Supplementary Information

Peptomer Substrates for Quantitative Pattern-Recognition Sensing of Proteases

Mariah J. Austin,[a] Hattie C. Schunk,[a,b] Natalie Ling,[a] Adrianne M. Rosales[a*]

[a] McKetta Department of Chemical Engineering, University of Texas at Austin, Austin, TX 78712, United States

[b] Biomedical Engineering Department, University of Texas at Austin, Austin, TX 78712, United States

[*] Corresponding author: arosales@che.utexas.edu

Table of Contents

1 Experimental

Materials

L-amino acids, sarcosine, and modified lysines were all Fmoc-protected and purchased from Chem-Impex International, Inc. (Wood Dale, IL), along with Rink amide resin, *O*-(1H-6-Chlorobenzotriazol-1-yl)-*N*,*N*,*N'*,*N'*-tetramethyluronium hexafluorophosphate (HCTU, ≥ 99%), bromoacetic acid (≥ 99%), and *N,N'*-Diisopropylcarbodiimide (DIC, ≥ 99%). Peptoid submonomers, including 1-(2-aminoethyl) pyrrolidine (98%) and glycinamide hydrochloride (98%) were purchased from Acros Organics (Fir Lawn, NJ), as well as triisopropylsilane (TIPS, 98%). Sodium hydroxide (NaOH, ≥97%) and all solvents were purchased from Fisher Scientific (Hampton, NH) at the following purity levels: dimethylformamide (DMF, ≥99.8%), *N*-methyl-2-pyrrolidone (NMP, ≥99%), trifluoroacetic acid (TFA, ≥99.5%), diethyl ether (ether, ≥99%), acetonitrile (ACN, ≥99.9%), 2-propanol (IPA, ≥99.5%), dimethylsulfoxide (DMSO, ≥99.7%). Tris-HCl (≥ 99%) was purchased from Fisher Scientific. Calcium chloride ($CaCl_2$, ≥ 96%) was purchased from Acros Organics. Proteases were all commercially sourced as listed in Table S1.

Table S1: Protease manufacturer information

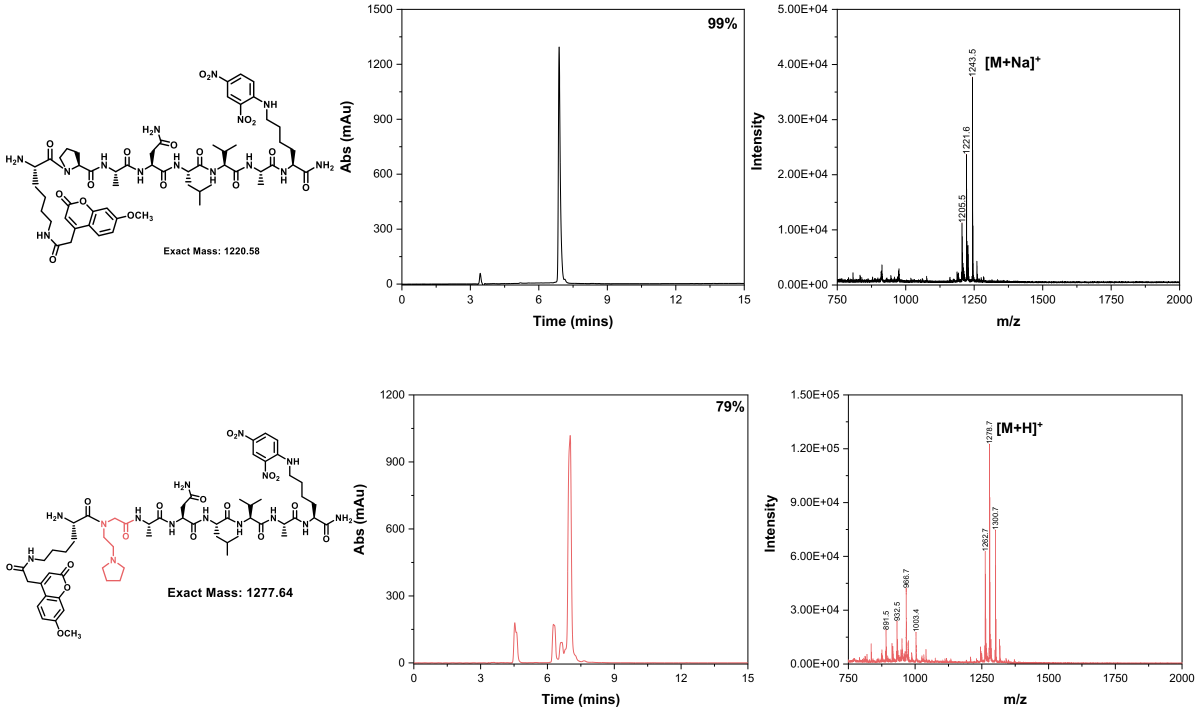
Protease	Manufacturer	Supplier	Product Name	Product/ Catalog Number	Lot Number
Collagenase	Gibco™	ThermoFisher Scientific	Collagenase, Type I, powder	17100017	2357188
Thermolysin	Promega	Promega	Thermolysin	V4001	000394563/ 0000365086 (Dispensed)
Proteinase K	Promega	Promega	Proteinase K	V3021	0000391893/ 0000346043 (Dispensed)
Chymotrypsin	Sigma	Millipore-Sigma	α-Chymotrypsin from bovine pancreas	C4129	SLCH1926
Elastase	Alpha Aesar	ThermoFisher Scientific	Elastase, porcine pancreas	J61874	Q21H009
Papain	Sigma	Millipore-Sigma	Papain from papaya latex	P4762	SLCJ3582

Substrate Preparation

Substrates were synthesized for previous studies and repurposed here.[1] Briefly, peptides and peptomers were all synthesized using Rink Amide polystyrene resin (0.43 mmol g^{-1}) on a Prelude X automated peptide synthesizer (Gyros Protein Technologies) at a scale of 50 µmol. Fmoc groups were removed from the resin and subsequent amino acids by washing twice with 20% piperidine in DMF. Peptide residues and sarcosine utilized Fmoc-protected amino acids (250 mM, 5x molar excess) coupled using HCTU activator (250 mM, 5x molar excess) and NMM (500 mM, 10x molar excess). Coupling steps were performed twice. Fmoc-protected sarcosine was used to generate the *F NAla* peptomer. The *A NPro* and *C NAsn* peptomers included peptoid residues installed according to previously published submonomer synthesis methods[2] using 1-(2-

aminoethyl) pyrrolidine and glycinamide, respectively. First, bromoacylation occurs via addition of bromoacetic acid (1.2 M in DMF) and DIC at a molar ratio of 1:0.93.[3] The bromine is then displaced by a primary amine (2M in NMP) to install the entire peptoid residue. Upon completion of synthesis, substrates were cleaved from the resin using a cleavage cocktail comprised of 95% TFA/ 2.5% water/ 2.5% TIPS for substrates with a sidechain protecting group (those with Asn(Trt)), or 95% TFA/ 5% water. The resin was then filtered off, and the substrates were purified.

