NetAllergen, a random forest model integrating MHC-II presentation propensity for improved allergenicity prediction

Yuchen Li[1], Morten Nielsen[1,2], Carolina Barra[1,*]

[1] Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Argentina

* Corresponding author - carolet@dtu.dk

Abstract

Allergy is a pathological immune reaction towards innocuous protein antigens. Although only a small fraction of plant or animal proteins induce allergy, the atopic disorders affect millions of children and adults, and cost billions in healthcare systems worldwide. In-silico predictors can aid in the development of more innocuous food sources. Previous allergenicity predictors are based on sequence similarity, common structural domains, and amino acid physicochemical features. However, these predictors strongly rely on similarity to previous allergens and fail to accurately predict allergenicity when protein sequence similarity diminishes. Furthermore, allergen is a wide terminology that may include different compounds, hindering the classification task. To overcome these limitations, we used allergens collected from AllergenOnline, a curated database of IgE inducing allergens, carefully removed allergen redundancy with a novel protein partitioning pipeline, and developed a new allergen prediction method, NetAllergen, introducing MHC presentation propensity as a novel feature. NetAllergen was demonstrated to out-perform a sequence similarity-based Blast method as well as previous allergenicity predictors when similarity to known allergens is diminished. NetAllergen is available as a web-service

23 (services.healthtech.dtu.dk/service.php?NetAllergen-1.0) and can predict allergenicity from a
24 protein sequence.

25 Introduction

26 Currently, allergy affects as many as 20% of the worldwide population[1,2] and only in the US,
27 healthcare treatments costs 24.8 billion dollars annually[3]. Moreover, the prevalence of allergic
28 diseases has increased in the past decades, and it is likely to rise in the future[4,5].
29
30 Allergy is characterised by a pathological immune reaction, triggered by normally ubiquitous and
31 innocuous protein antigens in susceptible individuals[1]. In addition, allergy is a wide terminology
32 that encompasses a number of diseases primarily sharing a IgE-dependent pathway[6]. Allergy is
33 triggered through ingestion, inhalation or direct contact with the allergen[4]. Once the allergen
34 contacts the skin or mucosal tissues, antigen-presenting cells (APCs) uptake and digest allergens
35 into small peptides. Peptides with specific motifs are next presented by the major
36 histocompatibility complex (MHC) class II alleles to activate T helper (Th) 2 cells. In susceptible
37 individuals, Th2 cells further stimulate B cells to produce IgE antibodies in the first sensitization
38 stage. An individual develops an allergy when the immune system is re-exposed to the allergen.
39 The allergen is targeted by multiple IgE antibodies, which bind to Fc receptors in basophils and
40 mast cells activating the degranulation process[6]. Why and how only a small fraction of proteins
41 induce allergy is largely unknown. However, it has been suggested that allergens share specific
42 features, that according to the route and level of exposure, genetic predisposition, and
43 environmental factors can develop into allergies[7,8].
44

45 Different approaches have been applied to identify common patterns in allergens to predict
46 allergenicity. Early predictors were relational databases that clustered allergens according to their
47 protein sequence, or later to common structural domains, biochemical and functional
48 properties[7,9,10]. Allermatch, a sequence-comparing method, identifies allergens by finding shared
49 amino acids segments to known allergens[11]. Likewise, Blast has also been applied to search query
50 sequences against allergen databases in AllerTool[12]. Later, motif-profile predictions, such as in
51 WebAllergen, improved prediction by searching for similarities in potential functional domains[13].
52 Finally, machine learning-based methods with various features have been widely applied to predict
53 allergenicity. AlgPred was developed using support vector machines (SVMs) combining IgE
54 epitopes, motifs with amino acids and dipeptide compositions as features[14]. AllerHunter predicts
55 allergenicity using SVM with iterative pairwise sequence similarity as feature vectors[15]. In
56 addition to Blast, faster alignment tools were employed on Allerdictor, which uses K-mer
57 frequencies in a SVM[16]. The auto- and cross-covariance of Z-descriptors, the principal components
58 derived from 29 physicochemical properties matrix for 20 amino acids, were adopted in AllerTOP,
59 and AllerTOP v2 used the same transformation of E-descriptors which takes the principal
60 components from a matrix containing 237 properties[17,18]. AllergenFP was developed based on
61 descriptor fingerprints, which is based on Tanimoto coefficient similarity search[19]. Q-repeats, 3D
62 structural similarities, and sequential similarities were incorporated into the decision tree in
63 AllerCatPro[20]. AlgPred2 predicts sequences with hybrid scores by integrating Blast, and motif
64 features using a Random Forest model[21].
65
66 Although using different algorithms and alignment methods, current allergenicity predictors
67 highly rely on sequence similarity to known allergens, probably due to the FAO/WHO guidelines

68 for computational assessment of allergenic potential and the insurmountable task of IgE epitope
69 mapping towards allergens[5,20,22]. MHC class II antigen presentation plays an essential role in the
70 allergic response as allergen peptides need to be presented on MHC-II in order to develop Th2 and
71 IgE producing immune responses. However, to the best of our knowledge, yet no predictor has
72 used MHC presentation in allergenicity prediction. Recently, genome wide association studies
73 (GWAS) have revealed that several allergies are strongly associated with specific HLA alleles, in
74 particular HLA-DQ and HLA-DR alleles[23–25]. For example, allergic rhinitis has shown a strong
75 association with HLA-DRB1*04:01, HLA-DQB1*02:02, and HLA-DQB1*03:01[24]. The allele
76 HLA-DQA1*01:03-DQB1*06:01 was found to be associated with wheat allergy, and patients who
77 had celiac diseases carried HLA-DQA1*05:01-DQB1*02:01 and HLA-DQA1*03:01-
78 DQB1*03:02[23]. In allergic respiratory disease researches, HLA-DRB1*15:01 had association
79 with mite (Der p I)[26] and pollen (Bet v 1) allergens[27]. However, the same allele was negatively
80 associated with gelatine[28] and cow's milk[29]. These associations not only point to HLA class II
81 specific alleles, but also suggest that specific amino acid positions in the peptide binding groove
82 shared among different alleles provide susceptibility or protection against different types of
83 allergies[23,24]. Together these findings strongly suggest that differential MHC-II presentation might
84 be involved in allergy.
85
86 In this study, we developed a random forest predictor, NetAllergen, using MHC class II peptide
87 presentation propensity for the relevant and strongly associated HLA-DR and HLA-DQ alleles in
88 combination with physicochemical properties, structural information, and evolutionary
89 information, and investigated the relevance of MHC presentation propensity for allergenicity
90 predictions. NetAllergen was benchmarked against a sequence-based Blast model and other state-

91 of-the-art predictors and demonstrated to achieve unprecedented performance particularly for
92 allergens that share limited similarity to previously known and well characterised allergens.

93 Materials and Methods

94 Training dataset

95 2171 allergen protein IDs were retrieved from AllergenOnline, a curated database of IgE
96 producing allergens (On September 2020)[30]. Available allergen fasta sequences were downloaded
97 from NCBI and filtered to include protein sequences that ranged in length from 50 to 1000 amino
98 acids. To ensure that no highly redundant sequences existed in the dataset, a Hobohm 1[31] algorithm
99 with 90% sequence identity was applied to the dataset, gathering 686 allergic protein sequences.
100 In Hobohm 1, the protein sequences were firstly sorted from longer to shorter length. The first and
101 longest protein sequence was added to a "unique list" and the subsequent sequence was queried
102 against the protein in the unique list using Blast alignment tool. Sequence percentage identity was
103 determined as the number of amino acids that have an exact match with regard to the shorter of
104 the DB/query protein length. If the sequence identity was greater than 90% to any protein in the
105 unique list, the query sequence was discarded, otherwise appended to the unique list. This strategy
106 was repeated for each allergen in the sorted list.
107
108 For each allergen, five non-allergen protein candidates were selected by searching matching
109 species names with the allergen, and excluding results with some keywords (allergen, hypothetical,
110 uncertain, predicted, probable, putative) in NCBI. Additionally, sequence lengths were limited to
111 the ±20% of the allergen length. Considering possible annotation issues and duplicates, candidate

112 sequences were searched against allergen and adopted non-allergen datasets. The candidates were
113 finally randomly selected with E-values higher than $1x10^{-10}$ against the positives (to avoid
114 including unannotated proteins with high similarity to the positives) and $1x10^{-100}$ (to avoid
115 including duplicates) against negative sequences. The number of non-allergens collected per
116 allergen subset may vary (below five), because of the limited set of proteins for some species
117 applying the current filters. The negative dataset consisted of 2477 non-allergic protein sequences.

