International Journal of *Molecular Sciences*

Article

Improvement of Epitope Prediction Using Peptide Sequence Descriptors and Machine Learning

Cristian R. Munteanu [1,2,3], Marcos Gestal [1,3,*], Yunuen G. Martínez-Acevedo [1,4], Nieves Pedreira [1], Alejandro Pazos [1,2] and Julián Dorado [1,3]

[1] RNASA-IMEDIR, Computer Science Faculty, University of A Coruna, 15071 A Coruña, Spain

[2] Biomedical Research Institute of A Coruña (INIBIC), University Hospital Complex of A Coruña (CHUAC), 15006 A Coruña, Spain

[3] Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), Campus de Elviña s/n, 15071 A Coruña, Spain

[4] Unidad Profesional Interdisciplinaria de Biotecnología, National Polytechnic Institute (IPN), Ticoman, 07340 Mexico City, Mexico

* Correspondence: marcos.gestal@udc.es; Tel.: +34-981167000 (ext. 1379)

Received: 31 July 2019; Accepted: 30 August 2019; Published: 5 September 2019

Abstract: In this work, we improved a previous model used for the prediction of proteomes as new B-cell epitopes in vaccine design. The predicted epitope activity of a queried peptide is based on its sequence, a known reference epitope sequence under specific experimental conditions. The peptide sequences were transformed into molecular descriptors of sequence recurrence networks and were mixed under experimental conditions. The new models were generated using 709,100 instances of pair descriptors for query and reference peptide sequences. Using perturbations of the initial descriptors under sequence or assay conditions, 10 transformed features were used as inputs for seven Machine Learning methods. The best model was obtained with random forest classifiers with an Area Under the Receiver Operating Characteristics (AUROC) of 0.981 ± 0.0005 for the external validation series (five-fold cross-validation). The database included information about 83,683 peptides sequences, 1448 epitope organisms, 323 host organisms, 15 types of in vivo processes, 28 experimental techniques, and 505 adjuvant additives. The current model could improve the in silico predictions of epitopes for vaccine design. The script and results are available as a free repository.

Keywords: epitopes; machine learning; protein sequences; qualitative structure–activity relationships

1. Introduction

The term proteome was used for the first time in 1994 and was defined as the total proteins expressed by a genome in a tissue or in a cell. It is said that its main characteristic is to be dynamic, due to its components, as proteins vary depending on the tissue, cell, or cell compartment, and these can change in response to their microenvironment—for instance stress, temperature, drug action, etc. Proteomics unlocks the paths to the search for clinically useful biomarkers of diseases, treatment response, and aging. Currently, proteomics research has application in different fields of science and industry, so a large amount of information contained in public databases is available [1]. In addition, the fast growth of bioinformatics techniques and their applications, in conjunction with the substantial quantity of experimental information, have created a big impact on immunology analysis.

Vaccination is one of the most effective techniques to prevent infectious diseases. Upon vaccination, a foreign antigen interacts with the host immune system, which evokes an immune response. The immunizing agent could also be an intact inactivated version of the pathogens or parts of the pathogens that are able to generate a strong immunogenic response. The understanding of the antigen molecular interactions resulted in the development of improved peptide vaccines. Peptide vaccines consist in the discovery and synthesis of B-cell and T-cell epitopes. These peptides are able to induce a specific immune response. Epitopes are recognized by the immune system parts as antibodies, B cells, and T-cells [2–4]. The epitope fragments could be found in foreign and self-proteins. They can be categorized depending on their structure and how they integrate with the paratope [5]. T-cell epitopes need a bond with the major histocompatibility (MHC) molecules in order to induce an immune response, and they can be found on the surface of an antigen-presenting cell. The key role in the immune system is represented by the recognition of the epitopes by the T-cells. Small variations of the normal immunity function could have a grave impact on the organism. In specific cases of autoimmune diseases, the peptides of the native cells are miss-recognized by the T-cells as foreign peptides. Thus, native peptide could be attacked, resulting in a modification of the normal tissues [6].

A possible method to find new T-cell epitopes is the prediction of MHC binding needed for the T-cell recognition [7]. This is extremely difficult to be put into practice by experimental techniques using an extensive set of alleles. Thus, the bioinformatics solutions are needed for developing prediction methods of epitopes. Data-driven solutions are the most successful for the prediction of T-cell epitopes. The T-cell epitope in silico prediction is usually based on previous information such as the peptide-binding specificity to MHC alleles [8]. Different prediction algorithms for T-cell epitopes have been constructed using peptide sequence and experimental affinity. The Immune Epitope Database (IEDB—http://www.iedb.org) [9] is a freely available resource that classifies experimental data on immune epitopes (the molecular targets of adaptive immune responses) studied in humans, non-human primates, and other animal species in the context of infectious diseases, allergy, autoimmunity, and transplantation, with the aim of keeping the findings of recent research available and updated.