Substrates dissolved in cleavage cocktail were added dropwise to a ten-fold volume excess of chilled ether, centrifuged to collect the precipitate, and then washed twice with fresh ether. After drying overnight to remove trace ether, substrates were each dissolved in a mixture of 25% acetonitrile/ 75% water with 0.1% TFA and purified with a semi-prep C18 column on a Dionex UltiMate 3000 UHPLC with a 15-minute gradient from 25-100% acetonitrile at 10 mL min^{-1}. Fractions were collected by their UV signal at 214 nm and the samples were re-purified until only a single peak remained (1-3 purification cycles). Purified substrates were then lyophilized and analyzed by HPLC with an analytical C18 column and via matrix-assisted laser desorption/ionization- time of flight (MALDI-TOF) mass spectrometer using a Bruker autoflex maX instrument to assess purity and confirm molecular weight, respectively. Chemical structures, analytical HPLC traces, and MALDI spectra are shown in Figure S1.



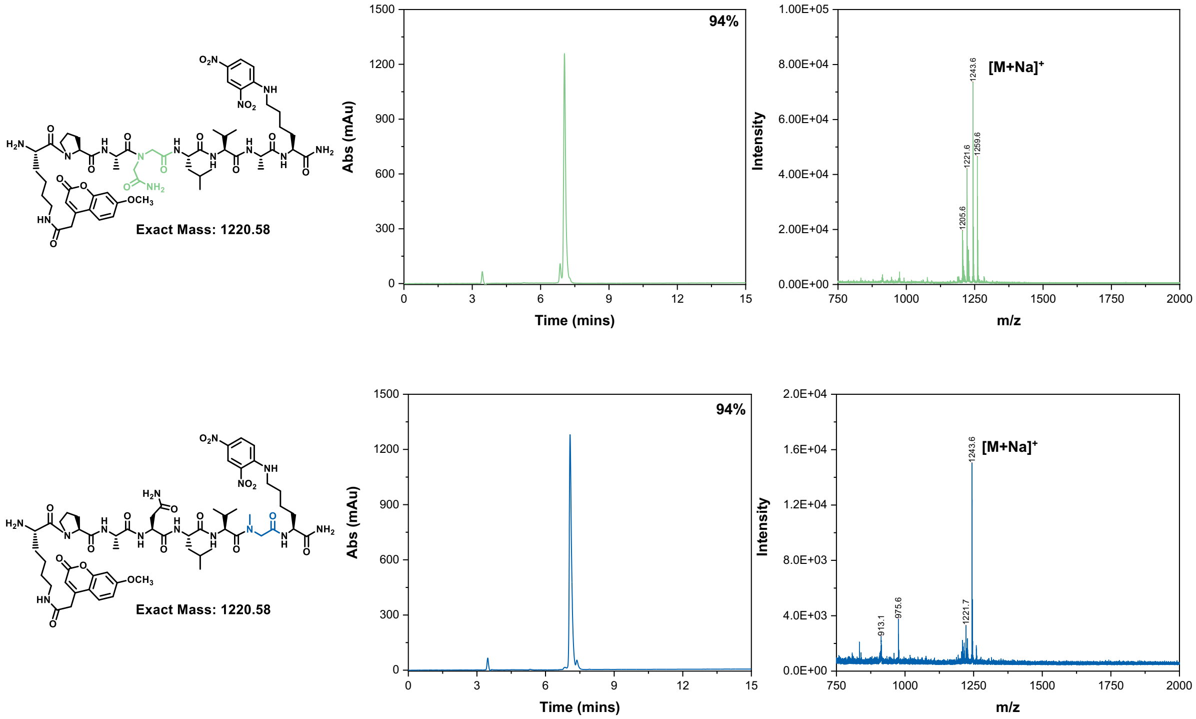


Figure S1: Substrate purity traces.

Peptide (black), *A NPro* (pink), *C NAsn* (green), *F NAla* (blue). Chemical structures with colored peptoid substitution and expected mass (left), analytical HPLC traces with purity determined by integration (center), and MALDI spectra to confirm molecular weight (right). HPLC trace have a solvent peak between 3-4 minutes that was excluded from purity determinations.

Degradation assays

Substrate cleavage was monitored in real-time by tracking the fluorescent signal of 7-methoxycoumarin-4-acetic acid (Mca) as it was liberated from the dinitrophenyl (Dnp) quencher by enzymatic hydrolysis. Substrates were first dissolved in DMSO at a concentration of 1 mM, as measured by the Dnp's absorbance at 363 nm on a NanoDrop OneC Microvolume UV-Vis Spectrophotometer using an experimentally derived extinction coefficient of $\varepsilon_{363\ \mathrm{nm}}$= 16,900 cm^{-1} M^{-1}. Substrates were then diluted to 20 µM with 10% DMSO in our universal protease buffer (50 mM Tris HCl, 10 mM $CaCl_2$, pH 7.8). Protease solutions were made by dissolving proteases at roughly 1 mg mL^{-1} by weight in the same universal protease buffer. Once dissolved and equilibrated, the protease concentration was measured on by NanoDrop using the Scopes measurement parameters for quantifying protein concentration. This measurement was performed in triplicate, and the average value was then used to dilute the solution to 80 µg mL^{-1}. Seven serial dilutions were carried out to give protease solutions at eight different concentrations spanning three orders of magnitude. Next, 50 µL of substrate solution was combined with either 50 µL of buffer (controls) or 50 µL of enzyme solution (samples) in triplicate in a black 96-well plate. The plate was oscillated for 10 seconds to mix and then read on a BioTek Synergy H1 Multi-Mode

Microplate Reader at Ex./Em. 325/392 nm for three hours. Fluorescence values of the controls were subtracted from the sample wells for fluorescence plots.

