategy for allergen-targeted interventions, though concerns remain regarding off-target effects and safety. Our pipeline combines CHOPCHOP and CRISPOR to assess these properties computationally, yielding gRNAs with high predicted efficiency and low predicted off-target risk. Agreement between two prediction tools does not address the underlying concern, since both infer activity from sequence context and neither measures editing in cells; the predictions identify candidates worth testing rather than reducing the risk itself.

This study also underscores the novelty of applying an integrated bioinformatics and CRISPR design pipeline to human allergy datasets. Previous research has largely focused on allergen modification in plants or animal models[10] [15] or on preliminary biomarker discovery. Here we apply the pipeline to human allergy datasets, prioritizing candidate genes from human transcriptomic and genomic data and proposing corresponding gRNAs as starting points for experimental evaluation.

Limitations include reliance on publicly available datasets, which may introduce population and tissue-specific biases, and the absence of experimental validation, which remains essential to confirm in silico predictions. The two source datasets are a particular constraint in this respect. GSE169376 profiles nasal mucosa and GSE245608 profiles skin tape-strips from children with allergic asthma, so a gene called differentially expressed in one is not interchangeable evidence with a gene called in the other; the consolidated list should not be read as identifying genes implicated across allergic tissues. GSE169376 also contributes only three control samples, which makes its adjusted p-values unstable and its fold-change estimates noisy, so the differential expression calls derived from it are less reliable than those from the larger dataset and carry correspondingly less weight in the downstream prioritization.Further limitations apply specifically to the gRNA component. No experimental validation of any kind was performed: editing efficiency, indel spectra, specificity, cell viability and downstream function are all unmeasured, and CHOPCHOP and CRISPOR infer cutting from sequence context rather than measuring activity in cells. MIT and CFD are alignment-based heuristics that do not account for chromatin accessibility or cell-type-specific activity and are known to miss off-target sites detectable only by empirical assays such as GUIDE-seq or CIRCLE-seq, so a low predicted off-target score is not evidence of specificity. The composition of the candidate set is itself a constraint: four of the six genes (RPL23A,

FAU, H2AC20 and MRPS10) encode ribosomal, histone or mitochondrial ribosomal proteins with essential housekeeping roles, and disrupting them would be expected to affect cell viability broadly rather than allergic signalling specifically. Allergic disease also involves airway epithelium, skin and several immune lineages, and no delivery vehicle with the tissue restriction this would require is proposed or evaluated here; systemic or germline editing raises safety and ethical considerations beyond the scope of this study. Any therapeutic application would require, at minimum, editing in relevant primary cells, empirical off-target profiling, functional assays of allergic readouts and animal models before clinical work could be considered. Within those limits, the integration of transcriptomic, genomic and machine-learning analyses provides a documented and reproducible route from a large candidate space to a small set of testable genes and gRNAs.

CONCLUSION

This study integrated transcriptomic profiling, network analysis, pathway enrichment, GWAS data and machine learning to prioritize candidate genetic targets for allergic diseases. Six genes — EPRS1, RPL23A, FAU, SIRT1, H2AC20 and MRPS10 — were ranked highest by the integrated evidence score and are proposed as candidates for experimental investigation; of these, EPRS1 and RPL23A also appear among the network hub genes. No causal or regulatory role in allergic disease is established by this analysis. By incorporating CRISPR gRNA design and dual-layer computational specificity assessment, we further present a set of gRNAs with high predicted efficiency and low predicted off-target risk that can serve as starting points for experimental evaluation.

This work is a computational prioritization study. It does not bridge the gap between prediction and application, but it narrows a large candidate space to a small set of genes and gRNAs that can be taken into the laboratory, and it documents the reasoning and the thresholds by which that narrowing was done. Whether any of these candidates has a functional role in allergic disease, and whether the proposed gRNAs edit their targets efficiently and specifically in relevant cells, are open questions that only experimental work can answer.

ABBREVIATIONS:

AI (Artificial Intelligence), **AUC** (Area Under the Curve), **CFD** (Cutting Frequency Determination, an off-target scoring metric in CRISPOR), **CRISPR** (Clustered Regularly Interspaced Short Palindromic Repeats), **CRISPR-Cas9** (CRISPR-associated protein 9), **DEG(s)** (Differentially Expressed Gene(s)), **DNN** (Deep Neural Network), **FC/log$_2$FC** (Fold Change/log$_2$ Fold Change), **gRNA(s)** (guide RNA(s)), **GWAS** (Genome-Wide Association Study), **IgE** (Immunoglobulin E), **MIT score** (Massachusetts Institute of Technology specificity score for gRNA off-target analysis), **PPI** (Protein–Protein Interaction), **RF** (Random Forest), **ROC** (Receiver Operating Characteristic), **shRNA** (Short Hairpin RNA), **SNP(s)** (Single Nucleotide Polymorphism(s)), **SVM** (Support Vector Machine), **WAO** (World Allergy Organization), and **XGB** (Extreme Gradient Boosting).