The computational techniques are aimed at finding mathematical relationships in the form of equations between datasets that have been obtained based on experimental work or have been calculated using theoretical considerations. Studies of correlations between physical properties and chemical or biological properties require quality biological data, define relevant chemical descriptors, and choose a suitable model to predict the biological function of peptides. The molecular information about a peptide could be its 3D structure or its simple amino acid sequence. The computational methods to find a mathematical model to predict peptide function using its molecular information are numerous, from linear to non-linear algorithms. The qualitative structure–activity/property relationships (QSAR/QSPR) [10,11] represent very useful methods that are able to predict molecular activity or property using molecular descriptors. In order to find the mathematical function/algorithm that is able to map the inputs (molecule descriptors/characteristics) to the output (molecule biological activity), linear and non-linear Machine Learning (ML) methods are used. In order to encode the changes on B-cell epitope activity after multiple variations/perturbations in factors such as the sequences of the peptides, host organism, source organism, immunological process, and experimental technique, the perturbation theory (PT) is introduced to mix the original molecular descriptors with these factors [12,13]. This feature transformation through mixing information is an alternative to the classical ML one-hot representation of the categorical features (binary features for each category).

Different groups developed ML models to predict B-cell epitope activity after multiple variations/perturbations in experimental conditions. Gonzalez-Diaz et al. [14] developed the first linear model for B-epitope prediction using 200,000 cases of perturbations such as structural changes in peptides from assay involving 500 source organisms, 50 host organisms, 10 biological processes, and 30 experimental techniques. The reported accuracy, sensitivity, and specificity were >90% (training and validation series). Vázquez-Prieto et al. [15] compared the PT-ML models for B-cell epitopes prediction using different physicochemical molecular properties of peptide sequences. All these models do not include important factors such as adjuvant additives in vaccine design. In addition, the datasets of the studies were based on a previous version of IEDB.

Martinez-Arzate et al. presented the latest published model for epitopes [16]. This model was based on Shannon's information as peptide molecular descriptors and experimental condition perturbations. Thus, the dataset was based on 1,048,190 pairs of query and reference peptide sequences and perturbations in sequence or assay conditions: 1448 epitope organisms, 323 host organisms, 15 types of in vivo processes, 28 experimental techniques, and 505 adjuvant additives. This linear model was characterized by modest values of accuracy, sensitivity, and specificity (71–80%). Therefore, these results were improved by the current study using non-linear ML methods, better metrics for unbalanced datasets (area under the receiver operating characteristics—AUROC [17]), reproducible open-access python scripts, and multiple dataset splits (n-fold cross-validation) for statistical significance of the results.

2. Results

The main Jupyter notebook uses pipelines from sklearn (python) [18], standardization scaling of data (standard deviation units), five-fold cross-validation (outer validation using stratified folds for unbalanced classes), and seven Machine Learning methods, such as k-nearest neighbors algorithm (KNN) [19], support vector machine (SVM linear and SVM non-linear based on radial basis functions, RBF) [20], logistics regression (LR) [21], decision tree (DT) [22], random forest (RF) [23], and XGBoost—an optimized distributed gradient boosting library (XGB) [24]. The performance of the models was characterized applying the Area Under the Receiver Operating Characteristics (AUROC) [17]. In order to be able to reproduce the project results, all the scripts, datasets and results are available as a GitLab repository (https://gitlab.com/muntisa/machine-learning-for-peptide-epitopes/).

The AUROC values for the used ML methods are shown in Table 1. The best classifier using the default parameters of all the ML methods and class weights for unbalanced classes was provided by random forest (10 estimators/decision trees) with an AUROC of 0.973 ± 0.001 (see 1-Epitope-classifiers-7ML.ipynb).

Table 1. Area Under the Receiver Operating Characteristics (AUROC) values for seven Machine Learning (ML) methods (five-fold cross-validation (CV)).

ML	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	SD
KNN	0.900	0.898	0.900	0.898	0.897	0.899	0.0013
SVM linear	0.785	0.800	0.789	0.799	0.790	0.792	0.0067
SVM	0.866	0.863	0.864	0.866	0.862	0.864	0.0019
LR	0.818	0.816	0.816	0.816	0.814	0.816	0.0015
DT	0.923	0.923	0.923	0.923	0.923	0.923	0.0003
RF	0.974	0.973	0.973	0.972	0.971	**0.973**	**0.0010**
XGB	0.892	0.890	0.890	0.889	0.887	0.890	0.0017

ML = Machine Learning; SD = standard deviation; KNN = KNeighborsClassifier, SVM linear = SVC (kernel="linear"), SVM = SVC (kernel="rbf"), LR = LogisticRegression, DT = DecisionTreeClassifier, RF = RandomForestClassifier, XGB = XGBClassifier; the best AUROC value and the corresponding SD are bolded.

Figure 1 shows the box-plot of the AUROC values for all the ML methods. The AUROC values for the five folds have very small variations (SD between 0.0003 for DT and 0.0067 for SVM linear). This suggests that the AUROCs for all the ML methods are stable within each fold. In addition, the high difference between the RF and the other methods (box-plots are far from overlapping) demonstrated that the results were statistically significant.

