Computers in Biology and Medicine 136 (2021) 104746

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

ChAlPred: A web server for prediction of allergenicity of chemical compounds

Neelam Sharma, Sumeet Patiyal, Anjali Dhall, Naorem Leimarembi Devi, Gajendra P. S. Raghava [*]

Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi, 110020, India

ARTICLE INFO

Keywords:
Chemical allergens
Allergy
Machine learning
PaDEL
FDA-approved drugs
DrugBank
IEDB
ChEBI
Chemical descriptors
Fingerprints

ABSTRACT

Background: Allergy is the abrupt reaction of the immune system that may occur after the exposure to allergens such as proteins, peptides, or chemicals. In the past, various methods have been generated for predicting allergenicity of proteins and peptides. In contrast, there is no method that can predict allergenic potential of chemicals. In this paper, we described a method ChAlPred developed for predicting chemical allergens as well as for designing chemical analogs with desired allergenicity.
Method: In this study, we have used 403 allergenic and 1074 non-allergenic chemical compounds obtained from IEDB database. The PaDEL software was used to compute the molecular descriptors of the chemical compounds to develop different prediction models. All the models were trained and tested on the 80% training data and evaluated on the 20% validation data using the 2D, 3D and FP descriptors.
Results: In this study, we have developed different prediction models using several machine learning approaches. It was observed that the Random Forest based model developed using hybrid descriptors performed the best, and achieved the maximum accuracy of 83.39% and AUC of 0.93 on validation dataset. The fingerprint analysis of the dataset indicates that certain chemical fingerprints are more abundant in allergens that include PubChemFP129 and GraphFP1014. We have also predicted allergenicity potential of FDA-approved drugs using our best model and identified the drugs causing allergic symptoms (e.g., Cefuroxime, Spironolactone, Tioconazole). Our results agreed with allergenicity of these drugs reported in literature.
Conclusions: To aid the research community, we developed a smart-device compatible web server ChAlPred (https://webs.iiitd.edu.in/raghava/chalpred/) that allows to predict and design the chemicals with allergenic properties.

1. Introduction

Allergy is an inappropriate reaction of the immune response when it misidentifies a harmless foreign substance as a threat [1–4]. These foreign substances are known as allergens, which could trigger several allergic reactions and lead to various allergic diseases. Different types of aeroallergens (e.g., pollens, spores, dust mites), food allergens (e.g., eggs, peanuts, tree nuts, genetically modified foods), and chemical allergens in personal care products (e.g., fragrances in the skin and hair care products, dyes, creams) [1,5,6] can lead to allergic symptoms such as allergic asthma, rhinitis, skin reactions and anaphylaxis [1,7]. Anaphylactic shock involves a series of allergic reactions from mild symptoms like itchy skin, rashes, facial swelling, irritation of the eyes leading to watery eyes and nose to severe symptoms like shortness of breath, lack of consciousness, weak pulse, nausea, vomiting, which can even lead to death if untreated [8,9]. Certain studies show that allergic diseases are much more prevalent in developed countries than in developing countries [10–13]. In the last few years, the increment in the occurrence of allergic diseases have increased the expenses of the treatment and also negatively influenced the status of life of a huge population [14].

There is a wide variety of molecules that can pose a threat as allergens. It includes chemical molecules [15], macromolecules such as proteins/peptides [4,16], lipids [17], carbohydrates [18], nucleic acid (mRNA vaccines) [19] and some engineered nanoparticles [20]. These molecules can stimulate some allergic reactions like asthma [17], food

[*] Corresponding author.
E-mail addresses: neelams@iiitd.ac.in (N. Sharma), sumeetp@iiitd.ac.in (S. Patiyal), anjalid@iiitd.ac.in (A. Dhall), leimarembi@gmail.com (N.L. Devi), raghava@iiitd.ac.in (G.P.S. Raghava).

https://doi.org/10.1016/j.compbiomed.2021.104746
Received 26 May 2021; Received in revised form 4 August 2021; Accepted 4 August 2021
Available online 8 August 2021

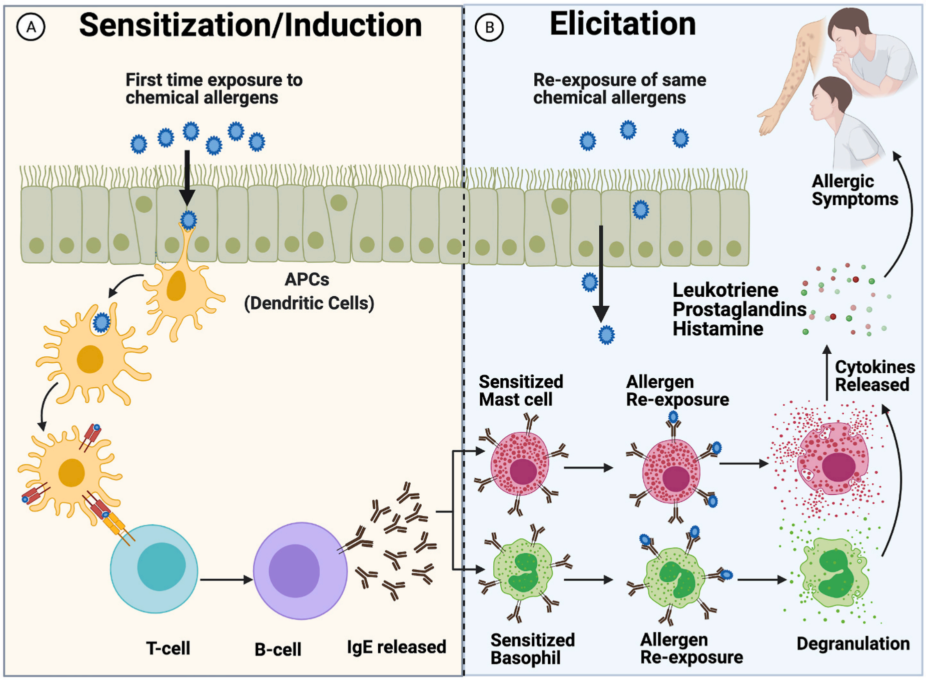

Fig. 1. The mechanism of the allergy caused by chemical allergens.

allergies [18], anaphylaxis [19] and chronic kidney diseases [20]. In the past, several methods have been proposed to predict protein allergens from genetically engineered foods, vaccines and therapeutics. All the available methods are based on protein/peptide allergens, such as the recently developed method AlgPred 2.0 [1,21]. Many other methods such as AllerTool [3], AllerHunter [22], AllerTOP [23], AllerTOPv2 [2], PREAL [24], AllergenFP [25], AllerCatPro [5] are heavily used by the scientific community. These methods are especially used in clinical researcher for designing proteins with desired allergenicity. In contrast, there is no method for predicting allergenic potential of the chemicals. Despite the fact, day-to-day life, the human body is exposed to innumerable chemical substances, such as makeup, soaps, perfumes, lotions, hair dyes, preservatives in food, metals in the jewellery [26]. Many of these chemical products are known to provoke allergic reactions, causing skin sensitization in some people, which results in skin or contact dermatitis, and some may cause the sensitization of the respiratory tract leading to occupational asthma, which could be lethal [7,8].