118 Feature constructions

119 **MHC-II presentation propensity scores**

120 NetMHCIIpan 4.0[32] was used to predict the number of peptides with a likelihood of being
121 presented on MHC class II with a percentile rank below 10%. Here, predictions were made for all
122 possible peptides in the allergen protein sequence using a sliding window of length 15. From these
123 numbers, two different features were collected. Binder 1 includes the counts binders to HLA-
124 DRB1*04:01, HLA-DQA1*02:01-DQB1*02:02, HLA-DQA1*04:01-DQB1*03:01, HLA-
125 DQA1*01:03-DQB1*06:01, HLA-DQA1*03:01-DQB1*03:02, and HLA-DQA1*05:01-
126 DQB1*02:01 and Binder 2 the counts for HLA-DRB1*15:01. These sets of alleles were selected
127 as they were previously demonstrated to be positively (Binder 1) or negatively (Binder 2)
128 associated with allergies. Final MHC presentation propensity scores were calculated by summing
129 the binders over different allele settings normalised based on protein sequence length.

130

131 **Physicochemical properties**

132 The physicochemical properties collected here included average residue weight, average charge,
133 and isoelectric point. Amino acids were also classified into tiny, small, aliphatic, aromatic, non-

134 polar, polar, charged, basic, and acidic groups by size, structure and charge. These were obtained
135 from Pepstats API in the EMBL-EBI tools[33]. Additionally, hydropathy indices were calculated
136 using the Kyte-Doolittle scale[34]. Positive scores on the Kyte-Doolittle scale indicate that the amino
137 acid is hydrophilic, while negative scores indicate that the amino acid is hydrophobic. A window
138 size of 9 equally weighted amino acids was used to determine position-specific hydropathy indices.
139 Based on the sequential indices, we calculated an overall score for a protein by counting the
140 number of hydrophobic or hydrophilic residues in a sliding window of 5. If all consecutive residues
141 within a window have the same tendency to be hydrophobic or hydrophilic, a 1 was added to the
142 overall score. To eliminate the influence of length on feature value, the scores were normalised by
143 the protein sequence length.
144
145 **Local structural features**
146 Structural features were represented here by alpha-helix, beta-sheet, relative solvent accessibility
147 (RSA) and disordered regions. These features were predicted by using NetSurfP 2.0 [19]. NetSurfP
148 predicts secondary structures specific to a given position. To obtain a single count for each
149 structural property, the occurrences in a sliding window of seven amino acids was counted. The
150 occurrences of 5 predicted alpha-helix residues and 3 predicted beta-sheet residues within the
151 sliding window summarised both alpha-helix and beta-sheet counts per protein. For RSA, the
152 average value in the same window size was calculated and compared to a threshold of 0.3. If the
153 threshold was exceeded, the average relative solvent accessibility was added 1. Additionally, under
154 the same window size setting, the presence of more than three residues with disorder higher than
155 0.5 was considered to contribute to the overall disorder. Likewise, these structural features were
156 normalised by the sequence length.

157

158 **Amino acid compositions and Evolutionary information**

159 The relative amino acid frequencies in the protein sequence were represented as 20 features
160 regarded as amino acid compositions (AAC). Further, the evolutionary information was
161 represented by the position-specific scoring matrix (PSSM). The PSI-Blast from the NCBI-Blast
162 2.12.0 tool packages was performed to produce the PSSM[35]. To search query sequences with PSI-
163 Blast, the sequences in the training datasets were used as a Blast database. The E-value threshold
164 was set to 0.001. We replaced the PSSMs of query sequences with no hits with the default scoring
165 matrix BLOSUM62 due to the database limit. We applied the autocovariance (AC) transformation
166 to vectorize the PSSM for a protein.

167 $$AC_{i,l} = \frac{1}{L-l}\sum_{j=1}^{L-l}(P_{i,j} - \bar{P}_i)(P_{i,j+l} - \bar{P}_i)$$

168 where i, i1, or i2 represents a specific amino acid, l is the lag to calculate the covariance, L is the
169 protein sequence length, j is the amino acid position, Pi,j is a score of amino acid i at position j in
170 the PSSM. As a result of vectorizing the PSSM, 20 autocovariance features are generated.

171 Model development

172 Clustering

173 Allergen proteins sharing sequence similarity were clustered together using Blast E-score and
174 minimum spanning tree (MST). First, each allergen in the dataset was aligned against all the other
175 allergens using BlastP (from BLAST 2.12.0) using a leave-one-out (LOO) database that consisted
176 of 685 allergens leaving the query sequence out. After comparing every pair of sequences, a
177 squared matrix of E-values is obtained. Next, the threshold E-value $1x10^{-10}$ was used to construct

178 a binary similarity matrix. If the E-value was lower than $1x10^{-10}$ it was assigned to 1, if bigger than
179 0. Based on the binary matrix, clusters were generated by the minimum spanning tree algorithm,
180 implying that all the sequences sharing similarity to at least one common neighbour (value of 1)
181 were clustered together. While the use of E-value score is dependent on sequence and database
182 length, its use is more robust and independent of the alignment length when the database is fixed
183 compared to percentage of identity.
184
185 **Redundancy reduction**
186 Redundancy in the training dataset was further reduced by removing the proteins that after
187 clustering with MST had more than 20 edges (several numbers of edges were tried to find the
188 optimal, see results section). The final positive dataset had 537 allergic proteins after removing
189 sequence redundancy.
190
191 **Data partitioning**
192 After clustering the positive allergens, all the non-allergens associated with each allergen (sharing
193 the same species and length range) were grouped together. The clusters from MST were randomly
194 distributed into five partitions with similar sizes. Ten different random seeds were used to generate
195 different partitions.
196
197 **Cross-validation**
198 A nested cross-validation was used for model training and performance evaluation. Based on the
199 data partitioning, the outer layer of nested cross-validation had five folds, and the inner layers had
200 four folds to ensure that partitions were mutually independent. Hyperparameter tuning and model
201 selection were carried out on the training and validation datasets. The selected optimal models

predicted on the outer test datasets. The inner four folds provided an ensemble prediction by an average of the four predictions. By concatenating the predictions from the five outer test sets, we obtained the final machine learning prediction results for the entire dataset.

Normalisation

In our study, we performed standardisation on our data to eliminate the effect of different feature scales introduced by average residue weights and other features. All features were transformed by z-score before model training. The z-scores were calculated based upon the means and standard deviations of each training dataset. Data in test datasets was normalised by the means and standard deviations from the training datasets which were used for model construction.

Random forest

In this study, random forest (RF) were applied for model constructions. We implemented grid searches for hyperparameter tuning to determine a set of hyperparameters that would best fit data. The hyperparameters include the number of estimators, the maximum depth of the decision tree, the minimum samples for split and leaf, and impurity criterion. We evaluated all combinations of hyperparameters in the grid and the set of hyperparameters with the best performance was selected. Ten models were generated from the ten different seeds used for partitioning. The final model was made as an ensemble of the ten different models.

Baseline model

Allergens in training datasets were used to construct the Blast database. Sequences in the outer folds were searched against these databases and the E-values were extracted as the predictions of the Blast-baseline model.

225 ## Evaluation datasets

226 To validate the models, two independent validation datasets were collected. The first one was
227 collected from the AlgPred2 validation dataset[21]. Here identical duplicates were removed leaving
228 765 allergen and 2015 non-allergen sequences. Since the publicly available AlgPred2 method was
229 retrained adding the validation dataset, this data set cannot be used to validate AlgPred2.
230 Therefore, we constructed a new validation dataset to further evaluate and benchmark against
231 AlgPred2. Here, allergens were retrieved from Allergen Nomenclature[36], Allerbase[37],
232 COMPARE[38], Allergome[39], and SDAP[10] databases (On December 2021). Allergenic protein
233 sequences that were identical to positive sequences in the training dataset were excluded. Next,
234 the same procedure to collect non-allergens in the training dataset was applied here to obtain the
235 negative dataset. This validation dataset contained 1065 positive and 4914 negative proteins.

