International Journal of
Molecular Sciences

MDPI

Communication

Predicting Structural Susceptibility of Proteins to Proteolytic Processing

Evgenii V. Matveev [1,2,3], Vyacheslav V. Safronov [4], Gennady V. Ponomarev [1,2] and Marat D. Kazanov [1,2,3,5,*]

[1] Skolkovo Institute of Science and Technology, Moscow 121205, Russia
[2] A.A. Kharkevich Institute for Information Transmission Problems, Moscow 127051, Russia
[3] Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Moscow 117998, Russia
[4] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119991, Russia
[5] Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey
* Correspondence: mkazanov@gmail.com; Tel.: +90-216-483-90-00

Abstract: The importance of 3D protein structure in proteolytic processing is well known. However, despite the plethora of existing methods for predicting proteolytic sites, only a few of them utilize the structural features of potential substrates as predictors. Moreover, to our knowledge, there is currently no method available for predicting the structural susceptibility of protein regions to proteolysis. We developed such a method using data from CutDB, a database that contains experimentally verified proteolytic events. For prediction, we utilized structural features that have been shown to influence proteolysis in earlier studies, such as solvent accessibility, secondary structure, and temperature factor. Additionally, we introduced new structural features, including length of protruded loops and flexibility of protein termini. To maximize the prediction quality of the method, we carefully curated the training set, selected an appropriate machine learning method, and sampled negative examples to determine the optimal positive-to-negative class size ratio. We demonstrated that combining our method with models of protease primary specificity can outperform existing bioinformatics methods for the prediction of proteolytic sites. We also discussed the possibility of utilizing this method for bioinformatics prediction of other post-translational modifications.

Keywords: regulatory proteolysis; proteases; protease substrates; substrate identification

Citation: Matveev, E.V.; Safronov, V.V.; Ponomarev, G.V.; Kazanov, M.D. Predicting Structural Susceptibility of Proteins to Proteolytic Processing. *Int. J. Mol. Sci.* **2023**, *24*, 10761. https://doi.org/10.3390/ijms241310761

Academic Editor: Emilia Pedone

Received: 29 May 2023
Revised: 16 June 2023
Accepted: 26 June 2023
Published: 28 June 2023

1. Introduction

After synthesis, proteins within living cells undergo a wide range of chemical modifications collectively referred to as post-translational modifications (PTMs) [1]. To date, about a dozen types of PTMs are known, including phosphorylation, acetylation, glycosylation, ubiquitination, methylation, and others [2]. Unlike most types of post-translational modifications, which involve the addition of small chemical groups, proteolysis is an irreversible post-translational modification that catalyzes the hydrolysis of the peptide bond [3]. To perform cleavage, a protease needs to bind to the polypeptide chain in the vicinity of the cleaved peptide bond [4]. The ability of a protease to bind to a polypeptide chain in a specific amino acid context is known as protease specificity, which is an intrinsic property of the protease's active site pocket [5]. The broad or narrow specificity of a protease refers to its ability to cleave a wide or restricted range of substrate sequences, respectively. As proteases have evolved to fulfill specific biological functions, the variations in protease specificity can be attributed to evolutionary adaptation [6]. Thus, proteases with broad specificity usually participate in processes such as protein degradation or processing, breaking down proteins into smaller peptides and amino acids. On the other hand, regulatory proteases typically possess narrow specificity, allowing precise cleavages of specific proteins in specific cellular contexts to regulate signaling pathways or protein activation/inactivation [7]. The 3D structure of substrates is not crucial for digestive proteases since the substrate is cleaved

into short peptides, losing its native 3D structure. However, it plays a significant role in regulatory proteolysis, as potential cleavage sites can be shielded within the hydrophobic core of the protein [8–10].

Our understanding of regulatory proteolytic networks in multicellular organisms is still limited, making it crucial to conduct further research in this direction. Experimental work to identify protease substrates and cleavage sites is highly labor-intensive. However, the guidance provided by bioinformatics predictions can significantly facilitate this process. To date, numerous bioinformatics methods exist for predicting proteolytic cleavage sites [11–20]. However, many of these methods primarily focus on modeling protease sequence preferences near the cleavage site and often do not incorporate information about the 3D structures of potential substrates [21]. While some methods incorporate specific structural features of the substrate into their prediction models [17,22], to our knowledge, only one method considers the 3D structure of the substrate as an input [23]. However, even this method does not provide separate probabilities for the structural susceptibility of the peptide bonds of the considered protein to proteolysis. The Nickpred method [8], which was originally developed for this purpose, is no longer available. To address this gap in the field, we developed a method for predicting the susceptibility of protein regions to proteolysis based on the known 3D structure of the potential substrate.

