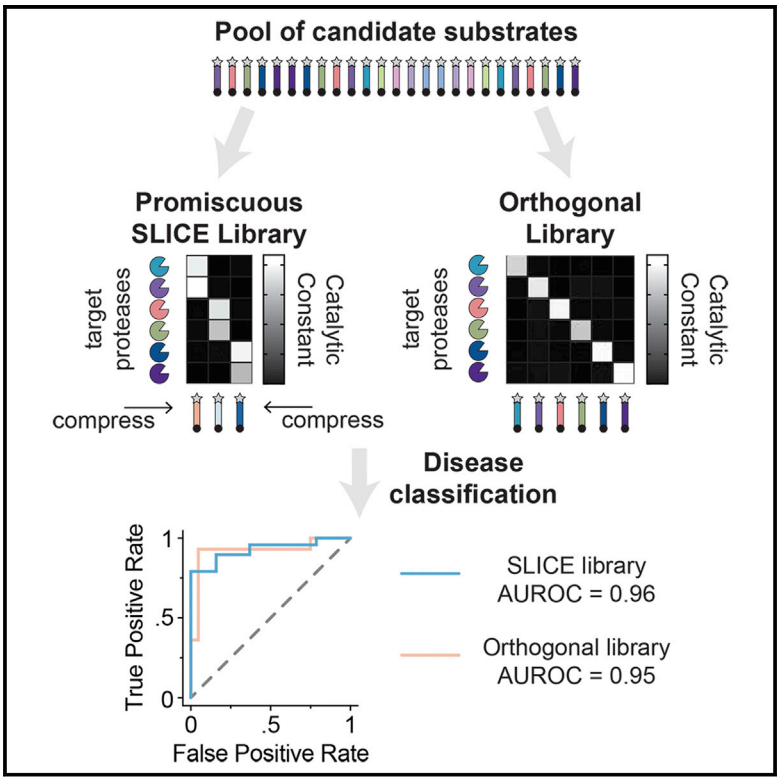
Cell Reports
Methods

Article

Embracing enzyme promiscuity with activity-based compressed biosensing

Graphical abstract



Highlights

- Substrate selection is a key step in designing protease-activatable drugs or sensors
- Promiscuous peptide substrates are typically discarded due to lack of specificity
- A computational method, SLICE, is developed to make use of promiscuous substrates
- Mixtures of 11 proteases are classified with high accuracy using SLICE substrates

Authors

Brandon Alexander Holt, Hong Seo Lim, Anirudh Sivakumar, ..., Haley Liakakos, Peng Qiu, Gabriel A. Kwong

Correspondence

gkwong@gatech.edu

In brief

Holt et al. present a computational method to select promiscuous peptide substrates—which are typically discarded—for the design of protease-activatable drugs or diagnostics. Using this method, they demonstrate that as few as two promiscuous substrates can accurately classify complex mixtures of 11 proteases in plasma.

Holt et al., 2023, Cell Reports Methods *3*, 100372
January 23, 2023
https://doi.org/10.1016/j.crmeth.2022.100372

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Embracing enzyme promiscuity with activity-based compressed biosensing

Brandon Alexander Holt,[1,8] Hong Seo Lim,[1,8] Anirudh Sivakumar,[1] Hathaichanok Phuengkham,[1] Melanie Su,[1] McKenzie Tuttle,[1] Yilin Xu,[1] Haley Liakakos,[1] Peng Qiu,[1,8] and Gabriel A. Kwong[1,2,3,4,5,6,7,8,9,*]

[1]Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech College of Engineering and Emory School of Medicine, Atlanta, GA 30332, USA
[2]Parker H. Petit Institute of Bioengineering and Bioscience, Atlanta, GA 30332, USA
[3]Institute for Electronics and Nanotechnology, Georgia Tech, Atlanta, GA 30332, USA
[4]Integrated Cancer Research Center, Georgia Tech, Atlanta, GA 30332, USA
[5]Georgia ImmunoEngineering Consortium, Georgia Tech and Emory University, Atlanta, GA 30332, USA
[6]Emory School of Medicine, Atlanta, GA 30332, USA
[7]Emory Winship Cancer Institute, Atlanta, GA 30322, USA
[8]These authors contributed equally
[9]Lead contact
*Correspondence: gkwong@gatech.edu
https://doi.org/10.1016/j.crmeth.2022.100372

MOTIVATION Proteases drive key biological processes, and their dysregulation underlies pathological conditions like cancer and inflammatory diseases. Protease-activatable sensors and therapies are under development, yet their design typically requires screening for peptide substrates specific to target proteases, which becomes increasingly difficult with multiple target proteases because many peptides can be promiscuously digested by multiple proteases. Drawing from a signal processing technique called compressed sensing, we developed a computational method for selecting libraries of promiscuous substrates that can classify distinct protease mixtures without relying on specific substrates. Using this method, we showed that a panel as small as two substrates could accurately differentiate plasma samples that contained different mixtures of 11 proteases.

SUMMARY

The development of protease-activatable drugs and diagnostics requires identifying substrates specific to individual proteases. However, this process becomes increasingly difficult as the number of target proteases increases because most substrates are promiscuously cleaved by multiple proteases. We introduce a method—substrate libraries for compressed sensing of enzymes (SLICE)—for selecting libraries of promiscuous substrates that classify protease mixtures (1) without deconvolution of compressed signals and (2) without highly specific substrates. SLICE ranks substrate libraries using a compression score (*C*), which quantifies substrate orthogonality and protease coverage. This metric is predictive of classification accuracy across 140 *in silico* (Pearson $r = 0.71$) and 55 *in vitro* libraries ($r = 0.55$). Using SLICE, we select a two-substrate library to classify 28 samples containing 11 enzymes in plasma (area under the receiver operating characteristic curve [AUROC] = 0.93). We envision that SLICE will enable the selection of libraries that capture information from hundreds of enzymes using fewer substrates for applications like activity-based sensors for imaging and diagnostics.

INTRODUCTION

Proteases are a major class of enzymes; more than 600 enzymes, comprising ~3% of the human genome,[1] are classified as proteases due to their ability to hydrolyze peptide bonds and degrade proteins (i.e., proteolysis). Protease activity is a driver of important biological processes, ranging from development and differentiation[2] to pathological conditions such as cancer, neurodegenerative disorders, and inflammatory diseases.[3] However, due to the irreversible nature of proteolysis, protease activity is tightly regulated via mechanisms such as inhibitory prodomains, cofactor binding, and protein inhibitors.[4] Given this degree of posttranslational regulation, quantifying protease activity, rather than transcriptomic or proteomic analyses, is

often required to understand the biological roles of proteases.[5] This has motivated the development of activity-based sensors that have been applied to early disease diagnostics—for example with imaging probes[6–8] and synthetic biomarkers in urine[9–11] and breath[12]—as well as therapies including protease inhibitors[13,14] and masked biologics.[15–17] The two primary compositions of activity-based sensors are (1) substrates that produce a signal upon proteolysis and (2) probes that bind active proteases.[4,18] For the former approach, a major bottleneck is substrate design, which involves screening for peptide substrates that are specific to the target protease (Figure 1, step 1). However, finding substrates with high specificity becomes increasingly difficult as the number of target enzymes increases because most proteases are characterized by promiscuous activity.[3,19]

To accelerate the process of designing specific substrates, methods to generate and screen libraries of peptide sequences have been developed, including positional scanning libraries,[20,21] peptide microarrays,[22,23] fluorogenic peptides,[24] other mixture-based peptide libraries,[25,26] and multiplex mass spectrometry assays.[27] These libraries are either degenerate or diversified at certain positions based on consensus cleavage motifs from the literature[28] or computational approaches to predict peptide sequences based on the structure of the active site of a target protease[29,30] (Figure 1, step 2). To generate potentially novel specific substrates, high-throughput evolution-based methods display and iteratively screen randomized peptide sequences on the surface of bacteria (e.g., CLiPS)[31,32] or bacteriophages (e.g., phage display)[33] and have been extended for screening endogenous protease activity[34] (Figure 1, steps 3–4). To further increase substrate specificity, approaches have been developed to broaden the chemical diversity of peptide libraries, such as via the introduction of non-natural amino acids[35,36] or cyclic peptide libraries.[37] In cases where protease-substrate kinetics are known, signal deconvolution algorithms can infer the activity levels of individual enzymes in a complex mixture[24,38]; this approach works well on controlled reactions involving recombinant enzymes. With these methods, libraries of up to 10–20 substrates, each of which have unique molecular barcodes, have been constructed to sense dysregulated protease activity for early detection of disease.[9,11,39] However, the current paradigm in substrate design methods is to favor specific substrates over promiscuous candidates.