Thus it is important to understand the allergenicity of chemicals in order to develop prediction methods. Broadly, allergic reaction caused by small chemical compounds is developed in two phases; sensitization and elicitation. The first phase is initiated when a sensitized individual is exposed to a chemical allergen in sufficient amount, and via a proper route, then it will lead to immunological priming. In the context of allergy, immunological priming is called as sensitization or induction, which means that the mast cells and basophils are loaded with IgE antibodies against the chemical allergen [1,27,28]. In the second phase, the re-exposure to the same chemical compound at the same or different site will provoke an accelerated and more aggressive secondary immune response. This secondary immune response is called as elicitation, which results in an allergic reaction. The already sensitized mast cells and basophils result in releasing cytoplasmic granules, and inflammatory molecules, such as, leukotriene, prostaglandins, histamine etc., leading to a mild allergic reaction to sudden death from anaphylactic shock [1,27,28]. The mechanism of allergy caused by chemical allergens is depicted in Fig. 1.

In this study, first time systematic attempt has been made to develop in silico models for predicting allergic potential of chemicals. We obtained experimentally validated chemical-based allergens and non-allergens from well-established database IEDB. These chemicals were analyzed to understand chemical groups or fingerprints (FP) responsible for causing allergenicity. In order to derive the rules and to understand the relationship between allergenicity and structure of chemicals, we compute wide range of descriptors using PaDEL software [29]. These descriptors can be divided broadly in three categories; 2D descriptors, 3D descriptors and fingerprints. Finally, we developed machine learning based models for predicting allergenicity of chemicals using different types of descriptors. Our best models have been integrated into the webserver; it allows user to predict allergenicity of chemicals as well as to generate analogs of desired allergenicity https://webs.iiitd.edu.in/raghava/chalpred/.

2. Methods

2.1. Dataset collection and descriptors generation

In this study, we have collected allergenic and non-allergenic chemical compounds from the Immune Epitope Database (IEDB) [30] and the structure for the same compounds were downloaded from Chemical Entities of Biological Interest (ChEBI) database [31]. We obtained a total of 519 chemical compounds with allergenic properties from IEDB. On the other hand, we have taken 2211 non-allergenic chemical compounds with a filter of non-peptidic; No IgE; No histamine; No hypersensitivity; No allergy; No Cancer from the IEDB database. The chemical compounds with allergenic properties were considered as a positive dataset (allergens), and compounds with non-allergenic properties were taken as a negative dataset (non--allergens). Further, compound Ids were used to download the 2D and 3D structure files for 519 allergen and 2211 non-allergen chemical compounds. However, out of 2730 compounds, only 403 positive and 1074 negative compound structures were available in ChEBI. The final dataset contained 403 positive and 1074 negative chemical compounds. This dataset was divided into 80:20 ratio, where 80% of the data was used for training and 20% data validation. Our training dataset comprises of 320 allergens and 859 non-allergens, whereas our validation dataset comprises of 83 allergens and 215 non-allergens.

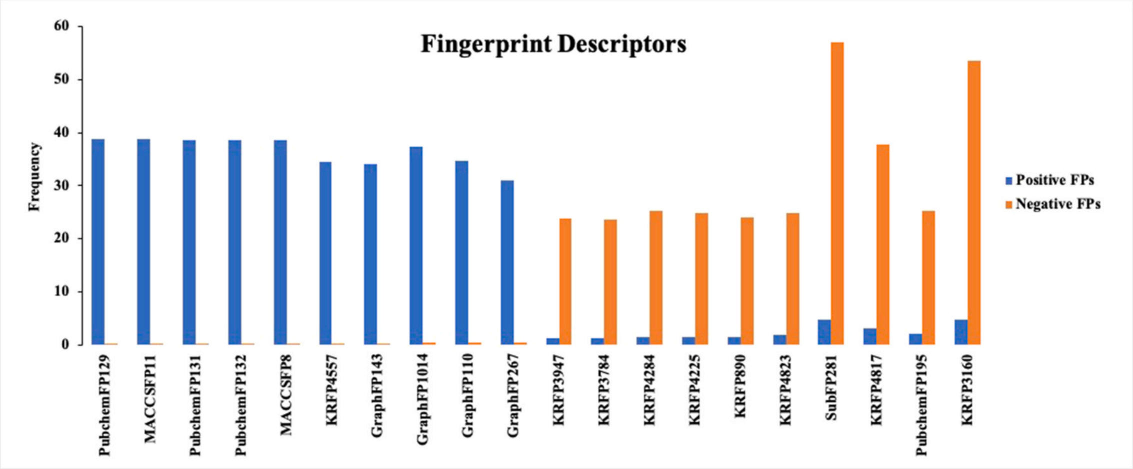

Fig. 2. Shows frequency of top 10 positive and 10 negative fingerprints in allergens and non-allergens.

2.2. Generation of descriptors

The chemical descriptors/features of allergen and non-allergen chemical compounds were computed using PaDEL software [29]. It can compute number of molecular descriptors, such as 2D, 3D and different types of fingerprints for a single chemical compound. It has computed 729 2D descriptors, 431 3D descriptors, and 16092 binary fingerprint-based descriptors for 403 allergen and 1074 non-allergen chemical compounds. These 2D, 3D, and FP descriptor files were further used to develop different machine learning models.

2.3. Dataset preprocessing

The values of 2D, 3D and FP descriptors are in different range, thus we have performed preprocessing in order to normalize the values. In this study, we used a well-established standard scaler method. The normalization and preprocessing were performed using a standard scaler package of scikit learn, i.e., sklearn.preprocessing.StandardScaler, which is based on a z-score normalization algorithm [32].

