PNAS

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Prediction and design of protease enzyme specificity using a structure-aware graph convolutional network

Changpeng Lu[a], Joseph H. Lubin[b], Vidur V. Sarma[a], Samuel Z. Stentz[c], Guanyang Wang[d], Sijian Wang[a,d], and Sagar D. Khare[a,b,1]

Edited by David Baker, University of Washington, Seattle, WA; received March 5, 2023; accepted August 14, 2023

Site-specific proteolysis by the enzymatic cleavage of small linear sequence motifs is a key posttranslational modification involved in physiology and disease. The ability to robustly and rapidly predict protease–substrate specificity would also enable targeted proteolytic cleavage by designed proteases. Current methods for predicting protease specificity are limited to sequence pattern recognition in experimentally derived cleavage data obtained for libraries of potential substrates and generated separately for each protease variant. We reasoned that a more semantically rich and robust model of protease specificity could be developed by incorporating the energetics of molecular interactions between protease and substrates into machine learning workflows. We present Protein Graph Convolutional Network (PGCN), which develops a physically grounded, structure-based molecular interaction graph representation that describes molecular topology and interaction energetics to predict enzyme specificity. We show that PGCN accurately predicts the specificity landscapes of several variants of two model proteases. Node and edge ablation tests identified key graph elements for specificity prediction, some of which are consistent with known biochemical constraints for protease:substrate recognition. We used a pretrained PGCN model to guide the design of protease libraries for cleaving two noncanonical substrates, and found good agreement with experimental cleavage results. Importantly, the model can accurately assess designs featuring diversity at positions not present in the training data. The described methodology should enable the structure-based prediction of specificity landscapes of a wide variety of proteases and the construction of tailor-made protease editors for site-selectively and irreversibly modifying chosen target proteins.

protease specificity | machine learning | geometric machine learning | protein design | yeast surface display

Significance

Enzymes that can precisely and selectively read, write, and edit DNA have revolutionized biochemical sciences and technologies. The availability of similar enzymes for site-selectively "editing" proteins would have broad impact. Proteases are a large class of enzymes that have the ability to site-selectively cleave target proteins and therefore serve as examples to learn the rules of site-specific recognition, and as starting points for designing targeted cleavage. We present a geometric machine learning approach that uses protein structure and energetics to enable protease–substrate specificity prediction and design targeting alternative substrates. These studies set the stage for the large-scale prediction and design of tailor-made proteases that can site-selectively edit (cut) any chosen target protein, associated, for example, with a disease state.

Multispecificity, the specific recognition and nonrecognition of multiple substrates by protease enzymes, is critical for many biological processes and diseases (1–5). For example, the selective recognition and cleavage of host and viral target sites by viral and host protease enzymes is critical for the lifecycle of many RNA viruses, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (6–10). Identifying proteolytic targets of proteases would, therefore, provide deeper insights into the mechanisms and biological functions of proteases (3, 11). As protease inhibitors are often designed to mimic substrates, the ability to predict substrates may also aid inhibitor design against novel viruses (12–15). Furthermore, the ability to infer the global landscape of protease specificity, i.e., the set of all substrate sequence motifs that are recognized (and not recognized) by a given enzyme, would also enable the selection or design of bespoke proteases with specificities to degrade chosen biotechnologically relevant or disease-related targets (16–20).

Current experimental methods for protease substrate cleavage site identification involve assaying libraries of potential substrates for cleavage, one protease variant at a time (1, 21–26). Apart from being labor-intensive and time-consuming, only limited sampling of the protease:substrate sequence diversity is possible. Therefore, the development of rapid, cost-effective and generalizable computational approaches for precise prediction of specificity is valuable. Most current computational approaches for protease specificity prediction involve detecting and/or learning patterns in known substrate sequences using techniques ranging from inferred substitution matrices (27–32) to supervised machine learning (ML) (33–40). In some approaches [e.g., Procleave (41)], the accessibility of substrates depending on their solvent exposure and secondary structure assignment are also considered during prediction. We previously developed a supervised ML–based approach for specificity prediction in which protease–substrate interaction energy terms for the interface were considered (42, 43). We found that energetic terms play an important role in helping rank probabilities of cleavage. Similarly, inclusion of energetics in ML models was found to increase classification accuracy for identifying metalloenzymes (44). While computational approaches have successfully guided experiments in finding novel

Author contributions: C.L., J.H.L., S.W., and S.D.K. designed research; C.L., J.H.L., V.V.S., and S.Z.S. performed research; C.L., J.H.L., V.V.S., G.W., and S.W. contributed new reagents/analytic tools; C.L., J.H.L., V.V.S., G.W., S.W., and S.D.K. analyzed data; and C.L., V.V.S., J.H.L., and S.D.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: sagar.khare@rutgers.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2303590120/-/DCSupplemental.

Published September 20, 2023.

cleavage sites and obtaining a better understanding of protease–substrate interactions, these black-box approaches do not provide physical/chemical insight into the underlying basis for a particular experimentally observed specificity profile, nor are they robust to substitutions in the protease, requiring retraining for every protease variant, thus making these unsuitable for guiding protease design. Thus, there is need for interpretable and generalizable computational models of protease specificity.

We reasoned that a more semantically rich model of specificity would encompass both substrate sequence and the explicit energetics of the protease–substrate complex. Specificity depends on the residue-level interactions between enzymes and substrates, and for this reason, we hypothesized that a high-resolution energetic representation of a protease–substrate complex will have a high predictive value. As the energies of a protein are a consequence of sequence, we anticipated that a sufficiently granular and accurate energetic representation may obviate the need for sequence features. Using energies rather than sequence-based models for protease specificity naturally enables the design of proteases by training on directed evolution trajectories aimed at altering protease specificity for benchmarking (45). To encode the topology and energetic features of protease–substrate complexes for modeling specificity landscapes, here we develop a Protein Graph Convolutional Network (PGCN). PGCN uses experimentally derived data and a physically intuitive structure-based molecular interaction energy graph to pose specificity prediction as a classification problem. We find that PGCN consistently performs as well as or better than other previously used ML models for substrate specificity prediction especially when using energy-only features. A single PGCN model can effectively predict specificities for multiple protease variants, and ablation tests enable identification of critical subgraph patterns responsible for observed specificity patterns, highlighting the interpretability of the model. We then use PGCN to guide the design of protease libraries aimed at cleaving noncanonical substrates for TEV (Tobacco Etch Virus) protease, and experimentally validate these predictions using a yeast surface display-based assay. Importantly, designs included residue positions and substitutions not present in the training set, speaking to the high generalizability of PGCN.

Results

Overview of PGCN. We present a PGCN, which models protein structures and their complexes as fully connected graphs encoding sequence and single-residue and pairwise interaction energies generated using Rosetta (46). For the protease–substrate complexes, the substrate peptide is recognized by the protease for cleavage or rejection in the active site (Fig. 1*A*). The enzyme–substrate graph (Fig. 1*B*) is fed into a graph convolutional neural network, which outputs a probability of cleavage for a given complex (Fig. 1*C*). Our protease specificity dataset consists of experimentally determined cleavage information, i.e., lists of cleaved and uncleaved peptides for the wild type and variants of two viral proteases, NS3/4 protease of the Hepatitis C Virus (referred to as HCV protease in the following) (Dataset S1), and TEV protease (Dataset S2) obtained from Pethe et al. (42) and Packer et al. (45). The pools of experimentally confirmed cleaved and uncleaved substrates were randomly split into 80% training, 10% validation, and 10% test datasets.

PGCN Performs Better than Baseline ML Models for Substrate Specificity Prediction for Various Feature Encodings. To evaluate the performance of PGCN predicting substrate specificity, we first trained and tested models for specificity landscapes of WT (wild-type) and three HCV protease variants, A171T, D183A, and R170K/A171T/D183A (Table 1). We further combined all HCV protease variant data and trained and tested a single PGCN model on this combined set to explore how sensitive PGCN is in discriminating specificity changes upon small structural changes in the protease.

In benchmark tests, PGCN outperformed other ML models for all HCV variants using sequence features only (Fig. 2*A*), achieving more than 90% test accuracy for all datasets, including the combined dataset. We evaluated PGCN performance using different metrics besides accuracy, including F1 score, Precision, Recall, Area under curve (AUC), and Average Precision (AP), all standard evaluation metrics for ML tasks with imbalanced data (47–50). PGCN had the highest F1, Recall, and AP scores of the benchmarked methods (example: 93.53% F1, 96.85% Precision, 90.44% Recall, 97.90% AUC, 96.05% AP for A171T protease using sequence features only) (see details in Dataset S3).

We then evaluated PGCN's performance when using energy features. In these tests, we used either Rosetta energy information only, or sequence and Rosetta energy information together as features used in PGCN, see *Materials and Methods* for Rosetta energies details. As shown in Fig. 2 *B* and *C*, PGCN always performed the best with either energy features only or complete sequence and energy features. This result is remarkable because previous energy-based scoring approaches for protease–peptide interactions, which involved weighted sums of different energy terms, did not perform as well as sequence-based learning approaches (42, 43). A key difference between other energy-based models and PGCN is how calculated energies of interaction are used as features. In all models other than PGCN, energies are learned in simple linear combinations, while PGCN takes advantage of graph representation to encode intermolecular energies in an implicitly nonlinear relationship. Therefore, our results show that graph-based convolution of individual energy terms is a promising approach for combining biophysical analysis and data-driven modeling in a way that addresses some of the limitations of each.

Having demonstrated good performance in predicting substrate specificities when provided training data including a large pool of substrates and sparse protease diversity, we sought to evaluate PGCN performance in predictions involving greater protease diversity. Therefore, we trained PGCN on the engineered TEV protease dataset, which had a larger set of protease variants than our HCV set, although fewer substrates per variant were experimentally assayed (45) (Table 1 and *SI Appendix*, Table S1). We trained on this TEV dataset using the same three sets of feature encodings, either sequence-only information, energy-only information, or complete sequence and energy information.

