ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Protease Activity Analysis: A Toolkit for Analyzing Enzyme Activity Data

Ava P. Soleimany,*[,¶] Carmen Martin-Alonso,[¶] Melodi Anahtar,[¶] Cathy S. Wang, and Sangeeta N. Bhatia*

Cite This: *ACS Omega* 2022, 7, 24292−24301

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Analyzing the activity of proteases and their substrates is critical to defining the biological functions of these enzymes and to designing new diagnostics and therapeutics that target protease dysregulation in disease. While a wide range of databases and algorithms have been created to better predict protease cleavage sites, there is a dearth of computational tools to automate analysis of *in vitro* and *in vivo* protease assays. This necessitates individual researchers to develop their own analytical pipelines, resulting in a lack of standardization across the field. To facilitate protease research, here we present Protease Activity Analysis (PAA), a toolkit for the preprocessing, visualization, machine learning analysis, and querying of protease activity data sets. PAA leverages a Python-based object-oriented implementation that provides a modular framework for streamlined analysis across three major components. First, PAA provides a facile framework to query data sets of synthetic peptide substrates and their cleavage susceptibilities across a diverse set of proteases. To complement the database functionality, PAA also includes tools for the automated analysis and visualization of user-input enzyme−substrate activity measurements generated through *in vitro* screens against synthetic peptide substrates. Finally, PAA supports a set of modular machine learning functions to analyze *in vivo* protease activity signatures that are generated by activity-based sensors. Overall, PAA offers the protease community a breadth of computational tools to streamline research, taking a step toward standardizing data analysis across the field and in chemical biology and biochemistry at large.

INTRODUCTION

Proteases play essential roles in diverse biological processes ranging from development to differentiation, and dysregulated protease activity is a driver of a variety of pathological conditions including cancer, neurodegeneration, and infectious diseases.[1] Because proteases most proximally exert their function through their *activity*, understanding protease activity, rather than transcript or protein expression, is required to elucidate the biological roles of proteases and to harness these enzymes as diagnostic and therapeutic targets. To this end, molecular tools such as activity-based probes (ABPs), short synthetic peptide substrates, and noninvasive enzyme activity sensors have been developed to measure protease activities *in vitro*, i.e., of recombinant proteases or enzymes present in biospecimens,[2−6] as well as *in vivo*, i.e., within the disease microenvironment.[7−11] Beyond their use as a discovery tool, sensors that quantify protease activity are being applied directly for early detection and monitoring of disease,[5,12−18] biological imaging,[7,19] and drug discovery.[20,21] Furthermore, proteolytic cleavage of peptide linkers is being used to trigger disease-specific activation of engineered activity-based diagnostics[11] and therapeutics,[22−25] all of which inherently rely on assessments of protease activity for their design and optimization. To support the development of these new activity-based tools and to advance the study of protease biology at large, a wide range of databases and algorithms have been created to better identify protease substrates and cleavage sites, providing a clear demonstration of how protease research can benefit from computational tools.[26−28]

Rapidly identifying, designing, and characterizing new peptide substrates and activity-based sensors remains a major bottleneck toward advancing these applications. This is due to the promiscuous nature of protease activity, the combinatorial number of synthetically accessible substrates, and the dearth of methods to automate protease activity analysis and substrate design. Current protease databases and analytic tools also focus exclusively on endogenous substrates and cleavage sites,[29−31] despite the fact that synthetic activity-based sensors and large-scale libraries of synthetic peptides are now standard tools for measuring and quantifying protease activity *in vivo* and *in vitro*. Furthermore, the development of these experimental and molecular methods has not been accompanied by scalable, modular data analytic workflows. The creation of computa-

Received: March 15, 2022
Accepted: June 10, 2022
Published: July 6, 2022

tional packages similar to what exists for genomics (e.g., Bioconductor) could enable the standardized analysis of data generated from both *in vitro* and *in vivo* protease-based experiments. This would greatly benefit researchers by accelerating experimental workflows, informing diagnostic and therapeutic design, and facilitating biological insight into protease dysregulation in disease.

To address these needs, we present Protease Activity Analysis (PAA), a toolkit that addresses the need for computational methods to accelerate data analysis of enzyme activity data in biochemistry, chemical biology, and bioengineering (Figure 1). PAA contains a searchable database of

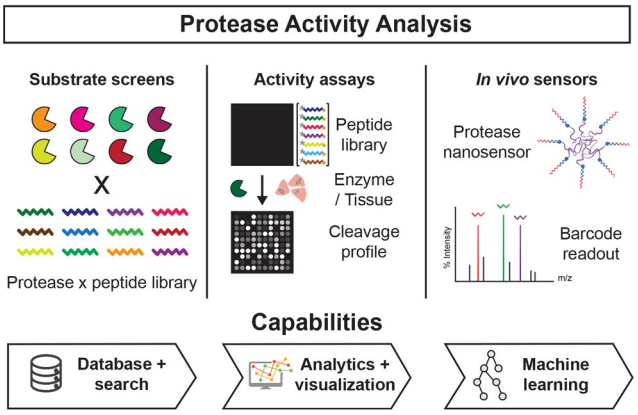

Figure 1. Overview of Protease Activity Analysis (PAA). The PAA package is designed to analyze data from large-scale substrate screens, enzyme activity assays, and *in vivo* enzyme activity sensors. Key package capabilities include searchable databases, where users can both query preloaded protease-substrate data sets published as part of PAA or import new data sets privately for their own use; data analytics and visualization functions, for facile and automated analysis of protease activity data; and machine learning models, for classification analysis of activity-based sensor data.

existing protease activity data, curated from over a decade of published works from our group, along with modular analytics that enable users to query these data sets for enzymes or substrates of interest. Through PAA's framework, users can additionally create and search new databases using their own protease-substrate screening data. PAA enables analytic standardization via functions that automate the quantification and visualization of user-input data from *in vitro* protease activity screens and *in vivo* protease-activated nanosensors. The package is accompanied by step-by-step tutorials that detail the functionalities provided by PAA in an open-source repository (https://github.com/apsoleimany/protease_activity_analysis). PAA's Python-based implementation provides a modular framework that is easy to interface with other software packages and can be readily integrated into broader data analytic workflows.

RESULTS

PAA provides scalable and modular analysis capabilities for data sets of enzyme activity measurements. Specifically, PAA supports three core analytic workflows (Figure 1): (1) analysis and query of databases of peptide substrate sequences and their cleavage susceptibilities; (2) analysis of substrate screens using recombinant enzymes or biospecimens; and (3) analysis of measurements from protease-responsive *in vivo* nanosensors. Across all three workflows, PAA provides preprocessing, visualization, machine learning, and search functionalities.