For the best ML method, RF, a test was performed using different numbers of trees, from 5 to 1000 (five-fold CV) (see Figure 2). Even if for the classic QSAR model, RF could reach a constant performance after only 50 trees [25], we tested this assumption with our complex model based on difference between the perturbations of the molecular descriptors under experimental conditions. Even with five trees, the classifier was able to obtain an AUROC of 0.963. After 100 trees, AUROC was higher than 0.98, but no significant improvements were obtained by increasing the number of trees to 200–1000. By increasing the number of trees from 10 to 20, from 20 to 40, or from 40 to 100, statistically significant improvements were obtained for the AUROC values. Therefore, we chose the best models as RF100 (RF with 100 estimators/trees), with an AUROC average of 0.981 ± 0.0005 (see 2-Epitope-RF-trees.ipynb).

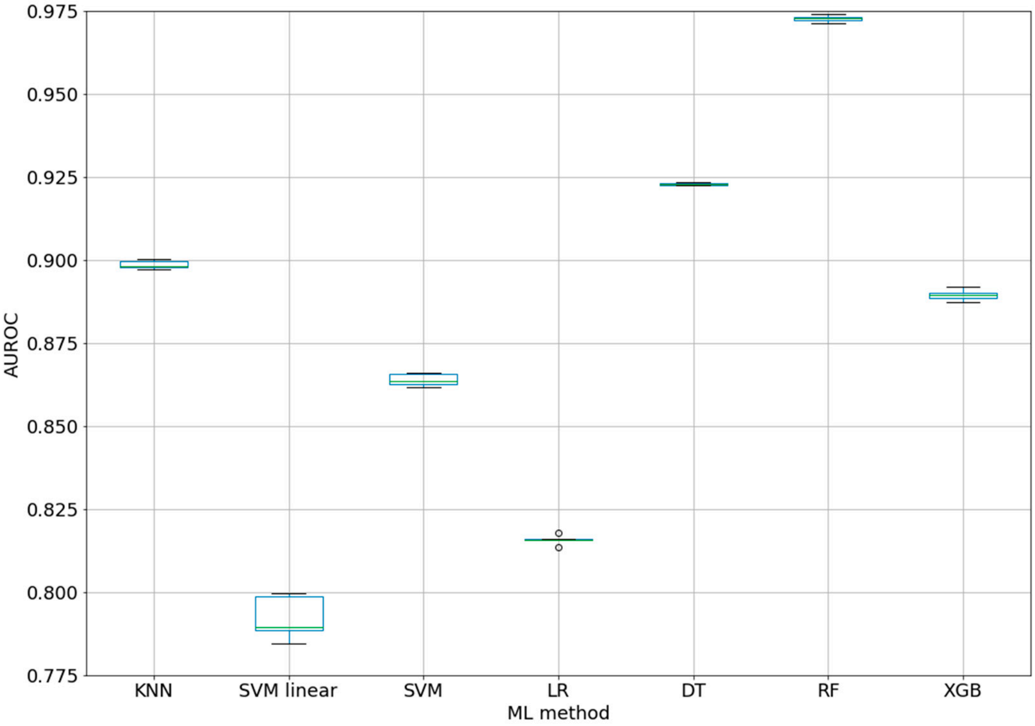

Figure 1. Box-plot for AUROC values of ML classifiers (five-fold CV).

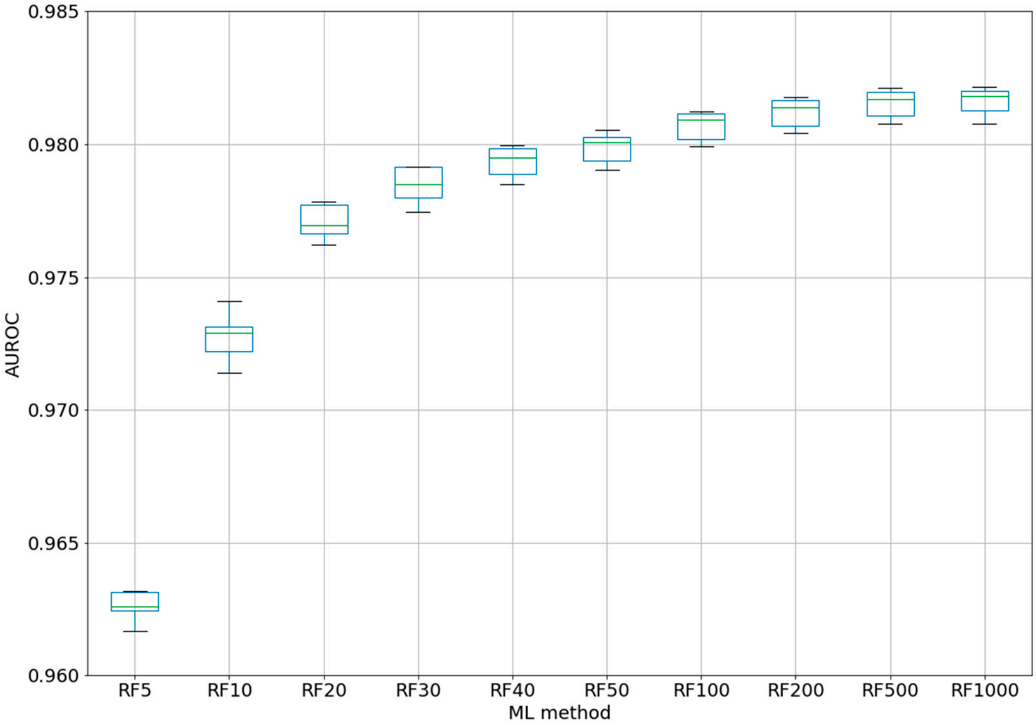

Figure 2. Box-plot for AUROC values of RF classifiers with different trees (five-fold CV). RF*n* = Random Forest with *n* trees (*n* = 5, 10, 20, 30, 40, 50, 100, 200, 500, 1000).