236 ## Evaluation metrics

237 The predictive performance was evaluated by the area under the receiver operating characteristic
238 curve (AUROC or AUC), AUC 0.1 and the positive predictive value (PPV). AUC 0.1 was
239 calculated as the area integrated to 10% false positive rate, thus indicating a value of 0.1 as random
240 performance. The higher AUC or AUC 0.1, the better model performance. For the PPV, the
241 predicted values were firstly sorted in descending order, next PPV was obtained by counting the
242 number of predicted values with positive targets for the first n instances and normalised by n,
243 where n was the number of allergens.
244
245 Bootstrapping was used to statistically compare the predictions of the different models. In this
246 case, to find the significant differences between models, paired predictions from two models were

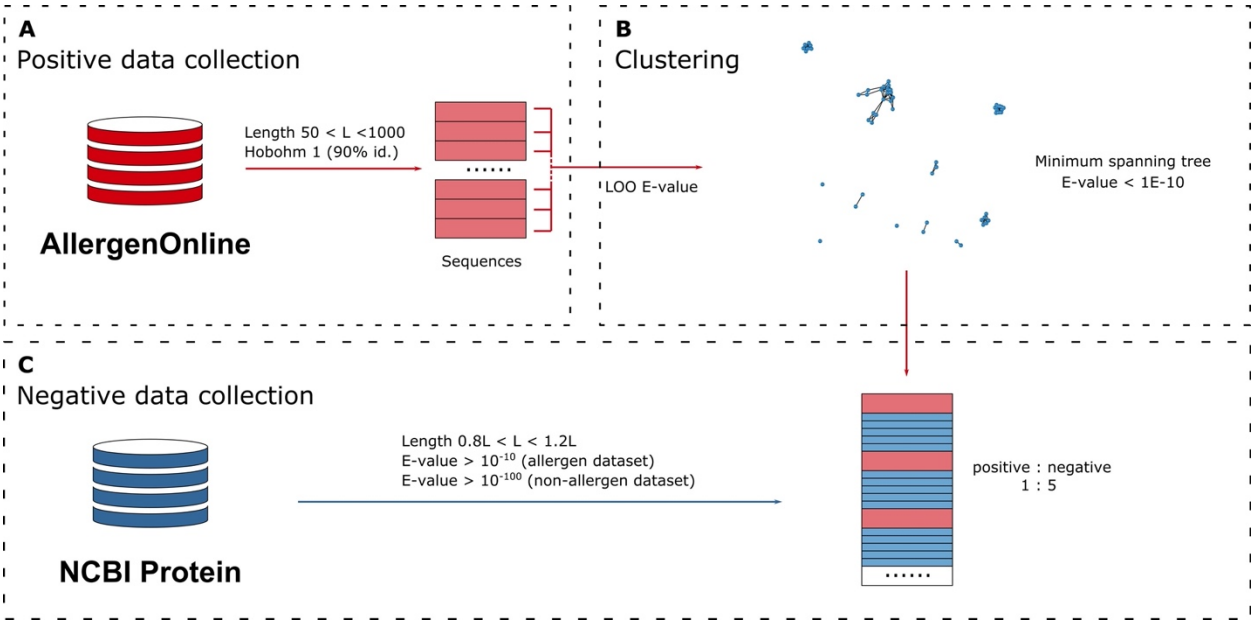
247 randomly sampled with replacement. Next the AUC of resampled data was calculated for each
248 model. This was repeated 10,000 times. In the final step, the p-value for model 1 outperforming
249 model 2 was calculated by counting the occurrences of AUC from model 2 greater than model 1
250 and divided by the number of samplings.

251 Results

252 Data redundancy reduction

253 One of the big and current challenges in machine learning approaches is data leakage, which
254 originates from data used for the model to learn, with high redundancy to what is later evaluated
255 on[40]. This issue not only over-estimates model accuracy but also curbs its generalisation and
256 therefore hinders its prediction ability. In particular, the multiple allergen databases that exist
257 contain a great number of homologous sequences or iso-allergens[37]. While highly relying in
258 sequence similarity may facilitate the task of finding homologous candidate allergens in
259 unexplored species, it might derange the effort to find new protein families or unrelated unknown
260 allergens. Hence, we first developed a pipeline to efficiently remove similar sequences from the
261 allergens dataset using Blast sequence alignment and minimum spanning tree (MST), which
262 connects into single clusters all proteins sharing sequence similarity over a threshold.
263
264 Allergen protein sequences were collected from AllergenOnline[30] and filtered to remove atypically
265 short and long sequences, highly redundant or identical sequences (detailed in Methods section,
266 Figure 1A). After this, 686 allergen sequences were clustered by MST using leave-one-out (LOO)
267 Blast sequence similarity E-values lower than $1x10^{-10}$ to define neighbours (Fig 1B). Non-allergens

268 are not trivial to obtain since they are not experimentally validated. Here, for each allergen, we
269 collected five protein sequences not annotated as allergens from the same species of the allergen
270 as described in material and methods (Fig 1C).
271



272

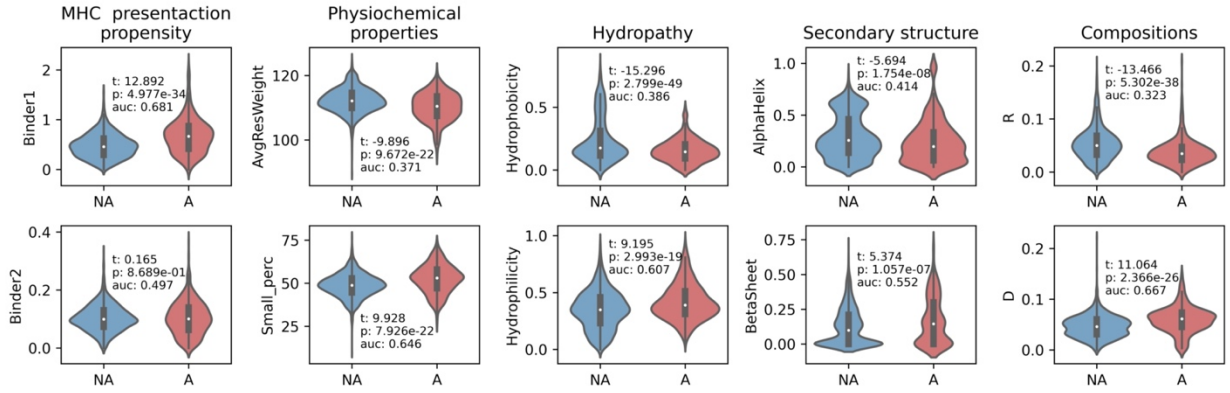
273 **Figure 1. Data collection. A.** Positive data collection. Allergens were collected from the AllergenOnline database
274 and filtered by length (50-100) and similarity (90% identity). **B.** Clustering. Clusters were generated using the
275 minimum spanning tree algorithm and distances from Blast leave one out (LOO) E-values (detailed in methods
276 section). **C.** Negative data collection. Non-allergen sequences (blue) were collected from Uniprot and added in a 5:1
277 ratio to positive allergens (red) searching for the same species in each allergen. Candidates that meet the length (range
278 20% allergen) and similarity (E-values) requirements were selected. The similarity was determined by searching
279 against the allergen dataset and non-allergen proteins that had been adopted.
280

281 ## MHC presentation propensity

282 In order to induce Th2 responses and IgE production, allergens are first digested by antigen
283 presenting cells and selected peptides are presented on MHC-II complexes. Therefore, we

284 hypothesised that allergens must contain MHC-II presentable peptides to induce allergy. Recently,
285 GWAS studies of patients with allergic rhinitis, wheat allergy and other food allergies have shown
286 strong association to specific MHC class II alleles, including HLA-DR and HLA-DQ[23,24,26].
287 Therefore, NetMHCIIpan was used to predict MHC presentation potential for allergens and non-
288 allergens on a set of selected alleles[41]. MHC-II presentation propensity was collected by adding
289 the counts of predicted binders for a set of relevant alleles derived from genetic associations found
290 in allergies. In particular two features were created grouping all positive and negative allele
291 associations called binder 1 (including HLA-DRB1*04:01, HLA-DQA1*02:01-DQB1*02:02,
292 HLA-DQA1*04:01-DQB1*03:01, HLA-DQA1*01:03-DQB1*06:01, HLA-DQA1*03:01-
293 DQB1*03:02, HLA-DQA1*05:01-DQB1*02:01), and binder 2 (including HLA-DRB1*15:01)
294 respectively.
295
296 In addition to MHC-II presentation propensity, previously described classification features for
297 allergen/non-allergens were included (for details see methods). Those include composition of
298 amino acids, physicochemical properties, structural features (alpha helix, beta sheets, coiled
299 regions and relative solvent accessibility), and evolutionary properties. Feature distributions for
300 allergen and non-allergens were illustrated by violin plots for the most significant features
301 displaying the median, quantiles, and densities of each set of data (Figure 2). T-statistic and p-
302 values were calculated to statistically investigate the feature differences. For binder 1, the t-statistic
303 and p-value were 12.892 and 4.97×10^{-34} respectively, indicating a higher MHC-II presentation
304 propensity on allergens compared to non-allergens. Differences on binder 2 were not statistically
305 significant, suggesting that HLA-DRB1*15:01 presentation propensity was less discriminative for
306 this dataset. When comparing the average residue weight, the t-test had a negative t-statistic value