PAA Supports Searchable Enzyme−Substrate Databases. Identifying and characterizing which substrates are robustly and specifically cleaved by proteases of interest, such as those that are overexpressed in a specific cancer, is critical to discovery and engineering efforts that seek to understand and exploit protease activity. Indeed, the rapid rise and promise of engineered conditionally activated diagnostics and therapeutics, which almost universally incorporate a protease-cleavable peptide linker as the "trigger" for disease-specific activation, has motivated the need for tools and methods that identify synthetic peptide substrates that are maximally cleaved in diseased tissues and/or by target proteases. To this end, in PAA we present a *SubstrateDatabase* data structure that provides a facile framework for curating and querying data sets of enzyme−substrate activity, which often take the form of fluorometric assays of proteolytic cleavage of synthetic, fluorogenic peptide substrates. These assays measure the kinetics of enzyme activity over time and can be used to assess both the efficiency of an enzyme for a particular substrate, by quantifying the initial rate of the reaction, as well as the specificity of a substrate for a protease, by comparing the substrate's cleavage against other proteases screened.

To demonstrate these capabilities, we have created a publicly available database that incorporates data generated by our group from six independent recombinant protease screens against fluorogenic peptide substrates. The database consists of 150 unique synthetic peptide substrates and their cleavage susceptibilities across a set of 77 distinct recombinant proteases spanning the metallo, serine, cysteine, and aspartic catalytic classes. The substrates published as part of PAA were identified based on the literature and designed to query the activity of disease-associated proteases, including those in cancer, infection, and thrombosis. As such, there is an overrepresentation of metallo- and serine-sensitive substrates in the data set, which is open-sourced as a part of PAA (Figure 2B).[13,32] However, users can also import their own data into PAA for individual use, instantiating *SubstrateDatabase* data structures that can be readily queried and analyzed.

A guide to the core analytic and visualization functions related to the database can be found at https://github.com/apsoleimany/protease_activity_analysis/tree/master/tutorials. This step-by-step guide showcases how to load and query the protease-substrate database that is published with this work. Briefly, to instantiate a *SubstrateDatabase*, the user inputs raw data matrices of activity measurements (i.e., $n \times k$ where n is the number of substrates screened, and k is the number of conditions, e.g., proteases, assayed) for screens to be included in the database; a file that maps substrate names or labels to their corresponding sequences; as well as an optional file that maps amino acids to different colors based on properties of interest (e.g., hydrophobicity, chemistry, and identity) (Figure 2A). The *SubstrateDatabase* object first identifies overlapping substrates or proteases across multiple screens and aggregates all the data available for each unique substrate and protease into one simple data structure. In this way, protease-substrate activity assay data for proteases, substrates, or sequences of interest can be easily and efficiently queried. For example, if a user wants to identify potential substrates for a specific protease, they can input the protease name, and PAA will output a ranked list of substrates predicted to be efficiently and specifically cleaved by the protease of interest. The predictive

https://doi.org/10.1021/acsomega.2c01559
ACS Omega 2022, 7, 24292−24301

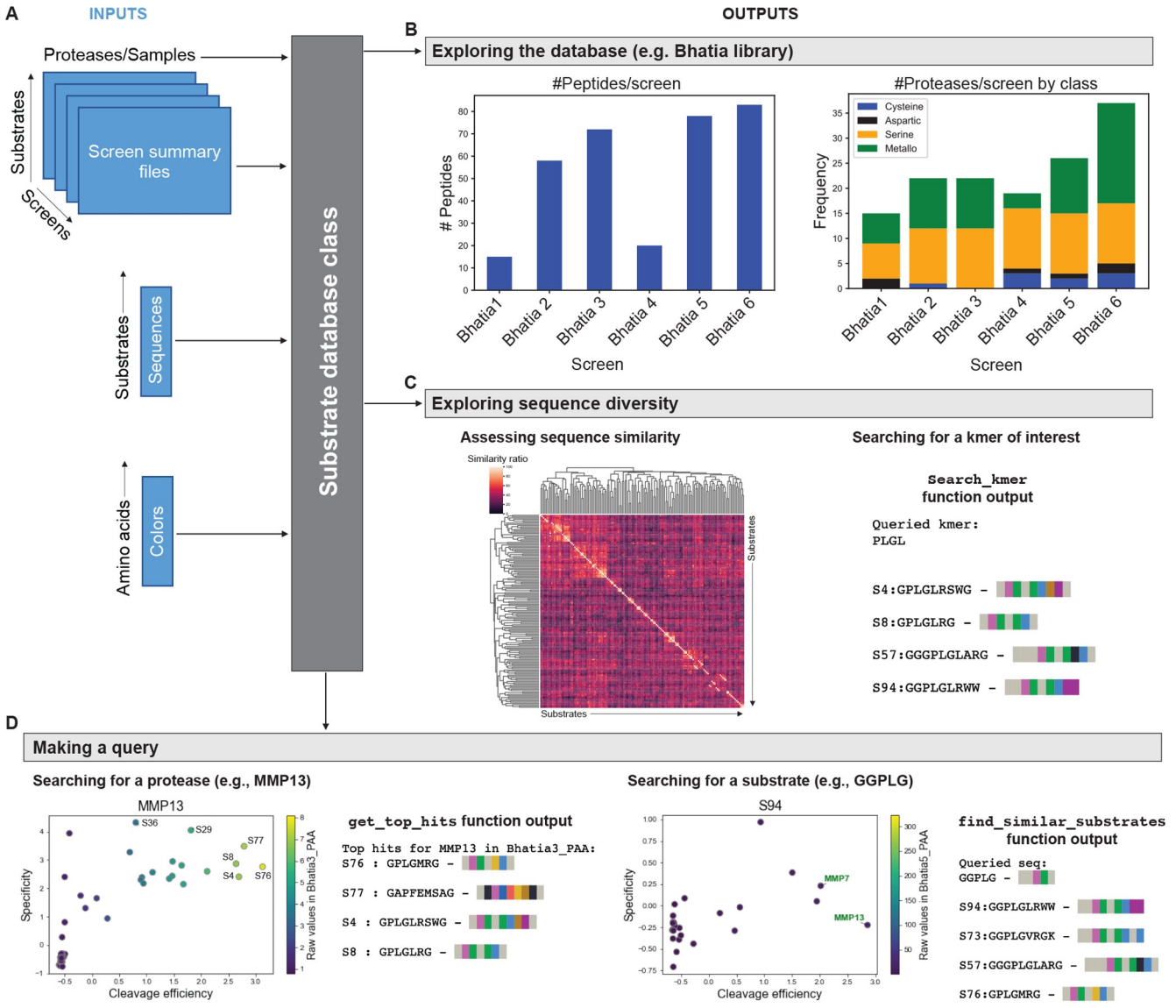

Figure 2. PAA provides an infrastructure for queryable databases of enzyme substrates. (A) *In vitro* activity screen summary files, a substrate sequence file, and an amino acid color map file provide data inputs for a *SubstrateDatabase*. (B) Sample use of the *SubstrateDatabase* to query a database of 150 unique substrate sequences screened against a diverse set of recombinant proteases. Summary plots of number of substrates and proteases across six independent screens comprising the database. (C) Metrics of sequence diversity include hierarchical clustering of pairwise sequence similarity scores as well as the ability to search for *k*-mers of interest. (D) Sample outputs of querying the database for a protease of interest (e.g., MMP13) and a sequence or cleavage motif of interest (e.g., "GGPLG").