3. Discussion

The prediction of new epitopes represents a challenge for the vaccine design. In a previous study [16], a linear classifier (linear discriminant analysis (LDA) [26]) was proposed with accuracy, sensitivity, and specificity between 71 and 80% (training and test subsets). The classifier was able to predict the epitope activity of a query peptide under a set of experimental conditions and using a reference peptide. Therefore, the input features consisted of peptide molecular descriptors calculated with S2SNet software and the derived features that mixed original peptide descriptors with experimental data applying the perturbation theory. The study published the dataset containing 1,048,190 pairs of query and reference peptide sequences. The dataset was based on 83,683 peptides sequences, 1448 epitope organisms, 323 host organisms, 15 types of in vivo processes, 28 experimental techniques, and 505 adjuvant additives. The model demonstrated the power of QSAR models using peptide descriptors and the perturbation theory.

The proposed linear classifier has several limitations. The relationships between the molecular properties and their activity are not always linear. In fact, most of the relationships in nature are not linear. This could explain the maximum performance of only 80% (accuracy). The non-linear Machine Learning methods could offer non-linear relationships between the molecular properties and their activity/property. This is the main reason of the current study, where several non-linear methods were tested. Therefore, seven ML methods were used, including KNN, linear SVM, non-linear SVM using radial basis function (RBF) kernels, LR, and tree-based methods, such as DT, RF, and XGB.

Due to an unbalanced dataset (the number of instances is different in the classes), the accuracy metrics was not the most accurate. Consequently, there was a need for better metrics, such as AUROC and the use of class weighting in ML training. In addition, a single split of data did not provide the best statistical results. Thus, all the current calculations used five-fold CV. The published dataset was corrected by eliminating the duplicate instances too. The availability of an open repository with the dataset, scripts, and results offer the possibility to reproduce all the results just by executing these scripts.

The statistical significance is very important in ML, and boxplots were therefore created with the scripts in order to present the distribution of the AUROC values with each fold (split of dataset in training and test subset). It was therefore possible to check the spread of the results, the median value, and the outlier values. Thus, we were able to choose random forest as the best ML method. In addition, we checked how the number of trees/estimators in RF could influence the AUROC values. The results demonstrated limitations in increasing the number of trees to 100. From this value (0.981), the gain in AUROC was not statistically significant and the computational effort was not proportional.

The feature importance of this model for five-fold CV is shown in Figure 3. The query epitope activity class (output variable) was predicted using the observed activity of a reference epitope, three query epitope perturbations of S2SNet descriptors, and six differences between the perturbations of the S2SNet descriptors of query and the reference epitope sequence (under the same experimental conditions). Thus, the observed reference epitope activity (ε_r) was the most important input feature, followed by the perturbation of the Shannon entropies of the query epitope sequence for $k = 5$ and $k = 0$ in the sequences and organisms, ${}^{q}\theta_5$(Seq) and ${}^{q}\theta_0$(Org). The next two features referred to differences between perturbations in the sequences and organisms: $\Delta\theta_5$(Seq), $\Delta\theta_0$(Org). It was observed that, for the epitope activity, the most important information was encoded into the sequence and depended on the organism. On the contrary, the difference between the perturbations of Shannon entropies for query and reference epitopes in the same adjuvant additives conditions ($\Delta\theta_0$(Adju)) was less important for this model.

In conclusion, the current study improved the previous linear classifier using a non-linear classifier (RF), better metrics, such as AUROC, statistical significance using five-fold cross-validation, and offering the script for the reproducibility of the results. This methodology could be used to improve the in silico screening of peptides for a new epitope activity.

