bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TAXyl: An in-silico method for predicting the thermal activity for xylanases from GH10 and GH11 families

Mehdi Foroozandeh Shahraki[a], Kiana Farhadyar[a], Kaveh Kavousi[a], Mohammad Hadi Azarabad[a], Amin Boroomand[b], Shohreh Ariaeenejad[c]*, and Ghasem Hosseini Salekdeh[c,d]*

[a] Laboratory of Complex Biological Systems and Bioinformatics (CBB), Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran

[b] School of Natural Sciences, University of California Merced, Merced, California, United States of America

[c] Department of Systems and Synthetic Biology, Agricultural Biotechnology Research Institute of Iran (ABRII), Agricultural Research Education and Extension Organization (AREO), Karaj, Iran

[d] Department of Molecular Sciences, Macquarie University, Sydney, NSW, Australia

*Corresponding authors:

Ghasem Hosseini Salekdeh (h_salekdeh@abrii.ac.ir; hsalekdeh@yahoo.com) and Shohreh Ariaeenejad (sh.ariaee@abrii.ac.ir; shariaee@gmail.com)

Department of Systems Biology, Agricultural Biotechnology Research Institute of Iran (ABRII), Karaj, Iran. P. O. Box: 31535-1897, Tel.: +98 26 32703536, Fax: +98 26 32704539

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Xylanases are a class of enzymes with numerous industrial applications and are involved in the degradation of xylose polysaccharide, which is present in lignocellulosic biomass. The optimum temperature of enzymes is the indicator of their thermal activity and is an essential factor to be considered when choosing an appropriate biocatalyst for a particular purpose. Therefore, in-silico prediction of this enzymatic attribute is a significant cost and time-effective step in the effort to identify and characterize novel enzymes. The objective of this study was to develop an accurate computational method to predict the thermal activity status of xylanases from glycoside hydrolases families 10 and 11, the most prevalent known xylanase families. Here we present TAXyl (Thermal Activity Prediction for Xylanase), a new sequence-based machine learning method that has been trained using a selected combination of various physicochemical protein features. This ensemble of four supervised learning algorithms discriminates mesophilic, thermophilic, and hyper-thermophilic xylanases based on their optimum temperature with the process of soft-voting. TAXyl's performance was ultimately evaluated through multiple iterations of six-fold cross-validations, and it exhibited a mean accuracy of ~0.94, F1-score of ~0.91, and MCC of ~0.9. Additionally, the model was tested on previously unseen data and depicted relatively similar performance. To the best of our knowledge, this tool is the most accurate and practical prediction tool currently available and operating on this class of enzymes. TAXyl is freely accessible as a web-service at http://arimees.com/ and provides users with several features to facilitate the characterization of GH10 and GH11 xylanases.

Keywords:

xylanase, machine learning, optimum temperature, ensemble learning, sequence-based classification

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1. Introduction

Endo-1,4-beta-xylanase (EC 3.2.1.8) catalyzes the degradation of xylan, a component of hemicellulose, into xylooligosaccharides and D-xylose. Xylanases are currently being used for a broad spectrum of industrial applications, such as pulp and paper, food, textiles, biofuel, animal feed, and beverages (1). A considerable ratio of known xylanases are from glycoside hydrolases (GH) families 10 and 11 (2).

One of the essential attributes of enzymes is their optimum temperature, in which they exhibit their maximum relative activity. An enzyme's optimum temperature the indicator of its thermal activity (3). On the one hand, high temperature increases substrates' solubility and bioavailability, accelerates molecular dynamics, and decreases the probability of microbial contamination significantly (4,5). On the other hand, numerous biological processes are carried out at mild or cold temperatures (6). Therefore, various research studies have been designed to predict the enzymes' optimum temperature in order to introduce novel appropriate biocatalysts for specific processes.

New enzymes are being discovered at an increased pace, thanks to the advances in sequencing technologies. Unlike culture-dependent methods that are unable to cultivate up to 99% of microorganisms, culture-independent methods such as metagenomics enable the extended exploration of the natural diversity within an environmental sample (7). Metagenomics is the direct analysis of genetic material found within an environmental sample (8,9). This field of study is a comparatively new culture-independent method to analyze microbial communities, their functional genes, and phylogenetic properties. Therefore, metagenomics provides access to almost all the genetic material inside an environmental sample. However, experimental functional annotation of newly identified genes is becoming a significant challenge (10,11). Utilizing an in-silico approach to address this problem can make a notable contribution. Consequently, the availability of metagenomic data is rapidly

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

increasing, hence employing computational methods instead of wet-lab experiments in order to identify new enzymes with specific properties can effectively reduce the costs and make the process much faster.

The primary structure of a protein is one of the most important factors affecting an enzyme's thermal activity, and there is a strong correlation between this functional property and its sequence. Many studies have used sequence similarity-based methods in order to predict different properties of proteins. However, there are many cases where sequence similarity does not directly correlate with the functional resemblance, as proteins with a nearly similar primary structure can sometimes depict unique properties or functional similarity can be observed among proteins with different amino acid sequences (12). Machine learning approaches have been successfully applied to predict various properties of proteins such as tertiary structure (13),(14), function (15), localization (16), thermal stability (17), etc.(18),(19). These computational methods are capable of learning more complex relationships between the primary structure of proteins and their different properties (20–22).

Numerous studies have presented *in-silico* methods for predicting enzymatic attributes. Discrimination between thermophilic and mesophilic proteins by using machine learning methods was the focus of a study by Gromiha and Suresh (23). Similarly, Tang et al. used support vector machines (SVM) to develop a two-step method for discriminating thermophilic proteins (24) and amino acid compositions have been the basis of a statistical method for a similar task (25). Pucci et al. presented a statistical approach to predict thermostability (26), and Jia et al. designed a thermostability predictor tool (27). AcalPred is another similar study that utilizes SVM to classify acidic and alkaline enzymes based on their primary structure (28). In another research, Ariaeenejad et al. applied a regression model based on a pseudo amino acid composition (PAAC) (29) to predict the optimum temperature and pH of xylanase in strains of Bacillus subtilis enzymes (30). Genetic Algorithm-Artificial Neural Network

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(GA-ANN) have been employed for the optimization of xylanase production for industrial purposes (31,32). In order to find features with the highest correlation with the thermostability of proteins, K-mean algorithm clustering method and a decision tree have been employed (33). Panja et al. found that the prevalence of smaller non-polar amino-acids, more hydrophobicity, and salt-bridges are some shared characteristics among most thermophilic proteins (34). Moreover, feature selection methods such as recursive feature elimination (RFE) have been previously used by some studies in order to choose best sequence-extracted features. As an instance, Kumar et al. employed RFE with SVM to classify the enzymes' function into different classes and subclasses (35).