2. Results

2.1. The Three-Dimensional Structure of a Protease Substrate Determines the Susceptibility of Protein Regions to Proteolysis

Our aim was to develop a method that predicts the susceptibility of protein regions to proteolysis. This method takes the three-dimensional structure of a potential protease substrate as input and provides a cleavage susceptibility score for each peptide bond of the protein. Our predictive model was constructed using experimentally verified proteolytic events extracted from the CutDB [24]. We mapped the proteolytic sites from CutDB onto the available 3D structures of substrates extracted from the Protein Data Bank (PDB) [25] to create a training set. For each peptide bond, we calculated the structural features that were identified in earlier studies as predictors of proteolytic susceptibility [26,27]. The set of structural features used for prediction included well-known predictors such as solvent accessibility, secondary structure, and B-factor, as well as additional features developed by our group, including loop length and regions of flexible N- and C-termini (see Figure 1 and Methods for a complete list of features). We estimated the prediction quality of the developed method using the cross-validation technique on the training set and a separate testing dataset, which was collected from recent literature on proteolytic cleavage experiments. To maximize the prediction quality, we applied eight machine learning methods and ultimately selected Linear Discriminant Analysis (Figure 1B). Our training set was highly imbalanced, consisting of 445 positive examples (cleavage sites) and 68,840 negative examples. Therefore, we sampled the negative class in various proportions relative to the size of the positive class and examined the impact of the positive-to-negative class size ratio on the prediction quality (Figures 1C and S1). We found that the quality of the prediction was generally independent of the class size ratio; thus, to simplify calculations, we chose a 1:1 class size ratio (Figure 1C). We also visualized the predicted cleavage susceptibility scores on the 3D structures of substrates and confirmed that the method assigns higher scores to protruded loops, regions with a high B-factor, N- and C- flexible termini, and solvent-accessible regions (Figure 1D), as expected from earlier studies [26,27].