rankings output by PAA strongly align with empirical results (Figure S1), lending strength to the power of PAA to help optimize protease activity experiments using *in silico* methods. Similarly, given a substrate as the user query, PAA can identify proteases that have been shown to robustly and/or specifically cleave that substrate. Note that, for the public data set published as part of PAA, substrate names correspond to names assigned by our group for specific sequences. These names map to existing nomenclature from previously published works for easy reference.[12,13,32]

Despite the fact that PAA contains cleavage data for a large number of substrates, in many cases the user will have a query sequence of interest that is not already included in the database. In the absence of an exact match, the *SubstrateDatabase* can retrieve the top-*k* substrates most similar to the query sequence, as quantified by different sequence similarity metrics. PAA offers two different sequence similarity metrics: the Levenshtein distance similarity ratio and the partial ratio. The former is strictly based on the Levenshtein distance that can be computed as the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one amino acid sequence into another. The partial ratio metric works similarly but instead takes the shortest sequence and compares it with all substrings of the same length. This partial ratio is particularly useful when two substrates contain the same amino acid cleavage motif (e.g., "PLG") but are flanked by different spacers at the N- or C-terminus (e.g., "GG" or "GS" spacers), as they will still be assigned high similarity estimates. After calculating the similarity between the user's input sequence, PAA returns the *k* most similar sequences and the values of the similarity metrics (Figure 2D).

https://doi.org/10.1021/acsomega.2c01559
ACS Omega 2022, 7, 24292−24301

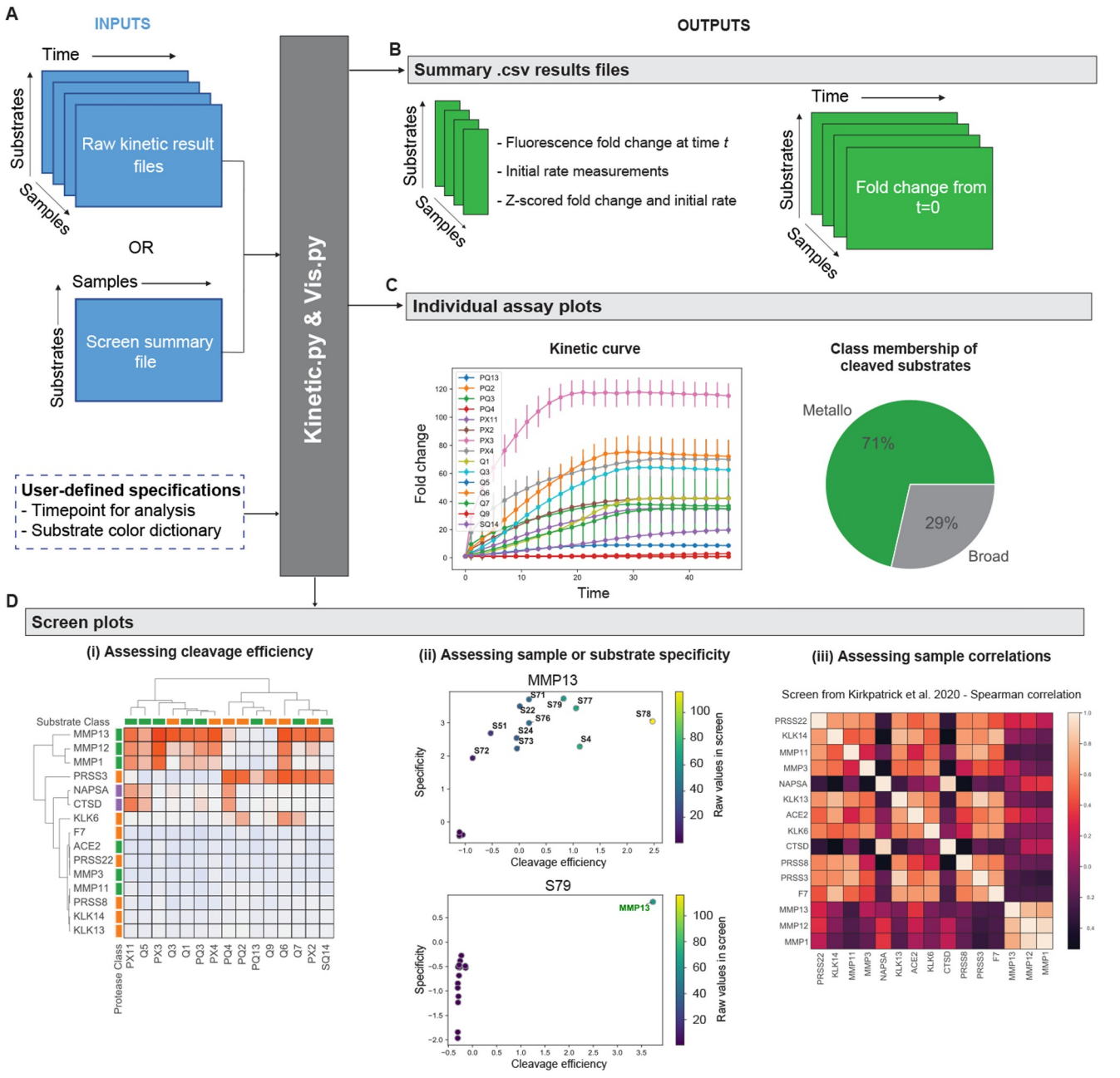

Figure 3. PAA automates analysis of *in vitro* assays of protease activity. (A) Fluorescence intensity measurements, together with user-defined specifications such as time points for analysis, are provided as inputs for construction of a *KineticDataset* . (B) *KineticDataset* automatically generates and saves key output activity measurements, such as initial activity rates and fold increases in substrate turnover as a function of time. (C) Retrospective analysis of 15 lung-cancer-associated proteases screened against a panel of 14 Forster resonance energy transfer (FRET)-paired substrates,[13] with representative plots for MMP13 shown, including line plots of fold change intensity over time for each substrate and a pie chart summarizing substrate cleavage susceptibility. (D) For the same study,[13] comprehensive assessment of cleavage efficiency and specificity across recombinant proteases and substrates. (i) Fluorescence fold changes were subject to hierarchical clustering to cluster proteases (vertical) by their substrate specificities and substrates (horizontal) by their protease specificities. (ii) Specificity versus efficiency (SvE) plots compare standard scores across substrates (efficiency; *x*-axis) against standard scores across proteases (specificity; *y*-axis). SvE analysis for the protease MMP13 shows promiscuous activity across substrates (top). SvE analysis for the substrate S79 highlights that it is specifically cleaved by MMP13 relative to other proteases assayed (bottom). (iii) Pairwise correlation analysis of initial rates across all substrates for recombinant proteases in the screen, measured as the Spearman rank correlation coefficient. Heatmap identifies the highest correlation of substrate cleavage between MMP1 and MMP12, among all proteases in the analyzed data set.[13]

Furthermore, the database also incorporates informative metrics on sequence diversity across substrates (Figure 2C). Such estimates can be very useful during library optimization to characterize the degree of redundancy and orthogonality between substrates in a given peptide library. Alternative metrics have been recently described by others to achieve similar goals.[33] To this end, PAA incorporates the function *get_similarity_matrix* that performs hierarchical clustering of pairwise similarity scores between all substrate sequences in the database and affords a compact visualization of sequence

https://doi.org/10.1021/acsomega.2c01559
ACS Omega 2022, 7, 24292−24301

diversity. In addition, the *search_kmer* function allows the user to readily find all substrates in the data set that contain a given *k*-mer motif of interest, such as the metalloprotease cleavage motif "PLGL" (Figure 2C). By integrating data from both of these analyses, PAA can help guide library optimization by allowing the user to make inferences about clusters of substrates that may have similar protease cleavage susceptibilities based on similarity scores and known cleavage motifs. Additionally, PAA can identify substrates that rank most distinct from others in the library and thus may be favored or disfavored based on the application at hand.