2.4. Selection of descriptors

It has been shown in past studies, that all the descriptors are not significant [33–35]. Hence, it is important to find out the most relevant features from the vast number of descriptors. There are many feature selection techniques available; however, in this study, we have used the variance threshold-based method, correlation-based method and SVC-L1-based feature selection technique to select the significant features. Firstly, we have removed the low variance features from all descriptor files using the VarianceThreshold feature selection method from the sklearn package [32]. It is used to filter-out low variance 2D, 3D and FP descriptors from the positive and negative data. Initially, there were 729 2D, 431 3D, and 16092 FP descriptors. After removing low variance features, we were left with 286 2D, 362 3D, and 1957 FP descriptors.

Secondly, we have used the correlation-based feature selection method for the removal of highly correlated features. We have developed a python script to compute the pairwise correlation of all descriptors of each dataset. Then we have removed those features that had a correlation of greater than or equal to 0.6 (≥0.6). In this way, remaining were those features which have a correlation less than 0.6 (<0.6) with each other. As a result, we were left with 34 descriptors out of 286 descriptors for 2D, 8 descriptors out of 362 descriptors for 3D, and 210 descriptors out of 1957 FP descriptors. In order to get a highly significant feature set, we further tried to reduce the feature vector size using the most popular feature selection method, i.e., SVC-L1. It implements the support vector classifier (SVC) with linear kernel, penalized with L1 regularization. It selects the non-zero coefficients and then implements the L1 penalty to choose the relevant features from the large feature vector to reduce dimensions [36,37]. Based on this technique, we get the most important feature set, i.e., 14 descriptors out of 34 descriptors for 2D, 6 out of 8 descriptors for 3D and 22 FP descriptors out of 957 descriptors. The information regarding the selected descriptors is tabulated in Supplementary Table 1.

2.5. Machine learning models

In this study, different machine learning techniques have been used for the classification of allergen and non-allergen chemical compounds. Logistic Regression (LR) [38], k-nearest neighbors (KNNs) [39], Decision Tree (DT) [40], Gaussian Naive Bayes (GNB) [41], XGBoost (XGB) [42], Support Vector Classifier (SVC) [43], and Random Forest (RF) [44] were implemented to develop the classification models.

2.6. Cross-validation and evaluation parameters

Several studies in the past have used 80:20 ratio for the division of the complete dataset [21,37]. In the present study, to evaluate the developed machine learning models, we have applied 5-fold cross-validation on 80% of the training data for the internal training, testing and model evaluation [45,46]. In 5-fold CV, the training data is divided into 5-sets, where four sets were used for the training and fifth set was utilized for the testing purposes. The same process is repeated five times, so that each set of positive and negative data is used for training and testing purposes. The performance of machine learning models was evaluated using the standard evaluation parameters. Threshold dependent and independent parameters both were used to measure the performance. Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), Matthews correlation coefficient (MCC) are threshold-dependent parameters, whereas the area under receiver operating characteristic curve (AUC) is a threshold-independent parameter. These performance evaluation parameters are well-defined in the literature and have been extensively used in assessing the performance of the model [21,47,48].

$$Sensitivity\ (Sens) = \frac{TP}{TP + FN} \times 100 \tag{1}$$

$$Specificity\ (Spec) = \frac{TN}{TN + FP} \times 100 \tag{2}$$

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + TN + FN + FP} \times 100 \tag{3}$$

Table 1
The performance of ML-based models developed using 14 (2D) descriptors and 6 (3D) descriptors.

					2D Descriptors					
ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
XGB	81.68	82.11	81.99	0.90	0.60	80.25	76.64	77.63	0.89	0.52
KNN	81.37	81.99	81.82	0.90	0.59	81.48	79.91	80.34	0.88	0.57
RF	81.99	81.29	81.48	0.90	0.59	83.95	81.31	82.03	0.90	0.61
LR	80.12	81.05	80.80	0.88	0.57	81.48	77.57	78.64	0.88	0.54
DT	79.50	79.65	79.61	0.85	0.55	67.90	76.64	74.24	0.80	0.42
GNB	78.57	78.36	78.42	0.86	0.53	81.48	77.57	78.64	0.87	0.54
SVC	78.26	77.78	77.91	0.87	0.52	85.19	78.04	80.00	0.88	0.58
					3D Descriptors					
RF	79.14	78.69	78.81	0.88	0.54	75.33	81.19	79.66	0.85	0.53
KNN	77.61	77.87	77.80	0.85	0.51	68.83	80.73	77.63	0.83	0.47
XGB	76.69	76.11	76.27	0.86	0.49	77.92	79.36	78.98	0.86	0.53
SVC	73.31	72.25	72.54	0.81	0.42	62.34	73.85	70.85	0.77	0.33
LR	68.41	71.31	70.51	0.73	0.36	70.13	72.48	71.86	0.76	0.38
GNB	68.41	70.61	70.00	0.75	0.36	64.94	73.39	71.19	0.75	0.35
DT	69.33	68.38	68.64	0.76	0.34	71.43	60.55	63.39	0.72	0.28

DT, Decision Tree; *GNB*, Gaussian Naive Bayes; *KNN*, k-nearest neighbors; *LR*, Logistic Regression; *RF*, Random Forest; *SVC*, Support Vector Classifier; *XGB*, XGBoost; *Sens*, Sensitivity; *Spec*, Specificity; *Acc,* Accuracy; *AUC*, Area under receiver operating characteristic curve; MCC, Matthews correlation coefficient.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

where FP, FN, TP, and TN are false positive, false negative, true positive, and true negative respectively.

3. Results

3.1. Fingerprints based analysis

In order to understand, importance of each fingerprint in classification of allergens and non-allergens, we have computed the prediction ability of each fingerprint. We used our in-house scripts, to check discrimination ability of fingerprint based descriptors calculated by PaDEL. We ranked the fingerprints according to their probabilities for correctly classifying the chemical as allergen and non-allergen. Based on ranking, we identified the most important 20 fingerprints. Ten fingerprints are highly present in allergens and were called positive fingerprints, whereas other 10 which are highly present in non-allergens were called as negative fingerprints. Fig. 2 depicts the frequency of top 10 positive and 10 negative fingerprints in allergens and non-allergens. These 10 positive fingerprints are highly abundant in allergens but negligible in non-allergens. Similarly, 10 negative fingerprints are highly abundant in non-allergens but negligible in allergens. The complete information regarding these top fingerprints is provided in Supplementary Table 2.

