WebLog Journal of Allergy

Research Article
Published: 05 Nov, 2026

Sequence Similarity, Deduplication and Feature Choice in the Evaluation of Allergen Protein Classifiers

Binod Bhattarai[1], Dipesh Gurung[2] and R N Thakur[2]*

[1]Department of Computer Science, Islington College, Kathmandu, Nepal

[2]Lord Buddha Education Foundation, Kathmandu, Nepal

WebLog Open Access Publications
Article ID : wja.2026.k0502
Author : Dipesh Gurung

Abstract

Background: Allergen protein classifiers are often evaluated on benchmarks built from overlapping sources, where duplicates and homologues can inflate performance.

Methods: We conducted a computational robustness and reproducibility study of the public AlgPred 2 dataset (repository commit 533225b). Exact-sequence auditing reduced 20,150 records to 13,294 unique sequences; deduplicating the training files left 10,514 development sequences (2,454 allergens and 8,060 putative non-allergens). Logistic regression and random forest classifiers using amino-acid composition and log length (feature set A) or five added physicochemical descriptors (feature set B) were evaluated under random splitting and MMseqs2 sequence groups at 30%, 40% and 60% identity with 80% bidirectional coverage, with three seeds and nested tuning.

Results: Random forest Matthews Correlation Coefficient (MCC) was 0.84-0.85 under random splitting and 0.67-0.72 (set A) under audited grouped splitting; logistic regression stayed near 0.52-0.55 in every strategy. Initial groups left detected cross-partition matches, which were merged before fitting; a post hoc higher-sensitivity search still found residual related allergens in 30% splits. Added descriptors changed mean MCC by -0.010 to +0.021 with inconsistent direction. On the reserved authors-defined holdout (2,780 unique sequences), random forest MCC was 0.617 (A) and 0.596 (B); 225 sequences had a detected match of at least 40% identity to development data, and MCC for the other 2,555 was 0.568 and 0.535.

Conclusions: Deduplication, grouped splitting with similarity auditing and simple baselines materially changed the interpretation of performance; the added descriptors gave no reliable benefit. These findings concern benchmark evaluation, not patient diagnosis or food safety.

Keywords: Allergen Protein Classification; Sequence Similarity; Data Leakage; Deduplication; Matthews Correlation Coefficient; Reproducibility; MMseqs2

OPEN ACCESS

Correspondence:
Dipesh Gurung, Lord Buddha Education Foundation, Kathmandu, Nepal, E-mail: dipesh.gurung@lbef.edu.np

Received Date: *15 Oct 2026*
Accepted Date: *03 Nov 2026*
Published Date: *05 Nov 2026*

Citation:
Bhattarai B, Gurung D, Thakur RN. Sequence Similarity, Deduplication and Feature Choice in the Evaluation of Allergen Protein Classifiers. WebLog J Allergy. wja.2026.k0502. https://doi.org/10.5281/zenodo.24223971

Copyright *© 2026 Dipesh Gurung. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Abbreviations

AAC: Amino-Acid Composition; AP: Average Precision; AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient

Introduction

Allergen protein classification is used to prioritise candidate proteins for experimental assessment, for example when screening novel or engineered proteins. Computational classifiers are trained on collections of sequences labelled as allergens or non-allergens, and their reported accuracy depends heavily on how the evaluation data were assembled. AlgPred 2.0 combined a similarity search, epitope mapping, motif and machine-learning components, and reported an area under the receiver operating characteristic curve of 0.98 and a Matthews correlation coefficient of 0.85 for its best model on a validation set that was described as sharing no more than 40% similarity with other partitions [1]. The method extended an earlier server [2]. Later deep learning work trained on the same or related collections, including a convolutional network [3] and protein-language-model embeddings [4].

Biological sequence data pose a well-known evaluation problem: related sequences placed on both sides of a split allow a model to reproduce labels of near neighbours, and homology itself is a strong predictor [5]. Guidance on avoiding leakage and on reporting validation is available [6, 7], and dedicated splitting tools exist [8]. For allergens specifically, a recent preprint that proposed a protein-language-model framework (Applm) built similarity-controlled benchmarks and reported that, in earlier allergen studies, inter-split similarity was often not controlled and that clustering with

CD-HIT [9] did not guarantee the intended separation, including for AlgPred 2.0 partitions [10]. Sequence-aware evaluation of allergen classifiers is therefore not new, and we do not claim it as such.

The question addressed here is narrower and practical. For lightweight, fully reproducible classifiers built on amino-acid composition and a few physicochemical descriptors, how much of the apparent performance depends on (i) exact duplicate sequences and the class imbalance created by removing them, (ii) the sequence-identity threshold used to group sequences and the thoroughness with which cross-partition similarity is audited, and (iii) the choice of descriptors? Descriptor sets such as protein-level molecular weight, isoelectric point, hydropathy, aromaticity and instability index are cheap to compute and are often added in the expectation of better discrimination, but their value under similarity-controlled evaluation has been less clearly documented than the value of leakage control itself.

We therefore present a computational robustness and reproducibility study, not a new prediction algorithm. Our contributions are: a full accounting of exact duplicates in the public AlgPred 2 files and its effect on class balance; a search-based audit and repair of MMseqs2 sequence groups at three identity thresholds, with detected crossings reported before and after repair; a paired feature-set comparison on identical test sequences across three seeds; a locked-model evaluation of the authors-defined holdout that reports both the full unique set and the subset without a detected close training relative; and a check of the grouping used for bootstrap uncertainty estimates. Analysis code, sequence identifiers, split definitions, manifests and checksums are available from the corresponding author on reasonable request.