All ML models are able to learn some patterns for TEV data if considering sequence features only, but tree-based approaches, SVM (support vector machine), and ANN (artificial neural network) achieved lower accuracy when considering energy features (Fig. 2 *B* and *C*). PGCN's performance is stable among different feature encodings, with accuracies of 86.86%, 86.62%, and 87.72% when using sequence-only (Fig. 2*A*), energy-only (Fig. 2*C*), and sequence+energy features (Fig. 2*B*), respectively. PGCN takes advantage of encoding residue-level pairwise energies into edges of graphs that enable PGCN to learn the local environments of residues at each GCN layer. Furthermore, PGCN with complete sequence and energy features outperforms the models with reduced features (e.g., SVM), supporting our hypothesis that the prediction of protease specificity benefits from both sequences of substrates and physical energies of interaction between enzyme and substrate.

To ensure that PGCN performance, especially with sequence features, is not dominated by memorization of substrate sequence patterns during training (and detecting similar patterns in the test

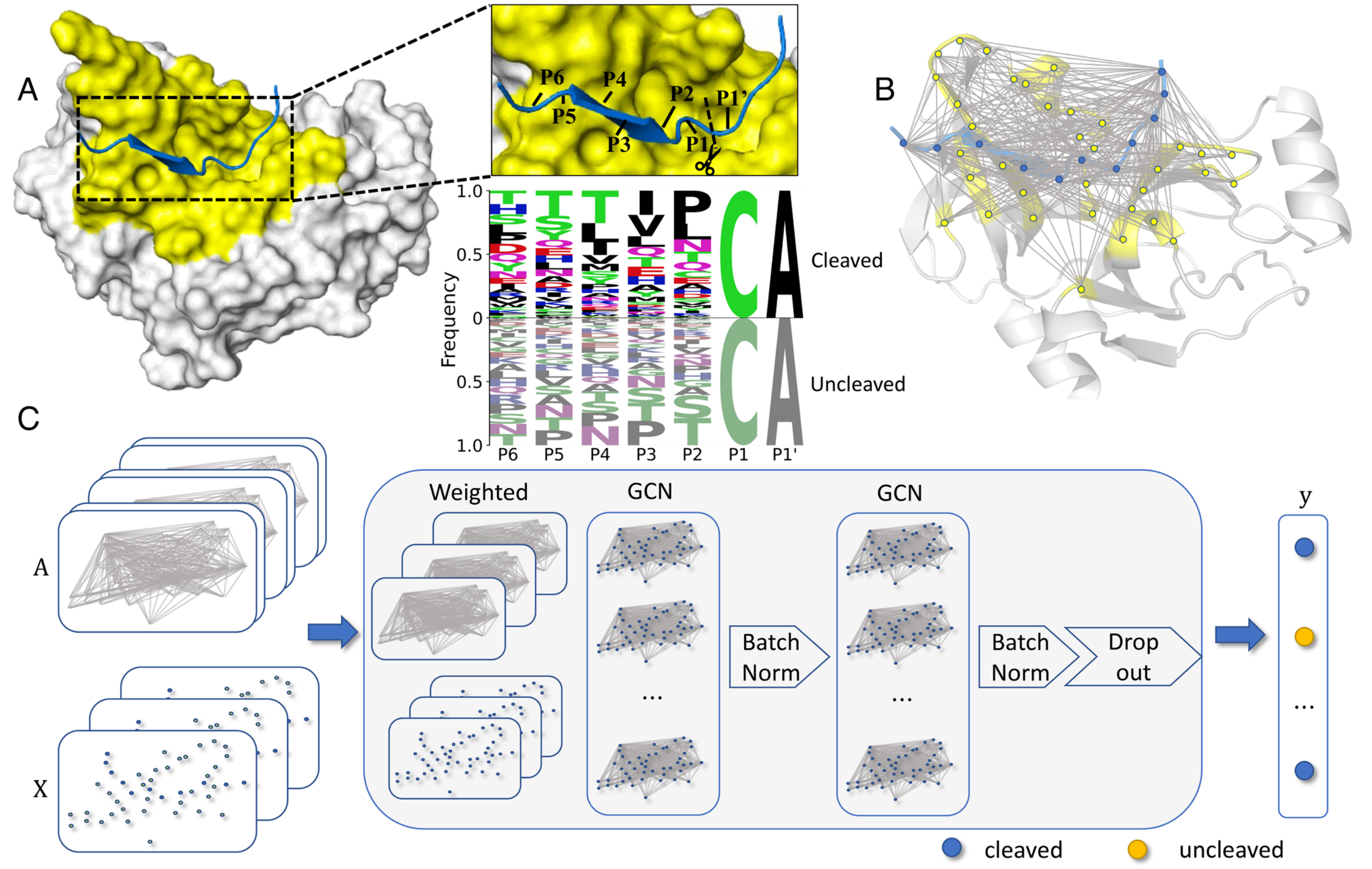


Fig. 1. Architecture of PGCN (*A*) Peptide substrate (blue) in the binding pocket (yellow) of HCV protease (gray). The seven-residue substrate spans P6 to P1′, with cleavage between P1 and P1′. The logo plot indicates the substrate sequences in the training set, where P1 and P1′ were kept constant, and P6 to P2 were variable. (*B*) Molecular depiction of the nodes and edges as a graph. Each substrate (blue) and binding pocket (yellow) amino acid constitutes a node of the graph. Gray lines between pairs of residues denote edges between pairs of nodes. (*C*) PGCN model architecture. Nodes are represented as a $N \times F$ matrix of nodes and node features. Edges are represented as a $N \times N \times M$ tensor of node pairs and edge features, flattened by the weighted sum of overall edge features. The PGCN model ultimately outputs probabilities of the given substrate belonging to each class, cleaved and uncleaved.

set), we also trained PGCN models using a train, validation, test split strategy based on Kmeans clustering of substrate sequences in cleaved and uncleaved pools such that substrate sequences in each set are sequence-distant from the other two sets. We find that PGCN still has the highest performance for the TEV Combined dataset regardless of which feature encoding is considered, and it dominates the prediction with an accuracy of 86.41% when using sequence+energy features compared with other ML models (the best accuracy: 75.96% for SVM) (Fig. 2*D*). Similar results are obtained for HCV protease (*SI Appendix*, Fig. S1). Thus, we conclude that PGCN-based discrimination is not based on memorizing or learning (nearest-neighbor) substrate sequence patterns, and therefore employ node-edge ablation tests to further investigate the sources of PGCN performance.

Node-Edge Importance Analysis to Obtain Physical Insights from PGCN. One advantage of PGCN is that the nodes and edges correspond directly to physical amino acid residues and their relationships. Therefore, we reasoned that we could identify important residues and interactions by identifying nodes and edges found to be critical for PGCN performance. To identify the prediction strength of each graph component by PGCN, we perturbed feature values of each node (or edge) across all sample graphs and computed accuracy again (see *Node/Edge*

Table 1. Summary of input dataset

	Training			Validation			Test			
Protease variant	Cleaved	Uncleaved	Total	Cleaved	Uncleaved	Total	Cleaved	Uncleaved	Total	Total
HCV WT	1,566	4,307	5,873	175	559	734	191	544	735	7,342
HCV A171T	2,905	7,659	10,564	366	954	1,320	373	948	1,321	13,205
HCV D183A	3,538	5,953	9,491	422	764	1,186	390	797	1,187	11,864
HCV Triple*	2,496	2,974	5,470	315	369	684	324	360	684	6,838
HCV Combined	10,404	20,995	31,399	1,319	2,606	3,925	1,338	2,587	3,925	39,249
TEV Combined	2,111	2,229	4,340	259	283	542	238	305	543	5,425

*Mutations made for HCV Triple: R170K, A171T, D183A.

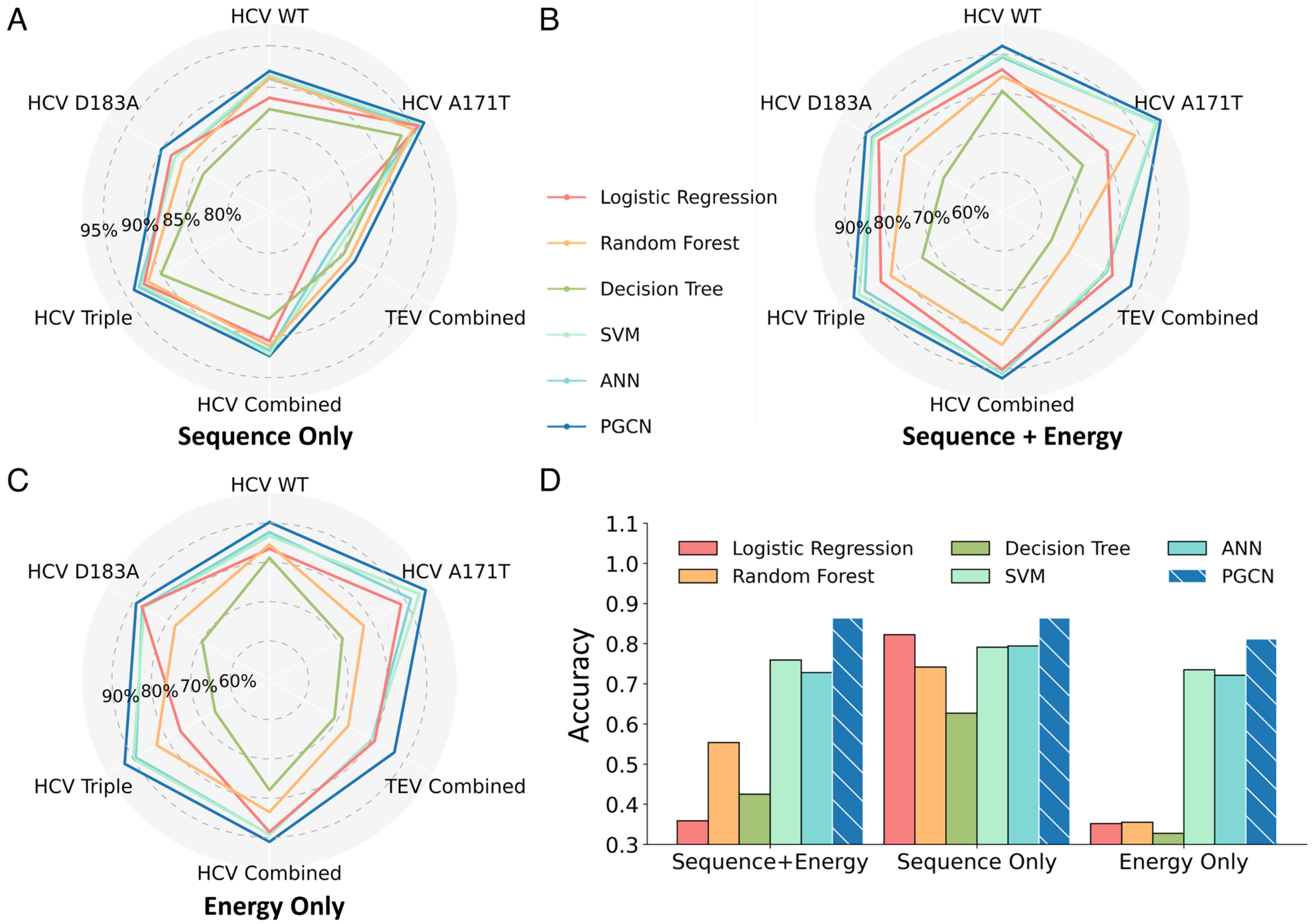


Fig. 2. PGCN performance. We evaluate models on six datasets, four consisting of a single HCV variant (WT, A171T, D183A, or Triple (R170K, A171T, D183A)), with various substrates, and two, Combined, one of which pools the other four, and the other consists of 10 TEV variants. (*A*–*C*) The radar plots show polygon patterns of average test accuracies across three seeds of benchmarked ML models (labeled in different colors) on the five datasets. The highest accuracy is on the polygon periphery. (*D*) Accuracy barplot of model prediction performance on the TEV Combined protease specificity data. In this analysis, the substrate were clustered based their sequence to ensure that training, validation, and test sets have distinct sequence patterns.