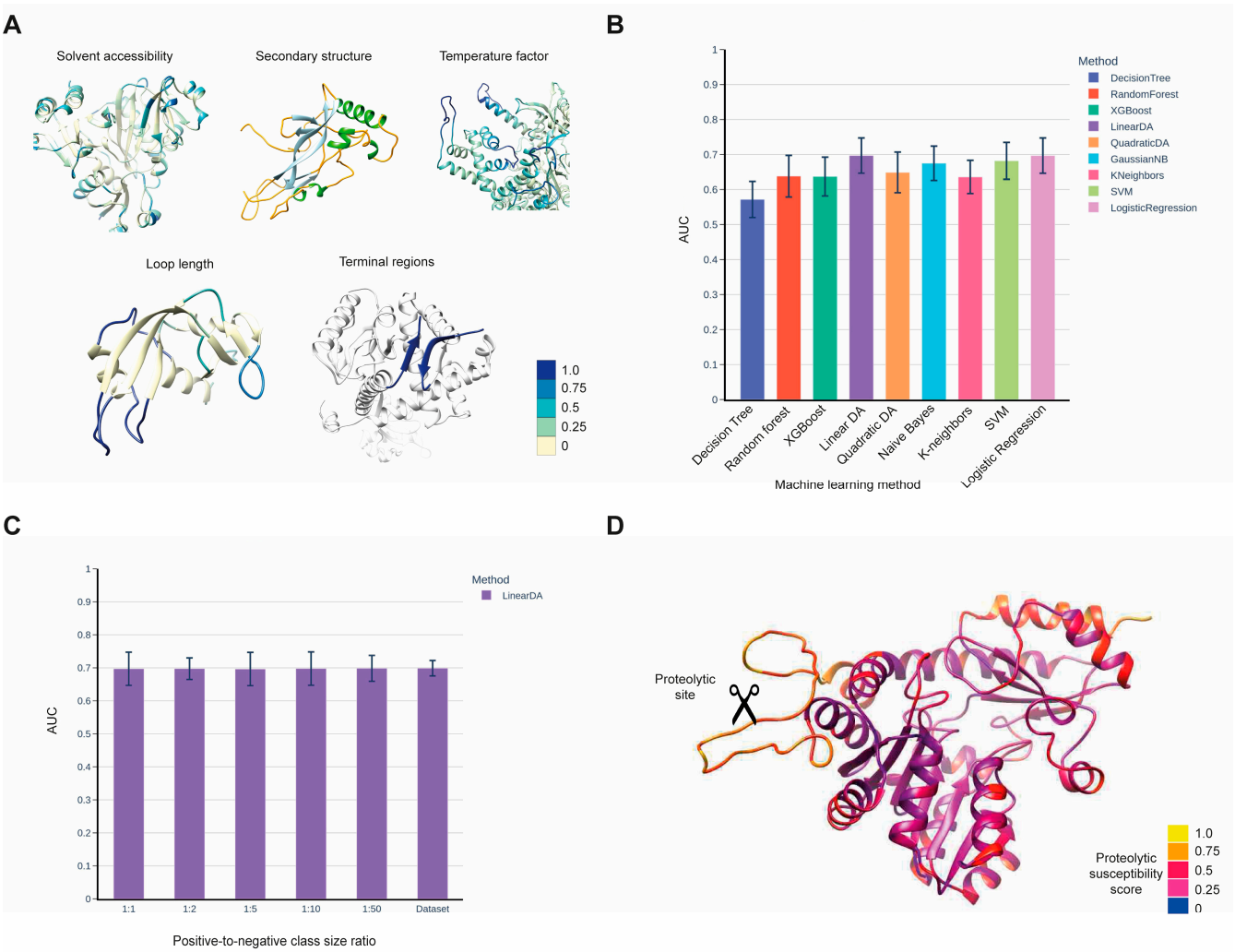

Figure 1. (**A**) List of structural features used in the method, along with examples of their distribution along the protein polypeptide chain visualized within the substrate structures. The color bar represents a color scale ranging from 0 to 1, indicating numerical features such as solvent accessibility, temperature factor, and loop length, as well as binary features such as terminal regions. The secondary structure is visualized using a different color scheme: helices are shown in green, beta strands in light blue, and loops in yellow. (**B**) Prediction quality, measured using the Area Under the ROC Curve (AUC), of various machine learning methods calculated via cross-validation using the training set of CutDB proteolytic events mapped onto PDB structures. Negative class examples were sampled to achieve a 1:1 positive-to-negative class size ratio. (**C**) Dependence of the method's prediction quality on different positive-to-negative class ratios. (**D**) Visualization of the proteolytic susceptibility probabilities predicted by our method for the 3D structure of the protease substrate.

2.2. Extension of the Training Set with AlphFold Models Improves the Quality of Prediction

Recent progress in artificial intelligence has led to breakthroughs in various fields of study, including computational molecular biology. Thus, the recently introduced protein structure prediction method AlphaFold has significantly outperformed other methods in this field and has demonstrated a prediction quality comparable to experimental methods [28]. Later, AlphaFold was applied to the entire human proteome, and high-quality predicted 3D structures were made publicly available in the AlphaFold Protein Structure Database [29]. We used 3D protein structures predicted by AlphaFold to expand our training set and construct a new model with the aim of comparing its prediction quality to the previous version of the model, which was solely based on PDB 3D structures. However, not all structural features extracted from PDB 3D structures are available in AlphaFold models,

notably, experimental-specific features such as the temperature factor (B-factor). Therefore, we reconstructed our initial PDB-based model, excluding the experiment-specific features, and compared its performance with the model constructed using the training set extended with AlphaFold-predicted 3D structures (Figure 2A). The latter model demonstrated better prediction quality and the difference between the median values of the models' Area Under the Curve (AUC) of the Receiver Operating Characteristic curve (ROC) [30] was 0.05. It is worth noting that the AlphaFold method provides values for the prediction confidence for each amino acid position of the protein. We analyzed whether this feature alone could predict susceptibility to proteolysis and found that it possesses substantial predictive power (Figure 2B).

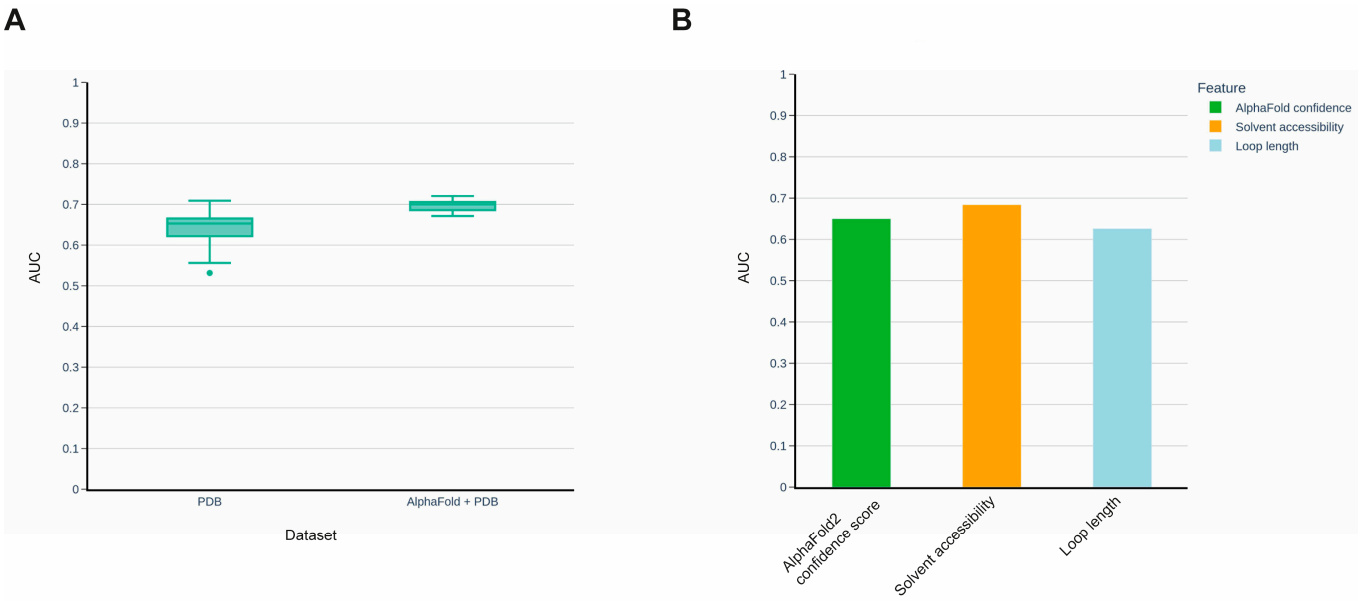

Figure 2. (**A**) Improvement in prediction quality of the method after extension of the training set using AlphaFold models. (**B**) Comparison of AlphaFold confidence score with solvent accessibility and loop length in predicting proteolytic sites.