Dataset collection

10 μM substrate fluorescence traces were fit to an exponential plateau fit in MATLAB according to the equation:

$$y = A - Ae^{-kx}$$

where y was the fluorescence reading, left in units of RFU, A was the saturation value in RFU, k was the kinetic constant in min^{-1}, and x was time in min. For the first fit of each trace, A was constrained to be between 2 x 10^{6} and 7 x 10^{6} RFU based on observed saturation values corresponding to full cleavage. All other traces on that plot (each with a unique substrate-protease combination, but with multiple protease concentrations) then used the same A value for consistency, since not all curves reached saturation within three hours. The k value was constrained to be greater than 1 x 10^{-5} min^{-1} meaning substrates without detectable cleavage would reach this lower limit. This condition was only observed for low concentrations of chymotrypsin and papain cleaving one substrate, thus did not significantly affect feature development.

Multivariate data analysis

After curve fitting, the log$_{10}$ value of each k-constants was calculated and triplicates were averaged, resulting in an array with four features and eight samples for each protease. For PCA, the data matrix was directly fed to the 'pca' function in MATLAB, and the outputs 'coeff,' 'score,' 'latent,' and 'explained' were used to construct visuals in lower dimensional space. For LDA, log$_{10}$ values were first normalized using the minimum and maximum values, then input to the 'fitcdiscr' function, which resulted in a mapping of all eight samples for each protease. To test the model, holdout cross-validation was conducted using the 'cvpartition' function with 'holdout' directed as the method. The holdout value was 0.375, representing 3/8 samples for each protease, as they were stratified by protease identity. The selection of which three concentrations, however, was random and different for each protease. A new model was then trained, again using 'fitcdiscr' but in this case only using 5/8 samples for each protease. The 'predict' function was then carried out with the 'identity,' 'score,' and 'cost' recorded. The predicted identities were used to assemble a confusion matrix representing the classification accuracy, and the scores were used to quantify the posterior probabilities for each prediction. Finally, LDA-classified samples were fit by multiple regression to predict their concentrations. Multiple regression was subjected to the same holdout cross-validation, thus each curve was fit using five sample concentrations. Regression was carried out with log$_{10}$ values using MATLAB's 'regress' function, which determined coefficients to be applied for each feature. The regression fit used was determined by the protease identity predicted by LDA. Finally, predicted values were converted back to concentration by raising 10 to the power of the log value calculated. MATLAB code is available at github.com/mariahaustin/Quantitative-Protease-Pattern-Recognition.

2 Supporting Data

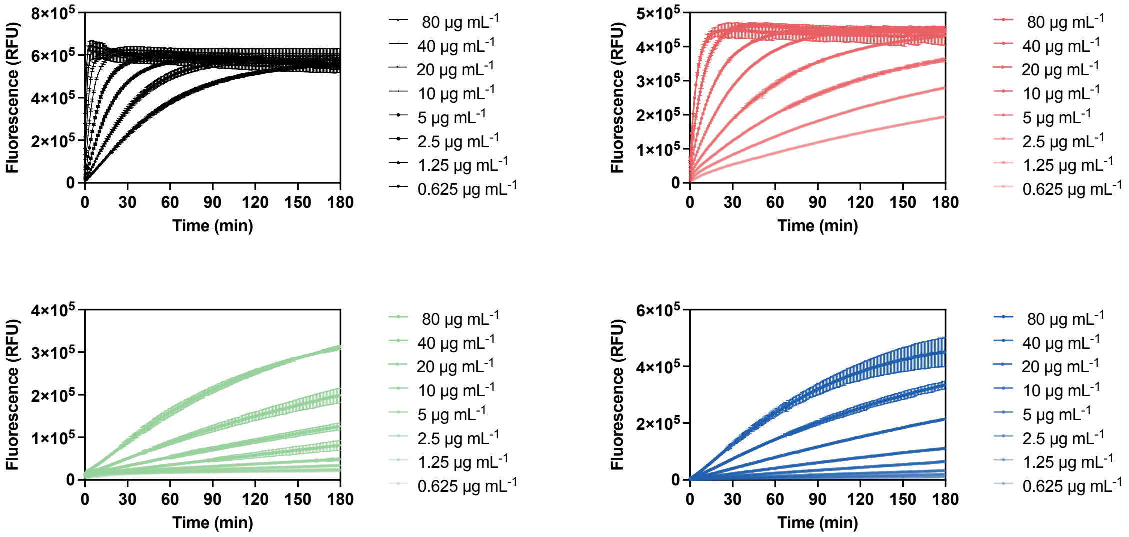


Figure S2: Proteinase K fluorescence cleavage tracking

Peptide (black), *A NPro* (pink), *C NAsn* (green), *F NAla* (blue)

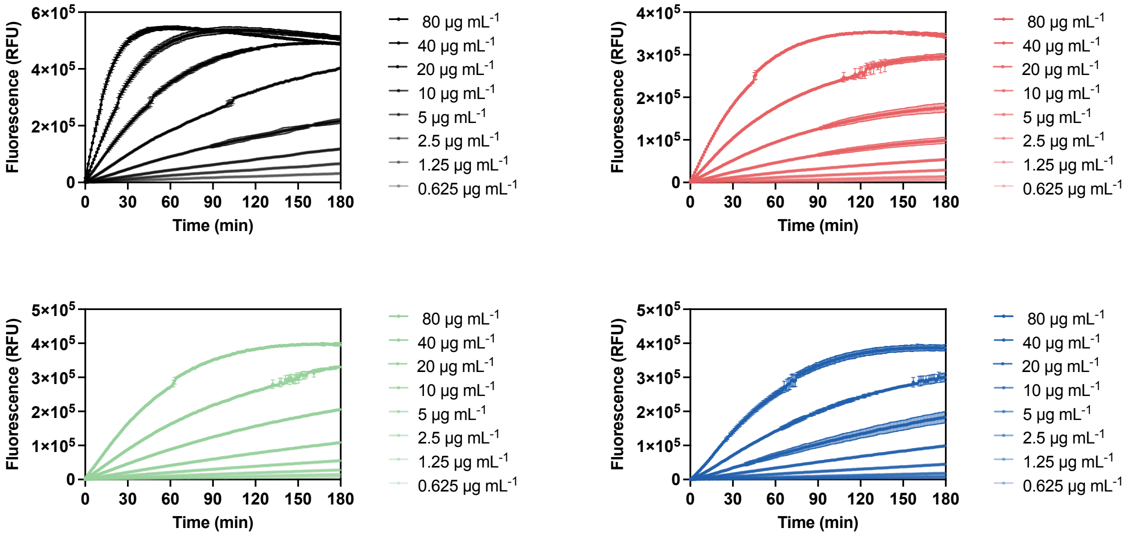


Figure S3: Elastase fluorescence cleavage tracking

Peptide (black), *A NPro* (pink), *C NAsn* (green), *F NAla* (blue)