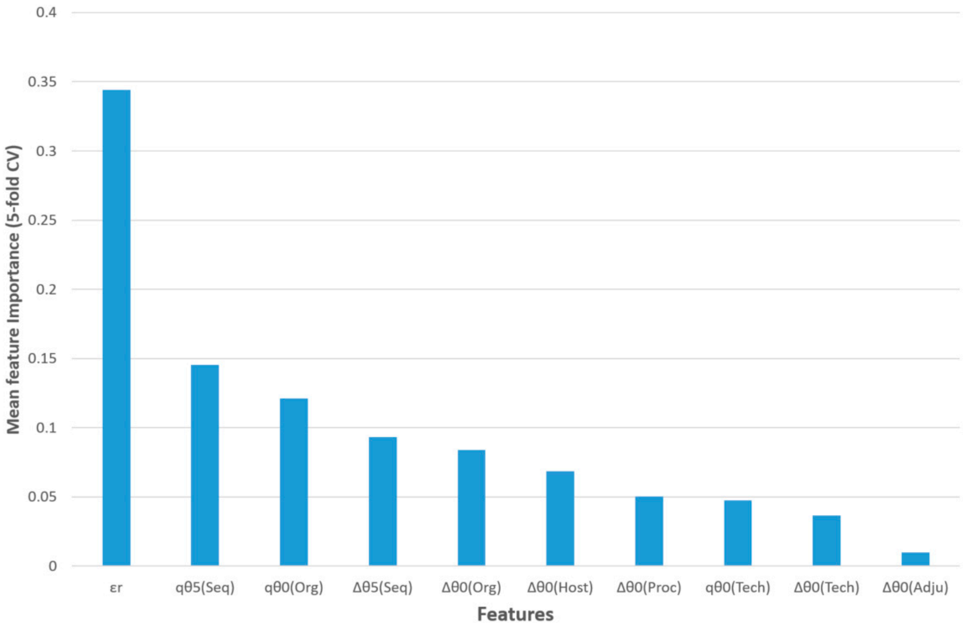

Figure 3. Feature importance for the best RF classifier.

4. Materials and Methods

The current work improved the classifier for epitope prediction using a dataset published in a previous paper [16]. The dataset contains information about the linear B-cell epitopes reported in the IEDB database (http://www.iedb.org): 83,683 peptides sequences, 1448 epitope organisms, 323 host organisms, 15 types of in vivo processes, 28 experimental techniques, and 505 adjuvant additives. The features are based on S2SNet peptide descriptors, query and reference peptide sequences and experimental conditions. Using the perturbation theory, the transformed descriptors were calculated (see Ref. [16]). Thus, the proposed classifiers are QSAR models (peptide structure – activity relationships) using perturbations of the molecular descriptors.

The current study presented a state-of-the-art classifier for in silico B-cell epitope prediction before its synthesis in order to save time and money in vaccine design. The classifier represents a Quantitative Structure-Activity Relationship (QSAR) [27] model obtained with Machine Learning methods. The structure of query and reference peptide sequences were transformed into descriptors of sequence recurrence networks using S2SNet software [28]. The initial descriptors were transformed into perturbation descriptors using the perturbation theory for specific experimental conditions. The dataset was available from Ref [16].

The dataset presented in the previous work [16] was used to evaluate the level of epitope activity as a class (query sequence). This output was assessed using information about a pair of epitope sequences: query and references epitopes. The final features of the models are: the observed activity of a reference epitope, the perturbations of Shannon entropies of the query epitope Star graphs under specific experimental conditions, and the differences between these perturbations for query and reference epitopes in the same conditions. Thus, the prediction of a new epitope activity (query epitope) should use an already known activity of a reference epitope and perturbations of graph-type molecular descriptors of both epitope sequences.

The 10 input features are ε_r, ${}^q\theta_5$(Seq), ${}^q\theta_0$(Org), ${}^q\theta_0$(Tech), $\Delta\theta_5$(Seq), $\Delta\theta_0$(Host), $\Delta\theta_0$(Adju), $\Delta\theta_0$(Proc), $\Delta\theta_0$(Org), and $\Delta\theta_0$(Tech). Figure 4 shows a flow of the methodology: from the query-reference epitope sequences, the molecular descriptors and their perturbations under experimental conditions were calculated. ε_r is the observed value of the epitope activity for the reference (r) peptide sequence. The Shannon entropy information measures for query (q) and reference (r) sequences are ${}^q\theta_k$ and ${}^r\theta_k$ (k = natural powers of the Markov matrix used in the S2SNet software). ${}^q\theta_k(c_j)$ represents the perturbations of ${}^q\theta_k$ for c_j factor/experimental conditions such as peptides sequences (Seq), epitope organisms (Org), host organisms (Host), in vivo processes (Proc), experimental techniques (Tech), and adjuvant additives (Adj). Thus, ${}^{q}\theta_{5}$(Seq), ${}^{q}\theta_{0}$(Org), ${}^{q}\theta_{0}$(Tech) represent the perturbations of peptide Shannon entropies for query sequence with Seq, Org and Tech. $\Delta\theta_{k}(c_{j})$ represents the differences between the perturbations of Shannon entropies for query and reference peptides for c_j factors/experimental conditions ${}^{q}\theta_{k}(c_{j}) - {}^{r}\theta_{k}(c_{j})$. Thus, $\Delta\theta_{5}$(Seq), $\Delta\theta_{0}$(Host), $\Delta\theta_{0}$(Adju), $\Delta\theta_{0}$(Proc), $\Delta\theta_{0}$(Org), and $\Delta\theta_{0}$(Tech) are the corresponding differences of perturbations for Seq, Host, Adj, Proc, and Org. For details, please check Ref. [16]