Here, we embrace enzyme-substrate promiscuity by developing a substrate design method—substrate libraries for compressed sensing of enzymes (SLICE)—for selecting complementary promiscuous substrates to compile libraries of activity-based sensors that can classify distinct protease mixtures without specific substrates or signal deconvolution (Figure 1, step 5). Rather, SLICE, inspired by the signal processing technique compressed sensing,[40–42] evaluates different combinations of substrates to find the most complementary library that maximally senses all target proteases. We accomplish this by designing a compression score, *C*, which scores substrate libraries according to two features: (1) substrate orthogonality, which measures the uniqueness of protease-substrate kinetics, and (2) protease coverage, which measures the total fraction of target proteases sampled. In a simulated disease-detection challenge based on a melanoma gene microarray dataset,[43] *C* was predictive of classification accuracy across 140 *in silico* libraries (Pearson $r = 0.71$) and 55 *in vitro* libraries (Pearson $r = 0.55$). Further, we used SLICE to design a 2-substrate library ($C = 0.94$) that classified 28 complex samples containing one of two distinct 11-protease mixtures in the presence of murine plasma with high accuracy (area under the receiver operating characteristic curve [AUROC] = 0.93). Looking forward, producing smaller libraries will reduce the number of readouts, the overall cost, and the processing time, which is ideal for imaging- and activity-based diagnostics. We envision that SLICE will enable the selection of promiscuous substrate libraries that capture information from hundreds of enzymes using fewer activity-based sensors than is currently possible.

RESULTS

Computational pipeline for evaluating classification performance of simulated substrate libraries

Given an initial pool of candidate substrates, our goal was to develop a method for predicting which libraries of promiscuous substrates should be selected to accurately classify distinct mixtures of proteases. Therefore, we sought to create a simulation pipeline for evaluating the classification performance of substrate libraries with known protease-substrate cleavage kinetics (e.g., catalytic constants [k_{cat}]). To simulate a disease detection problem, we used a microarray gene expression dataset[43] containing data on 162 extracellular proteases in a murine melanoma model (Figure S1A). We calculated average protease gene expression profiles for healthy (day 1) and disease (day 7) samples and then generated Gaussian-distributed populations of 200 simulated samples from healthy and disease conditions (i.e., 100 simulated samples for each condition) (Figure 2, part 1a). These populations were generated by adding up to two standard deviations of random noise to the average expression profiles, as this noise level is sufficient so that measuring a single protease would be insufficient to accurately classify healthy and disease, while measuring all proteases would lead to high accuracy. After performing principal-component analysis on the simulated samples, we observed that the first two principal components represent >80% of variance and provide a clear separation between the healthy and disease groups, meaning that the two groups can be easily classified using all protease measurements simultaneously. Given the challenge of sensing the activity of all proteases simultaneously, we use libraries of promiscuous substrates to measure combinations of proteases. To simulate promiscuous substrate libraries, we randomly generated k_{cat} for all pairwise combinations of proteases and substrates (Figure 2, part 1b). These values were generated by randomly selecting 10 to 30 proteases to cut a given substrate and then assigning Gaussian-distributed k_{cat} values (normalized between 0 and 1) to each of these proteases. We calculated a vector of product formation rates, V_{max}, for each substrate across all simulated samples by multiplying matrix P, which contains the gene expression levels of all 162 proteases for every simulated sample, by the vector k_{cat}, which contains k_{cat} for each protease with a given substrate (Figure 2, part 2). We used a random forest model for classifying the simulated

CellPress
OPEN ACCESS

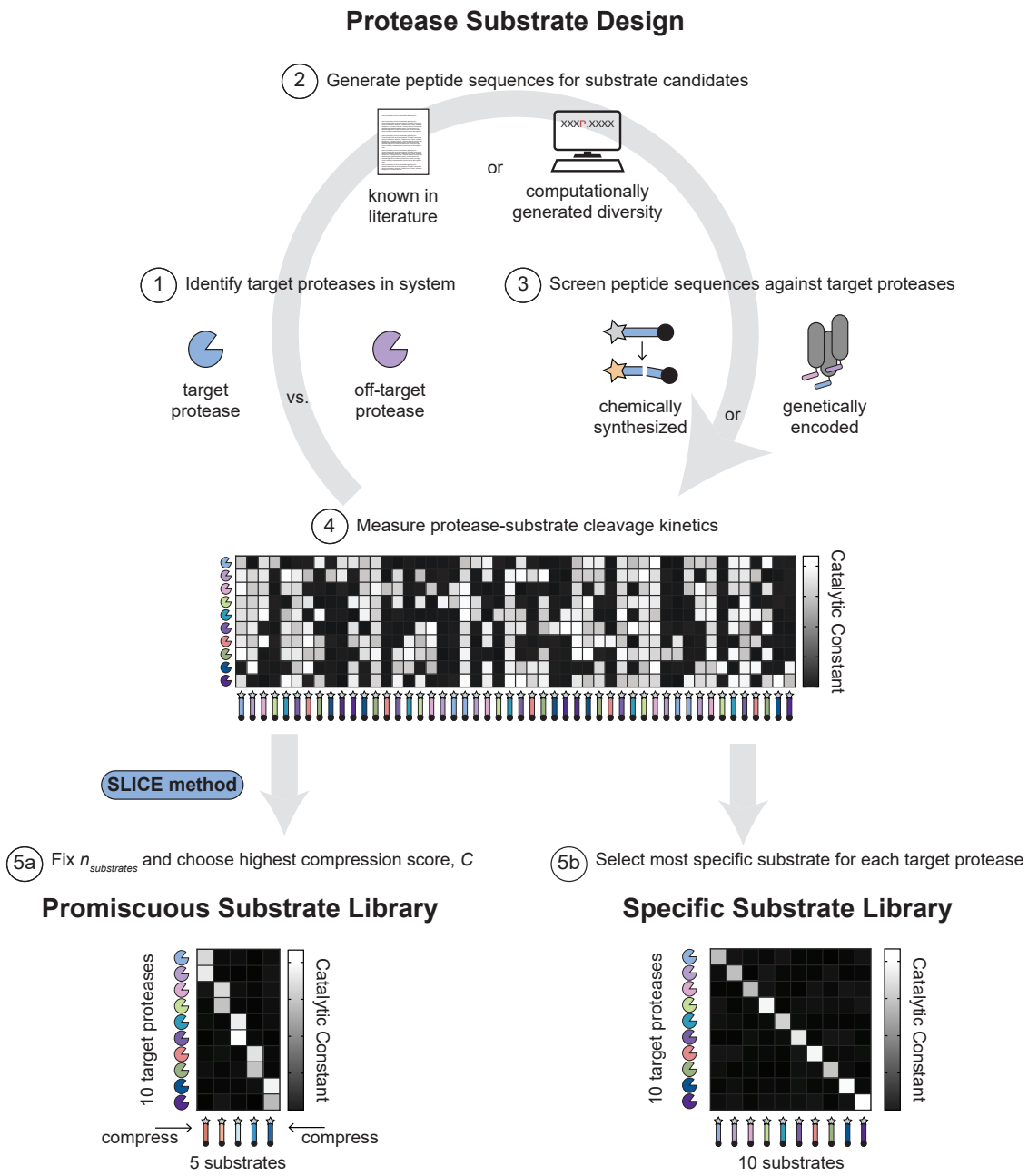


Figure 1. Conceptual overview of protease substrate design using the SLICE method

(1) Identify which proteases in the system being probed are considered target proteases (blue Pacman) and which are off-target proteases (purple Pacman).

(2) Generate candidate peptide sequences that can be used as substrates for target proteases. Peptide sequences can be acquired from the literature (paper icon) or computationally generated (computer icon). Computationally generated diversity includes degenerate libraries as well as predicted sequences derived from computational modeling software.

(3) Screen candidate peptide sequences against all protease targets via chemically synthesized activity-based sensors (e.g., fluorogenic probes, peptide microarrays, etc.) or genetically encoded libraries (e.g., phage display, bacteria display, etc.).

(4) Heatmap of cleavage kinetics, quantified by the catalytic constant, k_{cat}, for all protease-substrate pairs (rows = proteases, columns = substrates).

(5a) An example promiscuous substrate library that has fewer substrates ($n_{substrates} = 5$) than proteases ($n_{proteases} = 10$). The compression score, C, represents the score assigned to the library by the SLICE method, with 1 being the highest score and 0 the lowest.

(5b) An example specific substrate library that has the same number of substrates as proteases ($n_{substrates} = n_{proteases} = 10$).

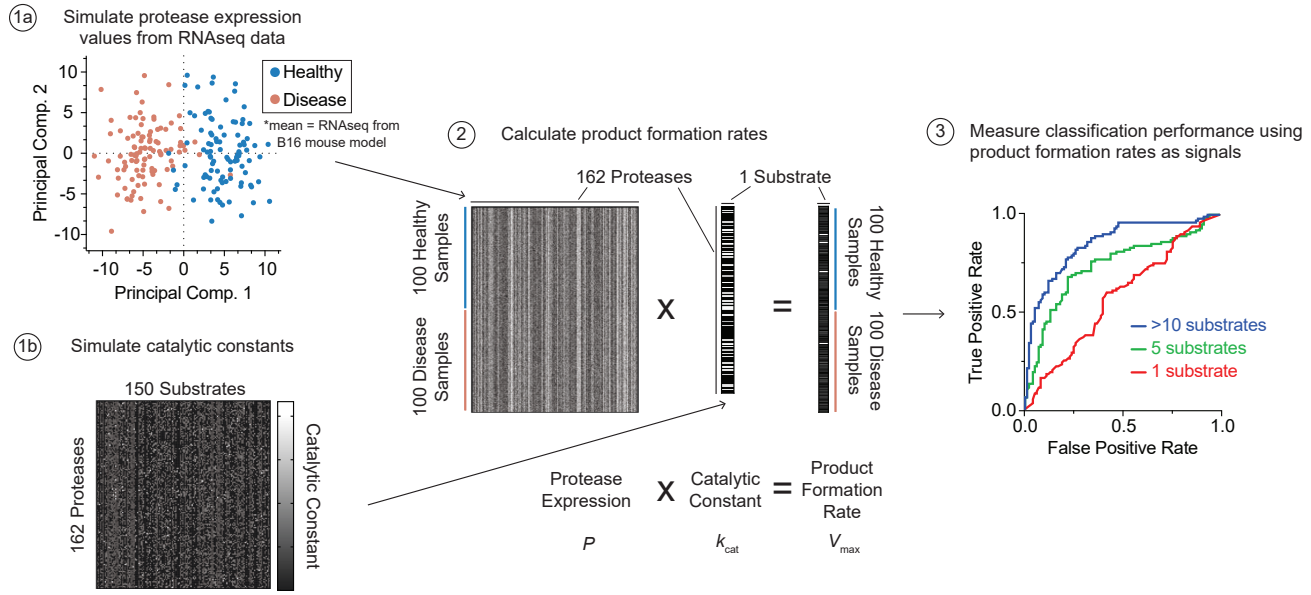


Figure 2. Computational pipeline for evaluating classification performance of simulated substrate libraries