The increasing use of new high throughput technologies can rapidly produce enormous amounts of data, while the processes of getting access to the protein's tertiary and quaternary structures are much slower. Therefore, an accurate sequence-based approach with acceptable agility is in demand. The objective of this study was to design and implement a multi-step method for the classification of the thermal activity of xylanases from glycoside hydrolases families 10 and 11 based on their optimum temperature. Since most of the available data in the literature belong to GH10 and GH11 families, we focused on the members of these two protein families. To the best of our knowledge, this is the first time that a combination of different protein descriptors is calculated, selected, and used to train multiple machine-learning algorithms to make predictions on xylanase optimum temperature by an ensemble voting method. We presented TAXyl (Thermal Activity prediction for xylanase), a prediction web-server for the thermal activity of xylanases. The performance of TAXyl was evaluated using multiple cross-validation tests and also on previously unseen data.

2. Methods

2.1. Dataset preparation

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A new dataset from GH families 10 and 11, which constitutes a considerable ratio of known xylanases, had to be collected. Even though this makes the scope narrower, it can help the final estimation be more accurate and reliable. The National Center for Biotechnology Information (NCBI) database was explored by searching for thermoactive or thermostable xylanases, and 254 results were found. After removing the records without the exact optimum temperature report, the remaining sequences were divided into two groups of families 10 and 11. Afterward, the BRENDA database was explored for xylanases with reported optimum temperature, and newly collected data were added to the previous dataset. The Uniprot website was finally searched with the same strategy, and new samples were collected.

Redundant or highly similar samples were removed using the CD-Hit tool, which clusters highly-homologous sequences, with a 0.9 cut-off (36). In some protein sequences, a few amino acids are unknown. These residues are represented by the character "X." Since the existence of such noise could potentially interfere with the learning process and feature extraction tools are designed for 20 amino acid residues, all unknown amino acids were removed from the sequences.

The final dataset consisted of 145 different xylanases from GH families 10 and 11 with optimum temperature ranging from 25°C to 95°C. These samples were labeled accordingly into three different categories: 1) "Non-thermoactive" with optimum temperature below 50°C; 2) "thermoactive" with the optimum temperature between 50°C and 75°C, and 3) "hyper-thermoactive" with the optimum temperature above 75°C. Fig. 1 shows the proportion of each thermal class and GH family in the dataset.

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

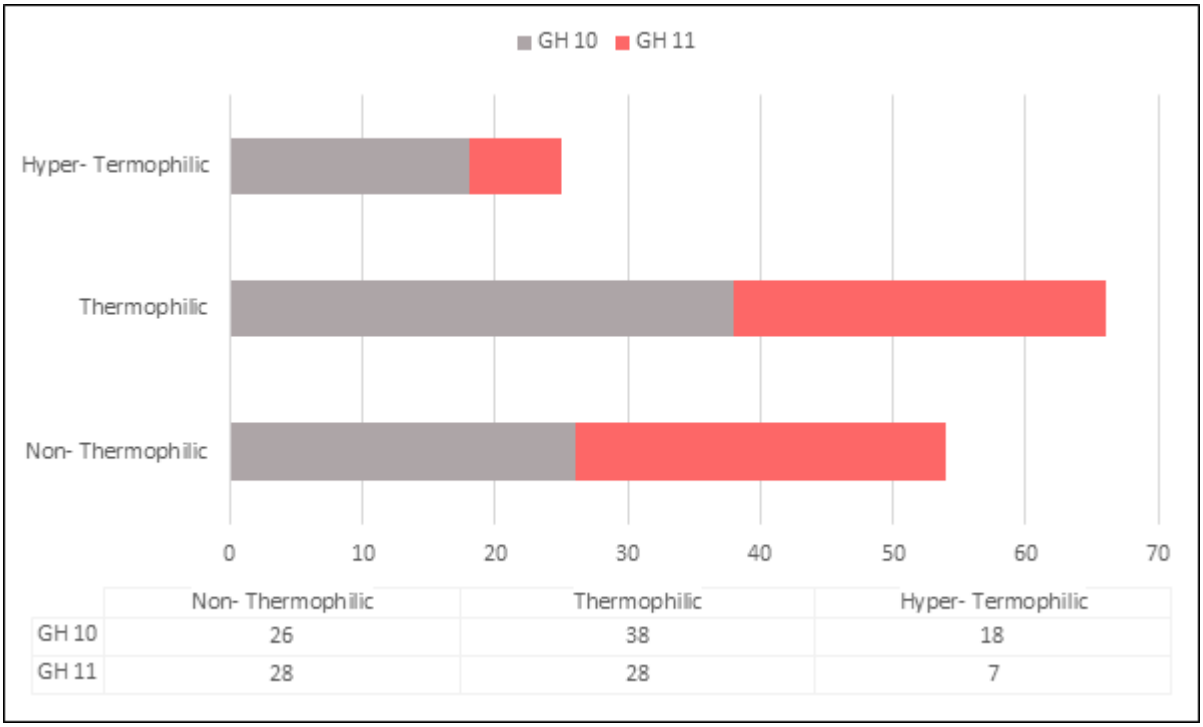


Figure 1 The dataset is divided into three categories. *Samples in the dataset are from GH10 and GH11 families with optimum temperatures ranging from 25°C to 95°C. These enzyme samples are divided into three thermal activity classes. This figure illustrates the proportion of each class and GH family in the dataset.*

2.2. Feature extraction

Using an appropriate set of features is undoubtedly one of the most crucial elements in creating an efficient classifier. Because there is not much precise evidence regarding the most related features to thermal activity, in this study, various protein descriptors were computed using the PyDPI python package (37).