Importance in Materials and Methods). The decrease in accuracy upon perturbation is used to measure the (relative) importance of node *i* (or edge *j*) in the PGCN graph.

We normalized the calculated importance of node/edge by the overall accuracy of the prediction and aggregated the normalized importance by feature type (node or edge) to see how the features used by PGCN for training affected the classification. There are two types of nodes (protease, substrate) and three types of edges (protease–protease, substrate–substrate, and intermolecular) depending on the types of nodes that are connected by a given edge. When the sequence is the only feature (nothing on edges), as expected only peptide nodes contribute to accuracy for single-variant sets (Fig. 3*A*). However, for datasets in which protease diversity is also sampled ("Combined" dataset in Fig. 3*A*), protease nodes, typically sites of substitutions, are also detected as contributors to accuracy. When energy features are considered either solely or together with sequences, protease nodes make significantly greater contributions (Fig. 3*A*), indicating that protease residue energies are sensitive to the changes in their environment. In the same vein, when the sequence information is excluded, the dependence on edge features increases while the overall accuracy of prediction is not significantly affected. Leveraging energy information allows broader attention to residue–residue interactions as more edges are deemed significantly important, as shown in *SI Appendix*, Fig. S2. These observations show that sequences are an abstraction that PGCN uses as a shortcut when available, but the same information can be learned from energy.

Next, we visualized the positions of important nodes and edges in the HCV protease structure. For the WT and each variant protease, a key edge was found to be between the P2 residue of the substrate and the catalytic base H72, which presumably reflects the proper positioning of the substrate in the active site. We also observed that some important nodes/edges were different between WT and variant proteases. For example, the protease edge R138-D183 is prominent for the wild type (Fig. 3*B*), but it is not a significant interaction when either A171T (Fig. 3*C*) or D183A (Fig. 3*D*) or the Triple variant A171T, D183A, and R170K (Fig. 3*E*). When D183A mutation is introduced (Fig. 3 *D* and *E*), side chain orientations of some intermolecular edges, e.g., P6-R138, were different even though the protease node D183A itself did not significantly influence classification for substrate specificity. Conversely, we found that some other intermolecular edges such as P3-I147, P2-A171(T), P4-A171(T), P4-V173, and P6-V173 are at least two times more important for models trained on D183A and Triple variants than those for models trained on the wild type and A171T data (Dataset S4). Also, protease node R170(K) shows its importance only if D183A is mutated (Fig. 3 *D* and *E*). However, the important edge and node lists are not additive: for example, edges 138 to 173 and 96 to 170 are insignificant for the Triple variant (Fig. 3*E*), while they are two of the most important protease edges for the model trained on D183A

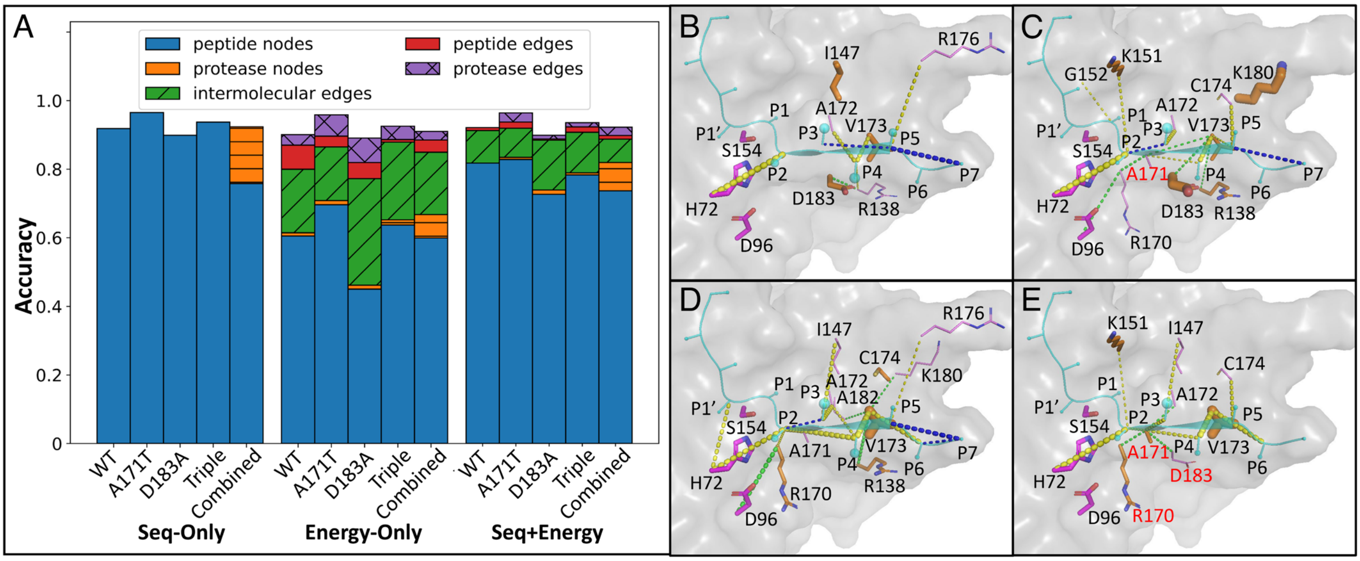


Fig. 3. Node/edge importance analysis for HCV protease. (*A*) Relative importance of major physical groups of nodes/edges. Node/edge importance is analyzed by the decrease in accuracy upon perturbation normalized by the original test accuracy of the PGCN model trained on each of five HCV protease variants using sequence-only (S), energy-only (E), or both (S+E). All nodes and edges in the PGCN graph are grouped into five major physical groups, and each group's total relative importance is aggregated as the sum of importance scores from nodes (or edges) of the specified type. (*B*–*E*) Structural representations of node/edge importance for HCV protease (PDB ID 3M5N; 3M5L), calculated by the energy-only PGCN model trained on *B* wild type; (*C*) A171T; (*D*) D183A; (*E*) Triple data. Only nodes/edges whose relative importance scores are within the first 25% quantile are displayed in the zoom-in protease structure (gray), with the substrate (cyan) as the center and the catalytic triad (magenta) as the reference. Among those nodes that meet the criteria, the relative importance levels of protease nodes (orange) are shown by the thickness of corresponding residue side chains, while that of peptide nodes (cyan) is reflected by the sizes of corresponding residue spheres at CB. For those important edges that meet the criteria, different groups are highlighted in different colors, including peptide edges (blue), protease edges (green), and intermolecular edges (yellow). All residues related to node/edge importance have labels in residue identifiers with one-letter codes (colored in red if mutation sites), including protease residues that are only related to edges (violet).

(Fig. 3*D*). When taking the model trained on the HCV combined set into consideration, although most of the important nodes and edges for single-variant sets are equally important for the combined set (such as node V173, edge P2-H72, P4-V173, etc.), there are also some important nodes/edges that are only useful for the prediction of substrate specificity within individual variants or the wild type, such as P3-I147 (Dataset S4). Thus, the model is able to classify to approximately the same level of accuracy using overlapping but distinct sets of node and edge features.

Similar trends were apparent in the node/edge sensitivity analyses for TEV protease predictions (Fig. 4*A*). Several positions that are the sites of amino acid substitutions in the TEV variants had high importance, such as D148, S170, and N177 (Fig. 4*B*). Strong signals of some important interactions were also identified. For example, the interaction between the P2 residue of the substrate and the catalytic base H46 is consistently important across all TEV variants and the wild type (Fig. 4*C*), the same as in the HCV prediction described above. In addition, several other intermolecular edges, e.g., P3-S170 (Fig. 4*C*), intraprotein edges around the S3 pocket (Fig. 4*D*) and the S1 pocket (Fig. 4*E*) were also found to be of high importance.

Taken together, the analyses for HCV and TEV protease variants follow a series of general rules. First, the intermolecular edge between P2 and the histidine in the catalytic triad (P2-H72 for HCV in Fig. 3 *B*–*E*, and P2-H46 for TEV in Fig. 4*C*) is always one of the most important edges, reflecting the proper positioning of the substrate in the active site. Second, those interactions that presumably provide the H-bonds that template the substrate into the required β-sheet conformation were also identified as prominent edges, such as P4-V173 for HCV (Fig. 3 *B*–*E*), P3-S170, P3-F217 for TEV (Fig. 4*C*). This is consistent with the well-known observation that protease substrates always adopt an extended conformation in the active site with beta-sheet complementation (51). Some important nodes/edges form interconnected clusters that are consistent with the canonical substrate binding pockets (S pockets), for example, the S3 subpocket for TEV includes residues 170, 172, 217 (Fig. 4*D*); the S1 pocket for TEV including interactions with 148, 167, 170 (Fig. 4*E*). Thus, we argue that the PGCN models have learned to discriminate between cleaved and uncleaved substrates based on criteria that can have an interpretable biophysical basis in some cases, and reflect nonobvious statistical relationships between various interactions in other cases.