Materials and Methods

Source data and exact-sequence audit

Sequence files were retrieved from the authors' AlgPred 2 repository at commit 533225b2e7ee766b3cad43dc73bd00c0e098d7a6 [11] (GPL-3.0 license; an archived record is cited in the repository documentation [12]). The four files (training allergens, training non-allergens, validation allergens, validation non-allergens) contain 20,150 records in total. File URLs, sizes and SHA256 checksums were recorded. In the primary description, allergens were compiled from several databases and earlier datasets, and non-allergens were reviewed Swiss-Prot entries retrieved with a query that excluded allergen-, allergenic- and cancer-related annotation, so the negative labels are putative [1]. The primary description also states an 80%/20% division of clusters into training and validation data; the validation files are therefore a holdout from the same assembled corpus and not an independent external dataset.

Records were checked for invalid residues, length (minimum 50 residues), identical sequences within and across labels, and identical sequences shared between training and validation files. Exact duplicates within a label were collapsed to one record. No identical sequence carried conflicting labels, no record contained invalid residues, and no sequence appeared in both the training and validation files. The deduplicated training sequences form the development corpus. The deduplicated validation sequences were reserved as the authors-defined holdout and were never used for model selection.

Dataset access: AlgPred 2 source sequence files at the analyzed commit.

Features

Feature set A contained the 20 amino-acid fractions and the natural logarithm of one plus sequence length (21 features). Feature set B added molecular weight (average masses, unmodified linear protein), theoretical isoelectric point, mean hydropathy (GRAVY), aromaticity (the combined fraction of F, W and Y) and the instability index (26 features), computed with the ProtParam module of Biopython 1.88 [13, 14] using published definitions [15, 16]. Finite values, composition sums of one, positive molecular weights and the aromaticity identity were verified for all 10,514 development sequences.

Models, tuning and baselines

Logistic regression (standardised inputs, maximum 3,000 iterations) and random forest [17] (300 trees, square-root feature subsampling) were fitted with scikit-learn 1.9.1 [18]. The logistic grid contained inverse regularisation strength C in {0.01, 0.1, 1, 10} and class weights {none, balanced}. The forest grid contained maximum depth {12, unlimited}, minimum leaf size {5, 1} and class weights {none, balanced}. Within each outer training partition, settings were selected by five-fold inner cross-validation maximising MCC, and the best setting was refitted on the whole outer training partition. The decision threshold was fixed at 0.5. Two diagnostics were run on every outer split: a prior-probability dummy classifier and a logistic regression using only log length (tuned with the logistic grid). They were run for set A inputs only.

Partitioning strategies

Four strategies were evaluated with seeds 17, 42 and 2026: random stratified splitting, and splitting by sequence groups defined at 30%, 40% and 60% identity. In every strategy approximately 20% of development sequences (2,103 or 2,104 sequences, of which 491 were allergens) formed the outer test set; the remainder was divided into five inner folds. Grouped strategies used stratified group splitting so that whole groups stayed together in the outer and inner partitions. Sequence groups were obtained by clustering allergens and non-allergens jointly with MMseqs2 [19, 20] (identity from alignment, at least 80% coverage of both sequences, sensitivity 7.5, connected-component clustering, single-step workflow; binary checksum and version are recorded in the package). Identity thresholds around 30-40% lie near the range in which sequence relationships become difficult to infer [21].

Cross-partition similarity audit and group repair

Cluster membership alone does not guarantee that no similar pair spans two partitions. An independent all-against-all MMseqs2 search of the 10,514 development sequences (minimum identity 0.3, at least 80% bidirectional coverage, sensitivity 7.5, maximum 11,000 hits, E-value 100) was therefore used to list detected matches. For each split, detected pairs at or above the strategy threshold (0.4 for random splits, shown for reference) that crossed an outer or inner partition boundary were counted. In the first set of grouped splits, such crossings were found. Original groups linked by any qualifying detected match were merged (union-find on the audit matches), splits were regenerated, and the audit was repeated against the same saved match list before any model was fitted. Original manifests and audit files were preserved. Because the repair merged every qualifying pair from this list, a repeat audit against the same list returns zero crossings by construction; it verifies the implementation, not the completeness of the search. We therefore added a post hoc re-search with higher sensitivity and a longer hit list (sensitivity 8.5, E-value

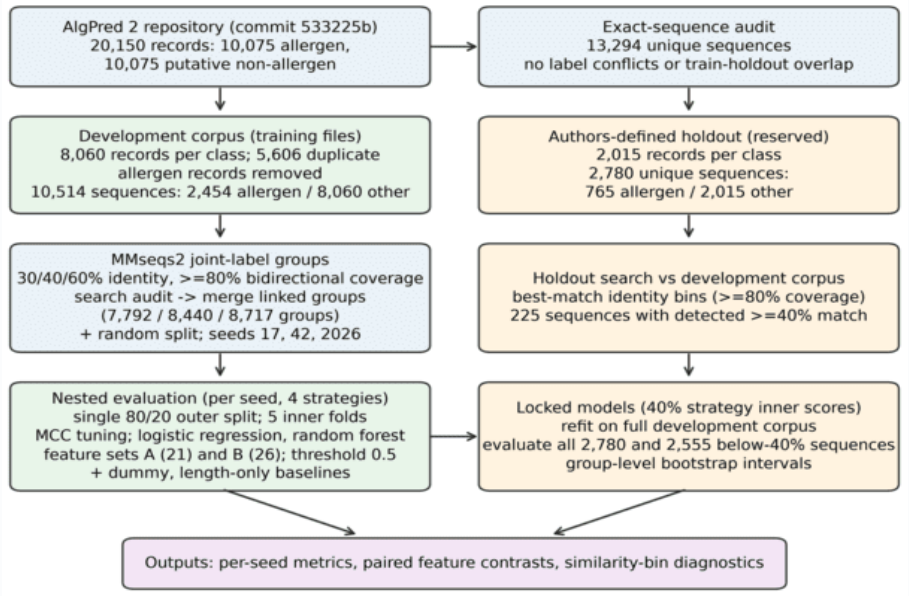


Figure 1: Dataset preparation and evaluation workflow. Source records were audited for exact duplicates, split into a development corpus and a reserved authors-defined holdout, grouped with MMseqs2 at 30%, 40% and 60% identity, audited for cross-partition matches and repaired by merging linked groups, and used for nested model evaluation and locked-model holdout evaluation. Numbers are sequence counts.