3.2. Prediction models using 2D/3D/FP descriptors

Several models have been developed for predicting chemical allergens using different kinds of chemical descriptors like 2D, 3D and FP descriptors. Several machine learning approaches have been used for developing prediction models, it includes RF, KNN, XGB, SVC, LR, GNB, and DT. The models developed by using these machine learning techniques were optimized by tuning different parameters on training dataset using five-fold cross validation. Firstly, we have developed the machine learning models using 14 descriptors selected from 2D descriptors. XGB algorithm perform better than other ML models and achieved maximum accuracy (81.99% and 77.63%), AUC (0.90 and 0.89) on the training and validation datasets, respectively. Similarly models were developed using 6 features selected from 3D descriptors. RF-based model outperformed the other methods and achieved an accuracy of 78.81% and AUC 0.88 on training dataset as well as accuracy of 79.66% and AUC 0.85 on the validation dataset (Table 1). In order to develop models using fingerprint, we selected 22 out of total 16092 fingerprints. Our RF-based model on 22 fingerprints achieved maximum AUC 0.92 and 0.92 on the training and validation datasets, respectively (Table 2).

3.3. Prediction models using hybrid features

In addition to the prediction models developed using single descriptor, we have also developed the ML-based hybrid model by combining all three types of descriptors, i.e., 2D, 3D and FP. The hybrid model developed using 42 features containing 2D (14 features), 3D (6 features) and FP (22 features). The RF-based model had achieved maximum AUC 0.94 and 0.93 on the training and validation datasets,

Table 2
The performance of ML-based models developed using 22 (FP) descriptors.

ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	85.06	85.11	85.10	0.92	0.66	86.67	85.52	85.81	0.92	0.67
XGB	85.37	85.11	85.18	0.92	0.66	85.33	85.52	85.47	0.90	0.66
LR	83.84	83.82	83.83	0.91	0.64	81.33	81.45	81.42	0.86	0.58
SVC	83.54	83.00	83.15	0.91	0.62	82.67	80.54	81.08	0.86	0.58
KNN	82.93	83.12	83.07	0.90	0.62	85.33	80.54	81.76	0.87	0.60
GNB	79.57	79.37	79.42	0.88	0.55	70.67	81.45	78.72	0.83	0.49
DT	79.88	78.90	79.17	0.86	0.54	77.33	76.92	77.03	0.83	0.49

DT, Decision Tree; *GNB*, Gaussian Naive Bayes; *KNN*, k-nearest neighbors; *LR*, Logistic Regression; *RF*, Random Forest; *SVC*, Support Vector Classifier; *XGB*, XGBoost; *Sens*, Sensitivity; *Spec*, Specificity; *Acc,* Accuracy; *AUC*, Area under receiver operating characteristic curve; MCC, Matthews correlation coefficient.

Table 3
The performance of ML-based hybrid models developed after combining all descriptors.

42 Hybrid Descriptors (2D+3D + FP)										
ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	85.63	86.00	85.90	0.94	0.68	87.95	82.55	84.07	0.93	0.66
SVC	84.06	84.01	84.03	0.91	0.64	93.98	78.77	83.05	0.92	0.66
KNN	84.06	83.66	83.77	0.92	0.63	80.72	81.60	81.36	0.92	0.58
XGB	83.75	83.78	83.77	0.92	0.63	85.54	79.72	81.36	0.92	0.60
LR	83.75	83.08	83.26	0.91	0.62	85.54	82.08	83.05	0.89	0.63
GNB	84.06	79.70	80.88	0.89	0.59	73.49	83.02	80.34	0.87	0.54
DT	79.06	78.76	78.85	0.87	0.53	81.93	80.66	81.02	0.88	0.58
36 Hybrid Descriptors (2D + FP)										
RF	87.5	87.28	87.34	0.94	0.71	84.34	83.02	83.39	0.93	0.63
XGB	85	84.95	84.96	0.93	0.66	81.93	81.13	81.36	0.91	0.59
LR	84.06	84.48	84.37	0.91	0.64	84.34	83.96	84.07	0.90	0.64
KNN	83.44	84.13	83.94	0.93	0.63	81.93	82.08	82.03	0.91	0.60
SVC	84.06	83.08	83.35	0.90	0.63	86.75	81.60	83.05	0.91	0.63
GNB	84.06	79.00	80.37	0.88	0.58	77.11	83.96	82.03	0.88	0.58
DT	80	79.00	79.27	0.86	0.54	86.75	76.89	79.66	0.88	0.58

DT, Decision Tree; *GNB*, Gaussian Naive Bayes; *KNN*, k-nearest neighbors; *LR*, Logistic Regression; *RF*, Random Forest; *SVC*, Support Vector Classifier; *XGB*, XGBoost; *Sens*, Sensitivity; *Spec*, Specificity; *Acc,* Accuracy; *AUC*, Area under receiver operating characteristic curve; MCC, Matthews correlation coefficient.

respectively with balanced sensitivity and specificity. Further, we exclude 3D descriptors, and developed the prediction models using only 36 features (2D and FP). As depicted in Table 3, the RF-based model has obtained an AUC of 0.94 on the training dataset and 0.93 on the validation dataset. It indicates that 36 features are sufficient to achieve highest performance, which is the best model.

Table 4
FDA-approved drug molecules predicted by our server (ChAlPred) causing allergic symptoms.

Drug Bank ID	FDA-Approved Drugs	Prediction	Allergic Symptoms
DB01112	Cefuroxime	Allergen	Anaphylactic Reaction [51]
DB00421	Spironolactone	Allergen	Skin allergy, drug rash with eosinophilia and systemic symptoms (DRESS) induced by spironolactone [52]
DB00859	Penicillamine	Allergen	Skin allergy [53]
DB05013	Ingenol mebutate	Allergen	Skin allergy [54]
DB01007	Tioconazole	Allergen	Contact hypersensitivity [55]
DB06209	Prasugrel	Allergen	Hypersensitivity Skin Reaction [56]
DB01330	Cefotetan	Allergen	Cefotetan-induced anaphylaxis [57]
DB01331	Cefoxitin	Allergen	Allergic reactions [58]
DB04854	Febuxostat	Allergen	Hypersensitivity Reactions (HSRs) [59]
DB09212	Loxoprofen	Allergen	Type I allergic reaction, eosinophilic coronary periarteritis [60]
DB00973	Ezetimibe	Allergen	Angioedema allergic reaction [61]
DB00390	Digoxin	Allergen	Cutaneous hypersensitivity [62]
DB00493	Cefotaxime	Allergen	Immediate Hypersensitivity Reactions [63]
DB01150	Cefprozil	Allergen	Immediate Hypersensitivity Reactions [63]
DB00833	Cefaclor	Allergen	Immediate Hypersensitivity Reactions [63]
DB00689	Cephaloglycin	Allergen	Immediate Hypersensitivity Reactions [63]
DB00438	Ceftazidime	Allergen	Immediate Hypersensitivity Reactions [63]
DB00267	Cefmenoxime	Allergen	Immediate Hypersensitivity Reactions [63]
DB00703	Methazolamide	Allergen	Skin allergy [64]
DB13154	Parachlorophenol	Allergen	Allergic contact dermatitis [65]