(1a) Plot of first two principal components from principal-component analysis on microarray gene expression data of 162 protease genes in day 1 (healthy, blue) and 7 (disease, red) mouse tissue samples in a B16 melanoma model. To simulate, 100 samples and 100 disease samples are computationally generated as a Gaussian distribution from a single biological sample.

(1b) Heatmap of simulated catalytic constatnts, k_{cat}, for every pairwise combination between 162 proteases and 150 substrates (white = high, black = low).

(2) Visualization of how product formation rates, V_{max}, are calculated using protease concentrations, P, and k_{cat}. The result of this calculation is a product formation rate per substrate per sample.

(3) Receiver operating characteristic (ROC) curves as a measure of healthy versus disease classification performance using product formation rates as features of observations used to train a random forest model. Blue trace is an ROC curve when using signals (i.e., product formation rates) from 11 substrates (green trace = 5 substrates, red trace = 1 substrate).

samples (i.e., healthy versus disease) and used 5-fold cross-validation by aggregating predictions of an unseen fold (i.e., test set) based on the model trained by the other 4-folds (i.e., training set). To quantify classification performance, we calculated the AUROC resulting from applying the trained model to the test set (Figures 2, part 3, and S1B). We observed a clear trend that increasing the number of substrates in a library resulted in increased classification power. With this pipeline, we can evaluate the classification performance for a substrate library with known k_{cat} in a simulated disease detection problem as a proxy for true classification power.

A compression score for promiscuous substrate library selection

Since a promiscuous substrate can be cleaved by multiple proteases, the net signal of a substrate represents some weighted combination of product formation rates from multiple proteases. Therefore, measuring the signal of a promiscuous substrate compresses the product formation rates (i.e., activity) from multiple proteases into one feature. We sought to design a compression score, C, that selects for the most complementary set of promiscuous substrates that maximally senses the proteases of interest. To account for this, C is a weighted sum of two metrics—substrate orthogonality, $S_{orth.}$, and protease coverage, $P_{cov.}$ (Figure 3A; Equation 1).

$$C = \omega S_{orth.} + (1 - \omega) P_{cov.} \quad \text{(Equation 1)}$$

C operates on a 2D matrix of kinetic constants (e.g., k_{cat}, product formation rates, etc.) for all pairwise combinations of protease (rows) and substrate (columns); the score outputs one value ranging between 0 and 1, with 1 being the optimal score (Figure 3B). $S_{orth.}$, which is the cosine distance metric (Figure S2), quantifies the orthogonality of the columns, or how unique each of the substrates are from one another in the protease space. For example, substrate libraries with high $S_{orth.}$ will have columns that are different from one another, whereas the columns will be more similar in libraries with low $S_{orth.}$ (Figure 3B, y axis). Conversely, $P_{cov.}$ quantifies how many rows have at least one element with a high value, or how many proteases are collectively sampled by a library. For example, substrate libraries with high $P_{cov.}$ will have a high value in all rows, whereas libraries with low $P_{cov.}$ will include rows of only low values (Figure 3B, x axis). To verify that C is predictive of classification performance, we used the computational pipeline described in Figure 2 to evaluate the classification performance of 140 simulated substrate libraries. We found that C demonstrated a strong correlation with classification performance (Pearson's r = 0.71) with substrate libraries where $C < 0.2$ provided little useful information (i.e., 0.5 < AUROC < 0.6) and libraries where $C > 0.9$ demonstrated strong classification performance (i.e., AUROC > 0.85)

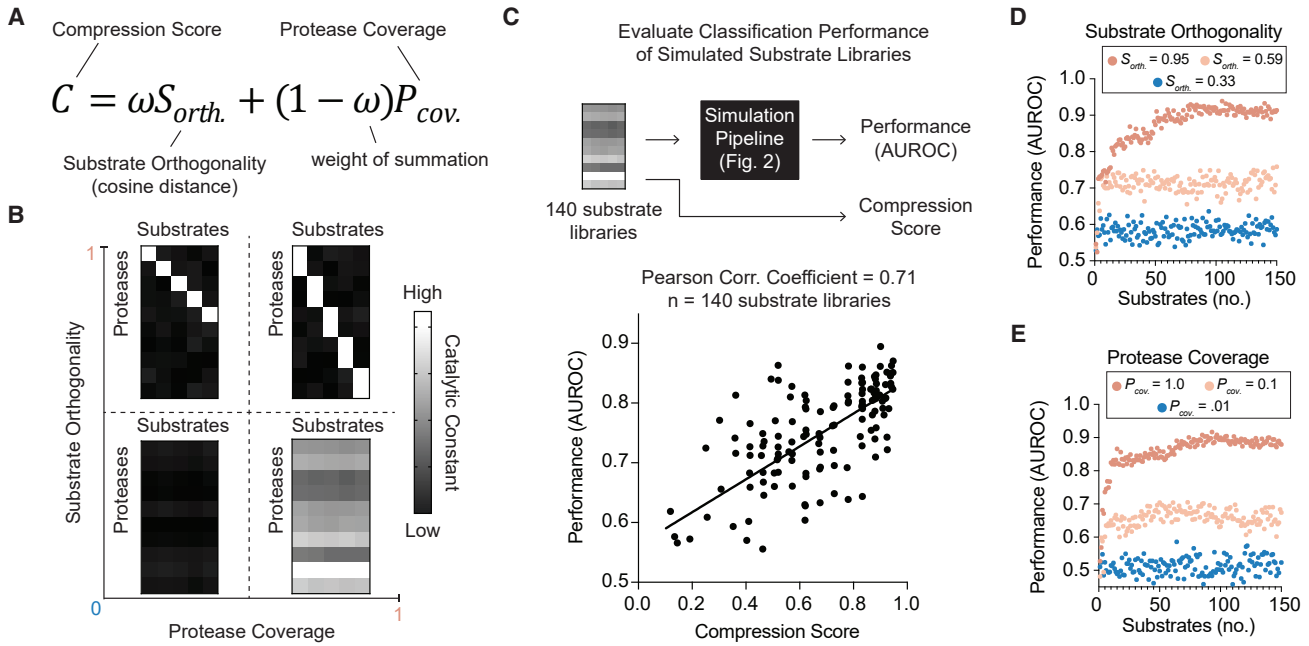


Figure 3. A compression score for promiscuous substrate selection

(A) Equation used to calculate the compression score, *C*. Substrate orthogonality, $S_{orth.}$, which is quantified by the cosine distance metric, and protease coverage, $P_{cov.}$, which quantifies the fraction of proteases that are sampled by a substrate library, are combined according to the weight of summation, ω. All variables range from 0 to 1.

(B) Schematic showing four example substrate libraries and their relative magnitude in $S_{orth.}$ (y axis) and $P_{cov.}$ (x axis) space. Each substrate library is represented with a heatmap of catalytic constats, k_{cat}, (white = high, black = low) for all protease (rows) and substrate (columns) combinations.

(C) (Top) Schematic showing pipeline for calculating *C* and classification performance for 140 simulated substrate libraries. (Bottom) Plot of correlation between *C* (x axis) and classification performance (AUROC, y axis). Black line is line of best fit. Each dot represents the performance of one substrate library averaged over 5 repeats.

(D and E) Plots showing classification performance (AUROC, y axis) versus substrate library size (number of substrates, x axis) for changing value of $S_{orth.}$ (D) and $P_{cov.}$ (E). Each dot represents the performance of one substrate library.

(Figure 3C). To verify that both $S_{orth.}$ and $P_{cov.}$ contribute to *C* independently, we independently fixed each variable and observed the change in classification performance across varying substrate library sizes (i.e., $1 < n_{substrates} < 150$). We found that increasing both $S_{orth.}$ (Figure 3D) and $P_{cov.}$ (Figure 3E) independently increased classification performance from 0.5–0.6 to >0.9 across all substrate library sizes tested. With *C*, we can rank-order and select the optimal set of promiscuous substrates where the kinetic constants toward the relevant protease targets are known.