All recombinant proteases included in the database are of human origin. However, because overlapping cleavage sites have been identified between protease orthologs,[34] we anticipate that the human sequences in the database will prove useful to researchers studying proteases in model organisms. To enable users to search for orthologous protease genes across species, PAA includes a function, *species_to_species*, that builds off of the comprehensive "Mammalian Degradome Database"[35] to facilitate mapping of genes across species of interest (human, chimpanzee, mouse, and rat). For instance, the function will map the human protease "*MMP9*" to its mouse ortholog "*Mmp9*" while alerting the user that human "*GZMH*" does not have a mouse ortholog.[36] Furthermore, while PAA features data from our group published as a resource and example for the *SubstrateDatabase* data structure, users can use PAA as a local package to analyze their own data sets, which will keep all data private to the user. Then, users can implement PAA's functionalities and modular methods to analyze their private or internally generated data sets, as well as other data sets of interest.

PAA Provides Modular Methods to Analyze *In Vitro* Substrate Screens. As highlighted by the rich information that could be harnessed from PAA's public data set of protease−substrate activity measurements (Figure 2), *in vitro* enzyme activity assays are vital to characterizing protease activities and their dysregulation in disease.[1,10,37] Despite the broad prevalence of these assays throughout the protease biology and chemical biology communities, as well as the consideration that such assays will only become larger in size as high-throughput screening becomes more common, data analysis remains, to the best of our knowledge, largely manual. There is a dearth of computational tools to automate the analysis and visualization of enzyme activity data sets. PAA introduces a way to represent, store, and analyze these data sets automatically through the *KineticDataset* data object, which contains a suite of functions for rapidly preprocessing, visualizing, and analyzing these data (Figure 3).

The *KineticDataset* class is equipped to take in raw data files from enzyme activity assays (e.g., cleavage of fluorogenic substrates) generated directly by measurement instruments (e.g., fluorimeters; Figure 3A, Figure S2). Raw files consist of matrices of activity measurements for each sample to be analyzed (i.e., $n \times t$, where n is the number of substrates screened and t is the number of time points recorded). The class automatically generates key output activity measurements, including initial rates (intensity/min^{-1}) and fold changes at user-defined time points across substrates (Figure 3B). The resulting measurements can then be visualized with line plots that depict changes in raw fluorescence intensity and fold change intensity over time for each substrate (Figure 3C). Furthermore, users can define the catalytic class of each screened protease (e.g., metallo- versus serine-), to visualize the cleavage susceptibilities of their substrates by different protease classes (Figure 3C). This may prove useful if a certain protease class is known to be associated with a particular disease state, such as metalloproteases and cancer.

PAA also supports inputs from retrospective screens, for which a matrix summarizing cleavage efficiencies across a set of samples may have already been produced (Figure 3A). Based on these summary matrices, PAA can be used to cluster samples (e.g., recombinant enzymes, cell, or tissue lysates) of interest based on their activity patterns; to identify substrates that are cleaved with increased specificity by a given sample; and to examine correlations in cleavage patterns across screened samples (Figure 3D). In particular, specificity versus efficiency analyses ("SvE" plots) enable identification of optimal protease−substrate pairs from *in vitro* activity assays (Figure 3D). SvE plots are generated by calculating z-scores across the screened substrates, which serve as a surrogate metric for cleavage efficiency, and z-scores across the screened proteases, which serve as a surrogate metric for specificity. By plotting these metrics against each other, optimal substrate−protease pairs can be rapidly identified from large sets of screening data by identifying substrates that score high for both of these metrics (Figure 3D).[18] In addition, annotation by the raw activity measurement values for each protease−substrate pair reflects the absolute cleavage rate of a substrate of interest and overcomes the relative nature of standard scoring. Altogether, these analyses enable rapid assessment of substrate cleavage efficiency and specificity as well as robust identification of differential or overlapping activity signatures across different enzymes or tissue types (Figure 3D).

A step-by-step tutorial of the core data input, processing, visualization, and analytic functions can be found at https://github.com/apsoleimany/protease_activity_analysis/tree/master/tutorials. The demonstrations and the results presented in Figure 3 feature a retrospective analysis of a previously reported *in vitro* screen of a panel of lung-cancer-associated proteases against a panel of 14 peptide substrates[13] (Figure 3). We showcase the modularity of these functionalities through analysis of a second independent *in vitro* protease screen from the literature[33] (Figure S2), demonstrating that PAA extends to data sets from different experimental setups and research groups.

Together, *KineticDataset* and the visualization functionality provided by PAA streamline the aggregation, visualization, and analysis of *in vitro* activity measurements. In particular, these analyses facilitate the assessment of cleavage efficiency and specificity as well as the identification of differential and overlapping activity signatures among different enzymes or tissue types.

PAA Enables Machine Learning Analysis of *In Vivo* Activity Data. Because proteases play critical functional roles in a variety of disease processes, recent years have seen the emergence of new classes of activity-based diagnostics that are engineered to measure the activity of endogenous enzymes at the site of disease and to generate an output signal that can be read out externally.[11,38] To this end, our group has developed activity-based nanosensors, probes that detect the activity of aberrant proteases within the body and generate exogenous urinary reporters that reflect the degree of proteolytic cleavage encountered *in vivo*.[12−14,16,18,39−43] These nanosensors consist of an inert scaffold whose surface is decorated with peptide substrates, designed to be cleaved by proteases dysregulated in the disease state of interest. Each substrate is marked with a