The PyDPI computed protein features are 15 descriptor types, which are from six main groups. Amino acid composition (AAC), dipeptide composition (2AAC) and tripeptide composition (3AAC), represent their fraction in the protein sequence. Unlike previous feature groups, pseudo amino acid composition (PAAC) and amphiphilic pseudo amino acid composition (APAAC), try to evade missing the sequence-order information and combine that with the composition data (38),(39). Conjoint triad

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

features (CTF) cluster twenty amino acids into several classes based on their dipoles and the number of side chains (40). CTF considers the properties of an amino acid and its neighboring ones while regarding any amino acid triads as a unit. Other features are calculated by taking structural and physiochemical properties into account. Three autocorrelation descriptors, including normalized Moreau−Broto, Geary, and Moran autocorrelations, attempt to describe the amount of correlation among peptide or protein sequences. Composition, transition, distribution descriptors (CTD), sequence-order coupling number (SOCN), and quasi-sequence-order (QSO) delineate the distribution pattern of amino acids along a protein sequence in terms of structural and physicochemical attributes.

As a result, a feature vector with 10,074 descriptors was calculated for each protein sequence. Many of these features were duplicates and were removed afterward. Due to the different ranges of the descriptors, all raw values had to be rescaled into the same range. This transformation was done using MinMaxScaler.

2.3. Feature selection

Different protein descriptors were used separately to train the same models and depicted different prediction performances. These individually evaluated features were then combined and submitted to steps of feature selection which resulted in a significant improvement in prediction performmance. All features do not contribute equally to the final prediction. Thus, a process of feature selection is necessary in order to find the most relevant descriptors for optimum temperature classification and to remove ones of less importance.

Filter feature selection methods utilize a statistical measure to rank features based on their score. For this step, an f-test was applied as the filter method, and the best features were chosen using the SelectKBest method for the next selection step. Recursive feature elimination (RFE) trains a machine learning model using input features and ranks them based on their contribution to the prediction. The

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RFE method was executed with logistic regression, a classification algorithm, and most relevant features were selected. As a result, the final feature vector was prepared with 512 dimensions.

2.4. Model selection and training

A plethora of classification algorithms are currently available. Each algorithm is built based on different theories. In order to find the best methods for our problem, various classification algorithms were tested, including multilayer perceptron (MLP), decision tree, Gaussian Naïve Bayes, Gaussian process, AdaBoost, KnearestNeighbors and support vector machine (SVM), all of which were applied from the sci-kit learn python package (41) Four of these supervised learning algorithms with the best accuracies were chosen. These four classifiers were MLP, SVM, Gaussian process classifier, and Gaussian Naïve Bayes.

An MLP with four hidden layers was constructed to address this classification. SVM classifiers form hyperplanes that categorize inputs into different classes. The hyperplane can be constructed using various kernel functions, including linear, radial basis, and polynomial. Our support-vector machine used a radial basis function (RBF) kernel to predict the protein's thermal activity. Naïve Bayes is based on Bayes theorem and can be applied for this classification problem. This algorithm is relatively fast since it only calculates the probability of each class and the probability of each class given different sets of inputs. Gaussian Naïve Bayes is a popular supervised learning algorithm for dealing with continuous data, and this method was used for this problem (42). Gaussian process classifier is another classification algorithm that was employed with the RBF kernel (43). Afterward, an ensemble classifier was implemented to decide the final output based on soft voting among the four mentioned classifiers (44). In the process of soft voting, each model returns an array representing the probabilities of each class, and the ensemble classifier decides the final answer based on the weighted average of class probabilities.

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The pipeline mentioned above required several parameter tuning steps, all of which were done using the GridSearchCV method. Grid searching is the process of testing the model with various hyperparameters and finding the optimum configuration. The sci-kit learn python package was used several times during this study (41)

2.5. Evaluation criteria

Since this is a multi-class classification problem, accuracy, macro-recall, macro-precision, the macro-f1 score, and the Matthews correlation coefficient were used as evaluation metrics. These metrics were calculated using the following formulas:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TN}{TN+FP} \quad (3)$$

$$F1 = 2.\frac{Precision*Recall}{Precision+Recall} \quad (4)$$

$$MCC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{[TP+FP][TP+FN][TN+FP][TN+FN]}} \quad (5)$$

Here, TP (true positive) and TN (true negative) are respectively positive and negative examples that were correctly predicted. Similarly, FP (false positive) and FN (false negative) are positive and negative examples that were mistakenly classified. Matthew's correlation coefficient represents the correlation coefficient between the predictions and the actual values (45) The "macro" prefix refers to getting the average of each metric across the three different classes.

For the evaluation step, samples were randomly split into two subsamples (85% as the training set and 15% as the test set) five times, and each time 20 iterations of the six-fold cross-validation were done on the training set. Subsequently, the models were tested on the unseen subsample (test set). This

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

means that our classifiers were evaluated through 100 iterations of six-fold cross-validations and were tested on unseen data five times to assure that the models are robust. In the six-fold cross-validation step, the training set was randomly split into six equal subsamples, five of which were used as training sets, and one subsample was then used for testing the models with different evaluation metrics. This was done six times, leaving out one subsample each time for testing. Figure 2 is the schematic diagram of the steps of development and evaluation of the current prediction model.