2.3. Comparison to Other Proteolytic Site Prediction Methods

To the best of our knowledge, our method is the first to estimate the structural susceptibilities of protein regions to proteolysis regardless of specific proteases. Thus, there is currently no method available to directly compare prediction qualities. However, if we add knowledge on protease specificity into our method, for example by using a position-specific scoring matrix (PSSM) [31,32], we can compare our method with 3D structure-based methods that also incorporate information on protease specificity. To this end, we chose Procleave [23], the most recent and reliable method for the identification of proteolytic sites, for comparison. This method can predict cleavage sites for 27 proteases, including matrix metalloproteases, cathepsins, and other proteinases. We generated PSSM matrices for these proteases using data from the MEROPS database (see Section 4 [33]. To integrate the structural susceptibility predicted by our method with protease sequence specificity, we created a dataset that included two features: the structural score and the PSSM score. This dataset was used to train the prediction model using data from CutDB (Figure 3A). The obtained model demonstrated improved performance on the testing set compared to the Procleave method (Figure 3B). The AUC ROC mean values were 0.962 and 0.937, respectively, while the respective median values were 0.97 and 0.966. However, the statistical difference estimated using the Wilcoxon test was not found to be significant.

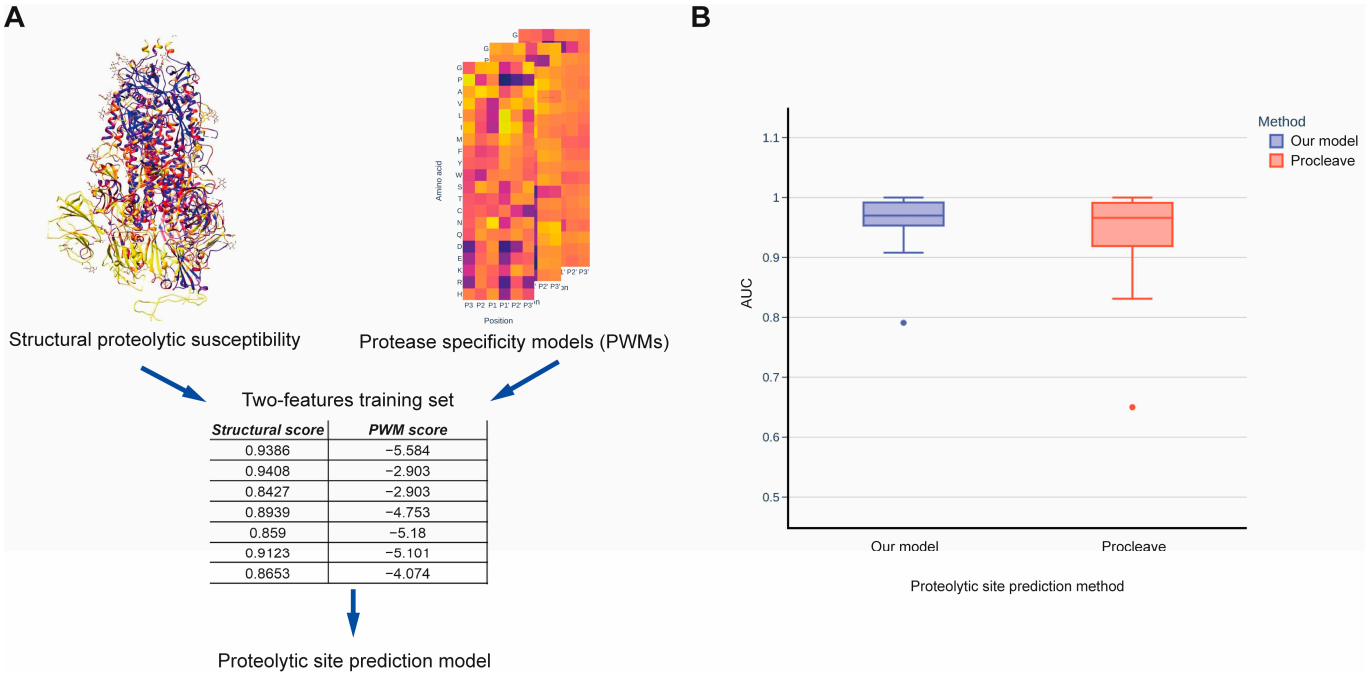

Figure 3. (**A**) A schematic representation of combining the proteolytic susceptibility probabilities predicted by our 3D structure-based method with protease sequence specificity models for comparison with other proteolytic site prediction methods. (**B**) Comparison of prediction quality between our method combined with protease sequence specificity models and the Procleave method.