Exploring PGCN Generalizability by Cross-Test and Leave-One-Out Test Analyses. To investigate if the models learned by PGCN for a single-variant protease could be further generalized to protease variants outside the training data, we designed cross-test analysis for HCV datasets and leave-one-out analysis for the TEV dataset. First, we cross-tested prediction accuracy for each PGCN model trained on one of five HCV datasets on the other four. In each cross-test experiment, we employed the identical training subset utilized in the previous evaluation of PGCN's performance. Subsequently, we conducted testing on the other four HCV datasets, excluding substrates that exhibited redundancy in both the training and test data. Although the AUC of the self-test PGCN model is consistently higher than that of the cross-test model for each test dataset, cross-test PGCN models still have a good ability to discriminate protease specificity, reaching at least 81% AUC score when testing the combined datasets excluded the trained variant, shown in Fig. 5*A*. Other metrics of performance are included in Dataset S7.

Next, to measure PGCN's ability to generalize over protease variants with multiple substitutions, we chose three TEV variants, **N176I**, **I138T**/N171D/N176T and E107D/D127A/S135F/**R203Q/K215E** (Var2), which have one or two unique mutation, respectively, among all of ten TEV variants for the leave-one-out test purpose. For each of these three variants, we divided all of its substrates in cleaved and uncleaved pools into the test dataset, and split the remained data into training and

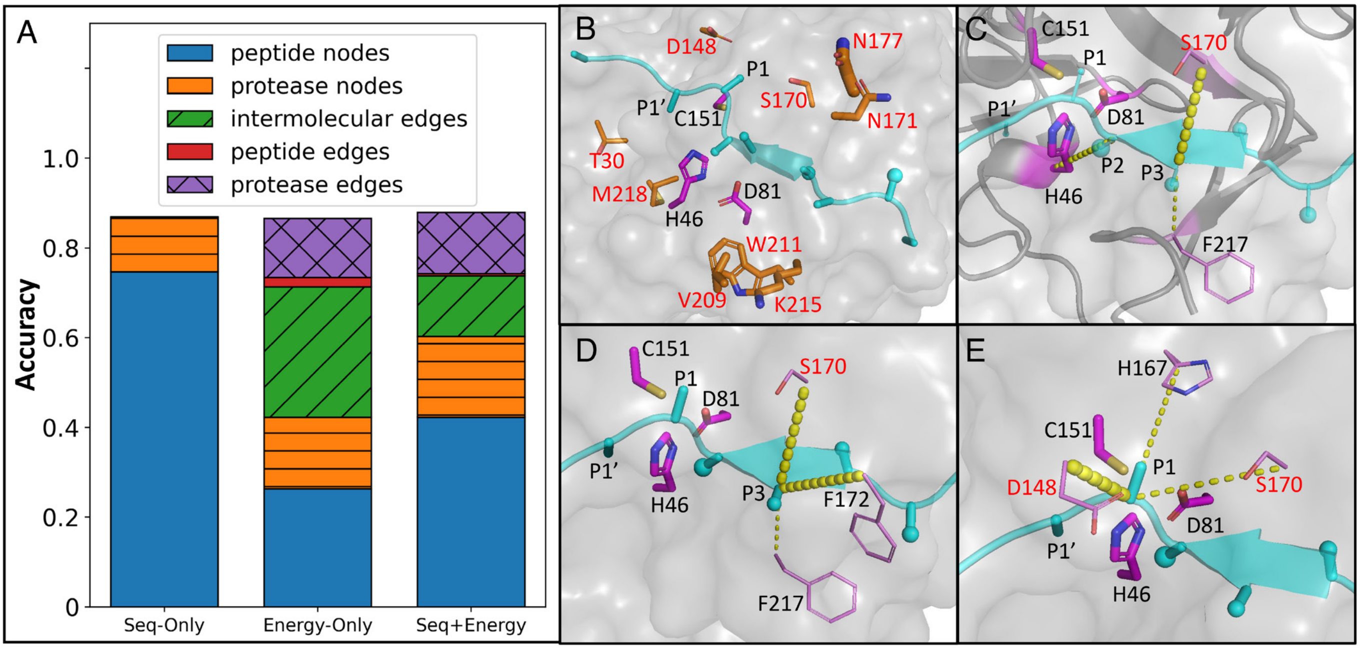


Fig. 4. Node and edge importance contribution for TEV. (*A*) Relative importance of major physical groups of nodes/edges. (*B*–*E*) Partial structural representations of node and edge importance in the context of TEV (PDB ID 1LVB; 1LVM) to show: (*B*) important protease nodes; (*C*) intermolecular edges that presumably form hydrogen bonds; (*D*) the subpocket surrounding P3, including P3-S170, P3-F172, P3-F217; (*E*) the subpocket surrounding P1, including intermolecular edges P1-D148, P1-H167, P1-S170. The same color setting is used as in Fig. 3. The catalytic triad in each structure is just the reference of relative position. All mutations are highlighted in red.

validation datasets based on the ratio of 9:1. In Fig. 5*B*, PGCN models always have the highest AUC (96.63% AUC for N176I as an example) compared with the best baseline ML model (SVM: 75.36% AUC for N176I). For other variants, PGCN also outstands among other ML models, as shown in *SI Appendix*, Fig. S3.

Therefore, from the cross-test analysis and leave-one-out analysis, we conclude that PGCN models have the ability of transfer learning for the prediction of protease specificity compared with other ML models; on the other hand, PGCN models self-trained on the single mutation can still maximize its power for prediction of the dataset containing the same mutation.

Experimental Evaluation of PGCN Generalizability Using Protease Design. To further test if PGCN is able to generalize its classification ability to protease variants that are not in the training dataset, we turned to TEV protease specificity design. WT TEV protease demonstrates a preference for an ENLYFQ/X motif at the P6-P1′ positions of its canonical substrate (X = A,G,S) (45, 52, 53). We aimed to design proteases against altered substrates with single-residue substitutions within the canonical recognition motif (P6: KNLYFQ/A, P2: ENLYYQ/A). A substitution of K at P6 or Y at P2 resulted in no cleavage of the substrate by WT TEV (*SI Appendix*, Fig. S4), providing a well-posed problem for designs using PGCN – given a set of designs, predict which designed variants would lead to cleavage of P6 and P2 target substrates.

We applied Rosetta-based computational design (*SI Appendix, Computational Design Process for TEV Protease*) to propose sequences (4,320 P6-targeted designs and 280 P2-targeted designs) that included stabilizing interactions with the target substrates (Fig. 6*A*). We then used our pretrained TEV protease model to score designs, and identified those with a high predicted probability of cleavage (Fig. 6*B*). PGCN selected 200 of 280 P2-targeted designs and 126 of 4,320 P6-targeted designs with high predicted cleavage

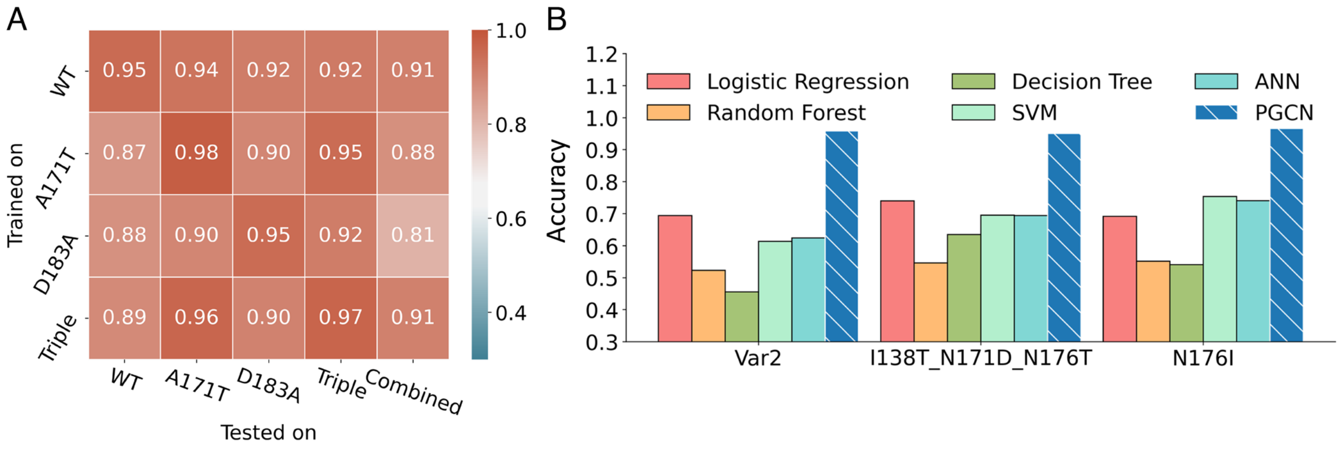
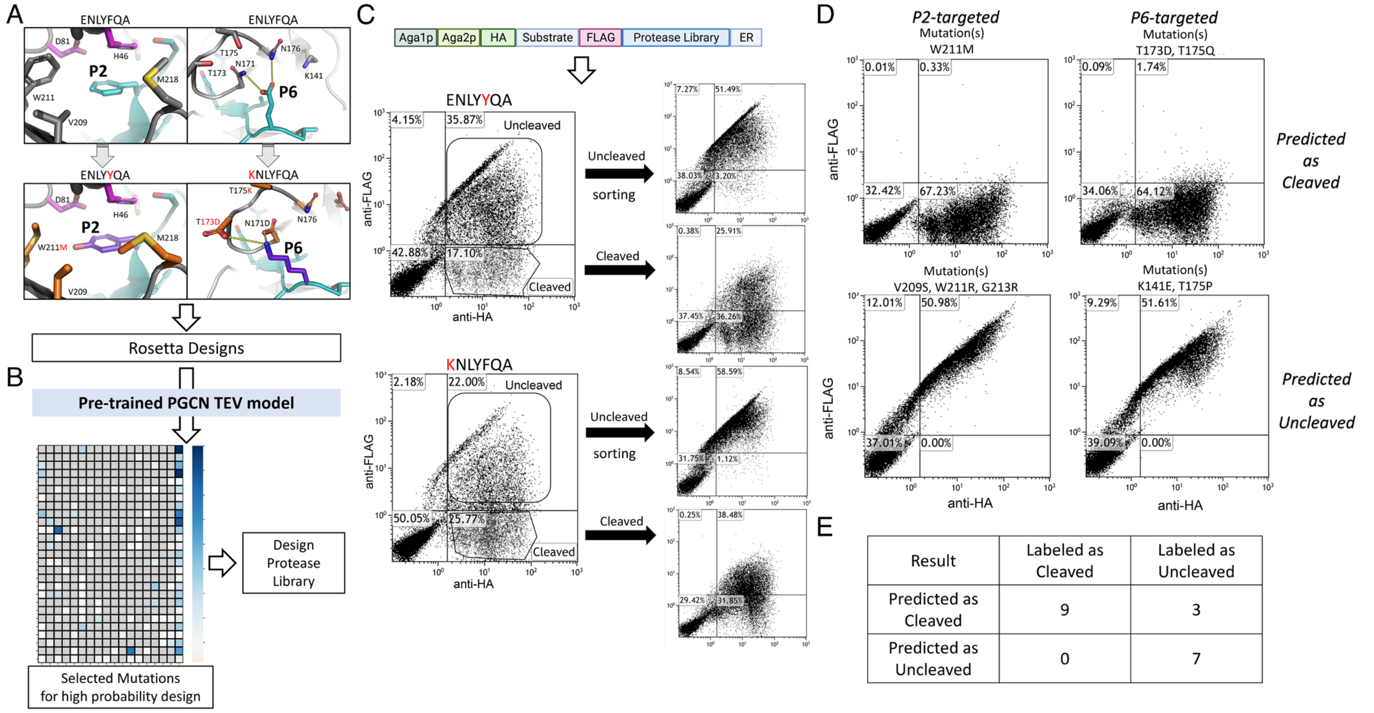


