





## Article

# GPpred: A Novel Sequence-Based Tool for Predicting Glutamic Proteases Using Optimized Hybrid Encodings

Ahmad Firoz <sup>1,2,†</sup>, Adeel Malik <sup>3,†</sup> , Nitin Mahajan <sup>4</sup> , Hani Mohammed Ali <sup>1,2</sup> , Majid Rasool Kamli <sup>1,5,\*</sup>   
and Chang-Bae Kim <sup>6,\*</sup>

<sup>1</sup> Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia; aakram@kau.edu.sa (A.F.); hmohammedali@kau.edu.sa (H.M.A.)

<sup>2</sup> Princess Dr. Najla Bint Saud Al-Saud Center for Excellence Research in Biotechnology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

<sup>3</sup> Institute of Intelligence Informatics Technology, Sangmyung University, Seoul 03016, Republic of Korea; adeel@procarb.org

<sup>4</sup> Wugen, St. Louis, MO 63110, USA; dr.nitin20@yahoo.com

<sup>5</sup> Center of Excellence in Bionanoscience Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia

<sup>6</sup> Department of Biotechnology, Sangmyung University, Seoul 03016, Republic of Korea

\* Correspondence: mkamli@kau.edu.sa (M.R.K.); evodevo@smu.ac.kr (C.-B.K.)

† These authors contributed equally to this work.

**Abstract:** Glutamic proteases (GPs) represent one of the seven peptidase families described in the MEROPS database of peptidases (also known as proteases, proteinases, and proteolytic enzymes). Currently, the GP family is divided into six sub-families (G1–G6) distributed across three clans (GA, GB, and GC). A glutamic acid and another variable amino acid are the catalytic residues in this family. Members of the GP family are involved in a wide variety of biological functions. For example, they act as bacterial and plant pathogens, and are involved in cancer and celiac disease. These enzymes are considered potential drug targets given their crucial roles in numerous biological processes. Characterizing GPs provides insights into their structure–function relationships, enabling the design of specific inhibitors or modulators. Such advancements directly contribute to drug discovery by identifying novel therapeutic targets and guiding the development of potent and selective drugs for various diseases, including cancers and autoimmune disorders. To address the challenges associated with labor-intensive experimental methods, we developed GPpred, an innovative support vector machine (SVM)-based predictor to identify GPs from their primary sequences. The workflow involves systematically extracting six distinct feature sets from primary sequences, and optimization using a recursive feature elimination (RFE) algorithm to identify the most informative hybrid encodings. These optimized encodings were then used to evaluate multiple machine learning classifiers, including K-Nearest Neighbors (KNNs), Random Forest (RF), Naïve Bayes (NB), and SVM. Among these, the SVM demonstrated a consistent performance, with an accuracy of 97% during the cross-validation and independent validation. Computational methods like GPpred accelerate this process by analyzing large datasets, predicting potential enzyme targets, and prioritizing candidates for experimental validation, thereby significantly reducing time and costs. GPpred will be a valuable tool for discovering GPs from large datasets, and facilitating drug discovery efforts by narrowing down viable therapeutic candidates.

**Keywords:** glutamic proteases; machine learning; support vector machine; recursive feature elimination



**Citation:** Firoz, A.; Malik, A.; Mahajan, N.; Ali, H.M.; Kamli, M.R.; Kim, C.-B. GPpred: A Novel Sequence-Based Tool for Predicting Glutamic Proteases Using Optimized Hybrid Encodings. *Catalysts* **2024**, *14*, 894. <https://doi.org/10.3390/catal14120894>

Academic Editors: Chia-Hung Kuo, Chwen-Jen Shieh and Yung-Chuan Liu

Received: 28 October 2024

Revised: 28 November 2024

Accepted: 28 November 2024

Published: 5 December 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Peptidases are a diverse family of hydrolases found in all living organisms. They are responsible for degrading peptide bonds in various biological processes [1]. For example, they break down malfunctioning proteins in cytosol, lysosome, plasma membranes, and extracellular space. They may also have regulatory roles in biological processes essential for

maintaining cell homeostasis [2]. Several peptidases have multiple domains; however, the proteolytic activity is limited to a single structural domain [3]. Therefore, only the sequence and structure of this individual domain is exploited to assign a proteolytic enzyme to a family and clan. Currently, the MEROPS [1] database encompasses seven peptidase families, which are named according to the main catalytic residues such as aspartic, cysteine, metallo, serine, glutamic, threonine, and asparagine. Among these proteolytic enzymes, glutamic proteases (GPs) were previously classified as the A4 family of aspartic endopeptidases, and were initially thought to be present only in fungi [4,5]. Subsequently, based on their three-dimensional (3D) structure and catalytic mechanism, GPs were identified as a novel protease family, the Eqolisins, a name deduced from the catalytic amino acids, glutamic acid (E) and glutamine (Q). Later [4], these GPs were also found in bacteria, archaea, and viruses [6,7].

Currently, the MEROPS database divides GPs into three main clans (GA, GB, and GC), and six sub-families (G1–G6). Among these, the sub-families G1 and G2 belong to the clans GA and GB, respectively, whereas G4 and G6 are grouped together within the clan GC. In contrast, G3 and G5 are not assigned to any specific clan, and, therefore, remain as unassigned categories. The G1 sub-family mainly includes enzymes that are found in plant pathogens. Scytalidoglutamic peptidase or eqolisin is the representative member of the G1 sub-family, and has a novel fold consisting of a  $\beta$ -sandwich composed of two seven-stranded anti-parallel  $\beta$ -sheets [4]. Eqolisin also has a unique catalytic dyad represented by amino acids glutamine 53 (Gln53) and glutamic acid 36 (Glu36). The representative member of the G2 sub-family is a pre-neck appendage protein from *Bacillus* phage  $\phi$ 29 with glutamic acid (Glu) and aspartic acid (Asp) as its catalytic residues [8]. This bacteriophage  $\phi$ 29 infects *Bacillus subtilis*, and has 12 “appendages” (gene product 12 or gp12) connected to its neck. These appendages are implicated in host cell recognition and entry. The G3 sub-family includes a novel GP known as “Pro2-Glu”, which was identified in strawberry mottle virus (SmoV) of the *Secoviridae* family. Pro2-Glu processes the RNA2 polyprotein at two different cleavage sites by exploiting two highly conserved glutamic acid residues which are essential for its activity. This enzyme shows structural similarities with fungal and bacterial GPs exhibiting a lectin fold [8]. Another member of the G3 sub-family is “neprosin”, first identified in the insectivorous tropical pitcher plants of the *Nepenthes* species [9]. Neprosins have a  $\beta$ -sandwich structure, contain two highly conserved catalytic Gln residues, and have propyl endopeptidase activity [10]. Neprosin is involved in the hydrolysis of proline-rich gliadin, a constituent of gluten that drives celiac disease. The G4 sub-family includes Tiki peptidase, which adopts the catalytic mechanism of the erythromycin esterase family. In addition to two highly conserved glutamic acids (glu85 and glu205), Tiki peptidases also have two highly conserved histidine residues (His58 and His331) with catalytic activity [11–14]. Tiki proteins inhibit the Wnt signaling pathway by acting as Wnt proteases, affecting Wnt solubility by its amino-terminal cleavage. Ras-converting enzyme (Rce1) and Microcystinase (MlrA) are two representative GPs from the G5 sub-family. The 3D structure of Rce1 comprises eight transmembrane  $\alpha$ -helices and two peripheral membrane  $\alpha$ -helices. The conserved residues Glu140, Glu141, His173, His227, and Asn231 are necessary for its catalytic activity. Rce1 plays an essential role in cancer and many infectious diseases. Therefore, it is a potential therapeutic target [15]. In contrast, MlrA is the first enzyme involved in the biodegradation of microcystins (MCs), one of the most abundant toxins associated with freshwater algal blooms [16]. The amino acid residues E172, H205, H260, and N264 are responsible for the catalytic activity of MlrA [17]. Finally, the representative member of the G6 sub-family of GPs is Ras/Rap1 site-specific endopeptidase (RRSP), a toxin protein found in many pathogenic bacteria such as *Vibrio vulnificus*, *Aeromonas hydrophila*, and *Photobacterium damela* [18]. Structural analysis of RRSP protein indicated that the C2 catalytic domain adopts a typical  $\alpha+\beta$  TIKI fold with the highly conserved catalytic residues (His3902, His4030, Glu3900, and Glu3930) [19].