307 at the significant level. These results corresponded to the composition of small-size amino acids
308 and suggested that allergens are characterised by a higher proportion of small-size amino acids.
309 Likewise, allergens were found to be more likely hydrophilic and less hydrophobic, as previously
310 described[7], and containing less arginine and more aspartic acid. The t-statistic for alpha-helix was
311 negative, and beta-sheet positive which indicated the allergens are depleted in alpha-helices and
312 enriched to beta-sheet compared to non-allergens confirming previous observations[42].
313
314 We next determined the AUC of each single feature to estimate how well each feature
315 discriminated when used as an individual prediction. Here, AUC values below 0.5 would suggest
316 that allergic proteins have lower feature values than non-allergic proteins. As expected, the results
317 of these analyses aligned with the t-statistics above, i.e. allergens had more MHC-II binder1, more
318 small-size amino acids, less alpha-helices, more beta sheets, more hydrophilic proteins and
319 containing more aspartic acid and less Arginine (Figure 2).
320



321
322 **Figure 2. Distribution and individual-feature performances on selected features for allergens and non-**
323 **allergens.** T-test was used to compare the significant difference between allergens (red, A) and non-allergens (blue,
324 NA). The AUC was calculated with feature values and the targets. "small_perc" corresponds to the relative abundance

325 of small-size amino acids (A, C, D, G, N, P, S, T, V) in a given protein sequence. Likewise, D and R correspond to
326 the frequencies of Aspartic acid and Arginine per protein.
327

328 Cross validation performances

329 After showing that MHC presentation propensity is a discriminatory feature for allergens, we
330 sought to incorporate this feature into machine learning models in order to improve state of the art
331 allergenicity predictors. Each allergen set from MST clustering and its negative associated
332 sequences were kept together, and the clusters were randomly distributed into five partitions
333 (Figure 3). A five-fold nested cross-validation scheme was performed, and the best random forest
334 hyperparameters were selected in the inner loop to avoid overfitting (Fig 3B). MHC presentation
335 propensity was combined with different sets of features, optimal performance on the training data
336 was obtained with a random forest model with 60 features (60F) (Supp. Table 2). The 60F model
337 included MHC presentation propensity, physicochemical features, amino acid compositions, auto-
338 covariances and structural features (for details refer to Material and Methods section). Finally, the
339 partitions from MST clusters were repeated ten times using different random seeds and a final
340 random forest model ensemble was generated. The ensemble outperformed significantly the
341 individual models in each random seed (Fig 3C).
342

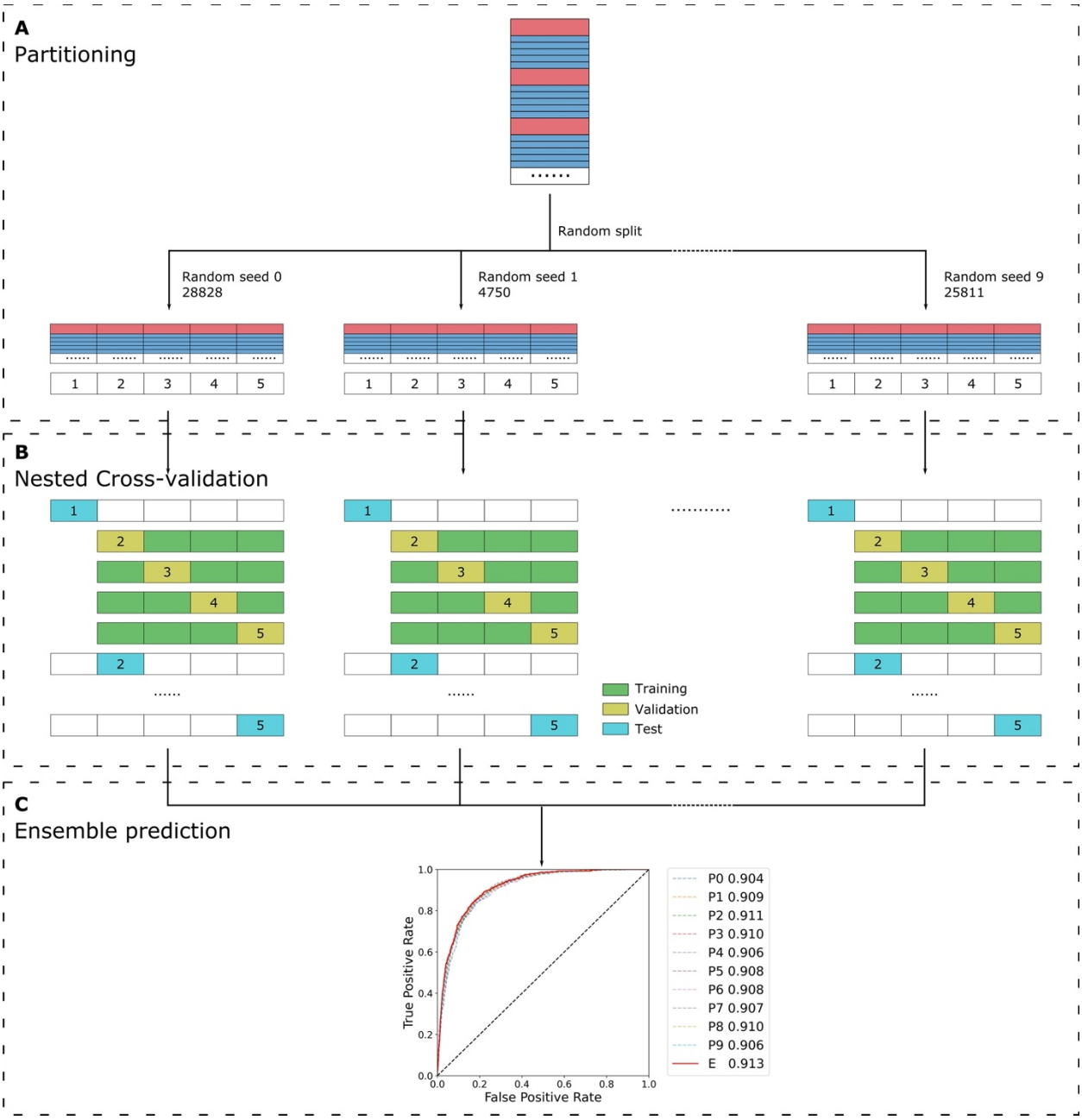


Figure 3. Model development pipeline of NetAllergen. A. Partitioning. Five partitions randomly combining the clusters obtained from minimum spanning tree (MST) clustering were distributed using ten different random seeds. Every allergen (red bar) and its five associated negatives (blue bars) share the same partition. **B.** Cross Validation. A nested cross-validation setup was used using five folds in the outer layer and four folds in the inner layer. **C.** Ensemble prediction. AUC Prediction performance of the individual models for each seed and the ensemble performance based on the average of the ten model predictions. Ensemble AUC was significantly higher than the individual partitions (P, p-value < 0.05) except when comparing to P3, which it was not significant (p-value 0.911).

Since we have previously observed that data redundancy in the training dataset greatly influences the trained model partitions (data not shown), we assessed further removing clusters which contained highly similar sequences to further cut-out redundancy in the dataset. Cross-validation predictive performance was evaluated to find the optimal filter for removing the allergens with a higher number of neighbours in the tree network after MST clustering (Figure 4). The removal of highly connected vertices in the tree reduced the bias effect of highly similar sequences and improved the model performance. However, when applying more stringent filters (removing vertices with fewer than 20 neighbours) caused a loss of informative allergen sequences as well as the performance. Hence, we applied a filter of 20 and removed the sequences that had more than 20 neighbours in the MST clusters.