Fig. 5. PGCN generalizability. (*A*) AUC for HCV cross-test among HCV WT, HCV A171T, HCV D183A, HCV Triple data using sequence+energy features. (*B*) AUC for leave-one-out tests of three TEV variants which have unique mutations among all TEV variants.



Result	Labeled as Cleaved	Labeled as Uncleaved
Predicted as Cleaved	9	3
Predicted as Uncleaved	0	7

Fig. 6. Pipeline for TEV protease design, including procedures of (*A*) computational design, (*B*) PGCN prediction, and (*C*) yeast-based assay testing using FACS, and (*D*) flow cytometry–based analysis of individual colonies. (*A*) S2 and S6 Pockets corresponding to altered substrate residues were inferred from the crystal structure of a TEV protease–substrate complex, and redesigned using Rosetta. (*B*) Rosetta-generated designs were evaluated using a pretrained PGCN model and mutations enriched in high-scoring designs were identified and used to generate combinatorial protease libraries. (*C*) These libraries were screened using the YESS assay with P6 and P2 variant substrates, and pools of cells corresponding to cleaved and uncleaved protease:substrate variant pairs were isolated using FACS. (*D*) Individual colonies isolated from cleaved and uncleaved pools were tested using flow cytometry. 2-D scatter plots for positive P2-targeted (*Top-Left*) and P6-targeted (*Top-Right*) designs and negative designs (P2: *Bottom-Left*, P6: *Bottom-Right*) are shown. (*E*) Comparison table of PGCN prediction and experimental results on 19 clonally tested designs.

probability. Based on visual inspection, we selected 96 energetically favorable designs targeting P6 and 18 energetically favorable designs targeting P2 for further evaluation. To cost-effectively test these designs in experiments as well as generate negative (uncleaved) examples for classification, we identified amino acid substitutions that were enriched in the selected designs (*SI Appendix*, Fig. S5) and generated a combinatorial library sampling these protease substitutions. The library of selected designs was subjected to our previously reported version of the YESS (Yeast endosomal sequestration screening) assay (54) (Fig. 6*C*). Briefly, the assay detects cleavage by the presence/absence of HA and FLAG® tags on either side of the chosen substrate using anti-tag fluorescently labeled antibodies, and fluorescence-activated cell sorting (FACS) is used to isolate pools of "cleaved" and "uncleaved" cells. Several individual colonies from plating the cleaved and uncleaved pools of the P6- and P2-targeted TEV libraries were clonally validated (Fig. 6*D*). We selected 19 designs for clonal validation, 9 from the cleaved pool and 10 from the uncleaved pool. As shown in Fig. 6*E*, all 9 cleaved designs are correctly predicted by PGCN with high confidence (probability > 0.75) whereas 7 out of the 10 uncleaved designs are correctly predicted. Predictions are robust across 10 different Rosetta FastRelax-generated models of designs (*SI Appendix*, Fig. S6).

Discussion

Current computational methods for protease specificity prediction largely rely on statistical pattern detection in datasets of known substrates and nonsubstrates for learning specificity, and are therefore not generalizable for use in protease design, and do not provide insights into the underlying biophysical bases of substrate–protease molecular recognition. We developed PGCN to include residue-level energies as features and decompose the Rosetta-computed energies into the node and edge features. PGCN shows high accuracy in classification tasks for HCV and TEV protease variants. A PGCN model utilizing sequence and energy-based features and trained on experimental data for TEV protease was used to evaluate and select designed TEV protease variants for cleaving noncanonical substrates. Evaluated protease diversity included residue positions and/or amino acid substitutions not present in the training dataset. Experimental validation showed that PGCN scoring led to high-accuracy selection of cleaved designs. Thus, PGCN was able to be generalized and our studies show proof-of-principle for the challenging task of protease design using ML-enabled structure-aware computational modeling.

As energies implicitly encode distance information, we also compared the contribution of energy information and distance information in models along with sequence features. The models incorporating additional distance information did not exhibit greatly improved accuracy of prediction (*SI Appendix*, Table S5). However, we found that PGCN can achieve comparable accuracy to sequence+energy models when utilizing sequence+distance features. This can be attributed to the primary role of sequence features in the model's ability to discriminate (Figs. 3 and 4). Thus, for specificity prediction where sequence information is available for training, distance features perform equivalent to energy. However, energy features are required for design tasks where sequence and distance features have limited utility as interresidue distance does not change during design and sequence at all positions is not known a priori.

The physically based graphical structure of PGCN enables overcoming to some extent the black-box nature of ML methods as applied to protein modeling. Sensitivity analysis of nodes/

edges that are mapped to a residue (a node) or the link between two residues (an edge) helps identify residues and pairwise residues that are most influential for selectivity, suggesting that the latent space of the PGCN model is rich and has the potential to further our understanding of molecular bases of protease selectivity. For example, identified important intermolecular edges influence the relative placement of the scissile bond with respect to the catalytic base, a key geometric requirement for the acylenzyme formation step protease hydrolysis. Furthermore, PGCN identified subgraphs or networks of interacting residues that are key for specificity in recognizing a single peptide residue. Current implementation of PGCN gives as output probabilities for a binary classification, thus the results are qualitative. A quantitative or semiquantitative prediction (i.e., ranking with respect to a known variant) of the catalytic parameters of the enzyme may become possible with protease activity datasets of sufficiently large size generated using experiments in which proteolysis is measured as a function of time. Due to its physical grounding, PGCN may be generally applicable in the specificity prediction and design of other proteins that bind peptide substrates, especially when only small experimental datasets (a few hundred or thousand peptide substrates) are available (55). As PGCN only needs the structure of one protein:peptide complex as input, the availability of predicted structures of protein–peptide complexes opens the door to tackling specificity modeling on a large scale, provided accurate complex structures can be built (56, 57). These efforts are currently ongoing in our laboratory.

Materials and Methods

Protease Specificity Data.***HCV protease.*** The experimental dataset for HCV protease was obtained in previous yeast surface display experiments conducted in our lab using yeast surface display coupled with deep sequencing (42). This method allowed for rapid sampling of many candidate P6-P2 sequences with a given protease (WT HCV or one of three variants shown in *SI Appendix*, Fig. S7*A*) and determining each of those substrates to be either cleaved or not cleaved. Specifically, it sampled 7,342 substrates for wild type (1,932 of which were confirmed as cleaved), 13,208 substrates for A171T variant (3,644 of which were confirmed as cleaved), 11,864 substrates for D183A variant (4,350 of which were confirmed as cleaved) and 6,838 substrates for R170K/A171T/D183A variant (3,135 of which were confirmed as cleaved). All data points for HCV proteases are combined into a new single dataset (named *HCV Combined*). For each protease variant, substrate identity within each pool is less than 80% to avoid overfitting to the input because of data redundancy with a number of samples for each variant shown in Table 1. Experiments enable the sampling of P6-P2 substrates across all amino acids (*SI Appendix*, Fig. S7 *B*–*F*). As the datasets for WT, A171T, and D183A are imbalanced (the number of uncleaved samples is at least twice of the cleaved samples), metrics besides accuracy, such as precision, recall, F1 score, etc., were also considered during the analysis.

TEV protease. The experimental dataset for TEV protease was obtained from directed evolution and deep sequencing profiling data of the phage substrate display collected in Packer et al. (45). Briefly, substrates of nine designed TEV variants (*SI Appendix*, Table S1) were profiled based on single-mutation substrate libraries, and each variant has between 4,000 and 6,000 substrates, including 2,000 to 3,000 cleaved sequences. L2F variant and WT variant were also profiled based on a triple-mutation substrate library (three substrate positions are randomized simultaneously). Up to ~55,000 cleaved sequences and ~80,000 uncleaved sequences for the L2F variant were obtained, while ~30,000 cleaved sequences and ~40,000 uncleaved sequences were obtained for the WT. However, due to the noise in the high-throughput sequencing assays, we observed considerable overlap between the cleaved and uncleaved pools for all variants. Therefore, we developed a data processing pipeline (*SI Appendix*, Fig. S8) to preprocess raw deep sequencing data to identify nonoverlapping substrate sets (*SI Appendix, TEV Deep Sequencing Data Processing*). Sequence data were filtered based on empirical threshold to minimize overlap between cleaved and uncleaved populations (*SI Appendix*, Figs. S9–S12), resulting in a final tally of 5,425 high-confidence substrates among ten TEV protease types (including the wild type or variants), 48% of which are determined to be cleaved.

Observed substitutions in the engineered TEV variants are dispersed within the TEV protease structure (Fig. 7*A*). Additionally, there is variation in both P1 and P1′ positions of substrates shown in Fig. 7*B*, in contrast with HCV data. Other than Q/S, other P1/P1′ combinations, such as H/I, N/S, H/G, and Q/W, have been included. The top frequent amino acids at all positions in the cleaved population are the same as in the uncleaved population, and those from P6 to P1′ consist of ENLYFQ/S, known as the canonical sequence of TEV protease. Thus, the differences between cleaved and uncleaved sets are subtle.