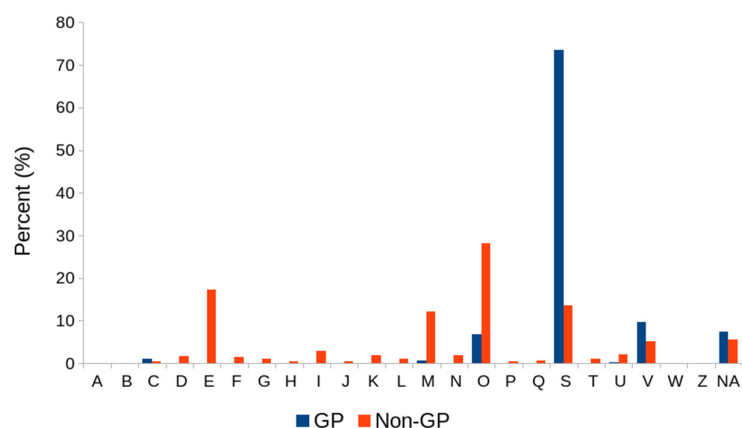
One common feature of the GP family of proteases is that their catalytic residue includes a glutamic acid and an additional variable amino acid [20]. GPs are involved in

various biological functions, such as acting as plant and bacterial pathogens and playing essential roles in several diseases, including celiac disease and cancer. This suggests that GPs are potential drug targets. With the rapid increase in potential GP sequences in public databases, a significant challenge lies in their annotation. Experimental methods for identifying and classifying GPs are time-consuming and costly, making computational approaches a valuable alternative for accurately identifying GP enzymes based on their primary sequences. Currently, available methods like BLAST [21,22] rely on sequence similarity, which limits their effectiveness in cases where the target sequence shares similarities with known GP sequences. Consequently, these methods struggle to detect novel GP enzymes. In contrast, machine learning (ML) approaches offer promising solutions for developing predictive models to classify GPs efficiently. In this study, we develop a novel predictor known as GPpred, which utilizes an SVM classifier and optimal hybrid encodings to differentiate between GPs and non-GPs. GPpred performed exceptionally well on both cross-validation and independent datasets. These insights could be beneficial to researchers investigating this unique category of enzymes, and aid in discovering potential therapeutic targets.

## 2. Results

### 2.1. Overview of the Dataset

A total of 7448 sequences, comprising 3724 GP (positive) and an equal number of randomly chosen non-GP (negative) proteins obtained from the MEROPS database, make up the final non-redundant dataset. We employed EggNOG mapper for their functional annotation and to compare various cluster of orthologous groups (COG) categories between positive and negative datasets. A distinguishable variation in the COG profile of GP and non-GP proteins was observed (Figure 1).



**Figure 1.** Distribution of different COG categories in both non-GP and GP datasets.

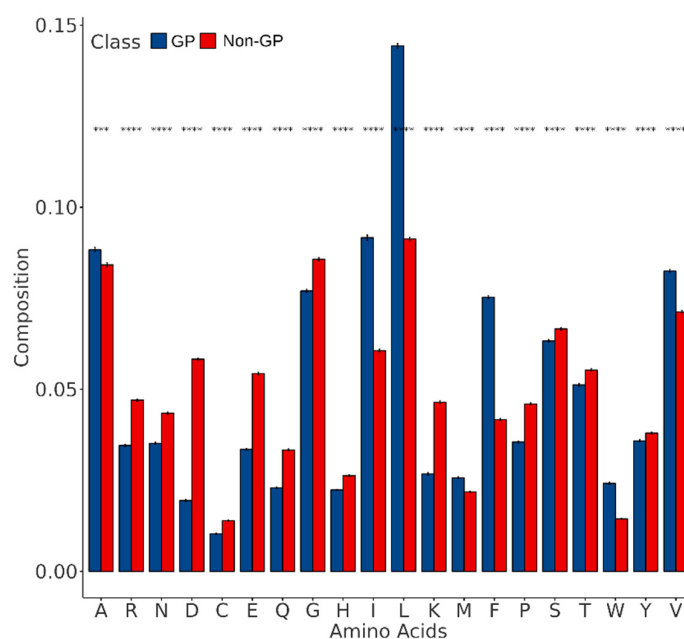
COG categories on the x-axis stand for energy generation and conversion (C); chromatin structure and dynamics (B); cell cycle control, cell division, and chromosomal partitioning (D); RNA processing and modification (A); enzyme transport and metabolism (H); carbohydrate transport and metabolism (G); amino acid transport and metabolism (E); nucleotide transport and metabolism (F); lipid transport and metabolism (I); protein turnover, chaperones, post-translational modification, and cell wall/membrane/envelope biogenesis (O); translation, ribosomal structure, and biogenesis (J); transcription (K); replication, recombination, and repair (L); cell motility (N); function unknown (S); signal transduction mechanisms (T); extracellular structures (W); defence mechanisms (V); intracellular trafficking, secretion, and vesicular transport (U); inorganic ion transport and metabolism (P); secondary metabolite biosynthesis, transport, and catabolism (Q); and not applicable (NA).

Approximately 73% of the GP sequences were classified as having unknown functions (S). Additionally, no hits were found in the EggNOG database for about 7.44% of the GP sequences. Among the annotated COG category defense mechanisms (V) and chaperone

functions (O), post-translational modification and protein turnover were the most over-represented categories, comprising 10% and 7%, respectively. In contrast, COG category O, at 28.2%, was the most enriched category in the negative dataset, followed by amino acid metabolism and transport (E). Interestingly, only 13.5% of the negative dataset belonged to category (S), about 60% less than the positive dataset. Another top COG category in the negative dataset was cell wall/membrane/envelop biogenesis (M).

## 2.2. Amino Acid Composition Comparison Between Glutamic and Non-Glutamic Proteases

We compared the AACs of both (i.e., positive and negative) datasets to observe any compositional variations between the GP and non-GP enzymes. We observe that 7/20 amino acids in the GP proteins are distinctly dominant (Wilcoxon's test;  $p < 0.05$ ) (Figure 2).



**Figure 2.** Amino acid composition (AAC) difference between GP and non-GP sequences. Asterisks above the bars indicate the  $p$ -value (\*\* =  $p < 0.01$ ; and \*\*\*\* =  $p < 0.0001$ ).

Specifically, the hydrophobic residues such as alanine (A), valine (V), isoleucine (I), leucine (L), and methionine (M) were over-represented in the GPs. Additionally, two residues with aromatic groups, such as tryptophan (W) and phenylalanine (F), were also enriched in the GP enzymes. In contrast, the non-GP proteins exhibited the dominance of three positively charged amino acids, arginine (R) histidine (H), and lysine (K), and two negatively charged amino acids, aspartic acid (D) and glutamic acid (E). Moreover, all the residues with polar uncharged side chains, such as proline (P), asparagine (N), cysteine (C), serine (S), glutamine (Q), and threonine (T), were dominant in the non-GPs. Hence, the lowest frequency of positively charged amino acids and residues with polar, uncharged side chains in the GP enzymes may be the most crucial property for classification. These distinctive compositional variations between the GPs and non-GPs suggest that our model could use differences in the amino acids as a reasonable approach to classify GP from non-GP sequences.

## 2.3. Assessment of ML Classifiers Utilizing the Original Feature Sets

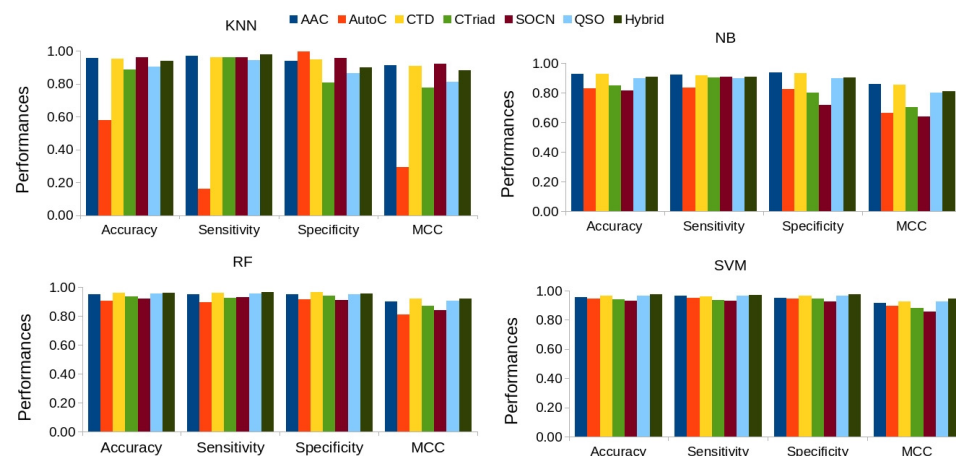
We assessed the performance of a disparate set of features and ML algorithms on separating GP from non-GP proteins by employing four different classifiers (KNN, SVM, NB, and RF) and seven features exhibiting a protein sequence's fundamental physicochemical and compositional properties. We generated 28 models (4 ML classifiers  $\times$  7 feature) and compared their performances by employing a 10-fold CV strategy. The top three models

exploited an SVM classifier and used hybrid and QSO- and CTD-based features (Table 1). Among these three models, the SVM classifier using hybrid features achieved an ACC of 97%, and its MCC was 0.947. The other two SVM-based classifiers exhibited an ACC of 96% with an MCC of 0.92. The ACC and MCC of the remaining 25 models were between 58 and 96% and 0.291 and 0.923, respectively. The model with the KNN- and AutoC-based encodings performed the poorest.