3. Discussion

In this study, we presented a method for estimating structural susceptibility to proteolysis protein regions based on the known three-dimensional structure of a protein. It is known that the 3D structure of the protease substrate significantly influences the ability of protease to cleave a protein's peptide bonds [8–10,26,34,35]. Indeed, it is a common opinion that protein regions in the hydrophobic core of a protein are hardly accessible to proteolytic processing while the 3D structure of the protein is intact [8–10]. Another protein property that influences proteolytic processing is the secondary structure: our [26,27] and other [8–10,34,35] studies showed that loops are cleaved more easily than helices, and helices are cleaved more easily than beta-sheets. These and other known structural preferences of limited proteolysis seem universal for different types of proteases, contrary to protease sequence specificity [26]. Although several proteolysis prediction tools use specific structural features, there is, to our knowledge, no method that estimates the general susceptibility to proteolysis of protein regions based on known 3D structure. We developed and presented here such a type of method to fill this gap in the field.

We incorporated into the method structural features that influenced proteolytic processing according to current knowledge from our [26,27] and previous [8–10,34,35] studies. We also developed two additional structural features—loop length and N- and C- termini regions—based on our earlier observations [26]. To maximize prediction quality, we tried several machine learning methods and chose Linear Discriminant Analysis, which showed the best results for our task. Together, these efforts allowed us to develop a method demonstrating a quality of prediction comparable with state-of-the-art proteolytic site prediction tools, such as Procleave, when combined with protease primary specificity models.

Our method estimates the susceptibility of being proteolytically processed for each peptide bond of the considered protein with a known 3D structure. As the number of cleavages performed in a protein by a particular protease depends on its colocalization and the colocalization duration [36,37], we did not apply any threshold to the predicted cleavage probability. Thus, our method did not classify protein peptide bonds into presumably cleaved and uncleaved ones. Moreover, after the first cleavage, a substrate can change its

conformation or even become denatured; thus, predicted proteolytic sites can lose their confidence [38]. A new round of prediction is preferable if the 3D structure of the protein after conformations induced by the first cleavage changes is known.

Future research in proteolytic site prediction could focus on developing a method that combines prediction of structural susceptibility of protein regions, using the methodology developed in this study as a universal component applicable to all proteases, along with protease-specific models as plug-in modules. In this study, we demonstrated the relevance of this approach. In conclusion, we speculate that the scope of our method extends beyond the prediction of proteolytic events to encompass other post-translational modifications. Indeed, our method assigns higher probabilities of proteolytic sites to the hydrophilic exterior rather than the hydrophobic core of the protein, to protruded loops rather than helices and beta-sheets, and to flexible protein regions instead of the stable parts of the protein structure. This trend may hold true for various other post-translational modifications.

4. Materials and Methods

4.1. Data Collection and Processing

Information on experimentally verified proteolytic events was extracted from CutDB [39]. For each proteolytic event, we considered three attributes: the substrate identifier, the position of the proteolytic site within the substrate sequence, and the protease MEROPS code (Supplemental File S1). In total, we extracted 4576 proteolytic events related to 2062 unique substrates cleaved by 457 proteases. All substrate sequences were collected into a single FASTA file and then queried against the PDB database [25] using BLAST [40]. If the retrieved results included structures with a sequence identity of over 90% of the queried substrate sequence, the top structure in the list was associated with the protease substrate. Otherwise, the substrate was categorized as unmapped. In total, we found 585 three-dimensional structures of substrates associated with 1499 proteolytic events cleaved by 256 proteases. To map substrate amino acid positions into the 3D structure, we aligned substrate and 3D structure sequences using Clustal Omega [41]. The number of proteolytic sites upon mapping decreased by more than twofold (777 proteolytic sites, 323 structures, 183 proteases), as many of them were mapped into disordered regions. Since some of the secondary proteolytic cleavages observed in the experiments could occur after the loss of the intact substrate's 3D structure, we visualized the cleavage sites on the 3D structures using Chimera [42] and performed manual curation. We excluded proteolytic events if there were multiple cleavages attributed to a single publication and if they were predominantly located within the hydrophobic core of the substrate, indicating a potential loss of the 3D structure during the experiment. The final training dataset comprised 445 proteolytic events, specifically associated with peptide bonds in 190 3D structures of substrates that underwent proteolytic processing by 130 proteases. The average number of proteolytic events per protein in the training set was 2.34, with a median of one.

4.2. Structural Features

The selection of structural features for use in the prediction model was based on our earlier studies [26,27] as well as other relevant research in the field [8–10,34,35]. Primary structural features were solvent accessibility, secondary structure, and temperature factor (B-factor). Solvent accessibility and secondary structure were obtained using the DSSP tool [43]. B-factor was extracted from the PDB files of protease substrates' 3D structures. Based on our earlier observations of an increased density of cleavage sites in long protruded loops and C- and N-protein termini, we introduced two specific structural features associated with these observations. First, we added the length of the loop as a feature and assigned it to all peptide bonds within the loop. Second, we defined the regions of the C- and N-protein termini as unstructured terminal protein regions that are adjacent to the regular secondary structure elements, with the exception of short ones (see comments in the source code). Solvent accessibility, loop length, and B-factor were normalized using min-max scaling. The secondary structure was converted into three binary features that indicated

the presence or absence of specific types of secondary structures. C- and N-protein termini were encoded as a binary variable. A single binary variable was used to represent the C- and N-protein termini.