In terms of the degree of protease variation in the TEV dataset, up to 23 different substitutions are present in comparison with the TEV WT protease (Fig. 7 *C*, *Left*). As depicted in the Sankey diagram (Fig. 7*C*), among 10 TEV protease variants (including wild type), R203Q, K215E, and I138T are present in one variant; T17S, H28L, T30A, N68D, F132L, T146S, D148P, S153N, F162S, S170A, N171D, N177M, V209M, W211I, M218F (cyan), and K229E are present in two variants; N176T (orange) is present in three variants; E107D, D127A, S135F (magenta) are present in four variants. L2F variant has the largest number of amino acid substitution sites among all variants and accounts for the majority of the dataset.

Protease–Substrate Complex Model Generation Using Rosetta. For HCV protease, models of protease–substrate complexes were based on crystal structures of inhibitor-bound (PDB code: **3M5N**) and peptide-bound inactive (PDB code: **3M5L**) variants of HCV protease (58). The structures were superimposed, and the peptide substrate of 3m5l was copied into the unmutated active site of 3m5n, replacing the inhibitor, and then the complex was minimized using the Rosetta FastRelax protocol (59) with coordinate constraints on all Cα atoms. A similar process was used with crystal structures of inhibitor-bound (PDB code: **1LVM**) and peptide-bound inactive (PDB code: **1LVB**) for TEV (52).

Mutant complexes were generated by substituting the appropriate residues in the WT complex and then minimizing all interfacial side chains (interfacial residues were defined as those with Cα-Cα distance <5.5 Å, or with distance <9 Å and Cα-Cβ vectors at an angle <75°) with an immobile backbone. We generated models for each possible substrate (3.2×10^5 possibilities within P5-P2) as part of a complex including the peptide from P7 to P4′ bound to each of the four HCV variants and the peptide from P7 to P3′ bound to each of the ten TEV variants in PyRosetta (23), using the Rosetta FastRelax protocol (59). Coordinate constraints minimized the movement of C_α for the peptide backbone, and the protease backbone was held fixed. Distance, angle, and dihedral constraints were applied to the catalytic triad (H72, D96, and S154 for HCV and H46, D81, and C151 for TEV) and the P1 and P1′ residues to enforce the catalytic geometry required for cleavage. A single FastRelax trajectory was performed for each protease–substrate complex.

Protein Graph Representation. The protein complex obtained from Rosetta modeling was encoded as a fully connected graph (i.e., there is an edge between every pair of nodes). The nodes of this graph are the amino acids of the substrate and the binding pocket of the protease, and the edges represent the pairwise residue interactions between nodes (Fig. 1*B*). Each node contains the features of a single residue, including a 1-hot encoder for amino acid type, a binary variable to indicate whether a residue is part of the substrate or the protease, and all 1-body Rosetta energy terms (60). These 1-body terms include statistical energy terms describing the likelihood of backbone (rama, omega) and side-chain torsions (fa_dun) compared to the values observed for high-resolution protein structures, and backbone-dependent probability of observing a given amino acid (p_aa_pp) and a reference energy (ref) that models the unfolded state. Each edge contains relational features between the pair of residue nodes it connects, including binary indications of whether the edge is between a substrate and a protease residue and whether the residues are covalently bonded (i.e., adjacent residues in sequence), and all Rosetta 2-body energy terms. Rosetta 2-body energies measure various types of interactions, including Van der Waals interactions, electrostatics, solvation, hydrogen bonding. List and brief explanations of node and edge features are listed in *SI Appendix*, Table S2, and further details of Rosetta energy functions used to compute these values can be found in ref. 60. Since residues that are far away from the substrate are less likely to

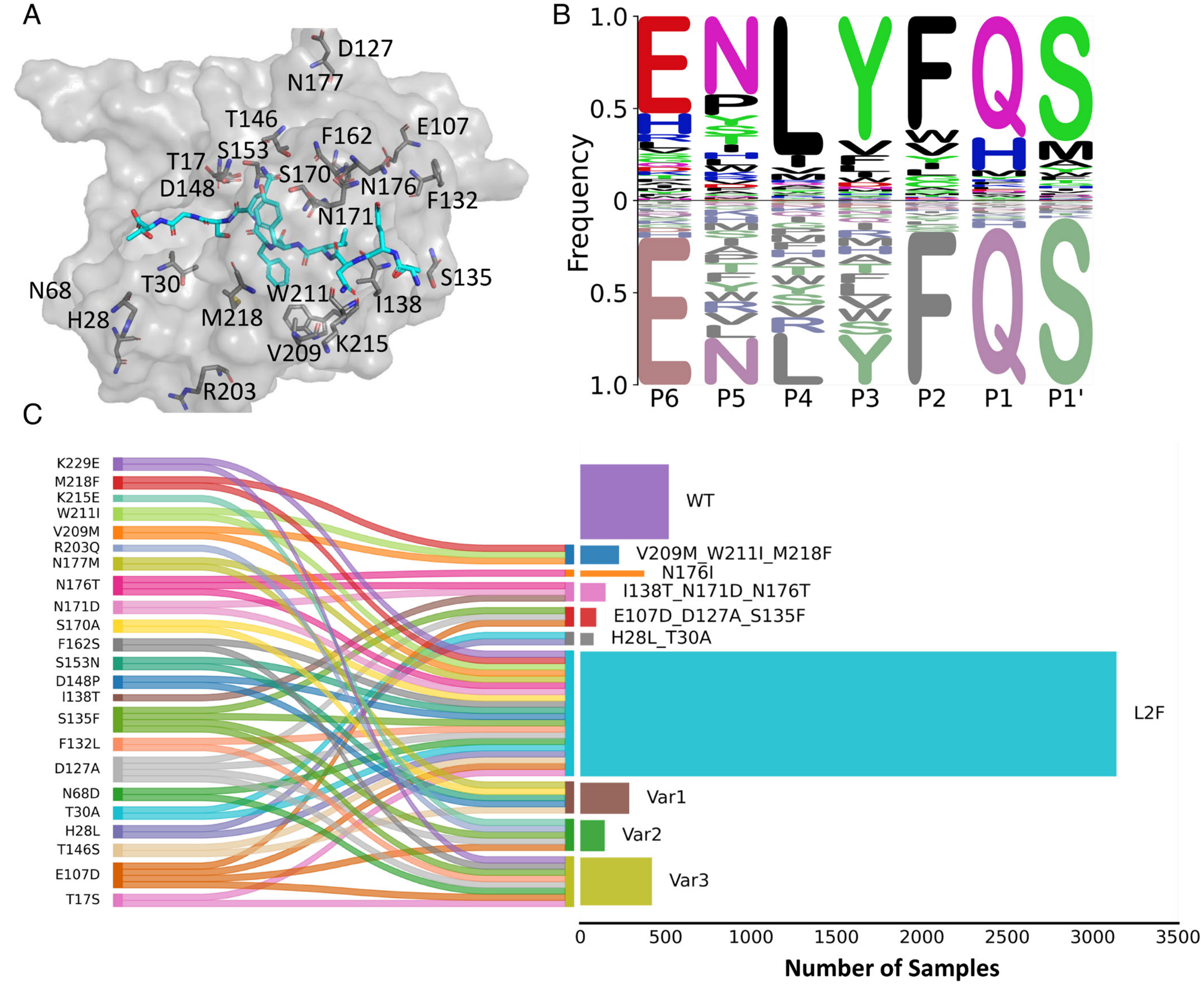


Fig. 7. TEV input data variety. (*A*) TEV protease mutation sites (gray) are shown all together in the TEV WT protease structure, located around the substrate (cyan). (*B*) Substrate sequence logos for TEV input data from P6-P1′ positions. *X* axis splits data into cleaved (above *x* axis) and uncleaved (below *x* axis) populations, where the higher frequency of amino acids appeared in the cleaved population, the higher it is located in the logo plot; the lower if considering the uncleaved population. (*C*) Sankey diagram of mutation sites for TEV variants, combined with a horizontal barplot, showing the number of samples for different TEV variants (along *x* axis). Variants are named based on their mutation sites except for the following four variants: Var1 (T146S_D148P_S153N_S170A_N177M), Var2 (E107D_D127A_S135F_R203Q_K215E), Var3 (T17S_N68D_E107D_D127A_F132L_S135F_F162S_K229E), and L2F. See Dataset S5 for complete mutation sites in the table.

influence specificity and more likely to introduce noise into the model, we only consider a subgraph including the substrate and interfacial residues that are within 10 Angstroms from the substrate (see *SI Appendix*, Tables S3 and S4 for node indices for HCV/TEV input graphs), selected based on proximity and sidechain orientation toward the substrate.

The node features for HCV protease are encoded into a $N \times F$ node feature matrix X, where $N = 34$ is the number of nodes, and $F = 20 + 8$ is the number of node features if combining both sequence and energy features for HCV data. The edge features are encoded as a M-length list of $N \times N$ edge feature matrices, where $M = 8$ is the number of edge features.

For TEV data, the number of nodes is $N = 47$, while node and edge feature encodings are the same as HCV data. To better introduce the architecture of our model, we use HCV data under combined (Sequence + Energy) feature encoding as the reference in the following paragraphs.

Transformation of Energy Features. Rosetta energies are negative for favorable interactions and positive for unfavorable interactions and penalties. However, for learning, it is preferable if all features are positive values when updating node weights during convolution. To accomplish this conversion, we performed a two-step transformation. First, we transformed each element $e_{i,j,q}$ in the edge tensor to a modified Boltzmann weight, $e'_{i,j,q}$ (Eq. **1**), where *i* and *j* are the nodes comprising the edge and *q* is the edge feature, numbered 1 to 8.

$$e'_{i,j,q} = \exp(-e_{i,j,q}) \cdot \qquad [1]$$

This converts negative Rosetta energies to larger positive values and positive Rosetta energies to small positive values, thereby assigning more appropriate weights to stabilize interactions. Second, since each edge of the graph is shared by two nodes, we followed Kipf (61) to further normalize each edge weight by degrees of two ends of nodes to match normalized Laplacian. The normalized Laplacian is written in the matrix format, which is similar to Eq. **2**,

$$\tilde{E}'_q = D_q^{-1/2} E'_q D_q^{-1/2}. \qquad [2]$$

Herein, E'_q represents qth edge feature matrix, and each element of the matrix E'_q is $e'_{i,j,q'}$ derived from Eq. **1**. D_q is the diagonal degree matrix of $E'_{q'}$ of which diagonal element is the sum of qth edge features for edges that are linked to a specific node.