**Table 1.** Performance comparison of 28 ML models using the original feature sets on training data.

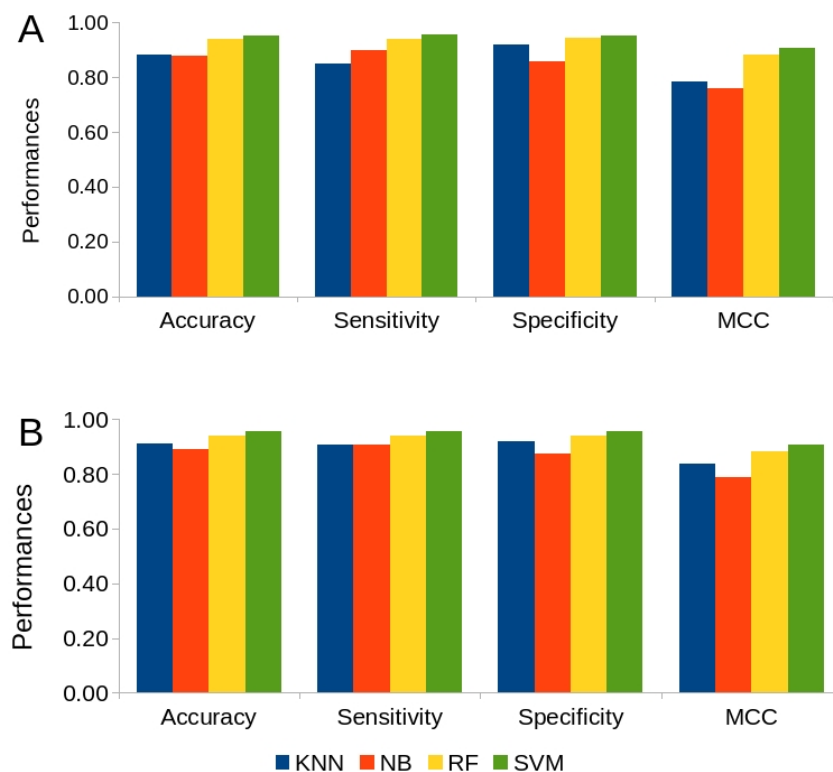
Method	Features	Accuracy	Sensitivity	Specificity	MCC
SVM	Hybrid	0.973	0.970	0.977	0.947
SVM	QSO	0.964	0.964	0.964	0.928
SVM	CTD	0.964	0.960	0.967	0.927
RF	CTD	0.961	0.958	0.965	0.923
RF	Hybrid	0.961	0.964	0.957	0.922
KNN	QSO	0.960	0.962	0.958	0.921
SVM	AAC	0.958	0.965	0.951	0.916
KNN	AAC	0.956	0.970	0.941	0.911
KNN	CTD	0.954	0.959	0.949	0.908
RF	QSO	0.954	0.957	0.951	0.908
RF	AAC	0.950	0.951	0.949	0.900
SVM	AutoC	0.948	0.951	0.944	0.895
SVM	CTriad	0.942	0.937	0.946	0.883
KNN	Hybrid	0.939	0.977	0.901	0.881
RF	CTriad	0.934	0.926	0.942	0.868
NB	AAC	0.930	0.923	0.936	0.859
SVM	SOCN	0.929	0.933	0.926	0.858
NB	CTD	0.927	0.920	0.933	0.853
RF	SOCN	0.920	0.929	0.911	0.840
KNN	SOCN	0.905	0.945	0.866	0.813
NB	Hybrid	0.906	0.910	0.902	0.812
RF	AutoC	0.906	0.895	0.918	0.812
NB	QSO	0.900	0.900	0.900	0.799
KNN	CTriad	0.884	0.960	0.808	0.777
NB	CTriad	0.851	0.903	0.799	0.705
NB	AutoC	0.831	0.835	0.827	0.663
NB	SOCN	0.813	0.907	0.720	0.638
KNN	AutoC	0.580	0.162	0.998	0.291

We then evaluated each ML classifier's performance over seven feature encodings to determine each classifier's effectiveness in predicting the GP sequences (Figure 3).



**Figure 3.** Comparing the performance of each classifier on seven original feature encodings during cross-validation.

The results demonstrate that in both ACC and MCC, the SVM and RF perform better than the other two classifiers (Figure 4A). In particular, the SVM-based classifier exhibited a slightly better average MCC than that of RF. Furthermore, the average ACC of these two classifiers is around 6–7% higher than that of the KNN- and NB-based classifiers. Comparatively speaking, the average MCC of the SVM and RF classifiers is between 10 and 14% better than that of the NB-based and KNN classifiers.



**Figure 4.** Overall performance of all ML classifiers irrespective of the feature encodings on training data using (A) control features and (B) optimal features.

Furthermore, regardless of the classifier employed, we sought to identify the top-performing feature descriptors for GP sequence classification. Our analysis shows that the CTD-based features outperformed the others, with an of 0.951. In contrast, the of the AAC, QSO, and hybrid features was approximately 94%, and their respective MCC scores were 0.903, 0.896, 0.889, and 0.891, respectively. The remaining encodings' performances ranged from 81 to 90%.

#### 2.4. Assessment of ML Classifiers After Feature Selection

In anticipation of any potential enhancements in the performances of the four ML models, we used the RFE method on each feature encoding. The findings show that the RFE technique substantially reduces the feature dimension size for nearly all of the encodings (Figure S1). No dimension reduction was observed for AAC (20D), suggesting that the composition of all 20 amino acids is essential for the prediction. In particular, RFE decreased the AutoC and the hybrid encodings' feature dimensions by 70% and 97%, respectively. A dimension size decrease between 17 and 60% was observed in the remaining features. We then employed these optimal features to train all four ML classifiers to assess and compare their performances.



By comparing the performance of the models based on RFE and control features, we observed that most of the models achieved similar performances. Remarkably, an and of 0.971 and 0.942 were shown by the top two models that used RFE-based optimum features (Table 2). Both these models were based on optimal hybrid encodings and exploited the RF and SVM classifiers, respectively. The next two top-performing models were based on SVM using optimal CTD and QSO encodings, and achieved an ACC of 96%.