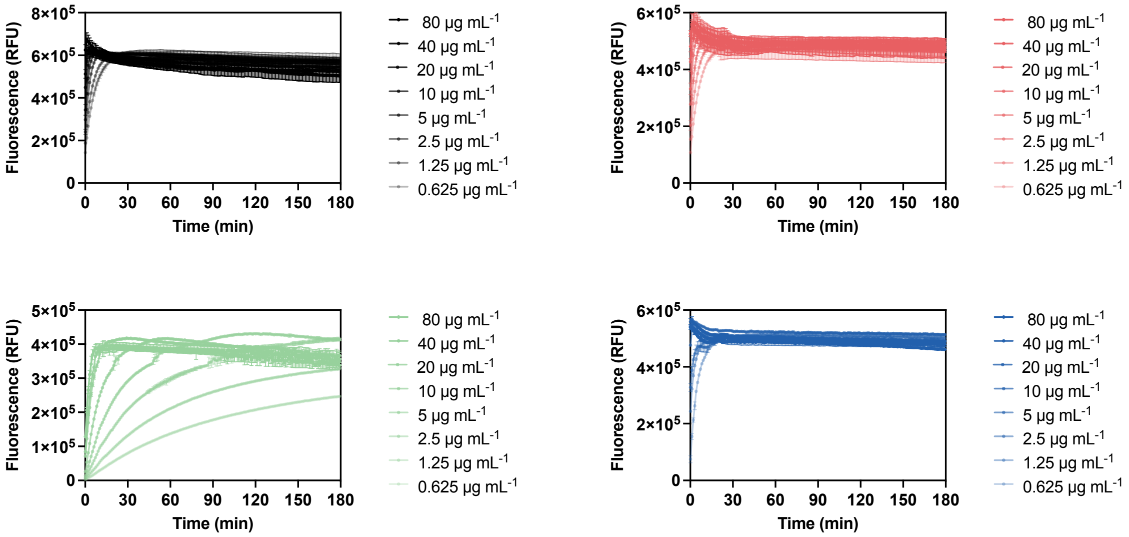


Figure S4: Thermolysin fluorescence cleavage tracking

Peptide (black), *A NPro* (pink), *C NAsn* (green), *F NAla* (blue)

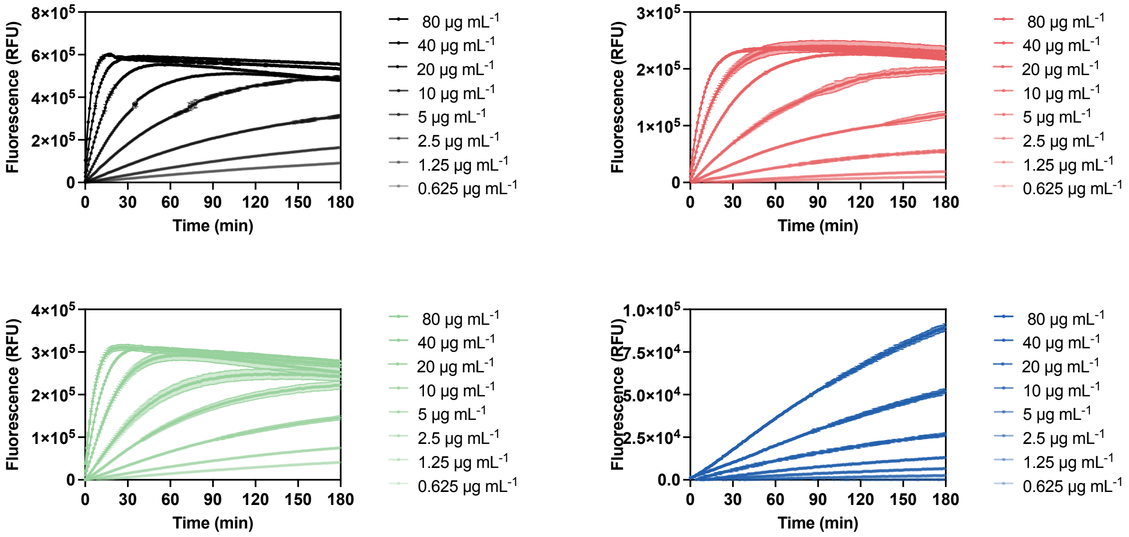


Figure S5: Chymotrypsin fluorescence cleavage tracking

Peptide (black), *A NPro* (pink), *C NAsn* (green), *F NAla* (blue)

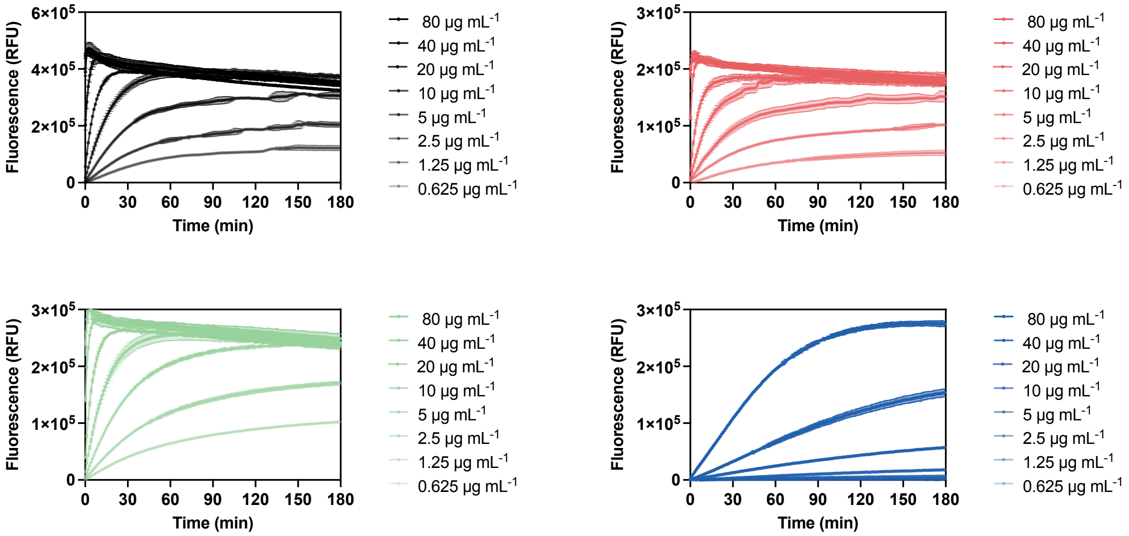


Figure S6: Papain fluorescence cleavage tracking

Peptide (black), *A NPro* (pink), *C NAsn* (green), *F NAla* (blue)

Table S2: Degradation rates used as features for multivariate data analysis

Degradation traces were individually fit to the exponential plateau function, resulting in a k-constant representing the rate of degradation. The $\log_{10}$ value of the k-constants were then calculated, and the average of the triplicate measurements was used as the input for PCA, LDA, and multiple regression.