After the transformation and normalization, the original edge data are transformed into a list of M matrices labeled by $\tilde{E}'_1, \tilde{E}'_2, \ldots, \tilde{E}'_q$. Each matrix is of the size N by N, where $N = 34$ denotes the number of nodes. Next, PGCN transforms the $M = 8$ edge matrices to a weighted sum $N \times N$ matrix E. In other words, $E = \sum_{i=1}^{q} \square\ w_i \tilde{E}_i'$ where $w_1,\ w_2,\ \ldots,\ w_q$ are learned weights. Those weights are initialized within the range of [0, 1) and updated throughout the training process.

PGCN. After PGCN has transformed edge feature matrices into the matrix E with the size of N by N, each node from the node feature matrix X with the size of N by N is able to learn from all other nodes based on weights assigned by edge features. E matrix is used as the weight matrix for convolution. Then the weight matrix E is fed to a GCN (61) layer (Fig. 1*C*), which in fact results in that each edge feature matrix $\tilde{E}_i'$ performs matrix multiplication with node feature matrix X, and independent multiplication outputs are concatenated and linearly transformed to N by F dimension, which aligns with the idea of multihead attention (62).

To reduce computational complexity, we used two GCN layers, and the number of hidden nodes of the second layer is the same as the first GCN layer. Therefore, the output of two GCN layers is

$$H = \sigma(E\sigma(EXW_1)W_2), \qquad [3]$$

where σ is the nonlinear activation function (63), here we use $\sigma = ReLU(x) = max(x, 0)$ element-wise applied on each output of the GCN layer, W_1, W_2 are learned weight matrices both with the size of $F \times C$ for hidden layers with $C = 20$ feature maps. Each GCN layer is followed by a BatchNorm layer (64), which aims at avoiding slow convergence.

Next, PGCN drops out a proportion of hidden nodes over nodes to avoid the overfitting problem (65). Finally, PGCN flattens the output matrix $H \in R^{N \times C}$ from the dropout layer into a one-dimensional vector $H' \in R^{1 \times NH}$. Then we transform H' to the expected dimension by applying a linear layer,

$$Y = H'W_3 + B, \qquad [4]$$

where $Y \in R^{1 \times 1}$ is the output, $W_3 \in R^{NH \times 1}$ is the learned vector, $B \in R^{1 \times 1}$ is the learned bias. Herein, we could apply the sigmoid as the activation function to calculate probabilities of being cleaved/uncleaved.

We followed a previously described approach to initialize weights (25). For training, PGCN does backpropagation to update all parameters mentioned above and the set of tuning hyperparameters: batch size, learning rate, dropout rate, and weight decay coefficient. The weight decay coefficient is a part of the L2 regularization term that multiplies the sum of learned weights for the antioverfitting procedure. Learned parameters are updated through epochs. The trained PGCN model is used for testing, in which test data pass through each layer of the PGCN model but skip the dropout process. The loss function is cross-entropy loss. PGCN is trained on training datasets using PyTorch, and tested on validation sets for hyperparameter tuning. PGCN performances are reported on test datasets.

Comparison with Baseline ML methods. We compared the prediction performance obtained for the PGCN with that obtained from five other ML methods. We used the Scikit-learn 0.20.1 (66) to implement logistic regression (lg), random forest (rf), decision tree (dt), SVM classification, and Tensorflow 1.13.1 (67) for ANN. The ANN model in this experiment is a one-layer fully connected neural network with 1,024 hidden nodes and allows a dropout rate between 0.1 to 0.9. To better compare performances between PGCN and other ML models, energy features are formed by residue-level energies, including single residue energies and pairwise energies flattened into a 1-dimensional vector, together with the protease type identifier (encoded in 10-dimensional one hot encoder) and sequence one hot encoder. In this case, PGCN and other ML models have the same feature encoding on energies to ensure proper comparison. We also followed the same rule as PGCN of splitting data into training, validation, and test datasets.

Node/Edge Importance. We would like to derive biological insights about important residues or relationships between pairs of residues that contribute to the discrimination of cleaved and uncleaved substrates in PGCN. Since the test graphs for a PGCN model all come from the same protein family, they share the same graph structure. Therefore, we can discover the important residue (or a pair of residues) of the same node (or edge) across all test graphs. To efficiently determine the importance of a specific node (or edge), we perturbed values of each node feature for the same node (or edge feature for the same edge) across all test samples and inspected how much the test accuracy drops. By doing this, we avoid retraining the PGCN and the time complexity of perturbation is $O(1)$. By following the procedure above, we were able to evaluate the importance of all nodes and edges on test graphs. We further normalized the change of accuracy by (Original_Accuracy – Perturbed_Accuracy)/Original_Accuracy.

Combinatorial Library Preparation and Yeast-Surface Display. Oligonucleotides containing degenerate codons at positions K141, N171, T173, T175, N176 (P6-targeted residues) or V209, W211 and M218 (P2-targeted residues) of the TEV protease sequence were purchased from IDT Inc. Application of degenerate codons increased the theoretical library size from 96 to 432 in the P6 targeted library and from 18 to 48 in the P2 targeted library. The double-stranded insert DNA sequence (varying from 150-300 bp in length) coding for the combinatorial amino acid library was assembled through overlap assembly PCR followed by agarose gel extraction and column purification. The integrity of the assembled insert was verified by Sanger sequencing through Genewiz Inc.

An LY104 vector backbone (obtained from Y. Li, B. Iverson, and G. Georgiou at University of Texas at Austin) containing the gene sequence for TEV protease and P6-P2 region of the corresponding substrate was linearized through PCR with primers to create sufficient overlap with the insert sequence for effective homologous recombination. Electrocompetent EBY100 yeast cells were transformed with the DNA library through electroporation on a Micropulser™ electroporation apparatus at 1.8 kV. The cultures were grown overnight in selective dextrose casamino acid media. While the O.D of the cultures was less than 6, the cells were resuspended in selective galactose casamino acid induction media to induce display of the constructs on the surface of yeast.

Library FACS and cytometric analysis of individual designs. The induced combinatorial yeast-surface displayed libraries were tested separately using flow cytometry. 3×10^7 cells were pelleted at 2,250 rcf for 3 min and washed with 1 mL of PBS + 0.1%BSA at 3,000 rcf for 5 min. Washed cells were incubated with antibody stains (1:25 of anti-FLAG(DYKDDDDK)-PE(Phycoerythrin), 130-101-576, and 1:50 of anti-HA-fluorescein isothiocyanate (FITC) from Miltenyi, 130-120-722) for 1 h at 4 °C. Following incubation, the cells were washed with 1 mL PBS with 0.1% BSA, pelleted and then resuspended in 1 mL PBS. Samples were diluted to achieve a final concentration of 5×10^6 cells/mL following which, FITC (anti-HA) and PE (anti- FLAG) intensities were detected using the Beckman Coulter Gallios flow cytometer.

Gates for cell sorting of cleaved and uncleaved populations were defined using the MoFlo Astrios Cell Sorter. Cells from the two gates–cleaved and uncleaved underwent one round of sorting and were collected until a cell count of 10^6 was reached. DNA was collected from each population by using a Zymoprep Kit (Omega).

Data, Materials, and Software Availability. All related analytical results in this study are provided in supporting information. All scripts to generate data and pre-trained HCV/TEV models, all classification files for cleavage activities, and HCV/TEV input datasets for PGCN model selection are available at Zenodo (68). TEV designs for PGCN screening and for flow cytometry analysis are also available at Zenodo with https://doi.org/10.5281/zenodo.7653923 (68). All other source data are available on request from the authors. All codes and scripts to replicate PGCN results are available in https://doi.org/10.5281/zenodo.7653923 (68). See instructions in https://github.com/Nucleus2014/protease-gcnn-pytorch/ (69).

ACKNOWLEDGMENTS. We acknowledge NSF Molecular Foundations of Biotechnology grant CHE2226816 (to S.W. and S.D.K). J.H.L. was funded by the Rutgers NIH Biotechnology Training Program (T32 GM008339 and GM135141). S.Z.S. was funded by the Rosetta Commons Research Experience for Undergraduates.

Author affiliations: [a]Institute for Quantitative Biomedicine, Rutgers–The State University of New Jersey, Piscataway, NJ 08854; [b]Department of Chemistry and Chemical Biology, Rutgers–The State University of New Jersey, Piscataway, NJ 08854; [c]Verily Life Sciences, Boulder, CO 80302; and [d]Department of Statistics, Rutgers–The State University of New Jersey, Piscataway, NJ 08854