https://doi.org/10.1021/acsomega.2c01559
ACS Omega 2022, 7, 24292−24301

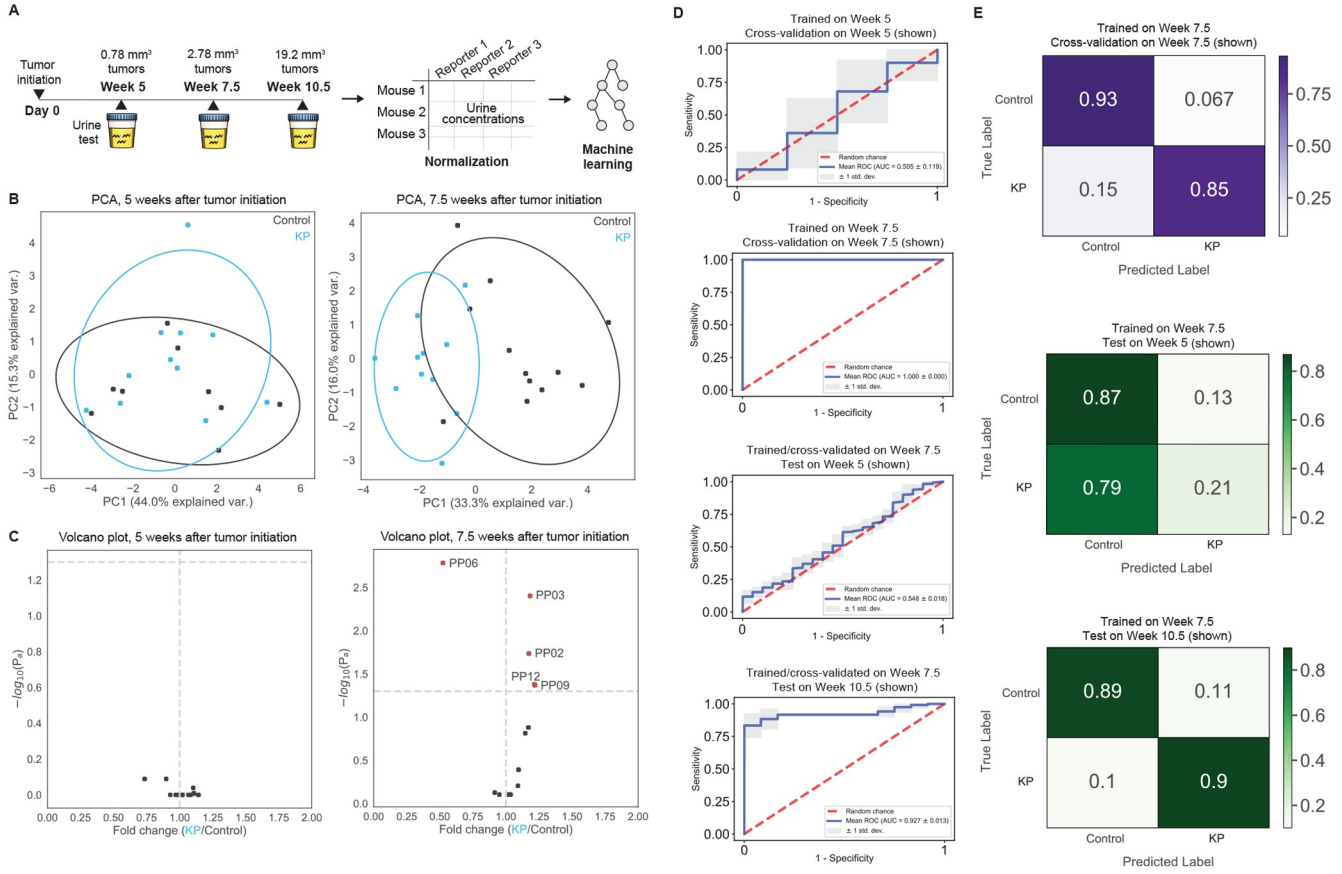

Figure 4. PAA enables automated machine learning analysis of *in vivo* activity data from activity-based nanosensors. (A) In a previously published study, activity-based nanosensors were administered at three different time points after tumor initiation in a mouse model of lung adenocarcinoma.[13] Dysregulated protease activity in the cancerous lungs triggered the release of mass-encoded reporters into the urine. The urinary reporter concentrations were measured with mass spectrometry. PAA enables analysis, visualization, and machine learning on these data. (B,C) PAA automates analysis to visualize differences in reporter enrichment among conditions, such as different sample classes, e.g., wild-type control (Control) and lung cancer (KP) mice, and time points, e.g., 5 and 7.5 weeks after tumor initiation in KP mice. (B) Principal component analysis (PCA) can reduce the dimensionality of the feature space to discover differential activity signatures across conditions. (C) Volcano plots identify nanosensors driving these signatures, by comparing the fold change in reporter concentrations between two classes (x-axis) against their statistical significance ($-\log_{10}(P_{adj})$; y-axis). (D) PAA evaluates the diagnostic potential of these activity signatures through automated training, validation, and testing of machine learning models, for example on the classification of healthy control and KP lung cancer mice. (E) Multiclass classifiers can also be trained, tested, and visualized using PAA.

mass-encoded peptide barcode, which is released upon interaction of the nanosensor with target proteases and then concentrated in the urine. Upon collection of urine, the relative concentrations of each reporter are quantified using mass spectrometry. Multiple sensors can be multiplexed simultaneously by barcoding each unique peptide substrate with a different mass-encoded peptide reporter.[5,12−14,18,16,32] This multiplexing generates $n \times k$ matrices of urinary reporter measurements, where n is the number of samples and k is the number of sensors/reporters. The reporters serve as input features, and thusPAA automates data analysis of these input matrices and enables training of downstream machine learning classifiers.

Considering the flexibility afforded by this approach, we implemented a modular data framework, *SyneosDataset* for analysis and machine learning on these mass-barcoded reporter measurements. Our framework provides a variety of capabilities directly tied to biological, diagnostic, and analytic questions of interest. These capabilities include differential enrichment analysis of reporters between conditions (i.e., identifying which reporters are associated with disease versus healthy states); unsupervised dimensionality reduction; binary and multiclass classification; feature, sample, and data specification for all analyses; recursive feature elimination; and associated visualizations. To demonstrate these capabilities, we created a step-by-step guide, published as an implementation tutorial, that details input data requirements, analytic functions, and visualization options. This guide recapitulates the findings of previously published work demonstrating the noninvasive detection of localized lung cancer in mice using a 14-plex activity-based nanosensor panel.[13] For the original study, the authors analyzed the *in vivo* data using unique scripts created specifically for their analysis. In our demonstration, we show how the same raw data can be parsed, analyzed, and visualized using the modular functions available in PAA (Figure 4). We developed modular machine learning functions that enabled the creation of additional diagnostic classifiers (Figure 4), thus demonstrating how PAA can be used to derive new insights from existing data.

Briefly, in the original study, a 14-plex nanosensor panel was administered into a mouse model of lung adenocarcinoma at 5, 7.5, and 10.5 weeks after tumor initiation and in parallel

https://doi.org/10.1021/acsomega.2c01559
ACS Omega 2022, 7, 24292−24301

healthy controls (Figure 4A). After collecting the urine from each mouse, the urinary reporter concentrations were quantified using mass spectrometry. In the original study, the authors sought to determine the earliest stage at which the nanosensor panel could detect lung cancer. PAA streamlined normalization, statistical analysis, and machine learning of the nanosensor data into a single computational pipeline. We used this pipeline to verify that PAA's modular workflow could recapitulate the original findings. PAA automated dimensionality reduction, specifically principal component analysis (PCA), to compare the urinary signals from each disease state across the tested time points (Figure 4B), and differential enrichment analysis to identify significant reporters (Figure 4C). In the example, one PCA plot (5 weeks after tumor initiation) shows an overlap between the two conditions, reflected in the volcano plot without any reporters being significantly differentially enriched (Figure 4B). In contrast, the PCA plot showing separation between clusters (7.5 weeks after tumor initiation) corresponds to differentially enriched reporters that drive the separation between groups, as reflected in the corresponding volcano plot (Figure 4C). With PAA, all graphs can be generated using a single function call in one line of code, making such analyses easily accessible, automated, and standardized.