1000), counted residual crossings in the repaired splits at 80% and at 50% coverage of both sequences, and compared allergen recall for outer-test sequences with and without a residual match to the outer training partition. Outer-test MCC was also recomputed after removing flagged sequences; because flagged sequences differ in class composition, this is a diagnostic and not a corrected estimate. The original splits and fits were not changed.

Metrics and paired comparisons

The primary metric was MCC [22, 23]. Secondary metrics were the area under the receiver operating characteristic curve (AUC), Average Precision (AP), sensitivity, specificity, balanced accuracy and Brier score. Per-run confusion counts were recomputed from saved predictions and reconciled with the stored metrics. Feature sets A and B were compared on identical test sequences within each seed, strategy and model (paired contrasts of MCC). Contrasts between strategies are descriptive only because each strategy has a different outer test set.

Uncertainty

Intervals were obtained by resampling whole sequence groups with replacement (1,000 replicates, fixed random seed, percentile 95% intervals) and summing group-level confusion counts. Paired feature contrasts resampled the same groups for both feature sets. The original analysis used audited 40% groups for every strategy. Because the independence unit should reflect each strategy, the bootstrap was repeated using the strategy's own audited groups (40% groups for the random strategy) and, as a conservative sensitivity analysis, audited 30% groups for every strategy; the original results were retained. These intervals condition on the fitted models and the realised split; they do not include uncertainty from tuning, partition choice or label noise, and they are reported per seed without pooling.

Similarity bins

For each outer test sequence, the highest identity to any sequence in the corresponding outer training partition was taken from the saved search results (at least 80% bidirectional coverage), and sequences were assigned to four bins: no detected match at 30% identity, 30 to below 40%, 40 to below 60%, and at least 60%. Bins with fewer than 20 sequences of either class were flagged sparse. MCC was treated as undefined when a bin contained only one class, and sensitivity for the allergen class is shown instead.

Authors-defined holdout

Hyperparameters for the four model-feature combinations were locked before the holdout was scored, by averaging inner-fold MCC across the three seeds of the prespecified 40% grouped strategy using development data only. Each locked model was refitted on all 10,514 development sequences and evaluated on the 2,780 unique holdout sequences. Holdout performance was not used for any selection. Holdout sequences were searched against the development corpus with the same MMseqs2 settings; the best detected identity defined the bins above, and results are given for all holdout sequences and for the 2,555 sequences with no qualifying match or a best match below 40%. Bootstrap intervals resampled groups formed from the audited 40% development groups together with detected holdout–development and holdout–holdout links of at least 40%.

Reproducibility

Python 3.12.14 with pinned package versions was used on Windows 11, with MMseqs2 run through an Ubuntu subsystem. Scripts, manifests, checksums and exact commands are available from the corresponding author on reasonable request. Raw sequences are not redistributed; retrieval instructions are given instead. The overall workflow is summarised in Figure 1.

Results and Discussion

Dataset composition and deduplication

The repository contained 20,150 records (Table 1), which collapsed to 13,294 unique sequences. All 6,856 removed duplicate records were allergen records: 5,606 from training and 1,250 from validation. The training allergen file held 8,060 records but only

Table 1: Dataset flow from the source repository to the development corpus and authors-defined holdout.

Stage	Records	Unique sequences	Exact duplicates removed	Max copies / composition
Original training, allergen	8,060	2,454	5,606	19
Original training, putative non-allergen	8,060	8,060	0	1
Original validation, allergen	2,015	765	1,250	14
Original validation, putative non-allergen	2,015	2,015	0	1
Total	20,150	13,294	6,856	19
Development corpus		10,514		2,454 allergen / 8,060 non-allergen (23.3% allergen)
Authors-defined holdout		2,780		765 allergen / 2,015 non-allergen (27.5% allergen)

Duplicates were identical sequences within one label; none were found across labels or between training and validation files. Allergen and non-allergen labels follow the source repository; non-allergens are putative.

2,454 unique sequences, with up to 19 copies of one sequence; 772 sequences occurred once and the remainder occurred between two and 19 times. The validation allergen file showed the same pattern (2,015 records, 765 unique sequences, up to 14 copies). The non-allergen files contained no repeated sequences. The balanced 1:1 composition of the source files therefore did not survive exact deduplication: the development corpus contained 2,454 allergens and 8,060 non-allergens (23.3% allergens), and the holdout 765 and 2,015 (27.5%).

The consequences are practical. Accuracy becomes uninformative (a classifier that always predicts non-allergen scores 76.7% accuracy on the development corpus), the average-precision baseline equals the allergen prevalence, and MCC, which was the primary metric here, is preferred under imbalance [23]. A record-level random split of the original training allergen file would also place an exact copy of most held-out allergens in the training part: in a simple illustration with three seeds, 87.0-89.3% of held-out allergen records had an identical sequence in the training portion. The cause of the repeated records is not established by the primary description, which reports assembly from several databases and earlier datasets [1]; overlapping source collections are one possibility, but we do not infer intent or error and we report the redundancy as observed. The generic identifiers in the files do not allow source or species to be recovered, so source or species effects could not be examined.