Exhaustive scoring of substrate libraries *in vitro* with SLICE

To demonstrate the process of constructing a substrate library with SLICE experimentally, we selected a candidate pool of 11 substrates compiled from commercial products or published sequences[28,44,45] with known cleavage activity from matrix metalloproteases (MMPs), cathepsins, or complement proteases (Table S1). We focused on these protease classes as they have been shown to be dysregulated in pathologies like cancer[46] and organ transplant rejection[47] and have been targets of activity-based sensors.[6,9,11,39] We designed fluorogenic probes for these substrate sequences by flanking each with a fluorophore and quencher such that peptide cleavage would result in a measurable increase in fluorescence (Figure 4A, part 1). We performed cleavage assays for all 11 substrates with 11 proteases (121 unique protease-substrate pairs), including the target protease classes and other proteases (e.g., KLK2, thrombin, etc.), which were included to account for promiscuity of protease substrates. We then extracted the product formation rates (i.e., initial velocity) as representative kinetic parameters (Figures 4A, part 2, and S3). All substrates showed increasing signals with at least one protease (>2-fold increase in fluorescence after 60 min), indicating cleavage activity, while some protease-substrate pairs with negligible activity showed slightly decreased signals (<25%) due to photobleaching of uncleaved substrates. We observed that although the substrate sequences were known to target MMPs, cathepsins, and complement proteases, the off-target proteases used in these experiments also showed a propensity to cleave these sequences, which can be attributed to substrate promiscuity. To visualize the distribution of scores for these libraries, we exhaustively enumerated all libraries with sizes ranging from 2 to 10 and computed the $S_{orth.}$, $P_{cov.}$, and *C* scores for all those libraries (Figure 4B, part 1). We found that this candidate pool of substrates produced libraries high in $P_{cov.}$ but low in $S_{orth.}$ (Figure 4B, part 2). We found that the mean *C* score of 0.66 (n =

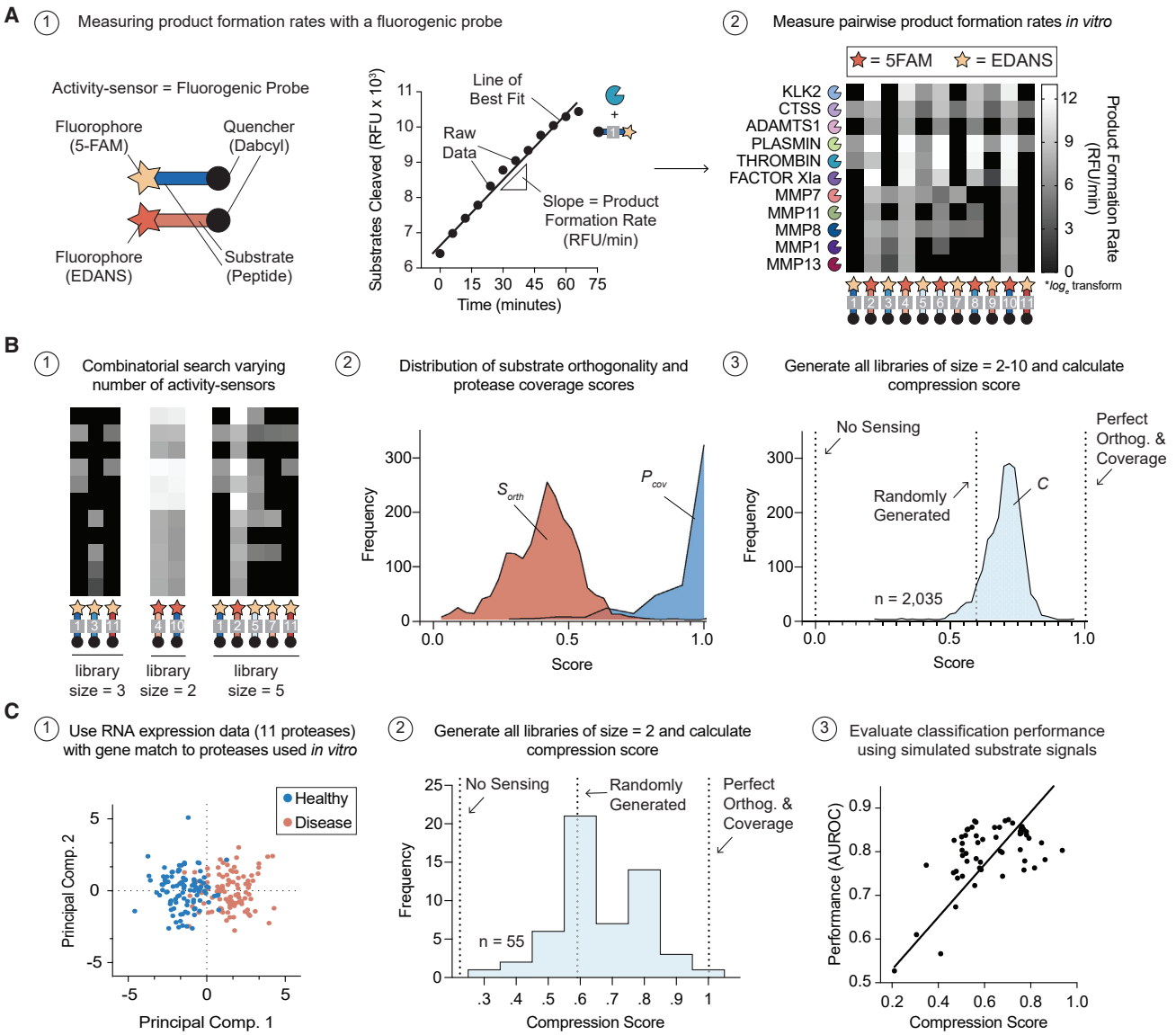


Figure 4. Exhaustive scoring of substrate libraries *in vitro* with SLICE

(A) (1, left) Schematic of activity sensor or fluorogenic probe. Activity sensor comprises a peptide substrate (blue and red bar) flanked with a fluorophore (yellow star = 5-FAM, red star = EDANS) and a quencher (black circle = Dabcyl). Upon cleavage, the fluorophore and quencher separate, which results in an increase in fluorescent signal. (1, right) Cleavage assay of thrombin and substrate-1 showing the increase in number of substrates cleaved (y axis) over time (x axis). Black dots are raw data. The slope (triangle) of the line of best fit (black line) is calculated as the product formation rate. Relative fluorescence unit (RFU)/min is used as RFU correlates with the number of substrates cleaved. (2) Heatmap showing all pairwise combinations of product formation rates as measured from independent cleavage assays. Proteases are in rows, and substrates are in columns. Data are natural log transformed.

(B) (1) Schematic showing that all unique combinations of substrates, with library sizes ranging from 2 to 10, are scored with SLICE. (2) Histogram showing the distribution of $S_{orth.}$ (red distribution) and $P_{cov.}$ (blue distribution) scores. (3) Histogram showing the distribution of the compression score, *C* (light blue distribution). Vertical dashed lines depict the score of various controls. "No sensing" depicts the score of a library where kinetic constant = 0 for all protease-substrate pairs. "Randomly generated" depicts the score of a library where kinetic constants are randomly generated. "Perfect orthog. & coverage' depicts the score of a library where all proteases are sampled, and each substrate has no overlapping kinetic constants.

(C) (1) Principal-component analysis of 11 proteases selected from 162 found in original B16 study. Proteases selected as either exact match or as member of same family as 11 proteases used in our study (A, part 2). Each dot represents one simulated sample (red = disease, blue = healthy). (2) Histogram showing the distribution of *C*s (light blue distribution) for all substrate libraries of size 2 (i.e., 2 substrates). (3) Plot showing correlation between *C* (x axis) and classification performance (y axis, AUROC). Black line shows line of best fit.