The reporter concentrations can then be used to train binary and multiclass machine learning classifiers that can be used to diagnose disease (Figure 4D,E). PAA is capable of performing classification using a variety of algorithms, including support vector machines, random forests, and linear regression. This allows the user to benchmark methods rigorously and determine the best statistical learning method for their data. The user can also specify which sets of reporters, class labels, or individual labels should be used to train and test the classifiers. In the demonstration, we have shown that the reporter concentrations collected 5 weeks after tumor initiation are unable to yield a learned representation indicative of lung cancer, whereas by 7.5 weeks, the activity-based nanosensors generate an activity dataset that can be used to accurately diagnose lung cancer (Figure 4D,E). More generally, PAA provides a modular, streamlined data analytic workflow for measurements from *in vivo* protease nanosensors and can readily be applied to new data sets for automated statistical and machine learning analysis.

DISCUSSION

PAA advances computational methods to accelerate data analysis in biochemistry, chemical biology, and bioengineering. PAA represents a toolkit for users to automate the analysis of protease activity measurements generated *in vitro* through substrate screens or *in vivo* through noninvasive enzyme activity sensors. Here, we focus on the analysis of screens against synthetic, fluorescent-quenched peptide substrates (for the former) and of urinary reporter measurements from activity-based nanosensors (for the latter).

However, the modular methods and concepts presented in PAA readily extend to other data sets, particularly in terms of the tools for analysis of substrate screening data, as shown in Figure S2. Additional analytic functions for protease−substrate screening data, such as modeling time to cleavage saturation, prediction of competitive interactions between pairs of peptides,[44] deconvolution of signals from mixtures of proteases,[45] and identification of optimal substrate sets using principles from information theory,[33] will expand PAA's abilities to automate optimal enzyme substrate selection and design. Future work could extend PAA's machine learning capabilities to include neural network models for classification analysis,[46] methods for assessment of distribution shift and data set bias,[47−49] as well as approaches for quantification of predictive confidence.[50−55]

Not only does PAA contain valuable analysis tools, but it also includes a publicly available database of 150 unique synthetic peptides and their cleavage susceptibilities across a set of 77 distinct recombinant proteases, together with an interface to query this database for proteases, substrates, or sequences of interest. PAA's database can be readily expanded through publication and addition of new protease data, for example through high-throughput screening efforts that expand its coverage to additional enzymes. Users can use PAA as a local package to upload and analyze their own data sets, keeping all data private and leveraging PAA's functionalities to query and analyze their data. PAA's modular database functionality and public dataset could be of great interest in the context of nomination of protease-cleavable peptide linkers, for example for protease-activated diagnostics and therapeutics. By focusing on synthetic substrates that directly measure protease activity and providing modular data science functionalities through an accessible software package, PAA's database and analytic capabilities directly complement existing tools for assessing protease cleavage patterns.[29−31,45] Being implemented and released as a Python package, PAA can be further developed and integrated into larger data analytic workflows. We envision that PAA will accelerate analysis workflows for biologists, biochemists, and engineers interested in understanding and leveraging protease activity to better understand, detect, and treat disease.

METHODS

PAA's core relies on *NumPy*, *SciPy*, *Matplotlib*, *pandas*, *seaborn*, and *scikit-learn*. The Python-based implementation allows for flexible use, easy interfacing to machine learning and data analytic packages, and object-oriented programming. PAA's open-source code is available at https://github.com/apsoleimany/protease_activity_analysis and is published under the MIT license. PAA is organized and built as a package for ease of use and to facilitate developer integration.

The demonstrations described in this work are stored as Jupyter notebooks available in the PAA repository. These include: (1) querying and analysis of protease−substrate databases; (2) analysis and aggregation of *in vitro* screens of recombinant proteases and tissue lysates against synthetic peptide substrates; and (3) analysis and machine learning classification of urinary reporter signatures from *in vivo* activity-based nanosensors. The data sets used in these demonstrations were generated by our research group and are published together with PAA.

All code was written in the Python programming language. The PAA package is compatible with Mac OS, Windows, and Linux operating systems.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c01559.

Code availability; Figure S1: Empirical validation of predictions for substrate ranking; Figure S2: Automated

https://doi.org/10.1021/acsomega.2c01559
ACS Omega 2022, 7, 24292−24301

analysis of protease activity data from the literature (PDF)

AUTHOR INFORMATION

Corresponding Authors

Ava P. Soleimany − *Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts 02139, United States; Program in Biophysics, Harvard University, Boston, Massachusetts 02115, United States; Microsoft Research New England, Cambridge, Massachusetts 02142, United States;* orcid.org/0000-0002-8601-6040; Email: avasoleimany@microsoft.com

Sangeeta N. Bhatia − *Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts 02139, United States; Department of Electrical Engineering and Computer Science, MIT, Cambridge, Massachusetts 02139, United States; Howard Hughes Medical Institute, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-1293-2097; Email: sbhatia@mit.edu

Authors

Carmen Martin-Alonso − *Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts 02139, United States*

Melodi Anahtar − *Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-3475-3871

Cathy S. Wang − *Department of Biological Engineering, MIT, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-0910-4562

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c01559

Author Contributions

¶A.P.S., C.M.-A., and M.A. contributed equally. Conceptualization: A.P.S., C.M.A., and M.A. Data curation: A.P.S., C.M.A., M.A., and C.S.W. Formal analysis: A.P.S., C.M.A., M.A., and C.S.W. Investigation: A.P.S., C.M.A., M.A., and C.S.W. Methodology: A.P.S., C.M.A., and M.A. Project administration: A.P.S., C.M.A., M.A., and S.N.B. Resources: S.N.B. Software: A.P.S., C.M.A., M.A., and C.S.W. Supervision: A.P.S., C.M.A., M.A., and S.N.B. Validation: A.P.S., C.M.A., M.A., and C.S.W. Visualization: A.P.S., C.M.A., and M.A. Writing − original draft: A.P.S., C.M.A., and M.A. Writing − review and editing: A.P.S., C.M.A., M.A., C.S.W., and S.N.B.

Notes

The authors declare the following competing financial interest(s): S.N.B. reports compensation for cofounding, consulting, or board membership in Glympse Bio, Satellite Bio, CEND Therapeutics, Catalio Capital, Intergalactic Therapeutics, Port Therapeutics, Vertex Pharmaceuticals, and Moderna and receives sponsored research funding from Johnson and Johnson, Revitope, and Owlstone.

ACKNOWLEDGMENTS

The authors thank H. Fleming, L. Hao, N.-S. Harzallah, and H. Ko for their thoughtful feedback, pilot testing, and valuable discussions on this work. A.P.S. acknowledges support from the NIH Molecular Biophysics Training Grant NIH/NIGMS T32 GM008313 and the National Science Foundation Graduate Research Fellowship. C.M.A. acknowledges support from "La Caixa" Foundation Postgraduate Fellowship Abroad. M.A. acknowledges support from the National Science Foundation Graduate Research Fellowship. C.S.W. acknowledges support from the National Science Foundation Graduate Research Fellowship. S.N.B. is a Howard Hughes Medical Institute Investigator.