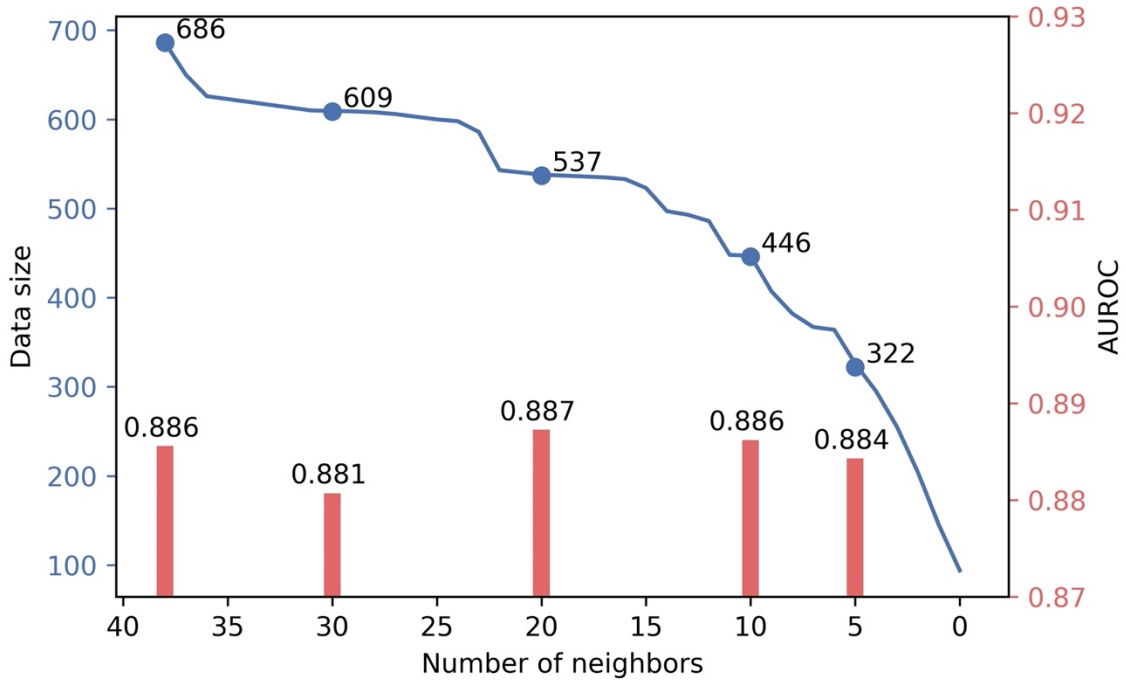


Figure 4. AUC performances and data sizes for models with different pruning filters for the number of

365 **neighbours.** The AUROC was evaluated on the minimal common subset (the dataset with a filter of 5).

366

367 MHC presentation propensity is amongst the most important features

368 The advantage of simple machine learning (ML) models such as random forest is their ease of
369 feature importance interpretation. Here, we calculated feature importance as the mean decrease of
370 Gini impurity for each specific feature in all the trees. A higher value indicates a better split of the
371 data using that feature. We explored the contribution of the different features in the random forest
372 model of 60F and in a reduced model of 20F (Performance shown in Supp. Table 2) only including;
373 MHC presentation propensity (MHC), structural properties (NSP), and physicochemical features
374 (PCP). MHC II presentation propensity binder 1 was the most important feature in the simpler 20F
375 model (Fig 5A). The frequency of small amino acids (Small), average relative solvent accessibility
376 (AvgRSA), binder 2, acidic amino acids, and alpha helix and beta sheet were also informative
377 features to discriminate allergens and non-allergens. In the 60F model, although the binder 1 was
378 no longer the most important feature, it still played an important role in the prediction (Fig 5B).
379 The binder 1 feature thus showed consistency of importance in both the 20F and the 60F model.
380 In contrast, the binder 2 feature became less important in the 60F model. Further, the amino acid
381 composition of arginine, aspartic acid, and lysine, together with the auto-covariance of PSSM on
382 Glutamine (AC_Q) were found to be among the most discriminatory features in the 60F model.
383 The difference in feature importance between shared features of the 60F and 20F models may
384 indicate feature correlations on the amino acid composition features.

385

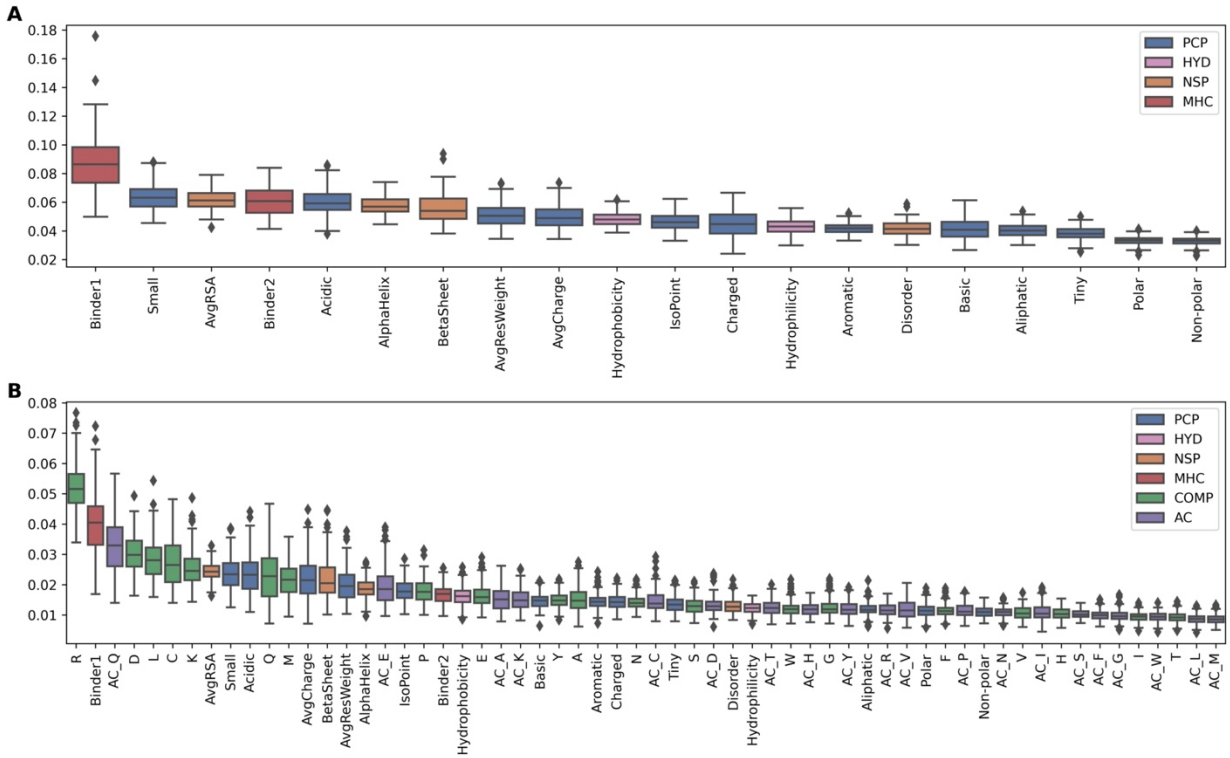


Figure 5. Median feature importance for random forest (RF) models. Medians are calculated from cross-validation partitions and 10 random seeds in each random forest as the mean decrease of Gini impurity for **A.** 20 features (20F) model, and **B.** 60F model. Individual features were grouped into physicochemical properties (PCP, blue), hydropathy (HYD, pink), structural information (NSP, yellow), MHC-II presentation propensity (MHC, red), amino acid compositions (COMP, green), and evolutionary information (AC, purple). Average relative solvent accessibility (AvgRSA).