Audited sequence groups

Clustering the development corpus gave 7,808, 8,445 and 8,730 groups at 30%, 40% and 60% identity. The independent search nevertheless detected matches spanning outer or inner partitions in the initial grouped splits (44, 22 and 90 directed pairs summed over the three seeds at 30%, 40% and 60%; maximum crossing identity 0.79, 0.79 and 0.96). Merging groups linked by qualifying matches gave 7,792, 8,440 and 8,717 groups (largest group 246, 246 and 224 sequences). In the regenerated splits, no detected pair at or above the threshold crossed an outer or inner boundary for any of the nine grouped splits (Supplementary Tables S5-S6). Random splits retained 130,845 detected directed pairs at or above 40% identity across outer boundaries (summed over seeds; maximum identity 1.0). This contrast should be read as detected separation under one search configuration, not as proof that no related pair exists; the zero for repaired splits is true by construction for the match list used in the repair.

A post hoc re-search with higher sensitivity showed that the separation was incomplete (Supplementary Tables S15-S17). At the strategy threshold and with at least 80% coverage of both sequences, it found 433-439 directed crossing pairs per 30% split, 224, 4 and 4 for the 40% splits (seeds 17, 42, 2026) and 18-21 per 60% split. Averaged over seeds, 220 of 2,104 outer test sequences (10.5%) at 30%, 39 of 2,103 (1.8%) at 40% and 8 (0.4%) at 60% had such a match to the outer training partition, and nearly all of them were allergens (217 of 491 test allergens at 30%). In 30% splits, allergen recall of the random forest (set A) was 0.940 for allergens with a residual match and 0.443 for those without (logistic regression 0.991 and 0.618), and removing flagged sequences lowered MCC from 0.674 to 0.471; the corresponding changes were 0.680 to 0.662 at 40% and 0.725 to 0.721 at 60%. The 30% estimate therefore substantially reflects allergens with detectable relatives, which limits interpretation of the relatively small difference from the 40% estimate, and it should not be read as performance on unrelated sequences. The 40% and 60% estimates were affected little (largely a single seed at 40%), and the random-split versus grouped contrast remained large (random forest A at 40% after filtering: 0.662 versus 0.843 with random splitting). With the coverage requirement relaxed to 50%, 2,441-3,787 crossing pairs per 30% or 40% split were found, including pairs with 100% identity over partial coverage, so partial-length and domain-level similarity is not controlled by 80%-coverage grouping.

Performance across strategies

All 72 expected fits (48 primary and 24 diagnostic) were present and their stored metrics matched values recomputed from the saved predictions. Table 2 and Figure 2 summarise the results. For random forest with set A, mean MCC was 0.843 under random splitting and 0.674, 0.680 and 0.725 under 30%, 40% and 60% grouping; the corresponding AUC values were 0.974 and 0.905-0.930. Set B gave the same pattern (0.853 for random splitting; 0.679-0.715 for grouped strategies). Grouped estimates were therefore about 0.12-0.17 MCC lower than random-split estimates. The prespecified hypothesis that random splitting gives a higher MCC than 40% grouping held for all six-forest seed and feature-set combinations (random minus 40%: +0.137 to +0.197) but for none of the six logistic combinations (-0.015 to -0.026). Because each strategy evaluates a different test set, this is a descriptive contrast of evaluation designs; it is consistent with, but does not prove, a contribution of related training sequences to the random-split estimate (see similarity bins below). Differences among the three grouped thresholds were smaller than the per-seed ranges for some comparisons and should not be over-interpreted.

Logistic regression behaved differently. Its mean MCC stayed between 0.519 and 0.553 across strategies and feature sets, with high sensitivity and lower specificity (mean sensitivity about 0.79 and specificity about 0.81 in set A; class-balanced weighting was chosen in 46 of the 48 primary runs, all 24 forests and 22 of 24 logistic

Table 2: Test performance by partition strategy, model and feature set (mean of three seeds; MCC range across seeds in parentheses).

Strategy	Model	Set	MCC (seed range)	Mean per-seed 95% interval	AUC	AP	Sens.	Spec.
Random	Forest	A	0.843 (0.830–0.857)	0.79 to 0.88	0.974	0.950	0.83	0.98
Random	Forest	B	0.853 (0.846–0.857)	0.80 to 0.89	0.974	0.951	0.85	0.98
Random	Logistic	A	0.537 (0.533–0.542)	0.43 to 0.62	0.876	0.675	0.79	0.81
Random	Logistic	B	0.527 (0.513–0.535)	0.40 to 0.62	0.875	0.684	0.61	0.90
30% groups	Forest	A	0.674 (0.661–0.690)	0.30 to 0.81	0.905	0.797	0.66	0.96
30% groups	Forest	B	0.679 (0.662–0.699)	0.29 to 0.81	0.911	0.812	0.66	0.96
30% groups	Logistic	A	0.532 (0.524–0.543)	0.21 to 0.70	0.851	0.574	0.78	0.81
30% groups	Logistic	B	0.535 (0.523–0.555)	0.21 to 0.70	0.854	0.587	0.79	0.81
40% groups	Forest	A	0.680 (0.660–0.704)	0.34 to 0.80	0.924	0.821	0.67	0.96
40% groups	Forest	B	0.701 (0.686–0.709)	0.36 to 0.81	0.923	0.827	0.69	0.96
40% groups	Logistic	A	0.553 (0.548–0.561)	0.24 to 0.71	0.867	0.599	0.83	0.80
40% groups	Logistic	B	0.549 (0.533–0.560)	0.24 to 0.70	0.871	0.614	0.83	0.79
60% groups	Forest	A	0.725 (0.715–0.740)	0.44 to 0.85	0.930	0.864	0.73	0.96
60% groups	Forest	B	0.715 (0.711–0.723)	0.42 to 0.84	0.929	0.862	0.72	0.96
60% groups	Logistic	A	0.522 (0.512–0.530)	0.25 to 0.69	0.857	0.580	0.77	0.81
60% groups	Logistic	B	0.519 (0.506–0.530)	0.24 to 0.68	0.859	0.594	0.77	0.81
Random	Length only	log length	0.248 (0.231–0.269)	–	0.694	0.349	0.65	0.64
30% groups	Length only	log length	0.345 (0.339–0.350)	–	0.721	0.380	0.77	0.64
40% groups	Length only	log length	0.336 (0.321–0.351)	–	0.726	0.391	0.77	0.62
60% groups	Length only	log length	0.306 (0.258–0.340)	–	0.708	0.373	0.73	0.63