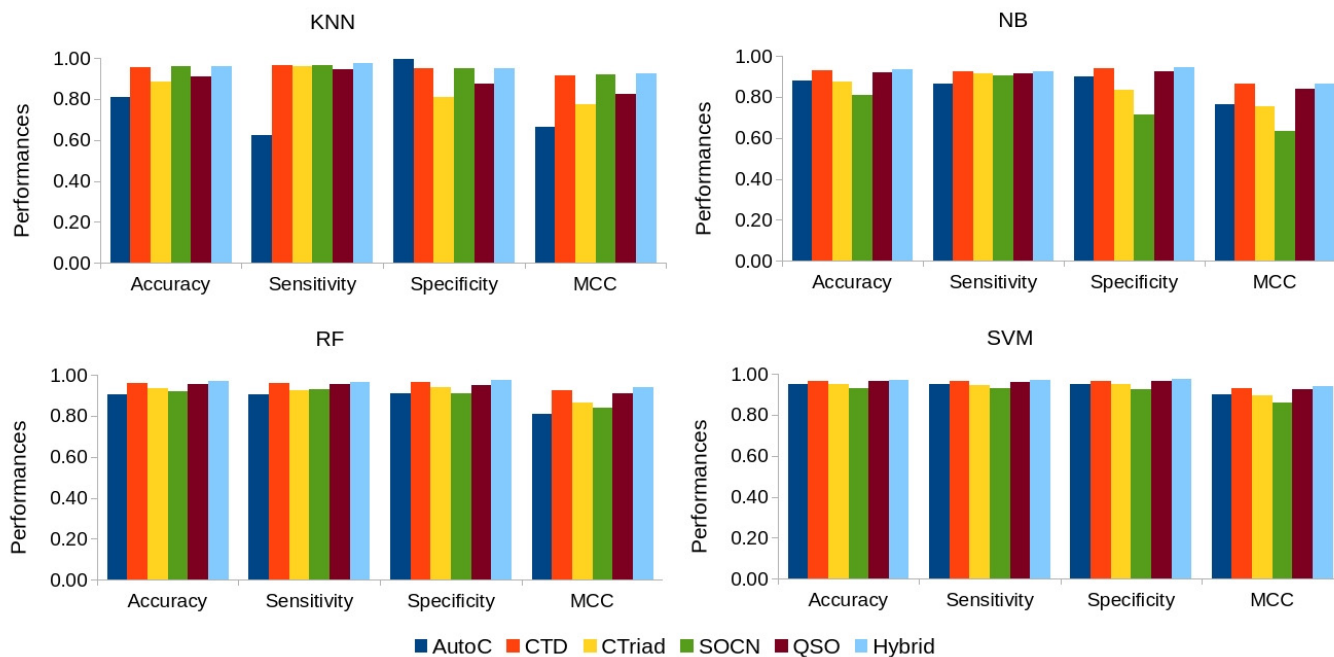
**Table 2.** Performance comparison of various ML models using RFE-based optimal feature sets on training data.

Method	Features	Accuracy	Sensitivity	Specificity	MCC
RF	Hybrid	0.971	0.967	0.975	0.942
SVM	Hybrid	0.971	0.968	0.974	0.942
SVM	CTD	0.966	0.966	0.965	0.931
SVM	QSO	0.963	0.962	0.963	0.926
RF	CTD	0.962	0.960	0.963	0.923
KNN	Hybrid	0.961	0.975	0.948	0.923
KNN	QSO	0.959	0.967	0.951	0.918
KNN	CTD	0.956	0.963	0.950	0.913
RF	QSO	0.954	0.957	0.951	0.909
SVM	AutoC	0.951	0.951	0.950	0.901
SVM	CTriad	0.948	0.947	0.948	0.895
NB	Hybrid	0.933	0.923	0.944	0.867
RF	CTriad	0.933	0.925	0.942	0.867
NB	CTD	0.931	0.925	0.937	0.862
SVM	SOCN	0.929	0.931	0.927	0.857
NB	QSO	0.920	0.914	0.926	0.840
RF	SOCN	0.918	0.928	0.909	0.837
KNN	SOCN	0.910	0.945	0.876	0.823
RF	AutoC	0.906	0.903	0.909	0.812
KNN	CTriad	0.884	0.960	0.807	0.776
NB	AutoC	0.882	0.864	0.900	0.764
NB	CTriad	0.875	0.914	0.836	0.752
KNN	AutoC	0.808	0.623	0.993	0.662
NB	SOCN	0.810	0.906	0.714	0.632

It should be noted that three of the five top models were based on the SVM classifier, while the remaining two utilized RF. However, no significant difference in the top-performing models' performance was observed when models based on original and dimension-reduced features were compared. Although RFE did improve the performance of some models, such as the KNN-based model using optimal AutoC encodings, by about 22% (Tables 1 and 2), this improvement was still poorer than that in the top-performing models.

Subsequently, seven optimal feature encodings were evaluated for each ML classifier, and their average performance was determined (Figure 5). We observed that the SVM classifier using RFE-based optimal features emerged as the best-performing classifier. The ACC and MCC achieved by this model were 0.954 and 0.909, respectively (Figure 4B). The performance achieved by the other three classifiers ranged between 89.2% (NB) and 94.1% (RF). Similarly, their MCC scores ranged between 0.786 and 0.882.

Again, focusing on individual feature encodings, regardless of the classifier used, the models using optimal hybrid encodings based on RFE performed slightly better than the average ACC of 95.9% for all the other models, and the corresponding MCC was 0.919 (Table S2). This represents a minor improvement of about 2% compared to the average performance of the models using original hybrid encodings. A significant jump of 8% was observed when the optimized AutoC features were used, and its MCC exhibited an increase of 12%.



**Figure 5.** Comparing the performance of each classifier on six optimal feature encodings during cross-validation. Since no dimension reduction was observed for AAC, it was excluded from this analysis.

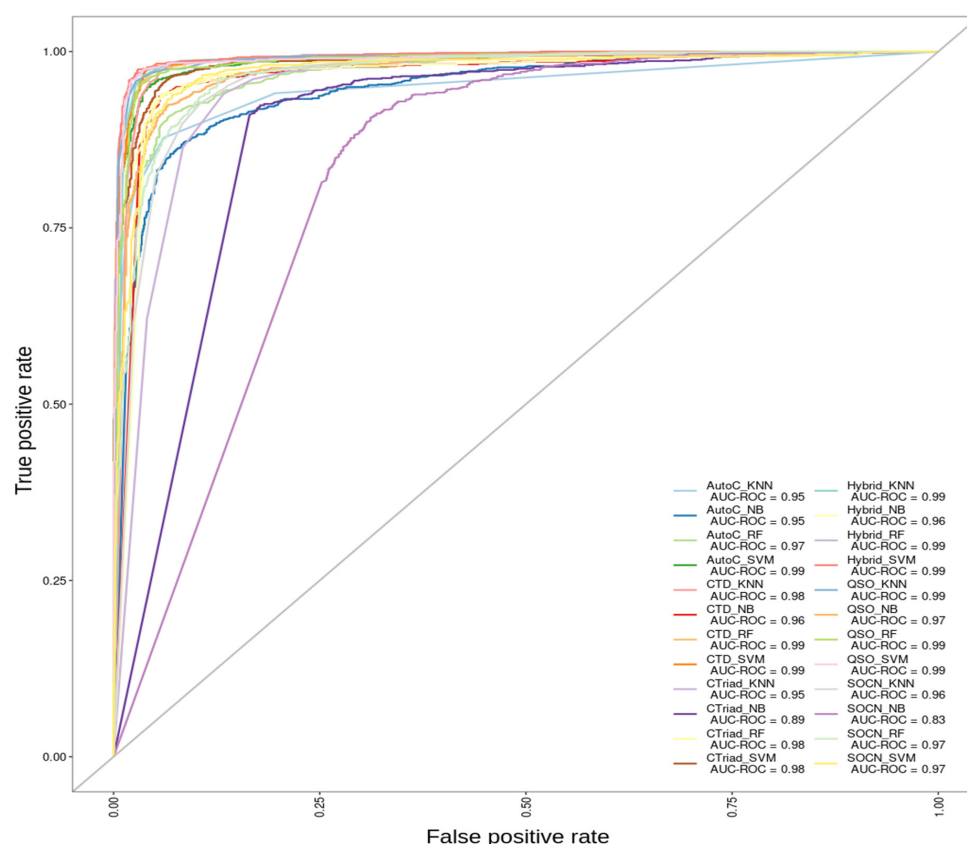
### 2.5. Assessment of ML Classifiers Using Optimal Feature Encodings on Independent Validation Datasets

We then used an independent validation test dataset (VS1) to evaluate the performance of all of the ML models. We considered the consistent performance of both independent and cross-validation, especially in terms of the ACC, rather than concentrating solely on the performance of VS1. When the top two models are considered, the independent validation results were consistent with the performance of the training datasets (Tables 2 and 3). The two top-ranking models in the training and validation datasets exhibited similar performances. For example, the highest-ranking models in the training and validation datasets were based on the SVM and RF classifiers and exploited optimal hybrid encodings. Specifically, both of these models achieved an ACC of 97% and an MCC of 0.942 during cross-validation. However, during independent validation, the SVM-based model exhibited a slightly better performance of 97.2% than the 97% shown by the RF-based model. The MCC scores achieved by these models were 0.944 and 0.940, respectively. To visually assess the performance of various encodings, an ROC curve was generated by plotting the true positive rate (TPR) against the false positive rate (FPR) (Figure 6). A higher AUC score in such plots reflects a better classifier performance. The figure shows that nine classifiers using optimal features achieved the highest AUC of 0.99, with four of these based on the SVM classifier, including the consistently well-performing SVM model using optimal hybrid encodings. These results highlight the potential of the SVM-based model with hybrid features to deliver a robust performance. Therefore, the SVM-based classifier was selected as the final model due to its consistent performance during training and independent validation.