Protease	**Concentration (μg mL^{-1})**	***Peptide***	***A NPro***	***C NAsn***	***F NAla***
Collagenase	80	-0.72	-0.79	-2.05	-1.17
Collagenase	40	-1.01	-1.12	-2.40	-1.50
Collagenase	20	-1.38	-1.53	-2.78	-1.87
Collagenase	10	-1.66	-1.83	-3.10	-2.30
Collagenase	5	-1.96	-2.20	-3.41	-2.63
Collagenase	2.5	-2.25	-2.45	-3.69	-2.98
Collagenase	1.25	-2.59	-2.88	-3.96	-3.30
Collagenase	0.625	-2.89	-3.28	-4.26	-3.65
Proteinase K	80	0.59	-0.61	-2.10	-2.05
Proteinase K	40	0.05	-0.89	-2.42	-2.31
Proteinase K	20	-0.34	-1.22	-2.68	-2.59
Proteinase K	10	-0.66	-1.47	-2.89	-2.91
Proteinase K	5	-0.99	-1.76	-3.10	-3.17
Proteinase K	2.5	-1.28	-2.02	-3.23	-3.47
Proteinase K	1.25	-1.52	-2.25	-3.35	-3.69
Proteinase K	0.625	-1.77	-2.48	-3.40	-3.93
Elastase	80	-1.11	-1.58	-1.75	-1.85
Elastase	40	-1.42	-1.97	-2.10	-2.19
Elastase	20	-1.75	-2.36	-2.42	-2.53
Elastase	10	-2.12	-2.70	-2.78	-2.86
Elastase	5	-2.52	-3.02	-3.10	-3.22
Elastase	2.5	-2.85	-3.31	-3.41	-3.62
Elastase	1.25	-3.13	-3.66	-3.69	-3.91
Elastase	0.625	-3.46	-4.05	-3.95	-4.12
Thermolysin	80	2.17	2.12	-0.13	2.12
Thermolysin	40	1.15	1.78	-0.38	2.11
Thermolysin	20	0.95	1.11	-0.49	2.10
Thermolysin	10	0.71	0.81	-0.90	2.12
Thermolysin	5	0.58	0.67	-1.23	1.60
Thermolysin	2.5	-0.09	-0.09	-1.51	0.79
Thermolysin	1.25	-0.40	-0.42	-1.85	-0.05
Thermolysin	0.625	-0.59	-0.68	-2.14	-0.60
Chymotrypsin	80	-0.46	-0.91	-0.65	-2.48
Chymotrypsin	40	-0.81	-1.18	-0.92	-2.76
Chymotrypsin	20	-1.14	-1.48	-1.24	-3.07
Chymotrypsin	10	-1.57	-1.90	-1.74	-3.40
Chymotrypsin	5	-1.88	-2.35	-2.02	-3.71
Chymotrypsin	2.5	-2.32	-2.76	-2.38	-4.18
Chymotrypsin	1.25	-2.69	-3.28	-2.75	-5.00
Chymotrypsin	0.625	-2.99	-3.59	-3.06	-4.99
Papain	80	2.14	1.90	2.11	-1.76
Papain	40	2.09	1.25	1.43	-2.37
Papain	20	0.47	0.64	0.61	-2.89
Papain	10	-0.69	-0.83	-0.64	-3.43
Papain	5	-1.16	-1.33	-1.14	-3.85
Papain	2.5	-1.76	-1.89	-1.59	-4.24
Papain	1.25	-2.22	-2.28	-2.09	-4.31
Papain	0.625	-2.53	-2.66	-2.45	-5.00

Table S3: Iterations of holdout cross-validation

Protease classification by LDA and concentration estimations by multiple regression were carried out with holdout cross-validation applied ten times. Misclassifications are labeled in red. For each iteration, the percent classification accuracy (correctly classified samples/total samples), the percent of concentration predictions within bounds (predicted concentrations between the concentrations above and below the actual concentration—*e.g.,* for a sample with actual concentration of 2.5 µg mL^{-1} a predicted concentration was in bounds if it was between 1.25 and 5 µg mL^{-1}), and average percent error in predicted concentrations for correctly classified proteases.

Iteration	Protease	Predicted Protease	Predicted Concentration (µg mL^{-1})	Actual Concentration (µg mL^{-1})	Within Bounds	Error (µg mL^{-1})	% Error (Abs)
1	Collagenase	Collagenase	3.85	2.50	1	1.35	54%
1	Collagenase	Collagenase	2.09	1.25	1	0.84	67%
1	Collagenase	Collagenase	0.90	0.63	1	0.28	45%
1	Proteinase K	Proteinase K	9.51	10.00	1	-0.49	5%
1	Proteinase K	Proteinase K	1.46	1.25	1	0.21	17%
1	Proteinase K	Proteinase K	0.81	0.63	1	0.19	30%
1	Elastase	Elastase	83.73	80.00	1	3.73	5%
1	Elastase	Elastase	18.49	20.00	1	-1.51	8%
1	Elastase	Elastase	0.62	0.63	1	0.00	0%
1	Thermolysin	Thermolysin	9.11	40.00	0	-30.89	77%
1	Thermolysin	Thermolysin	2.72	2.50	1	0.22	9%
1	Thermolysin	Collagenase	65.88	0.63	0	65.26	10441%
1	Chymotrypsin	Chymotrypsin	21.02	20.00	1	1.02	5%
1	Chymotrypsin	Chymotrypsin	9.51	10.00	1	-0.49	5%
1	Chymotrypsin	Chymotrypsin	2.16	2.50	1	-0.34	14%
1	Papain	Papain	5.25	10.00	1	-4.75	47%
1	Papain	Papain	2.97	5.00	1	-2.03	41%
1	Papain	Papain	1.54	2.50	1	-0.96	38%
1	**Classification Accuracy**	**94%**		**Within Bounds**	**89%**	**Average Error**	**27%**

Iteration	Protease	Predicted Protease	Predicted Concentration (µg mL^{-1})	Actual Concentration (µg mL^{-1})	Within Bounds	Error (µg mL^{-1})	% Error
2	Collagenase	Collagenase	64.41	80	1	-15.59	19%
2	Collagenase	Collagenase	21.17	20	1	1.17	6%
2	Collagenase	Collagenase	9.11	10	1	-0.89	9%
2	Proteinase K	Proteinase K	31.81	80	0	-48.19	60%
2	Proteinase K	Proteinase K	17.41	20	1	-2.59	13%
2	Proteinase K	Proteinase K	10.72	10	1	0.72	7%
2	Elastase	Elastase	73.95	80	1	-6.05	8%
2	Elastase	Elastase	4.58	5	1	-0.42	8%