To further investigate the importance of MHC II features, models removing the binder 1 and binder 2 were trained and evaluated using cross-validation. When the MHC II features were removed from the 60F model (58F model), AUC was reduced from 0.913 to 0.911 and AUC 0.1 from 0.509 to 0.505 (Table 1). Only the AUC difference of 58F and 60F was however found to be statistically significant. For the 20F model however, when taking out the MHC presentation propensity features (18F), the reduction in prediction power was found to be significant for all performance metrics (AUC, AUC 0.1 and PPV) (Table 1). Consistently with feature importance in those models, this

401 result might indicate that the addition of amino acid compositions and autocovariances could be
402 correlated with MHC presentation propensity.
403
404 **Table 1. Predictive performance and significant difference of models removing MHC II features**
405

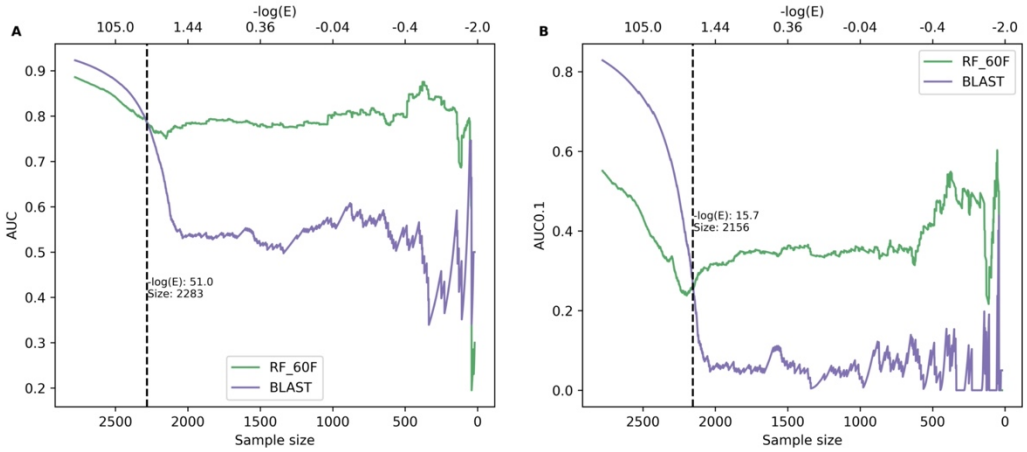
Model	AUC	p-value	AUC 0.1	p-value	PPV	p-value
58F	0.911	0.022	0.505	0.288	0.657	0.612
60F	0.913		0.509		0.646	
18F	0.858	< 0.001	0.317	0.003	0.516	0.016
20F	0.869		0.348		0.544	

406 Note: All p-values were obtained using bootstrapping.
407

408 Model comparison

409 Sequence similarity to previously known allergens is the main driver of previously developed
410 allergenicity predictors[5,20,22]. Therefore, we have constructed a sequence similarity baseline model
411 using the Blast algorithm. In the Blast-baseline model, the predicted value for each potential
412 allergen from the validation dataset was taken from the best hit in the training dataset. This baseline
413 model was compared to the random forest model of 60F on the AlgPred2 independent data set (see
414 Methods). Overall, the Blast-baseline model had a higher performance (AUC of 0.923) on the full
415 validation dataset compared to the random forest of 60F (AUC 0.886, Supp Table 3). This very
416 high performance of the Blast-baseline model clearly suggests the high similarity of the allergens
417 in the validation dataset towards allergens in the training data set. However, it is expected that the
418 Blast-baseline model will decrease performance on unknown or unrelated new allergen sequences.
419 To investigate this and quantify how model-performance correlates with sequence similarity to the
420 training dataset, we sorted the validation dataset from high to low similarity. Here, similarity is

421 defined as the lowest E-value between the query protein and the positive training data-set. We
422 observed that the Blast-baseline model performed slightly better than our 60F model for highly
423 query entries with high similarity to the training data, but that the performance drops dramatically
424 when sequence similarity decreases (Figure 6). On the contrary, the random forest model (RF_60F)
425 demonstrated a more consistent AUC performance on all the dataset and outperformed Blast on
426 proteins that were less similar (E-value > $1x10^{-50}$) to sequences in the allergen database (Figure
427 6A). The same trend can be observed for AUC 0.1 (Figure 6B).
428



429
430 **Figure 6. Model performance for variable datasets with decreasing similarity thresholds on AlgPred2**
431 **validation dataset.** The sequences were sorted by similarities against allergens in the training dataset with a
432 descending order. The vertical line indicated the changes of performance and split the curves into areas of higher and
433 lower similarities. For the rightmost curve, the number of sequences reduced as the decreasing similarity, which
434 resulted in that the evaluation was less robust because of noise. Blast-baseline model (BLAST). **A.** AUC. **B.** AUC 0.1.
435

436 In order to compare to state of the art allergenicity predictors, we benchmark the random forest
437 model against AlgPred2[21]. Since AlgPred2 was available in a re-trained format using the validation
438 dataset used in their publication, we collected a second and independent validation dataset of

allergens that were published after the AlgPred2 method was retrained (detailed information in Methods). Due to a temporary inaccessibility of NetSurfP 2.0, the full feature space could not be constructed. Thus, we benchmarked against AlgPred2 and Blast, a reduced 56F model by removing the four secondary structure features. Removing the structural features displayed a significant drop in performance (Table 2). As in the earlier benchmark, the new independent validation set was ranked from higher to lower similarity Blast E-values against the positive sequences in the combined training data sets of the RF and AlgPred2 models. Performance evaluation was measured in a similar fashion as before for the baseline model. Both AUC and AUC 0.1 performances behaved similarly as before, showing a higher performance of AlgPred2 and Blast for the highly similar sequences (Figure 7). In a similar fashion both AlgPred2 and Blast performances dropped below that of the RF_56 model when similarity to training-allergens E-values were higher than $1x10^{-42}$. This result suggests the 56F model had superior predictive power on novel sequences which were not included in current allergen databases.

Table 2. Cross validation performance of models with (60F) or without structural features (56F)

Model	AUC	p-value	AUC 0.1	p-value	PPV	p-value
RF_60F	0.913	< 0.001	0.509	< 0.001	0.646	0.066
RF_56F	0.903		0.481		0.639	

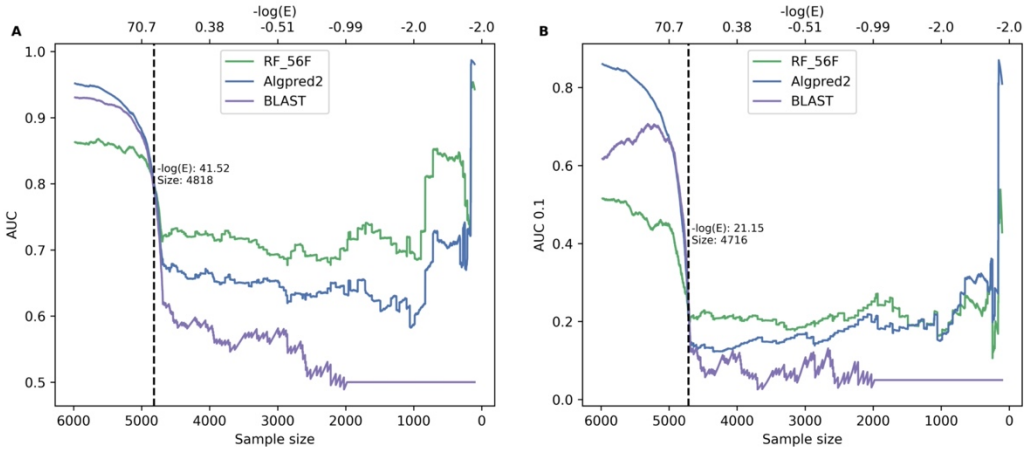


Figure 7. Model performance with different similarity thresholds on the new validation dataset. The similarities were represented by -log(E-value) which were obtained from the Blast-baseline model(BLAST). The searching database consisted of the positive sequences from our and AlgPred2 dataset. The noise in decreasing data size led to the less robust curves at the rightmost part. **A.** AUC. **B.** AUC0.1.

Web server

The 56F random forest model is deployed as a web server considering the inaccessibility and computational cost of NetSurfP 2.0. The allergenicity predictor is available at https://services.healthtech.dtu.dk/service.php?NetAllergen-1.0. The program accepts a fasta file of protein sequences and provides a prediction result table (Figure 8).