REFERENCES

(1) López-Otín, C.; Bond, J. S. Proteases: multifunctional enzymes in life and disease. *J. Biol. Chem.* **2008**, *283*, 30433−30437.

(2) Withana, N. P.; Garland, M.; Verdoes, M.; Ofori, L. O.; Segal, E.; Bogyo, M. Labeling of active proteases in fresh-frozen tissues by topical application of quenched activity-based probes. *Nat. Protoc.* **2016**, *11*, 184−191.

(3) Ivry, S. L.; Sharib, J. M.; Dominguez, D. A.; Roy, N.; Hatcher, S. E.; Yip- Schneider, M. T.; Schmidt, C. M.; Brand, R. E.; Park, W. G.; Hebrok, M.; et al. Global protease activity profiling provides differential diagnosis of pancreatic cysts. *Clin. Cancer Res.* **2017**, *23*, 4865−4874.

(4) Soleimany, A. P.; Kirkpatrick, J. D.; Su, S.; Dudani, J. S.; Zhong, Q.; Bekdemir, A.; Bhatia, S. N. Activatable zymography probes enable in situ localization of protease dysregulation in cancer. *Cancer Res.* **2021**, *81*, 213−224.

(5) Soleimany, A. P.; Kirkpatrick, J. D.; Wang, C. S.; Jaeger, A. M.; Su, S.; Naranjo, S.; Zhong, Q.; Cabana, C. M.; Jacks, T.; Bhatia, S. N. Multiscale profiling of enzyme activity in cancer. *bioRxiv* **2021**.

(6) Aung, A.; Cui, A.; Soleimany, A. P.; Bukenya, M.; Lee, H.; Cottrell, C. A.; Silva, M.; Kirkpatrick, J. D.; Amlashi, P.; Remba, T.; et al. Spatially regulated protease activity in lymph nodes renders B cell follicles a sanctuary for retention of intact antigens. *bioRxiv* **2021**, DOI: 10.1101/2021.11.15.468669.

(7) Weissleder, R.; Tung, C.-H.; Mahmood, U.; Bogdanov, A. In vivo imaging of tumors with protease-activated near-infrared fluorescent probes. *Nat. Biotechnol.* **1999**, *17*, 375−378.

(8) Jiang, T.; Olson, E. S.; Nguyen, Q. T.; Roy, M.; Jennings, P. A.; Tsien, R. Y. Tumor imaging by means of proteolytic activation of cell-penetrating peptides. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17867−17872.

(9) Blum, G.; Von Degenfeld, G.; Merchant, M. J.; Blau, H. M.; Bogyo, M. Noninvasive optical imaging of cysteine protease activity using fluorescently quenched activity-based probes. *Nat. Chem. Biol.* **2007**, *3*, 668−677.

(10) Sanman, L. E.; Bogyo, M. Activity-based profiling of proteases. *Annu. Rev. Biochem.* **2014**, *83*, 249−273.

(11) Soleimany, A. P.; Bhatia, S. N. Activity-based diagnostics: an emerging paradigm for disease detection and monitoring. *Trends Mol. Med.* **2020**, *26*, 450−468.

(12) 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. Mass-encoded synthetic biomarkers for multiplexed urinary monitoring of disease. *Nat. Biotechnol.* **2013**, *31*, 63−70.

(13) Kirkpatrick, J. D.; Warren, A. D.; Soleimany, A. P.; Westcott, P. M.; Voog, J. C.; Martin-Alonso, C.; Fleming, H. E.; Tammela, T.; Jacks, T.; Bhatia, S. N. Urinary detection of lung cancer in mice via noninvasive pulmonary protease profiling. *Sci. Transl. Med.* **2020**, DOI: 10.1126/scitranslmed.aaw0262.

(14) Kirkpatrick, J. D.; Soleimany, A. P.; Dudani, J. S.; Liu, H.-J.; Lam, H. C.; Priolo, C.; Henske, E. P.; Bhatia, S. N. Protease activity sensors enable real-time treatment response monitoring in lymphangioleiomyomatosis. *Eur. Respir. J.* **2022**, *59*, 2100664.

(15) Hao, L.; Rohani, N.; Zhao, R. T.; Pulver, E. M.; Mak, H.; Kelada, O. J.; Ko, H.; Fleming, H. E.; Gertler, F. B.; Bhatia, S. N. Microenvironment-triggered multimodal precision diagnostics. *Nat. Mater.* **2021**, *20*, 1440−1448.

(16) Cazanave, S. C.; Warren, A. D.; Pacula, M.; Touti, F.; Zagorska, A.; Gural, N.; Huang, E. K.; Sherman, S.; Cheema, M.; Ibarra, S.; et al. Peptide-based urinary monitoring of fibrotic nonalcoholic steatohe-