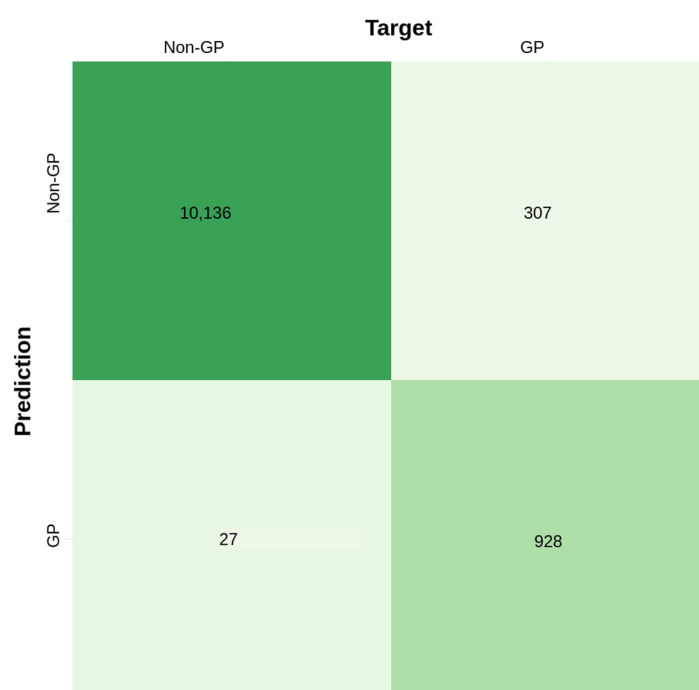
**Table 3.** Performance comparison of various ML models using RFE-based optimal feature sets on VS1.

Method	Features	Accuracy	Sensitivity	Specificity	MCC
SVM	Hybrid	0.972	0.972	0.971	0.944
RF	Hybrid	0.970	0.970	0.970	0.940
SVM	QSO	0.969	0.973	0.965	0.938
SVM	CTD	0.966	0.970	0.962	0.933
RF	CTD	0.962	0.963	0.960	0.923
KNN	Hybrid	0.961	0.976	0.946	0.923
KNN	QSO	0.961	0.970	0.952	0.922
KNN	CTD	0.958	0.972	0.944	0.916
RF	QSO	0.955	0.970	0.940	0.910
SVM	AutoC	0.951	0.961	0.942	0.903
SVM	CTriad	0.945	0.945	0.945	0.891
NB	Hybrid	0.939	0.941	0.937	0.878
RF	CTriad	0.938	0.938	0.938	0.876
NB	CTD	0.938	0.940	0.936	0.876
SVM	SOCN	0.931	0.941	0.921	0.862
NB	QSO	0.923	0.934	0.913	0.847
RF	SOCN	0.915	0.934	0.897	0.831
RF	AutoC	0.914	0.923	0.905	0.828
KNN	SOCN	0.906	0.950	0.863	0.816
NB	AutoC	0.889	0.880	0.899	0.779
KNN	CTriad	0.880	0.972	0.787	0.773
NB	CTriad	0.859	0.933	0.784	0.725
KNN	AutoC	0.818	0.645	0.991	0.678
NB	SOCN	0.797	0.907	0.687	0.608



**Figure 6.** Comparison of binormal receiver operating characteristic (ROC) curves for various prediction models tested on the independent dataset VS1. Higher scores reflect superior performance of the corresponding model.

To evaluate the robustness of our method, we retrieved a diverse set of non-GP sequences from the UniProt [23] database. For this analysis, sequences ranging in length from 40 to 550 amino acids were selected (refer to Section 2.1). These sequences were combined with the 1117 positive sequences from the VS1 dataset, and CD-HIT was applied to the combined dataset to remove redundancy at a 50% sequence similarity threshold. The resulting independent validation dataset (VS2) consisted of 11,398 sequences, including 955 GP and 10,443 non-GP proteins (Table S1). Testing the VS2 dataset with GPpred demonstrated a high ACC of 0.97, correctly predicting 928 out of 955 GP sequences (Figure 7). However, 307 out of the 10,443 non-GP sequences were misclassified as GP enzymes. A decrease in the MCC to 0.84 was observed, which is expected given the large number of negative sequences in the VS2 dataset. These results suggest that GPpred is a reliable and robust predictor that can identify GP sequences, as evidenced by its consistently high accuracy (97%) and ability to correctly classify most GP sequences (928 out of 955) in a significantly imbalanced dataset.



**Figure 7.** Confusion matrix displaying the prediction results on the additional independent dataset VS2. The matrix illustrates the distribution of outcomes for each class (GP and non-GP).

### 2.6. Availability as a Standalone Program

At present, no standalone application or web server for GP protein prediction utilizing sequence-derived optimum features is available. This constraint limits the practical application of such approaches for additional experimental characterization and annotation. To fill this gap between the sequencing and functional annotation of possible GPs, we have developed a free standalone program that is simple to use and incorporates our approach: <https://procarb.org/GPpred/> (accessed on 2 December 2024).

## 3. Discussion

Glutamic proteases (GPs) are proteolytic enzymes characterized by a glutamic acid residue in their active site. First identified in 2004, they were described as the sixth type of catalytic protease [4]. This family is characterized by catalytic residues that include a glutamic acid and a variable amino acid, and were initially found in pathogenic fungi that affect plants and humans [24]. They display various biological functions, such as roles in plant and bacterial pathogens and associations with celiac disease and cancer, indicating their potential as drug targets [20]. This makes their identification and characterization

indispensable. Thus, this study introduces an SVM-based predictor designed to predict GPs utilizing sequence-based optimal hybrid feature encodings.

Over the past few decades, significant progress has been achieved by ML-based techniques in learning from complex data and then predicting using unknown data [25]. The application of ML in biological sciences has surged, addressing various difficulties like predicting the structure of proteins [26], protein classification [27,28], peptide therapeutics [29,30], binding site prediction [31,32], image classification [33], and drug discovery [34]. Recognizing the absence of ML-based models for GP prediction, using sequence-derived optimum features to evaluate four widely used ML classifiers, GPpred was developed using sequence-based optimal encodings. Recently, similar approaches have been constructed for other proteolytic enzymes, including sortases, asparagine peptide lyase, and C10 family proteases [35–37].

GPpred was systematically created by evaluating the effectiveness of different classifiers on a non-redundant, balanced dataset, and finally, the model that showed the best consistency was selected through independent validation and training. Recognizing that high-dimensional features can include non-redundant or superfluous elements impacting classifier performance, we used RFE to identify the most suitable optimal feature set and ML classifier. RFE is a feature selection technique aimed at determining the optimal subset of features based on the trained model and classification accuracy [38]. In traditional RFE, features are removed one by one, starting with the least important feature that reduces the classification accuracy when eliminated after constructing the classification model. While GPpred demonstrated an excellent performance, there is room for further improvement. For instance, future enhancements could involve developing models with larger datasets, exploring other feature encodings, and creating ensemble-based models [39]. A limitation of this work is that GPpred cannot identify specific GP sub-families (e.g., G1, G2, or G3). The accurate prediction of these sub-families needs a significant enough number of sequences from each member. This is possible only when more data becomes available in public databases. Nevertheless, GPpred is the only publicly accessible predictor for detecting GP proteins.

## 4. Methods

### 4.1. Dataset Construction

MEROPS (v.12.4) peptidase database [1] was used to extract the protein sequences of positive (GP) and negative (non-GP) datasets. GP sub-families G1, G2, G3, G4, G5, and G6 represent the positive dataset, whereas the other proteolytic enzyme families were considered the negative dataset. Since most of the positive dataset enzymes showed sequences of length between 40 and 550 amino acids, we selected a negative dataset exhibiting a similar sequence length range. Both the positive and negative datasets were combined, and we subsequently applied CD-HIT v4.8.1 [40] with a 60% sequence identity cut-off to remove redundant sequences. With around 146,000 negative and 3724 positive sequences, this redundancy-reduced sample was highly unbalanced. Consequently, we chose 3724 negative sequences at random from the list of non-GP proteins to create a balanced dataset. Subsequently, we used the “createDataPartition” function from the caret [41] package to divide the final dataset into two sets: one for independent validation (VS1 = 2234 sequences) and one for training (5214 sequences). This ensures class stratification during the split, preserving the proportion of each class in both the training and validation sets (Table S1).