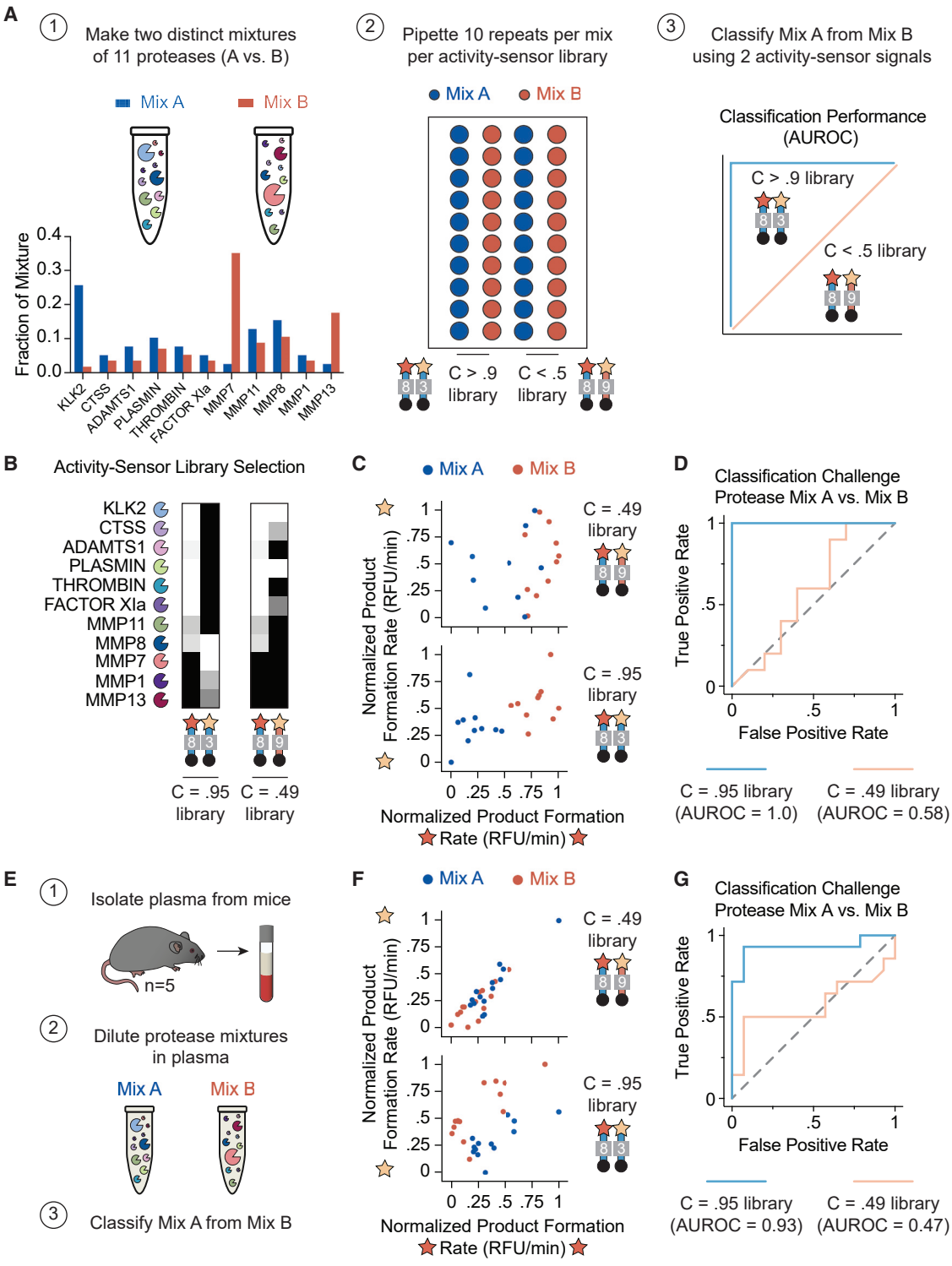


Figure 5. Experimental validation of substrate library design with SLICE

(A) Schematic of experimental workflow: (1) Two mixtures (A = blue, B = red) of 11 proteases are randomly generated. Each mixture is represented with a test tube containing 11 proteases (Pacman shape). Relative size of protease roughly represents the relative concentration. Actual relative concentrations are plotted in bar graph below (A = blue bars, B = red bars). (2) Schematic of experimental well plate containing samples of protease mixtures (1 circle = 1 well). Both mixtures are independently pipetted 10 times each (blue well = mix A, red well = mix B) to create a population with variance due to pipetting error. One library is introduced to all 20 samples (10 of mixture A, 10 of mixture B), and the product formation rates of both activity-based sensors in the library are measured. (3) Schematic graph (not

(legend continued on next page)

2,035 libraries) was higher than the benchmark score of randomly generated libraries (C = 0.6), meaning that real substrates tended to be more promiscuous than randomly generated substrates (Figure 4B, part 3). To validate that C is predictive of substrate library performance using empirically derived kinetic constants (i.e., product formation rates), we repeated the pipeline described in Figure 2 using the product formation rates found in Figure 4A. We trimmed down the list from 162 to 11 protease genes that were either from the same family or an exact match to the 11 proteases used in our experiments and simulated 100 healthy and 100 disease samples (Figure 4C, part 1). To fix library size, we calculated the distribution of Cs of all libraries comprising only two substrates (Figure 4C, part 2) and found that this distribution closely matched the score distribution for all library sizes (Figure 4B, part 3). We evaluated the classification performance for all 55 libraries of size 2 and found that C correlated with the AUROC (Figure 4C, part 3; Pearson's r = 0.55). Here, we demonstrated that constructing a library with SLICE involves (1) selecting a candidate pool of substrates that broadly recognize known protease targets, (2) measuring a kinetic parameter for each protease-substrate pair, and (3) identifying the optimal library/libraries by evaluating C.

Experimental demonstration and validation of substrate library design with SLICE

To validate the efficacy of a promiscuous substrate library designed with SLICE, we created an *in vitro* classification challenge for detecting dysregulated protease activity. To represent the two classification groups (i.e., protease mixture A versus protease mixture B; Figure 5A, part 1), we randomly generated two distinct mixtures of the same 11 target proteases from previous experiments (Figures 3 and 4). We incubated the library separately with 10 hand-pipetted repeats of both mixtures to introduce variance in the protease concentrations within the same group (Figure 5A, part 2). To evaluate the classification performance, we used the product formation rates of each substrate as the observations used to train a random forest model and calculated the AUROC for all test set samples in all 5-fold cross-validation iterations (Figure 5A, part 3). As a negative control, we tested a library with a low C (C < 0.5) to benchmark the performance of the SLICE library (C > 0.9) (Figure 5A). The kinetic parameter heatmap for the SLICE library (C = 0.95) showed that there is at least one substrate that can sense each protease, and the substrates only overlapped on one protease target (i.e., MMP8). Conversely, the negative control library (C = 0.49) does not sense 3 proteases (i.e., MMP1, MMP7, MMP13), and the substrates overlap on 4 protease targets (i.e., KLK2, CTSS, plasmin, factor XIa) (Figure 5B). These results validate that the scoring system (i.e., C) used in the SLICE method accurately represents $P_{cov.}$ and $S_{orth.}$ (Figure S4). We first assessed whether both cleavage signals of a two-substrate library could be monitored simultaneously using 5-FAM and EDANS fluorophores. Cleavage of 5-FAM- and EDANS-labeled substrates resulted in signal only in the expected fluorescence channel with no detectable crosstalk. Furthermore, the presence of an EDANS substrate had no significant effect on the cleavage fluorescence of a 5-FAM substrate, nor did a 5-FAM substrate affect cleavage signals of an EDANS substrate (Figure S5). Therefore, we proceeded to incubate each library with all 20 protease mixtures (i.e., 10 repeats of mixture A, 10 repeats of mixture B), and we plotted the results from each mixture in substrate space (i.e., x axis = product formation rate of 5-FAM substrate, y axis = EDANS substrate) (Figure S6). We observed that the SLICE library (C = 0.95) provided strong separation between mixture A and mixture B when compared with the negative control (C = 0.49) (Figure 5C). These results were confirmed by AUROC analysis, where the SLICE library (C = 0.95) classified all twenty mixtures with perfect accuracy (AUROC = 1) while compressing the dimensionality from 11 proteases to 2 substrates. By comparison, the negative control library (C = 0.49) showed worse classification performance (AUROC = 0.58), which held true across all temporal endpoints tested (Figures 5D and S7). Further, we found that the same substrate signal (i.e., substrate-8) that resulted in a negative feature importance score in the negative control (C = 0.49) library produced a positive feature importance score in the SLICE (C = 0.95) library (Figure S8). This demonstrates that while promiscuous substrates can be detrimental to certain libraries, pairing them with complementary substrates can improve the overall classification performance of the library. Finally, we tested whether classification performance is retained in a complex biological sample containing plasma isolated from mice (n = 5; Figure 5E). Plasma contains endogenous proteases (e.g., coagulation and complement proteases) and protease inhibitors that may contribute background noise and increase the challenge of classification.[48,49] The SLICE (C = 0.95) and negative control (C = 0.49) libraries were incubated with 28 protease mixtures consisting of 14 repeats of either mixture A or B mixed with plasma. Plasma from five biological replicates was used as opposed to a single mouse in order to further introduce variance across the samples. The SLICE library again classified the two mixtures with higher accuracy than the

real data) showing that the library with a high compression score, C, (C > 0.9) should have high classification performance (blue line), whereas the library with low C (C < 0.5) should have low classification performance (orange line).

(B) Heatmaps showing the product formation rates for the library with the highest C (C = 0.95 library) and the library with the lowest C (C = 0.49 library) (white = high product formation rate, black = low product formation rate).

(C) Plot of the resulting product formation rates for each activity sensor after incubation with protease mixtures (1 dot = 1 mixture; blue dot = mixture A, red dot = mixture B). The product formation rates from activity-based sensors using 5-FAM are plotted on the x axis, and product formation rates from EDANS are plotted on the y axis. The top plot shows the results when using the C = 0.49 library, and the bottom plot shows the results when using the C = 0.95 library. Rates were normalized from 0 to 1 for visualization.

(D) AUROC plot showing the results of classifying mixture A from mixture B when using the C = 0.95 library (blue trace) or the C = 0.49 library (orange trace).

(E) Schematic of workflow to test classification in citrated plasma.

(F) Plot of product formation rates for each activity sensor after incubation with protease mixture A or B in the presence of citrated plasma (plasma was isolated from 5 mice, and assay was performed with 2–3 technical replicates each, for total of n = 14).

(G) AUROC plot showing classification results in plasma.

negative control library (Figures 5F and 5G; AUROC = 0.93 versus 0.47). Here, we demonstrated that the SLICE method can select for substrate libraries and assign *C* that accurately predicts their classification performance when differentiating complex protease activity.