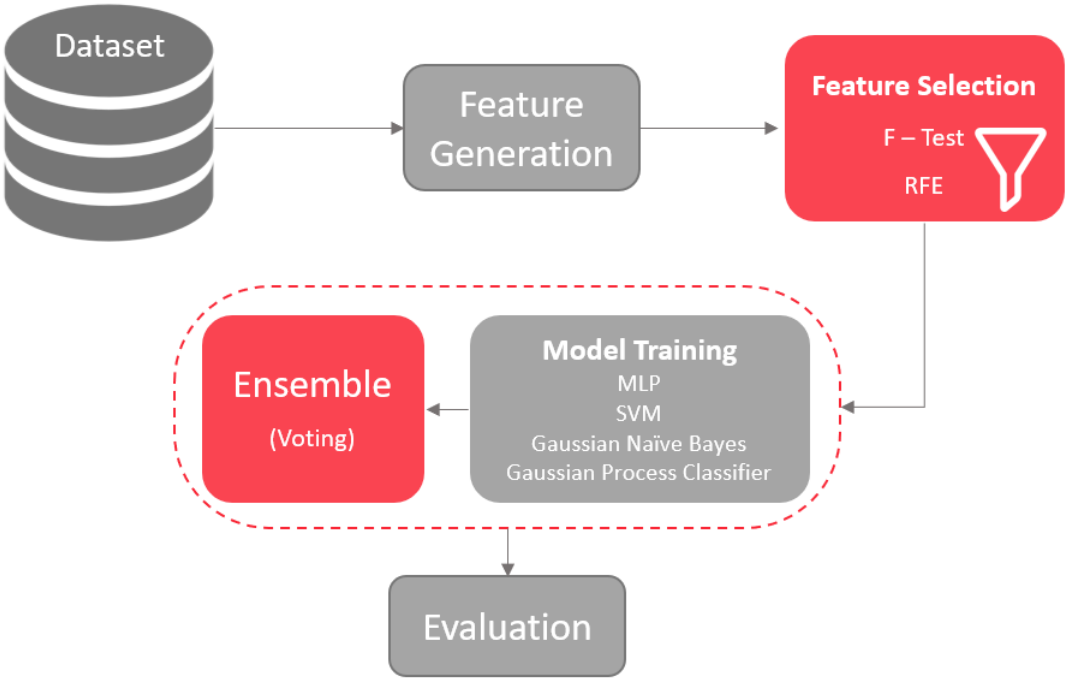


Figure 2 Schematic diagram of the workflow *. The figure above illustrates the steps for the development and evaluation of proposed prediction model.*

3. Results

3.1. Feature importance analysis and feature selection

Generated protein descriptors encapsulate various molecular and sequential information with different degrees of relevance to the enzyme's thermal activity. In order to obtain a better perception of the relationship between different descriptors and the prediction performance, the same model was

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

trained using different sets of features. A summary of generated features, their feature groups, and their dimensions are presented in Table 1.

Table 1 A summary of different generated features from enzyme sequences that were used in this study

Feature Group	**Number of Descriptors**
Amino acid, Dipeptide and Tripeptide compositions	8420
Autocorrelations (Moreau–Broto, Geary, and Moran)	720
Composition, Transition, Distribution	147
Conjoint triad	512
Quasi-Sequence-order and Sequence-order coupling number	190
Pseudo Amino Acid Composition and Amphiphilic Pseudo Amino Acid Composition	85

The average of metrics for different models were calculated and then compared to combined features that went through feature Selection steps. Amino acid composition, Pseudo amino acid composition, dipeptide composition, amphiphilic pseudo amino acid composition, and Quasi sequence order were ranked in descending order by their performances when implemented individually. This implies their order in terms of importance and relevance to thermal activity. A complete representation of feature importance analysis is presented in Table 2 and Fig. 3. Our results showed that a selected combination of different descriptors augments the prediction performance noticeably.

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2 : Mean of different evaluation metrics for the same models when trained with different sets of features. Feature groups are sorted in an ascending order of accuracy.

	Accuracy	Macro-Recall	Macro-Precision	Macro-F1	MCC
SOCN	0.383	0.33	0.231	0.245	-0.009
MoreauBroto	0.397	0.385	0.351	0.332	0.064
CTF	0.445	0.433	0.409	0.385	0.13
3AAC	0.458	0.461	0.437	0.391	0.19
Moran	0.467	0.441	0.428	0.403	0.16
Geary	0.481	0.459	0.443	0.417	0.195
CTD	0.488	0.472	0.455	0.427	0.205
QSO	0.557	0.544	0.537	0.505	0.327
APAAC (lambda=15)	0.569	0.556	0.555	0.529	0.327
2AAC	0.588	0.566	0.591	0.546	0.362
PAAC (lambda=15)	0.597	0.604	0.602	0.572	0.383
AAC	0.605	0.603	0.612	0.582	0.387
Combined + Feature Selection	**0.905**	**0.882**	**0.895**	**0.877**	**0.855**