3.4. *Webserver interface*

We have developed a user-friendly web server named ChAlPred for the prediction of chemicals as allergens and non-allergens. In this server, we have provided the three modules: (i) predict, (ii) draw and (iii) analog design module. The Predict module allows the user to submit the chemical compounds in different formats, such as SMILE, SDF and MOL formats, to predict whether the chemical could be allergenic or non-allergenic. The Draw module allows the user to draw or modify a molecule in an interactive way using Ketcher [49] and submit the molecule to the machine learning models to predict whether the modified compounds will be allergenic or not. The Analog design module can be used to generate analogs based upon a combination of a given scaffold, building blocks and linkers. The server subsequently predicts the generated analogs as allergenic or non-allergenic. The web server has been designed using a responsive HTML template and browser compatibility for different OS systems.

3.5. *Case study: potential allergenic FDA-approved drugs*

In order to identify the FDA-approved drugs that can cause allergic reactions to the person, we have downloaded a total of 2675 FDA drug molecules from the DrugBank Database [50]. Out of 2675, we have only considered 1102 drugs which are approved. From 1102 drug molecules, the 2D structures were available only for 842 drugs. Finally, we have the structures of 842 FDA-approved drug molecules, which were used to identify that which drug molecules could be allergenic and non-allergenic. For this, we have used the “Predict” module of the “ChAlPred” web server using the default parameters. The model has predicted 114 drug molecules to be allergenic. Several studies done in the past have also supported our findings that some of these drugs can cause allergy in the patient when administered. We have identified 20 drug molecules which are used to cure some diseases but also tend to cause allergic symptoms. Table 4 depicts the information of the drug molecules which cause some allergic reactions.

4. Discussion and conclusion

One of the major challenges in the field of drug discovery is side effect or adverse reaction of drugs. In the past, number of drugs have been already withdrawn from market due to their adverse effects. A wide range of toxicities are responsible for side-effect of drugs, it may be cytotoxicity, immuno-toxicity, hemo-toxicity, liver toxicity or allergenicity [66]. Identification of toxicity is costly, time consuming and

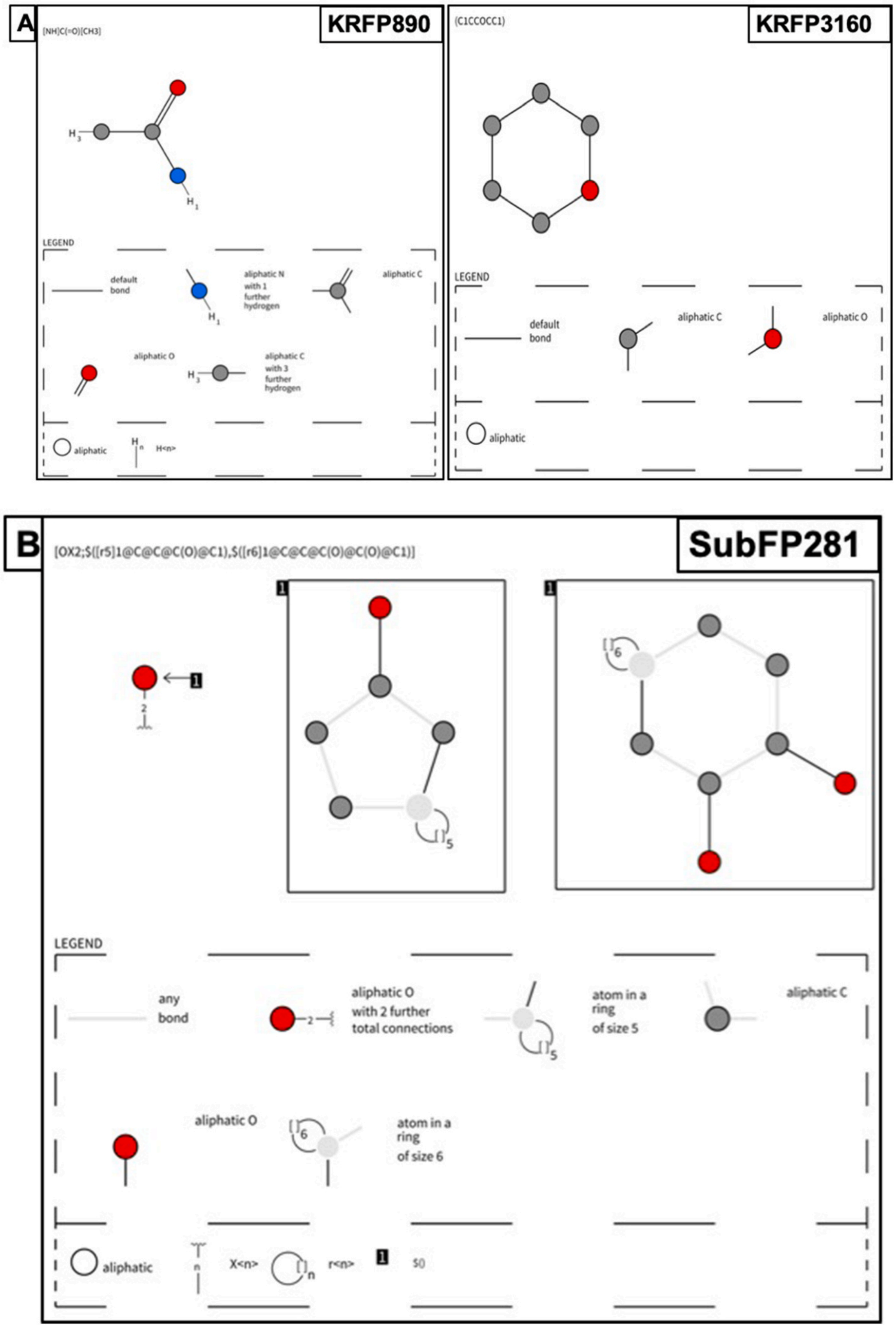

Fig. 3. Chemical structure representation of the fingerprints (A) KRFP890, KRFP3160 (B) SubFP281, found abundant in non-allergic compounds.

tedious task. Thus, there is a need to predict these toxicities using in silico methods. Numerous tools have been developed to estimate the toxicity of the chemicals using different methodologies, such as The Toxicity Estimation Software Tool (TEST). It uses Quantitative Structure-Activity Relationships (QSARs) to estimate the toxicity of chemicals [67]. VegaQSAR [68], Toxtree [69], and PreADMET [70] are the other tools based on the QSAR model for toxicity prediction of the chemical molecules. Machine learning-based tools such as ToxiM, developed by Sharma et al., predict the toxicity and toxicity-related properties of small chemical molecules using machine learning approaches [26], ProTox-II [71].