DISCUSSION

Here, we develop a method, SLICE, for compiling libraries of promiscuous substrates that sense protease activity for classification or diagnostic applications. This method involves (1) selecting a candidate pool of substrates that sense the target proteases, (2) measuring a kinetic parameter (e.g., k_{cat}, V_{max}, etc.) for each protease-substrate pair, and (3) identifying the optimal library of a fixed size by evaluating the *C*. The advantages of this method are that it enables the use of fewer promiscuous substrates (i.e., specific substrates not required) than the number of target proteases. By comparison, the current paradigm is to search for substrates that are specific to one protease and use approximately the same number of substrates as proteases. With these methods, all off-target protease activity is considered background noise, which is traditionally filtered out via chemical[50] or computational methods.[24,38] As the number of enzyme targets increases, it becomes increasingly difficult to maintain specificity across all substrates. Further, since each substrate requires a unique reporter, the number of simultaneous readouts becomes limited by cost (e.g., mass barcodes[9]) or physical restrictions (e.g., fluorescence[13]).

It is suggested that protease promiscuity bolsters fitness by (1) providing alternative evolutionary starting points and (2) increasing biological efficiency (i.e., multiple functions per enzyme).[19] We proposed that embracing protease promiscuity could leverage the ubiquity of substrates that recognize multiple targets. Serving as inspiration for the SLICE method, compressed sensing (CS) is a signal processing technique that utilizes measurements of a mixture of multiple target signals to recover information of individual signals.[51] A well-known application of CS is the single-pixel camera, which demonstrated the ability to efficiently handle high-dimensional datasets (e.g., hyperspectral imaging, video, etc.) and inspired the use of CS in magnetic resonance imaging[40] and imaging transcriptomics.[41,42] CS utilizes compressed signals, which are a composite of multiple different signals; this mirrors how the total number of cleaved copies of a promiscuous substrate results from a weighted combination of different proteases. However, a major difference is that our method does not require the deconvolution of compressed signals (i.e., cleaved substrate signals) as, unlike conventional CS, our approach aims to achieve high classification performance and not to reconstruct the original signal (i.e., individual protease activities). One consequence of this is while CS requires that the original signal is sparse, our approach may apply to cases where protease expression is not sparse. Future iterations of SLICE could incorporate (1) CS features (e.g., sparsity, incoherence) for substrate selection metrics (i.e., *C*) and (2) deconvolution of the compressed signals. However, we found that compressed signals are often sufficient for achieving high classification accuracy and would be preferable for applications such as point-of-care[52] or imaging diagnostics,[42] where fewer signals reduces the overall cost and processing time.

We envision that SLICE will be useful for applications where obtaining precise activity values per protease is less important than detecting systems-level changes, such as disease staging, classification, and diagnosis. Measuring protease activity at a systems level accounts for activation, deactivation, and inhibition by other proteases and proteinase inhibitors in native biological systems, which can occur in serum and in pathological settings like cancer and coagulation.[48,49,53,54] The ultimate application of SLICE would be a universal substrate library that is constructed by running all candidate substrates through a standardized test, which measures k_{cat} against all >600 recombinant human proteases. From this library, various sublibraries targeting different groups of proteases could be extracted on a per-application basis. For example, a diagnostic activity-sensor library could be extracted from the universal library by defining disease-specific target proteases ideally in pathologies that can be diagnosed using blood or plasma samples, such as coagulation disorders[55] or cancer.[56,57] While *in vitro* protease activity measurements may not fully account for the dynamic states of proteases *in vivo*,[58] future work could improve this by creating more robust *in vitro* tests that sample proteases under multiple states (e.g., redox, fluid dynamics, etc.) or developing *in vivo* tests that isolate the activity from individual proteases.

Further, other classes of enzymes also exhibit promiscuity,[59] which means that the design rules presented in this work can likely be extended to other promiscuous enzymes such as kinases or phosphatases and their activity-based sensors.[60,61] For example, candidate substrates would be mapped onto sensors that exhibit phosphorylation- or dephosphorylation-dependent changes in signals (e.g., fluorescence).[60] These sensors would be used to measure enzyme-substrate kinetics and generate an activity matrix, which could be processed using the SLICE method. In conclusion, we present SLICE as a method for embracing the use of promiscuous substrates for detecting changes in protease activity, as an alternative approach to the use of specific substrates. Given the ubiquity of promiscuous substrates and the motivation to sense biological activity, we anticipate that the ideas presented here will have broad applicability to the field of enzyme sensing at large.

Limitations of the study

This study focused on a set of 11 proteases to demonstrate selection of promiscuous substrates using the SLICE method. Extension to larger panels of proteases, especially those dysregulated in the context of disease, is warranted in future studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2022.100372.

ACKNOWLEDGMENTS

The authors thank Dr. Melissa Kemp (Georgia Tech and Emory) for their helpful discussions regarding the article. This work was funded by an NIH Director's New Innovator Award (award no. DP2HD091793); an R01 from the NCI (5R01CA237210); and a U01 from the NCI and NIBIB (1U01CA265711), as well as projects from the NSF (CCF1552784 and CCF2007029). B.A.H is supported by the NSF GRFP (grant no. DGE-1650044), the National Institutes of Health GT BioMAT Training Grant under award no. 5T32EB006343, and the Georgia Tech President's Fellowship. G.A.K. holds a Career Award at the Scientific Interface from the Burroughs Welcome Fund. P.Q. is an ISAC Marylou Ingram Scholar, a Carol Ann and David D. Flanagan Faculty Fellow, and a Wallace H. Coulter Distinguished Faculty Fellow. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

B.A.H., H.S.L., P.Q., and G.A.K. conceived the idea. B.A.H., H.S.L., A.S., H.P., P.Q., and G.A.K. designed experiments and interpreted results. B.A.H., A.S., H.P., M.S., M.T., Y.X., and H.L. carried out the experiments. B.A.H., H.S.L., and P.Q. wrote and ran the code. B.A.H., H.S.L., A.S., P.Q., and G.A.K. wrote the manuscript.

DECLARATION OF INTERESTS

G.A.K. is cofounder of Glympse Bio and Port Therapeutics. This study could affect his personal financial status. The terms of this arrangement have been reviewed and approved by Georgia Tech in accordance with its conflict-of-interest policies.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: January 12, 2022
Revised: October 11, 2022
Accepted: December 6, 2022
Published: December 30, 2022

REFERENCES

1. Bond, J.S. (2019). Proteases: history, discovery, and roles in health and disease. J. Biol. Chem. *294*, 1643–1651.
2. Barrett, A.J., Rawlings, N.D., and Woessner, J.F. (2004). Introduction. In Handbook of Proteolytic Enzymes, Second Edition, A.J. Barrett, N.D. Rawlings, and J.F. Woessner, eds. (Academic Press). pp. xxxiii–xxxv.
3. López-Otín, C., and Bond, J.S. (2008). Proteases: multifunctional enzymes in life and disease. J. Biol. Chem. *283*, 30433–30437.
4. Sanman, L.E., and Bogyo, M. (2014). Activity-based profiling of proteases. Annu. Rev. Biochem. *83*, 249–273.
5. Turk, B. (2006). Targeting proteases: successes, failures and future prospects. Nat. Rev. Drug Discov. *5*, 785–799.
6. Yim, J.J., Harmsen, S., Flisikowski, K., Flisikowska, T., Namkoong, H., Garland, M., van den Berg, N.S., Vilches-Moure, J.G., Schnieke, A., Saur, D., et al. (2021). A protease-activated, near-infrared fluorescent probe for early endoscopic detection of premalignant gastrointestinal lesions. Proc. Natl. Acad. Sci. USA *118*. e2008072118.
7. Edgington, L.E., Berger, A.B., Blum, G., Albrow, V.E., Paulick, M.G., Lineberry, N., and Bogyo, M. (2009). Noninvasive optical imaging of apoptosis by caspase-targeted activity-based probes. Nat. Med. *15*, 967–973.
8. Jiang, T., Olson, E.S., Nguyen, Q.T., Roy, M., Jennings, P.A., and Tsien, R.Y. (2004). Tumor imaging by means of proteolytic activation of cell-penetrating peptides. Proc. Natl. Acad. Sci. USA *101*, 17867–17872.
9. Kwong, G.A., von Maltzahn, G., Murugappan, G., Abudayyeh, O., Mo, S., Papayannopoulos, I.A., Sverdlov, D.Y., Liu, S.B., Warren, A.D., Popov, Y., et al. (2013). Mass-encoded synthetic biomarkers for multiplexed urinary monitoring of disease. Nat. Biotechnol. *31*, 63–70.
10. Kwong, G.A., Ghosh, S., Gamboa, L., Patriotis, C., Srivastava, S., and Bhatia, S.N. (2021). Synthetic biomarkers: a twenty-first century path to early cancer detection. Nat. Rev. Cancer *21*, 655–668.
11. Kirkpatrick, J.D., Warren, A.D., Soleimany, A.P., Westcott, P.M.K., Voog, J.C., Martin-Alonso, C., Fleming, H.E., Tammela, T., Jacks, T., and Bhatia, S.N. (2020). Urinary detection of lung cancer in mice via noninvasive pulmonary protease profiling. Sci. Transl. Med. *12*. eaaw0262.
12. Chan, L.W., Anahtar, M.N., Ong, T.-H., Hern, K.E., Kunz, R.R., and Bhatia, S.N. (2020). Engineering synthetic breath biomarkers for respiratory disease. Nat. Nanotechnol. *15*, 792–800.
13. Neefjes, J., and Dantuma, N.P. (2004). Fluorescent probes for proteolysis: tools for drug discovery. Nat. Rev. Drug Discov. *3*, 58–69.
14. Bachovchin, D.A., and Cravatt, B.F. (2012). The pharmacological landscape and therapeutic potential of serine hydrolases. Nat. Rev. Drug Discov. *11*, 52–68.
15. Desnoyers, L.R., Vasiljeva, O., Richardson, J.H., Yang, A., Menendez, E.E.M., Liang, T.W., Wong, C., Bessette, P.H., Kamath, K., Moore, S.J., et al. (2013). Tumor-specific activation of an EGFR-targeting probody enhances therapeutic index. Sci. Transl. Med. *5*. 207ra144.
16. Holt, B.A., Tuttle, M., Xu, Y., Su, M., Røise, J.J., Wang, X., Murthy, N., and Kwong, G.A. (2022). Dimensionless parameter predicts bacterial prodrug success. Mol. Syst. Biol. *18*. e10495.
17. Mansurov, A., Hosseinchi, P., Chang, K., Lauterbach, A.L., Gray, L.T., Alpar, A.T., Budina, E., Slezak, A.J., Kang, S., Cao, S., et al. (2022). Masking the immunotoxicity of interleukin-12 by fusing it with a domain of its receptor via a tumour-protease-cleavable linker. Nat. Biomed. Eng. *6*, 819–829.
18. Edgington, L.E., Verdoes, M., and Bogyo, M. (2011). Functional imaging of proteases: recent advances in the design and application of substrate-based and activity-based probes. Curr. Opin. Chem. Biol. *15*, 798–805.
19. Khersonsky, O., and Tawfik, D.S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu. Rev. Biochem. *79*, 471–505.
20. Schneider, E.L., and Craik, C.S. (2009). Positional scanning synthetic combinatorial libraries for substrate profiling. Methods Mol. Biol. *539*, 59–78.
21. Backes, B.J., Harris, J.L., Leonetti, F., Craik, C.S., and Ellman, J.A. (2000). Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. Nat. Biotechnol. *18*, 187–193.
22. Salisbury, C.M., Maly, D.J., and Ellman, J.A. (2002). Peptide microarrays for the determination of protease substrate specificity. J. Am. Chem. Soc. *124*, 14868–14870.
23. Szymczak, L.C., Kuo, H.-Y., and Mrksich, M. (2018). Peptide arrays: development and application. Anal. Chem. *90*, 266–282.
24. Miller, M.A., Barkal, L., Jeng, K., Herrlich, A., Moss, M., Griffith, L.G., and Lauffenburger, D.A. (2011). Proteolytic Activity Matrix Analysis (PrAMA) for simultaneous determination of multiple protease activities. Integr. Biol. *3*, 422–438.