### 4.2. Feature Encodings

Using the “protr” [42] package, sequences of different lengths were represented as fixed-length feature vectors. We used six distinct features and their hybrid (that includes all six encodings) to build the machine learning models. These comprise the following: amino acid composition (AAC); autocorrelation (AutoC); composition (C), transition (T), and distribution (D) (CTD); conjoint triad (CTriad); quasi-sequence order (QSO); and sequence order coupling number (SOCN). These encodings, frequently employed in protein science,

show a sequence's primary physical, chemical, and compositional characteristics [27,37,43–47]. Each of these encodings are described in the following.

#### 4.2.1. Amino Acid Composition (AAC)

For AAC, the frequency of each of the 20 naturally occurring amino acids was computed for each leishmanolysin-like and non-leishmanolysin protein sequence. The AAC is represented as a fixed-length 20-dimensional feature vector, defined as

$$AAC(i) = \frac{R_i}{L}$$

where  $R_i$  denotes the count of amino acids of type  $i$  and  $L$  represents the length of the protein sequence.

#### 4.2.2. Autocorrelation (AutoC)

The AutoC feature can extract information on the physicochemical properties of a protein sequence, potentially enhancing the model performance [47]. AutoC descriptors are generally categorized into three main types:

(i) Moran AutoC encodings, represented as

$$I(d) = \frac{\frac{1}{N-d} \sum_{i=1}^{N-d} (P_i - \bar{P})(P_{i+d} - \bar{P})}{\frac{1}{N} \sum_{i=1}^N (P_i - \bar{P})^2} d = 1, 2, \dots, 30$$

where  $d$  is the autocorrelation lag,  $P_i$  and  $P_{i+d}$  denote the amino acid properties at positions  $i$  and  $i + d$ , respectively, and  $nlag$  is the maximum lag value.

(ii) Moreau–Broto AutoC descriptors, represented as

$$I(d) = \frac{\frac{1}{N-d} \sum_{i=1}^{N-d} (P_i - \bar{P})(P_{i+d} - \bar{P})}{\frac{1}{N} \sum_{i=1}^N (P_i - \bar{P})^2} d = 1, 2, \dots, 30$$

(iii) Geary AutoC descriptors, represented as

$$C(d) = \frac{\frac{1}{2(N-d)} \sum_{i=1}^{N-d} (P_i - P_{i+d})^2}{\frac{1}{N} \sum_{i=1}^N (P_i - \bar{P})^2} d = 1, 2, \dots, 30$$

where  $\bar{P}$  represents the average value of the property  $P$ , denoted as  $\bar{P} = \sum_{i=1}^N P_i / N$

#### 4.2.3. Composition (C), Transition (T), and Distribution (D) (CTD)

Since their initial application in protein folding classification [45,48], CTD descriptors have been widely used in developing various bioinformatics tools [27,35,37,46]. In CTD descriptors, the 20 naturally occurring amino acids are classified into three groups—polar, neutral, and hydrophobic—based on seven distinct physicochemical properties: solvent accessibility, polarizability, polarity, hydrophobicity, charge, normalized van der Waals volume, and secondary structure.

#### 4.2.4. Conjoint Triad (CTriad)

CTriad encodings were first utilized to predict protein–protein interactions [49]. The CTriad descriptor is derived by evaluating the properties of a single amino acid and its neighboring residues, treating each group of three consecutive residues as a triad. These triads are categorized by the classes of the amino acids they contain—specifically, triads are distinguished when all three residues fall into the same amino acid class. In CTriad, a protein sequence is represented in a vector space that encapsulates amino acid features. This vector space is simplified by clustering the 20 amino acids based on dipole moments

and side chain volumes, resulting in a 343-dimensional feature vector for any given protein sequence. Mathematically, CTriad is defined as follows:

$$d_i = \frac{f_i - \min\{f_1, f_2, \dots, f_{343}\}}{\max\{f_1, f_2, \dots, f_{343}\}}, i = 1, 2, \dots, 343$$

where  $f_i$  ( $i = 1, 2, \dots, 343$ ) is the frequency of occurrence of each triad.

#### 4.2.5. Quasi-Sequence Order (QSO)

Given the wide variety of sequence order patterns in biological sequences, it is impractical to directly incorporate this information into an ML classifier [43]. To overcome this challenge, QSO encoding is utilized to indirectly integrate sequence order information [50]. QSO encodings are derived using the Grantham distance matrix and the Schneider–Wrede distance matrix for each pair among the 20 naturally occurring amino acids [42,51]. Grantham matrix provides information on chemical distances, while the Schneider–Wrede matrix captures physicochemical properties, including polarity, hydrophobicity, and hydrophilicity [52].

#### 4.2.6. Sequence Order Coupling Number (SOCN)

The  $d$ -th rank sequence order coupling number is defined as follows:

$$\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2 \quad d = 1, 2, \dots, \text{maxlag}$$

where the maximum lag is  $d_{i,i+d}$ , and the sequence length must be at least the value of maxlag.

### 4.3. Feature Selection

Feature selection is an imperative part of an ML task due to its potential to enhance the model performance. Here, we use recursive feature elimination (RFE) strategy, which attempts to identify the most optimal subgroup of descriptors based on the classification model and its accuracy [38] applied RFE on all seven feature encodings (including the hybrid features), using the RF function (rfFuncs) within the rfeControl function. Throughout the RFE process, we generated several subsets of training data of different dimension size (e.g., 10–1390 with a step size of 10), and evaluated all the models using a 10-fold CV approach. Accuracy was utilized to evaluate and determine the optimal feature subset.

### 4.4. Machine Learning Classifiers

A well-known R package (2021) [53] known as caret [41] is used to apply four popular machine learning classifiers: KNN, NB, RF, and SVM. Here is a brief overview of the four classifiers:

#### 4.4.1. K-Nearest Neighbor (KNN)

KNN is also commonly known as a distance-based model because it is among the most straightforward and one of the quickest machine learning classifiers [54]. In KNN, an attempt is made to find the  $k$ -nearest examples in a reference set [55]. The Euclidean distance is utilized in KNN to determine the distance between incidences.

#### 4.4.2. Naïve Bayes (NB)

By assuming that the predictors are independent, the Naive Bayes (NB) makes learning simple [56]. Based on the Bayes theorem, this approach assumes that a feature's presence in a class is unrelated to the presence of any other features. As a result, each feature contributes equally to the final result [57].

#### 4.4.3. Random Forest (RF)

Initially developed by [58], Random Forest (RF) is a popular learning method used for both regression and classification tasks. RF stands out among other classifiers for its ease of training, fast prediction, and interpretability [59]. This technique constructs multiple decision trees, each built from a subset of randomly selected descriptors from the full feature set [60]. The random selection reduces bias and reduces the correlation between the unpruned trees [61]. Combining the resampled set with a randomly selected feature vector, a decision tree is constructed [62,63] bagging technique produces a training feature set with reconfigured examples. A final prediction is determined by majority vote once all of the decision trees' predictions have been combined [64].

#### 4.4.4. Support Vector Machine (SVM)

To maximize the generalization performance, SVMs make use of statistical learning theory and the structural risk minimization principle [50]. SVM maximizes the distance between the two classes it creates from the training data and the hyperplanes [65]. By utilizing machine learning theory, SVM reduces overfitting and improves prediction accuracy, frequently yielding better results than other classifiers [66,67].

#### 4.5. Evaluation Metrics

The four widely used measures that provide an assessment of the binary classification quality are Matthews' correlation coefficient (MCC), accuracy (ACC), specificity (Sp), and sensitivity (Sn). These measurements are frequently used to assess machine learning models' performance. As a result, we evaluated our ML models' performance using these indicators on both the training and independent validation sets to assess their ability to generalize and their performance during development. Mathematically, these measures can be represented as follows:

$$\left\{ \begin{array}{l} Sn = \frac{TP}{TP+FN} \\ Sp = \frac{TN}{TN+FP} \\ ACC = \frac{TP+FN+TN+FP}{TP+FN+TN+FP} \\ MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{array} \right.$$

where true positive, true negative, false positive, and false negative are represented by *TP*, *TN*, *FP*, and *FN*.