2	Elastase	Elastase	1.39	1.25	1	0.14	11%
2	Thermolysin	Thermolysin	16.27	80	0	-63.73	80%
2	Thermolysin	Thermolysin	47.78	40	1	7.78	19%
2	Thermolysin	Thermolysin	2.72	2.5	1	0.22	9%
2	Chymotrypsin	Chymotrypsin	90.27	80	1	10.27	13%
2	Chymotrypsin	Chymotrypsin	6.45	5	1	1.45	29%
2	Chymotrypsin	Chymotrypsin	0.83	0.625	1	0.21	33%
2	Papain	Papain	5.25	10	1	-4.75	47%
2	Papain	Papain	2.97	5	1	-2.03	41%
2	Papain	Papain	1.54	2.5	1	-0.96	38%
2	**Classification Accuracy**	**100%**		**Within Bounds**	**89%**	**Average Error**	**25%**
Iteration	**Protease**	**Predicted Protease**	**Predicted Concentration (μg mL^{-1})**	**Actual Concentration (μg mL^{-1})**	**Within Bounds**	**Error (μg mL^{-1})**	**% Error**
3	Collagenase	Collagenase	9.12	80	0	-70.88	89%
3	Collagenase	Collagenase	16.11	40	0	-23.89	60%
3	Collagenase	Collagenase	6.15	10	1	-3.85	38%
3	Proteinase K	Proteinase K	52.88	40	1	12.88	32%
3	Proteinase K	Proteinase K	11.47	10	1	1.47	15%
3	Proteinase K	Proteinase K	2.49	2.5	1	-0.01	0%
3	Elastase	Elastase	18.79	20	1	-1.21	6%
3	Elastase	Elastase	10.39	10	1	0.39	4%
3	Elastase	Elastase	2.45	2.5	1	-0.05	2%
3	Thermolysin	Thermolysin	32.52	20	1	12.52	63%
3	Thermolysin	Thermolysin	4.45	5	1	-0.55	11%
3	Thermolysin	Thermolysin	2.66	2.5	1	0.16	6%
3	Chymotrypsin	Chymotrypsin	94.77	80	1	14.77	18%
3	Chymotrypsin	Chymotrypsin	5.31	5	1	0.31	6%
3	Chymotrypsin	Chymotrypsin	2.00	2.5	1	-0.50	20%
3	Papain	Papain	10.55	20	1	-9.45	47%
3	Papain	Papain	3.35	1.25	0	2.10	168%
3	Papain	Papain	0.65	0.625	1	0.02	4%
3	**Classification Accuracy**	**100%**		**Within Bounds**	**83%**	**Average Error**	**33%**
Iteration	**Protease**	**Predicted Protease**	**Predicted Concentration (μg mL^{-1})**	**Actual Concentration (μg mL^{-1})**	**Within Bounds**	**Error (μg mL^{-1})**	**% Error**
4	Collagenase	Collagenase	64.92	80	1	-15.08	19%
4	Collagenase	Collagenase	3.04	2.5	1	0.54	22%
4	Collagenase	Collagenase	0.54	0.625	1	-0.08	14%

4	Proteinase K	Proteinase K	45.84	80	1	-34.16	43%
4	Proteinase K	Proteinase K	19.52	20	1	-0.48	2%
4	Proteinase K	Proteinase K	2.36	2.5	1	-0.14	6%
4	Elastase	Elastase	39.78	40	1	-0.22	1%
4	Elastase	Elastase	1.39	1.25	1	0.14	11%
4	Elastase	Elastase	0.81	0.625	1	0.19	30%
4	Thermolysin	Thermolysin	12.99	80	0	-67.01	84%
4	Thermolysin	Thermolysin	108.75	40	0	68.75	172%
4	Thermolysin	Thermolysin	9.20	10	1	-0.80	8%
4	Chymotrypsin	Chymotrypsin	54.68	80	1	-25.32	32%
4	Chymotrypsin	Chymotrypsin	33.32	40	1	-6.68	17%
4	Chymotrypsin	Chymotrypsin	0.70	0.625	1	0.08	12%
4	Papain	Papain	578.79	80	0	498.79	623%
4	Papain	Papain	10.35	2.5	0	7.85	314%
4	Papain	Papain	1.57	0.625	0	0.94	151%
4	**Classification Accuracy**	**100%**		**Within Bounds**	**72%**	**Average Error**	**87%**
Iteration	**Protease**	**Predicted Protease**	**Predicted Concentration (μg mL^{-1})**	**Actual Concentration (μg mL^{-1})**	**Within Bounds**	**Error (μg mL^{-1})**	**% Error**
5	Collagenase	Collagenase	3.85	2.5	1	1.35	54%
5	Collagenase	Collagenase	2.09	1.25	1	0.84	67%
5	Collagenase	Collagenase	0.90	0.625	1	0.28	45%
5	Proteinase K	Proteinase K	9.51	10	1	-0.49	5%
5	Proteinase K	Proteinase K	1.46	1.25	1	0.21	17%
5	Proteinase K	Proteinase K	0.81	0.625	1	0.19	30%
5	Elastase	Elastase	83.73	80	1	3.73	5%
5	Elastase	Elastase	18.49	20	1	-1.51	8%
5	Elastase	Elastase	0.62	0.625	1	0.00	0%
5	Thermolysin	Thermolysin	9.11	40	0	-30.89	77%
5	Thermolysin	Thermolysin	2.72	2.5	1	0.22	9%
5	Thermolysin	Collagenase	65.88	0.625	0	65.26	10441%
5	Chymotrypsin	Chymotrypsin	21.02	20	1	1.02	5%
5	Chymotrypsin	Chymotrypsin	9.51	10	1	-0.49	5%
5	Chymotrypsin	Chymotrypsin	2.16	2.5	1	-0.34	14%
5	Papain	Papain	5.25	10	1	-4.75	47%
5	Papain	Papain	2.97	5	1	-2.03	41%
5	Papain	Papain	1.54	2.5	1	-0.96	38%
5	**Classification Accuracy**	**94%**		**Within Bounds**	**89%**	**Average Error**	**27%**