25. Schilling, O., and Overall, C.M. (2008). Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. Nat. Biotechnol. *26*, 685–694.

26. Klein, T., Eckhard, U., Dufour, A., Solis, N., and Overall, C.M. (2018). Proteolytic cleavage—mechanisms, function, and "omic" approaches for a near-ubiquitous posttranslational modification. Chem. Rev. *118*, 1137–1168.

27. O'Donoghue, A.J., Eroy-Reveles, A.A., Knudsen, G.M., Ingram, J., Zhou, M., Statnekov, J.B., Greninger, A.L., Hostetter, D.R., Qu, G., Maltby, D.A., et al. (2012). Global identification of peptidase specificity by multiplex substrate profiling. Nat. Methods *9*, 1095–1100.

28. Rawlings, N.D., Barrett, A.J., and Bateman, A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. *40*, D343–D350.

29. Ochoa, R., Magnitov, M., Laskowski, R.A., Cossio, P., and Thornton, J.M. (2020). An automated protocol for modelling peptide substrates to proteases. BMC Bioinf. *21*, 586.

30. Boyd, S.E., Pike, R.N., Rudy, G.B., Whisstock, J.C., and Garcia de la Banda, M. (2005). PoPS: a computational tool for modeling and predicting protease specificity. J. Bioinform. Comput. Biol. *3*, 551–585.

31. Rice, J.J., Schohn, A., Bessette, P.H., Boulware, K.T., and Daugherty, P.S. (2006). Bacterial display using circularly permuted outer membrane protein OmpX yields high affinity peptide ligands. Protein Sci. *15*, 825–836.

32. Stach, N., Kalinska, M., Zdzalik, M., Kitel, R., Karim, A., Serwin, K., Rut, W., Larsen, K., Jabaiah, A., Firlej, M., et al. (2018). Unique substrate specificity of SplE serine protease from Staphylococcus aureus. Structure *26*, 572–579.e4.

33. Whitney, M., Crisp, J.L., Olson, E.S., Aguilera, T.A., Gross, L.A., Ellies, L.G., and Tsien, R.Y. (2010). Parallel in vivo and in vitro selection using phage display identifies protease-dependent tumor-targeting peptides. J. Biol. Chem. *285*, 22532–22541.

34. Pleiko, K., Põšnograjeva, K., Haugas, M., Paiste, P., Tobi, A., Kurm, K., Riekstina, U., and Teesalu, T. (2021). In vivo phage display: identification of organ-specific peptides using deep sequencing and differential profiling across tissues. Nucleic Acids Res. *49*, e38.

35. Kaman, W.E., Voskamp-Visser, I., de Jongh, D.M.C., Endtz, H.P., van Belkum, A., Hays, J.P., and Bikker, F.J. (2013). Evaluation of a D-amino-acid-containing fluorescence resonance energy transfer peptide library for profiling prokaryotic proteases. Anal. Biochem. *441*, 38–43.

36. Poreba, M., Salvesen, G.S., and Drag, M. (2017). Synthesis of a HyCoSuL peptide substrate library to dissect protease substrate specificity. Nat. Protoc. *12*, 2189–2214.

37. Chen, S., Lovell, S., Lee, S., Fellner, M., Mace, P.D., and Bogyo, M. (2021). Identification of highly selective covalent inhibitors by phage display. Nat. Biotechnol. *39*, 490–498.

38. Zhuang, Q., Holt, B.A., Kwong, G.A., and Qiu, P. (2019). Deconvolving multiplexed protease signatures with substrate reduction and activity clustering. PLoS Comput. Biol. *15*. e1006909.

39. Mac, Q.D., Sivakumar, A., Phuengkham, H., Xu, C., Bowen, J.R., Su, F.-Y., Stentz, S.Z., Sim, H., Harris, A.M., Li, T.T., et al. (2022). Urinary detection of early responses to checkpoint blockade and of resistance to it via protease-cleaved antibody-conjugated sensors. Nat. Biomed. Eng. *6*, 310–324.

40. Lustig, M., Donoho, D.L., Santos, J.M., and Pauly, J.M. (2008). Compressed sensing MRI. IEEE Signal Process. Mag. *25*, 72–82.

41. Cleary, B., Cong, L., Cheung, A., Lander, E.S., and Regev, A. (2017). Efficient generation of transcriptomic profiles by random composite measurements. Cell *171*, 1424–1436.e18.

42. Cleary, B., Simonton, B., Bezney, J., Murray, E., Alam, S., Sinha, A., Habibi, E., Marshall, J., Lander, E.S., Chen, F., and Regev, A. (2021). Compressed sensing for highly efficient imaging transcriptomics. Nat. Biotechnol. *39*, 936–942.

43. Matsushita, H., Hosoi, A., Ueha, S., Abe, J., Fujieda, N., Tomura, M., Maekawa, R., Matsushima, K., Ohara, O., and Kakimi, K. (2015). Cytotoxic T lymphocytes block tumor growth both by lytic activity and IFNγ-dependent cell-cycle arrest. Cancer Immunol. Res. *3*, 26–36.

44. Wijeyewickrema, L.C., Yongqing, T., Tran, T.P., Thompson, P.E., Viljoen, J.E., Coetzer, T.H., Duncan, R.C., Kass, I., Buckle, A.M., and Pike, R.N. (2013). Molecular determinants of the substrate specificity of the complement-initiating protease, C1r. J. Biol. Chem. *288*, 15571–15580.

45. Holt, B.A., and Kwong, G.A. (2020). Protease circuits for processing biological information. Nat. Commun. *11*, 5021.

46. López-Otín, C., and Hunter, T. (2010). The regulatory crosstalk between kinases and proteases in cancer. Nat. Rev. Cancer *10*, 278–292.

47. Stites, E., Le Quintrec, M., and Thurman, J.M. (2015). The complement system and antibody-mediated transplant rejection. J. Immunol. *195*, 5525–5531.

48. Anderson, N.L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R.S., Conrads, T.P., Veenstra, T.D., Adkins, J.N., Pounds, J.G., Fagan, R., and Lobley, A. (2004). The human plasma proteome: a nonredundant list developed by combination of four separate sources. Mol. Cell. Proteomics *3*, 311–326.

49. Armstrong, P.B. (2006). Proteases and protease inhibitors: a balance of activities in host-pathogen interaction. Immunobiology *211*, 263–281.

50. Badeau, B.A., Comerford, M.P., Arakawa, C.K., Shadish, J.A., and DeForest, C.A. (2018). Engineered modular biomaterial logic gates for environmentally triggered therapeutic delivery. Nat. Chem. *10*, 251–258.