## 5. Conclusions

Glutamic proteases are also known as acidic proteases and typically exhibit optimal activity at low pH levels. Initially believed to be primarily in filamentous fungi, GPs have since been discovered in bacteria and archaea. A glutamic protease originating from a plant virus, such as the strawberry mottle virus, has also been identified. Many GPs are considered promising drug targets because of their important roles in many diseases. For this reason, the precise identification and categorization of these GPs may provide insightful information for developing various therapeutic strategies. Herein, a new predictor called GPpred was developed, employing SVM and optimal hybrid encodings to distinguish between GPs and non-GPs. GPpred showed superior performance during cross-validation and independent validation. Currently, GPpred is the only tool for predicting GP sequences, serving as a valuable resource for identifying potential novel GPs. This could provide valuable insights for researchers studying this distinctive class of enzymes.

**Supplementary Materials:** The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/catal14120894/s1>: Table S1: Summary of the final datasets used to develop GPpred. Table S2: Performance comparison of each control and original feature sets for GP protein classification. Scores for each feature represent the average of four machine learning



classifiers. The standard deviation ranges between 0.007–0.369. Figure S1: Overview of dimension reduction by RFE based approach, and its comparison with the original dimension size.

**Author Contributions:** Methodology, A.F. and A.M.; Software, A.M.; Validation, N.M. and H.M.A.; Formal analysis, A.F.; Investigation, A.F. and N.M.; Resources, H.M.A. and C.-B.K.; Data curation, N.M.; Writing—review & editing, A.M., M.R.K. and C.-B.K.; Visualization, M.R.K.; Supervision, M.R.K. and C.-B.K.; Project administration, H.M.A.; Funding acquisition, A.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Institutional Fund Projects under Grant No. (IFPIP:431-130-1443). The authors gratefully acknowledge the technical and financial support provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

**Data Availability Statement:** All the data utilized in this study can be downloaded at <https://procarb.org/GPpred/>.

**Conflicts of Interest:** Nitin Mahajan is employed by Wugen, St. Louis, MO, USA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

## References

1. Rawlings, N.D.; Barrett, A.J.; Thomas, P.D.; Huang, X.; Bateman, A.; Finn, R.D. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **2018**, *46*, D624–D632. [[CrossRef](#)] [[PubMed](#)]
2. Kos, J. Peptidases: Role and Function in Health and Disease. *Int. J. Mol. Sci.* **2023**, *24*, 7823. [[CrossRef](#)] [[PubMed](#)]
3. Rawlings, N.D.; Bateman, A. How to use the MEROPS database and website to help understand peptidase specificity. *Protein Sci.* **2021**, *30*, 83–92. [[CrossRef](#)] [[PubMed](#)]
4. Fujinaga, M.; Cherney, M.M.; Oyama, H.; Oda, K.; James, M.N. The molecular structure and catalytic mechanism of a novel carboxyl peptidase from *Scytalidium lignicolum*. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3364–3369. [[CrossRef](#)] [[PubMed](#)]
5. Sims, A.H.; Dunn-Coleman, N.S.; Robson, G.D.; Oliver, S.G. Glutamic protease distribution is limited to filamentous fungi. *FEMS Microbiol. Lett.* **2004**, *239*, 95–101. [[CrossRef](#)]
6. Jensen, K.; Oestergaard, P.R.; Wilting, R.; Lassen, S.F. Identification and characterization of a bacterial glutamic peptidase. *BMC Biochem.* **2010**, *11*, 47. [[CrossRef](#)]
7. Mann Krin, S.; Chisholm, J.; Sanfaçon, H. *Strawberry Mottle Virus* (Family *Secoviridae*, Order *Picornavirales*) Encodes a Novel Glutamic Protease to Process the RNA2 Polyprotein at Two Cleavage Sites. *J. Virol.* **2019**, *93*, e01679-18. [[CrossRef](#)]
8. Xiang, Y.; Leiman, P.G.; Li, L.; Grimes, S.; Anderson, D.L.; Rossmann, M.G. Crystallographic Insights into the Autocatalytic Assembly Mechanism of a Bacteriophage Tail Spike. *Mol. Cell* **2009**, *34*, 375–386. [[CrossRef](#)]
9. Lee, L.; Zhang, Y.; Ozar, B.; Sensen, C.W.; Schriemer, D.C. Carnivorous Nutrition in Pitcher Plants (*Nepenthes* spp.) via an Unusual Complement of Endogenous Enzymes. *J. Proteome Res.* **2016**, *15*, 3108–3117. [[CrossRef](#)]
10. Ting, T.Y.; Baharin, A.; Ramzi, A.B.; Ng, C.-L.; Goh, H.-H. Neprosin belongs to a new family of glutamic peptidase based on in silico evidence. *Plant Physiol. Biochem.* **2022**, *183*, 23–35. [[CrossRef](#)]
11. Morar, M.; Pengelly, K.; Koteva, K.; Wright, G.D. Mechanism and Diversity of the Erythromycin Esterase Family of Enzymes. *Biochemistry* **2012**, *51*, 1740–1751. [[CrossRef](#)] [[PubMed](#)]
12. Zhang, X.; Abreu, J.G.; Yokota, C.; MacDonald, B.T.; Singh, S.; Coburn, K.L.; Cheong, S.M.; Zhang, M.M.; Ye, Q.Z.; Hang, H.C.; et al. Tiki1 Is Required for Head Formation via Wnt Cleavage-Oxidation and Inactivation. *Cell* **2012**, *149*, 1565–1577. [[CrossRef](#)] [[PubMed](#)]
13. Sanchez-Pulido, L.; Ponting, C.P. Tiki, at the head of a new superfamily of enzymes. *Bioinformatics* **2013**, *29*, 2371–2374. [[CrossRef](#)] [[PubMed](#)]
14. Zhang, X.; MacDonald, B.T.; Gao, H.; Shamashkin, M.; Coyle, A.J.; Martinez, R.V.; He, X. Characterization of Tiki, a New Family of Wnt-Specific Metalloproteases. *J. Biol. Chem.* **2016**, *291*, 2435–2443. [[CrossRef](#)] [[PubMed](#)]
15. Hampton, S.E.; Dore, T.M.; Schmidt, W.K. Rce1: Mechanism and inhibition. *Crit. Rev. Biochem. Mol. Biol.* **2018**, *53*, 157–174. [[CrossRef](#)]
16. Bouaïcha, N.; Miles, C.O.; Beach, D.G.; Labidi, Z.; Djabri, A.; Benayache, N.Y.; Nguyen-Quang, T. Structural Diversity, Characterization and Toxicology of Microcystins. *Toxins* **2019**, *11*, 714. [[CrossRef](#)]
17. Xu, Q.; Fan, J.; Yan, H.; Ahmad, S.; Zhao, Z.; Yin, C.; Liu, X.; Liu, Y.; Zhang, H. Structural basis of microcystinase activity for biodegrading microcystin-LR. *Chemosphere* **2019**, *236*, 124281. [[CrossRef](#)]
18. Antic, I.; Biancucci, M.; Zhu, Y.; Gius, D.R.; Satchell, K.J.F. Site-specific processing of Ras and Rap1 Switch I by a MARTX toxin effector domain. *Nat. Commun.* **2015**, *6*, 7396. [[CrossRef](#)]
19. Jang, S.Y.; Hwang, J.; Kim, B.S.; Lee, E.-Y.; Oh, B.-H.; Kim, M.H. Structural basis of inactivation of Ras and Rap1 small GTPases by Ras/Rap1-specific endopeptidase from the sepsis-causing pathogen *Vibrio vulnificus*. *J. Biol. Chem.* **2018**, *293*, 18110–18122. [[CrossRef](#)]