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

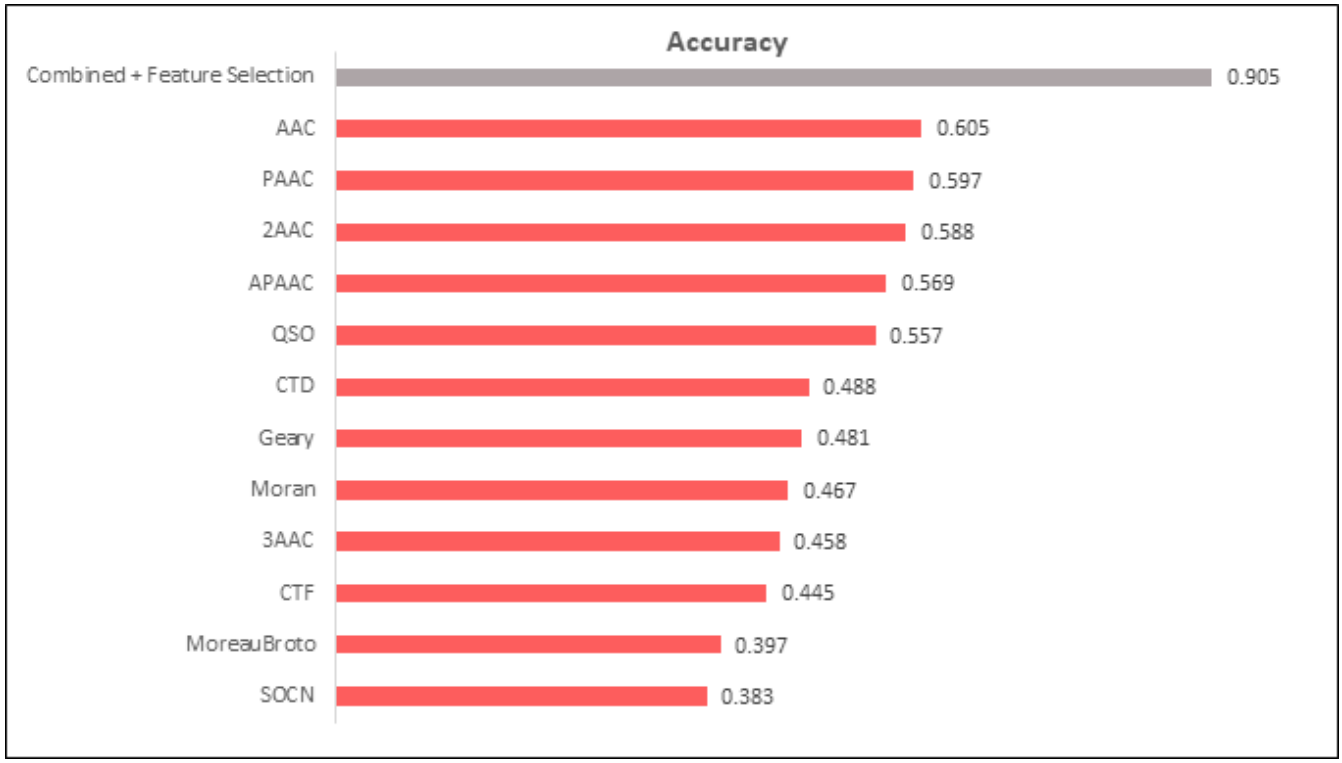


Figure 3 : Feature Importance Analysis. *This chart represents the importance and relevance of each generated feature set with the task of thermal activity prediction. Mean of accuracy for the same model, trained with different feature sets, are provided and compared. This representation implies the importance of feature selection steps in order to obtain a better prediction performance.*

In the process of filter feature selection, both chi-2 test and f-test were both used as filter methods, and the f-test showed a slightly better result based on final evaluations. For the filter feature selection, we used the SelectKBest method, and the value of K was chosen to be 3500 based on the better prediction performance after multiple tests. Afterward, 3500 selected features were used in the feature selection step with the RFE method. In the RFE step, logistic regression was employed, and 512 most contributing features were chosen as the final feature vector. Again, the dimension of 512 was opted due to better performance.

3.2. Model performance

As it is shown in Table 3, our four models depicted reasonably high performance, achieving acceptable scores at different evaluation metrics. Although all four models mostly depicted agreement

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

on correct predictions, in case of mistake, outputs were diverse. This diversity is because of the different algorithms that each model implements to make the prediction. Therefore, an ensemble method such as Voting Classifier could be a promising synergistic approach to get a higher overall performance since it enables us to exploit multiple learning algorithms for a single prediction. Soft voting refers to the process in which the final output is determined based on the computed probabilities of all models for each class. Since our four models had different capabilities, voting was executed by assigning uneven weights to each model. Reported performance metrics were computed through 100-time six-fold cross-validation tests. In each cross-validation (CV) iteration, the dataset was shuffled with different random seeds before splitting. The ensemble method demonstrated a slight improvement, out-performing the most accurate individual model, which was MLP. Table 3 demonstrates the performance of different individual classification methods in comparison to the ensemble voting classifier.

Table 3 A numerical comparison of different models' classification performance *through 100 iteration of six-fold cross-validation.*

	Accuracy	**Macro-Recall**	**Macro-Precision**	**Macro-F1**	**MCC**
Multi-Layer Perceptron	0.932	0.905	0.936	0.908	0.893
Gaussian Naive Bayes	0.846	0.798	0.855	0.8	0.755
Support Vector Machine	0.925	0.9	0.925	0.901	0.881

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gaussian Process	0.844	0.748	0.716	0.717	0.759
Voting Classifier	**0.940**	**0.917**	**0.940**	**0.919**	**0.906**

3.3. Testing the model on unseen test data

As described, our final model was chosen to be the ensemble classifier due to the achievement of better classification scores. The thermal activity prediction for xylanases (TAXyl) was tested using unseen test data (15% of the initial dataset, which was reserved at the beginning). This step was executed five times, and each time, the data was shuffled entirely. Table 4 shows the comparison of TAXyl evaluation during cross-validation and holdout method tests and Fig. 4 illustrates the TAXyl's performance through multiple cross-validation tests.

Table 4 Comparison of cross-validation and holdout method test results. *TAXyl's performance during 100 iterations of 6-fold cross validations and five iterations of test (holdout) validations on previously unseen data.*

	Accuracy	**Macro-Recall**	**Macro-Precision**	**Macro-F1**	**MCC**
Test Performance	0.94	0.92	0.95	0.93	0.91
CV Performance	0.94	0.91	0.94	0.91	0.9