4.3. Training Set Processing and Selection of the Machine Learning Method

To create the training set, we computed the mentioned structural features for every amino acid in each substrate. Next, we assigned the structural features calculated for the amino acid at P1 position of the cleavage sites (Schechter–Berger notation) [44] to each peptide bond, as this position has previously been identified as the most important from a structural perspective [26,45–47]. We applied multiple machine learning methods from the scikit-learn library [48], such as Random Forest, Decision Trees, Naïve Bayes, SVM, Logistic Regression, XGBoost, Linear and Quadratic Discriminant Analysis, to identify the optimal method for our task (Figure 1B). The quality of the models was assessed via the AUC ROC metric using a 10-fold cross-validation technique. We also varied the proportions of the negative class relative to the positive class to assess the impact of the positive–negative class size ratio on the prediction quality and determine the optimal ratio (Figure 1C). We found that the quality of prediction was generally independent of the class size ratio; therefore, we chose a 1:1 class size ratio. Among the applied machine learning methods, Linear Discriminant Analysis demonstrated the best quality of prediction.

4.4. Extending Training Set with AlphaFold Models

Structure models were downloaded from the AlphaFold Protein Structure Database [29]. BLAST [40] was used to query the remaining protease substrates against the AlphaFold models. Filtering of the BLAST search results, mapping of the cleavage sites into AlphaFold models, and visualization followed a similar procedure as described above for the search against PDB. The numbers of proteolytic sites, substrates, and associated proteases at each filtering step were as follows: 3168, 1209, and 317 after the BLAST search step; 2925, 1209, and 317 after the mapping step; and 2918, 1205, and 314 after the curation step, respectively. In this dataset, proteins had an average of 2.42 proteolytic events, with a median of one event per protein.

4.5. Combining the Method with Protease Sequence Specificity Models

A testing set of proteolytic events was created using the MEROPS database [33]. We selected proteolytic events that were added to the database after the release of the Procleave method [23]. At each filtering step, the counts of proteolytic sites, substrates, and associated proteases were as follows: 213, 129, and 3 after extraction from MEROPS; 81, 48, and 3 after BLAST search; 43, 27, and 3 after mapping; and 28, 18, and 3 after the curation process, respectively. In the testing set, the average number of proteolytic events per protein was 2.34, while the median value was one. Protease sequence specificity models, in the form of PSSM matrices [31], were constructed following the method described in [49]. To combine the predicted values of structural susceptibility to proteolysis generated by our method with the sequence specificity scores generated by PSSM models, we created a training set that included these two features and applied the Naïve Bayes method. The obtained model was applied to the testing set and compared with the Procleave results.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms241310761/s1.

Author Contributions: Conceptualization, M.D.K.; methodology, M.D.K.; software, E.V.M.; data analysis, E.V.M., V.V.S., G.V.P. and M.D.K.; data curation, E.V.M. and M.D.K.; writing—original draft preparation, M.D.K.; writing—review and editing, M.D.K.; visualization, E.V.M., G.V.P. and M.D.K.; supervision, M.D.K.; project administration, M.D.K.; funding acquisition, M.D.K. All authors have read and agreed to the published version of the manuscript.

Funding: The reported study was funded by RFBR, project number 20-04-60066.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data and source code are available at: https://github.com/KazanovLab/ProteolysisStructuralPrediction (accessed on 15 June 2023). Development scripts are available at: https://github.com/EugeneVlg02/ProteolysisStructuralPrediction_development (accessed on 15 June 2023).