20. Oda, K.; Wlodawer, A. Overview of the Properties of Glutamic Peptidases That Are Present in Plant and Bacterial Pathogens and Play a Role in Celiac Disease and Cancer. *Biochemistry* **2023**, *62*, 672–694. [[CrossRef](#)]
21. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)] [[PubMed](#)]
22. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [[CrossRef](#)] [[PubMed](#)]
23. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2004**, *32*, 115–119. [[CrossRef](#)] [[PubMed](#)]
24. Oda, K. New families of carboxyl peptidases: Serine-carboxyl peptidases and glutamic peptidases. *J. Biochem.* **2012**, *151*, 13–25. [[CrossRef](#)]
25. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [[CrossRef](#)]
26. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [[CrossRef](#)]
27. Firoz, A.; Malik, A.; Ali, H.M.; Akhter, Y.; Manavalan, B.; Kim, C.B. PRR-HyPred: A two-layer hybrid framework to predict pattern recognition receptors and their families by employing sequence encoded optimal features. *Int. J. Biol. Macromol.* **2023**, *234*, 123622. [[CrossRef](#)]
28. Dao, F.-Y.; Liu, M.L.; Su, W.; Lv, H.; Zhang, Z.Y.; Lin, H.; Liu, L. AcrPred: A hybrid optimization with enumerated machine learning algorithm to predict Anti-CRISPR proteins. *Int. J. Biol. Macromol.* **2023**, *228*, 706–714. [[CrossRef](#)]
29. Kurata, H.; Tsukiyama, S.; Manavalan, B. iACVP: Markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model. *Brief. Bioinform.* **2022**, *23*, bbac265. [[CrossRef](#)]
30. Manavalan, B.; Patra, M.C. MLCPP 2.0: An Updated Cell-penetrating Peptides and Their Uptake Efficiency Predictor. *J. Mol. Biol.* **2022**, *434*, 167604. [[CrossRef](#)]
31. Firoz, A.; Malik, A.; Joplin, K.H.; Ahmad, Z.; Jha, V.; Ahmad, S. Residue propensities, discrimination and binding site prediction of adenine and guanine phosphates. *BMC Biochem.* **2011**, *12*, 20. [[CrossRef](#)] [[PubMed](#)]
32. Malik, A.; Ahmad, S. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. *BMC Struct. Biol.* **2007**, *7*, 1. [[CrossRef](#)] [[PubMed](#)]
33. Ullah, M.; Han, K.; Hadi, F.; Xu, J.; Song, J.; Yu, D.J. PSL-HDeep: Image-based prediction of protein subcellular location in human tissue using ensemble learning of handcrafted and deep learned features with two-layer feature selection. *Brief. Bioinform.* **2021**, *22*, bbab278. [[CrossRef](#)] [[PubMed](#)]
34. Deng, J.; Yang, Z.; Ojima, I.; Samaras, D.; Wang, F. Artificial intelligence in drug discovery: Applications and techniques. *Brief. Bioinform.* **2021**, *23*, bbab430. [[CrossRef](#)]
35. Malik, A.; Subramaniam, S.; Kim, C.B.; Manavalan, B. SortPred: The first machine learning based predictor to identify bacterial sortases and their classes using sequence-derived information. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 165–174. [[CrossRef](#)]
36. Malik, A.; Kamli, M.R.; Sabir, J.S.; Rather, I.A.; Phan, L.T.; Kim, C.-B.; Manavalan, B. APLpred: A machine learning-based tool for accurate prediction and characterization of asparagine peptide lyases using sequence-derived optimal features. *Methods* **2024**, *229*, 133–146. [[CrossRef](#)]
37. Malik, A.; Mahajan, N.; Dar, T.A.; Kim, C.-B. C10Pred: A First Machine Learning Based Tool to Predict C10 Family Cysteine Peptidases Using Sequence-Derived Features. *Int. J. Mol. Sci.* **2022**, *23*, 9518. [[CrossRef](#)]
38. Jeon, H.; Oh, S. Hybrid-Recursive Feature Elimination for Efficient Feature Selection. *Appl. Sci.* **2020**, *10*, 3211. [[CrossRef](#)]
39. Qiu, W.-R.; Xu, A.; Xu, Z.-C.; Zhang, C.-H.; Xiao, X. Identifying Acetylation Protein by Fusing Its PseAAC and Functional Domain Annotation. *Front. Bioeng. Biotechnol.* **2019**, *7*, 311. [[CrossRef](#)]
40. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [[CrossRef](#)]
41. Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [[CrossRef](#)]
42. Xiao, N.; Cao, D.S.; Zhu, M.F.; Xu, Q.S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **2015**, *31*, 1857–1859. [[CrossRef](#)] [[PubMed](#)]
43. Chou, K.-C. Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. *Biochem. Biophys. Res. Commun.* **2000**, *278*, 477–483. [[CrossRef](#)] [[PubMed](#)]
44. Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Bioinform.* **2001**, *43*, 246–255. [[CrossRef](#)]
45. Dubchak, I.; Muchnik, I.; Holbrook, S.R.; Kim, S.H. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 8700–8704. [[CrossRef](#)]
46. Malik, A.; Kamli, M.R.; Sabir, J.S.; Phan, L.T.; Kim, C.-B.; Manavalan, B. RDR100: A Robust Computational Method for Identification of Krüppel-like Factors. *Curr. Bioinform.* **2024**, *19*, 584–599. [[CrossRef](#)]
47. Chen, C.; Zhang, Q.; Ma, Q.; Yu, B. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemom. Intell. Lab. Syst.* **2019**, *191*, 54–64. [[CrossRef](#)]
48. Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S.H. Recognition of a protein fold in the context of the SCOP classification. *Proteins Struct. Funct. Bioinform.* **1999**, *35*, 401–407. [[CrossRef](#)]

49. Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 4337–4341. [[CrossRef](#)]
50. Akbar, S.; Hayat, M.; Tahir, M.; Chong, K.T. cACP-2LFS: Classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach. *IEEE Access* **2020**, *8*, 131939–131948. [[CrossRef](#)]
51. Ong, S.A.K.; Lin, H.H.; Chen, Y.Z.; Li, Z.R.; Cao, Z. Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinform.* **2007**, *8*, 300. [[CrossRef](#)] [[PubMed](#)]
52. van den Berg, B.A.; Reinders, M.J.; Roubos, J.A.; Ridder, D. de SPiCE: A Web-Based Tool for Sequence-Based Protein Classification and Exploration. *BMC Bioinform.* **2014**, *15*, 93. [[CrossRef](#)] [[PubMed](#)]
53. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
54. Ahmad, A.; Akbar, S.; Hayat, M.; Ali, F.; Khan, S.; Sohail, M. Identification of Antioxidant Proteins Using a Discriminative Intelligent Model of K-Spaced Amino Acid Pairs Based Descriptors Incorporating with Ensemble Feature Selection. *Biocybern. Biomed. Eng.* **2020**, *42*, 727–735. [[CrossRef](#)]
55. Shen, H.; Chou, K.C. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 288–292. [[CrossRef](#)]
56. Rish, I. *An Empirical Study of the Naive Bayes Classifier*; T.J. Watson Research Center: Yorktown Heights, NY, USA, 2001.
57. Abbas, Z.; Tayara, H.; Chong, K.T. Alzheimer’s disease prediction based on continuous feature representation using multi-omics data integration. *Chemom. Intell. Lab. Syst.* **2022**, *223*, 104536. [[CrossRef](#)]
58. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
59. Jo, T.; Cheng, J. Improving protein fold recognition by random forest. *BMC Bioinform.* **2014**, *15*, S14. [[CrossRef](#)]
60. Li, J.; Wu, J.; Chen, K. PFP-RFSM: Protein fold prediction by using random forests and sequence motifs. *J. Biomed. Sci. Eng.* **2013**, *6*, 10. [[CrossRef](#)]
61. Waris, M.; Ahmad, K.; Kabir, M.; Hayat, M. Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix. *Neurocomputing* **2016**, *199*, 154–162. [[CrossRef](#)]
62. Hayat, M.; Khan, A.; Yeasin, M. Prediction of membrane proteins using split amino acid and ensemble classification. *Amino Acids* **2012**, *42*, 2447–2460. [[CrossRef](#)]
63. Ma, X.; Guo, J.; Sun, X. DNABP: Identification of DNA-Binding Proteins Based on Feature Selection Using a Random Forest and Predicting Binding Residues. *PLoS ONE* **2016**, *11*, e0167345. [[CrossRef](#)] [[PubMed](#)]
64. Sabooh, M.F.; Iqbal, N.; Khan, M.; Khan, M.; Maqbool, H. Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou’s PseKNC. *J. Theor. Biol.* **2018**, *452*, 1–9. [[CrossRef](#)] [[PubMed](#)]
65. Ahmed, S.; Arif, M.; Kabir, M.; Khan, K.; Khan, Y.D. PredAoDP: Accurate identification of antioxidant proteins by fusing different descriptors based on evolutionary information with support vector machine. *Chemom. Intell. Lab. Syst.* **2022**, *228*, 104623. [[CrossRef](#)]
66. Akbar, S.; Hayat, M. iMethyl-STTNC: Identification of N6-methyladenosine sites by extending the idea of SAAC into Chou’s PseAAC to formulate RNA sequences. *J. Theor. Biol.* **2018**, *455*, 205–211. [[CrossRef](#)]
67. Ali, F.; Arif, M.; Khan, Z.U.; Kabir, M.; Ahmed, S.; Yu, D.-J. SDBP-Pred: Prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM. *Anal. Biochem.* **2020**, *589*, 113494. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.