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

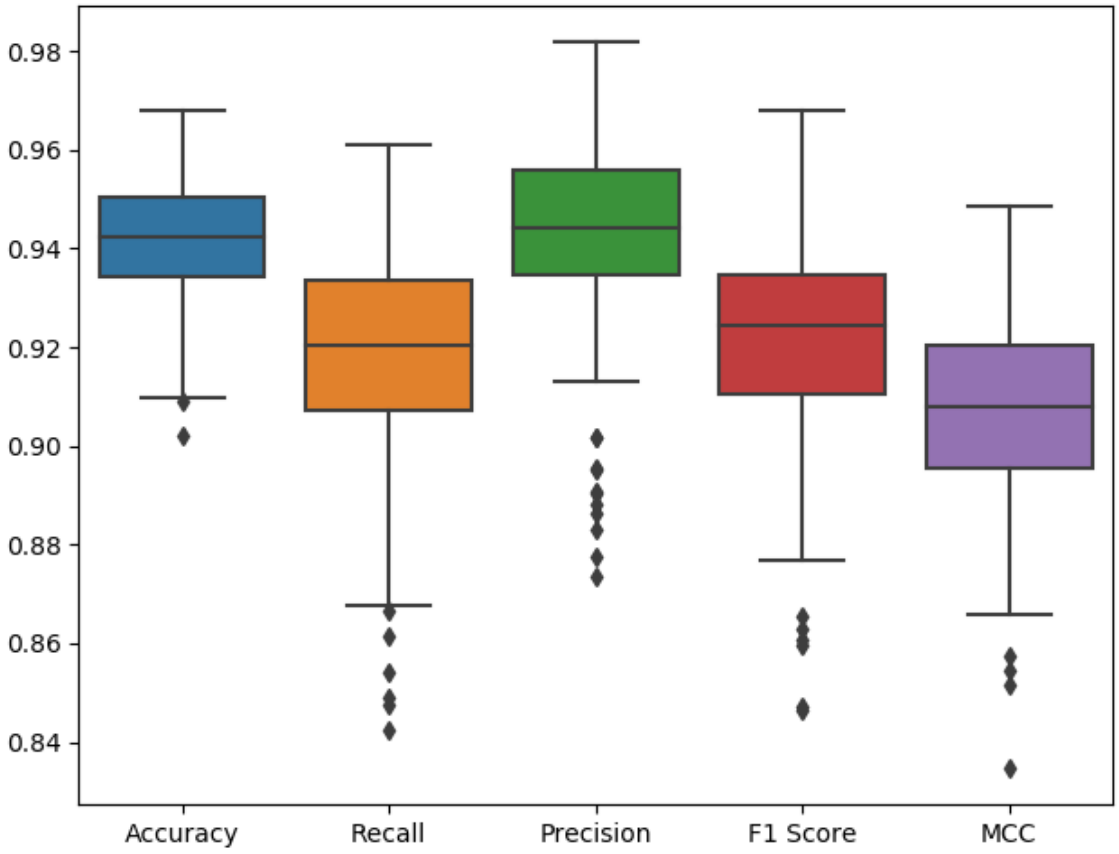


Figure 4 Box plot of TAXyl evaluation metrics during 100 times six-fold cross validations.

3.4. Online Server

TAXyl is freely available at http://arimees.com. This web service is capable of getting inputs in the forms of the FASTA file, amino acid sequence, or protein entry of xylanases from GH families 10 and 11 and returns their probable thermal activity status. TAXyl also enables users to download and export the selected features for their inserted protein sequences in CSV format for other machine learning applications. This web service is also accessible from the CBB lab website (http://cbb.ut.ac.ir), under databases and tools sub-menu.

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4. Discussion and conclusion

Xylanases are carbohydrate-active enzymes responsible for the hydrolysis of xylan polysaccharide into xylose. This group of biocatalysts have multiple industrial applications and therefore have high commercial value (1). Determining the optimum temperature of activity plays an essential role in choosing the appropriate enzymes for a specific task, making this enzymatic property substantially important. Since the advent of next-generation sequencing and its accelerating improvements, getting access to metagenomic data is becoming increasingly easier and more affordable, and the only possible way to analyze these enormous amounts of data is by fast and accurate computational methods.

In this study, we presented a novel method based on an ensemble of machine learning algorithms to predict the optimum temperature of xylanases activity from GH families 10 and 11. Because of the limited number of xylanases with a reported optimum temperature in the literature, we explored different learning methods to find ones with an acceptable interpretation of the data. TAXyl uses sequence-based and length-independent protein descriptors to train four different supervised machine learning algorithms and employs a voting classifier to integrate all individual models to obtain a more accurate prediction. As demonstrated, the ensemble method which benefits from the synergistic combination of various information sources can slightly improve the performance by taking several learning methods into account for a single task of decision making. Furthermore, the voting classifier exhibited less variance in its metrics in comparison to other methods, which stems from its greater flexibility and robustness.

In comparison to similar previous studies, TAXyl out-performs the model developed by Gromiha et al. on the prediction of GH10 and GH11 xylanases by providing the CV accuracy of 94% over 89% (23). Similarly, our model's performance was higher than the statistical method which was based on amino acid compositions (25). Although TAXyl and the two-step discrimination model of Tang et al.

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

had a relatively similar accuracy (24), our model, unlike non of the above, extends the classification ability to the third class which are the hyper-thermoactive enzymes.

Our results indicated the competence of computational methods to address common problems in bioinformatics and the capability of sequence-based and length-independent protein descriptors for training supervised learning algorithms to effectively predict general enzymatic features (26). It is clear that various protein attributes are related to their structural and functional properties. With a proper multi-step method, it is possible to predict them with minimal cost in time and expenses. We observed that amino acid composition, pseudo amino acid composition and sequence order descriptors are among the most relevant protein features for thermal activity prediction (33) Moreover, our findings indicate that the feature selection steps amended the performance considerably by enabling us to exploit the best descriptors from different feature groups and pruning the trivial ones. As presented in Table 3, all individual models were capable of reasonably accurate classifications. When testing the classifiers separately, MLP and SVM demonstrated a better performance than Gaussian Naïve Bayes and Gaussian process classifier. However, the ensemble classifier depicted a better performance by combining all four classifiers' discrimination ability.

Advancements in sequencing technologies and metagenomics have revolutionized our access to a wealth of sequence data from enzymes with possible industrial applications. In comparison with our previous two studies, in which xylanase enzymes from the metagenomic source were identified and characterized by experimental techniques (10,11), TAXyl enabled the identification of more putative thermoactive xylanases and enhanced the extended exploration of the metagenomic data as well as validating our two previously characterized enzymes. This tool can be implemented to significantly reduce the number of potential candidates of xylanases with a specific thermal activity profile before engaging the wet-lab experiments. Another potential usage for such tools is in facilitating the

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

engineering of enzymes through directed evolution to obtain biocatalysts with higher thermal stability targeted at particular industrial purposes. In case of sufficient data availability, a possible direction for future works would undoubtedly be developing similar tools to predict the structural, functional, and thermodynamic properties of other enzyme families. The TAXyl web-service is available and provides users with a reasonably accurate approximation of any GH10 or GH11 xylanase thermal activity status.

Supporting Information

Dataset which was used for this study. (.xlsx)

Availability and implementation:

Prediction online webservice: http://arimees.com/

Codes: https://github.com/mehdiforoozandeh/TAXyl

Acknowledgments

This research was supported by a grant from the Agricultural Biotechnology Research Institute of Iran (ABRII), Institute of Biochemistry and Biophysics (IBB).