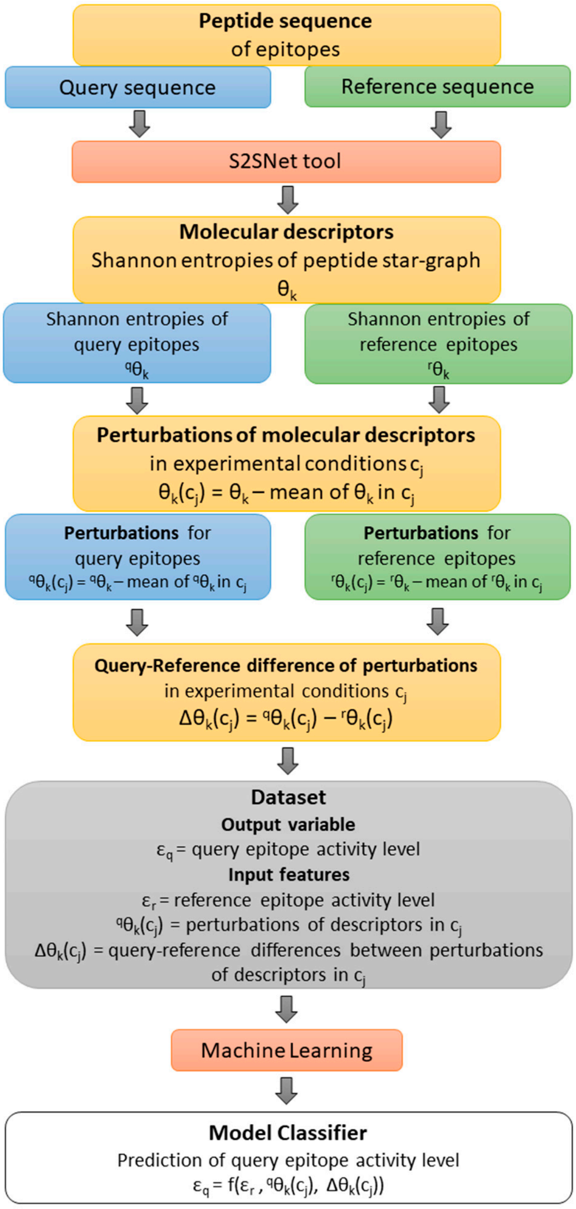
Peptide sequence
of epitopes
Query sequence
Reference sequence
S2SNet tool
Molecular descriptors
Shannon entropies of peptide star-graph
θ_k
Shannon entropies of
query epitopes
${}^{q}\theta_k$
Shannon entropies of
reference epitopes
${}^{r}\theta_k$
Perturbations of molecular descriptors
in experimental conditions c_j
$\theta_k(c_j) = \theta_k$ – mean of θ_k in c_j
Perturbations for
query epitopes
${}^{q}\theta_k(c_j) = {}^{q}\theta_k$ – mean of ${}^{q}\theta_k$ in c_j
Perturbations for
reference epitopes
${}^{r}\theta_k(c_j) = {}^{r}\theta_k$ – mean of ${}^{r}\theta_k$ in c_j
Query-Reference difference of perturbations
in experimental conditions c_j
$\Delta\theta_k(c_j) = {}^{q}\theta_k(c_j) - {}^{r}\theta_k(c_j)$
Dataset
Output variable
ε_q = query epitope activity level
Input features
ε_r = reference epitope activity level
${}^{q}\theta_k(c_j)$ = perturbations of descriptors in c_j
$\Delta\theta_k(c_j)$ = query-reference differences between perturbations
of descriptors in c_j
Machine Learning
Model Classifier
Prediction of query epitope activity level
$\varepsilon_q = f(\varepsilon_r, {}^{q}\theta_k(c_j), \Delta\theta_k(c_j))$

Figure 4. Methodology flow for building models to predict epitope activity level.

Due to the classification task, the epitopes were grouped into two classes according to the intensity of the immunogenic response related to this epitope (ε). The resulting classes are positive-high epitopes ($\varepsilon = 1$) and positive-intermediate epitopes ($\varepsilon = 0$). Since the biological assays were used to determine the intensity of the immunogenic response under different experimental conditions cj, ε was described as a function of cj. ${}^{q}\theta_{k}(c_{j})$ / ${}^{r}\theta_{k}(c_{j})$ are perturbations of ${}^{q}\theta_{k}$ / ${}^{r}\theta_{k}$ for c_j factor/experimental conditions such as 83,683 peptides sequences (Seq), 1448 epitope organisms (Org), 323 host organisms (Host), 15 types of in vivo processes (Proc), 28 experimental techniques (Tech), and 505 adjuvant additives

(Adj). $\Delta\theta_k(c_j)$ are the differences between the perturbations of Shannon entropies for query and reference peptides for c_j factors/experimental conditions (${}^{q}\theta_k(c_j) - {}^{r}\theta_k(c_j)$). After duplicate removal (dataset preprocessing), the dataset contains 709,100 instances and 10 features.

Given that the corrected dataset was slightly unbalanced and we used different ML methods, there was a need of class weighting for the Machine Learning classifiers (Class 0: 0.77, Class 1: 1.42 or Class 0: 1.00, Class 1: 1.84). Seven Machine Learning methods were used: k-nearest neighbors algorithm (KNN) [19], support vector machine (SVM linear and SVM non-linear based on radial basis functions, RBF) [20], logistics regression (LR) [21], decision tree (DT) [22], random forest (RF) [23], and XGBoost—an optimized distributed gradient boosting library (XGB) [24]. A more detailed explanation about the design methodology can be found in the previous works of the authors [29,30]. All the calculations used python/sklearn with Jupyter notebooks. The pipelines contain random stratified five-fold splits (specific to unbalanced datasets to maintain the same ratio of instances by class in the splits) for an outer five-fold cross-validation. Five-fold CV means that for each fold, 80% of the dataset was the training subset and 20% the test subset. For the same reason of the unbalanced classes, the accuracy metrics was removed with the Area Under the Receiver Operating Characteristics (AUROC) [17]. The project offers the possibility of reproducing the results by making available all the scripts, dataset, and results at https://gitlab.com/muntisa/machine-learning-for-peptide-epitopes/.