In contrast, no tool has been developed for predicting allergenicity of chemicals. In this current work, we have collected chemical compounds with their well-defined molecular descriptors utilising publicly available databases such as IEDB and ChEBI. The data yielded several descriptors which was reduced using various feature selection methods. We sorted the most important feature set, i.e., 14 descriptors for 2D, 6 descriptors for 3D and 22 FP descriptors. Based on these selected features (14 2D and 22 FP), we have successfully employed several machine learning approach and found that Random Forest attained a highest AUC of 0.94 and 0.93 in the training as well as validation dataset. In addition, fingerprints based analysis suggests that two positive fingerprints, i.e., PubChemFP129 (Extended Smallest Set of Smallest Rings (ESSSR) ring set ≥ 1 any ring size 4) and GraphFP1014 are highly present in allergenic chemical compounds, and three negative fingerprints, i.e., Klekota-Roth fingerprints (KRFP890 ([!#1][NH]C(=O)[CH3], KRFP3160 (C1CCOCC1)) and Substructure fingerprint (SubFP281 [OX2; $([r5]1@C@C@C(O)@C1),$([r6]1@C@C@C(O)@C(O)@C1)]) are abundant in non-allergenic chemical compounds as shown in Fig. 3(A) and (B).

FDA-approved drugs analysis have shown that few drugs which are used for treatment of certain diseases, are also causing allergy as the side effect. Literature evidences have shown that administration of FDA-approved drugs such as, Cefuroxime [51], Spironolactone [52], Penicillamine [53] can cause allergic reactions like, skin allergies, anaphylactic reactions, hypersensitivity. For instance, A case report has shown that 60 year old patient was experiencing anaphylactic reaction after given an antibiotic cefuroxime [72]. Another report by Kinsara has shown that Spironolactone, a potassium sparing diuretic was given to a patient which was diagnosed with idiopathic cardiomyopathy, and he developed a macular rashes on both the arms [73]. A clinical study by Zhu et al., reported that the patients with Wilson disease were given D-penicillamine (DPA) medication at first, but later they developed neurological symptoms as well as allergies [53].

We can see that these medications can cause a variety of allergic reactions in patients, some of which can be fatal. To prevent these problems, there is dire need for predicting the allergenicity of chemical compounds before using them for treatment purposes. Eventually, we built a freely available webserver namely ChAlPred, for predicting allergenic and non-allergenic chemical compounds using machine learning techniques based on their 2D, 3D and FP molecular descriptors. We hope that this study will be helpful in the future for designing the drug molecules with no allergenic properties.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors contribution

Conception and design: Neelam Sharma, Gajendra P. S. Raghava.
Development of methodology: Neelam Sharma, Sumeet Patiyal, Anjali Dhall, Gajendra P. S. Raghava.
Acquisition of data: Neelam Sharma.
Analysis and interpretation of data and results: Neelam Sharma, Sumeet Patiyal, Anjali Dhall, Gajendra P. S. Raghava.
Webserver Implementation: Neelam Sharma, Sumeet Patiyal.
Writing, reviewing, and revision of the manuscript: Neelam Sharma, Anjali Dhall, Naorem Leimarembi Devi, Gajendra P. S. Raghava.

Data availability

All the datasets generated for this study are available at the "ChAlPred" webserver, https://webs.iiitd.edu.in/raghava/chalpred/dataset.php.

Biorxiv link

DOI: https://doi.org/10.1101/2021.05.21.445101 https://www.biorxiv.org/content/10.1101/2021.05.21.445101v1.full.pdf.

Declaration of competing interest

The authors declare no competing financial and non-financial interests.

Acknowledgement

The authors are thankful to Department of Science and Technology (DST-INSPIRE), Department of Biotechnology (DBT) for fellowships, financial support, and IIIT-Delhi for providing the infrastructure and facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104746.

References

[1] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, G.P.S. Raghava, AlgPred 2.0: an Improved Method for Predicting Allergenic Proteins and Mapping of IgE Epitopes, Brief Bioinform., 2020, p. bbaa294.

[2] I. Dimitrov, I. Bangov, D.R. Flower, I. Doytchinova, v. AllerTOP, 2–a server for in silico prediction of allergens, J. Mol. Model. 20 (2014) 2278.

[3] Z.H. Zhang, J.L. Koh, G.L. Zhang, K.H. Choo, M.T. Tammi, J.C. Tong, AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins, Bioinformatics 23 (2007) 504–506.

[4] H.X. Dang, C.B. Lawrence, Allerdictor: fast allergen prediction using text classification techniques, Bioinformatics 30 (2014) 1120–1128.

[5] S. Maurer-Stroh, N.L. Krutz, P.S. Kern, V. Gunalan, M.N. Nguyen, V. Limviphuvadh, F. Eisenhaber, G.F. Gerberick, AllerCatPro-prediction of protein allergenicity potential from the protein sequence, Bioinformatics 35 (2019) 3020–3027.

[6] K.K. Isaacs, M.R. Goldsmith, P. Egeghy, K. Phillips, R. Brooks, T. Hong, J. F. Wambaugh, Characterization and prediction of chemical functions and weight fractions in consumer products, Toxicol Rep 3 (2016) 723–732.

[7] I. Kimber, D.A. Basketter, G.F. Gerberick, C.A. Ryan, R.J. Dearman, Chemical allergy: translating biology into hazard characterization, Toxicol. Sci. 120 (Suppl 1) (2011) S238–S268.

[8] U.S. Food & Drug Administration, Allergens in cosmetics, Available at, https://www.fda.gov/cosmetics/cosmetic-ingredients/allergens-cosmetics, 2020. (Accessed 13 May 2021). December 11.

[9] T.W. Mak, M. Saunders, B.D. Jett, Immune hypersensitivity, in: Primer to the Immune Response. Academic Cell., 2014, pp. 487–516.

[10] Molecular mechanism behind why allergies are more common in developed countries discovered: British Society for Immunology, Available at, https://www.immunology.org/news/molecular-mechanism-allergies-discovered, 2017. (Accessed 13 May 2021). accessed.