Abbreviations

AAC, Amino acid composition; 2AAC, Dipeptide composition; 3AAC, Tripeptide composition; ANN, Artificial Neural Networks; APAAC, amphiphilic pseudo amino acid composition; CTD, Composition, Transition, Distribution; CTF, conjoint triad features; GH, Glycoside Hydrolase; MCC, Matthews Correlation Coefficient; MLP, Multi-Layer Perceptron; PAAC, Pseudo amino acid composition; QSO,

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Quasi-sequence Order; RBF, Radial Basis Function; RFE, Recursive Feature Elimination; SOCN, Sequence order coupling number; SVM, Support-Vector Machine

References

1. Kumar D, Kumar SS, Kumar J, Kumar O, Mishra SV, Kumar R, et al. Xylanases and their industrial applications: A review. Biochemical and Cellular Archives. 2017.

2. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J. 1991;

3. Liu L, Wang L, Zhang Z, Wang S, Chen H. Effect of codon message on xylanase thermal activity. J Biol Chem. 2012;

4. Kumar V, Verma D, Archana A, Satyanarayana T. Thermostable bacterial xylanases. In: Thermophilic Microbes in Environmental and Industrial Biotechnology: Biotechnology of Thermophiles. 2013.

5. Kumar S, Dangi AK, Shukla P, Baishya D, Khare SK. Thermozymes: Adaptive strategies and tools for their biotechnological applications. Bioresource Technology. 2019.

6. Collins T, Gerday C, Feller G. Xylanases, xylanase families and extremophilic xylanases. FEMS Microbiology Reviews. 2005.

7. Schloss PD, Handelsman J. Biotechnological prospects from metagenomics. Current Opinion in Biotechnology. 2003.

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

8. Thomas T, Gilbert J, Meyer F. Metagenomics: A guide from sampling to data analysis. In: The Role of Bioinformatics in Agriculture. 2014.

9. Gharechahi J, Salekdeh GH. A metagenomic analysis of the camel rumen's microbiome identifies the major microbes responsible for lignocellulose degradation and fermentation. Biotechnol Biofuels. 2018;

10. Ariaeenejad S, Hosseini E, Maleki M, Kavousi K, Moosavi-Movahedi AA, Salekdeh GH. Identification and characterization of a novel thermostable xylanase from camel rumen metagenome. Int J Biol Macromol. 2019;

11. Ariaeenejad S, Maleki M, Hosseini E, Kavousi K, Moosavi-Movahedi AA, Salekdeh GH. Mining of camel rumen metagenome to identify novel alkali-thermostable xylanase capable of enhancing the recalcitrant lignocellulosic biomass conversion. Bioresour Technol. 2019;

12. Sadowski MI, Jones DT. The sequence-structure relationship and protein function prediction. Current Opinion in Structural Biology. 2009.

13. Cheng J, Tegge AN, Baldi P. Machine Learning Methods for Protein Structure Prediction. IEEE Rev Biomed Eng. 2008;

14. Dehzangi A, Phon Amnuaisuk S, Ng KH, Mohandesi E. Protein fold prediction problem using ensemble of classifiers. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2009.

15. Kulmanov M, Khan MA, Hoehndorf R. DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;

16. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

17. Wu LC, Lee JX, Huang H Da, Liu BJ, Horng JT. An expert system to predict protein thermostability using decision tree. Expert Syst Appl. 2009;

18. Rawat P, Kumar S, Michael Gromiha M. An in-silico method for identifying aggregation rate enhancer and mitigator mutations in proteins. Int J Biol Macromol. 2018;

19. Zhang G, huihua G, Yi L. Stability of halophilic proteins: From dipeptide attributes to discrimination classifier. Int J Biol Macromol. 2013;

20. Shen H Bin, Chou KC. Ensemble classifier for protein fold pattern recognition. In: Bioinformatics. 2006.

21. Feng P-M, Ding H, Chen W, Lin H. Naïve Bayes Classifier with Feature Selection to Identify Phage Virion Proteins. Comput Math Methods Med. 2013;

22. Cai Y, Yang H, Li W, Liu G, Lee PW, Tang Y. Multiclassification Prediction of Enzymatic Reactions for Oxidoreductases and Hydrolases Using Reaction Fingerprints and Machine Learning Methods. J Chem Inf Model. 2018;

23. Gromiha MM, Suresh MX. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins Struct Funct Genet. 2008;

24. Tang H, Cao RZ, Wang W, Liu TS, Wang LM, He CM. A two-step discriminated method to identify thermophilic proteins. Int J Biomath. 2017;

25. Zhang G. A simple statistical method for discrimination of Thermophilic and Mesophilic Proteins based on amino acid composition. In: International Journal of Bioinformatics Research and Applications. 2013.

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

26. Pucci F, Dhanani M, Dehouck Y, Rooman M. Protein thermostability prediction within homologous families using temperature-dependent statistical potentials. PLoS One. 2014;9(3).

27. Jia L, Yarlagadda R, Reed CC. Structure based thermostability prediction models for protein single point mutations with machine learning tools. PLoS One. 2015;

28. Lin H, Chen W, Ding H. AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes. PLoS One. 2013;

29. Chou K-C. Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. Curr Proteomics. 2009;6(4):262–74.

30. Ariaeenejad S, Mousivand M, Dezfouli PM, Hashemi M, Kavousi K, Salekdeh GH. A computational method for prediction of xylanase enzymes activity in strains of Bacillus subtilis based on pseudo amino acid composition features. PLoS One. 2018;

31. Kumar V, Chhabra D, Shukla P. Xylanase production from Thermomyces lanuginosus VAPS-24 using low cost agro-industrial residues via hybrid optimization tools and its potential use for saccharification. Bioresour Technol. 2017;

32. Kumar V, Kumar A, Chhabra D, Shukla P. Improved biobleaching of mixed hardwood pulp and process optimization using novel GA-ANN and GA-ANFIS hybrid statistical tools. Bioresour Technol. 2019;

33. Ebrahimi M, Ebrahimie E. Sequence-Based Prediction of Enzyme Thermostability Through Bioinformatics Algorithms. Curr Bioinform. 2010;

34. Panja AS, Bandopadhyay B, Maiti S. Protein thermostability is owing to their preferences to non-polar smaller volume amino acids, variations in residual physico-chemical properties and more salt-bridges. PLoS One. 2015;

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

35. Yadav SK, Bhola A, Tiwari AK. Classification of enzyme functional classes and subclasses using support vector machine. In: 2015 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management, ABLAZE 2015. 2015.

36. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics. 2010;

37. Cao DS, Liang YZ, Yan J, Tan GS, Xu QS, Liu S. PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. J Chem Inf Model. 2013;

38. Shen HB, Chou KC. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem. 2008;373(2):386–8.

39. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005;

40. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci. 2007;

41. Pedregosa FABIANPEDREGOSA F, Alexandre Gramfort N, Michel V, Thirion BERTRANDTHIRION B, Grisel O, Blondel M, et al. Scikitlearn: Machine Learning in Python Gaël Varoquaux. J Mach Learn Res. 2011;

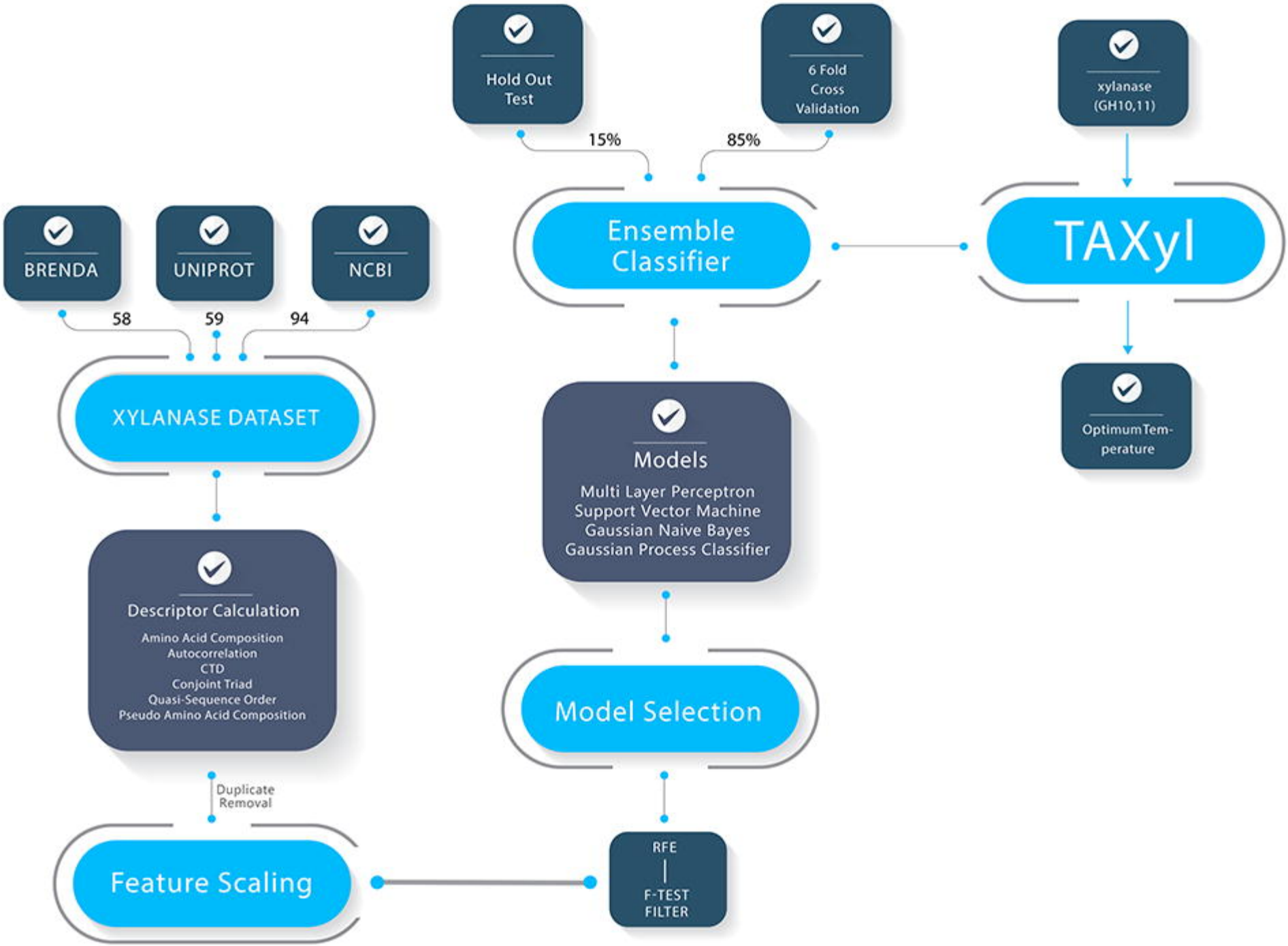
42. John GH, Langley P. Estimating Continuous Distributions in Bayesian Classifiers George. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. 1995.

43. Rasmussen CE. Gaussian Processes in Machine Learning. In 2004.

44. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q. Boosting: An ensemble learning tool for compound classification and QSAR modeling. J Chem Inf Model. 2005;

bioRxiv preprint doi: https://doi.org/10.1101/826040; this version posted November 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

45. Powers DMW. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. J Mach Learn Technol. 2011;


Hold Out Test
6 Fold Cross Validation
xylanase (GH10,11)
15%
85%
Ensemble Classifier
TAXyl
BRENDA
UNIPROT
NCBI
58
59
94
XYLANASE DATASET
Models
Multi Layer Perceptron
Support Vector Machine
Gaussian Naive Bayes
Gaussian Process Classifier
OptimumTem-perature
Descriptor Calculation
Amino Acid Composition
Autocorrelation
CTD
Conjoint Triad
Quasi-Sequence Order
Pseudo Amino Acid Composition
Model Selection
Duplicate Removal
Feature Scaling
RFE
F-TEST FILTER