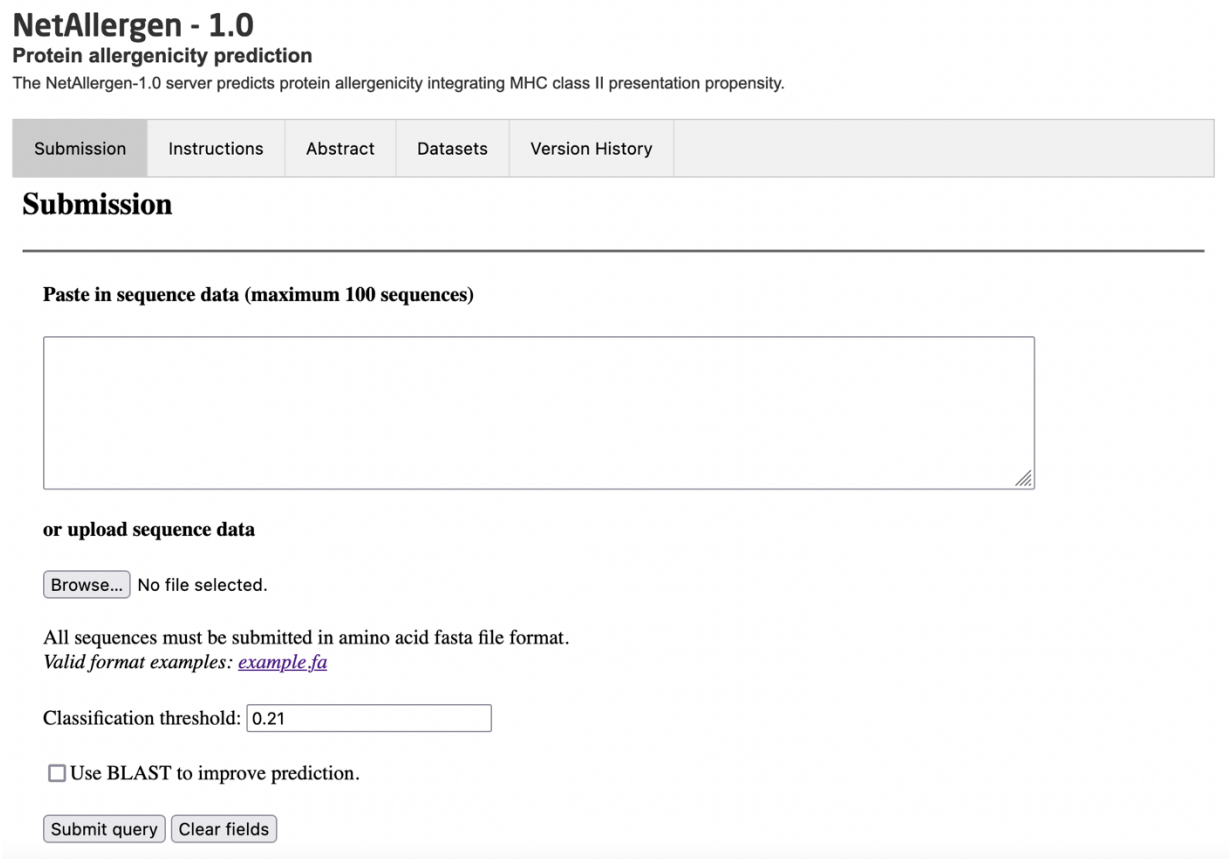


Figure 8. Web service of NetAllergen-1.0.

Discussion

Understanding the nature of allergenicity is of high importance as it contributes to the development of future diagnostic tools and therapeutic strategies. Previous allergen predictors have focused on answering the question of what makes an allergen. We have here proposed a first attempt of answering the question of why only some individuals develop allergies towards ubiquitous proteins. The answer may relate to presentation on particular HLA class II alleles.

477 Here, we have developed a new allergen predictor called NetAllergen which incorporates a
478 significant novel feature regarding MHC class II presentation. NetAllergen was trained on a
479 carefully curated dataset where redundant allergen sequences were removed. The method was
480 demonstrated to perform better than a state-of-the-art predictor AlgPred2, and a Blast-baseline
481 model when sequence similarity towards known allergens is reduced.

482 Although it has long been recognised that allergy has a strong HLA association component[43], still
483 very little is known about the mechanisms that trigger allergy. Recently, large genome-wide
484 association studies pointed to conserved specific amino acid positions in the HLA-II binding
485 groove determining protective or risk variants for certain allergies[23,24]. This observation suggests
486 that MHC-II presentation in those alleles might have an important role in allergies. We have
487 therefore hypothesised that allergen protein sequences might have an increased number of
488 presentation motifs towards those specific alleles compared to non-allergens. Here, we showed
489 that this is indeed the case. HLA presentation propensity to binder 1 alleles (HLA-DRB1*04:01,
490 HLA-DQA1*02:01-DQB1*02:02, HLA-DQA1*04:01-DQB1*03:01, HLA-DQA1*01:03-
491 DQB1*06:01, HLA-DQA1*03:01-DQB1*03:02, and HLA-DQA1*05:01-DQB1*02:01) was the
492 single stronger discriminant feature for allergens (AUC=0.68), compared to previously reported
493 hydrophilicity (AUC=0.61) or beta sheets (AUC=0.55). In addition, incorporation of this novel
494 feature into random forest models boosted performance for allergenicity prediction.
495

496 The advantage of random forest compared to other more complex models in machine learning is
497 its ease of interpretability. Therefore, we have picked this simple model to better analyse feature
498 importance. Also, here, MHC presentation propensity binder 1 was found to be one of the most
499 important features when analysing which features best split the data in the random forest.

500 Moreover, removing MHC class II presentation features from the model (18F), significantly
501 reduced performance compared to a model including those features (20F) in several performance
502 measures (Table 1). This suggested that MHC-II presentation is an informative classification
503 allergen feature. However, when incorporating additional features such as amino acid
504 compositions, and evolutionary information of the amino acids, those improvements were only
505 moderate). In the 60F models, Arginine frequency was the most important feature followed by
506 MHC-II presentation binder 1, auto-covariance of glutamine (AC_Q), and amino acid
507 compositions of Aspartic acid, Leucine, Cysteine and Lysine. While MHC-II presentation features
508 are based on the count of presentation motifs to specific MHC II alleles, the amino acid
509 compositions count the percentage of amino acids in a protein sequence. They both capture
510 important amino acids information in two distinct approaches, explaining why the 60F model did
511 not show very significant improvements against the 58F model.
512
513 In addition to the MHC II binders, the exploration revealed other important features. The average
514 residue weight was negatively correlated to allergenicity and small amino acids composition (A,
515 C, D, G, N, P, S, T, V) had positive correlation. In particular, Threonine and Valine are branched
516 on the beta-carbon and thus are more often found on beta strands than in alpha helices. On the
517 other hand, Proline is also disfavoured on alpha helices as its imino ring is not able to adopt normal
518 helical main chain conformations[44]. These findings align with previous observations also found
519 here that allergens have a higher abundance of beta sheets and less alpha helices compared to non-
520 allergens[42].
521

522 Other properties, such as acidic amino acids, were significant but lower than some single amino
523 acid compositions (such as AC_Q, R or D). Compositions related to Arginine and Aspartic acid
524 tended to have respectively a negative and positive correlation to allergy (Figure 2), in agreement
525 with more acid amino acid compositions. Auto-covariance (AC) represents the average covariance
526 between the sequence and itself. Hence, the coefficients of AC_Q indicated that Glutamine might
527 be present in a large number of repetitions in the sequence. This information suggested by AC of
528 Glutamine was similar to the Q-repeats features used in AllerCatPro, thus aligning with previous
529 findings for allergens[20].
530
531 Many important applications in bioinformatics, including sequence alignment and machine
532 learning algorithms, employ sequence weighting schemes to mitigate the effects of non-
533 independent homologous sequences over-represented in a dataset[45,46]. In particular, allergen
534 datasets are highly redundant as many homologous proteins are discovered in closely related
535 species. Moreover, strict sequence similarity guidelines proposed by the FAO/WHO to assess
536 allergenicity in-silico may have favoured the discovery of iso-allergens. As part of the
537 development of NetAllergen, we have proposed a novel pipeline to deal with highly redundant
538 datasets and avoid data leakage issues. Here we have showcased how this pipeline can be applied
539 to real-life datasets aiding the development of data-driven prediction methods. We believe this
540 proposed pipeline to be broadly applicable and expect it to be of general use for the community.
541

Key points

- NetAllergen-1.0, a random forest allergenicity predictor, incorporates novel MHC II presentation features outperforming state-of-the-art classifiers for allergens that share less similarities with known allergens.
- Here we developed a new sequence redundancy reduction pipeline to prevent data leakage between training and test and ensure the model could effectively learn patterns from less similar protein sequences.
- MHC class II presentation propensity of allergen-specific HLA II alleles is highly associated with protein allergenicity.
- Feature exploration in the random forest reveals the importance of physicochemical properties, secondary structures and motifs for differentiating allergens from non-allergens.

Data Availability Statement

The data used in this manuscript is available at the web-server link. https://services.healthtech.dtu.dk/service.php?NetAllergen-1.0.

Authorship

YL collected the data, developed the pipeline, trained the models, analysed and interpreted the data; MN analysed and interpreted the data, provided critical revision of the article; and CB

analysed and interpreted the data, provided critical revision of the article, and conceived the work. All authors wrote the paper and approved the final version to be published.

Competing interest

All authors declare no competing interest.

566 # References

567 1. Yang, L., Fu, J. & Zhou, Y. Research Progress in Atopic March. *Front Immunol* **11**, 1907 (2020).