REFERENCES

1. GBD 2021 Asthma and Allergic Diseases Collaborators, Global, regional, and national burden of asthma and atopic dermatitis, 1990-2021, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. Lancet Respir Med 2025, 13 (5), 425-446.
2. Wang, J.; Zhou, Y.; Zhang, H.; Hu, L.; Liu, J.; Wang, L.; Wang, T.; Zhang, H.; Cong, L.; Wang, Q., Pathogenesis of allergic diseases and implications for therapeutic interventions. Signal Transduct Target Ther 2023, 8 (1), 138.
3. Mahesh, P. A.; Moitra, S.; Mabalirajan, U.; Garg, M.; Malamardi, S.; Vedanthan, P. K.; Christopher, D. J.; Agrawal, A.; Krishna, Mamidipudi T., Allergic diseases in India – Prevalence, risk factors and current challenges. Clinical & Experimental Allergy 2023, 53 (3), 276-294.
4. Lv, J. J.; Kong, X. M.; Zhao, Y.; Li, X. Y.; Guo, Z. L.; Zhang, Y. J.; Cheng, Z. H., Global, regional and national epidemiology of allergic disorders in children from 1990 to 2019: findings from the Global Burden of Disease study 2019. BMJ Open 2024, 14 (4), e080612.
5. Reinmuth-Selzle, K.; Kampf, C. J.; Lucas, K.; Lang-Yona, N.; Fröhlich-Nowoisky, J.; Shiraiwa, M.; et al., Air Pollution and Climate Change Effects on Allergies in the Anthropocene: Abundance, Interaction, and Modification of Allergens and Adjuvants. Environ Sci Technol 2017, 51 (8), 4119-4141.
6. Pawankar, R., Allergic diseases and asthma: a global public health concern and a call to action. World Allergy Organ J 2014, 7 (1), 12.
7. Moffatt, M. F.; Gut, I. G.; Demenais, F.; Strachan, D. P.; Bouzigon, E.; Heath, S.; von Mutius, E.; Farrall, M.; Lathrop, M.; Cookson, W., A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010, 363 (13), 1211-1221.
8. Wjst, M., Exome variants associated with asthma and allergy. Sci Rep 2022, 12 (1), 21028.
9. Zhang, Y.; Huang, Y.; Chen, W. X.; Xu, Z. M., Identification of key genes in allergic rhinitis by bioinformatics analysis. J Int Med Res 2021, 49 (7), 03000605211029521.
10. Brackett, N. F.; Pomés, A.; Chapman, M. D., New Frontiers: Precise Editing of Allergen Genes Using CRISPR. Front Allergy 2021, 2, 821107.
11. Zhu, L.; Wang, X.; Li, F.; Song, J., PreAcrs: a machine learning framework for identifying anti-CRISPR proteins. BMC Bioinformatics 2022, 23 (1), 444.
12. Liu, B.; Yang, Z.; Liu, Q.; Zhang, Y.; Ding, H.; Lai, H.; Li, Q., Computational prediction of allergenic proteins based on multi-feature fusion. Front Genet 2023, 14, 1294159.
13. Dileepan, M.; Ha, S. G.; Rastle-Simpson, S.; et al., Pulmonary delivery of ORMDL3 short hairpin RNA – a potential tool to regulate allergen-induced airway inflammation. Exp Lung Res 2020, 46 (7).
14. Wang, M.; Strand, M. J.; Lanser, B. J.; Santos, C.; Bendelja, K.; Fish, J.; Esterl, E. A.; Ashino, S.; Abbott, J. K.; Knight, V.; Gelfand, E. W., Expression and activation of the steroidogenic enzyme CYP11A1 is associated with IL-13 production in T cells from peanut allergic children. PLoS One 2020, 15 (6), e0233563.

15. Biswas, S.; Wahl, N. J.; Thomson, M. J.; Cason, J. M.; McCutchen, B. F.; Septiningsih, E. M., Optimization of Protoplast Isolation and Transformation for a Pilot Study of Genome Editing in Peanut by Targeting the Allergen Gene Ara h 2. Int J Mol Sci 2022, 23 (2), 837.