Iteration	Protease	Predicted Protease	Predicted Concentration (μg mL^{-1})	Actual Concentration (μg mL^{-1})	Within Bounds	Error (μg mL^{-1})	% Error
6	Collagenase	Collagenase	116.23	80	1	36.23	45%
6	Collagenase	Collagenase	50.43	40	1	10.43	26%
6	Collagenase	Collagenase	4.81	5	1	-0.19	4%
6	Proteinase K	Proteinase K	45.84	80	1	-34.16	43%
6	Proteinase K	Proteinase K	19.52	20	1	-0.48	2%
6	Proteinase K	Proteinase K	2.36	2.5	1	-0.14	6%
6	Elastase	Elastase	38.76	40	1	-1.24	3%
6	Elastase	Elastase	17.97	20	1	-2.03	10%
6	Elastase	Elastase	5.01	5	1	0.01	0%
6	Thermolysin	Thermolysin	80.20	80	1	0.20	0%
6	Thermolysin	Thermolysin	10.78	40	0	-29.22	73%
6	Thermolysin	Thermolysin	1.12	1.25	1	-0.13	11%
6	Chymotrypsin	Chymotrypsin	7.91	10	1	-2.09	21%
6	Chymotrypsin	Chymotrypsin	4.68	5	1	-0.32	6%
6	Chymotrypsin	Chymotrypsin	2.07	2.5	1	-0.43	17%
6	Papain	Chymotrypsin	27838.94	40	0	27798.94	69497%
6	Papain	Papain	0.00	20	0	-20.00	100%
6	Papain	Papain	0.00	0.625	0	-0.62	100%
6	**Classification Accuracy**	**94%**		**Within Bounds**	**78%**	**Average Error**	**27%**

Iteration	Protease	Predicted Protease	Predicted Concentration (μg mL^{-1})	Actual Concentration (μg mL^{-1})	Within Bounds	Error (μg mL^{-1})	% Error
7	Collagenase	Collagenase	68.87	80	1	-11.13	14%
7	Collagenase	Collagenase	3.25	2.5	1	0.75	30%
7	Collagenase	Collagenase	1.45	1.25	1	0.20	16%
7	Proteinase K	Proteinase K	42.57	40	1	2.57	6%
7	Proteinase K	Proteinase K	2.33	2.5	1	-0.17	7%
7	Proteinase K	Proteinase K	0.59	0.625	1	-0.04	6%
7	Elastase	Elastase	95.52	80	1	15.52	19%
7	Elastase	Elastase	43.61	40	1	3.61	9%
7	Elastase	Elastase	0.64	0.625	1	0.01	2%
7	Thermolysin	Thermolysin	26.32	20	1	6.32	32%
7	Thermolysin	Thermolysin	9.66	10	1	-0.34	3%
7	Thermolysin	Thermolysin	0.68	0.625	1	0.05	8%
7	Chymotrypsin	Chymotrypsin	94.77	80	1	14.77	18%
7	Chymotrypsin	Chymotrypsin	5.31	5	1	0.31	6%
7	Chymotrypsin	Chymotrypsin	2.00	2.5	1	-0.50	20%

7	Papain	Papain	638.50	20	0	618.50	3092%
7	Papain	Papain	5.20	10	1	-4.80	48%
7	Papain	Papain	6.26	0.625	0	5.64	902%
7	**Classification Accuracy**	**100%**		**Within Bounds**	**89%**	**Average Error**	**236%**

Iteration	Protease	Predicted Protease	Predicted Concentration (μg mL^{-1})	Actual Concentration (μg mL^{-1})	Within Bounds	Error (μg mL^{-1})	% Error
8	Collagenase	Collagenase	68.87	80	1	-11.13	14%
8	Collagenase	Collagenase	3.25	2.5	1	0.75	30%
8	Collagenase	Collagenase	1.45	1.25	1	0.20	16%
8	Proteinase K	Proteinase K	39.31	40	1	-0.69	2%
8	Proteinase K	Proteinase K	11.73	20	1	-8.27	41%
8	Proteinase K	Proteinase K	4.10	5	1	-0.90	18%
8	Elastase	Elastase	83.57	80	1	3.57	4%
8	Elastase	Elastase	2.25	2.5	1	-0.25	10%
8	Elastase	Elastase	0.86	0.625	1	0.23	37%
8	Thermolysin	Thermolysin	1.30	2.5	1	-1.20	48%
8	Thermolysin	Thermolysin	0.47	1.25	0	-0.78	62%
8	Thermolysin	Collagenase	192.54	0.625	0	191.91	30706%
8	Chymotrypsin	Chymotrypsin	62.00	80	1	-18.00	22%
8	Chymotrypsin	Chymotrypsin	23.46	20	1	3.46	17%
8	Chymotrypsin	Chymotrypsin	4.78	5	1	-0.22	4%
8	Papain	Papain	5.37	5	1	0.37	7%
8	Papain	Chymotrypsin	7.20	1.25	0	5.95	476%
8	Papain	Chymotrypsin	4.21	0.625	0	3.58	573%
8	**Classification Accuracy**	**83%**		**Within Bounds**	**78%**	**Average Error**	**22%**

Iteration	Protease	Predicted Protease	Predicted Concentration (μg mL^{-1})	Actual Concentration (μg mL^{-1})	Within Bounds	Error (μg mL^{-1})	% Error
9	Collagenase	Collagenase	61.51	80	1	-18.49	23%
9	Collagenase	Collagenase	5.61	5	1	0.61	12%
9	Collagenase	Collagenase	1.12	1.25	1	-0.13	10%
9	Proteinase K	Proteinase K	38.14	80	0	-41.86	52%
9	Proteinase K	Proteinase K	2.33	2.5	1	-0.17	7%
9	Proteinase K	Proteinase K	1.23	1.25	1	-0.02	2%
9	Elastase	Elastase	18.81	20	1	-1.19	6%
9	Elastase	Elastase	10.32	10	1	0.32	3%
9	Elastase	Elastase	1.28	1.25	1	0.03	2%
9	Thermolysin	Thermolysin	10.81	40	0	-29.19	73%