568 2. Cianferoni, A. Non-IgE Mediated Food Allergy. *Curr Pediatr Rev* **16**, 95–105 (2020).

569 3. Gupta, R. *et al.* The economic impact of childhood food allergy in the United States. *JAMA Pediatr*
570 **167**, 1026–1031 (2013).

571 4. Sicherer, S. H. & Sampson, H. A. Food allergy: A review and update on epidemiology, pathogenesis,
572 diagnosis, prevention, and management. *J Allergy Clin Immunol* **141**, 41–58 (2018).

573 5. European Food Safety Authority. Workshop on allergenicity assessment – prediction. *EFS3* **18**,
574 (2021).

575 6. Yu, W., Freeland, D. M. H. & Nadeau, K. C. Food allergy: immune mechanisms, diagnosis and
576 immunotherapy. *Nat Rev Immunol* **16**, 751–765 (2016).

577 7. Scheurer, S., Toda, M. & Vieths, S. What makes an allergen? *Clin Exp Allergy* **45**, 1150–1161
578 (2015).

579 8. du Toit, G., Tsakok, T., Lack, S. & Lack, G. Prevention of food allergy. *J Allergy Clin Immunol* **137**,
580 998–1010 (2016).

581 9. Ivanciuc, O., Schein, C. H. & Braun, W. Data mining of sequences and 3D structures of allergenic
582 proteins. *Bioinformatics* **18**, 1358–1364 (2002).

583 10. Ivanciuc, O., Schein, C. H. & Braun, W. SDAP: database and computational tools for allergenic
584 proteins. *Nucleic Acids Res* **31**, 359–362 (2003).

585 11. Fiers, M. W. E. J. *et al.* Allermatch, a webtool for the prediction of potential allergenicity according to
586 current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics* **5**, 133 (2004).

587 12. Zhang, Z. H. *et al.* AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in
588 proteins. *Bioinformatics* **23**, 504–506 (2007).

589 13. Riaz, T., Hor, H. L., Krishnan, A., Tang, F. & Li, K.-B. WebAllergen: a web server for predicting
590 allergenic proteins. *Bioinformatics* **21**, 2570–2571 (2005).

591 14. Saha, S. & Raghava, G. P. S. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes.
592 *Nucleic Acids Res* **34**, W202-209 (2006).

593 15. Muh, H. C., Tong, J. C. & Tammi, M. T. AllerHunter: A SVM-Pairwise System for Assessment of
594 Allergenicity and Allergic Cross-Reactivity in Proteins. *PLoS ONE* **4**, e5861 (2009).
595 16. Dang, H. X. & Lawrence, C. B. Allerdictor: fast allergen prediction using text classification techniques.
596 *Bioinformatics* **30**, 1120–1128 (2014).
597 17. Dimitrov, I., Flower, D. R. & Doytchinova, I. AllerTOP--a server for in silico prediction of allergens.
598 *BMC Bioinformatics* **14 Suppl 6**, S4 (2013).
599 18. Dimitrov, I., Bangov, I., Flower, D. R. & Doytchinova, I. AllerTOP v.2--a server for in silico prediction
600 of allergens. *J Mol Model* **20**, 2278 (2014).
601 19. Dimitrov, I., Naneva, L., Doytchinova, I. & Bangov, I. AllergenFP: allergenicity prediction by descriptor
602 fingerprints. *Bioinformatics* **30**, 846–851 (2014).
603 20. Maurer-Stroh, S. *et al.* AllerCatPro-prediction of protein allergenicity potential from the protein
604 sequence. *Bioinformatics* **35**, 3020–3027 (2019).
605 21. Sharma, N. *et al.* AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of
606 IgE epitopes. *Brief Bioinform* **22**, bbaa294 (2021).
607 22. Pomés, A., Mueller, G. A. & Chruszcz, M. Structural Aspects of the Allergen-Antibody Interaction.
608 *Front Immunol* **11**, 2067 (2020).
609 23. Noguchi, E. *et al.* HLA-DQ and RBFOX1 as susceptibility genes for an outbreak of hydrolyzed wheat
610 allergy. *Journal of Allergy and Clinical Immunology* **144**, 1354–1363 (2019).
611 24. The 23andMe Research Team *et al.* Genome-wide association and HLA fine-mapping studies
612 identify risk loci and genetic pathways underlying allergic rhinitis. *Nat Genet* **50**, 1072–1080 (2018).
613 25. Marenholz, I. *et al.* Genome-wide association study identifies the SERPINB gene cluster as a
614 susceptibility locus for food allergy. *Nat Commun* **8**, 1056 (2017).
615 26. Moffatt, M. F. *et al.* Association between quantitative traits underlying asthma and the HLA-DRB1
616 locus in a family-based population sample. *Eur J Hum Genet* **9**, 341–346 (2001).
617 27. Van Overtvelt, L. *et al.* Assessment of Bet v 1-Specific CD4 $^{+}$ T Cell Responses in Allergic and
618 Nonallergic Individuals Using MHC Class II Peptide Tetramers. *J Immunol* **180**, 4514–4522 (2008).
619 28. Sakaguchi, M. *et al.* Analysis of HLA in children with gelatin allergy: Sakaguchi et al : HLA in children
620 with gelatin allergy. *Tissue Antigens* **59**, 412–416 (2002).

621 29. Dimitrov, I. & Doytchinova, I. Associations between Milk and Egg Allergens and the HLA-DRB1/DQ
622 Polymorphism: A Bioinformatics Approach. *Int Arch Allergy Immunol* **169**, 33–39 (2016).
623 30. Goodman, R. E. *et al.* AllergenOnline: A peer-reviewed, curated allergen database to assess novel
624 food proteins for potential cross-reactivity. *Mol Nutr Food Res* **60**, 1183–1198 (2016).
625 31. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of representative protein data sets.
626 *Protein Sci* **1**, 409–417 (1992).
627 32. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0:
628 improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration
629 of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
630 33. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids*
631 *Res* **47**, W636–W641 (2019).
632 34. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J Mol*
633 *Biol* **157**, 105–132 (1982).
634 35. Klausen, M. S. *et al.* NetSurfP-2.0: Improved prediction of protein structural features by integrated
635 deep learning. *Proteins* **87**, 520–527 (2019).
636 36. Pomés, A. *et al.* WHO/IUIS Allergen Nomenclature: Providing a common language. *Mol Immunol*
637 **100**, 3–13 (2018).
638 37. Kadam, K., Karbhal, R., Jayaraman, V. K., Sawant, S. & Kulkarni-Kale, U. AllerBase: a
639 comprehensive allergen knowledgebase. *Database (Oxford)* **2017**, (2017).
640 38. van Ree, R. *et al.* The COMPARE Database: A Public Resource for Allergen Identification, Adapted
641 for Continuous Improvement. *Front. Allergy* **2**, 700533 (2021).
642 39. Mari, A. & Scala, E. Allergome: a unifying platform. *Arb Paul Ehrlich Inst Bundesamt Sera Impfstoffe*
643 *Frankf A M* 29–39; discussion 39-40 (2006).
644 40. Gibney, E. Could machine learning fuel a reproducibility crisis in science? *Nature* d41586-022-02035-
645 w (2022) doi:10.1038/d41586-022-02035-w.
646 41. Reynisson, B. *et al.* Improved Prediction of MHC II Antigen Presentation through Integration and
647 Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. *J. Proteome Res.* **19**, 2304–
648 2315 (2020).

649 42. Dall'antonia, F., Pavkov-Keller, T., Zangger, K. & Keller, W. Structure of allergens and structure
650 based epitope predictions. *Methods* **66**, 3–21 (2014).
651 43. Levine, B. B., Stember, R. H. & Fotino, M. Ragweed hay fever: genetic control and linkage to HL-A
652 haplotypes. *Science* **178**, 1201–1203 (1972).
653 44. Betts, M. J. & Russell, R. B. Amino Acid Properties and Consequences of Substitutions. in
654 *Bioinformatics for Geneticists* (eds. Barnes, M. R. & Gray, I. C.) 289–316 (John Wiley & Sons, Ltd,
655 2003). doi:10.1002/0470867302.ch14.
656 45. De Maio, N. *et al.* A phylogenetic approach for weighting genetic sequences. *BMC Bioinformatics* **22**,
657 285 (2021).
658 46. Nielsen, H., Tsirigos, K. D., Brunak, S. & von Heijne, G. A Brief History of Protein Sorting Prediction.
659 *Protein J* **38**, 200–216 (2019).
660