[11] D.A. Santos, Why the World Is Becoming More Allergic to Food, 2018. *BBC News* King's College London: BBC. Available at, https://www.bbc.com/news/health-46302780. (Accessed 13 May 2021). accessed.

[12] W. Loh, M.L.K. Tang, The epidemiology of food allergy in the global context, Int. J. Environ. Res. Publ. Health 15 (2018).

[13] E. Hossny, M. Ebisawa, Y. El-Gamal, S. Arasi, L. Dahdah, R. El-Owaidy, C. A. Galvan, B.W. Lee, M. Levin, S. Martinez, R. Pawankar, M.L.K. Tang, E.H. Tham, A. Fiocchi, Challenges of managing food allergy in the developing world, World Allergy Org. J. 12 (2019) 100089.

[14] G. Obermeyer, F. Ferreira, Can we predict or avoid the allergenic potential of genetically modified organisms? Int. Arch. Allergy Immunol. 137 (2005) 151–152.

[15] I. Kimber, D.A. Basketter, R.J. Dearman, Chemical allergens–what are the issues? Toxicology 268 (2010) 139–142.

[16] R.E. Goodman, S.L. Hefle, S.L. Taylor, R. van Ree, Assessing genetically modified crops to minimize the risk of increased food allergy: a review, Int. Arch. Allergy Immunol. 137 (2005) 153–166.

[17] M.G. Del Moral, E. Martinez-Naves, The role of lipids in development of allergic responses, Immune Netw. 17 (2017) 133–143.

[18] S.P. Commins, T.A. Platts-Mills, Allergenicity of carbohydrates and their role in anaphylactic events, Curr. Allergy Asthma Rep. 10 (2010) 29–33.

[19] R. Rubin, Allergic reactions to mRNA vaccines, J. Am. Med. Assoc. 325 (2021) 2038.

[20] N.B. Alsaleh, J.M. Brown, Engineered nanomaterials and type I allergic hypersensitivity reactions, Front. Immunol. 11 (2020) 222.

[21] S. Saha, G.P. Raghava, AlgPred: prediction of allergenic proteins and mapping of IgE epitopes, Nucleic Acids Res. 34 (2006) W202–W209.

[22] H.C. Muh, J.C. Tong, M.T. Tammi, AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins, PloS One 4 (2009), e5861.

[23] I. Dimitrov, D.R. Flower, I. Doytchinova, AllerTOP–a server for in silico prediction of allergens, BMC Bioinf. 14 (Suppl 6) (2013) S4.

[24] J. Wang, D. Zhang, J. Li, PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection, BMC Syst. Biol. 7 (Suppl 5) (2013) S9.

[25] I. Dimitrov, L. Naneva, I. Doytchinova, I. Bangov, AllergenFP: allergenicity prediction by descriptor fingerprints, Bioinformatics 30 (2014) 846–851.

[26] A.K. Sharma, G.N. Srivastava, A. Roy, V.K. Sharma, ToxiM: a toxicity prediction tool for small molecules developed using machine learning and chemoinformatics approaches, Front. Pharmacol. 8 (2017) 880.

[27] M. Saito, R. Arakaki, A. Yamada, T. Tsunematsu, Y. Kudo, N. Ishimaru, Molecular mechanisms of nickel allergy, Int. J. Mol. Sci. 17 (2016).

[28] I. Kimber, R.J. Dearman, D.A. Basketter, D.R. Boverhof, Chemical respiratory allergy: reverse engineering an adverse outcome pathway, Toxicology 318 (2014) 32–39.

[29] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474.

[30] R. Vita, S. Mahajan, J.A. Overton, S.K. Dhanda, S. Martini, J.R. Cantrell, D. K. Wheeler, A. Sette, B. Peters, The immune Epitope database (IEDB): 2018 update, Nucleic Acids Res. 47 (2019) D339–D343.

[31] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, C. Steinbeck, ChEBI in 2016: improved services and an expanding collection of metabolites, Nucleic Acids Res. 44 (2016) D1214–D1219.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, W.R. Prettenhofer, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[33] H. Singh, S. Singh, D. Singla, S.M. Agarwal, G.P. Raghava, QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest, Biol. Direct 10 (2015) 10.

[34] S.K. Dhanda, D. Singla, A.K. Mondal, G.P. Raghava, DrugMint: a webserver for predicting and designing of drug-like molecules, Biol. Direct 8 (2013) 28.

[35] A. Dhall, S. Patiyal, N. Sharma, N. LDevi, Gajendra P.S. Raghava, Computer-aided Prediction of Inhibitors against STAT3 for Managing COVID-19 Associate Cytokine Storm, PREPRINT (Version 1) Available at Research Square, 2021.

[36] J. Tang, S. Alelyani, H. Liu, Feature selection for classification: a review, Data Classif Algorithms Appl 37 (2014) 1871–1874.

[37] A. Dhall, S. Patiyal, N. Sharma, S.S. Usmani, G.P.S. Raghava, Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19, Briefings Bioinf. 22 (2021) 936–945.

[38] J. Tolles, W.J. Meurer, Logistic regression: relating patient characteristics to outcomes, J. Am. Med. Assoc. 316 (2016) 533–534.

[39] A. Mucherino, P.J. Papajorgji, P.M. Pardalos, K-nearest neighbor classification, in: Data Mining in Agriculture, vol. 34, Springer Optimization and Its Applications, 2009, pp. 83–106.

[40] G.I. Webb, J. Fürnkranz, J. Fürnkranz, et al., Decision tree, Encycl. Mach. Learn. 63 (2011) 263–267.

[41] H. Zhang, Exploring conditions for the optimality of naïve bayes, Int. J. Pattern Recogn. Artif. Intell. 19 (No. 02) (2005) 183–198.

[42] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[43] C. Zhang, Xiaojian Shao, Dewei Li, Knowledge-based support vector classification based on C-SVC, Procedia Comput. Sci. 17 (2013) 1083–1090.

[44] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.

[45] A. Dhall, S. Patiyal, H. Kaur, S. Bhalla, C. Arora, G.P.S. Raghava, Computing skin cutaneous melanoma outcome from the HLA-alleles and clinical characteristics, Front. Genet. 11 (2020) 221.

[46] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, G.P.S. Raghava, AntiCP 2.0: an Updated Model for Predicting Anticancer Peptides, Brief Bioinform, 2020 bbaa153.

[47] S. Patiyal, P. Agrawal, V. Kumar, A. Dhall, R. Kumar, G. Mishra, G.P.S. Raghava, NAGbinder: an approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence, Protein Sci. 29 (2020) 201–210.