1. S. Tang *et al.*, Mechanism-based traps enable protease and hydrolase substrate discovery. *Nature* **602**, 701–707 (2022).
2. A. Erijman, Y. Aizner, J. M. Shifman, Multispecific recognition: Mechanism, evolution, and design. *Biochemistry* **50**, 602–611 (2011).
3. M. Vizovišek *et al.*, Protease specificity: Towards in vivo imaging applications and biomarker discovery. *Trends Biochem. Sci.* **43**, 829–844 (2018).
4. S. D. Mason, J. A. Joyce, Proteolytic networks in cancer. *Trends Cell Biol.* **21**, 228–237 (2011).
5. L. E. Sanman, M. Bogyo, Activity-based profiling of proteases. *Annu. Rev. Biochem.* **83**, 249–273 (2014), 10.1146/annurev-biochem-060713-035352.
6. S. Seth, J. Batra, S. Srinivasan, COVID-19: Targeting proteases in viral invasion and host immune response. *Front. Mol. Biosci.* **7**, 215 (2020).
7. B. Meyer *et al.*, Characterising proteolysis during SARS-CoV-2 infection identifies viral cleavage sites and cellular targets with therapeutic potential. *Nat. Commun.* **12**, 5553 (2021).
8. L. Zhang *et al.*, Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368**, 409–412 (2020).
9. B. Luan, T. Huynh, X. Cheng, G. Lan, H.-R. Wang, Targeting proteases for treating COVID-19. *J. Proteome Res.* **19**, 4316–4326 (2020).
10. B. E. Turk, L. L. Huang, E. T. Piro, L. C. Cantley, Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nat. Biotechnol.* **19**, 661–667 (2001).
11. A. Doucet, G. S. Butler, D. Rodríguez, A. Prudova, C. M. Overall, Metadegradomics: Toward in vivo quantitative degradomics of proteolytic post-translational modifications of the cancer proteome. *Mol. Cell. Proteomics* **7**, 1925–1951 (2008).
12. A. A. Agbowuro, W. M. Huston, A. B. Gamble, J. D. A. Tyndall, Proteases and protease inhibitors in infectious diseases. *Med. Res. Rev.* **38**, 1295–1331 (2018).
13. A. K. Patick, K. E. Potts, Protease inhibitors as antiviral agents. *Clin. Microbiol. Rev.* **11**, 614–627 (1998).
14. B. Turk, Targeting proteases: Successes, failures and future prospects. *Nat. Rev. Drug Discov.* **59**, 785–799 (2006).
15. J. Breidenbach *et al.*, Targeting the main protease of SARS-CoV-2: From the establishment of high throughput screening to the design of tailored inhibitors. *Angew. Chem. Int. Ed.* **60**, 10423–10429 (2021).
16. R. P. Dyer, G. A. Weiss, Making the cut with protease engineering. *Cell Chem. Biol.* **29**, 177–190 (2022).
17. M. Pogson, G. Georgiou, B. L. Iverson, Engineering next generation proteases. *Curr. Opin. Biotechnol.* **20**, 390–397 (2009).
18. J. L. Guerrero, P. S. Daugherty, M. A. O'Malley, Emerging technologies for protease engineering: New tools to clear out disease. *Biotechnol. Bioeng.* **114**, 33–38 (2017).
19. C. A. Denard *et al.*, YESS 2.0, a tunable platform for enzyme evolution, yields highly active TEV protease variants. *ACS Synth. Biol.* **10**, 63–71 (2021).
20. T. R. Blum *et al.*, Phage-assisted evolution of botulinum neurotoxin proteases with reprogrammed specificity. *Science* **371**, 803–810 (2021).
21. D. J. Matthews, J. A. Wells, Substrate phage: Selection of protease substrates by monovalent phage display. *Science* **260**, 1113–1117 (1993).
22. J. Zhou *et al.*, Deep profiling of protease substrate specificity enabled by dual random and scanned human proteome substrate phage libraries. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25464–25475 (2020).
23. L. Wang, K. Main, H. Wang, O. Julien, A. Dufour, Biochemical tools for tracking proteolysis. *J. Proteome Res.* **20**, 5264–5279 (2021).
24. L. E. Araya, I. V. Soni, J. A. Hardy, O. Julien, Deorphanizing Caspase-3 and Caspase-9 substrates in and out of apoptosis with deep substrate profiling. *ACS Chem. Biol.* **16**, 2280–2296 (2021).
25. B. J. Backes, J. L. Harris, F. Leonetti, C. S. Craik, J. A. Ellman, Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. *Nat. Biotechnol.* **182**, 187–193 (2000).
26. W. J. L. Wood, A. W. Patterson, H. Tsuruoka, R. K. Jain, J. A. Ellman, Substrate activity screening: A fragment-based method for the rapid identification of nonpeptidic protease inhibitors. *J. Am. Chem. Soc.* **127**, 15521–15527 (2005).
27. S. E. Boyd, R. N. Pike, G. B. Rudy, J. C. Whisstock, M. G. De La Banda, PoPS: A computational tool for modeling and predicting protease specificity. *J. Bioinform. Comput. Biol.* **3**, 551–585 (2005).
28. C. Backes, J. Kuentzer, H. P. Lenhof, N. Comtesse, E. Meese, GraBCas: A bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Res.* **33**, W208–W213 (2005).
29. M. Ayyash, H. Tamimi, Y. Ashhab, Developing a powerful in silico tool for the discovery of novel caspase-3 substrates: A preliminary screening of the human proteome. *BMC Bioinf.* **13**, 1–14 (2012).
30. J. Verspurten, K. Gevaert, W. Declercq, P. Vandenabeele, SitePredicting the cleavage of proteinase substrates. *Trends Biochem. Sci.* **34**, 319–323 (2009).
31. Z. Liu *et al.*, GPS-CCD: A novel computational program for the prediction of calpain cleavage sites. *PLoS One* **6**, e19001 (2011).
32. T. Lohmüller *et al.*, Toward computer-based cleavage site prediction of cysteine endopeptidases. *Biol. Chem.* **384**, 899–909 (2003).
33. F. Li *et al.*, DeepCleave: A deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics* **36**, 1057–1065 (2019).
34. B. R. Southey, A. Amare, T. A. Zimmerman, S. L. Rodriguez-Zas, J. V. Sweedler, NeuroPred: A tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. *Nucleic Acids Res.* **34**, W267–W272 (2006).
35. J. Song *et al.*, PROSPERous: High-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* **34**, 684–687 (2018).
36. J. Song *et al.*, PROSPER: An integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One* **7**, e50300 (2012).
37. J. Song *et al.*, iProt-Sub: A comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.* **20**, 638–658 (2019).
38. L. J. K. Wee, T. W. Tan, S. Ranganathan, CASVM: Web server for SVM-based prediction of caspase substrates cleavage sites. *Bioinformatics* **23**, 3241–3243 (2007).
39. J. Song *et al.*, Cascleave: Towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* **26**, 752–760 (2010).
40. M. Piippo, N. Lietzén, O. S. Nevalainen, J. Salmi, T. A. Nyman, Pripper: Prediction of caspase cleavage sites from whole proteomes. *BMC Bioinf.* **11**, 320 (2010).
41. F. Li *et al.*, Procleave: Predicting protease-specific substrate cleavage sites by combining sequence and structural information. *Genom. Proteom. Bioinf.* **18**, 52–64 (2020).
42. M. A. Pethe, A. B. Rubenstein, S. D. Khare, Data-driven supervised learning of a viral protease specificity landscape from deep sequencing and molecular simulations. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 168–176 (2019).
43. M. A. Pethe, A. B. Rubenstein, S. D. Khare, Large-scale structure-based prediction and identification of novel protease substrates using computational protein design. *J. Mol. Biol.* **429**, 220–236 (2017).
44. R. Feehan, M. W. Franklin, J. S. G. Slusky, Machine learning differentiates enzymatic and non-enzymatic metals in proteins. *Nat. Commun.* **12**, 3712 (2021).
45. M. S. Packer, H. A. Rees, D. R. Liu, Phage-assisted continuous evolution of proteases with altered substrate specificity. *Nat. Commun.* **81**, 956 (2017).
46. A. Leaver-Fay *et al.*, Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
47. G. Hoang, A. Bouzerdoum, S. Lam, "Learning pattern classification tasks with imbalanced data sets" in *Pattern Recognition,* P.-Y. Yin, Ed. (IntechOpen, 2009), 10.5772/7544.
48. C. Singh Tumrate *et al.*, Classification of imbalanced data: Review of methods and applications. *IOP Conf. Ser. Mater. Sci. Eng.* **1099**, 012077 (2021), 10.1088/1757-899X/1099/1/012077.
49. Y. Sun, A. K. C. Wong, M. S. Kamel, Classification of imbalanced data: A review. *Int. J. Pattern Recogn. Artif. Intell.* **23**, 687–719 (2011), 10.1142/S0218001409007326.
50. B. Mirza *et al.*, Machine learning and integrative analysis of biomedical big data. *Genes* **10**, 87 (2019).
51. J. D. A. Tyndall, T. Nall, D. P. Fairlie, Proteases universally recognize beta strands in their active sites. *Chem. Rev.* **105**, 973–999 (2005).
52. J. Phan *et al.*, Structural basis for the substrate specificity of Tobacco Etch Virus protease. *J. Biol. Chem.* **277**, 50564–50572 (2002).
53. R. B. Kapust, J. Tözsér, T. D. Copeland, D. S. Waugh, The P1′ specificity of tobacco etch virus protease. *Biochem. Biophys. Res. Commun.* **294**, 949–955 (2002).
54. Q. Li *et al.*, Profiling protease specificity: Combining yeast ER sequestration screening (YESS) with next generation sequencing. *ACS Chem. Biol.* **12**, 510–518 (2017).
55. A. Motmaen *et al.*, Peptide binding specificity prediction using fine-tuned protein structure prediction networks. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2216697120 (2022), 10.1101/2022.07.12.499365.
56. W. Gao, S. P. Mahajan, J. Sulam, J. J. Gray, Deep learning in protein structural modeling and design. *Patterns* **1**, 100142 (2020).
57. T. Tsaban *et al.*, Harnessing protein folding neural networks for peptide-protein docking. *Nat. Commun.* **13**, 176 (2022). 10.1101/2021.08.01.454656.
58. K. P. Romano, A. Ali, W. E. Royer, C. A. Schiffer, Drug resistance against HCV NS3/4A inhibitors is defined by the balance of substrate recognition versus inhibitor binding. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 20986–20991 (2010).
59. M. D. Tyka *et al.*, Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405**, 607–618 (2011).
60. R. F. Alford *et al.*, The rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
61. T. N. Kipf, M. Welling, "Semi-supervised classification with graph convolutional networks" in *5th International Conference Learn. Representation ICLR 2017–Conference Track Proceedings* (2016).
62. A. Vaswani, "Attention is all you need" in *31st Conference on Neural Information Processing Systems* (NIPS, Long Beach, CA, USA, 2017).
63. X. Glorot, A. Bordes, Y. Bengio, "Deep sparse rectifier neural networks" in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, PMLR (2011), vol. 15, pp. 315–323.
64. S. Ioffe, C. Szegedy, "Batch normalization accelerating deep network training by reducing internal covariate shift" in *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning* (2015), pp. 448–456.
65. N. Srivastava, G. Hinton, A. Krizhevsky, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929—1958 (2014).
66. F. Pedregosa *et al.*, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
67. M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning" in *OSDI'16: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation* (2016), pp. 265–283.
68. C. Lu *et al.*, Prediction and Design of Protease Specificity Using a Structure-Aware Graph Convolutional Network. Zenodo. https://doi.org/10.5281/zenodo.7653923. Deposited 16 February 2023.
69. C. Lu, protease-gcnn-pytorch. Github. https://github.com/Nucleus2014/protease-gcnn-pytorch/. Deposited 7 August 2020.