9	Thermolysin	Thermolysin	2.50	2.5	1	0.00	0%
9	Thermolysin	Thermolysin	1.12	1.25	1	-0.13	11%
9	Chymotrypsin	Chymotrypsin	22.20	20	1	2.20	11%
9	Chymotrypsin	Chymotrypsin	3.17	1.25	0	1.92	154%
9	Chymotrypsin	Chymotrypsin	0.97	0.625	1	0.34	55%
9	Papain	Papain	21.00	10	0	11.00	110%
9	Papain	Papain	9.06	5	1	4.06	81%
9	Papain	Papain	10.50	1.25	0	9.25	740%
9	**Classification Accuracy**	**100%**		**Within Bounds**	**72%**	**Average Error**	**75%**
Iteration	**Protease**	**Predicted Protease**	**Predicted Concentration (μg mL^{-1})**	**Actual Concentration (μg mL^{-1})**	**Within Bounds**	**Error (μg mL^{-1})**	**% Error**
10	Collagenase	Collagenase	84.69	80	1	4.69	6%
10	Collagenase	Collagenase	15.23	20	1	-4.77	24%
10	Collagenase	Collagenase	1.10	1.25	1	-0.15	12%
10	Proteinase K	Proteinase K	634.93	80	0	554.93	694%
10	Proteinase K	Proteinase K	51.25	40	1	11.25	28%
10	Proteinase K	Proteinase K	1.47	1.25	1	0.22	17%
10	Elastase	Elastase	38.82	40	1	-1.18	3%
10	Elastase	Elastase	18.00	20	1	-2.00	10%
10	Elastase	Elastase	0.62	0.625	1	0.00	1%
10	Thermolysin	Thermolysin	14.88	10	1	4.88	49%
10	Thermolysin	Thermolysin	9.79	5	1	4.79	96%
10	Thermolysin	Thermolysin	1.10	1.25	1	-0.15	12%
10	Chymotrypsin	Chymotrypsin	73.19	80	1	-6.81	9%
10	Chymotrypsin	Chymotrypsin	44.65	40	1	4.65	12%
10	Chymotrypsin	Chymotrypsin	24.86	20	1	4.86	24%
10	Papain	Papain	1.94	2.5	1	-0.56	22%
10	Papain	Chymotrypsin	6.05	1.25	0	4.80	384%
10	Papain	Chymotrypsin	3.39	0.625	0	2.77	443%
10	**Classification Accuracy**	**89%**		**Within Bounds**	**83%**	**Average Error**	**64%**

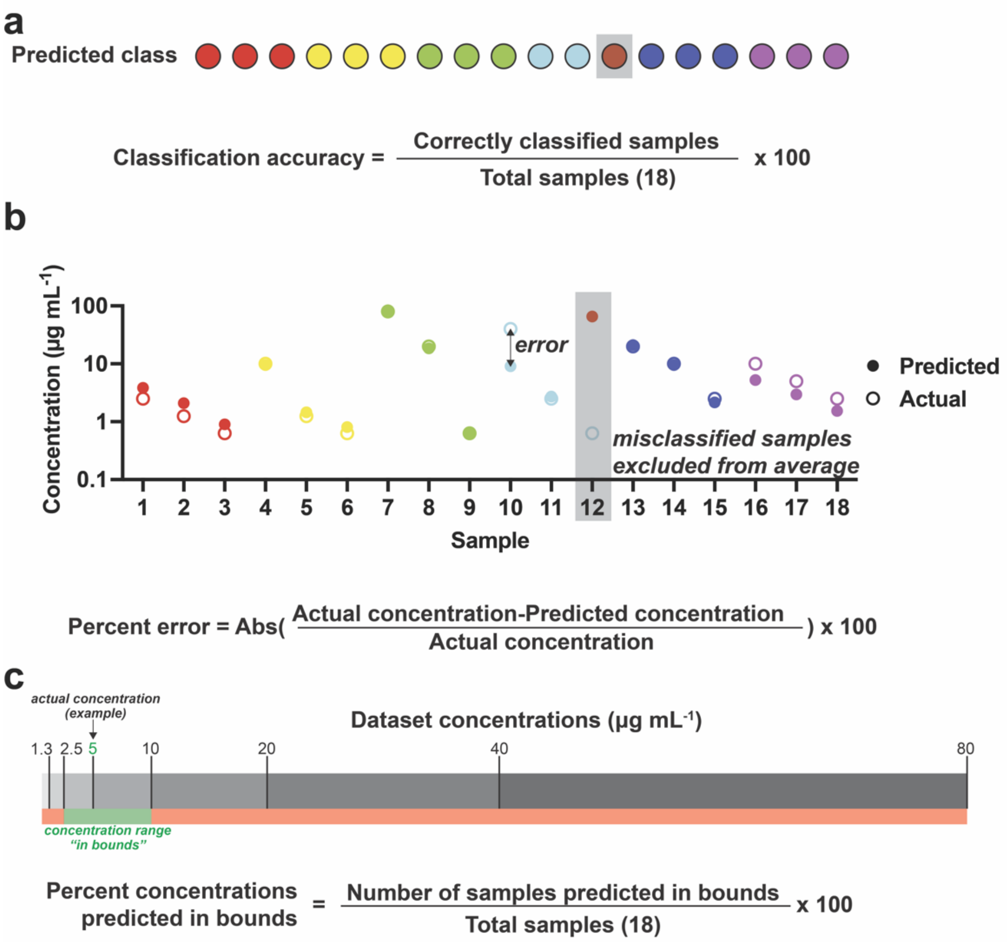

a
Predicted class
Classification accuracy = Correctly classified samples / Total samples (18) x 100
b
Concentration (µg mL-1)
100
10
1
0.1
error
misclassified samples
excluded from average
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Sample
● Predicted
○ Actual
Percent error = Abs(Actual concentration-Predicted concentration / Actual concentration) x 100
c
actual concentration
(example)
Dataset concentrations (µg mL-1)
1.3
2.5
5
10
20
40
80
concentration range
"in bounds"
Percent concentrations predicted in bounds = Number of samples predicted in bounds / Total samples (18) x 100

Figure S7: Metric extraction methodology

For each iteration of holdout cross-validation, three metrics were quantified: (a) Classification accuracy is the percentage of correctly classified samples out of the total, which is 18 samples for each round. (b) The percent error in concentration predicted is calculated for each sample. The percent error is then averaged for each of the samples that had correct classification (as misclassification led to highly erroneous concentrations that qualify as outliers for the metric). (c) The predicted concentrations were determined to be "in bounds" if they were within two-fold of the actual concentration, since the training dataset was comprised of serially diluted concentrations. The percentage in bounds is tabulated for all 18 samples in a holdout iteration, regardless of correct classification. Classification accuracy and percent in bounds are given as a percentage of the 18 samples in each holdout round. Percent error is individually calculated for each sample, then averaged for the 18 samples in the round.

3 References

1 M. J. Austin, H. Schunk, C. Watkins, N. Ling, J. Chauvin, L. Morton and A. M. Rosales, *Biomacromolecules*, 2022, **23**, 4909–4923.

2 R. N. Zuckermann, J. M. Kerr, W. H. Moosf and S. B. H. Kent, *J. Am. Chem. Soc.*, 1992, **114**, 10646–10647.

3 B. C. Lee, R. N. Zuckermann and K. A. Dill, *J. Am. Chem. Soc.*, 2005, **127**, 10999–11009.