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

References

1. Barber, K.W.; Rinehart, J. The ABCs of PTMs. *Nat. Chem. Biol.* **2018**, *14*, 188–192. [CrossRef]
2. Conibear, A.C. Deciphering Protein Post-Translational Modifications Using Chemical Biology Tools. *Nat. Rev. Chem.* **2020**, *4*, 674–695. [CrossRef] [PubMed]
3. López-Otín, C.; Bond, J.S. Proteases: Multifunctional Enzymes in Life and Disease. *J. Biol. Chem.* **2008**, *283*, 30433–30437. [CrossRef] [PubMed]
4. Turk, B. Targeting Proteases: Successes, Failures and Future Prospects. *Nat. Rev. Drug Discov.* **2006**, *5*, 785–799. [CrossRef] [PubMed]
5. Ratnikov, B.I.; Cieplak, P.; Gramatikoff, K.; Pierce, J.; Eroshkin, A.; Igarashi, Y.; Kazanov, M.; Sun, Q.; Godzik, A.; Osterman, A.; et al. Basis for Substrate Recognition and Distinction by Matrix Metalloproteinases. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E4148–E4155. [CrossRef]
6. Fuchs, J.E.; von Grafenstein, S.; Huber, R.G.; Margreiter, M.A.; Spitzer, G.M.; Wallnoefer, H.G.; Liedl, K.R. Cleavage Entropy as Quantitative Measure of Protease Specificity. *PLoS Comput. Biol.* **2013**, *9*, e1003007. [CrossRef]
7. Neurath, H.; Walsh, K.A. Role of Proteolytic Enzymes in Biological Regulation (A Review). *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 3825–3832. [CrossRef]
8. Hubbard, S.J.; Beynon, R.J.; Thornton, J.M. Assessment of Conformational Parameters as Predictors of Limited Proteolytic Sites in Native Protein Structures. *Protein Eng.* **1998**, *11*, 349–359. [CrossRef]
9. Hubbard, S.J.; Campbell, S.F.; Thornton, J.M. Molecular Recognition: Conformational Analysis of Limited Proteolytic Sites and Serine Proteinase Protein Inhibitors. *J. Mol. Biol.* **1991**, *220*, 507–530. [CrossRef]
10. Hubbard, S.J.; Eisenmenger, F.; Thornton, J.M. Modeling Studies of the Change in Conformation Required for Cleavage of Limited Proteolytic Sites. *Protein Sci.* **1994**, *3*, 757–768. [CrossRef]
11. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M.R.; Appel, R.D.; Bairoch, A. Protein Identification and Analysis Tools on the ExPASy Server BT. In *The Proteomics Protocols Handbook*; Walker, J.M., Ed.; Humana Press: Totowa, NJ, USA, 2005; pp. 571–607. ISBN 978-1-59259-890-8.
12. Garay-Malpartida, H.M.; Occhiucci, J.M.; Alves, J.; Belizário, J.E. CaSPredictor: A New Computer-Based Tool for Caspase Substrate Prediction. *Bioinformatics* **2005**, *21* (Suppl. S1), i169–i176. [CrossRef]
13. Backes, C.; Kuentzer, J.; Lenhof, H.-P.; Comtesse, N.; Meese, E. GraBCas: A Bioinformatics Tool for Score-Based Prediction of Caspase- and Granzyme B-Cleavage Sites in Protein Sequences. *Nucleic Acids Res.* **2005**, *33*, W208–W213. [CrossRef]
14. Wee, L.J.K.; Tan, T.W.; Ranganathan, S. CASVM: Web Server for SVM-Based Prediction of Caspase Substrates Cleavage Sites. *Bioinformatics* **2007**, *23*, 3241–3243. [CrossRef] [PubMed]
15. Verspurten, J.; Gevaert, K.; Declercq, W.; Vandenabeele, P. SitePredicting the Cleavage of Proteinase Substrates. *Trends Biochem. Sci.* **2009**, *34*, 319–323. [CrossRef] [PubMed]
16. Song, J.; Tan, H.; Shen, H.; Mahmood, K.; Boyd, S.E.; Webb, G.I.; Akutsu, T.; Whisstock, J.C. Cascleave: Towards More Accurate Prediction of Caspase Substrate Cleavage Sites. *Bioinformatics* **2010**, *26*, 752–760. [CrossRef]
17. Barkan, D.T.; Hostetter, D.R.; Mahrus, S.; Pieper, U.; Wells, J.A.; Craik, C.S.; Sali, A. Prediction of Protease Substrates Using Sequence and Structure Features. *Bioinformatics* **2010**, *26*, 1714–1722. [CrossRef]
18. Song, J.; Tan, H.; Perry, A.J.; Akutsu, T.; Webb, G.I.; Whisstock, J.C.; Pike, R.N. PROSPER: An Integrated Feature-Based Tool for Predicting Protease Substrate Cleavage Sites. *PLoS ONE* **2012**, *7*, e50300. [CrossRef] [PubMed]
19. Song, J.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Akutsu, T.; Haffari, G.; Chou, K.-C.; Webb, G.I.; Pike, R.N.; Hancock, J. PROSPERous: High-Throughput Prediction of Substrate Cleavage Sites for 90 Proteases with Improved Accuracy. *Bioinformatics* **2018**, *34*, 684–687. [CrossRef]
20. Li, F.; Chen, J.; Leier, A.; Marquez-Lago, T.; Liu, Q.; Wang, Y.; Revote, J.; Smith, A.I.; Akutsu, T.; Webb, G.I.; et al. DeepCleave: A Deep Learning Predictor for Caspase and Matrix Metalloprotease Substrates and Cleavage Sites. *Bioinformatics* **2020**, *36*, 1057–1065. [CrossRef]
21. Li, F.; Wang, Y.; Li, C.; Marquez-lago, T.T.; Leier, A.; Rawlings, N.D.; Haffari, G.; Revote, J.; Akutsu, T.; Chou, K.; et al. Twenty Years of Bioinformatics Research for Protease-Specific Substrate and Cleavage Site Prediction: A Comprehensive Revisit and Benchmarking of Existing Methods. *Briefings Bioinform.* **2018**, *20*, 2150–2166. [CrossRef]
22. Kumar, S.; Ratnikov, B.I.; Kazanov, M.D.; Smith, J.W.; Cieplak, P. CleavPredict: A Platform for Reasoning about Matrix Metalloproteinases Proteolytic Events. *PLoS ONE* **2015**, *10*, e0127877. [CrossRef]