KNN is one of the most commonly known non-parametric classifiers in the ML field, which assigns an unclassified sample to the same class as the nearest of k samples in the training set [19]. This project uses $k = 5$. In SVM, the input data is non-linearly mapped to a higher dimensionality space, where a linear decision surface can be established [20] using Gaussian radial basis (RBF) kernel functions. LR [21] represents a linear model able to estimate the probability of a binary response using different factors.

DT is a set of decision rules inferred from the features into a tree structure rules (the paths from root to leaf represent classification rules) [22]. RF represents its aggregating decision trees (in parallel) [23]. Thus, RF was characterized by low-bias, low correlation between individual trees, and high variance. Another tree-based ensemble method is XGB—sequential trees [31] with weak classifiers to correct errors.

In the first step, the AUROC of the seven ML methods were statistically tested using the boxplots of the AUROC values for all folds. In the second step, the best ML classifier was tested for different hyperparameters (ex: for RF—number of trees/estimators). The scripts automatically calculated all the results and plotted the boxplots.

Author Contributions: Conceptualization, C.R.M. and M.G.; methodology, C.R.M.; software, C.R.M.; data analysis with Machine Learning and scripting, C.R.M., M.G., N.P., and J.D.; writing—original draft preparation, C.R.M. and Y.G.M.-A.; writing—review and editing, C.R.M., M.G., N.P., and J.D.; visualization, C.R.M.; supervision, A.P. and J.D.; project administration, A.P. and J.D.

Funding: This research and the APC were funded by NAME OF FUNDER, grant number PI17/01826 (Collaborative Project in Genomic Data Integration (CICLOGEN) funded by the Instituto de Salud Carlos III from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER)—"A way to build Europe."

Acknowledgments: This project was also supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia ED431D 2017/16 and "Drug Discovery Galician Network" Ref. ED431G/01 and the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23), and finally by the Spanish Ministry of Economy and Competitiveness through the funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union. Additional support was offered by the Consolidation and Structuring of Competitive Research Units—Competitive Reference Groups (ED431C 2018/49) and Accreditation, Structuring, and Improvement of Consolidated Research Units and Singular Centers (ED431G/01), funded by the Ministry of Education, University and Vocational Training of the Xunta de Galicia endowed with EU FEDER funds. Last, the authors also acknowledge research grants from the Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P), the Basque government (IT1045-16), and support of Ikerbasque, the Basque Foundation for Science.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ML	Machine Learning
QSAR	Quantitative Structure-Activity Relationship
MHC	Major histocompatibility
IEDB	Immune Epitope Database
CV	Cross-validation
S2SNet	Sequences to Star Networks software
LDA	Linear discriminant analysis
KNN	k-nearest neighbors algorithm
SVM	Support vector machine
RBF	Radial basis functions
LR	Logistics Regression
DT	Decision tree
RF	Random forest
XGB	Optimized distributed gradient boosting
AUROC	Area Under the Receiver Operating Characteristics

References

1. Wilkins, M. Proteomics data mining. *Expert Rev. Proteomics* **2009**, *6*, 599–603. [CrossRef]
2. Patronov, A.; Doytchinova, I. T-cell epitope vaccine design by immunoinformatics. *Open Biol.* **2013**, *3*, 120139. [CrossRef]
3. Andreotti, R. Performance of two Bm86 antigen vaccin formulation against tick using crossbreed bovines in stall test. *Rev. Bras. Parasitol. Vet.* **2006**, *15*, 97–100.
4. Rodriguez, M.; Rubiera, R.; Penichet, M.; Montesinos, R.; Cremata, J.; Falcon, V.; Sanchez, G.; Bringas, R.; Cordoves, C.; Valdes, M.; et al. High level expression of the B. microplus Bm86 antigen in the yeast Pichia pastoris forming highly immunogenic particles for cattle. *J. Biotechnol.* **1994**, *33*, 135–146. [CrossRef]
5. Huang, J.; Honda, W. CED: A conformational epitope database. *BMC Immunol.* **2006**, *7*, 7. [CrossRef]
6. Walsh, S.J.; Rau, L.M. Autoimmune diseases: A leading cause of death among young and middle-aged women in the United States. *Am. J. Public Health* **2000**, *90*, 1463. [CrossRef]
7. Koşaloğlu-Yalçın, Z.; Lanka, M.; Frentzen, A.; Logandha Ramamoorthy Premlal, A.; Sidney, J.; Vaughan, K.; Greenbaum, J.; Robbins, P.; Gartner, J.; Sette, A. Predicting T cell recognition of MHC class I restricted neoepitopes. *Oncoimmunology* **2018**, *7*, e1492508. [CrossRef]
8. Sanchez-Trincado, J.L.; Gomez-Perosanz, M.; Reche, P.A. Fundamentals and Methods for T- and B-Cell Epitope Prediction. *J. Immunol. Res.* **2017**, *2017*, 2680160. [CrossRef]
9. Vita, R.; Overton, J.A.; Greenbaum, J.A.; Ponomarenko, J.; Clark, J.D.; Cantrell, J.R.; Wheeler, D.K.; Gabbard, J.L.; Hix, D.; Sette, A. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **2014**, *43*, D405–D412. [CrossRef]
10. Concu, R.; Dea-Ayuela, M.A.; Perez-Montoto, L.G.; Bolas-Fernandez, F.; Prado-Prado, F.J.; Podda, G.; Uriarte, E.; Ubeira, F.M.; Gonzalez-Diaz, H. Prediction of enzyme classes from 3D structure: A general model and examples of experimental-theoretic scoring of peptide mass fingerprints of Leishmania proteins. *J. Proteome Res.* **2009**, *8*, 4372–4382. [CrossRef]
11. Ran, T.; Liu, Y.; Li, H.; Tang, S.; He, Z.; Munteanu, C.R.; Gonzalez-Diaz, H.; Tan, Z.; Zhou, C. Gastrointestinal Spatiotemporal mRNA Expression of Ghrelin vs Growth Hormone Receptor and New Growth Yield Machine Learning Model Based on Perturbation Theory. *Sci. Rep.* **2016**, *6*, 30174. [CrossRef]
12. Ferreira da Costa, J.; Silva, D.; Caamaño, O.; Brea, J.M.; Loza, M.I.; Munteanu, C.R.; Pazos, A.; García-Mera, X.; González-Díaz, H. Perturbation Theory/Machine Learning Model of ChEMBL Data for Dopamine Targets: Docking, Synthesis, and Assay of New l-Prolyl-l-leucyl-glycinamide Peptidomimetics. *ACS Chem. Neurosci.* **2018**, *9*, 2572–2587. [CrossRef]
13. Kleandrova, V.V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J.M.; Speck-Planche, A.; Cordeiro, M.N. Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environ. Sci. Technol.* **2014**, *48*, 14686–14694. [CrossRef]