Set A, composition plus log length; set B, set A plus five physicochemical descriptors. MCC: Matthews Correlation Coefficient; AUC: Area Under the receiver operating characteristic Curve; AP: Average Precision (allergen prevalence 0.233); Sens: Sensitivity; Spec: Specificity at threshold 0.5. Intervals are means of per-seed 95% group-bootstrap limits using strategy-matched groups; they condition on fitted models and are not pooled. Strategies use different outer test sets, so contrasts between rows are descriptive. The dummy classifier had MCC 0 and AUC 0.5 in all strategies (Supplementary Table S7).

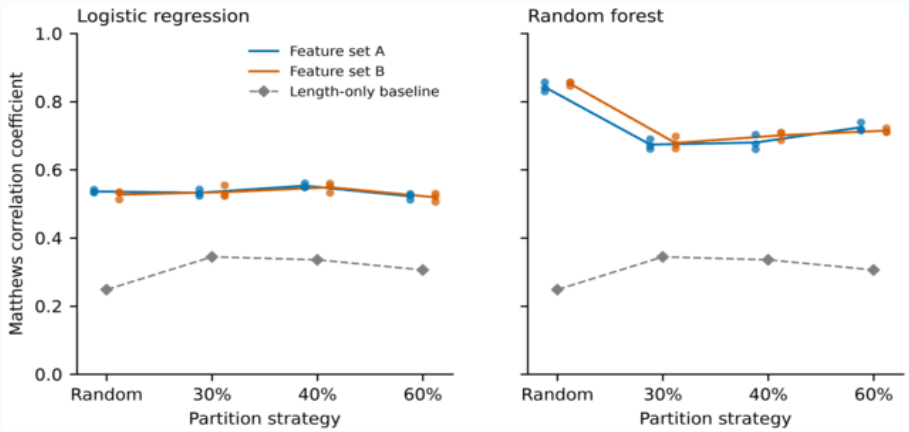


Figure 2: Matthews Correlation Coefficient (MCC) by partition strategy, model and feature set. Points are individual seeds (17, 42, 2026); lines join seed means. Blue, feature set A; orange, feature set B; grey dashed line, length-only logistic baseline. Strategies use different outer test sets, so comparisons across strategies are descriptive.

models), whereas the random forest showed higher specificity (about 0.96) and lower sensitivity. The low-capacity linear model was thus insensitive to the evaluation design but also never reached the grouped-split performance of the forest. The length-only diagnostic reached a mean MCC of 0.248 under random splitting and 0.345 at most under grouped splitting, and the dummy classifier scored 0 (Table 2; Supplementary Table S7), so sequence length alone carries some signal and should be controlled when interpreting composition models.

Effect of added descriptors

Paired contrasts of set B minus set A on identical test sequences (Table 3, Figure 3) were small and inconsistent. The mean MCC difference for random forest was +0.010, +0.005, +0.021 and -0.010 for random, 30%, 40% and 60% strategies, and for logistic regression -0.009, +0.002, -0.005 and -0.003. Of the 24 seed-level contrasts (four strategies, two models, three seeds), the matched-group bootstrap interval excluded zero in three (random forest at random, 40% and 60%, the last being negative); four excluded zero

Table 3: Paired difference in MCC between feature set B and feature set A on identical test sequences.

Strategy	Model	Mean B−A	Seed range	Seeds B>A	Intervals excluding 0 (matched groups)	Intervals excluding 0 (30% groups)
Random	Forest	+0.010	+0.000 to +0.016	3/3	1/3	1/3
Random	Logistic	−0.009	−0.020 to +0.000	1/3	0/3	0/3
30% groups	Forest	+0.005	+0.001 to +0.008	3/3	0/3	0/3
30% groups	Logistic	+0.002	−0.007 to +0.012	2/3	0/3	0/3
40% groups	Forest	+0.021	+0.006 to +0.048	3/3	1/3	1/3
40% groups	Logistic	−0.005	−0.015 to +0.002	1/3	0/3	1/3
60% groups	Forest	−0.010	−0.029 to +0.003	1/3	1/3	1/3
60% groups	Logistic	−0.003	−0.006 to +0.001	1/3	0/3	0/3

B-A: MCC of feature set B minus MCC of feature set A. Per-seed 95% group-bootstrap intervals; counts are of seeds (out of three) whose paired interval excluded zero. Intervals are unadjusted for multiplicity and are not significance tests.

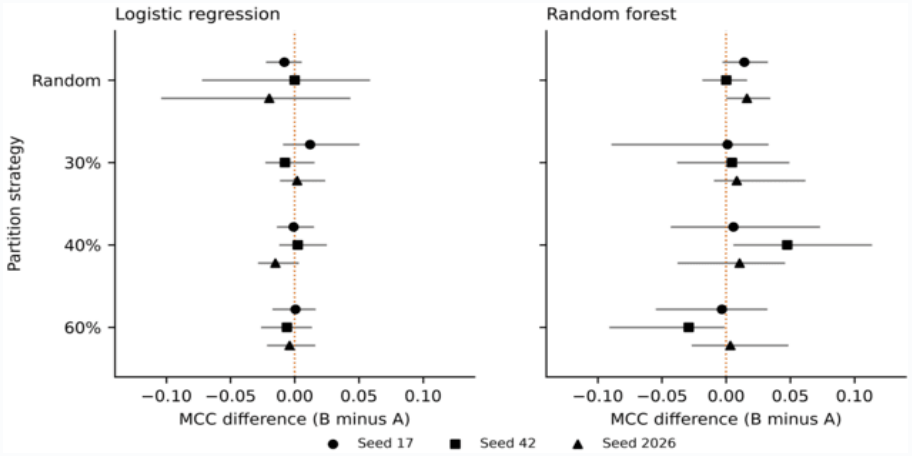


Figure 3: Paired difference in MCC between feature set B and feature set A on identical test sequences, by seed. Points are point estimates and horizontal lines are 95% group-bootstrap intervals using strategy-matched groups. The dotted line marks no difference. Intervals are unadjusted and condition on fitted models.