23. Li, F.; Leier, A.; Liu, Q.; Wang, Y.; Xiang, D.; Akutsu, T.; Webb, G.I.; Smith, A.I.; Marquez-Lago, T.; Li, J.; et al. Procleave: Predicting Protease-Specific Substrate Cleavage Sites by Combining Sequence and Structural Information. *Genom. Proteom. Bioinforma.* **2020**, *18*, 52–64. [CrossRef] [PubMed]
24. Igarashi, Y.; Eroshkin, A.; Gramatikova, S.; Gramatikoff, K.; Zhang, Y.; Smith, J.W.; Osterman, A.L.; Godzik, A. CutDB: A Proteolytic Event Database. *Nucleic Acids Res.* **2007**, *35*, D546–D549. [CrossRef] [PubMed]
25. Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data. *Nucleic Acids Res.* **2019**, *47*, D520–D528. [CrossRef] [PubMed]
26. Kazanov, M.D.; Igarashi, Y.; Eroshkin, A.M.; Cieplak, P.; Ratnikov, B.; Zhang, Y.; Li, Z.; Godzik, A.; Osterman, A.L.; Smith, J.W. Structural Determinants of Limited Proteolysis. *J. Proteome Res.* **2011**, *10*, 3642–3651. [CrossRef]
27. Belushkin, A.A.; Vinogradov, D.V.; Gelfand, M.S.; Osterman, A.L.; Cieplak, P.; Kazanov, M.D. Sequence-Derived Structural Features Driving Proteolytic Processing. *Proteomics* **2014**, *14*, 42–50. [CrossRef]
28. 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]
29. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50*, D439–D444. [CrossRef]
30. Fawcett, T. An Introduction to ROC Analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874. [CrossRef]
31. Wasserman, W.W.; Sandelin, A. Applied Bioinformatics for the Identification of Regulatory Elements. *Nat. Rev. Genet.* **2004**, *5*, 276–287. [CrossRef]
32. Boyd, S.E.; Pike, R.N.; Rudy, G.B.; Whisstock, J.C.; de la Banda, M.G. PoPS: A Computational Tool for Modeling and Predicting Protease Specificity. *J. Bioinform. Comput. Biol.* **2005**, *3*, 551–585. [CrossRef] [PubMed]
33. 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]
34. Novotný, J.; Bruccoleri, R.E. Correlation among Sites of Limited Proteolysis, Enzyme Accessibility and Segmental Mobility. *FEBS Lett.* **1987**, *211*, 185–189. [CrossRef]
35. Fontana, A.; Fassina, G.; Vita, C.; Dalzoppo, D.; Zamai, M.; Zambonin, M. Correlation between Sites of Limited Proteolysis and Segmental Mobility in Thermolysin. *Biochemistry* **1986**, *25*, 1847–1851. [CrossRef] [PubMed]
36. Daniel, R.M.; Cowan, D.A.; Morgan, H.W.; Curran, M.P. A Correlation between Protein Thermostability and Resistance to Proteolysis. *Biochem. J.* **1982**, *207*, 641–644. [CrossRef]
37. Parsell, D.A.; Sauer, R.T. The Structural Stability of a Protein Is an Important Determinant of Its Proteolytic Susceptibility in Escherichia Coli. *J. Biol. Chem.* **1989**, *264*, 7590–7595. [CrossRef] [PubMed]
38. Klein, T.; Eckhard, U.; Dufour, A.; Solis, N.; Overall, C.M. Proteolytic Cleavage—Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modi Fi Cation. *Chem. Rev.* **2018**, *118*, 1137–1168. [CrossRef]
39. Igarashi, Y.; Heureux, E.; Doctor, K.S.; Talwar, P.; Gramatikova, S.; Gramatikoff, K.; Zhang, Y.; Blinov, M.; Ibragimova, S.S.; Boyd, S.; et al. PMAP: Databases for Analyzing Proteolytic Events and Pathways. *Nucleic Acids Res.* **2009**, *37*, D611–D618. [CrossRef]
40. 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]
41. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [CrossRef]
42. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef] [PubMed]
43. Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637. [CrossRef] [PubMed]
44. Schechter, I.; Berger, A. On the Active Site of Proteases. 3. Mapping the Active Site of Papain; Specific Peptide Inhibitors of Papain. *Biochem. Biophys. Res. Commun.* **1968**, *32*, 898–902. [CrossRef] [PubMed]
45. Schilling, O.; Overall, C.M. Proteome-Derived, Database-Searchable Peptide Libraries for Identifying Protease Cleavage Sites. *Nat. Biotechnol.* **2008**, *26*, 685–694. [CrossRef]
46. Alves, M.F.M.; Puzer, L.; Cotrin, S.S.; Juliano, M.A.; Juliano, L.; Bromme, D.; Carmona, A.K. S3 to S3′ Subsite Specificity of Recombinant Human Cathepsin K and Development of Selective Internally Quenched Fluorescent Substrates. *Biochem. J.* **2003**, *373*, 981–986. [CrossRef]
47. Debela, M.; Magdolen, V.; Schechter, N.; Valachova, M.; Lottspeich, F.; Craik, C.S.; Choe, Y.; Bode, W.; Goettig, P. Specificity Profiling of Seven Human Tissue Kallikreins Reveals Individual Subsite Preferences. *J. Biol. Chem.* **2006**, *281*, 25678–25688. [CrossRef]

48. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
49. Nishida, K.; Frith, M.C.; Nakai, K. Pseudocounts for Transcription Factor Binding Sites. *Nucleic Acids Res.* **2009**, *37*, 939–944. [CrossRef]