14. Gonzalez-Diaz, H.; Perez-Montoto, L.G.; Ubeira, F.M. Model for vaccine design by prediction of B-epitopes of IEDB given perturbations in peptide sequence, in vivo process, experimental techniques, and source or host organisms. *J. Immunol. Res.* **2014**, *2014*, 768515. [CrossRef]
15. Vázquez-Prieto, S.; Paniagua, E.J.A.; Ubeira, F.M.; González-Díaz, H. QSPR-Perturbation Models for the Prediction of B-Epitopes from Immune Epitope Database: A Potentially Valuable Route for Predicting "In Silico" New Optimal Peptide Sequences and/or Boundary Conditions for Vaccine Development. *Int. J. Pept. Res. Ther.* **2016**, *22*, 445–450. [CrossRef]
16. Martínez-Arzate, S.G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Díaz-Albiter, H.M.; Vázquez-Chagoyán, J.C.; González-Díaz, H. PTML Model for Proteome Mining of B-Cell Epitopes and Theoretical–Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *J. Proteome Res.* **2017**, *16*, 4093–4103. [CrossRef]
17. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *J. Pattern. Recogn.* **1997**, *30*, 1145–1159. [CrossRef]
18. Hao, J.; Ho, T.K. Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *J. Educ. Behav. Stat.* **2019**, *44*, 348–361. [CrossRef]
19. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
20. Patle, A.; Chouhan, D.S. SVM kernel functions for classification. In Proceedings of the 2013 International Conference on Advances in Technology and Engineering (ICATE), Mumbai, India, 23–25 January 2013; pp. 1–9.
21. Peduzzi, P.; Concato, J.; Kemper, E.; Holford, T.R.; Feinstein, A.R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **1996**, *49*, 1373–1379. [CrossRef]
22. Swain, P.H.; Hauska, H. The decision tree classifier: Design and potential. *IEEE Trans. Geosci. Electron.* **1977**, *15*, 142–147. [CrossRef]
23. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
24. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
25. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef]
26. Cristianini, N. Fisher Discriminant Analysis (Linear Discriminant Analysis). In *Dictionary of Bioinformatics and Computational Biology*; Wiley-Liss: Hoboken, NJ, USA, 2004.
27. Kubinyi, H. *QSAR: Hansch analysis and related approaches. Methods and Principles in Medicinal Chemistry*; VCH: Weinheim, Germany, 2008; Volume 1.
28. Munteanu, C.R.; Magalhaes, A.; Duardo-Sánchez, A.; Pazos, A.; González-Díaz, H. S2Snet: A tool for transforming characters and numeric sequences into star network topological indices in chemoinformatics, bioinformatics, biomedical, and social-legal sciences. *Curr. Bioinform.* **2013**, *8*, 429–437. [CrossRef]
29. Fernandez-Lozano, C.; Gestal, M.; Munteanu, C.R.; Dorado, J.; Pazos, A. A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ* **2016**, *4*, e2721. [CrossRef]
30. Aguiar-Pulido, V.; Seoane, J.A.; Gestal, M.; Dorado, J. Exploring Patterns of Epigenetic Information with Data Mining Techniques. *Curr. Pharm. Des.* **2013**, *19*, 779–789. [CrossRef]
31. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **2001**, *29*, 1189–1232. [CrossRef]