under the conservative 30% grouping. These are unadjusted per-seed intervals from overlapping data, so no significance claim is made. The descriptors were not reliably useful: the largest mean gain (+0.021 MCC, random forest at 40%) appeared in one strategy and reversed in sign at 60%. For logistic regression, AP was slightly higher with set B in all four strategies and AUC in three of four, which indicates a small ranking benefit that did not translate into a consistent MCC gain at the fixed threshold of 0.5. The mixed picture agrees with the practical expectation that composition and length already capture much of what coarse physicochemical summaries encode.

Uncertainty and the bootstrap grouping check

Per-seed 95% intervals for MCC were wide for grouped strategies (mean width about 0.46 for 40% groups) and narrower for random splits (0.15), because grouped test sets contain whole large groups whose composition differs from seed to seed. Inspection of the original bootstrap showed that audited 40% groups had been used for all strategies, which is finer than the splitting unit at 30% (about 1,685 resampling units *versus* 1,557) and coarser than the unit at 60% (about 1,720 versus 1,740). Repeating the bootstrap with matched groups changed the mean interval width by at most 0.006 (30%: 0.497 to 0.502; 60%: 0.432 to 0.426), and the conservative 30% grouping gave similar widths (0.459 at 40%). Conclusions were therefore not sensitive to this choice, but the matched intervals are reported in Tables 2 and 3. Interval limits in Table 2 are means of per-seed limits and are not a pooled interval, because the seeds' test sets overlap.

Similarity bins

Bins were strongly associated with class (Figure 4, Supplementary Table S9). In random outer tests (seed means), the bin with a detected match of at least 60% contained about 407 allergens and no non-allergens, while the bin without a detected match at 30% held about 1482 sequences of which only about 45 were allergens. Allergens therefore usually had close relatives in training whereas non-allergens usually did not, which is consistent with the redundancy of the allergen source collections. Sensitivity of the random forest for allergens was 0.927 in the at-least-60% bin and 0.307 where no match was detected; the 30-<40% bin (about 11 allergens) was sparse. In 60% grouped tests the 40-<60% bin held about 342 allergens but only about 16 non-allergens (sparse for non-allergens), with allergen sensitivity 0.876 against 0.424 where no match was detected. Because bins are confounded with class and several are sparse, we draw no strong conclusion beyond the qualitative pattern that allergen recall rises when a close training relative exists. A bin of 'no detected match' means no match at 30% identity and 80% coverage was found; it does not demonstrate absence of homology.

Authors-defined holdout

Locked settings were selected from inner-fold scores on development data only (mean inner MCC 0.468 and 0.471 for logistic A and B, and 0.543 and 0.569 for random forest A and B; Supplementary Table S10). On the full unique holdout (2,780 sequences; Table 4, Figure 5), MCC was 0.553 and 0.558 for logistic

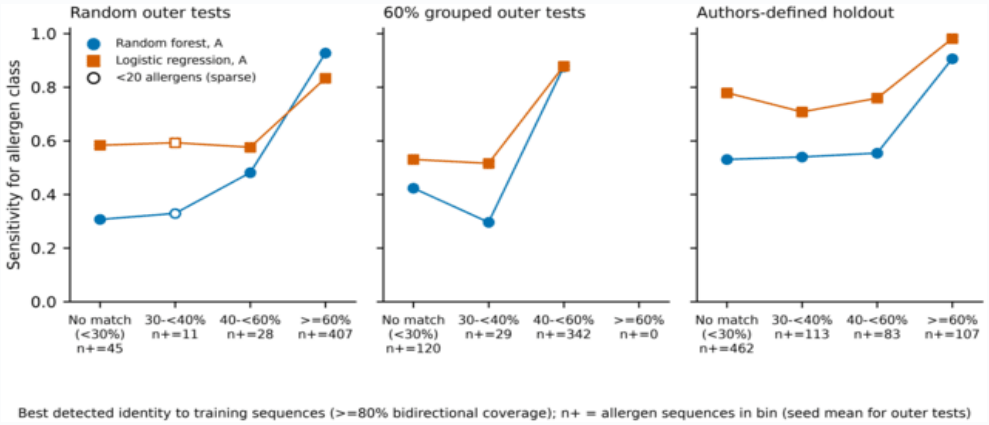


Figure 4: Sensitivity for the allergen class by best detected identity to training sequences (at least 80% bidirectional coverage) for random-split outer tests, 60% grouped outer tests (seed means) and the authors-defined holdout. Open symbols mark bins with fewer than 20 allergens; n+ is the number of allergens in each bin (seed mean for outer tests). Bins are strongly associated with class (for example, the at-least-60% bin of random-split tests contains no non-allergens), and the 60% grouped panel has no bin at or above 60% by construction. Specificity and MCC are given in Supplementary Table S9.

Table 4: Locked-model performance on the authors-defined holdout, full unique set and subset without a detected match of at least 40% identity to development sequences.