patitis by mass-barcoded activity-based sensors. *Sci. Transl. Med.* **2021**, 13.
(17) Bekdemir, A.; Tanner, E. E.; Kirkpatrick, J.; Soleimany, A. P.; Mitragotri, S.; Bhatia, S. N. Ionic liquid-mediated transdermal delivery of thrombosis-detecting nanosensors. *Adv. Healthcare Mater.* **2022**, *11*, 2102685.
(18) Anahtar, M.; Chan, L. W.; Ko, H.; Rao, A.; Soleimany, A. P.; Khatri, P.; Bhatia, S. N. Host protease activity classifies pneumonia etiology. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2121778119.
(19) Widen, J. C.; Tholen, M.; Yim, J. J.; Antaris, A.; Casey, K. M.; Rogalla, S.; Klaassen, A.; Sorger, J.; Bogyo, M. AND-gate contrast agents for enhanced fluorescence-guided surgery. *Nat. Biomed. Eng.* **2021**, *5*, 264−277.
(20) Bachovchin, D. A.; Brown, S. J.; Rosen, H.; Cravatt, B. F. Identification of selective inhibitors of uncharacterized enzymes by high-throughput screening with fluorescent activity-based probes. *Nat. Biotechnol.* **2009**, *27*, 387−394.
(21) Rut, W.; Groborz, K.; Zhang, L.; Sun, X.; Zmudzinski, M.; Pawlik, B.; Wang, X.; Jochmans, D.; Neyts, J.; Młynarski, W.; et al. SARS-CoV-2 M pro inhibitors and activity-based probes for patient-sample imaging. *Nat. Chem. Biol.* **2021**, *17*, 222−228.
(22) Desnoyers, L. R.; Vasiljeva, O.; Richardson, J. H.; Yang, A.; Menendez, E. E.; Liang, T. W.; Wong, C.; Bessette, P. H.; Kamath, K.; Moore, S. J. Tumor-specific activation of an EGFR-targeting probody enhances therapeutic index. *Sci. Transl. Med.* **2013**, DOI: 10.1126/scitranslmed.3006682.
(23) Kavanaugh, W. M. Antibody prodrugs for cancer. *Expert Opin. Biol. Ther.* **2020**, *20*, 163−171.
(24) Trang, V. H.; Zhang, X.; Yumul, R. C.; Zeng, W.; Stone, I. J.; Wo, S. W.; Dominguez, M. M.; Cochran, J. H.; Simmons, J. K.; Ryan, M. C.; et al. A coiled-coil masking domain for selective activation of therapeutic antibodies. *Nat. Biotechnol.* **2019**, *37*, 761−765.
(25) Millar, D. G.; Ramjiawan, R. R.; Kawaguchi, K.; Gupta, N.; Chen, J.; Zhang, S.; Nojiri, T.; Ho, W. W.; Aoki, S.; Jung, K.; et al. Antibody-mediated delivery of viral epitopes to tumors harnesses CMV-specific T cells for cancer therapy. *Nat. Biotechnol.* **2020**, *38*, 420−425.
(26) Li, F.; Leier, A.; Liu, Q.; Wang, Y.; Xiang, D.; Akutsu, T.; Webb, G. I.; Smith, A. I.; Marquez-Lago, T.; Li, J.; et al. Procleave: predicting protease-specific substrate cleavage sites by combining sequence and structural information. *Genomics, Proteomics Bioinf.* **2020**, *18*, 52−64.
(27) Song, J.; Wang, Y.; Li, F.; Akutsu, T.; Rawlings, N. D.; Webb, G. I.; Chou, K.-C. iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Briefings Bioinf* **2019**, *20*, 638−658.
(28) Li, F.; Wang, Y.; Li, C.; Marquez-Lago, T. T.; Leier, A.; Rawlings, N. D.; Haffari, G.; Revote, J.; Akutsu, T.; Chou, K.-C.; et al. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. *Briefings Bioinf* **2019**, *20*, 2150−2166.
(29) Rawlings, N. D.; Barrett, A. J.; Finn, R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **2016**, *44*, D343−D350.
(30) Fortelny, N.; Yang, S.; Pavlidis, P.; Lange, P. F.; Overall, C. M. Proteome TopFIND 3.0 with TopFINDer and PathFINDer: database and analysis tools for the association of protein termini to pre-and post-translational events. *Nucleic Acids Res.* **2015**, *43*, D290−D297.
(31) Uzozie, A. C.; Smith, T. G.; Chen, S.; Lange, P. F. Sensitive identification of known and unknown protease activities by unsupervised linear motif deconvolution. *Anal. Chem.* **2022**, *94*, 2244.
(32) Dudani, J. S.; Ibrahim, M.; Kirkpatrick, J.; Warren, A. D.; Bhatia, S. N. Classification of prostate cancer using a protease activity nanosensor library. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 8954−8959.
(33) Holt, B. A.; Lim, H. S.; Su, M.; Tuttle, M.; Liakakos, H.; Qiu, P.; Kwong, G. A. Embracing enzyme promiscuity with activity-based compressed biosensing. *bioRxiv* **2022**, DOI: 10.1101/2022.01.04.474983.
(34) Puente, X. S.; Sánchez, L. M.; Overall, C. M.; López-Otín, C. Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet.* **2003**, *4*, 544−558.
(35) Pérez-Silva, J. G.; Español, Y.; Velasco, G.; Quesada, V. The Degradome database: expanding roles of mammalian proteases in life and disease. *Nucleic Acids Res.* **2016**, *44*, D351−D355.
(36) Trapani, J. A. Granzymes: a family of lymphocyte granule serine proteases. *Genome Biol.* **2001**, *2*, 1−7.
(37) Dudani, J. S.; Warren, A. D.; Bhatia, S. N. Harnessing protease activity to improve cancer care. *Annu. Rev. Cancer Biol.* **2018**, *2*, 353−376.
(38) Kwong, G. A.; Ghosh, S.; Gamboa, L.; Patriotis, C.; Srivastava, S.; Bhatia, S. N. Synthetic biomarkers: a twenty-first century path to early cancer detection. *Nat. Rev. Cancer* **2021**, 1−14.
(39) Warren, A. D.; Kwong, G. A.; Wood, D. K.; Lin, K. Y.; Bhatia, S. N. Point-of-care diagnostics for noncommunicable diseases using synthetic urinary biomarkers and paper microfluidics. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 3671−3676.
(40) Kwon, E. J.; Dudani, J. S.; Bhatia, S. N. Ultrasensitive tumour-penetrating nanosensors of protease activity. *Nat. Biomed. Eng.* **2017**, *1*, 1−10.
(41) Loynachan, C. N.; Soleimany, A. P.; Dudani, J. S.; Lin, Y.; Najer, A.; Bekdemir, A.; Chen, Q.; Bhatia, S. N.; Stevens, M. M. Renal clearable catalytic gold nanoclusters for in vivo disease monitoring. *Nat. Nanotechnol.* **2019**, *14*, 883−890.
(42) Mac, Q. D.; Mathews, D. V.; Kahla, J. A.; Stoffers, C. M.; Delmas, O. M.; Holt, B. A.; Adams, A. B.; Kwong, G. A. Non-invasive early detection of acute transplant rejection via nanosensors of granzyme B activity. *Nat. Biomed. Eng.* **2019**, *3*, 281−291.
(43) He, J.; Nissim, L.; Soleimany, A. P.; Binder-Nissim, A.; Fleming, H. E.; Lu, T. K.; Bhatia, S. N. Synthetic Circuit-Driven Expression of Heterologous Enzymes for Disease Detection. *ACS Synth. Biol.* **2021**, *10*, 2231−2242.
(44) Leung, D.; Abbenante, G.; Fairlie, D. P. Protease inhibitors: current status and future prospects. *J. Med. Chem.* **2000**, *43*, 305−341.
(45) Miller, M. A.; Barkal, L.; Jeng, K.; Herrlich, A.; Moss, M.; Griffith, L. G.; Lauffenburger, D. A. Proteolytic Activity Matrix Analysis (PrAMA) for simultaneous determination of multiple protease activities. *Integr. Biol.* **2011**, *3*, 422−438.
(46) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.
(47) Lipton, Z.; Wang, Y.-X.; Smola, A. Detecting and correcting for label shift with black box predictors. *International Conference on Machine Learning* **2018**, 3122−3130.
(48) Amini, A.; Soleimany, A. P.; Schwarting, W.; Bhatia, S. N.; Rus, D. Uncovering and mitigating algorithmic bias through learned latent structure. *Proceedings of the 2019 AAAI/ACM Conference on AI*, Ethics, and Society 2019, 289−295.
(49) Koh, P. W.; et al. Wilds: A benchmark of in-the-wild distribution shifts. *International Conference on Machine Learning* **2021**, 5637−5664.
(50) Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning* **2016**, 1050−1059.
(51) Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems* **2017**, *30*, 6405−6416.
(52) Amini, A.; Schwarting, W.; Soleimany, A.; Rus, D. Deep evidential regression. *Advances in Neural Information Processing Systems* **2020**, *33*, 14927−14937.
(53) Hie, B.; Bryson, B. D.; Berger, B. Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Syst* **2020**, *11*, 461−477.
(54) Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. Evidential deep learning for guided molecular

https://doi.org/10.1021/acsomega.2c01559
ACS Omega 2022, 7, 24292−24301

property prediction and discovery. *ACS Cent. Sci.* **2021**, *7*, 1356−1367.
(55) Kompa, B.; Snoek, J.; Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ. Digit. Med.* **2021**, *4*, 1−6.

https://doi.org/10.1021/acsomega.2c01559
ACS Omega 2022, 7, 24292−24301