51. Taghouti, M. (2020). Chapter 10 - compressed sensing. In Computing in Communication Networks, F.H.P. Fitzek, F. Granelli, and P. Seeling, eds. (Academic Press), pp. 197–215.

52. Vashist, S.K. (2017). Point-of-Care diagnostics: recent advances and trends. Biosensors *7*, 62.

53. Olson, O.C., and Joyce, J.A. (2015). Cysteine cathepsin proteases: regulators of cancer progression and therapeutic response. Nat. Rev. Cancer *15*, 712–729.

54. Massberg, S., Grahl, L., von Bruehl, M.L., Manukyan, D., Pfeiler, S., Goosmann, C., Brinkmann, V., Lorenz, M., Bidzhekov, K., Khandagale, A.B., et al. (2010). Reciprocal coupling of coagulation and innate immunity via neutrophil serine proteases. Nat. Med. *16*, 887–896.

55. Menegatti, M., and Palla, R. (2020). Clinical and laboratory diagnosis of rare coagulation disorders (RCDs). Thromb. Res. *196*, 603–608.

56. Chen, X., Gole, J., Gore, A., He, Q., Lu, M., Min, J., Yuan, Z., Yang, X., Jiang, Y., Zhang, T., et al. (2020). Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat. Commun. *11*, 3475.

57. Goebel, C., Louden, C.L., McKenna, R., Onugha, O., Wachtel, A., and Long, T. (2020). Blood test shows high accuracy in detecting stage I non-small cell lung cancer. BMC Cancer *20*, 137.

58. Finn, N.A., and Kemp, M.L. (2014). Systems biology approaches to enzyme kinetics: analyzing network models of drug metabolism. In Enzyme Kinetics in Drug Metabolism: Fundamentals and Applications, S. Nagar, U.A. Argikar, and D.J. Tweedie, eds. (Humana Press), pp. 317–334.

59. Leveson-Gower, R.B., Mayer, C., and Roelfes, G. (2019). The importance of catalytic promiscuity for enzyme design and evolution. Nat. Rev. Chem *3*, 687–705.

60. Sharma, V., Wang, Q., and Lawrence, D.S. (2008). Peptide-based fluorescent sensors of protein kinase activity: design and applications. Biochim. Biophys. Acta *1784*, 94–99.

61. Zhang, J.-F., Liu, B., Hong, I., Mo, A., Roth, R.H., Tenner, B., Lin, W., Zhang, J.Z., Molina, R.S., Drobizhev, M., et al. (2021). An ultrasensitive biosensor for high-resolution kinase activity imaging in awake mice. Nat. Chem. Biol. *17*, 39–46.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Custom peptide library	Genscript; this manuscript	Table S1
Human recombinant kallikrein 2	Prospec	Cat#:ENZ-719
Human recombinant cathepsin S	R&D Systems	Cat#1183-CY-010
Human recombinant ADAMTS1	R&D Systems	Cat#2197-AD-020
Human plasmin	Prolytix	Cat#HCPM-0140
Human alpha-thrombin	Prolytix	Cat#HCT-0020
Human factor Xia	Prolytix	Cat#HCXIA-0160
Human recombinant MMP11	Enzo Life Sciences	Cat#BML-SE282
Human recombinant MMP8	Enzo Life Sciences	Cat#BML-SE255
Human recombinant MMP7	Enzo Life Sciences	Cat#BML-SE181
Human recombinant MMP1	Enzo Life Sciences	Cat#BML-SE180
Human recombinant MMP13	Enzo Life Sciences	Cat#BML-SE246
Experimental models: Organisms/strains		
Mouse: C57BL6/J	The Jackson Laboratory	RRID:IMSR_JAX:000664
Software and algorithms		
MATLAB	MathWorks	https://www.mathworks.com/products/matlab.html
GraphPad Prism	GraphPad Software	https://www.graphpad.com/scientific-software/prism/
Custom code for SLICE and other analysis	This paper	Open Science Framework: https://doi.org/10.17605/OSF.IO/D36EV

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be addressed by the lead contact, Gabriel Kwong (gkwong@gatech.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data supporting the findings of this study are available within the paper and its supplemental information files. Raw data in this paper is available from the lead contact upon request.
- The code supporting the findings of this study is publicly available at Open Science Framework: https://doi.org/10.17605/OSF.IO/D36EV.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

For mouse plasma used in *in vitro* cleavage assays, plasma was isolated from wild-type C57BL/6J mice (Jackson Labs, female, ~8 weeks). These mice display a healthy phenotype and require normal animal maintenance and care. All animal procedures were approved by the Georgia Tech Institutional Animal Care and Use Committee (protocol no. KWONG-A100193).

METHOD DETAILS

Cleavage assays

All protease cleavage assays were performed with a BioTek Cytation 5 Imaging Plate Reader, taking fluorescent measurements at 485/528 nm (excitation/emission) for read-outs measuring peptide substrates terminated with fluorophore 5FAM (5-

Carboxyfluorescein) with quencher Dabcyl. Substrate sequences were identified from previous literature (e.g., substrate-2[44] and substrate-4[45]) or commercial products (e.g., substrate-7 (AnaSpec) and substrate-9 (Enzo Life Sciences)), or were generated from consensus sequences compiled by the MEROPS peptidase database.[28] In all conditions, substrate (20 μM) was added to protease (250 nM) in PBS for each well of a 384-well microplate immediately before reading began. For the *in vitro* classification challenge, two substrates (20 μM) were mixed and added to a mixture of 11 proteases in PBS. For the classification challenge in plasma, citrated plasma was isolated from C57BL/6J mice and added to the mixture of substrates and proteases to 25% of the reaction volume. Plasma was used to generate a more complex biological condition as plasma contains serum proteases and protease inhibitors that could contribute to noise in the classification. Kinetic measurements were taken every minute over the course of 60–120 min at 37 C. Activity RFU measurements were normalized to time 0 measurement, and as such later time points (after time-0) represent fold change in signal. Initial velocity, V0, or product formation rate (RFU/min) is calculated through the line of best fit on the changes in RFU in the first 7 min after the time adjustment. For the classification challenges, product formation rates were normalized between 0 and 1 for each probe solely for data visualization, and unnormalized rates were used for classification. All fluorogenic peptide substrates were purchased from Genscript.

Simulation pipeline for evaluating classification performance of simulated libraries

The simulated disease detection challenge is generated based on a melanoma gene microarray dataset [Matsushita, H. et al.]. This mouse microarray gene expression dataset contains two conditions, healthy(day1) and disease(day7), providing a few samples per condition. Among all known proteases and their corresponding genes, there were 162 proteases genes present in this dataset, which is why the simulated disease detection challenge focused on 162 proteases. For each of the two conditions, an average protease gene expression profile across the samples was calculated, which served as a proxy for the protease activity profile for each condition in this simulated classification challenge.

The average protease gene expression profiles of the two conditions (healthy and disease) were used to generate simulated healthy and disease data points. More specifically, 100 samples are randomly generated where each sample is simulated by adding random Gaussian noise (centered around 0 with a SD of 2) to the average protease expression data. We generated 100 healthy sample based on the protease gene expression profile of day 1, and another 100 disease samples based on the protease gene expression profile of Day 7. So, a total of 200 samples are simulated, half healthy and half diseased.

The noise level (SD2 mentioned above) was chosen such that the multi-variate machine learning classifier Random Forest performed well (i.e., correctly classify healthy vs disease), while uni-variate classifier based on individual protease profiles does not perform well. This choice of noise level represents a situation where we can classify well if we can accurately measure all protease, but cannot classify well if we can only measure one protease. Such noise level would allow us to test whether measuring a few substrates could achieve good classification close to the scenario where we measure all proteases.

For each simulated substrate, the number of proteases that can cleave the substrate is randomly generated between 10–30, and the set of proteases that can cleave the substrate is randomly chosen among the 162 proteases. We create a vector with a length of 162, where each element(protease) is assigned either 0 if not chosen, or 1 if chosen. A set of random values are then assigned to the chosen proteases, which represent the cleavage activity of the chose proteases with respect to the substrate. These random values are generated from Gaussian distributions, and then normalized so that they sum up to 1. These values serve as simulated Kcat values.

Panel number 2 of Figure 2 shows a matrix multiplication of two matrices. The first (leftmost) matrix contains the simulated expression levels for all 162 proteases in the dataset (columns) for the 200 simulated healthy or diseased samples (rows). The expression of any given protease appears similar across healthy and diseased samples due to the Gaussian noise, so that a library of substrates measuring many proteases would likely be necessary for accurate classification. The second (middle) matrix is a simulated substrate vector of Kcat values that describes which proteases can cleave this substrate with what efficiency. This matrix multiplication produces 200 values, which represent the simulated measurements of the 200 simulated samples, if we apply one simulated substrate to sense/measure the 200 samples. When we compare these 200 values against the ground truth label of healthy vs disease, we can draw ROC and compute AUROC, which quantifies the classification power for an individual substrate, as shown in panel number 3 of Figure 2.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was performed using custom MATLAB code (Open Science Framework: https://doi.org/10.17605/OSF.IO/D36EV) and/or using GraphPad Prism. Statistical tests and sample sizes are stated in the figure caption. Unless otherwise stated in the caption, center is defined as mean and error bars depict mean ± SEM, and significance is defined based on p-value <0.05.