Model	Set	Scope	N	Allergens	MCC (95% interval)	Sens.	Spec.	AUC	AP
Logistic	A	All unique	2,780	765	0.553 (0.472–0.614)	0.79	0.80	0.862	0.708
Logistic	A	<40% or no match	2,555	575	0.505 (0.416–0.581)	0.77	0.80	0.850	0.659
Logistic	B	All unique	2,780	765	0.558 (0.477–0.622)	0.80	0.80	0.865	0.713
Logistic	B	<40% or no match	2,555	575	0.509 (0.421–0.586)	0.77	0.80	0.852	0.663
Forest	A	All unique	2,780	765	0.617 (0.510–0.691)	0.59	0.96	0.904	0.818
Forest	A	<40% or no match	2,555	575	0.568 (0.443–0.662)	0.53	0.96	0.893	0.755
Forest	B	All unique	2,780	765	0.596 (0.491–0.673)	0.56	0.96	0.909	0.817
Forest	B	<40% or no match	2,555	575	0.535 (0.396–0.638)	0.49	0.96	0.896	0.744

Models were locked using inner-fold scores on development data only and refitted on all 10,514 development sequences. Intervals resample sequence groups (1,000 replicates) and condition on the fitted models. "<40% or no match": best detected identity below 40% or no qualifying match (80% bidirectional coverage). Abbreviations as in Table 2.

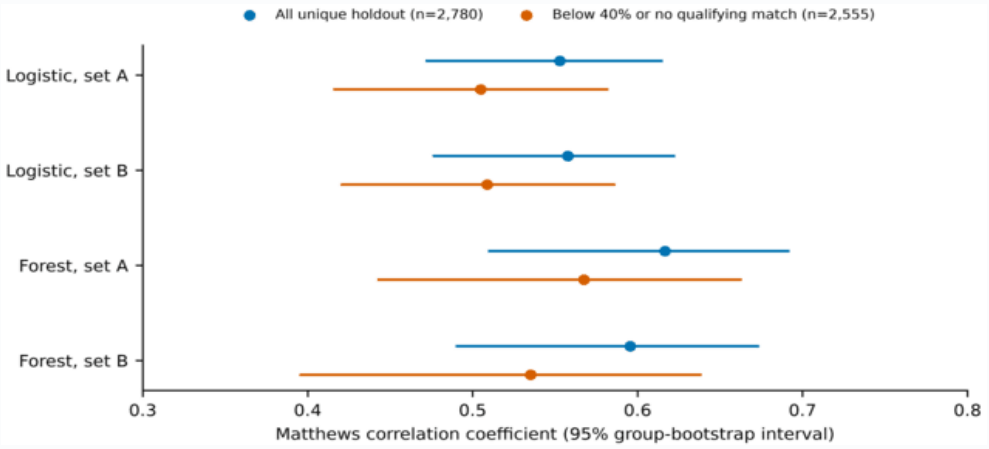


Figure 5: MCC of locked models on the authors-defined holdout, for all 2,780 unique sequences (blue) and the 2,555 sequences with no detected match of at least 40% identity to development sequences (orange). Horizontal lines are 95% group-bootstrap intervals conditional on the fitted models.

regression and 0.617 and 0.596 for random forest. Holdout search found 225 sequences (8.1%) with a detected match of at least 40% identity and 80% coverage to development data: 118 in the 40-<60% bin (83 allergens) and 107 at 60% or more (all allergens). For the 2,555 sequences without such a match, MCC was lower: 0.505 and 0.509 for logistic regression and 0.568 and 0.535 for random forest, reductions of 0.049 (A) and 0.060 (B) for the forest, with overlapping intervals. Allergen sensitivity of the forest was 0.907 in the at-least-60% bin and 0.530–0.554 elsewhere (Supplementary Table S11). Feature set A was nominally better than B for the forest on the holdout, but the intervals overlap broadly and the comparison is a single fitted model per feature set. The holdout MCC of the forest (A, all sequences: 0.617)

is lower than the 40% grouped development estimate (0.680) and far lower than the random-split estimate (0.843). The below-40% set is an overlap-filtered analysis of the same corpus, not a new independent dataset, and the holdout remains an authors-defined sample from the assembled data.

Comparison with previous work

AlgPred 2.0 reported its best model at AUC 0.98 and MCC 0.85 on the validation set [1]. Our random-split forest estimates (MCC 0.84-0.85) are numerically similar, whereas unique-sequence grouped and holdout estimates are considerably lower. We did not reproduce the authors' pipeline: their models, features, similarity search and record-level validation files (which contain repeated allergen records) differ from ours, so the figures are not directly comparable and we make no claim that the original result is wrong. What our analysis does show is that, under the same corpus, evaluation choices alone move forest MCC across a range of about 0.3. The primary study states that no two proteins in different sets exceed 40% similarity under CD-HIT-based partitioning [1]; we detected 225 holdout sequences with a match of at least 40% identity under a different tool and alignment-based definition. The two statements are not contradictory because the definitions and search sensitivity differ, and a recent preprint reports similar concerns about CD-HIT-based separation in allergen datasets [10].

Deep learning classifiers built on these collections report higher accuracy than the lightweight models studied here. AllergenAI used a random division and five-fold cross-validation, together with a separate check on newly deposited allergens [3], and AllerTrans reported results on the AlgPred 2.0 training and validation sets using protein-language-model embeddings; the extent of redundancy control was not established in our literature comparison [4]. Applm is closest in spirit, with benchmarks designed around novel allergens, high-similarity homologues and mutation effects [10]. Our contribution beyond these studies is not a new similarity-controlled benchmark or a more accurate model. It is a documented, low-compute and fully scripted accounting of how exact deduplication, the grouping threshold, a post-clustering match audit with repair, simple baselines and descriptor choice jointly affect conclusions for classical feature sets, with uncertainty that follows the grouping used. Whether the same sensitivities apply to embedding-based models was not tested.