[48] V. Kumar, P. Agrawal, R. Kumar, S. Bhalla, S.S. Usmani, G.C. Varshney, G.P. S. Raghava, Prediction of cell-penetrating potential of modified peptides containing natural and chemically modified residues, Front. Microbiol. 9 (2018) 725.

[49] Life Sciences Open Source, Ketcher 2.0 (accessed at 13 May, 2021), https://lifescience.opensource.epam.com/ketcher/index.html#ketcher-2-0.

[50] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082.

[51] P. Del Villar-Guerra, B. Moreno Vicente-Arche, S. Castrillo Bustamante, C. Santana Rodriguez, Anaphylactic reaction due to cefuroxime axetil: a rare cause of anaphylaxis, Int. J. Immunopathol. Pharmacol. 29 (2016) 731–733.

[52] P.D. Ghislain, A.D. Bodarwe, O. Vanderdonckt, D. Tennstedt, L. Marot, J. M. Lachapelle, Drug-induced eosinophilia and multisystemic failure with positive patch-test reaction to spironolactone: DRESS syndrome, Acta Derm. Venereol. 84 (2004) 65–68.

[53] X.Q. Zhu, L.Y. Li, W.M. Yang, Y. Wang, Combined dimercaptosuccinic acid and zinc treatment in neurological Wilson's disease patients with penicillamine-induced allergy or early neurological deterioration, Biosci. Rep. 40 (2020).

[54] U.S. Food & Drug Administration, FDA Drug Safety Communication: FDA warns of severe adverse events with application of Picato (ingenol mebutate) gel for skin condition; requires label changes (accessed at 13 May, 2021), https://www.fda.gov/drugs/drug-safety-and-availability/fda-drug-safety-communication-fda-warns-severe-adverse-events-application-picato-ingenol-mebutate.

[55] H. Heikkila, S. Stubb, S. Reitamo, A study of 72 patients with contact allergy to tioconazole, Br. J. Dermatol. 134 (1996) 678–680.

[56] S.H. Kim, S.D. Park, Y.S. Baek, S.Y. Lee, S.H. Shin, S.I. Woo, D.H. Kim, J. Kwan, Prasugrel-induced hypersensitivity skin reaction, Kor. Circ. J. 44 (2014) 355–357.

[57] Y.H. Nam, E.K. Hwang, G.Y. Ban, H.J. Jin, H.S. Yoo, Y.S. Shin, Y.M. Ye, D.H. Nahm, H.S. Park, S.K. Lee, Immunologic evaluation of patients with cefotetan-induced anaphylaxis, Allergy Asthma Immunol. Res. 7 (2015) 301–303.

[58] D.J. Crotty, X.J. Chen, M.R. Scipione, Y. Dubrovskaya, E. Louie, J.A. Ladapo, J. Papadopoulos, Allergic reactions in hospitalized patients with a self-reported penicillin allergy who receive a cephalosporin or meropenem, J. Pharm. Pract. 30 (2017) 42–48.

[59] K.S. Ma, J.C. Wei, W.H. Chung, Correspondence to 'Hypersensitivity reactions with allopurinol and febuxostat: a study using the Medicare claims data, Ann. Rheum. Dis. (2020), https://doi.org/10.1136/annrheumdis-2020-218090.

[60] S. Ichimata, Y. Hata, N. Nishida, An autopsy case of sudden unexpected death with loxoprofen sodium-induced allergic eosinophilic coronary periarteritis, Cardiovasc. Pathol. 44 (2020) 107154.

[61] T. Lu, T. Grewal, Ezetimibe: an unusual suspect in angioedema, Case Rep. Med. 2020 (2020) 9309382.

[62] S.J. Martin, D. Shah, Cutaneous hypersensitivity reaction to digoxin, J. Am. Med. Assoc. 271 (1994) 1905.

[63] M.H. Kim, J.M. Lee, Diagnosis and management of immediate hypersensitivity reactions to cephalosporins, Allergy Asthma Immunol. Res. 6 (2014) 485–495.

[64] Y. Xu, M. Wu, F. Sheng, Q. Sun, Methazolamide-induced toxic epidermal necrolysis in a Chinese woman with HLA-B5901, Indian J. Ophthalmol. 63 (2015) 623–624.

[65] T.S. Sonnex, R.J. Rycroft, Allergic contact dermatitis from orthobenzyl parachlorophenol in a drinking glass cleaner, Contact Dermatitis 14 (1986) 247–248.

[66] H. Yang, L. Sun, W. Li, G. Liu, Y. Tang, Silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts, Front. Chem. 6 (2018) 30.

[67] U.S. EPA, User's guide for T.E.S.T. (version 5.1) (toxicity estimation software tool): a program to estimate toxicity from molecular structure, in: Chemical Characterization and Exposure Division Cincinnati, 2020. Avaiable at, https://www.epa.gov/sites/production/files/2016-05/documents/600r16058.pdf 2020. (Accessed 13 May 2021).

[68] Q.S.A.R. VEGA, Laboratory of Environmental Chemistry and Toxicology, 2021. https://www.vegahub.eu/portfolio-item/vega-qsar/. (Accessed 23 July 2021). accessed.

[69] G. Patlewicz, N. Jeliazkova, R.J. Safford, A.P. Worth, B. Aleksiev, An evaluation of the implementation of the Cramer classification scheme in the Toxtree software, SAR QSAR Environ. Res. 19 (2008) 495–524.

[70] PreADMET. https://preadmet.bmdrc.kr/toxicity-prediction/, 2015. (Accessed 23 July 2021) accessed.

[71] P. Banerjee, A.O. Eckert, A.K. Schrey, R. Preissner, ProTox-II: a webserver for the prediction of toxicity of chemicals, Nucleic Acids Res. 46 (2018) W257–W263.

[72] J. Gu, S. Liu, Y. Zhi, Cefuroxime-induced anaphylaxis with prominent central nervous system manifestations: a case report, J. Int. Med. Res. 47 (2019) 1010–1014.

[73] A.J. Kinsara, Spironolactone- induced rash: a case report and review, J. Clin. Cardiol. Diagn. 1 (2) (2018) 1–2.

Neelam Sharma (NS) is currently working as Ph.D. in Bioinformatics from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Sumeet Patiyal (SP) is currently working as Ph.D. in Bioinformatics from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Anjali Dhall (AD) is currently working as Ph.D. in Bioinformatics from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Dr. Naorem Leimarembi Devi (NLD) is currently working as Research associate in the Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Prof. Gajendra P. S. Raghava (GPSR) is currently working as Professor and Head of Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India