Limitations

Labels are putative: non-allergens were defined by database annotation and may include unrecognised allergens, and all metrics are conditional on these labels. The source holdout comes from the same assembled corpus and shares its composition and curation; it is not an external or prospective validation. Species, source and accession metadata were unavailable, so source shortcuts could not be examined. Exact duplicate removal handled identical sequences only; the cause of the redundancy remains unconfirmed.

Similarity was assessed with one prespecified search configuration (MMseqs2, 80% bidirectional coverage, sensitivity 7.5); a more sensitive post hoc search found residual related allergens, especially in 30% splits, so grouped estimates at 30% in particular overstate performance on unrelated sequences, and local similarity at lower coverage was not controlled. We did not regroup and refit with the more sensitive match list. Group repair at one threshold does not control similarity at lower thresholds. Only logistic regression and random forest with two descriptor sets were studied; all conclusions are specific to them. The decision threshold was fixed at 0.5 and no calibration or threshold analysis was performed. Intervals are per-seed, conditional on fitted models and do not include tuning or split uncertainty. Comparisons between strategies are descriptive, similarity bins are confounded with class and partly sparse, and holdout bootstrap groups were formed at 40% only. The work concerns evaluation of sequence-based classifiers and does not address patient diagnosis, food safety, causal IgE epitopes or prospective clinical usefulness.

Conclusion

In the public AlgPred 2 data, exact deduplication converted a balanced benchmark into one with about 23% allergens, grouped splitting lowered random-forest MCC by roughly 0.12-0.17 relative to random splitting while leaving logistic regression essentially unchanged (a more sensitive post hoc search showed that residual related allergens remained mainly in the 30% splits), and five added physicochemical descriptors gave no reliable benefit. A locked-model evaluation of the authors-defined holdout gave lower estimates still, particularly for the sequences without a detected close training relative. Studies of allergen protein classifiers should report exact-duplicate handling, audit grouped partitions for residual similarity with more than one search sensitivity, include simple baselines such as sequence length, treat negative labels as putative, and describe the grouping used for uncertainty. The results support careful benchmark evaluation; they do not establish diagnostic or clinical validity.

Acknowledgements

The authors thank the creators of the public datasets and open-source software cited in this manuscript.

Author Contributions

B.B.: conceptualization, methodology, investigation, formal analysis, visualization, writing - original draft, and writing - review and editing. D.G.: methodology, validation, interpretation of findings, and writing - review and editing. R.N.T.: supervision, conceptual and methodological guidance, and writing - review and editing. All authors reviewed and approved the final manuscript.

Funding

This research received no specific funding.

Competing Interests

The authors declare no competing interests.

Ethics Statement

This study used only previously published, publicly available protein-sequence records. No new participants were recruited, no clinical intervention was performed, and no identifiable patient data were collected. Ethics review was not required.

Use of AI Tools

An AI assistant supported code development, computational execution, analysis, literature retrieval and drafting. The authors verified the analyses, citations and text, take full responsibility for the content, and did not list the AI tool as an author.

Additional Information

Correspondence and requests for materials should be addressed to D.G. (dipesh.gurung@lbef.edu.np).

Data Availability

The protein-sequence datasets analysed in this study are publicly available from the AlgPred 2 repository at commit 533225b2e7ee766b3cad43dc73bd00c0e098d7a6 [11]; see the reference list for access details and terms of use. Fixed sequence identifiers, checksums, split assignments and per-sequence derived results generated for this study are available from the corresponding author on reasonable request. Raw protein sequences are not redistributed in the submission package; any redistribution remains subject to applicable source terms. Dataset: AlgPred 2 dataset files

Code Availability

Analysis code is available from the corresponding author on reasonable request.

References

1. Sharma N, Patiyal S, Dhall A, Pande A, Arora C, Raghava GPS. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. Brief Bioinform. 2021; 22(4): bbaa294.
2. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res. 2006; 34: W202-W209.
3. Liu J, Negi SS, Yang C, Zhou X, Schein CH, Braun W, et al. AllergenAI: a deep learning model predicting allergenicity based on protein sequence. BMC Bioinformatics. 2025; 26: 279.
4. Sarlakifar F, Malek H, Allahyari FN. AllerTrans: a deep learning method for predicting the allergenicity of protein sequences. Biol Methods Protoc. 2025; 10(1): bpaf040.
5. Hamp T, Kassner R, Seemayer S, Vicedo E, Schaefer C, Achten D, et al. Homology-based inference sets the bar high for protein function prediction. BMC Bioinformatics. 2013; 14(Suppl 3): S7.
6. Bernett J, Blumenthal DB, Grimm DG, Haselbeck F, Joeres R, Kalinina OV, et al. Guiding questions to avoid data leakage in biological machine learning applications. Nat Methods. 2024; 21(8): 1444-1453.
7. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, Capriotti E, et al. DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021; 18(10): 1122-1127.
8. Joeres R, Blumenthal DB, Kalinina OV. Data splitting to avoid information leakage with DataSAIL. Nat Commun. 2025; 16: 3337.
9. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13): 1658-1659.
10. Wong BSH, Kim JM, Fung SH, Xiong Q, Ao KFK, Wei J, et al. Driving accurate allergen prediction with protein language models and generalization-focused evaluation. arXiv. 2025.
11. Raghava GPS. algpred2. GitHub Repository.
12. Sharma N, Raghava G. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. Zenodo. 2026.
13. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25(11): 1422-1423.
14. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein identification and analysis tools on the ExPASy server. Humana Press. 2005.
15. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982; 157(1): 105-132.
16. Guruprasad K, Reddy BVB, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. 1990; 4(2): 155-161.
17. Breiman L. Random forests. Mach Learn. 2001; 45(1): 5-32.
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011; 12: 2825-2830.
19. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017; 35(11): 1026-1028.
20. Steinegger M, Soding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018; 9: 2542.
21. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999; 12(2): 85-94.
22. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405(2): 442-451.
23. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020; 21: 6.